Regression Analysis on the Wine Classification and Quality

This research both uses classification and multi-linear regression method to determine the Vinho Verde wine quality, to provide other wine firms a way to evaluate their product quality.

YICHAO, IVAN, DAI

Wenzhou-Kean University

In this project, the researchers aim to build several models to classify the wine type and predict white wine quality. All the datasets were collected through the UCI public data source. The project first constructed the classification model based on the Classification and Regression Tree (CART) algorithm; then, the researcher will update the decision tree by choosing the best complexity parameter. The classification model performance is excellent; the overall classification accuracy is higher than 98%. Moreover, the research also uses 5 different methods, including Ordinary Least Square (OLS), Akaike information criterion (AIC), Ridge, least absolute shrinkage and selection operator (LASSO), and Elastic net (EN), which aims to eliminate the multicollinearity and choose the best predictive model based on the root square to mean error and to adjust R square. Although the linear relationship is significant, the model can only explain 31% variability of the white wine quality. Further researches can be conducted to select more correlated variables and improve model performance

CCS CONCEPTS • Linear regression • Logistic regression • Decision tree

Additional Keywords and Phrases: Classification, wine quality, dimension reduction, model selectin, shrinkage.

1 INTRODUCTION

Alcoholic drinks have become indispensable in modern times, but there are also different classification levels of other wines. The producer needs to classify the various quality of wine to set the price. Moreover, customers also require the way how they rank. The process of making a glass of wine is very complicated. The winemaker must strictly maintain the alcohol content, density, PH value, sulfur dioxide, and other ingredients in the process. Variations in the range of the components are likely to make a difference in the wine's quality. In this project, we will use the Vinho Verde wine data, including the white wine and red wine, to do the classification analysis and regression analysis. The corresponding data about the ingredients used in making a bottle of wine have been collected from Paulo Cortez and other producers. We hope the research results can help other wine producers evaluate their product qualifications and adjust their production process.

2 METHODOLOGY AND PROCESS

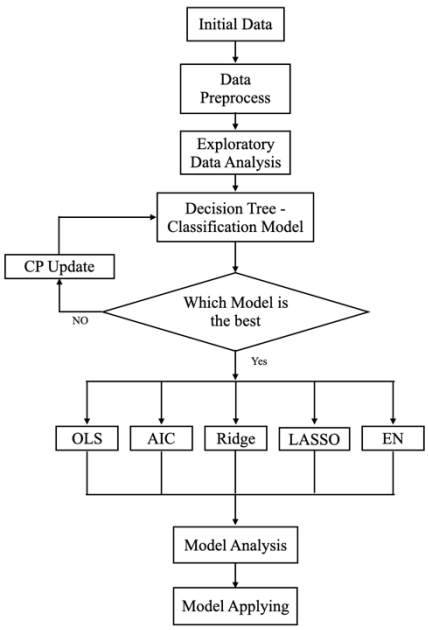

Figure 1: Flow chart

The whole research process have been showing at the following flow chart (*figure 1*). The main process in the regression analysis on wine classification and quality are shown as follow:

1) Collect initial data: Initial datasets required for this project have already collected through the producers of Vinho Verde.

2) Data Preprocess: There are two datasets recording white wine and red wine. We need to preprocess these data to combine these two dataset and add one categorical variable to classify two wine types.

3) Exploratory data analysis: Before start to build the model, the project we do some descriptive statistic to explore some potential relationship among the variables, including scatter plot, boxplot, and even cluster plot.

4) Classification model: the project will build the classification model and plot the classification & regression tree through programming. Also, the project the model accuracy by using the test set. If the accuracy is smaller than 90%, we will use the principal component analysis (PCA) method to update the model, until the model accuracy is higher than the threshold.

5) Multi-regression model for red and white wine. The project we separately build the predictive regression model for both type of wine quality (response variable). The researchers will also check the root mean square error and R square to evaluate the model.

6) Model Applying: The project are going to take some practical examples, to discuss how the models built in the research can be used in reality.

3 PREPROCESSED DATA

After preprocessing data, the final data example are shown in the following (*Table 1*). The whole data sets have 6497 observation without any missing value. There are totally 13 variables in the data sets, which measure the main component of wine. The full datasets is in the fold named as 'wine.csv'.

fixed.acidity	volatile.acidity	citric.acid	…	alcohol	quality	type
7.6	0.180	0.36	…	10.30	5	White
6.4	0.260	0.21	…	9.50	5	White
…	…	…	…	…	…	…
8.9	0.610	0.49	…	9.30	5	Red

Table 1: Example Data

The following table show a brief introduction to the data description.

	Data Type	**Description**
fixed.acidity	Numeric	The fixed acids found in wines are tartaric, malic, citric, and succinic.
Volatile.acidity	Numeric	The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
citric.acid	Numeric	Small quantities, citric acid can add 'freshness' and flavor to wines
residual.sugar	Numeric	The amount of sugar remaining after fermentation stops
chlorides	Numeric	The amount of salt in the wine
free.sulfur.dioxide	Numeric	The free form of SO2 exists in equilibrium between molecular SO2 and bisulfite ion; it prevents microbial growth and the oxidation of wine
total.sulfur.dioxide	Numeric	Amount of free and bound forms of S02;
density	Numeric	The density of water, depending on the percent alcohol and sugar content
pH	Numeric	Describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic)
sulphates	Numeric	A wine additive which acts as an antimicrobial and antioxidant
alcohol	Numeric	The percent alcohol content of the wine
quality	Numeric	Score between 0 and 10

Table 2: Data description

4 EXPLORATORY DATA ANALYSIS (EDA)

There are two parts in Exploratory data analysis. First, it will explore relationship of different variables against the wine type, which prepare for the classification model. Second, we will explore the relationship among the numeric variables, to see the correlation and whether these is a multi-colinear method if we build up the regression model.

4.1 EDA in wine classification

The following boxplot and density distribution (*figure 2*) showing the residual sugar difference of wine components variables against the wine type. Residual sugar is the amount of sugar remaining after fermentation stops. We can see white wine has a wider spread the red wine has. Generally, white wine have a higher level of residual sugar than red wine has, the distribution of residual sugar in red wine are extremely right skewed but highly density around some value. Over 75% of the red wine are stick together around 2, which is rational to have a lot outliers on the right tail of the distribution. We can see the degree of the skewness of these distribution through the following tables. We can see both of the distribution have high value of the skewness. Moreover, because the residual of red wine are more density, the outlier in red wine results in a higher skewness.

