Pattern Recognition Assignment

1.Objective

The purpose of this study is to perform a comparative analysis of multiple machine learning models on the **Wine Quality dataset** (red and white wine) to identify patterns and key predictors of wine quality. Both **classification** (quality as high/low) and **regression** (predicting exact quality score) tasks are included.

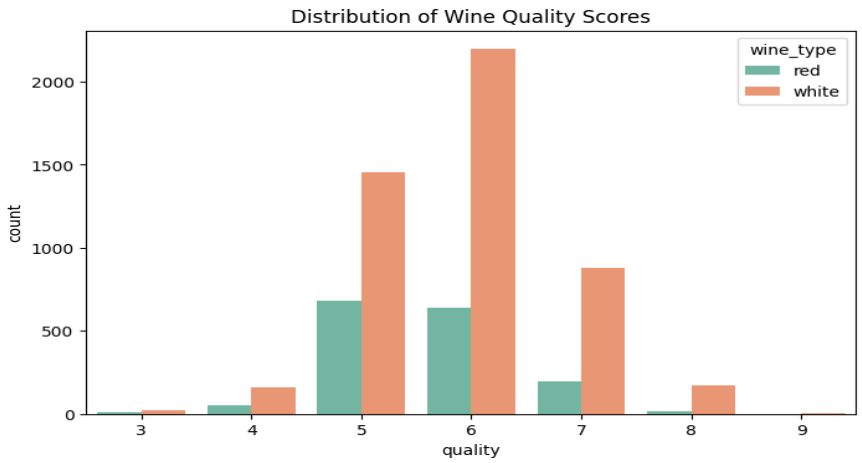
2.Dataset Description

- **Source:** UCI Machine Learning Repository – Wine Quality Dataset
- **Size:** ~6,500 records (red + white wine)
- **Features:** 11 physicochemical variables (alcohol, sulphates, citric acid, pH, etc.) + wine type
- Target **Variables:**
  - *Classification:* Wine quality → High (≥6), Low (<6)
  - *Regression:* Wine quality score (0–10)

```
Dataset shape: (6497, 13)
wine_type
white    4898
red      1599
Name: count, dtype: int64
```

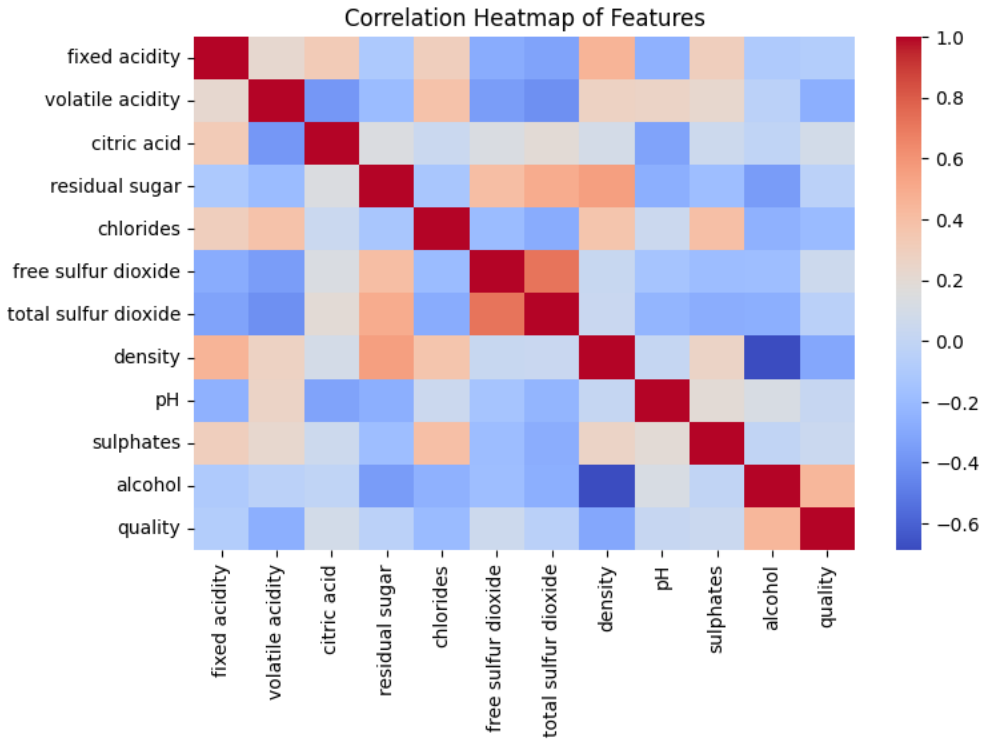
3. Exploratory Data Analysis (EDA)

