Machine Learning Approach and Adolescent Health

Implementation of Machine Learning Algorithms to Explore Adolescent Health, BMI and Weight Perception Using COMPASS Study

by

Qianyu Zhang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Public Health Sciences

Waterloo, Ontario, Canada, 2024

© Qianyu Zhang 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisor(s):	**Scott T. Leatherdale** Professor, School of Public Health Sciences University of Waterloo **Helen H. Chen** Professor, School of Public Health Sciences University of Waterloo
Internal Member(s):	**James Wallace** Associate Professor, School of Public Health Sciences University of Waterloo **Abel Torres Espin** Assistant Professor, School of Public Health Sciences University of Waterloo
Internal-External Member:	**Liqun Diao** Research Assistant Professor, Statistics and Actuarial Science University of Waterloo
External Examiner:	**Carolyn McGregor AM** Professor, Faculty of Business and Information Technology University of Ontario Institute of Technology

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

For the studies in Chapters 4, 5 and 6, I was responsible for developing the research question, data preparation and aggregation, statistical analysis, ML model development, and writing all sections of the manuscripts. Drs Chen, Espin, Leatherdale, and Wallace provided guidelines and editing suggestions, especially on data processing and model development.

Chapter 4 is planned to be submitted to the BMC Public Health Journal.

Chapter 5 is planned to be submitted to Frontiers in Public Health Journal.

Chapter 6 is planned to be submitted to the International Journal of Pediatrics.

Abstract

Introduction

Adolescence is critical for behaviour formation and physical, social, and psychological changes. The risk related to unhealthy behaviours occurring during this stage can emerge in early adulthood and could lead to adverse outcomes. The COVID-19 pandemic has significantly impacted the youth's characteristics and the school's learning environment. Exploring the profile of Canadian youth before and after COVID-19 helps understand the factors that influence student engagement in the multi-domain setting, including school activities, mental health, healthy eating, etc. Among the various aspects of student profile, childhood obesity and weight perceptions are two important topics for Canadian youth's health and well-being. Obesity in children is a complex public health concern in Canada. Almost 1 in 7 children are considered obese, and 75% of obese children are obese in their adulthood when they grow up. Additionally, a youth's perception of their weight is often connected to their BMI status. Studies have shown that adolescents who are overweight and obese tend to underestimate their true weight, and the ones with normal weight could overestimate their weight status. Both overweight and obese or misperception of body weight are risk factors for developing non-communicable diseases in adulthood and can lead to worse mental health issues, such as lower self-esteem. Therefore, observing these two aspects and related impactful factors is necessary for informing youth obesity preventive intervention. Compared to traditional statistics, machine learning has proven effective in handling the complex relationships between multidomain variables and in detecting nonlinear relationships between the predictor variables and the target outcome. This thesis explores behavioural patterns, specifically BMI and weight perception, among the Canadian youth participating in the COMPASS study.

Objective and Methods

The dissertation utilized Wave 7 (2018-19) and Wave 9 (2020-21), and linked data from Wave 9 to 11 (2020-23) of the COMPASS study. This ongoing survey-based cohort study includes grade 9-12 students attending secondary schools in Alberta, British Columbia, Ontario, and Quebec in Canada. The survey questions cover multiple aspects of student health behaviours, including eating, sedentary behaviour, alcohol, tobacco, and marijuana use, bullying, academic performance, physical activity, BMI, and school connectedness.

Study 1 applied k-means clustering analysis to establish student behaviour profiles, followed by a Random Forest model (RF) to identify the factors associated with the characteristics before and after COVID-19.

Among the various student-level behaviour factors, Study 2 focused on self-reported BMI by predicting youth BMI status and identified associated factors using six supervised machine learning classifiers, K-Nearest Neighbour, Logistic Regression, Support Vector Machine,

Random Forest, Multinomial Naïve Bayes, and Extreme Gradient Boosting (XGBoost) between Complete Case Analysis (CCA) and multiple imputation (MI) approaches.

Since weight perception was identified as the most significant factor associated with BMI prediction in Study 2, Study 3 used weight perception as the outcome variable. This study explored the transition pattern of student weight perception and the associated factors through the Markov Chain, Multinomial Logistic Regression, and time-series deep learning models, including Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). The Shapley Additive Explanations (SHAP) analysis was applied for all three studies to assess the model interpretability.

Results

Study 1 found that in both pre-COVID (2018-19) and post-COVID (2020-21) cohorts, clusters were distinguished as closely involved with substance use and lower psychological well-being, including personal relationships with their families or unwillingness to discuss their mental health with an adult at school. However, the number of students in these clusters was relatively low compared to the other clusters, comprising only 0.5% of the pre-COVID cohort and 0.2% of the post-COVID cohort.

In sub-clusters from the pre-COVID cohort that excluded substance use factors, students who skipped more than 20 classes, did not complete homework, and were involved in bullying were more likely to be grouped together in one sub-cluster compared to students without these behaviours. Additionally, those intending to lose weight, perceiving themselves as overweight,and being less active had higher odds of being in one sub-cluster compared to those without weight change intentions and who perceived their weight as about the right. In the sub-clusters of the post-COVID cohort, students with adverse mental health well-being, higher anxiety scores, self-reported overweight or obese BMI, and those failing to meet the guidelines of at least 60 minutes of physical activity per day were more likely to be grouped together in one sub-cluster compared to the students with the opposite behaviours. Moreover, students with increased cannabis use, less physical activity, lower flourishing scores and higher anxiety scores were more likely to be in one group relative to students who reported no change in cannabis use.

Study 2 discovered that, among both CCA and MI approaches, XGBoost achieved the best performance in predicting BMI compared to those using Support Vector Machine (SVM) and Random Forest (RF). In CCA, it achieved an overall accuracy and ROC-AUC value of 0.64 and 0.78. The model performance was similar in MI and yielded an overall accuracy of 0.64 and ROC-AUC value of 0.79. Based on the SHAP summary plot, the most important predictors were weight perception, gender, and weight change intention in all three BMI classes, and the result was consistent in both CCA and MI. According to the SHAP summary plot, the most significant predictors in all three BMI classes were further explained by weight perception, gender and

weight change intention. Students who perceived themselves as slightly or very overweight and were trying to lose weight were classified into the overweight and obese class, while model prediction on health weight BMI involved perceiving the right weight and not trying to change their weight. Additionally, gender was emerged as the most significant variable in the unknown BMI class, with model predicting that females and gender minorities were more likely to be categorized in the unknown weight class compared to male students.

Study 3 used the MI approaches on the linked dataset. It found that although there was a moderate probability of transitioning from underweight or overweight to about the right weight and a minor probability of changing from underweight to overweight, most youths were more likely to stay in the same weight perception as the previous year.

In addition, Study 3 identified predictors associated with weight perception transition over time. Students with underweight BMI, the intention of gaining weight and disagreement with discussing their problems with families in Year 1 had a higher likelihood of perceiving themselves as underweight in Year 2. While overweight and obese or unknown BMI students who had the intention of losing weight and decreased time on social media after COVID-19 in Year 1 were more likely to have an overweight perception in Year 2. Compared to peers who stayed at perceiving about the right weight in Year 2, underweight BMI students had higher odds of remaining underweight perception. In contrast, overweight BMI students with losing weight intention were likely to continue with an overweight perception and transition from about the right weight to an overweight perception. Students who wanted to gain weight were more likely to remain underweight or change from underweight to about the right weight. Notably, students with more participation in muscle-strengthening exercises had a lower likelihood of remaining overweight perception.

Both deep learning models have shown the top significant variables in Year 1 and Year 2 were weight change intention, BMI, skipped breakfast for losing weight, and days in muscle training exercises predicting third-year weight perceptions. The intention of losing weight, higher BMI status such as overweight and obese or unknown, and more days in skipping breakfast for losing weight predicted students towards an overweight perception. In contrast, among the right weight perception group, students tended to stay the same weight or not try to do any weight change, had healthy BMI, had a lower number of days skipping breakfast, and had more days in muscle training. These results were consistent with the weight perception transition-associated predictors from the Multinomial logistic regression.

Conclusion

As the first study that focused on Canadian youth behaviour profiles, prediction of BMI and weight perception incorporating multiple machine learning techniques, the result of the dissertation disseminates key messages to the stakeholders who want to understand student

behaviour profiles and focus on the preventive intervention on youth obesity. This research emphasized the importance of the school's environment on youth's behaviour and the tri-directional relationships among BMI, weight change intention, and weight perception, suggesting future analysis to emphasize these variables when developing healthy behaviour-related strategies. Overall, it illustrated the necessity of establishing educational programs related to BMI and weight perception at school to raise awareness of self-esteem and body image acceptance.

Acknowledgements

The breeze of the wind flew through the shadow of the leaves.

The shining sun gently spread the warmness around the geese and their children.

I walked past TJB with a coffee and a hoagie from Eby Street.

A sandwich filled with Philly-style roasted pork and melted cheddar with garlic mayo.

Blinking thoughts on getting the Earl Grey New York roll at Roux Bakehouse, chicken pot pie at Mom's Bread and Pie Co., or the back Mamba Roll at The Humble Lotus?

That is a good question!

I am happy and smiling to enjoy the amazing food with an appreciated heart while closing my PhD journey.

I am grateful for my two awesome supervisors, Scott and Helen. Scott, you have been supportive of my decisions since day one. Whenever I have an idea for my research, you are willing to listen and provide me with constructive feedback. Thank you so much for supporting me throughout this journey and assigning me the research assistant position, allowing me to be involved in publication. Helen, your HLTH 650 inspired the machine learning idea for my research. Thank you for always making time to chat with me when I have questions about the model implementation and results. Your advice on my dissertation was incredible and guided me throughout my research. You have been a starry light during this journey. I am grateful for the help of my committee members, Abel and Jim. Thank you so much, Abel, for your suggestions on model preprocessing and interpretation of clustering and deep learning models. Jim, thank you so much for taking the time to share your knowledge of data visualization while I was building my thesis.

I am grateful for the COMPASS team. Angie, Katie, Kai, and Jay, thank you so much for preparing the raw cohort datasets that allowed me to start my research and answered all my questions related to the data. Thank you so much, Mahmood, for taking the time to discuss my statistics question. My statistical analysis could not have been consolidated without your expertise. Thank you so much, Mohammad, for explaining to me the model preprocessing steps.

I am grateful for my parents, Wenjie and Yi. You have been supportive since I started high school in Massachusetts. Thank you so much for being such loving, understanding, and knowledge-sharing friends. I am grateful to my family, Olenka, Vesper, Winston, and Freya. Thank you so much for accompanying me and bringing so much joy to me since the beginning of this journey.

I am grateful for this PhD journey and the knowledge and experience I have obtained.

Table of Contents

Author's Declaration iii
Statement of Contributions iv
Abstract v
Acknowledgements ix
List of Figures xv
List of Tables xvii
List of Abbreviations xix
Chapter 1 1
Background 1
1.1 Adolescent health before and after COVID-19 1
1.2 Childhood Obesity 2
1.3 Childhood Obesity Risk Factors 3
1.4 Childhood Obesity Interventions 5
1.5 The COMPASS System 6
1.5.1 Measures in COMPASS 8
1.6 NHANES database 10
1.6.1 In-Home Interviews 11
1.6.2 Mobile Examination Center (MEC) Visit 11
1.7 Thesis Scope and Implication 12
1.8 Thesis Organization 13
Chapter 2 14
Analytic Approach 14
2.1 Traditional Statistical Analysis in Childhood Obesity 14
2.1.1 Drawbacks of Statistical Modelling 15
2.2 Why Machine Learning (ML)? 17
2.3 Types of Machine Learning Algorithms 17
2.4 Unsupervised Learning 18
2.4.1 K-Means Clustering 18
2.4.1.1 Determine the Number of K 19
2.4.1.2 Related Research Using K-Means Clustering 22

2.4.2 Hierarchical Clustering 23
2.5. Supervised Learning 24
2.5.1 Overview of Classification Algorithms 25
2.5.2 Supervised ML Research in Childhood Obesity 27
2.6 Deep Learning Model in Time Series Analysis 28
2.6.1 Overview of Deep Learning Models 29
2.6.2 DL Research in Childhood Obesity 32
2.7 Markov Chain Model (MCM) 33
Chapter 3 36
Study Rational and Research Questions 36
3.1 Study 1: Clustering Analysis of COMPASS Student's Behaviours in Pre-COVID and Post-COVID Cohort 36
3.1.1 Rationale 36
3.1.2 Research Questions 37
3.2 Study 2: Building an Interpretable Machine Learning Model for Selecting Modifiable Attributes of Childhood Obesity 37
3.2.1 Rationale 37
3.2.2 Research Questions 38
3.3 Study 3: Exploring the Longitudinal Weight Perception Transition and Associated Factors Using the Markov Chain Model and Deep Learning Approach 38
3.3.1 Rationale 38
3.3.2 Research Questions 39
Chapter 4 40
4.1 Abstract 40
4.2 Introduction 41
4.3 Materials and Methods 42
4.3.1 Study Dataset 42
4.3.2 Variables of Interest 42
4.3.3 Data Preprocessing 45
4.4 Results 47
4.4.1 Descriptive Statistics 47
4.4.2 K-means Clustering and Model Performance 53
4.4.3 High-Level Clustering Analysis Interpretation in 2018-19 and 2020-21 55
4.4.4.1 Pre-COVID (2018-19) 60
4.4.4.2 Post-COVID (2020-21) 70

4.5 Discussion 79
4.6 Conclusions 80
Chapter 5 82
Building an Interpretable Machine Learning Model for Selecting Modifiable Attributes of Childhood Obesity 82
5.1 Abstract 82
5.2 Introduction 83
5.3 Materials and Methods 85
5.3.1 Study Cohort 85
5.3.2 Predictors and Outcome 85
5.3.3 Data Preparation 86
5.3.4 Data Preprocessing 87
5.3.5 Machine Learning Model Development 88
5.4. Results 89
5.4.1 Baseline Characteristics 89
5.4.2 Model Performance 93
5.4.3 SHAP Analysis 96
5.5 Discussion 114
5.6 Conclusion 116
Chapter 6 117
Exploring the Longitudinal Weight Perception Transition and Associated Factors Using the Markov Chain Model and Deep Learning Approach 117
6.1 Abstract 117
6.2 Introduction 118
6.3 Materials and Methods 120
6.3.1 Study Cohort 120
6.3.2 Data Cleaning 120
6.3.3 Feature Selection for Predictors 121
6.3.4 Data Preprocessing 122
6.3.5 Markov Chain Model (MCM) and Statistical Analysis 122
6.3.6 Deep Learning Data Preprocessing and Architecture 123
6.3.7 Hyperparameters Tuning and Model Evaluation 125
6.3.8 Model Interpretation 126
6.4 Results 126
6.4.1 Baseline Characteristics 126

6.4.2 Transition Probabilities in Weight Perception 134
6.4.3 The Association Between Year 1 Predictors and Weight Perception in Year 2 136
6.4.4 The Association Between Year 1 Predictors and Weight Perception Transition from Year 1 to Year 2 139
6.4.5 Model Prediction on Year 3 Weight Perception 141
6.4.6 SHAP Analysis 143
6.5 Discussion 154
6.6 Conclusion 156
Chapter 7 157
Final Discussion 157
7.1 Study 1 Findings 157
7.1.1 Study 1 RQ1: What factors are associated with the clustering of student profiles in the 2018-19 (pre-COVID) and 2020-21 (post-COVID) COMPASS cohorts? 157
7.1.2 What are the similarities or differences of the factors associated with the clustering of student behaviour pre-COVID-19 and post-COVID-19? 158
7.2 Study 2 Findings 159
7.2.1 Study 2 Q1: What important factors are associated with youth BMI status in the 2020-21 COMPASS student cohort using ML classifiers from the CCA and MI approach? 160
7.2.2 Study 2 Q2: How do the results of ML-generated factors compare with the common factors leading to childhood obesity in the previous literature? What are the similarities and differences? 160
7.3 Study 3 Findings 162
7.3.1 Study 3 Q1: How do the estimated probabilities of weight perception change from Year 1 (2020-21) to Year 2 (2021-22) and from Year 2 (2021-22) to Year 3 (2022-23) using the Markov Chain Model? 162
7.3.2 Study 3 Q2: What factors are associated with the weight perception in Year 2 and the weight perception transitional change from Year 1 to Year 2? 163
7.3.3 Study 3 Q3: In the Deep Learning approach, what are the important factors associated with predicting weight perception in Year 3 from the previous two years? 164
7.4 Strengths 165
7.5 Limitations 167
7.6 Contribution to Public Health 170
7.7 Contribution to ML Research 171
7.8 Future Work 172
7.9 Conclusion 173

References 175
Appendices 199
Appendix A: The COMPASS Data Dictionary for Representative Domains 200
Appendix B: Study 1 Supplementary Materials 205
Appendix C: Study 2 Supplementary Materials 280
Appendix D: Study 3 Supplementary Materials 300
Appendix E: The COMPASS Student Questionnaire (2020-21) 323

List of Figures

Figure 2.6 RNN computation of training loss that shows the input state to output state **30**

Figure 5.1 SHAP beeswarm plots for health weight, overweight and obese, and unknown in CCA **99**

Figure 5.2 Boxplot of weight perception and gender SHAP values in each BMI status (CCA) **101**

Figure 5.3 SHAP Waterfall plots for each BMI Class **103**

Figure 5.4 SHAP summary plot for all BMI status in MI **105**

Figure 5.5 SHAP beewswarm plots for each BMI status in MI **106**

Figure 5.6 Boxplot of weight perception and gender SHAP values in each BMI status (MI) ... **108**

Figure 5.7 SHAP waterfall plots of each BMI status in MI **110**

Figure 5.8 SHAP beeswarm plots of overweight and obese females vs males **112**

Figure 5.9 SHAP beeswarm plots of unknown weight females vs males **113**

Figure 6.1 RNN structure **124**

Figure 6.2 LSTM structure **125**

Figure 6.3 Markov Chain Transition Probabilities of Weight Perception from Year 1 to Year 2 ..... **135**

Figure 6.4 Markov Chain Transition Probabilities of Weight Perception from Year 2 to Year 3 ..... **136**

Figure 6.5 SHAP summary plot for each weight perception status in RNN **145**

Figure 6.6 SHAP summary plot for each weight perception status in LSTM **148**

Figure 6.6.1 SHAP beeswarm plot for each weight perception status in LSTM **151**

Figure S1.2 (A-E) 2018-2019: SHAP summary plot for sub clusters in high-level cluster 0 .... **255**

Figure S1.3 (A-D) 2018-2019: SHAP summary plot for sub clusters in high-level cluster 2 .... **260**

Figure S1.4 (A-D) 2020-2021: SHAP summary plot for high-level cluster **264**

Figure S1.5 (A-E) 2020-2021: SHAP summary plot for sub clusters in high-level cluster 0 .... **268**

Figure S1.6 (A-C) 2020-2021: SHAP summary plot for sub clusters in high-level cluster 1 .... **273**

Figure S1.7 (A-D) 2020-2021: SHAP summary plot for sub clusters in high-level cluster 3 .... **276**

Figure S2.1 Top 50 feature important variables selected by Random Forest classifier **296**

Figure S2.2 Mean Convergence plot for MICE imputed iterations **297**

Figure S3.1 Year 1 variable's feature importance score **301**

Figure S3.2 Year 2 variable's feature importance score **302**
Figure S3.3 Year 3 variable's feature importance score **303**
Figure S3.4 Convergence plot of 20 iterations **313**
Figure S3.5 Convergence plot of 30 iterations **314**
Figure S3.6 Convergence plot of 40 iterations **314**
Figure S3.7 Convergence plot of 50 iterations **315**
Figure S3.8.1 Density plot of Year 1 on selected variables **316**
Figure S3.8.2 Density plot of Year 2 on selected variables **317**
Figure S3.8.3 Density plot of Year 3 on selected variables **318**
Figure S3.9 (A) – (C) Feature importance SHAP beeswarm plot for each weight perception status in RNN **320**

List of Tables

Table 4.1 Descriptive Statistics of COMPASS Study Cohort 2018-19 **47**
Table 4.2 Descriptive Statistics of COMPASS Study Cohort 2020-21 **50**
Table 4.3 Cohorts 2018-19 and 2020-21 K-means Clustering Result **53**
Table 4.4 Model Performance for Random Forest Model **54**
Table 4.5 High Level Cluster Year 2018-19 **56**
Table 4.5.1 Logistic Regression Analysis **57**
High Level Cluster 2018-19 **57**
Table 4.6. High Level Cluster Year 2020-21 **58**
Table 4.5.2 Cluster 0 Sub Cluster Year 2018-19 **61**
Table 4.5.2.1 Multinomial Logistic Regression Analysis **64**
Sub Cluster from Cluster 0 (2018-19) **64**
Table 4.5.3 Cluster 2 Sub Cluster Year 2018-19 **66**
Table 4.5.3.1 Multinomial Logistic Regression Analysis **69**
Sub Cluster from Cluster 2 (2018-19) **69**
Table 4.6.1 Cluster 0 Sub Cluster Year 2020-21 **72**
Table 4.6.1.1 Multinomial Logistic Regression Analysis **74**
Sub Cluster from Cluster 0 (2020-21) **74**
Table 4.6.2 Cluster 1 Sub Cluster Year 2020-21 **75**
Table 4.6.2.1 Multinomial Logistic Regression Analysis **76**
Sub Cluster from Cluster 1 (2020-21) **76**
Table 4.6.3 Cluster 3 Sub Cluster Year 2020-21 **77**
Table 4.6.3.1 Multinomial Logistic Regression Analysis **78**
Sub Cluster from Cluster 3 (2020-21) **78**
Table 5.4 Baseline Characteristics of COMPASS Cohort 2020-21 **89**
Table 5.5 F1 Score of Six Machine Learning Classifiers on the Three Preprocessing Strategies **94**
Table 5.6 Overall Model Performance in Predicting BMI Status **94**
Table 5.7 Performance Metrics of each BMI Status in CCA **95**
Table 5.8 Performance Metrics of each BMI status in MI **95**

Table 6.4 Descriptive Statistics of Year 1 Cohort (2020-21) by Weight Perception **127**

Table 6.5 Descriptive Statistics of Year 2 Cohort (2021-22) by Weight Perception **129**

Table 6.6 Descriptive Statistics of Year 3 Cohort (2022-23) by Weight Perception **132**

Table 6.7 Markov Chain Transition Probabilities of Weight Perception from Year 1 to Year 2 **135**

Table 6.8 Markov Chain Transition Probabilities of Weight Perception from Year 2 to Year 3 **136**

Table 6.9 Multinomial Logistic Regression of Year 1 Predictors Associated with Year 2 Weight Perception **137**

Table 7 Multinomial Logistic Regression of Year 1 Predictors Associated with Transition in Weight Perception from Year 1 to Year 2 ..... **140**

Table 8 Model Performance in Predicting Year 3 Weight Perception in CCA and MICE Imputed Datasets **142**

Table 9 Model Performance in Predicting Year 3 Underweight, About the Right Weight, and Overweight in CCA Datasets **142**

Table 9.1 Model Performance in Predicting Year 3 Underweight, About the Right Weight, and Overweight in MICE Imputed Datasets **142**

Table S1.1 Cohort 2018-2019 Input Variable for Clustering Analysis **205**

Table S1.2 Cohort 2020-2021 Input Variable for Clustering Analysis **230**

Table S2.1 All Variables **280**

Table S2.2 Hyperparameters of Machine Learning Models for BMI Prediction **295**

Table S3.1 F1 and ROC AUC Score for Determining the Number of Features in XGBoost Classifier **300**

Table S3.2 Final Variables for Cohort Analysis **304**

Table S3.3 Percentage of Missing for the Selected Variables in Each Year **307**

Table S3.4 RNN and LSTM Parameters for Bayesian Optimization **312**

Table S3.5 Optimized Parameters for the Selected Model **319**

List of Abbreviations

AIC Akaike information criterion

ANN Artificial neural network

AUCROC/ROC-AUC Area under the curve of the ROC curve

BMI Body Mass Index

CHI Calinski-Harabasz Index

CNN Convolutional Neural Network

CI Confidence Interval

CCA Complete Case Analysis

DB Davis-Bouldin Index

DL Deep Learning

DT Decision Tree

EHR Electronic Health Record

KNN K-Nearest Neighbors

LSTM Long Short-Term Memory

LR Logic Regression

MAE Mean Absolute Error

MAR Missing at Random

MCA Multiple Correspondence Analysis

MCAR Missing Completely at Random

MCM Markov Chain Model

MEC Mobile Examination Center

MI Multiple Imputation

MICE Imputation by Chained Equations

ML Machine Learning

MNAR Missing Not at Random

MSE Mean Square Error

NB Naive Bayes

NHANES National Health and Nutrition Examination Survey

PCA Principal Component Analysis

RF Random Forest

RNN Recurrent Neural Network

SM Statistical Modelling

SVM Support Vector Machine

XGBoost Extreme Gradient Boosting

Chapter 1
Background

The Background section began with a literature review of why improving adolescent health is important and the effect of COVID-19 on their behaviour formation. Among various behaviours, the subsequent sections focused on childhood obesity, highlighting its impact on public health, followed by the risk factors contributing to this concern and current interventions targeting childhood obesity. After reviewing adolescent health and factors related to childhood obesity, the chapter discussed the data source used for this thesis, introduced a comparable study from the US regarding collected data measures, and concluded with sections outlining the thesis scope and structure.

1.1 Adolescent health before and after COVID-19

Adolescence is critical for behaviour formation and involves physical, social, and psychological changes [1]. Health-related behaviour changes, such as substance use, unhealthy eating and lack of physical exercise, can occur during this period. The risk of unhealthy behaviours during this stage can persist into early adulthood and lead to adverse outcomes [2, 3]. For instance, nutrition and substance use, including drugs, smoking and alcohol use, are key determinants that affect the health of young adulthood. Low- and middle-income countries have been impacted by both malnutrition as well as nutritional disorders due to the prevalence of being overweight and obesity [4, 5]. Substance use is also a significant health concern among adolescents. Young adults who start smoking tend to continue this habit throughout their lifetime. Alcohol consumption is also reported as a major issue among 10 to 24-year-olds, accounting for 7% of disability-adjusted life years [1]. Young adults are more prone to drug abuse, leading to several negative outcomes such as poor mental health and unemployment [6, 7]. More importantly, lifestyles and health-related behaviours adopted between the ages of 12 and 18 continue to remain through older ages [8].

During this sensitive stage of behaviour development, COVID-19 was declared a global pandemic by the World Health Organization (WHO) on March 11, 2020, affecting adolescents' daily routines. Limitations were placed on large social gatherings and sports events, and social quarantine was implemented to prevent the spread of the virus [9, 10]. More than 1.576 billion students from 18 countries were affected by the school's closure and switched to online learning [8]. The changes in the school environment have restricted youth participation in physical activity, increased sedentary time and affected their quality of life [8]. The disruption of social interaction with their community networks and peers has exacerbated difficulties in emotional

regulation [10]. Several studies have discovered a significant increase in depression symptoms, anxiety or stress for future uncertainty [9, 10, 11, 12].

Besides mental health well-being, there was a significant decline in physical activity and increased sedentary behaviour among youth [8, 13]. A Korean national adolescent health survey examined the changes due to COVID-19 on adolescent health and discovered an increase in the perception of chubby body type and the intention to lose weight, as well as a decrease in average daily sleep duration [8]. This finding was also reflected in a systematic review of youth eating behaviours impacted by COVID-19. Studies have reported the perception of body shape as obese and an increase in BMI during the pandemic compared to pre-pandemic [14].

Additionally, a scoping review identified a decline in physical activity time in cross-sectional and longitudinal studies. All studies investigating the change in the proportion of youth meeting the physical guidelines indicated a decrease [13]. Another behaviour impacted by COVID-19 is eating habits. Studies reported a decline in well-balanced dietary intake and an increase in the consumption of comfort food such as desserts and sweets [15, 16]. Consumption of food in high sugar and fat acted as a coping mechanism for stress related to the pandemic [18].

Given the multi-dimensional aspects of behaviour emerging during adolescence, it is essential to investigate various factors and compare potential changes before and after COVID-19. The first study focuses on multiple perspectives and profiles of students (Grades 9 – 12) based on their characteristics. Investigating these changes provides an understanding of monitoring and informing strategies for school interventions to promote adolescent health and well-being.

1.2 Childhood Obesity

Several behavioural aspects are critical determinants of adolescents' health. This thesis focuses on childhood obesity and weight perception, as these two phenomena are associated with various factors, including individual behavioural, environmental, or social. Identifying modifiable factors that can be addressed in early childhood before leading to obesity in adulthood is valuable.

Childhood obesity is a worldwide health concern. According to Statistics Canada, in 2018, the prevalence of overweight and obesity in children between the ages of 5 and 11 was projected to be 27.1% in Canada [19]. Childhood obesity is an important factor for non-communicable diseases, including cardiovascular disease, hypertension, type 2 diabetes, and non-alcoholic fatty liver disease, affecting both developed and developing countries [20, 21]. Being obese during childhood often leads to obesity in adults [21]. Besides the health burden,

obesity also has an economic impact. According to the statistics among Grade 5 children in Nova Scotia, the total health costs were 21% higher in obese children compared to healthy weight children over three years [22]. In another review article, the economic burden of childhood obesity was projected to cost $13.62 billion directly and $49.02 billion indirectly per year by 2050 worldwide [212].

The definition of overweight and obese is defined as excess body fat [21]. Body mass index (BMI) is an indirect measure of adiposity. It is calculated using weight (kilograms) divided by height (metres squared). The commonly accepted cut-off points for BMI are <18.5 for underweight, 18.5 - 24.9 for normal weight, 25 - 29.9 for overweight, and ≥ 30 as obese [23]. The Public Health Agency of Canada uses the World Health Organization (WHO) growth reference charts for children. Specifically, overweight children are those with a BMI between one or two standard deviations above the mean. Obese children are categorized as having a BMI two standard deviations above the mean [24]. Other direct measurement methods include ultrasound, air displacement plethysmography, and dual-energy X-ray absorptiometry [25]. However, these are limited to small cohort studies that require high precision due to their high costs [25].

1.3 Childhood Obesity Risk Factors

Several aspects are related to childhood obesity, including biological, behavioural, environmental, and psychological. Below are the risk factors from each aspect that influence this public health concern.

Biological: Genetic factors increase the risk for obesity in childhood. This factor influences metabolism, and the heritage of obese parents also impacts obesity in children [26]. Studies reported that genetic variants from different genes increase hunger and food intake, resulting in excessive weight gain [27]. During infancy, formula feeding instead of breastfeeding has been associated with increased weight, persisting into later stages. BMI has been found to be 25-40% heritable [20].

Behavioural: Behavioural causes of obesity usually refer to dietary intake and physical activity. Parents or caregivers strongly influence the types of food children consume in early life. For example, they may encourage children to complete a processed and less nutritious meal even though they may not want to [27]. In a later stage, peer interactions and food options at school are critical factors influencing food choices [28]. Dietary factors include the consumption of sugary beverages, snacks and fast food. Fast food tends to be high in calories and low in nutritional value [20]. Increased consumption of fast food has been linked to obesity. Participants who consumed more calories from fast food were associated with weight gain [20]. Sugary

drinks, including soda, juice, and other sweetened beverages, as well as snacks such as baked goods, candy, or chips, are also high in calories and significantly contribute to becoming overweight [29]. In addition to dietary factors, the level of physical activity is another important factor in childhood obesity. Overweight and obese children are less active and engage in sedentary lifestyles such as television watching and playing video games [20, 30]. Among the 26 studies reviewed in a scoping review on the impact of video gaming on childhood obesity, 46% reported a positive association between gaming and obesity, while 53% found no association [261]. Notably, certain physically active video games seemed to contribute to reducing weight gain [261]. This shows that playing video games has inconsistent results in childhood obesity. Similarly, social media appears to have a bi-directional effect on childhood obesity. Some evidence suggests that increased screen time viewing energy-dense drinks and fast-food content or food advertisements contributes to obesity in children [262]. On the other hand, interactive social media might promote physical activity to prevent obesity [262]. However, more studies, such as randomized control trials, need to be implemented to explore this effect. Studies conducted globally have discovered that television watching has an inverse relationship with participation in physical activity, and this leisure activity is demonstrated as a main cause of obesity among children [30]. On the contrary, increased physical activity has a protective effect, reducing the extra energy needed for weight gain. For instance, resistance training has been shown to promote gaining lean body mass and participating in moderate activities for 150 – 250 minutes per week helps prevent weight gain [25].

Environmental: The environmental aspects include access to fast food in the community, walkability, and the quality of lunches and snacks at school. A food desert refers to areas with limited access to nutritious and affordable food. Evidence has shown that the density of fast-food restaurants in the neighbourhood is associated with obesity prevalence [31]. The safety of the living environment also influences the activity level. Among children who did not walk or bike to schools, most parents indicated that the neighbourhood was not safe for walking due to heavy traffic or fears of child abduction [29]. Green spaces and recreational facilities act as a protective effect on children's weight [50]. More access to green space encourages more physical activity and less screen time [51]. Since adolescents spend most of their time at school, they are affected by the quality of food offered at school [32, 33]. Among the 20 studies in a systematic review, the availability of healthy food options, including fruit and vegetables offered during breaks, school gardens, and school breakfast programs, has significantly decreased the likelihood of obesity. In contrast, unhealthy food options, such as the vending machines and soft drinks in cafeterias, were associated with high BMI [32]. Cultural differences have also been shown to contribute to childhood obesity among racial and ethnic groups [260]. Examples of racial and ethnic minorities include non-Hispanic White, non-Hispanic Black and Mexican-American children [260]. Some ethnic minority mothers have a different interpretation of defining a healthy child since thinness is considered poor health in certain cultures. The

neighbourhood in which the racial and ethnic minority children reside could be less favourable for them to access food markets and recreational facilities [260].

Psychological: To align with the scope of the third study of the thesis, the psychological aspect focuses on body dissatisfaction and self-esteem. Body dissatisfaction or misperception refers to the discrepancy between an individual's actual and perceived ideal body size. There are also gender differences in weight perception [34]. Females tend to overestimate their weight, while males underestimate it [34]. Low self-esteem related to body dissatisfaction often leads to depression and anxiety [34]. Overweight and obese children have a higher proportion of body dissatisfaction compared to normal-weight children, particularly females, who exhibit greater body dissatisfaction than males [35]. Notably, the relationship between body dissatisfaction and BMI seems to be bi-directional. The negative perception of one's body appearance and weight-related eating disorders have been identified as risk factors for obesity, and a significant association between body dissatisfaction and obesity has been reported [36, 37]. Research has also shown that being obese at an early age increases the risk of poorer body image as well as low self-esteem [38].

1.4 Childhood Obesity Interventions

The current intervention for childhood obesity focuses on community, school, behavioural, and family-based interventions. Community support is critical to introducing events to promote healthy eating, educate, and improve people's awareness of healthy lifestyles [26]. Examples include establishing groceries in areas with limited access and infrastructure changes in the neighbourhood to add more walkable areas and more facilities for physical activity [26]. More accessible and safe community environments encourage children to engage in physical activity [26]. To tackle the fear of safety for parents, urban planning is necessary to have more walkable areas and shorter distances to schools. Collaboration efforts from the local and national governments are also needed to create liveable communities to promote healthier lifestyles among children [45].

Nowadays, children spend more time playing video games and using smartphones than watching TV. Reducing screen time on computers and video games appears to reduce weight gain [262]. Due to this rapid shift from television watching to social media content, future research could investigate the impact of social media exposure on weight gain in children [262] Robinson et al. [262] also mentioned that active video games could be an intervention for weight control programs, but more observational studies are required to examine the associations.

School is another focus for building the intervention program since youth spend most of their time at school [26, 41]. The interventions include nutritional and physical aspects for

weight control. From a nutritional perspective, schools can provide balanced meals and healthy food by introducing nutritious food choices. One of the programs in Canada is called the school nutrition policy. This program is designed to improve the quality of food and beverage intake to reduce the risks of being overweight and obese. It consists of multiple components, such as providing health and nutrition education, specific selections of foods or beverages, dietary counselling and health services [41]. Various studies indicate the effectiveness of the school nutrition policy as it is positively associated with improving dietary choices both inside and outside of school [41]. Another type of intervention at school is physical activity-focused, which motivates students to walk more and participate in non-competitive activities such as dancing [26]. An example of this intervention strategy includes daily structured moderate to vigorous physical activity, sending letters to parents to raise awareness of physical activity, educational programs to promote physical activity, and school events aimed at encouraging active lifestyles [40].

Family plays an important role in developing children's behaviours and lifestyles [26]. Family-based intervention has been recommended by experts, suggesting that the family system provides greater changes among children since the shape of childhood behaviour happens daily [42, 43]. The family intervention generally consists of behavioural controls or monitoring of exercise and diet, education on nutrition and physical activity, and psychoeducational parenting groups [45]. The ' Stoplight Diet ' is a well-known successful family-based intervention that continuously results in weight loss longitudinally. This program has targeted nutritional and physical activity education, parenting skills training, and the at-home monitoring of physical and nutritional behaviours [45]. Combining parenting skills and education regarding nutrition and physical activity has proven more effective results than the intervention focusing on one component [46].

1.5 The COMPASS System

In the earlier section of this chapter, BMI is denoted as an indirect measurement of childhood obesity. Even though objective measurements tend to be accurate, this method may not be feasible in large studies since such measurements would require access to the appropriate equipment and trained professionals [47]. Often, the measurement of BMI is based on the self-reported weight and height of the youth population [47].

The COMPASS system is an ongoing cohort study, focusing on collecting hierarchical longitudinal data among grade 9 to 12 students attending secondary school. Initially, 90 schools were recruited (79 in Ontario and 11 in Alberta in Wave 2, 2013-14) to follow annually regarding the characteristics of the built environment within the community surrounding schools and the school's policy and programs [48]. Using Wave 7 (2018-19) COMPASS sample as an

example to demonstrate the school's demographics, 138 schools participated in the COMPASS study [263]. Among these participating schools, there were 61 schools from Ontario, 52 schools from Quebec, 15 schools from British Columbia, 8 schools from Alberta, and 2 schools from Nunavut [263]. From this sample, there were 32 schools defined as Wave 1 schools that completed their first data collection in 2012-13, 24 (Wave 2) schools that completed the first data collection in 2013-14, one (Wave 4) school that completed the first data collection in 2015-16, 23 (Wave 5) schools that completed the first data collection in 2016-17, 39 (Wave 6) schools that completed the first data collection in 2017-18, and 19 (Wave 7) schools that completed the first data collection in 2018-19 [263].

The inclusion criteria for recruiting a school required it to be a secondary school with at least 100 students from grades 9 to 12 and granted permission to use active information, passive consent, and parental permission protocols [48]. The mentioned consent protocols were also applied to the eligible students during student-level recruitment. Parents or guardians were provided with the information letter and could contact the COMPASS recruitment coordinator if they wished to withdraw their child [48]. Students were also allowed to withdraw or decline participation from the study during the consent process or data collection [48]. Details about the data can be found on the website (www.compass.uwaterloo.ca) [49].

Before the COMPASS system, there was inadequate information available to understand school characteristics such as programs, policy, built environment, and their impacts on the various behaviours and outcomes of the youth population [48]. It is crucial to fill this knowledge gap by developing school-based interventions to improve adolescent health. Therefore, the primary purpose of establishing the COMPASS system is to improve youth preventative research by providing real-world evidence and engaging multiple stakeholders to generate adolescent health initiatives [48, 49]. The COMPASS system collected data to strengthen the capacity of research focusing on advancing adolescent health from multiple perspectives, such as substance use intervention, promotion of healthy diet and physical activity, and obesity prevention [49].

For collecting student-level data, the 12-page COMPASS Student Questionnaire (C_q) was distributed during class time among the participating schools on an annual basis [48]. The C_q was designed anonymously by generating a unique self-generated code for each student within a school that protects the privacy of the participants and allows for the linking of the responses of students for longitudinal research [48]. The survey is machine-readable and designed to be completed between 30 to 40 minutes [48]. The survey questions include multiple health indicators and behavioural aspects, including demographic information such as age, grade, biological sex, gender, race, ethnicity, income, and home life such as how many people live in the household, healthy eating measures, health and weight for measuring BMI, sedentary lifestyles such as screen time, physical activity, and substance use, including alcohol drinking,

smoking, tobacco and marijuana use, bullying, academic achievement, and sleep patterns [48, 49].

In addition to the student-level questionnaire, the COMPASS School Programs and Policies Questionnaire (SPP) was sent to the school administrator to assess the school programs and policy environment [48]. This survey evaluates the change and the existing school policy, current resources, and programs related to the students' health, focusing on the behavioural perspective from C_q [48]. Moreover, the COMPASS School Environment Application (Co-SEA) and COMPASS built environment (C-BED) were developed to monitor the built environment within the school and the surrounding school community-based environment that influences the students' behaviours [48].

1.5.1 Measures in COMPASS

This section introduces the various domains addressed in the 2020-21 COMPASS survey questionnaire. The domains include demographics, weight-related measures, sedentary lifestyles, physical activity, substance use, school-related measures, mental health, healthy eating, and COVID-19 variables. The measures related to these domains are shown below. Some variables were pre-processed for analysis; the details are in each study's data preparation and preprocessing sections. The COMPASS data dictionary for all measures and a COMPASS survey (2020-21) are in the Appendix.

- Demographic: Age, grade, biological sex, gender, race and ethnicity
- Weight-related measures: weight perception, weight-change intention
- Sedentary lifestyles: screen time spent playing video games, watching TV, doing homework, surfing the internet, browsing and scrolling through social media (data collection for these variables started only in the 2021-22 survey)
- Physical activity: participation in school-organized sports, hard or moderate physical activity
- Substance use: alcohol, smoking and tobacco use, vaping and e-cigarette use, cannabis and opioid use
- School-related measures: academic performance, bullying, school connectedness
- Mental health: Generalized Anxiety Disorder 7-item Scale [54], flourishing scale [55]
- Healthy eating: reasons for skipping breakfast, consumption of snacks, protein, vegetables, sugary drinks, food purchased at a fast-food restaurant, water consumption
- COVID-19 related measures: knowledge of COVID-19 and life changed caused by COVID-19 in boredom, loneliness, anxiety, time spent with my friend in

person, time spent in video games, social media, cigarette use, vaping, alcohol and cannabis use

From the above representative measures, certain variables are collected differently depending on the survey questions. In the demographic domain, sex refers to the one assigned at birth. Self-identified gender refers to an individual's perception of their identity, which is not confined to the binary terms of female and male. [254]. Specific COMPASS questions related to these two measures are addressed below.

Before the 2021-22 COMPASS survey, a sex question was asked "Are you female or male?". The binary responses were either "Female" or "Male". The same question was asked in the 2020-21 COMPASS survey, but the response included "I describe my gender in a different way", and "I prefer not to say". Due to the self-identified categories, this variable was later referred to in the second study as gender. Starting in the 2021-22 COMPASS survey, a more detailed question regarding gender was asked "Which gender do you most identify with?", the response categories included "Girl/Women", "Non-binary person", "Two-Spirit", "Boy/Man", "I describe my gender differently", and "I prefer not to say". The question about sex was phrased more directly starting in the same year. Instead of asking, "Are you female or male?", the sex question focused on "What sex were you assigned at birth"? The categories included "Female", "Male", and "I prefer not to say".

Similarly, starting in the 2021-22 COMPASS survey, questions were asked for more details about race and ethnicity. Race was surveyed as "What race best describes you"? and the categories included "Black", "East Asian", "Latino", "Middle" Eastern", "South Asian", "Southeast Asian", "White", "Another category", "I don't know', and "I prefer not to say". To account for multiple responses to this question, the derived ethnicity included an additional category called "Multiethnic" alongside ones mentioned in race. "Another category" refers to the Indigenous ethnicity.

Another measure that is surveyed differently is healthy eating. In the 2018-19 COMPASS survey, the questionnaire collected data on the consumption of lunch, snacks, sugar-sweetened beverages at school and food on the weekend and yesterday (detailed questions are in the Appendix). These questions were not surveyed after this year, and the data collected then focused on fast food, restricted food intake, fruit and veggies. For instance, the fast-food question asked, "During that week, on which days did you do the following?" The response options included Monday to Sunday.

Besides the changes in the survey questions, the questionnaire distribution in Quebec was slightly different. The Quebec survey was divided into Modules A, B, and C in 2023. Module A

contains all questions for all participants. In Module B, since this was based on random sampling, only 50% of the sub-sample of participants were questioned. For example, 50% of the students in Quebec in the 2022-23 COMPASS survey received questions about weight perception. Module C contains more diverse optional questions, and schools can choose one more module. In 2024, Module A remains the same as last year, but Module B is removed. One of the five Module C options is distributed randomly to the schools that are able to add a second option in Module C.

1.6 NHANES database

This section introduces another study that assessed obesity prevalence compared to the COMPASS study. The COMPASS study includes the various dimensions of youth behaviours on self-reported student-level data, while the National Health and Nutrition Examination Survey (NHANES) directly measured the heights and weights of the participants, in addition to medical visits and lab data collected alongside survey data [68]. Medical and physically collected data are not currently available in the COMPASS study.

NHANES is a national ongoing survey conducted since 1999 among the non-institutionalized civilian household population in the United States. This health-related program is conducted by the Centers for Disease Control and Prevention (CDC) National Center for Health Statistics (NCHS) [56]. On an annual basis, the survey is conducted among 5000 persons living across 15 counties in the country, and the data are released every two years. The survey contains a home-administered questionnaire and a follow-up health examination conducted at the mobile examination centers (MECs) [57]. The inclusion of the sample population has been updated throughout the NHANES cycles. The oversampling of the Mexican American population and a supplemental sample of pregnant women were added in 1999-2006 (the sample of pregnant women was discontinued in the 2007-2010 sample), the Hispanic population was added in the 2007-2020 sample, followed by the oversampling of the Asian population in the 2011-2014 design [57]. The diverse sample includes all ages from birth, including those who are lower-income and aged 80 and over [57]. The primary goal of NHANES is to assess the nutritional and health situations of children and adults in the United States (US) [58]. The other goals of this program include providing estimates on the prevalence of health conditions and characteristics in the US population, studying the relationship between nutrition, diet and health, and assessing the environmental, social, and behavioural factors influencing disease occurrences [66, 67].

1.6.1 In-Home Interviews

All members aged 0 to 16 and older than 60 are eligible to participate in the selected household. If the household has more than one eligible participant, ages 20 to 59, one will be randomly selected. Two participants will be randomly selected if there are four or more eligible participants [60]. Participation is also voluntary; both the in-house interview and the MEC staff obtained consent from the participants through written and verbal consent (if conducted by telephone during 2021-2023). For children, participation consent is provided by their parents or guardians. The in-person interview at home begins after informed consent has been obtained through the computer-assisted personal interview system [60].

The NHANES interview includes demographics such as age, biological sex, gender, income, race, ethnicity, housing and family characteristics, socioeconomic status, dietary, and health-related questions [56, 59]. The questionnaire is collected at both the individual and family/household levels. For example, in the NHANES 2019-2020 questionnaire, some of the sections from the individual-level questionnaire include Diet Behaviour and Nutrition, Weight History, and Diabetes. For the Family Questionnaire, Smoking and Food Security sections are included. The full list and detailed questionnaire instruments are in [62]. Selected measures for the domains mentioned above are:

- Diet Behaviour and Nutrition: breastfeeding, frequency of milk drinking, school lunch and breakfast, and intake of prepared food
- Diabetes and Weight history: insulin (if any), weight, height, and intention to lose weight
- Family-related measures: smoking (frequency or use) within the household, food insecurity in the household (if any)

1.6.2 Mobile Examination Center (MEC) Visit

Besides the in-home interview as self-reported data, NHANES also collects physical measurements. After the in-home interview, the sample participants are scheduled to visit the MEC. The MEC collects dental, medical, and physiologic measurements and laboratory test data. Examples of medical data include anthropometry, blood pressure, body measurement, and audiometry [62]. For collecting the lab data, the tests are distributed to collect the participant's biological specimens [63]. All the procedures are administered by trained medical personnel and staff. The full list of the procedures can be found in [62, 63].

Within the examined data, the body measurement monitors the growth of infants and children of all ages. Head circumference is measured from birth through six months old, and recumbent length is measured from birth through 47 months. Standing height and waist

circumference are measured from 2 years and older, upper leg length is measured from 8 years and older, upper arm length and middle upper arm circumference are measured from 2 months and older, and hip circumference is measured from 12 years and older [64]. For calculating the BMI-related categories, the participant's weight, height, age, and gender are considered. The BMI is calculated as weight (kg) divided by height (meters squared) and rounded to one decimal place. The BMI category, children and adolescents between 2 and 19 years old, is classified based on the CDC growth chart [65]. The four categories are defined as follows [64]:

- Underweight: BMI < 5th percentile
- Normal weight: BMI 5th - < 85th percentile
- Overweight: BMI between 85th – 95th percentiles
- Obese: BMI ≥ 95 percentile

1.7 Thesis Scope and Implication

This thesis is intended to be viewed as a guide for using the machine learning (ML) approach to drive decision-making by stakeholders aiming to improve adolescent health. The intended stakeholders of this thesis include public health professionals from the Health Agency of Canada, health policymakers such as provincial and territorial government officials, school principals, and researchers interested in using ML for public health research. The thesis aims to disseminate messages to the stakeholders to enhance their understanding of critical behaviour-related factors that drive adolescent health outcomes, further assisting them in developing targeted strategies, interventions, and school-based programs to improve the wellness of adolescents and prevent public health problems such as childhood obesity.

The COMPASS study focuses on multiple perspectives on high school student behaviours and school characteristics, filling the gap in evidence-based youth data. However, this study also has some limitations in the data collection among the surveyed provinces. For example, the self-reported data were collected using passive consent, which assumes that the children are eligible to participate if the parents do not contact and withdraw their students [48]. This leads to the exclusion of the school boards from Waterloo, Toronto, and Peel, resulting in a loss of sample size. In contrast, Quebec uses census-based consent, and the school boards can be directly contacted without approval. Therefore, the data collected in the COMPASS study may not represent all provinces in Canada. Despite the limitations, COMPASS still serves as an inclusion cohort study that guides the understanding of school-related characteristics associated with adolescent health behaviours [48].

Similar to the initial objective of starting the COMPASS study, this thesis enhances current youth studies by emphasizing youth populations that spend most of their time at school

and further exploring the potentially modifiable behaviours using advanced analytics. Previous adolescent health studies have provided limited evidence on ML predictive models in a school-based cohort study focusing on adolescents cross-sectionally and longitudinally. The ML approach used in this thesis is capable of handling multi-dimensional factors to explore influences on youth behaviour simultaneously. This provides additional data and granularity to existing adolescent research, which often focuses on specific variables, such as those related solely to physical activity or healthy eating. From a practical standpoint, the result of this thesis allows the aforementioned audience to visualize the important factors simultaneously associated with outcomes that affect youth behaviours through advanced modelling. It also enhances current practices by incorporating multiple aspects of behaviour domains collected from the school environment. Having interpretable results generated from the ML approach is also meaningful for the stakeholders to understand the associated factors cross-sectionally and in terms of historical trajectories in the outcome change, such as weight perception longitudinally. Understanding these perspectives informs the stakeholders and helps them evaluate the short-term and long-term effectiveness of their current programs or initiatives. The model from this thesis is also retrainable and flexible, allowing for adding more parameters based on specific requirements from the schools or public health professionals if new measures are introduced from the COMPASS study or if they are interested in a different focus besides weight-related behaviours.

1.8 Thesis Organization

The three studies are organized as separate manuscripts with the main theme of youth behaviour health. Chapter 1 introduces the background of adolescent health, followed by an overview of childhood obesity and the COMPASS system with related student questionnaire domains. Chapter 2 discusses the machine learning methods and the statistical approach. Chapter 3 presents the rationales and the research questions for each study. In Chapter 4, Study 1 begins with a general approach to clustering student behaviours from multiple COMPASS questionnaire domains to observe youth profiles before and after COVID-19. Chapter 5 contains Study 2, which focuses on self-reported BMI by building an XGBoost ML model and investigating the factors associated with BMI prediction. Since weight perception is the most important predictor of BMI prediction, Study 3 in Chapter 6 further examines its transitional probabilities and utilizes deep learning models to capture the temporal factors predicting weight perception. Chapter 7 highlights the results of the three studies by answering each research question mentioned in Chapter 3, discussing strengths and limitations, exploring contributions to public health and ML, and outlining future work.

Chapter 2

Analytic Approach

The general approach for the statistical analysis of childhood obesity assesses the BMI as the outcome and indicator of youth obesity. Mondal et al. [70] suggest that current analyses on youth obesity prediction can be categorized as regression analysis, risk factor-based prediction, and machine learning prediction. The risk-based model examines relative risk factors for obesity prediction, including biological, family, community, and environmental factors [70].

This section highlights the current analytical approach to assess the prevalence and trend of childhood obesity. It also summarizes the literature review's traditional statistics and machine learning approach. The description of the machine learning approach and other models focuses on the methods used in the three studies.

2.1 Traditional Statistical Analysis in Childhood Obesity

In a systematic review on predicting childhood overweight status using maternal and early life risk factors, logistic regression was applied in seven out of nine studies [69]. In another review of the statistical models to predict childhood and adolescent obesity, logistic regression was the most applied among the 14 mentioned studies. In studies on youth obesity using logistic regression or multivariable logistic regression approaches, the primary outcomes were either binary (obese and non-obese children) or classified the status as normal weight, overweight or obese based on the CDC standardized growth curves or the growth reference data from the World Health Organization [71, 72]. As part of the risk-based model, the logistic regression analysis has been implemented in several studies to determine the risk and the magnitude of factors that contributed to the development of childhood obesity. Heerman et al. [72] applied multiple logistic regression to estimate the odds of obesity after 36 months among predictors of parental characteristics, child baseline BMI, energy intake, physical activity and food security. In a longitudinal study to predict the future risk of youth obesity in Japan, Sonoda et al. [259] categorized obesity as the participants having a larger percentage of overweight status, equal to 20%. The binary outcome categorized obese or non-obese. The definition of obesity (binary) was calculated based on the percentage of overweight, greater and equal to 20% in overweight defined obesity or not. The predictors, including family histories of hypertension and diabetes, lack of sleep and exercise, time spent in sedentary activities, and medical measures, were included in the multivariable logistic regression.

Another approach was regression analysis, where the primary outcome is treated as continuous, such as percentile and z-score. In a systematic review of maternal BMI and youth obesity, 23 studies reported using child BMI or z-score as the continuous outcome [73]. Ardic et al. [74] investigated the change in BMI z-score in 3-year-old children during the COVID-19 pandemic and the associated factors using linear regression analysis. Pei et al. [75] transformed the anthropometric weight and height measurement into BMI z-score based on the sex-specific and age-specific WHO standards. Linear regression was used to examine the effects of birth and standardized BMI score and the association between the covariates, such as maternal smoking during pregnancy and this outcome. Additionally, continuous BMI was seen as an outcome in evaluating the intervention in randomized-control settings. In a systematic review of the randomized-control trials on the effectiveness of weight management intervention, the weight changed from baseline to 12 months [76]. Llargues et al. [77] analyzed the differences in BMI scores in a randomized prospective study to assess the effectiveness of the school-based intervention on physical activity and eating habits. In a clinical intervention in a quasi-experimental field trial, the BMI z-scores were assessed before and after the intervention, both retrospectively and longitudinally [78].

Stratification was also commonly implemented, followed by logistic regression. From the above-mentioned youth obesity research in Japan, the authors performed stratification on biological sex (females and males) on obese and non-obese children [259]. They first identified the key factors associated with the development of obesity, then constructed another stratified model with and without the overweight as a factor and included the significant factors from the sex-adjust logistic regression model as independent variables. In addition, obesity prevalence was also performed by stratifying between racial and ethnic groups. For example, Skinner et al. [256] used the ordinary least squares regression to predict the obesity and severe obesity prevalence based on racial and ethnic groups, differentiating among White, African American, Hispanic, Asian American, and Other (with and without Asian American). The sub-group analysis revealed the obesity prevalence in different represented groups and accounts for the potential results regarding associated risk factors related to childhood obesity.

2.1.1 Drawbacks of Statistical Modelling

Certain disadvantages exist when using the traditional statistical model (SM), such as logistic and linear regression.

In the prediction model, the application's success depends on choosing appropriate input variables. Large standard errors and inaccurate confidence intervals affect the true association [79]. Collinearity or multicollinearity is a potential concern, referring to the highly correlated

predictors. For instance, weight and BMI are highly related [80, 81]. One way of handling this issue is to drop one of the variables to avoid the impact of the imprecision in the regression model. Another concern for using logistic regression is that the basic assumptions need to be met, including 1) Independence of observations, 2) Linearity for any continuous predictor, meaning there needs to be a linear relationship between the variables and the logit-transformed outcomes, 3) Absence of multicollinearity, as it is essential to avoid predictors that are highly correlated, which causes large standard errors for the slopes, and 4) Absence of influential outliers, meaning the predicted and actual values need to be similar to each other. Otherwise, the accuracy of the model could be compromised [84].

Similar to logistic regression, the result of the linear regression may be misleading if the assumptions are not met [85]. Assumptions include 1) the linear relationship must exist between the independent and the dependent variables, 2) Homoscedasticity, meaning the variance of the residual should not vary for any value of predictor, 3) All observations in the samples need to be independent of each other, and 4) Normality of residuals, meaning the residuals should be normally distributed [86]. Multicollinearity also needs to be addressed before implementing the regression models. Another concern is overfitting in regression analysis. Overfitting refers to the increased complexity of the model by including excessive predictors, interactions or non-linear terms [87]. Hawkins [88] illustrated that using irrelevant variables violates parsimony if fewer variables are enough to explain the outcome. Adding unnecessary predictors could lead to inaccurate predictions based on undetected errors. Therefore, linear regression is considered inflexible and prone to complex relationships within the dataset.

For the traditional SM, multivariable analysis is one of the most common approaches to identifying the association between risk factors and outcomes [250]. One of the major limitations of this analytic model is that it is prone to the effects of correlation if it exists among the inputted variables, and these covariates are not independent of one another [250]. The impact of the correlation effect results in a loss of statistical power and increased standard error of regression coefficient estimation. This leads to a biased estimation of the regression coefficients with inaccurate magnitude and direction since these variables are related to each other [250]. As stated above, the traditional SM is more restricted by the assumption that assumed variables are not correlated. Variable selection methods, including forward selection and backward elimination, help reduce the effect of multicollinearity in SM [251]. The former selects the variables that maximize the decrease of the residual sum of squares. The latter chooses the variables that least increase the residual sum of squares [251].

2.2 Why Machine Learning (ML)?

There are some advantages of ML over SM. First, ML operates well even when encountering massive datasets, including electronic health records (EHR), millions of patient charts, imaging, and social media data [75, 89, 91]. Some ML algorithms perform better than traditional regression methods in scenarios with many independent variables and a limited number of observations [89]. Second, pre-selecting predictors is critical in SM to avoid collinear variables. An example from the medical field is the relationship between tumour stage and the surgical completion rate of tumour removal [91]. ML algorithms are flexible enough to learn complex relationships in situations where there is a lack of understanding of the phenomena [92]. An example is that little knowledge exists to explain the association between neurobiological disease and the progression of the brain. Using ML can help reveal the multiple aspects of possible influences on disease mechanisms [92]. Third, the ML approach can be applied to different healthcare settings to predict patient diagnosis, guide decision-making in treatment, and predict side effects to improve patients' overall quality of life [89].

Regarding the impact of the multicollinearity effect on ML, the model could have less generalizability, resulting in overfitting [251]. The difference between ML and SM is that certain ML methods, such as a neural network, are superior in handling this issue compared to traditional SM [251]. In addition to using variable selection as mitigation, a method called FA-ANN (factor analysis-artificial neural network) has been used to tackle multicollinearity [252]. This model performed with the highest accuracy compared to regression analysis [251, 2.52]. This shows that statistical assumptions do not constrain the ML approach, and sophisticated models such as deep learning can handle correlated and high dimensional data [251, 252]. There are also additional mitigation methods, such as regularization and penalty mechanisms, that can be used within the deep learning model to complement the model selection.

Additionally, neural networks, including the Convolutional Neural Network (CNN), can detect the most representative variables in the convolutional and pooling layers [251]. The attention layer implemented in the Long Short-Term (LSTM) model can also capture variable importance based on their assigned weights to overcome the problem of multicollinearity [251]. Therefore, ML provides a more robust analytic approach in handling the potential multicollinearity effect than SM, which does not necessarily need to adopt a feature selection process and can produce a better estimation of the parameters [251].

2.3 Types of Machine Learning Algorithms

Several ML algorithms exist based on the types of tasks to be accomplished. These can be classified as supervised, unsupervised, semi-supervised, and reinforcement learning. The

following sections explain the logistics of supervised and unsupervised learning since this thesis did not use semi-supervised or reinforcement learning techniques. For a brief overview, semi-supervised learning combines labelled and unlabelled data from supervised and unsupervised methods. This technique is useful when there are many unlabelled and less labelled data. It is applied in the areas of fraud detection, text classification and machine translation [93]. Reinforcement learning uses machines or agents to automatically evaluate the optimal behaviours by either rewarding or minimizing the risk as one of the insights. Some applications include robotics and automation in driving [93].

2.4 Unsupervised Learning

Unsupervised learning refers to analyzing unlabelled data without human intervention. It has been widely used for discovering meaningful trends, structure, and grouping observations. The common tasks are clustering, feature learning, and dimensionality reduction [93]. Clustering involves grouping participants who share common attributes into clusters and has been applied to several areas, including biomedical image analysis, diabetes studies, and cardiac disease. Common clustering methods include k-means clustering, density-based clustering, and hierarchical clustering, among others [94, 95]. Among the 58 eligible studies in a systematic review investigating clustering algorithms for identifying patterns, 19 used partition-based clustering, such as k-means or k-medoids, and 13 used hierarchical clustering [95]. The application of clustering extends to multiple research areas, including image segmentation, customer segmentation, healthcare, climate science, and criminal profiling [95].

2.4.1 K-Means Clustering

As one of the most popular methods for clustering analysis, this algorithm aims to distribute the samples within the dataset into k clusters so that the objects, based on the closest mean within the clusters, have more similarities than the characteristics of the other clusters [96]. The k-means clustering process can be followed in the below steps [96].

Step 1: Group the dataset into k subsets based on the pre-defined k number of clusters

Step 2: Randomly choose k data points as the distinct centroids (the centers of the clusters)

Step 3: Assign data points to the nearest cluster centroid by calculating the Euclidean distance (d) between two points to quantify the data point’s similarities, represented by the formula [97]:

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

The distance is calculated between two points in the n-dimensional space. x_i represents the first coordinate of point X in the i-th dimension, where X = $(x_1, x_2, \ldots, x_n)$. y_i represents the second coordinate of point Y in the i-th dimension, where Y = $(y, y_2, \ldots, y_n)$.

Step 4: Update the centroid based on the mean of the data points in all the new clusters formed. Denoted as [98]:

$$c_k' = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i \quad (2)$$

c_k' is the updated centroid of cluster k (lowercase k refers to the individual cluster). n_k is the number of data points assigned to cluster k, where k =1, 2, 3…, with n as the total number of clusters). x_i represents the data point in the i-th data point in cluster k.

Step 5: Repeat 3-4 until all centroids stop changing or the maximum number of iterations has been achieved.

2.4.1.1 Determine the Number of K

Even though k-means is a straightforward way to identify grouping, pre-defining the number of clusters is challenging [96, 99]. The methods below are commonly used for determining k.

Elbow Method: This evaluation method determines the optimal number of clusters, k, by calculating the sum squared error (SSE) for each cluster until the maximum number of clusters has been reached. SSE refers to the sum of the average Euclidean distance of each datapoint against the centroid [101]. By comparing the difference between the SSE, the value that drops dramatically and forms an elbow angle is considered the optimal number of clusters [101]. The SSE is represented below [101]:

$$SSE = \sum_{i=1}^{k} \sum_{x_i \in C_i} || x_i - c_i' ||^2 \quad (3)$$

In equation (3), k represents the number of clusters, x_i is the number of data points in the dataset, c_i is the i-th cluster and c_i' is the new centroid of the i-th cluster.

***Silhouette method*:** This method utilizes the Silhouette index to evaluate the clustering result. The formula is defined as:

$$s(\chi_i) = \frac{b\,(\chi_i) - a(\chi_i)}{max\,\{b(\chi_i),\ a(\chi_i)\}} \quad (4)$$

$a(\chi_i)$ represents the average distance of a data point to all the other data points in the cluster, and χ_i is the element in the cluster, referred to as the intra-cluster distance. $b(\chi_i)$ is the minimum average distance from the objective to all points in other clusters, referred to as the inter-cluster distance [102, 103]. The Silhouette value calculated from the above formula ranges between the value of -1 to 1. A positive value is desired as the intra-cluster distance is smaller than the inter-cluster distance. The greater the value, the more likely the data points are correctly clustered. In contrast, the negative value of $b\,(\chi_i) - a(\chi_i)$ is not ideal because it indicates that the within dissimilarity is larger than the between-cluster dissimilarity [102]. Therefore, a higher value in the Silhouette index indicates the clusters are well-separated, whereas a value closer to 0 indicates overlapping clusters and negative values mean the data points are misclassified [102, 104].

Calinski-Harabasz (CH) Index: This is commonly known as the Variance Ratio Criterion. The value can be calculated as the ratio of the dispersion between clusters and the dispersion within clusters. The CH index for k clusters is defined by the formula below [108]:

$$\mathrm{CH}\ (k) = \left[\frac{B(k)\ (\mathrm{n}-\mathrm{k})}{W(k)\ (\mathrm{k}-1)}\right] \quad (5)$$

k represents the number of clusters. n refers to the total number of data points in the dataset. B(k) and W(k) indicate the inter-cluster divergence and intra-cluster divergence [108]. B(k) calculates the distance between cluster centroids, showing if the clusters are distinguishable from each other. W(k) quantifies the dispersion of data points within each cluster [100]. The higher the value of B(k), the more dispersion or variability between the clusters, while the smaller the value of W(k) is, the closer the relationship is within the cluster. Therefore, a larger ratio of the CH index indicates better clustering and well-separation between clusters [108].

Davis-Bouldin (DB) Index: The DB index is another method for determining the clusters. It is calculated as the average similarity of each cluster compared to the cluster with the most similarity. The concept was originally defined, and the detailed formulas for the steps were used in the research on cluster separation measures by Davies & Bouldin [105]. Here are the steps [106, 107]

Step 1: Calculate the intra-cluster dispersion, meaning the average distance between each data point within the cluster and its centroid.

Step 2: Calculate the distance between the centroids of each cluster and the closet cluster

Step 3: Calculate the ratio between the within clusters distance (average distance of the data points in the cluster to the cluster center) and the between cluster distance, which is the distance between the centroids of the clusters to their closest cluster.

Step 4: Repeat Step 1 to Step 3

Step 5: Calculate the average ratio for all the defined clusters.

The formula is represented by [107]:

$$\mathrm{DB} = \frac{1}{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{\max}{\mathrm{j} \neq \mathrm{i}} \left(\frac{\mathrm{s_i} + \mathrm{s_j}}{\mathrm{d_{ij}}}\right) \quad (6)$$

k indicates the number of clusters, s_i is the average within-cluster distance in the cluster i and s_j is the average within-cluster distance in cluster j, and $\mathrm{d_{ij}}$ is the between clusters distance of cluster i and cluster j. The higher ratio of the DB index shows poorer separation among the defined clusters, and the lowest DB index indicates better separation [106, 107].

***Gap Statistic*:** The Gap Statistic uses the Monte Carlo simulations of the sampling process to generate the reference dataset under the null hypothesis, assuming no distinct clustering. It compares the within-cluster dispersion, which is the sum of squared Euclidean distance within individual clusters, as defined in equation (1), for different numbers of clusters (k) with the expected value under the null reference distribution [109]. The Gap Statistic steps are illustrated below, as shown in [109]:

$$Gap_n(k) = \mathrm{E}_n^* \log(W_k^*) - \log\ (W_k) \quad (7)$$

$$sd(k) = \sqrt{\frac{1}{B}\sum_b\ (log(\mathrm{W}_{kb}^*) - \overline{w})^{\,2}} \quad (8)$$

$$s_k = \sqrt{1 + \frac{1}{B}} \cdot sd(k) \quad (9)$$

$$Gap(k) \geq Gap(k+1) - S_{K+1} \quad (10)$$

In equation (7), $\mathrm{E}_n^* \log(W_k)$ represents the expected value of log-transformed W_k^* (within-cluster dispersion for k clusters) under the null reference distribution. n refers to the total number of data points in the dataset. The Gap statistic, $Gap_n\ (k)$, calculates the difference between the observed $\log W_k$ and its expected value under the null hypothesis, indicating the deviation of the observed cluster from the expected values.

Equation (8) shows the standard deviation of the Gap statistic, B is the number of B reference datasets through bootstrapping (B) and $log(\mathrm{W}_{kb}^*)$ is the log within-cluster dispersion for the b-th reference datasets. The average log within-cluster dispersion (W_k^*) across all reference datasets is shown as $\overline{w}$. Equation (9) adjusted for standard deviation for variability and S_{K+1} is the adjusted standard error for k + 1 clusters (Equation (10)). The optimal number of clusters (smallest k) is chosen as adding more clusters (k + 1) does not improve the quality of the k clusters [109].

2.4.1.2 Related Research Using K-Means Clustering

Research has adopted k-means clustering to group observed populations with similar behavioural factors to study the impact of these clusters on the outcome of interests. This method has been used for pattern identification and the clustering of phenotypes [95, 99].

Quick [111] applied k-means clustering to identify the home environment of parents with preschool children in a cross-sectional study. Among the selected variables in socioeconomic and health-related behaviours of obesity, the mother's intake of fruit and vegetables, time in sedentary screen time, intake of sugary drinks, and levels of physical activity were used as inputs for clustering analysis. K-means clustering partitioned two clusters that distinguish mothers who were less healthy and inactive and mothers who were healthy and active [111]. Among the latter cluster, mothers consumed more fruits and vegetables with a lower-level intake of sugary beverages daily, spent less than four hours per day in sedentary screen time (categorical (Yes/No)), and were more active in physical exercises than the prior cluster.

Besides the obesity-related behaviour factors, Orsoni et al. [112] applied the k-means clustering to identify students' cognitive abilities among attention, visual perception, general intelligence and the awareness of phonology in Italy. Investigating the insights of students' cognitive abilities helps teachers create more personalized teaching strategies [112]. In addition to k-means clustering, the researchers utilized the Artificial Neural Network to predict the generated clusters. Nine clusters were generated based on working memory and visuospatial attention values. The characteristics of the nine groups range from very high, high, average, low

and very low performance in working memory and visuospatial attention [112]. Similarly, Kundu et al. [115] clustered students' risk profiles using multi-domain factors using k-means. Three clusters were generated, corresponding to youth groups as the low risk that contains higher fruit and vegetable consumption, no substance use, the highest level of physical activity, moderate risk group as the second lowest number of students engaging in the low-risk group, and high-risk group who involved more in violence, the lowest level of physical activity and highest engagement in screen time.

Additionally, k-means clustering has also been applied in clinical practices. Lin et al. [113] and Violán et al. [114] have utilized this technique to discover the clinical characteristics of the patient's cohort. The prior study discovered the multimorbidity patterns of metabolic disorders, mental health behaviour-related disorders, and hypertension diseases in men and women based on diagnosis in electronic health records [113]. The latter classified patients with metabolic obesity from Shanghai Tenth People's Hospital based on the clinical measure of glucose AUC, insulin and uric acid. Four groups of patients were clustered, ranging from the healthy metabolic obesity group, which had the lowest level of comorbidities, to the hypometabolic obesity group, which had the lowest glucolipid metabolism and extremely high glucose [114].

2.4.2 Hierarchical Clustering

Unlike k-means clustering, hierarchical clustering establishes the hierarchy of clusters in a tree structure in either an agglomerative or divisive approach based on the distances between the subgroups of the clusters [93, 115]. Agglomerative refers to the "bottom-up" approach, as the data points within the clusters and pairs of clusters were combined and moved up the hierarchy until clusters have been incorporated into a single large cluster [93, 99]. In contrast, divisive is a "top-down" approach, all data points start in one cluster and split recursively into sub-clusters using dissimilarity measurement [93, 95]. After either of the two approaches, the group characteristics for the number of clusters can be decided by slicing the dendrogram [93]. The advantage of hierarchical clustering compared to k-means is that it provides more informative cluster results for decision-making by providing the tree-structure hierarchy [93]. There are also some drawbacks to this technique. This clustering technique cannot handle extremely large datasets [95]. Like k-means, hierarchical clustering is sensitive to outliers and skewed distribution in the datasets.

An example of using hierarchical clustering is in customer segmentation. Hung et al. [116] applied a bottom-up hierarchical clustering technique to group customers into different categories based on their shared traits in location, age, and spending habits. The dendrogram visualized the customers into three groups: medium to high credit limits but limited purchases,

the second group with lowest credit limits and least purchases, and the third group with highest spending with low to high credit limits [116]. The result of the hierarchical clustering provides insights into customers' characteristics for designing market strategies [116].

2.5. Supervised Learning

Unlike unsupervised learning, supervised learning utilizes predictors as inputs and outputs as labels [93]. This is performed using training data to predict the labels based on the sets of input data [75, 93]. This learning type is divided into classification and regression objectives [93]. Using childhood obesity as an example, the regression task would use a continuous BMI as the predicted label, while the classification task focuses on the categorical form of whether the youth is obese or not. Classification problems can be further divided into binary and multiclass classification. Binary classification refers to the classification task using two categories of labels, such as "yes and no" or 'disease detected and not detected' in medical research [93]. A multiclass classification task includes labels with more than two categories [93]. For example, study 2 used the youth's BMI status, healthy weight, overweight and obese, and unknown weight as the predicted label. Another example of a scenario that applies multiclassification is text classification, where the categories include different types of news text [93].

To evaluate the performance of the supervised model, performance metrics are used to analyze how well the ML predicts the outcome of interest [69]. In the case of a classification task, common metrics are accuracy, precision, recall, and F1 score. Accuracy refers to the ratio of predicted correct samples over the total samples. Precision refers to the accuracy of the positive predictions over the numbers of true positive and false positive, calculated as true positive over the sum of true positive and false positive. Recall refers to the correct identification of the true positive cases, calculated as true positive over the sum of true positive and false negative. The F1 score is the harmonic mean of precision and recall [75]. Additionally, the area under the curve of the ROC curve (AUCROC) is used to measure the separability power of the classifiers [75]. To evaluate the performance in regression, common measures are mean squared error or mean absolute error [75].

Besides the evaluation metrics, loss functions are a critical part of ML that measure the difference between the model-predicted and actual output. The loss function types differ based on regression or classification problems [253]. The common one for regression is called the Mean Square Error (MSE), known as L2 loss. It is the average of the square difference between the ML model's predicted and actual results [253]. Another regression loss function is Mean Absolute Error (MAE), or L1 loss. Instead of the square difference, this function calculates the average absolute difference between the ML model predicted and the actual values [253]. Compared to MSE, MAE is less sensitive to outliers since it weighs the error terms equally

regardless of their magnitude [253]. For classification tasks, Binary Cross-Entropy is a performance measure commonly used in ML models. It quantifies the prediction that represents the probability of a data sample belonging to a binary class with a value between 0 and 1. If the classification problem involves multiple-class labels, Categorical Cross-Entropy is appropriate to use [253].

This dissertation's theme is supervised learning with a focus on classification tasks. The overview of the supervised learning models focuses on ones that solve classification problems instead of regressions.

2.5.1 Overview of Classification Algorithms

The following ML algorithms were chosen based on a review of studies that identified attributes associated with youth overweight and obesity. Among the eligible 28 studies from the Chatterjee et al. [117] review, K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), and Extreme Gradient Boosting (XGBoost), are the most common classifiers that were used in the literature review on youth obesity and related risk factors, achieving both high accuracy and specificity as well as working in the multi-classification environment [117, 118, 119, 120, 121, 122, 123].

Naïve Bayes (NB)*:* This algorithm is based on Bayes' theorem, assuming that the attributes of a given class are unrelated to the values of other attributes [93]. It is capable of handling binary and multiple-class problems, and it has been utilized in spam filtering, text classification, and childhood obesity research due to maintaining high accuracy in processing large databases [93, 124]. One of the advantages of this classifier is the small amount of training data required for estimating the parameters [93]. The established posterior probability of the outcome variable based on the predictor variables, according to Bayes theorem, can be represented as:

$$P\,(y = j|x_1, \dots, x_n) = \frac{P\,(x_1, \dots, x_n|y = j)\,P(y=j)}{P\,(x_1,\dots,x_n)} \qquad (11)$$

P (y = j) is the initial or prior probability of y taking on the value of j, and $P\,(x_1, \dots, x_n|y = j)$ is the posterior probability after observation.

K-Nearest Neighbor (KNN)*:* KNN is non-generalizing learning algorithm that stores all instances related to the training data in n-dimensional space [93]. It is used for classification problems and generates a boundary to classify the data according to the closest neighbors [124, 125]. This algorithm determines the data point category based on similarity among its k closest neighbors. The measure of closeness is generally based on the Euclidean distance [125]. The

hyperparameter k is used to calculate the nearest k neighbors for predictions [69]. KNN is suitable for both classification and regression tasks [124]. The advantages of this algorithm include simplicity, effectiveness and easy implementation.

Logistic Regression (LR): LR is widely used in ML applications as supervised learning and statistical models [93]. It has been utilized to benchmark other machine learning classifiers [75]. LR assumes that the predictor variables and the probability (log odds) of the outcome variable (e.g., obese or not) have a linear relationship [126]. LR uses the logistic function to estimate the probability and is represented by the sigmoid function [93]:

$$g(z) = \frac{1}{1+\exp(-z)} \quad (12)$$

LR works well when the dataset is linearly separable, but this also shows that LR is affected by the assumption of the linearity between the dependent and independent variables. It has been commonly used in classification but can also be adapted for regression problems [93].

Support Vector Machine (SVM): SVM is a supervised ML learning technique used in regression and classification analysis. It generates a maximum margin separator to determine boundaries with the largest distance [69, 75]. SVM creates a hyperplane in a high dimensional space to separate the data points from different categories [69]. The purpose of SVM is to identify the hyperplane that can help classify the data point in a dimension that contains several features. SVM can handle both binary and multi-class classification cases, where the algorithm learns from the data input of potential attributes contributing to the development of obesity status [127]. In the case of binary classification, SVM separates the hyperplane calculated from the support vectors and maximum margin separator [69]. The advantage of SVM is that it can generalize the new samples better even though the inputs contain bias. Due to its non-parametric features and local operation, it does not need to be limited to linear data and can handle large datasets in training [128].

Random Forest (RF)*:* RF is an ensemble classification technique that uses random data subsets and creates several individual decision trees [69, 93]. It builds trees by randomly selecting features in addition to choosing random data subsets [69]. Within RF, individual trees predict the category, and then the model prediction is based on the highest votes for a particular category [69]. RF operates based on the Gini index and fits several decision trees in a parallel position [69, 93]. By collecting the results from each tree, RF can reduce variance, handle noise, and be more resilient to outliers, making it more robust in prediction than decision trees [69, 93, 129, 130]. This algorithm can be applied to both classification and regression problems [93]. In classification, the output is determined by the majority of the trees [130].

Extreme Gradient Boosting (XGBoost)*:* XGBoost is a popular boosting algorithm. In boosting, various models are utilized in training sets in order, and the new model assesses the residual errors and misclassification from the previous model [130]. XGBoost uses the parallel tree gradient boosting algorithm. It also includes advanced regularization, such as L1 and L2, to overcome overfitting and improve the model's generalizability [93]. XGBoost is considered a powerful ML algorithm for handling large datasets, but it lacks interpretability and can be affected by overfitting [93, 130].

2.5.2 Supervised ML Research in Childhood Obesity

Given the complexity of childhood obesity with multi-domain impactful factors, ML has been utilized to handle the high-dimensional relationships between these variables and overcome the limitations of the SM, such as generalized linear models. Variables used in this type of SM are either reduced or similar since the number of predictors is more than the number of instances, creating limitations in predictive power [75]. ML has the capability of discovering complex and non-linear relationships between the predictors and outcomes automatically [75]. Several studies using ML approaches on youth obesity have been summarized below.

Abdullah et al. [132] applied the classification techniques of the Bayesian Network, Neural Networks, SVM, and Decision Tree (J48) to predict childhood obesity at the age of 12 among 153 primary schools in Malaysia. The input data was based on the student questionnaires, which addressed the students' sociodemographic, dietary assessment, and physical activity. The authors compared the performance of the implemented classifier and found J48 to outperform others [132].

In a study to evaluate the value of a BMI screening program in the US, Zare et al. [39] used the kindergarten BMI z-scores as the main predictor among public school students to predict the risk of childhood obesity at grade 4. This study employed a Decision Tree (DT), Artificial Neural Network (ANN), Random Forest, and Logistic Regression for the classification [39]. The model further generated the feature importance from the classifiers and discovered the important variables, including kindergarten BMI z-score (most important), family income, and parent’s educational level [39].

In addition to the data acquired from the BMI screening program, [45] predicted early childhood obesity using a clinical-based cohort. Specifically, this study utilized the attributes in a hospital database, targeting youth between the ages of three and five. Attributes include the children’s age, gender, height, weight, birth weight, physical activity, parental data, and habit of food consumption before age five [45]. The authors applied SVM and ANN to predict whether

children were either obese or not obese before the age of five. The model performance showed robustness in prediction, with ANN achieving the highest accuracy of 96.81%, 96.73% in precision, 97.36% in recall, and specificity of 92.85% [45].

The prediction of youth obesity also incorporated the electronic health record (EHR) as the data source in clinical research. Pang et al. [83] generated a cohort of youth encounter records from the Children's Hospital of Philadelphia. This study used the BMI status before the age of 2 and predicted the BMI status between ages three and seven using Decision Tree, Bernoulli Naïve Bayes, Gaussian Naïve Bayes, LR, SVM, and Neural Network and XGBoost [83]. XGBoost outperformed the other prediction models, achieving the ROC-AUC score of 0.81. This study further acquired the top 20 important factors for interpreting the XGBoost prediction and identified weight, weight-for-height, race, and body temperature as the important factors in the predictive model [83].

Using a Youth Risk Behaviour Surveillance System in 2015, Zheng & Ruggiero [121] focused on using KNN, an ANN, an improved DT, and binary LR for predicting obesity among high school students. This cross-sectional study collected data from students in Tennessee attending grade 9 to 12 public and private high schools. Three health-related categories, including factors of sedentary behaviours, energy intake, and physical activity, were used as inputs to train the ML models [121]. The binary LR resulted in the lowest model performance, while ANN achieved an overall accuracy of 84.22% and specificity of 93.44% [121]. They also discovered that increased levels of physical activity and eating breakfast had significantly decreased the risk of obesity, and consumption of sugary drinks and more sedentary time increased the risk of obesity [121].

All the above studies have validated that ML is robust for predicting youth obesity in school and clinical settings. More importantly, ML has shown promising performance using EHR data, containing millions of data records and massive medical measurements, including procedure and drug prescriptions [75]. Its application of big data can further support healthcare management and clinical decision-making in large populations [83].

2.6 Deep Learning Model in Time Series Analysis

The deep learning (DL) model is a subset of the machine learning approach within artificial neural networks [93]. Unlike traditional ML models, which are unable to improve performance even with the increase in data points in the training set, DL contains computational architecture with processing layers to overcome this saturation and process complex data structures such as speech, image, and text [75, 93]. Common DL models include Multi-layer Perceptron, Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and

Convolutional Neural Network (CNN) [93]. Another advantage of DL models over traditional ML methods is that many have sequential layers that can be stacked to learn the hierarchical representation of the various aspects of interrelated data [75].

The applications of DL models have been incorporated well with time series analysis in forecasting [133]. Time series forecasting refers to estimating the future of time series and certain events evolve over time based on sequential, time-ordered data for decision-making [133]. This forecasting in healthcare has been frequently utilized in predicting diabetes mellitus, blood glucose concentration and cancer growth. Besides the utilization in the medical field, the adoption of DL in healthcare has also started with childhood obesity [75]. Previous research on youth obesity prediction focused on the obesity status at a single time point without encountering the temporal change in the child's health pattern [134]. For instance, Zare et al. [39] predicted the childhood obesity status at grade 4, Berge & Everts [45] at age five using the prior year's inputs, and Abdullah et al. [132] assessed youth obesity at the age of 12. Since youth weight status varies at different time points, it is necessary to predict the risk of youth obesity based on the temporal relationships within the data [134].

Additionally, Gupta et al. [134] mentioned that single-point predictions based on non-longitudinal pattern data are a major limitation due to the rapid change in child growth. In the reviewed article, fifteen studies demonstrated the significant relationship between rapid youth growth, such as weight gain during adolescence, and future risk of being overweight or obese [135]. Instead of using longitudinal data, previous thesis studies extracted predictor measures as snapshots of youth behaviour [38, 45, 132] at particular times for developing ML prediction models. Moreover, another clinical study on type-2 diabetes mellitus used time series data analysis with deep learning models. The rationale for using time series analysis is to capture changes in patient measurements to predict diabetes conditions [136]. To overcome the limitations of traditional ML models and investigate factors in a longitudinal format for predicting youth obesity, it is necessary to develop DL models to capture the temporal sequence.

2.6.1 Overview of Deep Learning Models

The section below focuses on the overviews of LSTM and RNN, the major models used in Study 3's analysis.

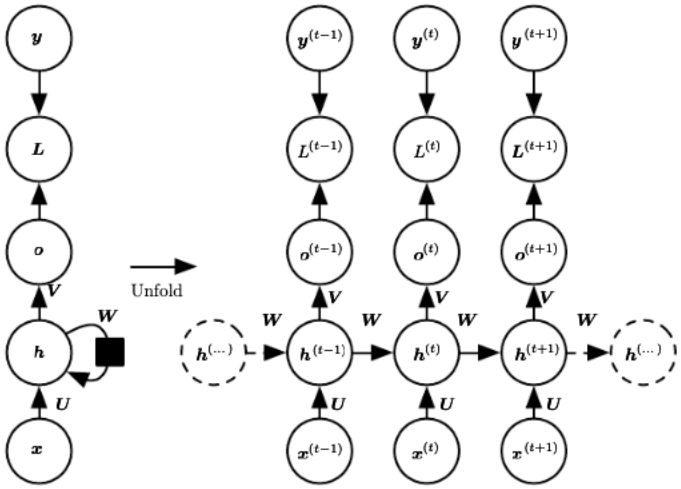
Recurrent Neural Networks (RNN)

RNN has been implemented in time series analysis [75]. It aims to explore the relationship between current time samples and past samples. RNN uses the neurons from the previous time steps and maps the inputs to the output sequences [133]. The basic structure of RNN contains the input, hidden, and output layers. The flow of the unfolded RNN can be represented as [137]:

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}; W) \quad (13)$$

In this equation, x is the input data at t, t refers to the time step, h represents the hidden state in the network, and W represents the weights of the network connections, including input to hidden, hidden to hidden, and hidden to output [137]. The inputs of x are incorporated into the hidden state (h), and the output processes the information from the input and hidden state [137]. The RNN computational graph for processing the input state, hidden state, and output state is shown below [138]:

Figure 2.6 RNN computation of training loss that shows the input state to output state, from [138]



This graph computes the training loss of RNN where L indicates loss and measures the distance each output (o) is from the related training target y [138]. The input x at the previous time step and when fed into the hidden state expressed in the term of a weight matrix U, the weight matrix W represents the hidden-to-hidden recurrent connection, and the weight matrix V represents the hidden-to-output connection, while $o(t-1)$ is the output of the network [138]. Simple RNN can capture the short-term dependence, but since it is trained by gradient descent and backpropagation through a time algorithm, it is less capable of memorizing long-term information or dependencies and suffers from varnishing gradients [133, 137]. The generated gradients from the backpropagation can be either small or large, causing unstable training in RNN.

Long Short-Term Memory (LSTM)

To overcome the vanishing gradient issue in RNN, LSTM has been introduced and has become one of the most successful computing methods for long sequential aspects with several steps [133]. The LSTM model contains three main components: the input gate, forget gate, and output gate [137]. The steps of the LSTM mechanism can be represented as the below formula [139]:

$$\text{Input gate: } I_t = \sigma\left(X_t W_{xi} + H_{t-1} W_{hi} + b_i\right) \quad (14)$$

$$\text{Forget gate: } F_t = \sigma\left(X_t W_{xf} + H_{t-1} W_{hf} + b_f\right) \quad (15)$$

$$\text{Output gate: } O_t = \sigma\left(X_t W_{xo} + H_{t-1} W_{ho} + b_o\right) \quad (16)$$

$$\text{Cell State: } C_t = F_t \odot C_{t-1} + I_t \odot \tanh\left(X_t W_{xc} + H_{t-1} W_{hc} + b_c\right) \quad (17)$$

$$\text{Hidden state: } H_t = O_t \odot \tanh(C_t) \quad (18)$$

Within the formula, while W_x represent the weight parameters and b_x represents the bias parameters. The forget gate determines which information should be removed from the cell state [137]. This layer is controlled by the sigmoid function, with weights ranging between 0 and 1 [137, 140]. Information is discarded if the value is closer to 0, while information is retained if the value is closer to 1. A new candidate value is created by the tanh layer, which is multiplied by the input gate to decide what information to retain in the cell state. $F_t \odot C_{t-1}$ represents the forget gate F_t multiplied with the previous C_{t-1} as the "forgetting" information. The output gate O_t with the sigmoid fuction multiples the current cell state C_t through the tanh function, resulting in values between -1 and 1, which represent the hidden state H_t. The output gate will decide whether to pass the value in H_t or discard it [137, 140].

An attention layer can also be embedded into the LSTM layer to pay more attention to the important features and improve the model prediction performance [141]. In Kaji et al. research [141], the attention vector (a_i) learned the weight for each feature and passed it to the output layer of the network. The authors calculated the attention (a_k) by the weights (W_k) and the k features in time steps through a softmax activation function, shown below [141]:

$$a_k = softmax\left(W_k x_k\right) \quad (19)$$

The final input was weighted by the attention vector a_k on feature k overtime, represented as [141]:

$$y_k = a_k \odot x_k \quad (20)$$

2.6.2 DL Research in Childhood Obesity

The rise of DL models in ML has been shown in predicting childhood obesity studies, specifically using EHR data [75]. The studies selected below have focused on LSTM or RNN. Other DL models for childhood obesity include the Artificial Neural Network and Convolutional Neural Network [144].

In the EHR data setting, Gupta et al. [134] applied an RNN architecture with LSTM cells to predict childhood obesity using both static and dynamic variables. They converted the 30-day time-point visit data within a two-year observation period into 25 time steps per patient [134]. The models were designed to predict obesity (1–3 years in advance) for patients aged 3 to 20 using EHR data. For example, the observation window was from ages 1, 2, and 3, while the prediction window targeted obesity risk at ages 3, 4, and 5. Each patient was required to have at least five years of data. In addition to the model architecture, the authors implemented an attention layer to better interpret the model predictions. They ranked important features based on the three most important visits and discovered that prior-year BMI and obesity status (obese or not obese) were the most important predictors of future obesity. Over the 3-year input window, the LSTM model achieved AUC values of 0.80, 0.93, and 0.92, outperforming the baseline models of random forest and linear regression.

Beyond predicting childhood obesity, research using RNN and LSTM also explores predicting the success of obesity interventions. Hasan et al. [142] analyzed communication between patients and providers during motivational interviews. They predicted the success of the program by using RNN with LSTM-based models, as well as Markov Chain and Hidden Markov Models. The conversation between providers, obese adolescents, and their caregivers was coded to capture patient behavioural responses. The goal of the program was to encourage behavioural change and improve self-efficacy through counselling [142]. Successful communication codes included positive change talk or language of commitment, while unsuccessful codes included talk of negative change or commitment. The final sample consisted of 5,143 sequences, of which 82.15% were positive and 17.85% were negative communication codes [142]. Due to the imbalanced data, the researchers developed deep learning models using undersampling and oversampling approaches. The RNN-LSTM model without target replication resulted in higher F1 scores compared to the Markov Chain and Hidden Markov Model, yielding F1 scores of 0.8677 with undersampling and 0.8381 with oversampling [142].

In another study on childhood obesity using EHR data, Cheng et al. [143] utilized LSTM to analyze a longitudinal dataset. The study included 73,957 historical encounter records from a sample of 6,614 children aged 0 to 4. Sex-stratified models were developed to predict outcomes

for boys, girls, and a combined group. Time steps were created from 2, 3, 5, or 8 previous encounters, followed by the prediction of BMI based on these historical data points. The predictors included static factors, such as the patient's race, birth weight, and mother's age at birth, along with time-varying factors, such as sleep quality and visit type. The model was trained on 80% of the data and validated on 20% [143]. This study used mean absolute error as the evaluation metric, with results showing that the error was lowest for models with five or eight historical encounters compared to those with two or three encounters [143]. Specifically, the mean absolute errors for the five-history encounter models were 1.02, 1.06, and 0.98 for boys, girls, and combined cohorts, respectively. In contrast, the mean absolute errors for the two-history encounter models were 1.49 and 1.60 for boys and girls, respectively.

In summary, DL models that emphasize temporal sequences are highly effective in analyzing the dynamic patterns and changing factors that occur across multiple stages of childhood to early adulthood. These models can incorporate the diverse types and sizes of data extracted from EHR or other healthcare systems [134].

2.7 Markov Chain Model (MCM)

This section gives an overview of the MCM and its applications in real-world scenarios. MCM was used in Study 3 to account for the transition probability in weight perception over time among COMPASS students.

The MCM consists of discrete-time and continuous-time variants. This introduction focuses on the discrete-time MCM to align with the approach in Study 3. The discrete-time MCM models the possibility of progression at distinct time points, such as days, months, or years [145]. The MCM is a stochastic process represented by a sequence of events, where the probabilities of the event depend only on the state achieved by the previous event [146]. It calculates the probabilities of events that occur as the state transitions to other states or remains the same as the previous state [149]. The discrete-time MCM can be expressed by the formula below [147]:

$$P(A_{n+1} = a | A_1 = a_1, A_2 = a_2, \dots A_n = a_n) = P\,(A_{n+1} = a | A_n = a_n) \quad (21)$$

In this formula, the probabilities of A_{n+1} depend on the current state of A_n . This illustrates that the movement of the discrete-time MCM reflects the state of the condition at the end of each cycle [145]. The transition probabilities can also be represented in a matrix form, called the transition probabilities matrix, as shown below [148]:

$$P = \begin{bmatrix} P11 & P12 & \cdots & P1r \\ P21 & P21 & \cdots & P2r \\ \cdots & \cdots & \cdots & \cdots \\ P21 & Pr2 & \cdots & Prr \end{bmatrix}$$

In this matrix P, the states are either 1, 2, or r. Furthermore:

$$\sum_{k=1}^{r} P_{ik} = \sum_{k=1}^{r} P(X_{m+1} = K \,|X_m = i) = 1 \quad (22)$$

The probability of P_{ij} for all values of K in rows must equal 1 when $P_{ij} \geq 0$

The MCM has been widely implemented in studying chronic diseases [145]. Since chronic illness is often described as an independent health state, the MCM is suitable for mapping the progression of diseases across different states [145]. Several studies have used the MCM to assess the transition status of overweight and obesity among youth and adults.

