**Charotar University of Science and Technology [CHARUSAT]**

**Chandubhai S. Patel Institute of Technology [CSPIT]**

**U & P U. Patel Department of Computer Engineering**

**Lab Manual**

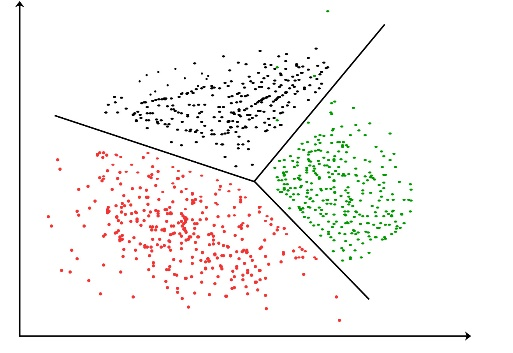
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Subject code | : | CE350 | Semester | : | 6 | Academic Year | : | 2020-2021 |
| Subject name | : | DATA WAREHOUSE AND DATA MINING | | | | | | |

**Practical – 8**

**AIM: Cluster Inspection and Hierarchical Clustering**

**Prerequisite knowledge:**

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

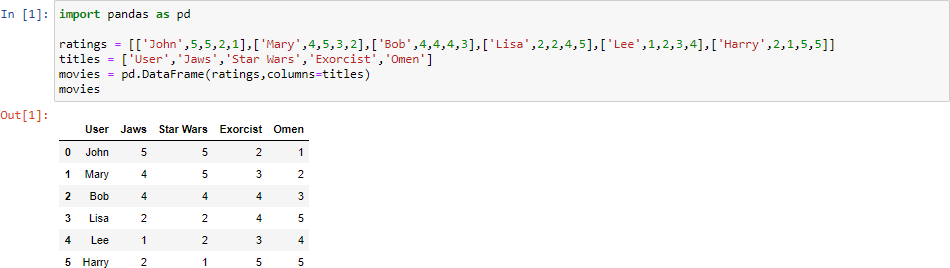
1) K-means Clustering – It is the simplest unsupervised learning algorithm that solves clustering problem. K-means algorithm partition n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster .  


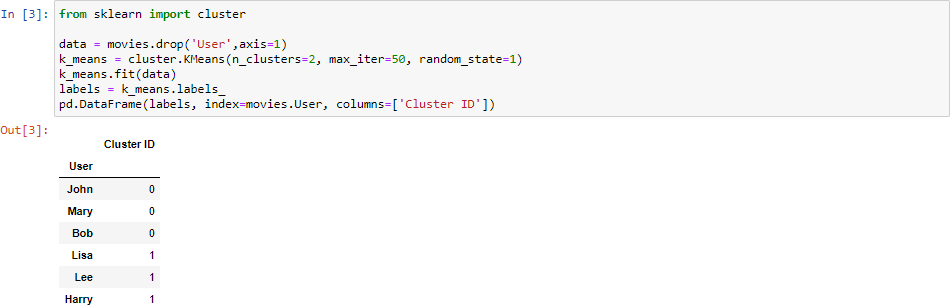
2) Hierarchical Clustering: A Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data points as a separate cluster. Then, it repeatedly executes the subsequent steps:

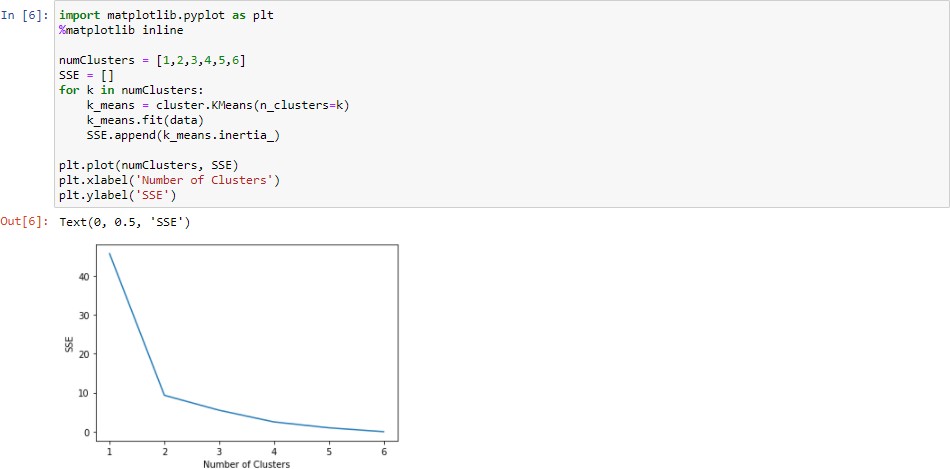
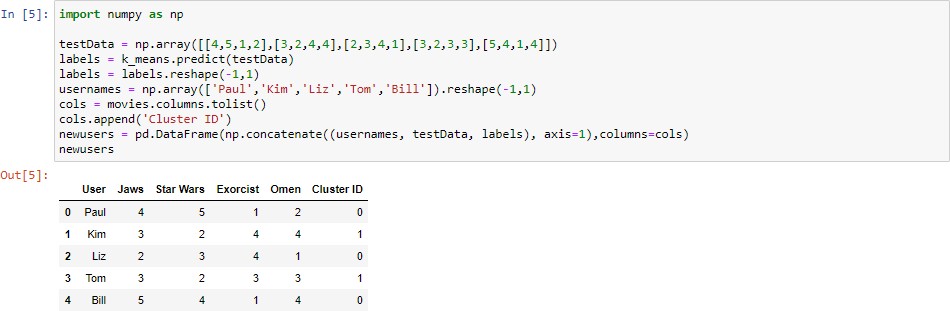
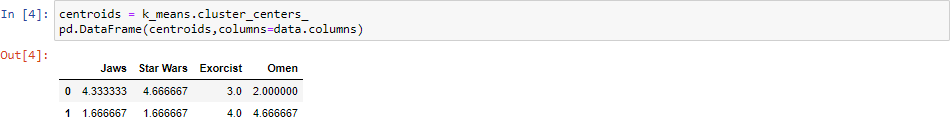
1. Identify the 2 clusters which can be closest together, and
2. Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top-down view).

