Machine Learning Project – Red Wine Quality Prediction

Kurtis Shane Wilhoite

TABLE OF CONTENTS

1 INTRODUCTION

1.1 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 inherent 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.

The notebook file which is discussed further can be found from here. (Wilhoite, 2021)
The dataset which is used further can be found here. (Cortez, Cerdeira, Almeida, Matos, & Reis, 2009)

1.2 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.

2 ANALYSIS AND VISUALIZATION

2.1 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.

```
dataset.describe()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Figure 1 - 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).

2.2 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.

```
[23] #How volatile acidity affects quality
     sns.barplot(x='quality', y='volatile acidity', data = dataset)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f72cd2e9cd0>
```

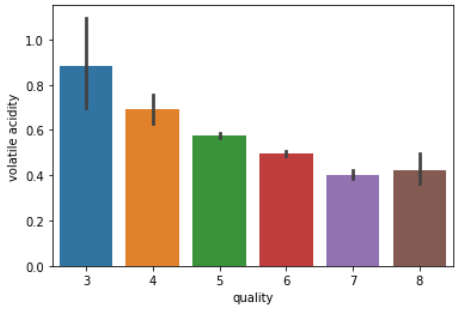

Figure 2- Volatile Acidity vs Quality

```
#How citric acidity affects quality
sns.barplot(x='quality', y='citric acid', data = dataset)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f72d6166d50>
```

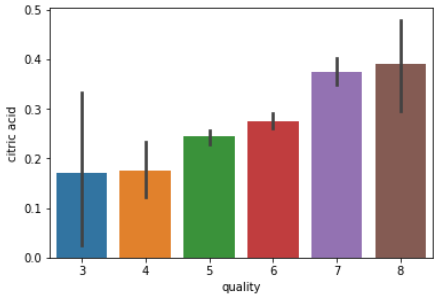

Figure 3- Citric Acid vs Quality

If we compare 2 of the 3 types 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 as it rises (*Figure 3*).

```
[22] #How Fixed acidity affects quality
     sns.barplot(x='quality', y='fixed acidity', data = dataset)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f72cd350f10>
```

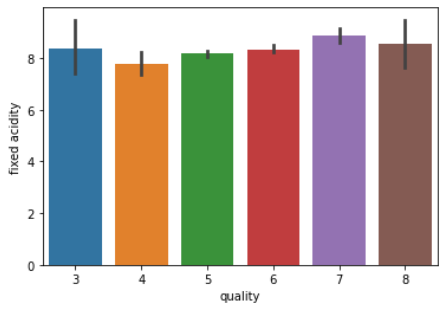

Figure 4- Fixed Acidity vs Quality

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.

```
#Total correlation heatmap to show full-scale of correlation between each attribute
datasetCorrelation = dataset.corr()
plt.figure(figsize=(9,9))
sns.heatmap(datasetCorrelation, cbar=True, square=True, annot=True, fmt='.2f', cmap="Reds")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f72ca4db410>
```

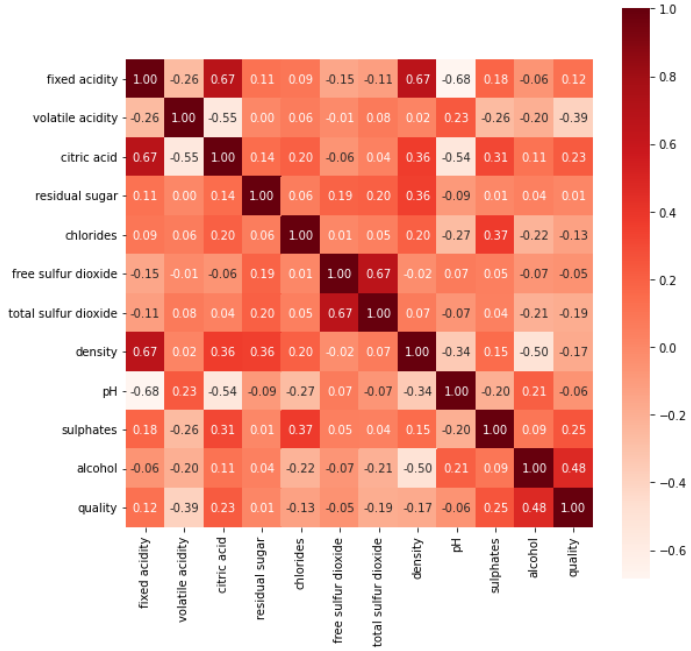

Figure 5- Correlation Heatmap

For an in-depth analysis of all correlations, a heatmap gives a clear indication of how each attribute can affect each other (*Figure 5*). 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.

3 IMPLEMENTATION AND RESULTS

3.1 RANDOM FOREST CLASSIFIER IMPLEMENTATION

