GEORGE MASON UNIVERSITY
College of Engineering and Computing

FINAL PROJECT REPORT

WINE QUALITY

Applied Statistics and Visualizations for Analytics (STAT-515-002)

Prof. Richard Sigman

Due date: Dec 8,2022

PRAVALLIKA AVULA
KINNERA VEERAPANENI
ANURAG GOUD KOUKUNTLA

1) PROJECT DESCRIPTION:

The main goal of this final project is to choose a data set from any source that is prepared for analysis and conduct some research using concepts, such as graphing, multivariate analysis (regressions)to create some suitable models that can help us predict the necessary information and the effects of response variables on the predictor variable in the data set.

The first step in the data analysis process is to comprehend the dataset as well as the reasons for and relationships between variables. The next step is to clean the dataset by performing exploratory analysis to check for null values and ensure that it is error-free. Making visualizations with subsets of data for better data understanding is the third step. The fourth step is to determine the response variable and any potential predictor variables. The final step is to develop an appropriate prediction model. This is the entire data analysis procedure.

2) DATA SET:

SOURCE: This project has two datasets namely red and white wine samples from the north of Portugal.

https://archive.ics.uci.edu/ml/datasets/wine+quality

CONTEXT:

Wine was always thought of as a luxury item, but today a larger variety of customers are enjoying it. With 3.17% of the market in 2005, Portugal ranked tenth in the world for wine exports. From 1997 to 2007 the country's Vinho Verde wine exports (from the northwest region) climbed by 36%. The wine business is making investments in new technology to help its expansion in both the winemaking and wine-selling operations. In this perspective, wine certification and quality evaluation are essential components. This study will focus on Vinho Verde, a distinctive beverage from Portugal's Minho (northwest) region. The two most popular varieties from the designated Vinho Verde region, white and red (rosé is also produced), will be examined in this work. Quality evaluation is often part of the certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices). While sensory tests mostly rely on human expertise, physicochemical laboratory tests frequently used to classify wine include assessment of density, alcohol, or pH levels. It should be emphasized that because taste is the sense of the human body that is least known, classifying wines is a challenging task. Furthermore, the connections between physicochemical analysis and sensory analysis are intricate and still little understood. Using only protected designation of origin samples that had been evaluated at the official certification entity, data were gathered from May 2004 to February 2007. (CVRVV). The CVRVV is an inter-professional group with the mission of enhancing vinho Verde quality and marketing. A computerized system (iLab) that automatically controls the wine sample testing procedure from producer requests through laboratory and sensory analysis was used to capture the results. Analytical or sensory tests are identified by their respective entries in the final database, which was produced as a single page (.csv).

VARIABLES:

- FIXED ACIDITY - most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- VOLATILE ACIDITY - the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- CITRIC ACID - found in small quantities, citric acid can add 'freshness' and flavor to wines
- RESIDUAL SUGAR - the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet.
- CHLORIDES - Amount of salt present in wine
- FREE SULFUR DIOXIDE - the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine.

- TOTAL SULFUR DIOXIDE - amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- DENSITY - the density of water is close to that of water depending on the percent alcohol and sugar content
- PH - describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- SULPHATES - a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
- ALCOHOL - Percent of alcohol present in wine
- QUALITY (score between 0 and 10) (UCI Machine Learning Repository: Wine Quality Data Set, n.d.)

The database was altered during the preprocessing step so that each row now contained a unique wine sample (with all tests). Only the most popular physicochemical tests were chosen to avoid excluding cases. Two datasets were created with 1599 red and 4898 white instances because the red and white tastes are very distinct from one another (Cortez et al., 2009).

