Wine Quality Prediction

by- Mugdha Paithankar

OVERVIEW

“Beer is made by men, wine by God.” and we all want a sip as divine as its creator! It is probably one of the oldest drinks in the history of mankind. But what gives wines their particular flavor, or color? How is white wine different from red wine? What goes into making a “high quality” wine. These are the questions we aim to find answers for in this project. The analysis will thus enable winemakers to predict how well their wine will be perceived. wine enthusiasts and sellers to select wines based on particular taste preferences or cost.

DATA

The dataset has been obtained from Kaggle and it consists of a comprehensive list of red and white wines features which include volatile acidity, pH, alcohol content, citric acid content, residual sugars, chlorides along with the corresponding quality rating which ranges between 1-10.

But what do these features mean in real life?

1. fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
2. volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
3. citric acid: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

4. 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
5. chlorides: the amount of salt in the wine
6. 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
7. total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
8. density: the density of water is close to that of water depending on the percent alcohol and sugar content
9. 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
10. sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
11. alcohol: the percent alcohol content of the wine.

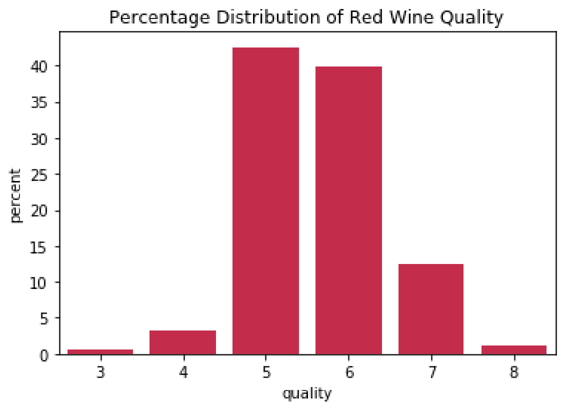
All features are measured in g/dm^3 except for total + free sulfur dioxide and alcohol which is in mg/dm^3 and % by volume respectively.

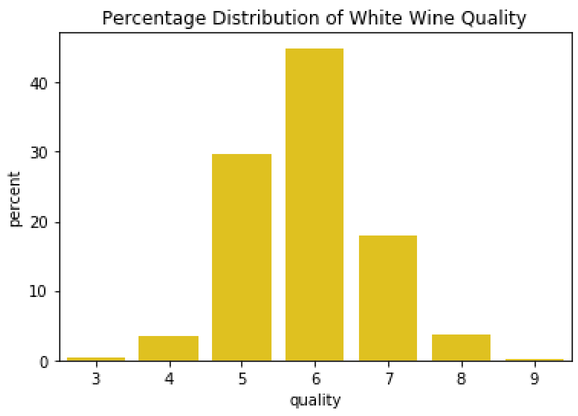
We will be examining red and white wines separately in this report.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) which will help us visualize the trends and behavior of the wine quality dataset and ultimately give us necessary pointers as we dive in getting statistical significance for the features and finally build a model for quality prediction of wines.

- **Bar Charts:** Let us first visualize the quality distribution of these wines.

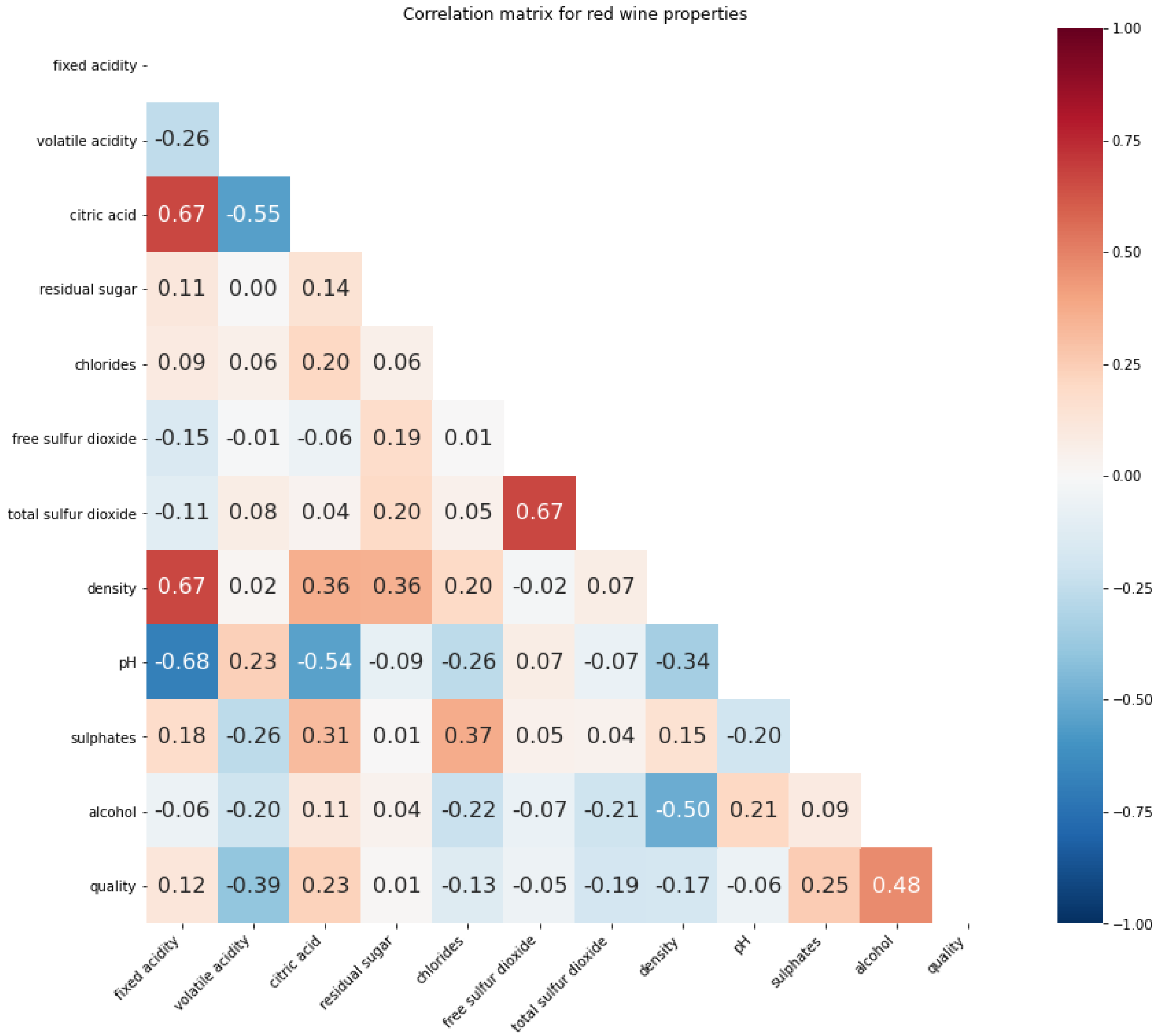




As we can see, a majority of red and white wines (about 85%) have a quality rating of 5 and 6. There are no 9s for red wine, very few wines (less than 1%) with a quality rating of 3, and only 2% (for red wines) with a rating of 8. Since we have very few wines of exceptional quality or poor quality, it might add some bias or make it tricky to build a model which accurately selects wines of high (7 or above rating) or low quality (rating below 3).

