Experiment No.9
Clustering, Classification and Association Data Mining using WEKA tool
Date of Performance:
Date of Submission:

Aim: To implement clustering , classification and association data mining by using WEKA

Objective: Simulate K-Means Algorithm, Single Linkage AlgorithmDecision tree induction and apriori algorithm by using WEKA

Theory:

WEKA, formally called Waikato Environment for Knowledge Learning, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from agricultural domains. WEKA supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. The basic premise of the application is to utilize a computer application that can be trained to perform machine learning capabilities and derive useful information in the form of trends and patterns. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

1) K-Means Algorithm using WEKA

EXAMPLE:

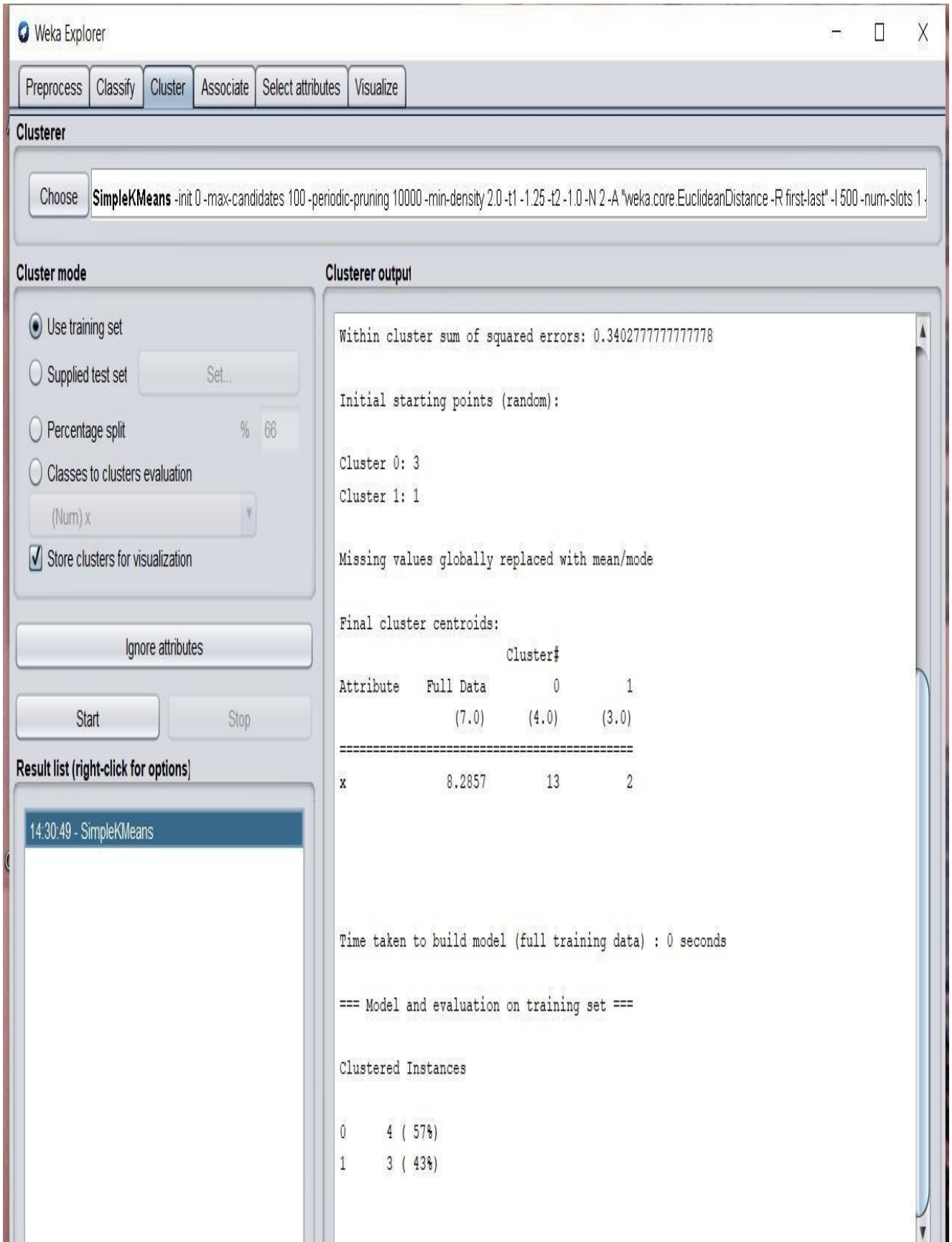
Dataset: D = {1, 2, 3, 8, 9, 10, 25}

1. Randomly assign means m1 = 3 and m2 = 10
   k1 = {1,2,3} k2 = {8,9,10,25}
2. m1 = 2 and m2 = 13
   k1 = {1,2,3} k2 = {8,9,10,25}

