Clustering Analysis Report

Overview

This report presents an in-depth analysis of clustering results obtained using the DBSCAN algorithm. Key performance metrics, optimal parameters, and visualizations are detailed to assess the quality and structure of the clusters formed.

1. Number of Clusters Formed

The clustering algorithm utilized was DBSCAN. The optimal parameters, determined through evaluation, were:
- Epsilon (ε): 0.1
- Minimum Samples: 6
The number of clusters formed, excluding noise points (label = -1), is inferred from the dataset and visualizations.

2. Davies-Bouldin (DB) Index

Best DB Index: 0.0358
A lower Davies-Bouldin Index indicates better clustering quality, characterized by compact and well-separated clusters.

3. Other Clustering Metrics

The following metrics were used to evaluate clustering performance:

- Silhouette Score: 0.9727
  This high value indicates dense and well-separated clusters.

- Calinski-Harabasz Score: 7777.16
  A high score reflects well-defined clusters.

- Adjusted Rand Index (ARI): 1.0
  Demonstrates perfect agreement between the predicted clusters and the true labels.

- Homogeneity Score: 1.0
  Indicates that each cluster contains only data points belonging to a single class.

- Completeness Score: 1.0
  Confirms that all data points of a single class are assigned to the same cluster.

- V-Measure Score: 1.0
  Achieves a perfect combined score of homogeneity and completeness.

- Fowlkes-Mallows Index: 1.0
  Suggests a balanced trade-off between precision and recall in clustering.

4. Principal Component Analysis (PCA)

PCA was employed to reduce the dataset's dimensionality while retaining 95% of the variance.
Number of PCA components selected: 4

5. Optimal DBSCAN Parameters

- Epsilon (ε): 0.1
- Minimum Samples: 6

These parameters were determined to yield the best clustering results based on the evaluation metrics.

6. Visualization of Clusters

A scatterplot depicting the clustering results was generated using the first two PCA components.
Noise points (label = -1) were excluded from the visualization for clarity.
The plot demonstrates distinct cluster boundaries, further confirming the effectiveness of the chosen DBSCAN parameters.