STSCI 4740 Final Report

Group: Aditya Vinodh, Muhammad Jee, Veda Kamaraju, Javohir Abdurazzakov, Yung Hsueh Lee

1. Introduction

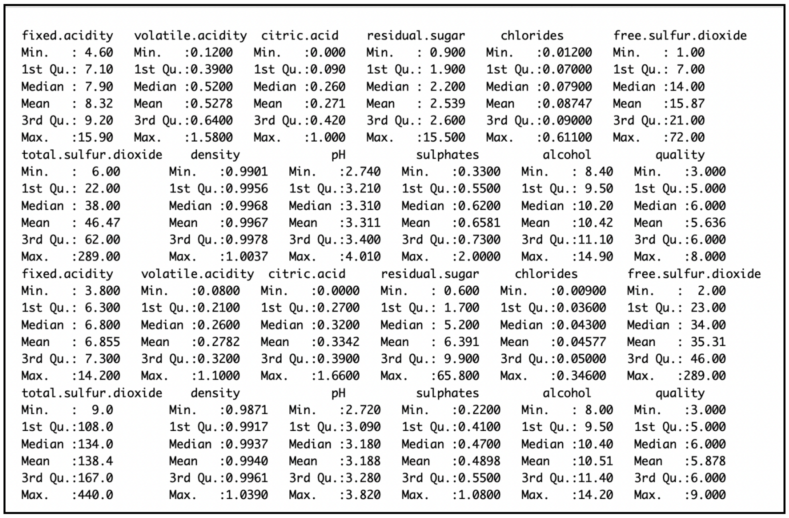
Many companies want to have the best-tasting wine in order to satisfy customers and be at the top of the wine market. Wine certification and quality assessment are crucial for ensuring consumer satisfaction and market competitiveness. However, there are many factors that determine what makes a wine of good quality. Our team aims to make a model that can predict the quality of a wine based on the 12 variables present in the dataset we used. These 12 predictors are the following: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and color.

Fixed acidity represents the total concentration of acids, particularly tartaric acid, influencing taste and stability. Volatile acidity measures the presence of volatile acids like acetic acid, indicating potential spoilage or fermentation issues. Citric acid, a naturally occurring acid found in citrus fruits, contributes to acidity and freshness, enhancing fruitiness. Residual sugar denotes the remaining sugar after fermentation, affecting sweetness levels. Chlorides reflect salt compound concentrations, impacting taste and mouthfeel. Free sulfur dioxide, an antimicrobial agent, prevents oxidation and spoilage, while total sulfur dioxide encompasses overall sulfur dioxide content, influencing aroma and stability. Density indicates mass per unit volume, offering insights into body and mouthfeel. pH measures acidity or alkalinity, affecting taste, stability, and microbial activity. Sulfates, as preservatives, inhibit microbial growth and oxidation, enhancing longevity. Alcohol content, derived from fermentation, impacts body, flavor, and warmth perception. Finally, color distinguishes between red and white wines, influencing consumer preferences. Each variable plays a distinct role in shaping wine quality and perception, contributing to comprehensive analysis and predictive modeling.

After a thorough exploratory data analysis, we decided non-linear models would perform better than linear models, which prompted us to compare the efficacy of five distinct models—Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Regularized Polynomial Regression, and Generalized Additive Models (GAM)—in predicting the quality of wine based on its composition. Our findings hold implications for wine producers, certification entities, and oenologists, offering actionable insights to enhance decision-making processes and product quality. The remainder of this report details the methods employed, presents the model design and results produced, and concludes with implications of the variables with the most impact.

