Experiment 7 - Perform data Pre-processing task and Demonstrate performing Classification, Clustering, Association algorithm on data sets using data mining tool

Name:	**Shaikh Shadab**	**Rollno**	**:**	**17DCO74**
Class :	**TE.CO**	**Batch**	**:**	**B3**

#Theory

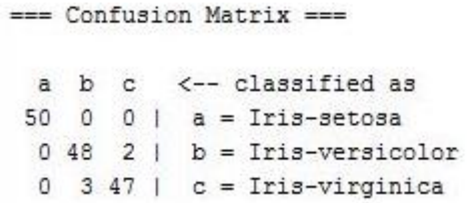
- Weka Tool

  Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization. Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive nature. The name is pronounced like this, and the bird sounds like this. Weka is open source software issued under the GNU General Public License. It is possible to apply Weka to process big data and perform deep learning!

- Testing methods
  - Use training set - Means we will test our knowledge on the same data we learned. Not very accepted because we can just make build our code to memorize the training instances (which will be in the test).
  - Supplied test set - it is an external file that we can use as training set. It can be used when we want/need to test the algorithm's knowledge against a specific test set.
  - Percentage split: Splits the data and separates x% of the data for learning and the rest of it for testing. It is useful when your algorithm is slow.
  - Cross-validation (CV): Works like many percentage splits. we fold the data in 10 folds (for example) and repeat 10 (because it is 10-folds) the following process: Use 9 folds for learning and leave 1-fold out for testing. Every time leaving a different fold for testing. This is the most used testing method in papers. They say "anything over 5 folds is acceptable", but no one has any good explanation for that.

- Confusion Matrix

  A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy identification of confusion between classes e.g. one class is commonly mislabelled as the other. Most performance measures are computed from the confusion matrix.



```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  3 47 |  c = Iris-virginica
```

  For example, from the above table we can say that the actual class was a i.e Iris-setosa and predicted class a count was 50, b is 0 and c is 0.similaryly the actual class was c i.e Iris-virginica and predicted class a count was 0 b is 3 and c is 47 (using CV 2 fold)

- True positive, True negative, False positive, False negative
  - When the actual class was true and we predict it also as true is TP
  - When the actual class was false and we predict it also as false is TN
  - When the actual class was false and we predict it as true is FP (Type I error)
  - When the actual class was true and we predict it as false is FN (Type II error)

  In simpler terms, A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class.

  A false positive is an outcome where the model incorrectly predicts the positive class. And a false negative is an outcome where the model incorrectly predicts the negative class.

  Real life example of TP (correctly identified)
  Umpire gives a Batsman NOT OUT when he is NOT OUT.
  Real life example of TN (correctly rejected)
  Umpire gives a Batsman OUT when he is OUT.
  Real life example of FP (Type I error) (incorrectly identified)
  Umpire gives a Batsman NOT OUT when he is OUT.
  Real life example of FN (Type II error) (incorrectly rejected)
  Umpire gives a Batsman OUT when he is NOT OUT.

- Sensitivity, Specificity, significance

  True positive rate (sensitivity) - measures the proportion of actual positives that are correctly identified as such

  True negative rate (Specificity) - measures the proportion of actual negatives that are correctly identified as such

  False positive rate (significance) - proportion of all negatives that still yield positive test outcomes, i.e., the conditional probability of a positive test result given an event that was not present.

- Precision and Recall

  Precision tries to answer such type of queries:
  What proportion of positive identifications was actually correct?

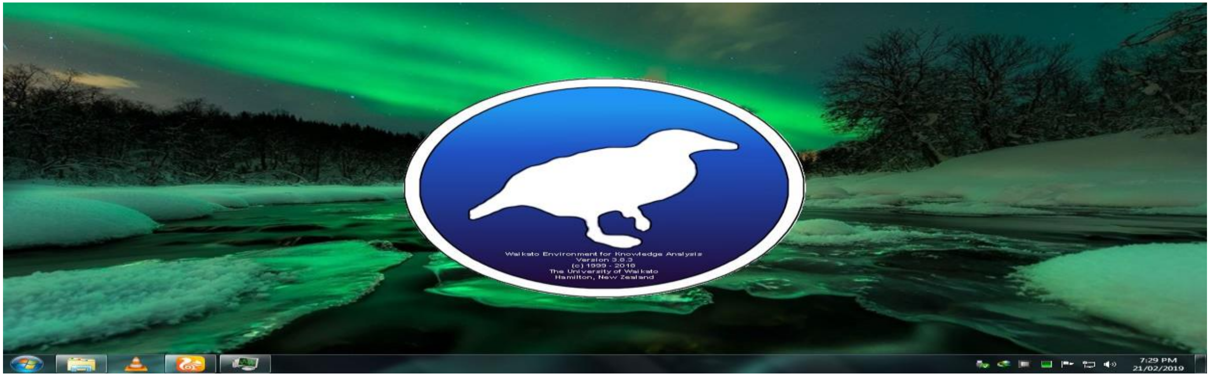
  $$\text{Precision} = \frac{TP}{TP + FP}$$

  Recall tries to answer such type of queries:
  What proportion of actual positives was identified correctly?

  $$\text{Recall} = \frac{TP}{TP + FN}$$

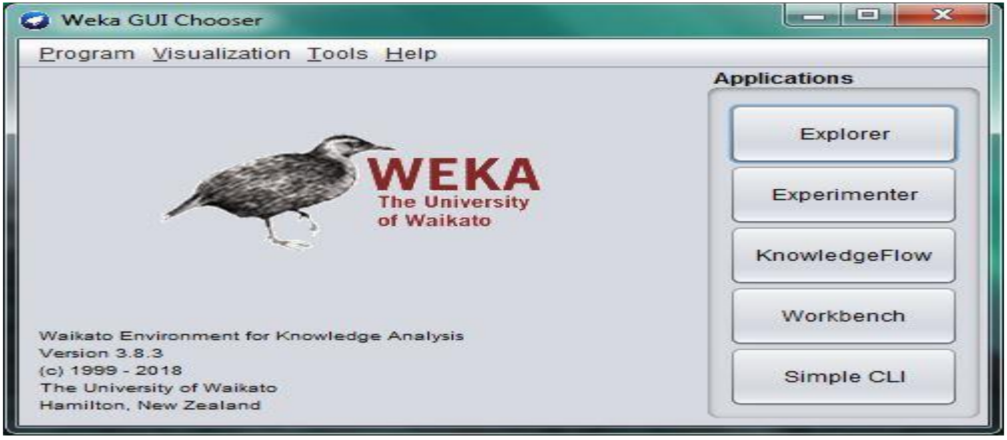
#Using Weka-tool for data pre-processing task

1. **Tool loading screen**



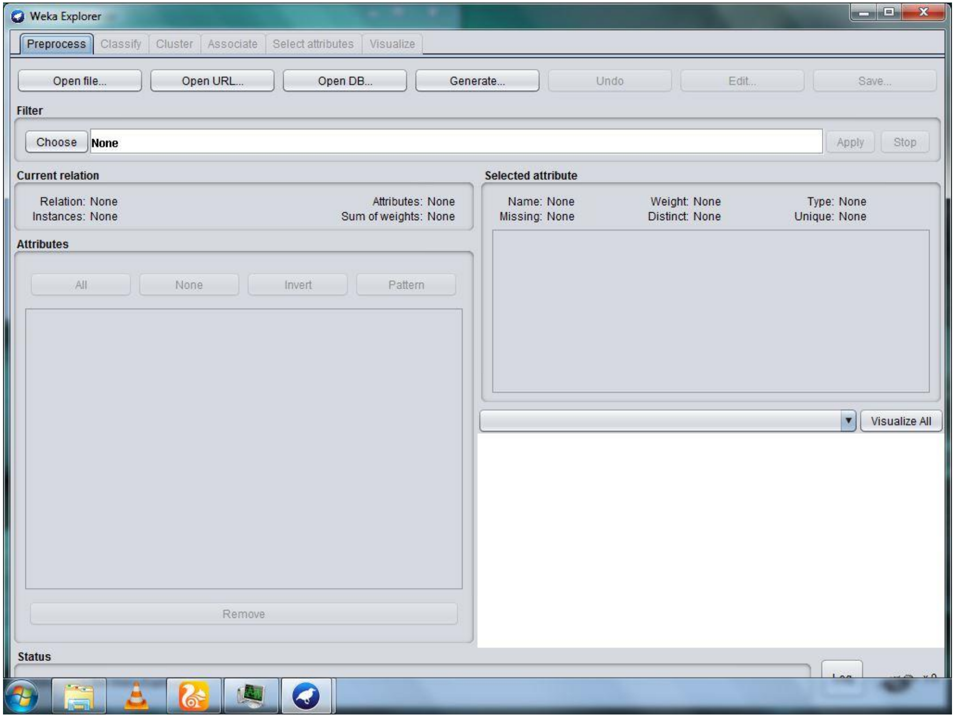
Waikato Environment for Knowledge Analysis
Version 3.8.3
(c) 1999 - 2018
The University of Waikato
Hamilton, New Zealand
7:29 PM
21/02/2019