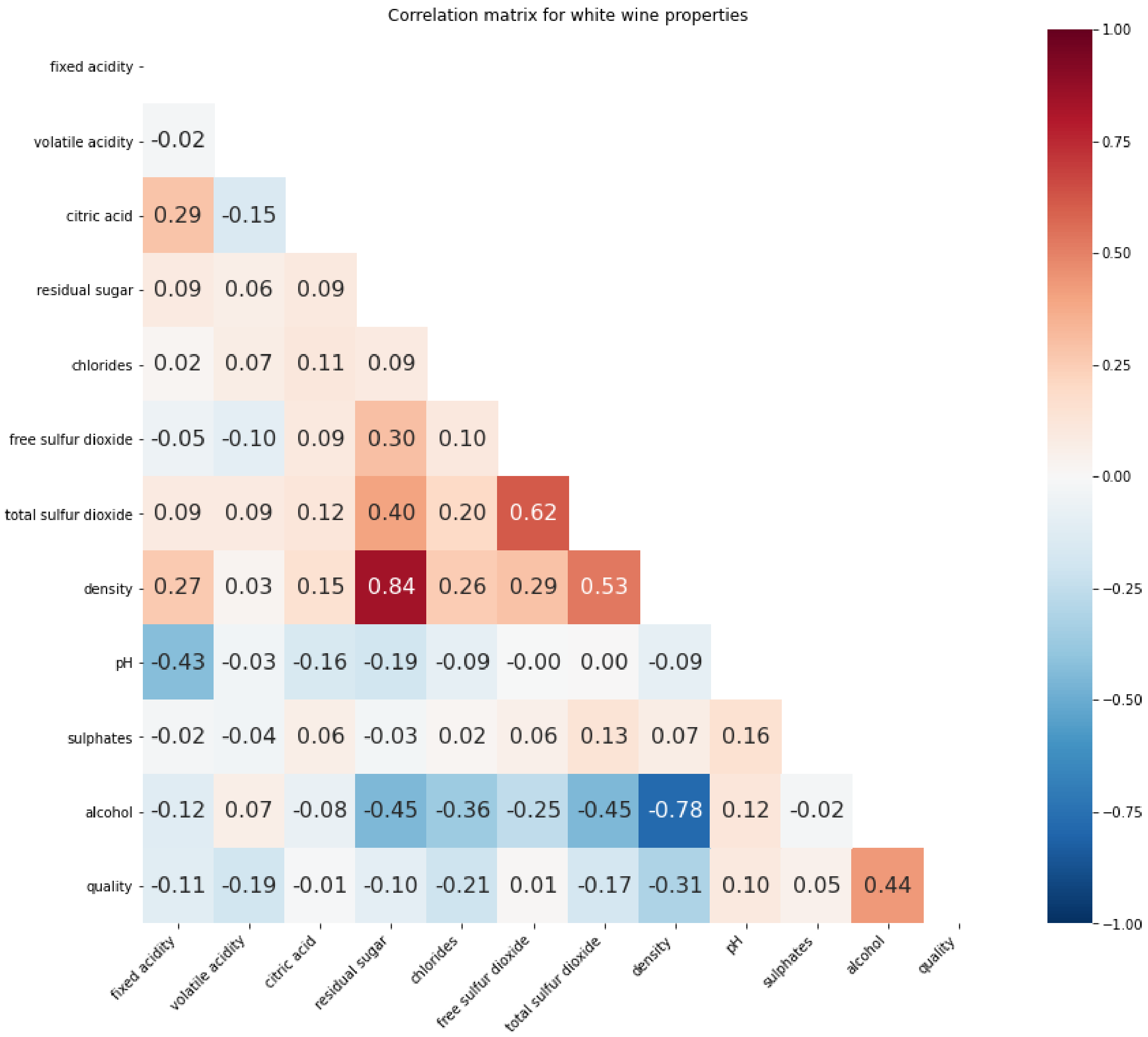
- **Correlation Matrix:** Correlation matrix is a good way to summarize a large amount of data in order to see patterns. We can check if the variables are highly correlated with each other. We can better understand the relationships between our features, which is why I plotted correlation matrices for red and white wines to explore how different features have an effect on the wine quality.

1. Red Wine correlations


Correlation matrix for red wine properties

The fixed acidity feature of red wines seems to have strong/ moderate correlations with pH, density and citric acid. Free sulfur dioxide has strong correlation with total sulfur dioxide, which might be expected. Alcohol has a correlation of 0.48, which is the maximum correlation any feature has with quality! pH has the least correlation of -0.06 with the quality feature of red wines.

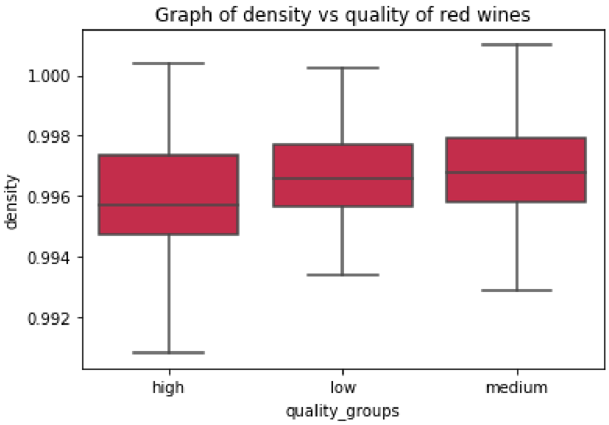
2. White Wine correlations

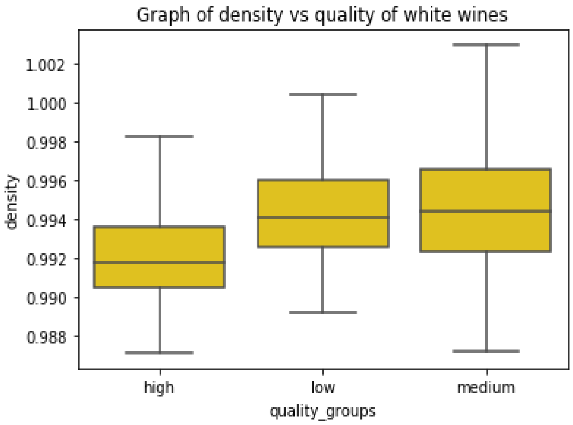


White wines seem to have strong/moderate correlations between residual sugar, alcohol with density, alcohol and residual sugar, pH and fixed acidity, alcohol and density with quality. We can say there is multicollinearity between density, residual sugar and alcohol. (correlation > 0.7)

- **Box Plots and Tukey Test:** Next, we visualize how the quality rating of wines changes with each feature, seeing if we can find any patterns.We try to find answers for questions like: do highly rated wines have greater alcohol content? Do low rated wines have higher acidity? For this reason, we have divided the quality ratings of red and white wines into 3 categories: Low (1-4), Medium (5-6) and High (7-10) and we make boxplots for all possible statistically significant features. Significance testing for the box plots below is done via a Tukey test, which compares the means of all treatments to the mean of every other treatment.

1. Density VS Quality

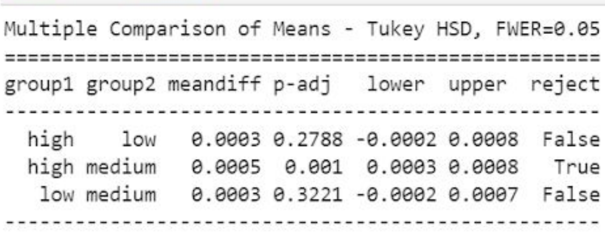




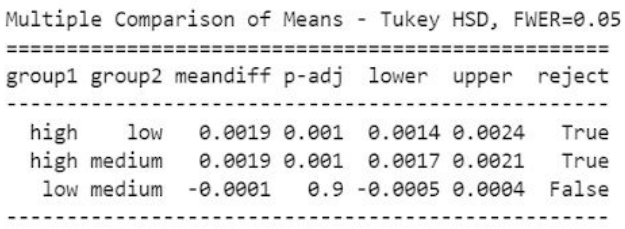
Background: Sweeter wines generally have higher densities. High quality red wines seem to have slightly lesser densities compared to low and medium quality ones. Lesser density could be an indicator of more acidity/lesser pH.

Analysis: The density of white wines does not vary much with respect to quality groups, with high quality white wines having only slightly lower density, as is the case with red wines. There are 3 outliers in medium quality white wines who have significantly higher densities.

