Redesigning Common Mind & Business Towards Excellence

3P urpose rocess eople Culture

sns INSTITUTIONS

www.snsgroups.com

1st GenAI Powered Design Thinking FrameWork

Build an Entrepreneurial Mindset Through Our Design Thinking FrameWork

SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

23AMB201 - MACHINE LEARNING

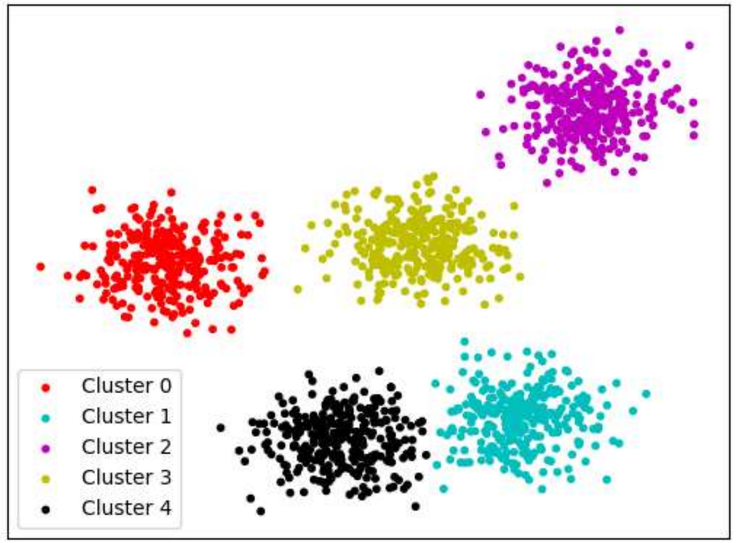
II YEAR IV SEM

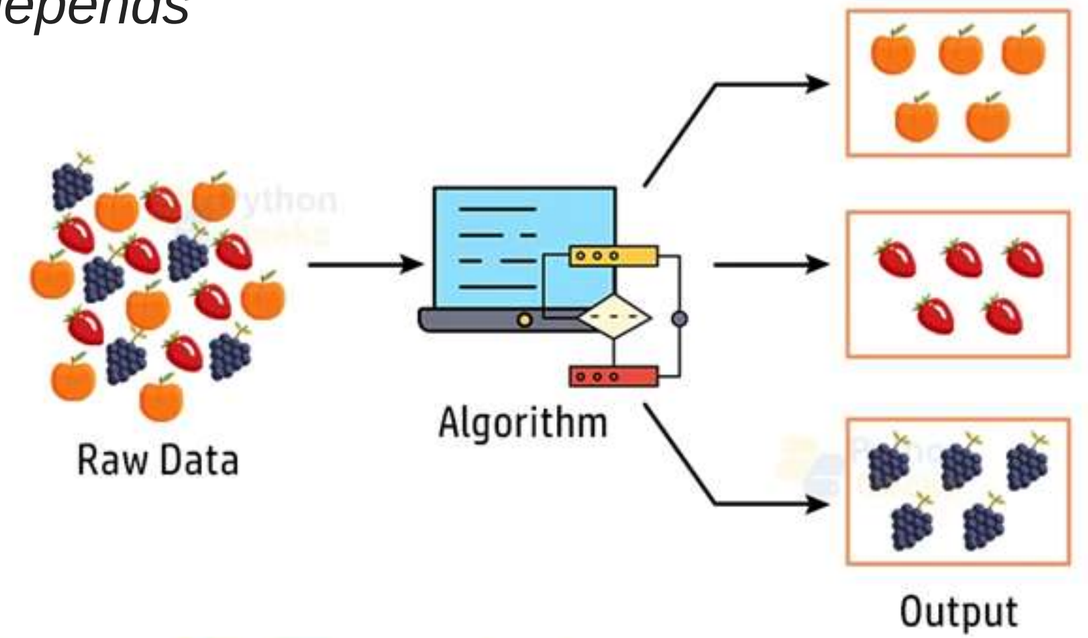
UNIT IV – UNSUPERVISED LEARNING ALGORITHM

TOPIC 24 – Hierarchical Clustering

Clustering

Grouping the same items together depends distance.





Types of Clustering

Hierarchical algorithms: - Find successive clusters

1. **Agglomerative** ("bottom-up"): Begins with each element as a separate cluster and merge them into successively larger clusters.
2. **Divisive** ("top-down"): Begins with the whole set and proceed to divide it into successively smaller clusters.

Partitional clustering: Partitional algorithms determine all clusters at once.

