DATA ANALYSIS WITH WEKA

Author:

Nagamani Mutteni

Asst.Professor

MERI

Topic: Data Analysis with Weka

Course Duration: 2 Months

Objective:

Everybody talks about Data Mining and Big Data nowadays. Weka is a powerful, yet easy to use tool for machine learning and data mining. This course provides a deeper account of data mining tools and techniques. The emphasis is on principles and practical data mining using Weka, rather than mathematical theory or advanced details of particular algorithms. Students will work with multimillion-instance datasets, classify text, experiment with clustering, association rules, etc.

Assessment criteria: After completion of program students are awarded certificate after clearing a MCQ based examination.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION TO WEKA

1.1 Introduction:

What is WEKA?

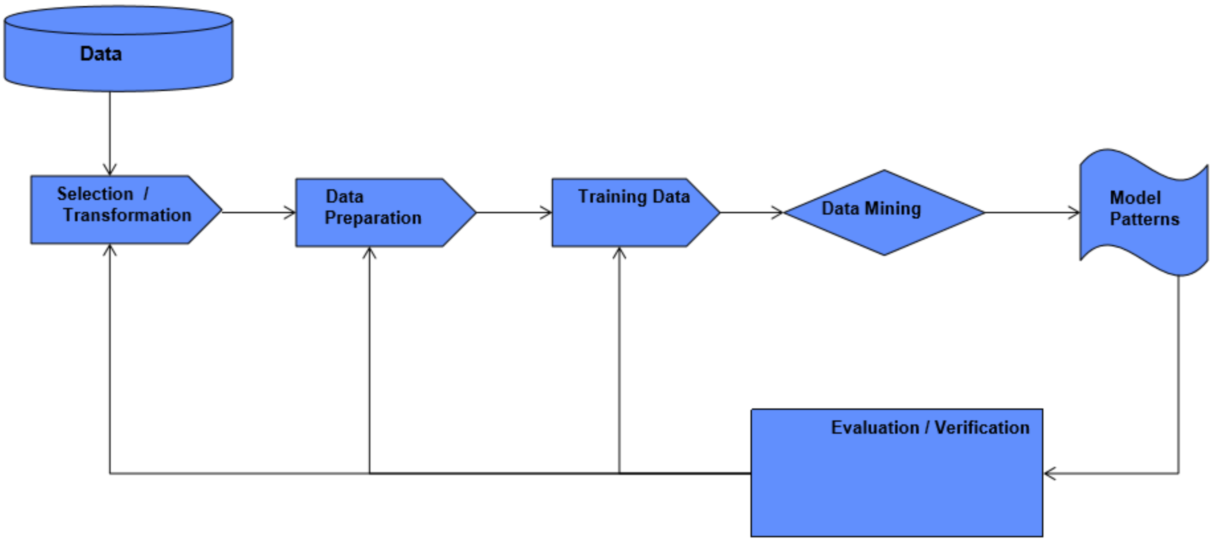
WEKA, formally called **Waikato Environment for Knowledge Learning**, is a computer program that was developed at the University of Waikato in New Zealand for the purpose of identifying information from raw data gathered from different domains.

WEKA supports many different standard data mining tasks such as data pre-processing, 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.