We performed the Tukey test to determine if the relationship between the quality groups is statistically significant.



```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj  lower  upper  reject
---------------------------------------------------
  high    low   0.0003 0.2788 -0.0002 0.0008  False
  high medium   0.0005  0.001  0.0003 0.0008   True
   low medium   0.0003 0.3221 -0.0002 0.0007  False
---------------------------------------------------
```



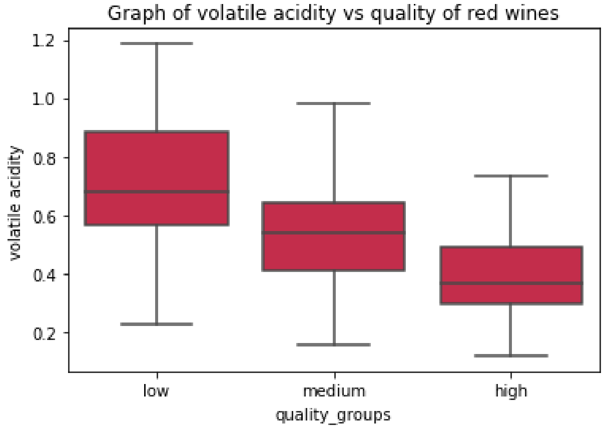
```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj  lower  upper  reject
---------------------------------------------------
  high    low   0.0019 0.001  0.0014 0.0024   True
  high medium   0.0019 0.001  0.0017 0.0021   True
   low medium  -0.0001   0.9 -0.0005 0.0004  False
---------------------------------------------------
```

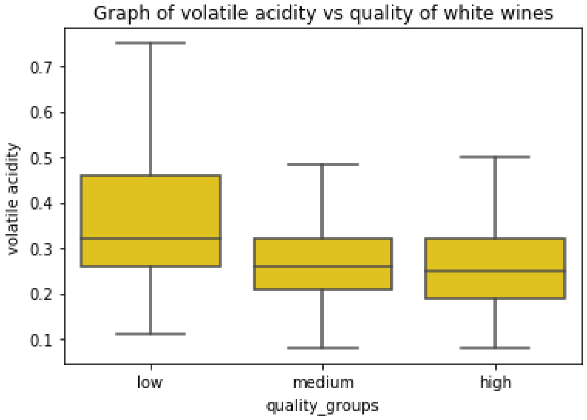
The results of the Tukey test for quality groups of red and white wine for density feature are shown in the figures above respectively.

For red wines, the Tukey test indicates that there could be a statistically significant relationship between the medium and high quality groups of red wine, since the Null hypothesis rejection is true. This indicates that the change in density is more marked for red wines rated between 5-7 (medium) and those rated above 7 (high quality).

Tukey test for white wines indicates that there could be a marked change in densities between the high and low quality white wines (which could be expected) and between high and medium quality white wines too. This might mean that overall high quality white wines seem to have more density.

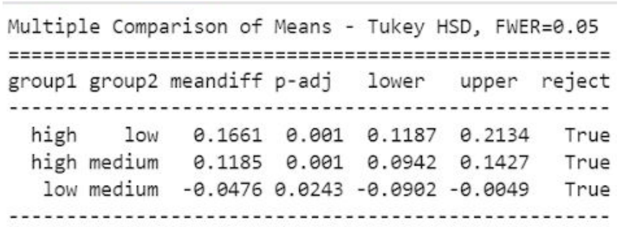
2. Volatile Acidity VS Quality



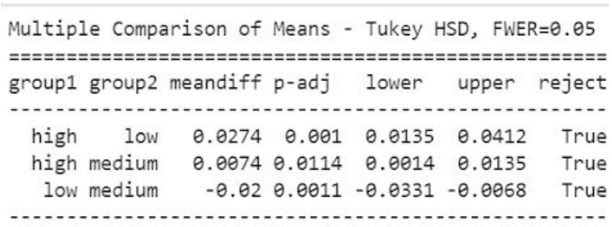


Background: Volatile acidity could be an indicator of spoilage, or errors in the manufacturing processes — caused by things like damaged grapes or wine exposed to air. This causes acetic acid bacteria to enter and thrive, and give rise to unpleasant tastes and smells. It is reasonable to see that volatile acidity content is lesser in high quality red wines. It is most in low quality red wines. There is one outlier in low and high quality groups.

Analysis: The change in volatile acidity with quality groups is not as distinct as it is with red wines. For white wines, the volatile acidity is only slightly higher for low quality compared to medium quality wines. Even with high quality white wines the volatile acidity is not significantly lower but only slightly lower with fewer outliers. It is said that wine experts can often tell the volatile acidity just by smelling it. However, it seems that in the case of white wines, it might be a bit difficult to distinguish because of these smaller changes. All groups were found to have statistically significantly different means (p < .005).



```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj  lower   upper  reject
----------------------------------------------------
  high    low   0.1661  0.001  0.1187  0.2134   True
  high medium   0.1185  0.001  0.0942  0.1427   True
   low medium  -0.0476 0.0243 -0.0902 -0.0049   True
----------------------------------------------------
```

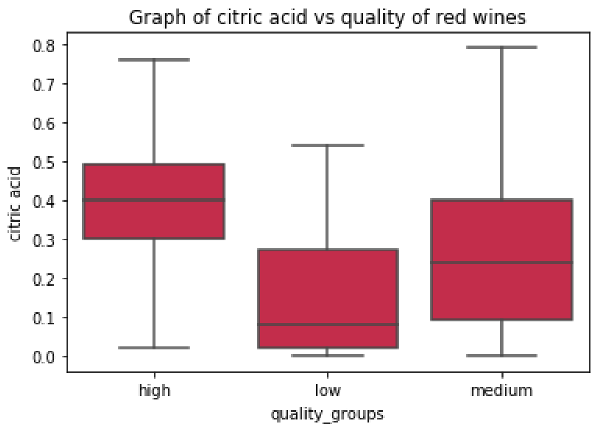


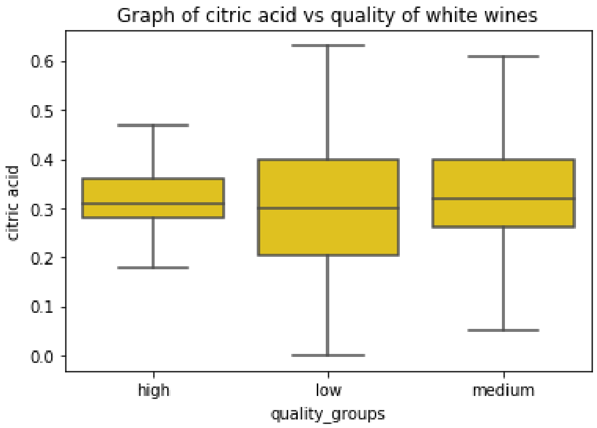
```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj  lower   upper  reject
----------------------------------------------------
  high    low   0.0274  0.001  0.0135  0.0412   True
  high medium   0.0074 0.0114  0.0014  0.0135   True
   low medium    -0.02 0.0011 -0.0331 -0.0068   True
----------------------------------------------------
```

The results of the Tukey test for quality groups of red and white wine for volatile acidity feature are shown in the figures above respectively.

For red and white wines, there seems to be a statistically significant relationship between all the quality groups. The volatile acidity for all the quality groups seems to be markedly different from each other! The mean difference is negative for low and medium quality red and white wines. So higher quality red and white wines seem to have more volatile acidity. As the wine quality rating increases, the volatile acidity content might be increasing too!

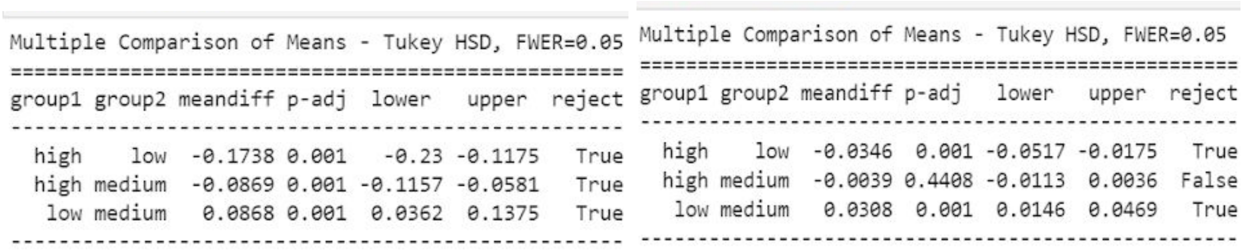
3. Citric Acid VS Quality





Background: Citric acid is generally found in very small quantities in wine grapes. It acts as a preservative and is added to wines to increase acidity, complement a specific flavor or prevent ferric hazes. Citric acid content seems to be slightly higher in high quality red wines. It can be added to finished wines to increase acidity and give a “fresh” flavor.

