# Depression Level Classification in NHANES dataset

Qiaochu Chen, Menghui Sun, and Ruixuan Wang

## Abstract

**Objective** Major Depressive Disorder is a prevalent mental disease but often goes undetected, resulting in treatment delays. This project aims to compare the performance of different machine learning methods in screening depression and figure out the essential non-psychiatric factors for predicting depression. **Methods** A total of 26,175 adult records collected between 2005 and 2018 from NHANES were included in the project, with 22 demographic, nutrition, health-related risk factors and chronic disease features. A PHQ-9 score assesses the status of depression. Three methods to deal with imbalanced data, i.e. Stratified, Down-sampling and SMOTE, were applied. Six machine learning classifiers, i.e. Logistic Regression, Naive Bayes, eXtreme Gradient Boosting(XGBoost), Support Vector Machine(SVM), K Nearest Neighbors(KNN), and Random Forest, were evaluated using R. Among them, the two methods based on using a multitude of decision trees, XGBoost and Random Forest, are used to assess the feature importance. **Results** 1. XGBoost appeared to be the best-performing model with the highest measures in AUC (0.80596/0.809466) under Stratified and Down-sampling, and SVM was close in performance to XGBoost under all sampling methods. 2. General health conditions are the best predictive factor; age, income, marital status and employment are also crucial in prediction. **Conclusion** Our results indicate that applying machine learning methods to screening depression is feasible with moderate performance. The model can be used for joint screening and detecting depression in communities or by non-psychiatric physicians.

## Introduction and problem statement

Major Depressive Disorder(MDD) is one of the most common mental disorders in the United States. According to the CDC report, during 2013–2016, 8.1% of adults in America had depression in a given 2-week period, which indicates a high prevalence of depression across the country[1]. Depression is usually characterized by hopelessness, anxiety, tiredness and lack of energy. Further studies showed that depression rates and severity are significantly associated with non-psychiatric factors, such as gender, age, nutrition, obesity, race, and chronic disease[3]. In terms of diagnosing whether a patient has an MDD, the Patient Health Questionnaire (PHQ), a self-administered version of the PRIME-MD diagnostic instrument for common mental disorders, is usually used. Specifically, PHQ-9 allows for criteria-based diagnoses of depression such that a PHQ-9 score $\geq 10$ is clustered as MDD, while $< 10$ indicates no MDD[2]. Therefore, our primary goal was to build a predictive machine learning model using non-psychiatric factors to predict the diagnosis of MDD. The model aids in screening depression in the community without using psychiatric factors and understanding which non-psychiatric factors play an essential role in predictions.

## Methods

**Data overview**: Based on the motivation, we decided to use the public dataset, the National Health and Nutrition Examination Survey (NHANES), initiated in 1999, to determine the prevalence of illnesses and their risk factors in the United States. The NHANES interview includes demographic, socioeconomic, dietary, and health-related questions, such as PHQ. The examination consists of medical, dental, and physiological measurements and laboratory tests administered by highly trained medical personnel. The noninstitutionalized U.S. civilian population of all ages living in all 50 states and Washington, D.C., is the sample population for the NHANES. Each year, the poll evaluates a sample of 5,000 nationally representative people[4]. Specifically, we formed a study population using the participants in the PHQ-9 survey. Their corresponding diagnosis based on PHQ-9 scores were obtained as the true labels in this study. Then, their non-psychiatric factors

were collected and applied to several machine learning models to predict whether they were diagnosed with MDD. Compare the predictions with the true labels to determine the most appropriate predictive model. Therefore, the study population contains 26,175 adult records (2005-2018) in the United States with their non-psychiatric 22 features/predictors (10 continuous and 12 categorical variables) in four perspectives: (i) Demographics: Age (>20), Gender, Family income, Education level, Marital Status, Race/Ethnicity, Health insurance, Occupation, Total number of people in the household. (ii) Nutrition: Carbohydrate, Protein, Polyunsaturated fatty acid (PUFA), Folic acid, Iron. (iii) Health-Related Risk Factors: Body Mass Index (BMI), Tobacco use, General health condition, Vigorous work activity. (iv) Chronic Disease: Hypertension, High cholesterol, Diabetes, Asthma. These variables are selected based on the data integrity in the NHANES database and literature review. The outcome of this study, the depression level, is measured by each subject's PHQ-9 score.

