Wine Quality — Prediction and Comparison Using Decision Tree and Random Forest

1. Introduction

1.1. Brief Problem Description

The primary objective of this report was to use machine learning techniques for predicting sensory scores based on 11 distinct physicochemical variables. I aimed to emulate human sensory and assessment by employing advanced modelling approaches. Furthermore, I intended to conduct a comparative analysis of the performance and accuracy of two different modelling approaches across two subsets and the joint dataset. Additionally, I sought to delve deeper into understanding the underlying and relevant features that significantly affect taste scores. Through these efforts, I wish to gain comprehensive insights into the intricate relationship between physicochemical predictors and sensory scores to enhance my understanding of flavor perception mechanisms, which can also improve my comprehension and application of modelling.

1.2. Dataset introduction

The two datasets under consideration are red and white variants of the Portuguese "Vinho Verde" wine. These datasets present classification or regression challenges, whose ordered classes are not evenly distributed. For example, the dataset contains significantly more normal instances compared to excellent or poor ones. Outlier detection algorithms could be applied to explore the few excellent or poor wines. Additionally, there is uncertainty if all input variables are relevant (Cortez et al., 2009).
Number of Instances: red wine - 1599; white wine - 4898
Number of Attributes: 11 + output attribute
Attribute information:
Input variables (based on physicochemical tests):

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 – alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

Missing Attribute Values: None

1.3. Modelling approaches

Treating the target variable as a categorical variable, I decided to employ the classification tree and random forest to model and investigate the problem.

2. Data Cleaning and Exploratory Data Analysis

All variables from 1 to 11 are continuous and represented as double types. The 12th output variable whose type is integer can be treated as either a continuous variable or categorical variable. Here, I converted its type, initially an integer, to a factor. Additionally, I introduced another factor variable "type" to distinguish between red wine and white wine for merging the two datasets.

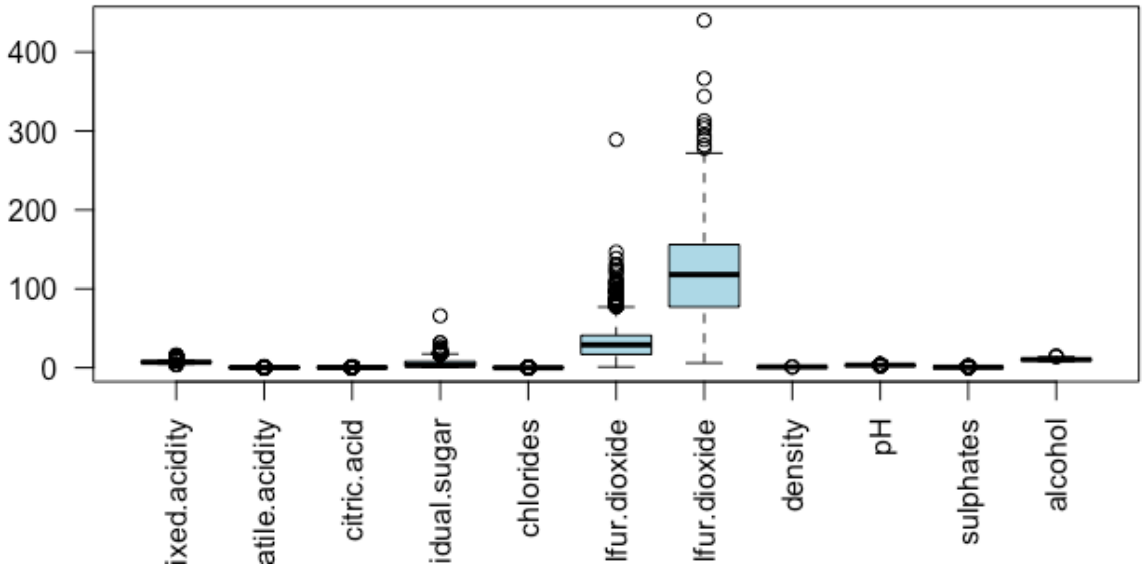


Figure 1. Boxplot of Physicochemical Predictors

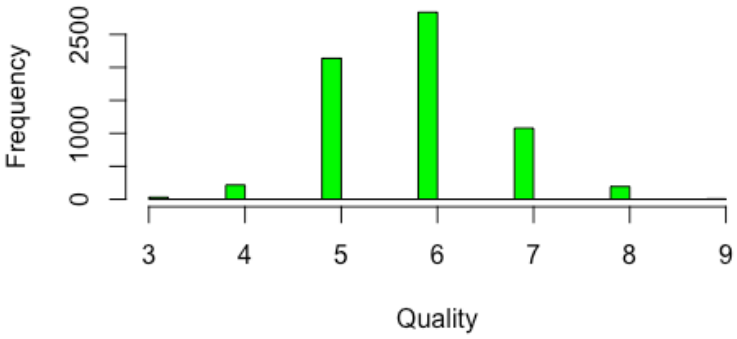
While the dataset has no missing values, it contains numerous outliers due to unbalanced classes. Deleting or transforming these outliers may result in the loss of valuable information. Since the models chosen are robust and have the ability to handle outliers, I planned to retain them, preserving more information to enhance the generalization of my model.

For the initial analysis, I merged the red wine dataset and the white wine datasets, introducing a new column labelled "type" to differentiate between the two varieties, focusing on two different model approaches comparison. Red wine is represented as 1 and white wine as 0. Subsequently, I respectively applied the modelling techniques to compare the features of the two wine types. Across all sub-datasets and joint datasets, I allocated 90% of the data to training data to fit the model and 10% to testing data to predict and test the model (See Appendix for details).