Analysis: Citric acid content does not vary a lot according to quality groups for white wines. For red wine, however, citric acid seems to have a strong positive relationship with wine quality. There are however, a few outliers in each quality group, which could be investigated to understand better what causes them to have such deviation from the group.



```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj  lower   upper  reject
----------------------------------------------------
  high    low  -0.1738 0.001   -0.23 -0.1175   True
  high medium  -0.0869 0.001 -0.1157 -0.0581   True
   low medium   0.0868 0.001  0.0362  0.1375   True
----------------------------------------------------
```

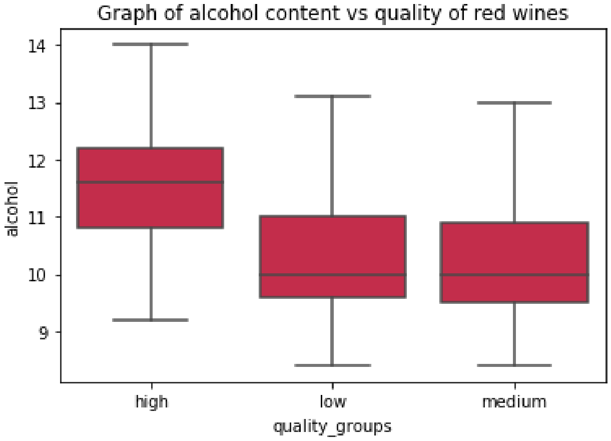
```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
  high    low  -0.0346  0.001 -0.0517 -0.0175   True
  high medium  -0.0039 0.4408 -0.0113  0.0036  False
   low medium   0.0308  0.001  0.0146  0.0469   True
----------------------------------------------------
```

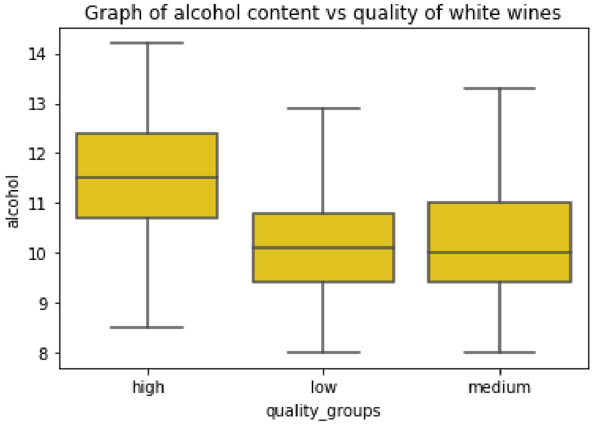
The results of the Tukey test for quality groups of red and white wine for citric acid feature are shown in the figures above respectively.

Citric acid content seems to vary significantly between low, medium and high quality groups for red wines. The mean difference is negative for high quality red wines when compared to low and medium quality ones. So maybe high quality red wines have a lower amount of citric acid! The citric acid content for low quality groups of white wines seems to be different when compared to other quality groups.

As with red wines, even for low and high quality white wines, the mean difference is negative, meaning high quality white wines might have lower amounts of citric acid in them! If we extrapolate this logic then it is to be expected that the mean difference is positive for low and medium quality white wines. This means citric acid content is more for lower quality white wines!

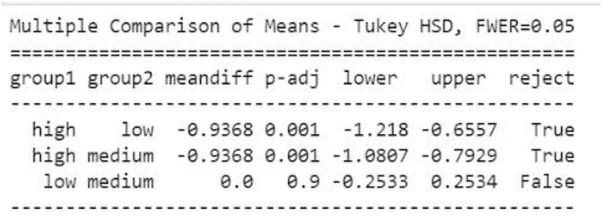
4. Alcohol VS Quality



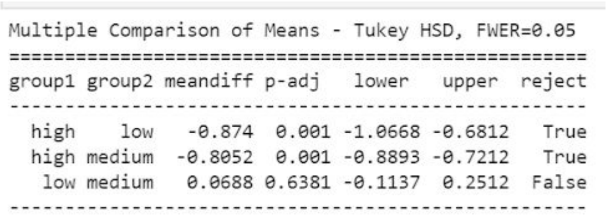


Background: Alcohol produces a desirable psychedelic effect that many find enjoyable in their wine. Too much can have undesirable side effects.

Analysis: High quality red wines (8 and above) have a higher alcohol content. Low and medium quality wines might have roughly the same range of alcohol content. There seem to be a few outliers in medium quality red wines, where despite high alcohol content, their quality rating is below 8. Alcohol content is slightly more in high quality white wines.



```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj  lower   upper  reject
----------------------------------------------------
  high    low  -0.9368 0.001  -1.218 -0.6557   True
  high medium  -0.9368 0.001 -1.0807 -0.7929   True
   low medium      0.0   0.9 -0.2533  0.2534  False
----------------------------------------------------
```



```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower   upper  reject
----------------------------------------------------
  high    low   -0.874  0.001 -1.0668 -0.6812   True
  high medium  -0.8052  0.001 -0.8893 -0.7212   True
   low medium   0.0688 0.6381 -0.1137  0.2512  False
----------------------------------------------------
```

The results of the Tukey test for quality groups of red and white wine for quality feature are shown in the figures above respectively.

The alcohol content for high quality red and white wines seems to vary significantly when compared to other quality groups. There seems to be a negative difference in means indicating that the alcohol content might be lower for higher quality red and white wines!

FEATURE SELECTION AND STATISTICAL ANALYSIS

VIF Scores

I then checked VIF scores for the features. The variance inflation factor (VIF) quantifies the extent of correlation between one predictor and the other predictors in a model. It is used for diagnosing collinearity/multicollinearity. Higher values signify that it is difficult to assess accurately the contribution of predictors to a model. The more VIF increases, the less accurate our coefficients will be as indicators of the effect of a variable. In general, a VIF above 10 indicates high correlation and might be a cause for concern. But the VIF scores for the red and white wine dataset were below 5!

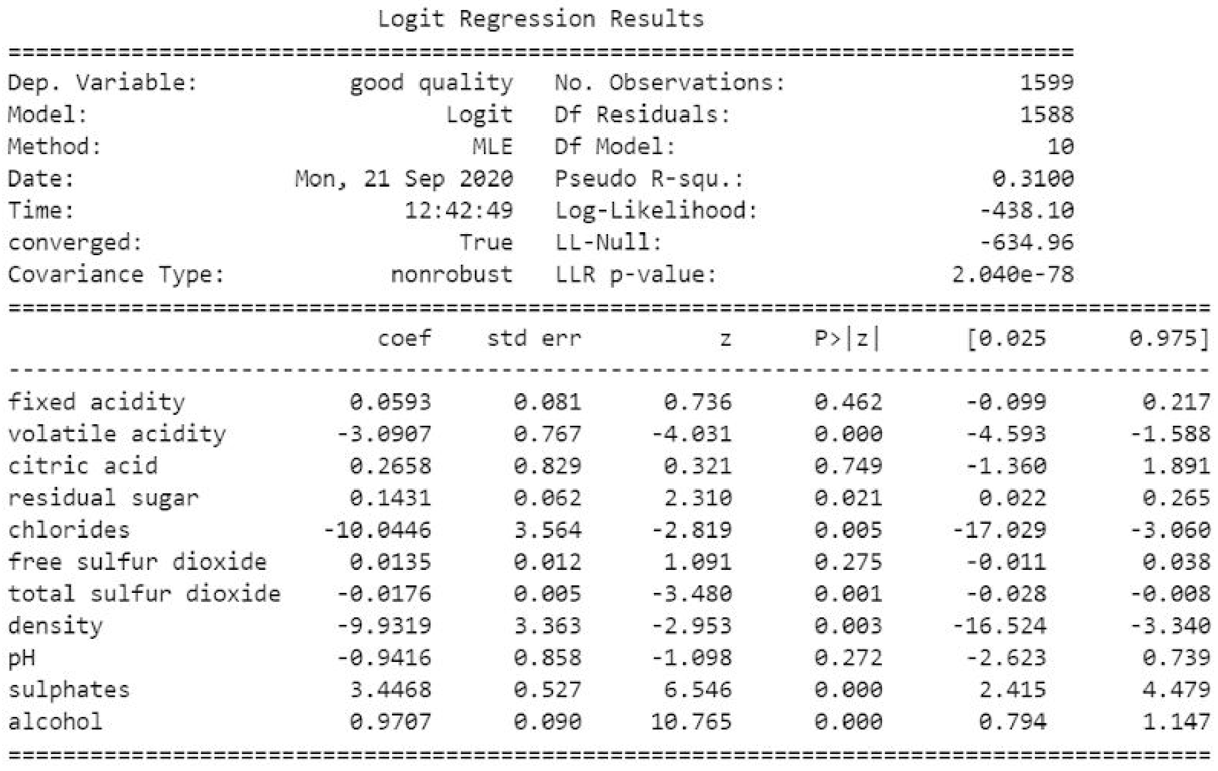
Recursive Feature Elimination (RFE) using logistic regression

RFE is a wrapper-type feature selection algorithm. This means that a different machine learning algorithm is given and used in the core of the method, is wrapped by RFE, and used to help select features. RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains.

