STATS415 Group Project

Group 8: Yihao Sun, Tianhong Gao, Jiayi Zhao, Nicholas Charles Gaty

December 2 2023

1 Part I: Prediction on one's probability of suffering from depression

1.1 Introduction

Depression, a severe psychological disorder, is increasingly gaining attention due to the significant harm it causes to both physical and mental health, as well as its various, complex, and unpredictable causes. We believe that, in addition to the existing PHQ-9 questionnaire, by observing a person's lifestyle and physical health, we can early on distinguish a group of individuals with a high likelihood of depression. This can subsequently aid in better diagnosis and prevention of the worsening of the condition. Therefore, in this report, we attempt to answer the question of what is the probability of an individual currently suffering from depression by developing a model that can predict the probability of individuals being diagnosed with depression based on questionnaire results about their personal lifestyle habits and health.

1.2 Data

1.2.1 Data Set Selection

Based on the handbook[3], we grasp that sleep disorder, dramatic weight loss, and disordered diet behavior are major symptoms of depression shown in daily life, while low physical activity participation, alcohol use, low income, and extraordinary blood pressure may indicate those major symptoms. Thus, we select related questionnaire data sets from [1]2017-March 2020 Pre-Pandemic Questionnaire Data - Continuous NHANES, those picked data sets are respondent's income status(P_INQ), sleep quality(P_SLQ), diet behavior(P_DBQ), physical activity participation(P_PAQ), alcohol use(P_ALQ), blood pressure status(P_BPQ), weight history(P_WHQ), and depression(P_DPQ).

1.2.2 Feature Selection and Data Cleaning

The questionnaires in the data sets follow a "Skip Logic" structure, where respondents skip certain questions based on previous responses, leading to many empty entries. To mitigate the impact of missing answers, only questions answered by over 95% of respondents are selected. Missing values, "don't know," and "refused" in predictors are handled by creating a "not answered" choice for multiple-choice questions, filling in mean values for meaningful data, and recording the miss in a new column. Timestamps in the sleep disorder data set are transformed into doubles representing hours. After merging data sets, dummy features are created for multiple-choice questions. The PHQ_9 questionnaire's scores are summed up, with scores above 4 indicating depression[2]. That is to say, if a respondent has PHQ_9 score larger than 4, the respondent would be regarded as a depression patient. The distributions of total and PHQ_9s are analyzed.

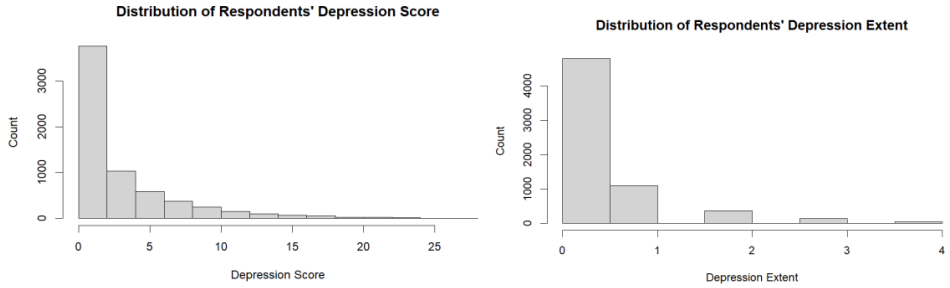


Figure 1: Distributions of Respondents' PHQ_9s and Extents

1.3 Method

Recognizing that only a small proportion, approximately 3.8%[3], of the total population suffers from depression, our data set is notably skewed towards the healthy side, making predicting the exact PHQ_9 score exceedingly challenging. Consequently, rather than treating inaccurately predicted PHQ_9 scores as a definitive binary distinction of 0% or 100% depression possibility, our approach focuses on predicting the probability of a respondent experiencing depression based on their approximated PHQ_9 score. For instance, if a respondent has an approximated PHQ_9 score of 1, commonly associated with an absence of depression, there remains a possibility of experiencing depression due to the inherent inaccuracies in the score approximation. We estimate this probability with the following methods.

