A

Project Report

On

Wine Preference Prediction

# 

# **Guided by**

**Dr Syed Khasim**

**Submitted by**

**Saloni Reddy 20BCI7167**

**Sundar Raj Godina 20BCD7183**

Table of Contents

**Abstract** ........................................................................................................................................ 2

**Introduction** ................................................................................................................................. 2

**Objective** ...................................................................................................................................... 3

**About the collection of Dataset** .................................................................................................. 4

**Methodology** ............................................................................................................................... 4

**Implementation** ........................................................................................................................... 5

Import dataset and read dataset .................................................................................... 5

Exploratory data analysis ………........................................................................................ 5

Classification into categories .…….................................................................................... 10

Separate data into testing and training sets ………………….………………………..…………………..11

Building the model with random forest package …………………………………..…………………….11

Classification using decision trees …………………………………….…………………..…………………….12

## Correlation ……………………………………………………………………….……………………..………………….14

## Conclusion ....................................................................................................................................15

**References** ....................................................................................................................................15

**Abstract**

As an alcoholic beverage, wine has remained prevalent for thousands of years, and the quality assessment of wines has been significant in wine production and trade. For this project we will start with a general idea of the wine consumption throughout the world, including dataset analysis. A general description and exploratory analysis of the dataset will be presented which will be followed by the application of Decision Trees and Random Forests to perform prediction and correlation analysis. In addition, it compares the accuracy of both the models to suggest a better model for analysis and prediction of wine quality and thus improve the robustness of the proposed model. For predicting values of new data points, the model uses ‘correlation’, assigning a new point to values based on how close it resembles the points on the training set. The idea of this study is illustrating the different forecasting tools, comparing them, and analyzing the behavior of predictions.

**Introduction**

Despite the effect of the pandemic and extreme weather, the production of wine continues to grow. The United States, the world’s largest consumer of wine, consumes 33 million hectoliters of wine each year. Italy, France, and Spain are the largest three wine-producing countries in the world, producing 20.8 million hectoliters, 20.2 million hectoliters, and 13.6 million hectoliters of wine in 2020. As a commodity that is distributed in large quantities, wine has a well-established industrial production process and distribution network. The color of the wine, and especially the flavor and fragrance, define the quality and even the price of the wine. Wine contains about 97% of water and ethanol, the remaining 3% consists of organic acids, sugars, acetic acid, acetaldehyde, glycerol, higher alcohols, sorbitol and polyols, methanol, sulfites, amino acids, volatile esters, minerals, phenols. These chemical compositions determine the organoleptic quality (appearance, aroma, and flavor) of the wine. The distribution of chemical components in the wine is influenced by many factors, such as the quality of the grapes, the weather conditions, and the geographical environment in which the grapes are grown, as well as the temperature, humidity, and duration of the fermentation process. Fluctuations in these chemicals can be observed by examining the physicochemical properties of the wine. These physicochemical properties like pH value, alcohol content, sulfur content, anthocyanin content, and other properties of wine can greatly reflect its quality. It is important to note that these physicochemical characteristics do not independently affect the quality of the wine. Research on wine has revealed the interesting fact that there is an intrinsic link between the physicochemical features of the wine and that two or more features may work together to influence certain qualities of a wine. For example, the presence of ethanol can mask the acidity in the wine as well as lessen the concentration of some odors. The presence of organic acids lowers the pH of the wine and brings out the sourness. So, the content of ethanol and the content of organic acids, as well as the pH of the wine, have an impact on the presentation of sourness in the wine.

Traditionally, wine quality assessment has relied on manual sensory evaluation. Experienced tasters look at the appearance, aromas, and flavors of the wine in order to facilitate a comprehensive score. However, the disadvantages of this method are obvious: first of all, manual sensory evaluation relies heavily on personal perception, and different tasters may give very different results. Second, due to the lack of standardized quality evaluation criteria, manual sensory evaluation can no longer be adapted to mass production. The experience of a taster comes from the analysis of data during his or her career. There is a very limited amount of data that one can remember and process. Unlike humans, artificial intelligence (AI) has the ability to remember and learn from massive amounts of data. After trained by large amount of data, AI can work regardless of subjective factors to adapt to modern massive production.

**Objective**

The objectives of this project are to experiment with different classification methods to see which yields the highest accuracy and hence determine which features are the most indicative of a good quality wine. We also aim to visualize the different features which affect the wine quality and conclude with determining which features in the dataset are the best quality wine indicators to our model’s red wine quality.

