Wine Quality Prediction Project (Regression + Classification)

1. Project Overview

This project analyzes the Wine Quality dataset using regression and classification models.

We explore model performance, feature importance, and real-world interpretations.

2. Data Understanding

- 6497 rows, 11 numeric features

- Target: quality (integer 3–9)

- No missing values

- Imbalanced distribution (most wines are 5 or 6)

3. Exploratory Data Analysis

- Correlation heatmap reveals alcohol and volatile acidity are strong predictors.

- Boxplots show outliers in sulfur dioxide and residual sugar.

- Target distribution confirms limited high■quality samples.

4. Regression Models

Linear Regression

- MAE: ~0.56

- RMSE: ~0.74

- R^2: ~0.26

Interpretation: Linear models cannot capture non-linear chemical relationships.

Ridge / Lasso Regression

- Similar performance to Linear Regression.

- Regularization improves coefficient stability but not accuracy.

Random Forest Regressor

- MAE: ~0.43

- RMSE: ~0.60

- R^2: ~0.50

Interpretation: Handles outliers and non-linearity effectively. Best regression model.

5. Feature Importance (Random Forest)

Top predictors:

1. Alcohol

2. Volatile acidity

3. Sulphates

4. Free sulfur dioxide

5. Total sulfur dioxide

Interpretation:

Higher alcohol $\rightarrow$ higher quality.

Higher volatile acidity $\rightarrow$ lower quality.

Sulfur compounds influence preservation and taste.

6. Classification Model (Good vs Bad Wine)

Target conversion:

- Good (1): quality ≥ 6

- Bad (0): quality < 6

Logistic Regression Results

- Accuracy: ~73%

- F1 Score (Good): 0.79

- F1 Score (Bad): 0.60

Interpretation:

Good wines are more common; logistic regression predicts them well.

Imbalanced data affects the minority (bad wines).

7. Regression vs Classification Summary

Regression

Best for predicting the exact quality score.

Random Forest achieves $R^2 \approx 0.50$.

Classification

Best for practical categorization: good vs bad.

Logistic Regression achieves ~73% accuracy.

8. Final Conclusion

- Random Forest is the best regression model.
- Logistic Regression performs well for classification.
- Alcohol and volatile acidity are the strongest quality indicators.
- Combining regression and classification provides a full understanding.

This project demonstrates a complete real-world ML pipeline:

EDA $\rightarrow$ Regression $\rightarrow$ Classification $\rightarrow$ Feature Interpretation $\rightarrow$ Insights.