After scaling the red and white wines dataset within the range of 0 to 1 using sklearn’s MinMaxScaler, we implemented sklearn’s Recursive Feature Elimination (RFE) to check if any of the wine features are unnecessary or redundant. Features were ranked using RFE’s ranking_ and support_ attribute. RFE returned all the 11 features as important. So I built a Logistic Regression model using statsmodels, with all the 11 features.

Logistic Regression using statsmodels

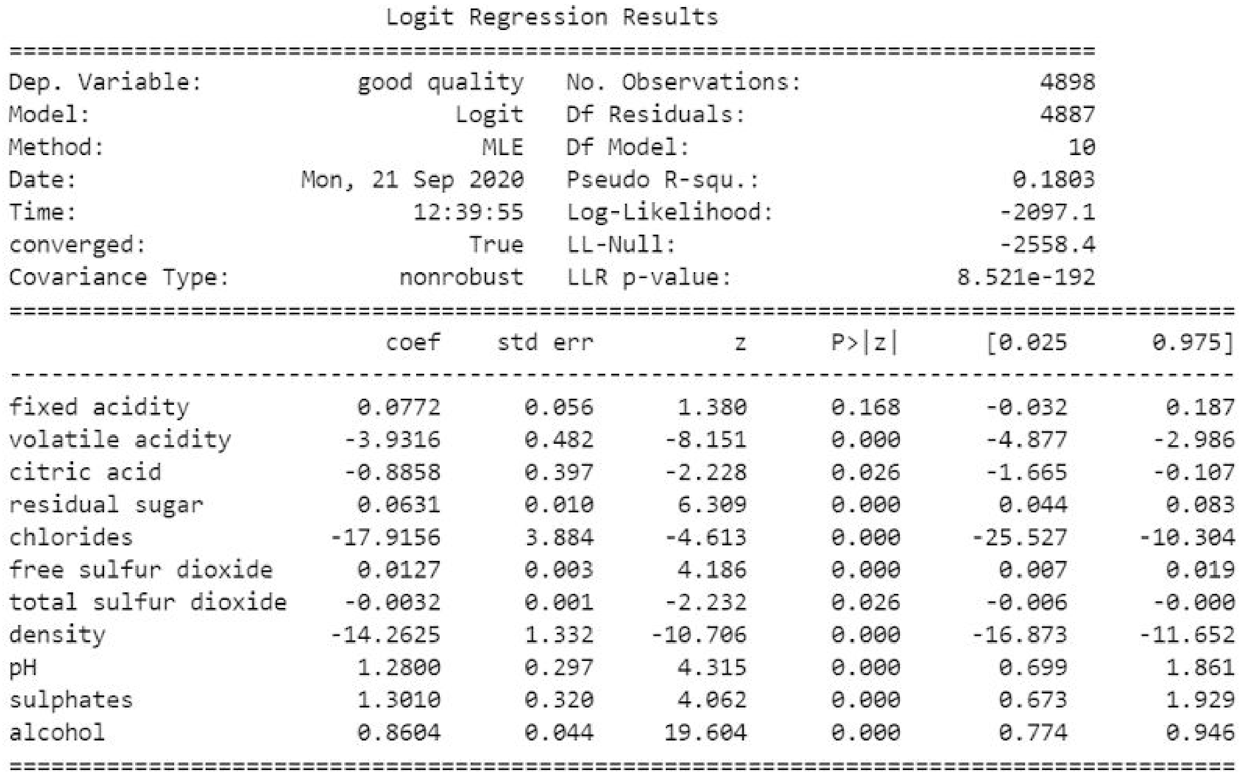
The aim was to build a classifier which can differentiate between good and poor quality wines. For this reason all wines rated 7 and above were classified as “good quality” and the ones below 7 as “poor quality”.



```
                           Logit Regression Results
==============================================================================
Dep. Variable:           good quality   No. Observations:                 1599
Model:                          Logit   Df Residuals:                     1588
Method:                           MLE   Df Model:                           10
Date:                Mon, 21 Sep 2020   Pseudo R-squ.:                  0.3100
Time:                        12:42:49   Log-Likelihood:                -438.10
converged:                       True   LL-Null:                       -634.96
Covariance Type:            nonrobust   LLR p-value:                 2.040e-78
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
fixed acidity            0.0593      0.081      0.736      0.462      -0.099       0.217
volatile acidity        -3.0907      0.767     -4.031      0.000      -4.593      -1.588
citric acid              0.2658      0.829      0.321      0.749      -1.360       1.891
residual sugar           0.1431      0.062      2.310      0.021       0.022       0.265
chlorides              -10.0446      3.564     -2.819      0.005     -17.029      -3.060
free sulfur dioxide      0.0135      0.012      1.091      0.275      -0.011       0.038
total sulfur dioxide    -0.0176      0.005     -3.480      0.001      -0.028      -0.008
density                 -9.9319      3.363     -2.953      0.003     -16.524      -3.340
pH                      -0.9416      0.858     -1.098      0.272      -2.623       0.739
sulphates                3.4468      0.527      6.546      0.000       2.415       4.479
alcohol                  0.9707      0.090     10.765      0.000       0.794       1.147
========================================================================================
```

Logistic Regression results for unscaled red wine data

After scaling the data, all red wine features except for fixed acidity, citric acid, free sulfur dioxide and density were found to have a statistically significant effect on quality.



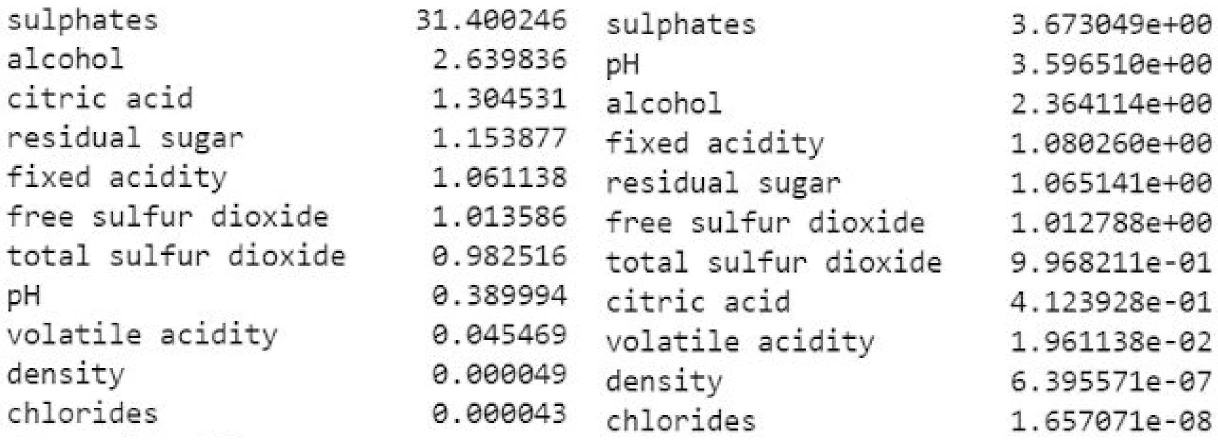
```
                           Logit Regression Results
==============================================================================
Dep. Variable:           good quality   No. Observations:                 4898
Model:                          Logit   Df Residuals:                     4887
Method:                           MLE   Df Model:                           10
Date:                Mon, 21 Sep 2020   Pseudo R-squ.:                  0.1803
Time:                        12:39:55   Log-Likelihood:                -2097.1
converged:                       True   LL-Null:                       -2558.4
Covariance Type:            nonrobust   LLR p-value:                8.521e-192
========================================================================================
                           coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
fixed acidity            0.0772      0.056      1.380      0.168      -0.032       0.187
volatile acidity        -3.9316      0.482     -8.151      0.000      -4.877      -2.986
citric acid             -0.8858      0.397     -2.228      0.026      -1.665      -0.107
residual sugar           0.0631      0.010      6.309      0.000       0.044       0.083
chlorides              -17.9156      3.884     -4.613      0.000     -25.527     -10.304
free sulfur dioxide      0.0127      0.003      4.186      0.000       0.007       0.019
total sulfur dioxide    -0.0032      0.001     -2.232      0.026      -0.006      -0.000
density                -14.2625      1.332    -10.706      0.000     -16.873     -11.652
pH                       1.2800      0.297      4.315      0.000       0.699       1.861
sulphates                1.3010      0.320      4.062      0.000       0.673       1.929
alcohol                  0.8604      0.044     19.604      0.000       0.774       0.946
========================================================================================
```