**About the Dataset**

In the wine quality dataset, there are several features that will be used to classify the quality of wine, including many chemicals which we need to have a basic understanding of. The attributes/features include:

* **volatile acidity :**   Volatile acidityis thegaseous acids present in wine
* **fixed acidity** :   Primary fixed acids found in wine are tartaric, succinic, citric, and malic
* **residual sugar :**Amount of sugar left after fermentation
* **citric acid :** It is weak organic acid, found in citrus fruits naturally
* **chlorides :**Amount of salt present in wine
* **free sulfur dioxide :**   So2 is used for prevention of wine by oxidation and microbial spoilage
* **total sulfur dioxide :** Total SO2 concentration present in the final product
* **pH :**In wine pH is used for checking acidity
* **density :** Density of the fermented wine
* **sulphates**:    Added sulfites preserve freshness and protect wine from oxidation, and bacteria
* **alcohol :**   Percent of alcohol present in wine

**Methodology**

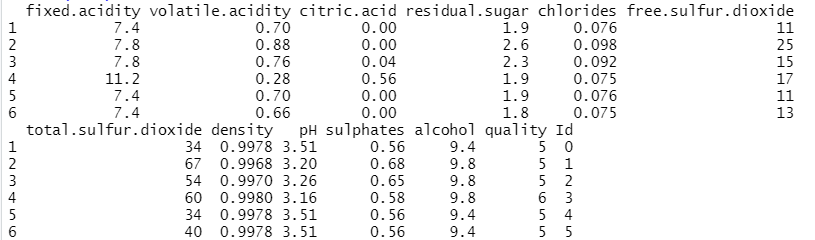
* The abstract and introduction gives a brief overview of the different aspects involved within this project. This includes some financial theory of Wine as global market, a concise explanation of the different algorithms used in this report and an overview of what we aim to find out by the end.
* We then proceed to understand the process in which the algorithm was implemented and how the data was used to obtain results. It will also describe how the different models were used to arrive at results. All the thoughts and decisions made within the implementation.
* The findings are then presented in the results section, where they are presented numerically through graphs and tables & Error calculations.
* The results are then analyzed. This means that one considers the error calculations and the results one arrived at and then one attempts to explain why the results look the way they do, and whether or not one can answer the research question based on the results one has arrived at.
* The conclusion then consists of a final statement which states whether or not the research question has been answered. It also must answer what implications the results one arrived at have had for the future of the concerned areas.

**Implementation**

**IMPORT AND READ WINE DATASET**

wine <- read.csv('WineQT.csv')

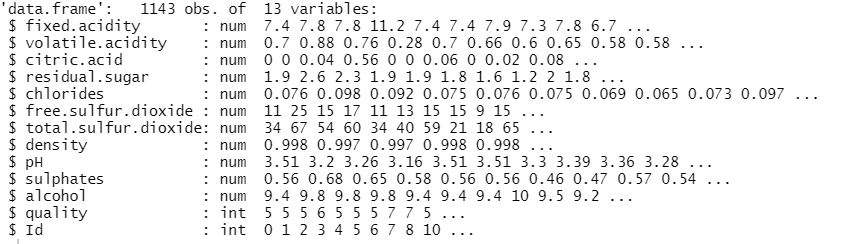
head(wine)



**EXPLORATORY DATA ANALYSIS**

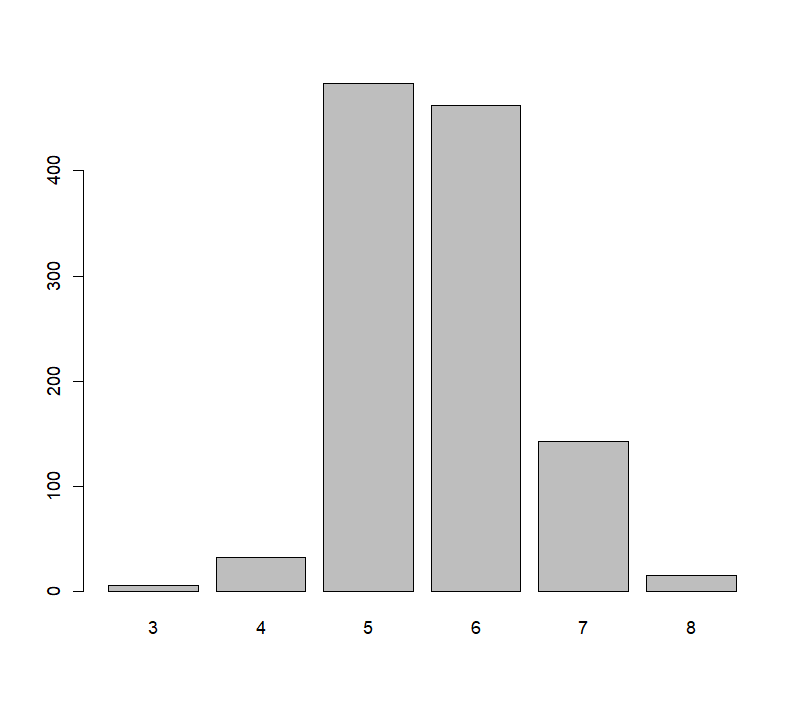
Viewing details of WineQT dataset:

str(wine)



