Data Preprocessing Pipeline Report

Wine Quality Dataset

Submitted By:

- Mirunalini A.R.A
  B.E CSE (AI & ML) – Honours
  SCSVMV, Enathur

- Shivani Reddy
  B.E CSE (Cybersecurity) – Honours
  SCSVMV, Enathur

Subject: Data Warehousing and Mining
Instructor: Mr.K. Harshawardhan Prof Dept of CSE
Date of Submission: 16-02-2026

---

1. Introduction

Data preprocessing plays a critical role in improving machine learning model performance. Raw datasets often contain duplicates, outliers, inconsistent formats, and features on different scales, which can negatively impact model accuracy and generalization.

This project implements a complete end-to-end preprocessing pipeline on the Wine Quality dataset. The objective is to analyze and compare the impact of various preprocessing techniques including cleaning, transformation, feature selection, and dimensionality reduction on model performance.

---

2. Dataset Description

The dataset was obtained from the UCI Machine Learning Repository.

- Dataset: Wine Quality (Red Wine)
- Total Samples: 1599
- Features: 11 physicochemical properties
- Target Variable: Quality score (converted to binary classification)

Target conversion:

- Quality ≥ 6 → Good Wine (1)
- Quality < 6 → Bad Wine (0)

All features are numerical.

---

3. Exploratory Data Analysis (Before Preprocessing)

Key Observations:

- No missing values were found in the dataset.
- Some duplicate records were present.
- Several features showed skewed distributions.
- Boxplots revealed the presence of outliers.
- Correlation analysis showed:
    - Alcohol positively correlated with quality.
    - Volatile acidity negatively correlated with quality.

Visualizations included:

- Missing value heatmap
- Feature distribution histograms
- Correlation heatmap
- Boxplot for outlier detection

---

4. Data Cleaning

4.1 Duplicate Removal

Duplicate records were identified and removed to prevent redundancy and bias.

4.2 Outlier Removal (IQR Method)

Outliers were detected using the Interquartile Range (IQR) method and removed.

Impact:

- Reduced extreme values
- Stabilized feature distributions
- Improved data consistency

Visualizations:

- Boxplot before outlier removal
- Boxplot after outlier removal
- Statistical summary comparison

---

5. Data Transformation

Before applying transformations, the dataset was split into:

- 80% Training data
- 20% Testing data

This prevents data leakage.

5.1 Z-Score Normalization

StandardScaler was applied to standardize features:

- Mean = 0
- Standard Deviation = 1

5.2 Min-Max Normalization

MinMaxScaler scaled features between 0 and 1.

Observations:

- Scaling changed feature ranges.
- Correlation structure remained unchanged.
- Standardization improved model compatibility.

Visualizations:

- Histogram before scaling
- Histogram after scaling
- Correlation heatmap comparison

---

6. Feature Selection

Two methods were applied:

6.1 Correlation-Based Selection

Features with significant correlation to the target were retained.

6.2 Recursive Feature Elimination (RFE)

RFE selected the top 5 most predictive features using Logistic Regression.

Observations:

- Alcohol and volatile acidity were strong predictors.
- Reduced feature space with minimal accuracy loss.

Visualizations:

- Feature importance bar chart
- Accuracy comparison graph

---

7. Dimensionality Reduction (PCA)

Principal Component Analysis (PCA) was applied after scaling.

- Reduced 11 features to 2 principal components.
- Preserved approximately **(insert your value)%** of total variance.

Observations:

- PCA compressed the dataset effectively.
- Slight reduction in accuracy due to information loss.
- Enabled 2D visualization of class separation.

Visualizations:

- Explained variance plot
- 2D scatter plot before PCA
- 2D scatter plot after PCA

---

8. Model Performance Comparison

Logistic Regression was trained at different stages.

Dataset Stage	Accuracy

Raw Scaled Data	X.XX
Correlation Selected	X.XX
RFE Selected	X.XX
After PCA	X.XX

(Insert your actual values)

Performance Comparison Chart was generated.

---

9. Discussion

- Data cleaning improved dataset reliability.
- Scaling significantly improved model stability.
- Feature selection reduced dimensionality with minimal performance drop.
- PCA reduced dimensionality further but slightly reduced accuracy.
- The preprocessing pipeline demonstrated measurable impact on model behavior.

This shows that preprocessing is not optional — it directly affects model performance.

---

10. Conclusion

This project successfully implemented a complete data preprocessing pipeline including:

- Data cleaning
- Feature transformation
- Feature selection

- Dimensionality reduction
- Model comparison

Results demonstrate that proper preprocessing improves model robustness and interpretability. While dimensionality reduction simplifies the dataset, excessive compression may slightly reduce accuracy.

Overall, structured preprocessing leads to better-performing and more reliable machine learning models.

---

11. References

- UCI Machine Learning Repository – Wine Quality Dataset
- Scikit-learn Documentation
- Python Pandas Documentation
- Seaborn & Matplotlib Documentation