

Project Proposal: Predicting Red Wine Quality

By: Taya Bhavsar

I. Background of Data-set

The Wine Quality dataset focuses on red variants of Portuguese "Vinho Verde" wine from northern Portugal. The dataset contains 1,599 samples with 11 physicochemical features (such as acidity, pH, sulfur dioxide, alcohol content, etc.) and a quality rating from 0-10 based on sensory evaluation. This is an imbalanced dataset with more normal-quality wines than excellent or poor ones, making it suitable for both classification and regression approaches.

The Wine Quality dataset contains one dependent variable and eleven independent variables. The **dependent variable** is **quality**, a discrete score ranging from 0 to 10 based on sensory evaluation by wine experts, representing the overall assessment of each wine sample. The **independent variables** consist of eleven physico-chemical properties measured through laboratory testing: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. All predictor variables are continuous measurements with no missing values, providing objective and reproducible data for modeling. The quality score can be analyzed as either a continuous outcome for regression tasks or converted into categories (e.g., low, medium, high quality) for classification purposes, making this dataset versatile for multiple analytical approaches.

Main Objective

I aim to determine which physicochemical properties most accurately predict red wine quality using the UCI Wine Quality dataset, which contains 1,599 Portuguese "Vinho Verde" red wine samples. Each sample includes 11 continuous chemical measurements (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content) and expert quality ratings from 3-8 on a 10-point scale.

Research Questions

1. Can wine quality be accurately predicted using only three chemical measurements?
2. Is alcohol content alone sufficient to predict quality, or are acidity measures necessary?
3. What is the minimum alcohol percentage for a wine to have >50% probability of being rated good quality?

Methodology

I will apply three complementary modeling approaches to predict wine quality:

Logistic Regression will classify wines into binary categories ("Good" quality ≥ 6 vs. "Bad" quality < 6) to identify the top three most significant chemical predictors and determine if alcohol content dominates quality prediction.

Classification Trees will categorize wines into multi-class quality tiers (Low: 3-4, Medium: 5-6, High: 7-8) and identify critical thresholds for key variables like alcohol percentage and volatile acidity that separate quality categories.

Gradient Boosting will leverage the dataset's continuous features, class imbalance, and non-linear relationships to demonstrate effectiveness at feature selection and capturing complex chemical interactions, using both binary classification and potential regression formulations.

Expected Outcomes

What I hope to gain from this analysis is the outcome that it will rank all eleven physicochemical properties by predictive importance, identify optimal chemical combinations for high-quality wines, and provide actionable thresholds for wine production. The dataset's clean structure (no missing values), inherent class imbalance, and documented feature redundancy make it ideal for comparing model performance and demonstrating the strengths of different data mining approaches.

II. Preliminary Analysis

Exploratory Data Analysis: Wine Quality Dataset

This dataset contains 1,599 red wine samples characterized by 11 physicochemical properties and a quality rating. The quality scores range from 3 to 8 on a 0-10 scale, with a mean of 5.636 and median of 6, indicating that most wines in the sample are of average quality. Notably, no wines received ratings below 3 or above 8, suggesting either a filtered dataset or genuine absence of extremely poor or exceptional wines.

The chemical composition reveals several interesting patterns. Fixed acidity averages 8.32 g/L with substantial variation (4.6-15.9), while volatile acidity remains relatively low at 0.53 g/L on average, which is generally favorable for wine quality. The pH levels cluster consistently around 3.31, confirming the expected acidic nature of red wine. Alcohol content shows moderate diversity with a mean of 10.42% ABV, ranging from 8.4% to 14.9%, suggesting the dataset includes various wine styles.

Sulfur compounds display considerable variability, with total sulfur dioxide ranging from 6 to 289 mg/L, though the mean sits at 46.47 mg/L. This wide range indicates some wines may have preservation concerns or measurement anomalies. Similarly, residual sugar levels are generally low (mean: 2.54 g/L), consistent with dry red wines, though the maximum of 15.5 g/L suggests a few sweeter outliers exist.