**Pre-processing**: ***Missing values***: 375 subjects did not have a complete PHQ-9 questionnaire, which makes it impossible to calculate a valid PHQ-9 score, so they were excluded from the study. Hence the population narrowed down to 25,800 adults. ***Outcome reconstruction***: We used 10 as the cutoff score on the PHQ-9 score to label whether a subject has an MDD [2]. We define a score less than 10 as no MDD (labeled as 0, the negative class) and $\geq 10$ as having a disorder (labeled as 1, the positive class). Hence, the outcome becomes a binary variable. ***Variable correlation***: Pearson's correlation coefficients between each pair of predictors are less than 0.9; that is, there are no highly correlated variables[5]. We assumed that we have independent predictors for fitting the models. ***One-hot encoding***: One-hot encoding converts categorical variables into a form we can fit into the machine learning models. Specifically, we convert each categorical variable into a new categorical column with one-hot and assign a binary value of 1 or 0 to those columns. Therefore, the original 12 categorical variables are expanded to 28 dummy numeric variables upon completing one-hot encoding. Up to now, all variables are numeric (38 in total), which can fit into all our proposed methods. ***Data standardization***: The ranges and scales of input of numeric variables vary greatly. For example, the range of carbohydrates in dietary intake is between 0 and around 1815 gm, but the total number of people in households varies from 1 to 7. Therefore, we standardize input variables by the z-score method to ensure the data's internal consistency and avoid errors and troubles during the model training process. ***Data splitting and imbalance dataset***: It is noticeable that the dataset was imbalanced as there are 23,593 (91.46%) subjects who do not have an MDD and 2,207 (8.54%) subjects with an MDD. Therefore, the random sampling technique is inappropriate for this data set. It is highly likely that the resulting training set only contains records from the majority class, which is useless for model learning. Hence, we first stratified sampling with 80% as training and 20% as testing to keep the data distribution. Then, utilized two resampling techniques to deal with the imbalance issue in training and compared it with the original stratified data in terms of model performance (i) Applied down-sampling on the stratified training set yielding the majority class an equal amount to the minority class. (ii) Applied Synthetic Minority Oversampling Technique (SMOTE) on the stratified training set, which yields the minority class an equal amount as the majority class. For consistency in comparison and reliability of the testing, we used the same stratified testing set for evaluating the model fitting without any transformation (no down-sampling and oversampling).

**Model Analysis**:

***Proposed models***: We used supervised machine learning models for this study because of the known true labels of each subject. The proposed models include Logistic Regression, Naive Bayes classifier, eXtreme Gradient Boosting(XGBoost), Support Vector Machine(SVM), K Nearest Neighbors(KNN), and Random Forest. One major consideration for KNN is choosing the number of the nearest neighbors, k. To determine the optimal k, we applied 5-fold cross-validation to tune this parameter when training the models. We picked the one with the highest AUC score for building the final model and evaluating it with the testing set. Similarly, we tune the two essential parameters in the Random Forest: the number of input features a decision tree has available to consider at any given point in time (mtry) and the number of decision trees that are combined for final prediction (ntree)[7-8]. We also tune the number of decision trees in the final model (nrounds) and the max depth of a decision tree (max depth) for XGBoost[9-10]. The results of parameter tuning are shown in **Table.1**.

***Performance Metrics***: Because of the imbalanced dataset, mean accuracy is a misleading metric for model evaluation. Even if all records are naively predicted as the majority class, there is still an excellent accuracy

of 91.46%. This situation is called the accuracy paradox[6]. According to the prediction results on the testing set, we comprehensively compare different classification evaluation metrics between the six proposed models. These metrics are the F1 score, balanced accuracy and the area under the ROC curve (AUC). Balanced accuracy and AUC are sufficient for an imbalanced dataset to evaluate a classification machine learning model. However, another performance metric was argued to be effective for an exceedingly imbalanced dataset, which explains why F1 scores are used[11]. All works were done by R 4.2.1.