3) EXPLORATORY ANALYSIS:

In the Exploratory Data Analysis, we must understand the structure, quality, variability, and co-variability of the dataset, as well as some of the dataset's interesting relationships.

```
> colSums(is.na(red_df))
      fixed.acidity    volatile.acidity         citric.acid      residual.sugar
                  0                   0                   0                   0
          chlorides free.sulfur.dioxide total.sulfur.dioxide             density
                  0                   0                   0                   0
                 pH           sulphates             alcohol             quality
                  0                   0                   0                   0
> dim(red_df)
[1] 1599   12
> #Data overview
> str(red_df)
'data.frame':   1599 obs. of  12 variables:
 $ fixed.acidity       : num  7.4 7.8 7.8 11.2 7.4 7.4 7.9 7.3 7.8 7.5 ...
 $ volatile.acidity    : num  0.7 0.88 0.76 0.28 0.7 0.66 0.6 0.65 0.58 0.5 ...
 $ citric.acid         : num  0 0 0.04 0.56 0 0 0.06 0 0.02 0.36 ...
 $ residual.sugar      : num  1.9 2.6 2.3 1.9 1.9 1.8 1.6 1.2 2 6.1 ...
 $ chlorides           : num  0.076 0.098 0.092 0.075 0.076 0.075 0.069 0.065 0.073 0.071 ...
 $ free.sulfur.dioxide : num  11 25 15 17 11 13 15 15 9 17 ...
 $ total.sulfur.dioxide: num  34 67 54 60 34 40 59 21 18 102 ...
 $ density             : num  0.998 0.997 0.997 0.998 0.998 ...
 $ pH                  : num  3.51 3.2 3.26 3.16 3.51 3.51 3.3 3.39 3.36 3.35 ...
 $ sulphates           : num  0.56 0.68 0.65 0.58 0.56 0.56 0.46 0.47 0.57 0.8 ...
 $ alcohol             : num  9.4 9.8 9.8 9.8 9.4 9.4 9.4 10 9.5 10.5 ...
 $ quality             : int  5 5 5 6 5 5 5 7 7 5 ...
>
```

Figure 1: R code for data cleaning for red wine

```
> colSums(is.na(white_df))
      fixed.acidity    volatile.acidity         citric.acid      residual.sugar
                  0                   0                   0                   0
          chlorides free.sulfur.dioxide total.sulfur.dioxide             density
                  0                   0                   0                   0
                 pH           sulphates             alcohol             quality
                  0                   0                   0                   0
> dim(white_df)
[1] 4898   12
> #Data overview
> str(white_df)
'data.frame':   4898 obs. of  12 variables:
 $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
 $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
 $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
 $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
 $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
 $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
 $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
 $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
 $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
 $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
 $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
 $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
>
```

Figure 2: R code for data cleaning for red wine

The above figure gives you the process that we followed for the cleaning of the data set by finding the null values in all the columns. As there are no null values in the data set, we didn't omit any values. The summary of the cleaned dataset named red_df is also derived. Similarly, we did the same for white wine dataset.

I have made further analysis and visualizations to understand the interaction between our numeric variables of interest and our dependent variable of quality.

