COMPASS Vision: An Interactive Visualization Supporting the Exploration of COMPASS System Youth Health Data

Alex Pawelczyk, Oliver Kavelman, Calum Bruton, and Gagandeep Varya

University of Waterloo

CS792/HLTH612: Data Structures and Standards in Health Informatics

Dr. Helen Chen

April 29th, 2020

Background and Challenges Addressed

The COMPASS longitudinal study began in 2012 with a primary objective of improving youth prevention research and practice (Reel, Bredin & Leatherdale, 2018). Adapted from the Canadian Cancer Society's School Health Action Planning and Evaluation System (SHAPES) framework, the COMPASS system was developed to address knowledge gaps in school-based prevention research and to provide a comprehensive research, evaluation, and knowledge exchange system (Leatherdale et al., 2014). The system facilitates the collection, translation, and exchange of student-level and school-level health data from high-school students at participating institutions (124 schools across four provinces and one territory as of Year 6 - 2018) (Reel, Bredin & Leatherdale, 2018). In COMPASS research, student participants are asked about aspects of their personal health behaviour such as substance use, physical activity, and mental health via anonymous, annual questionnaires (Leatherdale et al., 2014). Participating institutions are first evaluated based on their existing health policies and programs and subsequently undergo a facility evaluation (conducted by COMPASS researchers) that examines health influencing characteristics of their internal and external environment (Leatherdale et al., 2014). Since COMPASS research captures data from multiple sources (student questionnaires, environmental assessments, policy evaluations), the COMPASS dataset is diverse and applicable to many health-related disciplines. COMPASS research has produced dozens of academic publications covering topics including environmental health, health promotion, and preventative medicine (University of Waterloo, 2020a). Analysis of COMPASS data has also allowed researchers to identify and target school-specific health priorities for participating institutions.

Contextually relevant information is crucial for developing meaningful interventions that target modifiable risk factors for chronic diseases and health behaviour. Early COMPASS publications noted the substantial variability across Canadian jurisdictions, not only in terms of youth physical activity levels, substance use, and mental wellbeing, but also with regards to healthy school environments and policies (Leatherdale et al., 2014). Context specific adaptation activities are supported by COMPASS research and generate additional practice-based evidence that can be reapplied to similar settings (University of Waterloo, 2020a). With COMPASS data, youth health interventions are better informed and can be optimized by adapting programs or policies based on recognized capacities and needs.

Recently, an interactive COMPASS data visualization tool was developed that allows users to explore health risk factor trends from data gathered in Year 6 of the study (2017/2018). The tool currently provides one level of granularity (provincial level) for select COMPASS indicators including tobacco use, obesity rates and physical activity levels. Although the geographic visualizations can be customized to some extent by specifying attributes like gender or grade, data exploration capabilities with the existing tool are limited.

Increasingly, machine learning (ML) techniques are being applied to public health datasets to explore associations, identify disease patterns and predict health outcomes (Mooney and Pejaver, 2018). Supervised learning algorithms have been developed to predict hospital readmissions and tuberculosis transmission while cluster analysis has been used to conduct public health surveillance and associate patient characteristics with clinical outcomes (Mooney and Pejaver, 2018). ML techniques may also be used for generating hypotheses from large datasets and could be used to inform healthy policies (Mooney and Pejaver, 2018). Interpretability of model output and of models themselves has been noted as a concern where ML has been applied to public health and clinical medicine (Mooney and Pejaver, 2018). Early clinical applications of ML were criticized for 'black boxing' decision making processes but this issue can now be mitigated with the use of interpretable models (e.g. decision trees and GLM's) and model-agnostic methods (e.g. LIME and SHAP) (Mooney and Pejaver, 2018). Model-based collaborative filtering (CF) with alternating least squares (ALS) has been applied elsewhere to provide recommendations based on user 'likes' or attributes (Piccardi et al., 2018).

Although it is known that healthy habits during adolescence tend to persist into adulthood, the amount of research focusing on social correlates of youth health behaviour (e.g. peers relations, parental support, school programs) has been inadequate (Lau, Faulkner, Qian & Leatherdale, 2016). The COMPASS study addresses this topic as it relates to Canadian high school students but is not scaled or intended to produce population-level statistics. COMPASS Vision applies ML techniques to the COMPASS dataset to enhance data exploration capabilities of the existing visualization tool and identify complex associations between variables. With a visualization tool supported by ML algorithms, COMPASS researchers are provided with the opportunity to interact with the COMPASS dataset in a new and meaningful fashion. Specifically, the primary contributions of this work are as follows: (1) interactive visualization maps that aim to support evidence-based decision making process of COMPASS stakeholders, (2) an alternating least squares (ALS) machine learning model for predicting the "preference" that students are likely to have towards to important COMPASS health variables, and (3) Decision Tree, Random Forest, and Gradient Boosted Tree approaches to predicting COMPASS student health factors.

Currently, COMPASS data is shared through direct collaboration with external youth prevention researchers and using the existing data visualization map, albeit with limited functionality. COMPASS Vision will support youth prevention research in Canada by improving the accessibility and interpretability of high-level COMPASS data. Ideally, the improved visualizations will: (1) expose national and regional trends in youth health behaviour, (2) inform youth prevention interventions, and (3) reveal the significant between-school variability that exists in terms of youth health risk factors (Leatherdale et al., 2014). The ML techniques provided by COMPASS Vision will assist COMPASS researchers in identifying associations among the dozens of measured variables and with defining high-risk student groups. Since healthy behaviour (and the development of many chronic diseases) is often determined by a multitude of interrelated social, individual and environmental factors, identifying particularly impactful risk factors is challenging (Huang, Chen & Lee, 2007). This, and the complexity of the COMPASS dataset, makes ML an attractive solution. At the very least, COMPASS Vision ML capabilities provide a novel strategy for manipulating COMPASS data and exposing associations between risk factors.

Main Use Cases and Functions

COMPASS Vision aims to combine the powers of machine learning and data visualization to create an evidence-based platform for exploring important COMPASS student health factors. Since these types of factors may vary among schools, COMPASS Vision leverages the powers of dynamic map visualizations to present users with region-specific information about the COMPASS students. Upon its full completion, COMPASS Vision will enable users to explore clusters of students based on their predicted health-risk factors, identify important health analytics information on a per-province basis, and learn from detailed visualizations of machine learning model predictions. However, the initial COMPASS Vision prototype is limited in these functionalities. The homepage uses Google Maps marker clustering to visualize COMPASS school participation statistics, a Google Geo Chart displays only one important health analytics query (smoking susceptibility by province), and machine learning visualizations show summaries of evaluation metrics, rather than actual predictions or the model learning process. Nonetheless, the initial prototype can still be used to uncover meaningful information, and this section details three use case scenarios for the first COMPASS Vision prototype.

