PREDICTION AND ANALYSIS OF WINE QUALITY

Introduction

Wines have started seeing an increase in consumption from the mid-1990s[1]. The primary market for wine consumption and production was Europe. Now, the rise in wine demand can be seen in mid-African and Asian countries. More than 10000 wineries alone are in the US in 2020, which produces approximately 10% of the global wine volume[2]. With more than 270 million hectoliters of the production volume, wines are produced in more than 60 countries. Italy, Spain, the USA, Argentina, and China are among the five largest wine producers. France, Italy, Portugal, Switzerland, and Luxembourg are among the world's higher wine consumers. However, the USA and China are the largest Wine aggregators because of the increasing number of wine consumers.

Wines are generally associated with various cultures in society. It can be served as an aperitif, in a three-course meal, or just socializing with friends. The drink's color varies from dark red to sparkling whites, sometimes with a pinkish blush tone of rose varieties. Winegrowing and production is a long-term business and requires considerable capital. Thus, it is crucial to understand how the different types of wines produced are preferred by the customer to increase sales and profits of the businesses.

Therefore, the project's objective is to build a classification model that can classify wines into different segments based on quality rating. Also, to identify the relationship between various properties and components of wine, which leads to a better quality wine.

[1] https://www.yourwineestate.com/yourwineestate/main# !global-wine-market
[2] https://www.statista.com/topics/1541/wine-market/

To perform the analysis, we used the dataset from UCI Machine Learning[3]. The most crucial factor is to find the perfect combination of wine properties that results in highly rated wines. The higher the quality rated, the better is the wine. The scale of the quality is from 0 to 10. To transform it into a classification problem, we have assumed that the quality of wine greater than 7 is 'Good Quality' and otherwise 'Bad Quality.' To predict the quality, we have developed a classification model, which indicates $\sim 87\%$ accuracy. In the report, we have also analyzed a different set of variables in Red wine and White wine based on the feature selection methodology from random forest and principal component techniques and the relationship of these variables with quality in respective wines.

The following section discusses the predictors/variables, the classification model selection criteria, PCA analysis, and the model results.

Data Description

The data consists of approx. six thousand observations with 13 variables, out of which two are categorical - Type and Quality; and the remaining eleven are continuous variables - Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulfates and Alcohol. The target variable 'Quality' is classified into Good and Bad by clustering the wines from one to six on the quality scale as Bad and seven and above as Good (See Figure 12(b)).

Let's examine the distribution of each of the numeric variables in the dataset.

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

1) Fixed Acidity - The histogram exhibits a positive skew with a mean of around 7.21 and a standard deviation of roughly 1.29, having a fixed acidity between 6 to 8 for most of the wines. Further analyzing the distribution via boxplot, it became evident that the skewness was majorly due to red wines, in which there were many outliers above the mean. (Refer Figures 1a and 1b).

2) Volatile Acidity - This variable is highly positively skewed, which illustrates that a large number of wines have a volatile acidity less than 0.5. This depicts that the average number of wines consists of 0.33 volatile acidity and 50% of the wines have volatile acidity below 0.29. The high positive skewness was majorly contributed by white wine as it has a large number of outliers above the mean compared to red wines, and there were no outliers below for both the types of wine, which clearly explains the skewness. (Refer Figures 2a and 2b).

3) Citric Acid - This variable is positively skewed, with many wines containing citric acid between 0.1 to 0.5. 50% of the wines have citric acid less than 0.31, and the average number of wines have a citric acid of 0.3187. The boxplot confirms this by showing many outliers above the mean in the case of white wine and a small number of outliers below. In contrast, the wine did not contribute to citric acid's positive skewness as there was only one outlier above the mean and no outlier below. (Refer to Figure 3a and 3b).

4) Residual Sugar - This variable has a huge concentration between 2 to 12 and a very long positive tail. The boxplot confirms this as there were many outliers above the mean for the red wines compared to white wines exhibiting a tight concentration around

low mean and large positive outliers. 50% of the wines have residual sugar concen- tration below 3, which is less than the average number of wines with a residual sugar concentration of 5.44. (Refer to Figure 4a and 4b).

5) Chlorides - Similar to the residual sugar, the chlorides also are very highly positively skewed with a long positive tail depicting that a large number of wines have the chlo- rides concentration in the range of 0.01 to 0.1. This means that average wines have chlorides of only 0.05, and 75% of the wines have chlorides below 0.065. This was clearly illustrated in the box plot. The maximum number of outliers was above the mean for red wine compared to white wines and only two outliers below the mean, which concludes that red wine was the major contributor to the positive skewness of chlorides. (Refer to Figure 5a and 5b).

6) Free Sulfur Dioxide - The free sulfur dioxide has a large concentration between 1 to 50 with a long positive tail, where 50% of the wines had a concentration less than 29 and average wines had a concentration of 30.53. The major contributor to high positive skewness was white wine due to many outliers above the mean with no outlier below. (Refer to Figure 6a and 6b)

7) Total Sulfur Dioxide - The majority of wines have total sulfur dioxide concentration in the range of 90 to 200, where average wines contain a concentration of 115.7 and 50% of the wines have a concentration less than 118. Since the distribution isslightly positive, there were some outliers in white wine and red wine above the mean and negligible outliers below the mean. (Refer to Figure 7a and 7b)

8) Density - It was observed that density is highly positively skewed but does not have a long positive tail, and due to this, there are very few outliers in both red and white wines above the mean; the only difference is that red wines have a few outliers below the mean as well which are not present in white wines. This indicates that a large number of wines have a density between 0.99 to 1. (Refer Figure 8a and 8b)

9) The pH - pH level exhibits an approximately normal distribution with a mean of around 3.2 and a standard deviation of about 0.16, where most wines have a pH concentration between 3.0 to 3.4. Further analyzing the boxplots reveals that the mean is about 3.2 and the normality of pH, which also holds since there are outliers both above and below the mean in both red and white wines. The white wines have more outliers above the mean than red wine, which confirms that it is the major contributor to slight positive skew in pH. (Refer to Figure 9a and 9b)

10) Sulfates - The sulfates are also slightly right-skewed with a long positive tail, and the maximum concentration is between 0.3 to 0.7 for most wines. The red wine majorly contributes to sulfates' skewness due to many outliers above the mean compared to white wines, and 50% of the wines have a concentration of sulfates less than 0.51. (Refer to Figure 10a and 10b)

11) Alcohol - The distribution of alcohol is not very telling. The boxplot can be considered an approximately normal distribution as there were no outliers in the case of white wines and only three outliers in the red wine above the mean with no outliers below the mean. Therefore, we can say that 50% of the wines have 10.30 alcohol content, and the maximum concentration ranges between 9 to 12 for most wines. (Refer to Figure

11a and 11b)

We constructed the correlation matrix for each of the numerical variables to ensure that the predictors are uniquely determined and not influenced by each other (See Figure 13). No two variables had the correlation coefficient higher than $\sim$ 0.8-0.85 threshold; therefore, no collinear relationship between the variables was observed.

Model Selection

Feature Selection Process

To build our model, we first classified the categorical variables such as type and quality into factors using as.factor() function in R studio. This process re-codes categorical variables into a series of binary dummy variables, which can then be input into the classification models.

Although outliers were depicted in many predictors, we chose not to remove them from any variables because estimates of predictive power are more accurate by retaining outliers. Removing the outlier would merely ignore the data that the model does not fit well and reduce the model's ability to predict accurately in a test set.

We then performed a random forest and principal component analysis to select the best features for the prediction of wine quality. In the model, the target variable is the quality of wines and the remaining 12 variables were considered as predictors.

The random forest methodology suggested that the most important features contributing to the quality of red and white wines are alcohol, volatile acidity, residual sugar and sulfates,

pH, and free sulfur dioxide. This is because if we remove these variables, then on average, the model's accuracy will decrease by 101,71.89,68.69,72.76,67.28 and 70.74 respectively, and these variables also decrease the Gini node impurity by 330.15, 179.99,178.81,167.35,162.45 and 159.43 respectively.

