MOHAMMED RAFIK.M
192124179

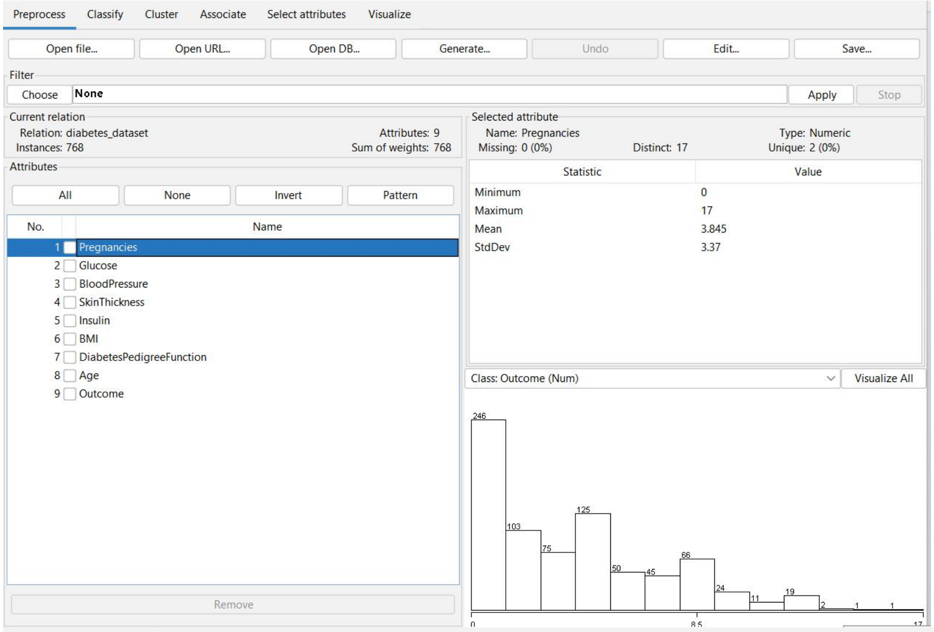
CSA1668 - DWDM FOR PATTERN ANALYSIS

WEKA EXPERIMENTS

1.DATA PREPROCESSING AND PREPARATION FOR KNOWLEDGE ANALYSIS USING WEKA.

DATA PRE-PROCESSING:

It is a data mining technique which is helpful in transforming the raw data into useful and efficent data.



Filter: None

Current relation
Relation: diabetes_dataset
Instances: 768
Attributes: 9
Sum of weights: 768

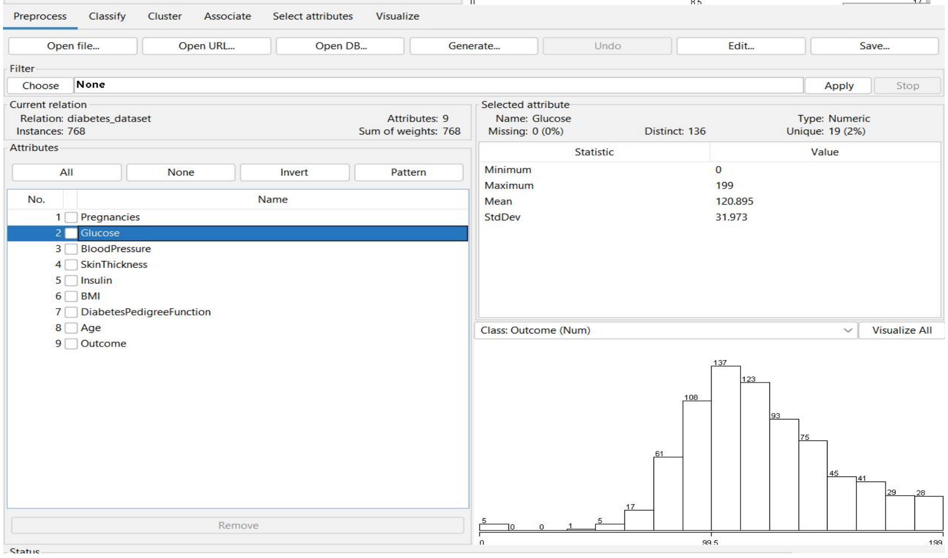
Attributes

No.	Name
1	Pregnancies
2	Glucose
3	BloodPressure
4	SkinThickness
5	Insulin
6	BMI
7	DiabetesPedigreeFunction
8	Age
9	Outcome

Selected attribute
Name: Pregnancies
Missing: 0 (0%)
Distinct: 17
Type: Numeric
Unique: 2 (0%)

Statistic	Value
Minimum	0
Maximum	17
Mean	3.845
StdDev	3.37

Class: Outcome (Num)



Filter: None

Current relation
Relation: diabetes_dataset
Instances: 768
Attributes: 9
Sum of weights: 768

Attributes

No.	Name
1	Pregnancies
2	Glucose
3	BloodPressure
4	SkinThickness
5	Insulin
6	BMI
7	DiabetesPedigreeFunction
8	Age
9	Outcome

Selected attribute
Name: Glucose
Missing: 0 (0%)
Distinct: 136
Type: Numeric
Unique: 19 (2%)

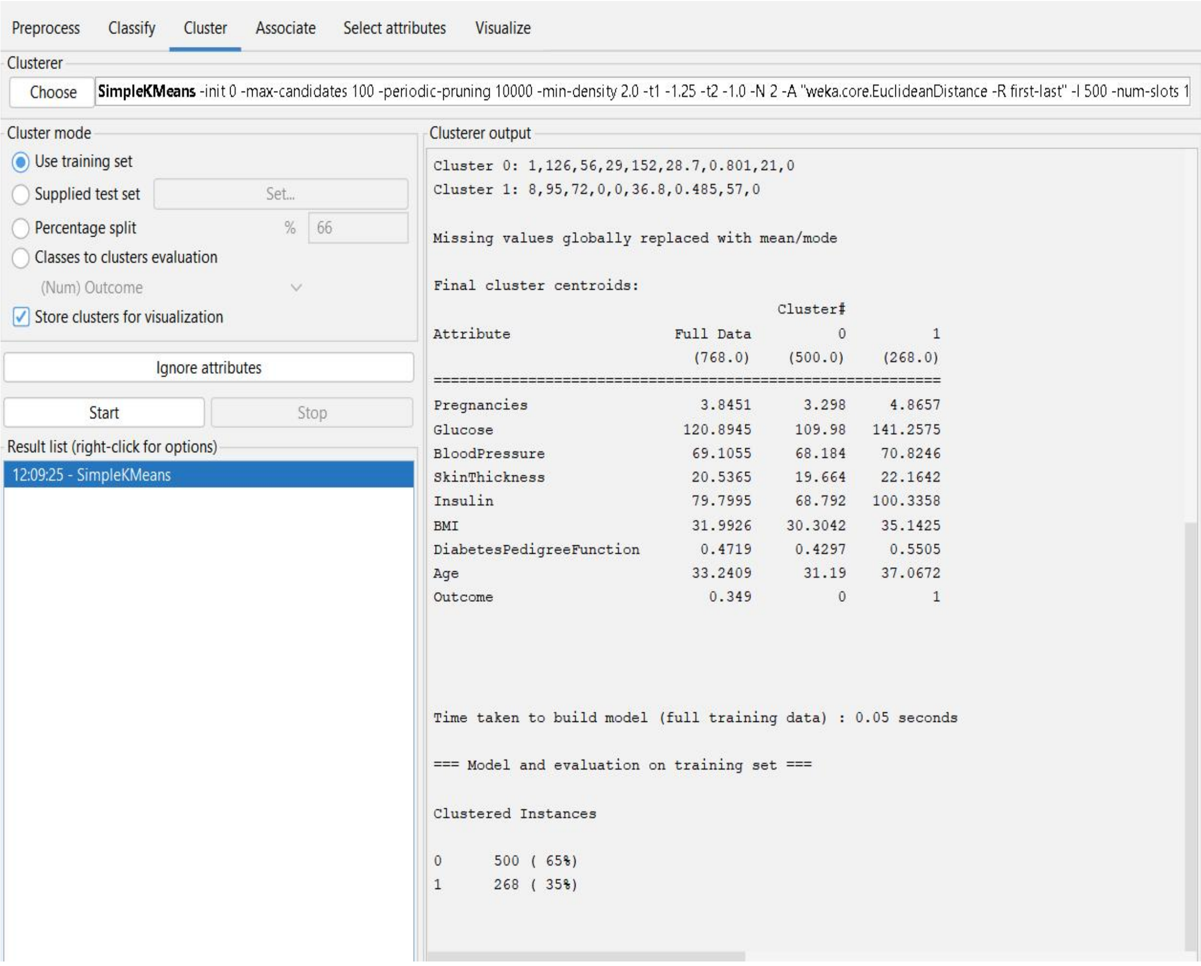
Statistic	Value
Minimum	0
Maximum	199
Mean	120.895
StdDev	31.973

Class: Outcome (Num)

2. K-MEANS CLUSTER ANALYSIS USING WEKA.

K-Means Clusters:

L-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.