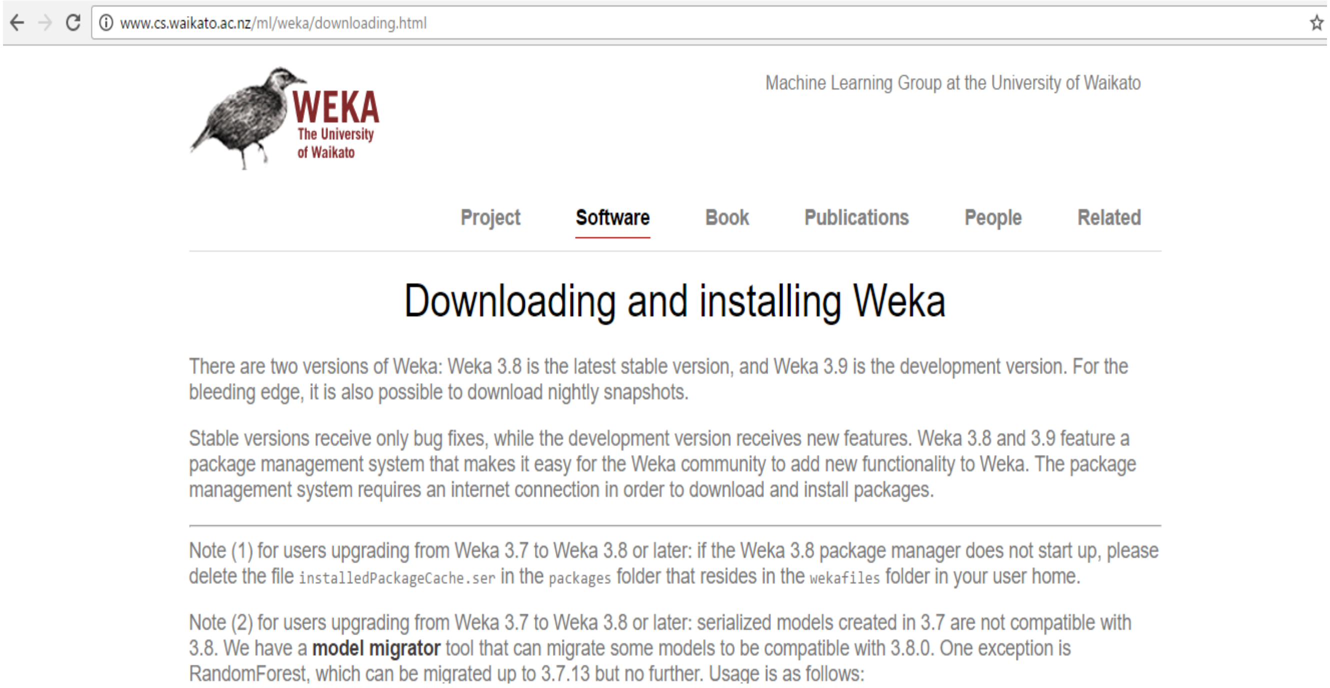
WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces.

1.2 KDD Process:



1.3 Installation of Weka:

The weka can be explored from the different sites, one of the sites is http://www.cs.waikato.ac.nz/m1/weka/downloading.html


www.cs.waikato.ac.nz/ml/weka/downloading.html
WEKA
The University of Waikato
Machine Learning Group at the University of Waikato
Project
Software
Book
Publications
People
Related

Downloading and installing Weka

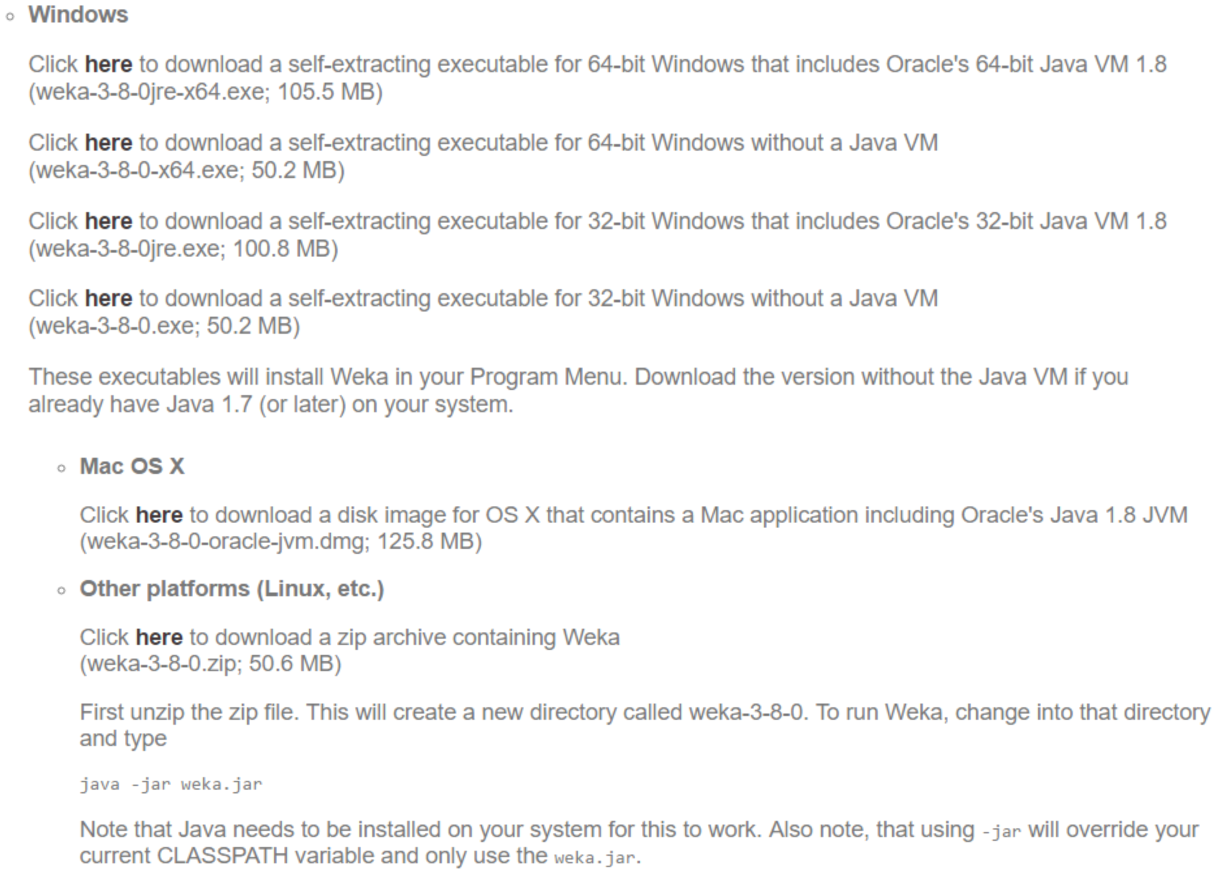
There are two versions of Weka: Weka 3.8 is the latest stable version, and Weka 3.9 is the development version. For the bleeding edge, it is also possible to download nightly snapshots.

