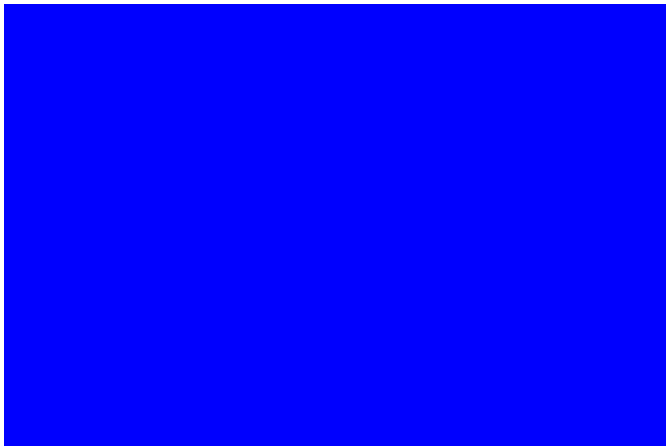
Machine Learning

Dimensionality Reduction

The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction

Higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information



Algorithm:

Step 1: Get the data from $m \times n$ matrix A

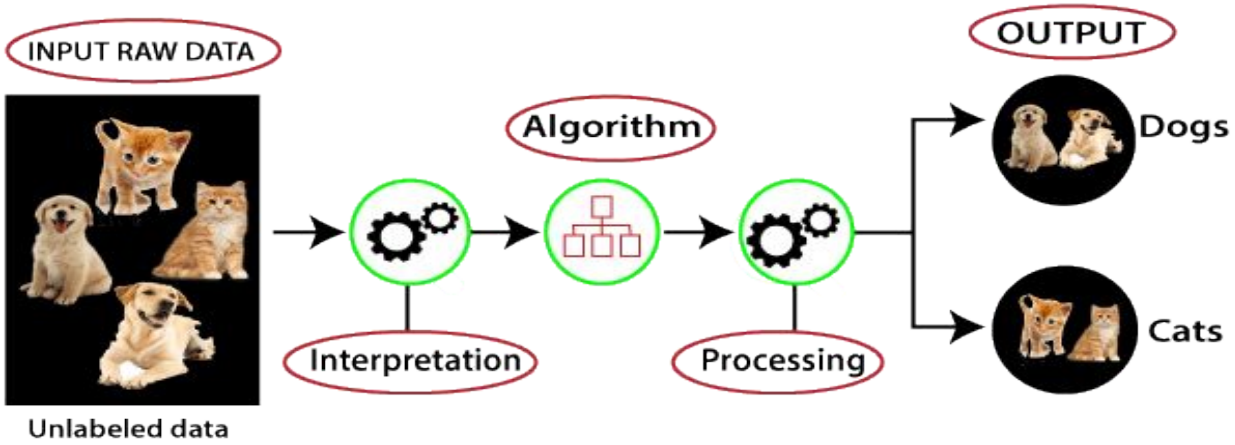
Step 2: Calculate the covariance matrix

Step 3: Calculate the eigenvectors and eigenvalues of the covariance matrix

Step 4: Choosing principal components and forming a feature vector

Step 5: Deriving the new data set and forming the clusters

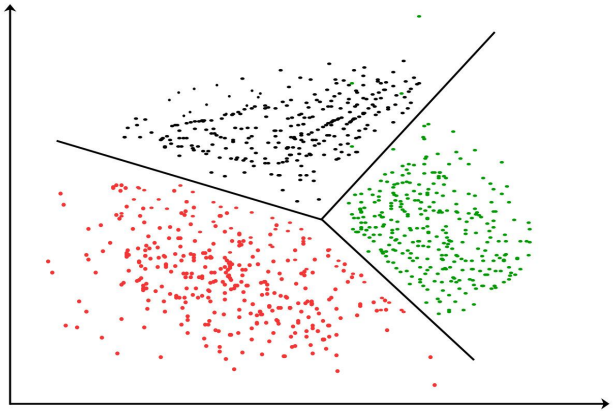
Unsupervised Learning -



- **Clustering**
- **Association**

Clustering -

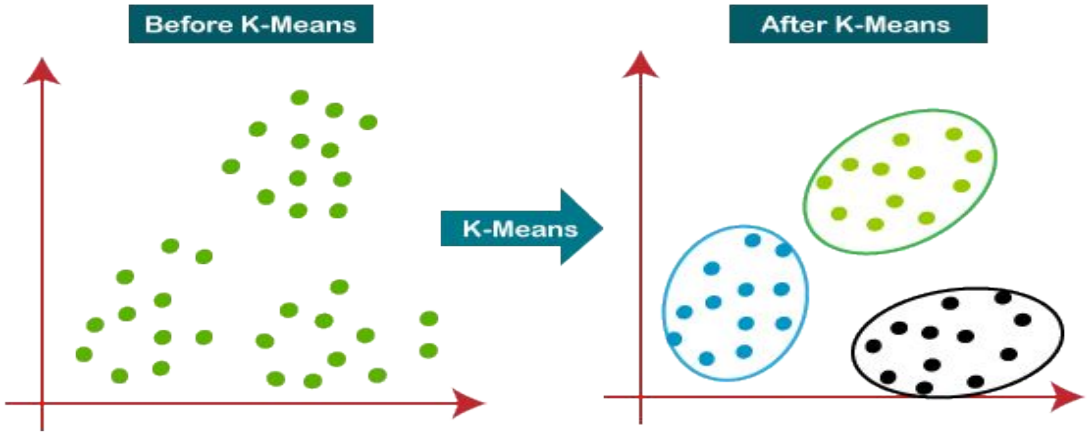
A method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group



- **K-means Clustering**
- **Hierarchical Clustering**