```
[51] #Seperating Quality from other attributes
     attributes = dataset.drop('quality',axis=1)
     quality = dataset['quality']

     #Training and Testing Splits
     attributes_train, attributes_test, quality_train, quality_test = train_test_split(attributes, quality, test_size=0.25, random_state=43)

     #Model Training
     model = RandomForestClassifier()
     model.fit(attributes_train, quality_train)

     RandomForestClassifier()
```

```
[52] #Accuracy test
     attributes_test_prediction = model.predict(attributes_test)
     test_accuracy = accuracy_score(attributes_test_prediction, quality_test)
     print('Accuracy: ', test_accuracy)

     Accuracy:  0.7275
```

Figure 6 - 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 reasonable accuracy of 72%.

3.2 RANDOM FOREST CLASSIFIER RESULTS

```
#Data Input - Edit this to make changes!
#In order: Fixed Acidity, volatile acidity, citric acidity, sugar, chlorides, FSO2, SO2, Density, pH, sulphates, alcohol
wine_data = (8.31,0.52,0.27,2.53,0.087,15.87,46.46,0.99,3.31,0.65,10.42)

wine_data = np.asarray(wine_data)
wine_data = wine_data.reshape(1, -1)

#Modify data to predict only 1 drink
prediction = model.predict(wine_data)
print(prediction)

[6]
```

Figure 7- 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.

3.3 Linear Regression Implementation

```
[46] wine_train, wine_test = train_test_split(dataset, test_size=0.15, random_state=43)

     def train_and_pred(quality, attributes):
         reg = LinearRegression().fit(wine_train[attributes], wine_train[quality])
         quality_pred = reg.predict(wine_test[attributes])
         return quality_pred


     prediction = train_and_pred('quality', attributes.columns)
     wine_test['prediction'] = prediction
     wine_test.head(10)
```

Figure 8- Linear Regression Source Code

To test a far more inaccurate method of achieving the earlier process, Linear Regression was used to see if general trends could still be produced or if the data was simply being pulled in too many directions to call for a proper outcome and to show if Linear regression could be worked on further in the future to produce adequate results. The data is once again split into test and training and then each attribute overall is broken down into comparing it's effects on quality via linear regression which compares correlations.

3.4 Linear Regression Results

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	prediction
547	10.6	0.31	0.49	2.5	0.067	6.0	21.0	0.99870	3.26	0.86	10.7	6	6.196261
1197	7.7	0.57	0.21	1.5	0.069	4.0	9.0	0.99458	3.16	0.54	9.8	6	5.463080
500	7.8	0.52	0.25	1.9	0.081	14.0	38.0	0.99840	3.43	0.65	9.0	6	5.181661
631	10.4	0.28	0.54	2.7	0.105	5.0	19.0	0.99880	3.25	0.63	9.5	5	5.622290
1128	10.0	0.43	0.33	2.7	0.095	28.0	89.0	0.99840	3.22	0.68	10.0	5	5.600249
782	9.0	0.82	0.05	2.4	0.081	26.0	96.0	0.99814	3.36	0.53	10.0	5	4.990955
1121	6.6	0.50	0.00	1.8	0.062	21.0	28.0	0.99352	3.44	0.55	12.3	6	6.249671
426	6.4	0.67	0.08	2.1	0.045	19.0	48.0	0.99490	3.49	0.49	11.4	6	5.632425
832	10.4	0.44	0.42	1.5	0.145	34.0	48.0	0.99832	3.38	0.86	9.9	3	5.697635
145	8.1	0.67	0.55	1.8	0.117	32.0	141.0	0.99680	3.17	0.62	9.4	5	4.842693

Figure 9- Linear Regression Results

As predicted, Linear regression was not optimal for prediction of this varied of input within multiple different attributes, as it's more suited for attributes with higher correlation and less varied quantities separating the quality and prediction. However, that was never the intent of using Linear Regression, but instead to see if the general correlations stayed the same throughout.

