CLUSTERING ALGORITHMS

AGENDA

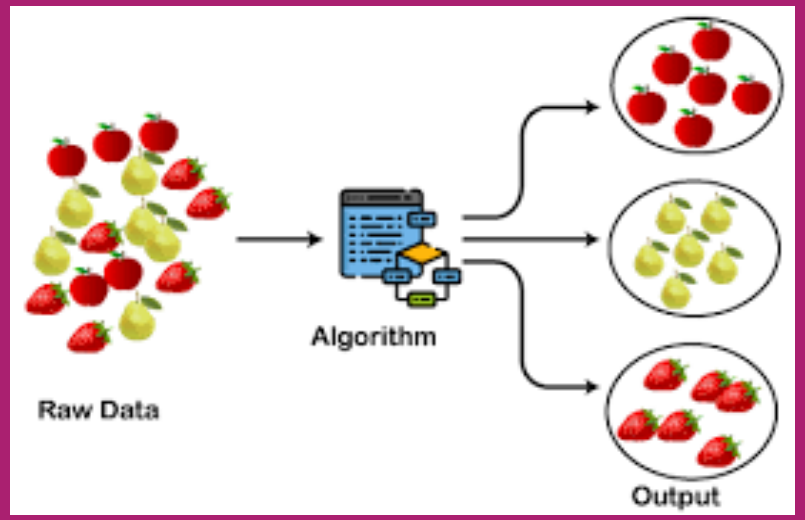
Introduction

Types

Top clustering algorithms

CLUSTERING IS AN UNSUPERVISED MACHINE LEARNING TASK.

A CLUSTER IS A GROUP OF DATA POINTS THAT ARE SIMILAR TO EACH OTHER BASED ON THEIR RELATION TO SURROUNDING DATA POINTS..



THERE ARE DIFFERENT TYPES OF CLUSTERING ALGORITHMS THAT HANDLE ALL KINDS OF UNIQUE DATA

DENSITY BASED

DISTRIBUTION BASED

CENTROID BASED

HIERARCIAL BASED

CENTROID BASED

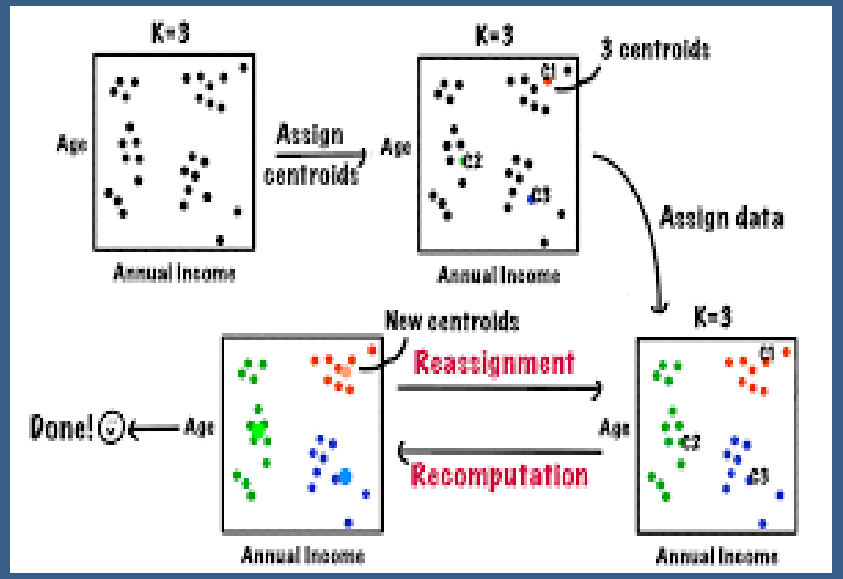
A centroid is a data point that represents the center of the cluster.

Centroid-based clustering organizes the data into non-hierarchical clusters.

Sensitive to initial conditions and outliers.

