Capstone 2: Project Report

Del Wester

3/24/2021

Wine Quality

Many wine brands are seeking new ways to maximize the success of their wines. Before making any decisions, it might be helpful to know which features contribute to a wine's quality. Knowing these features can enable a brand to make more intelligent decisions when making it. But what exactly are these features? Using ML techniques with wine data retrieved from the following website, I plan to answer this question. https://archive.ics.uci.edu/ml/datasets/wine+quality

Data Wrangling

The dataset for this project was wrangled by another party prior to beginning this project.

Using red and white wine samples, inputs include objective tests (PH values) and the output is based on sensory data (wine tasting by experts). Using a median of at least 3 evaluations, each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Several data mining methods were applied to model these datasets under a regression approach to determine wine quality.

Objective: Find if any of the features other than quality can be used to distinguish quality.

Two datasets were created, using red and white wine samples. The input included objective tests (e.g. PH values) and the output was based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent). Several data mining methods were applied to model these datasets under a regression approach. The support vector machine model achieved the best results. Several metrics were computed: MAD, confusion matrix for a fixed error tolerance (T),etc. Also, we plotted the relative importance of the input variables (as measured by a sensitivity analysis procedure).

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. 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.). These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines.

Number of Attributes: 12 + output attribute

Attribute information:

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

12 - type

Output variable (based on sensory data):

13 - quality (score between 0 and 10)

Missing Attribute Values: None

Exploratory Data Analysis

The Wine Quality data includes mostly continuous data with a few categorical columns. Exploratory data analysis can be used to derive relationships between the wine quality and the various features available from the wine's profile and suggest improvements to the profiles that would increase the wine's quality. This analysis takes into consideration a certain spectrum of wines related to the two specific types – red and white.

The quality feature of this set was the result of wine tasters opinions, ranging from 3 to 9, with the higher numbers being higher quality. The spread is shown below:

"quality" value counts:

```
6   2836
5   2138
7   1079
4    216
8    193
3     30
9      5
```

Type Distribution - Red: 1599 White: 4898

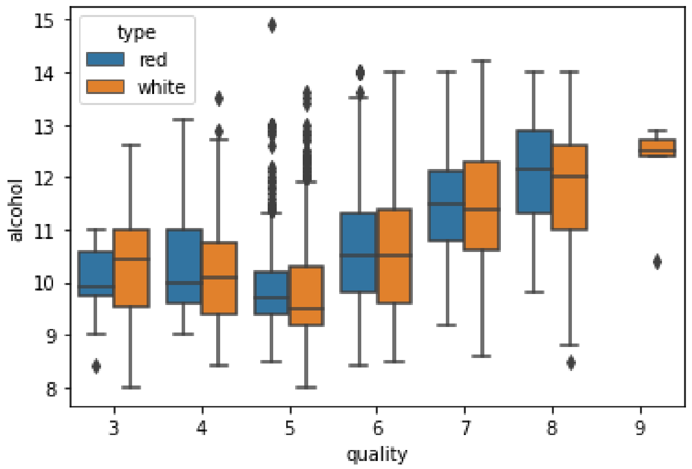
These will be split into a binary feature in the preprocessing step, with 7 and higher being high quality and everything else being lower quality. The plots for the output variable vs each of the important features are below.

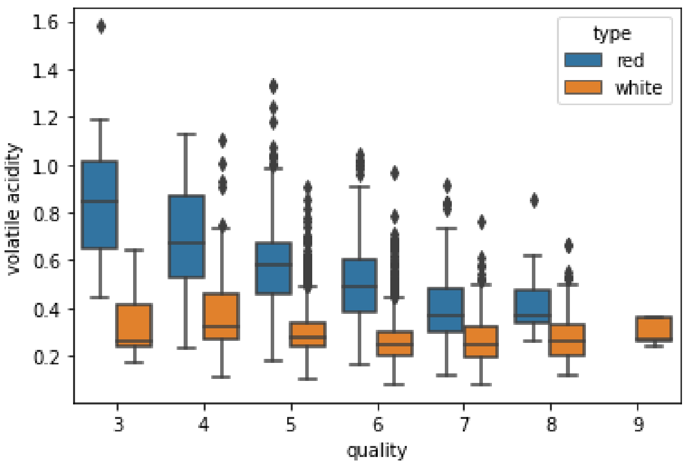
Top Three Important Features:

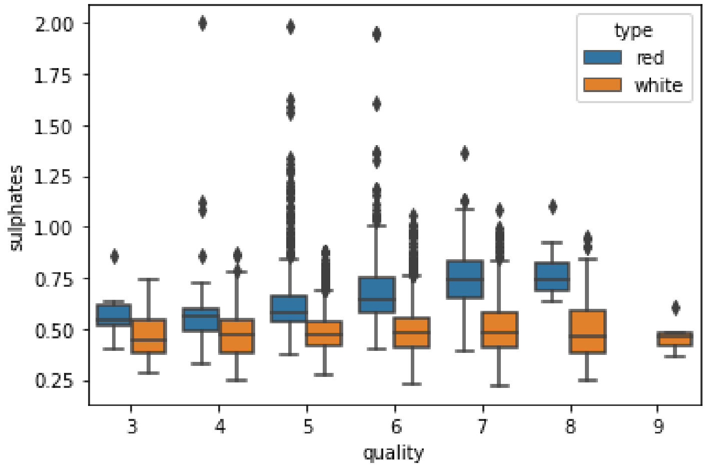
Red – alcohol, volatile acidity, sulphates

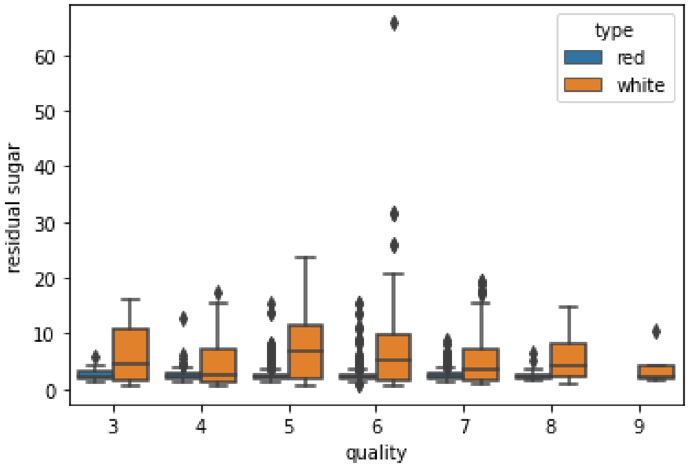
White - alcohol, volatile acidity, residual sugar

For both wines, as the alcohol content rises, so does the quality. Volatile acidity presents a very different picture. For red, it tends to be lower in high quality. For white, it doesn’t change much at all. Sulphates for red high quality tend to be higher and residual sugars in white lower.