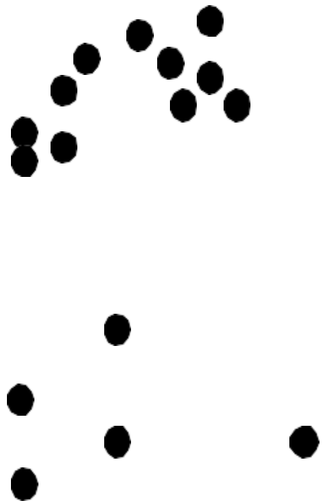
They include:

1. K-means and derivatives
2. Fuzzy c-means clustering

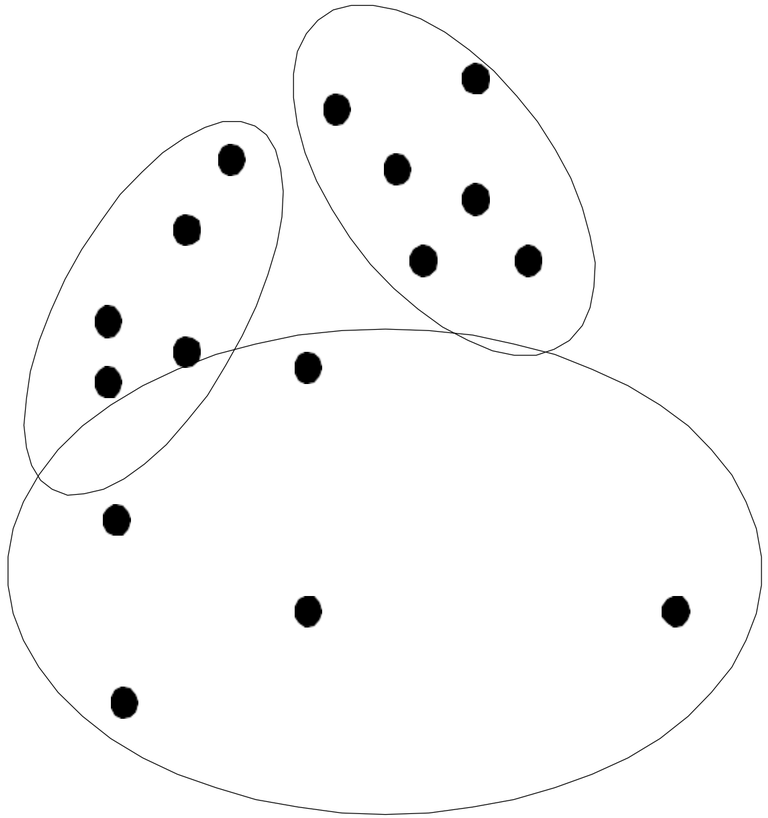
Density based clustering

Fuzzy clustering

Partitional Clustering



Original Points

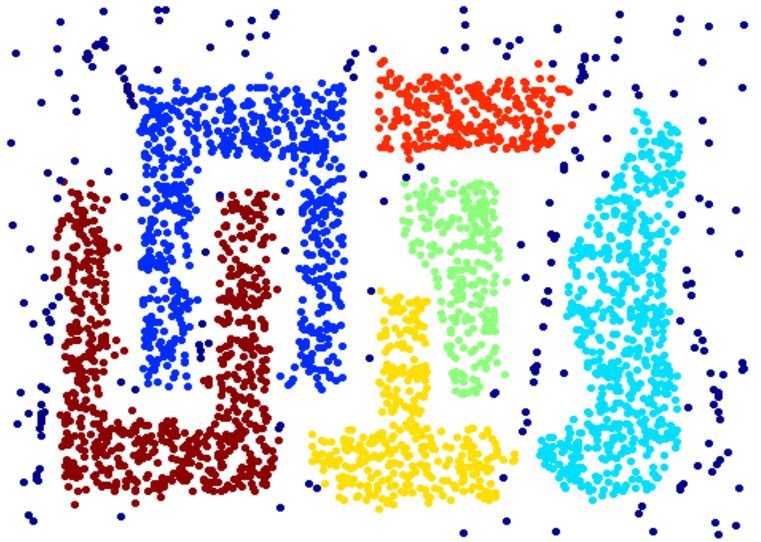


A Partitional Clustering

Density based Clustering

Original Points

Clusters

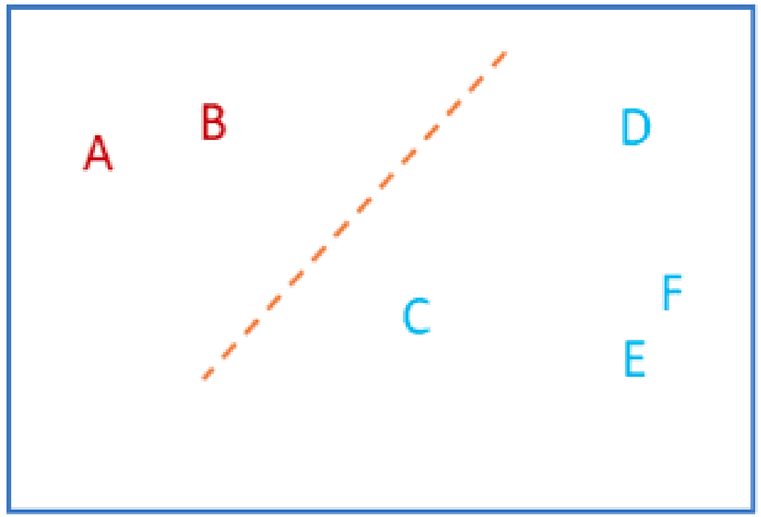


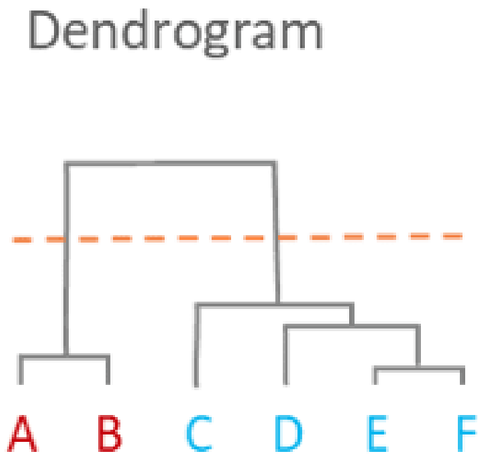
Hierarchical Clustering

