Predicting Depression in At-Risk Adults

Problem statement formation:

How can the NIMH decrease the median time delay for mental health treatment through the development of an app that screens individuals for mental illnesses?

Context: According to the NIMH, one in five Americans live with a mental illness. This can include both mental, behavioral, or emotional disorders. In addition, according to the NIMH data from 2017, among the 46.6 million adults with any mental illness, 19.8 million (42.6%) received mental health services in the past year. Furthermore, in 2017, the NIMH identified that 4.5% of all U.S. adults experienced a serious mental illness (SMI), meaning that the SMI substantially interfered with or limited one or more major life activities (Substance Abuse and Mental Health Services Administration, 2018). As the effects of mental illness can be severe, it is prudent to diagnose and treat individuals with mental illness in a timely manner. There are significant delays in diagnosing and treating a mental illness after initial onset (McLaughline, 2004).

Criteria for Success: The creation of an app that will screen individuals for depressive disorders and refer them to a psychiatrist immediately. Success will be measured by assessing the delay in treatment in upcoming years. It is expected that the median delay time will decrease from 11 years until the first contact with a psychiatrist, to less than one year.

Scope of solution space: This app will be used for adults (age 18 and up) in the United States who believe that they are struggling with their mental health. In the future, this can expand to other countries when data is collected from the WHO.

Constraints: There is a lot of missing data in the BRFSS. Potentially, this could be because people refused to answer due to sensitivity to the topic. Because of this, there could be a confounding variable and missing data will have to be dealt with carefully. In addition, the budget needs to be kept in mind because funding will be allocated from the NIMH for this project according to the likelihood of success.

Stakeholders: Those invested in this project will be the NIMH, the SAMHSA, psychiatric treatment providers and insurance companies.

Method: I will solve this problem by developing a classification model to predict the presence of a mental illness in individuals using the app. The app will narrow the predictors down to approximately 10 questions for ease of use.

Deliverables: The deliverables include an app that functions appropriately for adults in the United States, as well as a slide deck that explains how the app was developed and verified. In addition, all the code will be available in a GitHub repo for each step of the project.

Data Collection

Data is acquired from the 2010 BRFSS, which includes over 200 items and over 200,000 responses. However, after reviewing the data included related to depression items, the sample size decreases to about 80,000. Twelve features were extracted from the dataset, with 11 potential predictors and one outcome variable (depression diagnosis). The initial description of the dataset is displayed below.

	count	mean	std	min	25%	50%	75%	max
EMTSUPRT	436221.0	1.974605	1.444904	1.0	1.0	1.0	2.0	9.0
LSATISFY	435968.0	1.688021	0.973586	1.0	1.0	2.0	2.0	9.0
ADPLEASR	79466.0	61.909974	38.298421	1.0	10.0	88.0	88.0	99.0
ADDOWN	79420.0	65.177739	37.122756	1.0	14.0	88.0	88.0	99.0
ADSLEEP	79391.0	50.676638	40.693882	1.0	5.0	88.0	88.0	99.0
ADENERGY	79367.0	36.939232	39.828500	1.0	3.0	14.0	88.0	99.0
ADEAT1	79335.0	59.207676	39.124483	1.0	8.0	88.0	88.0	99.0
ADFAIL	79298.0	74.269868	30.735371	1.0	88.0	88.0	88.0	99.0
ADTHINK	79266.0	73.963553	30.872160	1.0	88.0	88.0	88.0	99.0
ADMOVE	79219.0	79.497684	24.815877	1.0	88.0	88.0	88.0	99.0
ADANXEV	79160.0	1.892724	0.513281	1.0	2.0	2.0	2.0	9.0
ADDEPEV	79104.0	1.843459	0.528095	1.0	2.0	2.0	2.0	9.0

Data Cleaning:

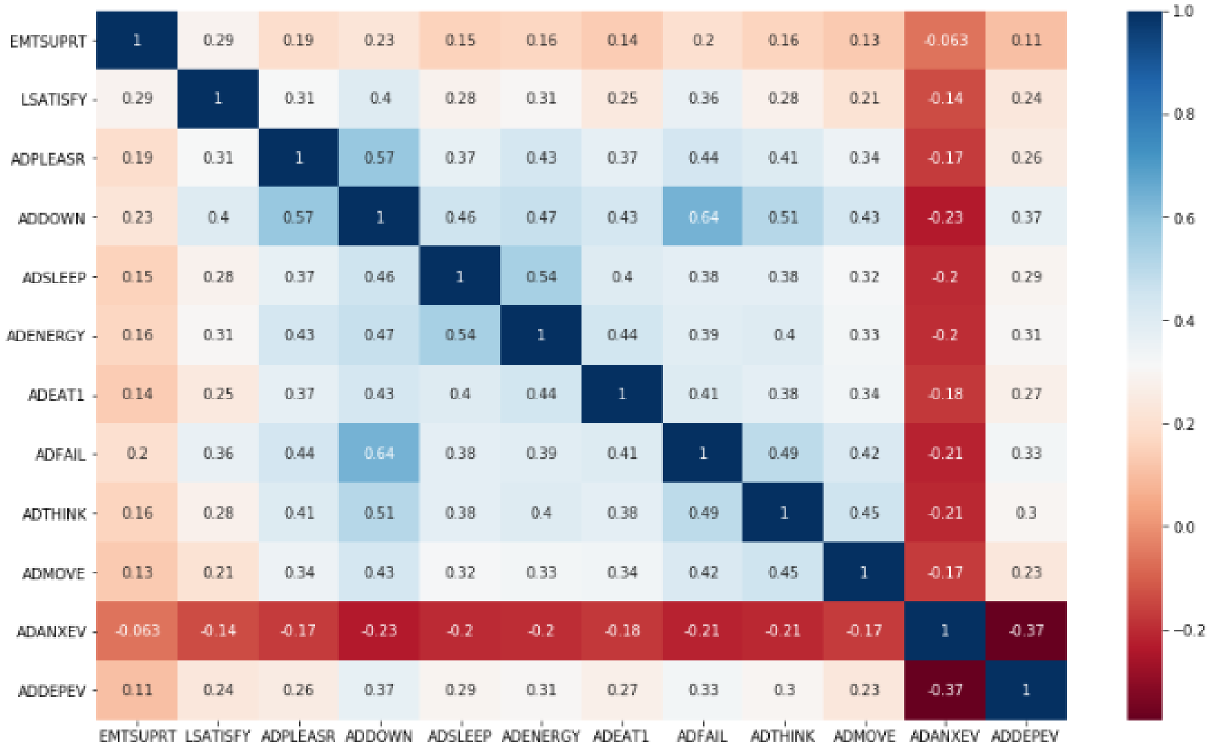
There were several values indicating missing responses for the variables of interest. (Note, values of 9 and 99 as the max values in the table above). For instance, two of the options were 'Unsure' and 'Refused'. The items were all related to mental health, and specifically depression. Therefore, the unsure and refused responses were recoded as missing values and missing values were dropped because it is not possible to predict depression with those responses. The final description of the dataset is displayed below.

Statistic	EMTSUPRT	LSATISFY	ADPLEASR	ADDOWN	ADSLEEP	ADENERGY	ADEAT1	ADFAIL	ADTHINK	ADMOVE	ADANXEV	ADDEPEV
count	72432	72432	72432	72432	72432	72432	72432	72432	72432	72432	72432	72432
mean	1.869173846	1.60524354	1.656850563	1.279172189	2.90297106	3.638516125	2.0926248	0.90177	1.00484592	0.62218357	1.880287718	0.17674509
std	1.263659016	0.72213965	3.488506703	3.07974991	4.55925337	4.694482309	3.9814594	2.805918	2.96215376	2.39009693	0.414176561	0.38145546
min	1	1	0	0	0	0	0	0	0	0	1	0
25%	1	1	0	0	0	0	0	0	0	0	2	0
50%	1	2	0	0	0	2	0	0	0	0	2	0
75%	2	2	2	1	4	5	2	0	0	0	2	0
max	9	9	14	14	14	14	14	14	14	14	9	1

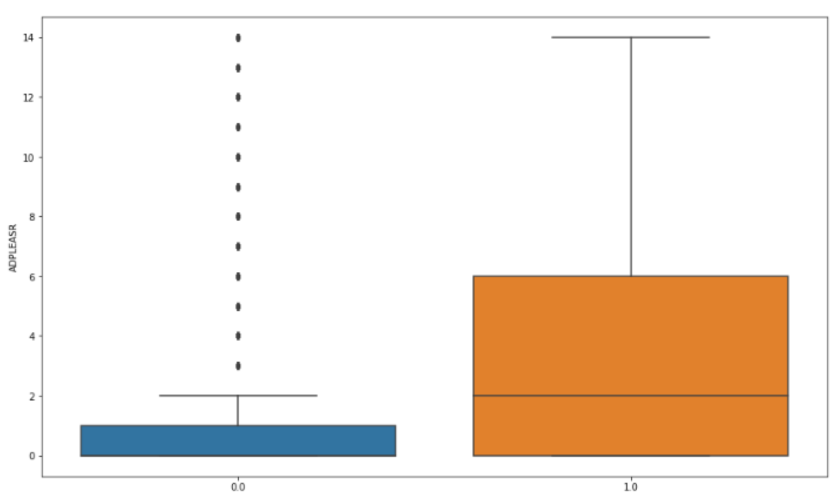
Dummy variables were created for the categorical variables (EMTSUPRT, LSATISFY), which had yes and no responses. The other variables were continuous and asked participants about how many days they experienced specific symptoms over the pas two weeks. Therefore, the numerical variables were not standardized because they were all on the same scale of 0-14.

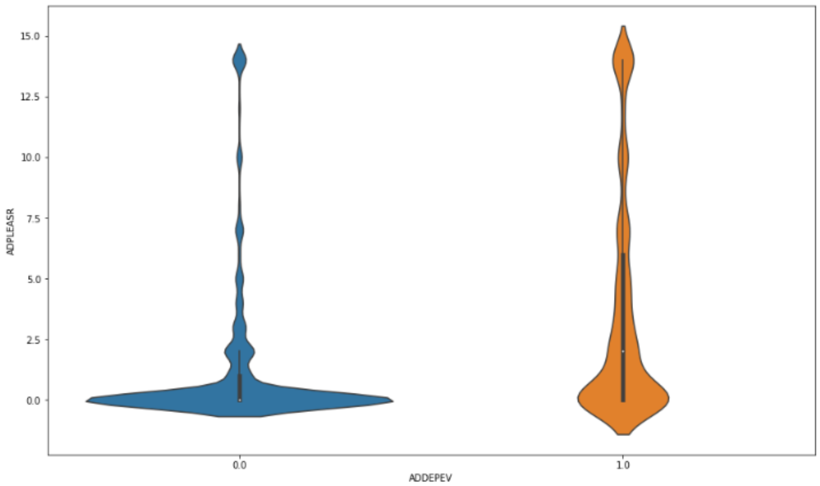
Exploratory Data Analysis:

A heatmap was created to visualize the relationship between variables of interest.



The data were visualized through bar charts for categorical variables, and boxplots and violin plots for the continuous variables. A few examples are displayed below.





As displayed in the charts, it is apparent that those diagnosed with depression (1.0) reported greater variability in their responses to items and overall higher scores, which makes sense because the items ask about symptoms of negative feelings such as feeling like a failure, or lack of pleasure.

Preprocessing

The data were split into the X and y variables (predictors and outcome). Then X and y were each split into 80/20 train/test datasets to analyze the data for the modeling stage of the project. The indices were reset for each dataset to ensure that the index did not skew the modeling results.

Modeling

This is a supervised learning classification problem. Therefore, I have tested the following classification models:

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Decision Tree
- Random Forest

First, logistic regression was used with default values and it did an OK job at predicting the classes with 86% accuracy and 68% precision. However, I wanted to test additional models to see if we could improve upon the precision because that is an important metric in this situation. Precision tells us the proportion of positive cases that are correct. This is important for diagnosing depression because we don't want a high false positive rate.

The next model I tried was KNN. With the number of neighbors set to 5, the model was optimized, yet performed more poorly than the logistic regression model initially run. The accuracy was 84% and the precision was 56%, which is significantly lower.

Next, I tried a decision tree. I started with the default values which yielded a subpar model with an accuracy of 82% and precision 49%. Then I applied hyperparameter tuning to adjust the criterion and max depth. I ended up deciding on the entropy model with a max depth of 5 for the optimal results, with accuracy of 87% and precision of 70%.

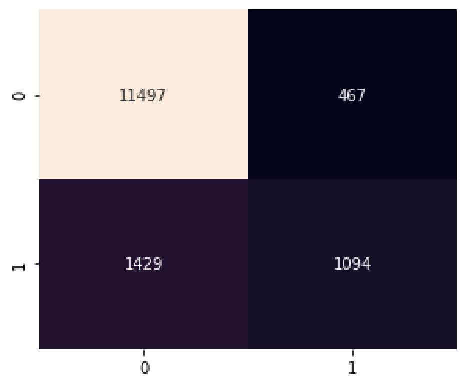
Random forest was also assessed, but it did not perform as well as the decision tree.

Model Selection

DecisionTreeClassifier(criterion='entropy', max_depth=5)

Accuracy: 0.8691240422447711

Precision: 0.700832799487508



Conclusions

The above results demonstrated that a decision tree could accurately and precisely predict those with depression from high-risk adults in the US. This can be applied to build an app for individuals to screens themselves for depression before seeing a psychiatrist. In addition, this can be used as a computer program in colleges to screen students for potential depressive symptoms. Furthermore, this can be expanded to include individuals worldwide or under the age of 18 as more data become available.