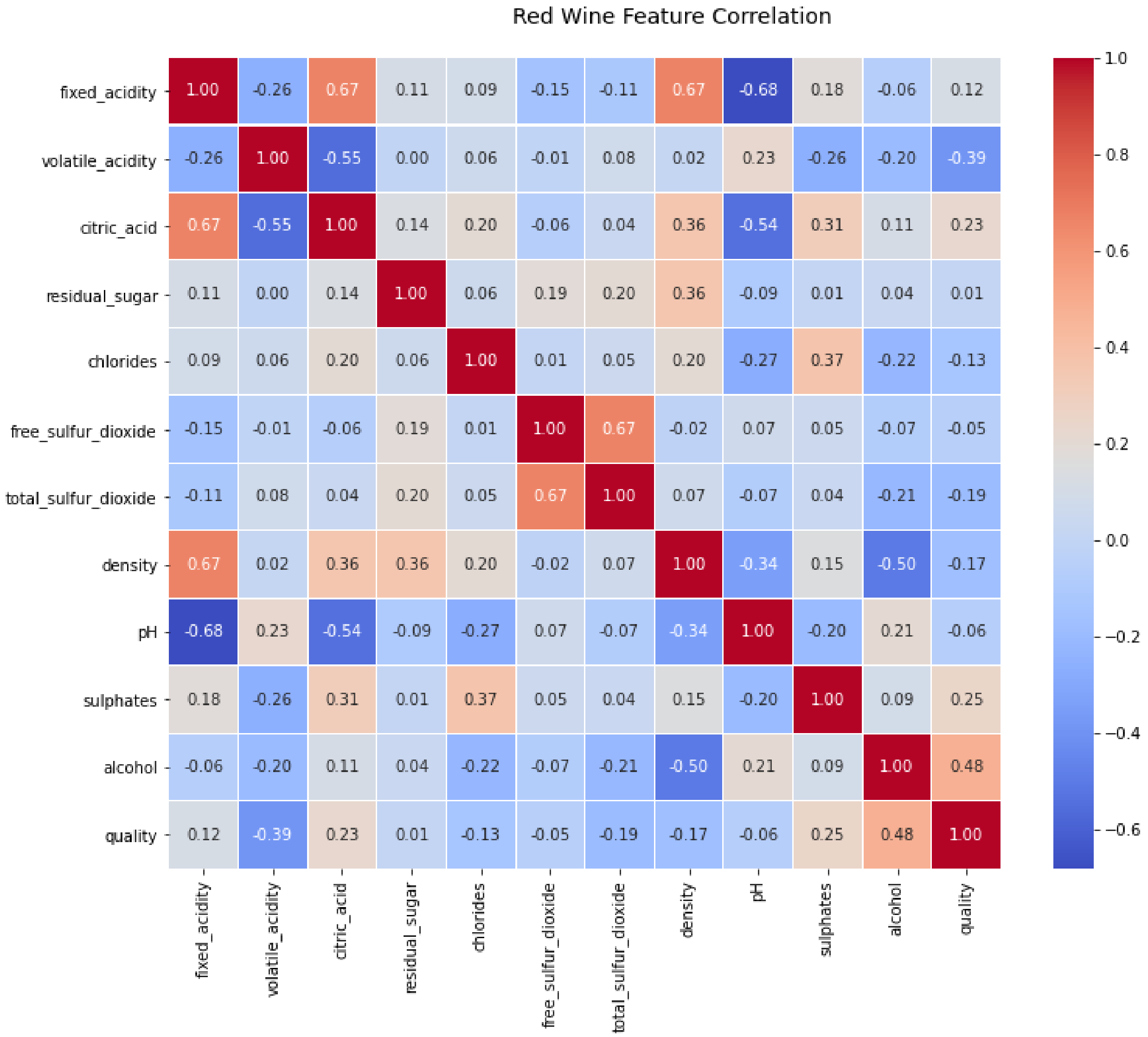








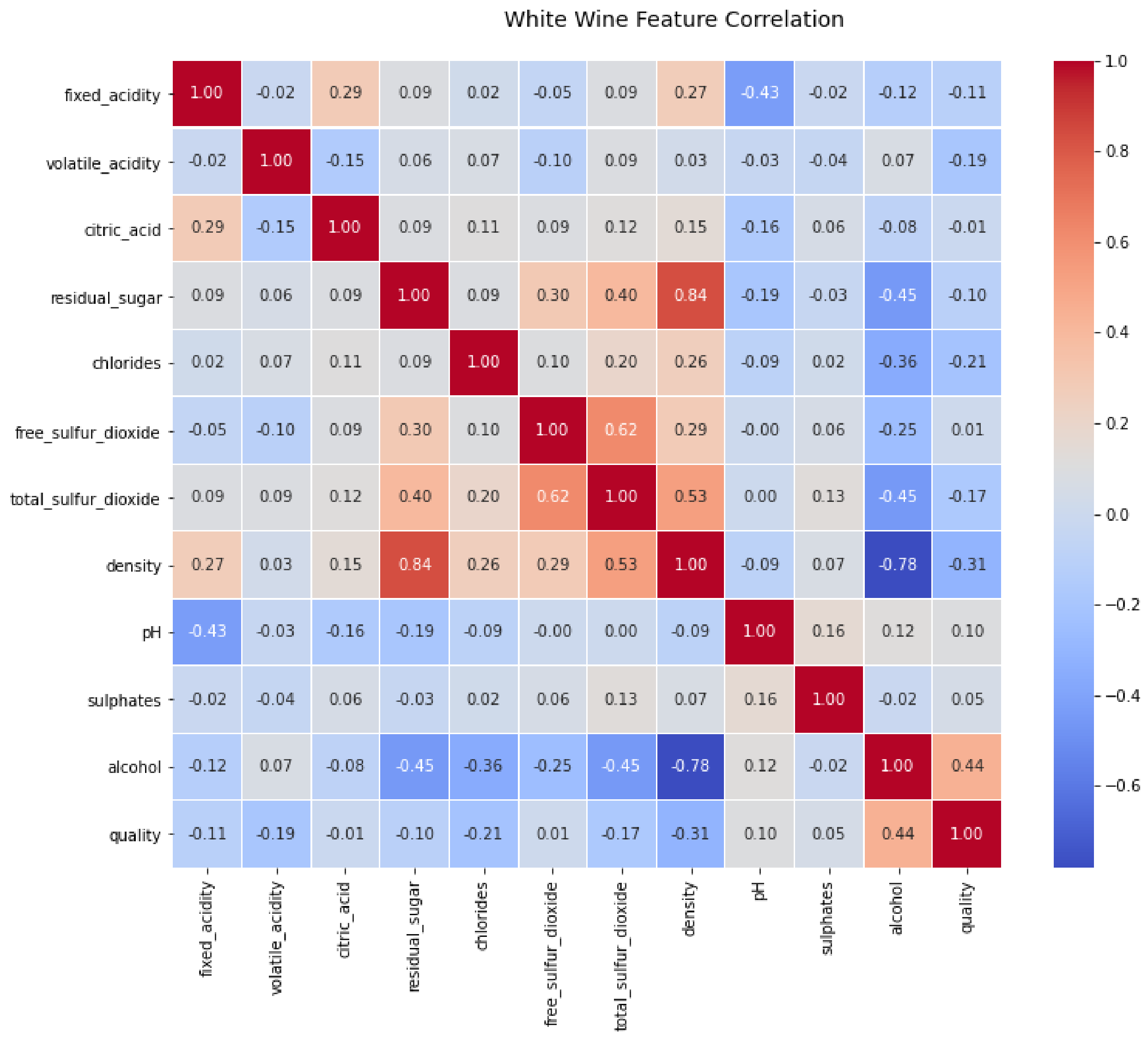
Correlation Matrix



Red Wine Feature Correlation

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
fixed_acidity	1.00	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile_acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.20	-0.39
citric_acid	0.67	-0.55	1.00	0.14	0.20	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
residual_sugar	0.11	0.00	0.14	1.00	0.06	0.19	0.20	0.36	-0.09	0.01	0.04	0.01
chlorides	0.09	0.06	0.20	0.06	1.00	0.01	0.05	0.20	-0.27	0.37	-0.22	-0.13
free_sulfur_dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00	0.67	-0.02	0.07	0.05	-0.07	-0.05
total_sulfur_dioxide	-0.11	0.08	0.04	0.20	0.05	0.67	1.00	0.07	-0.07	0.04	-0.21	-0.19
density	0.67	0.02	0.36	0.36	0.20	-0.02	0.07	1.00	-0.34	0.15	-0.50	-0.17
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05	0.04	0.15	-0.20	1.00	0.09	0.25
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07	-0.21	-0.50	0.21	0.09	1.00	0.48
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05	-0.19	-0.17	-0.06	0.25	0.48	1.00

Reds appear to have correlations between fixed acidity and pH, fixed acidity and density, fixed acidity and citric acid, and free sulfur dioxides and total sulfur dioxides.