Stable versions receive only bug fixes, while the development version receives new features. Weka 3.8 and 3.9 feature a package management system that makes it easy for the Weka community to add new functionality to Weka. The package management system requires an internet connection in order to download and install packages.

Note (1) for users upgrading from Weka 3.7 to Weka 3.8 or later: if the Weka 3.8 package manager does not start up, please delete the file `installedPackageCache.ser` in the `packages` folder that resides in the `wekafiles` folder in your user home.

Note (2) for users upgrading from Weka 3.7 to Weka 3.8 or later: serialized models created in 3.7 are not compatible with 3.8. We have a **model migrator** tool that can migrate some models to be compatible with 3.8.0. One exception is RandomForest, which can be migrated up to 3.7.13 but no further. Usage is as follows:

There are different options to launch weka depending the operating systems



- **Windows**

  Click **here** to download a self-extracting executable for 64-bit Windows that includes Oracle's 64-bit Java VM 1.8 (weka-3-8-0jre-x64.exe; 105.5 MB)

  Click **here** to download a self-extracting executable for 64-bit Windows without a Java VM (weka-3-8-0-x64.exe; 50.2 MB)

  Click **here** to download a self-extracting executable for 32-bit Windows that includes Oracle's 32-bit Java VM 1.8 (weka-3-8-0jre.exe; 100.8 MB)

  Click **here** to download a self-extracting executable for 32-bit Windows without a Java VM (weka-3-8-0.exe; 50.2 MB)

  These executables will install Weka in your Program Menu. Download the version without the Java VM if you already have Java 1.7 (or later) on your system.

  - **Mac OS X**

    Click **here** to download a disk image for OS X that contains a Mac application including Oracle's Java 1.8 JVM (weka-3-8-0-oracle-jvm.dmg; 125.8 MB)

