Machine Learning in Prediction of Mental Health during COVID-19 among Canadians

Vu Hoang Anh Nguyen
Thompson Rivers University
805 TRU Way
Kamloops, BC, Canada
vnguyen19@mytru.ca

Quang Nguyen
Thompson Rivers University
805 TRU Way
Kamloops, BC, Canada
nguyenq19@mytru.ca

Cuong Phan
Thompson Rivers University
805 TRU Way
Kamloops, BC, Canada
phanc19@mytru.ca

Shaojie Ma
Thompson Rivers University
805 TRU Way
Kamloops, BC, Canada
shaojie-ma@mytru.ca

Ziqing Wang
Thompson Rivers University
805 TRU Way
Kamloops, BC, Canada
wangz133@mytru.ca

ABSTRACT
The coronavirus disease 2019 (COVID-19) pandemic and its immediate aftermath present a serious threat to the mental health of Canadian residents. In this paper, we utilize survey data from Canadian Perspectives Survey Series 2: Monitoring the Effects of COVID-19 and apply several statistical and machine learning models and techniques such as Chi-Squared Test, Bayesian Networks, Synthetic Minority Oversampling Technique, Random Forests, Support Vector Machines, Extreme Gradient Boosting, and Naive Bayes to analyze the impacts the COVID-19 pandemic has had on the Canadians' mental health.

CCS Concepts
I.2.4 [**Artificial Intelligence**] Knowledge Representation Formalisms and Methods

Keywords
COVID-19; Mental Health; Machine Learning; Feature Selection

1. INTRODUCTION
After 2 years of initial reporting, the coronavirus pandemic has raged across the world. Besides the obvious impact on physical health, the pandemic is likely to negatively affect the mental health and well-being of Canadians. In tandem with living amidst a global pandemic, stress, social isolation, and the associated financial crisis may result in significant adverse mental health effects. In this paper, we focus on learning a ranked list of factors that could indicate a predisposition to a mental disorder during the COVID-19 pandemic and explore how these predictors might interact in identifying individuals who are at a greater risk of psychological distress.

To begin the analysis, we select variable MH_05 as our target variable. This variable indicates how respondents describe their mental health since the outbreak of COVID-19 on a scale of 1 to 5. On this scale, 1 is excellent, 2 is very good, 3 is good, 4 is fair, and 5 is poor. The remainder of this paper is structured as follows: In the following section, we discuss the methodology, describing the experimental framework used to find the top predictors of an individual's mental health. Section 3 presents the procedure of experiments and computational results. Section 4 discusses and analyzes the top predictors of Canadians' emotional well-being obtained by Machine Learning methods. Section 5 concludes the paper by summarizing our overall findings and specifying future work to be done.

2. METHODOLOGY
2.1 Data Set
This investigation focuses on data from the second wave of our cross-sectional survey, "Monitoring the effects of COVID-19". We extract the data of 4,600 respondents, who answer a number of questions about their mental health, behaviour, and labour market activities. To summarize, the data set comprises 57 variables on demographics, mental health impacts, behaviours and health impacts, labour market impacts, and food security.

2.2 Data Preprocessing
Data preprocessing is an important step that helps improve the quality of data to extract meaningful insights from the data. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data pre-processing includes data learning, normalization, transformation, feature extraction and selection, etc [7]. Our target dataset, which is collected by Statistics Canada, is categorical and encoded. Therefore, we do not need to go through the normalization, transformation, or feature extraction process. However, it is important to identify and correctly handle the missing values in the data before any statistical or machine learning model is employed. Otherwise, inaccurate conclusions and inferences may be made from the data. A common approach for dealing with missing features involves imputation, the process of replacing missing data with substituted values. Multiple imputation is an approach that aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining results obtained from each of them. In other words, multiple imputation breaks imputation out into three steps: imputation (multiple times), analysis (staging how the results should be combined), and pooling (integrating the results into the final imputed matrix) [12]. A popular multiple imputation algorithm is called MICE (Multiple Imputation by Chained Equation), and a Python implementation thereof is available as part of the `fancyimpute` package [1].