- **Hierarchical clustering** is a powerful unsupervised machine learning algorithm used to group data points into a hierarchy of clusters. It is particularly useful when the number of clusters is not predefined, and it helps to visualize the data's structure through a **Dendrogram**, which represents the nested clustering relationships.

Applications:

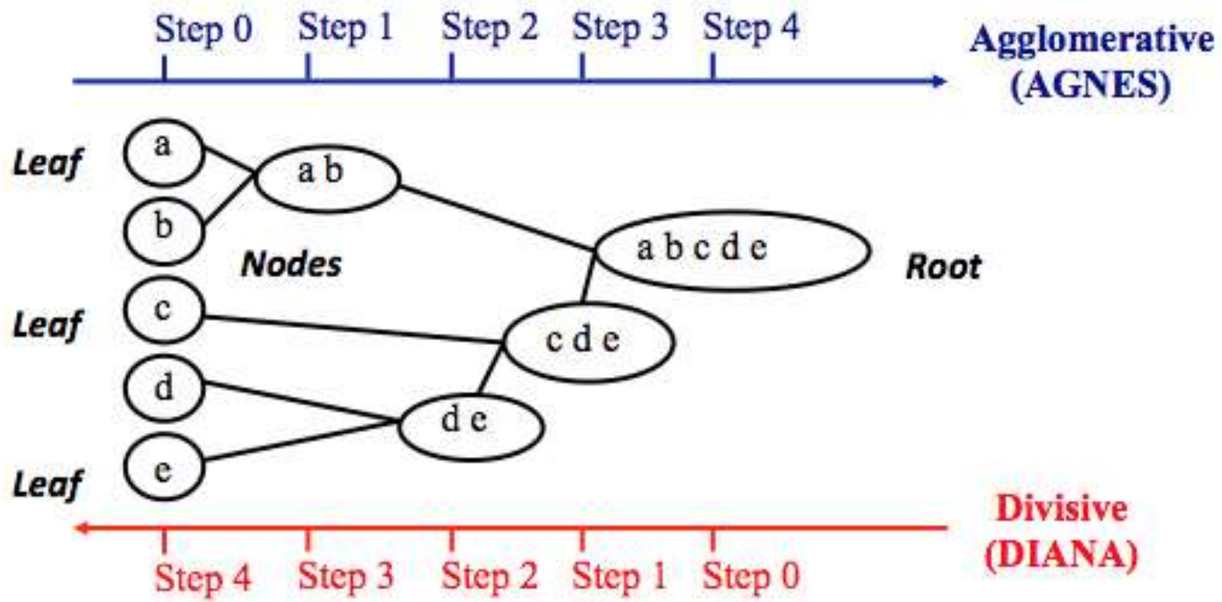
1. Biology
2. Marketing
3. Social network analysis





Types of Hierarchical Clustering

1. **Agglomerative:** Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**



How the Agglomerative Hierarchical clustering Work?

