Analysis on the Wine Classification and Quality

YICHAO, DAI*

Department of Finance, College of Business and Management, Wenzhou-Kean University, Wenzhou, China

MEILI, LIU

Institute of Artificial Intelligence and Research Center, Midea Group, Shenzhen, China

CHUN-TE, LEE*

School of Mathematical Sciences, College of Science and Technology, Wenzhou-Kean University, Wenzhou, China

JENG-ENG, LIN

Department of Mathematical Sciences, George Mason University, USA.

The paper aim to build several models to classify the wine type and predict the white wine quality. All the datasets were collected through the UCI public data source [1]. The project first built the classification model based on the Classification and Regression Tree (CART) algorithm, then the decision tree will be updated by choosing the best complexity parameter. The classification model performance is great, the overall classification accuracy is higher than 98%. Moreover, the paper use 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 mean error and adjusting 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 the model performance.

CCS CONCEPTS • Classification and Regression tree • Shrinkage regression

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

1 INTRODUCTION

Wine consumption occupies an important world market and have become indispensable in families across the world [2]. Therefore, it is necessary to establish an effective model to help distinguish the type and quality of wine to help pricing the wine. Also, there is a huge demand from wine producers that developing an effective machine learning algorithm to classify the types is necessary as the process of making wines in mass production is very complicated. In the wine-making process, the makers of the wine must strictly maintain the alcohol content, density, PH value, sulfur dioxide and other ingredients. Variations in the content of the ingredients are likely to make a difference in the quality of the wine [3]. This paper would use the data from the Vinho Verde wine including the white wine and red wine, to do the classification analysis as well as regression analysis. The corresponding data about the ingredients used in the making of a wine have been collected from the Paulo Cortez and other producers [1]. It is believed that the results of the research can help other wine producers to evaluate their product qualification and adjust their production process.

2 DATA DESCRIPTION

The following table (Table 1) shows a brief introduction to the data description.

	Data Type	Description
fixed.acidity	Numeric	The fixed acids found in wines are tartaric, malic, etc.
Volatile.acidity	Numeric	The amount of acetic acid in wine, which can lead to different 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.
total.sulfur.dioxide	Numeric	Amount of free and bound forms of S02;
density	Numeric	The density of water
pH	Numeric	Describes how acidic or basic a wine is.
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 1: Data description

3 EXPLORATORY DATA ANALYSIS

3.1 EDA on wine classification

Take a close look at the difference of the alcohol level in its boxplot and density distribution (Figure 1), 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 2), also shows the specific descriptive statistics about the alcohol in different type of wine. However, the distribution of the alcohol among different type of wine are similar, using the Student-t test to find out the 95% confidence interval of the difference of the average alcohol in different type of wine. Table 3 shows that the confidence interval is (-0.1539 , -0.0287), as a result, it has 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, it can conclude that there is a significant difference in alcohol between white wine and red wine.

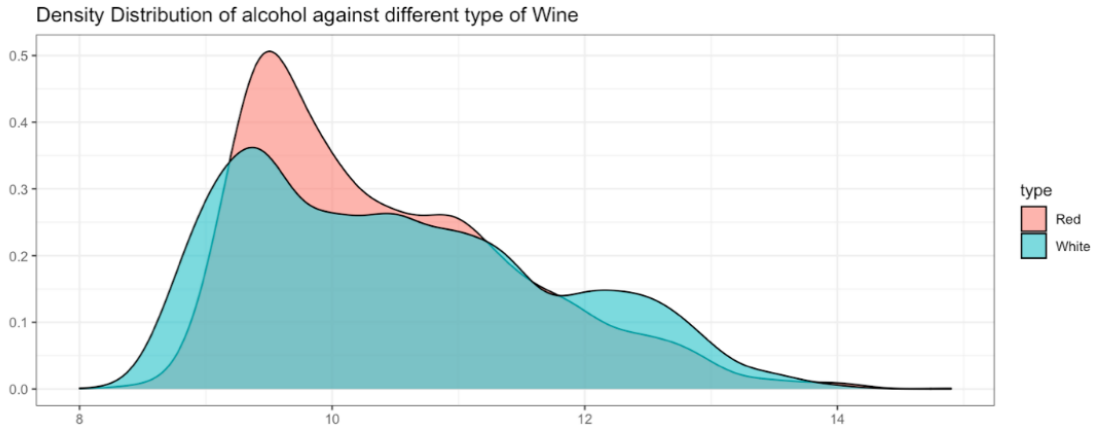

Figure 1: Boxplot and Density distribution of alcohol against different wine type

	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 2: Descriptive statistics on alcohol

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

Table 3: Confidence interval test on alcohol of two different type of wine

3.2 EDA on wine components correlation

The correlation matrix (Figure 2) shows the correlation of each two variables of wine components, 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 [4]. However, removing which one is the most important thing to decide.