Logistic Regression results for scaled red wine data

For white wines, after scaling the data, except for pH and sulphates, all other features had a significant effect on quality.

Odds Ratio on unscaled data

I then checked the odds ratio for the logistic regression model features. Each estimated coefficient is the expected change in the log odds of wine being of good quality (rated<7) for a unit increase in the corresponding predictor variable holding the other predictor variables constant at certain value. Each exponentiated coefficient is the ratio of two odds, or the change in odds in the multiplicative scale for a unit increase in the corresponding predictor variable holding other variables at a certain value.



```
sulphates               31.400246   sulphates               3.673049e+00
alcohol                  2.639836   pH                      3.596510e+00
citric acid              1.304531   alcohol                 2.364114e+00
residual sugar           1.153877   fixed acidity           1.080260e+00
fixed acidity            1.061138   residual sugar          1.065141e+00
free sulfur dioxide      1.013586   free sulfur dioxide     1.012788e+00
total sulfur dioxide     0.982516   total sulfur dioxide    9.968211e-01
pH                       0.389994   citric acid             4.123928e-01
volatile acidity         0.045469   volatile acidity        1.961138e-02
density                  0.000049   density                 6.395571e-07
chlorides                0.000043   chlorides               1.657071e-08
```

Odds ratio for red and white wines (left and right) respectively.

Odds ratio was highest for sulphates in case of red wines and white wines. It was the least for chlorides in case of red wines and white wines.

An odds ratio of 31.40 for sulphates in case of red wines means we will see a 3140% increase in the odds of a red wine being of good quality for a 1 g/dm^3 increase in sulphates! Similarly, there will be a 163% increase in the odds of a red wine being of good quality for a 1 % by volume increase in alcohol content. And a 30% increase in the odds of a red wine being of good quality for a 1 g/dm^3 increase in citric acid. But the p-value was not significant for citric acid indicating that the odds ratio won't be statistically significant.

For white wines, 267% increase in the odds of a white wine being of good quality for a 1 g/dm^3 increase in sulphates. And 259% and 136% increase in odds for good quality white wine for 1 mg/dm^3 increase in pH and 1 % by volume alcohol content! Since the p-value was < 0.01, these results are statistically significant.

We now move on to the machine learning section of the project.

MACHINE LEARNING

In the wines dataset, there were 1382 red wines of “poor quality”, rated below and 7 and 217 of high quality. 3838 white wines were “poor quality” and 1060 of high quality.

We used sklearn’s train_test_split to split the already standardised dataset into train and test sets.

The models were trained using sklearn’s Logistic Regression, XGBoost, Decision Tree and Random Forest.

All models were tuned for hyper parameters to get the best possible performance. GridSearchCV was used for hyper parameter tuning. AUC score was used to evaluate the model performance.

Summary

A summary of all the models is displayed below.

Red wines

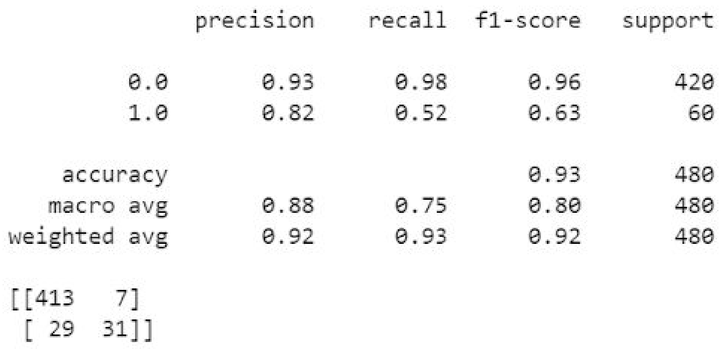
Model Type	Parameters with Grid Search	AUC score
Logistic Regression	'C: 5.179474679231202, penalty: l2	0.8033
XGBoost	learning_rate=0.05, max_depth=4, n_estimators=180	0.9089
Decision Tree	max_depth=4, max_leaf_nodes=6, min_samples_leaf=1, min_samples_split=2	0.7652
Random Forest	‘max_depth : 20, 'max_features': 0.25, 'min_samples_split': 2, 'n_estimators': 150	0.9182

White wines

Model Type	Parameters	AUC score
Logistic Regression	‘C': 268.2695795279727, 'penalty': 'l2'	0.8033
XGBoost	learning_rate=0.1, max_depth=3, n_estimators=100	0.9071
Decision Tree	max_depth=13, max_leaf_nodes=41, min_samples_leaf=1, min_samples_split=2	0.7434
Random Forest	'max_features': 0.25, 'min_samples_split': 2, 'n_estimators': 250	0.9190

The best performing model for red and white wines, was the Random Forest classifier. For both the wine types, Random Forest is able to correctly classify high and poor quality wines, ~92% of the time!

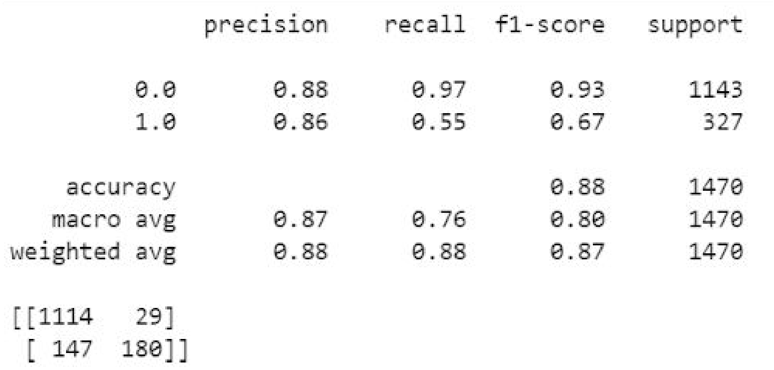
Classification report and confusion matrix of random forest model for red wines:



```
              precision    recall  f1-score   support

         0.0       0.93      0.98      0.96       420
         1.0       0.82      0.52      0.63        60

    accuracy                           0.93       480
   macro avg       0.88      0.75      0.80       480
weighted avg       0.92      0.93      0.92       480

[[413   7]
 [ 29  31]]
```

