Evaluation of Different Imputation and Machine Learning Techniques for Prediction of Depression

Caroline Sklaver
Rough Draft - 07/02/2020
George Washington University Department of Data Science

1. INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC), about 1 in every 6 adults in the US will experience depression during their lifetime, which is about 16 million adults each year. Depression is characterized as a long-term affect in mood that can take form in a variety of symptoms such as feeling sad, loss of interest in activities, changes in appetite and sleeping patterns, increased fatigue, feeling worthless, difficulty concentrating, or thoughts of suicide ("What Is Depression?," 2017). Depression is caused by a complex combination of genetic, physiological, environmental, and social factors and is a leading cause of disability and major contributor to the global burden of disease ("Depression," 2020). Depression is more common in women and is associated with physical comorbidities such as cardiovascular disease, stroke, cancer and other chronic conditions as well as mental health comorbidities such as anxiety disorder, and alcoholism (Belmaker & Agam, 2008).

Despite the widespread prevalence of depression, it frequently goes undiagnosed and untreated (Majidi, Khadembashi, Etemad, Jafari, & Khodakarim, 2019), contributing to the overall burden of disease. Previous research has shown associations between depression and dietary practices including caffeine consumption (Iranpour & Sabour, 2019), physical activity, metabolic syndrome, BMI (Liu, Ozodiegwu, Yu, Hess, & Bie, 2017), and other chronic conditions such as asthma, kidney disease (Patel, Patel, & Baptist, 2017), , osteoporosis (Cizza, Primma, & Csako, 2009), and more. These associations were found mainly using linear regression analysis to identify single variable correlations with depression. Machine learning algorithms have been used to classify the severity of depression (Kessler et al., 2016) as well as therapeutic outcomes (Lee et al., 2018). Kessler et al. 2016 used ensemble regression trees and 10-fold cross-validated penalized regression to predict the severity and persistence of patients with major depressive disorder. Meanwhile other uses of machine learning algorithms include classification algorithms to predict therapeutic outcomes (Chekroud et al., 2016; Lee et al., 2018) and deep learning models used on neuroimaging data to identify potential depression biomarkers (Gao, Calhoun, & Sui, 2018).

1.1 Problem Statement

The literature fails to predict depression within the population using both demographic and physiological indicators. In this project, I aim to use National Health and Nutrition Examination Survey (NHANES) data to build a model that can predict depression among adults in the US. This data has been used in machine learning projects that successfully used shallow convolutional networks to predict Coronary Heart Disease (Dutta, Batabyal, Basu, & Acton, n.d.) and both support vector machines and ensemble modeling to predict risk of Diabetes (Semerdjian & Frank, 2017). Additionally, this project includes a component of data imputation techniques. NHANES data has many inconsistencies and missing values. This project will

explore different approaches handling missing values and will evaluate model performance using each method.

2. MATERIALS & METHODS

2.1 NHANES Data

The CDC's National Center for Health and Statistics (NCHS) provided NHANES data every two-years dating back to 1999. The NHANES is a stratified, multistage probability sample of the civilian, non-institutionalized US population amounting to about 5,000 participants each two-year period (NHANES overview). This survey is unique in that it combines both interviews and physical examinations divided into 4 categories as outlined in Figure 1. For this project, I have included data from 2005-2016 surveys as they are the most heterogenous and complete with respect to the depression questionnaire portion. Data was downloaded from the CDC website and converted from XPT to CSV files using python script found on GitHub (Wyatt, 2016).

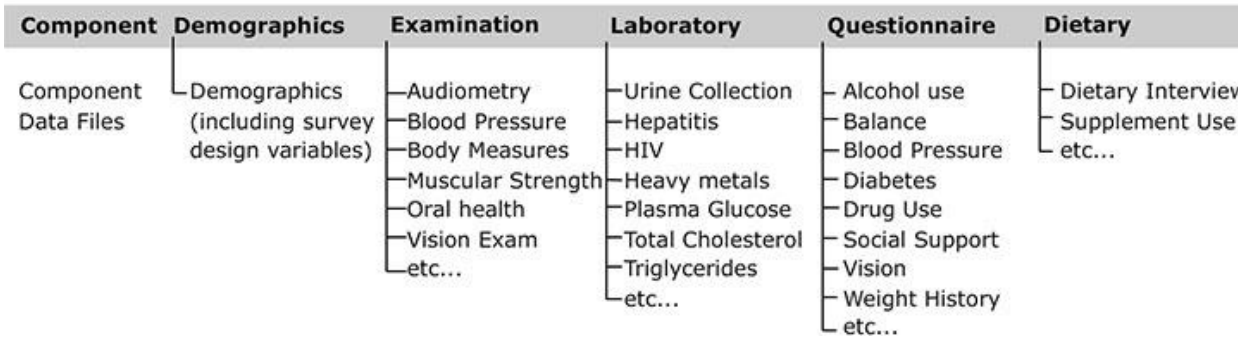

Fig. 1: Overview of NHANES Components
Retrieved from https://wwwn.cdc.gov/nchs/nhanes/tutorials/module1.aspx

