Introduction

Quality control and assurance is tantamount for making consistent premium products. In order to better optimize quality, data about the products are often cataloged and used to improve future innovations. When developing wine on a large scale, quality control extends to measuring multiple physicochemical characteristics of the wine. Precise control and understanding of these characteristics on the overall quality of the wine can improve the overall product and subsequently the price charged.

This project is designed to accomplish two major tasks using wine characteristic data from a Portuguese winery. The first task is to use the physicochemical data of the wines to build a classification model to predict overall quality of the wine. As discussed below, this analysis will be done on both red and white wines individually and differences between quality predictions will be gathered and noted. The second task is to unify the data from both red and white wines to predict the color of the wine based on the physicochemical characteristics. The goal of this is to determine which characteristics are closely related to red or white wines and could be applied to wines that fall within that spectrum (e.g. if a rosé wine will be more similar to a red or white wine).

The datasets used for this analysis come from a report of the Vinho Verde style of Portuguese wine published by Cortez et. al.[1] This dataset contains 6497 total entries divided into two separate mini datasets of red (1599 entries) and white (4898 entries) wine. All the entries contain physicochemical variables as inputs and sensory data as output (e.g. quality of wines). No information was given on the types of grapes, brand of wine, price of wines or other proprietary data.

The dataset contains 12 variables containing 11 input variables based on physicochemical tests and one output variable based on sensory data. Nine of the input variables are based on the amounts of fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, sulfates, and alcohol dissolved in the wine. The last two input variables are the pH and density of the wine solution. All variables had values that were of the “float” type. The output variable, quality, was listed as integers with values from 0 to 10.

Fixed acidity, volatile acidity, citric acid, and pH variables all indicate how acidic or tart the wine is and the relative qualities of that acid. Citric acid is a common additive to lower the pH of wine during the winemaking process and is a weak preservative with a “fresh” flavor.[2] Residual sugar represents the relative sweetness of the wine and how much sugar is left after fermentation. The level of chlorides in the wine reflects the extraction process of the juice used in the process and is often influenced by the geographic, geologic, and climate conditions of the vine culture

[1] https://archive.ics.uci.edu/ml/datasets/wine+quality
[2] https://wineserver.ucdavis.edu/industry-info/enology/methods-and-techniques/common-chemical-reagents/citric-acid

(e.g. terroir). The level of chloride also represents the amount of sodium ions in the wine which is often regulated at a country level.[3]

Sulfur dioxide is added to wines to help preserve the subtle characteristics by resisting oxidation but may have undesired side effects if in high amounts. Free sulfur dioxide and total sulfur dioxide can estimate how much of this compound is in the wine.[4] Sulfates are a measure of water quality used in the brewing process. Some winemakers that have “softer” water (water with fewer minerals) may supplement by adding chemicals like calcium sulfate to make the water “harder” (water with more minerals) which can impact desired taste of the finished product.[5] Alcohol represents the percentage of alcohol in the finished wine.

To predict the quality of the wines the analysis of the red wine and white wine data sets were processed identically but kept separate. Two sets of data analysis and modeling were performed but the processing methods were identical to both. To generate a model to predict wine color, the red wine and white wine data sets were merged together, as described below in the data wrangling section.

Data Wrangling

Two major transformations were performed on the data sets. The first transformation was converting the output variable “quality” from a numeric score to a category. This assignment was set to have a score of 7 or higher to be considered a “good” wine while a score of 6 or lower to be a “bad” wine. This transformation was performed on both of the red and white wine data sets. The breakdown of the wines considered good vs bad can be seen below in Figure 1, with 13.6% of red wines and 21.6% of white wines considered to be “good”.

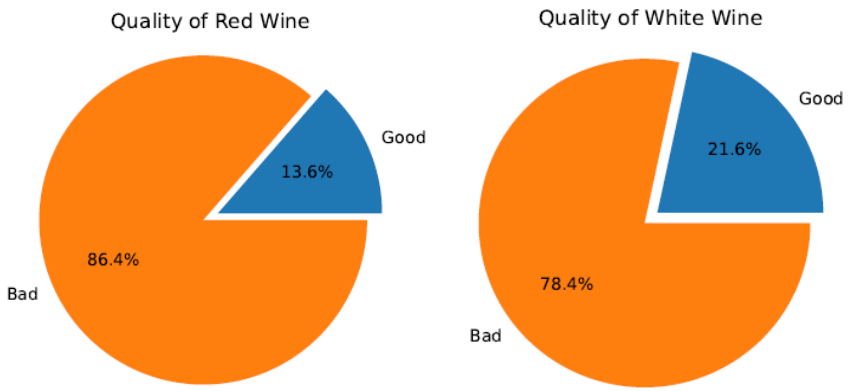

Figure 1: Percentage of Wine Quality in Red Wine (left) and White Wine (right) Data Sets

[3] https://www.awri.com.au/wp-content/uploads/2018/08/s1530.pdf, https://www.oiv.int/public/medias/2604/oiv-ma-d1-03.pdf
[4] https://agrifoodecon.springeropen.com/articles/10.1186/s40100-015-0038-1, https://daily.sevenfifty.com/how-sulfites-affect-a-wines-chemistry/
[5] https://www.winespectator.com/articles/difference-between-sulfites-sulfates-wine-54706

For the color prediction task, each of the data sets had an additional column added to reflect the "color" of the wine: red for the red wine and white for the white wine. The modified red and white wine data sets were then merged together. The transformed quality values were included in this data set for analysis. For the merged data set, 24.6% of the total wines were red and 75.4% were white, as seen in Figure 2.

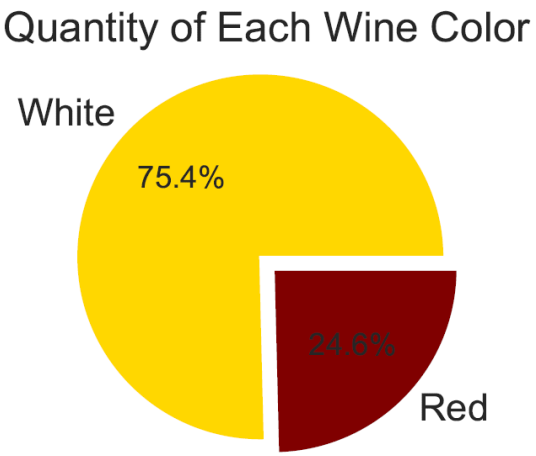

Figure 2: Percentage of Wine Color in Combined Data Set

Exploratory Data Analysis

Red Wine Quality

The quality scores of the red wines were plotted against the various input variables. The quality was not classified into good and bad for this analysis, but instead the original scores between 0 and 10 were kept. The amount of each quality score was variable with values set at the extremes (e.g. 3 and 8 for red wine) occurring at a lower frequency compared to those closer to the median (e.g. 5 or 6). As such, it should be expected to have larger error bars on these more extreme scores. This behavior is observed in most of the plots below within the data analysis. Additionally, this report only includes bar chart plots but box plots were also generated and can be viewed in the notebook.

Some features showed no significant variation between quality values. Most notably, fixed acidity, residual sugars, density, and pH had mostly consistent values across all values of quality. These features were still included in the analysis but most likely did not have a strong influence, as will be seen in the feature importance below.

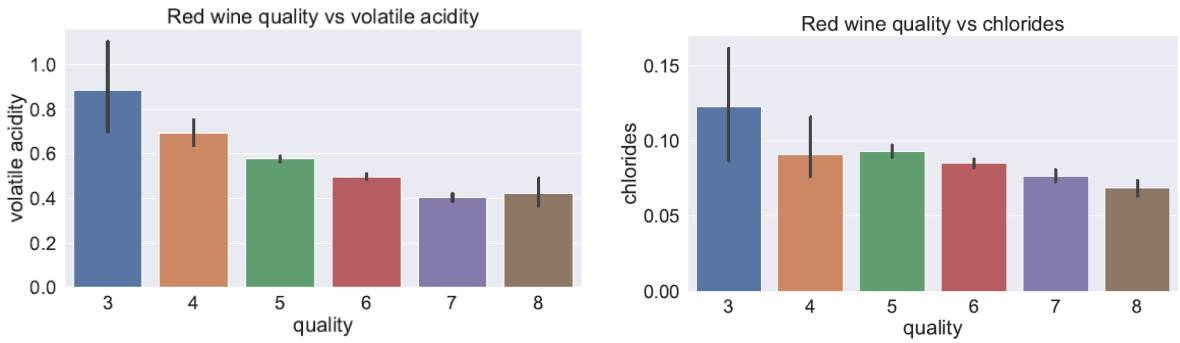

Figure 3: Volatile Acidity vs Quality (left), Chlorides vs Quality (right)

Volatile acidity and chlorides have an inverse relationship with the quality of the wine: more volatile acids and chlorides were present in wines assigned with lower quality. This relationship is seen in Figure 3. Conversely, wines with relatively more citric acid, sulfates, and alcohol were assigned with higher quality as seen in Figure 4. Higher sulfates probably implies that the water used was either "harder" or adjusted to improve taste of the final wine. Similarly, citric acid may impart the desired fresh flavor to compliment the overall wine characteristics. The increase of alcohol in the wine may also reflect on the overall taste of the finished products.

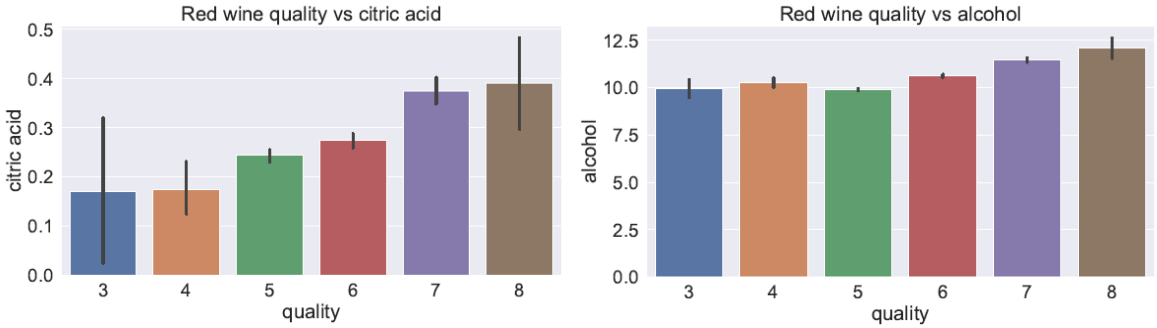

Figure 4: Citric Acid vs Quality (left), Alcohol vs Quality (right)

The amount of sulfur dioxide (represented in both "free sulfur dioxide" and "total sulfur dioxide") was represented with an upside-down U shaped curve increasing from quality values of 3 to a maximum at 5 then descending again (Figure 5). This implies that wines assigned as "bad" have a wide range of sulfur dioxide and that good wines tend to have a smaller window. As some individuals are sensitive to higher values of sulfur dioxides, higher levels of sulfur dioxide in lower quality wines is not surprising.

