

RULE-BASED & STATISTICAL MODELS

IMPLEMENTATION |
VISUALIZATION | EVALUATION

USING
RED WINE QUALITY DATASET

S A L M Mihiliya Jayasiri
23113610

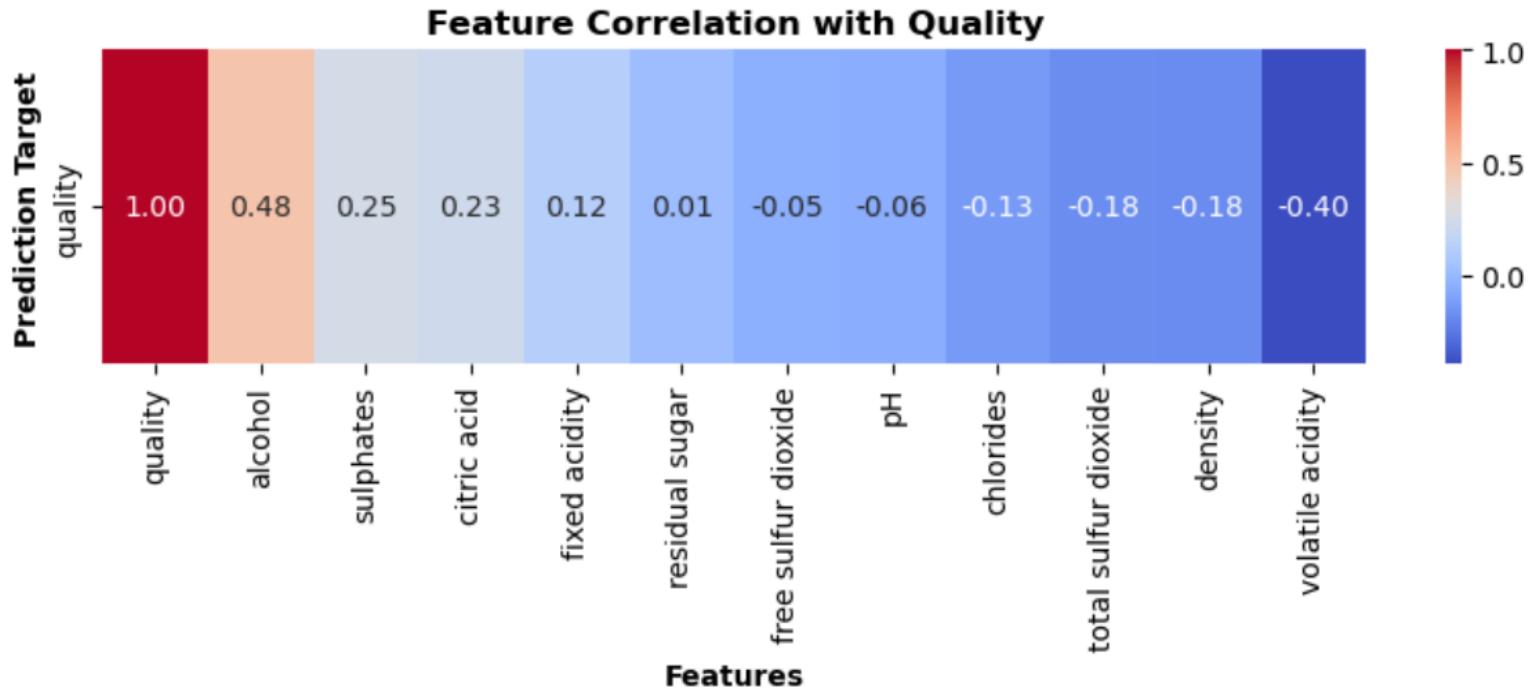


Dataset Selection & Relevance

Wine Quality – Red Wine Variant (UCI Repository)

Dataset:

- **Samples:** 1599 red wine records of Portuguese Vinho Verde.
- **Features:**
 - 11 physicochemical features
 - A sensory quality score (0–10)
- **Target:** Wine quality, rated by experts.



Prediction Relevance:

- Fully numeric and interpretable.
- Ideal for comparing rule –based vs statistical models.
- Clear chemical to quality relationship support predictive modelling.