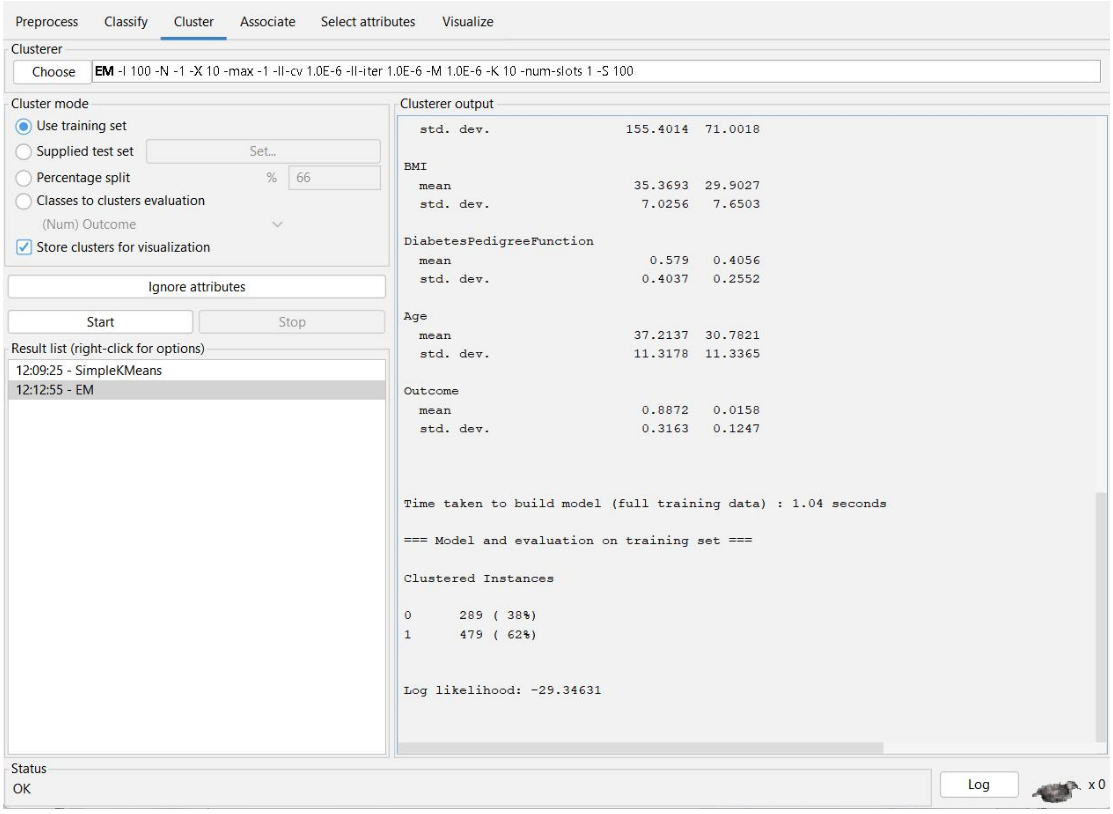
SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) Outcome
☑ Store clusters for visualization

12:09:25 - SimpleKMeans

```
Cluster 0: 1,126,56,29,152,28.7,0.801,21,0
Cluster 1: 8,95,72,0,0,36.8,0.485,57,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute                   Full Data          0          1
                              (768.0)    (500.0)    (268.0)
===========================================================
Pregnancies                    3.8451      3.298     4.8657
Glucose                      120.8945     109.98   141.2575
BloodPressure                 69.1055     68.184    70.8246
SkinThickness                 20.5365     19.664    22.1642
Insulin                       79.7995     68.792   100.3358
BMI                           31.9926    30.3042    35.1425
DiabetesPedigreeFunction       0.4719     0.4297     0.5505
Age                           33.2409      31.19    37.0672
Outcome                         0.349          0          1




Time taken to build model (full training data) : 0.05 seconds

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

Clustered Instances

0      500 ( 65%)
1      268 ( 35%)
```

3. DATA ANALYSIS BY EXPECTATION MAXIMISATION ALGORITHM USING WEKA.



Clusterer: 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

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Num) Outcome
- ☑ Store clusters for visualization

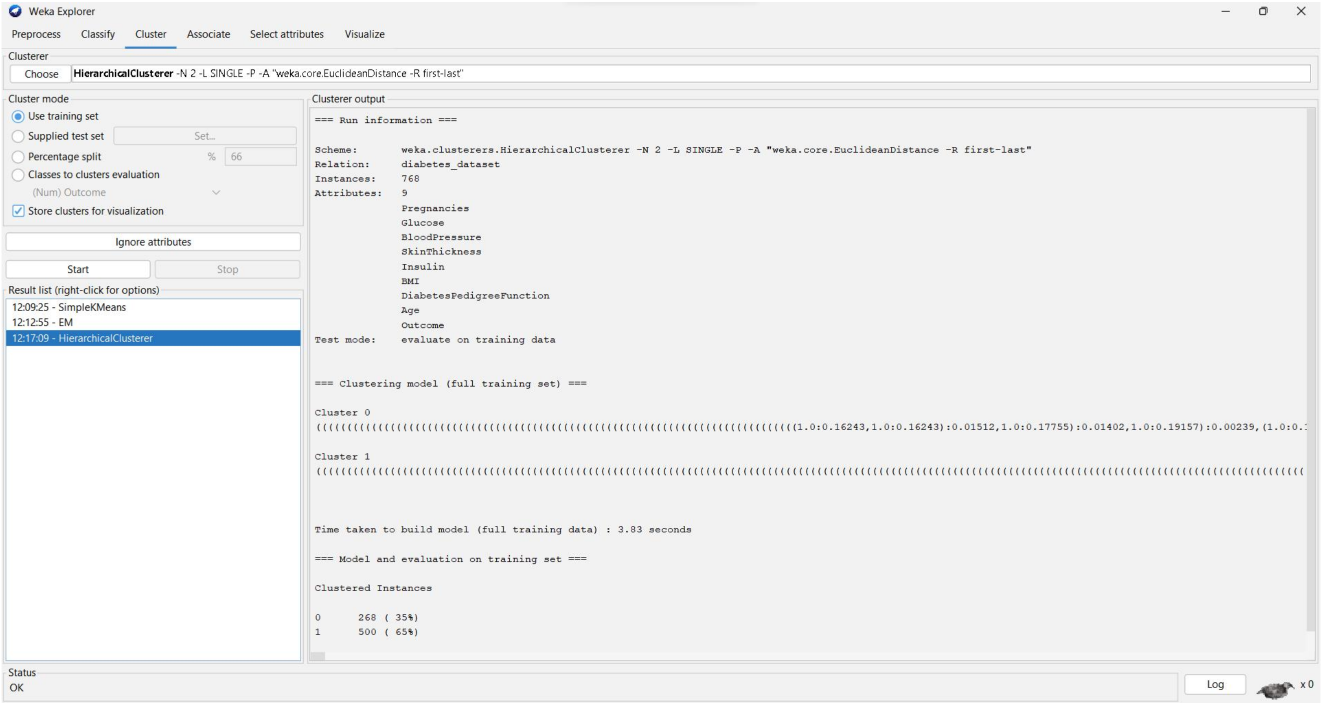
Result list:

- 12:09:25 - SimpleKMeans
- 12:12:55 - EM

```
  std. dev.                155.4014  71.0018

BMI
  mean                      35.3693  29.9027
  std. dev.                  7.0256   7.6503

DiabetesPedigreeFunction
  mean                        0.579   0.4056
  std. dev.                  0.4037   0.2552

Age
  mean                      37.2137  30.7821
  std. dev.                 11.3178  11.3365

Outcome
  mean                       0.8872   0.0158
  std. dev.                  0.3163   0.1247



Time taken to build model (full training data) : 1.04 seconds

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

Clustered Instances

0       289 ( 38%)
1       479 ( 62%)


Log likelihood: -29.34631
```

Status: OK

4.DATA ANALYSIS BY COBWEB-HIERARCHAL CLUSTERING ALGORITHM USING WEKA.



