Name:- Shantanu S. Mangale
B.E B A1242 DMW

LP - II

Assignment No. 2

Title:- Consider a suitable dataset. for clustering of data instances in different groups, apply different clustering techniques (minimum 2). ~~Viat~~ Visualize the clusters using suitable tool.

Problam Definition:-

Visualize the cluster using Suitable tool.

Pre-requisite:-

i) Basic concepts of ETL.
ii) Knowledge about R tool.

S/W & H/W Requirements:-

R- tool, PIV, 2GB RAB, 500 GB HDD

Learning Objectives:-

Use R functions to create K-means Clustering models & Heirarchical clustering models.

Outcomomes:-

Visualize the effectiveness of the K-means clustering algorithm & hierarchical clustering using graphic capabilities in R.

Theory :-

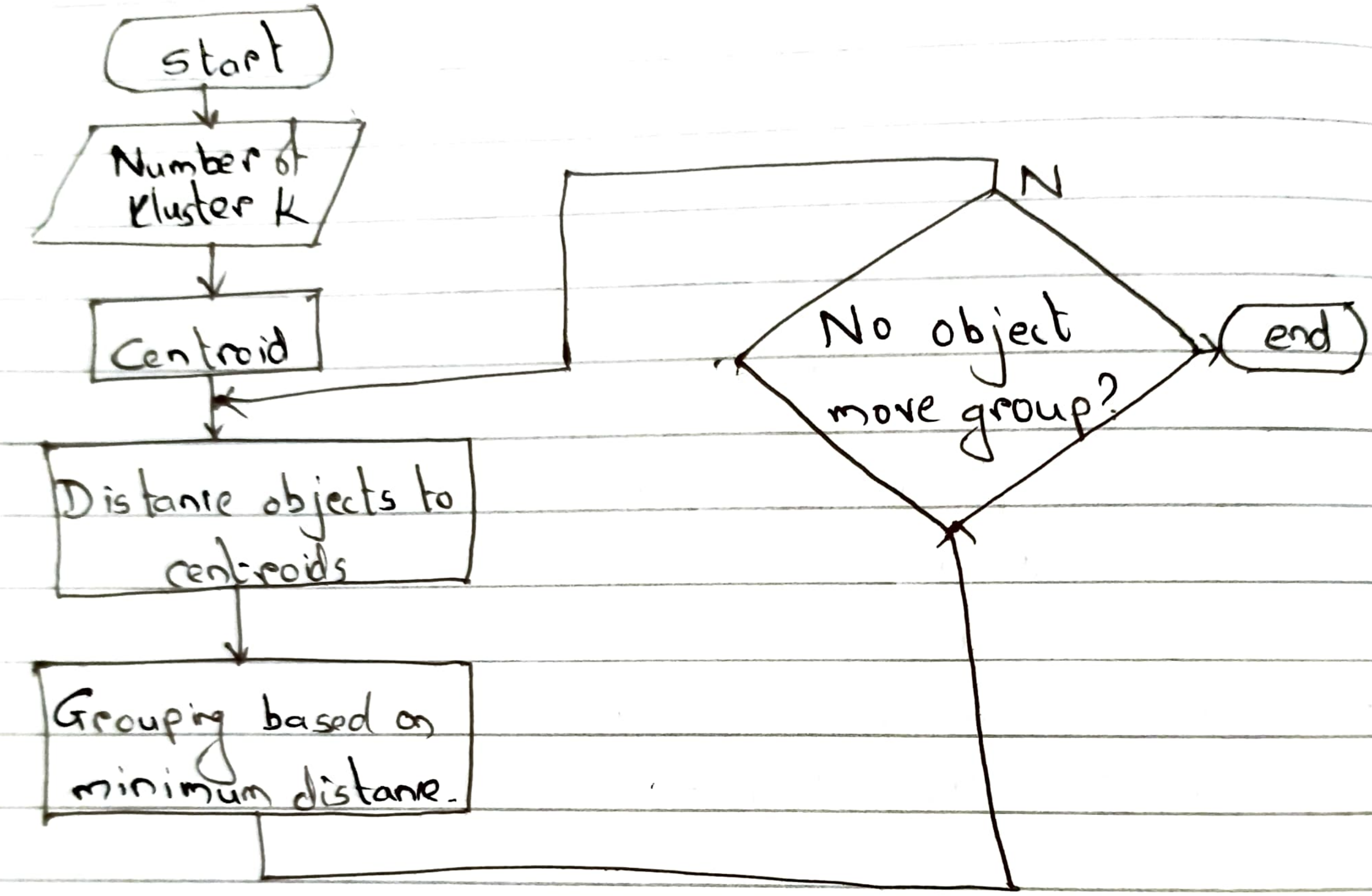
- What is K-means clustering?

