Reviewing the Quality of Portuguese Wine based on Chemical Factors

Ottilie Mitchell - 20318500

25/11/2020

Background

Europe has the most established wine market and remains the world's consumption centre, consuming more than 50% of the total volume of wine (Elfman, 2020). Portuguese wine has become increasingly popular and is now the world's 9th largest wine exporter (ViniPortugal, 2020). Wine is considered a "luxury" good and carries certain connotation of success in business and for pleasure. Determining wine quality is therefore very important.

Wine quality is a subjective factor based on personal taste and is particularly difficult to pinpoint. Marketing academics have led to a focus on perceived quality rather than actual quality (Oude Ophuis & van Trijp, 1995). Perceived quality is the "customers' perception of the overall quality of a product with respect to its intended purpose, relative to alternatives" (Aaker, 1991). For the purpose of this study, the quality of the wine is based on experts' opinions, scored on a scale 0 (very bad) – 10 (very good).

Many studies have been undertaken which investigate wine quality based on the relationship to purchase intent (Botonaki & Tsakiridou, 2004), but not considering the chemical factors.

Wine is made up of thousands of compounds which originate from the grapes they are made from and from the wine making process. The process and grapes differ in the production of red and white wines. Red wine is made from dark red or black grapes and the skins are kept on during the fermentation process. White wine is primarily made with white grapes and the skins are removed before fermentation. These differences during the making process affects the chemical components of the wine. (Puckette, 2017)

The present report considers wine quality from an expert's perspective and specifically examines whether we can predict wine quality based on chemical factors. The report will also study if there is a significant difference between chemical components in red and white wine.

Objectives

- Predict the quality of wine and determine the best model
- Determine which chemical components influence wine quality
- Are the chemical components affecting wine quality the same in red and white wine?

Methods

Data set

The data set is made up of 6498 rows, each of which represent a different Portuguese wine, and 13 columns which include the wine type, chemical factors of each wine and the quality.

The variable names are:
1. Type – wine type, either red or white
2. Fixed acidity - acids involved with wine that do not evaporate readily (g/dmˆ3)
3. Volatile acid - the amount of acetic acid, related to creating a vinegar taste (g/dmˆ3)
4. Citric acid - citric acid found I small quantities can add freshness (g/dmˆ3)
5. Residual sugar - the amount of sugar remaining after fermentation stops (g/dmˆ3)
6. Chlorides - the amount of salt in wine (g/dmˆ3)
7. Free sulfur dioxide - the free form of SO2 exists (g/dmˆ3)
8. Total sulfur dioxide - amount of free and bound forms of SO2 (g/dmˆ3)
9. Density - the density of wine is close to that of water depending on alcohol and sugar content (g/cmˆ3)
10. pH - describes how acidic (0) or alkaline (14) the wine is
11. Sulphates - additive to wine which can contribute to SO2 levels (g/dmˆ3)
12. Alcohol - the amount of alcohol in wine (% by volume)
13. Quality - the quality of wine on a scale from 0-10

The full data set can be found in my GitHub page: https://github.com/OttilieM/Red-Wine

Data preparation and exploratory data analysis

It is important to gain an understanding of the data itself and to test hypothese. This involves examining numerical summaries and graphing data. It also gives the opportunities to highlight any outliers or missing values that may affect further analysis of the data. Data was processed initially to see if there were any missing values or zero values.

The data was split by type (red and white) to allow for further analysis between them.

Histograms are the best way for showing the spread of data and is easily compared between the types of wine. Boxplots were also explored for displaying the spread of data, however, they did not show the results as clearly.

Correlograms are graphs which summaries the correlations between each variable by creating a correlation matrix. They are very useful to highlight the most correlated variables, which are determined by colour. In this case we were interested in how variables correlated with quality. (Friendly, 2002) Corrographs were done using the "corrplot" package in R (Kuhn, 2008).

Training data and test data were created using a 70%:30% split.

Calculating success rate

For each model the Root Mean Square Error (RMSE) was calculated to give a standardised result for each model. The RMSE calculates the average magnitude of the errors in a set of forecasts (Chai & Draxler, 2014). In this data set, it shows how inaccurate the quality result of each wine was.

