Alicia Wilson, Teagan Milford, Aditya Chawla,
Joseph Bridges, Sankalp Kulkarni, Carson Mullen

Introduction

This report identifies the physicochemical characteristics of red and white wine that most contribute to the quality of the wine. Eleven predictors were considered: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol.

Description of Project Goals

Description

The dataset is from the UCI Machine Learning Repository, titled “Wine Quality.” It contains data about red and wine variants of a Portuguese wine called “Vinho Verde.” The purpose of this report is to answer the question: which physiochemical traits are predictive of wine being high-quality?

Importance

There are countless wineries across the world - over 4,800 in California alone - that depend heavily on the quality of their product. For restaurant purchasing managers and sommeliers, being able to deduce the exact qualities that make a wine good and predict the overall wine quality based on quantitative metrics is the key to success in a $400 billion industry.

Initial Cleaning

The dataset had 6,497 wines, and 34 of these had NaN/Missing observations. Since the number of wines with NaN values was only 0.5% of the data, they were dropped completely. This left 6,463 wines to analyze.

Exploratory Data Analysis

Initial Correlations and Scatter Plots

A correlation matrix and heatmap were created first (Figure 1). Notably, free sulfur dioxide and total sulfur dioxide correlate 0.72. This raises a concern about multicollinearity. There were no moderate to high correlations in the target variable, quality (Figure 1, Table 1). Alcohol had the highest correlation of 0.44. The remaining correlations were weak at less than 0.3. To visualize the correlation matrices for quality, scatterplots were created (Figures 2.1-2.11), with quality on the y-axis and the respective chemicals on the x-axis. Aside from alcohol, there were no strong patterns.

Splitting Wine into Type

The data was split by type, showing 4,870 white and 1,593 red wines. Using this split, more correlations were calculated and compared directly to the combined wine correlations (Tables 1.1-1.3). Some coefficients are more robust when separated. For example, volatile acidity increases from -0.27 to -0.39 when red wine is separated. Note, as the coefficient for certain chemicals increases for one wine type, it decreases for the other. This raises the question of differing chemical importance for wine type. To further explore this question a table of mean values was produced (Table 2.1). While quality, density, pH, and alcohol mean values were similar, the remaining variables differed drastically. For example, the mean residual sugar for white was three times the red. A more extensive table was created, keeping the red/white split, but separating the mean values by quality (Table 2.2). To visualize this table, corresponding bar charts were created (Figures 3.1-3.11), revealing that high-quality red wines have a different chemical makeup than high-quality white wines. The same can be said for low-quality wines. For instance, while mean sulphates increased with the quality of red wine, they stayed level for white (Figure 3.1). For modeling, the data remained split between red and white.

Understanding the Target Variable: Quality

There is a highly imbalanced distribution across different quality ranges (Figure 4). Consequently, quality was binarized, where 0 = poor quality (quality ratings below 6) and 1 = good quality (quality ratings including and above 6). This solution tries to avoid biased predictions toward the majority class and simplifies the overarching problem (Figure 4.2). Yet, there is still a large imbalance. Red has 1,376 poor-quality wines and 217 good-quality wines, while white has 3,816 and 1,054, respectively.

Data Preparation I: For Logistic Regression and KNN Models

Outlier Detection and Treatment:

A combination of boxplots (Figure 5) and an outlier table (Table 3) were leveraged to flag any outliers. For both types of wines, none of the variables had >10% of outliers.

Since the cleaned data only contains 6,463 observations, outliers were treated with Winsorization. A 5% threshold was used since Table 3 revealed that most of the columns have less than 2.5% values as outliers; thus, the impact of outliers was reduced by replacing the lower 2.5% observations with the 2.5th percentile and higher 2.5% observations with the 97.5th percentile. Table 4 shows the pre- and post-Winsorization results.

Multicollinearity Detection and Treatment:

While KNN is not directly impacted by multicollinearity, it can be influenced indirectly by assigning more weight to the highly collinear predictors and interactions. Consequently, multicollinearity was checked using the Variance Inflation Factor (VIF). Density has a high VIF value for both red and white wines (Figure 6.3), so this variable was removed (Table 5). Total sulfur dioxide was also removed since it is a sum of “Bound” and “Free” sulfur dioxide, and free sulfur dioxide is already included in the data.

Standardization of the predictors:
The predictors were normalized using Z-transform for robust logistic regression and KNN model performances (Figure 6).

Data Preparation II: For All Models
Synthetic Minority Over-Sampling Technique (SMOTE)
Although the data for logistic regression and KNN were subject to Data Preparation I, and the data for XG Boost and Naïve Bayes were not, all of the models' data were subject to SMOTE. This technique addresses the problem of quality imbalance by generating synthetic samples for the minority classes. New samples, which were combinations of feature vectors from the minority quality classes, were created. Note, there are now four primed data sets in use: two for logistic regression and KNN models (one each for red/white), and two for XG Boost and Naïve Bayes (one each for red/white).

Models
All models provide an accuracy rating, an F1 Score, and an AUC score for comparison purposes.

Logistic Regression:
The data was split into train and test. A logistic regression was then applied, returning an accuracy of 82.8% for red wine and 72.6% for white wine. The most important features for predicting a red wine using logistic regression are alcohol and sulphates, while the most important features for white are alcohol and residual sugar (Tables 5.1-5.2). The F1 score for red wine is 0.8294, and 0.7406 for white wine, while the AUC score is 0.8863 and 0.7889, respectively (Figures 7.1-7.6).

K-Nearest Neighbors (KNN):
Using the train/test split, two KNN models were built (one for red, one for white). While an initial k value of 15 was used, cross-validation later revealed that the optimal k value for both datasets is 5 (Figures 8.1-8.2). With this new k value, the accuracy results are incredibly high: 0.981 (red) and 0.967 (white). The F1 score for red wine is 0.976, and 0.986 for white wine, while the AUC score is 0.998 and 0.998, respectively (Figures 8.3-8.4).

XG Boost:
XGBoost used a tree-based model with all of the features and outliers in the training data. The accuracy was high, again: 0.9375 (red) and 0.900 (white). From this model, it is clear that alcohol plays a key role in determining the quality of both types of wine. A greater presence of sulphates increases the quality of red wine. Conversely, for white wine,

volatile acidity, and residual sugar play a key role in determining the wine quality. The F1 score for red wine is 0.91, and 0.94 for white wine, while the AUC score is 0.97 and 0.98, respectively (Figures 9.2- 9.3)

Naïve Bayes:
Since each predictor is a continuous variable, each predictor was split into 5 quantiles for classification. After resampling, the priors were 0.497 and 0.503 for the negative and positive classes respectively for white wine; for red wine, the priors were 0.501 and 0.499 for the negative and positive classes. The Naïve Bayes models had an accuracy of 0.745 for white wine and 0.810 red, while the ROC AUC score is 0.80 and 0.87, respectively (Figures 10.1-10.2).

The Feature importance analysis revealed that red wines in the top quintile were 26 times more likely to be good quality rather than poor quality (Table 6). In contrast, white wines in the middle quintile were 3.5 times more likely to be poor quality. Alcohol percentage dominated the feature importance of white wines, while sulphates, citric acidity, and volatile acidity were other important factors for red wines.

Conclusion
KNN performed the best of all models in every measure, followed closely by XG Boost. This could be for many reasons, such as dataset noise, the size of the data, or localized patterns amongst variables.

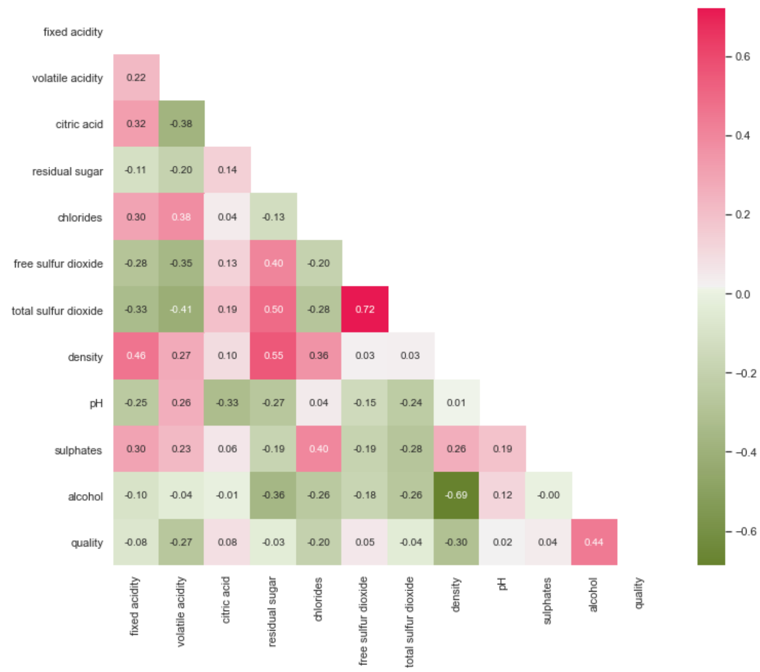
When determining the quality of wine there are key elements to focus on. For example, the Naïve Bayes, XG Boost, and Logistic Regression rank alcohol as the most important predictor. Behind alcohol, sulphates, and volatile acidity were the next important for red wines, which is supported by the EDA. Red wine quality decreased as volatile acidity increased, while it increased when citric acidity and sulphate content increased. White wine was similarly affected by volatile acidity, but less so by citric acidity and sulphates. In practice, sulphates can create a "stinkiness" to a wine, which can play in a red wine's flavor as it exhibits a richer, more complex flavor to some. Volatile acidity has to do with the "vinegar" flavor in a wine, with a higher vinegar flavor putting taste testers off of wine. Taste testers seem to like a citric acid flavor, which may have to do with how well a wine preserves the "fruitiness" of its fermented grapes.

