

Wine Quality

Exploring the Relationship Between Chemical Properties and Wine Quality

by Shu-Yu, Lin

2.2.1 Introduction

The wine quality datasets used in this analysis originate from the Vinho Verde region in the northwest of Portugal, a well-known area for producing young, fresh, and slightly effervescent wines. The region's unique climate and grape varieties contribute to the distinct characteristics of its red and white wines. These datasets were made publicly available through the UCI Machine Learning Repository (Dua & Graff, 2017) and were initially compiled by researchers for a study on modeling wine quality using machine learning techniques (Cortez et al., 2009).

The data comprises two separate files: one for red wine and one for white wine. Each entry in the datasets represents a different wine sample, characterized by 11 physicochemical properties, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. In addition to these input features, each sample includes sensory-based quality scores ranging from 0 to 10, which are rated by wine tasters.

The main objective of the dataset is to support predictive modelling of wine quality based on measurable chemical properties, offering a valuable resource for research in data science, food chemistry, and quality control. It allows for classification and regression analysis, helping to uncover the relationships between chemical composition and perceived wine quality. These insights could assist winemakers in refining production processes and maintaining consistent product standards.

After performing the correlation analysis on the merged dataset, the relationship between the input features and the target variable appeared relatively weak. This led to the hypothesis that red and white wines may have different quality standards and influential chemical characteristics. As a result, I split the dataset and analyzed it separately. Feature importance was explored using heatmaps and tree-based methods such as Mean Decrease Accuracy and Mean Decrease Impurity, revealing subtle differences in which attributes most significantly impact wine quality for red versus white wines. That is, feature selection plays a critical role in improving model performance and interpretability, especially in high-dimensional datasets (Guyon & Elisseeff, 2003). For example, alcohol and sulphates play stronger roles in red wine quality, while residual sugar and pH are more relevant for white wines.

To better reflect practical applications, the quality score was treated as a categorical variable rather than a continuous one. The scores were grouped into quality levels, transforming the task into a classification problem. To model each dataset, two machine learning algorithms were applied: Classification and Regression Trees (CART) with appropriate pruning techniques and Random

Forests. These models were chosen for their interpretability, robustness, and suitability for classification tasks involving structured tabular data. This approach follows the methodology introduced by Cortez et al. (2009), who originally applied decision tree-based models to predict wine preferences from physicochemical features. CART, in particular, is known for its simplicity and transparency, making it useful when model interpretability is essential (Timofeev, 2004). While the random forest, has been shown to perform well in wine classification tasks, including wine type prediction, as demonstrated by Cao, Chen, and Lin (2022). In addition, the random forest was selected for its ability to handle high-dimensional data and its consistent performance in various classification domains—including applications beyond the wine industry, such as remote sensing (Belgiu & Drăguț, 2016). The final section of the report compares the performance of these models on red and white wine data separately, providing insights into their effectiveness and the unique characteristics of each wine type.

2.2.2 Data Cleaning and Exploratory Data Analysis

The dataset provided was clean and well-structured, with no missing values or obvious outliers. After confirming that all variables were loaded correctly as numerical data, the dataset was split into two groups: red wine (1,599 samples) and white wine (4,898 samples).

An initial exploratory analysis was conducted using bar plots to visualize the distribution of quality scores in both wine types. It was observed that the majority of red and white wines were rated as quality levels 5 and 6, followed by level 7. Extremely low (quality = 3) and extremely high (quality = 9) scores were rare, indicating a class imbalance issue in the dataset. (Figure 1)