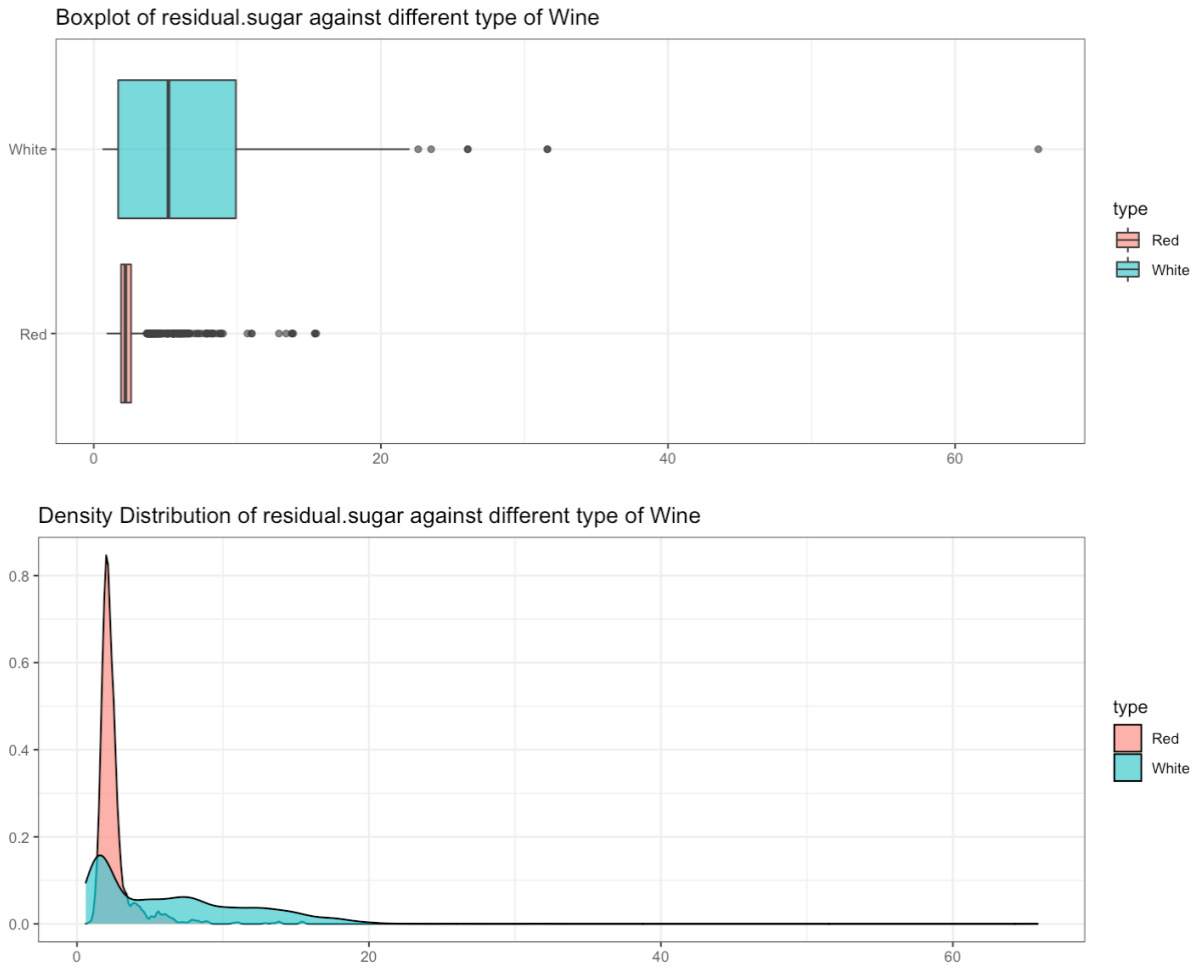

Figure 2: Boxplot and Density distribution of residual sugar against wine type

	Minimum	Maximum	Median	Std	Skewness
Red Wine	0.9	15.5	2.2	1.409928	4.53214
White Wine	0.6	65.8	5.2	5.072058	1.076434

Table 3: Descriptive statistics on alcohol

Take a close look at the difference of the alcohol level in its boxplot and density distribution (*figure 3*), we can observe that the median alcohol level of red wine are a little bit lower, and its distribution is a little more concentrated between 9 and 10, which means it has less variability than white wine has. The following table (*table 4*), also show the specific descriptive statistics about the alcohol in different type of wine. However, we can observe the distribution of the alcohol among different type of wine are similar, we use the Student-t test to find out the 95% confidence interval of the difference of the average alcohol in different type of wine. We can see the output of the confidence interval in *table 5* Because the confidence interval is (-0.1539 , -0.0287), as a result, we have 95% confidence that the difference is in this range, as a result, 0 does not included in this interval. Moreover, by increasing the confidence level to 99%, the confidence interval still does not include 0. Consequently, we can conclude that there is a significant difference in alcohol between white wine and red wine.

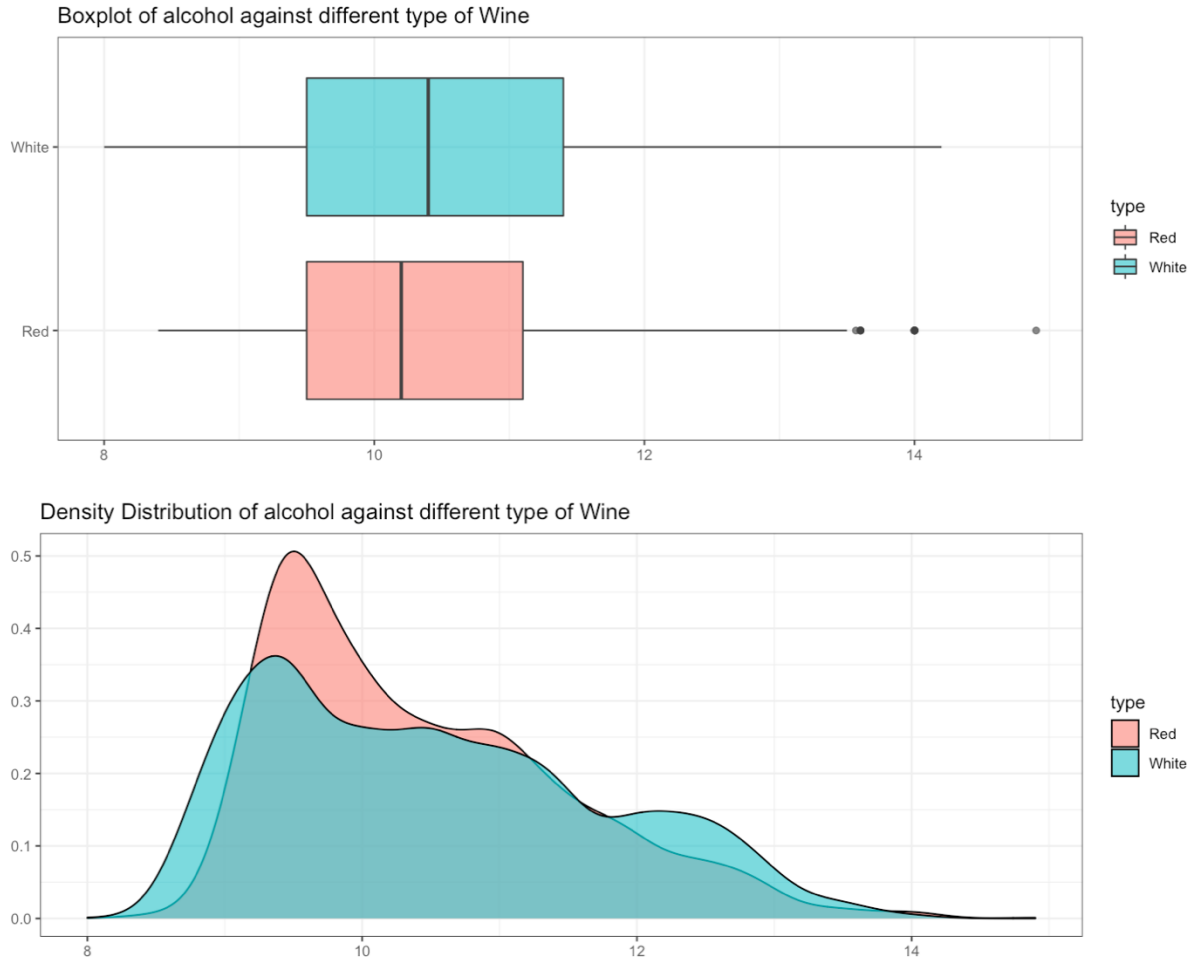

Figure 3: Boxplot and Density distribution of alcohol against different wine type

	Minimum	Maximum	Median	Mean	Std
Red Wine	8.40	14.90	10.20	10.42	1.07
White Wine	8.00	14.20	10.40	10.51	1.23

Table 4: Descriptive statistics on alcohol

	Interval
95% CI (Red - White):	(-0.1539 , -0.0287)
99% CI (Red - White):	(-0.1736 , -0.0089)

Table 5: Confidence interval test on alcohol of two different type of wine

4.2 EDA in wine components correlation

The following correlation matrix (*figure 4*) shows the correlation of each two variables of wine components, we can see most of explanatory variables have a relatively weak correlation, however, the density and alcohol have a relative strong correlation. In the OLS multi-regression model, highly correlated variables would add increase the standard error of the coefficients and increase the bias of the models, which result in multicollinearity problems. The easy way to eliminate this effect is by removing one of the those correlated explanatory variables. However, removing which one is the most important thing to decide.

Figure 4: Correlation Matrix

Correlations focus on bivariate relationships to assess relevancy and redundancy. A predictor deemed to be irrelevant or redundant based on bivariate correlations must be dropped. On the other hand, a predictor identified as being relevant and non-redundant based on bivariate correlations may still not be included in the model. This is because a predictor may still be irrelevant or redundant in a multivariate sense. To examine this possibility, we can examine statistical significance of regression coefficients and variance inflating factor (VIF).

$$VIF_i = \frac{1}{1 - R_i^2}, \quad where \ R_i^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

Where R_i^2 is the coefficient of determination of the regression equation.

Threat of collinearity can also come from linear relationships between sets of variables. One way to assess the threat of multicollinearity in a linear regression is to compute the Variance Inflating Factor (VIF). 1<VIF<Inf. VIF>10 indicates serious multicollinearity while 5<VIF<10 may warrant examination. The following table show VIF of different explanatory variables:

	VIF		**VIF**		**VIF**
fixed.acidity	2.61	chlorides	1.23	pH	2.17
volatile.acidity	1.15	free.sulfur.dioxide	1.83	sulphates	1.14
citric.acid	1.16	total.sulfur.dioxide	2.25	alcohol	6.86
residual.sugar	11.95	density	25.70		

Table 6: VIF

As a result, we can see that residual.sugar, density, and alcohol have relative large VIF, which means regression model may remove some of the these variables to avoid multi-colinear problems or may use shrinkage (Lasso regression) to eliminate the multi-colinear problems.

