HERIOT WATT UNIVERSITY

Group

- Dubai CW PG Thursday Group 6

Group Members

- Pratibha Yadubanshi
- Mohamed Aman
- Faizan Watare
- Fardeen Khan

F21DL Coursework Part 2 – Clustering

In this report, we present the results and findings from Part 2 of the coursework. The primary objectives were to apply various clustering algorithms, explore their performance, evaluate the optimal number of clusters, and compare clustering results with Bayesian classification. This part of the coursework allowed us to conduct experiments based on the lectures from week 7 – 9.

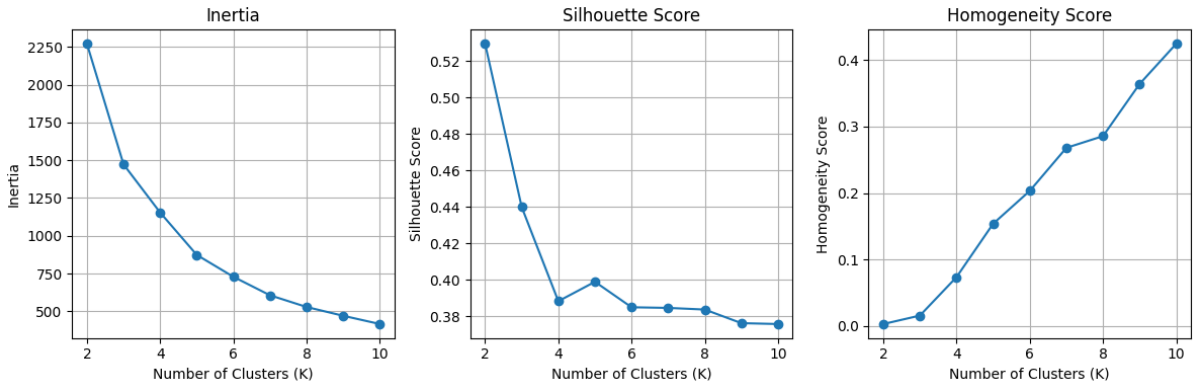
Data Preprocessing

We imported the complete dataset and extracted only 2 classes from it (class 1 and 2) as both had almost same number of datapoints (around 2250). Then applied feature selection based on correlation. The top 10 features for each class were selected as it gave the highest scores in our part1 of the coursework. We normalized the dataset using Min-Max scaler and used principal component analysis (PCA) to reduce the dimensionality to two components for selecting most infomative featuers and easy visualization. This dimensionality reduced dataset is used for further analysis.

K-Means Clustering

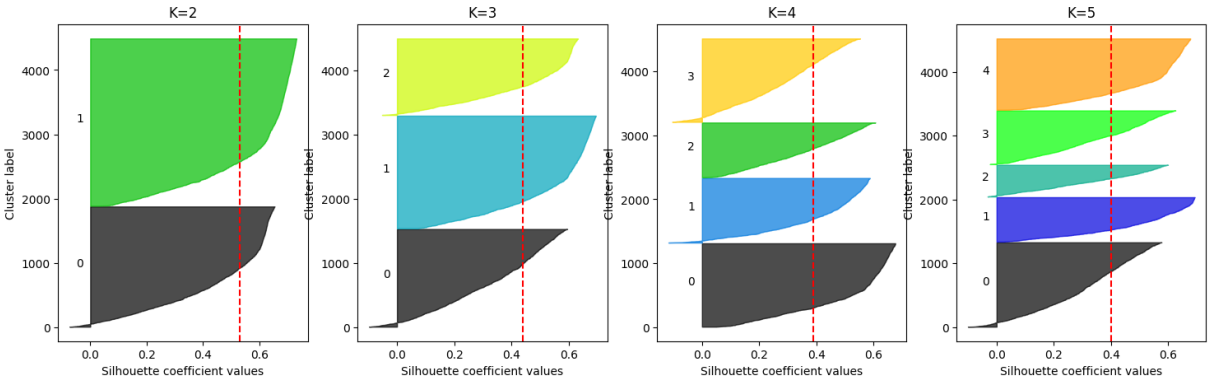
We began our clustering analysis by using the K-Means algorithm. After applying K-Means clustering with varying values of k (number of clusters), we assessed the quality of the clustering through three key metrics: Inertia, Silhouette Score, and Homogeneity Score. These metrics provided insights into the effectiveness and the ability of K-Means to correctly group data points. The results are;

	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
Inertia	2270	1472	1153	873	728	606	528	470	413
Silhouette Score	0.53	0.44	0.39	0.40	0.38	0.38	0.38	0.38	0.38
Homogeneity Score	0.003	0.016	0.074	0.154	0.203	0.268	0.286	0.364	0.425



These metrics revealed that K-Means performed relatively well for k=2, with a high Silhouette Score suggesting meaningful clustering. The low Inertia indicated tight clusters. However, as k increased, the clustering quality seemed to deteriorate.