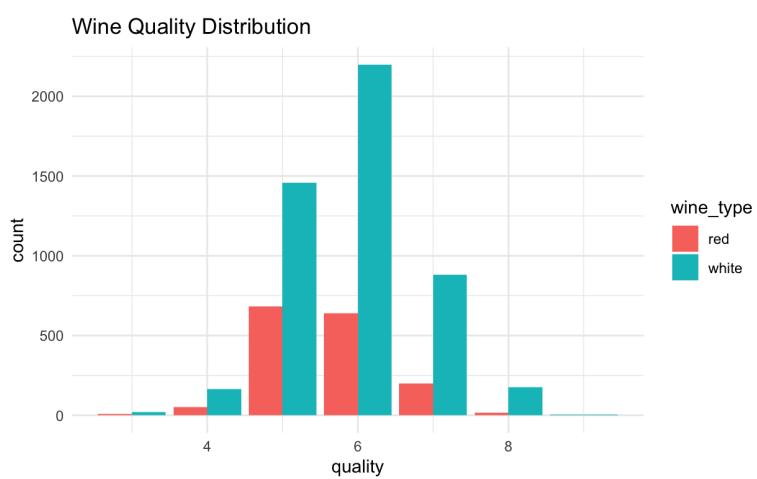


Figure 1: Wine Quality Score Distribution by Type

Following the initial inspection, early experiments were conducted using CART models with the original quality score as the target. Despite tuning hyperparameters, pruning, and adjusting feature weights, the models consistently struggled to classify the extreme quality levels (3, 4, 8, and 9). Most predictions were concentrated around classes 5, 6, and 7.

While the majority of these central classes were correctly classified, there was still a significant amount of misclassification among them. For example, several quality 5 wines were predicted as quality 4, and many wines labelled as quality 6 were misclassified as either 5 or 7. To understand this behaviour, boxplots were used to examine the feature distributions of classes 5, 6, and 7. These plots showed a high degree of overlap in several features—such as volatile acidity, alcohol, and sulphates—making it difficult for the model to form distinct boundaries between these adjacent classes. The high similarity in feature values contributed to the misclassification, even for the most frequent classes.

Furthermore, despite relatively better classification rates for classes 5, 6, and 7, the overall model performance remained unsatisfactory. Key evaluation metrics—including accuracy, Cohen's kappa, sensitivity, and balanced accuracy—were all noticeably low. These results indicated that the model was not only struggling with the minority classes but also failed to achieve reliable performance across the entire classification spectrum. To address both the class imbalance and the difficulty of distinguishing overlapping quality levels, the original 0–10 quality scores were re-categorized into three broader groups: low, medium, and high. A new column, `quality_category`, was created as a categorical variable to be used in classification models. This transformation simplified the prediction task and better reflected practical scenarios where broader quality groupings may be sufficient.

Subsequent CART models were trained using this categorical target. Feature importance was also evaluated using tree-based methods (Mean Decrease Accuracy and Mean Decrease Impurity), and separate analyses were performed for red and white wines to account for their differing characteristic patterns. Random Forest classifiers were applied in parallel to provide a performance benchmark and evaluate consistency with CART results.

2.2.3 Modelling

To address the wine quality classification task, I applied two supervised machine learning models as required: Classification and Regression Trees (CART) from List 1 and Random Forests from List 2. Both models were developed separately for red and white wines due to their distinct feature distributions and underlying quality characteristics.

Red Wine Modelling

Based on insights from the heatmap, pairplot, and boxplot analyses, I initially selected the top five features showing the highest correlation with wine quality. (Figure 2 and Figure 3) However, some of these features exhibited significant overlap across low, medium, and high-quality groups, as seen in the boxplots. Therefore, I re-evaluated and finalized the following five features for model training: alcohol, volatile acidity, sulphates, citric acid and total sulfur dioxide.

