DMDW LAB – 5

AIM: Demonstrate performing clustering on data sets

THEORY:

Clustering is an unsupervised learning technique used to group data points into clusters or groups based on their similarity or proximity to one another. Unlike supervised learning, where data is labelled, clustering aims to discover hidden patterns or structures within the data. Following are the characteristics of clustering:

* Unsupervised Learning: Clustering is an unsupervised learning method, meaning it doesn't rely on predefined labels. Instead, it identifies groupings based on data similarities.
* Applications: Clustering has a wide range of practical applications, including:
* Customer Segmentation: Grouping customers with similar behavior or preferences for targeted marketing.
* Image Segmentation: Identifying and separating objects or regions in images.
* Anomaly Detection: Detecting unusual or rare data points within a dataset.
* Key Objective: The primary goal of clustering is to maximize the similarity within clusters and minimize the similarity between clusters.
* Choice of Algorithm: The choice of clustering algorithm depends on the dataset's nature and the desired outcomes.

Clustering algorithms:

* K-Means:

K-Means divides data into K clusters by minimizing the variance within each cluster and works well for spherical clusters.

* Hierarchical Clustering (Agglomerative and Divisive):

Hierarchical clustering builds a tree-like structure of clusters, allowing for a flexible number of clusters and visual interpretation.

* DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

DBSCAN identifies clusters based on data density, making it effective at detecting clusters of various shapes and sizes.

* Gaussian Mixture Models (GMM):

GMM models clusters as Gaussian distributions and assigns data points to the most probable cluster, suitable for complex data distributions.

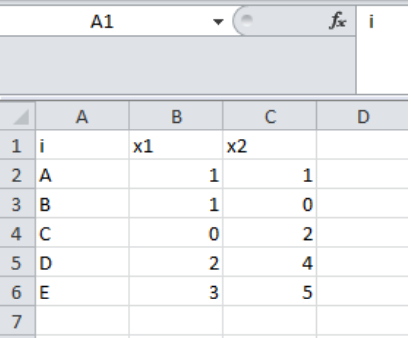
* Self-Organizing Maps (SOM):

Self-Organizing Maps reduce high-dimensional data to a lower-dimensional map, preserving data topology for visualization and understanding.

OUTPUT:

Steps to perform clustering on WEKA:

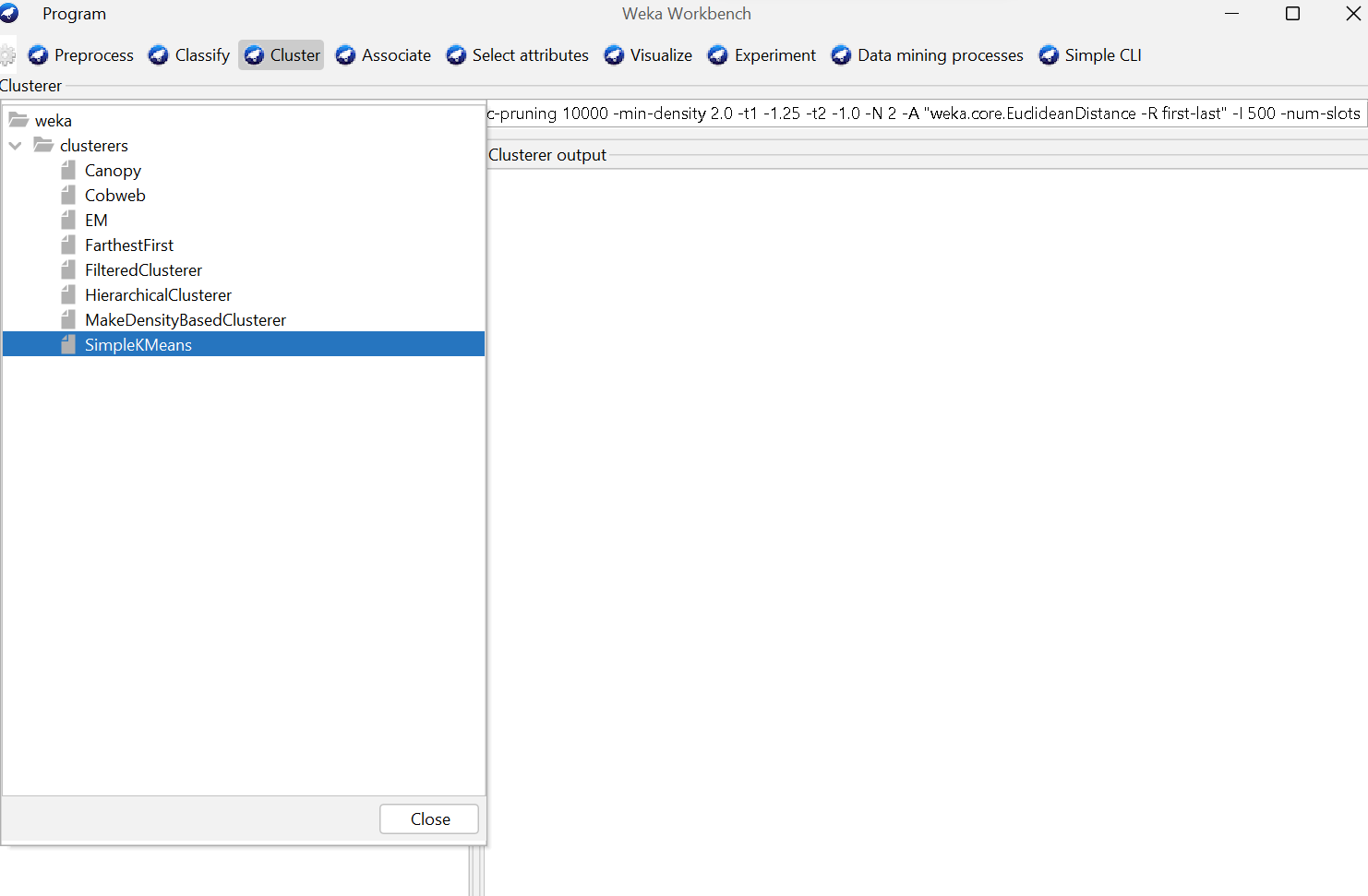
Step1: Create a csv for the data.