Moreover, the histogram revealed that the distribution of quality scores was skewed, with fewer occurrences observed for scores 3, 4, 8, and 9. For improving the accuracy of prediction, it would be beneficial to have more data representing these scores. I planned to discretize the quality value into three ranges to simplify and generalize my model (Actually, there are only 3~8 scores in the red wine dataset and 3~9 scores in the white wine dataset).

Figure 2. the Quality Distribution of All Wine Dataset

Bad: quality score is less than or equal to 5.
Medium: quality score is equal to 6.
Good: quality score is greater than or equal to 7.



Through the graph below, we can find the relationships between the response variable and predictor variables and the relationships between predictor variables.

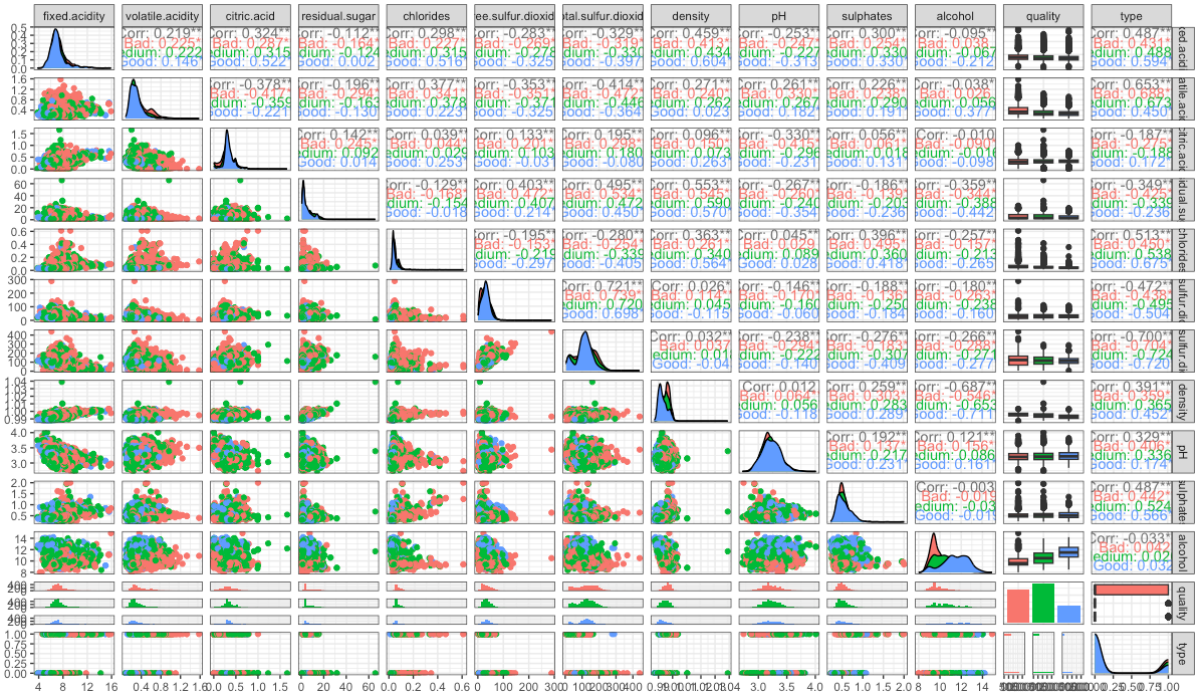


Figure 3. the Relationships among the Variables

3. Modelling

3.1 Classification Tree

Overview of Classification Tree

I utilized the "CART" library to construct the decision tree. "CART" serves as a general term describing various tree algorithms, including Breiman's CART algorithm. There are two primary categories of trees:

1. Classification Trees: These decision trees are employed for classification tasks and predict categorical responses.

2. Regression Trees: These decision trees are utilized to model numeric response variables using regression techniques.

Operation on Classification Tree

Here, I used a classification tree since I treated the quality variable as a categorical value. And I employed the caret library, which provides built-in cross-validation functionality. Cross-validation is crucial to prevent overfitting by optimizing the selection of leaf nodes in the decision tree. This process is akin to pruning, ensuring a more robust model. I set the 10-fold cross validation here. Initially, I constructed and evaluated the model using the merged dataset. Subsequently, I trained and tested the

two datasets separately. For now, I only talked about the merged data here, emphasizing the model comparison. Two sub-datasets and this merged dataset will be thoroughly compared in the model comparison section. After fitting the model by using the training dataset extracted from the origin dataset and feeding all 12 variables into the model (including the “type” feature), I then utilized the remaining testing data to predict and generate the results and the confusion matrix.

Results of Classification Tree

The final tree model for “*quality ~ .”* does the following:

1. It initially splits the training data into two subsets. Take all the points for which “alcohol” is greater than or equal to 10.625 and classify 13% of these points as “Bad”, 50% of them as “Medium” and 37% as “Good”. The remaining points are further used for the next split.
2. It divides all observations for which the “volatile.acidity” is less than 0.275. For those meeting this condition, it classifies 35% as “Bad”, 50% of them as “Medium”, 15% as “Good”. The remaining points are classified 63% as “Bad”, 35% of them as “Medium”, 2% as “Good”.

