Assignment No. 8

Aim :

Implement K-means algorithm for clustering to create a cluster on any dataset using Python

Objectives :

- To understand concept of clustering

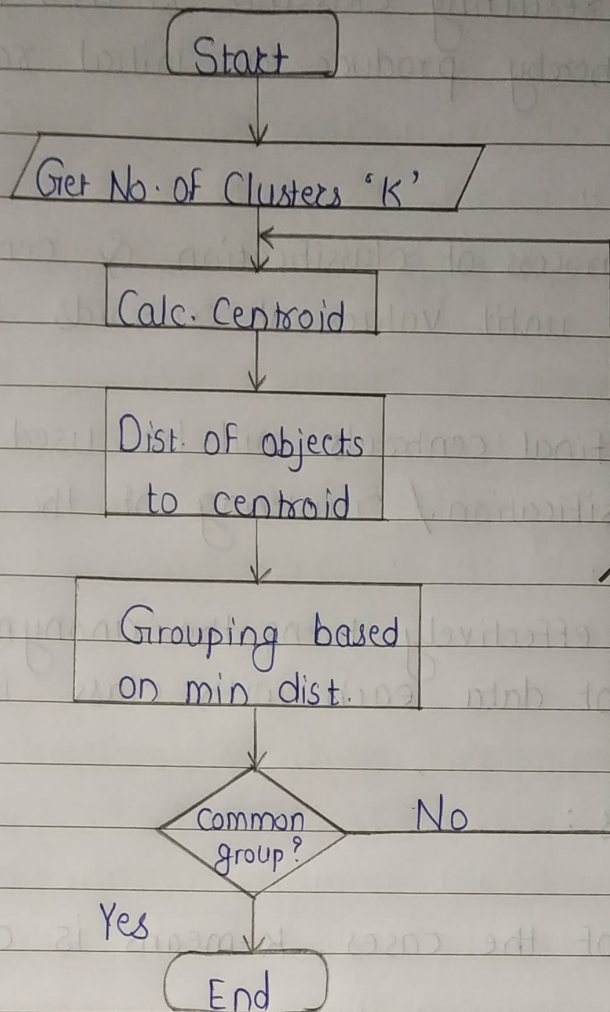- To implement K-means clustering algorithm

Theory :

K-Means Clustering :

It aims to partition any 'n' observations into k clusters in which each observation belong to cluster with nearest mean, serving as prototype of cluster.

The less variation we have within the clusters, the more homogeneous the data points are within same cluster.

It is an iterative algorithm.

Flowchart :

```
                    ┌─────────┐
                    │  Start  │
                    └─────────┘
                         │
                         ↓
         ╱───────────────────────────╲
        ╱  Get No. of Clusters 'K'     ╲
        ╲───────────────────────────╱
                         │
                         ↓
              ┌────────────────────┐
              │  Calc. Centroid    │
              └────────────────────┘
                         │
                         ↓
              ┌────────────────────┐
              │  Dist. of objects  │
              │   to centroid      │
              └────────────────────┘
                         │
                         ↓
              ┌────────────────────┐
              │  Grouping based    │
              │  on min dist.      │
              └────────────────────┘
                         │
                         ↓
                    ◇─────────◇
                   ╱  Common   ╲     No
                   ╲  group ?   ╱───────────►
                    ◇─────────◇
                         │
                       Yes
                         ↓
                    ┌─────────┐
                    │   End   │
                    └─────────┘
```

Working :

K-Means algorithm attempts to split a given dataset into fixed number of clusters.

Initially, K centroids are chosen.

They are picked up randomly in initial stage such that they are all unique.

These centroids are used to train KNN classifier.

The resulting classifier is used to classify the data & thereby produce an initial randomized set of clusters.

The process of classification & centroid adjustment is replaced until values of centroids stabilize.

The final centroids will be used to produce the final classification / clustering of the input data.

This effectively turns the anonymous data points into set of data each with class identity.

Advantages :

- Most of the cases, K-means is computationally faster.

- Relatively simple to implement.

- Scales to large dataset

- Easily adopts to new examples

- Generalizes to cluster of different set shapes & sizes.

Disadvantages :

- Choosing K manually

- Centroid can be dragged by outliers

- Scaling with number of dimenssions

- Dependent on initial values

Conclusion :

Therefore, we have implemented & understood the concept of K-means clustering algorithm which is one of the most important partitioning algorithm.