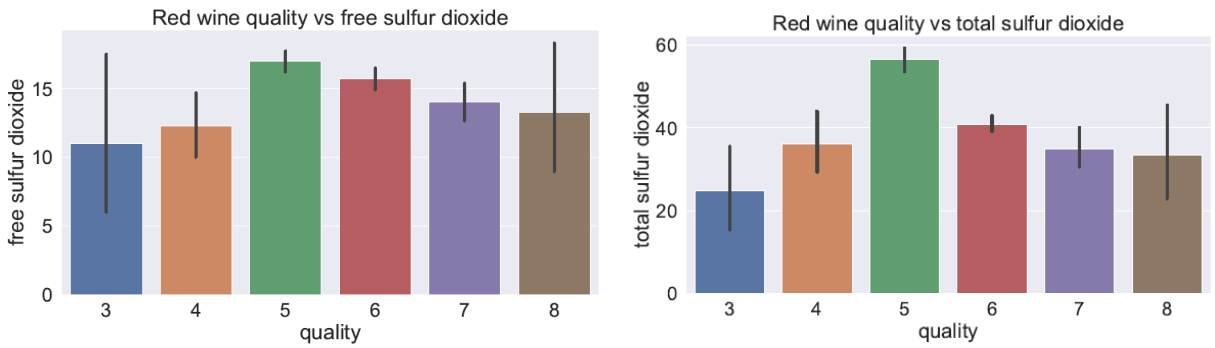

Figure 5: Free Sulfur Dioxide vs Quality (left), Total Sulfur Dioxide vs Quality (right)

The heat map of features for the quality of red wine used 12 variables: 11 features and the outcome variable, “quality”. The heat map was generated using the standard Pearson’s R function pulled from the scipy.stats library and can be seen in Figure 6. Based on the heat map, there are a few features that should be noted for collinearity. Values that typically have correlation numbers higher than 0.7 or lower than -0.7 can be omitted as there is a risk for multicollinearity. In general, most of the heatmap in Figure 6 is not at a high risk of correlation and we can assume these features are mostly independent.

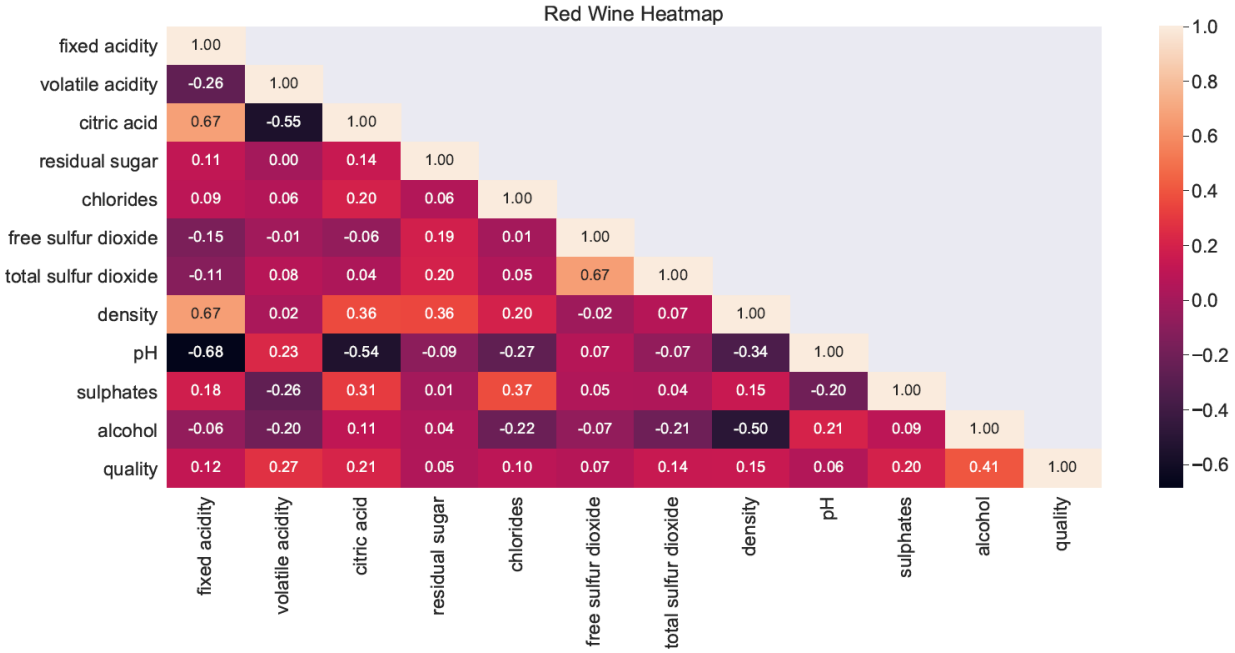

Figure 6: Heat Map of Features for Red Wine Quality Generated Using Pearson’s R

Most of the notable correlations involved similar characteristics of variables. For instance, an increase in citric acid will have an effect on the fixed or volatile acidity as seen with correlation values up to 0.67. Since pH is a measure of the acidity, a strong correlation in fixed acidity and citric acid can be seen as well with values of -0.68 and 0.54, respectively. Free sulfur dioxide

was correlated with the total sulfur dioxide amounts, as a portion of the total sulfur dioxide was reacted to preserve the freshness of the wine. Lastly, density is highly correlated with more ions and compounds dissolved in the wine, which is reflected in significant correlation values for fixed acidity, citric acid, residual sugars, and chlorides.

For the quality, no major correlations are present aside from alcohol percentage. A preliminary hypothesis based on the exploratory data analysis would imply the tasters interpreted higher alcohol present in red wine as higher quality.

White Wine Quality

The quality scores of the white wines were plotted against the various input variables. The quality was not classified into good and bad for this analysis, but instead the original scores between 0 and 10 were kept. The amount of each quality score was variable with values set at the extremes (e.g. 3 and 9 for red wine) occurring at a lower frequency compared to those closer to the median (e.g. 5 or 6). As such, it should be expected to have larger error bars on these more extreme scores. This behavior is observed in most of the plots below within the data analysis. Additionally, this report only includes bar chart plots but box plots were also generated and can be viewed in the notebook.

