CSI 777 Project

Analysis Of Red Wine Data For Feature Selection And Classification

Submitted by

Ajay Kulkarni G01024139
Emmanuel Essiaw G00754904

Table of contents

Abstract

The main objective of the project was to identify the features which majorly affect the quality of the red wine and then by using those features classify the red wines. Red wine data consist of 12 variables and 1599 data points. Among those 12 variables 'quality' is the dependent variable. We started this project by cleaning the data which includes finding the empty values and removal of outliers. After cleaning we implemented Multiple Linear Regression and Random Forest for feature selection. These two algorithms helped us for build two models. The accuracy of the prediction of these models is then evaluated using Random Forest and Support Vector Machine. Further, we used these models to classify the red wines using three different methods C5.0, Support Vector Machine (SVM) and CART (Classification And Regression Trees).

Chapter 1: Introduction

Red Wine data has been studied in this project to select the best features of the red wine. After selecting the best features, we have used the selected features to classify the quality of the red wines. The red wine data has been taken from the UCI Machine Learning Repository. This data is related to the Portuguese "Vinho Verde" wine. The data consist of 12 variables and 1599 data points. The 12 variables which are present in data are given below along with their significance,

- Fixed acidity – Acids give the sourness or tartness "a fundamental feature in wine taste". Reduced acidity makes for a "flat" taste. Examples: tartaric, malic, citric and succinic acids. They are all found in grapes except for succinic.

- Volatile acidity – Too much volatile acidity is undesirable. The main acid of the issue is acetic acid. This acid is to be distilled from the wine, leaving only fixed acids.

- Citric acid – Most citric acid in the grapes is consumed during fermentation. It is sometimes added to wine to acidify it and give it "freshness". However, this can lead to microbial growth. Tartaric acid is sometimes used instead.

- Residual sugar – This is the sugar left in the wine from the grapes. These are the sweet wines. The Brix scale is used to track sugar development and determine when to harvest.

- Chlorides – The saltiness of wine. Chlorides are a result of the grapes used.

- Free sulfur dioxide – Not bound to any compound in the wine and also known as sulfites. It acts as an anti-microbial agent which limit the growth of harmful yeast/bacteria.

- Total sulfur dioxide – All sulfur dioxide in the wine, bound and free.

- Density – A comparison of the weight of a certain volume of wine to an equivalent volume of water.

- pH – It describes the acidity. If pH is lower, higher the acidity and it relates to the fixed acidity of the wine. There are "buffer acids" that do not contribute to acidity but help keep the pH level.

- Sulfates – Sulfates are mineral salts containing sulfur. They can be a byproduct of animal or plant decay as well as industrial processes. Sulfates may be connected with fermenting nutrition, which affects wine aroma.

- Alcohol – Sugars are converted to alcohol in the yeast fermentation. Too much alcohol compared to other components leads “hot” wine.

- Quality – Quality is the parameter which describes the quality of the wine in-between 1 to 10. The best wine has a quality rating as 8 and worst wine has a quality rating as 3.

From the above 12 explained variables in the data, quality is the dependent variable which is dependent on other variables.

To analyze the wine data, we started with the Exploratory data analysis and data preprocessing which is explained in chapter 2. After preprocessing we used that data for feature selection and detail explanation of feature selection is provided in chapter 3. Based on the selected features we built the models and test those models for predicting the quality of the red wines. Details about model building are explained in chapter 4. Further, we used those model to classify the red wines and results of classification are explained in chapter 5. The findings of this analysis are concluded in chapter 6 of the report.

Chapter 2: Exploratory Data Analysis and Data Preprocessing

Data preprocessing mainly deals with the checking the quality of the data and make the data ready for the further analysis. In data preprocessing we analyzed the data for missing values and outliers. In this chapter, different plots are presented to search the incomplete values in the data and to determine the quality of the data. The data which we obtained was in a .csv file and we have used MS-Excel initially to import the data and then we have used R to analyze the data. All the variables in the data are plotted using histograms and boxplots. Further, we also have calculated correlations using Pearson's correlation method to enhance the understanding of variables.

2.1 Histograms

The Histogram is a graphical representation of the distribution of numerical data. The purpose of the histogram is to graphically summarize the distribution of a data. We have used the histograms for studying the spread and skewness of the data. The histograms for all the variables are as follows:

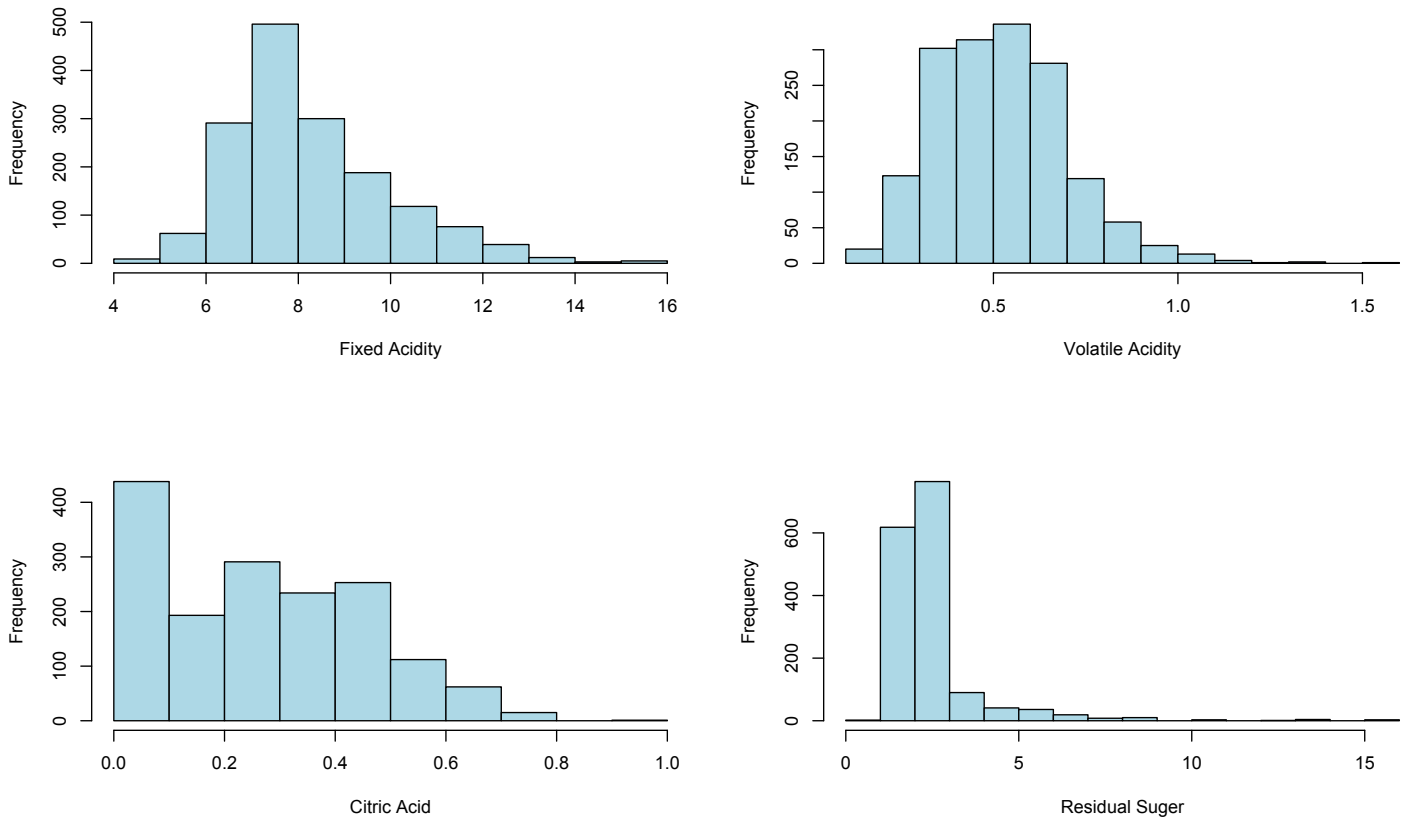


Figure 1: Histograms for Fixed acidity, Volatile acidity, Citric acid and Residual sugar

