K means Clustering

K means Clustering :

- K-Means Clustering is an Unsupervised Machine Learning algorithm, which groups the unlabeled dataset into different clusters.
- It is the process of teaching a computer to use unlabeled, unclassified data and enabling the algorithm to operate on that data without supervision.
- Without any previous data training, the machine's job in this case is to organize unsorted data according to parallels, patterns, and variations.

Goal:

- The goal of clustering is to divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups.

- It is essentially a grouping of things based on how similar and different they are to one another.

- To achieve this, we will use the K-means algorithm; an unsupervised learning algorithm.
- ‘K’ in the name of the algorithm represents the number of groups/clusters we want to classify our items into.
- The algorithm will categorize the items into k groups or clusters of similarity.
- To calculate that similarity, we will use the Euclidean distance as a measurement.

The algorithm works as follows:

- First, we randomly initialize k points, called means or cluster centroids.
- We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that cluster so far.
- We repeat the process for a given number of iterations and at the end, we have our clusters.

- The “points” mentioned above are called means because they are the mean values of the items categorized in them
- To initialize these means, we have a lot of options.
- An intuitive method is to initialize the means at random items in the data set.
- Another method is to initialize the means at random values between the boundaries of the data set (if for a feature *x,* the items have values in [0,3], we will initialize the means with values for *x* at [0,3]).

The above algorithm in pseudocode is as follows:

```
Initialize k means with random values

--> For a given number of iterations:

--> Iterate through items:

--> Find the mean closest to the item by calculating
the euclidean distance of the item with each of the means

--> Assign item to mean

--> Update mean by shifting it to the average of the items in
that cluster
```

Hierarchical Clustering

- Hierarchical clustering analysis is a method of clustering analysis that seeks to build a hierarchy of clusters
- i.e. tree-type structure based on the hierarchy.
- **Connectivity-based clustering:** This type of clustering algorithm builds the cluster based on the connectivity between the data points.
- Example: Hierarchical clustering

Hierarchical clustering

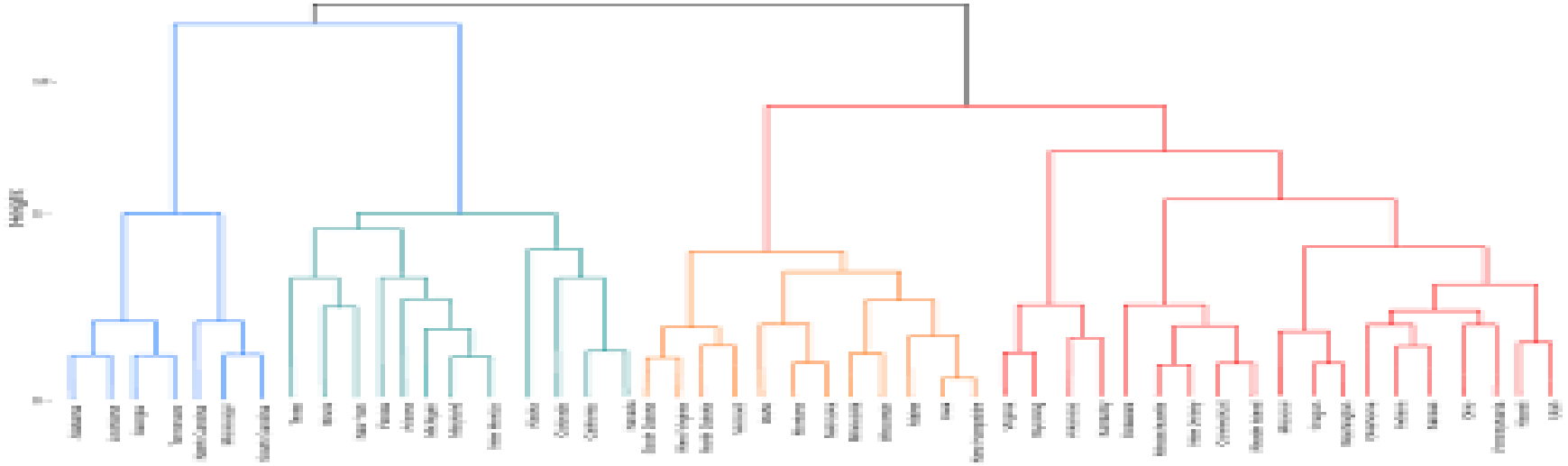
- Hierarchical clustering is a connectivity-based clustering model that groups the data points together that are close to each other based on the measure of similarity or distance.

- The assumption is that data points that are close to each other are more similar or related than data points that are farther apart.

- A dendrogram, a tree-like figure produced by hierarchical clustering, depicts the hierarchical relationships between groups.

Dendogram

- Individual data points are located at the bottom of the dendrogram, while the largest clusters, which include all the data points, are located at the top.
- In order to generate different numbers of clusters, the dendrogram can be sliced at various heights.



- The dendrogram is created by iteratively merging or splitting clusters based on a measure of similarity or distance between data points.
- Clusters are divided or merged repeatedly until all data points are contained within a single cluster, or until the predetermined number of clusters is attained.
- We can look at the dendrogram and measure the height at which the branches of the dendrogram form distinct clusters to calculate the ideal number of clusters.
- The dendrogram can be sliced at this height to determine the number of clusters.

