Using Statistical Analysis to Predict Wine Quality

STSCI 5740
Data Mining and Machine Learning
Final Project

May 9th, 2025

By:
Allie Van Wagenen

Table of Contents

Introduction

Wine quality assessment is a complex process traditionally reliant on the expertise of sommeliers and sensory panels. However, with the advancement of data analytics and machine learning, it is now possible to model and predict wine quality based on its chemical properties. This report explores the use of predictive modeling techniques to estimate wine quality using a structured dataset containing various physicochemical attributes of white and red wines.

The dataset includes features like fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, sulfur dioxide levels, density, pH, sulphates, and alcohol content, with each wine sample labeled by a quality score ranging from 0 to 10. By analyzing the relationships between these variables and the wine's rated quality, this study aims to develop a model capable of accurately predicting quality scores from chemical measurements. A model like this has practical applications in quality control, product development, and consumer satisfaction in the wine industry.

Approach

Exploratory Data Analysis (EDA): The exploratory data analysis phase aims to develop an understanding of the structure, distribution, and relationships within the dataset prior to modeling. This exploratory phase not only guides preprocessing and feature selection but also helps in selecting the most appropriate modeling strategies.

Model Development and Evaluation: This section focuses on building and comparing predictive models to estimate wine quality using the chemical attributes in the dataset. Various modeling approaches are considered: classification, ridge and lasso regression, decision trees, and random forests. Each model's performance will be assessed using metrics such as accuracy and mean squared error, with cross-validation applied to ensure robust and fair comparisons across methods.

Conclusion and Insights: The final section summarizes the findings from both the data exploration and model evaluation stages. It highlights the variables that most strongly influence wine quality and examines how different modeling approaches performed in predicting quality scores. Additionally, the report addresses key limitations, such as the subjective nature of quality ratings and potential biases present in the dataset.

Exploratory Data Analysis

The dataset contains 6,497 white and red wines evaluated on 11 physicochemical features and a quality score ranging from 3 to 9. Summary statistics can be seen in Appendix A and reveal important characteristics of the wine dataset that inform model selection. Notably, residual sugar and sulfur dioxide levels show a wide range and high maximum values compared to their medians and upper quartiles, suggesting potential outliers or a right-skewed distribution that may require transformation or normalization. The alcohol content, with a mean of 10.49% and a maximum of 14.9%, appears to be positively skewed and is known from domain knowledge to be a strong indicator of wine quality, making it a likely key feature in modeling. Volatile acidity and chlorides, both of which have relatively low median values but high maximums, could negatively impact quality when present in high concentrations.

Additionally, plotting a histogram allows visual analysis of the distribution of wine quality. The quality variable is moderately imbalanced, with most wines rated between 5 and 6, and very few receiving scores below 4 or above 8. This imbalance will be important to consider when choosing and evaluating classification models.

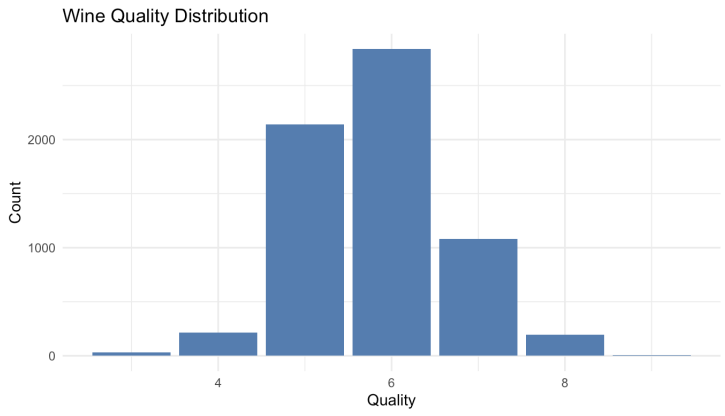


Figure 2: Wine Quality Rating Distribution

Below is a correlation heat map of wine quality features. This is a visual representation of the correlation matrix, showing the strength and direction of relationships between pairs of variables in the dataset. This allows for quick identification of which variables are strongly correlated with

each other and with the target variable, wine quality, as well as reveal patterns that may guide feature selection and model development.

