DATA SCIENCE WITH PYTHON : HIERARCHICAL CLUSTERING #1830

Presented by: Deepthi M

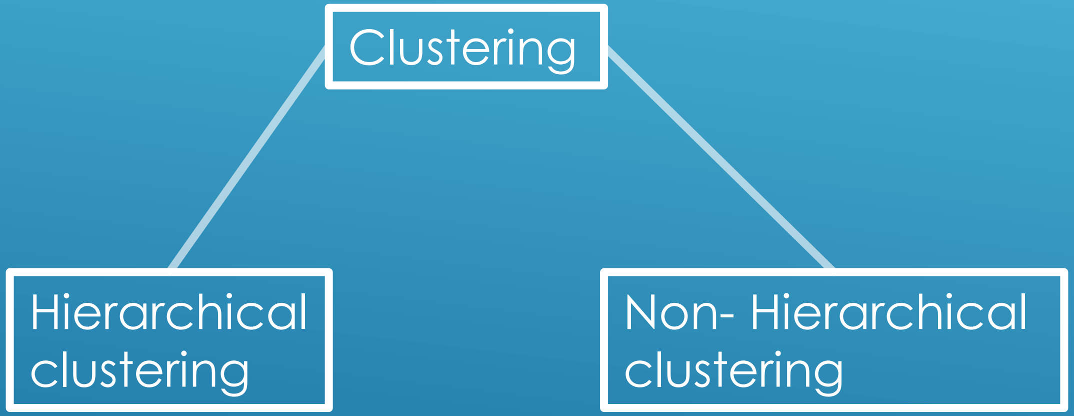
Batch Number: 05

Serial Number: 172

What is Clustering?

Clustering is a data mining technique which is used to group the data based on their similarities or differences.

Clustering types:
Hierarchical clustering
Non- Hierarchical clustering



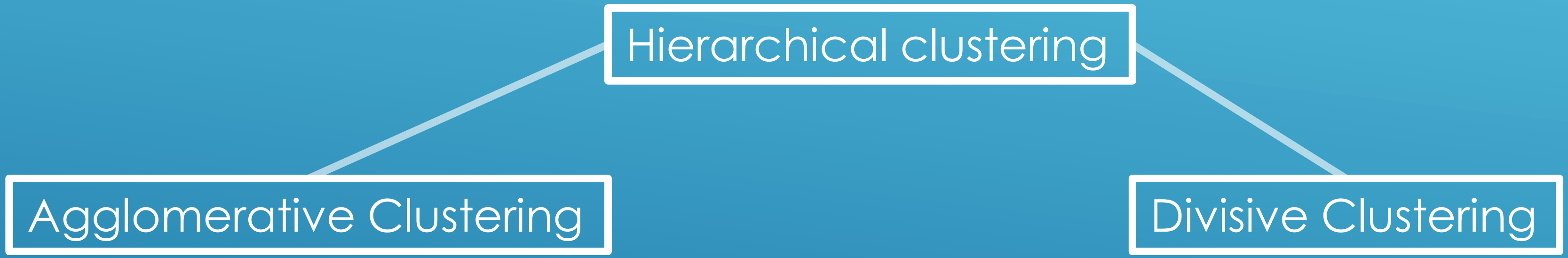
Clustering

Hierarchical clustering

Non- Hierarchical clustering

Hierarchical clustering

Hierarchical clustering follows a hierarchy. It can be categorized into two types: agglomerative clustering and divisive clustering



