PAGE NO :
DATE :

K-Means Clustering Algorithm

Input:
A dataset with n data points
$X = \{x_1, x_2, x_3 \ldots x_n\}$
Number of clusters : k

Algorithm:

1. Load the Dataset:
   - Read the .csv file to extract data points.
   - Store the data in a suitable structure

2. Preprocess the Data:
   - Handle missing values.
   - Normalize or standardize the features to bring them to a similar scale.

3. Initialize Centroids:
   - Randomly select k distinct data points from the dataset as initial centroids

4. Repeat until convergence or a set number of iterations

   a. Assign clusters:
      - For each data point, calculate the distance to each centroid
      - Assign the point to the cluster of the nearest centroid,

PAGE NO :
DATE :

6. Update Centroids:
- For each cluster, compute the mean of the points assigned to it.
- Update the centroid with this mean

5. Convergence check:
- If cluster assignments do not change or centroids remain the same, stop

6. Output:
- Final centroids of all clusters
- Cluster labels for each data point.

Principal Component Analysis (PCA)

1. Standardize the Data
- Given a dataset X of size $n \times d$,
n → no of samples.
d → number of features.

$$X_{centered} = X - \mu.$$

2. Compute the covariance Matrix:

- Calculate the covariance matrix of the centered data. This matrix captures the relationship between different features

$$C = \frac{1}{n-1} X_{centered}^T X_{centered}.$$