Prediction of Wine Quality (Red and White Wine) Using Machine Learning Models

STSCI 4740/5740 Final Project

Professor Ning

Nicholas Paschall (nbp33), Scarlett Wang (yw2448), Wei Yang (wy233), Shirley Zhang (xz499)

1. Introduction

As the demand for wine rises around the world, the supply of wine ranges increases as well. Finding methods to assess the quality of wine has become a vital task for the wine industry and consumers. To further study the assessment of wine quality, we use datasets on red and white vinho verde wine samples of the Portuguese "Vinho Verde" wine. The data contains physicochemical tests of the wine, provided by UCI Machine Learning Repository [1]. In this report, we seek to build separate models for red and white wine and use four different machine learning methods: multiple linear regression, gradient boosting, regression trees, and random forest. Among the four methods, the gradient boosting model outperformed other methods in predicting wine quality due to lower mean squared error (MSE), higher model interpretability and robustness.

2. Data Overview

2.1 Data Cleaning

In the original dataset, there are 6498 observations with 13 columns. The response variable is *quality,* and the remaining 12 variables are predictor variables: *type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates*, and *alcohol*. After making sure that there is no missing value, we noticed that all the predictor variables are continuous numerical values. We decided to standardize the numerical values with mean 0 and standard deviation 1. Standardizing the predictor variables can ensure that all features contribute equally to the model.

By looking at the table distribution of response value wine quality, we noticed the range is from 3 to 9, but quality 3 and quality 9 only contain 0.54% of the total data, which has much lower portions of data compared to other quality. As a result, we decided to drop those rows, with the remaining of 6462 observations.

Table 1: Distribution of Wine Quality

Quality	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Count	30	216	2138	2836	1079	193	5
Percentage	**0.46%**	3.32%	32.91%	43.65%	16.61%	2.97%	**0.08%**

Through graphing boxplot of all variables (Appendix A), we noticed that there are some extreme outliers in the predictor variables density, residual sugar, and free sulfur dioxide. Since extreme outliers will have negative impact to the accuracy of model fitting and data prediction,

we used Q1 - 1.5* IQR and Q3 + 1.5 * IQR to calculate the lower and upper bound of the data, then remove the data that is out of the range. In the end, we have 6454 observations.

2.2 Descriptive Statistics

Among the eleven predictor variables, we used side by side boxplot for red and white wine and found that the mean and standard deviation for each predictor variable is quite different. By separating the red and white wine, we can build more accurate models that can take these differences of predictor variables into account.

3. Feature Selection

3.1 Correlation Plot

The following correlation matrices were produced to gain a better understanding of the potential variables with significant effects on the quality of red and white wine, respectively.

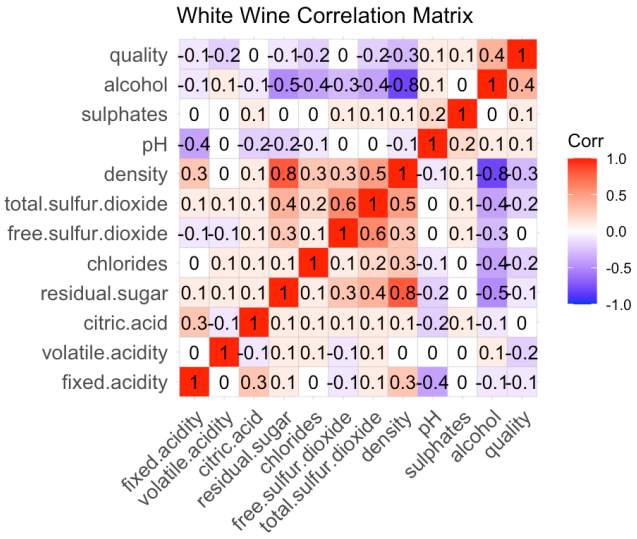


Figure 1: Whine Wine Correlation Matrix

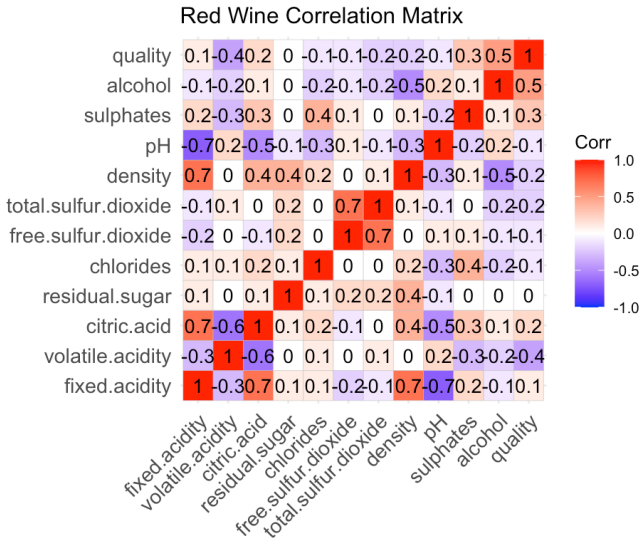


Figure 2: Red Wine Correlation Matrix

Within the white wine dataset, *alcohol*, *sulphates*, and *pH* are positively correlated with white wine quality, whereas *density*, *total sulfur dioxide*, *chlorides*, *residual sugar*, *volatile acidity*, and *fixed acidity* are negatively correlated. *Citric acid* and *free sulfur dioxide* show no correlation with quality. The most notable correlations present with the white wine quality are *alcohol* and *density*, which show correlations of 0.4 and -0.3, respectively. Some less notable ones include total *sulfur dioxide*, *chlorides*, and *volatile* which each show correlations of -0.2.

Within the red wine dataset, *alcohol*, *sulphates*, *citric acid*, and *fixed acidity* are positively correlated with red wine quality whereas *pH*, *density*, *total sulfur dioxide*, *free sulfur dioxide*, *chlorides*, and *volatile acidity* are negatively correlated. *Residual sugar* shows no correlation with red wine quality. The red wine appears to have stronger correlation among the variables compared to the white wine quality. Some notable correlations for red wine are *alcohol*, *volatile acidity*, and *sulphates*, exhibiting correlations of 0.5, 0.4, and 0.3, respectively. Less notable correlations include *citric acid*, *total sulfur dioxide*, and *density* which each show correlation so +/- 0.2.

Overall, it shows the trend that higher percentage of *alcohol* and *volatile acidity* in wine leads to better quality of wine.

3.2 Recursive Feature Elimination (RFE)

Although correlation matrices can detect potential linear relationships, it lacks in its ability to identify non-linear relationships. Due to this inability, we chose to use recursive feature elimination as a way to recognize some non-linear relationships. The RFE is a feature selection method recursively removing features by fitting a model, ranking features by importance, and eliminating the least important ones. Its advantage lies in the fact that it can indirectly relationships by identifying the features that contribute most to the model performance, regardless of whether their relationships with the target variables are linear or nonlinear.

By employing the RFE, we arrived at the following important features for red and white wine. We can see that only *alcohol* and *volatile acidity* are common predictor variables among red and white wine, which confirms the findings from the correlation plot.

Table 2: Important Feature for Red and White Wine

Red Wine	*alcohol, volatile acidity, total sulfur dioxide, density, sulphates*
White wine	*alcohol, volatile acidity, free sulfur dioxide, pH, residual sugar*

4. Machine Learning Models

4.1 Cross Validation

With the cleaned dataset, the dataset is highly imbalanced with the 4865 rows of white wine and 1589 rows of red wine. Therefore, we choose the cross validation method instead of

purely train-test split in all models to prevent the overfitting issue and to help create a more robust model.

We choose the parameter of *k-folds* = 5 for the cross validation process. For the consistency of the cross validation method across different models, we define the parameter *trControl* by *method = "cv"*, *number = 5*, and apply the parameter trControl in different models to make sure all models that need cross validation have the same procedure. We also set the seed = 123 to ensure the reproducibility of the code.

In addition, we use the same selected predictor variables across different models to ensure the consistency of model evaluation and comparison.

4.2 Multiple Linear Regression

To start our analysis from scratch, we want to build a simple model before creating other more flexible and advanced models for further analysis. In this case, a multiple linear regression model serves as a great benchmark model that allows us to predict wine quality given the selected predictor variables.

The red wine model has the following form with the MSE value of 0.41:

Quality = -5.47740 + (0.01555 * *density*) − (0.18573 * *volatile.acidity*) -(0.08406* *total.sulfur.dioxide*) + (0.11909 * *sulphates*) + (0.31971 * *alcohol*)

The white wine model has the following form with the MSE value of 0.54:

Quilty = 1.09807 + (0.04758 * *pH*) - (0.19748 * *volatile.acidity*) + (0.11876 * *residual.sugar*) + (0.08864 * *free.sulfur.dioxide*) + (0.46659 * *alcohol*)

From the summary of the model, we can see that all the coefficients of the predictors are significant, which confirms that there is a relative linear relationship between the predictor variables and the response quality.

