TY B.Tech. (CSE) – II [2020-21]
4CS462: PE-2 - Data Mining Lab.
Assignment No. [07]

Date: 16/10/2021

GroupID:	DM21G07	
Group Members:	2018BTECS00063	Aryan Mali
	2018BTECS00094	Shreya Singh
	2018BTECS00099	Ganesh Kasar

Title: Clustering in ML

Aim: Design and implement the following clustering algorithm:

a) Hierarchical clustering - AGNES & DIANA. Plot Dendrogram.

b) k-Means

c) k-Medoids (PAM)

d) DBSCAN

Introduction:

Clustering or cluster analysis is a machine learning technique, which groups the unlabelled dataset. It can be defined as "A way of grouping the data points into different clusters, consisting of similar data points. The objects with the possible similarities remain in a group that has less or no similarities with another group."

It does it by finding some similar patterns in the unlabelled dataset such as shape, size, colour, behaviour, etc., and divides them as per the presence and absence of those similar patterns.

It is an unsupervised learning method, hence no supervision is provided to the algorithm, and it deals with the unlabeled dataset.

After applying this clustering technique, each cluster or group is provided with a cluster-ID. ML systems can use this id to simplify the processing of large and complex datasets. The clustering technique is commonly used for statistical data analysis.

Theory:

Hierarchical Clustering Algorithms

Hierarchical clustering can be divided into two main types: agglomerative and divisive.

Agglomerative clustering:

It’s also known as AGNES (Agglomerative Nesting). It works in a bottom-up manner. That is, each object is initially considered as a single-element cluster (leaf). At each step of the algorithm, the two clusters that are the most similar are combined into a new bigger cluster (nodes). This procedure is iterated until all points are members of just one single big cluster (root) (see figure below). The result is a tree that can be plotted as a dendrogram.

Divisive hierarchical clustering:

It’s also known as DIANA (Divise Analysis) and it works in a top-down manner. The algorithm is an inverse order of AGNES. It begins with the root, in which all objects are included in a single cluster. At each step of the iteration, the most heterogeneous cluster is divided into two. The process is iterated until all objects are in their own cluster (see figure below).

K-means:

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. … Data points are clustered based on feature similarity.

PAM:

PAM stands for “partition around medoids”. The algorithm is intended to find a sequence of objects called medoids that are centrally located in clusters. Objects that are tentatively defined as medoids are placed into a set S of selected objects.

DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a popular learning method utilized in model building and machine learning algorithms. This is a clustering method that is used in machine learning to separate clusters of high density from clusters of low density.

It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers). The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster.

Procedure:

Use/extend the data analysis tool (menu-driven GUI) developed in Assignment No. 2

to perform the following task :

1. Design and implement the following clustering algorithm:

    a) Hierarchical clustering - AGNES & DIANA. Plot Dendrogram.

    b) k-Means

    c) k-Medoids (PAM)