Distribution of wine quality:

barplot(table(wine$quality))



Distribution of features if the dataset:

library(ggplot2)

library(gridExtra)

n1 <- qplot(x = fixed.acidity, data = wine,binwidth = 0.1) +

scale\_x\_continuous(breaks = seq(4, 16, 1))

n2 <- qplot(x = volatile.acidity, data = wine, binwidth = 0.01) +

scale\_x\_continuous(breaks = seq(0.12, 1.58, 0.1))

n3 <- qplot(x = citric.acid, data = wine, binwidth = 0.01) +

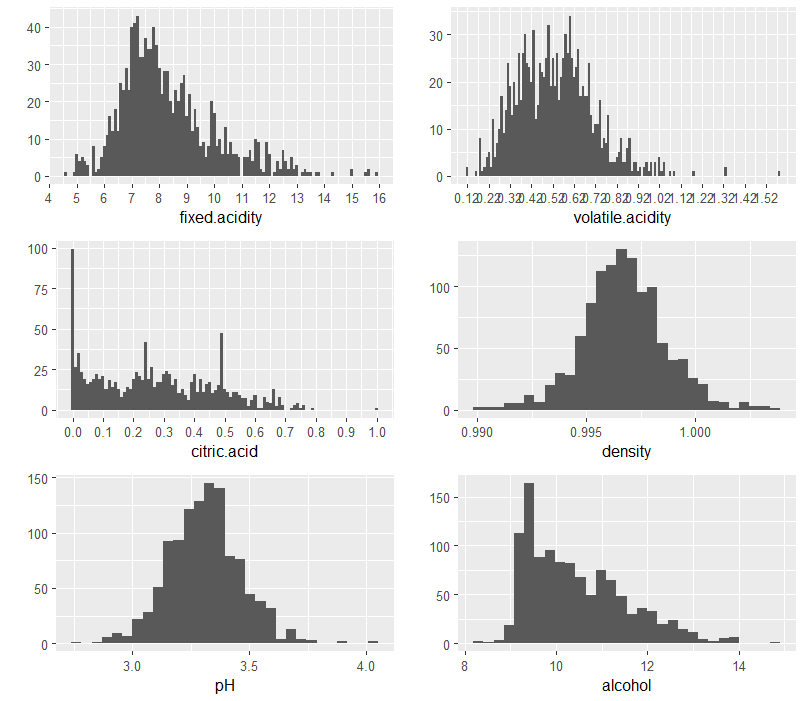
scale\_x\_continuous(breaks = seq(0, 1, 0.1))

n4 <- qplot(x = density, data = wine)

n5 <- qplot(x = pH, data = wine)

n6 <- qplot(x = alcohol, data = wine)

grid.arrange(n1, n2, n3, n4, n5, n6, ncol = 2)



Relationship between Wine Characteristics and Its Quality

1. Features increasing the quality of wine

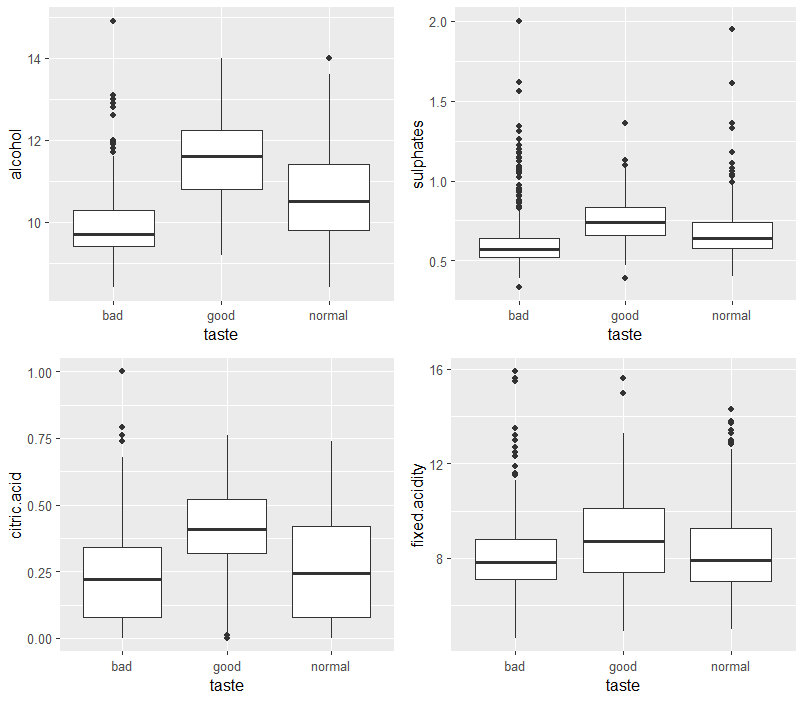
p1up = qplot(x = taste, y = alcohol, data = wine, geom = "boxplot")

