Wine Quality Prediction using Logistic Regression

1. Introduction

Wine quality assessment is traditionally performed by human tasters, but such evaluations are subjective and expensive. With the availability of physicochemical data of wines, machine learning offers an automated and objective method of predicting wine quality.This project applies Logistic Regression to predict whether a wine sample is of good quality (quality ≥ 7) or not good quality (quality ≤ 6), based on its chemical properties.

2. Dataset Description

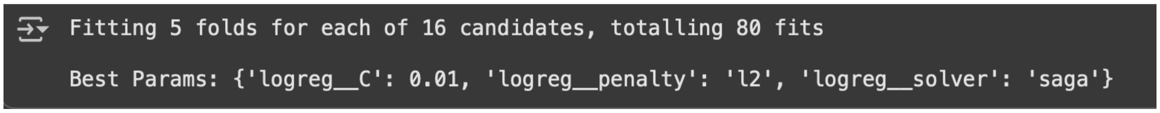
• Source: UCI Machine Learning Repository – Wine Quality Dataset• Variants: Red wine (~1,599 samples)• Features (11 continuous variables): Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol• Target variable: Quality score (0–10), converted into binary label: – Good wine: Quality ≥ 7 → label 1   – Not good wine: Quality ≤ 6 → label 0

3. Methodology

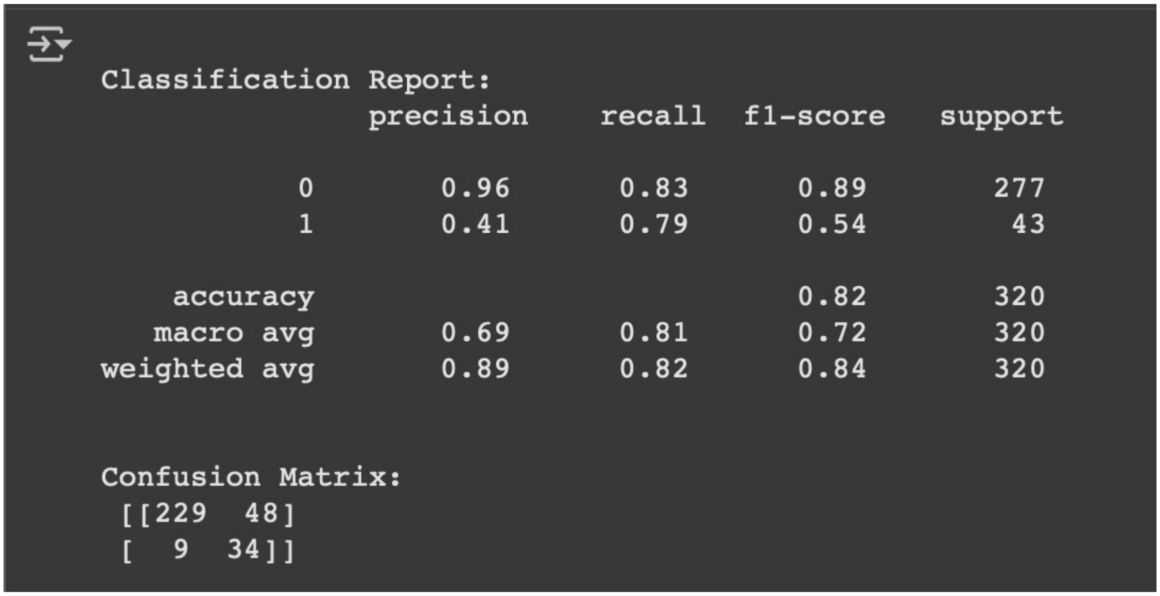
Preprocessing Steps:1. Data Loading: Imported CSV dataset into Pandas dataframe.2. Target Binarization: Converted quality scores into binary labels.3. Exploratory Data Analysis (EDA): Checked distribution of good vs not good wines, visualized correlations using a heatmap, observed imbalance in target classes.4. Train-Test Split: 80–20 split with stratification.5. Feature Scaling: Applied StandardScaler.Model Training:• Algorithm: Logistic Regression (Scikit-learn)• Class Imbalance Handling: Used class_weight='balanced'• Pipeline: Scaling + Logistic Regression• Hyperparameter Tuning: GridSearchCV with parameters: – Regularization strength C ∈ {0.01, 0.1, 1, 10}   – Penalty ∈ {L1, L2}   – Solver ∈ {liblinear, saga}• Evaluation Metrics: Accuracy, Precision, Recall, F1-score, Confusion Matrix

4. Results

Best Hyperparameters: {'C': 0.01, 'penalty': 'l2', 'solver': 'saga'}• Classification Performance:
    – Accuracy: ~82%   – Precision (Good wines): ~0.41   – Recall (Good wines): ~0.79   – F1-score: ~ 0.54• Feature Importance: – Alcohol and Sulphates contribute positively to good quality.   – Volatile acidity negatively impacts quality.



```
Fitting 5 folds for each of 16 candidates, totalling 80 fits

Best Params: {'logreg__C': 0.01, 'logreg__penalty': 'l2', 'logreg__solver': 'saga'}
```

