Machine Learning Project – Red Wine Quality Prediction

Kurtis Shane Wilhoite

Table of Contents

**No table of contents entries found.**

# Introduction

## Explanation of attributes

Wine, or in this case, specifically Red Portuguese “Vinho Verde” wine, contains many different attributes which may or may not affect the overall quality and taste. The dataset used contains the following attributes (and their meanings, for future reference):

1. **Fixed Acidity** – Consisting of the tartaric, malic, citric, and succinic acids in total. These acids, if lacking, can produce a “flat” taste to the wine.
2. **Volatile Acidity** – The total acidity of the wine. At too high levels, this can have a vinegar-like quality to the wine, too low and it dulls the wine’s taste.
3. **Citric Acid** – Part of the Fixed Acidity, found in small quantities typically. This can add a fresh taste to the wine, and a sourness if over used.
4. **Residual Sugar** – The amount of sugar remaining after fermentation. Most wines contain at least 1g/l. 45g/l and above is considered a sweeter wine.
5. **Chlorides** – Chloric intake, aka the salination or salt content of the wine. Too much can take away from the wine’s inherit sweetness, too little and it may dull the sweetness of the wine.
6. **Free Sulfur Dioxide** – The amount of free moving SO2. FSO2 can help prevent spoilage and oxidation in small amounts, but larger amounts can start to produce a bitter/chemical flavor and reduce the affects of “breathing” the wine.
7. **Total Sulfur Dioxide** – The total of SO2, Free moving and bound. Assists in spoilage prevention, in small amounts, it’s largely undetectable and helps shelf-life, in large amounts, it can give off a negative taste and scent.
8. **Density** – This is the consistency of the wine. A higher density creates a wine which is more viscous, or thicker, than that of water. This can affect the texture (or “mouth-feel”) and appetizing feeling of the wine if too high or too low.
9. **pH** – Describes the overall acidity of the wine. Scaled from 0 to 14 (High acidity to no acidity), Water is measured at a 7, wine typically ranges between 3-4. High acidity may taste tarter and crisper, low acidity may produce a smoother and rounded out taste.
10. **Sulphates** – A chemical additive which helps contribute to SO2 levels and increase shelf-life
11. **Alcohol** – Alcoholic content of the wine, can affect the flavor if too strong compared to sweetness. However, can affect flavor negatively if not present to a certain level as well, tasting more of a juice than of a wine.

These are being compared against the final attribute: Quality. The quality has been determined between 0 and 10 by the median evaluations of at least 3 wine experts. While taste is subjective, this was the best way to quantify it for analysis.

## Goal

Some of these qualities heavily affect the wine, and as such, it’d be ideal to be able to estimate each quantity and see what quality may be produced and rated by those seeking to produce a high-quality wine, such is the goal of this project, to create the ability to insert each level independently and have a predicted outcome of the wine flavor, as well as to attempt to produce a wine with quantities that fit an idealized wine. As well as analyzing how these attributes affect one another.

# Analysis and Visualization

## Base Data

Firstly, a basic average of all attributes should be collected, and an analysis of how much information is available is done. We’re given a total number of entries as “1599” different wines and their attributes.

Graphical user interface

Description automatically generated with medium confidence

Figure - Dataset Measurements

By using the describe function (*Figure 1*), we can find some interesting information about the dataset. The first being that the count is universal across the board and is the same as the number of columns, which indicates there are no missing columns at any point, so all wines have their appropriate entry.

There are several interesting points within this table, such as the fact that the average quality is 5.64, barely over half of the full possible score, meaning most wines lay somewhere in the middle of the possible scoring, with the maximum being 8 and minimum being 3. So, no wines were considered 0 across the average between wine experts, nor were any considered perfect, or even a nearly perfect score, of 9.

In addition to these things, the alcoholic content sitting around 10.4 on average, with a max of 14.9 and minimum of 8.4 shows that none of the wines are within the fortified wine range of 15.5% - 25% and are below the overall average of 11.6% (MasterClass Staff, 2021). As well, the residual sugars indicate that the wines are all Dry wines as well, containing less sugar content (Puckette, 2019).

## Correlations

When we compare various aspects on plots vs their quality counterparts, we can start to see correlations between the two different aspects and their reflection on the wine. Some aspects have a negative correlation, some have a positive correlation, and some have a neutral correlation, and their percentages largely don’t affect the quality of the wine.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure - Volatile Acidity vs Quality Figure - Citric Acid vs Quality

If we compare 2 of the 3 levels of acidity within the dataset, we can see a strong correlation between the quality of the wine and the respective levels. When Volatile acidity, which can cause a vinegar-like taste at high levels, is at a lower level, the quality tends to be higher (*Figure 2*). And on the other hand, when the citric acid is higher, it can lend a fresher taste to the wine, and thus, a tendency towards higher quality (*Figure 3*).

Chart, bar chart

Description automatically generated

Figure - Fixed Acidity vs Quality

The same cannot be said when comparing Fixed acidity to Quality (*Figure 4*) which gives us a rather neutral correlation. This is a fair assessment, as “Fixed acidity” can really refer to multiple different types of acidity within the wine, so not all attributes have a strong correlation to the quality of the wine.

Chart, treemap chart

Description automatically generated

Figure - Correlation Heatmap

For an in-depth analysis of all correlations, a heatmap gives a clear indication of how each attribute can affect each other. A correlation of 0 means there is no correlation, and positive numbers and negative numbers indicate a positive and negative correlation respectively. From this, we can also tell roughly what attributes can heavily affect the quality of the wine, with Alcohol having the strongest correlation positively, Volatile acidity having the strongest negative correlation, and residual sugar having the closest neutral correlation out of all attributes.

# Implementation and Results

## Random Forest Classifier Implementation

Graphical user interface, text, application

Description automatically generated

Figure - Source Code for RandomForestClassifier

First, the dataset is separated, breaking away the quality from the rest of the attributes. Then, the data is trained with a test size of 25%, and then model training is done through RandomForestClassifier and the training data is fitted.

To test the accuracy, we predict all the attributes given, and then give an accuracy score based on the real counterparts, we finally end up with a fair accuracy of 72%. Meaning that 72% of the time, it will have chosen the quality correctly.

## Random Forest Classifier Results

A picture containing text

Description automatically generated

Figure - User Input Results

To allow users to test their hand at having their own wine tested, the initial “wine\_data” variable is modified. In this case, for testing accuracy, the average of all attributes were input, which, when turned into an array and reformed into one entry and predicted, provides a quality of 6. Which is quite accurate, because the average quality is a 5.636, and because it cannot be a float, it would be a 6 in this case.

## Linear Regression Implementation

# Discussion

Do the results make sense

Is there some obvious reasoning behind the results

Why did some method work better than some other

# Conclusion

Can some conclusions be made?

How did the project go?

What could be done further on the subject?

# Appendices

Insert link to clean jupyter notebook that does all the work written about.

Insert link to dataset used.