It is recommended to buy a citrusy, not-so-vinegary, slightly stinky red wine with a good alcohol content to impress customers. And as far as white wines go, a high-alcohol content is the main consideration, although one must factor in the impact of the vinegar flavor when making purchasing decisions.

Alicia Wilson, Teagan Milford, Aditya Chawla,
Joseph Bridges, Sankalp Kulkarni, Carson Mullen

Figures and Tables
Figures

Figure 1:



Figures 2 (2.1 to 2.11):

Figure 2.1

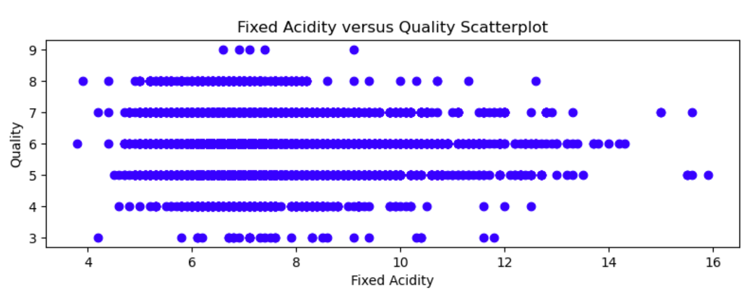


Figure 2.2

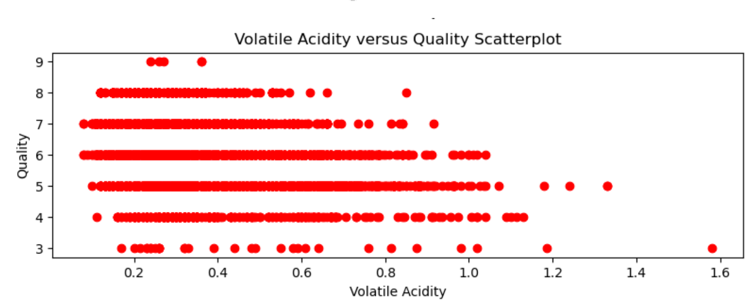


Figure 2.3

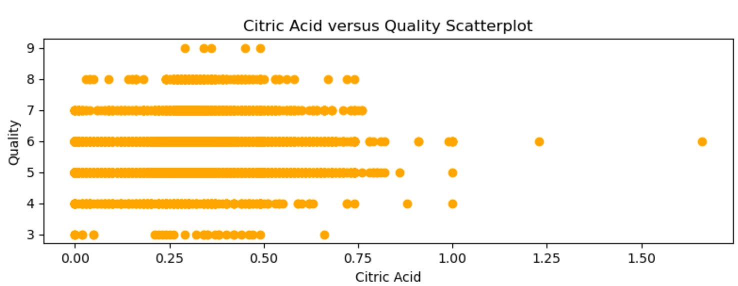


Figure 2.4

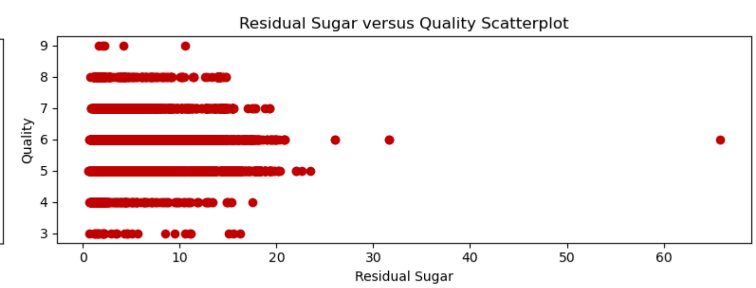


Figure 2.5

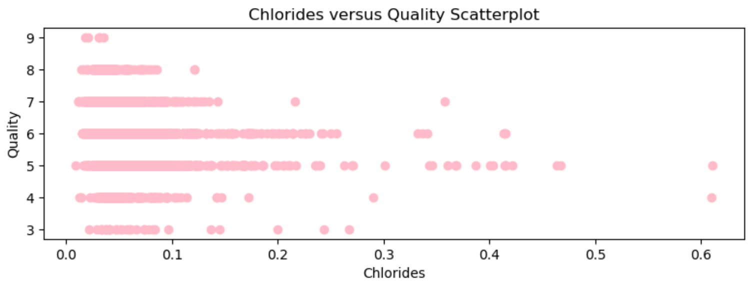


Figure 2.6

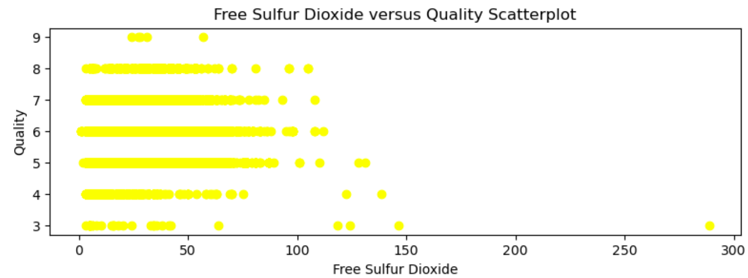


Figure 2.7

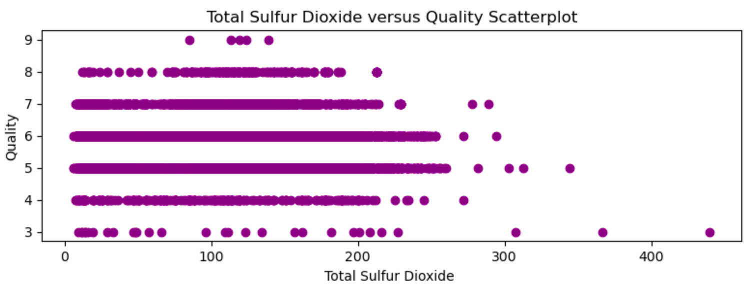


Figure 2.8

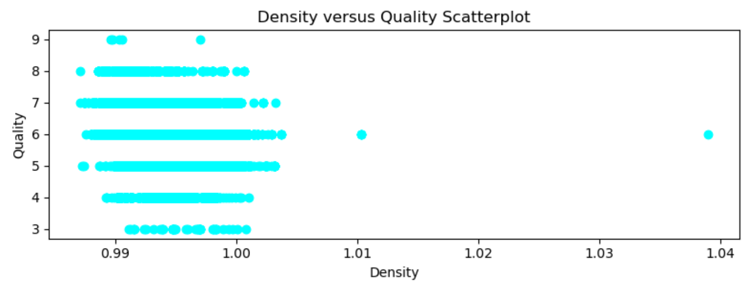


Figure 2.9

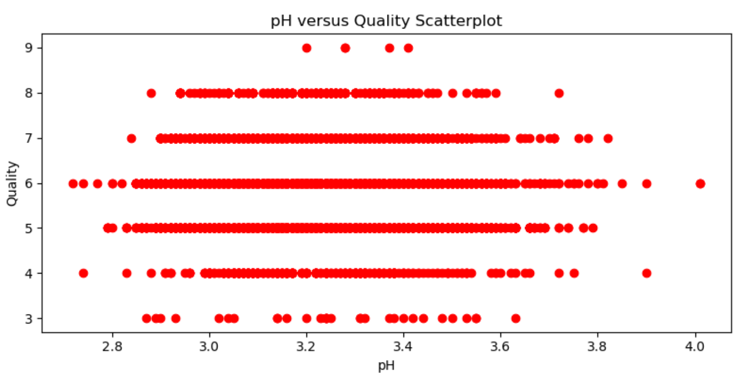


Figure 2.10

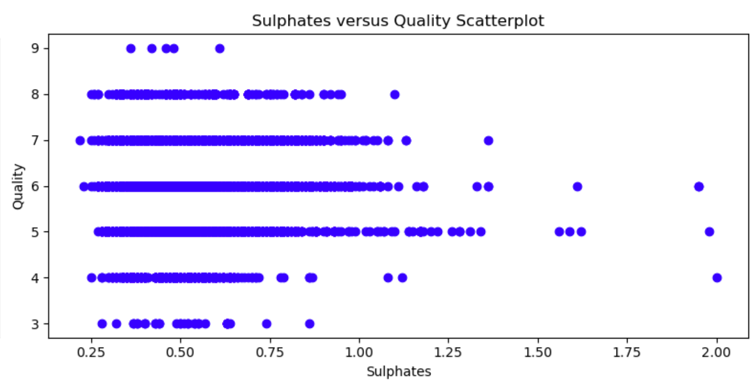
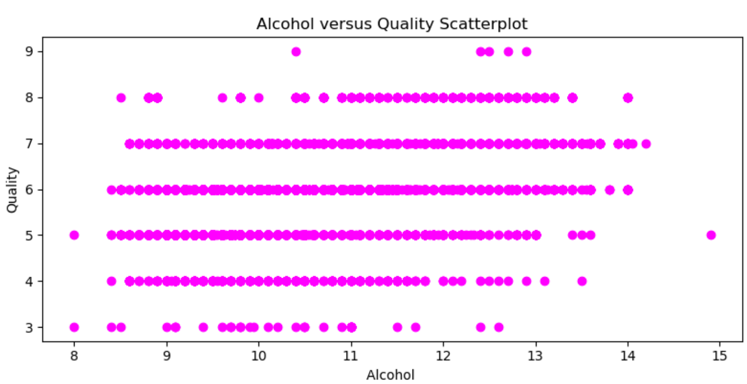


