Supplementary Material

Glossary

Standardization: Process of reducing numerical values to a 0-1 scale. It is achieved by first calculating the value of a variable, then subtracting the mean of the variable and finally dividing by the standard deviation of the variable

Normalization: Process of reducing numerical values to a 0-1 scale. It is achieved by first calculating the value of a variable, then subtracting the minimum value of the variable, and finally dividing by the maximum value of the variable minus the minimum value of the variable.

Holdout method: A machine learning evaluation method that is formed by splitting the data into a training set and a testing set. A large portion of the data goes to the training set and the remaining smaller portion of the data goes to the testing set.

Intermediate Results

When performing data exploration, we found out that all of the feature variables included outliers. We decided to detect and remove the outliers that are only 3 standard deviations away from the mean so that we don't remove many values from our data and end up having a very small dataset to work with. When plotting the correlation heatmap for both the standardized and normalized datasets, we concluded that multicollinearity was not present between the feature variables, therefore chose to compare Naïve Bayes and Decision Tree algorithms. Before hyperparameter optimization, we concluded that the Naïve Bayes classifier consistently performed better than the Decision Tree, but after hyperparameter optimization, we saw a drastic improvement in the Decision Tree algorithm, and a relatively stable improvement for the Naïve Bayes algorithm.

Implementation Details

Before training and testing the Naïve Bayes model, we decided to normalize our variables since Naïve Bayes can predict probability classifiers, therefore we had to have our values between 0-1. On the other hand, before training the Decision Tree model, we standardized our variables to have an accurate comparison of our variables. We chose to standardize to have a different approach to that of Naïve Bayes, so we can see to what extent this standardization process affects our final results, Finally, when performing hyperparameter optimization, we chose the hyperparameter optimization and auto functions in MATLAB so that it chose the best parameters to choose, which consisted of changing the distribution of the Naïve Bayes classifier and changing the leaf size of the Decision Tree classifier.

Other Graphs and Visualizations