Tree-Based Classification of Wine Type: Analyzing Accuracy and Interpretability in Ensemble Models

By: Jodi Barnes, Alexander Castronovo, Elise Eldridge, Chance Pickett

6th May 2025

Introduction

This project applies tree-based classification techniques to a dataset containing physicochemical measurements of red and white wines. The primary objective is to evaluate how well different decision-tree-based models can classify wine type based on these chemical characteristics, while also analyzing the trade-offs between interpretability and predictive accuracy.

Tree-based methods offer a flexible and intuitive framework for both classification and regression tasks. Basic decision trees generate human-readable rules but may suffer from high variance and limited accuracy. Ensemble techniques—such as bagging, random forests, and boosting—improve performance by combining multiple trees. Bagging averages across many bootstrapped models to reduce variance, while random forests add randomized feature selection at each split to decorrelate trees. Boosting builds trees sequentially, with each tree correcting errors made by its predecessors, and often delivers the strongest predictive performance among tree-based approaches.

To further interpret complex models, the project incorporates Shapley value analysis. This method quantifies the individual contribution of each variable to specific predictions, helping to clarify the behavior of more opaque ensemble models. By evaluating these methods side by side, the project aims to identify which features are most influential in determining wine type and assess how model complexity affects both accuracy and interpretability.

Data and Exploratory Data Analysis

From the dataset, wine class imbalance caused troubles with model sensitivity for the minority class. This was combatted by evaluating models using accuracy and variable importance rather than relying solely on class-specific metrics. The biggest thing that we discovered during our EDA was that alcohol, density, and volatile acidity showed the most pronounced differences between red and white wines, which were later confirmed as the most important features in the tree-based models.

Initial Modeling

The single classification tree model performed well and achieved a test accuracy of approximately 87%. The tree used two splits, one on alcohol and one on volatile acidity, which was easy to interpret. While this model was not as accurate as ensemble models, it effectively captured key differences between red and white wines and served as a strong baseline for other model comparisons.

Ensemble Methods and Tuning

Bagging was implemented using the randomForest package with mtry set equal to the total number of predictors. This approach provided a noticeable increase in accuracy compared to the single tree model and produced more stable predictions. The model consistently identified alcohol and density as top features, confirming earlier findings.

The random forest model also used randomForest, but with mtry tuned through cross-validation. Optimal performance was achieved near $\sqrt{p}$, which is theoretically expected. This model performed slightly better than bagging, showing improved accuracy and generalization while maintaining similar variable importance rankings.

Boosting was implemented using the gbm package, with tuning over the number of trees, interaction depth, and shrinkage. The best results came from pairing smaller learning rates with a larger number of trees.

Boosting produced the highest accuracy overall and demonstrated a strong ability to capture complex patterns in the data. This made it the most competitive model for final selection.

Model Diagnostics

Across all models, test set accuracy improved with model complexity. The single tree achieved about 87% accuracy, while bagging and random forest reached 93% and 94% respectively. Boosting performed the best, with a test accuracy near 95%. ROC curves followed a similar trend, with the single tree showing the lowest curve and boosting the highest, indicating stronger true positive rates at all thresholds. AUC values reflected this as well, with the single trailing and ensemble models producing near-perfect scores. Overall, ensemble methods showed consistent gains in both accuracy and ROC performance.

The bias-variance trade-off was clearly observed in the tuning process. Shallow trees and small ensembles exhibited higher bias, missing important patterns in the data. As tree depth increased, bias decreased, but at the risk of overfitting and increased variance. Ensemble size helped control this: adding more trees reduced variance and stabilized predictions. Random forest benefitted from tuning mtry, which helped balance the bias-variance trade-off. Boosting handled this particularly well, with small learning rates and deeper trees reducing bias while ensemble size managed variance effectively.

Final Model and Interpretation

All models performed very well with each achieving test accuracy above 90%. The most important variables were largely consistent across methods. Total sulfur dioxide, chlorides, and volatile acidity were the most influential for each model as shown by the Mean Decrease Accuracy graph. Removing these variables caused the mean accuracy to decrease significantly. The model with the best test accuracy was a non-bagged, non-boosted random forest with 99.6% test accuracy. The confusion matrix shows that it only predicted 13 wines wrong, with 11 of them being reds. The random forest model doesn't get much more accurate after 100 trees and flattens out completely at 300 trees. Shapley value analysis further validated these results, showing that alcohol, density, and volatile acidity consistently contributed the most to individual predictions.

Executive Summary

Alcohol, density, and volatile acidity were consistently identified as the most important features for distinguishing red and white wines. These patterns appeared across all models, with boosting confirming their importance through Shapley value analysis. Boosting outperformed other models in terms of accuracy and AUC, making it the most effective option. Higher alcohol levels were associated with white wine, while higher volatile acidity pointed to red, reflecting differences in fermentation processes. Despite class imbalance, model performance remained stable and required no resampling. These results show that a smalls et of well-measured chemical properties can accurately classify wine type. For practical use, producers should monitor these keys features for quality control. Future work should explore additional data, such as grape variety or production region, and test model performance on external datasets to evaluate generalizability.