FINAL PROJECT: WINE QUALITY

INTRODUCTION

This project is based on set of data of various chemical properties of wine and quality ranking of the wine samples as based on sensory evaluation by wine experts. There are two source datasets, representing red and white vine types of the Portuguese Vinho Verde wine. The following website provides more information about this specific type of wine: Link to Vinho Verde. The quality of the wine was assessed by taking the median of at least 3 evaluations by experts and is a number between 0 and 10 with 10 being the best and 0 the worst.

The breakdown of the data is such that the red wines dataset consists of 1599 records, while the white wines dataset consists of 4898 records. They both contain the same variables listed below:

- fixed acidity - numeric, continuous predictor
- volatile acidity - numeric, continuous predictor
- citric acid - numeric, continuous predictor
- residual sugar - numeric, continuous predictor
- chlorides - numeric, continuous predictor
- free sulfur dioxide - numeric, continuous predictor
- total sulfur dioxide - numeric, continuous predictor
- density - numeric, continuous predictor
- pH - numeric, continuous predictor
- sulphates - numeric, continuous predictor
- alcohol - numeric, continuous predictor
- quality - numeric, ordinal target

During the progress of this project, we have combined the two source datasets into one main set and assigned a dichotomous character variable: color (Red/White) to distinguish between the two types of wine. Breaking down this dataset into train and test sets, we utilized different predictor variables in classification models such as k nearest neighbors and random forest in attempt to come up with a reliable predictive model of the wine quality based on its chemical makeup. The sections below describe this process in further detail.

METHOD/ANALYSIS

We start our analysis by combining the two source data sets (Red Wine and White Wine) into one main set. Before combining them, We add a color variable to both sets and assign an identifier (Red/White) to distinguish between the two types of wine. Once combined, we then break down our data into two partitions, dfTrain and dfTest with an approximately 9:1 amount of records ratio between the two. We check the main dataset to see if any values are missing by plasing a logical check for null on all fields and then summing up the result. The outcome is a zero, meaning that there are no missing values in the source data. We then begin building our models by using the train set and test the results on the test set that we created.

We then proceed with a univariate analysis of our training set data by creating box plots and histograms of each of the eleven predictor variables. Both boxplots and histograms are broken up by wine type (Red/White). The boxplots indicate a presence of outliers in several predictors. They also show a wider spread of alcohol in both wines, a higher presence of total sulfur dioxide in the white wines and a higher presence of fixed and volatile acidity in the red wines. The histograms show a slight skewedness to the right and the alcohol histogram displays a shape with multiple peaks.

