Decision Trees

The CART Decision Tree used to classify numerical Wine Quality data

The UCI Wine Quality dataset provides an optimal foundation for demonstrating the capabilities of the Classification and Regression Trees (CART) algorithm, with several key characteristics making it particularly suitable for decision tree analysis. Unlike simpler datasets or those suited to linear models, this dataset's 1,599 samples and 11 physicochemical properties present complex, non-linear relationships that decision trees are uniquely positioned to capture. While alternative machine learning approaches like neural networks or support vector machines could be applied, they would lack the critical advantage of interpretability that decision trees offer – a crucial consideration when the goal is not just prediction but understanding the relationship between chemical properties and wine quality.

The choice of this dataset is further justified by its ordinal quality ratings (scored from 0 to 10), which naturally align with the hierarchical decision-making process of trees. Where linear regression might struggle with the discrete nature of quality ratings, and neural networks might obscure the decision-making process, decision trees can both handle the ordinal output and provide clear decision boundaries. The continuous nature of the input features (chemical properties) also suits CART's binary splitting mechanism, allowing the algorithm to discover optimal threshold points that have practical significance in wine production.

Beyond accuracy, the decision tree approach offers several distinct advantages for this specific application. First, the binary splitting mechanism provides clear, actionable thresholds for each chemical property, giving winemakers concrete guidelines for quality control. Second, the hierarchical structure naturally reveals feature importance through the ordering of splits, offering insights into which chemical properties most influence quality. Third, unlike black-box models, the tree structure can capture and display complex interaction effects between properties while maintaining interpretability – a crucial factor for practical application in the wine industry.

To rigorously evaluate the model's performance, we formulated and tested three specific hypotheses:

1. H1: Acidity levels would be primary predictors of wine quality due to their impact on taste and preservation

2. H2: Chemical properties would show clear threshold effects across multiple features

3. H3: Higher quality wines would emerge from multiple interacting favorable conditions

The implementation achieved an accuracy of 66% with carefully chosen constraints (max\_depth=5, min\_samples\_split=50). Testing our hypotheses against the results revealed several significant patterns. H1 was strongly confirmed through feature importance analysis: fixed acidity emerged as the most influential feature with 18 decision points, followed by volatile acidity with 13 points. This quantitative evidence definitively establishes acidity as the primary quality determinant. H2 was validated by examining the tree's structure, which revealed clear threshold effects particularly in acidity levels (fixed and volatile) and sulfur dioxide content. H3 was supported by analyzing the decision paths, which showed that high-quality predictions consistently required multiple favorable conditions to be met across different chemical properties.

A bar graph with blue squares

Description automatically generated

The feature importance analysis revealed a clear hierarchy of influence: after the dominant acidity measures (fixed: 18, volatile: 13), we observed moderate influence from citric acid (6 points), sulfates and total sulfur dioxide (5 points each), and lesser roles for alcohol, chlorides, and residual sugar (3 points each), with pH showing minimal impact (1 point). This pattern suggests that the original dataset design by Cortez et al. successfully captured the most relevant chemical properties for quality prediction, though some measured properties proved less influential than others.

The 66% accuracy requires careful contextualization. Given the subjective nature of wine quality assessment and the ordinal nature of the ratings, many model "errors" likely represent predictions of adjacent quality levels rather than severe misclassifications. The chosen tree constraints (depth and minimum samples) represent a conscious trade-off between model complexity and generalization ability, prioritizing robust, interpretable results over potentially overfitted higher accuracy scores.

Looking forward, our analysis suggests several potential improvements. The clear dominance of acidity-related features indicates that future data collection might benefit from more detailed acid profiling. The relatively low importance of some features suggests that a more focused model using primarily acid-related measurements might achieve similar accuracy with reduced complexity. Cross-validation across different vintages could verify the consistency of our identified chemical thresholds, while modifying the loss function to account for the ordinal nature of quality ratings could improve practical accuracy.

In conclusion, this implementation demonstrates both the suitability of the CART algorithm for wine quality analysis and the value of decision trees in providing interpretable, actionable insights. The quantitative feature importance analysis not only validates our hypotheses but also provides practical guidance for quality control in wine production. The clear hierarchy of feature importance, from dominant acidity measures to lesser-used properties, offers valuable insights for both theoretical understanding and practical application in the wine industry, justifying both our choice of dataset and algorithmic approach.

Recent advancements and a real-world application of Decision Trees