WEKA Code:

@RELATION iris

@ATTRIBUTE x NUMERIC
@DATA
1
2
3
8
9
10
25


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
(Num) x
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
14:30:49 - SimpleKMeans
Clusterer output
Within cluster sum of squared errors: 0.3402777777777778
Initial starting points (random):
Cluster 0: 3
Cluster 1: 1
Missing values globally replaced with mean/mode
Final cluster centroids:
Cluster#
Attribute Full Data 0 1
(7.0) (4.0) (3.0)
x 8.2857 13 2
Time taken to build model (full training data) : 0 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 4 (57%)
1 3 (43%)

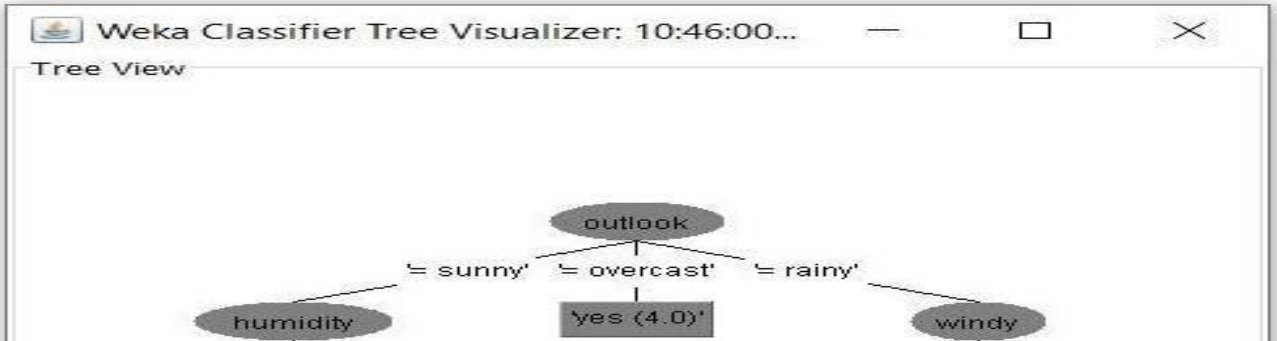
2) Decision Tree Induction using WEKA

A decision tree is a flowchart like tree structure, where each internal node(non-leaf node) denotes a test on an attribute,each branch represents an outcome of the test,and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node

