Clustering Algorithms with PCA for Dimensionality Reduction

1. Introduction

This report presents an analysis of clustering algorithms applied to the Wine dataset, with and without Principal Component Analysis (PCA) for dimensionality reduction.

2. Data Preparation and PCA

Dataset

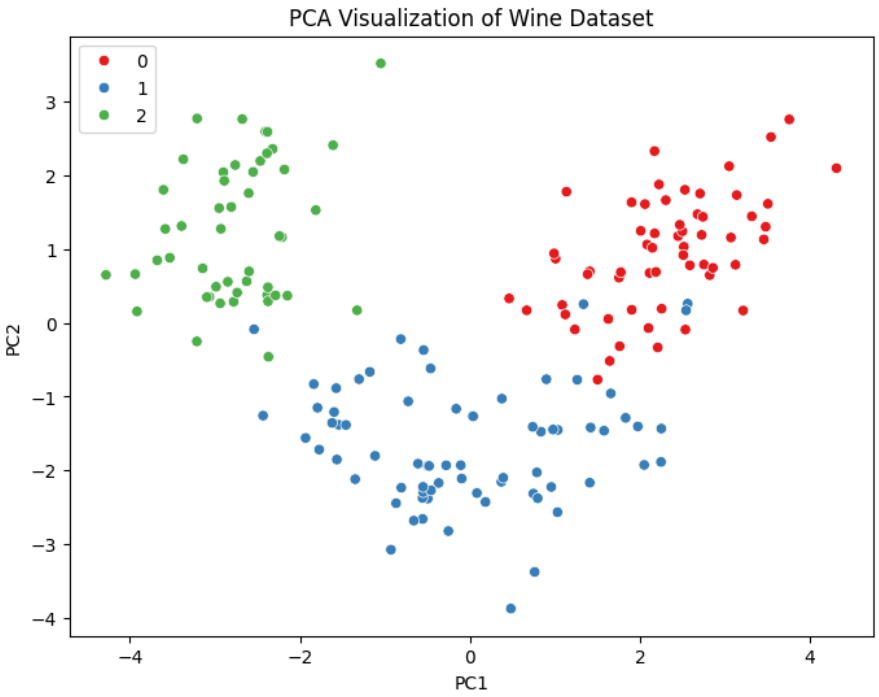
The dataset used is the **Wine dataset** from Scikit-learn, which consists of 178 samples with 13 numerical features describing different properties of wine.

Preprocessing

- The data was standardized using **StandardScaler** to ensure uniform feature scaling.
- No missing values were detected in the dataset.
- PCA was applied to reduce the dimensionality to **2 principal components** for visualization.

PCA Visualization

A scatter plot of the two principal components showed clear groupings among different wine classes, indicating that PCA effectively captured variance in the dataset.



3. Clustering Algorithm Implementation

Two clustering techniques were implemented:

- **K-Means Clustering** (with 3 clusters, as the dataset has 3 actual classes)
- **Hierarchical Clustering** (Agglomerative approach with 3 clusters)

Each algorithm was applied to both the **original dataset** and the **PCA-transformed dataset** to compare results.

4. Performance Evaluation

The clustering results were evaluated using:

- **Silhouette Score** (higher is better)
- **Davies-Bouldin Index** (lower is better)

Clustering Method	Dataset	Silhouette Score	Davies-Bouldin Index
K-Means	Original	0.2849	1.3892
K-Means	PCA	0.5602	0.5977
Hierarchical	Original	0.2774	1.4186
Hierarchical	PCA	0.5591	0.6013

5. Visualization and Insights

- The PCA-transformed dataset allowed for **better visualization** of clustering results in a 2D space.
- Clustering performance **Increased** when using PCA-reduced data, as seen in the higher silhouette scores and lower Davies-Bouldin indices.

6. Conclusion

- The PCA dataset performed **better** in terms of visualizing high-dimensional data and clustering evaluation metrics.