Clusterer: HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

- ● Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ○ Classes to clusters evaluation
- (Num) Outcome
- ☑ Store clusters for visualization

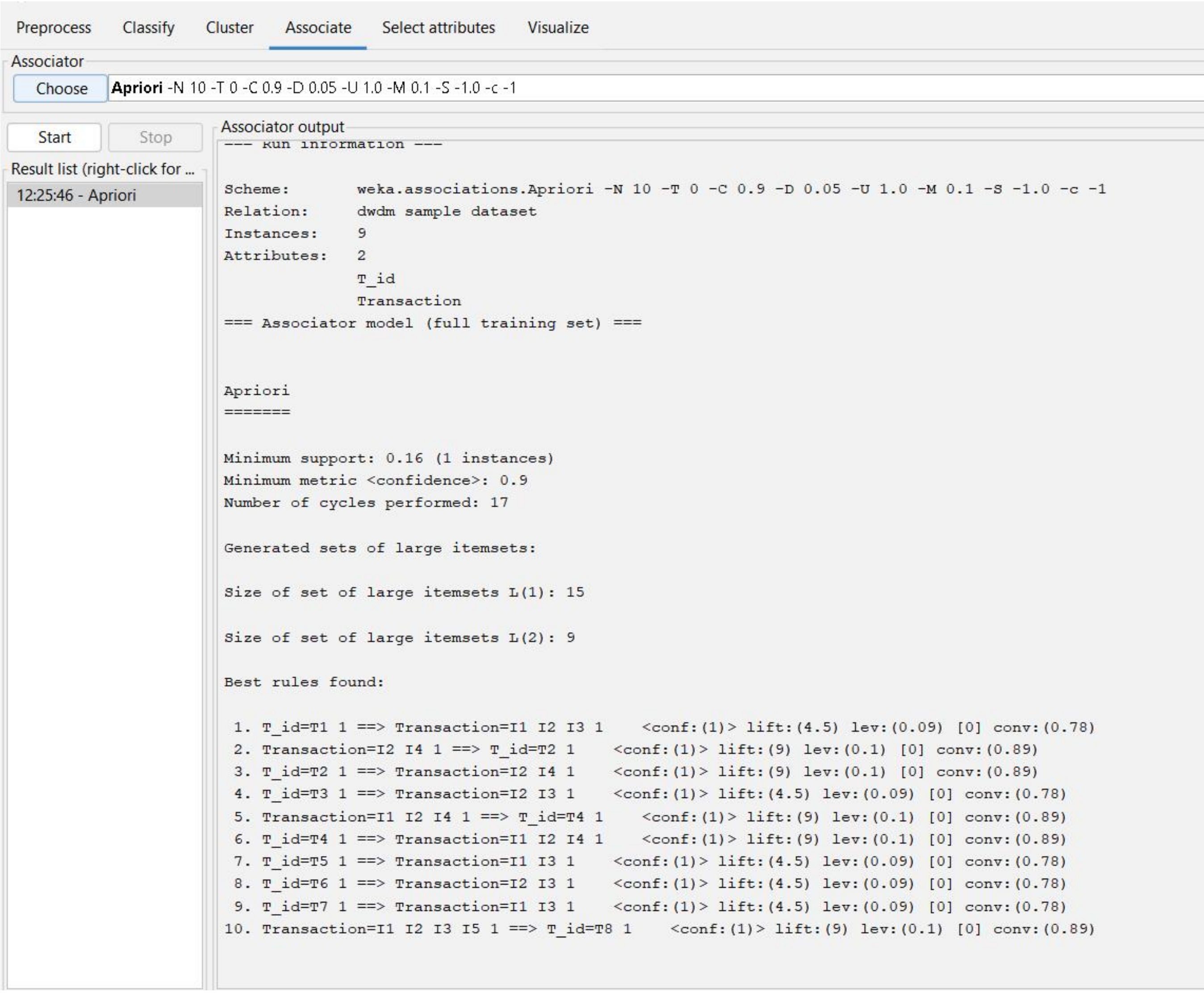
Result list:

- 12:09:25 - SimpleKMeans
- 12:12:55 - EM
- 12:17:09 - HierarchicalClusterer

```
=== Run information ===

Scheme:       weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Relation:     diabetes_dataset
Instances:    768
Attributes:   9
              Pregnancies
              Glucose
              BloodPressure
              SkinThickness
              Insulin
              BMI
              DiabetesPedigreeFunction
              Age
              Outcome
Test mode:    evaluate on training data


=== Clustering model (full training set) ===

Cluster 0
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((1.0:0.16243,1.0:0.16243):0.01512,1.0:0.17755):0.01402,1.0:0.19157):0.00239,(1.0:0.1

Cluster 1
((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((((



Time taken to build model (full training data) : 3.83 seconds

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

Clustered Instances

0       268 ( 35%)
1       500 ( 65%)
```

Status: OK

5 . KNOWLEDGE MINING USING ASSOCIATION RULE USING WEKA.



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

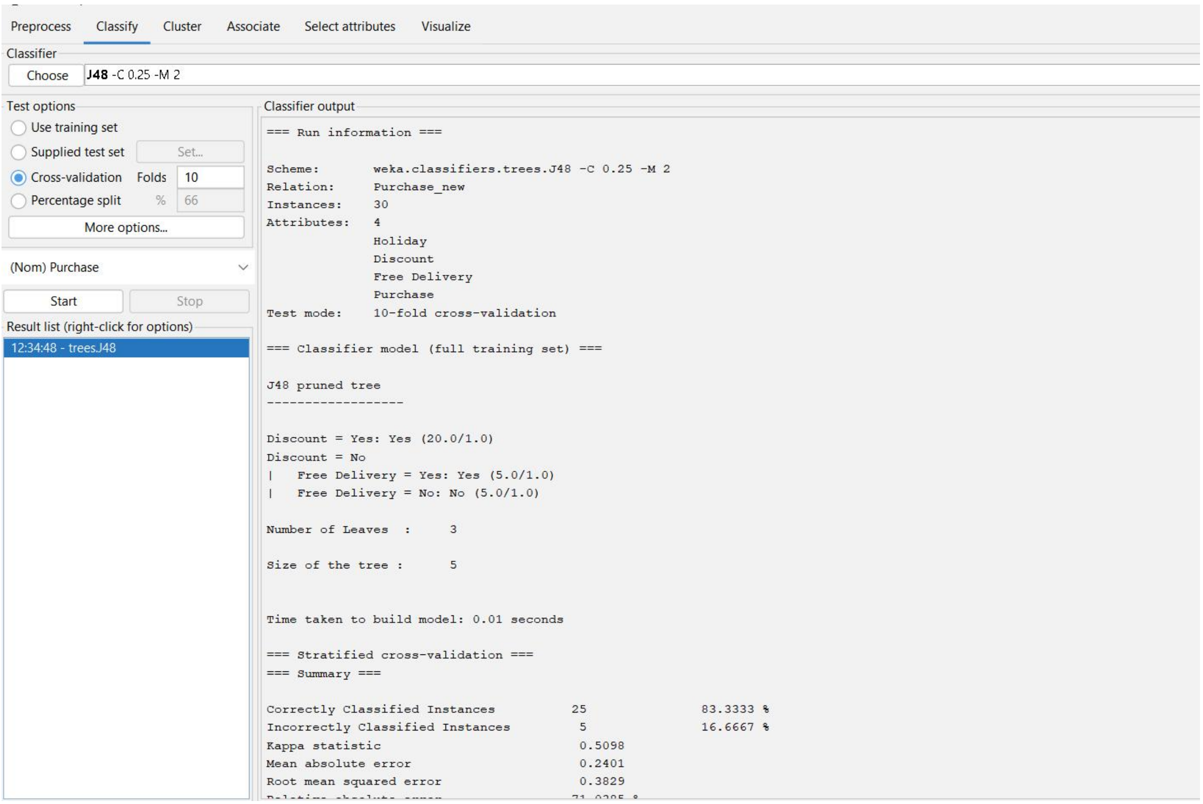
```
=== Run information ===

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:     dwdm sample dataset
Instances:    9
Attributes:   2
              T_id
              Transaction
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.16 (1 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 15

Size of set of large itemsets L(2): 9

Best rules found:

 1. T_id=T1 1 ==> Transaction=I1 I2 I3 1    <conf:(1)> lift:(4.5) lev:(0.09) [0] conv:(0.78)
 2. Transaction=I2 I4 1 ==> T_id=T2 1    <conf:(1)> lift:(9) lev:(0.1) [0] conv:(0.89)
 3. T_id=T2 1 ==> Transaction=I2 I4 1    <conf:(1)> lift:(9) lev:(0.1) [0] conv:(0.89)
 4. T_id=T3 1 ==> Transaction=I2 I3 1    <conf:(1)> lift:(4.5) lev:(0.09) [0] conv:(0.78)
 5. Transaction=I1 I2 I4 1 ==> T_id=T4 1    <conf:(1)> lift:(9) lev:(0.1) [0] conv:(0.89)
 6. T_id=T4 1 ==> Transaction=I1 I2 I4 1    <conf:(1)> lift:(9) lev:(0.1) [0] conv:(0.89)
 7. T_id=T5 1 ==> Transaction=I1 I3 1    <conf:(1)> lift:(4.5) lev:(0.09) [0] conv:(0.78)
 8. T_id=T6 1 ==> Transaction=I2 I3 1    <conf:(1)> lift:(4.5) lev:(0.09) [0] conv:(0.78)
 9. T_id=T7 1 ==> Transaction=I1 I3 1    <conf:(1)> lift:(4.5) lev:(0.09) [0] conv:(0.78)
10. Transaction=I1 I2 I3 I5 1 ==> T_id=T8 1    <conf:(1)> lift:(9) lev:(0.1) [0] conv:(0.89)
```