```
sns.lmplot(x='prediction', y='quality', data=wine_test)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fcccd7700d0>
```

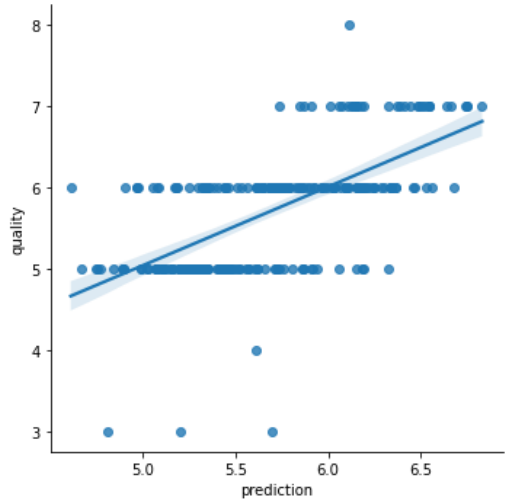

Figure 10-Prediction vs Actual Quality

```
sns.lmplot(x='prediction', y='alcohol', data=wine_test)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fcccd6f7290>
```

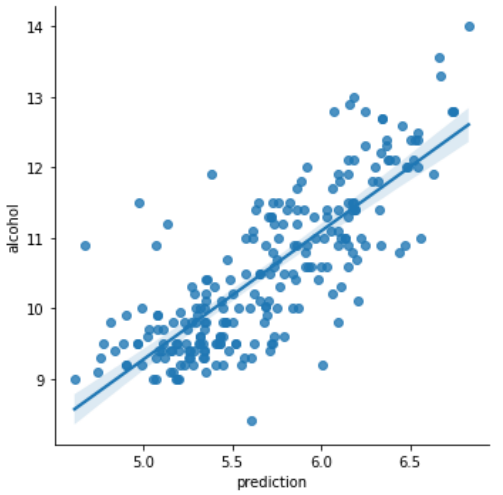

Figure 11-Prediction vs Alcohol Content

A general trend can be seen with the predictions vs the quality (*Figure 10*). The algorithm seems to play it safe, mostly only giving no predictions that are too high or too low to prevent itself from assuming outliers. If comparing the attribute with the highest effect on quality, Alcohol content (*Figure 11*), a general trend can be seen clearly as alcohol content goes up, the prediction rises, showing correlations are the same between these predictions and their real-life counterparts.

```
sns.lmplot(x='prediction', y='volatile acidity', data=wine_test)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fcccce44690>
```

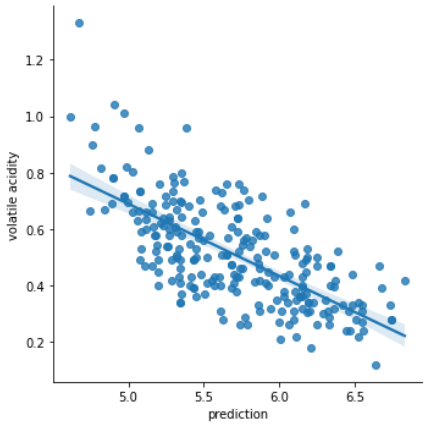

Figure 12-Quality prediction vs Volatile Acidity