We also performed the feature selection process using the principal component methodology. Computing the eigenvalues (See Table 2) for loading in the first six PCs, it was found that the three predictors do not account greatly in determining the quality of the wine. The three predictors are pH, Free Sulfur dioxide, and Fixed Acidity. The total sulfur and free sulfur are collinear (See Figure 15); therefore, we have included Total Sulfur in our analysis. Hence, the classification model was built using other predictors that showed correlation with the 'Quality' variable.

Classification Modelling

The model was built using the important features selected from random forest and PCA techniques to build our classification models. This is to understand how only including the most important variables may impact wines' quality and how accurate the predictions could be. We built and tested the accuracy of four classification models- Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.

Modeling on the Features selected by Random Forest and Principal Component Analysis.

To perform the model testing using the important features selected from the random forest and PCA, we performed cross-validation where we split the data into training and testing in 50:50 ratio and applied predict() and measured the accuracy of our predictions versus the test data.

We built a logistic regression model using glm() function and family = 'binomial,' and R took five iterations to find the maximum likelihood function. We also considered that if the probability of wine quality is >0.5, then it is predicted as Good Otherwise Bad, and after this, we measure the accuracy of our predictions. We calculated the R squared value using the training dataset, which came out to be 0.264. This indicates that the model explains 26% of the variability of the training data around its mean, which illustrates that the data points are less fitted to the regression line since the variance is less. Using the cross-validation test, we checked the accuracy of the model, which came out to be $\sim$ 82% with an error rate of
$\sim$ 17% with the important features selected using the random forest and $\sim$ 82% accuracy with an error rate of $\sim$ 18% were observed when tested with the most important features selected by PCA technique.

We then measured the accuracy of our predictions using a decision tree model. We first found the optimal cp value by testing the model with cp values of 0.01 for a normal tree with few branches and 0.0003 for an over fitted tree with multiple branches to achieve the

optimal tree. Through multiple testing, we found that the optimal cp value was 0.0073[4], which generated the best decision tree. Using cp value = 0.0073, we performed the cross- validation test to find the accuracy score of ∼ 82.6% with an error rate of ∼ 17.4% with features selected using both random and PCA techniques.

After this, we tested predictions using a random forest model. We have used optimal number of trees as 500 which would provide us with the best features and the maximum accuracy score and minimum error rate. The model generated an accuracy of ∼ 87% with only ∼ 12% error rate for the features selected by random forest technique and PCA techniques. It was also observed that the error rate generated by the cross-validation test and out of bag error were very close, and it was ∼ 12% and ∼ 13% , respectively.

To build the optimal boosting model, we used Bernoulli distribution since the dependent variable is categorical. After testing with multiple depths and number of trees, we found that the optimal interaction depth and ntree to be 3 and 10,000, respectively. Using these optimal hyperparameters, we performed a cross-validation test. We achieved an accuracy score of ∼ 84% with an error rate of ∼ 15.5% from features selected by random forest and
∼ 85% accuracy score with an error rate of ∼ 15% from the features selected using PCA technique.

On comparing the accuracy scores, we observed that Random Forest was the best model in predicting the quality of red and white wines in both the situations where it used the features selected from random forest technique and PCA technique. It generated the maximum accuracy of ∼ 87% with only ∼ 12% of the error rate. The accuracy and error rate of

[4] This value may change on multiple runs.

respective models are given in Table 3.

The results clearly illustrate that the most important wine attributes are alcohol concentration, volatile acidity, residual sugar, sulfates, pH, and free sulfur dioxide. So, these variables should be taken into consideration while preparing wines to further improve theirquality.

Principal Component Analysis

To understand the ideal mix of the components that make the wine quality better, we did PCA analysis on the dataset. Using the prcomp() function, we generated the value of loadings and the variation explained by each PCA component.

Initially, the PCA analysis was done on the entire dataset, and the good quality observations were marked with 'pink' color and bad quality with 'grey' color. However, there were issues to analyze the PCA components and the issues were two-fold.

First, there were a lot of good and bad quality observations that overlapped. (See Figure 17). This issue was resolved when we divided all the observations into Red and White wine. It was evident that there were two clusters of Red Wine and White Wine. This called for a separate analysis of Red Wine and White Wine. Second, the variation accounted by two components in the observations was $\sim 50\%$. Therefore, it was required to include one more component to account for the variables' maximum variation. (See Figure 16)

Furthermore, we used princomp() and plot3d() functions to plot observations and loadings of PCA components in 3-Dimensional for Red Wine and White Wine, respectively. (See Figure 19)

Managerial Implication

According to the recent figures, the wine market is expected to grow to 281 million cases worth $ 32.9 billion by 2022, with a CAGR of approx. 3% [5]. The real challenge of the wine producers is to come up with the perfect wine formula so that the manufacturers can compensate for the large capital cost incurred for setting up the production line of the Wines. Thus, it is important to analyze which components must have and help increase the quality of wines.

One common practice to measure the quality of the wine is to invite random people for testing on the vineyards. This at least requires initial sampling cost and production cost of the samples. However, we can leverage the model's power to identify the quality rating of wine with the properties before the production starts.

Figure 18 represents the plot of two PCA components having quality 'Good' and 'Bad'. It is difficult to interpret how different components affect the quality of a wine because the different types of wine i.e., Red Wine and White Wine, overlap. We decided to conduct a separate analysis for each type of wine. First, the properties that vary in Red and White Wine are analyzed using a 3-D PCA plot, in which we accounted for more than 60% of the variation in the observations.

Referring to Figure 18, it can be seen that the below components have more than average concentration in Red Wine than they are in White Wine:

1) Fixed Acidity

[5] https://www.toptal.com/finance/market-sizing/wine-industry

2) Sulfates

3)Chlorides

4)Volatile Acidity

The following components have less than average concentration in Red Wine than they are in White Wine:

1)Total and Free Sulfur Dioxide

2)Residual Sugar

There were few components whose concentration was equally distributed in Red Wine and White Wine because the following component direction was in the third direction. These components were:

1)Alcohol Content

2)Citric Acid 3)Density

4)pH Level

The points that were close to each other have similar characteristics, thus the observations in pink represented wine with good quality and ones in grey color bad quality wine.

Let's now discuss each type of wine and the components that lead to a better wine quality.

1) Red Wine:

a) The quality of the Red Wine is generally on the bad side when the all component of the wine has average concentration.

b) The properties that are positively correlated with the quality of Red Wine are: Alcohol and Sulfate. That means if the Alcohol and Sulfate level increases the quality of the wine also improves.

c) The properties that are negatively correlated with the quality of Red Wine are: Residual Sugar, Volatile Acidity, Chlorides, and Density. Quality decreases with the increase in the value of these variables.

d) Few properties do not affect the quality of Red Wine to a greater extent. These properties were pH, Fixed Acidity, Citric Acid, Free Sulfur, and Total Sulfur levels.

Note: These observations are used PCA (used 3 principal components and ac- counted $\sim$ 60% variance) for Red Wine observations only. (See Figure 21)

2) White Wine:

a) Similar to Red Wine, If the content of the properties of the White wine arekept on average, then the wine comes out to be of bad quality.

b) If we increase the content of Free and Total Sulfur Dioxide, Sulfates, and Alcohol, the quality of White wine will increase.

c) However, the increase in the properties such as Density, Chlorides, Volatile Acid- ity, and Residual Sugar will decrease White wine quality. Thus these variables

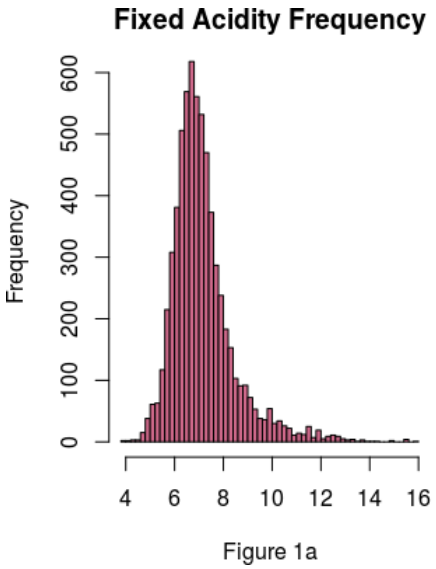
are negatively correlated with the quality of White wine.

d) The properties which influence the quality of White Wine the least were Fixed Acidity, pH, and Citric Acid levels. The direction of these loadings was orthogonal to quality loading.