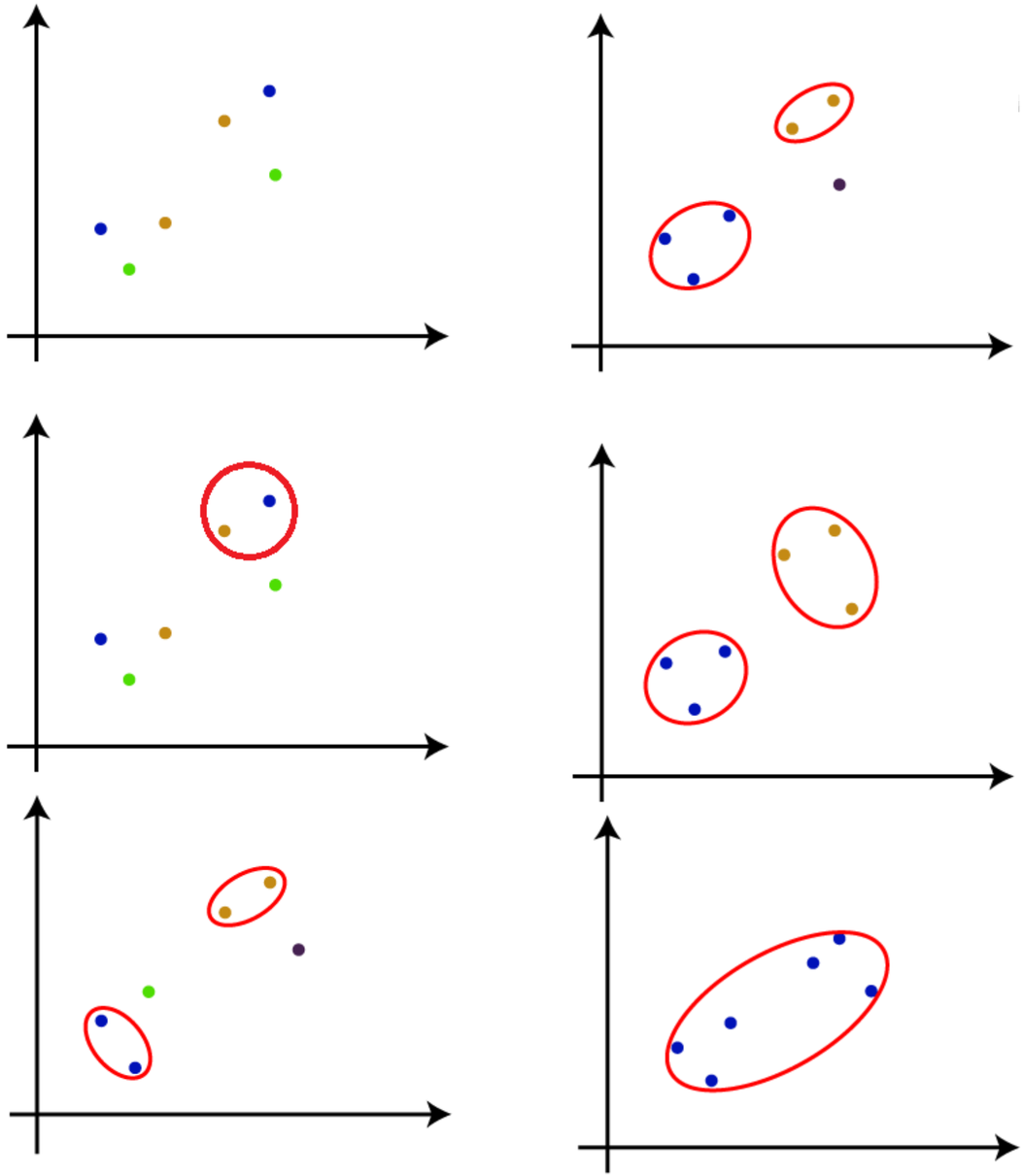
Step-1: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.

Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

Step-3: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.

Step-4: Repeat Step 3 until only one cluster left. So, we will get the following clusters. Consider the below images:

Step-5: Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem.



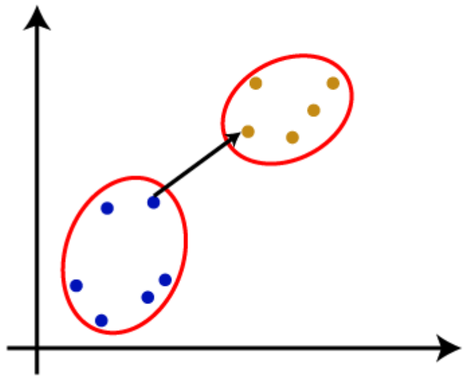
Measure for the distance between two clusters: Linkage Methods

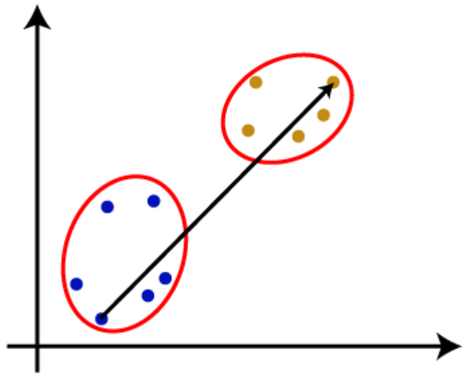
Single Linkage: It is the Shortest Distance between the closest points of the clusters. Consider the below image:

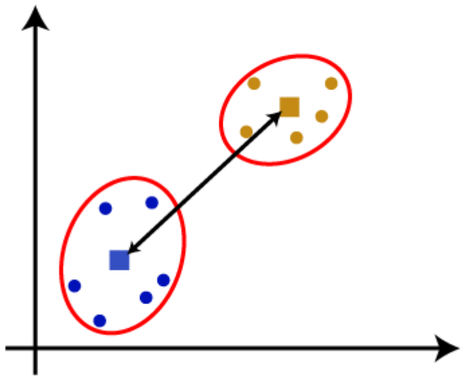
Complete Linkage: It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.

