Wine Quality Classification Project

Objective: Classify wine quality based on chemical properties.

Problem: Wine quality depends on chemical properties.

Dataset Description

The dataset is sourced from Kaggle (Wine Quality Dataset). It contains 1599 rows and 12 columns, including 11 chemical properties and 1 target variable (quality). Features include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. Quality scores range from 0 to 10, converted to binary labels: Good (quality >= 7) and Average-Bad (quality < 7).

Methods

Random Forest Classifier: A machine learning algorithm that builds multiple decision trees and combines their outputs for better accuracy and stability.
StandardScaler: Used to standardize features by removing the mean and scaling to unit variance, ensuring fair contribution of all features to the model.
Libraries: pandas for data manipulation, scikit-learn for modeling, seaborn and matplotlib for visualizations.
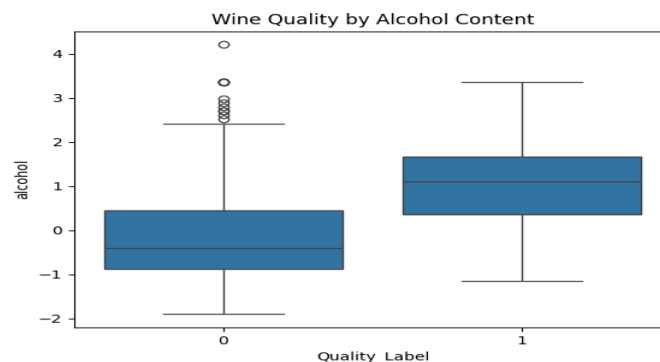
Exploratory Data Analysis (EDA)
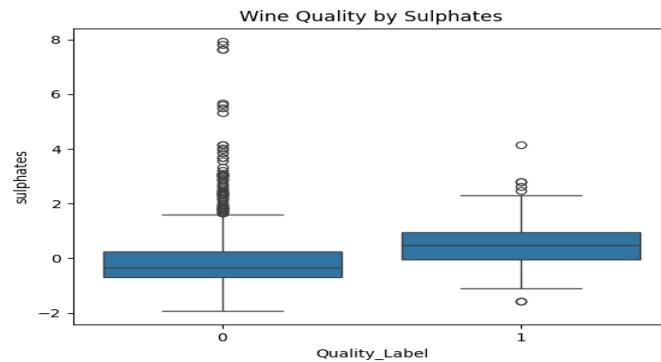
Wine Quality Distribution
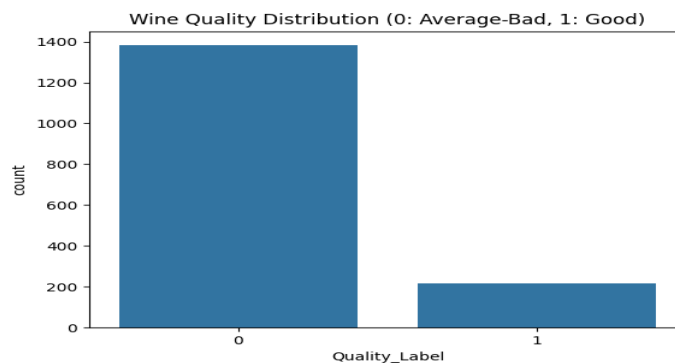
The distribution shows a class imbalance: Average-Bad wines (label 0) are more common than Good wines (label 1).
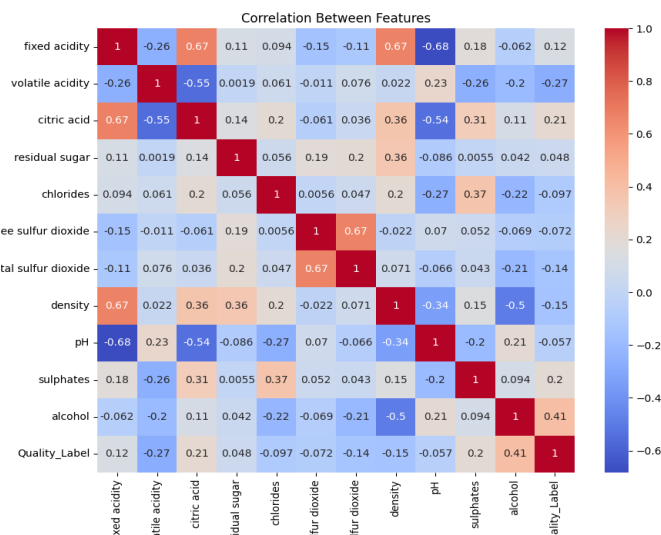
Alcohol Content vs. Quality

Good wines tend to have higher alcohol content compared to Average-Bad wines.

Sulphates vs. Quality

Good wines generally have higher sulphate levels.

Correlation Matrix

Alcohol and sulphates show a positive correlation with quality, while volatile acidity has a negative correlation.
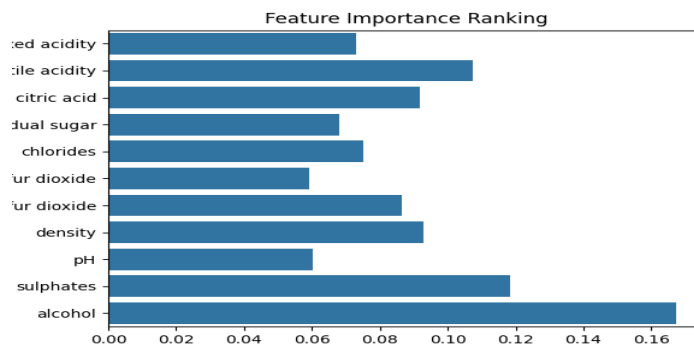
Results

Accuracy: 90%
Classification Report:
Average-Bad: Precision 0.92, Recall 0.97, F1-Score 0.94
Good: Precision 0.73, Recall 0.51, F1-Score 0.60
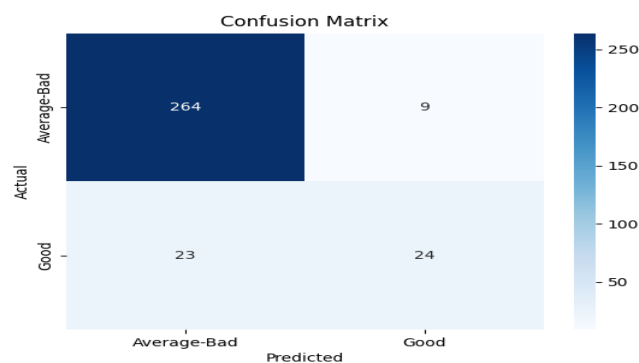Macro Average: F1-Score 0.77
Weighted Average: F1-Score 0.89

Feature Importance

Alcohol, sulphates, and volatile acidity are the most important features for predicting wine quality.

Confusion Matrix

The model correctly identifies most Average-Bad wines but struggles with Good wines due to class imbalance.

Detailed Analysis

The model achieves a high overall accuracy of 90%. However, the recall for the Good class (0.51) indicates that many good wines are misclassified as Average-Bad. This is likely due to class imbalance, as Good wines are underrepresented in the dataset (only about 13% of samples). Techniques like oversampling (SMOTE) or adjusting class weights could improve performance for the Good class.

Recommendations

For Producers: Optimize alcohol content (aim for higher levels) and increase sulphate levels to improve wine quality. Reduce volatile acidity, as it negatively impacts quality.
For Consumers: Choose wines with high alcohol content, higher sulphate levels, and low volatile acidity for better quality.
For Future Work: Address class imbalance using techniques like SMOTE or class weighting.
Explore other algorithms like XGBoost or SVM for potentially better performance.