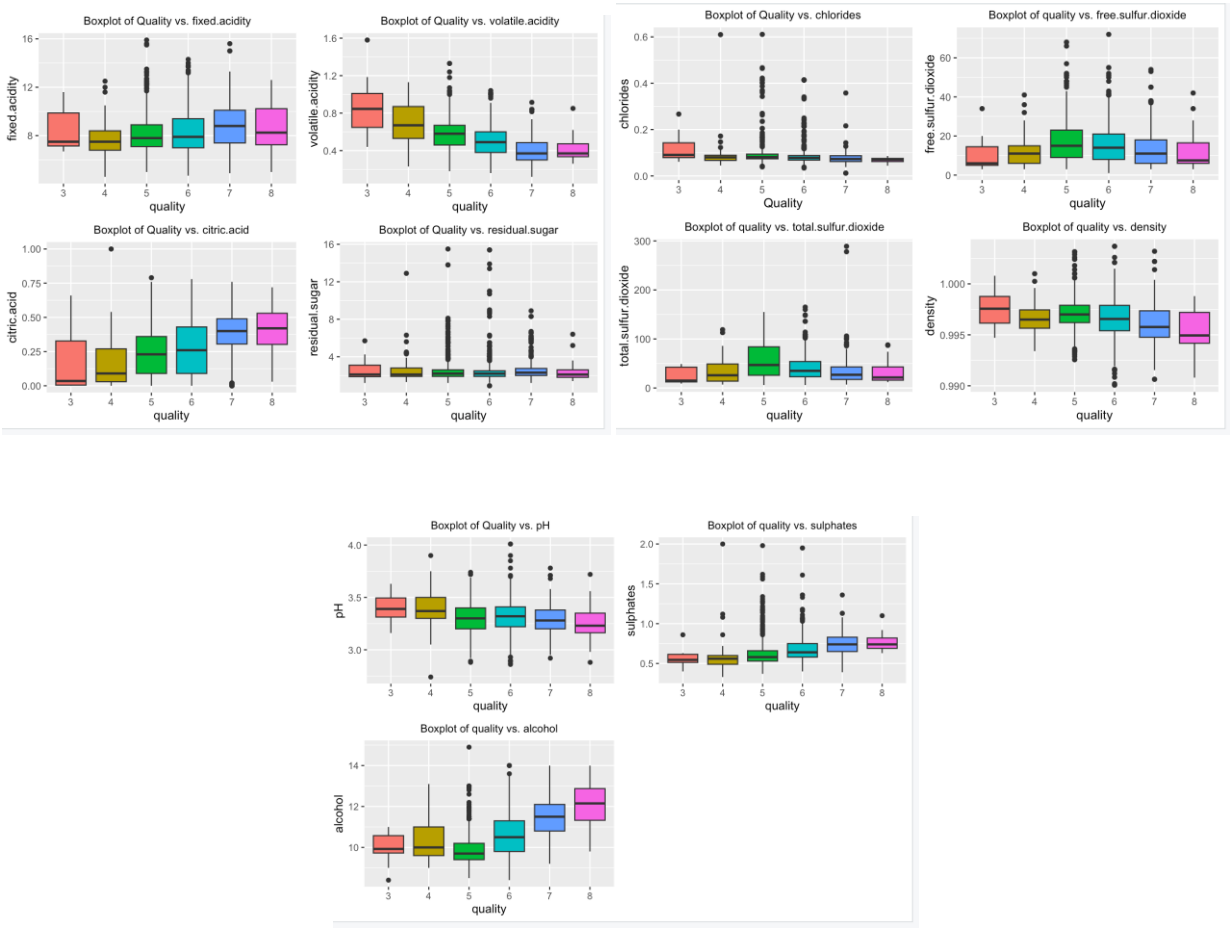

Figure 5: Box plots for relation between quality and variables - red wine

There are three distinct patterns that can be seen. The first is that quality and critic acid, alcohol, and sulphates have good correlations. Wines having a greater alcohol content should be highly rated in quality even though they may be less well-liked. Second, quality and volatility, pH, density, and acidity have a negative relationship. It seems sense that wines with less sweetness and less acidity are preferred in quality assessments. The variables- residual sugar, chlorides, and total sulfur dioxide do not significantly correlate with quality.

In the white wine dataset, it is observable that there is a positive correlation between quality and alcohol, pH, and volatile acidity. Whereas sulphates, and free Sulphur dioxide have negative relationship. It seems that citric acid and chlorides do not have much correlation with the quality of white wine.