Note: These observations used PCA (used 3 principal components and accounted $\sim$ 60% variance) for White Wine observations. See (See Figure 22)

Thus, with the model we can predict the quality of the wine around $\sim$ 87% accuracy and the PCA analysis will help judge the perfect blend of the Wine.

Appendix



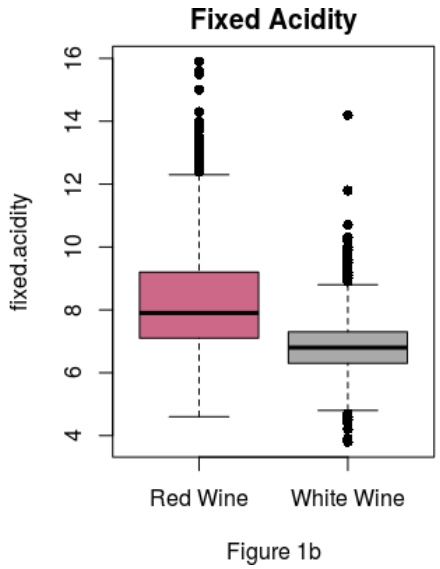
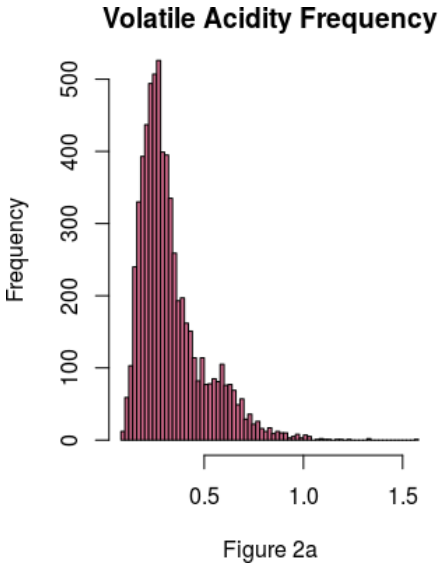