Types of Hierarchical Clustering

Basically, there are two types of hierarchical Clustering:

- Agglomerative Clustering
- Divisive clustering

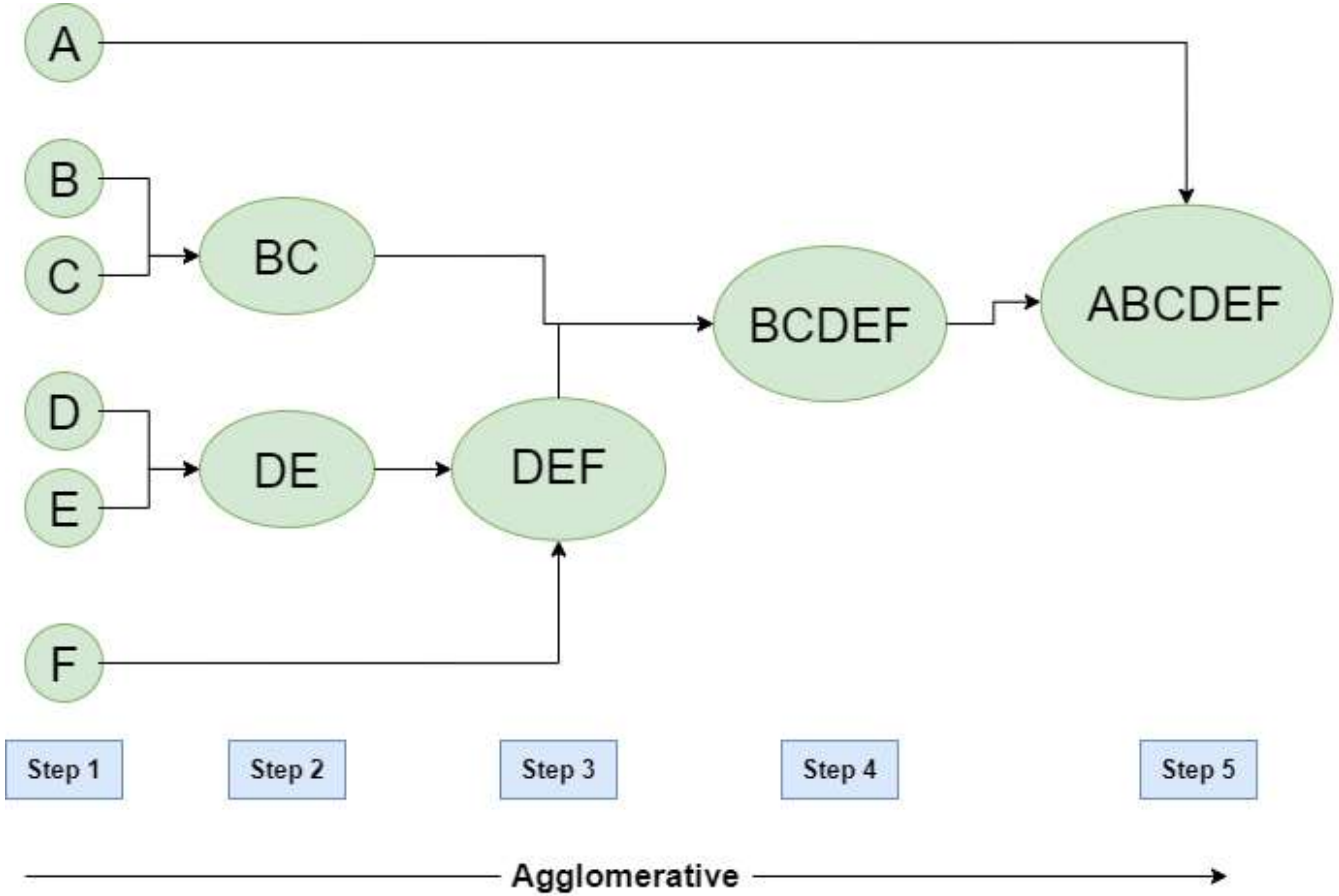
Hierarchical Agglomerative Clustering

- It is also known as the bottom-up approach or hierarchical agglomerative clustering (HAC).
- A structure that is more informative than the unstructured set of clusters returned by flat clustering.
- This clustering algorithm does not require us to prespecify the number of clusters.
- Bottom-up algorithms treat each data as a single cluster and then successively agglomerate(collect or form into a mass or group) pairs of clusters until all clusters have been merged into a single cluster that contains all data.

Algorithm :

given a dataset (d1, d2, d3, ...dN) of size N
compute the distance matrix
for i=1 to N:
  # as the distance matrix is symmetric about
  # the primary diagonal so we compute only lower
  # part of the primary diagonal
  for j=1 to i:
    dis_mat[i][j] = distance[di, dj]
each data point is a singleton cluster
repeat
  merge the two cluster having minimum distance
  update the distance matrix
until only a single cluster remains

Hierarchical Agglomerative Clustering



Hierarchical Divisive clustering

- It is also known as a top-down approach.
- This algorithm also does not require to prespecify the number of clusters.
- Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been split into singleton clusters.

Algorithm :

- given a dataset (d1, d2, d3, ...dN) of size N
- at the top we have all data in one cluster
- the cluster is split using a flat clustering method eg. K-Means etc
- repeat
- choose the best cluster among all the clusters to split
- split that cluster by the flat clustering algorithm
- until each data is in its own singleton cluster

Hierarchical Divisive clustering