

## EXPERIMENT NO. 6

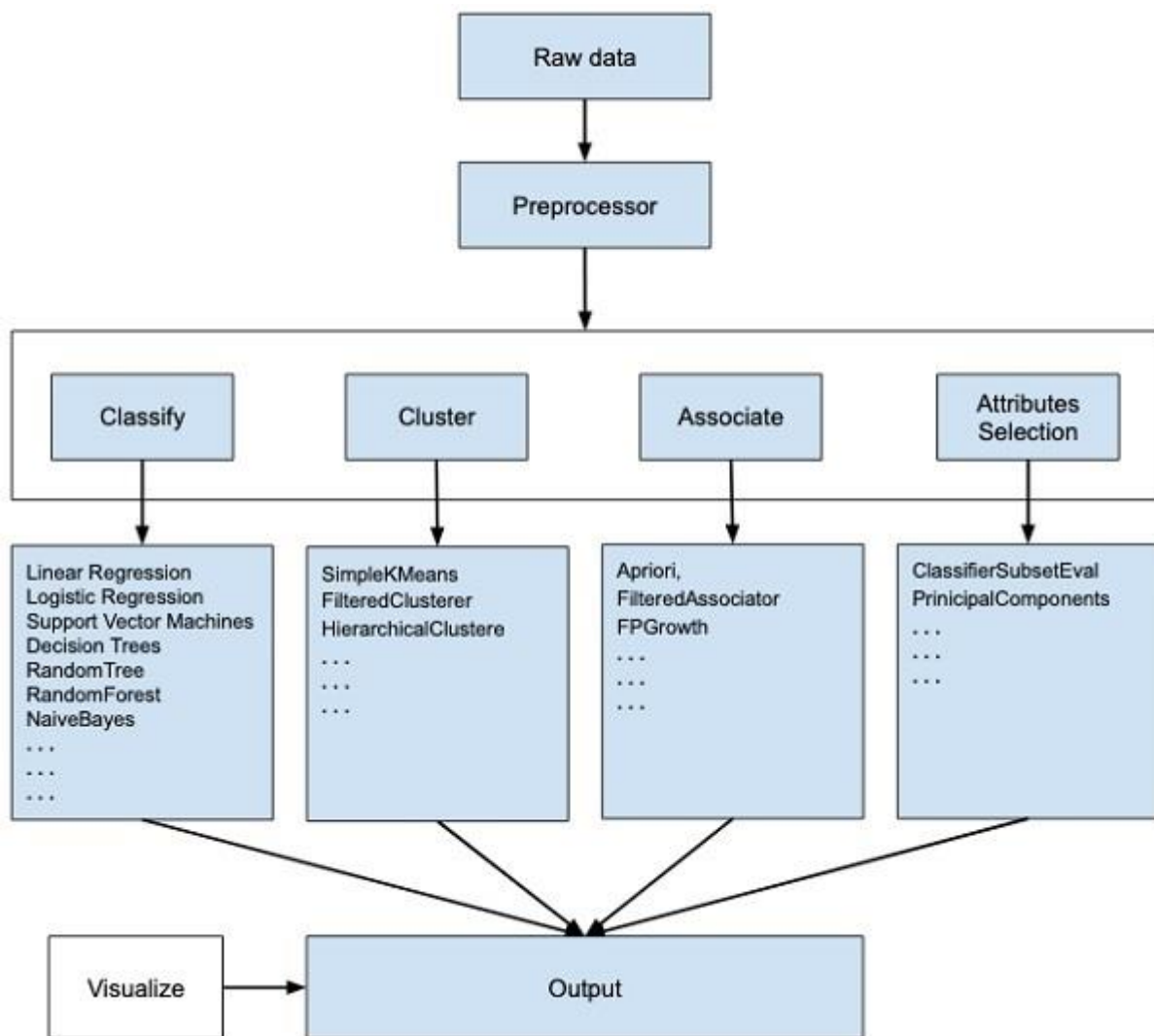
Aim: To perform data Pre-processing and implement Classification, Clustering and Association algorithms on data set.

Requirement: Windows O.S and Weka Tool.

Theory:

Weka Tool:

WEKA - an open source software provides tools for data pre-processing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram –



If you observe the beginning of the flow of the image, you will understand that there are many stages in dealing with Big Data to make it suitable for machine learning –

First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data pre-processing tools provided in WEKA to cleanse the data.