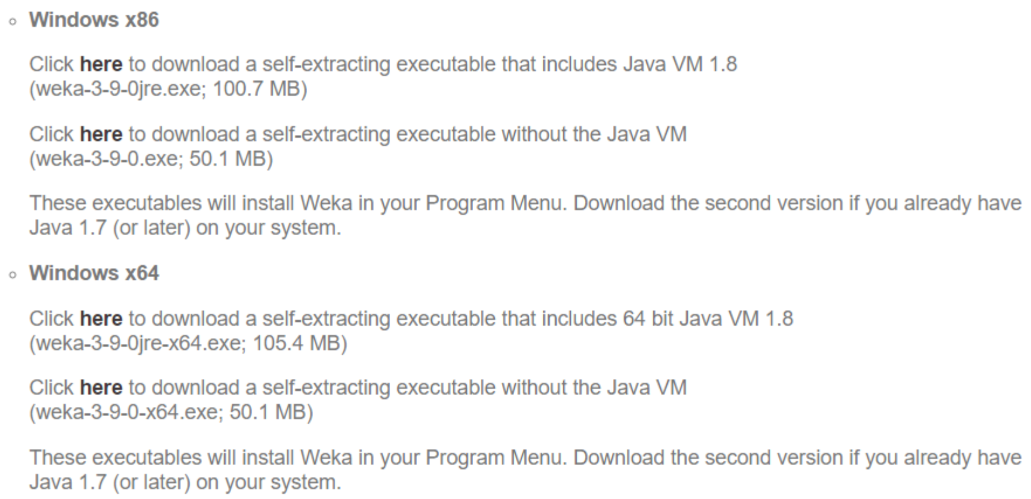
  - **Other platforms (Linux, etc.)**

    Click **here** to download a zip archive containing Weka (weka-3-8-0.zip; 50.6 MB)

    First unzip the zip file. This will create a new directory called weka-3-8-0. To run Weka, change into that directory and type

    ```
    java -jar weka.jar
    ```

    Note that Java needs to be installed on your system for this to work. Also note, that using `-jar` will override your current CLASSPATH variable and only use the `weka.jar`.



- **Windows x86**

  Click **here** to download a self-extracting executable that includes Java VM 1.8
  (weka-3-9-0jre.exe; 100.7 MB)

  Click **here** to download a self-extracting executable without the Java VM
  (weka-3-9-0.exe; 50.1 MB)

  These executables will install Weka in your Program Menu. Download the second version if you already have Java 1.7 (or later) on your system.

- **Windows x64**

  Click **here** to download a self-extracting executable that includes 64 bit Java VM 1.8
  (weka-3-9-0jre-x64.exe; 105.4 MB)

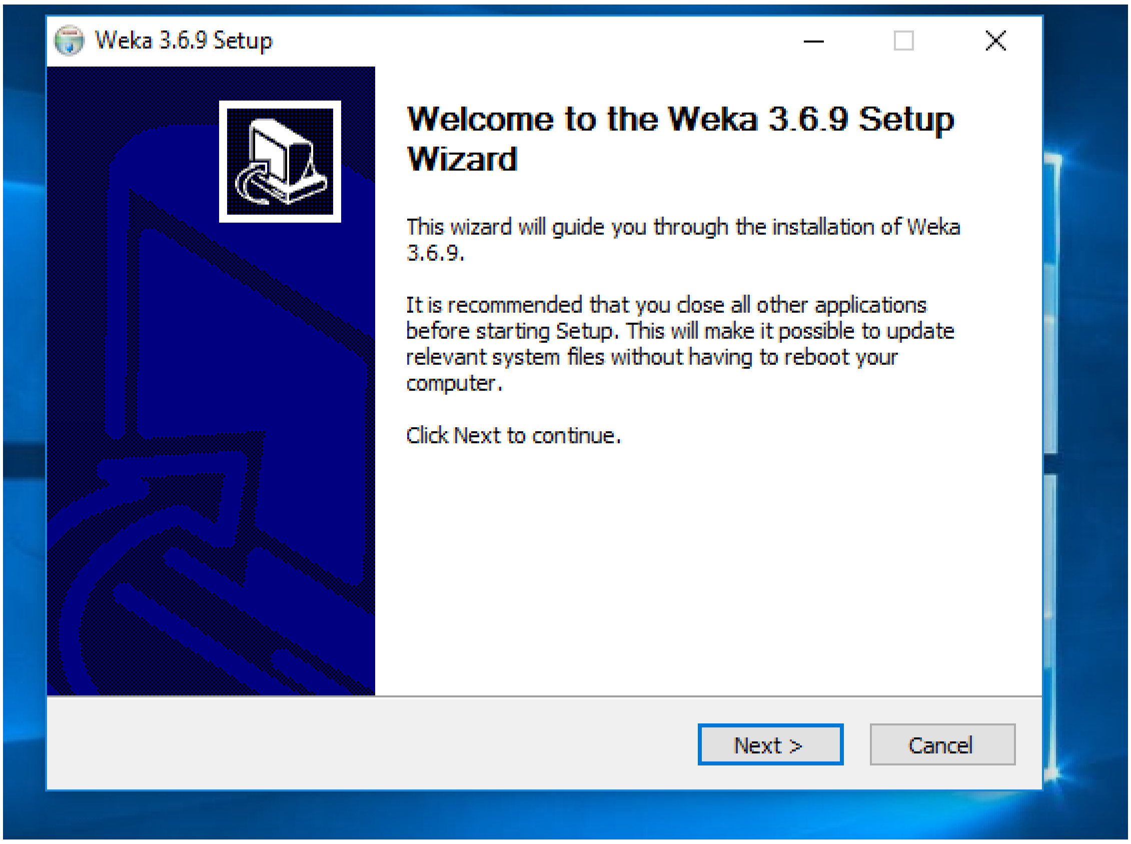
  Click **here** to download a self-extracting executable without the Java VM
  (weka-3-9-0-x64.exe; 50.1 MB)

