Name: Manisha Apchunde

Roll no: 281076

PRN:22420244

Batch: A3

Assignment 04

k-means clustering algorithm

Aim:

Apply appropriate ML algorithm on a dataset collected in a cosmetics shop showing details of customers to predict customer response for special offer. Create confusion matrix based on above data and find

a) Accuracy

b) Precision

c) Recall

d) F-1 score

Software used:

Programming Language: Python

1. Visual Studio Code

Libraries and packages used: NumPy, Matplotlib, scikit-learn

Theory:

Methodology:

- K-means Clustering is a popular unsupervised machine learning algorithm used for partitioning data into distinct clusters. It 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.
- The algorithm aims to minimize the variance within each cluster while maximizing the variance between clusters. The process involves iteratively

assigning data points to the nearest cluster centroid and updating the centroids based on the mean of the points assigned to each cluster.

- The k-means clustering algorithm mainly performs two tasks:
    1. Determines the best value for K center points or centroids by an iterative process.
    2. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Advantages:

1. Simplicity: K-means is straightforward to implement and easy to understand.
2. Efficiency: It is computationally efficient and scales well to large datasets.
3. Versatility: Suitable for a wide range of applications and data types.
4. Scalability: Performs well even with a large number of dimensions.
5. Interpretability: Results are easily interpretable, especially with low-dimensional data.

Disadvantages:

1. The choice of initial centroids can impact the final clustering results.
2. The algorithm requires specifying the number of clusters beforehand.
3. K-means assumes that clusters are spherical and of similar size.
4. Outliers can significantly affect the cluster centroids and the overall clustering outcome.

Applications:

1. Customer Segmentation – Group customers by behavior or demographics.
2. Medical Imaging – Identify regions in scans (e.g., tumors).
3. Market Basket Analysis – Cluster products often bought together.
4. Geographical Mapping – Group locations by population or delivery zones.
5. Gene Expression – Cluster similar gene/protein activity.
6. Document Clustering – Organize articles by topic.

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.

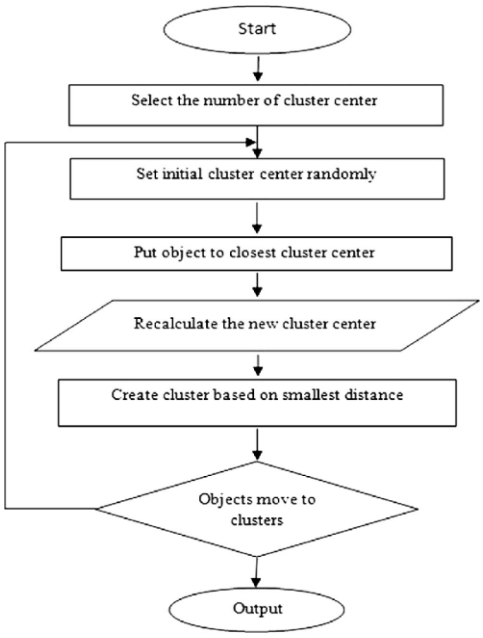
Step-4: Calculate the variance and place a new centroid of each cluster.

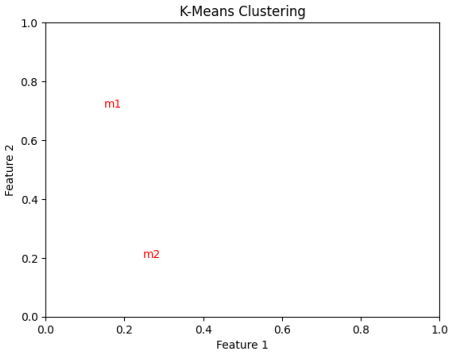
Step-5: Repeat the third steps, which means reassigning each datapoint to the new closest centroid of each cluster.

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, this assignment demonstrates the effectiveness of K-means clustering in partitioning data into distinct clusters based on similarity. We have explored its simplicity, efficiency, and versatility, showcasing its applicability across various domains such as customer segmentation, anomaly detection, and document clustering. However, the algorithm's performance is influenced by factors like initial centroid selection and the determination of the optimal number of clusters.