4.3 Regression Tree

The linear assumption in the multiple linear regression model might severely reduce the accuracy of our prediction if the linear assumption between the wine quality and predictor variables does not hold true. In the second model, we relax the linearity assumption and use a regression tree model to handle the non-linear relationships and feature interactions.

We chose cp=0.0124 and 0.0097 for red and white wine. ‘Cp’ stands for Complexity Parameter of the tree, and we choose the cp value for low error rate(around 0.68) according to the Appendix C.

According to the Appendix D, The regression result is quite straightforward in identifying the wine quality for both types. the MSE for redwine model is 0.49, and the MSE for white wine model is 0.58. For both white and red wine, *alcohol* is the most important identifier. The higher alcohol the wine contains, and the higher quality the wine is. The result aligns with the following random forest model.

For the second layer, *sulphate* is the major identifier for red wine. Any red wine with more than 9.871 alcohol, the *sulphate* larger than 3.80 and *volatile.acidity* larger than 2.82 has the wine quality larger than 6. For red wine with *alcohol* less than 9.87, the *volatile.acidity* larger than 1.91, the quality could still be 6 or higher. Other than that, the wine quality is around 5. For White wine, other than *alcohol* and *volatile.acidity*, *residual.sugar* is an important indicator to identify the wine with 8.82 alcohol. If the *residual.sugar* is larger than 2.48, then the white wine quality could be 6 or more.

4.4 Random Forest

Another model we implement is random forest, because it is good at handling the high-dimensional data, capturing non-linear relationships that could not be found from multiple linear regression models, and it could also provide feature importance rankings. Our wine dataset contains multiple features that may interact in complex ways to determine the wine class, so we think random forest can be another good option.

Random Forest has two types of models, the random forest classifier is for the categorical outcome, and the random forest regressor is for the linear outcome. In our case, we have the outcome variable from 3 to 9, the random forest classifier is not doing well in handling 7 categorical variables outcomes. Also, the evaluation metric of a random forest classifier is the Kappa value, and for the regressor is MSE. For the consistency in evaluating and comparing different models, we decided to use a random forest regressor for MSE value with better performance.

To test and get the best MSE result, we ran from “*mtry* = 2” to “*mtry* = 11”, and found out that “*mtry* = 2” has the lowest MSE score for both white wine and red wine dataset. Therefore,

we set "*mtry* = 2" to test over the variables. We also set the number of trees as 500. According to Appendix E, we found out that the larger number of trees, the lower error rate the model has. But both models' error rate remains when the number of trees is high enough. As a result, we found the model for white wine with best performance is 485 trees, for red wine is 449 trees.

One of the significant advantages of Random Forest is the ability to provide feature importance rankings. By examining these rankings, we identified the most influential features for predicting the wine class. We discovered that attributes such as *alcohol* content, *volatile acidity*, and *total sulfur dioxide* played crucial roles in determining the wine's classification. According to the Appendix F, for white wine, we get the importance ranking: *Alcohol* > *volatile.acidity* = *free.sulfur.dioxide* > *residual.sugar* > *pH*. For Red Wine, we get the ranking: *Alcohol* > *sulphates* > *volatile.acidity* > *total.sulfur.dioxide* > *density*. Therefore, our major takeaway would be that the ratio of alcohol is the most important variable to evaluate the wine quality. Secondly, for white wine, as a drink that is more sweet, residual sugar is included as one important variable.

The random forest is the model with slightly higher MSE compared to other models we have built so far. The MSE for White wine model is 0.59, and the MSE for the Red wine model is 0.54.

4.5 Gradient Boosting

In our final model, we tried another powerful machine learning model Gradient Boosting to predict wine quality. It works by combining the predictions of multiple weak learners from decision trees to build a stronger model with predictive performance.

We have 3 trees in total and every new tree added to the mix will learn from the mistakes of the previous models and try not to repeat them. Eventually, it will turn a weak learner into a strong learner, and make the model more accurate.

