LeTicia Cancel

DATA622 Homework #2

December 12, 2022

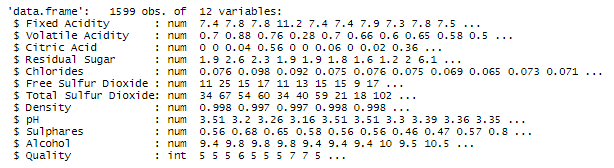
The dataset used for this assignment is the [Wine Quality Data Set](https://archive.ics.uci.edu/ml/datasets/wine+quality) from the UCI Machine Learning Repository. The datasets available are Red Wine Quality and White Wine Quality. The Decision Tree was built with the Red Wine dataset. The dataset has 1,599 observations and 12 variables. The variables are:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol

Output variable (based on sensory data):

1. quality (score between 0 and 10)

If we examine the structure of the dataframe, we see that the datatype of 11 out of the 12 variables are numerical and the output variable 'Quality' is an integer.

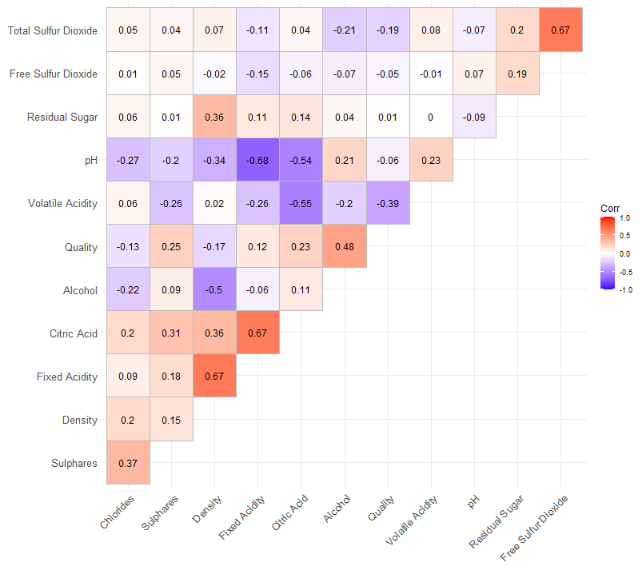


Since most of the variables are numbers we can check if any are correlated. Residual Sugar and Volatile Acidity have a correlation coefficient of zero so there is zero correlation between these two variables. The top negatively correlated variables are:

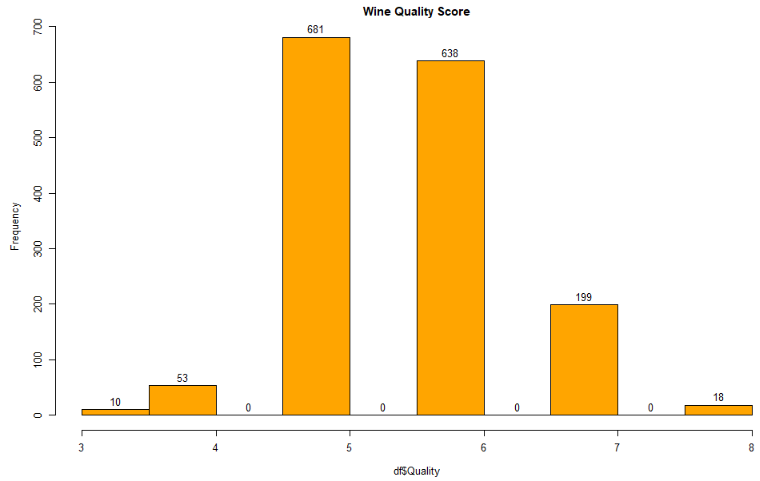
* pH and Fixed Acidity -0.68
* Volatile Acidity and Citric Acid -0.55
* pH and Citric Acid -0.54
* Alcohol and Density -0.5

The top positively correlated variables are:

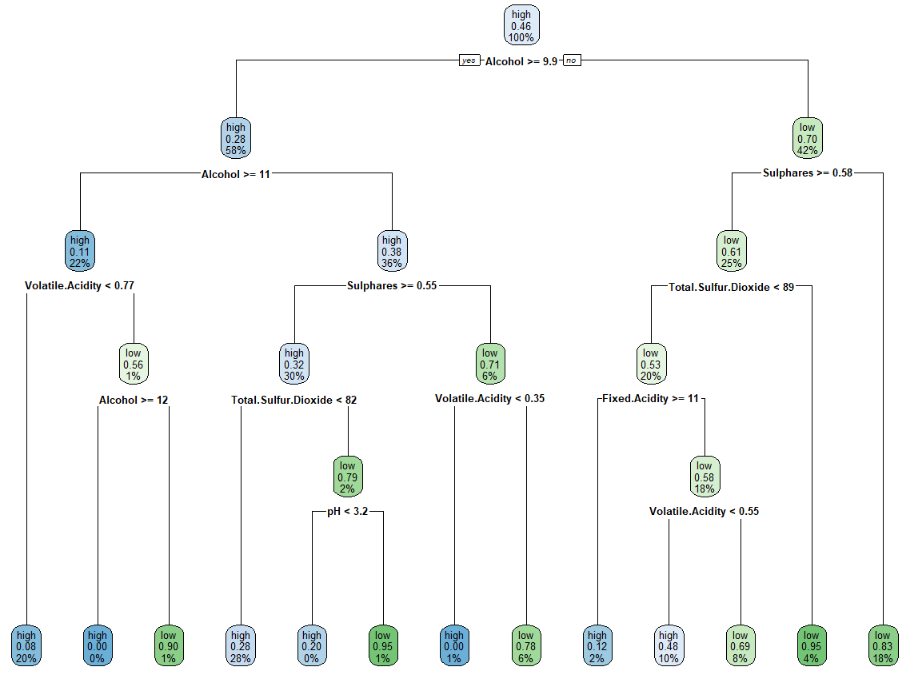
* Three variable pairs are tied for the top positive correlation coefficient of 0.67
* Total Sulfur Dioxide and Free Sulfur Dioxide
* Citric Acid and Fixed Acidity
* Density and Fixed Acidity
* Quality and Alcohol 0.48



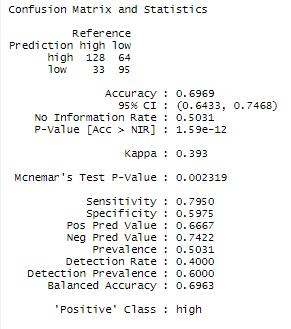
The Quality score will be used as the predictor for this analysis. The Quality score has a range between 0 and 10 with zero being the lowest score and 10 being the highest score. High = 855/53.47%, low = 744/46.53%, 1,599 total. Over 50% have a score of 6 or above. We will create a new column and group scores 6+ as "High" and all others as "Low".



A new column 'qualityScore' is added to the dataframe. All rows with a score of 6 or above is labeled as High and all scores below 6 are labeled as Low. The data is split 80/20 into test and training datasets. A Decision Tree is then created using the qualityScore as the predictor.



A confusion matrix is created with the predictions from the test dataset. The accuracy score is a little low at 69.69% but we will create a Random Forest and compare the results.



The accuracy of the Decision Tree is higher at 70.31%

