**A Statistical Study of Red Wine Quality**

By Team II

Xiaoguang Zeng , Di Wu, Yifeng Li, Lei Zhang, Xiaohan Yu, Yingbin Mei

# Introduction & Background

According to a report by Allied Market Research, the global red wine industry was valued at $182 billion in 2020 and is expected to grow to $278.5 billion by 2028.11 The global red wine industry is growing rapidly, and the demand for great tasting wines continues to increase. In this study, we will explore how different physicochemical properties affect wine quality using data mining methods. We found a dataset from the UCI Machine Learning Repository and analyzed it using supervised learning methods including multiple linear regression, tree-based classification, and random forest.1

# Research Question & Motivation

**Research question**

How do physicochemical properties affect the ranking of red wine quality?

## Motivation

Given the large market valuation of the growing red wine industry, the results of our study could have tremendous economic implications. Additionally, wine tasting is subjective and predictive models can assist wine-tasting experts in making consistent judgments. By conducting an in-depth statistical study, we hope to provide critical insights to help wineries make bettertasting wines and grow their revenues. Hence, we are motivated to explore how physicochemical properties affect the ranking of red wine quality.

# Dataset Description and Variable Introduction

## Data Set of Interest

The data set we would analyze is a red wine data set sourced from the UCI Machine Learning Repository.1 There are 2 data sets in this repository (one for white wine and another for red wine). We will focus on the red wine data set for this project.

This red wine dataset contains 12 columns and 1599 rows of data related to the

Portuguese *Vinho Verde* red wine, which is exported from the Northwest region of Portugal*.*1,2

Each row of data corresponds to one variant of the red wine sample from this region. There are 11 input variables, all of which are physicochemical data expressed in numerical values.1 These input variables were measured by physicochemical tests in the laboratory environment. The measured results of variables such as citric acid, chlorides, and sulfates show the concentrations of such chemicals in the wine.

**Input variables:**

## 1. pH (scale of 0 to 14)

Solutions with a pH higher than 7 are considered basic while solutions with a pH less than 7 are acidic. Red wines, in general, have a pH between 3.3 and 3.6.3 It is crucial to control the pH of wine because pH can impact the aroma and mouthfeel.4

## 2. Fixed acidity (g/dm3)

Fixed acidity refers to the concentration of acids that can be measured through a laboratory process called titration where acids are neutralized by the addition of basic solutions.5

## 3. Volatile acidity (g/dm3)

Volatile acidity, on the other hand, cannot be measured via titration, but rather must be measured through a process called steam distillation where acids can be evaporated from the wine.5 Acidity is important because excessive volatile acidity and fixed acidity can give rise to a sour taste in the wine while low acidity can make the wine too flat.5

## 4. Citric acid (g/dm3)

Citric acid plays an important role in the flavor of the wine because it can give a wine a

“fresh” flavor.6

## 5. Residual sugar (g/dm3)

Residual sugar is the amount of sugar that remains after the fermentation process, which can affect the sweetness of the wine.

## 6. Chlorides (g/dm3)

Chloride, which is an essential building block of salt, can affect the level of saltiness of the wine depending on its concentration.

## 7. Free sulfur dioxide (mg/dm3)

Free sulfur dioxide measures the concentration of sulfur dioxide molecules that are free to react with other chemicals to prevent oxidation of the wine.7

## 8. Total sulfur dioxide (mg/dm3)

Total sulfur dioxide refers to the concentration of free sulfur dioxide in addition to the ones that are already bound to other molecules.7 The concentration of both free and total sulfur dioxide is important because too high of a concentration can mask the wine's fruity aromas and give it a bitter flavor.7

## 9. Sulfates (g/dm3)

Sulfates are commonly added as preservatives in wines.8 It is critical to control the level of sulfates because low amounts of sulfates can “preserve the flavors,” while high amounts can adversely affect flavor and color.8

## 10. Alcohol (vol.%)

Alcohol content, measured in volume percentage, can increase the complexity of a wine’s flavor profile. Hence, it is also included as one of the input variables.9

## 11. Density (g/cm3)

Density measures the mass of the wine in grams per unit volume in cubic centimeters. Sweeter wine is typically associated with higher density. 10