Then, you would save the pre-processed data in your local storage for applying ML algorithms.

Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as **Classify**, **Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.

Note that under each category, WEKA provides the implementation of several algorithms. You would select an algorithm of your choice, set the desired parameters and run it on the dataset.

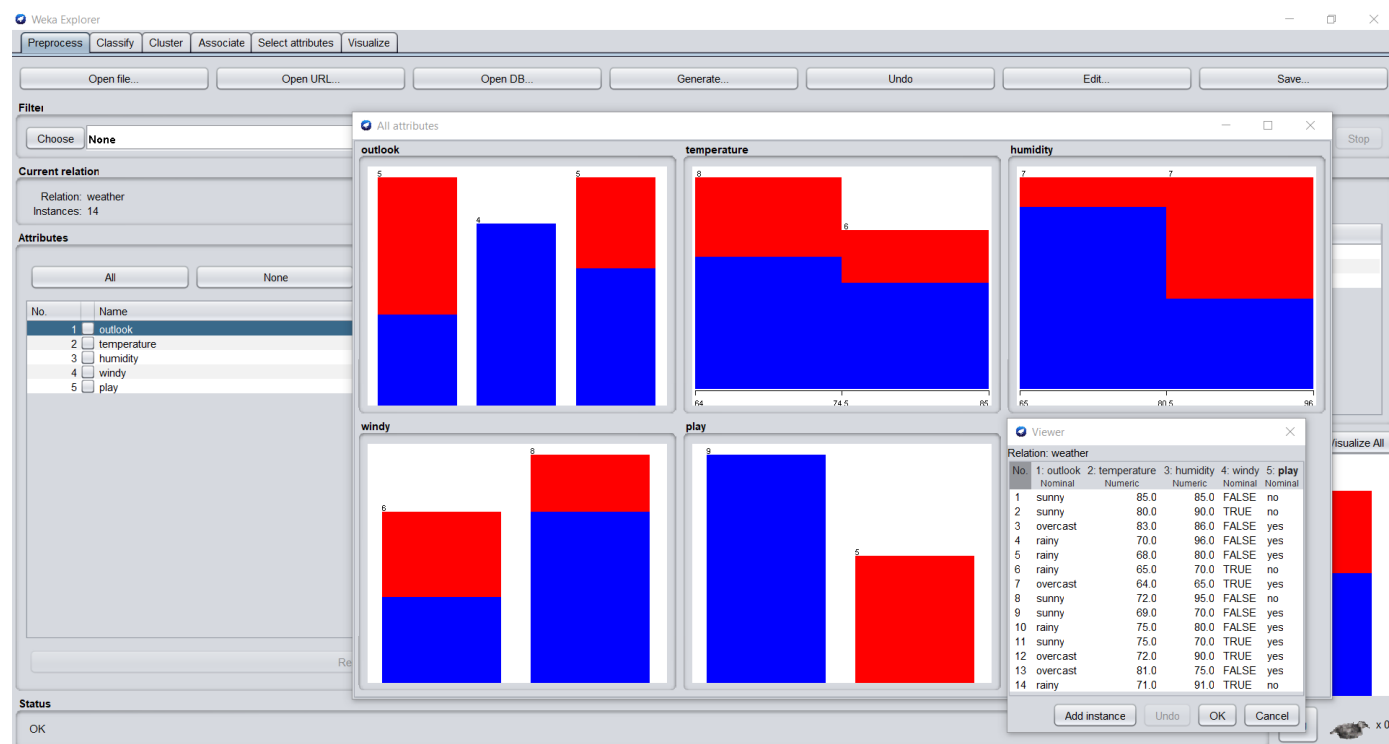
Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data.

