**Developing Depression Risk Prediction Models:**

Using the National Health and Nutrition Examination Survey Data 2013-2014

Group 2: Bei Wang, Fangyuan Zhao, Kat Hutcheson, Amro Hassan

1. **Introduction**

According to the WHO, depression has become one of the leading causes of disability worldwide (*Santomauro et al., 2020*). In the US, it is estimated that almost ten million people suffer from depression, while only one-third of them received appropriate treatment (*Bailey et al., 2019*). As the COVID-19 pandemic continues to negatively impact people's mental wellbeing, the burden of major depressive disorder is expected to grow (*Santomauro et al., 2020*).

Depression is known to have multifactorial causes including genetics, lifestyle, socioeconomic status, environmental stressors, and structural factors (*Bailey et al., 2019*). Despite considerable neurobiological, genetic, and psychological studies on the etiology of depression, our understanding of what predicts depression remains limited.

Social service and health care professionals are well-situated to assess depression risk of their clients or patients during brief encounters. But this has been challenging because people who are at higher risk of mental disorders might be reluctant to respond to standardized questionnaires for depression screening (e.g. PHQ-2 and PHQ-9). Moreover, there is no efficient and user-friendly predictive tool for this purpose.

This makes supervised learning techniques extremely helpful in developing a depression risk prediction model using a wide set of features. With such a model, organizations in the healthcare and social service industry can better screen people with higher risks of depression, allowing opportunities for early interventions and treatments.

The National Health and Nutrition Examination Survey (NHANES) is a great dataset to work on as it contains a wide variety of factors including measures of depression. In 2019, a model with similar intent has been built using the NHANES dataset (*Oh et al., 2019*). However, this model is hard to be administered in practical settings because it requires information based on lengthy and lab-based tests such as physically-collected biomarkers (e.g. Creatinine level in urine, cadmium and Glycohemoglobin level in blood). In addition, the full model contained 157 variables, making the implementation extremely time- and resource- consuming. More importantly, social distance during the pandemic urges the need of a model that uses information that can be easily and remotely collected (e.g., phone or online survey).

Therefore, our ***main goal*** is to build an efficient depression risk prediction model for social service and healthcare practitioners to use during brief encounters with people who might be at risk of depression. Our ***second aim*** is to examine whether people of different racial/ethnic backgrounds have different predictors. The existing model did not take this into account and it is known that people of diverse backgrounds might experience different risk factors and etiology of depression (*Bailey et al., 2019*).

The performance of the risk prediction model(s) will be measured by metrics including overall accuracy and the Area Under the Curve (AUC). The existing risk prediction model that used serial datasets from 1999 to 2012 and 157 variables had an AUC of 0.92. Thus, an acceptable performance of our model should be at least 0.80, given that we aim to narrow down to a much smaller set of features.

1. **Method**

**2.1 Data source, study design, and features**

We used the National Health and Nutrition Examination Survey (NHANES, 2013-2014) initiated by the Centers for Disease Control and Prevention (CDC) and the National Center for Health Statistics (NCHS). The NHANES aims to provide US national health data on the prevalence of major diseases and associated risk factors among adults and children.

The NHANES collects information on demographic, socioeconomic, dietary, health-related questions, medications, physical and psychological examination, and laboratory tests. The study oversamples non-Hispanic blacks, Hispanics, and recently non-Hispanic Asians, forming a diverse cohort. Details on study content and operation have been published elsewhere (*"NHANES - About the National Health and Nutrition Examination Survey", 2021*).

In our present project, we focus on NHANES survey questions that would be easy to ask in brief health care or social service encounters and does not require lengthy and expensive tests. Based on these criteria, we narrowed down to 1166 features in aspects of demographic, socioeconomic, dietary, socio-psychological, behavioral health, medical conditions, functional health, and life experience.

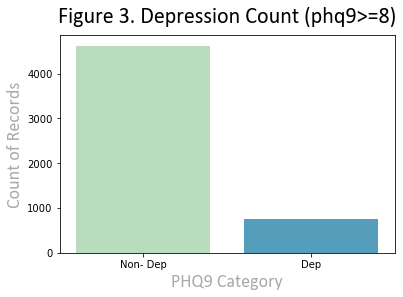
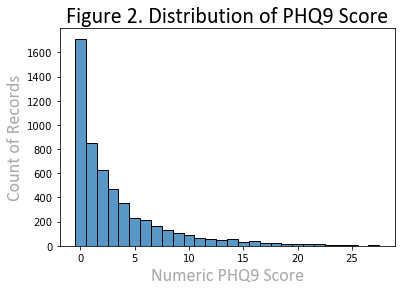
**2.2 Measurement of target: moderate depression**