Average Linkage: It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters.

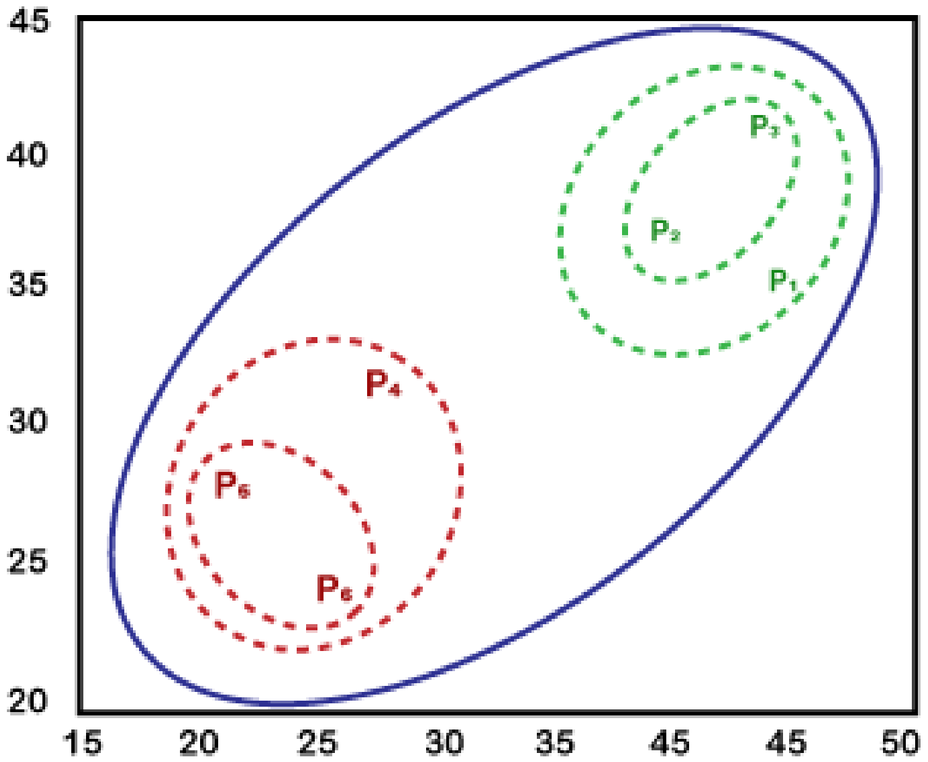
Centroid Linkage: It is the linkage method in which the distance between the centroid of the clusters is calculated.