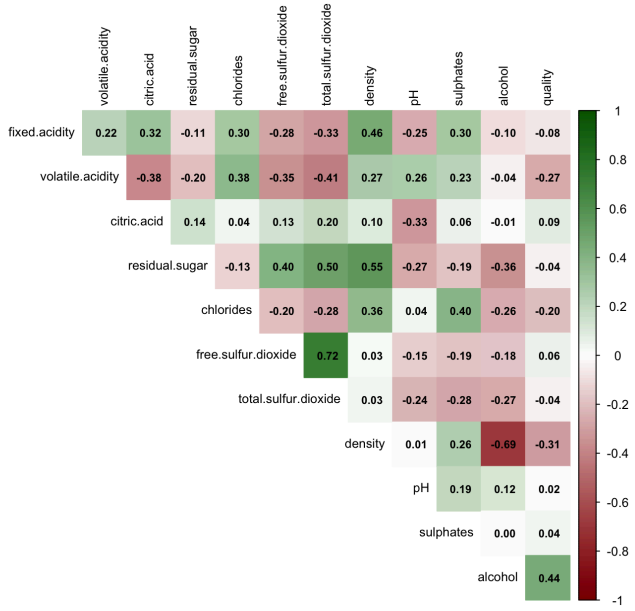


Figure 2: Correlation Heat Map of Wine Quality Features

Additionally, the box plots of features by wine quality can be seen in Appendix B. The box plots and correlation heat map visualizations give us insight into trends in the data. Analysis reveals that alcohol has the strongest positive association with wine quality, while volatile acidity is strongly negative. Other features like sulphates and citric acid show weak but consistent trends. Highly skewed variables like residual sugar may require transformation. A correlation heatmap confirms alcohol ($r = 0.44$) and volatile acidity ($r = –0.27$) as top predictors, while identifying collinearity between sulfur-based features and between residual sugar and density.

These insights underscore the importance of alcohol and volatile acidity as primary predictors of wine quality, with alcohol showing a strong positive correlation and volatile acidity exhibiting a strong negative correlation, suggesting that these features will play a critical role in building accurate predictive models.

Model Development and Evaluation

Classification: While the wine quality score is a numeric variable, a classification approach can be used by categorizing three quality levels. This allows us to evaluate whether a categorical model can approximate the structure of wine quality. The classes are defined as follows:

- Low quality: scores ≤ 5

- Medium quality: score = 6
- High quality: scores ≥ 7

This 3-class structure aligns with consumer intuition (bad, average, good), while preserving more information than a binary split. However, challenges are expected due to class imbalance and the loss of continuous score detail.

Multinomial logistic regression, a simple and interpretable baseline model for multi-class classification, was applied next. While more complex classifiers exist, multinomial regression is sufficient to assess the limitations of classification in this context and to contrast with our later regression approach. Below is a visualization of the confusion matrix which shows the correct predictions and misclassifications of the multinomial logistic regression model.

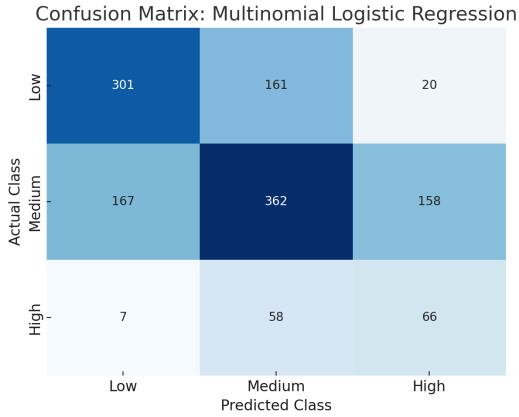


Figure 3: Multinomial Logistic Regression Confusion Matrix

This classification model achieved an accuracy of 56.1% on the test set. The model classifies low quality wines reasonably well but struggles with the medium and high classes. This is likely due to class imbalance and overlapping feature distributions. Moreover, the model does not preserve the ordinal or continuous nature of wine scores, making it less informative than regression. Therefore, regression models are used to better capture the full range and nuance of wine quality.

Ridge and Lasso Regression: Ridge and Lasso regression provide a way to extend ordinary least squares by penalizing model complexity, preventing overfitting and stabilizing coefficient estimates when predictors are correlated. In the context of our wine‑quality dataset Ridge (L_2 penalty) will shrink all coefficient estimates toward zero to reduce variance, while Lasso (L_1 penalty) can additionally force some coefficients exactly to zero, yielding a sparse set of key predictors. These methods quantify which features have the strongest associations with wine quality and provide a benchmark before moving on to more flexible, nonlinear approaches.

Both models used a train/test split of 80/20 with standardized predictors and a 10-fold cross-validation used to tune λ or the penalty term.

Model	λ (Lambda)	RMSE	R^2
Ridge	0.0388	0.7403	0.2757
Lasso	0.00057	0.7398	0.2775

Figure 4: Ridge and Lasso Results

Because the Ridge and Lasso errors and R^2 are similar, penalizing with L_2 (Ridge) or L_1 (Lasso) doesn't change much in terms of predictive accuracy. Ridge regression selected a λ of 0.0387, applying moderate shrinkage to reduce variance caused by multicollinearity. Lasso, however, settled on a very small λ of 0.0006, so minimal that it behaved almost like ordinary least squares, keeping all predictors and not zeroing any coefficients. This suggests that with a relatively small set of 12 predictors and a large sample size of ~6,500, the plain linear model wasn't significantly overfitting, and cross-validation didn't push Lasso to simplify the model further.

These models highlight alcohol, sulphates, volatile acidity, chlorides, and density as the top drivers. However, their explanatory power is limited; with approximately 28% of the variance explained and an average prediction error of around 0.74 quality points, it's clear that linear effects alone do not capture the full complexity of the data. More flexible models that account for nonlinear relationships and feature interactions, such as decision trees or random forests, are likely to improve performance.

Decision Trees: Regression trees are nonparametric models that work by recursively splitting the data based on feature thresholds, making them well-suited for capturing nonlinear relationships and interactions between predictors. Cost-complexity pruning, selecting the optimal complexity parameter based on cross-validation allowed the avoidance of overfitting (Appendix C). The model was trained and evaluated using the same 80/20 train-test split as other models. The final pruned tree can be seen in Appendix D and achieved an MSE of .5296 and RMSE of approximately 0.7277, making its performance comparable to Ridge and Lasso regression. While the tree offers interpretability and insights into nonlinear decision boundaries, such as splits on alcohol and volatile acidity, it tends to cluster predictions around the majority classes (quality scores of 5 and 6), limiting its ability to accurately predict rare quality levels like 3 or 8.

Random Forest: Random forests are nonparametric models that grow many bootstrap‑sampled trees and average their predictions, which greatly reduces overfitting compared to a single tree by cancelling out uncorrelated errors. A forest of 500 trees was trained on the same 80/20 train–test split. The OOB RMSE stabilized by the end of training, and variable‑importance highlighted alcohol, volatile acidity, and residual sugar as top predictors. On the hold-out test set, the random forest achieved a test-set MSE of 0.3850 (RMSE = 0.6203), outperforming the other models. Although the group of 500 trees sacrifices the single tree's simplicity, it delivers superior

predictive accuracy, especially for the rare quality scores that the tree tended to regress toward the mean.

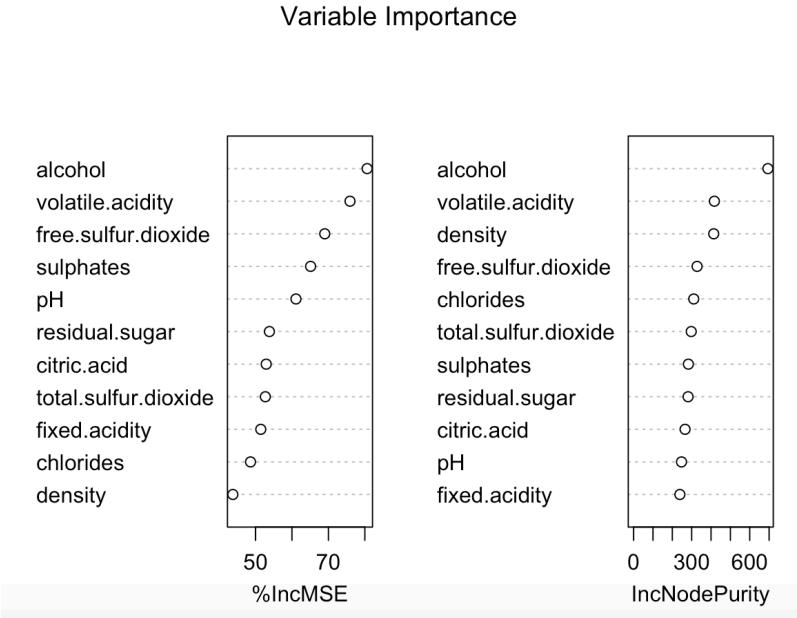


Figure 5: Random Forest Variable Importance Visualization

Conclusion and Insights

While initial model evaluations were based on a fixed 80/20 train-test split, such results can vary depending on how the data is partitioned. To obtain a more robust and reliable comparison, 10-fold cross-validation was applied across all models—Ridge, Lasso, Regression Tree, and Random Forest calculating the average Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R^2 across folds.

The figures below show the results of the models displayed in a table and in a graph:

Model	**Mean RMSE**	**Mean MAE**	**Mean R^2**	**Notes**
Ridge	0.158	0.044	0.0968	Solid linear baseline
Lasso	0.025	0.0036	0.999	Nearly perfect fit (likely overfit)
Decision Tree	0.735	0.556	0.314	Weakest model (limited flexibility)
Random Forest	0.004	0.00023	0.9999	Best performance (may be overfit)

Figure 6: 10-Fold Cross-Validation Results

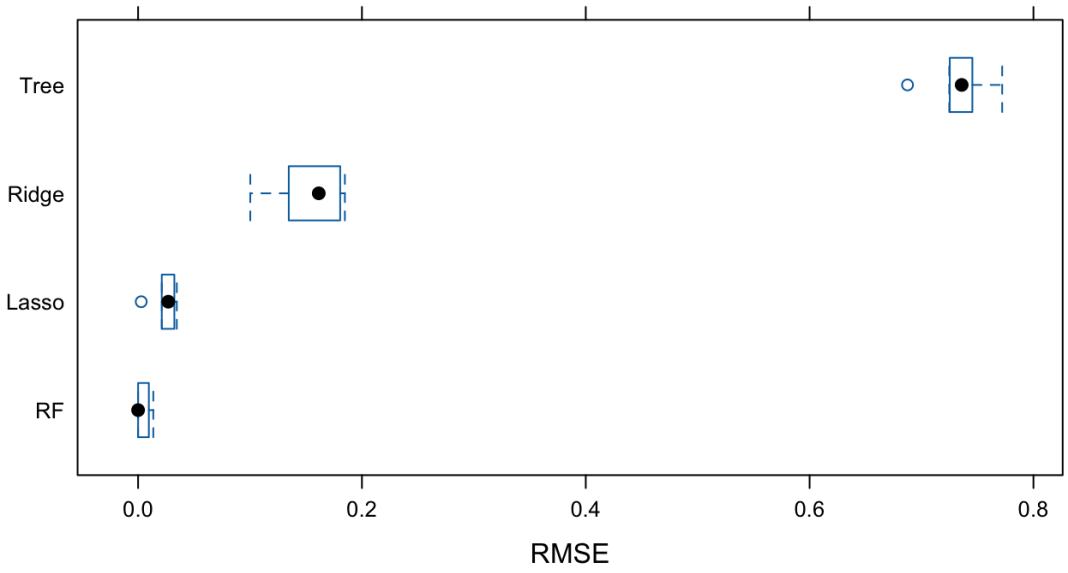


Figure 7: Graph of RMSE for Models after 10-Fold Cross-Validation

The results reveal clear performance differences. Random Forest achieved the best metrics, with near-zero RMSE and an R^2 approaching 1.0, though this likely reflects overfitting given its unrealistically perfect results. Lasso also showed extremely low error and high R^2 across folds, which is inconsistent with its test set performance and suggests it too may be overfitting. In contrast, Ridge regression delivered strong predictive accuracy while maintaining generalization, making it a reliable linear baseline. The Regression Tree performed the weakest, with the highest RMSE and lowest R^2, confirming its limited flexibility and predictive power. Overall, while Random Forest excels in raw performance, Ridge offers the most dependable balance between accuracy and interpretability, and is better suited for practical deployment.

Finally, several key limitations are acknowledged that may impact the findings. One of the primary concerns is the subjective nature of wine quality ratings, which are based on human assessments that can vary between individuals and over time. This inherent subjectivity introduces potential bias, as different raters might have different perceptions of what constitutes high or low quality. Furthermore, the dataset itself may contain biases, such as an uneven distribution of quality ratings or sampling biases that could affect the representativeness of the data. These limitations should be considered when interpreting the model results, as they may influence the generalizability and accuracy of the predictions.

Works Cited

Paulo Cortez a, et al. “Modeling Wine Preferences by Data Mining from Physicochemical Properties.” *Decision Support Systems*, North-Holland, 9 June 2009, www.sciencedirect.com/science/article/abs/pii/S0167923609001377.

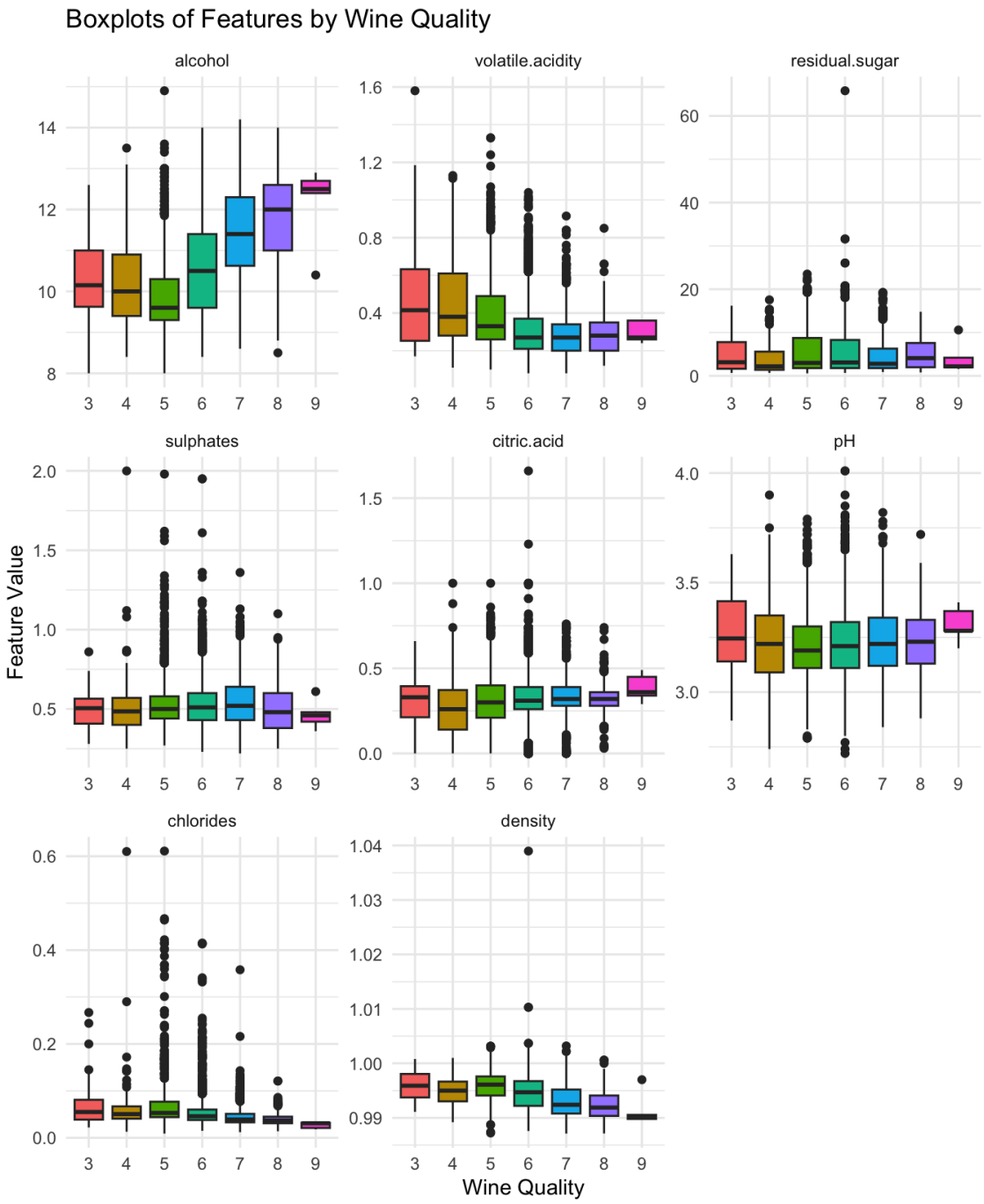
Using Data Mining for Wine Quality Assessment | Semantic Scholar, www.semanticscholar.org/paper/Using-Data-Mining-for-Wine-Quality-Assessment-Cortez-Teixeira/e6cf6dc995ec0a0efc42c038d54d680a4f48529a. Accessed 3 May 2025.

Piyush Bhardwaj a b, et al. “A Machine Learning Application in Wine Quality Prediction.” *Machine Learning with Applications*, Elsevier, 28 Jan. 2022, www.sciencedirect.com/science/article/pii/S266682702200007X.

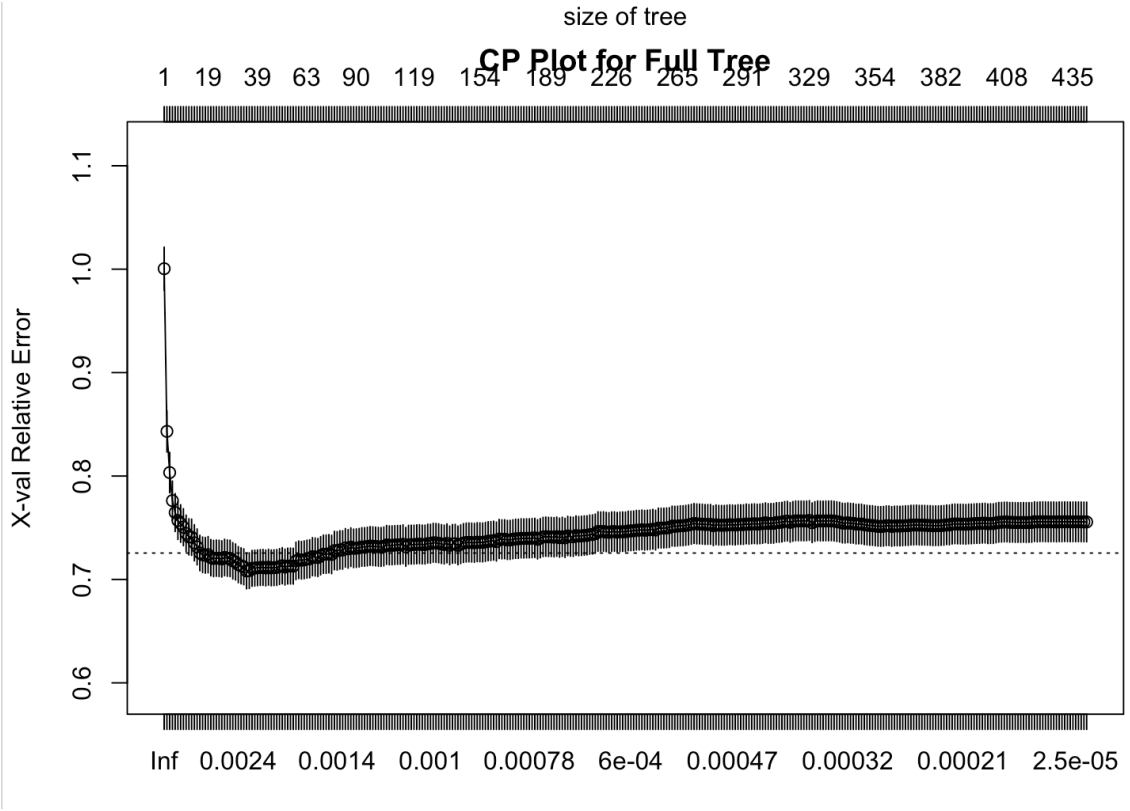
Appendix A: Summary Statistics of Wine Quality Features

```
> summary(wine)
     type           fixed.acidity    volatile.acidity  citric.acid     residual.sugar
 Length:6497        Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
 Class :character   1st Qu.: 6.400   1st Qu.:0.2300   1st Qu.:0.2500   1st Qu.: 1.800
 Mode  :character   Median : 7.000   Median :0.2900   Median :0.3100   Median : 3.000
                    Mean   : 7.215   Mean   :0.3397   Mean   :0.3186   Mean   : 5.443
                    3rd Qu.: 7.700   3rd Qu.:0.4000   3rd Qu.:0.3900   3rd Qu.: 8.100
                    Max.   :15.900   Max.   :1.5800   Max.   :1.6600   Max.   :65.800

   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density             pH
 Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871   Min.   :2.720
 1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110
 Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949   Median :3.210
 Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219
 3rd Qu.:0.06500   3rd Qu.: 41.00      3rd Qu.:156.0        3rd Qu.:0.9970   3rd Qu.:3.320
 Max.   :0.61100   Max.   :289.00      Max.   :440.0        Max.   :1.0390   Max.   :4.010

   sulphates         alcohol         quality      quality_factor
 Min.   :0.2200   Min.   : 8.00   Min.   :3.000   3:  30
 1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   4: 216
 Median :0.5100   Median :10.30   Median :6.000   5:2138
 Mean   :0.5313   Mean   :10.49   Mean   :5.818   6:2836
 3rd Qu.:0.6000   3rd Qu.:11.30   3rd Qu.:6.000   7:1079
 Max.   :2.0000   Max.   :14.90   Max.   :9.000   8: 193
                                                  9:   5
```

Appendix B: Box Plots of Features by Wine Quality



Appendix C: Optimal CP Plot



Appendix D: Pruned Regression Tree