p2up = qplot(x = taste, y = sulphates, data = wine, geom = "boxplot")

p3up = qplot(x = taste, y = citric.acid, data = wine, geom = "boxplot")

p4up = qplot(x = taste, y = fixed.acidity, data = wine, geom = "boxplot")

grid.arrange(p1up, p2up, p3up, p4up, ncol = 2)



1. Features increasing the quality of wine

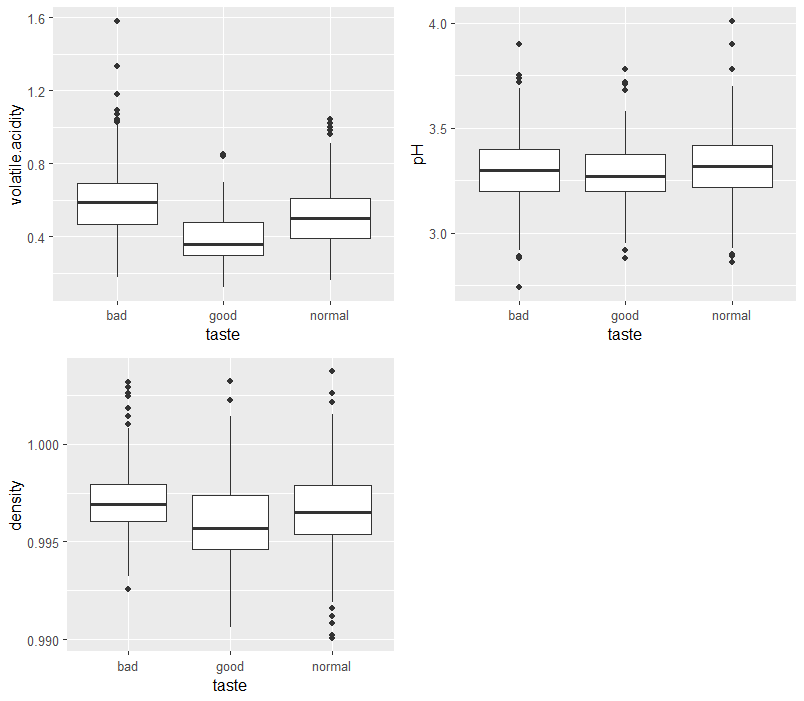
p1d = qplot(x = taste, y = volatile.acidity, data = wine, geom = "boxplot")

p2d = qplot(x = taste, y = pH, data = wine,geom = "boxplot")

p3d = qplot(x = taste, y = density, data = wine, geom = "boxplot")

p4d = qplot(x = taste, y = density, data = wine,geom = "boxplot")

grid.arrange(p1d, p2d, p3d, ncol = 2)



Multivariate Analysis:

wine$quality.factor <- factor(wine$quality)

p1<- ggplot(aes(x = log10(sulphates), y = alcohol, colour = quality.factor),data = wine) +

geom\_point(aes(size = quality.factor)) +

scale\_color\_brewer(type = 'div', palette="Set1") +

scale\_x\_continuous(lim=c(quantile(log10(wine$sulphates), 0.01),

quantile(log10(wine$sulphates), 0.99)))+

scale\_y\_continuous(lim=c(quantile(wine$alcohol, 0.01),

quantile(wine$alcohol, 0.99)))

p2<-ggplot(aes(x = fixed.acidity, y = density, colour = quality.factor), data=wine) +

geom\_point(size = 4) + #geom\_point() +

scale\_color\_brewer(type = 'div', palette="Set1") +

scale\_x\_continuous(lim=c(quantile(wine$fixed.acidity, 0.01),

quantile(wine$fixed.acidity, 0.99))) +

scale\_y\_continuous(lim=c(quantile(wine$density, 0.01),

quantile(wine$density, 0.99)))

p3 <- ggplot(aes(x = pH, y = total.sulfur.dioxide, colour = quality.factor), data = wine) +

geom\_point(aes(size = quality.factor)) +

scale\_color\_brewer(type = 'div', palette="Set1") +

scale\_x\_continuous(lim=c(quantile(wine$pH, 0.01),

quantile(wine$pH, 0.99))) +

scale\_y\_continuous(lim=c(quantile(wine$total.sulfur.dioxide, 0.01), quantile(wine$total.sulfur.dioxide, 0.99)))

p4 <- ggplot(aes(x = log10(total.sulfur.dioxide), y = log10(free.sulfur.dioxide), colour = quality.factor), data = wine) +

geom\_point(aes(size = quality.factor)) +

#geom\_point(aes(size = 12)) +

scale\_color\_brewer(type = 'div', palette="Set1") + scale\_x\_continuous(lim=c(quantile(log10(wine$total.sulfur.dioxide),

