**CSE 572 DATA MINING**

**Instructor- Prof. Arunabha Sen**

**ASSIGNMENT 3**

Group name: Sriharsha

Group ID: 15

Group members:

1. Sriharsha Uppu (1215318000)
2. Praveen Muruganandam (1215159608)
3. Chandrika Cuddapah Sudhinder (1215321133)
4. Nishant Trivedi (1214072053)

Clustering is a type of **unsupervised learning**. This is very often used when you don’t have labeled data.

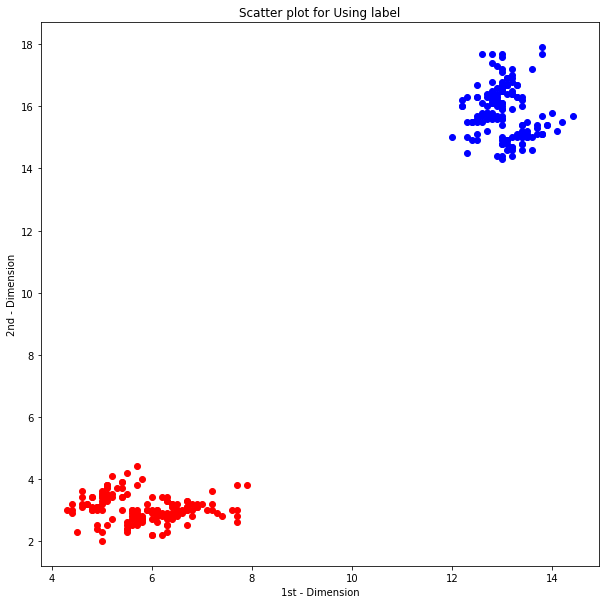
A)**K-Means Clustering** is one of the popular clustering algorithm. The goal of this algorithm is to find groups (clusters) in the given data.

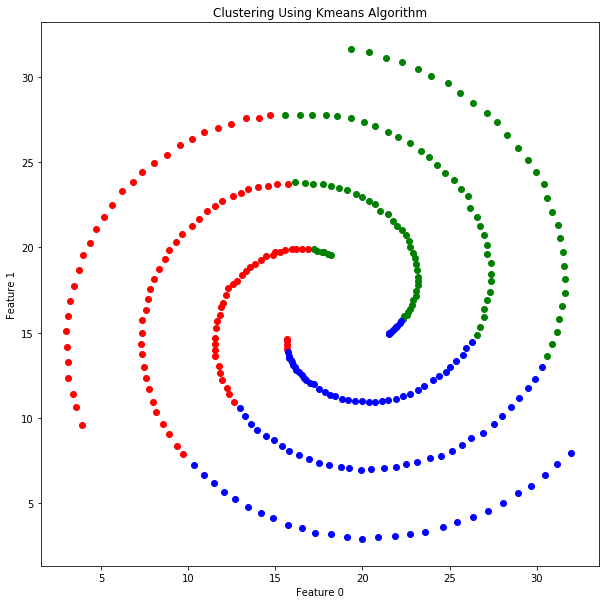
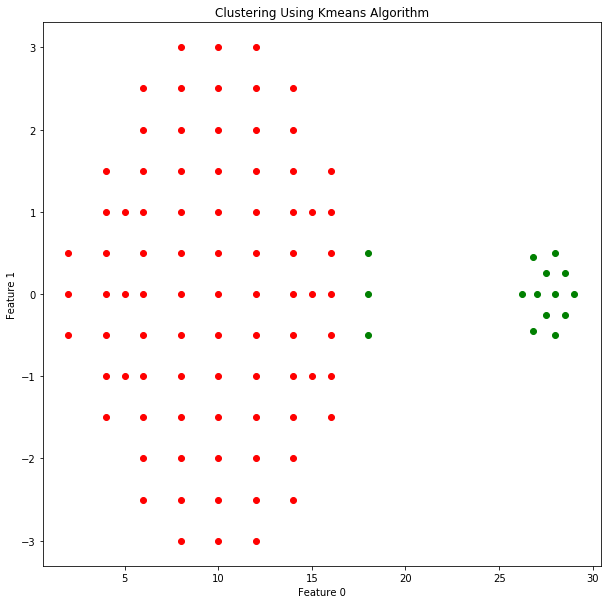
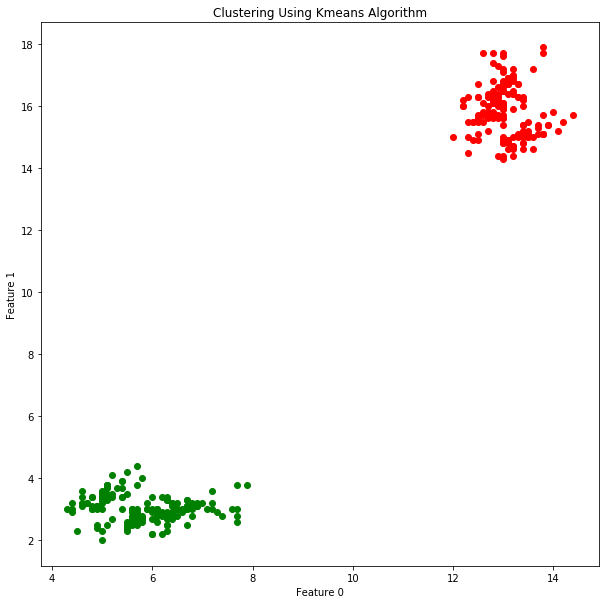
B) A clustering technique of distributing all the data-points in the dataset into k groups(clusters) such that **diameter** of the largest cluster is minimum among all possible ways of creating k clusters out of these data-points.

