Advanced Analyses on Wine Quality Dataset

Introduction

‘Vinho Verde’ wine is the perfect option for a summer afternoon with all kinds of food. It is one of the most renowned wine varieties in the country, as famous as Ports wine. However, ‘Vinho Verde’ is actually a small region in Northern Portugal. It literally means ‘green wine’ in Portuguese. The idea behind the name, ‘Vinho Verde’ is that it is harvested early and drunk young. Many residents in the area believe the name is derived from the lush natural environment (Signer, 2016).

There are several kinds of Vinho Verde wine. White Vinho Verde wine is the most prevalent type; red and rosé Vinho Verde wines are also offered in Portugal. The majority of the wines you will try are most likely a combination of different sorts of grapes and they usually contain 8.5% to 11% alcohol. However, some certain higher-quality wines comprise a specific type of grape (Costa, 2022).

The Wine dataset collects the numerous chemicals in wine and their effect on the quality of red variants of Portuguese "Vinho Verde" wine. Variation is a genetic mutation, which is due to the cumulative variation factor that occurs when the vine is cloned in large numbers. Over time, trees grow with the radiation of sunlight, and in the process, there will be a few genetic mutations, which are variants. In this report, we aim to figure out what chemicals significantly influence the quality of red variants of wine from Vinho Verde, Portuguese.

Data set source and detailed description

The data used for analysis is reported by Bengaluru Karnataka in a CSV file on Kaggle (https://www.kaggle.com/datasets/yasserh/wine-quality-dataset). It is about how the chemicals in wine will affect the quality of the red “Vinho Verde” wine in Portuguese. To be noticed, the number of qualities of the wines is unbalanced. For example, the number of normal wines is much more than the quantity of excellent or poor wines. Overall, there are a total of 1143 observations of 13 variables. Specifically, there are 11 input variables and 1 output variable; the other variable is the serial number of the wine. The inputs are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol; the output is wine's quality.

Variable Explanations:

1. Fixed acidity (tartaric acid - g/dm^3): fixed or non-volatile
2. Volatile acidity (acetic acid - g/dm^3): the amount of acid in wine, if the level is high, there will be a vinegar taste. Normally,1.2 g/dm^3 in whites and 1.4 g/dm^3 in reds.
3. The citric acid (g/dm^3): act as a preservative; small quantities add freshness and flavour to wines
4. Residual sugar (g/dm^3): the amount of sugar in the wine after fermentation stops (45g/dm^3 is considered sweet)
5. Chlorides (g/dm^3): the amount of salt in wine
6. Free sulfur dioxide (mg/dm^3): prevent microbial growth and the oxidation of wine
7. Total sulfur dioxide (mg/dm^3): when concentrations are over 50 mg/dm^3, people can smell the sulphur dioxide and taste in the wine
8. Density (g/dm^3): depends on the quality of alcohol and sugar contain in water
9. PH: describe the acidity or basicity of wine. Most wines are between 3 to 4.

10. Sulphates (g/dm³): act as an antimicrobial and antioxidant
11. Alcohol: the presence of alcohol contained in the wine
12. Wine quality: the bad or good of the wine quality, range from 0 to 10. The higher the score, the higher the quantity of wine

Methodology

Exploratory Analysis

We checked whether uncompleted cases existed in the dataset first. We decided to remove them from the dataset if we found missing values. We converted the class of wine quality from integer to factor with two levels,' 0' and '1', where '0' represents not good quality, and '1' represents good quality. We considered a wine quality as 'good' if the wine quality is rated equal to or higher than 7. We grouped the wines with quality lower than seven as not good wines. This process facilitated the wine quality prediction and transformed it into a classification problem. As a result, 156 observations were good wines, and 987 were not good wines. There is an unequal distribution of classes within the dataset. We used both oversampling and undersampling techniques to resample the response variable in the training set to solve this problem. We also kept the original response data for the testing set. Since the difference between the two classes is extensive, using only oversampling or undersampling is inappropriate. A histogram of the resampled response variable was plotted to show the balanced classes. Since the variable 'Id' is unnecessary for our analysis, we removed it, and the dataset contains 12 variables. We plotted a heatmap to demonstrate the relationships between variables and observe which variables strongly correlated with each other. We plotted 11 boxplots for each independent variable regarding the dependent variable since the response is binary; this was our initial investigation of significant variables of wine quality.