Figure 1a and Figure 1b: Distribution and boxplot of Fixed Acidity



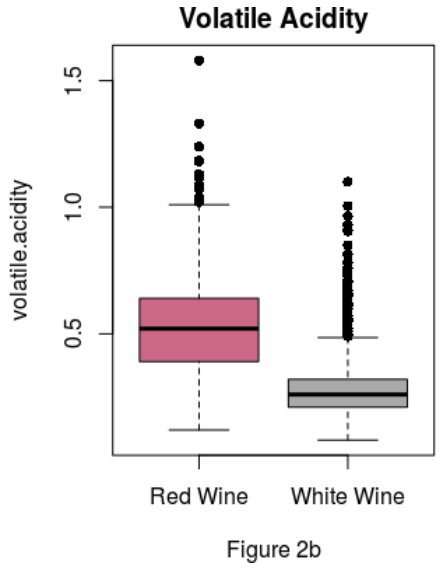


Figure 2a and Figure 2b: Distribution and boxplot of Volatile Acidity in Wines

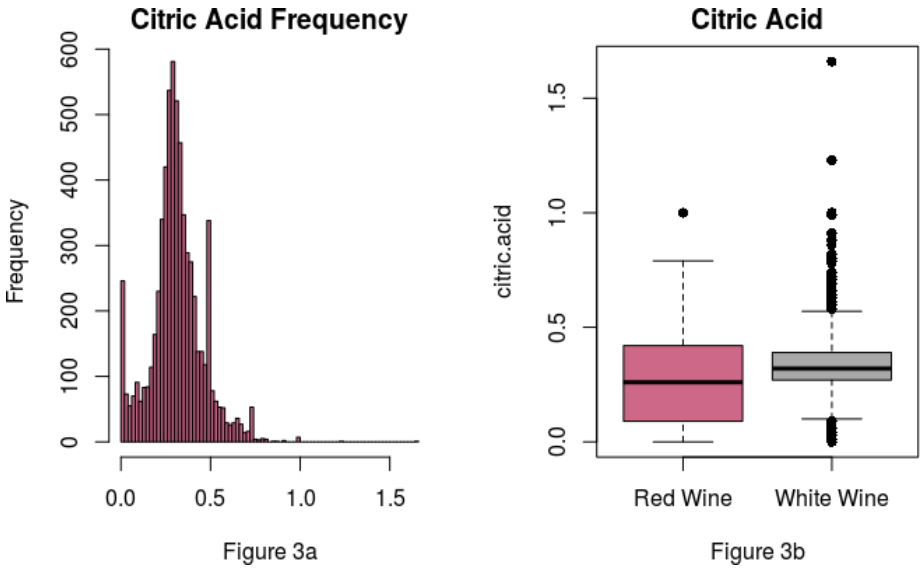


Figure 3a and Figure 3b: Distribution and boxplot of Citric Acid in Wines

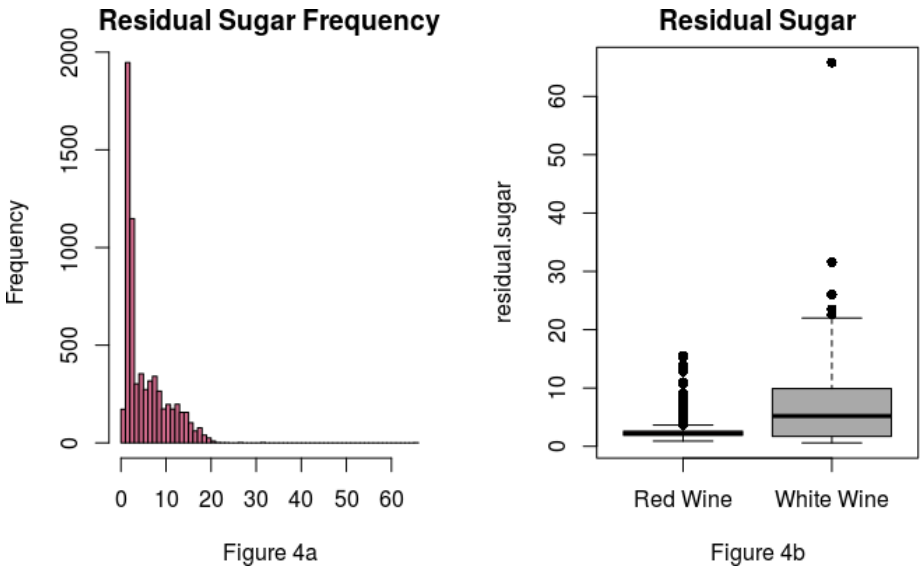


Figure 4a and Figure 4b: Distribution and boxplot of Residual Sugar in Wines

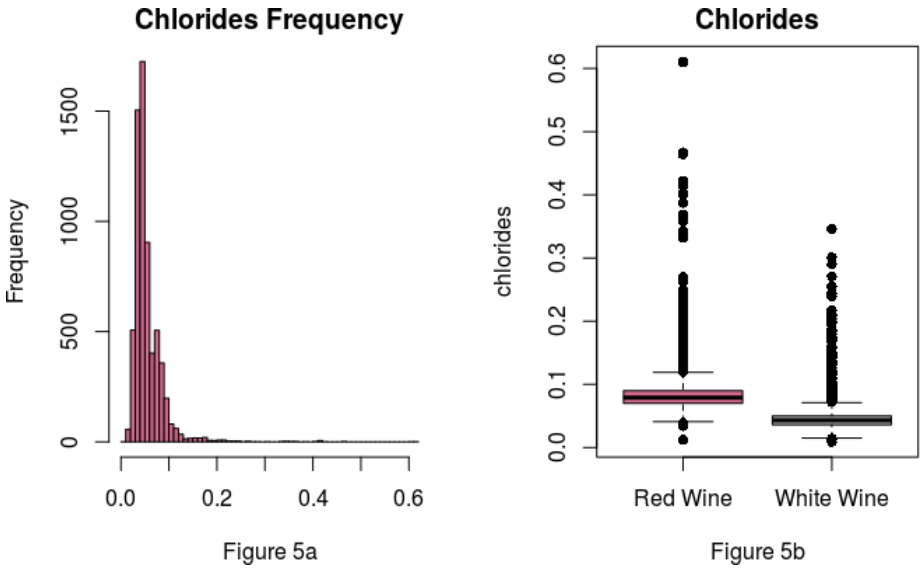


Figure 5a and Figure 5b: Distribution and boxplot of Chlorides in Wines

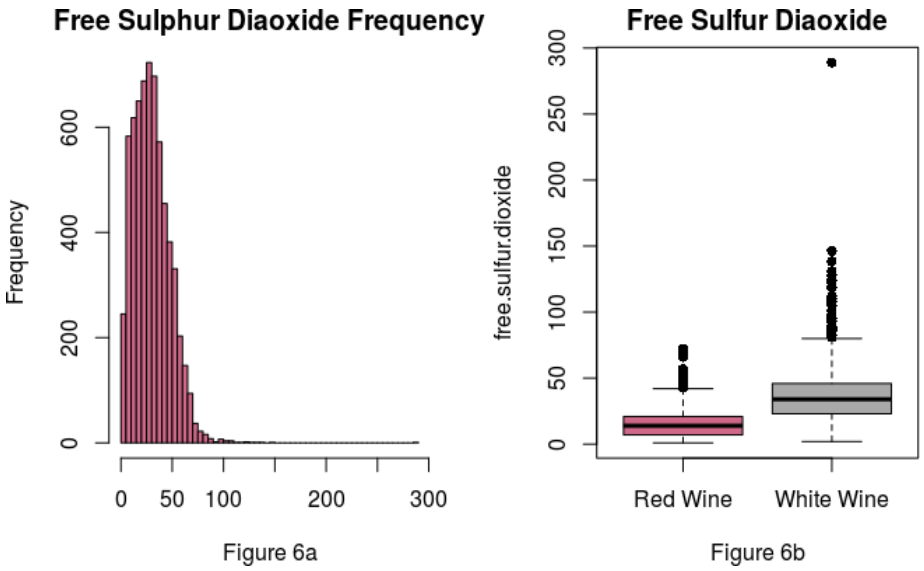


Figure 6a and Figure 6b: Distribution and boxplot of Free Sulfur Dioxide in Wines

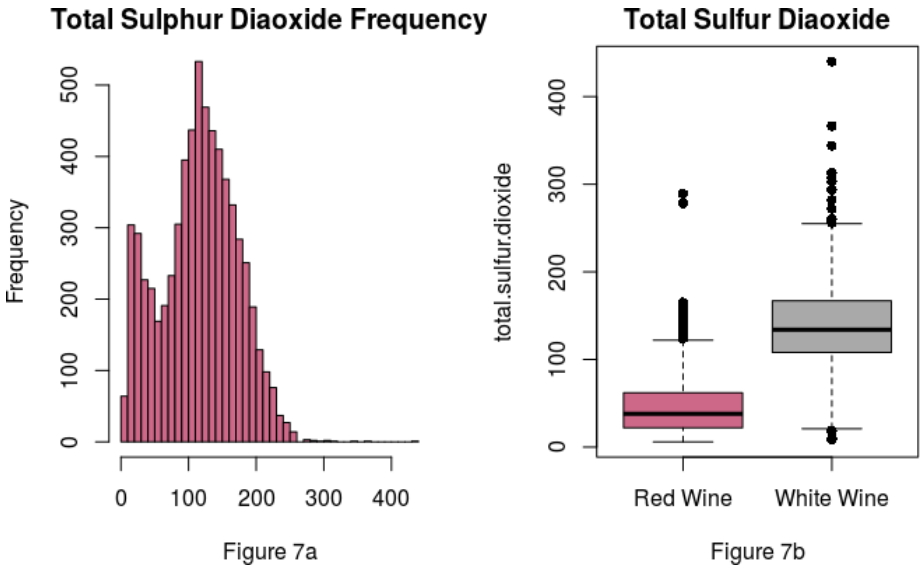


Figure 7a and Figure 7b: Distribution and boxplot of Total Sulfur Dioxide in Wines

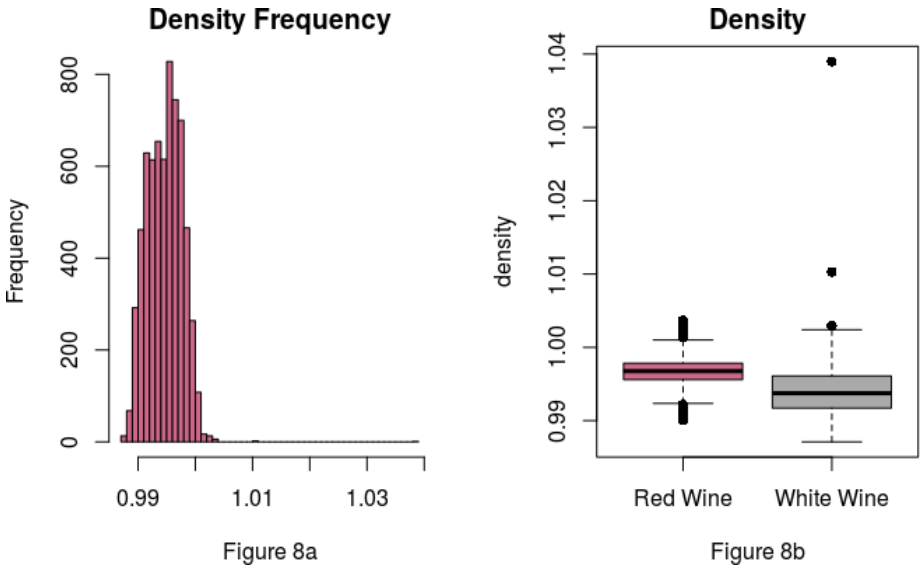


Figure 8a and Figure 8b: Distribution and boxplot of Density in Wines

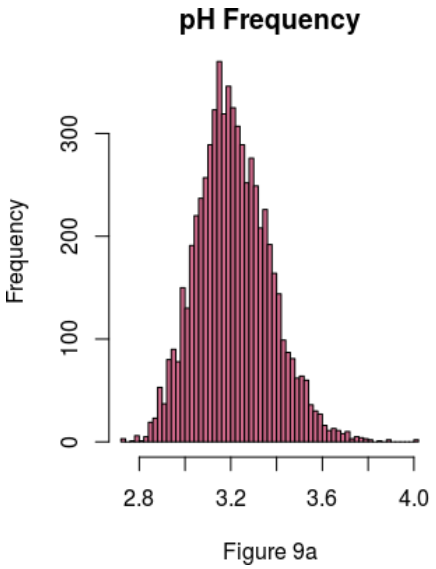


Figure 9a

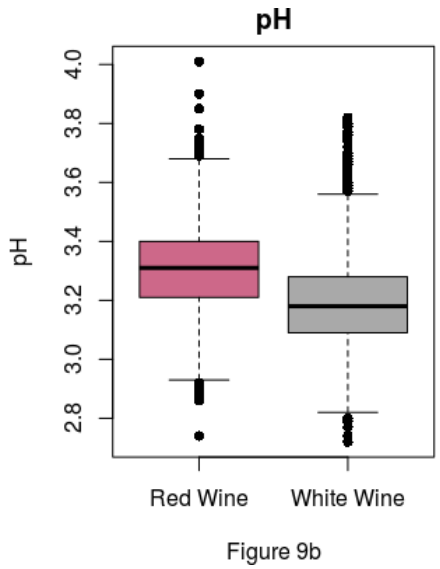


Figure 9b

Figure 9a and Figure 9b: Distribution and boxplot of pH Level in Wines

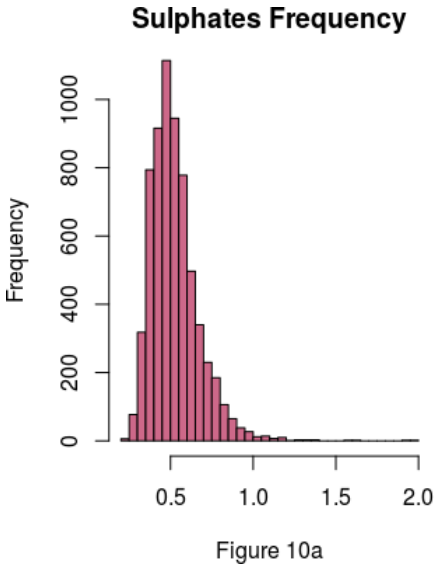


Figure 10a

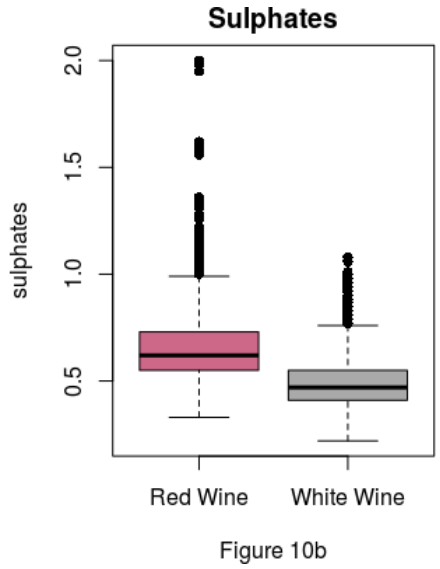


Figure 10b

Figure 10a and Figure 10b: Distribution and boxplot of Sulfates in Wines

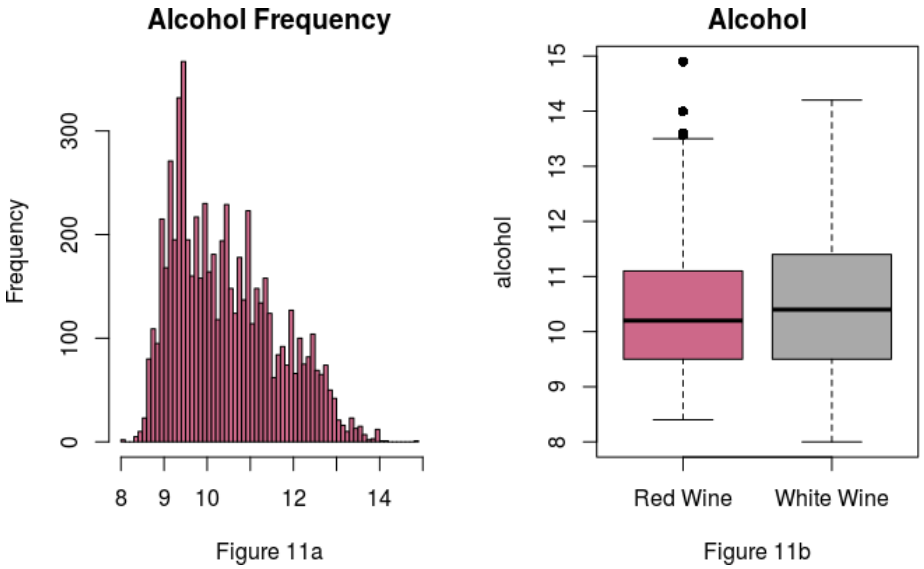


Figure 11a and Figure 11b: Distribution and boxplot of Alcohol Content in Wines

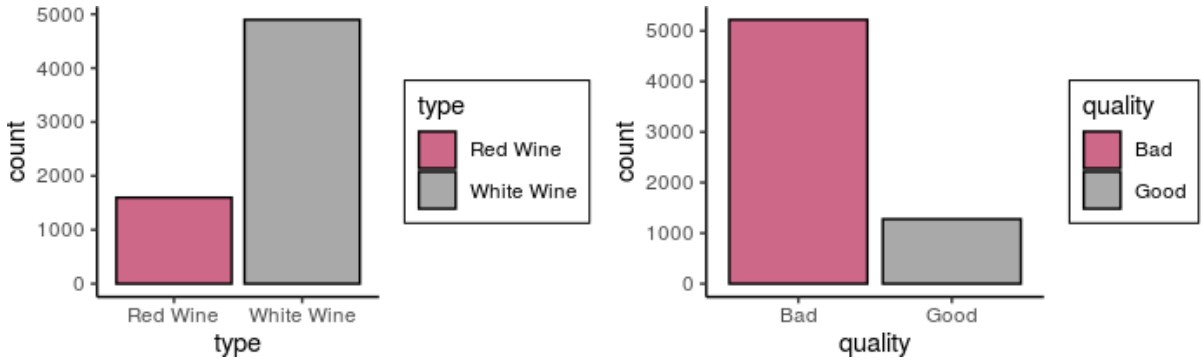


Figure 12a and Figure12b: Represents the distribution of Red/White Wine and Quality in Data

