# Introduction

The production and certification of wine is currently a complex process. There are many properties that can be controlled in the production of wine such as climate and weather for grape flavor and juiciness, the amount of sugars, salts, acidity, density, alcohol content and many others. Yet there is still much variability in the determination of the quality of wine. The quality of wine is subjective as wine rankings are scores based upon “highly” certified and specialized wine critics who rely on their taste buds and expertise. Consequently, the opinion and assessment of wine critics on wine creates a high degree of variability in the scoring process. Therefore, our goal is to create a predictive model that can predict the quality of wine based on their physicochemical characteristics.Knowing how each physicochemical feature will impact wine quality will help producers, distributors, and businesses in the red wine industry better assess their production, distribution, and pricing strategy.

# Description of Data

Our goal for this project is to conduct an analysis on Wine Quality data from the University of California, Irvine Machine Learning Repository.Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The data contain 4897 observations in white wine and 1599 observations in red wine data with 11 features in each. The physicochemical features are lab based. Quality is an ordinal variable with a possible ranking from 1 (worst) to 10 (best). Each variety of wine is blind tasted by three independent tasters and the final rank assigned is the median rank given by the tasters.  There are no missing or null values. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).Data is available at: https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

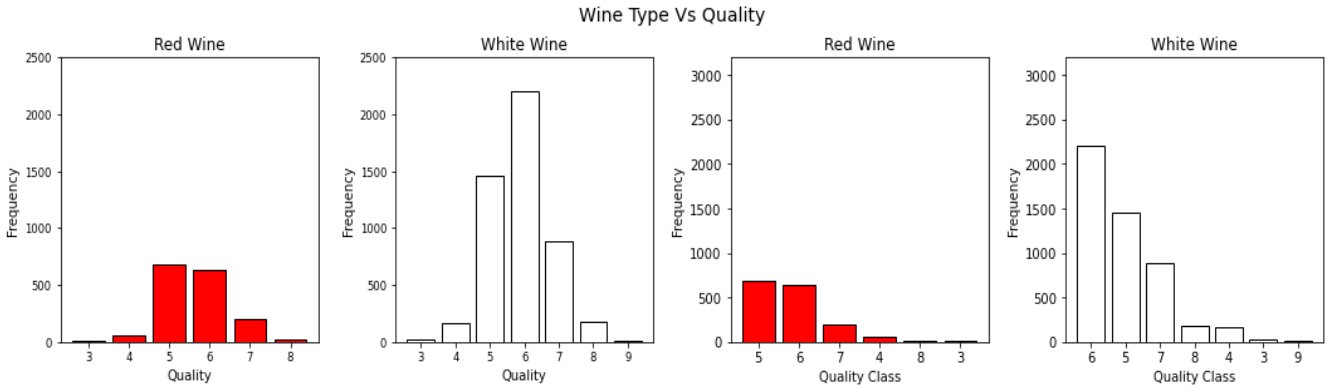
Features :

- wine type - 1599 Red and 4897 White wine
- fixed acidity - Most acids involved with wine are fixed or nonvolatile (Tartaric Acid - g/dm^3)
- volatile acidity - The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste (Acetic Acid - g/dm^3)
- citric acid - acts as a preservative to increase acidity (small quantities add freshness and flavor to wines) (g/dm^3)
- residual sugar - The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet (g/dm^3)
- chlorides - The amount of salt in the wine (Sodium Chloride - mg/dm^3)
- free sulfur dioxide - The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine (mg/dm^3)
- total sulfur dioxide - Amount of free and bound forms of S02 becomes evident in the nose and taste of wine (mg/dm^3)
- density - the density of water is close to that of water depending on the percent alcohol and sugar content (g/cm^3)
- pH - Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- sulphates - a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant (Potassium Sulphate g/dm^3)
- alcohol - the percent alcohol content of the wine
- quality - score between 0 and 10 given by wine experts

# Data Exploration

To explore the data we will first analyze the distribution of each feature to check for outliers or abnormal distributions. Next we will compare each feature to a target variable, quality, in order to isolate the features that will be most helpful in constructing a predictive model. We will then check for multicollinearity. Finally we will select which features to be used for our predictive model.

## Univariate Analysis




We will begin by taking a look at the distribution of the target variable,'Quality', in Red wine and white wine. First we know that we have 4898 white wine observations and 1599 red wine observations which is roughly a 3 to 1 ratio. Both are normally distributed. We can see that we have a large amount of 5, 6, and 7 versus the lower qualities 3 and 4 and higher qualities 8 and 9. For red wine we only have 63 observations below 5, and 18 observations with quality 8. For white wine only 183 below quality 5 and 180 above quality 7. It will be important to keep this data imbalance in mind as we move forward with modeling. With this in mind, it can be a good idea to create a new bucket variable for quality labeled poor, normal, excellent, where poor is quality 3 and 4, normal being quality 5, 6, and 7, and excellent being quality 8 and 9.

## Bivariate Analysis

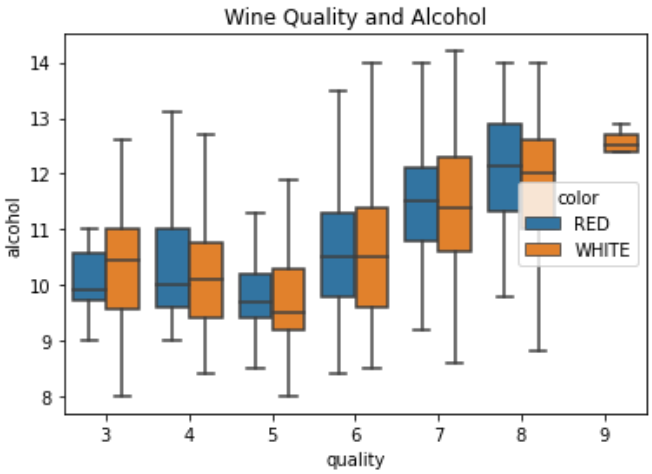



*Fig 5. [% of alcohol compared to the quality of red and white wine]*

**Context:**

Red wines are made of grapes that are usually harvested late in the season. These grapes have more sugar than the grapes used in white wines, so fermentation leads to a higher concentration of alcohol. Since alcohol is a product of fermentation, the riper the grapes are at the moment when yeast converts grape sugar into alcohol, the higher the wine's alcohol level is likely to be.