3.1 Distribution of Quality Scores



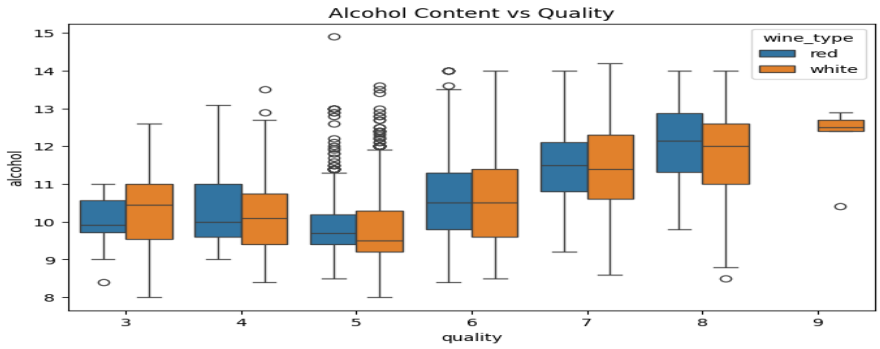
- White wine dominates dataset.
- Both red and white wines show more samples in medium quality range (5–6).

3.2 Correlation Heatmap



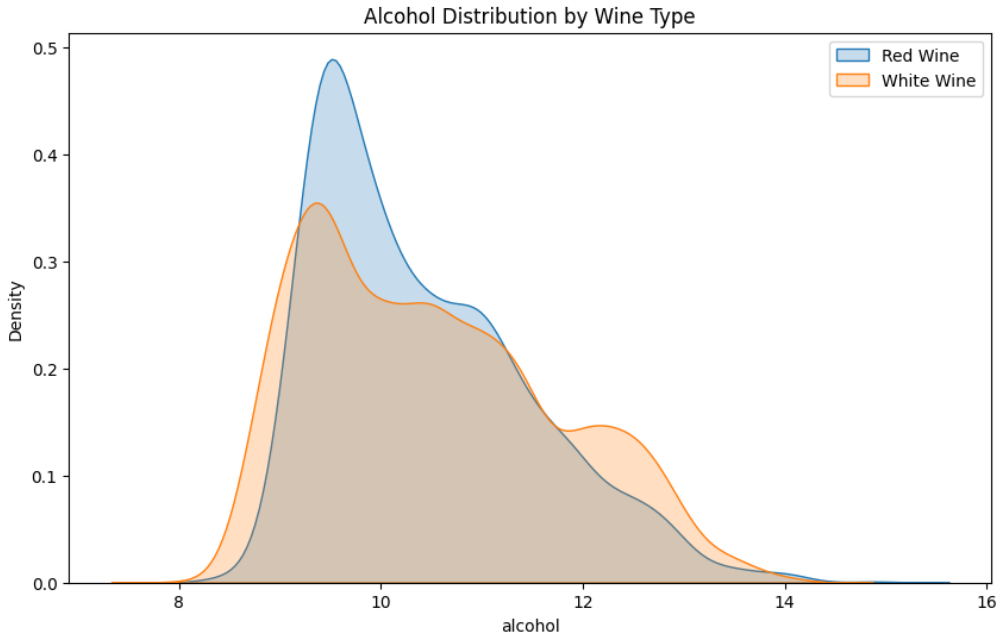
- **Alcohol** shows the strongest positive correlation with wine quality.
- **Volatile acidity** is negatively correlated.

