

Planning for a Crisis: Predicting Anxiety in a Population During COVID-19 Using Machine Learning

Bhawna KUMARI^{a,*}, Nidhi GOYAL^{a,*} and Christo ELMORR^{b,1}

^a Indian Institute of Technology, Kharagpur, India

^b School of Health Policy and Management, York University, Toronto, Canada

* Authors contributed equally

ORCID ID: Bhawna KUMARI <https://orcid.org/0009-0004-8581-3293>

Nidhi GOYAL <https://orcid.org/0009-0004-3921-9410>

Christo EL MORR <https://orcid.org/0000-0001-6287-3438>

Abstract. COVID-19 impact on population mental health has been reported around the world. Statistics Canada has conducted a survey among Canadian population to gauge mental health challenges they experienced, specifically in terms of anxiety. We create a machine learning model to predict anxiety symptoms as measured by the General Anxiety Scale among the sample of 45,989 respondents to the survey. Eight algorithms including Logistic Regression, Random Forest, Naive Bayes, K Nearest Neighbours, Adaptive boost, Multi linear perceptron, XGBoost and LightBoost. LightBoost provided the highest performing model AUC score (AUC=87.45%). In addition, the features “perception of mental health compared to before physical distancing”, “perceived life stress”, and “perceived mental health” were found to be the most important three features to predict anxiety. A limitation of this study is that the sample is not representative of the Canadian population. Preparing for virtual care interventions during a crisis need to take into considerations these factors.

Keywords. Machine learning, mental health, virtual care; tele mental health; anxiety

1. Introduction

COVID-19 had a high impact on population health [1] and different communities including healthcare workers [2] and students [3]. Covid-19 pandemic had worsened populations’ mental health, including anxiety. In Canada, healthcare professionals reported higher anxiety symptoms [4] and quarantine was associated with higher odds of mental distress [5]. Anxiety could be a debilitating condition. An inquiry into population’s anxiety is needed to understand, and plan for mental health interventions such as virtual care or tele-mental health. If we can understand better experience of anxiety and predict anxiety symptoms, policies can be implemented, and appropriate programs deploy in a tailored manner. ML has been used successfully to predict many types of diseases including stroke, COVID-19 [6], diabetes, breast cancer, kidney disease,

¹ Corresponding Author: Christo El Morr, School of Health Policy and Management York University Toronto, Canada, elmorr@yorku.ca

heart failure [7]. We have used machine learning (ML) and the Statistics Canada survey on Canadian's mental health during COVID-19 [8] to predict a significant level of anxiety.

2. Methods

2.1. Dataset

The analysis is based on the second cycle (45,989 respondents) of the Statistics Canada survey: Canadian Perspectives Survey Series - Public Use Microdata File (April to May 2020) [9]. All Canadians aged 15 and older, were the intended population for this self-administered five-minute online survey.

2.2. Features Used

We have used all variables provided by the dataset including gender, age group, community size, indigenous identity flag, immigration status, province of residence, rural/urban indicator, visible minority, mental health compared to before physical distancing, perceived life stress, perceived mental health, frequency of going to shopping at the grocery store or drugstore, frequency of using delivery service for groceries or drugstore, frequency of using a food delivery service for prepared food, concerns about: own health, member of household's health, vulnerable people's health, Canadian population's health, world population's health, overloading the health system, civil disorder, maintaining social ties, ability to cooperate and support one another, ability to cooperate after crisis, family stress from confinement, violence in the home, ability to meet financial obligations or essential needs, losing main job or self-employment income in the next 4 weeks. The survey measured anxiety using the General Anxiety disorder (GAD) [10].

GAD scores are as follows: 0-4 for minimal anxiety; 5-9 for mild anxiety; 10-14: for moderate anxiety and greater than 15 for severe anxiety. The binary "Generalized Anxiety Disorder Cut-point" variable was identified as the target variable: moderate to severe anxiety (GAD ≥ 10) vs. minimal to mild anxiety (GAD score < 10).

2.3. Pre-processing, Algorithm Selection and Performance Measurements

The missing values were replaced with NaN values which were then imputed using K-Nearest Neighbor (KNN) Imputer. We investigated the following algorithms: Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), KNN, XGBoost, AdaBoost, Multilayer Perceptron (MLP), and LightBoost. The performance measurement obtained were the receiver operating characteristic (ROC) area under the curve (AUC), Accuracy, Precision, and F1-score. We performed hyperparameter tuning using randomized search. MinMaxScaler was used to normalize the dataset.

Every model produces true positive (TP), true negative (TN), false positives (FP) and false negatives (FN). We have used the AUC to compare the models as both the True positive rate and false negative rate are important to know in a medical situation (i.e., mental health). In our case, we would like to arrive at a model that maximizes both the true positives rate known as sensitivity (TP/(TP+FP)), and the true negatives rate known

as specificity (TN/(TN+FN)), the AUC is a measure of performance across all possible classification thresholds; at the same threshold (e.g., 0.5) a higher AUC reflects a higher sensitivity and a higher specificity of a predictive model. Since in the case of mental health, the cost of a FP is as important as the cost of FN, then AUC is preferable over precision or recall. Finally, feature importance was evaluated using Python algorithms’ “feature_importances_”.

3. Results

The sample had 5.04% of people with age 15 to 24 years; 21.71% of people with age between 25 to 34 years; 26.84% of people with age 35 to 44 years; 19.55% of people with age 45 to 54 years; 16.18% of people with age 55 to 64 years; 10.67% of people with age 65 years and older. Minimal, and mild anxiety were reported by most participants (38.77% and 32.1% respectively), while moderate and severe anxiety were reported by 16.77% and 12.36 % respectively.

The best ROC score was obtained with the LightBoost model (90.45%), closely followed by AdaBoost (90.28%) and MLP (90.31%). The KNN model had the least ROC score (87.28%). The models’ performance measurements are presented in Table 1. LightBoost model proved to be the best performing the terms of both ROC score and Accuracy score.

Table 1. Models’ Performance measurements

Algorithm	AUC	Accuracy	Precision	F1-Score	Time in sec. train/test
LR	89.89%	83.81%	86.57%	88.96%	20.7 / 0.02
RF	89.84%	83.86%	83.86%	83.86%	31.83 / 0.52
Naive NB	87.28%	81.44%	83.30%	87.63%	0.16 / 0.04
KNN	87.28%	81.53%	86.06%	87.84%	0.01 / 3.25
XGBoost	89.52%	83.80%	83.82%	83.80%	0.47 / 0.01
MLP	90.31%	83.93%	83.93%	83.93%	77.5 / 0.03
LightBoost	90.45%	84.25%	84.25%	84.25%	0.42 / 0.06
AdaBoost	90.28%	83.75%	86.20%	86.20%	8.86 / 0.46

We computed the feature importance ranking method for LightBoost; ranking was computed as a score out of 1000. Perception of mental health compared to before COVID-19 (score 973), perceived life stress (score 901), perceived mental health (710) and ability to meet financial obligations or essential needs (623) were found to be the most important features in predicting anxiety. We have checked the LightBoost model for bias against indigenous, immigrants, female, and below 65 years old, using the Equal opportunity difference (EOD). LighBoost showed small bias against people 64 years or younger (-6.4%), female (-4.3%), immigrant (-1.8%), and indigenous (-9.05%).

4. Discussion

In our study, we are aiming to develop a predictive health and in such situations the AUC is a best measurement for performance. A test that has an AUC value between 80% and 90% is considered an excellent one, while a test with more than 90% AUC is considered outstanding [11]. While LightBoost achieved the highest AUC (90.45%) AdaBoost and MLP achieved closely similar AUC.

Worsening perception of mental health compared to before physical distancing, as well as negative perceived life stress and perceived mental health were all high predictors of anxiety; this is in line with findings from previous studies [12, 13]. When we have built a LighBoost model using only the 4 most important features, the AUC value decreased only slightly 88.93%; hence, answers to these 4 simple questions provide a robust predictive model.

This is important finding, as asking a simple question about subjective experiences can lead to have an insight into the anxiety symptoms. The ability to meet financial obligations or essential needs was also highly important in predict anxiety, this confirms prior studies that found income to be associated with anxiety among adolescent [14] or in the general population [15]. This might inform policy makers to address economic effects of crises (e.g., disasters, pandemics) to mitigate mental health impact, in addition to deploying virtual mental health tools [16].

While there are studies using ML to predict anxiety during the COVID-19 pandemic [17], this is the first study that analyzes the statistics Canada dataset related to anxiety during covid-19 using a machine learning approach. One limitation of this study is that the sample is not representative of the Canadian population. The relatively high bias of the model against indigenous population is another limitation. Debiasing the model is our next step.