Chen et al. [149] fitted a three-state MCM model on a longitudinal weight status dataset with complete assessments for 1494 elementary school children. The three states—healthy weight, overweight, and obese—monitored the changes in weight status from baseline to understand the transition probabilities over time, specifically during fall and summer [149]. The results showed that most children were more likely to stay in their previous state. They were also more likely to become overweight or obese during the summer compared to fall [149]. In terms of gender, socioeconomic status, and racial and ethnic groups, boys had a higher likelihood of becoming overweight or obese across all periods compared to girls. Non-Hispanic Black and Hispanic children, as well as those from lower socioeconomic backgrounds, had higher probabilities of transitioning to obesity [149].

In another cohort study among Chinese school children, a time-homogeneous three-state Markov model was applied [150]. Hu et al. [150] examined pattern changes among 2344 children aged six, using five years of physical examination data. They observed the transition probabilities among normal weight, overweight, and obesity, with results stratified by sex. Among children who initially had normal weight, there was a 6.3% probability of transitioning to overweight and a 0.8% probability of transitioning to obesity within one year of observation. In contrast, 24.8% of overweight children returned to normal weight, while 17.5% transitioned to obesity [150]. Overall, children were more likely to maintain their weight status during the first year, while 20.1% of children who were obese at baseline remained overweight over the five-year period [150].

Beyond using MCM to monitor children's weight status, research has also applied this method to adults. Tang et al. [151] observed weight status patterns in children and predictive

lipid markers among Chinese adults. The authors applied a multi-state MCM to the healthy weight, overweight, and obesity status of males and females, young and middle-aged. In addition to calculating transition probabilities, the study further examined the effect of covariates on the transition model, stratified by sex. Among the age groups, the weight status of young adults fluctuated more than that of middle-aged adults. Specifically, young male adults were more likely to develop obesity compared to middle-aged male adults. Regarding the impact of covariates, age, triglyceride levels, and high-density lipoprotein cholesterol were significant variables influencing male and female weight transitions. With a unit increase in high-density lipoprotein cholesterol, the probability of transitioning from overweight to healthy weight increased by 62% in females and 37% in males.

Chapter 3

Study Rational and Research Questions

This section provides an overview of the purpose of conducting the three studies and the corresponding research questions, written in manuscript style. It also provides an overview of the study rationale and research questions.

3.1 Study 1: Clustering Analysis of COMPASS Student's Behaviours in Pre-COVID and Post-COVID Cohort

3.1.1 Rationale

Adolescence is an essential period of youth behaviour development. Behaviours developed in terms of nutrition and substance use, including smoking, drugs, and alcohol use, were more likely to be maintained into adulthood [1,2,3]. During this critical developmental stage, COVID-19 negatively affected the youth's daily routine, including school closure, changes in the learning curriculum, and social quarantine [9, 10]. Social distancing prevented students from participating in physical activities and increased the likelihood of a sedentary lifestyle [8].

Clustering analysis has been commonly used to group samples with similar homogenous traits. It is suitable for describing the students' profiles and any hidden behaviour patterns. Previous studies have focused on clustering analysis in the medical area, such as multimorbidity characteristics among patients towards disease progression or clustering on a particular trait such as mental health, learning outcome, and cognitive abilities among students [112. 113, 114, 115]. For COVID-19 clustering research, studies were also focused on the single prospect of student behaviour, such as student academic performance during the pandemic [152, 153, 154]. However, limited research has been conducted on clustering students' profiles pre- and post-COVID-19. Research on the impact of the pandemic disruption is necessary to understand the change in various student behaviours compared to those in prior COVID-19.

This study aims to apply k-means clustering analysis followed by an RF classifier to cluster students with similar characteristics pre- and post-COVID-19 and identify the important factors associated with the established clusters. Compared to the previous study, this study analyzed potential covariates in multiple domains, including mental health, healthy eating, physical activity, screen time, and sedentary lifestyle, instead of focusing on specific aspects, providing a more comprehensive understanding of student behaviour profiles.

Additionally, this study will allow researchers and school principals to review the important factors for clustering prediction pre- and post-COVID-19. Understanding the difference between the two periods can lead to a more targeted school-based intervention to improve overall adolescent health.

3.1.2 Research Questions

1. What factors are associated with the clustering of student profiles in the 2018-19 (pre-COVID) and 2020-21 (post-COVID) COMPASS cohorts?

2. What are the similarities or differences of the factors associated with the clustering of student behaviour pre-COVID-19 and post-COVID-19?

3.2 Study 2: Building an Interpretable Machine Learning Model for Selecting Modifiable Attributes of Childhood Obesity

3.2.1 Rationale

Childhood obesity is a major public health concern. It is a critical risk factor for non-communicable diseases, including cardiovascular disease, hypertension, type 2 diabetes, and non-alcoholic fatty liver disease in both developed and developing countries [20, 21]. Obesity status during childhood increases the likelihood of transitioning into obesity in adulthood [21]. Due to the multi-dimensional risk factors for developing childhood obesity, traditional statistical models lack the robustness to handle the volume of input features and the complexity of non-linear relationships [82]. ML models have effectively handled large and complex datasets [90]. Therefore, analyzing the COMPASS dataset, which spans multiple domains of students' lives, using ML techniques is necessary. Study 2 focused on youth BMI, a weight-related measure from Study 1, and explored the factors associated with BMI prediction.

Study 2 utilized various classification ML techniques such as RF, SVM, Multinomial-NB, LR, KNN and XGBoost as potential candidates for building a predictive model on childhood obesity. Study 2 further applied the Shapley Additive Explanations Technique (SHAP) to address the interpretability challenges in ML models [208]. The findings illustrated the application of ML to youth obesity research in both complete case analysis and multiple imputation approaches. Additionally, Study 2 compared data preprocessing strategies, retaining

all variables, feature selection, and dimensionality reduction techniques, on model performance to ensure the most appropriate predictive model for defining youth obesity. Lastly, the SHAP plot explained the magnitude and direction of factors affecting health, weight, overweight, obesity, and unknown BMI, aiding in understanding their importance in model prediction. This prediction model provides insights into critical factors associated with COMPASS students' BMI status and informs strategies to prevent youth obesity.

3.2.2 Research Questions

1. What important factors are associated with youth BMI status in the 2020-21 COMPASS student cohort using ML classifiers from the CCA and MI approach?

2. How do the results of ML-generated factors compare with the common factors leading to childhood obesity in the previous literature? What are the similarities and differences?

3.3 Study 3: Exploring the Longitudinal Weight Perception Transition and Associated Factors Using the Markov Chain Model and Deep Learning Approach

3.3.1 Rationale

The relationship between weight perception and actual weight status among teenagers has been well-documented. The concern is that weight misperception among youth contributes negatively to their self-esteem and body dissatisfaction [37]. Teenagers who perceive themselves as overweight or obese tend to engage more in extreme weight loss programs, even when their BMI falls within the healthy range. Accurate weight perception is positively correlated with better adolescent health, as youth with accurate perceptions participate more in physical activities compared to those who perceive themselves as overweight [230]. Previous studies modelling dynamic changes during adolescence focused on weight changes over time rather than weight perception [149, 150]. Therefore, it is essential to understand the factors associated with youth weight perception, particularly how it changes over time. Since weight perception was the most important factor associated with BMI prediction in Study 2, Study 3 used weight perception as the outcome to observe its transitional pattern and time series analysis.

Study 3 had two objectives. The first objective was to analyze the trajectory of youth weight perception using a three-year linked dataset. This study used the MCM to reveal transition probabilities between underweight, about the right weight, and overweight categories.

Additionally, multinomial logistic regression was applied to further understand the significant factors associated with weight perception in Year 2 and the changes in weight perception from Year 1 to Year 2. The second objective was to apply deep learning techniques, such as LSTM and RNN, to predict weight perception in Year 3 based on predictors from the previous two years. Current studies on youth obesity using LSTM and RNN have focused more on clinical research, such as Electronic Health Record (EHR) data [134, 143, 142]. This study focused on sequential data from prior years to predict future weight perception, which is necessary to explore how temporal changes in predictors (e.g., physical activities, healthy eating, mental health) affect youth weight perception. Similar to Study 2, Study 3 employed SHAP analysis to interpret the importance of factors associated with predicting weight perception in Year 3.

The findings of Study 3 address gaps in understanding dynamic changes in youth weight perception over time. These results provide valuable insights to public health professionals, helping them develop targeted interventions for youth obesity at different time points, focusing on factors associated with self-weight perceptions.

3.3.2 Research Questions

1. Study 3 Q1: How do the estimated probabilities of weight perception change from Year 1 (2020-21) to Year 2 (2021-22) and from Year 2 (2021-22) to Year 3 (2022-23) using the Markov Chain Model?

2. What factors are associated with the weight perception in Year 2 and the weight perception transitional change from Year 1 to Year 2?

3. In the deep learning approach, what are the important factors associated with predicting weight perception in Year 3 from the previous two years?

Chapter 4
Clustering Analysis of COMPASS Student's Behaviours in Pre-COVID and Post-COVID Cohort

4.1 Abstract

Background
Assess clustering to identify subgroups based on multidimensional factors in a school setting that affect the youth population before and after COVID-19. Profiling youth cohorts in Canada is crucial to understanding behaviour and risk factors to guide prevention development efforts.

Method
The pre-COVID cohort (2018-19) and post-COVID cohort (2020-21) were from the COMPASS study. K-means clustering, Random Forest classifiers, and Multinomial Logistic Regression were applied to group students with similar behaviours. The goal was to identify the factors associated with the clustering and assess the strength of the association between the top five important factors and being in certain cluster groups.

Findings
High-risk clusters were distinguished by substance use and lower psychological well-being. In the pre-COVID cohort, students who skipped more than 20 classes, had incomplete homework, and engaged in bullying were more likely to be grouped into one sub-cluster compared to those without these behaviours. Students who tried to lose weight, had overweight perceptions, and were less physically active had a higher likelihood of being in a specific sub-cluster. In the post-COVID cohort, students with poor mental health, higher BMI, increased cannabis use, and lower physical activity were more likely to be grouped into one sub-cluster compared to students with opposite behaviours.

Conclusions
The results indicate that school engagement, mental health, substance use, and weight-related factors were significant indicators of student profiles in both pre- and post-COVID cohorts. The identified important factors provide critical insights for school principals and policymakers to develop targeted interventions focusing on specific student behaviours to prevent future adverse outcomes.

4.2 Introduction

Adolescence is a critical period for the formation of behaviours and for undergoing physical, social, and psychological changes. The risk of unhealthy behaviours developing during this stage can persist into early adulthood, leading to adverse outcomes [1, 2, 3]. For instance, nutritional issues such as malnutrition or substance use, including alcohol, smoking, and drugs, are significant factors affecting the long-term health of youth [4, 5]. Behaviours adopted between the ages of 12 and 18 tend to persist throughout life [1, 8].

During this key developmental period, the COVID-19 pandemic significantly affected schools and Canadian youth. Among the 5.7 million impacted children, 64% of youth aged 15-24 reported that their mental health was negatively affected during the pandemic [202]. Schools across Canada were closed, forcing a shift to remote learning. The school closures between 2020 and 2021 disrupted students' learning environments, mental health, physical activity levels, and eating habits [8, 203]. Declines in mental health, including increases in symptoms of depression and anxiety, were commonly reported during the COVID-19 pandemic. Several studies highlighted the negative impact of the pandemic on adolescent mental health, including decreases in quality of life due to difficulties with online learning and COVID-19-related worries [9, 10, 11, 12]. Besides mental health, a significant decline in physical activity and increased sedentary behaviour among youth due to social isolation were also reported [8, 9, 10]. Additionally, unhealthy eating patterns were exacerbated during home confinement, with evidence showing that many youths ate out of boredom and increased their consumption of unhealthy foods, such as snacks between meals [204].

Unsupervised machine learning algorithms can be employed without labelled training data to understand student behaviour profiles [93]. Among the various clustering analyses, k-means clustering is known for its popularity, efficiency, and simplicity in classifying study participants based on similarities [94]. It classifies groups into k clusters by calculating the cluster mean and grouping the data points based on the closest distance [96, 98]. Several studies have utilized k-means clustering to analyze heterogeneity in student characteristics and identify patient subgroups with a higher risk of developing certain disease complications [112, 113, 114, 115]. These studies successfully grouped patients or students based on similar traits and risk profiles, allowing patterns to be identified for intervention development or clinical practice.

Although previous studies have focused on specific aspects of youth behaviour, such as eating disorders or psychological well-being, few have considered multiple perspectives by profiling students based on their similarities using clustering analysis [112, 115]. Given the various behavioural aspects that arise during adolescence and the potential behavioural changes due to COVID-19 disruptions, it is essential to explore youth behaviour profiles before and after

the pandemic. This study aims to apply k-means clustering to group students attending grades 9–12 in the COMPASS study based on multiple domains, including substance use, mental health, physical activity, healthy eating, life changes due to COVID-19, and school connectedness. The primary objective is to investigate student behaviour profiles pre- (2018-19) and post-COVID-19 pandemic (2020-21) using k-means clustering, followed by a random forest machine learning model to identify the important factors predicting the clusters.

4.3 Materials and Methods

4.3.1 Study Dataset

This study acquired the dataset from the 2018-19 and 2020-21 cohorts of the COMPASS study (Cannabis, Obesity, Mental Health, Physical Activity, Alcohol use, and Sedentary Behaviour (COMPASS) study. This prospective ongoing study collects grade 9-12 students attending secondary schools in Alberta, British Columbia, Ontario, and Quebec in Canada [48]. The COMPASS study (2012-2027) collects student-level through the anonymous, self-administered COMPASS Student Questionnaire (C_q) annually on assessing healthy eating, sedentary behaviour, alcohol, tobacco, and marijuana use, bullying, academic achievement, physical activity, obesity, and school connectedness [48]. The participating students were under the permit of the use of active-information, passive-consent, and parental permission protocols [48]. Details about the dataset can be found on the study website (www.compass.uwaterloo.ca) [49].

4.3.2 Variables of Interest

To avoid discrepancies, the dataset includes only questions from all four provinces (Alberta, British Columbia, Ontario, and Quebec). The detailed list of final variables for each cohort can be found in Appendix B. The following behavioural variables were taken from each COMPASS questionnaire domain to represent the variables used for clustering analysis.

Mental Health

The mental health category included questions such as “How much do you agree with the following statement? I can talk about my problems with my friends”. The assessed responses were “strongly agree,” “agree,” “neither agree nor “disagree,” “disagree,” and “strongly disagree”. Anxiety and depression symptoms were evaluated by the Generalized Anxiety Disorder 7-item Scale (GAD7 [54]). Flourishing was measured based on Diener’s 8-item Flourishing Scale for how students perceived their psychological well-being [55]. The question asked, “How much do you agree with the following statements?”, and the response categories were “I lead a purposeful and meaningful life”, “My social relationships are supportive and

rewarding", "I am engaged and interested in my daily activities", "I actively contribute to the happiness and well-being of others", "I am competent and capable in the activities that are important to me", "I am a good person and live a good life", "I am optimistic about my future", "People respect me". The responses were "Strongly agree", "Agree", "Neither agree nor disagree", "Disagree", and "Strongly disagree". Another mental health-related variable was the Center for Epidemiologic Studies Depression Revised Scale (CESD) [23].

School Performance and Connectedness

The school-related category included bullying, academic achievement, and school connectedness scores. Bullying questions were about the ways students were bullied or how they bullied others. An example question was, "In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) I have not been bullied in the last 30 days" (Yes/No). An example question for academic achievement was, "In your current or most recent Math course, what is your approximate overall mark? (Think about last year if you have not taken math this year)", with the collapsed response being "80-100%", "50-70%", and "less than 50%". The school connectedness category involved questions about the support provided by schools in initiatives to ensure no one is bullied and to give the students the support they need to resist or quit tobacco. The overall school connectedness score was quantified using the National Longitudinal Study of Adolescent Health School Connectedness 5-item scale, ranging from 5 to 20, with higher scores indicating higher school connectedness [53].

Substance Use

This category included survey questions about smoking and tobacco use, vaping and e-cigarette use, alcohol use, cannabis use, and opioid use. An example survey question about cannabis use was, "If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have not used marijuana or cannabis in the last 12 months" (Yes/No). To determine smoking and tobacco use, the question was, "In the last 30 days, did you use any of the following (Mark all that apply)? I have not used any of these things in the last 30 days" (Yes/No). For opioid use, the question was, "Have you used or tried any of the following medications TO GET HIGH? A) Oxycodone (oxy, OC, APO, OxyContin, percs, roxies, OxyNEO)". The response categories were, "No, I have never done this", "Yes, I have done this in the last 12 months", and "Yes, I have done this, but NOT in the last 12 months".

Screen Time and Physical Activity

The screen time category referred to time spent on sedentary activities such as watching TV, playing video games, surfing the internet, and doing homework. For all the sedentary activities, the total time spent per day was derived from the question, "How much time per day do you usually spend doing the following activities?" (Hours or minutes). The physical activity category included measuring individual time spent doing physical activities and school-level

participation. The school level included questions such as, “Do you participate in before-school, noon hour, or after school physical activities organized by your school? (e.g., intramurals, non-competitive clubs)?”. The responses were “Yes,” “No,” and “None offered at school”. One of the individual-level physical activity questions asked about the average combined hard and moderate physical activity per day (in minutes).

Healthy Eating

The health eating category included questions about weekly consumption of fruit, water, snacks, vegetables, fast food purchased, and days of diet restriction. In the 2018-19 cohort, snacks were represented by questions, “From a usual school week (Monday to Friday), on how many days do you do the following? Eat snacks purchased from a vending machine, corner store, snack bar, or canteen off school property”. The response options ranged from “None” to “5 days”. For vegetable consumption, the question was asked, “Yesterday, from the time you woke up until the time you went to bed, how many servings of vegetables and fruits did you have? One 'Food Guide' serving of vegetables and fruit includes pieces of fresh vegetable or fruit, salad or raw leafy greens, cooked leafy green vegetables, dried or canned or frozen fruit, and 100% fruit or vegetable juice”. The response categories ranged from “None” to “9 or more servings”.

In the 2020-21 cohort, healthy eating responses were aggregated based on the total number of days in a week students participated in drinking water, eating food purchased at a fast-food place or restaurant, restricting food intake to change weight, and fruit and vegetable consumption. An example question was, “During the past week, on which days did you do the following? Eat snacks purchased from a vending machine, corner store, snack bar, or canteen”. if a student answered ‘Yes’ to this question from Monday to Sunday, the total number of days in a week of eating snacks would be seven. Skipping breakfast questions were also included. One representative question was, “If you do not eat breakfast every day, why do you skip breakfast? I am trying to lose weight” (Yes/No).

Weight

The weight-related category included the perspective on weight change intention, BMI, and weight perception. Weight change intention was assessed by asking, “Which of the following are you trying to do about your weight?”, with the responses being “Lose weight”, ‘Gain weight’, ‘Stay the same weight’, and ‘I am not trying to do anything about weight’. The derived BMI category (BMI_ACAT) is based on the WHO guidelines for ages 5 - 19 years of school-aged children and adolescents and was calculated from the valid response of sex, BMI (numeric version), and age [52]. The BMI category was self-reported by the student and collapsed as “underweight”, “healthy weight”, “overweight and obese” and “unknown” in this study. The weight perception was assessed by asking, “How do you describe your weight”? The response category were collapsed into “underweight”, “about the right weight”, and

"Overweight". Weight perception was also assessed by the question, "How do you describe your weight?". The collapsed categories were "Very underweight/Slightly underweight", "About the right weight", and "Slightly overweight/Very overweight".

COVID-19

This domain only appeared in and after the 2020-21 survey. It included questions about overall concern regarding COVID-19, feelings about the COVID-19 pandemic, its impact of COVID-19 on personal life, and knowledge about COVID-19. An example of the life impacted by COVID-19 was asked, "How has your life changed because of COVID-19? My stress has…", with response categories of "Increased", "Stayed the same/not applicable", and "Decreased". Similar questions were asked about boredom, loneliness, anxiety, time spent with my friend in person, anxiety, time spent on video games, social media, cigarette use, vaping, alcohol and cannabis use. To assess the knowledge about COVID-19, one question was, "Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) The use of hand sanitizer (e.g., Purell) helps prevent the transmission of COVID-19" (Yes/No).

4.3.3 Data Preprocessing

The two cross-sectional cohorts, 2018-19 and 2020-21, followed a complete case approach to handle missingness. This indicates that missingness within the variables was removed from the analysis, except for BMI. The missing BMI fields were classified as the 'unknown' category to avoid losing a large proportion of the sample size. The detailed percentage of missing in cohorts 2018-19 and 2020-21 can be found in Appendix B. All data preprocessing and clustering analyses were conducted in Python version 3.10.13.

A feature reduction step was necessary to prevent the curse of dimensionality to improve the learning algorithm's performance [22]. Due to the mixed data types in the dataset, principal component analysis (PCA) was not ideal for handling both quantitative and qualitative variables [205]. Instead, Factor Analysis for Mixed Data (FAMD) from the PRINCE library in Python was applied instead. Factor analysis is a dimensionality reduction technique that transforms the data into low-dimensional space [206]. It combined the features of both PCA and multiple correspondence analysis (MCA), partially performed on continuous and categorical variables. One study has shown that using FAMD on k-means clustering achieved the highest performance for large data sample sizes, especially when there were more categorical variables than continuous ones [206]. For the 2018-19 dataset, it started with the original 170 variables through FAMD and plotted the principal component index against the explained variance ratio. Different thresholds have been used in studies, but retaining 70% of the variability is common when analyzing reduced dimension quality [207]. The decision was to keep 112 components with an explained variance ratio of 0.8 (80%) from the original 172 variables. The same procedure was

applied to the 2020-21 dataset, and 100 components were retained from the original 150 variables.

K-means clustering analysis

To choose the optimal number of k, the elbow method through Within-Cluster-Sum of Squared Errors (WSS) calculation combined with the Kneedle algorithm that detects the elbow point, the Silhouette Coefficient, and Davies-Bouldin (DB) Index were the main evaluation methods to determine the number of clusters [105]. The numbers of clusters were chosen based on the elbow point, clusters with higher Silhouette Coefficient, and lower DB Index. The higher Silhouette Coefficient indicates a better formation of the clusters, and a lower DB Index was expected to achieve a well-separate number of clusters [102, 105]. K-means clustering was first applied to pre- and post-COVID cohorts using all predictor variables. Then, it was applied a second time, excluding the substance use variables, to define the factors associated with forming the sub-cohort. The COVID-19 question regarding the increased or decreased use of substance use variables were kept to observe the pandemic's impact. K-means clustering was conducted using scikit-learn (version 1.2.2) in the Python library.

Random Forest Model with Cluster as Outcome

The datasets with the clustering labels determined by k-means were split into 80% training and 20% testing data with the cluster column as the outcome. A supervised learning algorithm was used to explain the clustering result. The Shapely Additive exPlanations (SHAP) values quantified how each variable in the sample contributes to the label's prediction, and the summary plot explained the factor contribution to the model prediction [208]. Additionally, standardization was performed on the continuous variables to ensure these factors were scaled uniformly. Since the number of samples in each cluster was imbalanced, the SMOTETomek-based resampling method was used on the training dataset to improve the accuracy of the classifiers [209]. This method combined the SMOTE function's oversampling technique and the Tomek Link function's undersampling technique to address the imbalance issues [209]. The test datasets remained imbalanced to avoid bias and reflect the nature of the original dataset.

For a Random Forest (RF) classifier, hyperparameter tuning through a grid search with cross-validation of five folds was used to capture the optimal parameter. The fold with the maximized mean F1 score was chosen. After the hyperparameter tuning process, the random forest classifier with the optimized parameter was used on the training dataset and tested on the held-out testing dataset. To evaluate the classifier performance after hyperparameter tuning, accuracy, ROC-AUC score, F1 score, precision, and recall were reported. The SHAP plots with the TreeExplainer function were generated for the high-level and the sub-level clusters. The factors were ranked in descending order based on the SHAP mean absolute values according to

their importance in the model prediction of clusters. Further details on the SHAP plots are presented in Appendix B. SHAP was imported into the Python environment, version 0.42.1.

Statistical Analysis of Association

The top five important factors from each cluster based on the SHAP plots were selected for both multinomial and logistic regression analyses to assess the strength of association between the variable and being in each cluster group. The odds ratio and 95% CI were reported with the p-value significance level at $p < 0.05$. The ANOVA test was applied for the continuous variables, and the Chi-square test was applied for the categorical variables in relation to the clusters. This analysis aimed to understand which factor was associated with grouping students into certain clusters relative to others. This analysis was conducted in the SAS studio (Enterprise Edition).

4.4 Results

4.4.1 Descriptive Statistics

Tables 4.1 and 4.2 display the descriptive statistics of the 2018-19 and 2020-21 cohorts. The chosen variables were the representative questions mentioned in the predictor of interest section, including academic achievement, substance use, bullying, healthy eating, physical activity, school connectedness, screen time and weight. Demographic variables, including age, ethnicity, and gender, were also included. In the 2018-19 cohort, among the 74501 original samples, we had 32899 complete case records. In the 2020-21 cohort, there were a total of 28613 complete case records from the original 53469 samples.

Table 4.1 Descriptive Statistics of COMPASS Study Cohort 2018-19

Variable	Category	Total (N=74501)
		N (%) or Mean (SD)
Age		
	Mean (SD)	15.2 (1.5)
	Missing	541 (0.7)
Biological Sex		
	Female	36546 (49.1)
	Male	37126 (49.8)
	Missing	829 (1.1)
Province		
	AB	3301 (4.4)
	BC	10402 (14.0)

	ON	30675 (41.2)
	QC	30123 (40.4)
Ethnicity		
	White	56021 (75.2)
	Non-white	17818 (23.9)
	Missing	662 (0.9)
School-Related		
School Connectedness Score		
	Mean (SD)	18.5 (3.4)
	Missing	3088 (4.1)
Math Grade		
	80 – 100%	35099 (47.1)
	50 – 70%	33073 (44.4)
	Less than 50%	2757 (3.7)
	Missing	3572 (4.8)
Experienced Bullying in the Last 30 Days		
	Yes	13449 (18.1)
	No	61052 (81.9)
Substance Use		
Marijuana or Cannabis Use		
	Yes	19334 (26.0)
	No	53965 (72.4)
	Missing	1202 (1.6)
Smoking and Tobacco Use		
	Yes	15815 (21.2)
	No	58110 (78.0)
	Missing	576 (0.8)
Weight-Related		
BMI Status		
	Underweight	1397 (1.9)
	Healthy Weight	39388 (52.9)
	Overweight and Obese	12709 (17.1)
	Missing	21007 (28.2)
Weight Perception		
	Underweight	12140 (16.3)
	About the Right Weight	43893 (58.9)
	Overweight	17038 (22.9)
	Missing	1430 (1.9)
Weight Change Intention		
	Lose weight	22587 (30.3)
	Gain weight	11492 (15.4)
	Stayed the same weight	16593 (22.3)

	Not trying to do anything	22462 (30.1)
	Missing	1367 (1.8)
Screen Time		
Watching TV: Minutes per day		
	Mean (SD)	113.3 (94.2)
	Missing	602 (0.8)
Surfing Internet (Minutes per day)		
	Mean (SD)	116.6 (117.4)
	Missing	602 (0.8)
Doing Homework (Minutes per day)		
	Mean (SD)	90.0 (82.1)
	Missing	602 (0.8)
Physical Activity		
Participation in School Competitive Sport		
	Yes	26267 (35.3)
	No	45996 (61.7)
	None offered at my school	1075 (1.4)
	Missing	1163 (1.6)
Combined and Moderate Physical Activity (Minute per day)		
	Mean (SD)	108.2 (82.6)
	Missing	1777 (2.4)
Healthy Eating		
Skipped Breakfast for Losing Weight		
	Yes	3687 (4.9)
	No	70814 (95.1)
Snack Consumption (0 - 5 days)		
	Mean (SD)	0.4 (1.0)
	Missing	2240 (3.0)
Sugary Beverages Consumption (0 – 5 days)		
	Mean (SD)	1.4 (1.6)
	Missing	2337 (3.1)
Meat and Alternatives (0 – 5 or more servings)		
	Mean (SD)	2.1 (1.2)
	Missing	2833 (3.8)
Vegetables and Fruits (0 – 9 or more servings)		

	Mean (SD)	3.0 (2.0)
	Missing	2822 (3.8)
Mental Health		
Mental Well-being (Flourishing)	Mean (SD)	32.2 (5.7)
	Missing	4486 (6.0)
Anxiety (GAD7)	Mean (SD)	6.1 (5.6)
	Missing	6216 (8.3)

Table 4.2 Descriptive Statistics of COMPASS Study Cohort 2020-21

Variable	**Category**	**Total (N=53469)**
		N (%) or Mean (SD)
Age		
	Mean (SD)	15.0 (1.5)
	Missing	733 (1.4)
Gender		
	Female	26981 (50.5)
	Male	23658 (44.2)
	Gender-Minority	2046 (3.8)
	Missing	784 (1.5)
Province		
	AB	2282 (4.3)
	BC	5460 (10.2)
	ON	14040 (26.3)
	QC	31687 (59.3)
Ethnicity		
	White	42076 (78.7)
	Non-White	10511 (19.7)
	Unknown/Missing	882 (1.6)
School-Related		
School Connectedness Score		
	Mean (SD)	18.5 (3.4)
	Missing	6142 (11.5)
School Marks		
	80 – 100%	26431 (49.4)
	50 – 70%	20495 (38.3)
	Less than 50%	486 (0.9)
	Missing	6057 (11.3)
Experienced Bullying in the Last 30 Days		
	Yes	4554 (8.5)

	No	41365 (77.4)
	Missing	7550 (14.1)
COVID-19		
Life Changed Due to COVID-19: Boredom		
	Increased	32635 (61.0)
	Stayed the same	12112 (22.7)
	Decreased	2055 (3.8)
	Missing	6667 (12.5)
Life Changed Due to COVID-19: Stress		
	Increased	22761 (42.6)
	Stayed the same	21667 (40.5)
	Decreased	2462 (4.6)
	Missing	6579 (12.3)
Life Changed Due to COVID-19: Sleep		
	Increased	12564 (23.5)
	Stayed the same	19386 (36.3)
	Decreased	14774 (27.6)
	Missing	6745 (12.6)
Life Changed Due to COVID-19: Cannabis Use		
	Increased	2847 (5.3)
	Stayed the same	45393 (84.9)
	Decreased	5229 (9.8)
Physical Activity		
Participation in School Competitive Sport		
	Yes	12911 (24.1)
	No	33178 (62.1)
	None offered at my school	3164 (5.9)
	Missing	4216 (7.9)
Average combined physical activity per day		
	Mean (SD)	93.2 (80.1)
	Missing	6130 (11.5)
Screen Time		
Watching TV (Minutes per day)		
	Mean (SD)	119.3 (113.7)
	Missing	3080 (5.8)
Surfing Internet (Minutes per day)		
	Mean (SD)	120.9 (138.0)
	Missing	3080 (5.8)
Doing Homework (Minutes per day)		
	Mean (SD)	90.7 (91.3)

	Missing	3080 (5.8)
Healthy Eating		
Fast Food Purchased in the Past Week (Numbers of days)		
	Mean (SD)	1.1 (1.3)
	Missing	5257 (9.8)
Snacks Consumption in the Past Week (Numbers of days)		
	Mean (SD)	0.6 (1.3)
	Missing	5257 (9.8)
Vegetable Consumption in the Past Week (Numbers of days)		
	Mean (SD)	5.1 (2.5)
	Missing	5257 (9.8)
Fruit Consumption in the Past Week (Numbers of days)		
	Mean (SD)	4.8 (2.6)
	Missing	5257 (9.8)
Substance Use		
Smoking and Tobacco Use		
	Yes	8029 (15.0)
	No	41353 (77.3)
	Missing	4087 (7.6)
Marijuana or Cannabis Use		
	Yes	8752 (16.4)
	No	40154 (75.1)
	Missing	4563 (8.5)
Weight-Related		
BMI Status		
	Underweight	701 (1.3)
	Healthy Weight	21335 (39.9)
	Overweight and Obese	5725 (10.7)
	Missing	25708 (48.1)
Weight Perception		
	Underweight	8197 (15.3)
	About the Right Weight	31782 (59.4)
	Overweight	11587 (21.7)
	Missing	1903 (3.6)
Weight Change Intention		
	Lose weight	16748 (31.3)

	Gain weight	6645 (12.4)
	Stayed the same weight	11274 (21.1)
	Not trying to do anything	17116 (32.0)
	Missing	1686 (3.2)
Mental Health		
Mental Well-being (Flourishing)		
	Mean (SD)	31.6 (6.1)
	Missing	6181(11.6)
Anxiety (GAD7)	Mean (SD)	7.1 (5.9)
	Missing	7032 (13.2)

4.4.2 K-means Clustering and Model Performance

Table 4.3 presents the Silhouette coefficient and DB index for determining the numbers of k in both high-level and sub-level clusters. The sub-cluster k-means were not performed on the high-level Cluster 1 in 2018-19 and high-level Cluster 2 in 2020-21 cohorts since these two cohorts were considered high-risk in substance use. The purpose of sub-cluster k-means was to investigate other important variables that affect the rest of the clusters. Table 4.4 shows the model performance using clusters as labels in the RF model.

Table 4.3 Cohorts 2018-19 and 2020-21 K-means Clustering Result

Cohort (Pre-COVID)	**High-Level Cluster**	**Number of Students in High-Level Cluster N (%)**	**Silhouette Coefficient High-Level (High-Level)**	**Davies–Bouldin Index (High-Level)**	**Sub-Cluster**	**Number of Students in Sub-cluster N (%)**	**Silhouette Coefficient High-Level (Sub-Cluster)**	**Davies–Bouldin Index (Sub-Cluster)**
2018-19 (N=32899)	Cluster 0	24328 (73.95)	0.16	2.81	0	5464 (22.46)	0.0073	4.30
					1	3630 (14.92)		
					2	683 (2.81)		
					3	8,163 (33.55)		
					4	6,383 (26.24)		
	Cluster 1	155 (0.47)			N/A	N/A	N/A	N/A
	Cluster 2	8416 (25.58)			0	891 (10.59)	0.0247	4.42
					1	1375 (16.34)		
					2	2432 (28.90)		
					3	3718 (44.18)		
2020-21 (N=28613)	Cluster 0	3436 (12.01)	0.16	2.81	0	1136 (33.07)	0.0040	4.66
					1	834 (24.28)		
					2	474 (13.80)		
					3	860 (25.04)		
					4	132 (3.84)		

	Cluster 1	14692 (51.35)			0	5197 (35.37)	0.0676	4.18
					1	1228 (8.36)		
					2	8267 (56.27)		
	Cluster 2	56 (0.20)			N/A	N/A	N/A	N/A
	Cluster 3	10429 (36.45)			0	898 (8.61)	0.0436	4.44
					1	3479 (33.36)		
					2	5100 (48.90)		
					3	952 (9.13)		
[1] N/A represents the high-level clusters without the sub-cluster analysis								

Table 4.4 Model Performance for Random Forest Model

Cohort	**Cluster**	**Accuracy**	**Precision**	**Recall**	**F1 score**	**ROC-AUC**
2018-19 (N=32899)	High-Level (All)	0.97	0.98	0.85	0.90	0.92
	High-Level Cluster 0: Sub-Clusters	0.90	0.93	0.89	0.90	0.93
	High-Level Cluster 1: Sub-Clusters	N/A	N/A	N/A	N/A	N/A
	High-Level Cluster 2: Sub-Clusters	0.89	0.91	0.88	0.89	0.91
2020-21 (N=28613)	High-Level (All)	0.96	0.97	0.88	0.91	0.92
	High-Level Cluster 0: Sub-Clusters	0.84	0.88	0.78	0.81	0.87
	High-Level Cluster 1: Sub-Clusters	0.94	0.96	0.95	0.95	0.95
	High-Level Cluster 2: Sub-Clusters	N/A	N/A	N/A	N/A	N/A
	High-Level Cluster 3: Sub-Clusters	0.92	0.95	0.93	0.94	0.95
[1] N/A represents the high-level clusters without the sub-cluster analysis						

4.4.3 High-Level Clustering Analysis Interpretation in 2018-19 and 2020-21

The response categories were chosen from the top five important factors based on the SHAP plot for all the high-level clusters in the 2018-19 and 2020-21 cohorts.

In the pre-COVID cohort (2018-19), we discovered that the categories that most discriminative were substance use and mental health categories. Substance use categories included cannabis use, smoking and tobacco use, and opioid use. Based on Table 4.5, Cluster 1 students were high-risk with substance use (87.7 % of using marijuana or cannabis) compared to Cluster 0, who were primarily non-substance users (82.1% of not using marijuana or cannabis) and Cluster 2, who were between moderate and high substance users in a few related categories (77.2% of using marijuana or cannabis). In the mental health categories, Cluster 1 students had the highest responses in feeling depressed in the last seven days (34.2%) and did not feel comfortable talking to an adult at school (28.4%) compared to 5.9% and 15.6% in Cluster 0. Notably, only 0.5% of students were in Cluster 1, and most (73.9%) were in Cluster 0.

To understand how each factor influenced the likelihood of being in one cluster relative to the cluster with the most students, Logistic Regression was applied to compare Cluster 1 and Cluster 0. Table 4.5.1 shows that, relative to Cluster 0, using marijuana or cannabis in the last 12 months was associated with 20.08 times increased odds of being in Cluster 1. Additionally, students who felt depressed for 5-7 days in the last 7 days had 26.81 times higher odds of being in Cluster 1 relative to those who felt depressed for none or less than 1 day. Students who refused to discuss their mental health due to uncomfortableness also had an increased likelihood of 2.16 times being in Cluster 1 than the ones who did not choose these responses.

Table 4.6 represents the characteristics of high-level clusters in the post-COVID (2020-21) cohort. Similar to the pre-COVID cohort, substance use was also the most distinguished factor, separating Cluster 2 from the other clusters. For instance, Cluster 2 was the only one that 80.4% of students who tried Fentanyl to get high in the last 12 months. In the mental health category, students in Cluster 2 had the highest response in disagreement; 57.1% of them indicated they could not talk about their problems with their family relatives, while Cluster 1 only had 18.0% in this response. Due to the extremely small number of students in Cluster 2 (0.2%), there was not enough sample size to perform a logistic regression model for the post-COVID clusters since the maximum likelihood estimate would not be applicable.

Table 4.5 High Level Cluster Year 2018-19
Top Five Factors

N (%)	Cluster 0 N=24328 (73.9)	Cluster 1 N=155 (0.5)	Cluster 2 N=8416 (25.6)	Total N=32899 (100)	P-value
Cannabis Use					
Not used marijuana or cannabis in the last 12 months					<.0001[1]
No	4361 (17.9)	136 (87.7)	6500 (77.2)	10997 (33.4)	
Yes	19967 (82.1)	19 (12.3)	1916 (22.8)	21902 (66.6)	
Used marijuana or cannabis by eating or drinking it					<.0001[1]
No	24083 (99.0)	40 (25.8)	6165 (73.3)	30288 (92.1)	
Yes	245 (1.0)	115 (74.2)	2251 (26.7)	2611 (7.9)	
Used marijuana or cannabis by smoking					<.0001[1]
No	23212 (95.4)	30 (19.4)	2735 (32.5)	25977 (79.0)	
Yes	1116 (4.6)	125 (80.6)	5681 (67.5)	6922 (21.0)	
Difficult or easy for you to get marijuana					<.0001[1]
Difficult	2620 (10.8)	7 (4.5)	247 (2.9)	2874 (8.7)	
Easy	9733 (40.0)	142 (91.6)	7198 (85.5)	17073 (51.9)	
I do not know	11975 (49.2)	6 (3.9)	971 (11.5)	12952 (39.4)	
Smoking and Tobacco Use					
Not used any smoking and tobacco use in the last 30 days					<.0001[1]
No	2420 (9.9)	126 (81.3)	6161 (73.2)	8707 (26.5)	
Yes	21908 (90.1)	29 (18.7)	2255 (26.8)	24192 (73.5)	
Used blunt wraps in the last 30 days					<.0001[1]
No	24319 (100.0)	42 (27.1)	7995 (95.0)	32356 (98.3)	
Yes	9 (0.0)	113 (72.9)	421 (5.0)	543 (1.7)	
Used cigars in the last 30 days					<.0001[1]
No	24290 (99.8)	33 (21.3)	7802 (92.7)	32125 (97.6)	
Yes	38 (0.2)	122 (78.7)	614 (7.3)	774 (2.4)	
Mental Health					
I felt that everything I did was an effort in the last 7 days					<.0001[1]
None or less than 1 day	11299 (46.4)	72 (46.5)	2576 (30.6)	13947 (42.4)	
1 – 4 days	10788 (44.3)	33 (21.3)	4672 (55.5)	15493 (47.1)	
5 – 7 days	2241 (9.2)	50 (32.3)	1168 (13.9)	3459 (10.5)	
I felt depressed in the last 7 days					<.0001[1]
None or less than 1 day	15238 (62.6)	64 (41.3)	3053 (36.3)	18355 (55.8)	
1 – 4 days	7653 (31.5)	38 (24.5)	3691 (43.9)	11382 (34.6)	
5 – 7 days	1437 (5.9)	53 (34.2)	1672 (19.9)	3162 (9.6)	
Refused talk about mental health with an adult at school: Word would get out					<.0001[1]
No	20690 (85.0)	118 (76.1)	5741 (68.2)	26549 (80.7)	
Yes	3638 (15.0)	37 (23.9)	2675 (31.8)	6350 (19.3)	

Refused talk about mental health with an adult at school:					<.0001[1]
No one I feel comfortable talking to					
No	20522 (84.4)	111 (71.6)	6241 (74.2)	26874 (81.7)	
Yes	3806 (15.6)	44 (28.4)	2175 (25.8)	6025 (18.3)	

[1]Chi-Square p-value;

Table 4.5.1 Logistic Regression Analysis
High Level Cluster 2018-19

		OR (95% CI)	
Top Five Factors	Category	Cluster 0	Cluster 1
Not used marijuana or cannabis in the last 12 months			
	No	Ref	20.08 (11.50 – 35.21) ***
	Yes	Ref	Ref
Difficult or easy for you to get marijuana			
	Difficult	Ref	7.50 (2.01 – 27.58) ***
	Easy	Ref	19.33 (7.37 – 50.71) ***
	I do not know	Ref	Ref
Not used any smoking and tobacco use in the last 30 days			
	No	Ref	22.80 (13.78 – 37.72) ***
	Yes	Ref	Ref
I felt that everything I did was an effort in the last 7 days			
	None or less than 1 day	Ref	0.79 (0.42 – 1.49)
	1 – 4 days	Ref	0.27 (0.14 – 0.53) ***
	5 – 7 days	Ref	Ref
I felt depressed in the last 7 days			
	None or less than 1 day	Ref	Ref
	1 – 4 days	Ref	2.20 (1.28 – 3.76) ***
	5 – 7 days	Ref	26.81 (13.17 – 54.59) ***
Refused talk about mental health with an adult at school: No one I feel comfortable talking to			
	No	Ref	Ref
	Yes	Ref	2.16 (1.27 – 3.66) ***

P<0.0001 ***, P<0.01 **, P<0.05 *

Table 4.6. High Level Cluster Year 2020-21
Top Five Factors

N (%)	Cluster 0 N=3436 (12.0)	Cluster 1 N=14692 (51.3)	Cluster 2 N=56 (0.2)	Cluster 3 N=10429 (36.4)	Total N=28613 (100)	P-value
Academic Achievement						
Highest Education you Would Like to Get						<.0001[1]
Not University	1403 (40.8)	2916 (19.8)	32 (57.1)	2479 (23.8)	6830 (23.9)	
University and higher	1535 (44.7)	8098 (55.1)	15 (26.8)	6410 (61.5)	16058 (56.1)	
I don't know	498 (14.5)	3678 (25.0)	9 (16.1)	1540 (14.8)	5725 (20.0)	
Alcohol Use						
5 Drinks or More in the Last 12 months						<.0001[1]
Non drinker	619 (18.0)	14605 (99.4)	2 (3.6)	5680 (54.5)	20906 (73.1)	
More than once	2817 (82.0)	87 (0.6)	54 (96.4)	4749 (45.5)	7707 (26.9)	
Age of First Alcohol Drink						<.0001[1]
No drinker	251 (7.3)	13918 (94.7)	2 (3.6)	871 (8.4)	15042 (52.6)	
Less than 18	3185 (92.7)	766 (5.2)	54 (96.4)	9535 (91.4)	13540 (47.3)	
18 and older	0 (0.0)	8 (0.1)	0 (0.0)	23 (0.2)	31 (0.1)	
Cannabis Use						
Difficult or Easy to get marijuana						<.0001[1]
Difficult	147 (4.3)	2397 (16.3)	2 (3.6)	1188 (11.4)	3734 (13.1)	
Easy	2848 (82.9)	2352 (16.0)	49 (87.5)	5107 (49.0)	10356 (36.2)	
I do not know	441 (12.8)	9943 (67.7)	5 (8.9)	4134 (39.6)	14523 (50.8)	
Risk of cannabis use						<.0001[1]
No or low risk	2765 (80.5)	4884 (33.2)	40 (71.4)	6141 (58.9)	13830 (48.3)	
Medium or high risk	595 (17.3)	8518 (58.0)	13 (23.2)	4031 (38.7)	13157 (46.0)	
I don't know	76 (2.2)	1290 (8.8)	3 (5.4)	257 (2.5)	1626 (5.7)	
Smoking and Tobacco Use						
I have not Use Smoking or Tobacco in the last 30 days						<.0001[1]
No	923 (26.9)	16 (0.1)	35 (62.5)	66 (0.6)	1040 (3.6)	
Yes	2513 (73.1)	14676 (99.9)	21 (37.5)	10363 (99.4)	27573 (96.4)	
Opioid Use						
Tried Fentanyl to Get High in the last 12 Months						<.0001[1]
No	3355 (97.6)	14670 (99.9)	11 (19.6)	10414 (99.9)	28450 (99.4)	
Yes	81 (2.4)	22 (0.1)	45 (80.4)	15 (0.1)	163 (0.6)	
Tried Pain Relievers to Get High in the last 12 Months						<.0001[1]
No	2881 (83.8)	14261 (97.1)	13 (23.2)	10045 (96.3)	27200 (95.1)	
Yes	555 (16.2)	431 (2.9)	43 (76.8)	384 (3.7)	1413 (4.9)	
Tried Oxycodone to Get High in the Last 12 Months						<.0001[1]
No	3274 (95.3)	14670 (99.9)	12 (21.4)	10412 (99.8)	28368 (99.1)	
Yes	162 (4.7)	22 (0.1)	44 (78.6)	17 (0.2)	245 (0.9)	

Mental Health						
My Parent Expect Too Much of Me						<.0001[1]
Agree	1332 (38.8)	3519 (24.0)	31 (55.4)	2958 (28.4)	7840 (27.4)	
Neither agree nor disagree	1169 (34.0)	5399 (36.7)	9 (16.1)	3744 (35.9)	10321 (36.1)	
Disagree	935 (27.2)	5774 (39.3)	16 (28.6)	3727 (35.7)	10452 (36.5)	
I Can Talk About My Problems with My Family						<.0001[1]
Agree	1496 (43.5)	9257 (63.0)	19 (33.9)	5686 (54.5)	16458 (57.5)	
Neither agree nor disagree	809 (23.5)	2793 (19.0)	5 (8.9)	2266 (21.7)	5873 (20.5)	
Disagree	1131 (32.9)	2642 (18.0)	32 (57.1)	2477 (23.8)	6282 (22.0)	
I Can Talk About My Problems with My Friends						<.0001[1]
Agree	2525 (73.5)	10494 (71.4)	25 (44.6)	7907 (75.8)	20951 (73.2)	
Neither agree nor disagree	546 (15.9)	2720 (18.5)	10 (17.9)	1505 (14.4)	4781 (16.7)	
Disagree	365 (10.6)	1478 (10.1)	21 (37.5)	1017 (9.8)	2881 (10.1)	
I Have a Happy Home Life						<.0001[1]
Agree	2099 (61.1)	12399 (84.4)	23 (41.1)	8093 (77.6)	22614 (79.0)	
Neither agree nor disagree	819 (23.8)	1739 (11.8)	9 (16.1)	1622 (15.6)	4189 (14.6)	
Disagree	518 (15.1)	554 (3.8)	24 (42.9)	714 (6.8)	1810 (6.3)	
COVID						
Life Changed due to COVID-19: My Boredom						<.0001[1]
Increased	2824 (82.2)	9300 (63.3)	37 (66.1)	8133 (78.0)	20294 (70.9)	
Stayed the same	531 (15.5)	4617 (31.4)	9 (16.1)	2046 (19.6)	7203 (25.2)	
Decreased	81 (2.4)	775 (5.3)	10 (17.9)	250 (2.4)	1116 (3.9)	
Health Eating						
I don't Have for Breakfast						<.0001[1]
No	2558 (74.4)	13030 (88.7)	42 (75.0)	8667 (83.1)	24297 (84.9)	
Yes	878 (25.6)	1662 (11.3)	14 (25.0)	1762 (16.9)	4316 (15.1)	

[1]Chi-Square p-value

4.4.4 Sub-Level Clustering Analysis Interpretation in 2018-19 and 2020-21

To understand what other factors besides substance use, represent the student's profile before and after COVID-19, k-means clustering was applied to high-level Cluster 0 and 2 separately, excluding the substance-related variables in pre-COVID (2018-19). The evaluation methods defined five sub-clusters under high-level Cluster 0 (Table 4.5.2) and four sub-clusters under high-level Cluster 2 (Table 4.5.3).

In the post-COVID (2020-21), five sub-clusters were generated under high-level Cluster 0 (Table 4.6.1), three sub-clusters under high-level Cluster 1 (Table 4.6.2), and four sub-clusters under high-level Cluster 3 (Table 4.6.3). The factors chosen for all sub-clusters were also the top five important variables based on the SHAP plots.

Multinomial logistic regression was applied to identify the factors influencing the probability of being in a specific cluster. The sub-clusters with the highest numbers of students were chosen as the reference group. Relative to the reference group, the compared sub-clusters were either representative or had characteristics opposite to those of the other sub-clusters. For example, in high-level cluster 0 (2018-19), most students (33.6%) were in sub-cluster 3, so this sub-cluster was the reference group. One of the predictors was bullying others in cyber-attacks. Sub-Cluster 0 had none in bullying others, while Sub-Cluster 2 was the only one that had 9.4% participating in bullying others. Additionally, Sub-Cluster 2 had the highest proportion (14.0%) with no physically active friends. In comparison, Sub-Cluster 0 had the highest number of students with more than 5 physically active friends (59.2%) and the mean score of school connectedness (22.5). The rest of the sub-clusters were similar to Sub-Cluster 0. Therefore, Sub-Cluster 0 and 2 were selected to observe the relationship between the predictors and sub-clusters referencing Sub-Cluster 3.

This selection process was applied to all sub-cluster analyses in the 2018-19 and 2020-21 cohorts. For the predictors that involved derived variables, such as time spent surfing the internet in minutes or hours or sub-questions used for calculating the school connectedness score, the derived variables were included in the regression analysis. The interpretation of all sub-clusters focused on the significant factors from the multinomial logistic regression.

4.4.4.1 Pre-COVID (2018-19)

Sub-clusters from high-level Cluster 0:

The reference cluster was Sub-Cluster 3 (N=8168), and the comparison groups were Sub-Cluster 0 (N=5464) and Sub-Cluster 2 (N=683). Table 4.5.2.1 shows that the top five variables came from categories of academic achievement, screen time, bullying, healthy eating, physical activity, school connectedness, and weight. Students who had 1- 20 classes skipped in the last 4 weeks (OR=1.27, 95% CI = 1.04 - 1.56), often or usually with incomplete homework (OR=1.83, 95% CI = 1.42 - 2.38), and bullied by other students (OR = 10.08, 95% CI = 6.85 - 14.82) were more likely to be in Sub-Cluster 2 than students who never skipped class, completed their homework all the time, and did not experience bullying. In contrast, students with greater involvement in taking physical education classes at school (OR = 1.52, 95% CI = 1.26 - 1.84) and increased school connectedness score (OR = 2.44, 95% CI = 2.36 - 2.53) had a higher likelihood of being in Sub-Cluster 0. On the other hand, students who tried to lose weight (OR – 0.53, 95% CI = 0.45 - 0.63) and perceived their school were unsupportive in resisting tobacco (OR = 0.35, 95% CI = 0.30 - 0.40), access to healthy food/drinks (OR = 0.63, 95% CI = 0.51 -

0.78), and making sure no one was bullied (OR = 0.59, 95% CI = 0.47 - 0.74) were less likely to be in Sub-Cluster 0.

Sub-clusters from high-level Cluster 2:

As shown in Table 4.5.3.1, the reference cluster was Sub-Cluster 3 (N=3718), and the comparison groups were Sub-Cluster 1 (N=1375) and Sub-Cluster 2 (N=2432). Compared with the sub-clusters in high-level Cluster 0, we observed a similar effect in academic achievement and bullying categories. Students who skipped more than 20 classes with incomplete homework and were bullied by others had higher odds of being in Sub-Cluster 1 than students who did not participate in these activities.

Additional categories in this sub-cluster involved school-connected scores, weight perception and weight intention change. A one-unit increase in school connectedness score resulted in a 21% decrease in being in Sub-Cluster 1 (OR = 0.69, 95 % CI = 0.67 - 0.71) and a 50% increase in being in Sub-Cluster 2 (OR = 1.50, 95 % CI = 1.45 - 1.55). This indicated that Sub-Cluster 2 students had higher scores in school connectedness than the ones in Sub-Cluster 1. Among the weight-related variables, students who perceived themselves as underweight (OR = 1.86, 95% CI = 1.44 - 2.40) or overweight (OR = 1.61, 95% CI = 1.35 - 1.93) had a higher likelihood of being in Sub-Cluster 1 than those with about the right weight perception. Students who intended to lose weight were also more likely to be classified in Sub-Cluster 1 relative to not doing anything about their weights (OR = 1.26, 95% CI = 1.01 - 1.58).

Table 4.5.2 Cluster 0 Sub Cluster Year 2018-19
Top Five Factors

N (%)	Sub Cluster 0 N=5464 (22.5)	Sub Cluster 1 N=3630 (14.9)	Sub Cluster 2 N=683 (2.8)	Sub Cluster 3 N=8168 (33.6)	Sub Cluster 4 N=6383 (26.2)	Total N=24328 (100)	P-value
Screen: Homework							
Doing Homework: Minutes per day							<.0001[1]
Mean	95.4	111.2	87.5	116.8	73.2	98.9	
(SD)	(71.8)	(92.1)	(77.4)	(81.7)	(55.1)	(77.02)	
Academic Achievement							
N of class skipped in the last 4 weeks							<.0001[2]
zero class	4500 (82.4)	2636 (72.6)	506 (74.1)	6623 (81.1)	4217 (66.1)	18482 (76.0)	
1-20 classes	946 (17.3)	980 (27.0)	174 (25.5)	1537 (18.8)	2136 (33.5)	5773 (23.7)	
More than 20 classes	18 (0.3)	14 (0.4)	3 (0.4)	8 (0.1)	30 (0.5)	73 (0.3)	

Homework Incomplete							<.0001[2]
Never	2311 (42.3)	767 (21.1)	124 (18.2)	2197 (26.9)	1449 (22.7)	6848 (28.1)	
Seldom	2585 (47.3)	1791 (49.3)	360 (52.7)	4521 (55.4)	3766 (59.0)	13023 (53.5)	
Often of usually	568 (10.4)	1072 (29.5)	199 (29.1)	1450 (17.8)	1168 (18.3)	4457 (18.3)	
Overall mark in Math							<.0001[2]
80-100%	3535 (64.7)	1992 (54.9)	428 (62.7)	5207 (63.7)	3727 (58.4)	14889 (61.2)	
50-70%	1864 (34.1)	1535 (42.3)	245 (35.9)	2842 (34.8)	2589 (40.6)	9075 (37.3)	
Less than 50%	65 (1.2)	103 (2.8)	10 (1.5)	119 (1.5)	67 (1.0)	364 (1.5)	
Bullying							
Bullied others in Cyber Attacks							<.0001[2]
No	5463 (100.0)	3629 (100.0)	619 (90.6)	8168 (100.0)	6383 (100.0)	24262 (99.7)	
Yes	1 (0.0)	1 (0.0)	64 (9.4)	0 (0.0)	0 (0.0)	66 (0.3)	
Bullied others in Verbal Attacks							<.0001[2]
No	5463 (100.0)	3627 (99.9)	124 (18.2)	8168 (100.0)	6383 (100.0)	23765 (97.7)	
Yes	1 (0.0)	3 (0.1)	559 (81.8)	0 (0.0)	0 (0.0)	563 (2.3)	
Bullied By others in Cyber Attacks							<.0001[2]
No	5438 (99.5)	3429 (94.5)	628 (91.9)	8079 (98.9)	6360 (99.6)	23934 (98.4)	
Yes	26 (0.5)	201 (5.5)	55 (8.1)	89 (1.1)	23 (0.4)	394 (1.6)	
Healthy Eating							
Eat lunch at school in cafeteria							<.0001[2]
None	3366 (61.6)	2229 (61.4)	421 (61.6)	4941 (60.5)	4177 (65.4)	15134 (62.2)	
1-5 days	2098 (38.4)	1401 (38.6)	262 (38.4)	3227 (39.5)	2206 (34.6)	9194 (37.8)	
Physical Activity							
N of physically active closest friend							<.0001[2]
No friends	179 (3.3)	510 (14.0)	47 (6.9)	517 (6.3)	316 (5.0)	1569 (6.4)	
Less than 5 friends	2051 (37.5)	2341 (64.5)	353 (51.7)	4831 (59.1)	3429 (53.7)	13005 (53.5)	
More than 5 friends	3234 (59.2)	779 (21.5)	283 (41.4)	2820 (34.5)	2638 (41.3)	9754 (40.1)	
Taking Physical Class at School							<.0001[2]

Yes, taking this term	3781 (69.2)	1655 (45.6)	377 (55.2)	2882 (35.3)	5774 (90.5)	14469 (59.5)	
Yes, taking this year	765 (14.0)	522 (14.4)	107 (15.7)	1523 (18.6)	329 (5.2)	3246 (13.3)	
No	918 (16.8)	1453 (40.0)	199 (29.1)	3763 (46.1)	280 (4.4)	6613 (27.2)	
School Connectedness							
School Connectedness Score							<.0001[1]
Mean (SD)	22.5 (1.7)	16.2 (2.9)	18.6 (3.0)	18.7 (1.9)	18.9 (1.9)	19.2 (2.9)	
School Support: Resist/Quit Tobacco							<.0001[2]
Very supportive	4540 (83.1)	1333 (36.7)	325 (47.6)	4466 (54.7)	3988 (62.5)	14652 (60.2)	
Unsupportive	924 (16.9)	2297 (63.3)	358 (52.4)	3702 (45.3)	2395 (37.5)	9676 (39.8)	
School Support: Access to Healthy Food/Drinks							<.0001[2]
Very supportive	5084 (93.0)	2455 (67.6)	503 (73.6)	6533 (80.0)	5311 (83.2)	19886 (81.7)	
Unsupportive	380 (7.0)	1175 (32.4)	180 (26.4)	1635 (20.0)	1072 (16.8)	4442 (18.3)	
School Support: No one is bullied							<.0001[2]
Very supportive	5195 (95.1)	1921 (52.9)	411 (60.2)	6254 (76.6)	5458 (85.5)	19239 (79.1)	
Unsupportive	269 (4.9)	1709 (47.1)	272 (39.8)	1914 (23.4)	925 (14.5)	5089 (20.9)	
School Support: Physically Active Opportunities							<.0001[2]
Very supportive	5358 (98.1)	3105 (85.5)	631 (92.4)	7793 (95.4)	5884 (92.2)	22771 (93.6)	
Unsupportive	106 (1.9)	525 (14.5)	52 (7.6)	375 (4.6)	499 (7.8)	1557 (6.4)	
Screen Time							
Online Gambling in the last 30 days							0.0019[2]
Yes	66 (1.2)	40 (1.1)	16 (2.3)	89 (1.1)	49 (0.8)	260 (1.1)	
No	5398 (98.8)	3590 (98.9)	667 (97.7)	8079 (98.9)	6334 (99.2)	24068 (98.9)	
Weight							
Weight Change Intention							<.0001[2]

Lose weight	975 (17.8)	1702 (46.9)	210 (30.7)	2989 (36.6)	917 (14.4)	6793 (27.9)
Gain weight	816 (14.9)	339 (9.3)	144 (21.1)	1321 (16.2)	510 (8.0)	3130 (12.9)
Stay the same	1589 (29.1)	595 (16.4)	143 (20.9)	1685 (20.6)	2199 (34.5)	6211 (25.5)
Not doing anything about weight	2084 (38.1)	994 (27.4)	186 (27.2)	2173 (26.6)	2757 (43.2)	8194 (33.7)

[1]ANOVA F-test p-value; [2]Chi-Square p-value

Table 4.5.2.1 Multinomial Logistic Regression Analysis
Sub Cluster from Cluster 0 (2018-19)

		OR (95% CI)		
Top 5 Factors	**Category**	**Sub Cluster 0**	**Sub Cluster 2**	**Sub Cluster 3**
N of class skipped in the last 4 weeks				
	Zero class	Ref	Ref	Ref
	1-20 classes	0.71 (0.59 – 0.86) *	1.27 (1.04 – 1.56) *	Ref
	More than 20 classes	1.82 (0.20 – 16.97)	3.20 (0.73 – 14.08)	Ref
Homework Incomplete				
	Never	Ref	Ref	Ref
	Seldom	0.48 (0.42 – 0.56) ***	1.25 (1.00 – 1.56)	Ref
	Often of usually	0.45 (0.36 – 0.55) ***	1.83 (1.42 – 2.38) ***	Ref
				Ref
Overall mark in Math				
	80-100%	Ref	Ref	Ref
	50-70%	1.28 (1.12 – 1.48) **	0.86 (0.72 – 1.04)	Ref
	Less than 50%	2.12 (1.20 – 3.75) *	0.78 (0.38 – 1.59)	Ref
Bullied By others in Cyber Attacks				
	No	Ref	Ref	Ref
	Yes	0.52 (0.27 – 1.01)	10.08 (6.85 – 14.82) ***	Ref
				Ref
N of Physically Active Closest Friend				
	No friends	0.75 (0.55 – 1.03)	1.34 (0.94 – 1.91)	Ref

	Less than 5 friends	0.63 (0.55 – 0.72) ***	0.98 (0.82 – 1.17)	Ref
	More than 5 friends	Ref	Ref	Ref
Taking Physical Class at School				
	Yes, taking this term	1.78 (1.50 – 2.10)	1.26 (1.02 – 1.57)	Ref
	Yes, taking this year	1.52 (1.26 – 1.84) ***	0.96 (0.73 – 1.25)	Ref
	No	Ref	Ref	Ref
School Connected Score		2.44 (2.36 – 2.53) ***	1.03 (0.99 – 1.97)	Ref
School Support: Resist/Quit Tobacco				
	Very supportive	Ref	Ref	Ref
	Unsupportive	0.35 (0.30 – 0.40) ***	1.04 (0.86 – 1.26)	Ref
School Support: Access to Healthy Food/Drinks				
	Very supportive	Ref	Ref	Ref
	Unsupportive	0.63 (0.51 – 0.78) ***	1.13 (0.92 – 1.39)	Ref
School Support: No one is Bullied				
	Very supportive	Ref	Ref	Ref
	Unsupportive	0.59 (0.47 – 0.74) ***	2.02 (1.72 – 2.57) ***	Ref
Weight Change Intention				
	Lose weight	0.53 (0.45 – 0.63) ***	1.19 (0.95 – 1.49)	Ref
	Gain weight	0.85 (0.69 – 1.03)	1.06 (0.82 – 1.36)	Ref
	Stay the same	0.90 (0.75 – 1.07)	1.10 (0.86 – 1.41)	Ref
	Not doing anything about weight	Ref	Ref	Ref

P<0.0001 ***, P<0.01 **, P<0.05 *

Table 4.5.3 Cluster 2 Sub Cluster Year 2018-19
Top Five Factors

	N (%)					
	Cluster 0 N=891 (10.6)	Cluster 1 N=1375 (16.3)	Cluster 2 N=2432 (28.9)	Cluster 3 N=3718 (44.2)	Total N=8416 (100)	P-value
Screen: Homework						
Doing Homework: Minutes per day						<.0001[1]
Mean (SD)	69.5 (72.7)	83.6 (85.5)	66.4 (66.8)	82.6 (70.7)	76.7 (72.9)	
Academic Achievement						
N of class skipped in the last 4 weeks						<.0001[2]
zero class	365 (41.0)	586 (42.6)	1209 (49.7)	1487 (40.0)	3647 (43.3)	
1-20 classes	507 (56.9)	743 (54.0)	1205 (49.5)	2195 (59.0)	4650 (55.3)	
More than 20 classes	19 (2.1)	46 (3.3)	18 (0.7)	36 (1.0)	119 (1.4)	
Homework Incomplete						<.0001[2]
Never	70 (7.9)	122 (8.9)	331 (13.6)	332 (8.9)	855 (10.2)	
Seldom	346 (38.8)	507 (36.9)	1123 (46.2)	1760 (47.3)	3736 (44.4)	
Often of usually	475 (53.3)	746 (54.3)	978 (40.2)	1626 (43.7)	3825 (45.4)	
Overall mark in Math						<.0001[2]
80-100%	366 (41.1)	466 (33.9)	983 (40.4)	1381 (37.1)	3196 (38.0)	
50-70%	476 (53.4)	779 (56.7)	1367 (56.2)	2184 (58.7)	4806 (57.1)	
Less than 50%	49 (5.5)	130 (9.5)	82 (3.4)	153 (4.1)	414 (4.9)	
Bullying						
Bullied By others in Cyber Attacks						<.0001[2]
No	711 (79.8)	1084 (78.8)	2392 (98.4)	3542 (95.3)	7729 (91.8)	
Yes	180 (20.2)	291 (21.2)	40 (1.6)	176 (4.7)	687 (8.2)	
Physical Activity						
Were the last 7 Days a Typical Physical Activity Week						<.0001[2]
Yes	563 (63.2)	807 (58.7)	1782 (73.3)	2395 (64.4)	5547 (65.9)	
No, more active	118 (13.2)	197 (14.3)	192 (7.9)	403 (10.8)	910 (10.8)	

No, less active	210 (23.6)	371 (27.0)	458 (18.8)	920 (24.7)	1959 (23.3)	
Participate in Competitive School Sports						<.0001[2]
Yes	349 (39.2)	306 (22.3)	1154 (47.5)	1133 (30.5)	2942 (35.0)	
No	533 (59.8)	1045 (76.0)	1255 (51.6)	2554 (68.7)	5387 (64.0)	
None offered at my school	9 (1.0)	24 (1.7)	23 (0.9)	31 (0.8)	87 (1.0)	
School Connectedness						
School Connectedness Score						<.0001[1]
Mean (SD)	16.5 (3.14)	13.7 (3.44)	19.4 (2.99)	16.9 (2.16)	17.0 (3.32)	
I Feel Close to People at My School						<.0001[2]
Agree	615 (69.0)	473 (34.4)	2142 (88.1)	2534 (68.2)	5764 (68.5)	
Disagree	276 (31.0)	902 (65.6)	290 (11.9)	1184 (31.8)	2652 (31.5)	
Screen Time						
Online Gambling in the last 30 days						<.0001[2]
Yes	68 (7.6)	45 (3.3)	118 (4.9)	58 (1.6)	289 (3.4)	
No	823 (92.4)	1330 (96.7)	2314 (95.1)	3660 (98.4)	8127 (96.6)	
Weight						
Weight Change Intention						<.0001[2]
Lose weight	343 (38.5)	824 (59.9)	530 (21.8)	1755 (47.2)	3452 (41.0)	
Gain weight	188 (21.1)	155 (11.3)	594 (24.4)	493 (13.3)	1430 (17.0)	
Stay the same	144 (16.2)	134 (9.7)	576 (23.7)	722 (19.4)	1576 (18.7)	
Not doing anything about weight	216 (24.2)	262 (19.1)	732 (30.1)	748 (20.1)	1958 (23.3)	
Weight Perception						<.0001[2]
Underweight	184 (20.7)	221 (16.1)	367 (15.1)	514 (13.8)	1286 (15.3)	
About the right weight	438 (49.2)	521 (37.9)	1688 (69.4)	2019 (54.3)	4666 (55.4)	

Overweigh t	269 (30.2)	633 (46.0)	377 (15.5)	1185 (31.9)	2464 (29.3)

[1]ANOVA F-test p-value; [2]Chi-Square p-value;

Table 4.5.3.1 Multinomial Logistic Regression Analysis
Sub Cluster from Cluster 2 (2018-19)

		OR (95% CI)		
Top 5 Factors	**Category**	**Sub Cluster 1**	**Sub Cluster 2**	**Sub Cluster 3**
N of class skipped in the last 4 weeks				
	Zero class	Ref	Ref	Ref
	1-20 classes	0.84 (0.72 – 0.98) *	0.77 (0.68 – 0.88) **	Ref
	More than 20 classes	2.14 (1.21 – 3.79) **	0.76 (0.38 – 1.51)	Ref
Homework Incomplete				
	Never	Ref	Ref	Ref
	Seldom	0.88 (0.67 – 1.67)	0.59 (0.48 -0.73) ***	Ref
	Often of usually	1.29 (0.98 – 1.71)	0.47 (0.38 – 0.59) ***	Ref
				Ref
Overall mark in Math				
	80-100%	Ref	Ref	Ref
	50-70%	0.81 (0.69 – 0.95) **	1.07 (0.93 – 1.23)	Ref
	Less than 50%	1.84 (1.33 – 2.55) **	0.78 (0.55 – 1.11)	Ref
Bullied By others in Cyber Attacks				
	No	Ref	Ref	Ref
	Yes	4.67 (3.67 – 5.94) ***	0.53 (0.35 – 0.80) **	Ref
				Ref
Were the last 7 Days a Typical Physical Activity Week				
	Yes	Ref	Ref	Ref
	No, more active	1.18 (0.93 – 1.49)	0.82 (0.65 – 1.03)	Ref
	No, less active	1.15 (0.97 – 1.38)	0.69 (0.59 – 0.81) ***	Ref
				Ref
Participate in Competitive School Sports				
	Yes	0.55 (0.28 – 1.08)	1.37 (0.66 – 2.84)	Ref
	No	0.72 (0.37 – 1.41)	0.83 (0.40 – 1.72)	Ref
	None offered at my school	Ref	Ref	Ref
School Connected Score		0.69 (0.67 – 0.71) ***	1.5 (1.45 – 1.55) ***	Ref
I Feel Close to People at My School				
	Agree	Ref	Ref	Ref
	Disagree	1.38 (1.16 – 1.64) **	0.98 (0.82 – 1.18)	Ref
Weight Perception				
	Underweight	1.86 (1.44 – 2.40) ***	0.55 (0.45 – 0.67) ***	Ref
	About the right weight	Ref	Ref	Ref
	Overweight	1.61 (1.35 – 1.93) ***	0.58 (0.48 – 0.70) ***	
Weight Change Intention				
	Lose weight	1.26 (1.01 – 1.58) *	0.47 (0.38 – 0.56) ***	Ref

Gain weight	0.81 (0.60 – 1.10)	0.99 (0.81 – 1.23)	Ref
Stay the same	0.66 (0.50 – 0.86) **	0.80 (0.66 – 0.96) *	Ref
Not doing anything about weight	Ref	Ref	Ref

P<0.0001 ***, P<0.01 **, P<0.05 *

4.4.4.2 Post-COVID (2020-21)

Sub-clusters from high-level Cluster 0:

In Table 4.6.1.1, among the five sub-clusters, the reference cluster was Sub-Cluster 0 (N=1136), and the comparison groups were Sub-Cluster 1 (N=834) and Sub-Cluster 2 (N=474). Income, healthy eating and sleep appeared among the top five variables. Students with zero or unknown money spent on themselves each week (OR = 2.26, 95% CI = 1.63 - 3.15) and who skipped breakfast (OR = 1.35, 95% CI = 1.01 - 1.80) were more likely to be in Sub-Cluster 2 than those with more than $100 per week. The association between sleep and being in Sub-Cluster 2 was much stronger than in Sub-Cluster 1. Students who did not have an average of 8 hours of sleep had increased odds of being in either Sub-Cluster 2 (OR = 2.63, 95% CI = 2.02 - 3.42) compared to those with 8 hours of sleep Additionally, students with increased cannabis use due to COVID-19 (OR = 0.41, 95% CI = 0.32 – 0.53) and those who skipped breakfast (OR = 0.56, 95% CI = 0.43 – 0.73) were less likely to be in Sub-Cluster 1 compared to those with cannabis level staying the same and not skipping breakfast.

Sub-cluster from high-level Cluster 1:

In Table 4.6.2.1, the reference cluster was Sub-Cluster 2 (N=8267), and the comparison groups were Sub-Cluster 0 (N=5197) and Sub-Cluster 1 (N=1228). The top five categories included weight, physical activity, healthy eating, mental health and screen time. The association was focused on the life changes due to COVID rather than the opinions on COVID-19, such as ‘Mask can reduce the droplet transmission of COVID-19’. A higher likelihood of being in Sub-Cluster 0 and Sub-Cluster 1 was observed with each additional day of fast food purchased in the past week. This likelihood was 3% higher of being in Sub-Cluster 1 (OR = 1.17, 95% CI = 1.11 - 1.24) than in Sub-Cluster 0 (OR = 1.14, 95% CI = 1.09 - 1.19). In the BMI category, the likelihood of being in Sub-Cluster 0 for overweight and obese BMI students was higher (OR = 1.87, 95% CI = 1.57 - 2.21), while students with unknown BMI students were more likely to be in Sub-Cluster 1 (OR = 1.90, 95% = 1.64 - 2.19). In addition, one unit increased in the flourishing score had decreased odds of being in Sub-Cluster 0 (OR = 0.86, 95% CI = 0.86 - 0.87) or Sub-Cluster 1 (OR = 0.93, 95% CI = 0.92 - 0.95). Similarly, with one unit increase in

anxiety score, students were more likely to be in Sub-Cluster 0 (OR = 1.15, 95% CI = 1.14 - 1.16) or Sub-Cluster 1 (OR = 1.09, 95% CI = 1.07 - 1.10). The participation in physical activity was significantly associated with Sub-Cluster 0 only. Students who did not meet the 60-minute physical activity guidelines were more likely to be in Sub-Cluster 0. This result indicated that Sub-Cluster 0 or 1 students relative to Sub-Cluster 2 had higher BMI status and lower mental health well-being.

Sub-cluster from high-level Cluster 3:

In Table 4.6.3.1, the reference cluster was Sub-Cluster 2 (N=5100), and the comparison groups were Sub-Cluster 1 (N=3479) and Sub-Cluster 3 (N=952). The top five categories were the same as those in the high-level 1 sub-clusters. For students who did not meet the 60-minute physical activity per day (OR = 1.27, 95% CI = 1.08 - 1.50) and increased cannabis use due to COVID-19 (OR = 2.79, 95% CI = 1.64 - 4.75), the odds of being in Sub-Cluster 1 were higher. Mixed effects have been found in boredom due to COVID-19. Students with either increased boredom (OR = 2.38, 95% CI = 2.02 – 2.80) or decreased boredom (OR = 2.88, 95% CI = 1.89 - 4.41) had a higher likelihood of being in Sub-Cluster 1, but the likelihood of being in Sub-Cluster 1 was stronger for those with decreased boredom compared to increased boredom. Among the mental health categories, for each unit increase in flouring score, students were less likely to be in Sub-Cluster 1 (OR = 0.83, 95% CI = 0.82 - 0.84) or Sub-Cluster 3 (OR = 0.89, 95% CI = 0.86 - 0.92). For one unit increased in anxiety score, the odds of being in Sub-Cluster 1 (OR = 1.16, 95% CI = 1.14 - 1.17) were 9% higher than those of being in Sub-Cluster 3 (OR = 1.07, 95% CI = 1.04 - 1.11). These results indicated that both clusters contained students with higher anxiety scores. At the same time, lower levels of physical activity and increased cannabis use due to COVID-19 were more significantly associated with membership in Sub-Cluster 1.

Table 4.6.1 Cluster 0 Sub Cluster Year 2020-21
Top Five Factors

	N (%)						
Variables	Sub Cluster 0 N=1136 (33.1)	Sub Cluster 1 N=834 (24.3)	Sub Cluster 2 N=474 (13.8)	Sub Cluster 3 N=860 (25.0)	Sub Cluster 4 N=132 (3.8)	Total N=3436 (100)	P-value
Screen Time							
Surfing Internet: Minutes per day							<.0001[1]
Mean (SD)	144.4 (114.7)	117.5 (125.8)	191.1 (177.4)	191.5 (172.9)	160.2 (145.9)	156.7 (147.3)	
Watching TV: Minutes per day							<.0001[1]
Mean (SD)	126.0 (102.3)	110.6 (91.1)	149.3 (131.6)	157.8 (127.4)	126.4 (130.7)	133.5 (113.5)	
Doing Homework: Minutes per day							<.0001[1]
Mean (SD)	67.5 (70.1)	54.3 (66.7)	70.0 (75.6)	77.1 (86.8)	84.9 (110.4)	67.7 (76.9)	
Video Games: Minutes per day							<.0001[1]
Mean (SD)	59.2 (101.5)	119.7 (141.5)	83.7 (125.8)	107.8 (152.8)	152.0 (201.7)	93.0 (136.3)	
COVID							
Mask can reduce the droplet transmission of COVID-19							<.0001[2]
No	159 (14.0)	214 (25.7)	98 (20.7)	156 (18.1)	39 (29.5)	666 (19.4)	
Yes	977 (86.0)	620 (74.3)	376 (79.3)	704 (81.9)	93 (70.5)	2770 (80.6)	
COVID-19 is only found in individuals who show symptoms and signs of disease							<.0001[2]
No	1073 (94.5)	729 (87.4)	423 (89.2)	759 (88.3)	119 (90.2)	3103 (90.3)	
Yes	63 (5.5)	105 (12.6)	51 (10.8)	101 (11.7)	13 (9.8)	333 (9.7)	
Hand sanitizer helps prevent the transmission of COVID-19							<.0001[2]
No	231 (20.3)	241 (28.9)	95 (20.0)	138 (16.0)	41 (31.1)	746 (21.7)	
Yes	905 (79.7)	593 (71.1)	379 (80.0)	722 (84.0)	91 (68.9)	2690 (78.3)	
Life Changed due to COVID-19: My Cannabis Use N (%)							<.0001[2]
Increased	377 (33.2)	167 (20.0)	199 (42.0)	411 (47.8)	42 (31.8)	1196 (34.8)	
Stayed the same	562 (49.5)	470 (56.4)	200 (42.2)	314 (36.5)	48 (36.4)	1594 (46.4)	

Decreased	197 (17.3)	197 (23.6)	75 (15.8)	135 (15.7)	42 (31.8)	646 (18.8)	
Income							
Money Spend on Yourself or to Save Weekly							<.0001[2]
Zero or not known	188 (16.5)	192 (23.0)	181 (38.2)	377 (43.8)	36 (27.3)	974 (28.3)	
$1 - $100	369 (32.5)	286 (34.3)	159 (33.5)	321 (37.3)	43 (32.6)	1178 (34.3)	
More than $100	579 (51.0)	356 (42.7)	134 (28.3)	162 (18.8)	53 (40.2)	1284 (37.4)	
Healthy Eating							
Total Number of Days of Fast Food Purchased in the Past Week							0.1378[1]
Mean (SD)	1.7 (1.6)	1.8 (1.7)	1.9 (1.8)	1.7 (1.7)	1.6 (1.7)	1.7 (1.7)	
No Time for Breakfast							<.0001[2]
No	855 (75.3)	697 (83.6)	319 (67.3)	584 (67.9)	103 (78.0)	2558 (74.4)	
Yes	281 (24.7)	137 (16.4)	155 (32.7)	276 (32.1)	29 (22.0)	878 (25.6)	
Sleep							
At Least 8 hours of Sleep per Day (On Average)							<.0001[2]
Yes	800 (70.4)	481 (57.7)	204 (43.0)	348 (40.5)	78 (59.1)	1911 (55.6)	
No	336 (29.6)	353 (42.3)	270 (57.0)	512 (59.5)	54 (40.9)	1525 (44.4)	
Light Off and Gone to Sleep							<.0001[2]
AM	896 (78.9)	675 (80.9)	393 (82.9)	780 (90.7)	108 (81.8)	2852 (83.0)	
PM	240 (21.1)	159 (19.1)	81 (17.1)	80 (9.3)	24 (18.2)	584 (17.0)	

[1]ANOVA F-test p-value; [2]Chi-Square p-value

Table 4.6.1.1 Multinomial Logistic Regression Analysis
Sub Cluster from Cluster 0 (2020-21)

		OR (95% CI)		
Top 5 Factors	**Category**	**Sub Cluster 0**	**Sub Cluster 1**	**Sub Cluster 2**
Life Changed due to COVID-19: My Cannabis Use				
	Increased	Ref	0.41 (0.32 – 0.53) ***	0.94 (0.71 – 1.26)
	Stayed the same	Ref	Ref	Ref
	Decreased	Ref	1.03 (0.78 – 1.34)	0.91 (0.54 – 1.31)
Money Spend on Yourself or to Save Weekly				
	Zero or not known	Ref	1.25 (0.93 – 1.68)	2.26 (1.63 – 3.15) ***
	$1 - $100	Ref	Ref	Ref
	More than $100	Ref	0.68 (0.54 – 0.87) **	0.59 (0.43 – 0.80) **
No Time for Breakfast				
	No	Ref	Ref	Ref
	Yes	Ref	0.56 (0.43 – 0.73) ***	1.35 (1.01 – 1.80) *
At least 8 hours of sleep per day (on average)				
	Yes	Ref	Ref	Ref
	No	Ref	1.30 (1.04 – 1.62) **	2.63 (2.02 – 3.42) ***
Surfing Internet: Minutes per day		Ref	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00) ***
Watching TV: Minutes per day		Ref	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00)

P<0.0001 ***, P<0.01 **, P<0.05 *

Table 4.6.2 Cluster 1 Sub Cluster Year 2020-21
Top Five Factors

N (%)					
Variables	Sub Cluster 0 N=5197 (35.4)	Sub Cluster 1 N=1228 (8.4)	Sub Cluster 2 N=8267 (56.2)	Total N=14692 (100)	P-value
Healthy Eating					
Total Number of Days of Fast Food Purchased in the Past Week					<.0001[1]
Mean (SD)	1.0 (1.2)	1.0 (1.2)	0.7 (0.9)	0.8 (1.1)	
Mental Health					
Mental Wellbeing: Flourishing Score					<.0001[1]
Mean (SD)	29.3 (5.7)	32.4 (5.7)	34.8 (4.5)	32.6 (5.7)	
Anxiety: GAD7					<.0001[1]
Mean (SD)	9.2 (5.9)	6.2 (5.6)	3.9 (4.0)	6.0 (5.5)	
Physical Activity					
Average Hard and Moderate Physical Activity per day					<.0001[1]
Mean (SD)	74.5 (67.6)	100.7 (84.8)	96.3 (74.6)	89.0 (73.9)	
Meet the Guidelines for At Least 60 Minutes of Physical Activity per Day					<.0001[2]
Yes	2575 (49.5)	786 (64.0)	5319 (64.3)	8680 (59.1)	
No	2622 (50.5)	442 (36.0)	2948 (35.7)	6012 (40.9)	
Screen Time					
Surfing Internet: Minutes per day					<.0001[1]
Mean (SD)	136.2 (138.6)	115.2 (135.9)	75.0 (93.3)	100.0 (118.6)	
Weight					
BMI Category					<.0001[2]
Underweight	88 (1.7)	14 (1.1)	146 (1.8)	248 (1.7)	
About the right weight	1899 (36.5)	359 (29.2)	3436 (41.6)	5694 (38.8)	
Overweight and obese	523 (10.1)	112 (9.1)	652 (7.9)	1287 (8.8)	
Unknown	2687 (51.7)	743 (60.5)	4033 (48.8)	7463 (50.8)	

[1]ANOVA F-test p-value; [2]Chi-Square p-value

Table 4.6.2.1 Multinomial Logistic Regression Analysis
Sub Cluster from Cluster 1 (2020-21)

		OR (95% CI)		
Top 5 Factors	**Category**	**Sub Cluster 0**	**Sub Cluster 1**	**Sub Cluster 2**
BMI Category				
	Underweight	1.13 (0.77 – 1.64)	0.98 (0.55 – 1.75)	Ref
	About the right weight	Ref	Ref	Ref
	Overweight and obese	1.87 (1.57 – 2.21) ***	1.71 (1.35 – 2.17) ***	Ref
	Unknown	1.25 (1.13 – 1.39) ***	1.90 (1.64 – 2.19) ***	
Meet the Guidelines for At least 60 Minutes of Physical Activity per Day				
	Yes	Ref	Ref	Ref
	No	1.38 (1.22 -1.57) ***	1.12 (0.94 – 1.33)	Ref
Total number of days of Fast Food Purchased in the past week		1.14 (1.09 – 1.19) ***	1.17 (1.11 – 1.24) ***	Ref
Mental Wellbeing: Flourishing Score		0.86 (0.86 – 0.87) ***	0.93 (0.92 – 0.95) ***	Ref
Anxiety: GAD7		1.15 (1.14 – 1.16) ***	1.09 (1.07 – 1.10) ***	Ref
Average combined Hard and Moderate physical activity per day		1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00) **	Ref
				Ref
Surfing Internet: Minutes per day		1.00 (1.00 – 1.00) ***	1.00 (1.00 – 1.00) ***	Ref

P<0.0001 ***, P<0.01 **, P<0.05 *

Table 4.6.3 Cluster 3 Sub Cluster Year 2020-21
Top Five Factors

	N (%)					
Variables	Sub Cluster 0 N=898 (8.6)	Sub Cluster 1 N=3479 (33.4)	Sub Cluster 2 N=5100 (48.9)	Sub Cluster 3 N=952 (9.1)	Total N=10429 (100)	P-value
COVID						
Mask can reduce the droplet transmission of COVID-19						<.0001[1]
No	157 (17.5)	325 (9.3)	579 (11.4)	200 (21.0)	1261 (12.1)	
Yes	741 (82.5)	3154 (90.7)	4521 (88.6)	752 (79.0)	9168 (87.9)	
COVID-19 is only found in individuals who show symptoms and signs of disease						<.0001[1]
No	825 (91.9)	3286 (94.5)	4813 (94.4)	836 (87.8)	9760 (93.6)	
Yes	73 (8.1)	193 (5.5)	287 (5.6)	116 (12.2)	669 (6.4)	
Hand sanitizer helps prevent the transmission of COVID-19						<.0001[1]
No	184 (20.5)	484 (13.9)	897 (17.6)	236 (24.8)	1801 (17.3)	
Yes	714 (79.5)	2995 (86.1)	4203 (82.4)	716 (75.2)	8628 (82.7)	
Life Changed due to COVID-19: My Boredom						<.0001[1]
Increased	757 (84.3)	2954 (84.9)	3700 (72.5)	722 (75.8)	8133 (78.0)	
Stayed the same	119 (13.3)	450 (12.9)	1297 (25.4)	180 (18.9)	2046 (19.6)	
Decreased	22 (2.4)	75 (2.2)	103 (2.0)	50 (5.3)	250 (2.4)	
Life Changed due to COVID-19: My Cannabis Use						<.0001[1]
Increased	18 (2.0)	86 (2.5)	35 (0.7)	12 (1.3)	151 (1.4)	
Stayed the same	835 (93.0)	3339 (96.0)	5021 (98.5)	62 (6.5)	9257 (88.8)	
Decreased	45 (5.0)	54 (1.6)	44 (0.9)	878 (92.2)	1021 (9.8)	
Mental Health						
Mental Wellbeing: Flourishing Score						<.0001[2]
Mean (SD)	28.8 (6.0)	28.5 (5.7)	34.2 (4.4)	31.7 (5.6)	31.6 (5.8)	
Anxiety: GAD7						<.0001[2]
Mean (SD)	11.6 (6.1)	10.5 (5.7)	5.1 (4.3)	7.9 (5.7)	7.7 (5.8)	
Physical Activity						
Average Hard and Moderate Physical Activity per day						<.0001[2]

Mean (SD)	101.9 (79.0)	77.1 (66.3)	93.3 (70.2)	106.6 (83.3)	89.8 (71.7)	
Meet the Guidelines for At Least 60 Minutes of Physical Activity per Day						<.0001[1]
Yes	612 (68.2)	1816 (52.2)	3254 (63.8)	666 (70.0)	6348 (60.9)	
No	286 (31.8)	1663 (47.8)	1846 (36.2)	286 (30.0)	4081 (39.1)	

[1]Chi-Square p-value; [2]ANOVA F-test p-value

Table 4.6.3.1 Multinomial Logistic Regression Analysis
Sub Cluster from Cluster 3 (2020-21)

		OR (95% CI)		
Top 5 Factors	**Category**	**Sub Cluster 1**	**Sub Cluster 2**	**Sub Cluster 3**
Meet the Guidelines for At Least 60 Minutes of Physical Activity per Day				
	Yes	Ref	Ref	Ref
	No	1.27 (1.08 – 1.50) **	Ref	1.34 (0.86 – 2.09)
Life Changed due to COVID-19: My Boredom				
	Increased	2.38 (2.02 – 2.80) ***	Ref	1.01 (0.67 – 1.54)
	Stayed the same	Ref	Ref	Ref
	Decreased	2.88 (1.89 – 4.41) ***	Ref	2.42 (0.98 – 6.00)
Life Changed due to COVID-19: My Cannabis Use				
	Increased	2.79 (1.64 – 4.75) **	Ref	19.87 (9.26 – 42.63) ***
	Stayed the same	Ref	Ref	Ref
	Decreased	1.84 (1.09 – 3.09) *	Ref	-
Mental Wellbeing: Flourishing Score		0.83 (0.82 – 0.84) ***	Ref	0.89 (0.86 – 0.92) ***
Anxiety: GAD7		1.16 (1.14 -1.17) ***	Ref	1.07 (1.04 – 1.11) ***
Average combined Hard and Moderate Physical Activity per Day		1.00 (1.00 – 1.00) ***	Ref	1.00 (1.00 – 1.00) *

P<0.0001 ***, P<0.01 **, P<0.05 *
The - sign indicates the odds ratio was not interpretable

4.5 Discussion

Using k-means clustering analysis, this research identified three high-level clusters in the pre-COVID cohort (2018-19) and four high-level clusters in the post-COVID cohort (2020-21). Substance use was the key discriminating variable, and both cohorts contained a small proportion of students who were considered high-risk for the use of cannabis, smoking, tobacco, and opioids. These groups of students also felt depressed or were unwilling to discuss their mental health. Sub-clusters were also generated by k-means clustering for the remaining high-level clusters, excluding substance use-related variables. Substance use was also identified in another COMPASS study that clustered students into low to high-risk groups based on substance use involvement [155]. This study also indicated that the proportion of high-risk students was the smallest compared to most students in the three-year linked dataset [155]. In the high-risk group, students were significantly associated with higher mean CESD scores [155].

The top five factors in the pre-COVID cohort sub-clusters were more related to school topics, including physical activity, bullying, academic performance, and school connectedness. Across the sub-clusters in the pre-COVID cohort, distinct sub-clusters emerged, involving students who participated in school sports, seldom skipped classes, achieved higher grades with homework completed, and had higher scores in school connectedness than those who were less active, experienced bullying, and had lower academic achievement. The sub-clusters with less active students also had higher BMI, underweight or overweight perceptions, and intentions of losing weight. Most students in these sub-clusters disagreed with statements about schools supporting tobacco cessation, providing access to healthy food, or ensuring that no one was bullied.

These findings are consistent with previous literature. School engagement and peer relationships are vital for promoting healthy behaviours among students [157, 158]. Students who reported a more positive school climate were less involved in substance use and more engaged in physical activity [157]. In addition, students who perceived themselves as overweight were more likely to engage in weight loss activities than those who perceived themselves to be at a healthy weight [162]. The strong association between trying to lose weight and screen time also indicated that students attempting to lose weight had higher daily screen time than those trying to maintain or gain weight [36].

Comparing the pre-COVID sub-clusters, the relatively important factors in the post-COVID cohort were more related to individual-level physical activity, weight-related variables, and mental health domains such as flourishing and anxiety. BMI was a critical indicator both before and after the COVID-19 pandemic. Similar to findings from before the pandemic, sub-clusters with students who were overweight or had unknown BMI were less active and did not

meet the 60-minute physical activity guidelines. In other discernable sub-clusters, students with increased cannabis use due to COVID-19, higher BMI status, and who did not meet the 60-minute physical activity guidelines were more likely to be grouped in specific sub-clusters. Among these students, an inverse relationship between mental health well-being and anxiety was observed. Students with lower scores on the Flourishing scale had higher scores on the GAD-7 anxiety scale and were more likely to be in sub-clusters containing students with lower mental health well-being and higher anxiety. These findings align with the existing literature. In a scoping review on changes in cannabis use during COVID-19, 48% of studies found an increase in cannabis use, while only 16% reported a decrease. The review suggested that factors such as boredom, anxiety, and depression were associated with cannabis use [160]. Another study investigating youth substance use and physical activity before and during COVID-19 found a significant decrease in physical activity during the stay-at-home period, with no substantial decline in cannabis, tobacco, or alcohol use [210]. Our results are consistent with other studies showing a significant increase in youth and adolescent depression and anxiety during COVID-19, along with an inverse relationship between anxiety and mental health well-being [39, 40, 41].

