UNIT-4

Unsupervised learning : Algorithms for clustering:

K-Means,

Unsupervised Bayesian learning,

Criterion functions for clustering;

Hierarchical,

partitional and

online clustering methods.

Unsupervised learning

- In the previous topic, we learned supervised machine learning in which models are trained using labeled data under the supervision of training data. But there may be many cases in which we do not have labeled data and need to find the hidden patterns from the given dataset. So, to solve such types of cases in machine learning, we need unsupervised learning techniques.

What is Unsupervised Learning?

- As the name suggests, unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as:

- *Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.*

- Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. **The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format**.

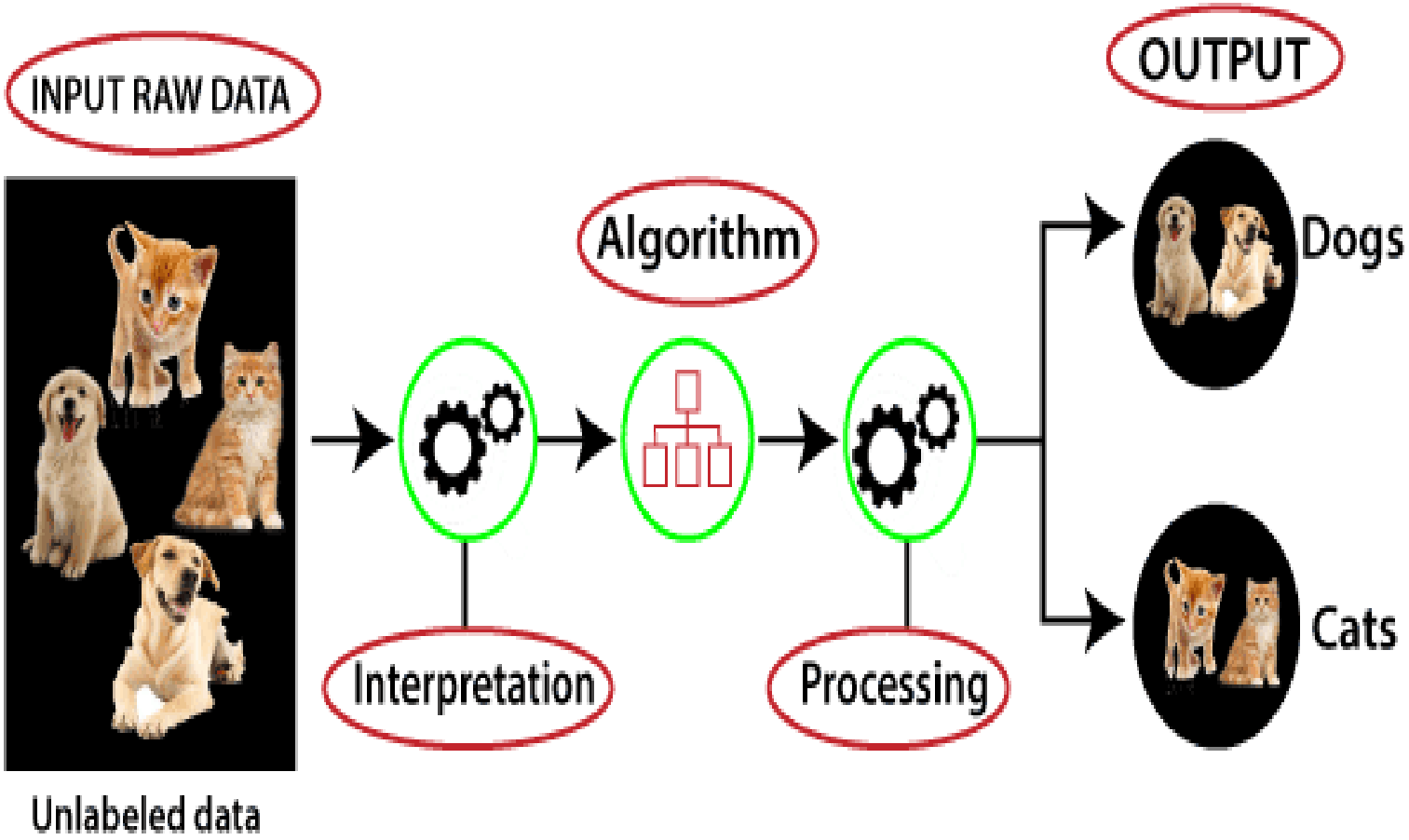
Example:

Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups according to similarities between images.

