
Classification Models of Wine Quality and Related Parameters

Siyuan Zhang sz264

Data Overview & Variable Distribution (EDA & Variable Exploration)

Dataset essentials

Dataset: red wine quality, 1,599 observations, zero missing values.

Target: wine-quality score (integer, 3–8; 70–80 % of samples cluster at 5–7).

Features: 11 chemical indicators—alcohol, volatile acidity, fixed acidity, citric acid, residual sugar, chlorides, free SO₂, total SO₂, density, pH, sulphates.

Target distribution

Scores 5–7 dominate (bar-plot & box-plot).

Discrete integer scale; median / mode \approx 5–6.

EDA confirms a clean, complete dataset with a clustered quality distribution. Chemical attributes—especially alcohol, acid families (volatile & fixed), and sulphates—are the most quality-informative inputs for downstream modelling.

GLM Analysis

Wine quality is discrete (3–8) but narrow and heavily concentrated at 5–7, behaving “quasi-continuous” ; a Gaussian GLM with identity link is an appropriate first-pass model.

Diagnostics

Residuals deviate from normality (Shapiro $p < 0.01$) and show mild heteroskedasticity (fan-shaped residual-vs-fitted).

Nevertheless, $R^2 \approx 0.36$ and significant coefficients yield high explanatory power.

Suggested figures

Bar chart of standardized GLM coefficients (color-coded by sign & significance)

Residual Q-Q plot & residual-vs-fitted plot

GLM delivers an interpretable linear map from chemistry to quality; alcohol, sulphates and fixed acidity are the most influential levers for winemaking adjustments.

Deep-Learning Model

Wine quality is driven by the joint, non-linear action of many chemical variables; linear models leave interaction effects on the table.

Feed-forward neural networks are universal approximators: they map numeric inputs to (quasi)continuous targets without assuming additivity or monotonicity.

The network automatically learns high-order interactions and non-linear mappings, giving superior fit when multi-dimensional features exhibit complex, opaque relationships to the target.

Random-Forest Model

Ensemble of decision trees : handles high-dimensional, non-linear, interacting predictors with minimal tuning.

Outlier-resistant, distribution-free, and delivers an internal feature-importance score—ideal for identifying which chemical traits drive wine quality.

Target can be treated as multi-class classification (quality cast to factor) or as regression (original integer scale)

Evaluation:

Classification: confusion matrix + accuracy

Regression: MSE / MAE / R^2

Variable importance: Mean Decrease Accuracy (or Gini)

The main reason I chose the Random Forest model is its strong resistance to over-fitting, which offers a good balance between interpretability and predictive power—perfect for small- to medium-sized datasets. My dataset contains about 1,500 observations, fitting that exact scale.

Core Drivers of Wine Quality

Wine-quality scores (3–8, mostly 5–7) are shaped by a handful of chemical traits that show consistent positive or negative associations across modelling approaches.

Model suitability at a glance

GLM: transparent coefficients, best when effects are roughly linear (alcohol, acids).

Deep Learning: captures complex interactions automatically; use when raw accuracy matters more than explanation.

Random forest: resists over-fitting, delivers clear importance ranking; good all-rounder for both prediction and insight.

Need interpretability → start with GLM or RandomForest variable-importance plot.

Need maximum accuracy on large data → consider deep Learning.

Across all methods, alcohol, sulphates and volatile acidity emerge as the “big-three” levers for quality control.

THANK YOU