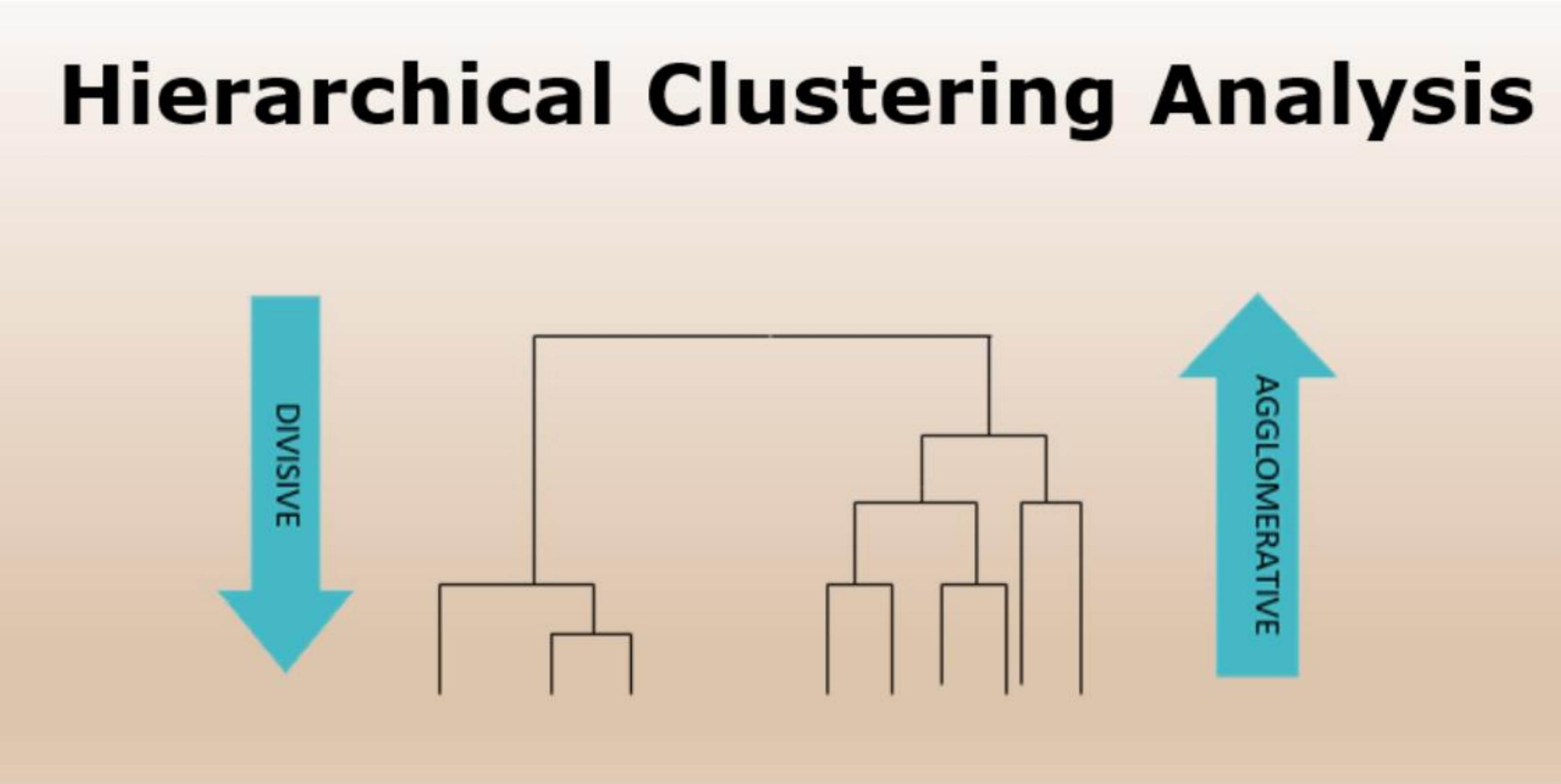
Hierarchical clustering

Agglomerative Clustering

Divisive Clustering

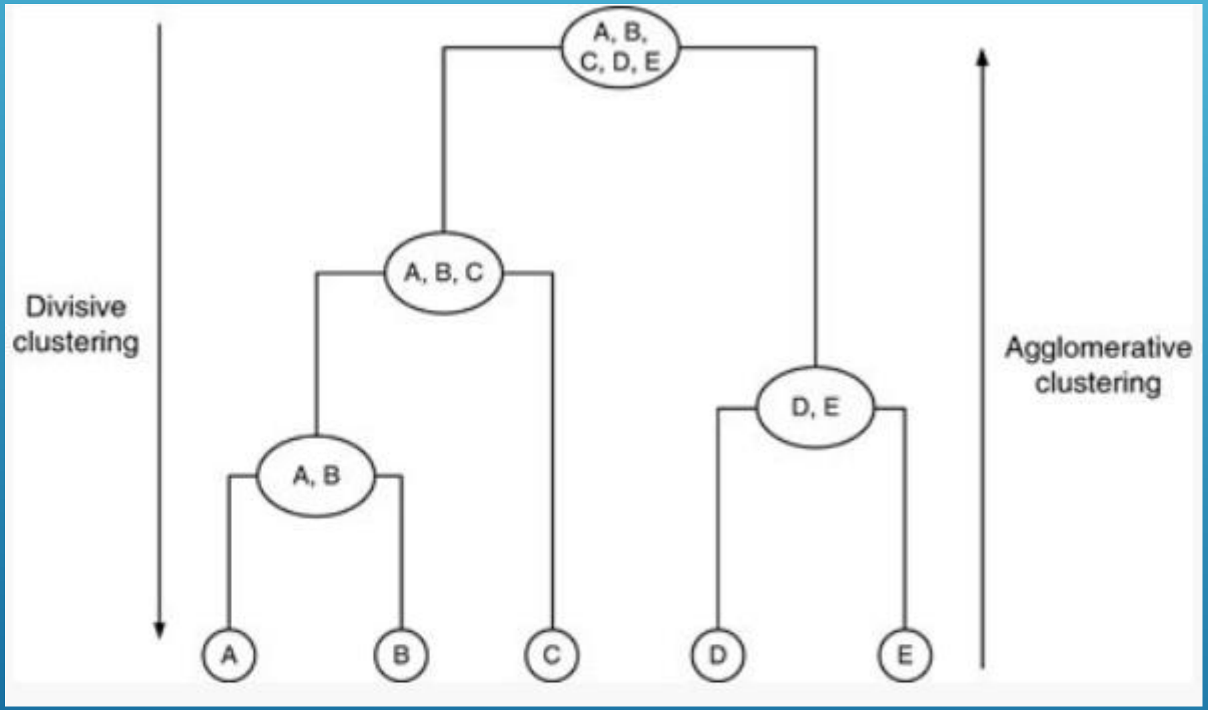
Agglomerative Clustering:

- Bottom-up approach.
- First data points are grouped separately and merged into a single cluster iteratively based on similarity.
- Distance used to measure the similarity between data points.



Divisive Clustering

- Top-down approach.
- First data points are grouped into a single cluster and separated into several clusters iteratively based on similarity.
- Distance used to measure the similarity between data points.



	Distance between 2 records
Numerical Data	Euclidean Distance Manhattan Distance
Categorical (Binary) Data	Simple Matching Coefficient Jaccard's Index Binary Euclidean Distance
Categorical (Multiple) Data	If the two categories are same then distance will be Zero •If the two categories are different then distance will be One

Distance between a record and a cluster; Or between 2 clusters

- Performed using Linkage functions and using distance measures (Euclidean, Manhattan)
- Linkage Functions

1. Single Linkage :

Minimum distance between members of the two clusters

2. Complete Linkage :

Greatest distance between members of the two clusters

3. Average Linkage :

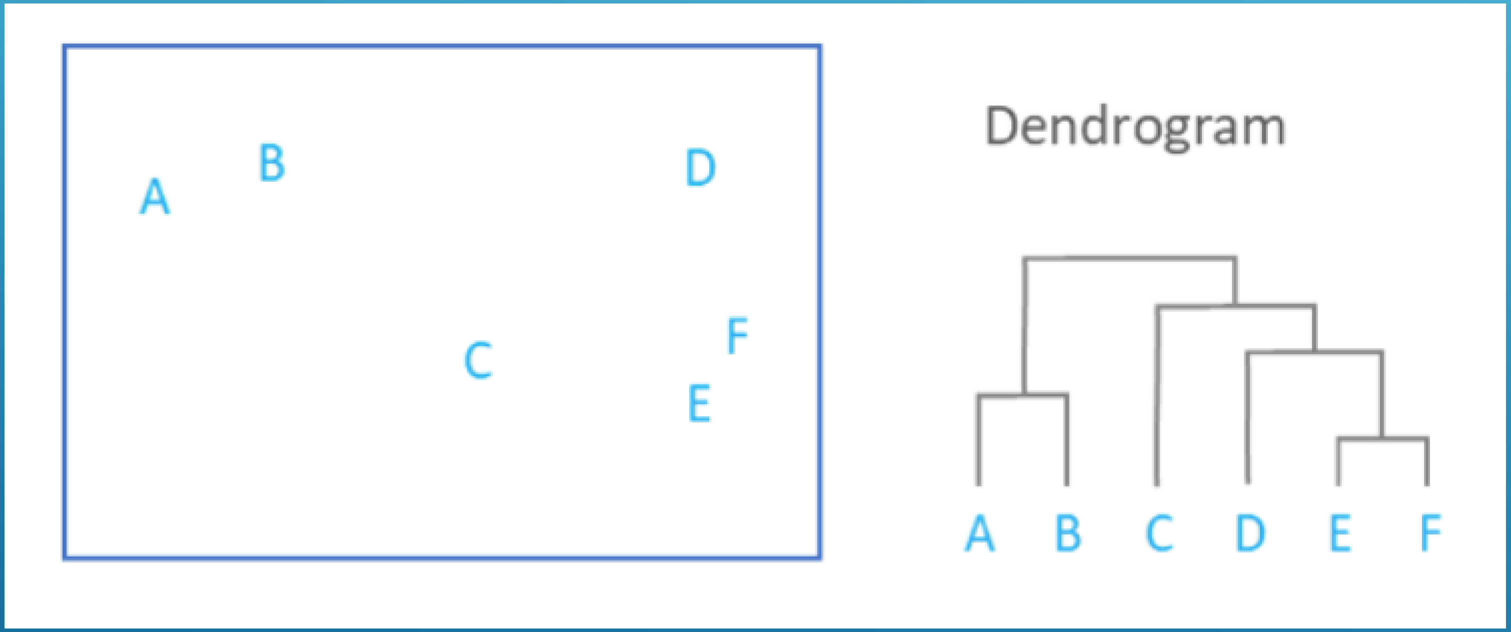
Average of all distances between members of the two clusters