5 MODEL BUILDING AND ANALYSIS

In this section, the project will first build the classification model based on the CART algorithm. Then the white-wine multi-regression model and red-wine multi-regression mode will be built separately.

5.1 Model Flow - Classification

Classification and Regression Tree (CART) algorithm, which explains how a targeted variable's value can be predicted by other variables' values. The model building process will follow the flow chart (*figure 5*).
The red wine will be set as 0.
The white wine will be set as 1.
This valuing process satisfies the model requirement. If the accuracy of predicted result of the models on test set is smaller than 0.9, we will update the model by using principal component analysis, to improve the classification accuracy.

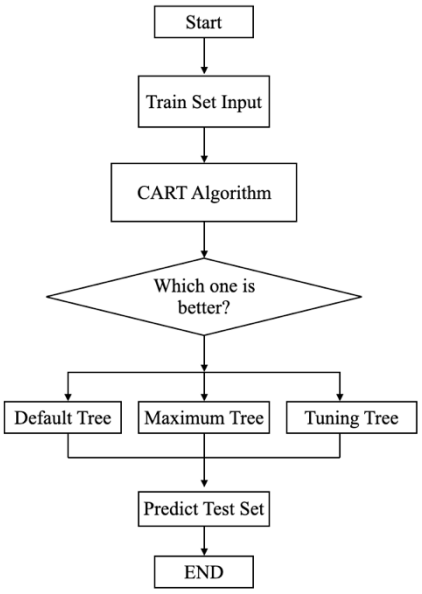

Figure 7: Classification Model Flow

5.2 Model Analysis – Classification

5.2.1 Default Tree

By using CART algorithm, we get the final model, showing in the following decision tree (*figure 6*). The function use the Gini impurity measure to split the note with a default complexity parameter (CP) equal to 0.001. The main classifier in the decision tree is total.sulfur.dioxide, chlorides, and volatile.acidity. The overall model performance is relative good. We apply the classification model to the test set, the following table show the performance of the classification model (table). The model accuracy is about 0.98, which is a relatively high-accuracy model. CP determine how complexity the classification model could be. We want to choose the CART classification model with the best CP.

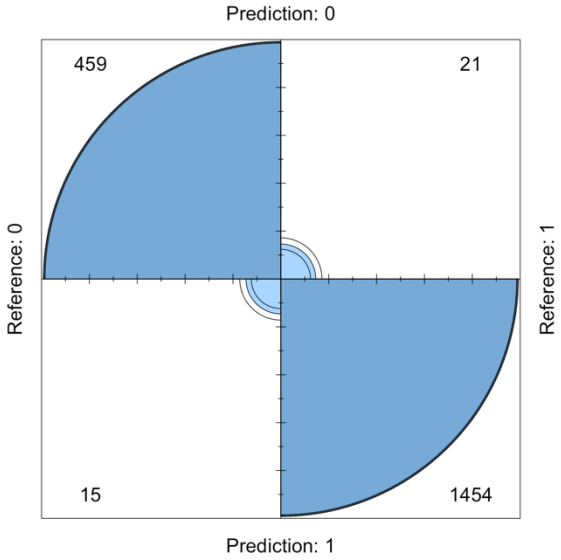

Figure 5: Confusion Matrix of Default Tree

Items	Value
Accuracy	0.9815
95% CI	(0.974, 0.987)
Sensitivity	0.9684
Specificity	0.9858
Balanced Accuracy	0.9771
Pos Pred Value	0.9563
Neg Pred Value	0.9898
Prevalence	0.2432
Detection Rate	0.2355
Detection Prevalence	0.2463

Table 8: Default Tree Performance

5.2.2 Maximum Tree

By setting the CP equal to 0, we can get the most complex classification model. However, the most complex model only have an model accuracy equal to 0.9805. The performance of the final model is showing as follow:

Items	Value
Accuracy	0.9805

95% CI	(0.9733, 0.9862)
Sensitivity	0.9742
Specificity	0.9825
Balanced Accuracy	0.9784
Pos Pred Value	0.9458
Neg Pred Value	0.9918
Prevalence	0.2391
Detection Rate	0.2329
Detection Prevalence	0.2463

Table 9: Maximum Tree Performance

5.2.3 Tuning Tree

In this section, we want to find the best classification model by setting different complexity parameter. Following plot show the corresponding model accuracy with different complexity parameter. The plot shows that when CP is equal to 0.0065, the model have a highest accuracy.

The following plot (*Figure 8)* show the decision tree with complexity parameter equal to 0.0065. We can see that the model is more complex than the previous, with a higher accuracy.