6. PREDICTION OF CATEGORICAL DATA USING DECISION TREE ALGORITHM USING WEKA.

ABOUT DATASET:

A dataset named Purchase is used in this decision tree which contains attributes of holidays, discount, free delievery and purchase.

RESULT:



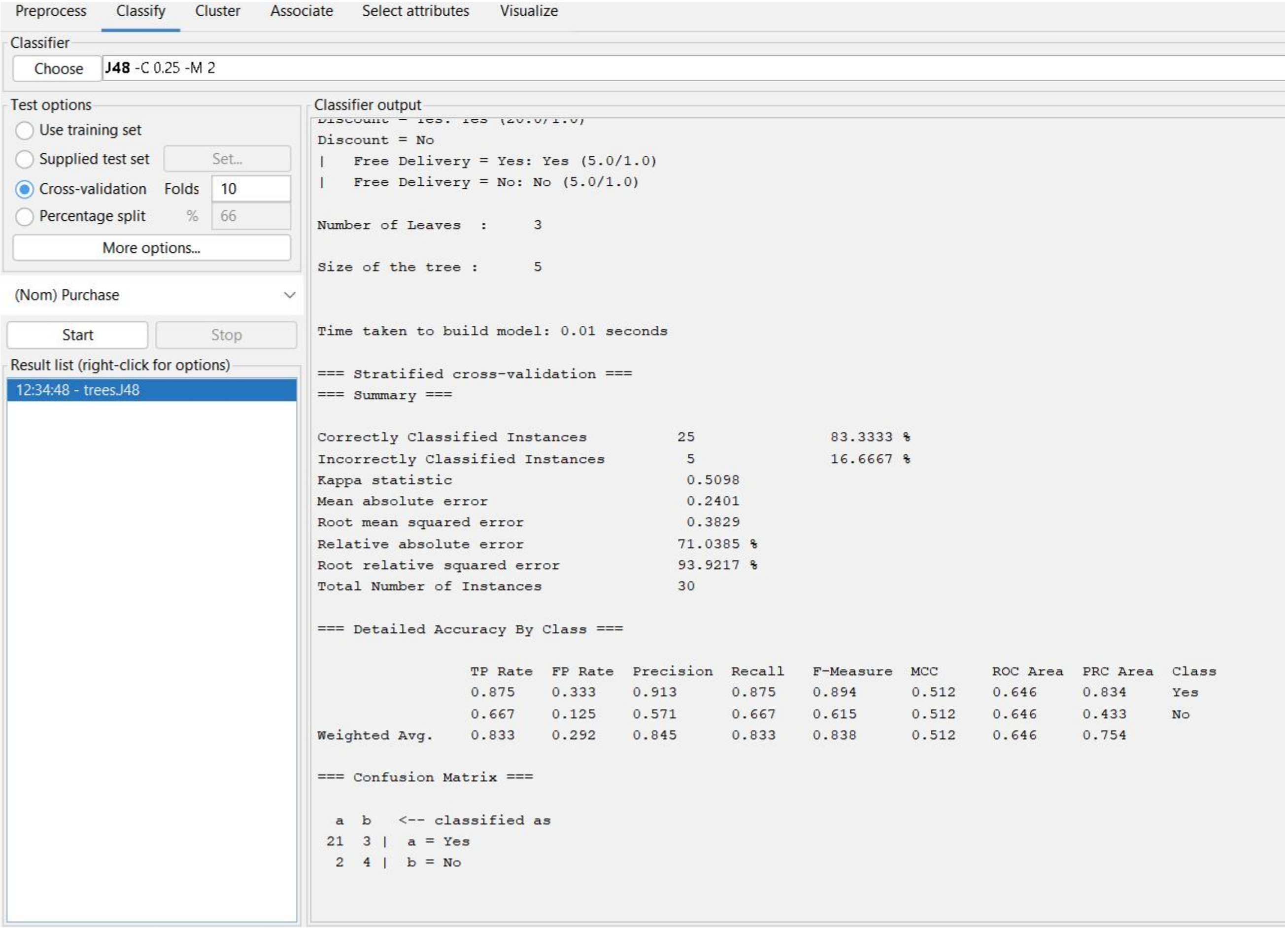
J48 -C 0.25 -M 2

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Purchase

12:34:48 - trees.J48

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Purchase_new
Instances:    30
Attributes:   4
              Holiday
              Discount
              Free Delivery
              Purchase
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

Discount = Yes: Yes (20.0/1.0)
Discount = No
|   Free Delivery = Yes: Yes (5.0/1.0)
|   Free Delivery = No: No (5.0/1.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          25               83.3333 %
Incorrectly Classified Instances         5               16.6667 %
Kappa statistic                          0.5098
Mean absolute error                      0.2401
Root mean squared error                  0.3829
```



J48 -C 0.25 -M 2

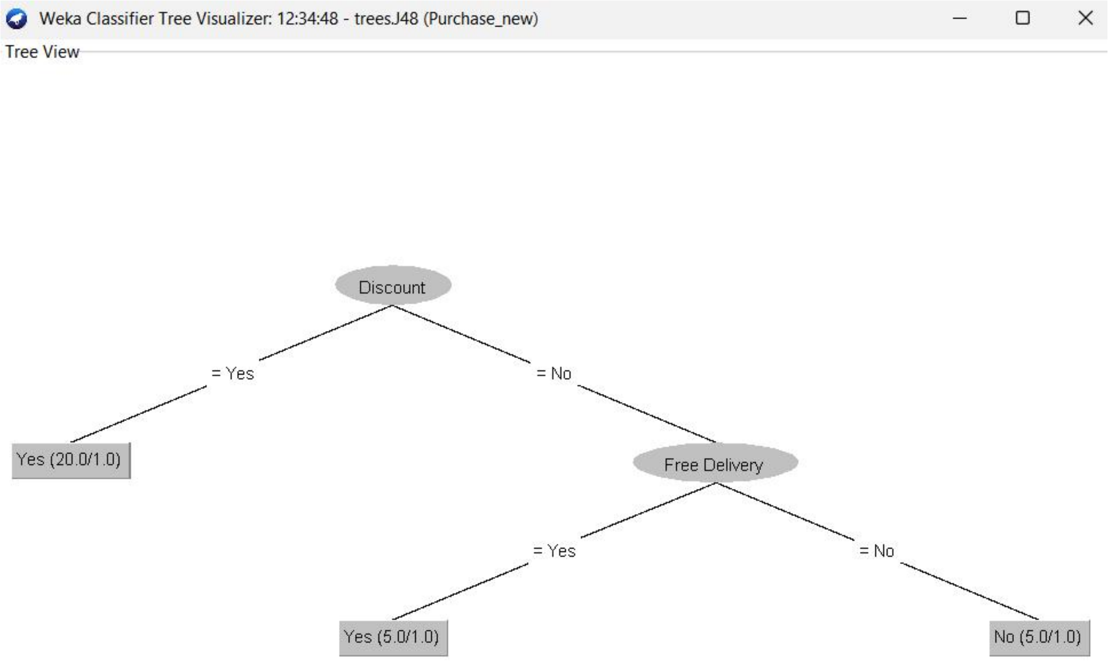
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Purchase

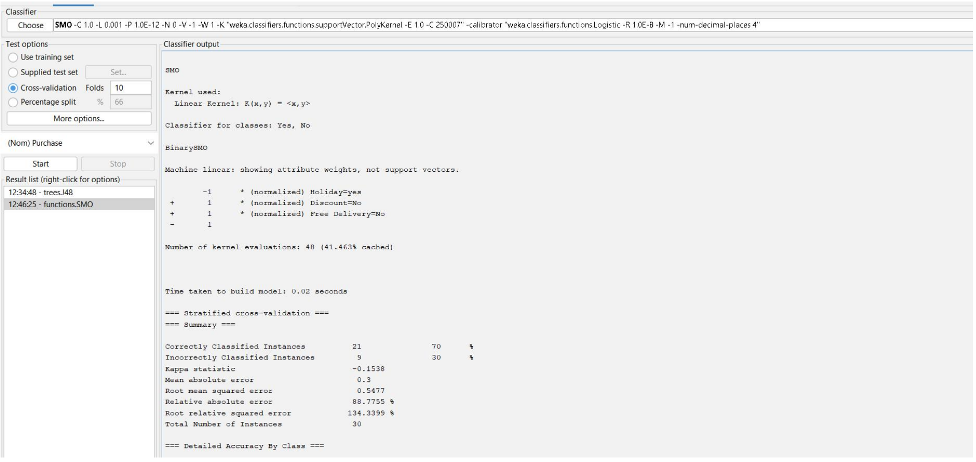
12:34:48 - trees.J48