There are several limitations to this study. First, the use of complete case analysis for both cohorts may not fully represent the entire cohort, as removing a significant amount of data may result in missing important predictors. More advanced imputation methods, such as Multivariate Imputation by Chained Equations, should be considered for future analysis to capture more student-level data. Second, k-means clustering was sensitive to complex data structures and non-spherical clusters [115, 199]. Even though the dimensionality reduction technique was applied to transform the data components into a lower dimension, some sub-clusters were still not homogeneous and not well-separated. This could be due to the inclusion of variables from multiple domains, creating difficulties in grouping similar factors. K-means clustering may not adapt well to different densities within clusters. Future analysis should focus on specific categories, such as mental health or physical activity, when clustering students' behaviours, rather than incorporating multiple aspects. Another solution to overcome the limitations of k-means is to use more advanced clustering strategies, such as fuzzy or density-based clustering. From a data perspective, using a linked dataset that follows the same students from pre-COVID to post-COVID, rather than two separate cohorts, would be beneficial. However, the sample size might be affected since part of the 2019-2020 survey was conducted online during the COVID-19 pandemic.

4.6 Conclusions

This study examined students' behaviour profiles before and after COVID-19 pandemic using k-means cluster analysis and SHAP plots to investigate the association between key factors and clustering groups. In both pre- and post-COVID high-level clusters, substance use and lower

mental health well-being were the two most significant variables distinguishing student groups. In the pre-COVID sub-clusters, important variables were more related to school activities, such as academic performance, connectedness, and engagement. In contrast, the post-COVID sub-clusters focused more on life changes due to COVID, individual physical activities, mental health well-being, and anxiety. The results of the study provide researchers and school principals with a clearer understanding of the key characteristics of different student groups. These insights could assist in developing school-based programs and health initiatives for targeted interventions. Further research should investigate whether focusing on specific behaviours could better distinguish certain student groups.

Chapter 5

Building an Interpretable Machine Learning Model for Selecting Modifiable Attributes of Childhood Obesity

5.1 Abstract

Objective: The study aims to identify the attributes associated with youth's body mass index (BMI) status by developing a machine learning predictive model among Canadian students.

Methods: The cross-sectional study analyzed the 53,469 grade 9-12 Canadian students participating in the COMPASS study during 2020-21. Data included survey questions on school connectedness, physical activities, healthy eating, demographics, sedentary lifestyles, and weight. Using complete case analysis (CCA) and multiple imputation (MI) approaches, the final predictive model on the multi-class BMI was generated from six machine learning algorithms, including K-Nearest Neighbour (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RT), Multinomial Naïve Bayes (mNB), and XGBboost. The final model was evaluated on the Area Under the Receiver Operating Characteristics Curve (ROC-AUC) performance metrics, F1 score, accuracy, precision, and recall. The importance of the predictors was examined using the Shapley Additive Explanations Technique (SHAP).

Results: XGBoost outperformed the other classifiers, achieving an overall accuracy of 0.64 and ROC-AUC of 0.78 or 0.79 in the CCA and MI approaches, respectively. Based on the SHAP summary plots from CCA and MI approaches, the most important factors associated with youth BMI predictions were weight perception, weight change intention, and gender.

Conclusion: Machine learning effectively predicted youth BMI and identified important attributes associated with this outcome. The interpretative model provides essential insights for school principals and policymakers in developing preventative programs in youth obesity and weight control programs.

5.2 Introduction

Childhood obesity is a global health concern. According to the World Health Organization (WHO), more than 390 million children and adolescents aged between 5 and 19 were overweight in 2022, with 160 million living with obesity [1]. In Canada, nearly 1 in 7 children and youth are considered obese, and 75% of obese children remain obese into adulthood [24]. Obesity in children is a complex public health issue associated with numerous risk factors beyond genetics, including physical inactivity, sedentary behaviour, poor dietary habits, and psychological contributors [20]. For example, parental obesity is a heritable risk for children [26]. From the behavioural perspective, unhealthy dietary intake, such as fast food consumption, has been linked to obesity. Children usually prefer food from fast-food restaurants due to its convenience and low costs despite its low nutritional value and high-calorie content [20]. Besides fast-food consumption, research indicated that the number of hours children spend in sedentary lifestyles, including watching TV, decreases the amount of time in physical activity [20, 30]. Additionally, most studies found that depression and anxiety derived from body dissatisfaction are significantly associated with obesity [20, 36, 37]. Since schools were closed during the COVID-19 pandemic, more than 80% of children have been affected by the school closure, and their risk of resulting childhood obesity has been increasing disproportionally. The confinement of schools has also increased children's sedentary time, and the global lockdown has lowered children's time spent in physical activity and increased consumption of sweets and desserts as coping mechanisms [13, 15, 18, 211].

Childhood obesity also leads to non-communicable diseases, including hypertension, type 2 diabetes, and cardiovascular disease in adulthood. The risks of developing cardiometabolic diseases are higher among children who are obese [20, 21]. Obesity occurs during childhood often transitions to obesity in adulthood [21]. In Canada, total health costs are higher among obese populations [22]. Based on a review of the economic burden of childhood overweight and obesity, the estimated annual direct and indirect healthcare costs related to childhood obesity are projected to be $13.62 billion in direct and $49.02 billion in yearly indirect costs by 2050 [212]. The severity of health and economic consequences related to obesity cannot be neglected.

Due to the complexity of the multidomain influencing factors related to obesity, traditional statistical models have limitations. They struggle with an increasing number of predictors and are less robust in their predictive power [91]. It is also constrained under assumptions such as the absence of multicollinearity or linearity [84]. The machine learning (ML) method has effectively managed complex relationships between multidimensional variables [75]. ML can automatically detect the nonlinear relationship between the predictor variables and the target outcome without observing the potential relationship between these variables [75]. With more predictor variables and fewer instances in the datasets, ML is capable

of handling this type of high-dimensional situation [75]. These advantages of ML demonstrate its potential in childhood obesity research. Several studies have applied various supervised learning models to predict youth obesity [39, 45, 83, 118, 121, 132]. Common supervised learning classifiers used in youth obesity and related risk factors research include K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and Extreme Gradient Boosting (XGBoost). Abdulla et al. [132] applied Bayesian Networks, Neural Networks, Support Vector Machine (SVM), and Decision Tree to predict grade 6 students' BMI based on the student-level questionnaire data. Dugan et al. [118] also utilized various ML classifiers, including Random Forest (RF), Random Tree, J48, ID3, and Naïve Bayes (NB), to model youth obesity after age of two.

Although ML techniques can address problems that traditional statistical approaches might struggle to answer, visualization is another crucial aspect to implement since it engages viewers in understanding the comprehensibility of ML models and the results generated from ML. One of the challenges of ML is providing interpretable results for users to apprehend the meaning of ML-generated results to provide recommendations [213]. Users often experience difficulties understanding the decisions made by ML [215]. Several techniques have been developed to improve the interoperability of the black box of machine learning models, including Shapley Additive exPlanation (SHAP) or Local Interpretable Model-agnostic Explanation (LIME) [208, 215]. SHAP analysis has been popularly adopted by ML algorithms such as XGBoost or LightGBM. The SHAP value represents each factor's contribution towards the outcome. The SHAP summary plot also presents the importance of the factors, sorted descending to provide an overview of the model prediction. [215].

ML models have proved to be more appropriate for identifying and ranking the essential risk factors while having a larger set of variables for future preventive interventions than the traditional statistical model that only works for short implementation [75]. Due to the predictive capabilities of ML, it is encouraged to apply this technique for larger datasets, considering aspects of sociocultural determinants of health and psychosocial factors in addition to physical inactivity [214]. However, fewer studies focus on risk factors among middle childhood populations, who spend most of their time at school. The school environment is also crucial in predicting youth obesity risk [214]. This research aims to address this gap by building an interpretable ML model to predict student BMI and applying SHAP analysis to identify key factors associated with youth obesity risk. The goal is to create an explainable model to inform school-based prevention, policy decisions, or medical interventions targeting childhood obesity.

5.3 Materials and Methods

5.3.1 Study Cohort

This study acquires the 2020-21 cohort dataset from the COMPASS study (Cannabis, Obesity, Mental Health, Physical Activity, Alcohol use, and Sedentary Behaviour (COMPASS) study. This prospective ongoing study collects grade 9-12 students attending secondary schools in Alberta, British Columbia, Ontario, and Quebec in Canada [48, 49]. The COMPASS study (2012-2027) collects student-level through the anonymous, self-administered COMPASS Student Questionnaire (C_q) annually on assessing healthy eating, sedentary behaviour, alcohol, tobacco, and marijuana use, bullying, academic achievement, physical activity, obesity, and school connectedness [48] The participating students were permitted to use active information, passive-consent, and parental permission protocols. [48] Details about the dataset can be found in print or on the website (www.compass.uwaterloo.ca) [49].

5.3.2 Predictors and Outcome

The outcome of the BMI was calculated based on height and weight. Students were asked, "How much do you weigh without your shoes on? (Please enter your weight in either pounds OR in kilograms)." and "How tall are you without your shoes on? (Please enter your height in either feet and inches OR in centimetres)". The derived BMI category was based on the WHO guidelines for ages 5- 19 years of school-aged children and adolescents and calculated from the valid response of sex, BMI (numeric version), and AGE [52]. Since the goal of the ML model focused on classification instead of regression, categorical BMI was employed as the outcome. This study excluded underweight BMI due to the small sample size (1.3%), merged students with overweight or obese to "Overweight and Obesity," and classified students without BMI category into the 'unknown' class.

The predictors include the variables from multiple aspects, focusing on demographics, physical activity, healthy eating, substance use, mental health, bullying, COVID-related, screen time, school-related, sleep-related, and weight-related. Questions distributed in all four provinces were kept, resulting in 144 final variables (Appendix C). Some representative variables from each category were the following.

Demographics included gender (female, male, I describe my gender differently, and I prefer not to say), ethnicity (Black, Asian, Latin American/Hispanic, and Other), age, province, and average money spent per week. Physical activity included average moderate to hard activity per day, participation in school sports, and meeting the 60-minute daily physical activity guidelines. Healthy eating consisted of reasons for skipping breakfast, the total number of days

in weekly fruit consumption, drinking water, eating snacks, restricting diet for weight change, vegetable consumption, and fast food purchased. These were derived variables aggregated from the weekly consumption. For example, “During the past week, on which days did you do the following (Mark all that apply) – Eat fruit (fresh, canned, dried, or frozen?” The response options were “Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday”. The consumption was calculated as the sum of the answers from Monday to Sunday.

Substance-related predictors included smoking and tobacco use, vaping and e-cigarette use, alcohol use, and cannabis use. One of the representatives asked, “Have you used a vape for any of the following reasons (Mark, all that apply)? I have not used a vape.” Mental health factors included the Generalized Anxiety Disorder 7-item Scale (GAD7 [54]), Flourishing [55], and self-reported 10-item Center for Epidemiologic Studies Depression Revised (CESD-R10 [216]).

School-related variables included school connectedness scores, bullying, academic performance such as average grade marks, and the highest education students would like to get. The school connectedness score was calculated by assessing student’s responses of strongly agree, agree, disagree, or strongly disagree among statements in “I feel close to people at school”, “I feel I am part of my school”, “I am happy to be at my school”, “I feel the teachers at my school treat me fairly”, “I feel safe in my school”, and “Getting good grades is important to me” [53]. Bullying included questions about being victims of bullying.

Screen time variables measured the average daily minutes spent watching TV, doing homework, surfing the internet, texting, and total daily sedentary. Sleep-related predictors included the average length of sleep per night in a week, sleep quality rating and meeting guidelines of at least 8 hours of sleep. COVID-related variables asked, “How has your life changed because of COVID-19?” addressing areas like stress, boredom, loneliness, anxiety, time spent on screen time, physical activity, substance use, physical activity and sleep. Lastly, weight-related variables included weight perception (“How do you describe your weight?”) and weight change intention (“Which of the following are you trying to do about your weight?”).

5.3.3 Data Preparation

The data preparation followed two approaches: complete case analysis (CCA) and multiple imputations (MI) to handle missing. For the complete case analysis, any missing appeared in predictors were removed from the analysis. The percentage of missingness in the predictors can be found in Appendix C.

Previous research on COMPASS data [174] has shown it is difficult to determine if the missing data is Missing Completely at Random (MCAR), Missing Not at Random (MNAR), or Missing at Random (MAR). Therefore, missingness was treated as MAR since the information needed for NMAR is not generally available [174]. Imputation by Chained Equations (MICE) works under the assumption of MAR, and this research used the miceforest python library with lightgbm random forest [217]. Incorporating MICE with Random Forest could accommodate interactions and nonlinearities and would not require much hyperparameter tuning [217, 218]. It also showed that the result with random forest MICE was more coverage on the confidence interval and more efficient [218]. This study independently tested the imputation with 5, 10, 20, 30, 40, 50 and 60 iterations. The mean convergence figure for each iteration was generated to visualize how well the data converged, and iteration 60 was implemented due to its better convergence. The illustration of the mean convergence among iterations 5, 20, and 60 is in the Appendix. Besides the missing data handling, Cramer's score was calculated between multi-class BMI and weight to measure the strength of their association. The score of 0.25 indicated a weak to moderate association between these two variables, so weight perception was retained as a predictor.

5.3.4 Data Preprocessing

The CCA and MI datasets were split into a randomly selected 80% training dataset and a 20% testing dataset. Standardization was applied for continuous variables. The outcome BMI category was significantly imbalanced, with 39.90% healthy weight samples, 19.71% overweight and obese, and 48.08% unknown. To overcome the issue of imbalance, the Synthetic Minority Oversampling Technique (SMOTE) – Tomek Links Method was applied to address the imbalance in the training datasets. Specifically, SMOTE generates synthetic data to increase samples from the minority class, and the Tomek links technique is an undersampling method to remove the random data from the majority class [209].

To experiment with the effect of feature selection and dimensionality reduction and decide which data preprocessing approach would be most ideal for the final model prediction, a Random Forest with feature selection approach and Linear Discriminant Analysis (LDA) were implemented. (LDA) is a supervised classification technique that aims to reduce the dimensionality of the dataset. This approach helps transform the original dataset into a new subset with lower dimensionality (fewer inputs) to increase the computing efficiency and avoid the degree of overfitting [189]. The purpose of feature selection is to extract the most relevant features from the high-dimensional dataset to minimize redundancy and noise. A Random Forest classifier was trained to retrieve the feature importance score of each input on the BMI category using the feature_importantces functions [189]. Each input was ranked in descending order, and

the top 50 variables were selected (Appendix C). The MI approach dataset followed the same data preprocessing steps as CCA.

5.3.5 Machine Learning Model Development

For the three preprocessing strategies, LDA-transformed, top 50 features, and All variables, six ML classifiers, including K-Nearest Neighbour (KNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (NB), and XGBboost were trained on each preprocessing method's dataset. This process involved a five-fold outer cross-validation and a three-fold inner stratified cross-validation for the grid search process. The classifier with the highest F1 score for each preprocessing strategy was selected for further hyperparameter tuning.

After selecting the best model for each preprocessing strategy, hyperparameter tuning was applied by using Optuna to search for the optimal hyperparameters. Optuna uses a Bayesian optimization algorithm, which is known for its higher performance and faster optimization time [219]. The three models for each preprocessing approach were compared based on the evaluation metrics, accuracy, precision, recall, ROC-AUC, and F1 score. Recall and precision refer to the probabilities of true positive cases over the actual number of positive cases and the probabilities of true positive cases among all predicted as positive. Accuracy refers to the percentage of correct predictions over the total number of predictions [196]. ROC-AUC represents how well the model can distinguish between positive and negative cases, and the F1 score refers to the harmonic mean of precision and recall [196]. The final model selection was based on the overall highest performance across these metrics. Once the final model was chosen, stratification was performed for females and males on BMI prediction (CCA model), excluding the gender minority (I describe my gender differently, and I prefer not to say). The reason was that females (50.5%) and males (44.2%) had comparable sample sizes, but gender minorities only accounted for 1.5% of the population.

Lastly, having interpretability behind the built ML models could assist health professionals significantly in understanding why the model made a specific prediction [215]. The final model implemented the SHAP technique to draw the feature importance plot and further interpret each input's contribution to model prediction based on the corresponding mean absolute SHAP values. The SHAP summary plot showed the global importance factors on their ranked importance, with the most relevant variables ranked from the top in the order of their influences on the outcome. To understand the individual level prediction for the BMI category, SHAP waterfall plots were applied to investigate the risk factors that predict healthy weight, overweight and obese, and unknown. All analyses were conducted in Python (Version 3.10.13).

All supervised learning models were implemented through the sciki-learn Python library (Version 1.2.2). SHAP version 0.42.1.

5.4. Results

5.4.1 Baseline Characteristics

Table 5.4 shows the baseline characteristics of students who contributed to the COMPASS survey in 2020-21. Among the total 53469 students, 10.7% (N=5725) of students self-reported as overweight and obese, while 48.1% of students did not report BMI information. Among the overweight and obese groups, 57.9% (N=3313) were males, and 42.1% (N=2412) were females. All gender minority students (N=2046, 8.0%) in the unknown BMI category did not have a response for BMI. Overweight and obese students had the highest number of responses (57.7%) in losing weight, followed by 35.3% of students within the unknown weight group. In the CCA approach, there was a total number of 28614 samples with completed records. Among these students, 46.0% (N = 13159) were in healthy weight group, 11.43% (N=3271) in overweight and obese group, and 41.2% (N=11798) in the unknown group.

Table 5.4 Baseline Characteristics of COMPASS Cohort 2020-21

N (%) or Mean (SD)

	Underweight (N=701)	**Healthy Weight (N=21335)**	**Overweight and Obese (N=5725)**	**Unknown (N=25708)**	**Total (N=53469)**
AGE					
Mean (SD)	14.9 (1.6)	15.2 (1.4)	15.2 (1.5)	14.7 (1.6)	15.0 (1.5)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	733 (2.9)	733 (1.4)
Gender					
Female	276 (39.4)	11772 (55.2)	2412 (42.1)	12521 (48.7)	26981 (50.5)
Male	425 (60.6)	9563 (44.8)	3313 (57.9)	10357 (40.3)	23658 (44.2)
Gender Minority	0 (0.0)	0 (0.0)	0 (0.0)	2046 (8.0)	2046 (3.8)
Missing	0 (0.0)	0 (0.0)	0 (0.0)	784 (3.0)	784 (1.5)
Province					
AB	25 (3.6)	983 (4.6)	300 (5.2)	974 (3.8)	2282 (4.3)
BC	90 (12.8)	2264 (10.6)	517 (9.0)	2589 (10.1)	5460 (10.2)
ON	162 (23.1)	5758 (27.0)	1870 (32.7)	6250 (24.3)	14040 (26.3)

QC	424 (60.5)	12330 (57.8)	3038 (53.1)	15895 (61.8)	31687 (59.3)
Ethnicity					
White	558 (79.6)	17872 (83.8)	4669 (81.6)	18977 (73.8)	42076 (78.7)
Non-White	143 (20.4)	3431 (16.1)	1042 (18.2)	5895 (22.9)	10511 (19.7)
Missing	0 （0）	32 (0.1)	14 (0.2)	836 (3.3)	882 (1.6)
Financial Comfortableness					
More comfortable	188 (26.8)	6314 (29.6)	1548 (27.0)	5476 (21.3)	13526 (25.3)
As comfortable	448 (63.9)	13635 (63.9)	3632 (63.4)	16933 (65.9)	34648 (64.8)
Less comfortable	60 (8.6)	1358 (6.4)	534 (9.3)	1885 (7.3)	3837 (7.2)
Missing	5 (0.7)	28 (0.1)	11 (0.2)	1414 (5.5)	1458 (2.7)
Weight-Related					
Weight Perception					
Underweight	500 (71.3)	4027 (18.9)	102 (1.8)	3568 (13.9)	8197 (15.3)
About the right weight	178 (25.4)	15105 (70.8)	2278 (39.8)	14221 (55.3)	31782 (59.4)
Overweight	19 (2.7)	2128 (10.0)	3321 (58.0)	6119 (23.8)	11587 (21.7)
Missing	4 (0.6)	75 (0.4)	24 (0.4)	1800 (7.0)	1903 (3.6)
Weight Change Intention					
Lose weight	38 (5.4)	4920 (23.1)	3300 (57.8)	8490 (33.0)	16748 (31.3)
Gain weight	272 (38.8)	3827 (17.9)	298 (5.2)	2248 (8.7)	6645 (12.4)
Stay the same	90 (12.8)	5350 (25.1)	999 (17.4)	4835 (18.8)	11274 (21.1)
Not trying anything about the weight	301 (42.9)	7220 (33.8)	1123 (19.6)	8472 (33.0)	17116 (32.0)
Missing	0 (0)	18 (0.1)	5 (0.1)	1663 (6.5)	1686 (3.2)
School-Related					
School Connectedness Score					
Mean (SD)	18.6 (3.5)	18.7 (3.2)	18.4 (3.4)	18.2 (3.5)	18.5 (3.4)
Missing	54 (7.7)	1430 (6.7)	503 (8.8)	4155 (16.2)	6142 (11.5)
School Marks					
80-100%	398 (56.8)	12730 (59.7)	2878 (50.3)	10425 (40.6)	26431 (49.4)
50-70%	248 (35.4)	7100 (33.3)	2309 (40.3)	10838 (42.2)	20495 (38.3)
Below 50%	4 (0.6)	84 (0.4)	46 (0.8)	352 (1.4)	486 (0.9)
Missing	51 (7.3)	1421 (6.7)	492 (8.6)	4093 (15.9)	6057 (11.3)

Bullied Last 30 Days					
Yes	65 (9.3)	1672 (7.8)	563 (9.8)	2254 (8.8)	4554 (8.5)
No	566 (80.7)	17680 (82.9)	4494 (78.5)	18625 (72.4)	41365 (77.4)
Missing	70 (10.0)	1983 (9.3)	668 (11.7)	4829 (18.8)	7550 (14.1)
Substance Use					
Marijuana or Cannabis Use					
Yes	80 (11.4)	4205 (19.7)	1285 (22.4)	3182 (12.4)	8752 (16.4)
No	586 (83.6)	16173 (75.8)	4103 (71.7)	19292 (75.0)	40154 (75.1)
Missing	35 (5.0)	957 (4.5)	337 (5.9)	3234 (12.6)	4563 (8.5)
E-cigarette Use					
Yes	173 (24.7)	8310 (39.0)	2452 (42.8)	6907 (26.9)	17842 (33.4)
No	496 (70.8)	12172 (57.1)	2972 (51.9)	15780 (61.4)	31420 (58.8)
Missing	32 (4.6)	853 (4.0)	301 (5.3)	3021 (11.8)	4207 (7.9)
COVID-19					
Life Changed due to COVID-19: Stress Has					
Increased	286 (40.8)	10035 (47.0)	2535 (44.3)	9905 (38.5)	22761 (42.6)
Stayed the same	326 (46.5)	8837 (41.4)	2385 (41.7)	10119 (39.4)	21667 (40.5)
Decreased	35 (5.0)	896 (4.2)	287 (5.0)	1244 (4.8)	2462 (4.6)
Missing	54 (7.7)	1567 (7.3)	518 (9.0)	4440 (17.3)	6579 (12.3)
Life Changed due to COVID-19: Boredom					
Increased	428 (61.1)	14458 (67.8)	3801 (66.4)	13948 (54.3)	32635 (61.0)
Stayed the same	181 (25.8)	4640 (21.7)	1185 (20.7)	6106 (23.8)	12112 (22.7)
Decreased	36 (5.1)	640 (3.0)	209 (3.7)	1170 (4.6)	2055 (3.8)
Missing	56 (8.0)	1597 (7.5)	530 (9.3)	4484 (17.4)	6667 (12.5)
Life Changed due to COVID-19: Sleep					
Increased	179 (25.5)	5223 (24.5)	1475 (25.8)	5687 (22.1)	12564 (23.5)
Stayed the same	296 (42.2)	8237 (38.6)	2110 (36.9)	8743 (34.0)	19386 (36.3)
Decreased	172 (24.5)	6258 (29.3)	1607 (28.1)	6737 (26.2)	14774 (27.6)
Missing	54 (7.7)	1617 (7.6)	533 (9.3)	4541 (17.7)	6745 (12.6)
Life Changed due to COVID-19: Cannabis Use					
Increased	33 (4.7)	1269 (5.9)	396 (6.9)	1149 (4.5)	2847 (5.3)

Stayed the same	621 (88.6)	18173 (85.2)	4718 (82.4)	21881 (85.1)	45393 (84.9)
Decreased	47 (6.7)	1893 (8.9)	611 (10.7)	2678 (10.4)	5229 (9.8)
Physical Activity					
Participation in School Competitive Sport					
Yes	127 (18.1)	6373 (29.9)	1697 (29.6)	4714 (18.3)	12911 (24.1)
No	504 (71.9)	12768 (59.8)	3389 (59.2)	16517 (64.2)	33178 (62.1)
None offered at my school	41 (5.8)	1403 (6.6)	371 (6.5)	1349 (5.2)	3164 (5.9)
Missing	29 (4.1)	791 (3.7)	268 (4.7)	3128 (12.2)	4216 (7.9)
Average Combined Hard and Moderate physical activity per day					
Mean (SD)	88.8 (73.7)	94.2 (75.3)	98.6 (79.2)	91.0 (84.7)	93.2 (80.1)
Missing	46 (6.6)	1313 (6.2)	392 (6.9)	4379 (17.0)	6130 (11.5)
Health Eating					
Diet restriction in the Past Week (Number of Days)					
Mean (SD)	0.8 (2.0)	1.3 (2.5)	2.2 (2.9)	1.4 (2.5)	1.5 (2.5)
Missing	49 (7.0)	1091 (5.1)	370 (6.5)	3747 (14.6)	5257 (9.8)
Fast Food Purchased in the Past Week (Numbers of Days)					
Mean (SD)	1.1 (1.5)	1.1 (1.3)	1.1 (1.4)	1.0 (1.3)	1.1 (1.3)
Missing	49 (7.0)	1091 (5.1)	370 (6.5)	3747 (14.6)	5257 (9.8)
Vegetable Consumption in the Past Week (Numbers of days)					
Mean (SD)	5.4 (2.3)	5.4 (2.3)	5.2 (2.5)	4.8 (2.7)	5.1 (2.5)
Missing	49 (7.0)	1091 (5.1)	370 (6.5)	3747 (14.6)	5257 (9.8)
Fruit Consumption in the Past Week (Numbers of days)					
Mean (SD)	4.9 (2.6)	5.1 (2.5)	4.9 (2.6)	4.5 (2.7)	4.8 (2.6)
Missing	49 (7.0)	1091 (5.1)	370 (6.5)	3747 (14.6)	5257 (9.8)
Screen Time					
Watching TV (Minutes per day)					
Mean (SD)	520.5 (310.3)	493.1 (281.9)	552.3 (328.3)	594.3 (442.3)	546.6 (372.3)
Missing	15 (2.1)	431 (2.0)	139 (2.4)	2495 (9.7)	3080 (5.8)
Surfing Internet (Minutes per day)					
Mean (SD)	107.4 (127.7)	109.1 (116.7)	121.1 (132.1)	131.8 (155.4)	120.9 (138.0)
Missing	15 (2.1)	431 (2.0)	139 (2.4)	2495 (9.7)	3080 (5.8)

Doing Homework (Minutes per day)					
Mean (SD)	95.1 (84.4)	88.9 (81.5)	85.6 (86.6)	93.3 (100.5)	90.7 (91.3)
Missing	15 (2.1)	431 (2.0)	139 (2.4)	2495 (9.7)	3080 (5.8)
Mental Health					
Mental Health Wellbeing (FLOURISH)					
Mean (SD)	31.8 (6.1)	32.3 (5.6)	31.8 (6.0)	31.0 (6.4)	31.6 (6.1)
Missing	55 (7.9)	1377 (6.5)	477 (8.3)	4272 (16.6)	6181 (11.6)
Anxiety (GAD7)					
Mean (SD)	6.9 (5.8)	6.9 (5.8)	6.9 (5.89)	7.3 (6.1)	7.1 (5.9)
Missing	70 (10.0)	1706 (8.0)	555 (9.7)	4701 (18.3)	7032 (13.2)
Depression (CESD)					
Mean (SD)	9.6 (6.5)	9.6 (6.5)	9.7 (6.6)	10.4 (6.9)	10.0 (6.7)
Missing	74 (10.6)	2158 (10.1)	674 (11.8)	5489 (21.4)	8395 (15.7)

5.4.2 Model Performance

After experimenting with the six ML classifiers, SVM, Random Forest, and XGBoost were the models that achieved the highest F1 scores corresponding to LDA-transformed, Top 50 variables, and all variables datasets when nested cross-validation performed across the six ML algorithms (Table 5.5). The optimal parameters from Optuna were used to train SVM, RF, and XGBoost on the training dataset (Appendix C). The details of the overall and model performance for healthy weight, overweight and obese, and unknown categories in both CCA and MI are in Tables 5.6, 5.7 and 5.8.

From the CCA results, XGboost had a higher ROC-AUC (0.78) than RF and SVM. In MI, SVM also performed worse than RF and XGBoost. XGBoost and RF had a similar performance. In XGBoost, the F1 score and ROC-AUC were higher in both CCA and MI in the overweight and obese categories than in RF. In XGBoost CCA, the F1 score was 0.43 and 0.80 for ROC-AUC score, while in MI, the F1 and ROC-AUC scores were 0.37 and 0.80. Both approaches outperformed RF CCA (F1= 0.40 and ROC-AUC=0.80) and RF MI (F1 = 0.34, ROC-AUC = 0.78). Therefore, XGBoost was selected as the final model for the SHAP analysis.

Table 5.5 F1 Score of Six Machine Learning Classifiers on the Three Preprocessing Strategies

	CCA			**MI**		
ML Classifiers (F1 score)	**LDA-transformed**	**Top 50**	**All variables**	**LDA-transformed**	**Top 50**	**All variables**
Support Vector machine	0.60	0.61	0.57	0.61	0.62	0.59
Logistic Regression	0.60	0.58	0.57	0.60	0.59	0.57
XGBoost	0.59	0.72	0.68	0.60	0.73	0.69
Random Forest	0.55	0.72	0.65	0.56	0.74	0.67
K-Nearest Neighbor	0.55	0.51	0.56	0.55	0.52	0.58
Multinomial Naive Bayes	0.39	0.50	0.49	0.39	0.50	0.50

Table 5.6 Overall Model Performance in Predicting BMI Status

Models	Preprocessing Strategies	Accuracy	Precision	Recall	F1	ROC-AUC
CCA						
Support Vector Machine	LDA-transformed	0.57	0.61	0.57	0.58	0.74
Random Forest	Top 50	0.64	0.63	0.64	0.63	0.77
XGBoost	All variables	0.64	0.63	0.64	0.63	0.78
MI						
Support Vector Machine	LDA-transformed	0.57	0.63	0.57	0.59	0.75
Random Forest	Top 50	0.64	0.64	0.65	0.64	0.78
XGBoost	All variables	0.64	0.64	0.64	0.64	0.79

Table 5.7 Performance Metrics of each BMI Status in CCA

CCA	Preprocessing Strategies	Performance Metrics	Healthy Weight	Overweight and Obese	Unknown
Support Vector Machine	LDA-Transformed	Overall Accuracy	0.57		
		Precision	0.67	0.31	0.62
		Recall	0.62	0.58	0.51
		F-1 score	0.64	0.40	0.56
		ROC-AUC	0.74	0.78	0.70
Random Forest	Top 50	Overall Accuracy	0.64		
		Precision	0.67	0.48	0.63
		Recall	0.73	0.34	0.63
		F-1 score	0.70	0.40	0.63
		ROC-AUC	0.77	0.80	0.75
XGBoost	All Variables	Overall Accuracy	0.64		
		Precision	0.67	0.41	0.64
		Recall	0.72	0.46	0.61
		F-1 score	0.69	0.43	0.63
		ROC-AUC	0.78	0.80	0.75

Table 5.8 Performance Metrics of each BMI status in MI

MI	Preprocessing Strategies	Performance Metrics	Healthy Weight	Overweight and Obese	Unknown
Support Vector Machine	LDA-Transformed	Overall Accuracy	0.57		
		Precision	0.62	0.28	0.71
		Recall	0.64	0.59	0.51
		F-1 score	0.63	0.37	0.60
		ROC-AUC	0.75	0.78	0.72
Random Forest	Top 50	Overall Accuracy	0.64		
		Precision	0.65	0.48	0.66
		Recall	0.64	0.27	0.74
		F-1 score	0.64	0.34	0.70
		ROC-AUC	0.78	0.78	0.76
XGBoost	All Variables	Overall Accuracy	0.64		
		Precision	0.64	0.45	0.68

		Recall	0.67	0.31	0.70
		F-1 score	0.65	0.37	0.69
		ROC-AUC	0.79	0.80	0.77

5.4.3 SHAP Analysis

The SHAP interpretation was based on the CCA approach (Figure 5 – 5.3). The MI approach generated a similar result to CCA (Figure 5.4 – 5.7).

Figure 5 shows the global importance by ranking the importance of each feature in descending order across BMI status. The top five variables that had the most influence on BMI were student weight perception, sex, the intention of changing their weight, age and school marks in the XGBoost prediction model. To understand the relative feature importance within each BMI category, the SHAP beeswarm plot for the healthy weight, overweight and obese, and unknown were used to quantify the influence of each predictor based on the SHAP values (Figure 5.1 A-C). All BMI classes ranked weight perception as the most important factor associated with the model's output. Using the overweight and obese SHAP plot as an example, the higher feature value (red) in weight perception indicated a positive SHAP value, classifying students with higher values in weight perception as being in this category. Based on the boxplots in Figure 5.2 B, students who perceived themselves as "very overweight" had the highest SHAP value. This showed that the perception of being very overweight had a positive contribution to being identified as the overweight and obese class.

On the contrary, the boxplot (Figure 5.2 A) that showed the distribution of SHAP value for students more likely to be in the healthy weight class, the weight perception of being 'very overweight' had the lowest SHAP value (negative). This showed that this weight perception decreased the model prediction for classifying students as a healthy weight. For the unknown class boxplot, perceiving oneself as very underweight or slightly underweight strongly increased the model's classification of students into unknown classes (Figure 5.2 C). Since gender was ranked as the most important factor for the unknown class, the boxplot (Figure 5.2 D) illustrated the effect of gender feature values on SHAP values. "I describe my gender in a different way" and "I prefer not to say" had a higher positive SHAP value compared to females and males. This indicated that students who chose the above two categories were more likely to be grouped into the unknown weight class.

To comprehensively understand the interpretability of individual student predictions, we also generated the SHAP Waterfall plots for each BMI class. The red arrows pointing to the right

refer to a positive contribution to predicting this student as healthy weight, overweight and obese, or unknown. In reverse, blue arrows indicated movement against prediction in the corresponding weight class. Each variable was ranked from top to bottom based on its importance in contributing to the model prediction.

In Figure 5.3 A, a healthy weight sample chose "3 = About the right weight" when asked, "How do you describe your weight?". This choice significantly (+0.57) pushed the model towards predicting this student as being in the healthy weight class. This positive influence was also seen in the response to school marks (1 = 90%-100%). However, when choosing "1 = Lose weight" for the question "Which of the following are you trying to do about your weight?", it moved in a negative direction, suggesting that the intention to lose weight was inversely related being classified as a healthy weight. For the sample in the overweight and obese class (Figure 5.3 B), the perception of very overweight was the major contributor to overweight and obese prediction (+1.5). Meanwhile, being female decreased (-0.41) the prediction of being overweight and obese. In the unknown class (Figure 5.3 C), older age and a goal of attaining a university degree as the highest education reduce the likelihood of being in the unknown weight class. Conversely, being female and perceiving oneself as slightly underweight pushed the prediction toward the classification of unknown weight for this student.

Compared with the overall population, SHAP plots also presented the overweight and obese and unknown BMI stratified by females and males. The two most important factors associated with these BMI predictions were still weight perception and weight change intention. This finding was consistent with the ones in the overall population. When looking at the top five factors in the overweight and obese plot (Figure 5.8), it was noted that sports-related activities and exercise were more important for males. At the same time, the risk and difficulties of substance use and sleep were significant for females. For example, the lower feature value (blue) in the competitive school sports indicated "Yes – participated," and the increased number of days in muscle training (red) had a positive SHAP value, contributing to the model predicting overweight and obese BMI. In females, lower feature levels in the risk of cannabis use, which was "no risk" or "low risk," resulted in a positive SHAP value, suggesting a direction towards the overweight and obese BMI.

In the unknown BMI plot (Figure 5.9), age and school marks were ranked among the top three factors in both females and males. This was slightly different than the overall population, where gender and weight perception were the top two most important factors (Figure 5.1 C). For males and females, younger age and lower school marks (high features in red, such as 50% or lower grades) had moved positively towards an unknown BMI prediction. There was no noticeable difference between these two groups among the top five factors, except in the categories of substance use. For example, males who were younger and consumed more than a

sip of alcohol contributed positively to this prediction. Conversely, females who chose either "easy" or "don't know" for the difficulty of obtaining marijuana were more likely to be classified into the unknown BMI category.

Figure 5. SHAP summary plot for all BMI status in CCA

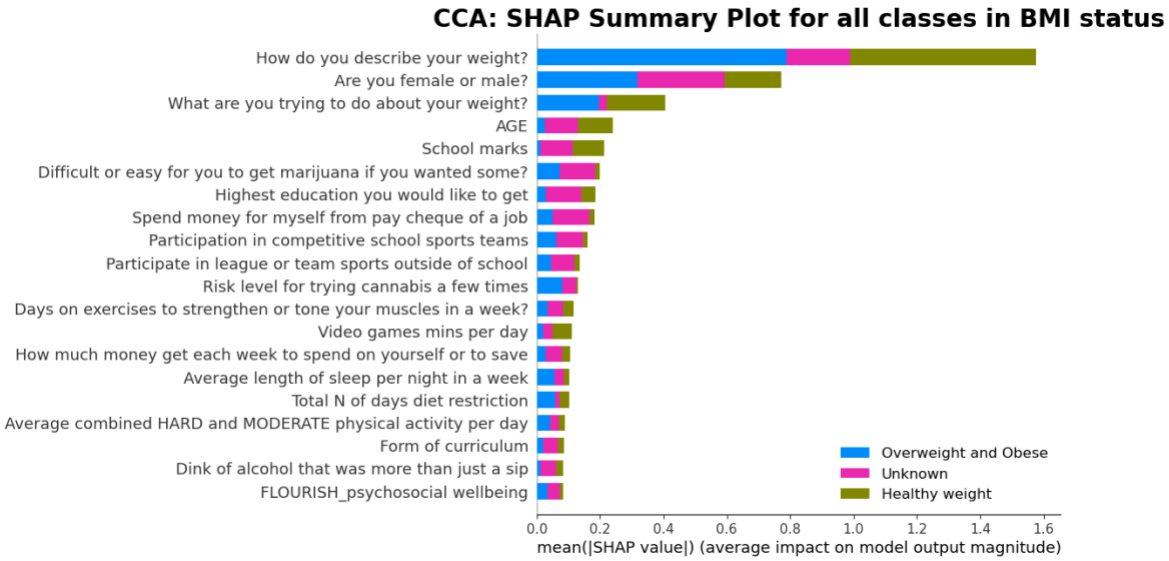
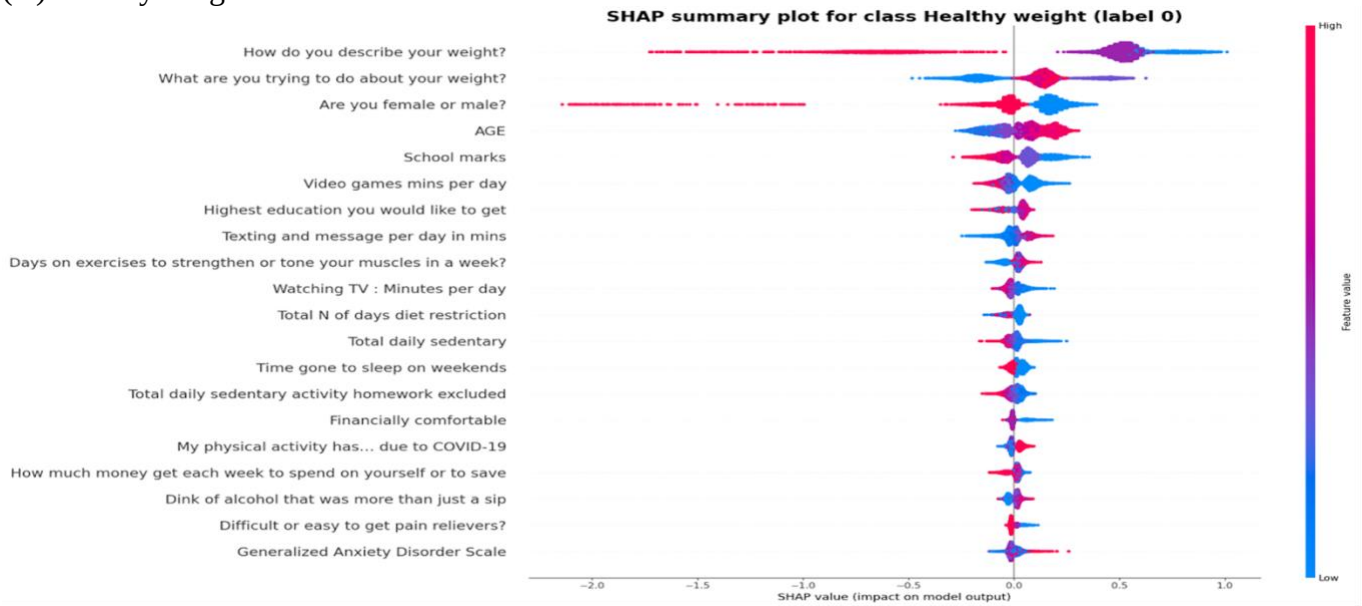
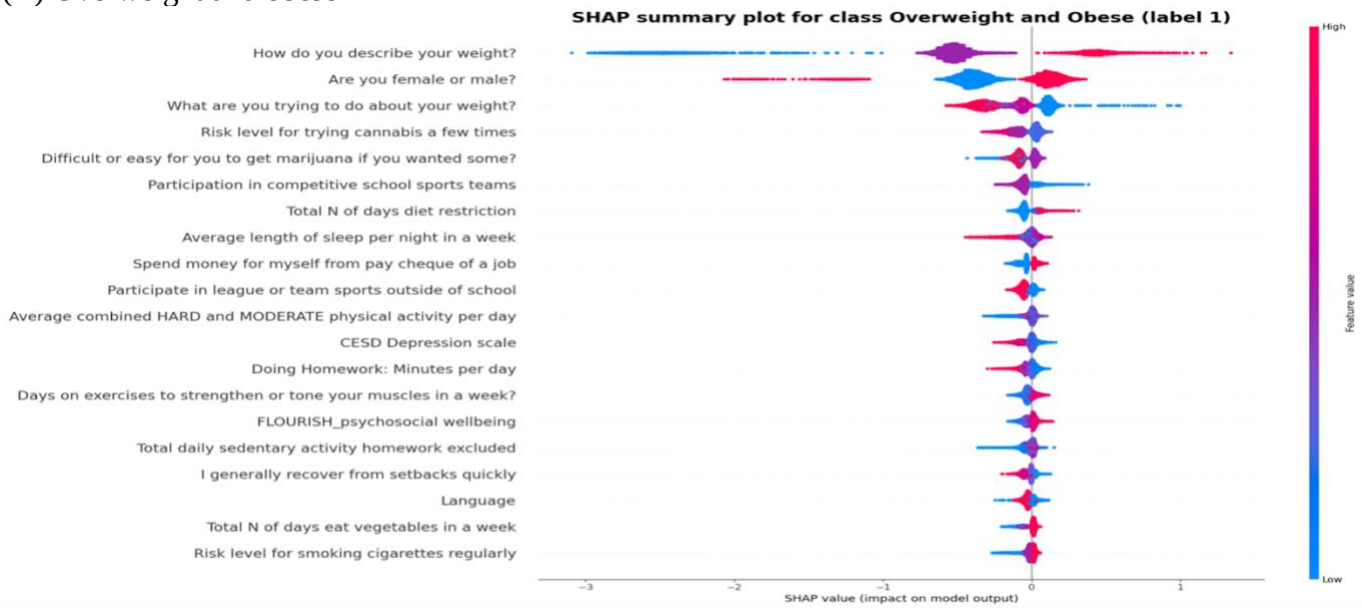


Figure 5.1 SHAP beeswarm plots for health weight, overweight and obese, and unknown in CCA

(A) Healthy weight



(B) Overweight and obese



(C) Unknown

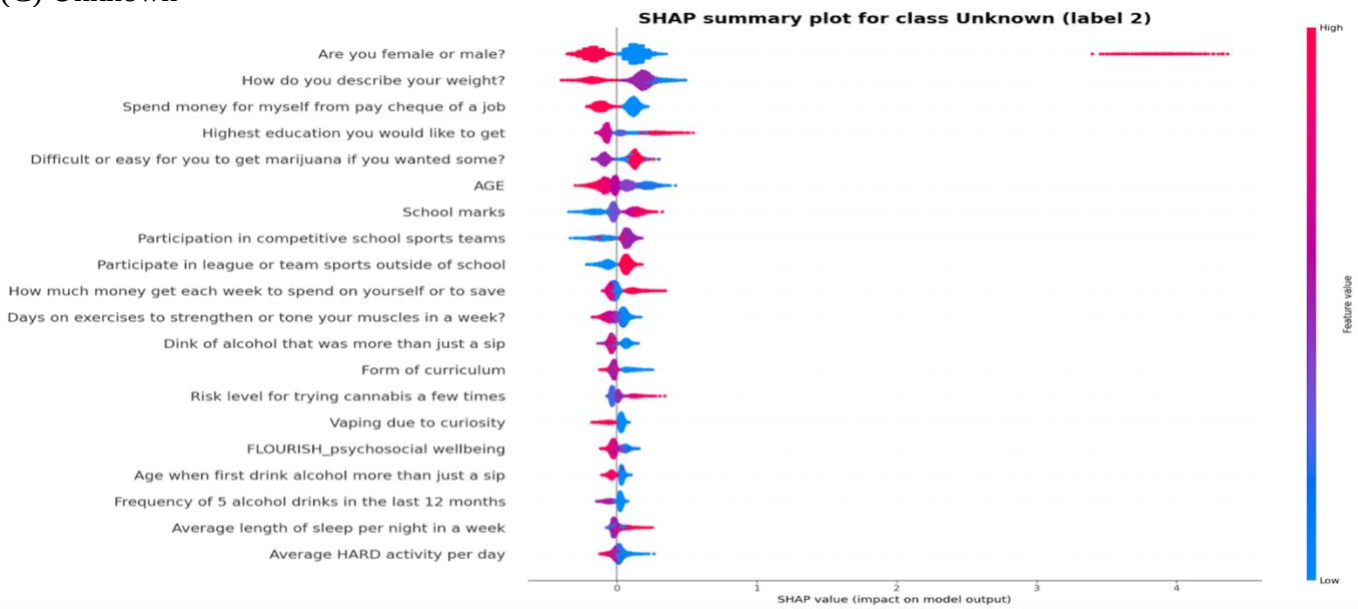
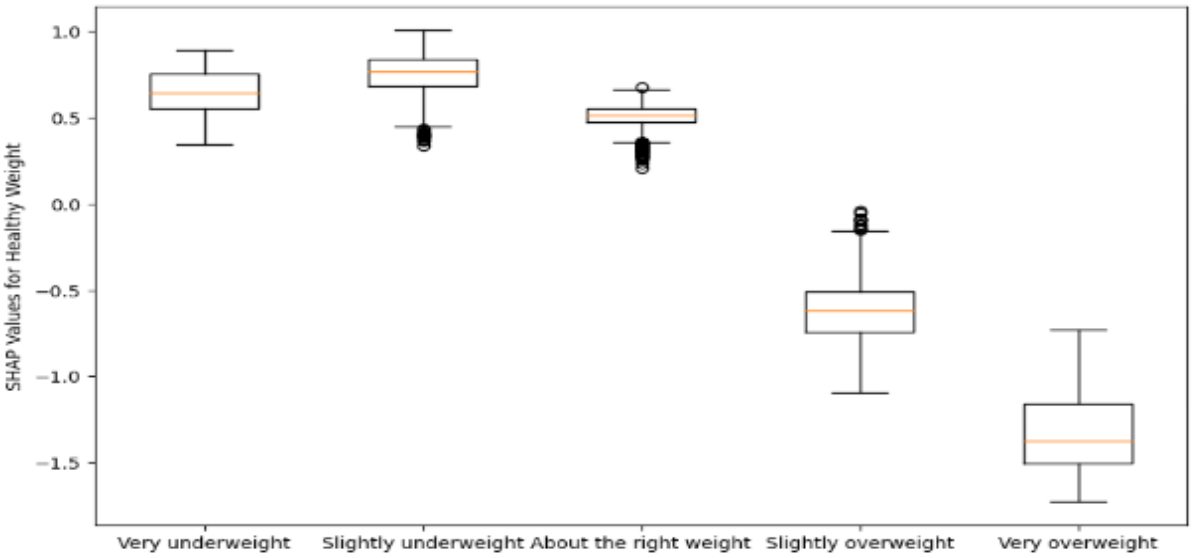
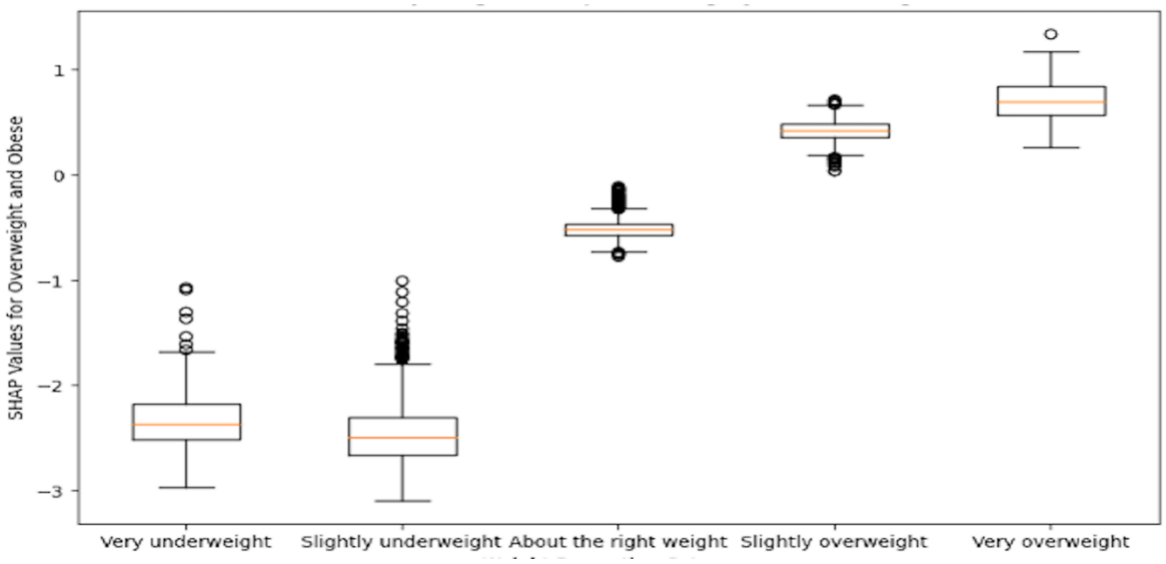

SHAP summary plot for class Unknown (label 2)
Are you female or male?
How do you describe your weight?
Spend money for myself from pay cheque of a job
Highest education you would like to get
Difficult or easy for you to get marijuana if you wanted some?
AGE
School marks
Participation in competitive school sports teams
Participate in league or team sports outside of school
How much money get each week to spend on yourself or to save
Days on exercises to strengthen or tone your muscles in a week?
Dink of alcohol that was more than just a sip
Form of curriculum
Risk level for trying cannabis a few times
Vaping due to curiosity
FLOURISH_psychosocial wellbeing
Age when first drink alcohol more than just a sip
Frequency of 5 alcohol drinks in the last 12 months
Average length of sleep per night in a week
Average HARD activity per day
0
1
2
3
4
SHAP value (impact on model output)
High
Feature value
Low

Figure 5.2 Boxplot of weight perception and gender SHAP values in each BMI status (CCA)

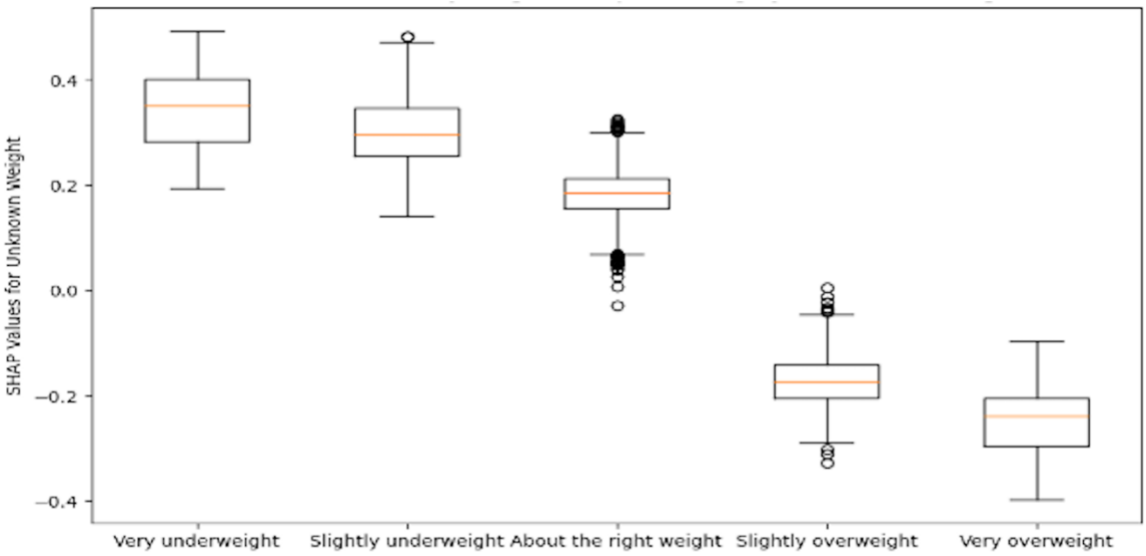
(A) Weight perception in healthy weight



(B) Weight perception in overweight and obese



(C) Weight perception in unknown



(D) Gender in unknown

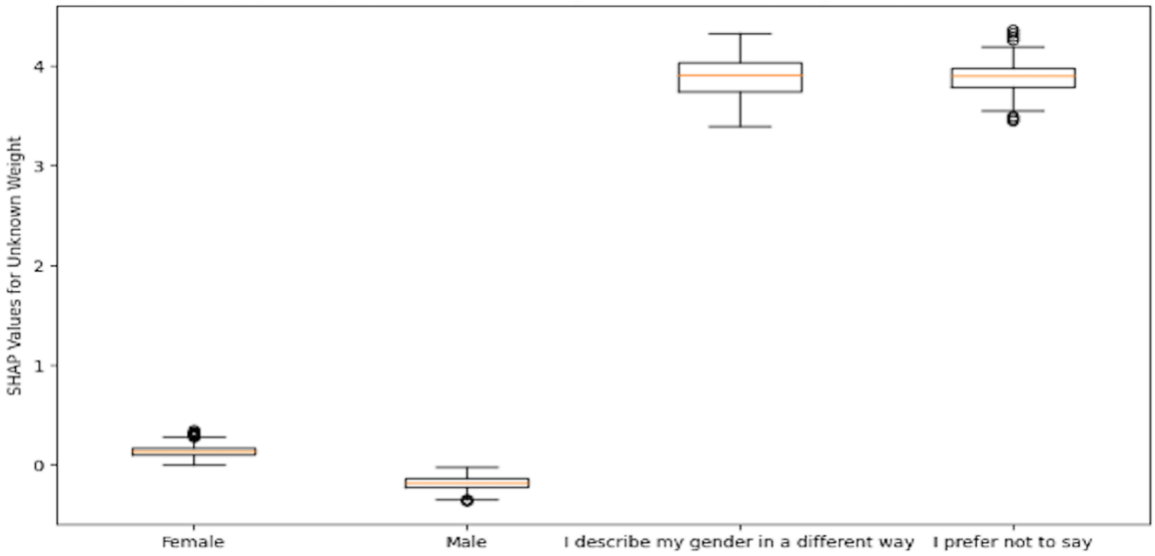
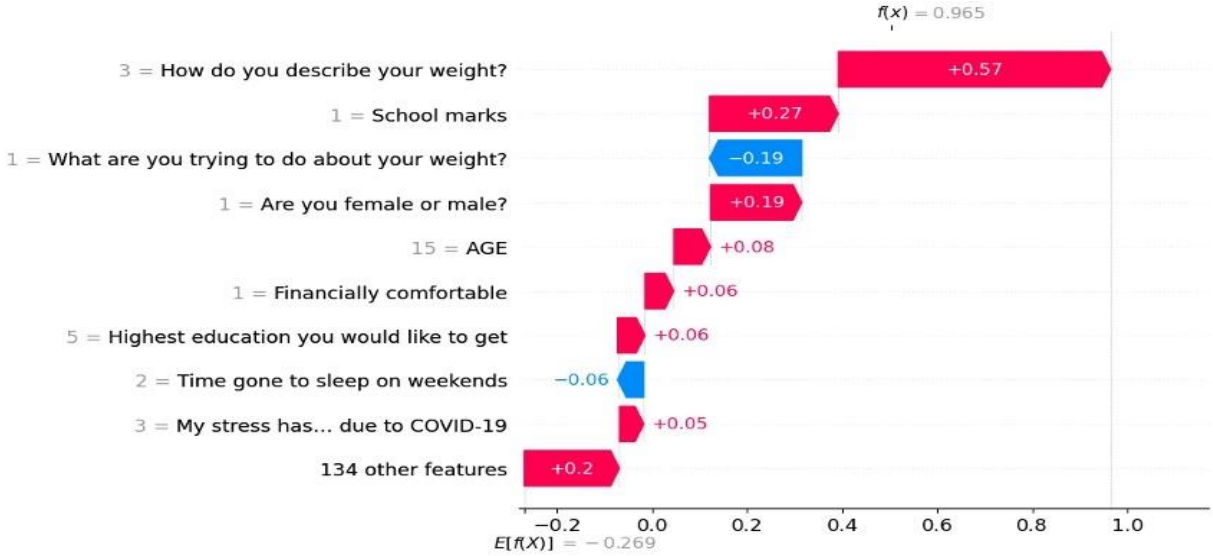
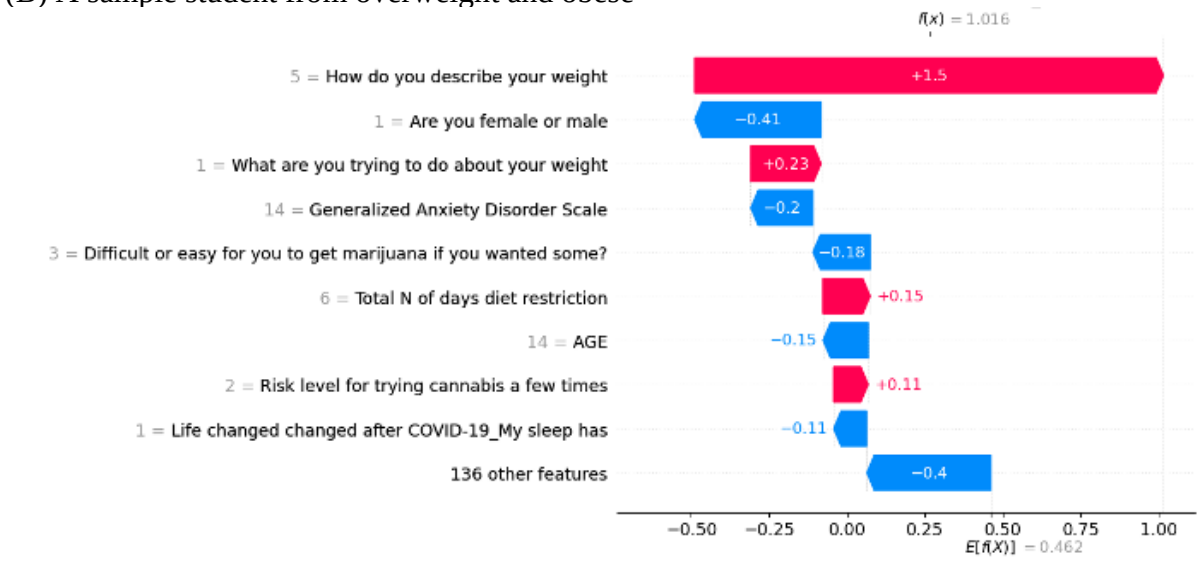


Figure 5.3 SHAP Waterfall plots for each BMI Class

(A) A sample student from healthy weight



(B) A sample student from overweight and obese



(C) A sample student from unknown

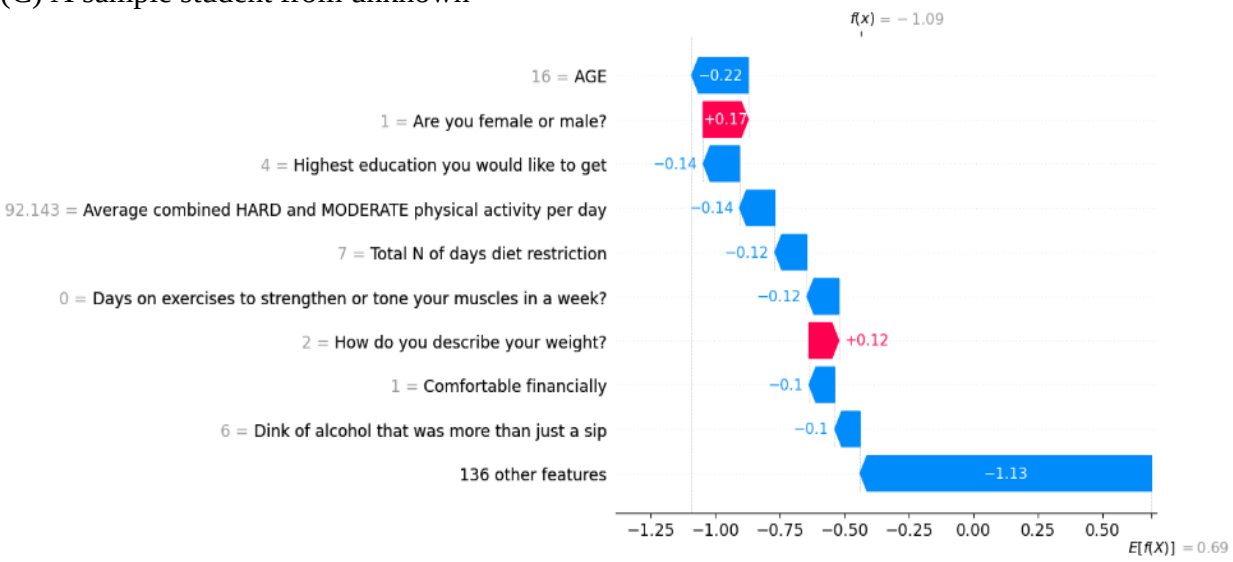


Figure 5.4 SHAP summary plot for all BMI status in MI

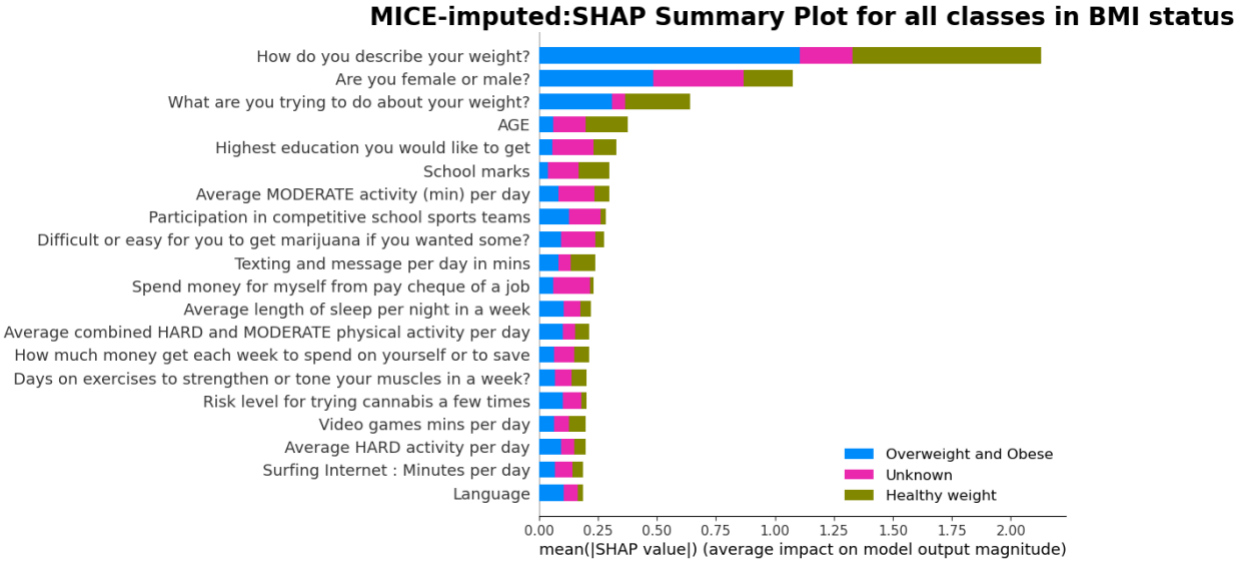
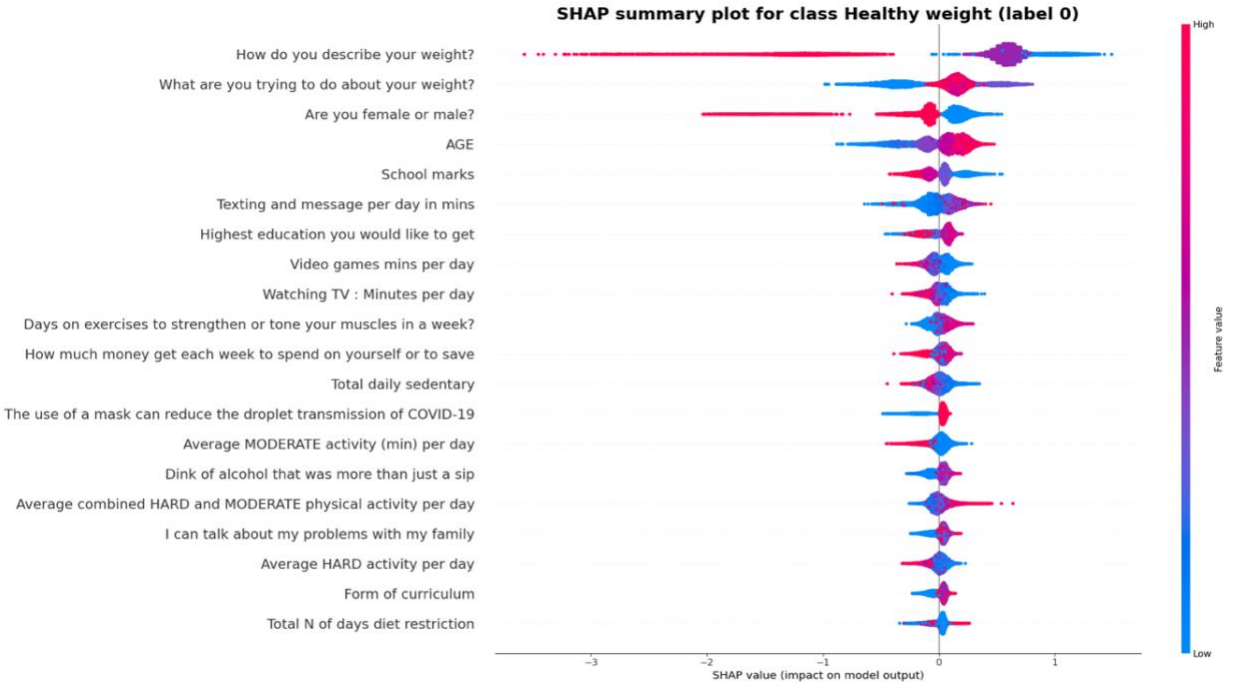
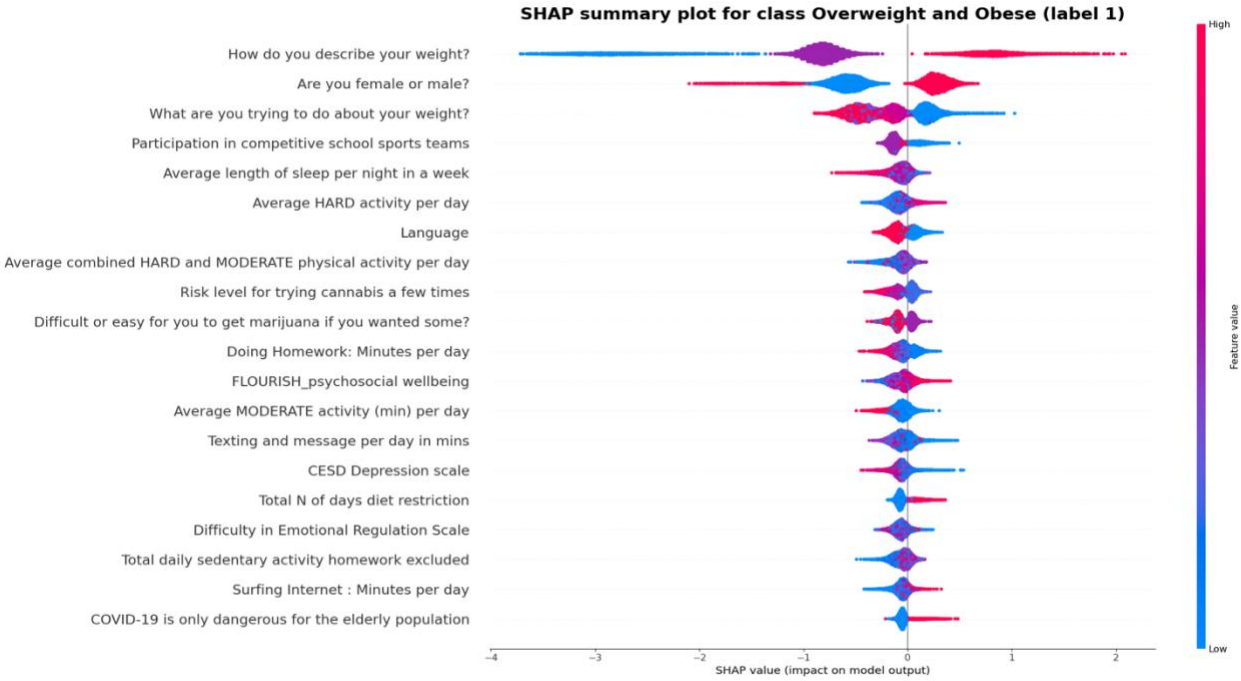


Figure 5.5 SHAP beewswarm plots for each BMI status in MI

(A) Healthy weight



(B) Overweight and obese



(C) Unknown

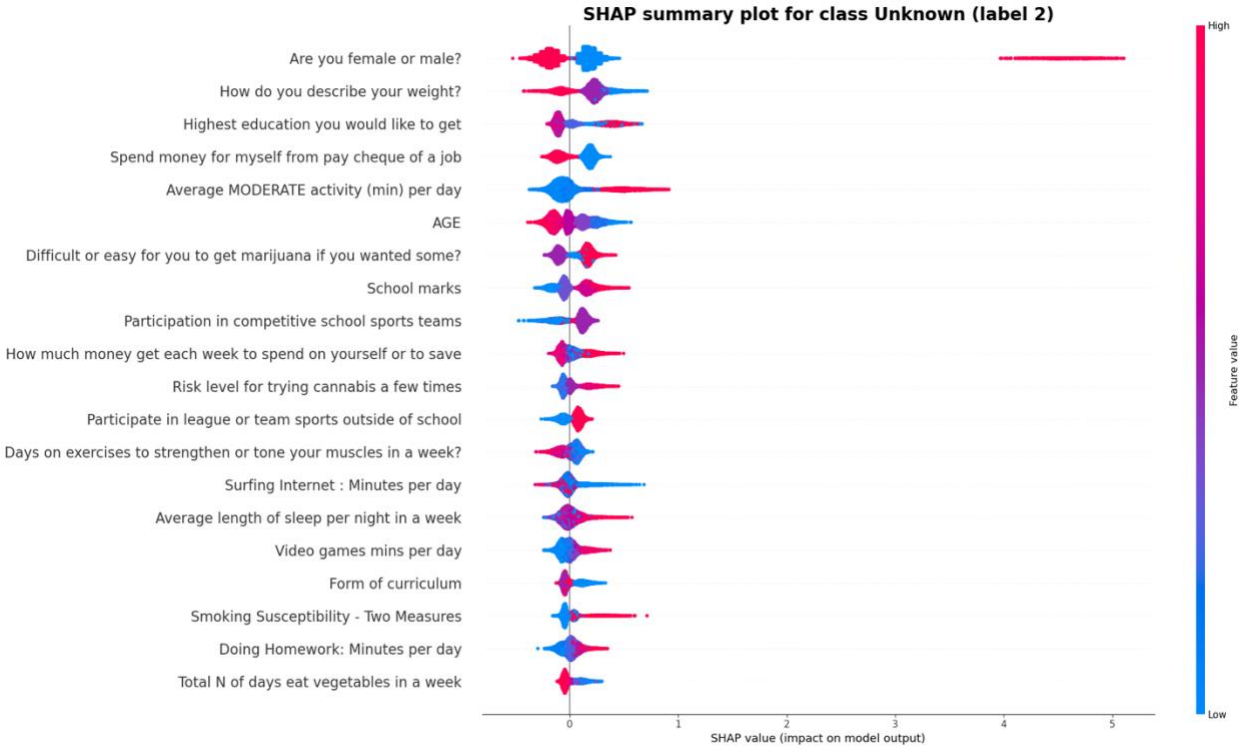
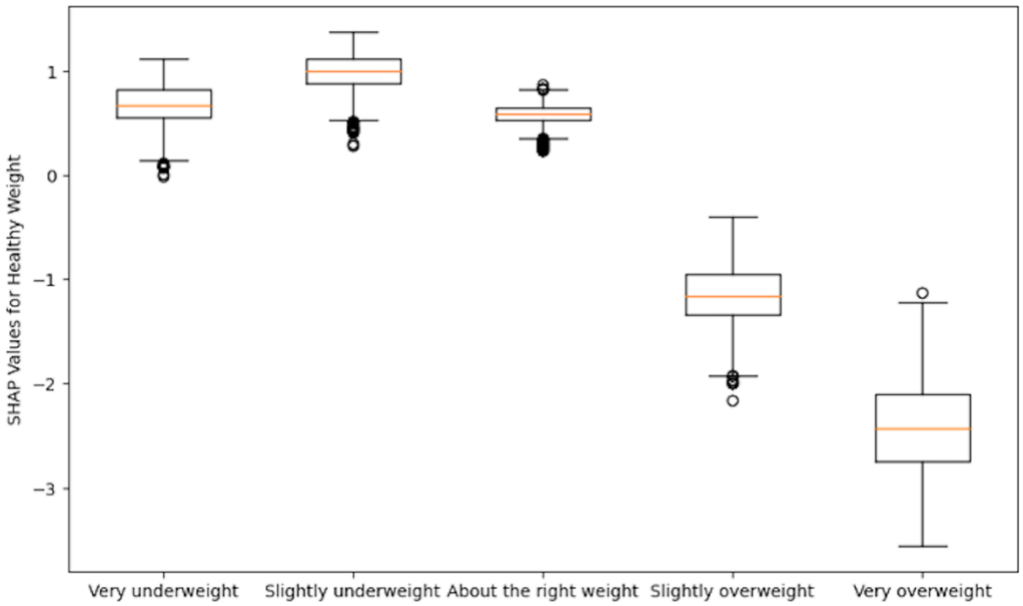
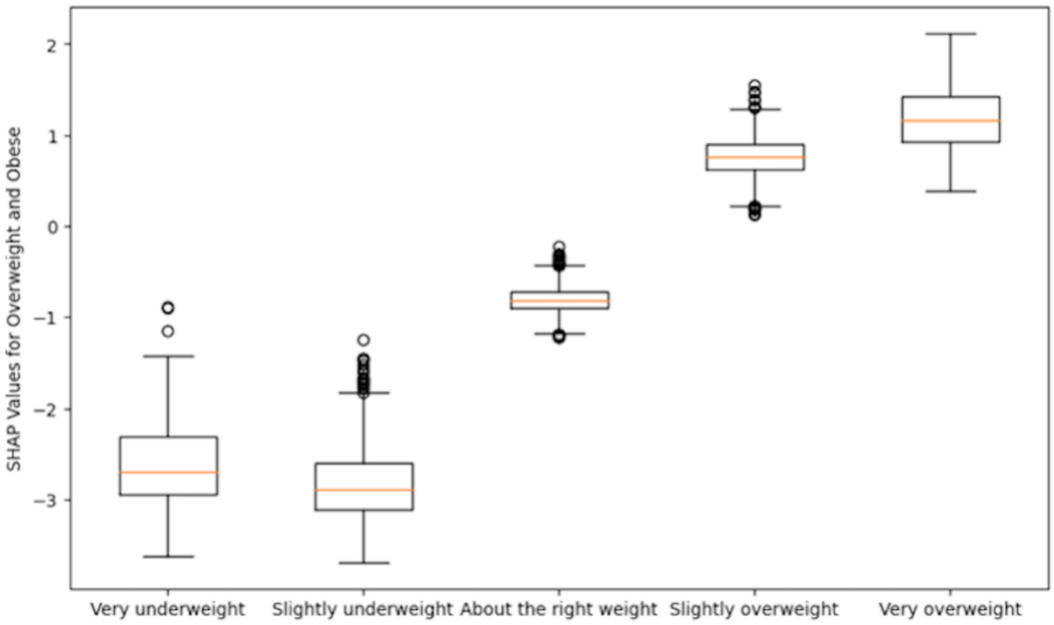

SHAP summary plot for class Unknown (label 2)
Are you female or male?
How do you describe your weight?
Highest education you would like to get
Spend money for myself from pay cheque of a job
Average MODERATE activity (min) per day
AGE
Difficult or easy for you to get marijuana if you wanted some?
School marks
Participation in competitive school sports teams
How much money get each week to spend on yourself or to save
Risk level for trying cannabis a few times
Participate in league or team sports outside of school
Days on exercises to strengthen or tone your muscles in a week?
Surfing Internet : Minutes per day
Average length of sleep per night in a week
Video games mins per day
Form of curriculum
Smoking Susceptibility - Two Measures
Doing Homework: Minutes per day
Total N of days eat vegetables in a week
0
1
2
3
4
5
SHAP value (impact on model output)
High
Feature value
Low

Figure 5.6 Boxplot of weight perception and gender SHAP values in each BMI status (MI)

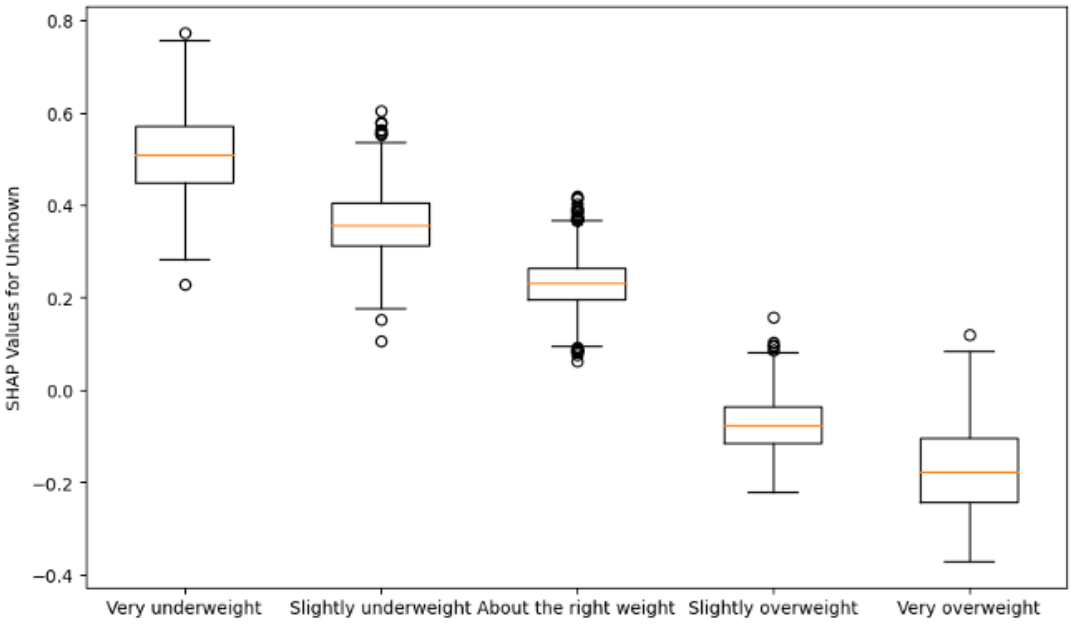
(A) Weight perception in healthy weight



(B) Weight perception in overweight and obese



(C) Weight perception in unknown



(D) Gender in unknown

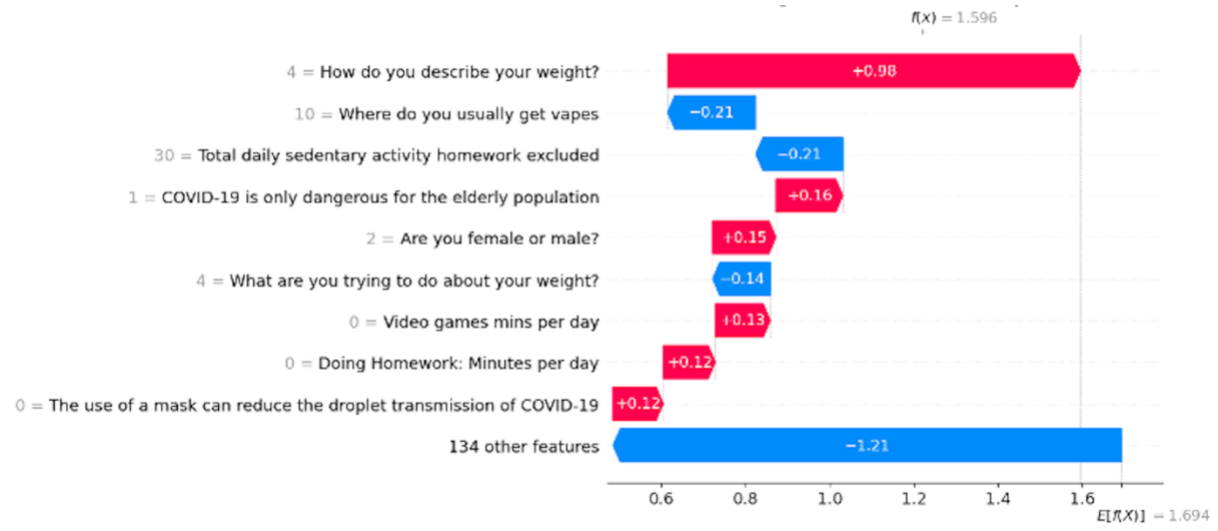


Figure 5.7 SHAP waterfall plots of each BMI status in MI

(A) A sample student from healthy weight



(B) A sample student from overweight and obese



(C) A sample student from unknown

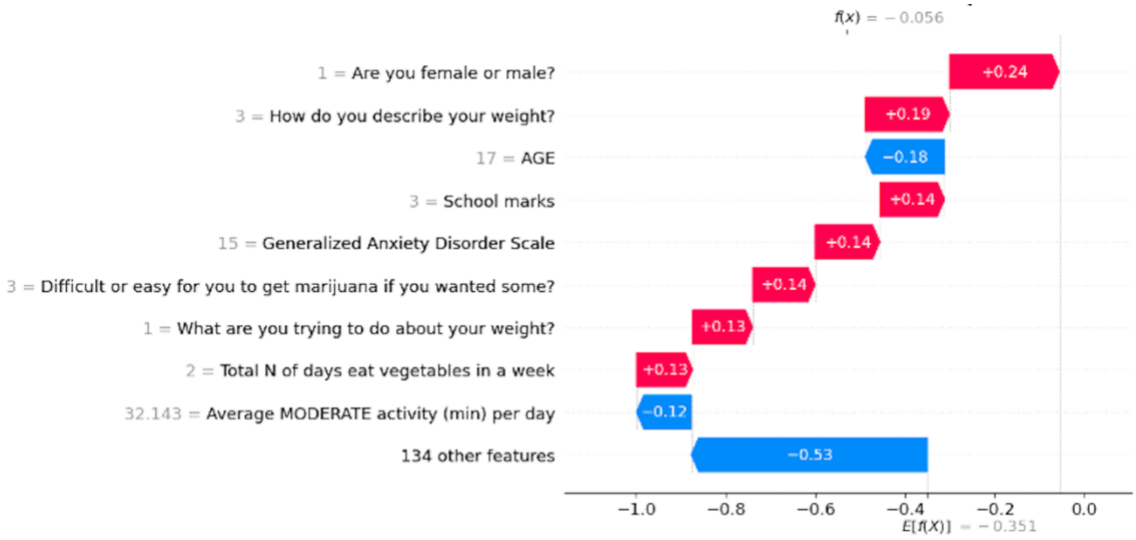
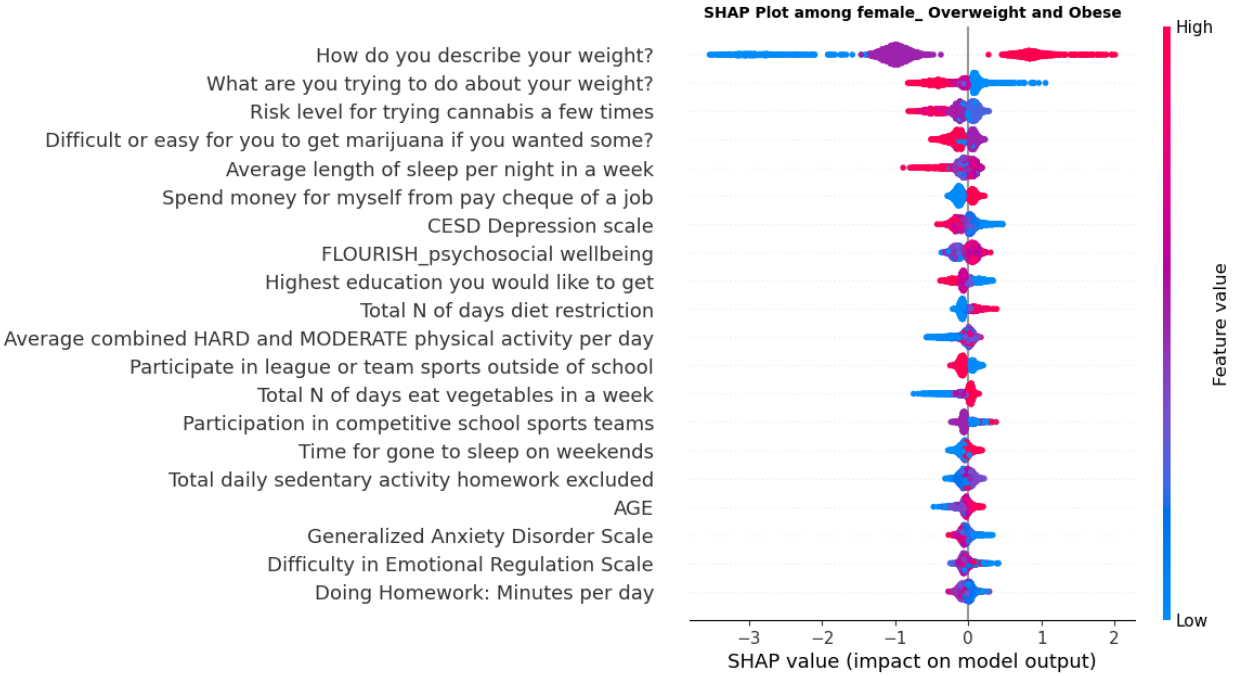


Figure 5.8 SHAP beeswarm plots of overweight and obese females vs males



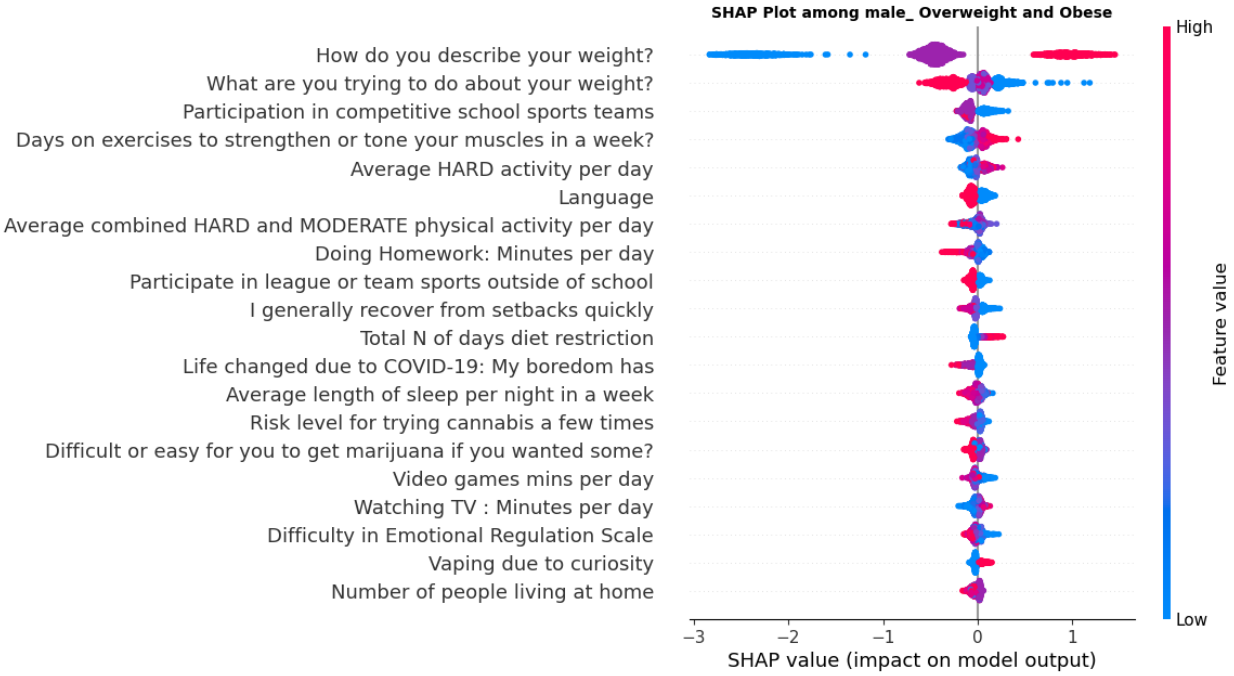
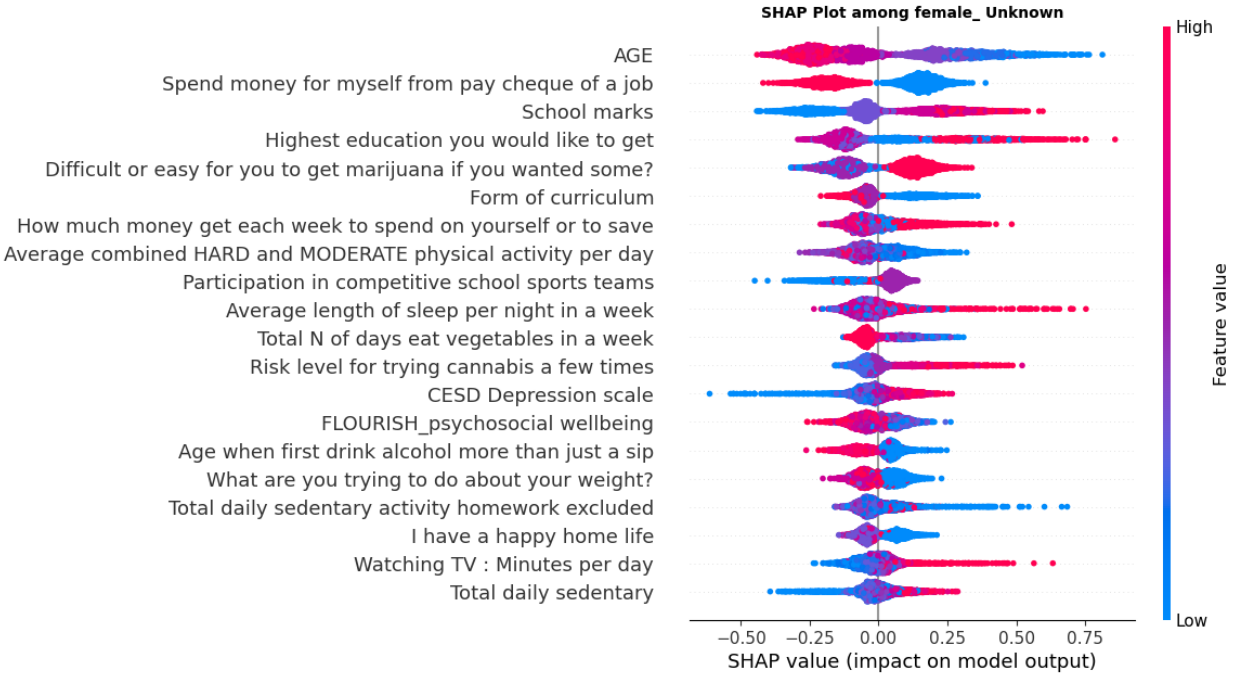
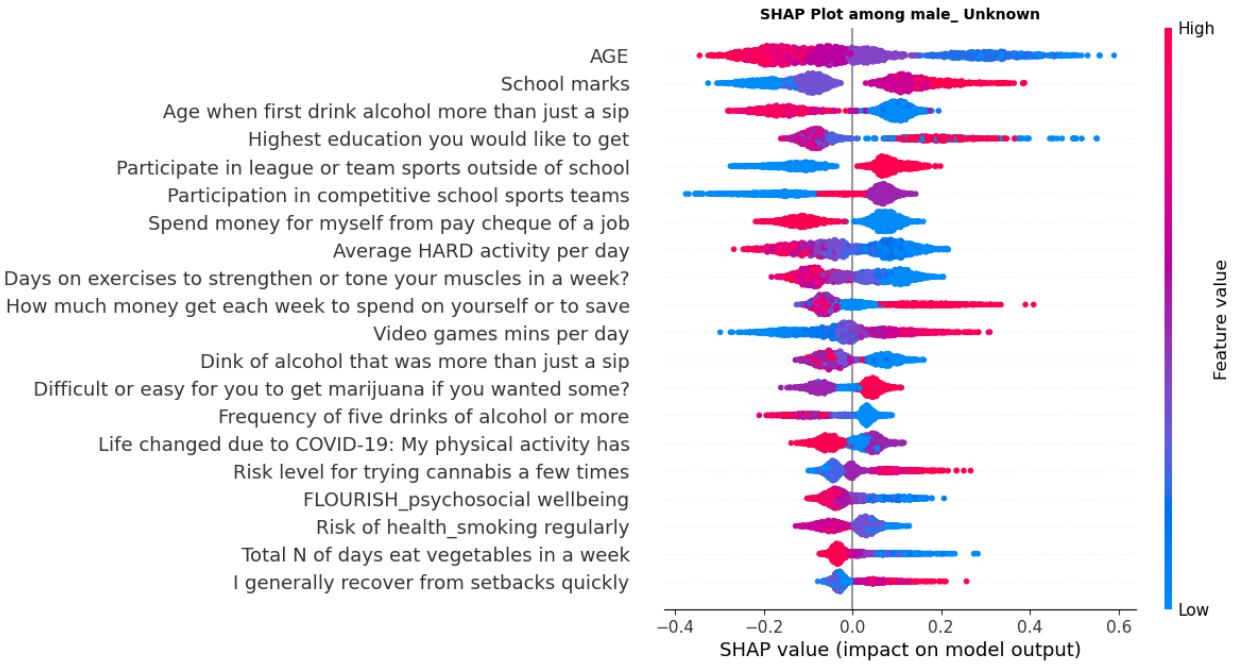


Figure 5.9 SHAP beeswarm plots of unknown weight females vs males





5.5 Discussion

In this study, we prepared three data processing strategies, LDA transformation, selecting the top 50 variables based on the feature importance, and using all variables as inputs for model training. We selected the top performer among the six classification candidates based on the F1 score for each strategy through cross-validation, followed by Optuna for hyperparameter tuning. For both CCA and MI approaches, training on all variables on XGBoost machine learning algorithms demonstrated higher F1 or ROC-AUC scores than SVM and RF. SVM was the top model for LDA-transformed inputs, and RF was the top one for the 50 features dataset. This study incorporated SHAP analysis and provided summary and waterfall plots for global and individual explanations to provide interpretable model prediction results. The summary plot provided insights into the degree of each factor's association with model prediction regarding the three BMI statuses. The top two significant variables were student weight perception and intention to change weight.