Figure 2.11



Figures 3 (3.1 to 3.11):

Figure 3.1

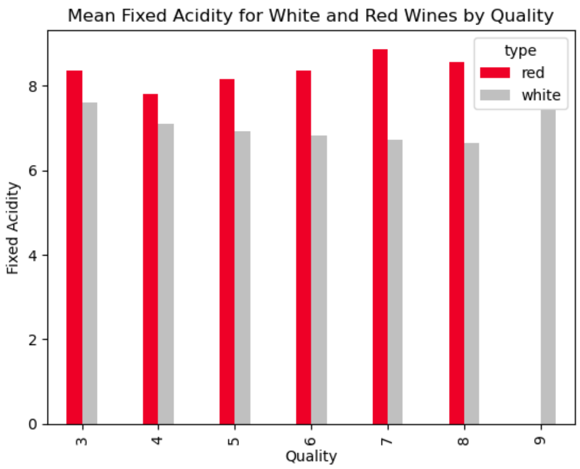


Figure 3.2

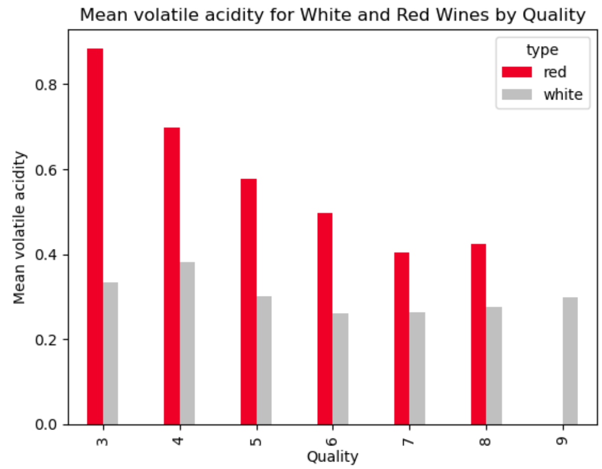


Figure 3.3

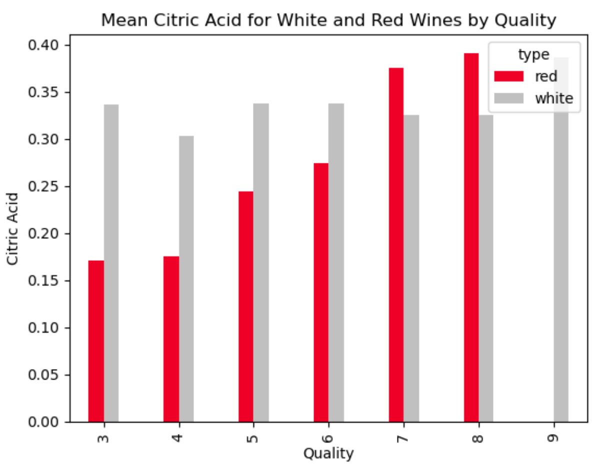


Figure 3.4

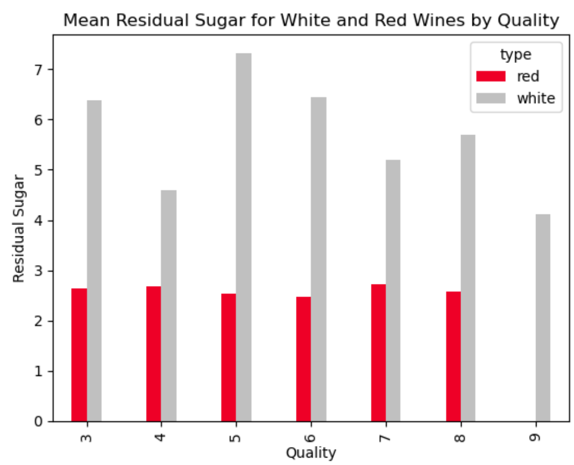


Figure 3.5

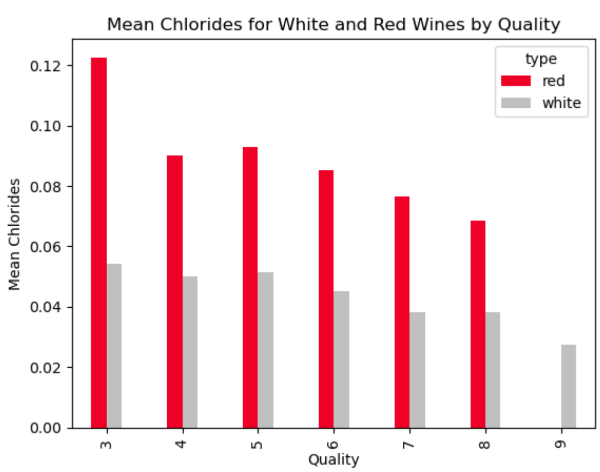


Figure 3.6

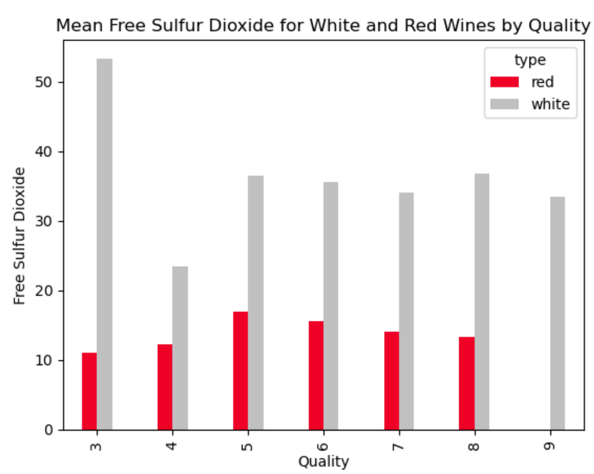


Figure 3.7

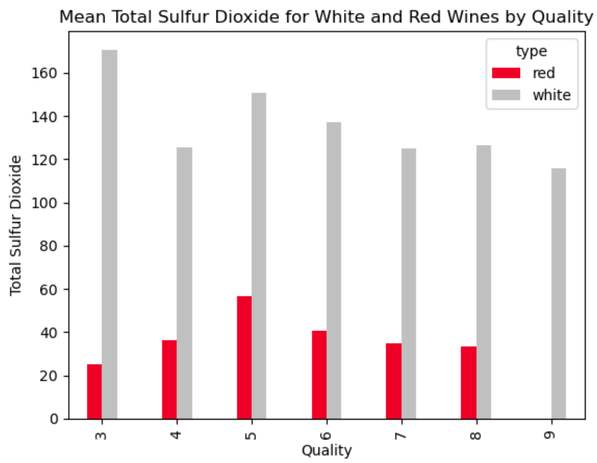


Figure 3.8

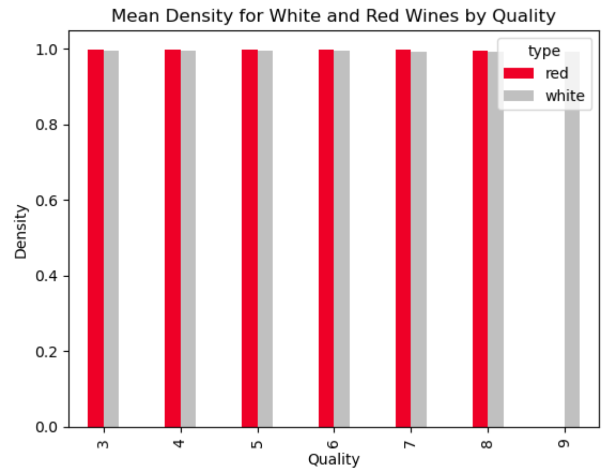


Figure 3.9

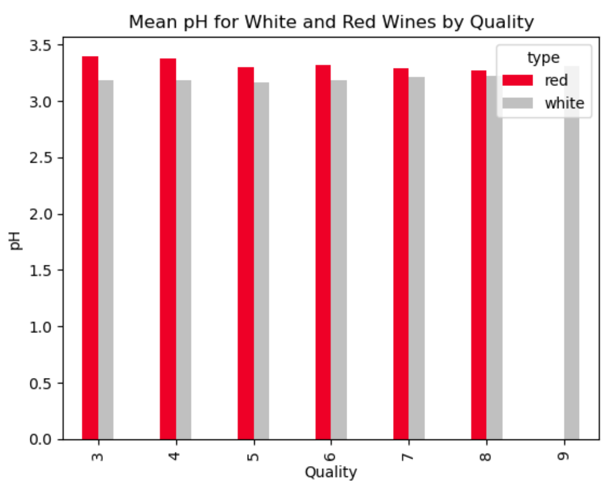


Figure 3.10