Confusion matrices were used to describe the performance of models. They show how well models perform using test data. Confusion matrices are easy to understand when using a binary classifier, therefore, they could not be used to show the results of the linear models.

The Akaike Information Criterion (AIC) is another method for evaluating how well a model fits the data. It uses the number of independent variables to build a model, and how well the model reproduced the data (Bevans, 2020). The smaller the AIC value, the better the model fits. AIC was used on linear and logistic regression results to see which performed the best.

Stepwise selection

Running models with irrelevant variables can lead to more complex functions. Stepwise selection is a way of looking at each variable to see how significant it is on the outcome variable, in this case wine quality.

Regsubsets() function can be used to perform forward and backward stepwise selection. This is useful to see which variables produce the best subset and therefore are more significant on the outcome variable.

Forwards stepwise selection starts with a model containing no variables and adds them to the model starting with the most significant. Variables are added until either all the variables are used, or until a pre-specified stopping rule is reached.

Backwards stepwise selection starts with all the variables in a model and removes them starting with the least significant variable. Variables are removed until there are no variables left, or until a pre-specified rule is reached.

Backwards stepwise is a better method when the variables are correlated with each other because unlike forward stepwise selection, all of the variables may be considered. Through exploratory data analysis it showed that some of the predictor variables were correlated, so therefore backward stepwise selection is the better method for this data set. (Choueiry, 2020)

Cross Validation

K-fold cross validation is a method used to validate subset selection. It uses test and training data to ensure an accurate estimate of which model of a given size is best. Firstly, a model is fitted only using training data. Subset selection is performed with each of the k training sets (k=10) and a matrix is created to store the results in. The predict() function cannot be used for regsubset(), so we must create our own "predict" formula. Finally, a loop is used to perform cross-validation, where predictions are made for each model size and then stored in the matrix. The mean errors that the cross validation found can then be plotted on a graph to clearly show which number of variables give the best model.

Although Subset selection and Cross validation are useful models to gain a better understanding of the significance of each variable, and potentially the best subset, all the variables were kept in the main dataset to run each model.

Linear regression

Linear regression is the simplest approach to supervised learning. It assumes that the dependence of the outcome variable on the input variables is linear; which is rarely the case. Although it is simplistic and not always very accurate, it can be useful to get a basic understanding of the relationships within the data. A more useful form of linear regression is multiple linear regression, where more than one predictor variable can be tested at one time. This was the best linear regression model, but only produced 53.69 mean standard error (MSE), showing that linear regression didn't fit our data very well.

The outcome variable quality is an integer, while the predictor values are continuous. Therefore, predictions made by a linear regression model will not be an integer, making it difficult to explain the variation in quality using a linear model.

To overcome this problem of predicting a single value, a new variable "rating" was created. This is a bivariable splitting quality into "good and "bad" wines based on quality being 6 or above for good and below 6 for bad. The test and train datasets needed to be updated to include this new variable.

Creating the "rating" variable enabled predictions of the classification of the wines using logistic regression and decision trees.

Logistic regression

This model is suitable to find a relationship between a categorical or bivariate variable and predictor variables. In this study, the relationship is between the chemical factors and the quality of the wine.

Simple logistic regression uses just one variable to predict the outcome variable, this in effect, is the same as linear regression. The difference between linear and logistic regression becomes obvious when multiple variables are involved in predicting an outcome variable. Logistic regression works by applying the "maximum likelihood"(ML) estimation method. This is where the variables for which the probability of the observed data is at its maximum is identified (Melesse et al., 2016). Once a logistic model is created, this can then be tested on unseen test data.

Multiple logistic regression was the best form of logistic regression for the wine dataset, including all of the predictor variables. This model gave an accuracy rate of 82.16% and RMSE 0.485. Although this method produced a very good result, decision trees were used to see if they could be improved further.

Decision trees

Decision trees are a non-parametric classifier that doesn't need any statistical assumptions regarding the distribution of data (Otukei and Blasche, 2009). They are known to produce more accurate results than other predictor models, but the performance of them can be improved further by pruning, boosting, bagging and random forests (Mahesh & Mather, 2003).