Based on the SHAP summary plot, it was clear that students who perceived themselves as slightly overweight or very overweight were classified into overweight and obese classes. When the individual profile contained perceiving as overweight and trying to lose weight, these students were more likely to be in the overweight and obese category. Adolescents who perceived themselves as overweight had a higher intention of participating in weight loss [162]. Previously reported results also showed that overweight and obese youth were more likely to try to lose weight than the non-overweight group [164]. Weight perception was also found to be significantly associated with weight gain. Self-identifying as overweight increased the likelihood of overeating, and this risk factor contributed to the weight increase and obstructed weight management [165].

In contrast, the prediction in the healthy weight student's profile included describing the weight as being about the right weight, not trying to do anything about their weight, and having an average of 90-100% of school marks. Additionally, females were more likely to be predicted in the healthy BMI group than males. Based on the stratified plots between females and males, males who exercised more were predicted to have overweight and obese BMI. These results align with the previous study. Regarding academic performance, healthy weight BMI children had a higher probability of reporting good grades than obese children [170]. While boys were reported to have a higher obesity prevalence than girls, studies also revealed that females, on average, tended to underreport their weight than males due to the influence of social media stigma on body image [166, 167, 168, 169]. This might explain why females were less likely to be predicted in the overweight and obese class from the SHAP summary plots. In comparison to females, playing sports and engaging in muscle training in males were more associated with

overweight and obese BMI prediction. It was also illustrated that males have the tendency and desire to build muscle and increase weight [174]. In a National Study of Adolescent to Adult Health, males (29.2%) had a much higher percentage of attempting weight gain relative to females (7.0%). Among the participants, 25.2% of males engaged in muscle-enhancing compared to only 3.8% of females [255].

It was also interesting to see that perceiving as very underweight and underweight most influenced the model prediction of the unknown class. Chau et al. [220] emphasized the linkage between reporting BMI refusal and health-related factors. In their study, participants who perceived themselves as underweight had 3.43 times increased odds of not reporting their BMI. In contrast, those feeling overweight had 1.77 times the odds of refusing to report after controlling for covariates [220]. Moreover, Guss et al. [172] illustrated that transgender youth had 2.4 times the odds (adjusted) of describing themselves as an underweight or healthy weight when they were overweight and obese compared to cisgender females. These findings aligned with the SHAP discovery on the relationship between weight perception of underweight and missing self-reported BMI.

Although why the gender minority group influenced the model prediction more towards unknown BMI than females and males is not confirmed, previous research has illustrated a significant relationship between missing BMI and poor body image and pointed out that transgender participants experienced higher levels of body dissatisfaction and negative self-perception due to social stigma on body image than cisgender populations [173, 174, 175]. These critical factors need to be considered to analyze the missing self-reported weight. Future studies may need to focus more on comparing weight perception and weight change intention among different gender groups.

This study has several limitations. First, there was a lack of representative risk factor data for predicting childhood obesity. COMPASS is a comprehensive youth dataset with multiple domains, including substance use, mental health, school connectedness, sleep, and physical activities. However, it does not contain factors highly associated with overweight and obesity, such as parental data. Parental data is an important indicator that affects the formation of childhood obesity [26, 27]. Studies that achieved superior model performance in classifying youth BMI include parental education level, children's medical history, family history of being overweight, and prenatal information [143, 195, 196, 197]. Second, the weight-related variables were self-reported, and discrepancies in height and weight have been noted in this reporting [191, 192]. Other studies also emphasized the misperception of weight. For instance, males were more likely to overreport their height, while females were more likely to underweight their weight [193]. Even though our study results showed that weight perception was the most significant variable in BMI prediction, discrepancies in self-identification could create

difficulties in developing targeted interventions to prevent youth obesity. Even though stratification was performed between females and males, factors associated with BMI prediction among gender minorities were unknown due to the extremely small sample size. Future wok should also consider conducting subgroup analysis on ethnicity to address potential different implications. Lastly, the model's generalizability is limited to students who participated in the COMPASS study, and the result may not be practical in other provinces or countries. Future studies on other youth populations are recommended to draw additional insights into other factors associated with youth BMI status.

5.6 Conclusion

This study utilized machine learning classifiers to predict childhood obesity (cohort 2020-21) and provided SHAP interpretation to explain the critical factors associated with the model's prediction of youth BMI. The results of this study disseminated essential implications for future research and obesity prevention initiatives, primarily focusing on addressing youth weight perception and weight change intention. This study encourages schools and policymakers to implement wellness support programs to reduce social stigma’s influence and improve student self-esteem and acceptance of body weight.

Chapter 6

Exploring the Longitudinal Weight Perception Transition and Associated Factors Using the Markov Chain Model and Deep Learning Approach

6.1 Abstract

Background

Weight perception has been found to be significantly related to weight status among youth. Misconception of weight status has resulted in extreme weight loss and other adverse behaviours. Examining the transitional probabilities of weight perception and the relevant factors is critical to understanding the influence on weight perception and promoting targeted interventions to reduce body dissatisfaction. However, limited research has been established on the weight perception transition over time and its associated factors. This study aims to identify the transition probabilities in weight perception over time and utilize deep learning models to predict weight perception and observe important variables associated with weight perception.

Methods

Using a three-year linked COMPASS dataset, this study applied the Markov Chain Model to observe the transition probabilities of weight perception. The multinomial logistic regression models assessed the significant factors in weight perception in Year 2 and the transition of weight perception from Year 1 to Year 2. To predict the weight perception in Year 3 and observe the temporal change of the associated factors, RNN and LSTM were implemented with SHAP analysis to interpret associated factors contributing to the model prediction.

Results

Most students were likely to maintain the same weight perception as the previous year. Students with higher BMI and an intention to lose weight were more likely to continue perceiving themselves as overweight or transition from perceiving themselves as being about the right weight to overweight. Conversely, students with healthy BMI who did not skip breakfast to lose weight were more likely to maintain an accurate weight perception. Additionally, muscle-strengthening exercises were found to have a protective effect by reducing the likelihood of maintaining an overweight perception.

Conclusions

This study utilized the Markov Chain Model and deep learning models to address the gaps in understanding weight perception transitions over time and the temporal effects of key factors associated with weight perception. Early intervention in schools, including programs advocating for accurate body image and improved self-esteem, is essential.

6.2 Introduction

Adolescence is a critical period marked by rapid physical development. This transformation could challenge body image acceptance and perception [221]. Body image is a multidimensional construct consisting of feelings, thoughts, and perceptions regarding an individual's own body and body of others [221, 222]. Weight perception was often represented as the subjective assessment of their bodies in terms of underweight weight, right weight, or overweight [162, 177]. Misconceptions about one's body can result from inaccurate evaluations of others [221]. Misidentification of weight status and body dissatisfaction is common among children and adolescents [223]. Weight misconception is known as individuals perceive their weight status differently than their true, objective weight status [224]. Study shows adolescents who are overweight and obese tend to underestimate their actual weight, and the ones with normal weight could overestimate their weight status. Females are also more likely to overestimate their weight status than males and engage in weight loss practices [162, 194, 225].

Several factors have been mentioned to influence adolescent's perception of weight. One of them is BMI, which is more connected with self-reported data [226]. Studies have shown that healthy weight or overweight teenagers who perceive their weight as overweight have a higher possibility of participating in weight loss activities. At the same time, there is a large proportion of teenagers who have a BMI overweight or at risk of overweight perceive themselves as underweight. They are also unlikely to engage in weight-loss activities such as dieting and physical exercises [226]. This discrepancy seems common among adolescents, who tend to overreport their height and underreport their weight [221]. In addition, inappropriate weight control practices are also linked to the under or overestimation of body weight. For instance, teenage boys are more likely to attempt gaining weight for an idealism in muscularity. As a result, healthy boys could also perceive themselves as overweight and intend to lose weight for a muscular body [227]. Dissatisfaction with the body happens more often among girls, who are more likely to participate in weight control practices. Some media and cultural norms advocated having a "thin" body, such as a V-shape upper body or lower waist-hip ratio [227, 228]. Due to the stigmatization of body image, individuals who perceive themselves as overweight engage more in the weight loss effort and could try extreme weight management strategies, including vomiting, skipping meals or fasting [182, 228].

On the other hand, the perception of the right weight positively affected the youth as they experienced fewer depressive symptoms longitudinally than overweight or underweight

perceptions [177, 229, 230]. Adolescents who perceive themselves at the right weight are more likely to engage in physical activities such as resistance training and school sports participation than those who perceive themselves as overweight [230]. Research also discovered that male and female adolescents with lower body dissatisfaction were active in sports [231, 232]. Based on the distorted effects resulting from misperception of weight status, it is necessary to develop targeted public health strategies to promote accurate weight perception among youth.

To account for the transition status, the Markov model has been utilized in previous childhood obesity studies. Hu et al. [150] and Chen et al. [149] implemented the Markov model to observe the transition probabilities and patterns of childhood weight status. They discovered that youth originally were normal weight or obese intended to maintain their first-stage weight status. Additionally, Tang et al. [151] used the multistate Markov model in a longitudinal study and reported dynamic change in weight status among young people. While many studies focus on the association of weight perception with risk factors cross-sectionally or discover the dynamic change in weight status, few research has investigated the transition pattern of weight perception and the significant factors over time.

Besides observing the transition probability using the Markov model, limited research has used time series machine learning methods to predict weight perception among youth besides childhood obesity. The time series classification predictive model uses longitudinal or sequential inputs across time on a target variable. Deep learning models are known to have the capability of capturing the time variation among the variables compared to the traditional classifiers, including random forest, support vector machine or logistic regression [234]. Several studies from the healthcare domain utilized time-series deep learning models such as RNN (Recurrent Neural Network) and long short-term memory (LSTM) in EHR obesity research [134, 235, 236, 237, 238]. For instance, Gupta et al. [134] used the LSTM network to train the model and predict obesity between 3 and 20 years old based on the previous year's EHR data. Additionally, a study utilized both LSTM and RNN to identify the keywords related to obesity weight status before and after COVID-19 from the national health and nutrition examination survey in Korea [236]. Based on the previous studies, RNN and LSTM have been widely applied for time series modelling, proved promising results, and were appropriate for capturing the changes of the outcome of interest longitudinally.

To the best of current knowledge, the deep learning model has not been utilized in weight perception research. Investigating the change in weight perception and the related factors could aid further intervention to focus on the important factors influencing weight perception. The accuracy of self-perception of weight status is critical for weight management, including nutritional and physical activity perspectives to reduce youth obesity [239]. The objectives of the studies aim to 1) explore the transition probabilities in weight perception over time using the

Markov Chain Model and identify the significant factors associated with weight perception through multinomial logistic regression and 2) build deep learning models (RNN and attention-based LSTM) to predict the Year 3 weight perception among a three-year linked COMPASS study.

6.3 Materials and Methods

6.3.1 Study Cohort

The three years linked datasets were Year 1 (2020-21), Year 2 (2021-22), and Year 3 (2022-23). These datasets came from the COMPASS data system. The COMPASS study is a prospective ongoing study that collects grade 9-12 students attending secondary schools in Alberta, British Columbia, Ontario, and Quebec [48]. The COMPASS study (2012-2027) collects student-level through the anonymous, self-administered COMPASS Student Questionnaire (C_q) annually on assessing healthy eating, sedentary behaviour, alcohol, tobacco, and marijuana use, bullying, academic achievement, physical activity, obesity, and school connectedness [48] The participating students were permitted to use active information, passive-consent, and parental permission protocols. [48] Details about the dataset can be found in print or on the website (www.compass.uwaterloo.ca) [49].

6.3.2 Data Cleaning

In this study, the dataset is a three-year linked longitudinal dataset, represented by Year 1 (2020-21), Year 2 (2021-22), and Year 3 (2022-23). The questions within the versions as "Appears at the end of the survey in Quebec", "EMMOWI Pilot version and Optional School Climate Module Only", "Non-Quebec Only", "Non-Quebec Only – Consolidated Responses", "Optional School Climate Module Only", and "Quebec Only" were excluded from the initial variable selection. The redundant variable refers to either questions being sub-sectional from a previous question or questions representing a similar perspective. For example, instead of the sub-sectional question, "Have you used a vape for any of the following reasons? (Mark all that apply) Curiosity / to try something new", the question "Have you ever tried a vape, also known as an e-cigarette? (e.g., JUUL, Vype, Suorin, Smok)" were included. In addition, the variables that create the derived variables were dropped from the dataset. For example, "Video Games: Minutes per day" was derived from the two questions, "How much time per day do you usually spend doing the following activities? Playing video/computer games; Hours" and "How much time per day do you usually spend doing the following activities? Playing video/computer games; Minutes".

6.3.3 Feature Selection for Predictors

The feature selection process was applied to choose the most relevant variables for weight perception. Using feature selection improves the predictive model's performance by including relevant features [189]. Gradient boosting tree models such as XGBoost have shown the capability to capture complex interactions based on their feature importance scores compared to traditional tree-based models [240]. XGBoost classifier was implemented in each year's cohort. F1 score and ROC AUC performance metrics were used to determine the number of features to be included. After determining the number of features, the variables were ranked based on their feature importance for each chosen number. The variables were retained as long as they appeared as an important feature in one of the years. After the feature selection, predictors came from various domains, including healthy eating, weight-related, physical activity, screen time, sleep, COVID-related, school engagement and substance use. The Appendix for Study 3 summarized the detailed feature selection and the final variables list. The data preprocessing on collapsing the predictor variables and the outcomes can be found in the data transformation section (Appendix D). Below are the representative questions from each domain.

In the physical activity category, questions included "On how many days in the last 7 days did you do exercises to strengthen or tone your muscles? (e.g., push-ups, sit-ups, or weight-training)", and "Do you participate in before-school, noon hour, or after school physical activities organized by your school? (e.g., intramurals, non-competitive clubs)". The diet-related variables include the aggregated variables such as the total number of days of eating, snacks, drinking water, fast food purchases, eating fruit, and restricting diet to lose weight. The mental health category included the Generalized Anxiety Disorder 7-item Scale (GAD7 [54]), self-reported 10-item Center for Epidemiologic Studies Depression Revised (CESD-R10 [216]), and Flourishing [55]. The COVID-related question asked, "How has your life changed because of COVID-19?" in terms of anxiety, physical activity, time spent on social media, vaping, cannabis use, loneliness, sleep and alcohol use. School engagement was assessed by the average school marks, bullying, and school connectedness scores. Lastly, the substance use question asked about the frequency of use of vaping, smoking, alcohol and cannabis.

The outcome weight perception was collapsed into three categories: slightly underweight/underweight, about the right weight, and slightly overweight/overweight. The other two weight-related factors were BMI and weight change intention: "Which of the following are you trying to do about your weight?" For BMI, the missing was categorized as a separate category (unknown).

6.3.4 Data Preprocessing

Missing data for weight perception was removed since it was only 5.7%. The remaining missing data was handled using Multivariate Imputation by Chained Equations (MI) in R studio 2023. The outcome was included in the MICE imputation process but was instructed to leave out of prediction. Gender and ethnicity were not provided in year 1 (2020-21), so the gender and ethnicity from subsequent years were used to fill in the missing data. The prediction matrix was created in R to indicate the imputation method for each variable [241]. The predictive meaning matching was used for continuous variables. Nominal categorical variables (unordered) proceeded with polynomial regression. Logistic regression was applied for dichotomous variables, and the ordinal categorical variables (ordered) response was imputed using the proportional odds model. Since the highest missing percentage for the predictor was 9.7%, 10 imputed datasets were chosen with 50 iterations. Details of missing percentages can be found in Appendix D. For selecting the imputed dataset, the mean differences between the imputed dataset and the complete case dataset were compared for the continuous variables. The imputed dataset with the smallest mean difference was selected. Detailed MI imputation information is provided in the MI section (Appendix D).

6.3.5 Markov Chain Model (MCM) and Statistical Analysis

Markov chains are commonly used to observe the state-to-state transition probabilities across subjects. This model has been applied to evaluate treatment programs for describing the chronic illness's progression through discrete-time homogeneous Markov chains [145]. It modelled the health state at the end of each time cycle without considering the progression between cycles [145]. Our study also focused on the discrete-time Markov model of the transition probabilities for the weight perception state across three years without the explanatory factors. The three states were underweight, about the right weight, and overweight.

In this study, the probability of the student's weight perception in the next year (X_{n+1}) depends on their current state X_n. For instance, if a student was classified as 'about the right weight' in Year 1, their weight perception in Year 2 is determined only based on their weight perception status in Year 1, represented by the following:

$$P\ (X_{n+1} = 'Underweight' \mid X_n =' \text{Overweight}')$$

$$P\ (X_{n+1} =' \textit{About the right weight}' \mid X_n =' \textit{About the right weight}')$$

$$P\ (X_{n+1} =' \textit{Overweight}' \mid X_n =' \text{Overweight}')$$

To visualize the transition probability of weight perception from Year 1 to Year 2 and Year 2 to year 3, the Graphviz graphing package was implemented in Python to represent the weight status transition probability over the year.

To evaluate the association between weight perception and the potential significant predictors, two multinomial logistic regression models were performed and adjusted for the demographic variable, including gender and ethnicity. The first model observed the association between the Year 1 predictor and Year 2 weight perceptions. In contrast, the second model investigated the Year 1 predictors associated with the transition of weight perception from Year 1 to Year 2. Since the COVID-19 pandemic variables only existed in Year 1 and Year 2, the statistical analysis for finding the associated factors focused on the weight perception transition group from Year 1 to Year 2.

The transition of weight perception from Year 1 to Year 2 was represented as nine groups. Group 1 = Remained underweight, Group 2 = Underweight to about the right weight, Group 3 = Underweight to overweight, Group 3 = Underweight to overweight, Group 4 = Remained about the right weight, Group 5 = About the right weight to underweight, Group 6 = About the right weight to overweight, Group 7 = Remained overweight, Group 8 = Overweight to underweight, and Group 9 = Overweight to about the right weight. The analysis explored the Year 1 predictors associated with weight perception changes from Year 1 to Year 2. The odds ratio and 95% confidence interval (CI) were reported to interpret the strength of the association between the significant predictors in Year 1, weight transition status in Year 2, and the weight transition from Year 1 to Year 2. The significance level for this research was set at $p<0.05$. The statistical analyses were conducted in the SAS studio environment (Enterprise Edition).

6.3.6 Deep Learning Data Preprocessing and Architecture

The Year 1 and Year 2 variables were used as the predictors, and Year 3 weight perception was used as the primary outcome. The demographic variables were not included in the model analysis due to the interest in time-variant features. Normalization was applied to the continuous variable to be more balanced and reduce the effect of any outliers in the dataset. The dataset was randomly split into 70% of students into training and 30% of them into testing datasets. The RNN and LSTM resampling was not straightforward because each two years of records per student must be retained. SMOTE resampling methods randomly increased the number of samplings in the minority group. Using the resampling could disturb the structure of the longitudinal data format. To address the imbalance in weight perception, the neural network calculated the class weight of each weight perception class, assigning more weight to the minority class than the majority class during model training. All analyses were conducted in Python (Version 3.10.13). The neural network models in this study used Tensorflow.keras (Version 2.10.0).

Recurrent Neural Network (RNN)

RNN is used for time series as a type of neural network [75]. The input layer receives and processes the information, and the hidden layer stores the memory based on the previous inputs to make predictions for the next sequence. The output layer would provide the result from the prediction [133]. A basic structure of RNN is shown in Figure 6.1 [140]. A simple RNN model was built as a benchmark to the LSTM model, creating one RNN recurrent hidden layer with experiments ranging from one to four for the subsequent dense layer. The proposed model was tested through iteration over the number of dense layers in the range. This testing was necessary to apply the ideal dense layers to the model development. The optimal number of dense layers was determined by the ones that achieved the highest ROC-AUC score. ROC-AUC was reasonable to apply due to the imbalanced numbers of samples in weight perception. Each dense layer was followed by a dropout layer to prevent model overfitting.

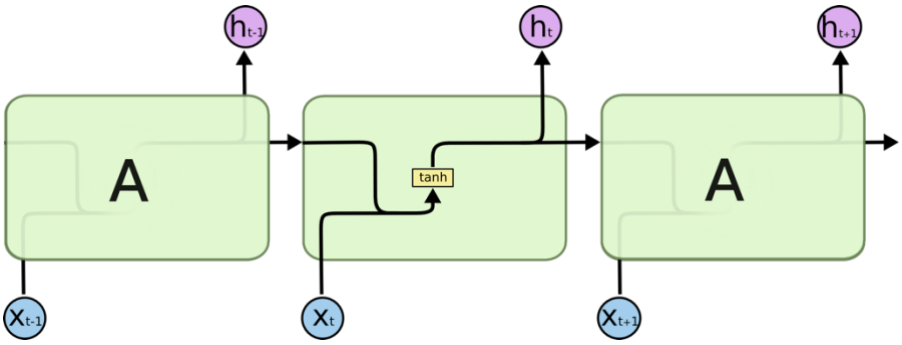


Figure 6.1 RNN structure

Long Short-Term Memory (LSTM)

LSTM is a more advanced example of an RNN designed to overcome the vanishing gradient issue. An illustration of LSTM is shown in Figure 2. Unlike RNN, the LSTM consists of a cell state, a forget gate, an input gate, and an output gate [137]. The forget gate from the LSTM allows the model to remove information, determine how much input information to keep, and decide the degree of total information for the output [137, 140]. Figure 6.2 shows the basic structure of LSTM [140].

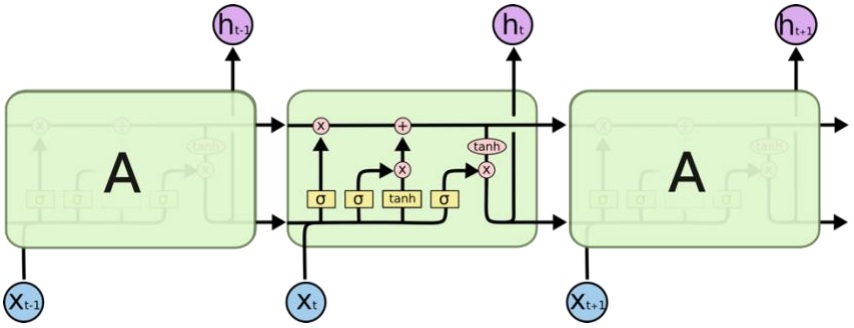


Figure 6.2 LSTM structure

A range of one to four LSTM and dense layers was experimented with to construct the LSTM model. Each configuration's performance was also evaluated using the ROC-AUC score. In addition to the LSTM layer and dense layer, the attention mechanism was applied after the LSTM layer to improve the model's prediction. Several studies have implemented the attention layer to improve the model performance and interpretability [139, 141, 242]. Adding the attention layer allows the model to pay more attention to inputs by assigning higher weights to the more important features [242]. Additionally, our model utilized the flatten layer to convert the multi-dimensional output from the attention layer into a 1D dimension for the dense layer to process. Adding this layer was essential for analyzing the temporal patterns by transforming the multi-dimensional form [243]. The final LSTM model structure contained the LSTM layers, determined by the ones with the highest mean ROC-AUC score: the attention layer, the flatten layer, dense layers with dropout layers, and the output layer.

6.3.7 Hyperparameters Tuning and Model Evaluation

Both the RNN and LSTM models were optimized through Bayesian Optimization to find the optimal parameters for the network. Bayesian Optimization is known for its efficacy and ability to handle complex functions of the objects compared to random search or grid search. This method is also suitable for deep learning models [244]. The optimal parameters for RNN and LSTM can be found in Appendix D. The Bayesian optimization process evaluates the above range of parameters based on the mean Area Under the Curve (AUC) score as the evaluation metrics through 10-fold cross-validation. This provided the internal validation by training the model in k-1 folds and the rest of the fold as validation. The Bayesian optimization process used ten iterations to search for the best hyperparameters.

RNN and LSTM were trained using the optimal hyperparameters obtained from the Bayesian optimization and trained on the testing dataset. In addition to implementing the optimal parameter, the early stopping and the call-back method were employed in the model training. For

both model training in RNN and LSTM, 20% of the data from the training dataset was withheld as the validation set. Both methods served the purpose of improving model performance while avoiding overfitting. Early stopping is a technique that allows for specifying the number of epochs and stops training when the model performance, AUC, stops improving on the validation set [245]. The model checkpoint callback was applied to save the best model observed during training. We tested the patience value in the 10, 20, 30, 40, and 50 range and saved the best model within each patience value for the testing dataset. A classification report was generated to evaluate the model performance, including accuracy, precision, recall, and F-1 score, along with the ROC-AUC for each weight perception class and the mean overall ROC-AUC.

6.3.8 Model Interpretation

The SHAP (SHapley Additive exPlanations) (version 0.42.1) in Python was applied to explain the model prediction and displayed the factors that associated with the model prediction on weight perception based on their importance. The SHAP value can quantify the input's contribution and magnitude to the outcome prediction [215]. The explainable analysis is useful for interpreting complex deep learning networks such as the LSTM model and explaining the model prediction [215, 246]. SHAP analysis for RNN and LSTM was implemented using DeepExplainer. The SHAP value was averaged for each time step to generate an overall contribution of factors associated with the prediction of weight perception in SHAP bar plots. To further display the magnitude of how each Year 1 or Year 2 factor contributed to the model prediction, the SHAP beeswarm plots were created to illustrate the factor associated with the model prediction on each weight perception in Year 3.

6.4 Results

6.4.1 Baseline Characteristics

Table 6.4 presents the cohort characteristics in the Year 1 cohort (2020-21) stratified by weight perception. The weight perception for most students was "About the right weight", followed by "Overweight", and the least number of students perceived as "Underweight". Among 63.3% of students perceived as "About the right weight" (N=3414), 44.6% had healthy weight BMI status, and most were non-smokers (96.3%) and had the highest average mean minutes (95.2) in combined hard and moderate physical activity per day. In contrast, among the 21% of overweight perception students (N=1141), 26.6% had overweight and obese BMI status. 95% of them were non-smokers and had the lowest mean average combined hard and moderate physical activity per day. Among 15.5% of underweight perception students (N=838), 8.1% of them have an underweight BMI status. The descriptive statistics for Year 2 and Year 3 cohorts can be viewed in Tables 6.5 and 6.6.

Table 6.4 Descriptive Statistics of Year 1 Cohort (2020-21) by Weight Perception

Year 1	Underweight N=838	About the right weight N=3414	Overweight N=1141	Total (N=5393)
		N (%) or Mean (SD)		
Gender				
Female	382 (45.6)	2025 (59.3)	633 (55.5)	3040 (56.4)
Male	440 (52.5)	1324 (38.8)	465 (40.8)	2229 (41.3)
Gender-Minority	16 (1.9)	65 (1.9)	43 (3.8)	124 (2.3)
Province				
AB	34 (4.1)	106 (3.1)	58 (5.1)	198 (3.7)
BC	53 (6.3)	132 (3.9)	65 (5.7)	250 (4.6)
ON	265 (31.6)	976 (28.6)	418 (36.6)	1659 (30.8)
QC	486 (58.0)	2200 (64.4)	600 (52.6)	3286 (60.9)
Ethnicity				
White	584 (69.7)	2549 (74.7)	781 (68.4)	3914 (72.6)
Non-White	254 (30.3)	865 (25.3)	360 (31.6)	1479 (27.4)
Weight-Related				
BMI Status				
Underweight	68 (8.1)	24 (0.7)	2 (0.2)	94 (1.7)
Healthy Weight	410 (48.9)	1523 (44.6)	205 (18.0)	2138 (39.6)
Overweight and Obese	5 (0.6)	213 (6.2)	304 (26.6)	522 (9.7)
Missing	355 (42.4)	1654 (48.4)	630 (55.2)	2639 (48.9)
Weight Change Intention				
Lose weight	50 (6.0)	812 (23.8)	867 (76.0)	1729 (32.1)
Gain weight	323 (38.5)	190 (5.6)	7 (0.6)	520 (9.6)
Stay the same	142 (16.9)	1025 (30.0)	99 (8.7)	1266 (23.5)
Not trying anything about the weight	323 (38.5)	1383 (40.5)	165 (14.5)	1871 (34.7)
Missing	0 (0.0)	4 (0.1)	3 (0.3)	7 (0.1)
Substance Use				
Marijuana Use				
Non User	758 (90.5)	3144 (92.1)	1003 (87.9)	4905 (91.0)
User	54 (6.4)	180 (5.3)	101 (8.9)	335 (6.2)
Missing	26 (3.1)	90 (2.6)	37 (3.2)	153 (2.8)

Smoking and Tobacco Use				
Non Smoker	807 (96.3)	3287 (96.3)	1084 (95.0)	5178 (96.0)
Smoker	13 (1.6)	52 (1.5)	24 (2.1)	89 (1.7)
Missing	18 (2.1)	75 (2.2)	33 (2.9)	126 (2.3)
Alcohol Consumption				
Non drinker	368 (43.9)	1532 (44.9)	468 (41.0)	2368 (43.9)
Drinker	446 (53.2)	1794 (52.5)	637 (55.8)	2877 (53.3)
Missing	24 (2.9)	88 (2.6)	36 (3.2)	148 (2.7)
School-Related				
School Mark				
80-100%	537 (64.1)	2183 (63.9)	669 (58.6)	3389 (62.8)
50-70%	242 (28.9)	1036 (30.3)	389 (34.1)	1667 (30.9)
Below 50%	20 (2.4)	54 (1.6)	24 (2.1)	98 (1.8)
Missing	39 (4.7)	141 (4.1)	59 (5.2)	239 (4.4)
School connectedness score				
Mean (SD)	19.2 (3.0)	19.4 (2.9)	18.3 (3.03)	19.2 (3.01)
Missing	46 (5.5)	142 (4.2)	61 (5.3)	249 (4.6)
Physical Activity				
Average Combined Hard and Moderate Physical Activity per Day				
Mean (SD)	95.0 (77.6)	95.2 (74.4)	87.0 (70.1)	93.4 (74.1)
Missing	39 (4.6)	172 (5.0)	69 (6.1)	280 (5.2)
Muscle-strengthening Training in the Last 7 Days				
Mean (SD)	2.4 (2.2)	2.7 (2.2)	2.2 (2.2)	2.5 (2.2)
Missing	19 (2.3)	90 (2.6)	39 (3.4)	148 (2.7)
Participation in School Physical				
Yes	222 (26.5)	971 (28.4)	249 (21.8)	1442 (26.7)
No	505 (60.3)	2034 (59.6)	718 (62.9)	3257 (60.4)
Non offered at school	92 (11.0)	311 (9.1)	130 (11.4)	533 (9.9)
Missing	19 (2.3)	98 (2.9)	44 (3.9)	161 (3.0)
Healthy Eating				
Diet restriction in the Past Week (Number of Days)				
Mean (SD)	0.9 (2.1)	1.1 (2.2)	2.6 (3.0)	1.4 (2.5)

Missing	33 (3.9)	119 (3.5)	48 (4.2)	200 (3.7)
Vegetable Consumption in the Past Week (Numbers of days)				
Mean (SD)	5.3 (2.4)	5.5 (2.3)	5.1 (2.5)	5.4 (2.4)
Missing	33 (3.9)	119 (3.5)	48 (4.2)	200 (3.7)
Fast Food Purchased in the Past Week (Numbers of Days)				
Mean (SD)	0.9 (1.1)	0.9 (1.1)	0.9 (1.1)	0.9 (1.1)
Missing	33 (3.9)	119 (3.5)	48 (4.2)	200 (3.7)
Mental Health				
Mental Health Wellbeing (FLOURISH)				
Mean (SD)	32.0 (5.7)	33.1 (5.4)	30.2 (6.2)	32.3 (5.7)
Missing	40 (4.8)	162 (4.7)	58 (5.1)	260 (4.8)
Anxiety (GAD7)				
Mean (SD)	6.8 (5.47)	6.0 (5.38)	8.0 (6.21)	6.6 (5.63)
Missing	63 (7.5)	221 (6.5)	67 (5.9)	351 (6.5)

Table 6.5 Descriptive Statistics of Year 2 Cohort (2021-22) by Weight Perception

Year 2	**Underweight N=916**	**About the right weight N=3310**	**Overweight N=1167**	**Total N=5393**
	N (%) or Mean (SD)			
Gender				
Female	375 (40.9)	1934 (58.4)	640 (54.8)	2949 (54.7)
Male	500 (54.6)	1267 (38.3)	445 (38.1)	2212 (41.0)
Gender-Minority	40 (4.4)	104 (3.1)	81 (6.9)	225 (4.2)
Missing	1 (0.1)	5 (0.2)	1 (0.1)	7 (0.1)
Province				
AB	40 (4.4)	94 (2.8)	64 (5.5)	198 (3.7)
BC	65 (7.1)	131 (4.0)	54 (4.6)	250 (4.6)
ON	313 (34.2)	902 (27.3)	444 (38.0)	1659 (30.8)
QC	498 (54.4)	2183 (66.0)	605 (51.8)	3286 (60.9)
Ethnicity				
White	663 (72.4)	2562 (77.4)	841 (72.1)	4066 (75.4)

Non-White	252 (27.5)	744 (22.5)	325 (27.8)	1321 (24.5)
Missing	1 (0.1)	4 (0.1)	1 (0.1)	6 (0.1)
Weight-Related				
BMI Status				
Underweight	64 (7.0)	21 (0.6)	0 (0.0)	85 (1.6)
Healthy Weight	557 (60.8)	1767 (53.4)	274 (23.5)	2598 (48.2)
Overweight and Obese	10 (1.1)	245 (7.4)	342 (29.3)	597 (11.1)
Missing	285 (31.1)	1277 (38.6)	551 (47.2)	2113 (39.2)
Weight Change Intention				
Lose weight	70 (7.6)	725 (21.9)	893 (76.5)	1688 (31.3)
Gain weight	426 (46.5)	301 (9.1)	14 (1.2)	741 (13.7)
Stay the same	100 (10.9)	1013 (30.6)	78 (6.7)	1191 (22.1)
Not trying anything about the weight	318 (34.7)	1267 (38.3)	181 (15.5)	1766 (32.7)
Missing	2 (0.2)	4 (0.1)	1 (0.1)	7 (0.1)
Substance Use				
Marijuana Use				
Non User	765 (83.5)	2852 (86.2)	942 (80.7)	4559 (84.5)
User	132 (14.4)	409 (12.4)	206 (17.7)	747 (13.9)
Missing	19 (2.1)	49 (1.5)	19 (1.6)	87 (1.6)
Smoking and Tobacco Use				
Non Smoker	865 (94.4)	3175 (95.9)	1105 (94.7)	5145 (95.4)
Smoker	37 (4.0)	93 (2.8)	46 (3.9)	176 (3.3)
Missing	14 (1.5)	42 (1.3)	16 (1.4)	72 (1.3)
Alcohol Consumption				
Non drinker	318 (34.7)	1059 (32.0)	342 (29.3)	1719 (31.9)
Drinker	581 (63.4)	2195 (66.3)	804 (68.9)	3580 (66.4)
Missing	17 (1.9)	56 (1.7)	21 (1.8)	94 (1.7)
School-Related				
School Mark				
80-100%	566 (61.8)	2129 (64.3)	681 (58.4)	3376 (62.6)
50-70%	299 (32.6)	1031 (31.1)	418 (35.8)	1748 (32.4)
Below 50%	22 (2.4)	60 (1.8)	30 (2.6)	112 (2.1)
Missing	29 (3.2)	90 (2.7)	38 (3.3)	157 (2.9)

School connectedness score				
Mean (SD)	18.3 (3.2)	18.9 (3.0)	17.7 (3.3)	18.5 (3.1)
Missing	29 (3.2)	92 (2.8)	41 (3.5)	162 (3.0)
Physical Activity				
Average Combined Hard and Moderate Physical Activity per Day				
Mean (SD)	95.3 (68.1)	99.4 (72.4)	87.0 (66.2)	96.1 (70.6)
Missing	44 (4.8)	115 (3.5)	64 (5.5)	223 (4.1)
Muscle-strengthening Training in the Last 7 Days				
Mean (SD)	2.4 (2.2)	2.5 (2.2)	2.2 (2.1)	2.5 (2.2)
Missing	21 (2.3)	59 (1.8)	33 (2.8)	113 (2.1)
Participation in School Physical				
Yes	332 (36.2)	1284 (38.8)	361 (30.9)	1977 (36.7)
No	535 (58.4)	1885 (56.9)	749 (64.2)	3169 (58.8)
Non offered at school	29 (3.2)	72 (2.2)	25 (2.1)	126 (2.3)
Missing	20 (2.2)	69 (2.1)	32 (2.7)	121 (2.2)
Healthy Eating				
Diet restriction in the Past Week (Number of Days)				
Mean (SD)	1.1 (2.3)	1.0 (2.2)	2.9 (3.1)	1.5 (2.6)
Missing	23 (2.5)	73 (2.2)	25 (2.1)	121 (2.2)
Vegetable Consumption in the Past Week (Numbers of days)				
Mean (SD)	5.1 (2.4)	5.4 (2.3)	5.1 (2.5)	5.3 (2.4)
Missing	23 (2.5)	73 (2.2)	25 (2.1)	121 (2.2)
Fast Food Purchased in the Past Week (Numbers of Days)				
Mean (SD)	1.2 (1.3)	1.1 (1.2)	1.2 (1.3)	1.1 (1.3)
Missing	23 (2.5)	73 (2.2)	25 (2.1)	121 (2.2)
Mental Health				
Mental Health Wellbeing (FLOURISH)				
Mean (SD)	31.0 (6.0)	32.6 (5.4)	29.1 (6.3)	31.6 (5.9)
Missing	30 (3.3)	120 (3.6)	42 (3.6)	192 (3.6)
Anxiety (GAD7)				
Mean (SD)	7.9 (6.0)	7.0 (5.7)	9.4 (6.5)	7.7 (6.0)
Missing	41 (4.5)	168 (5.1)	60 (5.1)	269 (5.0)

Table 6.6 Descriptive Statistics of Year 3 Cohort (2022-23) by Weight Perception

Year 3	Underweight N=924	About the right weight N=3183	Overweight N=1286	Total N=5393
		N (%) or Mean (SD)		
Sex				
Female	367 (39.7)	1861 (58.5)	775 (60.3)	3003 (55.7)
Male	537 (58.1)	1250 (39.3)	445 (34.6)	2232 (41.4)
Gender-Minority	20 (2.2)	72 (2.3)	65 (5.1)	157 (2.9)
Missing	0 (0.0)	0 (0.0)	1 (0.1)	1 (0.0)
Province				
AB	37 (4.0)	93 (2.9)	68 (5.3)	198 (3.7)
BC	62 (6.7)	135 (4.2)	53 (4.1)	250 (4.6)
ON	302 (32.7)	885 (27.8)	472 (36.7)	1659 (30.8)
QC	523 (56.6)	2070 (65.0)	693 (53.9)	3286 (60.9)
Ethnicity				
White	679 (73.5)	2516 (79.0)	937 (72.9)	4132 (76.6)
Non-White	243 (26.3)	662 (20.8)	347 (27.0)	1252 (23.2)
Missing	2 (0.2)	5 (0.2)	2 (0.2)	9 (0.2)
Weight-Related				
BMI Status				
Underweight	58 (6.3)	17 (0.5)	4 (0.3)	79 (1.5)
Healthy Weight	601 (65.0)	1978 (62.1)	317 (24.7)	2896 (53.7)
Overweight and Obese	18 (1.9)	234 (7.4)	406 (31.6)	658 (12.2)
Missing	247 (26.7)	954 (30.0)	559 (43.5)	1760 (32.6)
Weight Change Intention				
Lose weight	58 (6.3)	758 (23.8)	945 (73.5)	1761 (32.7)
Gain weight	456 (49.4)	458 (14.4)	20 (1.6)	934 (17.3)
Stay the same	102 (11.0)	928 (29.2)	95 (7.4)	1125 (20.9)
Not trying anything about the weight	308 (33.3)	1034 (32.5)	225 (17.5)	1567 (29.1)
Missing	0 (0.0)	5 (0.2)	1 (0.1)	6 (0.1)
Substance Use				
Marijuana Use				
Non User	703 (76.1)	2415 (75.9)	922 (71.7)	4040 (74.9)

User	203 (22.0)	687 (21.6)	340 (26.4)	1230 (22.8)
Missing	18 (1.9)	81 (2.5)	24 (1.9)	123 (2.3)
Smoking and Tobacco Use				
Non Smoker	855 (92.5)	3001 (94.3)	1188 (92.4)	5044 (93.5)
Smoker	53 (5.7)	123 (3.9)	80 (6.2)	256 (4.7)
Missing	16 (1.7)	59 (1.9)	18 (1.4)	93 (1.7)
Alcohol Consumption				
Non drinker	242 (26.2)	722 (22.7)	282 (21.9)	1246 (23.1)
Drinker	665 (72.0)	2392 (75.1)	984 (76.5)	4041 (74.9)
Missing	17 (1.8)	69 (2.2)	20 (1.6)	106 (2.0)
School-Related				
School Mark				
80-100%	516 (55.8)	1968 (61.8)	721 (56.1)	3205 (59.4)
50-70%	358 (38.7)	1075 (33.8)	485 (37.7)	1918 (35.6)
Below 50%	26 (2.8)	48 (1.5)	48 (3.7)	122 (2.3)
Missing	24 (2.6)	92 (2.9)	32 (2.5)	148 (2.7)
School connectedness score				
Mean (SD)	18.1 (3.3)	18.4 (3.1)	17.3 (3.3)	18.1 (3.2)
Missing	27 (2.9)	111 (3.5)	40 (3.1)	178 (3.3)
Physical Activity				
Average Combined Hard and Moderate Physical Activity per Day				
Mean (SD)	100.6 (77.9)	99.4 (69.9)	93.0 (75.4)	98.1 (72.7)
Missing	39 (4.2)	138 (4.3)	86 (6.7)	263 (4.9)
Muscle-strengthening training in the last 7 days				
Mean (SD)	2.4 (2.3)	2.6 (2.2)	2.1 (2.1)	2.5 (2.2)
Missing	20 (2.2)	84 (2.6)	36 (2.8)	140 (2.6)
Participation in School Physical				
Yes	312 (33.8)	1205 (37.9)	400 (31.1)	1917 (35.5)
No	565 (61.1)	1807 (56.8)	826 (64.2)	3198 (59.3)
Non offered at school	26 (2.8)	82 (2.6)	25 (1.9)	133 (2.5)
Missing	21 (2.3)	89 (2.8)	35 (2.7)	145 (2.7)
Healthy Eating				

Diet restriction in the Past Week (Number of Days)				
Mean (SD)	1.1 (2.3)	1.2 (2.4)	2.7 (3.0)	1.5 (2.6)
Missing	39 (4.2)	137 (4.3)	52 (4.0)	228 (4.2)
Vegetable Consumption in the Past Week (Numbers of days)				
Mean (SD)	4.9 (2.5)	5.4 (2.4)	5.0 (2.6)	5.2 (2.5)
Missing	39 (4.2)	137 (4.3)	52 (4.0)	228 (4.2)
Fast Food Purchased in the Past Week (Numbers of Days)				
Mean (SD)	1.6 (1.6)	1.3 (1.4)	1.5 (1.5)	1.4 (1.5)
Missing	39 (4.2)	137 (4.3)	52 (4.0)	228 (4.2)
Mental Health				
Mental Health Wellbeing (FLOURISH)				
Mean (SD)	30.8 (6.0)	32.1 (5.7)	28.8 (6.5)	31.1 (6.1)
Missing	48 (5.2)	177 (5.6)	67 (5.2)	292 (5.4)
Anxiety (GAD7)				
Mean (SD)	7.4 (5.9)	6.9 (5.6)	9.3 (6.3)	7.5 (5.9)
Missing	48 (5.2)	166 (5.2)	64 (5.0)	278 (5.2)

6.4.2 Transition Probabilities in Weight Perception

The transition matrix showed the likelihood of students transitioning from one weight perception state to another across the observed time. As this was a discrete-time Markov chain, the probability of moving from one state of weight perception only depended on the present state. Table 6.7 and Figure 6.3 show the weight transition probabilities from Year 1 to Year 2. Most students tended to remain unchanged in their weight perception over time. Students who perceived themselves as underweight had a 63.24% probability of remaining underweight in Year 2, those who perceived themselves as about the right weight had an 80.11% probability of staying at about the right weight in Year 2, and a 71.08% probability of remaining overweight in Year 2 for those who perceived as overweight.

Besides the stable state, there was a moderate probability of transitioning between weight perceptions, specifically from underweight to about the right weight and from overweight to about the right weight. Students who perceived themselves as underweight had a 33.53% chance of transitioning to about the right weight in Year 2, while those who perceived themselves as overweight had a 25.77% probability of transitioning to about the right weight in Year 2.

Additionally, the probability of transitioning from overweight to underweight (3.16%) and underweight to overweight (3.22%) was negligible. The transition probabilities from Year 2 to Year 3 followed the same patterns as those from Year 1 to Year 2. To visualize the transition probabilities from Year 2 to Year 3, please refer to Table 6.8 and Figure 6.4.

Table 6.7 Markov Chain Transition Probabilities of Weight Perception from Year 1 to Year 2

	Year 2 Weight perception		
Year 1 Weight perception	Underweight	About the right weight	Overweight
Underweight	**0.632458**	0.335322	0.032220
About the right weight	0.102519	**0.801113**	0.096368
Overweight	0.031551	0.257669	**0.710780**

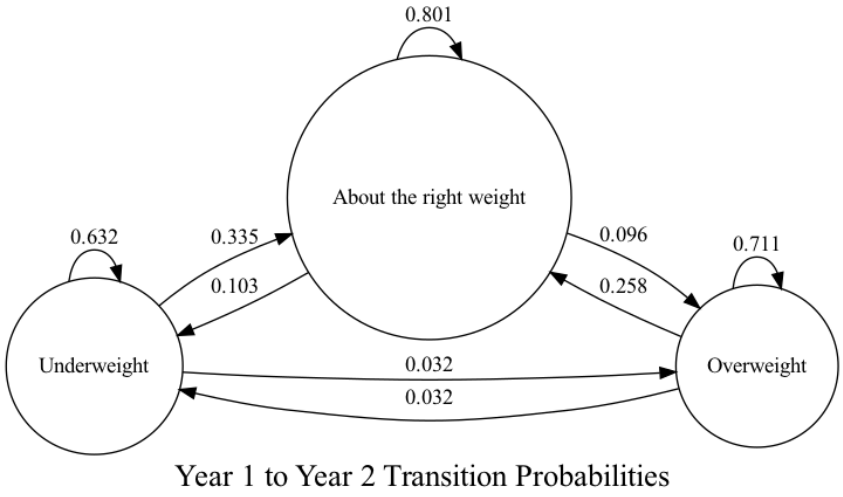


Figure 6.3 Markov Chain Transition Probabilities of Weight Perception from Year 1 to Year 2

Table 6.8 Markov Chain Transition Probabilities of Weight Perception from Year 2 to Year 3

	Year 3 Weight perception		
Year 2 Weight perception	Underweight	About the right weight	Overweight
Underweight	**0.648472**	0.317686	0.033843
About the right weight	0.090332	**0.800000**	0.109668
Overweight	0.026564	0.209083	**0.764353**

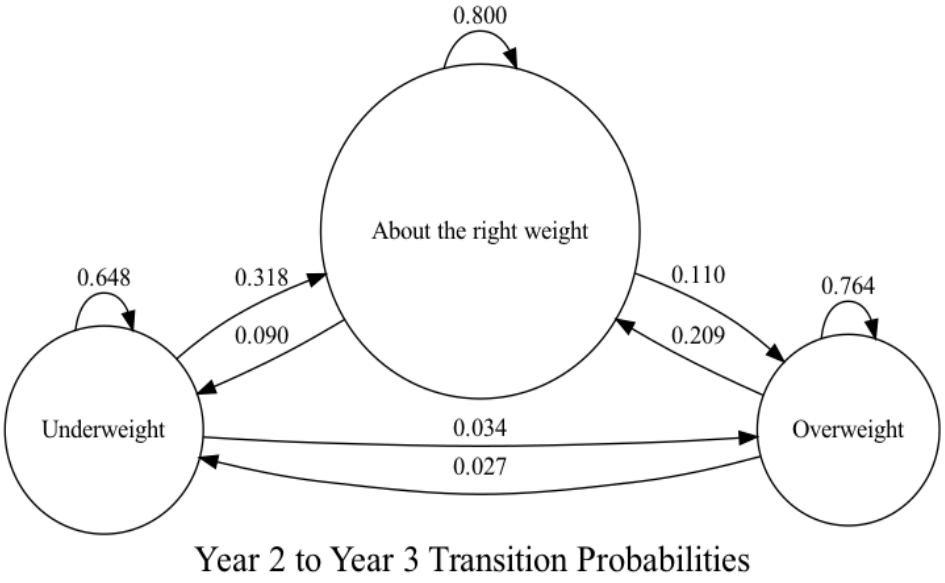


Figure 6.4 Markov Chain Transition Probabilities of Weight Perception from Year 2 to Year 3

6.4.3 The Association Between Year 1 Predictors and Weight Perception in Year 2

Table 6.9 shows the association between Year 1 predictors and Year 2 weight perception, using about the right weight as the reference group. The significant variables were gender, BMI, weight change intention, COVID-related, mental health, and physical activity.

Compared to male students, female students were less likely to be perceived as underweight (OR = 0.63, 95% CI: 0.51 - 0.78) and overweight (OR = 0.72, 95% CI 0.58 - 0.89). Students with an underweight BMI were more likely to perceive themselves as underweight compared to those with a healthy weight (OR = 4.9, 95% CI = 2.98 - 8.11), while overweight and obese BMI (OR = 6.66, 95% CI: 5.13 - 8.65) and unknown BMI (OR = 2.62, 95% CI: 2.15 -

3.19) students had higher odds of perceiving themselves as overweight. In addition to BMI, relative to students who were not trying to do anything about their weight, students who wanted to keep the same weight had decreased odds of underweight perception (OR = 0.63, 95% CI: 0.51 - 0.78). In contrast, students who tried to gain weight were more likely to perceive themselves as underweight (OR = 3.81, 95% CI: 3.01 - 4.83), while students who tried to lose weight had a higher likelihood of having an overweight perception (OR = 5.26, 95% CI: 4.23 - 6.53).

Decreasing time spent watching TV/movies or playing video games as life changed due to COVID-19 was positively associated with overweight perception. Students who had decreased time had higher odds of having overweight perception than those spending the same time (OR = 2.27, 95% CI = 1.39 - 3.70). Students who disagreed that they could discuss their problems with their families had higher odds of perceiving themselves as underweight (OR = 1.51, 95% CI: 1.15 - 1.98). Lastly, days in muscle training were positively associated with having about the right weight perception and negatively associated with having an overweight perception. For each day increase in muscle strengthening or toning, there were lower odds of having the overweight perception (OR = 0.89, 95% CI = 0.86 - 0.93).

Table 6.9 Multinomial Logistic Regression of Year 1 Predictors Associated with Year 2 Weight Perception

	Year 2 Weight Perception		
	Odds Ratio (95% CI)		
Year 1 Variables	**Underweight**	**About the right weight**	**Overweight**
Gender			
Female	0.63 (0.51 - 0.78) ***	Ref	0.72 (0.58 - 0.89) **
Male	Ref	Ref	Ref
Gender-Minority	1.09 (0.62 - 1.89)	Ref	1.31 (0.79 - 2.17)
BMI			
Underweight	4.9 (2.98 - 8.11) ***	Ref	1.02 (0.29 - 3.56)
About the right weight	Ref	Ref	Ref
Overweight and Obese	0.22 (0.12 - 0.39) ***	Ref	6.66 (5.13 - 8.65) ***
Unknown	0.82 (0.69 - 0.98) *	Ref	2.62 (2.15 - 3.19) ***

Intention of Weight Change			
Lose	0.31 (0.23 - 0.42) ***	Ref	5.26 (4.23 - 6.53) ***
Gain weight	3.81 (3.01 - 4.83) ***	Ref	0.33 (0.17 - 0.64) **
Stay the same weight	0.63 (0.51 - 0.78) ***	Ref	0.97 (0.75 - 1.24)
Not trying to do anything	Ref	Ref	Ref
Life changed after COVID: Time Spent Watching TV/Movies or Playing Video Games			
Increased	1.05 (0.84 - 1.30)	Ref	1.05 (0.85 - 1.30)
Stay the same/Not applicable	Ref	Ref	Ref
Decreased	1.10 (0.63 - 1.89)	Ref	2.27 (1.39 - 3.70) **
I Can Talk About My Problems With My Family			
Agree	1.16 (0.92 - 1.46)	Ref	0.86 (0.70 - 1.06)
Neither agree nor disagree	Ref	Ref	Ref
Disagree	1.51 (1.15 - 1.98) **	Ref	1.04 (0.82 - 1.36)
I am Worried About My Health (Regard COVID-19)			
True	0.74 (0.57 - 9.96) *	Ref	0.94 (0.73 - 1.19)
Neutral/ I don't know	Ref	Ref	Ref
False	0.81 (0.66 – 1.00) *	Ref	0.79 (0.65 - 0.97) *
Watching TV: Minutes per Day	1.0 (1.00 - 1.00)	Ref	1.00 (1.00 - 1.00) **
Video Games: Minutes per Day	1.00 (1.00 - 1.00) *	Ref	1.00 (1.00 - 1.00) *
Texting, messaging: Minutes per Day	1.00 (1.00 - 1.00) *	Ref	1.00 (1.00 - 1.00) **
Number of Days of Food Restriction in a Week with Intention of Weight Change	1.05 (1.00 - 1.10) *	Ref	1.04 (1.00 - 1.08) *
Number of Days Strengthen or Tone Muscles in a Week	0.98 (0.93 - 1.02)	Ref	0.89 (0.86 - 0.93) ***

P<0.0001 ***, P<0.01 **, P<0.05 *

6.4.4 The Association Between Year 1 Predictors and Weight Perception Transition from Year 1 to Year 2

The transition pattern refers to the change in weight perception from Year 1 to Year 2. The statistically significant factors associated with the change in weight perception were gender, BMI, weight change intention, days of food restriction, fruit consumption, texting, and muscle strength training. A group maintaining the perception of being about the right weight from Year 1 to Year 2 was used as the reference group. Table 7 shows the significance of predictors for each weight transition group (Year 1 to Year 2) compared to those remaining in the about the right weight perception group.

Females were less likely to change from overweight perception to underweight (OR = 0.13, 95% CI: 0.03 - 0.61) and remained overweight (OR = 0.54, 95% CI: 0.37 - 0.77). Gender minorities were more likely to change from about the right weight to overweight compared to males (OR = 3.41, 95% CI: 1.40 - 8.28).

Compared with healthy weight BMI, underweight BMI students had increased odds of remaining as underweight (OR = 10.34, 95% CI: 4.52 - 23.65). Overweight and obese BMI students had increased odds of changing from about the right weight to overweight (OR = 4.09, 95% CI: 2.37 - 7.04) perception and maintaining the overweight perception (OR = 17.62, 95% CI: 11.53 - 26.90). Unknown BMI students were more likely to change from about the right weight to overweight (OR = 2.11, 95% CI: 1.46 - 4.77) and remain in the overweight perception (OR = 3.43, 95% CI: 2.47 - 4.77).

Relative to those not trying to do anything, students with losing weight intention were less likely to remain underweight (OR = 0.13, 95% CI: 0.07 - 0.27), change from underweight to about the right weight (OR = 0.20, 95% CI: 0.09 - 0.44), and change from about the right weight to underweight (OR = 0.40, 95% CI: 0.23 - 0.71). However, these students were more likely to change from underweight to overweight (OR = 13.09, 95% CI: 1.41 - 121.36), about the right weight to overweight (OR = 3.13, 95% CI: 1.20 - 4.89), and remained as overweight (OR = 12.77, 95% CI: 8.71 - 18.73). On the contrary, students who wanted to gain weight were more likely to remain underweight (OR = 9.40, 95% CI: 6.25 - 14.13) and change from underweight to about the right weight (OR = 6.86, 95% CI: 4.13 - 11.38). For students who intended to stay at the same weight, they were less likely to remain underweight (OR = 0.57, 95% CI: 0.39 - 0.38) and change from underweight to about the right weight (OR = 0.48, 95% CI: 0.30 - 0.77).

In addition to weight change intention, the number of days in food restrictions was significantly associated with Group 8 (overweight to underweight). A one-day increase in food restriction was associated with a 54% increase in the odds of changing from overweight to

underweight (OR = 1.51, 95% CI: 1.21 - 1.89). It was also observed that with each additional daily serving of fruit, there was a 39% decrease in the odds of changing from underweight to overweight (OR=0.61, 95% CI: 0.42 - 0.89) and a 36% decrease in the odds of changing from overweight to underweight (OR = 0.74, 95% CI: 0.58 - 0.94). An inverse association was observed between muscle-strengthening exercises and maintaining an overweight perception. A one-day increase in muscle-strengthening and toning exercises was associated with a 16% decrease in the odds of maintaining an overweight perception (OR = 0.84, 95% CI: 0.78 - 0.90).

Table 7 Multinomial Logistic Regression of Year 1 Predictors Associated with Transition in Weight Perception from Year 1 to Year 2

Year 1 Variables	**Group 1**	**Group 2**	**Group 3**	**Group 4**	**Group 5**	**Group 6**	**Group 7**	**Group 8**	**Group 9**
	Remained Underweight	**Underweight to about the right weight**	**Underweight to overweight**	**Remained about the right weight**	**About the right weight to underweight**	**About the right weight to overweight**	**Remained as overweight**	**Overweight to underweight**	**Overweight to about the right weight**
Gender									
Female	0.77 (0.53 - 1.13)	1.08 (0.68 - 1.72)	1.72 (0.21 - 13.92)	Ref	0.61 (0.41 - 0.91) *	1.13 (0.72 - 1.77)	**0.54 (0.37 - 0.77) ****	**0.13 (0.03 - 0.61) ****	0.84 (0.52 - 1.35)
Male	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Gender-Minority	0.31 (0.08 - 1.21)	1.35 (0.34 - 5.27)	N/A	Ref	1.45 (0.54 - 3.86)	**3.41 (1.40 - 8.28) ****	0.92 (0.40 - 2.12)	1.04 (0.08 - 12.89)	1.67 (0.56 - 5.00)
BMI									
Underweight	**10.34 (4.52 - 23.65) *****	3.24 (1.07 - 9.81) *	N/A	Ref	1.58 (0.46 - 5.45)	N/A	N/A	N/A	1.63 (0.18 - 15.14)
About the right weight (Ref)	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Overweight and Obese	**0.04 (0.01 - 0.27) ****	0.18 (0.04 - 0.80) *	3.78 (0.18 - 79.15)	Ref	**0.25 (0.09 - 0.70) ****	**4.09 (2.37 - 7.04) *****	**17.62 (11.53 - 26.90) *****	4.49 (0.66 - 30.40)	**5.68 (3.39 - 9.51) *****
Unknown	0.82 (0.60 - 1.12)	0.79 (0.54 - 1.16)	6.85 (0.83 - 56.77)	Ref	0.73 (0.53 - 1.02)	**2.11 (1.46 - 4.77) *****	**3.43 (2.47 - 4.77) *****	4.76 (1.09 - 20.72)	1.52 (1.02 - 2.27) *
Weight Change Intention									
Lose weight	**0.13 (0.07 - 0.27) *****	**0.20 (0.09 - 0.44) *****	**13.09 (1.41 - 121.36) ***	Ref	**0.40 (0.23 - 0.71) ****	**3.13 (1.20 - 4.89) *****	**12.77 (8.71 - 18.73) *****	10.37 (1.40 - 77.17) *	**8.97 (5.38 - 14.94) *****
Gain weight	**9.40 (6.25 - 14.13) *****	**6.86 (4.13 - 11.38) *****	12.38 (0.51 - 300.02)	Ref	**2.46 (1.46 - 4.14) ****	0.55 (0.16 - 1.91)	N/A	6.70 (0.55 - 81.27)	N/A

Stay the same weight	**0.57 (0.39 - 0.83) ***	**0.48 (0.30 - 0.77) ****	1.99 (0.13 - 30.61)	Ref	0.83 (0.57 - 1.19)	1.31 (0.83 - 2.05)	1.12 (0.71 - 1.76)	6.06 (0.84 - 43.92)	1.44 (0.81 - 2.57)
Not trying to do anything	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref	Ref
Number of days of food restriction in a week	1.00 (0.92 - 1.08)	0.95 (0.86 - 1.05)	0.94 (0.67 - 1.30)	Ref	0.93 (0.85 - 1.03)	1.06 (0.98 - 1.14)	1.06 (1.00 - 1.12)	**1.51 (1.21 - 1.89) ****	1.01 (0.94 - 1.09)
Texting, messaging: Minutes per day	1.00 (1.00 - 1.00)	1.00 (1.00 - 1.00)	1.00 (0.99 - 1.01)	Ref	1.00 (1.00 - 1.00)	1.00 (1.00 - 1.00)	1.00 (1.00 - 1.00) **	0.99 (0.98 - 1.00)	1.00 (1.00 - 1.00)
Number of days in fruit consumption in the past week	0.96 (0.90 - 1.03)	1.10 (0.998 - 1.21)	**0.61 (0.42 - 0.89) ***	Ref	1.01 (0.94 - 1.09)	0.99 (0.91 - 1.07)	0.96 (0.90 - 1.02)	**0.74 (0.58 - 0.94) ***	0.95 (0.87 - 1.03)
Number of days strengthen or tone your muscles in a week	0.97 (0.90 - 1.05)	0.96 (0.90 - 1.08)	0.91 (0.61 - 1.37)	Ref	0.97 (0.90 - 1.06)	0.93 (0.86 - 1.02)	**0.84 (0.78 - 0.90) *****	0.71 (0.52 - 0.97) *	0.91 (0.84 – 1.00) *

[1] "NA" indicates no odds ratio was presented due to extremely small sample sizes

2 P<0.0001 ***, P<0.01 **, P<0.05 *

6.4.5 Model Prediction on Year 3 Weight Perception

Tables 8, 9, and 9.1 list the overall performance metrics for each weight perception category in RNN and LSTM models. Compared with CCA, MICE-imputed data achieved better overall model performance, specifically in ROC-AUC, Recall, and F1 scores across the three categories of weight perception. Notably, the mean ROC-AUC score in the RNN for the imputed dataset was 0.72, compared to 0.69 in CCA. The LSTM ROC-AUC score remained the same at 0.70 in both CCA and imputed datasets, but the ROC-AUC scores were higher in the imputed dataset for underweight, about the right weight, and overweight perceptions. Due to the superior model performance in the imputed dataset, the interpretation of the model results focused on the imputed dataset.

Table 8 Model Performance in Predicting Year 3 Weight Perception in CCA and MICE Imputed Datasets

Models	**Accuracy**	**Precision**	**Recall**	**F1**	**ROC-AUC**
CCA					
RNN	0.62	0.58	0.62	0.58	0.69
LSTM	0.60	0.57	0.60	0.58	0.70
MI					
RNN	0.60	0.61	0.60	0.60	0.72
LSTM	0.63	0.57	0.63	0.58	0.70

Table 9 Model Performance in Predicting Year 3 Underweight, About the Right Weight, and Overweight in CCA Datasets

CCA	Performance Metrics	Underweight	About the right weight	Overweight
RNN	Overall Accuracy	0.62		
	Precision	0.33	0.67	0.51
	Recall	0.02	0.79	0.59
	F-1 score	0.04	0.73	0.55
	ROC-AUC	0.65	0.64	0.78
	Mean ROC-AUC score	0.69		
LSTM	Overall Accuracy	0.60		
	Precision	0.34	0.68	0.49
	Recall	0.17	0.73	0.59
	F-1 score	0.22	0.70	0.53
	ROC_AUC	0.66	0.64	0.79
	Mean ROC-AUC	0.70		

Table 9.1 Model Performance in Predicting Year 3 Underweight, About the Right Weight, and Overweight in MICE Imputed Datasets

MICE	Performance Metrics	Underweight	About the right weight	Overweight
RNN	Overall Accuracy	0.60		
	Precision	0.30	0.72	0.57

	Recall	0.33	0.67	0.61
	F-1 score	0.31	0.69	0.59
	ROC_AUC	0.72	0.69	0.72
	Mean ROC-AUC	0.72		
LSTM	Overall Accuracy	0.63		
	Precision	0.25	0.66	0.57
	Recall	0.02	0.82	0.60
	F-1 score	0.03	0.73	0.58
	ROC_AUC	0.79	0.66	0.80
	Mean ROC-AUC	0.70		

6.4.6 SHAP Analysis

Figures 6.5 A-C and Figure 6.6 A-C represent the summary plots for RNN and LSTM models across the two-year time step to provide a global understanding of each factor's association in predicting weight perception in Year 3. Factors were ranked based on their importance in the model predictions.

Consistently in both RNN and LSTM models, the top five important factors in the underweight group were weight change intention ("What are you trying to do about your weight?"), BMI, total number of days in diet restriction, days in muscle training exercises, and whether the students eat breakfast. In addition to these factors, talking to family members about their problems appeared among the top five factors in the "about the right weight" group. In the overweight group, skipping breakfast to lose weight was among the top five factors, along with those identified in the underweight and "about the right weight" groups.

To account for the direction and magnitude of each factor's association with the prediction of weight perception in two time periods, SHAP beeswarm plots were utilized to illustrate how predictors from Year 1 and Year 2 influence the weight perception in Year 3. For instance, positive SHAP values indicated that the factors moved the prediction towards a specific weight perception group, whereas the negative values indicated moving away from the prediction. Since the overall important factors were similar in RNN and LSTM, the interpretation of the magnitude and contribution of the factors was based on the LSTM SHAP beeswarm plots (Figure 6.6.1 A - C). The RNN SHAP beeswarm plots can be viewed in Appendix D.

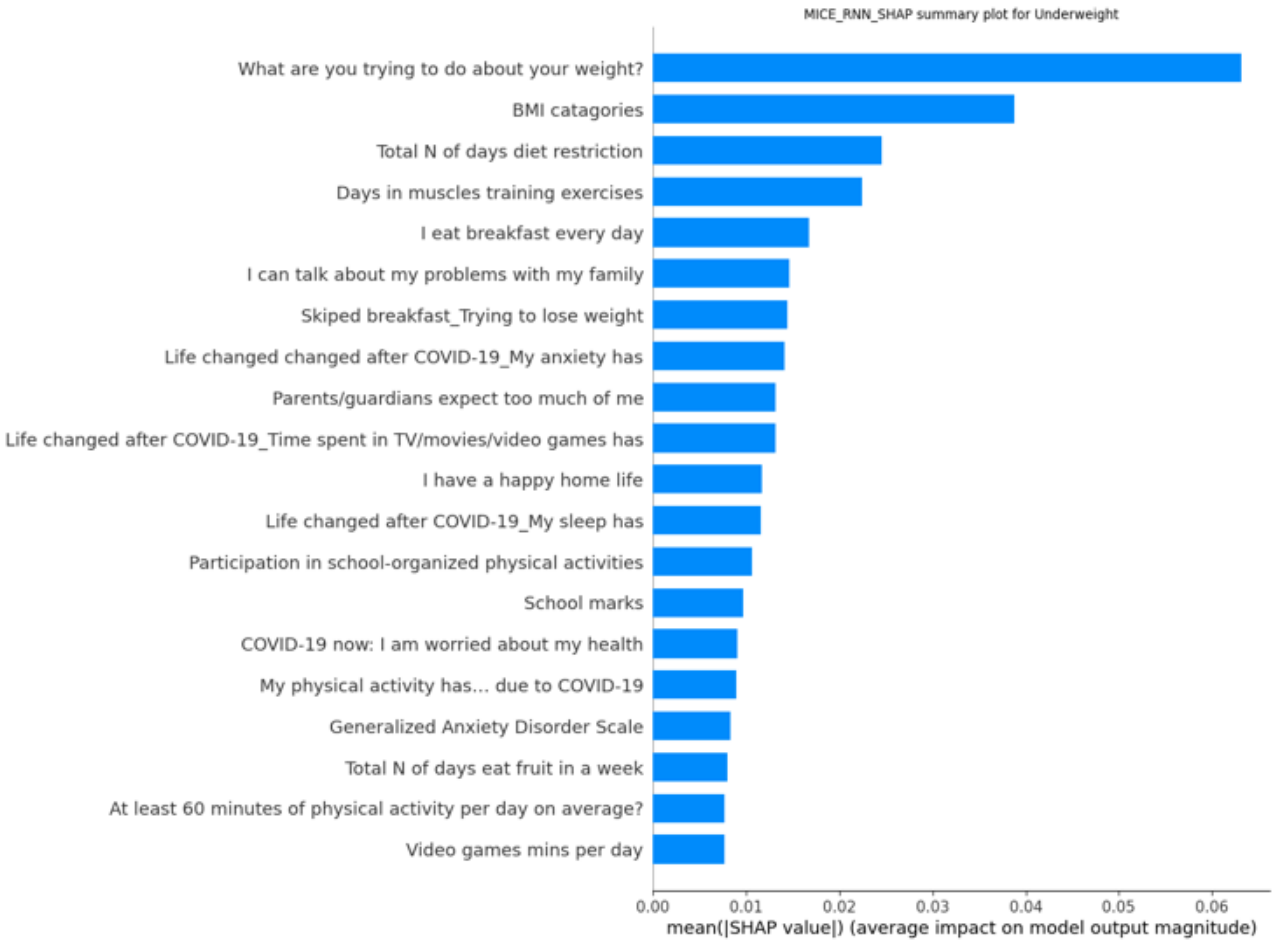
In the underweight perception (Figure 6.6.1 A), weight change intention was the most significant contributor. There was a high density of positive SHAP value (red) in the categories of intention of weight change. This reflected that students who chose the higher categories, such as gaining weight, staying the same, or not trying to do anything about weight, were more likely

to be predicted in the underweight group. The blue dots in skipping breakfast represented not skipping breakfast, resulting in a positive SHAP value, showing that not skipping breakfast to lose weight had a higher likelihood of being predicted as having an underweight perception. Regarding BMI categories, a lower value of BMI pushed the model prediction toward underweight perception. Additionally, increasing the number of days in muscle training and an increased physical activity level (lower value in blue) due to COVID-19 resulted in a positive SHAP value, meaning students involved in physical activity were predicted to belong to the underweight perception group.

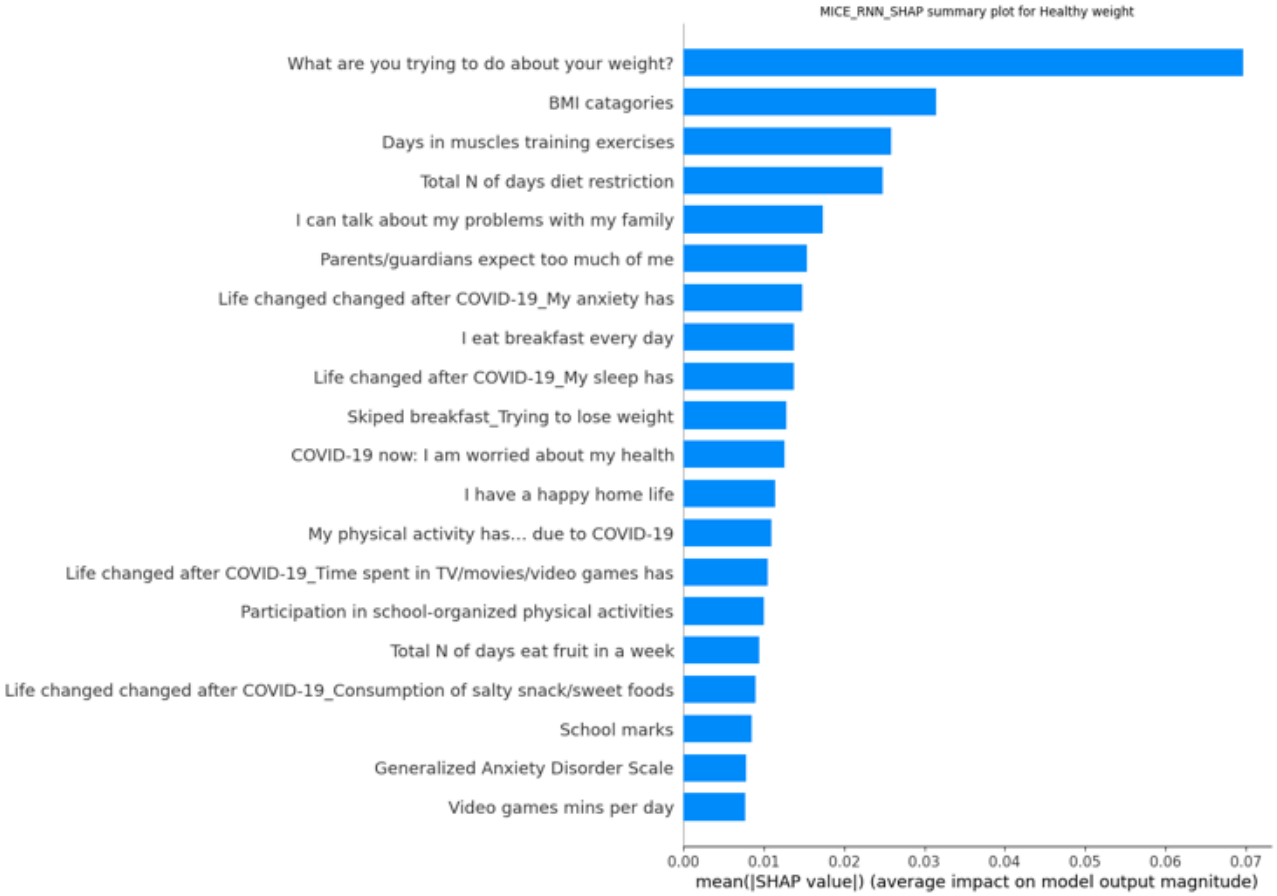
The effect of the variables mentioned in the underweight group was similar to that of the about the right weight group. Higher values in weight change intention, including gaining weight, staying the same, or not trying to do anything, not skipping breakfast to lose weight, and lower values in BMI categories such as healthy weight, more days in muscle training and increased physical activity levels during COVID-19 contributed towards classifying students into the about the right weight perception group. Notably, the average SHAP values for classifying the about the right group were overall higher among these predictors than in the underweight group. This suggests that these mentioned top factors contribute more significantly to predicting about the right weight perception. On the contrary, these variables showed the opposite direction in the overweight perception group. Regarding weight change intention, students who tried to lose weight (lower feature value in blue) were more likely to perceive themselves as overweight. In BMI, higher BMI categories showed a positive SHAP value, meaning either overweight and obese or unknown contributed to the prediction as having overweight perception. In addition, students who chose skipping breakfast to lose weight had a positive SHAP value, indicating they were more likely to be grouped into the overweight perception group. Students who were predicted to be in the overweight perception group also spent fewer days in muscle training exercises and decreased physical activities due to COVID-19.

Figure 6.5 SHAP summary plot for each weight perception status in RNN

(A) Underweight



(B) About the right weight



(C) Overweight

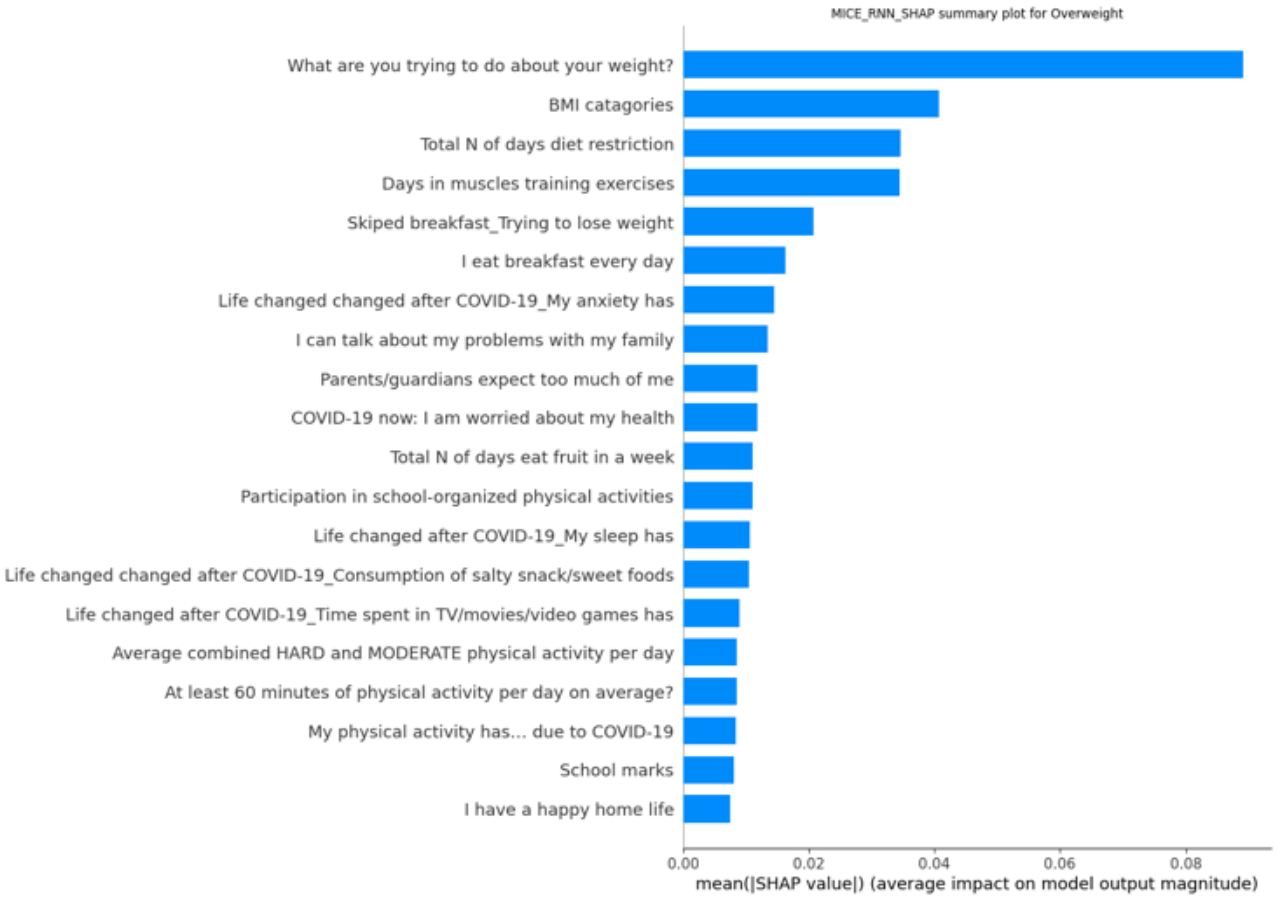
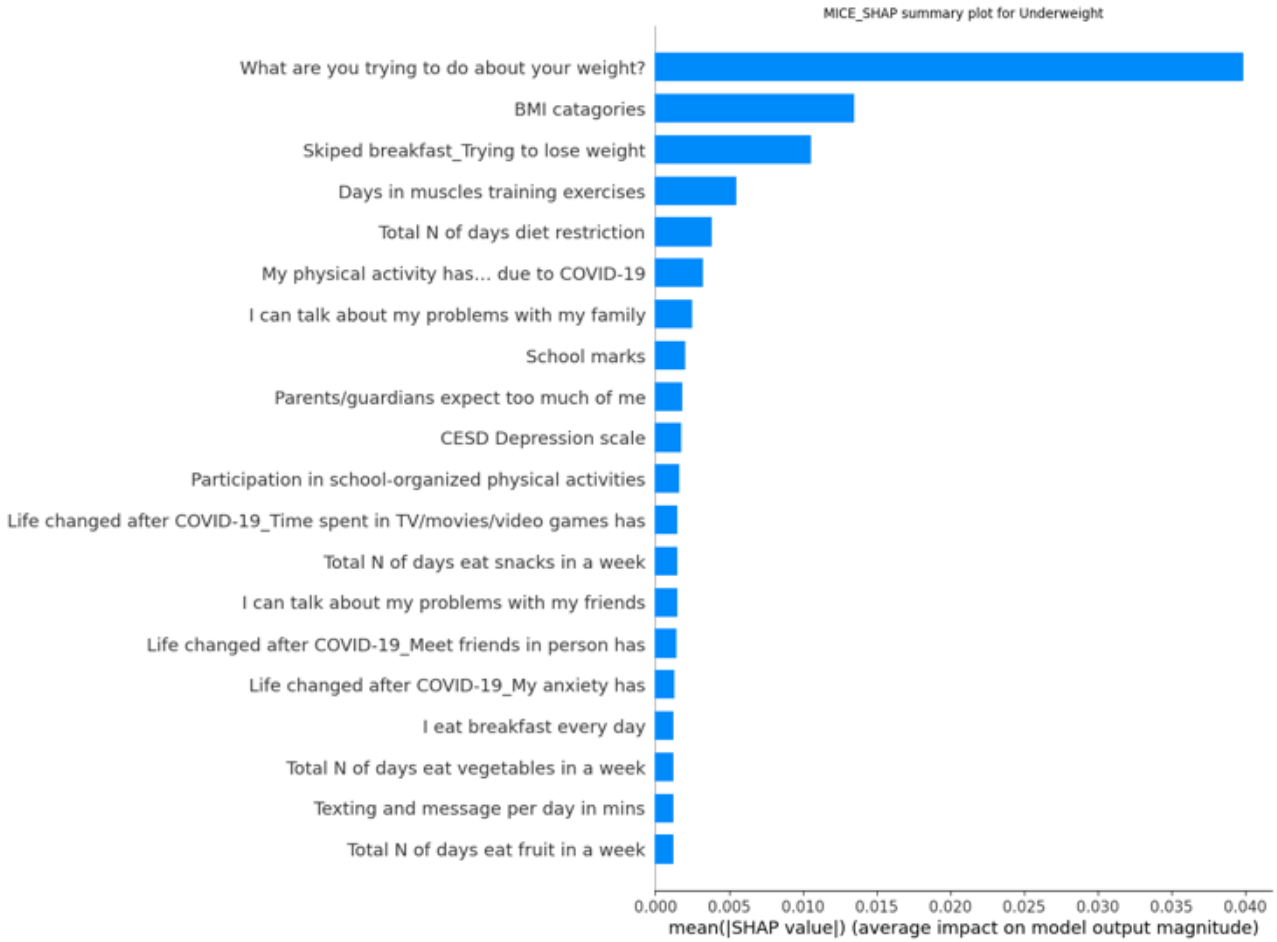
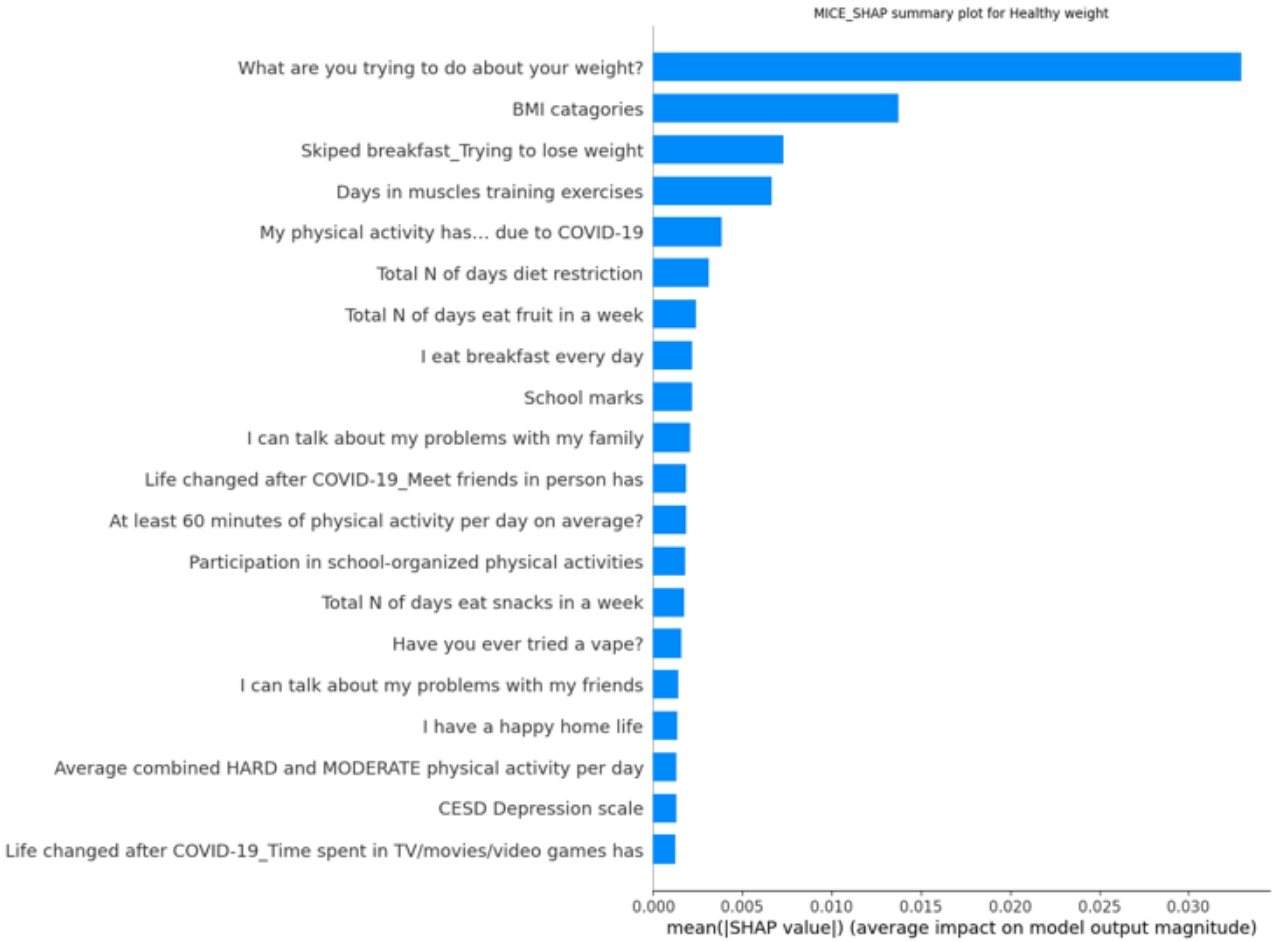


Figure 6.6 SHAP summary plot for each weight perception status in LSTM

(A) Underweight



(B) About the right weight



(C) Overweight

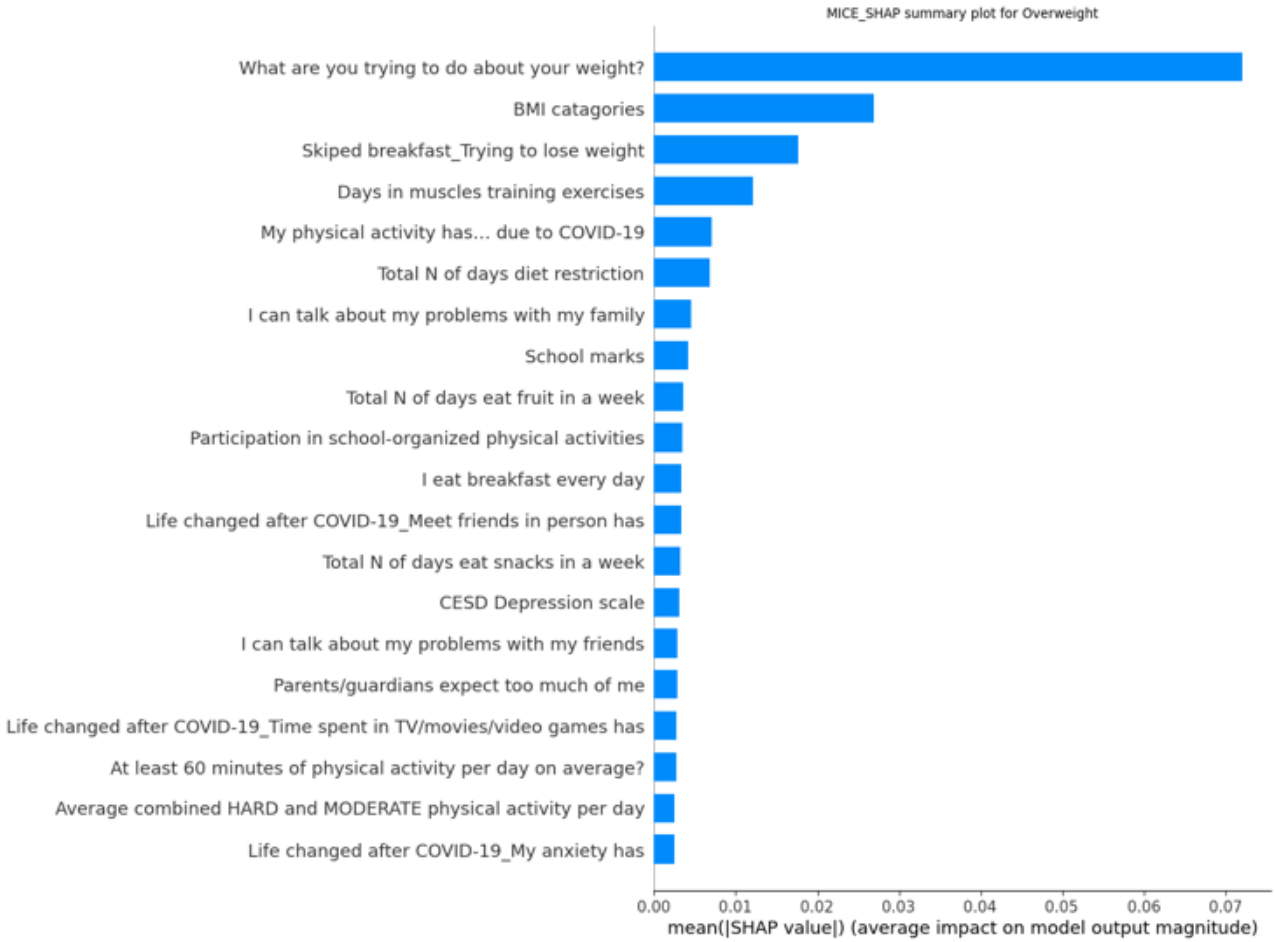
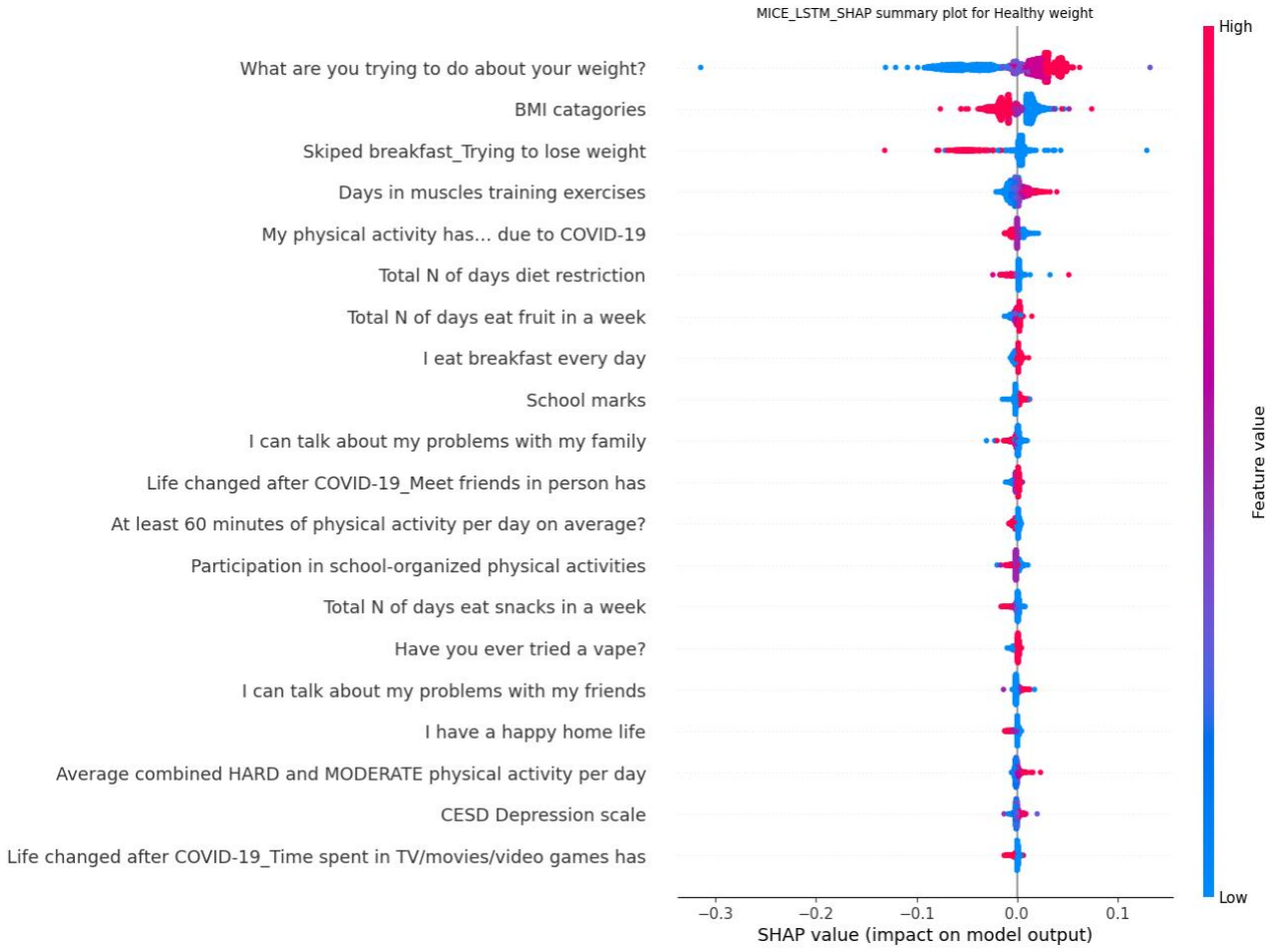


Figure 6.6.1 SHAP beeswarm plot for each weight perception status in LSTM

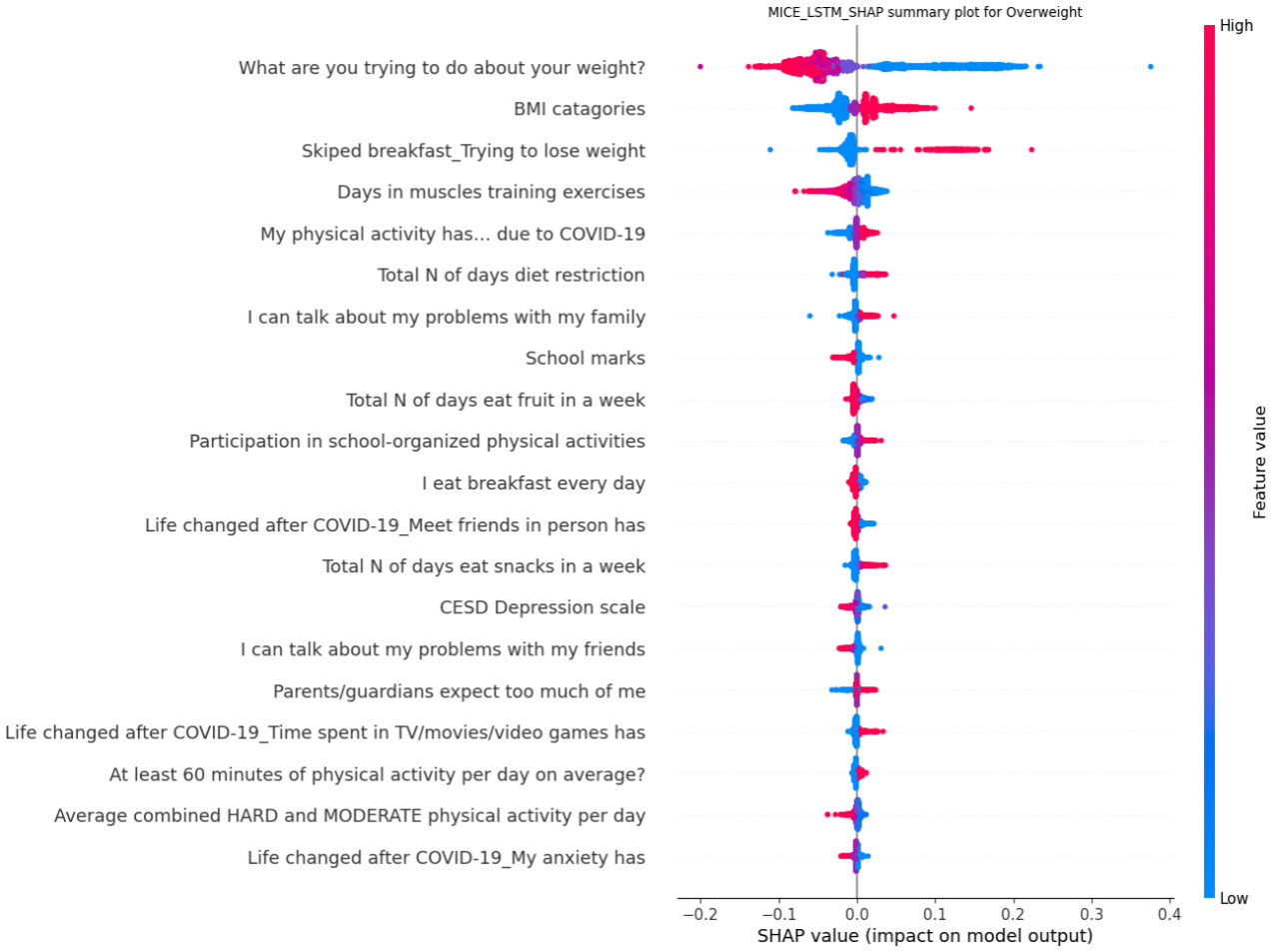
(A) Underweight



(B) About the right weight



(C) Overweight



6.5 Discussion

This study investigated the transition in weight perception and its prediction using three years of linked COMPASS data. The MCM estimated the probabilities of students moving from different weight perception states, illustrating that most students were more likely to maintain their past year's weight perception with the highest probabilities of staying at about the right weight. There was a moderate probability of students changing from underweight to about the right weight or overweight to about the right weight and a very mild probability of moving from overweight to underweight or underweight to overweight. This finding was consistent with similar studies investigating adolescent weight perception transition [176, 177]. Both studies showed that the weight perception transition was stable. Most adolescents maintained the same weight status perception in early and late adolescence. Specifically, Hahn et al. [176] showed that 15.5% transitioned to overweight in late adolescence, and the transition state from overweight to not overweight was least expected, further aligning with our study finding that the probability of being overweight to underweight was unlikely.

