**Data Mining Wine Quality**

**By**

**Ghaith Al Saifi**

**Rob Black**

**Caroline Gardner**

**Grace Howard**

**Marcus Spotanski**

**Table of Contents**  
SUMMARY..................................................................................................................................................3

DATASET ATTRIBUTES...........................................................................................................................3

METHODS USED........................................................................................................................................4

DATA TRANSFORMATION......................................................................................................................4

REGRESSION..............................................................................................................................................5

CODE USED................................................................................................................................................5

OUTPUT.......................................................................................................................................................9

DISCUSSION.............................................................................................................................................14

DECISION TREES.....................................................................................................................................16

CODE USED...............................................................................................................................................16

OUTPUT......................................................................................................................................................19

DISCUSSION..............................................................................................................................................22

MANAGERIAL IMPLEMENTATION......................................................................................................23

CONCLUSION............................................................................................................................................23

REFERENCES............................................................................................................................................24

**Summary**

For this project, our group decided on using a dataset containing a variety of characteristics about white and red wine. We wanted to investigate this data to see if there were any relationships between the wine characteristics and what type of wine it was (red or white). We wanted to look at if measurements of the characteristics could predict the type of wine used, as well as if the quality rating of wine differed between white and red.

Any relationships we discover could be beneficial to any vineyards or other companies that produce wine as connections between the wine characteristics could allow them to increase sales based on customer preference or make further inferences about how varying levels of ingredients can affect wine taste.

**Dataset Attributes**

Our wine dataset consisted of the 12 attributes listed below:

* Fixed acidity
* Volatile acidity
* Citric acid
* Residual sugar
* Chlorides
* Free sulfur dioxide
* Total sulfur dioxide
* Density
* pH
* Sulphates
* Alcohol
* Quality (score between 0 and 10)

The attributes above are objective tests except for the quality rating. Wine experts graded the wine quality between 0 (very bad) and 10 (excellent). The quality was determined by taking the median of at least three evaluations made by the wine experts. The two datasets (red & white) are from red and white variants of the Portuguese “Vinho Verde” wine. In accordance with privacy and logistic issues no data is available about grape types, wine brand or wine selling price.

**Methods Used**

Our dataset consists of 1599 rows of data about red wine and 4898 rows of data about white wine. The dataset analyses wine quality based on eleven variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. We used these variables to compare and analyze wine quality. This dataset was taken from the UCI Machine Learning Repository. We decided to analyze the data using regression and decision trees.

***Data Transformation***

All data used in this project from both the red and white wines was standardized by using the formula S = (x-min)/(max-min) where x is the datapoint being standardized. This process occurred in Microsoft Excel, then all the data was uploaded into R for each statistical calculation. For later classification calculations, an Outcome column was added to show each row of data’s type of wine (red or white).

**Regression**

Linear Regression is one of the most popular and commonly used methods of predictive analysis. The idea of this method is to determine if a relationship exists between an independent and dependent variable, and if it does, what kind of relationship. Linear Regression models are trained by using a training data set and a testing data set. The training data helps the model develop optimal weights to use which teaches the model how the variables affect each other. The testing data is then fed to the model and the model predicts the y values based on the given x. If the model is trained well, the line of best fit will encompass the majority of the given data points. When the regression is finished it will also give the user an r squared value. This value can fall between 0 and 1. The closer to 1, the stronger the linear relationship between the independent and dependent variables. Ideally when looking for a relationship between certain attributes, an r squared value closer to 1 is preferred.

***Code Used***

# Import the dataset, change this to your file path/correct sheet

whiteStd = read\_excel("CSCI 4890 Data Mining and Warehousing/Project/wine\_quality\_clean.xlsx",

sheet = "white-standardized")

redStd = read\_excel("CSCI 4890 Data Mining and Warehousing/Project/wine\_quality\_clean.xlsx",

sheet = "red-standardized")

# Linear Models

# Model 1: Quality vs Residual Sugar

model1\_white = lm(quality ~ `residual sugar`, data=whiteStd)

summary(model1\_white)

model1\_red = lm(quality ~ `residual sugar`, data=redStd)

summary(model1\_red)

# Shows that Residual Sugar is significant in White wines, but Reds when it

# comes to quality. This makes sense as White wines are sweeter than Reds.

# Model 2: Finding other significant variables for both Reds and Whites

model2\_white = lm(quality ~ `fixed acidity` + `volatile acidity` + `citric acid` +