```
Discount = No
|   Free Delivery = Yes: Yes (5.0/1.0)
|   Free Delivery = No: No (5.0/1.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          25               83.3333 %
Incorrectly Classified Instances         5               16.6667 %
Kappa statistic                          0.5098
Mean absolute error                      0.2401
Root mean squared error                  0.3829
```

```
Relative absolute error                 71.0385 %
Root relative squared error             93.9217 %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.875    0.333    0.913      0.875    0.894      0.512    0.646     0.834     Yes
                 0.667    0.125    0.571      0.667    0.615      0.512    0.646     0.433     No
Weighted Avg.    0.833    0.292    0.845      0.833    0.838      0.512    0.646     0.754

=== Confusion Matrix ===

  a  b   <-- classified as
 21  3 |  a = Yes
  2  4 |  b = No
```



7. PREDICTION OF CATEGORICAL DATA USING SMO ALGORITHM USING WEKA.



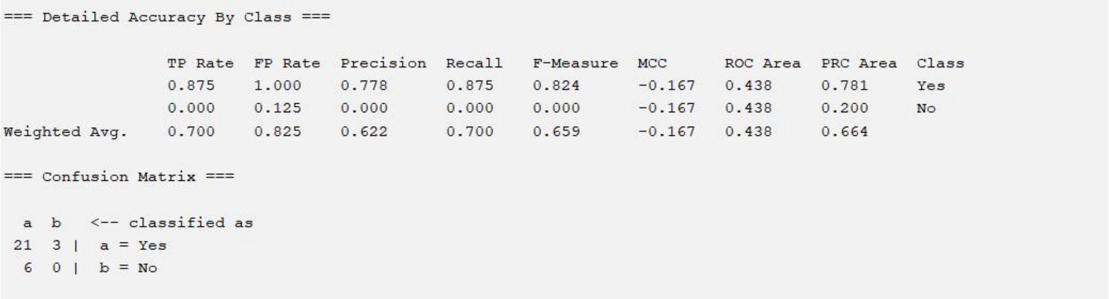
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

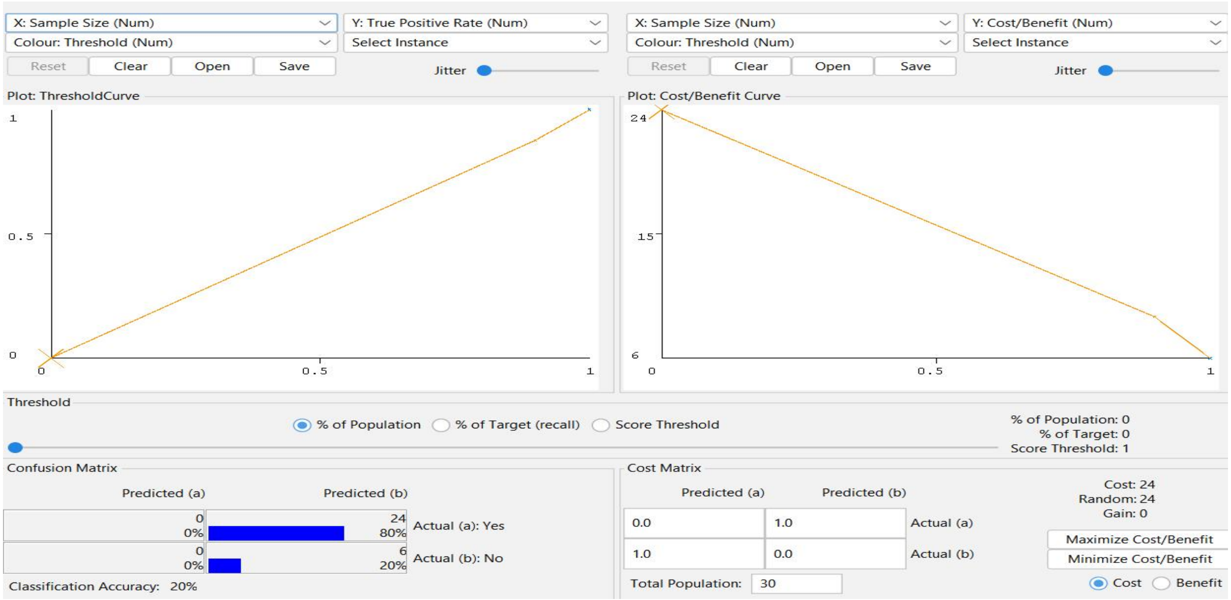
(Nom) Purchase

12:34:48 - trees.J48
12:46:25 - functions.SMO

```
SMO

Kernel used:
  Linear Kernel: K(x,y) = <x,y>

Classifier for classes: Yes, No

BinarySMO

Machine linear: showing attribute weights, not support vectors.

        -1       * (normalized) Holiday=yes
 +       1       * (normalized) Discount=No
 +       1       * (normalized) Free Delivery=No
 -       1

Number of kernel evaluations: 48 (41.463% cached)



Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          21               70      %
Incorrectly Classified Instances         9               30      %
Kappa statistic                         -0.1538
Mean absolute error                      0.3
Root mean squared error                  0.5477
Relative absolute error                 88.7755 %
Root relative squared error            134.3399 %
Total Number of Instances               30

=== Detailed Accuracy By Class ===
```



```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.875    1.000    0.778      0.875    0.824      -0.167   0.438     0.781     Yes
                 0.000    0.125    0.000      0.000    0.000      -0.167   0.438     0.200     No
Weighted Avg.    0.700    0.825    0.622      0.700    0.659      -0.167   0.438     0.664