Example:-

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Output:-



3) Apriori Algorithm using WEKA

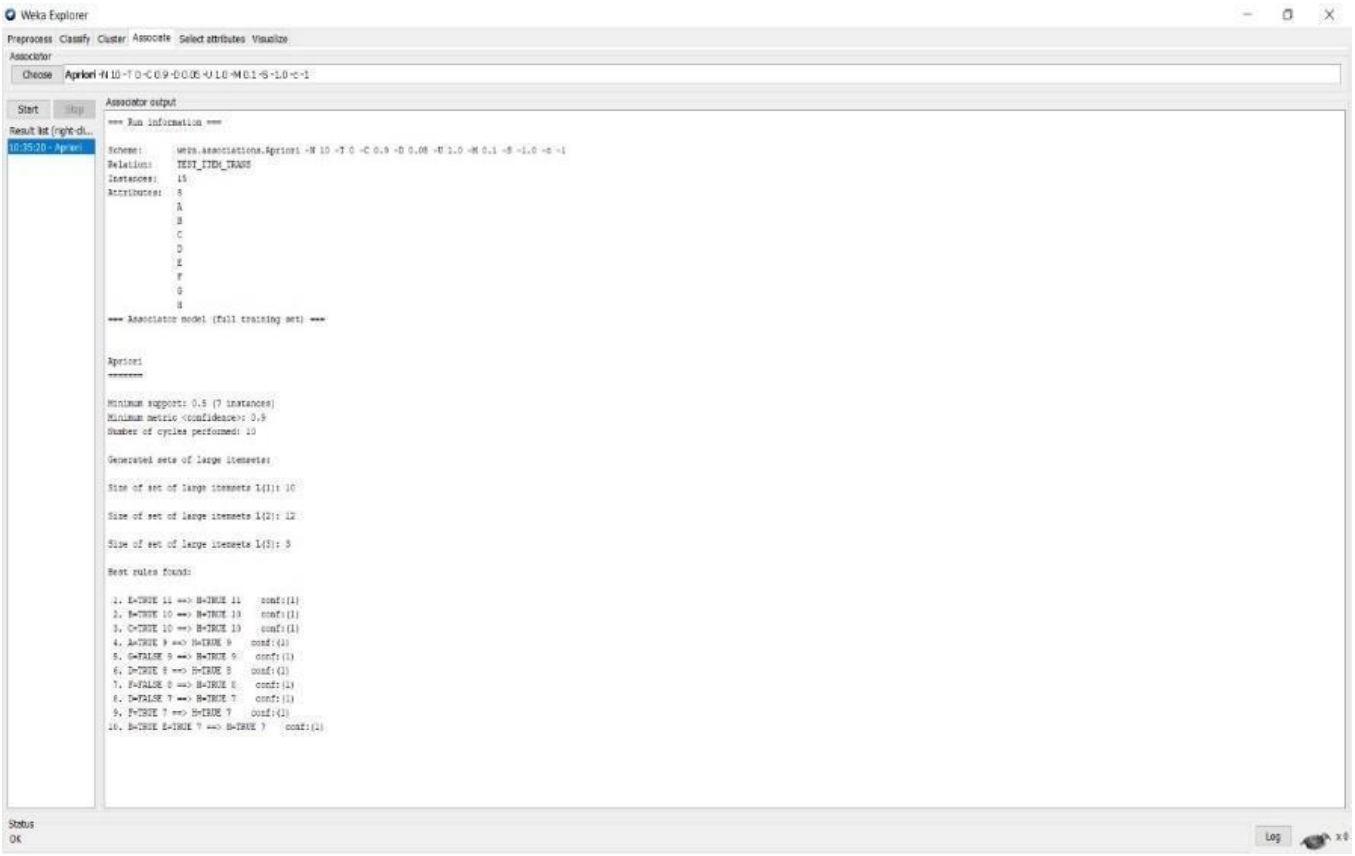
In this current world, globalization is the main feature of any environment. Everyone has to be update, fast and forward and information is the main element for it. For survival in this world it's the basic need to use and to store the information means to prepare a proper database or dataset to analyze. Using and storing the database is not an issue, but finding the relevant dataset or to analyze the meaningful dataset for a particular aspect, from the junkyard of the database is very big problem in analysis of a specific part of the database. To solve this problem the concept of data mining is used to abstracts the desirable information. Useful information from the large databases has been extracted in the form of the association rules. There are many algorithms have been developed to extract the association rules from the large databases. Apriori algorithm is the most popular algorithm to extract the association rules from the databases.

TID	Items
1	A,B,C,D,G,H
2	A,B,C,D,E,F,H
3	B,C,D,E,H
4	B,E,G,H
5	A,B,D,E,G,H
6	A,C,F,G,H
7	B,D,E,G,H
8	A,C,D,E,G,H
9	B,C,D,E,H
10	A,C,E,F,H
11	C,E,H
12	A,D,E,F,H

