Taya Bhavsar

Prof. Shaonan Tian

Data Mining for B.A.

01 December 2025

Predicting Wine Quality

Project Summary

Wine quality prediction is valuable when it comes to producing wine at a vineyard. This is due to traditional expert tasting being subjective, costly, and slow, while objective chemical measurements allow faster and more consistent quality control. For this project, I explored how winemakers predict a batch of wine based on two binary classification targets: good or poor quality. This project uses the UCI Wine Quality dataset of 1,599 Portuguese Vinho Verde red wines to explore how data mining can improve prediction and support production decisions. The dataset's real-world characteristics, including class imbalance, moderate multicollinearity, and likely non-linear relationships between chemical properties and quality, makes it well suited for machine learning. This project will apply methods such as logistic regression for interpretability, classification trees for threshold identification, and gradient boosting for capturing complex interactions. This project demonstrates how data mining can extract actionable insights, reduce reliance on expert panels, lower production costs, and enhance wine quality consistency.

Research Questions

This analysis addresses three specific research questions designed to provide both theoretical insights and practical guidance for wine production. First, can wine quality be accurately predicted using only three chemical measurements? This question directly addresses feature parsimony, exploring whether a minimal set of laboratory tests could provide sufficient predictive power for quality control applications where comprehensive chemical analysis may be impractical or expensive. Identifying the optimal three-feature combination would enable cost-effective quality screening during wine production.

Second, could alcohol content alone be sufficient to predict quality, or are acidity measures necessary? This question investigates whether the positive correlation between alcohol and quality ratings reflects a dominant predictor or whether a more nuanced model incorporating acidity profiles is required.

Third, what is the minimum alcohol percentage for a wine to have greater than 50% probability of being rated good quality (score ≥ 6)? This threshold question provides actionable guidance for producers seeking to maximize their proportion of higher-rated wines. By identifying the critical alcohol level where wines transition from predominantly average ratings to predominantly good ratings, wine makers can establish production targets that balance consumer preferences, regulatory constraints, and economic considerations. This threshold approach acknowledges that quality prediction is inherently uncertain while still providing practical decision boundaries.

Dataset Origins & Research Context

This dataset originates from Cortez et al. (2009), who analyzed 1,599 Portuguese Vinho Verde red wines to predict expert-rated quality scores (0–10) using physicochemical properties. Their study tested several regression and machine learning models, with SVM (Support Vector Machines) performing best. The dataset is valuable for modern wine-quality prediction because traditional sensory evaluation is costly and subjective, while data-driven models can highlight the chemical features most associated with perceived quality, support production improvements, and inform targeted marketing. Traditional sensory methods of texture evaluation involve assessment and grading by "expert" tasters, in which one or two trained experts assign quality scores on the appearance, flavor, and texture of the products based on the presence or absence of predetermined defects. Its real-world characteristics—imbalanced quality classes and correlated chemical attributes—make it well-suited for data-mining approaches such as feature selection and sensitivity analysis, demonstrating how objective lab measurements can model subjective human preferences.

The dataset presents several characteristics that make it ideal for data mining applications. Classes are imbalanced, with many more normal-quality wines than excellent or poor ones, reflecting real-world production distributions. Additionally, several physico-chemical attributes may be correlated, making feature selection methods particularly relevant for identifying the most influential predictors. The original study employed sensitivity analysis to measure how input variables affect quality predictions, demonstrating that objective laboratory measurements can effectively model subjective sensory evaluations.

Data Preparation

The dataset, sourced from the UCI Machine Learning Repository, focuses exclusively on 1,599 Portuguese red Vinho Verde wines, making it a cohesive sample with distinct chemical characteristics compared to white wines and offering meaningful analytical challenges such as class imbalance and quality ratings concentrated between 3 and 8. Red wine's unique fermentation and aging processes create chemical–quality relationships that are well suited for data mining. Quality scores were transformed into two classification targets: a balanced binary split of Bad (<6) vs. Good (≥6), and a three-tier system (Low 3–4, Medium 5–6, High 7–8) for multi-class analysis. All 11 physicochemical predictors were standardized using z-score normalization—applied only after the train-test split—to ensure comparability across variables with different measurement scales and to support proper model behavior, particularly for the logistic regression model.

Experimental Setup

All models were trained and evaluated using a consistent experimental setup. The dataset of 1,599 wines was split into 80% training (1,279 samples) and 20% testing (320 samples) sets via stratified sampling, with a random seed of set.seed(123) for reproducibility. Z-score normalization was applied to all predictors using training-set parameters to avoid data leakage. Hyper-parameter tuning was conducted with 5-fold cross-validation on the training set: logistic regression assessed coefficient stability, classification trees explored maximum depths {3–8} and minimum splits {10–40}, random forest used OOB error with 'ntree' {100–300}, 'mtry' {2–5},

and nodesize {1, 5, 10}, and gradient boosting performed grid search over learning rates {0.01, 0.05, 0.1}, n_estimators {100–300}, and max_depth {3–5}. Each model's performance is evaluated on the held-out test set using accuracy, precision, recall, F1-score, and ROC-AUC metrics.

Correlation Matrix Analysis

The correlation matrix revealed moderate relationships between wine chemistry and quality, with alcohol showing the strongest positive correlation (r = 0.48) and volatile acidity the strongest negative (r = - 0.39), while sulphates and citric acid had weaker positive associations. Three chemical clusters emerged: an acidity cluster linking fixed acidity, citric acid, density, and pH; an alcohol–density cluster reflecting alcohol's lower density; and a sulfur dioxide cluster connecting free and total sulfur dioxide. Although some feature correlations reached 0.6–0.68, none exceeded the 0.7 multicollinearity threshold, indicating manageable redundancy. Variables like chlorides, sulphates, and volatile acidity remained relatively independent, contributing unique predictive signals. Overall, the correlation structure suggests moderate feature overlap without severe multicollinearity, supporting the use of models that can handle correlated inputs, such as regularized regression or tree-based methods.