2. **Tool initial screen, selecting explorer**



Waikato Environment for Knowledge Analysis
Version 3.8.3
(c) 1999 - 2018
The University of Waikato
Hamilton, New Zealand

3. **Explorer screen**



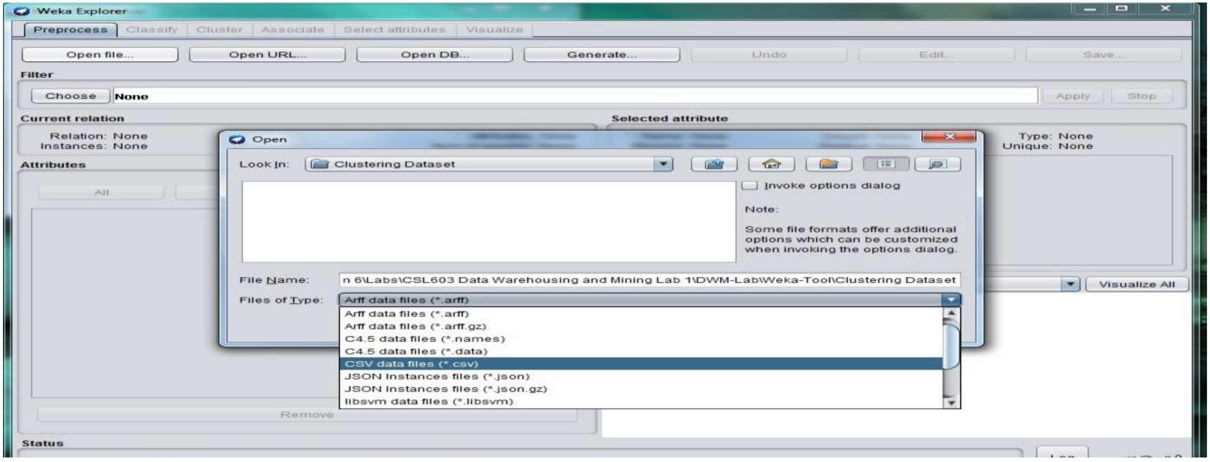
Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

#Performing Clustering

1. **Opening file, going to directory, selecting .csv format**



Look In: Clustering Dataset

File Name: n 6\Labs\CSL603 Data Warehousing and Mining Lab 1\DWM-Lab\Weka-Tool\Clustering Dataset

Files of Type: Arff data files (*.arff)

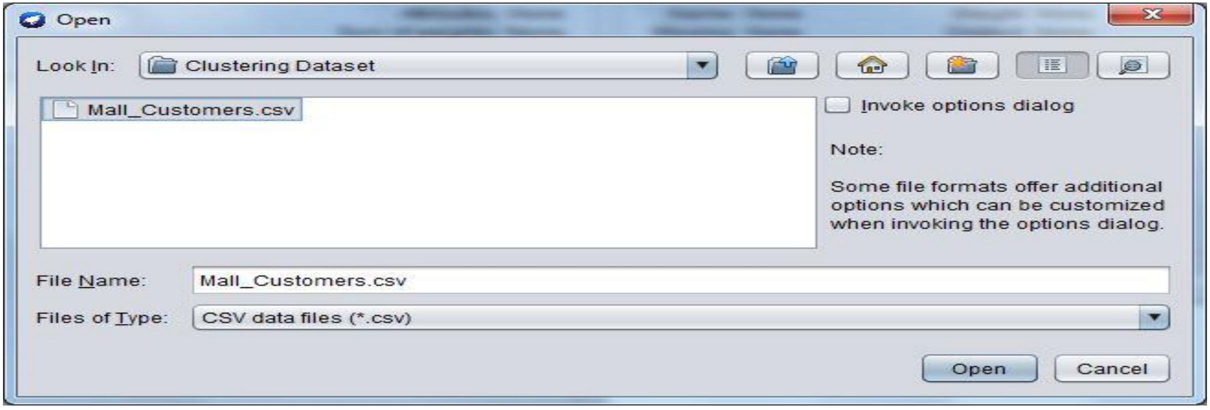
- Arff data files (*.arff)
- Arff data files (*.arff.gz)
- C4.5 data files (*.names)
- C4.5 data files (*.data)
- CSV data files (*.csv)
- JSON Instances files (*.json)
- JSON Instances files (*.json.gz)
- libsvm data files (*.libsvm)

□ Invoke options dialog

Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

2. **Selecting dataset for clustering (Mall_Customers.csv)**



Look In: Clustering Dataset

Mall_Customers.csv

□ Invoke options dialog

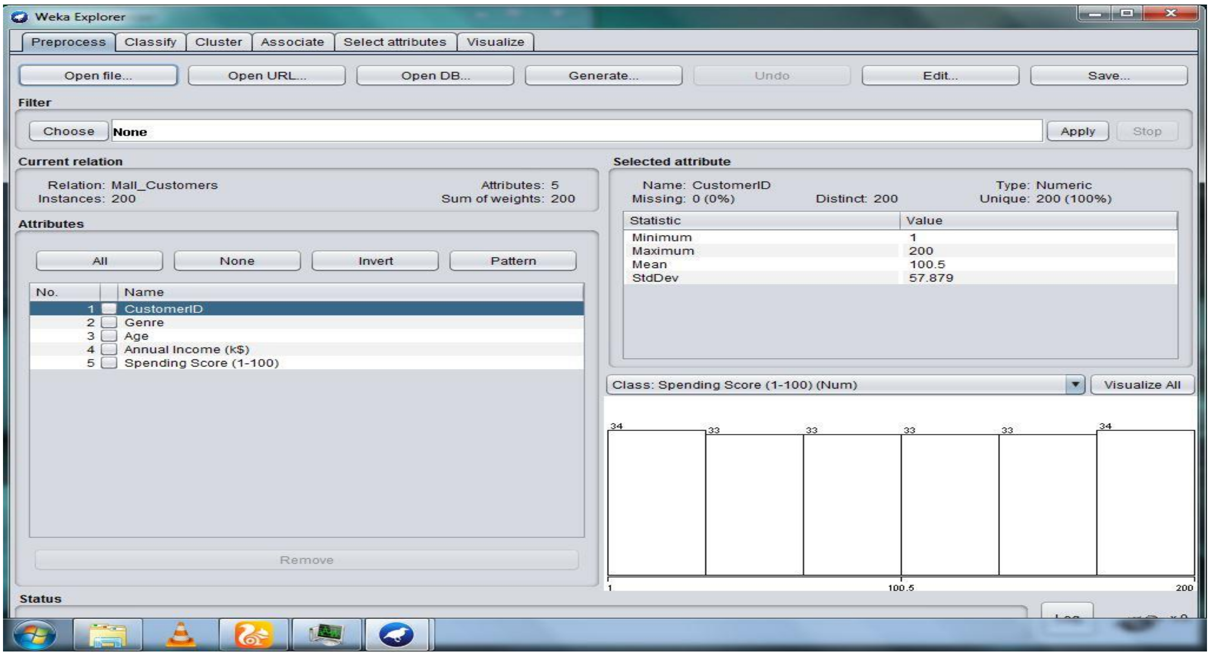
Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: Mall_Customers.csv

Files of Type: CSV data files (*.csv)

3. **Overview screen**



Current relation

Relation: Mall_Customers
Instances: 200
Attributes: 5
Sum of weights: 200

Filter: None

No.		Name
1	□	CustomerID
2	□	Genre
3	□	Age
4	□	Annual Income (k$)
5	□	Spending Score (1-100)

