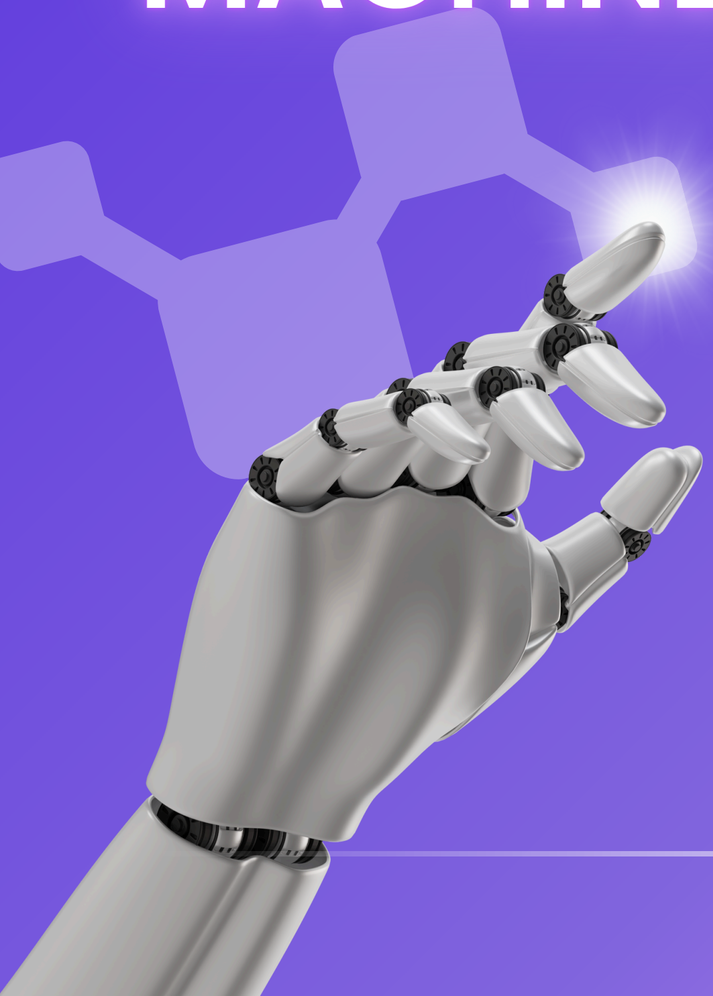
MACHINE LEARNING



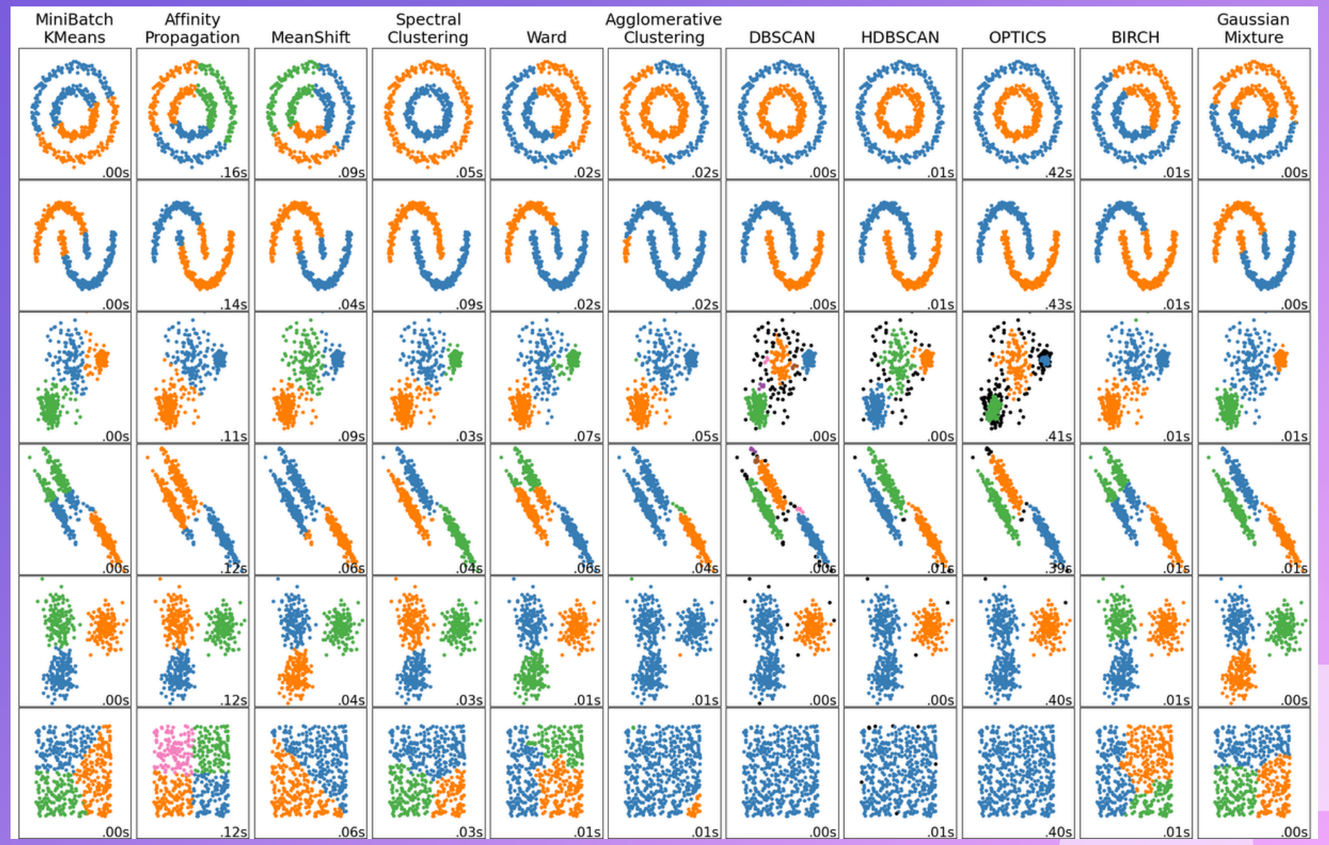
Clustering Algorithms

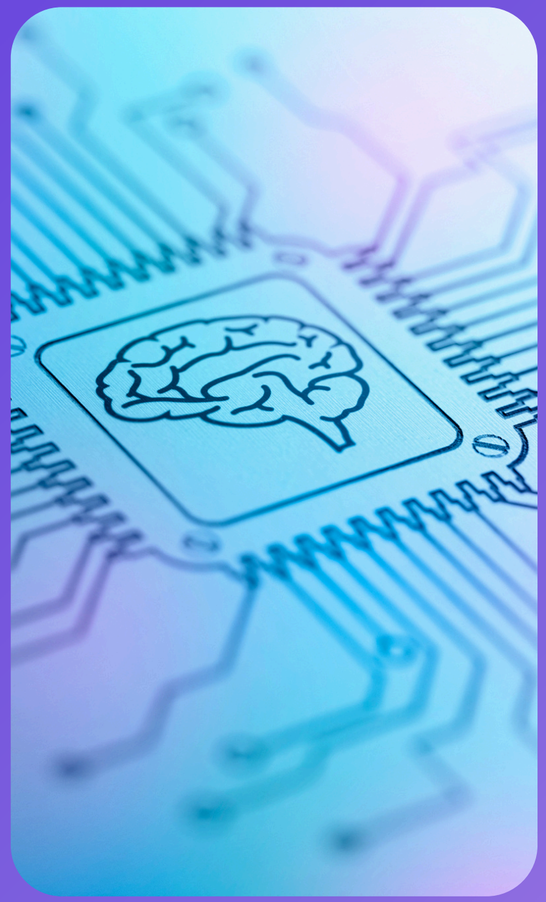
Devi Priya v

Introduction to Clustering Algorithm's

- Clustering is an unsupervised learning technique used to group similar data points.

- Helps in data segmentation, pattern recognition, and exploratory data analysis.