# The Patient Health Questionnaire (PHQ-9) was used in the NHANES (2013-2014) to evaluate depressive symptoms. We used the threshold of 8 to determine depression (PHQ-9 >=8: depression) (*Manea et al., 2012*). The PHQ-9 is a widely used and validated nine-question measurement assessing severity of depression. Each of the nine questions asked people to self-rate how often they have experienced the indicated symptoms of depression over the past two weeks on a 4-point likert scale (0= “not at all,” 1= “several days,” 2= “more than half the days,” and 3= “nearly every day). The scores on each question were summed, resulting in a total score ranging from 0 to 27, with higher scores indicating greater severity of depression.

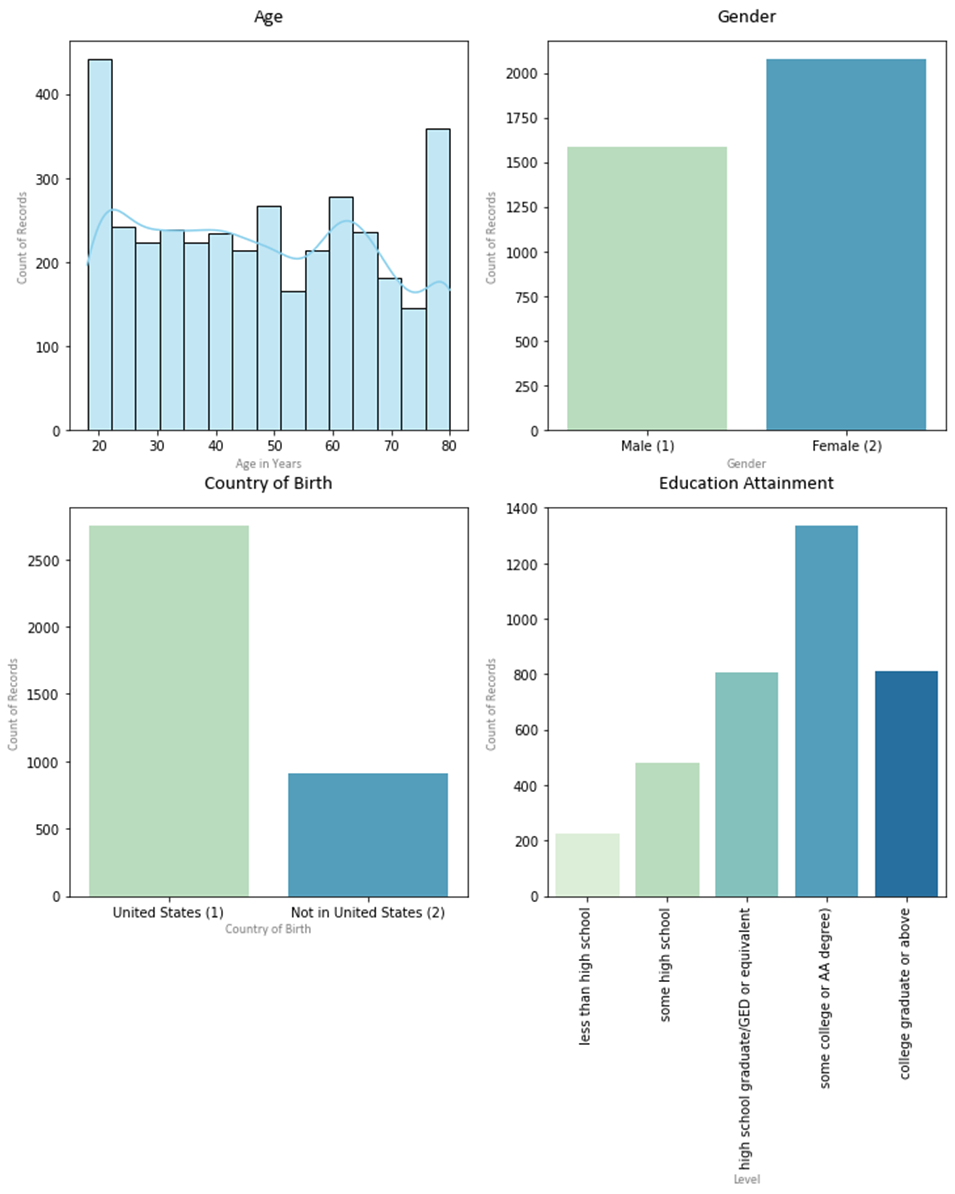
# **2.3 Exploration and visualization of target and feature variables**

One of our aims of using this data set was to develop race-specific risk prediction models. However, upon reviewing the data, it is skewed towards most respondents identifying as non-Hispanic Whites, leaving small sample/test power for us to achieve this aim.

