Assignment 4

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 belongs to?
b) What is the population of cluster around m2?
c) What is updated value of m1 and m2?.

S/W Packages and H/W apparatus used:

OS: **Ubuntu/Windows, Google Colab**
Packages: **Numpy, Pandas, Matplotlib and SkLearn**

Theory :

1. Methodology:

- **K-Means Clustering** : K-Means Clustering is an unsupervised learning algorithm that is used to solve clustering problems in machine learning or data science
- **What is the K-Means Algorithm?** :
    - K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
- It allows us to cluster the data into different groups and is a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

- It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters. The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.
- **The k-means clustering algorithm mainly performs two tasks**:
    - Determines the best value for K center points or centroids by an iterative process.
    - Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster. Hence each cluster has data points with some commonalities, and it is away from other clusters.

2. Advantages and Disadvantages & Limitation/Example:

1. **Advantages**:
    - **Simple and Intuitive:** K-means clustering is easy to understand and implement.
    - **Efficient**: It works well for large datasets and can handle high-dimensional data efficiently.
    - **Scalability:** K-means scales well with increasing dataset sizes.
    - **Interpretability:** Results are straightforward and easy to interpret.
2. **Disadvantages & Limitations/Example:**
    - **Sensitivity to Initial Centroids**: Results can vary depending on the initial centroid selection.
    - **Assumption of Spherical Clusters**: K-means assumes that clusters are spherical, which may not always be the case.
    - **Impact of Outliers**: Outliers can significantly affect the cluster centroids and result in suboptimal clustering.

- **Determining Number of Clusters**: The number of clusters needs to be specified beforehand, which can be subjective and challenging to determine.

Algorithm :

The working of the K-Means algorithm is explained in the below steps:

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

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

● **Step 3:** Assign each data point to its closest centroid, which will form the predefined K clusters.

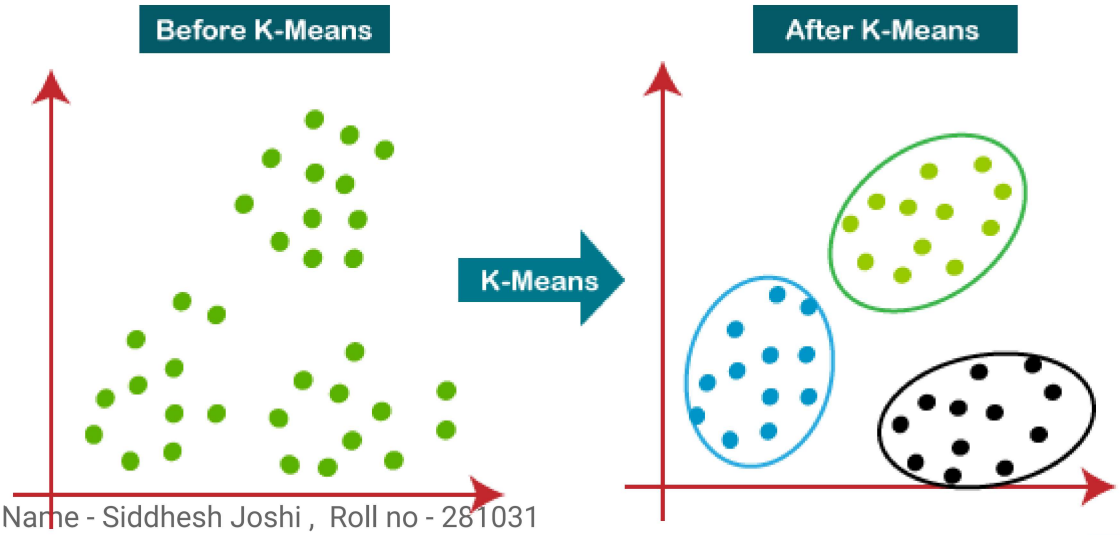
● **Step 4:** Calculate the variance and place a new centroid of each cluster.