Why use Unsupervised Learning?

- Unsupervised learning is helpful for finding useful insights from the data.
- Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
- Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
- In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

Working of Unsupervised Learning



- Here, we have taken an unlabeled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.
- Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

Types of Unsupervised Learning Algorithm:

- **Clustering**: Clustering is the **task of dividing the data points into a number of groups** such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

2. Association

- An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset. Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical example of Association rule is Market Basket Analysis.

The various types of clustering are:

- Connectivity-based Clustering (Hierarchical clustering)
- Centroids-based Clustering (Partitioning methods)
- Distribution-based Clustering
- Density-based Clustering (Model-based methods)
- Fuzzy Clustering
- Constraint-based (Supervised Clustering)

K-Means Clustering Algorithm
K-Means

- K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

What is K-Means Algorithm?

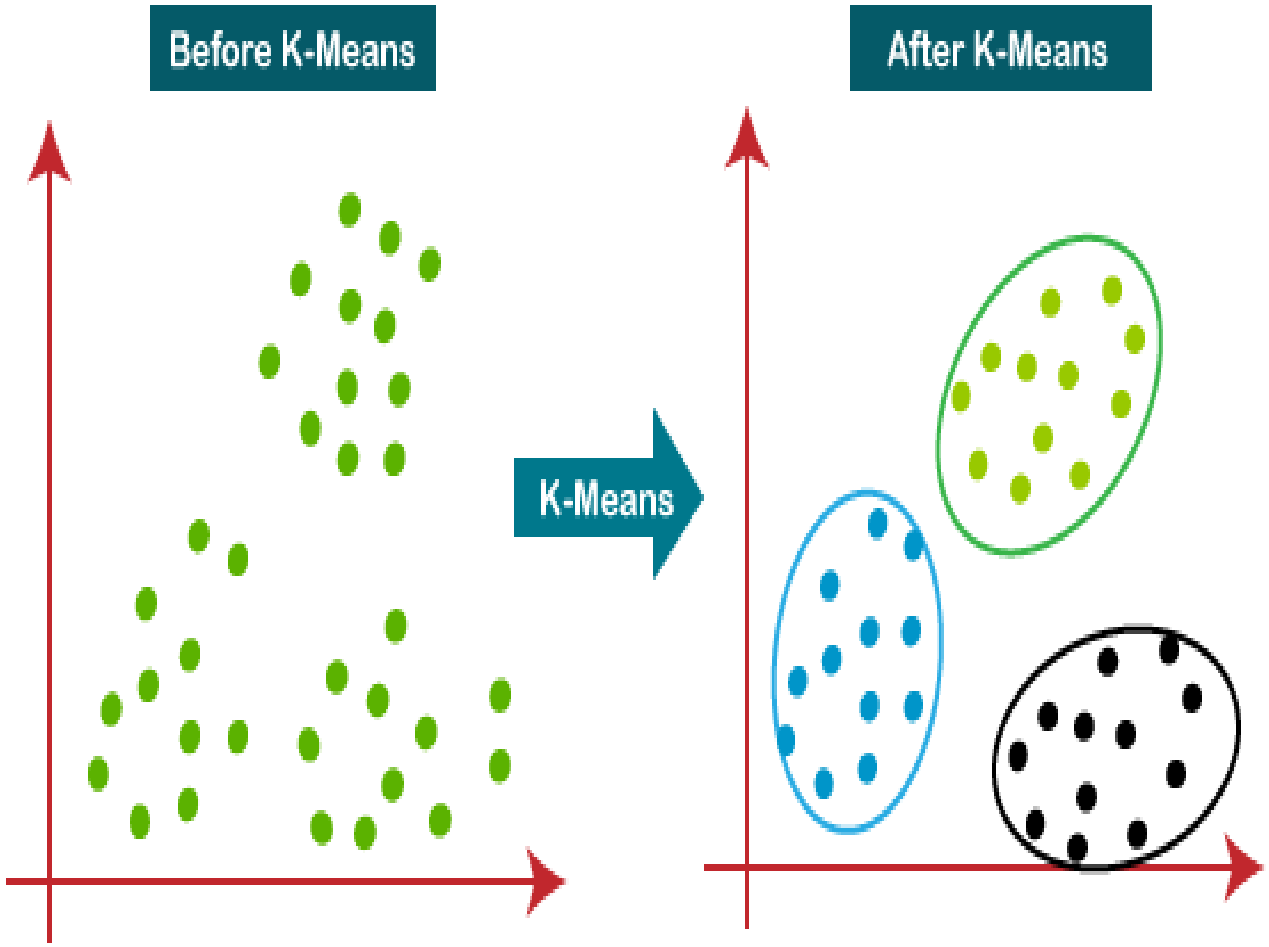
- K-Means Clustering is an Unsupervised Learning algorithm
- which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

