Machine Learning Project- Red wine

Group members: Aide Robles, Hao Hu, Sangeeta Rout, Lanelle Phillips

In our project we tried to figure out which factors would lead to the best quality wine. At the start, I assumed that a wine which had a high amount of alcohol would have the highest quality. I was not sure what other factors would lead to high quality wine.

Our dataset is robust because it uses ten different factors to predict the quality of wine.

We had many visualizations. They included 1 pair plot, 1 heat map, 2 bar plots and multiple confusion matrices.

We displayed our many columns to the user. Then we cleaned our data and made sure it did not have any invalid entries.

We used descriptive statistics to describe our data. It showed the stats of our columns. The statistics showed that total sulfur, free sulfur, alcohol, and fixed acidity had the highest quality of alcohol on average.

After we figured out how our variables were related to quality, then we figured out how our variables were related to each other. Our pair plot and heat map showed the correlations of the variables.

We used three machine learning algorithms: decision tree, random forest and xgboost.

We used a decision tree to inform us about the quality of alcohol based on certain features.

The decision tree is a non-parametric supervised learning method used for classification and regression of our data. We create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Afterwards, our decision tree is visualized. Then we plot the importance of the features on alcohol quality.

The plot shows that Alcohol, sulfates, volatile acidity, and chlorides are the most important factors for the quality of alcohol. Our decision tree has a testing accuracy of 61.25%.

We also have a confusion matrix to determine the performance of our classification model and to see if our prediction values align with our actual values.

Our next machine learning algorithm is random forest. Our testing accuracy was 71.87%. We also plotted a bar plot in conjunction with using random forests to see our feature importance. Fixed acidity, volatile acidity and citric acid were the most important features in predicting wine quality. We plot another confusion matrix and tune it several times.

Our next machine learning algorithm was the xgboost. It was the 1 library not covered in the scope of the class. We split our data, fit it into the model, make predictions of our test data and then come up with an accuracy which was 73.75%.

Finally, we use logistic regression for analysis. We split, standardize, train, use a PCA, then apply the logistic regression model to get an accuracy of 61% for our predicted values.

Our models had varying amounts of accuracy and each listed different features as important to the quality of alcohol. We could not arrive at a common consensus as to which features would lead to the highest quality of alcohol.

I determined two of the features as the most influential for high quality wine across all models: volatile acidity and alcohol. Our analysis has some limitations. Our testing accuracy is only 61% for the decision tree, 71% for the random classifier, 61% for the regression boost, and 73% for the xgboost. Additionally, models do not agree and the features that correlate to a higher wine quality differ across them. Two areas for improvement are ways to improve our accuracy and getting a common consensus about which features correlate to better wine qualities.