**Findings:**

As seen in Fig 5. Both red and white wines tend to have higher quality as the alcohol percentage increases. All the data points for white wine quality 9 have an alcohol percentage of 12% or more, while the lower qualities 3 and 4 tend to have lower concentrations of alcohol. For red wine a very similar pattern is observed, but we do not have any points for qualities higher than 8. It is interesting how alcohol does not give us much variation to distinguish whether the vine is white or red, but it makes a lot of difference in quality. Note that the higher the quality the higher the average alcohol concentration, increased by about 1% at each level. Although lower quality wines have the lowest standard deviation.

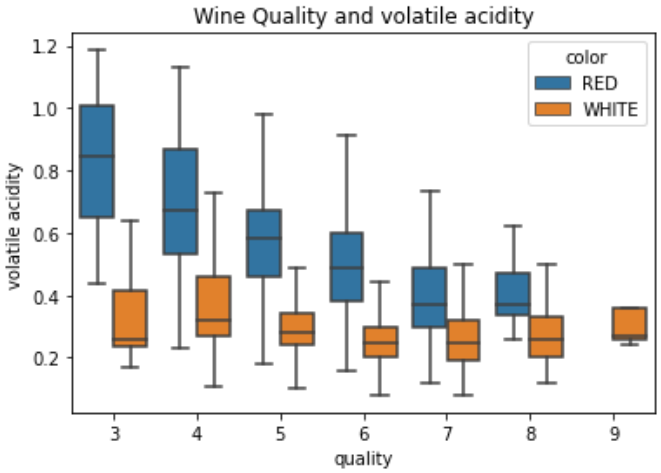



*Fig 6. [volatile acidity compared to the quality of red and white wine]*

**Context:**

Acetic acid in wine, often referred to as volatile acidity, or vinegar taint, and can be contributed by many wine spoilage yeasts and bacteria as well as oxygen. This can be from either a by-product of fermentation, or due to the spoilage of finished wine. The sensory threshold for acetic acid in wine is >700 mg/L, with concentrations greater than 1.2-1.3 g/L becoming unpleasant. There are different opinions as to what level of volatile acidity is appropriate for higher quality wine. Although too high a concentration is sure to leave an undesirable, 'vinegar' tasting wine, some wine's acetic acid levels are developed to create a more 'complex', desirable taste. The renowned 1947 Cheval Blanc is widely recognized to contain high levels of volatile acidity.

**Findings:** the higher the quality of wine, the lower the median volatile acidity for both wines. The spread begins to narrow to a lower concentration for red wines, as better quality red wines tend to be less acidic and less vinegar tasting than lower quality wines. White wines typically have low concentrations of VA. We will need to dive deeper to see why there is a difference in VA for red and white wines.

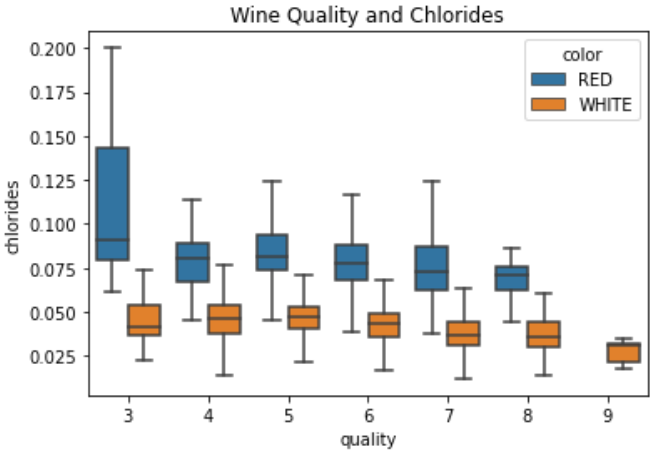



*Fig 7. [chlorides compared to the quality of red and white wine]*

**Context:**The average wine bottle contains 20-40mg per bottle of sodium, along with some organic acids, and they may have a key role on a potential salty taste of a wine, with chlorides being a major contributor to saltiness. The reason is salt can magnify the influence of tannins in red wines and acidity in high acid white wines. The result is to influence the balance equation making the wine taste slightly more astringent, more acidic and therefore, less sweet

**Findings:** A large amount of outliers are observed in the chloride values in poor red wine quality. As quality gets better, chloride levels stay around 7.5mg/L. White wines tend to have less chlorides on average than red wines, and chloride levels on average are very similar across all white wine qualities. It would be interesting to see why such outliers exist in red wine and if there is a relationship to quality as outliers affect correlation between variables

.

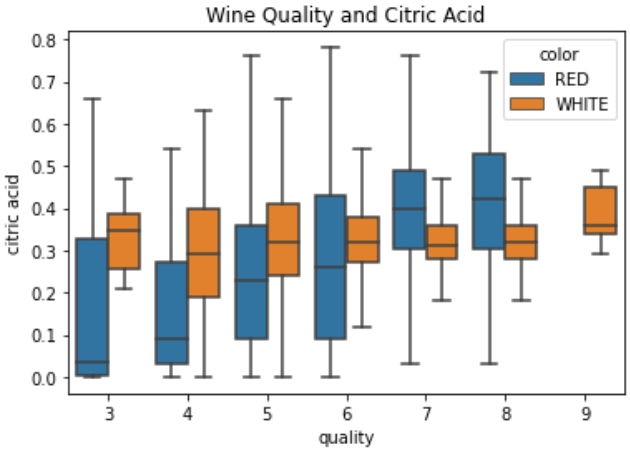



*Fig 8. [Citric Acid compared to the quality of red and white wine]*

**Context:**

Citric acid is often added to wines to increase acidity, complement a specific flavor or prevent ferric hazes. It can be added to finished wines to increase acidity and give a "fresh" flavor. The disadvantage of adding citric acid is its microbial instability. Since bacteria use citric acid in their metabolism, it may increase the growth of unwanted microbes. Often to increase acidity of wine, winemakers will add tartaric acid instead.

**Findings:**

Better red wines tend to have a higher median citric acid amount compared with lower quality wines. Specifically, wines with a rating > 6. As For white wine, as quality improves, spread in citric acid becomes more concentrated around 0.3 g/dm^3 to 0.5g/dm^3. We see a slightly positive correlation with citric acid and wine quality.