4. Centroid Linkage :

Distance between their centroids (centres)

An over view of hierarchical clustering:

- Number of cluster is upfront not decided.
- It is stable.
- Dendrogram visualization.
- Follows an hierarchy like top-down or bottom-up.

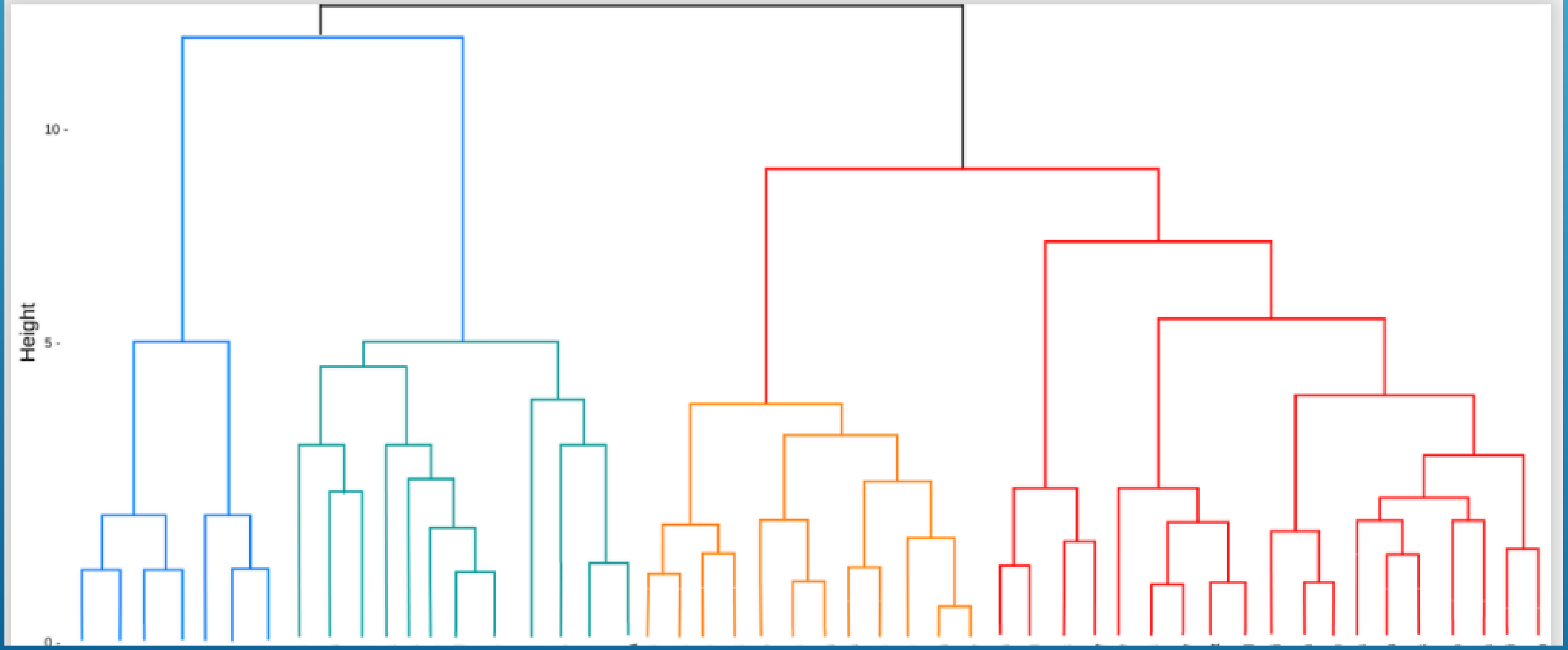


What makes clusters good:

- Intra-class similarity is high and inter –class similarity is low

Challenges:

- In accurate results as it involves human intervention to validate the output.
- Computational complexity due to large training data.



I hope now we have an overview on the topic:

- Hierarchical Clustering.

Thank you I hope you have enjoyed!!!