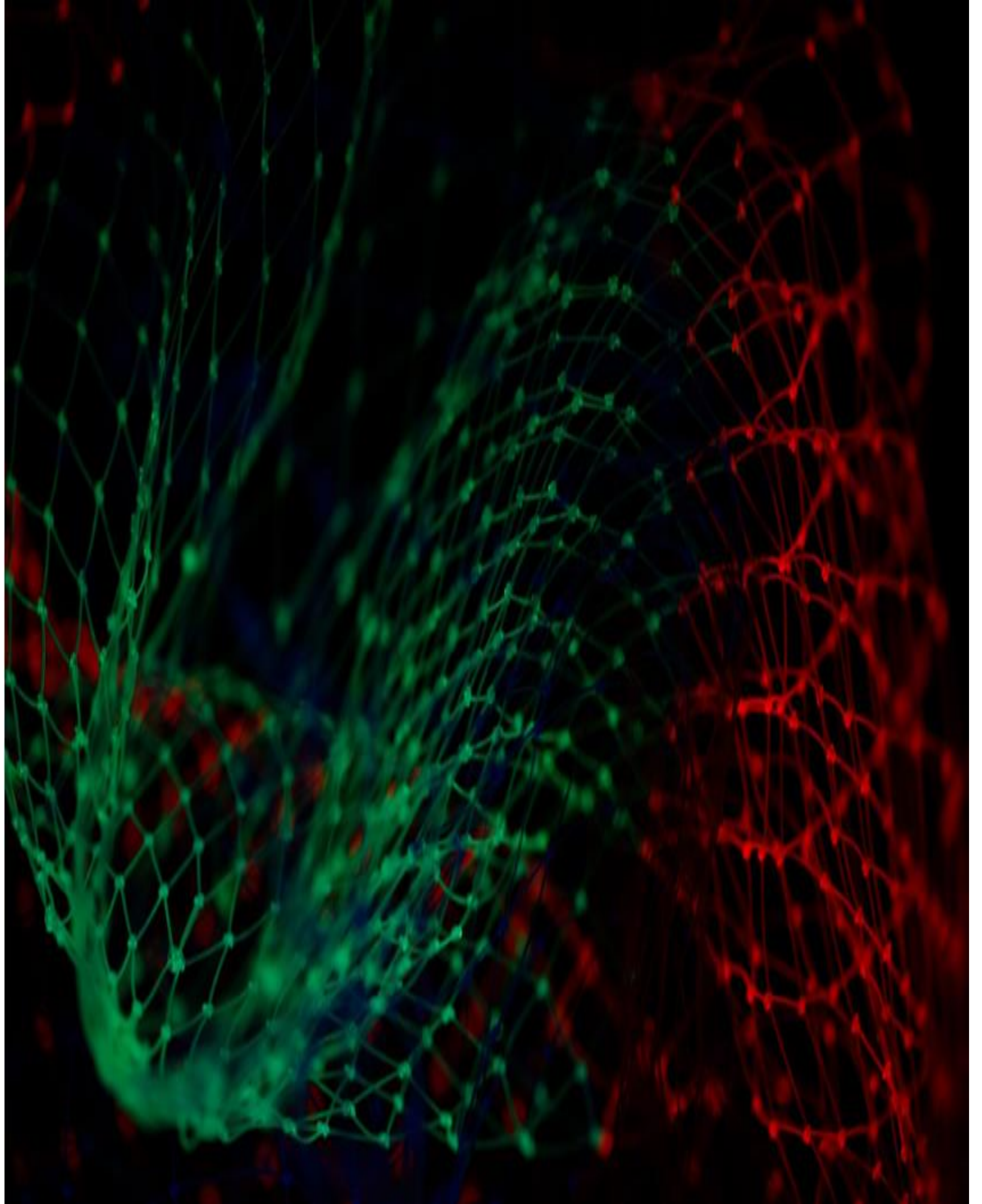
Machine Learning

CLUSTERING ALGORITHMS

KUPPUSAMY S



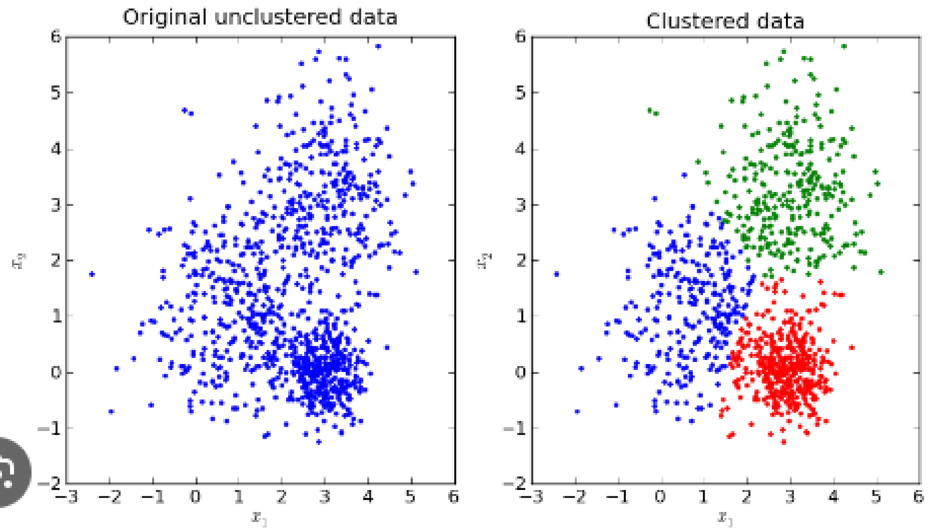
1.K-means Clustering

Definition:

- K-means clustering is a partitioning method that divides a dataset into K distinct, non-overlapping subsets (clusters).

- Clustering Approach- Partitional (flat)

- The algorithm aims to minimize the variance within each cluster.



Working Principle

- Select K initial cluster centroids randomly.
- Assign each data point to the nearest centroid, forming K clusters.
- Recalculate the centroids of each cluster based on the assigned points.
- Repeat the assignment and centroid update steps until convergence (no change in centroids or minimal change)

