Task 03: Clustering Report: Analysis and Results

Introduction:

Clustering is an unsupervised machine learning technique used to group similar data points together. In this report, we present the results of clustering performed on a dataset using different clustering algorithms: Agglomerative Clustering, Gaussian Mixture Model (GMM), DBSCAN, and K-Means. The task was to identify natural groupings in the data and evaluate the quality of these groupings using various clustering metrics.Taking 5 cluster points for each model.

Steps of Clustering Process:

1. **Data Preprocessing:**
    - **Handling Missing Values:** Any missing data points were either filled using mean imputation or removed from the dataset.
    - **Scaling:** The numerical features were scaled to ensure that the clustering algorithm wasn't biased by features with larger scales. This was done using the MinMaxScaler to normalize values between 0 and 1.

2. **Choosing the Clustering Algorithm:**
    - **Agglomerative Clustering:** A hierarchical clustering algorithm that builds the cluster tree bottom-up by merging the closest pairs of clusters.
    - **Gaussian Mixture Model (GMM):** A probabilistic model that assumes all data points are generated from a mixture of several Gaussian distributions.
    - **DBSCAN:** A density-based spatial clustering algorithm that groups together points that are close to each other based on a distance measurement.
    - **K-Means:** A centroid-based algorithm that divides the data into a predefined number of clusters by minimizing the variance within each cluster.
3. **Model Training:** The clustering algorithms were trained on the preprocessed dataset, with various configurations to achieve meaningful clusters.
4. **Evaluating Clustering Performance:** The clustering results were evaluated using the following metrics:
    - **Davies-Bouldin Index (DBI):** Measures the average similarity ratio of each cluster with the cluster that is most similar to it. Lower values indicate better clustering.
    - **Silhouette Score:** Measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A higher score indicates better-defined clusters.
    - **Calinski-Harabasz Index:** Evaluates clustering based on the ratio of between-cluster dispersion to within-cluster dispersion. A higher score indicates better clustering.
    - **Inertia:** Measures the compactness of the clusters. A lower value indicates tighter, more compact clusters.
    - **Dunn Index:** Measures the separation between clusters. A higher Dunn Index indicates better separation between clusters.

Results Summary:

The following are the evaluation metrics for each clustering algorithm:

1. **Agglomerative Clustering:**
    - **Davies-Bouldin Index (DBI):** 1.0718 (lower value is better)
    - **Silhouette Score:** 0.3103 (closer to 1 is better)
    - **Calinski-Harabasz Index:** 86.5253 (higher value is better)
    - **Dunn Index:** 0.1842 (higher value is better)
2. **Gaussian Mixture Model (GMM):**
    - **Davies-Bouldin Index (DBI):** 1.2998 (lower value is better)
    - **Silhouette Score:** 0.2509 (closer to 1 is better)
    - **Calinski-Harabasz Index:** 67.8857 (higher value is better)
    - **Inertia:** 11.570 (lower value is better)
    - **Dunn Index:** 0.0758 (higher value is better)

3. **DBSCAN:**
    - **Davies-Bouldin Index (DBI):** 0.8969 (lower value is better)
    - **Silhouette Score:** 0.2858 (closer to 1 is better)
    - **Calinski-Harabasz Index:** 16.3144 (higher value is better)
    - **Dunn Index:** 0.4988 (higher value is better)
4. **K-Means:**
    - **Davies-Bouldin Index (DBI):** 1.0913 (lower value is better)
    - **Silhouette Score:** 0.2733 (closer to 1 is better)
    - **Calinski-Harabasz Index:** 96.5446 (higher value is better)
    - **Dunn Index:** 0.0595 (higher value is better)

Visualizing the Results:

- A 2D scatter plot of the data points was created for each clustering algorithm, where each point is colored according to its assigned cluster. This visualization helps in understanding how well the data points are grouped into clusters.

Interpreting the Clusters:

- **Agglomerative Clustering:** The results showed a moderate clustering performance, with a decent Calinski-Harabasz Index, suggesting good clustering dispersion between clusters. However, the low Silhouette Score indicates that the separation between clusters could be improved.
- **Gaussian Mixture Model (GMM):** The GMM provided a more probabilistic approach to clustering but produced relatively less compact clusters compared to Agglomerative Clustering and DBSCAN. The low Silhouette Score suggests that the clusters weren't well-separated.
- **DBSCAN:** The results for DBSCAN were impressive with a lower Davies-Bouldin Index, indicating compact and distinct clusters. The higher Dunn Index showed better separation between clusters, though the Silhouette Score was relatively low, suggesting that the separation between clusters could be enhanced further. This algorithm excels in handling clusters with varying densities and shapes.
- **K-Means:** K-Means clustering performed moderately, with a relatively high Davies-Bouldin Index and Silhouette Score. Although the Calinski-Harabasz Index was the highest among the algorithms, indicating good overall dispersion, the Dunn Index was the lowest, showing poor separation between clusters. This suggests that while the clusters are well-defined in terms of intra-cluster variance, their separation is not optimal.

Conclusion:

Based on the evaluation metrics, **DBSCAN** demonstrated the best clustering performance in terms of cluster compactness and separation, though further tuning of its parameters (eps and min_samples) could potentially improve the Silhouette Score. **Agglomerative Clustering** and **Gaussian Mixture Model** showed moderate performance. **K-Means**, while performing well in terms of dispersion, struggled with cluster separation and compactness, making it less effective in this particular case.

DBSCAN's ability to handle clusters of varying shapes and densities provided a significant advantage, especially when dealing with non-convex clusters. For improving cluster separation in DBSCAN, fine-tuning its parameters could be key. Agglomerative Clustering and GMM could be considered for datasets with more well-separated, compact clusters.