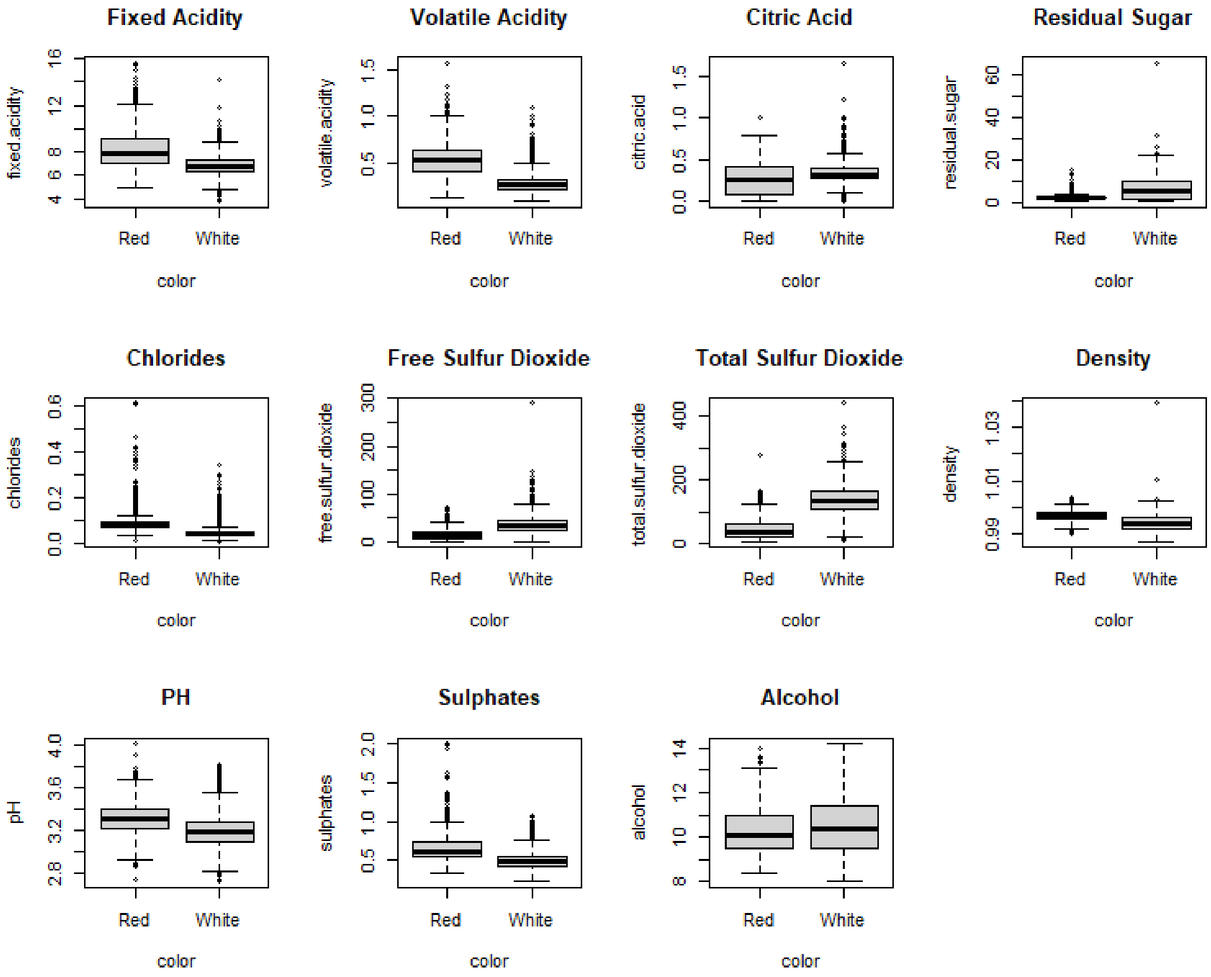
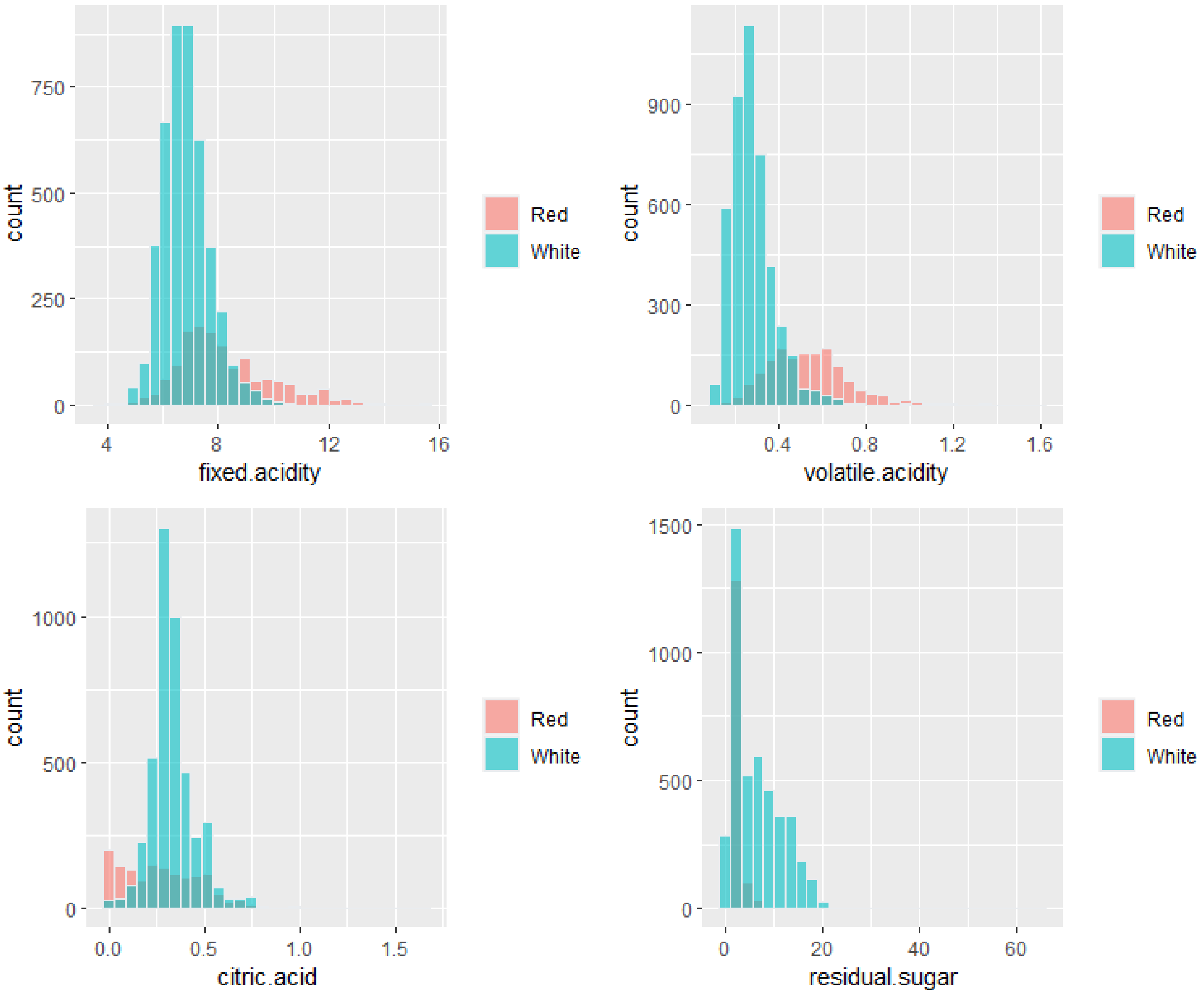
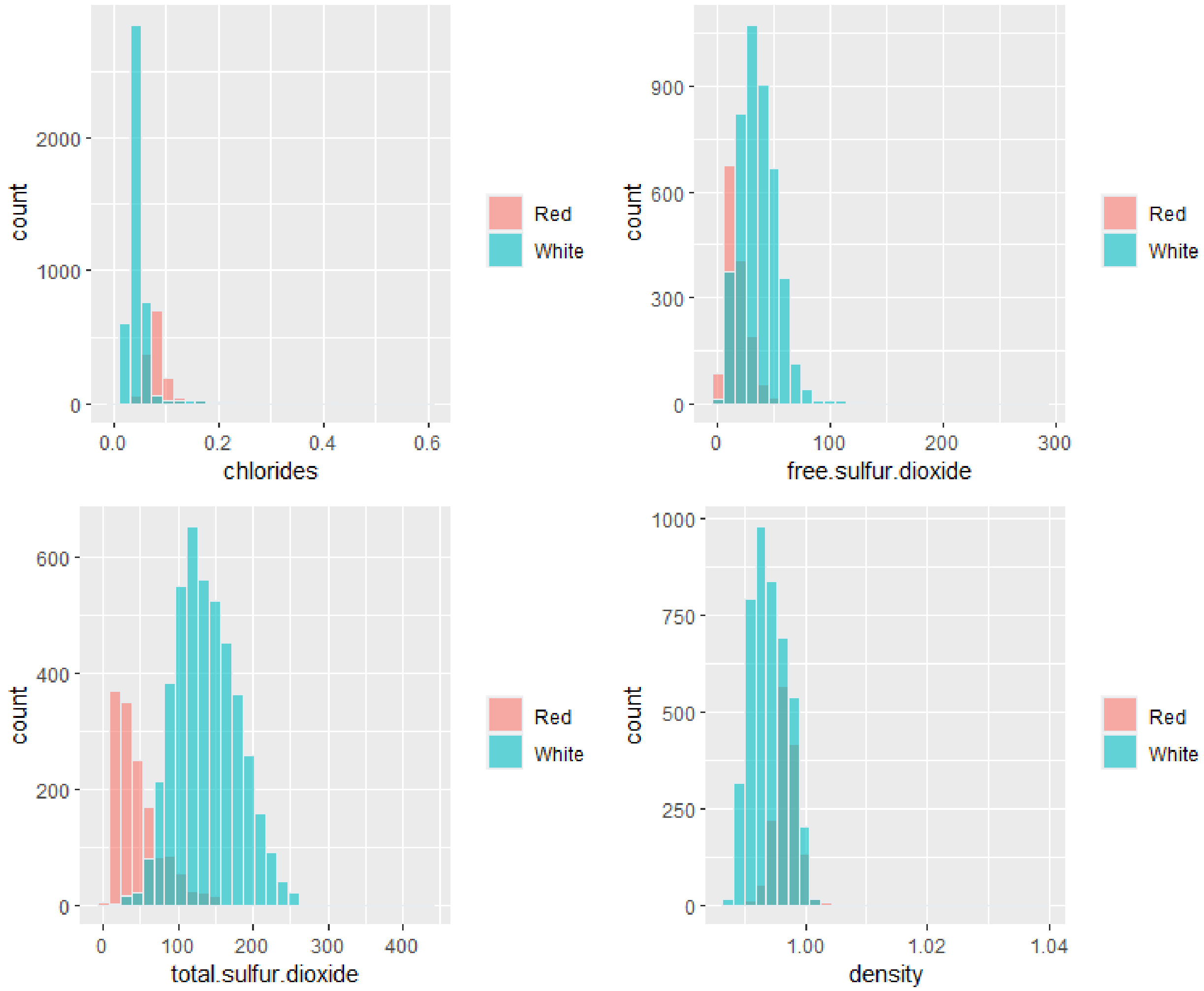