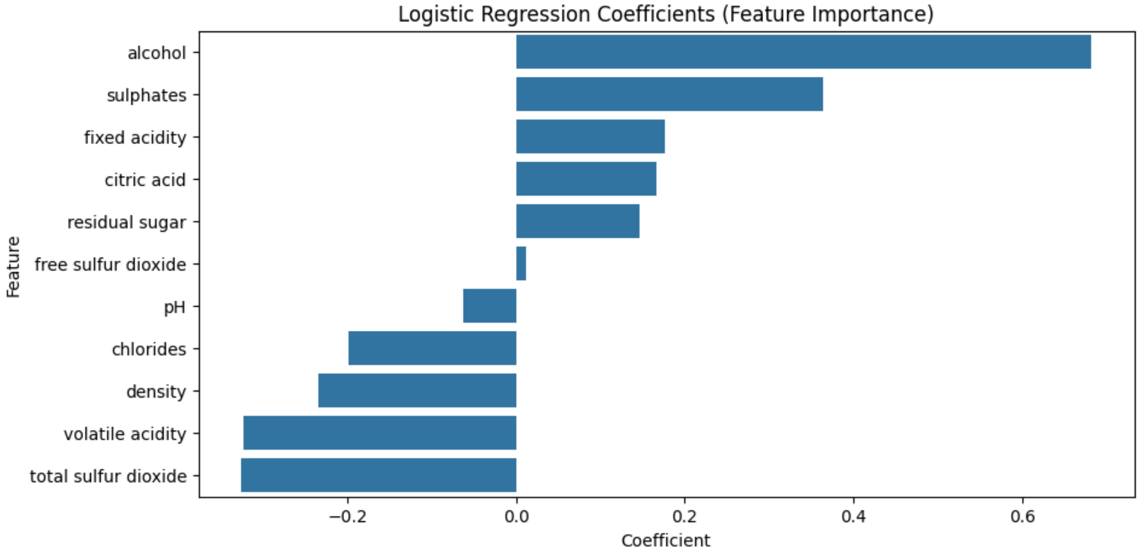


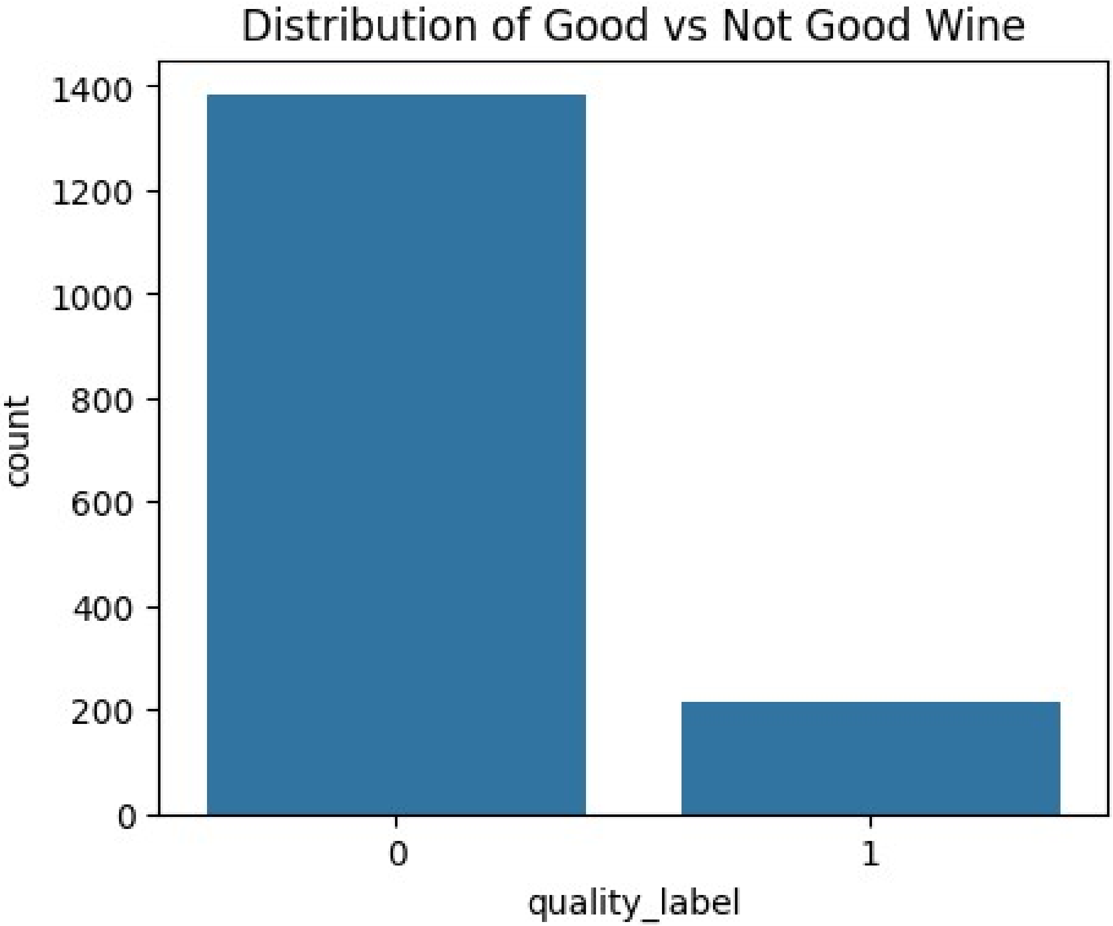
```
Classification Report:
               precision    recall  f1-score   support

           0       0.96      0.83      0.89       277
           1       0.41      0.79      0.54        43

    accuracy                           0.82       320
   macro avg       0.69      0.81      0.72       320
weighted avg       0.89      0.82      0.84       320


Confusion Matrix:
 [[229  48]
 [  9  34]]
```

