CSA1618 DWDM-DE

EXPERIMENT-22

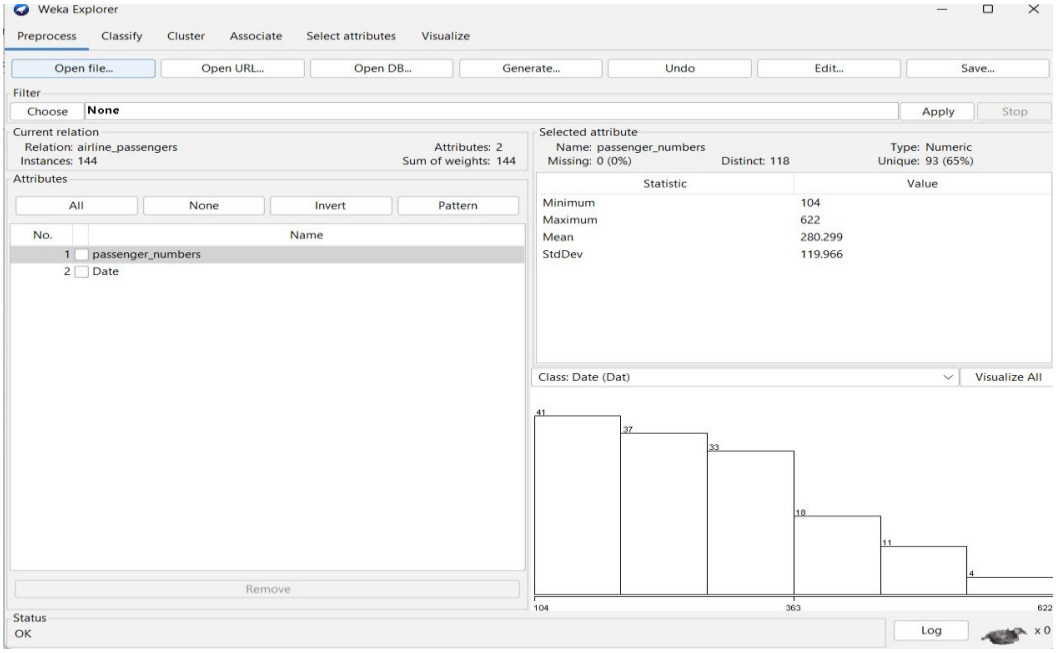
DATA SEGMENTATION BY K-MEANS CLUSTER USING WEKA AND R-TOOL

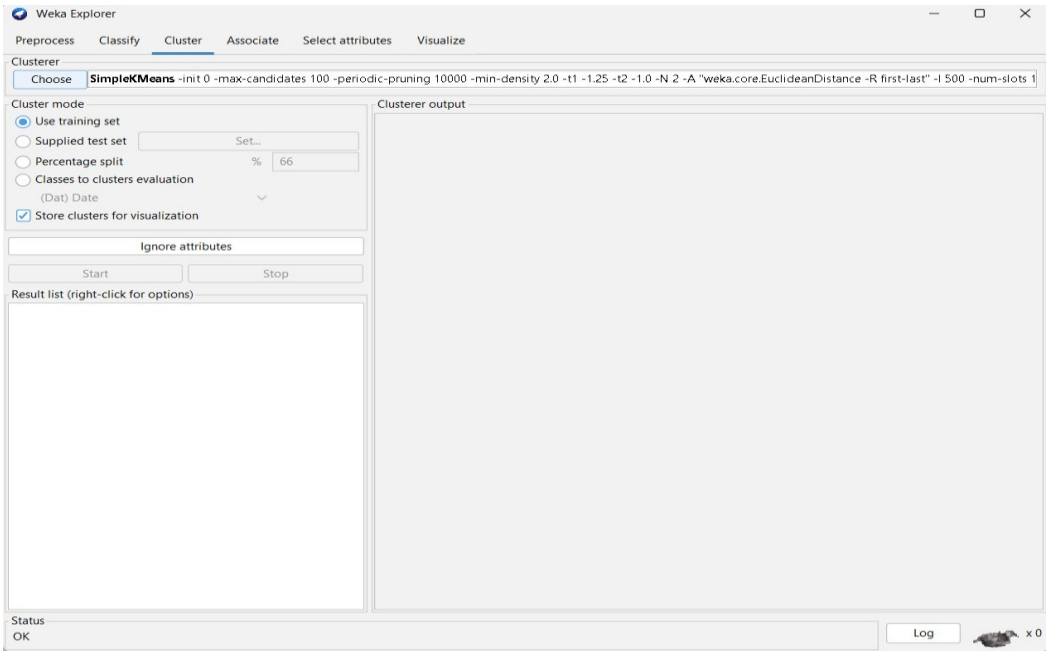
AIM:

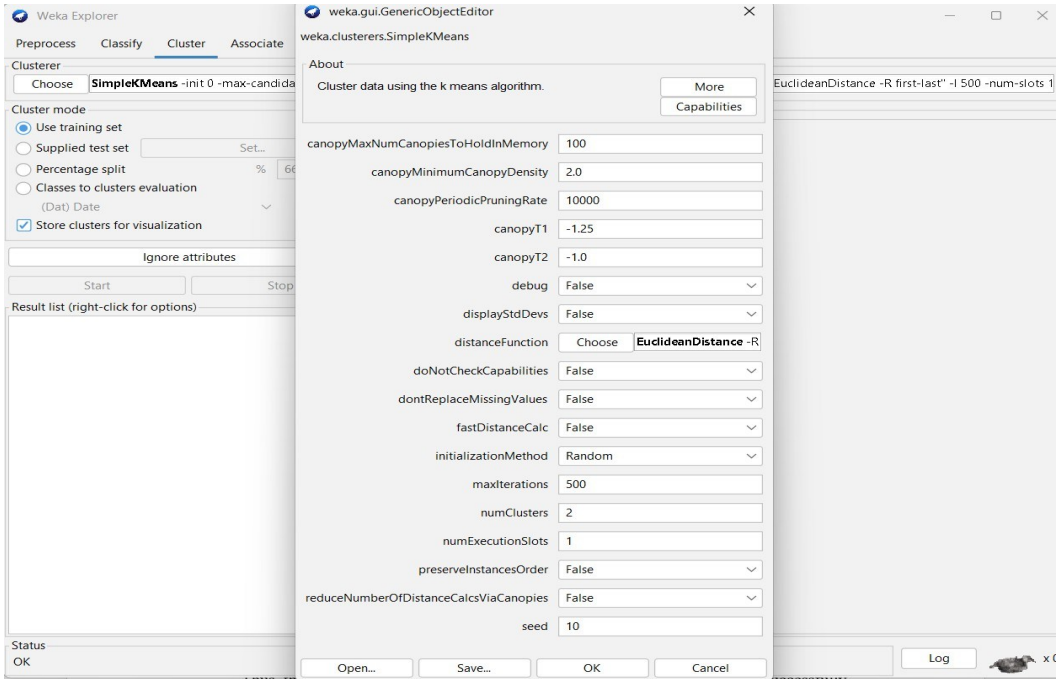
To create DataSegmentation by k-means cluster using weka and R-tool.

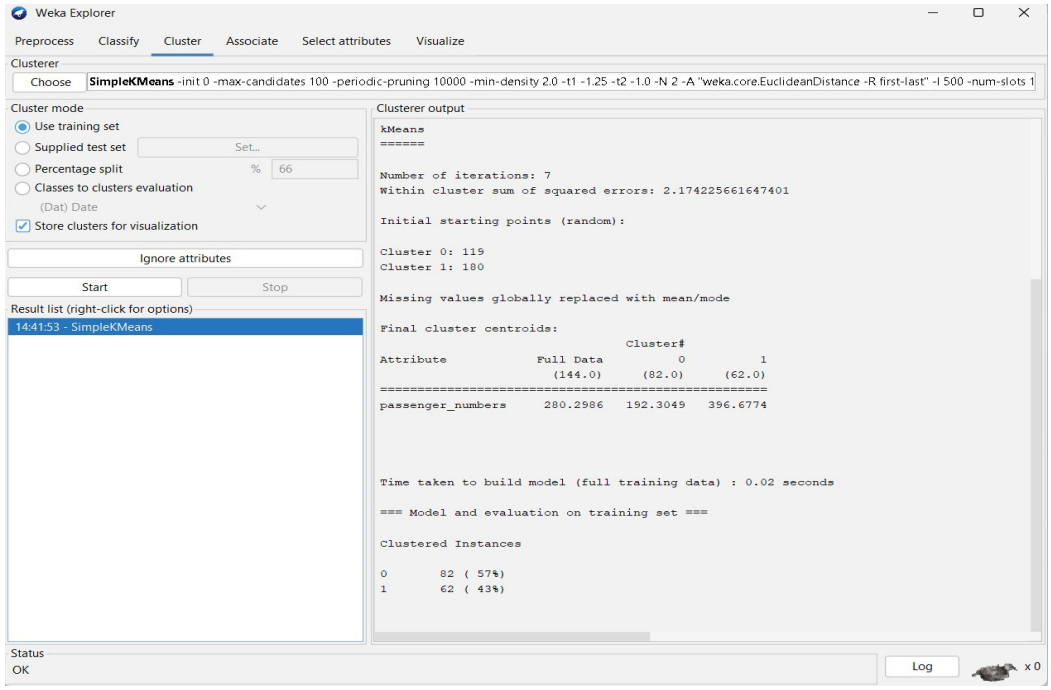
PROCEDURE:

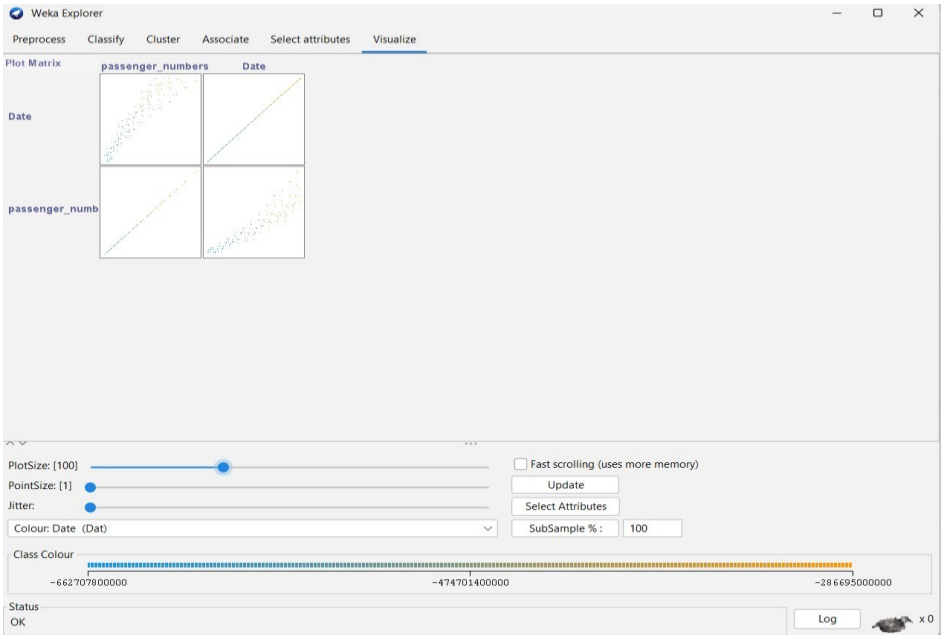
1. Download and install WEKA.
2. Open WEKA and Choose "Explorer" from the main menu.
3. Under Preprocess, Click on the open file button and select the dataset.
4. Click on the "Cluster" tab. In the Cluster mode section, select "Use training set".
5. Click "Choose" (next to the cluster algorithm) and Select SimpleKMeans (under weka.clusterers).
6. Click on "SimpleKMeans" to configure it - Set "numClusters" to the desired number of clusters (e.g., 2, 3, 5). Set "Seed" (random initialization, e.g., 10). Choose "Euclidean distance" (default).
7. Click "OK" and then "Start" to run clustering.
8. WEKA will display cluster assignments and statistics. Click "Visualize" to see how the clusters are distributed. Save the file.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: airline_passengers
Instances: 144
Attributes: 2
Sum of weights: 144
Attributes
All
None
Invert
Pattern
No.
Name
1 passenger_numbers
2 Date
Remove
Selected attribute
Name: passenger_numbers
Missing: 0 (0%)
Distinct: 118
Type: Numeric
Unique: 93 (65%)
Statistic
Value
Minimum 104
Maximum 622
Mean 280.299
StdDev 119.966
Class: Date (Dat)
Visualize All
41
37
33
18
11
4
104
363
622
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Dat) Date
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
Clusterer output
Status
OK
Log
x 0


Weka Explorer
Preprocess
Classify
Cluster
Associate
Clusterer
Choose
SimpleKMeans -init 0 -max-candida
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
Classes to clusters evaluation
(Dat) Date
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
Status
OK
Log
weka.gui.GenericObjectEditor
weka.clusterers.SimpleKMeans
About
Cluster data using the k means algorithm.
More
Capabilities
canopyMaxNumCanopiesToHoldInMemory 100
canopyMinimumCanopyDensity 2.0
canopyPeriodicPruningRate 10000
canopyT1 -1.25
canopyT2 -1.0
debug False
displayStdDevs False
distanceFunction Choose EuclideanDistance -R
doNotCheckCapabilities False
dontReplaceMissingValues False
fastDistanceCalc False
initializationMethod Random
maxIterations 500
numClusters 2
numExecutionSlots 1
preserveInstancesOrder False
reduceNumberOfDistanceCalcsViaCanopies False
seed 10
Open...
Save...
OK
Cancel
EuclideanDistance -R first-last" -I 500 -num-slots 1


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
% 66
Classes to clusters evaluation
(Dat) Date
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
14:41:53 - SimpleKMeans
Clusterer output
kMeans
======
Number of iterations: 7
Within cluster sum of squared errors: 2.174225661647401
Initial starting points (random):
Cluster 0: 119
Cluster 1: 180
Missing values globally replaced with mean/mode
Final cluster centroids:
Cluster#
Attribute Full Data 0 1
(144.0) (82.0) (62.0)
passenger_numbers 280.2986 192.3049 396.6774
Time taken to build model (full training data) : 0.02 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 82 (57%)
1 62 (43%)
Status
OK
Log
x 0



OBSERVATION:

K-means

=======

Number of iterations: 7

Within cluster sum of squared errors: 2.174225661647401 Initial

starting points (random):

Cluster 0: 119

Cluster 1: 180

Missing values globally replaced with mean/mode Final

cluster centroids:

		Cluster#	
Attribute	Full Data	0	1
	(144.0)	(82.0)	(62.0)
passenger_numbers	280.2986	192.3049	396.6774

Time taken to build model (full training data) : 0.02 seconds

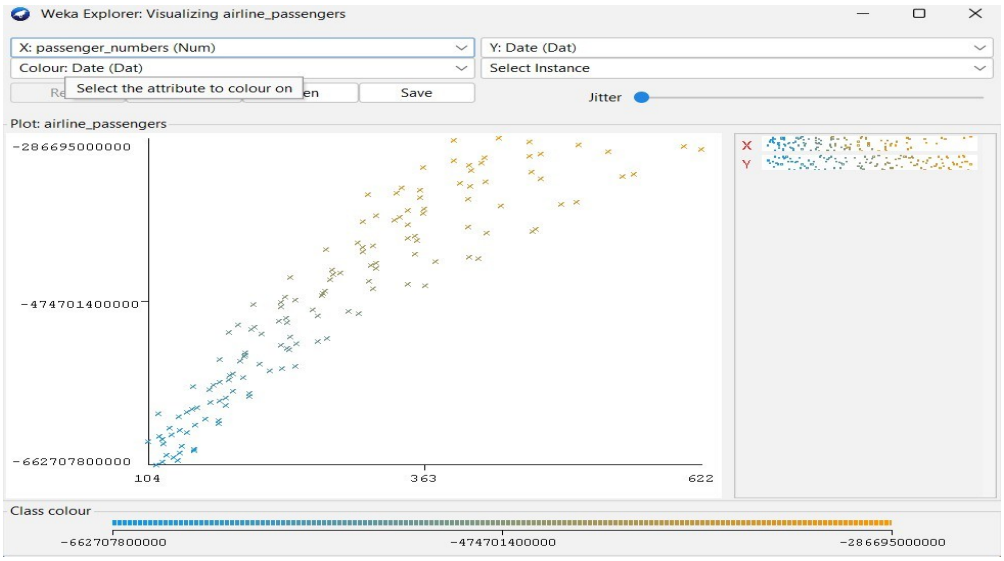
=== Model and evaluation on training set === Clustered

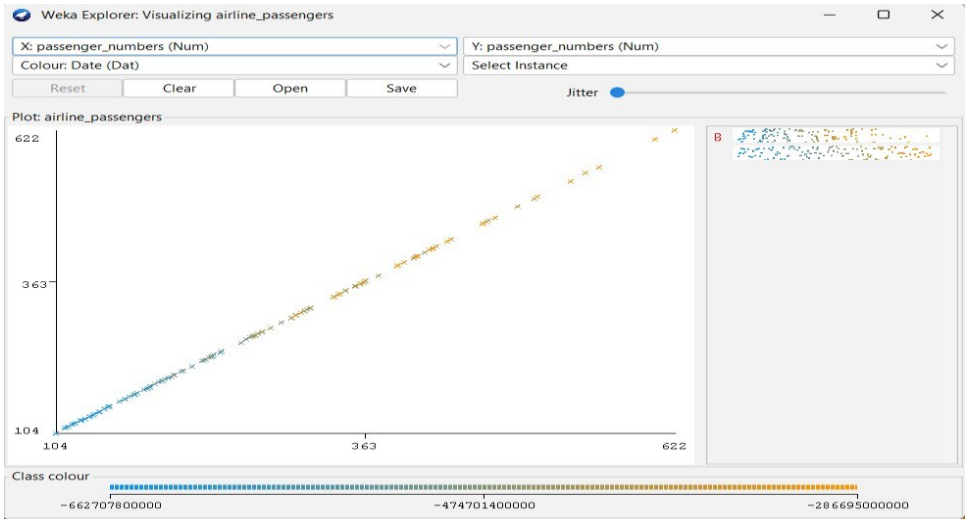
Instances

0 82 (57%)

1 62 (43%)

PLOT:





RESULT:

Thus, the K-means clustering analyzing using the weka tool has been successfully completed. In case of weka tool, the change in seed values lead to the decrease in the number of iterations.