The basic structure of decision trees starts with one root node, a number of internal nodes and a set of terminal nodes. The nodes are determined based on the significance of each variable, stating with the most significant. From each node, data is split based on a decision and proceeds to the next node, until they reach a terminal node. Firstly, the "tree" package was loaded which is used for producing both classification and regression trees. To simplify the process, new datasets were created removing type and quality variables. New test and train datasets also need to be created with the new "tree data".

Pruning Pruning can be undertaken to remove any unnecessary variables, to simplify the tree and can improve the error rate.

Bagging and RandomForest Bagging uses the randomForest package. Bagging essentially splits the training data into numerous sets (normally between 50-500) to create multiple trees. The trees are then culminated into an aggregated prediction (Breiman, 1996).

This method is useful as it is designed to increase stability and accuracy of algorithms. As it averages sets of data, it helps to reduce variance and minimise overfitting (Boehmke & Greenwell, 2020).

RandomForest works in exactly the same way as bagging, however, are modified to use the optimal number of variables. For classification trees, the optimal number of variables is given by sqrt(k).

Boosting Boosting algorithms use the "gbm" package in r. They differ to random forest, as it creates a number of "shallow" trees in sequence, learning and improving on the previous one. Boosting works best on models which have a high bias and low variance (Boehmke & Greenwell, 2020). Therefore, although boosting is a powerful tool, in this instance it did not improve on previous models as the dataset had low bias and high variance.

Importance function Importance plots are a really clear way of showing which variables are the most significant to predicting the outcome variable. It produces two plots, the first shows the mean decrease in accuracy and the second the mean decrease in Gini. Mean decrease in accuracy measures the accuracy decrease when the variable is not included. Mean decrease in Gini measures the decrease in the Gini index when the variable is removed. The Gini plot uses the training data, so therefore the mean decrease in accuracy is the more accurate plot.

Finally, decision trees were created separately for red and white wine to create Importance plots to see if the importance of variables differed between wine types.

Results

Which model predicted wine the most accurately?

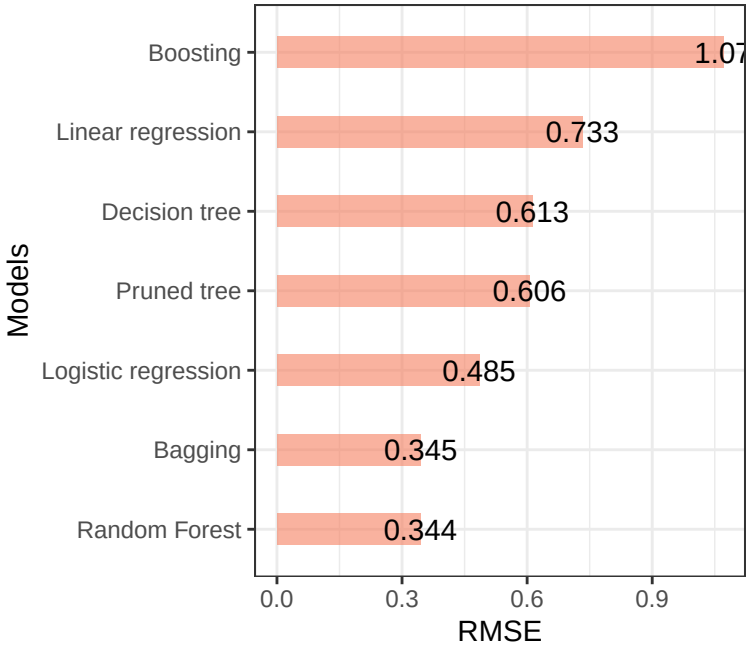

Figure 1: Accuracy of all models using RMSE

Having tested many models on my data, random forest produced the most accurate results when trying to predict the quality of wines. The model considered all the chemical factors of wine to determine wine quality. As you can see from figure 1, Random forest gave the lowest root mean squared error of 0.344. This indicates that this method was only mispredicting quality of wines by 0.344. After performing a confusion matrix on my test and training data, RandomForest was giving an accuracy rate of 88.1% - this accuracy rating was the best from all of the models tried.

Which variables are most important for predicting wine quality?

An Importance plot was used to show which variables are the most important for determining wine quality (figure 2).

Alcohol is clearly the most important variable for determining wine quality, showing 85 mean decrease in accuracy. Volatile acidity and residual sugar are the next most significant giving a decrease of 63 and 60 mean accuracy respectively. Total sulfur dioxide was the least significant variable for determining wine quality, decreasing the mean accuracy by just 41.

Are the chemical components affecting wine quality the same in red and white wine?

Histograms initially showed that the distribution of wine quality is similar in red and white wines. White wine quality has a gaussian skew, whereas red wine quality showed a very slight positive skew. No wines scored above 9 and lower than 3. (Figure 3 & Figure 4)

To investigate this further a corrograph was plotted for white and red wine to see if there were any obvious differences in correlation. (Figure 5 & Figure 6)

We can clearly see that there were some differences in correlation between quality and the predictor variables, between the two graphs.

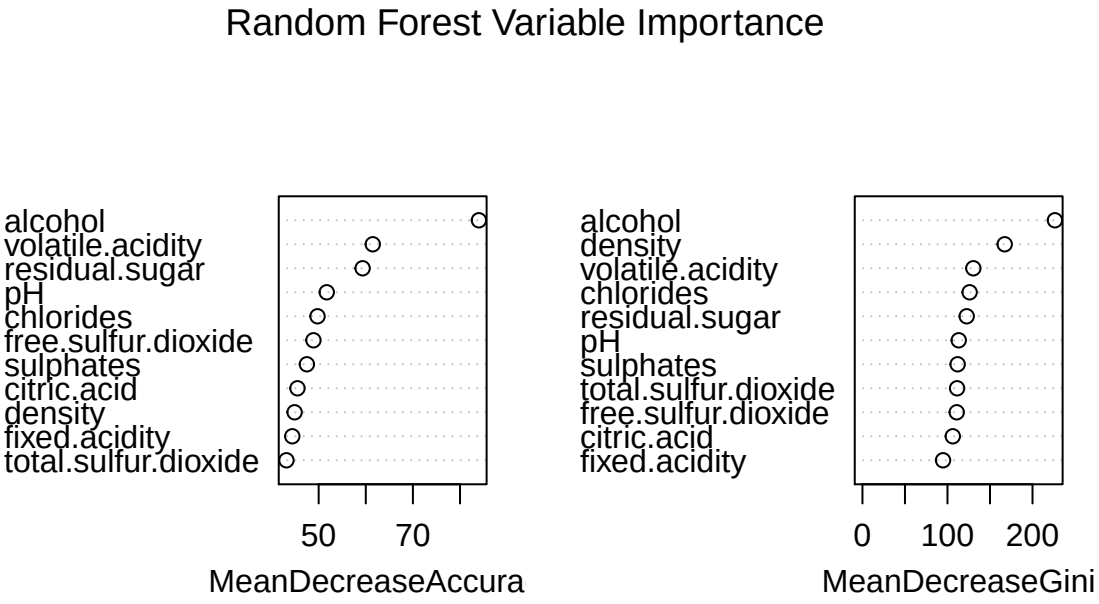

Figure 2: Variable Importance for Quality

Figure 3: White Wine Quality Distribution

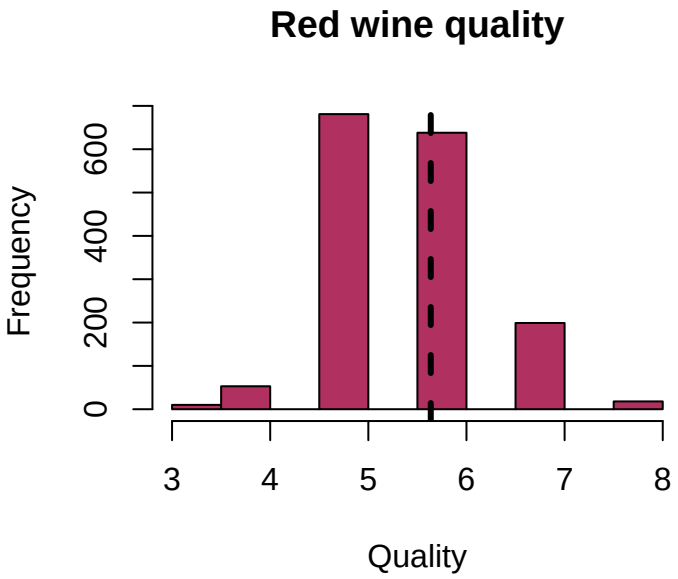

Figure 4: Red Wine Quality Distribution

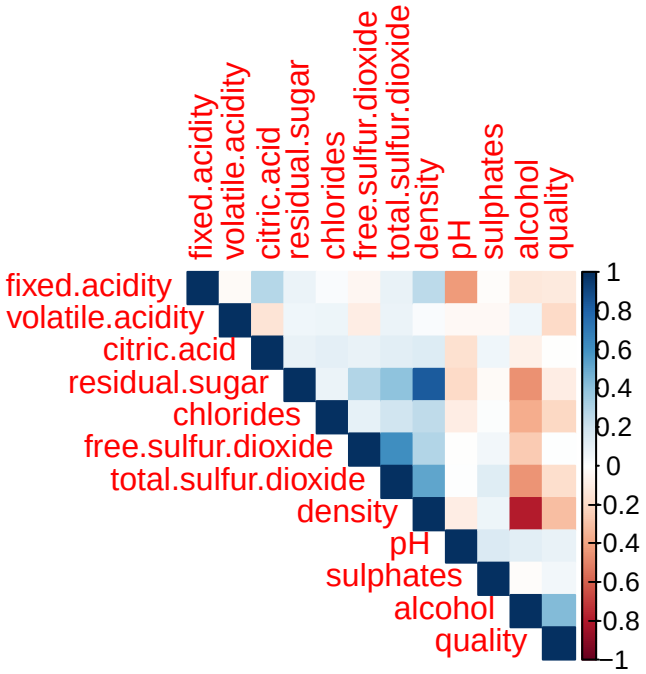

Figure 5: Correlations in White Wine

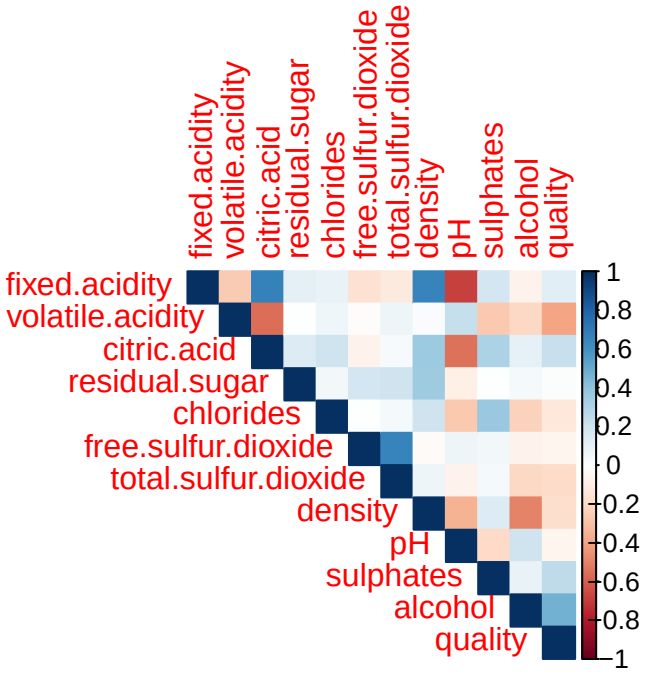

Figure 6: Correlations in Red Wine

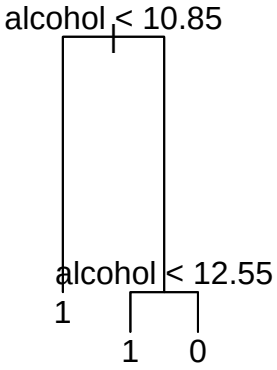

Figure 7: White Wine Tree

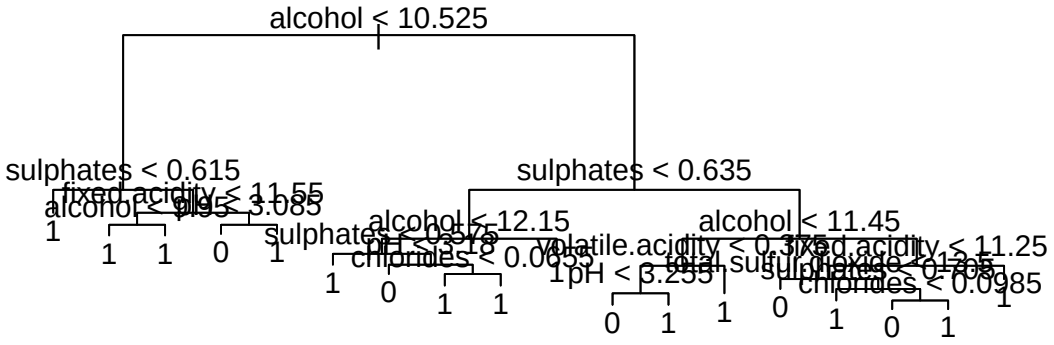

Figure 8: Red Wine Tree

The decision trees are clearly quite different between red and white wines, although are both still showing that alcohol is the most important variable for predicting quality. (Figure 7 & Figure 8)

Both of the importance plots also highlight how important the alcohol variable is in predicting wine quality, however, there are differences. (Figure 9 & Figure 10)

Firstly, we can see that the scale is different between the plots, alcohol is more important in white wine than red wine. Another major difference is that free sulfur dioxide and pH are important variables for white wine, and yet are the least significant variables for red wine.

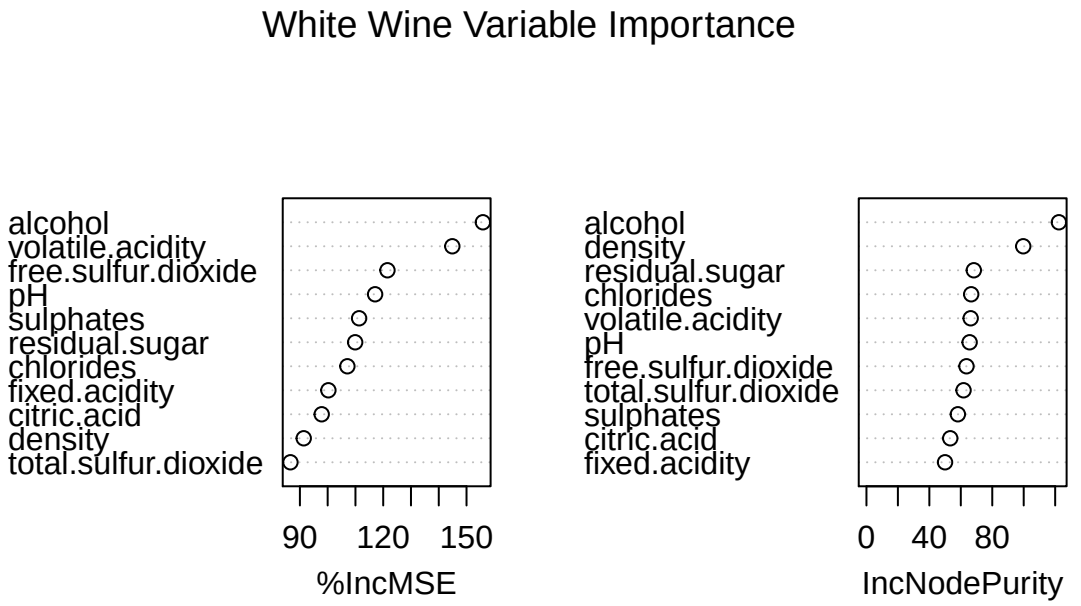

Figure 9: Variable Importance for White Wine

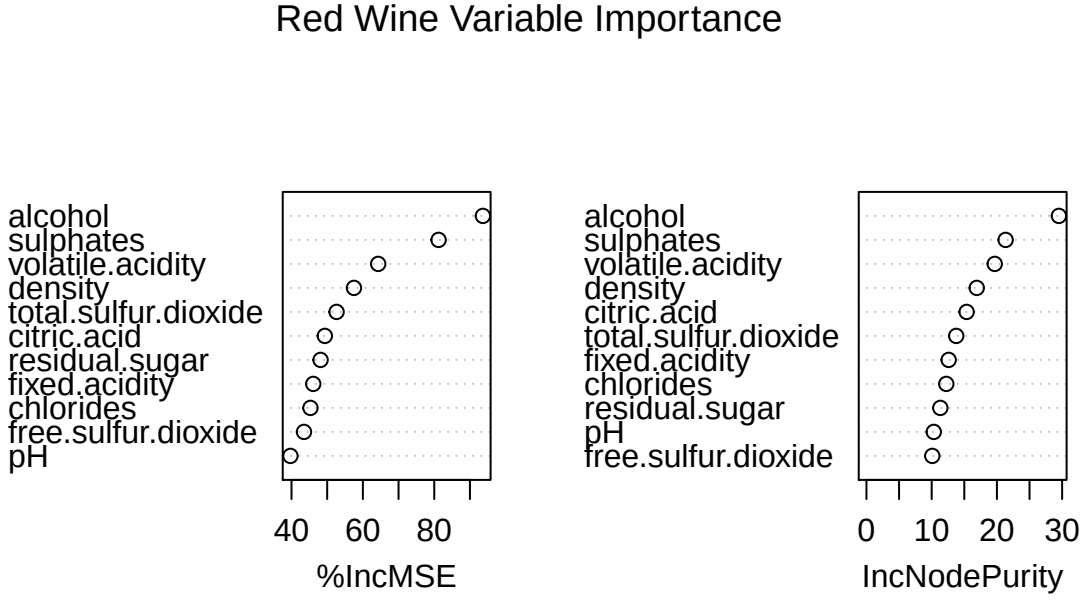

Figure 10: Variable Importance for Red Wine

Conclusion

Chemical factors of wine give sufficient information to accuracy predict wine through RandomForest. Alcohol was the most significant variable that influenced wine quality and total sulphur dioxide was the least significant. There are differences in which chemical factors affect white and red wine quality, although alcohol is still the most significant factor.

References

Aaker, D. (1991). Managing Brand Equity. California: Free Press

Bevans, R. (2020). *An intoduction to the Akaike information criterion.* Scribbr. Available at: https://www.scribbr.com/statistics/akaike-information-criterion/ (Accessed: 21th Nov. 2020)

Boehmke, B. & Greenwell, B.M. (2020). Hands-On Machine Learning with R. Florida: CRC Press

Botonaki, A. & Tsakiridou, E. (2004). Consumer response evaluation of a Greek quality wine. *Acta Agricola Scandinavia, Section C, Food Economics*, 1, 91–98.

Breiman, L. 1996a. Bagging Predictors. Machine Learning 24 (2). Springer: 123–40.

Chai, T. & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? - Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), pp. 1247-1250.

Choueiry, G. (2020). *Understanding Forward and backward Stepwise Regression.* Quantifying Health. Available at: https://quantifyinghealth.com/stepwise-selection/ (Accessed: 19th Nov. 2020)

Elfman, Z. (2020). *Libation Frontiers: A deep dive into the world wine industry.* Available at: https://www.toptal.com/finance/market-sizing/wine-industry (Accessed: 26th Oct. 2020)

Friendly, M. (2002). Corrgrams. *The American Statistician*, 54(4), pp.316-324.

Melesse, S., Sobratee, N. & Workneh, T. (2016). Application of logistic regression statistical technique to evaluate tomato quality subjected to different pre- and post-harvest treatments. *Biological Agriculture & Horticulture*, 32:4, 277-287.

Otukei, J.R. & Blaschke, T. (2009). Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *International Journal of Applied Earth Observation and Geoinformation*, 12:1, 27-31.

Puckette, M. (2017). Red wine vs white wine: The real differences. *Wine Folly.* Available at: https://winefolly.com/tips/red-wine-vs-white-wine-the-real-differences/ (Accessed: 26th Oct. 2020)

ViniPortugal. (2020). The wine sector. Available at: https://www.viniportugal.pt/WineSector (Accessed: 26th Oct. 2020)