\*

C) **Spectral-Clustering** - Use a Gaussian kernel for computing affinity score between two points. Use RBF Kernel method for graph construction**(k=5)**. Subtasks include computing diagonal Degree matrix, Eigen-vectors & Eigen-values.

The stages of the algorithm are:  
1) constructing a nearest neighbours graph (KNN graph) or radius based graph.  
2) Embed the data points in low dimensional space (spectral embedding) in which the clusters are more obvious with the use of eigenvectors of the graph Laplacian.  
3) Use the lowest eigen value in order to choose the eigenvector for the cluster

**Results for Task 1**:

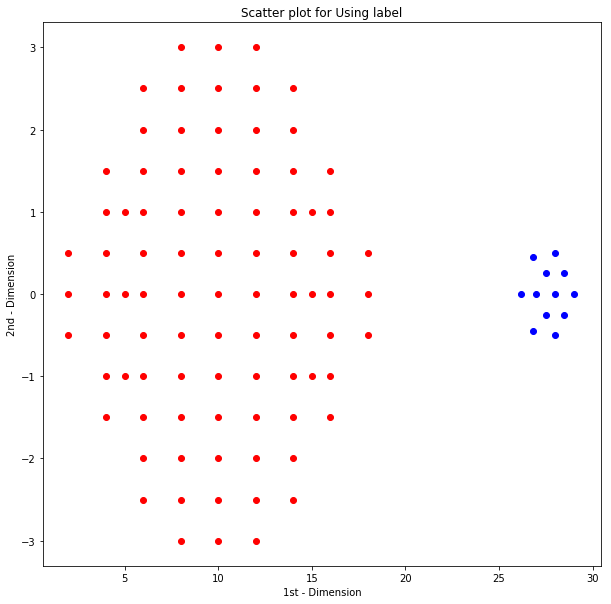
****

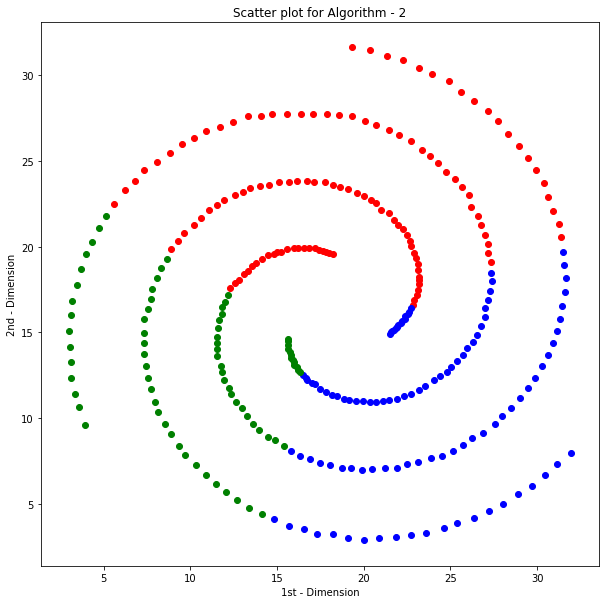
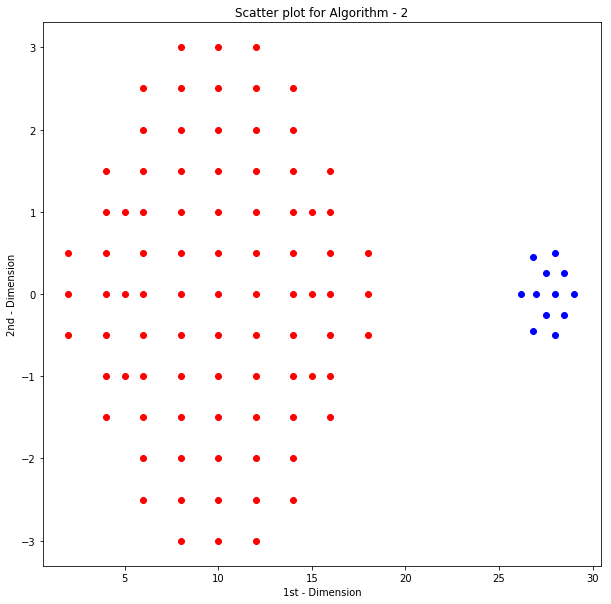
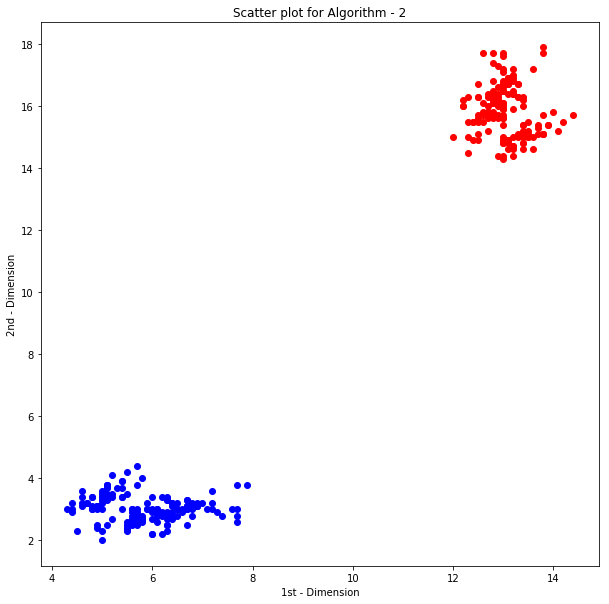
**Results for Task 2**:

For Dataset 2,

Algorithm C gave the worst performance. Spectral Clustering gave worst answers due to following reasons:

1. Cannot perform clustering on noisy or sparse data.
2. Prediction can make only one cluster model
3. The data in Dataset Is closely packed to each other with only slight difference.. hence the distance metric employed by the other 2 algorithms produces much better results than spectral which is used mainly when distance cannot be considered for clustering.





**Results for task 3**:

For Dataset 3,

Algorithm C gave the best performance. **Spectral Clustering** gave best answers due to following reasons:

The main problems with k-means are:

1) It makes assumption on the shape of the data (a round sphere, a radial basis).  
2) It requires multiple restarts at times to find the local minima (the best clustering).

**Spectral Clustering** algorithm helps to solve these two problems. This algorithm relies on the power of graphs and the proximity between the data points in order to cluster them, makes it possible to avoid the sphere shape cluster that the K means algorithm forces us to assume and it is also semi-convex.

K-means, as a data-clustering algorithm, is ideal for discovering globular clusters, where all members of each cluster are in close proximity to each other (in the Euclidean sense).

In contrast to data-clustering, we have graph-clustering techniques such as spectral clustering, where you don’t cluster data points directly in their native data space but instead form a similarity matrix where the (i,j)−(i,j)−th entry is some similarity distance you define between the ith and jth data points in your dataset. So, in a sense, spectral clustering is more general (and powerful) because whenever K-means is appropriate for use then so too is spectral clustering (just use a simple Euclidean distance as the similarity measure). The converse is not true though.