Within the Questionnaire lies the 'Mental Health – Depression Screener' which is comprised of the Patient Health Questionnaire nine-item (PHQ-9) depression screening instrument (Kroenke, Spitzer, & Williams, 2001) that assesses the frequency of depression symptoms over the past 2 weeks. Response categories include "not at all," "several days," "more than half the days," and "nearly every day" given a point ranging from 0 to 3. A total score was calculated ranging from 0 to 27, with a score of 10 or higher marked as 'depressed'. This method using the PHQ-9 survey has been proven extensively in the literature and is commonly used in clinical studies to define depression ("NHANES 2009-2010: Mental Health - Depression Screener Data Documentation, Codebook, and Frequencies," n.d.). Only data with complete PHQ-9 questionnaires were used in this project.

Features of this data frame were chosen based off prior associations found in the literature and are outlined in Table 1 of the Appendix. The complete data is 31,000 observations with 55 columns including respondent sequence number, and survey year. Responses of '1 – yes' and '2 – no' were converted to binary 1/0, respectively. All responses of '7 – don't know' or '9 – refused' were marked as an N/A value.

2.2 Data Imputation

Missing data was removed and imputed using two different methods: statistical and machine learning. The statistical methods include simple mean, median, and mode missing value imputation. Statistical imputations were implemented after dropping columns with a proportion of missing values exceeding an input threshold. Mean and median are simple imputations in which the missing values of continuous variable are replaced by the mean or median value of that variable. With discrete variables, missing values were imputed using the most frequent value or mode of that variable. Mean and median imputations with both 75% and 50% thresholds were evaluated.

The machine learning imputation methods aim to use available information within the dataset to estimate and substitute the missing values. In this way, I separated the values into complete data and incomplete data with missing values. The complete data (excluding the ultimate target variable 'depressed', ensuring the imputed variables are independent of the dependent target variable) was used to predict values of the next-most-complete column. These predicted values were then imputed into the column and used with the complete data to predict the missing values of the subsequent column. Thus, progressively predicting missing values with the imputed data. The machine learning models I have implemented in this section are multi-layer perceptron (MLP) k-nearest neighbors (KNN).

2.2.1 MLP

MLP is a supervised machine learning model that can estimate a non-linear function using multiple layers of computational units connected in a feed-forward way (Jerez et al., 2010; Pedregosa et al., 2011). The input layer is a set of neurons representing the input features and each of these neurons pass through hidden layers, which transforms the values from the previous layer using a weighted linear summation and a non-linear activation function. Then, the output layer transforms these values from the last hidden layer into output values (Pedregosa et al., 2011). MLP can be used for classification and regression, both using backpropagation (a form of gradient-descent). In classification, the MLP loss function is Cross-Entropy, while in regression MLP uses the Square Error loss function. In binary classification MLP uses the logistic function in the output layer, but if there are more than two classes, MLP uses the SoftMax function in the output layer (Pedregosa et al., 2011).

In this data imputation approach, MLP uses the different error and output functions described above depending on the type of target variable (continuous, binary, or multi-class). A simple architecture consisting of two hidden layers was used in this process.

2.2.2 KNN

The KNN algorithm predicts target values based on its resemblance to values in the training set (Kana, 2020). In comparison to the MLP method, KNN does not use the entire data, but rather uses the majority vote of *k* closest cases that are not missing values in the variable to be imputed (Jerez et al., 2010). These *k* cases are found by minimizing a distance measure (Manhattan, Euclidian, or Minkowski distance). The KNN model used for imputation used *k=5* with uniform weights and Euclidian distance to progressively predict and impute missing values.

2.3 Predictive Analysis

In classifying depression, six supervised machine learning algorithms were applied: Logistic Regression, Naïve Bayes, Decision Tree, Random Forest, Extreme Gradient Boosting, and MLP. Logistic regression was applied as a baseline standard statistical tool (Razzak, Imran, & Xu, 2020). MLP was implemented in two ways, using Scikit Learn and Keras packages. Each algorithm will be briefly outlined in this section, including their corresponding hyper-parameters. MLP was described in section 2.2.1 and has hyperparameters of learning rate schedule for weight updates and L2 penalty (regularization term) parameter. This will be followed by an explanation of the implementation of these models in this project.

2.3.1 Naïve Bayes

Naïve Bayes is a supervised learning algorithm based on Bayes' theorem that assumes all features are independent of each other and each pair of features are independent of each other given the value of the class variable. Naïve Bayes uses maximum likelihood for parameter estimation and is a simple, fast model. While it has been used for disease prevention (Razzak et al., 2020), the major limitation lies in the feature-independence assumption, thus it is sensitive to highly correlated features.