We divide our training methods into three parts. Exploratory data analysis with further data cleaning, fitting a ridge regression between predictors and the responding PHQ_9 in train set $y_{approx} = k_1 X$, and mapping the response y of the previously generated linear model to the depression probability. In the last step, we use the predicted scores to estimate the probability of depression, specifically the likelihood that an individual's actual PHQ_9 exceeds 4. The formula is given by $Pro(y(X) > 4) = k_2 y_{approx}(X)$, where y(X) represents the respondent's real PHQ_9 score, and $y_{approx}(X)$ represents the respondent's estimated PHQ_9 score obtained above. Finally, we test our model by comparing the predicted probability with the frequency in the test set.

1.3.1 Exploratory Data Analysis (Random Forest)

Before fitting our data with a model, we clean the data first. We enhance the data, redesign the path, and apply feature selection with random forest. Although data enhancement has limited progress, feature selection works well after we redesign the prediction path.

Noticing that variance is high in the data set because of too few samples with large PHQ_9 score, it is difficult to correctly tell those extents apart. As a result, we oversample those minorities with minor random noises to balance the data. However, oversampling introduces too much bias, leading to an unsatisfactory performance. Hence, we redesign the path and remove the extent.

In addition, we perform feature selection to remove redundant features. As there are about 80 dummy features and 30 numeric features, We select random forest model to do this job because it can deal with both kinds of them properly. 15 features are kept.

1.3.2 Approximate Linear Model (Cross-validated Ridge)

After doing all exploratory data analysis, we think cross-validated ridge regression most suitable, as it stabilizes the coefficients, prevents overfitting, and reduces noises. We fit the responding PHQ_9 y_{approx} to the hyper matrix of predictors X with cross-validated ridge regression model, where the optimal lambda is learned through cross validation. Given the large variance of the original data, we don't expect this

linear model to have very high accuracy in terms of MSE and R^2. Rather, this linear model serves as an approximation to predict the mathematical expectation of the response PHQ_9 if given a set of predictors.

1.3.3 Mapping to probability (LDA)

In this part, we maps the the response of the approximate linear model to the probability. We discovered that the predicted data follows a certain kind of distribution in figure 2. We suggest there is a linear relationship between the approximated PHQ_9 y_{approx} and the probability of depression. We assume the feature within each class have the same covariance matrix for all classes. Therefore, we choose LDA.

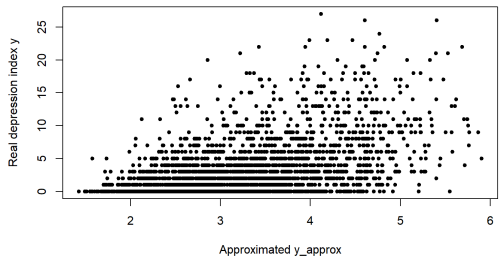


Figure 2: Distribution of Respondents' Actual PHQ_9 Versus Approximated Score

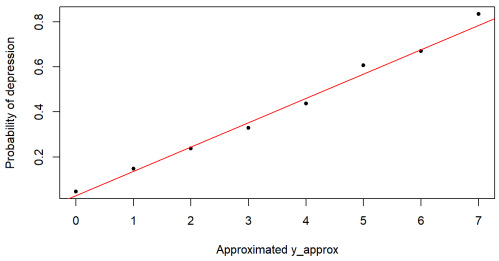


Figure 3: Probability of depression vs. Approximated y_{approx} (Tested)

After fitting on 10 different train sets, we find the relation is $Pro(y(X) > 4) = 0.102 \cdot y_{approx}$

1.3.4 Testing (Bootstrap)

Since we are testing on the probability instead of a certain value, we are very careful about the bias of the data. In order to avoid probability bias, we use bootstrap to generate 100 sets for test, with each set size equal to 1000. This can help us make sure that the model is really trying to predict the probability based on the features of each individual, instead of simply outputting the overall frequency of the whole data set.

When testing the model, there are three steps. First, we calculate the y_{approx} value with the first linear model. Then, we divide the data into small groups based one the value of their responding y_{approx}. The values that falls in the interval [y,y+1) is put into the same group. We then compute the frequency of depression within each group. Finally, we map the y_{approx} to get the predicted probability, and compare it with the real frequency.

1.4 Results

With random forest, we select 15 most important predictors. We use them to do the approximate linear prediction.

We then try to predict the approximate PHQ_9 with $y_{approx} = \beta X$. With linear regression, we have the coefficients of the variables given by coefficient table below.

We then maps the approximated y_{approx} to the probability of depression. By training on 10 different folds, we have $Pro(y > 4) = 0.102 y_{approx}$

We then test the result on 100 test sets generated by a test set independent from the training set. We get the mean R^2 is 0.98 and the minimum R^2 is 0.95. This suggests that our prediction on probability is unbiased and accurate.

Predictor name	Coefficient
INDFMMPI	-0.184696094
WHQ150	-0.001176355
WHQ040.1	0.862946332
sleepTimeD	-0.030072284
wakeTimeE	-0.056951151
wakeTimeD	0.037141343
SLD013	-0.034098930
SLQ120.3	0.885405683
SLQ120.4	0.363827047
WHD140	0.006975574
BPQ070.4	1.280794498
SLQ050.2	-0.858334031
SLQ050.1	0.858335375
ALQ121.9	0.840768859
BPQ090D.9	0.005382337

Figure 4: Coefficient of predictors

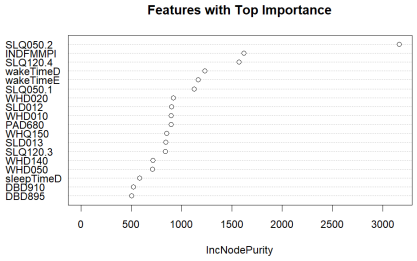


Figure 5: Features with Top Importance

1.5 Conclusion

Qualitatively, the top predictors that indicates higher probability of depression are

1. Poverty
2. Having sleeping trouble
3. Low stature and High weight
4. Long time sitting
5. Much alcohol use

Healthcare workers encountering individuals with the identified features should be alert to the possibility of depression, providing additional assistance for diagnosis and necessary therapy.

For individuals, it's better not to use this model as a self-diagnose tool for depression. Despite meeting several indicative features, there still remains a high likelihood of no or only mild depression. Maintaining a positive attitude and confidence contributes to good health. Eliminating unnecessary worries is a good way to keep depression away.

Quantitatively, we developed a way to predict the probability whether a person is suffering from depression. The accuracy is quite high (given $R^2 \geq 0.95$), and the result can be used for further biostatics research. Health organizations with big data can also implement this model in the system.

Research indicates that only 3.8% of the global population is experiencing depression. However, the model's data set reflects over 32% experiencing depression, likely influenced by biased questionnaire selection. Testing and parameter adjustments specific to local conditions are recommended before implementation, addressing potential issues in statistical analysis.

In summary, the model offers an effective method to predict individual depression probability. However, its application is advised for statistical research and primary healthcare worker diagnosis only. Direct use for diagnosis is cautioned due to potential variations in probability across communities and social backgrounds. A non-biased, inclusive environment is crucial, for the physical and mental health of everyone with and without depression.

2 Part II: Classification on high blood pressure

2.1 Introduction

This report focuses on developing a predictive model to classify individuals with high blood pressure using various health-related predictors. With the increase of issues related to high blood pressure, we thought about what data sets we could use to help predict such a health condition, and doing so we thought of the following question: "Which model is the most effective in predicting high blood pressure based on factors such as Diet, physical characteristics, lifestyle habits, and medical environment?"

We are using data sets from [1]2017-March 2020 Pre-Pandemic Questionnaire Data - Continuous NHANES. The responses to the "Ever told you had high blood pressure" question (BPQ020) in the "Blood Pressure & Cholesterol" data set are our response variable. We utilize seven data sets—Diet Behavior and Nutrition, Weight History, Physical Activity, Alcohol Use, Smoking - Cig Use, Health Insurance, and Blood Pressure & Cholesterol—to predict high blood pressure. These data sets provide valuable information on body characteristics, lifestyle habits, and health insurance. Access to regular health check-ups and services is a crucial factor considered in this study. Our objective is to identify the most effective model for predicting high blood pressure.

2.2 Data

When working with these data sets, our initial step involves selecting features strongly correlated with the target response as predictors. We employ the chi-squared test to assess this correlation. The test compares observed frequency distribution with the expected distribution for independent variables. A small p-value allows us to reject the null hypothesis of independence and choose the variable as a predictor. We set a p-value threshold of 0.00001. Additionally, we only consider features with a valid data proportion exceeding 70% to ensure a minimal proportion of invalid data.

The selected numerical predictors after this process are: Your current weight, your weight 1 year ago, your greatest weight over time. The corresponding questions are WHD020, WHD050, WHD140. We turn these predictors to dummy variables before applying the methods.

The selected categorical predictors after this process are:meal and food shopping responsibilities, transportation preferences (biking or walking), engagement in vigorous and moderate recreational activities, alcohol consumption frequency, excessive drinking habits, smoking history (at least 100 cigarettes), health insurance coverage (including gaps), prescription coverage, high cholesterol status, and prescribed medication use. The corresponding questions are DBQ930, DBA940, PAQ635, PAQ650, PAQ665, ALQ121, ALQ151, SMQ020, HIQ011, HIQ210, HIQ270, BPQ080, BPQ090D.

Below are some visualizations of the relationship between our selected predictors and response:

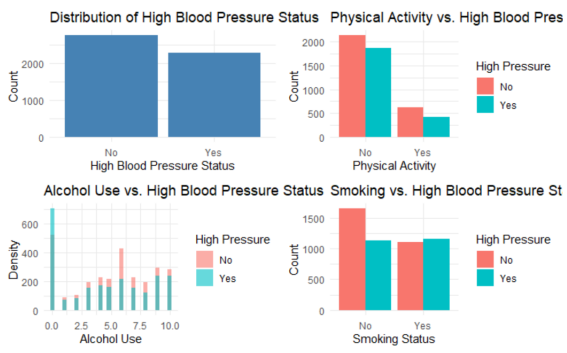


Figure 6: General Distributions

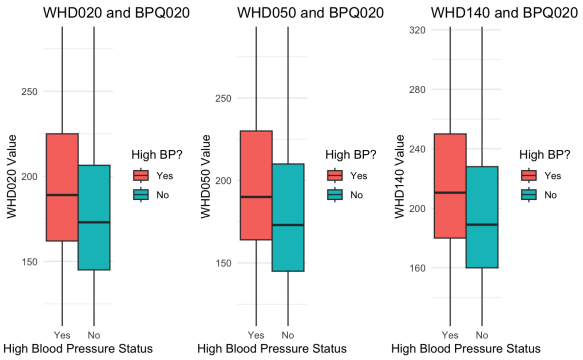


Figure 7: Box-plots of Weight History

We can intuitively observe the relationship between predictors and responses from the chart. For example, the proportion of individuals with high blood pressure is higher among smokers compared to non-smokers. Similarly, individuals with high blood pressure tend to have higher body weight than those with normal blood pressure.

In the following parts, we will attempt to fit models using these relevant data to predict whether an individual has high blood pressure or not.

2.3 Method

We utilized KNN, Trees, Random Forests, Logistic Model, and SVM, along with cross validation, to analyze the data. The data set was randomly split into an 80% training set and a 20% testing set. Cross validation helped identify optimal hyperparameters for models in the training set. After training the models, predictions were made on the test set. Performance evaluation, considering accuracy and ROC curve, were conducted for KNN, Trees, Random Forests, and SVM, as logistic models do not require hyperparameter tuning. Data scaling was applied to KNN and SVM.