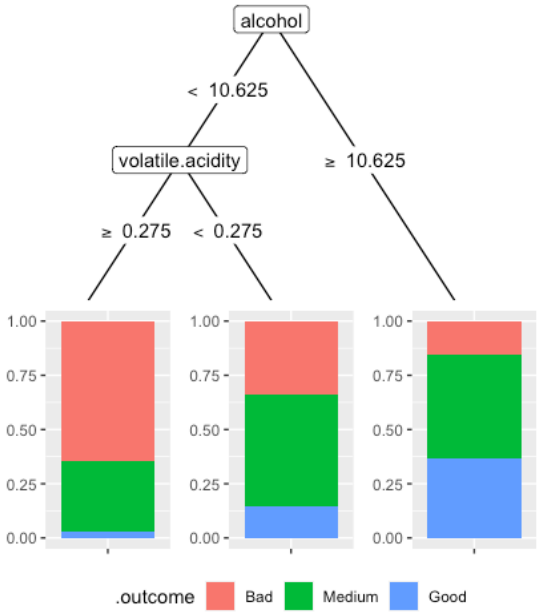


Figure 4. the plot of Classification Tree

From the CP table, it was evident that the variables “alcohol” and “volatile.acidity” played crucial roles in constructing the decision tree. The last model with 2 leaf nodes, indicated by the lowest relative error, did not require pruning. The confusion matrix provided insights into the corresponding relationship between prediction results and actual quality scores. The overall accuracy of 55.33% demonstrated that the model's performance is suboptimal. The variable importance table revealed that “alcohol”, “volatile.acidity” contributed more to this model, indicating that these two features significantly impact the quality score.

```
Variables actually used in tree construction:
[1] alcohol          volatile.acidity

Root node error: 3331/5897 = 0.56486

n= 5897

        CP nsplit rel error
1 0.110477      0   1.00000
2 0.080757      1   0.88952
3 0.014260      2   0.80877
```

```
rpart variable importance

                     Overall
alcohol              100.000
volatile.acidity      62.692
density               52.221
chlorides             46.381
citric.acid           33.488
free.sulfur.dioxide    7.059
residual.sugar         0.000
fixed.acidity          0.000
type1                  0.000
sulphates              0.000
pH                     0.000
total.sulfur.dioxide   0.000
```

```
Confusion Matrix and Statistics

          Reference
Prediction Bad Medium Good
    Bad    123     61    6
    Medium  86    209  115
    Good     0      0    0

Overall Statistics

               Accuracy : 0.5533
                 95% CI : (0.5125, 0.5936)
    No Information Rate : 0.45
    P-Value [Acc > NIR] : 2.445e-07

                  Kappa : 0.2328

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: Bad Class: Medium Class: Good
Sensitivity              0.5885        0.7741      0.0000
Specificity              0.8286        0.3909      1.0000
Pos Pred Value           0.6474        0.5098         NaN
Neg Pred Value           0.7902        0.6789      0.7983
Prevalence               0.3483        0.4500      0.2017
Detection Rate           0.2050        0.3483      0.0000
Detection Prevalence     0.3167        0.6833      0.0000
Balanced Accuracy        0.7086        0.5825      0.5000
```

Figure 5. CP Table, Confusion Matrix and Statistics, and Variable Importance Table of Classification Tree

3.2 Random Forest

Overview of Random Forest

Random forest is a widely used machine learning algorithm developed by Leo Breiman and Adele Cutler, known for its ability to combine the outputs of multiple decision trees to reach a single result. Renowned for its ease of use and flexibility, it is capable of handling both classification and regression problems. (IBM, 2024). Here, I used the random forest to tackle the classification problem.

Here is how to construct a random forest. A Random Forest is built by first obtaining a training dataset with observations and deciding on the number of trees for the forest. For each tree in the forest, a bootstrap sample of training of size should be taken. A tree starts to grow at random. For each split, a subset of features is randomly selected from all available features. The best split is picked by using the Gini Impurity This process continues until a stopping criterion is met (e.g. CP, entropy). Pruning is not needed. Ultimately, the ensemble of trees is returned, contracting the random forest.

Operation on Random Forest

Once again, I employed the **caret** library and applied the 10-fold cross-validation. Using the same approach for the classification tree, I fed the same training dataset and all 12 variables into the model (including type feature) to the random forest to ensure consistency in evaluation on later comparison with the performance of the classification tree model. Subsequently, I applied the fitted model to the same testing dataset to obtain the results.

Results of Random Forest

Looking at the results, I observed that the final value used for the model was mtry = 2, suggesting that only 2 variables out of the 12 predictors (including type) were used to achieve the best performance. When predicting the testing dataset, the confusion matrix showed that the total accuracy of the random forest model was 74.83%, showcasing satisfactory performance. Examining the variable importance, we can see that “alcohol”, “density” and “volatile.acidity” made significant contributions to the model, indicating that these three features have a more pronounced impact on the quality score.

```
Random Forest

5897 samples
  12 predictor
   3 classes: 'Bad', 'Medium', 'Good'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 5309, 5307, 5308, 5308, 5306, 5307, ...
Resampling results across tuning parameters:

  mtry  Accuracy   Kappa
   2    0.7307008  0.5696386
   7    0.7264644  0.5639946
  12    0.7218832  0.5572579

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was mtry = 2.
```

```
Confusion Matrix and Statistics

          Reference
Prediction Bad Medium Good
    Bad    164     46    2
    Medium  43    206   40
    Good     2     18   79

Overall Statistics

               Accuracy : 0.7483
                 95% CI : (0.7116, 0.7826)
    No Information Rate : 0.45
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.5986

 Mcnemar's Test P-Value : 0.03764

Statistics by Class:

                     Class: Bad Class: Medium Class: Good
Sensitivity              0.7847        0.7630      0.6529
Specificity              0.8772        0.7485      0.9582
Pos Pred Value           0.7736        0.7128      0.7980
Neg Pred Value           0.8840        0.7942      0.9162
Prevalence               0.3483        0.4500      0.2017
Detection Rate           0.2733        0.3433      0.1317
Detection Prevalence     0.3533        0.4817      0.1650
Balanced Accuracy        0.8310        0.7557      0.8056
```

```
rf variable importance

                      Overall
alcohol                100.00
density                 83.89
volatile.acidity        78.62
total.sulfur.dioxide    70.19
chlorides               69.02
free.sulfur.dioxide     67.40
sulphates               66.12
residual.sugar          64.90
citric.acid             63.94
pH                      63.77
fixed.acidity           58.24
type1                    0.00
```