Imbalance learning is also a challenging task. If the imbalanced data is not treated beforehand, then this will degrade the

performance of the classifier model. Resampling data is one of the most preferred approaches to deal with an imbalanced dataset. Synthetic Minority Oversampling Technique (SMOTE) is a data augmentation oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together [11].

2.3 Feature Selection

2.3.1 Chi-Squared Test

Since the problem is a classification problem, where the majority of variables are categorical, we use a statistical Chi-Squared test with a significance level of $\alpha = 0.05$ to determine whether the target variable, MH_05, is dependent or independent of the rest of the variables. The variables that are independent are considered as candidates for irrelevant features to the problem and they might be removed [10]. The `chi2` function in the `scikit-learn` library is an easy way to implement Chi-Squared Test.

2.3.2 Bayesian Network Analysis

Since mental health variables may have complex dependencies with potential confounding factors, mediation, and intercausal dependency, we extend our feature selection with Bayesian Network (BN) structure learning to improve the performance of the downstream analysis. A Bayesian Network is a representation of a joint probability distribution of a set of random variables with a possible mutual causal relationship [5]. The statistically equivalent signature (SES) algorithm is a method for feature selection inspired by the principles of constraint-based learning of Bayesian Networks. The SES algorithm is implemented in a homonym function included in the R package `MXM` [8].

2.4 Supervised Learning Models

Supervised Machine Learning (SML) is the search for algorithms that reason from externally supplied instances to produce general hypotheses, which then make predictions about future instances. Supervised classification is one of the tasks most frequently carried out by intelligent systems [9]. We use four supervised learning methods: Random Forests (RF), Support Vector Machines (SVMs), Extreme Gradient Boosting (XGBoost), and Naïve Bayes (NB). For all supervised machine learning models, we split the data into two subsets, called training set and test set. We train the model on the training set while the test set is held back from the algorithm. After we have found the optimal parameters of the model on the training set, we evaluate the trained model on the test set to find out how well the model performs on unseen data points. The algorithms are available in the `scikit-learn` Python machine learning library.

2.4.1 Random Forests

Random Forest is a commonly used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which comprises the construction of many "simple" decision trees in the training stage and the majority vote across them in the classification stage. Among other benefits, this voting strategy has the effect of correcting for the undesirable property of decision trees to overfit training data [3].

2.4.2 Support Vector Machines

Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. Support Vector machines can be defined as systems which use hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory [6].

2.4.3 Extreme Gradient Boosting

Extreme Gradient Boosting is a scalable and improved version of the gradient boosting algorithm designed for efficacy, computational speed and model performance. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion as other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function [4]

2.4.4 Naïve Bayes

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems, and it depends on the principle of Bayes' Theorem. The calculation of the probabilities for each hypothesis is simplified to make their calculation tractable. It assumes that the occurrence of a certain feature is independent of the occurrence of other features [2].

3. EXPERIMENTS RESULTS

We first target variable MH_05 (see Table 1), then fill in missing data by means of MICE and apply SMOTE to overcome class imbalances.

Table 1. Target Variable

MH_05: Perceived mental health				
1	2	3	4	5
Excellent	Very Good	Good	Fair	Poor

Next, we use the Chi-Squared test to find independent features and remove 26 variables. We also use Bayesian Network analysis to find irrelevant and redundant variables with intercausal dependency, removing 13 variables. After all the required steps for feature selection, we end up with 18 variables.