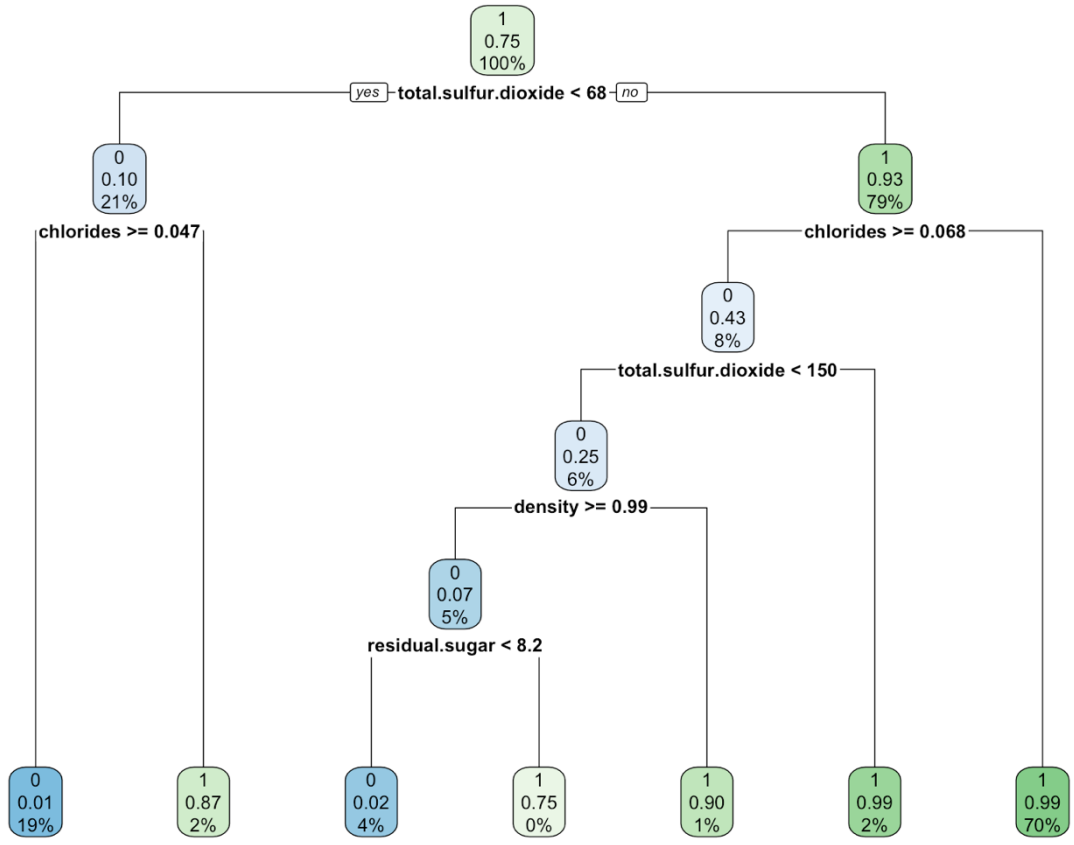

Figure 6: Tuning Decision Tree

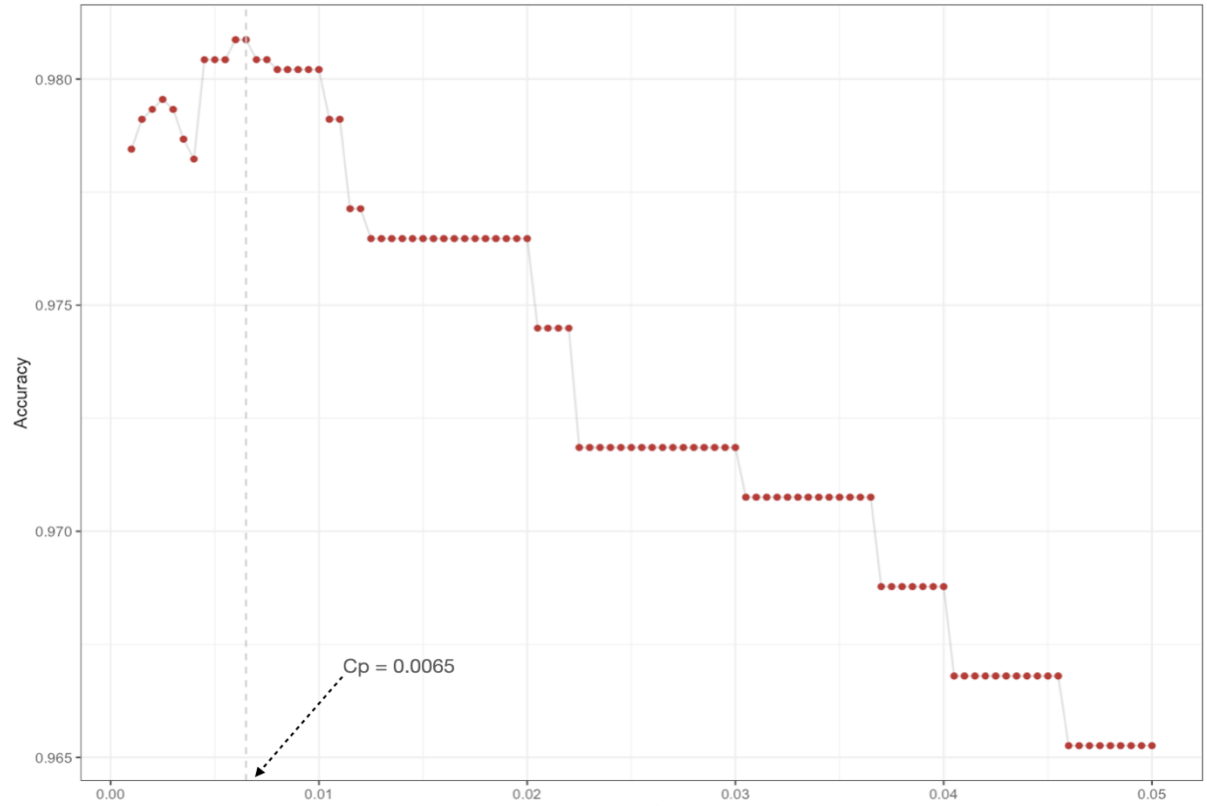

Figure 7: Tuning Tree Selection Process

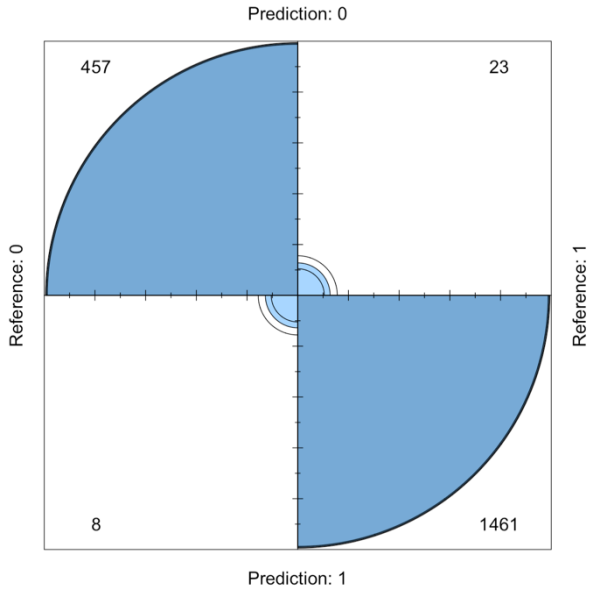

Figure 8: Confusion Matrix of Tuning Tree

Items	Value
Accuracy	0.9841
95% CI	(0.978, 0.989)
Sensitivity	0.9828
Specificity	0.9845
Balanced Accuracy	0.9836
Pos Pred Value	0.9521
Neg Pred Value	0.9946
Prevalence	0.2386
Detection Rate	0.2345
Detection Prevalence	0.2463

Table 6: Model Performance

Moreover, we can use the Receiver Operating Characteristics (ROC) to check the model's performance. Following plot show the ROC of the classification model. The area under the curve (AUC) could tell us how much model is capable of distinguishing between classes. The AUC for this classification model is calculated as 0.982. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes, consequently; the classification model for this project have a strong ability to distinguish the different type of wine. As a result, the producers can use this model to classify their wine, and then use the following multi-regression model to predict the wine quality.

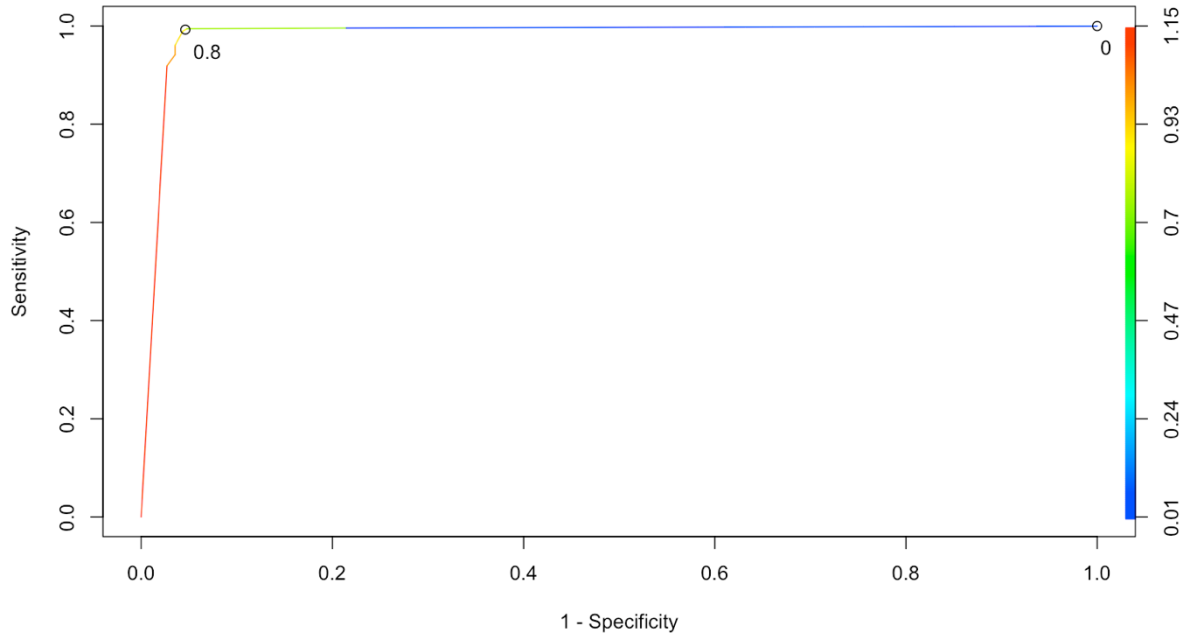

Figure 9: ROC and AUC

5.3 Model Flow – Predict White Wine Quality

In this section, the project will start to work out the predictive model on the white wine quality. In the previous section, we have found that there are some highly correlated variables in the dataset. As a result, the project will apply 5 different method including Ordinary Least Square (OLS), Akaike information criterion (AIC), Ridge, least absolute shrinkage and selection operator (LASSO), and Elastic net (EN), which can eliminate the multicollinearity problem and some other bias to build final model.

1) In the first method, we will use OLS method to use all the explanatory variables to predict the wine quality. The OLS multi-regression should first follow the 3 main assumptions: normality of errors, correlation and multicollinearity, and homoscedasticity. Non-normality residual can influence the capacity of the model forecasting. Also, multicollinearity problems may inflate the variance of the model and undermine the integrity of regression analysis. Homoscedasticity of errors make sure that reliabilities of the standard error of the OLS model. As a result, we will diagnostic the final to check the assumptions

2) In the second method we will use the AIC method to start with a full model, and select the best model based on the Akaike information criterion (AIC), which is calculated as follow:

$$AIC = 2k - \ln(\hat{L})$$

Where:

k: number of estimated parameters in the model.

$\hat{L}$: maximum value of the likelihood function for the model.

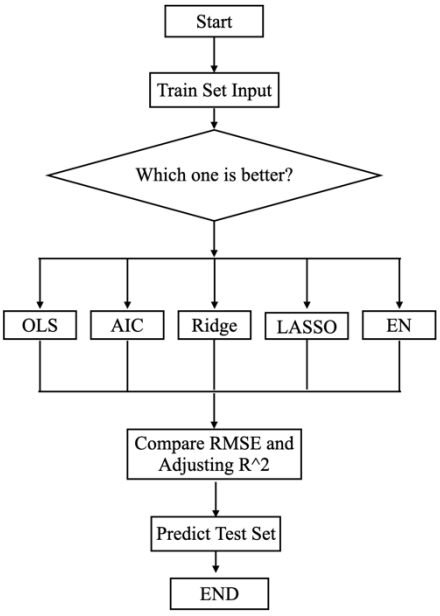

Figure 10: Flow Chart

The model selection process will delete one variable at one time to see whether the model can achieve a lower AIC, model with a lower AIC would be better. The final model will be selected if deleting any one of the variables cannot achieve a lower AIC.