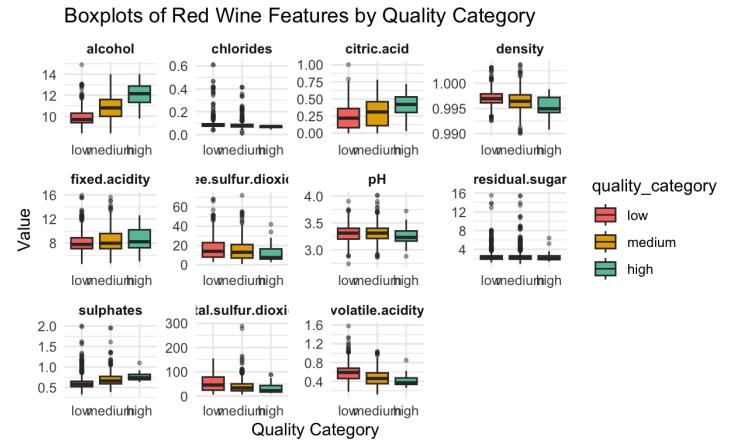
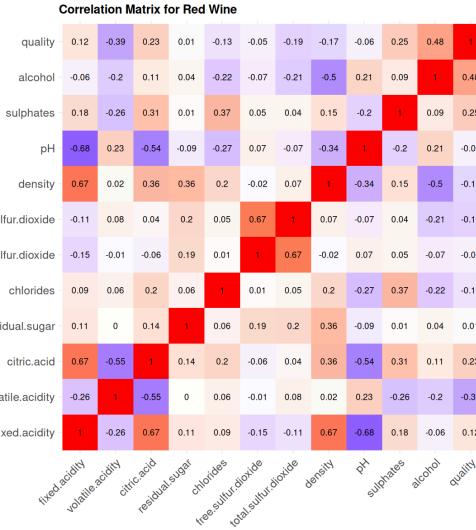


Figure 2: Red Wine Feature Correlation Heatmap

Figure 3: Boxplots of Red Wine Features by Quality Category

For the CART model, I carefully tuned hyperparameters, including `cp`, `min split`, `max depth`, and `class weights`, to address the class imbalance problem—especially the underrepresentation of high-quality samples. To ensure the decision tree remained interpretable and did not overfit, I constrained the depth to approximately 4–5 levels. This pruning strategy led to a cleaner and more generalizable tree structure. (Figure 4)

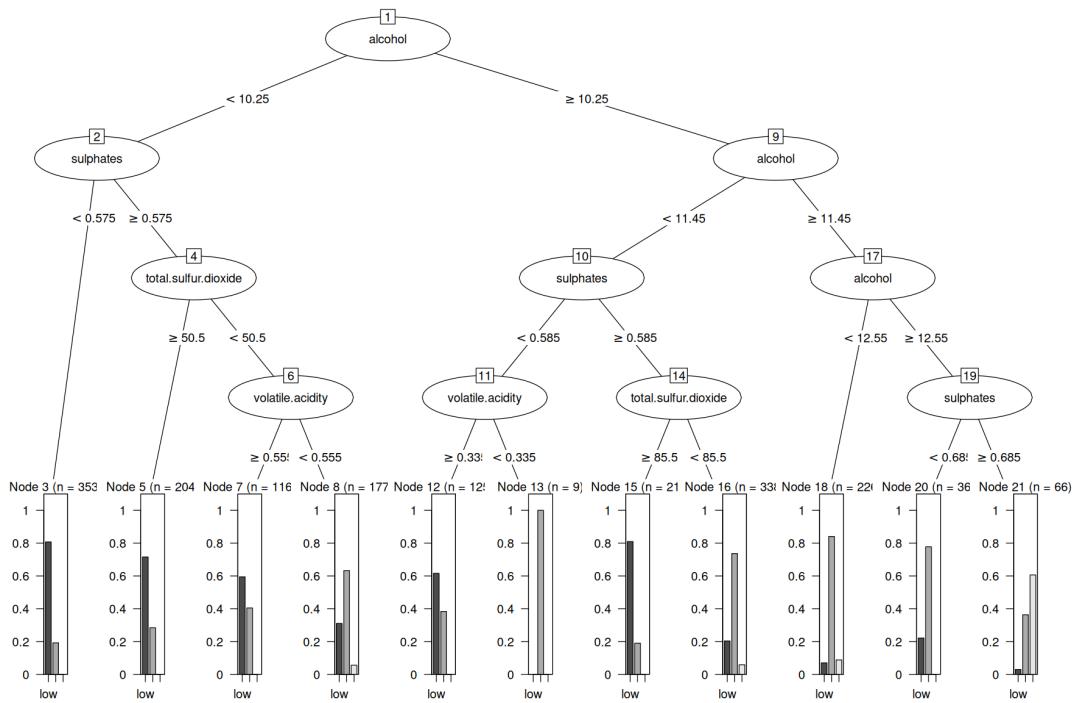


Figure 4: CART Classification Tree for Red Wine Quality

After parameter tuning, the CART model achieved an overall accuracy of 0.74 and a Kappa score of 0.50, indicating moderate agreement beyond chance. Sensitivity scores for each class were also acceptable, particularly for the low and medium categories. This suggests the tree was reasonably effective in differentiating between broad-quality classes.