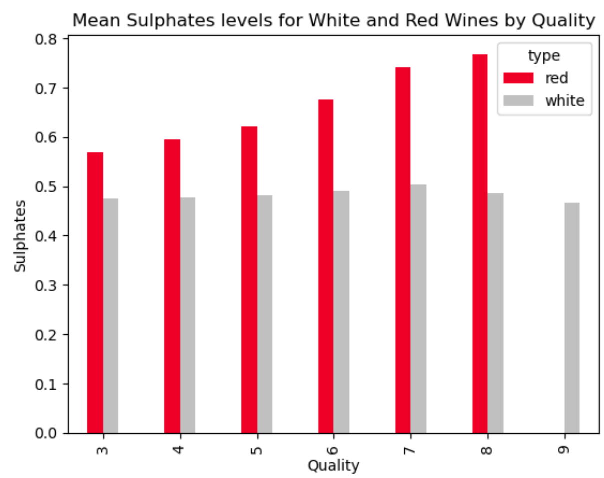
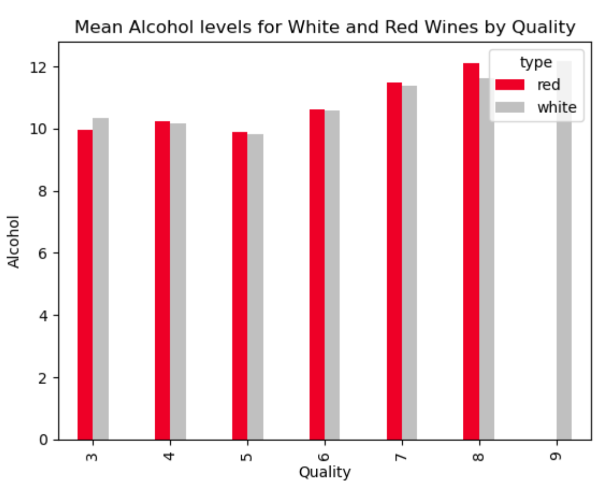


Figure 3.11



Figures 4.384595: Class Imbalance observed for "Quality" by type

Figure 4.1 Quality Categories for Red and White Wines

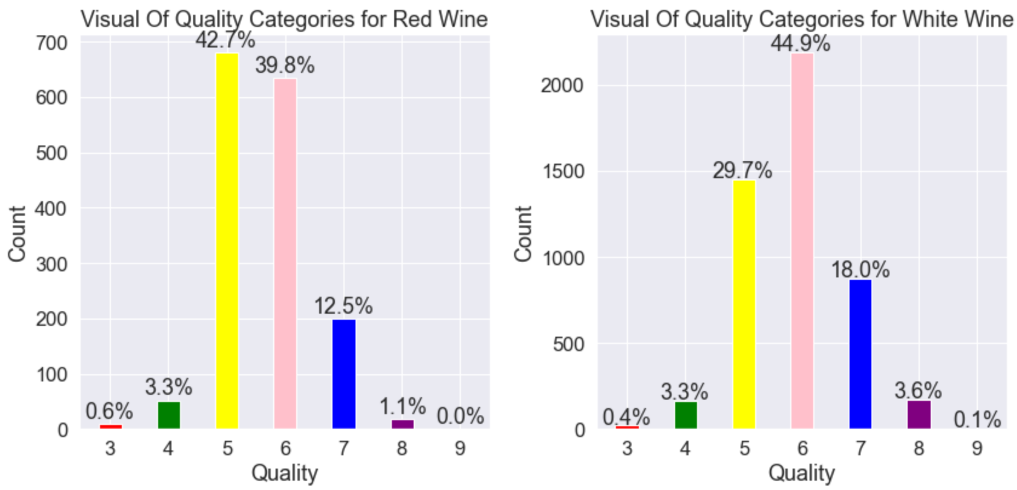


Figure 4.2: Value counts of minority class (1) before SMOTE

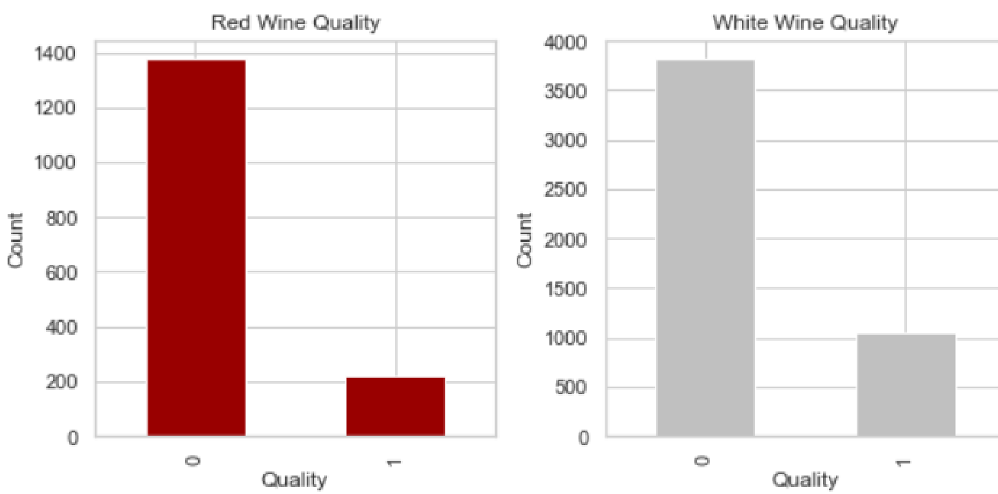
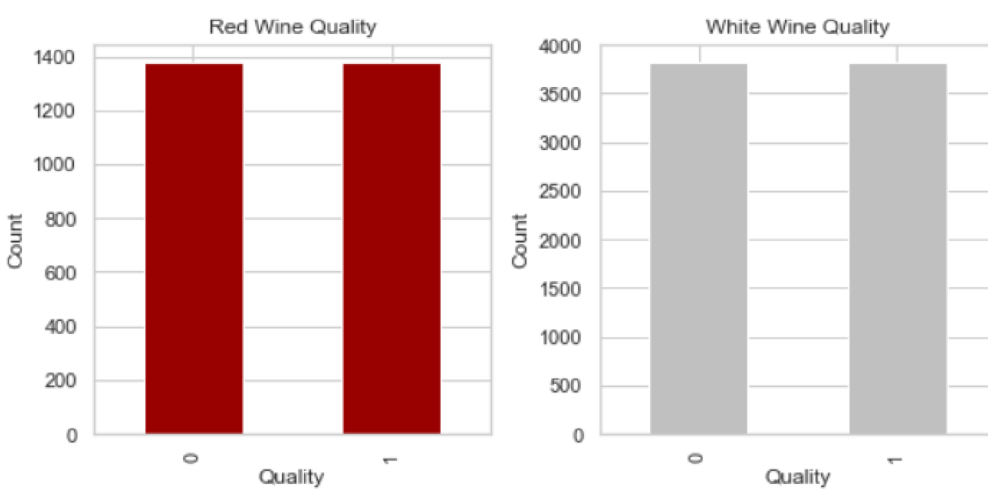


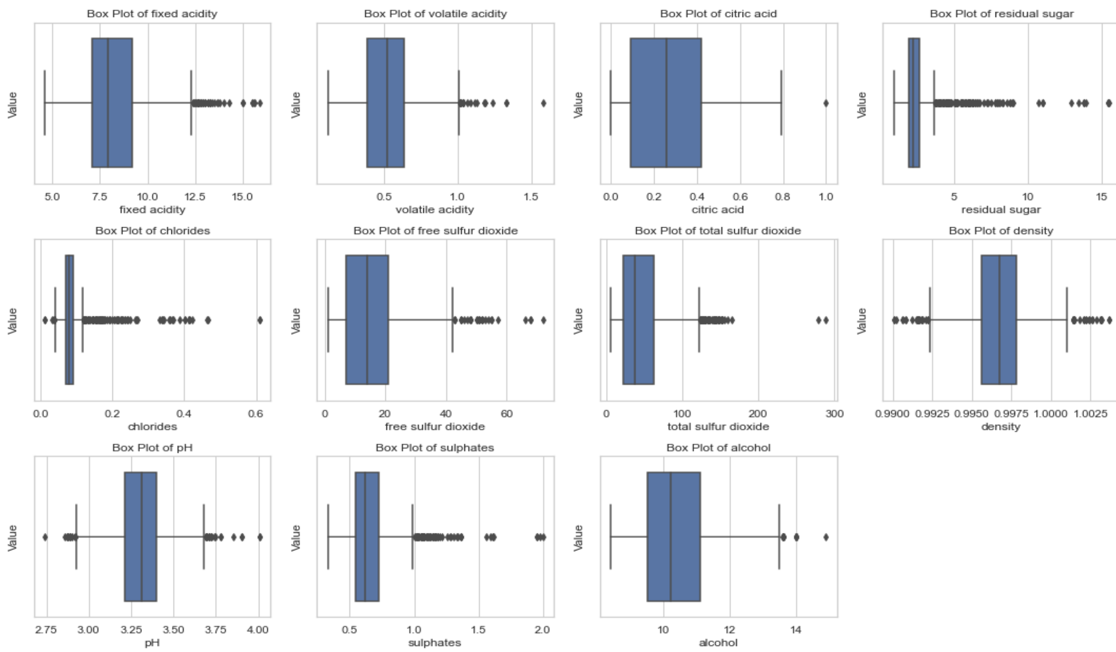
Figure 4.3 Value counts of minority class (1) after SMOTE for Naïve Bayes and XG Boost



