LAB Manual
PART A
(PART A: TO BE REFFERED BY STUDENTS)

Experiment No.07

A.1 Aim:

Implementation of K-means clustering using JAVA or WEKA.

A.2 Prerequisite:

Familiarity with the WEKA tool and programming languages.

A.3 Outcome:

After successful completion of this experiment students will be able to

Use classification and clustering algorithms of data mining.

A.4 Theory:

THEORY:

K -means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

Algorithmic steps for k-means clustering

Let X = $\{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and V = $\{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select '*c*' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).

PART B

(PART B: TO BE COMPLETED BY STUDENTS)

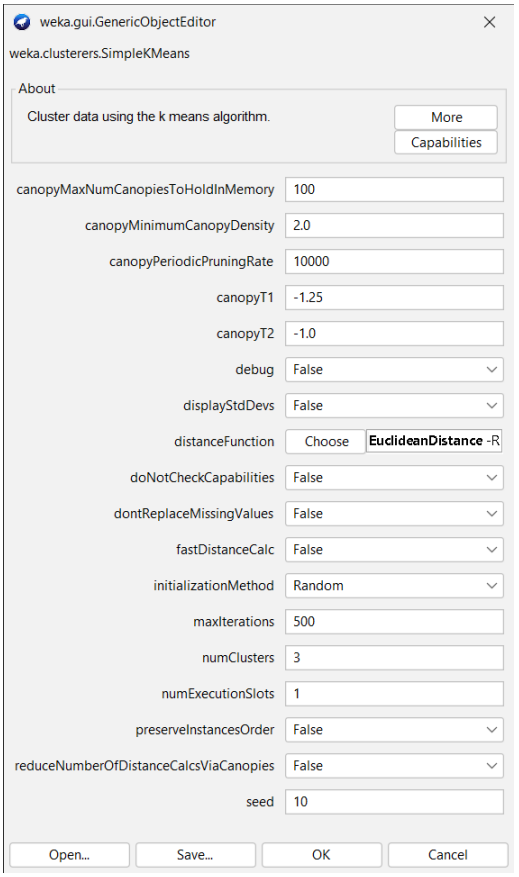
(Students must submit the soft copy as per following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case the there is no Black board access available)

Roll. No. A12	Name: SUFIYAN KHAN
Class: TE-A (AI&DS)	Batch: A1
Date of Experiment:	Date of Submission:
Grade:	

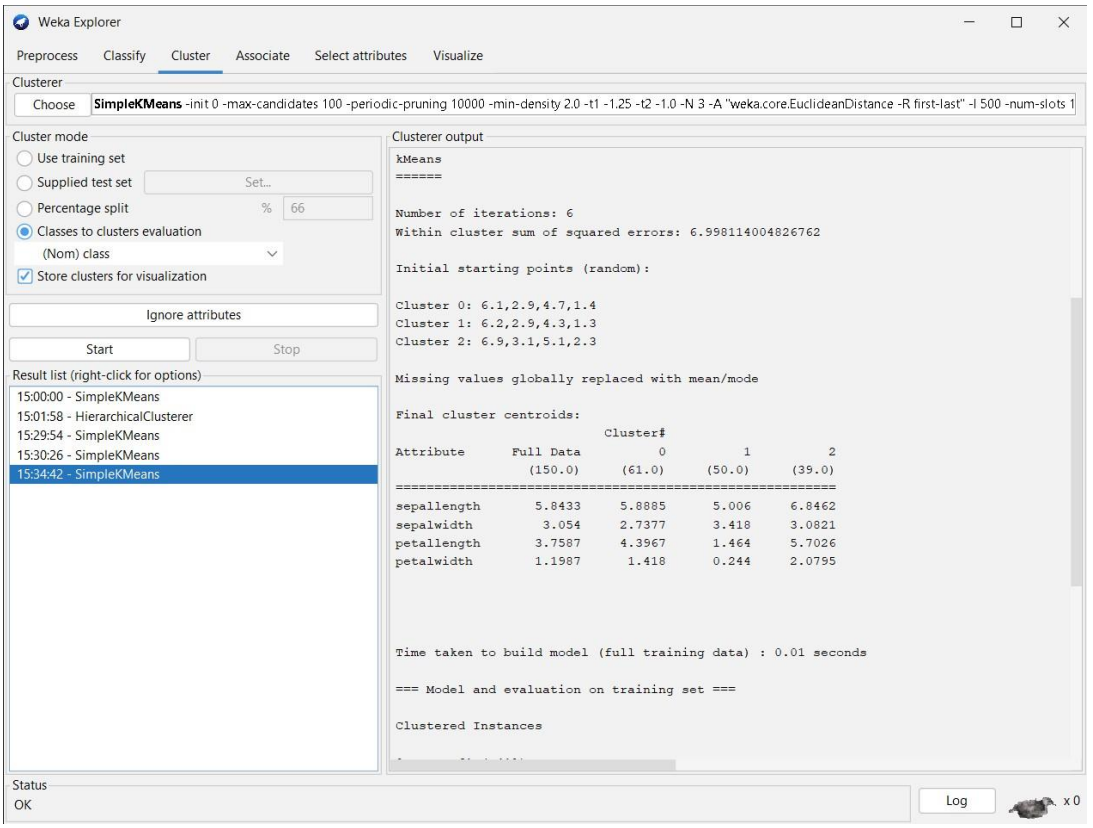
B.1 Software Code written by student:

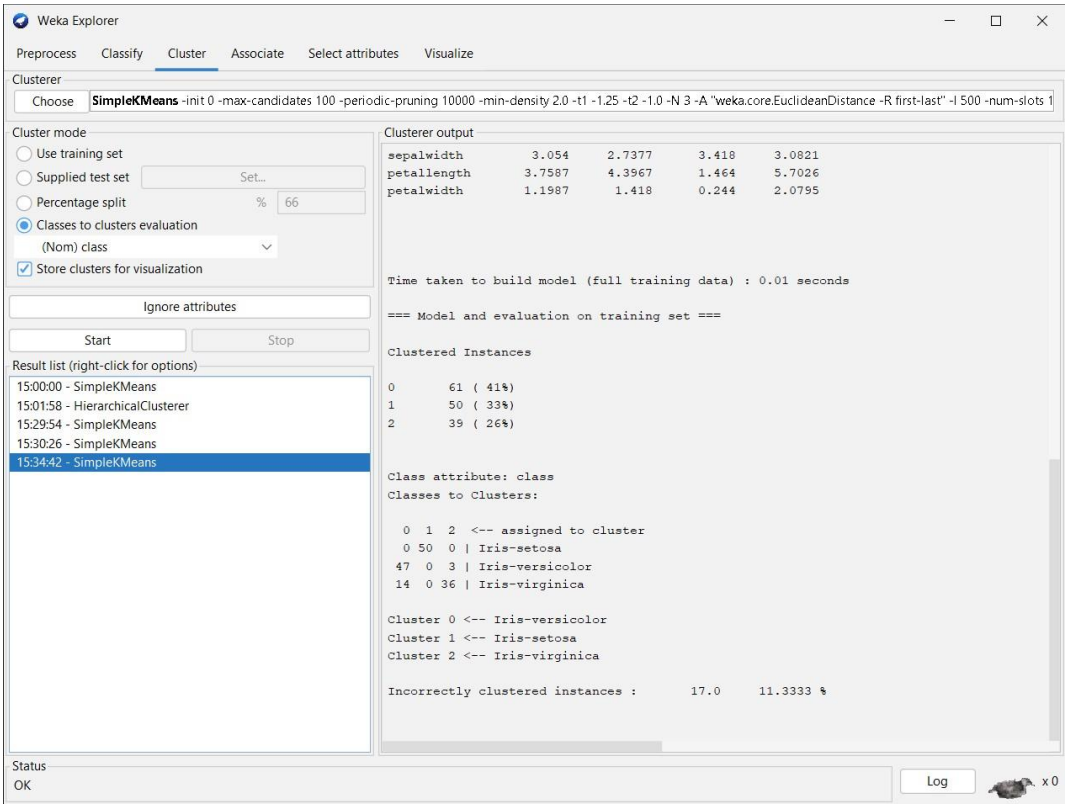
We are using the WEKA tool and iris.arff dataset to demonstrate the K-Means clustering.