We explored and visualized the distributions of PHQ9 and moderate depression (Figure 2-3).



We further explored several key features to describe the population, which includes race/ethnicity, age, gender, and country of origin (Figure 4). We found that there were more female respondents than males. There were greater proportions of 15-25 and 75-80 year-old respondents compared to others. We also found that the majority of the respondents attended or completed colleges (including associates degree). Additionally, the majority of respondents were born in the US (vs. another country).



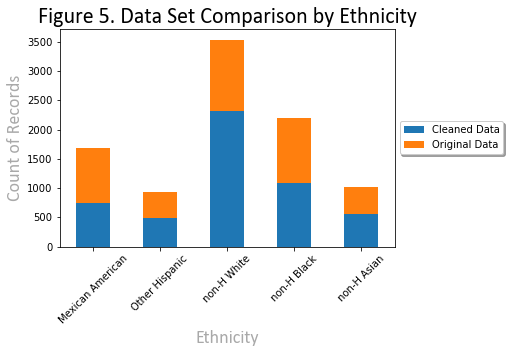
**2.4 Data preprocessing: missingness and imputation**Missingness in the original dataset was represented by combinations of 7s and 9s; we first cleaned the data by replacing them with NaN. We then dropped observations that have missingness in any of the PHQ-9 items, which results in 3657 unique observations. We also dropped features that have more than 20% of missing data and several non-numeric features. Remaining missing values were imputed using the most frequent value along each column using *Simpleimputer* from the sklearn package. 

Figure 5 compares the cleaned and original data by different racial/ethnic groups. There might be a pattern that non-white groups got greater proportions of missingness and were therefore dropped more compared to whites. Before removing missing data, Non-Hispanic white respondents were 38% of total records. Once removing records with majority missing data, Non-Hispanic white respondents increased to being 48% of the total records we used in our model.

**2.5 Data preprocessing: initial feature selection, normalization, and SMOTE**

To achieve our goal of building a model with questions that are brief and easy to ask, we went through a manual selection of features. This step includes dropping administration-related features (e.g., language of interview, interpreter code), features that require lengthy dietary assessments, and duplicated questions. All features were normalized by removing the mean and scaling to unit variance using *StandardScaler.* Our final dataset consists of 5372 unique observations and 211 features.

As shown in Figure 3, our target variable has an imbalanced classification, where there were much fewer minority classes (depression) compared to majority (non-depression). Synthetic Minority Oversampling Technique (SMOTE) is a commonly used technique to address this problem by creating a synthetic example of the minority class. Research suggests that the combination of SMOTE and under-sampling the majority class performs better than plain SMOTE (*Chawla et al., 2002*). We therefore oversample the minority class to have 40% of the number of examples of the majority class, then use random undersampling to reduce the number of examples in the majority class to have 50 percent more than the minority class (*Brownlee, 2021*).

**2.6 Feature selection with gradient boosting**

We used *XGBoost* to estimate the importance of features and selected the top 20 features with highest importance. Importance provides a score that indicates how useful or valuable each feature was in the construction of the boosted decision trees within the model. Importance is calculated from each single decision tree by the among that each attribute split improves the performance measure, weighted by the number of observations the node is responsible for. Feature importances were then averaged across all trees in the model (*Brownlee, 2021*).

**2.7 Model selection: 10-folds cross validation, grid search, and supervised learning models**

Because of the imbalanced class labels in our target variables, 40% of the data was set aside as the test set using the stratify parameter in *train\_test\_split* to retain the same proportion of class labels in test and training sets.

The training set is used to fit several candidate models with repeated 10-folds cross-validation. We evaluated these models using average accuracy scores and average Receiver Operation Characteristics *(*ROC) Area Under Curve (AUC) scores from the cross-validation. The AUC score is a more unbiased estimate with imbalanced classification like ours (Brownlee, 2021). After an optimal model is selected based on the best cross-validation results, we re-trained the selected model using the full training set. This trained model would be our final model and its performance would be evaluated for its AUC score using the held-out test set.

