Name: Allan Rodrigues
Class: TE IT A
Roll no: 59

St. Francis Institute of Technology, Mumbai-400 103

Department Of Information Technology

A.Y. 2021-2022

Experiment – 7: a)To Implement any one of the clustering algorithm using WEKA (K Means, Agglomerative, Divisive)

1. **Aim: :** To Implement any one of the clustering algorithm using WEKA (K Means, Agglomerative, Divisive)

2. **Objectives:** After study of this experiment, the students will be able to
   Implement K Means

3. **Outcomes:** After study of this experiment, the students will be able to

   CO 4: Design and Implement various clustering data mining techniques such as Partitioning methods, Hierarchical Methods, Density - Based methods along with identification and analysis of outlier.

4. **Prerequisite:** Introduction to all the three clustering algorithms & Problem solving approach.

5. **Requirements:** Personal Computer, Windows XP operating system/Windows 7, Internet Connection, Microsoft Word, WEKA tool.

6. **Theory:**
   a. What is Clustering in Data Mining?
   b. Difference Between Classification & Clustering
   c. Study of Algorithms for Clustering
   d. Implementation of Clustering Algorithms in WEKA

7. **Laboratory Exercise:** Implementation of Clustering Algorithm using WEKA, Printout of implementation along with coding and snapshot.

8. **Post-Experiments Exercise**
   a. **Questions:**
      - Difference between supervised and unsupervised learning
   b. **Conclusion:**
      - Summary of Experiment
      - Importance of Experiment
      - Application of Experiment

9. **Reference:** Data Mining: Concept & Techniques, 3rd Edition, Jiawei Han, Micheline Pei, Elsevier.

Q6)

1) What is clustering in data mining?

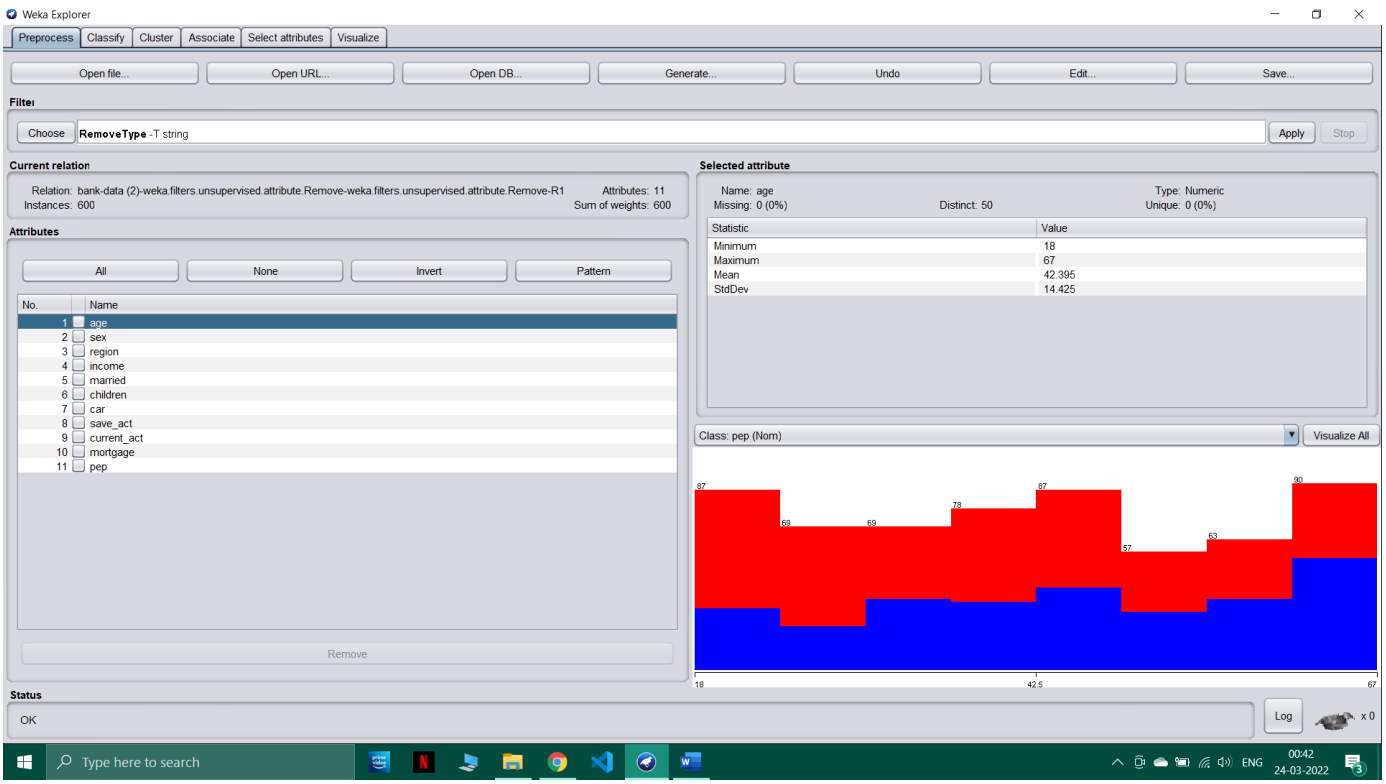
Cluster Analysis is the process to find similar groups of objects in order to form clusters.It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