In both white wine and red wine models, we generated 150 trees and the shrinkage parameter lambda is 0.1. The interaction depth d is 3, so each tree is a small tree with 3 splits. The model outcome results for both red wine and white wine are as follows:

	var <chr>	rel.inf <dbl>
alcohol	alcohol	47.619919
volatile.acidity	volatile.acidity	22.950685
free.sulfur.dioxide	free.sulfur.dioxide	14.242893
residual.sugar	residual.sugar	9.519144
pH	pH	5.667359

Figure 3: Relative Importance for White Wine

For the white wine, we used *alcohol*, *volatile acidity*, *free sulfur dioxide*, *pH*, and *residual sugar* to train the model using CV and got the MSE around 0.46. The most important variable would be *alcohol*, while the least important variable would be *pH*.

	var <chr>	rel.inf <dbl>
alcohol	alcohol	40.127569
volatile.acidity	volatile.acidity	19.814810
sulphates	sulphates	19.075515
total.sulfur.dioxide	total.sulfur.dioxide	12.395155
density	density	8.586951

Figure 4: Relative Importance for Red Wine

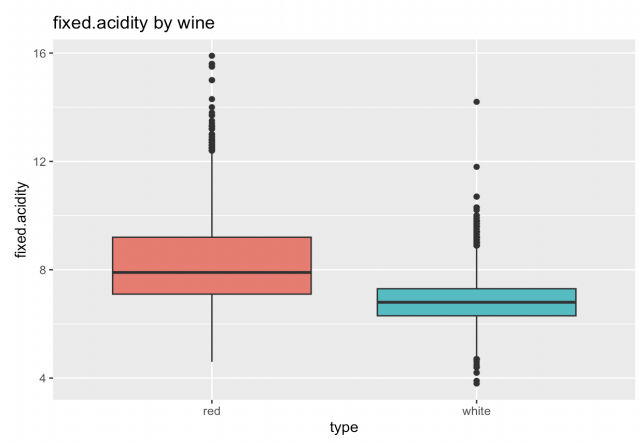
For red wine, we used *alcohol*, *volatile acidity*, *total sulfur dioxide*, *density*, and *sulphates* to train the model using CV, and got the MSE around 0.37, which is lower than the white wine. The most important variable would be *alcohol*, while the least important variable would be *density*. Thus, we can see that *alcohol* is an important evaluation metric for assessing the wine quality across different types of wine. Gradient Boosting model produced by far the smallest MSE for both red and white wine, outperformed multiple linear regression, regression trees, and random forest. The ability to capture complex relationships and interactions between the predictor variables and response quality, leading to improved predictive performance.

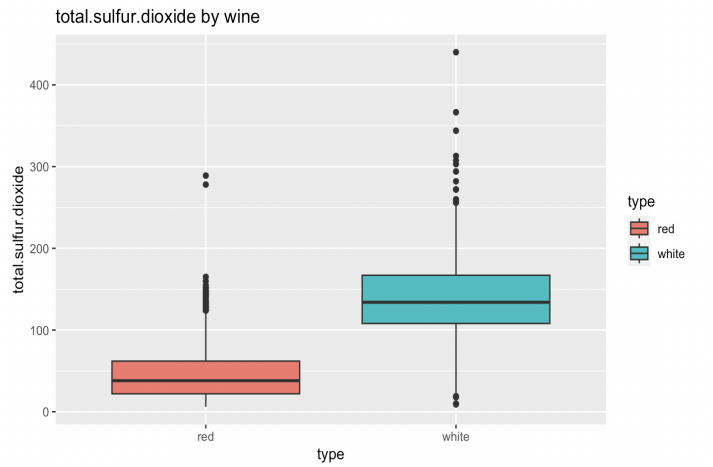
5. Conclusions

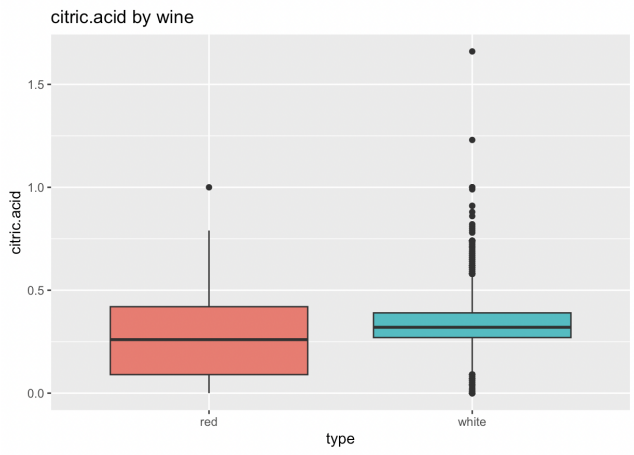
Splitting the wine dataset into red and white wine is necessary to build machine learning models as they have distinct predictor variables based on the correlation plot and recursive feature elimination methods. Among four machine learning models we implemented, we use the same predictor variables across the models so that we can keep consistency of model evaluation. The Regression Tree and Random Forest model has higher MSE among four machine learning models. Although the multiple linear regression model has slightly lower MSE, but it was quite an unstable model between white and red wine data. Lastly, the gradient boosting model delivered the lowest MSE and it is preferred due to its robustness and learned from the previous building model.