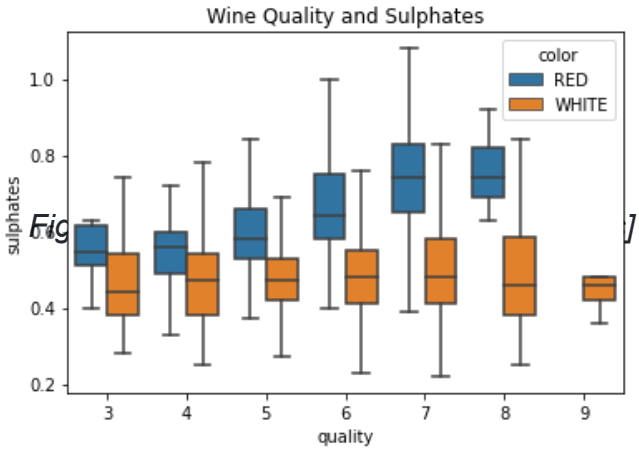



*Fig 9. [Sulphates compared to the quality of red and white wine]*

**Context:**
Wine is fermented using yeast, which produces sulphates as a byproduct, so almost all wine contains sulfites. Winemakers have been adding sulfur dioxide to wine since the 1800s. It has several effects on the winemaking process, including: Protecting against oxidation, which can affect the color and taste of wine. Wines with higher sugar content tend to need more sulfites to prevent secondary fermentation of the remaining sugar. Wines that are warmer release free sulfur compounds (the nasty sulfur smell) and can be “fixed” through decanting and chilling the wine.Wines with lower acidity need more sulfites than higher acidity wines. At pH 3.6 and above, wines are much less stable, and sulfites are necessary for shelf-life. Wines with more color (i.e., red wines) tend to need less sulfites than clear wines (i.e., white wines).

**Findings:**
For the whitewines, there are much lower concentrations of sulfur dioxides. Additionally, there is a slightly positive correlation with sulfates and wine quality. For qualities of 9 it's important to take note of the sulphate concentration of 0.5g/dm^3. For Red Wines, we see a trend in high quality red wines having sulphate concentration around 0.8g/dm^3 with sulphates increasing as quality increases.

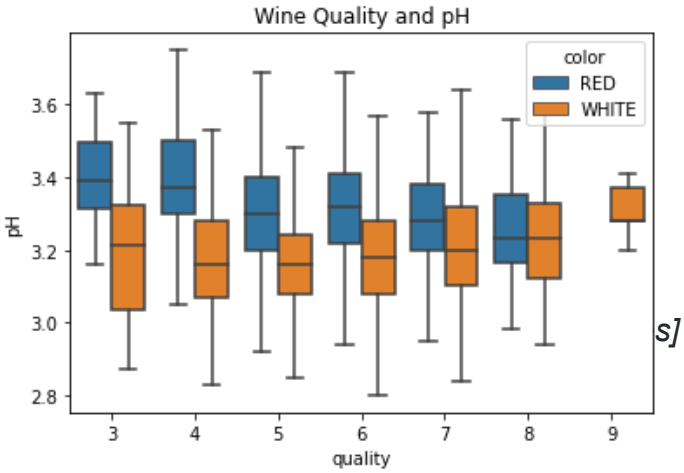



*Fig 10. [pH compared to the quality of red and white wine]*

**Context:**
Wine's chemical and biological stability are very dependent on pH value. Winemakers use pH as a way to measure ripeness in relation to acidity. Lower pH (high acidity) values are known to improve the stability and taste tart and crisp. Even slight changes in pH can influence a wine's taste, color, and smell. Wines with lower acidity (and thus higher pH numbers) are also more prone to bacterial problems, so winemakers usually

prefer a pH range of 3.0 to 3.5 with Most wine pH's fall around 3 or 4; about 3.0 to 3.4 is desirable for white wines, while about 3.3 to 3.6 is best for reds

**Findings:**
We see our findings are consistent with the context. Typical white wines pH lie between 3-3.4 with quality 9 being concentrated near 3.3-3.4. Red wines have a higher pH on average than white.

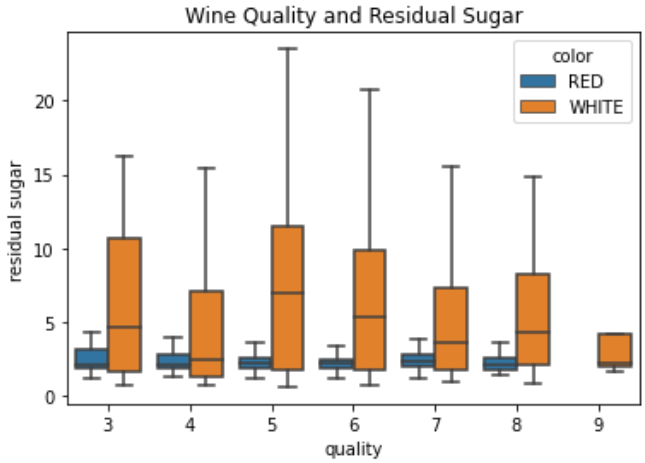


*Fig 11. [Total Residual Sugar in different wine qualities]*

**Context:**
Residual sugar consists mostly of the grape sugars which is a blend of glucose and fructose that are left over after the fermentation process but can also be created by small amounts added before bottling. During the fermentation process, yeast eats these sugars to make alcohol. That being said, it's possible to stop the fermentation before all the sugar gets consumed (through chilling or filtration). Residual sugar levels vary in different types of wine. In fact, many grocery store wines labeled as "dry" contain about 5 g/L of residual sugar. Noticeably sweet wines start at around 35 grams per liter of residual sugar and then go up from there. There are no sweet wines in this data.

**Findings:**
We see that in Fig 11, red wines are typically dry and have lower concentrations of residual sugar than white wines. Red wines tend to have residual sugar concentrations around 2.5grams/ liter. It is important to note in quality 9 lower concentrations of residual sugar in white wine and the large amount of outliers may affect correlation. White wine has greater variation, a greater median and one unusually large, sweet outlier. White wine still has a group with very low sugar levels. Perhaps these are dry white wines.

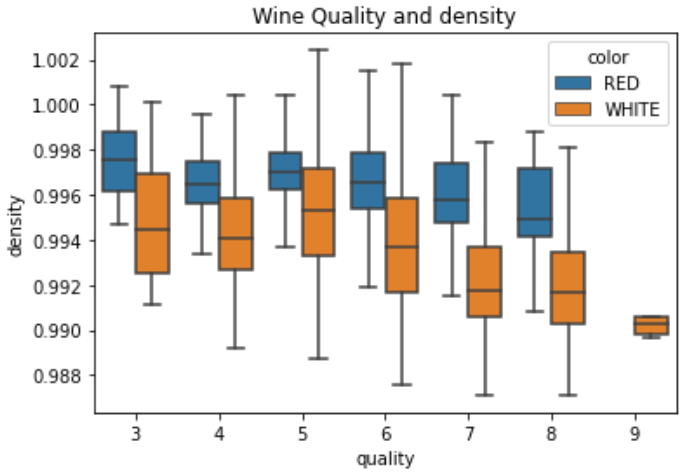



*Fig 12. [Density in different wine qualities]*

**Context:**
After grapes are harvested and crushed, the grape juice ferments in tanks for 7 to 14 days. During this time, the yeast present in the must (freshly crushed whole fruit juice) metabolize the sugar in the grape juice, resulting both in their proliferation and in the waste products ethanol and CO2, along with flavors.