We used multiple supervised learning models for classification of moderate depression. We started with a logistic regression model and a decision tree classification as our baseline models. We then used ensemble methods that synthesize the results of multiple learning algorithms to achieve better predictive performance over a single estimator. These ensemble models include random forests, adaptive boosting, and gradient boosting. Random forest trees build multiple decision trees using a bagging method, which results in average prediction ability of the trees. Gradient boosting sequentially builds decision trees using gradient descent to reduce the loss function (we applied *GradientBoostingClassifier* and *XGBoost*). Adaptive boosting builds weak learners sequentially and trains them using weighted training data. Once a predetermined number of weak learners have been reached or no further improvement can be made, the process stops. Grid search was used in tuning hyper-parameters of these models for the best performance.

1. **Result**

**3.1 Feature Selection: Top 20 Features using XGBoost**

Of the 5372 respondents in our sample, 748 (14%) had depression (defined as PHQ-9 score ≥ 8). Because the sample is imbalanced in terms of the outcome, we used SMOTE and under-sampling methods combined to balance the data, and then selected the 20 most important features using *XGBoost.* These selected features will proceed to train our model.

​​

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1. Top 20 ranked variables by feature importance using XGBoost after balancing data through SMOTE** | | | |
|  | **Variable** | **Feature Importance** | **Meaning of variable** |
| 0 | SLQ050 | 0.07114561 | Ever told doctor had trouble sleeping? |
| 1 | PFQ049 | 0.054363493 | Limitations keeping you from working |
| 2 | DLQ040 | 0.036361996 | Have serious difficulty concentrating ? |
| 3 | RIAGENDR | 0.027910536 | Gender |
| 4 | HSD010 | 0.024890268 | General health condition |
| 5 | DBQ700 | 0.023467826 | How healthy is the diet |
| 6 | CBQ550 | 0.022286935 | Eat at restaurants w/ waiter |
| 7 | DUQ370 | 0.02226003 | Ever use a needle to inject illegal drug |
| 8 | SMQ863 | 0.019702034 | Used nicotine replacement for the last 5 days? |
| 9 | OHQ770 | 0.018244049 | Past yr need dental but couldn't get it |
| 10 | MCQ010 | 0.016550215 | Ever been told you have asthma |
| 11 | DIQ160 | 0.0132357385 | Ever told you have prediabetes |
| 12 | MCQ080 | 0.012802899 | Doctor ever said you were overweight |
| 13 | INQ080 | 0.010637194 | Income from retirement/survivor pension |
| 14 | MCQ160F | 0.010082339 | Ever told you had a stroke |
| 15 | HSQ510 | 0.009599829 | SP have a stomach or intestinal illness? |
| 16 | PAQ665 | 0.009386042 | Moderate recreational activities |
| 17 | DIQ050 | 0.009236204 | Taking insulin now |
| 18 | MCQ160K | 0.009196441 | Ever told you had chronic bronchitis |
| 19 | WHQ040 | 0.008771916 | Like to weigh more, less or same |

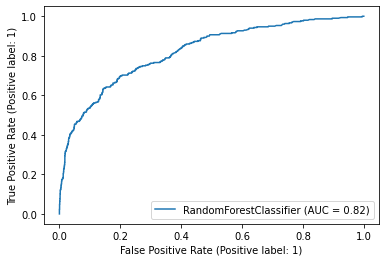
**3.2 Model Selection and Cross-validation results**

The detailed steps of model selection and comparison are presented in the corresponding Jupyter notebook. The following table summarizes the performances of the different models in the training set.

|  |  |  |
| --- | --- | --- |
| **Table 2. Model performance in training set with 10-fold cross validation** | | |
| **Model Name** | **Accuracy (sd)** | **Mean AUC**  **(with 10-fold cross validation)** |
| **Baseline model** | | |
| Logistic Regression | 0.778 (0.017) | 0.825 |
| Decision Tree | 0.857 (0.017) | 0.874 |
| **Tree-based ensemble classification models** | | |
| Random Forest | 0.885 (0.009) | 0.933 |
| AdaBoost | 0.889 (0.012) | 0.929 |
| Gradient Boosting | 0.887 (0.011) | 0.929 |
| XGBoost | 0.883 (0.014) | 0.917 |

As is shown above, the Random Forest model has the highest mean AUC while the AdaBoost model has the highest accuracy score. Given our project goals (i.e. depression risk prediction), we are more interested in AUC than accuracy scores because we care about the sensitivity and specificity of our model. In addition, the imbalanced nature of our dataset makes accuracy scores more likely to be biased. In fact, the accuracy scores of the tree-based ensemble models are extremely close. Therefore, we choose the Random Forest model as the final model.