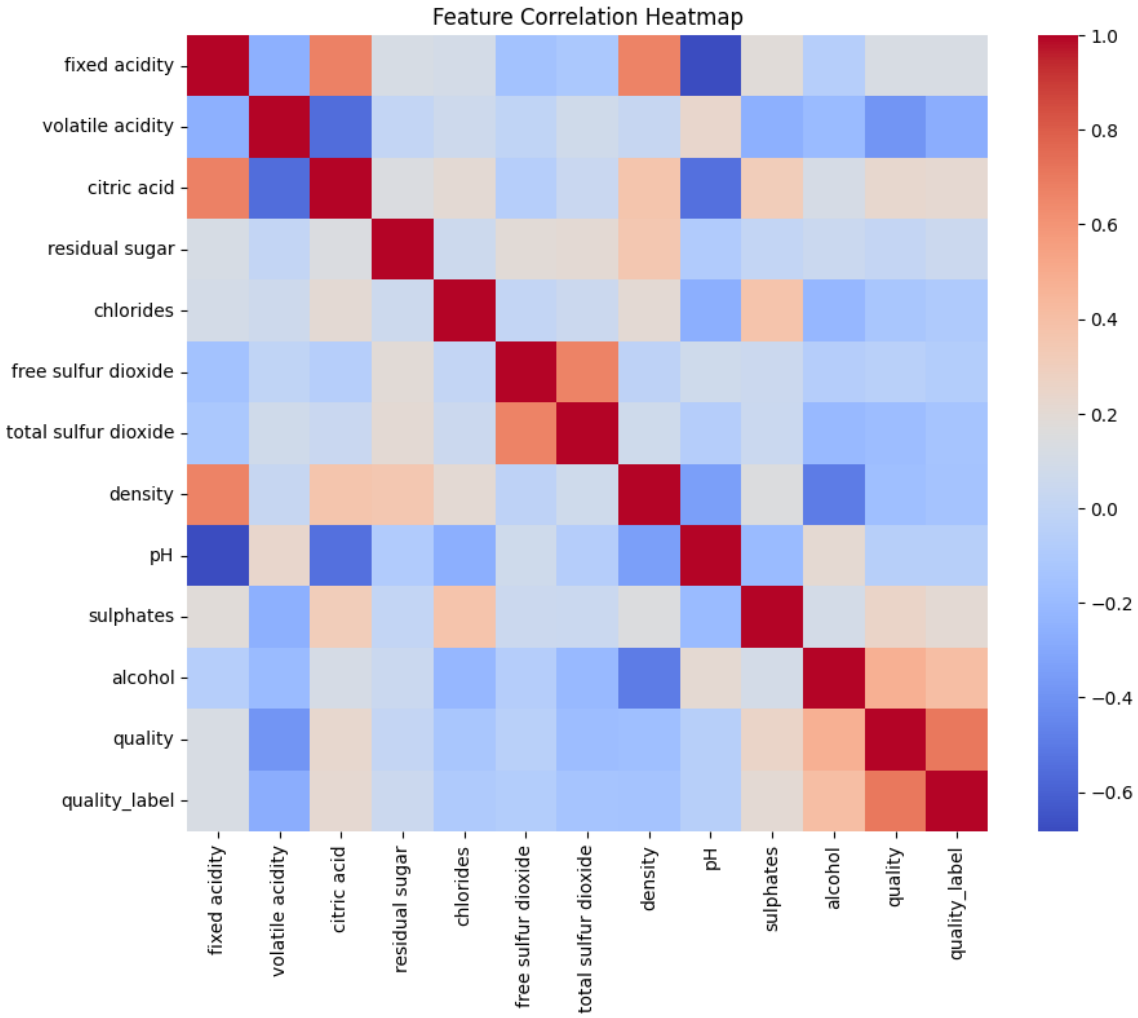
5. Conclusion

Logistic Regression is effective for wine quality prediction, achieving ~82% accuracy. The model highlights key physicochemical properties, such as alcohol content and volatile acidity, as strong predictors.While performance is reasonable, improvements can be made by handling imbalance with oversampling (SMOTE), trying nonlinear models like Random Forest or XGBoost, and applying feature engineering to capture complex relationships.This project demonstrates how Logistic Regression, despite being a simple linear model, provides interpretable and reasonably accurate results for wine quality classification.

6. Results and Plots






Feature Correlation Heatmap
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
quality_label
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6