KNN: This classification method is effective for predicting high blood pressure, given the similarity in outcomes for individuals sharing physical characteristics and unhealthy habits. Despite KNN's usual challenge with dimensionality, the use of 16 predictors in our case is modest and doesn't significantly impact accuracy.

Trees: These tools were particularly beneficial due to their interpret-ability and robustness. They are well at handling both numerical and categorical data and are capable of capturing non-linear relationships, which is crucial in data sets like those we used.

Logistic: We chose this method for its interpretability, especially in binary classification problems like ours - predicting the presence or absence of high blood pressure. Additionally, it is robust against the influence of outliers.

SVM: We chose SVM for its effectiveness in high-dimensional spaces, handling non-linear relationships with various kernels, making it suitable for our complex data with 16 selected predictors. The radial kernel, known for high flexibility, is selected for its applicability to our problem.

Random Forest: Random Forests are highly effective in predicting high blood pressure, managing complex, high-dimensional data with reduced overfitting and robustness against noise. Their capability to identify important features and handle non-linear relationships enhances their value in predictive modeling. Given our survey-based data's potential noise, we expect Random Forests to perform well in this context.

2.4 Results

After applying the aforementioned methods, we train the corresponding models on the training set and evaluate their performance on the test set. The results are presented below:

2.4.1 Prediction Tables

KNN Prediction Table

Predict \ Truth	Yes	No
Yes	148	69
No	82	206

Tree Prediction Table

Predict \ Truth	Yes	No
Yes	142	49
No	88	226

Logistic Prediction Table

Predict \ Truth	Yes	No
Yes	140	48
No	90	227

SVM Prediction Table

Predict \ Truth	Yes	No
Yes	0	0
No	230	275

Random Forest Prediction Table

Predict \ Truth	Yes	No
Yes	142	53
No	88	222

2.4.2 Parameter and Accuracy

KNN: The optimal k value obtained from cross validation is 19. The accuracy is 70.1% in test.

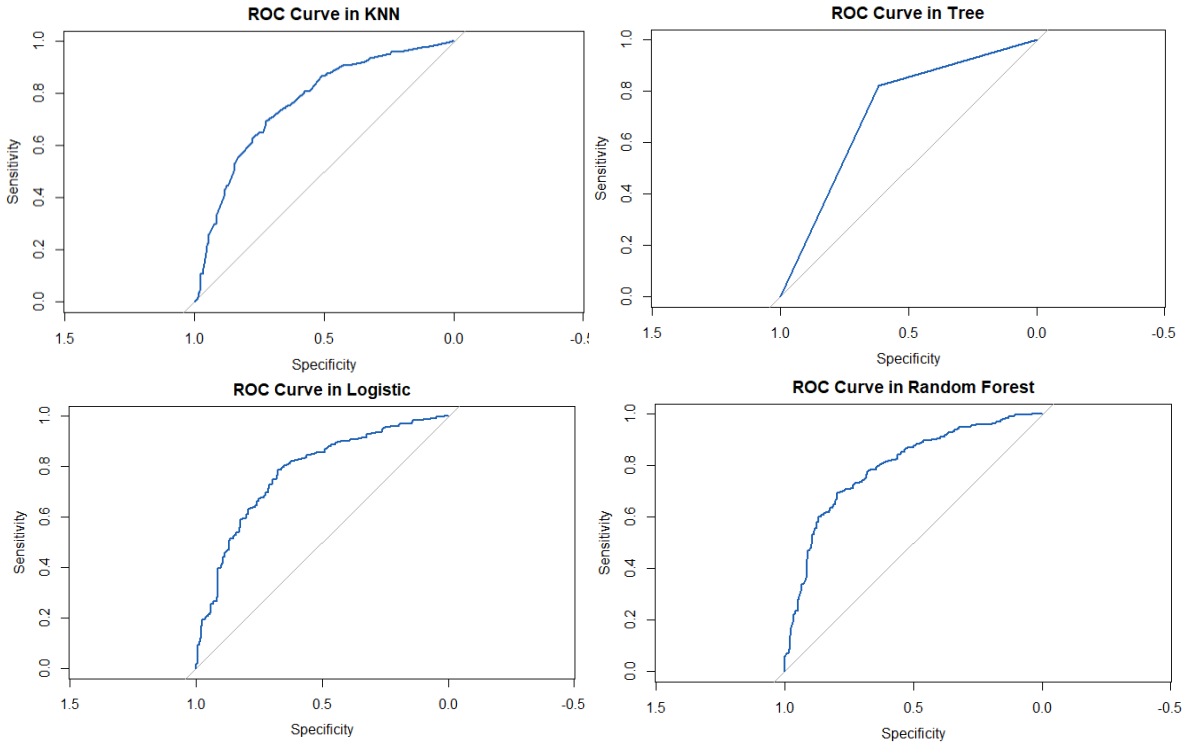
Trees: The optimal complexity parameter (cp) is 0.00266. The accuracy is 72.9% in test.