Figure 6. Random Forest Table, Confusion Matrix and Statistics, and Variable Importance Table

4. Model Comparison

4.1 Vertical model comparison

Comparing the plots of the classification tree and random variable, we can observe a notable performance gap between the two models. The accuracy of the classification tree was 55.33%, whereas the accuracy of the random forest was 74.83%, as confirmed by Figures 9 and 10. There was a significant improvement from the classification tree to the random forest, especially in predicting the “Good” quality of wine. In addition, Figures 7 and 8 underscored that the variables “alcohol”, “density” and “volatile.acidity” consistently emerged as significant contributors to both models, verified again. However, Figures 7 (with only 2 colors on the plot) and 9 (where all instances of the "Good" class were marked with crosses) highlighted a notable weakness in the classification tree model, particularly in predicting the "Good" quality of the testing dataset, where it performed poorly.

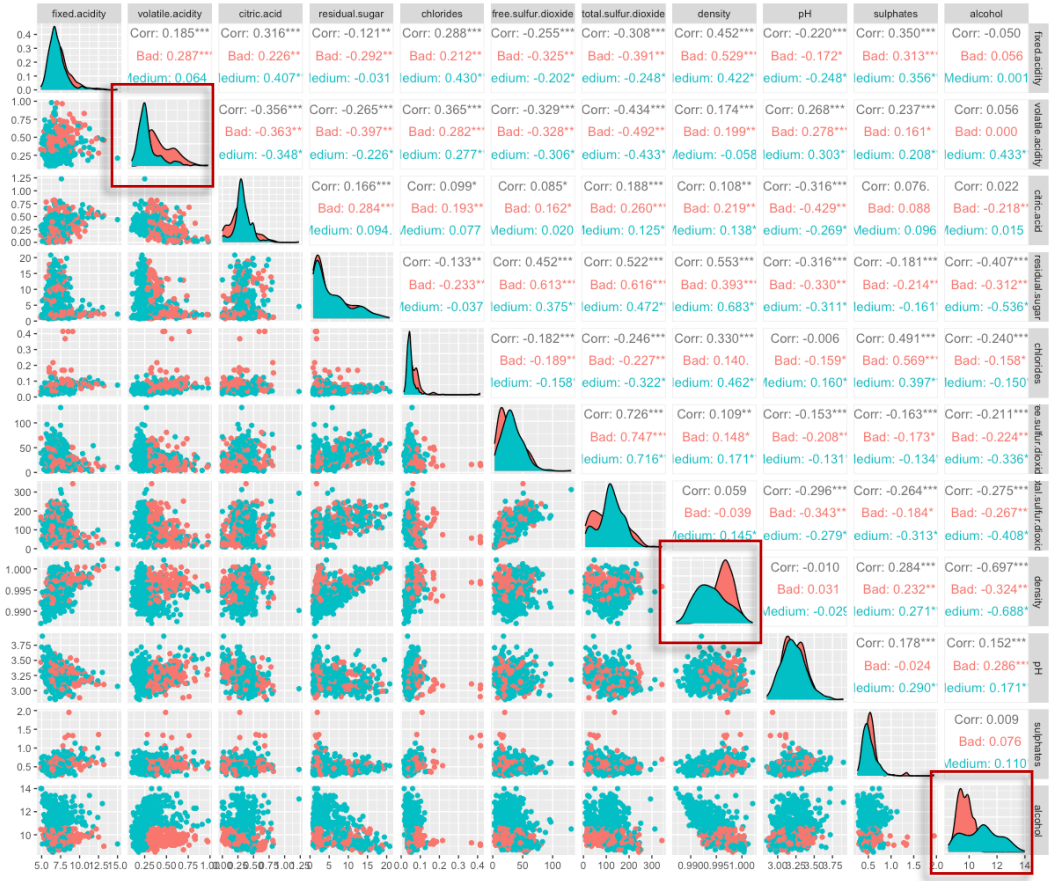


Figure 7. Classification Tree ggpairs plot

Figure 8. Random Forest ggpairs plot

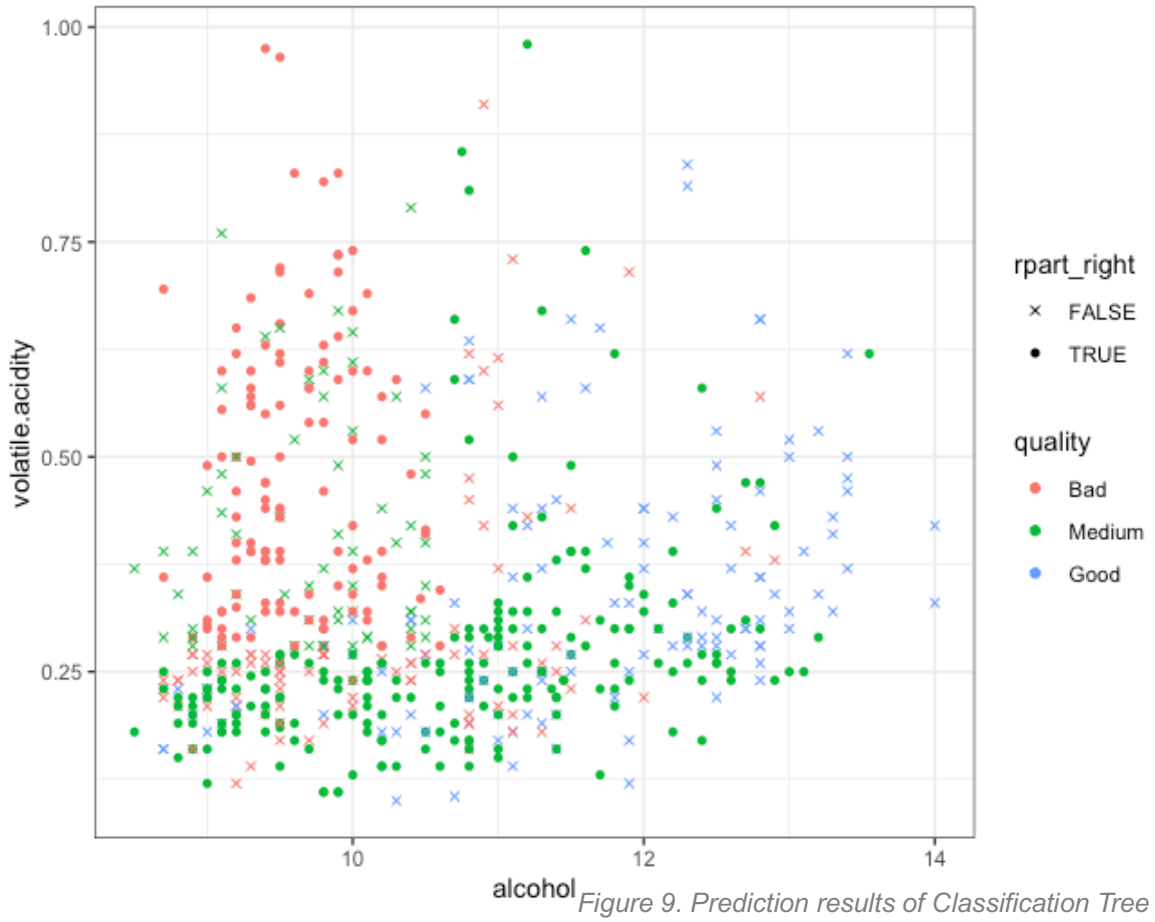


Figure 9. Prediction results of Classification Tree

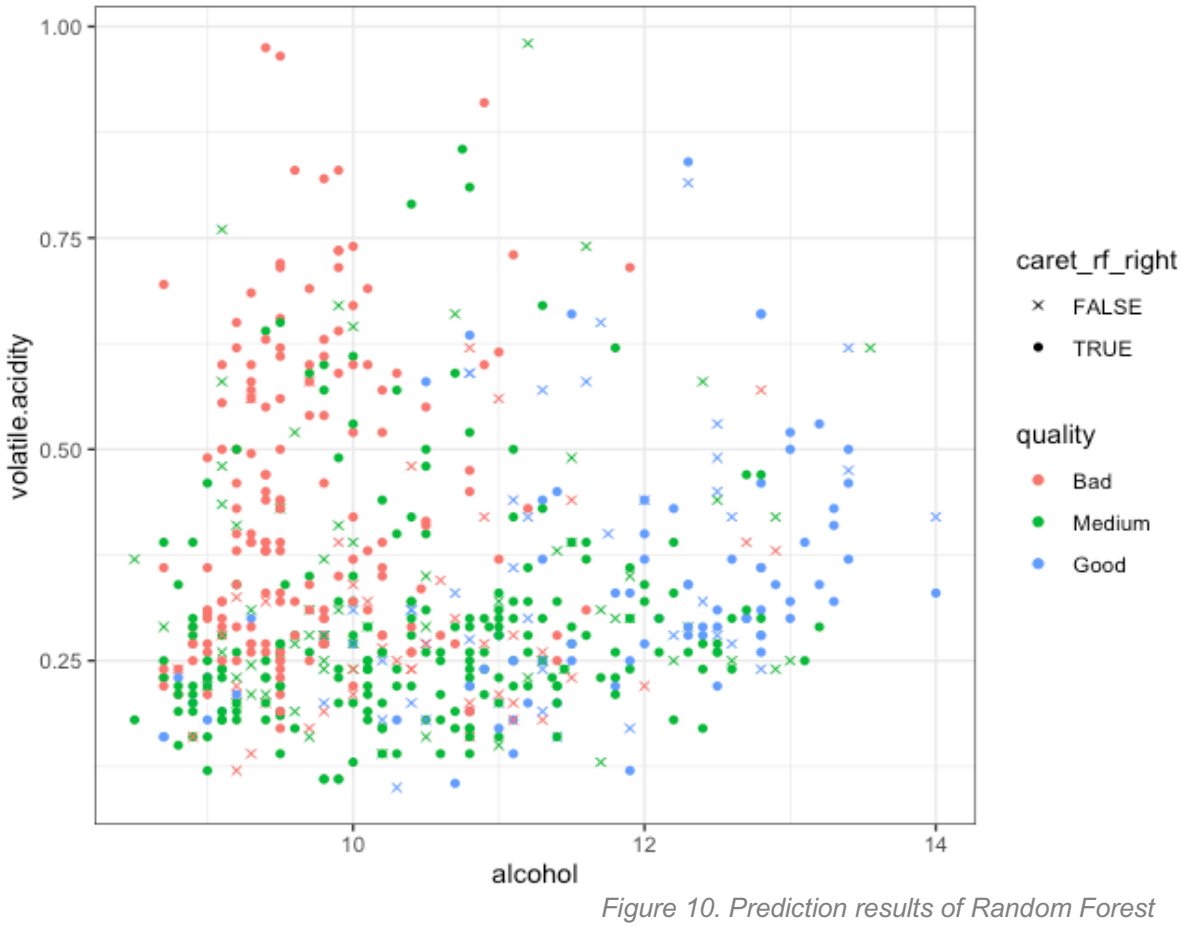


Figure 10. Prediction results of Random Forest

4.2 Horizontal data comparison

To further understand the disparities between red and white wines, I conducted training and testing using the red and white wine datasets respectively, employing the same two approaches as before. The data, model and results of these individual analyses were then compared to those of the combined dataset to distinguish the difference between red wine and white wine. The accuracy of the classification tree model for red wine was 57.5%, while for white wine it was 53.06%, and for the combined dataset it was 55.33%. Similarly, the accuracy of the random forest model for red wine was 71.25%, for white wine it was 72.45%, and for the combined dataset it was 74.83%.

It is worth noting that the quality range of red wine spans from 3 to 8.

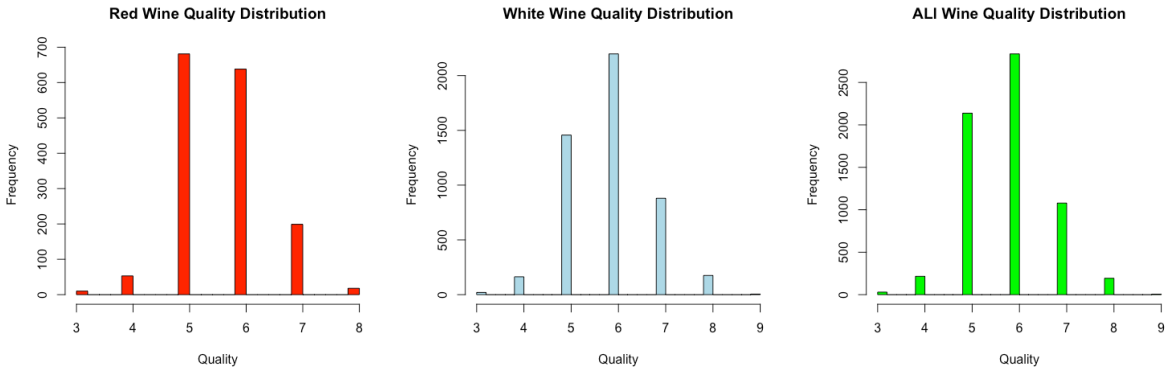


Figure 11. Quality Distribution of Red Wine, White Wine, and Combined Dataset

From Figure 12, the classification tree plots, and Figure 13, the importance tables, it was evident that the variables “sulphates” and “total.sulfur.dioxide” also played significant roles in growing the classification tree of red wine. These observations suggest that these variables may serve as key differentiators in taste between red and white wines.