Subsequently, a Random Forest model was built using the same five selected features. The feature importance plots (based on Mean Decrease Accuracy and Mean Decrease Gini) confirmed that alcohol, sulphates, and volatile acidity were the most influential variables in predicting red wine quality.(Figure 5) The ensemble approach improved overall stability and provided better class balance, though high-quality classification remained a challenge due to the limited sample size.This modeling approach is consistent with previous research that applied Random Forest algorithms to wine-related classification tasks, such as predicting wine type from physicochemical features (Cao, Chen, & Lin, 2022), further supporting the model's suitability in this domain.

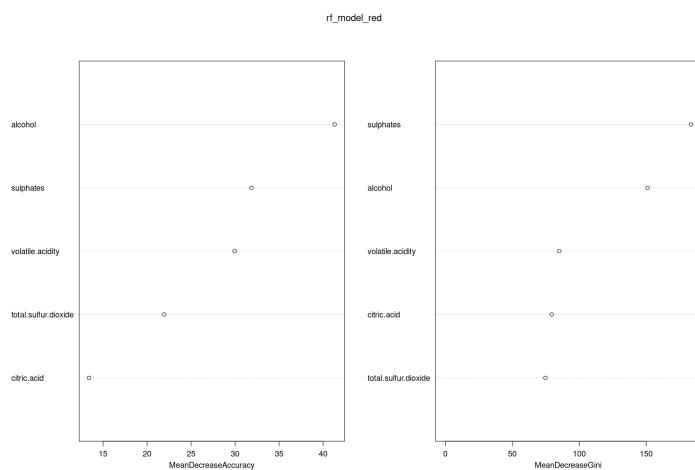


Figure 5: Variable Importance Plots for Red Wine Classification

White Wine Modelling

The same modelling process was followed for white wines. Starting from the heatmap and boxplot observations, I initially selected the top five features with the highest correlation. However, due to overlapping distributions in some variables, the final chosen features were: alcohol, volatile acidity, and total sulfur dioxide. (Figure 6 and Figure 7)

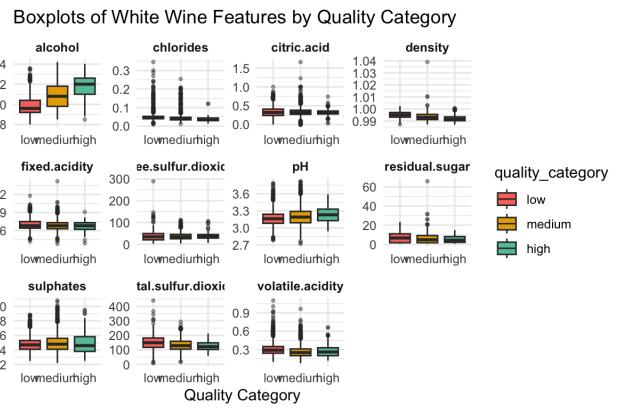
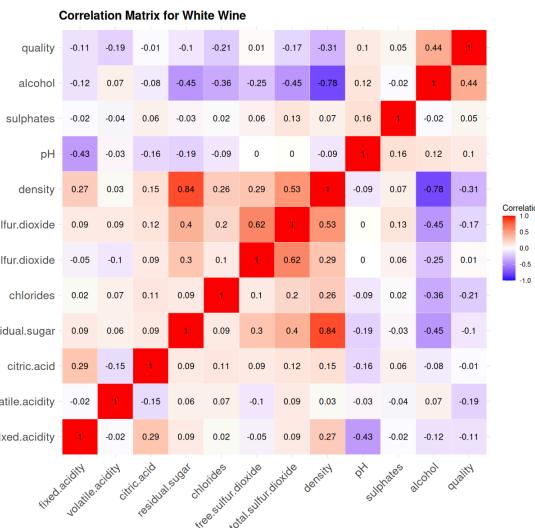