Figure 13: Correlation between different variables

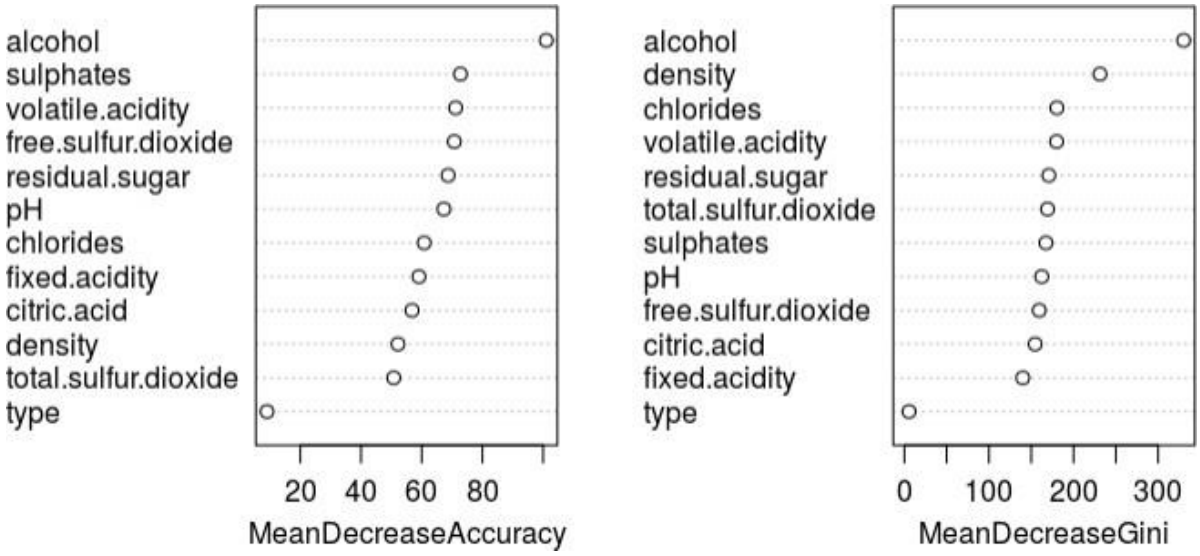


Figure 14: Variable Importance Results from Random Forest

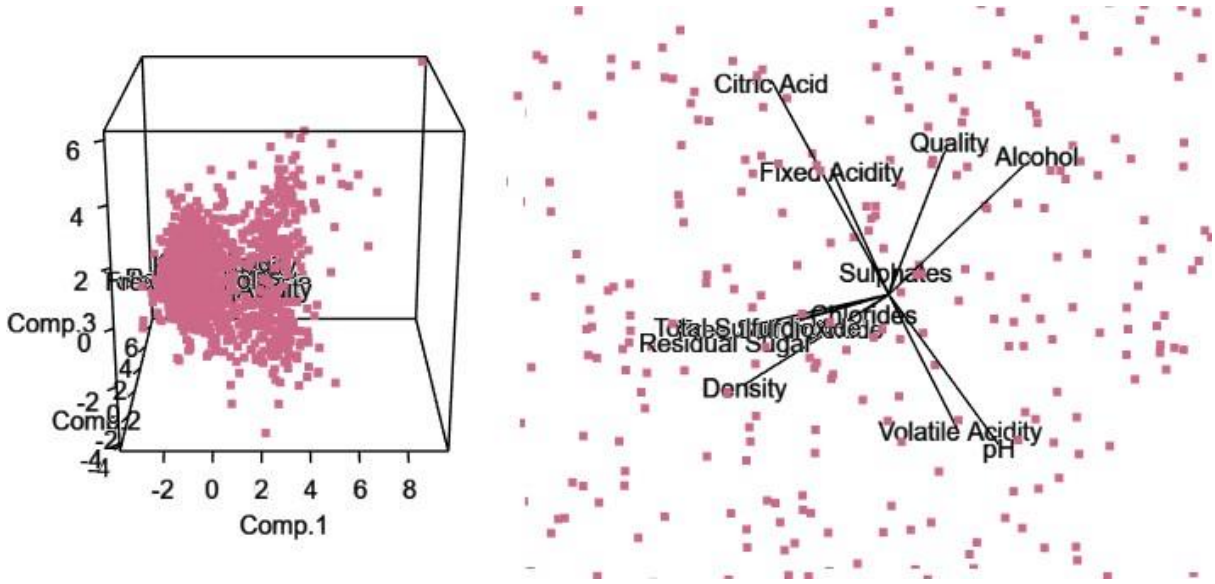


Figure 15: PCA analysis using 3 PCA component

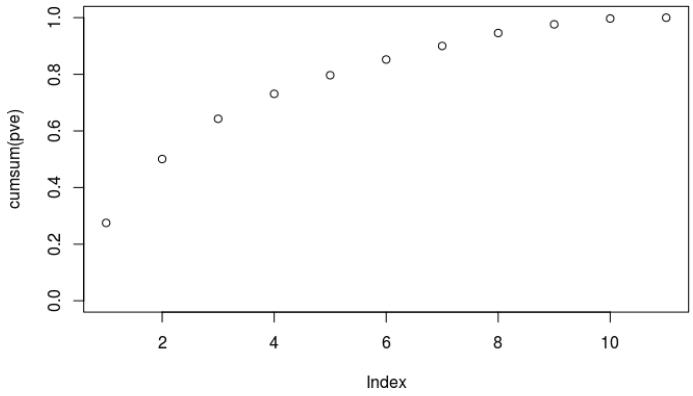


Figure 16: Cumulative Variation Distribution in PCA Analysis

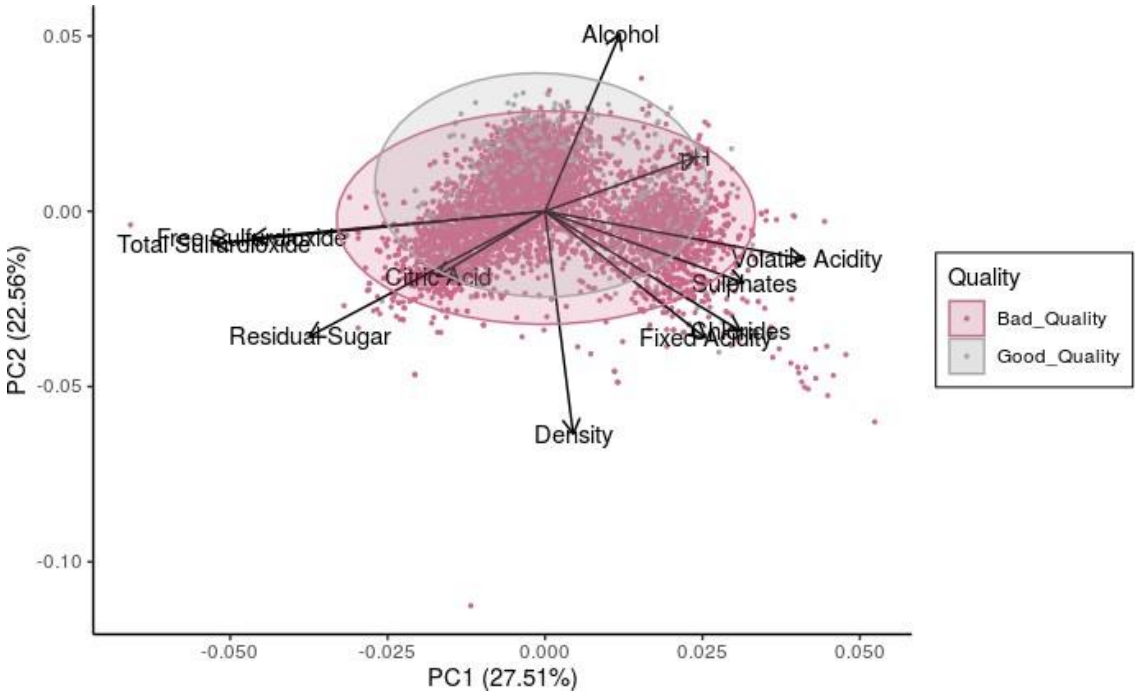


Figure 17: PCA Analysis

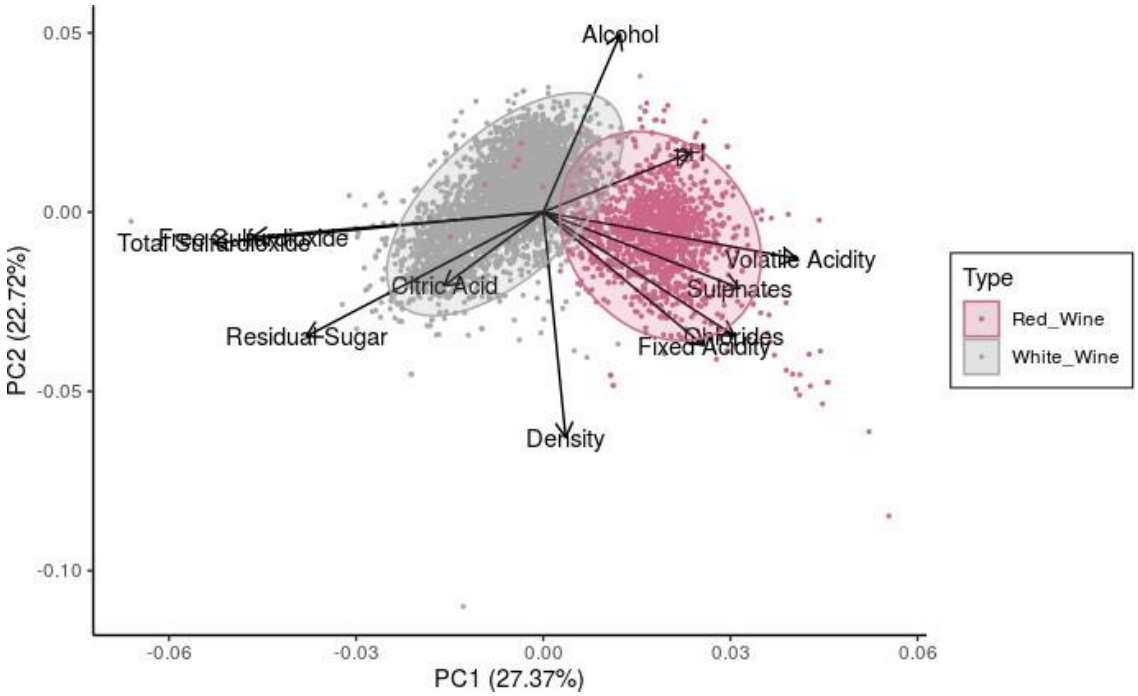


Figure 18: Plot of 2 PCA component with Loadings

Figure 19: 3-D plot of 3 PCA components with Loadings of Red and White Wine

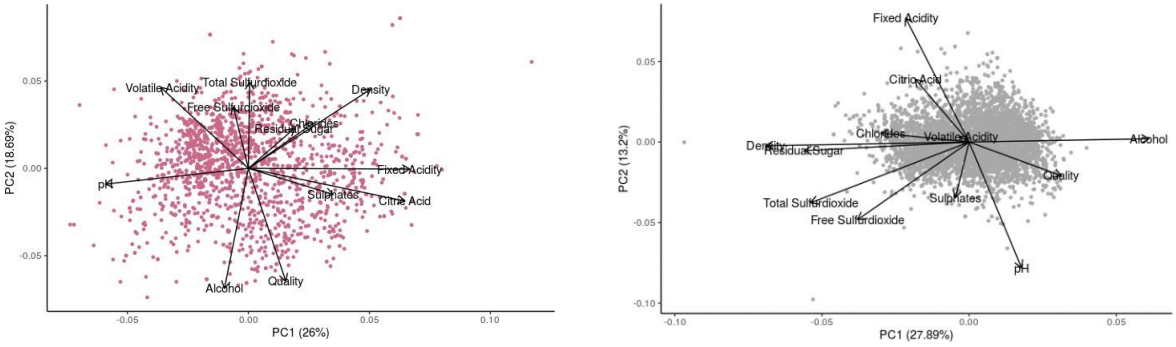


Figure 20 (a) and (b): 2-D Plot of 2 PC with loadings of (a) Red Wine and (b) White Wine