3) In the third method, we will use the Ridge regression to eliminate the effect of multicollinearity, which is one of regression method to use shrinkage. Ridge method will do the adjustment to the corresponding coefficients of each variables by using following formula:

$$\beta_{ridge} = \min\left[\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^2\right]$$

The function of lambda is similar with the lambda in LASSO method. If lambda is equal to 0, the equation is basically OLS equation, if lambda is greater than 0, the equation then put some constraint to the coefficients. However, ridge method may not good for features reduction.

4) In the fourth method, we will use the LASSO regression, which is another way to eliminate the effect of multicollinearity and reduce the features dimensions by using the shrinkage. LASSO method will do the adjustment to the corresponding coefficients of each variables by using following formula:

$$\beta_{lasso} = \min\left[\sum_{i=1}^{n}\left(y_i - \sum_{j} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right]$$

When $\lambda = 0$, no parameters are eliminated. The estimate is almost equal to the one found with OLS linear regression. As λ increases, more and more coefficients are set to zero and eliminated (theoretically, when $\lambda = \infty$, all coefficients are eliminated). As λ increases, bias increases. LASSO method selection process will choose the best lambda with its corresponding model as the final model.

5) In the fifth method, we will use the Elastic Net (EN) regression, which is a way to combine Ridge and LASSO method. The method may improve the model performance of the LASSO regression by reducing some bias, however, it cannot completely remove the multicollinearity problems.

At last, we will compare the root mean squared error (RMSE) and adjusting R square to check these model and decide which model is better. At last, we will use the final model to predict on the test set and calculated its RMSE. In the multi-regression model, it is important to compare the adjusting R square rather than the R square. Reducing feature dimensions would definitely reduce the R square, but it may improve the adjusting R square. It is calculated as following formula:

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

Where:

R^2 is the R square of the model.

n is the number of observation of the sample.

k is the number of the independent regressors, which is the explanatory variables in the model.

5.4 Model Analysis - Predict Quality

5.4.1 OLS Model of White Wine Quality Prediction

The initial model is showing below:

$$quality = \beta_0 * fixed.acidity + \beta_1 * volatile.acidity + \cdots + \beta_{11} * alcohol + \varepsilon \sim N(0, \sigma)$$

The β is the coefficient of the explanatory variables. ε is the error of the model. The model assume the error follow the normal distribution with a mean equal to 0.

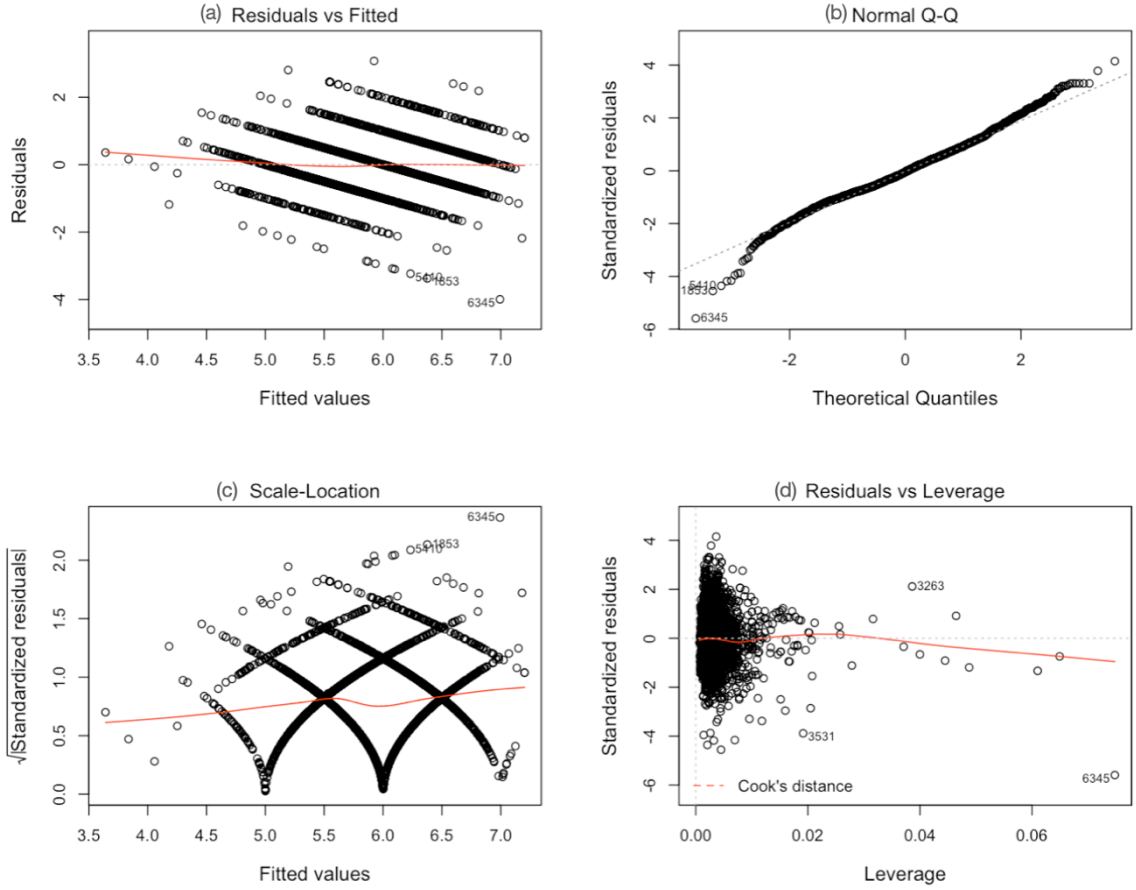

Figure 11: Diagnostics Plot

Fig. 12 illustrates the overall performance and diagnostic plots of the OLS model. In Fig. 12(a), residual and fitted plot is presented to show the pattern of residuals. It is shown that residuals are equally spread around a horizontal line without distinct patterns, suggesting that the model captures the linear relationship between predictor variables and an outcome variable and serves as a good indication where linear relationship was fully explained. It is noted that the residual plots can also give an indication from a 'good' model and a 'bad' model. The good model data are simulated in a way that meets the regression assumptions very well, while the bad model data are not. Fig. 12(b) demonstrates the normal quantile-Quantile (Q-Q) plot to check if residuals are normally distributed. The Q-Q plot is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as Normal distribution when the statistical analysis assumes the dependent variable is Normally distributed. As a result, one can see that residuals follow a straight dashed line to meet the assumption of normality. Fig. 12(c) shows the scale-location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how one can check the assumption of equal variance, i.e., homoscedasticity, and the results show that residuals appear randomly spread with a nearly horizontal line with equally (randomly) spread points being spotted. Therefore it is shown that our model meet the assumption of equal variance (homoscedasticity). In Fig. 12(d), the residuals against leverage plot is presented. This plot helps us to find influential cases (i.e., subjects) if any. Not all outliers are influential in linear regression analysis. Even though data have extreme values, they might not be influential to determine a regression line. That means, the results wouldn't be much different if we either include or exclude them from analysis, they still follow the trend in the majority of cases. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don't get along with the trend in the majority of the cases. Fig. 12(d) is the typical look when there is no influential case, or cases as one can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines.

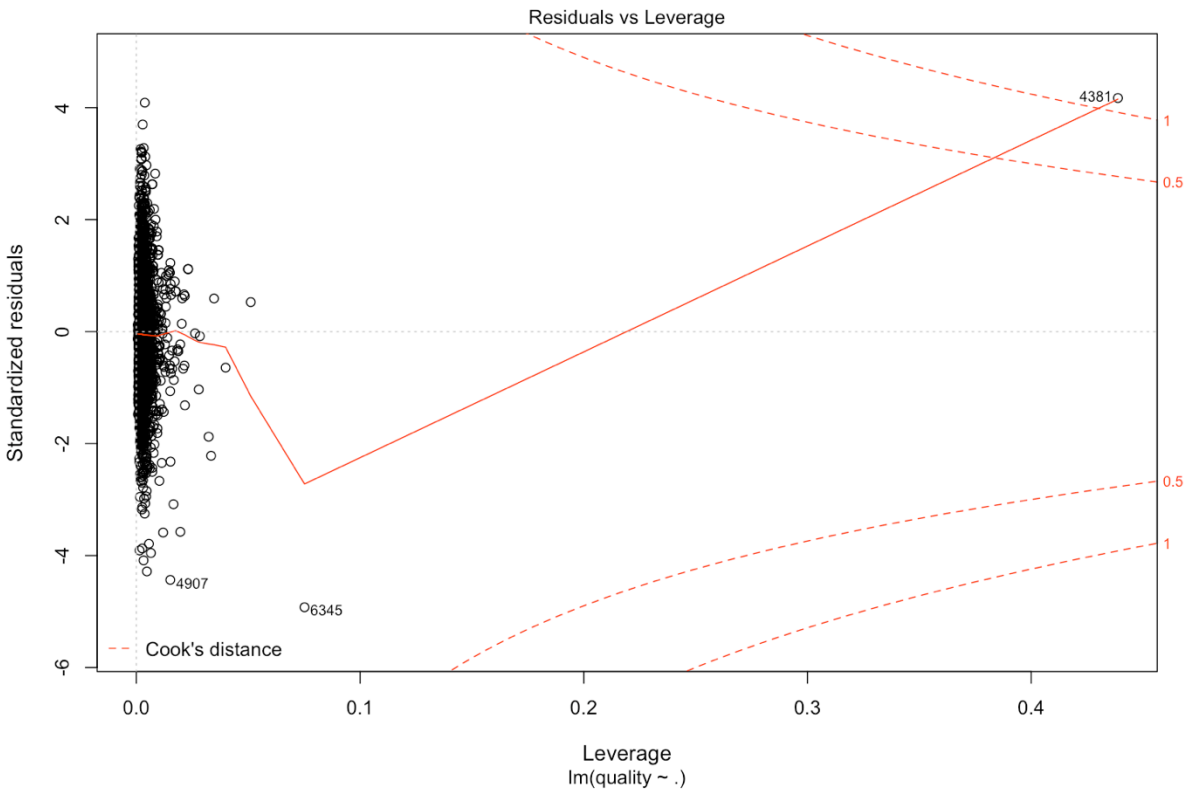

Figure 12: Comparison After removing the outlier

Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook's distance. When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases. Fig. 13 is an enlarged view of Fig 12(d) which shows an extreme value at the upper right corner (observation ID is equal to 4381). It is a case that is far beyond the Cook's distance lines with the other residuals appeared clustered on the left and the plot is scaled to show larger area than the previous plot. The plot identified the influential observation as #4381. If one excludes the 4381th case from the analysis, the adjusted R^2 changes from 0.2973 to 0.3032, show big impact.

After removing the influential point, it is shown that the residuals-leverage plot become normal. The summary in Table 10 shows the results after removing the outlier. It can be seen that the leverage line is almost around 0, which suggests that there is no more extremely influential points in the datasets. In the following method. We will continue use the train datasets without this outlier.

The table below show the summary of the full regression model and the model performance on the train set. The Estimated value is the coefficients of the full model.

	Estimated	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif.**
(Intercept)	2.51E+02	2.680e+01	9.37E+00	< 2e-16	***
fixed.acidity	1.53E-01	2.675e-02	5.73E+00	1.08E-08	***
Volatile.acidity	-1.75E+00	1.316e-01	-1.33E+01	< 2e-16	***
citric.acid	1.13E-02	1.120e-01	1.01E-01	9.20E-01	
residual.sugar	1.15E-01	1.013e-02	1.14E+01	< 2e-16	***
chlorides	-1.10E-01	6.401e-01	-1.72E-01	8.64E-01	
free.sulfur.dioxide	4.59E-03	9.911e-04	4.63E+00	3.75E-06	***
total.sulfur.dioxide	-4.94E-04	4.581e-04	-1.08E+00	2.81E-01	
density	-2.52E+02	2.715e+01	-9.30E+00	< 2e-16	***
pH	9.29E-01	1.312e-01	7.08E+00	1.75E-12	***
sulphates	8.67E-01	1.176e-01	7.38E+00	2.04E-13	***
alcohol	7.85E-02	3.395e-02	2.31E+00	2.09E-02	*

*Table 10: Regression Summary, Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1, Multiple R-squared: 0.3055, Adjusted R-squared: 0.3033, F-statistic: 124.8, p-value: < 2.2e-16*

	Adjusting R Square	**MSE**	**RMSE**
OLS FULL Model	0.303263	0.5493804	0.741202

Table 11: OLS Model Performance

5.4.2 AIC Model of White Wine Quality Prediction

The model will delete one variable at one time, to achieve a better model with a lower AIC. Backward AIC final model through the stepwise-backward method eliminates multicollinearity problems. The selection process is showing as the following table:

Step	**Var removed**	**R square**	**Adj. R square**	**AIC**	**RMSE**
1	citric.acid	0.3055	0.3035	7699.0038	0.7424
2	chlorides	0.3055	0.3037	7697.0304	0.7413

Table 12: AIC Backward Model Selection

The final model achieved by the stepwise backward method is as following table (*Table 12*) with a lowest AIC = 7697.0304. The estimated column show the estimated coefficients of the model. All the explanatory variables are linear significant. Moreover, the whole multi-regression model is in a significant level (F-statistics), and the adjusted R square is about 0.30, which means the model can explained 30% of variance of the white wine quality.

	Estimated	**Std. Error**	**t value**	**Pr(>|t|)**	**Signif.**
(Intercept)	2.59E+02	2.54E+01	1.02E+01	< 2e-16	***
fixed.acidity	1.57E-01	2.61E-02	6.03E+00	1.84e-09	***
Volatile.acidity	-1.78E+00	1.26E-01	-1.42E+01	< 2e-16	***
residual.sugar	1.18E-01	9.72E-03	1.21E+01	< 2e-16	***
free.sulfur.dioxide	3.94E-03	7.86E-04	5.01E+00	5.67e-07	***
density	-2.61E+02	2.58E+01	-1.01E+01	< 2e-16	***
pH	9.42E-01	1.28E-01	7.34E+00	2.69e-13	***
sulphates	8.63E-01	1.17E-01	7.37E+00	2.19e-13	***
alcohol	7.36E-02	3.35E-02	2.20E+00	0.0279	*

*Table 13: Regression Summary, Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1, Multiple R-squared: 0.3053, Adjusted R-squared: 0.3037, F-statistic: 187.8, p-value: < 2.2e-16*

	Adjusting R Square	**MSE**	**RMSE**
AIC Model	0.3036745	0.5495725	0.7413316

Table 14: AIC Model Performance

5.4.3 Ridge Model of White Wine Quality Prediction

The plot on the right of *figure 9* shows that most of R square were explained for quite heavily shrunk coefficients. But at the end, a little bit increase in R square will cause huge growth of some variables. It can be seen as a signal of model overfitting. *Figure 10* shows the model selection process of the Ridge regression, each lambda value will have the corresponding whole path of coefficients. The cross validation process will pick up the best model.

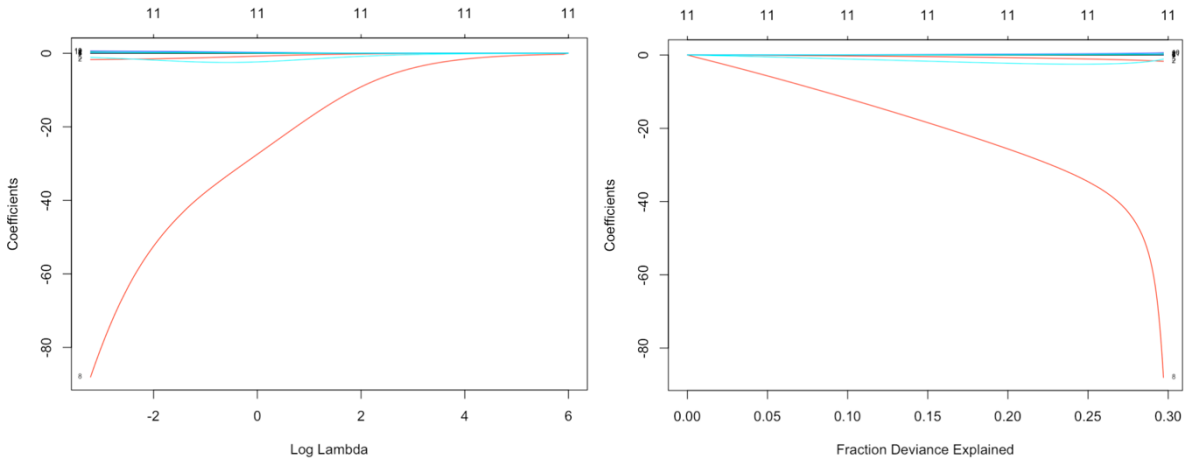

Figure 13: Ridge Model

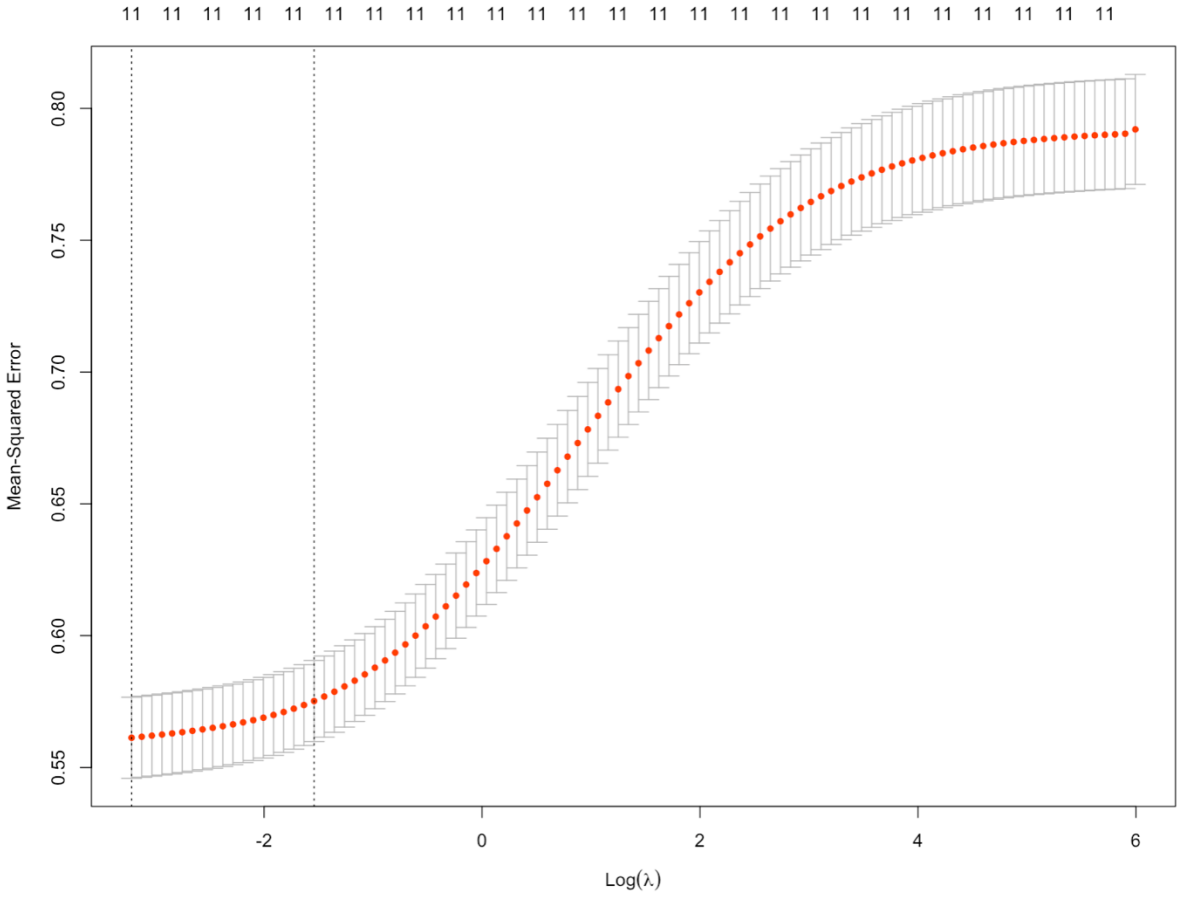

Figure 14: Ridge Model Selection

By looking at the *figure 11,* we can observe the model selection process of Ridge model. The model shown the relationship between the value of logarithm of lambda and the mean squared error (MSE) of the model. As a result, we can directly observe the best model is when the value of logarithm of lambda is near -3 with MSE roughly equal to 0.56. Table 4 show the coefficient of variables in Ridge model.

Variable	**Coef.**
(Intercept)	47.633600763
fixed.acidity	-0.015720818
Volatile.acidity	-1.370916986
citric.acid	0.058229507
residual.sugar	0.024968937

chlorides	-2.137396308
free.sulfur.dioxide	0.004535430
total.sulfur.dioxide	-0.001157541
density	-44.773660648
pH	0.182277294
sulphates	0.472124826
alcohol	0.223093170

Table 15: Ridge Model Coefficients

	Adjusting R Square	**MSE**	**RMSE**
Ridge Model	0.294901	0.555992	0.7456487

Table 16: Ridge Model Performance

5.4.4 LASSO Model of White Wine Quality Prediction

The LASSO model shows the similar *Figure 10* shows the model selection process of the LASSO regression, each lambda value will have the corresponding whole path of coefficients. The cross validation process will pick up the best model.

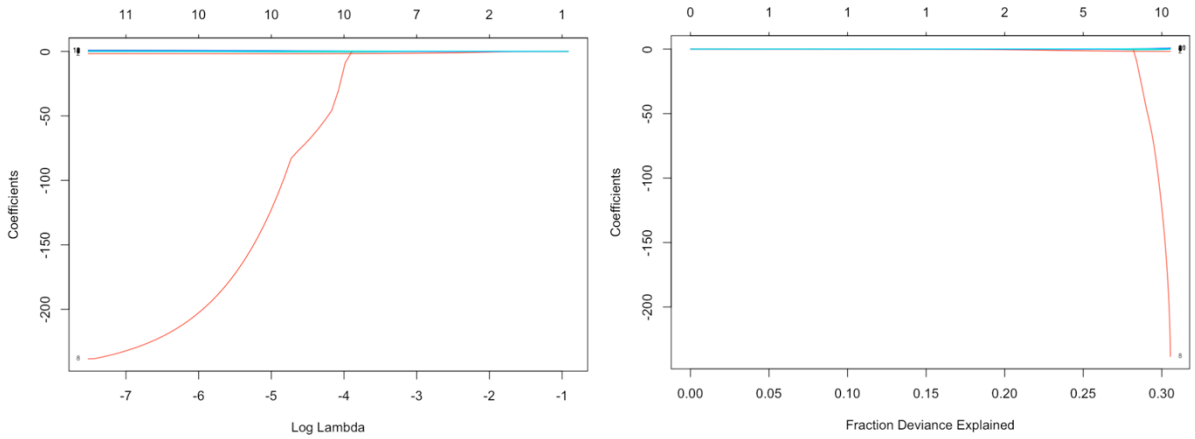

Figure 15: Lasso Model

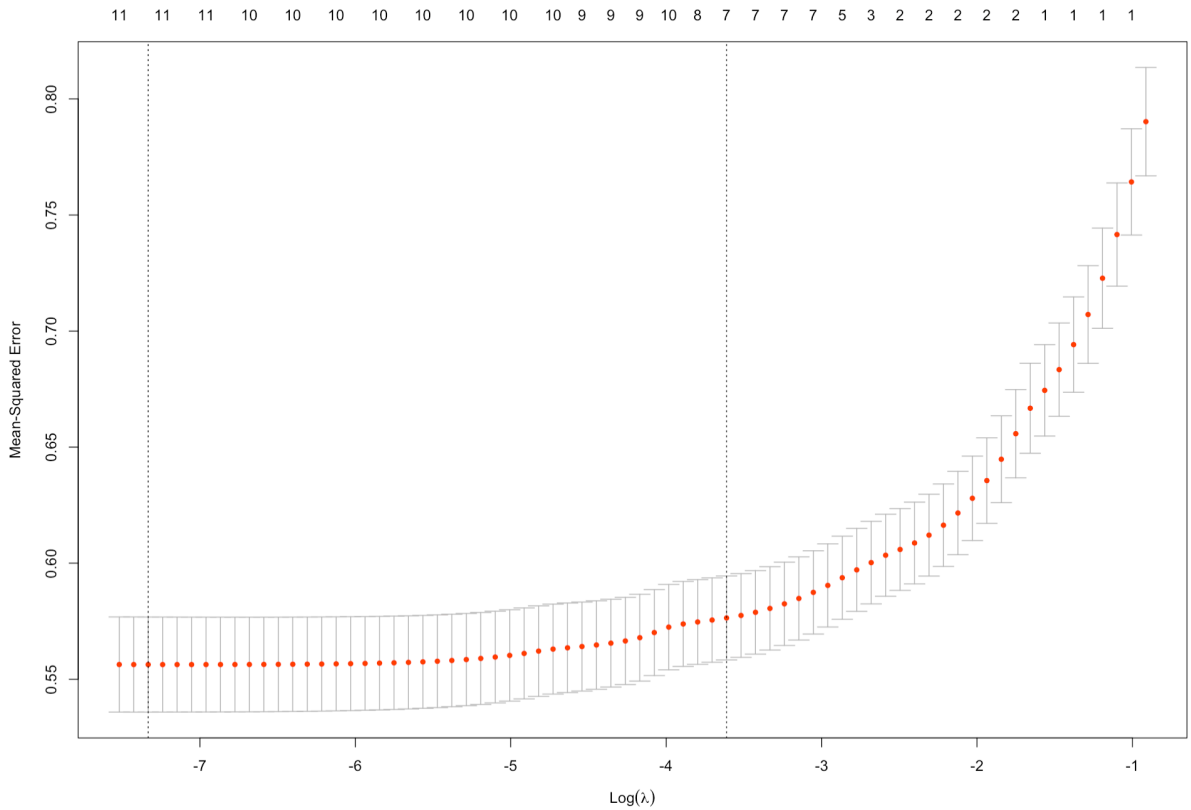

Figure 16: Lasso Model Selection

By looking at the *figure 11,* we can observe the model selection process of LASSO model. The model shown the relationship between the value of logarithm of lambda and the mean squared error (MSE) of the model. As a result, we can directly observe the best model is when the value of logarithm of lambda is near -5 with MSE roughly equal to 0.55. Table 17 shows the coefficient of variables in LASSO model. We can see some variable have been eliminated.