- **It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.**

- The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.
- **The k-means clustering algorithm mainly performs two tasks:**
- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



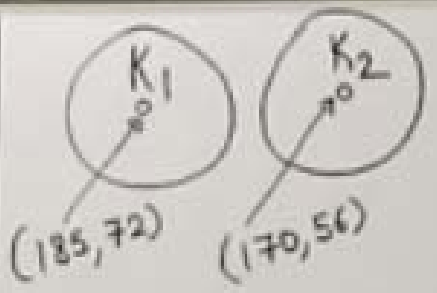
How does the K-Means Algorithm

- **Step-1:** Select the number K to decide the number of clusters.
- **Step-2:** Select random K points or centroids. (It can be other from the input dataset).
- **Step-3:** Assign each data point to their closest centroid, which will form the predefined K clusters.
- **Step-4:** Calculate the variance and place a new centroid of each cluster.

- **Step-5:** Repeat the third steps, which means reassign each datapoint to the new closest centroid of each cluster.
- **Step-6:** If any reassignment occurs, then go to step-4 else go to FINISH.
- **Step-7**: The model is ready.

K-means Algorithm

	Height	Weight
①	185	72
②	170	56
③	168	60
④	179	68
⑤	182	72
⑥	188	77
⑦	180	71
⑧	180	70
⑨	183	84
⑩	180	88
⑪	180	67
⑫	177	76



ED for ⑤ → K_1 → $\sqrt{(162-185)^2+(60-72)^2}$

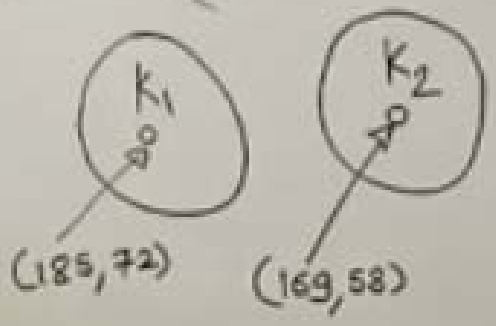
20.80

→ K_2 → $\sqrt{(162-170)^2+(60-56)^2}$

= 4.48

New Centroid Calculation :-

for K_2 = $\left(\frac{170+168}{2}, \frac{60+56}{2}\right)$ = (169, 58)



ED for ④ → K_1 = $\sqrt{(179-185)^2+(68-72)^2}$

= (6.32)

→ K_2 = $\sqrt{(179-169)^2+(68-58)^2}$

= 14.14

Euclidean Distance

$$\sqrt{(X_0-X_c)^2+(Y_0-Y_c)^2}$$

$K_1 \rightarrow \{1,4,5,6,7,8,9,10,11,12\}$

$K_2 \rightarrow \{2,3\}$

What is meant by hierarchical clustering?

- *Hierarchical clustering*, also known as *hierarchical cluster* analysis, is an algorithm that groups similar objects into groups called clusters.

- Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, **all files and folders on the hard disk are organized in a hierarchy**. There are two types of hierarchical clustering, **Divisive and Agglomerative.**

- A **Hierarchical clustering** method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data points as a separate cluster. Then, it repeatedly executes the subsequent steps:

1. Identify the 2 clusters which can be closest together, and
2. Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

The basic method to generate hierarchical clustering are:

- **1. Agglomerative:**
- **2. Divisive:**

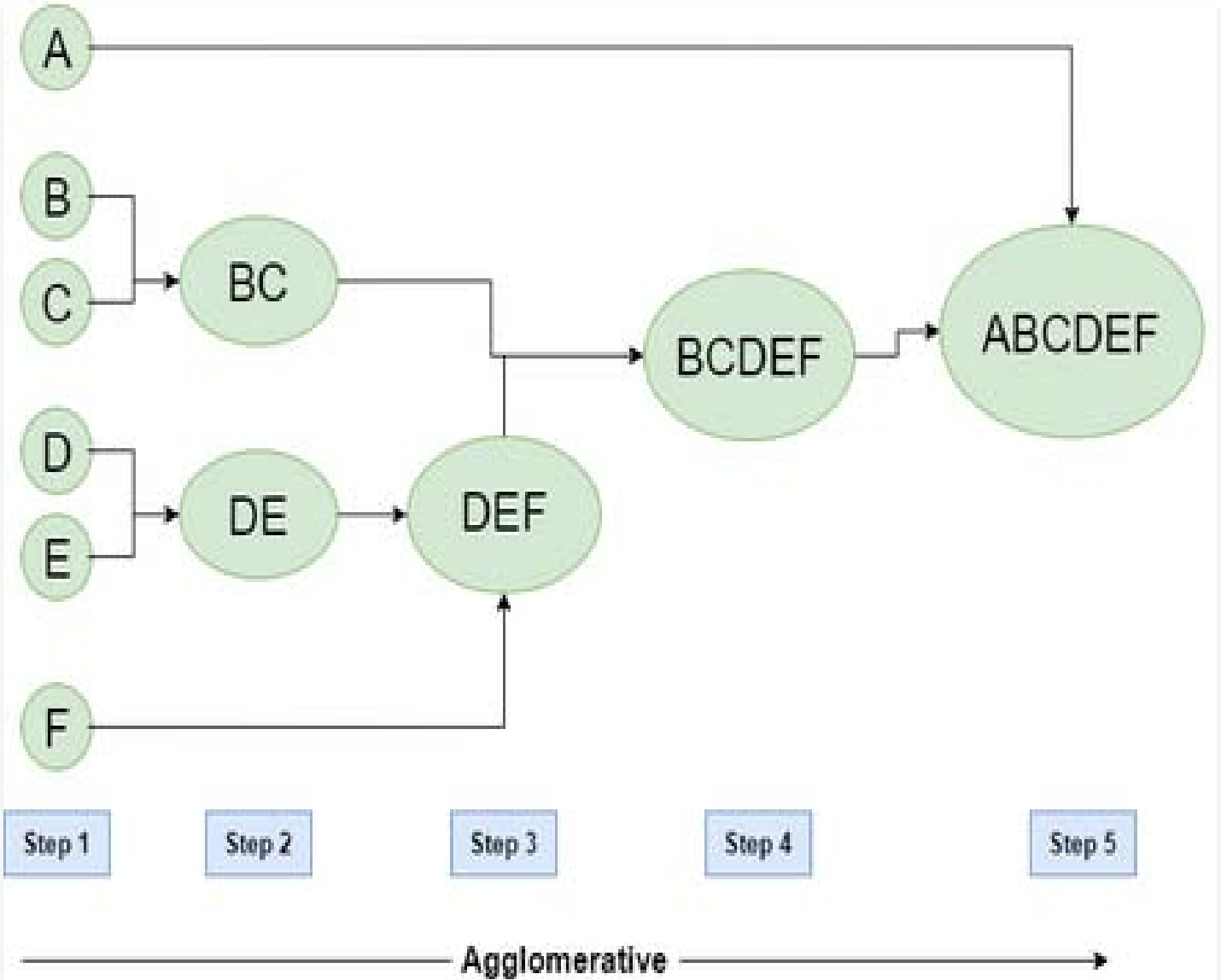
1. Agglomerative:

- Initially consider every data point as an **individual** Cluster and at every step, **merge** the nearest pairs of the cluster. (It is a bottom-up method). At first everydata set set is considered as individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed.

