DA Assignment 08
Report

Name: Vatsal Bhuva
Roll Number: IIT2022004

1. Objective

The aim was to cluster the Wheat Seeds dataset using the Partitioning Around Medoids (PAM) algorithm and evaluate how well the clusters match actual wheat types.

This involved several steps: loading and preprocessing the data, applying PCA for visualization, performing PAM clustering, and evaluating the results using appropriate metrics.

2. Data Loading & Exploration

The dataset was loaded from seeds_dataset.txt, which contains 210 samples with 7 physical features and a class label (1, 2, or 3).

Column names were added manually as the dataset lacked a header row. No missing values were found.

Initial exploration included visualizing distributions of each feature to check for scale differences or outliers.

3. Data Preprocessing

Standardization using StandardScaler was applied to ensure all features contributed equally to the distance-based PAM algorithm.

PCA was used for dimensionality reduction, primarily for visualization. Although the first two components explained only 88.98% of the variance, they were sufficient for 2D plotting.

A bar plot of explained variance ratio was generated to confirm how much information each component retained.

4. Clustering with PAM Algorithm

PAM clustering was implemented using the pyclustering.cluster.kmedoids module with k=3 (as the data represents three wheat types).

Initial medoids were randomly selected from the dataset, and the algorithm iteratively optimized the medoid selection.

Clustering results were plotted on the PCA-reduced dataset to visually inspect how well the clusters were separated.

5. Evaluation and Comparison

Three evaluation metrics were used:

Silhouette Score to assess cluster cohesion and separation.

Adjusted Rand Index (ARI) to compare clustering with actual labels.

Normalized Mutual Information (NMI) to measure information overlap between clusters and true classes.

A confusion matrix was generated to observe how true class labels mapped to predicted cluster IDs.

While cluster labels did not perfectly match true labels, moderate clustering performance was achieved.

6. Challenges Encountered

The default matplotlib backend (FigureCanvasAgg) was non-interactive, preventing plots from being displayed using plt.show().

Attempting to switch to the TkAgg backend failed due to the absence of the Tkinter module in the Python environment.

The PCA-reduced dimensions explained slightly less than the desired 95% variance, which could impact the accuracy of visual interpretation.

7. Conclusion

The PAM algorithm was able to reasonably cluster the wheat seeds into three groups based on physical features.

Preprocessing steps like standardization and PCA helped prepare the data for effective clustering and visualization.

Despite minor limitations, such as backend configuration and PCA variance shortfall, the project successfully demonstrated the use of PAM in unsupervised learning.