Experiment no.:- 6

Aim :- Perform data Pre-processing task and demonstrate classification, clustering algorithm on dataset using data mining tool (WEKA/ R Tool).

Theory:-

Firstly, install weka on your systems for Ubuntu use code in terminal

$ sudo apt update

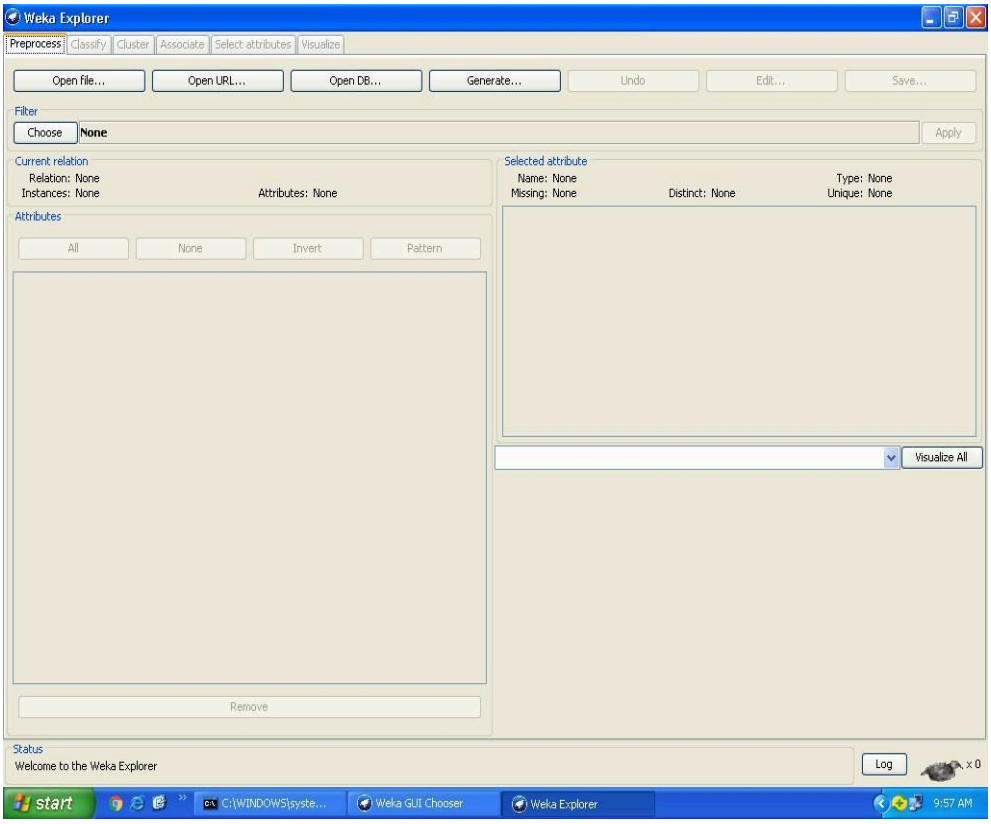
$ sudo apt install Weka

for windows

Browse and download through Chrome and complete the installation wizard.

Explore various options in Weka for Preprocessing data and apply (like Discretization Filters, Resample filter, etc.) n each dataset.

Preprocess Tab



Current Relation: Once some data has been loaded, the Preprocess panel shows a variety of in**formation. The Current relation box (the "current relation" is the** currently loaded data, which can be interpreted as a single relational table in database terminology) has three entries:

1. **Relation.** The name of the relation, as given in the file it was loaded from. Filters (described below) modify the name of a relation.

2. **Instances.** The number of instances (data points/records) in the data.