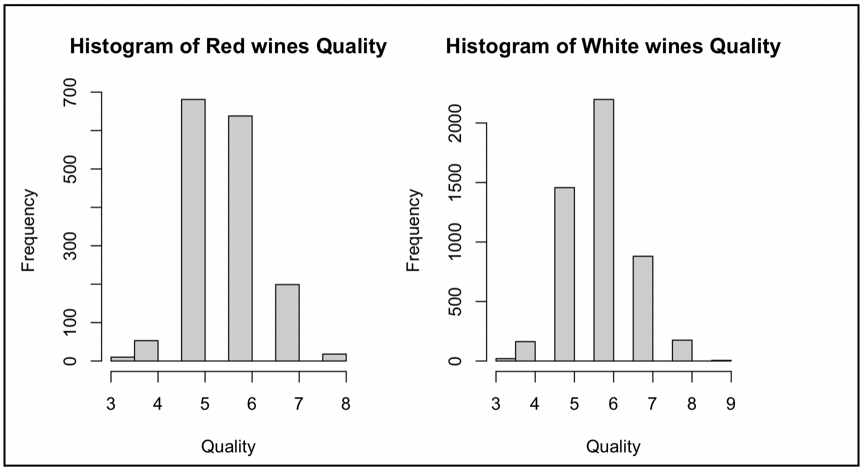
2. Exploratory Data Analysis

We began our Exploratory Data Analysis by conducting a preliminary review of the 2 datasets in terms of the number of data points, summary statistics of each predictor, and mean values of the outcome variable 'quality' in red and white wines. The number of observations of data available for red wines is 1599, while there are 4898 data points for white wines.