3.3 Alcohol vs Quality



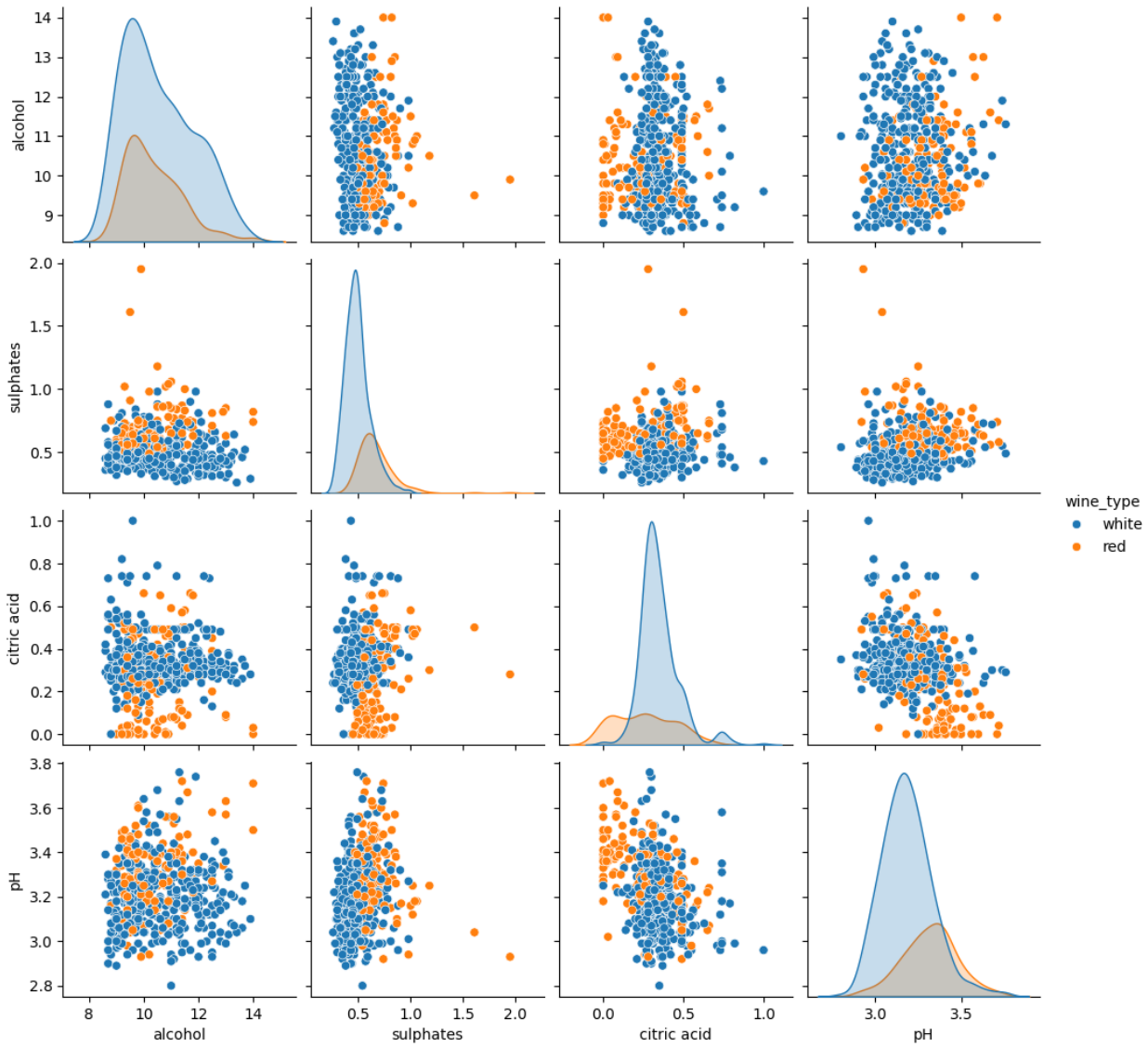
- Higher alcohol content generally corresponds to higher quality.

3.4 Alcohol Distribution



- Red and white wines show different alcohol distributions; whites tend to have higher alcohol content.

3.5 Pairplot



Alcohol and sulphates are promising features for distinguishing wine quality.