Variable	**Coef.**
(Intercept)	2.603527046
fixed.acidity	-0.022958065
Volatile.acidity	-1.656699060
citric.acid	.
residual.sugar	0.013366010
chlorides	-0.593160659
free.sulfur.dioxide	0.002849128
total.sulfur.dioxide	.
density	.
pH	.
sulphates	0.261196066
alcohol	0.342926332

Table 17: LASSO Model Coefficients

	R Square	**MSE**	**RMSE**
LASSO Model	0.3038571	0.5494391	0.7412416

Table18: LASSO Model Performance

5.4.5 EN Model of White Wine Quality Prediction

EN model is the combination of the LASSO and Ridge model. It depends on the mixed percentage. Following figure shows the different mixed percentage and its corresponding regularization parameter. The best mixed percentage is calculated as 0.1, and the best regularization parameter is 0.00098, showing at the bottom of the figure.

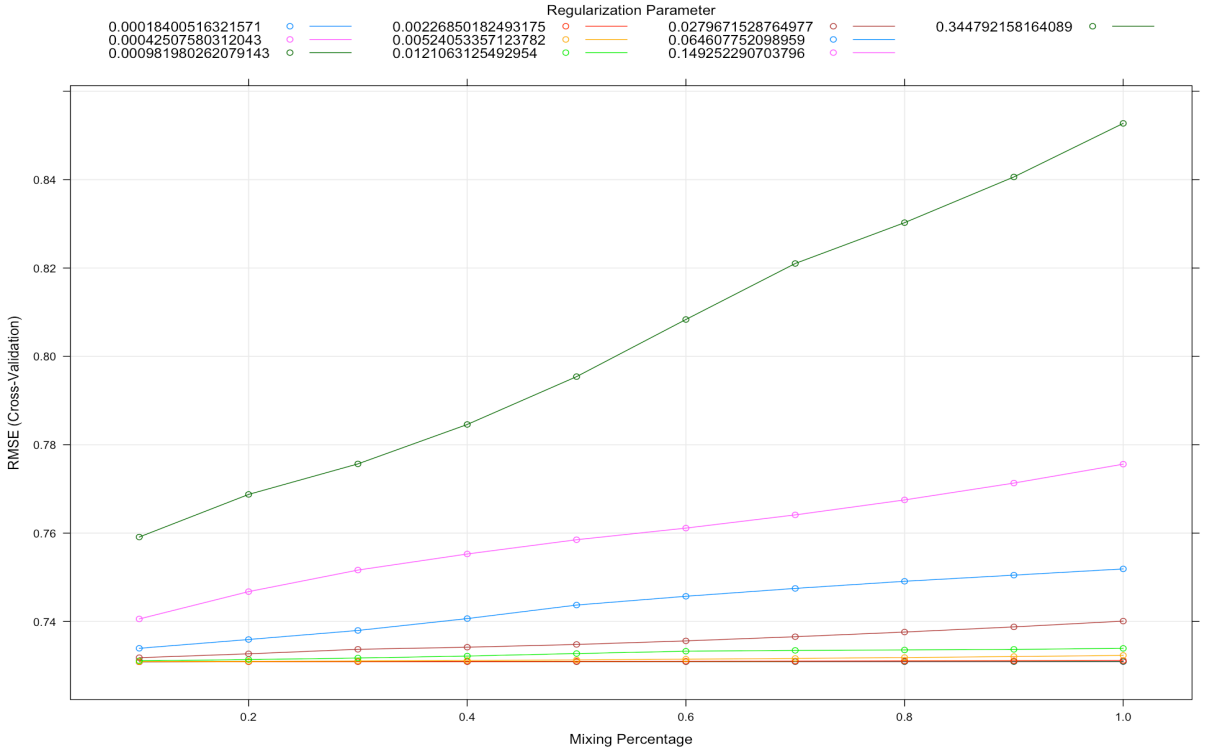

Figure 17: EN Model Selection

The following table show the EN regression model and its performance.

Variable	**Coef.**
(Intercept)	75.259277684
fixed.acidity	0.090872603
Volatile.acidity	-1.281723074
citric.acid	-0.125094770
residual.sugar	0.048605253
chlorides	-0.669037395
free.sulfur.dioxide	0.007021717
total.sulfur.dioxide	-0.002908846
density	-74.531726847
pH	0.430866308
sulphates	0.834843002
alcohol	0.245762405

Table 19: EN Model Coefficients

	R Square	**MSE**	**RMSE**
EN Model	0.306895	0.5299881	0.7280028

Table 20: EN Model Performance

5.4.6 Comparison Among the models

In this section, the project try to compare different models by using the adjusting R square and root mean squared error of each model. The following table show the comparison among these five models' performance on the training set:

	Adjusting R Square	MSE	RMSE
OLS Model	0.303263	0.549380	0.741202
AIC Model	0.303674	0.549573	0.741332
Ridge Model	0.294901	0.555992	0.745649
LASSO Model	0.303857	0.549439	0.741242
EN Model	0.306895	0.529988	0.728003

Table 21: Model comparison in the training set

Model with a higher R square means it can explain more variability of the response variables. We can see through the table that EN model can explain about 31% of the variability which is higher than other models. Moreover, Model with a lower root mean squared error (RMSE) suggests that model is more accurate. EN model have the lowest RMSE value which is equal to 0.73. Above all, the project should choose EN model as the predictive regression model of white wine.

5.4.7 Applying model to the test set

In this section, we will apply these models to the test set to see the model performance.

	MSE	RMSE
OLS Model	0.5917951	0.7692822
AIC Model	0.5906657	0.7685478
Ridge Model	0.5875058	0.7664893
LASSO Model	0.5907114	0.7685775
EN Model	0.564784	0.7515211

Table 22: Model comparison in the test set

As a result, the EN model's performance is the best among the models.

6 MODEL APPLICATION

BLANKA VINHO VERDE is a type of white wine Vinho Verde. Its detailed information is shown as following table:

Variable	Values
fixed.acidity	6.67
Volatile.acidity	0.28
citric.acid	0.34
residual.sugar	5.7
chlorides	0.04
free.sulfur.dioxide	36.80
total.sulfur.dioxide	124.5
density	0.96
pH	3.21
sulphates	0.48
alcohol	11.5

Table 23: BLANKA VINHO VERDE Detailed Information

First, we have the classification model to classify, and then use the EN predictive model to predict the wine quality. Following the decision tree, we can see that the predicted result is white wine.

Wine Type	Prob.
0: Red	0.008429597
1: White	0.9915704

Table 24: Classification Result

	Quality
EN Model Predict	8.547254

Table 25: Predicted Quality Result

The customers comments on the wine is 4.2/5, which is roughly equal to 8.4/10. The EN model predict the wine quality is roughly about 8.55. The predicted result is really close to the customer's rate. As a result, the model have a great power to predict the white wine quality.

7 FURTHER THINKING

In this project, we used the classification and regression (CART) algorithm with different complexity parameter to classify different type of the wine. Then, we 5 different method including Ordinary Least Square (OLS), Akaike information criterion (AIC), Ridge, least absolute shrinkage and selection operator (LASSO), and Elastic net (EN) to predict the wine quality. Further research can be conducted to examine how much alcohol a bottle of wine should have if the producers have make sure about the wine type and wine quality. It is important for the wine producer to know how much alcohol and what is the density of wine if they want to produce the wine with a certain level. Further research also can input more variables such as the wine price or sold condition to enrich the dataset. These researches can also examine the condition whether a wine can sold or not.

8 CONCLUSION

In this research, we have used the CART method to build the classification model. The model's performance is relatively sound; the accuracy of the model is higher than 98%. The producers can use this classification model to classify their wine type based on the three main variables, which are total. Sulfur.dioxide, chlorides, and volatile. acidity. Moreover, the research also builds several models to predict the wine quality. The EN model's performance is the best among the models. However, although the linear relationship is significant, the adjusting R square of the multi-regression model can only explain about 31% of the wine quality variability. Further research can use more related variables to update the predictive model.