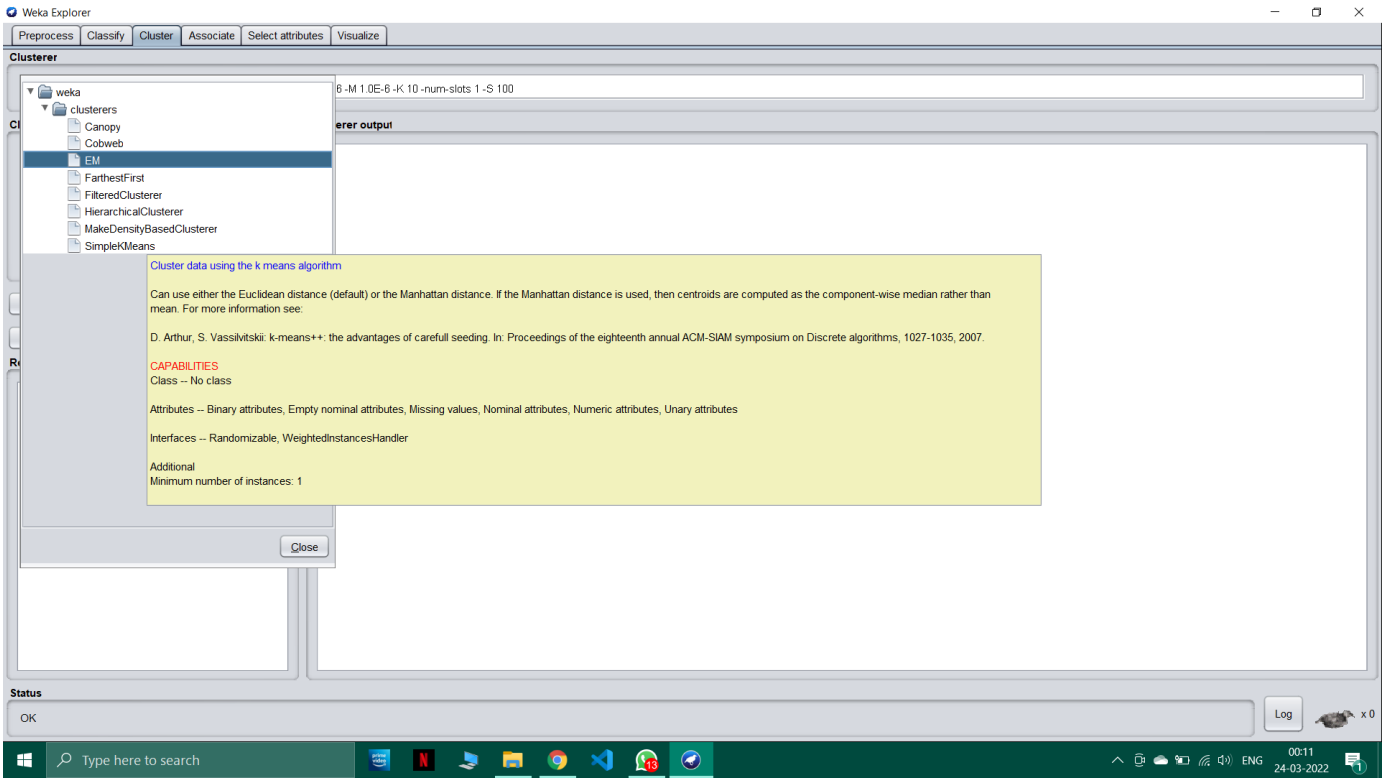
2) Difference between classification and clustering?

