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:

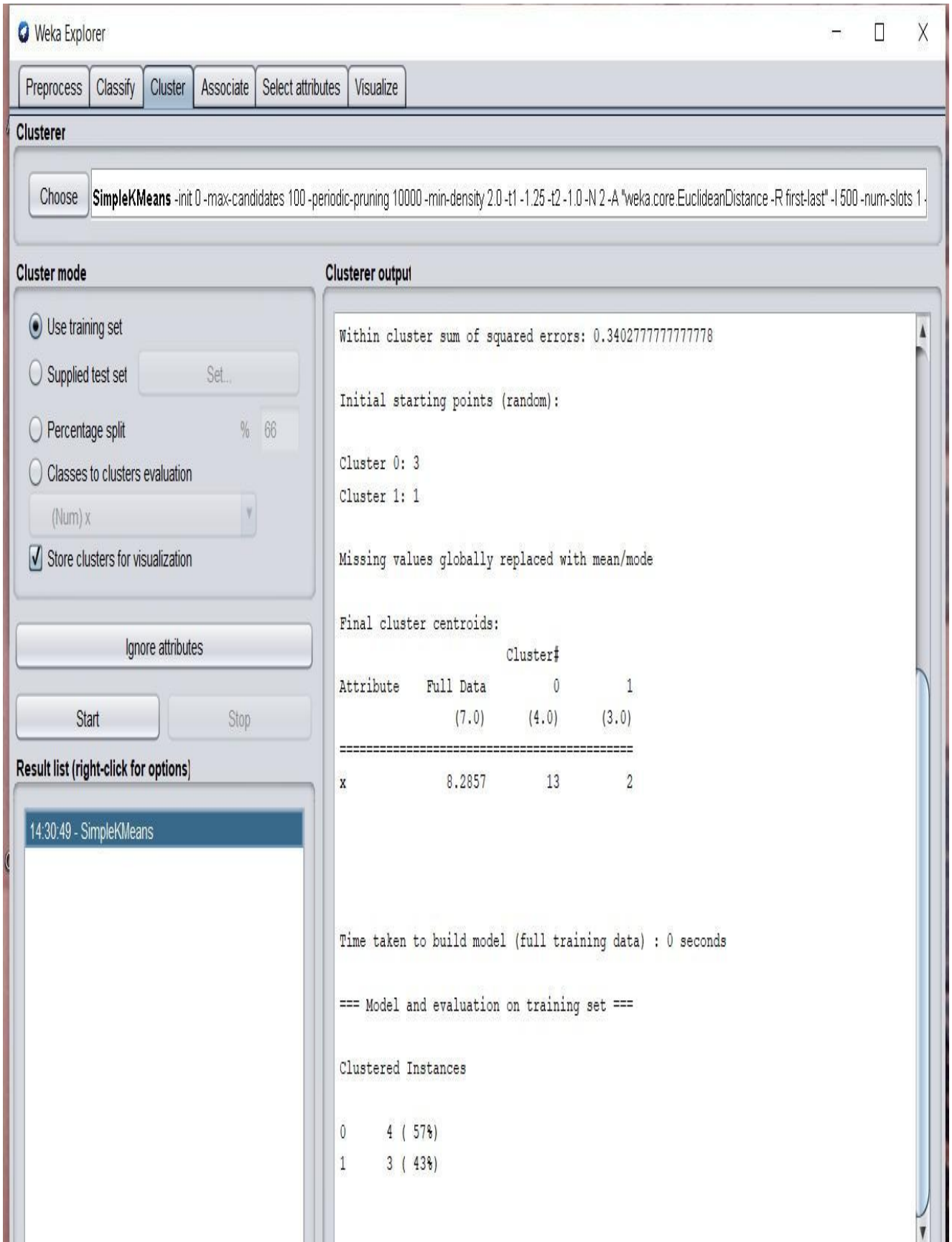
Dataset: D = {1, 2, 3, 8, 9, 10, 25}

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%)