  These executables will install Weka in your Program Menu. Download the second version if you already have Java 1.7 (or later) on your system.

Depending on the version click on the down load option. When we click on the download option setup of weka gets downloaded.

Click on setup and follow the below steps

Step 1:



Welcome to the Weka 3.6.9 Setup Wizard

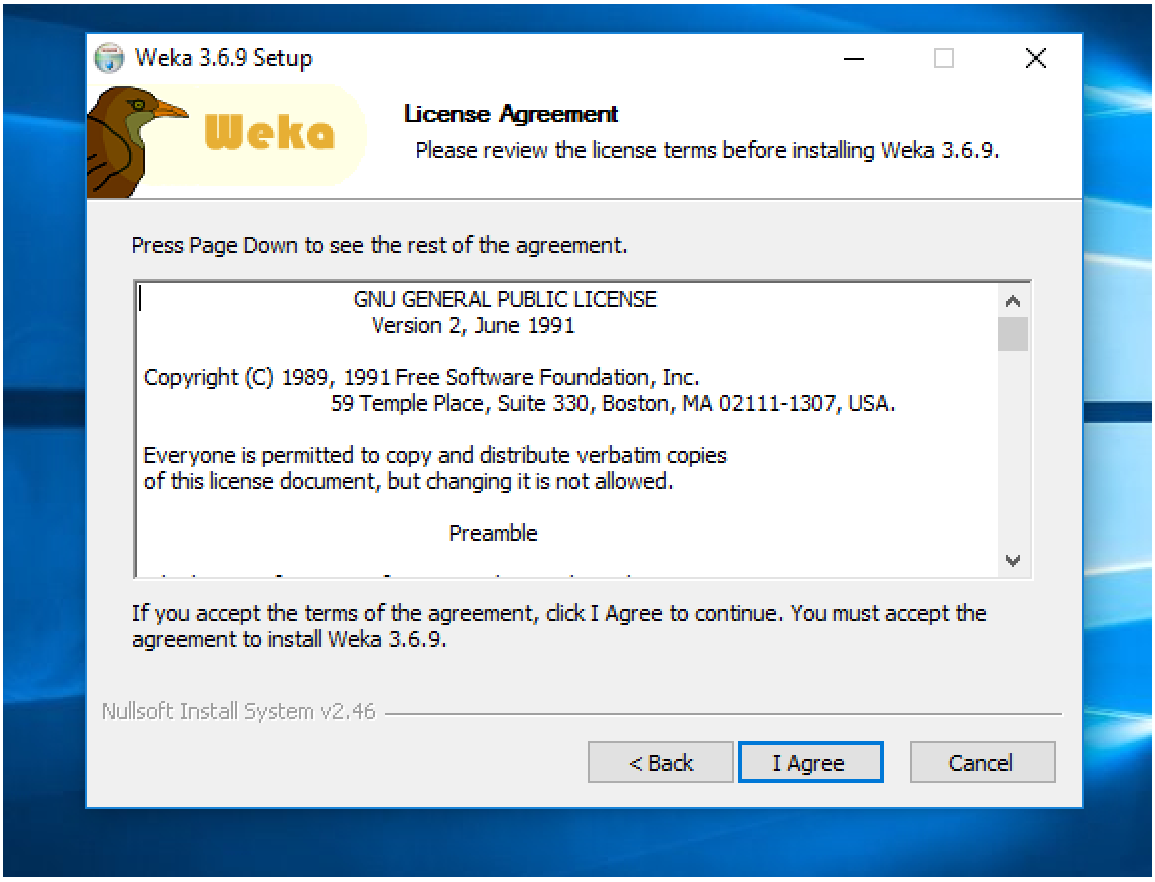
This wizard will guide you through the installation of Weka 3.6.9.