    d) DBSCAN

2. Tabulate the results with cluster validation accuracy

3. Use the following data sets from the UCI machine learning repository :

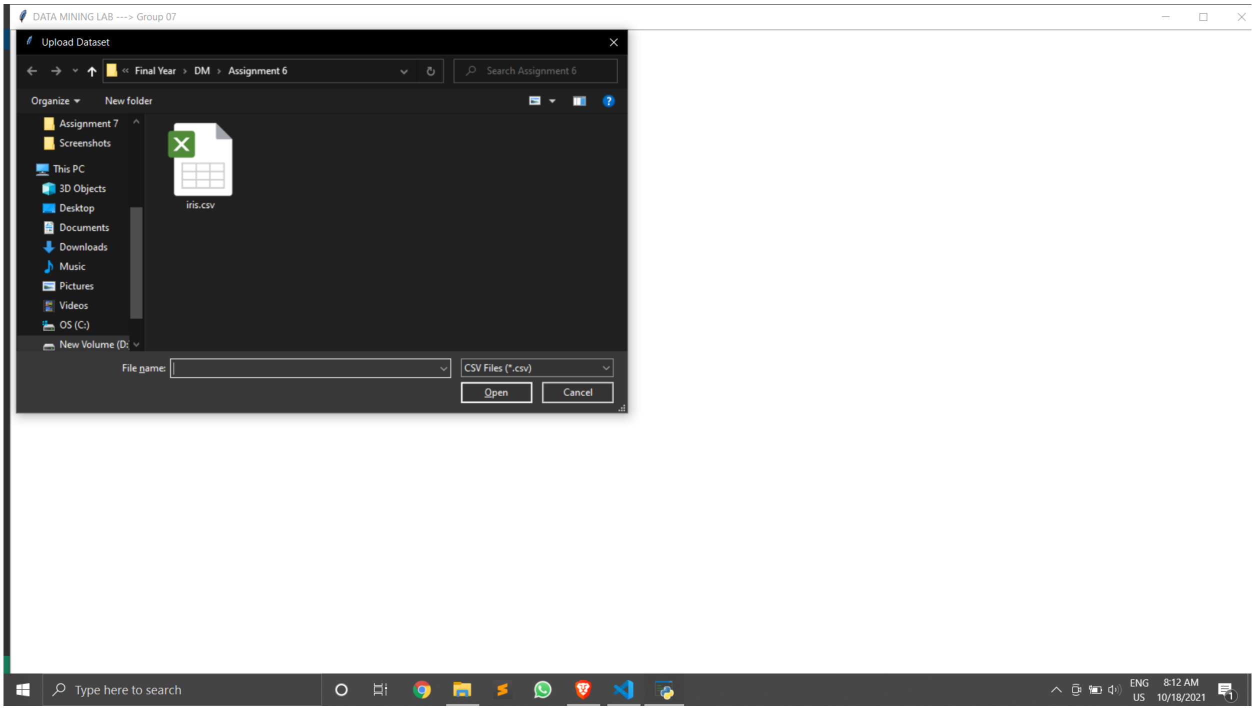
    a) IRIS

    b) Breast Cancer

    c) For DBSCAN, use US Census Data (1990) Data Set

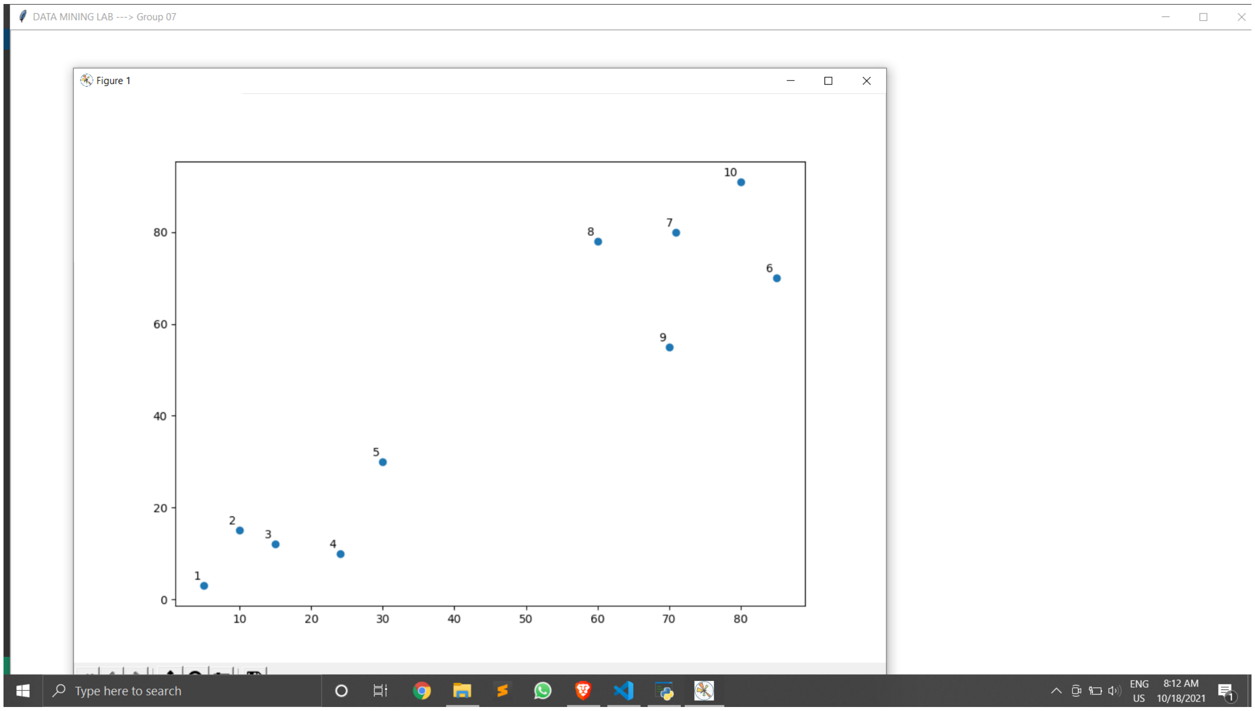
https://archive.ics.uci.edu/ml/datasets/US+Census+Data+(1990)

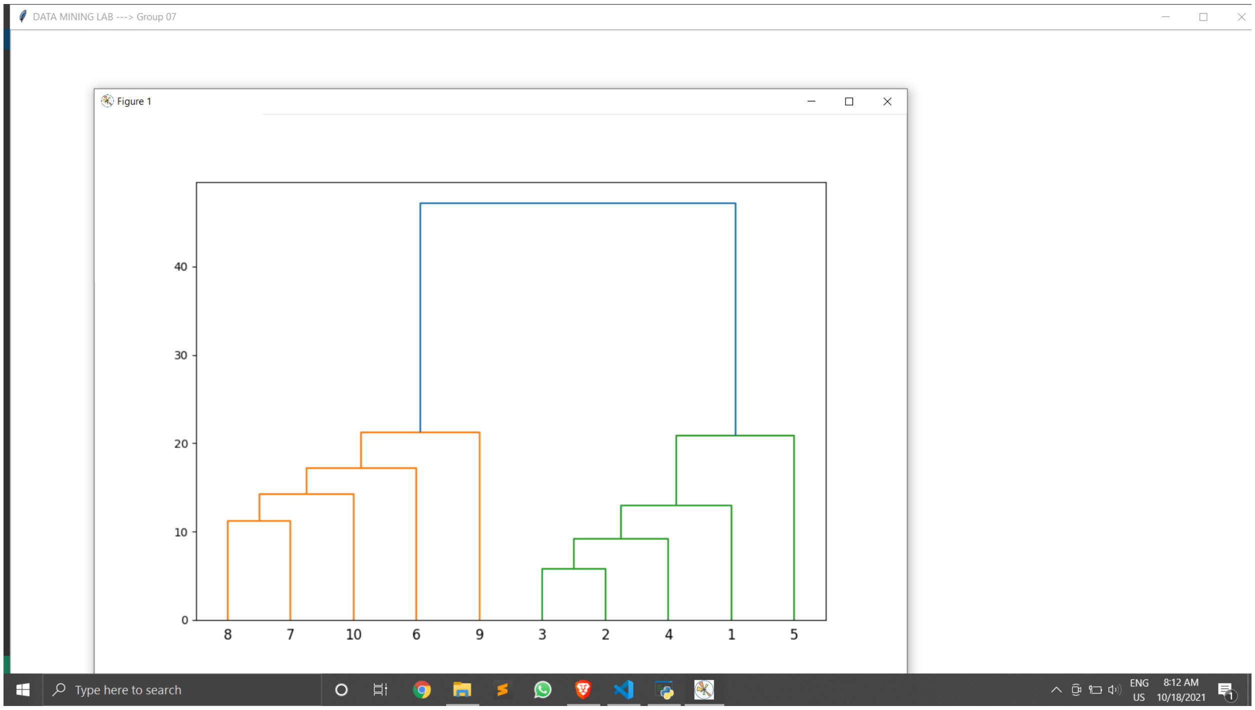
Results:

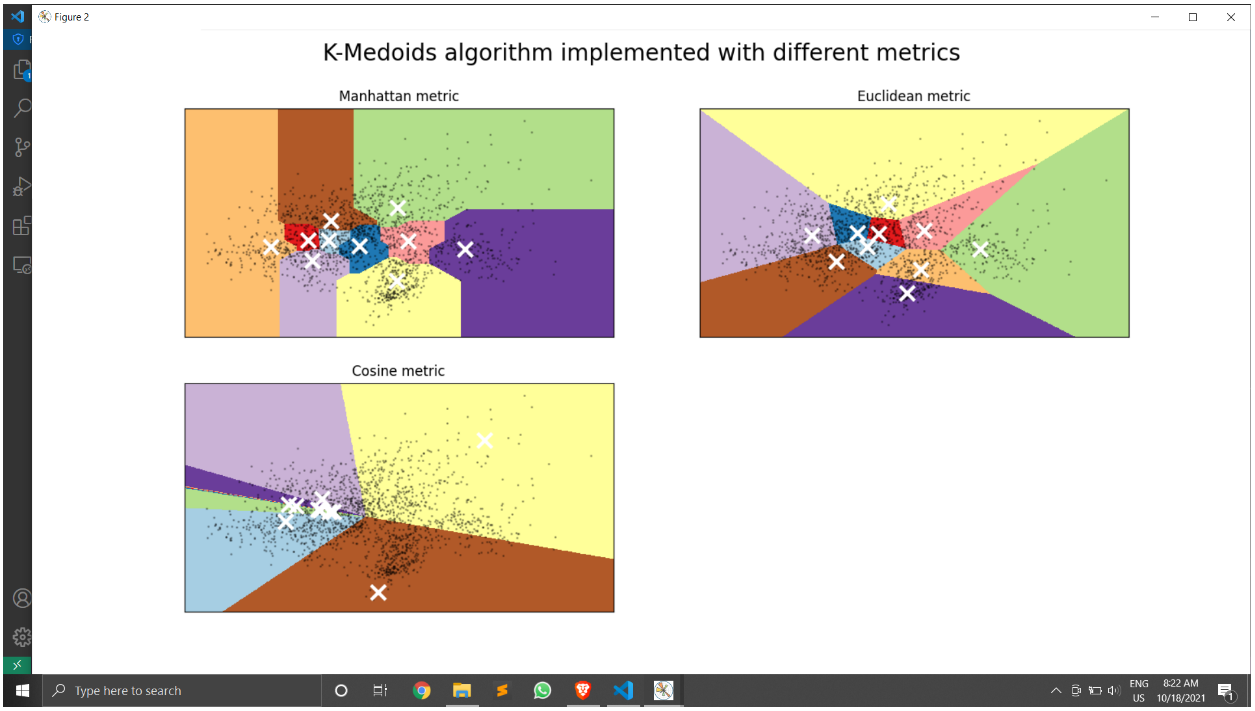


File name:

CSV Files (*.csv)







Conclusion:

Implemented

a) Hierarchical clustering - AGNES & DIANA. Plot Dendrogram.

b) k-Means

c) k-Medoids (PAM)

d) DBSCAN