To monitor this process, winemakers typically measure the density or specific gravity (SG) over time, determining a fermentation curve; this gives an insight into the fermentation conditions and their favorability to yeast growth

**Findings:**
Density seems to be correlated with multiple variables. The amount of sugar and alcohol in a wine affect the density. If sugar content is high,then the density will increase. The density of alcohol is smaller than water. Red wines typically have higher density than white wines. There are a lot of outliers in white wine density with a small concentration of quality 9 at .992 density

## Correlation Analysis

We will need to dive deeper into each feature to see which features are correlated with each other and its relationship to the quality of wine.

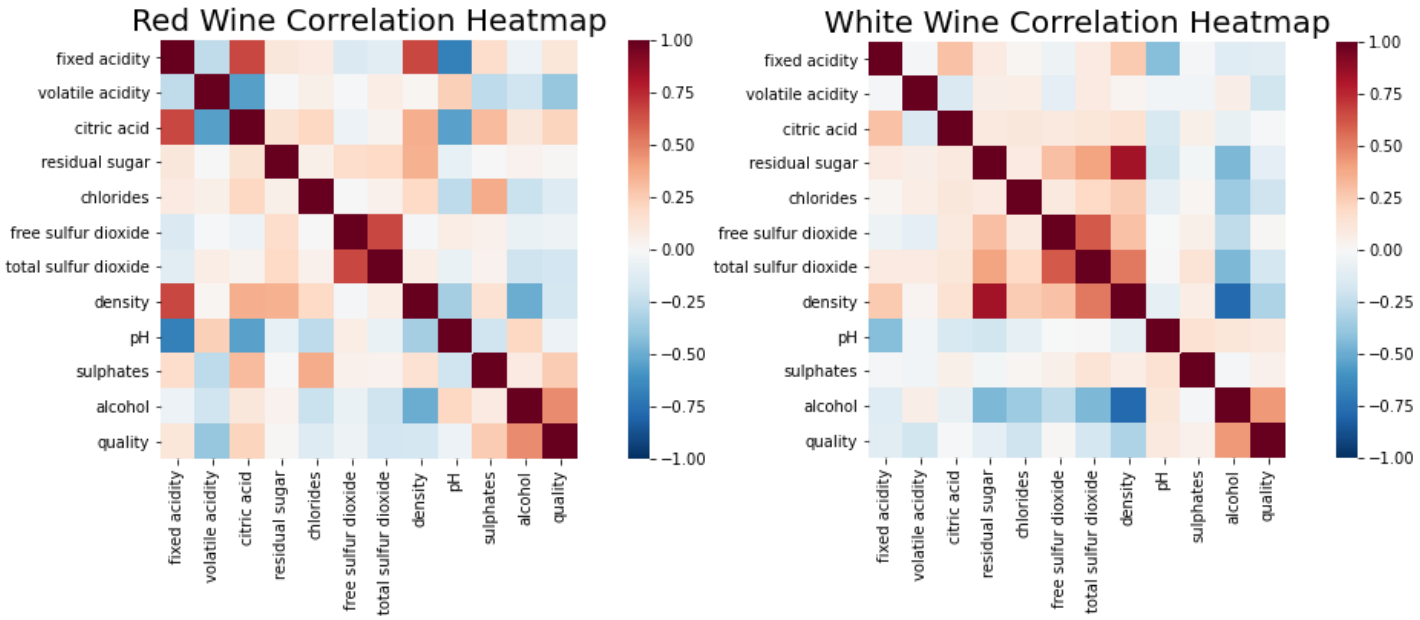



*Fig 13. [ Wine Correlation Matrix]*

Observations About The Correlation Analysis:

1. First we can observe for both wine qualities that quality is highly correlated with alcohol. However, alcohol and density are negatively correlated. Therefore, one of them can be used as a wine quality predictor. Moreover, it's the alcohol amount that reduces the density, due to chemistry, hence alcohol amount is a good choice as a wine quality predictor.
2. Both Wine quality are negatively correlated with the volatile acidity, as too high levels of it leads to vinegary taste, supporting the description about the data set.
3. As suspected from our previous findings, free SO2 and total SO2 are highly correlated with each other and negatively correlated with the volatile acidity.
4. pH is negatively correlated with fixed acidity, citric acid, total SO2 and residual sugar. The negative correlation with the residual sugar makes sense, since sugar turns into alcohol. Moreover, pH is positively correlated with the volatile acidity, which is a bit counter-intuitive as higher pH is typically lower acidity.
5. Residual sugar and density are also positively correlated, which makes sense, adding sugar increases density.
6. According to the description, sulphates are added to produce SO2 which acts as antimicrobial and antioxidant; total SO2 is also added for the same purpose, but why then they are negatively correlated, perhaps one is converted into another.

7. It is surprising to see a positive correlation between total SO2 and residual sugar, more SO2 is added to prevent sugar from being converted, and thus make sure that wine tastes a bit sugary.
8. It is nice to see volatile acidity is negatively correlated with SO2, as SO2 is added in the wine to prevent acetic acid formation.
9. As fixed acidity increases in Red Wine, the perceived quality seems to also increase and the opposite seems to happen with White Wine. As stated here, Red Wine tends to contain more tanins which add more bitterness and astringent qualities to wine. This combination of acidity and tanins could be well perceived by wine tasters. The lack of tanins in White Wine could make the perceived acidity in White Wine more apparent and therefore undesired at high levels. This could explain why winemakers do not seem to produce White Wine with fixed.acidity beyond 12g/dm3 and hence the lack of correlation with quality.
10. With white wine, it is interesting to note that alcohol is more negatively correlated with more variables than in red wine.Density, Chlories, and SO2 are much more correlated in white wine than red wine as well. There seems to be more multicollinearity in white wine than in red wine.

In Summary, When trying to extrapolate patterns in the physicochemical properties of wine, it is better to look at Red and White wines separately. Even though they share a lot of similar relationships, it is clear that in some cases, what works for one does not work for the other. We can confirm our findings that as we have higher qualities of wine, we have a higher concentration of alcohol. The correlation between alcohol and sugar content is much higher for Red wines than it is for white wines (boozy reds have more sugar than less boozy reds, while boozy whites have less sugar than less boozy whites). It is reasonable that less sweet wines and a lower level of acidity are favored in quality testing. In general, white wines exhibit more acidity than red wines. Acidity gives wine its crispness on the palate. A dry wine needs good levels of acid to provide liveliness and balance; sweet wine needs acidity so it does not seem distasteful. White wines mostly have a dry, crisp, fruity flavor. While red wines tend to have a rich and bitter taste. But the main difference is in the process of making wine. White wines are fermented without the grape skins, thus giving it that light, sweet taste.