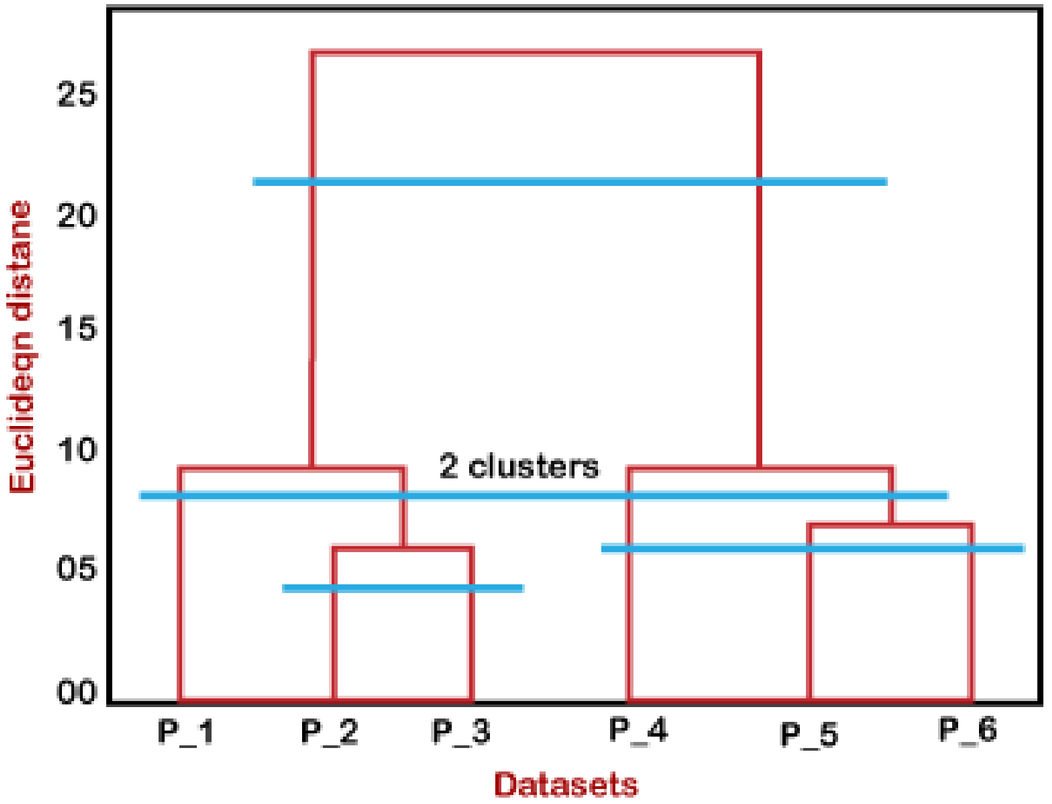






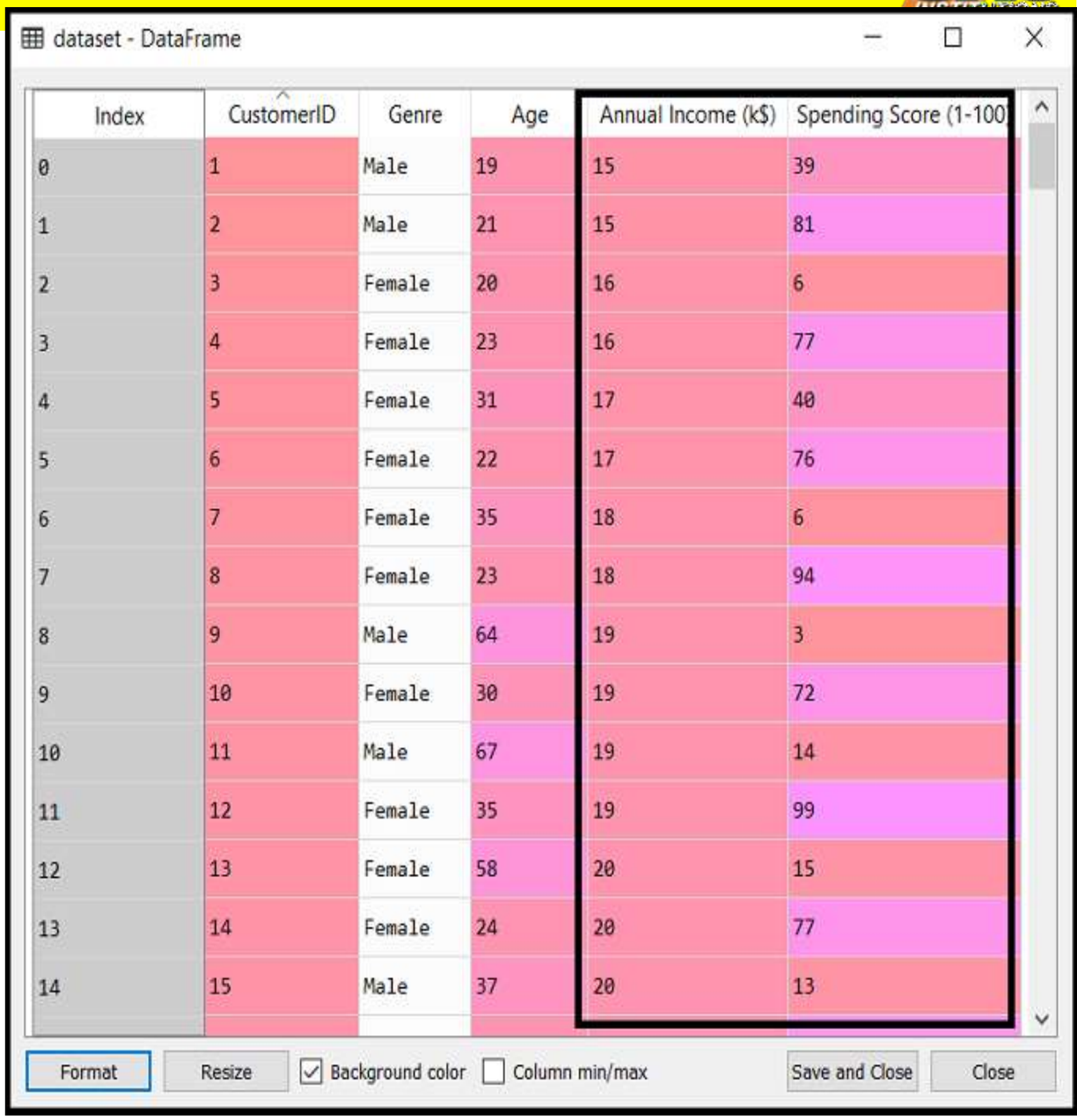
Woking of Dendrogram in Hierarchical clustering





