**Wine Quality Analysis**

Jihye Hong

**Aims:** Discover the correlations between sensory data(quality) and physicochemical data about wine quality. As we predict human wine taste preferences that are based on easily available analytical tests, it would be used to improve wine production.

**1. Data Exploration**

Each dataset for red and white wine demonstrates normal distribution. The distribution of wines shows the medium level of quality is more common than bad or good quality wine, thus overall data is inevitably unbalanced. In red wine data set, there is no quality level of 1, 2 and 9. (In this analysis, numerical wine quality was considered as 3 classes. Bad: 3,4 Medium:5,6 Good:7,8,9.) More or less all attributes have outliers, it can be legitimate objects so outliers also considered for the model.

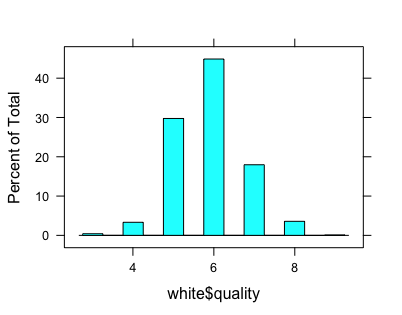
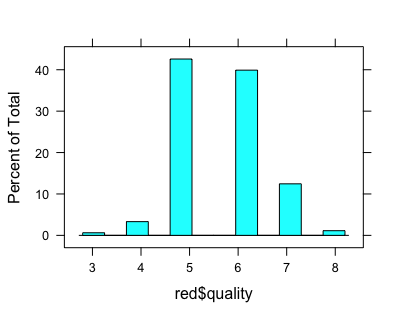


Figure 1.1 Distribution of wine quality

**2. Feature selection**

Each wine dataset has 11 input variables based on physicochemical tests and 1 output variable(quality) based on sensory data. As only important input variables were selected for building a model, it can achieve better performance and simplicity. Two different feature selection methods were applied to building a model. First, features were manually selected from the correlation plots (Figure 2.1). Based on the plots, top 5 attributes with highest correlations with quality were selected as predictors. Second, the best subset algorithm was used for automatic feature selection.

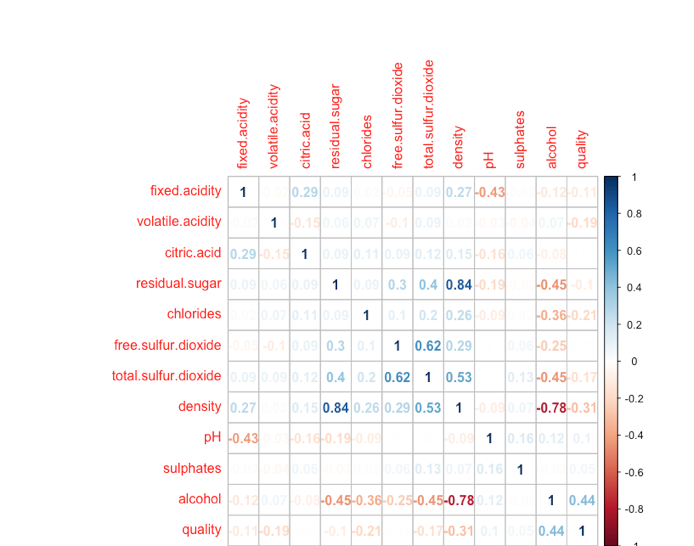
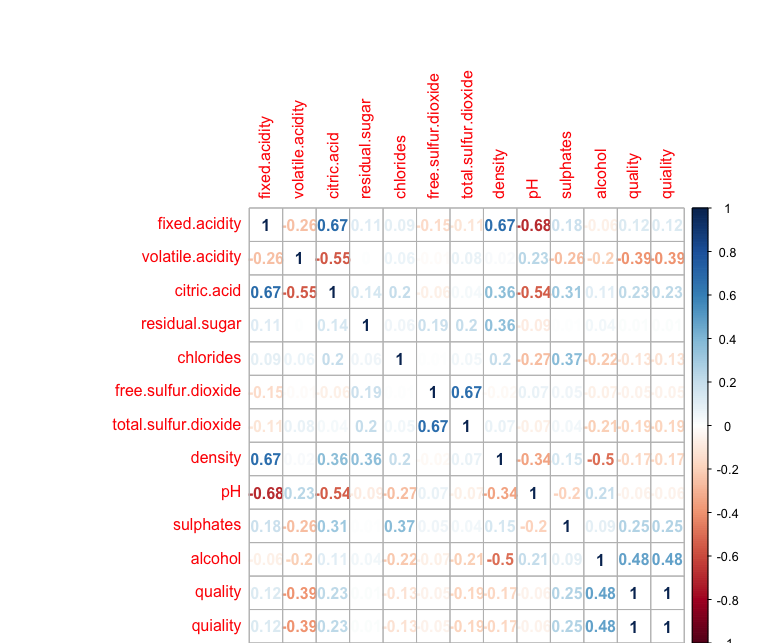


Figure 2.1 Correlation plots of red wine(left) and white wine(right)

Top 5 predictors for red wine: alcohol, volatile.acidity, sulphates, citiric.acid and total.sulful.dioxide

Top 5 predictors for white wine: alcohol, density, chlorides, volatile.acidity and total.sulfur.dioxide

**3. Model building**

In terms of algorithms used for model building, tree-based models were used based on the feature selection. As a result, automatic selection demonstrated a better performance than manual selection. So for building learning machines, predictors from automatic feature selection were applied.