Algorithm for Agglomerative Hierarchical Clustering is:

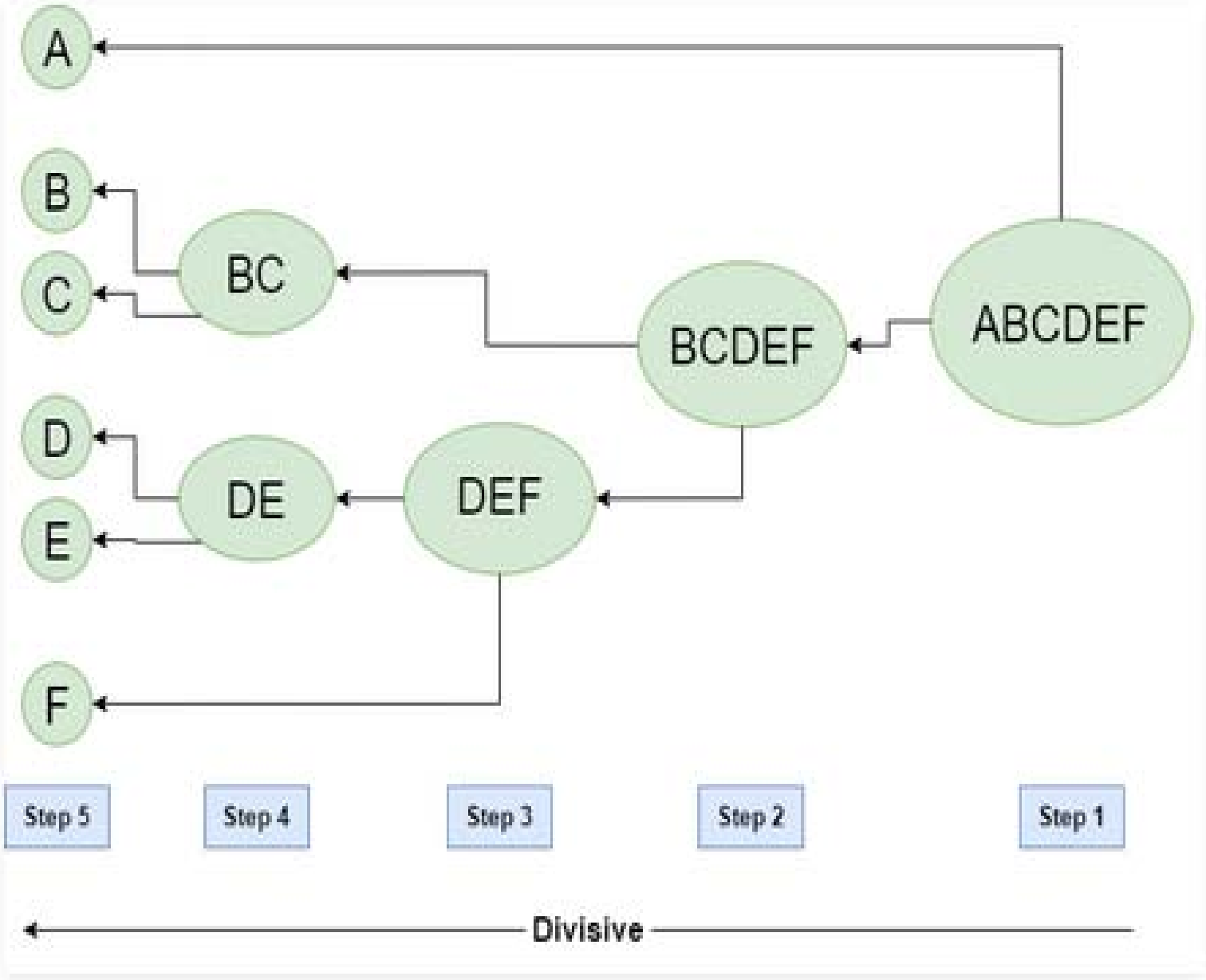
- Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
- Consider every data point as a individual cluster
- Merge the clusters which are highly similar or close to each other.
- Recalculate the proximity matrix for each cluster
- Repeat Step 3 and 4 until only a single cluster remains.