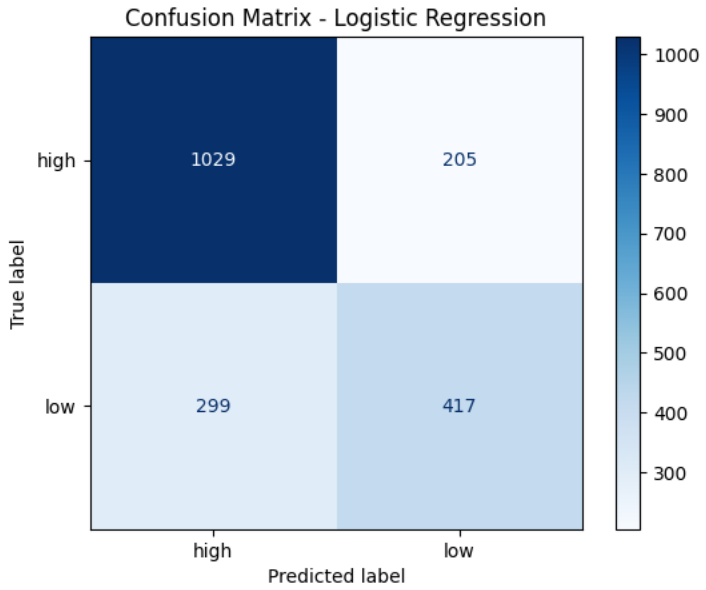
4. Classification Models

- Models trained: **Logistic Regression, SVM, KNN**
- Split: 70/30 (stratified), with standardization

```
Logistic Regression Accuracy: 0.742
Classification Report (Logistic Regression):
              precision    recall  f1-score   support

        high       0.77      0.83      0.80      1234
         low       0.67      0.58      0.62       716

    accuracy                           0.74      1950
   macro avg       0.72      0.71      0.71      1950
weighted avg       0.74      0.74      0.74      1950
```

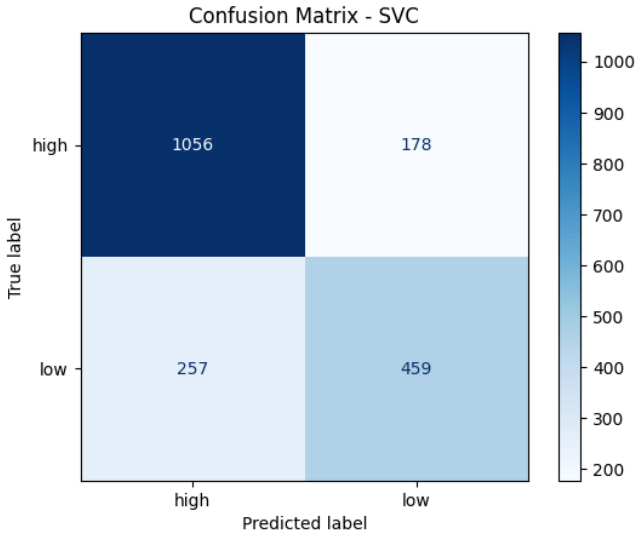


SVC:

```
SVC Accuracy: 0.777
Classification Report (SVC):
              precision    recall  f1-score   support

        high       0.80      0.86      0.83      1234
         low       0.72      0.64      0.68       716

    accuracy                           0.78      1950
   macro avg       0.76      0.75      0.75      1950
weighted avg       0.77      0.78      0.77      1950
```

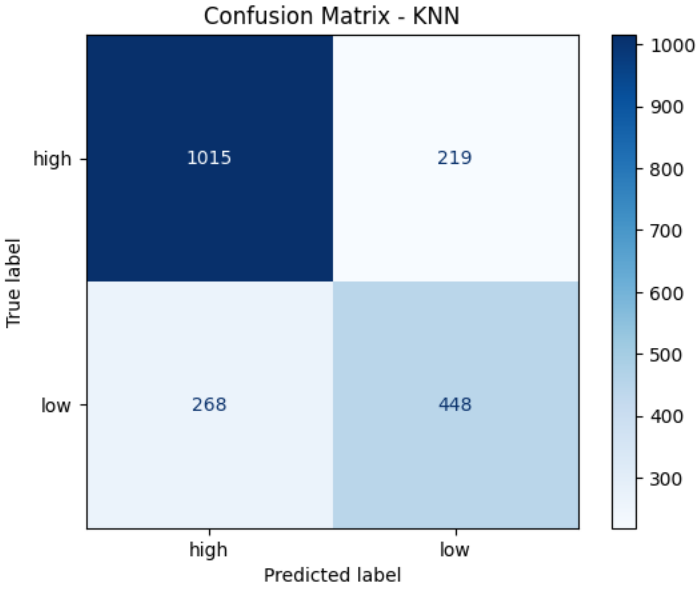


KNN:

```
KNN Accuracy: 0.750
Classification Report (KNN):
              precision    recall  f1-score   support

        high       0.79      0.82      0.81      1234
         low       0.67      0.63      0.65       716

    accuracy                           0.75      1950
   macro avg       0.73      0.72      0.73      1950
weighted avg       0.75      0.75      0.75      1950
```