Modeling

We generated three models to identify the significant variables of wine quality. Since we changed the data into a classification problem and the response is binary, we chose Random Forest, Logistic regression, and Classification Tree. Before modeling, we split the original data and resampled data into training and testing sets to obtain a more accurate result. The training set used for models was from the resampled dataset and the testing set used for prediction was from the original dataset to show honest performance results of models.

The first model is Random Forest. This model can automatically deal with the multicollinearity caused by highly correlated variables. By pruning the random forest, one of the two highly correlated variables will be removed from the model. Then, we plotted the variable importance of random forest to observe which was dominant in the wine quality. The second model is logistic regression. The function stepAIC() was used to choose the significant variables and it gave the model with the smallest AIC. Then, we checked the cook's distance for the logistic model to figure out whether the outliers were influential points. If they were, we decided to remove them from the training set and generate the model again without the influential points. We checked the variance inflation factor (VIF) to ensure there are no correlated variables in the model. Then, the variables left in the logistic model were considered as significant. The third model is the classification tree. This method immutes the correlation by nature. It only chose one of the perfect correlated features when it split. Similar to the random forest model, we pruned the classification tree as well. Then, we could figure out the important variables from the pruned tree.

Evaluation Metrics

We built a confusion matrix for each model by the predicted values and testing set. In the confusion matrix, TP, FP, TN, FN stand for true positive, false positive, true negative, and false negative, respectively. We computed the sensitivity, accuracy, and specificity, and plotted the ROC curves for each model to evaluate its performance on the testing set. Sensitivity is computed as : TP / (FN + TP). Accuracy is calculated as (TP + TN) / (TN + TP + FP + FN). Specificity is calculated as TN / (TN + FP). Since the problem caused by data imbalance has been solved before modeling, using accuracy to evaluate the models was applicable.

The ROC curve has two axes; the y-axis is the true positive rate (TP / TP + TN) and the x-axis is false positive rate (TN / TN + FN). The ROC curve demonstrates the trade-off between true positives and false positives. When comparing the models, we chose the model with the largest area under the ROC curve (AUC), which also gave the best accuracy. The AUC can be obtained by the following equation:

$$AUC = \int_0^1 ROC(t)\, dt$$

Results

Exploratory data analysis

When exploring the relationships between predictors and the response, we found that the fixed acidity, volatile acidity, citric acid, density, sulphates, and alcohol tended to have a larger impact on the wine quality. There was a huge difference between their mean with good quality and their mean with bad quality. In **Figure 1**, it shows the boxplot of alcohol and wine quality. Our initial investigation was that wines were likely to have good quality when they contained a higher percentage of alcohol. From the boxplots, we observed that volatile acidity and density had a negative relationship with wine quality. Whereas, sulphates, citric acid, alcohol and fixed acidity were positively related to wine quality.

