Experiment No.4

Title: Implementation of K- Means clustering algorithm for a given case study.

Batch: **Roll No.:** **Experiment No.: 4**

Aim: Implementation of K- Means clustering algorithm.

Resources needed: Any RDBMS, Java

Theory:

Cluster analysis or clustering is the process of partitioning a set of data objects (or observations) into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering.

Types of clustering:

Hierarchical algorithms:

Hierarchical algorithms find successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

Partitioning algorithms:

Partitioning algorithms typically determine all clusters at once, but can also be used as divisive algorithms in the hierarchical clustering e.g K-mean, K-medoid.

Density-based clustering algorithms:

Density-based clustering algorithms are devised to discover arbitrary-shaped clusters. In this approach, a cluster is regarded as a region in which the density of data objects exceeds a threshold. DBSCAN and OPTICS are two typical algorithms of this kind.

K-Means clustering Algorithm:

The *k*-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

The basic step of k-means clustering is simple. In the beginning determine number of cluster K and assume the centroid or center of these clusters. Take any random objects as the initial centroids or the first K objects in sequence can also serve as the initial centroids.

Then the K means algorithm will do the three steps below until convergence

Iterate until *stable* (= no object move group):

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance.

Procedure / Approach /Algorithm / Activity Diagram:

1. Download the dataset available at https://www.kaggle.com/code/vineetverma/clustering-bank-complaints/notebook. Identify attributes suitable for applying K-mean clustering
2. Apply some preprocessing techniques to drop some columns.
3. Draw a box plot to visualize the data.
4. Implement K-mean clustering on your dataset for clustering different types of complaints.
5. Evaluate the performance of your algorithm with suitable technique.

Results: (Program printout with output / Document printout as per the format)

Questions:

1. What are advantages and disadvantages of K-means clustering algorithm?

Outcomes:

Conclusion: (Conclusion to be based on the objectives and outcomes achieved)

Grade: AA / AB / BB / BC / CC / CD /DD

Signature of faculty in-charge with date

References:

Books/ Journals/ Websites:

1. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition