Wine Quality Classification Report

# Introduction

This project aims to classify red wine quality based on physicochemical properties using machine learning models. The dataset, sourced from the UCI Machine Learning Repository, contains laboratory analyses of **1,599 Portuguese “Vinho Verde” red wines**. Each sample includes **11 numerical attributes** describing its chemical composition and a **quality score** (an integer between 3 and 8) assigned by professional tasters. The task is a **multiclass classification** problem where we predict the quality rating using measurable chemical features.

# Dataset Description

A screenshot of a black screen

AI-generated content may be incorrect.

The dataset’s 11 input variables are objective physicochemical measurements:

* **Fixed acidity, volatile acidity, and citric acid** (g/dm³): Represent total and volatile acid levels, influencing tartness and spoilage.
* **Residual sugar** (g/dm³): Sugar remaining after fermentation. Red wines in this dataset are mostly dry.
* **Chlorides** (g/dm³): Reflect salt concentration; excessive amounts reduce quality.
* **Free and total sulphur dioxide** (mg/dm³): Antioxidants and preservatives preventing oxidation and spoilage.
* **Density** (g/cm³): Indicates sugar and alcohol content; alcohol decreases density while sugar increases it.
* **pH:** Measures acidity strength. Lower pH means higher acidity.
* **Sulphates** (g/dm³): Strengthen antioxidant effects; often correlate positively with quality.
* **Alcohol** (% by volume): A key determinant of flavour balance and body, usually positively associated with quality.

The **target variable** quality is an integer score rated by experts, typically ranging from 3 (poor) to 8 (excellent).

# Exploratory Data Analysis (EDA)

EDA showed that **alcohol** has the strongest **positive** correlation with quality, while **volatile acidity** correlates **negatively**, confirming that high acidity often reduces quality.

A screenshot of a computer

AI-generated content may be incorrect.A chart with different colored squares

AI-generated content may be incorrect.

A graph of a graph

AI-generated content may be incorrect.A graph of a graph of acidity

AI-generated content may be incorrect.**Sulphates** and **citric acid** show moderate positive correlations, and **density** shows a negative one.

A graph with blue and white bars

AI-generated content may be incorrect.A graph of sugar

AI-generated content may be incorrect.

Several features—such as **chlorides**, **residual sugar**, and **sulphur dioxide**—are **right-skewed**, warranting power transformations for variance stabilization.

A diagram of a box plot

AI-generated content may be incorrect.

A diagram of a box plot

AI-generated content may be incorrect.

A diagram of a box plot

AI-generated content may be incorrect.

A graph of a number of boxes

AI-generated content may be incorrect.

**Boxplots** were used to **compare feature distributions across wine quality levels**, highlighting **trends and differences** between low- and high-quality wines. They also **expose outliers** and reveal how median values shift with quality, providing visual evidence of potential predictors. Class imbalance was evident, necessitating stratified train-test splits and macro-F1 evaluation.

# 4. Baseline Models

Two baseline models were trained:

1. **Multinomial Logistic Regression (scaled)** – Accuracy: 0.59, Macro-F1: 0.28
2. **Random Forest (400 trees)** – Accuracy: 0.68, Macro-F1: 0.41

A screen shot of numbers

AI-generated content may be incorrect.

A chart of different colored squares

AI-generated content may be incorrect.A chart with different colored squares

AI-generated content may be incorrect.

The Random Forest model outperformed Logistic Regression, effectively capturing nonlinear feature relationships and handling skewed inputs. Feature importance analysis highlighted **alcohol, sulphates, volatile acidity, and density** as the most influential predictors.

## 5. Feature Engineering

Two feature-engineering strategies were applied:

**(A) Variance stabilization** using the **Yeo–Johnson transform** on skewed variables (chlorides, residual sugar, SO₂), to normalize distributions and help linear models.  
**(B) Domain-inspired features** created from chemical intuition:

* *sulfur\_ratio* = free SO₂ / total SO₂ (oxidation balance),
* *Total\_acidity* = fixed + volatile + citric acid,
* *acid\_sugar\_balance* = total\_acidity / (residual sugar + 1),
* *alcohol\_density* = alcohol / density.

These new ratios capture wine chemistry relationships known to affect taste and stability.

# 6. Model Re-evaluation

After engineering, both models were retrained:

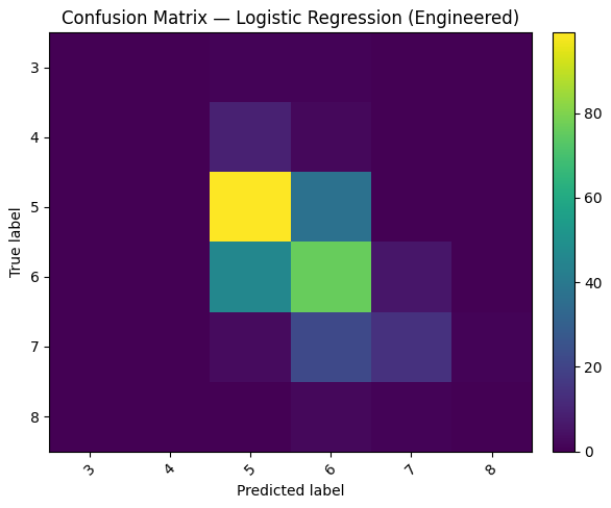
* **Logistic Regression:** Accuracy 0.59, Macro-F1 0.28 → 0.283 (slight F1 improvement).
* **Random Forest:** Accuracy 0.68 → 0.69, Macro-F1 0.41 (stable).

The improvements were modest but consistent, especially for underrepresented classes. The Random Forest remained the best-performing model overall

A screen shot of a computer code

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

# 7. Conclusion

This project demonstrated how physicochemical variables can reasonably predict wine quality. **Alcohol, volatile acidity, sulphates, and density** emerged as dominant quality indicators. **Feature engineering**,bespecially chemistry-inspired ratios, added interpretability and small but meaningful performance gains.

A graph of blue rectangular bars

AI-generated content may be incorrect.