## Results

Table 1: Hyper-parameters selection using 5-fold cross validation

| | KNN | Random Forest | | XGBoost | |
|---|---|---|---|---|---|
| | k | mtry | ntree | nrounds | max depth |
| Stratified-sampling | 93 | 5 | 500 | 250 | 1 |
| Down-sampling | 58 | 6 | 700 | 150 | 1 |
| SMOTE | 54 | 5 | 650 | 400 | 10 |

Table 2: Model performance comparison on testing set

| Model | Stratified-sampling | | | Down-sampling | | | SMOTE | | |
|---|---|---|---|---|---|---|---|---|---|
| | F1 score | Balanced Accuracy | AUC area | F1 score | Balanced Accuracy | AUC area | F1 score | Balanced Accuracy | AUC area |
| Logistic regression | 0.1737 | 0.5476 | 0.5476 | 0.3522 | 0.7152 | 0.7152 | 0.3599 | 0.7173 | 0.7173 |
| Naive Bayes | 0.2883 | 0.5761 | 0.6507 | 0.2610 | 0.5592 | 0.6721 | 0.2454 | 0.5523 | 0.6580 |
| XGBoost | 0.2235 | 0.5642 | 0.8060 | 0.3272 | 0.7356 | 0.8095 | 0.2138 | 0.6369 | 0.7055 |
| SVM | 0.0224 | 0.5056 | 0.7124 | 0.3181 | 0.7319 | 0.8004 | 0.2973 | 0.6142 | 0.7408 |
| KNN | 0.0090 | 0.5000 | 0.5034 | 0.1283 | 0.4910 | 0.4910 | 0.2654 | 0.7043 | 0.7045 |
| Random Forest | 0.1011 | 0.8123 | 0.5262 | 0.3210 | 0.5862 | 0.7304 | 0.2787 | 0.5869 | 0.6177 |

**Table.2** compares the performance metrics of the six proposed models developed on Stratified-Sampling, Down-Sampling and SMOTE-Sampling, respectively. F1 score, balanced accuracy were obtained based upon the confusion matrix with a fixed threshold, and AUC evaluated that classifier over all possible thresholds.

Overall speaking, the optimal performance is attributable to the XGBoost model. It outperforms other models with the highest measures in AUC (0.80596/0.809466) under Stratified and Down-sampling, Balanced Accuracy (0.735565) under Down-sampling, and F1 score (0.327179) under Down-sampling. The SVM model was close in performance to XGBoost under all sampling methods. Compared to SMOTE, Down-sampling did exert a more significant effect on the performance of SVM and XGBoost, with an observed increase of 20% in balanced accuracy for both and around 10% in AUC for SVM.

The following non-trivial results can be found in **Table.2** as well. Naive Bayes and XGboost gave decent AUC but failed to achieve acceptable balanced accuracy rates among stratified sampling. Additionally, Random Forest had a plausibly high balanced accuracy, which could originate from the incompetence of Balanced Accuracy and AUC to distinguish between specific types of errors[11]. Among SMOTE sampling, Logistic regression presented the highest balanced accuracy, while SVM generated the highest AUC.

Variable significance was assessed to determine how each predictor variable affects the model's accuracy. Each feature plays a different role during the decision tree construction in XGBoost. The importance is calculated for a single decision tree by the amount that each attribute split point improves the performance measure, weighted by the number of observations the node is responsible for[12]. **Fig.1** displays the comparative importance of 38 variables in the Down-sampling XGBoost model, scoring from 0 to 100. The feature with the highest importance is general health condition(HSD010), a self-reported feature. Studies before have reported depression symptoms are positively associated with bad global health[13]. Another reasonable explanation is that, unlike the other features, general health condition is the only subjective feature without numeric boundaries for ranking, which may reflect the moods or attitudes toward health and living conditions. Other features with high importance include age (RIDAGEYR), the ratio of family income to poverty (INDFMPIR), married marital status (DMDMARTL_1) and having work or business last week (OCD_150). A general explanation for the importance of these features is that depression reaches its lowest point at middle age

and will then increase with age due to retirement, economic hardship and loss of marriage[14]. Nutrition factors seem to play a certain role, with 4 predictors ranking 8-11 in the total 38 variables, which indicates the necessity of including nutrition factors in depression prediction.