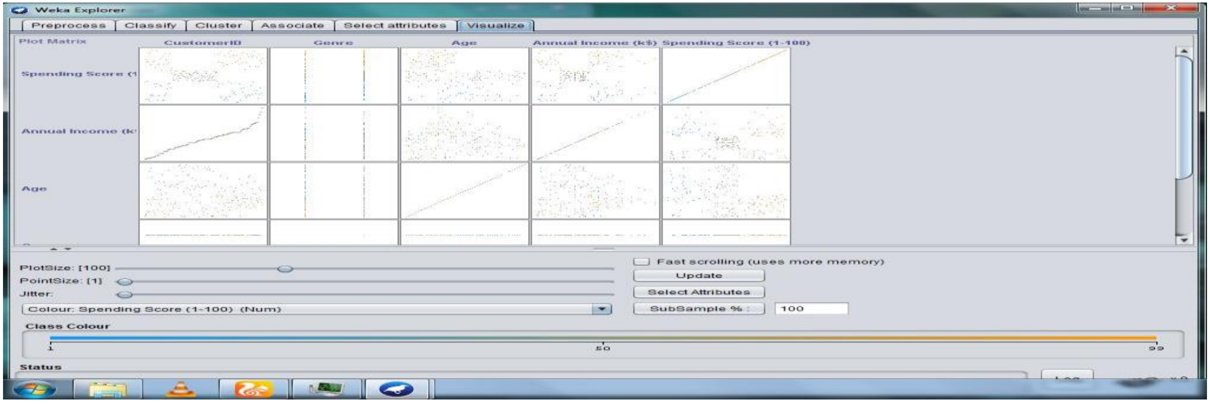
Selected attribute

Name: CustomerID
Missing: 0 (0%)
Distinct: 200
Type: Numeric
Unique: 200 (100%)

Statistic	Value
Minimum	1
Maximum	200
Mean	100.5
StdDev	57.879

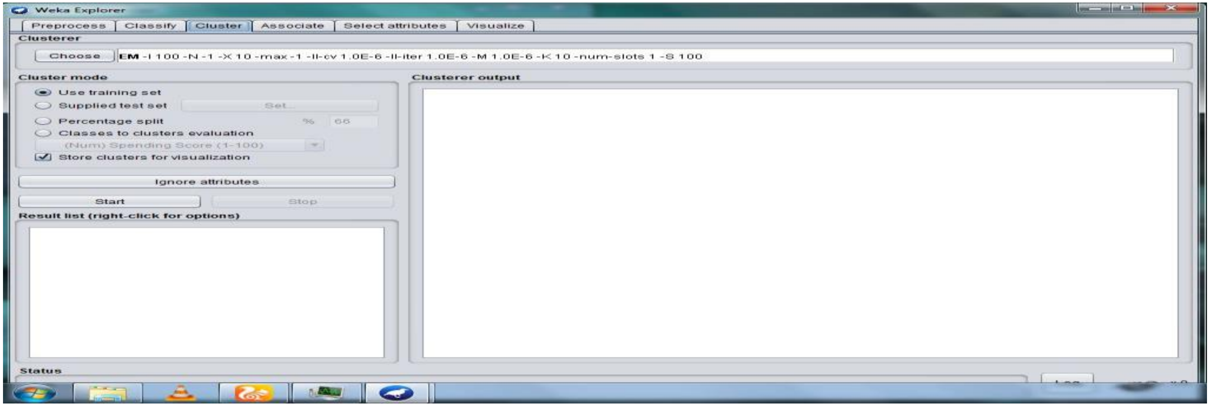
Class: Spending Score (1-100) (Num)

4. **Selecting visualise tab**



PlotSize: [100]
PointSize: [1]
Jitter:
Colour: Spending Score (1-100) (Num)
SubSample % : 100