Figure 6: White Wine Feature Correlation Heatmap

Figure 7: Boxplots of White Wine Features by Quality Category

As with red wine, the CART model for white wine underwent tuning for cp, min split, max depth, and class weights. Similar pruning strategies were used to control tree complexity and improve interpretability. The final tuned CART model reached an accuracy of 0.70 with reasonable sensitivity across the three quality groups. (Figure 8)

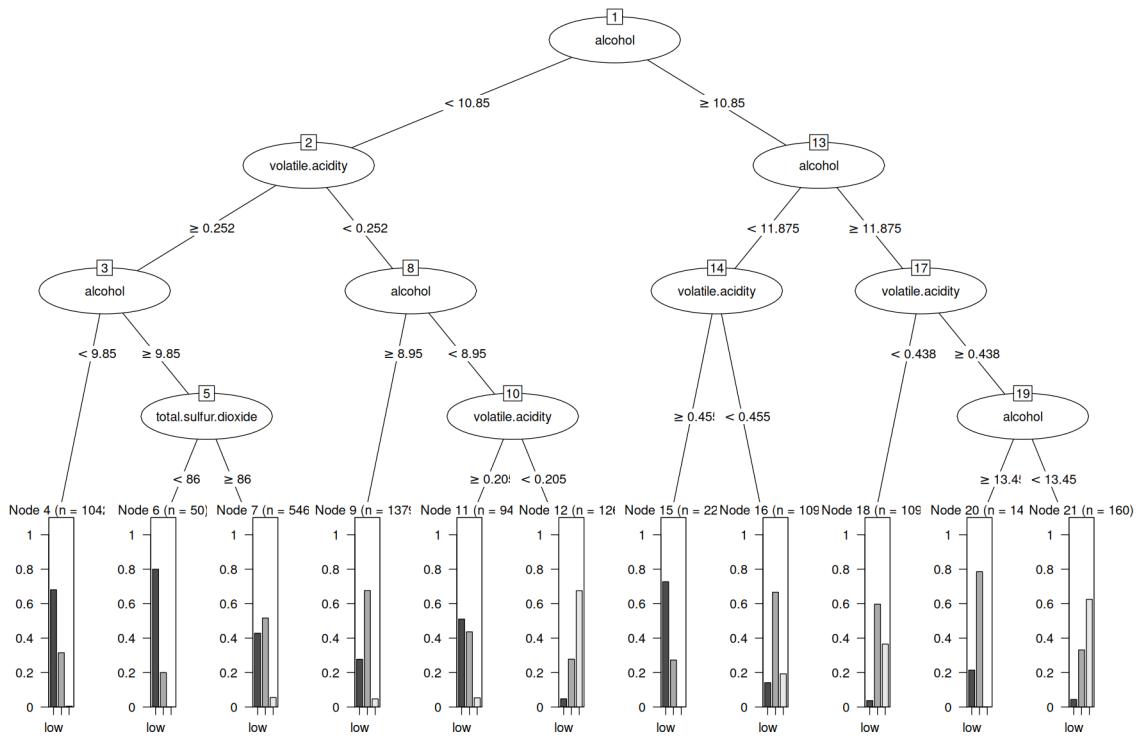


Figure 8: CART Classification Tree for White Wine Quality

The Random Forest model for white wine also confirmed the selected features' importance, with alcohol, density, and volatile acidity ranking highest based on both accuracy and Gini-based metrics. While the overall model performed well, correctly predicting low and medium qualities, the high-quality class remained difficult to classify due to its small sample size. Nonetheless, the

ensemble model delivered more robust and consistent predictions than the standalone CART model. (Figure 9)