Let’s say we have six data points **A, B, C, D, E, F.**



2. Divisive:

- We can say that the Divisive Hierarchical clustering is precisely the **opposite** of the Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren't comparable. In the end, we are left with N clusters.


A
B
C
BC
BCDEF
ABCDEF
D
E
DE
DEF
F
Step 5
Step 4
Step 3
Step 2
Step 1
Divisive

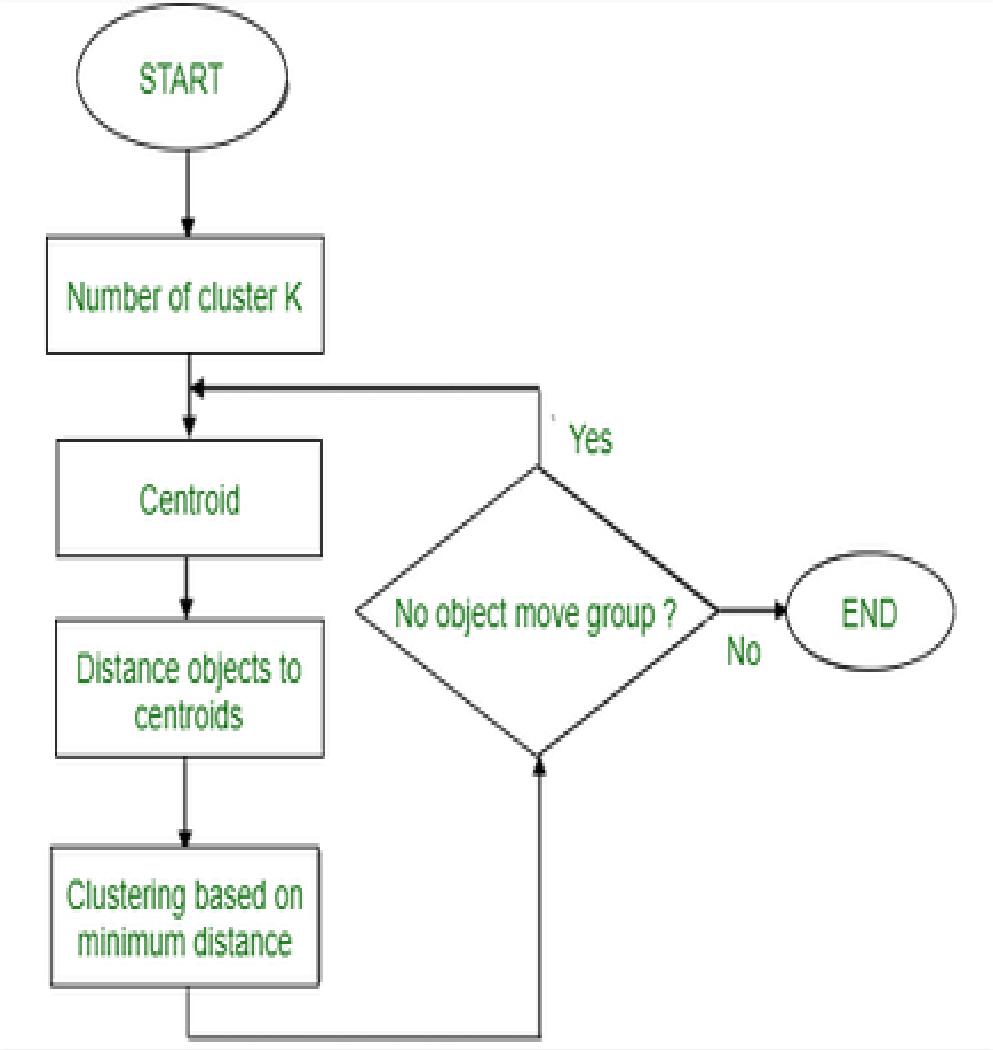
2. partitional clustering?