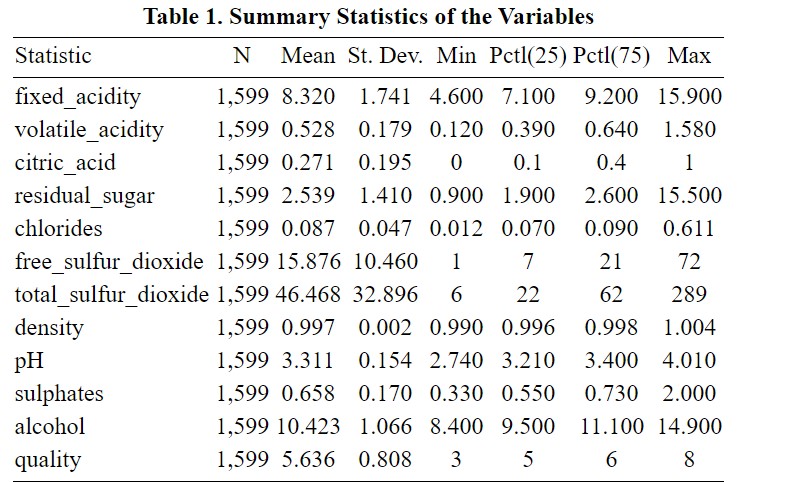
**Output variable:**

## 12. Quality (scale of 0 to 10)

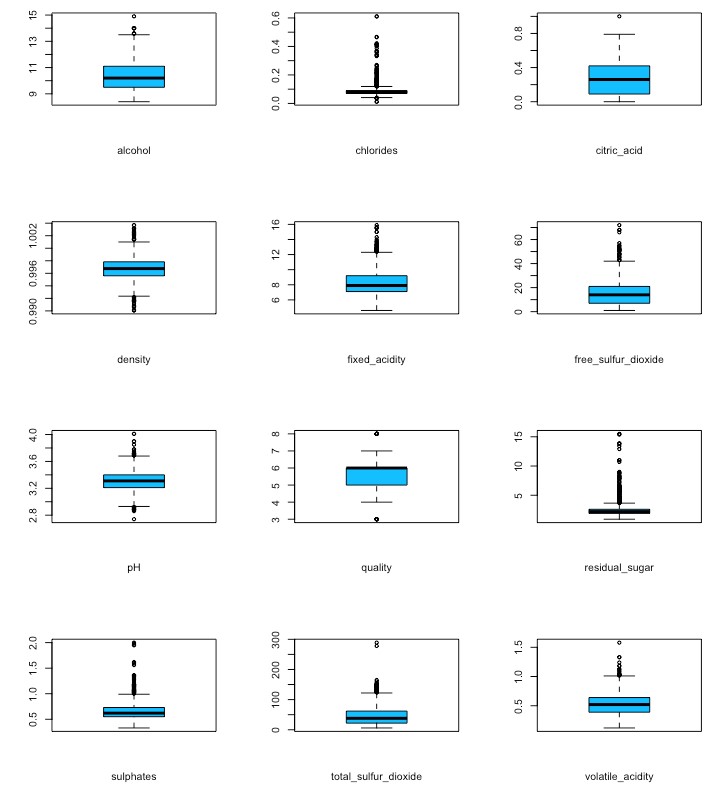
The output variable (overall quality of the wine) was measured by oenologists (wine tasting experts), based on a scoring scale of 0 (lowest quality) to 10 (highest quality). 2

# Data Summary Statistics

As shown in Table 1, summary statistics of the 12 variables and1599 observations are presented. Specifically, Table 1 details the mean, standard deviation, minimum, maximum, as well as the 25th percentile and 75th percentile values of each of the variables. It is important to note that the numeric score for quality ranges from 3 to 8 in our data set although the theoretical range is from 0 to 10.



To visualize the spread of the data of each variable, we created 12 box plots (Figure 1). Notably, chlorides and residual sugars have many data points that are outside of the interquartile range. But for variables such as quality, alcohol, and sulfates, most of the data points fall within the interquartile range.



**Figure 1. Box plots showing the quartiles and outliers of each variable.**

# Data Mining Method Description

Our goal is to build prediction models that will help predict the quality of the Portuguese *Vinho Verde* red wine using physicochemical variables. Our approach would be classified as supervised learning because the dependent variable is known. The 3 methods that we performed on our dataset are multiple linear regression, classification tree, and random forest.

Multiple linear regression was used because we want to make predictions on a continuous dependent variable (quality) based on a set of independent physicochemical variables. The Backward selection method was used to find the best linear model. Multiple linear regression is a great approach for our data set because looking at the coefficients of the regressors would allow us to easily interpret how a specific physicochemical variable affects red wine quality.

To supplement our analysis, we also used tree-based methods. While regression models can be easily influenced by outliers, tree-based methods are less sensitive to outliers. We decided to build a classification tree model because its results are easy to visualize. But before implementing the tree-based model, we categorized the dependent variable into 3 levels based on the following classification scheme: Low (3, 4), Medium (5, 6), and High (7, 8).12 We also randomly sampled half of the observations from the full data set to be the training data set and the other half to be the testing data set. Using cross-validation and plotting the error rate as a function of tree size, the optimal tree complexity was determined. Finally, the original tree was pruned using the optimal tree size. The performance of the pruned and unpruned trees was evaluated by comparing their respective accuracy rate.

To further improve model accuracy, we also tried building a random forest model. A random forest model can be more robust because it combines multiple tree models into a single model. Just like the classification tree, the random forest model was built using the training dataset, and the prediction was made on the testing data set. Again, the accuracy rate was computed to evaluate model performance. The importance of the variables was visualized to identify the physicochemical properties that affect red wine quality the most.

# Results and Model Performance

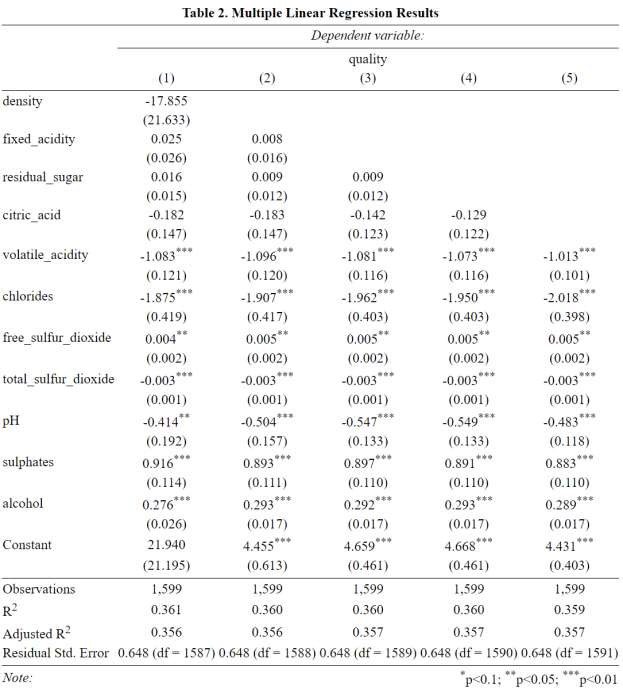
## Multiple Linear Regression

As shown in Figure 3, Linear Regression Model (1) is regressed on all of the independent variables. According to this model, several variables including density and fixed acidity are not statistically significant. To improve this model, we used backward selection to remove the variables that have the highest p-value one at a time until all the remaining variables are statistically significant.

Specifically, we removed density, fixed acidity, residual sugar, citric acid because they have large p-values which indicate they are not statistically significant in predicting the quality of red wine. Finally, we obtained Linear Regression Model (5), which is our best model, by regressing the dependent variable on volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH level, sulphates, and alcohol (Table 2). The coefficients on volatile acidity, chlorides, total sulfur dioxide, and pH level show that they have negative effects on red wine quality while sulphates and alcohol content have positive effects.

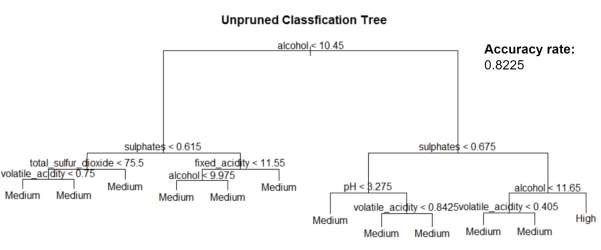
Specifically, the coefficient of -2.018 suggests that for each unit increase in chloride concentration, the quality of the red wine decreases by 2.018 points. On the other hand, for each additional unit increase in alcohol content (vol.%) and sulphates (g/dm3), the quality of the wine increases by 0.883 and 0.289 points respectively. Even though both free sulfur dioxide and total sulfur dioxide are statistically significant predictors, their effect on wine quality scores is negligible because their coefficients are close to 0 (Table 2).

Compared to Linear Regression Model (1), Linear Regression Model (5) shows a slightly larger adjusted R2 which indicates the later model fits better. In summary, because the adjusted R2 value is only 0.357, our final linear regression model only provides a moderate fit of our data.



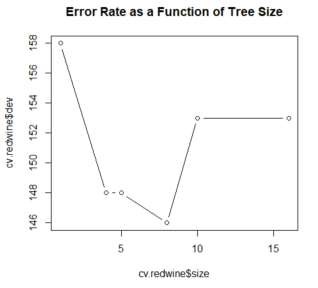
## Classification Tree

Figure 2 shows the unpruned tree that we obtained. The unpruned has 12 terminal nodes, and it is showing that alcohol and sulfates are acting as two of the most important variables in determining wine quality. This classification model is considered quite accurate because the accuracy rate is 0.8255, which is about 83% (Figure 2).



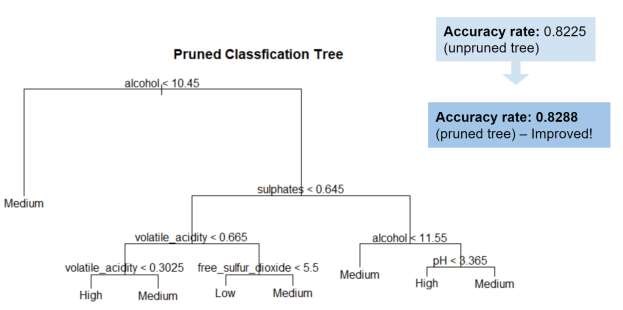
## Figure 2. Unpruned Classification Tree

Using cross-validation, the optimal tree complexity was determined. By plotting the error rate as a function of tree size, it is clear that the optimal tree size is 8 because it has the lowest classification error rate (Figure 3).



## Figure 3. Error Rate as a Function of Tree Size

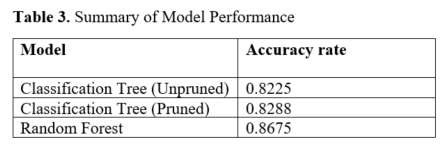
Based on the optimal tree size, we pruned the tree to have 8 terminal nodes. Results show that the 8-node tree has an improved accuracy rate. Like the unpruned tree, the pruned tree is also indicating that alcohol and sulfates play a key role in determining wine quality. The pruned tree suggests that if the alcohol content is less than 10.45 vol.%, it will be classified as medium. However, if it is more than 10.45 vol.%, we should follow the branch on the right side of the tree, and we will consider the sulphate concentrations to make further judgments about wine quality.

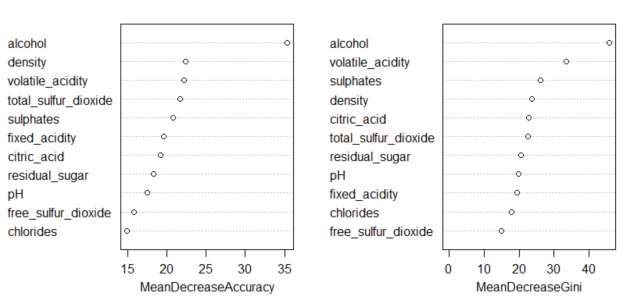


**Figure 4. Pruned Classification Tree**

## Random Forest

As shown in Table 3, the accuracy rate of the random forest reached 86.75%, which is higher than the accuracy rate from the previous classification tree models. This shows that the random forest model is indeed more robust than the previous classification tree models. The plot on the left ranks the importance of the variables based on the mean decrease in accuracy while the one on the right is based on the mean decrease in the Gini index. Overall, the two plots show that alcohol and volatile acidity have a heavy influence on red wine quality (Figure 5).





