Machine Learning Wine Quality Prediction & Clustering

1. Introduction

This report explores machine learning techniques applied to the UCI Wine Quality dataset. The tasks include regression, classification with dimensionality reduction (PCA), and clustering (K-Means). The goal is to build, evaluate, and interpret models for predictive, supervised and unsupervised learning.

2. Dataset Description

The dataset contains physicochemical measurements and quality scores (0–10) for Portuguese Vinho Verde wines. Red wine samples: 1,599, white wine samples: 4,898. After merging and removing 1177 duplicates. We have 5,320 rows × 13 columns.

The features included in the data set are: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphate, alcohol, quality and a new column 'type' which distinguishes red and white wine.

3. Methods

3.1 Regression (Linear vs Ridge)

For the regression task, the goal was to predict the numeric wine quality score. As a baseline, I used a simple mean predictor, which always outputs the average quality across the dataset.

then compared Multiple Linear Regression with Ridge Regression, experimenting with α values of 0.1, 1, and 10 to check if adding regularisation would improve performance.

Model accuracy was measured using RMSE with an 80/20 train–test split, and learning curves were plotted to show how error changed as the training size increased.

3.2 Classification with PCA

For the classification task, wines were labelled as high quality if their quality score was 7 or higher, and low quality otherwise. Before training, the features were standardised to ensure they were on the same scale.

Then applied Principal Component Analysis (PCA) and kept enough components to explain at least 90% of the variance, which reduced the dimensionality while retaining most of the information in the data.

Two models were tested: Logistic Regression and k-Nearest Neighbours (k = 1–15). Their performance was evaluated using accuracy, with an accuracy vs k curve plotted for k-NN, and the best k reported alongside its confusion matrix.

3.3 Clustering (K-Means)

For the clustering task, all features were standardised to ensure comparability across scales, and K-Means clustering was applied with k values from 2 to 6. The elbow method was used to select an appropriate number of clusters, and PCA was applied to reduce the data to two dimensions for visualisation. To interpret the clusters, then compared the distribution of red and white wines in each group and examined their average quality scores.

4.Exploratory data analysis

Alcohol histogram:

- Most wines have alcohol content between 9–12%, with fewer wines above 13%.
-

Sulphates histogram:

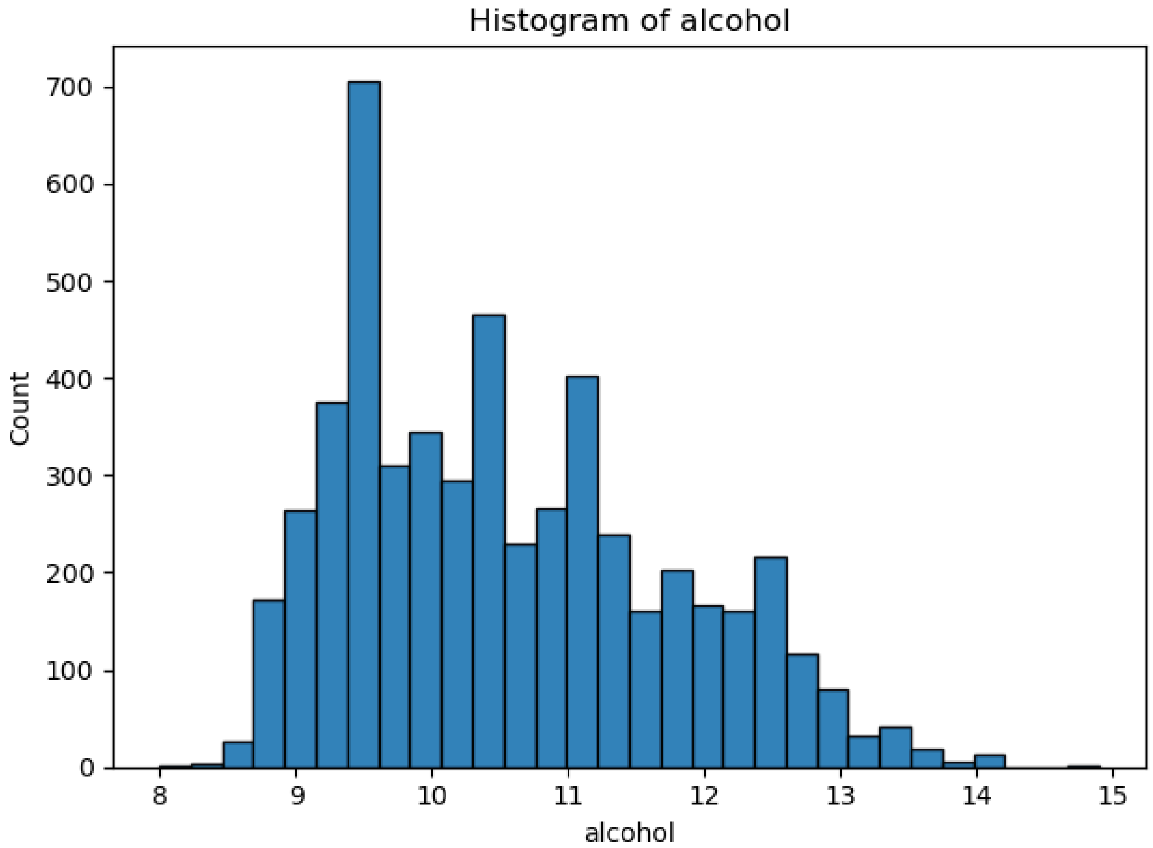
- Sulphates are concentrated between 0.4–0.7, with very few outliers above 1.0.