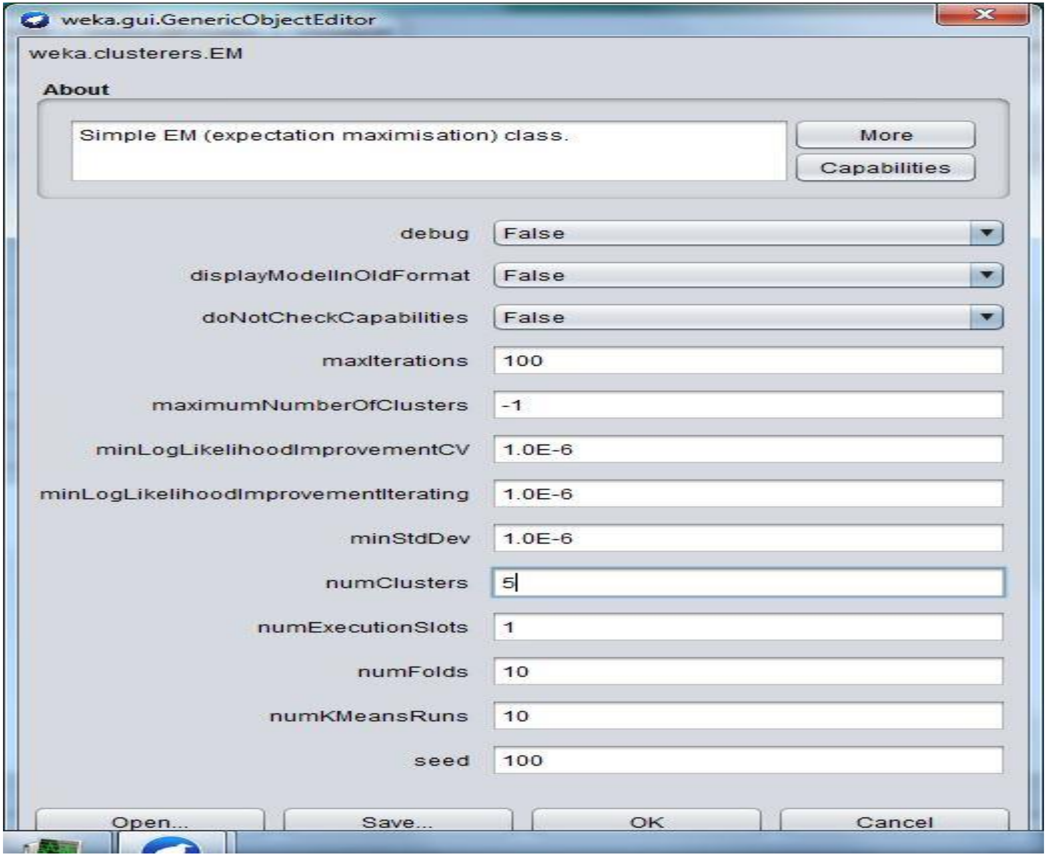
5. **selecting cluster tab and choosing use train dataset method**



```
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) Spending Score (1-100)
☑ Store clusters for visualization

6. **making number of cluster to 5**

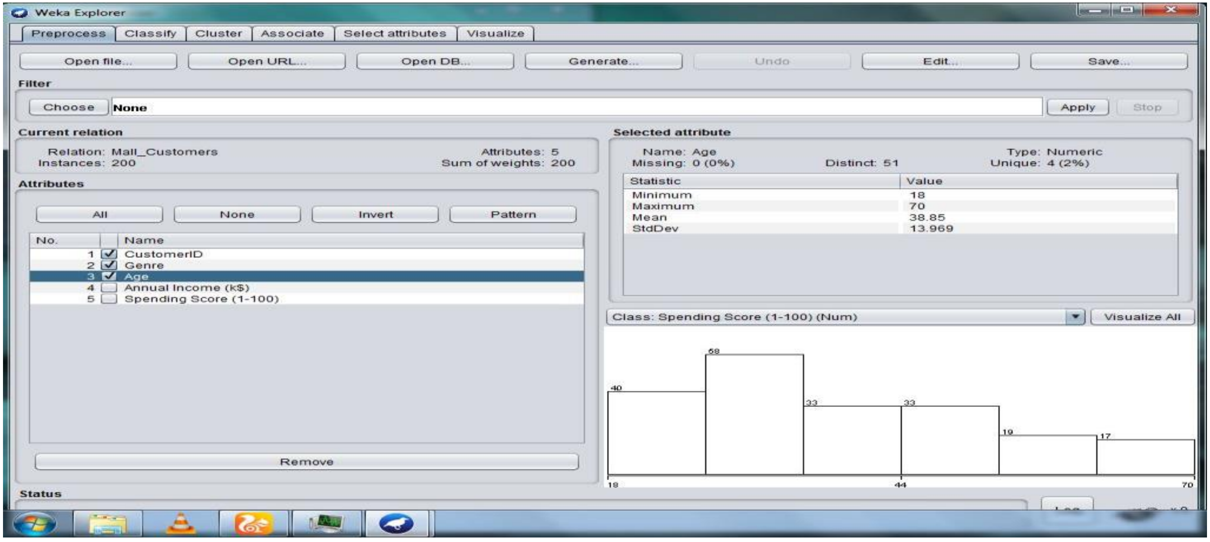


weka.clusterers.EM

Simple EM (expectation maximisation) class.

debug	False
displayModelInOldFormat	False
doNotCheckCapabilities	False
maxIterations	100
maximumNumberOfClusters	-1
minLogLikelihoodImprovementCV	1.0E-6
minLogLikelihoodImprovementIterating	1.0E-6
minStdDev	1.0E-6
numClusters	5
numExecutionSlots	1
numFolds	10
numKMeansRuns	10
seed	100

7. **Removing unwanted attribute (CustomerID, Genre,Age)**



Filter: None

Relation: Mall_Customers
Instances: 200
Attributes: 5
Sum of weights: 200

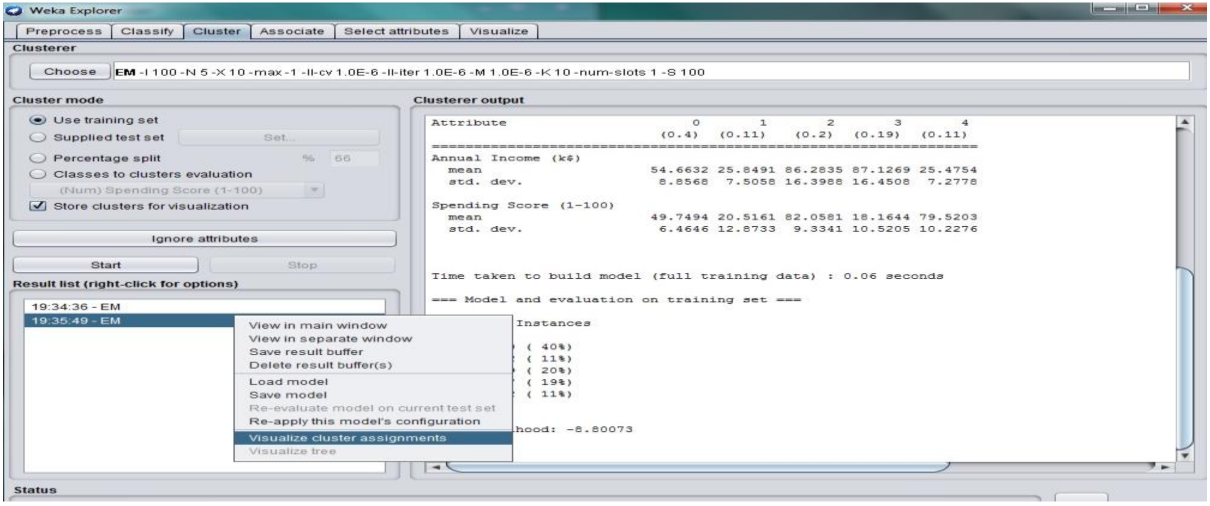
No.		Name
1	☑	CustomerID
2	☑	Genre
3	☑	Age
4	□	Annual Income (k$)
5	□	Spending Score (1-100)

Name: Age
Missing: 0 (0%)
Distinct: 51
Type: Numeric
Unique: 4 (2%)

Statistic	Value
Minimum	18
Maximum	70
Mean	38.85
StdDev	13.969

Class: Spending Score (1-100) (Num)

8. **right clicking on result buffer and clicking on visualise**



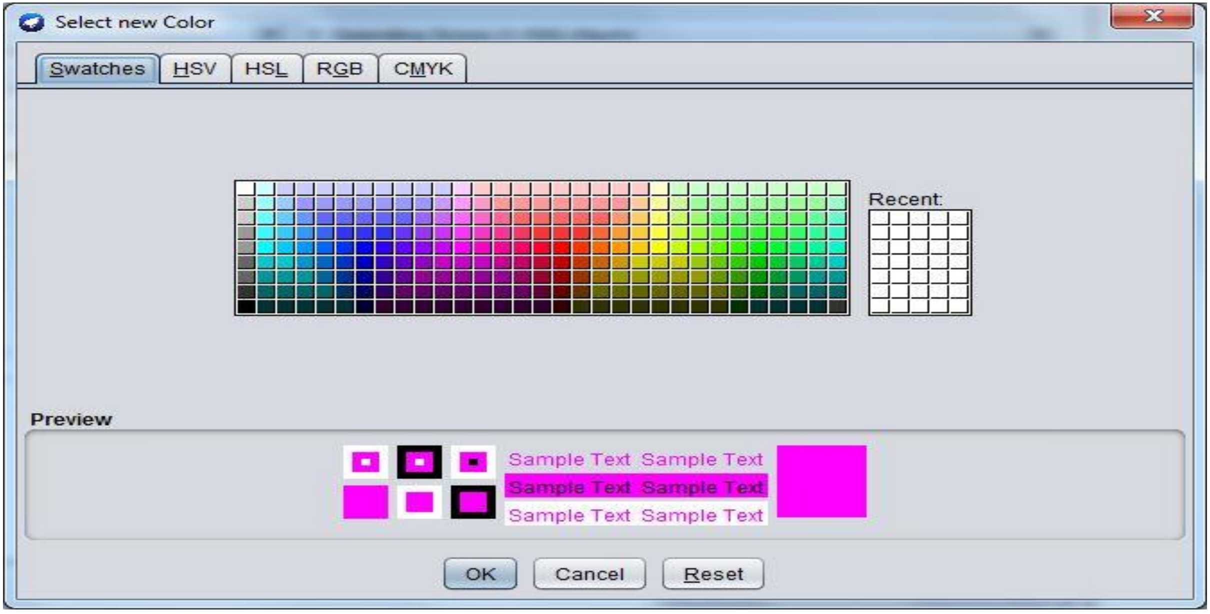
Clusterer: EM -I 100 -N 5 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Cluster mode:
● Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) Spending Score (1-100)
☑ Store clusters for visualization

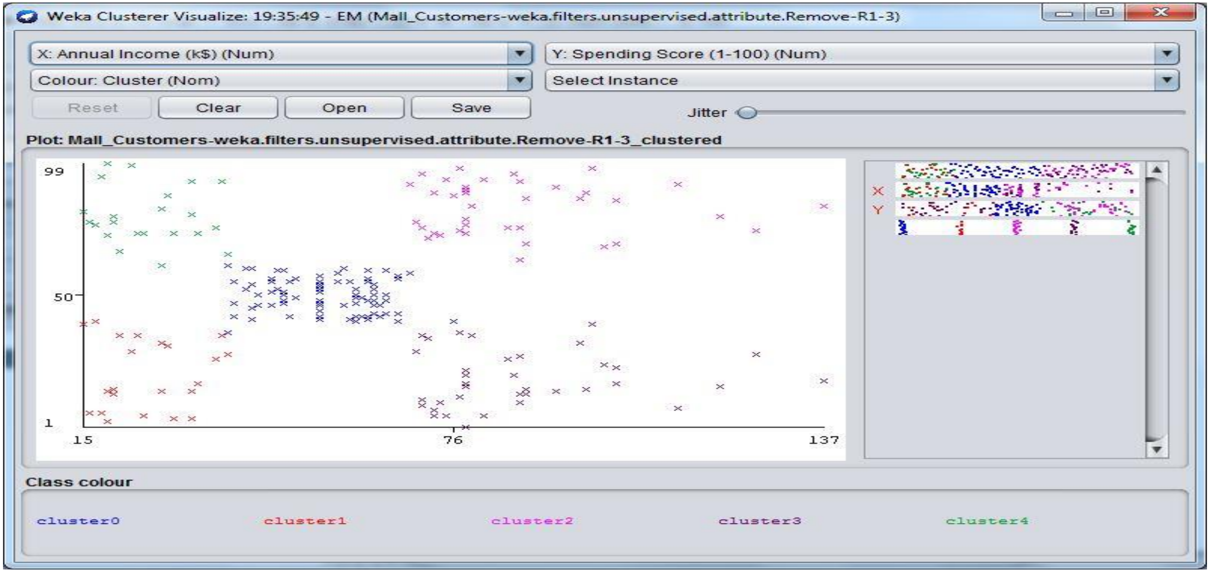
Result list:
19:34:36 - EM
19:35:49 - EM

```
Attribute                           0       1       2       3       4
                                (0.4)  (0.11)   (0.2)  (0.19)  (0.11)
======================================================================
Annual Income (k$)
  mean                        54.6632 25.8491 86.2835 87.1269 25.4754
  std. dev.                    8.8568  7.5058 16.3988 16.4508  7.2778

Spending Score (1-100)
  mean                        49.7494 20.5161 82.0581 18.1644 79.5203
  std. dev.                    6.4646 12.8733  9.3341 10.5205 10.2276