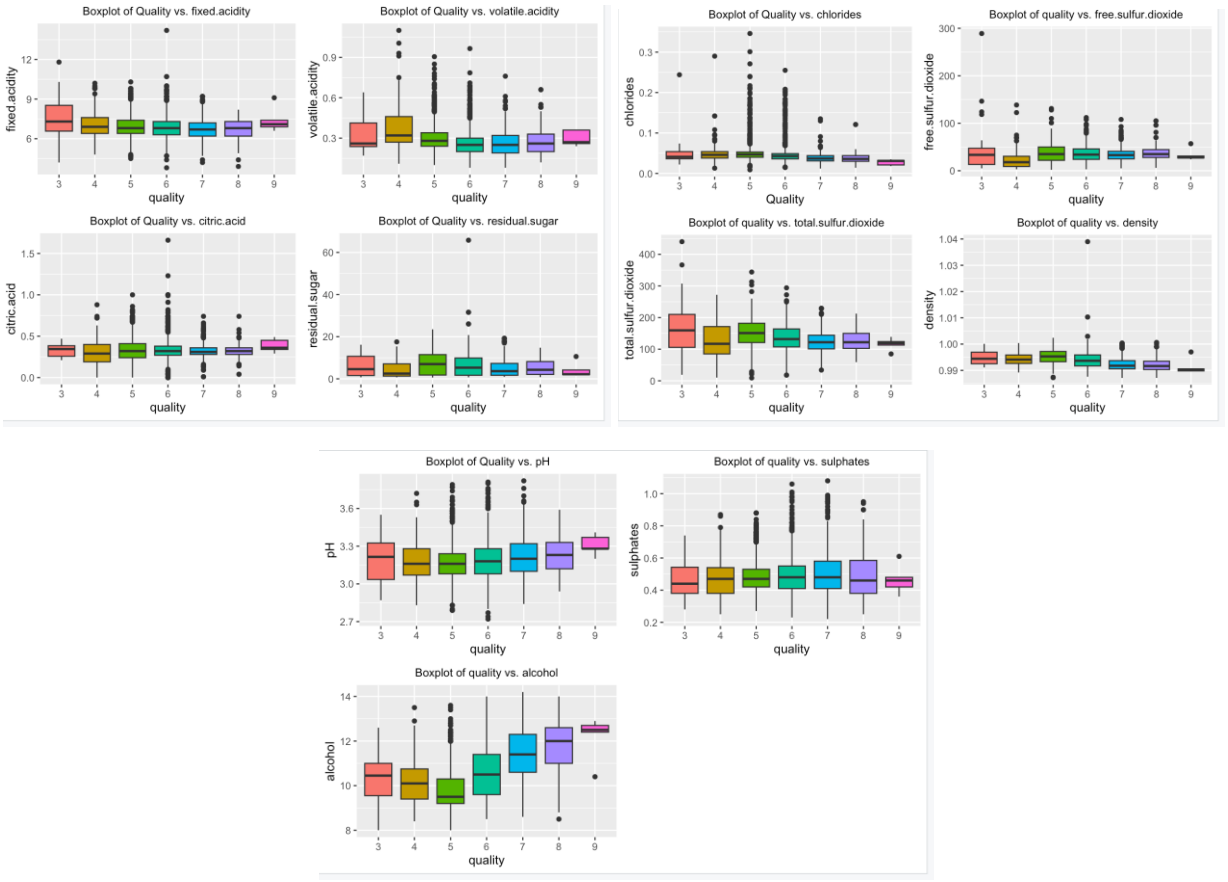

Figure 6: Box plots for relation between quality and variables - white wine

RESEARCH QUESTIONS:

- What wine features are important to get a promising result?
- Which regression method can give accurate results?

4) DATA ANALYSIS AND METHODS:

For this project we have used R studio software for data visualizations and analysis. These are the important libraries that were imported for our code:

1. GGally – 'GGally' extends 'ggplot2' by adding several functions to reduce the complexity of combining geometric objects with transformed data.
2. corrplot - Provides a visual exploratory tool on correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables.
3. Rpart - For building classification and regression trees
4. Metrics - Used in supervised machine learning.
5. Glmnet - Fits generalized linear and similar models via penalized maximum likelihood
6. Caret - Package contains tools for: data splitting. pre-processing.
7. Rattle - A Graphical User Interface for Data Mining using R
8. Party - Toolbox for recursive partitioning.