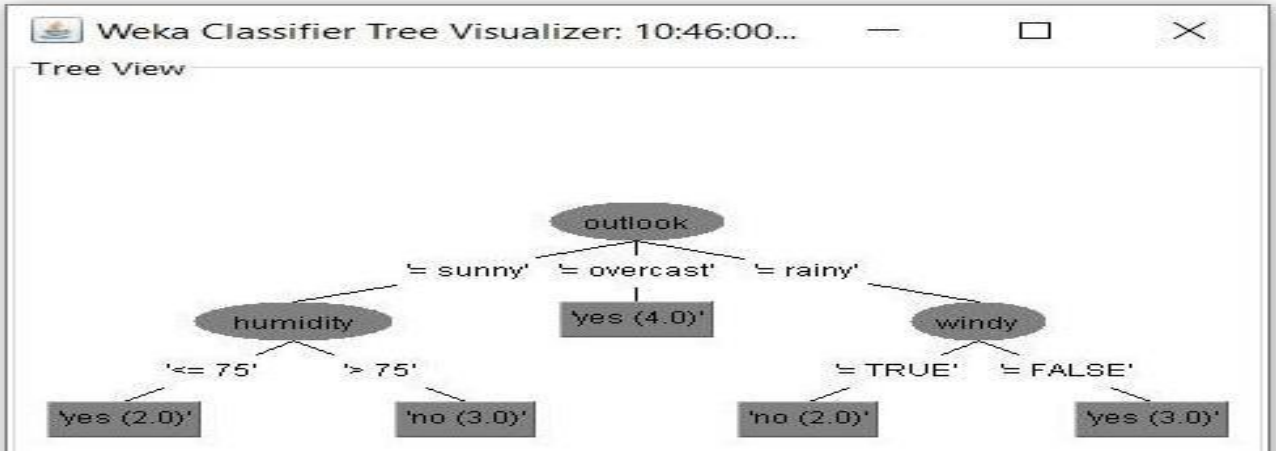
2) Decision Tree Induction using WEKA

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

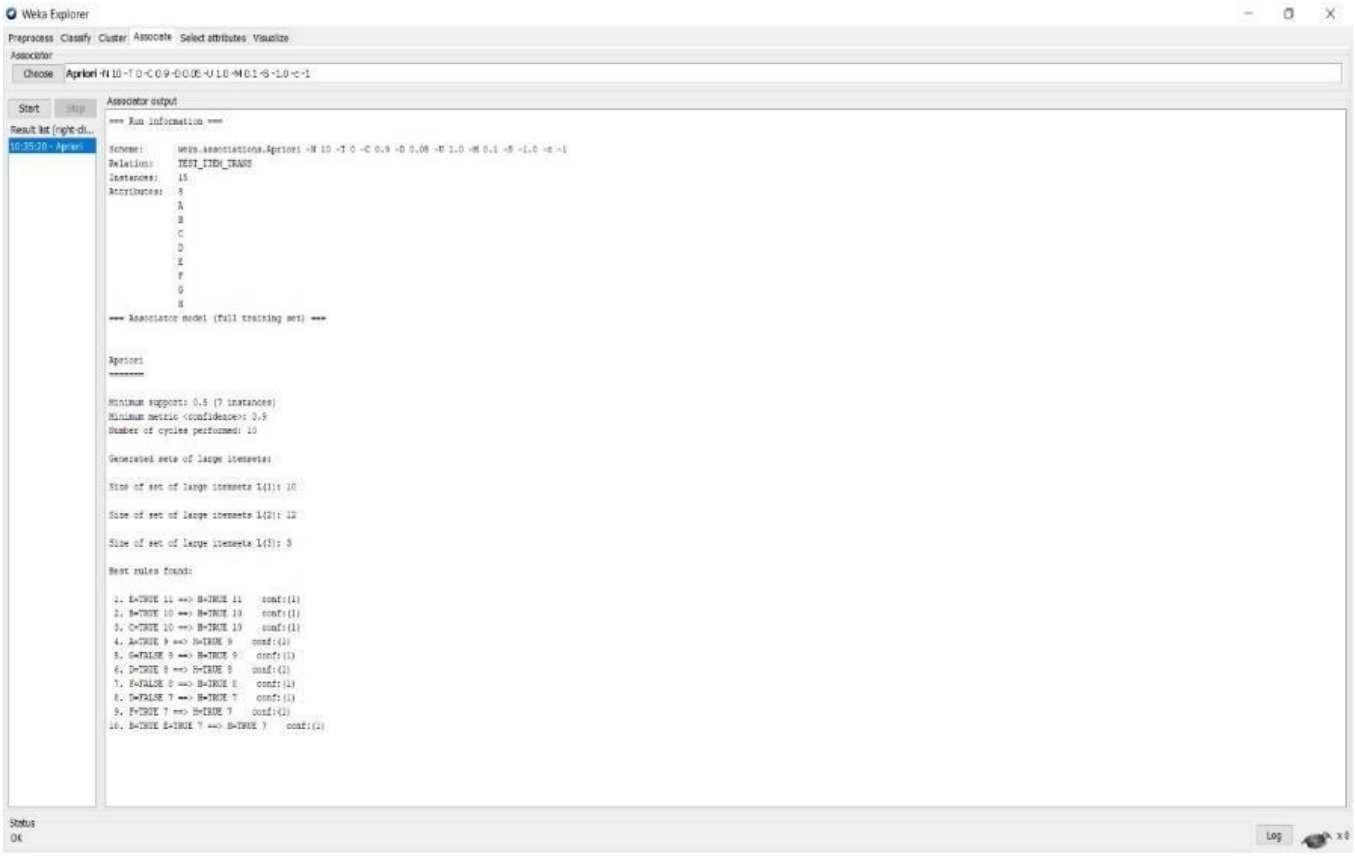
In this current world, globalization is the main feature of any environment. Everyone has to be update, fast and forward and information is the main