Time taken to build model (full training data) : 0.06 seconds

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

          Instances

            ( 40%)
            ( 11%)
            ( 20%)
            ( 19%)
            ( 11%)


        hood: -8.80073
```

9. **Changing colour for better representation**



10. selecting annual income on x and spending on y and visualising



Weka Clusterer Visualize: 19:35:49 - EM (Mall_Customers-weka.filters.unsupervised.attribute.Remove-R1-3)

X: Annual Income (k$) (Num)

Y: Spending Score (1-100) (Num)

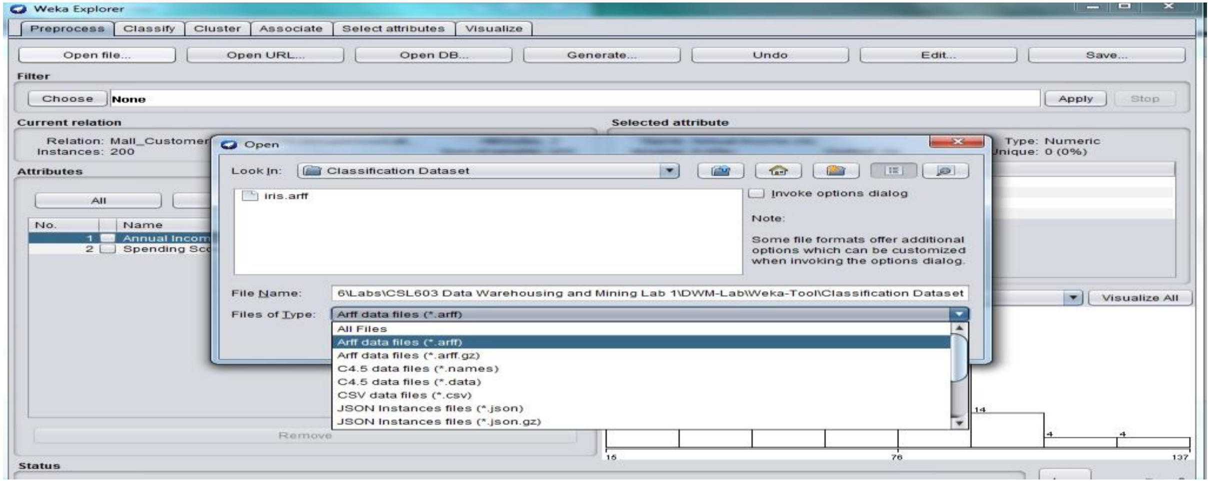
Colour: Cluster (Nom)

Select Instance

Plot: Mall_Customers-weka.filters.unsupervised.attribute.Remove-R1-3_clustered

#Performing Classification

1. **opening file, going to directory, selecting .arff file format**



Weka Explorer

Relation: Mall_Customer
Instances: 200

Type: Numeric
Unique: 0 (0%)

Look In: Classification Dataset

iris.arff

□ Invoke options dialog

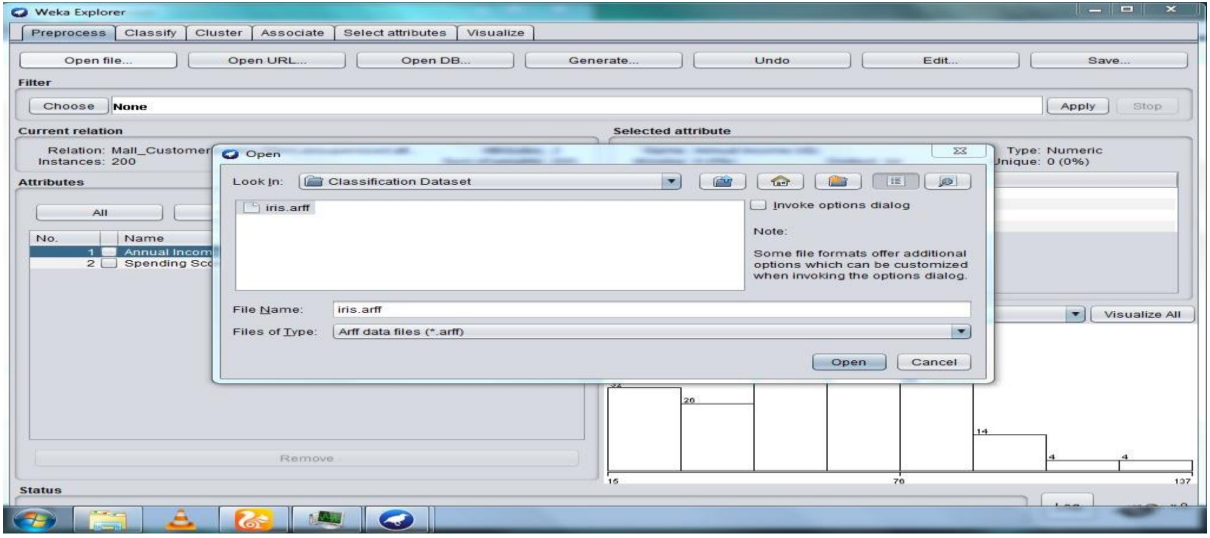
Note:
Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: 6\Labs\CSL603 Data Warehousing and Mining Lab 1\DWM-Lab\Weka-Tool\Classification Dataset

Files of Type: Arff data files (*.arff)

