A Final Project Report on

# *Wine Quality Determinators*

Prepared By

BA with R

Group 11

ShenJun Liu - sxl200043

Zhang Li- lxz200000

Srishti Vinaik- sxv210063

Fayad Hasan Khan- fxk200001

under the supervision of

Prof. Zhe Zhang

TABLE OF CONTENT

[Executive Summary 2](#_Toc89127655)

[Data Description 3](#_Toc89127656)

[Summary Statistics 4](#_Toc89127657)

[Data Processing 4](#_Toc89127658)

[Lowering impacts of outliers in each variable 4](#_Toc89127659)

[Training dataset and validation dataset 4](#_Toc89127660)

[Classification 4](#_Toc89127661)

[DATA VISUALIZATION AND EXPLORATION 5](#_Toc89127662)

[Data Distribution 5](#_Toc89127663)

[Correlation (Heatmap) 5](#_Toc89127664)

[DATA MINING 6](#_Toc89127665)

[Clustering 6](#_Toc89127666)

[Decision Tree 6](#_Toc89127667)

[Regression Analysis 6](#_Toc89127668)

[Linear Regression 6](#_Toc89127669)

[Logistic Regression 7](#_Toc89127670)

[FINDINGS AND RECOMMENDATIONS 7](#_Toc89127671)

[Findings 7](#_Toc89127672)

[Recommendation 8](#_Toc89127673)

[APPENDIX 9](#_Toc89127674)

[Variables Descriptions 9](#_Toc89127675)

[Figures 10](#_Toc89127676)

[Data Processing 10](#_Toc89127677)

[DATA VISUALIZATION AND EXPLORATION - Histogram 11](#_Toc89127678)

[DATA MINING - Quality Clustering 12](#_Toc89127679)

[Regression Analysis 15](#_Toc89127680)

[Codes 18](#_Toc89127681)

# Executive Summary

# Introduction

Red wine is one of the most favored types of alcoholic drink in the United States where it consumes the largest volume of wine of any country, at 33 million hectoliters in 2020.

Vinho Verde is a unique wine product from the Minho (northwest) region of Portugal. The general impression of the type of wine is described as light and fresh,complex, structured and mineral.

Vinho Verde wine quality can be determined by numeric variables which can be measured by physicochemical tests. Physicochemical tests, which are laboratory-based and take into consideration elements such as acidity, pH level, presence of sugar, and other chemical features, are another important factor in wine certification and quality assessment. The human quality of tasting might be linked to wine's chemical attributes so that the certification and quality assessment and assurance process is more controlled.The test seems to provide a more objective result than people’s feelings do.

Objective

What we are concerned about is the physicochemical tests of Vinho verde wine: What attributes determine the Vinho Verde wine? Can we use the results to predict human wine tasting reference?

Research

To answer the questions, we need to find out whether different levels of wine quality share the same degree of attributes. We also need to find relations between attributes and wine quality. Lastly, It is necessary to know to what extent the level of wine quality can be predicted by wine attributes. After all the analysis, we will give a finding summary in regards to what we observe from the outputs and recommendations on how to improve the wine production.

# Data Description

The dataset is taken from the UCI machine learning repository website. There are 12 attributes in the dataset that cover the features of each red wine along with the quality score. And all attributes’ characteristics are numerical, the number of observations is 1599. There is no missing value in the dataset. The dataset looks like the following:

Table

Description automatically generated

Table 1 - The Original Dataset

## Summary Statistics

From observing the results of the summary statistics table, the variables except total sulfur dioxide and free sulfur dioxide have relatively smaller standard deviations, which means the values are more concentrated than those two variables. Also, total sulfur dioxide and free sulfur dioxide have the largest range values, which are 283 and 67 separately.

Text, table

Description automatically generated

Table 2 - Summary Statistics

# Data Processing

## Lowering impacts of outliers in each variable

According to the dataset summary and box plot of each variable. The following variables: f**ixed.acidity, volatile.acidity, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, sulphates** contain huge number of outliers. Thus, we need to oversample the outliers to obtain a more balanced training set.

Any data values of the seven variables which are out of the range between Q1-1.5\*IQR and Q3+1.5\*IQR will be chosen with 0.01 probability. The rest of the values will be chosen with 0.99 probability.

## Training dataset and validation dataset

We randomly select 40% observations of all datasets as a training data set, 60% observations of all datasets as a validation data set.

## Classification

Before we conduct regression analysis, we wish to explore the dataset by developing a decision tree. Created a new column to illustrate whether the wine belongs to good quality or bad with column name “good quality”.