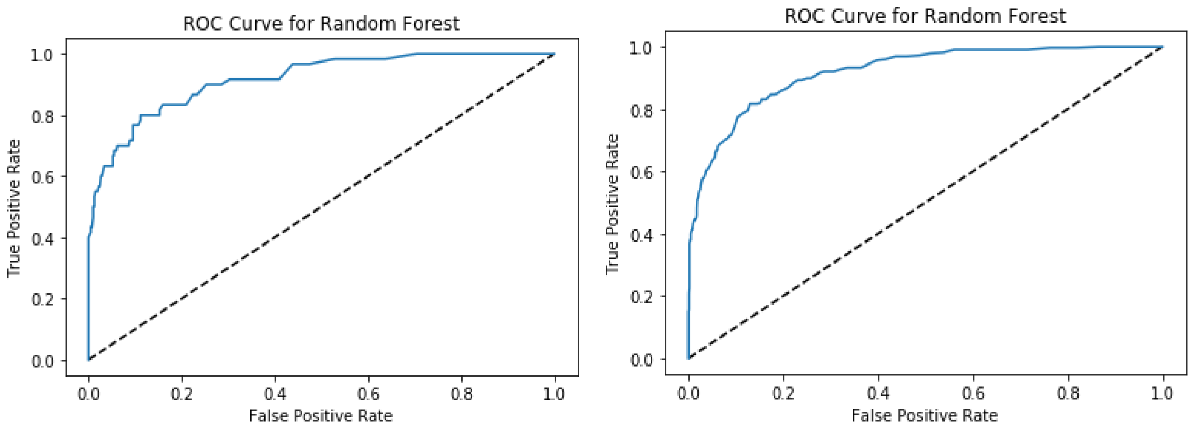
The random forest model for red wines has a precision of 0.82 for the positive class, it means when the model predicts a wine to be of high quality, it is correct 82% of the time! A recall of 0.52 for the positive class means, the model correctly identifies 52% of all high quality red wines in the dataset. As can be seen from the confusion matrix, out of 480 wines of the test set, the model predicts 29 red wines as FNs and 7 as FP. It means the model tends to classify high quality red wines (wines rated < 7) as low quality.

Classification report and confusion matrix of random forest model for white wines:



```
              precision    recall  f1-score   support

         0.0       0.88      0.97      0.93      1143
         1.0       0.86      0.55      0.67       327

    accuracy                           0.88      1470
   macro avg       0.87      0.76      0.80      1470
weighted avg       0.88      0.88      0.87      1470

[[1114   29]
 [ 147  180]]
```

The hyper parameter tuned, random forest model for white wines has a precision of 0.86 for the positive class, it means when the model predicts a wine to be of high quality, it is correct 86% of the time! A recall of 0.55 for the positive class means, the model correctly identifies 55% of all high quality white wines in the dataset. As can be seen from the confusion matrix, the model predicts more FNs (147), compared to FP (29). In this case it could indeed be better to get a high quality wine predicted as low quality, compared to a low quality wine being predicted as high quality!

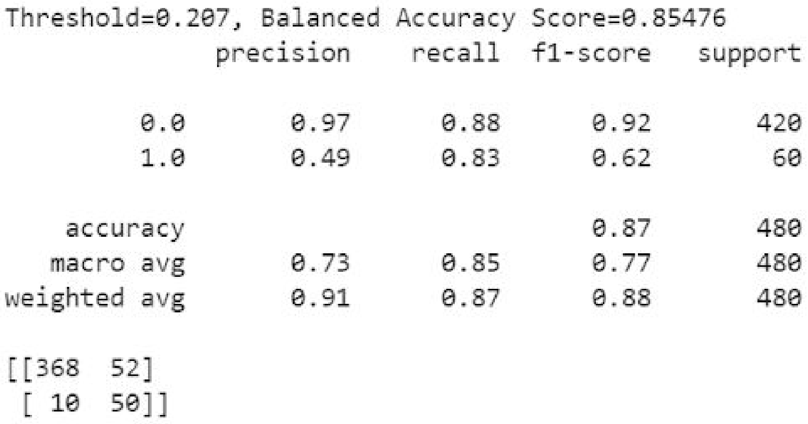


Red and White wine ROC Curves for Random Forest classifier model.

Custom Thresholding on random forest model

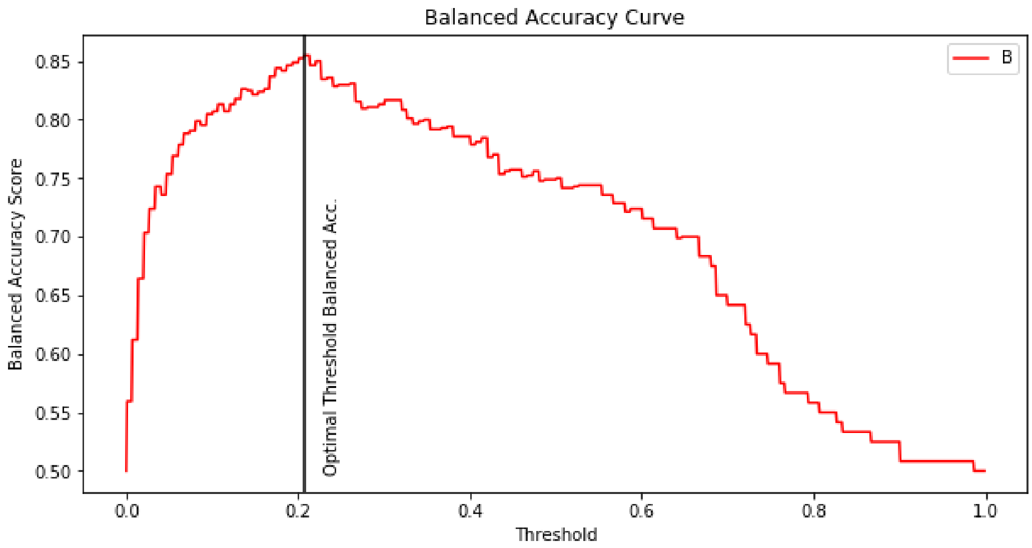
The default threshold for all ML models is 0.5. I tried to change it in order to get better model performance.

Results for red wine thresholding

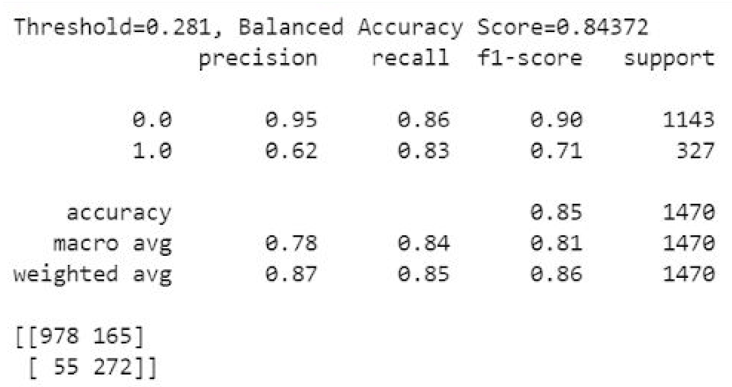


```
Threshold=0.207, Balanced Accuracy Score=0.85476
              precision    recall  f1-score   support

         0.0       0.97      0.88      0.92       420
         1.0       0.49      0.83      0.62        60

    accuracy                           0.87       480
   macro avg       0.73      0.85      0.77       480
weighted avg       0.91      0.87      0.88       480

[[368  52]
 [ 10  50]]
```

After checking the balanced accuracy score for thresholds ranging from 0 to 1, a threshold value of 0.207 gave the maximum balanced accuracy score of 0.85476. The recall for positive class now increased to 0.83, from 0.52 and precision decreased from 0.82 to 0.49. The model now misclassified 52 low quality red wines as high quality and 10 high quality red wines as low quality.