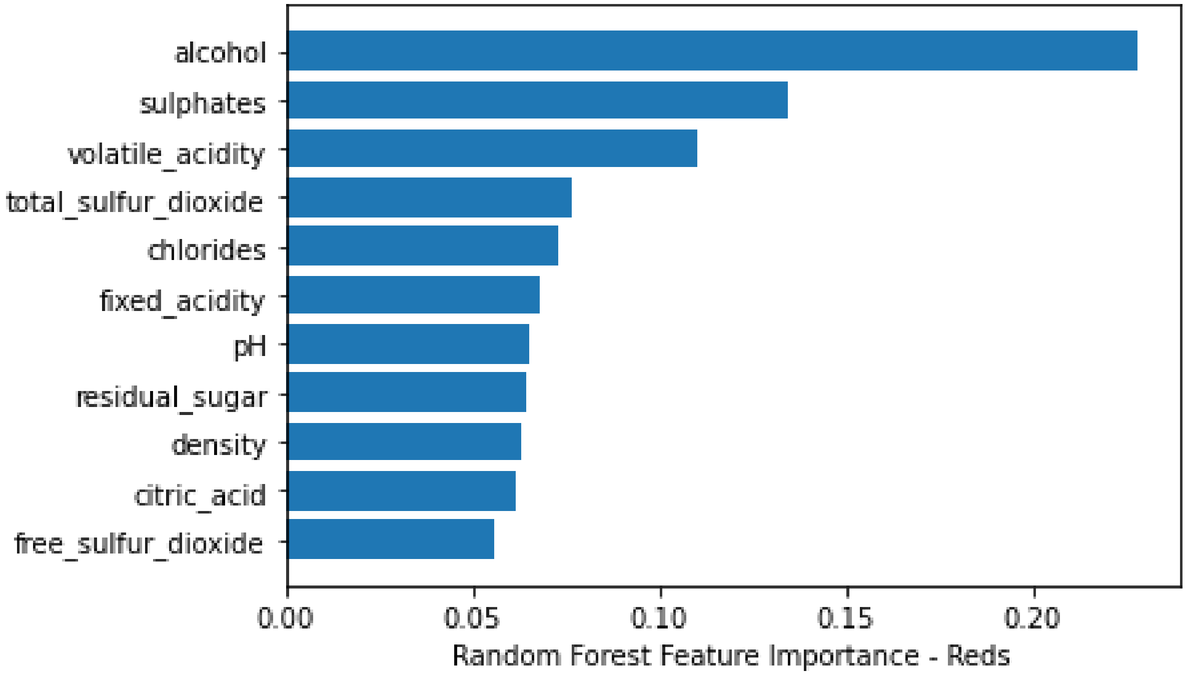


Whites have correlations between residual sugar and density and alcohol and density.

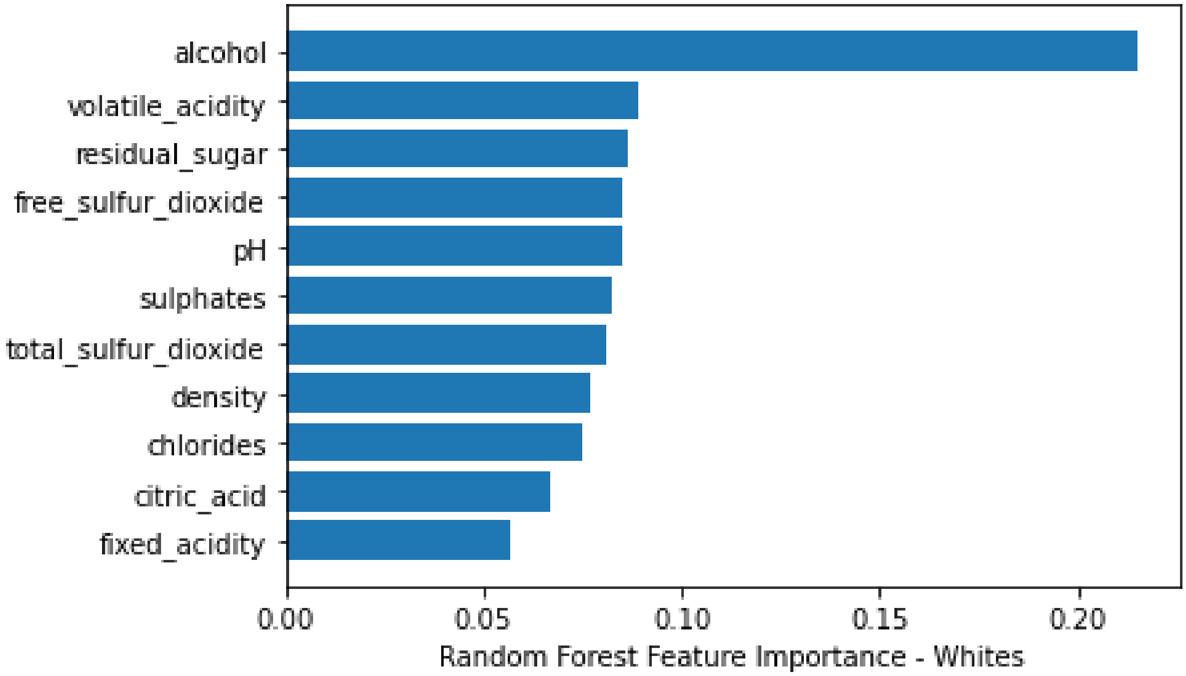
Preprocessing and Training

I started this section by splitting the data into high and low quality wines, with quality 7 and higher being high and the rest low, as well as segregating by type: red and white. 15 percent of the red wines in this data set are high quality and 25 percent of the whites. The data was then scaled using a Robust Scaler and I ran another correlation check to find any features to leave out of the final set due to high correlation as well as which features were of highest importance. I didn't drop any features.

Top Features - Red



Top Features – White



Modeling

I then tested and scored the following models on both sets of data using a gridsearch hyperparameter tuning for all three as well as running ROC_AUC scores and ROC curve: Random Forest Classifier, Logistic Regression, and KNeighbors Classifier. The Logistic Regression scored higher on both datasets than the others and was what I went with for the final model.

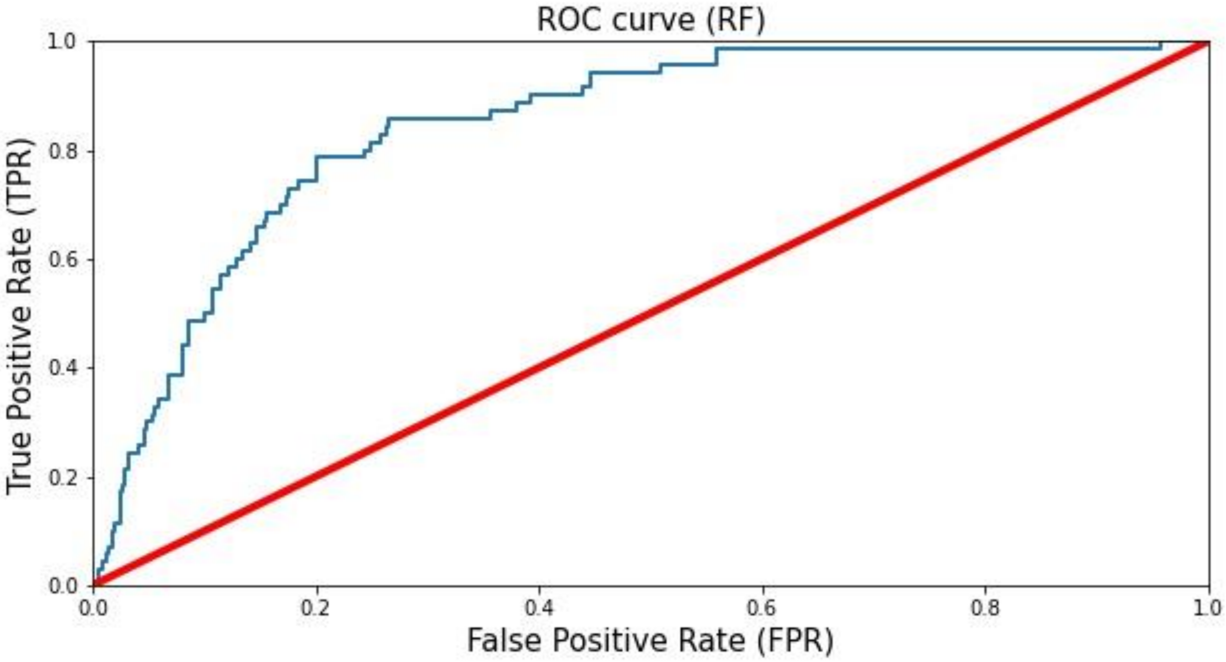
Logistic Regression Roc Auc scores:
Red: 0.84
White: 0.91