Step2: Open weka explorer and then select all the attributes in the table.

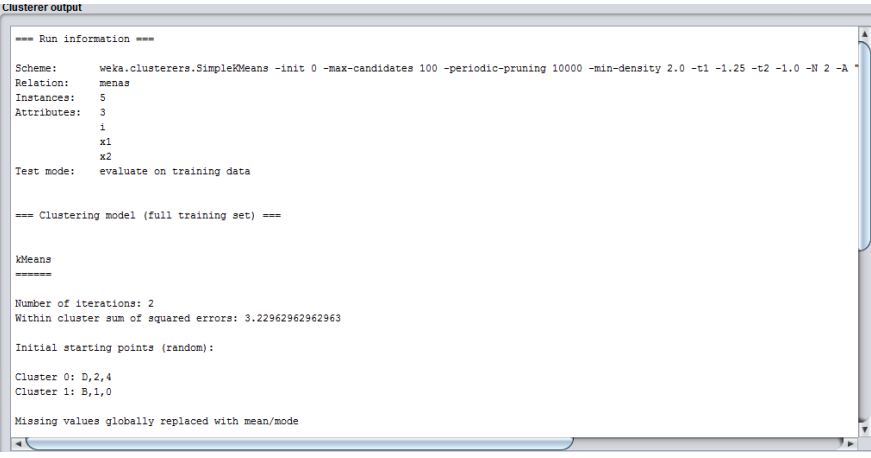


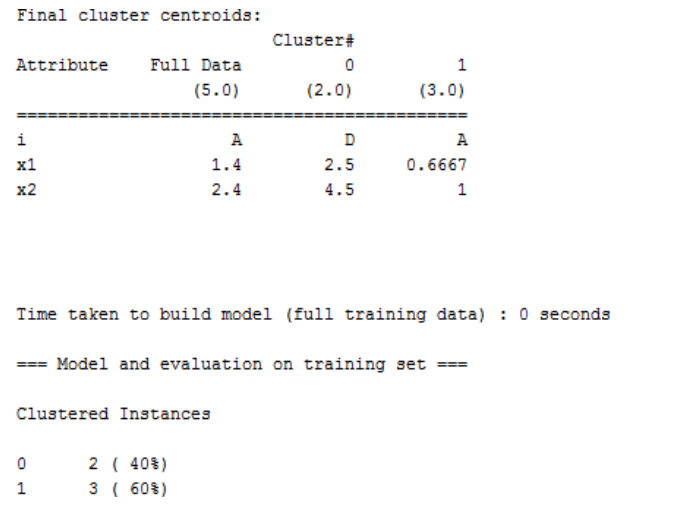
Step3: Open the cluster tab on the WEKA explorer



Here, select which clustering algorithm you want to use.

Step4: Say, we want to use normal K means clustering, it will give the following output





CONCLUSION:

Successfully implemented clustering on a dataset using the WEKA tool.