K-means clustering is a type of unsupervised learning which is ~~means~~ used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. Data points are clustered based on feature similarity. the results of the K-means clustering algorithm are:

i) The centroids of the K-clusters, which can be used to label new data.

ii) Labels for the training data (each data point is a assigned to a single cluster).

- steps to perform K-means Clustering.



• R Implemementations:

The k-means function, provided by the cluster package is used as follows:

```
k-means (x, centers, iter.max' = 10, nstart = 1,
         algorithm = c("Hartigan-Wong", Lloyd",
         "Forgy", "MacQueen"))
```

where the arguments are:

x → Anumeric matrix of data or an object that can be co-erced to sucha matrix (such as a numeric vector or a data frame with all numeric columns).

centers → Either the number of clusters or a set of initial (distinct) cluster centers. If a number a random set of (distinct) rows in x is chosen as the initial centers

iter.max → The maximum number of iterations allowed.

nstart → If centers is a number, n start gives the number of random sets that should be chosen.

algorithm → The algorithm to be used. It should be one of the following.

- IRIS dataset:

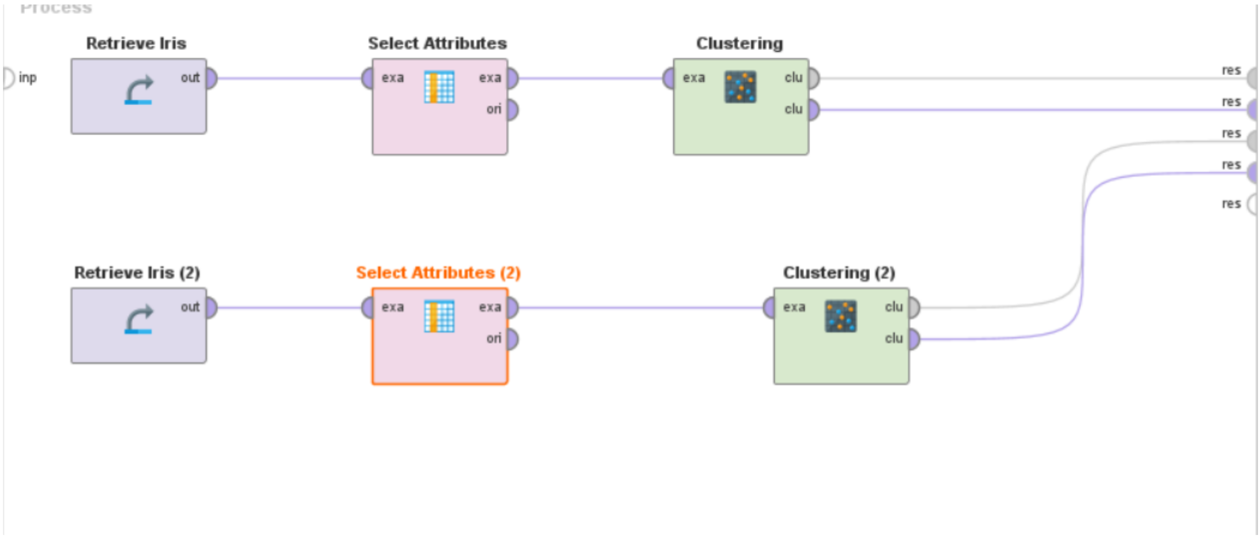
This is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each, where each class refers to a type of ~~its~~ iris plant.