This modelling consists of two parts, random forest and decision tree models for regression. These regression learning machines predicted the quality in a continuous scale rather than the discrete integer values. If the prediction result were processed as integers, it would cause poor accuracy as rounding the prediction values. Below are the prediction results of each model.

**Wine quality prediction test results**

|  |  |  |
| --- | --- | --- |
|  | Red wine | White wine |
| Random Forest error rate | 33.4% / | 37.8% |
| Regression tree error rate | 44.9% | 55.4% |

Random Forest outperformed Regression Decision Tree for both datasets.

**Classification tree results**

The percentage of correct predictions

Red wines: 82.5

White wines: 77.33769

Classification tree’s accuracy is significant. But here is a trick. The given dataset is extremely unbalanced and even if a tree predicts all prediction of objects as ‘medium’, it can achieve the accuracy level of 83%! This is a critical limitation of the classification model for wine quality dataset, as I applied several different models, all of them resulted that most of the prediction outputs were ‘medium’.

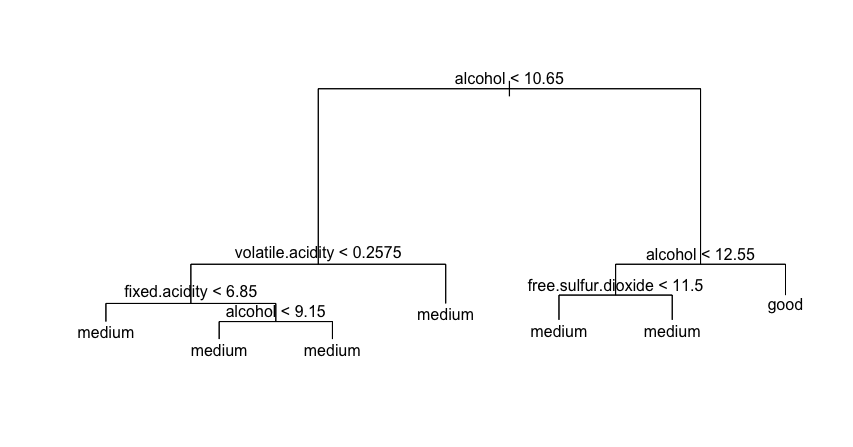
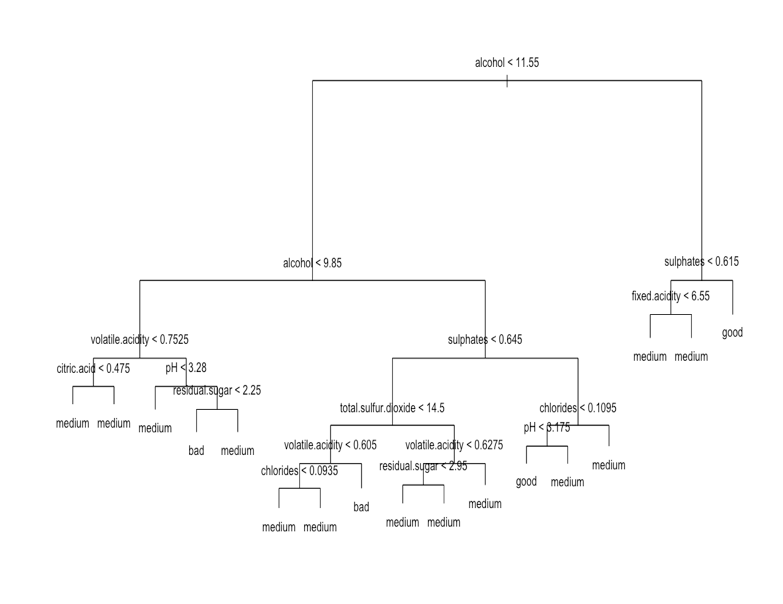


Figure 3.1 Classification tree of red wine(left) and white wine(right)

**4. Conclusions**

With this dataset, it’s hard to say the chemical properties of wines well support wine quality very much. Even though the random forest regression model achieved relatively high performance, it is not very significant (under 70%). I suppose if it has more dataset especially low and high quality wines for well balanced datasets, a model can be more accurate at the quality of wines.