**ASSIGNMENT: - 04**

**Problem Statement: -**

Write a program to do following:

We have given a collection of 8 points. P1= [0.1,0.6] P2= [0.15,0.71] P3= [0.08,0.9] P4= [0.16, 0.85] P5= [0.2,0.3] P6= [0.25,0.5] P7= [0.24,0.1] P8= [0.3,0.2]. Perform the k-mean clustering with initial centroids as m1=P1=Cluster#1=C1 and m2=P8=cluster#2=C2. Answer the following:

a) Which cluster does P6 belong to?

b) What is the population of a cluster around m2?

c) What is the updated value of m1 and m2?

**Software, library and package**

1. Software: Python
2. Library: scikit-learn (sklearn) - for k-means clustering algorithm
3. Package: NumPy (implicitly used by scikit-learn for numerical computations)

**Theory:**

**Methodology**: K-means clustering is an unsupervised machine learning algorithm that partitions data into K clusters based on similarity. The algorithm iteratively assigns data points to the nearest cluster centroid and updates the centroids until convergence.

**Advantages**:

1. **Simplicity**: K-means is easy to implement and understand, making it suitable for exploratory data analysis.
2. **Scalability**: It can handle large datasets efficiently.
3. **Versatility**: Works well with numerical data and is robust to outliers.

**Applications**:

1. **Customer Segmentation**: Identifying customer segments based on purchasing behavior.
2. **Image Compression**: Clustering similar pixels to compress images.
3. **Anomaly Detection**: Identifying outliers or anomalies in data.
4. **Recommendation Systems**: Grouping similar items or users in recommendation engines.

**Limitations**:

1. **Number of Clusters (K)**: The user needs to specify the number of clusters, which may not always be known a priori.
2. **Sensitive to Initial Centroids**: Results can vary based on the initial centroid placement.
3. **Assumes Spherical Clusters**: K-means works best with spherical clusters and struggles with non-linear or irregularly shaped clusters.

**Working/ Algorithm:**

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means reassigning each datapoint to the new closest centroid of each cluster.

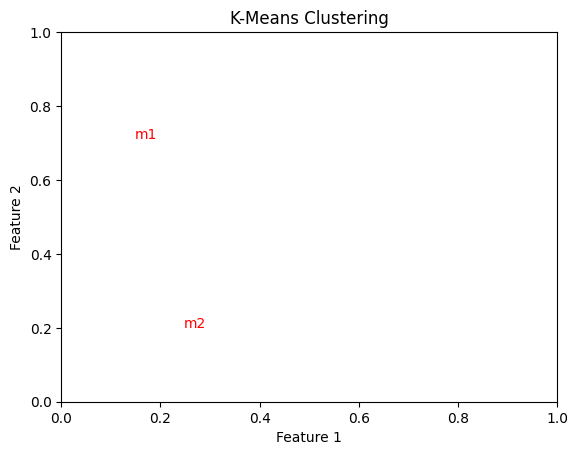
Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready

**Diagram:**

****





**Conclusion:**

In conclusion, k-means clustering is a versatile and easy-to-implement algorithm for unsupervised clustering tasks. Its advantages include simplicity, scalability, and versatility in handling large datasets and numerical data. However, it requires specifying the number of clusters (K), is sensitive to initial centroid placement, and assumes spherical clusters. Despite its limitations, k-means finds applications in various domains like customer segmentation, image compression, anomaly detection, and recommendation systems.