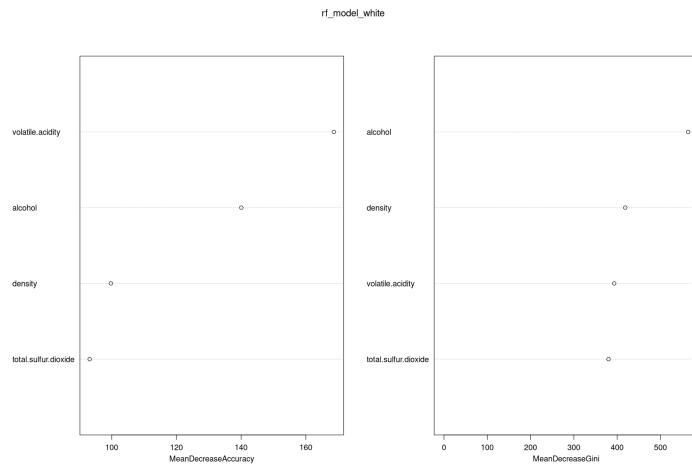


Figure 9: Variable Importance Plots for White Wine Classification

In both red and white wine CART trees, alcohol was consistently used as the top-level splitting feature, confirming its dominant role in predicting wine quality. Other features such as sulphates, volatile acidity, and total sulfur dioxide also contributed to mid- and lower-level splits, supporting earlier EDA findings. Interestingly, the white wine tree prioritized volatile acidity earlier than the red wine tree, suggesting subtle differences in how acidity affects perceived quality in different wine types.

Overall, both models successfully captured the dominant quality patterns in red and white wines, especially within the medium and low categories. The combination of EDA-guided feature selection, pruning strategies, and ensemble methods resulted in competitive performance, although challenges remain in classifying rare high-quality samples.

2.2.4 Model Comparison

Red Wine

To validate and compare the performance of the CART and Random Forest models, both were trained and evaluated using the same stratified data split and class weighting strategy. Model performance was assessed using several key metrics: accuracy, Cohen's kappa, balanced accuracy and ROC curves, particularly focusing on the high-quality wine class due to its underrepresentation.

Model	Accuracy	Kappa	Balanced_Accuracy
CART	0.7442151	0.5054705	0.7430647
Random Forest	0.8050314	0.6125000	0.7596787

Table 1: Model Performance Comparison (Accuracy, Balanced Accuracy, Kappa)

From this comparison, it is clear that Random Forest outperformed CART in all three evaluation metrics. The accuracy of Random Forest was about 6% higher, while the Kappa value—which accounts for agreement beyond chance—improved from 0.51 to 0.61, indicating stronger and more reliable predictions.

While both models performed reasonably well in classifying low- and medium-quality wines, the most significant differences emerged in the classification of high-quality wines—a minority class in both datasets. For this reason, the ROC curve analysis was focused specifically on the high-quality class, providing a more detailed comparison of each model's sensitivity and specificity when detecting this rare but essential class.

The performance bar plot clearly illustrates Random Forest's consistent advantage across all metrics. (Figure 10) The ROC curve for the high-quality class reveals that Random Forest also had better sensitivity and specificity in detecting high-quality wines, an essential yet complex class due to its small sample size. The Random Forest ROC curve reaches higher accurate favourable rates across various thresholds compared to the CART model. (Figure 11)