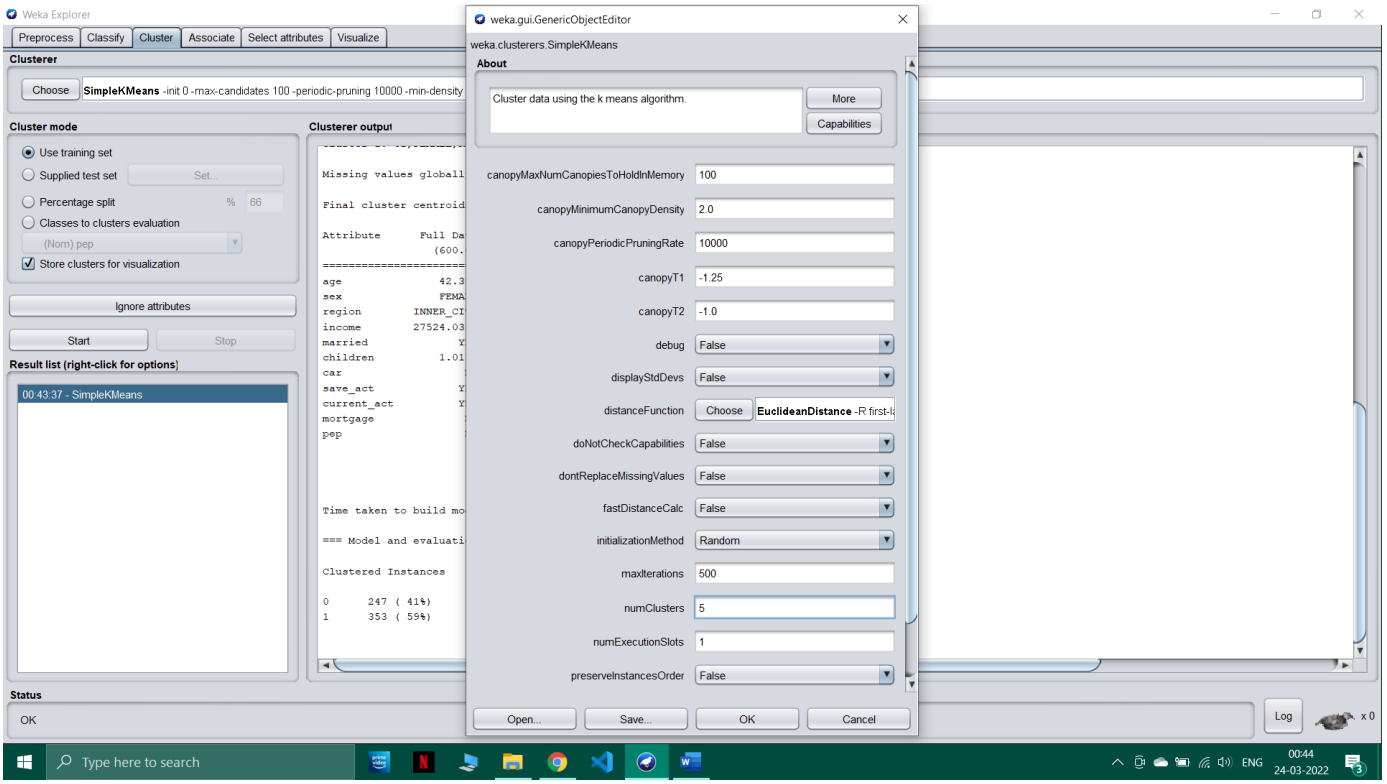
classification	clustering
It is used for supervised learning	It is used for unsupervised learning
Process of classifying the input instances based on their corresponding class labels	Grouping the instances based on their similarity without the help of class labels
It has labels so there is need of training and testing dataset for verifying the model created	There is no need of training and testing dataset
Examples : Logistic regression, Naive Bayes classifier, Support vector machines, etc.	Examples : k-means clustering algorithm, Fuzzy c-means clustering algorithm, Gaussian (EM) clustering algorithm, etc.