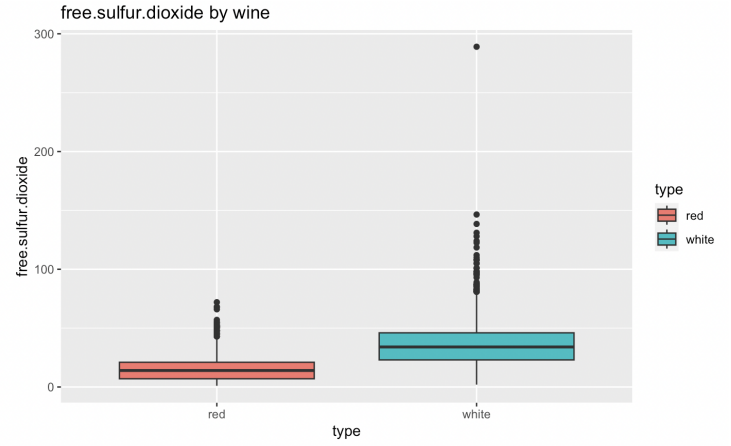
Appendix

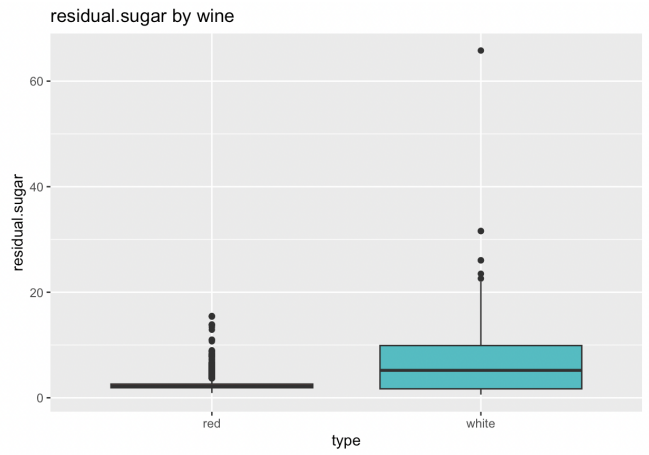
Appendix A. Boxplot of all variables

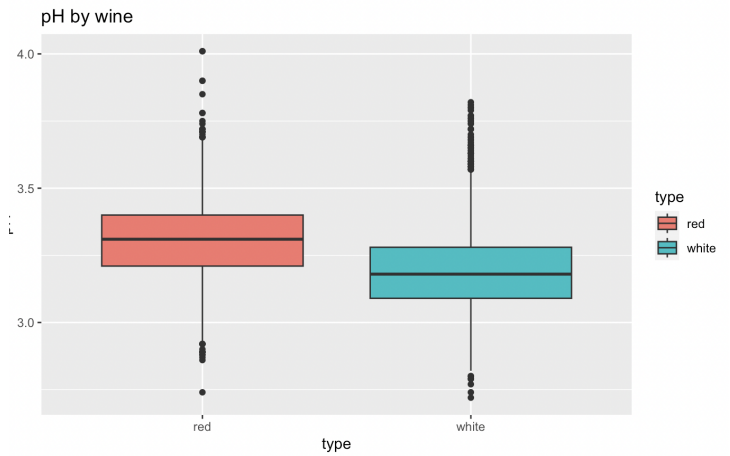


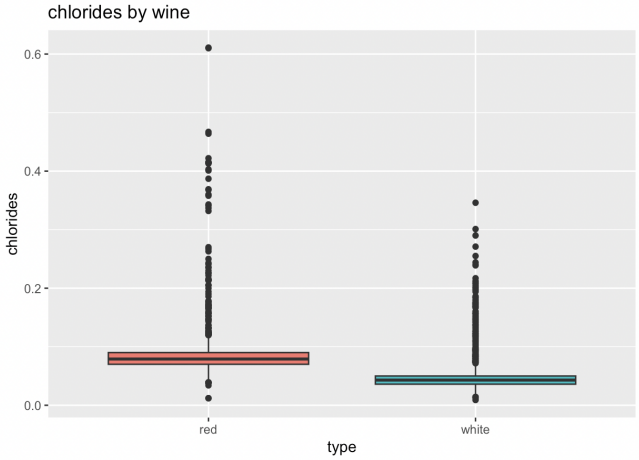


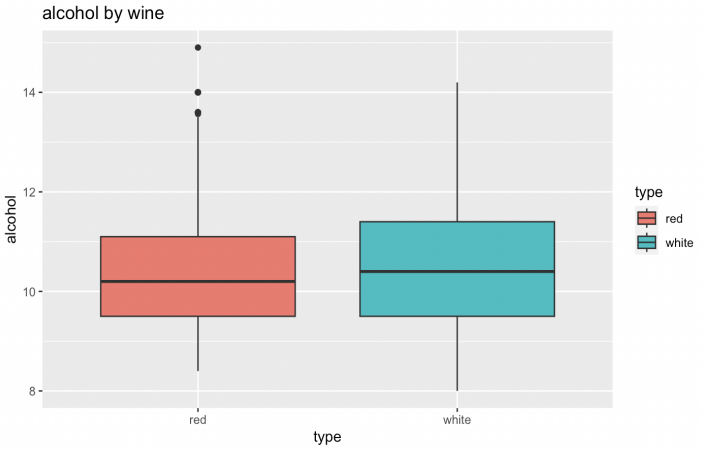


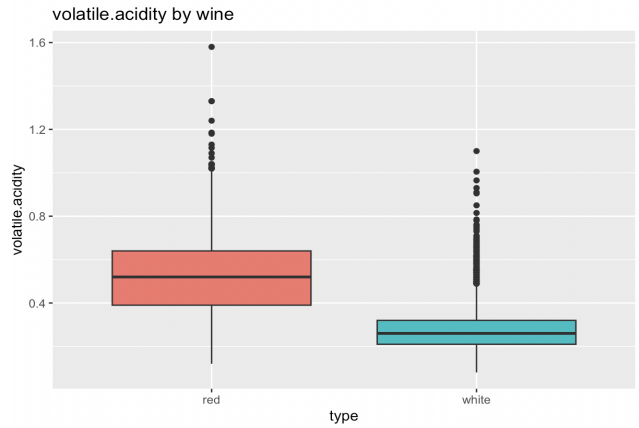


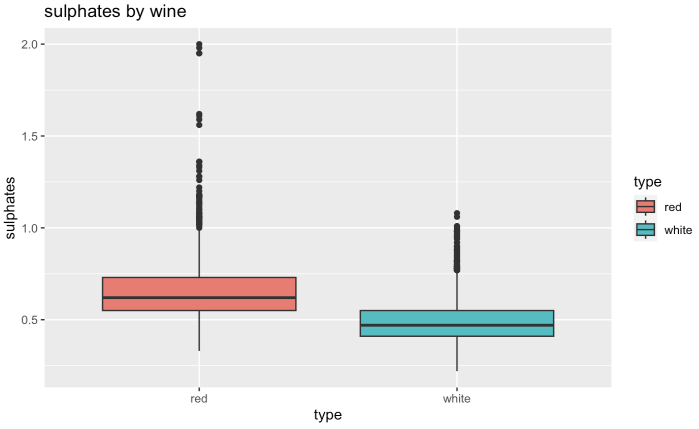

residual.sugar by wine
residual.sugar
60
40
20
0
red
white
type

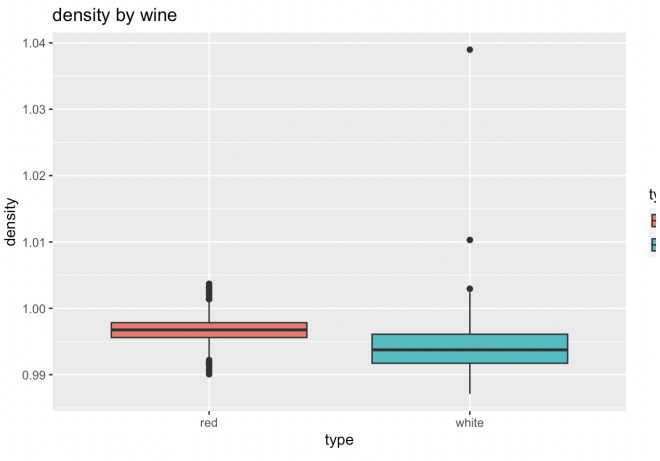

pH by wine
4.0
3.5
3.0
red
white