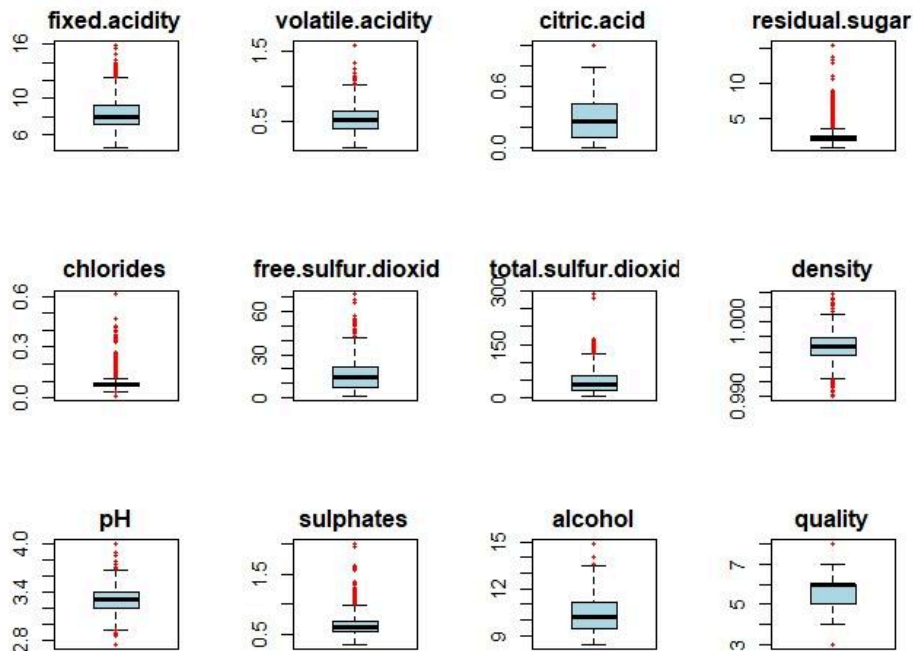
The dataset appears clean with no missing values, though several duplicate observations are present in the sample. Some extreme outliers warrant further investigation, particularly in chlorides (max: 0.611) and sulfur dioxide levels. Future analysis should focus on identifying which chemical properties most strongly correlate with quality ratings and whether outliers represent genuine wine characteristics or data collection issues.

Outlier Detection

The outlier analysis of the red wine quality dataset reveals that several physicochemical properties exhibit substantial outliers, most notably residual sugar, volatile acidity, and sulfur dioxide levels, which reflect natural variations in wine-making practices rather than data errors. Variables like **fixed acidity, citric acid, and chlorides** show moderate outliers that align with expected differences in grape composition and fermentation processes. In contrast, pH, alcohol content, and sulphates demonstrate relatively stable distributions with minimal extreme values. These outliers are oenologically plausible and represent legitimate diversity in wine chemistry stemming

from different production styles, preservation techniques, and intentional wine-making decisions.

Box Plot Outlier Detection



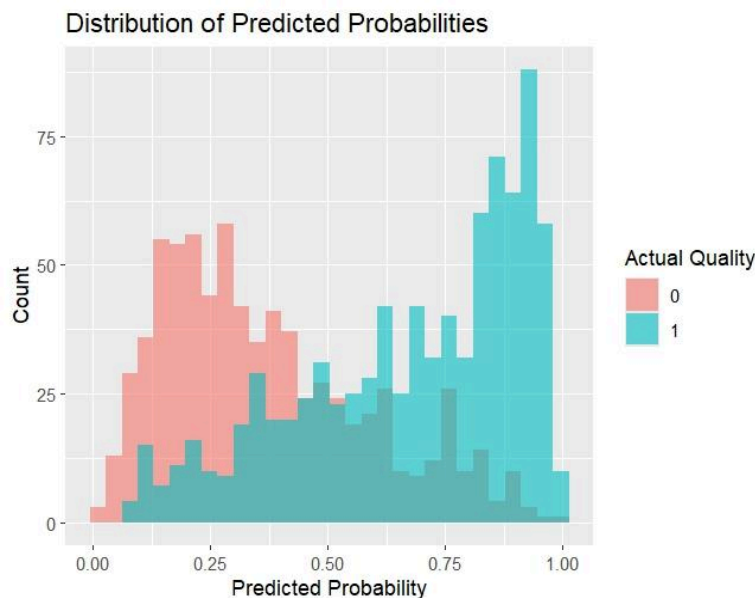
Logistic Regression

The logistic regression model identifies which chemical properties most strongly predict whether a wine will be rated as good quality (score of 6 or higher). The results show that alcohol content has the strongest positive impact on quality—wines with higher alcohol levels tend to be rated better. Sulphates also improve quality, likely due to their role in preservation and flavor enhancement. Conversely, volatile acidity (which indicates vinegar-like flavors) has a strong negative effect, meaning wines with higher acetic acid levels are rated lower. Interestingly, while some sulfur dioxide is beneficial for freshness, excessive amounts hurt quality ratings. Other factors like citric acid and chlorides show smaller but still significant effects. Overall, the model successfully distinguishes between good and poor quality wines based on these chemical measurements, with the dataset fairly balanced between the two quality categories (46.5% poor, 53.5% good wines).

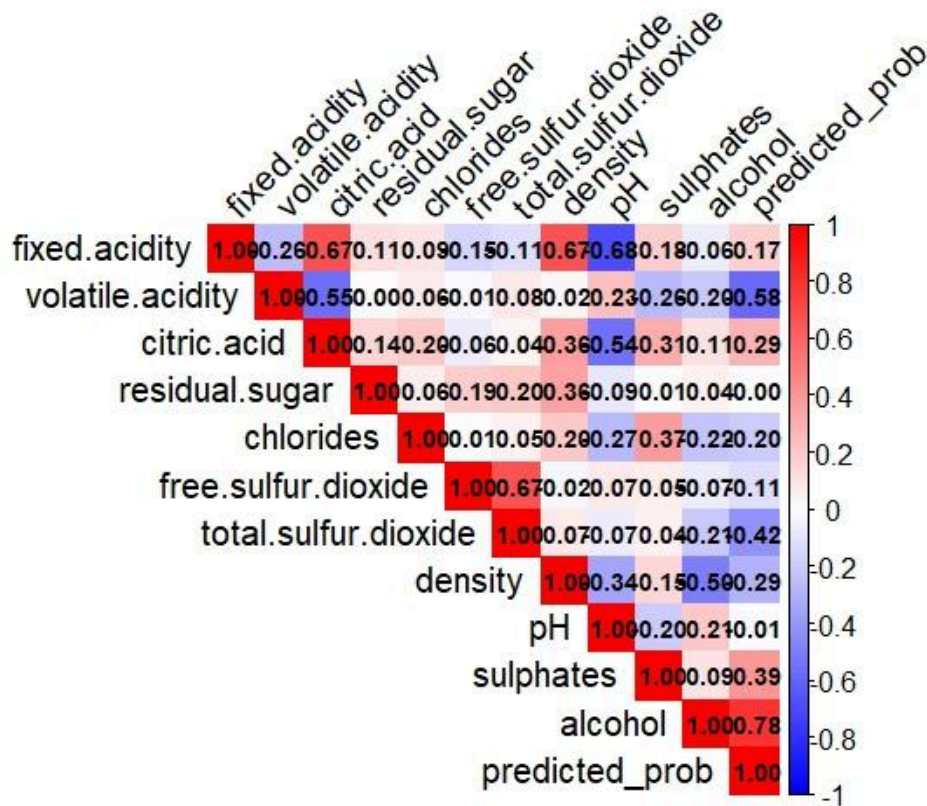
Predicted Probabilities

This calibration plot reveals poor model calibration, with predicted probabilities clustering at two extremes rather than spanning the full range. Low-quality wines (class 0, pink) are predominantly assigned probabilities between 0.15-0.35, while high-quality wines (class 1, teal) receive probabilities concentrated between 0.85-0.95. This bimodal distribution indicates the model is overly confident in its predictions, rarely assigning intermediate probabilities that would reflect genuine uncertainty. While the logistic regression identified several significant predictors—particularly alcohol content, sulphates, and volatile acidity—the stark separation suggests either well-separated classes in the feature space or potential overfitting that produces overly decisive classifications.

This calibration information provides a baseline for comparing logistic regression against boosting and classification trees. Tree-based methods often produce different probability distributions—individual trees generate step-like probabilities due to discrete splits, while boosting can create smoother, more calibrated estimates by combining multiple weak learners. By examining similar calibration plots for these methods, I can assess whether they suffer from the same overconfidence issue or provide better-distributed probabilities across the full range. Boosting may particularly reduce the bimodal clustering by iteratively focusing on misclassified observations, potentially yielding more nuanced probability estimates and revealing whether the extreme separation reflects actual data structure or is merely a logistic regression artifact.



Wine Quality Collinearity Matrix



This correlation analysis of the wine quality dataset reveals several key physicochemical relationships that will inform our modeling approach. The strongest correlations include fixed acidity with citric acid (0.67) and density (0.67), pH with fixed acidity (-0.68), and alcohol with density (-0.50), reflecting expected chemical and physical properties of wine. Notably, most correlations remain moderate (below 0.7), suggesting minimal multicollinearity concerns for predictive modeling. The primary feature clusters—the acidity-pH-citric acid group and the alcohol-density-sugar triangle—indicate where variables share information, which will guide feature selection and model interpretation. The relatively independent behavior of other variables like chlorides and sulphates suggests they may capture unique aspects of wine quality.

III. Proposal of Work towards Final Project

Project Timeline: November 10 - December 8, 2025

Week 1: November 10-16 - Data Preparation and Baseline Models

During the first week, I will complete the exploratory data analysis with visualizations examining key variable distributions, correlations, and outlier patterns. The target variable will be engineered into both binary format (Good ≥ 6 vs. Bad < 6) and multi-class categories (Low: 3-4, Medium: 5-6, High: 7-8). After establishing an 80/20 train/test split, I will implement and evaluate the base-line logistic regression model. This week will conclude with documentation of initial findings regarding the top three chemical predictors and their coefficients.

Week 2: November 17-23 - Classification Tree

The second week focuses on building classification tree models for both binary and multi-class quality predictions. I will perform hyperparameter tuning to optimize parameters such as maximum depth, minimum samples split, and minimum samples per leaf, followed by pruning analysis to balance model complexity and performance. Critical threshold values for alcohol percentage and volatile acidity levels that separate quality categories will be identified and documented. I will create visualizations of decision rules and feature importance rankings, then compare classification tree performance against the logistic regression baseline using appropriate evaluation metrics.

Week 3: November 24-30 - Gradient Boosting Implementation

During the third week, I will implement gradient boosting algorithms to address the dataset's inherent class imbalance using appropriate weighting strategies. Boosting-specific hyperparameters including learning rate, number of estimators, and maximum depth will be systematically tuned for optimal performance. I will analyze feature importance through SHAP values or native importance scores to understand which chemical properties contribute most to predictions. Alternative formulations will be tested, comparing direct regression on quality scores versus classification approaches, with particular attention to the model's effectiveness at capturing non-linear relationships and feature interactions.

Week 4: December 1-7 - Final Analysis and Report

The final week will be dedicated to conducting a comprehensive comparison across all three modeling approaches using metrics including accuracy, precision, recall, and F1-score. I will directly answer the research questions by identifying the minimum 3-feature model, assessing whether alcohol-only prediction is viable, and determining probability thresholds for quality ratings. Final visualizations comparing model performances and feature importance across methods will be created to support findings. The week concludes with writing the final report documenting methodology, results, and interpretations, along with preparing a presentation that summarizes key findings and provides practical recommendations for wine quality prediction. All deliverables will be submitted to the appropriate deadlines, including the clean data-set, trained models, comparative performance analysis, feature importance rankings, and final presentation.

References:

Dataset: Red Wine Quality

Cortez, Paulo, et al. "Wine Quality." UCI Machine Learning Repository, 2009,
<https://doi.org/10.24432/C56S3T>.
<https://archive.ics.uci.edu/dataset/186/wine+quality>

Introductory Paper

[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>