Aim: To implement k-means Algorithm on large dataset using Open source tool WEKA.

Objective: To make students well versed with open source tool like WEKA to implement k-means algorithm.

Theory:

- The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.
- A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters.
- A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression.
- Cluster analysis is an important human activity. Cluster analysis has been widely used in numerous applications including market research, pattern recognition, data analysis and image processing.
- Clustering is also called data segmentation in some applications because clustering partitions large data sets into groups according to their similarity.
- Clustering can also be used in outlier detection where outliers may be more interesting than common case.

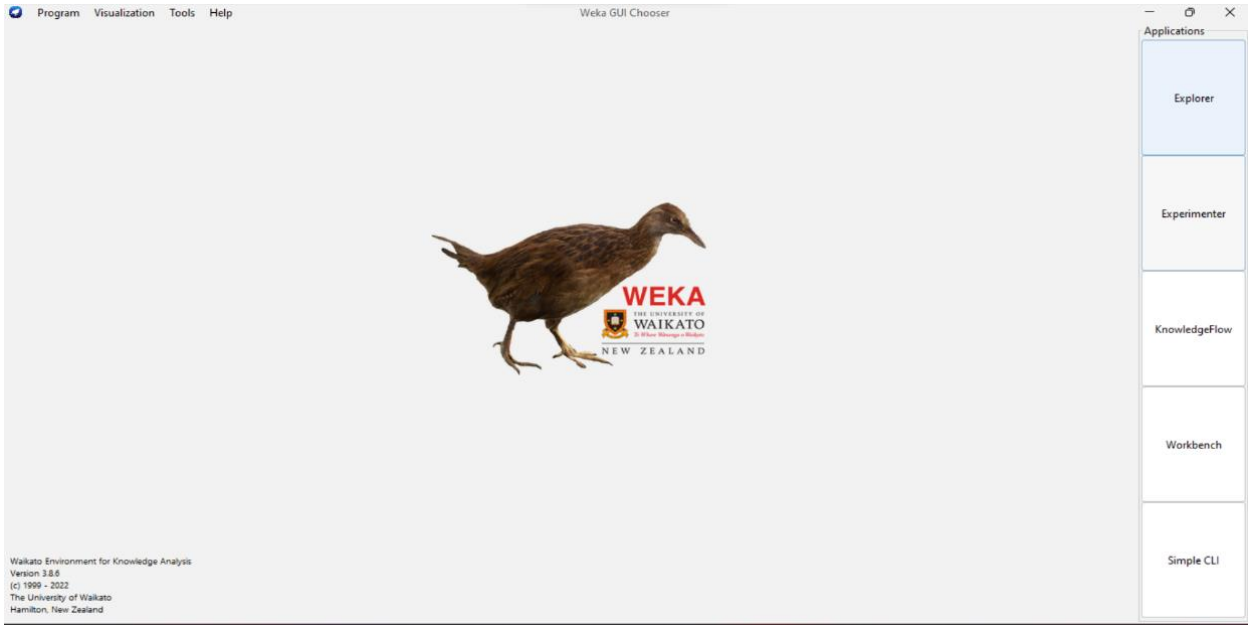
WEKA contains "clusterers" for finding groups of similar instances in a dataset. The clustering schemes available in WEKA are k-means, Cobwebs, DBSCAN, OPTICS. Clusters can be visualized and compared to true clusters. Evaluation is based on log likelihood if clustering scheme produces a probability distribution.In 'preprocess' window click on 'open file...' button to select data file. Choosing Clustering scheme: In the 'clusterer' box click on 'choose' button. In pull-down menu select WEKA Clusteres, and select the cluster scheme 'simple K means'. Some implementations of K -means only allow numerical values for attributes ; therefore we do not need to use a filter.

Once the clustering algorithm is chosen, right click on algorithm, 'weak.gui.GenericObjectEditor' comes up to the screen. Set the value in 'numclusters' box to number of clusters required. The seed value is used in generating a random number, which is used for making the initial assignments of instances to clusters. Before we run the clustering algorithm, we need to select 'cluster mode'. A new entry appears in the 'Result list' box on the left of the result. Run information gives the information about : the clustering scheme used, the relation name, the number of instances, number of attributes. The clustering model shows the centroid of each cluster and statistics on the number and percentage of instances assigned to different clusters. Cluster centroid is the mean vector of each cluster so each dimension value and centroid represents mean value for that dimension in the cluster. Thus centroids can be used to characterize the cluster.

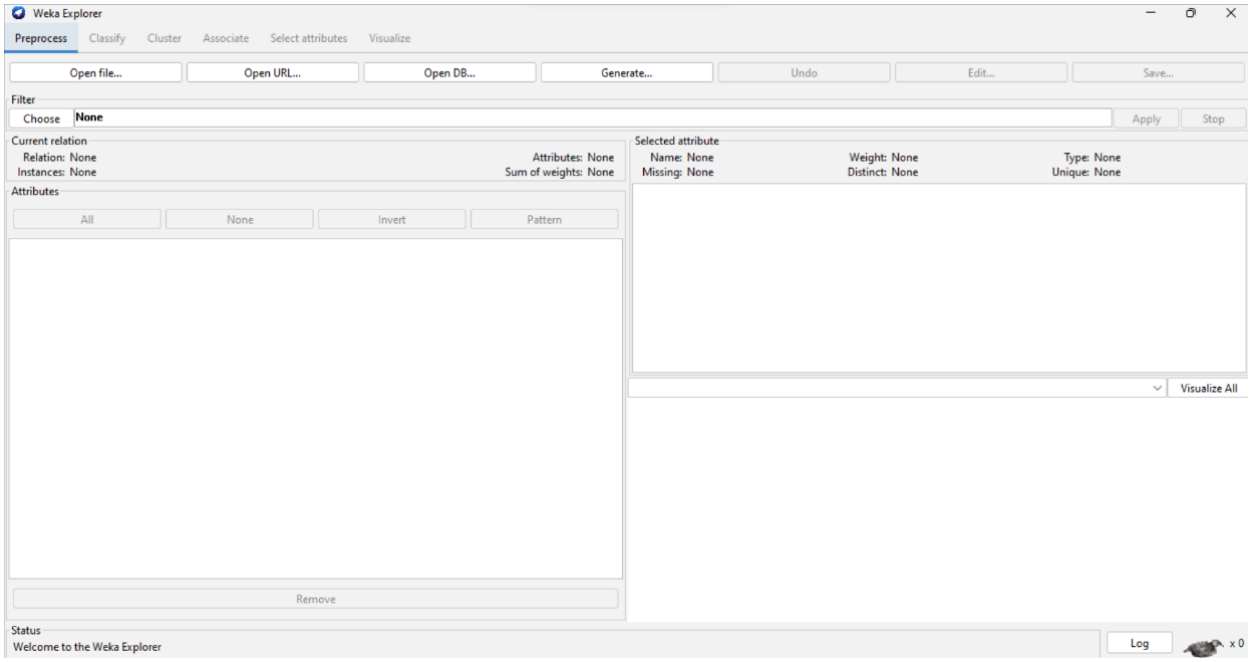
Another way of representation of results of clustering is through visualization. Right click on the entry in the 'Result list' and select ' Visualize cluster assignments' in the pull-down window. This brings up Weka clusterer visualize window. This window displays clusters in different colors for better visibility.

Output:

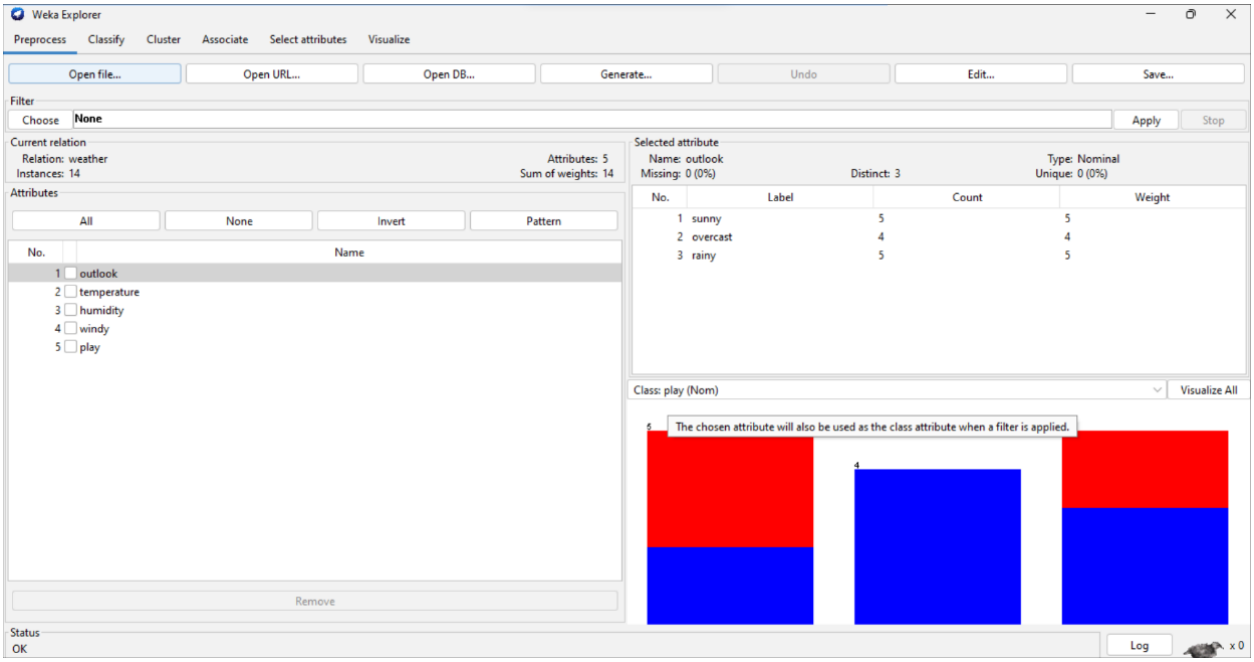
Step 1:Open WEKA, the following GUI should appear on your system.



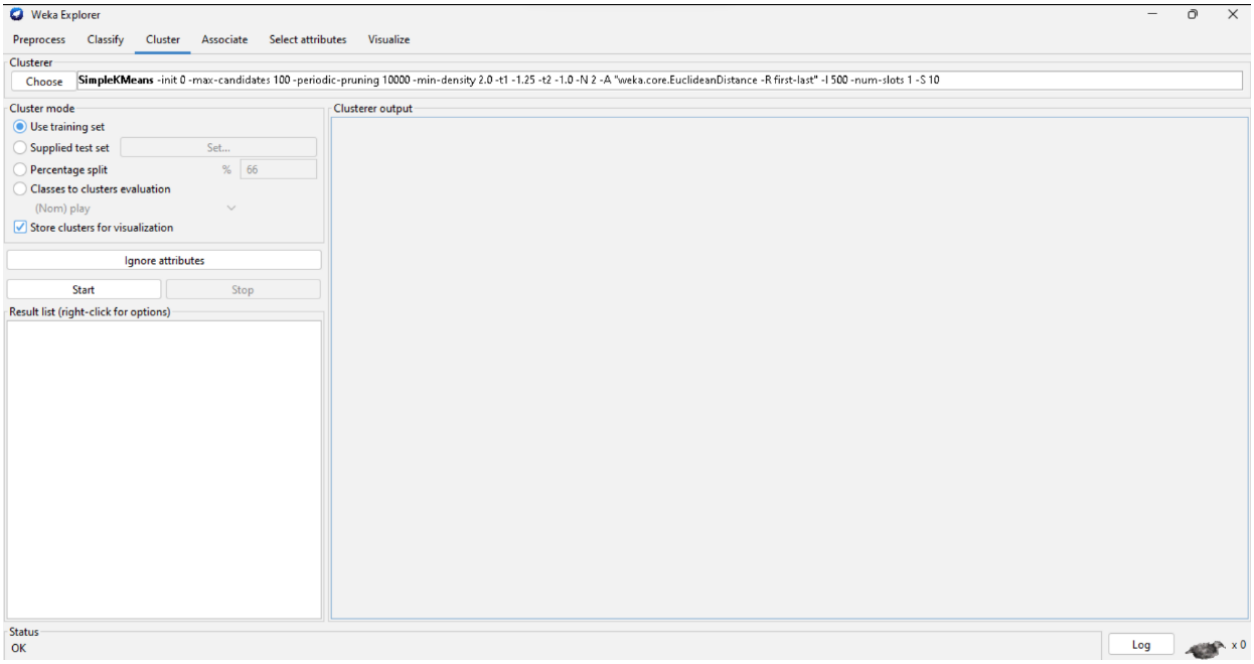
Step 2: Click on the explorer.



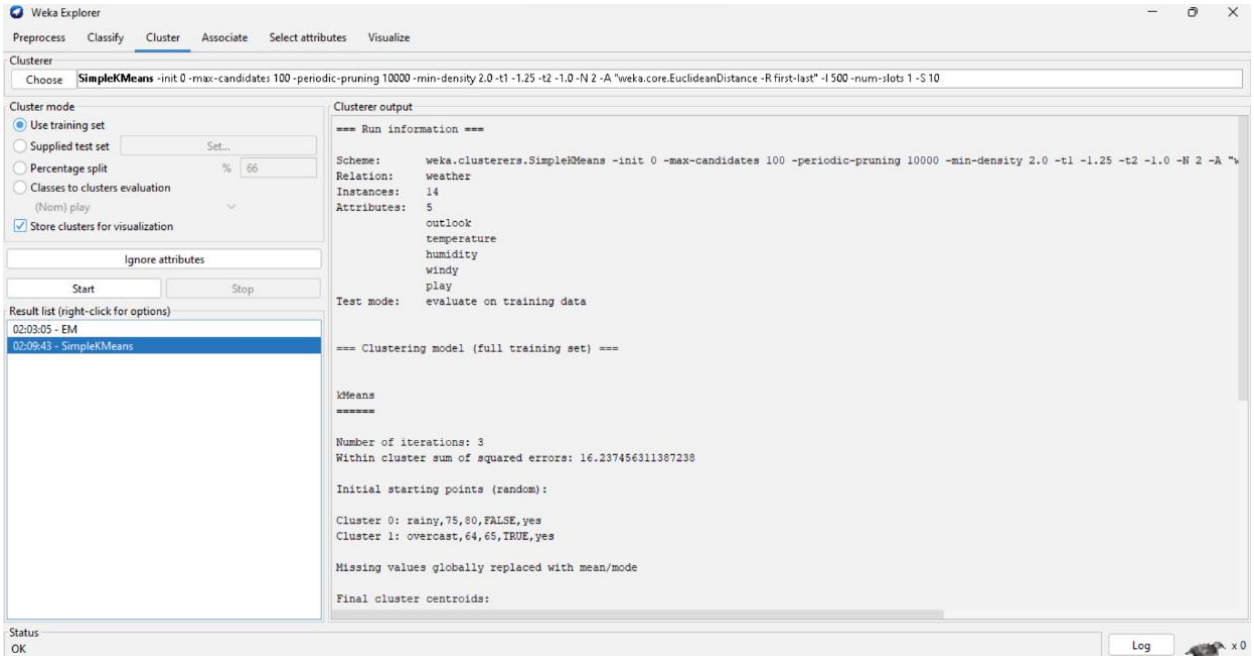
Step 3: Select 'weather.numeric.arff' dataset which already exists in the program files, from the 'Open File' section. The following screen should appear.

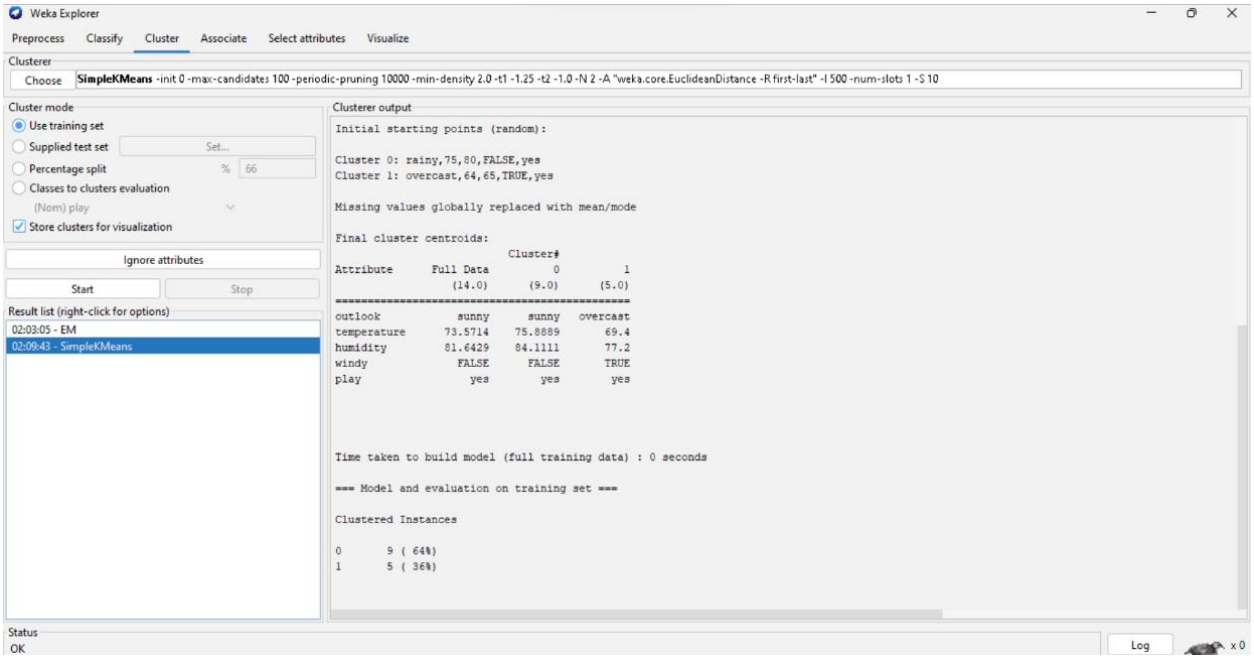


Step 4: Select 'simpleKmeans', under cluster->Choose->simpleKmeans and click on 'Start'.



Step 5: Click on 'Start' , the following output will appear.





Conclusion:

Thus, we have learned to implement k-means Algorithm on large dataset using Open source tool WEKA. WEKA contains "clusterers" such as k-means, Cobwebs, DBSCAN, OPTICS for finding groups of similar instances in a dataset.