element for it. For survival in this world it's the basic need to use and to store the information means to prepare a proper database or dataset to analyze. 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

Example

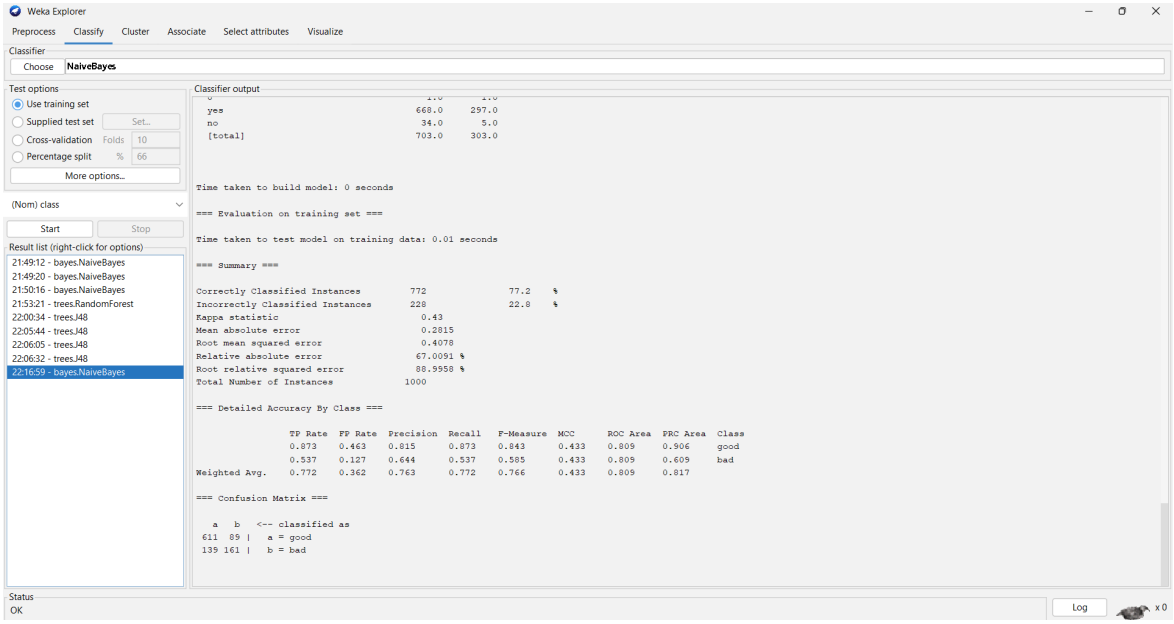
11	C,E,H
12	A,D,E,F,H
13	B,C,E,F,H
14	A,B,C,F,H
15	A,B,E,F,H