type
type
red
white


chlorides by wine
chlorides
0.6
0.4
0.2
0.0
red
white
type

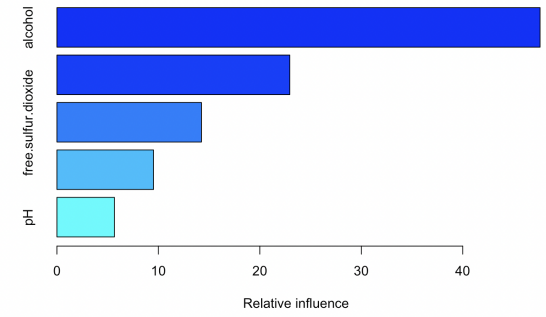

alcohol by wine
alcohol
14
12
10
8
red
white
type
type
red
white


volatile.acidity by wine
volatile.acidity
1.6
1.2
0.8
0.4
red
white
type

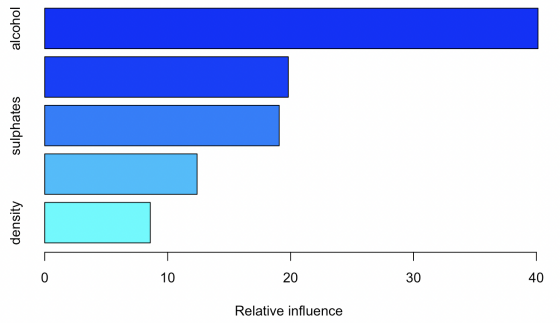

sulphates by wine
sulphates
2.0
1.5
1.0
0.5
red
white
type
type
red
white



Appendix B. Gradient Boosting - Variable importance plot

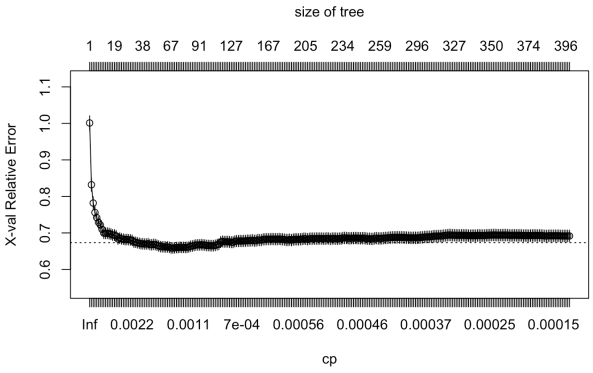