TWO CENTROID BASED ALGORITHMS ARE:

1.Kmeans-numerical data

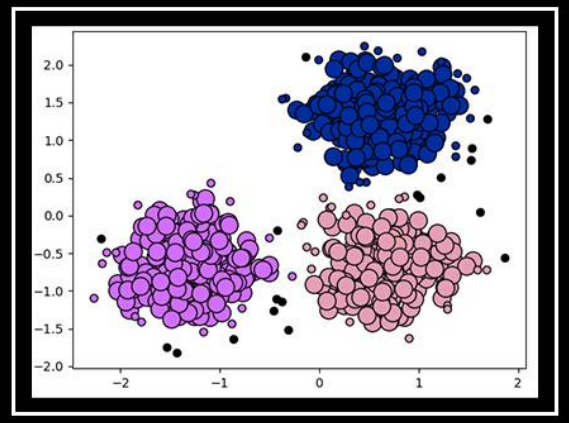


2.K mode-categorical data

DENSITY BASED

Data is grouped by areas of high concentrations of data points surrounded by areas of low concentrations of data points.

Clusters can be any shape.



1.DBSCAN Algorithm

Data points movement based on Euclidean diatance

2.Mean Shift Algorithm

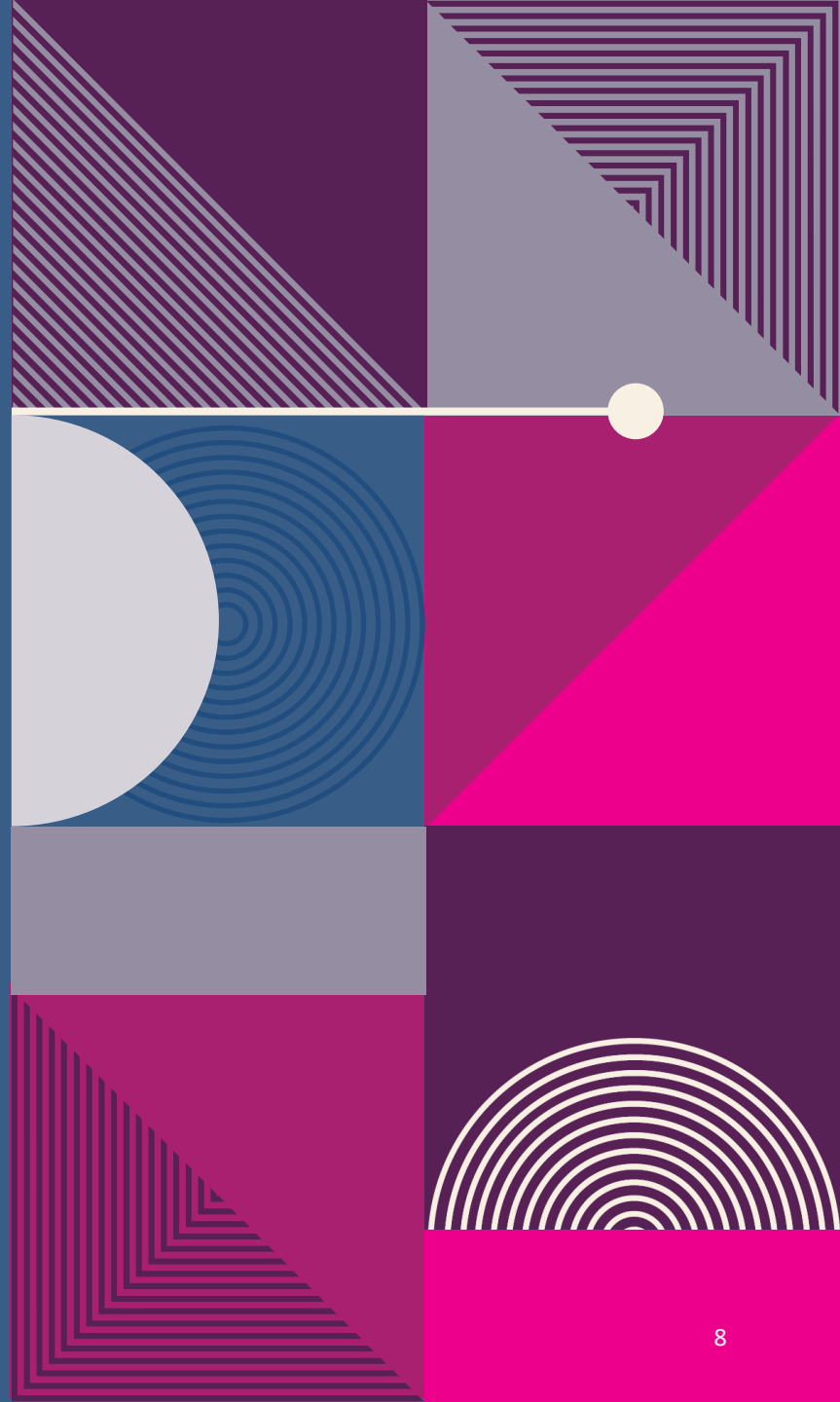
Data Point movement based on mean shift vector