=== Confusion Matrix ===

  a  b   <-- classified as
 21  3 |  a = Yes
  6  0 |  b = No
```



● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 1

Confusion Matrix

Predicted (a)	Predicted (b)	
0 0%	24 80%	Actual (a): Yes
0 0%	6 20%	Actual (b): No

Classification Accuracy: 20%

Cost Matrix

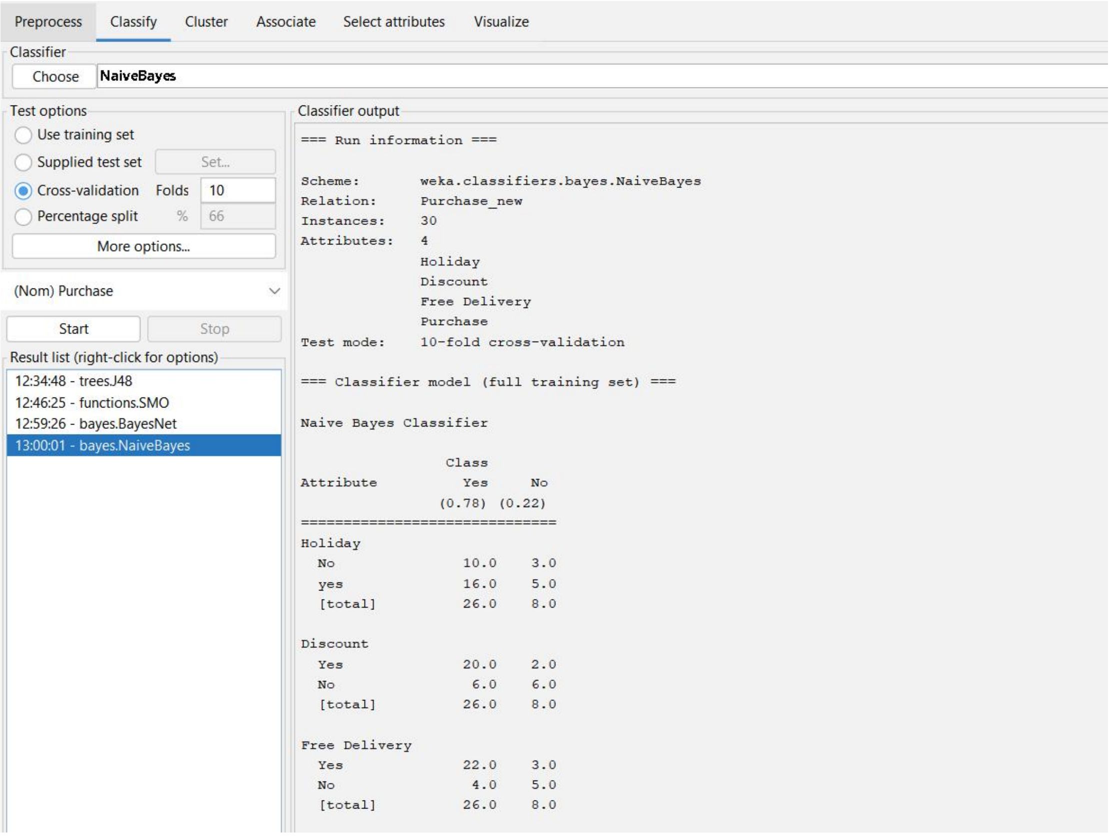
Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 30

Cost: 24
Random: 24
Gain: 0

● Cost ○ Benefit

8. PREDICTION OF CATERGORICAL DATA USING BAYESIAN ALGORITHM



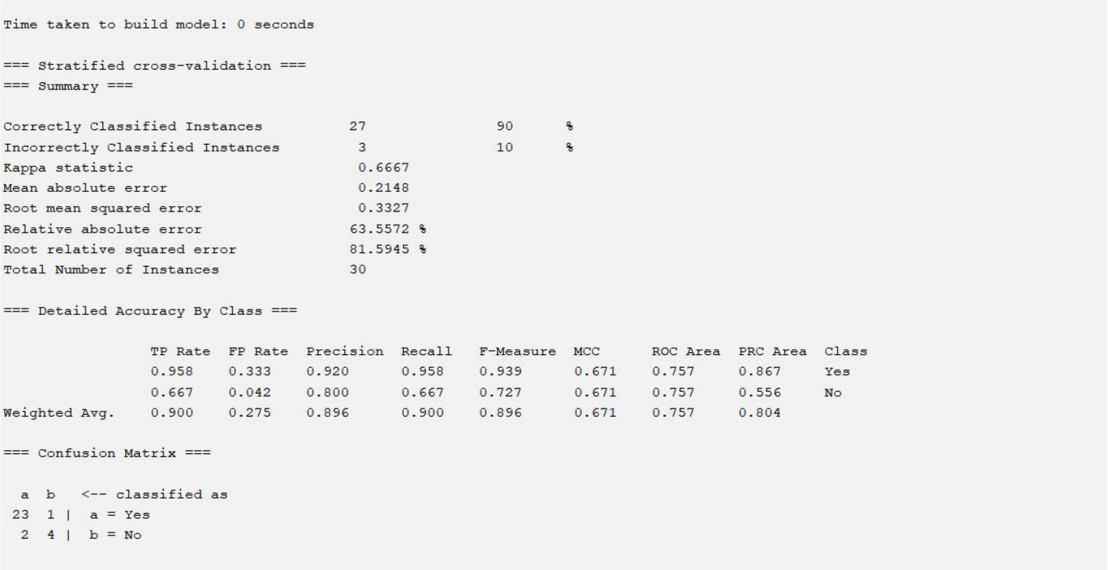
NaiveBayes

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) Purchase

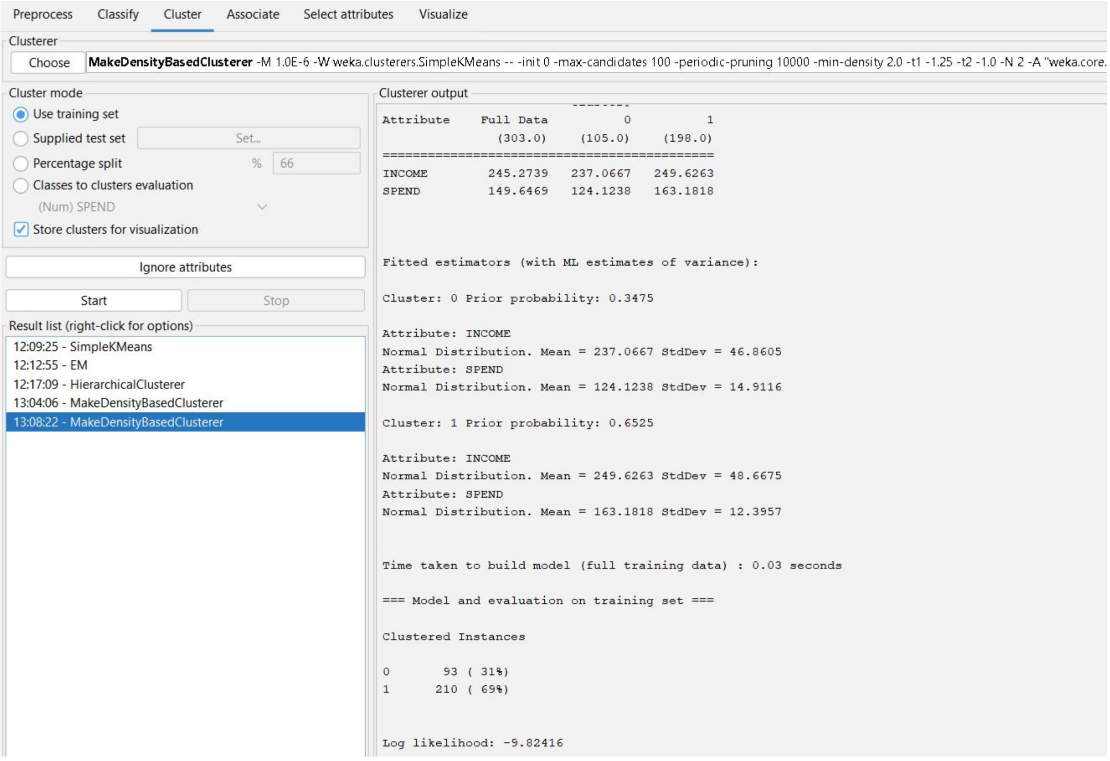
12:34:48 - trees.J48
12:46:25 - functions.SMO
12:59:26 - bayes.BayesNet
13:00:01 - bayes.NaiveBayes

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayes
Relation:     Purchase_new
Instances:    30
Attributes:   4
              Holiday
              Discount
              Free Delivery
              Purchase
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

Naive Bayes Classifier

                  Class
Attribute           Yes      No
                  (0.78)  (0.22)
===============================
Holiday
  No               10.0     3.0
  yes              16.0     5.0
  [total]          26.0     8.0

Discount
  Yes              20.0     2.0
  No                6.0     6.0
  [total]          26.0     8.0

Free Delivery
  Yes              22.0     3.0
  No                4.0     5.0
  [total]          26.0     8.0
```



```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          27               90      %
Incorrectly Classified Instances         3               10      %
Kappa statistic                          0.6667
Mean absolute error                      0.2148
Root mean squared error                  0.3327
Relative absolute error                 63.5572 %
Root relative squared error             81.5945 %
Total Number of Instances               30

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.958    0.333    0.920      0.958    0.939      0.671    0.757     0.867     Yes
                 0.667    0.042    0.800      0.667    0.727      0.671    0.757     0.556     No
Weighted Avg.    0.900    0.275    0.896      0.900    0.896      0.671    0.757     0.804

=== Confusion Matrix ===

  a  b   <-- classified as
 23  1 |  a = Yes
  2  4 |  b = No
```