`residual sugar` + chlorides + `free sulfur dioxide` +

`total sulfur dioxide` + density + pH + sulphates + alcohol,

data=whiteStd)

summary(model2\_white)

# Significant values include:

# High significance: Citric Acid, Residual Sugar, Density, PH, Sulfates, alcohol

# Medium significance: Fixed Acidity

model2\_red = lm(quality ~ `fixed acidity` + `volatile acidity` + `citric acid` +

`residual sugar` + chlorides + `free sulfur dioxide` +

`total sulfur dioxide` + density + pH + sulphates + alcohol,

data=redStd)

summary(model2\_red)

# Significant values include:

# High significance: Volatile Acidity, Chlorides, Total Sulfur Dioxide,

# Sulfates, Alcohol

# Low significance: Free sulfur Dioxide, pH

# Model 3: Only include significant variables for both white and red wines

model3\_white = lm(quality ~ `fixed acidity` + `citric acid` + `residual sugar` +

density + pH + sulphates + alcohol, data=whiteStd)

summary(model3\_white)

model3\_red = lm(quality ~ `volatile acidity` + chlorides +

`free sulfur dioxide` + `total sulfur dioxide` + pH +

sulphates + alcohol, data=redStd)

summary(model3\_red)

# Correlation values and graphs will show influence of each variable on Wine

# Quality for their respective types of wine.

# These graphs will show what higher rated wines should have ideally for each

# set of significant variable.

# Correlation Graphs

library(corrr)

library(knitr)

# White

whiteQualityCor = whiteStd %>%

correlate() %>%

focus(quality)

colnames(whiteQualityCor)[colnames(whiteQualityCor)=='term'] = 'variable'

# Table of Values

kable(whiteQualityCor[order(whiteQualityCor$quality),])

# Bar plot

whiteQualityCor %>%

mutate(variable = factor(variable, levels = variable[order(quality)])) %>%

ggplot(aes(x = variable, y = quality, fill=variable)) +

geom\_bar(stat = "identity") + ylab("Correlation with Quality") +

xlab("Variable") + ggtitle('Correlations to White Wine Quality') +

theme(axis.text.x=element\_blank(), axis.ticks.x=element\_blank())

# Red

redQualityCor = redStd %>%

correlate() %>%

focus(quality)

colnames(redQualityCor)[colnames(redQualityCor)=='term'] = 'variable'

# Table of Values

kable(redQualityCor[order(redQualityCor$quality),])

# Bar plot

redQualityCor %>%

mutate(variable = factor(variable, levels = variable[order(quality)])) %>%

ggplot(aes(x = variable, y = quality, fill=variable)) +

geom\_bar(stat = "identity") + ylab("Correlation with Quality") +

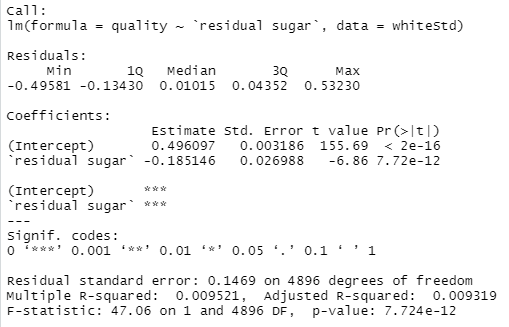
xlab("Variable") + ggtitle('Correlations to Red Wine Quality') +

theme(axis.text.x=element\_blank(), axis.ticks.x=element\_blank())

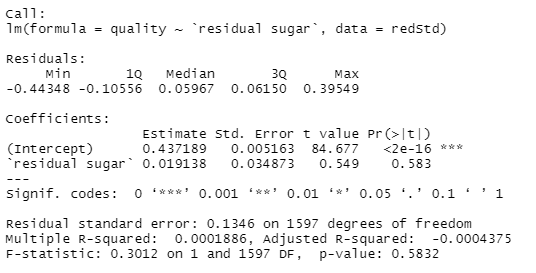
***Output***

**Model 1: Single Variable Regression Between Wine Quality and Residual Sugar**

**Summary of White:**

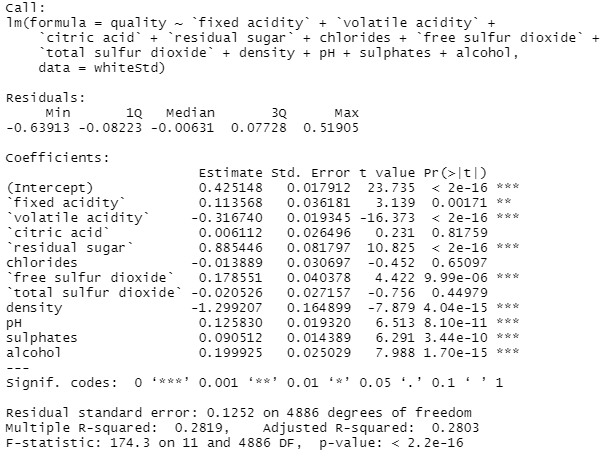


**Summary of Red:**

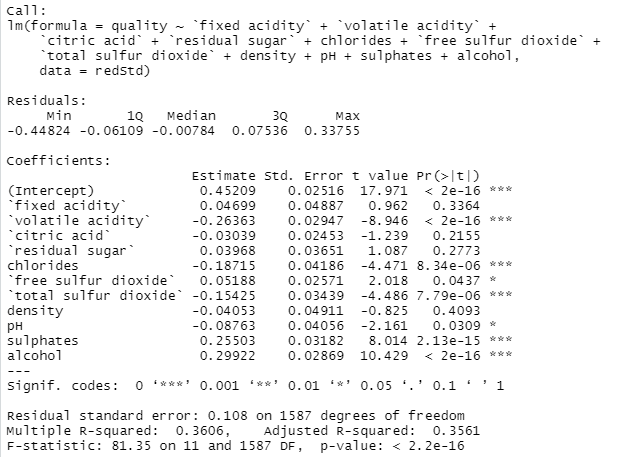


**Model 2: Multi-Variable Regression Between Wine Quality and All Variables**

**Summary of White:**

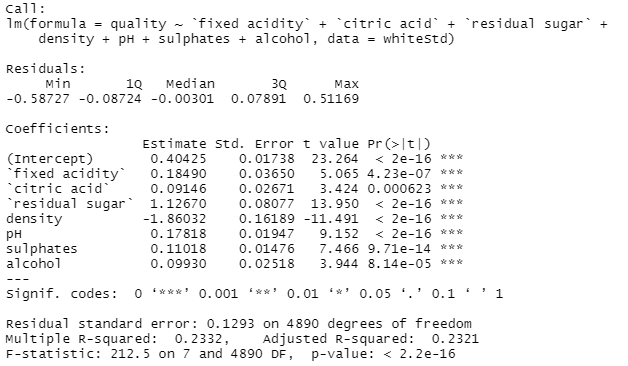


**Summary of Red:**

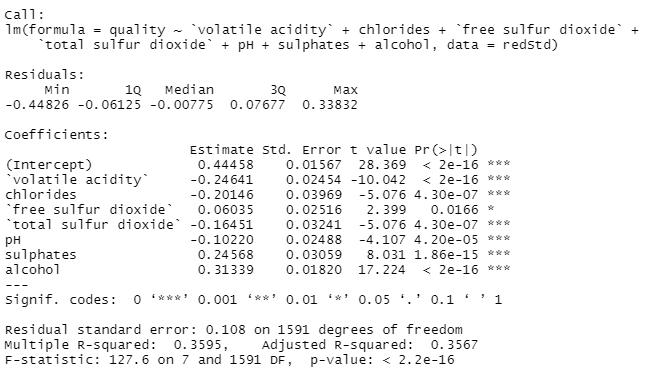


**Model 3: Multi-Variable Regression Between Wine Quality and Only Significant Variables**

**Summary of White:**

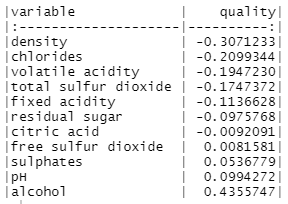


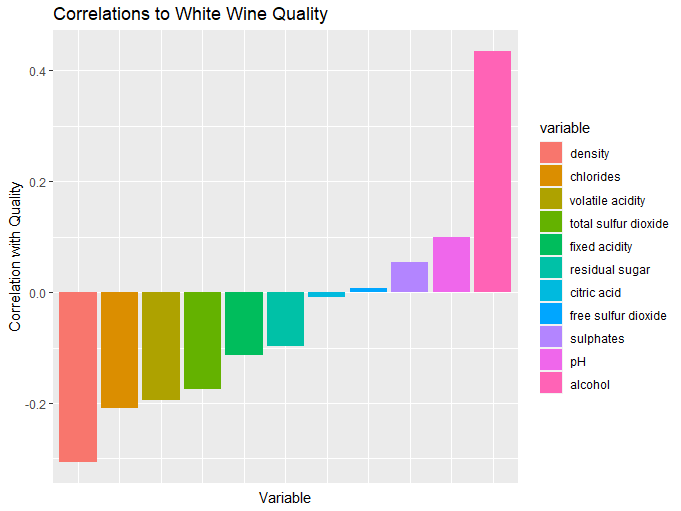
**Summary of Red:**



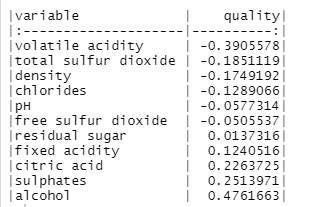
**Correlation Graphs and Tables of Wine Quality to Each Variable of Each Wine Type**

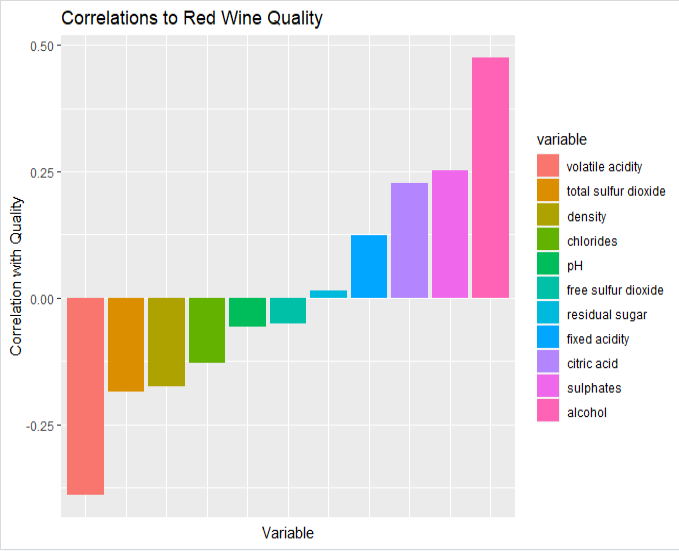
**White:**





**Red:**





***Discussion***

In Model 1 we compared wine quality versus residual sugars. We found that Red and White wines have a varying relationship with the same variable, giving some insight that this may be the case for the others as well. White wine did have a stronger correlation between these variables than red wine, however, red wine still showed this correlation.

Using Model 2 we found that the red and white wines share some significant variables that affect wine quality, and some that are strictly significant to only a specific wine type. Sulfates, and alcohol were highly significant in both forms of wine. In both types of wine, the higher the sulfates or alcohol content the higher quality a wine was assigned. Both red and white wine valued alcohol content the most, this variable had a positive correlation with quality, the higher the alcohol content, the higher the quality.

We also found that PH was highly significant in Whites, but was of low significance in Reds. In white wines a higher PH led to a higher quality, in reds PH was not a factor that affected the wine quality. Citric Acid, Residual Sugar, and Density were all highly significant in Whites, but not significant at all in Reds. These variables all had a negative correlation with quality. As the percentages of citric acid, residual sugar, and density increased the quality decreased. Of these variables density has the most significant effect on quality of white wines, while citric acid affected quality the least.

In our analysis of red wines, we found that Volatile Acidity, Chlorides, Total Sulfur Dioxide, and Free Sulfur Dioxide were all significant to Reds, but none were significant to Whites. All of these had a negative correlation, with volatile acidity being the most significant. All were highly significant except for Free Sulfur Dioxide which had the lowest negative correlation.

In Model 3 we included only the significant variables from Model 2 and showed that each of their levels of significance changed once the unsignificant variables from Model 2 were excluded. The correlation plots and graphs show a similar relationship to the significance levels to each of the respective wine types, indicating which variables could classify which types of wine. This gave us a visual representation of our findings from models 1 and 2 and further solidified our findings.

**Decision Tree**

Decision trees are a type of supervised machine learning used to categorize data or make predictions based on its training data. This model is tree-like in shape and shows several decisions and their possible outcomes. It is a useful algorithm when trying to understand the cause-and-effect relationships between data. Here, we look to determine key characteristics of each red and white wine. We will use 80% of the data set to train the algorithm and 20% to test it. In the end, we want an algorithm that can predict the type of wine based on the wine’s attributes.

***Code Used***

library(readxl)

library(data.table)

library(reshape2)

library(rpart)

library(rpart.plot)

library(dplyr)

library(caret)

set.seed(123)

whites <- read\_excel("wine\_quality\_clean.xlsx",

sheet = "white-standardized")

reds <- read\_excel("wine\_quality\_clean.xlsx",

sheet = "red-standardized")

reds$outcome = 'red'

whites$outcome = 'white'

wine = rbind.data.frame(reds, whites)

# Factorize wine dataset

wine$outcome = as.factor(wine$outcome)

# Create test and train data-sets

wine$id = 1:nrow(wine)

train = wine %>% sample\_frac(0.80)

test = wine %>% anti\_join(wine,train,by='id')

# Create Tree

# without density

wineTree = rpart(outcome~.-(id+density), data=train)

rpart.plot(wineTree)

# Prediction

pWine = predict(wineTree, train, type='class')

# Factorize colunms of predict and

confusionMatrix(data=pWine, train$outcome)

# With density

wineTreeDensity = rpart(outcome~.-id, data=train)

rpart.plot(wineTreeDensity)

# Prediction

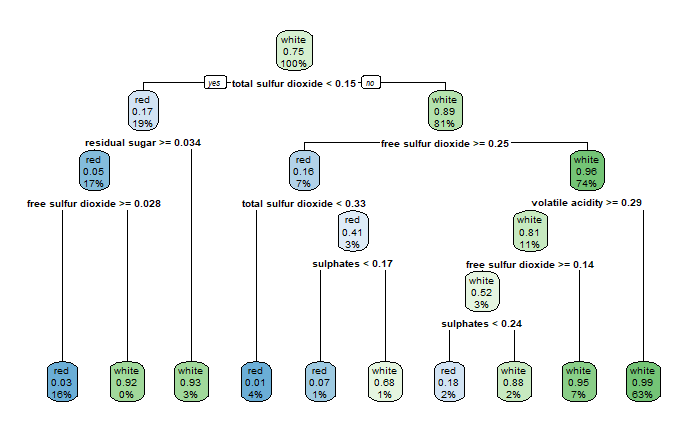
pWineDensity = predict(wineTreeDensity, train, type='class')

# Factorize colunms of predict and

confusionMatrix(data=pWineDensity, train$outcome)

***Output***

**Tree excluding Density**



**Confusion Matrix and Statistics**

**Reference**

**Prediction red white**

**red 1177 46**

**white 110 3865**

**Accuracy : 0.97**

**95% CI : (0.965, 0.9745)**

**No Information Rate : 0.7524**

**P-Value [Acc > NIR] : < 2.2e-16**

**Kappa : 0.9181**

**Mcnemar's Test P-Value : 4.558e-07**

**Sensitivity : 0.9145**

**Specificity : 0.9882**

**Pos Pred Value : 0.9624**

**Neg Pred Value : 0.9723**

**Prevalence : 0.2476**

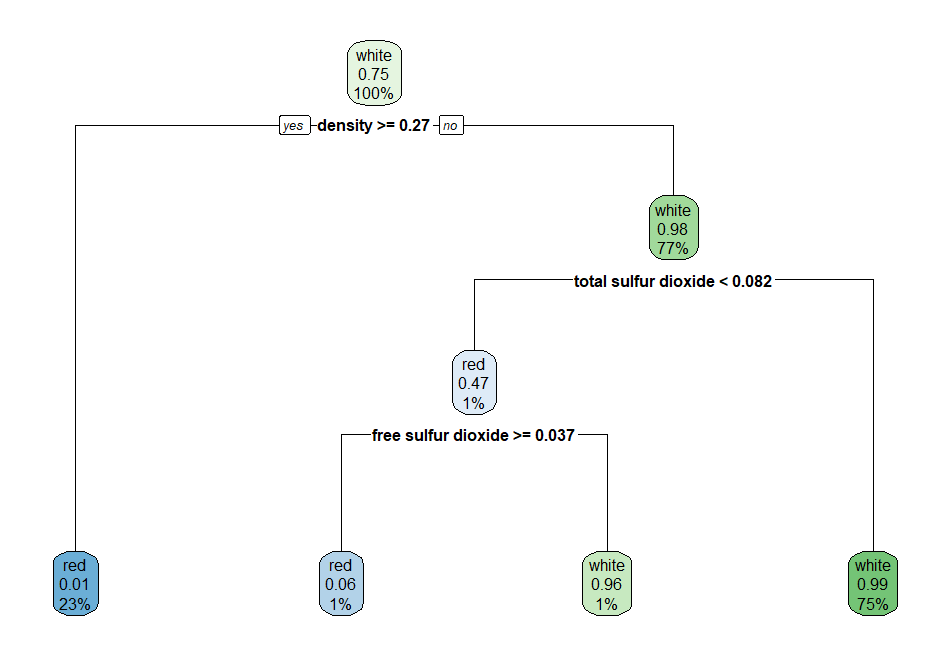
**Detection Rate : 0.2264**

**Detection Prevalence : 0.2353**

**Balanced Accuracy : 0.9514**

**'Positive' Class : red**

**Tree including Density**



***Confusion Matrix and Statistics***

***Reference***

***Prediction red white***

***red 1240 10***

***white 47 3901***

***Accuracy : 0.989***

***95% CI : (0.9858, 0.9917)***

***No Information Rate : 0.7524***

***P-Value [Acc > NIR] : < 2.2e-16***

***Kappa : 0.9703***

***Mcnemar's Test P-Value : 1.858e-06***

***Sensitivity : 0.9635***

***Specificity : 0.9974***

***Pos Pred Value : 0.9920***

***Neg Pred Value : 0.9881***

***Prevalence : 0.2476***

***Detection Rate : 0.2386***

***Detection Prevalence : 0.2405***

***Balanced Accuracy : 0.9805***

***'Positive' Class : red***

***Discussion***

In our initial findings, we found that density was a massive indicator of the type of wine. After changing the seed from both of our decision trees, we can determine that the three main indicators for the classification of each wine are its levels of density, total sulfur dioxide, and chlorides.

If we take a look at the first decision tree where we exclude the density from the algorithm, we will see that the first variable the algorithm checks for is the total sulfur dioxide. If the total sulfur dioxide is less than 0.15, there is a 16% chance that it is red wine and about 4% chance that it is white wine. In this case, the residual sugar and free sulfur dioxide variables will be used to determine the outcome. On the other hand, if the total sulfur dioxide is greater than 0.15, there is a 73% chance that it is a white wine and only 7% chance that it is a red wine. The variables free sulfur dioxide, volatile acidity, and sulphates will determine the outcome.

We can conclude that if the total sulfur dioxide is less than 0.15, it is four times more likely that it is a red wine, while it is about 10 times more likely that it is a form of white wine if the total sulfur dioxide is greater than 0.15. Although other variables are needed to accurately determine the outcome, total sulfur dioxide is a good starting point to split the data set in two.

A faster and more efficient way to determine the type of wine is by using density. Looking at the second decision tree graph, we can find that density is the determinant factor. If the wine’s density is less than 0.27, then it is a red wine. On the contrary, if the density is greater than 0.27, there is a 76% that it is a form of white wine and only a 1% chance that it a red wine. Comparing this decision tree to the first, the second has a height of 3, while the first’s height is five. Therefore, we can conclude that using density to determine the type of wine is faster and more efficient since there are less true-false expressions needed.

**Management Implementation**

Now that we can predict what type of wine to use and the quality rating of each type, how does this apply in the business world? To align themselves with customer preferences throughout their many market categories, wine producers must understand the primary influences on wine selection as well as the underlying motives. However, the identification of purchasing motivations is challenging due to the growing variety of local and international wine products and brands, as well as the growing disparity between wine types and costs. This has a significant impact on the capacity of wine producers to predict customer product preferences (Corduas and others, 2012, Introduction). Therefore, having those predictive algorithms will help wine producers predict customer behavior and act proactively.

**Conclusion**

All in all, we explored the different attributes in white and red wine. We used two algorithms that gave us two different insights, the regression model and the decision tree algorithms. The purpose of this research was to see how a vineyards or other companies that produce wine would make use of our research. Regression was used to study the relationships between the attributes, and we were able to point out five of those relations. On the other hand, decision trees were used to predict outcomes. After training and testing the algorithm, we found that the biggest attributes separating white and red wines is the density of the wine. Finally, those results can be used to predict customer purchasing behavior and act proactively.

**References**

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.

M. Corduas, L. Cinquanta, & C. Levoli. The Importance of wine attribute for purchase decisions: A study of Italian consumers’ perception. *Science Direct.* <https://www.sciencedirect.com/science/article/abs/pii/S0950329312002236>