I have only considered red wine data set for my research questions.

Coming to the research question, it is important to know the important variables to enhance the wine quality. To find the important variables, I created a random forest named redwineRF. The Mean of squared residuals of the random forest is 0.3154676 and % Var explained is 51.6.

The important factors are then plotted as bar graphs, and we can see them below. From the plot we can see that the important variables for red wine are Alcohol, Sulphates, Volatile acidity, Density, Total Sulphur dioxide. With the given dataset we can tell that the above variables are the most important to describe and enhance the quality of the wine. These top predictors are also important to build a model. It is accomplished through the use of MDI (Gini Importance or Mean Decrease in Impurity), which calculates the importance of each feature as the sum of the number of splits (across all trees) that include the feature, proportionally to the number of samples it splits.

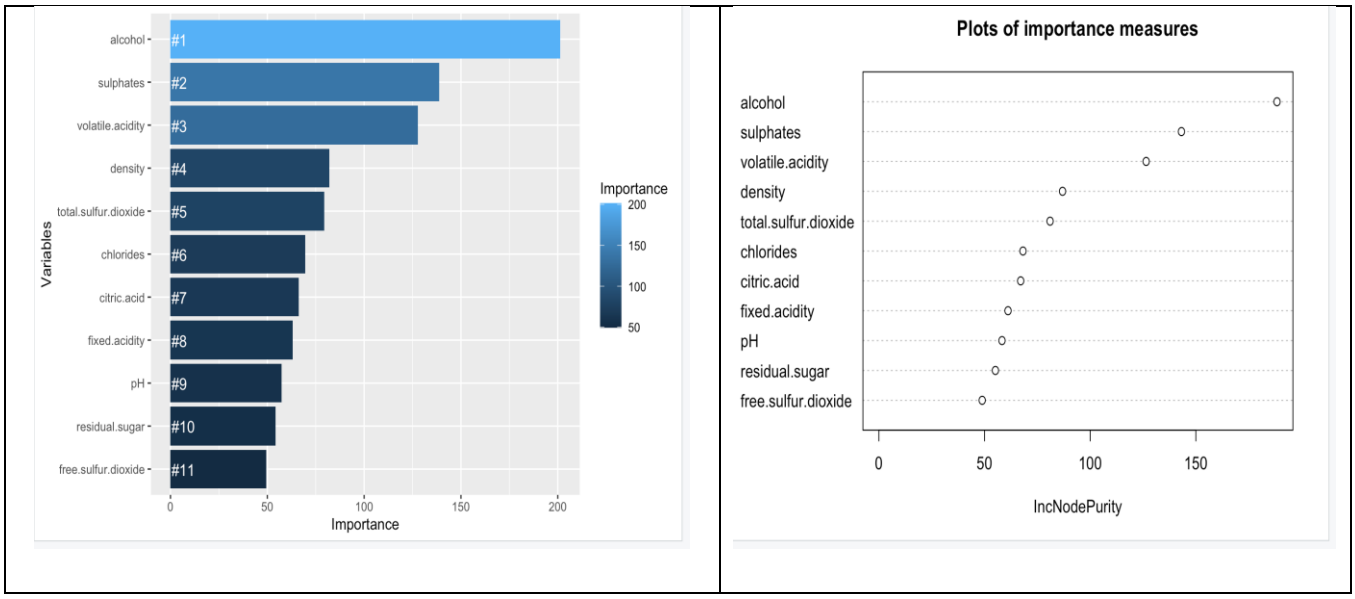

Figure 7: Importance plot

I conducted a correlation analysis between our independent variables and our dependent variable, quality, to determine which factors are most likely to have an impact on the quality of red wine. Following the analysis, a list of variables of interest with the strongest correlations to quality was produced.