Figure 2: 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. To examine this possibility, variance inflating factor (VIF) can examine statistical significance of regression coefficients [5].

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

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 [6]. 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 4: VIF

As a result, 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 (such as Lasso regression) to eliminate the multi-colinear problems [7].

4 MODEL BUILDING AND ANALYSIS

In this section, the paper aims to develop a classification model to distinguish between red wine and white wine based on the *Classification and Regression Tree* (CART) algorithm. A CART algorithm is a predictive model

algorithm, which explains how an outcome variable's values can be predicted based on other values. Here the target value of red wine is set to be 0 while the target value of white wine is set to be 1.

4.1 Classification Model on Wine Type

4.1.1 Maximum Tree

By using CART algorithm, some of the default values control the size of the trees, e.g. maximum depth, minimum samples and leaves and lead to fully grown and unpruned trees which can potentially be very large on some data sets. The default decision trees offer many benefits including the stopping criteria when a maximum depth of the tree is reached and automated tuning with optimal parameters. In our model the default function is Gini impurity which is used to measure the split of the variables and the complexity parameter (CP) is set up to be 0.001. It is noted that CP determines the complexity of the classification model. Maximum tree requires the CP value to be set to 0 and as a result zero will get the "most" complex classification model in the procedure [8]. Therefore, the CART algorithm under CP=0 produces the results with overall performance summarized in Table 5. It is seen that the model accuracy is around 0.9805, which suggests that our model algorithm is highly accurate and reliable.

Items	Value
Accuracy	0.9805
95% CI	(0.9733, 0.9862)
Sensitivity	0.9742
Specificity	0.9825
Pos Pred Value	0.9458
Neg Pred Value	0.9918

Table 5: Maximum Tree Performance

4.1.2 Tuning Tree

An optimal CP value can be estimated by testing different cp values and using cross-validation approaches to determine the corresponding prediction accuracy of the model. The best cp is then defined as the one that maximize the cross-validation accuracy. Our goal in this section is to execute the CART classification algorithm with the best selection of CP value. The graph in Fig. 7 demonstrates the results of accuracy of model for different complexity parameter. The plot shows that when CP is equal to 0.0065, the model have a highest accuracy. In addition, the graph in Fig. 8 shows the corresponding decision tree with complexity parameter being equal to 0.0065. It is noted that the model tree is more complex that the maximum tree with higher accuracy under the best CP value.

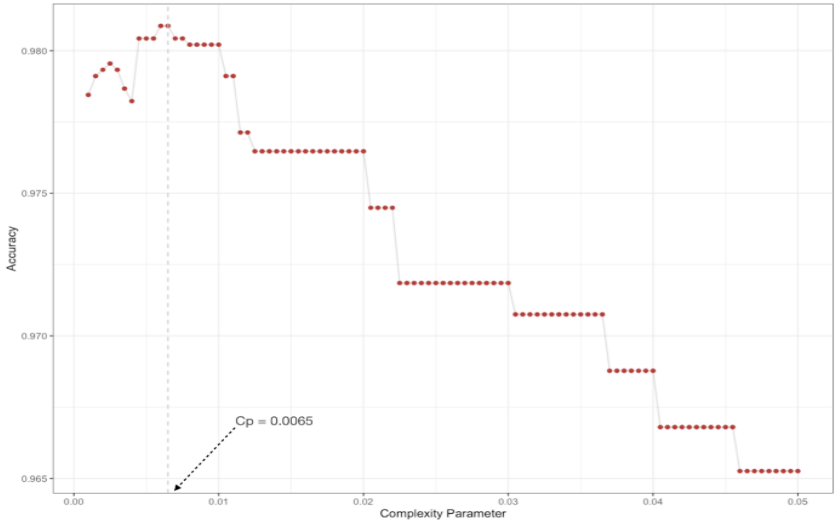

Figure 3: Tuning Tree Selection Process

Items	Value
Accuracy	0.9841
95% CI	(0.978, 0.989)
Sensitivity	0.9828
Specificity	0.9845
Pos Pred Value	0.9521
Neg Pred Value	0.9946

Table 6: Tuning-model Performance

In addition, it can be a effective way to employ the Receiver Operating Characteristics (ROC) and AUC (Area Under The Curve) method to check the performance of the model. ROC is a probability curve and AUC represents degree or measure of separability [9]. It tells how much model is capable of distinguishing between classes. The plot in Fig. 10 shows the sensibility of ROC curve against 1-specificity feature of the classification model. It is observed that the AUC for this classification model is 0.982. general, higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, higher the AUC, better the model is at distinguishing between white wine and red wine. Consequently, our classification model has a strong ability to distinguish different types 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.

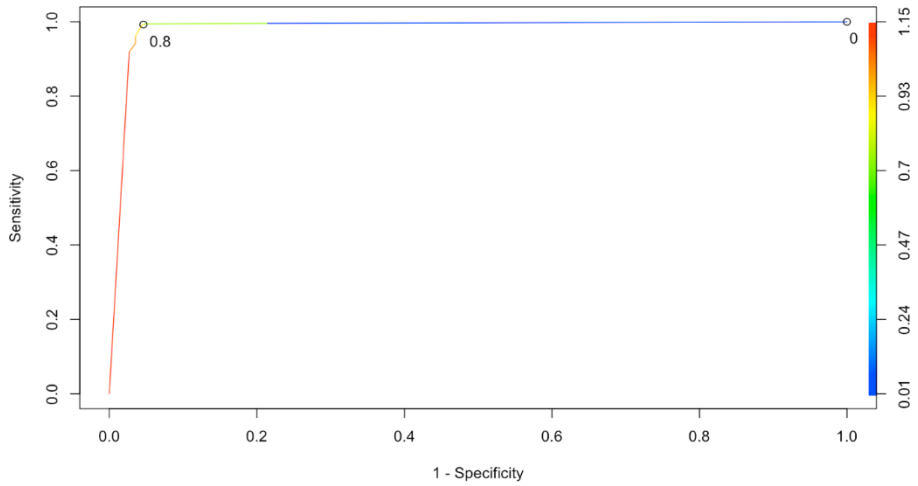

Figure 4: ROC and AUC

4.2 Numerical Experiments

4.2.1 OLS Model of White Wine Quality Prediction

The OLS multi-regression should first follow the 3 main assumptions: normality of errors, correlation and multicollinearity, and homoscedasticity [10]. Fig.5 illustrates the overall performance and diagnostic plots of the OLS model. In Fig.5(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. Fig.5(b) demonstrates the normal quantile-Quantile (Q-Q) plot to check if residuals are normally distributed. As a result, one can see that residuals follow a straight dashed line to meet the assumption of normality. Fig.5(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 meets the assumption of equal variance (homoscedasticity). In Fig.5(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. Fig.5

(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.