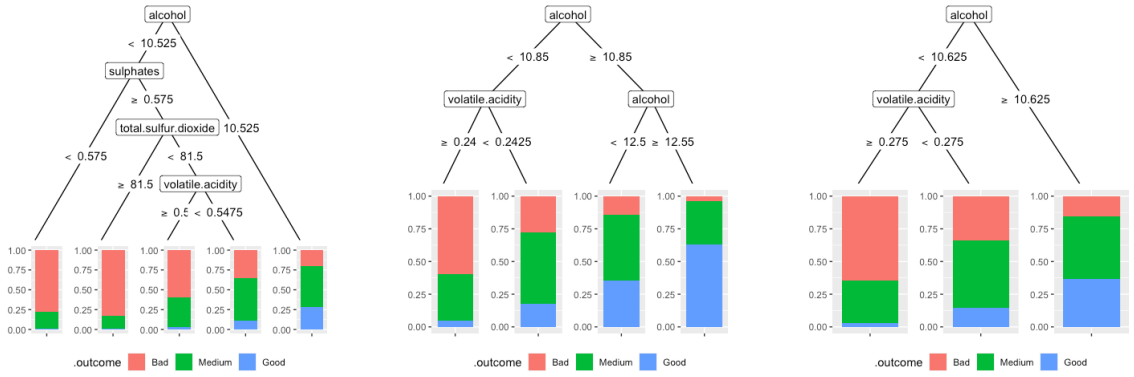


Figure 12. Classification Trees of Red Wine, White Wine, and Combined Dataset

```
> varImp(rpart)
rpart variable importance

                     Overall
alcohol              100.000
volatile.acidity      75.987
sulphates             62.677
total.sulfur.dioxide  55.379
density               25.717
fixed.acidity         22.186
chlorides             13.422
pH                     4.046
citric..acid           0.000
free.sulfur.dioxide    0.000
residual.sugar         0.000
> varImp(caret_rf)
rf variable importance

                     Overall
alcohol              100.0000
sulphates             65.5327
volatile.acidity      58.3401
total.sulfur.dioxide  42.8155
density               36.5405
chlorides             16.5937
citric.acid           12.5471
pH                    10.5077
fixed.acidity          9.4489
residual.sugar         0.3065
free.sulfur.dioxide    0.0000
```

```
rpart variable importance

                     Overall
alcohol              100.000
volatile.acidity      65.663
density               51.782
chlorides             44.126
total.sulfur.dioxide  22.713
free.sulfur.dioxide   20.720
citric.acid            9.972
pH                     6.027
fixed.acidity          5.361
sulphates              0.000
residual.sugar         0.000
> varImp(caret_rf)
rf variable importance

                     Overall
alcohol              100.000
density               71.743
volatile.acidity      54.257
free.sulfur.dioxide   43.851
total.sulfur.dioxide  39.360
residual.sugar        33.851
chlorides             30.405
pH                    20.806
citric.acid           19.707
sulphates              8.155
fixed.acidity          0.000
```

```
rpart variable importance

                     Overall
alcohol              100.000
volatile.acidity      62.692
density               52.221
chlorides             46.381
citric.acid           33.488
free.sulfur.dioxide    7.059
residual.sugar         0.000
fixed.acidity          0.000
type1                  0.000
sulphates              0.000
pH                     0.000
total.sulfur.dioxide   0.000
> varImp(caret_rf)
rf variable importance

                     Overall
alcohol              100.00
density               83.89
volatile.acidity      78.62
total.sulfur.dioxide  70.19
chlorides             69.02
free.sulfur.dioxide   67.40
sulphates             66.12
residual.sugar        64.90
citric.acid           63.94
pH                    63.77
fixed.acidity         58.24
type1                  0.00
```

Figure 13. Variable Importance Tables of Red Wine, White Wine, and Combined Dataset