Advantages and Disadvantages

Advantages:

- Simple and easy to implement.
- Efficient for large datasets.
- Works well with spherical-shaped clusters.

Disadvantages:

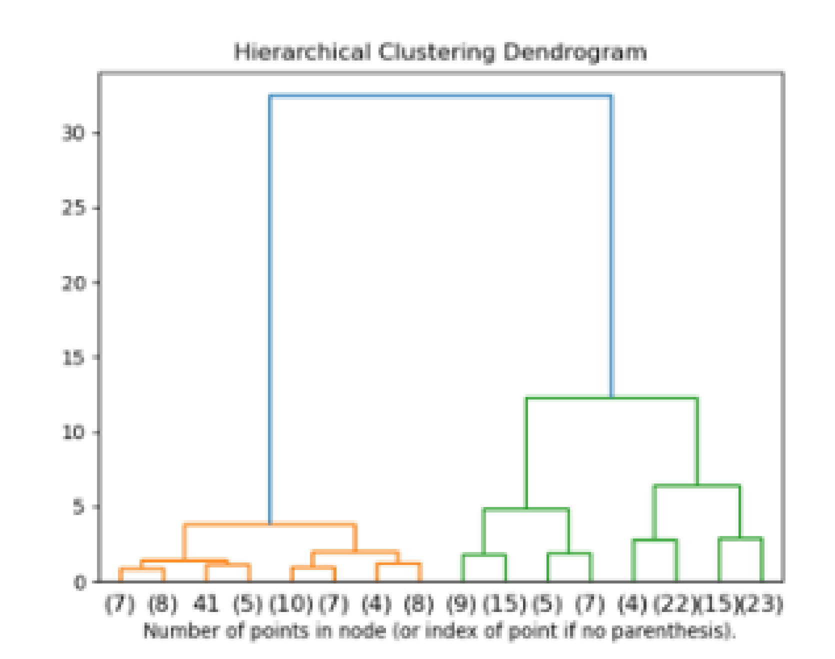
- Requires specifying the number of clusters (K) in advance.
- Sensitive to initial centroid placement.
- Not suitable for clusters with varying sizes or densities.

2.Agglomerative Clustering

Definition:

• Agglomerative clustering is a type of hierarchical clustering that builds nested clusters by merging pairs of data points successively.

• The process continues until all points belong to a single cluster or a stopping criterion is met.



Working Principle

• Start with each data point as its own cluster.

