**Predicting the Quality of Wine using their Physical Properties**

Dylan Tulett, [dtulett@bellarmine.edu](mailto:dtulett@bellarmine.edu)

Brian Johnston, [bjohnston2@bellarmine.edu](mailto:bjohnston2@bellarmine.edu)

**ABSTRACT**

Wine quality can be defined using its physical qualities. Using a dataset of Portuguese red wines and a Random Forest Model for machine learning, we can predict with nearly 60% accuracy the quality of wine on a 0 to 10 scale.

1. **INTRODUCTION**

Wine comes in many different varieties. Even when discussing color, there are categories of color, proof, and many others. When picking a wine to one’s own tastes, there are many factors to consider, though a large indicator of a good wine is a quality value, which can be calculated based on the physical components in the wine.

The data set we collected lists physical properties of wines, and then gives a quality value from 0 to 10. When applying machine learning algorithms, we will be predicting a non-binary value, so we decided to use the Random Forest Model.

1. **BACKGROUND**
   1. *Data Set Description*

This dataset can be found at <https://archive.ics.uci.edu/ml/datasets/wine+quality>. We decided to only look at the red wines, since there are clear differences in red and white wines, and so combining them or analyzing them separately would require 2 models. The dataset lists 10 physical properties of each of the 1599 wines, those being fixed acidity, volatile acidity, citric acid, residual sugars, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The data itself was collected on Portuguese “Vinho Verde” wines by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, who have their own paper on the data.

* 1. *Machine Learning Model*

Originally, we set out to use the kn Neighbor Model, but when analyzing, we realized that non-binary predictions would be far harder to achieve with this model. A similar circumstance arose with the Naïve Bayes Model, and then we settled on the Random Forest Model. This model allows our model to fit more easily with our predictive values.

The Random Forest Model combines various random decision trees and chooses the final value based on what was predicted the most from the random trees. Decision trees in the case of wine quality would take the form of starting based in one physical property, analyzing what that would mean in terms of quality and moving on to the next physical property, repeating until all properties have been analyzed, resulting in a final quality value.

1. **EXPLORATORY ANALYSIS**

This data set contains 11 variables with 1599 samples. The physical properties columns are all continuous variables, where the quality column is a discrete variable.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Fixed Acidity | Float, Continuous |
| Volatile Acidity | Float, Continuous |
| Citric Acid | Float, Continuous |
| Residual Sugar | Float, Continuous |
| Chlorides | Float, Continuous |
| Free Sulfur Dioxide | Float, Continuous |
| Total Sulfur Dioxide | Float, Continuous |
| Density | Float, Continuous |
| pH | Float, Continuous |
| Sulphates | Float, Continuous |
| Alcohol | Float, Continuous |
| Quality | Integer, Discrete |

T**able 2: Summary Statistics of Total Sulfur Dioxide**

|  |  |
| --- | --- |
|  | total sulfur dioxide |
| count | 1599.0 |
| mean | 46.46779237023140 |
| std | 32.89532447829900 |
| min | 6.0 |
| 25% | 22.0 |
| 50% | 38.0 |
| 75% | 62.0 |
| max | 289.0 |

T**able 3: Summary Statistics of Free Sulfur Dioxide**

|  |  |
| --- | --- |
|  | free sulfur dioxide |
| count | 1599.0 |
| mean | 15.874921826141300 |
| std | 10.46015696980970 |
| min | 1.0 |
| 25% | 7.0 |
| 50% | 14.0 |
| 75% | 21.0 |
| max | 72.0 |

T**able 4: Summary Statistics of Residual Sugar**

|  |  |
| --- | --- |
|  | residual sugar |
| count | 1599.0 |
| mean | 2.5388055034396500 |
| std | 1.4099280595072800 |
| min | 0.9 |
| 25% | 1.9 |
| 50% | 2.2 |
| 75% | 2.6 |
| max | 15.5 |

Notice in each of the summary tables for residual sugar, free sulfur dioxide, and total sulfur dioxide, all of the ranges are quite large. The max number is much larger than the 75th percentile for each of them.

**Figure 1: Alcohol Boxplot by Quality Score**

**Chart, box and whisker chart

Description automatically generated**

Notice the mean alcohol goes up with the higher quality red wines. This may have helped the ML model in predicting wine qualities.