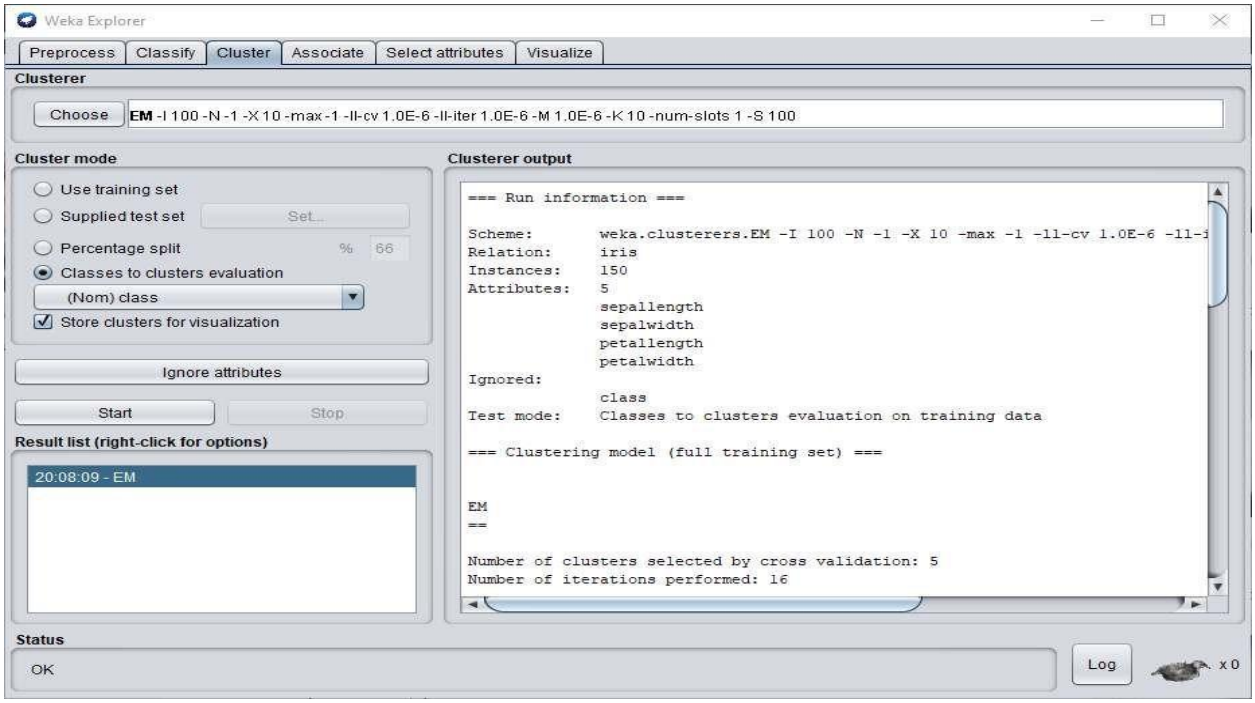
13	B,C,E,F,H
14	A,B,C,F,H
15	A,B,E,F,H