• Merge the closest pair of clusters based on a distance metric (e.g., Euclidean distance).

• Update the distance matrix and repeat until a single cluster remains or a predefined number of clusters is reached.

Advantages and Disadvantages

Advantages:

•Does not require specifying the number of clusters in advance.

•Captures a hierarchy of clusters, useful for understanding data structure.

Disadvantages:

•Computationally expensive for large datasets.

•Sensitive to the choice of distance metric and linkage criterion.

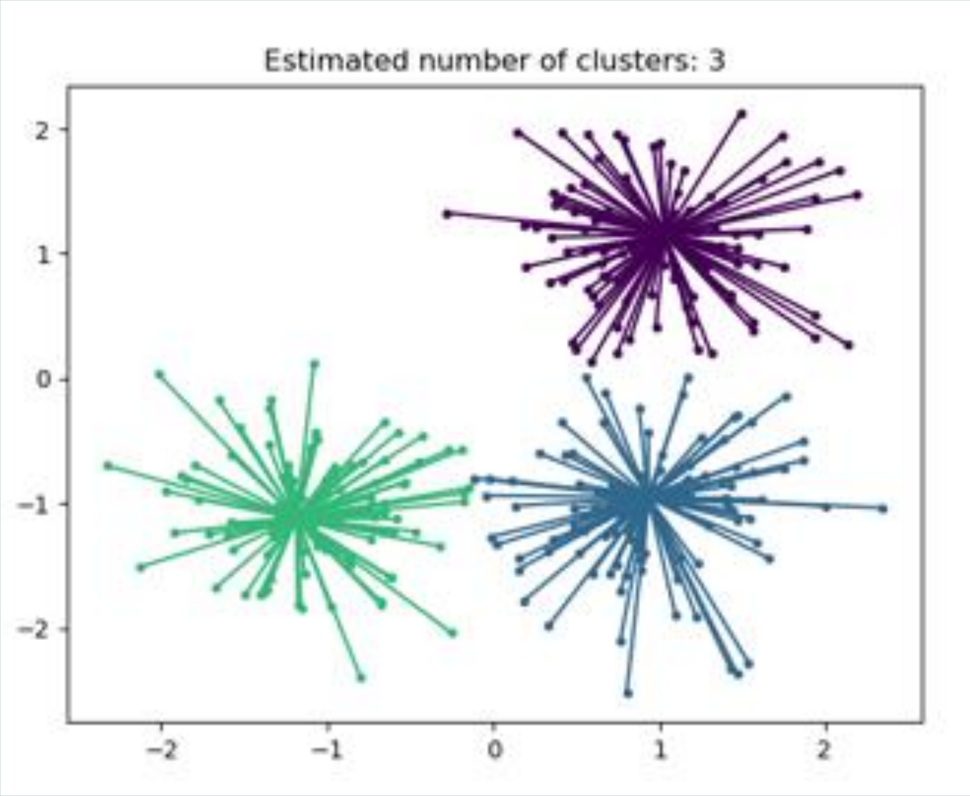
•Once a merge is made, it cannot be undone.

3.Affinity Propagation Clustering

Definition:

- Affinity Propagation is a clustering algorithm that identifies exemplars (representative points) by exchanging messages between data points.

- It does not require specifying the number of clusters beforehand.



Working Principle

- Each data point sends messages to all other points indicating how suitable they are as exemplars.
- Messages are updated iteratively based on "responsibility" and "availability" measures.
- Clusters are formed around points with the highest responsibility-availability scores.

Advantages and Disadvantages

Advantages:

•Automatically determines the number of clusters.
•Can find clusters of varying sizes and densities.

Disadvantages:

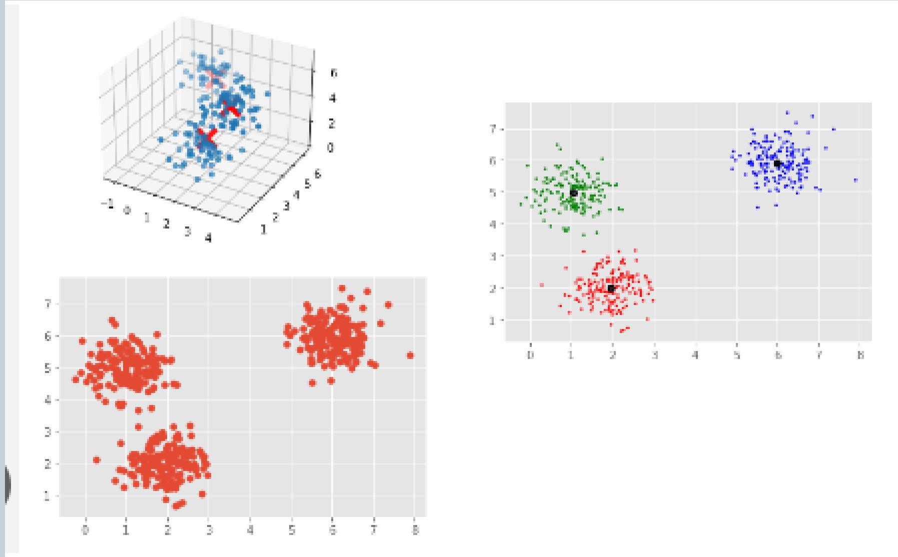
•Computationally expensive for large datasets.
•Sensitive to the choice of preference parameter.

4.MeanShift Clustering

Definition:

•MeanShift is a non-parametric, density-based clustering algorithm that does not require specifying the number of clusters.

•It works by shifting points towards areas of higher data density (modes of the density function).



Working Principle

- Place a window over each data point and compute the mean of points within the window.

- Shift the window to the mean and repeat this process until convergence.

- Points that converge to the same mode are assigned to the same cluster.

Advantages and Disadvantages

Advantages:

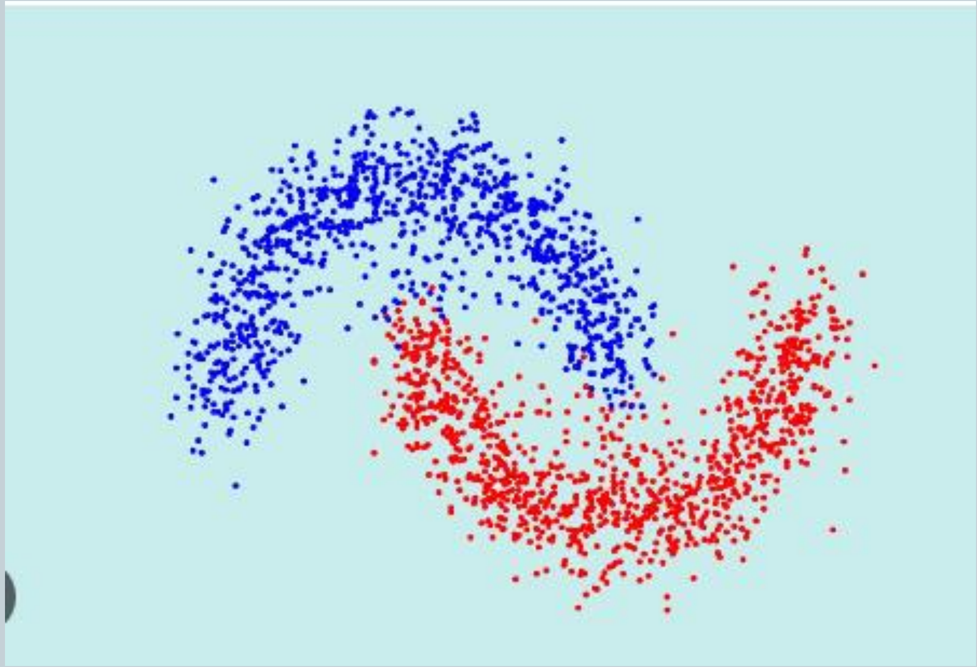
- Does not require pre-specifying the number of clusters.
- Can detect clusters of arbitrary shapes and sizes.

Disadvantages:

- Computationally expensive, especially for large datasets.
- Requires selecting a suitable bandwidth parameter for the window, which can be challenging.

5.Spectral Clustering

- **Definition:**
- Spectral Clustering is a graph-based algorithm that uses the eigenvalues of a similarity matrix to reduce dimensionality before clustering.
- It is effective for detecting non-convex clusters.



Working Principle

•Construct a similarity graph based on the data points.

•Compute the Laplacian matrix of the graph.

•Perform eigenvalue decomposition and select the top eigenvectors.

•Apply K-means clustering on the selected eigenvectors to form clusters.

Advantages and Disadvantages

Advantages:

•Suitable for non-convex clusters.

•Can handle complex data distributions.

Disadvantages:

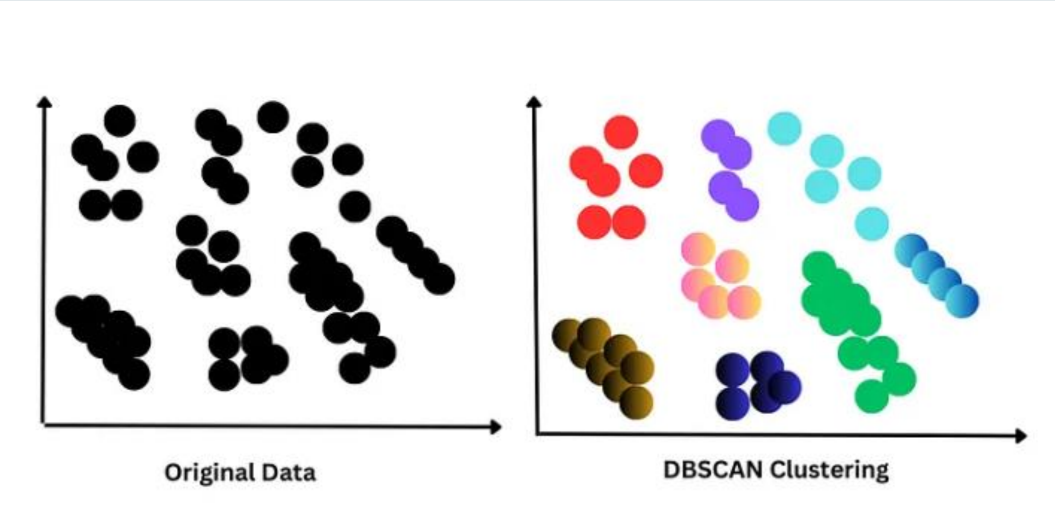
•Computationally expensive for large datasets.

•Requires a similarity matrix, which can be difficult to define.

6.DBSCAN Clustering

Definition:

•DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based algorithm that clusters points based on the density of their neighborhood.



•It can identify arbitrarily shaped clusters and outliers.

Working Principle

•Define core points with at least min_samples neighbors within eps distance.

•Expand clusters from core points by adding density-reachable points.

•Mark points not reachable by any core point as noise.

Advantages and Disadvantages

Advantages:

- Does not require specifying the number of clusters.
- Can find arbitrarily shaped clusters and detect outliers.

Disadvantages:

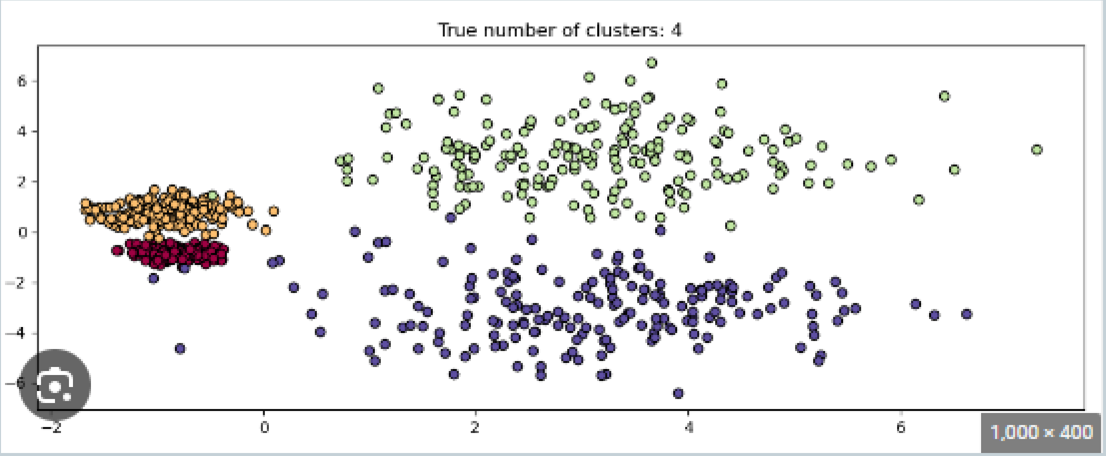
- Sensitive to parameter settings (eps and min_samples).
- Not suitable for datasets with varying densities.