0.01), quantile(log10(wine$total.sulfur.dioxide),0.99))) +

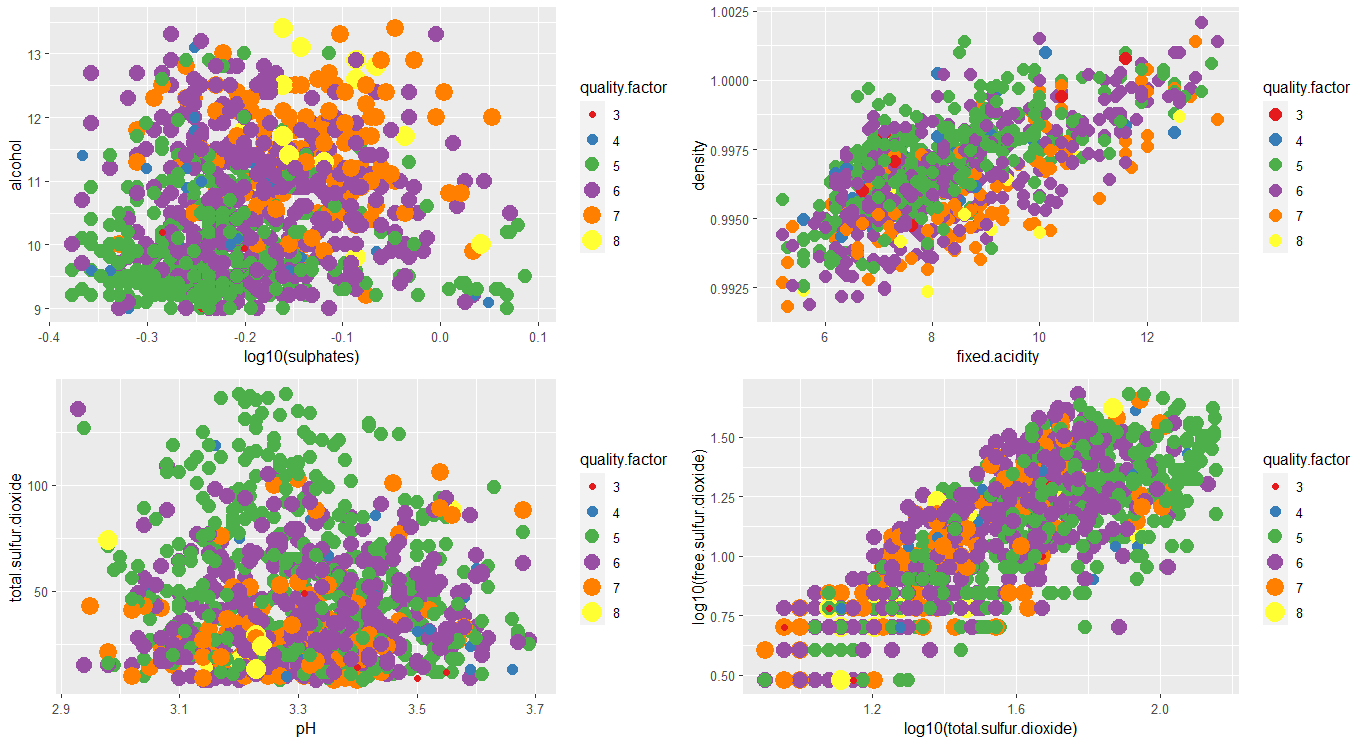
scale\_y\_continuous(lim=c(quantile(log10(wine$free.sulfur.dioxide),

0.01),

quantile(log10(wine$free.sulfur.dioxide),

0.99)))

grid.arrange(p1, p2, p3, p4, ncol = 2, nrow = 2)



**CLASSIFICATION INTO CATEGORIES**

Classifying wine into good, bad and normal:

wine$taste <- ifelse(wine$quality < 6, 'bad', 'good')

wine$taste[wine$quality == 6] <- 'normal'

wine$taste <- as.factor(wine$taste)

wine$taste



**SEPERATE DATA INTO TESTING AND TRAINING SETS:**

> set.seed(123)

> samp <- sample(nrow(wine), 0.6 \* nrow(wine))

> train <- wine[samp, ]

> test <- wine[-samp, ]

This will place 60% of the observations in the original dataset into train and the remaining 40% of the observations into test.