5. Conclusion

We have explained and discussed a model to predict anxiety for the general population in Canada based on a national survey. LighBoost and Adaboost models are both good candidates for anxiety symptom as they both provide comparable AUC and accuracy. In addition to virtual mental health applications, government financial support for vulnerable populations during times of crises could mitigate impact on anxiety.

Acknowledgment

The work was supported by a Mitacs Globalink Research Internship award.

References

- [1] Ghio D, Acosta E, Fisman D, Noymer A, Stilianakis NI, Assche SB. Population Health and COVID-19 in Canada: a Demographic Comparative Perspective. *Can Stud Popul.* 2021;48(2-3):131-7.
- [2] Tardif A, Gupta B, McNeely L, Feeney W. Impact of the COVID-19 Pandemic on the Health Workforce in Canada. *Healthc Q.* 2022;25(1):17-20.
- [3] Wang C, Zhao H, Zhang H. Chinese College Students Have Higher Anxiety in New Semester of Online Learning During COVID-19: A Machine Learning Approach. *Front Psychol.* 2020;11:587413.
- [4] Turna J, Patterson B, Goldman Bergmann C, Lamberti N, Rahat M, Dwyer H, et al. Mental health during the first wave of COVID-19 in Canada, the USA, Brazil and Italy. *Int J Psychiatry Clin Pract.* 2022;26(2):148-56.
- [5] Daly Z, Slemmon A, Richardson CG, Salway T, McAuliffe C, Gademmann AM, et al. Associations between periods of COVID-19 quarantine and mental health in Canada. *Psychiatry Res.* 2021;295:113631.
- [6] Gallo Marin B, Aghagholi G, Lavine K, Yang L, Siff EJ, Chiang SS, et al. Predictors of COVID-19 severity: A literature review. *Rev Med Virol.* 2021;31(1):1-10.

- [7] Mavrogiorgou A, Kleftakis S, Zafeiropoulos N, Mavrogiorgos K, Kiourtis A, Kyriazis D, editors. A Comparative Study of ML Algorithms for Scenario-agnostic Predictions in Healthcare. 2022 IEEE Symposium on Computers and Communications (ISCC); 2022 30 June-3 July 2022.
- [8] Statistics Canada. Crowdsourcing: Impacts of COVID-19 on Canadians' Mental Health Public Use Microdata File: Statistics Canada; 2020 [updated June 11, 2020]. Available from: <https://www150.statcan.gc.ca/n1/en/catalogue/13250002>.
- [9] Statistics Canada. Canadian Perspectives Survey Series (CPSS): Statistics Canada; 2020 [updated February 4, 2021]. Available from: <https://www150.statcan.gc.ca/n1/pub/13-25-0002/132500022020001-eng.htm>.
- [10] Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. 2006;166(10):1092-7.
- [11] Mandrekar JN. Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*. 2010;5(9):1315-6.
- [12] Mahalingam M, Jammal M, Hoteit R, Ayna D, Romanid M, Hijazi S, et al. A Machine Learning Study to Predict Anxiety on Campuses in Lebanon. 21st International Conference on Informatics, Management, and Technology in Healthcare (ICIMTH 2023); Athens, Greece2023.
- [13] Sisodia A, Jammal M, Christo EM. A Machine Learning Approach to Predict Poor Mental Health of Intimate Partner Violence Survivors. *International Conference on Advances in Computational Tools for Engineering Applications (ACTEA)*; Beirut, Lebanon2023.
- [14] Vine M, Vander Stoep A, Bell J, Rhew IC, Gudmundsen G, McCauley E. Associations between household and neighborhood income and anxiety symptoms in young adolescents. *Depress Anxiety*. 2012;29(9):824-32.
- [15] Thomson RM, Igelström E, Purba AK, Shimonovich M, Thomson H, McCartney G, et al. How do income changes impact on mental health and wellbeing for working-age adults? A systematic review and meta-analysis. *The Lancet Public Health*. 2022;7(6):e515-e28.
- [16] El Morr C. Virtual Communities, Machine Learning and IoT: Opportunities and Challenges in Mental Health Research. *International Journal of Extreme Automation and Connectivity in Healthcare*. 2019;1(1):4-11.
- [17] Albagmi FM, Alansari A, Al Shawan DS, AlNujaidi HY, Olatunji SO. Prediction of generalized anxiety levels during the Covid-19 pandemic: A machine learning-based modeling approach. *Inform Med Unlocked*. 2022;28:100854.