# DATA VISUALIZATION AND EXPLORATION

## Data Distribution

Most wine quality is among 5 to 7. According to the histogram of wine quality, we set a cut-off value of 6, which means any wine with 6 rating or above will be considered as good quality.

For the alcohol of wine, the maximum value is 14, minimum 10, and most of value are among 10 to 14, with average 10.48. And for the value distribution chart of pH, density, fixed acidity and volatile acidity, they all show normal distribution.

## 

## Correlation (Heatmap)

Chart, table, treemap chart

Description automatically generatedDetecting the underlying correlation between any two variables by using correlation matrix. There are strong positive relationships between free sulfur dioxide and total sulfur dioxide (0.67), fixed acidity and citric acid (0.67), fixed acidity and density (0.67) and alcohol and quality (0.48). And strong negative relationships between fixed acidity and pH (-0.68), volatile acidity and citric acid (-0.55), citric acid and pH (-0.54) and density and alcohol (-0.5).

Figure 2 - Correlation Heatmap

# DATA MINING

## Clustering

According to the quality score. We try to find if red wines share the same attributes when they are at the same quality score range. Since we have a large data set of 1599 observations, we tried to sample analysis by randomly selecting 300 observations. Using the Ward distance method generated a balanced dendrogram of the 300 observations.

by cutting that tree at h= 20, 13 clusters are produced and we produced a heatmap of these 13 clusters

From the generated heat map, we can see that clusters such as 2 which are high in pH value are very low in quality, same goes with the variable volatile acidity as well. Cluster 5,8 and 9 seem to have high quality wines in them and their pH value range is within low-medium and sulphate values are in the similar range as well. These high-quality clusters also share the same trait in case of residual sugar, chlorides, free sulfur dioxide and density variables.

## Decision Tree

Using all variables except “quality” to conduct a decision tree with entropy method when CP is a default value, the accuracy rate of the decision tree in the validation dataset is 0.8719.

After conducting cross-validation procedure, by right, the unseen data error should be the lowest as CP value = 0.0089928. However, when we apply the CP value in the decision tree, there is no change in accuracy rate both in training dataset and validation dataset.

We also conduct randomforest tree procedure to see importance level of each variable, the result shows below. The accuracy rate （0.8875） is a little bit higher than the normal decision tree.

# Regression Analysis

## Linear Regression

Using the linear regression model to find the correlation between attributes and quality.

Quality = β0 + β1\*fixed.acidity + β2\*volatile.acidity + β3\*citric.acid + β4\*residual.sugar + β5\*chlorides + β6\* free.sulfur.dioxide + β7\*total.sulfur.dioxide + β8\*density + β9\*pH + β10\*sulphates + β11\* alcohol

From the above regression, we can observe volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, pH, sulphates and alcohol are significantly correlated with dependent variables at a 10% significant level. Among them, residual sugar, sulphates and alcohol are positively correlated, which means as the increase of these values, Quality increases, while the other variables are negatively correlated. and for this model, adjusted R-squared is 0.3877, which means about 38% of quality value is explained by independent variables.

## 

## Logistic Regression

Using a logistic regression model to classify the data set using the same outcome variable, ’goodquality’ after considering all the variables at hand. We are trying to predict the outcome with best possible accuracy to find a good fit model for finding the attributes that contribute most to the quality of wine.

After running the regression model, we find that the log odds of goodquality wine is increased by 5.24 for each one-unit change in sulphates and its p-value indicates that it is somewhat significant in determining the quality. Similar effects of increasing on the log odds of good quality wine are interpreted for a unit-change in alcohol, density, free sulfur dioxide, and residual sugar.

Then, on performing backward, forward and stepwise selection process on the variables, we came up with a final model which has attributes like alcohol, volatile.acidity, sulphates, chlorides, total.sulfur.dioxide and free.sulfur.dioxide that affects the response variable more efficiently to come up with good enough result.

On comparing the performance of the logit model on the training and validation data sets by creating confusion matrices, we find the training data set has better accuracy than the validation data set. This can also be seen from the ROC curves plotted on both training and validation data sets which shows that the logit model using training data is better than the logit model using validation data.

# FINDINGS AND RECOMMENDATIONS

## Findings

1. The dataset includes three different kinds of acidity: fixed acidity, volatile acidity, and citric acid. According to the variable explanation, volatile acidity provides wine the unpleasant vinegar taste.

