Assignment 2

Title : Clustering

Problem : Consider a suitable dataset for clustering of data instances in different groups Apply different clustering techniques (minimum 2)

Software and Hardware :

R Studio / Jupiter Notebook
P IV
2 GB RAM
500 GB HDD

Learning Objective :
Use R functions / Scikit learn functions to create K means clustering models and hierarichal clustering models.

Theory :

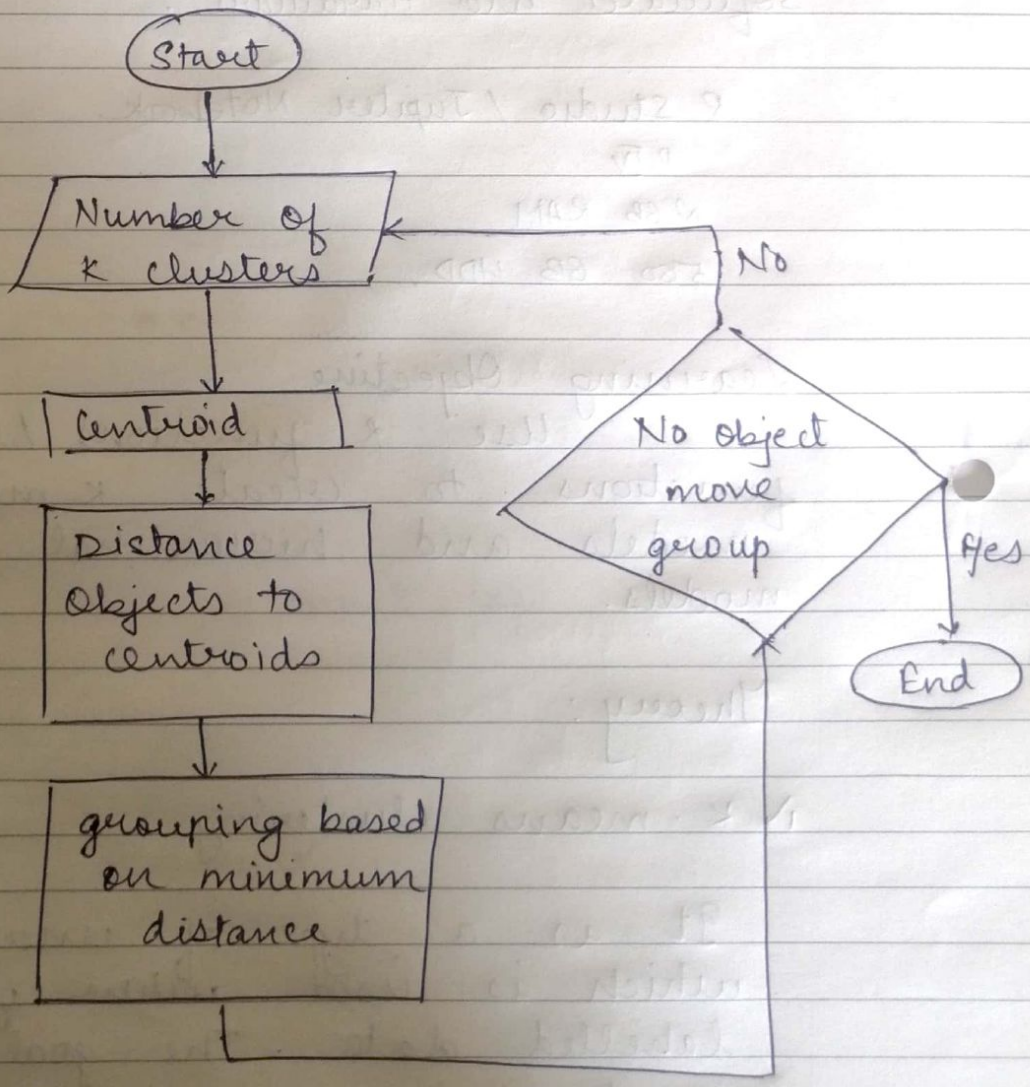
1) K - means clustering :

It is a type of unsupervised learning, which is used when you have unlabelled data. The goal of this algorithm is to find groups in the data, which

of groups represented by the variable K.

The algorithm works iteratively to assign each data point to one of K group based on features that are provided.