Some features showed no significant variation between quality values. Most notably, fixed acidity, density, and sulfates had mostly consistent values across all values of quality. These features were still included in the analysis but most likely did not have a strong influence, as will be seen in the feature importance below.

The values for pH were also mostly consistent as seen in Figure 7, however, there was a slightly higher pH value for the highest quality white wines. Citric acid has a mostly consistent amount between quality levels except for a small dip at quality score of 4 and a noticeable increase at a quality score of 9. While these values may be exciting for predicting quality against the amount of citric acid, the error bar present at a quality score of 9 was very large and should therefore be considered with a grain of salt.

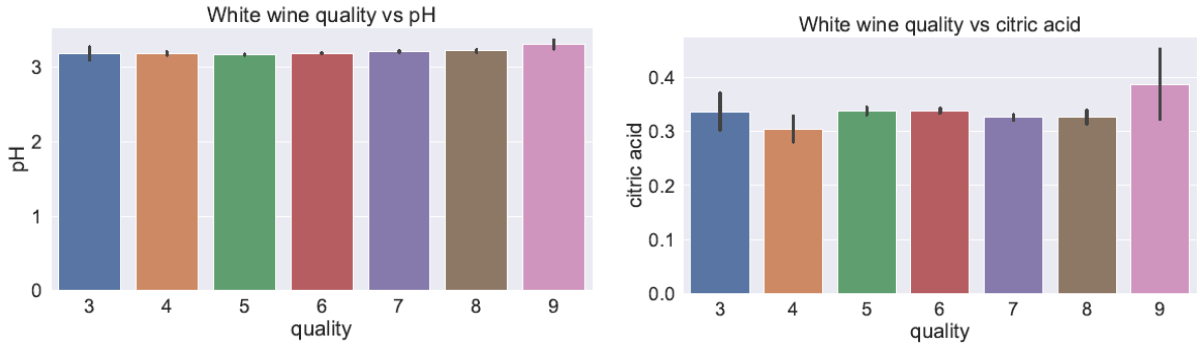

Figure 7: pH vs Quality (left), Citric Acid vs Quality (right)

Similarly, the plot of volatile acidity vs quality has an unusual shape where lower quality white wines have a larger range of volatile acidity compared to the higher quality wines.The residual sugar values have large variability both between individual quality scores and across the range of quality scores. This variability most likely has an impact in predicting quality for these types of wines.

The amount of sulfur dioxide (represented in both “free sulfur dioxide” and “total sulfur dioxide”) was represented in Figure 8. For “free sulfur dioxide”, the amount was mostly consistent between quality scores 5 through 9 with larger variability in scores 4 and lower. White wines can be more prone to over-oxidation which decreases the quality of the wine. Having more free sulfur dioxide could imply over-oxidation. Similarly to red wines, the white wines assigned as “bad” have a wide range of sulfur dioxide and that good wines tend to have a smaller window. The total sulfur dioxide for these wines tend to generally follow an inverse trend with white wine quality, with the exception being irregular behavior around quality scores 3 and 4.

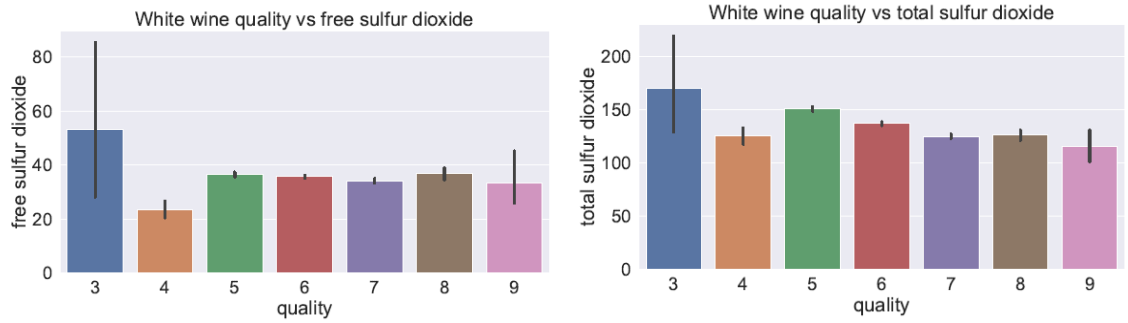

Figure 8: Free Sulfur Dioxide vs Quality (left), Total Sulfur Dioxide vs Quality (right)

Alcohol and chlorides did have a clear correlation to the quality scores, as seen in Figure 9. Similarly to red wine, the white wines with more alcohol were generally perceived to have a higher quality. The amount of chlorides present in the wine decreased the quality of the white wines. This makes sense as chlorides are often extracted from the grape skins. Too much extraction of the grape skins can discolor white wines and introduce more side characteristics not desirable in most white wines.

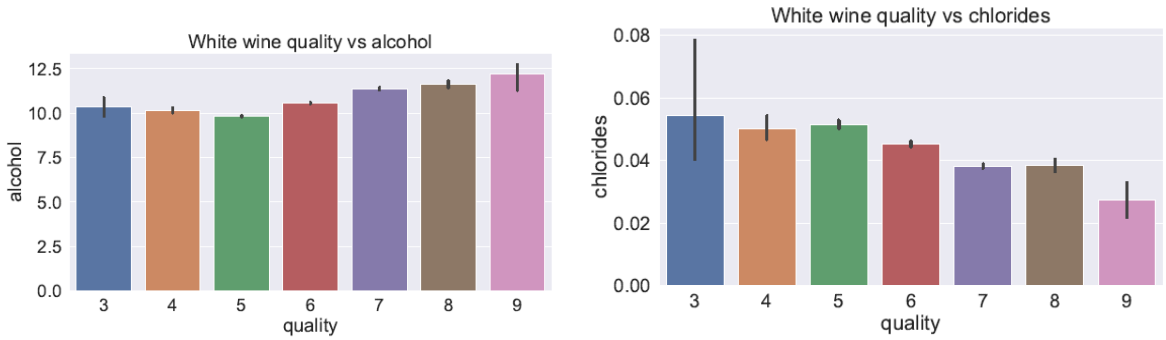

Figure 9: Alcohol vs Quality (left), Chlorides vs Quality (right)