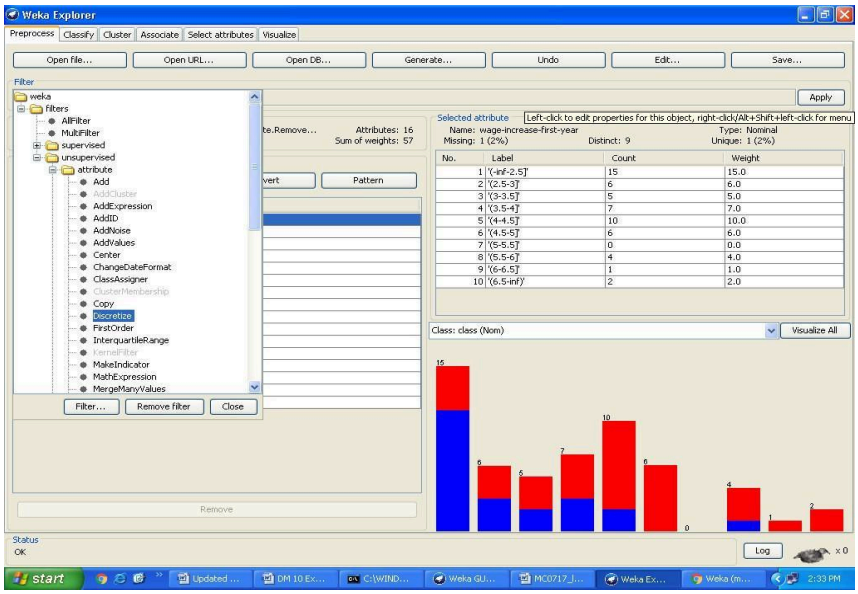
3. **Attributes.** The number of attributes (features) in the data.

PREPROCESSING

1. **All.** All boxes are ticked.
2. **None.** All boxes are cleared (unticked).
3. **Invert.** Boxes that are ticked become unticked and vice versa.

4. **Pattern.** Enables the user to select attributes based on a Perl 5 Regular Expression. E.g., .* id selects all attributes which name ends with id.

Working with Filters:-



The GenericObjectEditor Dialog Box

The GenericObjectEditor dialog box lets you configure a filter. The same kind of dialog box is used to configure other objects, such as classifiers and clusterers

(see below). The fields in the window reflect the available options.

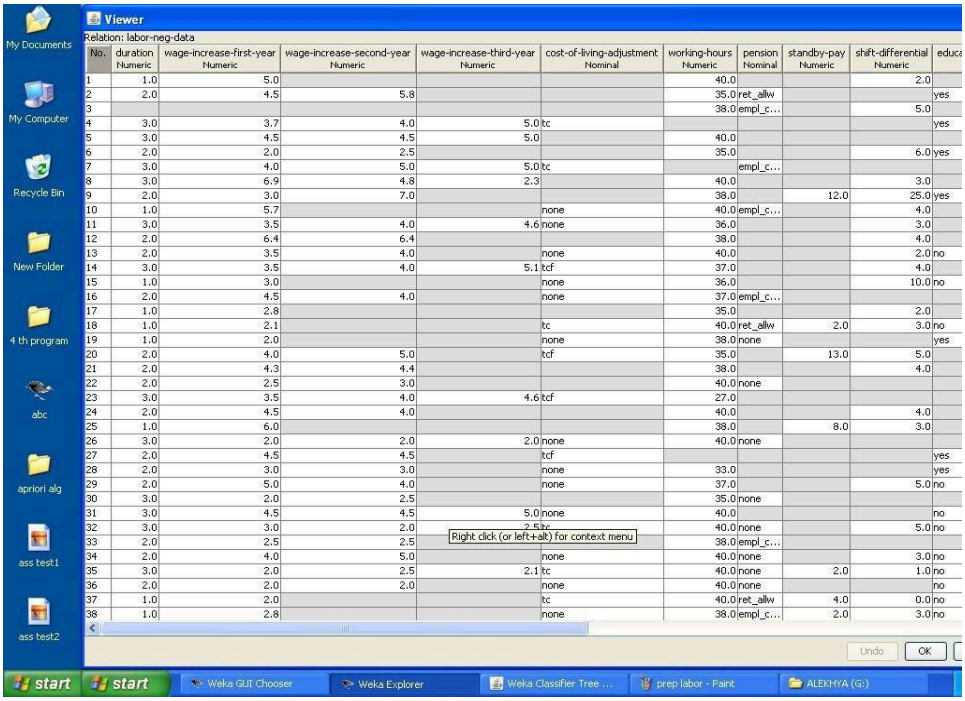
Right-clicking (or Alt+Shift+Left-Click) on such a field will bring up a popup menu, listing the following options:

Applying Filters

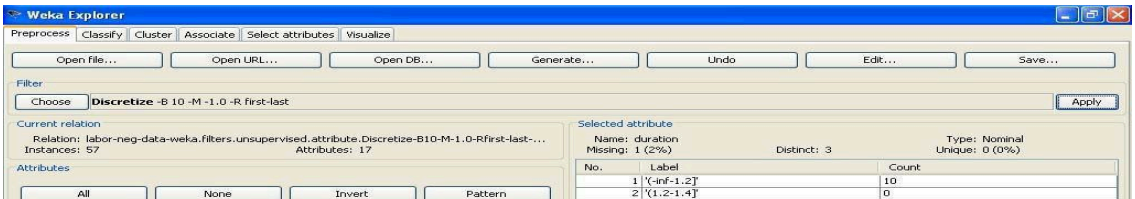
Steps for run preprocessing tab in WEKA

1. Open WEKA Tool.
2. Click on WEKA Explorer.
3. Click on Preprocessing tab button.
4. Click on open file button.
5. Choose WEKA folder in C drive.
6. Select and Click on data option button.
7. Choose labor data set and open file.
8. Choose filter button and select the Unsupervised-Discritize option and apply

Dataset labor.arff



The following screenshot shows the effect of discretization

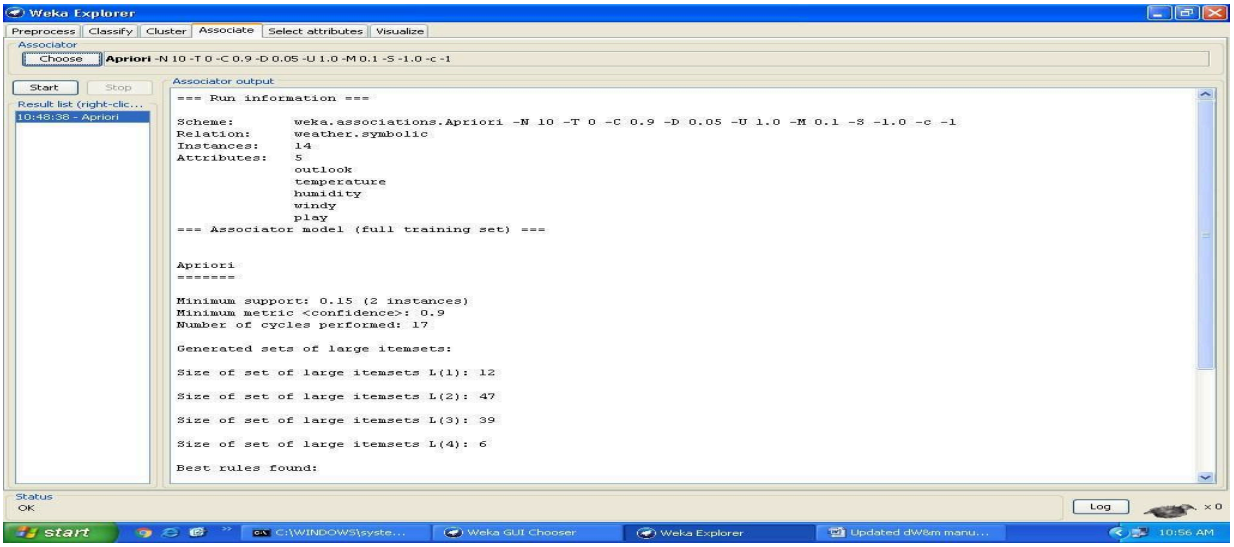


Association Tab:

Load each dataset into Weka and run Aprior algorithm with different support and confidence values. Study the rules generated.

Steps for run Aprior algorithm in WEKA

1. Open WEKA Tool.
2. Click on WEKA Explorer.
3. Click on Preprocessing tab button.
4. Click on open file button.
5. Choose WEKA folder in C drive.
6. Select and Click on data option button.
7. Choose Weather data set and open file.
8. Click on Associate tab and Choose Aprior algorithm
9. Click on start button.

