Complete Analysis Report :

1. Data Exploration Insights

- Dataset: 1599 red wine samples, 11 physicochemical features, and quality score (3–8).
- no null values
- There is 240 rows duplicated but if I remove them the evaluation gets lower

- Quality distribution is imbalanced, with most wines rated 5 or 6.

- It was better to do Binary classification for the evaluation.

- Strongest positive correlations with quality:

- Alcohol (+0.48)

- Volatile acidity (-0.39, negative correlation)

- Sulphates (+0.25)

- Weak correlations for residual sugar, free sulfur dioxide, and pH.

- But still with low correlations if I removed these features the evaluation gets lower.

2. Impact of K-Fold Cross-Validation

- Used 5-fold CV on the binary classification task.
- Fold accuracies: [0.7773, 0.7383, 0.7578, 0.7656, 0.7451]
- Average CV Accuracy: 0.7568
- This is slightly lower than the final test accuracy (0.7719), showing that the model generalizes well and is not overfitting.

3. Best Hyperparameters

- Using GridSearchCV (binary classification):
- C = 10
- Gamma = 0.1
- Kernel = RBF
- Best CV Accuracy ≈ 0.7615
- Using RandomizedSearchCV (example, if tested):
- Found similar parameters to GridSearchCV, confirming stability of best settings.

4. Performance Comparison

Model Setup	Accuracy	Precision	Recall	F1 Score
• Baseline SVM (default)	0.7625	0.7686	0.7625	0.7625

- After GridSearchCV tuning 0.8062 0.8089 0.8062 0.8064

5. Class Imbalance Handling

- Quality labels were skewed toward **5 and 6**.
- Binary transformation (≤5 vs ≥6) reduced imbalance and improved model performance.
- Experiments with class_weight=manual_weights improved overall accuracy than in class_weight='balanced'.
- In binary setup, imbalance was less severe and standard SVM performed well without class weights.