The heat map of features determining quality for the white wine was generated in an identical fashion to that for the red wine data set and can be seen in Figure 10. Based on the heat map, there are a few features that should be noted for collinearity. In general, most of the heatmap in Figure 10 is not at a high risk of correlation and we can assume these features are mostly independent. Those that were not predominantly relate to density and sulfur dioxide which will be discussed below.

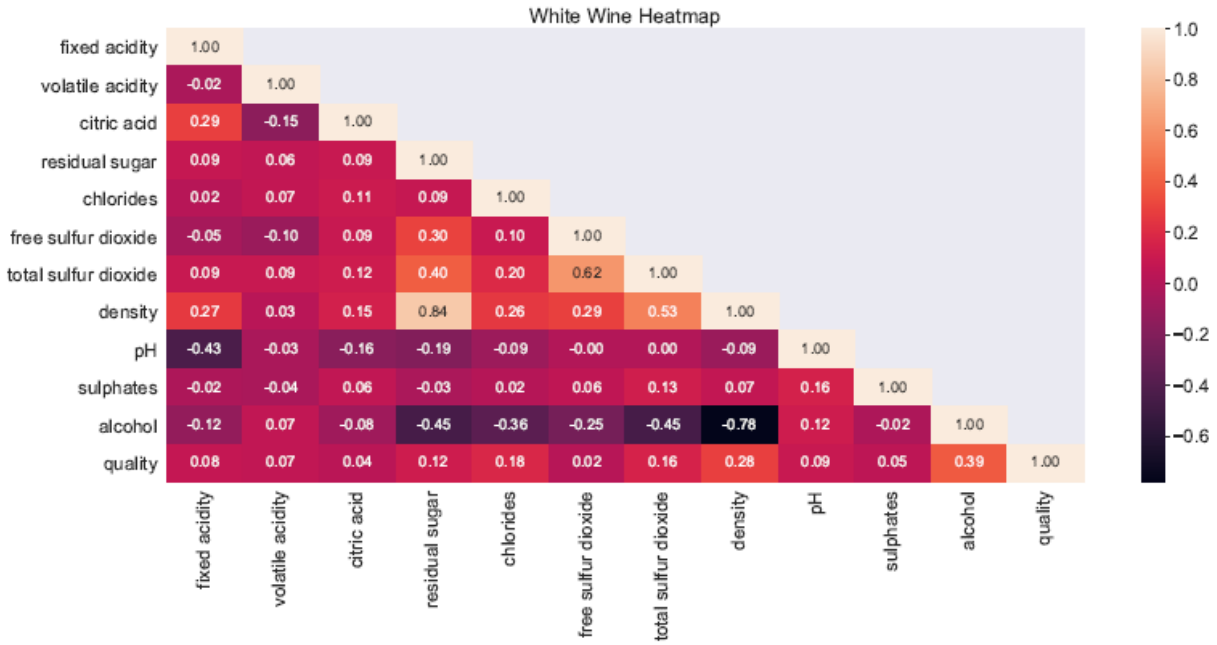

Figure 10: Heat Map of Features for White Wine Quality Generated from Pearson's R

Similarly to red wine, density is a function of everything dissolved in the wine. Unlike with red wine, there are higher relative levels of residual sugars in white wine which will more strongly influence the density, hence the correlation value of 0.84. Alcohol also impacts the density in the opposite way, with more alcohol decreasing the density of the liquid, shown with the negative correlation of -0.78. In a similar fashion to red wine, the white wine has a correlation between the free and total amounts of sulfur dioxide, with a correlation value of 0.62.

With all of these high correlations, there seems to be smaller amounts of correlation with the quality feature so no action will be taken at this stage for the red or white wine data sets to eliminate features. If this collinearity dropped the model accuracy and efficacy, removal of certain features may be pursued.

Feature Importance - Wine Quality

A standard methodology using the random forest classifier was selected to determine the feature importance for the data set.[6] To help further clarify which features had an influence in the outcome, a “random” feature was added to each client entry. This random feature was a number between 0 and 1 and was to provide a noise baseline.

For the determination of quality in the red wine data set (Figure 11), the most prominent feature seemed to be the amount of alcohol. Other features that proved to be important were the amount of sulfates, volatile acidity, and citric acid along with the density of the wine. Features like the free sulfur dioxide and pH were deemed not important while fixed acidity and residual sugar amounts were deemed to have little impact.