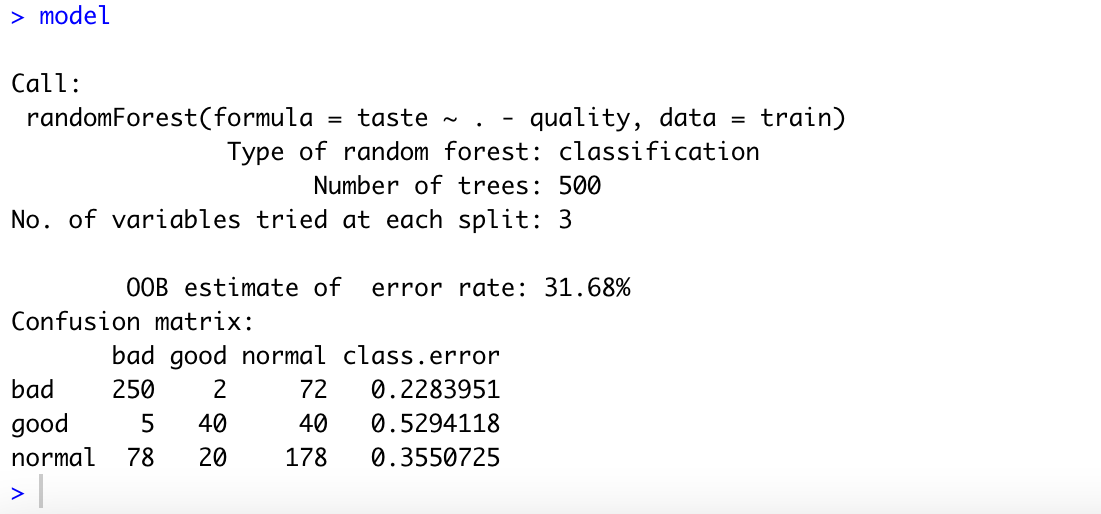
## BUILDING THE MODEL WITH RANDOMFOREST PACKAGE:

library(randomForest)

model <- randomForest(taste ~ . - quality, data = train)

install.packages("randomForest")

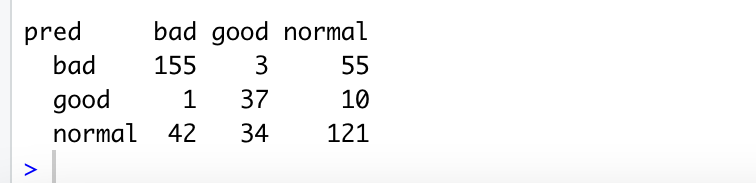
model



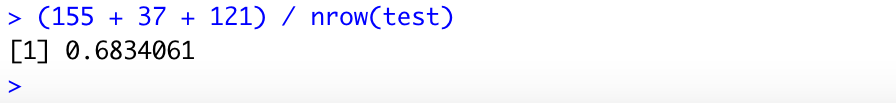
We can see that 500 trees were built, and the model randomly sampled 3 predictors at each split. It also shows a matrix containing prediction vs actual, as well as classification error for each class. Let’s test the model on the test data set.

pred <- predict(model, newdata = test)

table(pred, test$taste)



Now, we can test the accuracy:



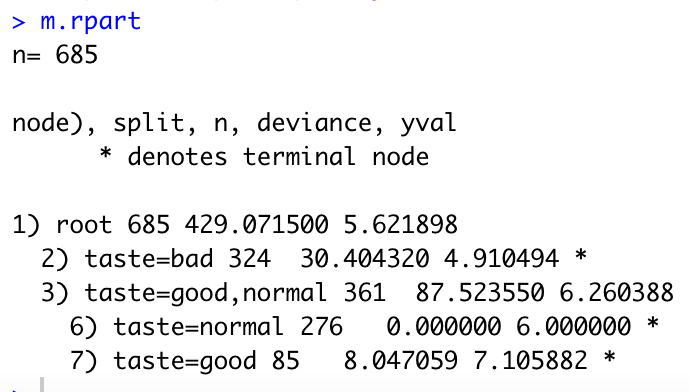
so, it has 68% accuracy.