7.HDBSCAN Clustering

Definition:

•HDBSCAN (Hierarchical DBSCAN) extends DBSCAN by converting it into a hierarchical clustering algorithm.

•It allows for varying cluster densities and provides a hierarchy of clusters.



Working Principle

- Construct a minimum spanning tree of the distance-weighted data points.
- Generate a hierarchy of clusters by varying the density threshold.
- Extract clusters using a stability measure to find the most persistent clusters.

Advantages and Disadvantages

Advantages:

•No need to specify the number of clusters.

•Handles varying cluster densities and noise.

Disadvantages:

•Computationally more expensive than DBSCAN.

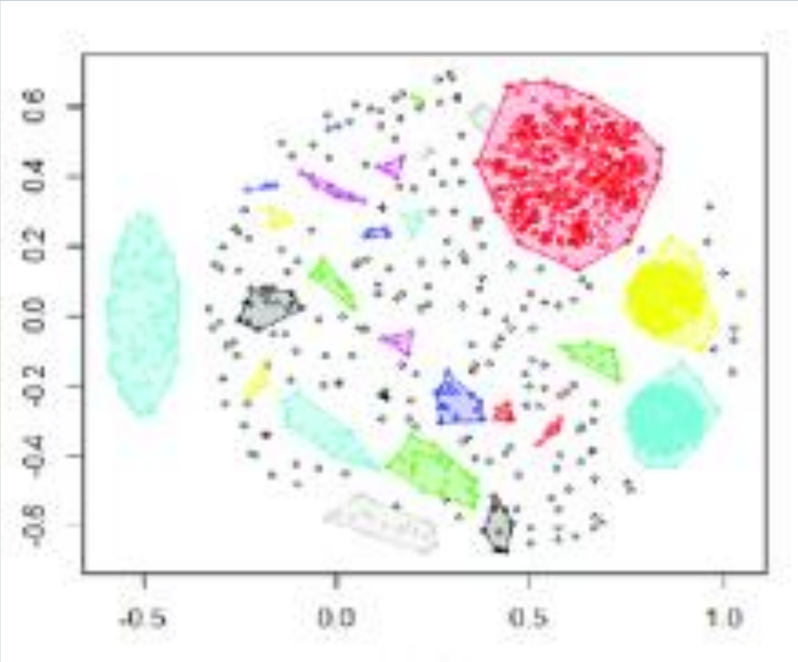
•Parameter tuning can be complex.

8.OPTICS Clustering

Definition:

•OPTICS (Ordering Points To Identify the Clustering Structure) is a density-based clustering algorithm that extends DBSCAN to handle varying densities.

•It produces an ordering of the points that reflects their density-based clustering structure.



Working Principle

- The algorithm orders data points based on their density reachability distance.
- A reachability plot is generated, and clusters are identified by valleys in the plot.
- Points with similar densities form clusters, while points with no neighbors are considered noise.

Advantages and Disadvantages

Advantages:

•Handles clusters of varying densities well.
•Does not require specifying the number of clusters.
•Can identify nested clusters and noise points.

•**Disadvantages:**

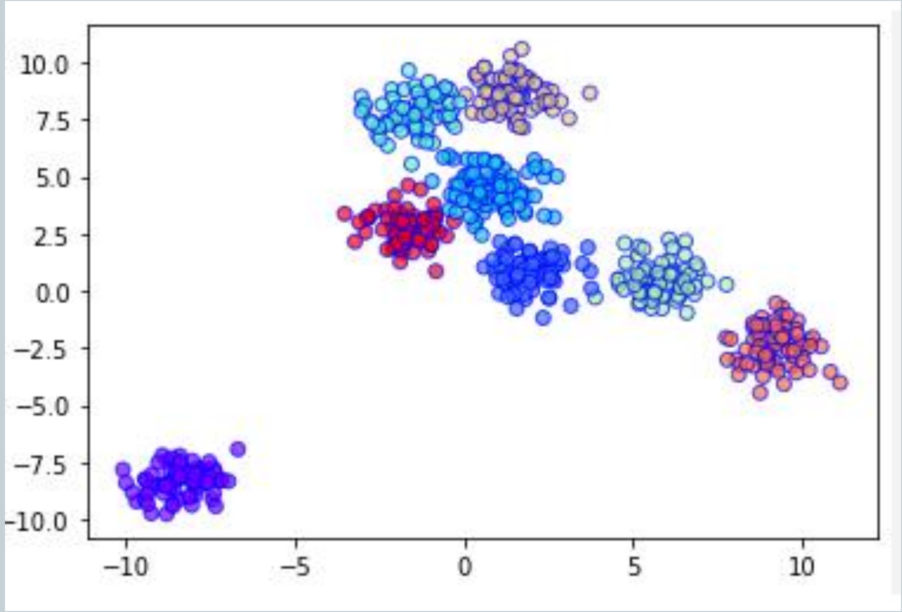
•Computationally expensive compared to DBSCAN.
•Interpreting reachability plots can be complex.