Use Case Scenario 1: University Researcher

Roger Smithfield is a current health informatics student at the University of Guelph. As part of his ongoing research, he is writing a report on the current state of health of Canadian high school students. He has heard about the COMPASS system in the past, and after conducting a quick Google search, he comes across the COMPASS Vision Web interface. Upon loading the homepage, he is presented with a map that shows summary statistics of schools that participate in the COMPASS study. As he plays around with the zooming-functionalities, he learns that each cluster displays a number which represents the number of schools displayed in that cluster. He also notices that clicking on a marker displays the number of students in that particular school. Roger then proceeds to the Geo Chart Analysis page and finds a map that shows the percentage of students in each province who are susceptible to smoking cigarettes. This is the credible evidence that he was looking for, and he records the statistics for each province. Roger also wants to investigate if differences in the number of students who attend a school may play a role in higher smoking susceptibility. Thus, he goes back to the homepage and uses the map to find the average number of students per school in British Columbia (province with highest number of students who are susceptible to smoking) and Alberta (province with the lowest number of students who are susceptible to smoking). He notes that British Columbia had on average 277 more students per school than Alberta, and plans to conduct further research to investigate if larger school sizes correlate to increased substance abuse among students.

Use Case Scenario 2: Parents of Canadian High School Students

Ms. Singh is a single mother who has recently moved from the United States to Waterloo with her two teenagers. She has a son in grade 9 and daughter in grade 11. Ms. Singh knows her divorce and their move has been tough on her children. She knows her children were not happy to leave their friends behind and start high school in a new city. She is worried they may be peer pressured into experimenting with drugs in order to fit in. Being a concerned and proactive parent, Ms. Singh decides to do some research on Canadian student substance abuse. She discovers the COMPASS study led by researchers from the University of Waterloo, and after coming across COMPASS Vision, she decides to use the Web interface to view what kind of risk factors her children could be exposed to at school and what they are most at risk for. She found that Ontario has the second highest percentage of students who are susceptible to smoking, where approximately one in every five students is at risk. Based on this information, Ms. Singh decides to have a conversation with her children and hopes to inform them of the risk factors.

Use Case Scenario 3: Policy Makers and School Stakeholders

With school board elections coming around soon, the policy makers, school board trustees, program administrators, and other personnel from Vancouver's District School Board would like to determine what school programs need to be allocated to which school based on student data. The policy makers are interested in data that can help them understand how school environment characteristics are impacting youth behaviour. Based on the student data, they will make a decision on funds allocation to substance abuse programs or mental health programs for participating schools. The policy makers want to identify and evaluate health behaviours and outcomes of high school students, so they use the COMPASS Vision interface to find new insights. For example, Vancouver's District School Board allocates a certain amount of funds towards educating students on the risks of smoking cigarettes and providing appropriate support programs. Thus, they use the COMPASS Vision Geo Chart to get updated statistics on the smoking susceptibility of British Columbia students. The team was alarmed to find that British Columbia has the highest percentage of students susceptible to smoking out of all the provinces that participated in the COMPASS study. Along with evidence obtained from other sources, the Vancouver District Board plans to use the numbers from COMPASS Vision in a report to justify why British Columbia needs to implement smoking awareness and support programs.

Data Sources and Standards

As previously mentioned, the COMPASS longitudinal study examines the relationship between school environmental characteristics and youth health behaviours. Data collection is an integral aspect of the COMPASS system and is foundational for subsequent processes (e.g. knowledge translation, intervention activities, system improvement). To preserve data integrity, strict protocols have been developed to ensure data collection is consistent across participating schools. Substantial efforts have also been made to streamline data collection methods so as to minimize interruptions to class time and reduce the burden of work on participating schools (Thompson-Haile & Leatherdale, 2013a).

COMPASS researchers make use of multiple data collection tools that have been specifically designed to capture actionable, context-specific data (Leatherdale, 2009). Student-level data, which forms the bulk of the COMPASS dataset, is gathered using the paper-based questionnaire C_q (Reel, Battista, Bredin et al., 2019). The 12-page questionnaire is completed anonymously and consists mainly of multiple-choice questions that inquire about physical characteristics, health behaviour and academic performance (Leatherdale et al., 2014). The data values generated through completion of the C_q are largely categorical however some continuous values are reported for select variables such as weight, height, and hours of physical activity. Eligible students at participating secondary institutions complete the questionnaire during class time on a prearranged ‘data collection day’ (Reel et al., 2019). For institutions that participate in COMPASS research across multiple years, C_q questionnaires are conducted annually. The questionnaire has been adapted several times over the course of the study in response to participant feedback and to better reflect emerging COMPASS research priorities (e.g. cannabis use among youth in wake of legalization) (Reel et al., 2019).

The characteristics of the schools participating in COMPASS research are evaluated using three data collection tools. Details regarding existing policies and programs are typically reported by having a knowledgeable school administrator complete the SPP Questionnaire (Leatherdale et al., 2014). The SPP is completed annually (at the same time as the C_q) and provides researchers with an overview of each schools’ policy environment (Leatherdale et al., 2014). Alternatively, the COMPASS School Environment Application (Co-SEA) is used to measure aspects of a school’s internal built environment as they relate to youth health and youth health behaviour (Leatherdale et al., 2014). Co-SEA is a software application that is used by COMPASS researchers as a direct observation tool when auditing participating schools for the presence of healthy or unhealthy physical features (e.g. vending machines, exercise facilities, drinking fountains) (Leatherdale et al., 2014). The contextual data captured by Co-SEA may exist as photographs, free-text, or categorical ratings (see Appendix 1 for sample screenshots of the CO-SEA application). To assess the external school environment for health influencing factors, data is obtained annually from the CanMap Route Logistics (CMRL) spatial information database and the Enhanced Points of Interest (EPOI) data resource (Leatherdale et al., 2014). By combining land-use and street network data from CMRL with opportunity structure location data from EPOI (e.g. presence of fast food outlets, tobacco retailers, parks, recreation facilities etc.), COMPASS researchers can remotely evaluate the physical environment that surrounds participating schools in terms of impact on student health (Leatherdale et al., 2014).

Notably, many participating schools are purposefully sampled (i.e. a non-random convenience sample) and as a result, COMPASS data is not extensible for use in population-level statistics (Leatherdale et al., 2014). English speaking secondary schools, with Grades 9 to 12 and more than 100 students per grade, are considered eligible for COMPASS and are approached by COMPASS researchers following school board approval (Thompson-Haile & Leatherdale, 2013b). No effort is made to obtain a jurisdictionally representative sample (either provincially or nationally) and years of participation in the study varies by school (Thompson-Haile & Leatherdale, 2013b). Recruitment coordinators collaborate with participating schools to select data collection dates and adapt data collection procedures as needed (Thompson-Haile & Leatherdale, 2013b). Whether COMPASS data is longitudinal or cross-sectional is partially dependent on whether a school participates in the study for more than one year and the grade/age of a participating student. For example, cross-sectional data is generated from cohorts of Grade 12 students that graduate after completing the COMPASS C_q once, whereas younger student cohorts may have the opportunity to complete C_q more than once, producing longitudinal data.

COMPASS research involves youth under the age of 18 and, as a result, parental/guardian consent is required for participation. Active-information passive-consent protocols have been approved by the Canadian Institute of Health Research and are useful for achieving high participation rates and reducing sampling bias while preserving student confidentiality (Thompson-Haile, Bredin & Leatherdale, 2013). The protocol provides parents with pamphlets that detail important COMPASS research and contact information (recruitment coordinator email and phone number) should they wish to withdraw their child from the study. Eligible students whose parents do not contact the recruitment coordinator within the two-week time frame provided are considered participants and complete the C_q questionnaire (Thompson-Haile, Bredin & Leatherdale, 2013). Students are permitted to withdraw participation at any point during the consent process or data collection period (Leatherdale et al., 2014).