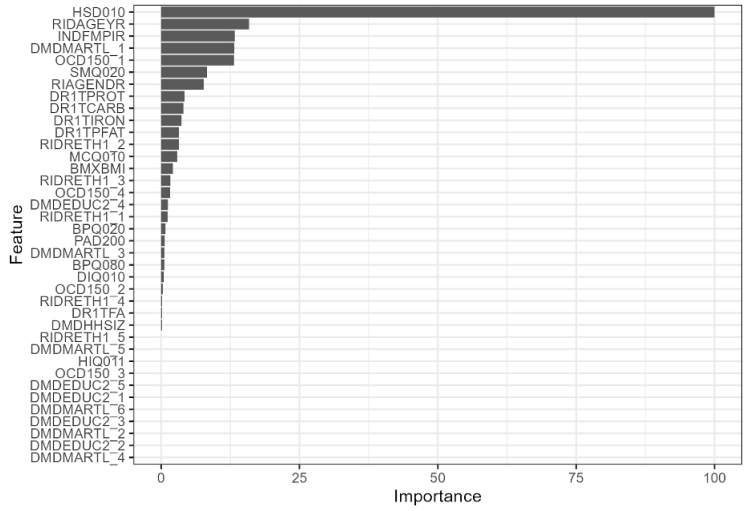


Figure 1: Figure 1. Feature Importance Selected by XGBoost (Down-sampling Data).

## Conclusion

This study was primarily aimed at developing and evaluating the predictive power of various machine learning methods. Our findings show that the machine learning model based on the XGBoost approach is able to reach the highest accuracy in predicting whether or not patients have major depressive disorder. Another objective of this study was to explore the predictability of non-psychiatric factors in diagnosing and classifying the depression cases from healthy cases. The most important feature driving the prediction of depression is the self-rated general health conditions. Age, income, marital status and employment are also indicative of depression, consistent with other studies. A possible real-world application of this study can be in the form of joint screening, where our model can be used by community healthcare providers and non-psychiatrist hospital departments cooperatively to improve the accuracy of depression recognition by non-psychiatrist physicians and prevent depression in the early stage[15].

## Limitation and future work

**Variation in Features**: There are thousands of features in the NHANES database, while the dataset we used in the project is only a small subset of the total data. We chose the features based on biomedical experiments and previous studies, which may cause a loss of the subjects' information. More highly correlated and informative non-psychiatric factors can be included to achieve a more reasonable ratio of the number of features to sample size[16]. **Stacking Model**: F1 scores were undesirable within the performance evaluation results we obtained. Using an ensemble machine learning algorithm called stacking could be a worthwhile future direction, which discovers the optimal way to combine the predictions from various well-performing machine learning models. The advantage of stacking is that it can use a variety of effective models to accomplish classification or regression tasks and produce predictions that perform better than any individual model in the ensemble[17]. Prior studies implied that the stacking method may have a higher true positive rate, lower false positive rate and higher F1 score than traditional learning methods[18]. **Resampling Techniques**: Although the ensemble resampling method like SMOTE was used to handle imbalanced issues in this study, We had to admit that SMOTE comes with a concerning nature: while generating synthetic examples, SMOTE does not weigh neighboring examples that can be from other classes. This can increase classes overlapping and can introduce additional noise[19]. To address this, modifications on SMOTE and more resampling methods can be taken into consideration for comparison purposes such as Up-Sampling,

Rose-Sampling, etc. **More Refined Classification**: Future studies should employ machine learning to analyze the forms of depression carefully. In this study, we only focused on predicting the occurrence of depression cases rather than trying to diagnose which category the suspected cases belonged in precisely.