**Figure 2: Residual Sugar Boxplot by Quality Score**

Chart, box and whisker chart

Description automatically generated

Interestingly, there doesn’t seem to be much change in mean or range for residual sugars. Despite this, it was useful in creating an efficient and mostly accurate ML model.

**Figure 2: Fixed Acidity Boxplot by Quality Score**

Chart, box and whisker chart

Description automatically generated

Similarly to alcohol, the fixed acidity tends to go up with quality. But it is not a strong as alcohol. This was however still very useful in the ML model despite the visually unclear trend.

**Figure 4: Fixed Acidity vs. Density Scatterplot**

**Chart, scatter chart

Description automatically generated**

**Figure 5: Fixed Acidity vs pH Scatterplot**

**Chart, scatter chart

Description automatically generated**

Both figure 4 and figure 5 showed interesting relationships that were worth noting. Both seem visually strong.

1. **METHODS**
   1. *Data Preparation*

The data we used needed to be normalized for proper usage, as we have density, which for liquids like wine tends to be about 1, and total sulfur dioxide, which had a range of 6 to 289. We kept all the columns since each of them affect the quality value. But only residual sugar, alcohol, and fixed acidity were used for the machine learning model.

* 1. *Experimental Design*

Table 5: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All 3 normalized features with a 60/20/20 split for train, validate, and test |
| 2 | All 3 normalized features with a 50/25/25 split for train, validate, and test |
| 3 | All 3 normalized features with a 70/15/15 split for train, validate, and test |
| 4 | All 3 raw features with a 50/25/25 split for train, validate, and test |
| 5 | All 3 raw features with a 60/20/20 split for train, validate, and test |

* 1. *Tools Used*

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Windows computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, and SKLearn 0.18.1.

Pandas was used to manipulate the data and gather summary statistics for each variable. Numpy was used to split the data for inputting into the models. Matplotlib was used to analyze the data, and Seaborn was used to visualize the data. SKLearn was used for the machine learning models to predict wine quality.

1. **RESULTS**
   1. *Classification Measures*

|  |  |
| --- | --- |
| **Measurement** | **Score** |
| Accuracy | 0.59 |
| Precision | 0.56 |
| Recall | 0.59 |
| F1 | 0.57 |

* 1. *Discussion of Results*

Our worst model was the Naïve Bayes Model. This makes sense, as many of the physical factors that were used in our predictions ended up being correlated, and thus we did not have completely independent variables. As for the best model, being the Random Forest Model, we believe the reason it worked so well is because it makes use of all the predictive values and combines previous knowledge to form a decision. This works best for the purposes of this problem, making the Random Forest a good choice.

* 1. *Problems Encountered*

Finding a model that fit a non-binary model well was a bit of a rough time, we tried a few models, with plenty of experiments each to see if there we any that kept an average of 55% or better, and we landed on Random Forest which luckily did give us a model above what we were looking for. Another issue we discovered was the ROC curve. The information given by an ROC curve is not influential when discussing a non-binary variable, so we could not use them for our models.

* 1. *Limitations of Implementation*

The model used only tested three variables. The ideal case would be using all ten of the physical traits to determine the wine’s quality, but with the computations available, three was more optimized for the processes at hand and worked well enough. Another case would be a different model, one that wasn’t discussed during class, perhaps, that’s better for modeling non-binary classifications. The model also may be limited to just the specific type of wine, or the specific region this wine data was collected from. It may not predict outside these boundaries well.

* 1. *Improvements/Future Work*

The next course of action from here would be to build the models with more variables to get a better fitting model. Other than that, there are other models that were not discussed during the course that may fit this question far better than the Random Forest Model, regardless of the inputs.

1. **CONCLUSION**

The problem that our model handled was predicting the quality of red wine based on the physical properties of the given wine. Using the given data and our model, the predictions are accurate to about 60% of the time. With a non-binary classification, accuracy above 50% is deemed a good model, and our goal was to look for something close to 60%, if we could not find better. Overall, this model is good enough, but could be better with some modifications. Choosing to follow the Random Forest Model allowed us to explore into the concepts of using machine learning for more than just binary problems and provided a decent model that may be a good fit for other, similar problems.

**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.