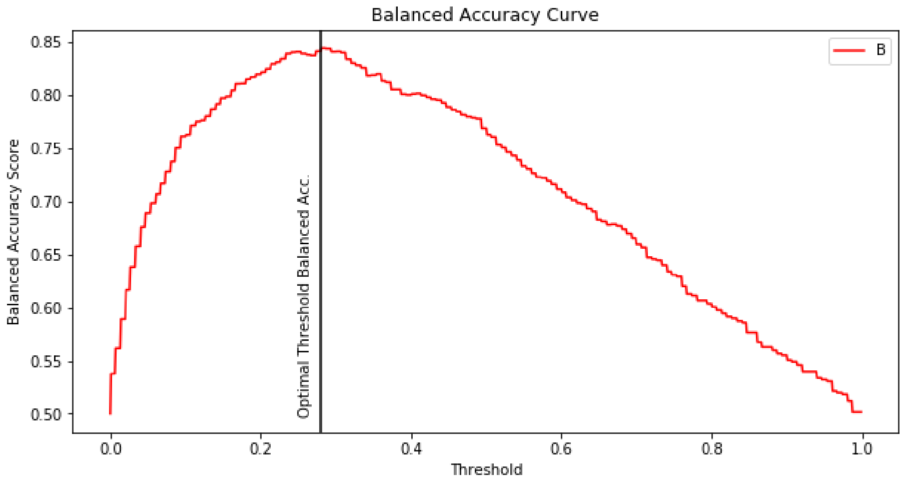


Results for white wine thresholding



```
Threshold=0.281, Balanced Accuracy Score=0.84372
              precision    recall  f1-score   support

         0.0       0.95      0.86      0.90      1143
         1.0       0.62      0.83      0.71       327

    accuracy                           0.85      1470
   macro avg       0.78      0.84      0.81      1470
weighted avg       0.87      0.85      0.86      1470

[[978 165]
 [ 55 272]]
```

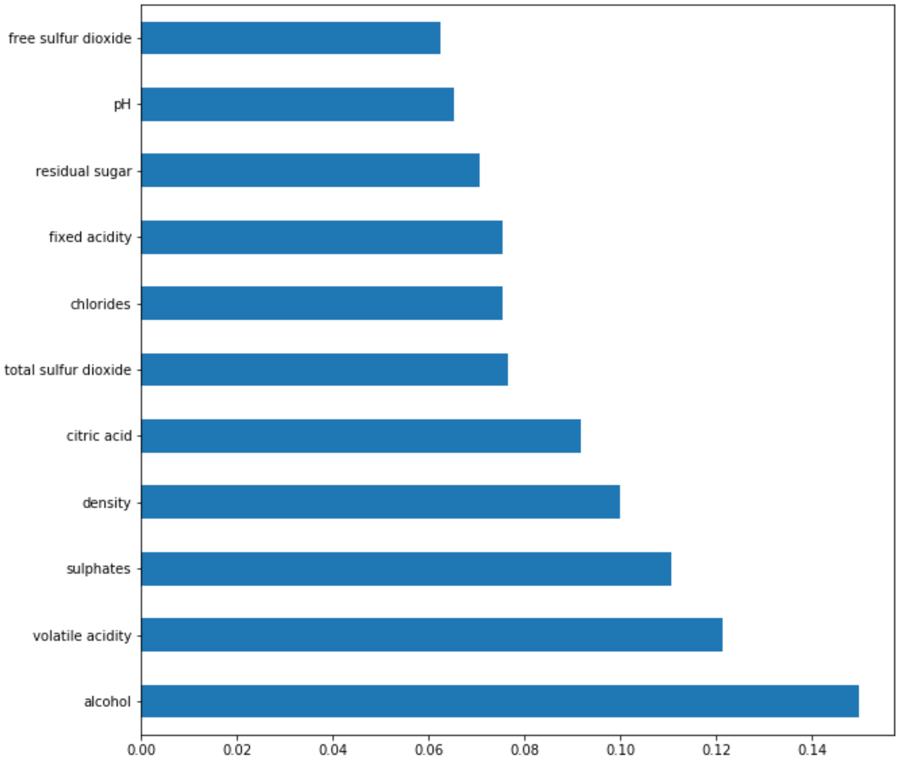
A maximum balanced accuracy score of 0.8472 was obtained for a threshold of 0.281. Precision reduced to 062 but recall increased to 0.83. The F1 score improved to 0.71 from 0.67. The model misclassified 165 white wines as high quality, when they were low quality and 55 white wines as low quality when they were of high quality.



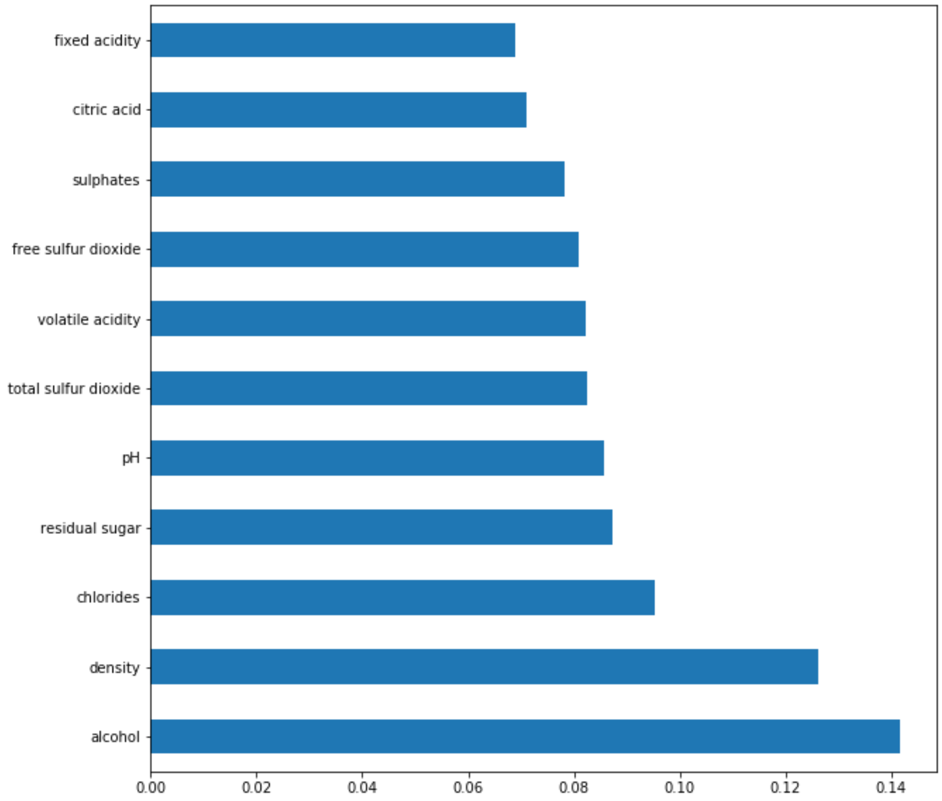
After thresholding, the model for red and white wines had an increase in recall value and a slight decrease in precision, but the accuracy did not see any improvement. It also had more misclassifications when compared to the earlier default threshold models.

Feature Importance using Random Forest

On the basis of the model built, I wanted to understand which wine features have the most influence on its quality. The feature_importances_ attribute of Random Forest Classifier was used for this purpose.



As in the image seen above, alcohol and volatile acidity have the most influence on the quality of red wines.



Alcohol and density have the most effect on the quality of white wines!

CONCLUSION

From the correlation matrix, box plots, Tukey test, logistic regression and the machine learning analysis, it is clear that alcohol content has the largest effect on the quality of red and white wines! Sulphate content is another important deciding factor for red wines whereas chlorides and residual sugars could play an important role in white wine quality. The odds ratio was high for pH and sulphates but the feature_importances_ attribute of random forest differed from that.

FUTURE DIRECTIONS

To conclude my project, I explored and visualized the red and white wine quality dataset of Kaggle using box plots and correlation matrix. Tukey test was performed for statistical analysis of the dataset. I built classifiers using Logistic Regression, XGBoost, Decision Trees and Random Forest which correctly predict 'high' (wines started > 7) and 'low' (wines rated below 7) quality wines. RFE technique was used on logistic regression to optimise feature selection. Grid searching and cross validation was used to optimize model performance. The best performing model was the Random Forest classifier with an AUC score of 91.70%. However, there is still room for improvement and with time and resources there are ways this model can be made to be better.

Here are a few of these potential topics:

- Get more data! The dataset had a total of 6497 wine ratings. More ratings could help improve the model
- The breadth of the dataset could be improved. This dataset has only red and white wines. Wines like Rose could be added.
- The number of features could be increased, we are focusing here on a few physicochemical properties but there may be others that are relevant as well or perhaps even confounding that would lead us to different results.
- For this project I focused on binary classification (either low or high rating), but implementing multiclass classification for machine learning with high, low and medium quality groups (like I did in boxplots) could also be informative.