**Figure 5. Variable Importance Plots**

# Conclusions

The optimal Linear Model (5) is made up of 7 statistically significant variables (alcohol, sulphates, volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, and pH). They account for about 36% of the variations in quality (adjusted R2 of 0.357). Notably, the coefficients on volatile acidity, chlorides, and pH level suggest that they have negative effects on red wine quality while sulphates and alcohol content have positive effects. Since the adjusted R2 is only about 0.357, the multiple linear regression model only provides a moderate fit. For treebased methods, the random forest model is the optimal model in terms of accuracy rate (86.75%). Both tree-based models overall show that alcohol, sulphates, and volatile acidity are the most important variables that affect red wine quality.

Even though our tree-based models provide high predictive accuracy, the linear regression model could be improved. This could be potentially achieved by removing outliers from some of the variables in our data set. Also, it would be interesting to use the K-Nearest Neighbor (KNN) method and compare it to our tree-based methods.

# Practical Implications

Our project has important implications in the wine industry. In many countries, wine products go through a certification process to ensure wine quality and safety.2 As part of this process, oenologists are hired to evaluate wine quality.2 But sometimes, oenologists make vastly different quality assessments compared to their peers because subjective factors are involved. If the model of our study can accurately forecast wine quality based on physicochemical properties, it can be applied to the certification process to help oenologists make faster and more consistent assessments.2

Additionally, our prediction model can also potentially help wineries within make higher quality wines. For instance, by adjusting the concentration of important physicochemical properties such as alcohol content, sulphates, and volatile acidity during the production process, wineries can optimize wine quality. Besides, leveraging the insights from our prediction model, wineries can create new products to meet the specific needs of customers from niche markets.2 This will not only benefit wineries but also red wine consumers.

## References

1. Wine Quality Data Set. UCI Machine Learning Repository. Accessed October 20, 2021.

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

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

doi:10.1016/j.dss.2009.05.016

1. Improving the Taste and Color of Wine with pH Control. Sensorex. Accessed October 25, 2021. https://sensorex.com/blog/2017/12/06/ph-improve-taste-colorwine/#How\_the\_pH\_of\_wine\_affects\_taste\_and\_smell

1. The Importance of pH in Wine Making. Winemaker’s Academy. Published September 13, 2013. Accessed October 25, 2021. https://winemakersacademy.com/importance-ph-winemaking/

1. Understanding Wine Acidity. Winemaker’s Academy. Published September 19, 2013. Accessed October 25, 2021. https://winemakersacademy.com/understanding-wine-acidity/

1. Citric Acid. UC Davis Viticulture & Enology. Accessed October 25, 2021.

https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/commonchemical-reagents/citric-acid

1. Maureen M. Total Sulfur Dioxide – Why it Matters, Too! Iowa State University Extension and Outreach. Published February 27, 2018. Accessed October 25, 2021.

https://www.extension.iastate.edu/wine/total-sulfur-dioxide-why-it-matters-too/

1. Savvy W. Sulfites in Your Wine: To Savor or Shun? Wine Cooler Direct. Published November 24, 2014. Accessed October 25, 2021. https://learn.winecoolerdirect.com/winesulfites/

1. How Alcohol Content Affects Wine. ayza. Published August 25, 2014. Accessed October 25, 2021. <https://ayzanyc.com/blog/how-alcohol-content-affects-wine/>

1. Nguyen D. Red Wine Quality Prediction Using Regression Modeling and Machine Learning. Towards Data Science. Published November 22, 2020. Accessed October 26, 2021.

https://towardsdatascience.com/red-wine-quality-prediction-using-regression-modeling-andmachine-learning-7a3e2c3e1f46

1. Chauhan S, Deshmuk R. Red Wine Market by Type (Cabernet Sauvignon Wine, Merlot

Wine, Syrah/Shiraz Wine, and Others), Pricing (Mass and Premium), and Distribution Channel (On-trade and Off-trade): Global Opportunity Analysis and Industry Forecast, 2021–2028. Alliedmarketresearch.com. Accessed October 23, 2021.

<https://www.alliedmarketresearch.com/red-wine-market-A13400>

1. Analysis of Wine Quality Data. Accessed October 26, 2021.

https://online.stat.psu.edu/stat508/book/export/html/804

1. Hlavac M. Published May 30, 2018. Accessed December 8, 2021.

https://cran.r-project.org/web/packages/stargazer/index.html.