Ensuring that the C_q is completed and submitted anonymously is another critical component of COMPASS data collection but introduces several technical challenges. Linking C_q responses from the same student across multiple years of participation is challenging, but necessary if longitudinal data is to be obtained (Bredin & Leatherdale, 2014). The cover page of the C_q guides students through the process of generating a unique code that allows COMPASS researchers to track individual students throughout the study (Bredin & Leatherdale, 2014). Self-generated identification codes for anonymizing questionnaire data have been used elsewhere in longitudinal studies and can provide highly accurate matching rates depending on the number and types of questions used to generate the unique pin (Kearney, Hopkins, Mauss, & Weisheit, 1984; Bredin & Leatherdale, 2014). COMPASS data linkage validation studies estimate the matching rate for questionnaires completed one year apart is well above 90% (Bredin & Leatherdale, 2014). Additional measures used to preserve anonymity include providing guidelines for the educators administering the questionnaires and envelopes to students so they may personally seal their completed C_q before submission (University of Waterloo, 2020b). The passive-consent process also provides anonymity as only the names of students who have been withdrawn from participation are recorded (a 'no permission' list), and no participant list is ever developed (Thompson-Haile & Leatherdale, 2013a).

Since COMPASS research is used to evaluate school policies/programs and identify general trends in youth health behaviour, the responses of individual students are not used in analysis or reporting of COMPASS data (University of Waterloo, 2020b). Completed questionnaires (paper based) are stored at the University of Waterloo for seven years while electronic data (devoid of personal information) is maintained indefinitely on a secure server (University of Waterloo, 2020b). Professor Scott Leatherdale, the principal investigator, maintains ownership of COMPASS data and is responsible for reviewing and approving COMPASS data usage applications (University of Waterloo, 2020c). Following approval, data usage agreement forms must be signed before data is shared via encrypted memory stick or secure file transfer (University of Waterloo, 2020c).

With the support of ML techniques, COMPASS Vision provides users with interpretable visualizations of underlying trends and associations within the COMPASS dataset. All visualizations and models were generated from student-level COMPASS data gathered in Year 7 (2018-2019) of the study. The C_q results for 75,000 students from 134 participating schools was exchanged using Sendit, The University of Waterloo's secure file transfer system. To maximize COMPASS Vision data exploration capabilities and optimize the underlying ML algorithms, a full set of student-level COMPASS variables was desired and no specific COMPASS variables were requested during the application process. Visualizations and ML models could likely be improved if historical COMPASS data (from previous years) were made available. As the COMPASS dataset continues to grow, the ML models applied in COMPASS Vision could be re-trained or improved to provide additional insight.

Architectural Components of COMPASS Vision

The primary aim of COMPASS Vision is to present users with meaningful information related to the COMPASS study. To do this, architectural components from data visualization and ML are combined into one user interface, and a summary of the implementation pipeline is presented in Figure 1. This section provides implementation details for the map visualizations, the alternating least squares (ALS) recommender system, and the Decision Tree (DT), Random Forest (RF), and Gradient Boosted Decision Tree (GBT) methods to predicting important student health factors.

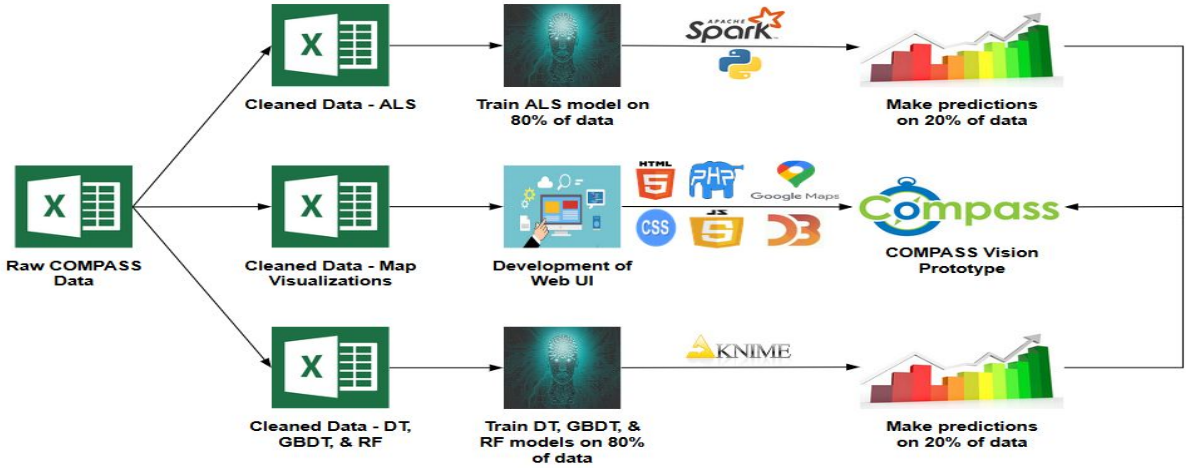

Raw COMPASS Data
Cleaned Data - ALS
Train ALS model on 80% of data
Spark
Make predictions on 20% of data
Cleaned Data - Map Visualizations
Development of Web UI
HTML
CSS
Google Maps
COMPASS Vision Prototype
Cleaned Data - DT, GBDT, & RF
Train DT, GBDT, & RF models on 80% of data
KNIME
Make predictions on 20% of data

Figure 1. COMPASS Vision implementation pipeline.

Data Cleaning

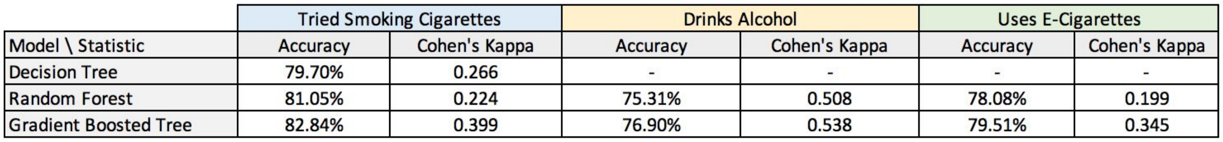
In preparation for machine learning and data analytics, a number of data preparation and cleaning steps were taken. First, the excel data was formatted in a new sheet concatenating data with its description, answer types, and a new short form and comprehensive name for easier interpretability and use. For example the short form of a question such as "MSNSLFB1" with a full form of "Choose the answer that best describes how you feel. Overall, I have a lot to be proud of" is very hard to work with, as in many programs only one can be chosen as a header. Therefore we created new headers, in this case for example we chose "Have a lot to be proud of". This way, when working with the data in something such as KNIME or within excel we could quickly work with features and interpret our findings. Next, we began removing features that added noise, such as School ID, Scan ID, minutes of exercise on specific days whose value we believe can be obtained from the simple summation average etc. This way our models will have less noise to work through, and won't attempt to learn on features that offer no insights, or value by our standards. Lastly, we converted specific variables to categorical or numeric based on their representation in the dataset. For example, weight and BMI were converted to numeric because of their ordinal nature, while many other responses that were answered with a number were actually representative of a mapping from a number to an answer where order has no importance. These are categorical variables and were treated as such. Because tree based classifiers were implemented, we did not have to use all numeric inputs and therefore did not one-hot encode these variables, but labelled them as string inputs.