The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

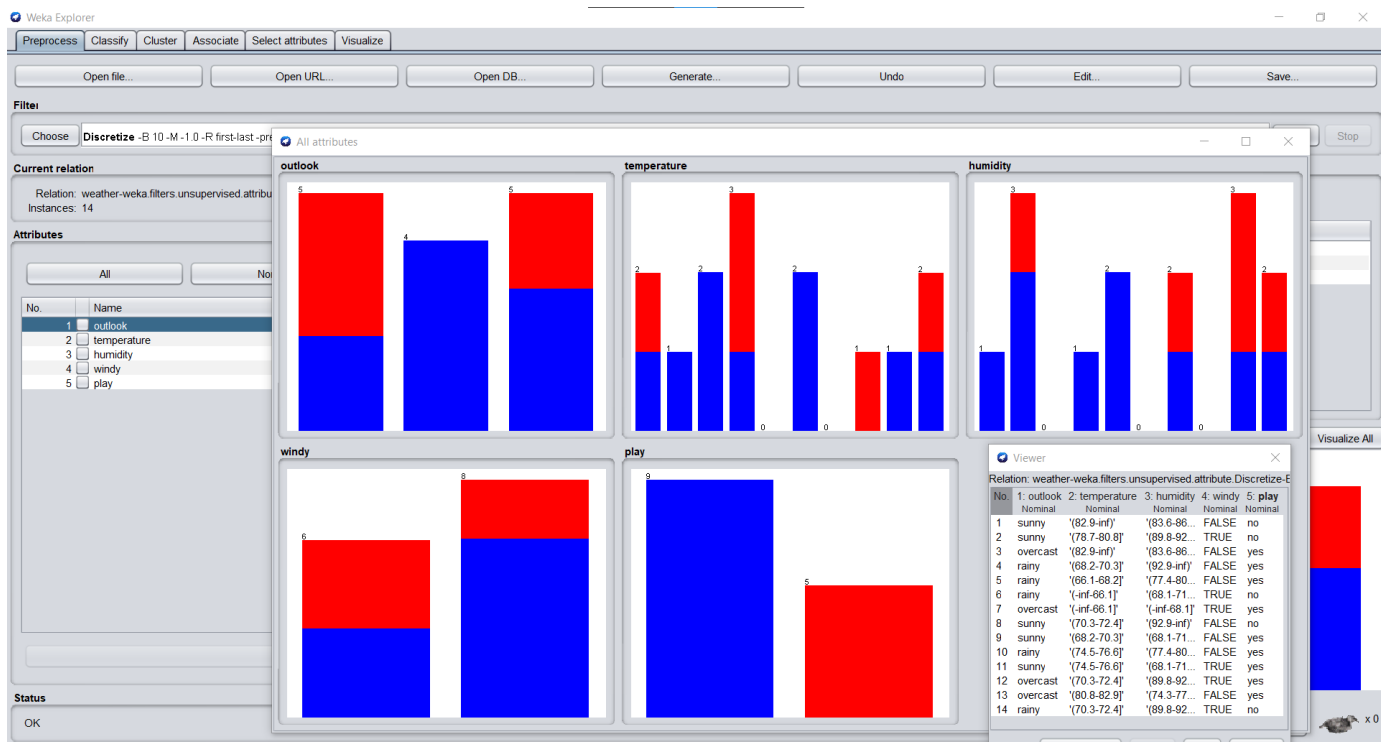
Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

