**Clustering Flow Diagram**

A diagram of a diagram

Description automatically generated

**Observation**

The clustering algorithms have been pre-processed using Principal Component Analysis (PCA) for dimensionality reduction. The data has been normalized according to standard scaler. Experimenting with different clustering algorithms has yielded that there is a significant improvement of 30% on the internal cluster validity (silhouette score) when combined with relative cluster validity (elbow method). The parameters for DBSCAN were fine tuned with optimal clusters from elbow method to align with data density improve cluster structure, leading to the highest Silhouette Score.

**t-SNE results**

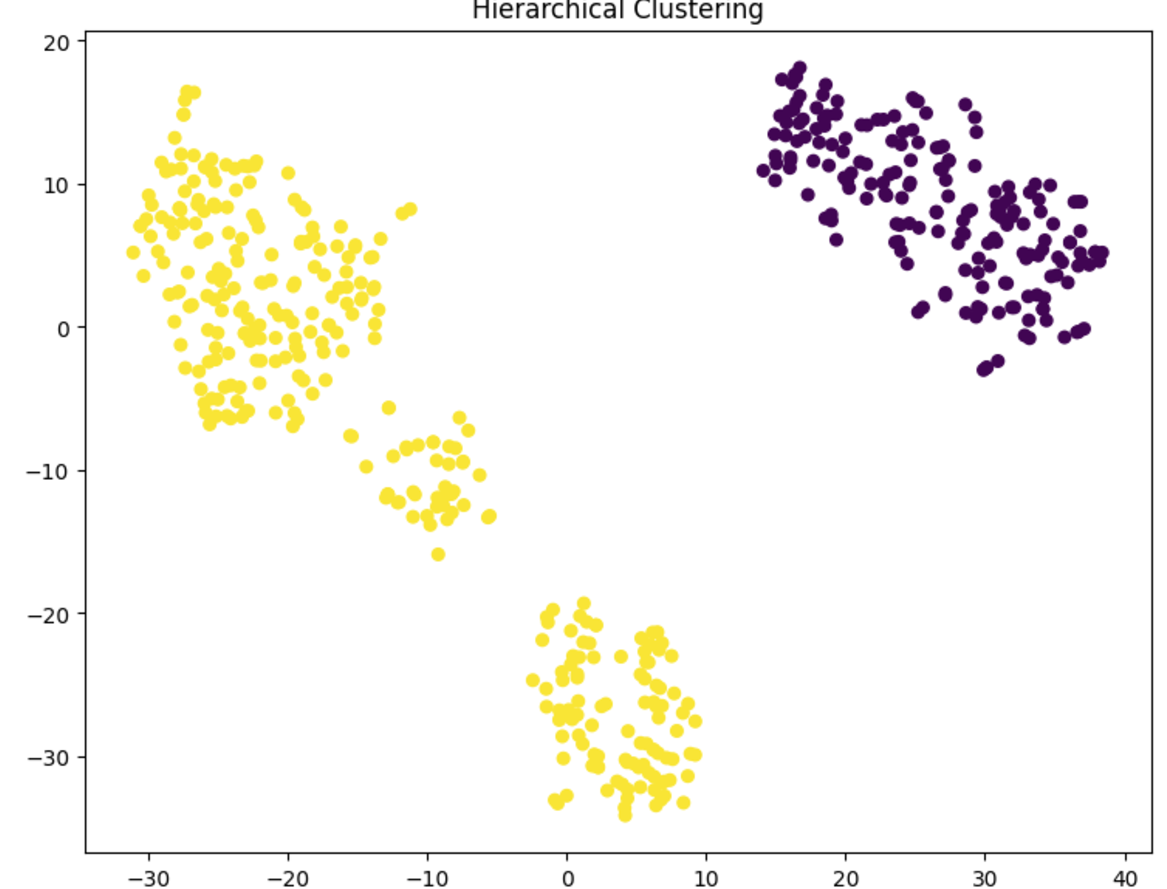
A diagram of a clustering gene

Description automatically generated with medium confidence

A yellow and purple dots

Description automatically generated

0.4586



A chart of yellow and purple dots

Description automatically generated

**Silhouette scores**

|  |  |  |
| --- | --- | --- |
| Clustering Algorithms | Internal Cluster Validity (Silhouette Score) | Relative Cluster Validity (Elbow method) |
| K-Means Algorithm | 0.4478 | Optimal k |
| Gaussian Mixture Model | 0.4430 | Optimal k |
| Hierarchical Clustering | 0.4250 | Optimal k |
| DBSCAN | 0.4586 | Evaluated silhouette score for each combination of eps and minimum samples using grid search. |