Clustering Results Report

This report provides a detailed analysis of the clustering results obtained using four algorithms: **K-Means**, **DBSCAN**, **Agglomerative Clustering**, and **Gaussian Mixture Models (GMM)**. Each algorithm's performance was evaluated using the **Silhouette Score** and **Davies-Bouldin Index (DB Index)** to measure cluster quality.

1. K-Means Clustering

- **Number of Clusters Formed:** 4
- **Silhouette Score:** 0.3991
- **Davies-Bouldin Index (DB Index):** 0.8001

Overview:

K-Means is a centroid-based clustering algorithm that divides the dataset into a predefined number of clusters (in this case, 4). Each point is assigned to the cluster with the nearest centroid, and centroids are iteratively updated to minimize within-cluster variance.

Interpretation:

- A **Silhouette Score** of **0.3991** indicates moderate cluster separation, where points are somewhat closer to their assigned cluster than to other clusters.
- A **DB Index** of **0.8001** suggests that the clusters are reasonably compact and well-separated, although there is room for improvement.

Strengths:

- K-Means is efficient and works well for spherical or evenly sized clusters.
- The algorithm successfully identified four distinct groups in the data.

Limitations:

- The moderate Silhouette Score implies that some overlap exists between clusters.
- Sensitivity to initialization and outliers may have impacted performance.

Suggestions:

- Experiment with different initialization methods or increase the number of iterations.

- Apply feature scaling or principal component analysis (PCA) to improve cluster separability.

2. DBSCAN Clustering

- **Number of Clusters Formed:** 2
- **Silhouette Score:** 0.3824
- **Davies-Bouldin Index (DB Index):** 1.0167

Overview:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm that identifies clusters as dense regions separated by sparse regions. Unlike K-Means, it does not require specifying the number of clusters.

Interpretation:

- A **Silhouette Score** of **0.3824** suggests that the clusters are less distinct compared to other methods, with some data points being ambiguously assigned.
- A **DB Index** of **1.0167** indicates relatively poor compactness and separation of clusters.

Strengths:

- DBSCAN can handle noise and discover clusters of arbitrary shapes.
- The algorithm effectively identified two dense regions.

Limitations:

- Sensitivity to the parameters eps (radius for neighborhood search) and min_samples (minimum points in a neighborhood).
- A relatively high DB Index and low Silhouette Score reflect suboptimal cluster quality.

Suggestions:

- Tune the eps and min_samples parameters to achieve better cluster formation.
- Visualize the data distribution to identify appropriate parameter values.
- Consider combining DBSCAN with dimensionality reduction to enhance performance.

3. Agglomerative Clustering

- **Number of Clusters Formed:** 2
- **Silhouette Score:** 0.5915
- **Davies-Bouldin Index (DB Index):** 0.2580

Overview:

Agglomerative Clustering is a hierarchical clustering method that iteratively merges or splits clusters based on a linkage criterion. This method does not require specifying the number of clusters initially but can be controlled by a distance threshold or predefined cluster count.

Interpretation:

- The **Silhouette Score** of **0.5915** indicates well-defined clusters with strong separation and cohesion.
- A **DB Index** of **0.2580** is the lowest among all methods, reflecting excellent compactness and separation.

Strengths:

- Best performance across all metrics, with clearly separated clusters.
- Handles different cluster shapes and densities effectively.

Limitations:

- Higher computational cost compared to K-Means for large datasets.
- The algorithm assumes clusters are hierarchical, which may not always hold true.

Suggestions:

- Agglomerative Clustering performed the best and is recommended for this dataset.
- Further experimentation with linkage methods (e.g., single, complete, average) could fine-tune results.

4. Gaussian Mixture Models (GMM)

- **Number of Clusters Formed:** 10
- **Silhouette Score:** 0.3006
- **Davies-Bouldin Index (DB Index):** 1.0338

Overview:

GMM is a probabilistic clustering method that models data as a mixture of Gaussian distributions. Unlike K-Means, it accounts for cluster variance and produces soft cluster assignments based on probability.

Interpretation:

- A **Silhouette Score** of **0.3006** indicates that clusters are not well-separated, with overlapping cluster boundaries.
- A **DB Index** of **1.0338** reflects poor compactness and separation.

Strengths:

- Flexibility in modeling clusters of varying shapes and sizes.
- Soft assignments provide probabilities for points belonging to each cluster.

Limitations:

- Overfitting due to the large number of clusters.
- High computational cost compared to K-Means and DBSCAN.

Suggestions:

- Reduce the number of clusters to avoid overfitting.
- Experiment with different covariance types (spherical, diag, full) for better results.
- Perform feature scaling and dimensionality reduction before applying GMM.

Overall Analysis

Algorithm	Clusters	Silhouette Score	DB Index	Performance Summary
K-Means	4	0.3991	0.8001	Moderate performance
DBSCAN	2	0.3824	1.0167	Poor separation of clusters
Agglomerative	2	0.5915	0.2580	Best performance overall
Gaussian Mixture	10	0.3006	1.0338	Overfitting, poor separation