SUPERVISED MACHINE LEARNING CLASSIFICATION OF DEPRESSION AMONG NEW YORK CITY RESIDENTS, 2018-2020

By: Carina P Doyle

B.A. Biology, New York University, 2019

Thesis Advisor: Dr. Siyu Heng, Ph.D.

A Master’s thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Biostatistics

School of Global Public Health

New York University

May 2024

**Abstract:**

**Introduction/Objective:**

This study aims to determine the most efficient machine learning classification technique for classifying depression among New York City residents in each of the years 2018, 2019, and 2020. Though previous studies have analyzed depression classified with machine learning, little research has identified efficient machine learning techniques for depression classification among New York City citizens. A deeper knowledge of machine learning classification for depression during these years could illuminate whether algorithms are equally good at classification across years.

**Methods:**

Using data from the New York City Department of Health and Mental Hygiene Community Health Survey in years 2018, 2019, and 2020, logistic regression, Naïve Bayes, K-nearest neighbors, decision trees, Random Forest, and XGBoost were performed on data from each year individually. Model performance was shown through accuracy and AUC values, and variable importance was displayed.

**Results:**

Following analyzing the AU-ROC values of each model by year, logistic regression was determined to be the most efficient machine learning technique for 2018 and 2020, while for 2019, XGBoost was the most reliable algorithm. For 2018, the variable that was most vital for classifying depression was general health status, while for 2019 and 2020, it was difficulty of daily activities.

**Conclusions:**

These results suggest that algorithmic performance may not be consistent year to year, as the most reliable model changed from logistic regression to XGBoost and back to logistic regression. Further research is needed to determine what may have caused this change.

**Key Words:** Machine Learning, Depression, New York City, Logistic Regression, XGBoost

*Abstract Word Count: 250*

*Main Text Word Count: 4,486*

*Number of Tables: 3*

*Number of Figures: 2*

**Introduction:**

Depression affects a large number of New York City citizens; major depressive disorder is likely the greatest source of disability for New York City inhabitants, and only 40% of those who suffer from this disease receive necessary treatment.1 Though mental health has always been a concern for New Yorkers, the Covid-19 pandemic greatly exacerbated the preexisting mental health problems the city faces, due to increased isolation and the vast death toll caused by the virus.2,3 In adolescents and young adults, rates of generalized anxiety and social anxiety increased; in females specifically, depression and pain/somatic symptoms increased.4 However, mental health symptoms may not be long-lasting; in one study, depression and anxiety symptoms were shown to be highest in April 2020, and these symptoms decreased over the summer.5

Machine learning is an efficient, inexpensive statistical analysis technique which utilizes computer programming to train an algorithm to make accurate classifications. A basic machine learning analysis includes feature selection to determine influential variables, developing and testing algorithms for classification, and assessing model reliability and accuracy. Machine learning algorithms vary in terms of reliability, depending on data quality, sample size, and variable types. Therefore, different machine algorithms may excel or fail at classifying values to an outcome, depending on the specifics of the outcome, the most influential variables, and the characteristics of the dataset.6

Many studies use multiple machine learning techniques when predicting an outcome, to determine the most efficient algorithm. In such studies, the algorithm that prevails as the most reliable varies. For example, one study determined that among front line healthcare workers, random forest was most effective for classifying depression7, and another study that analyzed mental health issues among adolescents also determined that random forest was the most efficient algorithm for classification.8 However, an additional study among adult United States citizens found that extreme gradient boosting was most reliable at classifying mental health outcomes, surpassing the effectiveness of random forest.9 More complicated models are not always preferred, as the simplicity of naïve bayes and logistic regression may sometimes perform equally well and be less computationally complicated.6,8

*Logistic Regression*

Logistic regression (LR), originally a statistical method, is implemented in machine learning for classification of variables that have a binary outcome.10 This binary outcome may be expressed as a probability, falling between zero and one. To ensure values fall within these boundaries, a logit (log odds) scale is used. Logistic regression may be used with multiple independent variables and the model can determine whether specific variables have greater influence in the classification of the binary outcome.11 Logistic regression has been proven to be effective for classification of values in Type 2 Diabetes11, protein interactions12, and hospital mortality.13 Some studies find that logistic regression, due to having greater requirements for use, is more accurate for classification, while others find that the requirements of logistic regression are not necessary to make accurate classifications, so other machine learning methods are preferred.14

*Naïve Bayes*

Naïve Bayes (NB) classifier functions by using Bayes theorem to classify data points depending on their attributes. This algorithm assumes that each predictor is independent, meaning that the value of one predictor does not influence the value of another predictor. Naïve Bayes is a part of the Bayesian Network of algorithms for classification but is known as both the simplest algorithm and the most reliable.15 Naïve Bayes has shown effectiveness at classifying outcomes of heart disease16, water quality17, and carcinogenicity of chemicals18, but has not been proven to effectively classify mental health outcomes. Furthermore, NB may not be the most effective model for machine learning, especially when the requirement of independent predictors is violated. Through providing the independent predictors are maintained, a researcher can ensure the effectiveness of a Naïve Bayes machine learning algorithm.15

*K-Nearest Neighbors*

K-Nearest Neighbors (KNN) classification uses Euclidian distance and K number of nearest data points in order to classify a value into a class. Using the number of datapoints that K is set to, the algorithm determines the distance between the datapoint being assigned and the K number of closest values in the feature space; the point is assigned to whichever class has the largest number of close values. Determining the best value for K is regarded as a tradeoff between using a low K value, which may increase the risk of underfitting the model, and a large K value, because as K increases, variance decreases and bias increases, and the model may become overfitted, matching the training dataset too closely and being unreliable for classification on a new dataset.19 20 Previously, KNN methods have proven effective at classifying diabetes diagnoses21, schistosomiasis22, and certain cancer diagnoses.23 KNN methods have been used to classify mental health outcomes but are rarely chosen as the most effective method at classifying this outcome.7,24

*Decision Trees*

Decision trees may be used in machine learning to classify variables into various outcomes. By following a path along a tree, from root to nodes to leaf, the algorithm can determine how a variable should be classified. However, decision tree methods work best when all predictor variables are binary, as this factor aids in the production of node splits to determine outcome. Conditionally probabilities may be generated for each node to determine the probability that an item reaching this point belongs to the correct group. Ideally, each node will account for values from ten to twenty data points. Any nodes or branches that are unreliable or do not accurately classify the outcome of interest may be pruned from the tree in order to increase overall reliability; however, decisions trees are often vulnerable to model overfitting.25 Studies have shown that decision trees are effective at classifying mortality26, Covid-19 case outcomes27, and supranuclear palsy28; additionally, studies have been completed using decision trees as one of many algorithms being tested to classify mental health, but decision trees are rarely determined to be the most effective algorithm for this outcome.7-9,24 Decision trees may be more effective at classifying mental health outcomes when combined with other techniques, as seen in Extreme Gradient Boosting and Random Forest.7-9,24

*Random Forest*

The Random Forest (RF) method utilizes multiple decision trees to generate a ‘forest’ of trees that the algorithm relies on for classification. Because the model has many different trees, which use different features that are randomly selected, randomness is incorporated into the model which may reduce overfitting, a common problem with the simple decision tree machine learning method. To create the subsets of data for individual trees, random forest uses a technique known as bagging, which comprises of randomly selecting datapoints from the dataset with replacement, so each individual sample is different but has overlap with other samples. Because the overall model is created from many different trees, classifications are determined through voting; each tree counts as one vote for classifying an outcome, and the class with the most votes is chosen.29,30 RF algorithms have proven effective at classifying macrosomia31, sarcoidosis, and tuberculosis32, as well as mental health decline.24

*Extreme Gradient Boosting*

Extreme gradient boosting (XGBoost) is a relatively new machine learning technique that has shown promising ability at classifying disease. Gradient boosting, which this type of model is based on, essentially comprises of building a first, weak model, from a technique such as decision trees, and then creating more elaborate models to fix any errors associated with this first model. In order to determine errors of the decision tree models, XGBoost employs a logistic loss function for classification.33 Studies have shown that this technique is reliable for classiftying mortality due to sepsis34, aneurysmal subarachnoid hemorrhage35, and some smoking-induced diseases.36 XGBoost is often used by studies to classify mental health outcomes, in conjunction with other machine learning techniques, and shows promising reliability.7,24

Many studies have analyzed the efficiency of the following machine learning algorithms at predicting mental health outcomes7-9,24, but to the researcher’s knowledge, no studies of machine learning prediction for mental health outcomes have yet been conducted on the New York City community. Predicting mental health outcomes in this population could lead to identifying the demographic variables that are most influential to mental health and to quicker diagnoses and treatment of mental illnesses based on contributing factors.37 Furthermore, utilizing data from multiple years, pre- and post-pandemic, will allow the researcher to determine whether the most reliable machine learning algorithm remains consistent across an event that may change the characteristics of those have depression.

**Research Question**

The objective of this study is to determine the most reliable supervised learning algorithms to predict poor mental health outcomes in New York City using NYC Community Health Survey data from 2018, 2019, and 2020.

**Methods**

*Data Source and Study Design*

The NYC Community Health Survey is a cross-sectional survey run by the New York City Department of Health and Mental Hygiene. This survey has been used to collect data on the health of New York City residents since 2009, including demographic information, general health, mental health, sexual wellness, and yearly current topics such as the 2012 hurricane and 2020 Covid-19 pandemic. Data from the years 2018 (n=10076), 2019 (n=8803), and 2020 (n=8781) will be used for this study. To complete this analysis, depression will be analyzed against other variables in the data (including gender, race, education, health status, and exercise) to determine the variables that are most important for classifying depression and the machine learning algorithm that is most reliable at classifying depression within this population.

*Participants and Procedures*

This sample is comprised of citizens of New York City who are aged 18+ in the years 2018, 2019, 2020 sampled through stratified random sampling of households in all five boroughs, using United Hospital Fund neighborhood definitions to create neighborhood estimates. Data is collected either through computer-assisted telephone interview (CATI) surveys. Surveys are conducted in a variety of languages, including English, Spanish, Chinese (traditional and simplified), Russian, and Haitian Creole. The response rates were 8.4% (2018), 7.2% (2019), and 7.4% (2020). These survey data were not weighted, as using multiyear weights is considered classified information to the DOHMH, and the researcher was unable to access the multiyear weight values.

*Depression*

The outcome variable for this study is depression. This outcome was assessed through survey diagnostic tools, the Personal Health Questionnaire Depression Scale (PHQ-8) in 2018 and the Kessler-6 index (K6) in 2019 and 2020. These tools are considered to be effective at determining which individuals are suffering from a depression condition.38,39

The PHQ-8 comprises of eight survey questions, assessing “Over the last 2 weeks, how often have you been bothered by:” with the following issues: 1-“little interest or pleasure in doing things”, 2-“feeling down, depressed or hopeless?”, 3-“trouble falling or staying asleep, or sleeping too much?”, 4-“feeling tired or having little energy”, 5-“poor appetite or overeating”, 6-“feeling bad about yourself – or that you are a failure or have let yourself or your family down”, 7-“trouble concentrating on things, such as reading the newspaper or watching TV”, and 8-“moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual.” Each question has the response options: “Not at all”, “Several days”, “More than half of days”, and “Nearly every day.” Each participant’s PHQ-8 score is then calculated from these responses, by assigning points to each response type, zero through three. Participants with scores equal to or greater than 10 are considered to be currently experiencing major depression. Following score calculation, a binary variable was created based on PHQ-8 score, with those with scores at or greater to 10 marked as currently experiencing depression, and those with scores less than 10 marked as currently not experiencing depression.

In the years 2019 and 2020, the NYC DOHMH moved toward using the K6 diagnostic tool instead of the PHQ-8. K6 screening is comprised of six mental health questions, each beginning with “During the past 30 days, how often did you feel…”, and ending with the following six phrases: 1-“So sad that nothing could cheer you up”, 2-“Nervous”, 3-“Restless or fidgety”, 4-“Hopeless”, 5-“That everything was an effort”, and 6-“Worthless.” The response options for each of these questions were “All of the time”, “Most of the time”, “Some of the time”, “A little of the time”, and “None of the time.” K6 score is calculated using these responses, with each response type corresponding to a value, from zero to four. K6 scores greater or equal to 13 were considered to be cases of depression. Following score calculation, a binary variable was created, coding those with a K6 score greater than or equal to 13 as having depression and those with a K6 score less than 13 as not having depression.

*Predictors*

Predictors were selected through supervised feature selection using LASSO feature selection. Important predictors include self-assessed general health status (“Excellent”, “Very Good”, “Good”, “Fair”, “Poor”), difficulty performing daily tasks due to physical, emotional, or mental problems (“Yes”, “No”), and nutrition in terms of number of cups of fruits and vegetables consumed daily (range of values, 0-50).

*Statistical Analysis*

All analyses will be performing in R Version 4.2.1. R and R Studio are open-source software that allow for high level statistical analyses to be performed through the R language. Relevant R packages that will be used for this project include dplyr (data recoding/formatting), haven (data import), gglot2 (data visualization), and RandomForest (a machine learning package designed for the Random Forest algorithm). Prior to any statistical analyses beginning, data will be pre-processed by K-nearest neighbors imputation of missing values and creating any new necessary variables. KNN imputation was chosen due to its effectiveness on datasets of a similar size to that of this study and due to its ability to impute multiple variables at once, considering variable relationships. Training and testing datasets will be randomly sampled from the overall dataset at a 70-30 ratio, in order to give the models more information on which to train. All analyses will be conducted for each year of data, separately, so datasets will not be merged.

*Data Balancing*

Balancing data through over/undersampling will be essential to gain successful results in this study. When learning models are trained on unbalanced data, there are often too few positive cases for the model to learn what identifies a case, and the model is unlikely to effectively classify the testing data. Balancing the training dataset to have a similar number of cases and controls allows models to better learn how to identify cases, making classifying testing data more reliable. The training dataset was balanced (as the outcome was relatively rare, at around 5%-10% of the population) using both oversampling of cases and undersampling of controls, in order to create a relatively even dataset for analysis. However, the testing dataset was not balanced to assess how the models perform on realistic data.

*Feature Selection*

Feature will be determined through the Least Absolute Shrinkage and Selection Operator) (LASSO) method. LASSO is a linear model that utilizes L1 regularization to shrink unimportant coefficients to 0. LASSO uses k-fold cross validation to ensure model is fit correctly. Once the LASSO model has run, the coefficients for each variable in the dataset are shown, with only important features having non-zero coefficients. For 2018, 25 features were selected, including general health status, race, age, employment, and sleep quality. For 2019, 39 features were selected, including general health status, difficulty of daily activities, education, and birth sex. For 2020, 40 features were selected, including general health status, difficulty of daily activities, age, and cycling frequency.

*Model Implementation*

Logistic regression, using a log odds scale, will determine probabilities of each participant being classified as a case or a control. The Naïve Bayes model assumes independence of predictors, and classifies individuals by their unique attributes. KNN functions through computing Euclidian distance from one participant to another in order to assign a value to this participant, with K being the number of neighbors accounted for when assigning value. The decision tree model will generate a path of decisions to arrive at a classification, with each node having a probability, and the final tree will be pruned to remove extraneous trees. Finally, XGBoost will be used as an iterative model that builds on itself through assessing model error with a log loss function and fixing issues of each previous model, resulting in a final, well-tuned model.

*Cross Validation*

10-fold cross validation was used to ensure that the final model was correctly trained and tuned, and will function correctly on the unseen testing dataset. 10-fold cross validation functions through partitioning the data into ten sets, training the model on nine of those sets, and using the final one set to test the function of the model. This process is repeated, ensuring each one of the partitions is used once as the held-out testing set, and models are evaluated for accuracy, and combined to create a single estimated accuracy value for the final model.

*Model Evaluation*

Machine learning models will be assessed through area under receiver-operating curves (AU-ROC), which analyzes the specificity, sensitivity, and accuracy of the model. Accuracy will be computed through dividing all correctly assigned values by the total population size. AU-ROC curves assess each of these factors, along with positive predictive value and negative predictive value and generates a curve to demonstrate model fitness; models with an AU-ROC value of 1 are considered perfect, but achieving such a high score is improbable. Models with AUC values from 0.7-0.8 are considered acceptable, and models with AUC values from 0.8-0.9 are considered excellent. Models with the highest AU-ROC value will be considered this study’s most reliable models.

**Results**

*Participant Characteristics (Univariate Analysis)*

Univariate analysis was vital for determining variable balance in these data. In 2018, about 10% of participants received a positive depression screening with the PHQ-8, while in around 5% received a positive depression screening with K6 in 2019 and 2010. Most participants in this study assessed that they were in good (32%, 30%, 30%) to very good (26%, 28%, 31%) health, and age of participants ranged, but most participants were at least 30 years old (86%, 84%, 84%). Birth sex was relatively fairly split, tending slightly towards females (57%, 56%, 56%). The most prevalent race of survey population was white, non-Hispanic (35%, 35%, 33%), and most participants were either married (39%, 35%, 33%) or never married (30%, 30%, 33%) (not divorced, widowed, etc.). Most participants were employed (45%, 49%, 46%), received health insurance through their employer (39%, 39%, 41%), and graduated college (42%, 46%, 46%). For 2019 and 2020, 17% and 13% of survey participants had difficulty performing daily tasks due to physical, mental, or emotional problems, respectively.

*2018 Models*

For the year 2018, with the outcome being depression as calculated by the PHQ-8 diagnostic tool, the logistic regression model (AUC=0.864) had the highest AUC value for classifying the outcome (**Table 2**); this pattern is confirmed through ROC curve comparisons (**Figure 1a**). The Naïve Bayes model had an AUC of 0.858, while K-nearest neighbors had an optimal K=7, with an AUC of 0.763. The decision tree algorithm had an AUC of 0.787, while Random Forest had an AUC of 0.833. Finally, the XGBoost method had an AUC of 0.855. For 2018, logistic regression had the highest AUC value, with naïve Bayes and XGBoost following. Random forest had the highest accuracy value at 0.853 (**Table 3**).

After determining that logistic regression had the highest AUC value for these data, features were analyzed for importance in depression classification. Feature importance plots (**Figure 2a**) determined that the most important feature was general health status (Poor), followed by sleep quality (Very Bad), and employment (Unable to work). Additional important features or variables levels included general health status (Fair), sleep quality in the past 30 days (Fairly Bad), had enough accessible food (Sometimes Not Enough), employment (Unemployed, >1 year), not able to get medical care in the past year (Yes), being unable to afford medical care in the past year (Yes), and exercising in the past 30 days (No).

*2019 Models*

For the year 2019, with the outcome being depression as calculated by the K6 questionnaire, the XGBoost algorithm (AUC=0.852) had the highest AUC value for classifying depression (**Table 2**), when analyzing AUC and ROC curve (**Figure 1b**). The logistic regression technique had an AUC of 0.845, and naïve Bayes had an AUC value of 0.842. K-nearest neighbors cross validation determined that K=7 was the most effective K for this sample, and the AUC value of this model was 0.751. The decision tree model had an AUC of 0.765, while Random Forest had an AUC of 0.836. Overall, all models showed relatively high AUC values, with XGBoost, logistic regression, and naïve Bayes having the highest AUC values for 2019. Random forest had the highest accuracy at 0.935 (**Table 3**).

Because the XGBoost model demonstrated the highest AUC, this model was used to determine feature importance for depression classification among 2019 survey participants. Feature importance plots (**Figure 2b**) demonstrated that difficulty performing daily acts due to physical, emotional, or mental issues (Yes) was the most important feature in 2019, with general health status (Poor) and age group also showing importance. Further analysis determined that age groups four (45-64) and five (65+) specifically were important features, compared to other age groups. Additional important features included marital status (Never married), number of cups of fruits and vegetables consumed per day, health insurance (No), employment (Unable to work), education (Less than HS), having cognitive decline (Yes), and experiencing rape (Yes).

*2020 Models*

For the year 2020, with the outcome being depression as calculated by the K6 questionnaire, the logistic regression method had the highest AUC value of 0.815 (**Table 2**), as confirmed by ROC curve plots (**Figure 1c**). The naïve Bayes model had an AUC of 0.805, while K-nearest neighbors had most success with K=10, with an AUC of 0.744. The final decision tree model following pruning had an AUC of 0.720, while the Random Forest iterative tree model demonstrated an AUC of 0.809. Finally, the XGBoost model had an AUC of 0.792. In all, most models had relatively high AUC scores, with logistic regression, Random Forest, and naïve Bayes models having the highest AUC values for 2020. Random forest had the highest accuracy of these models, at 0.912 (**Table 3**).

Using the logistic regression model shown to have the largest AUC, feature importance plots were generated. This plot (**Figure 2c**) demonstrates that having difficulty performing daily acts due to physical, emotional, or mental problems was the most important feature for depression classification, with general health status (Poor) and experiencing intimate partner violence through insults or controlling behavior (Yes) contributing as well. Other contributing features were average sugary drink consumption per day (>1), having two or more sodas per day (Yes), having to delay rent payment (Yes), being a heavy drinker (Yes), and being a woman who has exclusively had sex with other women in the past 12 months (No).

**Discussion**

Random forest prevailed as the most accurate machine learning method for each year in these data. These values are of interest, especially because they do not align with the most effective machine learning method as chosen by AUC. However, accuracy is not always the best performance metric to analyze when determining the most overall effective algorithm; because accuracy only measures the percentage of correctly assigned cases, it does not consider the difference between specificity and sensitivity. Because testing data is unbalanced, unlike the training data, AUC is preferred as a measure of performance for machine learning techniques, because it is a better metric of how well the model assessed the few cases in the data.

While each of these models performs best under certain circumstances, these results demonstrate that all models used for these analyses perform relatively well on these data. Most AUC values are no lower than 7.5, showing that these models all perform much better at classification than a randomly guessing model, inferring that all models were able to learn the trends of the data and the characteristics that indicate an individual may have depression. Additionally, many AUC values are above 0.8, demonstrating that these models performed excellently at classifying participants in these populations. However, logistic regression is the overall best performer in terms of AUC value, only being surpassed by XGBoost in 2019, with naïve Bayes and random forest following closely behind.

Analyzing feature importance for the most effective machine learning models is essential to determine which variables are most associated with depression in these samples. General health status is consistently associated with depression, indicating that those who are depressed may consider themselves unhealthy as well as ill. Additionally, many of the commonly appearing features are intuitive; having difficulty performing daily acts could easily be depressing, and nutrition has been closely linked to mental health. Finally, many of these features have an easily implementable treatment, which could aid health departments in reducing depression rates.

*Limitations*

Some limitations originate from the survey material used in this study. Though a mental health question was asked on this survey, there was not a question asking about any prior mental illness diagnoses, which may have been more accurate as it would rely on diagnoses from physicians, as opposed to a screening questionnaire. In addition, there may be some response bias in these survey data, as people may be unwilling to honestly answer questions surrounding their mental health due to stigma. As well, the Community Health Survey, while a comprehensive health survey, shows strong unbalance between those with and without depression, due to this outcome being uncommon in the population. Therefore, bias may have been introduced into the results due to the oversampling/undersampling techniques used in order to balance the training dataset. Additionally, the method in which depression was determined in these surveys is different between 2018 and 2019-2020, meaning that classification of what exactly constitutes depression may not be consistent across the years, so it may be difficult to compare them. Finally, the data used in these analyses are unweighted, due to the need for multiyear weights in order to weight all years together. Weighting years separately would have made comparison between the years difficult, so data remained raw.

**Conclusions**

Essentially, the most reliable machine learning algorithm in terms of classification of depression did not remain the same from 2018-2020, nor did it change only for 2020 to reflect Covid-19 trends. For 2018 and 2020, logistic regression was most reliable, while XGBoost was most efficient for 2019. In terms of classification accuracy, random forest was consistently the most accurate model for all of the years. Machine learning model performance may vary due to population characteristics in a dataset, but these characteristics are not apparent at this time; more research is needed to determine why 2019 is better served by XGBoost. In future research, training models on multiyear data may make for more reliable predictive models that can be used over many years. In additional, testing other algorithms, such as Support Vector Mechanics, could merit more accurate or reliable results.

**References**

1. Kepler KL, Perlman SE, Chernov C, Thorpe LE, Hoenig J, Norman C. The Burden of Depression in New York City Adults: Results from the 2013-14 NYC Health and Nutrition Examination Survey. *J Urban Health*. Dec 2018;95(6):832-836. doi:10.1007/s11524-018-0283-1

2. Davies EG, Gould LH, Le K, et al. Collateral Impacts of the COVID-19 Pandemic: The New York City Experience. *J Public Health Manag Pract*. Jul-Aug 01 2023;29(4):547-555. doi:10.1097/phh.0000000000001701

3. Jiwani Z, Cadwell OG, Parnes M, Brown AD. Uncertainty in the Time of Corona: Precautionary Adherence, Fear, and Anxiety in New York City During the COVID-19 Pandemic. *J Nerv Ment Dis*. Apr 1 2021;209(4):251-255. doi:10.1097/nmd.0000000000001308

4. Hawes MT, Szenczy AK, Klein DN, Hajcak G, Nelson BD. Increases in depression and anxiety symptoms in adolescents and young adults during the COVID-19 pandemic. *Psychological Medicine*. 2022;52(14):3222-3230. doi:10.1017/S0033291720005358

5. Hawes MT, Szenczy AK, Olino TM, Nelson BD, Klein DN. Trajectories of depression, anxiety and pandemic experiences; A longitudinal study of youth in New York during the Spring-Summer of 2020. *Psychiatry Res*. Apr 2021;298:113778. doi:10.1016/j.psychres.2021.113778

6. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak*. Dec 21 2019;19(1):281. doi:10.1186/s12911-019-1004-8