### Output:

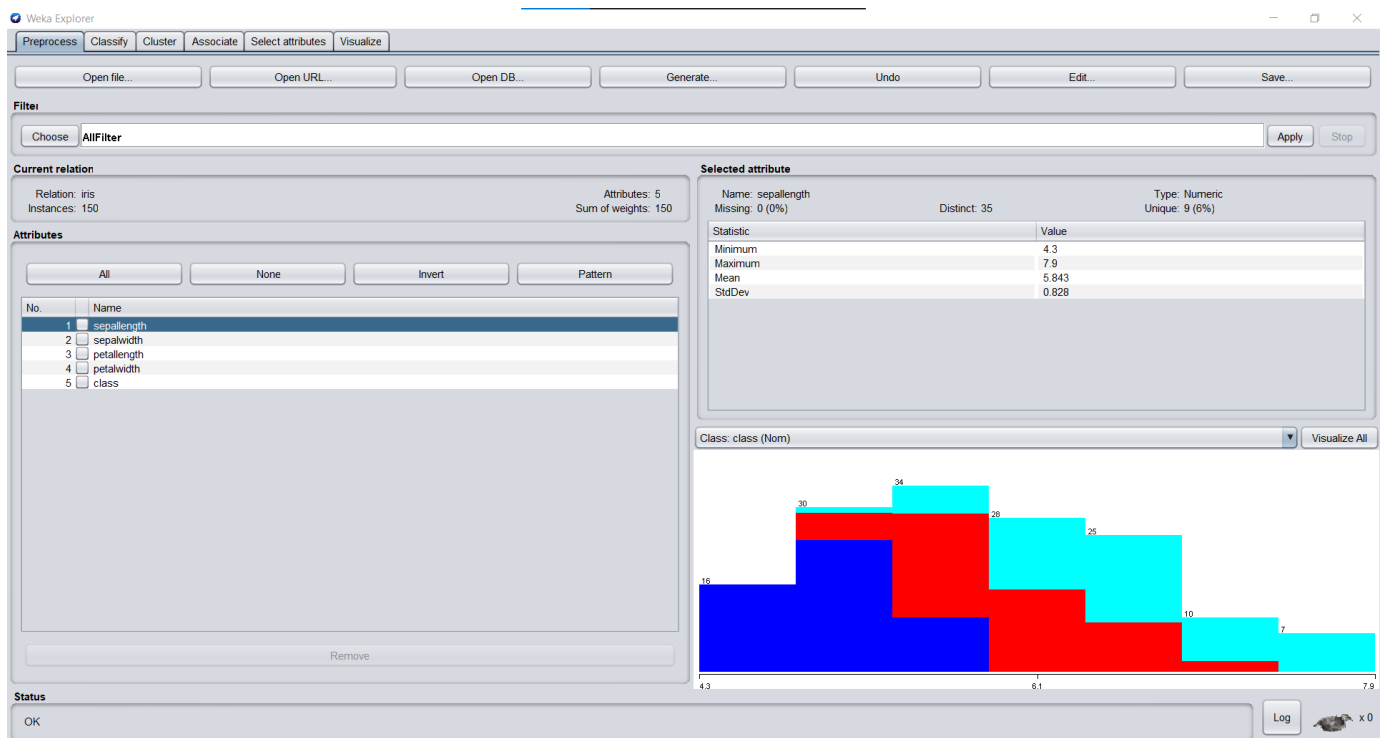
#### Data Set before Data Pre-processing (Data Discretization):



## Data Set after Data Pre-processing (Discretization):



## Iris Data Set:



## J48 Tree classification on iris data set:

**Weka Explorer**

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) class

Start Stop

**Result list (right-click for options)**

17.20.42 - trees.J48

**Classifier output**

```

=== Run information ===

Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    iris
Instances:   150
Attributes:  5
              sepalwidth
              sepalwidth
              petalwidth
              petalwidth
              class
Test mode:   10-fold cross-validation

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

J48 pruned tree
-----

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petalwidth <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petalwidth > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   |   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves :    5
  
```

**Status**

OK Log x0

**Weka Explorer**

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) class

Start Stop

**Result list (right-click for options)**

17.20.42 - trees.J48

**Classifier output**

```

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

Correctly Classified Instances      144           96 %
Incorrectly Classified Instances     6            4 %
Kappa statistic                     0.94
Mean absolute error                  0.035
Root mean squared error              0.1586
Relative absolute error              7.8705 %
Root relative squared error          33.6353 %
Total Number of Instances           150

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall  F-Measure  MDC   ROC Area  PRC Area  Cl
          0.980    0.000    1.000    0.980    0.990    0.985  0.990    0.987    Ir
          0.940    0.030    0.940    0.940    0.940    0.910  0.952    0.880    Ir
          0.960    0.030    0.941    0.960    0.950    0.925  0.961    0.905    Ir
Weighted Avg.  0.960    0.020    0.960    0.960    0.960    0.940  0.968    0.924

=== Confusion Matrix ===

 a  b  c  <-- classified as
49  1  0 | a = Iris-setosa
 0 47  3 | b = Iris-versicolor
 0  2 48 | c = Iris-virginica
  
```

**Status**

OK Log x0

## Data Clustering (K-Means/ K-Medoids) on iris Data Set:

**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 (Nom) class  
☒ Store clusters for visualization

Ignore attributes

Start Stop

**Result list (right-click for options)**

19:36:28 - SimpleKMeans

**Clusterer output**

```

=== Run information ===

Scheme:      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.EuclideanDistance" -R first-last" -I 500 -num-slots 1
Relation:    iris
Instances:   150
Attributes:  5
              sepalwidth
              sepalwidth
              petalwidth
              petalwidth
              class
Test mode:   evaluate on training data

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

KMeans
=====

Number of iterations: 7
Within cluster sum of squared errors: 62.1436882815797

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor

Missing values globally replaced with mean/mode