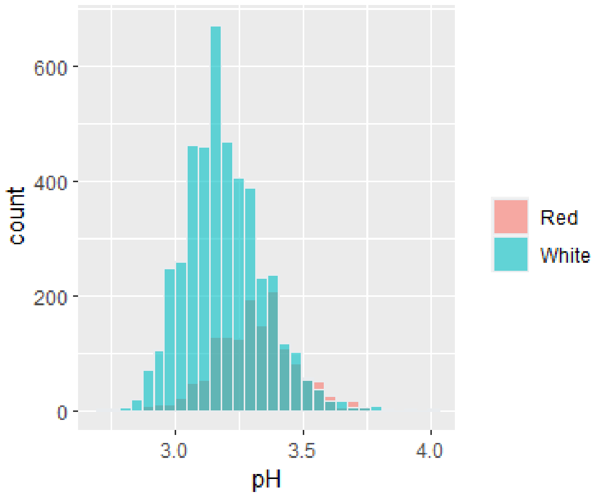
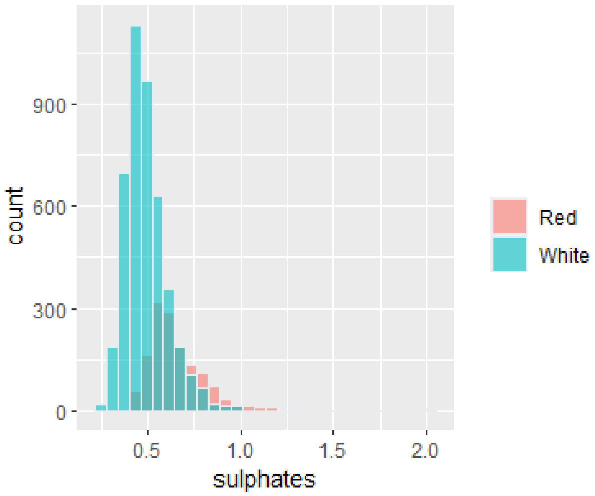


