LP-III Machine Learning (2024-25) **Assignment 4:** Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method.	
Student Name:	**Roll No. :**
Batch:	**Division :**

Assignment 4:
Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method.

Title: Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset using dataset available at https://www.kaggle.com/datasets/kyanyoga/sample-sales-data.

Aim: Implement K-Means clustering/ hierarchical clustering.

Prerequisites: Linear Regression, Random Forest, Decision Tree.

Theory:

Clustering:

Problem involves assigning the input into two or more clusters based on feature similarity. Similar groups based on their interests, age, geography, etc can be done by using Unsupervised Learning algorithms. A way of grouping the data points into different clusters, consisting of similar data points. It does it by finding some similar patterns in the dataset such as shape, size, color, behavior, etc., and divides them as per the presence and absence of those similar patterns. It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset. After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML system can use this id to simplify the processing of large and complex datasets.

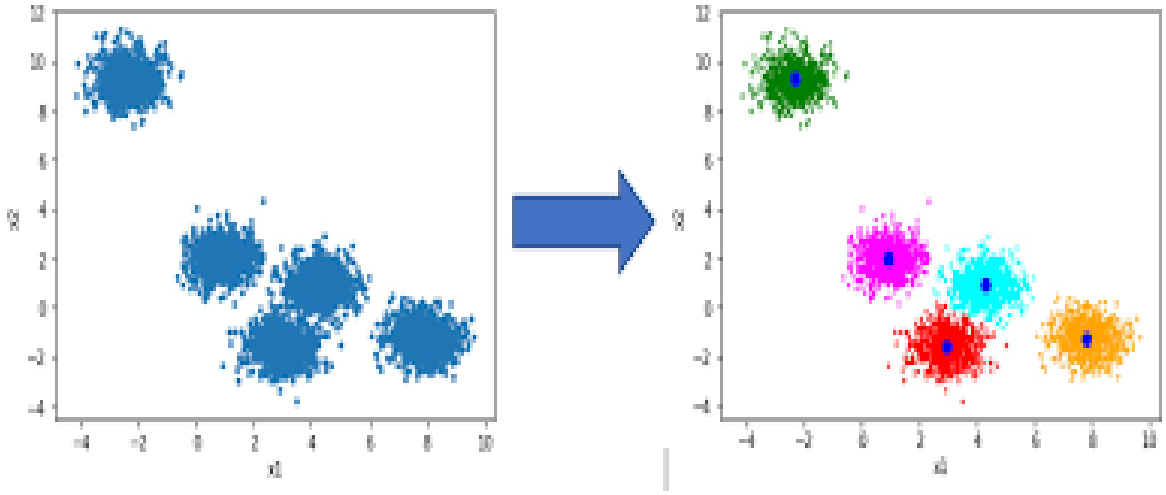

Figure 1: Clustering Algorithm

The clustering technique can be divided into following types:

1. Partitioning Clustering (Centroid-based Clustering)
2. Density-based Clustering
3. Hierarchical Clustering
4. Distributed Model-based Clustering
5. Fuzzy Clustering

K-means Clustering:

K-Means Clustering is an Unsupervised Learning Algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined 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 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.

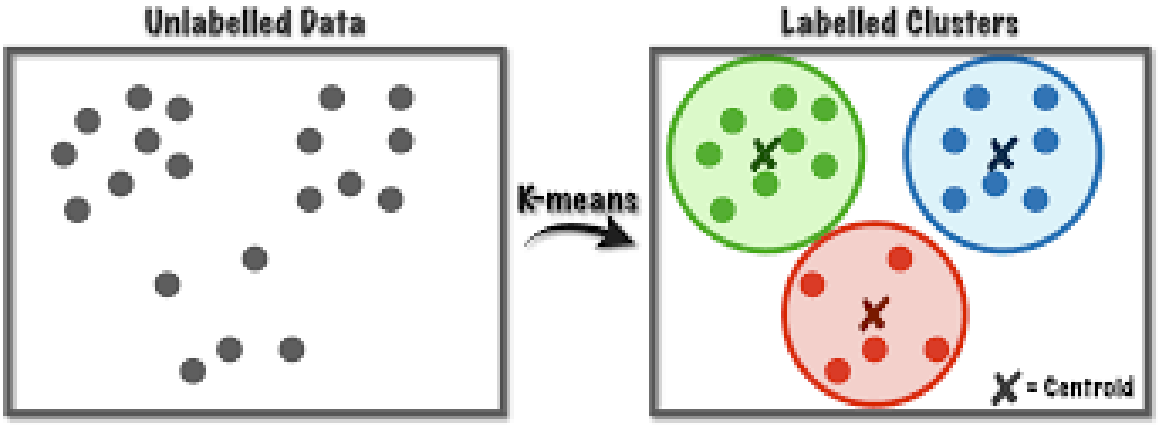

Figure 2: K-means Clustering Algorithm

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 reassign each data point 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.

Conclusion: We have implemented K-means clustering algorithms on sales data and determined number of clusters using Elbow method.

Questions:

1. What is Clustering?

2. Explain any one distance metric?

3. What is Elbow method?

4. Explain significance of k in k-means clustering?

5. List out types of clustering algorithms?