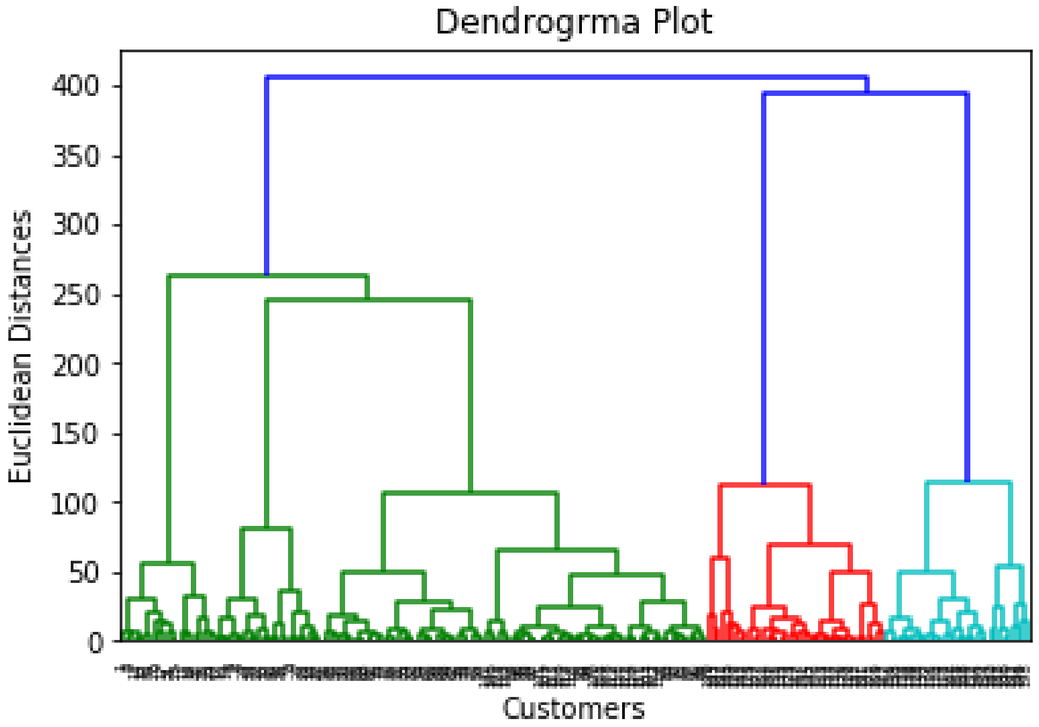
Python Implementation of Agglomerative Hierarchical Clustering

```
# Importing the libraries
import numpy as nm
import matplotlib.pyplot as mtp
import pandas as pd
# Importing the dataset
dataset = pd.read_csv('Mall_Customers_data.csv')
x = dataset.iloc[:, [3, 4]].values
import scipy.cluster.hierarchy as shc
dendro = shc.dendrogram(shc.linkage(x, method="ward"))
mtp.title("Dendrogrma Plot")
mtp.ylabel("Euclidean Distances")
mtp.xlabel("Customers")
mtp.show()
```



Index	CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	Female	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72
10	11	Male	67	19	14
11	12	Female	35	19	99
12	13	Female	58	20	15
13	14	Female	24	20	77
14	15	Male	37	20	13