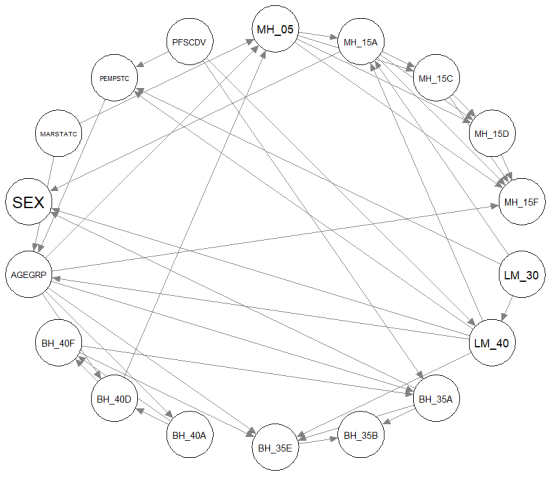

Figure 1. BN graph with selected variables

For supervised machine learning, we split the data into training-test sets (80%-20%) and train selected machine learning models (Random Forests, Support Vector Machines, Extreme Gradient Boosting, and Naïve Bayes) to find the most accurate and robust model. To evaluate how the models perform, we calculate the accuracy as the ratio of the number of correctly classified cases to the total of cases under evaluation.

- With Random Forests, the model obtains an accuracy of 82.07% on the test set. Figure 2 displays the feature importance scores of the RF model.
- The accuracy score of the Support Vector Machines model is 71.91% and feature importance scores are displayed in Figure 3.
- The Extreme Gradient Boosting accuracy score turns out to be 80.96% with feature importance scores displayed in Figure 4.
- Finally, the accuracy score of the Naïve Bayes model is 70.42%, and Figure 5 displays the feature importance scores of the model.