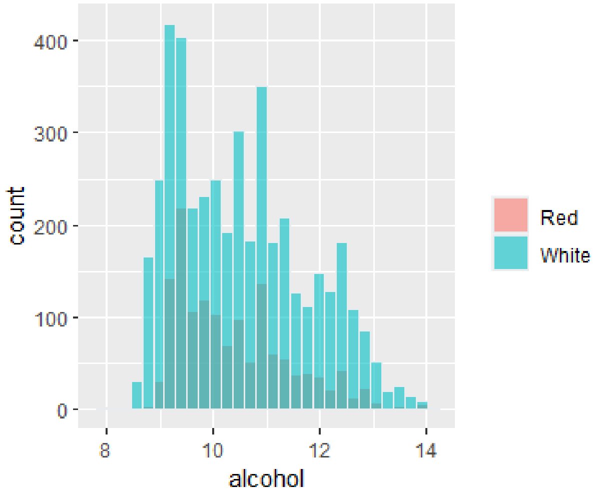
Figure 1: boxplots of predictor variables


count
750
500
250
0
4
8
12
16
fixed.acidity
Red
White
count
900
600
300
0
0.4
0.8
1.2
1.6
volatile.acidity
Red
White
count
1000
500
0
0.0
0.5
1.0
1.5
citric.acid
Red
White
count
1500
1000
500
0
0
20
40
60
residual.sugar
Red
White