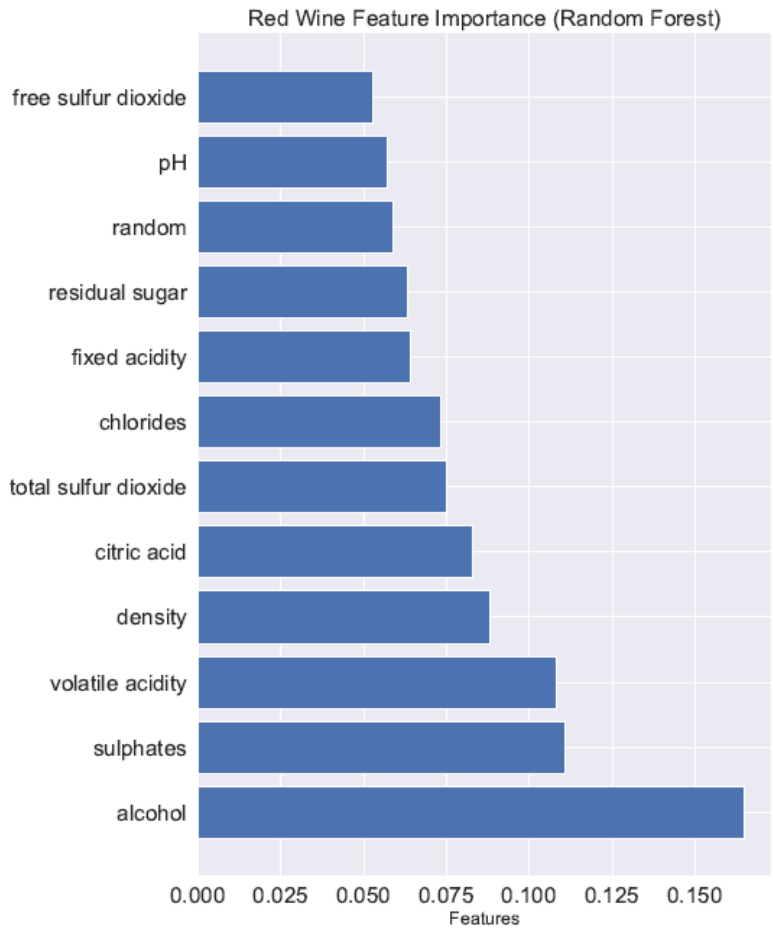

Figure 11: Feature Importance for Red Wine Quality Data Set using Random Forest Classifier

For the quality of the white wine data set (Figure 12), the two most prominent features were the amount of alcohol and the density. The other features within the top five most important included

[6] See notebook for exact method

the amount of chlorides, the amount of residual sugar, and the pH of the wine. The citric acid and fixed acidity features did not seem to have much impact on the model.

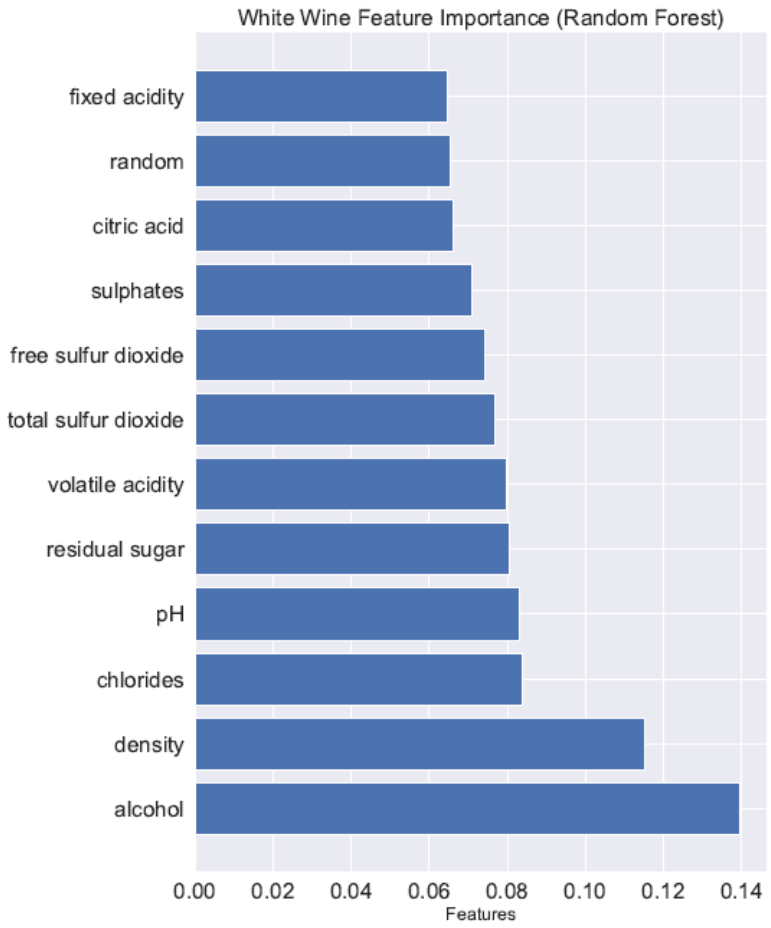

Figure 12: Feature Importance for White Wine Quality Data Set using Random Forest Classifier

Wine Color Analysis

The feature data for the combined red and white data sets exhibited the following average and median values seen below in Table 1. A majority of the features have an obvious difference between them. Specifically, red wines on average have higher fixed acidity, volatile acidity, and higher pH but lower amounts of citric acid compared to white wines. This implies that there is a different acid profile for the wines, probably from residual acids extracted from the grape skins. This also is supported by an increased amount of chlorides in red wines compared to white wines and could possibly explain the difference between sulfate levels. White wines tended to be much sweeter with a larger residual sugar amount compared to the red wines. White wines also tended to have levels of free and total sulfur dioxides much higher than that seen in red wines.

Density between the two wine types were very close in values. The alcohol content was similar, where the values were close but they were deemed different enough to include. Additionally,

data from the heat map below (Figure 13) correlated density and alcohol highly. Because of these worries about overfitting, this density feature was considered for omission from the final modeling for color analysis but ultimately left in. Quality is not reflected in Table 1 but it was included in the feature importance.