- This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. Its the data analysts to specify the number of clusters that has to be generated for the clustering methods.

- There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Mediods), CLARA algorithm (Clustering Large Applications) etc.

K-Mean (A centroid based Technique):

Flowchart:

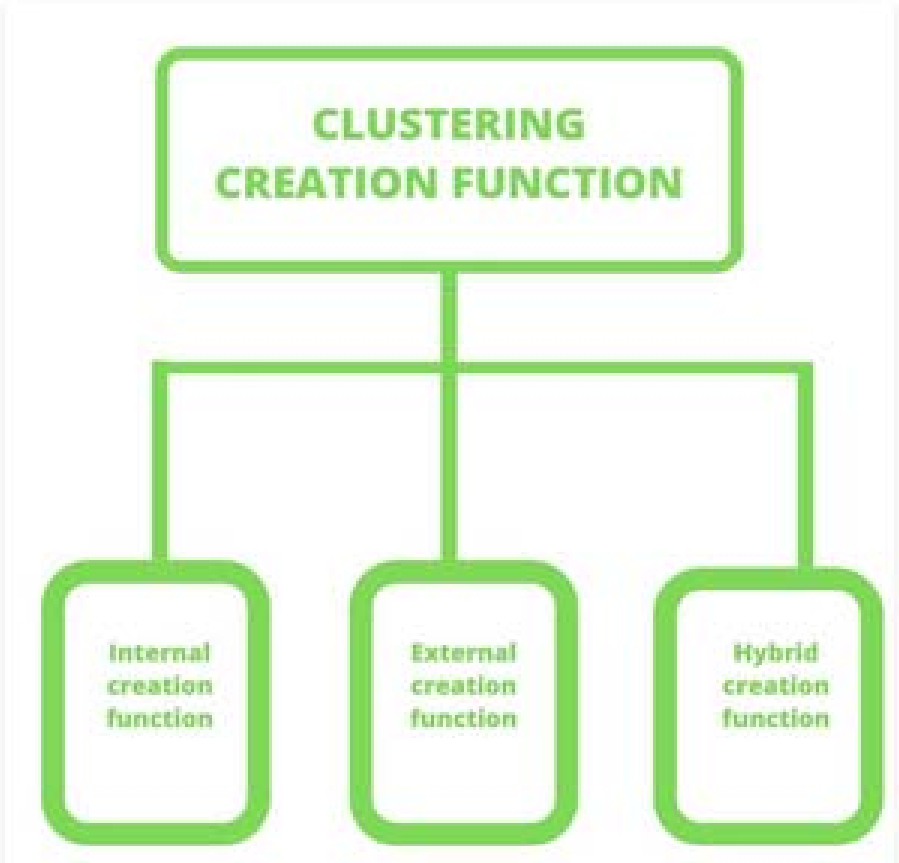


Criterion Function For Clustering

- Cluster examination isolates information into bunches (clusters) that are important, valuable, or both. In case significant bunches are the objective, at that point, the clusters ought to capture the common structure of the information. In a few cases, be that as it may, cluster investigation is as it were a valuable beginning point for other purposes, such as information summarization.

- A great clustering strategy will create tall quality clusters in which the quality of a clustering result too depends on both the similitude degree utilized by the strategy and its usage. The quality of a clustering strategy is additionally measured by its capacity to find a few or all of the covered up designs.

Criterion Function For Clustering –



- **Internal Criterion Function –** This class of grouping is an intra-clusterview. Internal basis work upgrades a capacity and measures the nature of bunching capacity different groups which are unique in relation to each other.

- **External Criterion Function –** This lesson of clustering measure is an inter-class view. External Basis Work optimizes a work and measures the quality of clustering capacity of different clusters which are diverse from each other.

- **Hybrid Criterion Function –** This work is utilized because it has the capacity to at the same time optimize numerous person Model Capacities not at all like as Inside Basis Work and Outside Basis Work.