Output

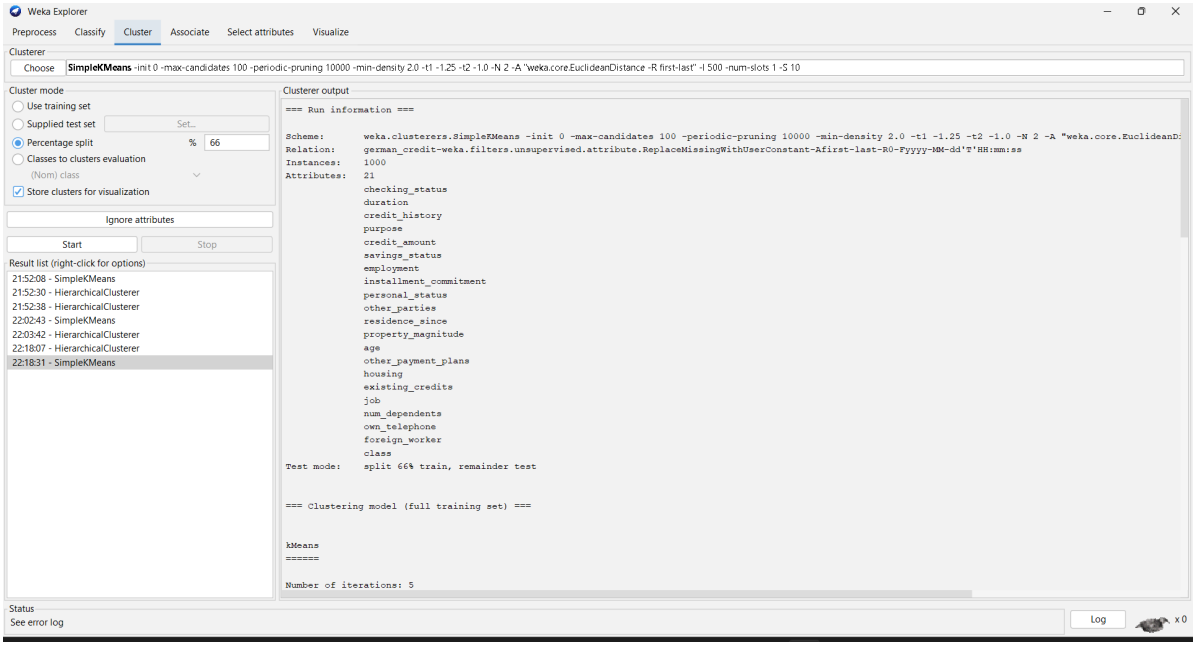

Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Associator
Choose Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Start Stop
Result list (right-cli...
10:35:20 - Apriori
Associator output
=== Run information ===
Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation: TEST_ITEM_TRANS
Instances: 15
Attributes: 8
A
B
C
D
E
F
G
H
=== Associator model (full training set) ===
Apriori
=======
Minimum support: 0.5 (7 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 10
Generated sets of large itemsets:
Size of set of large itemsets L(1): 10
Size of set of large itemsets L(2): 12
Size of set of large itemsets L(3): 5
Best rules found:
1. E=TRUE 11 ==> H=TRUE 11 conf:(1)
2. B=TRUE 10 ==> H=TRUE 10 conf:(1)
3. C=TRUE 10 ==> H=TRUE 10 conf:(1)
4. A=TRUE 9 ==> H=TRUE 9 conf:(1)
5. G=FALSE 9 ==> H=TRUE 9 conf:(1)
6. D=TRUE 8 ==> H=TRUE 8 conf:(1)
7. F=FALSE 8 ==> H=TRUE 8 conf:(1)
8. D=FALSE 7 ==> H=TRUE 7 conf:(1)
9. F=TRUE 7 ==> H=TRUE 7 conf:(1)
10. B=TRUE E=TRUE 7 ==> H=TRUE 7 conf:(1)
Status
OK
Log
x 0