```
All Files
Arff data files (*.arff)
Arff data files (*.arff.gz)
C4.5 data files (*.names)
C4.5 data files (*.data)
CSV data files (*.csv)
JSON Instances files (*.json)
JSON Instances files (*.json.gz)
```

2. **selecting dataset for classification (iris.arff)**



Weka Explorer

Relation: Mall_Customer
Instances: 200

Type: Numeric
Unique: 0 (0%)

Look In: Classification Dataset

iris.arff

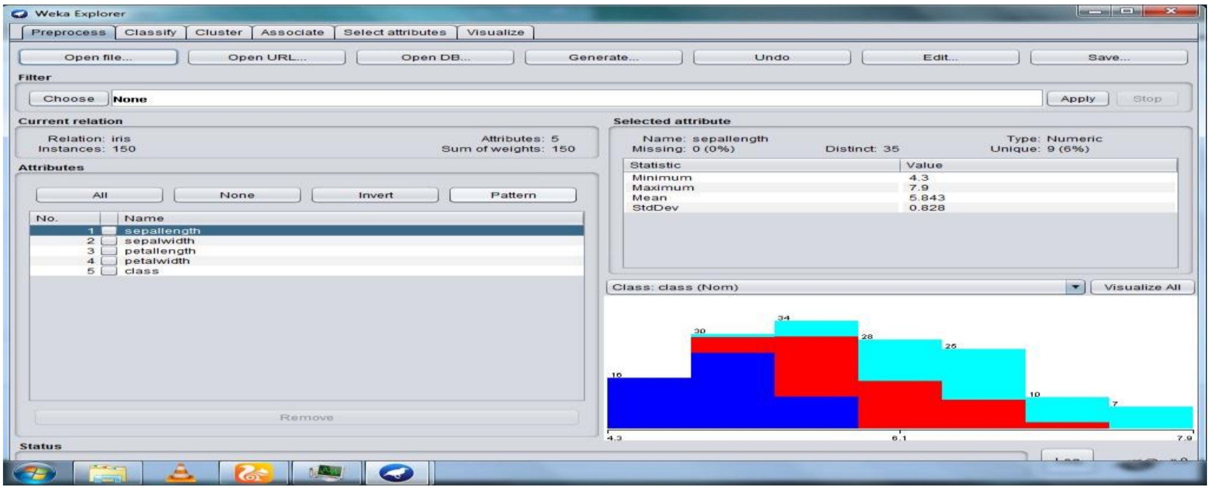
□ Invoke options dialog

Note:
Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: iris.arff

Files of Type: Arff data files (*.arff)

3. overview classification



Filter: None

Current relation

Relation: iris
Instances: 150
Attributes: 5
Sum of weights: 150

No.	Name
1	sepallength
2	sepalwidth
3	petallength
4	petalwidth
5	class

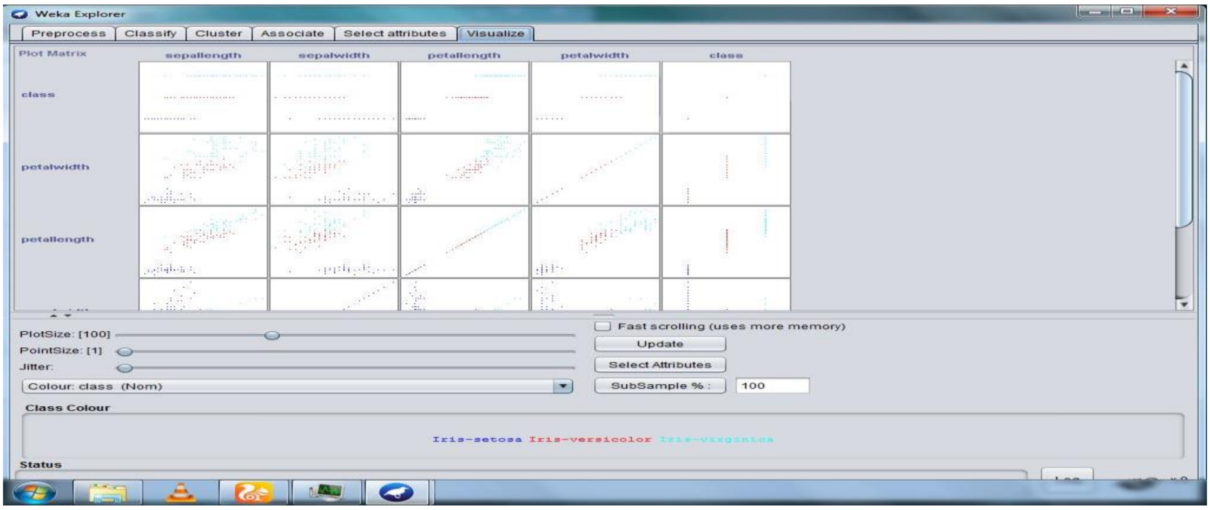
Selected attribute

Name: sepallength
Missing: 0 (0%)
Distinct: 35
Type: Numeric
Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom)

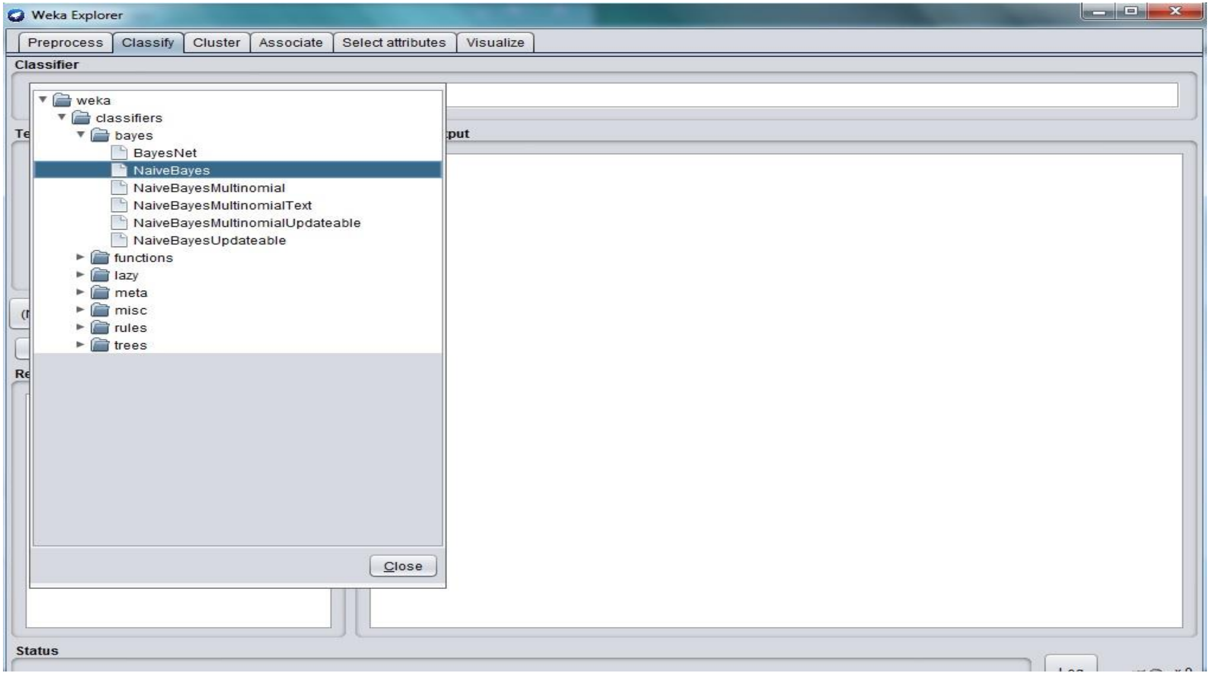