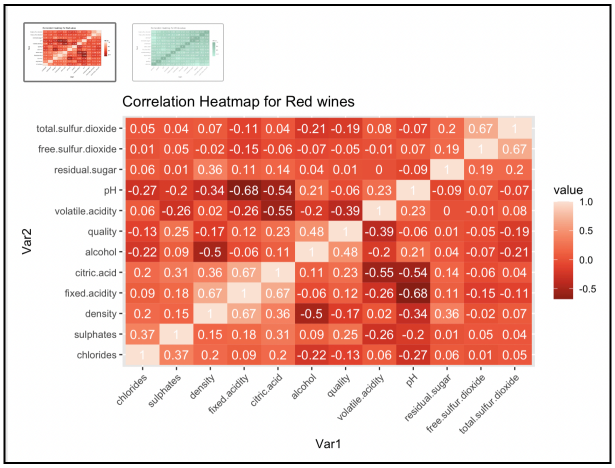
```
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00
 total.sulfur.dioxide    density             pH          sulphates         alcohol         quality
 Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40   Min.   :3.000
 1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50   1st Qu.:5.000
 Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200   Median :10.20   Median :6.000
 Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42   Mean   :5.636
 3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10   3rd Qu.:6.000
 Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90   Max.   :8.000
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar     chlorides        free.sulfur.dioxide
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600   Min.   :0.00900   Min.   :  2.00
 1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700   1st Qu.:0.03600   1st Qu.: 23.00
 Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200   Median :0.04300   Median : 34.00
 Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391   Mean   :0.04577   Mean   : 35.31
 3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900   3rd Qu.:0.05000   3rd Qu.: 46.00
 Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800   Max.   :0.34600   Max.   :289.00
 total.sulfur.dioxide    density             pH          sulphates         alcohol         quality
 Min.   :  9.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
 1st Qu.:108.0        1st Qu.:0.9917   1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
 Median :134.0        Median :0.9937   Median :3.180   Median :0.4700   Median :10.40   Median :6.000
 Mean   :138.4        Mean   :0.9940   Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
 3rd Qu.:167.0        3rd Qu.:0.9961   3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
 Max.   :440.0        Max.   :1.0390   Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

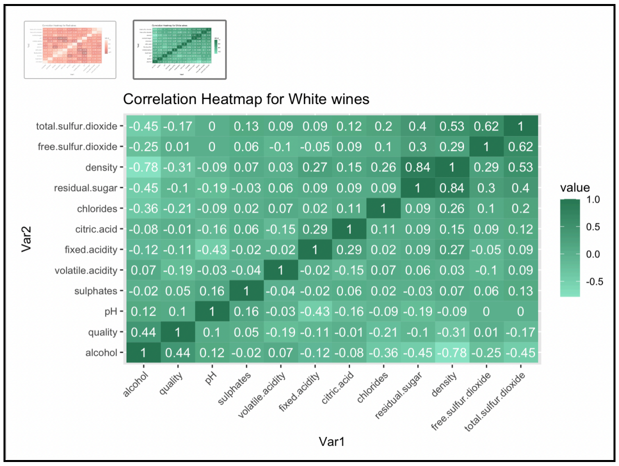
This is a summary of all the variables that make up both datasets. While it is difficult to draw any valuable conclusions from this directly, we see that the mean quality of both types of wine are fairly similar - 5.636 vs. 5.878. To further examine the distribution of the outcome variable between the 2 datasets, histograms in the following manner can be plotted:



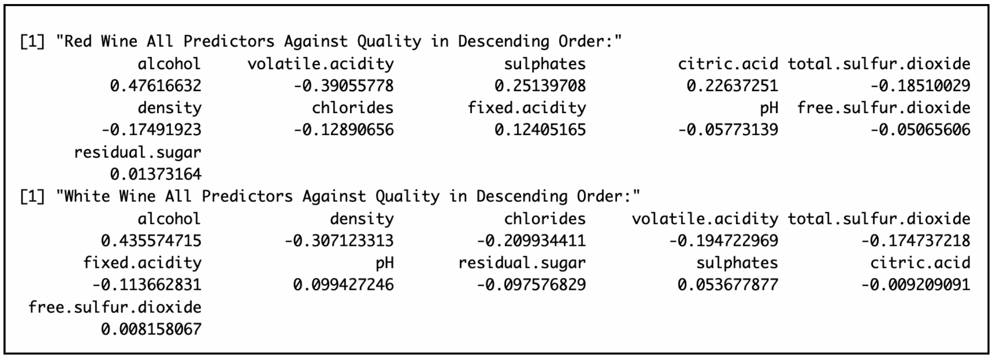
Both appear to show a normal distribution representing a peak between 5 and 6 and tapering at the ends. There seems to be slightly more variance in the quality of white wines as they spread across more evenly, while the red wines' qualities are very closely concentrated around 5 and 6. This gives us some evidence that the 2 datasets might have different underlying true models, and should therefore be approached separately.

The correlations between each of the variables were then looked at, to check for any noticeable interaction effects that stood out. This would provide us insight to construct our models while accounting for any confounding interactions that would otherwise distort the fitting of any model we apply. Thus, we plotted heatmaps to better visualize the interaction between variables in both datasets.



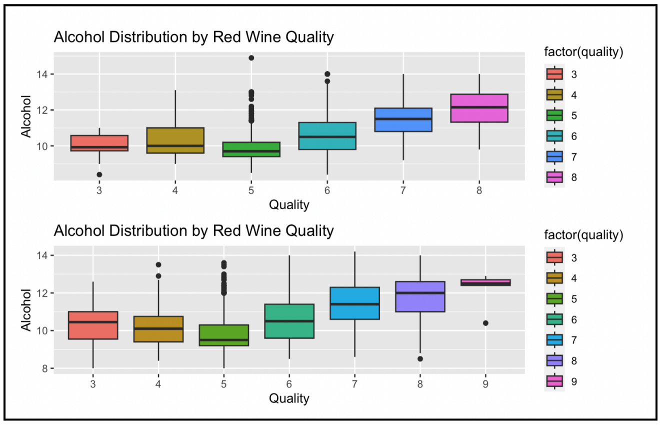


The correlations between each predictor and the outcome variable were then calculated individually. The predictors and their correlations were then sorted in decreasing order.

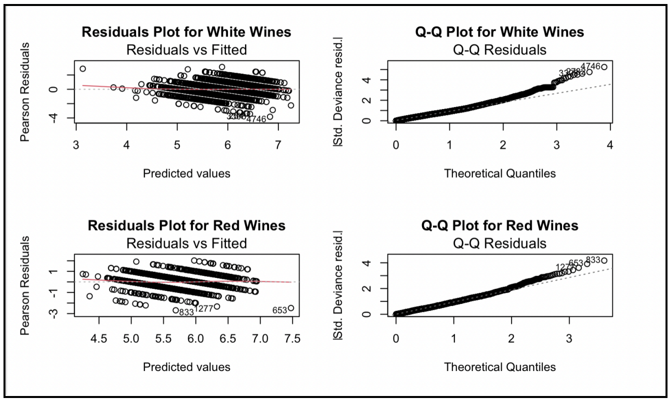