## Feature Importance

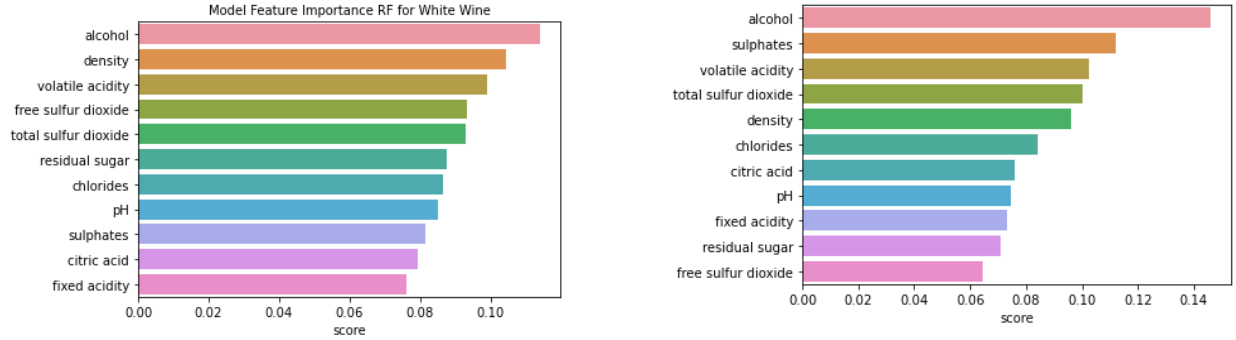



*Fig 14. [ Wine Feature Importance Graph]*

Diving deeper into each feature, we can see with the Random Forest model We can see that alcohol is the most important feature for quality in both red and white wines. Density is an important feature in white wine, while sulphates has more importance in red wine. Lastly Volatile acidity is important in both wines as well. We could see how the multicollinearity of some features affects our models going forward by dropping some less important features such as fixed acidity, residual sugar, and others.

## Pre-processing & Training

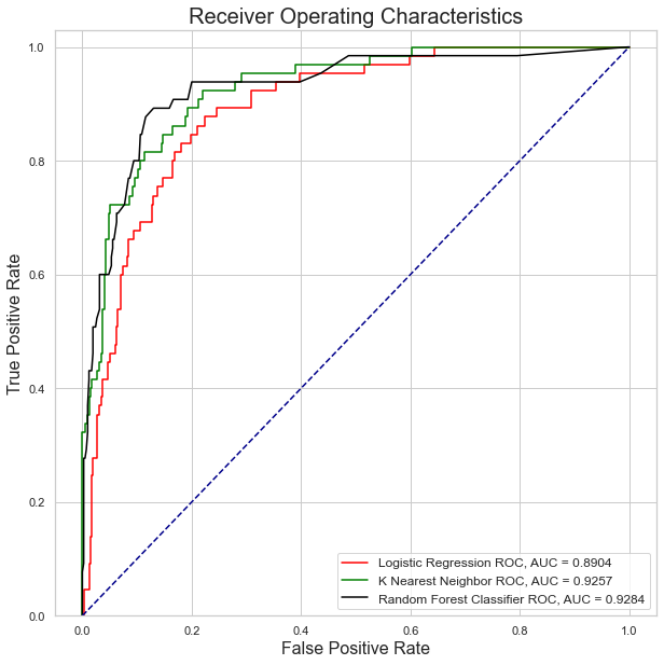
Machine learning techniques are used to predict wine quality in this study. Pre-processing is done on both wine datasets. The data is further divided into training (80%) and testing ( 20%) sets, with the training set being utilized to train the model utilizing Logistic Regression, Random Forest, and K-Nearest Neighbors Machine Learning Algorithms. The datasets are the imbalanced distribution of red wine and white wine where the separate classes are not equally represented. This imbalanced data can lead to overfitting and underfitting algorithms. The red wine's highest quality class 5 instances are 681 and white wine highest quality class 6 instances are 2198. Both datasets are unbalanced with the number of instances ranging from 5 in the minority class up to 681 in red wine and ranging from 6 in the minority class up to 2198 in the majority class. The highest quality scores are rarely paralleled to the middle classes. A good way to deal with the imbalanced datasets by applying the supervised synthetic minority oversampling technique SMOTE filter. SMOTE is an over-sampling technique in which a lesser amount of classes in the training set is over-sampled and creates the new sample form to relieve the class imbalance. Therefore, to solve the data imbalanced problem we used the SMOTE technique. Applying Smote increased accuracy for each model and reduced the number of mispredictions so we will continue our modeling with the balanced dataset.

## Model Selection - Red Wine

Compared 3 models: Logistic Regression, Random Forest, and K-Nearest Neighbors, grid searched for respective hyperparameter space for values that maximized algorithm performance and ordered relative performance of the algorithms by calculating the Receiver Operating Characteristic Area Under the Curve (ROC AUC) scores. Random Forest performed the best with a ROC AUC Score of and overall accuracy of

| **TABLE 1: ALGORITHM PERFORMANCE - RED WINE** | | |
|---|---|---|
| **Classifier** | **ROC-AUC Score** | **Best Hyperparameter Value** |
| Logistic Regression | 0.864 | C = 0.1 |
| K-Nearest Neighbors | 0.908 | N_neighbors = 44,<br>Weights = distance |
| Random Forest | 0.912 | Max_depth = 25<br>Max_features = 3,<br>Min_samples_leaf: 1<br>Min_samples_split: 2<br>N_estimators: 300 |

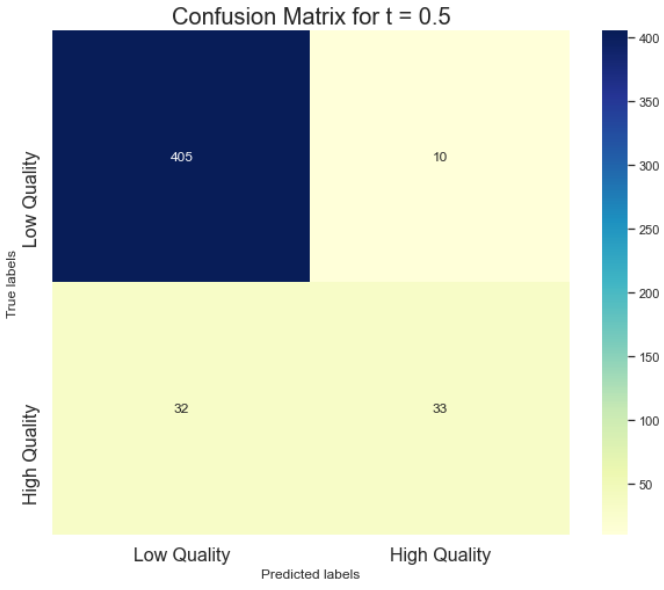
## Model Performance




The Random Forest model generated higher precision scores (0.80), compared to recall (0.51) and the weighted average f1 score (0.91). These results indicated the standard probability threshold of 0.5 generated fewer false positives compared to false negatives. Thus the model mislabeled bad quality wines as good quality (False Positives) more frequently than the model mislabeled good qualities wines as bad quality.