Logistic: The accuracy is 72.7% in test.

SVM: The optimal C parameter, obtained through cross-validation, is exceptionally small (1e-100), indicating an extremely soft margin. This suggests a high tolerance for misclassifications and a potential for a larger-margin decision boundary. The test accuracy is only 54.5%, revealing poor performance and significant bias as the model predicts every test point as not having high blood pressure.

Random Forest: The number of variables randomly sampled as candidates at each split in a single tree within a random forest (mtry) is 2 obtained from cross validation. The accuracy is 72.1% in test.

2.4.3 ROC



For the effective models (KNN, Trees, Logistic, and Random Forest), we compared their accuracy using ROC curves. A good model's ROC curve is in the upper-left corner, indicating a high true positive rate and low false positive rate. The plots show these models perform well, with a relatively small area in the upper-left corner.

2.5 Conclusions

In summary, our analysis demonstrates that predictive modeling for high blood pressure, utilizing 16 carefully chosen predictors, yields notable accuracy—approximately 72% using trees model, logistic or random forest, as detailed in the results. The actual performance of the applied trained model is expected to surpass these results when trained on the entire data set. ROC curves provide valuable insights, showing our models maintain high true positive rates while minimizing false positives, indicating robust predictive capability. Our comprehensive approach, considering accuracy metrics and ROC analysis, positions our predictive models as reliable tools for assessing and predicting high blood pressure based on the provided set of predictors.

Simultaneously, we acknowledge potential issues in our statistical analyses. Relying on data primarily from questionnaire surveys introduces uncertainty due to possible inaccuracies. This contributes to notable noise in the training data, impacting the SVM's poor performance. To improve prediction, incorporating more laboratory data as predictors is recommended. Caution is also needed in selecting important features from seven data sets, as the chi-squared test may lose accuracy with small sample sizes and low expected counts in some cells. For example, the disproportionate number of individuals with health insurance may affect the chi-squared test's effectiveness.

3 Contributions

In Part One, Yihao Sun meticulously selected predictors, performed thorough data cleaning, and explored various models, including ridge regression. Jiayi Zhao, also contributing to Part One, introduced an innovative method to enhance model performance by mapping linear model responses to probabilities. In Part Two, Tianhong Gao meticulously selected predictors, conducted essential data cleaning, and rigorously tested models like Logistic Regression and SVM. Nicholas Charles Gaty, also in Part Two, explored models like KNN and Trees, conducting rigorous tests and visualizing results.

References

[1] National Center for Health Statistics. "2017-March 2020 Pre-Pandemic Questionnaire Data - Continuous NHANES". In: *National Health and Nutrition Examination Survey* (2021-2022). URL: https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020.

[2] Janet B W Williams Kurt Kroenke Robert L Spitzer. "The PHQ-9: validity of a brief depression severity measure". In: *Journal of general internal medicine vol. 16,9* (2001). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1495268/.

[3] World Health Organization. "Depressive disorder (depression)". In: (2023). URL: https://www.who.int/news-room/fact-sheets/detail/depression/?gclid=Cj0KCQiA67CrBhC1ARIsACKAa8SPqUy3Bfqa3u5QPzr3T
wcB.