7. Samuelson KW, Dixon K, Jordan JT, Powers T, Sonderman S, Brickman S. Mental health and resilience during the coronavirus pandemic: A machine learning approach. *J Clin Psychol*. May 2022;78(5):821-846. doi:10.1002/jclp.23254

8. Tate AE, McCabe RC, Larsson H, Lundström S, Lichtenstein P, Kuja-Halkola R. Predicting mental health problems in adolescence using machine learning techniques. *PLoS One*. 2020;15(4):e0230389. doi:10.1371/journal.pone.0230389

9. Garriga R, Mas J, Abraha S, et al. Machine learning model to predict mental health crises from electronic health records. *Nat Med*. Jun 2022;28(6):1240-1248. doi:10.1038/s41591-022-01811-5

10. Stoltzfus JC. Logistic regression: a brief primer. *Acad Emerg Med*. Oct 2011;18(10):1099-104. doi:10.1111/j.1553-2712.2011.01185.x

11. Deberneh HM, Kim I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *Int J Environ Res Public Health*. Mar 23 2021;18(6)doi:10.3390/ijerph18063317

12. Prasasty VD, Hutagalung RA, Gunadi R, et al. Prediction of human-Streptococcus pneumoniae protein-protein interactions using logistic regression. *Comput Biol Chem*. Jun 2021;92:107492. doi:10.1016/j.compbiolchem.2021.107492

13. Li F, Xin H, Zhang J, Fu M, Zhou J, Lian Z. Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database. *BMJ Open*. Jul 23 2021;11(7):e044779. doi:10.1136/bmjopen-2020-044779

14. Song X, Liu X, Liu F, Wang C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *Int J Med Inform*. Jul 2021;151:104484. doi:10.1016/j.ijmedinf.2021.104484

15. Zhang Z. Naïve Bayes classification in R. *Ann Transl Med*. Jun 2016;4(12):241. doi:10.21037/atm.2016.03.38

16. Maheswari S, Pitchai R. Heart Disease Prediction System Using Decision Tree and Naive Bayes Algorithm. *Curr Med Imaging Rev*. 2019;15(8):712-717. doi:10.2174/1573405614666180322141259

17. Ilić M, Srdjević Z, Srdjević B. Water quality prediction based on Naïve Bayes algorithm. *Water Sci Technol*. Feb 2022;85(4):1027-1039. doi:10.2166/wst.2022.006

18. Zhang H, Cao ZX, Li M, Li YZ, Peng C. Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals. *Food Chem Toxicol*. Nov 2016;97:141-149. doi:10.1016/j.fct.2016.09.005

19. Guo G, Wang H, Bell D, Bi Y, Greer K. KNN Model-Based Approach in Classification. Springer Berlin Heidelberg; 2003:986-996.

20. Zhang S, Li X, Zong M, Zhu X, Wang R. Efficient kNN Classification With Different Numbers of Nearest Neighbors. *IEEE Trans Neural Netw Learn Syst*. May 2018;29(5):1774-1785. doi:10.1109/tnnls.2017.2673241

21. Prasad B, Gupta S, Borah N, Dineshkumar R, Lautre HK, Mouleswararao B. Predicting diabetes with multivariate analysis an innovative KNN-based classifier approach. *Prev Med*. Sep 2023;174:107619. doi:10.1016/j.ypmed.2023.107619

22. Zhou X, Wang H, Xu C, et al. Application of kNN and SVM to predict the prognosis of advanced schistosomiasis. *Parasitol Res*. Aug 2022;121(8):2457-2460. doi:10.1007/s00436-022-07583-8

23. Ehsani R, Drabløs F. Robust Distance Measures for kNN Classification of Cancer Data. *Cancer Inform*. 2020;19:1176935120965542. doi:10.1177/1176935120965542

24. Rezapour M, Hansen L. A machine learning analysis of COVID-19 mental health data. *Sci Rep*. Sep 2 2022;12(1):14965. doi:10.1038/s41598-022-19314-1

25. Kingsford C, Salzberg SL. What are decision trees? *Nature Biotechnology*. 2008/09/01 2008;26(9):1011-1013. doi:10.1038/nbt0908-1011

26. Elhazmi A, Al-Omari A, Sallam H, et al. Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU. *J Infect Public Health*. Jul 2022;15(7):826-834. doi:10.1016/j.jiph.2022.06.008

27. Migriño JR, Jr., Batangan ARU. Using machine learning to create a decision tree model to predict outcomes of COVID-19 cases in the Philippines. *Western Pac Surveill Response J*. Jul-Sep 2021;12(3):56-64. doi:10.5365/wpsar.2021.12.3.831

28. Koga S, Zhou X, Dickson DW. Machine learning-based decision tree classifier for the diagnosis of progressive supranuclear palsy and corticobasal degeneration. *Neuropathol Appl Neurobiol*. Dec 2021;47(7):931-941. doi:10.1111/nan.12710

29. Hu J, Szymczak S. A review on longitudinal data analysis with random forest. *Brief Bioinform*. Mar 19 2023;24(2)doi:10.1093/bib/bbad002

30. Rigatti SJ. Random Forest. *J Insur Med*. 2017;47(1):31-39. doi:10.17849/insm-47-01-31-39.1

31. Wang F, Wang Y, Ji X, Wang Z. Effective Macrosomia Prediction Using Random Forest Algorithm. *Int J Environ Res Public Health*. Mar 10 2022;19(6)doi:10.3390/ijerph19063245

32. Ma J, Yin H, Hao X, Sha W, Cui H. Development of a random forest model to classify sarcoidosis and tuberculosis. *Am J Transl Res*. 2021;13(6):6166-6174.

33. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, California, USA. <https://doi.org/10.1145/2939672.2939785>

34. Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med*. Dec 7 2020;18(1):462. doi:10.1186/s12967-020-02620-5

35. Wang R, Zhang J, Shan B, He M, Xu J. XGBoost Machine Learning Algorithm for Prediction of Outcome in Aneurysmal Subarachnoid Hemorrhage. *Neuropsychiatr Dis Treat*. 2022;18:659-667. doi:10.2147/ndt.S349956

36. Davagdorj K, Pham VH, Theera-Umpon N, Ryu KH. XGBoost-Based Framework for Smoking-Induced Noncommunicable Disease Prediction. *Int J Environ Res Public Health*. Sep 7 2020;17(18)doi:10.3390/ijerph17186513

37. Schaefer JD, Caspi A, Belsky DW, et al. Enduring mental health: Prevalence and prediction. *J Abnorm Psychol*. Feb 2017;126(2):212-224. doi:10.1037/abn0000232

38. Kubiak SP, Beeble M, Bybee D. Comparing the validity of the K6 when assessing depression, anxiety, and PTSD among male and female jail detainees. *Int J Offender Ther Comp Criminol*. Dec 2012;56(8):1220-38. doi:10.1177/0306624x11420106

39. Kroenke K, Strine TW, Spitzer RL, Williams JB, Berry JT, Mokdad AH. The PHQ-8 as a measure of current depression in the general population. *J Affect Disord*. Apr 2009;114(1-3):163-73. doi:10.1016/j.jad.2008.06.026

**Tables and Figures**

**Table 1. Univariate Analysis of Commonly Selected Features**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **N (%), Mean (Median)** | | |
| **Variable Name** | *2018 (n=10076)* | *2019 (n=8803)* | *2020 (n=8781)* |
| Currently have depression  Yes  No | 955 (9.48%)  9121 (90.52% | 463 (5.26%)  8340 (94.74%) | 516 (5.88%)  8265 (94.12%) |
| General Health Status  Excellent  Very Good  Good  Fair  Poor | 1547 (15.35%)  2644 (26.24%)  3223 (31.99%)  2045 (20.30%)  617 (6.12%) | 1431 (16.26%)  2486 (28.24%)  2711 (30.80%)  1676 (19.04%)  499 (5.67%0 | 1791 (20.40%)  2700 (30.75%)  2590 (29.50%)  1361 (15.50%)  339 (3.86%) |
| Age  18-24  25-29  30-44  45-64  65+ | 724 (7.19%)  689 (6.84%)  2348 (23.30%)  3407 (33.81%)  2908 (28.86%) | 654 (7.43%)  726 (8.25%)  2409 (27.37%)  2835 (32.20%)  2179 (24.75%) | 706 (8.04%)  733 (8.35%)  2456 (27.97%)  2965 (33.77%)  1921 (21.88%) |
| Birth Sex  Male  Female | 4338 (43.05%)  5738 (56.95%) | 3855 (43.79%)  4948 (56.21%) | 3850 (43.84%)  4931 (56.16%) |
| Race  White, non-Hispanic  Black, non-Hispanic  Hispanic  Asian/PI, non-Hispanic  Other, non-Hispanic | 3484 (34.58%)  2300 (22.83%)  2895 (28.73%)  1126 (11.18%)  271 (2.69%) | 3056 (34.72%)  1943 (22.07%)  2411 (27.39%)  1127 (12.80%)  266 (3.02%) | 2859 (32.56%)  1837 (20.92%)  2457 (27.98%)  1340 (15.26%)  288 (3.28%) |
| Marital Status  Married  Divorced  Widowed  Separated  Never married  Unmarried but living with partner | 3886 (38.57%)  1252 (12.43%)  985 (9.78%)  494 (4.90%)  3009 (29.86%)  450 (4.47%) | 3399 (38.61%)  1044 (11.86%)  742 (8.43%)  410 (4.66%)  2653 (30.14%)  555 (6.30%) | 3359 (38.25%)  961 (10.94%)  613 (6.98%)  396 (4.51%)  2850 (32.46%)  602 (6.86%) |
| Type of Health Insurance  Employer  Self-purchase  Medicare  Medicaid/Family Health+  Milit/CHAMPUS/Tricare/COBRA  Uninsured | 3638 (38.57%)  605 (6.00%)  2285 (22.68%)  2297 (22.80%)  363 (3.60%)  888 (8.81%) | 3433 (39.00%)  516 (5.86%)  1696 (19.27%)  1970 (22.38%)  304 (3.45%)  884 (10.04%) | 3577 (40.74%)  478 (5.44%)  1504 (17.13%)  1953 (22.24%)  339 (3.86%)  930 (10.56%) |
| Employment  Employed  Self-Employed  Unemployed, 1+ years  Unemployed, <1 year  Homemaker  Student  Retired  Unable to work | 4554 (45.20%)  918 (9.11%)  300 (2.98%)  312 (3.10%)  511 (5.07%)  358 (3.55%)  2262 (22.45%)  861 (8.55%) | 4305 (48.90%)  829 (9.42%)  246 (2.79%)  306 (3.48%)  415 (4.71%)  355 (4.03%)  1721 (19.55%)  626 (7.11%) | 4007 (45.63%)  773 (8.80%)  279 (3.18%)  893 (10.17%)  355 (4.04%)  389 (4.43%)  1484 (16.90%)  601 (6.84%) |
| Highest level of Education  Less than HS  High School grad  Some college  College grad | 1555 (15.43%)  2148 (21.31%)  2127 (21.11%)  4246 (42.14%) | 1152 (13.09%)  1779 (20.21%)  1855 (21.07%)  4017 (45.63%) | 1223 (13.93%)  1809 (20.60%)  1691 (19.26%) 4058 (46.21%) |
| Nutrition—cups of fruits and vegetables eaten yesterday  0-50 Range of Values | 2.597 (2) | 2.744 (2) | 2.336 (2) |
| Needed medical care in the past 12 months and did not get it  Yes  No | 1062 (10.54%)  9014 (89.46%) | 1044 (11.86%)  7759 (88.14%) | 1047 (11.92%)  7734 (88.08%) |
| Exercise in the past 30 days  Yes  No | 7299 (72.44%)  2777 (27.56%) | 6610 (75.09%)  2193 (24.91%) | 6370 (72.54%)  2411 (27.46%) |
| Sleep Quality in the past 30 days  Very good  Fairly good  Fairly bad  Very bad | 1402 (13.91%)  7654 (75.96%)  735 (7.29%)  285 (2.83%) |  |  |
| Had enough food accessible in the past 6 months  Had enough, wanted to eat  Had enough, not always wanted  Sometimes not enough  Often not enough | 7208 (71.54%)  1996 (19.81%)  653 (6.48%)  219 (2.17%) |  |  |
| Could not afford medical care in the past 12 months  Yes  No | 843 (8.37%)  9233 (91.63%) |  |  |
| Sexual Assault victim ever—rape  Yes  No |  | 686 (7.79%)  8117 (92.21%) |  |
| In the past year, experienced cognitive or memory issues that are getting worse?  Yes  No |  | 746 (8.47%)  8057 (91.53%) |  |
| Difficulty performing daily acts due to physical, mental, or emotional problems  Yes  No |  | 1489 (16.91%)  7314 (83.09%) | 1161 (13.22%)  7620 (86.78%) |
| Average # of sodas and sweetened drinks per day  None  <1  1  >1 |  |  | 4156 (47.33%)  2928 (33.34%)  492 (5.60%)  1205 (13.72%) |
| Two or more sodas consumed each day  Yes  No |  |  | 360 (4.10%)  8421 (95.90%) |
| In past 12 months, unable to pay rent/delayed in paying rent?  Yes  No |  |  | 1365 (15.54%)  7416 (84.46%) |
| Heavy drinker (>2 drinks a day for men, >1 drink a day for women)  Yes  No |  |  | 513 (5.84%)  8368 (95.30%) |
| Women who have exclusively had sex with women in the past 12 months  Yes  No |  |  | 84 (0.96%)  8697 (99.04%) |
| Experience intimate partner violence from current or past partner through insults/controlling behavior  Yes  No |  |  | 1193 (13.59%)  7588 (86.41%) |