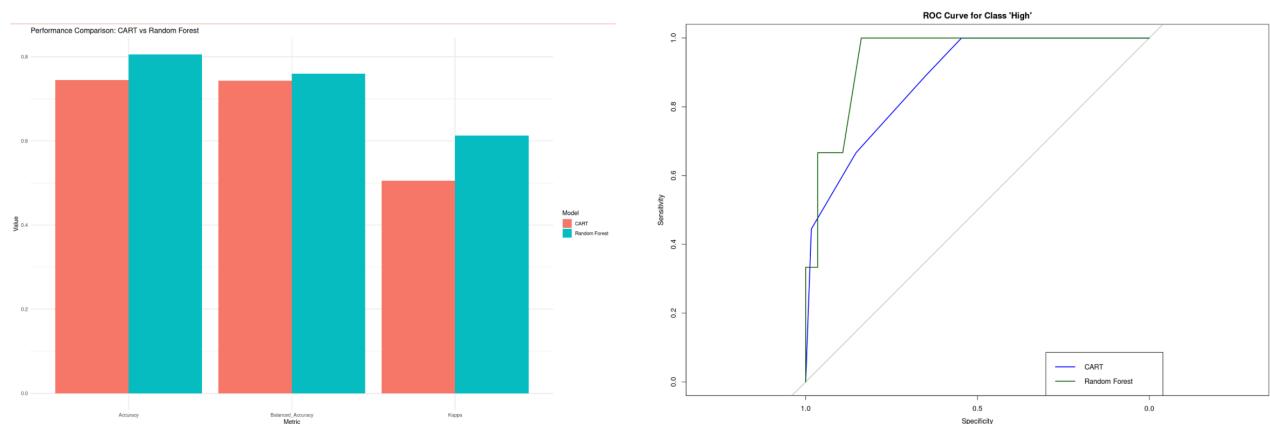


Figure 10 & Figure 11: ROC Curve for High-Quality Class (CART vs. Random Forest)

White Wine

To compare the effectiveness of CART and Random Forest models on the white wine dataset, both models were trained using the same stratified train-test split and adjusted with class weights to mitigate the impact of class imbalance. Model performance was evaluated using, Balanced

Accuracy and ROC curves, with particular attention paid to the high-quality class, which is notably underrepresented in the dataset.

Model	Accuracy	Kappa	Balanced_Accuracy
CART	0.7053900	0.3597927	0.6471165
Random Forest	0.7834525	0.5367274	0.7535841

Table 2: Model Performance Comparison (Accuracy, Balanced Accuracy, Kappa)

The results clearly show that the Random Forest model outperformed CART across all primary evaluation metrics. It achieved a higher overall accuracy (+7.8%), a more substantial kappa score (+17.7%), and better-balanced accuracy (+10.6%). These results suggest that Random Forest provides more stable and reliable predictions, especially in scenarios with imbalanced classes.

While both models performed reasonably well in classifying low-and medium-quality wines, the primary distinction between their performances lies in their ability to classify high-quality wines correctly. Due to the small number of high-quality samples, both models exhibited relatively low sensitivity for this class, but Random Forest demonstrated a noticeable advantage. (Figure 12)

To investigate this further, a ROC curve for the high-quality class was generated (Figure 13). The curve illustrates that Random Forest had significantly at various specificity levels, confirming its stronger capability in detecting this rare but important class. The improved area under the curve (AUC) for Random Forest further reinforces its superior discriminative power in this context.