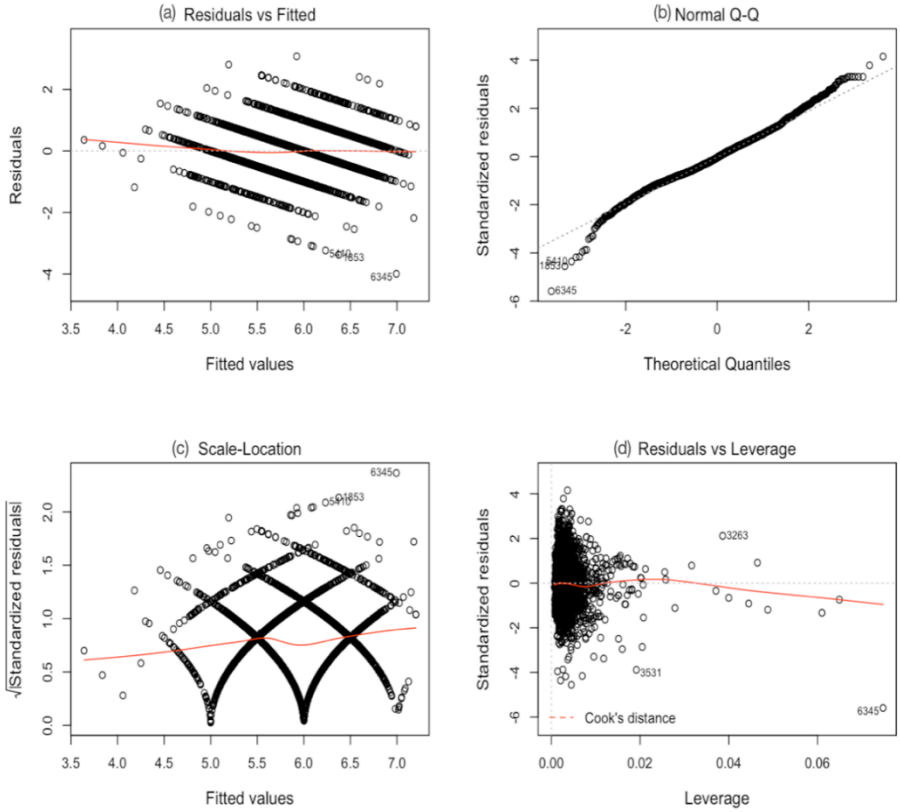

Figure 5: Diagnostics Plot

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 the study exclude those cases. Fig.6 is an enlarged view of Fig.5(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 becomes normal.

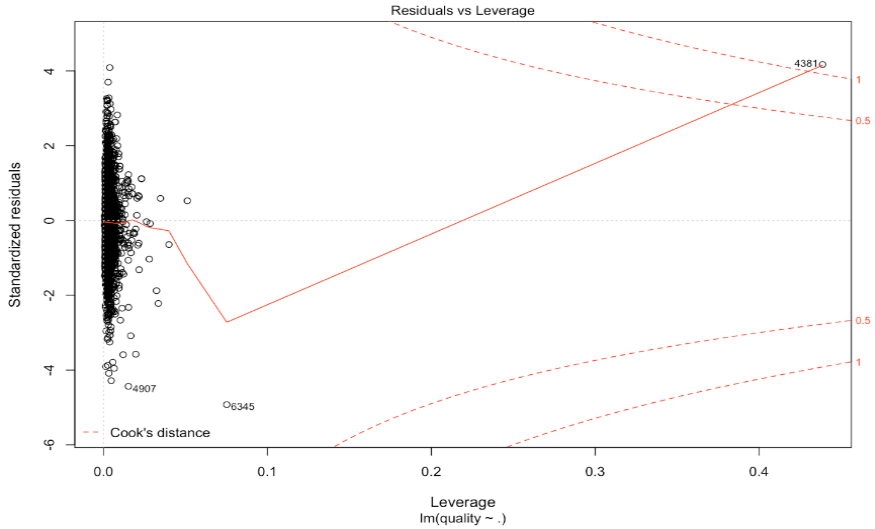

Figure 6: Comparison After removing the outlier

The leverage line is almost around 0, which suggests that there are no more extremely influential points in the datasets. In the following method. The paper would continue use the train datasets without this outlier.

4.2.2 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 7: AIC Backward Model Selection

The final model achieved by the stepwise backward method is as following table (*Table 12*) with a lowest $\mathrm{AIC} = 7697.0304$. 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.

4.2.3 Ridge & LASSO Model of White Wine Quality Prediction

The Ridge and LASSO algorithm are shrinkage regression which can eliminate the effect of multicollinearity [11]. Equation (2) show the Ridge equation. 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.

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

The plot on the right of Figure 7 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. Each lambda value will have the corresponding whole path of coefficients. The cross-validation process will pick up the best model.

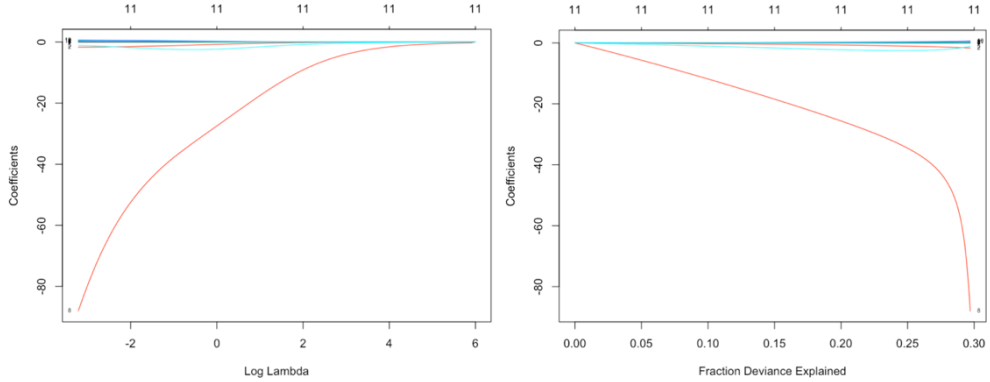

Figure 7: Ridge Model

The model selection process of LASSO model can be observed at Figure 8. The model shown the relationship between the value of logarithm of lambda and the mean squared error (MSE) of the model. As a result, 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, which shows some variable have been eliminated.

