EXPERIMENT NO: 5

Aim: To implement K-Means algorithm.

Description:

Clustering: Clustering is the method of dividing a set of abstract objects into groups. Points to Keep in Mind A set of data objects can be viewed as a single entity. When performing cluster analysis, we divide the data set into groups based on data similarity, then assign labels to the groups.

Simple-k means clustering: K-means clustering is a simple unsupervised learning algorithm. In this, the data objects ('n') are grouped into a total of 'k' clusters, with each observation belonging to the cluster with the closest mean. It defines 'k' sets, one for each cluster k n (the point can be thought of as the center of a one or two-dimensional figure). The clusters are separated by a large distance.

The data is then organized into acceptable data sets and linked to the nearest collection. If no data is pending, the first stage is more difficult to complete; in this case, an early grouping is performed. The 'k' new set must be recalculated as the barycenters of the clusters from the previous stage.

The same data set points and the nearest new sets are bound together after these 'k' new sets have been created. After that, a loop is created. The 'k' sets change their position step by step until no further changes are made as a result of this loop.

K-means clustering algorithm computes the centroids and iterates until we it finds optimal centroid. It assumes that the number of clusters are already known. It is also called **flat clustering** algorithm. The number of clusters identified from data by algorithm is represented by 'K' in K-means.

In this algorithm, the data points are assigned to a cluster in such a manner that the sum of the squared distance between the data points and centroid would be minimum. It is to be understood that less variation within the clusters will lead to more similar data points within same cluster.

Working of K-Means Algorithm

We can understand the working of K-Means clustering algorithm with the help of following steps −

Step 1 − First, we need to specify the number of clusters, K, need to be generated by this algorithm.

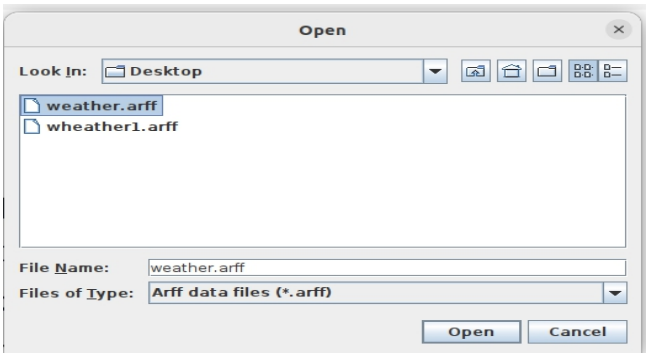
Step 2 − Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.

Step 3 − Now it will compute the cluster centroids.

Step 4 − Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more −

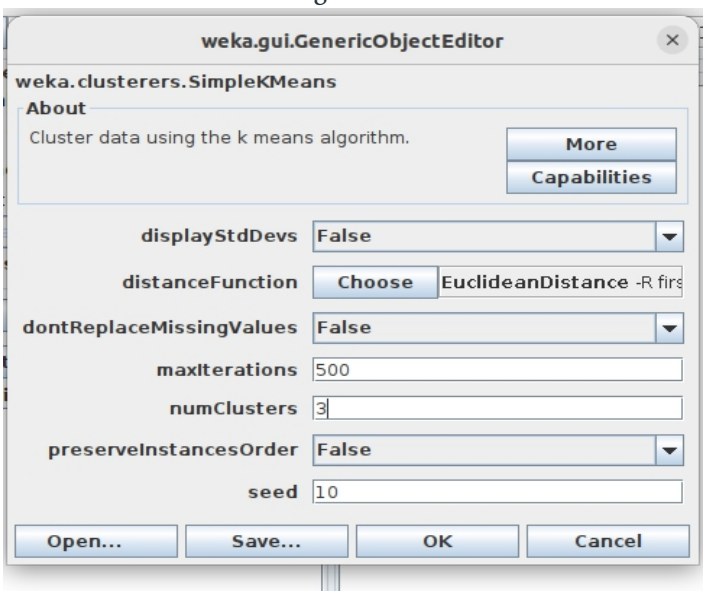
Procedure:

1. **Step 1:** In the preprocessing interface, open the Weka Explorer and load the required dataset, and we are taking the weather.arff dataset.

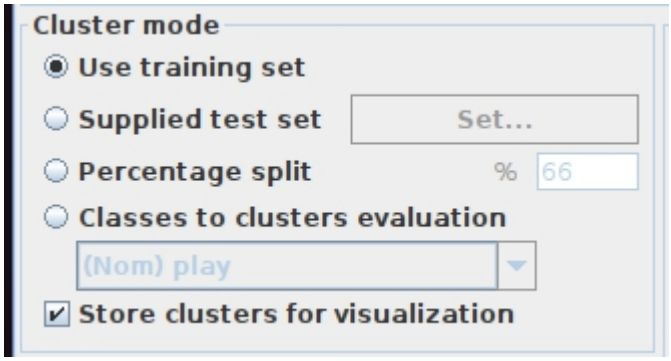


2. **Step 2:** Find the 'cluster' tab in the explorer and press the choose button to execute clustering. A dropdown list of available clustering algorithms appears as a result of this step and selects the simple-k means algorithm.

3. **Step 3:** Then, to the right of the choose icon, press the text button to bring up the popup window shown in the screenshots. We enter three for the number of clusters in this window and leave the seed value alone. The seed value is used to generate a random number that is used to make internal assignments of instances of clusters.



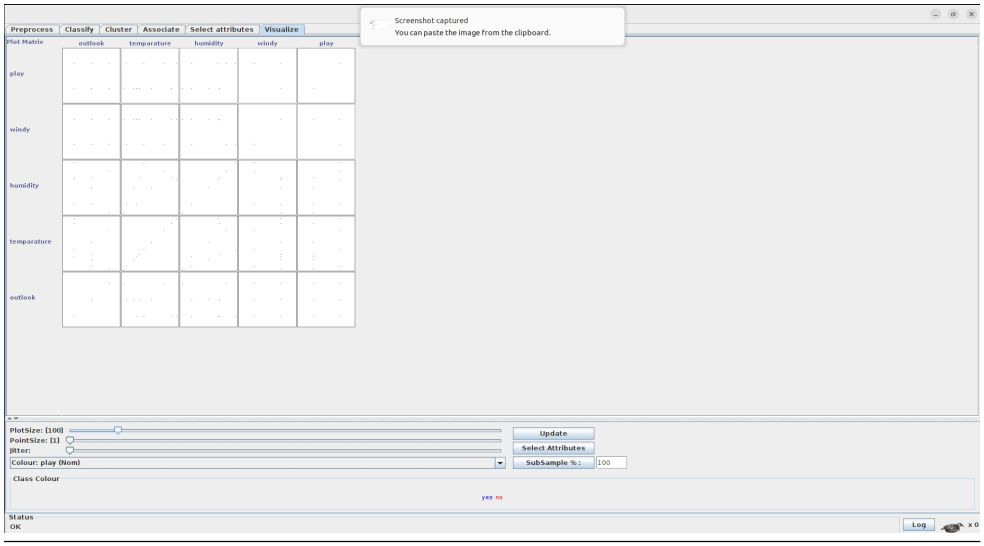
4. **Step 4:** One of the choices has been chosen. We must ensure that they are in the 'cluster mode' panel before running the clustering algorithm. The choice to use a training set is selected, and then the 'start' button is pressed. The screenshots below display the process and the resulting window.



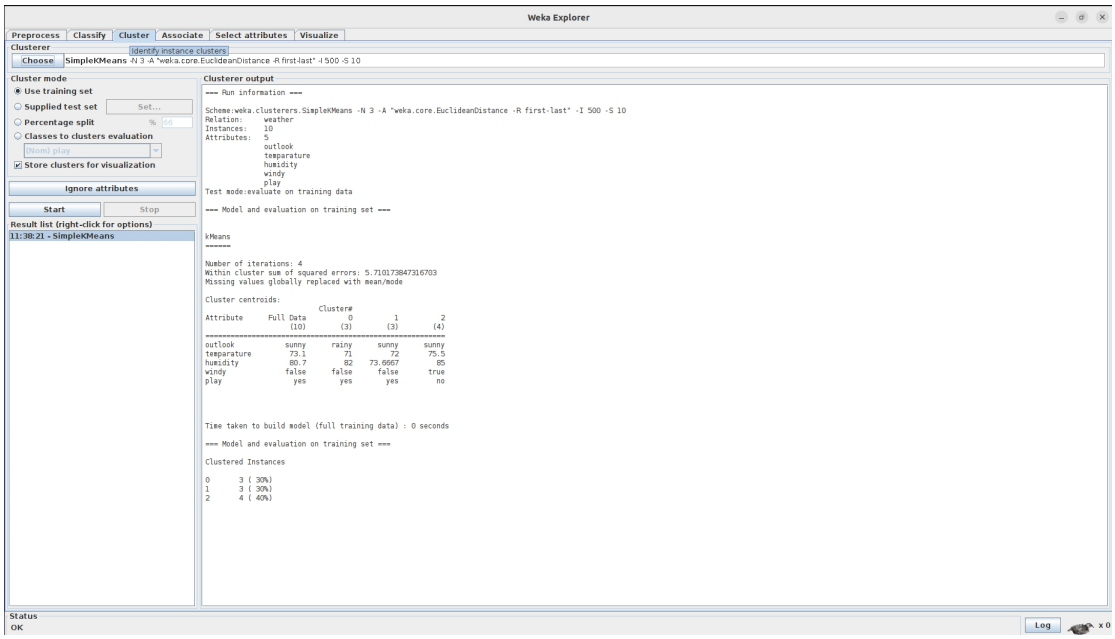
5. **Step 5:** The centroid of each cluster is shown in the result window, along with statistics on the number and percent of instances allocated to each cluster. Each cluster centroid is represented by a mean vector. This cluster can be used to describe a cluster.

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 5.710173847316703
Missing values globally replaced with mean/mode

Cluster centroids:
                         Cluster#
Attribute      Full Data         0          1          2
                    (10)       (3)        (3)        (4)
=============================================================
outlook            sunny      rainy      sunny      sunny
temparature         73.1         71         72       75.5
humidity            80.7         82    73.6667         85
windy              false      false      false       true
play                 yes        yes        yes         no



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       3 ( 30%)
1       3 ( 30%)
2       4 ( 40%)
```

6. **Step 6:** Another way to grasp the characteristics of each cluster is to visualize them. To do so, right-click the result set on the result. Selecting to visualize cluster assignments from the list column.



Output:



Result: This program has been successfully executed.

Viva Questions.

Q.1) What is the purpose of k-means algorithm?
Q.2) How do we use K means clustering algorithm in Weka?
Q.3) What does K mean in k-means algorithm?
Q.4) What type of algorithm is k-means?