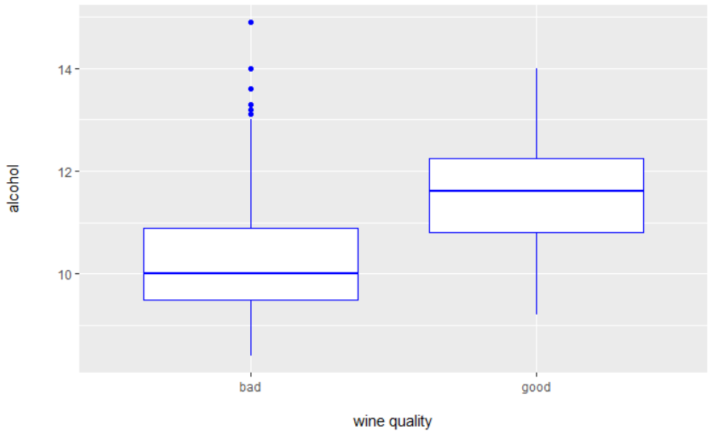


Figure 1. The boxplot of alcohol vs wine quality

Then, we had checked the correlation between predictors. From the correlation plot (**Figure 2**), there were both positive and negative correlations between the variables. The negative correlation meant when variable A had increased, variable B would decrease, vice versa. For example, when the percentage of fixed acidity in a bottle of wine has increased, the PH of the wine would decrease. Citric acid, density, and pH, were correlated with fixed acidity. Volatile acidity and pH were correlated with citric acid. Free sulfur dioxide was correlated with total sulfur dioxide.

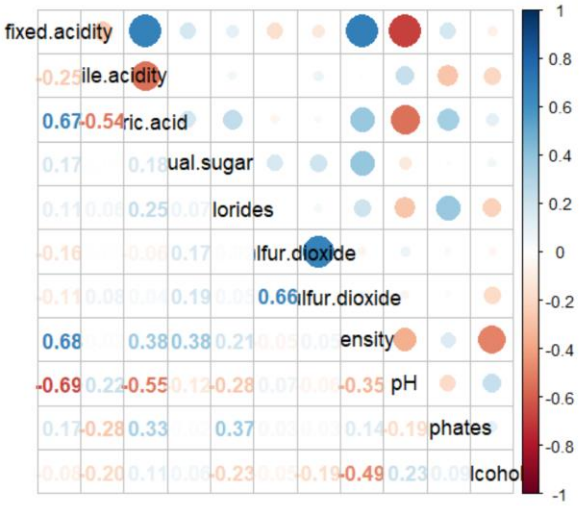


Figure 2. The correlation between the predictors

A histogram of the dependent variable was shown to visualize whether the two classes were balanced as well. It turned out that the count of bad wine quality was extremely larger than the count of good wine quality. Due to the imbalanced data sets, the machine learning algorithms could not get the sufficient information about the minority class (Kreiger, 2020); the outcomes would struggle with accurate prediction. Therefore, we would deal with the imbalanced data sets before modeling.

Before using the correct method to solve the imbalanced problem, we split the dataset of wine into a training set and a testing set, which were 80% and 20% respectively. Then, we left the testing set unchanged and made changes to the training set. The dataset for the training part only contained 110 good wines and 690 bad wines, which was 13.75% and 86.25% respectively. After building a decision tree to demonstrate how inadequately this problem could affect the prediction accuracy, we found the following interpretation. With the threshold value of 0.5, the precision is 0.622, which implied there are some false positives. The recall score was 0.469 and we had a higher number of false negatives. The F-value was low, which was 0.267, and suggested that the accuracy was weak in this model. Therefore, it was necessary to balance the dataset.

As the oversampling would give more information from the sample and undersampling would lose significant information from the sample, we used both oversampling and undersampling on the training set. Now the amount of bad quality was 595 and the amount of good quality was 548. Notice that, the total number of observations has remained unchanged, which was still 1143 observations. In the following steps, we used the resampled training set to generate models and tested the models on the unchanged testing set.

Main data analysis

Random Forest Tree

To start off, we chose to fit a random forest model on the training set. The output noted that the random forest included 500 trees and tried 3 variables at each split. The out-of-bag (OOB) error rate was really low, about 1.57%, so the train data set model accuracy was around 98%. Then we checked the test accuracy, which was about 87%. The classification error for bad quality is 3.19% and for good quality is 0.18%. From **Figure 3**, the error rate was stabilized with an increase in the number of trees. Although, as we mentioned before, there are multicollinearity in this data, the random forest did not affected by it (Bernard, Heutte & Adam, 2011). Also, random forest trees were not sensitive to outliers, so we did not have to worry about this problem in this model (Fawagreh, Gaber & Elyan, 2015).