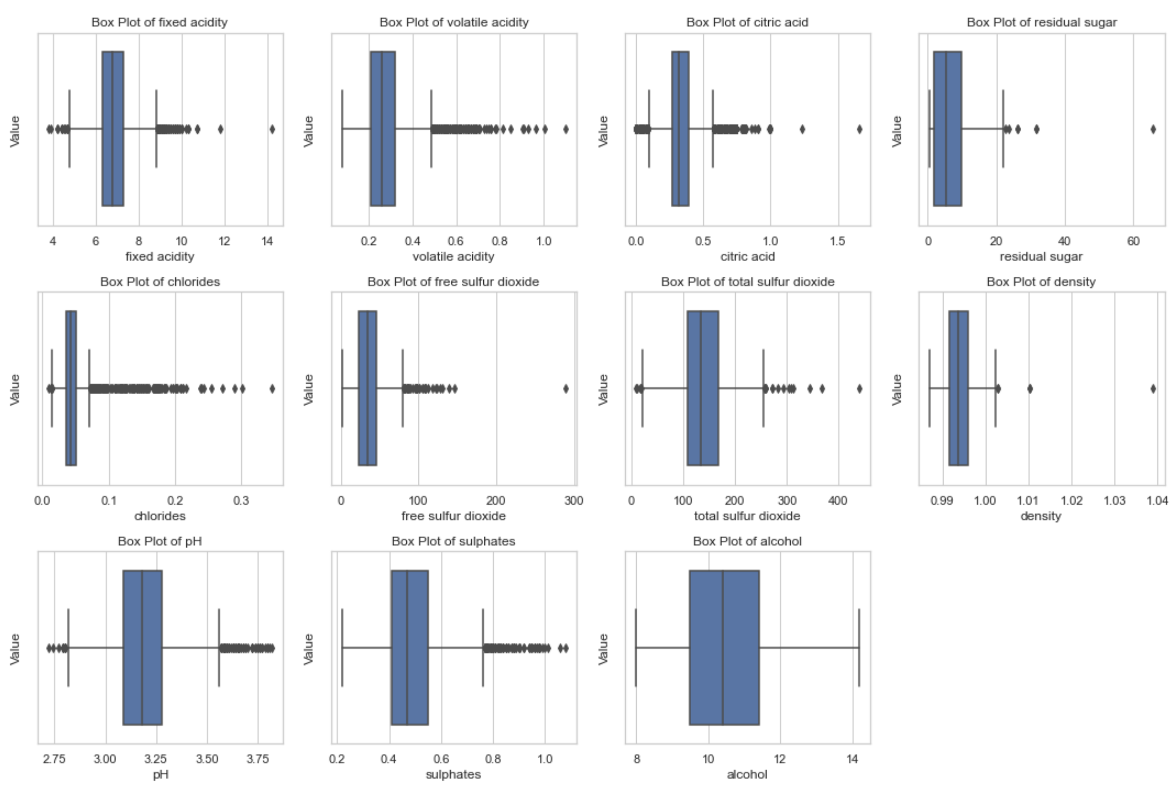
Alicia Wilson, Teagan Milford, Aditya Chawla,
Joseph Bridges, Sankalp Kulkarni, Carson Mullen