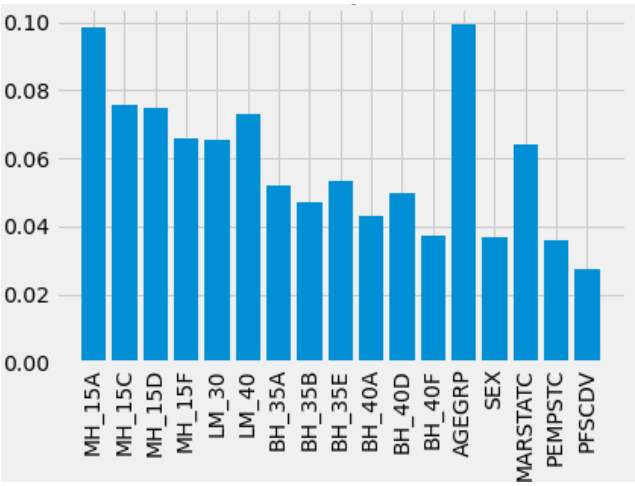

Figure 2. Feature scores of RF model

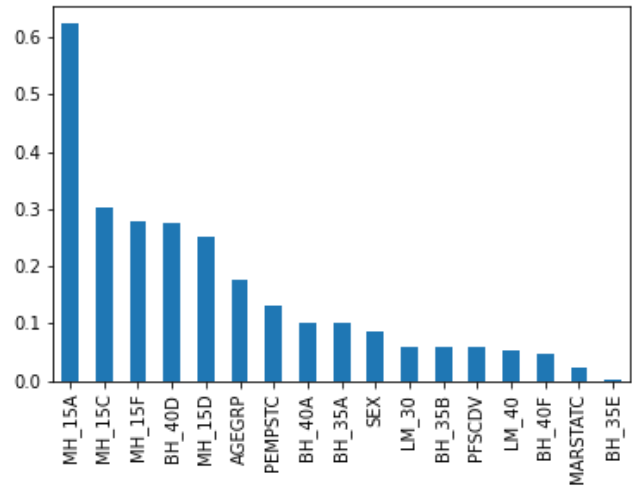

Figure 3. Feature scores of SVMs model

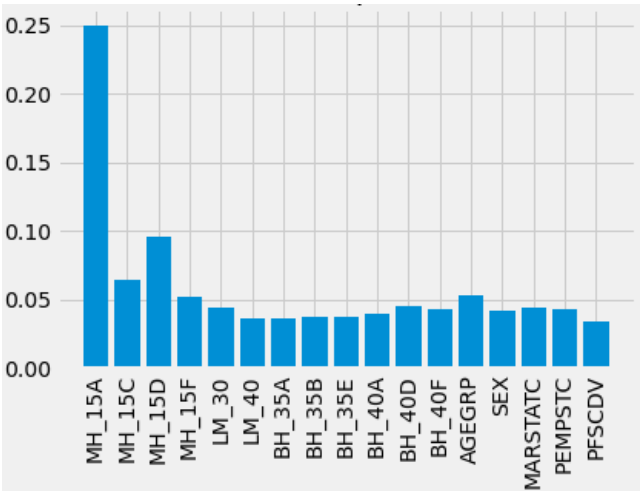

Figure 4. Feature scores of XGBoost model

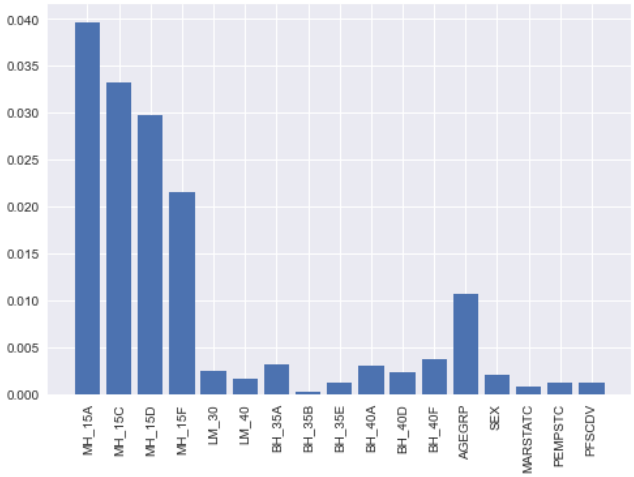

Figure 5. Feature scores of Naïve Bayes model

To get a better understanding of the relationship between the variables, a correlation heat map is provided in Figure 6.

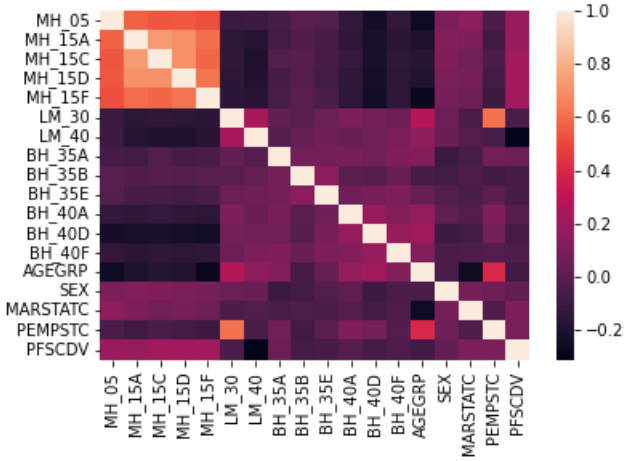

Figure 6. The correlation heat map for variables

4. Discussion

The most robust model is Random Forest with an accuracy of 82.07%, which means the feature importance scores of the model give well-grounded information about the top predictors. Extreme Gradient Boosting also has a satisfactory accuracy of 80.96%. Support Vector Machines and Naïve Bayes do an unsatisfactory prediction with the accuracy of 71.91% and 70.42% respectively.

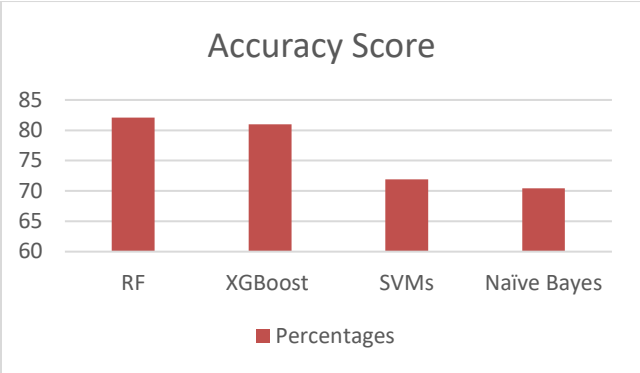

Figure 7. Accuracy of different models

Upon examination of the top predictors from the different models, we have identified the main variables which are highly predictive of the mental health of Canadians. Our analysis concludes that AGEGRP – the age group of respondents – is selected as the most important predictor of mental health by Random Forest, the strongest supervised machine learning model. Figure 5 shows that AGEGRP has a negative correlation with MH_05, indicating that younger Canadians tend to experience poorer mental health.