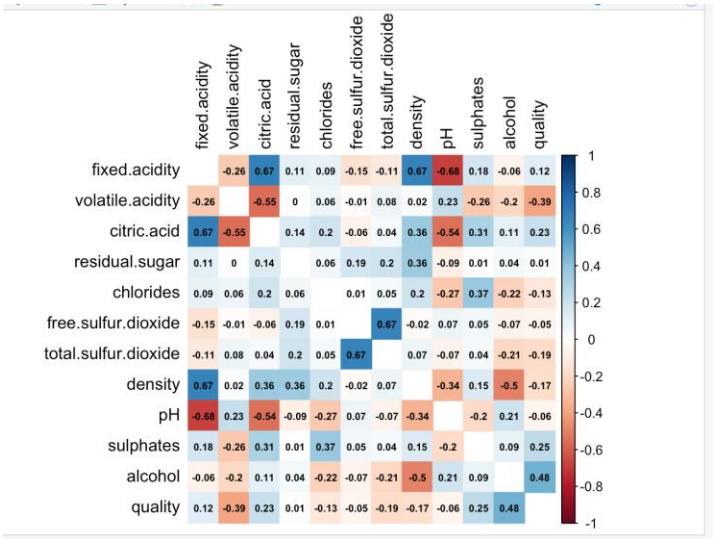

Figure 8: Co-relation matrix

We can tell that alcohol, volatile acidity, sulphates, total Sulphur dioxide has highest co-relation with quality of red wine.

To know the best model for accurate results I performed and compared three regression models namely Linear regression, Lasso regression, Random Forest. I built training and test datasets for building all the models. Below is the summary statistics of linear regression models with top 4 corelated variables namely alcohol, volatile acidity, sulphates, total Sulphur dioxide. The MSE, RMSE, MAE of the residuals is **0.403121, 0.6349181, 0.5233042** respectively.

```
> summary(red_model)

Call:
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
    total.sulfur.dioxide, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29129 -0.41439 -0.06258  0.44028  2.13996

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.0940601  0.2658198  11.640  < 2e-16 ***
alcohol               0.2826331  0.0232910  12.135  < 2e-16 ***
volatile.acidity     -1.0201384  0.1302268  -7.834 1.52e-14 ***
sulphates             0.5282305  0.1206150   4.379 1.35e-05 ***
total.sulfur.dioxide -0.0041631  0.0006662  -6.249 6.73e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6369 on 795 degrees of freedom
Multiple R-squared:  0.3058,    Adjusted R-squared:  0.3023
F-statistic: 87.55 on 4 and 795 DF,  p-value: < 2.2e-16
```

Figure 9: Summary of linear regression model

For the lasso regression with top 6 predictors namely alcohol, volatile acidity, sulphates, total Sulphur dioxide, density, chlorides the summary statistics is as follows.

```
> summary(red_model1)

Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + chlorides +
    total.sulfur.dioxide + sulphates + alcohol, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23168 -0.39914 -0.07055  0.45416  2.04570

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.0774808  0.2912846  10.565  < 2e-16 ***
fixed.acidity         0.0177703  0.0127942   1.389  0.16524
volatile.acidity     -0.9332460  0.1344362  -6.942 8.05e-12 ***
chlorides            -1.3764407  0.4649166  -2.961  0.00316 **
total.sulfur.dioxide -0.0040524  0.0006758  -5.997 3.06e-09 ***
sulphates             0.6873536  0.1345356   5.109 4.06e-07 ***
alcohol               0.2655413  0.0237330  11.189  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6333 on 793 degrees of freedom
Multiple R-squared:  0.3153,    Adjusted R-squared:  0.3101
F-statistic: 60.86 on 6 and 793 DF,  p-value: < 2.2e-16
```

Figure 10: Summary of Lasso regression model

The MSE, RMSE, MAE of the residuals is **0.3976104, 0.6305636, 0.5179756** respectively.