K-means Clustering

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.



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).

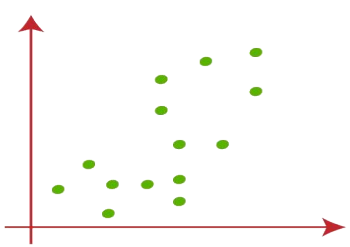
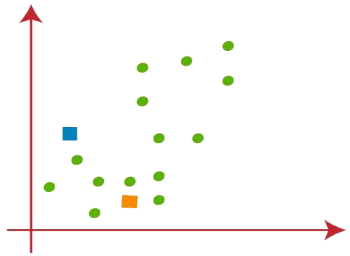
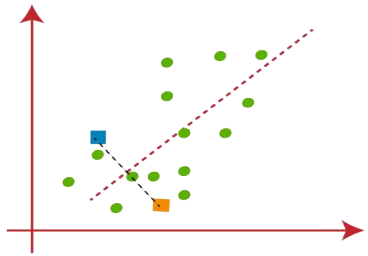
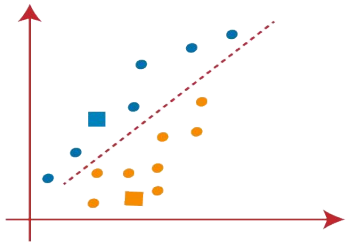
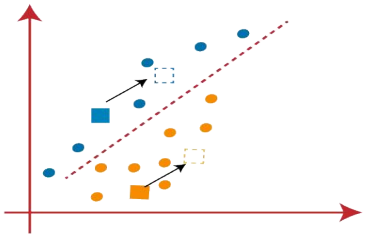
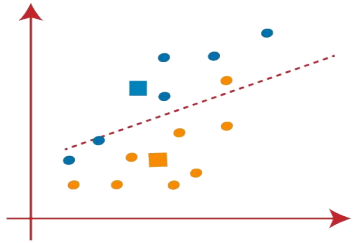
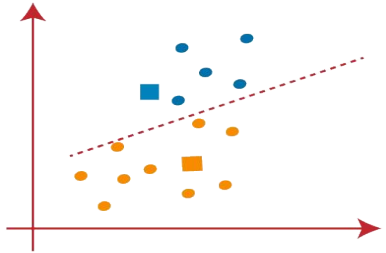
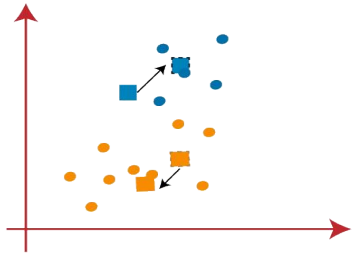
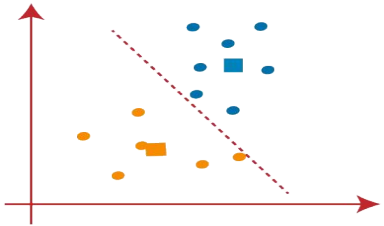
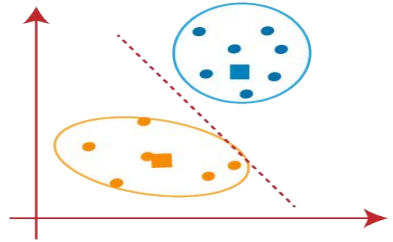
Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third steps, which means re-assign 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 = 2

Elbow Method

Ways to find the optimal number of clusters

WCSS stands for **Within Cluster Sum of Squares -** Total variations within a cluster

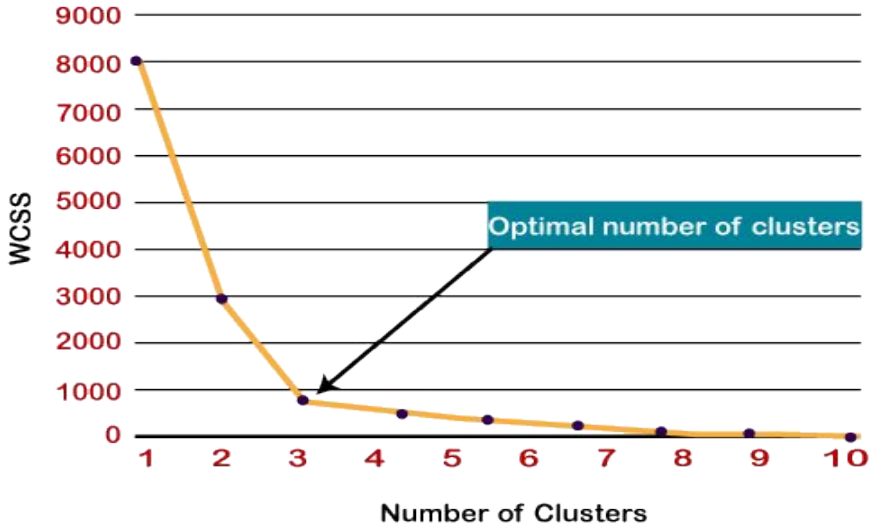
$$WCSS= \sum_{Pi \text{ in } Cluster1} distance(P_i\ C_1)^2 + \sum_{Pi \text{ in } Cluster2} distance(P_i\ C_2)^2$$

$\sum_{Pi \text{ in } Cluster1} distance(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1.

Distance is calculated using Euclidean distance or Manhattan distance

To find the optimal value of clusters, the elbow method follows the below steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.



Hierarchical Clustering

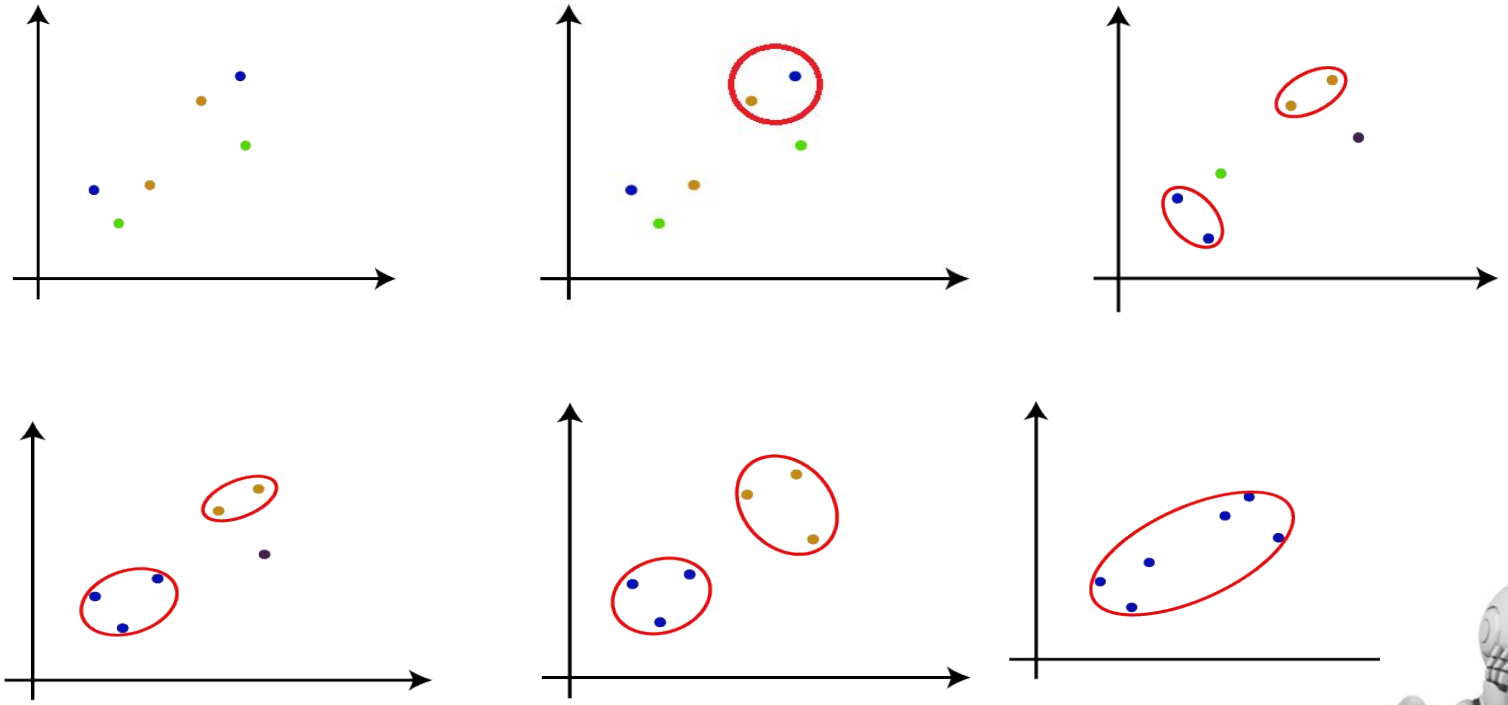
Used to group the unlabeled datasets into a cluster and also known as **hierarchical cluster analysis**

