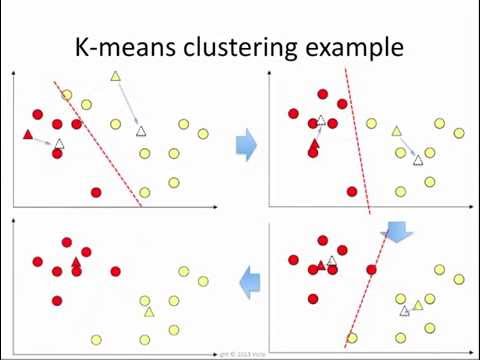
**Experiment No**

**Aim:** To implement k means using WEKA

**Theory:**

k-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



**K means Algorithm:**

* Step 1 - Pick K random points as cluster centers called centroids.
* Step 2 - Assign each xixi to nearest cluster by calculating its distance to each centroid.
* Step 3 - Find new cluster center by taking the average of the assigned points.
* Step 4 - Repeat Step 2 and 3 until none of the cluster assignments change.

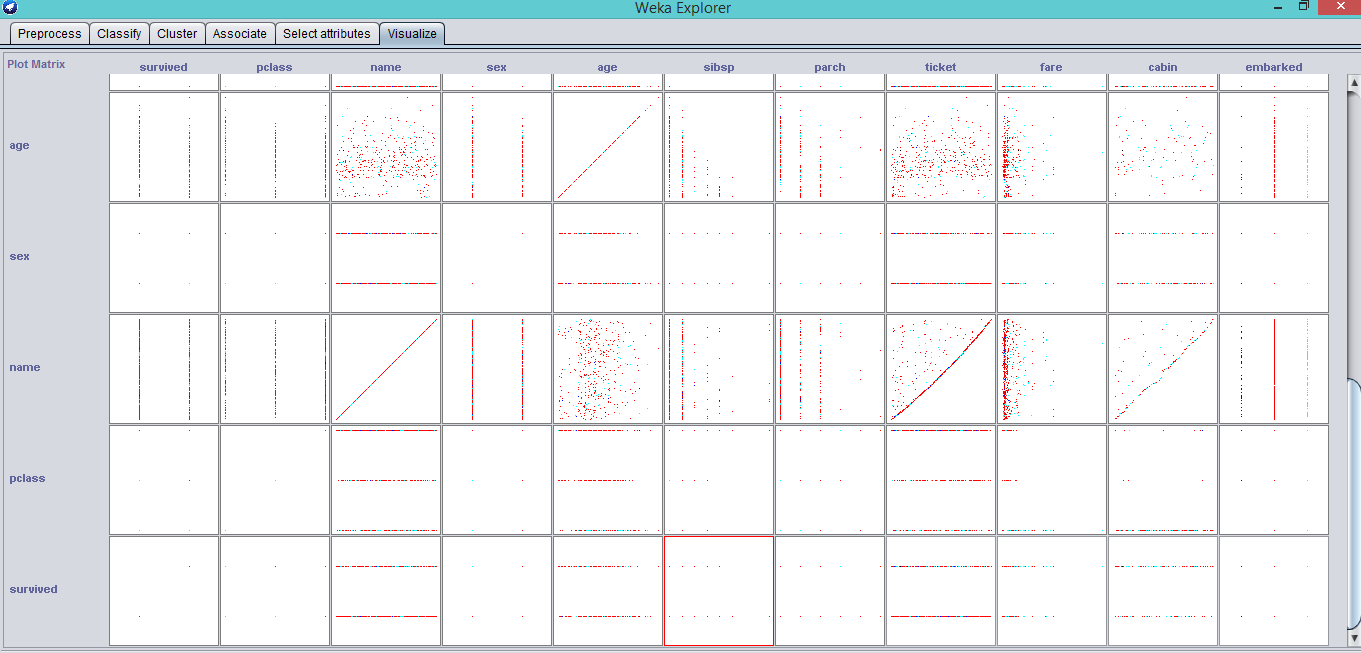
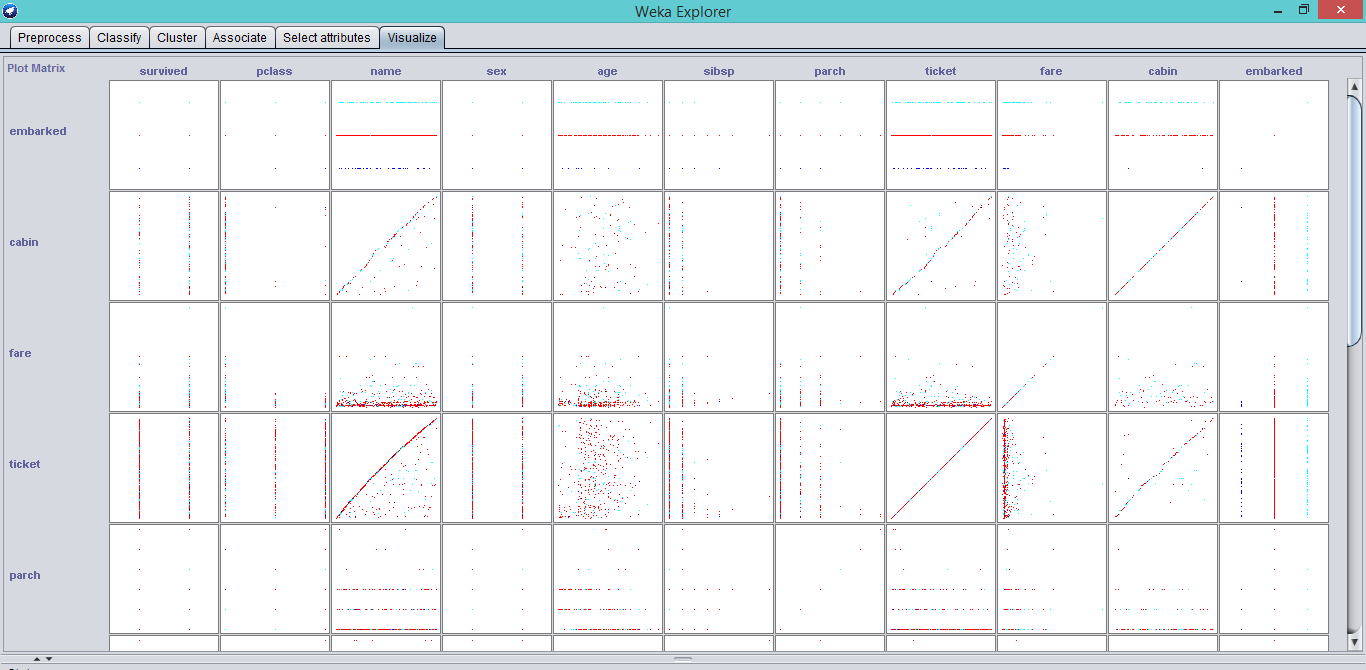
**1] Using training set (default).** After generating the clustering Weka classifies the training instances into clusters according to the cluster representation and computes the percentage of instances falling in each cluster. For example, the below clustering produced by k-means shows 40% (716instances) in cluster 0 and 60% (1079 instances) in cluster 1.

=== Stratified training-set ===   
=== Summary ===  
Correctly Classified Instances 600  
81.0325 %  
Incorrectly Classified Instances 169  
18.9675 %  
Kappa statistic 0.5714  
Mean absolute error 0.2911  
Root mean squared error 0.385  
Relative absolute error 61.5359 %  
Root relative squared error 79.1696 %  
Total Number of Instances 891  
=== Detailed Accuracy By Class ===  
 TP Rate FP Rate Precision Recall  
 0.953 0.418 0.785 0.953  
 0.582 0.047 0.884 0.582  
W.Avg. 0.81 0.276 0.823 0.81  
F-Measure ROC Area Class  
0.861 0.783 No  
0.702 0.783 Yes  
0.8 0.783  
=== Confusion Matrix ===  
 a b <-- classified as  
480 26 | a = No  
143 199 | b = Yes

**Cross**-**validation** is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it. In (10)k-fold **cross**-**validation**, the original sample is randomly partitioned into k equal size subsamples.

=== Stratified cross-validation ===   
=== Summary ===  
Correctly Classified Instances 722  
81.0325 %  
Incorrectly Classified Instances 169  
18.9675 %  
Kappa statistic 0.5714  
Mean absolute error 0.2911  
Root mean squared error 0.385  
Relative absolute error 61.5359 %  
Root relative squared error 79.1696 %  
Total Number of Instances 891  
=== Detailed Accuracy By Class ===  
 TP Rate FP Rate Precision Recall  
 0.953 0.418 0.785 0.953  
 0.582 0.047 0.884 0.582  
W.Avg. 0.81 0.276 0.823 0.81  
F-Measure ROC Area Class  
0.861 0.783 No  
0.702 0.783 Yes  
0.8 0.783  
=== Confusion Matrix ===  
 a b <-- classified as  
523 26 | a = No  
143 199 | b = Yes

Visualization:



**Conclusion:** Thus, k means clustering algorithm using WEKA has been implemented and Cross-validation Technique performs better than the training-set on the given dataset.