Heat map of correlation tells the same story that volatile acidity has the strongest correlation with wine quality. Meanwhile, three kinds of acidity correlated with each other effectively. As winery want to make a good wine, it is important to control the level of volatile acidity by balancing the two other acidities to obtain a pleasant taste.

1. We have 12 different attributes that contribute to the quality of wine in one or the other way, but the attributes that play an important role in predicting the quality of wine are total.sulfur.oxide, free.sulfur.dioxide, sulphates, alcohol, chlorides and volatile.acidity. Their concentration level in the wine varies significantly.
2. Free.sulfur.dioxide and total.sulfur.dioxide are associated with the concentration of SO2 in the wine where they are to be present in high and low levels respectively in wine to produce a good quality wine.
3. Alcohol and quality are strongly related as higher the concentration of alcohol content in wine, higher would be the quality of wine. This is also the most important variable that affects the quality of wine based on classification analysis.
4. The quantity of sulphates in the wine can be increased as it acts as an antimicrobial and antioxidant, which in turn increase the life of wine.
5. The chlorides as well volatile.acidity content can be decreased more to produce a better wine as the wine tastes much better then.
6. The increase in sugar content in the wines could also be a positive factor in producing good quality wines to consumers.
7. Heat map of correlation tells the same story that volatile acidity has the strongest correlation with wine quality. Meanwhile, three kinds of acidity correlated with each other effectively. As winery want to make a good wine, it is important to control the level of volatile acidity by balancing the two other acidity to obtain a pleasant taste.
8. For attracting more customers, the citric.acid levels can be increased to some extent so that different flavors wines can be produced and would increase market value. But this should not be compromised with the quality of wine.

## Recommendations

A random red wine has the large chance to reach a decent quality if it meets the following standards under this particular physicochemical test.

* Alcohol should be higher than 10. Similar to most vintage wine’s alcohol level, a red wine should keep alcohol level higher than 10.
* Sulphates should be higher than 0.8, which leverage wine's antioxidant, keep the fragrance last long and the good body of wine.
* Fixed acid should be between 6.6 and 6.9, providing the initial acids that balances the sweetness, alcohol and bitterness of wine.
* pH value should not exceed 3.3.
* Volatile.acidity should not exceed 0.43. It will affect the acid balanceness.
* Sugar level should be higher than 2.3. A good wine with tiny sweetness helps release of wine fragrance.

# APPENDIX

## Variables Descriptions

1. fixed acidity: most acids involved with wine or fixed or non-volatile (do not evaporate readily)

2. volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3. citric acid: found in small quantities, citric acid can add 'freshness' and flavour to wines

4. residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/litre and wines with greater than 45 grams/litre are considered sweet

5. chlorides: the amount of salt in the wine

6. 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

7. total sulfur dioxide: amount of free and bound forms of SO2; 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

8. density: the density of water is close to that of water depending on the percent alcohol and sugar content

9. 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

10. sulphates: a wine additive which can contribute to sulphur dioxide gas (SO2) levels, which acts as an antimicrobial and antioxidant

11. alcohol: the percent alcohol content of the wine

12. Output variable (based on sensory data): quality (score between 0 and 10)

## Figures

### Data ProcessingA screenshot of a computer Description automatically generated with low confidence

Before oversampling

A screenshot of a computer

Description automatically generated with low confidence

After Oversampling

Table

Description automatically generated

New Column for Quality

### DATA VISUALIZATION AND EXPLORATION - Histogram

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

### DATA MINING - Quality Clustering

A picture containing chart

Description automatically generated

Quality Clustering

A picture containing text, receipt

Description automatically generated

Decision Tree Analysis Result

Map

Description automatically generated

Decision Tree

Chart, scatter chart

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

Random Forest tree Analysis Result

### Regression Analysis

**Table

Description automatically generated**

Result of Linear Regression

**Table

Description automatically generated with low confidence**

Result of Logistic Regression

**Graphical user interface, text, application

Description automatically generated**

Stepwise Logistic Regression

**Text

Description automatically generated**

Confusion Matrix and Statistics of Training Dataset

Chart, line chart

Description automatically generated

The ROC Curve of Training Dataset

**Text

Description automatically generated with medium confidence**

Confusion Matrix and Statistics of Validation Dataset

Chart, line chart

Description automatically generated

The ROC Curve of Validation Dataset

## Codes

**Data exploration and visualization**

winequality.df <- read.csv("winequality.csv")

dim(winequality.df)

head(winequality.df)

View(winequality.df)

summary(winequality.df)

hist(winequalitytrain.df$fixed.acidity, main = 'fa.Dis', xlab = 'fixed.acidity')

hist(winequalitytrain.df$volatile.acidity, main = 'va.Dis', xlab = 'volatile.acidity ')

hist(winequalitytrain.df$citric.acid, main = 'ca.Dis', xlab = 'citric.acid')

hist(winequalitytrain.df$residual.sugar, main = 'rs.Dis', xlab = 'residual.sugar')

hist(winequalitytrain.df$chlorides, main = 'chlorides.Dis', xlab = 'chlorides')

hist(winequalitytrain.df$free.sulfur.dioxide, main = 'freesdio.Dis', xlab = 'free.sulfur.dioxide')

hist(winequalitytrain.df$total.sulfur.dioxide, main = 'totalsdio.Dis', xlab = 'total.sulfur.dioxide')

hist(winequalitytrain.df$density, main = 'density.Dis', xlab = 'density')

hist(winequalitytrain.df$pH, main = 'pH.Dis', xlab = 'pH')

hist(winequalitytrain.df$sulphates, main = 'sulphates.Dis', xlab = 'sulphates')

hist(winequalitytrain.df$alcohol, main = 'alcohol.Dis', xlab = 'alcohol')

hist(winequalitytrain.df$quality, main = 'quality.Dis', xlab = 'quality')

boxplot(winequality.df$fixed.acidity)

stack.df <- stack(winequality.df)

head(stack.df)

boxplot(stack.df$values ~ stack.df$ind,

col = rainbow(ncol(winequality.df)),, ylim = c(0, 150))

#randomly select 40% obs of all dataset as training data set.

set.seed(1)

winequalitytrain.index <- sample(c(1:dim(winegoodqualitytable.df)[1]), dim(winegoodqualitytable.df)[1]\*0.6)

winequalitytrain.index <- sample(c(1:dim(winequality.df)[1]), dim(winequality.df)[1]\*0.4, prob =ifelse((winequality.df$fixed.acidity<9.2)&(winequality.df$volatile.acidity<0.64)&(winequality.df$residual.sugar<2.6)&(winequality.df$chlorides<0.09)&(winequality.df$free.sulfur.dioxide<21)&(winequality.df$total.sulfur.dioxide<62)&(winequality.df$sulphates<0.73),0.9, 0.01))

winequalitytrain.df <- winegoodqualitytable.df[winequalitytrain.index, ]

winequalityvalid.df <- winegoodqualitytable.df[-winequalitytrain.index, ]

**heatmap:**

## heatmap with values

library(gplots)

heatmap.2(cor(winequality.df), Rowv = FALSE, Colv = FALSE, dendrogram = "none",

cellnote = round(cor(winequality.df),2),

notecol = "black", key = FALSE, trace = 'none', margins = c(10,10))

**Decision tree :**

wq.ct <- rpart(goodquality ~ ., data = winequalitytrain.df ,method = "class", minsplit = 1, xval = 5)

prp(wq.ct, type = 1, extra = 2, under = TRUE, split.font = 1, varlen = -10)

length(wq.ct$frame$var[wq.ct$frame$var == "<leaf>"])

#classification tree with entropy:

wq.info.ct <- rpart(goodquality ~ ., data = winequalitytrain.df, parms = list(split = 'information'), cp = 0.0089928, method = "class")

prp(wq.info.ct, type = 1, extra = 2, under = TRUE, split.font = 1, varlen = -10)

length(wq.info.ct$frame$var[wq.info.ct$frame$var == "<leaf>"])

#validate process: predicit training dataset given the decision tree made before:

wq.pred.train <- predict(wq.info.ct,winequalitytrain.df,type = "class")

#using confusion matrix to compare training

confusionMatrix(wq.pred.train, as.factor(winequalitytrain.df$goodquality))

#predicit validation dataset given the decision tree made before:

wq.pred.vali <- predict(wq.info.ct,winequalityvalid.df,type = "class")

confusionMatrix(wq.pred.vali, as.factor(winequalityvalid.df$goodquality))

#repeat the code for the validation set. then deeper tree

set.seed(1)

wq.cv.ct <- rpart(goodquality ~ ., data = winequalityvalid.df, method = "class", parms = list(split = 'information'), cp = 0.0001, minsplit = 1, xval = 5) # minsplit is the minimum number of observations in a node for a split to be attempted. xval is number K of folds in a K-fold cross-validation.

printcp(wq.cv.ct) # Print out the cp table of cross-validation errors. The R-squared for a regression tree is 1 minus rel error. xerror (or relative cross-validation error where "x" stands for "cross") is a scaled version of overall average of the 5 out-of-sample errors across the 5 folds.

pruned.ct <- prune(wq.cv.ct, cp = 0.0089928)

## random forest

rf <- randomForest(as.factor(goodquality) ~ ., data = winequalitytrain.df, ntree = 500,

mtry = 4, nodesize = 5, importance = TRUE)

## variable importance plot

varImpPlot(rf, type = 1)

## confusion matrix

rf.pred <- predict(rf, winequalityvalid.df)

confusionMatrix(rf.pred, as.factor(winequalityvalid.df$goodquality))

# linear regression model

winequalitytrain.df <- winequality.df[winequalitytrain.index, ]

reg <- lm(quality ~ ., data = winequalitytrain.df)

summary(reg)

**Logical Regression :**

winequality.df <- read.csv("winequality.csv")

View(winequality.df)

winegoodqualitytable.df <- winequality.df

View(winegoodqualitytable.df)

# replace column name quality to goodquality

colnames(winegoodqualitytable.df)[12] <- c("goodquality")

#convert numeric values of response variable goodquality to 2 levels 0 & 1

winegoodqualitytable.df$goodquality <- ifelse(winegoodqualitytable.df$goodquality>=6,1,0)

# partition data

set.seed(2)

winequalitytrain.index <- sample(c(1:dim(winegoodqualitytable.df)[1]), dim(winegoodqualitytable.df)[1]\*0.6)

winequalitytrain.index <- sample(c(1:dim(winequality.df)[1]), dim(winequality.df)[1]\*0.4, prob =ifelse((winequality.df$fixed.acidity<9.2)&(winequality.df$volatile.acidity<0.64)&(winequality.df$residual.sugar<2.6)&(winequality.df$chlorides<0.09)&(winequality.df$free.sulfur.dioxide<21)&(winequality.df$total.sulfur.dioxide<62)&(winequality.df$sulphates<0.73),0.9, 0.01))

winequalitytrain.df <- winegoodqualitytable.df[winequalitytrain.index, ]

winequalityvalid.df <- winegoodqualitytable.df[-winequalitytrain.index, ]

options(scipen=999)

# run logistic regression

logit.reg <- glm(goodquality ~ ., data = winequalitytrain.df, family = "binomial")

summary(logit.reg)

full.logit.reg <- glm(goodquality ~ ., data = winequalitytrain.df, family = "binomial")

empty.logit.reg <- glm(goodquality ~ 1,data = winequalitytrain.df, family= "binomial")

summary(empty.logit.reg)

# run backward selection process

backwards = step(full.logit.reg)

summary(backwards)

# run forward selection process

forwards = step(empty.logit.reg,scope=list(lower=formula(empty.logit.reg),upper=formula(full.logit.reg)), direction="forward",trace=0)

formula(forwards)

# run stepwise selection process

stepwise = step(empty.logit.reg,scope=list(lower=formula(empty.logit.reg),upper=formula(full.logit.reg)), direction="both",trace=0)

formula(stepwise)

# use predict() with type = "response" to compute predicted probabilities.

training\_prediction = predict(logit.reg, winequalitytrain.df, type = "response")

training\_prediction = ifelse(training\_prediction > 0.5, 1, 0)

# generate confusion matrix for training data

confusionMatrix(as.factor(training\_prediction), as.factor(winequalitytrain.df$goodquality))

validation\_prediction = predict(logit.reg, winequalityvalid.df, type = "response")

validation\_prediction = ifelse(validation\_prediction > 0.5, 1, 0)

# generate confusion matrix for validation data

confusionMatrix(as.factor(validation\_prediction), as.factor(winequalityvalid.df$goodquality))

# evaluating the model considered

library(ROCR)

pred\_training = prediction(training\_prediction, winequalitytrain.df$goodquality)

perf\_training = performance(pred\_training, "tpr", "fpr")

plot(perf\_training, col = "blueviolet", lwd = 4)+title("ROC curve for logit model using training data")

abline(a = 0, b = 1, col = "firebrick2")

# compute auc

perf\_training.auc = performance(pred\_training, "auc")

perf\_training.auc@y.values

pred\_validation = prediction(validation\_prediction, winequalityvalid.df$goodquality)

perf\_validation = performance(pred\_validation, "tpr", "fpr")

plot(perf\_validation, col = "blueviolet", lwd = 4)+title("ROC curve for logit model using validation data")

abline(a = 0, b = 1, col = "firebrick2")

# compute auc

perf\_validation.auc = performance(pred\_validation, "auc")

perf\_validation.auc@y.values