Furthermore, we have also found that the other top predictors of anxiety level identified by Random Forest are MH_15A (the frequency of feeling nervous, anxious or on edge), MH_15C (the frequency of worrying too much about different things), MH_15D (the frequency of trouble relaxing), and LM_40 (the ability to meet financial obligations or essential needs). MH_15A, MH_15C, and MH_15D have a positive correlation with the target variable when LM_40 has a negative one. Additionally, other models also select MH_15F (the frequency of being restless) and BH_40D (eating junk food or sweets) as important features. There is a positive correlation between MH_15F and MH_05 and a negative correlation between BH_40D and MH_05. Therefore, an individual who frequently feels nervous, worries too much, finds difficulty relaxing, loses financial ability, feels restless, and increasingly eats junk food tends to suffer from poor mental health.

5. Conclusion and Future Work

In this study, we present our analysis of the dataset “Monitoring the effects of COVID-19" collected by Statistics Canada. We utilize several machine learning and statistical models to analyze the data obtained. Through the interpretation of the many models, we have concluded that the most important factor in predicting mental health is the age group, followed by the frequency of feeling nervous, anxious or on edge, the frequency of worrying too much about different things, the ability to meet financial obligations or essential needs, the frequency of being restless, the increase of eating junk food or sweets. We hope that these findings can be utilized by Canadians to help preserve or control their mental health.

In future work, we would like to aggregate more data on economic and social activities, parenting, and substance use and stigma. With more data from a diverse range of topics, we would have the ability to try applying more complex and accurate models to the data and make stronger and more detailed conclusions about the impacts COVID-19 has on the mental health of Canadians.

6. ACKNOWLEDGMENTS

Our thanks to Dr. Yan Yan for guiding and supporting us with this project.

7. REFERENCES

[1] Bilogur, A. (2018, April 28). Simple techniques for missing data imputation. Kaggle. Retrieved April 25, 2022, from https://www.kaggle.com/residentmario/simple-techniques-for-missing-data-imputation/

[2] Brownlee, J. (2020, August 14). Naive Bayes for machine learning. Machine Learning Mastery. Retrieved April 25, 2022, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/

[3] Caie, P. D., Dimitriou, N., & Arandjelović, O. (2021). Precision Medicine in digital pathology via Image Analysis and machine learning. Artificial Intelligence and Deep Learning in Pathology, 149–173. https://doi.org/10.1016/b978-0-323-67538-3.00008-7

[4] Fafalios, S., Charonyktakis, P., & Tsamardinos, I. (2020). Gradient Boosting Trees.

[5] Horný, M., & Lin, M.-Y. (2018). Bayesian networks. Handbook of Machine Learning, 77–96. https://doi.org/10.1142/9789813271234_0005

[6] Jakkula, V.R. (2011). Tutorial on Support Vector Machine (SVM).

[7] Kotsiantis, S. , Kanellopoulos, D. , Pintelas, P. (2007). 'Data Preprocessing for Supervised Leaning'. World Academy of Science, Engineering and Technology, Open Science Index 12, International Journal of Computer and Information Engineering, 1(12), 4104 - 4109.

[8] Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., & Tsamardinos, I. (2017). Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. Journal of Statistical Software, 80(7). https://doi.org/10.18637/jss.v080.i07

[9] Osisanwo, F. Y., Akinsola , J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: Classification and comparison. International Journal of Computer Trends and Technology, 48(3), 128–138. https://doi.org/10.14445/22312803/ijctt-v48p126

[10] Rezapour, M., & Hansen, L. (2022). A machine learning analysis of COVID-19 Mental Health Data. https://doi.org/10.21203/rs.3.rs-1129807/v1

[11] Satpathy, S. (2021, January 6). Overcoming Class Imbalance using SMOTE Techniques. Analytics Vidhya. Retrieved April 25, 2022, from https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

[12] Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ, 338(jun29 1). https://doi.org/10.1136/bmj.b2393