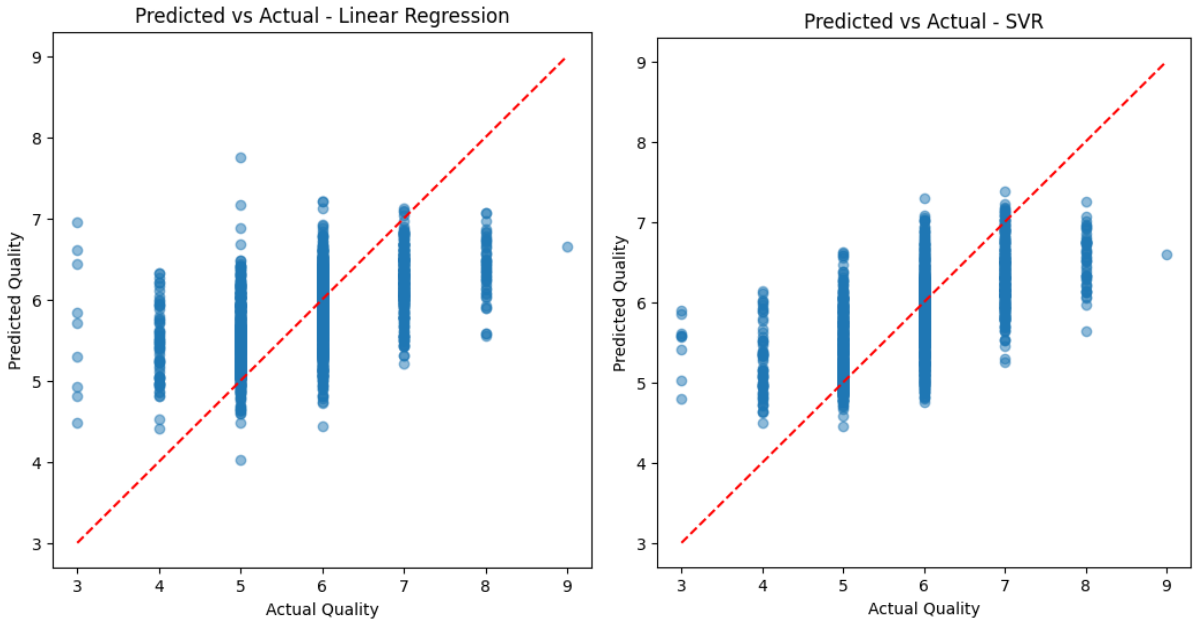
5. Regression Models

- Models trained: **Linear Regression, SVR**

5.1 Results

```
Linear Regression -> MSE: 0.533, R²: 0.270
```

```
SVR -> MSE: 0.449, R²: 0.384
```



- Regression models capture trends but underperform due to non-linear relationships.

6. Conclusion

This study demonstrates the challenges of predicting wine quality using traditional machine learning models. While exploratory analysis confirmed that **alcohol content** and **volatile acidity** are the strongest indicators of wine quality, the predictive performance of baseline models remains modest:

- **Classification models** (Logistic Regression, SVM, KNN) achieved **~75% accuracy**, reflecting the difficulty of distinguishing "high" vs "low" quality wines due to overlapping feature distributions and class imbalance.
- **Regression models** (Linear Regression, SVR) achieved $\mathbf{R^2 \approx 0.25–0.4}$, indicating they capture only a small portion of the variance in wine quality scores.

Key Insights

- **Alcohol content** consistently shows a positive influence on higher quality ratings.
- **Volatile acidity** is negatively correlated with quality, reinforcing domain knowledge from winemaking.
- Simple linear models struggle because wine quality is influenced by **complex, nonlinear interactions** between chemical features.

Final Note

While baseline models provide useful insights, they are not sufficient for accurate quality prediction. Future improvements could include:

- Applying **ensemble methods** (Random Forest, Gradient Boosting, XGBoost) for better handling of nonlinearity.
- **Feature engineering** to capture interaction terms (e.g., alcohol × sulphates).
- Using **class balancing techniques** (SMOTE, weighted losses) to address label imbalance.

Overall, the study highlights that wine quality prediction is a **multifactorial, nonlinear problem** and that advanced machine learning methods will likely outperform the baseline models tested here.