In addition to the weight perception transition probabilities, certain factors were identified as significantly associated with Year 2 weight perception and the nine weight perception transition groups in both multinomial logistic regression and time series machine learning models. Time-variant factors such as students' BMI status, intention of weight change, days in diet restriction, and muscle training exercises were significantly associated with the Year 2 weight perception and weight perception transition. These factors were also discovered in the SHAP summary plot in the RNN and LSTM model, with student's weight change intention and BMI ranked as the top two most important factors associated with the model prediction for each weight perception. Relative to students who were not trying to do anything about their weight, students who intended to lose weight were more likely to move from underweight to overweight, remain overweight and move from about the right weight to overweight. Moreover, the intention of gaining weight had a higher likelihood of moving from underweight to about the right weight perception and maintaining underweight perception. Studies also reported that the attempt to lose weight happened among adolescents with an overweight perception [162, 178].

Notably, BMI status had a critical contribution to the transitioning of weight perception states. Students with BMI classified as overweight and obese or unknown were more likely to change from about the right weight to overweight and were more likely to stay in the overweight perception. Meanwhile, underweight BMI students were more likely to remain underweight. The results aligned with research that found a positive relationship between BMI status and weight perception. Participants with higher BMI were dissatisfied with their weight and had higher odds of perceiving themselves as overweight, while those with lower BMI had lower odds of perceiving themselves as overweight [179, 180].

Regarding BMI, diet restriction was predominately significant in group 9, changing from overweight perception to underweight perception. Increasing the number of days of restricting diet had higher odds of moving from overweight to underweight perception. Eating less was reported as a healthy weight control strategy among individuals who perceive themselves as overweight. However, this strategy was also indicated as disordered eating, such as fasting and meal skipping [176, 224, 247].

Moreover, muscle strengthening and toning exercises appeared to have a protective effect on the weight perception transition group. More days of muscle training exercise resulted in 16% lower odds of maintaining the overweight perception. Similarly, students who had more days in muscle training exercises in Year 1 had higher odds of perceiving as having about the right weight in Year 2. This modifiable effect of exercise was observed in several studies, specifically on maintaining about the right weight and reducing the likelihood of body dissatisfaction [177, 181].

One interesting turning point discovered only from the model with Year 2 weight perception outcome was the inverse relationship between life changes due to COVID-19 on the level of watching TV/movies and playing video games and weight perception. Surprisingly, the decreased level of spending time in screen time in Year 1 had increased the odds of perceiving as overweight. This suggests using social media might have a beneficial effect on weight perception. Social media, including texting, TV, and video games, had mixed effects on body image and weight perception. Messages on social media and advertisements on the idealization of a thin body image were reported in weight-based internalized stigma, negative self-perception, and higher body dissatisfaction [248, 249]. On the contrary, some studies have argued about the positive influence of social media, especially social movements, regarding programs that raise awareness of weight bias or promote self-acceptance of various body sizes through TikTok and Instagram. These media-based movements were shown to reduce the negative effects of the thin-ideal context and positivity related to body satisfaction [185, 186].

There are several limitations to this study. First, many students were not included in the analysis since the weight perception question was only received by 50% in the Quebec survey. Due to the reduced sample size, the result of this analysis might not be generalized to represent the whole population across all provinces regarding weight perception. Additionally, since this focuses only on schools that participated in the COMPASS survey, the generalizability is also limited to the students from those specific schools. It may not be used as an indicator outside of COMPASS participants. Second, the machine learning models only included the time-variant variables and did not include the demographic variables. Thus, the prediction of weight perception and associated factors might not be comprehensive. As indicated in the statistical model, baseline characteristics such as gender were significantly associated with weight perception transition patterns. There were also limitations to the lack of stratification analysis on

whether factors associated with weight perception varied among different racial and ethnic groups. Future studies should consider performing subgroup analyses. Moreover, the two deep learning models suffered from the imbalance of samples in weight perception despite the effort of using class weight during the model prediction process. There was no other COMPASS cohort available for external validation to test the model's prediction of weight perception. In the future, the aim is to explore the prediction of weight perception using additional years of linked data, rather than just two, to improve model performance and investigate advanced resampling methods that could retain students' longitudinal records.

6.6 Conclusion

The study examined the transition probabilities between three states of weight perception and the factors associated with the transition. The study further utilized deep learning models to predict future weight perception, maintaining the temporality of the predictors in the previous two years. Results illustrated that students with higher BMI, intending to lose weight, and spending more days on diet restriction were more likely to have overweight perceptions, while muscle training had a protective effect, and social media might have a mitigating effect on weight perception. To the best of current knowledge, this was the first study that utilized both approaches to examine weight perception transitions and the prediction of weight perception over time by observing the associated factors. The results of the study emphasize the importance of understanding the relationship between risk or modifiable factors and weight perception to overcome negative self-perception on weight and body image. Future studies should consider assessing media-based interventions for promoting healthy weight perceptions and investigate the effects of the mentioned predictors on weight perception among gender or racial and ethnic groups.

Chapter 7

Final Discussion

This section discusses the main findings from Studies 1 to 3, how they respond to the research question, their strengths and limitations, contributions, and future work.

7.1 Study 1 Findings

- Study 1 focused on clustering students' profiles based on various behavioural measures that influence adolescent health

7.1.1 Study 1 RQ1: What factors are associated with the clustering of student profiles in the 2018-19 (pre-COVID) and 2020-21 (post-COVID) COMPASS cohorts?

From the ML perspective, Study 1 applied k-means clustering two times and generated high-level and sub-level clusters from the COMPASS pre-COVID cohort (2018-19) and post-COVID cohort (2020-21) and implemented the RF models to identify the important factors contributing to each cluster.

Based on the top five important variables from the SHAP plots, high-level clusters discriminated student groups by their risk of substance use, including using marijuana, access to marijuana, smoking and tobacco use, and discussion about mental health conditions or depression in both cohorts. The multi/logistic regression further presented the strength of the association. Relative to those with no involvement in substance use and high mental health well-being, students who engaged in substance use, felt depressed in the last seven days and refused to discuss their mental health were more likely to be in one cluster. In the pre-COVID high-level clusters, 0.47% of the students over the total number of complete records (N=32899) were considered at high risk of substance use compared to students who were considered at no risk of substance use (73.9%). In the post-COVID cohort, only 0.2% of students were in the high-risk cluster relative to the no-risk cluster, which contained the most students. In terms of mental health, among the high-risk cluster in pre-COVID, 34.2 % of them felt depressed for 5-7 days in the past week, while 42.9% of students disagreed that they had a happy home life in the high-risk cluster.

Regarding the sub-clusters without the substance use variables, the major factors that were important in forming the sub-clusters in the pre-COVID cohort were school-based covariates, including physical activity at school, academic performance, bullying, and weight-

related factors. Students who completed their homework on time, seldom skipped classes, had higher school connectedness scores, participated in school sports with more than 5-7 physically active friends, and did not experience bullying were more likely to be in one sub-cluster compared to students who had lower school connectedness score, were less active, experienced bullied, often had incomplete homework and skipped classes. Among these sub-clusters that are considered lower academic performance and less active, students had either an underweight or overweight perception with no intention to stay the same weight. In the post-COVID cohort sub-clusters, the top five important factors emphasized the impact of the COVID-19 pandemic on life change, weight status, individual level of physical activity, and mental health-related categories. Students who had overweight and obese or unknown BMI, increased cannabis use due to COVID, were less active, increased the number of days in purchased fast food, and had higher scores in anxiety and lower scores in flourishing had a significantly higher likelihood of belonging to one sub-cluster.

7.1.2 What are the similarities or differences of the factors associated with the clustering of student behaviour pre-COVID-19 and post-COVID-19?

The involvement of substance use and mental health-related issues were significant factors in both the pre-COVID and post-COVID cohorts. Students who were at high risk of substance use and not mental well-being were clustered in the same group. These factors have been consistent regardless of the period before and after the pandemic. Weight-related variables have been significantly associated with the sub-clusters in both cohorts. Regardless of the effect of the pandemic, students who intended to change their weight, perceived as either underweight or overweight, or BMI identified as overweight and obese, were more likely to be in clusters compared to healthy weight BMI and the right weight perception.

For the difference in the covariates, school-related factors, including academic performance, school connectedness, and physical engagement at school, had a significant association with the clustering. In contrast, changes due to COVID-19 substance use, individual-level physical activity, and mental health in terms of flourishing and anxiety were the important factors associated with the clustering in the post-COVID cohort.

Comparing with current literature: The findings on substance use and mental health issues as the high-level cluster's important factors were consistent in the current literature. Regarding the small proportion of students being high risk of substance use, a study researching the pattern of polysubstance use among COMPASS students also discovered that the high-risk group constituted the lowest population (11.5%) at baseline in the three-year linked cohort pre-COVID compared to most students in low risk (32.5%) [155]. Similar to the mental health factor in the high-level cluster, this study also discovered that the high-risk group was significantly associated

with the highest mean scores on the Center for Epidemiologic Studies Depression Revised Scale (CESD) [155]. The post-COVID cohort further showed the clustering of substance use, aligning with research that students with multi-polysubstance use maintained this pattern over time. Weight-related factors did not vary by the pandemic; youth with overweight and obese BMI with weight change intention were more likely to be clustered. It was known that overweight perception was engaged in more weight-loss activity [162]. Sub-clusters with higher BMI also did not engage frequently in physical activity, regardless of the disruption of the pandemic.

The major difference in the pre-pandemic sub-cluster was the variables related to school activities. Several research studies also showed the importance of the school environment and academic performance at school. In our findings, students with higher academic performance and more participation in school sports were more likely to be clustered together than clusters with low performance and less active physical activity. In Borgavakar & Shrivastava's research using k-means clustering [156], clusters were identified on academic performance, and the characteristics in the low-performance cluster were low grades and incomplete assignments, agreeing with the Study 1 findings. Additionally, it was mentioned that school engagement is associated with health-related behaviours; specifically, students who felt more connected to schools had higher levels of physical activity, while students who had experienced bullying had lower levels of engagement at school [157, 158, 159].

Moreover, the post-COVID cohort revealed that COVID-19 increased substance use and these students had higher scores in anxiety and lower scores on the flourishing scale. This finding was also consistent with the scoping review that 48% of the studies reported an increase in cannabis during COVID-19 among the 129 documents, and anxiety was one of the major contributing factors to the increased cannabis use [160]. Aligning with the studies on mental health research, the inverse relationship was observed between mental health symptoms of anxiety and mental health wellness during COVID-19, specifically higher mental health symptoms associated with lower well-being as well as increased prevalence of anxiety during the pandemic [161, 162]. Additionally, physical activity was also found to decline substantially during the quantitative period [210]. Overall, the important factors from the pre- and post-COVID clusters emphasized the school engagement improvement towards adolescent mental health well-being, physical activities and substance use interventions.

7.2 Study 2 Findings

- Study 2 focused on the weight-related measure, BMI, from the various behavioural measures mentioned in Study 1

7.2.1 Study 2 Q1: What important factors are associated with youth BMI status in the 2020-21 COMPASS student cohort using ML classifiers from the CCA and MI approach?

Study 2 applied supervised learning algorithms to predict youth BMI status. In both the CCA and MI approaches, XGBoost outperformed the SVM model with dimensionality reduction and the RF model with the top 50 important variables. In CCA, XGBoost achieved in an overall 64% accuracy, 63% precision, 64% recall, 63% F1 score, and 78% ROC-AUC score. In MI, XGBoost had an overall 64% accuracy, precision, recall, F1 score, and 79% of ROC-AUC score. The predicted outcomes were BMI status of healthy weight, overweight and obese, and unknown. Based on the global SHAP summary plot, the top three variables in all three categories were weight perception, gender, and weight change intention. Moreover, school marks, age, and participation in school sports were among the top ten important factors across all BMI statuses.

Among healthy weight and overweight and obese BMI, weight perception was ranked as the most important factor, while gender was the most critical factor for predicting unknown BMI. The SHAP plot also provided magnitude for interpretation. Specifically, older students who achieved higher school marks and described their weight as being about the right weight positively contributed to the model prediction as healthy weight compared to the other two BMI categories. In contrast, perceiving themselves as slightly overweight and overweight positively affected the model prediction of overweight and obese BMI. In addition, having a weight change intention towards losing weight contributed towards the prediction as overweight and obese compared to staying the same or not trying to change their weight.

Another notable finding was that gender was the most important feature for unknown weight prediction. Genders other than females and males were more likely to be classified as having unknown BMI. Regarding weight perception, the gender minority group who were predicted as having unknown BMI perceived themselves as either very underweight or slightly underweight. In terms of age, older students were less likely to be classified in the unknown BMI compared to younger students. Through the stratification between females and males, the factors associated with overweight and obese BMI in males were related to sports participation and muscle training compared to females.

7.2.2 Study 2 Q2: How do the results of ML-generated factors compare with the common factors leading to childhood obesity in the previous literature? What are the similarities and differences?

Comparing with the current literature: The top three important factors, weight perception, gender, and weight change intention, are consistent with findings in the current

literature on youth obesity and the clustering results from Study 1. Weight perceptions were significantly associated with the intention of weight change [162]. Study 1 showed that students with overweight perceptions and the intention of losing weight were more likely to be in one cluster. Meanwhile, self-perception of being overweight was more likely to be associated with engagement in weight loss activities [162, 163, 164]. Having overweight and obese BMI was also more likely to try to lose weight [164]. Moreover, the perception of overweight was found to be predictive of increased weight gain over time and a higher likelihood of engaging in overeating to cope with stress [165].

In terms of gender, the SHAP plot revealed that males were predicted to be overweight and obese compared to females. A study observing the prevalence of obesity among boys and girls in 188 counties found that 65% of them reported a higher obesity rate in boys [166]. However, it should be noted that, due to the nature of self-reported BMI, females are more likely to underreport their weight and therefore obtain a lower BMI compared to males [167, 168, 169]. Current media promotes a 'thinner body' image and weight concerns, which are known to be associated with misconceptions of body weight. Adolescents influenced by the social stigma of body image are more likely to underestimate their body weight and height [169].

Other than the top three important factors, the findings on school marks, age, and sports participation were consistent with the current literature. Academic performance has been associated with childhood obesity among the 32 studied countries, where healthy-weight students were more likely to report good academic performance than obese children [170]. Another COMPASS study also found that students with overweight BMI had lower school achievement [171].

An insightful finding was the relationship between gender, underweight perception, and unknown BMI. Genders other than females and males were more likely to be classified as unknown, and they perceived themselves as underweight or slightly underweight. This was supported by Guss et al. [172], who claimed that transgender youth had higher odds of self-perception as an underweight or healthy weight when they were actually overweight and obese compared to cisgender females. Although the exact reason behind the non-disclosure of BMI-related information among gender minorities is unknown, it has been emphasized that body dissatisfaction and younger age contribute to missing BMI data, and the transgender population experiences higher body dissatisfaction compared to the cisgender population [173, 174, 175]. Moreover, those perceiving themselves as underweight were more likely not to report BMI, with the likelihood being stronger than that of the overweight perception [220]. Moreover, perceiving oneself as underweight was more likely to lead to not reporting BMI, with the likelihood being stronger than that of the overweight perception [220]. Our finding that an underweight perception was predictive of unknown BMI aligns with this point.

In summary, our results from the ML approach discovered that the factors associated with predicting youth BMI fall into the psychological aspect among the commonly peer-reviewed risk factors listed. Common behavioural factors such as physical activity at school were ranked less important than weight perception and intention to change weight. However, physical activity, such as muscle training, was the key factor associated with overweight and obese BMI in males. Studies indicate that males are more likely to engage in muscle-build activities and gain weight [174, 255]. Since weight perception was the most important factor in all BMI categories, establishing accurate weight perception on this modifiable factor could benefit healthy weight management. In the meantime, it would also be worth reviewing the key factors that have a different influence on males and females when developing adolescent obesity-related interventions.

7.3 Study 3 Findings

- Weight perception was the most important factor associated with BMI prediction. Therefore, Study 3 chose this variable as the main outcome for transitional pattern and time series analysis

7.3.1 Study 3 Q1: How do the estimated probabilities of weight perception change from Year 1 (2020-21) to Year 2 (2021-22) and from Year 2 (2021-22) to Year 3 (2022-23) using the Markov Chain Model?

For the transition change of weight perception derived from the MCM, most students were more likely to stay in the same weight perception state over time. From Year 1 to Year 2, the probability of remaining the same weight perception as the previous year was 63.24% for underweight, 80.11% for about the right weight, and 71.08% for overweight. From Year 2 to Year 3, the probabilities of staying underweight, about the right weight, and overweight were 64.84%, 80.00%, and 76.44%.

There was a moderate likelihood of transitioning from underweight to about the right weight and overweight to about the right weight. The probability of changing from underweight to about the right weight was 33.53% from Year 1 to Year 2 and 31.77% from Year 2 to Year 3. For the transition from overweight to about the right weight, the transition probability was 25.77% from Year 1 to Year 2 and 20.91% from Year 2 to Year 3.

Moreover, there was a negligible or small likelihood of transitioning from underweight to overweight and overweight to underweight. From Year 1 to Year 2, the transition probability from underweight to overweight was 3.22% and 3.38% from Year 2 to Year 3. The probability

of changing from overweight to underweight was 3.16% from Year 1 to Year 2 and 2.66% from Year 2 to Year 3.

Based on the result, it appeared that the probability of dynamic weight perception change did not fluctuate during the two-year period. However, there was a moderate likelihood of changing to about the right weight in the next year if previously having an overweight or underweight perception. However, it is unlikely that the students who considered themselves underweight would change to overweight or overweight to underweight.

7.3.2 Study 3 Q2: What factors are associated with the weight perception in Year 2 and the weight perception transitional change from Year 1 to Year 2?

The two multinomial logistic models were applied to observe the association between the Year 1 predictors and the Year 2 weight perception and the Year 1 predictors on the nine transition groups, indicating the change of weight perception from Year 1 to Year 2.

In the first model with Year 2 weight perception as the outcome, gender, BMI, weight change intention, talking about their problems with their family, concerns about health, life changed after COVID on social media/playing video games, and weekly muscle-strength training in Year 1 were significantly associated with weight perception. Relative to about the right weight, females were less likely to perceive themselves as underweight or overweight than males. Compared to healthy BMI, underweight BMI and overweight and obese BMI students were more likely to perceive themselves as underweight and overweight. Students who tried to gain weight had higher odds of perceiving themselves as underweight, while losing weight intention resulted in higher odds of having an overweight perception. Additionally, increasing the number of days in muscle training exercise decreased the probability of having an overweight perception. Interestingly, students who spent less time on social media after COVID-19 had higher odds of perceiving themselves as overweight.

In the second model, with weight transition state from Year 1 to Year 2 as the outcome, gender, BMI, weight change intention, number of days in food restrictions, and muscle training were the significant factors associated with the transition of weight perception status. Females, students with overweight and obese BMI, or those with unknown BMI, were more likely to transition from about the right weight to overweight or maintain overweight from Year 1 to Year 2. Among the students trying to lose weight, they were more likely to change from underweight or about the right weight to overweight. Increasing the number of days in food restriction was associated with higher odds of transitioning from overweight to underweight, while more days in muscle training reduced the odds of remaining an overweight perception.

7.3.3 Study 3 Q3: In the Deep Learning approach, what are the important factors associated with predicting weight perception in Year 3 from the previous two years?

RNN and LSTM were implemented to predict the Year 3 weight perception based on Year 1 and Year 2 predictors. The predictors were the time-variant variables, excluding the baseline characteristics such as gender or ethnicity.

Based on the SHAP analysis from RNN, the top three important features across the underweight, about the right, and overweight perception were weight change intention, BMI, and total number of days in diet restriction. Other than the top three features, days in muscle training, being able to discuss problems with family, and skipping breakfast to lose weight were also significant factors associated with the model prediction. The same variables were also discovered in the LSTM model.

The SHAP analysis plot from LSTM was displayed to interpret the magnitude of the association. For weight change intention in a higher category, including gaining weight, staying the same, or not trying to do anything about the weight, students were predicted as underweight or about the right weight perceptions. Lower or healthy BMI and not skipping breakfast to try to lose weight positively contributed to the model's prediction of being underweight or about the right weight perception. In contrast, having a higher BMI, an intention to lose weight, and an increased number of days skipping breakfast pushed the model prediction toward classify these students as having overweight perceptions. Additionally, fewer days in muscle training were predicted to have an overweight perception. The results of the deep learning models confirmed the similar significant variables associated with the transition of weight perception compared to the ones mentioned in multinomial logistic regression.

Comparing with the current literature: Based on the MCM, most students tend to retain the same weight perception over time. This was consistent with the two studies investigating the weight perception transition in adolescence in one year of change or from early to late adolescence [176, 177]. Both studies illustrated that most youths maintain the same weight perception status in early and late adolescence. Research on weight perception using MCM is limited, but existing studies on the transition probability of BMI status also confirmed that the weight status tends to remain constant over time [149, 150].

Both multinomial logistic regression and deep learning models have shown that BMI, weight change intention, and muscle training were the major factors in the transition of weight perception. Specifically, overweight and obese BMI with losing weight intentions at baseline were more likely to transition from about the right weight to overweight or remain overweight. Similar to the results of Study 3, other studies also reported that the attempt to lose weight

occurred among adolescents with an overweight perception [162, 178]. Additionally, existing literature also found that participants with higher BMI and body dissatisfaction had higher odds of perceiving themselves as overweight, while those with lower BMI were less likely to perceive themselves as overweight [179, 180]. Moreover, muscle training exercise was a protective factor against overweight perception for both the second-year weight perception and the transition over time. This physical activity has been agreed upon as a method to manage weight among participants with an overweight perception, mitigate body dissatisfaction, and maintain a healthy weight perception. The modifiable effect of exercise has been observed in several studies, specifically for maintaining about the right weight and reducing the likelihood of body dissatisfaction [177, 178, 181, 182].

The decreased time spent watching TV/video games after COVID-19 increased the likelihood of an overweight weight perception in Year 2, which differed from the current literature on weight perception. Existing literature reported that increasing video game time was associated with an overweight perception or higher BMI among youth [177, 183]. Body dissatisfaction related to social media is linked to the stigmatization of body image among females [198]. However, a review article also noted that 40% of the studies observed the modifiable effect between active video games and youth obesity among 2433 articles, suggesting academic interest in using them as an intervention [184]. Although no positive information has been found between video games and weight perception, some studies argued that specific social movements, including programs, increased the awareness of weight bias or promoted self-acceptance of various body sizes. These media-based movements can reduce weight bias and body dissatisfaction [185, 186]. The body appreciation resulting from the positive influence of social media could encourage a more accurate perception of youth weight.

7.4 Strengths

There are several strengths of this dissertation. First, this study comprehensively examines multiple aspects of student behaviours as features to generate the results throughout the three studies. COMPASS data system assessed various elements of student-level questionnaires annually, enabling the studies to apply advanced modelling techniques to observe student behaviours. Previous studies used clustering analysis involving single or few aspects of student characteristics such as admission data, mental health, cognitive activity or online learning during the pandemic [111, 154, 187, 188]. None of the studies incorporated different perspectives, including healthy eating, physical activity, substance use, screen time, weight-related status, and school engagement, to perform clustering. The COMPASS data system has acquired longitudinal student-level data since 2012, and having annual information provides the capacity to observe temporal changes in features for time series modelling, such as deep learning models.

The second strength is the comprehensive process of handling missing data, ML data preprocessing for optimizing model performance and improving ML model interpretation. In both Study 2 and Study 3, multiple imputations by chain equation (MICE) were applied to handle the missingness using miceforest library in Python and the MICE package in R under the assumption of missing at random. Instead of using the standard iterations, the two studies generated and compared convergence plots to determine the suitable numbers of iterations based on the datasets. The handling of missing data in the BMI category variable was approached with caution. Given the high proportion of missing data, using the MI approach, would not be valid under this large proportion of missing. Therefore, missingness was categorized separately and used in the ML prediction model to investigate further the factors associated with the unknown BMI.

In Study 1, feature extraction was employed to compress data through FAMD to reduce the dimensionality. This factorial analysis was able to compress both continuous and categorical data. To retain the most information, 80% of the explained variance ratio was maintained, followed by component transformation. It is also known that choosing the number of k in the k-means process is challenging [96,99]. Rather than relying on a single evaluation method, the elbow method, Silhouette coefficient, and DB index were used as the main evaluation methods. The majority agreement from all three methods determined the final number of k. The k-means clustering was performed in two trajectories to understand student profiles with and without substance use. Moreover, multinomial LR and RF were built using the k-means clusters as the outcome to understand the strength of the association between the important factors on the student's behaviours within each cluster. In Study 2, rather than using a single preprocessing strategy, this study compared the model performance based on the F1 score among six ML classifiers in the dataset with the top 50 important features, the dataset with all variables, and the dataset with LDA processed before selecting the best-performing model before fit on the testing dataset.

In Study 3, feature selection was applied from baseline to Year 3. To determine the number of variables required for the analysis, each year's variables were evaluated based on the ones that achieved higher F1 and ROC-AUC scores. Selecting the more relevant variables by training XGBoost model allowed a simpler model and reduced the issues of overfitting [189]. Study 3 also tested different combinations of hidden layers in RNN and LSTM to design the network model structure. To further enhance model performance, an attention-based layer was added to instruct the model to focus on the most relevant inputs.

Lastly, the ML models are known for their complexity and the nature of being a 'black box' [190]. SHAP analysis was conducted on the XGBoost model in Study 2 and the two deep learning models in Study 3 to improve model interpretability. This process explained the reasons

behind the model prediction on each weight status and weight perception by viewing the SHAP summary plots and SHAP waterfall for individual observations. This approach can also assist non-technical professionals in interpreting the ML results for their decision-making.

The rigorous process of handlining missing data and applying various techniques in data preprocessing steps ensured the training datasets were well-represented before the ML modelling. Additionally, the visualizations of the ML model's results improve interpretation through SHAP plots, allowing users to better understand the meaning behind model predictions. All the ML preprocessing steps have built credibility and confidence in optimizing the models.

7.5 Limitations

Input Data: Throughout the studies, the nature of the datasets was self-reported student survey data. Self-reported weight and height, which are used to calculate BMI, contained higher variabilities and underestimated the prevalence of youth obesity compared to measured BMI [191,192]. For weight-related subjective measures, overweight youth tended to underreport their weight. Due to the social impression of the "ideal" body image, females are more likely to underreport their weight, while males are more likely to overreport their height [193]. A similar pattern has been observed in weight perception, as misperception of body weight is common in adolescents, especially among overweight and obese youth [194]. It is also acknowledged that self-reported BMI is influenced by some weight perception since students were asked to report their weight without shoes. However, the correlation (measured between 0 and 1) test result from Study 2 showed a weak correlation based on the Cramer's V score (0.25). It would be beneficial to incorporate objective measures such as weight and height to validate the self-reported data results. Although measured BMI is reliable, it is usually not feasible to collect data from a large student population or at the national level due to time constraints and required resources [47]. A mitigation plan for addressing the bias in the self-reported data would involve using another cohort study that contains objective BMI measurement and performing sensitivity analysis to see how much change between self-reported and objectively measured BMI. The details of how to approach this plan are in the Future Work section.

Moreover, parental data such as parents' eating behaviour, weight or maternal-level factors were not collected from the COMPASS data. The influence of parent-related data is a representative indicator of the youth's mental and physical health, affecting students' behaviour formation related to obesity [26, 27]. ML studies focusing on youth obesity that achieved robust model performance included parental-level information such as parents' education level and occupation [195, 196, 197]. Missing these representative factors could result in a lack of understanding of how family members' behaviours influence health-related perspectives among youths. In addition to the lack of family-related data, input data related to watching TV and

surfing the internet does not specify whether these activities relate to social media use. Data collection on “browsing and scrolling through social media” only started in the 2021-22 survey. This could lead to undefined information on whether social media activities impacted health-related behaviours among youth in the previous cohort year.

The input data was also subject to inconsistencies in data collection. The more comprehensive gender and racial and ethnic categories were only implemented starting in the 2022-23 cohort, but not in earlier years like the 2020-21 cohort, resulting in gaps in the linked data. Due to the different distribution approach in Quebec relative to the other provinces, only 50% of their students received weight perception questions in Year 3. This inconsistency has resulted in missing data and reduced sample sizes within the linked dataset. Additionally, the results of the ML model apply only to COMPASS participants. The generalizability of the ML models in all three studies needs to be validated using other years of COMPASS datasets. For example, the first study employed only the CCA approach, and the clustering result needs to be compared with another pre-COVID cohort or post-COVID cohort. Since only certain provinces are included in the COMPASS study for the survey years used in this thesis (Alberta, British Columbia, Ontario, and Quebec), the result of the ML model may not represent other geographic locations in Canada that have not participated in the COMPASS study. Due to the passive consent model, some of the school boards, including Waterloo, Toronto and Peel, were also not included in the study, resulting in the loss of some key information on the students from these regions.

Lastly, this study lacks stratification by subgroups, such as gender-minorities, and comparisons across different racial and ethnic groups. Due to the small sample sizes in some categories and the potential risk of overfitting, stratification by these subgroups was not performed. For example, gender minorities accounted for only 3.8% of the overall population, and all of them had no BMI. Similarly, the gender minorities represented 2.3% of the overall sample in Study 3. Regarding ethnicity, most participants were White, comprising 79% in Study 2 and 72% in Study 3. Given these distributions, stratification could have compromised the statistical power and increased the risk of overfitting certain subgroups using the ML approach. As an experimental observation, Study 2 revealed the difference among the critical factors associated with BMI prediction in males and females, given these two groups' comparable sample sizes. Study 3 included gender and ethnicity as covariates to account for their effects associated with Year 2 weight perception and the change of weight perception over time. Based on Study 2’s experiment on BMI prediction among females and males, we observed some differences between the associated factors. The boxplot for the unknown BMI also showed that gender minorities had the highest positive SHAP value in this category, indicating it could be worth exploring what other factors are critical. However, due to the extremely small sample sizes, the BMI prediction focusing on gender minorities was not performed.

The limited representation of racial and ethnic groups due to the extremely imbalanced samples (White vs. non-White) could result in a lack of understanding of how cultural background affects BMI or weight perception. A study analyzing NHANES data from 1999 to 2016 found that among children aged 2 to 19, White and Asian children had significantly lower obesity prevalence compared to their Hispanic, African or other race peers [256]. In a Canadian study, the prevalence of obesity in children and youth is also higher among the Aboriginal population (First Nations off-reserve, Inuit and Métis) [257]. According to the self-reported data from the Aboriginal People Survey in 2006 among aged 6 to 14 years participants, the obesity prevalence was high, resulting in 16.9% in Métis, 20.0% in the off-reserve First Nations and 25.6% among Inuit. These statistics show that racial and ethnic groups are important and should be considered in terms of the prevalence of overweight and obesity among youth populations. Besides BMI, one study explored the misperception of weight among adolescents, measuring their objective weight and weight perceptions. It illustrated that African Americans had a higher probability of underestimating their weight relative to Whites. This effect was particularly observed among overweight and obese girls and boys [258]. Despite the limitation of the majority of the White populations in the COMPASS survey, the data collection efforts, if possible, should emphasize recruitment to increase samples from other racial and ethnic groups in order to address potential cultural differences or other factors that influence the BMI status.

ML model perspective: Although k-means clustering is flexible and efficient in computation, k-means clustering performance is sensitive to complex data structures, especially outliers. Since it randomly selects the initial centroid and only detects spherical clusters, the quality of clusters could be questioned [115, 199]. Study 1 implemented several evaluation methods to select the initial k and applied FAMD to reduce data dimensionality. However, heterogeneity was still detected within the cluster, meaning some sub-clusters were not discernible. It was suspected that multiple aspects of student's behaviours created difficulties for the k-means clustering to detect well-separated clusters due to the complexity, and each student may have various responses in healthy eating, physical activity, screen time or school engagement. The student's substance level was the most distinguished factor, which clearly separated the high-level clusters.

Another limitation in Study 2 is the impact of missingness on BMI status. Although missing was categorized as an unknown category, achieving promising predictive power may not be practical due to the complex nature of missingness in BMI. BMI can be either Missing at Random (MAR) or Missing not at Random (MNAR). MAR refers to missing data that depends on the observed data, while MNAR relates to the observed data and the missing values themselves [200]. Even though multiple imputation is valid under the assumption of MAR, the missing pattern of BMI could be MNAR [201]. For instance, participants with missing BMI tend

to have higher BMI [201]. Therefore, the model's performance in predicting the missingness in BMI as a separate category is challenging, and a more advanced model, such as semi-supervised learning, would be needed to tackle this issue. The details of how to approach this plan are in the Future Work section.

The last limitation of the ML approach is the use of temporal sequence models like LSTM. LSTM is designed for long sequential prediction with multiple time steps [133]. The data in Study 3 only contained two time steps, which may not fully capture the association of predictors and weight perception over time. For future research, it would be worth considering using four-year linked data and observing whether BMI or other important factors found in Study 3 are still the important variables in weight perception over time. However, steps beyond four may not be practical for the COMPASS data system as students would graduate after Grade 12. For future studies, the LSTM or RNN model should include static variables such as demographics to provide a more nuanced understanding of the effects on weight perception over time. Stratification among gender, racial and ethnic groups should also be performed to capture the dynamic change in weight perception over time.

7.6 Contribution to Public Health

This dissertation also adds novelty to the current public health research on youth behaviour profiles, childhood obesity, and the transition of youth weight perception over time. Each study's result is derived from the previous one. Study 1 revealed the key factors associated with clustering before and after COVID-19, adding more depth to youth behaviour research and filling the gaps in understanding the differences and similarities of student's characteristics before and after the pandemic. Compared with the ML studies in childhood obesity in the literature review, none have included the unknown BMI as a separate outcome class for building the predictive model. In Study 2, the important variables such as weight perception and gender associated with unknown BMI would provide critical messages for public health professionals to examine the association between the factors and missing BMI. To the best knowledge, Study 3 is the first to predict future youth weight perception and account for the temporal effect of the previous year's predictors. The results of all three studies disseminate key messages to public health professionals and various stakeholders, informing them of early intervention plans for body dissatisfaction at school and improving youth self-esteem and body acceptance, which can lead to healthy dieting and increased in physical activity.

Study 1 revealed that, regardless of the pandemic, substance use remained the most distinctive variable among the youth population. Although the proportion of students was relatively small compared to the rest of the population, it is necessary for the school principals to address this behaviour. A previous study on risk profiles of substance uses among COMPASS

students emphasized the likelihood of transitioning from single use of substances to higher-level use [155]. It appears that students with moderate risk of substance use could have a tendency towards high risk, highlighting the need for school engagement programs. Before quarantine, students were clustered based on their academic performance, school connectedness and peer relationships. This further suggests that school-based intervention could be effective since students spend most of their time at school.

Study 2 and Study 3 results show that BMI and weight perception are bi-directional factors, meaning how students perceive their weight status and their self-reported BMI affect each other interchangeably. Rather than focusing solely on single behavioural factors such as nutritional eating or physical activity levels, it is suggested that reducing body dissatisfaction may be a critical step for establishing healthy weight management. From the transitional probabilities, most students intended to maintain the same weight perception as in the previous state. Thus, it could be beneficial to implement positive social media programs that promote self-acceptance of body image starting in Grade 9. Moreover, initiating a school-based curriculum on accurately perceiving weight status and self-measuring weight and height could help reduce weight misperception. Targeting misconceptions of weight status is necessary to prevent eating disorders and improve physical activity engagement. Addressing over- and underreporting could enable more targeted intervention programs to prevent childhood obesity.

For public health policymakers, stakeholders from The Public Health Agency of Canada and school principals as the potential audience, this thesis addresses the limitation of ML by providing inoperable plots to justify the associated factors and help support their initiatives. It would also be valuable for health policymakers to understand the key aspects of targeting current public health campaigns and compare these with the current policy focus on weight-related behaviour. For public health professionals, this thesis offers an overall view of the most important factors in predicting BMI and weight change intention, providing information for their decision-making in refining public health and intervention strategies, considering national or provincial initiatives to reduce weight misperception. Given that youth frequently use smartphones, promoting self-esteem through social media campaigns could be worth exploring. For school principals, the factors ranked in order of importance for BMI and weight perception prediction can inform whether school activities focused on improving adolescent health are effective and guide necessary adjustments. This can help ensure that interventions effectively address the needs of the student population within schools.

7.7 Contribution to ML Research

The three studies have demonstrated the potential of utilizing ML modelling in student-level survey data to research youth behaviour-related public health concerns. While applying

advanced techniques is feasible, building a high-quality training dataset is crucial regardless of whether a supervised or unsupervised learning model is used. Future research using self-reported data should proceed with caution. The survey questions present in one year may not exist in another province or be consistent across different years, and high proportions of missing certain variables (e.g., BMI) may require a combination of techniques to handle them. Besides internal validation to prevent model overfitting, such as cross-validation, researchers are encouraged to compare the model's performance with that of other cohorts to evaluate its feasibility.

It would also be reasonable to conduct stratified analyses among provinces participating in the COMPASS study to account for different effects on model prediction in BMI and weight perception. This approach allows for a more targeted early intervention for various youth populations. In summary, the ML approach is a powerful tool. However, due to the complexity of youth obesity and the potential bias in self-reported data, multiple ML models approach and validation strategies should be employed to ensure generalizability. Regarding a meaningful message to the ML data scientist as the audience, the strategies utilized in this thesis, such as resampling to overcome the imbalances, handlining missing data through multiple imputations, improving model performance by experimenting with optimal training datasets and neural networks architecture, serve as valuable resources for future ML analytic projects in adolescent health.

7.8 Future Work

From the data collection perspective, provincial leaders, public health, and medical professionals should work together to link measured BMI with the self-reported level data. National-level efforts, such as increasing funding for data surveillance, would benefit the inclusion of the family-level or home environment data. Including the parental diet behaviour and information allows for the investigation of more potential factors contributing to childhood obesity. With the measured BMI, a sensitive analysis can be performed to evaluate intervention efforts in changing BMI among different age groups. Survey data collection can also be standardized to enhance consistency across distributed questions and reduce missingness.

From the ML perspective, there are several mitigations for improving future related studies on adolescent health. Regarding future research using k-means clustering, the below steps can be used to improve homogeneity within the clustering. 1) Retain only relevant variables from each behavioural measure. For example, remove detailed questions on the mode of use and opinions on the risk of marijuana and keep only whether participants have used marijuana in the past days when assessing substance use behaviour. 2) Establish clusters within each behavioural measure and select variables that better represent each measure to overcome the variance and

noise on k-means. This would help overcome student heterogeneity and improve the cluster quality by focusing on one main theme at a time. For instance, physical activities are used as the main theme, and variables are chosen based on the sedentary lifestyle indicators. 3) Utilize more advanced clustering techniques to reduce the effect of outliers, such as density-based or fuzzy clustering [115, 155].

Future research addressing the bias in self-reported unknown BMI, due to the high proportion of missing data, could use advanced ML models like semi-supervised learning models. Example steps include 1) Apply supervised learning to train the model with known BMI label, 2) Use the trained model with labelled data to predict the missing category in BMI, 3) Combine the training dataset of the labelled data with the predicted unlabelled data, 4) Retrain the combined dataset through another classifier and compare the model performance with the baseline model using only labelled data.

Additionally, benchmarking against another study that uses objective measures for sensitivity analysis, or employing inverse probability weighting, could help address bias in self-reported data. For example, Model 1 could use self-reported BMI, and Model 2's BMI could be adjusted based on objective measures. Suppose the average BMI in the objective measure was higher in females by 2-3 kg. Then, Model 2 could add 2 kg to each observation and explore whether the factors associated with BMI prediction remain the same. Missingness probabilities can also be calculated by categorizing missing as 0 and not missing as 1. Using a logistic regression model, these probabilities can be calculated, and lower weights can be assigned to observations with lower probabilities of missingness. Implementing these weights in the ML model ensures the model gives more attention to data with a lower probability of being missing.

For future longitudinal research in time series analysis, four-year linked datasets could be used to capture changes in factors related to students' weight perception. ML efforts could also be applied to compare the effect of interventions on weight perception or BMI before and after implementation. In addition to time series analysis, stratification by gender, racial and ethnic groups should be conducted if more samples from minority groups can be collected. Observing the different effects within these subgroups while considering cultural differences is essential.

7.9 Conclusion

This thesis employed a rigorous process using unsupervised and supervised learning ML models. Utilizing the COMPASS dataset, the predictive model explored multiple perspectives of student-level behaviours, focusing on BMI and weight perception. Building the ML predictive model cross-sectionally and longitudinally enhances current adolescent research on BMI prevalence and weight misperception by considering various behavioural domains among

students. The research has provided valuable insights for both the public health and ML communities on adolescent studies. This thesis also acknowledges the limitations of self-reported data, including the impact of missing BMI on ML models and bias in data, while outlining directions for further enhancement of advanced ML models and more exploratory analyses on certain subgroups to mitigate bias. This thesis has demonstrated the potential of implementing ML analytics to support public health initiatives and inform stakeholders to focus on more critical factors to build targeted interventions to improve adolescent health, ultimately contributing to a healthier future for younger generations.

References

1. Salam RA, Das JK, Lassi ZS, Bhutta ZA. Adolescent health and well-being: Background and methodology for review of potential interventions. Journal of adolescent health. 2016 Oct 1;59(4):S4-10. doi: 10.1016/j.jadohealth.2016.07.023

2. Gohari MR, Patte KA, Ferro MA, Haddad S, Wade TJ, Bélanger RE, Romano I, Leatherdale ST. Adolescents' Depression and Anxiety Symptoms During the COVID-19 Pandemic: Longitudinal Evidence From COMPASS. Journal of Adolescent Health. 2023 Sep 29.

3. Liu F, Yang D, Liu Y, Zhang Q, Chen S, Li W, Ren J, Tian X, Wang X. Use of latent profile analysis and k-means clustering to identify student anxiety profiles. BMC psychiatry. 2022 Dec;22(1):1-1.

4. Lobstein T., Baur L., Uauy R. Obesity in children and young people: A crisis in public health. Obes Rev. 2004;5(Suppl. 1):4–85.

5. Popkin BM, Gordon-Larsen P. The nutrition transition: worldwide obesity dynamics and their determinants. International journal of obesity. 2004 Nov;28(3):S2-9.

6. Murali V, Oyebode F. Poverty, social inequality and mental health. Advances in psychiatric treatment. 2004 May;10(3):216-24.

7. Pompili M, Serafini G, Innamorati M, Dominici G, Ferracuti S, Kotzalidis GD, Serra G, Girardi P, Janiri L, Tatarelli R, Sher L. Suicidal behavior and alcohol abuse. International journal of environmental research and public health. 2010 Apr;7(4):1392-431.

8. Kim NH, Lee JM, Yoo E. How the COVID-19 pandemic has changed adolescent health: physical activity, sleep, obesity, and mental health. International Journal of Environmental Research and Public Health. 2022 Jul 28;19(15):9224.

9. Yarger HA, Shariq D, Hickey AC, Giacobbe E, Dziura SL, Redcay E. Examining Adolescents' Mental Health Before and During the COVID-19 Pandemic. Merrill-Palmer quarterly (Wayne State University. Press). 2023 Jan;69(1):1.

10. Pedrini L, Meloni S, Lanfredi M, Ferrari C, Geviti A, Cattaneo A, Rossi R. Adolescents' mental health and maladaptive behaviors before the Covid-19 pandemic and 1-year after: analysis of trajectories over time and associated factors. Child and adolescent psychiatry and mental health. 2022 Jun 10;16(1):42.

11. Zhou J, Yuan X, Qi H, Liu R, Li Y, Huang H, et al. Prevalence of depression and its correlative factors among female adolescents in China during the coronavirus disease 2019 outbreak. Global Health. 2020;16(1):1–6.

12. Odriozola-González P, Planchuelo-Gómez Á, Irurtia MJ, de Luis-García R. Psychological symptoms of the outbreak of the COVID-19 confinement in Spain. J Health Psychol. 2020;27(4):825–35.

13. Paterson DC, Ramage K, Moore SA, Riazi N, Tremblay MS, Faulkner G. Exploring the impact of COVID-19 on the movement behaviors of children and youth: A scoping review of evidence after the first year. Journal of sport and health science. 2021 Dec 1;10(6):675-89.

14. Brakspear L, Boules D, Nicholls D, Burmester V. The impact of COVID-19-related living restrictions on eating behaviours in children and adolescents: a systematic review. Nutrients. 2022 Sep 4;14(17):3657.

15. Horikawa C, Murayama N, Kojima Y, Tanaka H, Morisaki N. Changes in selected food groups consumption and quality of meals in Japanese school children during the COVID-19 pandemic. Nutrients. 2021 Aug 10;13(8):2743.

16. Al Hourani H, Alkhatib B, Abdullah M. Impact of COVID-19 lockdown on body weight, eating habits, and physical activity of Jordanian children and adolescents. Disaster medicine and public health preparedness. 2022 Oct;16(5):1855-63.

17. World Health Organization. Obesity and overweight. Available from: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight. Accessed June 15, 2024

18. Pourghazi F, Eslami M, Ehsani A, Ejtahed HS, Qorbani M. Eating habits of children and adolescents during the COVID-19 era: A systematic review. Frontiers in nutrition. 2022 Oct 18;9:1004953.

19. Statistics Canada. Archived - Children's body mass index - World Health Organization classification system. 2018. Available from: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310032201&pickMembers%5B0%5D=3.1&cubeTimeFrame.startYear=2011&cubeTimeFrame.endYear=2015&referencePeriods=20110101%2C20150101. Accessed January 18, 2024

20. Sahoo K, Sahoo B, Choudhury AK, Sofi NY, Kumar R, Bhadoria AS. Childhood obesity: causes and consequences. Journal of family medicine and primary care. 2015 Apr 1;4(2):187-92.

21. Marcos-Pasero H, Colmenarejo G, Aguilar-Aguilar E, Ramírez de Molina A, Reglero G, Loria-Kohen V. Ranking of a wide multidomain set of predictor variables of children obesity by machine learning variable importance techniques. Scientific Reports. 2021 Jan 21;11(1):1910.

22. Janssen I. The public health burden of obesity in Canada. Canadian journal of diabetes. 2013 Apr 1;37(2):90-6.
23. Rueda-Clausen CF, Poddar M, Lear SA, Poirier P, Sharma AM. Canadian Adult Obesity Clinical Practice Guidelines: Assessment of People Living with Obesity. Available from: https://obesitycanada.ca/guidelines/assessment. Accessed June 19, 2024.

24. Rao DP, Kropac E, Do MT, Roberts KC, Jayaraman GC. Childhood overweight and obesity trends in Canada. Health promotion and chronic disease prevention in Canada: research, policy and practice. 2016 Sep;36(9):194.

25. Romieu I, Dossus L, Barquera S, Blottière HM, Franks PW, Gunter M, Hwalla N, Hursting SD, Leitzmann M, Margetts B, Nishida C. Energy balance and obesity: what are the main drivers?. Cancer causes & control. 2017 Mar;28:247-58.

26. Karnik S, Kanekar A. Childhood obesity: a global public health crisis. Int J Prev Med. 2012 Jan 1;3(1):1-7.

27. Kansra AR, Lakkunarajah S, Jay MS. Childhood and adolescent obesity: a review. Frontiers in pediatrics. 2021 Jan 12;8:866.

28. Lowe CJ, Morton JB, Reichelt AC. Adolescent obesity and dietary decision making—a brain-health perspective. The Lancet Child & Adolescent Health. 2020 May 1;4(5):388-96.Anderson PM, Butcher KE. Childhood obesity: Trends and potential causes. *Future Child.* 2006;16:19–45.

29. Anderson PM, Butcher KF. Childhood obesity: trends and potential causes. The Future of children. 2006 Apr 1:19-45.

30. Kapil U, Bhadoria AS. Television viewing and overweight and obesity amongst children. Biomed J. 2014 Sep 1;37(5):337-8.

31. Lee A, Cardel M, Donahoo WT. Social and environmental factors influencing obesity. Endotext [Internet]. 2019 Oct 12.

32. Hermans RC, De Bruin H, Larsen JK, Mensink F, Hoek AC. Adolescents' responses to a school-based prevention program promoting healthy eating at school. Frontiers in Public Health. 2017 Nov 20;5:309.

33. Goncalves VS, Figueiredo AC, Silva SA, Silva SU, Ronca DB, Dutra ES, Carvalho KM. The food environment in schools and their immediate vicinities associated with excess weight in adolescence: A systematic review and meta-analysis. Health & Place. 2021 Sep 1;71:102664.

34. Shin A, Nam CM. Weight perception and its association with socio-demographic and health-related factors among Korean adolescents. BMC public health. 2015 Dec;15:1-9.
35. Côté M, Legendre M, Aimé A, Brault MC, Dion J, Bégin C. The paths to children's disordered eating: The implications of BMI, weight-related victimization, body dissatisfaction and parents' disordered eating. Clinical Psychology in Europe. 2020 Mar;2(1).

36. Rawana JS, Morgan AS, Nguyen H, Craig SG. The relation between eating-and weight-related disturbances and depression in adolescence: A review. Clinical child and family psychology review. 2010 Sep;13:213-30.

37. Moradi M, Mozaffari H, Askari M, Azadbakht L. Association between overweight/obesity with depression, anxiety, low self-esteem, and body dissatisfaction in children and adolescents: a systematic review and meta-analysis of observational studies. Critical Reviews in Food Science and Nutrition. 2021 Dec 31;62(2):555-70.

38. Wardle J, Waller J, Fox E. Age of onset and body dissatisfaction in obesity. Addictive behaviors. 2002 Jul 1;27(4):561-73.

39. Zare S, Thomsen MR, Nayga Jr RM, Goudie A. Use of machine learning to determine the information value of a BMI screening program. American journal of preventive medicine. 2021 Mar 1;60(3):425-33.

40. Santina T, Beaulieu D, Gagné C, Guillaumie L. Tackling childhood obesity through a school-based physical activity programme: a cluster randomised trial. International Journal of Sport and Exercise Psychology. 2021 May 4;19(3):342-58.

41. Critch JN. School nutrition: Support for providing healthy food and beverage choices in schools. Paediatrics & Child Health. 2020 Feb 6;25(1):33-8.

42. Berge JM. A review of familial correlates of child and adolescent obesity: what has the 21st century taught us so far?. International journal of adolescent medicine and health. 2009 Oct;21(4):457-84.

43. Kitzmann KM, Beech BM. Family-based interventions for pediatric obesity: methodological and conceptual challenges from family psychology.

44. Chatterjee K, Jha U, Kumari P, Chatterjee D. Early prediction of childhood obesity using machine learning techniques. InAdvances in Communication and Computational Technology: Select Proceedings of ICACCT 2019 2021 (pp. 1431-1440). Springer Singapore.

45. Berge JM, Everts JC. Family-based interventions targeting childhood obesity: a meta-analysis. Childhood Obesity. 2011 Apr 1;7(2):110-21.

46. Ben-Sefer E, Ben-Natan M, Ehrenfeld M. Childhood obesity: current literature, policy and implications for practice. International nursing review. 2009 Jun;56(2):166-73.

47. Allison C, Colby S, Opoku-Acheampong A, Kidd T, Kattelmann K, Olfert MD, Zhou W. Accuracy of self-reported BMI using objective measurement in high school students. Journal of nutritional science. 2020 Jan;9:e35.

48. Leatherdale ST, Brown KS, Carson V, Childs RA, Dubin JA, Elliott SJ, Faulkner G, Hammond D, Manske S, Sabiston CM, Laxer RE. The COMPASS study: a longitudinal hierarchical research platform for evaluating natural experiments related to changes in school-level programs, policies and built environment resources. BMC public health. 2014 Dec;14:1-7.

49. The COMPASS System. Leatherdale ST (2022). https://uwaterloo.ca/compass-system/ Accessed Dec 15, 2023.

50. Daniels KM, Schinasi LH, Auchincloss AH, Forrest CB, Roux AV. The built and social neighborhood environment and child obesity: A systematic review of longitudinal studies. Preventive medicine. 2021 Dec 1;153:106790.

51. Jia P, Shi Y, Jiang Q, Dai S, Yu B, Yang S, Qiu G, Yang S. Environmental determinants of childhood obesity: a meta-analysis. The Lancet Global Health. 2023 Mar 1;11:S7.

52. Onis MD, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J. Development of a WHO growth reference for school-aged children and adolescents. Bulletin of the World health Organization. 2007 Sep;85(9):660-7. doi: 10.2471/blt.07.043497

53. García-Moya I, Bunn F, Jiménez-Iglesias A, Paniagua C, Brooks FM. The conceptualisation of school and teacher connectedness in adolescent research: a scoping review of literature. Educational review. 2019 Jul 4;71(4):423-44. doi: 10.1080/00131911.2018

54. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. Archives of internal medicine. 2006 May 22;166(10):1092-7. doi: 10.1001/archinte.166.10.1092

55. Diener E, Wirtz D, Tov W, Kim-Prieto C, Choi D, Oishi S, et al. New well-being measures: short scales to assess flourishing and positive and negative feelings. Soc Indic Res. 2010;97(2):143–56. https://doi.org/10.1007/s11205-009-9493-y.

56. Ahluwalia N. Nutrition monitoring of children aged birth to 24 mo (B-24): Data collection and findings from the NHANES. Advances in Nutrition. 2020 Jan 1;11(1):113-27.

57. Johnson CL, Dohrmann SM, Burt VL, Mohadjer LK. National health and nutrition examination survey: sample design, 2011-2014. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics; 2014.

58. National Health and Nutrition Examination Survey. About the National Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm Accessed: June 20, 2024.

59. Smith SS. NCHS Dataline. Public Health Reports. 2015 Mar 1:176-8.

60. Zipf G, Chiappa M, Porter KS, Ostchega Y, Lewis BG, Dostal J. Health and nutrition examination survey plan and operations, 1999-2010.

61. National Health and Nutrition Examination Survey. NHANES 2019-2020 Questionnaire Instruments. Accessed: June 20, 2024. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/questionnaires.aspx?BeginYear=2019

62. NHANES 2017-March 2020 Pre-Pandemic Examination Data. Accessed June 20, 2024. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination

63. NHANES 2017-March 2020 Pre-Pandemic Laboratory Data. Accessed June 20, 2024. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory

64. National Health and Nutrition Examination Survey 2017-March 2020 Data Documentation, Codebook, and Frequencies. Accessed June 20, 2024. https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/P_BMX.htm#Protocol_and_Procedure

65. National Center for Health Statistics. Percentile Data Files with LMS Values. Accessed: June 20, 2024. https://www.cdc.gov/growthcharts/percentile_data_files.htm

66. Ford ES, Giles WH, Dietz WH. Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey. Jama. 2002 Jan 16;287(3):356-9.

67. National Health and Nutrition Examination Survey. Interviewer Procedures Manual. Accessed: June 20, 2024. https://wwwn.cdc.gov/nchs/data/nhanes/2013-2014/manuals/intrvwr_proc_manual.pdf.

68. National Academies of Sciences, Medicine Division, Committee on Evaluating Approaches to Assessing Prevalence, Trends in Obesity. Assessing prevalence and trends in obesity: navigating the evidence. National Academies Press; 2016 Oct 30.

69. Mondal PK, Foysal KH, Norman BA, Gittner LS. Predicting childhood obesity based on single and multiple well-child visit data using machine learning classifiers. Sensors. 2023 Jan 9;23(2):759.

70. Ziauddeen N, Roderick PJ, Macklon NS, Alwan NA. Predicting childhood overweight and obesity using maternal and early life risk factors: a systematic review. Obesity Reviews. 2018 Mar;19(3):302-12.

71. Heerman WJ, Sommer EC, Slaughter JC, Samuels LR, Martin NC, Barkin SL. Predicting early emergence of childhood obesity in underserved preschoolers. The Journal of pediatrics. 2019 Oct 1;213:115-20.

72. Weir CB, Jan A. BMI Classification Percentile and Cut Off Points. StatPearls. 2022. Treasure Island. FL. 2023.

73. Scott-Sheldon LA, Hedges LV, Cyr C, Young-Hyman D, Khan LK, Magnus M, King H, Arteaga S, Cawley J, Economos CD, Haire-Joshu D. Childhood obesity evidence base project: a systematic review and meta-analysis of a new taxonomy of intervention components to improve weight status in children 2–5 years of age, 2005–2019. Childhood Obesity. 2020 Sep 1;16(S2):S2-21.

74. Ardic C, Uzun K, Karakullukcu A, Karakullukcu S. Changes in body mass index-z scores in 3-year-old children during the COVID-19 pandemic: a 2-year retrospective cohort study. Revista da Associação Médica Brasileira. 2023 Sep 25;69:e20230336.

75. Colmenarejo G. Machine learning models to predict childhood and adolescent obesity: a review. Nutrients. 2020 Aug 16;12(8):2466.

76. Madigan CD, Graham HE, Sturgiss E, Kettle VE, Gokal K, Biddle G, Taylor GM, Daley AJ. Effectiveness of weight management interventions for adults delivered in primary care: systematic review and meta-analysis of randomised controlled trials. Bmj. 2022 May 30;377.

77. Llargues E, Franco R, Recasens A, Nadal A, Vila M, Pérez MJ, Manresa JM, Recasens I, Salvador G, Serra J, Roure E. Assessment of a school-based intervention

78. Gortmaker SL, Polacsek M, Letourneau L, Rogers VW, Holmberg R, Lombard KA, Fanburg J, Ware J, Orr J. Evaluation of a primary care intervention on body mass index: the Maine Youth Overweight Collaborative.

79. Ranganathan P, Pramesh CS, Aggarwal R. Common pitfalls in statistical analysis: logistic regression. Perspectives in clinical research. 2017 Jul 1;8(3):148-51.

80. Bayman EO, Dexter F. Multicollinearity in logistic regression models. Anesthesia & Analgesia. 2021 Aug 1;133(2):362-5.

81. Nahhas RW. Introduction to Regression Methods for Public Health Using R. Dayton: Creative Common. 2023.

82. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology. 1996 Dec 1;49(12):1373-9.

83. Pang X, Forrest CB, Lê-Scherban F, Masino AJ. Prediction of early childhood obesity with machine learning and electronic health record data. International journal of medical informatics. 2021 Jun 1;150:104454.

84. Stoltzfus JC. Logistic regression: a brief primer. Academic emergency medicine. 2011 Oct;18(10):1099-104.

85. Aggarwal R, Ranganathan P. Common pitfalls in statistical analysis: Linear regression analysis. Perspectives in clinical research. 2017 Apr 1;8(2):100-2.

86. Simple Linear regression. Correlation and Regression with R. Accessed on June 21, 2024. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html#:~:text=There%20are%20four%20assumptions%20associated,are%20independent%20of%20each%20other.

87. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosomatic medicine. 2004 May 1;66(3):411-21.

88. Hawkins DM. The problem of overfitting. Journal of chemical information and computer sciences. 2004 Jan 26;44(1):1-2.

89. Bennett M, Kleczyk EJ, Hayes K, Mehta R. Evaluating similarities and differences between machine learning and traditional statistical modeling in healthcare analytics. InArtificial Intelligence Annual Volume 2022 2022 May 31. IntechOpen.

90. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science. 2001;16(3):199-231

91. Rajula HS, Verlato G, Manchia M, Antonucci N, Fanos V. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. Medicina. 2020 Sep 8;56(9):455.

92. Bzdok D, Krzywinski M, Altman N. Machine learning: a primer. Nature methods. 2017 Nov 11;14(12):1119.

93. Sarker IH. Machine learning: Algorithms, real-world applications and research directions. SN computer science. 2021 May;2(3):160.

94. Bhattacharya M, Ehrenthal D, Shatkay H. Identifying growth-patterns in children by applying cluster analysis to electronic medical records. In2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2014 Nov 2 (pp. 348-351). IEEE.

95. Chaudhry M, Shafi I, Mahnoor M, Vargas DL, Thompson EB, Ashraf I. A systematic literature review on identifying patterns using unsupervised clustering algorithms: A data mining perspective. Symmetry. 2023 Aug 31;15(9):1679.

96. Ashabi A, Sahibuddin SB, Salkhordeh Haghighi M. The systematic review of K-means clustering algorithm. InProceedings of the 2020 9th International Conference on Networks, Communication and Computing 2020 Dec 18 (pp. 13-18).

97. Paul E. Black, "Euclidean distance", in *Dictionary of Algorithms and Data Structures* [online], Paul E. Black, ed. 17 December 2004. (accessed TODAY) Available from: https://www.nist.gov/dads/HTML/euclidndstnc.html. Accessed September 14, 2024.

98. Nigro L. Performance of parallel K-means algorithms in Java. Algorithms. 2022 Mar 29;15(4):117.

99. Parlett-Pelleriti CM, Stevens E, Dixon D, Linstead EJ. Applications of unsupervised machine learning in autism spectrum disorder research: a review. Review Journal of Autism and Developmental Disorders. 2023 Sep;10(3):406-21.

100. GeeksforGeeks. CalinskiHarabasz Index in R [Internet]. GeeksforGeeks. Available from: https://www.geeksforgeeks.org/calinski-harabasz-index-in-r/. Accessed on September 15, 2024.

101. Umargono E, Suseno JE, Gunawan S. K-Means clustering optimization using the elbow method and early centroid determination based-on mean and median. InProceedings of the International Conferences on Information System and Technology 2019 Mar (pp. 234-240). Setubal, Portugal: SCITEPRESS—Science and Technology Publications.

102. Shutaywi M, Kachouie NN. Silhouette analysis for performance evaluation in machine learning with applications to clustering. Entropy. 2021 Jun 16;23(6):759.

103. Wu S, Chow TW. Clustering of the self-organizing map using a clustering validity index based on inter-cluster and intra-cluster density. Pattern Recognition. 2004 Feb 1;37(2):175-88.

104. Lletı R, Ortiz MC, Sarabia LA, Sánchez MS. Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes. Analytica Chimica Acta. 2004 Jul 5;515(1):87-100.

105. Kleine D. Detecting knee- / elbow points in a graph of a function [Internet]. Medium. Towards Data Science; 2021. Available from: https://towardsdatascience.com/detecting-knee-elbow-points-in-a-graph-d13fc517a63c#:~:text=The%20algorithm%20%E2%80%9CKneedle%E2%80%9D%20detects%20those. Accessed September 14, 2024.

106. Davies-Bouldin Index [Internet]. GeeksforGeeks. 2023. Available from: https://www.geeksforgeeks.org/davies-bouldin-index/. Accessed on June 22, 2024.

107. Machine Learning Evaluation Mastery: How to Use Davies-Bouldin Index and Dunn Index for Clustering Problems [Internet]. GPTutorPro. 2024. Available from: https://gpttutorpro.com/machine-learning-evaluation-mastery-how-to-use-davies-bouldin-index-and-dunn-index-for-clustering-problems/. Accessed on June 22, 2024.

108. Wang X, Xu Y. An improved index for clustering validation based on Silhouette index and Calinski-Harabasz index. InIOP Conference Series: Materials Science and Engineering 2019 Jul 1 (Vol. 569, No. 5, p. 052024). IOP Publishing.

109. K-means Cluster Analysis · UC Business Analytics R Programming Guide [Internet]. uc-r.github.io. Available from: https://uc-r.github.io/kmeans_clustering#:~:text=The%20gap%20statistic%20compares%20the,simulations%20of%20the%20sampling%20process. Accessed on September 15, 2024

110. Tran QN, Arabnia H, Mohammed EA, Naugler C, Far BH. Emerging trends in computational biology, bioinformatics and system biology. Waltham, MA: Elsevier Inc. 2015.

111. Quick V. Clustering of obesity-related risk behaviors among families with preschool children using a socioecological approach: cross-sectional study. JMIR Pediatrics and Parenting. 2018 Apr 25;1(1):e10320.

112. Orsoni M, Giovagnoli S, Garofalo S, Magri S, Benvenuti M, Mazzoni E, Benassi M. Preliminary evidence on machine learning approaches for clusterizing students' cognitive profile. Heliyon. 2023 Mar 1;9(3).

113. Lin Z, Feng W, Liu Y, Ma C, Arefan D, Zhou D, Cheng X, Yu J, Gao L, Du L, You H. Machine learning to identify metabolic subtypes of obesity: a multi-center study. Frontiers in Endocrinology. 2021 Jul 14;12:713592.

114. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Vigués M, Pujol-Ribera E, Valderas JM. Multimorbidity patterns with K-means nonhierarchical cluster analysis. BMC family practice. 2018 Dec;19:1-1.

115. Gao CX, Dwyer D, Zhu Y, Smith CL, Du L, Filia KM, Bayer J, Menssink JM, Wang T, Bergmeir C, Wood S. An overview of clustering methods with guidelines for application in mental health research. Psychiatry Research. 2023 May 27:115265.

116. Hung PD, Lien NT, Ngoc ND. Customer segmentation using hierarchical agglomerative clustering. InProceedings of the 2nd International Conference on Information Science and Systems 2019 Mar 16 (pp. 33-37).

117. Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview. *Sensors*, *20*(9), 2734. https://doi.org/10.3390/s20092734

118. Dugan TM, Mukhopadhyay S, Carroll A, Downs S. Machine learning techniques for prediction of early childhood obesity. Applied clinical informatics. 2015;6(03):506-20.

119. Selya AS, Anshutz D. Machine learning for the classification of obesity from dietary and physical activity patterns. Advanced Data Analytics in Health. 2018:77-97.120.

120. Jindal K, Baliyan N, Rana PS. Obesity prediction using ensemble machine learning approaches. InRecent Findings in Intelligent Computing Techniques: Proceedings of the 5th ICACNI 2017, Volume 2 2018 (pp. 355-362). Springer Singapore.

121. Zheng, Z.; Ruggiero, K. Using machine learning to predict obesity in high school students. In Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Kansas City, MO, USA, 13–16 November 2017; pp. 2132–2138.

122. Dunstan J, Aguirre M, Bastías M, Nau C, Glass TA, Tobar F. Predicting nationwide obesity from food sales using machine learning. Health informatics journal. 2020 Mar;26(1):652-63.123.

123. DeGregory KW, Kuiper P, DeSilvio T, Pleuss JD, Miller R, Roginski JW, Fisher CB, Harness D, Viswanath S, Heymsfield SB, Dungan I. A review of machine learning in obesity. Obesity reviews. 2018 May;19(5):668-85.124.

124. Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. Decision Analytics Journal. 2022 Jun 1;3:100071.125.

125. Ozturk Kiyak E, Ghasemkhani B, Birant D. High-Level K-Nearest Neighbors (HLKNN): A Supervised Machine Learning Model for Classification Analysis. Electronics. 2023 Sep 10;12(18):3828.

126. Wong JE, Yamaguchi M, Nishi N, Araki M, Wee LH. Predicting overweight and obesity status among Malaysian working adults with machine learning or logistic regression: retrospective comparison study. JMIR Formative Research. 2022 Dec 7;6(12):e40404.

127. Cheng X, Lin SY, Liu J, Liu S, Zhang J, Nie P, Fuemmeler BF, Wang Y, Xue H. Does physical activity predict obesity—A machine learning and statistical method-based analysis. International Journal of environmental research and public Health. 2021 Apr 9;18(8):3966.

128. Karamizadeh S, Abdullah SM, Halimi M, Shayan J, javad Rajabi M. Advantage and drawback of support vector machine functionality. In2014 international conference on computer, communications, and control technology (I4CT) 2014 Sep 2 (pp. 63-65). IEEE.

129. Kursa MB. Robustness of Random Forest-based gene selection methods. BMC bioinformatics. 2014 Dec;15:1-8.130.

130. Talekar B, Agrawal S. A detailed review on decision tree and random forest. Biosci. Biotechnol. Res. Commun. 2020 Dec 28;13(14):245-8.

131. An R, Shen J, Xiao Y. Applications of artificial intelligence to obesity research: scoping review of methodologies. Journal of Medical Internet Research. 2022 Dec 7;24(12):e40589.

132. Abdullah FS, Manan NS, Ahmad A, Wafa SW, Shahril MR, Zulaily N, Amin RM, Ahmed A. Data mining techniques for classification of childhood obesity among year 6 school children. InRecent Advances on Soft Computing and Data Mining: The Second International Conference on Soft Computing and Data Mining (SCDM-2016), Bandung, Indonesia, August 18-20, 2016 Proceedings Second 2017 (pp. 465-474). Springer International Publishing.

133. Casolaro A, Capone V, Iannuzzo G, Camastra F. Deep Learning for Time Series Forecasting: Advances and Open Problems. Information. 2023 Nov 4;14(11):598.

134. Gupta M, Phan TL, Bunnell HT, Beheshti R. Obesity Prediction with EHR Data: A deep learning approach with interpretable elements. ACM Transactions on Computing for Healthcare (HEALTH). 2022 Apr 7;3(3):1-9.

135. Monteiro PO, Victora CG. Rapid growth in infancy and childhood and obesity in later life–a systematic review. Obesity reviews. 2005 May;6(2):143-54.

136. Alhassan Z, McGough AS, Alshammari R, Daghstani T, Budgen D, Al Moubayed N. Type-2 diabetes mellitus diagnosis from time series clinical data using deep learning models. InArtificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27 2018 (pp. 468-478). Springer International Publishing.

137. Das S, Tariq A, Santos T, Kantareddy SS, Banerjee I. Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research. Machine Learning for Brain Disorders. 2023 Jul 23:117-38.

138. Heaton J. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. Genetic programming and evolvable machines. 2018 Jun;19(1):305-7.

139. Gandin I, Scagnetto A, Romani S, Barbati G. Interpretability of time-series deep learning models: A study in cardiovascular patients admitted to Intensive care unit. Journal of Biomedical Informatics. 2021 Sep 1;121:103876.

140. Olah C. Understanding LSTM Networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed 15 Jun 2024.

141. Kaji DA, Zech JR, Kim JS, Cho SK, Dangayach NS, Costa AB, Oermann EK. An attention based deep learning model of clinical events in the intensive care unit. PloS one. 2019 Feb 13;14(2):e0211057. doi: 10.1371/journal.pone.0211057

142. Hasan M, Kotov A, Carcone AI, Dong M, Naar S. Predicting the outcome of patient-provider communication sequences using recurrent neural networks and probabilistic models. AMIA Summits on Translational Science Proceedings. 2018;2018:64.

143. Cheng ER, Steinhardt R, Ben Miled Z. Predicting childhood obesity using machine learning: Practical considerations. BioMedInformatics. 2022 Mar 8;2(1):184-203.

144. Siddiqui H, Rattani A, Woods NK, Cure L, Lewis RK, Twomey J, Smith-Campbell B, Hill TJ. A survey on machine and deep learning models for childhood and adolescent obesity. IEEE Access. 2021 Nov 25;9:157337-60.

145. Craig BA, Sendi PP. Estimation of the transition matrix of a discrete-time Markov chain. Health economics. 2002 Jan;11(1):33-42. doi: 10.1002/hec.654

146. Bamakan SM, Nurgaliev I, Qu Q. Opinion leader detection: A methodological review. Expert Systems with Applications. 2019 Jan 1;115:200-22.

147. Elavarasan D, Vincent DR, Sharma V, Zomaya AY, Srinivasan K. Forecasting yield by integrating agrarian factors and machine learning models: A survey. Computers and electronics in agriculture. 2018 Dec 1;155:257-82.

148. Pishro-Nik H. Introduction to probability, statistics, and random processes. Blue Bell, PA, USA: Kappa Research, LLC; 2014 Aug.

149. Chen TA, Baranowski T, Moreno JP, O'Connor TM, Hughes SO, Baranowski J, Woehler D, Kimbro RT, Johnston CA. Obesity status transitions across the elementary years: use of M arkov chain modelling. Pediatric obesity. 2016 Apr;11(2):88-94.

150. Hu X, Tan L, Wang Z, Zhang J. Transition patterns of weight status: A cohort study of Chinese school-age children. Frontiers in Public Health. 2022 Nov 7;10:942307.

151. Tang X, Zhang H, Zhao Y, Lei F, Liu Q, Hu D, Li G, Song G. Transition patterns of weight status and their predictive lipid markers among Chinese adults: a longitudinal cohort study using the Multistate Markov Model. Diabetes, Metabolic Syndrome and Obesity. 2021 Jun 14:2661-71.

152. Rusydiyah EF, Virgiannada HR, Ridwan M, Nugroho BA, Rahman MR. Clustering of learning media user data during Covid-19 Pandemic using K-Means method based on multicultural culture in Indonesia. Journal of Innovation in Educational and Cultural Research. 2021 Dec 15;2(2):67-76.

153. Ahmad M, Arshad NI, Sarlan AB. An Analysis of Students' Academic Performance Using K-Means Clustering Algorithm. InInternational Conference of Reliable Information and Communication Technology 2021 Dec 22 (pp. 309-318). Cham: Springer International Publishing.

154. Arifin[1] TS, Natassa N, Khoirunnisa D, Hendrowati R. The Level of Student Satisfaction with the Online Learning Process During a Pandemic Using the K-means Algorithm. Inform. 2021 Jul;6(2).

155. Yang Y. Phenotyping Risk Profiles of Substance Use and Exploring the Dynamic Transitions in Use Patterns: Machine Learning Models using the COMPASS Data.

156. Arora RK, Badal D. Evaluating student's performance using k-means clustering. Int. J. Comput. Sci. Technol. 2013 Apr;4(2).

157. Carter M, McGee R, Taylor B, Williams S. Health outcomes in adolescence: Associations with family, friends and school engagement. Journal of adolescence. 2007 Feb 1;30(1):51-62.

158. Fitzgerald A, Fitzgerald N, Aherne C. Do peers matter? A review of peer and/or friends' influence on physical activity among American adolescents. Journal of adolescence. 2012 Aug 1;35(4):941-58.

159. Fletcher EC, Dumford AD. The Relationship of School Characteristics and Bullying Between Career Academy and Comprehensive High School Students. Journal of Education for Students Placed at Risk (JESPAR). 2023 Jul 3;28(3):306-25.

160. Mehra K, Rup J, Wiese JL, Watson TM, Bonato S, Rueda S. Changes in self-reported cannabis use during the COVID-19 pandemic: a scoping review. BMC public health. 2023 Nov 1;23(1):2139.

161. Crabtree SA, Captari LE, Hall EL, Sandage SJ, Jankowski PJ. Mental health symptoms, well-being and experiences of the COVID-19 pandemic: A mixed-methods practice-based study. Counselling and Psychotherapy Research. 2023 Mar;23(1):134-51.

162. Bhurtun DD, Jeewon R. Body weight perception and weight control practices among teenagers. International Scholarly Research Notices. 2013;2013(1):395125.

163. Malinauskas BM, Raedeke TD, Aeby VG, Smith JL, Dallas MB. Dieting practices, weight perceptions, and body composition: a comparison of normal weight, overweight, and obese college females. Nutrition Journal. 2006 Dec;5:1-8.

164. Wang Y, Liang H, Chen X. Measured body mass index, body weight perception, dissatisfaction and control practices in urban, low-income African American adolescents. BMC public health. 2009 Dec;9:1-2.

165. Robinson E, Haynes A, Sutin A, Daly M. Self-perception of overweight and obesity: a review of mental and physical health outcomes. Obesity science & practice. 2020 Oct;6(5):552-61.

166. Shah B, Cost KT, Fuller A, Birken CS, Anderson LN. Sex and gender differences in childhood obesity: contributing to the research agenda. BMJ Nutrition, Prevention & Health. 2020 Dec;3(2):387.

167. Luo J, Thomson CA, Hendryx M, Tinker LF, Manson JE, Li Y, Nelson DA, Vitolins MZ, Seguin RA, Eaton CB, Wactawski-Wende J. Accuracy of self-reported weight in the Women's Health Initiative. Public health nutrition. 2019 Apr;22(6):1019-28. doi: 10.1017/S1368980018003002

168. Olfert MD, Barr ML, Charlier CM, Famodu OA, Zhou W, Mathews AE, Byrd-Bredbenner C, Colby SE. Self-reported vs. measured height, weight, and BMI in young adults. International journal of environmental research and public health. 2018 Oct;15(10):2216. doi: 10.3390/ijerph15102216

169. Al-Hazzaa HM, Al-Awadhi BA, Al-Dashti YA, Alajmi FA, Almansour FD, Al-Haifi AR. Adolescent's self-reported weight and its association with media impact on decision to lose weight and body thinness perception. Scientific Reports. 2022 Apr 8;12(1):5908.

170. Devaux M, Vuik S. The relationship between childhood obesity and educational outcomes.

171. Livermore M, Duncan MJ, Leatherdale ST, Patte KA. Are weight status and weight perception associated with academic performance among youth?. Journal of eating disorders. 2020 Dec;8(1):1-0.

172. Guss CE, Williams DN, Reisner SL, Austin SB, Katz-Wise SL. Disordered weight management behaviors, nonprescription steroid use, and weight perception in transgender youth. Journal of Adolescent Health. 2017 Jan 1;60(1):17-22. doi: 10.1016/j.jadohealth.2016.08.027

173. Fonseca H, de Matos MG, Guerra A, Gomes-Pedro J. Emotional, behavioural and social correlates of missing values for BMI. Archives of Disease in Childhood. 2009 Feb 1;94(2):104-9.

174. Doggett A. Body Mass Index and Missing Data: Examining the Levels, Patterns, and Impacts of Missing Data in a Large Cohort Study of Canadian Youth.

175. Coelho JS, Suen J, Clark BA, Marshall SK, Geller J, Lam PY. Eating disorder diagnoses and symptom presentation in transgender youth: a scoping review. Current psychiatry reports. 2019 Nov;21:1-0.

176. Hahn SL, Burnette CB, Hooper L, Wall M, Loth KA, Neumark-Sztainer D. Do Weight Perception Transitions in Adolescence Predict Concurrent and Long-Term Disordered Eating Behaviors?. Journal of Adolescent

177. Patte KA, Qian W, Leatherdale ST. Predictors of One-Year Change in How Youth Perceive Their Weight. Journal of obesity. 2020;2020(1):7396948.

178. Fredrickson J, Kremer P, Swinburn B, De Silva A, McCabe M. Weight perception in overweight adolescents: associations with body change intentions, diet and physical activity.

179. Fowler BA. The relationship of body image perception and weight status to recent change in weight status of the adolescent female. Adolescence. 1989 Oct 1;24(95):557.

180. Feng X, Wilson A. Association between community average body mass index and perception of overweight. Social Science & Medicine. 2022 Feb 1;294:114694.

181. Gaspar MJ, Amaral TF, Oliveira BM, Borges N. Protective effect of physical activity on dissatisfaction with body image in children–A cross-sectional study. Psychology of Sport and Exercise. 2011 Sep 1;12(5):563-9.

182. Haynes A, Kersbergen I, Sutin A, Daly M, Robinson E. A systematic review of the relationship between weight status perceptions and weight loss attempts, strategies, behaviours and outcomes. Obesity Reviews. 2018 Mar;19(3):347-63.

183. Goodman W, Jackson SE, McFerran E, Purves R, Redpath I, Beeken RJ. Association of video game use with body mass index and other energy-balance behaviors in children. JAMA pediatrics. 2020 Jun 1;174(6):563-72.

184. Lu AS, Kharrazi H, Gharghabi F, Thompson D. A systematic review of health videogames on childhood obesity prevention and intervention. GAMES FOR HEALTH: Research, Development, and Clinical Applications. 2013 Jun 1;2(3):131-41.

185. Clark O, Lee MM, Jingree ML, O'Dwyer E, Yue Y, Marrero A, Tamez M, Bhupathiraju SN, Mattei J. Weight stigma and social media: evidence and public health solutions. Frontiers in nutrition. 2021 Nov 12;8:739056.

186. Webb JB, Vinoski ER, Bonar AS, Davies AE, Etzel L. Fat is fashionable and fit: A comparative content analysis of Fatspiration and Health at Every Size® Instagram images. Body image. 2017 Sep 1;22:53-64. doi: 10.1016/j.bodyim.2017.05.003

187. Battaglia OR, Di Paola B, Fazio C. K-means clustering to study how student reasoning lines can be modified by a learning activity based on Feynman's unifying approach. Eurasia Journal of Mathematics, Science and Technology Education. 2017 Jun 15;13(6):2005-38.

188. Cahapin EL, Malabag BA, Santiago Jr CS, Reyes JL, Legaspi GS, Adrales KL. Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms. Bulletin of Electrical Engineering and Informatics. 2023 Dec 1;12(6):3647-56.

189. Raschka S, Liu YH, Mirjalili V. Machine Learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python. Packt Publishing Ltd; 2022 Feb 25.

190. Lin ZQ, Shafiee MJ, Bochkarev S, Jules MS, Wang XY, Wong A. Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. arXiv preprint arXiv:1910.07387. 2019 Oct 16.

191. Fonseca H, Silva AM, Matos MG, Esteves I, Costa P, Guerra A, Gomes-Pedro J. Validity of BMI based on self-reported weight and height in adolescents. Acta Paediatrica. 2010 Jan;99(1):83-8.

192. Brener ND, Billy JO, Grady WR. Assessment of factors affecting the validity of self-reported health-risk behavior among adolescents: evidence from the scientific literature. Journal of adolescent health. 2003 Dec 1;33(6):436-57.193.

193. Merrill RM, Richardson JS. Peer reviewed: validity of self-reported height, weight, and body mass index: findings from the National Health and nutrition examination survey, 2001-2006. Preventing chronic disease. 2009 Oct;6(4).

194. Datar A, Chung PJ. Accuracy of weight perceptions in a nationally representative cohort of US 8th grade adolescents. Academic pediatrics. 2016 Apr 1;16(3):267-74.

195. Liu H, Wu YC, Chau PH, Chung TW, Fong DY. Prediction of adolescent weight status by machine learning: a population-based study. BMC Public Health. 2024 May 20;24(1):1351.

196. Gozukara Bag HG, Yagin FH, Gormez Y, González PP, Colak C, Gülü M, Badicu G, Ardigò LP. Estimation of obesity levels through the proposed predictive approach based on physical activity and nutritional habits. Diagnostics. 2023 Sep 14;13(18):2949.

197. Cheng ER, Cengiz AY, Miled ZB. Predicting body mass index in early childhood using data from the first 1000 days. Scientific Reports. 2023 May 31;13(1):8781.

198. Groesz LM, Levine MP, Murnen SK. The effect of experimental presentation of thin media images on body satisfaction: A meta-analytic review. International Journal of eating disorders. 2002 Jan;31(1):1-6.

199. Ikotun AM, Ezugwu AE, Abualigah L, Abuhaija B, Heming J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. Information Sciences. 2023 Apr 1;622:178-210.

200. Doggett A, Chaurasia A, Chaput JP, Leatherdale ST. Learning from missing data: examining nonreporting patterns of height, weight, and BMI among Canadian youth. International Journal of Obesity. 2022 Sep;46(9):1598-607.

201. Zhou Y, Aryal S, Bouadjenek MR. Review for Handling Missing Data with special missing mechanism. arXiv preprint arXiv:2404.04905. 2024 Apr 7.

202. Statistics Canada. School Closures and COVID-19: Impacts on children [Internet]. [place unknown: publisher unknown]; 2021 May 03. Accessed: Dec 8, 2023. Available from: https://www150.statcan.gc.ca/n1/pub/71-607-x/2021009/sc-fe-eng.htm

203. Tandon S, Patte KA, Goldfield GS, Leatherdale ST, Chaput JP. Loneliness during COVID-19 and its association with eating habits and 24-hour movement behaviours in a sample of Canadian adolescents. Preventive Medicine Reports. 2023 Jun 16:102287.

204. Ammar A, Brach M, Trabelsi K, Chtourou H, Boukhris O, Masmoudi L, Bouaziz B, Bentlage E, How D, Ahmed M, Müller P. Effects of COVID-19 home confinement on eating behaviour and physical activity: results of the ECLB-COVID19 international online survey. Nutrients. 2020 Jun;12(6):1583.

205. Giannetti C, Ransing RS, Ransing MR, Bould DC, Gethin DT, Sienz J. A novel variable selection approach based on co-linearity index to discover optimal process settings by analysing mixed data. Computers & Industrial Engineering. 2014 Jun 1;72:217-29.s

206. Costa E, Papatsouma I, Markos A. Benchmarking distance-based partitioning methods for mixed-type data. Advances in Data Analysis and Classification. 2023 Sep;17(3):701-24.

207. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences. 2016 Apr 13;374(2065):20150202.

208. HAKIM ML. Penerapan Local Interpretable Model Agnostic Explanation Lime pada Analisis Sentimen Pengguna TWITTER terhadap Kebijakan Pajak.

209. Wang ZH, Wu C, Zheng K, Niu X, Wang X. SMOTETomek-based resampling for personality recognition. IEEE access. 2019 Sep 9;7:129678-89.

210. Chaffee BW, Cheng J, Couch ET, Hoeft KS, Halpern-Felsher B. Adolescents' substance use and physical activity before and during the COVID-19 pandemic. JAMA pediatrics. 2021 Jul 1;175(7):715-22.

211. Hauerslev M, Narang T, Gray N, Samuels TA, Bhutta ZA. Childhood obesity on the rise during COVID-19: A request for global leaders to change the trajectory. Obesity (Silver Spring, Md.). 2022 Feb;30(2):288. doi: 10.1002/oby.23307

212. Ling J, Chen S, Zahry NR, Kao TS. Economic burden of childhood overweight and obesity: A systematic review and meta-analysis. Obesity reviews. 2023 Feb;24(2):e13535.

213. Mureddu F, Schmeling J, Kanellou E. Research challenges for the use of big data in policy-making. Transforming government: People, process and policy. 2020 Nov 23;14(4):593-604. doi: 10.1108/TG-08-2019-0082

214. LeCroy MN, Kim RS, Stevens J, Hanna DB, Isasi CR. Identifying key determinants of childhood obesity: a narrative review of machine learning studies. Childhood Obesity. 2021 Apr 1;17(3):153-9. doi: 10.1089/chi.2020.0324

215. Nohara Y, Matsumoto K, Soejima H, Nakashima N. Explanation of machine learning models using shapley additive explanation and application for real data in hospital. Computer Methods and Programs in Biomedicine. 2022 Feb 1;214:106584.

216. Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for depression in well older adults: evaluation of. Prev Med. 1994;10:77-84.

217. Multiple Imputation with Random Forests in Python. Wilson (2020). Available from: https://towardsdatascience.com/multiple-imputation-with-random-forests-in-python-dec83c0ac55b. Accessed: January 12, 2024.

218. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. American journal of epidemiology. 2014 Mar 15;179(6):764-74.

219. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. InProceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining 2019 Jul 25 (pp. 2623-2631).

220. Chau N, Chau K, Mayet A, Baumann M, Legleye S, Falissard B. Self-reporting and measurement of body mass index in adolescents: refusals and validity, and the possible role of socioeconomic and health-related factors. BMC Public Health. 2013 Dec;13:1-6.

221. Toselli S, Zaccagni L, Rinaldo N, Mauro M, Grigoletto A, Maietta Latessa P, Marini S. Body image perception in high school students: The relationship with gender, weight status, and physical activity. Children. 2023 Jan 10;10(1):137.

222. Gruszka W, Owczarek AJ, Glinianowicz M, Bąk-Sosnowska M, Chudek J, Olszanecka-Glinianowicz M. Perception of body size and body dissatisfaction in adults. Scientific Reports. 2022 Jan 27;12(1):1159.223.

223. Riahi R, Motlagh ME, Heshmat R, Qorbani M, Daniali SS, Kelishadi R. Body weight misperception and psychological distress among children and adolescents: The CASPIAN-V study. Osong Public Health and Research Perspectives. 2019 Oct;10(5):315.224.

224. Sonneville KR, Thurston IB, Milliren CE, Gooding HC, Richmond TK. Weight misperception among young adults with overweight/obesity associated with disordered eating behaviors. International Journal of Eating Disorders. 2016

225. Viner RM, Haines MM, Taylor SJ, Head J, Booy R, Stansfeld S. Body mass, weight control behaviours, weight perception and emotional well being in a multiethnic sample of early adolescents. International journal of obesity. 2006 Oct;30(10):1514-21.

226. Brener ND, Eaton DK, Lowry R, McManus T. The association between weight perception and BMI among high school students. Obesity research. 2004 Nov;12(11):1866-74.

227. Shin A, Nam CM. Weight perception and its association with socio-demographic and health-related factors among Korean adolescents. BMC public health. 2015 Dec;15:1-9.

228. Kriaucioniene V, Raskiliene A, Petrauskas D, Petkeviciene J. Trends in eating habits and body weight status, perception patterns and management practices among first-year students of Kaunas (Lithuania) Universities, 2000–2017. Nutrients. 2021 May 11;13(5):1599.

229. Kim Y, Austin SB, Subramanian SV, Kawachi I. Body weight perception, disordered weight control behaviors, and depressive symptoms among Korean adults: The Korea National Health and Nutrition Examination Survey 2014. PLoS One. 2018 Jun 14;13(6):e0198841.

230. Patte KA, Laxer RE, Qian W, Leatherdale ST. An analysis of weight perception and physical activity and dietary behaviours among youth in the COMPASS study. SSM-Population Health. 2016 Dec 1;2:841-9.

231. Añez E, Fornieles-Deu A, Fauquet-Ars J, López-Guimerà G, Puntí-Vidal J, Sánchez-Carracedo D. Body image dissatisfaction, physical activity and screen-time in Spanish adolescents. Journal of health psychology. 2018 Jan;23(1):36-47.

232. Neumark-Sztainer D, Goeden C, Story M, Wall M. Associations between body satisfaction and physical activity in adolescents: Implications for programs aimed at preventing a broad spectrum of weight-related disorders. Eating disorders. 2004 Jun 1;12(2):125-37.