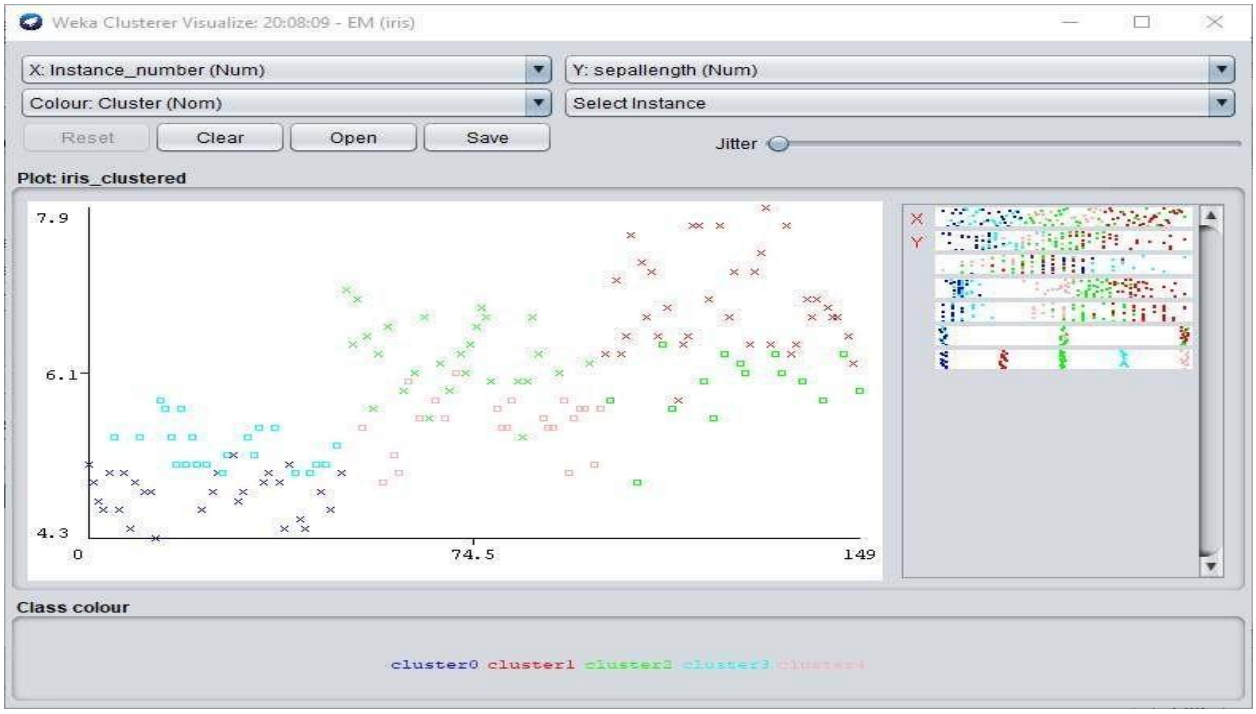
Example

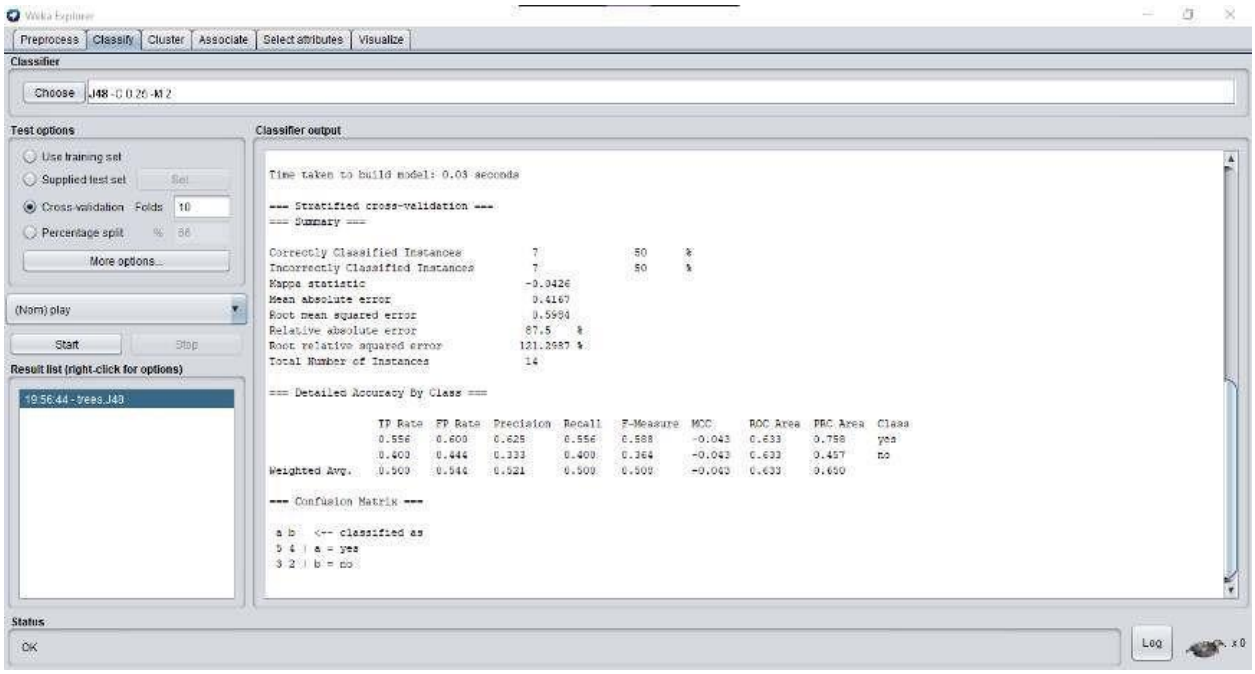
Output