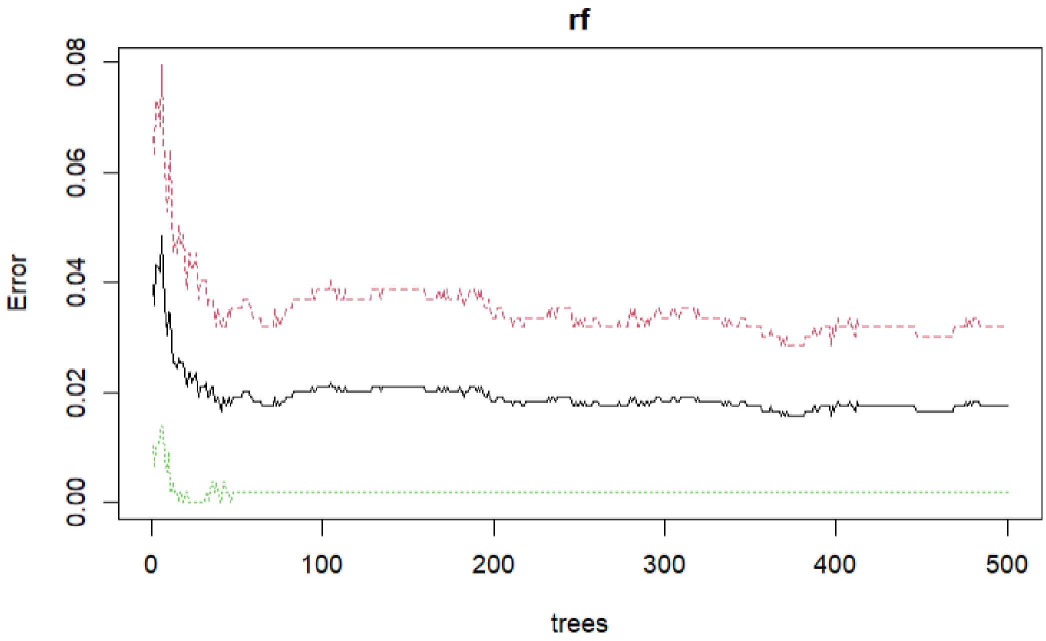


Figure 3. Error rate for the number of the trees

We want to know the importance that the classifier had assigned, so we check both mean decrease accuracy and mean decrease Gini (**Figure 4**). According to the mean decrease accuracy plot, we observed a rough estimate of the loss in prediction performance when the particular variable was omitted from the training set; the 3 most important variables were alcohol, sulphates and volatile acidity. To be noticed, there was a gap between the third and the fourth variables, which were voltie acidity and citric acid respectively. The mean decrease in Gini, which measured the node impurity, showed that the 3 most important variables are alcohol, sulphates and volatile acidity as well. To be noticed, there was a large gap between the variable of alcohol and sulphates. For both plots, the alcohol was the most important feature.