4. Visualise tab screen



PlotSize: [100]
PointSize: [1]
Jitter:
Colour: class (Nom)
□ Fast scrolling (uses more memory)
SubSample % : 100

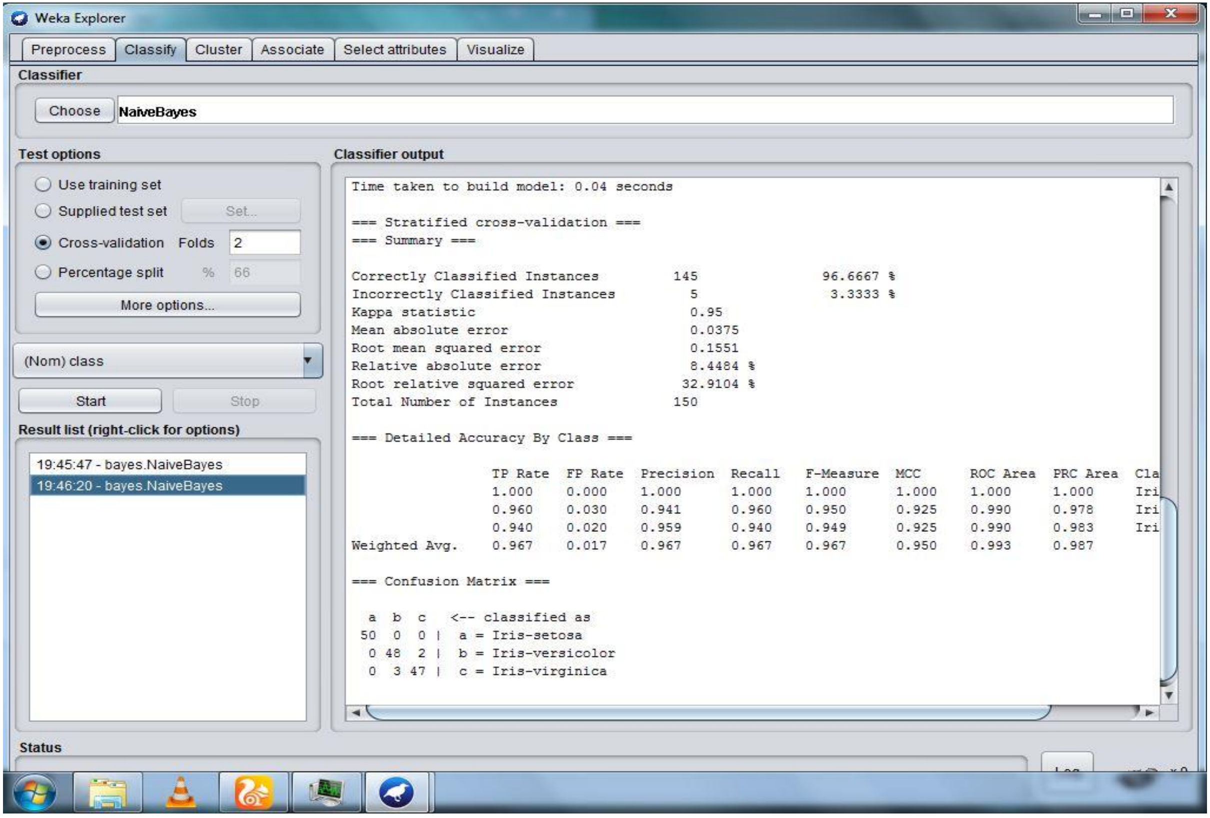
Iris-setosa Iris-versicolor Iris-virginica

5. selecting classify tab, choosing naivebayes, with cross validation of 2 fold



- weka
  - classifiers
    - bayes
      - BayesNet
      - NaiveBayes
      - NaiveBayesMultinomial
      - NaiveBayesMultinomialText
      - NaiveBayesMultinomialUpdateable
      - NaiveBayesUpdateable
    - functions
    - lazy
    - meta
    - misc
    - rules
    - trees

6. Summary of result and confusion matrix



Classifier: NaiveBayes

Test options:

○ Use training set

○ Supplied test set

● Cross-validation Folds 2

○ Percentage split % 66

(Nom) class

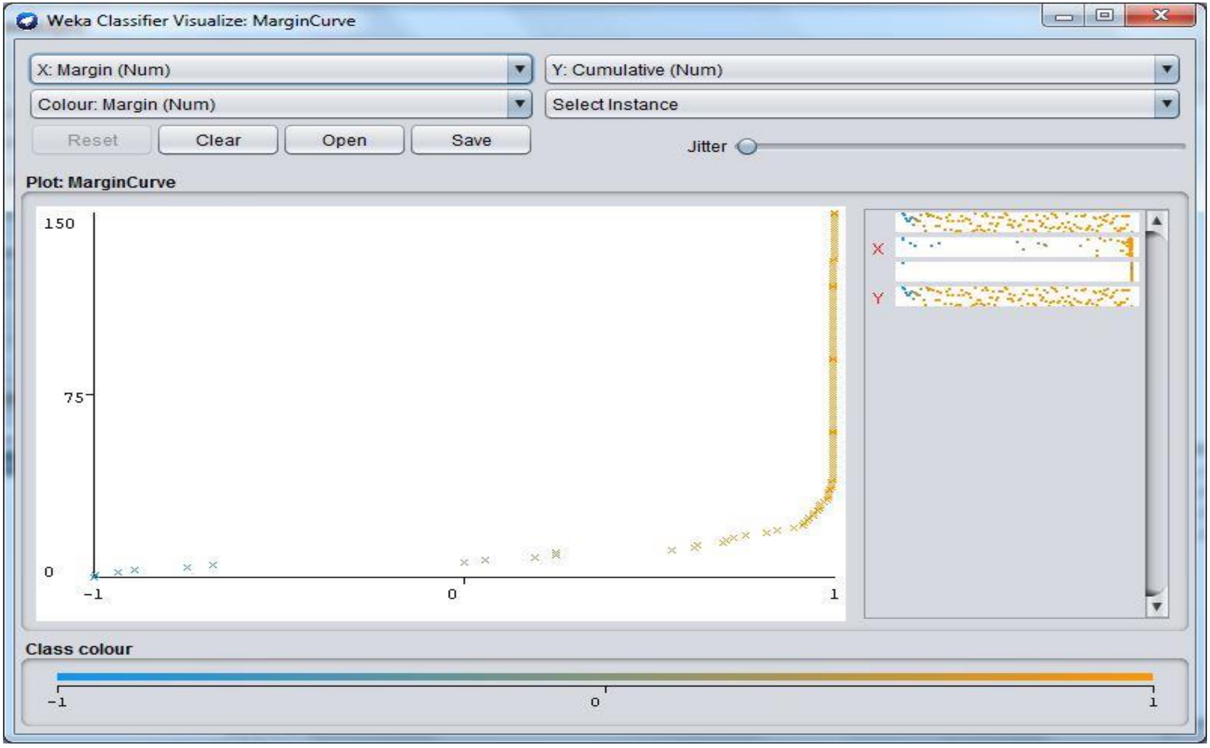
Result list:

19:45:47 - bayes.NaiveBayes

19:46:20 - bayes.NaiveBayes

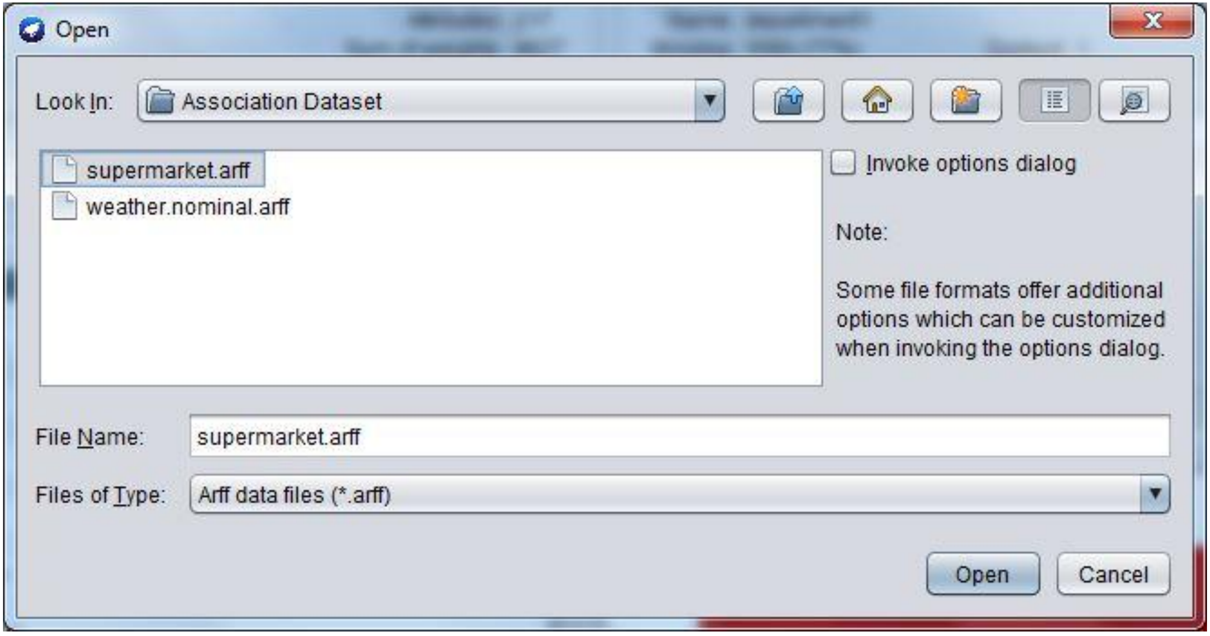
```
Time taken to build model: 0.04 seconds