To evaluate the quality of clusters generated by K-Means, we used the silhouette score, which measures how well data points are assigned to their own cluster compared to other clusters. Below is the visualization of silhouette scores for different values of k (number of clusters). This visualization allowed us to identify the optimal number of clusters, as indicated by the peak silhouette score:



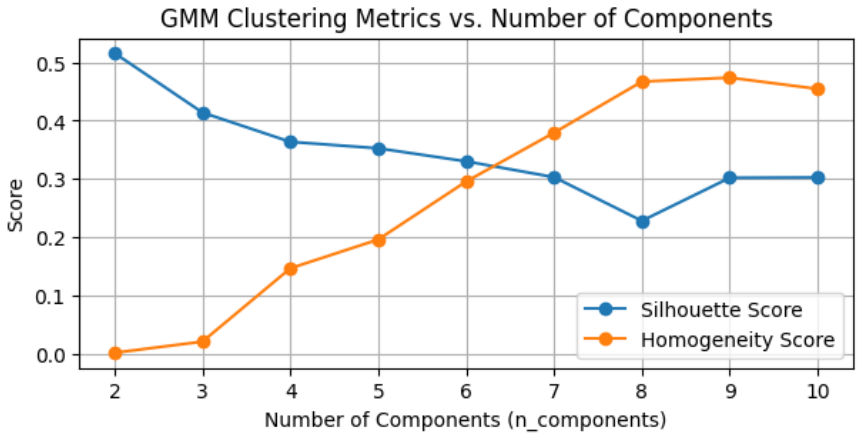
As observed in the plot, the silhouette score is highest when k=2. This suggests that dividing the data into two clusters is most appropriate, as it maximizes the separation between clusters while ensuring the consistency of data points within each cluster.

Gaussian Mixture Model (GMM)

We extended our analysis by applying a Gaussian Mixture Model (GMM). Similar to K-Means, we evaluated GMM using metrics like Silhouette Score and Homogeneity Score, but GMM also provided insights into the number of components that best fit the data.

The results for GMM were as follows:

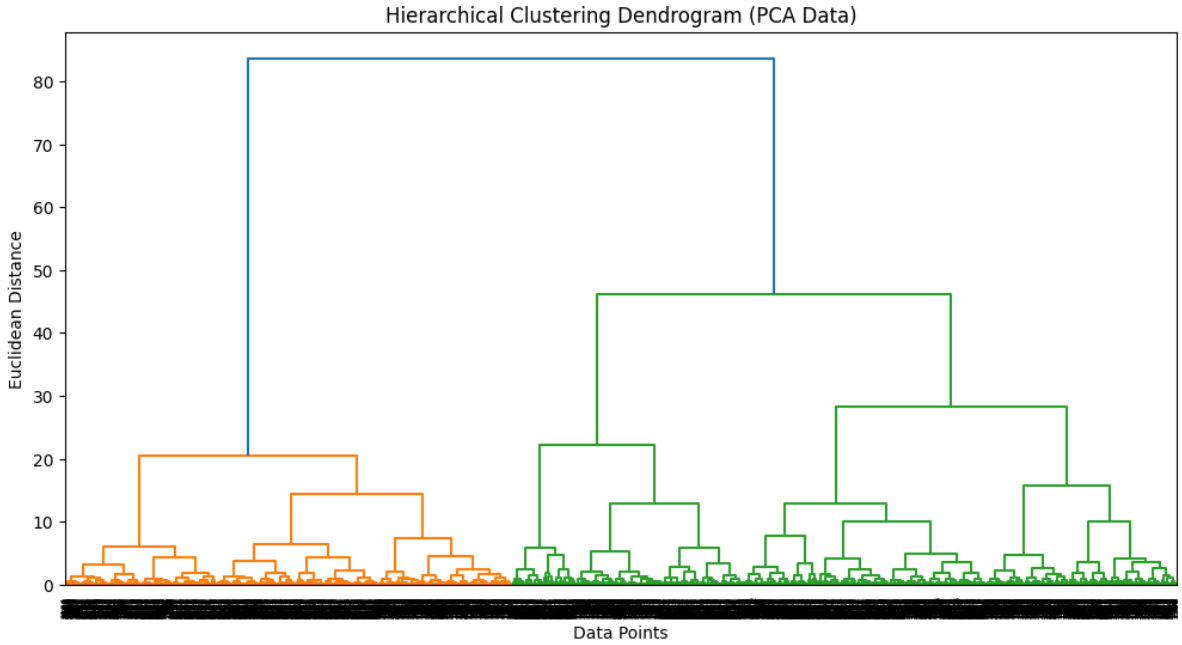
	n_components								
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Silhouette Score	0.52	0.41	0.36	0.35	0.33	0.30	0.23	0.30	0.30
Homogeneity Score	0.001	0.020	0.146	0.196	0.295	0.379	0.467	0.474	0.454



GMM revealed a positive performance, particularly with n_components=2, which was consistent with the results from K-Means. GMM successfully estimated the number of components that best represented the data. However, the silhouette scores for GMM seemed to be lower compared to K-Means.

