## Clustering Algorithms:-

↳ Algo is based on kind of data using.

↳ Such as, some algo used to find num. of clusters & some required to find min. dist. bet$^n$ observations.

### 1) K-means Algorithm:

↳ One of the most popular & widely used.

↳ Divide samples into different clusters of equal variances.

↳ Aim: Divide dataset into K clusters, pre-defined.

↳ Algo initialize K centroids randomly & assign data points to nearest centroid.

↳ Update centroid as mean of all data points in cluster.

↳ Efficient & work well for roughly spherical & equal sized clusters.

### 2) Hierarchical Clustering:

↳ Create tree-like structure (dendrogram) of cluster

↳ Start consider each point separate cluster & iteratively merge closest clusters until remain only 1 / pre-defined num. of clusters.

↳ Agglomerative (bottom-up)

↳ Divisive (top-down)

↳ Useful for hierarchical data / represent data at different levels.

## 3) DBSCAN :

↳ DBSCAN – Density-Based Spatial Clustering of Applications with Noise.

↳ Based on concept of density.

↳ Define clusters as high density data points separated by lower density.

↳ Identify: - core points (dense areas),
- border points (near core points, but less dense areas)
- Noise points.

↳ Effective discover arbitrary shape clusters & handling noise.

## 4) Agglomerative Clustering :

↳ Start with individual data point as clusters & merge iteratively based on linkage criterion.

↳ Common linkage methods include:
- single linkage (min distance)
- Complete linkage (max distance)
- Average linkage (average distance)

↳ Create hierarchical representation of clusters similar to hierarchical clustering.

↳ Intuitive & work well when num. of clusters not predetermined.

## 5) Gaussian Mixture Model (GMM):

↳ Represent clusters as probability distribution.

↳ Assume that, points generated from mixture of Gaussian distributions.

↳ Algo estimate params of these Gaussian distributions (mean, covariances & weights) using Expectation - Maximization (EM) Algo.

↳ Effective for modelling data with complex distributions & identify clusters of varying shapes & orientations.

## ⊙ Applications of clustering :-

① Identify cancer cells - Divide cancerous & non-cancerous cells.

② In search Engines - Grouping similar data into clusters, far from other dissimilar objects.

③ clustering segmentation - Market research to segment customers based on choices & preferences.

④ Biology - Classify species of plants & animals.

⑤ Libraries - Classify books on topics & informat.

## Advantages of clustering :-

① Unsupervised Learning – Unlabelled data. Suitable obtaining labeled data expensive / impractical.

② Pattern Discovery – Hidden patterns, structures or natural grouping within data.

③ Data Reduction – Reduce dimensionality, easy visualize & understand. Help complex datasets.

④ Anomaly Detection – Outliers / data points other than normal clusters, do not fit into any cluster cla anomalies.

⑤ Customer Segmentation – Help business understand customer clusters, create strategies & improve customer satisfaction.

⑥ Recommendation System – Group users / items based on preferences, easy recommend products, services / content to users.

⑦ Flexibility – Different algorithms suit different data & clustering scenarios. Flexibility allow choose suitable algo for specific problem.

## Limitations of clustering :-

① choosing num of clusters (K) – Inappropriate value lead to poor results.

② Sensitivity to Initialization – Some algo, like K-means sensitive to initialization of centroids.

③ Scalability – Some algo like hierarchical, computationally expensive & may not scale large datasets.

④ Cluster Shape – Algo like K-means assume spherical clusters & equally sized, may not suitable non-spherical & unevely sized clusters.

⑤ Outliers – Sensitive to outliers.

⑥ Curse of dimensionality – High-dimensional data sparsely spaced, difficult define distances betn data points.