Figure 5 : Outlier detection using Box Plots

<u>Figure 5.1: Red Wine Box Plots:</u>



<u>Figure 5.2: White Wine Box Plots:</u>



Figures 6.1 to 6.3: Distributions of variables before and after standardization

Figure 6.1: Fixed Acidity

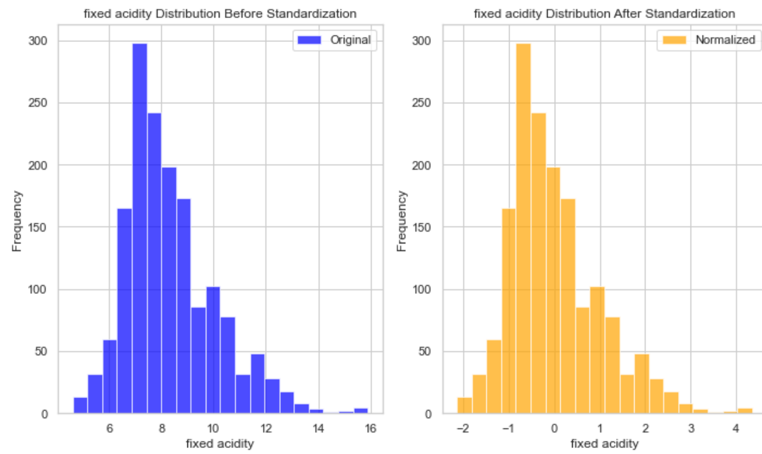


Figure 6.2: Volatile Acidity

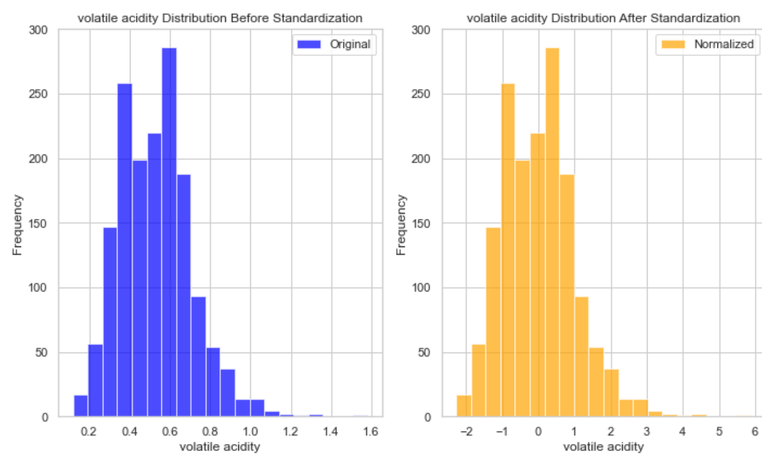


Figure 6.3: Citric Acid

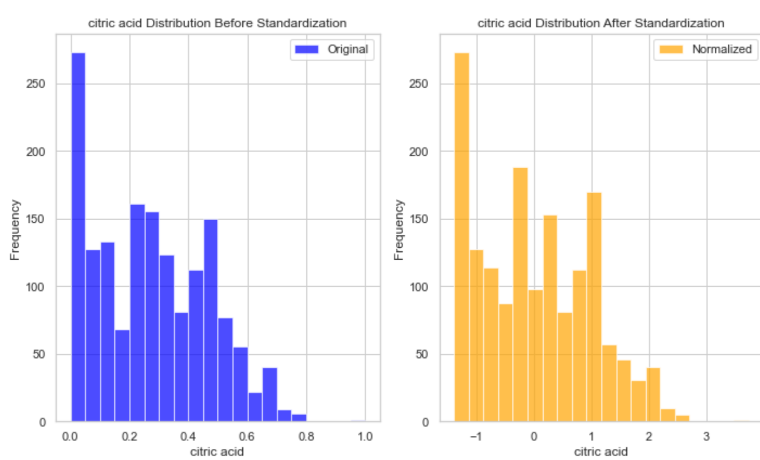


Figure 7: Logistic Regression

Figure 7.1: White Wine Confusion Matrix

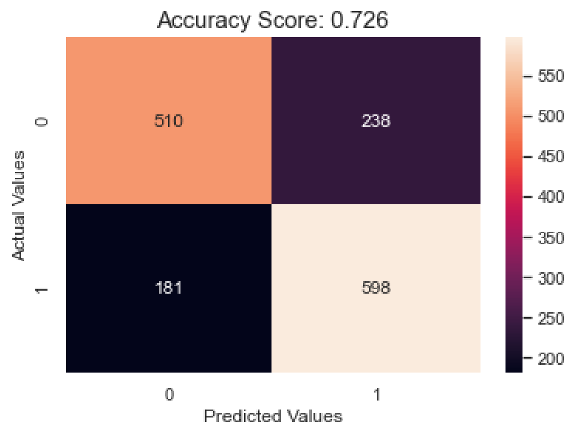


Figure 7.2: Red Wine Confusion Matrix

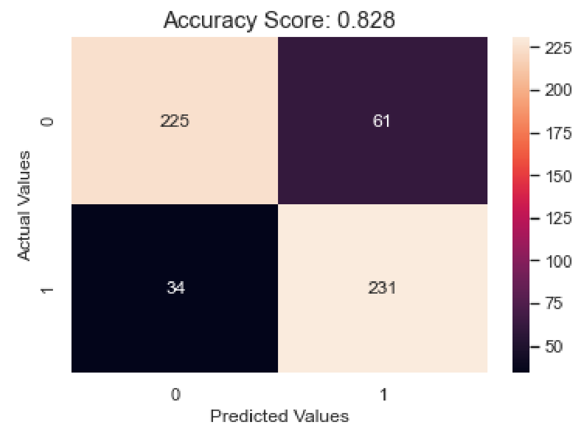


Figure 7.3: White Wine ROC Curve

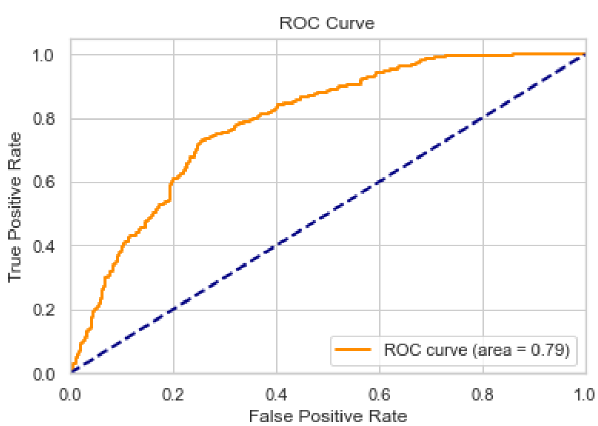


Figure 7.4: Red Wine ROC Curve

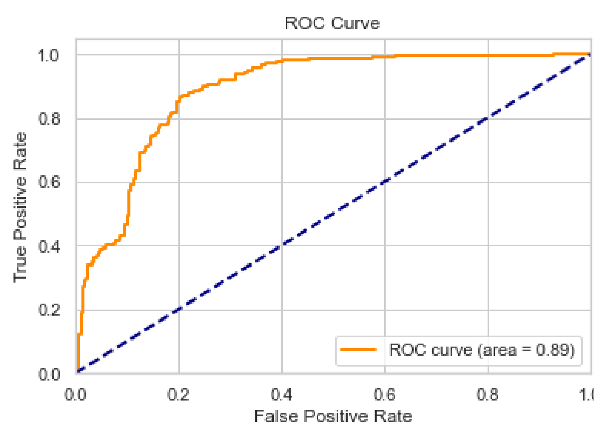


Figure 7.5: White Wine Precision Recall Curve

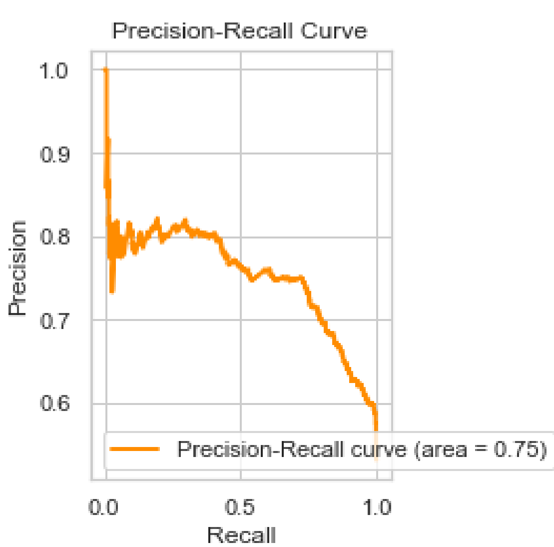
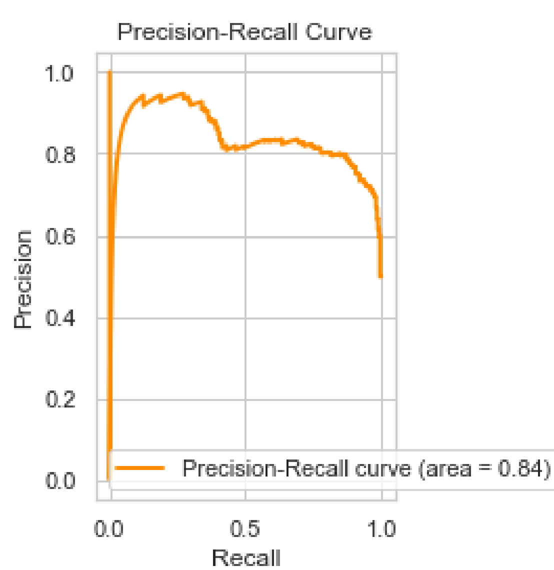


Figure 7.6: Red Wine Precision Recall Curve



Figures 8.1 to 8.4: KNN Model Figures

Figure 8.1: Red Wine Cross Validation:

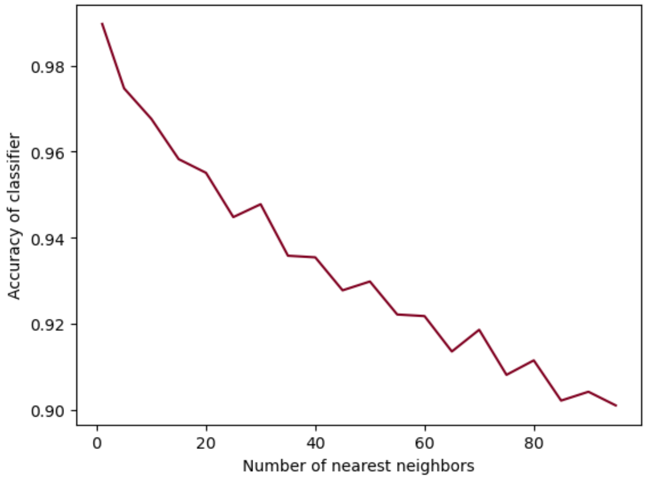


Figure 8.2: White Wine Cross Validation:

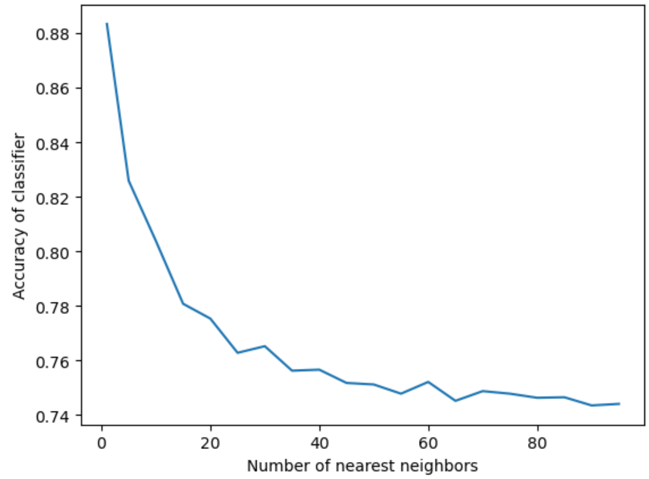


Figure 8.3: White Wine ROC Curve

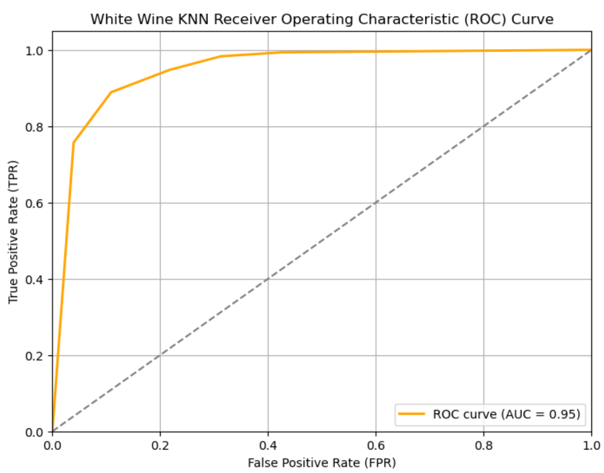
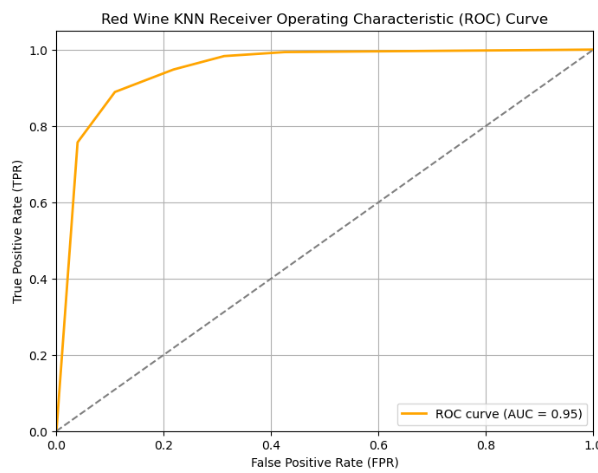


Figure 8.4: Red Wine ROC Curve



Figures 9.1 to 9.3: XGBoost Model Figures

Figure 9.1: White Wine ROC Curve

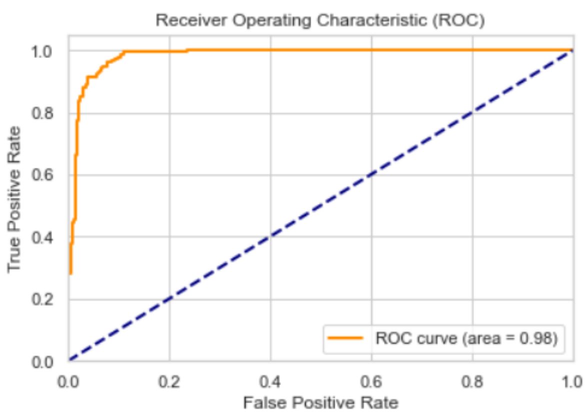


Figure 9.2: Red Wine ROC Curve

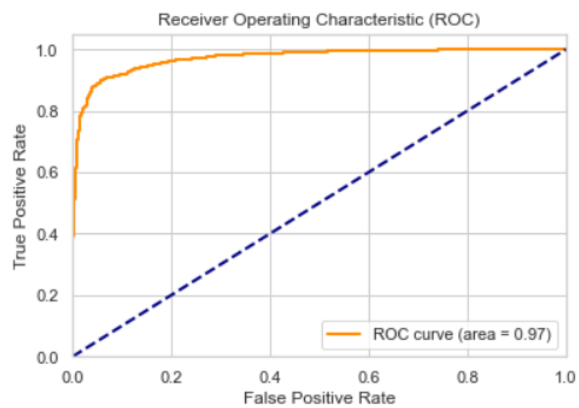


Figure 9.3: Feature Importance Comparison Plot

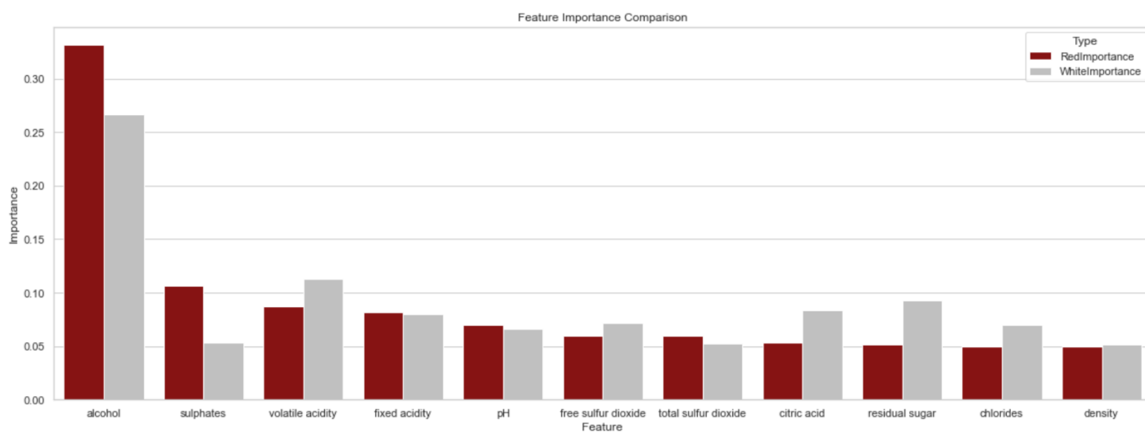


Figure 10 Naïve Bayes Models

Figure 10.1: White Wine ROC Curve

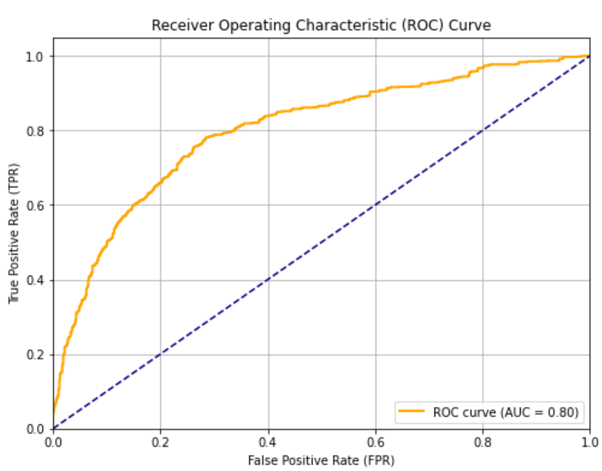
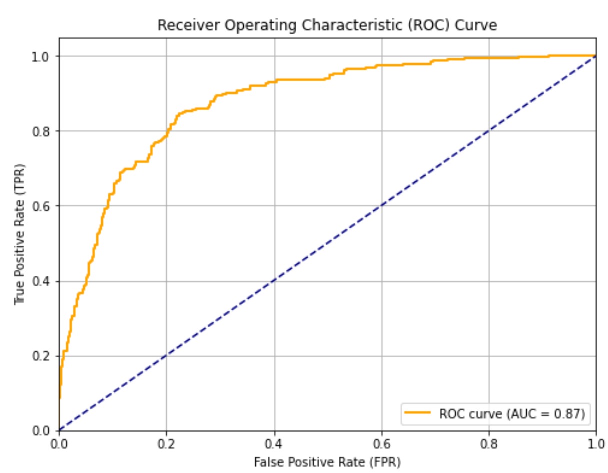


Figure 10.2: Red Wine ROC Curve



Tables

Tables 1.1 to 1.3: Correlations

Table 1.1: Combined Wine Quality Correlations

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
quality	-0.076174	-0.266677	0.084926	-0.034654	-0.200553	0.054924	-0.041598	-0.304447	0.018403	0.039054	0.444637	1.0

Table 1.2: White Wine Quality Correlations

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
quality	-0.112172	-0.196657	-0.010079	-0.094923	-0.210754	0.007747	-0.174597	-0.305481	0.097292	0.054241	0.435383	1.0

Table 1.3: Red Wine Quality Correlations

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
quality	0.123564	-0.391075	0.226345	0.01531	-0.127827	-0.05288	-0.186888	-0.174402	-0.056138	0.251363	0.477355	1.0

Tables 2.1 to 2.2: Mean Value Splits

Table 2.1: Mean Red/White Split Mean Values

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
type												
red	8.326365	0.527659	0.271551	2.538512	0.087497	15.841808	46.463905	0.996752	3.310590	0.657866	10.419617	5.636535
white	6.855123	0.278071	0.334199	6.394343	0.045771	35.317146	138.340144	0.994026	3.188154	0.489700	10.516772	5.878029

Table 2.2: Mean Red/White Split Mean Values by Quality Rating

		fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
type	**quality**											
red	**3**	8.360000	0.884500	0.171000	2.635000	0.122500	11.000000	24.900000	0.997464	3.398000	0.570000	9.955000
	4	7.805769	0.697115	0.175769	2.671154	0.090058	12.230769	36.346154	0.996537	3.379231	0.595000	10.250962
	5	8.170294	0.576853	0.243897	2.528456	0.092753	16.989706	56.555882	0.997105	3.304515	0.620912	9.899265
	6	8.357729	0.497169	0.274732	2.478628	0.085063	15.619874	40.771293	0.996626	3.317555	0.674748	10.622923
	7	8.872362	0.403920	0.375176	2.720603	0.076588	14.045226	35.020101	0.996104	3.290754	0.741256	11.465913
	8	8.566667	0.423333	0.391111	2.577778	0.068444	13.277778	33.444444	0.995212	3.267222	0.767778	12.094444
white	**3**	7.600000	0.333250	0.336000	6.392500	0.054300	53.325000	170.600000	0.994884	3.187500	0.474500	10.345000
	4	7.112963	0.381358	0.303272	4.591975	0.050173	23.478395	125.540123	0.994269	3.184691	0.477222	10.152778
	5	6.934945	0.301920	0.337990	7.322548	0.051584	36.438881	150.845994	0.995257	3.168909	0.481699	9.810345
	6	6.838472	0.260357	0.337946	6.449108	0.045189	35.650503	137.025160	0.993961	3.188408	0.491070	10.578484
	7	6.735486	0.262931	0.325657	5.206857	0.038174	34.131429	125.052571	0.992454	3.213394	0.502891	11.372210
	8	6.659770	0.275201	0.325632	5.697126	0.038368	36.764368	126.459770	0.992253	3.218046	0.486782	11.629310
	9	7.420000	0.298000	0.386000	4.120000	0.027400	33.400000	116.000000	0.991460	3.308000	0.466000	12.180000

Table 3: Count of outliers for each column

Table 3.1: Red Wine

	Outlier count
fixed acidity	49
volatile acidity	19
citric acid	1
residual sugar	154
chlorides	111
free sulfur dioxide	30
total sulfur dioxide	55
density	45
pH	35
sulphates	59
alcohol	13

Table 3.2: White Wine

	Outlier count
fixed acidity	118
volatile acidity	182
citric acid	266
residual sugar	7
chlorides	206
free sulfur dioxide	50
total sulfur dioxide	19
density	5
pH	75
sulphates	124
alcohol	0

Tables 4.1 to 4.2: Winsorized Statistics

Table 4.1: Red Wine

```
Original Statistics for Red Wine:
                      Original Minimum  Original Maximum
fixed acidity                  4.60000          15.90000
volatile acidity               0.12000           1.58000
citric acid                    0.00000           1.00000
residual sugar                 0.90000          15.50000
chlorides                      0.01200           0.61100
free sulfur dioxide            1.00000          72.00000
total sulfur dioxide           6.00000         289.00000
density                        0.99007           1.00369
pH                             2.74000           4.01000
sulphates                      0.33000           2.00000
alcohol                        8.40000          14.90000

Winsorized Statistics for Red Wine:
                      Winsorized Minimum  Winsorized Maximum
fixed acidity                    6.10000              11.800
volatile acidity                 0.27000               0.840
citric acid                      0.00000               0.600
residual sugar                   1.50000               5.100
chlorides                        0.05400               0.127
free sulfur dioxide              4.00000              35.000
total sulfur dioxide            11.00000             113.000
density                          0.99358               1.000
pH                               3.06000               3.570
sulphates                        0.47000               0.930
alcohol                          9.20000              12.500
```

Table 4.2: White Wine

Original Statistics for White Wine:

	Original Minimum	Original Maximum
fixed acidity	3.80000	14.20000
volatile acidity	0.08000	1.10000
citric acid	0.00000	1.66000
residual sugar	0.60000	65.80000
chlorides	0.00900	0.34600
free sulfur dioxide	2.00000	289.00000
total sulfur dioxide	9.00000	440.00000
density	0.98711	1.03898
pH	2.72000	3.82000
sulphates	0.22000	1.08000
alcohol	8.00000	14.20000

Winsorized Statistics for White Wine:

	Winsorized Minimum	Winsorized Maximum
fixed acidity	5.60000	8.300
volatile acidity	0.15000	0.460
citric acid	0.18000	0.540
residual sugar	1.10000	15.700
chlorides	0.02700	0.067
free sulfur dioxide	11.00000	63.000
total sulfur dioxide	75.00000	212.000
density	0.98963	0.999
pH	2.96000	3.460
sulphates	0.34000	0.710
alcohol	8.90000	12.700

Tables 5.1 to 5.2: VIF

Table 5.1 Pre VIF

VIF for red_wine DataFrame:

	Feature	VIF
0	fixed acidity	6.968489
1	volatile acidity	1.920236
2	citric acid	3.192142
3	residual sugar	1.746056
4	chlorides	1.293189
5	free sulfur dioxide	2.113477
6	total sulfur dioxide	2.395339
7	density	6.242264
8	pH	3.086011
9	sulphates	1.285805
10	alcohol	3.180446

VIF for white_wine DataFrame:

	Feature	VIF
0	fixed acidity	2.553361
1	volatile acidity	1.119840
2	citric acid	1.148013
3	residual sugar	12.453515
4	chlorides	1.533300
5	free sulfur dioxide	1.845177
6	total sulfur dioxide	2.375466
7	density	31.234441
8	pH	2.040998
9	sulphates	1.127897
10	alcohol	9.252880

Table 5.2 Post VIF

```
VIF for red_wine DataFrame:
              Feature       VIF
0       fixed acidity  2.706642
1    volatile acidity  1.844281
2         citric acid  3.019066
3      residual sugar  1.115999
4           chlorides  1.254553
5 free sulfur dioxide  1.068024
6                  pH  2.101133
7           sulphates  1.219607
8             alcohol  1.269899

VIF for white_wine DataFrame:
              Feature       VIF
0       fixed acidity  1.305252
1    volatile acidity  1.053850
2         citric acid  1.136212
3      residual sugar  1.444180
4           chlorides  1.455953
5 free sulfur dioxide  1.195722
6                  pH  1.287186
7           sulphates  1.040567
8             alcohol  1.799796
```

Now this variable set does not have multicolinearity

Tables 5.1 to 5.2: Logistic Regression Coefficients

Table 5.1: Red Wine Logistic Regression Coefficients

	Red Wine Logistic Regression Coefficients
fixed acidity	0.453466
volatile acidity	-0.833760
citric acid	-0.365775
residual sugar	0.110526
chlorides	-0.324368
free sulfur dioxide	-0.162080
pH	-0.095740
sulphates	0.552123
alcohol	1.308776

Table 5.2: White Wine Logistic Regression Coefficients

	White Wine Logistic Regression Coefficients
fixed acidity	0.089200
volatile acidity	-0.628413
citric acid	-0.197986
residual sugar	0.431270
chlorides	-0.398669
free sulfur dioxide	0.117365
pH	0.126882
sulphates	0.178926
alcohol	1.218876

Tables 6.1 to 6.2: Feature of Importance for Naïve Bayes

Table 6.1: Red Wine Feature of Importance

	Positive class	Negative class	Positive/Negative Ratio	Importance
Q("alcohol_binned_(8.399000000000001, 9.7]")	0.001600	0.037301	0.042890	3.149106
Q("sulphates_binned_(0.329, 0.57]")	0.005552	0.033833	0.164112	1.807208
Q("volatile_acidity_binned_(0.61, 1.58]")	0.005364	0.030366	0.176654	1.733560
Q("alcohol_binned_(11.9, 14.9]")	0.031903	0.005811	5.490383	1.702998
Q("volatile_acidity_binned_(0.119, 0.324]")	0.029362	0.007310	4.016563	1.390427
Q("sulphates_binned_(0.72, 0.8]")	0.028797	0.008529	3.376561	1.216858
Q("citric_acid_binned_(-0.001, 0.1]")	0.009599	0.027366	0.350762	1.047649
Q("density_binned_(0.989, 0.995]")	0.026162	0.009560	2.736776	1.006781
Q("alcohol_binned_(11.198, 11.9]")	0.025598	0.009934	2.576663	0.946495
Q("total_sulfur_dioxide_binned_(60.0, 289.0]")	0.010540	0.026804	0.393230	0.933361
Q("alcohol_binned_(9.7, 10.5]")	0.010823	0.025773	0.419913	0.867707
Q("citric_acid_binned_(0.1, 0.266]")	0.010446	0.024649	0.423801	0.858492
Q("volatile_acidity_binned_(0.324, 0.4]")	0.027197	0.012184	2.232282	0.803024
Q("citric_acid_binned_(0.49, 1.0]")	0.023151	0.010403	2.225393	0.799934
Q("sulphates_binned_(0.8, 2.0]")	0.024939	0.011528	2.163393	0.771678
Q("sulphates_binned_(0.57, 0.64]")	0.011669	0.024461	0.477063	0.740107
Q("chlorides_binned_(0.011, 0.0644]")	0.023527	0.011340	2.074671	0.729803
Q("fixed_acidity_binned_(4.598999999999999, 7.1]")	0.012328	0.025398	0.485396	0.722789
Q("volatile_acidity_binned_(0.5, 0.61]")	0.011481	0.021743	0.528040	0.638584
Q("citric_acid_binned_(0.39, 0.49]")	0.025598	0.013777	1.858002	0.619502
Q("pH_binned_(3.4, 4.01]")	0.012516	0.022587	0.554152	0.590316
Q("fixed_acidity_binned_(10.0, 15.9]")	0.022586	0.012559	1.798461	0.586931
Q("alcohol_binned_(10.5, 11.198]")	0.020986	0.012090	1.735840	0.551492
Q("density_binned_(0.998, 1.004]")	0.013928	0.022962	0.606583	0.499914
Q("chlorides_binned_(0.091, 0.611]")	0.013834	0.022118	0.625461	0.469267

Table 6.2: White Wine Feature of Importance

	Positive class	Negative class	Positive/Negative Ratio	Importance
Q("alcohol_binned_(12.1, 14.2]")	0.028692	0.008162	3.515326	1.257132
Q("alcohol_binned_(7.999, 9.5]")	0.009361	0.031282	0.299254	1.206462
Q("alcohol_binned_(9.5, 10.454]")	0.008043	0.024076	0.334075	1.096391
Q("density_binned_(0.986, 0.991]")	0.027104	0.009391	2.886007	1.059874
Q("chlorides_binned_(0.05, 0.346]")	0.009158	0.024384	0.375601	0.979228
Q("alcohol_binned_(11.2, 12.1]")	0.026157	0.010279	2.544663	0.933998
Q("density_binned_(0.994, 0.996]")	0.010375	0.026159	0.396612	0.924798
Q("total_sulfur_dioxide_binned_(167.15, 440.0]")	0.010240	0.025511	0.401400	0.912796
Q("density_binned_(0.996, 1.039]")	0.010713	0.025067	0.427385	0.850071
Q("residual_sugar_binned_(1.6, 2.934]")	0.023893	0.010826	2.207069	0.791665
Q("chlorides_binned_(0.008, 0.0325]")	0.025448	0.011577	2.198118	0.787602
Q("density_binned_(0.991, 0.992]")	0.025177	0.011577	2.174765	0.776921
Q("chlorides_binned_(0.0325, 0.0374]")	0.024637	0.011953	2.061177	0.723277
Q("chlorides_binned_(0.043, 0.05]")	0.012639	0.024247	0.521278	0.651473
Q("citric_acid_binned_(-0.001, 0.27]")	0.014532	0.027662	0.525338	0.643713
Q("citric_acid_binned_(0.3, 0.334]")	0.019263	0.011031	1.746337	0.557521
Q("residual_sugar_binned_(10.4, 65.8]")	0.013315	0.022778	0.584556	0.536903
Q("citric_acid_binned_(0.27, 0.3]")	0.022643	0.013319	1.700067	0.530667
Q("citric_acid_binned_(0.39, 1.66]")	0.014025	0.022949	0.611131	0.492443
Q("free_sulfur_dioxide_binned_(47.0, 289.0]")	0.013282	0.021344	0.622255	0.474406
Q("residual_sugar_binned_(2.934, 6.297]")	0.022203	0.014002	1.585759	0.461063
Q("fixed_acidity_binned_(7.4, 14.2]")	0.012673	0.020012	0.633271	0.456857
Q("free_sulfur_dioxide_binned_(1.999, 22.0]")	0.015005	0.022505	0.666735	0.405362
Q("sulphates_binned_(0.58, 1.08]")	0.020412	0.013933	1.464983	0.381844
Q("pH_binned_(3.32, 3.82]")	0.020210	0.013797	1.464791	0.381712

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016.

The advocacy group for the California Wine Industry. Wine Institute. https://wineinstitute.org/.

Vantage Market Research. *Wine market size USD 698.54 billion by 2030*. Vantage Market Research. https://www.vantagemarketresearch.com/industry-report/wine-market.