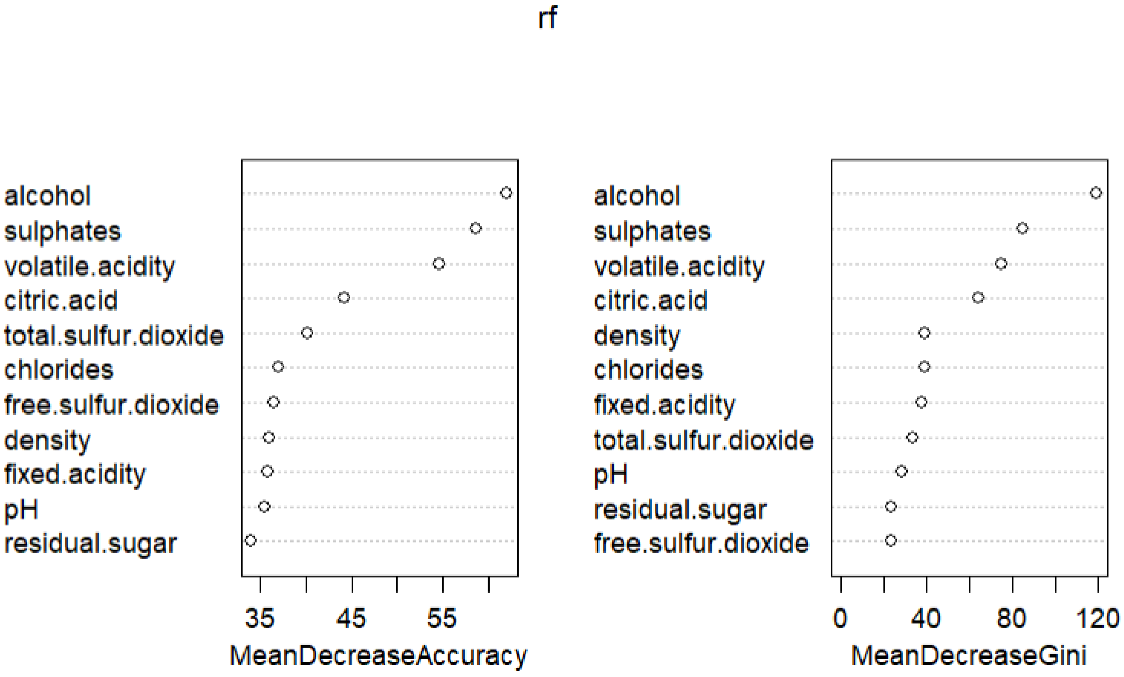


Figure 4. Importance parameter for random forest model

Then we checked the area under the curve (AUC), sensitivity, specificity and accuracy, which were 0.906, 0.633, 0.915 and 0.875 respectively. As the value of AUC was higher than 0.9 and the accuracy is more than 0.8, this model was considered a good performance.

Logistic Regression

Since the logistic regression could not handle both high correlations and was affected by outliers in the data, we had to use other methods to generate the best model. After building the initial logistic regression model with all variables, we checked the variance inflation factor (VIF) (Bertsimas & King, 2017). The fixed acidity was exceedingly correlated with other variables since its VIF (= 10.584) was over 10. Therefore, we used Akaike's Information Criteria (AIC) with both backward and forward directions to find the model with the smallest AIC. Ten variables were left in the model: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, sulphates, and alcohol. The AIC was 966.78. Except for the density, all other predictors were significant.

We did not discard the density even though it was not substantial. The reason was that the AIC would increase by removing it. We rechecked the VIF and were all under 6.5, which became adequate. Then, we checked the influential points. The plot of the cook's distance (**Figure 5**) showed that no influential points were needed to be concerned and removed from the model (Peng & So, 2002).

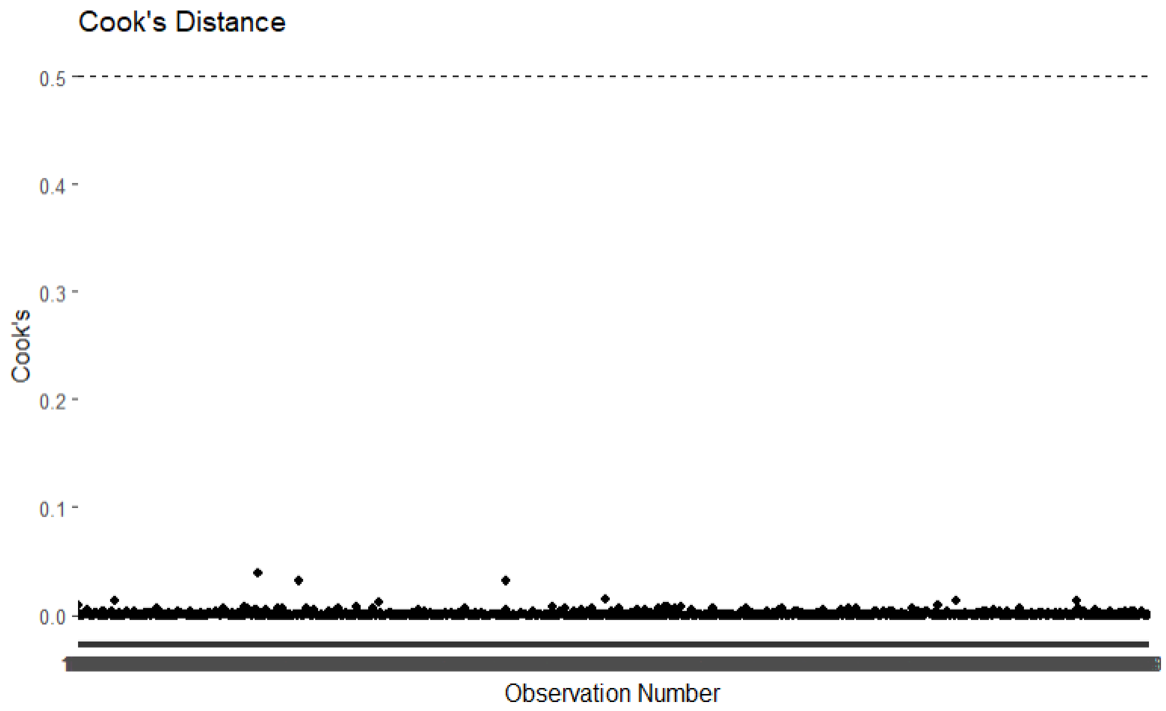


Figure 5. Plot of cook’s distance

We computed the area under the curve (AUC), sensitivity, specificity and accuracy, which were 0.865, 0.714, 0.799 and 0.781 respectively. As the value of AUC was higher than 0.8 and the accuracy was close to 0.8, we considered that this model had an acceptable performance.

Classification Tree

Although we noted during earlier stages of the project that there was some correlation between our predictors, the regression tree was one of the main types of the decision tree, which could immunize multicollinearity by nature (Shannon & Banks, 1999). Therefore, we did not have to worry about the correlated variables in this model. Also, there were some outliers in this dataset. However, the decision tree handled the outliers automatically. Hence we did not have to worry about the influential points (Kreiger,2020).

First, we used all the predictors to build a sizeable initial classification tree model. We used a small value of complexity parameters to ensure that the tree was large enough. We found that only the alcohol, chlorides, citric acid, density, fixed acidity, sulphates, total sulfur dioxide,

and volatile acidity were used in the tree. As we wanted the lowest test error, we pruned the tree. We could notice from the final pruned tree (**Figure 6**) that it had 12 terminal nodes. Each terminal node showed the predicted quality of wines' in that node along with the percentage of observations from the original data set that belonged to that node.
Then we checked the area under the curve (AUC), sensitivity, specificity, and accuracy, which were 0.807, 0.714, 0.793, and 0.781, respectively. As the value of AUC was higher than 0.8 and the accuracy was close to 0.8, this model was considered acceptable performance.

From the pruned tree, the wines which had a percentage of alcohol greater than 10, a content of sulphates more than 0.6, and grams per decimeter to the third power of volatile acidity less and equal than 0.6, were the most likely to have good qualities. The least likely group to have bad quality wine were the percentage of alcohol less than 10, the content of fixed acidity less than 12, the amount of salt in wine greater and equal than 0.06, and the grams per decimeter to the third power of citric acid less than 0.7.