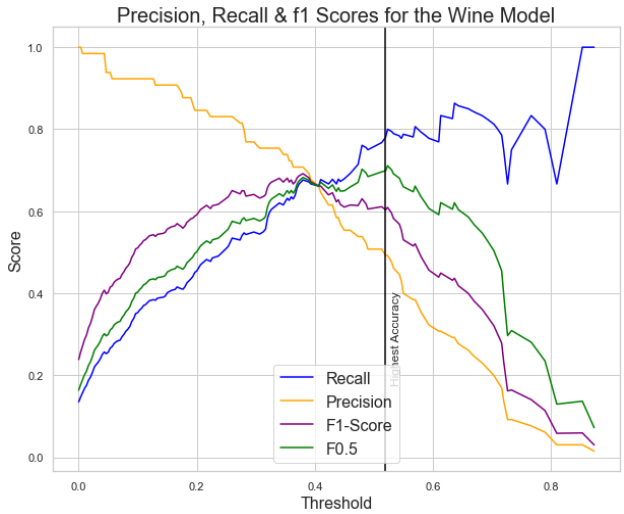
| | Low Quality | High Quality | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| precision | 0.938967 | 0.722222 | 0.914583 | 0.830595 | 0.909616 |
| recall | 0.963855 | 0.600000 | 0.914583 | 0.781928 | 0.914583 |
| f1-score | 0.951249 | 0.655462 | 0.914583 | 0.803355 | 0.911194 |
| support | 415.000000 | 65.000000 | 0.914583 | 480.000000 | 480.000000 |

### Model Optimization - Adjusting the Threshold



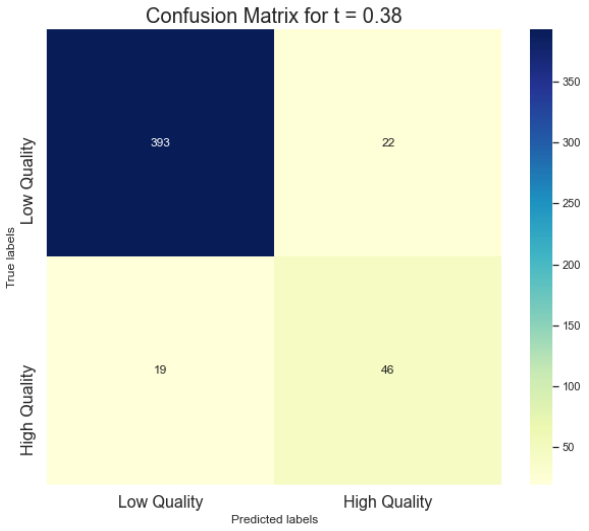

The default cutoff probability of 0.5 for threshold of individual cases lends equal weight to both precision and recall. However, different scenarios potentially favor one over the other. Thresholding involves modeling a range of potential binary classification cutoffs values and observing the effects of associated classification metrics. The F-beta metric incorporates a beta parameter which lends emphasis to either precision or recall. When beta is set to one, F-beta is equivalent to the f1 score and both recall and precision are emphasized equally. When we care more about minimizing false positives than minimizing false negatives, we would want to select a beta value of < 1 for the F-beta score.On the other hand, when the priority is to minimize false negatives, we would want to select a beta value of >1 for the F-beta score.

## Practical Business Applications - Red Wine




### Scenario 1 - Minimize Model's probability of mislabeling bad wines

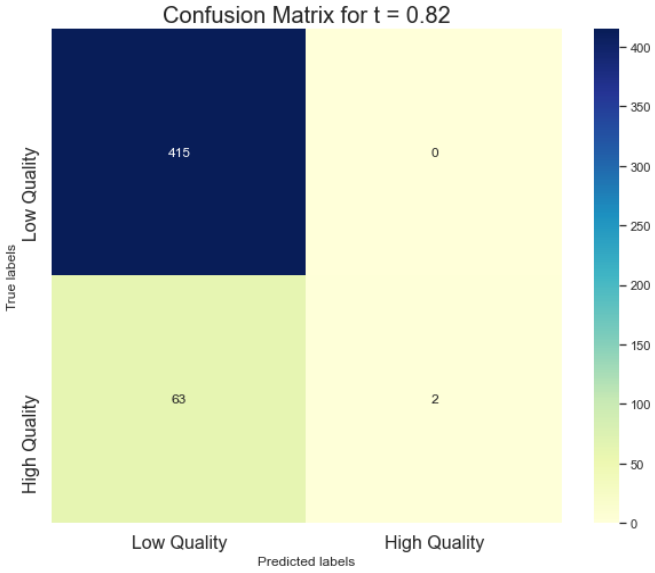
I explored the wine quality's model's capability to real-world business applications associated with predictive modeling. Essentially we would want to look at the ability of our model to accurately identify good quality wines and bad quality wines for wine manufacturers




and distributors to efficiently and effectively be able to output only those good quality wines.

In order to avoid overlooking good wine, we would want to focus on minimizing false positives. In order to simulate such a scenario, I adjusted the beta parameter of the F-beta metric to emphasize precision (Beta < 1) using beta = 0.5 and then quantified F-beta for probability threshold values ranging between 0-1. The threshold value producing the largest F-beta score of 0.9146 is .52. This resulted in higher false positives by 12, increase in True high quality wines by 13 and slightly lower false negatives 13.