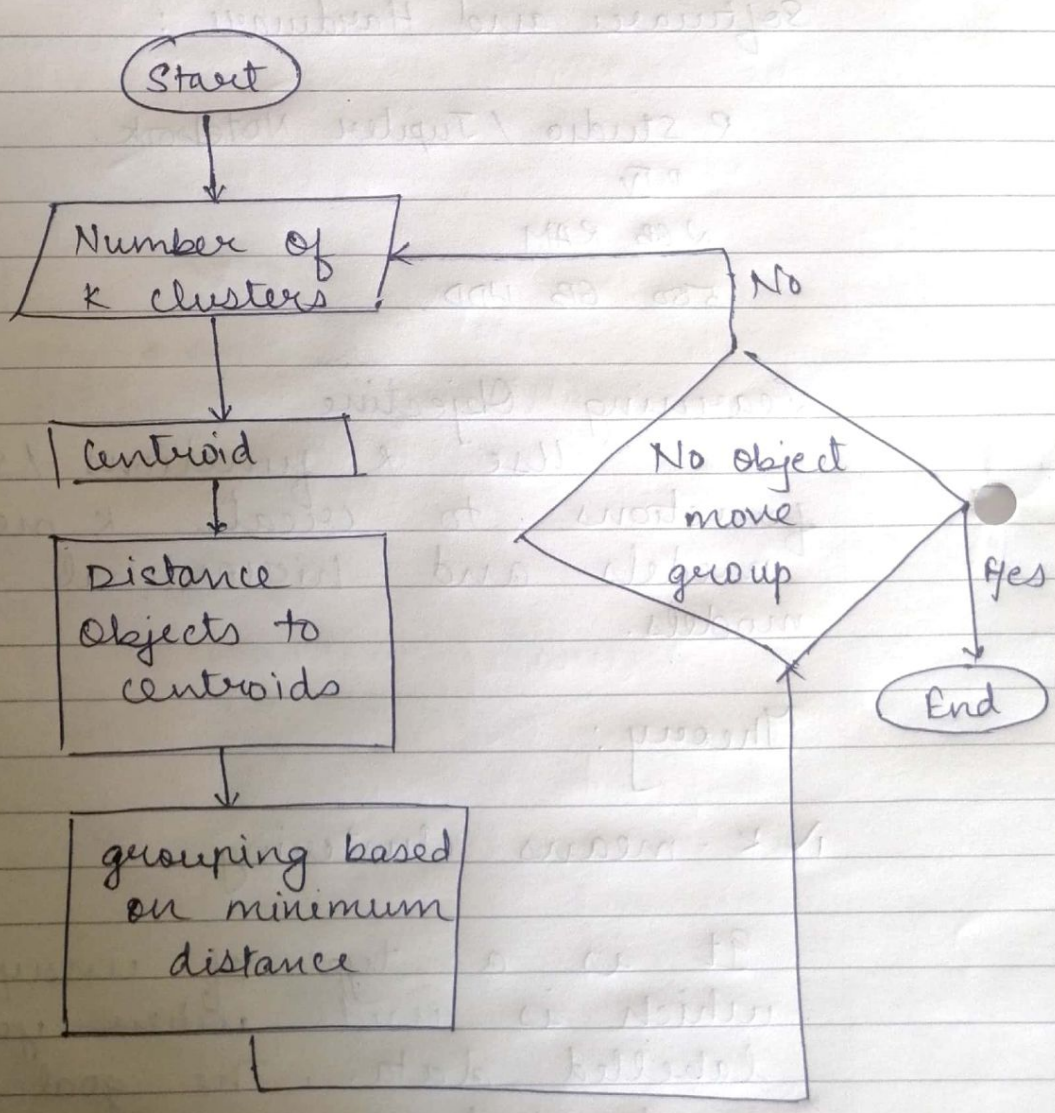
The centroids of the K cubes, which can be used to label new data



of groups represented by the variable K.