2.3.2 Decision Tree

Decision Tree (DT) is a model that predict the target value by using tree-like graphs based on sorting and learning decision rules from the data features. Each node represents the feature to be classified, while each branch is the value that node can assume. DTs can be used for both classification and regression analysis and are commonly used in disease prevention (Razzak et al., 2020). Hyper-parameters of DT models are minimum sample splits and minimum sample leaves.

2.3.3 Random Forest

Random Forest (RF) is a supervised machine learning algorithm based on ensemble learning. A RF classifier fits a number of DTs to a sample of the training data and uses majority vote to get predictions. RF uses averaging to improve the accuracy of prediction and to limit over-fitting. RF have proven to be successful in predicting Parkinson's Disease (Geetha, Professor, Head, & Sivagami, 2011; Razzak et al., 2020). Hyper-parameters in RF are the same as that of DT.

2.3.4 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a boosting model that utilizes tree ensembles at its base. Boosting is a process of improving prediction accuracy by assigning weights to previous outcomes, beginning with a weak learner. XGBoost grows classification trees and attempts to minimize the misclassification rate in subsequent trees. This is a powerful method in machine learning. Hyperparameters include step size, minimum loss reduction, and L2 regularization term on weights.

2.3.5 Keras Neural Network

This analysis also included an artificial neural network built using python keras package. The model is a sequential model with three dense layers. The activation function of the first two layers is ReLU and the last being sigmoid. Binary cross entropy is used to evaluate the model performance.

In this project, in order to reach the optimal solution, grid search was used for hyper-parameter tuning of each model. All training data was standardized and split into 80% training 20% testing data. Training data was then split into a 20% validation set, which was used in evaluation. For each of the models and set of parameters, evaluation was computed using 10-fold cross-validation. Each model and hyper-parameter tuning was built and evaluated for each data imputation technique. F1-micro score is used for evaluation. F1-score conveys the balance between precision and recall, with precision being the number of True Positives divided by the number of True Positives and False Positives, and recall being the number of True Positives divided by the number of True Positives and False Negatives (sensitivity).

3. RESULTS

Results are shown in Table 1 represented by F1-score of each model and imputation strategy with the best-performing hyper-parameters.

Table 1: Algorithm F1-micro Scores by Imputation Strategy

	Model F1-micro Score					
Imputation Strategy	MLP	XGBoost	Random Forest	Decision Tree	Logistic Regression	Naive Bayes
KNN	0.9266	0.9272	0.9251	0.8723	0.7442	0.8048
MLP	0.9265	**0.927367**	0.9252	0.8716	0.7437	0.8029
Mean - Drop >75% NA	0.925373	0.927366	0.925214	0.871516	0.742914	0.80327
Mean - Drop >50% NA	0.9262	0.9268	0.9251	0.8734	0.7429	0.8022
Median - Drop >75% NA	0.9259	0.9271	0.9249	0.8712	0.7445	0.8038

Table 2: Keras Deep Learning MLP Results

	Keras MLP Model	
Imputation Strategy	Test Loss	Test Accuracy
KNN	0.2327	0.9250

MLP	0.2328	0.9249
Mean - Drop >75% NA	0.2329	0.9252
Mean - Drop >50% NA	0.2334	0.9260
Median - Drop >75% NA	0.2323	0.9245

4. DISCUSSION

- data is very unbalanced, thus it is difficult to make conclusions from these results

6. APPENDIX

Table 1.

Chosen Features
Gender
Age
Race/Ethnicity?
What is the highest grade or level of school you have completed or the highest degree you have received?
Marital status?
Annual household income?
Total number of people in the Household?
Ever have 4/5 or more drinks every day?
Doctor told you have diabetes or prediabetes?
In general, how healthy is your overall diet?
How many of those meals did you get from a fast-food or pizza place in the past week?
Have serious difficulty hearing?
Have serious difficulty seeing?
Ever used cocaine/heroin/methamphetamine?
Covered by health insurance?
Ever told you had weak/failing kidneys?
Ever been told you have asthma?
Doctor ever said you had arthritis?
Ever told had congestive heart failure?
Ever told you had coronary heart disease?
Ever told you had heart attack?
Ever told you had a stroke?
Ever told you had thyroid problem?
Ever told you had chronic bronchitis?
Ever told you had any liver condition?
Ever told you had COPD?
Ever told you had cancer or malignancy?
Broken or fractured a hip?
Ever told had osteoporosis/brittle bones?
Vigorous recreational activities?
Moderate recreational activities?

Minutes sedentary activity?

Past week number of days cardiovascular exercise?

Past week number of days strengthened muscles?

Number of sex partners/lifetime?

Describe sexual orientation.

Ever told by doctor have sleep disorder?

Smoked at least 100 cigarettes in life?

Caffeine Intake (mg/day)

Systolic: Blood pressure (mm/Hg)

Diastolic: Blood pressure (mm/Hg)

*Body Mass Index (kg/m**2)*

Waist Circumference (cm)

Total Bone Mineral Density (g/cm^2)

Direct HDL-Cholesterol (mg/dL)

LDL-cholesterol (mg/dL)

Total Cholesterol (mg/dL)

Triglyceride (mg/dL)

Glycohemoglobin (%)

Herpes Simplex Virus Type 2

HIV antibody test result

References

Belmaker, R. H., & Agam, G. (2008). Major Depressive Disorder. *New England Journal of Medicine*, *358*(1), 55–68. https://doi.org/10.1056/NEJMra073096

Centers for Disease Control and Prevention (CDC). (2017). *NHANES 2005–2016*. Retrieved from https://wwwn.cdc.gov/nchs/nhanes/default.aspx

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., … Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

Chollet, F., & others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

Cizza, G., Primma, S., & Csako, G. (2009). Depression as a risk factor for osteoporosis. *Trends in Endocrinology and Metabolism*, *20*(8), 367–373. https://doi.org/10.1016/j.tem.2009.05.003

Depression. (2020). Retrieved July 2, 2020, from https://www.who.int/news-room/fact-sheets/detail/depression

Dutta, A., Batabyal, T., Basu, M., & Acton, S. T. (n.d.). *An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction*.

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, *24*(11), 1037–1052. https://doi.org/10.1111/cns.13048

Geetha, R., Professor, R., Head, &, & Sivagami, G. (2011). *Parkinson Disease Classification using Data Mining Algorithms*. *International Journal of Computer Applications* (Vol. 32).

Iranpour, S., & Sabour, S. (2019). Inverse association between caffeine intake and depressive symptoms in US adults: data from National Health and Nutrition Examination Survey (NHANES) 2005–2006. *Psychiatry Research*, *271*, 732–739. https://doi.org/10.1016/j.psychres.2018.11.004

Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, *50*(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Kana, M. (2020). Handling Missing Data for Advanced Machine Learning | Towards AI — Multidisciplinary Science Journal. Retrieved July 2, 2020, from https://medium.com/towards-artificial-intelligence/handling-missing-data-for-advanced-machine-learning-b6eb89050357

Kessler, R. C., Van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., … Zaslavsky, A. M. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry*, *21*(10), 1366–1371. https://doi.org/10.1038/mp.2015.198

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., … McIntyre, R. S. (2018, December 1). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*. Elsevier B.V. https://doi.org/10.1016/j.jad.2018.08.073

Liu, Y., Ozodiegwu, I. D., Yu, Y., Hess, R., & Bie, R. (2017). An association of health behaviors with depression and metabolic risks: Data from 2007 to 2014 U.S. National Health and Nutrition Examination Survey. *Journal of Affective Disorders*, *217*, 190–196. https://doi.org/10.1016/j.jad.2017.04.009

Majidi, M., Khadembashi, N., Etemad, K., Jafari, M., & Khodakarim, S. (2019). Associated factors with major depression: a path analysis on NHANES 2013–2014 study. *International Journal of Culture and Mental Health*. https://doi.org/10.1080/17542863.2018.1563623

NHANES 2009-2010: Mental Health - Depression Screener Data Documentation, Codebook, and Frequencies. (n.d.). Retrieved June 29, 2020, from https://wwwn.cdc.gov/Nchs/Nhanes/2009-2010/DPQ_F.htm

Patel, P. O., Patel, M. R., & Baptist, A. P. (2017). Depression and Asthma Outcomes in Older Adults: Results from the National Health and Nutrition Examination Survey. *Journal of Allergy and Clinical Immunology: In Practice*, *5*(6), 1691-1697.e1. https://doi.org/10.1016/j.jaip.2017.03.034

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved from http://jmlr.org/papers/v12/pedregosa11a.html

Razzak, M. I., Imran, M., & Xu, G. (2020). Big data analytics for preventive medicine. *Neural Computing and Applications*, *32*(9), 4417–4451. https://doi.org/10.1007/s00521-019-04095-y

Semerdjian, J., & Frank, S. (2017). An Ensemble Classifier for Predicting the Onset of Type II Diabetes. Retrieved from http://arxiv.org/abs/1708.07480

What Is Depression? (2017). Retrieved July 2, 2020, from https://www.psychiatry.org/patients-families/depression/what-is-depression

Wyatt, M., Johnston, T., Papas, M., Taufer, M. Development of a Scalable Method for Creating Food Groups Using the NHANES Dataset and MapReduce. In Proceedings of the ACM Bioinformatics and Computational Biology Conference (BCB), pp. 1 – 10. Seattle, WA, USA. October 2 – 4, 2016.