Assignment-6

* Aim:- Demonstration of clustering Rule process on the data-set iris.arff using simple k means,

* Course Outcomes:-

CO 1:- Demostrate the classification, clustering and etc in the Large data sets.

CO 3:- Ability to add mining algorithms as an component to the exiting tools.

* Software & Hardware Requirements:-

Sr No	Requirements	Software	Hardware
1	Operating System	Windows 11	2GHz Processor 8GB RAM
2	Weka	version -V 3.7.6	

* Theory:-

* Clustering in the Data Mining:-
Clustering refers to the process of the grouping the similar data together based on

certain features or the attributes. this process is often used to discover hidden patterns or the structures within the data sets and is an essential technique in the exploratory data analysis, pattern Recognition and data mining tasks.

→ The main objective of the clustering is to partition a datasets into the subsets, known as the clusters, such data points within the same cluster are more similar to each other than to those in the other clusters.

→ Classification of the clustering :-

1) Un Supervised Learning :-

→ Un supervised learning is an type of the Learning where the algorithm is trained on the unlabelled data.

→ In the Clustering, the algorithm aims to discover inherent patterns or the structures within the data without the prior knowledge of the group memberships.

2) Grouping Similar data points :-

→ Clustering algorithm analyse the features of data points and group them into clusters based on their similarities.

→ Similarity is typically measured using the distance metric, such as Euclidean distance or the cosine similarity which similarities the dissimilarity between data points.

3) Partitioning the datasets:-
→ The goal of the clustering is to divide the dataset into the distinct partitions or the subsets, called clusters, where each cluster represent a group of the data points that share similar characteristics.

4) Characteristics of clusters:-
→ Clusters may vary in the size, shape and density depending on the distribution of the data points in the feature space.
→ Clusters should exhibits high extra intra-cluster similarity and low inter-cluster similarity.

5) Applications of the clustering:-
1) Customer Segmentation in marketing:-
Grouping customers based on their purchasing behavior or the demographics

2) Image Segmentation in the Computer vision:-
→ Partitioning an Image into the regions with the similar pixel characteristics

3) Document clustering in the natural language Processing:-
→ Organizing text documents into the topic-Based clusters for the analysis or the Retrival.

* K- Means Clustering algorithm:-

→ K means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in the machine learning or the data Science.

→ K means Clustering is an unsupervised learning algorithm which groups the unlabelled dataset into the different clusters. Here K defines the number of pre-defined clusters that need to be created in the process as if K=2, there will be two clusters, and for K=3, there will be three clusters and so on.

→ It allows us to cluster the data into the different groups and convenient way to discover the categories of the groups in the unlabelled dataset as it own without the need for any thing.

→ It is an centroid - Based algorithm, where the clusters is associated with the centroid. The main aim of the algorithm is to classify and minimizes the sum of the distances between the data point and their corresponding clusters.

→ This algorithm takes an unlabelled dataset into the input, divides the dataset into the K- numbers of the clusters, and repeat the process untill it does not find the best clusters. The value of the K should be predetermined in this algorithm.

The K-means Clustering algorithm mainly performs two tasks:-

- Determines the best value for k center points or the centroids by the Iterative processes
- Assigns each data points to its closest k-centers Those data points where are near to the particulas k-center, create a cluster.

* Working of the k-means Algorithm:-

Step 1:- Select the number k to decide the number of the clusters
Step 2:- Select the Random k points or the centroids.
Step 3:- Assign each data points to their closest centroid which will form the predefined k-clusters.
Step 4:- Calculate the variance and place a new centroid of the each clusters.
Step 5:- Repeat the third step, which means assign each data point to the new closet centroid of the each cluster.
Step 6:- If any reassignment occurs, then go to the step 4, else go to finish.
Step 7:- The model is ready

6(60

* Iris Arff (Attribute relation file format)

→ The Iris dataset is an popular dataset in the machine learning. It contain the 150 samples of the iris flowers, each with four features: Sepal length, sepal width, pedal length and pedal width. The dataset is often used for the classification tasks.

→ The Dataset is avaiable in the form of the ARFF, which stands for the attribute relation file format, commonly used for the weka machine learning software.

→ The Iris dataset consists of the 150 instances where each Instance Represents the flowers. There are four attributes or features for each instance:-

1) Sepal length (In centrimeters)
2) Sepal width (In centrimeters)
3) Petal length (In centimeters)
4) Petal width (In centrimeters)

→ Each Instance is also labelled with one of three classes, representing the species of the iris

1) Iris - setosa
2) Iris - vericolor
3) Iris - vigrinica

→ The Dataset is well known in the machine learning for its simplicity and the effective ness in the demonstrating various algorithm

→ The Iris dataset is commonly used for the supervised learning tasks, particularly classification.
→ It serves as an foundamental datasets for the learning and testing classification algorith
→ Due to its small size, simplicity and well defined classes, if often used for the educational purposes and as a benchmark dataset for evaluating new algorithms.

* Where it is used?
→ The Iris dataset is used for the clustering and visualization, especially for the technique like principal component analysis (PCA) to reduce dimensionality and visualize the data in the lower dimensions.

* Conclusion:-
→ Through this practical, we obversed, the application of the K-means clustering algorithm on the iris.arff dataset provided credential sights into the underlying structure of the data. By grouping similar data points into the clusters K-means faciliated a deeper understanding the datasets patterns and Relationships.