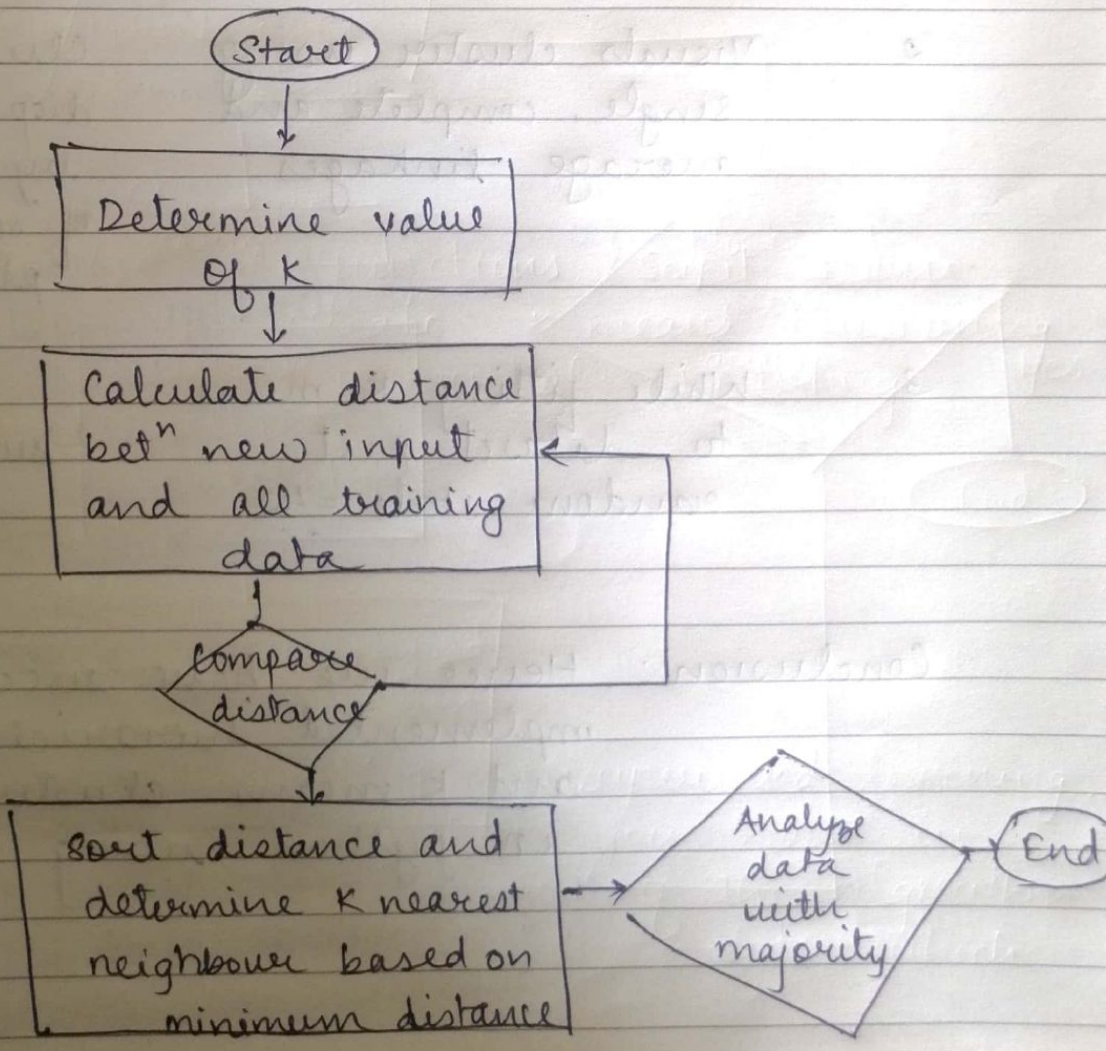
The algorithm works iteratively to assign each data point to one of K group based on features that are provided.

The centroids of the K cubes, which can be used to label new data



K-nearest neighbour (KNN) clustering.

- It is a supervised classification algorithm
- It takes bunch of labelled points and uses them to learn how to label other points.
- To label a new point, it looks at the labelled points closet that new point which are its nearest neighbours and has those neighbours vote.
- Here K in KNN is number of neighbour it checks.



Test Cases

Sr.no	Description	Expected O/p	Actual O/p
1	In KNN clustering method (a unsupervised algo) we created confusion matrix & classification report based on Euclidian distance K clusters are formed	No of clusters rendered = 5	success
2	Visuals cluster using single, complete and average linkages	Clusters displayed by means of scatter plot	success
3	While fitting K means to dataset put random state = 42	success	success

Conclusion: Hence we have successfully implemented hierarichal clustering and K means clustering algorithm in python using jupyter notebooks