Types of Clustering

- *Hard Clustering: Each data point belongs to one cluster (e.g., K-Means).*
- *Soft Clustering: Data points can belong to multiple clusters (e.g., Fuzzy C-Means).*
- *Hierarchical Clustering: Organizes data in a tree structure.*

K-Means Clustering

- *Centroid-based algorithm that partitions data into K clusters.*

Steps:

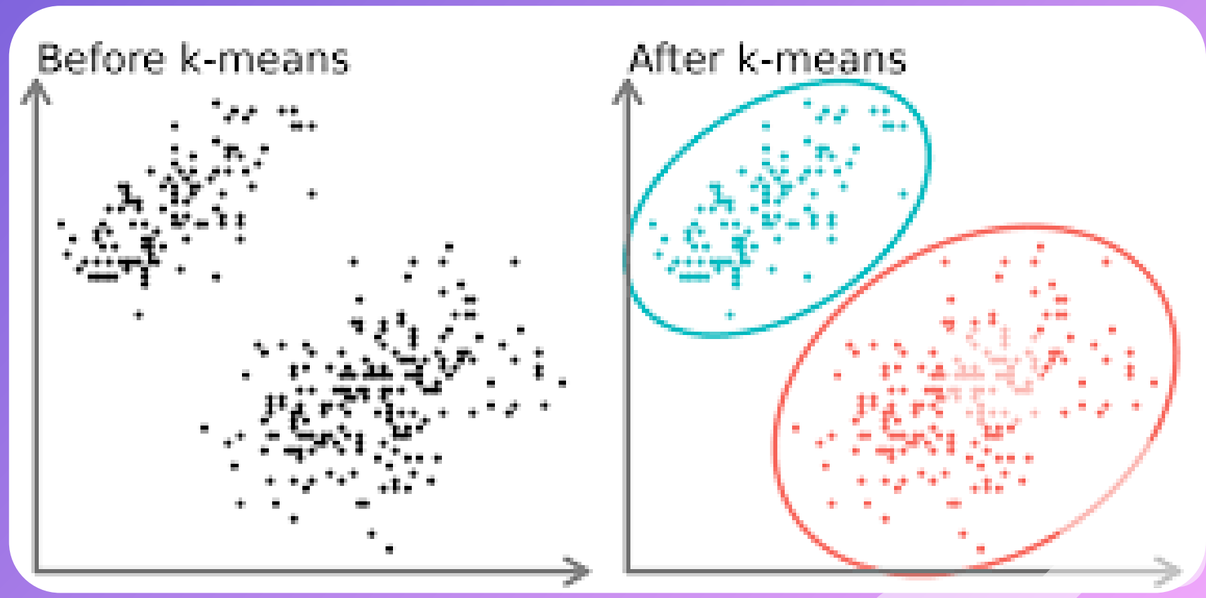
1. Choose K cluster centers.
2. Assign each data point to the nearest cluster.
3. Update centroids until convergence.

Advantages:

Simple and scalable.

Disadvantages:

- Sensitive to initial centroids and outliers



Hierarchical Clustering

• *Builds a tree-like hierarchy of clusters.*

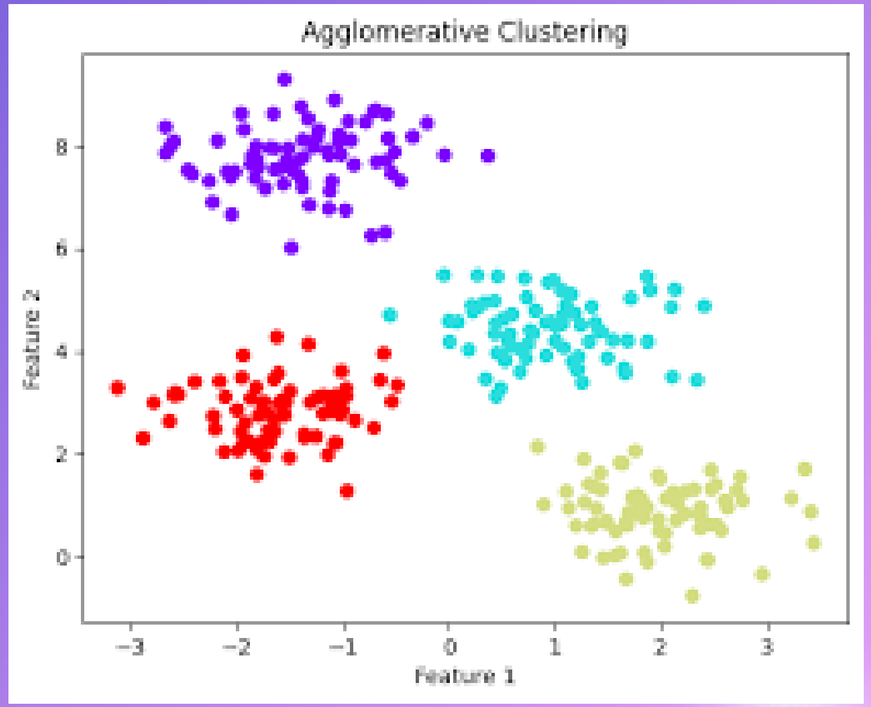
• **Types:**

- Agglomerative (bottom-up approach).

- Divisive (top-down approach).

• **Advantages: No need to specify K.**

• **Disadvantages: Computationally expensive.**

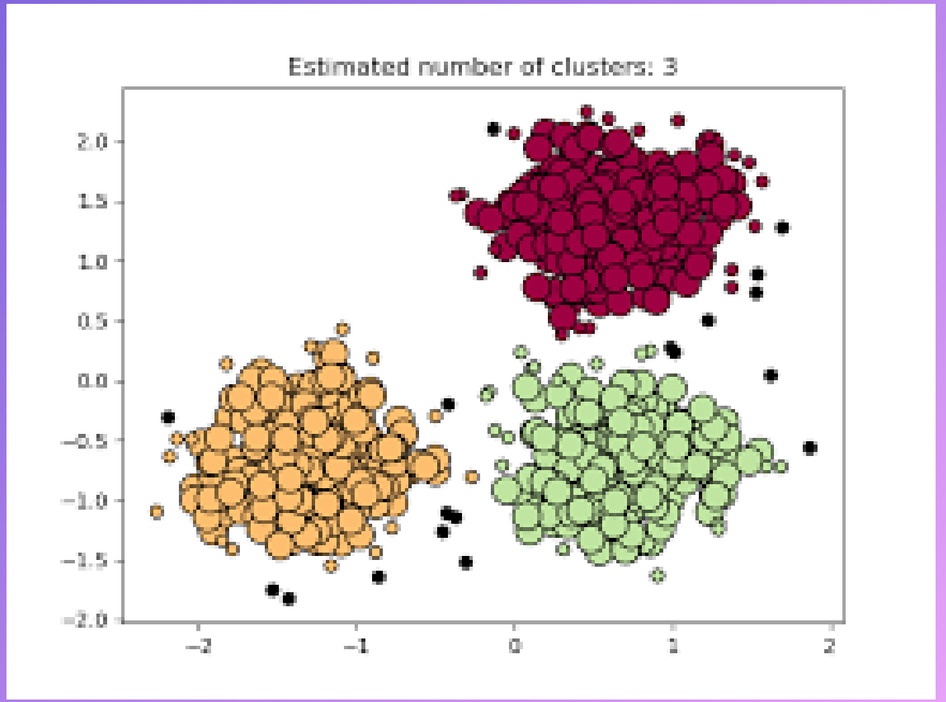


DBSCAN (Density-Based Clustering)

- Groups dense regions and identifies noise points.

Parameters

- Epsilon (ε) : Defines neighborhood radius.
- MinPts : Minimum points in a dense region.