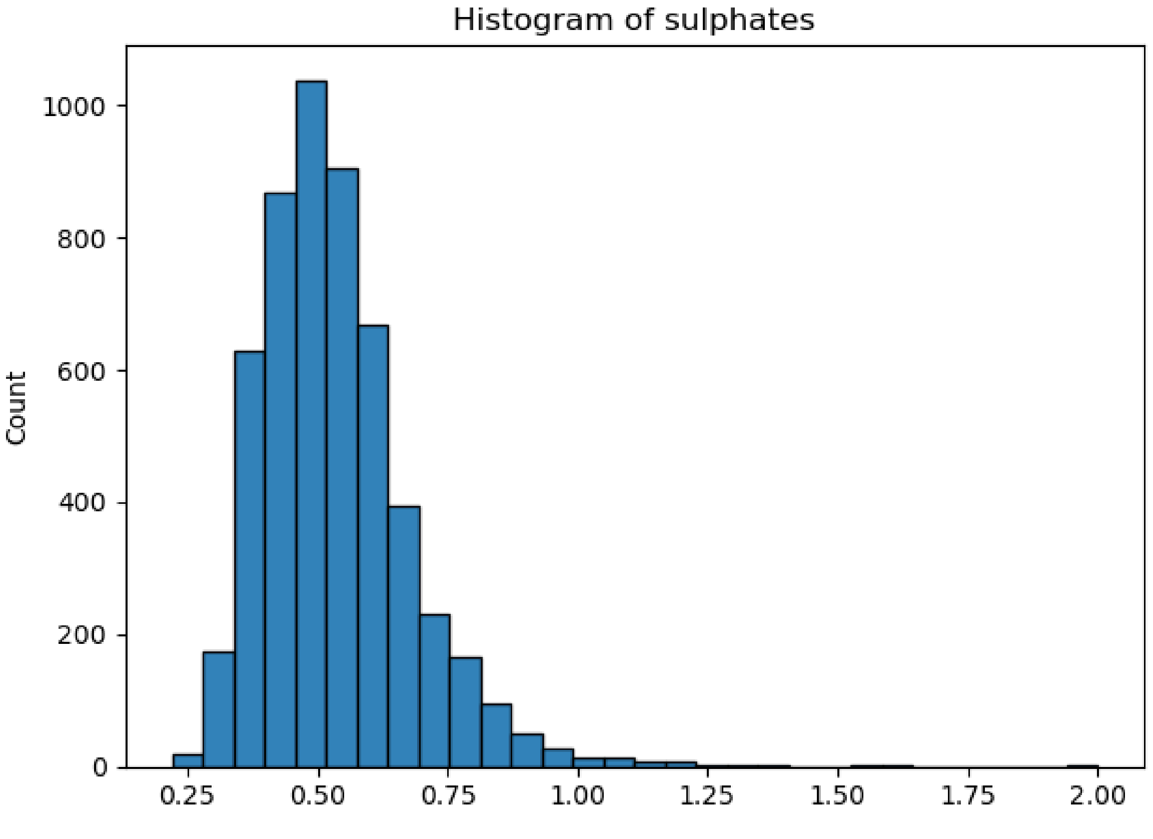
Citric acid histogram:

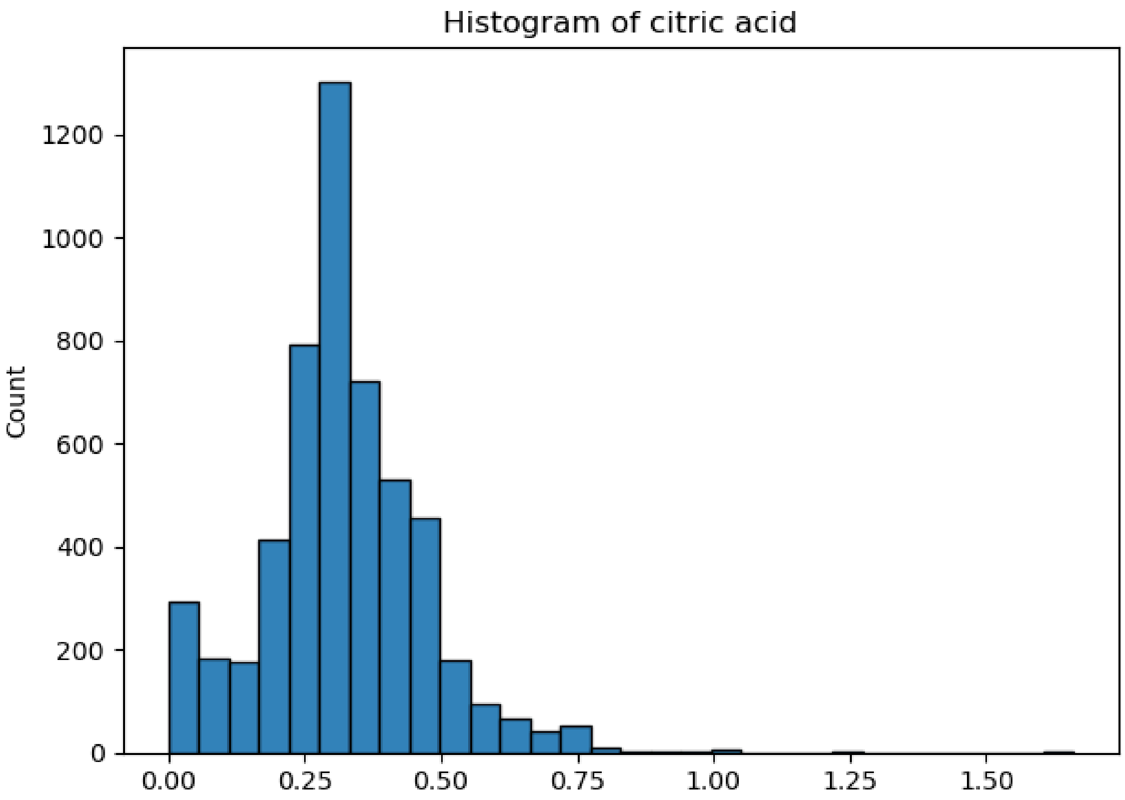
- The majority of wines cluster around 0.2–0.4, while many have near-zero citric acid.

Alcohol vs Quality scatter plot:

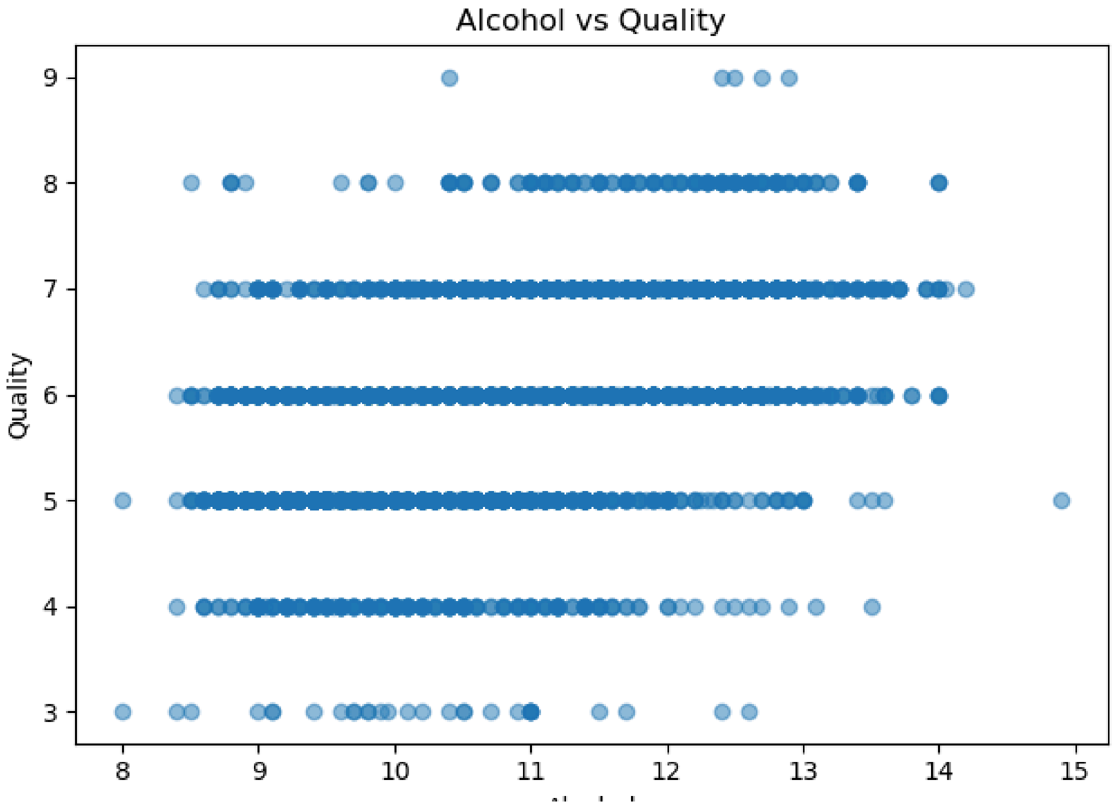
- Higher alcohol levels generally correspond to higher quality ratings, though the relationship is not perfectly linear.


Histogram of alcohol


Histogram of sulphates



Histogram of citric acid



Alcohol vs Quality

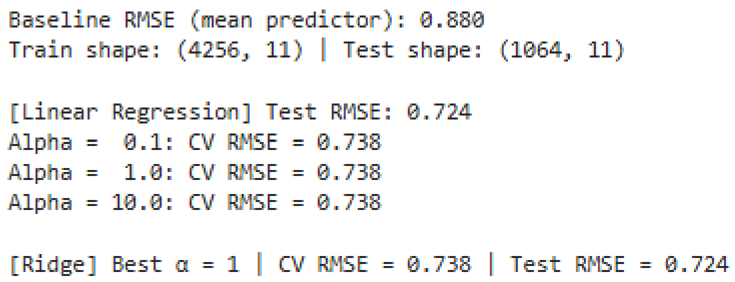
4. Results

4.1 Regression

The baseline model (mean predictor) produced an RMSE of 0.88, which represents the error when predicting the average quality for all wines. Both Linear Regression and Ridge Regression achieved lower errors, with a Test RMSE of 0.724, demonstrating a clear improvement over the baseline model.

Ridge regression was tuned using cross-validation with alpha = {0.1, 1, 10}, all the values gave nearly identical performance, so I picked alpha = 1 for the report and graph.

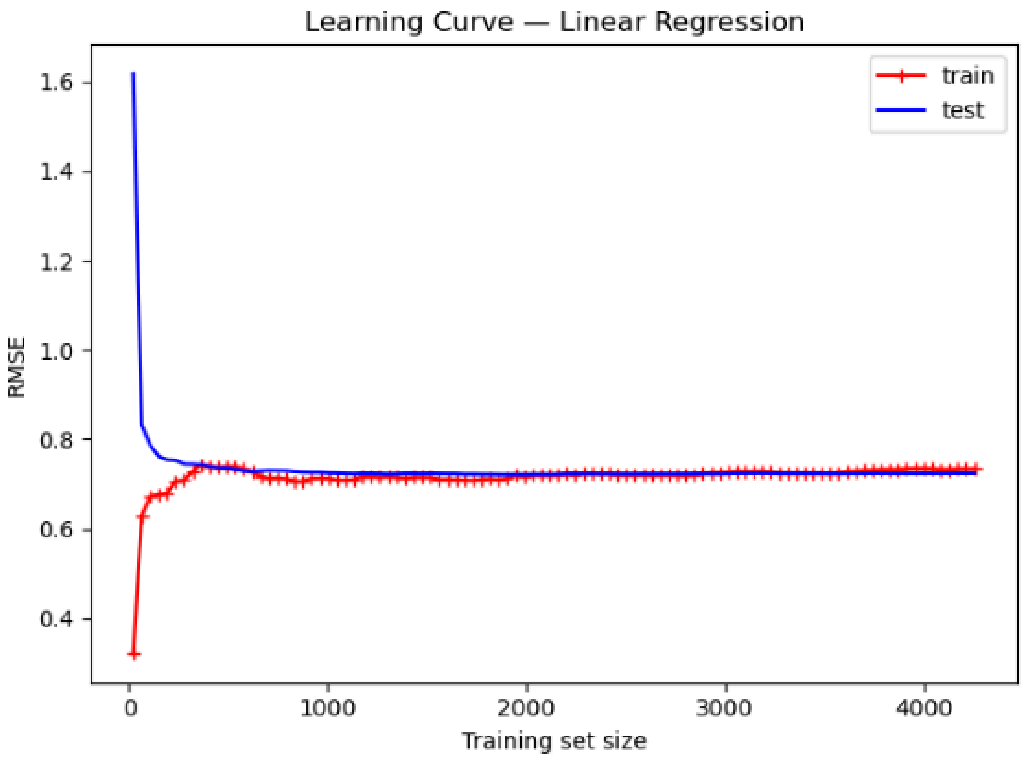
Learning curves for both models converged in the range of 0.72–0.73, indicating stable generalisation and little evidence of overfitting or underfitting as shown in the learning curve graph. .

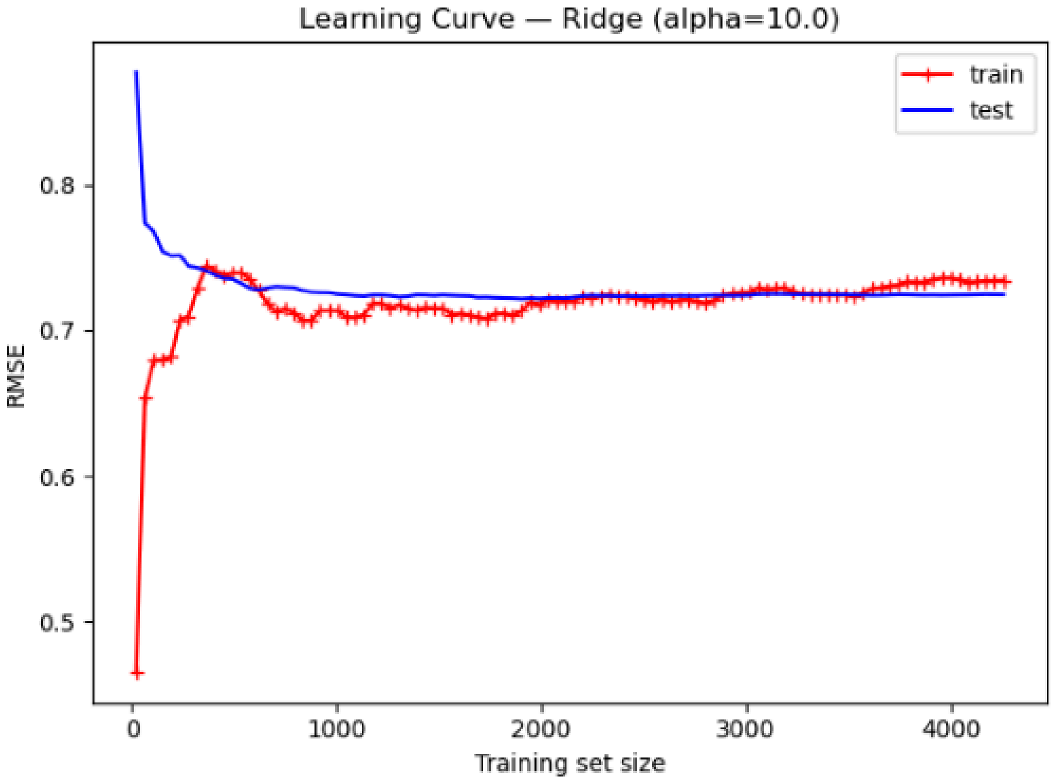


```
Baseline RMSE (mean predictor): 0.880
Train shape: (4256, 11) | Test shape: (1064, 11)

[Linear Regression] Test RMSE: 0.724
Alpha =  0.1: CV RMSE = 0.738
Alpha =  1.0: CV RMSE = 0.738
Alpha = 10.0: CV RMSE = 0.738

[Ridge] Best α = 1 | CV RMSE = 0.738 | Test RMSE = 0.724
```


Learning Curve — Linear Regression
train
test
RMSE
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0
1000
2000
3000
4000
Training set size

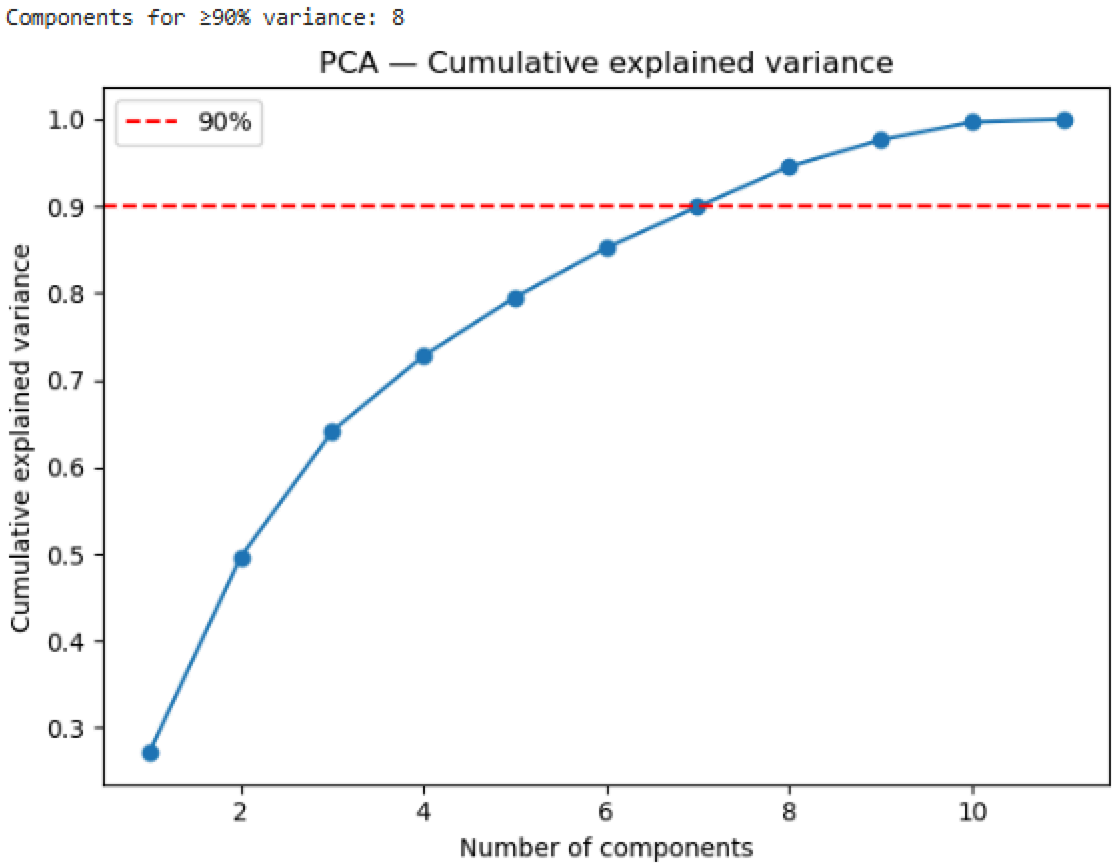

Learning Curve — Ridge (alpha=10.0)
train
test
RMSE
0.8
0.7
0.6
0.5
0
1000
2000
3000
4000
Training set size

4.2 Classification

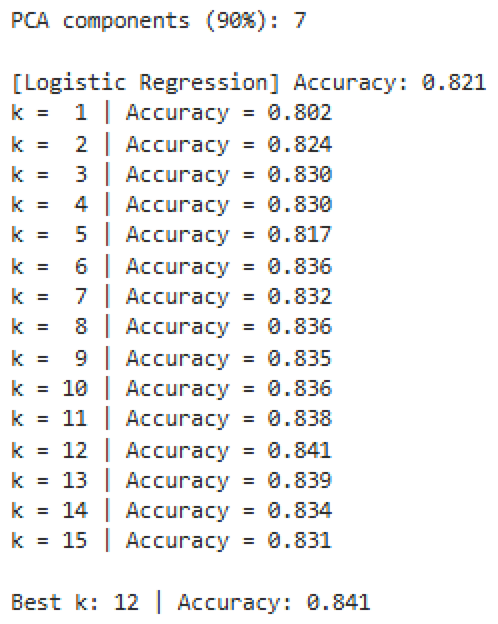
Applying PCA showed that 7–8 components were enough to explain at least 90% of the variance in the data.

Using these reduced features, Logistic Regression achieved an accuracy of 0.821. The k-Nearest Neighbours classifier performed slightly better, reaching a peak accuracy of 0.841 at k = 12.

The confusion matrix indicates that most low-quality wines were classified correctly, while some high-quality wines were misclassified as low quality, reflecting a tendency towards false negatives where high quality wins were predicted as low quality



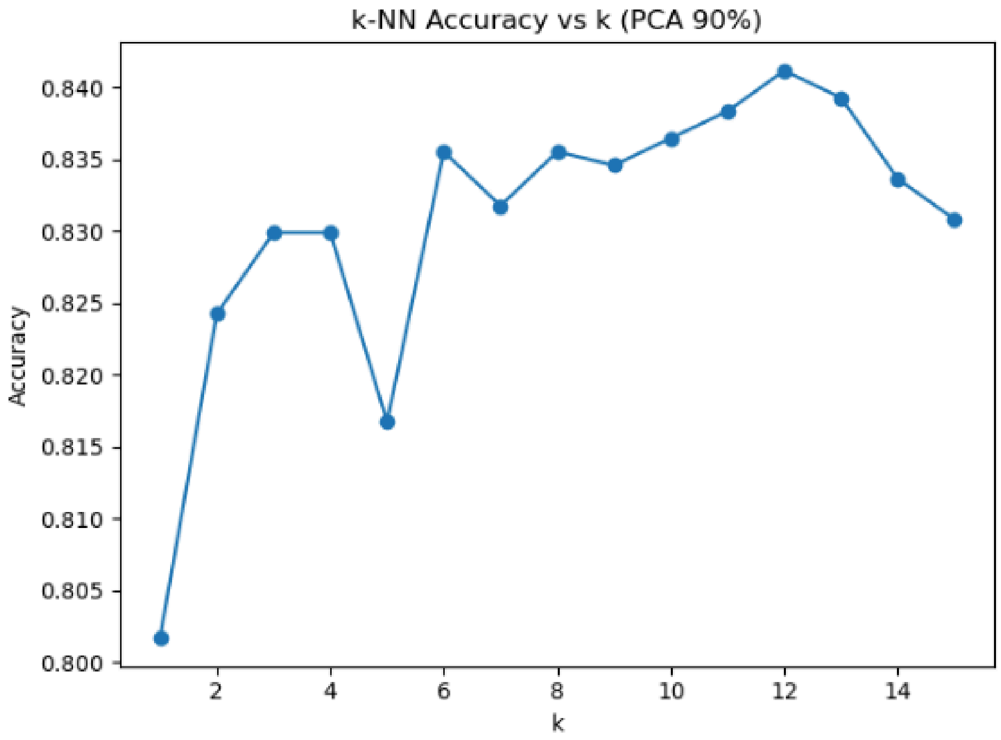
```
Components for ≥90% variance: 8
```



```
PCA components (90%): 7

[Logistic Regression] Accuracy: 0.821
k =  1 | Accuracy = 0.802
k =  2 | Accuracy = 0.824
k =  3 | Accuracy = 0.830
k =  4 | Accuracy = 0.830
k =  5 | Accuracy = 0.817
k =  6 | Accuracy = 0.836
k =  7 | Accuracy = 0.832
k =  8 | Accuracy = 0.836
k =  9 | Accuracy = 0.835
k = 10 | Accuracy = 0.836
k = 11 | Accuracy = 0.838
k = 12 | Accuracy = 0.841
k = 13 | Accuracy = 0.839
k = 14 | Accuracy = 0.834
k = 15 | Accuracy = 0.831

Best k: 12 | Accuracy: 0.841
```



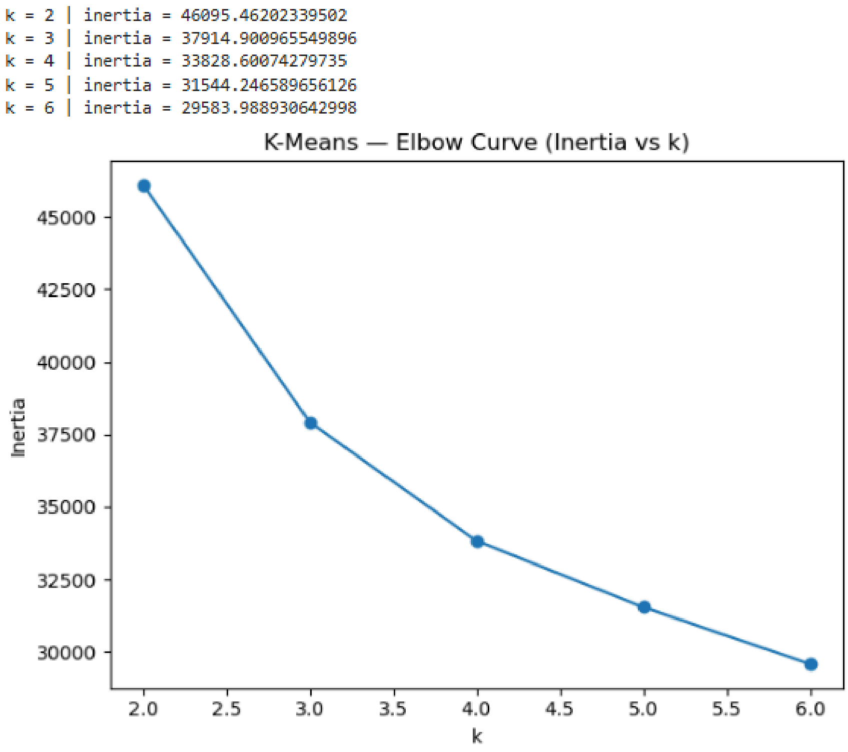
```
Confusion Matrix (k = 12):
[[832  30]
 [139  63]]
```

4.3 Clustering

From the elbow curve, the largest drop in inertia occurred up to k = 3–4, suggesting this range as a good choice for clustering. I selected k = 4 to better highlight the separation and quality differences between red and white wines.

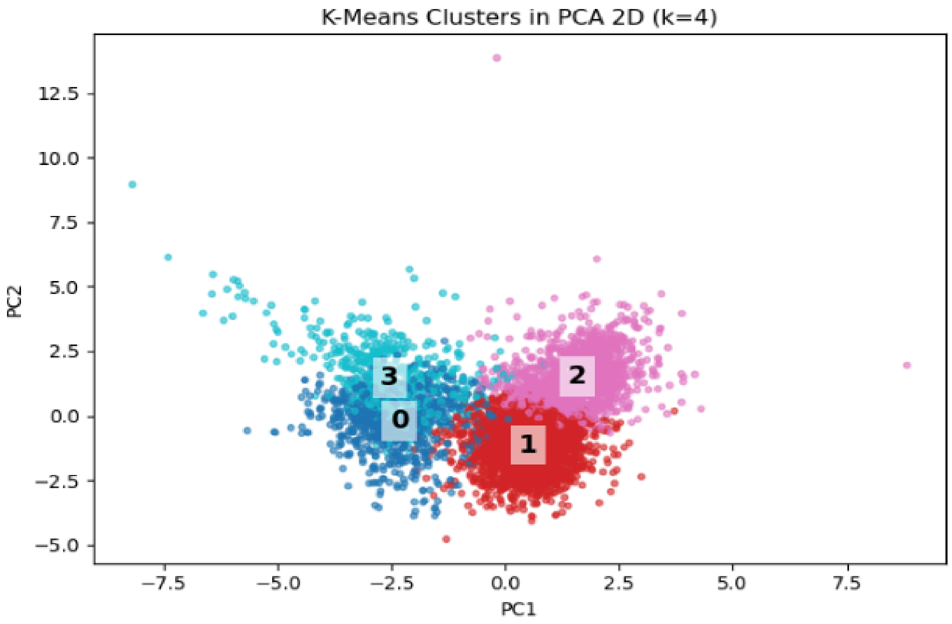
The PCA scatter plot (PC1 vs PC2) showed clear separation between groups, even though only about 50% of the variance is captured in 2D. Using four clusters helped distinguish red wines from white wines.

The quality summary indicated that cluster 1, which is mostly white wines, had the highest mean quality (6.1), while cluster 0, mainly red wines, had the lowest (5.4). This indicates that, within these clusters, white wines tended to have a higher average quality compared to red wines in this dataset.



```
k = 2 | inertia = 46095.46202339502
k = 3 | inertia = 37914.900965549896
k = 4 | inertia = 33828.60074279735
k = 5 | inertia = 31544.246589656126
k = 6 | inertia = 29583.988930642998
```

```
Explained variance (PC1+PC2): 0.497
```



```
Counts by cluster and type:
 type      red  white
cluster
0          774     80
1           38   2435
2            4   1402
3          543     44

Cluster quality summary:
           count      mean
cluster
0            854  5.384075
1           2473  6.065507
2           1406  5.571124
3            587  5.795571
```

6. Conclusion

In this project, I applied regression, classification, and clustering techniques to the UCI Wine Quality dataset. Both Linear and Ridge regression produced an RMSE of 0.724, a clear improvement over the baseline mean predictor. For classification, PCA reduced the feature set to 7–8 components while keeping performance strong at 90% variance, with k-NN achieving the best accuracy at k = 12 with a 0.841 accuracy . K-Means clustering with *k = 3–4* revealed meaningful groupings that roughly reflected wine type and average quality. Overall, the results suggest that even relatively simple models can capture important patterns in the dataset, although the imbalance in high-quality wines remains a limitation for predictive accuracy.

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. UCI Machine Learning Repository – Wine Quality Dataset.