A significant amount of data cleaning was also necessary to implement the ALS model. Data cleaning began by identifying the most interesting questions from C_q and placing them, along with all the student responses, into a reduced excel spreadsheet. Then, a program was written in Java to convert the reduced dataset into the format needed by the ALS model. The Apache Spark ALS model takes in three columns of numeric data as input: student id, COMPASS variable id, and COMPASS variable rating. Thus, for each student in the reduced spreadsheet and for each question that they responded to, if they indicated that they had tried a particular COMPASS variable (e.g., had tried smoking cigarettes, drinking alcohol, suffered from depression, etc.), then their numeric response was normalized to the range of [1, 5], where a higher rating indicates a more significant use or suffering from a COMPASS variable. If a student indicated that they have not tried or experienced a particular COMPASS variable, then a rating of 0 was given. All student-variable pairs with non-zero ratings were then inserted into the final ALS dataset in the necessary three-column format. Since the aim of the ALS model is to recommend preferences to students of variables that they have not tried yet, any student-variable combinations with ratings of 0 were excluded from the ALS dataset. The final ALS dataset consists of 1,895,064 student-variable pairings and their corresponding ratings, with a total of 74 distinct COMPASS variables.

Decision Tree, Random Forest, and Gradient Boosted Trees

We utilized three machine learning algorithms in our analysis to determine their prediction capabilities on features in the dataset while also using them for their unique abilities to give insights on the dataset. First, we used a simple decision tree to get an initial understanding of the prediction capacity of a feature, and because of its high interpretability and visualization abilities. Secondly, we used a random forest, which is a strong ensemble predictor made of multiple decision trees that are created on a subset of records and features. Random forests are very useful because they also act as a feature selector. Due to the fact that every tree makes splits based on node importance (such as the information gained) we can determine how important a feature is by the number of times it is used as a node within one of the decision trees. Lastly, we used gradient boosted trees which are an ensemble of decision trees that learn sequentially by changing the weights of importance of records that are more poorly predicted. GBT's are less interpretable than our other methods but have proven to be a very strong predictor. During each prediction problem we removed features related to the variable of interest we were trying to classify. For example for many of the prediction problems we were focused on the use of narcotics, such as cigarettes and alcohol. In these problems we removed the students answers to other questions about these, as we felt they gave away too much information, and we wanted to find underlying factors that may make people use them instead of seeing the rather obvious correlation between drinking and smoking.

The first prediction problem we tackled was that of predicting students that smoke cigarettes. For this problem we found that we could achieve an accuracy of 82.84% and a Cohen's kappa of 0.3999 using GBTs. We are very happy with this result as given the confusion matrix in Figure 2, we can see that the predictor only guesses that the student smokes about one eighth of the time, but is right 66% of the time it predicts this. While this only accounts for about 40% of all smoking students, this is a very hard problem domain and we think this is a great result.

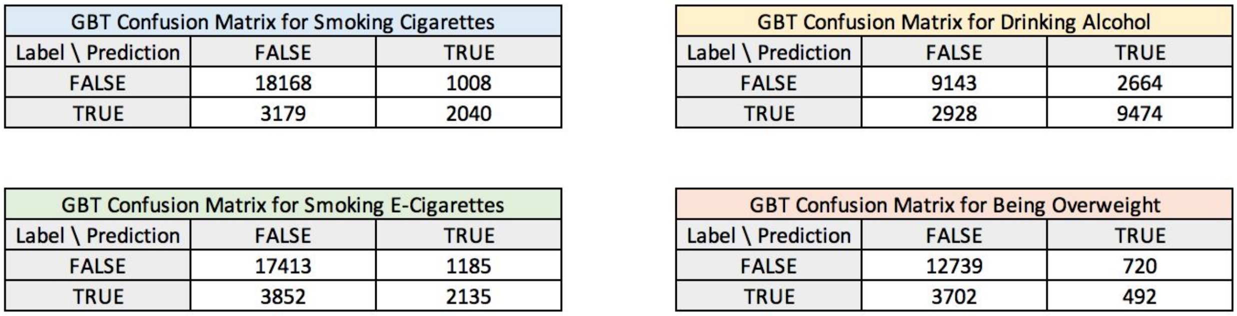


	Tried Smoking Cigarettes		Drinks Alcohol		Uses E-Cigarettes	
Model \ Statistic	Accuracy	Cohen's Kappa	Accuracy	Cohen's Kappa	Accuracy	Cohen's Kappa
Decision Tree	79.70%	0.266	-	-	-	-
Random Forest	81.05%	0.224	75.31%	0.508	78.08%	0.199
Gradient Boosted Tree	82.84%	0.399	76.90%	0.538	79.51%	0.345

Figure 2. COMPASS Feature Prediction Results

The second and third prediction problems we tacked were predicting students who currently drink alcohol regularly and those who use e-cigarettes. Once again, for both of these problems we found gradient boosted trees to be the most successful option, obtaining an accuracy and Cohen's kappa of 79.9% and 0.538, and 79.51% and 0.345 respectively. We found that the predictor for students using e-cigarettes had a very similar result as for that of predicting students smoking cigarettes. With almost the same number of predictions towards the 'true' class. On the other hand for drinking alcohol we saw a relatively equal distribution and accuracy of predictions between both classes. This is likely because of the more balanced nature of this problem.

Lastly we decided to predict whether a student was overweight or obese. This ended up being our hardest prediction task, originally obtaining ~60% accuracy using GBT's. To try and improve this result we utilized the Synthetic Minority Sampling Technique (SMOTE) which is a statistical technique for increasing the number of undersampled test cases. After utilizing SMOTE we were able to achieve an accuracy of 74.95% however this result mainly came from a decrease in overall guesses for the 'true' class. Realistically, the preference in a model may come down to a preference in false positives or true negatives.



GBT Confusion Matrix for Smoking Cigarettes		
Label \ Prediction	FALSE	TRUE
FALSE	18168	1008
TRUE	3179	2040

GBT Confusion Matrix for Drinking Alcohol		
Label \ Prediction	FALSE	TRUE
FALSE	9143	2664
TRUE	2928	9474

GBT Confusion Matrix for Smoking E-Cigarettes		
Label \ Prediction	FALSE	TRUE
FALSE	17413	1185
TRUE	3852	2135

GBT Confusion Matrix for Being Overweight		
Label \ Prediction	FALSE	TRUE
FALSE	12739	720
TRUE	3702	492

Figure 3. COMPASS Feature Prediction Confusion Matrices

Explicit CF with Alternating Least Squares (ALS)

A typical CF system analyzes relationships between users and interdependencies among items (e.g., movies, songs, products) to identify new user-item associations (Hu, Koren & Volinsky, 2008). These user-item associations are generated from information on the past behavior of users, such as the way they rate products or their transaction histories. In the context of

COMPASS Vision, COMPASS students correspond to users, COMPASS variables (e.g., survey questions related to whether a student has smoked cigarettes, drank alcohol, suffered from depression, etc.) correspond to items, and indicating the use of a COMPASS variable corresponds to providing a rating for an item. Similar to how Netflix recommends movies to a user that they have not seen yet and that they are likely to enjoy, CF can be applied to the COMPASS dataset to identify variables that a student may be at high risk of acquiring.

To develop a CF system for the COMPASS dataset, raw COMPASS data needs to be converted into a format that is interpretable by the Apache Spark ALS model. Given a table with three columns (student id, variable id, and numeric variable rating), the ALS model can generate predictions for the variables that students have not tried yet. To obtain this table, the raw COMPASS dataset was manually examined and a total of 76 interesting survey questions and their numeric ratings were extracted into a seperate file. Then, for each student in the reduced dataset, the values of each column (i.e., all COMPASS variables) were analyzed. If the value indicates the use of a COMPASS variable, then the numeric rating is normalized to a range of [1,5], and a new record is inserted into the ALS dataset. This record consists of a numeric student id, a numeric variable id, and the normalized rating of that variable. However, any variables that a student has not tried are excluded from the ALS dataset, and the aim of the ALS model is to predict the “preference” that a student may have towards these variables.

In the backend of the ALS model, the ALS dataset is initially represented as a COMPASS student-variable matrix R, in which $R_{ij} >= 1$ if COMPASS student p_i indicated the use of a COMPASS variable q_j, and $R_{ij} = 0$ otherwise. Then, the ALS algorithm is used (via Apache Spark’s machine learning library, *MLlib*) to decompose matrix M into two factor matrices, Q and P, such that $R \approx QP^T = R'$. The rows of Q and P represent COMPASS variables and COMPASS students in k latent dimensions, respectively (we use the *MLlib* default value of $k = 10$). This means that $R' = QP^T$ captures the similarity of each COMPASS student-variable paring with respect to the k latent dimensions. Thus, matrix R' can be used to predict (or “recommend”) variables that a student may be at risk of acquiring.

After training on 80% of the data and testing on the remaining 20%, the ALS model was capable of generating predictions while maintaining a root mean squared error (RMSE) of 1.293. This means that the predictions of the model were on average within 1.293 points (25.9%) of the true ratings given by the students. Moreover, the ALS model is capable of generating a ranked list of variables for each student in which the highest ranked variable indicates the variable that the student is most likely to enjoy or prefer. The ALS model can also apply a similar logic to the COMPASS variables and generate a ranked list of students for each variable.

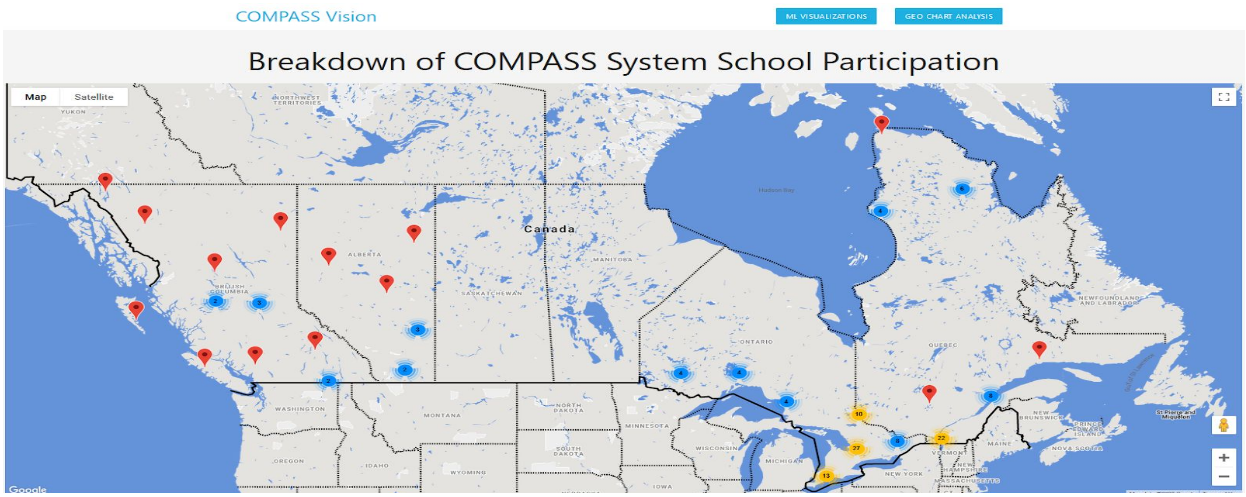


Figure 4. Homepage of COMPASS Vision Web interface.

COMPASS Vision Interactive Visualizations

Upon loading the COMPASS Vision homepage, users are presented with the map visualization shown in Figure 4. This map was created using the *Google Maps JavaScript API* and supports a variety of interactive functionalities. For example, users can zoom in and out on areas of interest, and they can drag the map to navigate throughout the map. The map also uses the *MarkerClustererPlus* library to create and manage per-zoom-level clusters for large amounts of markers. Currently, red markers represent different schools participating in the COMPASS study, and these markers are dynamically grouped together as a user zooms out on the map. Similarly, zooming in on the map gradually begins to reveal more markers. Clicking on a cluster drills down into the cluster by dynamically zooming in and focusing the map on the center of the cluster. Clicking on a marker presents users with an information window displaying the province names, school id, and number of students from that school. The map also provides users with options to view the map in satellite and street view modes.

Under the ‘Geo Chart Analysis’ tab of the COMPASS Vision Web UI (Figure 5), users are presented with a map that shows the percentage of students in each province that are susceptible to smoking cigarettes. The color of each province corresponds to the percentage of students being represented in that province, where a darker shade of blue indicates a larger percentage of students, and a lighter shade of blue indicates a smaller percentage of students. A color scale is provided in the bottom left corner of the map to provide users with the symbolic tie between the numeric values that the different shades of color represent. When a user hovers over a province, a tooltip is displayed that shows the province name and the percentage of students that are at high risk of smoking cigarettes. Moreover, a marker appears on the color scale that points to the shade of color and percentage of students that is being depicted for the current province that is being hovered over.

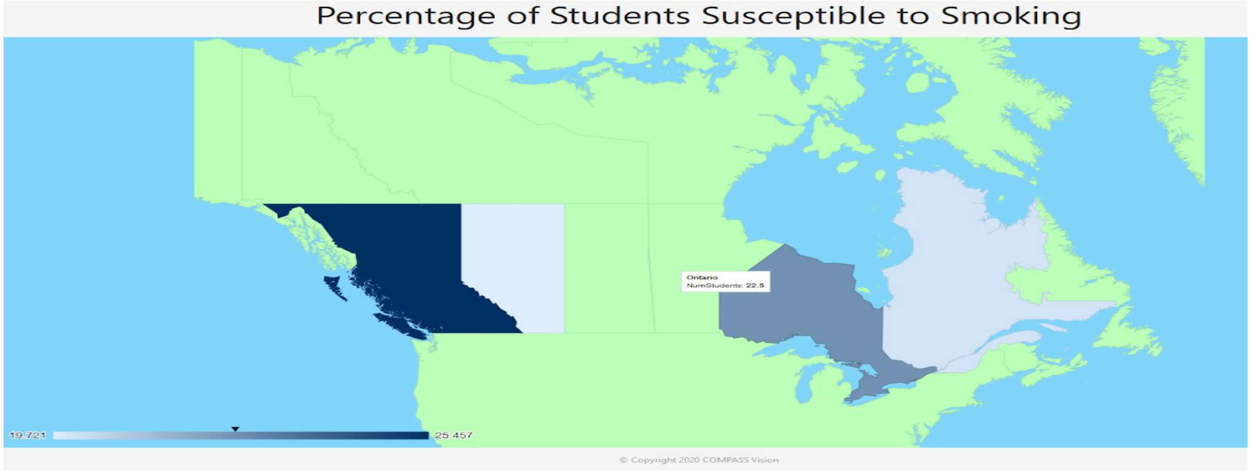

Percentage of Students Susceptible to Smoking
Ontario
NumStudents: 22.5
19.721
25.457
© Copyright 2020 COMPASS Vision

Figure 5. Geo Chart showing important evidence of Canadian youth smoking susceptibility.