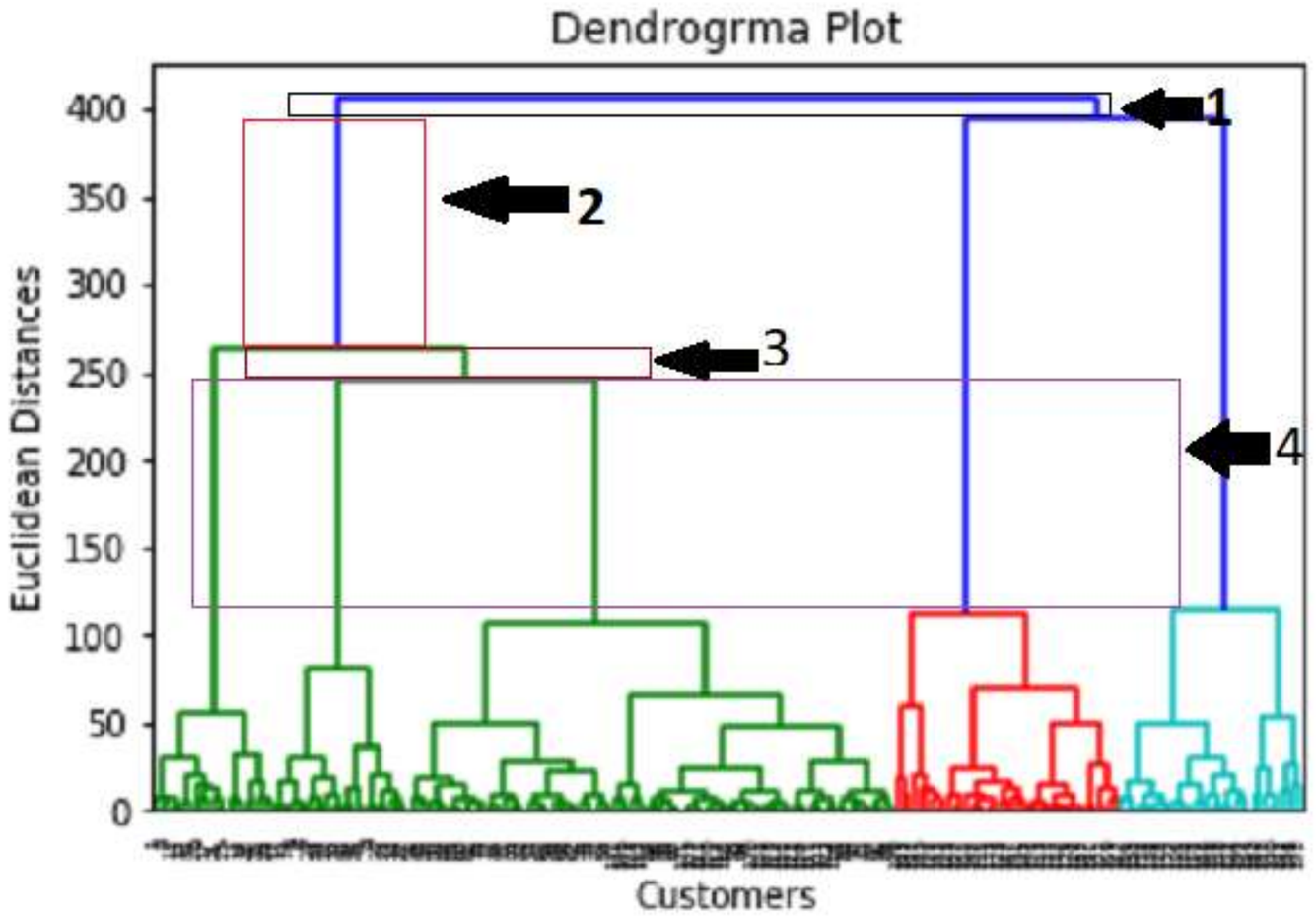
Output



The 4th distance is looking the maximum, so according to this, **the number of clusters will be 5**(the vertical lines in this range)

The optimal number of clusters will be 5 and train the model in the next step

find the **maximum vertical distance** that does not cut any horizontal bar



Training the hierarchical clustering model

1. The 4th distance is looking the maximum, so according to this, **the number of clusters will be 5**(the vertical lines in this range)
2. **The optimal number of clusters will be 5 and train the model in the next step**

```
#training the hierarchical model on dataset
from sklearn.cluster import AgglomerativeClustering
hc= AgglomerativeClustering(n_clusters=5, affinity='euclid
ean', linkage='ward')
y_pred= hc.fit_predict(x)
```

Output of training samples

Visualizing the clusters

```
#visulaizing the clusters
mtp.scatter(x[y_pred == 0, 0], x[y_pred == 0, 1], s = 100, c = 'blue', label = 'Cluster 1')
mtp.scatter(x[y_pred == 1, 0], x[y_pred == 1, 1], s = 100, c = 'green', label = 'Cluster 2')
mtp.scatter(x[y_pred== 2, 0], x[y_pred == 2, 1], s = 100, c = 'red', label = 'Cluster 3')
mtp.scatter(x[y_pred == 3, 0], x[y_pred == 3, 1], s = 100, c = 'cyan', label = 'Cluster 4')
mtp.scatter(x[y_pred == 4, 0], x[y_pred == 4, 1], s = 100, c = 'magenta', label = 'Cluster 5')
mtp.title('Clusters of customers')
mtp.xlabel('Annual Income (k$)')
mtp.ylabel('Spending Score (1-100)')
mtp.legend()
mtp.show()
```

References

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, —Learning from Data, AML Book Publishers, 2012.

P. Flach, —Machine Learning: The art and science of algorithms that make sense of data‖, Cambridge University Press, 2012.

https://www.appliedaicourse.com/blog/hierarchical-clustering-in-machine-learning/

https://www.tpointtech.com/hierarchical-clustering-in-machine-learning

https://www.analyticsvidhya.com/blog/2022/11/hierarchical-clustering-in-machine-learning/