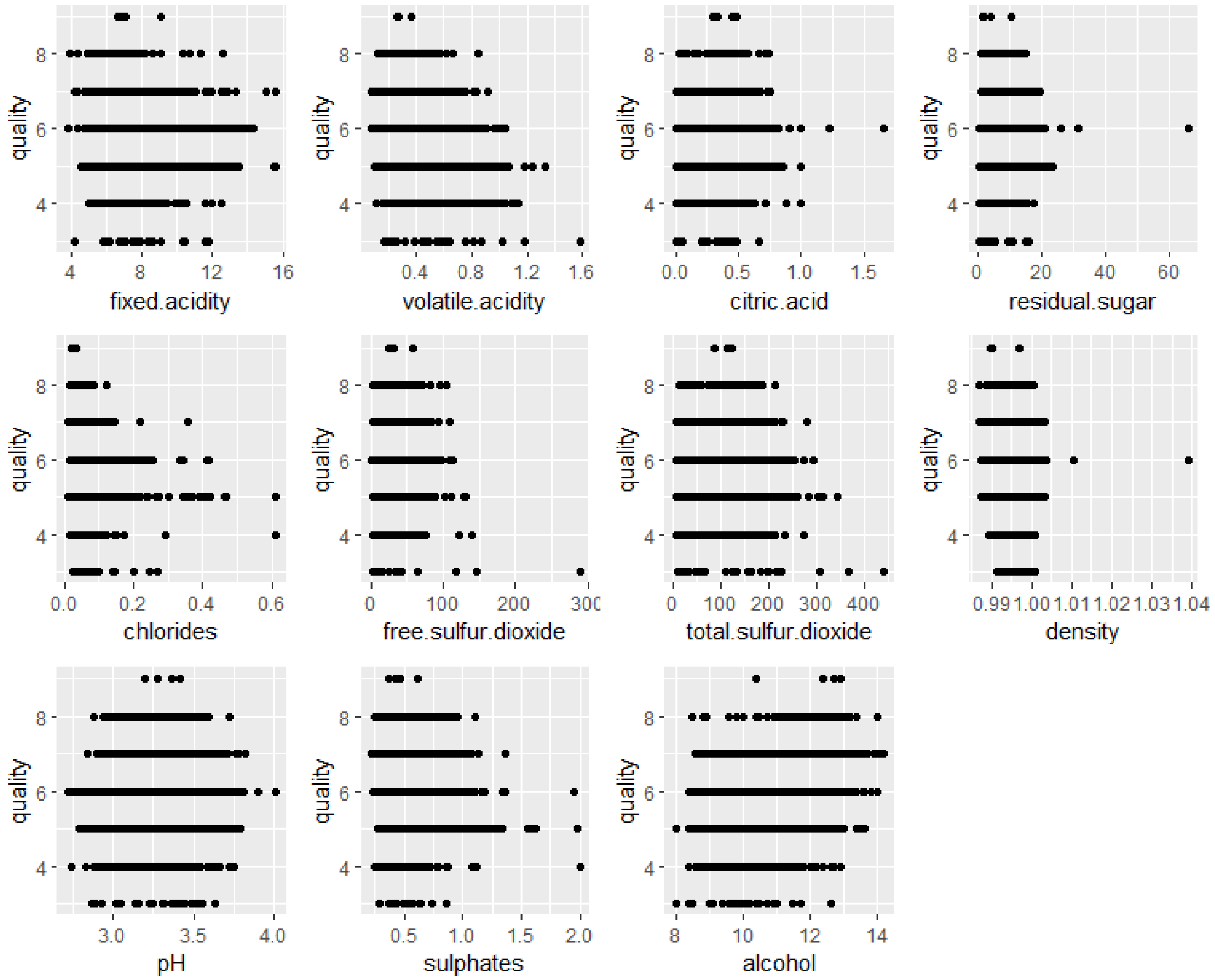

count
2000
1000
0
0.0
0.2
0.4
0.6
chlorides
Red
White
count
900
600
300
0
0
100
200
300
free.sulfur.dioxide
Red
White
count
600
400
200
0
0
100
200
300
400
total.sulfur.dioxide
Red
White
count
1000
750
500
250
0
1.00
1.02
1.04
density
Red
White