233. Tambalis KD, Panagiotakos DB, Psarra G, Sidossis LS. Prevalence, trends and risk factors of thinness among Greek children and adolescents. Journal of preventive medicine and hygiene. 2019 Dec;60(4):E386.

234. Wang WK, Chen I, Hershkovich L, Yang J, Shetty A, Singh G, Jiang Y, Kotla A, Shang JZ, Yerrabelli R, Roghanizad AR. A systematic review of time series classification techniques used in biomedical applications. Sensors. 2022 Oct 20;22(20):8016.

235. Purushotham S, Meng C, Che Z, Liu Y. Benchmarking deep learning models on large healthcare datasets. Journal of biomedical informatics. 2018 Jul 1;83:112-34.

236. Eom G, Byeon H. Development of keyword trend prediction models for obesity before and after the COVID-19 pandemic using RNN and LSTM: analyzing the news big data of South Korea. Frontiers in Public Health. 2022 Apr 29;10:894266.

237. Kim HH, Kim Y, Park YR. Interpretable conditional recurrent neural network for weight change prediction: algorithm development and validation study. JMIR mHealth and uHealth. 2021 Mar 29;9(3):e22183.

238. Eom G, Byeon H. Development of keyword trend prediction models for obesity before and after the COVID-19 pandemic using RNN and LSTM: analyzing the news big data of South Korea. Frontiers in Public Health. 2022 Apr 29;10:894266.

239. Fan M, Jin Y, Khubchandani J. Overweight misperception among adolescents in the United States. Journal of pediatric nursing. 2014 Nov 1;29(6):536-46.

240. Alsahaf A, Petkov N, Shenoy V, Azzopardi G. A framework for feature selection through boosting. Expert Systems with Applications. 2022 Jan 1;187:115895.

241. Katitas A. Getting Started with Multiple Imputation in R. University of Virginia Library-Statlab Articles https://uvastatlab. github. io/2019/05/01/getting-started-with-multiple-imputation-in-r. 2019 May 1.

242. Kang Q, Chen EJ, Li ZC, Luo HB, Liu Y. Attention-based LSTM predictive model for the attitude and position of shield machine in tunneling. Underground Space. 2023 Dec 1;13:335-50.

243. Nazir A, He J, Zhu N, Qureshi SS, Qureshi SU, Ullah F, Wajahat A, Pathan MS. A deep learning-based novel hybrid CNN-LSTM architecture for efficient detection of threats in the IoT ecosystem. Ain Shams Engineering Journal. 2024 Jul 1;15(7):102777.

244. Habtemariam ET, Kekeba K, Martínez-Ballesteros M, Martínez-Álvarez F. A Bayesian optimization-based LSTM model for wind power forecasting in the Adama district, Ethiopia. Energies. 2023 Feb 28;16(5):2317.

245. Brownlee J. Use early stopping to halt the training of neural networks at the right time. Machine Learning Mastery. 2018 Dec;9.

246. Maarif MR, Saleh AR, Habibi M, Fitriyani NL, Syafrudin M. Energy Usage Forecasting Model Based on Long Short-Term Memory (LSTM) and eXplainable Artificial Intelligence (XAI). Information. 2023 Apr 29;14(5):265.

247. Shi Z, Lien N, Kumar BN, Holmboe-Ottesen G. Perceptions of weight and associated factors of adolescents in Jiangsu Province, China. Public health nutrition. 2007 Mar;10(3):298-305.

248. Selensky JC, Carels RA. Weight stigma and media: An examination of the effect of advertising campaigns on weight bias, internalized weight bias, self-esteem, body image, and affect. Body Image. 2021 Mar 1;36:95-106.

249. Hawkins N, Richards PS, Granley HM, Stein DM. The impact of exposure to the thin-ideal media image on women. Eating disorders. 2004 Jan 1;12(1):35-50.

250. Yoo W, Mayberry R, Bae S, Singh K, He QP, Lillard Jr JW. A study of effects of multicollinearity in the multivariable analysis. International journal of applied science and technology. 2014 Oct;4(5):9.

251. Chan JY, Leow SM, Bea KT, Cheng WK, Phoong SW, Hong ZW, Chen YL. Mitigating the multicollinearity problem and its machine learning approach: a review. Mathematics. 2022 Apr 12;10(8):1283.

252. Garg A, Tai K. Comparison of regression analysis, artificial neural network and genetic programming in handling the multicollinearity problem. In2012 proceedings of international conference on modelling, identification and control 2012 Jun 24 (pp. 353-358). IEEE.

253. Alake R. Loss functions in machine learning explained [Internet]. DataCamp; Available from: https://www.datacamp.com/tutorial/loss-function-in-machine-learning. Accessed September 12, 2024.

254. Government of Canada. What is gender? What is sex? [Internet]. Canadian Institutes of Health Research. 2023. Available from: https://cihr-irsc.gc.ca/e/48642.html. Accessed September 19, 2024

255. Nagata JM, Ganson KT, Griffiths S, Mitchison D, Garber AK, Vittinghoff E, Bibbins-Domingo K, Murray SB. Prevalence and correlates of muscle-enhancing behaviors among adolescents and young adults in the United States. International Journal of Adolescent Medicine and Health. 2022 Apr 18;34(2):119-29.

256. Skinner AC, Ravanbakht SN, Skelton JA, Perrin EM, Armstrong SC. Prevalence of obesity and severe obesity in US children, 1999–2016. Pediatrics. 2018 Mar 1;141(3).

257. Public Health Agency of Canada. Obesity in Canada – Prevalence among aboriginal populations - Canada.ca [Internet]. Canada.ca. 2011. Available from: https://www.canada.ca/en/public-health/services/health-promotion/healthy-living/obesity-canada/prevalence-among-aboriginal-populations.html. Accessed September 14, 2024.

258. Mbogori T, Arthur T, Kimmel K. Gender and Racial/Ethnic Disparities in the Perception of Weight Status, Health Status and Diet Quality Among Adolescents in the USA (P04-114-19). Current Developments in Nutrition. 2019 Jun 1;3:nzz051-P04.

259. Sonoda R, Tokiya M, Touri K, Tanomura Y, Yada K, Funakoshi Y, Saito I. A point system to predict the future risk of obesity in 10-year-old children. Environmental Health and Preventive Medicine. 2023;28:25-.

260. Peña MM, Dixon B, Taveras EM. Are you talking to ME? The importance of ethnicity and culture in childhood obesity prevention and management. Childhood Obesity. 2012 Feb 1;8(1):23-7.

261. Kracht CL, Joseph ED, Staiano AE. Video games, obesity, and children. Current obesity reports. 2020 Mar;9:1-4.

262. Robinson TN, Banda JA, Hale L, Lu AS, Fleming-Milici F, Calvert SL, Wartella E. Screen media exposure and obesity in children and adolescents. Pediatrics. 2017 Nov 1;140(Supplement_2):S97-101.

263. Reel B, Battista K, Leatherdale ST. COMPASS Year 7 and 8 Recruitment and Retention: Technical Report Series. (2021); 8(1): Waterloo, Ontario: University of Waterloo. Available at: https://uwaterloo.ca/compass-system/publications#technical. Accessed September 14, 2024

Appendices

Appendix A: The COMPASS Data Dictionary for Representative Domains

Weight-Related Measures

BMI: Self-reported weight was asked, "How much do you weigh without your shoes on? (Please enter your weight in either pounds or kilograms)". Self-report height was asked, "How tall are you without your shoes on? (Please enter your height in either feet and inches or centimetres)". The derived BMI category is based on the WHO guidelines for ages 5- 19 years of school-aged children and adolescents and calculated from the valid response of sex, BMI (numeric version), and age [52]. The response categories were "Underweight", "Healthy Weight", "Overweight", and "Obese".

Weight Perception: Weight perception was assessed by the question, "How do you describe your weight?". The response categories were "Very underweight", "Slightly underweight", "About the right weight", "Slightly overweight", and "Very overweight".

Weight Change Intention: The question was asked, "Which of the following are you trying to do about your weight?" The response categories were "Lost weight," "Gain weight," "Stay the same weight," and "I am not trying to do anything about my weight."

Demographics

Age, grade, and sex were reported. Ethnicity was assessed by asking, "What race best describes you?" The response categories were "Black", "East Asian", "Latino", "Middle Eastern", "South Asian", "White", "Another category", "I don't know", and "I prefer not to say". Other measures include income, "About how much money do you usually get each week to spend on yourself or to save?", the response categories were "Zero", "$1 to $5", "6 to $10", "$11 to $20", "$21 to $40", "$41 to $100", "More than $100", "I do not know how much money I get each week".

Healthy eating

Skipping breakfast: "If you do not eat breakfast every day, why do you skip breakfast (Mark all that apply". The response options were "I eat breakfast every day", "I don't have time for breakfast", "The bus comes too early", "I sleep in", "I'm not hungry in the morning", "I feel sick when I eat breakfast", "I'm trying to lose weight", "There is nothing to eat at home", and "Other".

Lunch at school: “In a usual school week (Monday to Friday), on how many days do you do the following? Eat lunch at school - lunch packed and brought from home”. The response options were “None”, “1 day”, “2 days”, “3 days”, “4 days”, and “5 days”.

Snacks consumption: “In a usual school week (Monday to Friday), on how many days do you do the following? Eat snacks purchased from a vending machine in your school”. The response options were “None”, “1 day”, “2 days”, “3 days”, “4 days”, and “5 days”.

Sugary-drinks consumption: “In a usual school week (Monday to Friday), on how many days do you do the following? Drink sugar-sweetened beverages (soda-pop, Kool-Aid, Gatorade, etc.) Do not include diet/sugar-free drinks”. The response options were “None”, “1 day”, “2 days”, “3 days”, “4 days”, and “5 days”.

Protein consumption: “Yesterday, from the time you woke up until the time you went to bed, how many servings of meats and alternatives did you have? One 'Food Guide' serving of meat and alternatives includes cooked fish, chicken, beef, pork, or game meat, eggs, nuts or seeds, peanut butter or nut butters, legumes (beans), and tofu”. The response options were “None”, “1 serving”, “2 serving”, “3 serving”, “4 serving”, and “5 or more serving”.

Vegetable consumption: “Yesterday, from the time you woke up until the time you went to bed, how many servings of vegetables and fruits did you have? One 'Food Guide' serving of vegetables and fruit includes pieces of fresh vegetable or fruit, salad or raw leafy greens, cooked leafy green vegetables, dried or canned or frozen fruit, and 100% fruit or vegetable juice”. The response options were “None”, “1 serving”, “2 serving”, “3 serving”, “4 serving”, and “5 serving”, “6 serving”, “7 serving”, “8 serving”, and “9 or more serving”.

Physical Activity

Physical activity at school: “Do you participate in before-school, noon hour, or after school physical activities organized by your school? (e.g., intramurals, non-competitive clubs)?”. The responses were “Yes”, “No”, and “None offered at school”

Level of Physical Activity: “Mark how many minutes of MODERATE physical activity you did on each of the last 7 days. This includes physical activity during physical education class, lunch, after school, evenings, and spare time. Do not include time spent doing hard physical activities”. The same question applied to the HARD physical activity. The derived continuous variable from this question combined the hours (0, 1, 2, 3, and 4) response and

minutes (0, 15, 30, and 45) to calculate the total amount of time spent on HARD or MODERATE activities as well as average time per week (total amount of time/7).

Sedentary Lifestyle

Surfing internet: The surfing internet (continuous) per minute was derived from the questions "How much time per day do you usually spend doing the following activities? Surfing the internet; Hours", and "How much time per day do you usually spend doing the following activities? Surfing the internet; Minutes". The response in hours ranged from 0 hours to 9 hours, and minutes in (0, 15, 30, and 45). The same type of questions was asked for "Doing homework: Minutes per day", "Watching TV: Minutes per day", and "Taking on the phone: Minutes per day".

School-related Measures

School connectedness: It was calculated based on questions in "I feel close to people at school", "I feel I am part of my school", "I am happy to be at my school", "I feel the teachers at my school treat me fairly', 'I feel safe in my school', and 'Getting good grades is important to me". The responses were "1=strongly agree", "2=agree", "3=disagree", and "4=strongly disagree" [53]. The connectedness score was calculated on a revised scale, ranging from 6 to 24.

Academic Achievement: "On average, what marks do you usually get in school? (Choose only one)". The responses were "90% - 100% (mostly As)", "80% - 89% (mostly Bs)", "70% - 79% (mostly Cs), "60% - 69% (mostly Ds), "50% - 59% (mostly Ds)", and "Below 50% (Mostly Fs)".

Bullying: The bullying domain includes questions on ways students were bullied, the reasons for being bullied and bullying frequency. An example question was "In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) I have not been bullied in the last 30 days". The responses were "Yes" and "Questions not answered" (This is considered "No").

Mental Health

Generalized Anxiety Disorder 7-item Scale [54]: It was assessed based on the questions "Over the last 2 weeks, how often have you been bothered by the following problems?", among "Feeling nervous, anxious, or on edge", "Not being able to stop or control worrying", "Worrying too much about different things", "Trouble relaxing.", "Being so restless that it's hard to sit still", "Becoming easily annoyed or irritable", "Feeling afraid as if something awful might happen". The responses were "0 = Not at all", "1 = Several days", "2 = Over half the days", and "3 =

Nearly every day". The scores sum the score from 'Not at all" to "Nearly every day" with the range between 0 to 21.

Flourishing Scale [55]: It was assessed based on the questions "How much do you agree with the following statements?" among "I lead a purposeful and meaningful life", "My social relationships are supportive and rewarding", "I am engaged and interested in my daily activities", "I actively contribute to the happiness and well-being of others", "I am competent and capable in the activities that are important to me", "I am a good person and live a good life", "I am optimistic about my future", "People respect me". The responses were "Strongly agree", "Agree", "Neither agree nor disagree", "Disagree", and "Strongly disagree".

Substance Use

This domain asked about the frequency of the below substance use activity in the last 30 days or last 12 months, age at first use, and substance use as risky behaviour.

Alcohol Use: An example question from the drinking frequency was assessed as "In the last 12 months, how often did you have a drink of alcohol that was more than just a sip?". The response categories were "I have never drunk alcohol", "I did not drink alcohol in the last 12 months", "I have only had a sip of alcohol", "Less than once a month", "Once a month", "2 or 3 times a month", "Once a week", "2 or 3 times a week", "4 to 6 times a week", and "Every day".

Cannabis Use: An example question from the use frequency was assessed as "In the last 12 months, how often did you use marijuana or cannabis? (a joint, pot, weed, hash)". The response categories were "I have never used marijuana", "I used marijuana, but not in the last 12 months", "Less than once a month", "Once a month", "2 or 3 times a month", "Once a week", "2 or 3 times a week", "4 to 6 times a week", and "Every day".

Smoking and Tobacco Use: An example question from the use frequency was assessed as "In the last 12 months, how often did you smoke one or more cigarettes?". The responses categories were "I have never smoked a cigarette", "I did not smoke a cigarette in the last 12 months", "I have only had a few puffs of a cigarette", "Less than once a month", "Once a month", "2 or 3 times a month", "Once a week", "2 or 3 times a week", "4 to 6 times a week", and "Every day".

Opioid Use: "Have you used or tried any of the following medications TO GET HIGH? A) Oxycodone (oxy, OC, APO, OxyContin, percs, roxies, OxyNEO)?" The response categories were "No, I have never done this", "Yes, I have done this in the last 12 months", and "Yes, I have done this, but NOT in the last 12 months".

COVID-19

This domain only appeared in and after the 2020-21 survey. This domain included questions about overall concern regarding COVID-19, feelings about the COVID-19 pandemic, and impact of COVID-19 on personal life, and knowledge about COVID-19.

Impact of COVID-19: "How has your life changed because of COVID-19? My stress has…", the response categories were "Increased", "Stayed the same/not applicable", and "Decreased". The same type of questions were asked about boredom, loneliness, anxiety, time spent with my friend in person, time spent in video games, social media, cigarette use, vaping, alcohol and cannabis use.

Knowledge of COVID-19: "Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) The use of hand sanitizer (e.g., Purell) helps prevent the transmission of COVID-19". The response categories were "Yes" and "Questions not answered" (The question net answered is considered as "No").

Appendix B: Study 1 Supplementary Materials

Table S1.1 Cohort 2018-2019 Input Variable for Clustering Analysis

2018-19 Variable	Question	Data Type	Missing Percentage	Category
CESD	Depression Revised Scale (CESD R-10)	float64	14.57	Mental Health
PFMSKLA1	How do you usually travel to and from school? (If you use two or more modes of travel, choose the one that you spend most time doing) From school	category	10.93	Demographics
GAD7	Anxiety Disorder (GAD7)	float64	8.34	Mental Health
DERS	Difficulties in Emotion Regulation Scale (DERS)	float64	7.56	Mental Health
OSUPRTD1	How supportive is your school of the following? Giving students the support they need to resist or quit tobacco	category	7.08	School Connectedness
OSUPRTE1	How supportive is your school of the following? Giving students the support they need to resist or quit drugs and/or alcohol	category	7.02	School Connectedness
MHFEELI1	On how many of the last 7 days did you feel the following ways? I felt lonely.	category	6.51	Mental Health
MHFEELG1	On how many of the last 7 days did you feel the following	category	6.50	Mental Health

	ways? My sleep was restless.			
OSUPRTC1	How supportive is your school of the following? Making sure no one is bullied at school	category	6.42	School Connectedness
MHFEELH1	On how many of the last 7 days did you feel the following ways? I was happy.	category	6.32	Mental Health
MHFEELD1	On how many of the last 7 days did you feel the following ways? I felt that everything I did was an effort.	category	6.14	Mental Health
MHFEELE1	On how many of the last 7 days did you feel the following ways? I felt hopeful about the future.	category	6.07	Mental Health
FLOURISH	Flourishing Scale_self-perceived success	float64	6.02	Mental Health
MHFEELJ1	On how many of the last 7 days did you feel the following ways? I could not get "going".	category	5.99	Mental Health
MHFEELF1	On how many of the last 7 days did you feel the following ways? I felt fearful.	category	5.94	Mental Health
OSUPRTB1	How supportive is your school of the following? Making sure students have access to healthy foods and drinks	category	5.73	School Connectedness
MHFEELC1	On how many of the last 7 days did you feel the following	category	5.52	Mental Health

	ways? I felt depressed.			
OSUPRTA1	How supportive is your school of the following? Making sure there are opportunities for students to be physically active	category	5.47	School Connectedness
PTOSKLA1	How do you usually travel to and from school? (If you use two or more modes or travel, choose the one that you spend most time doing) To school	category	5.31	Demographics
OLIKEDA1	What is the highest level of education you would like to get?	category	5.03	Academic Achievement
OCONCTC2	How strongly do you agree or disagree with each of the following statements? I am happy to be at my school.	category	4.96	School Connectedness
MHFEELB1	On how many of the last 7 days did you feel the following ways? I had trouble keeping my mind on what I was doing.	category	4.96	Mental Health
SEVRQTA1	Have you ever tried to quit smoking cigarettes?	category	4.91	Smoking & Tobacco Use
OCONCTE2	How strongly do you agree or disagree with each of the following statements? I feel safe in my school.	category	4.91	School Connectedness

OMTHMKA1	In your current or most recent Math course, what is your approximate overall mark? (Think about last year if you have not taken math this year)	category	4.79	Academic Achievement
OWILLEA1	What is the highest level of education you think you will get?	category	4.78	Academic Achievement
OENGMKA1	In your current or most recent French/English course, what is your approximate overall mark? (Think about last year if you have not taken English this year)	category	4.76	Academic Achievement
OCONCTD2	How strongly do you agree or disagree with each of the following statements? I feel the teachers at my school treat me fairly.	category	4.76	School Connectedness
MHFEELA1	On how many of the last 7 days did you feel the following ways? I was bothered by things that usually don't bother me.	category	4.75	Mental Health
OCONCTF2	How strongly do you agree or disagree with each of the following statements? Getting good grades is important to me.	category	4.66	School Connectedness

MBUL30A1	In the last 30 days, how often have you been bullied by other students?	category	4.61	Bullying
MBLG30A1	In the last 30 days, how often have you taken part in bullying other students?	category	4.60	Bullying
GGAMBLA1	In the last 30 gays, did you gamble online for money?	category	4.55	Screen
OCONCTB2	How strongly do you agree or disagree with each of the following statements? I feel I am part of my school.	category	4.54	School Connectedness
HEMTOFC1	In a usual school week (Monday to Friday), on how many days do you do the following? Eat lunch at school - lunch packed and brought from home	category	4.36	Healthy Eating
DVHRDAVE	Average HARD activity per day (Hard activity total / 7)	float64	4.28	Physical Activity
OCONCTA2	How strongly do you agree or disagree with each of the following statements? I feel close to people at my school.	category	4.27	School Connectedness
AFNTNLA1	Have you used or tried any of the following medications TO GET HIGH? B) Fentanyl (china white,	category	4.20	Opioid Use

	synthetic heroin, china girl)			
DVOCONCT	School Connectedness Score	float64	4.14	School Connectedness
OINCOMA1	How often do you go to class without your homework complete?	category	4.03	Academic Achievement
HEMTOFB1	In a usual school week (Monday to Friday), on how many days do you do the following? Eat breakfast provided to you as past of a school program	category	3.99	Healthy Eating
OMISSHA1	In the last 4 weeks, how many days of school did you miss because of your health?	category	3.95	Academic Achievement
OSKIP0A2	In the last 4 weeks, how many classes did you skip when you were not supposed to?	category	3.93	Academic Achievement
GSMSKLA1	Did you attend this school last year?	category	3.92	Demographics
HMEATSA1	Yesterday, from the time you woke up until the time you went to bed, how many servings of meats and alternatives did you have? One 'Food Guide' serving of meat and alternatives includes cooked fish, chicken, beef, pork, or game meat, eggs, nuts or seeds, peanut	category	3.80	Healthy Eating

	butter or nut butters, legumes (beans), and tofu.			
AOXYCDA1	Have you used or tried any of the following medications TO GET HIGH? A) Oxycodone (oxy, OC, APO, OxyContin, percs, roxies, OxyNEO)	category	3.80	Opioid Use
HFRVEGA1	Yesterday, from the time you woke up until the time you went to bed, how many servings of vegetables and fruits did you have? One 'Food Guide' serving of vegetables and fruit includes pieces of fresh vegetable or fruit, salad or raw leafy greens, cooked leafy green vegetables, dried or canned or frozen fruit, and 100% fruit or vegetable juice.	category	3.79	Healthy Eating
HGRAINA1	Yesterday, from the time you woke up until the time you went to bed, how many servings of grain products did you have? One 'Food Guide' serving of grain products includes bread, bagels, flatbread such as tortilla, pita, cooked rice or pasta, and cold cereal.	category	3.67	Healthy Eating

HMILKSA1	Yesterday, from the time you woke up until the time you went to bed, how many servings of milk and alternatives did you have? One 'Food Guide' serving of milk or milk alternatives includes milk, fortified soy beverage, reconstituted powdered milk, canned (evaporated) milk, yogurt or kefir (another type of cultured milk product), and cheese.	category	3.66	Healthy Eating
HESTOSD1	On a usual weekend (Saturday and Sunday), on how many days do you do the following? Eat snacks purchased from a vending machine, corner store, snack bar, or canteen	category	3.45	Healthy Eating
HEMTOFD1	In a usual school week (Monday to Friday), on how many days do you do the following? Eat lunch at school - lunch purchased in the cafeteria	category	3.42	Healthy Eating
HESTOSC1	On a usual weekend (Saturday and Sunday), on how many days do you do the following? Eat foods purchased at a	category	3.38	Healthy Eating

	fast food place or restaurant			
HESTOSB1	On a usual weekend (Saturday and Sunday), on how many days do you do the following? Eat lunch	category	3.36	Healthy Eating
HESTOSF1	On a usual weekend (Saturday and Sunday), on how many days do you do the following? Drink high energy drinks (Red Bull, Monster, Rock Star, etc.)	category	3.25	Healthy Eating
HESTOSH1	On a usual weekend (Saturday and Sunday), on how many days do you do the following? Drink coffee or tea without sugar	category	3.24	Healthy Eating
MOPTIMI1	How much do you agree with the following statements? I generally recover from setbacks quickly	category	3.23	Mental Health
AEVRETB1	How old were you when you first had a drink of alcohol that was more than a sip?	category	3.23	Alcohol Use
HESTOSE1	On a usual weekend (Saturday and Sunday), on how many days do you do the following? Drink sugar-sweetened beverages (soda pop, Kool-Aid, Gatorade, etc.) Do not include	category	3.22	Healthy Eating

	diet/sugar-free drinks			
MFAMSUB1	How much do you agree or disagree with the following statements?My parents/guardians expect too much of me	category	3.22	Mental Health
HESTOSG1	On a usual weekend (Saturday and Sunday), on how many days do you do the following? Drink coffee or tea with sugar (include cappuccino, frappuccino, iced-tea, iced-coffees, etc.)	category	3.16	Healthy Eating
HEMTOFH1	In a usual school week (Monday to Friday), on how many days do you do the following? Drink sugar-sweetened beverages (soda-pop, Kool-Aid, Gatorade, etc.) Do not include diet/sugar-free drinks	category	3.14	Healthy Eating
APRSCPA1	Have you used or tried any of the following medications TO GET HIGH? C)Other prescription pain relievers (codeine, morphine, Tylenol 3)	category	3.09	Opioid Use
HEMTOFI1	In a usual school week (Monday to Friday), on how many days do you do	category	3.04	Healthy Eating

	the following? Drink high-energy drinks (Red Bull, Monster, Rock Star, etc.)			
MFAMSUC1	How much do you agree or disagree with the following statements? I can talk about my problems with my family	category	3.04	Mental Health
HEMTOFE1	In a usual school week (Monday to Friday), on how many days do you do the following? Eat lunch purchased at a fast food place or restaurant	category	3.02	Healthy Eating
HEMTOFG1	In a usual school week (Monday to Friday), on how many days do you do the following? Eat snacks purchased from a vending machine, corner store, snack bar, or canteen off school property	category	3.01	Healthy Eating
HEMTOFJ1	In a usual school week (Monday to Friday), on how many days do you do the following? Drink coffee or tea with sugar (include cappuccino, frappuccino, iced-tea, iced-coffees, etc.)	category	2.98	Healthy Eating
HESTOSA1	On a usual weekend (Saturday and Sunday), on how	category	2.82	Healthy Eating

	many days do you do the following? Eat breakfast			
APRHRDA1	Do you think it would be difficult or easy to get pain relievers (Oxycodone, Fentanyl, codeine, etc.) if you wanted some?	category	2.78	Opioid Use
HEMTOFF1	In a usual school week (Monday to Friday), on how many days do you do the following? Eat snacks purchased from a vending machine in your school	category	2.77	Healthy Eating
HEMTOFA1	In a usual school week (Monday to Friday), on how many days do you do the following? Eat breakfast	category	2.77	Healthy Eating
HEMTOFK1	In a usual school week (Monday to Friday), on how many days do you do the following? Drink coffee or tea without sugar	category	2.74	Healthy Eating
MFAMSUD1	How much do you agree or disagree with the following statements? I can talk about my problems with my friends	category	2.52	Mental Health
DVACTAVE	Average combined HARD and MODERATE physical activity per	float64	2.39	Physical Activity

	day (combined activity total / 7)			
PA_LEVEL	respondents who meet the guidelines for at least 60 minutes of physical activity per day	category	2.39	Physical Activity
AHWHRDA2	Do you think it would be difficult or easy for you to get marijuana if you wanted some?	category	2.27	Cannabis Use
MFAMSUA1	How much do you agree or disagree with the following statements? I have a happy home life	category	2.24	Mental Health
PHRDTPA2	Were the last 7 days a typical week in terms of the amount of physical activity that you usually do?	category	2.23	Physical Activity
PEDUCAA1	Are you taking a physical education class at school this year?	category	1.98	Physical Activity
HDSCWTA1	How do you describe your weight?	category	1.92	Weight
P5FRNDA2	Your closest friends are the friends you like to spend the most time with. How many of your closest friends are physically active?	category	1.92	Physical Activity
HWEGHTA1	Which of the following are you trying to do about your weight?	category	1.83	Weight
PSTRENA3	On how many days in the last 7 days did you do exercises to strengthen or tone	category	1.80	Physical Activity

	your muscles? (e.g., push-ups, sit-ups, or weight-training)			
PSPORTC5	Do you participate in before-school, noon hour, or after school physical activities organized by your school? (e.g., intramurals, non-competitive clubs)	category	1.69	Physical Activity
AOFTMJA2	In the last 12 months, how often did you use marijuana or cannabis? (a joint, pot, weed, hash)	category	1.61	Cannabis Use
PSPORTD5	Do you participate in competitive school sports teams that compete against other schools? (e.g., junior varsity or varsity sports)	category	1.56	Physical Activity
PSPORTD4	Do you participate in league or team sports outside of school?	category	1.54	Physical Activity
ADRINKA1	In the last 12 months, how often did you have a drink of alcohol that was more than just a sip?	category	1.53	Alcohol Use
SUSECGA2	Have you ever tried an electronic cigarette, also known as an e-cigarette?	category	1.52	Vaping & E-Cigarette Use
GMONEYA1	About how much money do you usually get each week to spend on yourself or to save? (Remember to include all money	category	1.45	Demographics

	from allowances and jobs like baby-sitting, delivering papers, etc.)			
SECG30A1	On how many of the last 30 days did you use an e-cigarette?	category	1.39	Vaping & E-Cigarette Use
SSUSNYA1	At any time during the next year do you think you will smoke a cigarette?	category	1.37	Smoking & Tobacco Use
S5FRNDA1	Your closest friends are the friends you like to spend the most time with. How many of your closest friends smoke cigarettes?	category	1.35	Smoking & Tobacco Use
SSUSFOA1	If one of your best friends was to offer you a cigarette, would you smoke it?	category	1.34	Smoking & Tobacco Use
SEX3	Are you female or male?	category	1.11	Demographics
GRADE	What grade are you in?	category	1.10	Demographics
DSUSCEPT	Smoking Susceptibility	category	0.93	Smoking & Tobacco Use
SLST30A1	On how many of the last 30 days did you smoke one or more cigarettes?	category	0.87	Smoking & Tobacco Use
SHUND0A1	Have you ever smoked 100 or more whole cigarettes in your life?	category	0.85	Smoking & Tobacco Use
PPHON	Talking on phone : Minutes per day	float64	0.81	Screen Time
PTVMO	Watching TV : Minutes per day	float64	0.81	Screen
PVIDE	Video Games : Minutes per day	float64	0.81	Screen Time

PHMWK	Doing Homework: Minutes per day	float64	0.81	Screen Time
DTDSCRNT	Total daily sedentary activity, with homework excluded	float64	0.81	Screen Time
PSURF	Surfing Internet : Minutes per day	float64	0.81	Screen Time
PTEXT	Texting, messaging : Minutes per day	float64	0.81	Screen Time
PSLEP	Sleeping : Minutes per day	float64	0.81	Sleep
DTDSEDNA	Total daily sedentary	float64	0.81	Screen Time
SPUFF0A1	Have you ever tried cigarette smoking, even just a few puffs?	category	0.77	Smoking & Tobacco Use
AGE	How old are you today?	category	0.73	Demographics
AEVRMJB1	How old were you when you first used marijuana or cannabis?	category	0.44	Cannabis Use
A5DRNKC1	In the last 12 months, how often did you have 5 drinks of alcohol or more on one occasion?	category	0.33	Alcohol Use
GETHNCH1	How would you describe yourself? Latin American/Hispanic	category	0.00	Demographics
GETHNCM1	How would you describe yourself? Other	category	0.00	Demographics
MBULLGC1	In the last 30 days, in what ways did you bully other students? (Mark all that apply) Verbal attacks (e.g., teased, threatened, or	category	0.00	Bullying

	spread rumours about them)			
MBULLGD1	In the last 30 days, in what ways did you bully other students? (Mark all that apply) Cyber-attacks (e.g., sent mean text messages or spread rumours about them on the internet)	category	0.00	Bullying
MBULLGE1	In the last 30 days, in what ways did you bully other students? (Mark all that apply) Stole from them or damaged their things	category	0.00	Bullying
HSKBRKI1	If you do not eat breakfast every day, why do you skip breakfast? Other:	category	0.00	Healthy Eating
HSKBRKH1	If you do not eat breakfast every day, why do you skip breakfast? There is nothing to eat at home	category	0.00	Healthy Eating
HSKBRKG1	If you do not eat breakfast every day, why do you skip breakfast? I'm trying to lose weight	category	0.00	Healthy Eating
LANGUAGE	Questionnaire language administered.	category	0.00	Demographics
PROVINCE	Province in which the school resides	category	0.00	Demographics
DVMDRAVE	Average MODERATE activity per day (Mod_activity_total / 7)	float64	0.00	Physical Activity

GETHNCL1	How would you describe yourself? White	category	0.00	Demographics
GETHNCC1	How would you describe yourself? Black	category	0.00	Demographics
BMI_ACAT	BMI catagories	category	0.00	Weight
HSKBRKF1	If you do not eat breakfast every day, why do you skip breakfast?I feel sick when I eat breakfast	category	0.00	Healthy Eating
HSKBRKE1	If you do not eat breakfast every day, why do you skip breakfast? I'm not hungry in the morning	category	0.00	Healthy Eating
HSKBRKD1	If you do not eat breakfast every day, why do you skip breakfast? I sleep in	category	0.00	Healthy Eating
HSKBRKC1	If you do not eat breakfast every day, why do you skip breakfast? The bus comes too early	category	0.00	Healthy Eating
HSKBRKB1	If you do not eat breakfast every day, why do you skip breakfast? I don't have time for breakfast	category	0.00	Healthy Eating
HSKBRKA1	If you do not eat breakfast every day, why do you skip breakfast? I eat breakfast every day	category	0.00	Healthy Eating
MBULLGB1	In the last 30 days, in what ways did you bully other students? (Mark all that apply) Physical attacks	category	0.00	Bullying

	(e.g., beat up, pushed, or kicked them)			
GETHNCN1	How would you describe yourself? Asian	category	0.00	Demographics
MBULLDC1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Verbal attacks (e.g., getting teased, threatened, or having rumours spread about you)	category	0.00	Bullying
MBULLGA1	In the last 30 days, in what ways did you bully other students? (Mark all that apply) I did not bully other students in the last 30 days	category	0.00	Bullying
MBRHLPD1	If you had concerns regarding your mental health, are there any reasons why you wouldn't talk to an adult at school (e.g., a school social worker, child and youth worker, counsellor, psychologist, nurse, teacher, or other staff person)? (Mark all that apply) Prefer to handle problems myself.	category	0.00	Mental Health
MBRHLPB1	If you had concerns regarding your mental health, are there any reasons why you wouldn't talk to an adult at	category	0.00	Mental Health

	school (e.g., a school social worker, child and youth worker, counsellor, psychologist, nurse, teacher, or other staff person)? (Mark all that apply) Worried about what others would think of me (e.g., I'd be too embarrassed)			
MBRHLPA1	If you had concerns regarding your mental health, are there any reasons why you wouldn't talk to an adult at school (e.g., a school social worker, child and youth worker, counsellor, psychologist, nurse, teacher, or other staff person)? (Mark all that apply) I would have no problem talking to an adult at school about my mental health.	category	0.00	Mental Health
SEVTRYK2	In the last 30 days, did you use any of the following? (Mark all that apply) Roll-your-own cigarettes (tobacco only)	category	0.00	Smoking & Tobacco Use
SEVTRYO2	In the last 30 days, did you use any of the following? (Mark all that apply) Loose tobacco mixed with marijuana	category	0.00	Smoking & Tobacco Use
SEVTRYP1	In the last 30 days, did you use any of	category	0.00	Vaping & E-Cigarette Use

	the following? (Mark all that apply) E-cigarettes (electronic cigarettes that look like cigarettes/cigars, but produce vapour instead of smoke)			
SEVTRYL2	In the last 30 days, did you use any of the following? (Mark all that apply) Smokeless tobacco (chewing tobacco, pinch, snuff, or snus)	category	0.00	Smoking & Tobacco Use
SEVTRYG4	In the last 30 days, did you use any of the following? (Mark all that apply) Nicotine patches, nicotine gum, nicotine lozenges, or nicotine inhalers	category	0.00	Smoking & Tobacco Use
SEVTRYH7	In the last 30 days, did you use any of the following? (Mark all that apply) Hookah (water-pipe) to smoke tobacco	category	0.00	Smoking & Tobacco Use
SEVTRYH8	In the last 30 days, did you use any of the following? (Mark all that apply) Hookah (water-pipe) to smoke herbal sheesha/shisha	category	0.00	Smoking & Tobacco Use
SEVTRYN2	In the last 30 days, did you use any of the following? (Mark all that apply) Blunt wraps (a sheet or tube made of tobacco	category	0.00	Smoking & Tobacco Use

	used to roll cigarette tobacco)			
SEVTRYJ2	In the last 30 days, did you use any of the following? (Mark all that apply) I have not used any of these things in the last 30 days	category	0.00	Smoking & Tobacco Use
AMJTYPD1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have not used marijuana or cannabis in the lsat 12 months	category	0.00	Cannabis Use
AMJTYPC1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have used it by eating or drinking it (e.g., in brownies, cookies, candies, tea)	category	0.00	Cannabis Use
AMJTYPB1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have used it by vaping it	category	0.00	Vaping & E-Cigarette Use
AMJTYPA1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have used it by smoking it (e.g., in a joint, a pipe, a bong)	category	0.00	Cannabis Use

MBRHLPC1	If you had concerns regarding your mental health, are there any reasons why you wouldn't talk to an adult at school (e.g., a school social worker, child and youth worker, counsellor, psychologist, nurse, teacher, or other staff person)? (Mark all that apply) Lack of trust in these people - word would get out.	category	0.00	Mental Health
MBRHLPE1	If you had concerns regarding your mental health, are there any reasons why you wouldn't talk to an adult at school (e.g., a school social worker, child and youth worker, counsellor, psychologist, nurse, teacher, or other staff person)? (Mark all that apply) Don't think these people would be able to help.	category	0.00	Mental Health
MBULLDE1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Had someone steal from you or damage your things	category	0.00	Bullying
MBRHLPF1	If you had concerns regarding your	category	0.00	Mental Health

	mental health, are there any reasons why you wouldn't talk to an adult at school (e.g., a school social worker, child and youth worker, counsellor, psychologist, nurse, teacher, or other staff person)? (Mark all that apply) Wouldn't know who to approach.			
MBULLDD1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Cyber-attacks (e.g., being sent mean text messages or having rumours spread about you on the internet)	category	0.00	Bullying
SCHOOLID	SchoolID	category	0.00	Demographics
MBULLDB1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Physical attacks (e.g., getting beaten up, pushed, or kicked)	category	0.00	Bullying
MBULLDA1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) I have not been bullied in the last 30 days	category	0.00	Bullying
SSUSMTA1	Do you think in the future you might try smoking cigarettes?	category	0.00	Smoking & Tobacco Use

SUSECGB1	Have you used e-cigarettes for any of the following reasons? (Mark all that apply) I have not used e-cigarettes	category	0.00	Vaping & E-Cigarette Use
SUSECGC1	Have you used e-cigarettes for any of the following reasons? (Mark all that apply) Curiosity / to try something new	category	0.00	Vaping & E-Cigarette Use
SUSECGD1	Have you used e-cigarettes for any of the following reasons? (Mark all that apply) I can use e-cigarettes in places where smoking is not allowed	category	0.00	Vaping & E-Cigarette Use
SUSECGE1	Have you used e-cigarettes for any of the following reasons? (Mark all that apply) To smoke fewer cigarettes	category	0.00	Vaping & E-Cigarette Use
SUSECGF1	Have you used e-cigarettes for any of the following reasons? (Mark all that apply) To help me quit smoking cigarettes	category	0.00	Vaping & E-Cigarette Use
SUSECGG1	Have you used e-cigarettes for any of the following reasons? (Mark all that apply) I have used e-cigarettes for some other reason	category	0.00	Vaping & E-Cigarette Use
SEVTRYA2	In the last 30 days, did you use any of the	category	0.00	Smoking & Tobacco Use

	following? (Mark all that apply) Pipe tobacco			
SEVTRYB5	In the last 30 days, did you use any of the following? (Mark all that apply) Cigarillos or little cigars (plain or flavoured)	category	0.00	Smoking & Tobacco Use
SEVTRYB6	In the last 30 days, did you use any of the following? (Mark all that apply) Cigars (not including cigarillos or little cigars, plain or flavoured)	category	0.00	Smoking & Tobacco Use
MBRHLPG1	If you had concerns regarding your mental health, are there any reasons why you wouldn't talk to an adult at school (e.g., a school social worker, child and youth worker, counsellor, psychologist, nurse, teacher, or other staff person)? (Mark all that apply) There is no one I feel comfortable talking to.	category	0.00	Mental Health
scanID	scanID	category	0.00	Demographics

Table S1.2 Cohort 2020-2021 Input Variable for Clustering Analysis

Variable (2020-21)	**Question**	**Data Type**	**Missing Percentage**	**Category**

CESD	Depression Revised Scale (CESD R-10)	float64	15.7	Mental Health
SEVTRYM1	In the last 30 days, did you use any of the following? (Mark all that apply) Heated tobacco product (a device that heats tobacco instead of burning it, such as IQOS or Heatstick)	category	14.61	Smoking & Tobacco Use
SEVTRYO2	In the last 30 days, did you use any of the following? (Mark all that apply) Loose tobacco mixed with marijuana	category	14.61	Smoking & Tobacco Use
SEVTRYB6	In the last 30 days, did you use any of the following? (Mark all that apply) Cigars (not including cigarillos or little cigars, plain or flavoured)	category	14.61	Smoking & Tobacco Use
SEVTRYB5	In the last 30 days, did you use any of the following? (Mark all that apply) Cigarillos or little cigars (plain or flavoured)	category	14.61	Smoking & Tobacco Use
SEVTRYG4	In the last 30 days, did you use any of the following? (Mark all that apply) Nicotine patches, nicotine gum, nicotine lozenges, or nicotine inhalers	category	14.61	Smoking & Tobacco Use
SEVTRYJ2	In the last 30 days, did you use any of the following? (Mark all that apply) I have not used any of these	category	14.61	Smoking & Tobacco Use

	things in the last 30 days _ Cigarillos or little cigars (plain or flavoured) _ Cigars (not including cigarillos or little cigars, plain or flavoured) _ Loose tobacco mixed with marijuana _ Heated tobacco product (a device that heats tobacco instead of burning it, such as IQOS or Heatstick) _ Smokeless tobacco (chewing tobacco, pinch, snuff, or snus) _ Nicotine patches, nicotine gum, nicotine lozenges, or nicotine inhalers _ I have not used any of these things in the last 30 days			
SEVTRYL2	In the last 30 days, did you use any of the following? (Mark all that apply) Smokeless tobacco (chewing tobacco, pinch, snuff, or snus)	category	14.61	Smoking & Tobacco Use
CKNOWLJ1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) COVID-19 is only dangerous for the elderly population.	category	14.51	COVID

CKNOWLF1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) Hand washing for at least 20 seconds (with soap and water) helps prevent the transmission of COVID-19	category	14.51	COVID
CKNOWLH1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) When a person coughs, the use of a mask can reduce the droplet transmission of COVID-19.	category	14.51	COVID
CKNOWLI1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) COVID-19 is only found in individuals who show symptoms and signs of disease.	category	14.51	COVID
CKNOWLG1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) The use of hand sanitizer (e.g., Purell) helps prevent the transmission of COVID-19.	category	14.51	COVID
MBULLDC1	In the last 30 days, in what ways were you	category	14.12	Bullying

	bullied by other students? (Mark all that apply) Verbal attacks (e.g., getting teased, threatened, or having rumours spread about you)			
MBULLDA1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) I have not been bullied in the last 30 days	category	14.12	Bullying
MBULLDD1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Cyber-attacks (e.g., being sent mean text messages or having rumours spread about you on the internet)	category	14.12	Bullying
MBULLDF1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Social bullying (e.g., being purposefully excluded from a group, being humiliated by others with gestures or graffiti)	category	14.12	Bullying
MBULLDE1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Had someone steal from you or damage your things	category	14.12	Bullying

MBULLDB1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Physical attacks (e.g., getting beaten up, pushed, or kicked)	category	14.12	Bullying
DVHRDAVE	Average HARD activity per day (Hard activity total / 7)	float64	13.6	Physical Activity
GAD7	Anxiety Disorder (GAD7)	float64	13.15	Mental Health
DERS	Difficulties in Emotion Regulation Scale (DERS)	float64	12.89	Mental Health
CLIFCND1	How has your life changed because of COVID-19? My anxiety has…	category	12.82	COVID
CLIFCNI1	How has your life changed because of COVID-19? My physical activity has…	category	12.74	COVID
CLIFCNH1	How has your life changed because of COVID-19? My time spent surfing/posting on social media has…	category	12.65	COVID
CLIFCNC1	How has your life changed because of COVID-19? My loneliness has…	category	12.65	COVID
CLIFCNO1	How has your life changed because of COVID-19? My consumption of salty snack (e.g., chips) and sweet foods (e.g., candy, ice cream, cake) has…	category	12.62	COVID

CLIFCNJ1	How has your life changed because of COVID-19? My sleep has…	category	12.61	COVID
CLIFCNG1	How has your life changed because of COVID-19? Time spent watching TV/movies or playing video games has…	category	12.57	COVID
CLIFCNF1	How has your life changed because of COVID-19? Time spent communicating with my friends online has…	category	12.54	COVID
CLIFCNE1	How has your life changed because of COVID-19? Time spent with my friends in person has…	category	12.5	COVID
CLIFCNB1	How has your life changed because of COVID-19? My boredom has…	category	12.47	COVID
CLIFCNA1	How has your life changed because of COVID-19? My stress has…	category	12.3	COVID
FLOURISH	Flourishing Scale_self-perceived success	float64	11.56	Mental Health
DVOCONCT	School Connectedness Score	float64	11.49	School connectedness
PA_LEVEL_AVG	respondents who meet the guidelines for at least 60 minutes of physical activity per day on average	category	11.46	Physical Activity
DVACTAVE	Average combined HARD and MODERATE	float64	11.46	Physical Activity

	physical activity per day (combined activity total / 7)			
OMARKSA1	On average, what marks do you usually get in school? (Choose only one)	category	11.33	Academic Achievement
OLIKEDA1	What is the highest level of education you would like to get?	category	11.07	Academic Achievement
SEVRQTA2	Have you ever tried to quit smoking cigarettes?	category	10.83	Smoking & Tobacco Use
AMJTYPA1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have used it by smoking it (e.g., in a joint, a pipe, a bong)	category	10.64	Cannabis Use
AEVRETB1	How old were you when you first had a drink of alcohol that was more than just a sip?	category	10.39	Alcohol Use
MOPTIMI1	How much do you agree with the following statements? I generally recover from setbacks quickly	category	10.28	Mental Health
AFNTNLA1	Have you used or tried any of the following medications TO GET HIGH? B) Fentanyl (china white, synthetic heroin, china girl)	category	10.21	Opioid Use
QCHEFFF1	For each of the following behaviours,	category	10.12	Cannabis Use

	please indicate the extent to which you believe an individual would be exposed to risks to his or her physical health: S/he smokes cannabis (marijuana, hashish) regularly			
QCHEFFE1	For each of the following behaviours, please indicate the extent to which you believe an individual would be exposed to risks to his or her physical health: S/he has tried cannabis (marijuana, hashish) a few times	category	10.12	Cannabis Use
QCHEFFD1	For each of the following behaviours, please indicate the extent to which you believe an individual would be exposed to risks to his or her physical health: S/he smokes e-cigarettes regularly	category	10.07	Vaping & E-Cigarette Use
APRSCPA1	Have you used or tried any of the following medications TO GET HIGH? C)Other prescription pain relievers (codeine, morphine, Tylenol 3)	category	10.02	Opioid Use
MFAMSUB1	How much do you agree or disagree with the following statements?My parents/guardians	category	9.94	Mental Health

	expect too much of me			
ARISKBB1	In the last 30 days, how many times have you done the following? Rode in/on a vehicle when the driver (you or someone else) had been using marijuana or cannabis	category	9.93	Cannabis Use
QCHEFFC1	For each of the following behaviours, please indicate the extent to which you believe an individual would be exposed to risks to his or her physical health: S/he has tried e-cigarettes a few times	category	9.87	Vaping & E-Cigarette Use
tot_fruit	Total number of days of Eat fruit (fresh, canned, dried, or frozen) in the past week	float64	9.83	Healthy Eating
tot_vege	Total number of days of Eat vegetables (raw or cooked, fresh, canned, or frozen) in the past week	float64	9.83	Healthy Eating
tot_restrict_diet	Total number of days of Restrict/alter your food intake with the intention of changing your weight/shape in the past week	float64	9.83	Healthy Eating
tot_snacks	Total number of days of eat snacks purchased from a vending machine, corner store, snack bar, or canteen in the past week	float64	9.83	Healthy Eating

tot_water	Total number of days of drink water in the past week	float64	9.83	Healthy Eating
tot_fast_food	Total number of days of Eat food purchased at a fast food place or restaurant in the past week	float64	9.83	Healthy Eating
AOXYCDA1	Have you used or tried any of the following medications TO GET HIGH? A) Oxycodone (oxy, OC, APO, OxyContin®, percs, roxies, OxyNEO®)	category	9.78	Opioid Use
MFAMSUD1	How much do you agree or disagree with the following statements? I can talk about my problems with my friends	category	9.75	Mental Health
MFAMSUC1	How much do you agree or disagree with the following statements? I can talk about my problems with my family	category	9.74	Mental Health
APRHRDA1	Do you think it would be difficult or easy to get pain relievers (Oxycodone, Fentanyl, codeine, etc.) if you wanted some?	category	9.72	Opioid Use
AHWHRDA2	Do you think it would be difficult or easy for you to get marijuana if you wanted some?	category	9.56	Cannabis Use

MFAMSUA1	How much do you agree or disagree with the following statements? I have a happy home life	category	9.43	Mental Health
ARISKBA1	In the last 30 days, how many times have you done the following? Rode in/on a vehicle when the driver (you or someone else) had been drinking alcohol	category	9.42	Alcohol Use
AEVRMJB1	How old were you when you first used marijuana or cannabis?	category	9.03	Cannabis Use
AMJTYPD1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have not used marijuana or cannabis in the last 12 months	category	8.73	Cannabis Use
AMJTYPB1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have used it by vaping it	category	8.73	Vaping & E-Cigarette Use
AMJTYPC1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have used it by eating or drinking it (e.g., in brownies, cookies, candies, tea)	category	8.73	Cannabis Use
HSKBRKG1	If you do not eat breakfast every day,	category	8.67	Healthy Eating

	why do you skip breakfast? I'm trying to lose weight			
HSKBRKA1	If you do not eat breakfast every day, why do you skip breakfast? I eat breakfast every day	category	8.67	Healthy Eating
HSKBRKB1	If you do not eat breakfast every day, why do you skip breakfast? I don't have time for breakfast	category	8.67	Healthy Eating
HSKBRKH1	If you do not eat breakfast every day, why do you skip breakfast? There is nothing to eat at home	category	8.67	Healthy Eating
HSKBRKC1	If you do not eat breakfast every day, why do you skip breakfast? The bus comes too early	category	8.67	Healthy Eating
HSKBRKD1	If you do not eat breakfast every day, why do you skip breakfast? I sleep in	category	8.67	Healthy Eating
HSKBRKE1	If you do not eat breakfast every day, why do you skip breakfast? I'm not hungry in the morning	category	8.67	Healthy Eating
HSKBRKF1	If you do not eat breakfast every day, why do you skip breakfast?I feel sick when I eat breakfast	category	8.67	Healthy Eating
A5DRNKC1	In the last 12 months, how often did you have 5 drinks of	category	8.55	Alcohol Use

	alcohol or more on one occasion?			
AOFTMJA2	In the last 12 months, how often did you use marijuana or cannabis? (a joint, pot, weed, hash)	category	8.53	Cannabis Use
ADRINKA1	In the last 12 months, how often did you have a drink of alcohol that was more than just a sip?	category	8.41	Alcohol Use
PSPORTC5	Do you participate in before-school, noon hour, or after school physical activities organized by your school? (e.g., intramurals, non-competitive clubs)	category	8.28	Physical Activity
PSTRENA3	On how many days in the last 7 days did you do exercises to strengthen or tone your muscles? (e.g., push-ups, sit-ups, or weight-training)	category	8.28	Physical Activity
SECGGTA1	Where do you usually get vapes (including pods, e-juice or e-liquid)? (Choose only one)	category	8.2	Vaping & E-Cigarette Use
SUSECGC2	Have you used a vape for any of the following reasons? (Mark all that apply) Curiosity / to try something new	category	8	Vaping & E-Cigarette Use
SUSECGB2	Have you used a vape for any of the following reasons? (Mark all that apply)	category	8	Vaping & E-Cigarette Use

	I have not used a vape			
SUSECGD2	Have you used a vape for any of the following reasons? (Mark all that apply) I can vape in places where smoking is not allowed	category	8	Vaping & E-Cigarette Use
SUSECGH2	Have you used a vape for any of the following reasons? (Mark all that apply) It's cool / a lot of peole I know use them	category	8	Vaping & E-Cigarette Use
SUSECGF2	Have you used a vape for any of the following reasons? (Mark all that apply) To help me quit smoking cigarettes	category	8	Vaping & E-Cigarette Use
SUSECGI2	Have you used a vape for any of the following reasons? (Mark all that apply) To relax or relieve stress and anxiety	category	8	Vaping & E-Cigarette Use
SUSECGJ2	Have you used a vape for any of the following reasons? (Mark all that apply) For the nicotine high	category	8	Vaping & E-Cigarette Use
SUSECGK2	Have you used a vape for any of the following reasons? (Mark all that apply) I really like the flavours	category	8	Vaping & E-Cigarette Use
SUSECGG2	Have you used a vape for any of the following reasons? (Mark all that apply)	category	8	Vaping & E-Cigarette Use

	I have used a vape for some other reason			
SECGSHA1	When you vape, how often do you share the device with someone else or use someone else's vape?	category	7.98	Vaping & E-Cigarette Use
SLP_AVG	Average length of sleep per night in a week	float64	7.94	Sleep
SLP_GUIDE	Respondents who meet the guidelines for at least 8 hours of sleep per day (on average)	category	7.94	Sleep
SECG30A2	On how many of the last 30 days did you use a vape?	category	7.92	Vaping & E-Cigarette Use
PSPORTD5	Do you participate in competitive school sports teams that compete against other schools? (e.g., junior varsity or varsity sports)	category	7.88	Physical Activity
SUSECGA3	Have you ever tried a vape, also known as an e-cigarette? (e.g., JUUL, Vype, Suorin, Smok)	category	7.87	Vaping & E-Cigarette Use
DSUSCEPT3	Smoking Susceptibility - Three Measures	category	7.86	Smoking & Tobacco Use
PSPORTD4	Do you participate in league or team sports outside of school?	category	7.75	Physical Activity
SLST30A1	On how many of the last 30 days did you smoke one or more cigarettes?	category	7.69	Smoking & Tobacco Use
SPUFF0A1	Have you ever tried cigarette smoking,	category	7.64	Smoking & Tobacco Use

	even just a few puffs?			
PSLPEZA1	During the past week, what time have you usually turned out the light and gone to sleep… on weekends? (Select the time using the drop-down lists for hour, minutes and AM/PM) - AM/PM	category	7.02	Sleep
PSLPDZA1	During the past week, what time have you usually turned out the light and gone to sleep… on weekdays? (Select the time using the drop-down lists for hour, minutes and AM/PM) - AM/PM	category	6.82	Sleep
PWAKEZA1	During the past week, what time have you usually woken up in the morning… on weekends? (Select the time using the drop-down lists for hour, minutes and AM/PM) - AM/PM	category	6.79	Sleep
PWAKDZA1	During the past week, what time have you usually woken up in the morning… on weekdays? (Select the time using the drop-down lists for hour, minutes and AM/PM) - AM/PM	category	6.64	Sleep
DTDSEDNA_12	Total daily sedentary	float64	5.76	Screen Time
PTEXT_12	Texting, messaging : Minutes per day	float64	5.76	Screen Time

DTDSCRNT_12	Total daily sedentary activity, with homework excluded	float64	5.76	Screen Time
PSURF_12	Surfing Internet : Minutes per day	float64	5.76	Screen Time
PHMWK_12	Doing Homework: Minutes per day	float64	5.76	Screen Time
PVIDE_12	Video Games : Minutes per day	float64	5.76	Screen Time
PTVMO_12	Watching TV : Minutes per day	float64	5.76	Screen Time
PSLPQTA1	During the past week, how would you rate your sleep quality overall (how well you sleep)?	category	5.14	Sleep
HDSCWTA1	How do you describe your weight?	category	3.56	Height & Weight
GMONEYC1	Where do you get money to spend on yourself or to save? My parents/guardians give me money (e.g., an allowance)	category	3.38	Demographics
GMONEYB1	Where do you get money to spend on yourself or to save? I do not usually get any money to spend on myself or to save	category	3.38	Demographics
GMONEYE1	Where do you get money to spend on yourself or to save? I get paid cash for occasional work (babysitting, mowing lawns, shovelling snow, etc.)	category	3.38	Demographics
GMONEYD1	Where do you get money to spend on yourself or to save? I get a pay cheque from a job (working	category	3.38	Demographics

	evenings or weekends at a restaurant, store, etc.)			
OWEIGHA1	How much do you weigh without your shoes on? (Please enter your weight in either pounds OR in kilograms)	category	3.21	Height & Weight
HWEGHTA1	Which of the following are you trying to do about your weight?	category	3.15	Height & Weight
OHEIGHA1	How tall are you without your shoes on? (Please enter your height in either feet and inches OR in centimetres)	category	2.83	Height & Weight
QCHCOMF1	Would you say that you and your family are more or less comfortable financially than the average student in your class?	category	2.73	Demographics
QCOBEDA1	In your house, do you have your own bedroom?	category	2.35	Demographics
GATHOMA1	Including you, how many people are currently living at your home?	category	2.33	Demographics
GMONEYA1	About how much money do you usually get each week to spend on yourself or to save? (Remember to include all money from allowances and jobs like baby-sitting, delivering papers, etc.)	category	2.32	Demographics

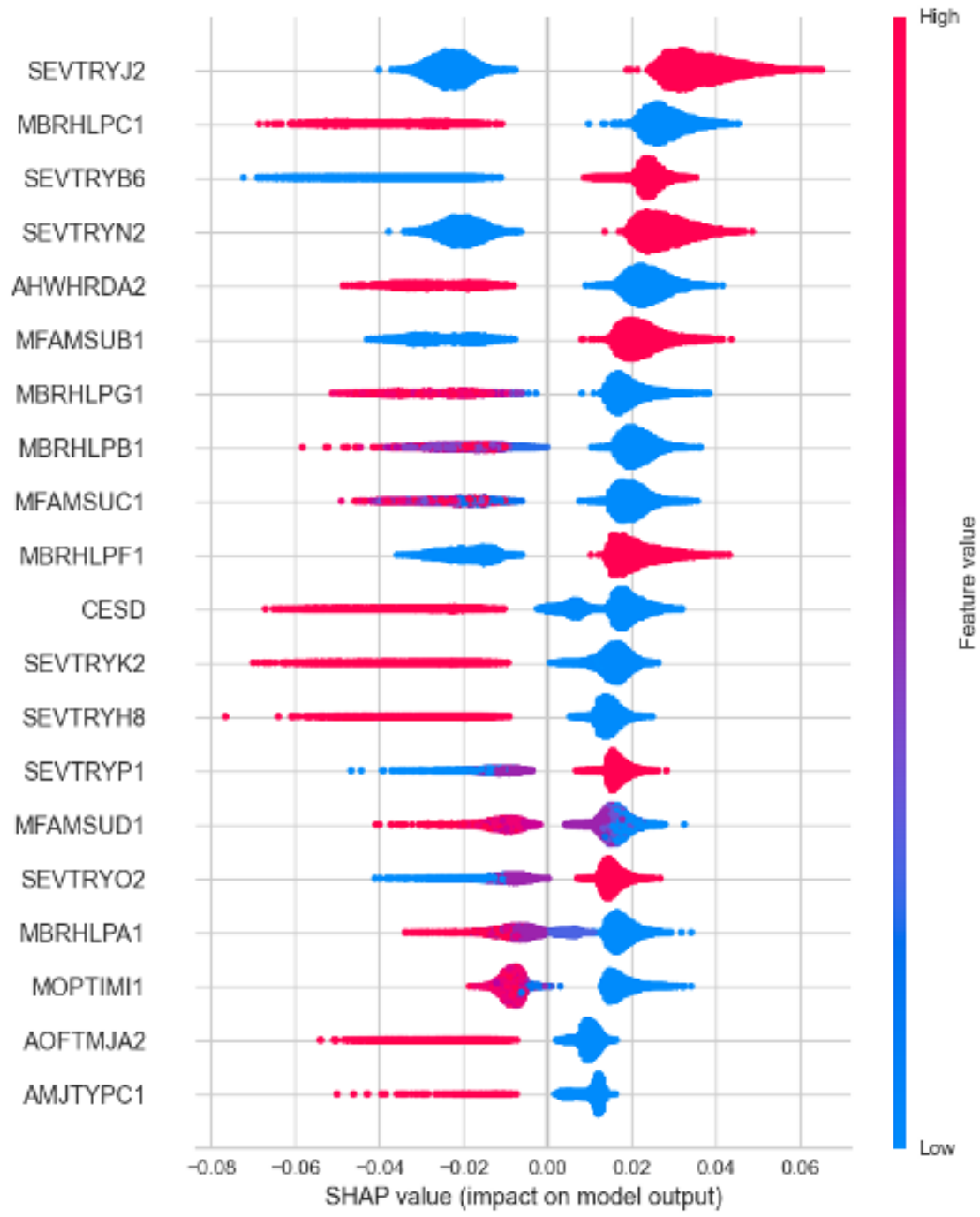
QCHNGRA1	Do you sometimes go to bed hungry because there is not enough money to buy food?	category	2.27	Demographics
CCURRIA1	What form of curriculum delivery are you currently participating in?	category	1.71	Demographics
GETHNCM1	How would you describe yourself? Other	category	1.65	Demographics
GETHNCL1	How would you describe yourself? White	category	1.65	Demographics
GETHNCH1	How would you describe yourself? Latin American/Hispanic	category	1.65	Demographics
GETHNCC1	How would you describe yourself? Black	category	1.65	Demographics
GETHNCN1	How would you describe yourself? Asian	category	1.65	Demographics
SEX4	Are you female or male?	category	1.47	Demographics
AGE	How old are you today?	category	1.37	Demographics
GSMSKLA2	Did you attend this school last year?	category	0.95	Demographics
province1	province	category	0	Demographics
LANGUAGE	Questionnaire language administered.	category	0	Demographics
PSURFHA3	How much time per day do you usually spend doing the following activities? Surfing the internet; Hours	category	0	Screen
DVMDRAVE	Average MODERATE activity	float64	0	Physical Activity

	per day (Mod_activity_total / 7)			
DSUSCEPT2	Smoking Susceptibility - Two Measures	category	0	Smoking & Tobacco Use
BMI_ACAT	BMI catagories	category	0	Weight
CLIFCNN1	How has your life changed because of COVID-19? My cannabis use has…	category	0	COVID
CLIFCNM1	How has your life changed because of COVID-19? My alcohol use has…	category	0	COVID
CLIFCNL1	How has your life changed because of COVID-19? My vaping has…	category	0	COVID
CLIFCNK1	How has your life changed because of COVID-19? My cigarette use has…	category	0	COVID
PSURFMA2	How much time per day do you usually spend doing the following activities? Surfing the internet; Minutes	category	0	Screen
ARISKBD1	In the last 30 days, how many times have you done the following? Used marijuana or cannabis when you were all by yourself	category	0	Cannabis Use
ARISKBC1	In the last 30 days, how many times have you done the following? Drank alcohol when you were all by yourself	category	0	Alcohol Use
SCHOOLID	School ID	float64	0	Demographics

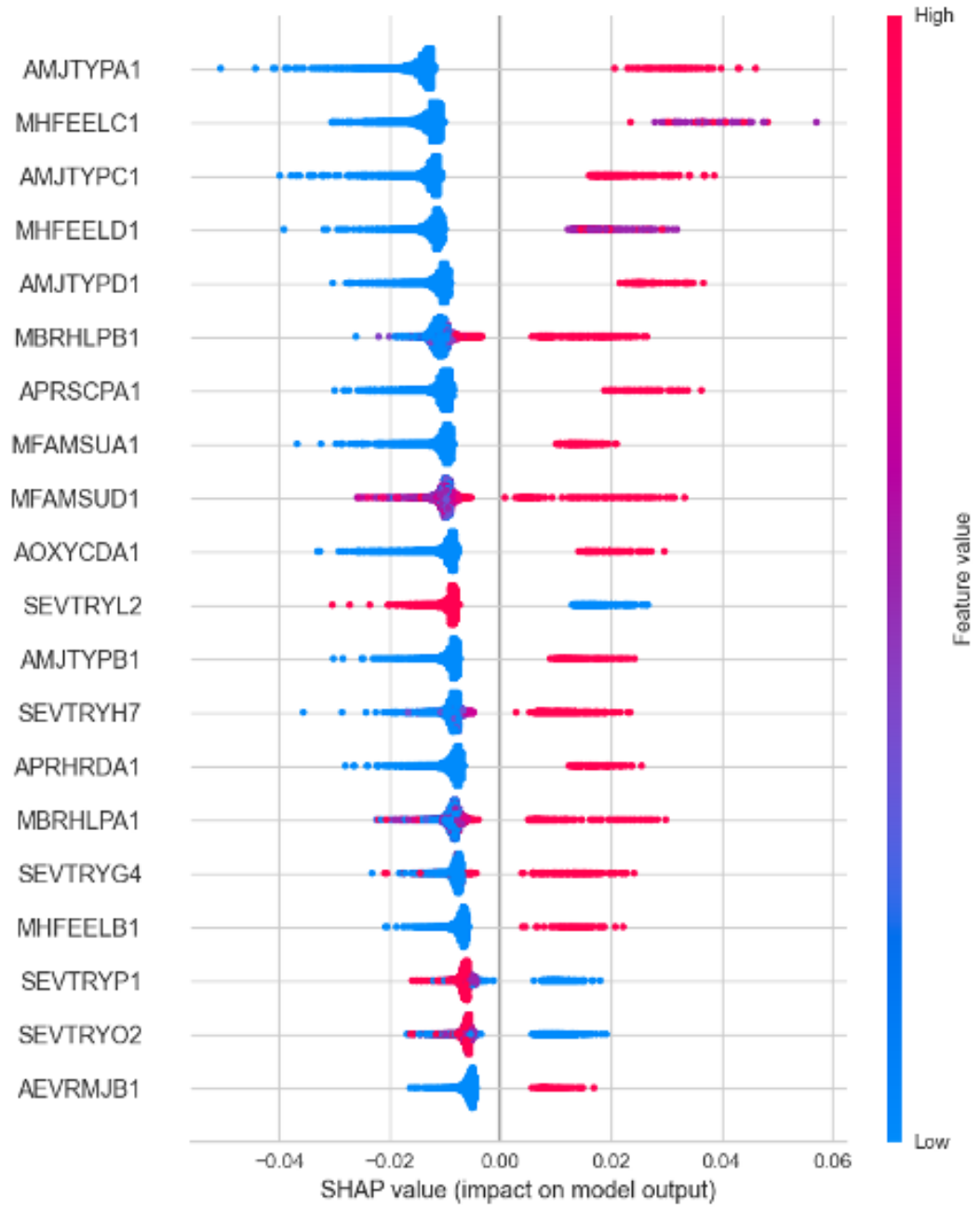
scanID	ScanID	float64	0	Demographics

Figure S1.1 (A-C) 2018-2019: SHAP summary plot for high level clusters

A. SHAP summary plot for high level Cluster 0



B. SHAP summary plot for high level Cluster 1



C. SHAP summary plot for high level Cluster 2

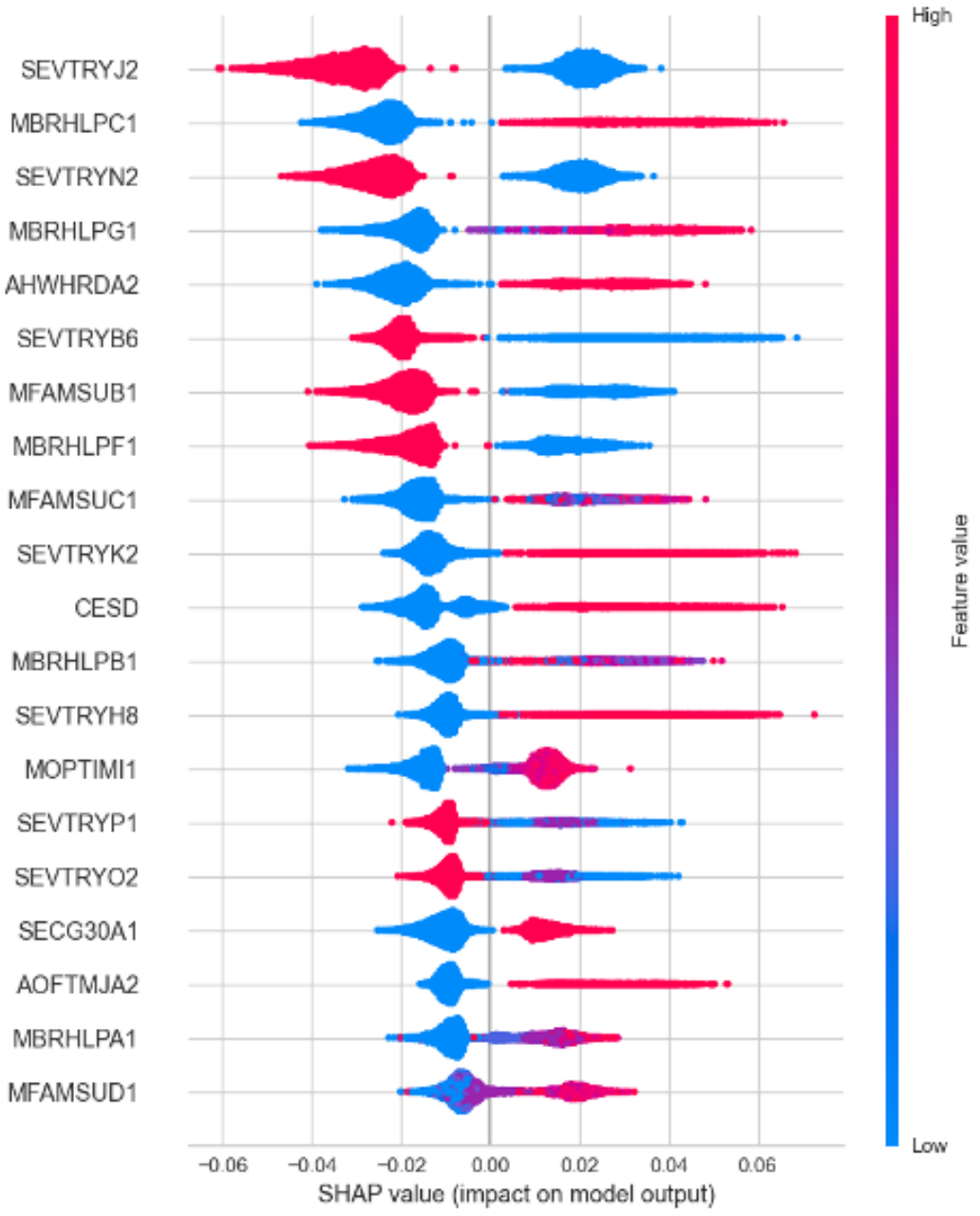
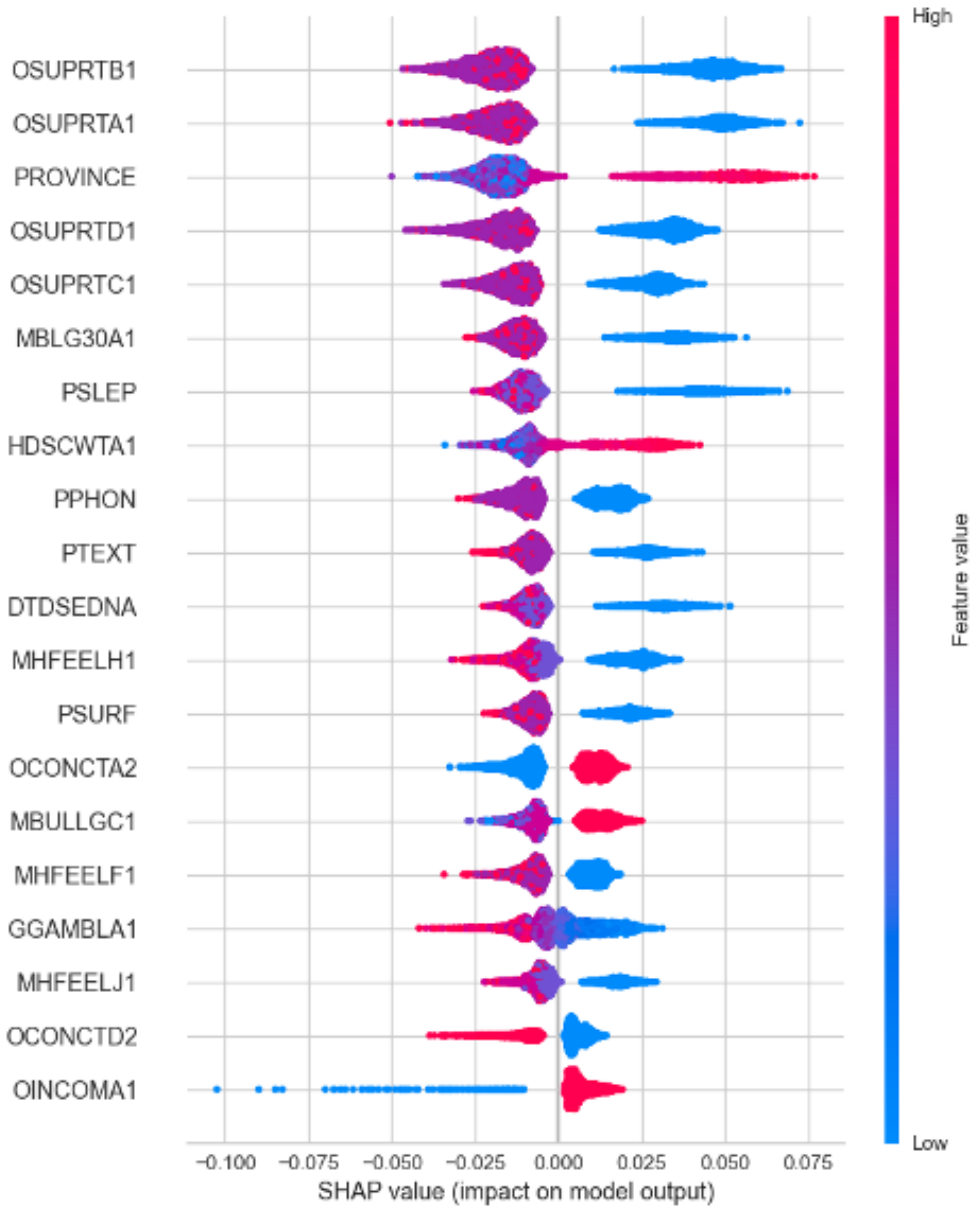
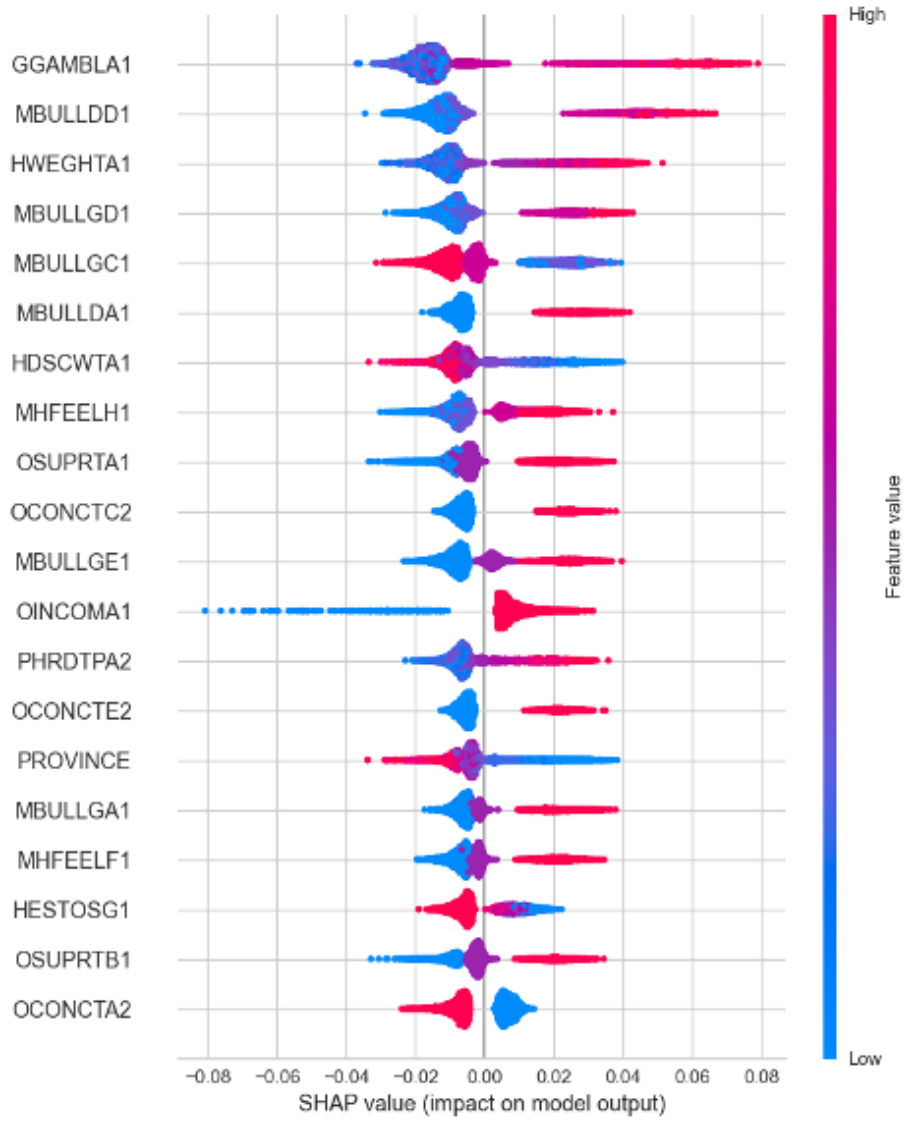


Figure S1.2 (A-E) 2018-2019: SHAP summary plot for sub clusters in high-level cluster 0

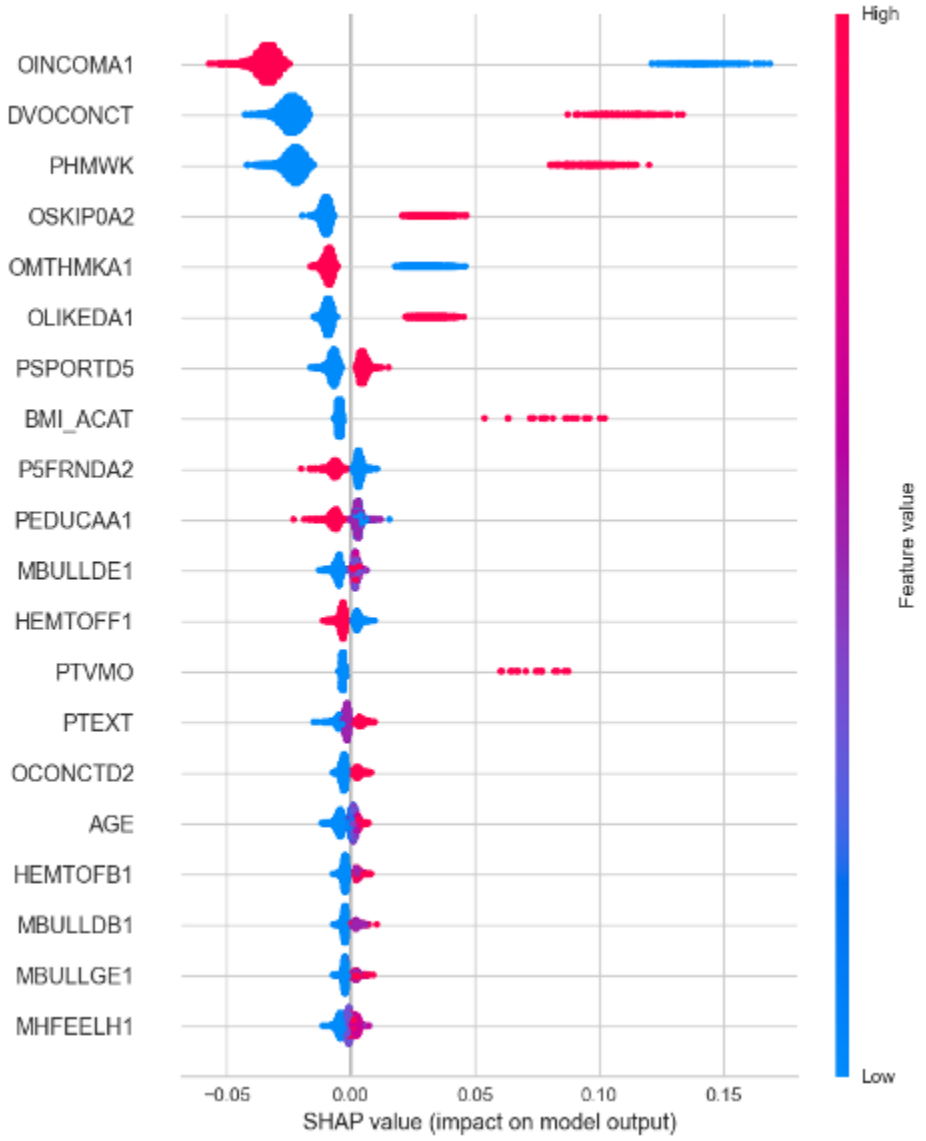
A. SHAP summary plot for Sub Cluster 0



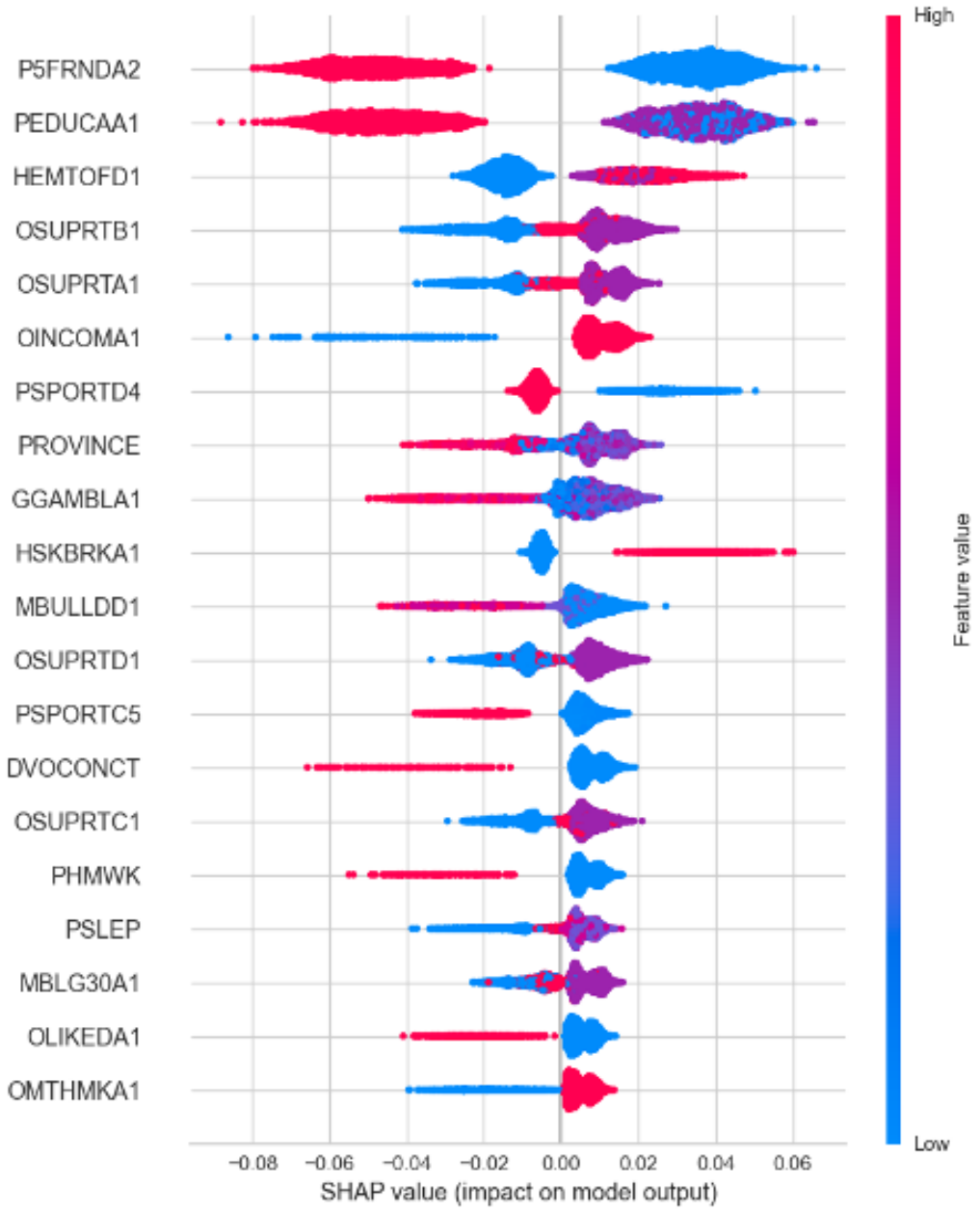
B. SHAP summary plot for Sub Cluster 1



C. SHAP summary plot for Sub Cluster 2



D. SHAP summary plot for Sub Cluster 3



E. SHAP summary plot for Sub Cluster 4

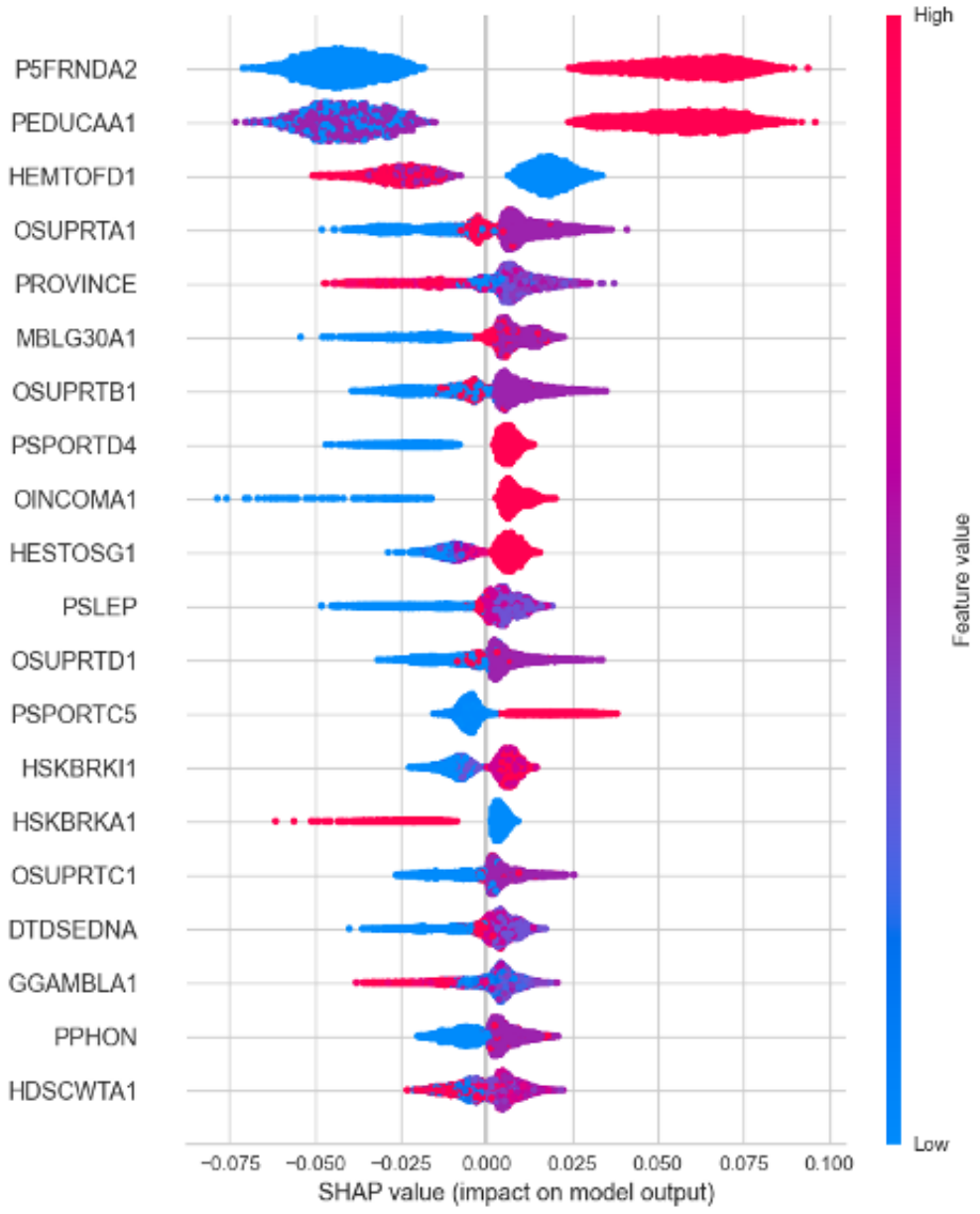
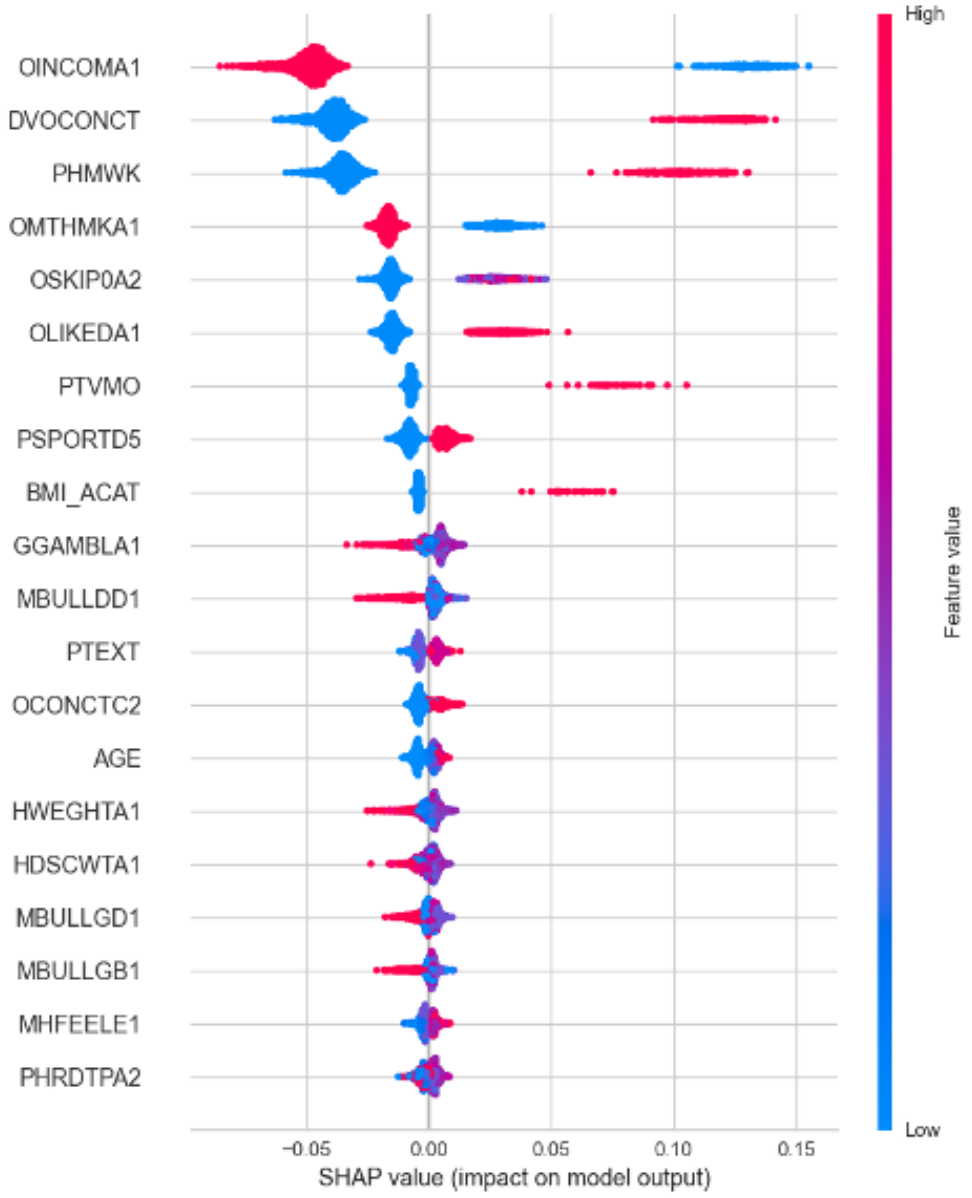
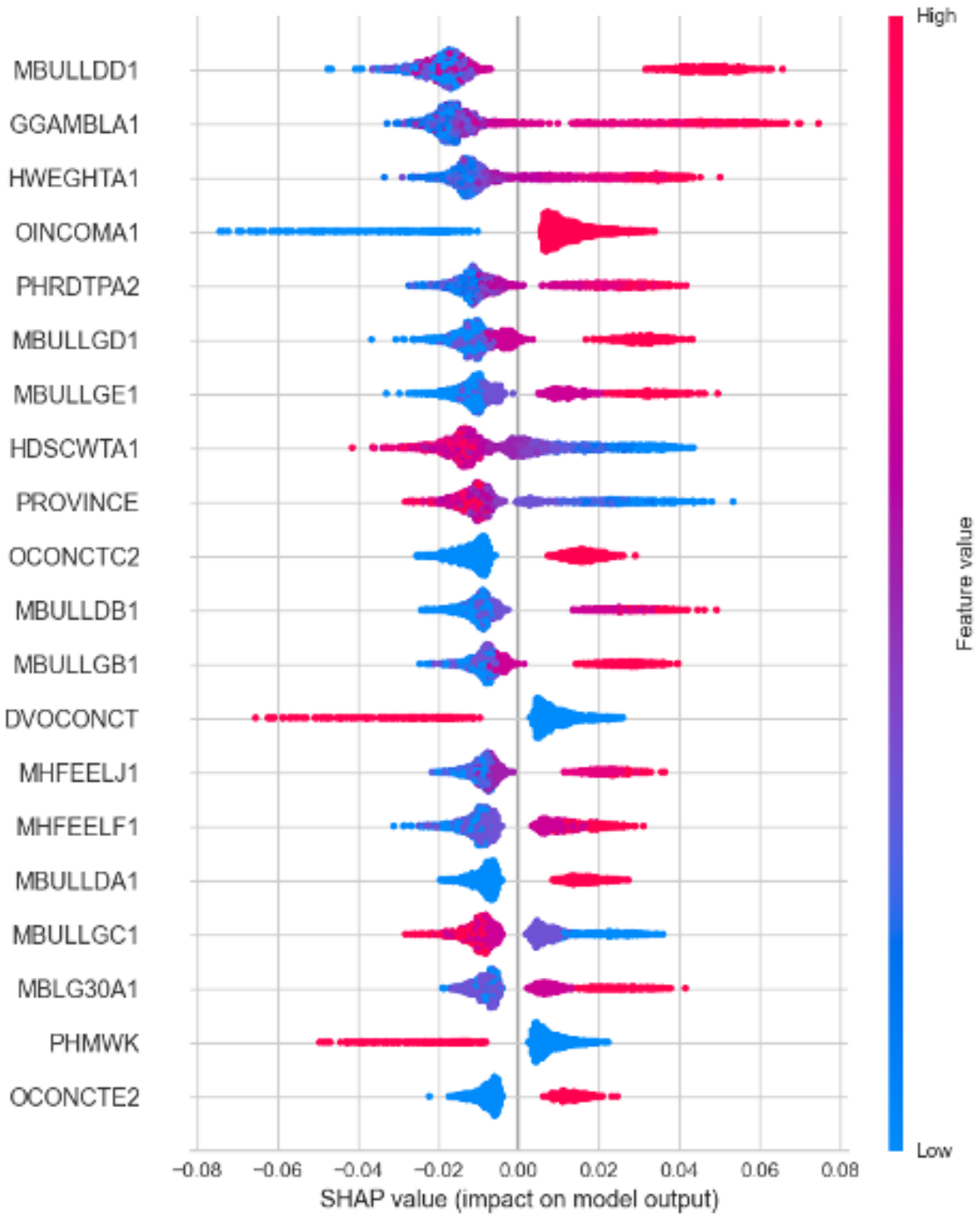