	Red Wine		White Wine	
	Average	Median, 50%	Average	Median, 50%
Fixed Acidity	8.32	7.90	6.85	6.80
Volatile Acidity	0.53	0.52	0.28	0.26
Citric Acid	0.27	0.26	0.33	0.32
Residual Sugar	2.54	2.20	6.37	5.20
Chlorides	0.087	0.079	0.046	0.043
Free Sulfur Dioxide	15.9	14.0	35.3	34.0
Total Sulfur Dioxide	46.5	38.0	138	134
Density	0.997	0.997	0.994	0.994
pH	3.31	3.31	3.19	3.18
Sulphates	0.658	0.620	0.490	0.470
Alcohol	10.42	10.20	10.51	10.40

Table 1: Average and Median Values for Red and White Wine Features

The heat map of features determining wine color was generated in an identical fashion to that for the red and white wine data sets and can be seen in Figure 13. Based on the heat map, there are a few features that should be noted for collinearity. In general, most of the heatmap in Figure 10 is not at a high risk of correlation and we can assume these features are mostly independent.

There were many instances of moderately high to high correlations between features. The most notable correlations that were concerning were density and alcohol correlating strongly at a value of -0.69 and the free sulfur dioxide and the total sulfur dioxide at 0.72. Residual sugar had some moderate correlations between free sulfur dioxide, total sulfur dioxide, and density. In general, these correlations are similar to what was observed in individual wines and are noted for model assessment later.

Additionally, there were multiple instances where features were correlated to the outcome variable, “color”. The features fixed acidity, volatile acidity, chlorides, and sulfates were all

moderately negatively correlated to the color of the wine. Total sulfur dioxides, free sulfur dioxides, and residual sugar were positively correlated with the color of the wine. These large correlation numbers were noted for when the model was assessed.

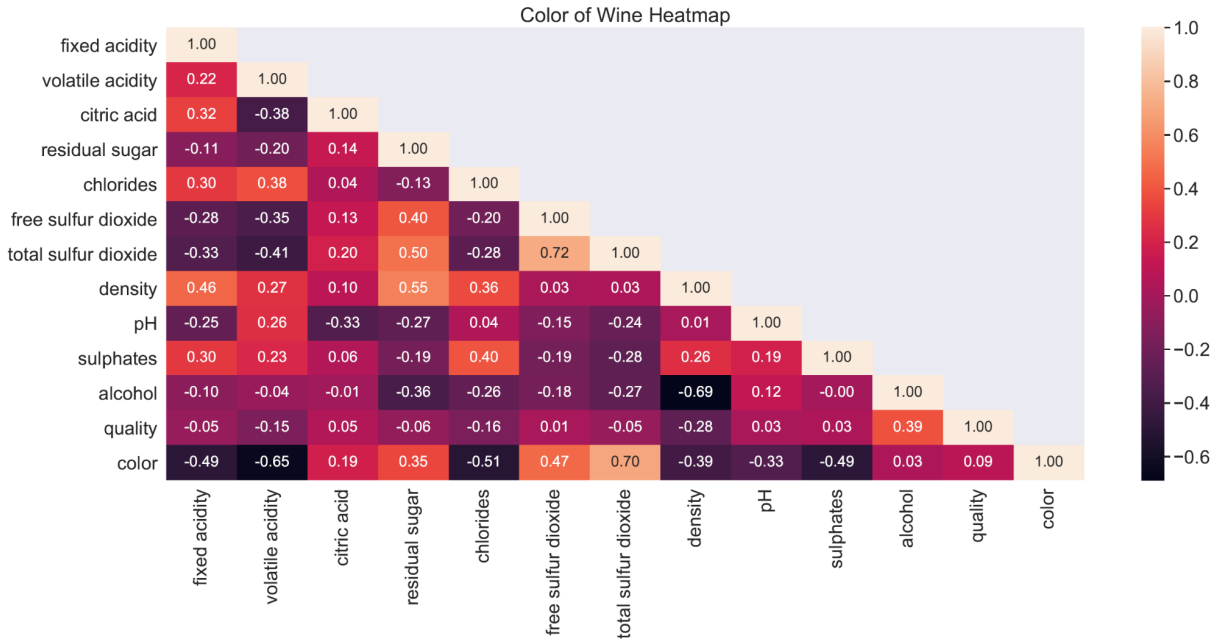

Figure 13: Heat Map of Features for Wine Color Generated from Pearson's R

Feature Importance - Wine Color

The feature importance to determine wine color was performed in an identical fashion as the red and white wine data sets individually. The combined data set was subjected to the random forest classifier with the same "random" feature to provide the background. The feature importance plot can be seen in Figure 14.

The major features deemed important for determining wine color by the random forest model were "total sulfur dioxide" and "chlorides" by a wide margin. The next three important features were "volatile acidity", "density", and "residual sugar". These features are likely illustrated in the differences between the white and the red wines in Table 1. The notable difference is that density was very similar between red and white wines.

The amount of alcohol, citric acid, and the pH of the wine were classified to be not very important in the feature analysis. The feature "quality" was deemed to be a very unreliable feature, scoring lower than the "random" baseline. As such, the "quality" feature was not included in the final model for determining wine color.