Variable	Coef.	Variable	Coef.
fixed.acidity	-0.022958065	total.sulfur.dioxide	.
Volatile.acidity	-1.656699060	density	.
citric.acid	.	pH	.

residual.sugar	0.013366010	sulphates	0.261196066
chlorides	-0.593160659	alcohol	0.342926332

Table 8: LASSO Model Coefficients

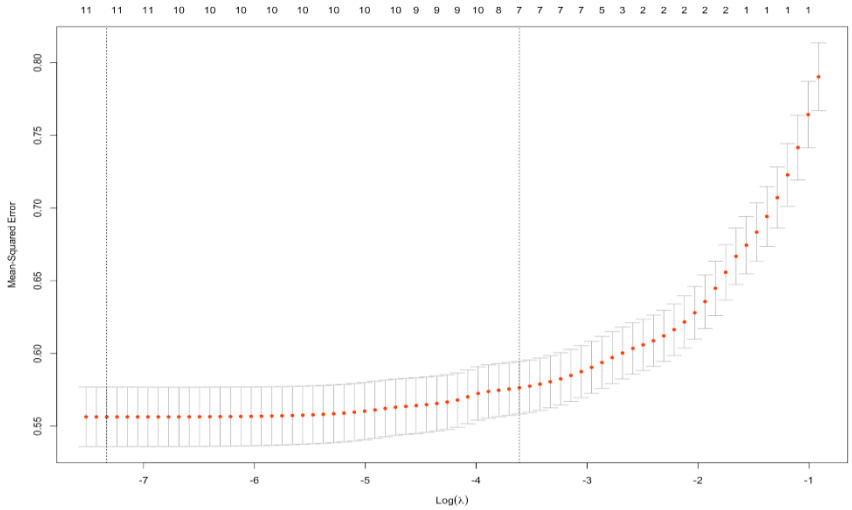

Figure 8: Lasso Model Selection

4.2.4 EN Model of White Wine Quality Prediction

EN model is the combination of the LASSO and Ridge model [11]. It depends on the mixed percentage. Figure 9 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.

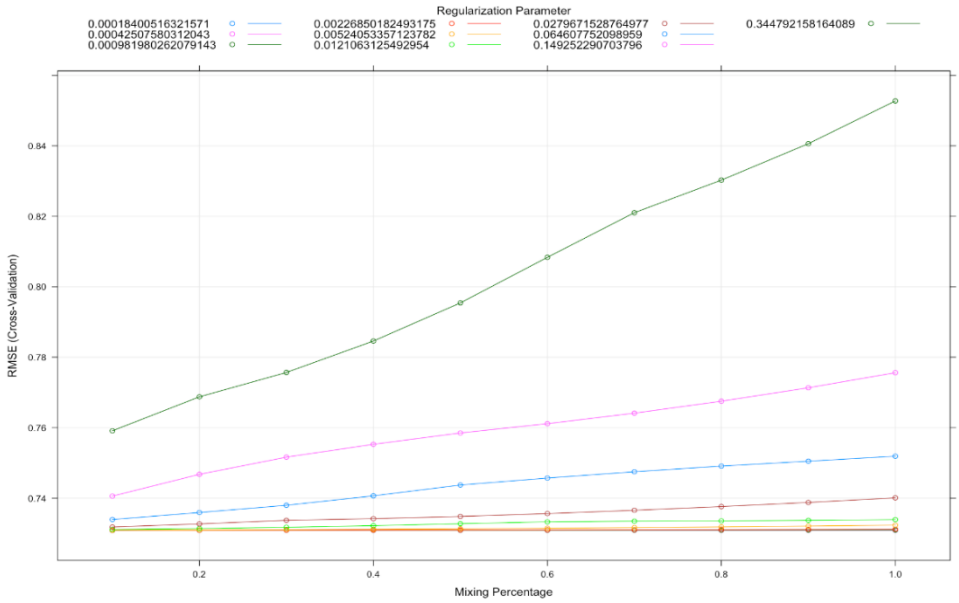

Figure 9: EN Model Selection

4.2.5 Comparison Among the models

In this section, the project tries to compare different models by using the adjusting R square and root mean squared error of each model. Table 9 shows 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 9: Model comparison in the training set

Model with a higher R square means it can explain more variability of the response variables [12]. 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.