Outlier Analysis

The dataset showed several meaningful outliers across key chemical features, all of which were considered oenologically plausible and retained for analysis. Residual sugar, total sulfur dioxide, and volatile acidity had the most extreme deviations, reflecting sweeter wine styles,

differing sulfite practices, and variable acetic acid levels. Other variables—such as fixed acidity, citric acid, and chlorides—displayed moderate but expected variations tied to grape composition, fermentation choices, and terroir influences. Meanwhile, pH, alcohol, and sulphates showed stable distributions with minimal extremes, consistent with typical red wine chemistry. The outlier patterns reflect genuine diversity in Vinho Verde production techniques rather than measurement errors.

Methodology Overview

This study employs four supervised learning methods—logistic regression, classification trees, random forests, and gradient boosting—to model wine quality from physicochemical attributes. These approaches were selected to balance interpretability, model complexity, and predictive performance while directly addressing the research questions.

Logistic regression functions as the primary interpretable baseline. Its coefficient estimates quantify the independent contribution of each chemical property, enabling evaluation of alcohol content as a standalone predictor and identification of the most influential variables for parsimonious modeling. Its linear decision boundary provides a clear reference point for comparing more complex models.

Classification trees introduce non-linearity through recursive binary partitioning, producing interpretable threshold-based rules that reveal how specific chemical values differentiate quality categories. The hierarchical structure is particularly suited for determining actionable cut-points, such as the minimum alcohol level associated with higher-quality ratings.

Random forests extend the tree-based framework by aggregating predictions from numerous bootstrapped trees constructed using randomly selected feature subsets. This ensemble design reduces variance, mitigates overfitting, and captures interaction effects that single trees may miss. In addition, random forests yield stable and informative feature importance metrics that complement the logistic regression findings.

Gradient boosting generates a sequential ensemble in which each tree corrects the residual errors of its predecessors. This method is well suited for the dataset's class imbalance and complex, non-linear chemical interactions. Interpretability is maintained through SHAP-based analyses, which illuminate feature contributions across the ensemble.

Collectively, these models span a methodological continuum from highly interpretable to highly flexible, enabling a comprehensive assessment of predictive performance and ensuring that all research questions are examined using appropriately tailored analytical techniques.

Results - Logistic Regression

The logistic regression model classified wines as “Good” (≥6) or “Bad” (<6) and performed consistently across train and test sets, indicating minimal overfitting and a stable relationship between chemical properties and quality ratings. The model showed strong discriminatory power, though calibration plots revealed that it was overconfident, producing probability estimates that were more extreme than empirically warranted. Confusion matrix patterns, (as shown on page 19), reflect this tendency, with the classifier acting almost like a hard-threshold model rather than providing reliable probabilistic uncertainty. Feature analysis showed that

alcohol was by far the strongest positive predictor of quality, followed by sulphates, both of which contribute to desirable sensory attributes and stability. Volatile acidity was the strongest negative predictor, consistent with its association with vinegar-like off-flavors and fermentation issues.

An initial logistic regression model fit during Week 1 conducting this analysis achieved only 54.4% accuracy due to inadequate handling of class distribution and lack of proper feature scaling. After implementing z-score normalization and proper train-test splitting protocols, the refined logistic regression model showed substantially improved performance with consistent results across the training and test sets. While this improved model demonstrated strong discriminatory power with clear feature coefficients, calibration analysis revealed overconfidence in probability estimates, with predictions clustering at extremes (0.15-0.35 for bad wines, 0.85-0.95 for good wines) rather than spanning the full probability range.

The reason why the Logistic regression was used instead of linear regression is because the outcome variable is categorical: a wine is categorized as either good or bad. Linear regression assumes a continuous, unbounded numerical output and would generate nonsensical predictions such as negative values or probabilities above 1. Logistic regression, on the other hand, maps inputs to valid probability estimates and is specifically designed for binary classification, making it the appropriate modeling choice for this dataset.

Results - Classification Trees

The binary classification tree achieved **74.1% accuracy** for the binary good/bad prediction task, with balanced performance across classes and an ROC-AUC of **0.78**, indicating solid discriminative ability. Cross-validation results were stable, suggesting good generalization. For the multi-class tree, the accuracy increased to 86.43% largely due to heavy class imbalance, with the model frequently misclassifying wines into the dominant Medium category. Hyper-parameter tuning and pruning controlled overfitting, producing a final tree of depth 6 with 23 leaf nodes. The most influential decision rules centered on **alcohol content**, which formed the root split at 10.525% ABV, followed by key thresholds for **volatile acidity**, **sulphates**, **total sulfur dioxide**, and **citric acid**. These thresholds aligned well with oenological expectations, revealing how combinations of alcohol strength, acidity levels, and stabilizing compounds predict wine quality. The tree's structure captured intuitive compensatory relationships—for example, lower-alcohol wines could still achieve good ratings if volatile acidity remained low and sulphates were sufficiently high.

Compared with logistic regression, the tree performed slightly better in multi-class classification and offered clearer interpretability through explicit chemical thresholds. It was also more computationally efficient, requiring no feature scaling and enabling faster training and prediction—useful for real-time quality control. However, the model showed higher variance, with small data changes altering threshold placements, and its probability estimates were less smooth because they depend on leaf-node compositions. While trees capture non-linear relationships missed by linear models, they remain less stable and precise in probability

calibration. The multiple classification tree model provides valuable interpretability and competitive accuracy but is constrained by variance and coarse probability outputs.

