**Assignment – 4**

# Agglomerative Hierarchical Clustering

**Problem statement:** The objective of this project is to conduct k-means clustering on a collection of 8 points in a two-dimensional space. The clustering process begins with the definition of initial centroids, set as P1 = [0.1,0.6] and P8 = [0.3,0.2]. After clustering, we determine the cluster to which point P6 = [0.25,0.5] belongs, thereby revealing its cluster affiliation. Additionally, the population of the cluster around centroid P8 is ascertained. Finally, the centroids' positions, denoted as m1 and m2, are updated based on the newly formed clusters, refining the clustering solution. The goal is to help us understand how to do clustering using the K- Means Clustering algorithm.

**Software used:**

1. Python 3.x
2. Google Colab

**Libraries and packages used:** NumPy, Matplotlib, scikit-learn, pandas, seaborn , scipy

**Theory:**

**Methodology:**

• Agglomerative Hierarchical Clustering: Agglomerative Hierarchical Clustering is an unsupervised machine learning algorithm that groups similar data points into clusters by iteratively merging them together based on the similarity measure.

• Algorithm Steps:

1. Start with each point as a separate cluster.

2. Calculate the distance between all pairs of clusters.

3. Merge the two closest clusters.

4. Repeat steps 2 and 3 until only one cluster remains.

• Advantages:

1. Simple and intuitive approach to clustering.

2. Does not require the number of clusters to be specified beforehand.

3. Hierarchical structure provides insights into data relationships.

• Disadvantages:

1. Computationally expensive for large datasets.

2. Memory-intensive for storing distance matrices.

3. Difficult to determine the optimal number of clusters.

4. Sensitive to noise and outliers.

5. May not scale well to high-dimensional data.

**Working / Algorithm:**

1)Start with each data point as a separate cluster.

2)Calculate the distance between all pairs of clusters.

3) Merge the two closest clusters.

4)Update the distance matrix.

5)Repeat steps 2-4 until only one cluster remains.

**Conclusion:**

In conclusion, this assignment demonstrates the application of Agglomerative Hierarchical Clustering in partitioning data into clusters based on similarity. We have explored its simplicity, hierarchical nature, and applications across various domains such as gene expression analysis, image segmentation, and document clustering. However, the algorithm's performance is influenced by factors like computational complexity, sensitivity to noise, and the difficulty of determining the optimal number of clusters.