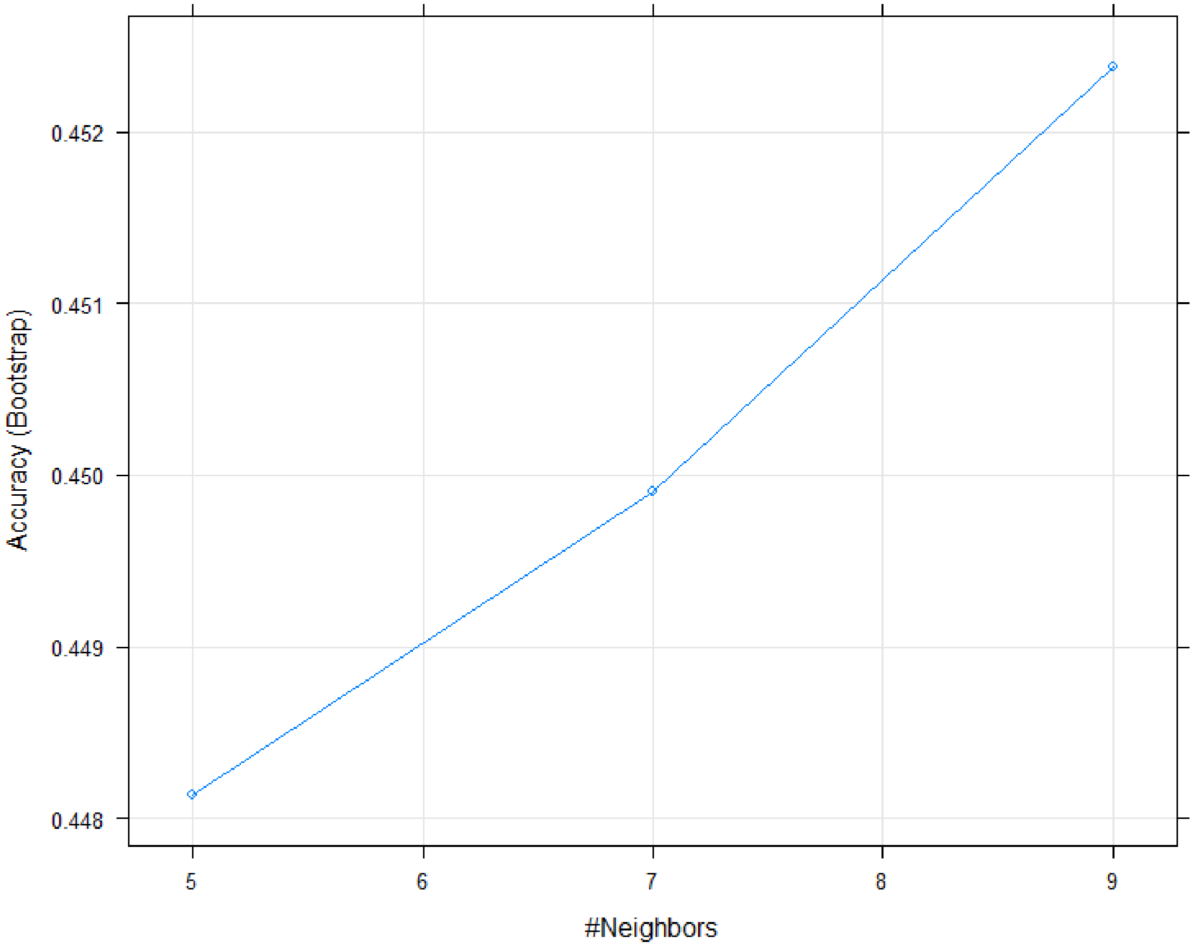




We then perform a bi-variate analysis of predictor variables against the target quality variable to check for any potential correlations between them and based on the plots, we do not see anything of significance:



We then continue to build our first prediction model. Before doing this, we add a factor type variable to both, the train and the test set, representing the quality target variable as a factor. This is needed for quality to be utilized in our trained models. The first model is based on the k nearest neighbors algorithm. We utilize all predictors in building this model and we plot the result, which does not seem to be very good based on the



plot:

We then produce a predictor set by applying the trained knn model to the test set and calculate the accuracy of the model by using a confusion matrix. The accuracy result of approximately 48% is not a good one.

Our next model is a linear discriminate analysis algorithm. Again, we train a model by utilizing all predictors in the train set and then develop a predictor set by applying the trained lda model against the test set and use a confusion matrix to calculate the accuracy of the model, which does only slighter better than the knn model at around 54%.

We proceed to creating a classification tree to see if there is an identifiable pattern for effects of predictors on the target. To achieve this, we use the rpart function with all predictor variables and set the complexity parameter to 0.004 so as not to overwhelm the visual. We then plot the tree and label it. Unfortunately, the tree does not show a significant predictor:

Next, we train a classification tree model against the test set using all predictors. We then visualize the result of this model, observing that it does only slighter better than the knn and the lda models we created previ-

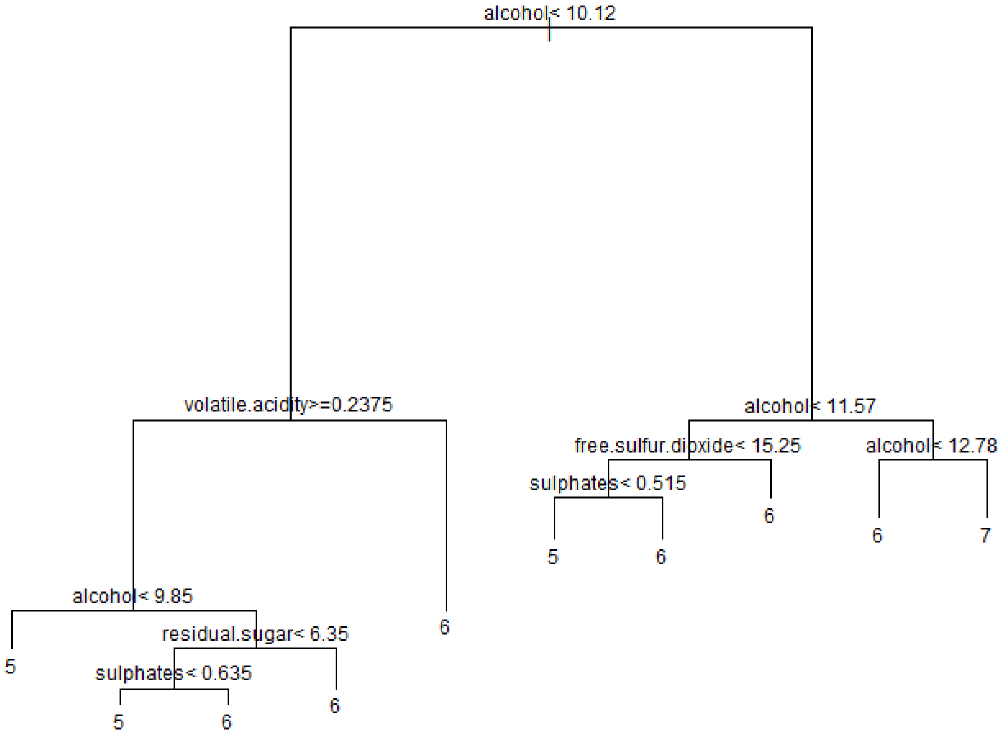
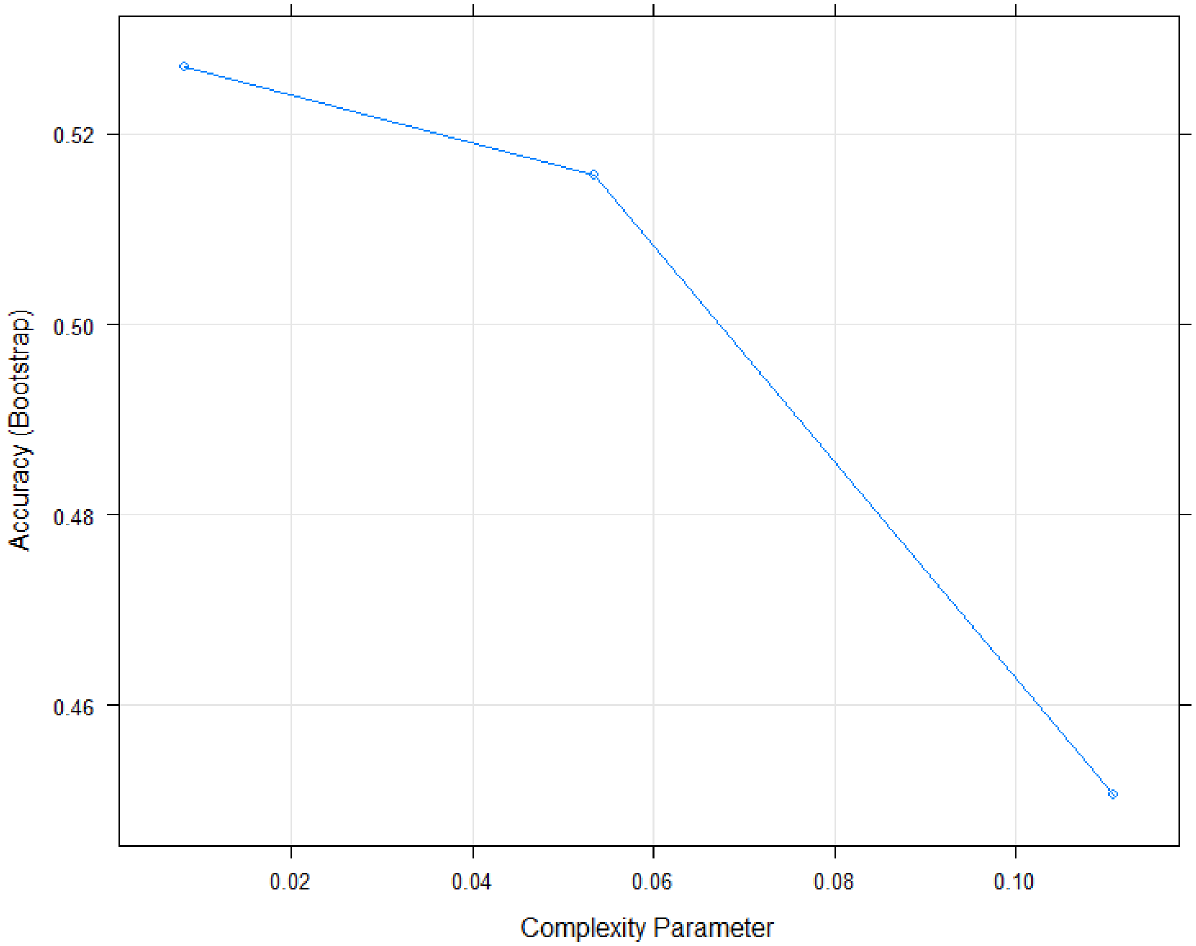