The ‘ML Visualizations’ page of the COMPASS Vision Web UI aims to visualize important results and predictions made by the COMPASS Vision ML models. The current prototype displays interactive bar graphs that visualize the accuracy and Cohen’s Kappa scores of the DT, RF, and GBT models for predicting cigarette smoking, e-cigarette smoking, alcohol drinking, and obesity. These graphs were implemented using D3.js, a JavaScript library that enables the binding of arbitrary data to a Document Object Model (DOM), and then applies data-driven transformations to the document to create interactive visualizations of data (Bostock, 2019). Figure 6 shows that numbers are displayed on each bar representing their respective y-axis values, and hovering over a bar displays a guideline at height of the current bar that spans horizontally across the graph. Moreover, when hovering over one bar, the numbers in all other bars are updated to show the difference in value (either positive or negative) compared to the bar that is being hovered over. This helps users easily compare the differences between the results of the ML models in ways that static graphs are not capable of doing.

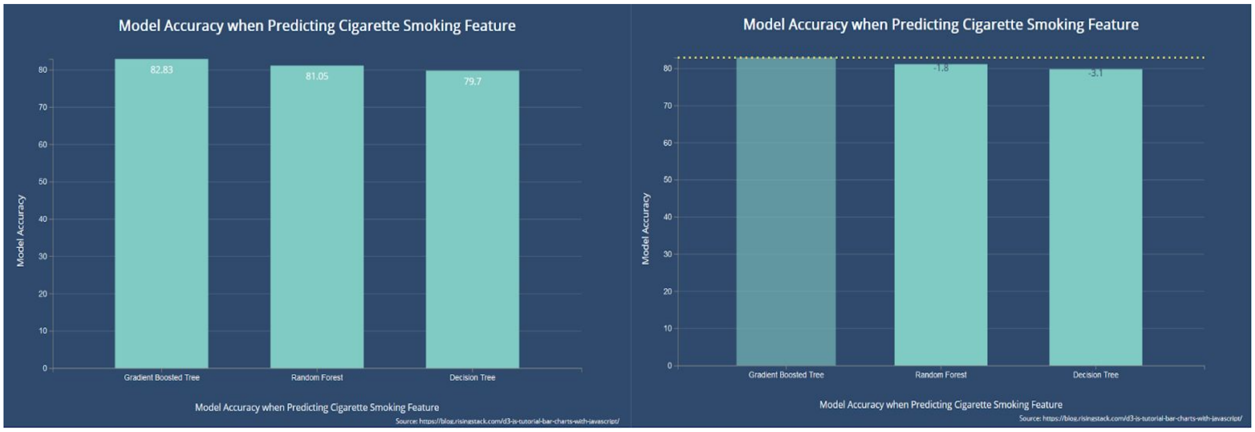


Figure 6. Visualization of GBT, RF, and DT model accuracies when classifying students as either cigarette smokers or non-smokers.

Evaluation

This section outlines the evaluation of the initial COMPASS Vision prototype. Based on a team analysis and feedback given during our project demonstration, we first detail the major strengths of COMPASS Vision. We then identify the limitations of the current prototype and discuss methods for mitigating these limitations with future research.

COMPASS Vision Strengths

One of the strengths of COMPASS Vision comes from the use of ALS, DT, RF, and GBT approaches to predicting important student health factors. The ALS approach focused on predicting the likelihood of students enjoying non-acquired health factors based on their pre-existing health factors, while the DT, RF, and GBT approaches used interesting features of students to classify them into either one of two groups (e.g., smoker or non-smoker). All four models were evaluated using popular ML metrics to measure their effectiveness, and initial results show that most of the models had relatively accurate predictions. Moreover, we were able to gain some interesting insights into the data that would not have been possible without the use of these ML methods. For example, when predicting students who smoked, we were able to view the features used most within the random forest and found that the most important feature was the number of days in the week that the student drank energy drinks. While this feature did not stick out based on a simple correlation matrix, it became significantly important in prediction when mixed with other features such as the number of classes skipped in the last four weeks. When predicting students who are drinking alcohol regularly we found that how the student travels to and from school was particularly important. Perhaps because they are more likely to drink with their friends if they can walk home instead of driving. It is insights like these that we would not be able to find otherwise that make using ML so interesting and valuable in this context. While more analysis would need to be done before taking action, it is insights like these that lead to knowledge translation and exchange activities, and subsequently intervention activities such as effective policy changes to keep kids safe and healthy.

Another strength of COMPASS Vision includes the use of interactive and dynamic visualizations that help users gain insights into COMPASS data. The homepage of the COMPASS Vision interface features a dynamic map that provides a visual summary of participating schools in the COMPASS study, and this provides a backbone for the future visualization of COMPASS analytics and ML model predictions. The initial prototype also features an easy to learn user interface, provides good usability, and the code is scalable to a large number of data points. Moreover, a custom *Mapster* library was implemented using the help of a course on the Google Maps Javascript API (Envato Tuts+, 2017). Mapster is valuable because it provides a framework that can easily be customized to support different tasks throughout the development of future COMPASS map visualizations. Some important functions include adding markers to the map, attaching events to markers (e.g., show a tooltip when a marker is clicked on), and removing markers based on a given key. Since the Google Maps JavaScript API does not provide a method for storing and keeping track of markers, Mapster includes a custom *List* data structure that manages this important functionality. Customizing different map options (e.g., min and max zoom levels, locations of buttons, style, etc.) is also easy to do, and making additions to future COMPASS Vision prototypes will be more efficient because of the Mapster library.

Finally, COMPASS Vision undertook a challenging data cleaning process, and some of the work that we did may be useful to future COMPASS researchers. For example, the raw COMPASS data was formatted in a new excel sheet that concatenated data with its description, answer types, and a new short form and comprehensive name for easier interpretability and use. One of the most challenging aspects of this project was cleaning the raw data, and we plan to submit this newly formatted data to the COMPASS research team. The hope is that by having this clean dataset, COMPASS researchers can spend less time processing data and more time analyzing it. We also plan to provide all necessary code to COMPASS researchers so they can learn how the machine learning models were implemented, and perhaps they can customize them for their individual research needs. The same goes for the code for the Web interface and all of its visualizations, which was designed in a way to make it flexible across different use cases.

COMPASS Vision Limitations

Although COMPASS Vision provides a solid foundation for the visualization of COMPASS data, the current tool is limited in its ability to provide significant evidence that could help in the decision making process of Canadian policy makers. For example, the map that summarizes COMPASS student participation (Figure 4) does not provide any meaningful information related to the health of students. However, future prototypes plan to enhance this map by visualizing the most “preferred” variables of each student as predicted by the ALS model. First, we plan to train a new ALS model using the entire dataset (rather than 80%) and extract the most preferred variable of each student. Then, all school markers will be given custom school icons and all students will be given custom student icons based on the ALS model’s prediction of their most preferred variable. Each school marker will be plotted on its appropriate location on the map, and the student markers will be placed into clusters around their respective schools. Each cluster represents a group of students from a particular school who may be at risk of acquiring a COMPASS variable, and clicking on a cluster will drill down to reveal all the student markers for that cluster. Users are then able to click on individual student markers to reveal more student-specific information (e.g., their top 5 recommended COMPASS variables). This type of visualization would provide powerful insights into COMPASS student data, and it would help direct the appropriate student support programs to the schools where they are needed most.