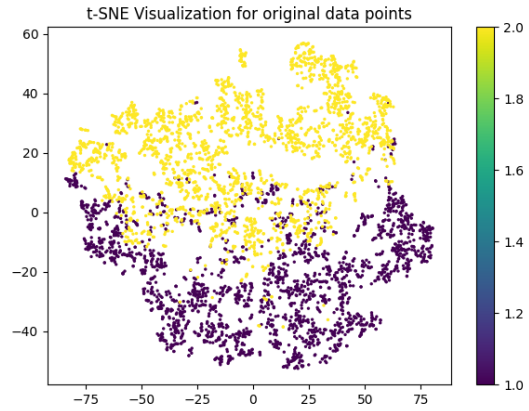
Hierarchical Clustering

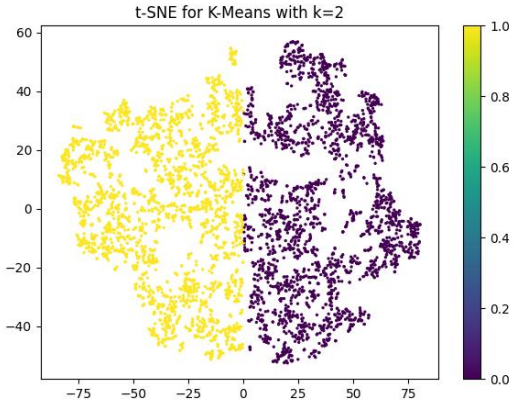
In this step, we ran hierarchical clustering and visualized the results through a dendrogram. This hierarchical approach provided insights into the hierarchical structure of the data and allowed us to identify possible clusters within data.



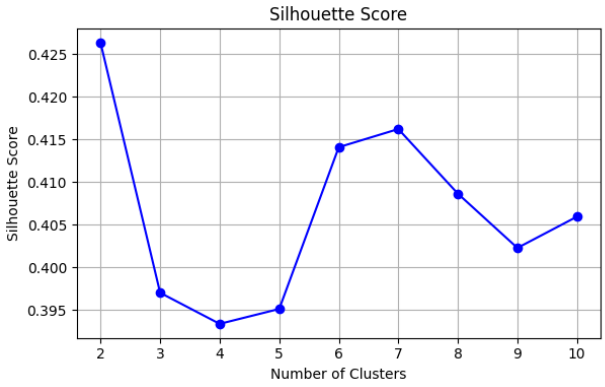
Additional Analysis using t-SNE

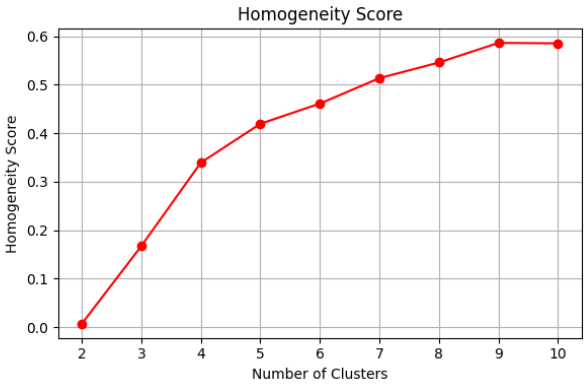
As an additional exploration, we applied t-SNE, a dimensionality reduction technique that aimed to visualize the data in a lower-dimensional space. We visualized the original data points using t-SNE visualization. We then applied K-Means clustering to the t-SNE-transformed data. We experimented with different values of K to identify the optimal number of clusters.





The t-SNE metrics (silhouette score and homogeneity score) for varying values of k (the number of clusters) were as follows:





Conclusion

K-Means and GMM both demonstrated relatively good performance with k=2, indicating meaningful clustering within the data. The Silhouette Score and Homogeneity Score suggested that this was the optimal number of clusters for our dataset.

Hierarchical clustering provided a hierarchical view of the data, offering potential insights into its structure.

t-SNE provided meaningful visualizations.

- Github Link: https://github.com/amannuhman/F21DL-CW

Thank You