**Predicting Red Wine Quality**

**Tyler S. Perry, Mehmet Ozmen, Fabio Nasseh**

**Introduction**

The wine industry has a vested interest in determining the quality of their wines in order to more efficiently price and market different red wine varieties. Typically, determining the quality or rating of wine is a process that involves panels of sommeliers and wine critics coming to consensus based on a number of personal preferences as well as objective factors **[1]**. We will attempt to determine how different variables provided to us in the dataset impact wine quality, and we will predict the quality of different Vinho Verde red wine varieties based on some or all of the variables[[1]](#footnote-0) provided to us in the dataset **[2]** using classification models and machine learning. This information will help producers and distributors of red wine assess their pricing and distribution strategy.

**Dataset**

The dataset we used was *Red Wine Quality* by UCI Machine Learning available for download from Kaggle. This dataset contains physicochemical and sensory data on 1,599 different red wine variants of the Portuguese “Vinho Verde” wine. We downloaded the data from Kaggle as a CSV file.

**Tools Used**

The CSV file was imported into Google Sheets for basic data cleaning. Then, we used the “CSV import” function in Orange3 to create data visualizations to better understand the data. We used Weka to preprocess the data and apply classification experiments.

**Data Acquisition**

The dataset is an Open Database, available for free on Kaggle. We downloaded the dataset as a CSV file and opened it in Google Sheets.

**Data Analysis and Results**

***Basic Characteristics of the Dataset***

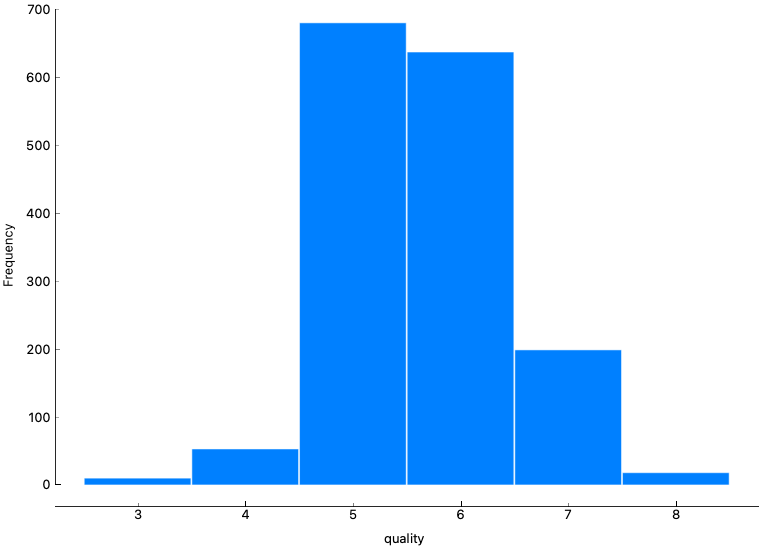
In order to gain a better understanding of the data, we used Orange3 to find the feature statics for each attribute provided in the dataset. The mean, median, minimum value, and maximum value for each attribute is listed below in Table 1.

Since our dependent variable is *quality*, we used Orange3 to run a distribution graph of the attribute. We found that six wine qualities are represented in the dataset, ranging from 3 to 8. The distribution is normal, with the quality of most wines falling in the middle. The distribution plot is shown below in Figure 1.

**Table 1.** Feature Statistics for all 12 attributes provided in the dataset

| **Attribute** | | **Mean** | **Median** | **Min** | **Max** |
| --- | --- | --- | --- | --- | --- |
| 1 | fixed acidity | 8.3196 | 7.9 | 4.6 | 15.9 |
| 2 | volatile acidity | 0.5278 | 0.52 | 0.12 | 1.58 |
| 3 | citric acid | 0.2710 | 0.26 | 0.00 | 1 |
| 4 | residual sugar | 2.5389 | 2.2 | 0.9 | 15.5 |
| 5 | chlorides | 0.0875 | 0.079 | 0.012 | 0.611 |
| 6 | free sulfur dioxide | 15.8749 | 14 | 1 | 72 |
| 7 | total sulfur dioxide | 46.4678 | 38 | 6 | 289 |
| 8 | density | 0.9967 | 0.9967 | 0.99 | 1 |
| 9 | pH | 3.3111 | 3.31 | 2.74 | 4.01 |
| 10 | sulphates | 0.658 | 0.62 | 0.33 | 2 |
| 11 | alcohol | 10.423 | 10.2 | 8.4 | 14.9 |
| 12 | quality | 5.64 | 6 | 3 | 8 |

**Figure 1.** Distribution of *wine quality* ***(***[***click to enlarge***](https://drive.google.com/file/d/10wloiQnP-pPiDfX6mZ7PqrLKsDSq0lny/view?usp=sharing)*)*



To determine the effect of each variable on quality, we used Orange3 to run a correlation analysis of our independent variables against the dependent variable, *quality*. The variables with the highest correlation were *alcohol*, *volatile acidity*, and *sulphates*. The variables with the lowest correlation were *pH*, *free sulfur dioxide,* and *residual sugar*. The r value of each variable against wine quality is seen below in Table 2; an explanation[[2]](#footnote-1) of each variable is shown in Table 3.

**Table 2.** Correlation of each attribute on *quality*.

| **X** | **Y** | **Correlation** | **Correlation Coefficient (r)** |
| --- | --- | --- | --- |
| quality | volatile acidity | negative correlation | -0.390 |
| quality | total sulfur dioxide | negative correlation | -0.185 |
| quality | density | negative correlation | -0.175 |
| quality | chlorides | negative correlation | -0.129 |
| quality | pH | negative correlation | -0.058 |
| quality | free sulfur dioxide | negative correlation | -0.051 |
| quality | residual sugar | positive correlation | 0.014 |
| quality | fixed acidity | positive correlation | 0.120 |
| quality | citric acid | positive correlation | 0.230 |
| quality | sulphates | positive correlation | 0.251 |
| quality | alcohol | positive correlation | 0.476 |