From Figures 14 and 15, it's noticeable that the white wine classification tree can predict “Good” quality data, whereas the red wine model struggled to do so This discrepancy might explain why the joint dataset model also struggled to predict "Good" quality from the testing data, as the attribute was skewed towards red wine. Of course, these Figures demonstrate the superior performance of random forest models overall.

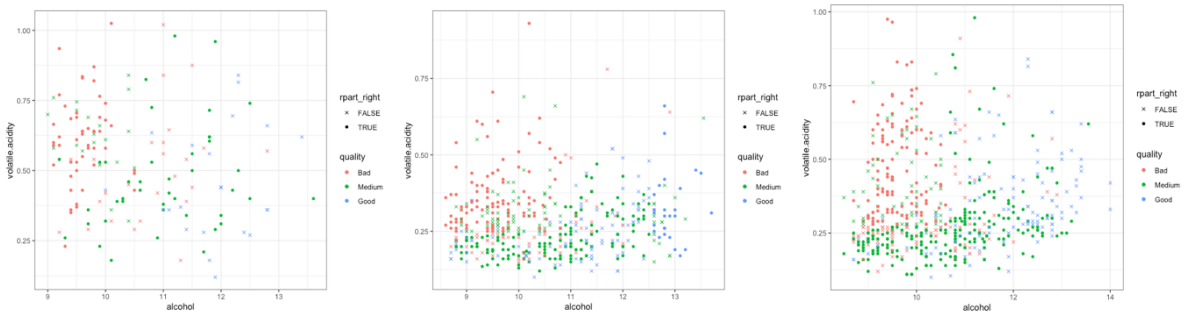


Figure 14. Prediction results of Classification Tree for Red Wine, White Wine, and Combined Dataset

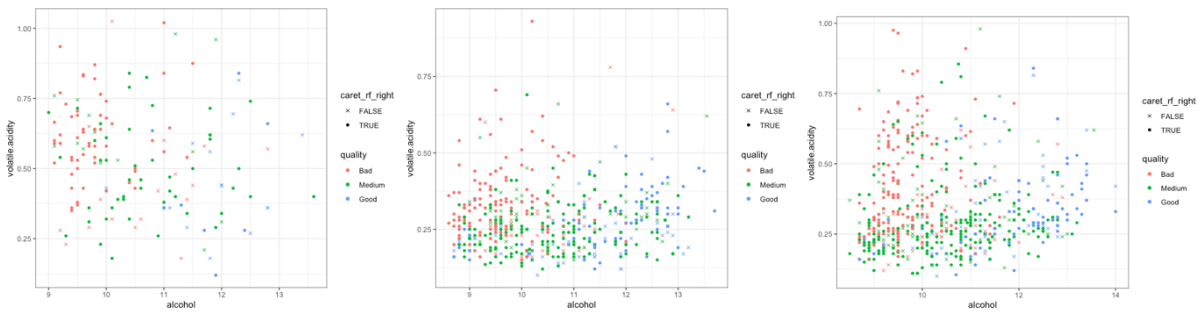


Figure 15. Prediction results of Random Forest for Red Wine, White Wine, and Combined Dataset

5. Results and Conclusion

To sum up, through the results of the two approaches, the features “alcohol”, “density” and “volatile.acidity” based on physicochemical tests have more relationship with quality score, suggesting that these attributes play a crucial role in determining the perceived quality of wine by human senses. There was also a slight difference between red wine and white wine. “sulphates” and “total.sulfur.dioxide” will also influence the quality score, compared to white wine. For red wine enthusiasts, the sulphates and total sulfuric dioxide might be the one of underlying chemical reasons why they would give better comments on red wine.

Regarding the model performance, the random forest outperformed classification trees in prediction accuracy. Of course, a random forest has a larger power and computational cost than a classification tree. However, it's worth noting that classification trees struggled in predicting "Good" quality, possibly due to limited data availability for this category in the training dataset, which resulted in the model not capturing the feature of the data with “Good” quality. Or this is the tricky difference between computer cognition and human sensory perception.

Limitations

Imbalanced Data Distribution

From the distribution of quality variable, we can see there are few data on 3,4,8 and 9. The reason behind might be this is a sensory score, which could be attributed to the subjective nature of sensory scoring, contrasting with the objective physicochemical tests. Therefore, most score tend to fall within the medium range, limiting the availability of data for training models, particularly for extreme quality ratings like "very bad" and "very excellent.". This lack of diverse data points may influence the ability of models to generalize effectively. You can see although there are 0 to 10 level of score, there are even no data of 0, 1, 2 and 10 scores. The scope could be narrowed down.

Outliers

The dataset contains numerous outliers. I roughly retained them directly to obtain more information and improve the generalizability of the model performance. While decision tree and random forest models are known for their robustness to outliers, their

presence could potentially affecte the results and accuracy of my models, and perhaps they have indeed affected them.

Limited Domain Knowledge

Another limitation is my superficial knowledge of wine, such as its chemistry and sensory attributes, may have influenced the data preprocessing, model selection, and interpretation of results. Collaborating with domain experts such as chemists or wine enthusiasts could enhance the comprehension and handling of the data, leading to more informed analysis and decision-making.

Further Plan

Data Supplementation

The data representing excellent and extremely poor-quality scores needs to be introduced to enable the model to capture more comprehensive information for better training and prediction of data.

Introducing New Features

In addition to physicochemical data, some sensory data, including sensory data such as colour and flavour could provide valuable insights into wine quality. Introducing such data may enrich the predictive models and lead to more accurate quality prediction.

Exploration of other Advanced Models

Considering the complexity of the dataset, exploring more sophisticated models such as neural networks could yield better performance. Alternatively, quality scores can be viewed as continuous values and regression modelling can be attempted to see if there is a better model performance.

Other Interesting Response Variable

Apart from predicting the quality score of wine, these experimental data could be leveraged to predict wine vintage, offering a different perspective on wine analysis. Focusing on a more objective target variable, such as vintage, could expand the scope of the analysis and explore new research directions in the field.

Appendix:

Part of combined dataset

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality	type
1	7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	Bad	1
2	7.8	0.880	0.00	2.60	0.098	25	67	0.9968	3.20	0.68	9.8	Bad	1
3	7.8	0.760	0.04	2.30	0.092	15	54	0.9970	3.26	0.65	9.8	Bad	1
4	11.2	0.280	0.56	1.90	0.075	17	60	0.9980	3.16	0.58	9.8	Medium	1
5	7.4	0.700	0.00	1.90	0.076	11	34	0.9978	3.51	0.56	9.4	Bad	1
6	7.4	0.660	0.00	1.80	0.075	13	40	0.9978	3.51	0.56	9.4	Bad	1
7	7.9	0.600	0.06	1.60	0.069	15	59	0.9964	3.30	0.46	9.4	Bad	1
8	7.3	0.650	0.00	1.20	0.065	15	21	0.9946	3.39	0.47	10.0	Good	1
9	7.8	0.580	0.02	2.00	0.073	9	18	0.9968	3.36	0.57	9.5	Good	1
10	7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	Bad	1
11	6.7	0.580	0.08	1.80	0.097	15	65	0.9959	3.28	0.54	9.2	Bad	1
12	7.5	0.500	0.36	6.10	0.071	17	102	0.9978	3.35	0.80	10.5	Bad	1
13	5.6	0.615	0.00	1.60	0.089	16	59	0.9943	3.58	0.52	9.9	Bad	1
14	7.8	0.610	0.29	1.60	0.114	9	29	0.9974	3.26	1.56	9.1	Bad	1
15	8.9	0.620	0.18	3.80	0.176	52	145	0.9986	3.16	0.88	9.2	Bad	1
16	8.9	0.620	0.19	3.90	0.170	51	148	0.9986	3.17	0.93	9.2	Bad	1
17	8.5	0.280	0.56	1.80	0.092	35	103	0.9969	3.30	0.75	10.5	Good	1
18	8.1	0.560	0.28	1.70	0.368	16	56	0.9968	3.11	1.28	9.3	Bad	1
19	7.4	0.590	0.08	4.40	0.086	6	29	0.9974	3.38	0.50	9.0	Bad	1
20	7.9	0.320	0.51	1.80	0.341	17	56	0.9969	3.04	1.08	9.2	Medium	1
21	8.9	0.220	0.48	1.80	0.077	29	60	0.9968	3.39	0.53	9.4	Medium	1
22	7.6	0.390	0.31	2.30	0.082	23	71	0.9982	3.52	0.65	9.7	Bad	1
23	7.9	0.430	0.21	1.60	0.106	10	37	0.9966	3.17	0.91	9.5	Bad	1
24	8.5	0.490	0.11	2.30	0.084	9	67	0.9968	3.17	0.53	9.4	Bad	1
25	6.9	0.400	0.14	2.40	0.085	21	40	0.9968	3.43	0.63	9.7	Medium	1
26	6.3	0.390	0.16	1.40	0.080	11	23	0.9955	3.34	0.56	9.3	Bad	1
27	7.6	0.410	0.24	1.80	0.080	4	11	0.9962	3.28	0.59	9.5	Bad	1
28	7.9	0.430	0.21	1.60	0.106	10	37	0.9966	3.17	0.91	9.5	Bad	1
29	7.1	0.710	0.00	1.90	0.080	14	35	0.9972	3.47	0.55	9.4	Bad	1
30	7.8	0.645	0.00	2.00	0.082	8	16	0.9964	3.38	0.59	9.8	Medium	1
31	6.7	0.675	0.07	2.40	0.089	17	82	0.9958	3.35	0.54	10.1	Bad	1
32	6.9	0.685	0.00	2.50	0.105	22	37	0.9966	3.46	0.57	10.6	Medium	1

Reference:

IBM (2024), What is random forest?

Available at: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems. (Accessed on 12th February 2024)

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. ISSN: 0167-9236.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016

[Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf

[bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib