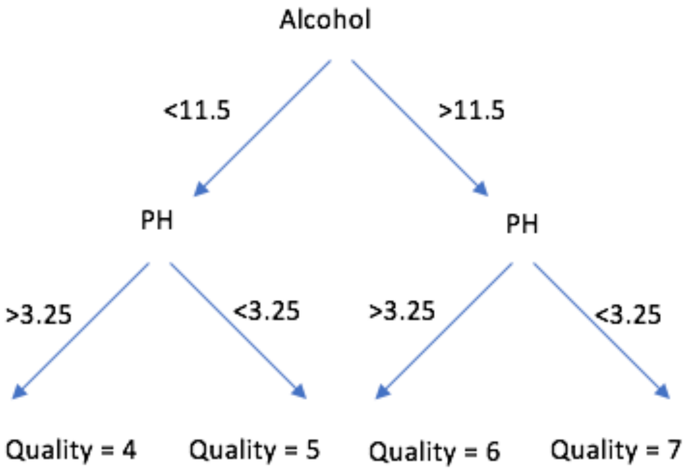
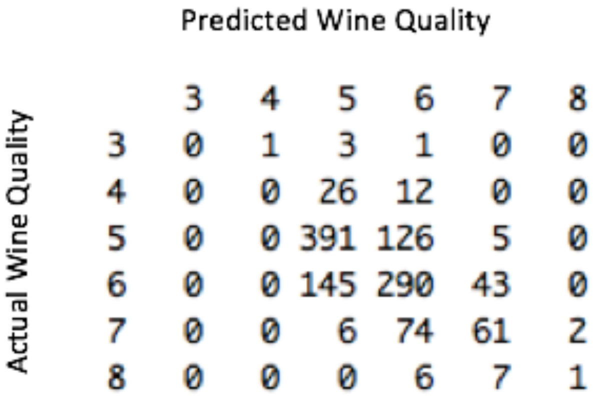
For this project, I chose to use a random forest classification model to try and predict wine quality. A random forest model is made up of a large group of decision trees. A decision tree can be used to classify wines into different quality rating groups based on a few different variable conditions. This can be illustrated by the graph to the right.



I used a classification model because the wine quality ratings can be viewed as discrete numbers since they are whole numbers and not rated on a continuous scale. I randomly split the original wine data set into a training dataset (25%) and a testing dataset (75%). Using quality as the output variable and all of the other variables in the training dataset as inputs, I created the model with 1000 trees using a package in R. I then used the model to predict the wine quality of the testing dataset to determine the accuracy of the model. The model was able accurately wine quality about 62% of the time. The predicted value was no more than 1 off from the true quality about 97% of the time. A sample of the results can be seen to the right.

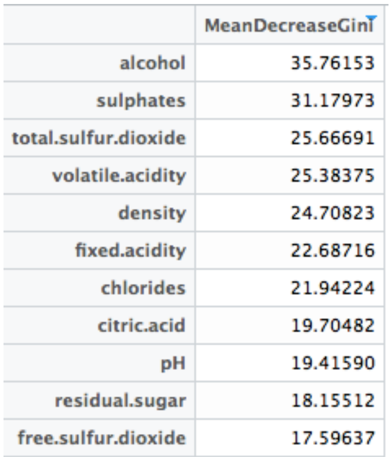


Predicted Wine Quality

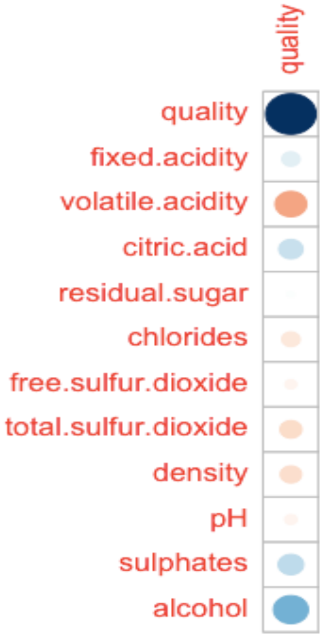
Actual Wine Quality	3	4	5	6	7	8
3	0	1	3	1	0	0
4	0	0	26	12	0	0
5	0	0	391	126	5	0
6	0	0	145	290	43	0
7	0	0	6	74	61	2
8	0	0	0	6	7	1

I also tried out a couple other models. The first was a multiple linear regression model. This model only used the statistically significant variables as inputs, which were volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol. Since model creates continuous predictions, I rounded the predictions to the nearest whole number. The accuracy rate for this model was 60%. I ultimately decided against using a regression model because I felt a classifier model was a better fit for the data, although both models do have pretty similar accuracy percentages. I also tried a single decision tree model. This, however, resulted in the model only predicting wine qualities of 5 and 6 with an accuracy of only 55.25%.

Based on the random forest model the two most important variables were alcohol and sulphates, with total sulfur dioxide, volatile acidity, and density also being important aspects. The chart at the bottom right shows a correlation plot between wine quality and each of the variables. Large blue circles represent a strong positive correlation and large red circles represent strong negative correlation. Higher alcohol and sulphate content generally translates to higher quality ratings. On the flip side, higher total sulfur dioxide, volatile acidity, and densities generally results in lower wine quality ratings.



	MeanDecreaseGini
alcohol	35.76153
sulphates	31.17973
total.sulfur.dioxide	25.66691
volatile.acidity	25.38375
density	24.70823
fixed.acidity	22.68716
chlorides	21.94224
citric.acid	19.70482
pH	19.41590
residual.sugar	18.15512
free.sulfur.dioxide	17.59637


quality
quality
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
free.sulfur.dioxide
total.sulfur.dioxide
density
pH
sulphates
alcohol