```

**Status**

OK Log x0

**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 (Nom) class  
☒ Store clusters for visualization

Ignore attributes

Start Stop

**Result list (right-click for options)**

19:36:28 - SimpleKMeans

**Clusterer output**

```

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute      Full Data      Cluster#      0      1
                (150.0)      (100.0)      (50.0)
=====
sepalwidth     5.8433         6.262         5.006
sepalwidth     3.054          2.872         3.418
petalwidth     3.7587         4.906         1.464
petalwidth     1.1987         1.676         0.244
class          Iris-setosa   Iris-versicolor   Iris-setosa

Time taken to build model (full training data) : 0.01 seconds

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

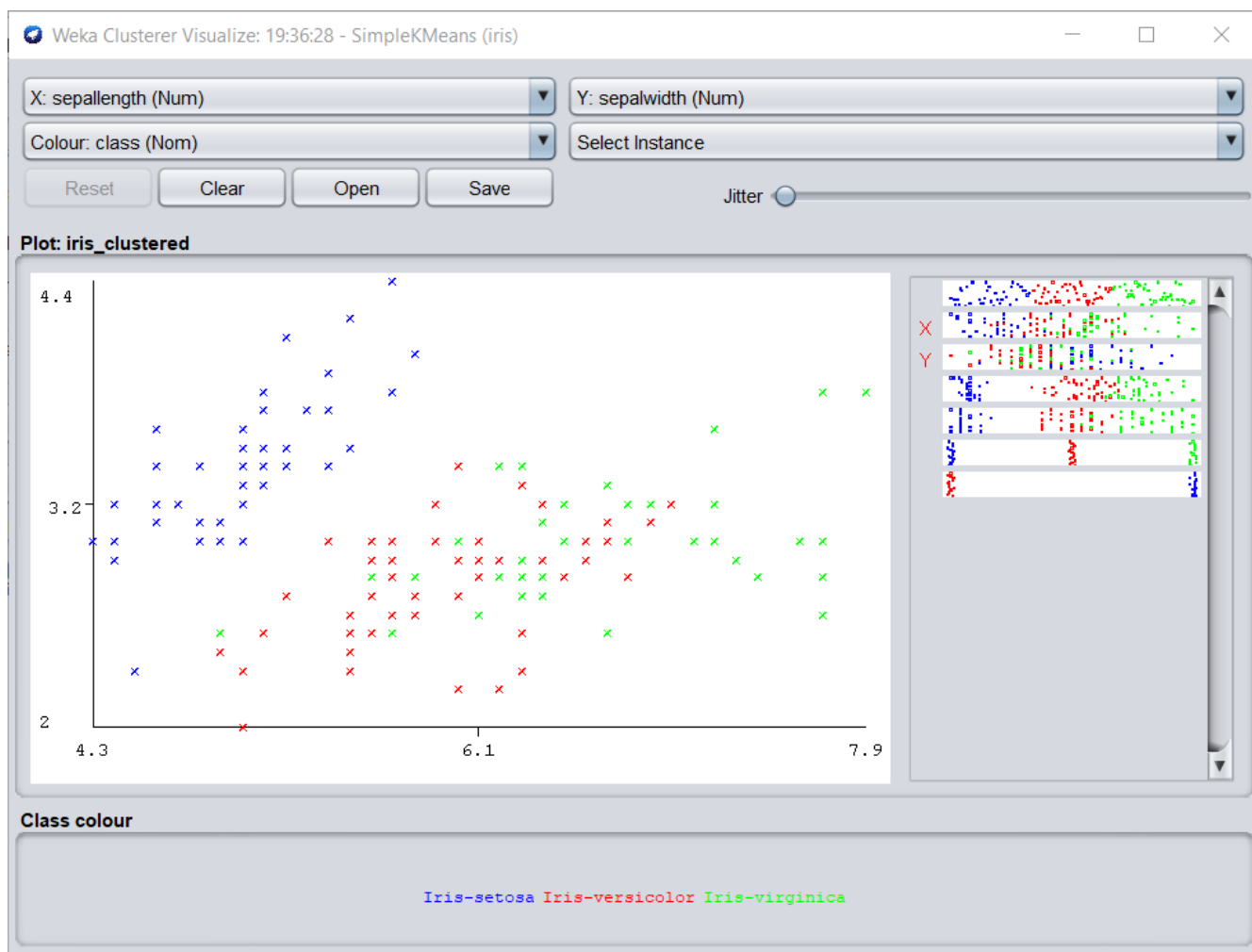
Clustered Instances

0      100 ( 67%)
1       50 ( 33%)

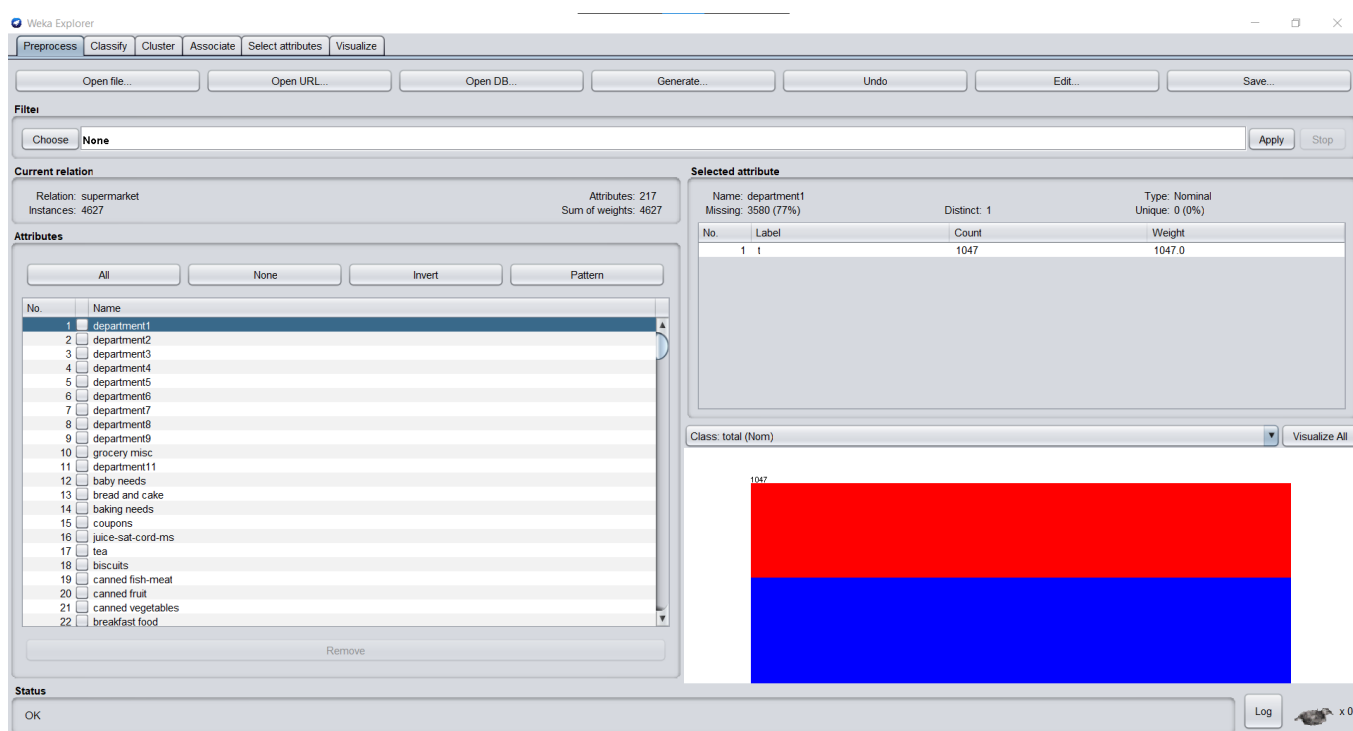
```