**Table 2. AU-ROC Curve Values for Years 2018, 2019, 2020**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **AU-ROC Value By Year** | | |
| **Learning Model** | *2018* | *2019* | *2020* |
| Logistic Regression | **0.864** | 0.845 | **0.815** |
| Naïve Bayes | 0.858 | 0.842 | 0.805 |
| K-Nearest Neighbors | 0.763 | 0.751 | 0.744 |
| Decision Trees | 0.787 | 0.765 | 0.720 |
| Random Forest | 0.833 | 0.836 | 0.809 |
| XGBoost | 0.855 | **0.852** | 0.792 |

**Table 3. Classification Accuracy By Algorithm for Years 2018, 2019, 2020**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy Value By Year** | | |
| **Learning Model** | *2018* | *2019* | *2020* |
| Logistic Regression | 0.797 | 0.835 | 0.782 |
| Naïve Bayes | 0.811 | 0.885 | 0.861 |
| K-Nearest Neighbors | 0.820 | 0.806 | 0.855 |
| Decision Trees | 0.719 | 0.863 | 0.855 |
| Random Forest | **0.853** | **0.935** | **0.912** |
| XGBoost | 0.804 | 0.877 | 0.842 |

**Figure 1a: ROC Curves of Supervised Learning Models for 2018**

**A graph of different colored lines

Description automatically generated**

**Figure 1b: ROC Curves of Supervised Learning Models for 2019**

**A graph of a curve

Description automatically generated**

**Figure 1c: ROC Curves of Supervised Learning Models for 2020**

**A graph of a curve

Description automatically generated**

**Figure 2a: Importance Plots for Classification of Depression, 2018**

**A graph of a number of variables

Description automatically generated with medium confidence**

**Figure 2b: Importance Plots for Classification of Depression, 2019**

A graph with purple bars

Description automatically generated

**Figure 2c: Importance Plots for Classification of Depression, 2020**

A graph with blue bars

Description automatically generated