Advantages

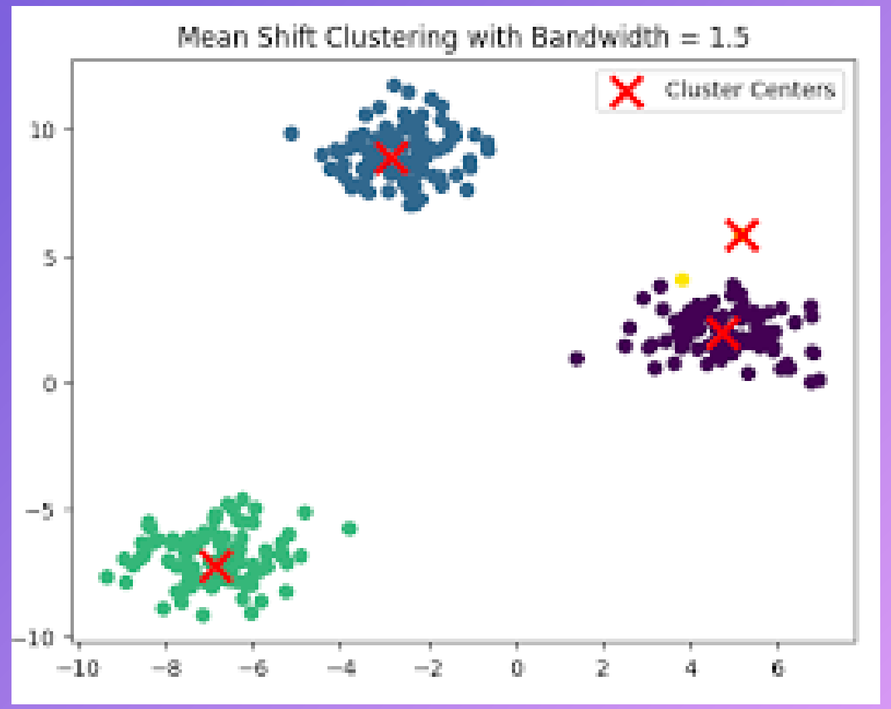
- Detects arbitrary shapes.

Disadvantages

- Sensitive to parameters.

Mean Shift Clustering

- Centroid-based approach that shifts points towards higher density regions.
- Works well without specifying K.
- Computationally expensive.

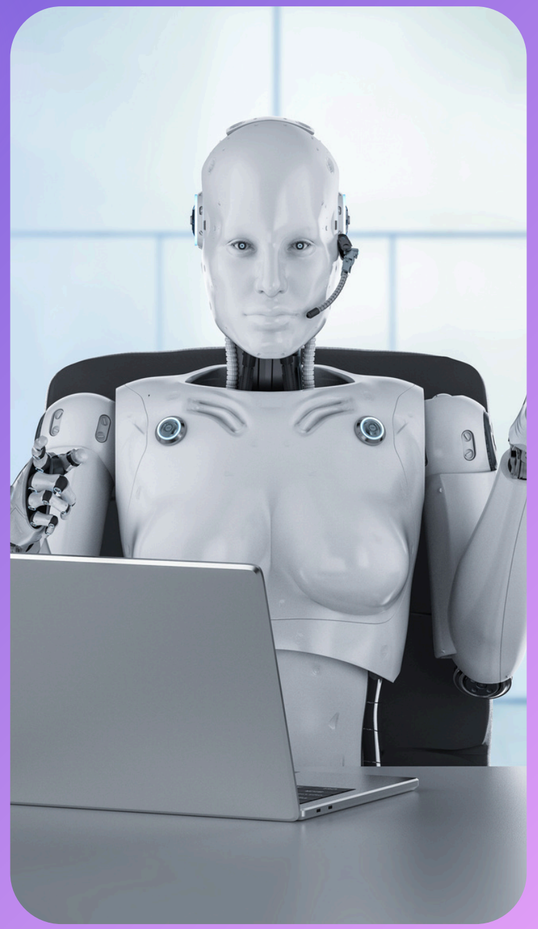


Evaluating Clustering Performance

- ***Metrics:***
  - Silhouette Score.
  - Davies-Bouldin Index.
  - Dunn Index.
  - Inertia (for K-Means).

Applications of Clustering

- Customer segmentation in marketing.
- Anomaly detection in cybersecurity.
- Image segmentation in computer vision.
- Recommender systems.



CONCLUSION

- Clustering is a fundamental unsupervised learning technique.
- Various algorithms exist with different strengths and weaknesses.
- Choosing the right method depends on the dataset and use case.