## Contribution

**Qiaochu Chen**: Model fitting for Logistic Regression and Naive Bayes, data processing, presentation slides editing and report writing. **Menghui Sun**: Model fitting for Random Forest and KNN, data processing, presentation slides editing and report writing. **Ruixuan Wang**: Model fitting for XGBoost and SVM, data processing, presentation slides editing and report writing.

## Reference


[1] Brody, D. J., Pratt, L. A., & Hughes, J. P. (2018). Prevalence of Depression Among Adults Aged 20 and Over: United States, 2013-2016. NCHS data brief, (303), 1–8.
[2] Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. Journal of general internal medicine, 16(9), 606–613.
[3] Rao, T. S., Asha, M. R., Ramesh, B. N., & Rao, K. S. (2008). Understanding nutrition, depression and mental illnesses. Indian journal of psychiatry, 50(2), 77–82.
[4] CDC. National Health and Nutrition Examination Survey. URL:https://www.cdc.gov/visionhealth/vehss /data/national-surveys/national-health-and-nutrition-examination-survey.html.
[5] Chen, C.F. and Rothschild, R. (2010), "An application of hedonic pricing analysis to the case of hotel rooms in Taipei", Tourism Economics, Vol. 16 No. 3, pp. 685-694.
[6] Wikipedia. Accuracy paradox. url: https://en.wikipedia.org/wiki/Accuracy_paradox.
[7] Ellis, C. Between academic research experience and industry experience (2022) Mtry in random forests, Crunching the Data. Available at: https://crunchingthedata.com/mtry-in-random-forests/.
[8] Ellis, C. Between academic research experience and industry experience (2022) Number of trees in random forests, Crunching the Data. https://crunchingthedata.com/number-of-trees-in-random-forests.
[9] He, T. (2016, March 10). Tong he. XGBoost, https://xgboost.ai/rstats/2016/03/10/xgboost.html
[10] XGBoost parameters (no date) XGBoost Parameters - xgboost 2.0.0-dev documentation, https://xgboos t.readthedocs.io/en/latest/parameter.html.
[11] Korstanje, J. (2021, August 31). The F1 score. Medium. Retrieved December 16, 2022, from https: //towardsdatascience.com/the-f1-score-bec2bbc38aa6
[12] https://notebook.community/minesh1291/MachineLearning/xgboost/feature_importance_v1
[13] Castro-Costa, E., Lima-Costa, M. F., Carvalhais, S., Firmo, J. O., & Uchoa, E. (2008). Factors associated with depressive symptoms measured by the 12-item General Health Questionnaire in community-dwelling older adults (The Bambuí Health Aging Study). Revista brasileira de psiquiatria (Sao Paulo, Brazil : 1999), 30(2), 104–109.
[14] Mirowsky, J., & Ross, C. E. (1992). Age and Depression. Journal of Health and Social Behavior, 33(3), 187–205.
[15] Cepoiu M, McCusker J, Cole MG, Sewitch M, Belzile E, Ciampi A. Recognition of depression by non-psychiatric physicians–a systematic literature review and meta-analysis. J Gen Intern Med. 2008 Jan;23(1):25-36. doi: 10.1007/s11606-007-0428-5. Epub 2007 Oct 26. PMID: 17968628; PMCID: PMC2173927.
[16] Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. R. (2005, April 15). Optimal number of features as a function of sample size for various classification rules. OUP Academic. Retrieved December 16, 2022, from https://academic.oup.com/bioinformatics/article/21/8/1509/249540
[17] Brownlee , J. (2021, April 27). Stacking Ensemble Machine Learning With Python. Machine Learning Mastery. https://ttps://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python//
[18] Li, Zhang, Liu, Zhang, Zhao, Gong, & Fu. (2022, March 2). Developing stacking ensemble models for multivariate contamination detection in water distribution systems. Developing Stacking Ensemble Models for Multivariate Contamination Detection in Water Distribution Systems - ScienceDirect. https: //www.sciencedirect.com/science/article/abs/pii/S0048969722013766
[19] Paul, S. (2018, October). Diving Deep with Imbalanced Data. Data Camp. Retrieved December 16, 2022, from https://www.datacamp.com/tutorial/diving-deep-imbalanced-data