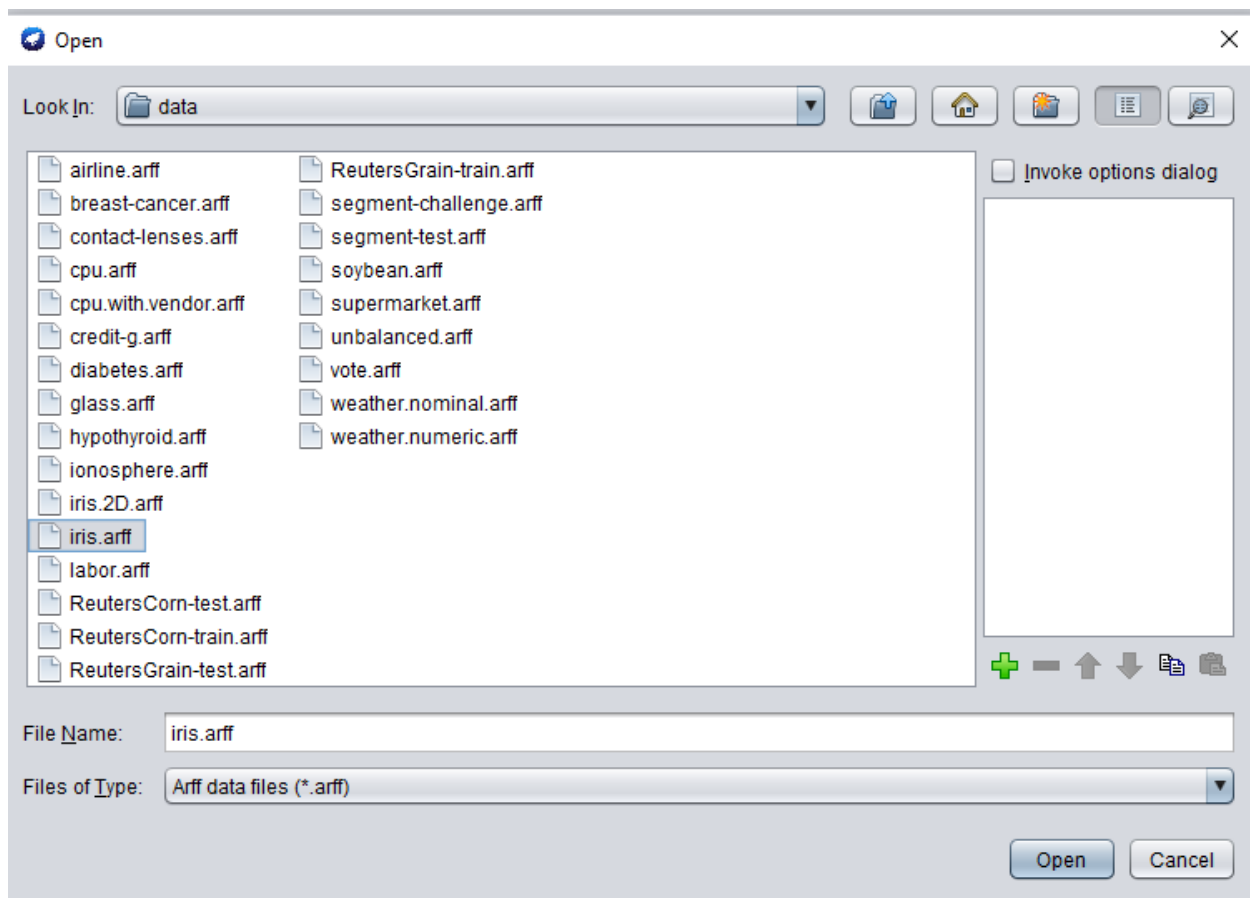


Week 8: Demonstrate performing clustering of data sets. Load each dataset into Weka and run simple k-means clustering algorithm with different values of k (number of desired clusters). Study the clusters formed. Observe the sum of squared errors and centroids, and derive insights. Explore other clustering techniques available in Weka. Explore visualization features of Weka to visualize the clusters. Derive interesting insights and explain.

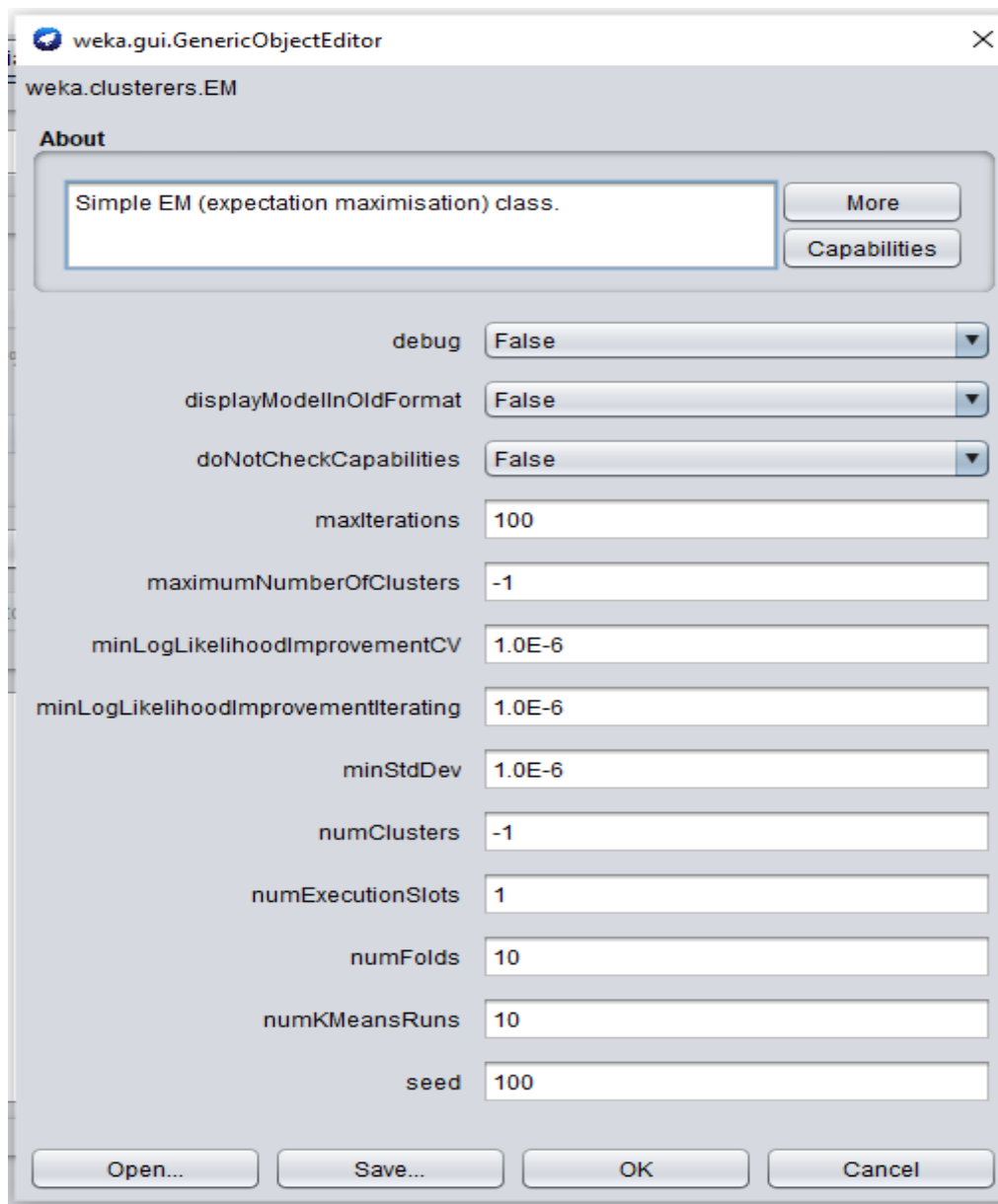
Steps to be followed:

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



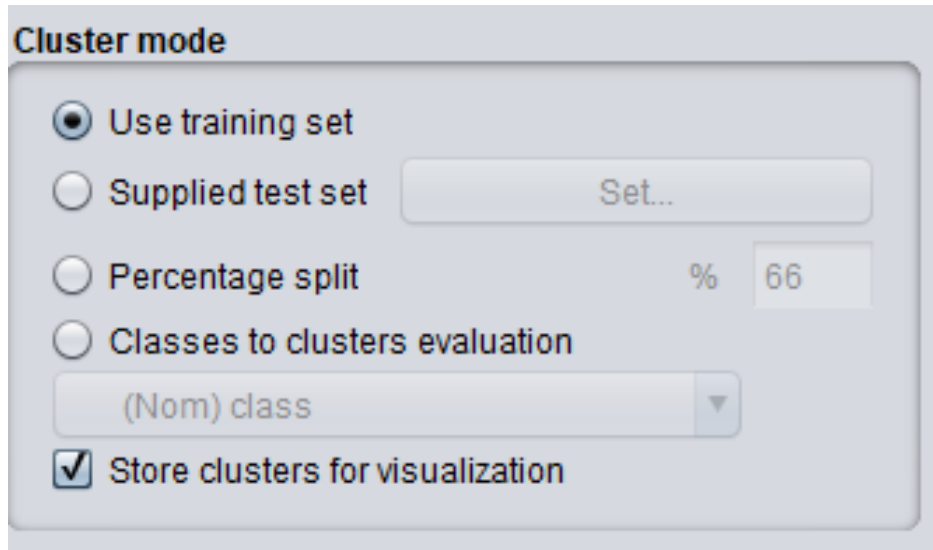
Step 2: Find the 'cluster' tab in the explorer and press the choose button to execute clustering. A dropdown list of available clustering algorithms appears as a result of this step and selects the simple-k means algorithm.

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



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.



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.

Number of clusters selected by cross validation: 4
Number of iterations performed: 16

Attribute	Cluster			
	0 (0.32)	1 (0.33)	2 (0.2)	3 (0.14)
=====				
sepal.length				
mean	5.897	5.006	6.9426	6.1304
std. dev.	0.5279	0.3489	0.498	0.2943
sepal.width				
mean	2.7519	3.418	3.1103	2.8088
std. dev.	0.3103	0.3772	0.2952	0.2361
petal.length				
mean	4.2267	1.464	5.8559	5.0993
std. dev.	0.445	0.1718	0.4626	0.2462
petal.width				
mean	1.3134	0.244	2.1495	1.8254
std. dev.	0.1864	0.1061	0.232	0.2152
class				
Iris-setosa	1	51	1	1
Iris-versicolor	48.1125	1	1.0182	3.8693
Iris-virginica	2.0983	1	31.0375	19.8641
[total]	51.2108	53	33.0557	24.7335

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.