9. DATA ANALYSIS BY DENSITY BASED CLUSTERING ALGORITHM USING WEKA.



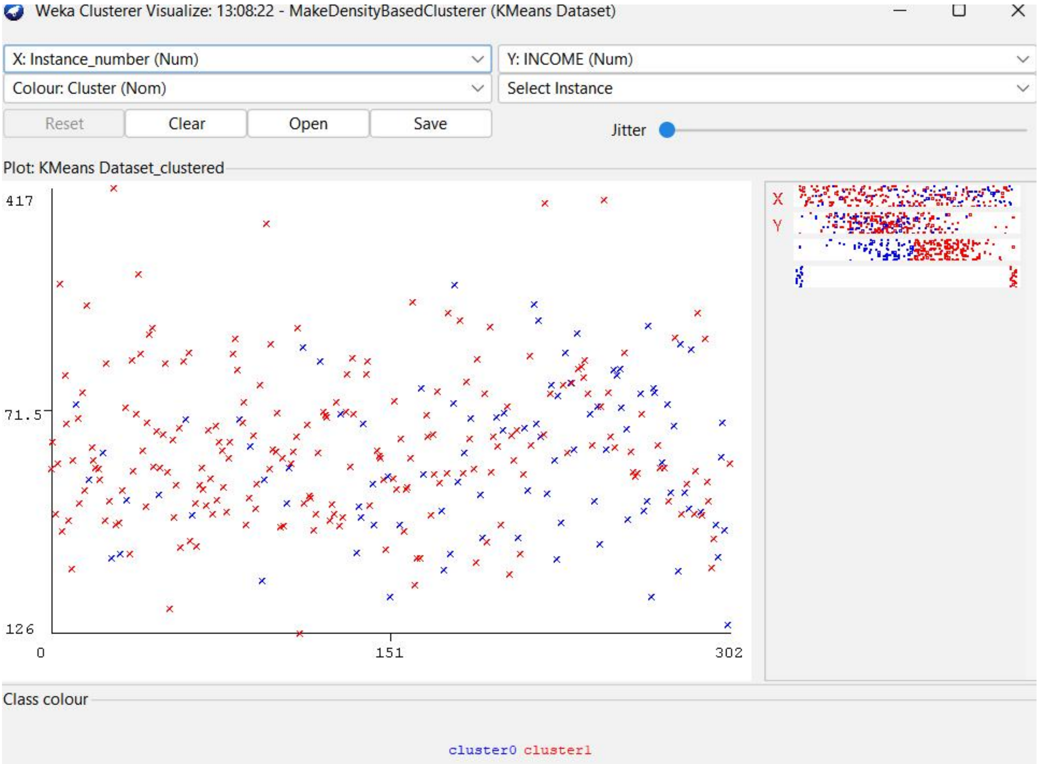
MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.

● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) SPEND
☑ Store clusters for visualization

Result list (right-click for options)

12:09:25 - SimpleKMeans
12:12:55 - EM
12:17:09 - HierarchicalClusterer
13:04:06 - MakeDensityBasedClusterer
13:08:22 - MakeDensityBasedClusterer

```
Attribute    Full Data         0          1
               (303.0)    (105.0)    (198.0)
============================================
INCOME        245.2739   237.0667   249.6263
SPEND         149.6469   124.1238   163.1818



Fitted estimators (with ML estimates of variance):

Cluster: 0 Prior probability: 0.3475

Attribute: INCOME
Normal Distribution. Mean = 237.0667 StdDev = 46.8605
Attribute: SPEND
Normal Distribution. Mean = 124.1238 StdDev = 14.9116

Cluster: 1 Prior probability: 0.6525

Attribute: INCOME
Normal Distribution. Mean = 249.6263 StdDev = 48.6675
Attribute: SPEND
Normal Distribution. Mean = 163.1818 StdDev = 12.3957


Time taken to build model (full training data) : 0.03 seconds

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

Clustered Instances

0       93 ( 31%)
1      210 ( 69%)


Log likelihood: -9.82416
```



Weka Clusterer Visualize: 13:08:22 - MakeDensityBasedClusterer (KMeans Dataset)

X: Instance_number (Num)

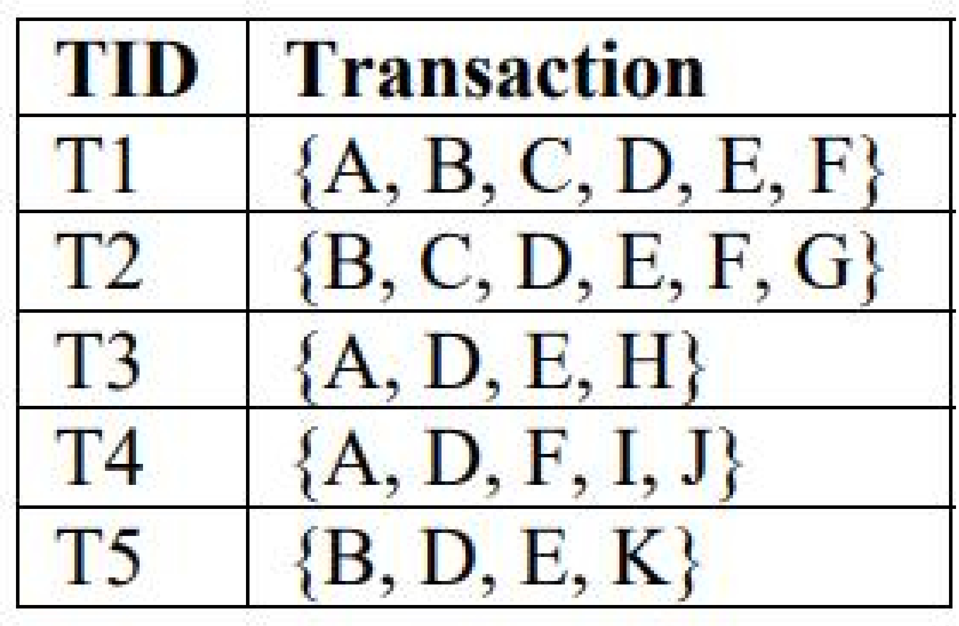
Y: INCOME (Num)

Colour: Cluster (Nom)