```
[1] "Red Wine All Predictors Against Quality in Descending Order:"
           alcohol     volatile.acidity            sulphates          citric.acid total.sulfur.dioxide 
        0.47616632          -0.39055778           0.25139708           0.22637251          -0.18510029 
           density            chlorides        fixed.acidity                   pH  free.sulfur.dioxide 
       -0.17491923          -0.12890656           0.12405165          -0.05773139          -0.05065606 
    residual.sugar 
        0.01373164 
[1] "White Wine All Predictors Against Quality in Descending Order:"
           alcohol              density            chlorides     volatile.acidity total.sulfur.dioxide 
       0.435574715         -0.307123313         -0.209934411         -0.194722969         -0.174737218 
     fixed.acidity                   pH       residual.sugar            sulphates          citric.acid 
      -0.113662831          0.099427246         -0.097576829          0.053677877         -0.009209091 
free.sulfur.dioxide 
       0.008158067 
```

As seen above, the variable 'alcohol' seems to be the most important feature in predicting the quality of both red and white wines. Therefore, we decided to visualize its association with wine quality through 2 boxplots to investigate any obvious trends.



To gauge the linearity of the model, we applied preliminary linear models to look at the distribution of residuals of each predicted value. A Q-Q plot was also made to examine whether residuals followed a normal distribution.



3. Methodology

This section outlines the feature selection, model selection, and evaluation metric approaches.

3.1. Model Selection

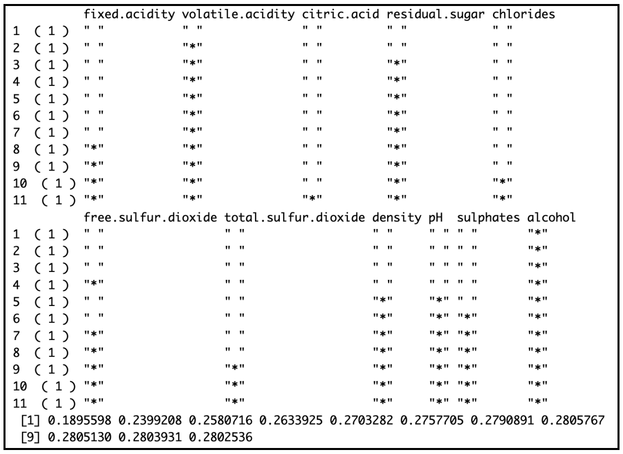
To effectively model the relationship between the 12 physicochemical features and wine quality, we explored a range of predictive techniques that could handle the characteristics of this dataset:

- Random Forest models were employed to capture potential non-linear patterns and leverage the high-dimensional feature space.
- K-Nearest Neighbors (KNN), a non-parametric algorithm, was chosen for its flexibility in modeling complex relationships without assumptions about the underlying data distribution.
- Logistic Regression enabled classification of wine quality into binary categories (high vs. low) based on the compositional features.
- Regularized Polynomial Regression allowed us to account for non-linear associations by raising predictors to the third degree, while Lasso regularization prioritized relevant features and mitigated overfitting.
- Generalized Additive Models (GAMs) offered a flexible approach by permitting non-linear relationships between predictors and wine quality, while considering each predictor's contribution separately through an additive technique.

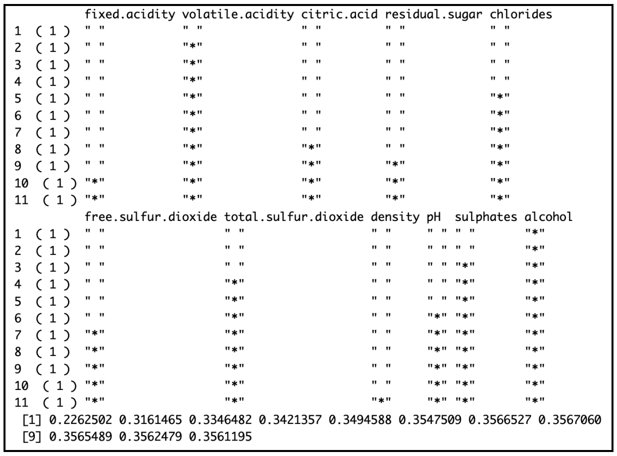
By exploring this diverse set of models, we aimed to identify the most effective approach for accurately predicting wine quality based on the unique characteristics of the dataset.

3.2. Feature Selection

To identify the most relevant predictors for accurately modeling wine quality, we utilized three feature selection strategies tailored to each model:
We began by using best subset selection, an exhaustive approach to examine all possible combinations of predictors to identify the most impactful subset based on the Bayesian Information Criterion (BIC).

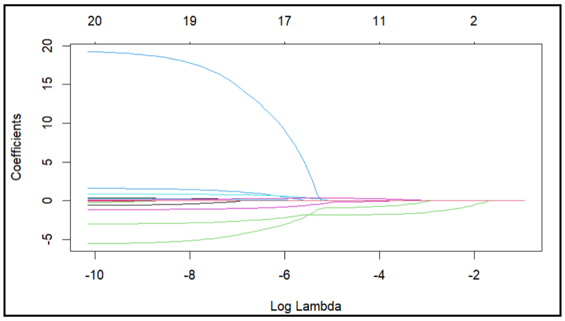