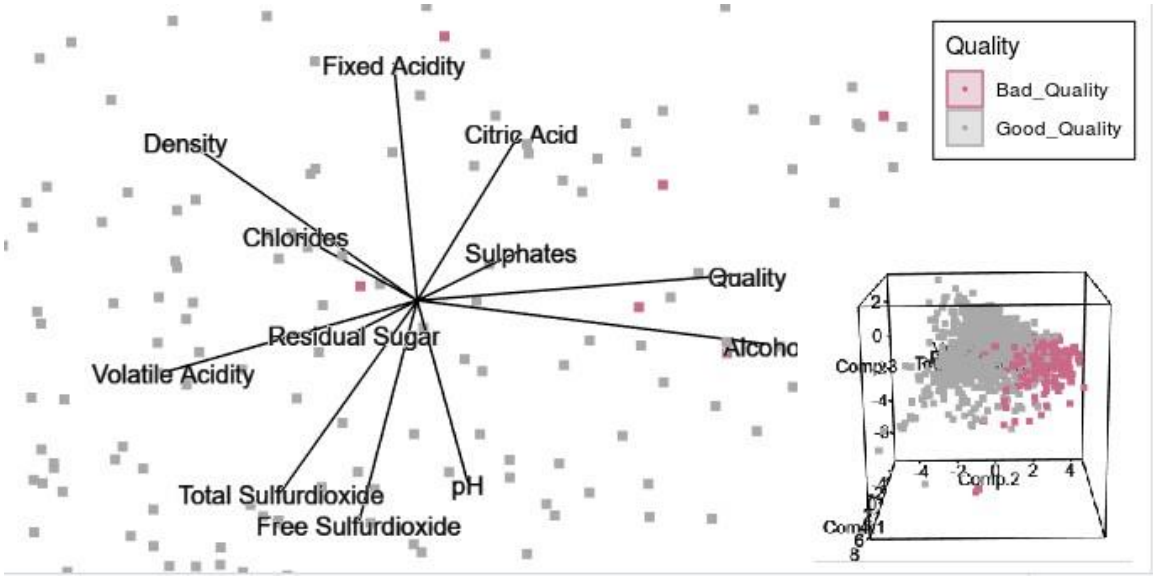


Figure 21: 3-D plot of 3 PCA components with Loadings of Red Wine

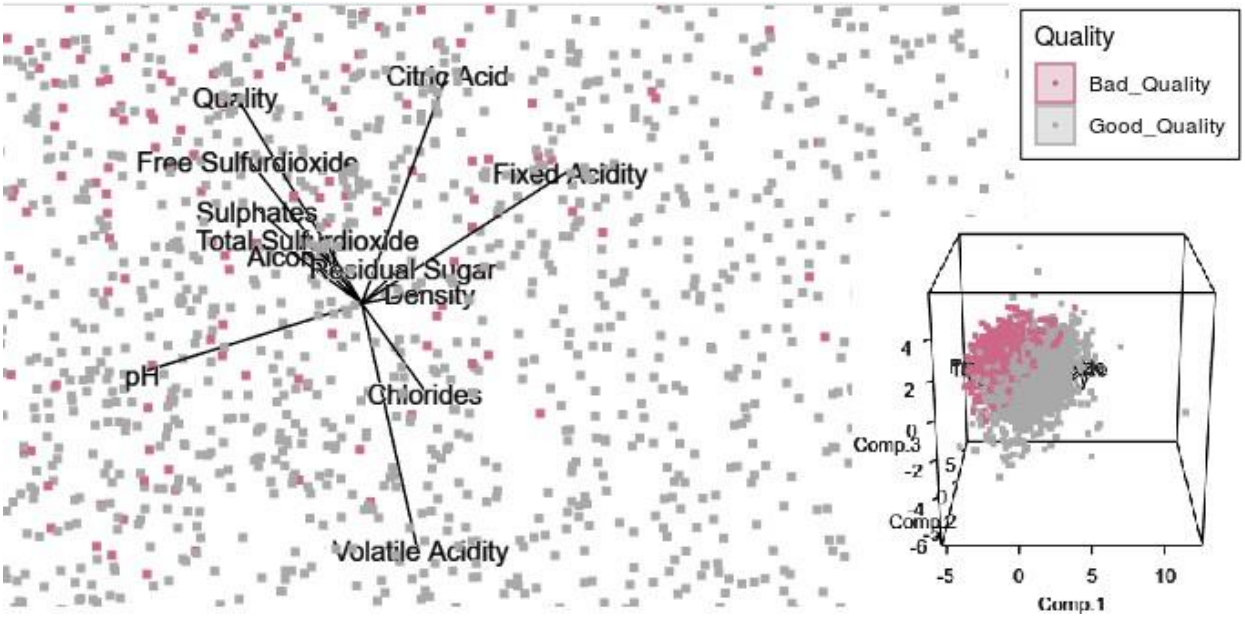


Figure 22:3-D plot of 3 PCA components with Loadings of Red Wine

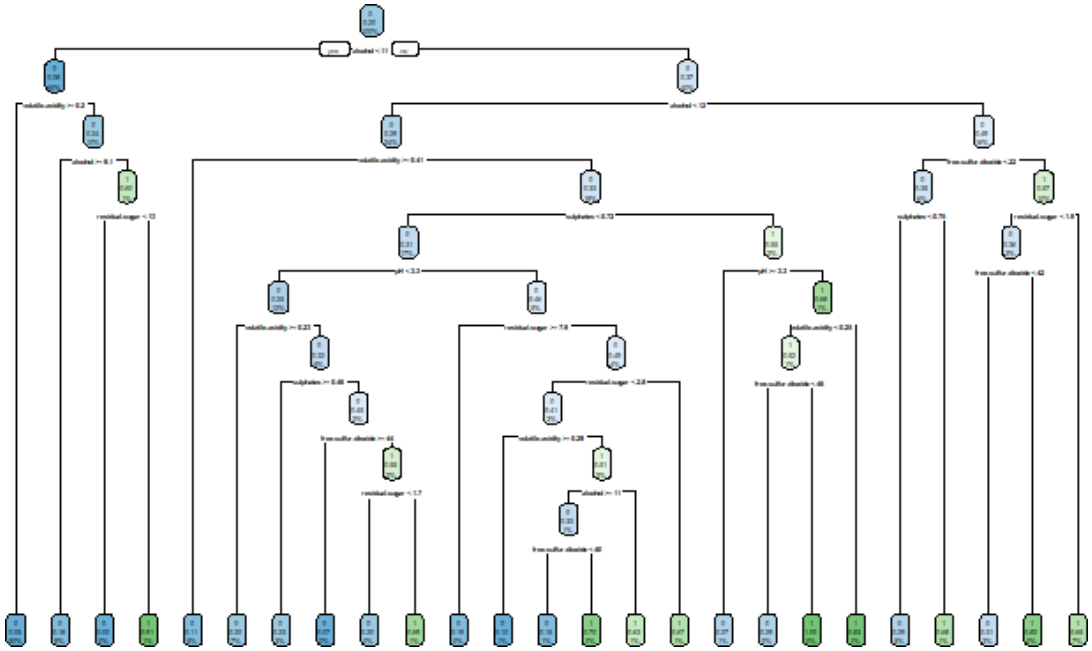


Figure 23: Decision Tree when important features selected using Random For- est

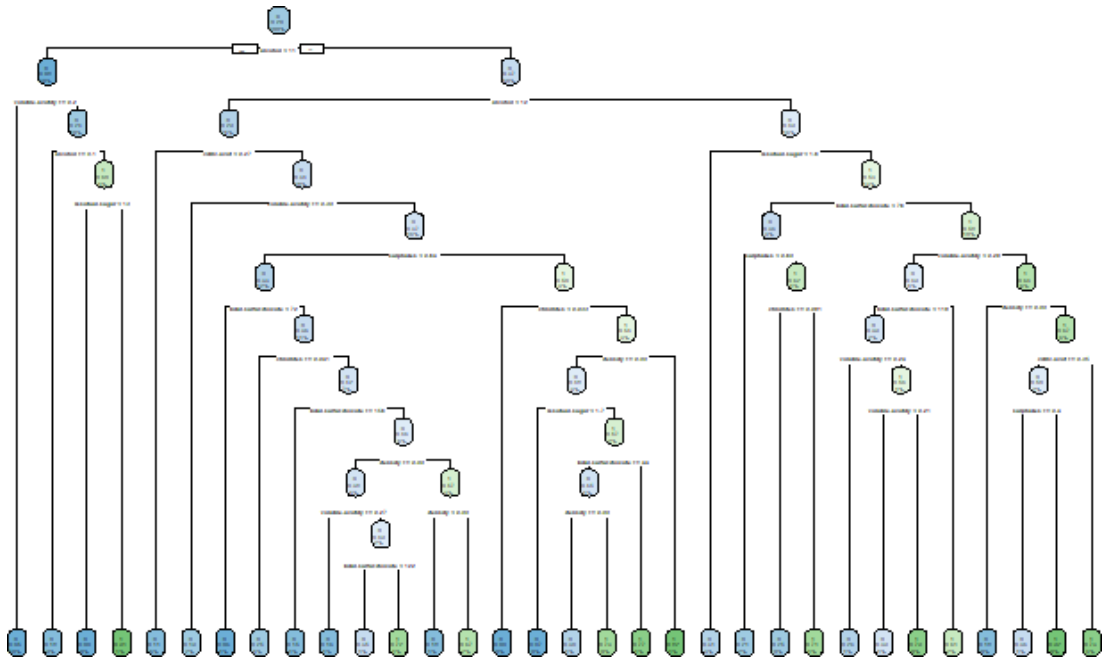


Figure 24: Decision Tree when important features selected using PCA

Variable Importance Results

	Bad	Good	MeanDecreaseAccuracy	MeanDecreaseGini
type	6.055	9.309	9.020	5.520
fixed.acidity	38.128	46.950	59.087	140.036
volatile.acidity	41.204	71.082	71.189	179.999
citric.acid	33.992	53.050	56.811	154.636
residual.sugar	49.128	45.987	68.697	170.812
chlorides	34.782	52.712	60.817	180.288
free.sulfur.dioxide	44.931	63.083	70.740	159.431
total.sulfur.dioxide	26.531	57.129	50.761	169.223
density	35.863	45.697	52.177	231.274
pH	49.043	59.123	67.288	162.456
sulphates	42.688	67.713	72.763	167.353
alcohol	57.005	82.189	101.033	330.152

Table 1: Variable Importance Results from Random Forest

```
Rotation (n x k) = (11 x 11):
                        PC1         PC2         PC3         PC4         PC5         PC6         PC7         PC8         PC9
Fixed Acidity        0.2404055 -0.33625270  0.42875169 -0.17428825  0.1627454  0.223779794  0.34719108 -0.31819709  0.3525534
 Volatile Acidity    0.3803532 -0.11085458 -0.31005749 -0.21057508 -0.1777464  0.491806424  0.35932016  0.15670958 -0.4971717
Citric Acid         -0.1482320 -0.18233031  0.59644458  0.25793921  0.1494505 -0.216932178  0.32188575  0.36001494 -0.4084847
Residual Sugar      -0.3477781 -0.32860673 -0.15988138 -0.17120910  0.3037986  0.238309976 -0.32641315  0.51022053  0.1017661
Chlorides            0.2881023 -0.31152449 -0.01519638  0.25589453 -0.6365240 -0.193511885 -0.02859125  0.43155977  0.2924838