9.BIRCH Clustering

Definition:

•BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is a hierarchical clustering algorithm designed for large datasets.

•It builds a Clustering Feature tree structure (CF tree) to summarize the data, and clusters are formed by traversing this structure.



Working Principle

- Build a CF tree, where each node represents a cluster of data points summarized by their centroid.

- Points are inserted into the tree, and sub-clusters are formed dynamically.

- Clusters are refined during the global clustering phase, using methods like K-means or agglomerative clustering.

Advantages and Disadvantages

Advantages:

- Scalable to large datasets.
- Can incrementally process new data points.
- Handles noise and outliers effectively.

Disadvantages:

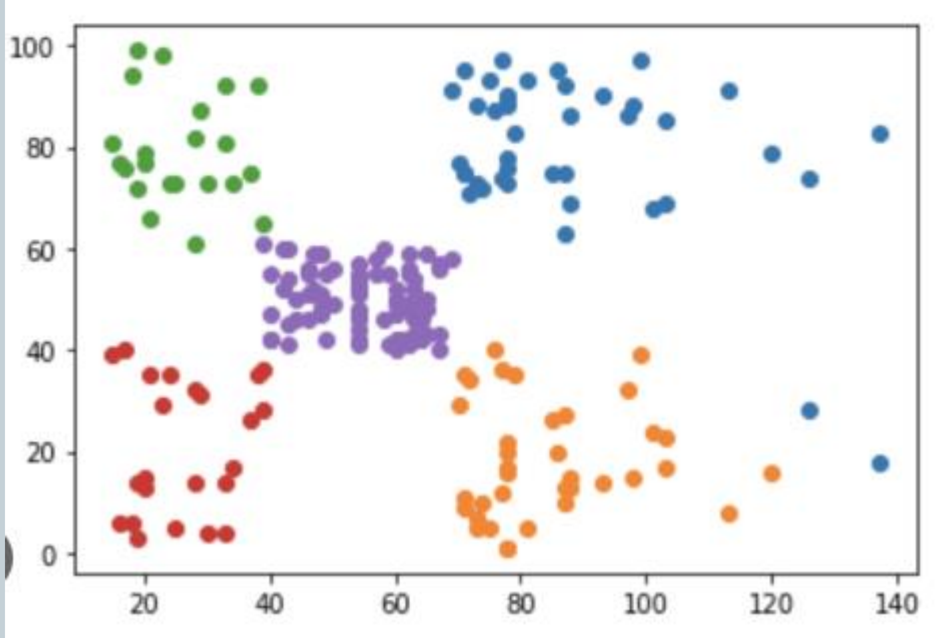
- Performance can degrade with high-dimensional data.
- The structure of the CF tree depends heavily on the input order of data.

10.BisectingKMeans Clustering

Definition:

•Bisecting K-means is a hierarchical (top-down) clustering algorithm that repeatedly splits clusters using K-means.

•It combines elements of both divisive hierarchical clustering and K-means partitioning.



Working Principle

- Start with all data points in a single cluster.

- At each iteration, the current cluster is bisected using K-means to form two sub-clusters.

- The bisection that minimizes intra-cluster variance is chosen, and this process continues until the desired number of clusters is reached.

Advantages and Disadvantages

Advantages:

- Often more efficient and scalable than standard K-means.
- Allows control over the number of splits and clusters.

Disadvantages:

- Sensitive to initial cluster centroids.
- May result in suboptimal splits if poor bisections are made early.