Code and output:

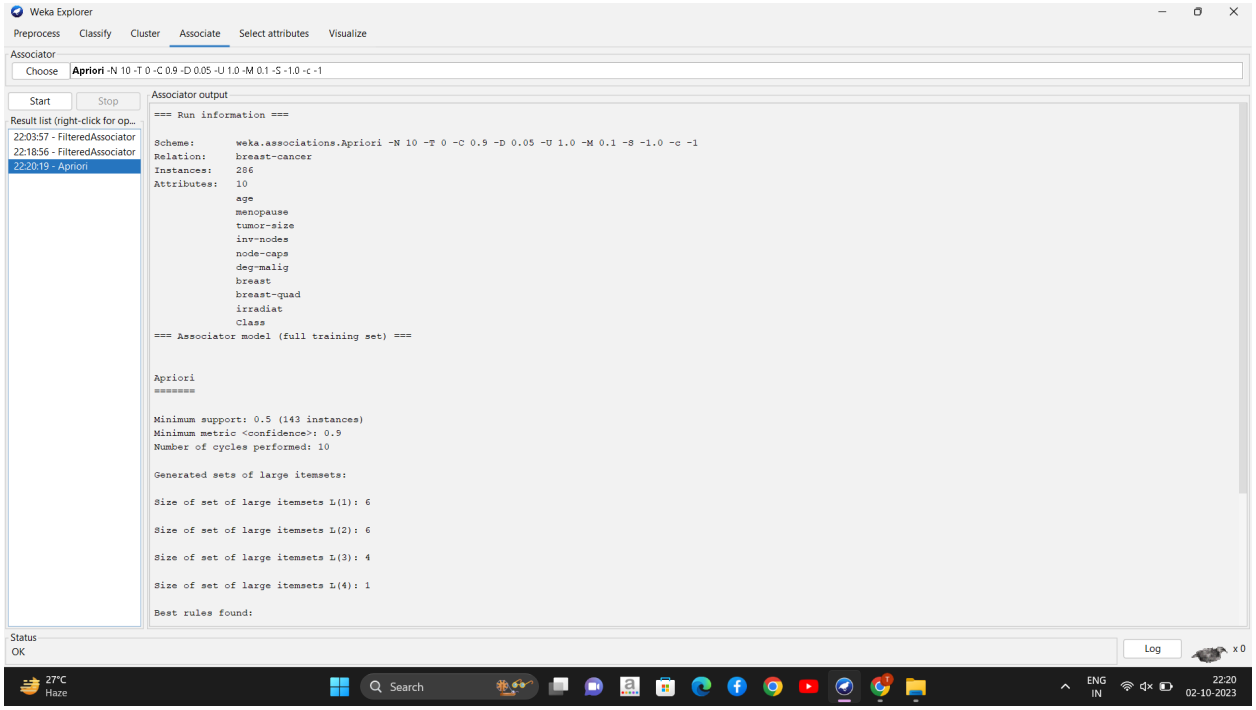
Classification:



Clustering:



Association mining:



Conclusion: WEKA streamlines machine learning with a user-friendly interface and an extensive algorithm library. It generates outputs encompassing model performance metrics, visualizations, and feature insights, aiding informed decision-making. The platform's simplicity lies in its comprehensive preprocessing tools, cross-validation automation, and accessibility for users of varying expertise, making it an effective solution for data analysis and model development.