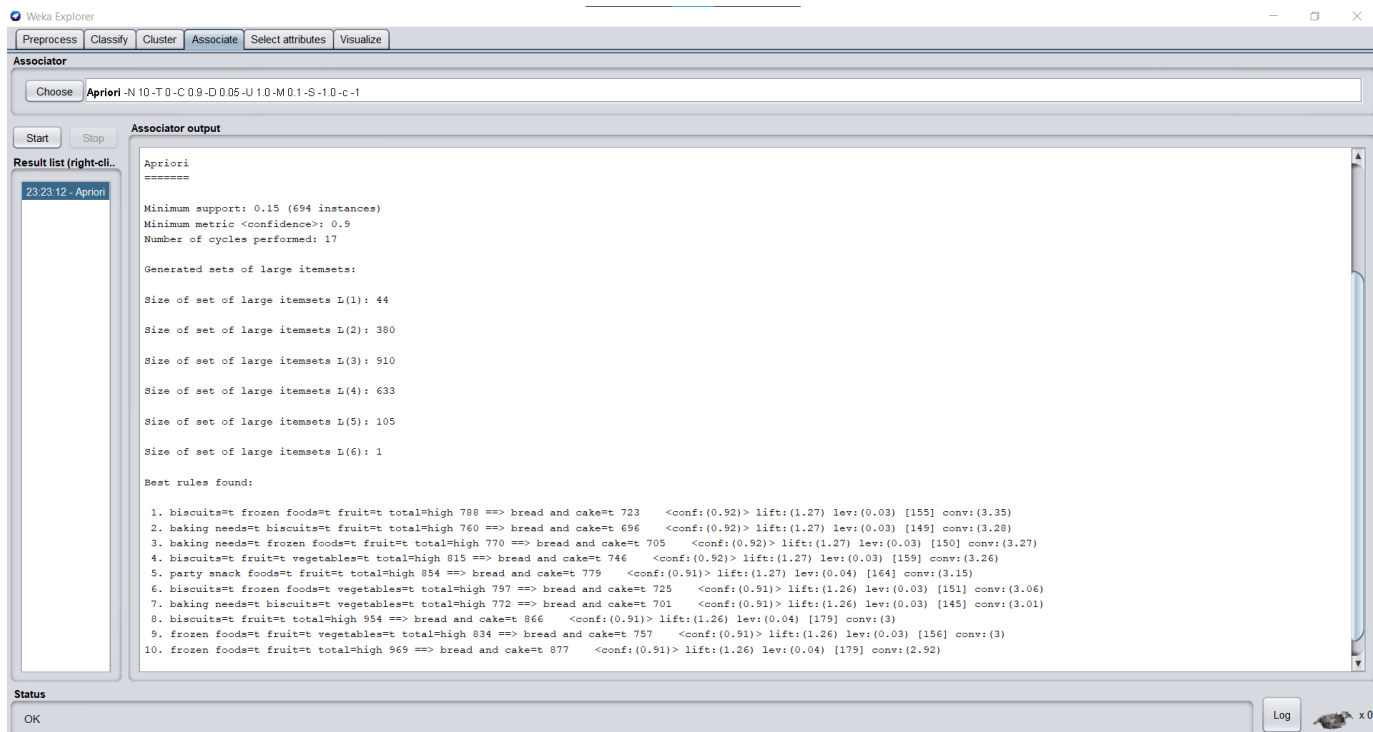
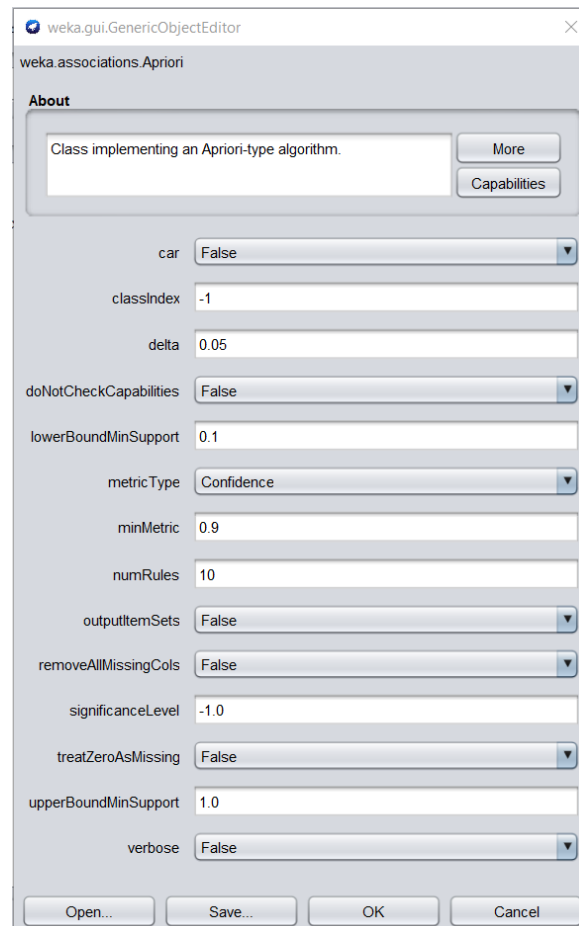
**Status**

OK Log x0



### Association algorithm (Apriori) on Supermarket Data Set:





**Conclusion:** We have successfully Pre-processed data and implemented Classification, Clustering and Association algorithms on data sets using Weka Tool.