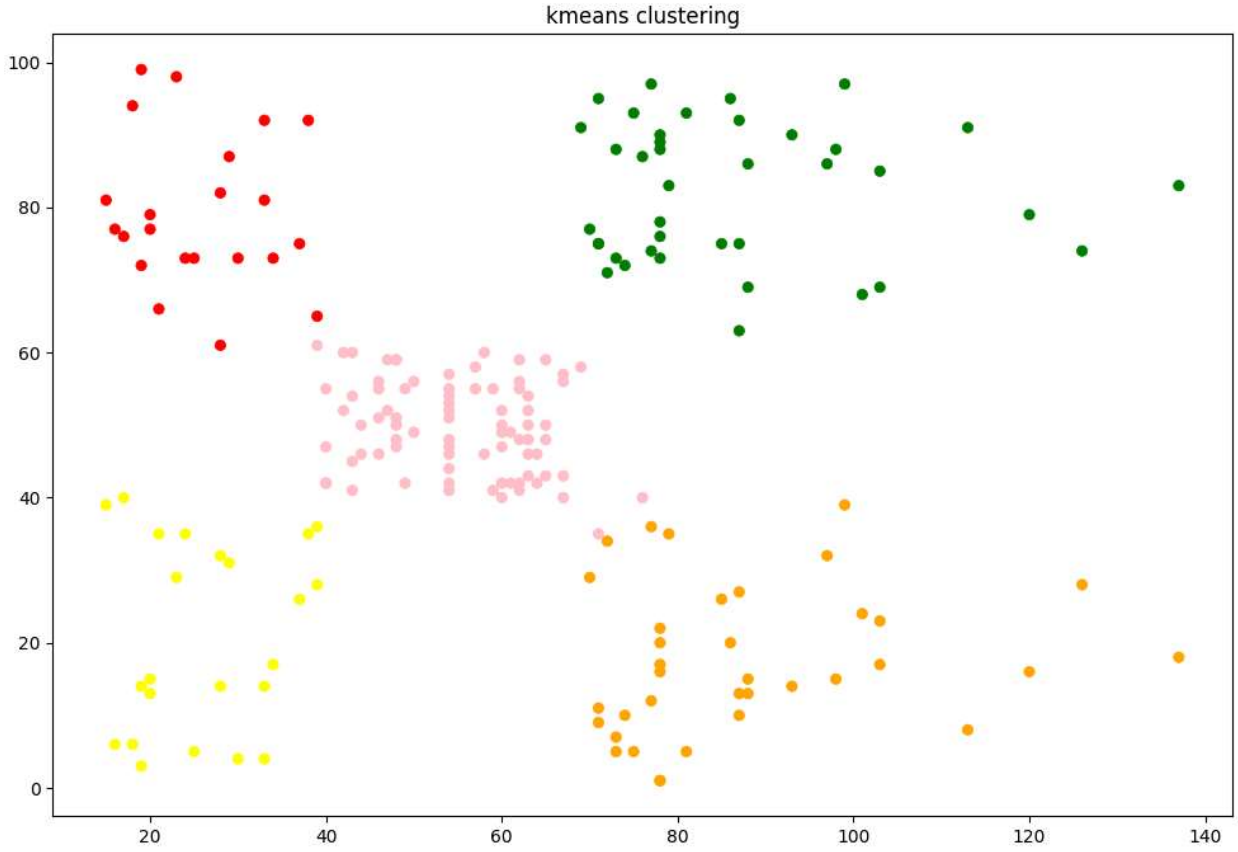
● **Step-5**: Repeat the third step, which means reassigning each datapoint to the new closest centroid of each cluster.

● **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 simple yet effective method for grouping data points into clusters. While it's straightforward to implement and scales well with large datasets, it requires careful consideration of initial centroids and assumes spherical clusters. Despite its limitations, K-means remains a popular choice for clustering tasks due to its efficiency and interpretability.