The ‘Geo Chart Analysis’ section of the COMPASS Vision interface is also limited in its data exploration and decision making support capabilities. The Geo Chart only shows information for one interesting query (smoking susceptibility by province), and future prototypes will expand this page to support multiple queries. After cleaning and analyzing the raw COMPASS data, we were able to generate a list of 20 interesting queries that group results based on provinces. Examples of these queries include the percentages of students who have drank by a certain age, tried marijuana, been bullied (verbally, physically, or electronically), and many more. Updating the current Geo Chart to support the visualization of all 20 queries would significantly increase the value of the COMPASS Vision interface as a whole. Adding different filtering options would also enhance data exploration capabilities (e.g., click on 'smoking' and 'grade 9' checkbox to update the map to show the percentage of grade 9 students in each province who have tried smoking cigarettes). We argue that these new functionalities will provide interested stakeholders with credible evidence related to important health factors of Canadian students.

The 'ML Visualizations' page of the initial COMPASS Vision prototype features dynamic bar graphs that visualize the evaluation results of our models. Although these graphs are better than a textual summary or static image of a graph, they fail to provide meaningful insights into the predictions made by the models. Thus, we plan to visualize the actual predictions of our models in future prototypes to help identify complex associations between COMPASS variables. One idea involves using the model-agnostic interpretability techniques of the SHAP framework to assign Shapley values to model features (Lundberg & Lee, 2017). SHAP values represent the degree of change in model output based on a feature's inclusion. By calculating a SHAP value for every feature in every sample, a range of SHAP values can be obtained and plotted for each feature to assess its contributions to predictive ability. We plan to incorporate SHAP into COMPASS Vision because the interpretability of SHAP values makes the framework well-suited to the field of medicine (and possibly population health). Moreover, previous research has applied SHAP to interpret predictions of hypoxemia during surgery and detection of acute intracranial hemorrhaging (Stiglic et al., 2020). Another idea involves using the Python machine learning library *dtreeviz* to visualize the results of the DT, RF, and GBT approaches of COMPASS Vision. Parr and Grover (2019) claim that the visualizations from *dtreeviz* illustrate how the feature space of a decision tree is split up at decision nodes, and this is the critical operation performed during decision tree model training.

Other limitations of COMPASS Vision stem from the many complexities related to the COMPASS source dataset. The COMPASS dataset is very complex and future COMPASS datasets will likely give rise to the same data cleaning challenges that were faced during the development of COMPASS Vision. For example, the script that was used to convert the COMPASS data into ALS format has many elements that are hardcoded and would potentially need to be modified to be compatible with evolving datasets. Standardization of the COMPASS dataset needs to be strongly considered to reduce the complexity of the data cleaning process for future COMPASS researchers. The COMPASS Vision prototype did not have access to student data over multiple years of the study, and this limited the predictive powers of our ML models. It would be interesting and useful to make predictions for a grade 9 student and then see if these predictions are accurate by the time they graduate high school. Future work hopes to gain access to this data and experiment with different prediction techniques as they apply to risk factor temporality.

COMPASS research methods also introduce limitations. Unfortunately, the ability of COMPASS Vision to provide multiple levels of granularity for geographic visualizations is hindered by the relatively small number of participating schools that are distributed over a large geographic range. Since the list of participating schools is confidential, precise locations cannot be indicated in any visualizations (University of Waterloo, 2020b). Additional granularity increases transparency, and although initial proposals for COMPASS Vision called for 'zooming capabilities' that would provide users with county- or municipality-specific statistics, this feature was largely unfeasible. With the exception of Ontario, participating schools per province are too few to provide this level of detail while preserving school anonymity. Grouping COMPASS data by province is also troublesome given the fact that schools are non-randomly selected for enrollment in COMPASS research (Leatherdale et al., 2014). COMPASS Vision users must take caution when interpreting visualizations that appear to show provincial-level statistics, since the sample of schools is not truly representative and external validity cannot be guaranteed.

Finally, the COMPASS dataset has a lot of human variability involved, and this poses a challenge for making accurate predictions. For instance, certain students may have different interpretations of the question responses on the COMPASS student survey (e.g., strongly agree vs. agree). Another possibility is that students may lie on certain questions because they are not comfortable with giving an honest answer. Some measures have been taken to validate questionnaire responses and mitigate response bias. The C_q questionnaire is brief and uses research-validated, self-report core outcome measures that have been specifically designed for high-school aged youth and can be benchmarked against existing public health guidelines (Bredin & Leatherdale, 2014). Additionally, the questionnaire has been updated several times to eliminate questions that were problematic or frequently skipped by student participants (Reel et al., 2019). The student data linkage process provides some ability to identify conflicting questionnaire responses that originate from a single participant but also introduces an imperfect linkage rate that generates additional error. Regardless, the consensus among stakeholders remains that COMPASS methodologies are sufficiently robust given the delicate balance of data accuracy and participant anonymity in longitudinal studies that concern youth (Battista, Qian, Bredin & Leatherdale, 2019).

Conclusion

The COMPASS study is a unique and insightful research endeavour that continues to contribute to our understanding of youth health behaviour and outcomes in Canada as they relate to school programs, policies and built environment characteristics. COMPASS Vision serves as a valuable asset to COMPASS research and will advance data exploration and knowledge translation activities with enhanced data visualizations and ML techniques. The technology consists of interactive data visualizations and ALS, DT, RF, and GBT ML models. The COMPASS Vision interactive visualizations provide users with summary statistics, zooming functionalities and data exploration opportunities. Once fully developed, visualizations will illustrate regional variability in youth health behaviour at multiple levels of granularity based on user-selected variables. The Geo Chart will also provide visualizations of ML model predictions. The ALS model has been trained to predict student 'preferences' for certain COMPASS variables and could likely be used to identify and define 'at-risk' groups. DT, RF and GBT were used to predict whether a student possesses a certain risk factor given their responses to other questionnaire items. These methods can also be used to identify trends among COMPASS variables and it was demonstrated that meaningful associations may exist between variables that are presumed to be disparate (in this case energy drink consumption per week and cigarette smoking). It is anticipated that future versions of COMPASS Vision will better combine the data visualization and ML aspects of what has been developed. Ideally, historical and future COMPASS data will be used to improve the models so they may illustrate temporality. Additional school-level data could also be included so that policy and physical environment differences can be displayed visually. It would also be beneficial to provide users with ML predictions and output as they interact with the visualizations and select variables of interest. Overall, it is believed that COMPASS Vision will assist COMPASS stakeholders in finding novel insights and predicting patterns or risk factors in youth health trends, while also serving as a form of evidence towards shaping youth program and policy decisions in Canadian high schools.