Random Forest Classifier Roc Auc scores:
Red: 0.79
White: 0.75

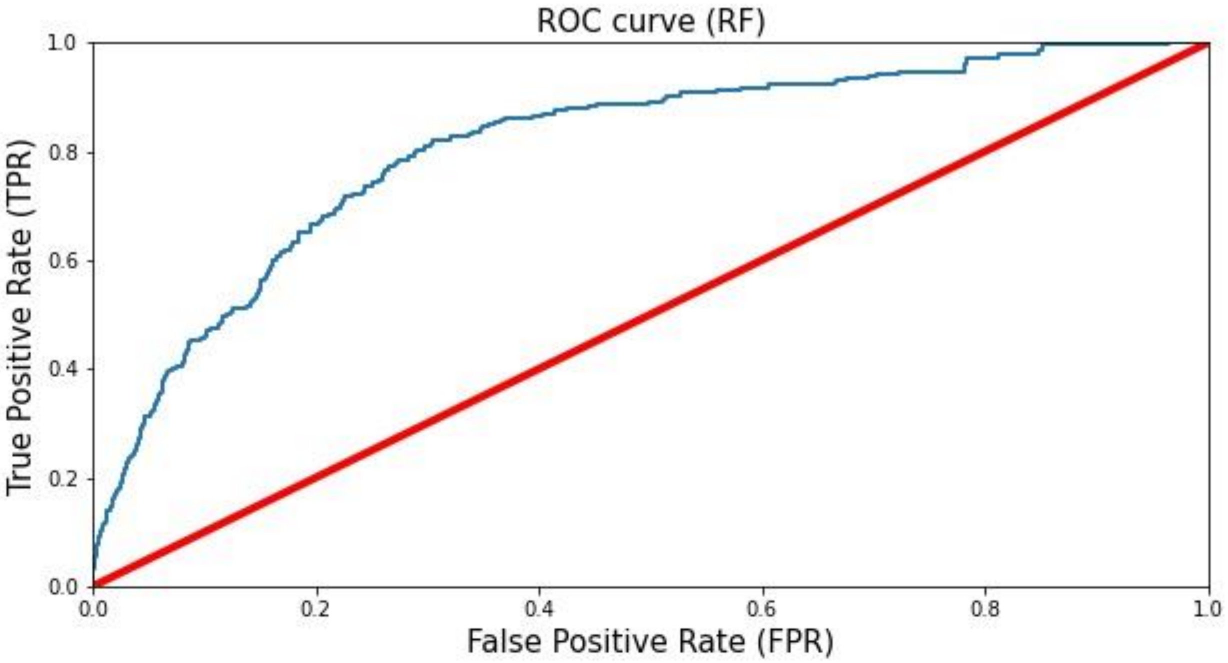
KNeighbors Classifier Roc Auc scores:
Red: 0.38
White: 0.77

I then ran the ROC curve using Logistic Regression

Red



White


ROC curve (RF)
True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Red:

```
                        precision    recall     f1-score   support


      Lo quality          0.88       0.99       0.93        410

      High quality        0.76       0.19       0.30         70


   accuracy                                       0.87        480

  macro avg             0.82       0.59          0.61        480

weighted avg            0.86       0.87          0.84        480
```

White:

```
                        precision        recall  f1-score   support


      Lo quality        0.82              0.99       0.89       1166

      High quality      0.72              0.14       0.24        304


   accuracy                                           0.81      1470

  macro avg             0.77              0.57       0.57       1470

weighted avg             0.80             0.81       0.76       1470
```

Conclusion

By analyzing the physicochemical data of red and white wines, I was able to create a model that can help industry producers, distributors, and sellers predict the quality of red wine products and have a better understanding of each critical feature. I found the Logistic Regression model performed better than the other two models. I determined three features most influential for both red and white wines. Red: volatile acidity, sulphates, and alcohol content. White: volatile acidity, residual sugars, and alcohol content. To be more specific, high-quality red wines seem to have lower volatile acidity, higher alcohol, and medium-to-high sulphates. Meanwhile, higher quality white wines also have low volatile acidity and high alcohol content, but differ in due to lower residual sugars.

This analysis comes with some limitations. First, the data set is unbalanced. A majority of the quality values were 5 and 6, which makes no significant contribution to finding an optimal model. These values make it harder to identify each features exact influence on a “high” or “low” quality of the wine, which was the main focus of this analysis. In order to improve the predictive model, more balanced data is needed. Another limitation worth mentioning is that the dataset only has 12 attributes, which reduces the accuracy of the predictive models. The solution for this is to include more relevant data features, such as the year of harvest, amount of brew time, or grape type. Different performance measures and/or machine learning techniques could also be utilized to find better performance and model comparisons.

Sources & Citations:

Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Available at: https://archive.ics.uci.edu/ml/datasets/wine+quality

Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009

Data source: http://www3.dsi.uminho.pt/pcortez/winequality09.pdf.