Select Instance

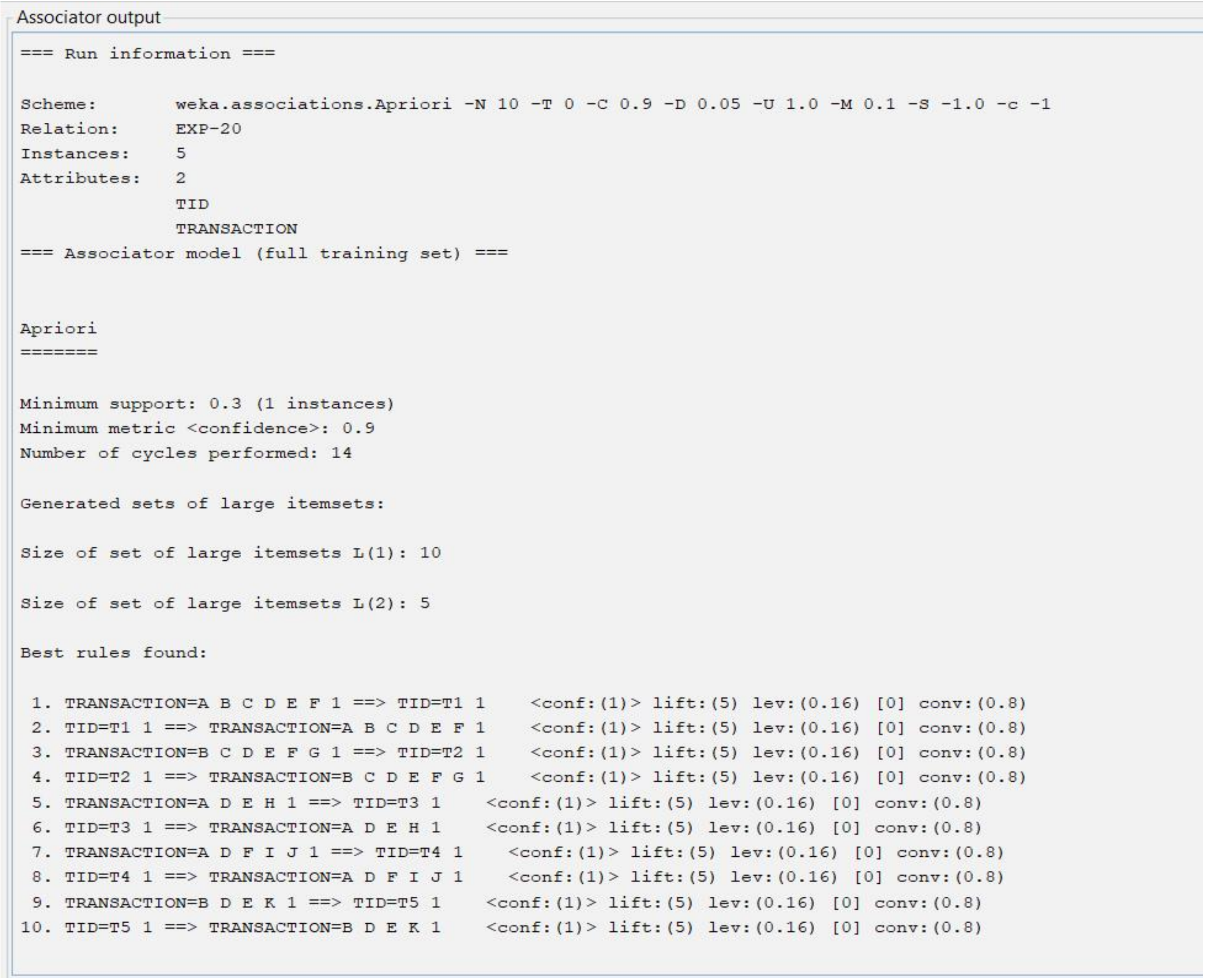
Plot: KMeans Dataset_clustered

10.GIVING THE FOLLOWING DATABASE WITH 5 TRANSACTIONS AND A MINIMUM SUPPORT THRESHOLD OF 60% AND A MINIMUM CONFIDENCE THRESHOLD OF 80%, FIND ALL FREQUENT ITEMSETS USING (A) APRIORI AND (B) FP-GROWTH.



TID	Transaction
T1	{A, B, C, D, E, F}
T2	{B, C, D, E, F, G}
T3	{A, D, E, H}
T4	{A, D, F, I, J}
T5	{B, D, E, K}

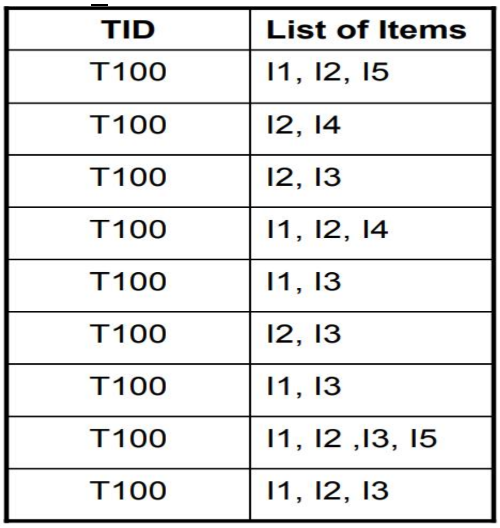
RESULT:



```
=== 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:     EXP-20
Instances:    5
Attributes:   2
              TID
              TRANSACTION
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.3 (1 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 14

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10

Size of set of large itemsets L(2): 5

Best rules found:

 1. TRANSACTION=A B C D E F 1 ==> TID=T1 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 2. TID=T1 1 ==> TRANSACTION=A B C D E F 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 3. TRANSACTION=B C D E F G 1 ==> TID=T2 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 4. TID=T2 1 ==> TRANSACTION=B C D E F G 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 5. TRANSACTION=A D E H 1 ==> TID=T3 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 6. TID=T3 1 ==> TRANSACTION=A D E H 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 7. TRANSACTION=A D F I J 1 ==> TID=T4 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 8. TID=T4 1 ==> TRANSACTION=A D F I J 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
 9. TRANSACTION=B D E K 1 ==> TID=T5 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
10. TID=T5 1 ==> TRANSACTION=B D E K 1    <conf:(1)> lift:(5) lev:(0.16) [0] conv:(0.8)
```

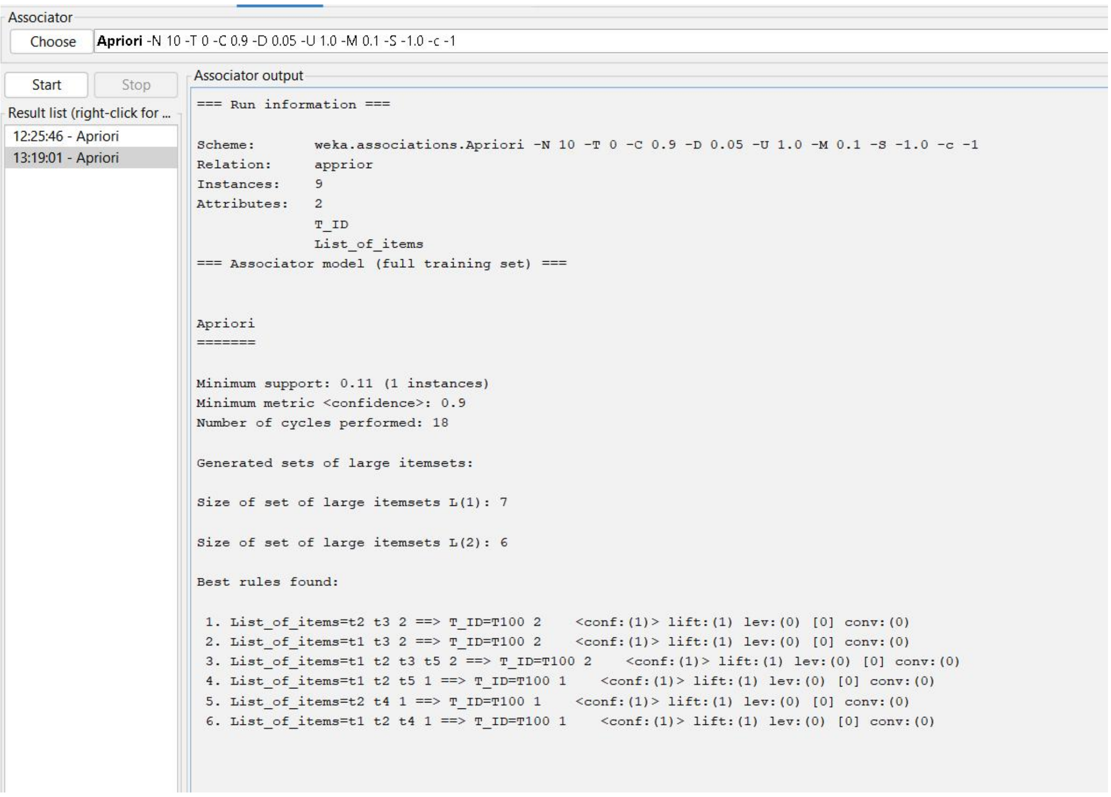
11.THE 'DATABASE' BELOW HAS NINE TRANSACTIONS. WHAT ASSOCIATION RULES CAN BE FOUND IN THIS SET, IF THE MINIMUM SUPPORT (I.E COVERAGE) IS 60% AND THE MINIMUM CONFIDENCE (I.E. ACCURACY) IS 80% ?

TRANS_ID ITEMLIST



TID	List of Items
T100	I1, I2, I5
T100	I2, I4
T100	I2, I3
T100	I1, I2, I4
T100	I1, I3
T100	I2, I3
T100	I1, I3
T100	I1, I2 ,I3, I5
T100	I1, I2, I3

RESULT:



Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

```
=== 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:     apprior
Instances:    9
Attributes:   2
              T_ID
              List_of_items
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.11 (1 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 7

Size of set of large itemsets L(2): 6

Best rules found:

 1. List_of_items=t2 t3 2 ==> T_ID=T100 2    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. List_of_items=t1 t3 2 ==> T_ID=T100 2    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. List_of_items=t1 t2 t3 t5 2 ==> T_ID=T100 2    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. List_of_items=t1 t2 t5 1 ==> T_ID=T100 1    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 5. List_of_items=t2 t4 1 ==> T_ID=T100 1    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 6. List_of_items=t1 t2 t4 1 ==> T_ID=T100 1    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
```