3.Affinity Based Algorithm

Data point movement based on similarity matrix, availability matrix, responsibility matrix.

4.OPTICS Algorithm

Each Data point is associated with Reachability Distance.



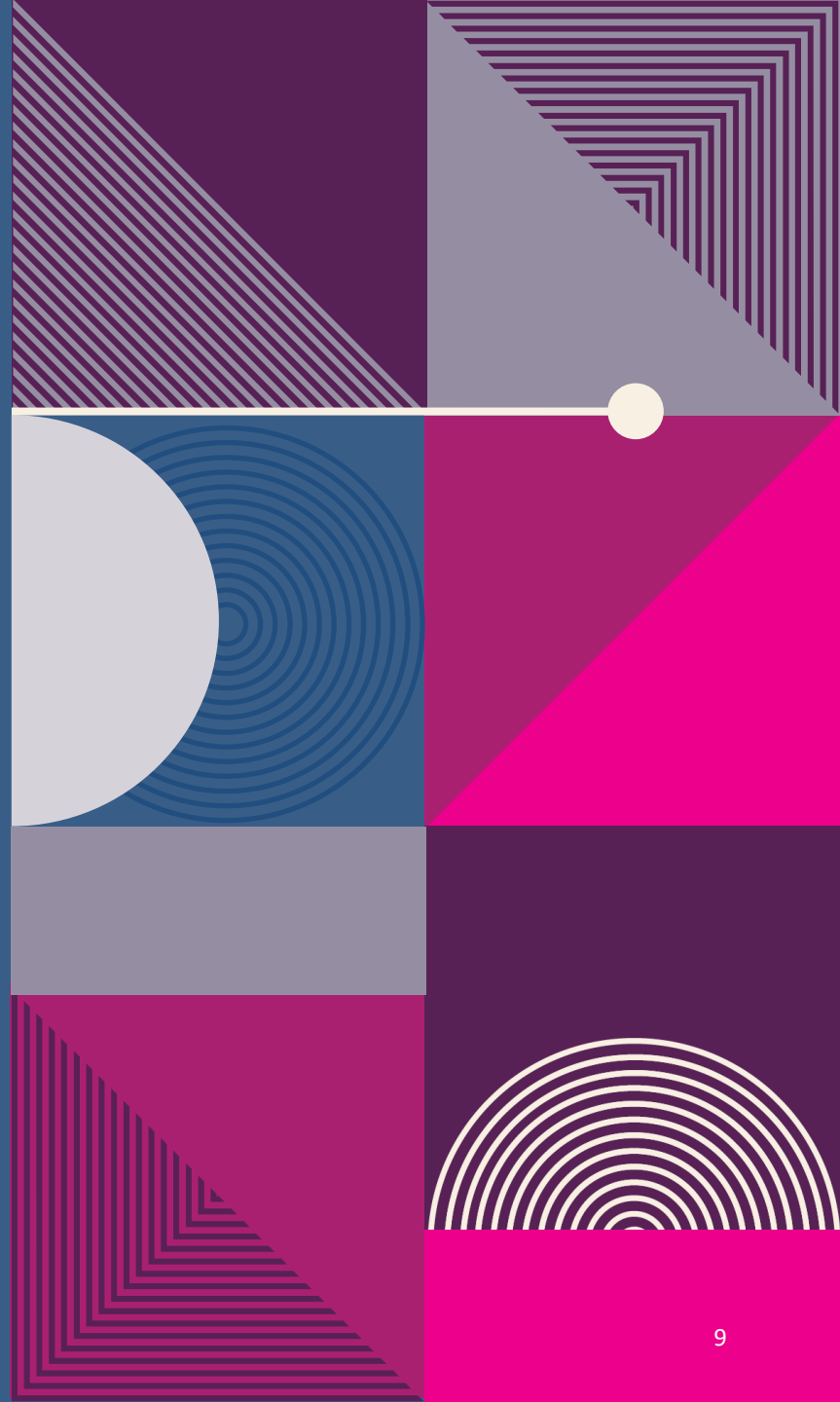
DISTRIBUTION BASED

All the data points are considered parts of a cluster based on the probability that they belong to a given cluster.

1.Gaussian Mixture Model

Each Gaussian Mixture distribution represents a cluster.