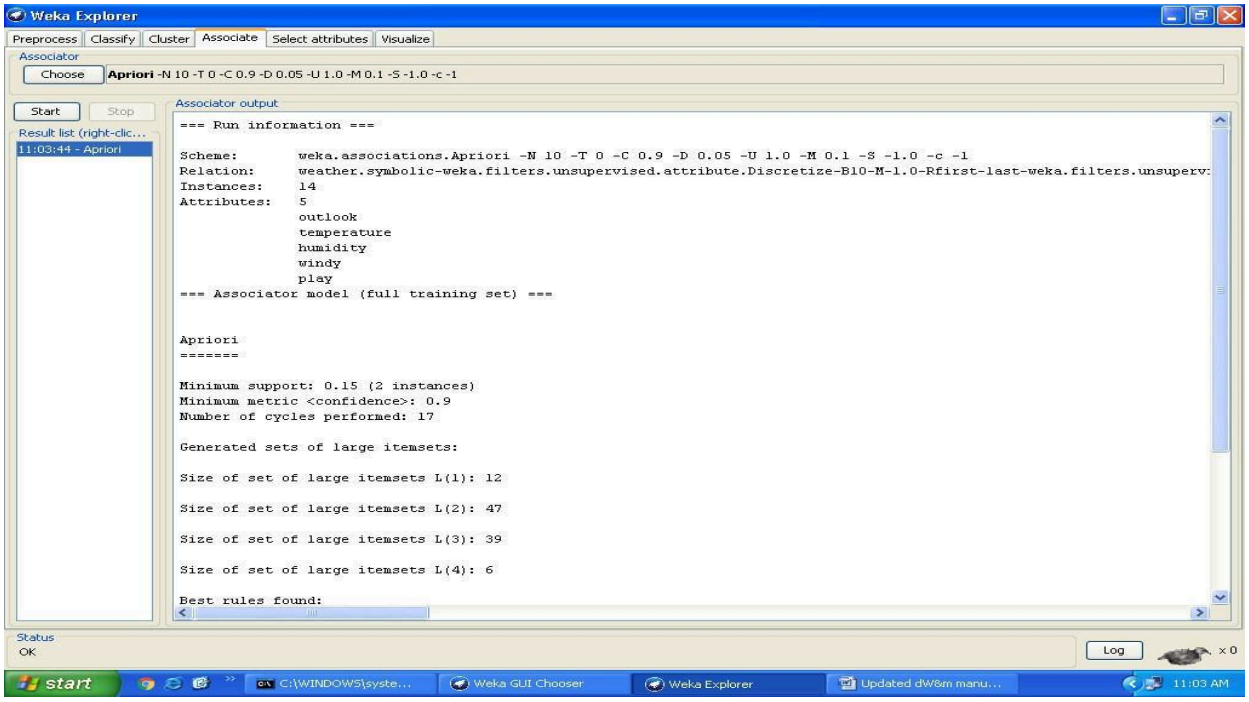


Apply different discretization filters on numerical attributes and run the Aprior association rule algorithm. Study the rules generated. Derive interesting insights and observe the effect of discretization in the rule generation process.

Steps for run Aprior algorithm in WEKA

1. Open WEKA Tool.
2. Click on WEKA Explorer.
3. Click on Preprocessing tab button.
4. Click on open file button.

5. Choose WEKA folder in C drive.
6. Select and Click on data option button.
7. Choose Weather data set and open file.
8. Choose filter button and select the Unsupervised-Discritize option and apply
9. Click on Associate tab and Choose Aprior algorithm
10. Click on start button.



Demonstrate performing classification on data sets.

Classification Tab

Selecting a Classifier

Test Options

The result of applying the chosen classifier will be tested according to the options that are set by clicking in the Test options box. There are four test modes:

1. **Use training set.** The classifier is evaluated on how well it predicts the class of the instances it was trained on.