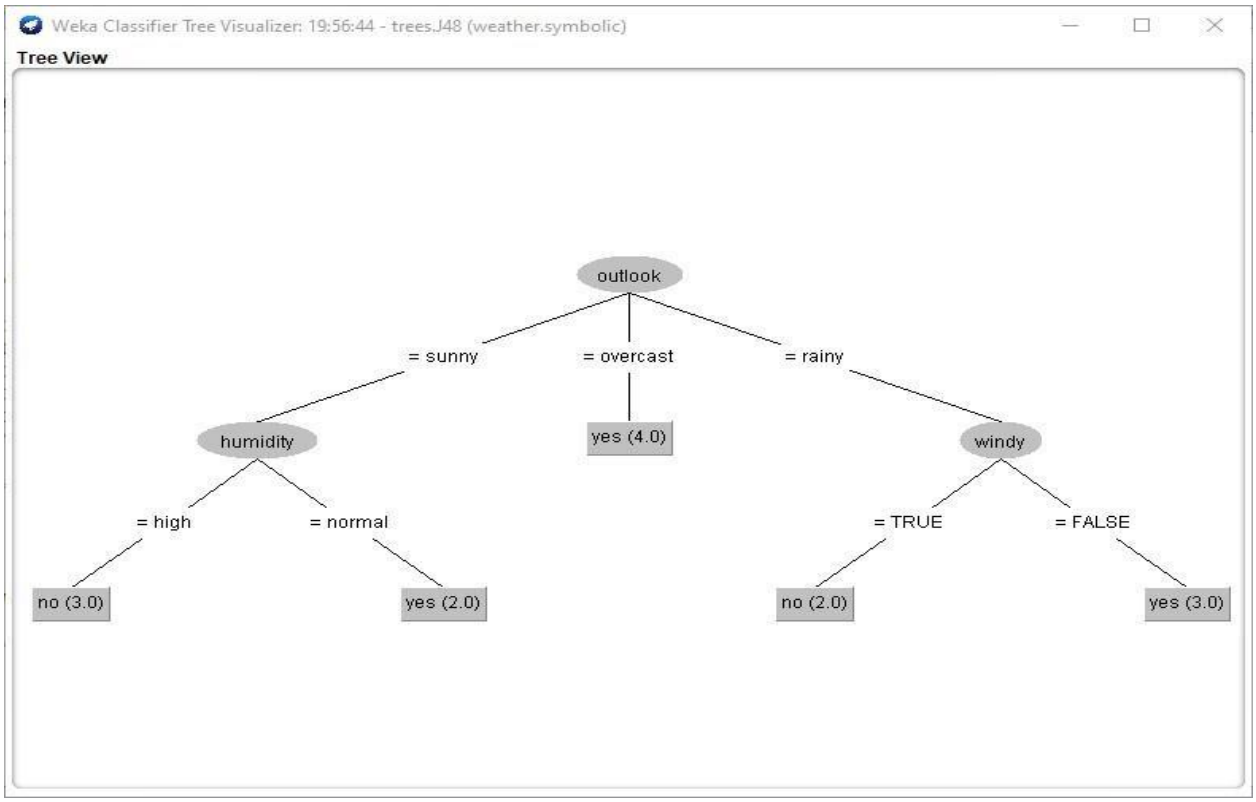


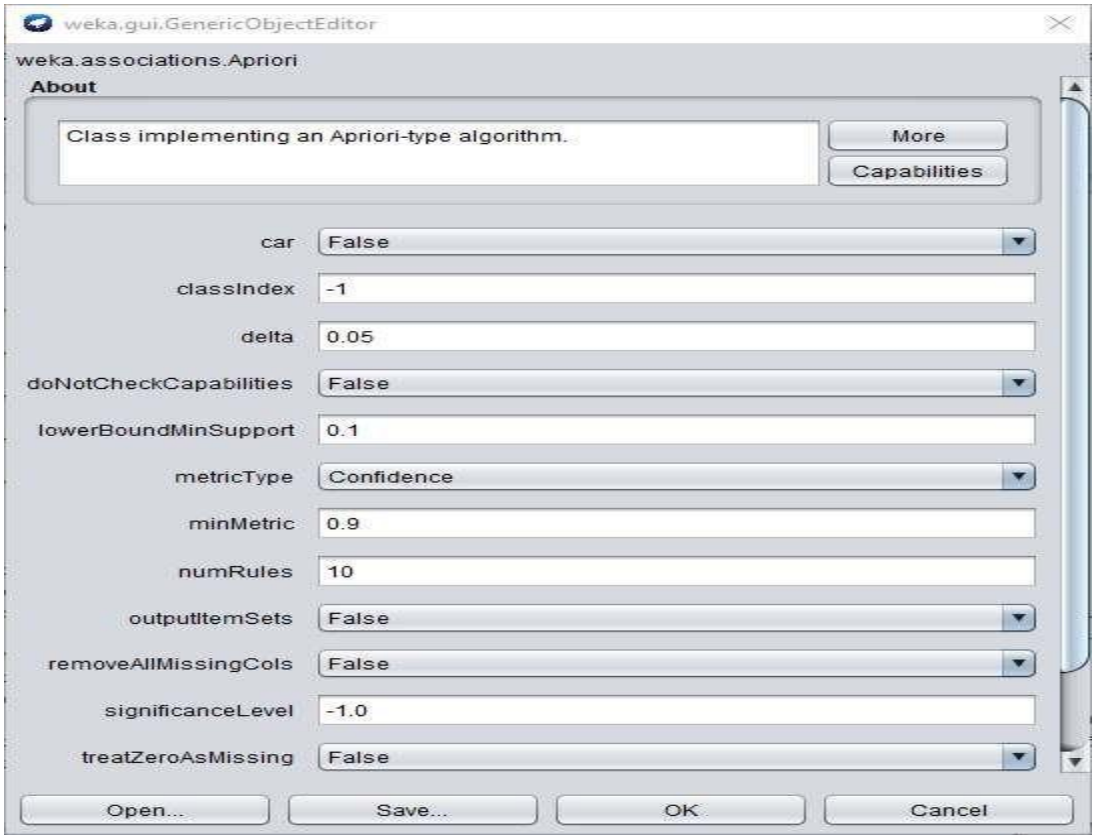
Code and output:

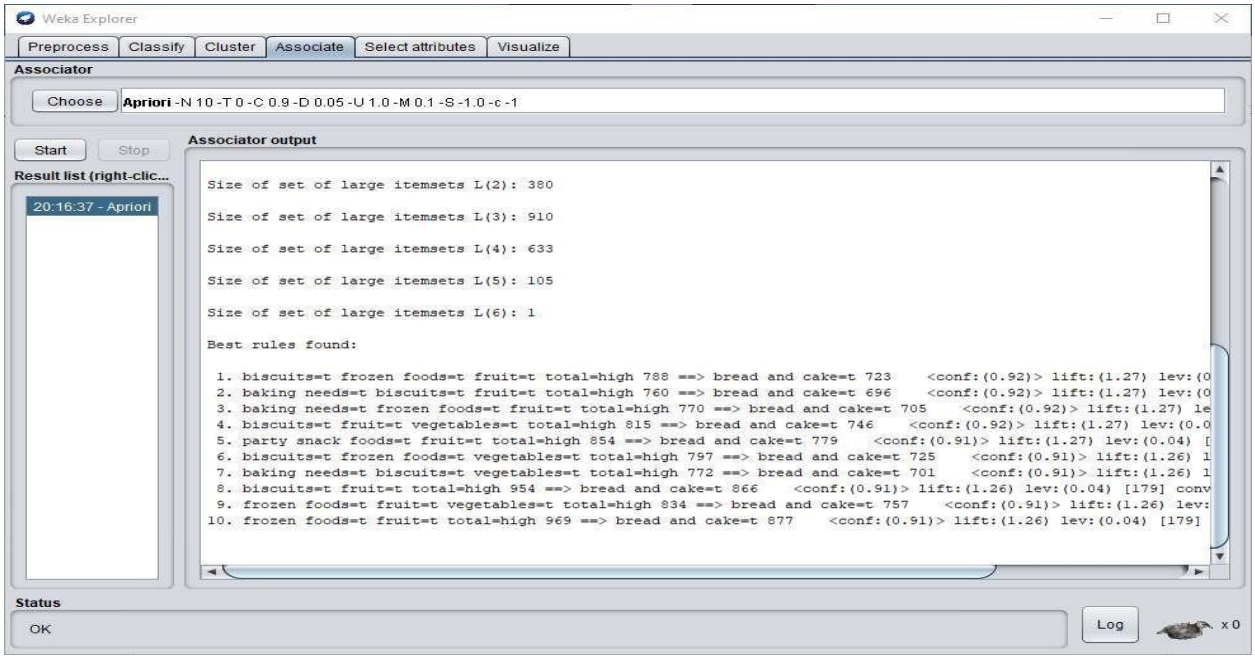

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Nom) class
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
20:08:09 - EM
Clusterer output
=== Run information ===
Scheme: weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-i
Relation: iris
Instances: 150
Attributes: 5
sepallength
sepalwidth
petallength
petalwidth
Ignored:
class
Test mode: Classes to clusters evaluation on training data
=== Clustering model (full training set) ===
EM
==
Number of clusters selected by cross validation: 5
Number of iterations performed: 16
Status
OK
Log
x 0