Key Correlations :

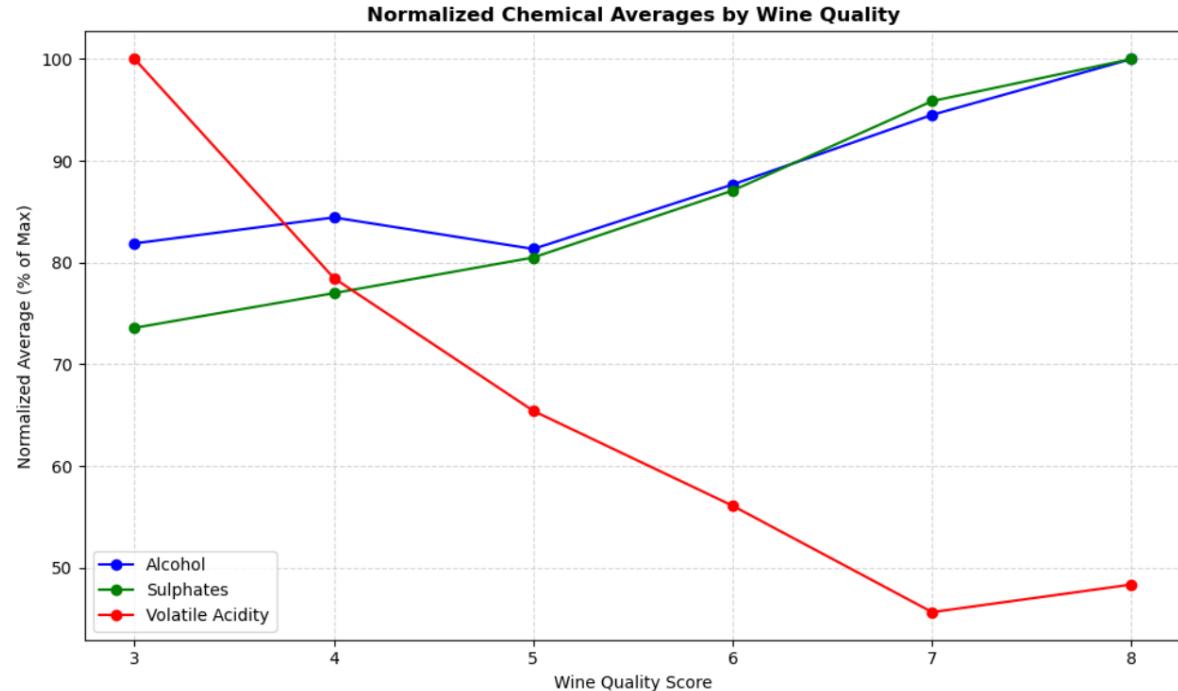
- **Positive:** Alcohol (strongest), sulphates, citric acid.
- **Negative:** Volatile acidity, density, sulfur dioxide.
- **Weak:** Fixed acidity, residual sugar, sulfur dioxide, pH, chlorides

Data Preprocessing

Step	Result
Import	1,599 records, 12 columns
Validation	All numeric
Missing Values	None
Duplicates	240 removed → 1,359 unique rows
Data Types	Float64 / Int64
Exception Handling	Implemented

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000	1359.000000
mean	8.310596	0.529478	0.272333	2.523400	0.088124	15.893304	46.825975	0.996709	3.309787	0.658705	10.432315	5.623252
std	1.736990	0.183031	0.195537	1.352314	0.049377	10.447270	33.408946	0.001869	0.155036	0.170667	1.082065	0.823578
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996700	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.430000	2.600000	0.091000	21.000000	63.000000	0.997820	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Rule-Based Model (Rule Derivation)



Most Correlating Features:

- Alcohol: **+0.48** (strong positive)
- Sulphates: **+0.25** (moderate positive)
- Volatile Acidity: **-0.40** (strong negative)

Interpretation:

- Higher alcohol/sulphates → higher quality
- Higher volatile acidity → lower quality

Rule-Based Model

Threshold Design:

(Based on histograms, and boxplots)

- **Alcohol (A):**

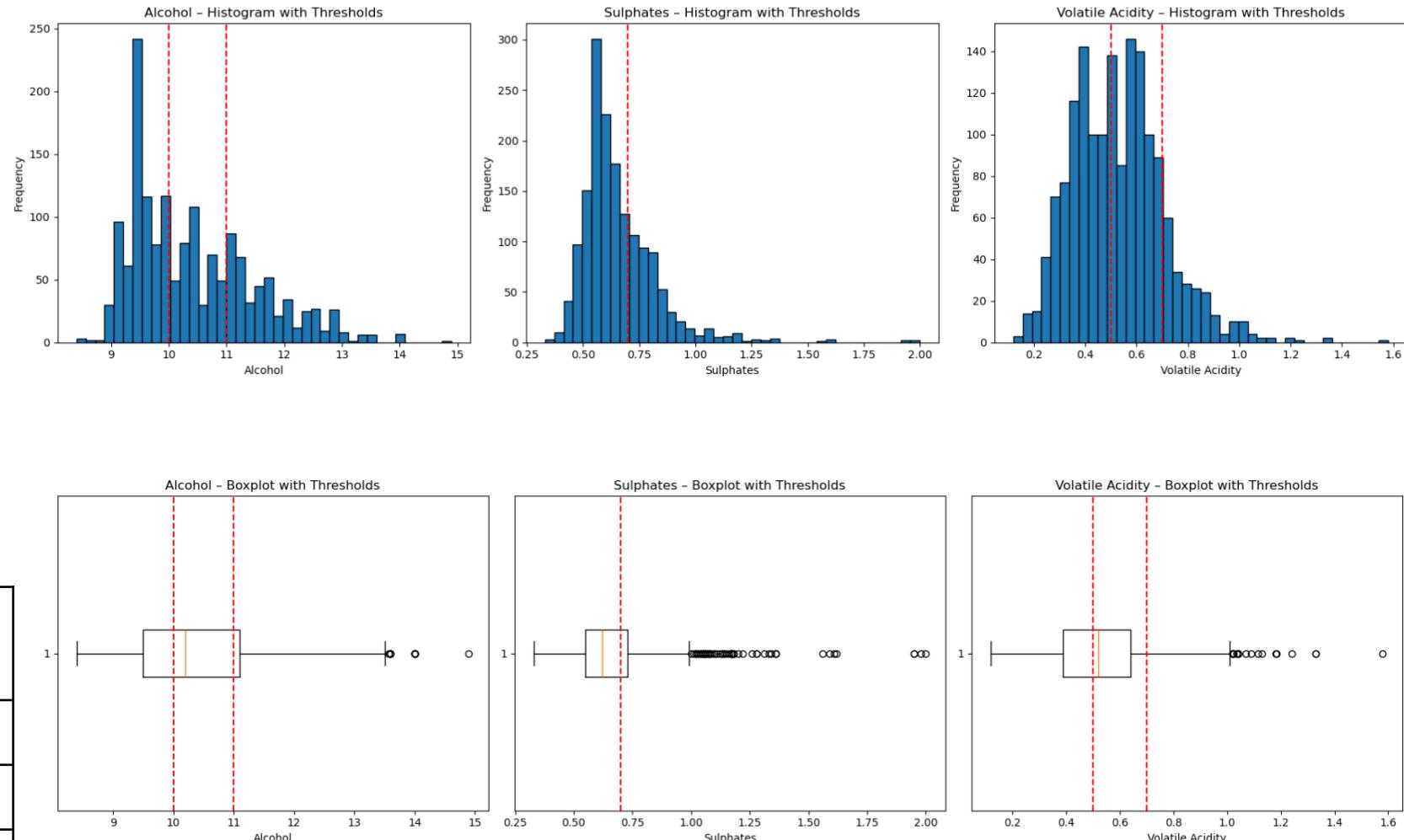
- $11 \rightarrow$ High quality
- $10 - 11 \rightarrow$ Medium
- $< 10 \rightarrow$ Low