## Scenario 2 - Maximizing Model's performance to detect High Quality Wine




Let's say we have a list of wines that we would like to start manufacturing and we want to make sure we are only producing the best wine possible. We would want to make sure we are predicting high quality wines as high and making sure no low quality wines are being misabled as high quality. In this instance, we also would not care for False negatives as we want to make sure we are only serving the best wine. To do so I examined the trade off between precision and recall and found the optimal threshold for this case would be 0.82.

| | 0 | 1 | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| precision | 0.868201 | 1.000000 | 0.86875 | 0.934100 | 0.886049 |
| recall | 1.000000 | 0.030769 | 0.86875 | 0.515385 | 0.868750 |
| f1-score | 0.929451 | 0.059701 | 0.86875 | 0.494576 | 0.811673 |
| support | 415.000000 | 65.000000 | 0.86875 | 480.000000 | 480.000000 |

The 0.90% probability threshold would yield a result where we achieve 100% precision for high quality wines and 100% recall for low quality wines. This would allow us to predict only the highest quality of wines.

## Model Selection - White Wine

Compared 3 models: Logistic Regression, Random Forest, and K-Nearest Neighbors, grid searched for respective hyperparameter space for values that maximized algorithm performance and ordered relative performance of the algorithms by calculating the Receiver Operating Characteristic Area Under the Curve (ROC AUC) scores. Random Forest performed the best with a ROC AUC Score 0.909 (TABLE 2) of and overall accuracy of 0.883

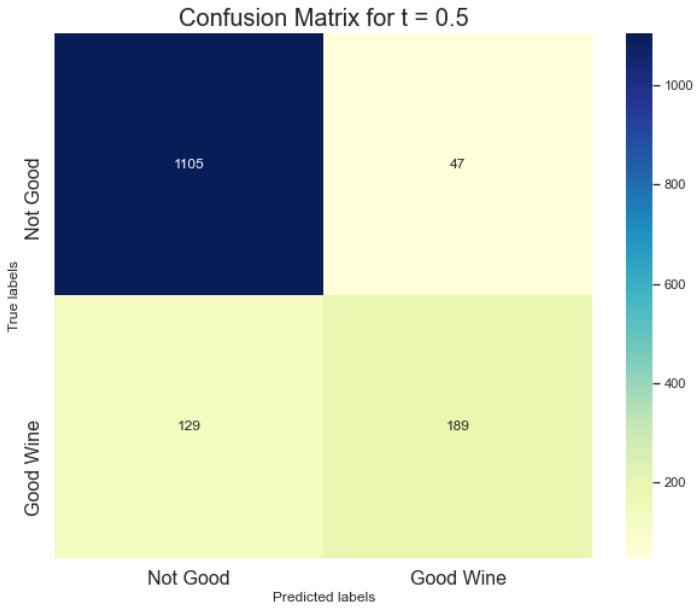
**TABLE 2: ALGORITHM PERFORMANCE - WHITE WINE**

| Classifier | ROC-AUC Score | Best Hyperparameter Value |
|---|---|---|
| Logistic Regression | 0.743 | C = 0.1 |
| K-Nearest Neighbors | 0.847 | N_neighbors = 48,<br>Weights = distance |
| Random Forest | 0.909 | Max_depth = 15<br>Max_features = 3,<br>Min_samples_leaf: 1<br>Min_samples_split: 2<br>N_estimators: 500 |

## MODEL PERFORMANCE


The Random Forest model generated higher precision scores (0.817), compared to recall (0.591) and the weighted average f1 score (0.686). These results indicated the standard probability threshold of 0.5 generated fewer false positives compared to false negatives. Thus the model mislabeled bad quality wines as good quality (False Positives) more frequently than the model mislabeled good qualities wines as bad quality.

| | Low Quality | High Quality | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| **precision** | 0.895462 | 0.800847 | 0.880272 | 0.848155 | 0.874994 |
| **recall** | 0.959201 | 0.594340 | 0.880272 | 0.776771 | 0.880272 |
| **f1-score** | 0.926236 | 0.682310 | 0.880272 | 0.804273 | 0.873469 |
| **support** | 1152.000000 | 318.000000 | 0.880272 | 1470.000000 | 1470.000000 |

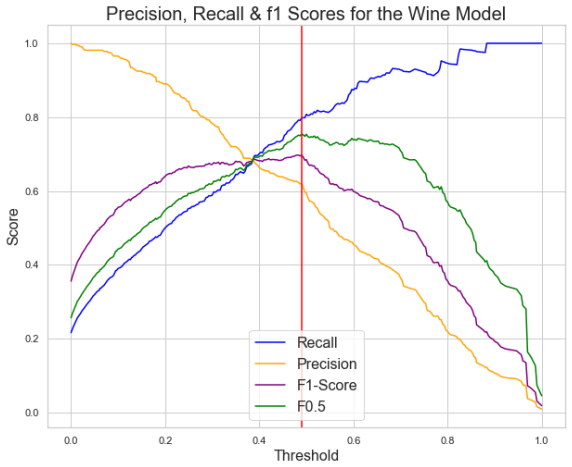



### Model Optimization - Adjusting the Threshold

The default cutoff probability of 0.5 for threshold of individual cases lends equal weight to both precision and recall. However, different scenarios potentially favor one over the other. Thresholding involves modeling a range of potential binary classification cutoffs values and observing the effects of associated classification metrics. The F-beta metric incorporates a beta parameter which lends emphasis to either precision or

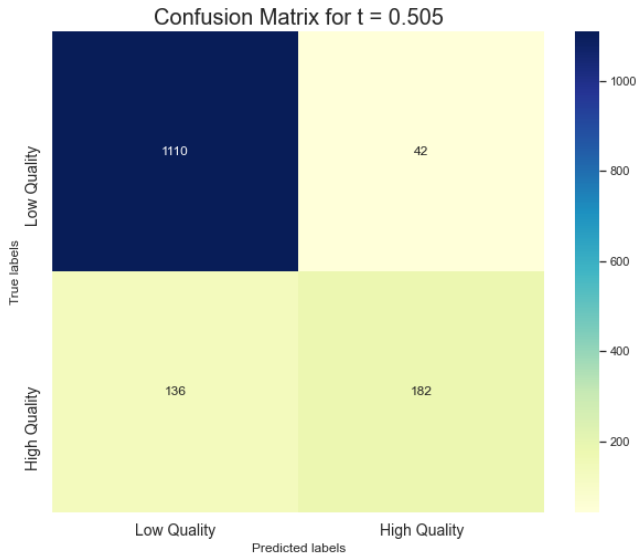
recall. When beta is set to one, F-beta is equivalent to the f1 score and both recall and precision are emphasized equally. When we care more about minimizing false positives than minimizing false negatives, we would want to select a beta value of < 1 for the F-beta score.On the other hand, when the priority is to minimize false negatives, we would want to select a beta value of >1 for the F-beta score.

