# Predicting Wine Quality Using Machine Learning

# Abstract

This project focuses on predicting wine quality using physicochemical properties with machine learning techniques. Both red and white wine samples were analysed to understand the relationship between these properties and wine quality. Key methods included exploratory data analysis, regression, and classification modelling. Random Forest Classifier and Regressor emerged as the best-performing models, identifying alcohol and density as the most influential features for quality prediction.

# 1. Introduction

Wine quality prediction is an essential task in the winemaking industry, where physicochemical attributes significantly influence consumer perception. This project aims to develop predictive models using machine learning to:

1. Accurately predict wine quality.
2. Identify the key features influencing wine quality.
3. Provide actionable insights for winemakers to improve quality.

The dataset includes red and white wine samples, offering a unified analysis of these wine types.

# 2. Dataset Description

The dataset contains physicochemical properties for red and white wines, with quality as the target variable. Below is a summary:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Type** | **Units** |
| fixed\_acidity | Fixed acidity level | Continuous | g/dm³ |
| volatile\_acidity | Volatile acidity level | Continuous | g/dm³ |
| citric\_acid | Citric acid level | Continuous | g/dm³ |
| residual\_sugar | Residual sugar level | Continuous | g/dm³ |
| chlorides | Salt content | Continuous | g/dm³ |
| free\_sulfur\_dioxide | Free SO₂ content | Continuous | mg/dm³ |
| total\_sulfur\_dioxide | Total SO₂ content | Continuous | mg/dm³ |
| density | Density of the wine | Continuous | g/cm³ |
| pH | pH level of the wine | Continuous | - (scale) |
| sulphates | Sulphates level | Continuous | g/dm³ |
| alcohol | Alcohol content | Continuous | % (v/v) |
| quality | Wine quality score (target variable) | Integer (0-10) | - |
| type | Type of wine (red/white) | Categorical | - |

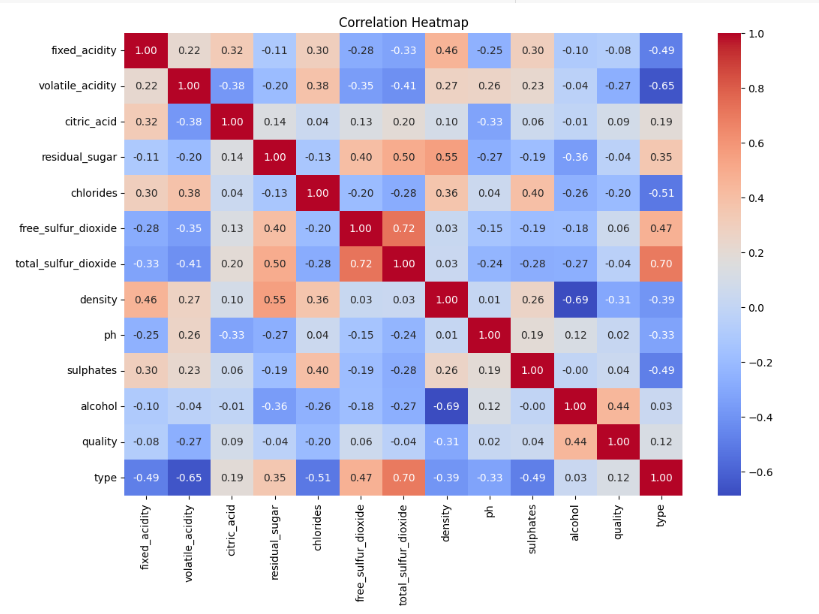
**Key Statistics**:

* Red Wine: 1,599 samples, mean quality = 5.64.
* White Wine: 4,898 samples, mean quality = 5.88.

# 3. Exploratory Data Analysis

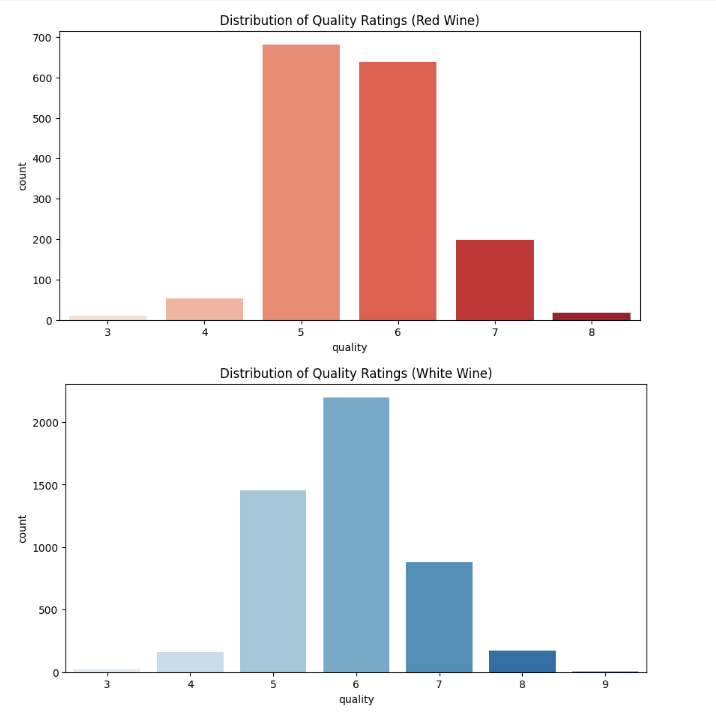
Key findings from the data:

1. **Correlation Heatmap**:



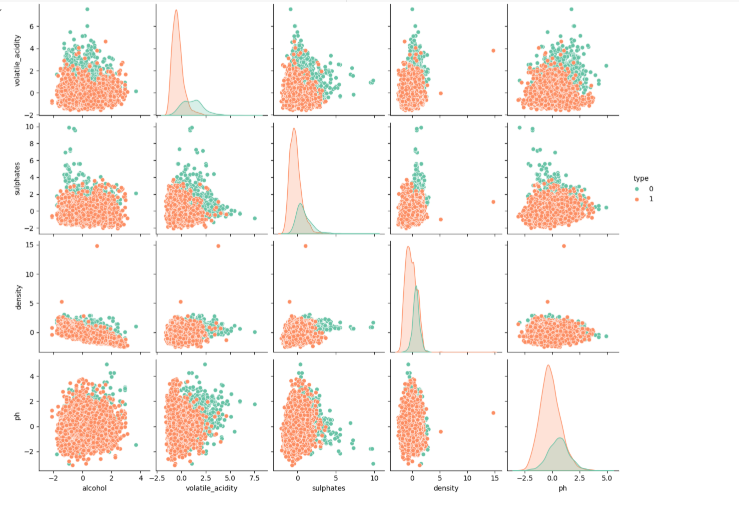
* 1. Alcohol and sulphates have the strongest positive correlations with quality.
  2. Volatile acidity and density have negative correlations with quality.

1. **Distribution**:



* 1. Quality ratings for both wine types are heavily centered around 5 and 6.
  2. Red wine shows a slightly narrower distribution compared to white wine.

1. **Pairplot Insights**:



* 1. Alcohol content and quality demonstrate a strong positive trend.
  2. Density clusters vary distinctly between wine types.

# 4. Methodology

1. **Data Preprocessing**:
   1. Combined red and white wine datasets with a type column.
   2. Standardized numerical features using StandardScaler.
   3. Encoded type as 0 (white) and 1 (red).
2. **Feature Selection**:
   1. Selected features based on correlation and domain knowledge: alcohol, volatile\_acidity, sulphates, density, type, ph, and chlorides.
3. **Modeling**:
   1. Built regression models (Linear Regression, Random Forest Regressor) to predict exact quality.
   2. Built classification models (Logistic Regression, Random Forest Classifier) to categorize quality into low, medium, and high.
4. **Evaluation**:
   1. Used RMSE and R² for regression models.
   2. Used accuracy, precision, recall, and F1-score for classification models.

# 5. Results

1. **Regression Models**:
   1. **Linear Regression**:
      1. RMSE: 0.7362, R²: 0.29.
   2. **Random Forest Regressor**:
      1. RMSE: 0.6453, R²: 0.45 (better fit).
2. **Classification Models**:
   1. **Logistic Regression**:
      1. Accuracy: 0.79, struggles with low-quality prediction.
   2. **Random Forest Classifier**:
      1. Accuracy: 0.84, significantly better precision and recall for all quality levels.
3. **Optimized Random Forest Classifier**:
   1. **Best Parameters**: max\_depth=30, n\_estimators=300, min\_samples\_split=2, min\_samples\_leaf=1.
   2. Accuracy: 0.843.
   3. High recall for medium-quality wines (0.94).
4. **Feature Importances**:
   1. Alcohol and density were the most critical predictors.
   2. Sulphates, volatile acidity, and pH also influenced predictions.

# 6. Discussion

* **Strengths**:
  + Random Forest models excelled in predictive accuracy and robustness.
  + Alcohol emerged as a reliable indicator of wine quality.
* **Limitations**:
  + Low recall for low-quality wines indicates class imbalance issues.
  + Potential improvement through oversampling or weighted modeling.

# 7. Conclusion

* Alcohol and density are the most influential features for predicting wine quality.
* Random Forest Classifier is the best model, achieving an accuracy of 0.843.
* Future work could explore addressing class imbalance and testing on additional datasets.

# 8. References

1. Dataset Source: [UCI Machine Learning Repository - Wine Quality](https://archive.ics.uci.edu/ml/datasets/Wine+Quality).