**Table 3.** Explanation of each attribute

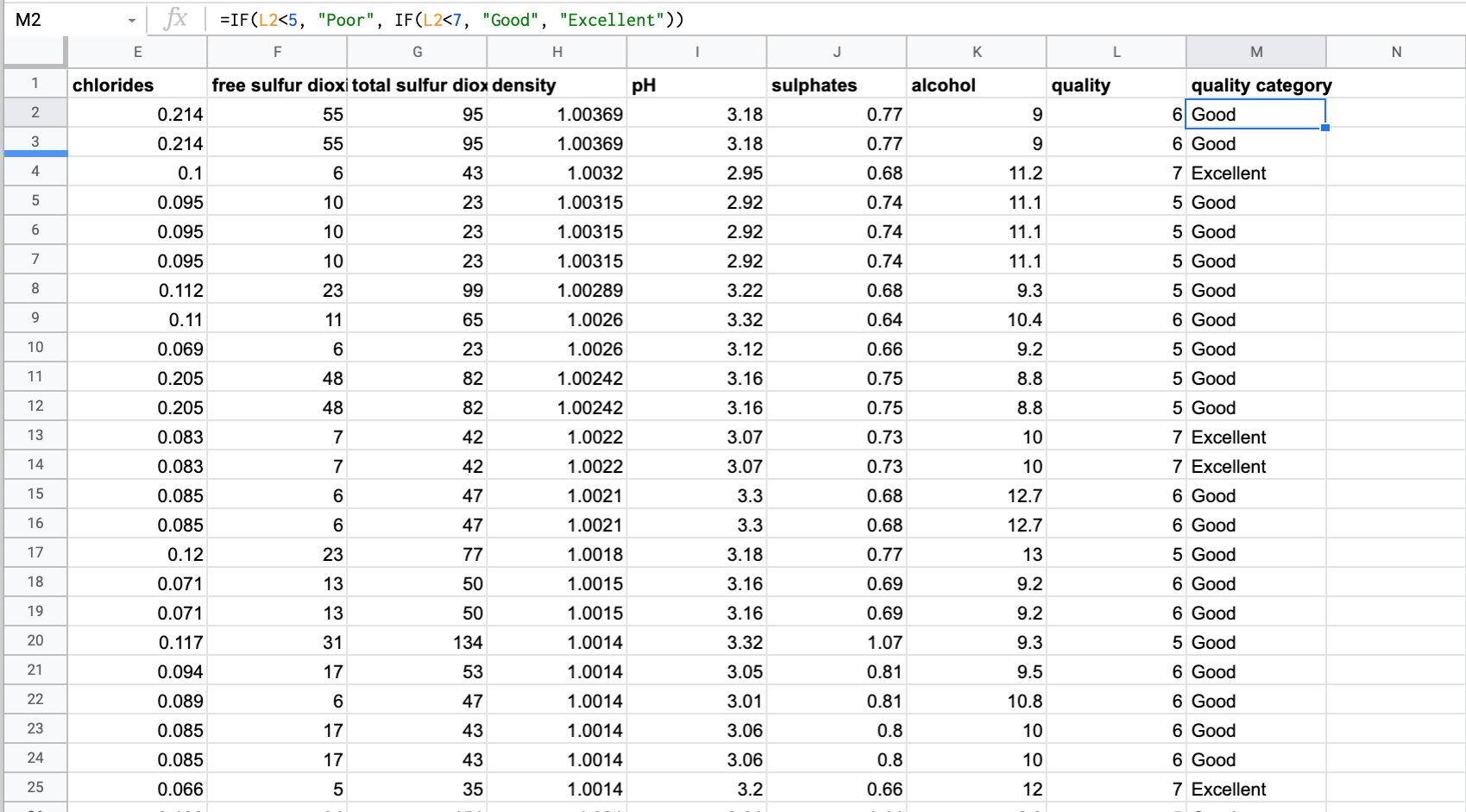
| **Volatile Acidity :** Acidic elements of a wine that are gaseous, rather than liquid, and therefore can be sensed as a smell. Volatile acids are produced through microbial action such as yeast fermentation, malolactic fermentation, and other fermentations carried out by spoilage organisms.  Amoun: **0.14 g/100 mL for red wine and 0.12 g/100 mL for white wines** |
| --- |
| **Total Sulfur Dioxide:** Preserves wine’s freshness and fruit characters by virtue of antioxidant, antimicrobial and anti-enzymatic properties.  Amount: **10 - 20 mg/L** |
| **Density:** Concentration of alcohol, sugar, glycerol, and other dissolved solids.  Amount: **1.080 - 1.090** |
| **Chlorides:** Amount of salt in the wine.  Amount: Max **606 mg/L** |
| **Ph:** Measure of the concentration of free hydrogen ions in solution. About 3.0 to 3.4 is desirable for white wines, while about 3.3 to 3.6 is best for reds.  Range: **3 - 4** |
| **Free Sulfur Dioxide:** The portion of SO2 that is free in the wine plus the portion that is bound to other chemicals in the wine such as sugar. It **prevents the wine from reacting with oxygen** which can cause browning and off-odors (oxidation), and it **inhibits the growth of bacteria and undesirable wild yeasts** in the grape juice and wine.  Amount: **25 mg/L on red wine and 30 mg /L on white wine** |
| **Residual Sugar:** The natural grape sugars left over in a wine after the alcoholic fermentation is complete.  Range: **0.2% - 10%** |
| **Fixed Acidity:** The combined sum of titratable and volatile acids present.  Range: **1,000 - 4,000 mg/L** |
| **Citric Acid:** Citric acid imparts a citric character that enhances the taste of many white and blush *wines*.  Range: **0.1 - 0.7 g/L** |
| **Sulfites:** Natural by-product of the fermentation process that work as a preservative against certain yeast and bacteria invasion. Sulfites are also added by the winemaker.  Range: Max **350ppm** |
| **Alcohol:** Unfortified wine is about **5.5% to 16%**, with an average of 11.6%. Fortified wines range from 15.5% to 25% ABV, with an average of 18%. |

***Data Preprocessing***

The dataset was already cleaned, contained no missing data, and no outliers. There was no need to apply additional data cleaning. Only discretization was performed, using Weka.

First, we created a new categorical attribute, *Quality Category*, using the *Wine Quality* attribute. We used an IF statement in Google Sheets to split the numerical *quality* values into three categories: *Poor*, *Good*, and *Excellent*. This process is shown below in Figure 2.

**Figure 2.** Creating a new categorical attribute, *wine quality.*



*We used an IF statement in google sheets to separate each red wine instance into a quality category based on its quality score. Wines scored at less than 5 were classified as “poor”, wines scored between 4 and 6 were classified as “good”, and wines scored higher than 6 were classified as “excellent.”*

After the csv file was imported into Weka, the nominal *quality* category was removed. Then, the remaining numerical variables were automatically discretized from a range of numeric attributes into nominal by using the first-last method with a precision of 10 in Weka.

***Data Modeling and Evaluation***

Three classification models, Random Forest, Random Tree, and Naïve Bayes were applied to the dataset. In Weka, the datasets are separated into training and testing sets by using 10-fold cross validation.

We applied the models first with all attributes, and then after removing irrelevant attributes (residual sugar, free sulfur dioxide, and pH**)** and compared the results. The results are shown below in Table 4.

**Table 4.** Classification Model Results

| Classification Model | | Accuracy | Precision | Recall | F1 |
| --- | --- | --- | --- | --- | --- |
| 1 | Random Forest, all attributes | 82.364% | 0.794 | 0.824 | 0.806 |
| 2 | Random Forest, only relevant attributes | **82.6767%** | 0.794 | **0.827** | 0.807 |
| 3 | Naive Bayes, all attributes | 82.0513% | **0.824** | 0.821 | 0.820 |
| 4 | Naive Bayes, only relevant attributes | 82.6141% | 0.827 | 0.826 | **0.824** |
| 5 | Random Tree, all attributes | 81.8637% | 0.790 | 0.819 | 0.802 |
| 6 | Random Tree, only relevant attributes | 82.4891% | 0.790 | 0.825 | 0.801 |

*We applied Random Forest, Naive Bayes, and Random Tree classification models on the data. First, we applied the models using all attributes, and then we removed residual sugar, free sulfur dioxide, and pH, since those attributes were irrelevant.*

**Discussion**

At the beginning of the project, we set out to accomplish two tasks: 1) determine which variables had the strongest impact on wine quality, and 2) develop a classification model to accurately predict the quality of different Vinho Verde red wine varieties using some of all of the variables provided to us in the dataset.

To accomplish the first task, we conducted a correlation analysis to calculate the relationship between all attributes and quality. We ranked all attributes based on their correlation coefficient. We determined that the three variables that had the largest impact on wine quality were *alcohol*, *volatile acidity*, and *sulphates.* We deemed three variables to be irrelevant due to their low correlation coefficient; those variables were *pH, free sulfur dioxide, and residual sugar*.

To accomplish the second task, three classification models were applied to the data: random tree, naïve bayes, and random forest. The accuracy of a classification experiment is measured by the percent of correctly identified instances. All models performed better after removing irrelevant attributes. We defined the three irrelevant attributes as the three attributes with the lowest correlation to quality. We found that after removing irrelevant attributes, the Random Forest model (Model 2) classified the instances into their appropriate class 82.68% of the time. By comparison, the Naive Bayes model (Model 4) classified the instances correctly 82.61% of the time, and the Random Tree model (Model 6) was correct 82.49% of the time.

However, accuracy isn’t always the best measure to use when determining the effectiveness of a model. In cases like ours, where class distribution is uneven, the F1 score is often more useful. The F1 score is the weighted average of precision and recall, and takes both false positives and false negatives into account. The Naive Bayes model (Model 4) scored the highest with a F-measure of 82.4% after irrelevant attributes were removed.

**Timeline for Completion**

**Week 1-3**

The first few weeks we spent familiarizing ourselves with the dataset. We read more in depth about the wine data set and the description of each property of the input variables. We created an account on Kaggle and explored Orange 3 and Weka. After that, we downloaded the dataset and began creating visualizations of the data.

**Week 4-6**

The next few weeks were spent on data preprocessing. The wine quality category attribute was created, and discretization was performed. The group met to discuss the correlations of each attribute on wine quality.

**Week 7-9**

After we got a better understanding of the data, we spent the next few weeks uncovering intermediate results and preparing for the midterm presentation. Many classification models were applied, and we settled on Random Forest and Naive Bayes to include in the final report. The group met on Zoom to put together our midterm presentation.

**Week 10-12**

After receiving instructor’s feedback on the midterm presentation, the group met again to retest the dataset after removing irrelevant attributes. Our final classification results were collected, and we began compiling the final report.

**Week 13**

The final report was completed, and the midterm presentation was updated to reflect new data.

**Team Workload and Roles**

Our team utilized a text message group chat to communicate throughout the week. We also utilized Zoom to meet synchronously and break down the worldload for each week. All work is done through a shared google drive folder, so that we are all working on one version of each file.

Although we all contributed wherever was needed; Fabio took lead on the weekly report writing, as well as a portion of the Data Analytics section, Tyler was assigned the task of data preprocessing/data analysis and drafting the final report, and Mehmet took on most of the Weka modeling.

**References**

1. Puckette, M. (2014, August, 18). *A Pragmatic Approach to Using Wine Ratings.* Wine Folly. <https://winefolly.com/tips/wine-ratings-explained/>
2. 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.

1. The variables provided to us in the dataset are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and wine quality. [↑](#footnote-ref-0)
2. *We researched different properties of wine provided in the dataset. We started by researching the input variables. These variables each need to be in a specific range to make up a good wine, therefore accurate measurements are required to be done in factory to ensure the quality of wine.*  [↑](#footnote-ref-1)