**Abstract:**

*Clustering is an important data mining technique. Clustering is used to gather the items in groups according to the features of the item. This report is about IRIS dataset clustering using K-Mean Clustering and hierarchical clustering algorithm in R statistical tool.*

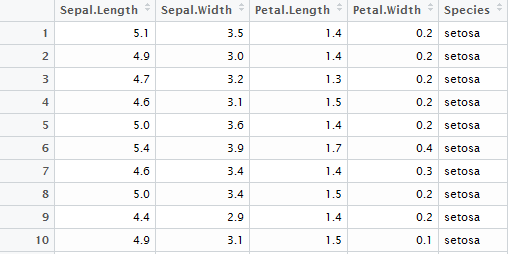
**Introduction:**

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory [data mining](https://en.wikipedia.org/wiki/Data_mining), and a common technique for [statistical](https://en.wikipedia.org/wiki/Statistics) [data analysis](https://en.wikipedia.org/wiki/Data_analysis), used in many fields, including [machine learning](https://en.wikipedia.org/wiki/Machine_learning), [pattern recognition](https://en.wikipedia.org/wiki/Pattern_recognition), [image analysis](https://en.wikipedia.org/wiki/Image_analysis), [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), [bioinformatics](https://en.wikipedia.org/wiki/Bioinformatics) and [data compression](https://en.wikipedia.org/wiki/Data_compression).

**K-Mean Clustering:**

  It is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bary centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move anymore.

**Training Data:** .



The dataset is taken of ‘IRIS’ which is already given in R tool and it can be accessed by just calling ‘iris’. It has four attributes and contains 3 type of classes. There are 150 instances. Three class are; *versicolor*, *setosa* and *virigenica*. Attributes containing sepal length, petal length, sepal width and petal width. It can judged from new instance from this data that new instance will be fall into which category.

**Solving in R:**

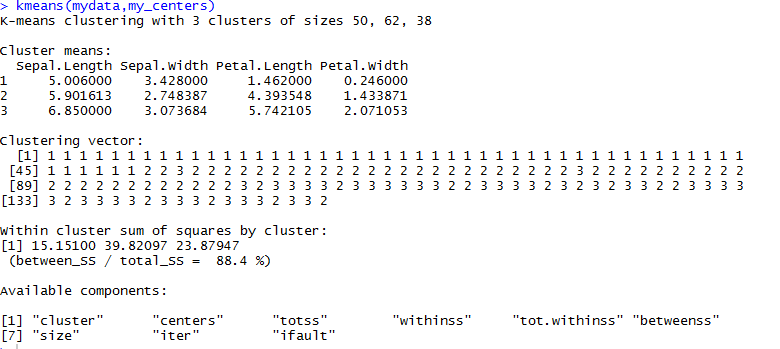
We have to load dataset into our variable, in which we can change something as we want. So we will separate class attribute because we don’t need it because we know that this data is already in three clusters. Then we call kmeans method which takes dataset and center instances in which it will make clusters.

mydata <- iris

mydata <- iris[,-5]

my\_centers <- iris[c(40,80,120),-5]

kmeans(mydata,my\_centers)



**Hierarchical Clustering:**

It is a method of [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) which seeks to build a [hierarchy](https://en.wikipedia.org/wiki/Hierarchy) of clusters. Strategies for hierarchical clustering generally fall into two types:

* **Agglomerative**: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
* **Divisive**: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In general, the merges and splits are determined in a [greedy](https://en.wikipedia.org/wiki/Greedy_algorithm) manner. The results of hierarchical clustering are usually presented in a [dendrogram](https://en.wikipedia.org/wiki/Dendrogram).

**Solving in R:**

We have taken same dataset IRIS. ‘dist’ function computes and returns the distance matrix computed by using the specified distance measure to compute the distances between the rows of a data matrix.

mydata <- iris

data <- dist(as.matrix(mydata))

hclus <- hclust(data)

plot(hclus)

The output shown in figure below