For the random forest with the top 10 predictors i.e., all the predictors the summary statistics is as show below.

```
> print(red_model3)

Call:
 randomForest(formula = quality ~ ., data = train, mtry = 3, importance = TRUE,
a.omit)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 0.3123345
                    % Var explained: 46.21
```

Figure 11: Summary of Random Forest model

The MSE, RMSE, MAE of the residuals is **0.3123345, 0.558869, 0.5192442** respectively.

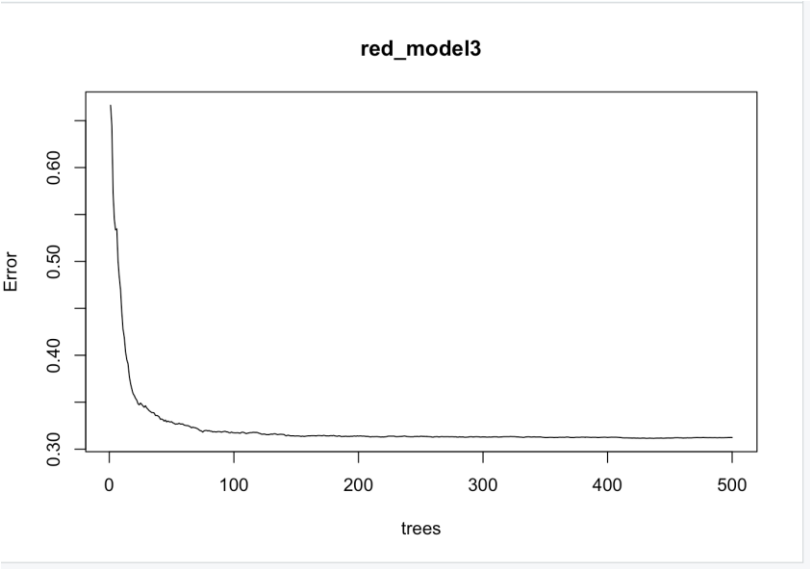

Figure 12: Random Forest model

Now I created a data frame with MSE, RMSE, MAE values of all three models to create a comparison dot plot as shown below.

Figure 13: Comparison of the models

From the above plots we can see that R-Squared value of random forest is more compared to other models. Likewise, RMSE, MSE values are lower for random forest compared to the other models. Hence, Random Forest is the best choice for accurate results. Although MAE is lower for Lasso regression and the second lowest is for the Random Forest. By comparing all the error values, I concluded that random forest is the optimum model.

CHALLENGES:

- The fundamental problem is that there are many predictors and variables, and it is difficult to determine which one is statistically significant enough to be a predictor.
- Most of the quality numbers were "regular" (5 and 6), which had no contribution in identifying the best model. We require additional balanced data to enhance our forecasting model.

CONCLUSION:

- Wines having a greater alcohol content should be highly rated in quality even though they may be less well-liked. Wines with less sweetness and less acidity are preferred in quality assessments.
- Independent variables with highest importance for red wine are alcohol, sulphates, volatile acidity, density and total Sulphur dioxide.
- By comparing R-Squared, RMSE, MSE, MAE values for all three regression models namely Linear regression, Random-forest regression, and Lasso regression I got to know that RMSE, MSE and MAE values are much lower for random-forest regression. Hence, Random Forest is the best choice for accurate results.

REFERENCES:

[1] UCI Machine Learning Repository: Wine Quality Data Set. (n.d.). UCI Machine Learning Repository: Wine Quality Data Set. Retrieved December 8, 2022, from https://archive.ics.uci.edu/ml/datasets/wine+quality

[2] Cortez, Almeida, & Matos. (2009, June 9). Modeling wine preferences by data mining from physicochemical properties. Modeling Wine Preferences by Data Mining From Physicochemical Properties - ScienceDirect. Retrieved December 7, 2022, from https://www.sciencedirect.com/science/article/pii/S0167923609001377