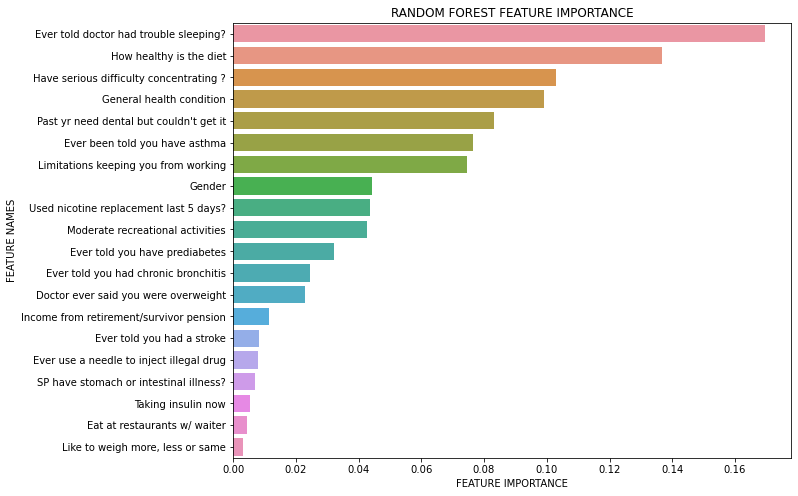
**3.3 Evaluating performance of the final model in testing set**



Our final model was evaluated for its AUC score using the testing set, which was NOT used in the training process to provide unbiased estimation. Our final best model has an AUC of 0.82. In general, 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding. This also meets with our goal of developing a model with an AUC no smaller than 0.8.

**3.4 Ranked feature importance of the final model**

Feature importance of the final model is presented in the graph below. We can see that sleep, behavioral health, functional health, and medical conditions play important roles in predicting depression.



1. **Discussion**

We used supervised learning methods to develop a 20-item predictive model for depression with satisfactory prediction performance (AUC= 0.82). The model demonstrated good predictive performance within a multiracial population. These questions can be administered within several minutes and does not require any in-person assessment, physical data collection, or lengthy examinations. This might assist practitioners to identify people who are at higher risk of depression and provides opportunities to optimize service/care (especially useful during the pandemic).

Our model has ***limitations***. We did not focus on drawing statistical inference and causality to shed light on risk factors of depression. As a consequence, our final model is subject to the problem of multicollinearity. Secondly, the imbalanced classification leads to a small number of true positives in classification results, limiting the sensitivity and usefulness of the model to predict depression. Additionally, although we resampled the training set and kept the test set unchanged, the use of resampling strategies still induced bias. Lastly, we were unable to find the evidence for whether people of different racial/ethnic backgrounds might have different predictors. Greater proportions of data were missing among non-white respondents, resulting in insufficient power to develop racial/ethnic-specific models.

**Reference**

Bailey, R. K., Mokonogho, J., & Kumar, A. (2019). Racial and ethnic differences in depression: current perspectives. *Neuropsychiatric disease and treatment*, *15*, 603.

Brownlee, J. (2021). Feature Importance and Feature Selection With XGBoost in Python. Retrieved 5 December 2021, from <https://machinelearningmastery.com/feature-importance-and-feature-selection-with-xgboost-in-python/>

Brownlee, J. (2021). ROC Curves and Precision-Recall Curves for Imbalanced Classification. Retrieved 5 December 2021, from <https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-imbalanced-classification/>

Brownlee, J. (2021). SMOTE for Imbalanced Classification with Python. Retrieved 6 December 2021, from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Cmaj*, *184*(3), E191-E196.

NHANES - About the National Health and Nutrition Examination Survey. (2021). Retrieved 5 December 2021, from <https://www.cdc.gov/nchs/nhanes/about_nhanes.htm>

Oh, J., Yun, K., Maoz, U., Kim, T. S., & Chae, J. H. (2019). Identifying depression in the National Health and Nutrition Examination Survey data using a deep learning algorithm. *Journal of affective disorders*, *257*, 623-631.

Santomauro, D. F., Herrera, A. M. M., Shadid, J., Zheng, P., Ashbaugh, C., Pigott, D. M., ... & Ferrari, A. J. (2021). Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic. *The Lancet*, *398*(10312), 1700-1712.