```
sns.lmplot(x='prediction', y='residual sugar', data=wine_test)
```

```
<seaborn.axisgrid.FacetGrid at 0x7fcccd572290>
```

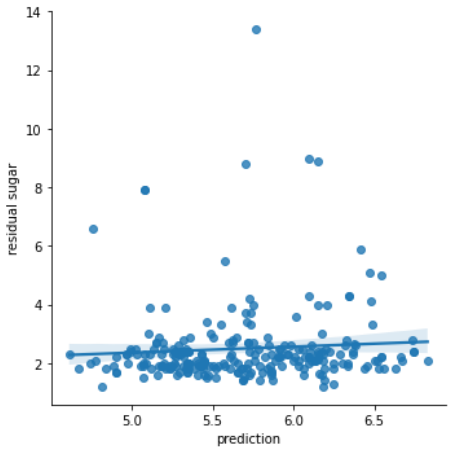

Figure 13-Quality prediction vs Residual Sugar

For further verification, Volatile acidity which has the strongest negative correlation (*Figure 12*) and Residual sugar, which has the most neutral correlation (*Figure 13*) were tested against their predictions. Which further show the correlation had generally remained the same within the predictions and that the inaccuracies present within Linear Regression were largely due to the lack of variety of entries and algorithm preferring to play estimates safe, than to guess toward outliers.

4 Discussion

Within the classification method, a relatively concrete accuracy could be obtained which maxed out around 72.5%. While this may seem relatively low, this is a prediction on a scale of 1-10, which throws in much more variation since there are so many other attributes which both negative, positively, and neutrally correlate with the quality.

Linear regression was tested with the assumption of an inaccurate reading in mind, rather to test if the correlations remained about the same, which was the case in all major aspects. Going back and comparing the predictions vs attributes and the actual qualities vs attributes reveals almost an entirely similar line of linear regression.

Figure 14-Quality vs Count per Quality

In both cases, the algorithms were highly resistant to outlier situations, almost exclusively preferring middle predictions, as that was where a significant majority of the qualities were held. This is rather natural for the algorithms to attempt due to the overwhelming evidence of quantities between 5-6 range vs other numbers (*Figure 14*). A vast majority was held at a middle quantity of 5-6, when compared to a high quality of 7 or above, and even fewer reached below into the 4s and under.

5 Conclusion

Overall, the correlations remained true in both methods used, however the classifier method was far more effective at producing accurate results due to its enhanced benefits over the rather, well…linear, uses applicable with linear regression. Considering the base dataset did not vary so heavily from the middle of quality predictions, it is rather accurate.

Due to the nature of the numbers, the lowest that any algorithms would produce would be a 3, and the highest would be of an 8. As humans, we have exterior knowledge we can use to produce higher and lower numbers than these, but the algorithms themselves only have the data to go from, which produced a minimum of a 3, and a maximum of an 8. The minority quantities of these as well would mean that it's hard to replicate a vastly higher or lower number as well.

If further data was obtained, both on the lower and higher ends, as opposed to the middle qualities of wine, the algorithms applied may be far more accurate, as such, furthering this project to be capable of producing a 10-quality wine would require further data.

One alternative to produce more accurate results as well is found within the study itself where the dataset was retrieved. Within the dataset, the quality is an average between a minimum of 3 wine experts. This lends more credibility to the quality's results. However, this also makes it where numbers will have naturally bottomed out around 8 as a maximum and 3 minimum. If given a dataset where all expert's judgments are given separately, it would allow for more variation and further accuracy in the results received, with the downside of receiving a bit more outliers which may affect accuracy negatively, but the additional data would overcome those potential discrepancies.

6 APPENDICES

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Wine Quality Data Set*. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/wine+quality

MasterClass Staff. (2021, July 29). *Learn About Alcohol Content in Wine: Highest to Lowest ABV Wines.* Retrieved from Master Class: https://www.masterclass.com/articles/learn-about-alcohol-content-in-wine-highest-to-lowest-abv-wines#what-does-alcohol-by-volume-abv-mean

Puckette, M. (2019, September 19). *What is residual sugar in wine?* Retrieved from Wine Folly: https://winefolly.com/deep-dive/what-is-residual-sugar-in-wine/

Wilhoite, K. (2021, December 6). *Red Wine Quality Machine Learning Project*. Retrieved from Github: https://github.com/KurtisWilhoite/RedWineQualityPrediction/blob/master/Red_Wine_Quality_Prediction_and_Analysis.ipynb