Conclusion:-

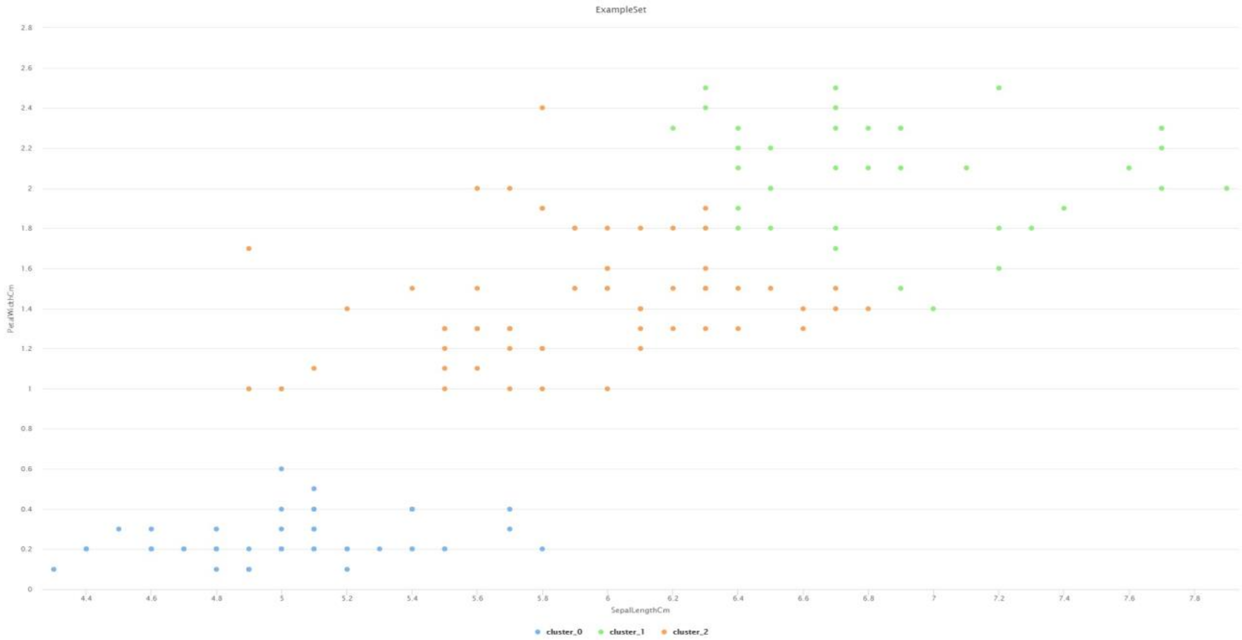
Visualized the effectiveness of the k- means clustering algorithm & hierarchical clustering using graphic capabilities in R.

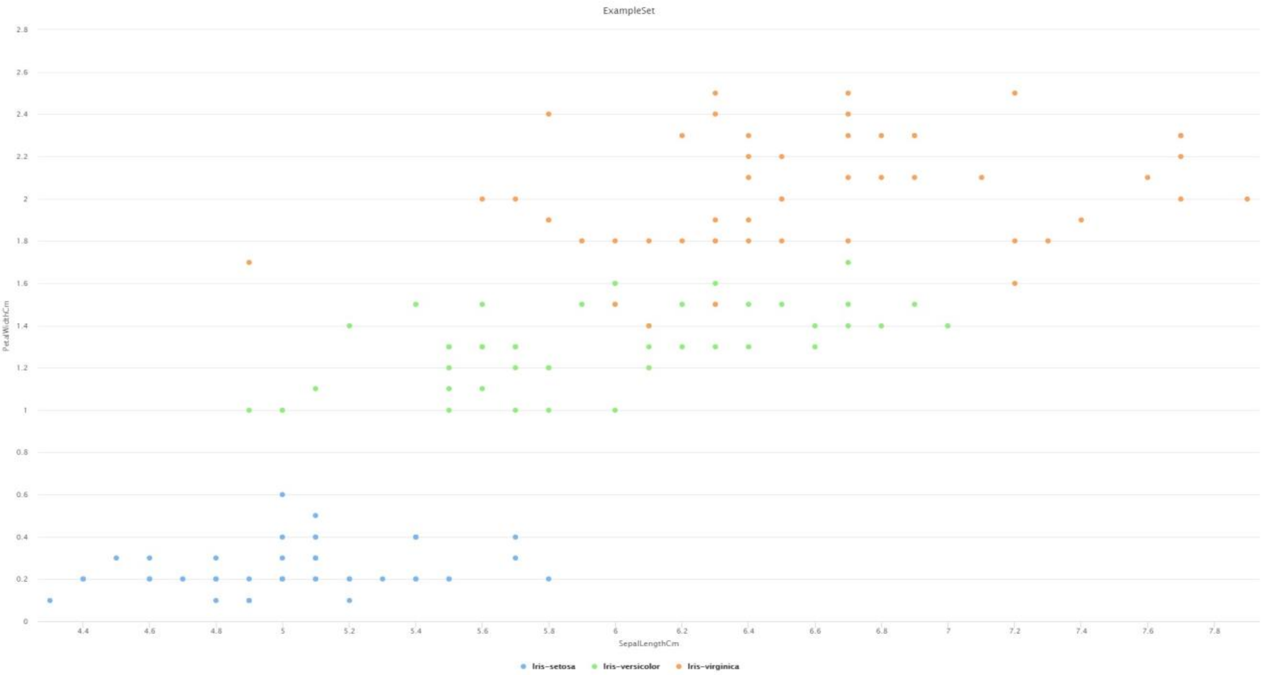
Output:

Clustering using K-means and DB Scan



Visualization:




ExampleSet
PetalWidthCm
2.8
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
4.4
4.6
4.8
5
5.2
5.4
5.6
5.8
6
6.2
6.4
6.6
6.8
7
7.2
7.4
7.6
7.8
SepalLengthCm
● Iris-setosa
● Iris-versicolor
● Iris-virginica