Figure S1.3 (A-D) 2018-2019: SHAP summary plot for sub clusters in high-level cluster 2

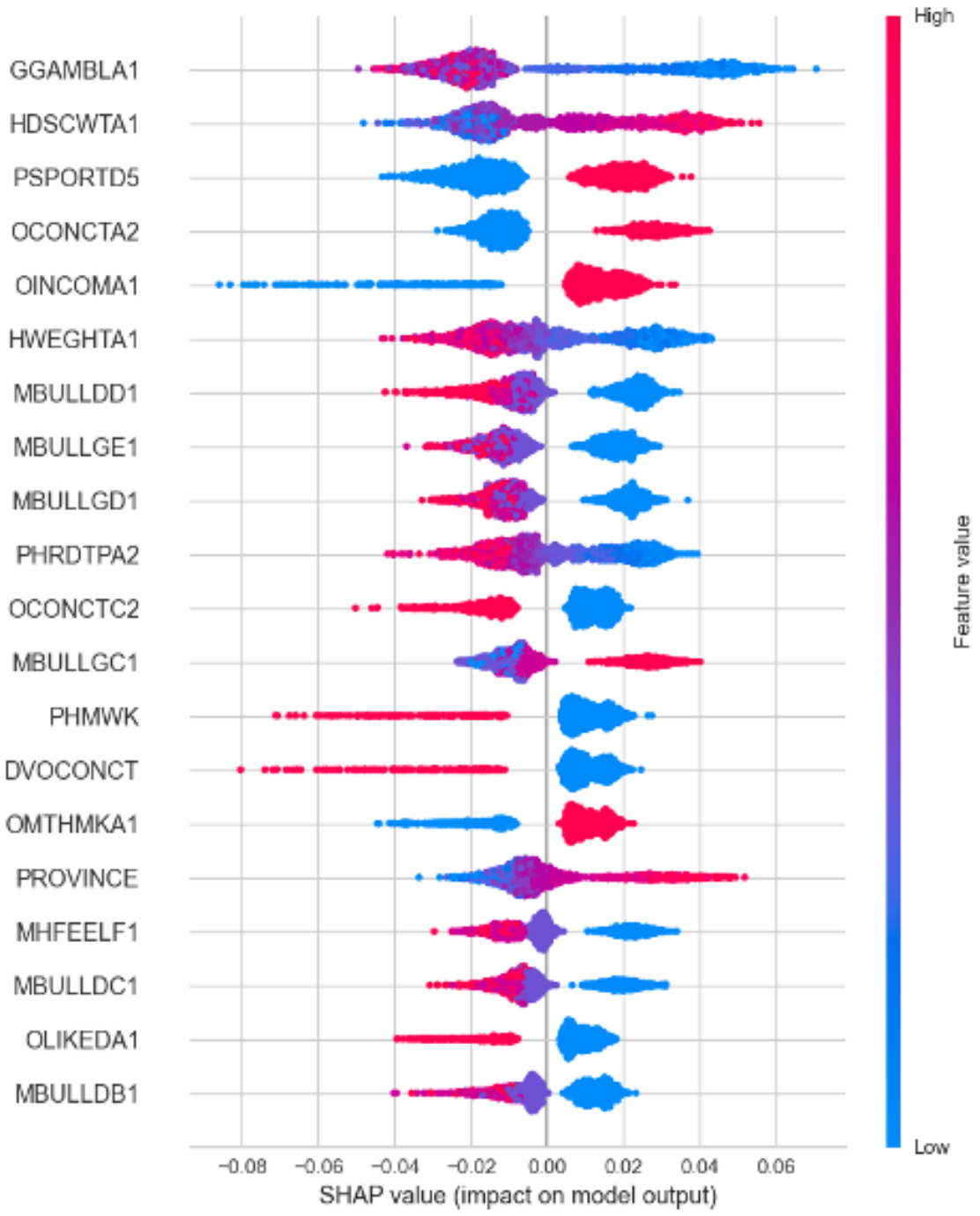
A. SHAP summary plot for Sub Cluster 0



B. SHAP summary plot for Sub Cluster 1



C. SHAP summary plot for Sub Cluster 2



D. SHAP summary plot for Sub Cluster 3

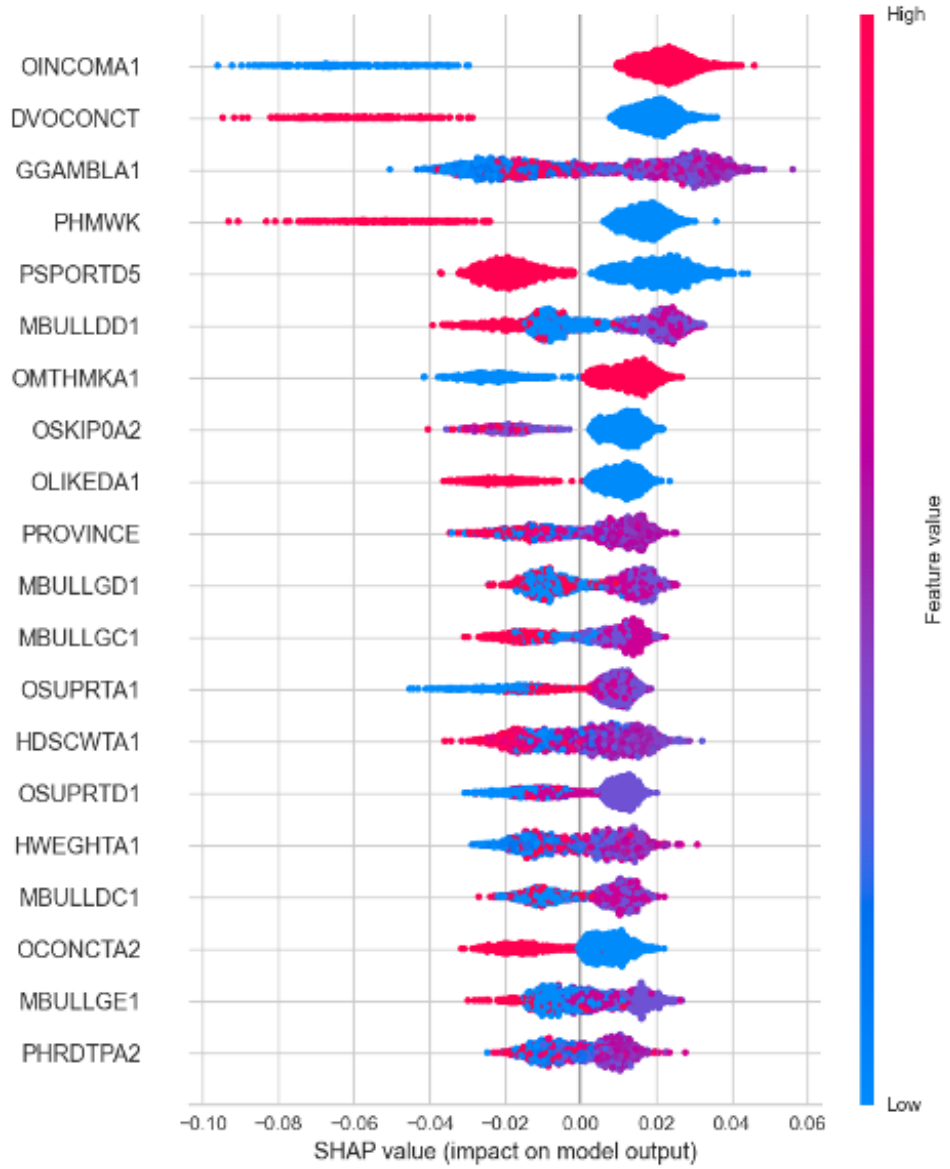
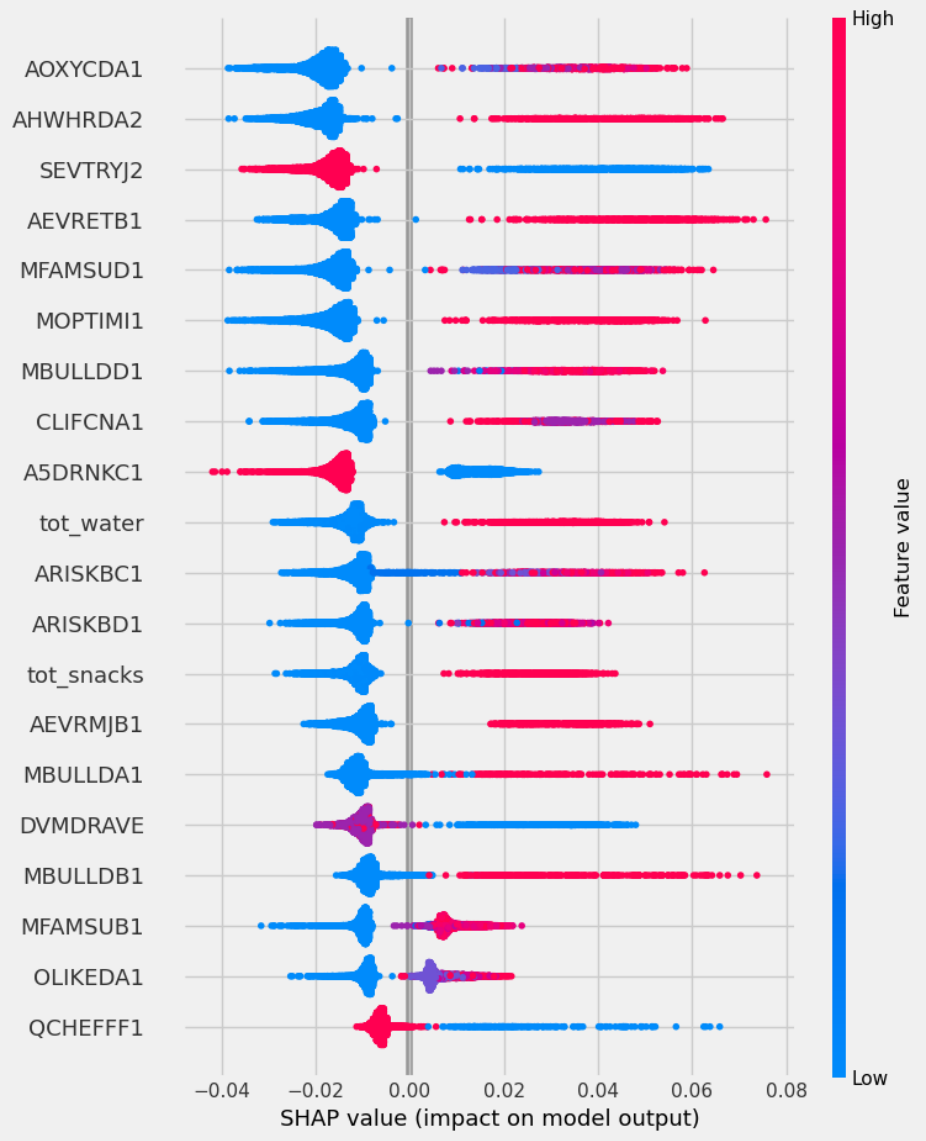
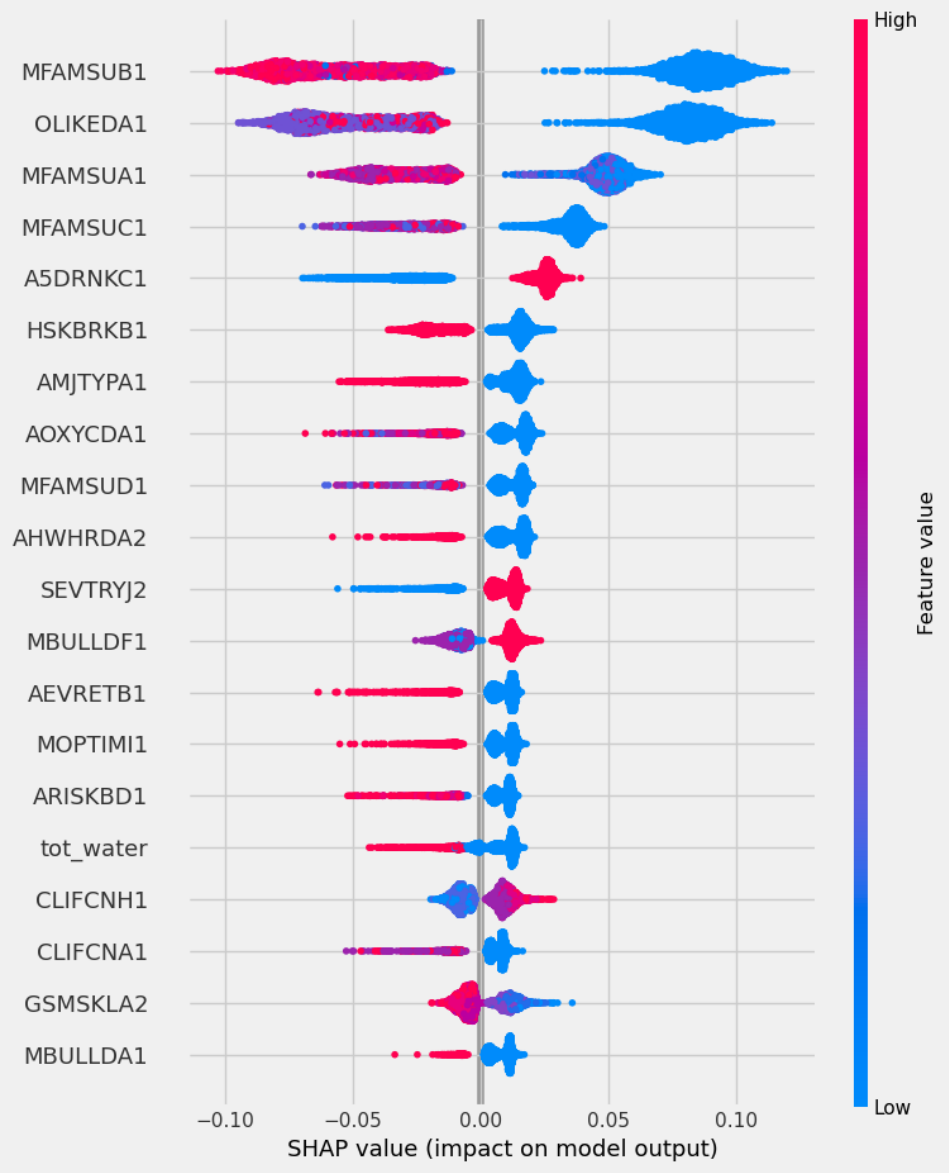


Figure S1.4 (A-D) 2020-2021: SHAP summary plot for high-level cluster

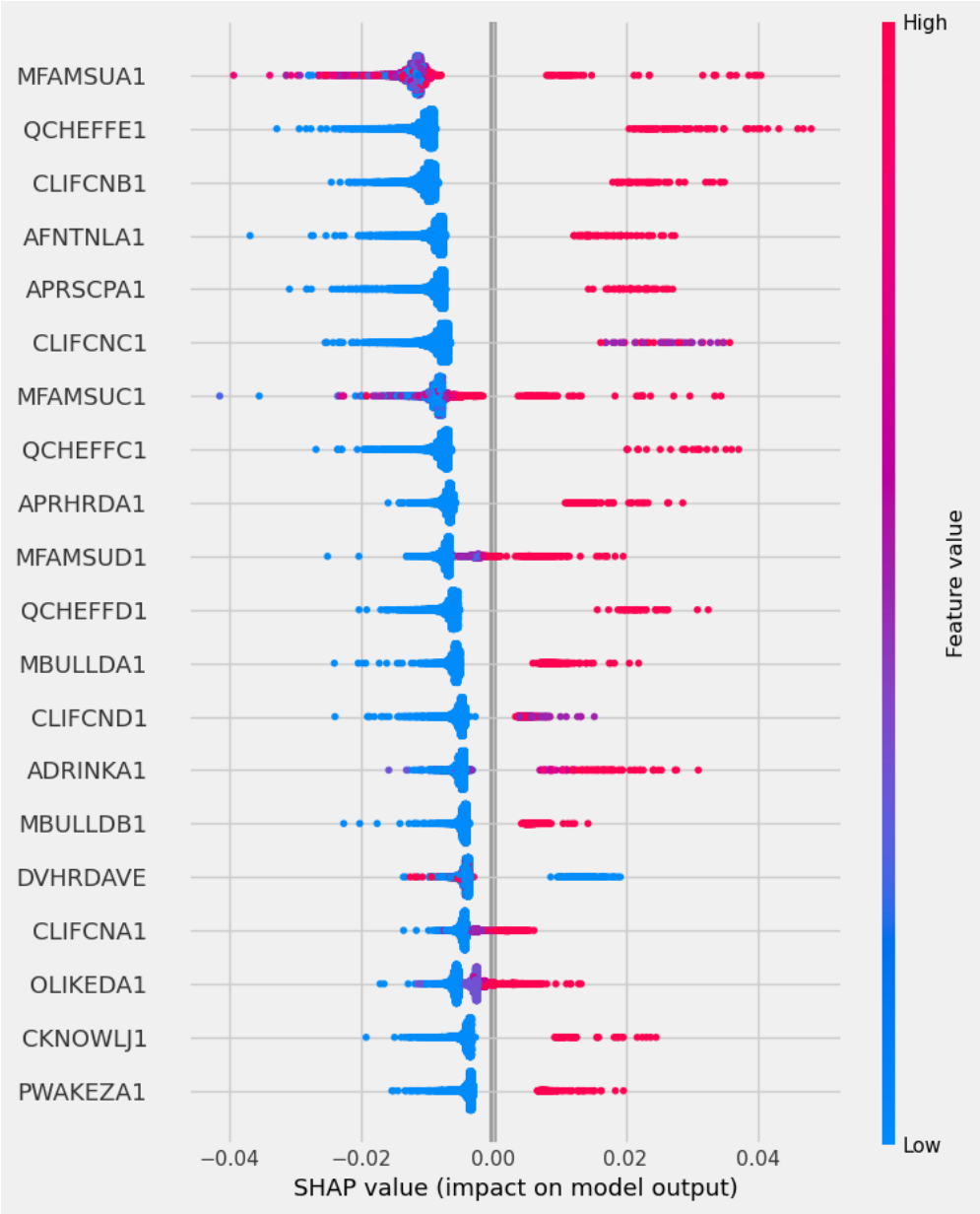
A. SHAP summary plot for High-level Cluster 0



B. SHAP summary plot for High-level Cluster 1



C. SHAP summary plot for High-level Cluster 2



D. SHAP summary plot for High-level Cluster 3

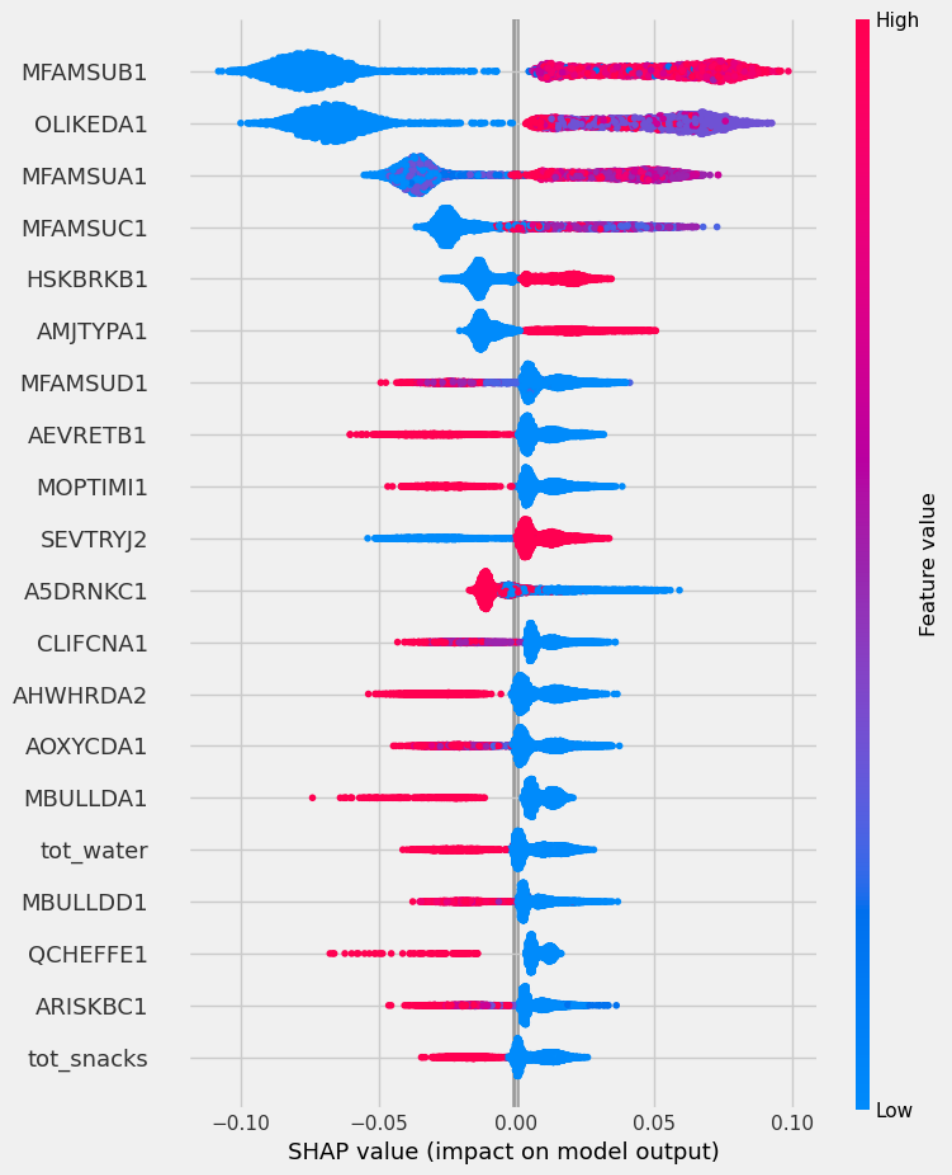
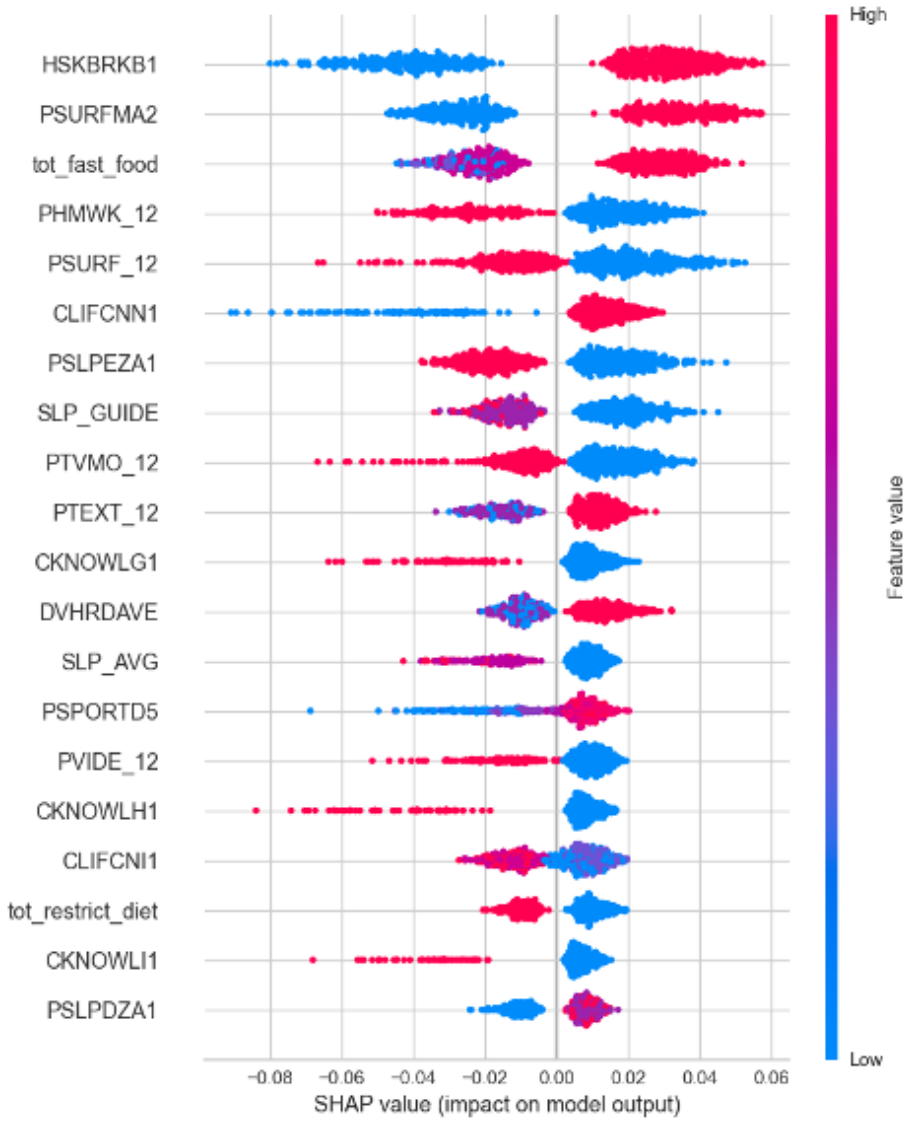
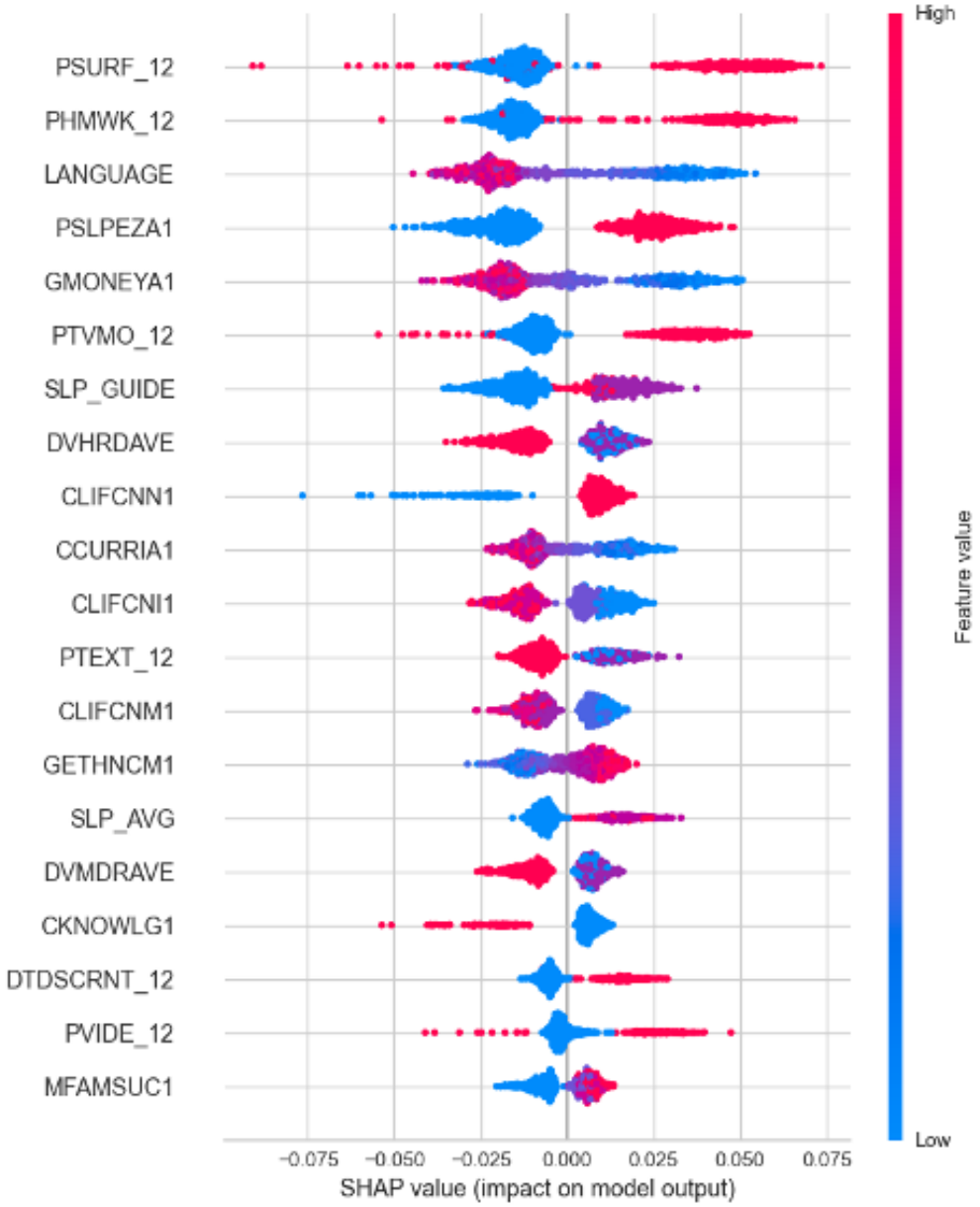


Figure S1.5 (A-E) 2020-2021: SHAP summary plot for sub clusters in high-level cluster 0

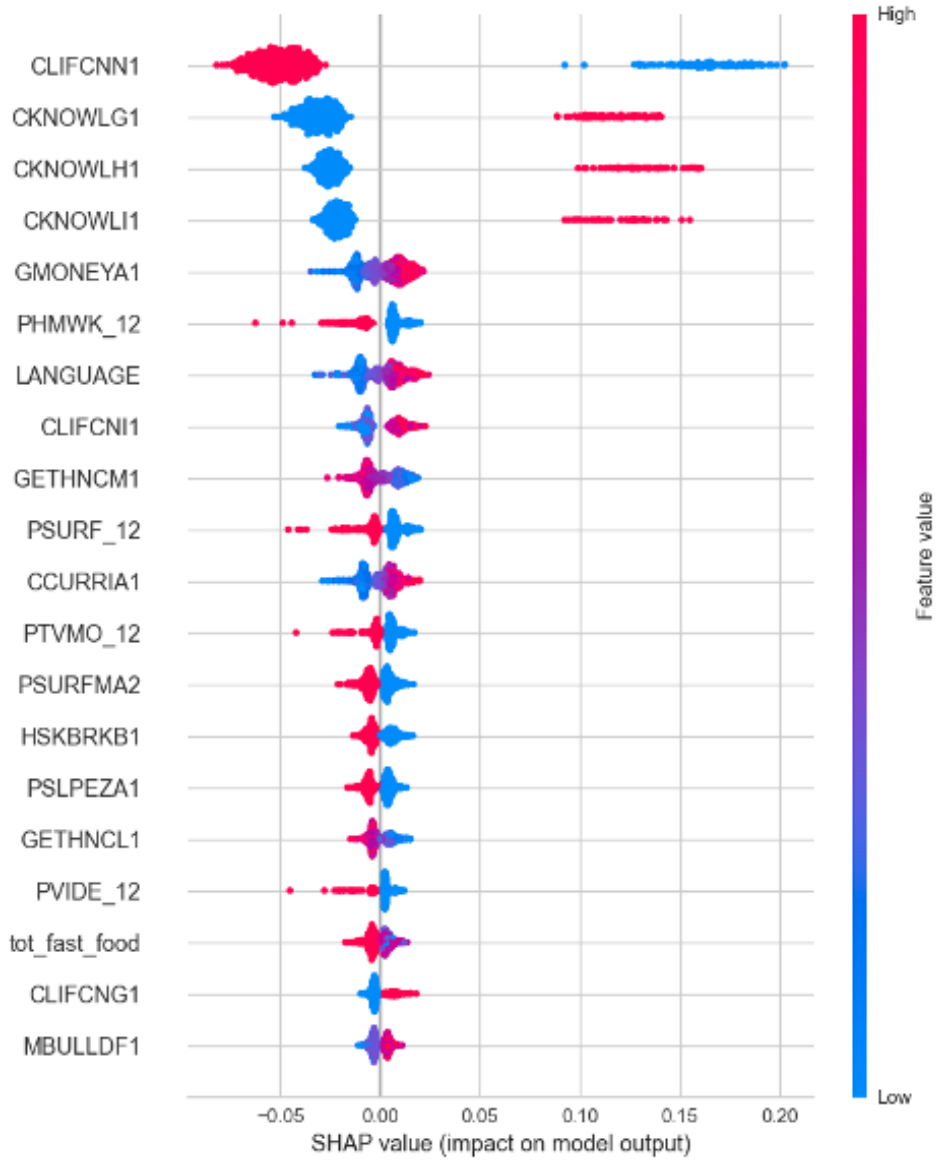
A. SHAP summary plot for Sub Cluster 0



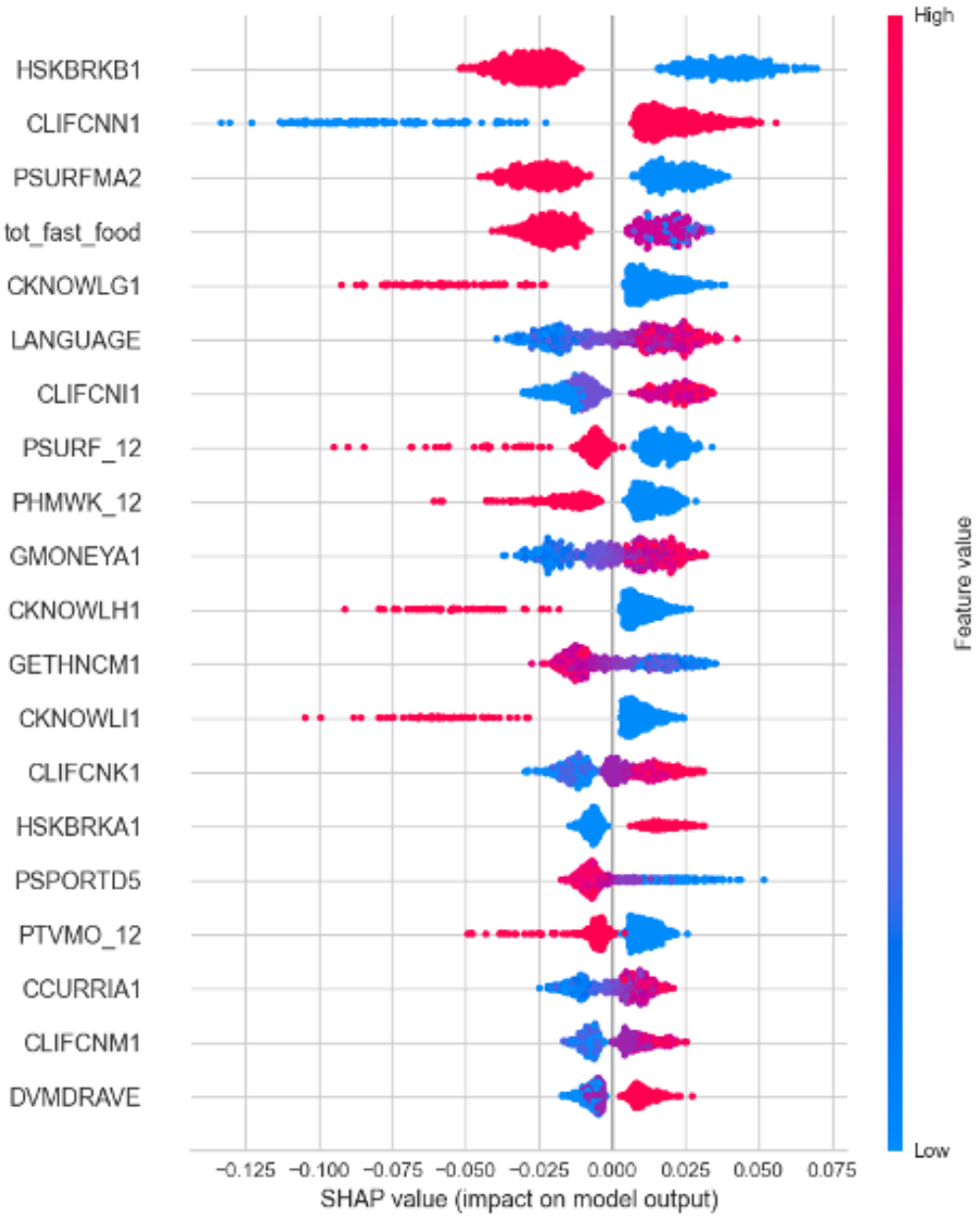
B. SHAP summary plot for Sub Cluster 1



C. SHAP summary plot for Sub Cluster 2



D. SHAP summary plot for Sub Cluster 3



E. SHAP summary plot for Sub Cluster 4

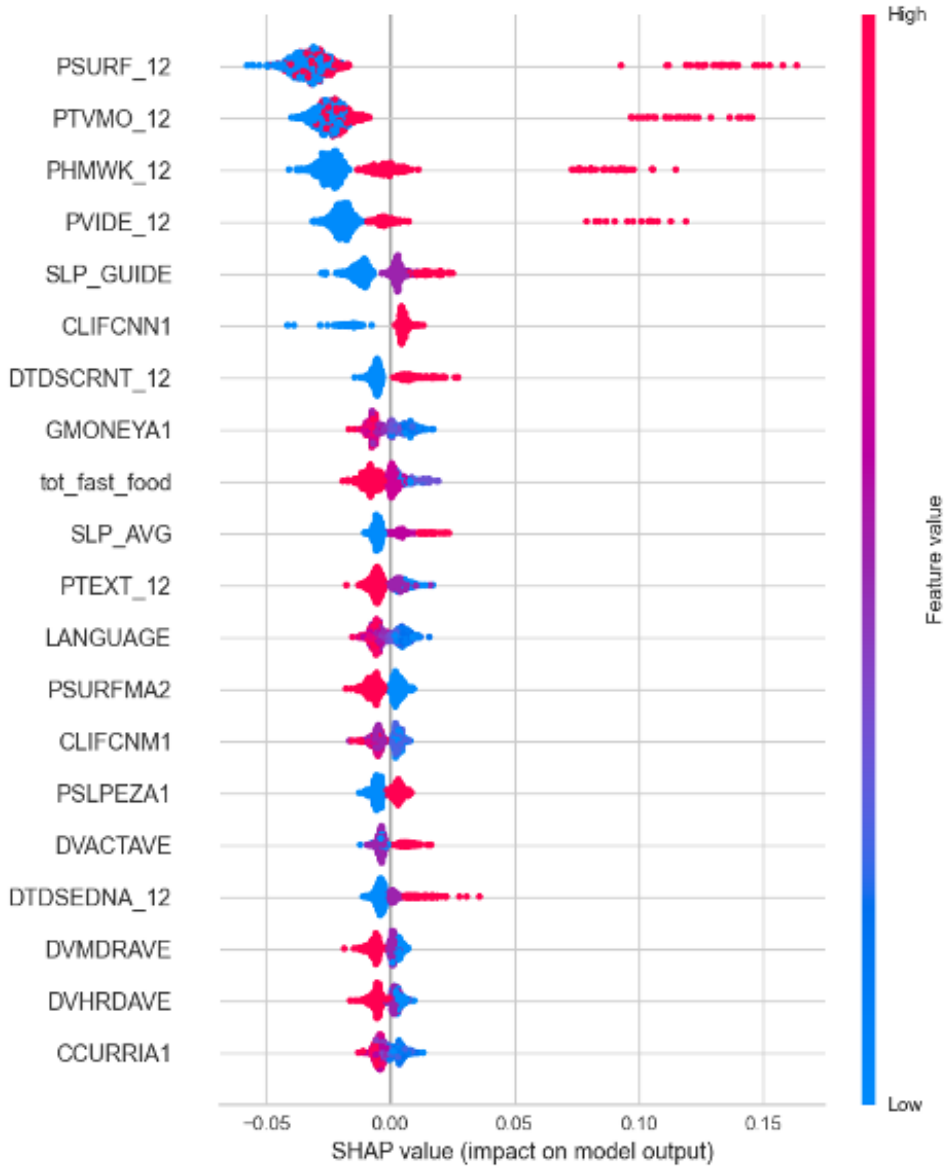
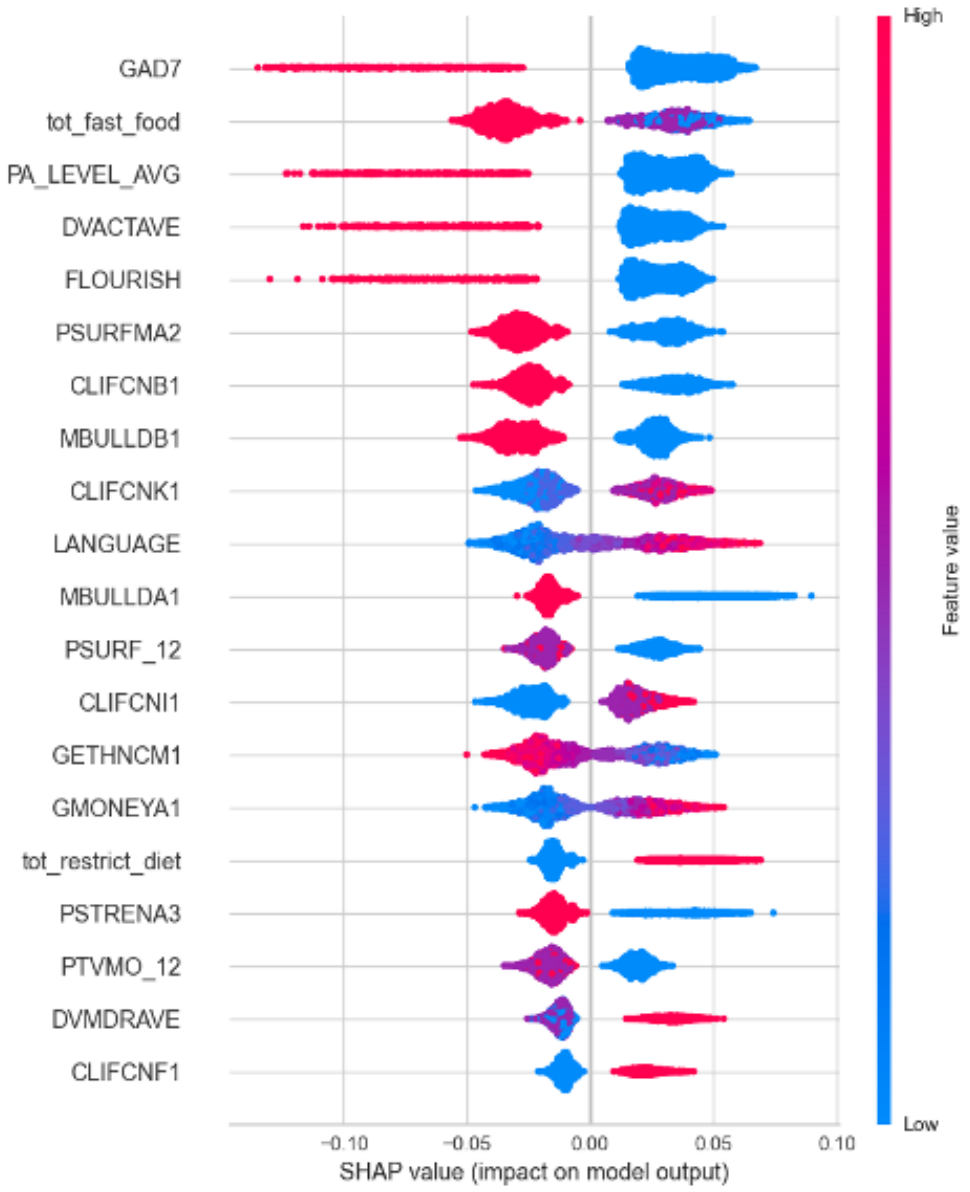
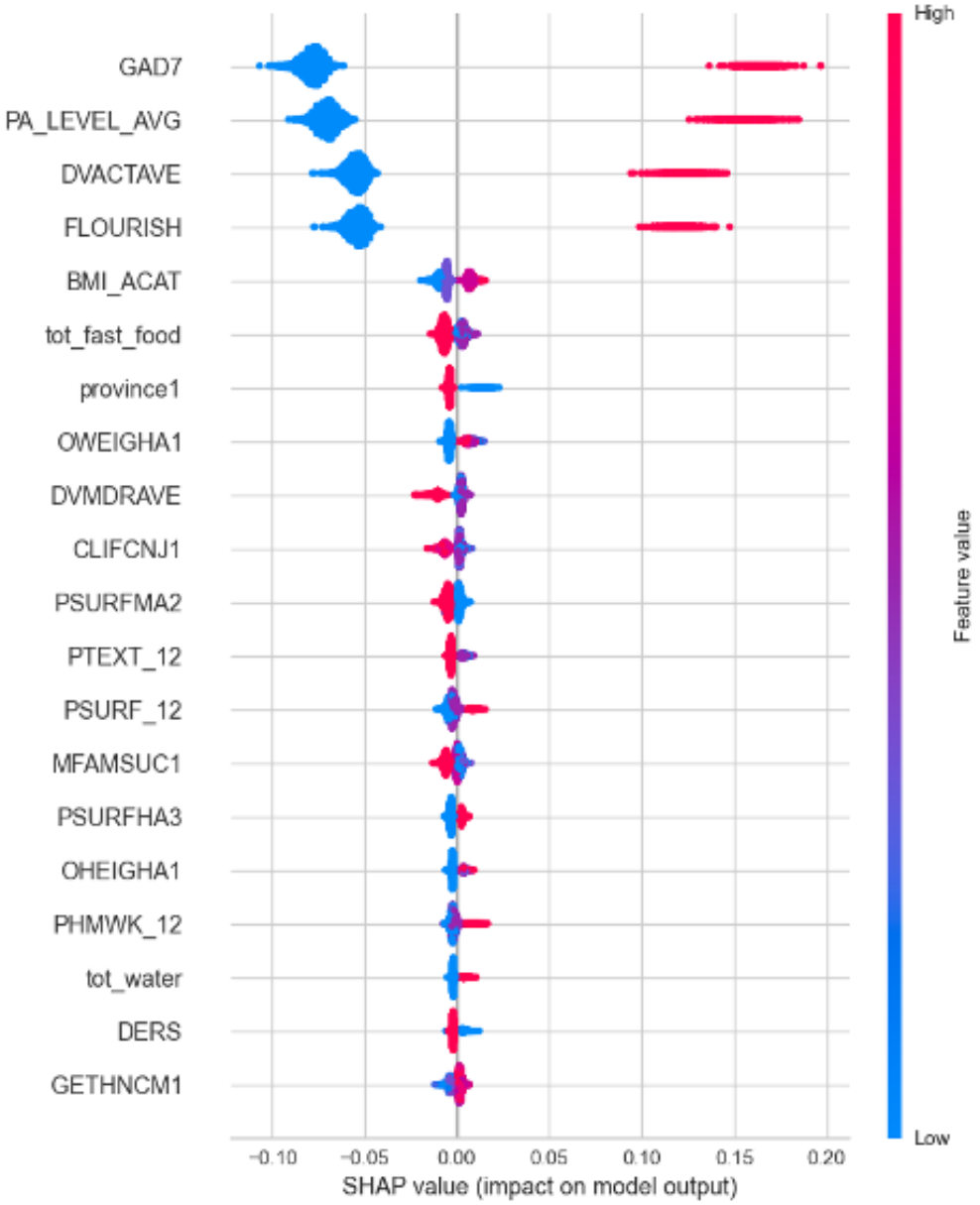


Figure S1.6 (A-C) 2020-2021: SHAP summary plot for sub clusters in high-level cluster 1

A. SHAP summary plot for Sub Cluster 0



B. SHAP summary plot for Sub Cluster 1



C. SHAP summary plot for Sub Cluster 2

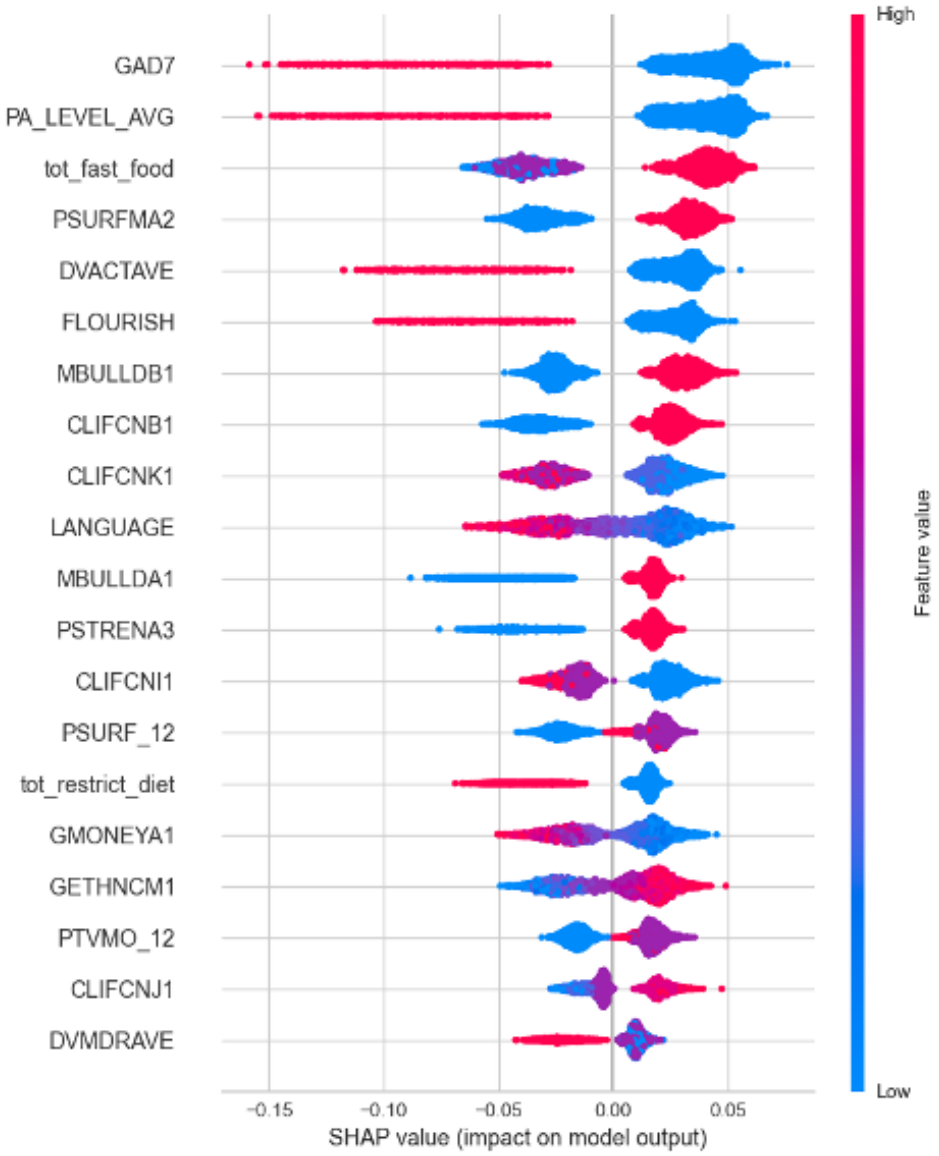
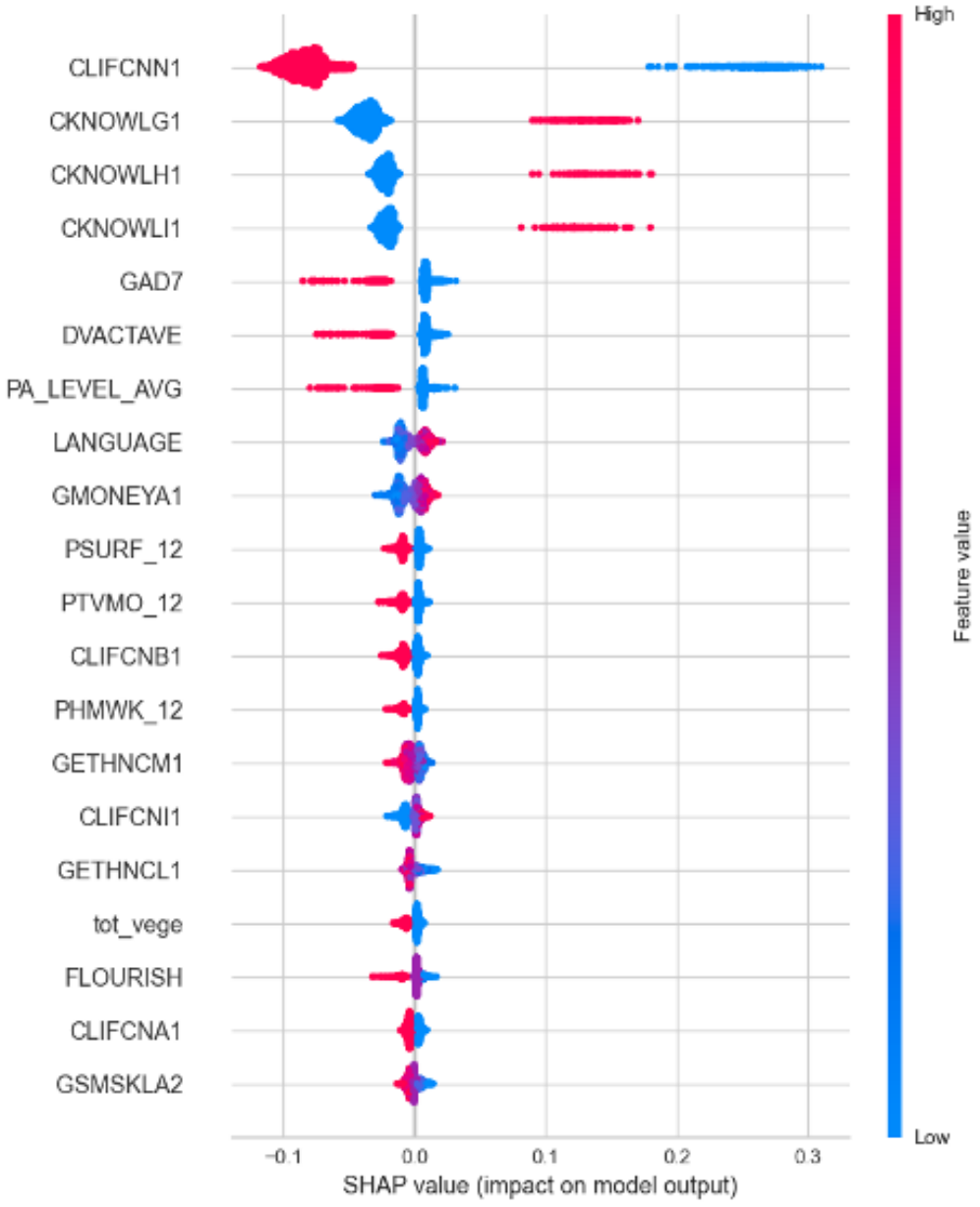
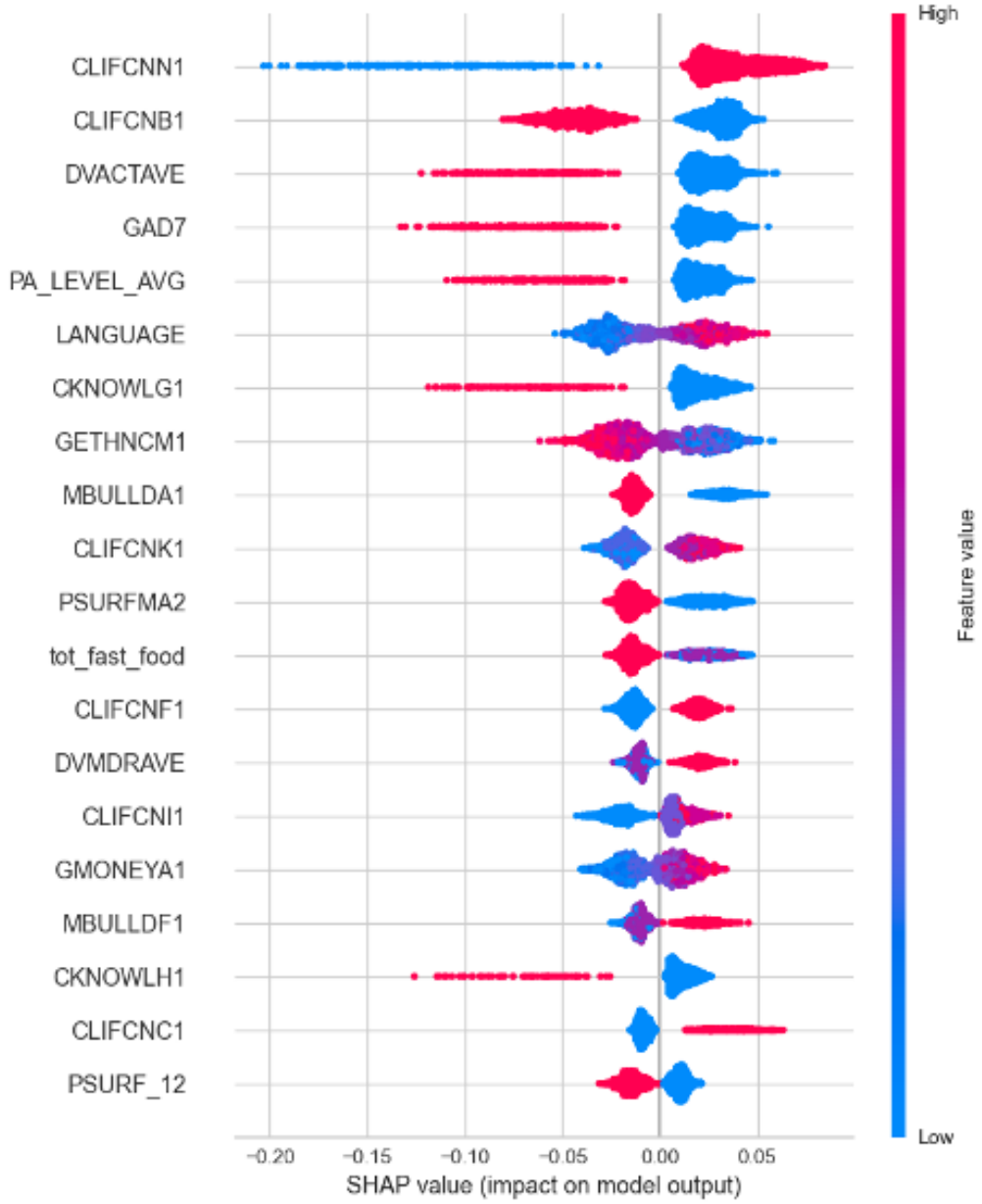


Figure S1.7 (A-D) 2020-2021: SHAP summary plot for sub clusters in high-level cluster 3

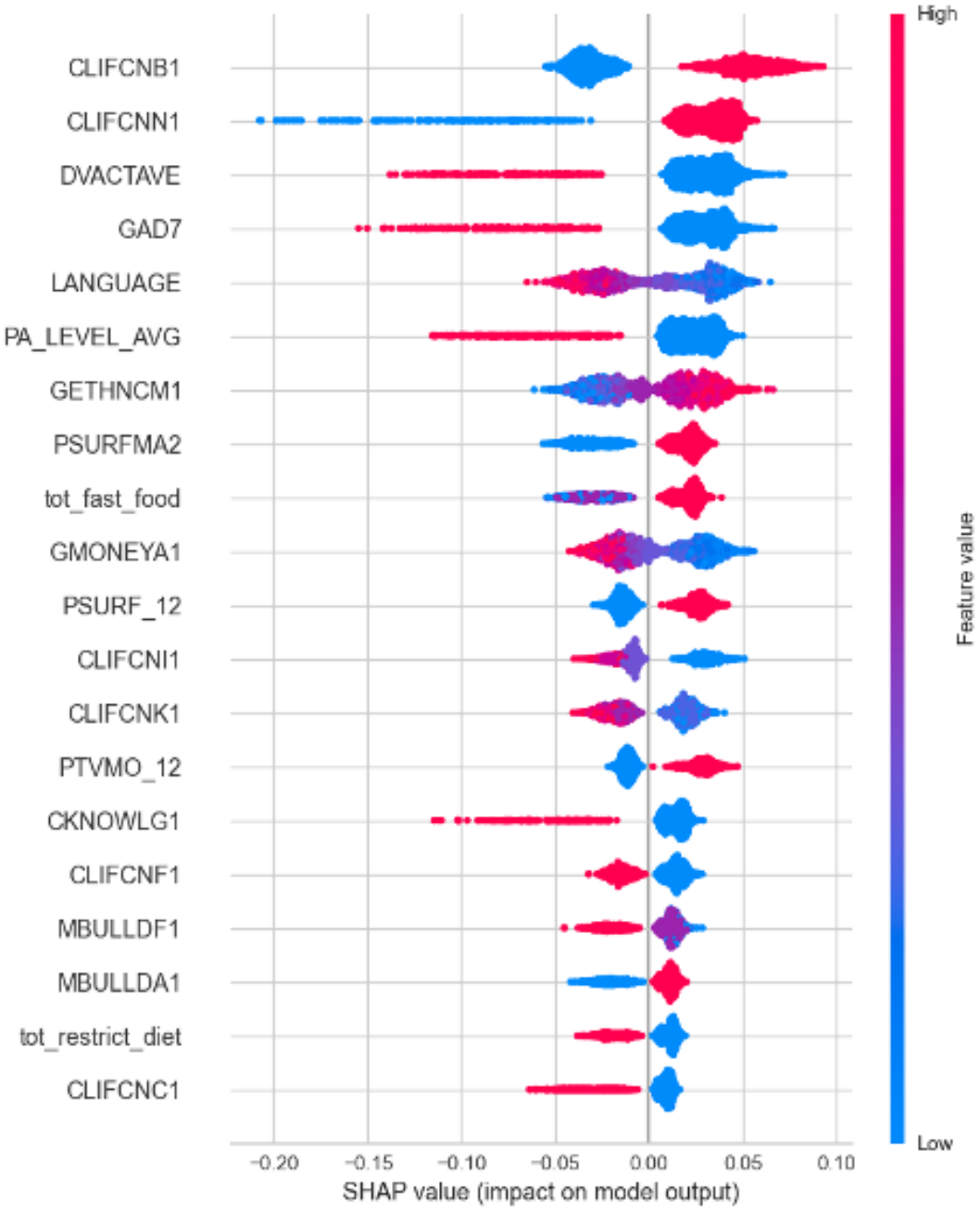
A. SHAP summary plot for Sub Cluster 0



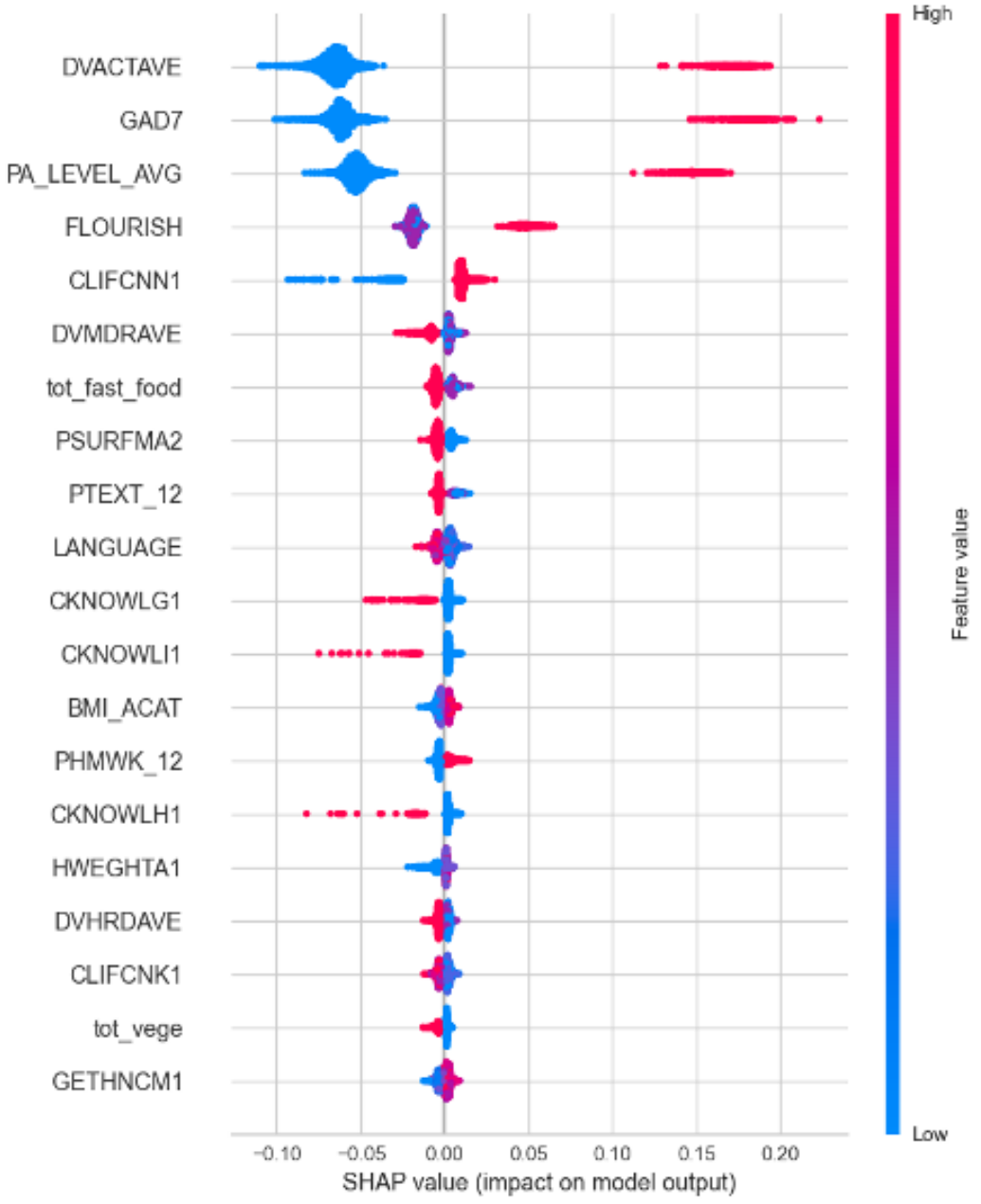
B. SHAP summary plot for Sub Cluster 1



C. SHAP summary plot for Sub Cluster 2



D. SHAP summary plot for Sub Cluster 3



Appendix C: Study 2 Supplementary Materials

Table S2.1 All Variables

Variable Names	Question	Category	Data Type	Percentage of Missing
CESD	10-item Center for Epidemiologic Studies Depression Revised	Mental Health	continuous	15.7
SEVTRYB5	In the last 30 days, did you use any of the following? (Mark all that apply) Cigarillos or little cigars (plain or flavoured)	Smoking & Tobacco Use	category	14.61
SEVTRYB6	In the last 30 days, did you use any of the following? (Mark all that apply) Cigars (not including cigarillos or little cigars, plain or flavoured)	Smoking & Tobacco Use	category	14.61
SEVTRYO2	In the last 30 days, did you use any of the following? (Mark all that apply) Loose tobacco mixed with marijuana	Smoking & Tobacco Use	category	14.61
SEVTRYM1	In the last 30 days, did you use any of the following? (Mark all that apply) Heated tobacco product (a device that heats tobacco instead of burning it, such as IQOS or Heatstick)	Smoking & Tobacco Use	category	14.61
SEVTRYL2	In the last 30 days, did you use any of the following? (Mark all that apply) Smokeless tobacco (chewing tobacco, pinch, snuff, or snus)	Smoking & Tobacco Use	category	14.61
SEVTRYG4	In the last 30 days, did you use any of the following? (Mark all that apply) Nicotine patches, nicotine gum,	Smoking & Tobacco Use	category	14.61

	nicotine lozenges, or nicotine inhalers			
SEVTRYJ2	In the last 30 days, did you use any of the following? (Mark all that apply) I have not used any of these things in the last 30 days	Smoking & Tobacco Use	category	14.61
CKNOWLF1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) Hand washing for at least 20 seconds (with soap and water) helps prevent the transmission of COVID-19	COVID	category	14.51
CKNOWLG1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) The use of hand sanitizer (e.g., Purell) helps prevent the transmission of COVID-19.	COVID	category	14.51
CKNOWLH1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) When a person coughs, the use of a mask can reduce the droplet transmission of COVID-19.	COVID	category	14.51
CKNOWLI1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) COVID-19 is only found in individuals who show symptoms and signs of disease.	COVID	category	14.51

CKNOWLJ1	Please indicate which of the following statements about COVID-19 you think are true. (Mark all the apply) COVID-19 is only dangerous for the elderly population.	COVID	category	14.51
MBULLDA1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) I have not been bullied in the last 30 days	Bullying	category	14.12
MBULLDB1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Physical attacks (e.g., getting beaten up, pushed, or kicked)	Bullying	category	14.12
MBULLDC1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Verbal attacks (e.g., getting teased, threatened, or having rumours spread about you)	Bullying	category	14.12
MBULLDD1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Cyber-attacks (e.g., being sent mean text messages or having rumours spread about you on the internet)	Bullying	category	14.12
MBULLDF1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Social bullying (e.g., being purposefully excluded from a group, being humiliated by others with gestures or graffiti)	Bullying	category	14.12

MBULLDE1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) Had someone steal from you or damage your things	Bullying	category	14.12
DVHRDAVE	Average HARD activity per day (Hard activity total / 7)	Physical Activity	continuous	13.6
GAD7	Generalized Anxiety Disorder 7-item Scale	Mental Health	continuous	13.15
DERS	Difficulties in Emotion Regulation Scale	Mental Health	continuous	12.89
CLIFCND1	How has your life changed because of COVID-19? My anxiety has…	COVID	category	12.82
CLIFCNI1	How has your life changed because of COVID-19? My physical activity has…	COVID	category	12.74
CLIFCNC1	How has your life changed because of COVID-19? My loneliness has…	COVID	category	12.65
CLIFCNH1	How has your life changed because of COVID-19? My time spent surfing/posting on social media has…	COVID	category	12.65
CLIFCNO1	How has your life changed because of COVID-19? My consumption of salty snack (e.g., chips) and sweet foods (e.g., candy, ice cream, cake) has…	COVID	category	12.62
CLIFCNJ1	How has your life changed because of COVID-19? My sleep has…	COVID	category	12.61
CLIFCNG1	How has your life changed because of COVID-19? Time spent watching TV/movies or playing video games has…	COVID	category	12.57

CLIFCNF1	How has your life changed because of COVID-19? Time spent communicating with my friends online has…	COVID	category	12.54
CLIFCNE1	How has your life changed because of COVID-19? Time spent with my friends in person has…	COVID	category	12.5
CLIFCNB1	How has your life changed because of COVID-19? My boredom has…	COVID	category	12.47
CLIFCNA1	How has your life changed because of COVID-19? My stress has…	COVID	category	12.3
FLOURISH	MOPTIM - Sum total score of items A-H (5 = Strongly Agree to 1 = Strongly Disagree).	Mental Health	continuous	11.56
DVOCONCT	School Connectedness Score	School connectedness	continuous	11.49
DVACTAVE	Average combined HARD and MODERATE physical activity per day (combined activity total / 7)	Physical Activity	continuous	11.46
PA_LEVEL_AVG	respondents who meet the guidelines for at least 60 minutes of physical activity per day on average	Physical Activity	category	11.46
OMARKSA1	On average, what marks do you usually get in school? (Choose only one)	Academic Achievement	category	11.33
OLIKEDA1	What is the highest level of education you would like to get?	Academic Achievement	category	11.07
SEVRQTA2	Have you ever tried to quit smoking cigarettes?	Smoking & Tobacco Use	category	10.83
AMJTYPA1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I	Cannabis Use	category	10.64

	have used it by smoking it (e.g., in a joint, a pipe, a bong)			
AEVRETB1	How old were you when you first had a drink of alcohol that was more than just a sip?	Alcohol Use	category	10.39
MOPTIMI1	How much do you agree with the following statements? I generally recover from setbacks quickly	Mental Health	category	10.28
AFNTNLA1	Have you used or tried any of the following medications TO GET HIGH? B) Fentanyl (china white, synthetic heroin, china girl)	Opioid Use	category	10.21
QCHEFFE1	For each of the following behaviours, please indicate the extent to which you believe an individual would be exposed to risks to his or her physical health: S/he has tried cannabis (marijuana, hashish) a few times	Cannabis Use	category	10.12
QCHEFFF1	For each of the following behaviours, please indicate the extent to which you believe an individual would be exposed to risks to his or her physical health: S/he smokes cannabis (marijuana, hashish) regularly	Cannabis Use	category	10.12
QCHEFFD1	For each of the following behaviours, please indicate the extent to which you believe an individual would be exposed to risks to his or her physical health: S/he	Vaping & E-Cigarette Use	category	10.07

	smokes e-cigarettes regularly			
APRSCPA1	Have you used or tried any of the following medications TO GET HIGH? C)Other prescription pain relievers (codeine, morphine, Tylenol 3)	Opioid Use	category	10.02
MFAMSUB1	How much do you agree or disagree with the following statements? My parents/guardians expect too much of me	Mental Health	category	9.94
ARISKBB1	In the last 30 days, how many times have you done the following? Rode in/on a vehicle when the driver (you or someone else) had been using marijuana or cannabis	Cannabis Use	category	9.93
QCHEFFC1	For each of the following behaviours, please indicate the extent to which you believe an individual would be exposed to risks to his or her physical health: S/he has tried e-cigarettes a few times	Vaping & E-Cigarette Use	category	9.87
tot_fruit	Total number of days of Eat fruit (fresh, canned, dried, or frozen) in the past week	Healthy Eating	continuous	9.83
tot_water	Total number of days of drink water in the past week	Healthy Eating	continuous	9.83
tot_snacks	Total number of days of eat snacks purchased from a vending machine, corner store, snack bar, or canteen in the past week	Healthy Eating	continuous	9.83
tot_restrict_diet	Total number of days of Restrict/alter your food intake with the intention of changing your	Healthy Eating	continuous	9.83

	weight/shape in the past week			
tot_vege	Total number of days of Eat vegetables (raw or cooked, fresh, canned, or frozen) in the past week	Healthy Eating	continuous	9.83
tot_fast_food	Total number of days of Eat food purchased at a fast food place or restaurant in the past week	Healthy Eating	continuous	9.83
AOXYCDA1	Have you used or tried any of the following medications TO GET HIGH? A) Oxycodone (oxy, OC, APO, OxyContin®, percs, roxies, OxyNEO®)	Opioid Use	category	9.78
MFAMSUD1	How much do you agree or disagree with the following statements? I can talk about my problems with my friends	Mental Health	category	9.75
MFAMSUC1	How much do you agree or disagree with the following statements? I can talk about my problems with my family	Mental Health	category	9.74
APRHRDA1	Do you think it would be difficult or easy to get pain relievers (Oxycodone, Fentanyl, codeine, etc.) if you wanted some?	Opioid Use	category	9.72
AHWHRDA2	Do you think it would be difficult or easy for you to get marijuana if you wanted some?	Cannabis Use	category	9.56
MFAMSUA1	How much do you agree or disagree with the following statements? I have a happy home life	Mental Health	category	9.43
ARISKBA1	In the last 30 days, how many times have you done the following? Rode in/on a vehicle	Alcohol Use	category	9.42

	when the driver (you or someone else) had been drinking alcohol			
AEVRMJB1	How old were you when you first used marijuana or cannabis?	Cannabis Use	category	9.03
AMJTYPB1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have used it by vaping it	Vaping & E-Cigarette Use	category	8.73
AMJTYPC1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have used it by eating or drinking it (e.g., in brownies, cookies, candies, tea)	Cannabis Use	category	8.73
AMJTYPD1	If you have used marijuana or cannabis in the last 12 months, how did you use it? (Mark all that apply) - I have not used marijuana or cannabis in the last 12 months	Cannabis Use	category	8.73
HSKBRKA1	If you do not eat breakfast every day, why do you skip breakfast? I eat breakfast every day	Healthy Eating	category	8.67
HSKBRKB1	If you do not eat breakfast every day, why do you skip breakfast? I don't have time for breakfast	Healthy Eating	category	8.67
HSKBRKC1	If you do not eat breakfast every day, why do you skip breakfast? The bus comes too early	Healthy Eating	category	8.67
HSKBRKD1	If you do not eat breakfast every day, why do you skip breakfast? I sleep in	Healthy Eating	category	8.67

HSKBRKE1	If you do not eat breakfast every day, why do you skip breakfast? I'm not hungry in the morning	Healthy Eating	category	8.67
HSKBRKF1	If you do not eat breakfast every day, why do you skip breakfast?I feel sick when I eat breakfast	Healthy Eating	category	8.67
HSKBRKG1	If you do not eat breakfast every day, why do you skip breakfast? I'm trying to lose weight	Healthy Eating	category	8.67
HSKBRKH1	If you do not eat breakfast every day, why do you skip breakfast? There is nothing to eat at home	Healthy Eating	category	8.67
A5DRNKC1	In the last 12 months, how often did you have 5 drinks of alcohol or more on one occasion?	Alcohol Use	category	8.55
AOFTMJA2	In the last 12 months, how often did you use marijuana or cannabis? (a joint, pot, weed, hash)	Cannabis Use	category	8.53
ADRINKA1	In the last 12 months, how often did you have a drink of alcohol that was more than just a sip?	Alcohol Use	category	8.41
PSTRENA3	On how many days in the last 7 days did you do exercises to strengthen or tone your muscles? (e.g., push-ups, sit-ups, or weight-training)	Physical Activity	category	8.28
PSPORTC5	Do you participate in before-school, noon hour, or after school physical activities organized by your school? (e.g., intramurals, non-competitive clubs)	Physical Activity	category	8.28

SECGGTA1	Where do you usually get vapes (including pods, e-juice or e-liquid)? (Choose only one)	Vaping & E-Cigarette Use	category	8.20
SUSECGB2	Have you used a vape for any of the following reasons? (Mark all that apply) I have not used a vape	Vaping & E-Cigarette Use	category	8.00
SUSECGC2	Have you used a vape for any of the following reasons? (Mark all that apply) Curiosity / to try something new	Vaping & E-Cigarette Use	category	8.00
SUSECGH2	Have you used a vape for any of the following reasons? (Mark all that apply) It's cool / a lot of peole I know use them	Vaping & E-Cigarette Use	category	8.00
SUSECGD2	Have you used a vape for any of the following reasons? (Mark all that apply) I can vape in places where smoking is not allowed	Vaping & E-Cigarette Use	category	8.00
SUSECGF2	Have you used a vape for any of the following reasons? (Mark all that apply) To help me quit smoking cigarettes	Vaping & E-Cigarette Use	category	8.00
SUSECGI2	Have you used a vape for any of the following reasons? (Mark all that apply) To relax or relieve stress and anxiety	Vaping & E-Cigarette Use	category	8.00
SUSECGJ2	Have you used a vape for any of the following reasons? (Mark all that apply) For the nicotine high	Vaping & E-Cigarette Use	category	8.00
SUSECGK2	Have you used a vape for any of the following reasons? (Mark all that apply) I really like the flavours	Vaping & E-Cigarette Use	category	8.00
SUSECGG2	Have you used a vape for any of the following reasons? (Mark all that	Vaping & E-Cigarette Use	category	8.00

	apply) I have used a vape for some other reason			
SECGSHA1	When you vape, how often do you share the device with someone else or use someone else's vape?	Vaping & E-Cigarette Use	category	7.98
SLP_AVG	Average length of sleep per night in a week	Sleep	continuous	7.94
SLP_GUIDE	Respondents who meet the guidelines for at least 8 hours of sleep per day (on average)	Sleep	category	7.94
SECG30A2	On how many of the last 30 days did you use a vape?	Vaping & E-Cigarette Use	category	7.92
PSPORTD5	Do you participate in competitive school sports teams that compete against other schools? (e.g., junior varsity or varsity sports)	Physical Activity	category	7.88
SUSECGA3	Have you ever tried a vape, also known as an e-cigarette? (e.g., JUUL, Vype, Suorin, Smok)	Vaping & E-Cigarette Use	category	7.87
DSUSCEPT3	Smoking Susceptibility - Three Measures	Smoking & Tobacco Use	category	7.86
PSPORTD4	Do you participate in league or team sports outside of school?	Physical Activity	category	7.75
SLST30A1	On how many of the last 30 days did you smoke one or more cigarettes?	Smoking & Tobacco Use	category	7.69
SPUFF0A1	Have you ever tried cigarette smoking, even just a few puffs?	Smoking & Tobacco Use	category	7.64
PSLPEZA1	During the past week, what time have you usually turned out the light and gone to sleep… on weekends? (Select the time using the drop-down lists for hour, minutes and AM/PM) - AM/PM	Sleep	category	7.02

PSLPDZA1	During the past week, what time have you usually turned out the light and gone to sleep… on weekdays? (Select the time using the drop-down lists for hour, minutes and AM/PM) - AM/PM	Sleep	category	6.82
PWAKEZA1	During the past week, what time have you usually woken up in the morning… on weekends? (Select the time using the drop-down lists for hour, minutes and AM/PM) - AM/PM	Sleep	category	6.79
PWAKDZA1	During the past week, what time have you usually woken up in the morning… on weekdays? (Select the time using the drop-down lists for hour, minutes and AM/PM) - AM/PM	Sleep	category	6.64
PTVMO_12	Watching TV : Minutes per day	Screen Time	continuous	5.76
PVIDE_12	Video Games : Minutes per day	Screen Time	continuous	5.76
PHMWK_12	Doing Homework: Minutes per day	Screen Time	continuous	5.76
PSURF_12	Surfing Internet : Minutes per day	Screen Time	continuous	5.76
PTEXT_12	Texting, messaging : Minutes per day	Screen Time	continuous	5.76
DTDSEDNA_12	Total daily sedentary	Screen Time	continuous	5.76
DTDSCRNT_12	Total daily sedentary activity, with homework excluded	Screen Time	continuous	5.76
PSLPQTA1	During the past week, how would you rate your sleep quality overall (how well you sleep)?	Sleep	category	5.14
HDSCWTA1	How do you describe your weight?	Weight-related	category	3.56
GMONEYB1	Where do you get money to spend on	Demographics	category	3.38

	yourself or to save? I do not usually get any money to spend on myself or to save			
GMONEYC1	Where do you get money to spend on yourself or to save? My parents/guardians give me money (e.g., an allowance)	Demographics	category	3.38
GMONEYD1	Where do you get money to spend on yourself or to save? I get a pay cheque from a job (working evenings or weekends at a restaurant, store, etc.)	Demographics	category	3.38
GMONEYE1	Where do you get money to spend on yourself or to save? I get paid cash for occasional work (babysitting, mowing lawns, shovelling snow, etc.)	Demographics	category	3.38
HWEGHTA1	Which of the following are you trying to do about your weight?	Weight-related	category	3.15
QCHCOMF1	Would you say that you and your family are more or less comfortable financially than the average student in your class?	Demographics	category	2.73
QCOBEDA1	In your house, do you have your own bedroom?	Demographics	category	2.35
GATHOMA1	Including you, how many people are currently living at your home?	Demographics	category	2.33
GMONEYA1	About how much money do you usually get each week to spend on yourself or to save? (Remember to include all money from allowances and jobs like baby-sitting, delivering papers, etc.)	Demographics	category	2.32

QCHNGRA1	Do you sometimes go to bed hungry because there is not enough money to buy food?	Demographics	category	2.27
CCURRIA1	What form of curriculum delivery are you currently participating in?	Demographics	category	1.71
GETHNCN1	How would you describe yourself? Asian	Demographics	category	1.65
GETHNCC1	How would you describe yourself? Black	Demographics	category	1.65
GETHNCH1	How would you describe yourself? Latin American/Hispanic	Demographics	category	1.65
GETHNCL1	How would you describe yourself? White	Demographics	category	1.65
GETHNCM1	How would you describe yourself? Other	Demographics	category	1.65
SEX4	Are you female or male?	Demographics	category	1.47
AGE	How old are you today?	Demographics	category	1.37
GSMSKLA2	Did you attend this school last year?	Demographics	category	0.95
LANGUAGE	Questionnaire language administered.	Demographics	category	0
ARISKBC1	In the last 30 days, how many times have you done the following? Drank alcohol when you were all by yourself	Alcohol Use	category	0
ARISKBD1	In the last 30 days, how many times have you done the following? Used marijuana or cannabis when you were all by yourself	Cannabis Use	category	0
CLIFCNK1	How has your life changed because of COVID-19? My cigarette use has…	COVID	category	0
CLIFCNL1	How has your life changed because of	COVID	category	0

	COVID-19? My vaping has…			
CLIFCNM1	How has your life changed because of COVID-19? My alcohol use has…	COVID	category	0
CLIFCNN1	How has your life changed because of COVID-19? My cannabis use has…	COVID	category	0
BMI_ACAT	BMI categories	Weight-related	category	0
DSUSCEPT2	Smoking Susceptibility - Two Measures	Smoking & Tobacco Use	category	0
DVMDRAVE	Average MODERATE activity per day (Mod_activity_total / 7)	Physical Activity	continuous	0
Province	Province in which the school resides	Demographics	category	0

Table S2.2 Hyperparameters of Machine Learning Models for BMI Prediction

Table 1.	**Models**	**Best Parameter (CCA)**	**Best F1 score (CCA)**	**Best Parameter (MI)**	**Best F1 score (MI)**
1. LDA Transformed	Support Vector Machine	'C': 0.6378015057531383, 'gamma': 0.01, 'kernel': 'rbf'	0.61	{'C': 22.45692481806743, 'gamma': 0.01, 'kernel': 'rbf'}	0.62
2. Top 50	Random Forest	'n_estimators': 652, 'max_depth': 48, 'min_samples_split': 8, 'min_samples_leaf': 1	0.74	{'n_estimators': 960, 'max_depth': 35, 'min_samples_split': 2, 'min_samples_leaf': 1}	0.76
3. ALL Variables	XGBoost	'lambda': 0.0039337708802886446, 'alpha': 0.0010826205343902869, 'gamma': 0.08132518865164108, 'colsample_bytree': 0.9132888202092534, 'subsample': 0.6913916110303424,	0.75	'lambda': 0.0026786850814010097, 'alpha': 0.2102566521481668, 'gamma': 0.003665865436795861, 'colsample_bytree': 0.5977345746480937, 'subsample': 0.8209658694556307,	0.77

		'learning_rate': 0.01844114492675588, 'n_estimators': 626, 'max_depth': 10, 'min_child_weight': 2, 'minority_multiplier': 1.0861510077250145		'learning_rate': 0.06319743442394166, 'n_estimators': 977, 'max_depth': 10, 'min_child_weight': 5, 'minority_multiplier': 1.369443617058948	

Figure S2.1 Top 50 feature important variables selected by Random Forest classifier

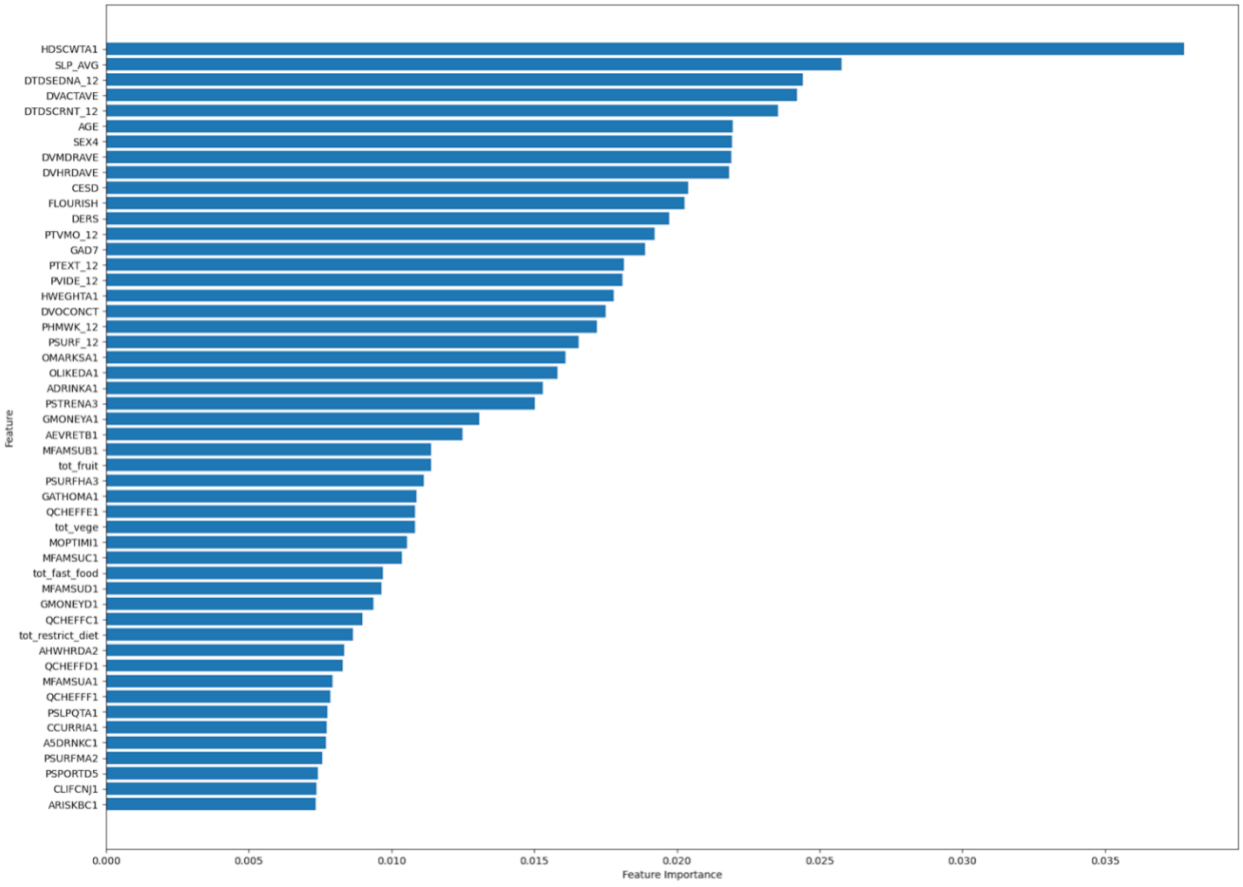
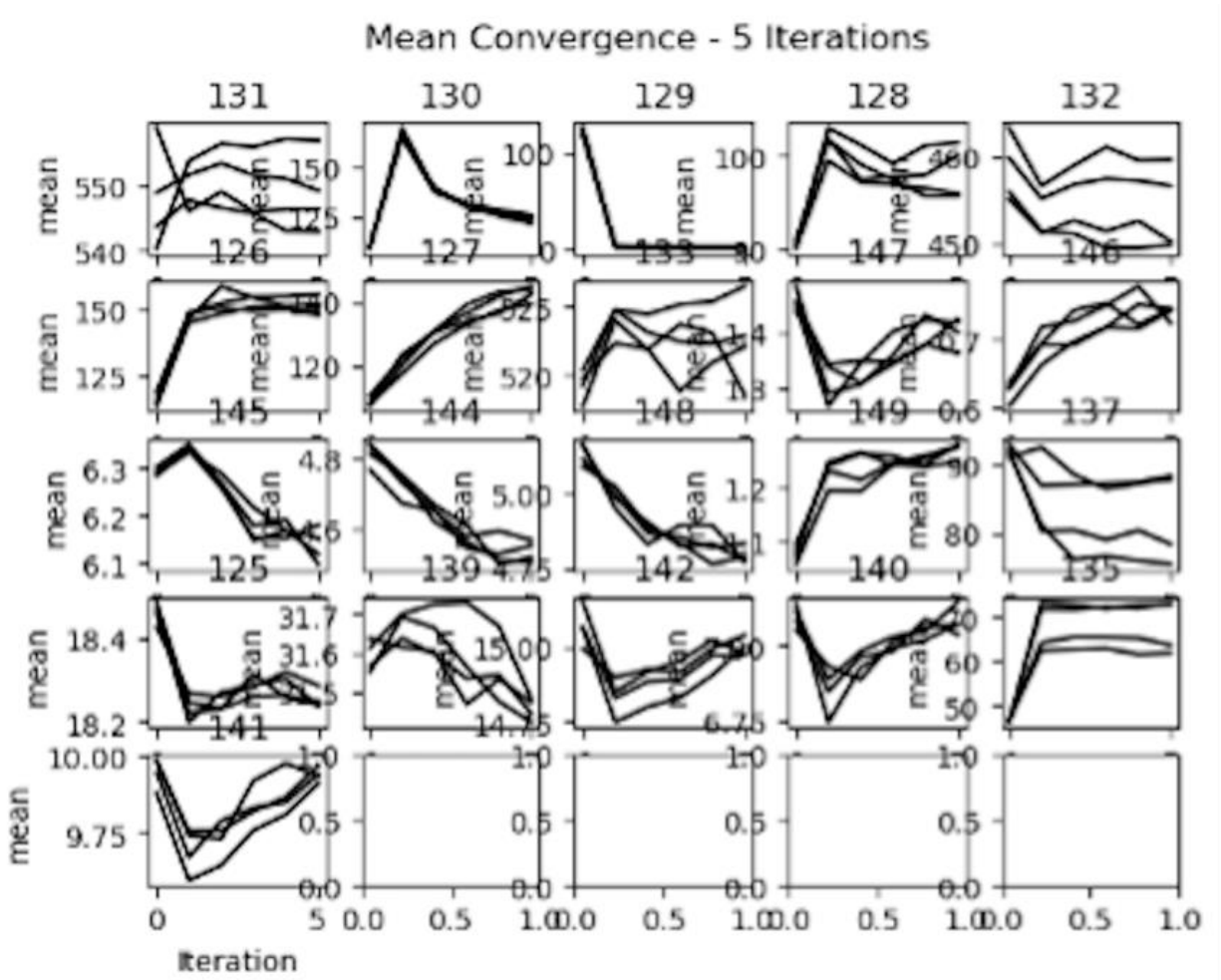
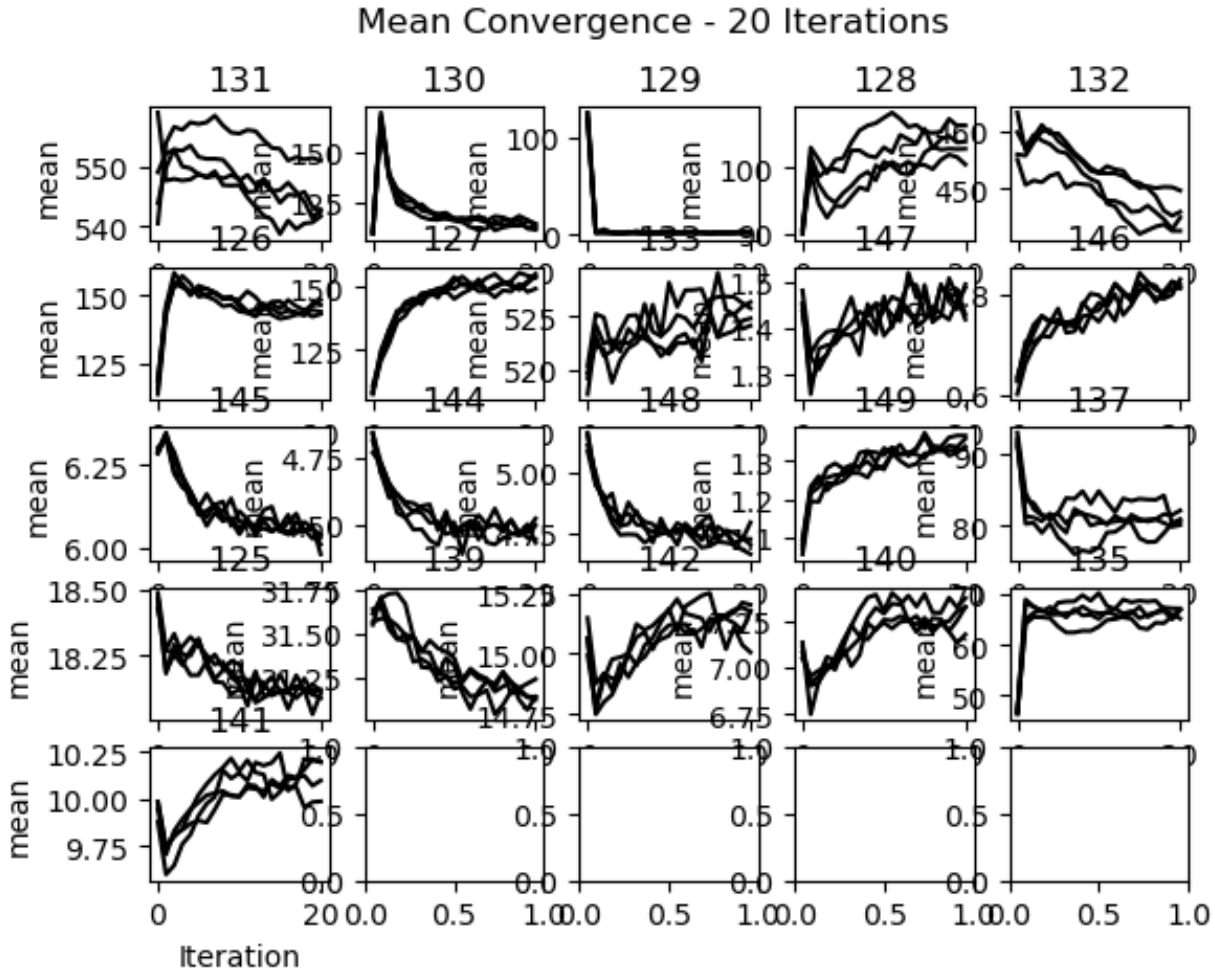


Figure S2.2 Mean Convergence plot for MICE imputed iterations

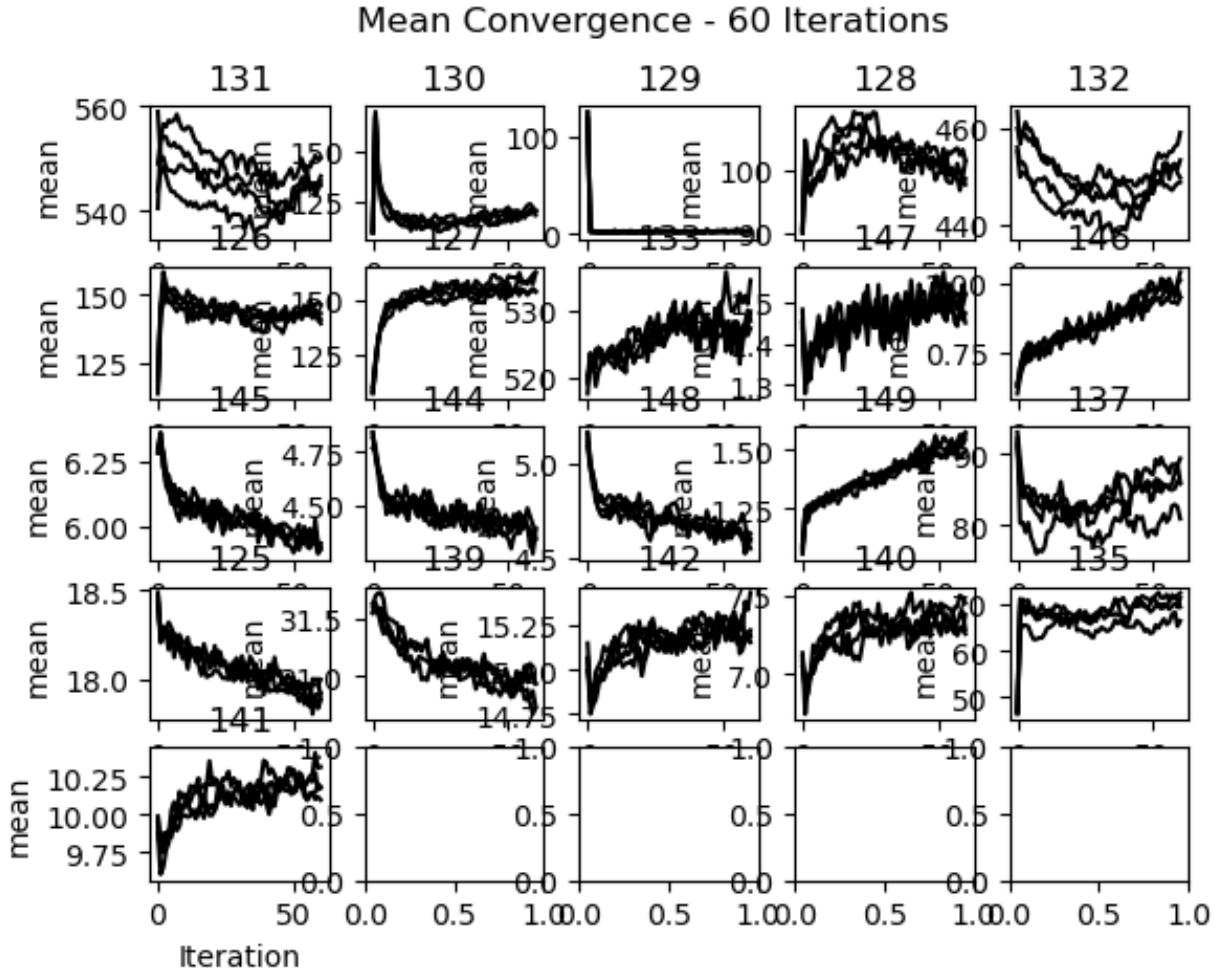
a. Mean convergence of 5 iterations



b. Mean convergence of 20 iterations



c. Mean convergence of 60 iterations



Appendix D: Study 3 Supplementary Materials

Table S3.1 F1 and ROC AUC Score for Determining the Number of Features in XGBoost Classifier

Year 1	F1 score	ROC AUC
3 features	70.0	79.0
4 features	69.4	78.7
40 features	**69.3**	**78.4**

Year 2	F1 score	ROC AUC
2 features	71.0	81.1
3 features	70.4	80.1
6 features	70.1	79.2
4 features	69.9	79.3
5 features	69.4	78.8
7 features	67.8	78.5
39 features	**67.5**	**78.3**

Year 3	F1 score	ROC AUC
2 features	63.6	77.6
34 features	**63.3**	**75.0**

Figure S3.1 Year 1 variable's feature importance score

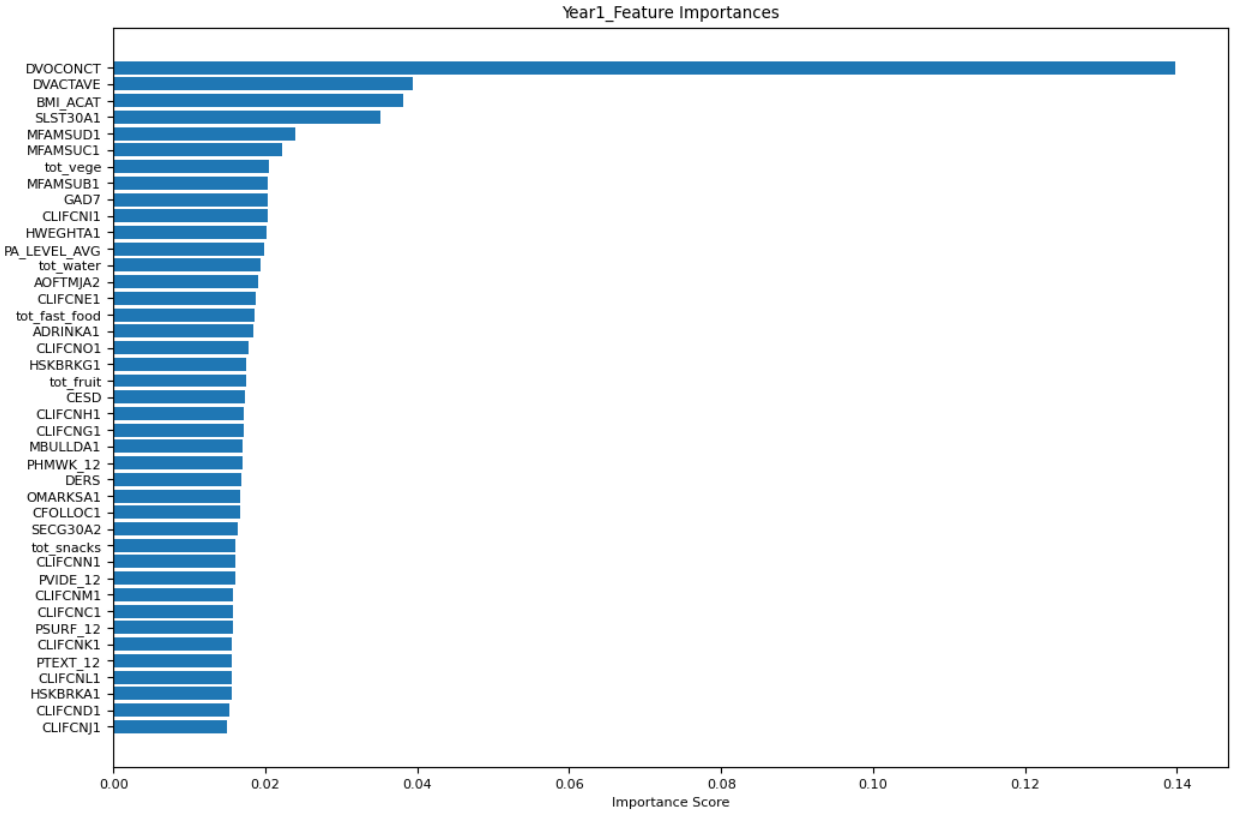


Figure S3.2 Year 2 variable's feature importance score

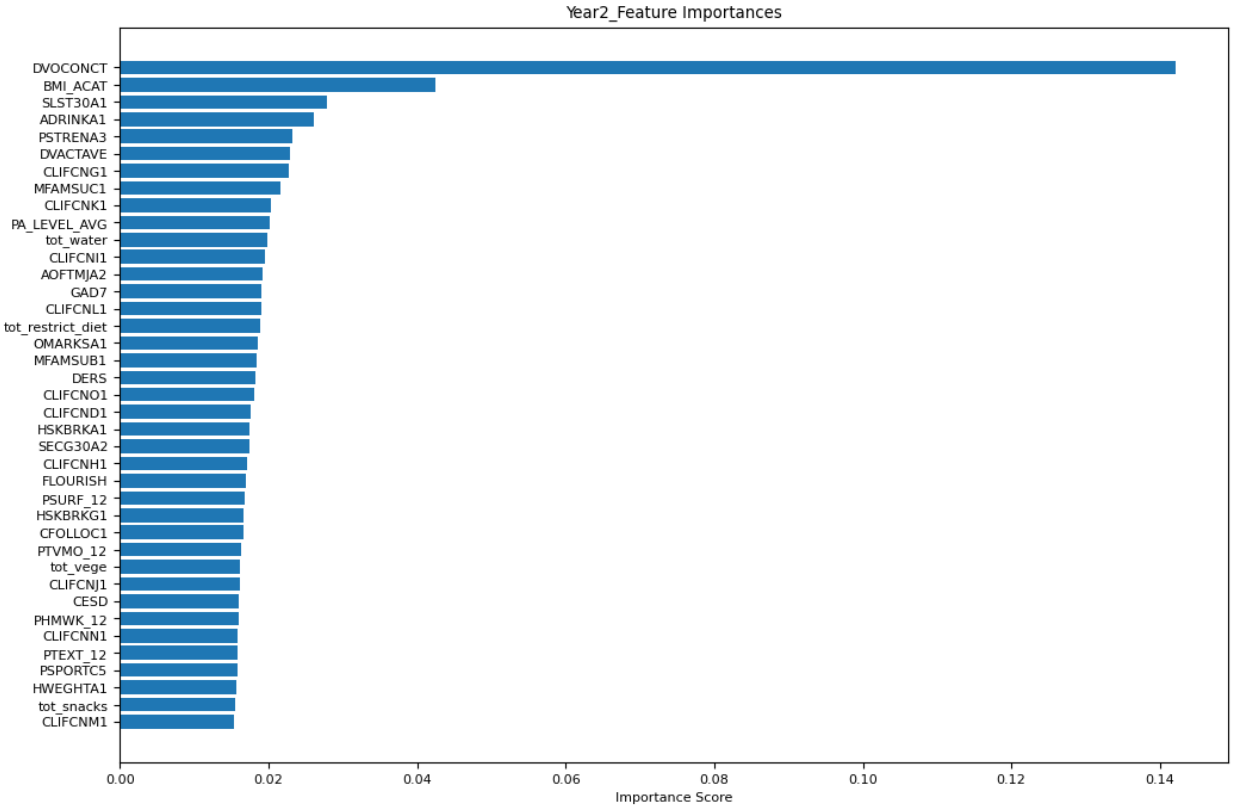


Figure S3.3 Year 3 variable's feature importance score

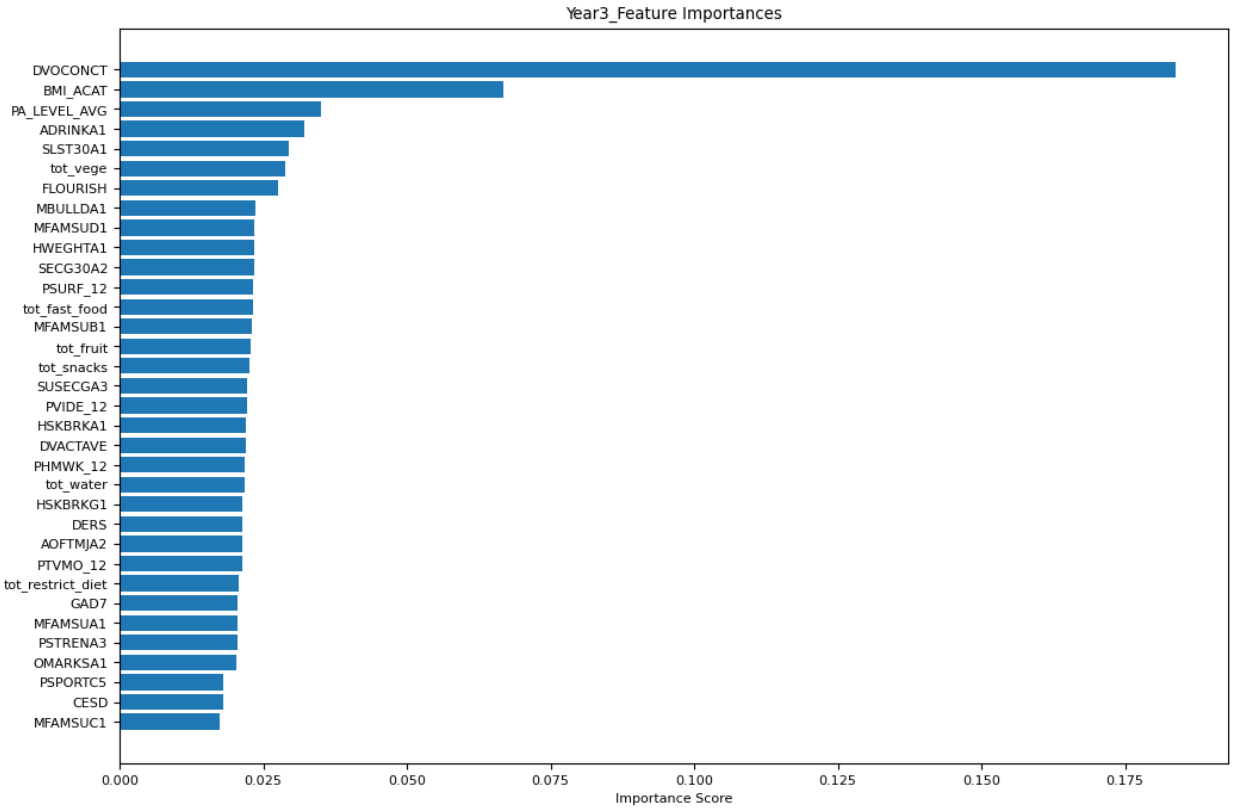


Table S3.2 Final Variables for Cohort Analysis

Variable	Question	Category
PA_LEVEL_AVG	respondents who meet the guidelines for at least 60 minutes of physical activity per day on average	Physical Activity
FLOURISH	The Flourishing Scale	Mental Health
PTVMO_12	Watching TV : Minutes per day	Screen Time
MFAMSUD1	How much do you agree or disagree with the following statements? I can talk about my problems with my friends	Mental Health
tot_restrict_diet	Total number of days of Restrict/alter your food intake with the intention of changing your weight/shape in the past week	Healthy Eating
HSKBRKG1	If you do not eat breakfast every day, why do you skip breakfast? I'm trying to lose weight	Healthy Eating
HSKBRKA1	If you do not eat breakfast every day, why do you skip breakfast? I eat breakfast every day	Healthy Eating
CLIFCNG1	How has your life changed because of COVID-19? Time spent watching TV/movies or playing video games has…	COVID
PSURF_12	Surfing Internet : Minutes per day	Screen Time
CFOLLOC1	How true are the following statements about COVID-19 for you right now? I am worried about my health	COVID
CLIFCNO1	How has your life changed because of COVID-19? My consumption of salty snack (e.g., chips) and sweet foods (e.g., candy, ice cream, cake) has…	COVID
CLIFCND1	How has your life changed because of COVID-19? My anxiety has…	COVID
SUSECGA3	Have you ever tried a vape, also known as an e-cigarette? (e.g., JUUL, Vype, Suorin, Smok)	Vaping & E-Cigarette Use
DERS	MDEPRS - Sum total score of items A-F (1 = Almost never to 5 = Almost always)	Mental Health
SLST30A1	On how many of the last 30 days did you smoke one or more cigarettes?	Smoking & Tobacco Use
CLIFCNI1	How has your life changed because of COVID-19? My physical activity has…	COVID
CLIFCNH1	How has your life changed because of COVID-19? My time spent surfing/posting on social media has…	COVID
PVIDE_12	Video Games : Minutes per day	Screen Time

CLIFCNL1	How has your life changed because of COVID-19? My vaping has…	COVID
MFAMSUA1	How much do you agree or disagree with the following statements? I have a happy home life	Mental Health
OMARKSA1	On average, what marks do you usually get in school? (Choose only one)	Academic Achievement
tot_fruit	Total number of days of Eat fruit (fresh, canned, dried, or frozen) in the past week	Healthy Eating
GAD7	Generalized Anxiety Disorder 7-item Scale	Mental Health
tot_vege	Total number of days of Eat vegetables (raw or cooked, fresh, canned, or frozen) in the past week	Healthy Eating
ADRINKA1	In the last 12 months, how often did you have a drink of alcohol that was more than just a sip?	Alcohol Use
CLIFCNN1	How has your life changed because of COVID-19? My cannabis use has…	COVID
tot_snacks	Total number of days of eat snacks purchased from a vending machine, corner store, snack bar, or canteen in the past week	Healthy Eating
PSTRENA3	On how many days in the last 7 days did you do exercises to strengthen or tone your muscles? (e.g., push-ups, sit-ups, or weight-training)	Physical Activity
BMI_ACAT	BMI catagories	Weight-related
tot_water	Total number of days of drink water in the past week	Healthy Eating
AOFTMJA2	In the last 12 months, how often did you use marijuana or cannabis? (a joint, pot, weed, hash)	Cannabis Use
CLIFCNC1	How has your life changed because of COVID-19? My loneliness has…	COVID
CLIFCNE1	How has your life changed because of COVID-19? Time spent with my friends in person has…	COVID
PTEXT_12	Texting, messaging : Minutes per day	Screen Time
tot_fast_food	Total number of days of Eat food purchased at a fast food place or restaurant in the past week	Healthy Eating
CLIFCNJ1	How has your life changed because of COVID-19? My sleep has…	COVID
CESD	10-item Center for Epidemiologic Studies Depression Revised	Mental Health
SECG30A2	On how many of the last 30 days did you use a vape?	Vaping & E-Cigarette Use
PHMWK_12	Doing Homework: Minutes per day	Screen Time

PSPORTC5	Do you participate in before-school, noon hour, or after school physical activities organized by your school? (e.g., intramurals, non-competitive clubs)	Physical Activity
HWEGHTA1	Which of the following are you trying to do about your weight?	Weight-related
MFAMSUB1	How much do you agree or disagree with the following statements?My parents/guardians expect too much of me	Mental Health
MBULLDA1	In the last 30 days, in what ways were you bullied by other students? (Mark all that apply) I have not been bullied in the last 30 days	Bullying
CLIFCNM1	How has your life changed because of COVID-19? My alcohol use has…	COVID
CLIFCNK1	How has your life changed because of COVID-19? My cigarette use has…	COVID
MFAMSUC1	How much do you agree or disagree with the following statements? I can talk about my problems with my family	Mental Health
DVOCONCT	School Connectedness Score	School connectedness
DVACTAVE	Average combined HARD and MODERATE physical activity per day (combined activity total / 7)	Physical Activity
SLP_AVG	Average length of sleep per night in a week	Sleep
HDSCWTA1_n_year3	How do you describe your weight?	Weight-related

Table S3.3 Percentage of Missing for the Selected Variables in Each Year

Variable	**Year 1**	**Year 2**	**Year 3**
CLIFCNN1	9.70	6.43	0.11
CLIFCNL1	9.36	5.93	0.11
CLIFCNK1	9.25	6.04	0.11
CLIFCNM1	9.12	5.84	0.11
CESD	8.94	6.23	6.82
MBULLDA1	6.84	4.49	5.04
GAD7	6.51	4.99	5.15
DERS	5.77	4.71	4.62
CLIFCND1	5.47	4.41	0.15
CFOLLOC1	5.41	4.01	0.85
CLIFCNI1	5.30	4.34	0.09
CLIFCNO1	5.27	4.17	0.09
CLIFCNH1	5.23	4.23	0.11
CLIFCNC1	5.23	4.21	0.13
CLIFCNG1	5.23	4.13	0.09
CLIFCNJ1	5.23	4.13	0.09
DVACTAVE	5.19	4.13	4.88
PA_LEVEL_AVG	5.19	4.13	4.88
CLIFCNE1	5.12	4.10	0.11
FLOURISH	4.82	3.56	5.41
DVOCONCT	4.62	3.00	3.30
OMARKSA1	4.43	2.91	2.74
tot_fast_food	3.71	2.24	4.23
tot_fruit	3.71	2.24	4.23
tot_restrict_diet	3.71	2.24	4.23
tot_snacks	3.71	2.24	4.23
tot_vege	3.71	2.24	4.23
tot_water	3.71	2.24	4.23
MFAMSUB1	3.54	2.11	2.54
MFAMSUC1	3.28	2.00	2.58
MFAMSUD1	3.24	2.10	2.54
HSKBRKA1	3.23	2.30	2.56
HSKBRKG1	3.23	2.30	2.56

MFAMSUA1	3.02	1.95	2.45
PSPORTC5	2.99	2.24	2.69
AOFTMJA2	2.84	1.61	2.28
PSTRENA3	2.74	2.10	2.60
ADRINKA1	2.74	1.74	1.97
SECG30A2	2.35	1.54	2.19
SUSECGA3	2.35	1.37	1.72
SLST30A1	2.34	1.34	1.72
PHMWK_12	1.59	1.13	1.61
PSURF_12	1.59	1.13	1.61
PTEXT_12	1.59	1.13	1.61
PTVMO_12	1.59	1.13	1.61
PVIDE_12	1.59	1.13	1.61
HWEGHTA1	0.13	0.13	0.11
BMI_ACAT	0.00	0.00	0.00
HDSCWTA1_n	0.00	0.00	0.00
scanid	0.00	0.00	0.00
year_n	0.00	0.00	0.00

Data Transformation

To establish the weight perception transition pattern, nine groups are created based on the year 1 weight perception (HDSCWTA1_y1) and year 2 weight perception (HDSCWTA1_y2), focusing on whether the students maintained their weight status in year 2 compared to year 1 or there was a change in year 2 compared to year 1. The following table represents the data processing for creating this variable.

Year 1 weight perception	Year 2 weight perception	Weight transition group	% of weight transition group
HDSCWTA1_y1 = 0 (underweight)	HDSCWTA1_y2 = 0 (underweight)	1	9.83
HDSCWTA1_y1 = 0 (underweight)	HDSCWTA1_y2 = 1 (about the right weight)	2	5.21
HDSCWTA1_y1 = 0 (underweight)	HDSCWTA1_y2 = 2 (overweight)	3	0.50
HDSCWTA1_y1 = 1 (about the right weight)	HDSCWTA1_y2 = 1 (about the right weight)	4	50.71

HDSCWTA1_y1 = 1 (about the right weight)	HDSCWTA1_y2 = 0 (underweight)	5	6.49
HDSCWTA1_y1 = 1 (about the right weight)	HDSCWTA1_y2=2 (overweight)	6	6.10
HDSCWTA1_y1 = 2 (overweight)	HDSCWTA1_y1 =2 (overweight)	7	15.03
HDSCWTA1_y1 = 2 (overweight)	HDSCWTA1_y2 = 0 (underweight)	8	0.67
HDSCWTA1_y1 = 2 (overweight)	HDSCWTA1_y2 = 1 (about the right weight)	9	5.45

Predictor category transformation

The predictors such as alcohol drinking or substance use below contain multiple categories. One of the categories in the smoking variable did not have any students in weight perception transition groups 3, 5, 6, and 7. Also, the interest of this study does not focus on exploring the effect of the number of days smoking/vaping or the difference between 'strongly agree' and 'agree'. Therefore, the categories for the variables below were collapsed for easier interpretation. After the data transformation, the maximum number of categories was four and only existed in BMI_ACAT (BMI status) and HWEGHTA1 (Intention of weight change). No further collapsing on these two variables since each category has its representation.

Question	Original response	Transformed response
SLST30A1: On how many of the last 30 days did you smoke one or more cigarettes?	1 = "None" 2 = "1 day" 3 = "2 to 3 days" 4 = "4 to 5 days" 5 = "6 to 10 days" 6 = "11 to 20 days" 7 = "21 to 29 days" 8 = "30 days (every day)" 96 = "Valid Skip"	1. = No smoker 'None' and 'Valid Skip' 2. = Smoker "1 day" "2 to 3 days" "4 to 5 days" "6 to 10 days" "11 to 20 days" "21 to 29 days" "30 days (every day)"

Question	Original response	Transformed response
ADRINKA: In the last 12 months, how often did you have a drink of alcohol that was more than just a sip?	1 = "I have never drunk alcohol" 2 = "I did not drink alcohol in the last 12 months" 3 = "I have only had a sip of alcohol" 4 = "Less than once a month" 5 = "Once a month" 6 = "2 or 3 times a month" 7 = "Once a week" 8 = "2 or 3 times a week" 9 = "4 to 6 times a week" 10 = "Every day"	1 = None drinker "I have never drunk alcohol" "I did not drink alcohol in the last 12 months" 2 = Drinker "I have only had a sip of alcohol" "Less than once a month" "Once a month" "2 or 3 times a month" "Once a week" "2 or 3 times a week" "4 to 6 times a week" "Every day"

Question	Original response	Transformed response
OMARKSA1: On average, what marks do you usually get in school?	1 = "90% - 100% (Mostly A+)" 2 = "80% - 89% (Mostly As or A-)" 3 = "70% - 79% (Mostly Bs)" 4 = "60% - 69% (Mostly Cs)" 5 = "50% - 59% (Mostly Ds)" 6 = "Below 50% (Mostly Fs)"	1= “90% - 100% (Mostly A+)" "80% - 89% (Mostly As or A-)" 2 = 70% - 79% (Mostly Bs)" "60% - 69% (Mostly Cs)" 3 = "50% - 59% (Mostly Ds)" "Below 50% (Mostly Fs)"

Question	Original response	Transformed response
MFAMSUD1: How much do you agree or disagree with the following statements? I can talk about my problems with my friends MFAMSUB1: How much do you agree or disagree with the following	1 = "Strongly agree" 2 = "Agree" 3 = "Neither agree nor disagree" 4 = "Disagree" 5 = "Strongly disagree"	1 = "Strongly agree" “Agree" 2 = "Neither agree nor disagree" 3 = "Disagree" "Strongly disagree"

statements? My parents/guardians expect too much of me MFAMSUA1: How much do you agree or disagree with the following statements? I have a happy home life MFAMSUC1: How much do you agree or disagree with the following statements? I can talk about my problems with my family		

Question	Original response	Transformed response
CFOLLOC1: How true are the following statements about COVID-19 for you right now? I am worried about my health	1 = "True" 2 = "Mostly true" 3 = "Neutral / I don't know" 4 = "Mostly false" 5 = "False"	1= "True" "Mostly true" 2 = "Neutral / I don't know" 3 = "Mostly false" "False"

Question	Original response	Transformed response
SECG30A2: On how many of the last 30 days did you use a vape?	1 = "None" 2 = "1 day" 3 = "2 to 3 days" 4 = "4 to 5 days" 5 = "6 to 10 days" 6 = "11 to 20 days" 7 = "21 to 29 days" 8 = "30 days (every day)" 96 = "Valid Skip"	1. = No user 'None' and 'Valid Skip' 2. = Vape user "1 day" "2 to 3 days" "4 to 5 days" "6 to 10 days" "11 to 20 days" "21 to 29 days" "30 days (every day)"

Table S3.4 RNN and LSTM Parameters for Bayesian Optimization

Models	Parameters	Parameter Range
RNN / LSTM	Units	(32, 256)
	Dropout rate	(0.1, 0.5)
	Learning rate	(0.01, 0.1)
	Epochs	(10,100)
	Batch size	(16, 128)
	Optimizers	Adam, RMSprop, Adadelta
	Activation	tanh, relu, sigmoid

MICE Imputation

The highest top three percentages of missing variables in Year 1 were CLIFCNN1 (9.7%), CLIFCNL1 (9.4%), and CLIFCNK1 (9.25%). Year 2, CLIFCNN1 (6.4%), CESD (6.2%), and CLIFCNK1 (6.0%). Year 3, CESD (6.8%), FLOURISH (5.4%), and GAD7 (5.2%). The COVID-19 pandemic variables were not asked in Year 3 and denoted as 97 to be kept in the dataset.

The max iteration in MICE chosen as 50 due to the convergence of the variables was less trend and mingled well in a stable state. The below represents the convergence plot for 20, 30, 40, and 50 iterations for a snapshot of certain variables. The 10 iterations were not displayed because the data failed to converge at 10 iterations. The imputed dataset in each year had the least mean difference compared to the original dataset among the continuous variables. The density plot for the continuous variables proved that the imputed dataset was mostly well-aligned with the original dataset. The original data was shown in blue, and the red colour was the imputed data.

Figure S3.4 Convergence plot of 20 iterations

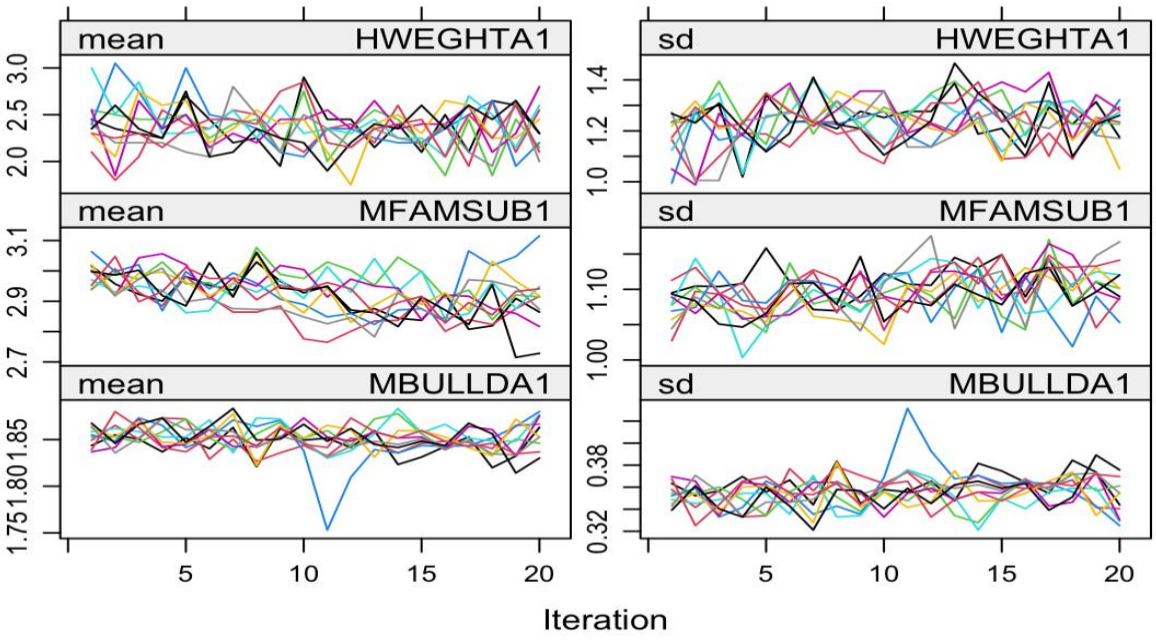


Figure S3.5 Convergence plot of 30 iterations

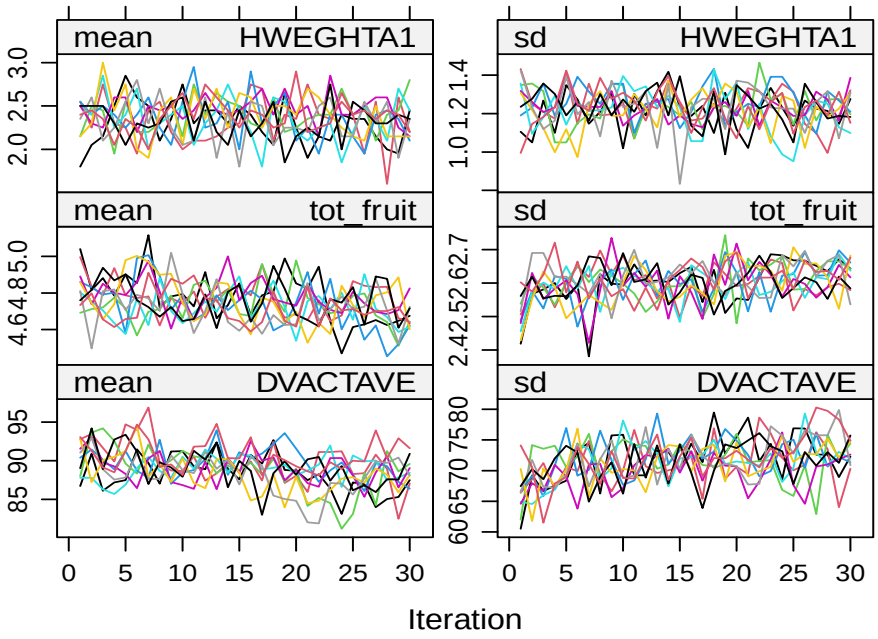


Figure S3.6 Convergence plot of 40 iterations

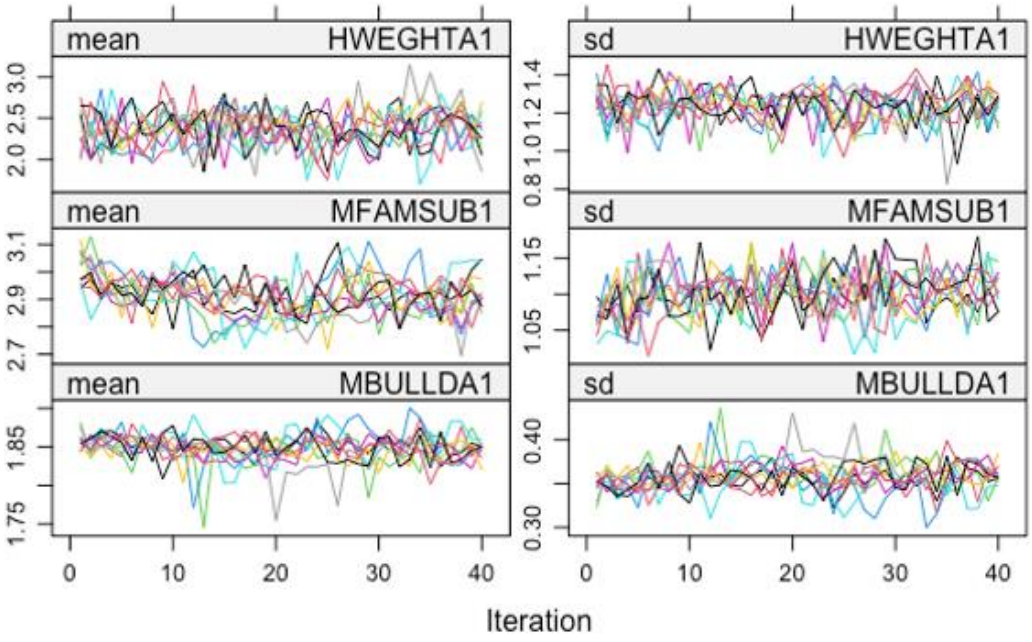


Figure S3.7 Convergence plot of 50 iterations

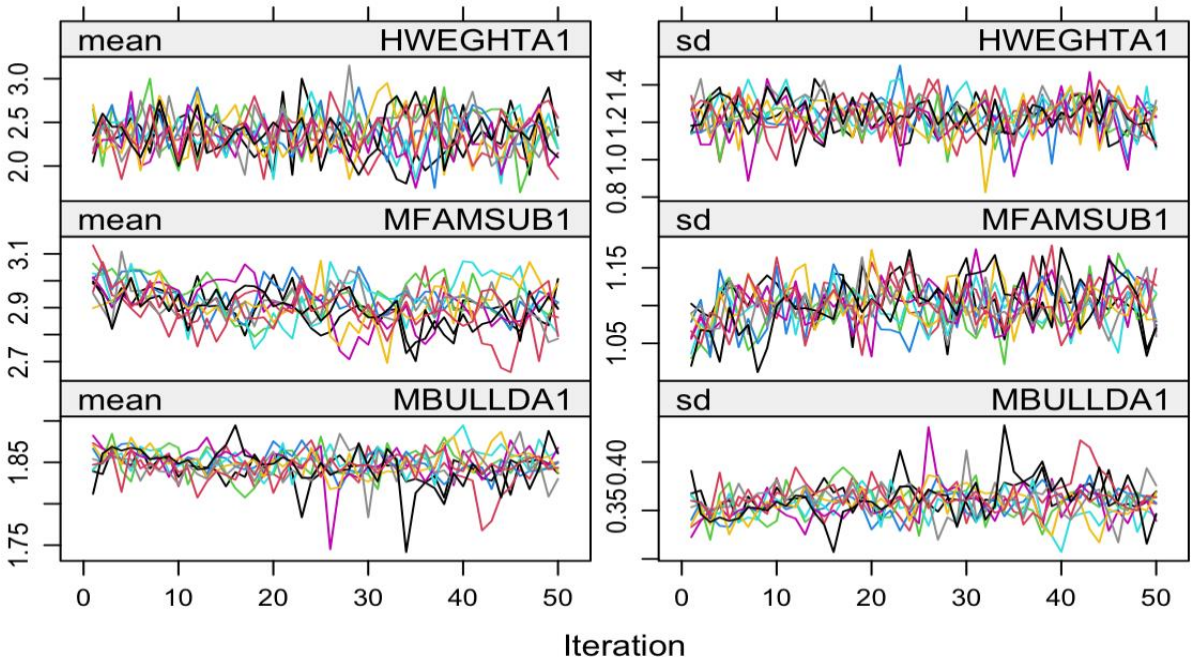


Figure S3.8.1 Density plot of Year 1 on selected variables

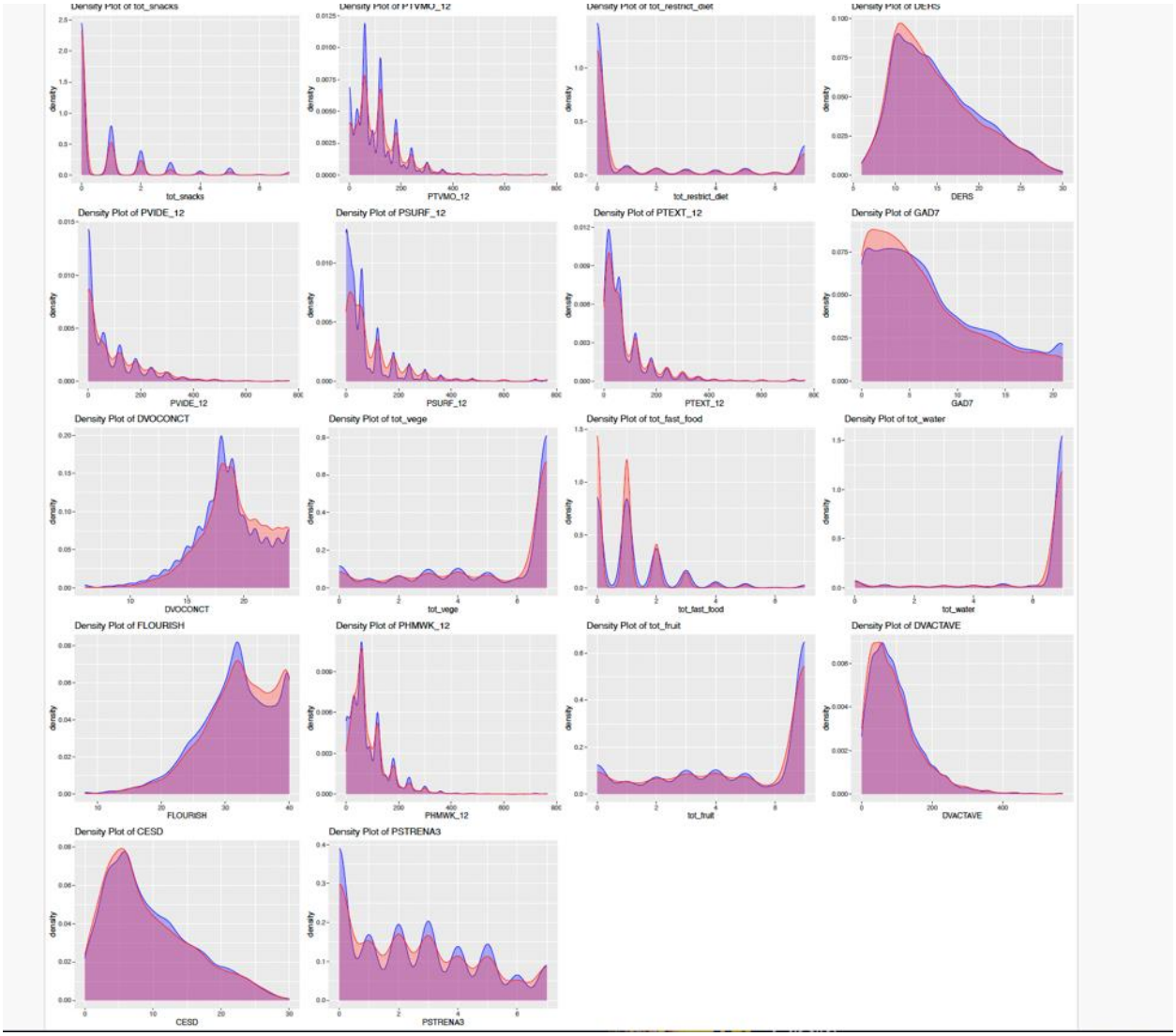


Figure S3.8.2 Density plot of Year 2 on selected variables

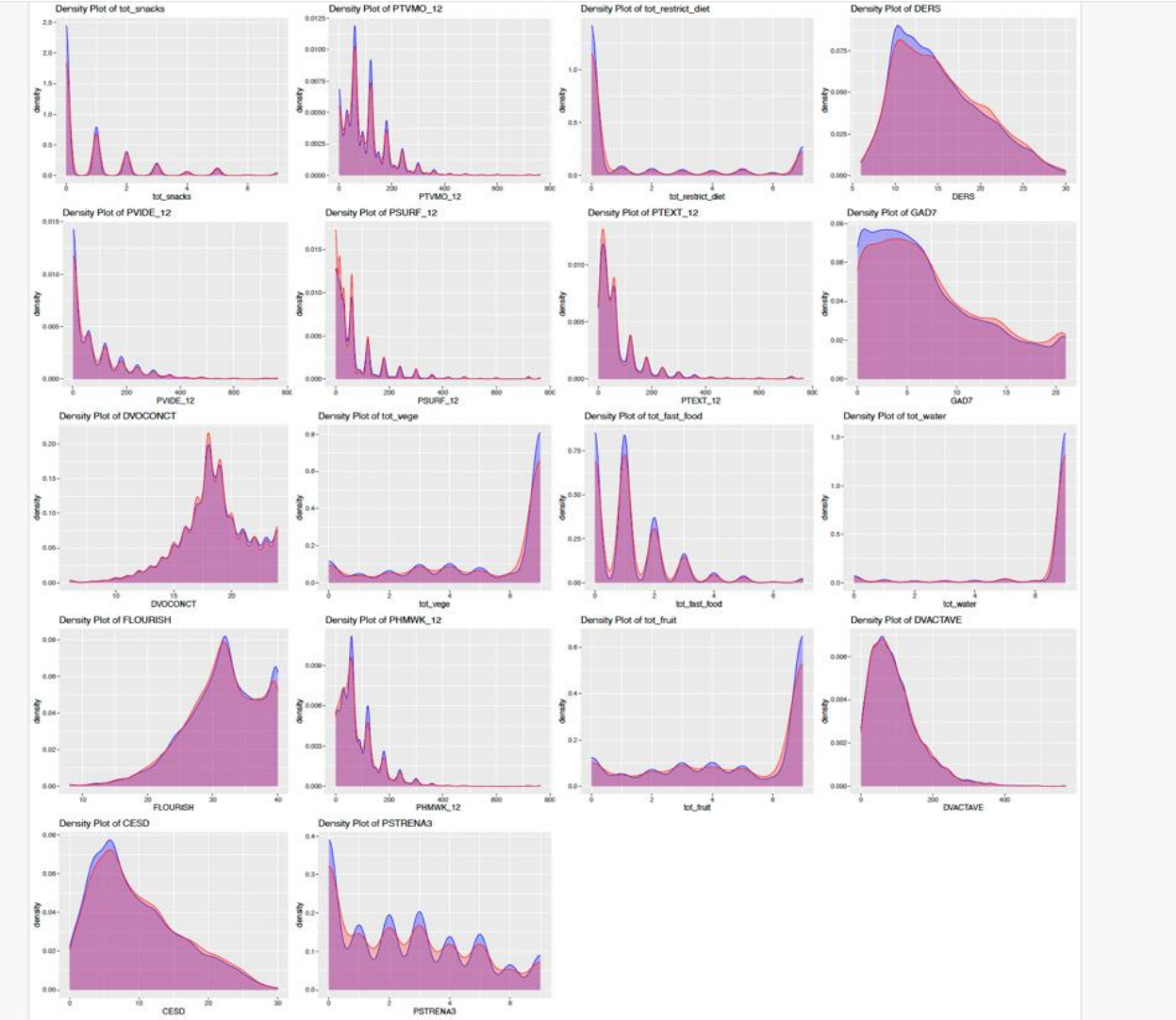


Figure S3.8.3 Density plot of Year 3 on selected variables

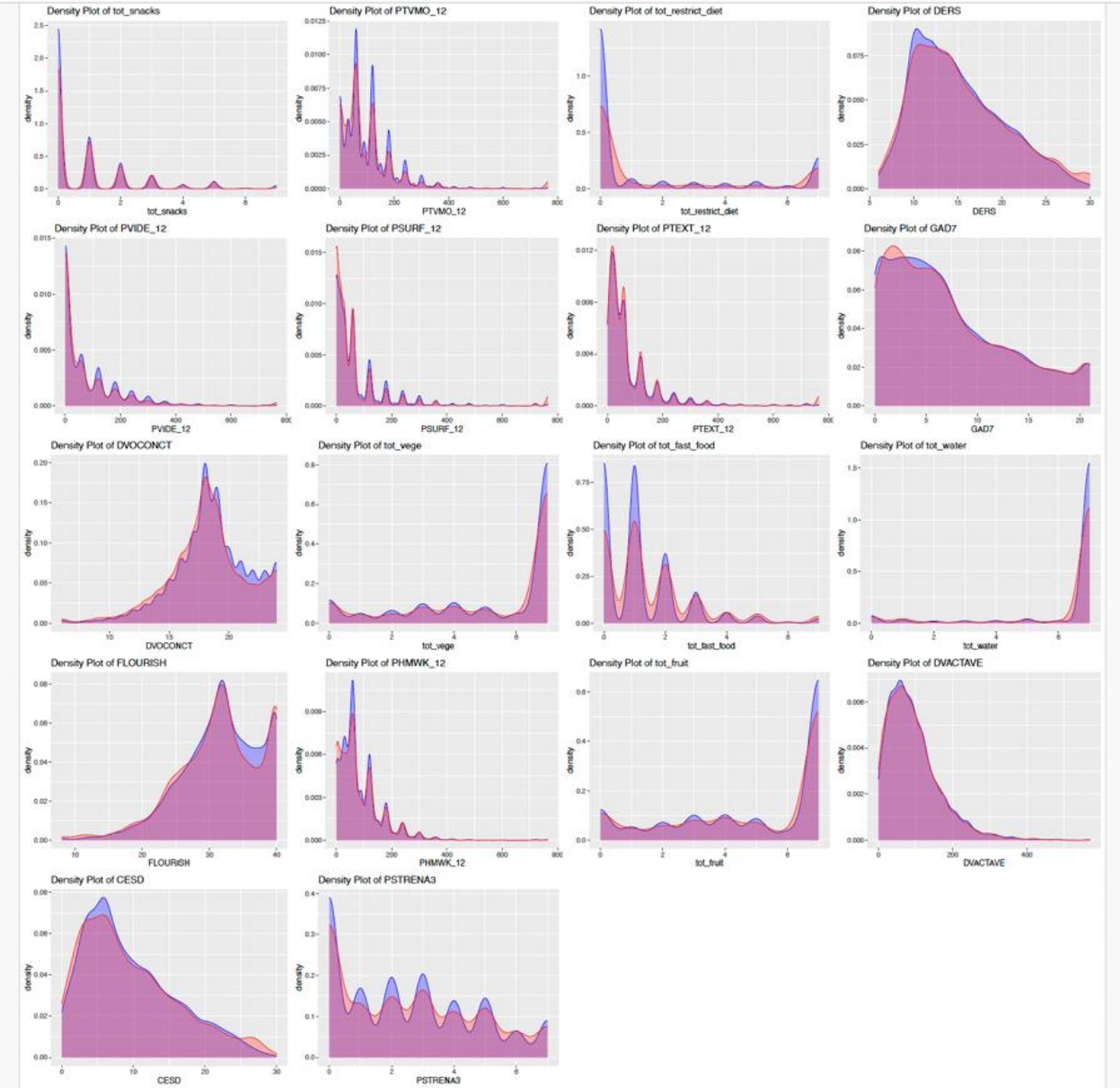
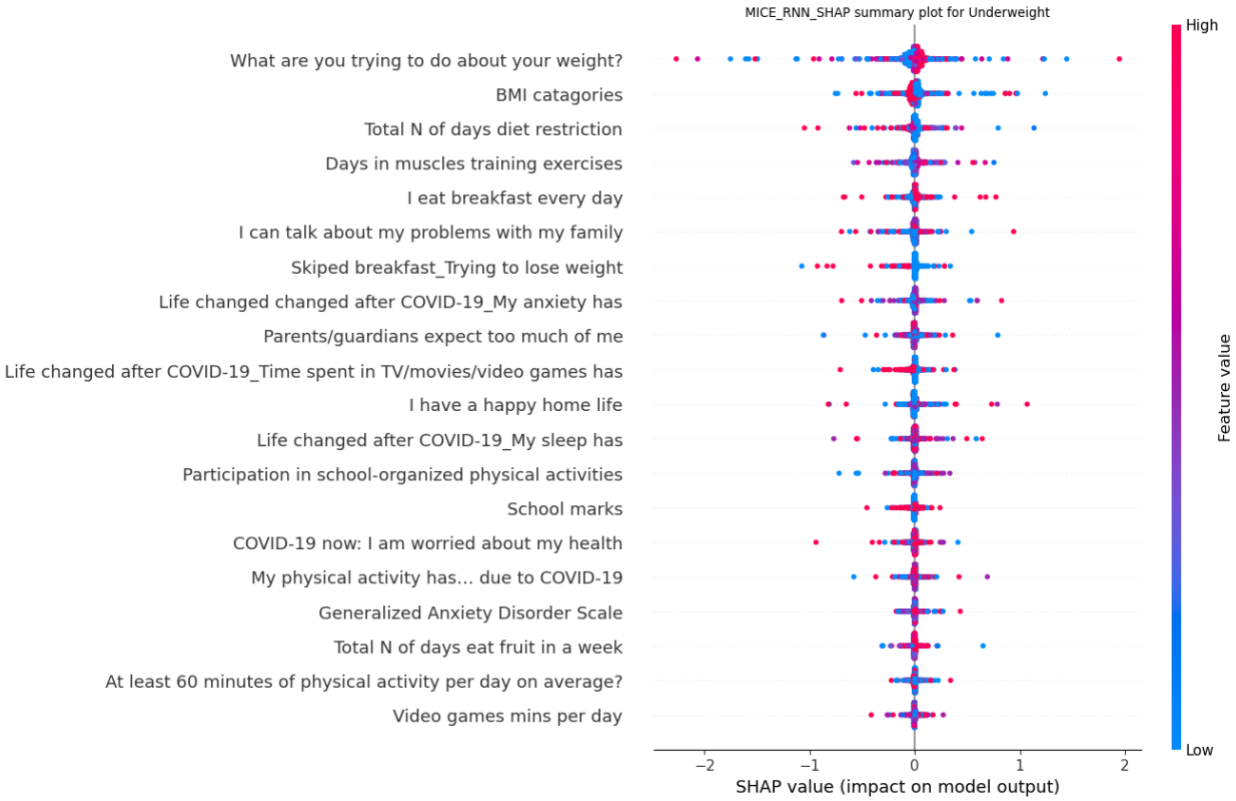


Table S3.5 Optimized Parameters for the Selected Model

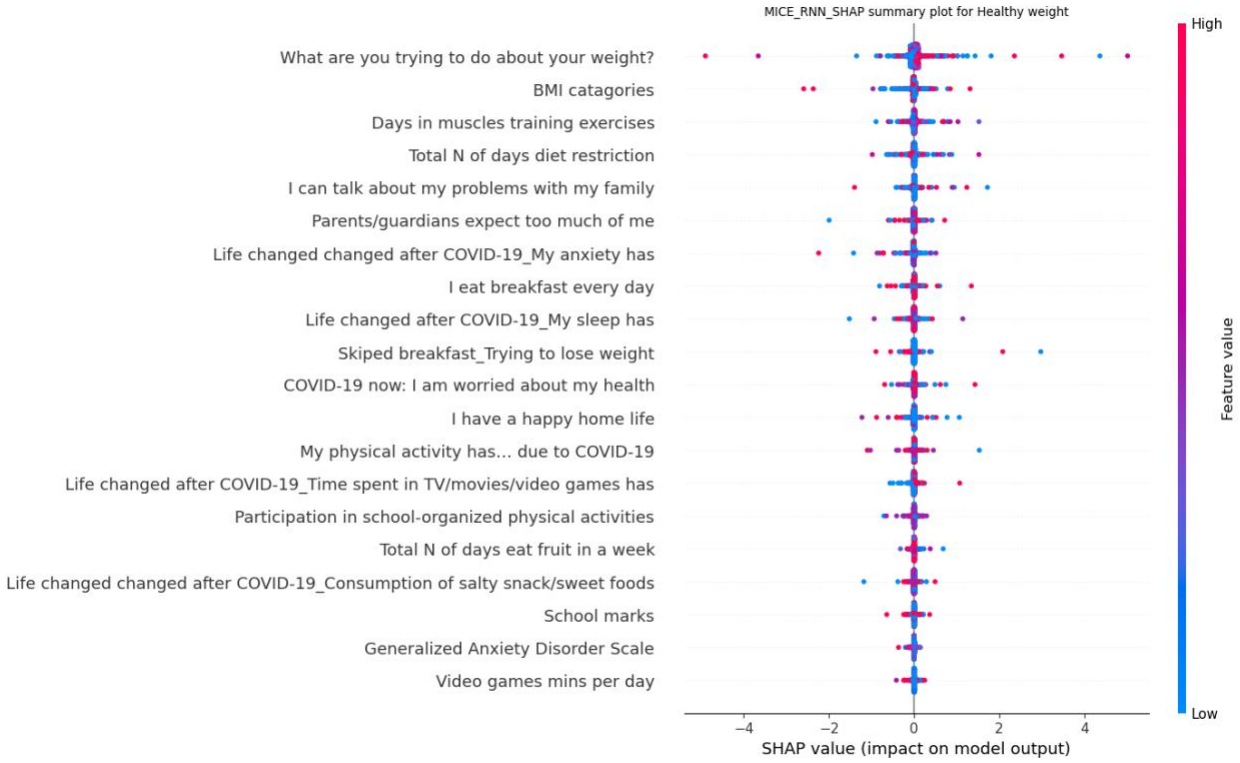
Model	Model Architecture	Optimized Parameters
RNN	RNN with 1 dense layer Recurrent-Layer Dense Layer 1 Dropout Layer 1 Output Layer	activation_idx: tanh batch_size: 16.314 dropout: 0.1 epochs: 50.0 learning_rate: 0.001 optimizer_idx: Adam units: 10.0
LSTM	LSTM with 2 LSTM and 2 dense layers LSTM layer 1 LSTM layer 2 Attention Layer Flatten Layer Dense Layer 1 Dropout Layer 1 Dense Layer 2 Dropout Layer 2 Output Layer	activation_idx: relu batch_size: 122.48 dropout: 0.393 epochs: 63.879 learning_rate: 0.016 optimizer_idx: Adam units: 45.011

Figure S3.9 (A) – (C) Feature importance SHAP beeswarm plot for each weight perception status in RNN

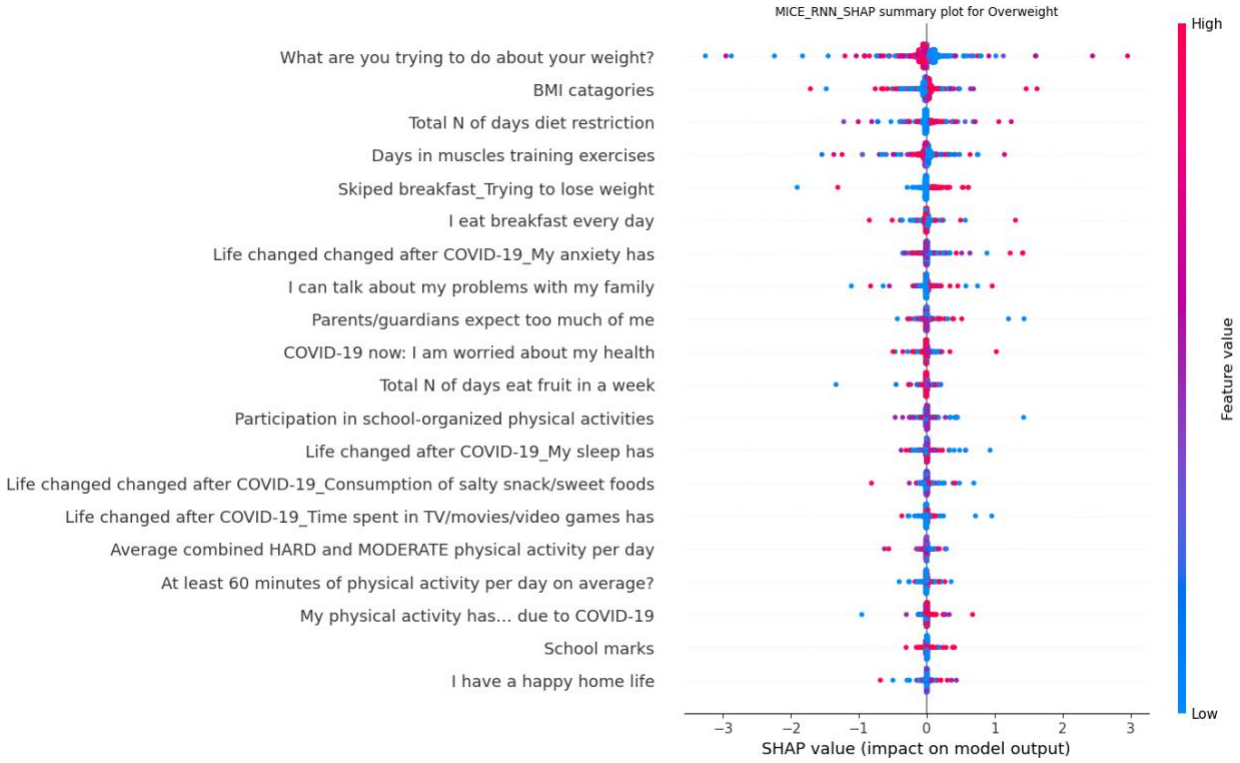
(A) Underweight



(B) About the right weight



(C) Overweight



Appendix E: The COMPASS Student Questionnaire (2020-21)

This is NOT a test. All of your answers are **anonymous** and your identity will be kept **confidential**. No one, not even your parents or teachers, will ever know what you answered. So, please be honest when you answer the questions.

Mark only **one option per question** unless the instructions tell you to do something else.

Choose the option that is the **closest** to what you think/feel is true for you. If you do not understand a question, or if you feel uncomfortable answering it, just leave it blank.

Note: The first five questions are only used to link data from one year to the next. They cannot be used to identify participants. Only University of Waterloo researchers have access to the responses, and they never have access to student names or other information. All responses are strictly anonymous.

Information collected during this survey is stored on a secure server at the University of Waterloo. Please note that when information is transmitted over the internet there remains a possibility of a third party gaining access to that information. That said, **because the survey does not ask for any identifying information (e.g. name), any information you provide will be anonymous.**

Do not copy. For reference only.
Copyright COMPASS 2020.

QA. What is the first letter of your **middle** name? (If you have more than one middle name, use your first middle name; if you do not have a middle name, select the letter "Z":	QB. In which month were you born: _____	QC. What is the **last** letter of your full **last** name: _____	QD. What is the **second** letter of your full **first** name: _____	QE. What is the **first** initial of your **mother's** first name? (Think about the mother you see the most.): _____
A J S B K T C L U D M V E N W F O X G P Y H Q Z I R	1 January 2 February 3 March 4 April 5 May 6 June 7 July 8 August 9 September 10 October 11 November 12 December	A J S B K T C L U D M V E N W F O X G P Y H Q Z I R	A J S B K T C L U D M V E N W F O X G P Y H Q Z I R	A J S B K T C L U D M V E N W F O X G P Y H Q Z I R

0. Did you attend <u>this</u> school last year?

☐ Yes, I attended the same school last year
☐ I am attending a new virtual school this year
☐ No, I was at another school last year

About You

1. **What grade are you in?**
   - ☐ Grade 9
   - ☐ Grade 10
   - ☐ Grade 11
   - ☐ Grade 12

2. **What form of curriculum delivery are you currently participating in?**
   - ☐ 100% in-person (I physically go in to school every day)
   - ☐ 100% online (I attend school remotely over the internet every day)
   - ☐ Alternating in-person and online (Some days I go in to school, some days I attend online)

3. **How old are you today?**
   - ☐ 12 years or younger
   - ☐ 13 years
   - ☐ 14 years
   - ☐ 15 years
   - ☐ 16 years
   - ☐ 17 years
   - ☐ 18 years
   - ☐ 19 years or older

4. **Are you female or male?**
   - ☐ Female
   - ☐ Male
   - ☐ I describe my gender in a different way
   - ☐ I prefer not to say

5. **How would you describe yourself?** (Mark all that apply)

Do not copy. For reference only.
Copyright COMPASS 2020.

☐ Asian
☐ Black
☐ First Nations (status or non-status), Métis, Inuit
☐ Latin American/Hispanic
☐ White
☐ Other

6. **About how much money do you usually get <u>each week</u> to spend on yourself or to save?**
(Remember to include all money from allowances and jobs like babysitting, delivering papers, etc.)

☐ Zero
☐ $1 to $5
☐ $6 to $10
☐ $11 to $20
☐ $21 to $40
☐ $41 to $100
☐ More than $100
☐ I do not know how much money I get each week

7. **Where do you get money to spend on yourself or to save?** *(Mark all that apply)*

☐ I do not usually get any money to spend on myself or to save
☐ My parents/guardians give me money (e.g., an allowance)
☐ I get a pay cheque from a job (working evenings or weekends at a restaurant, store, etc.)
☐ I get paid cash for occasional work (babysitting, mowing lawns, shovelling snow, etc.)

8. **<u>Including you,</u> how many people are currently living at your home?**

☐ One person
☐ Two people
☐ Three people
☐ Four people
☐ Five people
☐ Six or more people

9. **In your house, do you have your own bedroom?**

☐ Yes
☐ No

10. **Do you sometimes go to bed hungry because there is not enough money to buy food?**

☐ Yes
☐ No

11. **Would you say that you and your family are more or less financially comfortable than the average student in your class?**

Do not copy. For reference only.
Copyright COMPASS 2020.

☐ More comfortable
☐ As comfortable
☐ Less comfortable

Your Height and Weight

12. How tall are you without your shoes on? *(Please enter your height either in feet and inches OR in centimetres)*

☐ I do not know how tall I am
☐ I prefer not to say
☐ My height in feet and inches is… Feet: _____Inches: _____
☐ My height in centimetres is… Centimetres: _____

13. How much do you weigh without your shoes on? *(Please enter your weight either in pounds OR in kilograms)*

☐ I do not know how much I weigh
☐ I prefer not to say
☐ My weight in pounds is… Pounds: _____
☐ My weight in kilograms is… Kilograms: _____

14. How do you describe your weight?

☐ Very underweight
☐ Slightly underweight
☐ About the right weight
☐ Slightly overweight
☐ Very overweight

15. Which of the following are you trying to do about your weight?

☐ **Lose** weight
☐ **Gain** weight
☐ **Stay** the same weight
☐ I am **not trying to do anything** about my weight

Do not copy. For reference only.
Copyright COMPASS 2020.

Screen Time and Sleep

16. How much time <u>per day</u> do you *usually* spend doing the following activities? *For example: If you spend about 3 and a half hours watching TV each day, you will need to enter '3' in the hour box and '30' in the minute box.*

	Hours													Minutes			
a) Watching/streaming TV shows or movies	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45
b) Playing video/computer games	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45
c) Doing homework	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45
d) Surfing the internet	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45
e) Texting, messaging, emailing **(note: 50 texts = 30 minutes)**	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45

17. During the past week, what time have you usually turned out the light and gone to sleep…
(Select the time using the drop-down lists for hour, minutes and AM/PM)

	Hours													Minutes				AM/PM	
a) on <u>weekdays</u>?	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45	AM	PM
b) on <u>weekends</u>?	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45	AM	PM

18. During the past week, what time have you usually woken up in the morning…
(Select the time using the drop-down lists for hour, minutes and AM/PM)

	Hours													Minutes				AM/PM	
a) on <u>weekdays</u>?	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45	AM	PM
b) on <u>weekends</u>?	0	1	2	3	4	5	6	7	8	9	10	11	12	0	15	30	45	AM	PM

19. During the past week, how would you rate your sleep quality overall (how well you sleep)?

- ☐ Very good
- ☐ Fairly good
- ☐ Fairly bad
- ☐ Very bad

Do not copy. For reference only.
Copyright COMPASS 2020.

Physical Activity

HARD physical activities include jogging, team sports, fast dancing, jump-rope, and any other physical activities that increase your heart rate and make you breathe hard and sweat.
MODERATE physical activities include lower intensity activities such as walking, biking to school, and recreational swimming.

20. Mark how many minutes of HARD physical activity you did on each of the last 7 days. This includes physical activity during physical education class, lunch, after school, evenings, and spare time.
For example: If you did 45 minutes of hard physical activity on Monday, you will need to select the 0 hour option and the 45 minute option in the drop-down boxes next to Monday.

	Hours					Minutes			
Monday	0	1	2	3	4	0	15	30	45
Tuesday	0	1	2	3	4	0	15	30	45
Wednesday	0	1	2	3	4	0	15	30	45
Thursday	0	1	2	3	4	0	15	30	45
Friday	0	1	2	3	4	0	15	30	45
Saturday	0	1	2	3	4	0	15	30	45
Sunday	0	1	2	3	4	0	15	30	45

21. Mark how many minutes of MODERATE physical activity you did on each of the last 7 days. This includes physical activity during physical education class, lunch, after school, evenings, and spare time. Do not include time spent doing hard physical activities.
For example: If you did 1 hour and 30 minutes of moderate physical activity on Monday, you will need to select the 1 hour option and the 30 minute option in the drop-down boxes next to Monday.

	Hours					Minutes			
Monday	0	1	2	3	4	0	15	30	45
Tuesday	0	1	2	3	4	0	15	30	45
Wednesday	0	1	2	3	4	0	15	30	45
Thursday	0	1	2	3	4	0	15	30	45
Friday	0	1	2	3	4	0	15	30	45
Saturday	0	1	2	3	4	0	15	30	45
Sunday	0	1	2	3	4	0	15	30	45

22. On how many days in the last 7 days did you do exercises to strengthen or tone your muscles? (e.g., push-ups, sit-ups, or weight-training)

Do not copy. For reference only. Copyright COMPASS 2020.

☐ 0 days
☐ 1 day
☐ 2 days
☐ 3 days
☐ 4 days
☐ 5 days
☐ 6 days
☐ 7 days

23. Do you participate in before-school, noon hour, or after-school physical activities organized by your school? *(e.g., intramurals, non-competitive clubs)*

☐ Yes
☐ No
☐ None offered at my school

24. Do you participate in competitive sports teams that compete against other schools? *(e.g., junior varsity or varsity sports)*

☐ Yes
☐ No
☐ None offered at my school

25. Do you participate in league or team sports outside of school?

☐ Yes
☐ No
☐ There are none available where I live

Healthy Eating

26. If you do not eat breakfast every day, why do you skip breakfast? *(Mark all that apply)*

☐ I eat breakfast every day
☐ I don't have time for breakfast
☐ The bus comes too early
☐ I sleep in
☐ I'm not hungry in the morning
☐ I feel sick when I eat breakfast
☐ I'm trying to lose weight
☐ There is nothing to eat at home
☐ Other

27. During the past week, on which days did you do the following? *(Mark all that apply)*

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
a) Eat food purchased at a fast food place or restaurant	☐	☐	☐	☐	☐	☐	☐
b) Eat snacks purchased from a vending machine, corner store, snack bar, or canteen	☐	☐	☐	☐	☐	☐	☐

Do not copy. For reference only.
Copyright COMPASS 2020.

c) Restrict/alter your food intake with the intention of changing your weight/shape	☐	☐	☐	☐	☐	☐	☐
d) Drink water (plain)	☐	☐	☐	☐	☐	☐	☐
e) Eat fruit (fresh, canned, dried, or frozen)	☐	☐	☐	☐	☐	☐	☐
f) Eat vegetables (raw or cooked, fresh, canned, or frozen)	☐	☐	☐	☐	☐	☐	☐

Your Experience with Smoking and Vaping

28. Have you ever tried cigarette smoking, even just a few puffs?

- ☐ Yes
- ☐ No

29. If one of your best friends was to offer you a cigarette, would you smoke it?

- ☐ Definitely yes
- ☐ Probably yes
- ☐ Probably not
- ☐ Definitely not

30. At any time during the next year do you think you will smoke a cigarette?

- ☐ Definitely yes
- ☐ Probably yes
- ☐ Probably not
- ☐ Definitely not

31. Do you think in the future you might try smoking cigarettes?

- ☐ Definitely yes
- ☐ Probably yes
- ☐ Probably not
- ☐ Definitely not

32. On how many of the last 30 days did you smoke one or more cigarettes?

- ☐ None
- ☐ 1 day
- ☐ 2 to 3 days
- ☐ 4 to 5 days
- ☐ 6 to 10 days
- ☐ 11 to 20 days
- ☐ 21 to 29 days
- ☐ 30 days (every day)

33. Have you ever tried to quit smoking cigarettes?

Do not copy. For reference only.
Copyright COMPASS 2020.

☐ I have never smoked
☐ I have only smoked a few times
☐ I have never tried to quit
☐ I have tried to quit once
☐ I have tried to quit 2 or more times

Vapes (or vaporizers) are devices that produce vapour instead of smoke. They are sometimes called "**e-cigarettes"**, "e-cigs", or "vape pens". Vapes can have pods or tanks and can look like USB drives or pens. *Some examples of vapes are **JUUL**, **Vype**, **Suorin**, and **Smok***

34. Have you ever tried a vape, also known as an e-cigarette? (e.g., JUUL, Vype, Suorin, Smok)

☐ Yes
☐ No

35. Have you used a vape for any of the following reasons? *(Mark all that apply)*

☐ I have not used a vape
☐ Curiosity / to try something new
☐ It's cool / a lot of people I know use them
☐ I can vape in places where smoking is not allowed
☐ To help me quit smoking cigarettes
☐ To relax or relieve stress and anxiety
☐ For the nicotine high
☐ I really like the flavours
☐ I have vaped for some other reason

36. On how many of the last 30 days did you use a vape?

☐ None
☐ 1 day
☐ 2 to 3 days
☐ 4 to 5 days
☐ 6 to 10 days
☐ 11 to 20 days
☐ 21 to 29 days
☐ 30 days (every day)

37. Where do you *usually* get vapes (including pods, e-juice, or e-liquid)? *(Choose only one)*

☐ I buy them from a vape shop myself
☐ I buy them from a convenience store myself
☐ I buy them from another type of store myself
☐ I buy them on the Internet (including apps)
☐ I ask someone to buy them for me
☐ A family member gives them to me
☐ A friend gives them to me

Do not copy. For reference only.
Copyright COMPASS 2020.

☐ Someone else gives them to me
☐ I use someone's vape without their permission
☐ Other

38. When you vape, how often to do you share the device with someone else or use someone else's vape?

☐ Never
☐ Rarely
☐ Sometimes
☐ Most of the time
☐ Always

39. In the last 30 days, did you use any of the following? *(Mark all that apply)*

☐ Cigarillos or little cigars *(plain or flavoured)*
☐ Cigars (not including cigarillos or little cigars, *plain or flavoured*)
☐ Loose tobacco mixed with marijuana
☐ Heated tobacco product (a device that heats tobacco instead of burning it, such as IQOS or Heatstick)
☐ Smokeless tobacco (chewing tobacco, pinch, snuff, or snus)
☐ Nicotine patches, nicotine gum, nicotine lozenges, or nicotine inhalers
☐ I have not used any of these things in the last 30 days

Alcohol and Drug Use

A DRINK means: 1 regular sized bottle, can, or draft of beer; 1 glass of wine; 1 bottle of cooler; 1 shot of liquor (rum, whisky, etc); or 1 mixed drink (1 shot of liquor with pop, juice, energy drink).
Please remember that we will keep your answers ***completely confidential****.*

40. In the last 12 months, how often did you have a drink of alcohol that was more than just a sip?

☐ I have never drunk alcohol
☐ I did not drink alcohol in the last 12 months
☐ I have only had a sip of alcohol
☐ Less than once a month
☐ Once a month
☐ 2 or 3 times a month
☐ Once a week
☐ 2 or 3 times a week
☐ 4 to 6 times a week
☐ Every day

Do not copy. For reference only.
Copyright COMPASS 2020.

41. How old were you when you first had a drink of alcohol that was more than just a sip?

☐ I have never drunk alcohol
☐ I have only had a sip of alcohol
☐ I do not know
☐ 8 years or younger
☐ 9 years
☐ 10 years
☐ 11 years
☐ 12 years
☐ 13 years
☐ 14 years
☐ 15 years
☐ 16 years
☐ 17 years
☐ 18 years or older

42. In the last 12 months, how often did you have 5 drinks of alcohol or more on one occasion?

☐ I have never done this
☐ I did not have 5 or more drinks on one occasion in the last 12 months
☐ Less than once a month
☐ Once a month
☐ 2 to 3 times a month
☐ Once a week
☐ 2 to 5 times a week
☐ Daily or almost daily

43. In the last 12 months, how often did you use marijuana or cannabis? *(a joint, pot, weed, hash)*

☐ I have never used marijuana
☐ I have used marijuana but not in the last 12 months
☐ Less than once a month
☐ Once a month
☐ 2 or 3 times a month
☐ Once a week
☐ 2 or 3 times a week
☐ 4 to 6 times a week
☐ Every day

Do not copy. For reference only.
Copyright COMPASS 2020.

44. If you have used marijuana or cannabis in the last 12 months, how did you use it? *(Mark all that apply)*

☐ I have used it by smoking it (e.g., in a joint, a pipe, a bong)
☐ I have used it by vaping it
☐ I have used it by eating or drinking it (e.g., in brownies, cookies, candies, tea)
☐ I have not used marijuana or cannabis in the last 12 months

45. How old were you when you first used marijuana or cannabis?

☐ I have never used marijuana
☐ I do not know
☐ 8 years or younger
☐ 9 years
☐ 10 years
☐ 11 years
☐ 12 years
☐ 13 years
☐ 14 years
☐ 15 years
☐ 16 years
☐ 17 years
☐ 18 years or older

46. Do you think it would be difficult or easy for you to get marijuana if you wanted some?

☐ Difficult
☐ Easy
☐ I do not know

47. In the last 30 days, how many times have you done the following?

	Never	Once	Twice	3 or more times	I do not know
a) Rode in/on a vehicle when the driver (you or someone else) had been drinking alcohol	☐	☐	☐	☐	☐
b) Rode in/on a vehicle when the driver (you or someone else) had been using marijuana or cannabis	☐	☐	☐	☐	☐
c) Drank alcohol when you were all by yourself	☐	☐	☐	☐	☐

Do not copy. For reference only.
Copyright COMPASS 2020.

d) Used marijuana or cannabis when you were all by yourself	☐	☐	☐	☐	☐

48. Have you used or tried any of the following medications TO GET HIGH?

	NO, I have never done this	YES, I have done this in the last 12 months	YES, I have done this but NOT in the last 12 months
a) Oxycodone (oxy, OC, APO, OxyContin®, percs, roxies, OxyNEO®)	☐	☐	☐
b) Fentanyl (china white, synthetic heroin, china girl)	☐	☐	☐
c) Other prescription pain relievers (codeine, morphine, Tylenol 3)	☐	☐	☐

49. Do you think it would be difficult or easy to get pain relievers (Oxycodone, Fentanyl, codeine, etc.) if you wanted some?

☐ Difficult
☐ Easy
☐ I do not know

50. For each of the following behaviours, please indicate the extent to which you believe an individual would be exposed to risks to his or her physical or other health:

	No Risk	Low Risk	Medium Risk	High Risk	I do not know
a) S/he has tried e-cigarettes a few times	☐	☐	☐	☐	☐
b) S/he smokes e-cigarettes regularly	☐	☐	☐	☐	☐
c) S/he has tried cannabis (marijuana, hashish) a few times	☐	☐	☐	☐	☐
d) S/he smokes cannabis (marijuana, hashish) regularly	☐	☐	☐	☐	☐

Mental Health

51. How much do you agree or disagree with the following statements?

	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
a) I have a happy home life	☐	☐	☐	☐	☐
b) My parents/guardians expect too much of me	☐	☐	☐	☐	☐
c) I can talk about my problems with my family	☐	☐	☐	☐	☐

Do not copy. For reference only. Copyright COMPASS 2020.

d) I can talk about my problems with my friends	☐	☐	☐	☐	☐

52. How much do you agree or disagree with the following statements?

	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree
a) I lead a purposeful and meaningful life	☐	☐	☐	☐	☐
b) My social relationships are supportive and rewarding	☐	☐	☐	☐	☐
c) I am engaged and interested in my daily activities	☐	☐	☐	☐	☐
d) I actively contribute to the happiness and well-being of others	☐	☐	☐	☐	☐
e) I am competent and capable in the activities that are important to me	☐	☐	☐	☐	☐
f) I am a good person and live a good life	☐	☐	☐	☐	☐
g) I am optimistic about my future	☒	☐	☐	☐	☐
h) People respect me	☐	☐	☐	☐	☐
i) I generally recover from setbacks quickly	☐	☐	☐	☐	☐

53. Choose the answer that best describes how you feel.

	True	Mostly true	Sometimes true, sometimes false	Mostly false	False
a) In general, I like the way I am	☐	☐	☐	☐	☐
b) Overall, I have a lot to be proud of	☐	☐	☐	☐	☐
c) A lot of things about me are good	☐	☐	☐	☐	☐
d) When I do something, I do it well	☐	☐	☐	☐	☐
e) I like the way I look	☐	☐	☐	☐	☐

54. Over the last 2 weeks, how often have you been bothered by the following problems?

	Not at all	Several days	Over half the days	Nearly every day
a) Feeling nervous, anxious, or on edge	☐	☐	☐	☐
b) Not being able to stop or control worrying	☐	☐	☐	☐
c) Worrying too much about different things	☐	☐	☐	☐
d) Trouble relaxing	☐	☐	☐	☐
e) Being so restless that it is hard to sit still	☐	☐	☐	☐
f) Becoming easily annoyed or irritable	☐	☐	☐	☐

Do not copy. For reference only.
Copyright COMPASS 2020.

g) Feeling afraid as if something awful might happen	☐	☐	☐	☐

55. Please indicate how often the following statements apply to you:

	Almost Never	Sometimes	About half the time	Most of the time	Almost always
a) I have difficulty making sense out of my feelings	☐	☐	☐	☐	☐
b) I pay attention to how I feel	☐	☐	☐	☐	☐
c) When I'm upset, I have difficulty concentrating	☐	☐	☐	☐	☐
d) When I'm upset, I believe there is nothing I can do to make myself feel better	☐	☐	☐	☐	☐
e) When I'm upset, I lose control over my behaviour	☐	☐	☐	☐	☐
f) When I'm upset, I feel ashamed for feeling that way	☐	☐	☐	☐	☐

56. On how many of the last 7 days did you feel the following ways?

	None or less than 1 day	1 to 2 days	3 to 4 days	5 to 7 days
a) I was bothered by things that usually don't bother me	☐	☐	☐	☐
b) I had trouble keeping my mind on what I was doing	☐	☐	☐	☐
c) I felt depressed	☐	☐	☐	☐
d) I felt that everything I did was an effort	☐	☐	☐	☐
e) I felt hopeful about the future	☐	☐	☐	☐
f) I felt fearful	☐	☐	☐	☐
g) My sleep was restless	☐	☐	☐	☐
h) I was happy	☐	☐	☐	☐
i) I felt lonely	☐	☐	☐	☐
j) I could not get "going"	☐	☐	☐	☐

57. In general, how would you rate your mental health?

☐ Excellent
☐ Very good
☐ Good
☐ Fair
☐ Poor

Do not copy. For reference only.
Copyright COMPASS 2020.

Note. If you are a young person in Canada who needs support, you can reach out to Kids Help Phone's professional counsellors by calling 1-800-668-6868 or visiting kidshelpphone.ca. Their service is free, anonymous, confidential, and available 24/7/365.

Kids Help Phone

1-800-668-6868

Your School and You

58. How strongly do you agree or disagree with each of the following statements?

	Strongly Agree	Agree	Disagree	Strongly Disagree
a) I feel close to people at my school.	☐	☐	☐	☐
b) I feel I am part of my school.	☐	☐	☐	☐
c) I am happy to be at my school.	☐	☐	☐	☐
d) I feel the teachers at my school treat me fairly.	☐	☐	☐	☐
e) I feel safe in my school.	☐	☐	☐	☐
f) Getting good grades is important to me.	☐	☐	☐	☐

The following question is about bullying. **Bullying** can take many forms and usually involves intimidating, threatening, or otherwise hurtful behaviour. Bullying includes any sort of harassment related to sexual orientation, gender identity, gender expression, or sexual harassment.

59. In the last 30 days, in what ways were you bullied by other students? *(Mark all that apply)*

☐ I have not been bullied in the last 30 days
☐ Physical attacks (e.g., getting beaten up, pushed, or kicked)
☐ Verbal attacks (e.g., getting teased, threatened, or having rumours spread about you)
☐ Cyber-attacks (e.g., being sent mean text messages or having rumours spread about you on the internet)
☐ Social bullying (e.g., being purposefully excluded from a group, being humiliated by others with gestures or graffiti)
☐ Had someone steal from you or damage your things

60.On average, what marks do you usually get in school? *(Choose only one)*

☐ 90% - 100% *(Mostly A+)*
☐ 80% - 89% *(Mostly As or A-)*
☐ 70% - 79% *(Mostly Bs)*
☐ 60% - 69% *(Mostly Cs)*
☐ 50% - 59% *(Mostly Ds)*
☐ Below 50% *(Mostly Fs)*

61.What is the highest level of education you would like to get? *(Choose only one)*

Do not copy. For reference only. Copyright COMPASS 2020.

☐ Some high school or less
☐ High school diploma or graduation equivalency
☐ College/trade/vocational certificate
☐ University Bachelor's degree
☐ University Master's / PhD / law school / medical school / teachers' college degree
☐ I don't know

The COVID-19 Pandemic

The following questions deal with the COVID-19 pandemic. **COVID-19** is the infectious disease that is currently active in Canada and globally and is the reason why governments in Canada have introduced policies designed to protect the health of the population (e.g., promoting social/physical distancing, wearing a mask, banning large gatherings, telling people to stay home, etc.). **We are interested to know how you feel about this situation and what effects it is having on your daily life.**

62. How would you describe your current level of concern regarding COVID-19?

☐ Not at all concerned
☐ Somewhat concerned
☐ Very concerned

63. How true are the following statements about COVID-19 for you right now?

	True	Mostly True	Neutral / I don't know	Mostly False	False
a) I am nervous when I think about current circumstances	☐	☐	☐	☐	☐
b) I am calm and relaxed	☐	☐	☐	☐	☐
c) I am worried about my health	☐	☐	☐	☐	☐
d) I am worried about the health of my family members	☐	☐	☐	☐	☐
e) I feel stressed about leaving my house	☐	☐	☐	☐	☐
f) I am getting along well with my family	☐	☐	☐	☐	☐
g) I don't like having to stay home	☐	☐	☐	☐	☐
h) I am worried about getting behind in school work	☐	☐	☐	☐	☐
i) I am upset about missing events/activities (sports, concerts, parties)	☐	☐	☐	☐	☐
j) I am worried about my family being able to pay bills and expenses	☐	☐	☐	☐	☐
k) I am concerned about the future	☐	☐	☐	☐	☐
l) I think that COVID-19 represents very little risk to young people	☐	☐	☐	☐	☐
m) I am supportive of wearing a mask in indoor public spaces	☐	☐	☐	☐	☐

64. How seriously are you taking the new rules laid out by governments to reduce the spread of COVID-19?

Do not copy. For reference only.
Copyright COMPASS 2020.

☐ I take them very seriously (e.g. I wear a mask where I am supposed to, I do not go within 2 metres of people)

☐ I take them somewhat seriously (e.g. I sometimes wear a mask, I sometimes stay a full 2 metres away from people)

☐ I do not take them seriously (e.g. I don't wear a mask, I do not worry about staying 2 meters away from people)

65. How has your life changed because of COVID-19?

	Increased	Stayed the same / Not applicable	Decreased
My stress has…	☐	☐	☐
My boredom has…	☐	☐	☐
My loneliness has…	☐	☐	☐
My anxiety has…	☐	☐	☐
Time spent with my friends in person has…	☐	☐	☐
Time spent communicating with my friends online has…	☐	☐	☐
Time spent watching TV/movies or playing video games has…	☐	☐	☐
My time spent surfing/posting on social media has…	☐	☐	☐
My consumption of salty snack (e.g., chips) and sweet foods (e.g., candy, ice cream, cake) has...	☐	☐	☐
My physical activity has…	☐	☐	☐
My sleep has…	☐	☐	☐
My cigarette use has…	☐	☐	☐
My vaping has…	☐	☐	☐
My alcohol use has…	☐	☐	☐
My cannabis use has…	☐	☐	☐

66.Where have you been getting information/news about COVID-19? *(Mark all that apply)*

☐ News sources (TV news, news websites, newspapers)
☐ Social media (Twitter, Facebook, etc.)
☐ My school (teachers, principals, school staff)
☐ My parent(s)/guardian(s)
☐ My friends
☐ Government or Public Health websites
☐ I have been trying to limit my exposure to news about COVID-19

Do not copy. For reference only.
Copyright COMPASS 2020.

67. How have you been coping with changes related to COVID-19? *(Mark all the apply)*

- ☐ Staying connected with friends online
- ☐ Meeting up with my friends outside
- ☐ Getting exercise (e.g., getting outside to go for a walk or bike ride, working out at home)
- ☐ Keeping to a regular schedule (e.g., waking up, eating meals, and going to bed around the same time as usual)
- ☐ Spending time with my family (e.g., playing games, eating meals together, hanging out)
- ☐ Spending time with my dog/cat or other pet
- ☐ Spending time alone
- ☐ Studying or working on school work
- ☐ Reading, writing, playing music, or working on arts and crafts
- ☐ Playing video games, watching TV or movies, surfing the internet/social media
- ☐ Learning something new (painting, playing a musical instrument, a new language)
- ☐ Connecting with mental health professionals
- ☐ Eating junk food
- ☐ Cooking/baking
- ☐ Trying to help others
- ☐ Meditating or praying
- ☐ Using cannabis/marijuana
- ☐ Drinking alcohol
- ☐ Smoking cigarettes
- ☐ Vaping

68. Please indicate which of the following statements about COVID-19 <u>you think</u> are true. *(Mark all that apply)*

- ☐ Hand washing for at least 20 seconds (with soap and water) helps prevent the transmission of COVID-19
- ☐ The use of hand sanitizer (e.g., Purell) helps prevent the transmission of COVID-19.
- ☐ When a person coughs, the use of a mask can reduce the droplet transmission of COVID-19.
- ☐ COVID-19 is only found in individuals who show symptoms and signs of disease.
- ☐ COVID-19 is only dangerous for the elderly population.

Do not copy. For reference only
Copyright COMPASS 2020.

69. Since the beginning of the COVID-19 pandemic, have you done the following?

	Always	Sometimes	Never
a) I cancelled or postponed meetings with friends, eating-out, and sports events	☐	☐	☐
b) I clean and disinfect items that can be easily touched with hands (i.e. door handles and surfaces)	☐	☐	☐
c) I washed my hands more often	☐	☐	☐
d) I avoided coughing around people	☐	☐	☐
e) I discussed measures to prevent infection with family, friends and/or health care professionals	☐	☐	☐
f) I discussed what to do in case of infection, with family, friends or health care providers	☐	☐	☐
g) I avoided places with a large number of people gathered (public spaces, grocery stores, etc.)	☐	☐	☐

70. How supportive are you of the safety precautions your school has taken to protect against COVID-19?

	Supportive	Neutral	Unsupportive	My school does not do this
a) Requiring students to wear masks	☐	☐	☐	☐
b) Separating students into cohorts	☐	☐	☐	☐
c) Limiting interactions with classmates	☐	☐	☐	☐
d) Increasing availability of cleaning products like hand sanitizer	☐	☐	☐	☐
e) Changes to school sports activities (intramurals, varsity sports)	☐	☐	☐	☐

71. How strongly do you agree or disagree with the following statements?

	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree
a) If a vaccine against COVID-19 was offered to me, I would take it	☐	☐	☐	☐	☐
b) It is important for everyone to get the COVID-19 vaccine	☐	☐	☐	☐	☐
c) I believe that, in general, vaccines are safe	☐	☐	☐	☐	☐
d) I believe that, in general, vaccines are effective	☐	☐	☐	☐	☐

Do not copy. For reference only. Copyright COMPASS 2020.