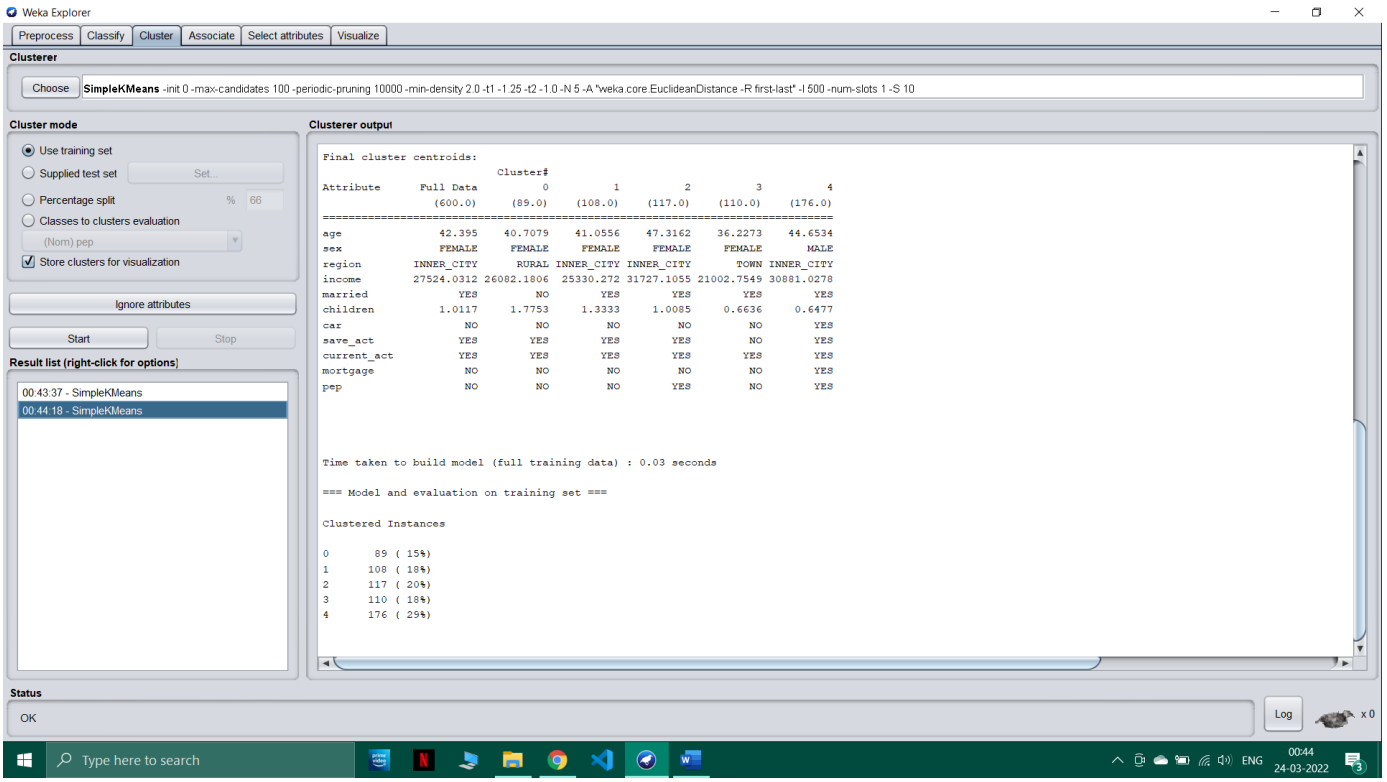
3) Study of algorithms for clustering

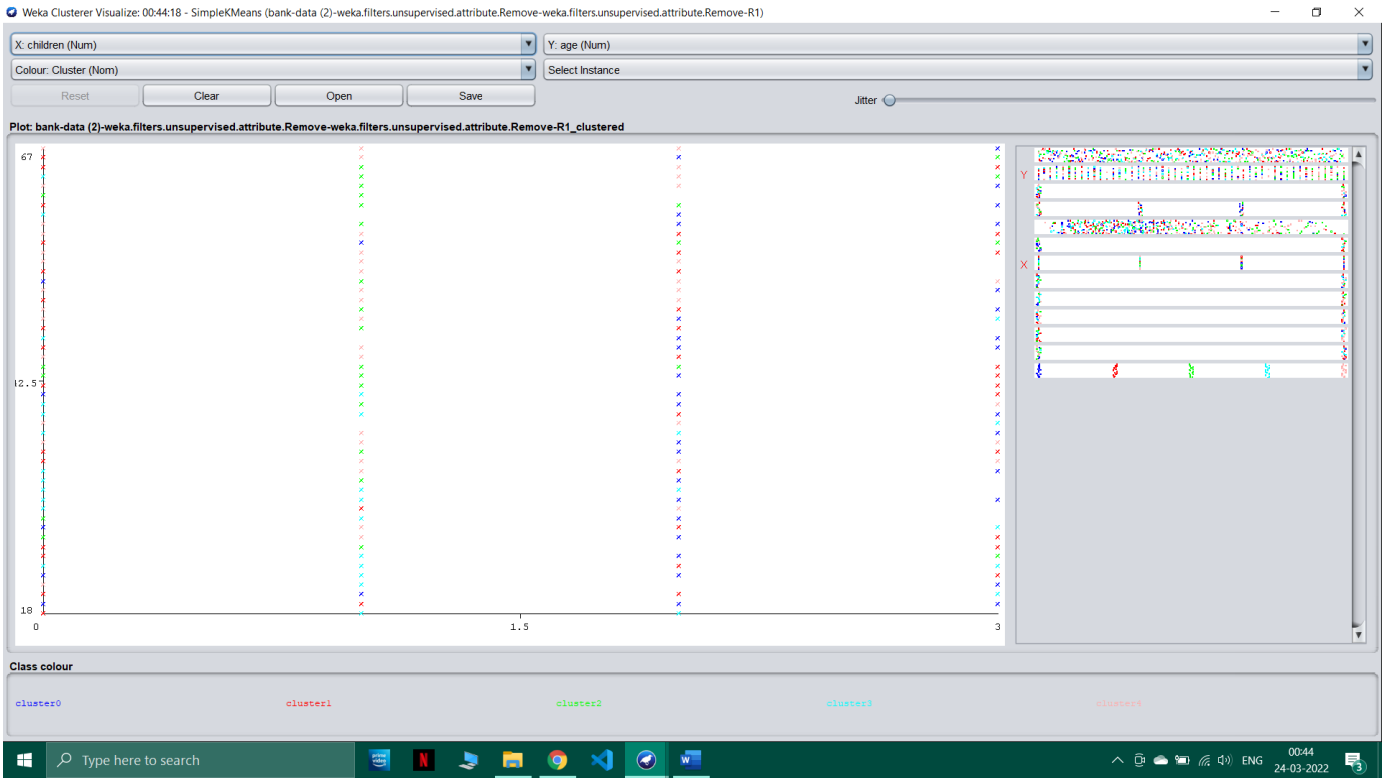
- K-means clustering algorithm
- DBSCAN clustering algorithm
- Gaussian Mixture Model algorithm
- Agglomerative Hierarchy clustering algorithm

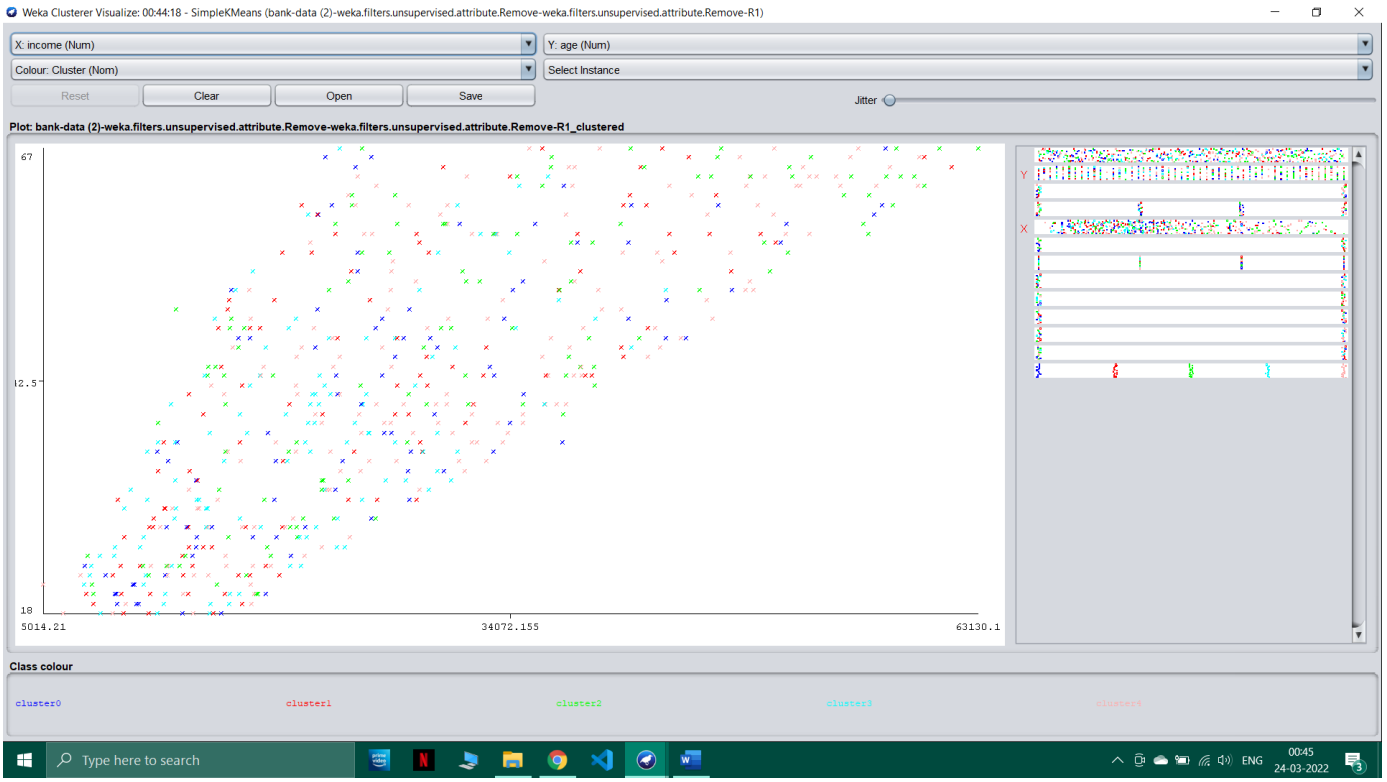



Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
weka
clusterers
Canopy
Cobweb
EM
FarthestFirst
FilteredClusterer
HierarchicalClusterer
MakeDensityBasedClusterer
SimpleKMeans
6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster data using the k means algorithm
Can use either the Euclidean distance (default) or the Manhattan distance. If the Manhattan distance is used, then centroids are computed as the component-wise median rather than mean. For more information see:
D. Arthur, S. Vassilvitskii: k-means++: the advantages of carefull seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, 1027-1035, 2007.
CAPABILITIES
Class -- No class
Attributes -- Binary attributes, Empty nominal attributes, Missing values, Nominal attributes, Numeric attributes, Unary attributes
Interfaces -- Randomizable, WeightedInstancesHandler
Additional
Minimum number of instances: 1
Close
Status
OK
Log
x 0
Type here to search
ENG
00:11
24-03-2022


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) pep
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:43:37 - SimpleKMeans
Clusterer output
Missing values globall
Final cluster centroid
Attribute
Full Da
(600.
age
42.3
sex
FEMA
region
INNER_CI
income
27524.03
married
Y
children
1.01
car
save_act
Y
current_act
Y
mortgage
pep
Time taken to build mo
=== Model and evaluati
Clustered Instances
0
247 (41%)
1
353 (59%)
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 first-l
doNotCheckCapabilities
False
dontReplaceMissingValues
False
fastDistanceCalc
False
initializationMethod
Random
maxIterations
500
numClusters
5
numExecutionSlots
1
preserveInstancesOrder
False
Open...
Save...
OK
Cancel
Status
OK
Log
x 0
Type here to search
ENG
00:44
24-03-2022


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 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) pep
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
00:43:37 - SimpleKMeans
00:44:18 - SimpleKMeans
Clusterer output
Final cluster centroids:
Cluster#
Attribute Full Data 0 1 2 3 4
(600.0) (89.0) (108.0) (117.0) (110.0) (176.0)
age 42.395 40.7079 41.0556 47.3162 36.2273 44.6534
sex FEMALE FEMALE FEMALE FEMALE FEMALE MALE
region INNER_CITY RURAL INNER_CITY INNER_CITY TOWN INNER_CITY
income 27524.0312 26082.1806 25330.272 31727.1055 21002.7549 30881.0278
married YES NO YES YES YES YES
children 1.0117 1.7753 1.3333 1.0085 0.6636 0.6477
car NO NO NO NO NO YES
save_act YES YES YES YES NO YES
current_act YES YES YES YES YES YES
mortgage NO NO NO NO NO YES
pep NO NO NO YES NO YES
Time taken to build model (full training data) : 0.03 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 89 (15%)
1 108 (18%)
2 117 (20%)
3 110 (18%)
4 176 (29%)
Status
OK
Log
x 0
Type here to search
ENG
00:44
24-03-2022


Weka Clusterer Visualize: 00:44:18 - SimpleKMeans (bank-data (2)-weka.filters.unsupervised.attribute.Remove-weka.filters.unsupervised.attribute.Remove-R1)
X: children (Num)
Y: age (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: bank-data (2)-weka.filters.unsupervised.attribute.Remove-weka.filters.unsupervised.attribute.Remove-R1_clustered
67
42.5
18
0
1.5
3
Y
X
Class colour
cluster0
cluster1
cluster2
cluster3
cluster4
Type here to search
ENG
00:44
24-03-2022


Weka Clusterer Visualize: 00:44:18 - SimpleKMeans (bank-data (2)-weka.filters.unsupervised.attribute.Remove-weka.filters.unsupervised.attribute.Remove-R1)
X: income (Num)
Y: age (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: bank-data (2)-weka.filters.unsupervised.attribute.Remove-weka.filters.unsupervised.attribute.Remove-R1_clustered
67
42.5
18
5014.21
34072.155
63130.1
Y
X
Class colour
cluster0
cluster1
cluster2
cluster3
cluster4
Type here to search
ENG
00:45
24-03-2022

Q8)

Allan Rodrigues TE IT A-59

Rajdhani

DATE / /

Exp 7a - BiL

Q.8

a)

Supervised learning	Unsupervised learning
1) Algorithms are trained using labeled data	1) Algorithm are trained using unlabeled data.
2) It predicts the output	2) It finds hidden pattern in data
3) It produces an accurate result	3) It may give less accurate result as compared to supervised learning
4) Uses offline analysis	4) Uses real-time analysis data

b) Conclusion

We performed K-means algorithm on bank dataset in weka

Clustering helps in understanding the natural grouping in a dataset. Clustering quality depends on the method & the identification of hidden patterns

K-means can be applied to data that has a smaller number of dimensions, is numeric and is continuous, such as document clustering, customer segmentation, insurance fraud detection, etc