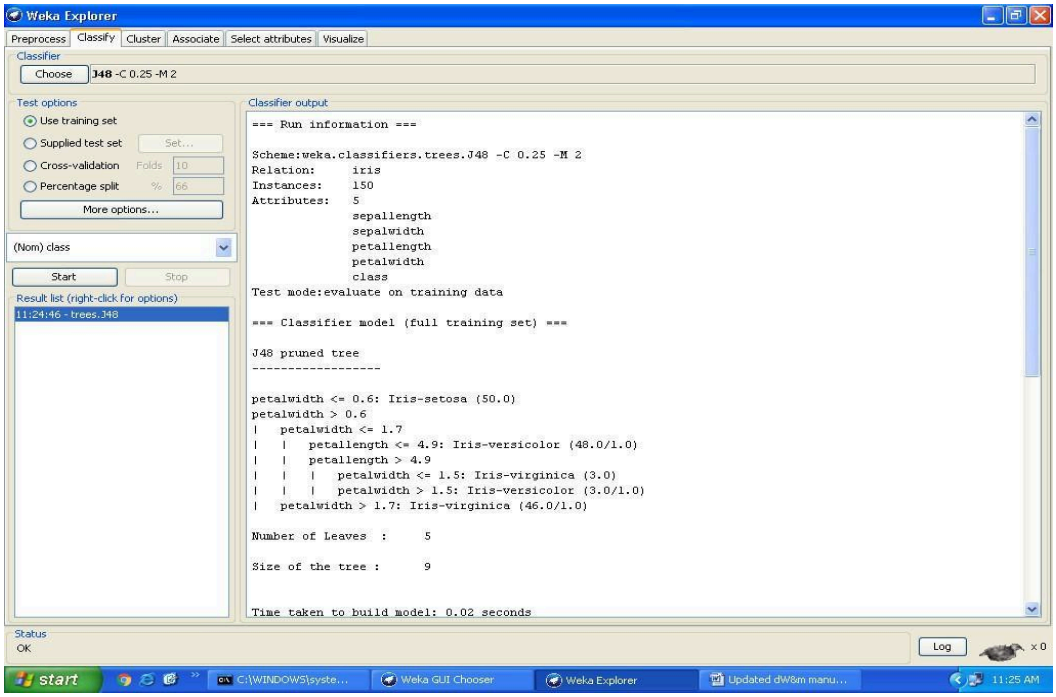
2. **Supplied test set.** The classifier is evaluated on how well it predicts the class of a set of instances loaded from a file. Clicking the Set... button brings up a dialog allowing you to choose the file to test on.

3. **Cross-validation.** The classifier is evaluated by cross-validation, using the number of folds that are entered in the Folds text field.
4. **Percentage split.** The classifier is evaluated on how well it predicts a certain percentage of the data which is held out for testing. The amount of data held out depends on the value entered in the
% field.

Load each dataset into Weka and run id3, j48 classification algorithm, study the classifier output. Compute entropy values, Kappa ststistic.

Steps for run ID3 and J48 Classification algorithms in WEKA

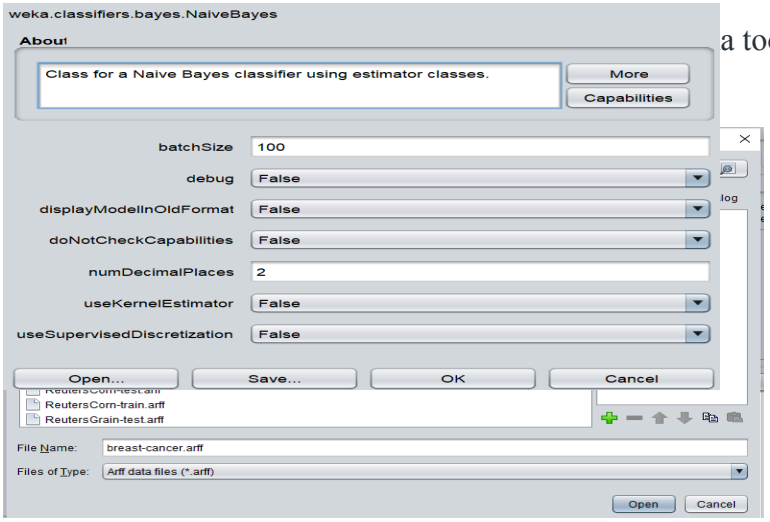
1. Open WEKA Tool.
2. Click on WEKA Explorer.
3. Click on Preprocessing tab button.
4. Click on open file button.
5. Choose WEKA folder in C drive.
6. Select and Click on data option button.
7. Choose iris data set and open file.
8. Click on classify tab and Choose J48 algorithm and select use training set test option.
9. Click on start button.
10. Click on classify tab and Choose ID3 algorithm and select use training set test option.
11. Click on start button.



Explore The Bayes' Classification Theorem:

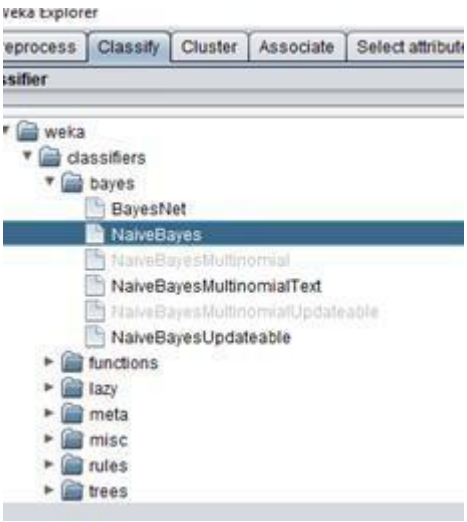
It is used to build a set of classification algorithms known as Naive Bayes classifiers. It is a family of algorithms that share a common concept, namely that each pair of features being classified is independent of the others.

Steps to be followed:



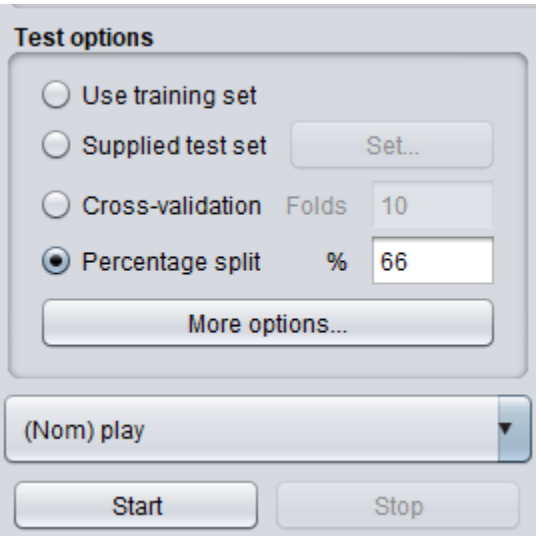
a tool using choose file option. Here we

Now we have to go to the classify tab on the top left side and click on the choose button and select the Naive Bayesian algorithm in it.



Now to change the parameters click on the right side at the choose button, and we are accepting the default values in this example.

We choose the Percentage split as our measurement method from the “Test” choices in the main panel. Since we don’t have a separate test data collection, we’ll use the percentage split of 66 percent to get a good idea of the model’s accuracy. Our dataset contains 14 examples, with h9 being used for training and 5 being used for testing.



To generate the model, we now click “start.” When the model is done, the evaluation statistic will appear in the right panel.

```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances           3               60      %
Incorrectly Classified Instances         2               40      %
Kappa statistic                          0
Mean absolute error                      0.4437
Root mean squared error                  0.5023
Relative absolute error                 93.8582 %
Root relative squared error            102.2471 %
Total Number of Instances                5

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    1.000    0.600      1.000    0.750      ?        0.667     0.867     yes
               0.000    0.000    ?          0.000    ?          ?        0.667     0.583     no
Weighted Avg.  0.600    0.600    ?          0.600    ?          ?        0.667     0.753

=== Confusion Matrix ===

 a b   <-- classified as
 3 0 | a = yes
 2 0 | b = no
```

Demonstrate performing clustering on data sets.

Cluster Tab

Selecting a Clusterer

By now you will be familiar with the process of selecting and configuring objects. Clicking on the clustering scheme listed in the Clusterer box at the top of the

window brings up a GenericObjectEditor dialog with which to choose a
new clustering scheme.

Cluster Modes

The Cluster mode box is used to choose what to cluster and how to evaluate

the results. The first three options are the same as for classification: Use training set, Supplied test set and Percentage split (Section 5.3.1)—except that now the data is assigned to clusters instead of trying to predict a specific class. The fourth mode, Classes to clusters evaluation, compares how well the chosen clusters match up with a pre-assigned class in the data. The drop-down box below this option selects the class, just as in the Classify panel.

An additional option in the Cluster mode box, the Store clusters for visualization tick box, determines whether or not it will be possible to visualize the clusters once training is complete. When dealing with datasets that are so large that memory becomes a problem it may be helpful to disable this option.

Learning Clusters

The Cluster section, like the Classify section, has Start/Stop buttons, a result text area and a result list. These all behave just like their classification counterparts. Right-clicking an entry in the result list brings up a similar menu, except that it shows only two visualization options: Visualize cluster assignments and Visualize tree. The latter is grayed out when it is not applicable.

Load each dataset into Weka and run simple k-means clustering algorithm with different values of k(number of desired clusters). Study the clusters formed. Observe the sum of squared errors and centroids, and derive insights.

Steps for run K-mean Clustering algorithms in WEKA

1. Open WEKA Tool.
2. Click on WEKA Explorer.
3. Click on Preprocessing tab button.
4. Click on open file button.
5. Choose WEKA folder in C drive.

6. Select and Click on data option button.
7. Choose iris data set and open file.
8. Click on cluster tab and Choose k-mean and select use training set test option.
9. Click on start button.

Output:

=== Run information ===

Scheme:weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500
-S 10
Relation: iris
Instances: 150

Test mode:evaluate on training data
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 7
Within cluster sum of squared errors:
62.1436882815797 Missing values globally
replaced with mean/mode

Cluster centroids:
Cluster#
Attribute Full Data 0 1
(150) (100) (50)
==========
Sepallength 5.8433 6.262 5.006
Sepalwidth 3.054 2.872 3.418
Petallength 3.7587 4.906 1.464
Petalwidth 1.1987 1.676 0.244
class Iris-setosa Iris-versicolor

Iris-setosa Time taken to

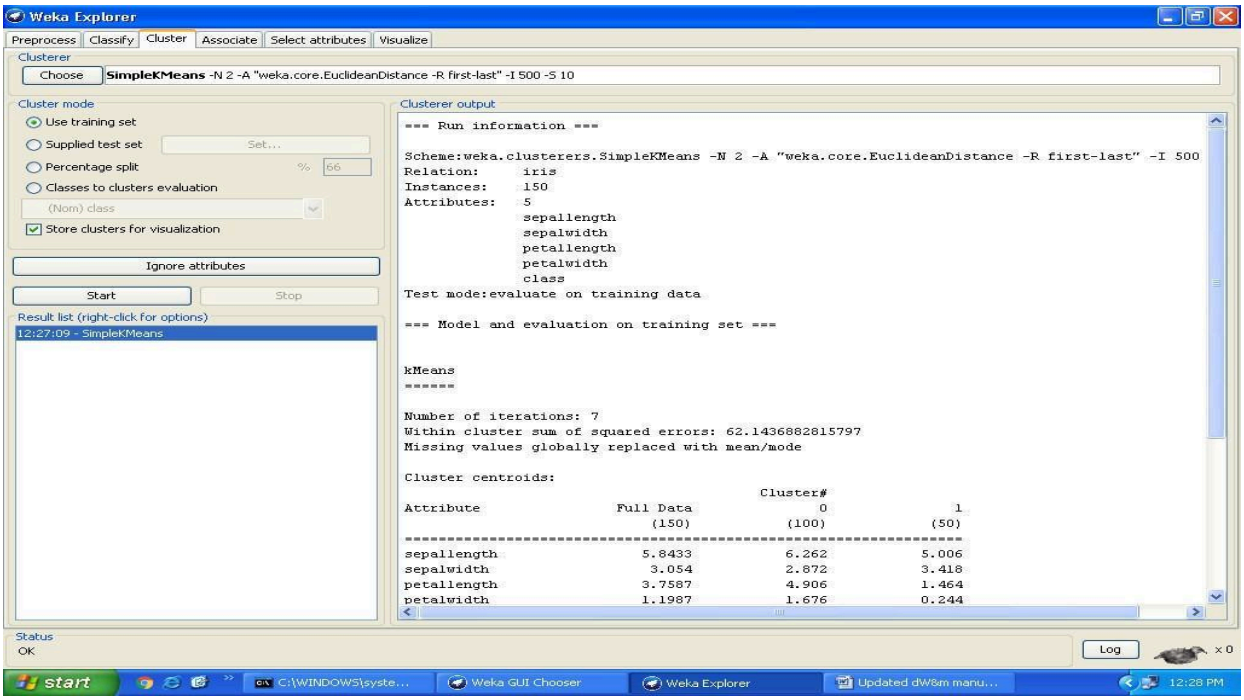
build model (full training data) : 0 seconds

=== Model and evaluation on

training set === Clustered

Instances

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



Result: The experiment had successfully performed data pre-processing task and demonstration of classification, Association and clustering al.