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

Correctly Classified Instances         145               96.6667 %
Incorrectly Classified Instances         5                3.3333 %
Kappa statistic                          0.95
Mean absolute error                      0.0375
Root mean squared error                  0.1551
Relative absolute error                  8.4484 %
Root relative squared error             32.9104 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Iri
                 0.960    0.030    0.941      0.960    0.950      0.925    0.990     0.978     Iri
                 0.940    0.020    0.959      0.940    0.949      0.925    0.990     0.983     Iri
Weighted Avg.    0.967    0.017    0.967      0.967    0.967      0.950    0.993     0.987

=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 48  2 |  b = Iris-versicolor
  0  3 47 |  c = Iris-virginica
```

7. Visualising margin curve



#Performing Association

- **Opening file selecting dataset for association (supermarket.arff)**



Look In: Association Dataset

supermarket.arff
weather.nominal.arff

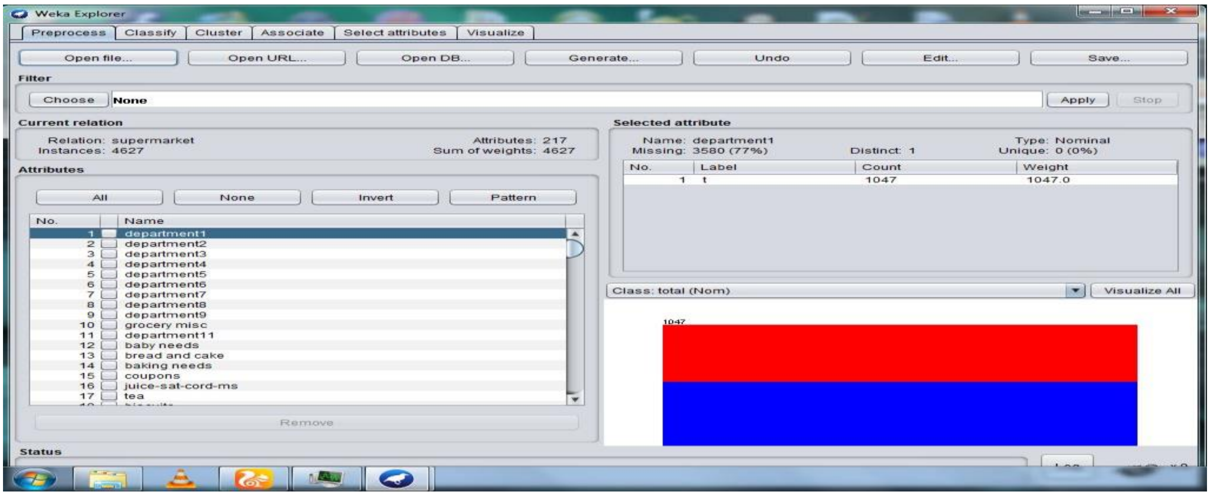
Note:

Some file formats offer additional options which can be customized when invoking the options dialog.

File Name: supermarket.arff

Files of Type: Arff data files (*.arff)

- **Overview screen**



Filter: None

Current relation

Relation: supermarket   Attributes: 217
Instances: 4627   Sum of weights: 4627

Attributes

No.	Name
1	department1
2	department2
3	department3
4	department4
5	department5
6	department6
7	department7
8	department8
9	department9
10	grocery misc
11	department11
12	baby needs
13	bread and cake
14	baking needs
15	coupons
16	juice-sat-cord-ms
17	tea

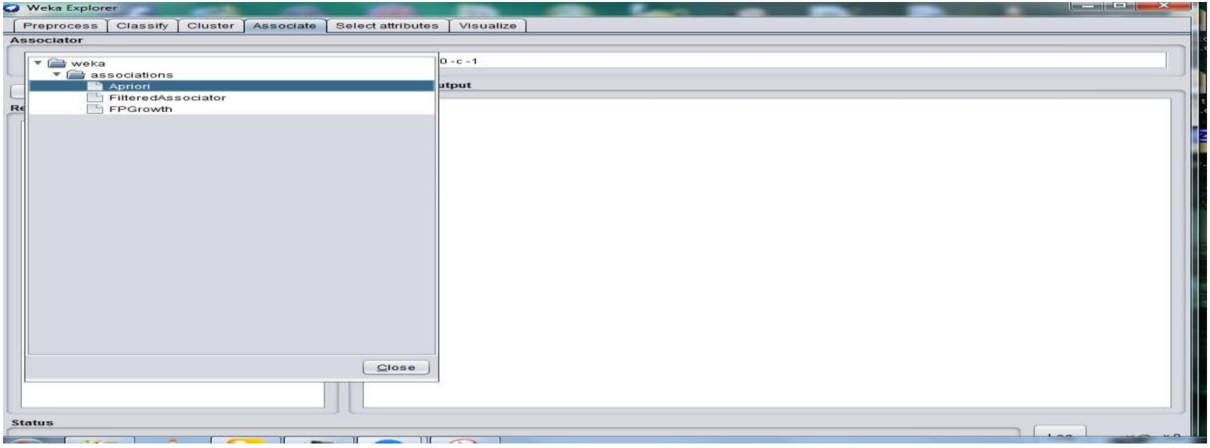
Selected attribute

Name: department1   Type: Nominal
Missing: 3580 (77%)   Distinct: 1   Unique: 0 (0%)

No.	Label	Count	Weight
1	t	1047	1047.0

Class: total (Nom)

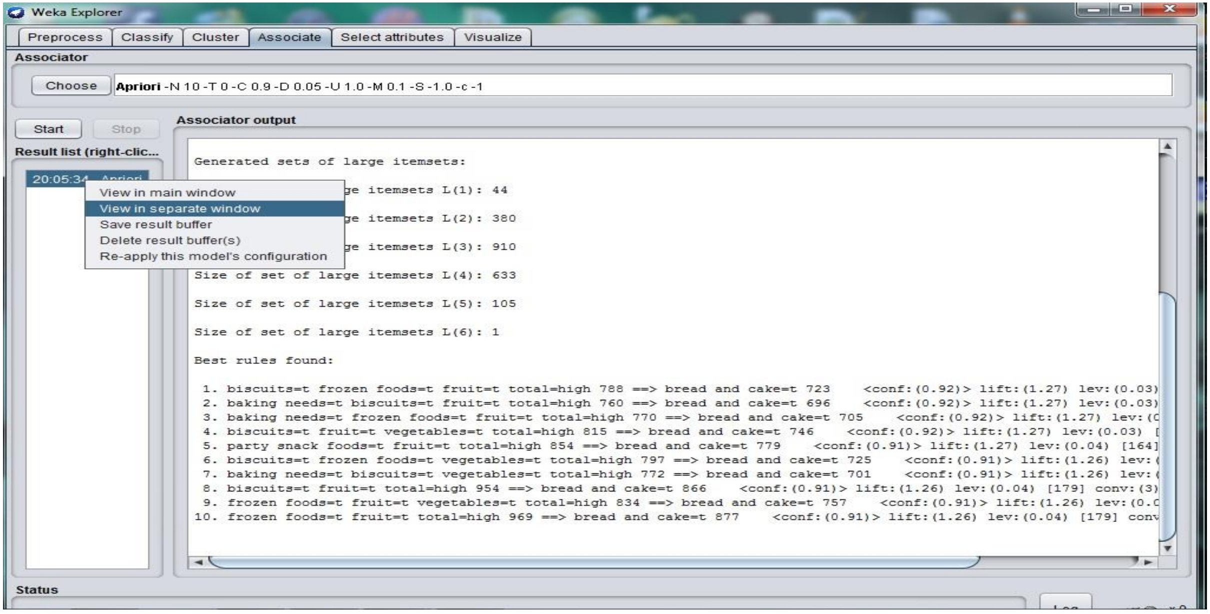
- **Choosing apriori and starting**



0 -c -1

weka
associations
Apriori
FilteredAssociator
FPGrowth

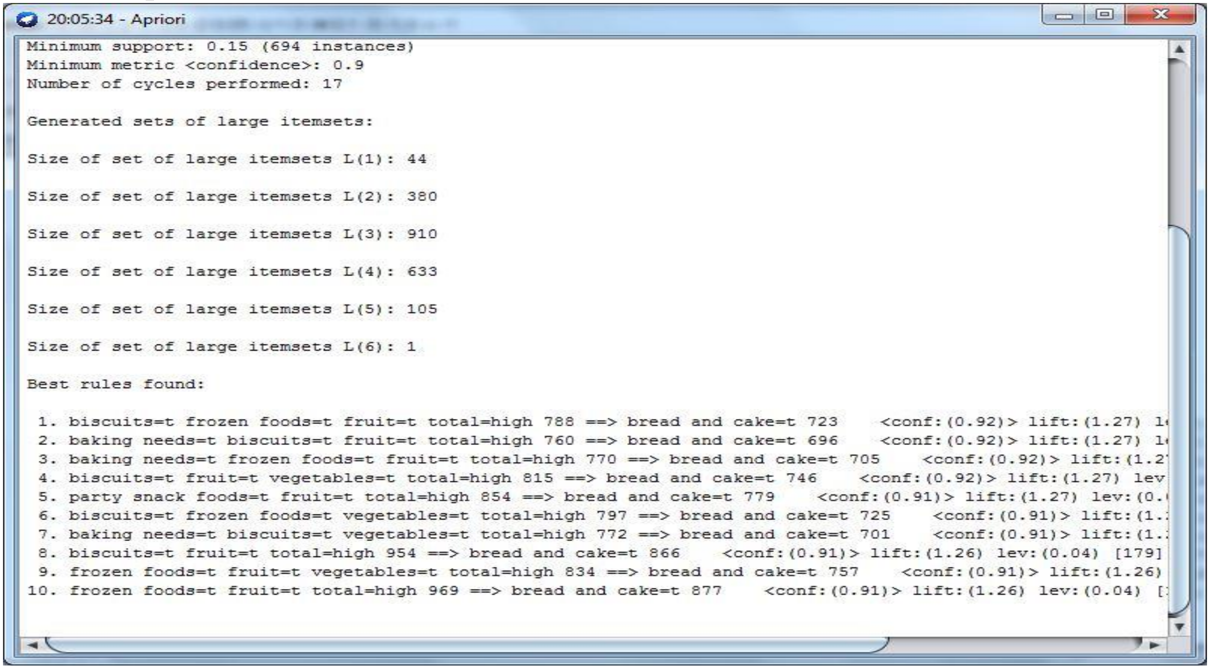
- **Selecting View in separate window from context menu of result buffer**



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

```
Generated sets of large itemsets:

                 ge itemsets L(1): 44

                 ge itemsets L(2): 380

                 ge 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:(0.03)
 2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696    <conf:(0.92)> lift:(1.27) lev:(0.03)
 3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705    <conf:(0.92)> lift:(1.27) lev:(
 4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746    <conf:(0.92)> lift:(1.27) lev:(0.03)
 5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779    <conf:(0.91)> lift:(1.27) lev:(0.04) [164]
 6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725    <conf:(0.91)> lift:(1.26) lev:
 7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701    <conf:(0.91)> lift:(1.26) lev:
 8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866    <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv:(3)
 9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757    <conf:(0.91)> lift:(1.26) lev:(0.
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877    <conf:(0.91)> lift:(1.26) lev:(0.04) [179] conv
```

- **Final rule pair of result**



```
Minimum support: 0.15 (694 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 17

Generated sets of large itemsets:

Size of set of large itemsets L(1): 44

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) l
 2. baking needs=t biscuits=t fruit=t total=high 760 ==> bread and cake=t 696    <conf:(0.92)> lift:(1.27) l
 3. baking needs=t frozen foods=t fruit=t total=high 770 ==> bread and cake=t 705    <conf:(0.92)> lift:(1.2
 4. biscuits=t fruit=t vegetables=t total=high 815 ==> bread and cake=t 746    <conf:(0.92)> lift:(1.27) lev
 5. party snack foods=t fruit=t total=high 854 ==> bread and cake=t 779    <conf:(0.91)> lift:(1.27) lev:(0.
 6. biscuits=t frozen foods=t vegetables=t total=high 797 ==> bread and cake=t 725    <conf:(0.91)> lift:(1.
 7. baking needs=t biscuits=t vegetables=t total=high 772 ==> bread and cake=t 701    <conf:(0.91)> lift:(1.
 8. biscuits=t fruit=t total=high 954 ==> bread and cake=t 866    <conf:(0.91)> lift:(1.26) lev:(0.04) [179]
 9. frozen foods=t fruit=t vegetables=t total=high 834 ==> bread and cake=t 757    <conf:(0.91)> lift:(1.26)
10. frozen foods=t fruit=t total=high 969 ==> bread and cake=t 877    <conf:(0.91)> lift:(1.26) lev:(0.04) [
```

#Conclusion

From this experiment we learned some very essential concept related to any data mining process. For data pre-processing we used an open source weka tool which provides every result and ample of GUI based so the work task for data analysis becomes easier. Lastly, we implemented classification, clustering, and association using this tool and visualised it appropriately