References

Battista, K., Qian, W., Bredin, C., & Leatherdale, S.T. (2019). *Student Data Linkage over Multiple Years.* Technical Report Series. 6(3): Waterloo, Ontario: University of Waterloo. Retrieved from:

https://uwaterloo.ca/compass-system/student-data-linkage-over-multiple-years

Bostock, M. (2019). D3.js - Data-Driven Documents. Retrieved from: https://d3js.org/

Bredin, C. & Leatherdale, S.T. (2014). Development of the COMPASS Student Questionnaire. COMPASS Technical Report Series. 2(2). Waterloo, Ontario: University of Waterloo. Retrieved from:https://uwaterloo.ca/compass-system/publications/development-compass-student-questionnaire#references

Envato Tuts+. (2017, July 17). *Custom Interactive Maps With the Google Maps API 01 Introduction* [Video]. YouTube. Retrieved from: https://www.youtube.com/watch?v=_L4IQoAWD9E&list=PLgGbWId6zgaXFR4SW_3qJ55cxmEqRNIzx

Huang, M.J., Chen, M.Y., & Lee, S.C. (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert systems with applications, 32*(3), 856-867. Retrieved from: https://doi.org/10.1016/j.eswa.2006.01.038

Hu, Y., Koren, Y., & Volinsky, C. (2008). Collaborative Filtering for Implicit Feedback Datasets. *2008 Eighth IEEE International Conference on Data Mining*. doi: 10.1109/icdm.2008.22

Kearney, K.A., Hopkins, R.H., Mauss, A.L., & Weisheit, R.A. (1984). Self-generated identification codes for anonymous collection of longitudinal questionnaire data. *Public Opinion Quarterly*, *48*(1B), 370-378. Retrieved from: https://academic.oup.com/poq/article-pdf/48/1B/370/5443196/48-1B-370.pdf

Lau, E.Y., Faulkner, G., Qian, W., & Leatherdale, S.T. (2016). Longitudinal associations of parental and peer influences with physical activity during adolescence: findings from the COMPASS study. *Health promotion and chronic disease prevention in Canada: research, policy and practice*, *36*(11), 235. Retrieved from:https://doi.org/10.24095/hpcdp.36.11.01

Leatherdale, S.T. (2009). Evaluating school-based tobacco control programs and policies: an opportunity gained and many opportunities lost. *The Canadian Journal of Program Evaluation*, *24*(3), 89. Retrieved from: https://search.proquest.com/docview/1038938258?pq-origsite=gscholar

Leatherdale, S.T., Brown, K.S., Carson, V. et al. (2014). 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 14*, 331. Retrieved from: https://doi.org/10.1186/1471-2458-14-331

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Retrieved from https://arxiv.org/abs/1705.07874

Mooney, S.J., & Pejaver, V. (2018). Big data in public health: terminology, machine learning, and privacy. *Annual review of public health*, *39*, 95-112. Retrieved from: https://www.annualreviews.org/doi/full/10.1146/annurev-publhealth-040617-014208

Parr, T., & Grover, P. (2019). How to visualize decision tree. Retrieved from: https://explained.ai/decision-tree-viz/

Piccardi, T., Catasta, M., Zia, L., & West, R. (2018). Structuring Wikipedia Articles with Section Recommendations. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR 18*. doi: 10.1145/3209978.3209984

Reel, B., Bredin, C. & Leatherdale. S.T. (2018). *COMPASS Year 5 and 6 School Recruitment and Retention:* Technical Report Series. 5(1): Waterloo, Ontario: University of Waterloo. Retrieved from: https://uwaterloo.ca/compass-system/publications/compass-year-5-and-6-school-recruitment-and-retention

Reel, B., Battista, K., Bredin, C., & Leatherdale, S.T. (2019). COMPASS Questionnaire Changes from Year 1 to Year 7: Technical Report Series. 6(1): Waterloo, Ontario: University of Waterloo. Retrieved from: https://uwaterloo.ca/compass-system/publications/compass-questionnaire-changes-year-1-year-7

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning based prediction models in healthcare. *arXivpreprintarXiv:2002.08596*.

Thompson-Haile, A. & Leatherdale, S.T. (2013a). Student-level Data Collection Procedures. COMPASS Technical Report Series. 1(5). Waterloo, Ontario: University of Waterloo. Retrieved from: www.compass.uwaterloo.ca.

Thompson-Haile, A. & Leatherdale, S.T. (2013b). School Board and School Recruitment Procedures. COMPASS Technical Report Series. 1(3). Waterloo, Ontario: University of Waterloo. Retrieved from: https://uwaterloo.ca/compass-system/publications/school-board-and-school-recruitment-procedures

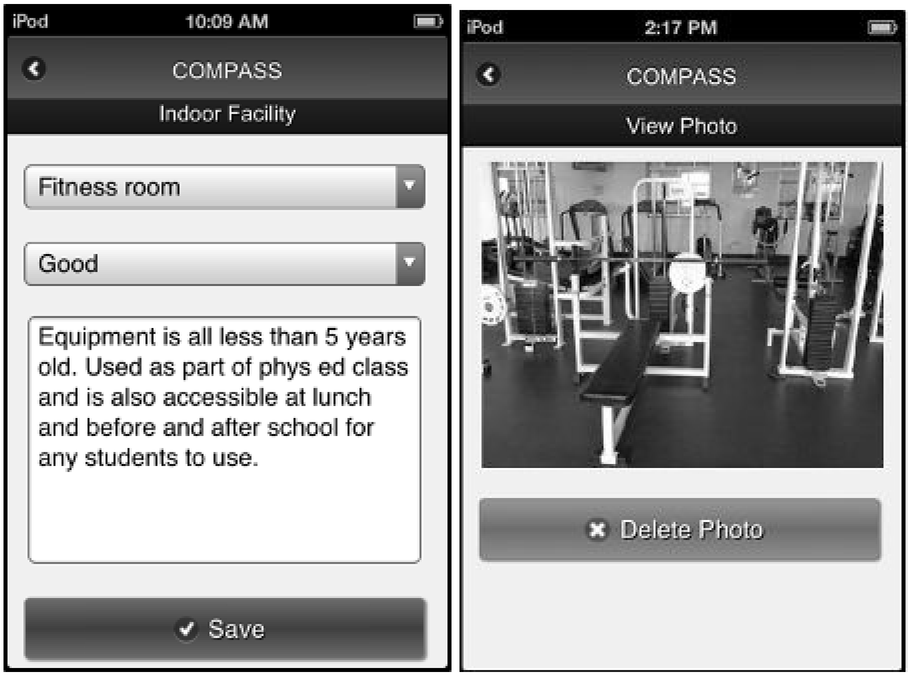
Thompson-Haile, A., Bredin, C. & Leatherdale, S.T. (2013). Rationale for using an Active-Information Passive-Consent Permission Protocol in COMPASS. COMPASS Technical Report Series. 1(6). Waterloo, Ontario: University of Waterloo. Retrieved from: https://uwaterloo.ca/compass-system/publications/rationale-using-active-information-passive-consent

University of Waterloo. (2020a). About the COMPASS system. Waterloo, Ontario: University of Waterloo. Retrieved from: https://uwaterloo.ca/compass-system/about

University of Waterloo. (2020b). COMPASS System: Confidentiality. Waterloo, Ontario: University of Waterloo. Retrieved from: https://uwaterloo.ca/compass-system/about/confidentiality

University of Waterloo. (2020c). COMPASS System: Information for researchers. Waterloo, Ontario: University of Waterloo. Retrieved from: https://uwaterloo.ca/compass-system/information-researchers

Appendix A: Sample screenshots of the Co-SEA application being used to record information about a school weight room (Leatherdale et al., 2014).



Fitness room

Good

Equipment is all less than 5 years old. Used as part of phys ed class and is also accessible at lunch and before and after school for any students to use.