Ability to handle overlapping clusters, model the covariance structure of the data, and provide probabilistic cluster assignments for each data point.



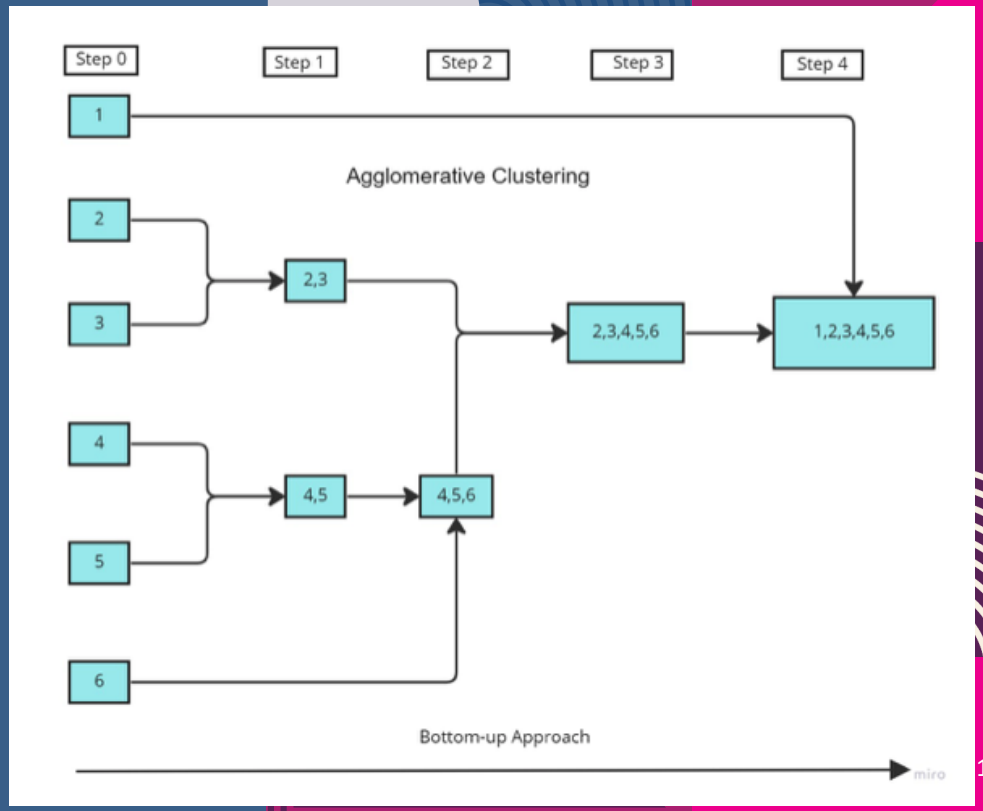
HIERARCHIAL BASED

BUILDS A TREE OF CLUSTERS SO EVERYTHING IS ORGANIZED BETWEEN THE CLUSTERS.

The 'ward' method is used to calculate the distances between the clusters. It minimizes the variance of the distances between the clusters being merged.
Dendrogram shows the hierarchical relationship .

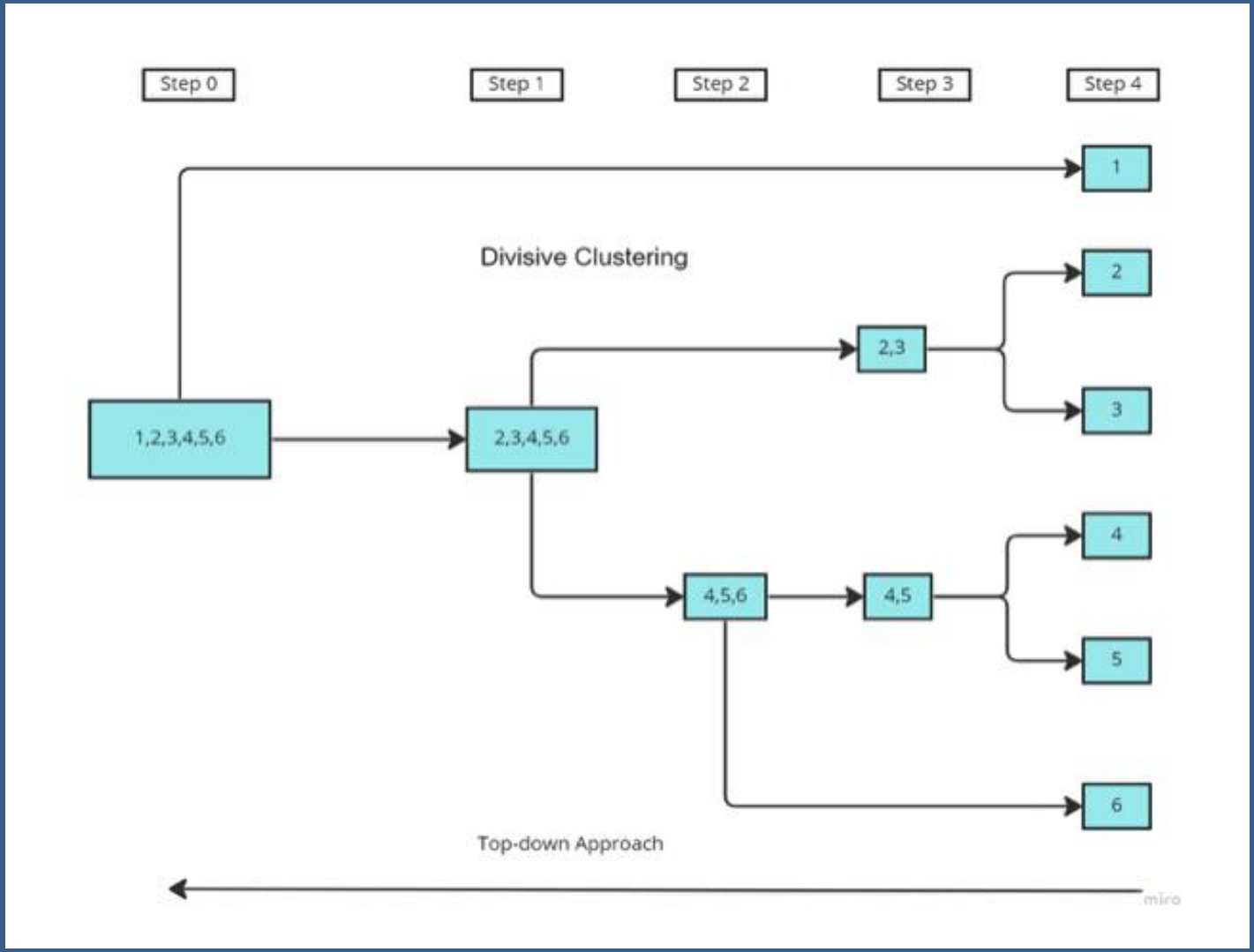
Agglomerative Clustering

It is a bottom-up approach that produces a dendrogram.



Divisive Clustering

It is a top-down approach.



Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH)

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that can cluster large datasets by first generating a small and compact summary of the large dataset that retains as much information as possible. This smaller summary is then clustered instead of clustering the larger dataset.

Based on factors such as Cluster feature.