Weka Clusterer Visualize: 20:08:09 - EM (iris)
X: Instance_number (Num)
Y: sepallength (Num)
Colour: Cluster (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: iris_clustered
7.9
6.1
4.3
0
74.5
149
Class colour
cluster0 cluster1 cluster2 cluster3 cluster4


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
19:56:44 - trees.J48
Classifier output
Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 7 50 %
Incorrectly Classified Instances 7 50 %
Kappa statistic -0.0426
Mean absolute error 0.4167
Root mean squared error 0.5984
Relative absolute error 87.5 %
Root relative squared error 121.2987 %
Total Number of Instances 14
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.556 0.600 0.625 0.556 0.588 -0.043 0.633 0.758 yes
0.400 0.444 0.333 0.400 0.364 -0.043 0.633 0.457 no
Weighted Avg. 0.500 0.544 0.521 0.500 0.509 -0.043 0.633 0.650
=== Confusion Matrix ===
a b <-- classified as
5 4 | a = yes
3 2 | b = no
Status
OK
Log
x 0


Weka Classifier Tree Visualizer: 19:56:44 - trees.J48 (weather.symbolic)
Tree View
outlook
= sunny
= overcast
= rainy
humidity
yes (4.0)
windy
= high
= normal
= TRUE
= FALSE
no (3.0)
yes (2.0)
no (2.0)
yes (3.0)


weka.gui.GenericObjectEditor
weka.associations.Apriori
About
Class implementing an Apriori-type algorithm.
More
Capabilities
car
False
classIndex
-1
delta
0.05
doNotCheckCapabilities
False
lowerBoundMinSupport
0.1
metricType
Confidence
minMetric
0.9
numRules
10
outputItemSets
False
removeAllMissingCols
False
significanceLevel
-1.0
treatZeroAsMissing
False
Open...
Save...
OK
Cancel


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Associator
Choose
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start
Stop
Result list (right-clic...
20:16:37 - Apriori
Associator output
Size of set of large itemsets L(2): 380
Size of set of large itemsets L(3): 910
Size of set of large itemsets L(4): 633
Size of set of large itemsets L(5): 105
Size of set of large itemsets L(6): 1
Best rules found:
1. biscuits=t frozen foods=t fruit=t total=high 788 ==> bread and cake=t 723 <conf:(0.92)> lift:(1.27) lev:(0
2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696 <conf:(0.92)> lift:(1.27) lev:(0
3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705 <conf:(0.92)> lift:(1.27) le
4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746 <conf:(0.92)> lift:(1.27) lev:(0.0
5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779 <conf:(0.91)> lift:(1.27) lev:(0.04) [
6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725 <conf:(0.91)> lift:(1.26) l
7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701 <conf:(0.91)> lift:(1.26) l
8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866 <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv
9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757 <conf:(0.91)> lift:(1.26) lev:
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877 <conf:(0.91)> lift:(1.26) lev:(0.04) [179]
Status
OK
Log
x 0

Conclusion:WEKA (Waikato Environment for Knowledge Analysis) is a popular open-source machine learning software that provides a user-friendly interface for various data mining and machine learning tasks.WEKA offers a graphical user interface (GUI) that allows users, even those without extensive programming or data science experience, to perform a wide range of machine learning tasks. WEKA provides a comprehensive set of preprocessing tools for data cleaning, transformation, and feature selection. WEKA simplifies the process of selecting machine learning algorithms and evaluating their performance.WEKA provides visualization tools that make it easier to understand and interpret the results of machine learning experiments. Once a model is trained, WEKA allows users to save it for future use. WEKA often provides parameter tuning options for algorithms, making it easier to fine-tune model performance without extensive manual experimentation. In summary, WEKA simplifies the machine learning process by offering a user-friendly interface, a rich set of preprocessing and modeling tools, visualization capabilities, and automated techniques for evaluation and parameter tuning. The outputs it generates, including cleaned datasets, performance metrics, and visualizations, assist users in making informed decisions throughout the entire machine learning workflow. This makes WEKA a valuable tool for both beginners and experienced data scientists in their data analysis and modeling endeavors.