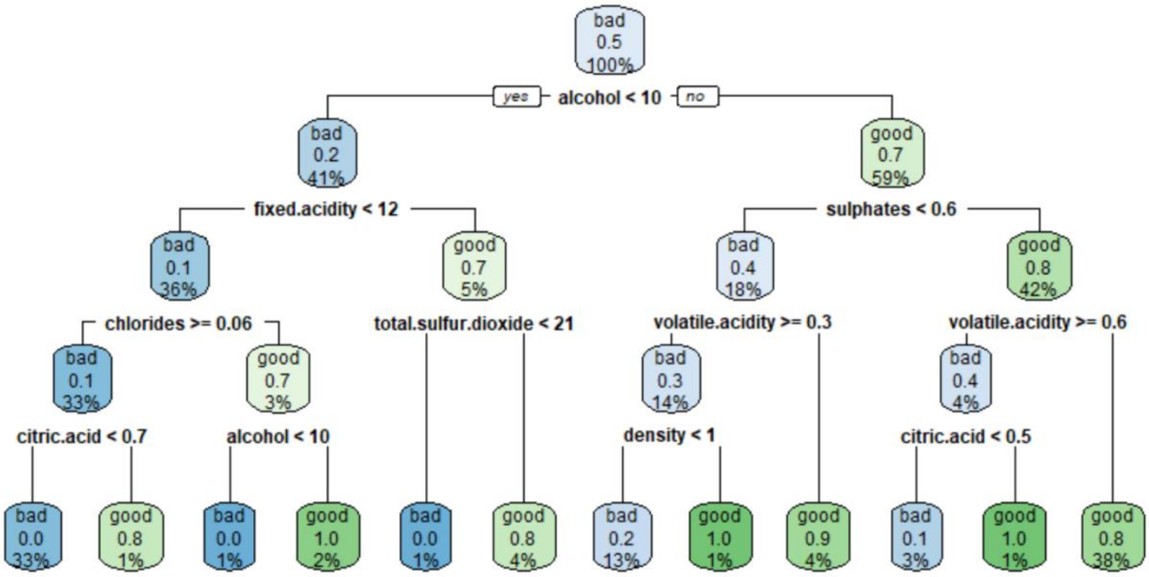
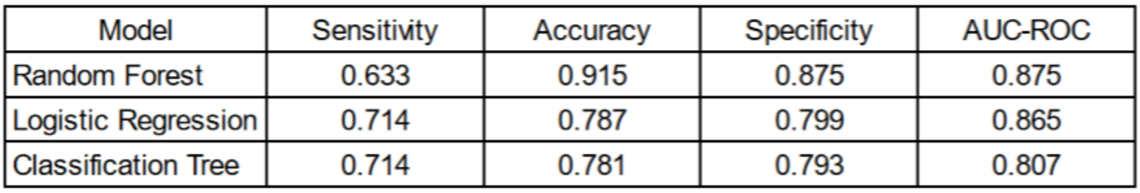


Figure 6. The pruned classification tree

Discussion
We discussed three models in this report, random forest, logistic regression, and classification tree. According to **Figure 7**, the ROC curves for the three models were plotted together for comparison. The red, green, and blue lines represented the random forest, logistic regression, and classification tree ROC curves, respectively. The random forest demonstrated the best performance among all three models since it had the greatest AUC (= 0.906).



Model	Sensitivity	Accuracy	Specificity	AUC-ROC
Random Forest	0.633	0.915	0.875	0.875
Logistic Regression	0.714	0.787	0.799	0.865
Classification Tree	0.714	0.781	0.793	0.807

Table 1. A table gathers the sensitivity, accuracy, specificity, AUC-ROC of three models

According to Table 1, the random forest model had a satisfying accuracy(= 0.915), much higher than the other two models (= 0.787 and 0.781). We struggled with choosing logistic regression or random forest. The logistic regression model also had a good performance on the ROC curve (AUC-ROC = 0.865) and a fair specificity (= 0.799). However, the random forest model had a huge advantage in accuracy and specificity. Although the low sensitivity (= 0.633) in the random forest model would lead to false-negative results, the sum of its sensitivity and specificity (= 0.633 + 0.875 = 1.508) was greater than 1.5, which can be considered pertinent. To sum up, we chose the random forest model as our best model.

It was evident that the variables alcohol, sulphates, and volatile acidity significantly influenced the wine quality. When the alcohol in the wine increased, wine quality was more likely to be good. As the sulphates in the wine increased, the wine quality tended to be better. If the volatile acidity decreased, the wine quality would be better.

