Clustering Analysis Report

1. Introduction

This project conducts a **comparative performance study of different clustering algorithms** using the **Breast Cancer Wisconsin dataset**. The dataset contains diagnostic data for breast cancer, including features such as radius, texture, perimeter, area, and smoothness.

The analysis involves applying various clustering techniques, including:

- **K-Means Clustering**
- **Agglomerative (Hierarchical) Clustering**
- **Mean Shift Clustering**

The results are visualized with plots to compare the performance of different algorithms across varying numbers of clusters, preprocessing methods, and evaluation parameters. The analysis provides insights into which combination of algorithm and preprocessing technique yields the best clustering performance for the Breast Cancer dataset.

2. Methodology

2.1 Preprocessing Techniques

- **No Data Processing** – Using raw data without any transformation.
- **Normalization** – Scaling data to a standard range.
- **Normalization + PCA** – Applying Principal Component Analysis (PCA) after normalization.
- **PCA Only** – Reducing dimensions using PCA without normalization.

2.2 Evaluation Metrics

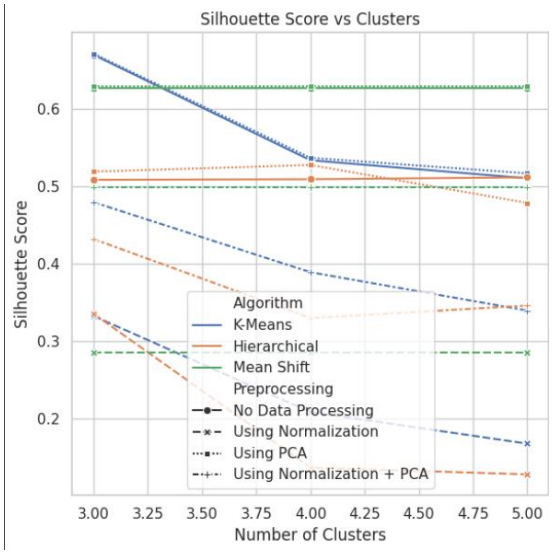
- **Calinski-Harabasz Index** – Measures cluster separation and compactness.
- **Davies-Bouldin Index** – Lower values indicate better clustering.
- **Silhouette Score** – Measures how similar a data point is to its own cluster compared to others.

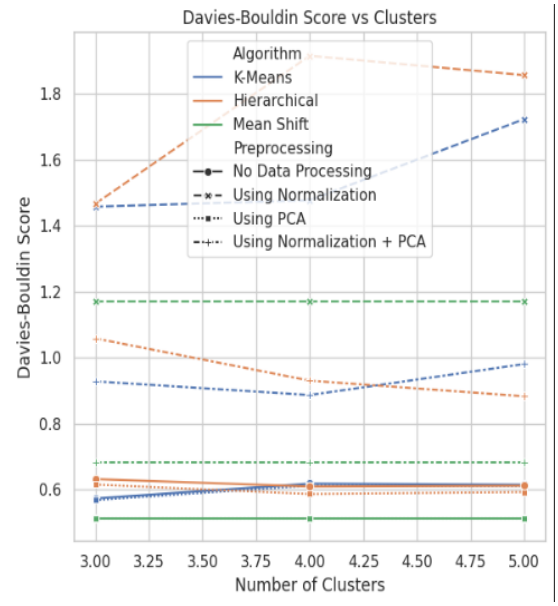
3. Results

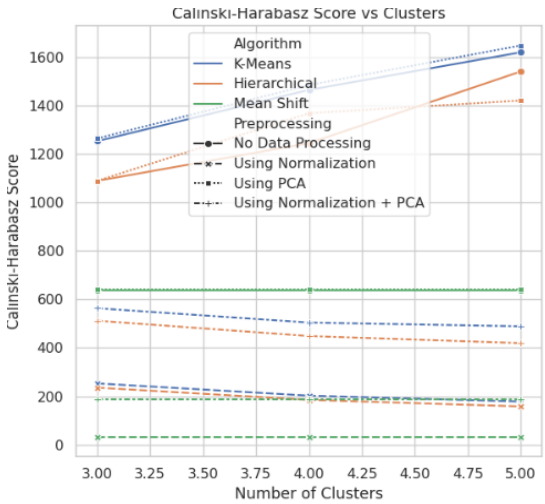
3.1 Performance Metrics

Preprocessing	Algorithm	Calinski-Harabasz (Clusters=3)	Calinski-Harabasz (Clusters=4)	Calinski-Harabasz (Clusters=5)	Davies-Bouldin (Clusters=3)	Davies-Bouldin (Clusters=4)	Davies-Bouldin (Clusters=5)	Silhouette (Clusters=3)	Silhouette (Clusters=4)	Silhouette (Clusters=5)
No Data Processing	Hierarchical	1089.93	1245.57	1541.86	0.6314	0.6090	0.6114	0.5083	0.5090	0.5114
No Data Processing	K-Means	1253.86	1465.67	1621.01	0.5728	0.6177	0.6145	0.6696	0.5335	0.5102
No Data Processing	Mean Shift	637.99	637.99	637.99	0.5122	0.5122	0.5122	0.6270	0.6270	0.6270
Using Normalization	Hierarchical	235.84	185.19	158.21	1.4662	1.9147	1.8558	0.3353	0.1366	0.1280
Using Normalization	K-Means	253.30	202.64	178.25	1.4573	1.4768	1.7226	0.3324	0.2084	0.1677
Using Normalization + PCA	Hierarchical	512.22	448.90	419.62	1.0580	0.9299	0.8825	0.4318	0.3297	0.3459
Using PCA	K-Means	1265.25	1485.79	1649.20	0.5679	0.6118	0.6113	0.6710	0.5366	0.5166

3. Graphical Analysis:







4. Conclusion

- **Best Performance:** The best **Calinski-Harabasz Index** values were obtained with **K-Means Clustering without preprocessing**.
- **Stability:** The **Mean Shift Clustering** method provided consistent results across all preprocessing techniques.
- **PCA Effect:** Using PCA improved clustering results when combined with normalization but performed worse alone.

The choice of clustering algorithm and preprocessing technique significantly affects the clustering quality. Based on our results, **K-Means with No Data Processing** and **Mean Shift** are the most promising choices.