**CLASSIFICATION USING DECISION TREES:**

The **rpart()** will be used to specify quality as the outcome variable and use the dot notation to allow all the other columns in the wine\_train data frame to be used in predictors.

m.rpart <- rpart(quality ~ ., data = train)

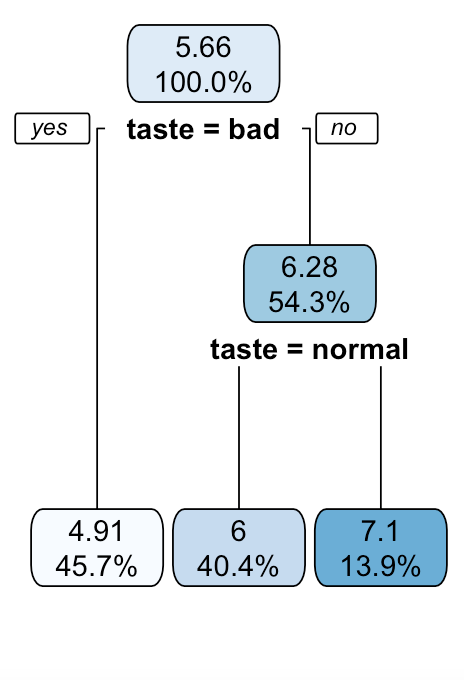
m.rpart



## Decision Tree Visualization

As one can see from the below, the visualization in the decision tree is much easier to read. Also, the digits parameter rounds all digits to the 3 places.

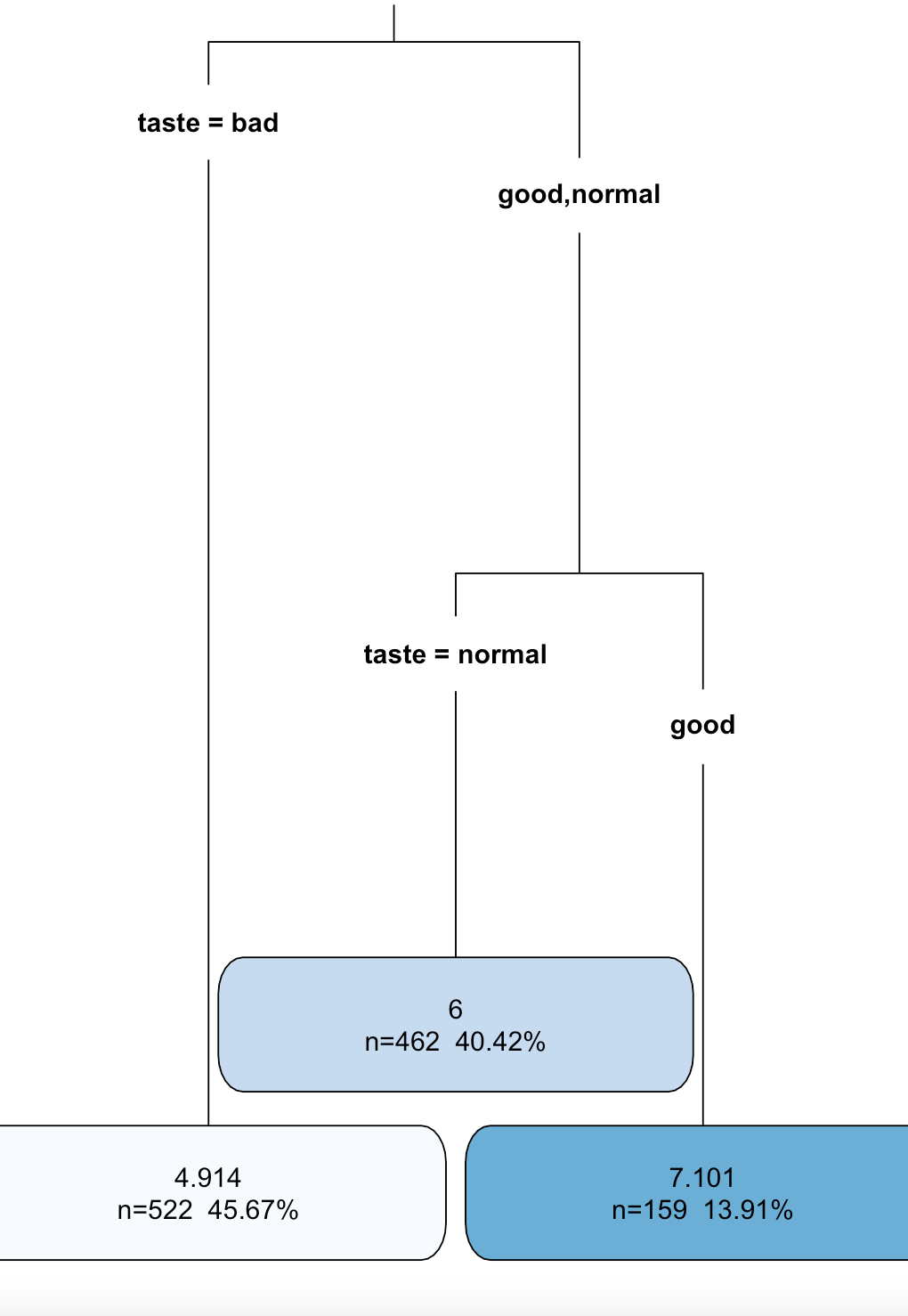
rpart.plot(m.rpart, digits = 3)



fallen.leaves() addition to the decision tree

This addition will show visualizations with the dissemination of regression tree results, as they are readily understood even without a mathematics background. The lead nodes are predicted values for the examples reaching that node.