## Practical Business Applications - White Wine



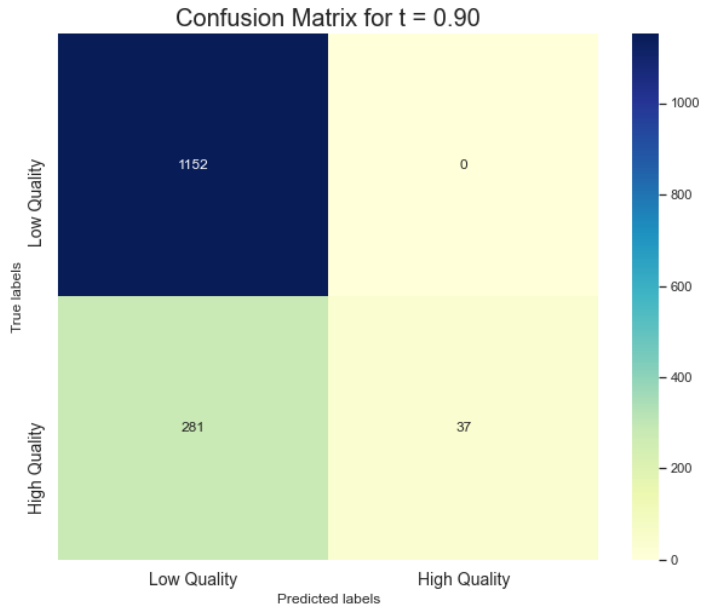

### Scenario 1 - Minimize Model's probability of mislabeling bad wines

I explored the wine quality's model's capability to real-world business applications associated with predictive modeling. Essentially we would want to look at the ability of our model to accurately identify good quality wines and bad quality wines for wine manufacturers and distributors to efficiently and effectively be able to output only those good quality wines.




In order to avoid overlooking good wine, we would want to focus on minimizing false positives. In order to simulate such a scenario, I adjusted the beta parameter of the F-beta metric to emphasize precision (Beta < 1) using beta = 0.5 and then quantified F-beta for probability threshold values ranging between 0-1. The threshold value producing the largest F-beta score of 0.8789 is .505. This resulted in lower false positives by 5, increase in True high quality wines by 2 and slightly higher false positives by 7. Precision played more of a factor at .8125 than the prior default threshold 0.5 at .801.

## Scenario 2 - Maximizing Model's performance to detect High Quality Wine




Let's say we have a list of wines that we would like to start manufacturing and we want to make sure we are only producing the best wine possible. We would want to make sure we are predicting high quality wines as high and making sure no low quality wines are being misabled as high quality. In this instance, we also would not care for False negatives as we want to make sure we are only serving the best wine. To do so I examined the trade off between precision and recall and found the optimal threshold for this case would be 0.90.

| | Low Quality | High Quality | accuracy | macro avg | weighted avg |
|---|---|---|---|---|---|
| **precision** | 0.803908 | 1.000000 | 0.808844 | 0.901954 | 0.846328 |
| **recall** | 1.000000 | 0.116352 | 0.808844 | 0.558176 | 0.808844 |
| **f1-score** | 0.891296 | 0.208451 | 0.808844 | 0.549873 | 0.743578 |
| **support** | 1152.000000 | 318.000000 | 0.808844 | 1470.000000 | 1470.000000 |

The 0.90% probability threshold would yield a result where we achieve 100% precision for high quality wines and 100% recall for low quality wines. This would allow us to predict only the highest quality of wines.

# Summary

Before considering any results which may be proposed, the limitations of the scope of this evaluation must be understood. First, it is important to note that the data set consists of wine metrics drawn from a selection of 6499 Portuguese “Vinho Verde” wines, of which 4899 bottles are white wine, and 1600 bottles are red wine, so any conclusions drawn are only applicable to this particular variety of wine. Also, because wine tasting is a qualitative, and to a degree subjective, judgment by a selection of professional sommeliers, the models developed in this research will be biased toward modeling the tastes of the particular sommeliers who generated the data, which may differ from the ratings which would have been given by different sommeliers.

In data exploration and visualization we look for features that may provide good prediction results. The best predictors have low distribution overlapping area and low correlation among them. For red wine, the best predictive features are Alcohol, density, and volatile acidity. For white wine, Additionally, wines with a low volatile acidity, low amounts of chlorides, high total sulfur dioxide, low density, and high alcohol content, are more likely to be of a higher quality as shown in the preceding visualizations, with alcohol content having the most marked association. This result would seem to line up with anecdotal evidence that a beverage with a high alcohol content quickly begins to taste better and be easier to drink the more one drinks it.. Maybe there are other properties not considered in this dataset that are better indicators for quality.The machine learning models used in this research aren’t able to predict red wine quality with high accuracy, specificity and sensitivity. This is partially explained by the low prevalence of quality levels 3, 4 and 8 and the large distribution overlapping area stratified by quality.

Besides this, the lack of information about how the dataset was created may impact the prediction of quality using the physicochemical properties as predictors. Such examples are the composition of grape varieties in each wine, the mix of experts that evaluated wine quality, or the production year, price range, manufacturer information brand etc.

## Recommendations for Future Implementation

To conclude my project, I used Random Forest to correctly predict as many high and low quality wines that I could using CV, thresholding, plots, and more. However there is still room for improvement.

The default Threshold of 0.5 gave us a good starting point for our model. However for different business cases we may want to adjust the threshold depending on what problem we are trying to solve.

For Scenario 1, the model was optimized to provide the most accurate model with emphasize on predicting high quality wines.The threshold value producing the largest F-beta score (beta= 0.5) of 0.8789 is .505. This resulted in lower false positives by 5, increase in True high quality

wines by 2 and slightly higher false positives by 7. Precision played more of a factor at .8125 than the prior default threshold 0.5 at .801.

For Scenario 2, I wanted to optimize the model for only high quality wine, meaning we want to predict high quality as high quality and no low qualities as high quality. To do this I explored the precision recall tradeoff and found the optimal point to achieve high precision for high quality is closer to a Threshold of 1. A threshold of .90 achieves this by giving us 100% for high quality and 100% recall for low quality with no false positives.

In summary this model can:

1) Look at the physicochemical features of a wine to see what features or combination of features are used to create a high quality wine vs low quality wine.
2) Optimize model to perform for most accurate predictions specifically for high quality.

Resources:

Steen, D. (2020, October 11). *Beyond the F-1 score: A look at the F-beta score*. Medium. Retrieved November 27, 2022, from https://medium.com/@douglaspsteen/beyond-the-f-1-score-a-look-at-the-f-beta-score-3743ac2ef6e3