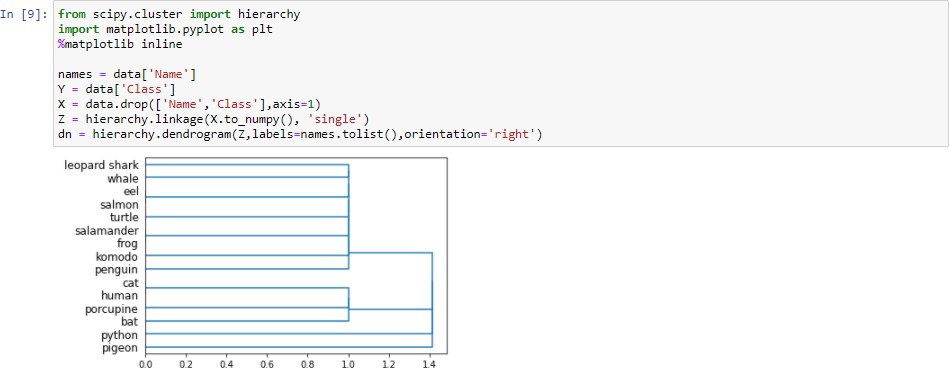
**Performing K-means clustering and Heirarchical clustering and visualize using plots:**



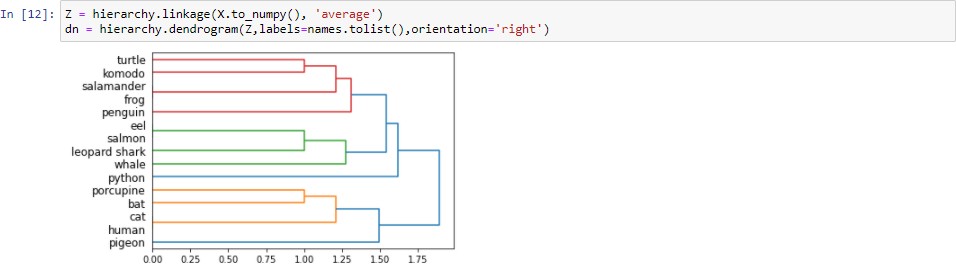




# Table Description automatically generated2)Hierarchical Clustering:



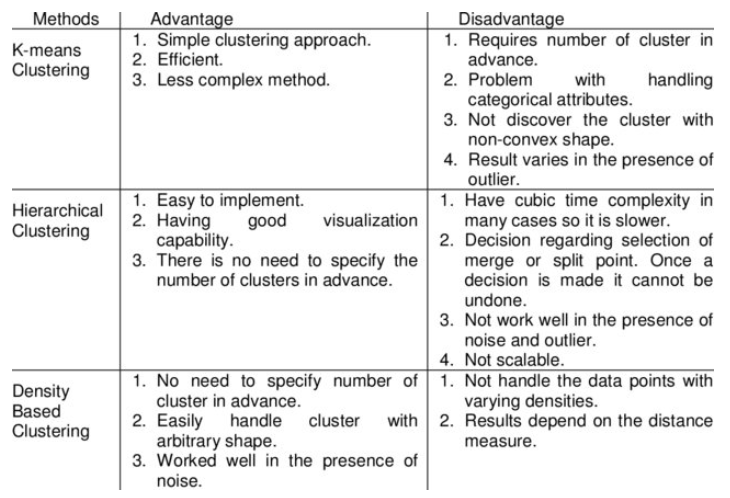
# Chart Description automatically generated



**D) Reference link of web material:**

* <https://www.researchgate.net/publication/258285203_A_survey_on_Data_Mining_approaches_for_Healthcare/figures?lo=1>
* <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>
* <https://www.geeksforgeeks.org/clustering-in-machine-learning/#:~:text=Clustering%20is%20the%20task%20of,data%20points%20in%20other%20groups>.

**E) Advantages/Disadvantages:**



**F) Question/Answers:**

Q: What is difference between K-means and Hierarchical Clustering?

* Hierarchical clustering can’t handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(n2).
* In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
* K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
* K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

Q: What are the applications of clustering?

Clustering has a large no. of applications spread across various domains. Some of the most popular applications of clustering are:

* Recommendation engines
* Market segmentation
* Social network analysis
* Search result grouping
* Medical imaging
* Image segmentation
* Anomaly detection

Q: How is closeness of clusters determined?

There are multiple metrics for deciding the closeness of two clusters :

* Euclidean distance: ||a-b||2 = √(Σ(ai-bi))
* Squared Euclidean distance: ||a-b||22 = Σ((ai-bi)2)
* Manhattan distance: ||a-b||1 = Σ|ai-bi|
* Maximum distance:||a-b||INFINITY = maxi|ai-bi|
* Mahalanobis distance: √((a-b)T S-1 (-b)) {where, s : covariance matrix}

**G) Conclusion:**

Thus, now we know how to evaluate our machine learning model using different performance measures and how to use Google Colab.

|  |  |  |  |
| --- | --- | --- | --- |
| **Prepared By:** | Rajiv Kumar Gupta | **Date:** | 16-04-2021 |