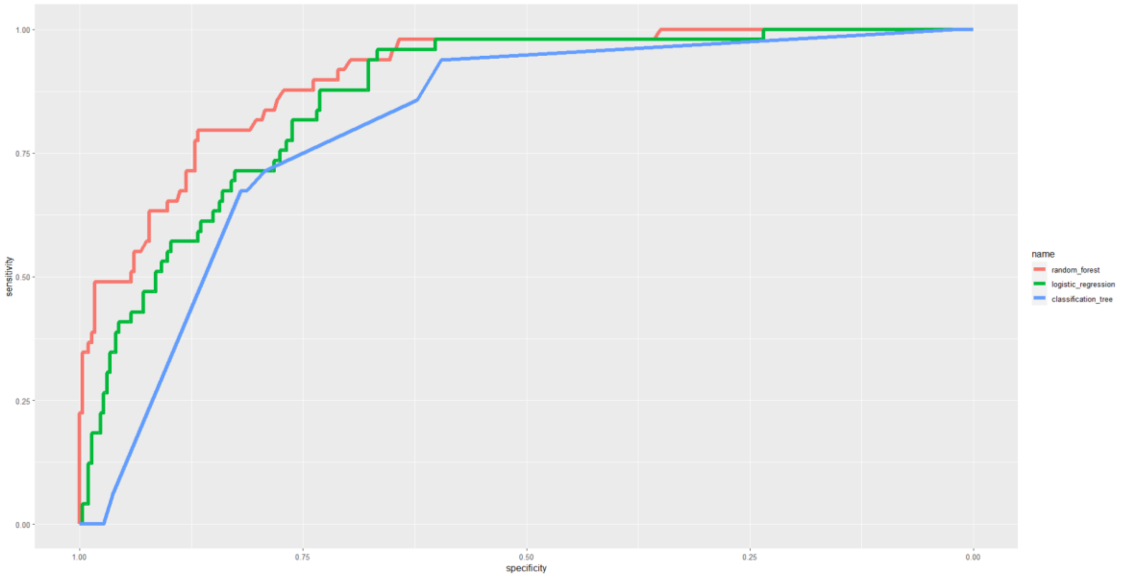


Figure 7. ROC curves for random forest (red), logistic regression (green), and classification tree (blue).

References

Bertsimas, D., & King, A. (2017). Logistic Regression: From Art to Science. Statistical Science, 32(3), 367–384. Retrieved from: http://www.jstor.org/stable/26408297

Bernard, S., Heutte, L., & Adam, S. (2011). A Study of Strength and Correlation in Random Forests. *International Conference on Intelligent Computing*, Changsha, China. Pp.186-191, DOI:10.1007/978-3-642-14831-6_25

Costa, G. (2022). *Protuguese Wine Guide: Vinho Verde.* Retreived from: Vinho Verde - Portuguese Wine Guide - Portugal.com

Fawagreh, K., Gaber, M. & Elyan, E. (2015). An Outlier Detection-based Tree Selection Approach to Extreme Pruning of Random Forests. *School of Computing Science and Digital MediaRobert Gordon University.* Retrieved from: https://doi.org/10.48550/arXiv.1503.05187

Kreiger, J. (2020). Evaluating a Random Forest model. Analytics Vidhya. Retreived from: https://medium.com/analytics-vidhya/evaluating-a-random-forest-model-9d165595ad56

Peng, C. & So, T. (2002). Logistic Regression Analysis and Reporting: A Primer, Understanding Statistics, 1:1, 31-70, DOI: 10.1207/S15328031US0101_04

Rachel, S. (2016). *7 Things You Need To Know About Vinho Verde.* Retrieved from: https://vinepair.com/wine-blog/7-things-you-need-to-know-about-vinho-verde/

Shannon, W. & Banks, D. (1999). Combining classification trees using MLE. STATISTICS IN MEDICINE. Retrieved from: https://doi.org/10.1002/(SICI)1097-0258(19990330)18:6<727::AID-SIM61>3.0.CO;2-2

Appendix for R code