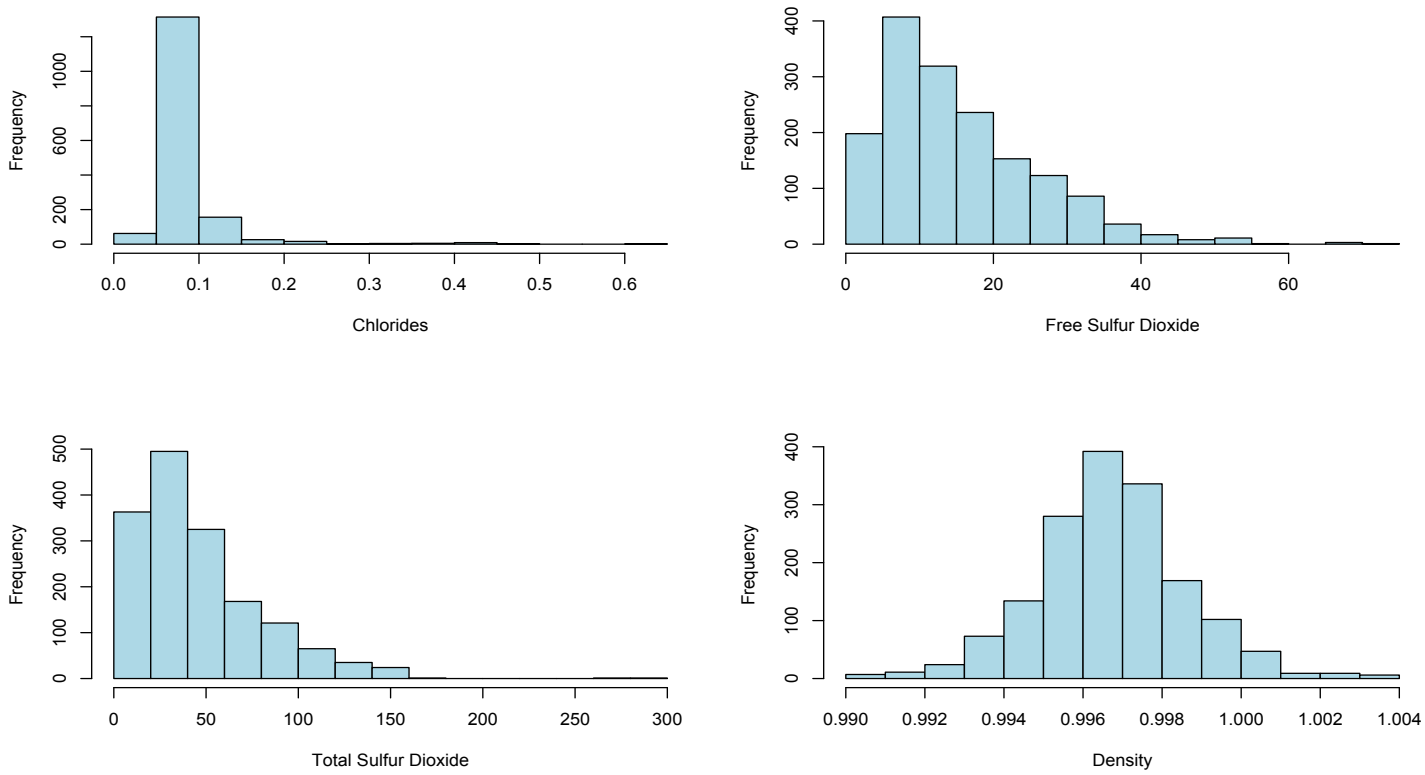
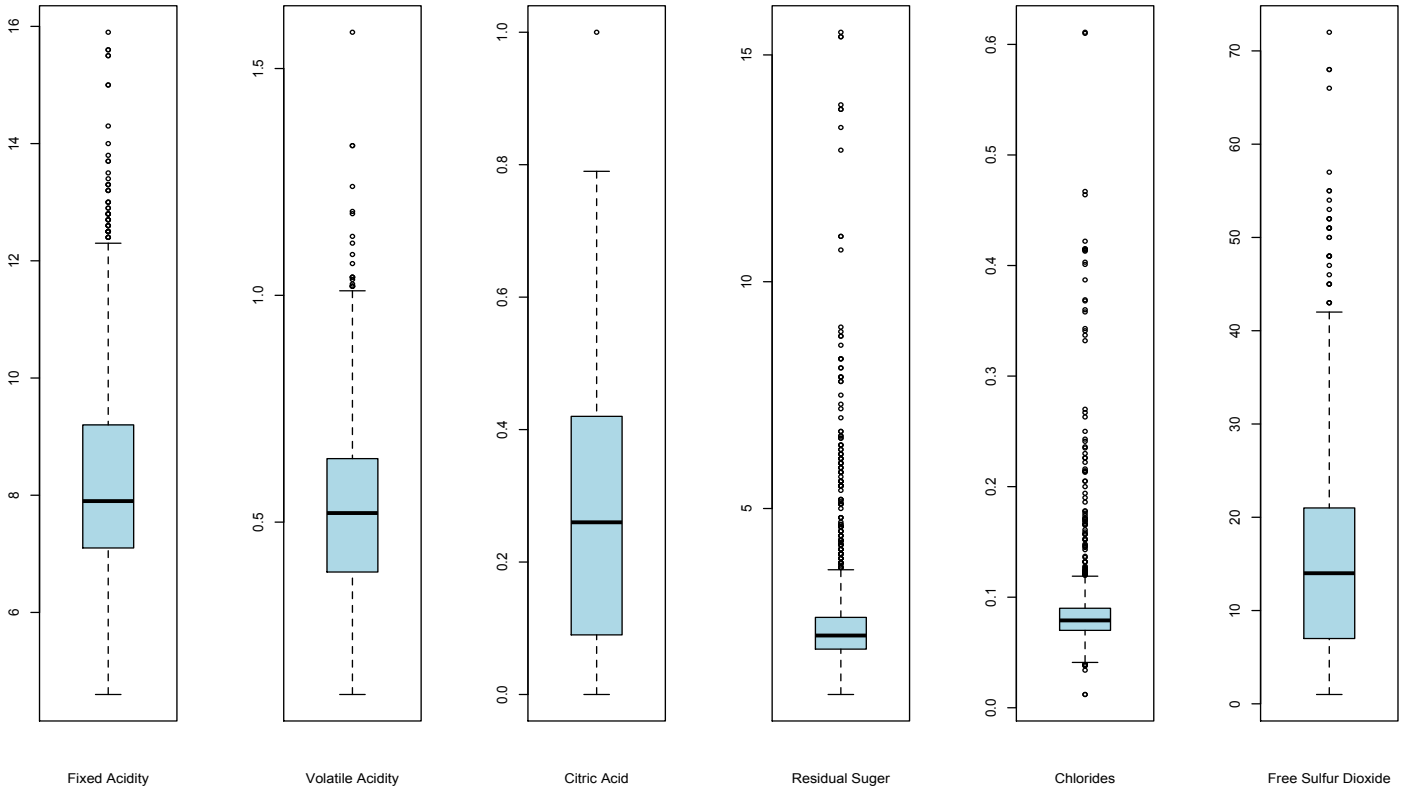


Figure 2: Histograms for Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide and Density

It can be observed from figure 1 and figure 2 that majority of the histograms are skewed right. The variables such as Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Sulphates, and Alcohol are right skewed. It is also observed from the histograms of variable pH and Density that, pH is slightly right skewed and Density is symmetric.

2.2 Box plots

Box plots are excellent tools for understanding the range of the data and detecting the outliers. The box plots also explain about the range, mean and median about the data. The main purpose of the box plots used here is to detect the outliers and to know about the data features like mean, median etc. Box plots for all the variables are given on the next page:



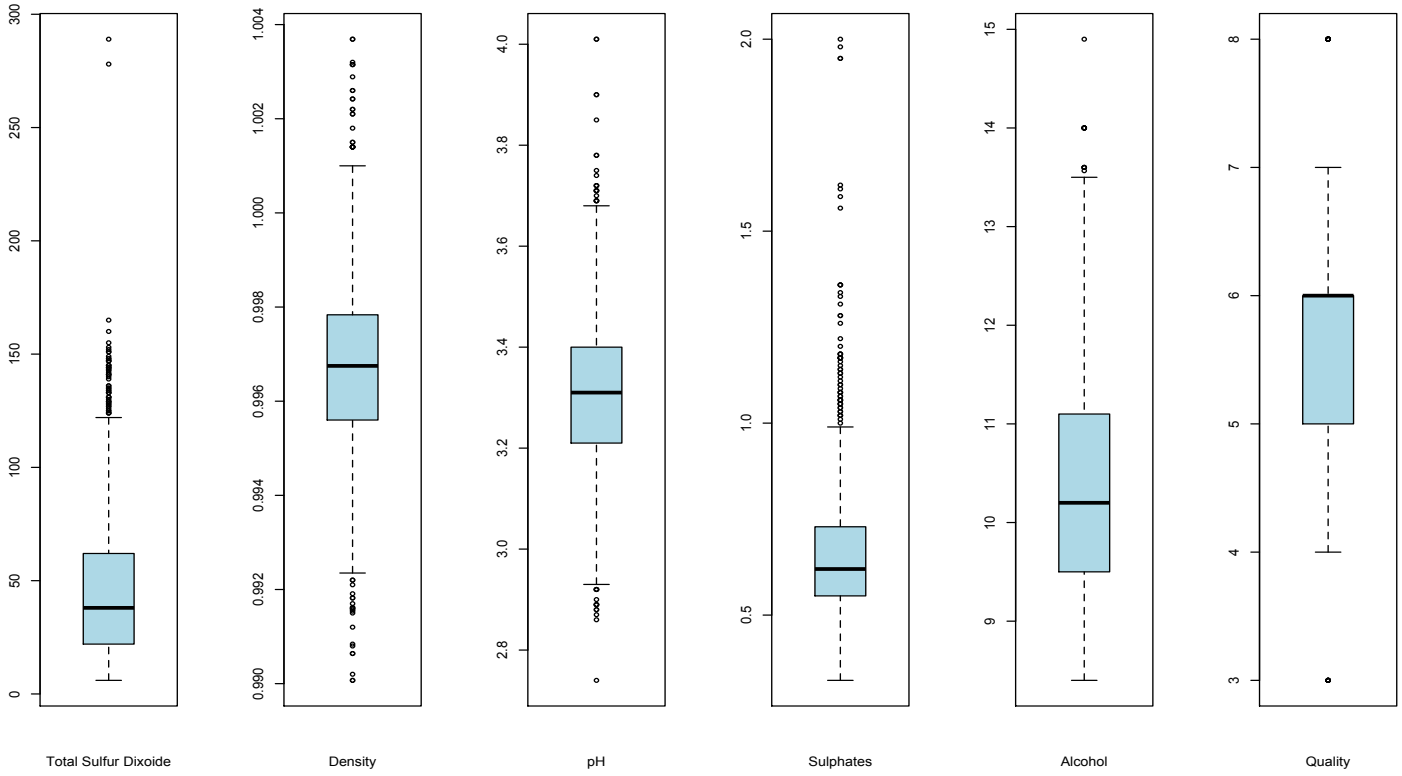


Figure 3: Box plots for all the variables

From the above plot, it is observed that more outliers are present in the variables Fixed acidity, Volatile acidity, Residual sugar, chlorides, Free sulfur dioxide, Total sulfur dioxide, pH, and Sulphates. On the other hand, variables like Citric acid and Alcohol have fewer outliers. It is also observed that for the majority of the variables outliers are present on the higher end of the plot. The summary of the data from the box plots can be given as follows:

Parameters	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Fixed Acidity	4.60	7.10	7.90	8.32	9.20	15.90
Volatile acidity	0.1200	0.3900	0.5200	0.5278	0.6400	1.5800
Citric Acid	0.0000	0.090	0.260	0.271	0.420	1.000
Residual Sugar	0.900	1.900	2.200	2.539	2.600	15.500
Chlorides	0.01200	0.07000	0.07900	0.08747	0.09000	0.61100
Free Sulfur Dioxide	1.00	7.00	14.00	15.87	21.00	72.00
Total Sulfur Dioxide	6.00	22.00	38.00	46.47	62.00	289.00
Density	0.9901	0.9956	0.9968	0.9967	0.9978	1.0037
pH	2.740	3.210	3.310	3.311	3.400	4.010
Sulphates	0.3300	0.5500	0.6200	0.6581	0.7300	2.000
Alcohol	8.40	9.50	10.20	10.42	11.10	14.90
Quality	3.000	5.000	6.00	5.636	6.000	8.000

Table 1: Summary of data with outliers

As it is observed from the box plots and summary of the data that it very important to remove the outliers from the data. It is also observed that majority of the outliers are present to the higher end of the plot. So we have only removed the outliers from the higher end of the values. The formula which we used for the removal of the outliers is as follows:

$$Max = Q_3 + 1.5 * IQR$$

The above method is called as Tukey's method and this method depends on the interquartile range of the data and 3rd quartile. So the values which are greater than Max are considered as outliers and are removed from the data. We also observed that there are no missing values present in the data and in this way prepared the data for the further analysis.

2.3 Pearson's correlation

The main purpose of the Pearson's correlation is getting the brief idea about the relationship of one variable with another variable. By using Pearson's correlation, we understand that which variables are more important for quality. The table for the correlations of all the variables is given below:

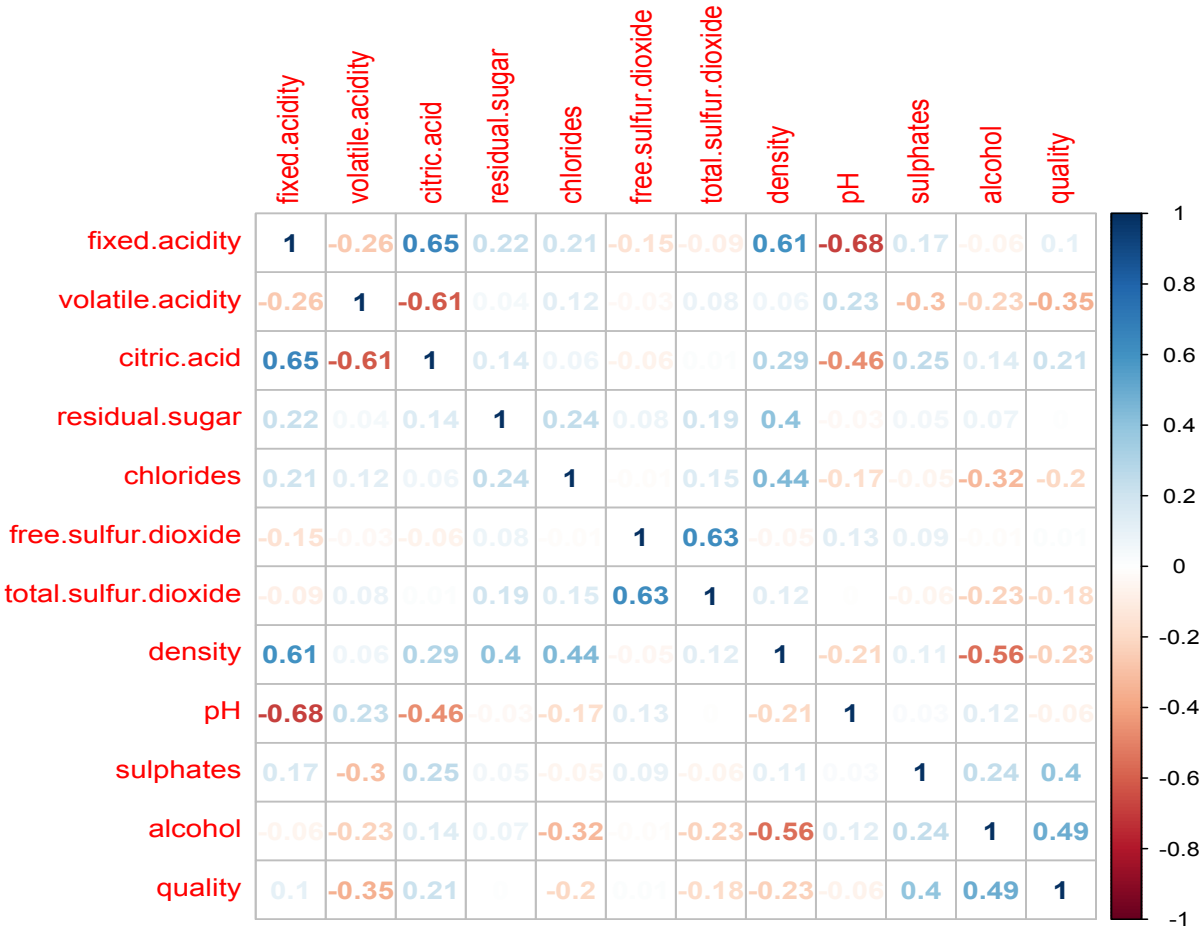


Table 2: Pearson's Correlation of all the variables

From the table 2, it can be observed that Volatile acidity (-0.35), Chlorides (-0.2), Total sulfur dioxide (-0.18) and density (-0.23) have a negative correlation with the quality of the wine. On the other hand, Sulphates (0.4), Alcohol (0.49), Fixed acidity (0.1) and Citric acid (0.21) have a positive correlation with the quality of the wine. There are some variables such as Residual sugar and Free sulfur dioxide which are not showing any relation with the quality of the wine. It is also observed that pH (-0.06) have a very less negative correlation with the quality.

Chapter 3: Feature Selection

Feature selection is the process of selecting different relevant parameters or variables which are most important to the data. Feature selection is also called as a variable selection or attribute selection. Feature selection techniques are used for the three reasons:

- Simplification of models to make them easier to interpret.
- Shorten training times.
- Enhanced generalization by reducing overfitting.

The main theme behind the feature selection is that there are many features present in the data. Some of them are irrelevant or redundant and if we remove those features it will not result in much loss of information. But if we select the most important features of the data then it can be used for building the good predictive model.

For the red wine, 11 features are present in the data and we need to select the features which majorly affect the quality of the red wine. The red wine data initially consist of 1599 data points but after removal of outliers, it reached to 1200. For feature selection, we have used two methods Multiple Linear Regression and Random Forest. The details about the feature selection are as follows:

3.1 Multiple Linear Regression

Multiple Linear Regression is implemented using an automated method known as stepwise forward selection. The method stepwise forward selection starts with no variables in the model and then testing the addition of each variable using a chosen model fit criterion. After adding the variable automatically, the model fit criterion is calculated and if the variable is giving the improvement in the fit then that variable is selected. So this process is repeated for all the variables and combinations of all the variables for selecting the best variables for the model.

Akaike Information Criterion (AIC) is used for the model comparison in stepwise forward regression. The AIC is a measure of the relative quality of statistical models for a given set of data. AIC estimates the quality of each model relative to each of the other model. So, AIC provides a means for model selection. The AIC value of the model can be computed as

$$AIC = 2K - 2 * \ln(L)$$

Where, L is the maximum likelihood function of the model and k is the number of estimated parameters in the model. The results obtained from the forward stepwise regression are given on the next page.

```
Start:  AIC=-512.32
quality ~ 1

                      Df Sum of Sq    RSS     AIC
+ alcohol              1   107.305 348.07 -736.05
+ sulphates            1    75.762 379.61 -663.18
+ volatile.acidity     1    56.099 399.27 -620.76
+ citric.acid          1    17.673 437.70 -543.57
+ chlorides            1    17.611 437.76 -543.46
+ total.sulfur.dioxide 1    15.783 439.59 -539.96
+ fixed.acidity        1     5.269 450.10 -520.10
<none>                             455.37 -512.32
+ pH                   1     0.958 454.41 -512.09
+ free.sulfur.dioxide  1     0.071 455.30 -510.46
+ residual.sugar       1     0.026 455.34 -510.37

Step:  AIC=-736.05
quality ~ alcohol

                      Df Sum of Sq    RSS     AIC
+ sulphates            1    39.232 308.83 -834.50
+ volatile.acidity     1    25.955 322.11 -799.15
+ fixed.acidity        1     8.605 339.46 -755.08
+ citric.acid          1     7.690 340.38 -752.81
+ pH                   1     6.119 341.95 -748.95
+ total.sulfur.dioxide 1     3.174 344.89 -741.74
<none>                             348.07 -736.05
+ residual.sugar       1     0.719 347.35 -735.79
+ chlorides            1     0.626 347.44 -735.56
+ free.sulfur.dioxide  1     0.000 348.07 -734.05

Step:  AIC=-834.5
quality ~ alcohol + sulphates

                      Df Sum of Sq    RSS     AIC
+ volatile.acidity     1   12.9869 295.85 -868.59
+ pH                   1    7.4250 301.41 -852.95
+ fixed.acidity        1    3.6315 305.20 -842.44
+ total.sulfur.dioxide 1    2.5133 306.32 -839.37
+ citric.acid          1    2.2487 306.58 -838.64
+ residual.sugar       1    0.8790 307.95 -834.90
<none>                             308.83 -834.50
+ chlorides            1    0.6733 308.16 -834.34
+ free.sulfur.dioxide  1    0.3293 308.50 -833.40


Step:  AIC=-868.59
quality ~ alcohol + sulphates + volatile.acidity

                      Df Sum of Sq    RSS     AIC
+ pH                   1   3.11900 292.73 -875.49
+ total.sulfur.dioxide 1   2.33740 293.51 -873.25
+ fixed.acidity        1   0.99585 294.85 -869.42
<none>                             295.85 -868.59
+ citric.acid          1   0.50495 295.34 -868.03
+ free.sulfur.dioxide  1   0.46543 295.38 -867.91
+ residual.sugar       1   0.41764 295.43 -867.78
+ chlorides            1   0.35461 295.49 -867.60

Step:  AIC=-875.49
quality ~ alcohol + sulphates + volatile.acidity + pH

                      Df Sum of Sq    RSS     AIC
+ citric.acid          1   2.62487 290.10 -881.06
+ total.sulfur.dioxide 1   2.28638 290.44 -880.08
+ chlorides            1   0.80504 291.92 -875.81
<none>                             292.73 -875.49
+ residual.sugar       1   0.47223 292.25 -874.85
+ free.sulfur.dioxide  1   0.24655 292.48 -874.20
+ fixed.acidity        1   0.05865 292.67 -873.66

Step:  AIC=-881.06
quality ~ alcohol + sulphates + volatile.acidity + pH + citric.acid

                      Df Sum of Sq    RSS     AIC
+ total.sulfur.dioxide 1   1.83846 288.26 -884.40
<none>                             290.10 -881.06
+ chlorides            1   0.57454 289.53 -880.73
+ fixed.acidity        1   0.42158 289.68 -880.28
+ free.sulfur.dioxide  1   0.37434 289.73 -880.15
+ residual.sugar       1   0.15599 289.95 -879.51

Step:  AIC=-884.4
quality ~ alcohol + sulphates + volatile.acidity + pH + citric.acid +
    total.sulfur.dioxide

                     Df Sum of Sq    RSS     AIC
<none>                           288.26 -884.40
+ chlorides           1   0.49961 287.76 -883.86
+ free.sulfur.dioxide 1   0.14402 288.12 -882.82
+ fixed.acidity       1   0.10235 288.16 -882.70
+ residual.sugar      1   0.00653 288.26 -882.42
```

Figure 4: Results from the forward stepwise regression

So after performing the analysis we got Alcohol, Sulphates, Volatile Acidity, pH, Citric acid and Total Sulfur Acid as the important features which can affect the quality of the red wine.

3.2 Random Forest

Random forests or Random Decision Forests are an ensemble learning method for classification, regression, and other tasks. Random Forests can be used for regression analysis and in fact called Regression Forests. They are an ensemble of different regression trees and are used for nonlinear multiple regression. Each leaf contains a distribution for the continues output variables. Random Forest package in R optionally produces two additional pieces of information: a measure of the importance of the predictor variables and a measure of the internal structure of the data. In this analysis, Variable importance of the Random Forest has been used for selecting important features of the data.

The Random Forest algorithm estimates the importance of a variable by looking at how much prediction error increases when data for that variable is permuted while all others are left unchanged. The necessary calculations are carried out tree by tree as the Random Forest is constructed. In the wine data, we used the variable importance function of the Random Forest and plot for the variable importance is given below

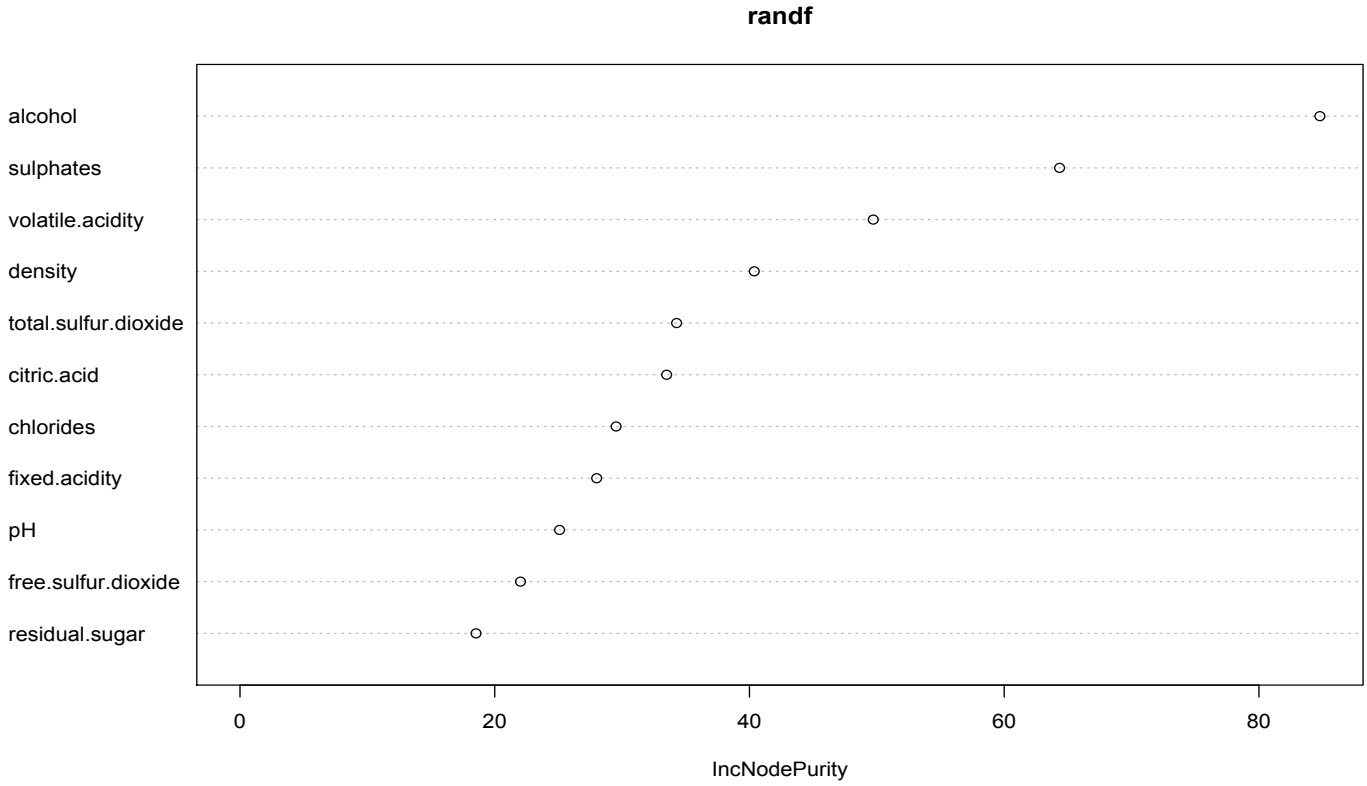


Figure 5: Variable importance plot using random forest

The importance of every variable can easily observe from the figure 5. It is observed that alcohol is the most important feature for the red wine and residual sugar is the least important feature for the red wine. From the above plot, we decided the criterion that we should select the variables whose importance is more than 40%. Thus, four variables with higher importance (Alcohol, Sulphates, Volatile Acidity and Density) are selected.

In this way by using Multiple Linear Regression and Random Forest we have selected the features. Multiple Linear regression method helped us to extract six important features and by using Random Forest we extracted four important features. In next chapter, we will be building two models based on these features and will select one of them based on the accuracy.

Chapter 4: Model Selection

Model building and selection of the model is completely dependent on the feature selection. We have selected the features from two methods and we have built two models based on the selected features. After building the models we have tested both the models using Random Forest and Support Vector Machine for predicting the wine quality. For analyzing the models, we divided the data into two parts training data (70%) and testing data (30%). The training data will be used by the model to train 70% of the data and remaining 30% data is used by the model for testing purpose i.e. predicting the quality of the wine. The details about model building and testing are given below.

4.1 Model Building

The six important features which we extracted from the Multiple Linear Regression are Alcohol, Sulphates, Volatile Acidity, pH, Citric acid and Total Sulfur Acid. So the first model which we built is given below along with its summary.

Model 1

quality ~ alcohol + sulphates + volatile.acidty + pH + citric.acid + total.sulfur.acid

```
Call:
lm(formula = quality ~ alcohol + sulphates + volatile.acidity +
    pH + citric.acid + total.sulfur.dioxide, data = t1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96088 -0.36139 -0.07625  0.44200  1.58761

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          4.4221297  0.6049650   7.310 6.28e-13 ***
alcohol              0.2857449  0.0225432  12.675  < 2e-16 ***
sulphates            1.7093110  0.1861059   9.185  < 2e-16 ***
volatile.acidity    -0.8912703  0.1605340  -5.552 3.80e-08 ***
pH                  -0.6621243  0.1758050  -3.766 0.000177 ***
citric.acid         -0.3975988  0.1585231  -2.508 0.012326 *
total.sulfur.dioxide -0.0017988  0.0007804  -2.305 0.021417 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5883 on 833 degrees of freedom
Multiple R-squared:  0.367,   Adjusted R-squared:  0.3624
F-statistic: 80.48 on 6 and 833 DF,  p-value: < 2.2e-16
```

Figure 6: Summary of model 1

Similarly, we extracted four important features Alcohol, Sulphates, Volatile Acidity and Density from the Random Forest method. So the second model we built can be given as,

Model 2

quality ~ alcohol + sulphates + volatile.acidty + density

4.2 Model Testing

4.2.1 Random Forest

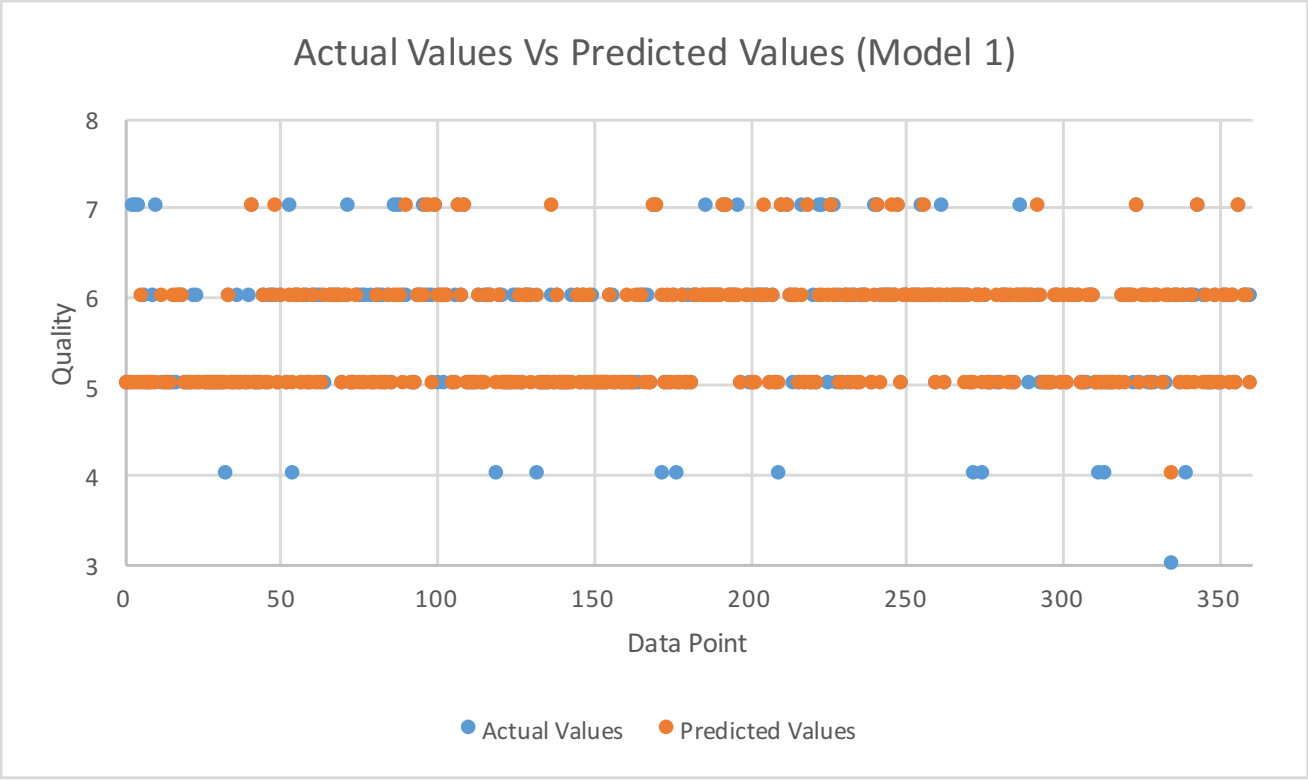


Figure 7: Random Forest results for Model 1

The result of the Random Forest method for model 1 is given in figure 7. In the above plot, blue dots indicate the actual values and orange dots indicate predicted values of the quality of the red wine. It can be observed from the plot that Model 1 is predicting most of the red wines which having quality ratings 5 and 6. It can also be observed that Model 1 is not predicting correctly the lower quality red wines having ratings 3 and 4 as well as higher quality wines having rating 7. Model 1 correctly predicts 246 points out of 360. So the percentage accuracy of the Model 1 can be given as 68.33%.

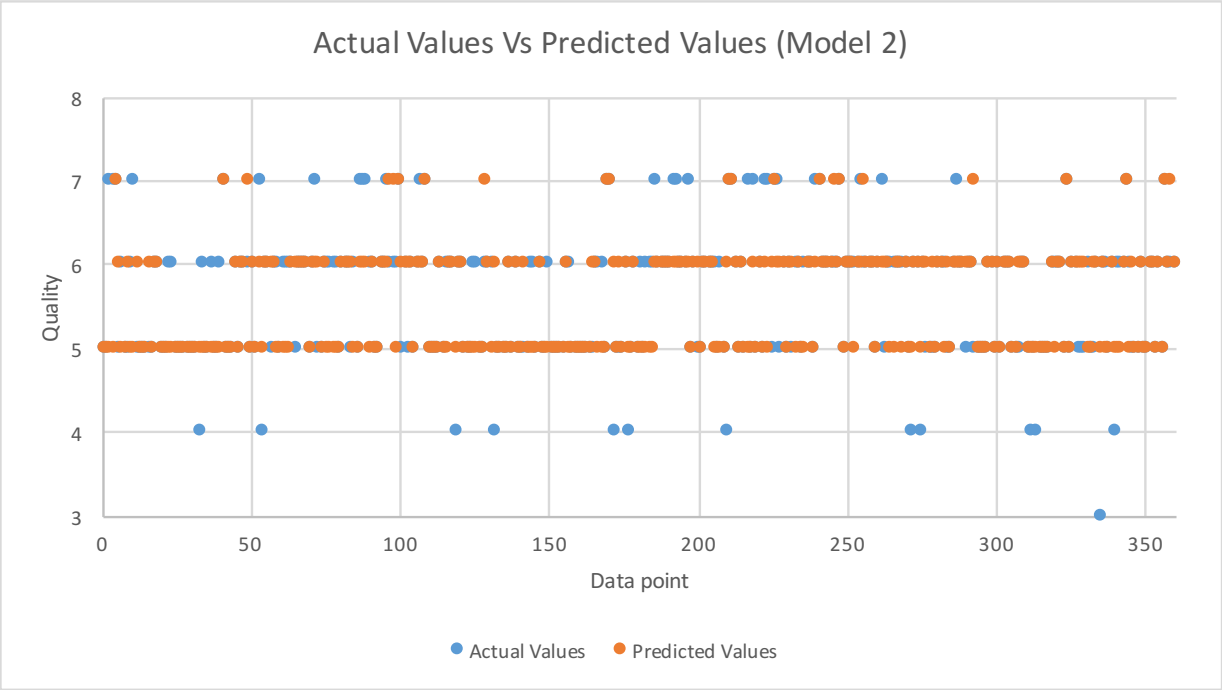


Figure 8: Random Forest results for Model 2

The result of the Random Forest method for model 2 is given in figure 8. From the observation of figure 8, we can say that the Model 2 is also not good in predicting the higher and lower wine quality ratings. Model 2 is correctly predicting 241 points out of 360. So the percentage accuracy of the model can be given as 66.94%.

4.2.2 Support Vector Machine

Support Vector Machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. In this project, we have used the SVM for selecting the best model among the two which can be used for better prediction of the quality of the wine. We have implemented SVM on both the models and results of both the models are given on next page.

Figure 9 indicates the result of Model 1 and figure 10 indicates the result of Model 2 using SVM. In both the plots, blue dots indicate the actual values and orange dots indicates the predicted values. It is observed from the figure 9 and figure 10 that both the models are good for predicting quality ratings 5 and 6 but they are not good for predicting the lower quality ratings (3 and 4) and higher quality rating (7). Model 1 is predicting 235 points correctly out of 360 data points. So the percentage accuracy for the Model 1 is about 65.27%. Similarly, Model 2 is predicting 229 points correctly among 360 data points. The percentage accuracy of the model 2 can be given as 63.61%.