Figure 2: classification tree



ously:

After this we create a prediction set by using the trained rpart model against the test set and use the confusion matrix to check the accuracy, which is in line with the observed model plot at about 54%.

We then move on to our final model - the random forest algorithm. We train this model again using all predictors against the train set and then build a predictor set against the test set and calculate the accuracy of the model by using the confusion matrix. Here we observe the best result of all the previously created models of accuracy around 71%. The random forest is our final model, although the result is still not very good in terms of prediction accuracy.

RESULTS

The following is a table of results of the various models created during the process of this project:

Method	Accuracy (%)
k nearest neighbors	48.1
linear discriminate analysis	53.8
classification tree	54.1
random forest	70.1
FINAL MODEL RESULT	**70.1**

Unfortunately based on the work in this project, we are not able to come up with a good model for a predictive algorithm for the quality of wine based on its chemical makeup. Our first three models all center around a 50% mark of accuracy, which is highly unreliable. We do improve the result by utilizing the random forest algorithm, providing us with a result of **70.1%**, a major improvement of around 20% from the other three models utilized in this analysis. However, while somewhat usable, this result leaves more to be desired in terms of reliability. The performance of our model is quite good, taking only seconds to run. This however may also be attributed to a someone small dataset of only a few thousand records.

Overall, we do observe a degree of connection between the chemical makeup and the resulting quality, albeit not a very strong one. Since quality was determined by experts through physical sensation such as taste and smell, it is possible that due to differences in palates, preferences and bias, chemical makeup does not play an overly important role in determining the quality of wine. It is also possible that there are other significant predictors, such as maker, type of grape, year of production, etc. that were not included in the source data that would present a major significance in predicting wine quality.

CONCLUSION

This project allowed us to examine the importance of various chemical ingredients of the Portugese Vinho Verde wine in its overall quality. Through examination of data, its analysis and training of four predictive models, we were not able to determine a very significant connection between these predictors and the target. There is a chance that taster bias and a lack of important variables plays a role in this result. This project is a very good beginning in building a more sophisticated model based on larger datasets with more variables to better determine a more reliable way of predicting wine quality.