White Wine

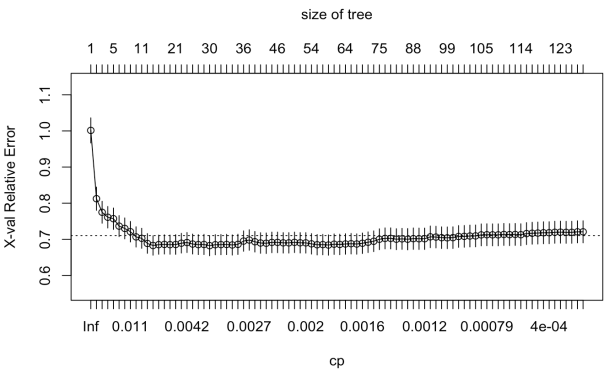


Red Wine

Appendix C. Regression Tree: Cp value VS Error rate

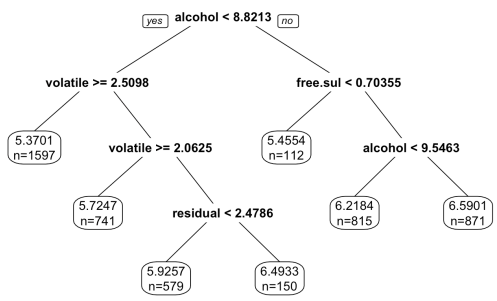


White Wine

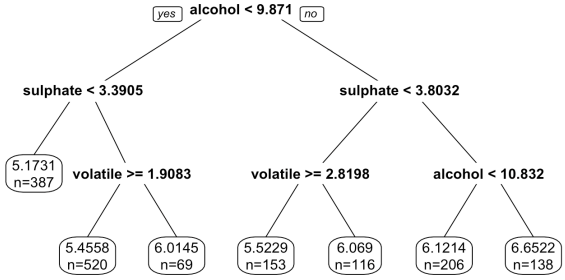


Red wine

Appendix D. Regression Tree for White Wine and Red Wine

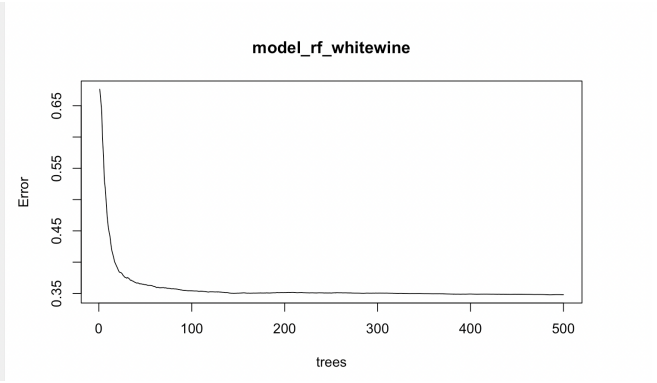


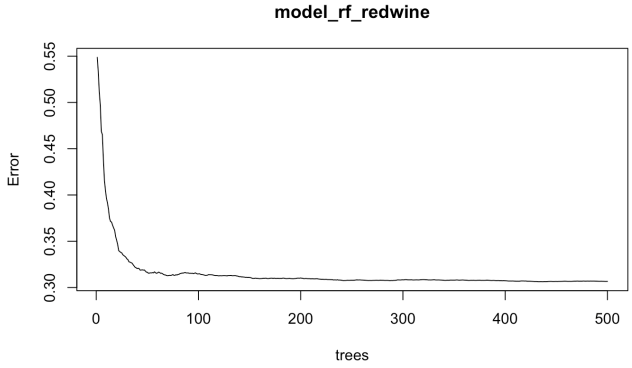
White Wine



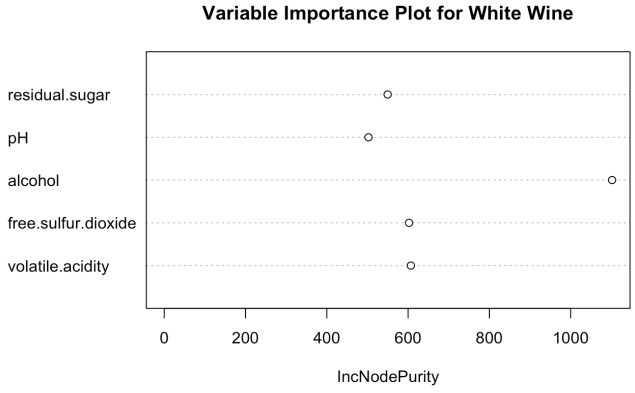
Red wine

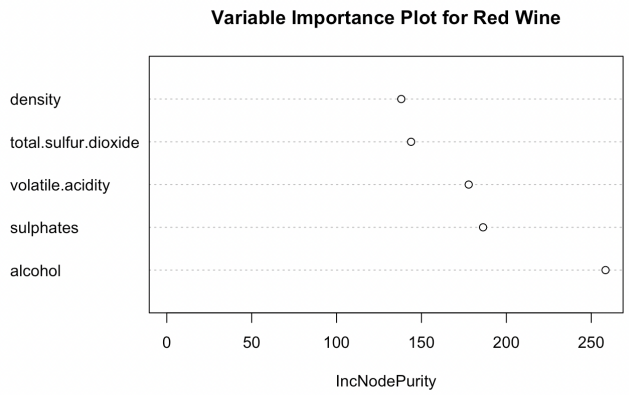
Appendix E. The number of Tree VS Error Rate For White Wine and Red Wine





Appendix F. The variable importance plot for White Wine and Red Wine





Reference

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.
https://archive.ics.uci.edu/ml/datasets/wine+quality

[2] Kniazieva, Yuliia. “Sommelier of the Digital Age: Machine Learning for Wine Quality Prediction.” Label Your Data, 15 February 2022,
https://labelyourdata.com/articles/machine-learning-for-wine-quality-prediction