After selecting the k-means clusterer we set the number of clusters to 3.



The centroid of each cluster is shown in the result window, along with statistics on the number and percent of instances allocated to each cluster. Each cluster centroid is represented by a mean vector. This cluster can be used to describe a cluster.

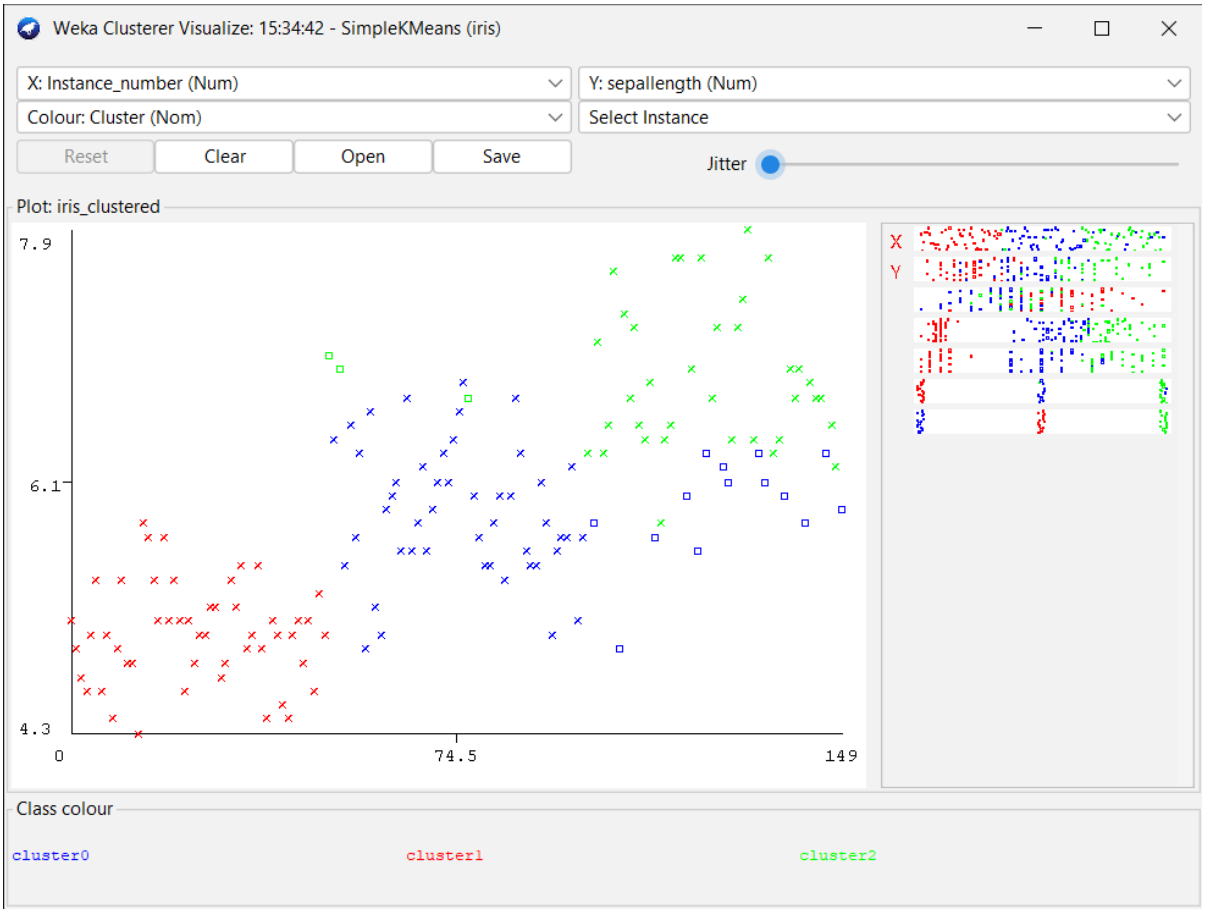




B.2 Input and Output:

Another way to grasp the characteristics of each cluster is to visualize them.

Here is the visualization.



B.3 Observations and learning:

The k-means clustering algorithm is a popular unsupervised machine learning technique used for partitioning a dataset into K distinct, non-overlapping subgroups or clusters.

B.4 Conclusion:

K-means clustering partitions data into K clusters based on proximity to centroids, with sensitivity to initializations, a need for determining optimal K, and an assumption of spherical clusters, making it important to preprocess and interpret results judiciously.

B.5 Question of Curiosity

Q1: What is Clustering? Types of clustering? Explain advantages and disadvantages of clustering.

Clustering is an unsupervised machine learning technique used to group similar data points together in a dataset. The goal is to discover inherent patterns or structures within the data without any predefined labels.

Types of Clustering:

1. K-means Clustering
2. Hierarchical Clustering.
3. DBSCAN.
4. GMM (distribution based).

Advantages of Clustering:

- Pattern Discovery:

Clustering helps identify natural groupings or patterns in data that may not be immediately apparent.

- Data Compression:

By representing clusters with their centroids, data can be effectively compressed.

- Anomaly Detection:

Outliers often don't fit well into any cluster, making them easier to identify.

- Simplifies Complex Data:

Clustering reduces the complexity of large datasets, making them easier to analyze and interpret.

- Facilitates Decision-Making:

Clusters can be used as a basis for making decisions or for further analysis.

Disadvantages of Clustering:

- Sensitivity to Initial Conditions:

Some clustering algorithms are sensitive to the initial placement of centroids, potentially leading to different results for different initializations.

- Determining Optimal K:

Selecting the right number of clusters (K) can be subjective and challenging, and an inappropriate choice can lead to suboptimal results.

- Assumption of Cluster Shape:

Some algorithms, like K-means, assume that clusters are spherical. If clusters have complex shapes, this assumption may not hold.

- Difficulty with Noisy Data:

Outliers or noisy data can significantly impact clustering results, especially in methods like K-means.

- Scalability:

Some clustering algorithms may not scale well to large datasets.

- Lack of Ground Truth:

In unsupervised learning, there's no ground truth to evaluate the quality of the clustering. Evaluation metrics are often heuristic.

Q2: Give the advantages and disadvantages of K- means clustering.

Advantages of K-means Clustering:

- Efficiency:

K-means is computationally efficient and is suitable for large datasets.

- Ease of Implementation:

It's relatively simple to implement and understand, making it a good starting point for clustering analysis.

- Scalability:

It can handle a large number of data points and features.

- Interpretability:

Results are easily interpretable, as each cluster can be described by its centroid.

- Well-Suited for Spherical Clusters:

It works well when clusters are roughly spherical in shape.

Disadvantages of K-means Clustering:

- Sensitive to Initializations:

Results can be highly dependent on the initial placement of centroids, which can lead to suboptimal solutions.

- Dependent on Number of Clusters (K):

Choosing an inappropriate value of K can lead to poor clustering results.

- Assumption of Spherical Clusters:

K-means assumes that clusters are spherical and have roughly equal variance in all dimensions, which may not always be the case in real-world data.

- Vulnerability to Outliers:

Outliers can significantly affect the clustering results, as the algorithm aims to minimize the sum of squares.

- Difficulty with Non-Linearly Separable Data:

It struggles with clusters that are not well-separated or have complex shapes.

- Lack of Robustness to Noise:

Noise or irrelevant features can impact the clustering results.

Q3: How is the number of cluster chosen?

The elbow method is used to determine the number of cluster which is actually the 'k' in k-means clustering.

This method involves running the k-means clustering algorithm on the dataset for a range of k values and calculating the sum of squared distances (inertia or WCSS - Within-Cluster Sum of Squares) for each k.

The "elbow point" in the plot of WCSS against k is the point where the rate of change of WCSS starts to slow down significantly. This can indicate a good value for k.