rpart.plot(m.rpart, digits = 4, fallen.leaves = TRUE, type = 3, extra = 101)

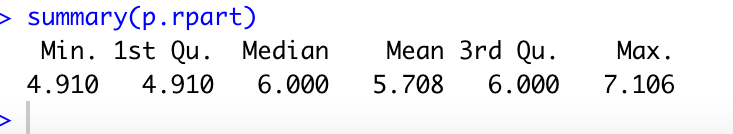


## Test Data Prediction

We must now make predictions on the test data, we use the **predict()** function. This will return the estimated numeric value for the outcome variable.

p.rpart <- predict(m.rpart, test)

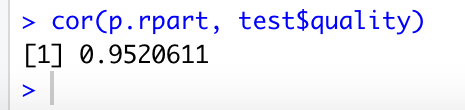
summary(p.rpart)



## CORRELATION

We can now check the correlation between the predicted and actual quality values provides a simple way to gauge the model’s performance. As one can see that the outcome is 0.54, which is acceptable but not ideal. This correlation only measures how strong the predictions are related to the true value. This is not a measure of how far off the predictions were from the true values.

cor(p.rpart, test$quality)

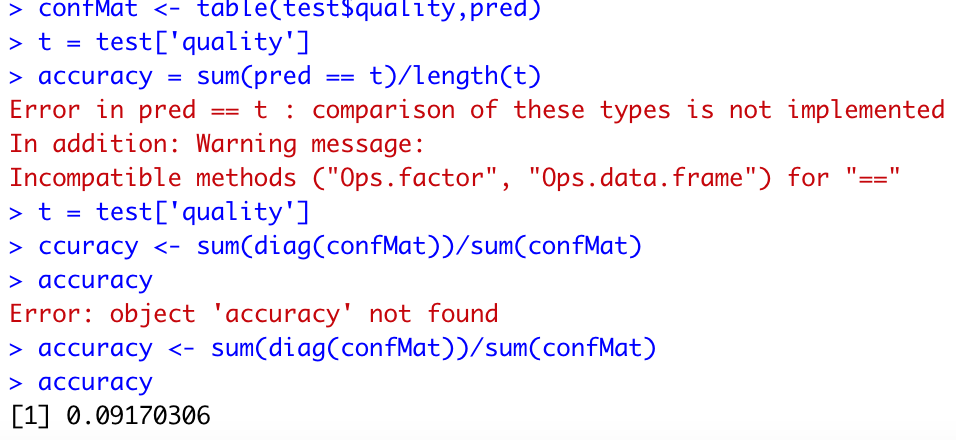


calculating accuracy of decision tree:

> t = test['quality']

> accuracy <- sum(diag(confMat))/sum(confMat)

> accuracy



Accuracy is only 9% which means that random forest is more accurate compared to the decision tree.

**Conclusion**

Discussing the wine quality issues in an overly complex and technical area of machine learning cannot go without a lyrical mood, of course. However, machine learning algorithms prove to be highly effective for wine quality assessment in the modern wine industry. Even though there’s still a lot of room for growth, we believe that ML can be safely used for product quality certification. Automationworks wonders for today’s manufacturers in terms of quality, time, speed, and effectiveness of the task assigned. Random forest model has better accuracy regarding wine quality prediction compared to decision tree model.

**References**

[**https://www.engineeringbigdata.com/white-wine-quality-analysis-with-regression-and-model-trees-in-r/**](https://www.engineeringbigdata.com/white-wine-quality-analysis-with-regression-and-model-trees-in-r/)

[**https://www.r-bloggers.com/2016/02/predicting-wine-quality-using-random-forests/**](https://www.r-bloggers.com/2016/02/predicting-wine-quality-using-random-forests/)

[**https://stackoverflow.com/questions/40080794/calculating-prediction-accuracy-of-a-tree-using-rparts-predict-method**](https://stackoverflow.com/questions/40080794/calculating-prediction-accuracy-of-a-tree-using-rparts-predict-method)

[**https://www.kaggle.com/datasets/yasserh/wine-quality-dataset**](https://www.kaggle.com/datasets/yasserh/wine-quality-dataset)