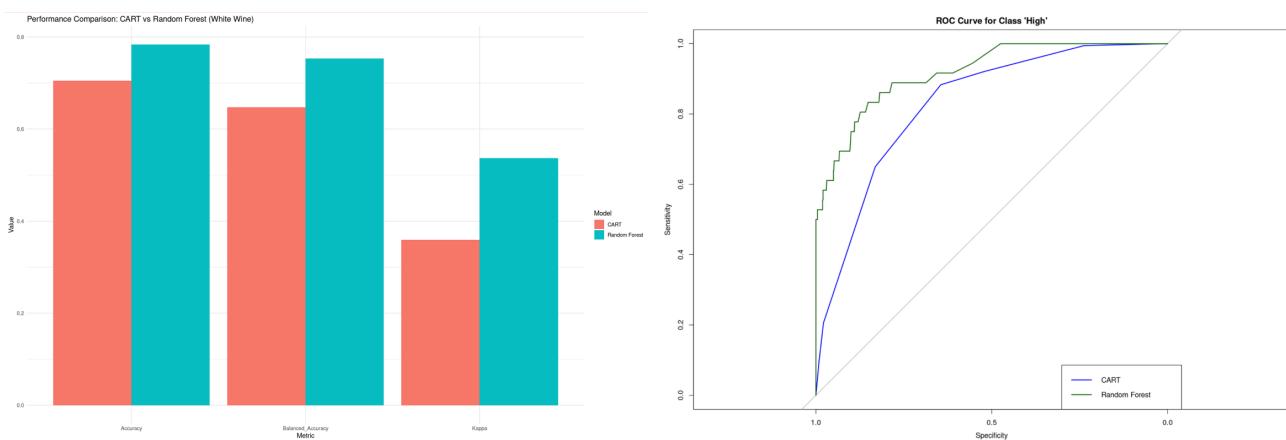


Figure 12 & Figure 13: ROC Curve for High-Quality Class (CART vs. Random Forest)

Overall, Random Forest consistently outperformed CART across both red and white wine datasets. It demonstrated higher predictive accuracy, more balanced classification across classes, and greater ability to identify rare high-quality wines. However, CART remains valuable for its interpretability, especially when explainability is critical for practical use. As noted by Timofeev (2004), CART models provide simple, rule-based decision structures that are easy to interpret and

communicate. Therefore, the final model choice depends on the application's priorities—Random Forest is recommended for performance-focused scenarios, while CART may be preferred when transparency and simplicity are paramount.

2.2.5 Results and Conclusion

This project aimed to classify wine quality using physicochemical properties, with a focus on comparing the performance of CART and Random Forest models on red and white wine datasets. The results demonstrated that Random Forest is consistently better than CART across key evaluation metrics such as accuracy, balanced accuracy, and Cohen's kappa. While both models performed adequately in distinguishing low- and medium-quality wines, the classification of high-quality wines remained a challenge, primarily due to severe class imbalance and overlapping feature distributions.

One key difficulty observed throughout the modelling process was that the feature distributions across most quality levels—especially grades 5, 6 and 7, were highly similar. Even when visualized through boxplots and pair plots, many physicochemical features such as density, acidity, or sulfur dioxide showed substantial overlap between classes. This overlap limited the model's ability to clearly separate quality categories, contributing to frequent misclassifications, particularly around the mid-range scores.

A significant strength of this project was the clear separation and tailored modelling for red and white wines, acknowledging their differing chemical profiles. In both cases, alcohol was a strong predictor, followed by volatile acidity, sulphates, and total sulfur dioxide. Using both performance metrics and ROC curves enabled a more detailed model evaluation, particularly for underrepresented classes.

However, several limitations were encountered. The dataset lacked sensory or contextual features (e.g., grape type, fermentation time), which may be critical to wine quality assessment. In addition, the highly imbalanced distribution of quality scores—especially the low representation of extreme classes (quality 3, 9)—posed significant challenges for classification models, even when weights were adjusted. Furthermore, the models were evaluated only on a single train-test split; using cross-validation could provide more robust performance estimates. Thus, for the next step, I would highly recommend adding inclusion features, such as production methods or sensory scores, to enrich the model's context and reduce reliance only on chemical measurements.

In conclusion, this study provided a solid foundation for wine quality prediction using supervised learning techniques. With further improvements in data diversity and model sophistication, such approaches can be valuable tools in supporting quality control and classification in the wine industry.

References

- Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31.
<https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Cao, Y., Chen, H., & Lin, B. (2022). Wine type classification using random forest model. *Highlights in Science, Engineering and Technology*, 4, 400–407.
<https://www.researchgate.net/publication/362436263> Wine Type Classification Using Random Forest Model
- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Modeling wine preferences by data mining from physicochemical properties* (Technical report). University of Minho.
<http://www3.dsi.uminho.pt/pcortez/wine5.pdf>
- Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547–553.
<https://doi.org/10.1016/j.dss.2009.05.016>
- Dua, D., & Graff, C. (2017). *Wine quality data set*. UCI Machine Learning Repository.
<https://archive.ics.uci.edu/dataset/186/wine+quality>
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157–1182.
<http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf>
- Timofeev, R. (2004). *Classification and regression trees (CART) theory and applications*. Humboldt University, Berlin, 54,48
<https://d1wqtxts1xzle7.cloudfront.net/38106508/timofeev-libre.pdf>