```
          fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
1  ( 1 )  " "           " "              " "         " "            " "      
2  ( 1 )  " "           "*"              " "         " "            " "      
3  ( 1 )  " "           "*"              " "         "*"            " "      
4  ( 1 )  " "           "*"              " "         "*"            " "      
5  ( 1 )  " "           "*"              " "         "*"            " "      
6  ( 1 )  " "           "*"              " "         "*"            " "      
7  ( 1 )  " "           "*"              " "         "*"            " "      
8  ( 1 )  "*"           "*"              " "         "*"            " "      
9  ( 1 )  "*"           "*"              " "         "*"            " "      
10  ( 1 ) "*"           "*"              " "         "*"            "*"      
11  ( 1 ) "*"           "*"              "*"         "*"            "*"      
          free.sulfur.dioxide total.sulfur.dioxide density pH  sulphates alcohol
1  ( 1 )  " "                 " "                  " "     " " " "       "*"    
2  ( 1 )  " "                 " "                  " "     " " " "       "*"    
3  ( 1 )  " "                 " "                  " "     " " " "       "*"    
4  ( 1 )  "*"                 " "                  " "     " " " "       "*"    
5  ( 1 )  " "                 " "                  "*"     "*" " "       "*"    
6  ( 1 )  " "                 " "                  "*"     "*" "*"       "*"    
7  ( 1 )  "*"                 " "                  "*"     "*" "*"       "*"    
8  ( 1 )  "*"                 " "                  "*"     "*" "*"       "*"    
9  ( 1 )  "*"                 "*"                  "*"     "*" "*"       "*"    
10  ( 1 ) "*"                 "*"                  "*"     "*" "*"       "*"    
11  ( 1 ) "*"                 "*"                  "*"     "*" "*"       "*"    
 [1] 0.1895598 0.2399208 0.2580716 0.2633925 0.2703282 0.2757705 0.2790891 0.2805767
 [9] 0.2805130 0.2803931 0.2802536
```



```
          fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
1  ( 1 )  " "           " "              " "         " "            " "      
2  ( 1 )  " "           "*"              " "         " "            " "      
3  ( 1 )  " "           "*"              " "         " "            " "      
4  ( 1 )  " "           "*"              " "         " "            " "      
5  ( 1 )  " "           "*"              " "         " "            "*"      
6  ( 1 )  " "           "*"              " "         " "            "*"      
7  ( 1 )  " "           "*"              " "         " "            "*"      
8  ( 1 )  " "           "*"              "*"         " "            "*"      
9  ( 1 )  " "           "*"              "*"         "*"            "*"      
10  ( 1 ) "*"           "*"              "*"         "*"            "*"      
11  ( 1 ) "*"           "*"              "*"         "*"            "*"      
          free.sulfur.dioxide total.sulfur.dioxide density pH  sulphates alcohol
1  ( 1 )  " "                 " "                  " "     " " " "       "*"    
2  ( 1 )  " "                 " "                  " "     " " " "       "*"    
3  ( 1 )  " "                 " "                  " "     " " "*"       "*"    
4  ( 1 )  " "                 "*"                  " "     " " "*"       "*"    
5  ( 1 )  " "                 "*"                  " "     " " "*"       "*"    
6  ( 1 )  " "                 "*"                  " "     "*" "*"       "*"    
7  ( 1 )  "*"                 "*"                  " "     "*" "*"       "*"    
8  ( 1 )  "*"                 "*"                  " "     "*" "*"       "*"    
9  ( 1 )  "*"                 "*"                  " "     "*" "*"       "*"    
10  ( 1 ) "*"                 "*"                  " "     "*" "*"       "*"    
11  ( 1 ) "*"                 "*"                  "*"     "*" "*"       "*"    
 [1] 0.2262502 0.3161465 0.3346482 0.3421357 0.3494588 0.3547509 0.3566527 0.3567060
 [9] 0.3565489 0.3562479 0.3561195
```

As seen above, we have a list of BIC from the different subsets, and by looking at the lowest BIC values, we can determine the set of predictors to incorporate into our future model analysis. We utilized best subset selection specifically in linear models, such as logistic regression, where an exhaustive search was computationally feasible. However, due to the computational intensity of this approach, we could not use it for models involving many predictor combinations or non-linear features, such as Random Forest or Generalized Additive Models (GAM), which demanded greater flexibility in feature handling.



K-Fold Cross-Validation proved particularly useful for identifying optimal features in models like Random Forest, K-Nearest Neighbors (KNN), and GAMs, where an exhaustive search was impractical. It ensured that our models used the most relevant features while reducing the risk of overfitting. For polynomial regression, we applied Lasso (L1 regularization), which prioritized the most important predictors by penalizing less relevant ones. This allowed us to minimize overfitting by reducing the feature set to the predictors most impactful to model accuracy.

3.3. Metric Selection

For most of our models, we calculated the Mean Squared Error (MSE) and Adjusted R-squared values. MSE quantifies the average squared difference between predicted and actual wine quality scores, with lower values indicating better model performance. Adjusted R-squared represents the proportion of variance in wine quality explained by the physicochemical features, adjusted for the number of predictors.

4. Results

This section presents the implementation and performance evaluation metrics for the different models explored to predict wine quality based on physicochemical features.

4.1. Random Forest

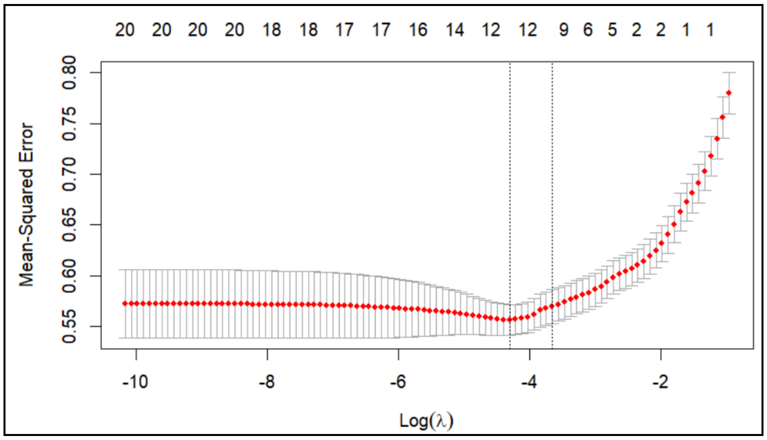
The results from Random Forest Modeling are shown below in the performance metrics tables, in which case we wanted to simulate the potential decision trees to make predictions and aim to optimize the combination of all predictors. Since LOOCV would be too computationally intensive, we decided to employ the K-Fold method to reduce the complexity. Again, since we ran the best subset selection before, we only trained the model based on the result from the best subset, which significantly increased our efficiency when training the model. We calculated the Adjusted R squared, MSE, and AIC as shown above. However, due to the random forest model's nature, it would not be ideal to look at AIC as it does not take into account maximum likelihood.

4.2. K-Nearest Neighbors

The results from KNN modeling are shown below in the performance metrics tables, in which we loop over different possible numbers of neighbors to find the best K that produces the smallest MSE. Again, it didn't take us much time to train the model since we already have the set of predictors from the best subset selection, which makes sure our model is more accurate without compromising the complex dimension of the dataset. We found the best k value to be 9 for white wine and 15 for red wine, with its respective Adjusted R, squared, MSE.

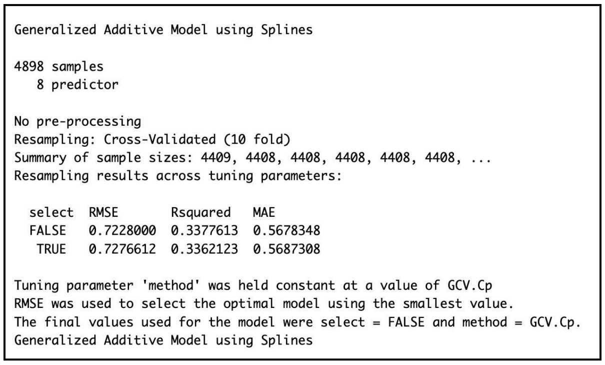
4.3. Regularized Polynomial Regression

For Polynomial Regression, we expanded the feature space through polynomial transformations. Specifically, we scaled up predictors to polynomial terms up to the third degree, effectively capturing potential non-linear relationships between predictors and wine quality. Next, to avoid overfitting, we used Lasso Regression. Here, we extracted the best lambda using K-Folds (10-Folds) and the lowest Test MSE.



4.4. Generalized Additive Model

We chose to use a Generalized Additive Model to explore the possibility that each feature followed varying underlying models, and should thus be approached using an additive technique. This allows us to apply non-linear relationships to the data with the additional flexibility of each predictor's contributions being considered separately. The smoothing parameter here is automatically tuned using the gam(), which uses a technique known as backfitting. Therefore, there was no need for manual tuning of hyperparameters in this case. We see that the model outputs a 10-fold cross-validation RMSE of 0.7267, an R-squared of 0.3355, and MAE of 0.5697 using the most optimal model where select = FALSE. This hyperparameter was automatically determined and signifies that the most optimal of the 2 possible GAM models apply additional penalties on the model curve to spaces where the effect of splining is null. When applied separately using the previously mentioned best subset for the red wines data, we obtain an RMSE of 0.6397, an R-squared of 0.3754, and MAE of 0.4966, which are not significant improvements from the white wine's GAM model.



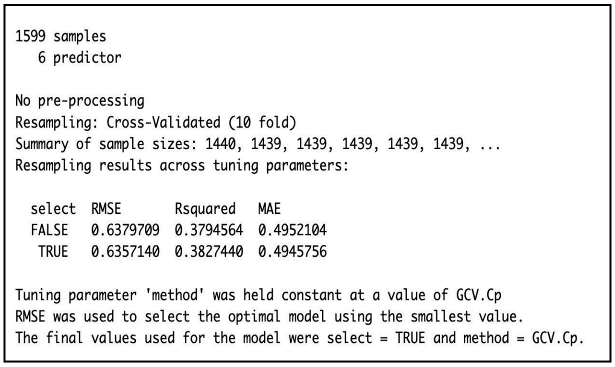
```
Generalized Additive Model using Splines

4898 samples
   8 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 4409, 4408, 4408, 4408, 4408, 4408, ...
Resampling results across tuning parameters:

  select  RMSE       Rsquared   MAE
  FALSE   0.7228000  0.3377613  0.5678348
   TRUE   0.7276612  0.3362123  0.5687308

Tuning parameter 'method' was held constant at a value of GCV.Cp
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were select = FALSE and method = GCV.Cp.
Generalized Additive Model using Splines
```



```
1599 samples
   6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 1440, 1439, 1439, 1439, 1439, 1439, ...
Resampling results across tuning parameters:

  select  RMSE       Rsquared   MAE
  FALSE   0.6379709  0.3794564  0.4952104
   TRUE   0.6357140  0.3827440  0.4945756

Tuning parameter 'method' was held constant at a value of GCV.Cp
RMSE was used to select the optimal model using the smallest value.
The final values used for the model were select = TRUE and method = GCV.Cp.
```

4.5. Logistic Regression

In this logistical model, we attempted to replicate similar logistic regression models we learned in class to predict the quality. In addition, in order to apply the model well, we needed to binarize the quality variable to fit a logistic regression model into High and Low. We decided to use LOOCV as the standard cross-validation measure and binarized the quality variable to fit the logistic regression. We decided that it would be best to fit the model directly rather than add a layer of variable selection. Best subset selection and stepwise selection performed worse and had been too computationally intensive. Thus, to fit a simple logistic regression model and achieve the most optimal outcome metrics, we programmed the metrics that are most relevant to a logistic regression model, including test MSE, AUC, Accuracy, Sensitivity, and Specificity, which encompass AUC, ROC, and a Confusion Matrix. The same process is applied to red wine.

Performance Metrics for White Wine Quality Prediction Models

	Random Forest	**Logistic Regression**	**KNN**	**Regularized Polynomial Regression**	**Generalized Additive Model**
MSE	0.0671829	0.1350475	0.7649898	0.5871471	0.528

Adj R^2	0.5408396	N/A	0.253898	0.278164	0.3378
Other Metrics	AIC: -13204.25	Accuracy: 0.8023683 AUC: 0.7938	N/A	N/A	MAE: 0.5678

Performance Metrics for Red Wine Quality Prediction Models

	Random Forest	**Logistic Regression**	**KNN**	**Regularized Polynomial Regression**	**Generalized Additive Model**
MSE	0.0698625	0.08422009	0.7356221	0.6482802	0.4041
Adj R^2	0.5156118	N/A	0.1702452	0.2804342	0.3827
Other Metrics	AIC: -4233.301	Accuracy: 0.8843027 AUC: 0.8822	N/A	N/A	MAE: 0.4946

5. Interpretation and Discussion:

To follow the flow of our report, we started with an interpretation of the dataset, including its key variables and relevance to our modeling. To further analyze the predictors and variables of the dataset, we performed an Exploratory Data Analysis wherein we compiled the most relevant figures and plots to provide unique insights on the dataset. From correlations to residuals, we were then able to start modeling. However, to take our modeling and analysis a step further, we wanted to perform a comparative study between different models and features adapted in class. Each one of us applied complex models on the dataset to achieve the most optimal predictions. However, for effective comparisons, we standardized the way in which we approached the models. First, we agreed to split the dataset into white and red wine, as provided by the UCI repository. We thoroughly researched the previous papers that utilized the two different datasets as inputs, so we wrote a model for white and red wine datasets each to add an extra layer of comparison. Next, we hoped to utilize the most optimal cross validation approach so we performed LOOCV. Since the dataset wasn't too large, we were able to run LOOCV. In some models that we performed, such as the GAM, we found that LOOCV in fact ran for several hours, so we found that K-Fold would work better for that particular GAM. The following step was variable selection. We agreed that best subset selection would be the best choice with the dataset. Best subset selection was in fact the best decision for most of the models, except logistic regression, where variable selection hadn't been necessary. To standardize our comparative analytics, our outcome metrics were test MSEs adjusted R squared (where appropriate). In some models, we found that there were other more relevant metrics that would provide us a better insight on the model, such as ROC, AUC, AIC, etc. Overall, we leveraged many lessons from class, MSEs, adjusted R squared, AUCs, AIC and BICs, cubic splines, Lasso, Logistic Regression, KNN, Random Forest, Generative Additive Models, and further EDA methods, to produce a consolidated report thoroughly and comparatively analyzing different models and different techniques for most optimal metrics.

We uploaded all the code files to a GitHub Repository:
https://github.com/java1202/STSCI-4740-Final-Project