Hierarchy of clusters in the form of a tree called **Dendrogram**

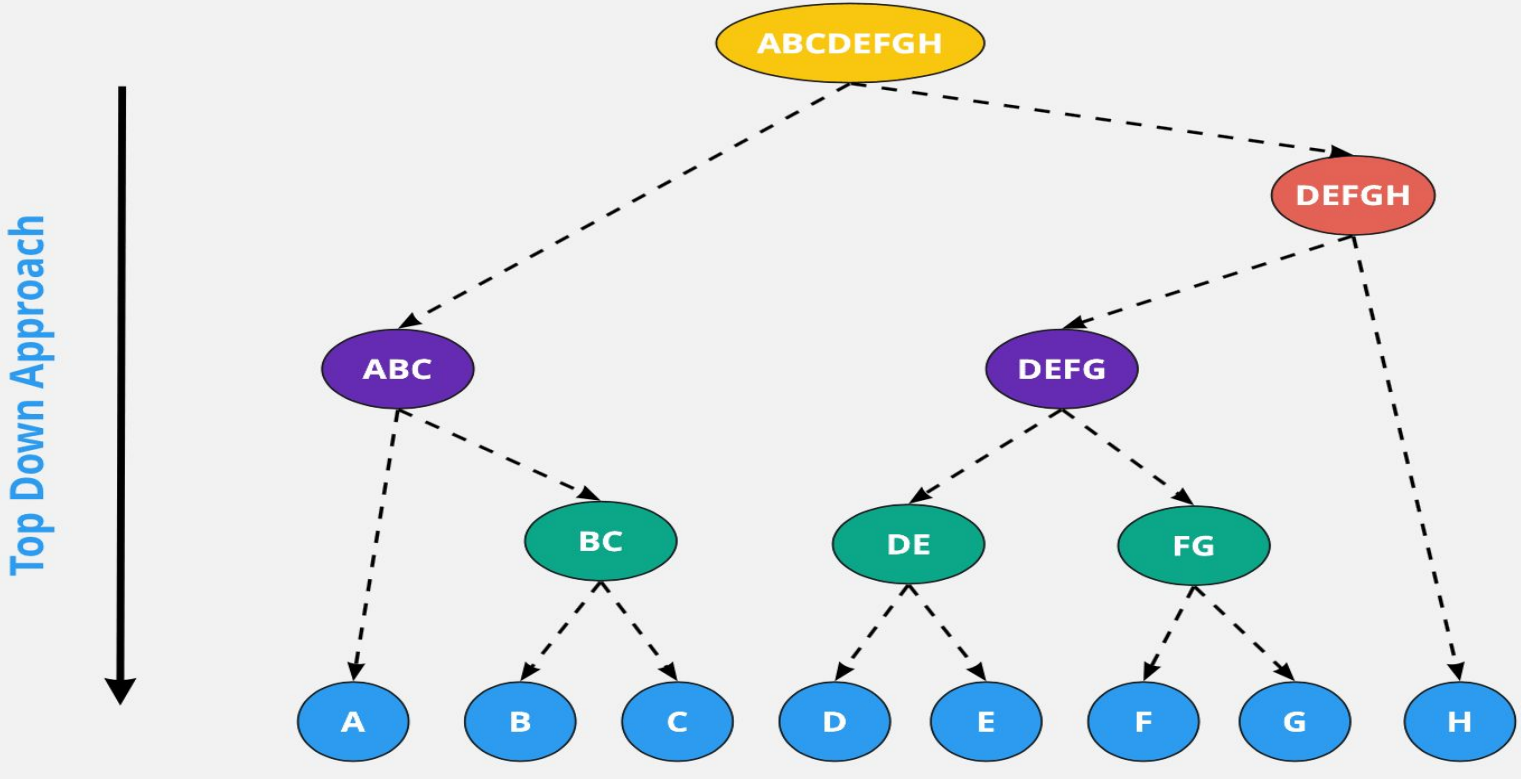
Two approach -

1. **Agglomerative: Bottom-up approach**
2. **Divisive: Top-down approach.**

Agglomerative Hierarchical clustering



Hierarchical Divisive Clustering



dataaspirant.com