Free Sulfurdioxide  -0.4300172 -0.07472421 -0.13132049  0.36519976 -0.2123986  0.344110353  0.32646679 -0.15440094  0.3616961
Total Sulfurdioxide -0.4874650 -0.09057693 -0.10649796  0.20735346 -0.1525423  0.145275128  0.16478904 -0.11567874 -0.3149460
Density              0.0509926 -0.58350056 -0.17427790 -0.07449712  0.3020817 -0.007750173  0.04325503  0.02087703  0.1128931
pH                   0.2203872  0.15889345 -0.44709028  0.42046878  0.4762270 -0.259436199  0.39538123  0.14327402  0.1291516
Sulphates            0.2962678 -0.19327243  0.08052573  0.62900441  0.1366672  0.303425747 -0.49733230 -0.25104393 -0.2104648
Alcohol              0.1006432  0.46851787  0.27048741  0.09796863  0.1304705  0.517493760  0.01006067  0.42048255  0.2490544
                       PC10        PC11
Fixed Acidity       -0.2791451273  0.338178311
 Volatile Acidity    0.1467045039  0.082018936
Citric Acid          0.2302541225 -0.003355235
Residual Sugar      -0.0053770486  0.435720514
Chlorides           -0.1929786837  0.042168336
Free Sulfurdioxide   0.4838490997  0.002480255
Total Sulfurdioxide -0.7152485414 -0.064450970
Density             -0.0009854516 -0.717382833
pH                  -0.1433794491  0.203243066
Sulphates            0.0371406300  0.077186709
Alcohol             -0.2064961243 -0.348141913
```

Table 2: Importance of PCA components and variance proportion

Accuracy Scores Using Random Forest Selected Features

	Models	Accuracy_Scores	Error_Rates
1	Logistic Regression	0.822	0.178
2	Decision Tree	0.826	0.174
3	Random Forest	0.877	0.123
4	Gradient Boosting	0.845	0.155

Table 3: Accuracy Scores and Error Rates Using Random Forest Feature Se- lection

Accuracy Scores Using PCA Selected Features

	Models_pca	Accuracy_Scores_pca	Error_Rates_pca
1	Logistic Regression	0.820	0.180
2	Decision Tree	0.826	0.174
3	Random Forest	0.876	0.124
4	Gradient Boosting	0.850	0.150

Table 4: Accuracy Scores and Error Rates Using PCA Feature Selection