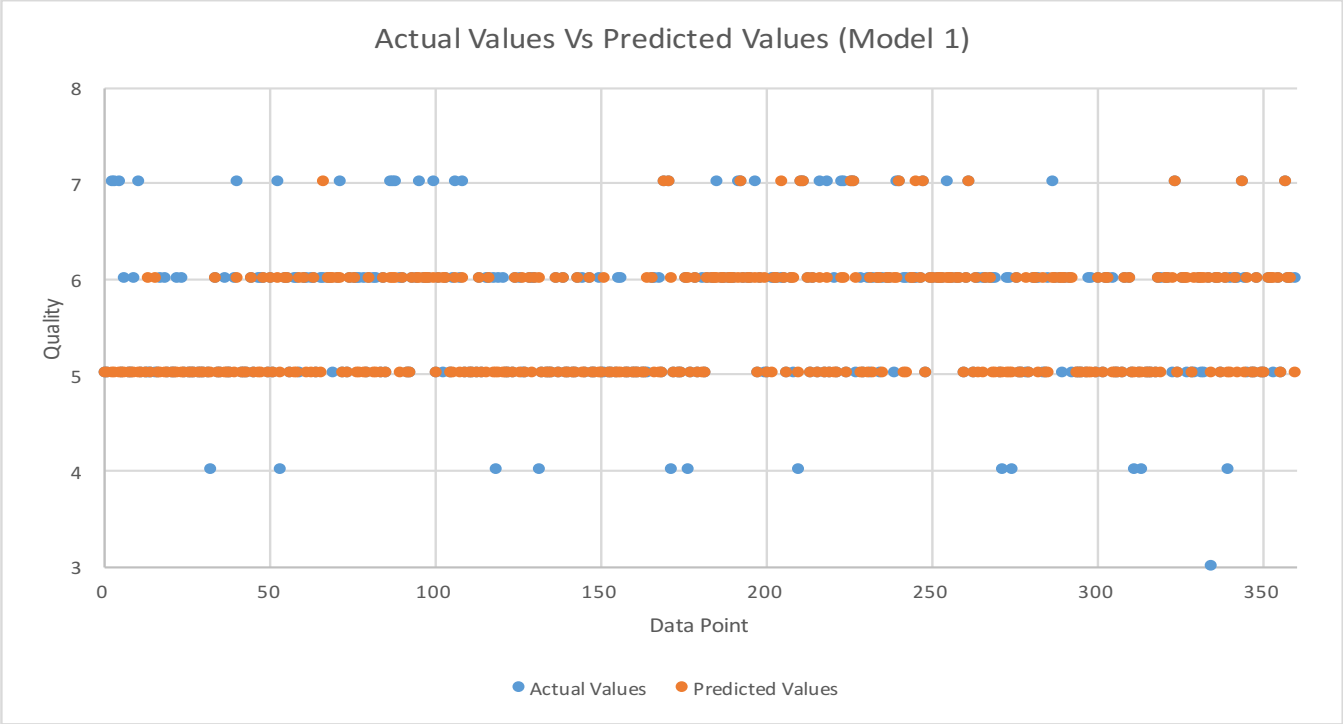


Figure 9: SVM results for Model 1

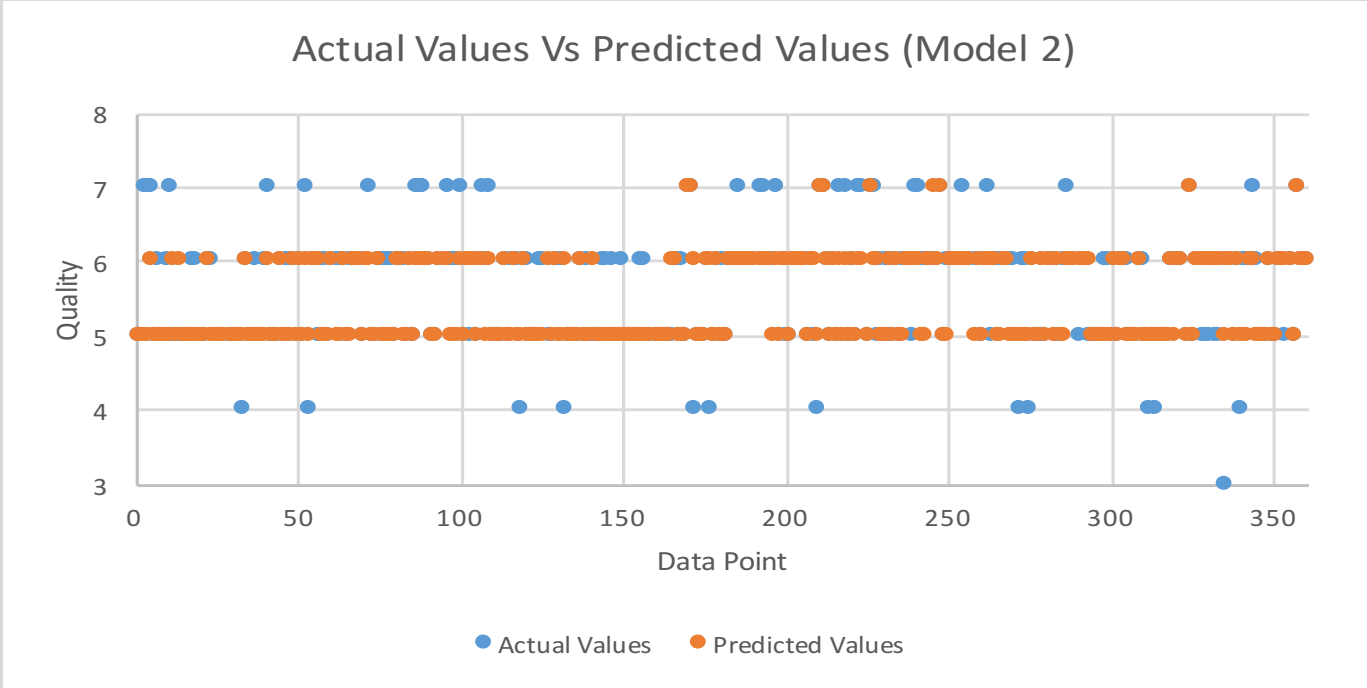


Figure 10: SVM results for Model 2

	Model 1	Model 2
Random Forest	**68.33%**	66.94%
SVM	**65.27%**	63.61%

Table 3: Summary of accuracy of the models

The summarized table of accuracy for both the models is given above. For both the methods, Model 1 is giving more accuracy than Model 2. So we selected Model 1 for the further analysis. Also from the above analysis, we can say that the selected six features (Alcohol, Sulphates, Volatile Acidity, pH, Citric acid and Total Sulfur Acid) are very important for predicting the quality of the red wine.

Chapter 5: Classification

Classification is the problem of identifying to which of a set of categories a new observation belongs on the basis of a training set of data whose category membership is known. The goal of the classification is to accurately predict the target class for each case in the data. To implement classification, first we need to convert a numerical variable into a categorical variable because 'quality' is numerical variable. So we divided qualities into two categories. The wines which having quality less than or equal to 5 are labeled as '0' and the wines which having quality greater than 5 are labeled as '1'. Different classification methods have been implemented on the selected model as well as on other models for verification. The number of wines in the testing data are 360. Class 0 contains 159 wines and class 1 contains 201 wines. Details about classification are as follows,

5.1 C5.0

The C5.0 method is a further extension of C4.5 and pinnacle of that line of methods. The C4.5 is an algorithm used to generate a decision tree and it is an extension of the ID3 algorithm. The C5.0 offers a number of improvement as compared to C4.5 in terms of speed, memory usage etc. We have used the C5.0 algorithm to classify the quality of the wines. The results of implementation of C5.0 for models are given below,

Model 1		
	0	1
0	134	73
1	25	128

Table 4: Confusion matrix for Model 1

Table 4 indicates the results for the model 1, it gives a correct prediction of 134 wines in class 0 and for 128 wines in class 1. So total accuracy of the model 1 can be given as 72.77%. We further implemented C5.0 algorithm to model 2 as well as on another model containing all the variables.

Model 2		
	0	1
0	118	49
1	41	152

Table 5: Confusion matrix for Model 2

Table 5 indicates the results for the model 2, it gives a correct prediction of 118 wines in class 0 and for 152 wines in class 1. So total accuracy of the model 2 can be given as 75.00%. For

checking the results of classification we also have implemented the C5.0 algorithm by considering all the variables in the model. The results of the classification are given below,

Model containing all variables C5.0(quality~.)		
	0	1
0	125	51
1	34	150

Table 6: Confusion matrix for Model containing all variables

This model is able to correctly classify 125 wines in category 0 and 150 wines in category 1. The accuracy of this model is given as 76.38%.

5.2 Support Vector Machine Classification

Model 1		
	0	1
0	128	59
1	31	142

Model 2		
	0	1
0	132	60
1	27	141

Model containing all variables SVM(quality~.)		
	0	1
0	132	56
1	27	145

Table 7: Confusion matrix using SVM for all the models

The results of classification using SVM for all the models are given above. Table 7 indicates that model 1 is classifying 128 wines to category 0 and 142 wines to category 1. Model 2 predicts 132 wines to category 0 and 141 wines to category 1. A model containing all the variables predicts 132 wines to category 0 and 145 wines to category 1. So the accuracy of the model 1 can be given as 75.00%, for model 2 it is 75.83% and accuracy for the model containing all the variables is calculated as 76.94%.

5.3 CART

CART stands for the Classification And Regression Trees algorithm that is used for classification and regression predictive model problems. We have implemented CART for all the models and results in terms of confusion matrix are given below,

Model 1		
	0	1
0	111	42
1	48	159

Model 2		
	0	1
0	111	42
1	48	159

Model containing all variables rpart(quality~.)		
	0	1
0	111	42
1	48	159

Table 8: Confusion matrix using CART method for all the models

The results of classification using CART for all the models are given above. Table 8 indicates that all the models are classifying 111 wines to category 0 and 159 wines to category 1. So accuracy is same for all the models which is 75.00%.

Chapter 6: Conclusion

Red wine data has been successfully analyzed for finding the important features which affect the quality of the red wine. We started feature selection using Multiple Linear Regression and Random Forest. Multiple Linear Regression helped us to extract six important features (Alcohol, Sulphates, Volatile Acidity, pH, Citric acid and Total Sulfur Acid) from the data and Random Forest helped us to extract four important features (Alcohol, Sulphates, Volatile Acidity and Density) from the data. After selecting these features two different models are built. The prediction accuracy for both the models are calculated using Random Forest and Support Vector Machine (SVM) and models are compared. In the case of Random Forest, the accuracy of model 1 is calculated as 68.33% and for model 2 the accuracy is about 66.94%. For SVM, the accuracy of the model 1 is 65.27% and for model 2 the accuracy is about 63.63%. So from the above analysis, we selected the first model containing six important features which are Alcohol, Sulphates, Volatile Acidity, pH, Citric acid and Total Sulfur Acid.

We have used the model 1 for classification and then compared the results of model 1 with model 2 containing four important features as well as with the model containing all the variables. For classifications, we have used three methods which are C5.0, Support Vector Machine, and CART. The results from all the three methods were not convincing and it became ambiguous to select the best model for the classification of the quality of the red wines.

Thus, we concluded that the quality of the red wine is majorly dependent on Alcohol, Sulphates, Volatile Acidity, pH, Citric acid and Total Sulfur Acid, but for classifying purpose these features are not sufficient.

References

1. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
   Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

2. https://prezi.com/s7kkx534rsns/wine-quality-analysis/

3. https://www.r-bloggers.com/predicting-wine-quality-using-random-forests/

4. https://www.r-bloggers.com/correlation-and-linear-regression/

5. http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/a-comparison-of-the-pearson-and-spearman-correlation-methods/

6. https://onlinecourses.science.psu.edu/stat857/node/223

7. http://scott.fortmann-roe.com/docs/MeasuringError.html

8. https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/

9. https://methodology.psu.edu/node/504

10. http://stats.stackexchange.com/questions/69452/how-does-stepwise-regression-work

11. http://www.svm-tutorial.com/2014/10/support-vector-regression-r/

12. https://discuss.analyticsvidhya.com/t/how-to-extract-important-variables-from-random-forest-model-using-varimpplot-in-r/1325/2

13. http://mlampros.github.io/2016/02/14/feature-selection/

14. http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html

15. https://en.wikipedia.org/wiki/C4.5_algorithm

16. https://en.wikipedia.org/wiki/Support_vector_machine

17. http://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/