4.2.6 Applying model to the test set

In this section, models would be applied to the test set to show the performance. As a result, the EN model's performance is the best among the models.

	RMSE
AIC Model	0.7685478
Ridge Model	0.7664893
LASSO Model	0.7685775
EN Model	0.7515211

Table10: Model comparison in the test set

5 MODEL APPLICATION

BLANKA VINHO VERDE is a one of white Vinho Verde. Its detailed information is shown as following table [13]:

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

Table 11: BLANKA VINHO VERDE Detailed Information

First, the study uses the classification model to classify, and then uses the EN predictive model to predict the wine quality. Following the decision tree, the predicted result is undoubtedly white wine with probability 99%.

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

Table 12: Classification 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.

	Quality
EN Model Predict	8.547254

Table 13: Predicted Quality Result

6 CONCLUSION

The purpose of this article is to establish an effective model to help distinguish the type and quality of wine in order to help wine pricing. CART algorithm has been used to build the classification model. The performance of the model is relatively good, 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 paper 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 variability of the wine quality. The results of the study could help other wine producers assess the quality of their products and adjust their production processes. Wine producers can use the models in this article and their own unique pricing models to determine the prices of their different categories of wine. Further research can update the prediction model with more relevant variables. If the producer has determined the type and quality of the wine, further research can be conducted to check how much alcohol a bottle of wine should contain. It is important for wine producers to know how much alcohol and the density of the wine if they want to produce a certain level of wine. Further research can also input more variables, such as wine prices or sales, to enrich the data set. The studies can also test whether the wine will sell.

REFERENCES

[1] 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. Retrieved from https://doi.org/10.1016/j.dss.2009.05.016

[2] Artero, A., Artero, A., Tarín, J. J., & Cano, A. (2015). The impact of moderate wine consumption on health. Maturitas, 80(1), 3–13. https://doi.org/10.1016/j.maturitas.2014.09.007

[3] Almeida, C. M. R., & Vasconcelos, M. T. S. D. (2004). Does the winemaking process influence the wine 87Sr/86Sr? A case study. Food Chemistry, 85(1), 7–12. https://doi.org/10.1016/j.foodchem.2003.05.003

[4] Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

[5] Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. Quality Engineering, 14(3), 391–403. https://doi.org/10.1081/QEN-120001878

[6] Stine, R. A. (1995). Graphical Interpretation of Variance Inflation Factors. The American Statistician, 49(1), 53–56. https://doi.org/10.1080/00031305.1995.10476113

[7] Martorell, A., Gutierrez‐Recacha, P., Pereda, A., & Ayuso‐Mateos, J. L. (2008). Identification of personal factors that determine work outcome for adults with intellectual disability. Journal of Intellectual Disability Research, 52(12), 1091–1101. https://doi.org/10.1111/j.1365-2788.2008.01098.x

[8] Banaszak, Z. A. (2006). CP-Based Decision Support for Project Driven Manufacturing. In J. Józefowska & J. Weglarz (Eds.), Perspectives in Modern Project Scheduling (pp. 409–437). Springer US. https://doi.org/10.1007/978-0-387-33768-5_16

[9] Saul, L. K., Weiss, Y., & Bottou, L. (2005). Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference. MIT Press.

[10] Banerjee, T. (2020). Forecasting Apple Inc. Stock Prices Using S P500– An OLS Regression Approach with Structural Break. 2020 IEEE 1st International Conference for Convergence in Engineering (ICCE), 306–310. https://doi.org/10.1109/ICCE50343.2020.9290495

[11] Verducci, J. S. (2007). Prediction and Discovery: AMS-IMS-SIAM Joint Summer Research Conference, June 25-29, 2006, Snowbird, Utah. American Mathematical Soc.

[12] Akossou, A., & R., P. (2013). Impact of data structure on the estimators R-square and adjusted R-square in linear regression. International Journal of Mathematics and Computation, 20, 84–93.

[13] BLANKA VINHO VERDE Information Page. https://www.vivino.com/blanka-branco/w/5396503?year=2017