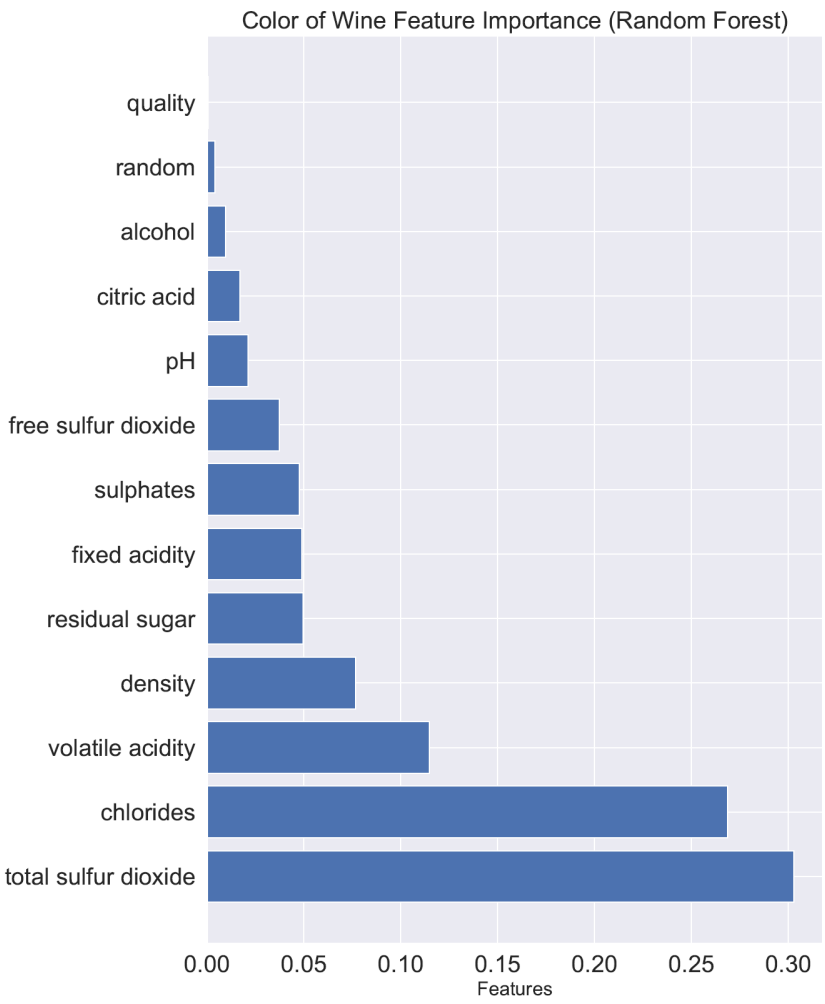

Figure 14: Feature Importance for Wine Color Data Set using Random Forest Classifier

Results

Three different classifier models were chosen for further analysis: Random Forest Classifier, Stochastic Gradient Descent Classifier, and Support Vector Classifier. Each of these underwent a grid search to tune hyperparameters based on time allotted and previous data. Each set of hyper parameters were tested with cv=5 to boost replicability. The results of the grid search on the hyperparameters for the three models are listed in Table 2. From the results of the three classifier models, the optimized Random Forest classifiers gave the best ROC-AUC scores for red wine quality, 0.924, and white wine quality, 0.912, respectively. The other two models did not perform as well. Because of this, the wine color data set was modeled using only the Random Forest classifier. The Random Forest classifier received a ROC-AUC score of 0.999 for classifying wine color under these conditions.

Data Set	Classifier	ROC-AUC Score	Best Hyperparameter Values
Quality - Red Wine	Random Forest	0.924	{'bootstrap': True, 'max_depth': 55, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 275}
	Stochastic Gradient Descent Classifier	0.892	{'alpha': 0.1, 'epsilon': 0.001, 'l1_ratio': 0.1, 'loss': 'log', 'max_iter': 1000.0}
	Support Vector Classifier	0.829	{'C': 0.3162277660168379, 'gamma': 'scale', 'kernel': 'poly', 'probability': True}
Quality - White Wine	Random Forest	0.912	{'bootstrap': True, 'max_depth': 55, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500}
	Stochastic Gradient Descent Classifier	0.784	{'alpha': 0.001, 'epsilon': 0.001, 'l1_ratio': 0.1, 'loss': 'log', 'max_iter': 1000.0}
	Support Vector Classifier	0.837	{'C': 1.3894954943731375, 'gamma': 'scale', 'kernel': 'rbf', 'probability': True}
Wine Color	Random Forest	0.999	{'bootstrap': True, 'max_depth': 55, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 500}

Table 2: Performance for Random Forest, Stochastic Gradient Descent, and Support Vector Classifiers[7] Note: ROC-AUC Scores were generated on test data.

[7] Please see the Jupyter notebook for a full workup of the data.

Conclusions

Wine Quality Analysis

The quality of both red and white wine was determined with relatively high accuracy in both cases, about 92%. Interestingly, the physicochemical features to determine what makes a good red wine are different from what makes a good white wine. This result may not be surprising to a good sommelier. The consistency of what makes a good quality red or white wine was observed from the physicochemical characteristics and should outline good quality control goals for the wine making process.

The data set could be bolstered by adding more red wines to make the two data sets more even. Additionally, the distribution of the score values were distributed in such a way that few of the lowest and highest quality wines were recorded. Additional entries of these extreme cases could bolster the evidence of what makes a truly good or bad wine of each type.

Wine Color Analysis

The ROC-AUC score for the classifier was very high. The model predicted a perfect 1.0 on the trial data and then predicted a 0.999 on the test data set. While this is suspicious, this is exactly the outcome desired: to produce a classifier that predicts the color of wine based on the characteristics. The features to choose what makes a good red or white wine were not necessarily the same as the features that were important to classify if a wine was red or white. More likely was the differences between red and white wines were easily identified by the classifier model.

To increase the efficiency for this model, minimizing the features to the important features listed in the analysis while removing extraneous features should be attempted. Using the top four most important features like total sulfur dioxide, chlorides, volatile acidity, and density should be able to predict a similar outcome.

The specific wine types for the wines were removed from the data set, specifically which types of white or red wines and which grape cultivars were used in the wine making process. Adding the subtlety of which wine types were present could have more fine tuning to what makes a good quality of each wine style rather than a general good red or white wine. Adding more types of wines of both red and white but also those intermediary wines like rosé wines would be an interesting study.

Generally, these analyses were performed on wines that were from a particular region in Portugal. It would be interesting to apply these results to wines of other regions of similar and different geological, geographic, and climate conditions to see whether the physicochemical properties would still produce the same profile for predicting good quality.