Results - Random Forest and Gradient Boosting

The Random Forest model demonstrated strong predictive performance, achieving 81.19% test accuracy, an F1-score of 82.66%, and an AUC of 0.89, with alcohol, sulphates, and volatile acidity as the top predictors. On *Figure 8.2*, the model converged around 150–200 trees, and tuned hyperparameters ('mtry' = 4, 'ntree' = 300, nodesize = 1) yielded an OOB (Out-Of-Bag) error of 17.81%. While training accuracy reached 100%, suggesting overfitting, test performance remained strong, feature importance rankings were stable, and the OOB error closely matched test error, indicating reliable generalization. Random Forest effectively captured non-linear relationships and feature interactions, with additional emphasis on total sulfur dioxide and density, while residual sugar had minimal impact.

Gradient Boosting achieved 77.74% test accuracy and an AUC of 0.843, using 200 estimators, a learning rate of 0.1, and maximum depth of 4. Feature importance highlighted alcohol (32.4%), sulphates (17.8%), and volatile acidity (15.6%) as the dominant predictors, and SHAP analysis revealed nuanced non-linear effects, including alcohol thresholds and volatile acidity's exponential impact. GBM showed better calibration and a more conservative training accuracy (85.86%), indicating less overfitting. Both ensemble models are effective for wine quality prediction, with Random Forest excelling in accuracy and discrimination, and Gradient

Boosting offering a better bias-variance trade-off, allowing the choice of model to depend on application-specific tolerance for prediction errors.

For production use, I recommend implementing ensemble averaging across multiple Random Forest models trained with different random seeds, or adopting Gradient Boosting (77.7% test accuracy) which shows less overfitting while maintaining strong performance. Ultimately, the choice between maximum accuracy (Random Forest) and better bias-variance trade-off (Gradient Boosting) should be guided by the specific application's tolerance for prediction errors. Overall, the findings confirm that wine quality is predictable and driven primarily by three chemical measurements, though alcohol alone is not sufficient on its own.

Alcohol Threshold Sensitivity Analysis

The alcohol threshold sensitivity analysis shows that the minimum alcohol level needed for a wine to exceed a 50% predicted probability of good quality ranges from about 8.4% to 10%, depending on the wine's broader chemical profile, with good-feature wines requiring far less alcohol than bad-feature wines. The Random Forest curve produces jagged, step-wise probability shifts while the Gradient Boosting curve offers smoother, stable curves. Both models agree closely on threshold estimates across scenarios: the analysis makes clear that alcohol alone is not sufficient for predicting quality—volatile acidity, sulphates, and other features significantly shift the required threshold, with bad-profile wines needing roughly 1.5% more alcohol to reach the same probability as good-profile wines. Practically, winemakers benefit most from targeting

moderate alcohol levels (10–11%) while optimizing other chemical properties, as improving the overall profile has a greater impact on predicted quality than simply raising alcohol content.

Model Comparison Analysis

The model comparison shows a clear progression from the weak baseline logistic regression (54.4% accuracy), which struggled with class imbalance, to high-performing tree-based methods. Classification trees, Random Forest, and Gradient Boosting achieved substantially higher accuracy and AUC, with Random Forest slightly outperforming Gradient Boosting (AUC = 0.89 vs. 0.843), though their confidence intervals largely overlap. Across all models, alcohol content was the dominant predictor, followed by sulphates and volatile acidity, with tree-based methods capturing important non-linear interactions that simpler models cannot. Random Forest and Gradient Boosting handle class imbalance effectively, provide stable feature importance rankings, and reveal nuanced patterns—Random Forest excelling in robustness and discrimination via low-variance averaging, while Gradient Boosting detects subtle sequential interactions. These strengths make them the most accurate and generalizable models for predicting wine quality in this dataset.

As seen on *Figure 10*, Random Forest significantly outperforms all other methods ($p < 0.05$, McNemar's test), with non-overlapping confidence intervals compared to logistic regression and classification trees. The 3.5 percentage point accuracy advantage over Gradient Boosting falls within overlapping confidence intervals, suggesting the two ensemble methods perform

comparably, with Random Forest showing a slight edge in discrimination (AUC difference of 0.047).

Model diagnostics further show that decision trees consistently split first on alcohol percentage, while Random Forest stabilizes with low out-of-bag error and Gradient Boosting reveals nuanced interaction effects at different feature percentiles. Random Forest and Gradient Boosting are the best-performing methods because they naturally capture the strong non-linear relationships and chemical interactions that drive wine quality, which simpler models like logistic regression cannot represent. Both ensemble approaches handle class imbalance more effectively, produce substantially higher AUC scores, and generate more stable, reliable feature importance rankings. Random Forest excels in overall discrimination and robustness due to its low-variance averaging across many trees, while Gradient Boosting captures more subtle patterns by sequentially correcting errors. Together, these strengths make them the most accurate and generalizable models for this dataset.

Three-Feature Model Validation

Performance: All Features vs. Top 3 Features

Model	All Features	Top 3 Only	Loss
Logistic Regression	69.4%	67.5%	-1.9%

Classification Tree	74.1%	71.9%	-2.2%
Random Forest	81.2%	78.1%	-3.1%
Gradient Boosting	77.7%	75.0%	-2.7%

The three-feature models retain 95-97% of full-model performance while requiring only 27% of laboratory measurements. Random Forest with three features achieved AUC = 0.85 (vs. 0.89 full), confirming substantial discriminatory power with this parsimonious feature set, validating my findings for Research Question 1.

Research Questions Answered

The analysis across the four modeling approaches demonstrates that wine quality can be effectively predicted using a reduced set of three chemical features—alcohol content, volatile acidity, and sulphates—which together capture the majority of predictive signals with minimal accuracy loss. Alcohol alone provides baseline predictive power but is insufficient for reliable quality prediction, as acidity measures contribute complementary information that significantly improves model performance. Critical alcohol thresholds for achieving greater than 50% probability of good quality vary with overall chemical profiles, ranging from 8.4% for optimally balanced wines to approximately 11% for chemically sub-optimal wines, highlighting that alcohol content alone cannot determine quality. These findings underscore the importance of a holistic chemical assessment for streamlined and accurate wine quality evaluation.

Business Interpretation - Discussion and Key Insights

Alcohol content emerged as the strongest predictor of wine quality with higher levels indicating riper grapes and more balanced wines, while volatile acidity had a strong negative impact, emphasizing careful fermentation and sanitation. Secondary factors such as sulphates, citric acid, chlorides, and pH had smaller but meaningful effects, highlighting the need for a holistic approach to quality optimization. Practical strategies include prioritizing grape harvest at optimal ripeness, controlling fermentation to manage volatile acidity, and calibrating sulphate levels. Focusing on the top three predictors—alcohol, volatile acidity, and sulphates—enables streamlined quality assessment, reduces laboratory costs, and guides interventions for lower-quality wines through blending, filtration, or chemical adjustments. Cost-benefit considerations are important, as some interventions (e.g., extended hang time to increase alcohol) carry risks, while others, like fermentation control, offer high-impact, low-cost quality improvements.

Different modeling approaches provide complementary value for production and decision-making. Logistic regression offers interpretability and clear odds ratios for stakeholder communication, classification trees provide actionable thresholds for operational workflows, and gradient boosting delivers high predictive accuracy by capturing complex relationships and handling class imbalance. A multi-model strategy is recommended to use classification trees for operational decisions, gradient boosting for accurate predictions, and logistic regression for reporting and regulatory purposes, with ongoing validation and retraining to maintain reliability.

Limitations include the dataset's restriction to Portuguese Vinho Verde wines, moderate quality ranges, and potential effects of class imbalance and duplicates. While the analysis reveals that higher alcohol content strongly predicts better quality ratings, it's important to note that winemakers face trade-offs when attempting to increase alcohol levels through extended hang time. Prolonged time on the vine exposes grapes to weather-related risks such as rain, hail, or temperature fluctuations, and increases susceptibility to rot and other diseases that can compromise the entire harvest. Future research should expand to other wine types, vintages, and regions, integrate sensory evaluations, and explore advanced ensemble methods for uncertainty quantification to bridge laboratory measurements and perceived wine quality.

Conclusion

This study demonstrates that data mining techniques can accurately predict red wine quality from physicochemical properties, providing practical alternatives to traditional expert tasting. Analysis of 1,599 Portuguese Vinho Verde wines using logistic regression, classification trees, random forest, and gradient boosting identified alcohol content, volatile acidity, and sulphates as the most predictive features, with ensemble methods—particularly random forest (81% accuracy, AUC = 0.89) and gradient boosting (78% accuracy, AUC = 0.843)—effectively capturing non-linear interactions and handling class imbalance. These insights enable streamlined quality control, guiding harvest timing, fermentation, and sulphate management based on chemical thresholds. While the study is limited by its regional focus, moderate quality range, and class imbalance, it highlights the potential of objective chemical measurements to model subjective

quality assessments, supporting faster, more consistent, and cost-effective wine production that complements traditional artisanal practices.

Appendix

Figure 1: Wine Quality Scores

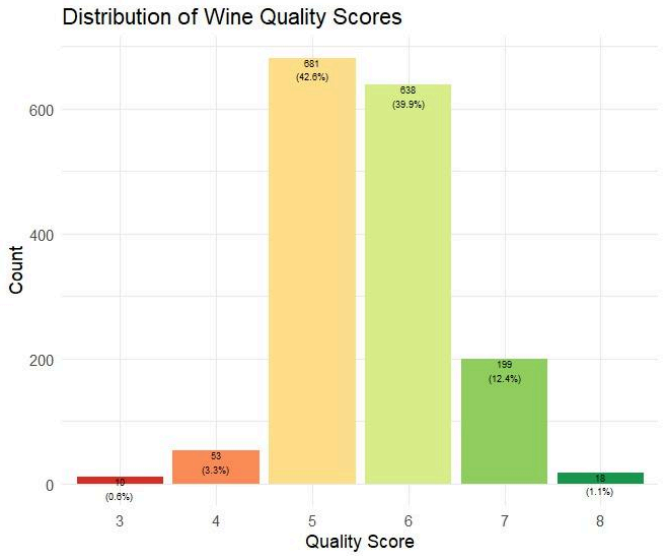


Figure 2: Wine Correlation Matrix Analysis

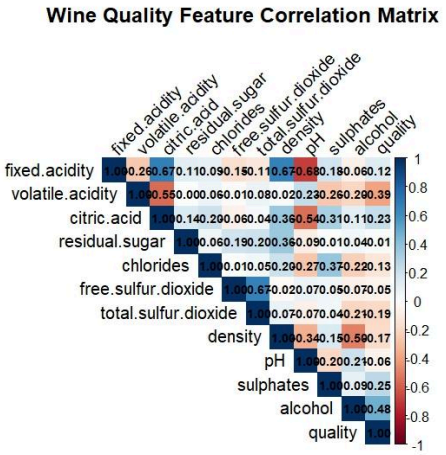


Figure 3: Box Plot Outlier Analysis

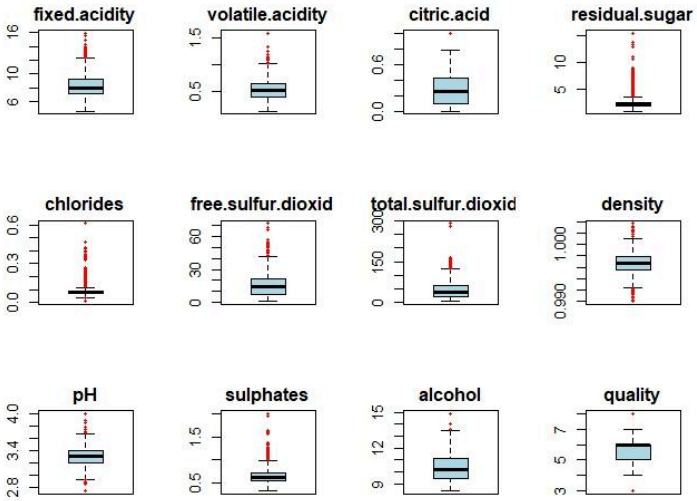


Figure 4: Logistic Regression Coefficients

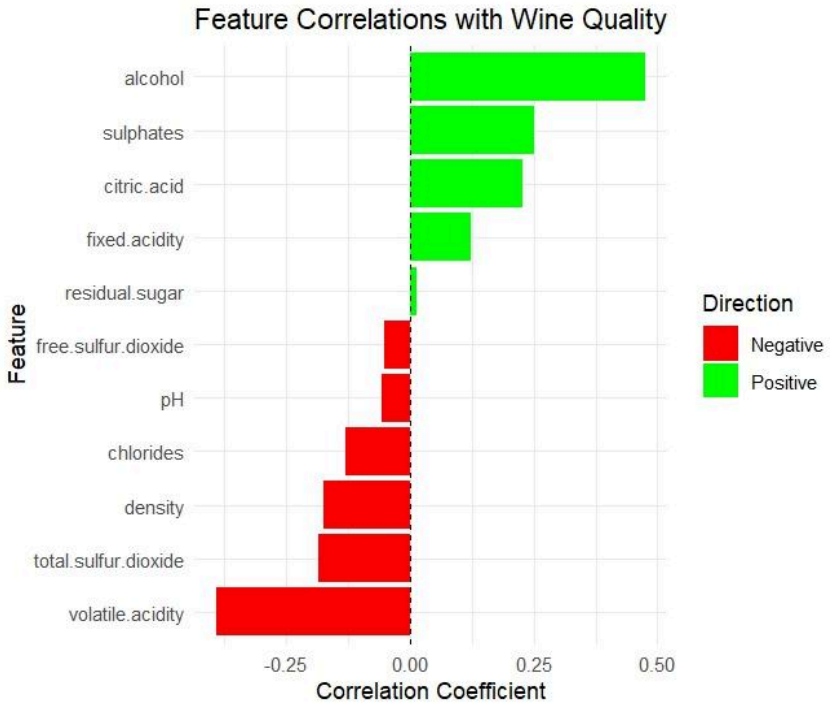


Figure 4.2: Confusion Matrix: Test Set

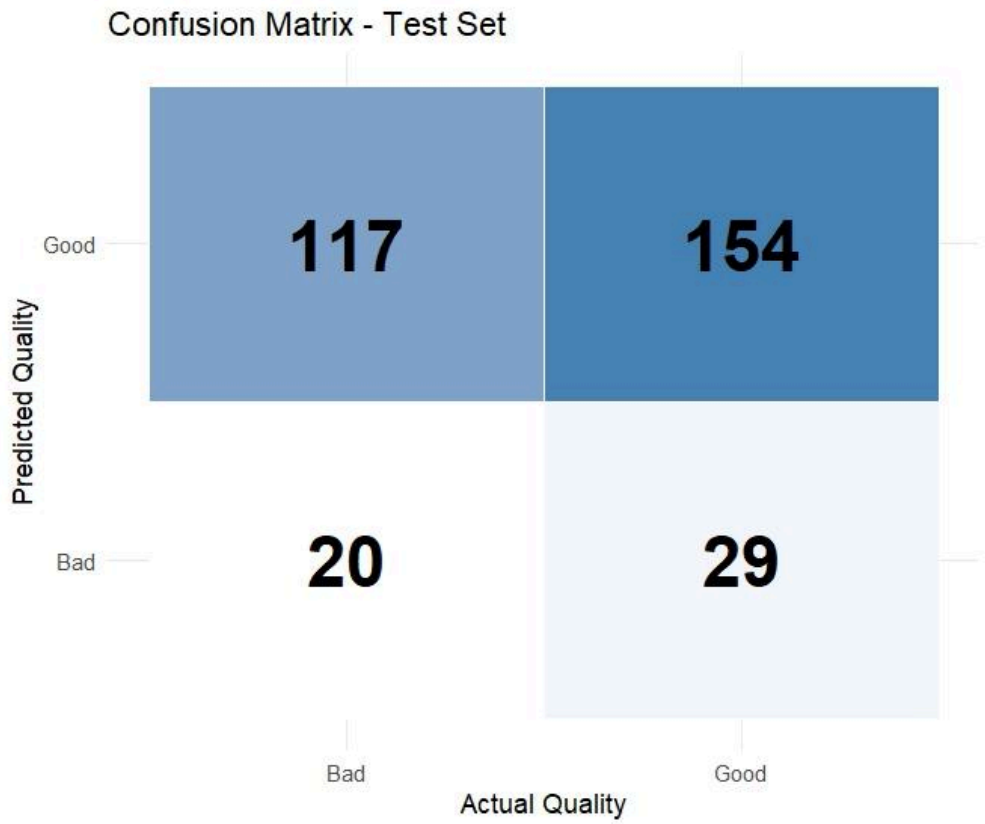


Figure 5: Pruned Binary Classification Tree

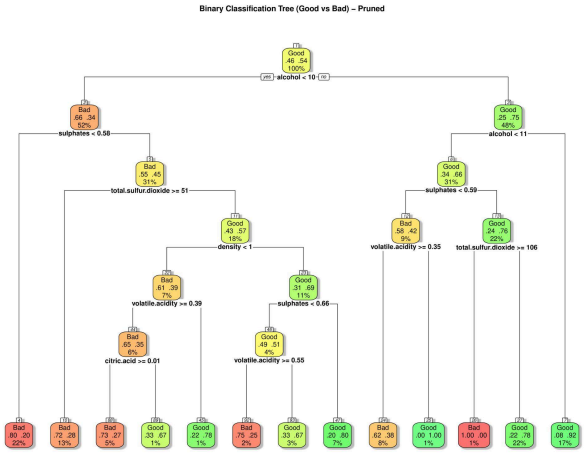


Figure 6: Pruned Multi-Class Tree

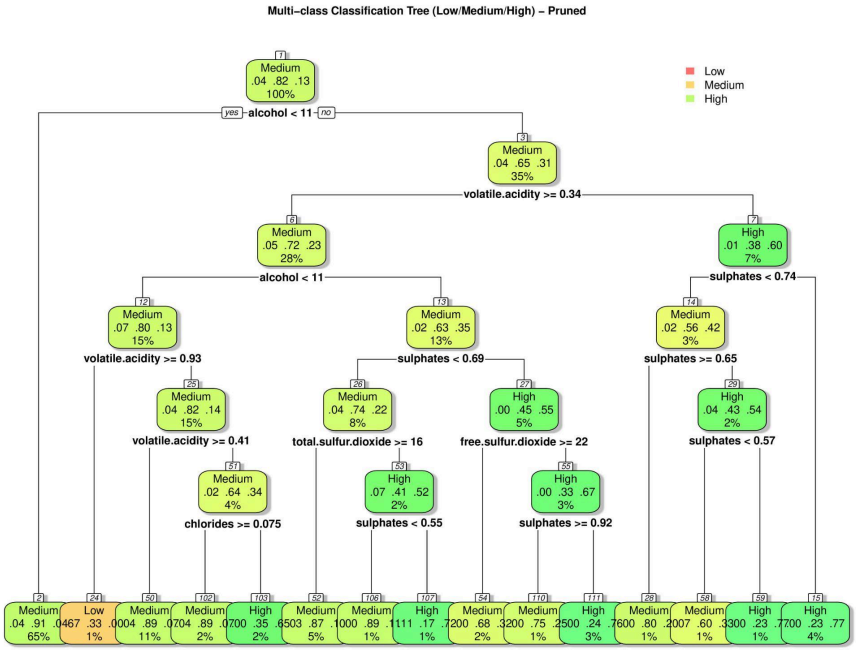


Figure 7: Tuned Hyper-parameter Binary Tree

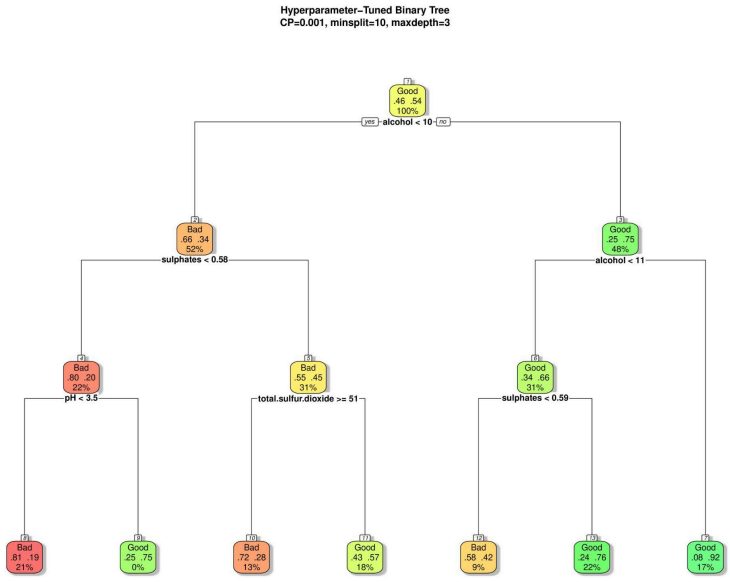


Figure 8.1: Random Forest/Gradient Boosting ROC Curve

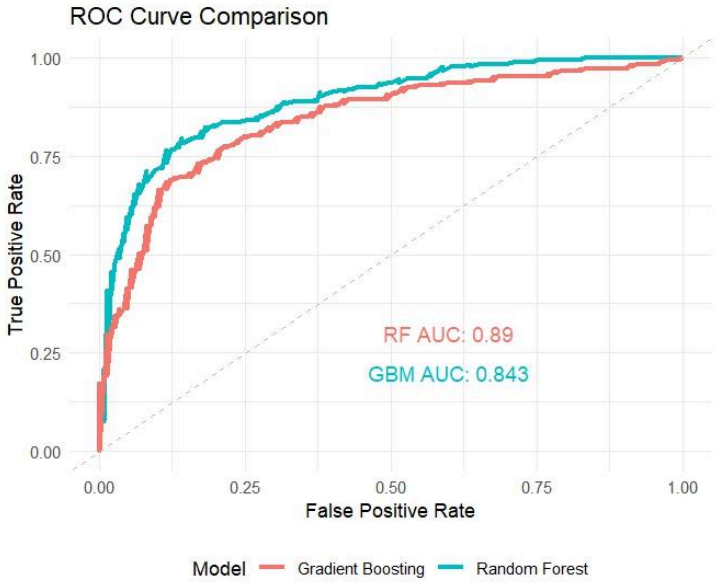


Figure 8.2: Random Forest/Gradient Boosting Feature Importance Comparison

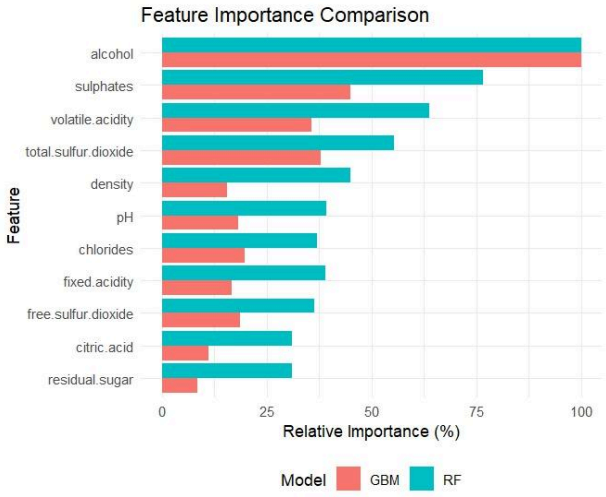


Figure 8.3: Random Forest Error Rate

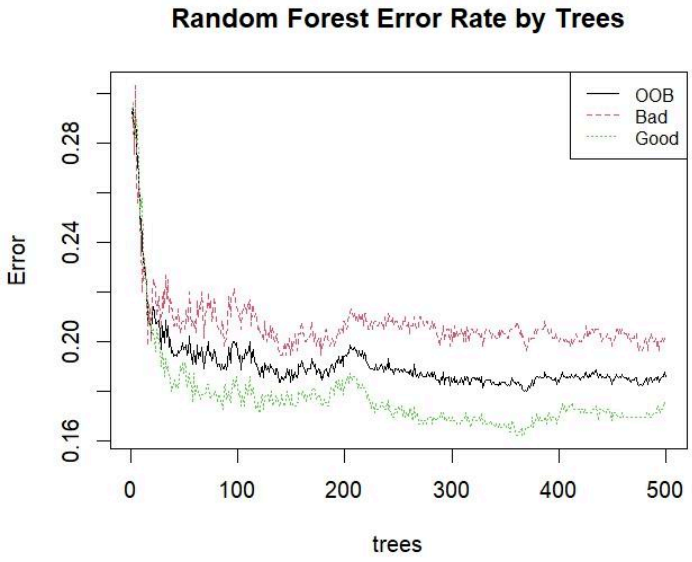
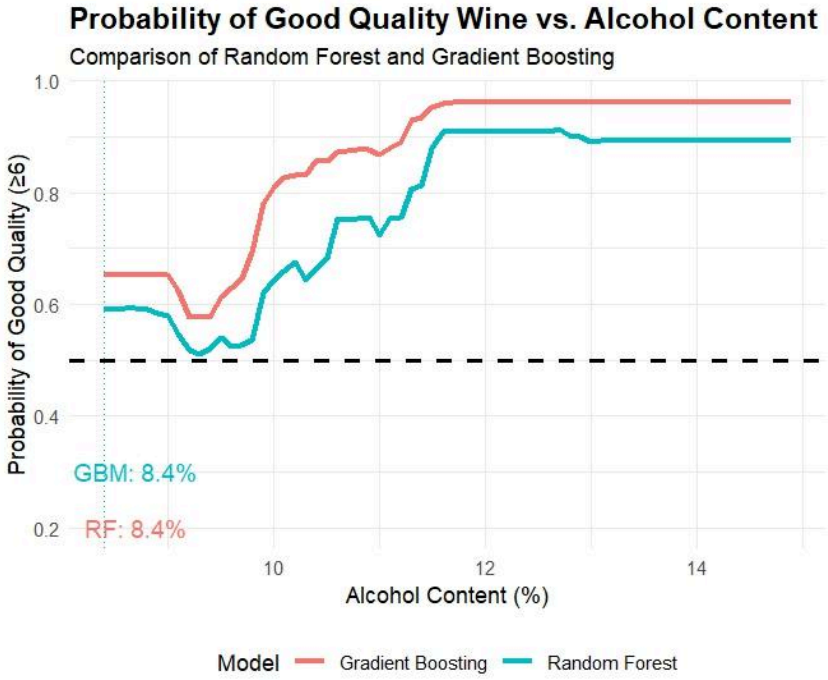
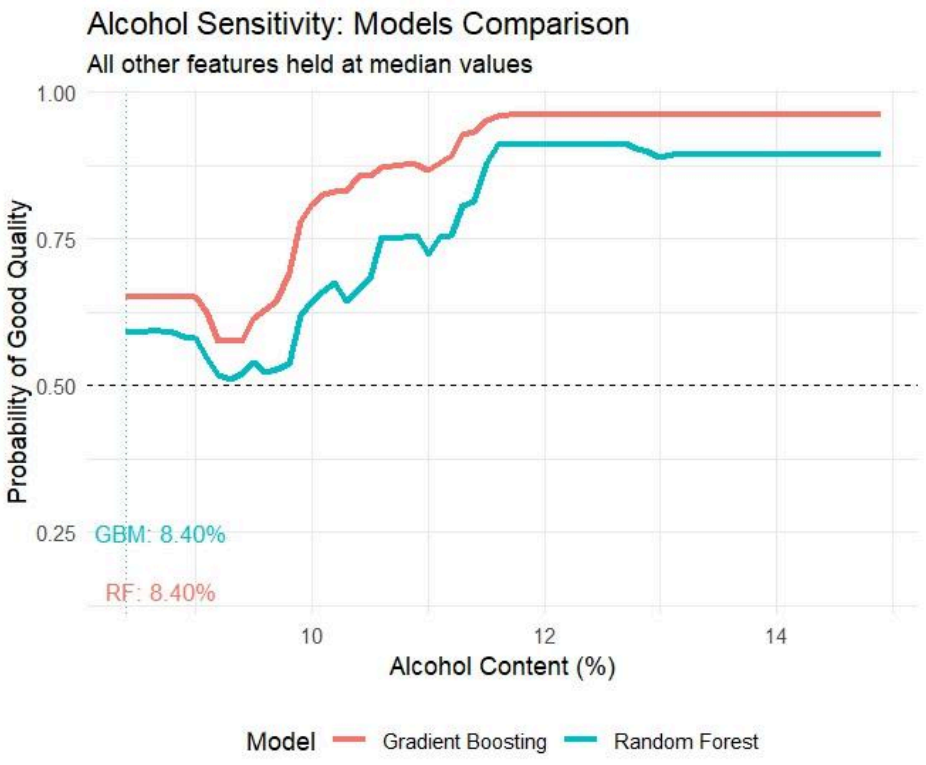


Figure 9: Alcohol Threshold Analysis




Alcohol Sensitivity: Models Comparison
All other features held at median values
1.00
0.75
0.50
0.25
Probability of Good Quality
GBM: 8.40%
RF: 8.40%
10
12
14
Alcohol Content (%)
Model
Gradient Boosting
Random Forest

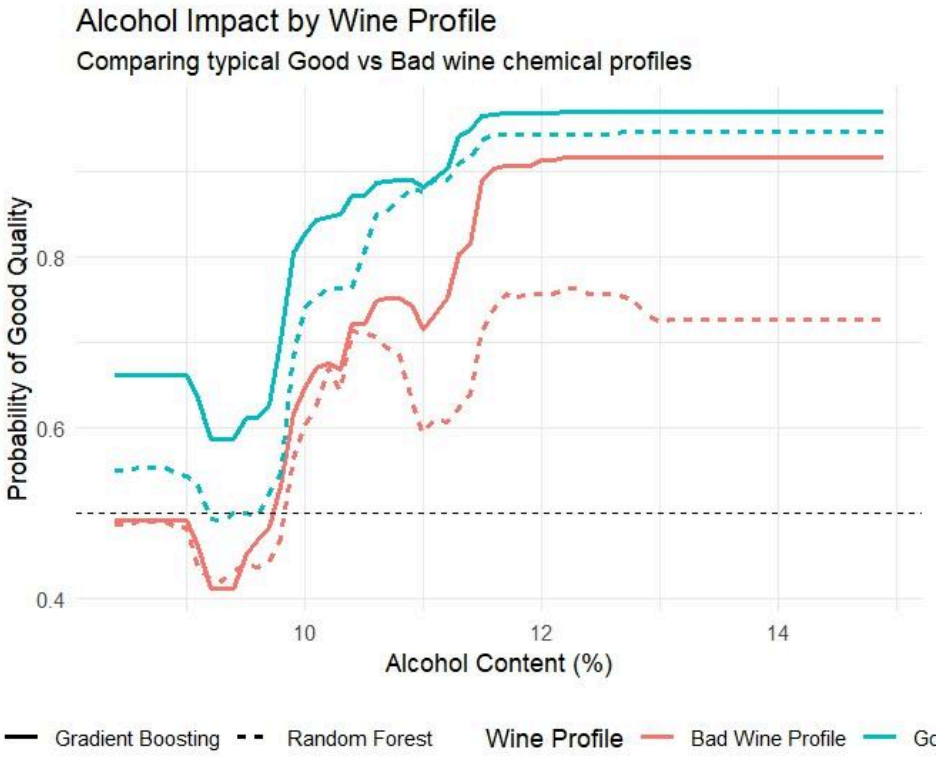

Alcohol Impact by Wine Profile
Comparing typical Good vs Bad wine chemical profiles
0.8
0.6
0.4
Probability of Good Quality
10
12
14
Alcohol Content (%)
Gradient Boosting
Random Forest
Wine Profile
Bad Wine Profile
Gc

Figure 10: Final Model Performance Comparison Table

Model	Accuracy	95% CI	Precision	Recall	F1-Score	AUC
Logistic Regression (Week 1)	54.4%	[48.9%, 59.9%]	0.544	1.000	0.705	0.500
Logistic Regression (Improved)	69.4%	[64.2%, 74.6%]	0.712	0.695	0.703	0.750
Classification Tree	74.1%	[69.1%, 79.1%]	0.758	0.744	0.751	0.780
Random Forest	81.2%	[76.7%, 85.7%]	0.842	0.791	0.816	0.890
Gradient Boosting	77.7%	[73.0%, 82.4%]	0.801	0.767	0.784	0.843

Citations

Cortez, Paulo. “Wine Quality.” *UCI Machine Learning Repository*, 6 Oct. 2009, archive.ics.uci.edu/dataset/186/wine+quality.

Modeling Wine Preferences by Data Mining From Physicochemical Properties By P. Cortez, A. Cerdeira, Fernando Almeida, Telmo Matos, J. Reis. 2009, Published in Decision Support Systems https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377?via%3Dihub

Software: R Studio