- **Sulphates (S):**

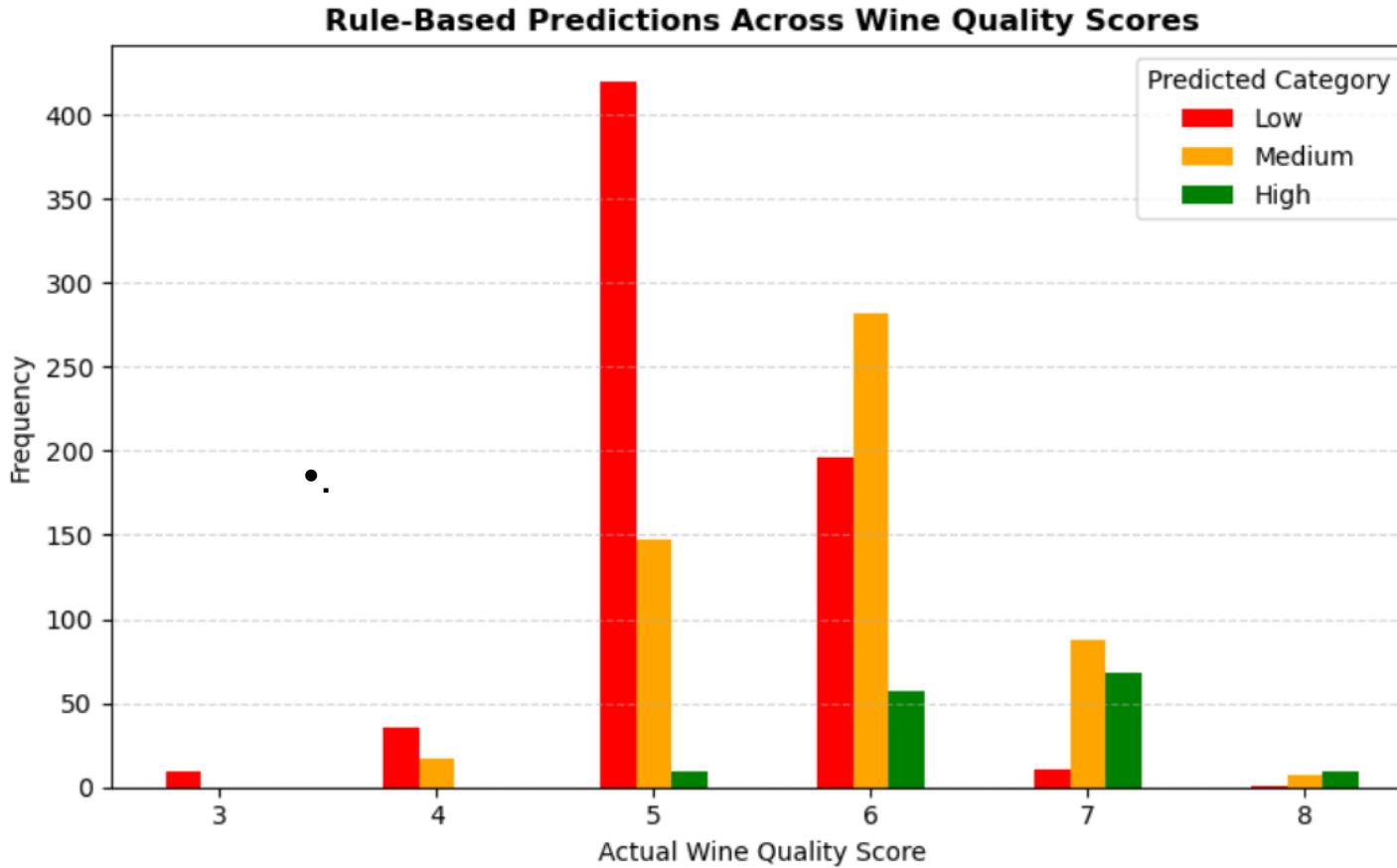
- $0.7 \rightarrow$ Supports high quality

- **Volatile Acidity (VA):**

- $0.7 \rightarrow$ Low quality
- $< 0.5 \rightarrow$ Medium quality

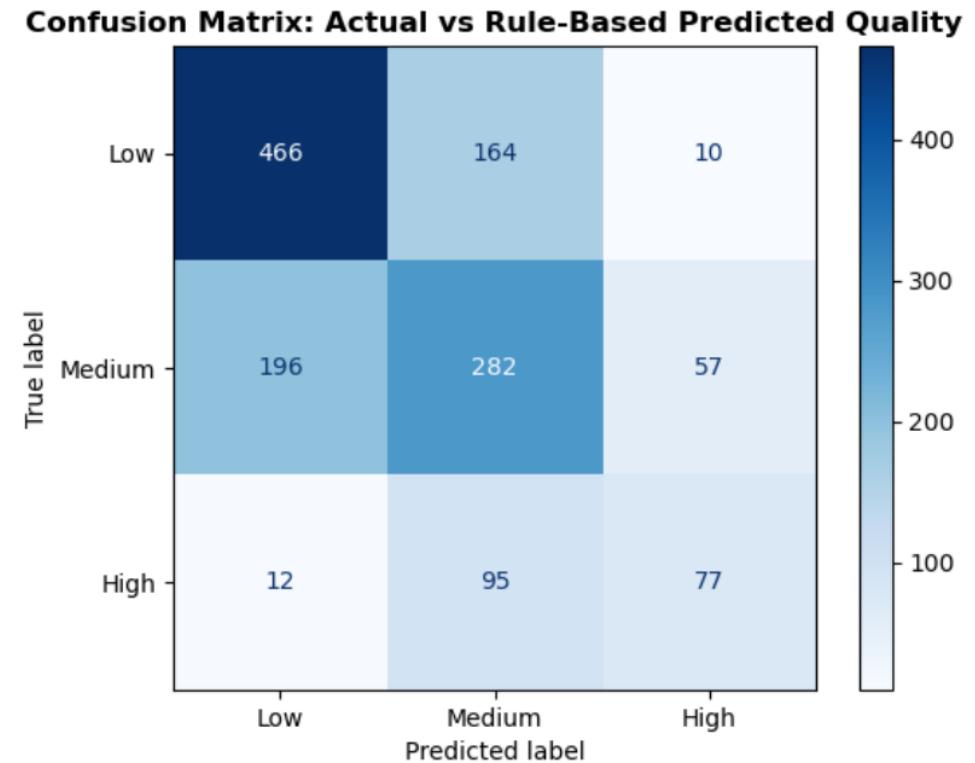


Rule-Based Model



Results:

- Scores 3–5 → mostly predicted as **Low**
- Scores 6–7 → mostly **Medium**
- Score 8 → predicted as **High**



Confusion Matrix Insight:

- Low-quality wines:** Accurately classified.
- Medium vs High:** Some overlap due to similar chemical profiles.
- Overall:** Clear trend, but limited precision for borderline cases

Statistical Model

Approach:

- $X \rightarrow$ used all 11 features; $y \rightarrow$ quality
 - Standardized inputs
 - Split data 80/20

Optimization:

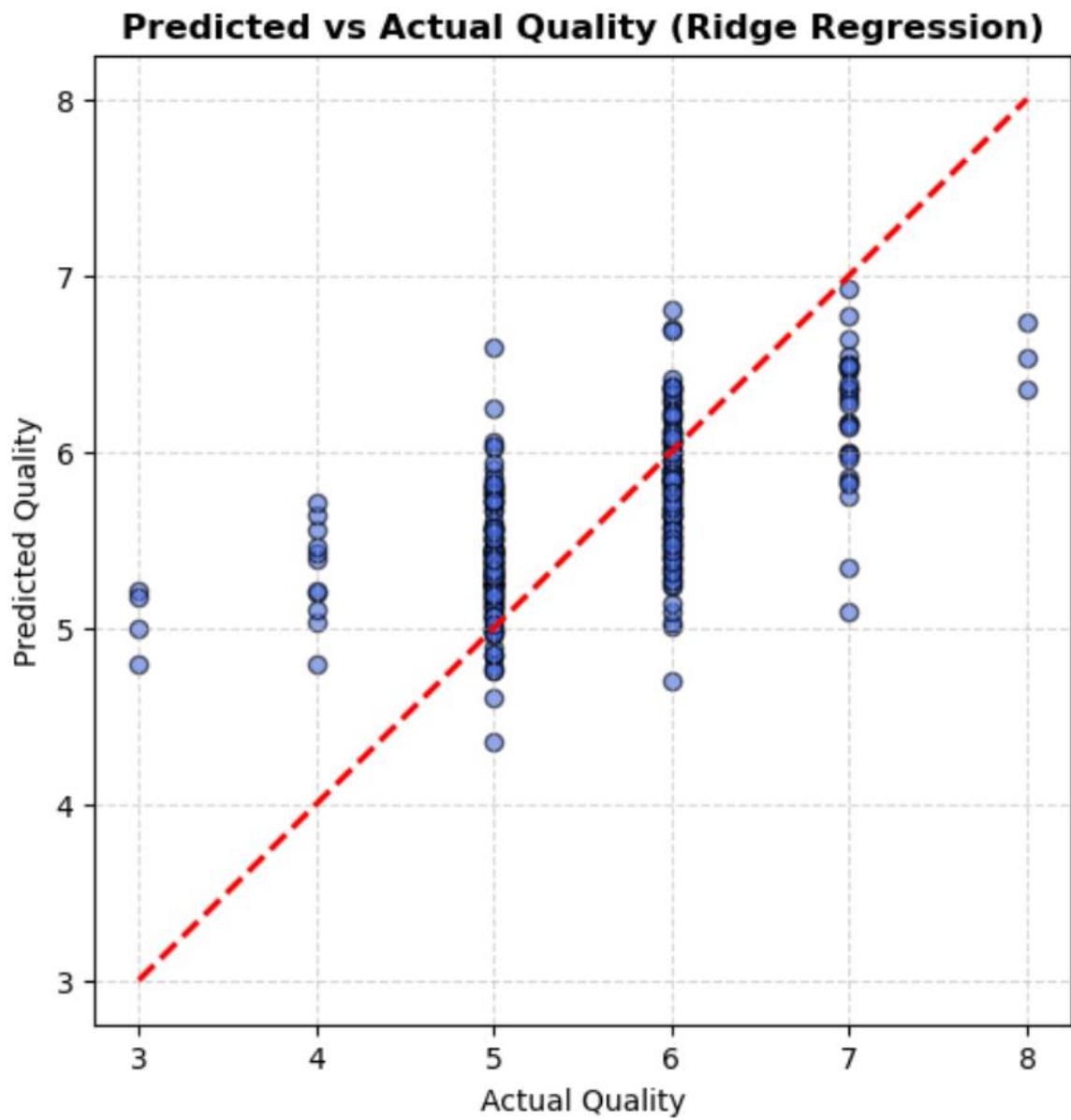
- Tuned regularization parameter α ; best result at $\alpha = 59$

Performance:

- Ridge MSE: **0.4275** → Slight improvement in accuracy and stability.
 - Linear MSE: **0.4310**

Interpretation:

- Regularization helped stabilize coefficients.
 - Scatter plot shows predictions closely match actual scores.



Comparative Evaluation & Reflection

Criterion	Rule-Based Model	Ridge Regression Model
Interpretability	Very high – intuitive thresholds	Moderate – requires statistical understanding
Predictive Accuracy	Qualitative – trend-following	Quantitative – lower MSE (0.4275)
Complexity	Low – simple visual rules	Moderate – scaling, tuning, regularisation
Transparency	Full – human-readable logic	Partial – abstract but explainable via coefficients
Scalability	Easy to adapt via thresholds	Needs retraining for new data

Key Findings

- **Alcohol** is the strongest quality indicator.
- **Sulphates** support higher quality.
- **Volatile acidity** lowers quality.
- **Rule-based model:** Clear and interpretable.
- **Ridge regression:** Better accuracy and more stable.

Limitations & Future Work

- Dataset size limits generalization.
- Explore non-linear models.
- Automate rule extraction for better interpretability + accuracy.