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Click on Next button

Step 2:



License Agreement

Please review the license terms before installing Weka 3.6.9.

Press Page Down to see the rest of the agreement.

GNU GENERAL PUBLIC LICENSE
Version 2, June 1991

Copyright (C) 1989, 1991 Free Software Foundation, Inc.
59 Temple Place, Suite 330, Boston, MA 02111-1307, USA.

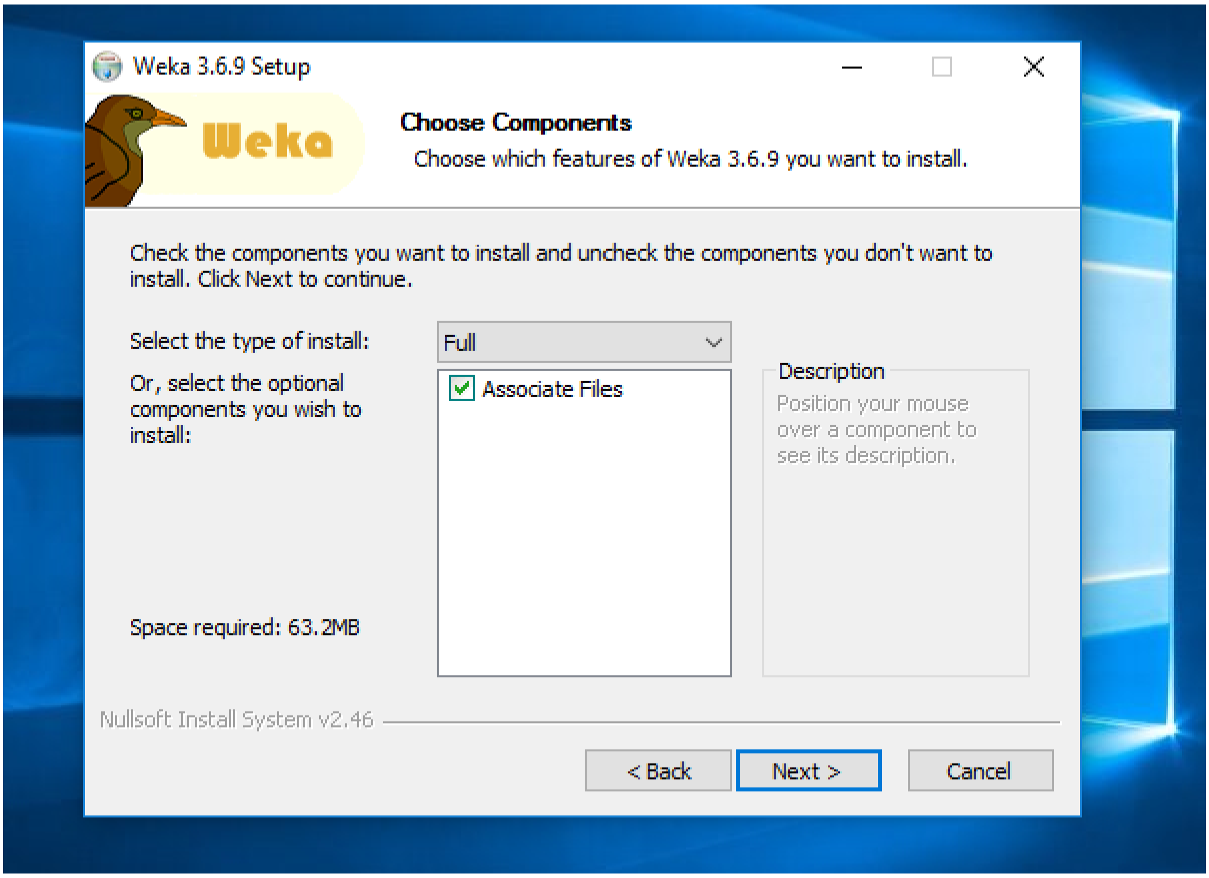
Everyone is permitted to copy and distribute verbatim copies
of this license document, but changing it is not allowed.

Preamble

If you accept the terms of the agreement, click I Agree to continue. You must accept the agreement to install Weka 3.6.9.

click on I Agree option

Step 3:



Weka 3.6.9 Setup

Choose Components

Choose which features of Weka 3.6.9 you want to install.

Check the components you want to install and uncheck the components you don't want to install. Click Next to continue.

Select the type of install: Full

Or, select the optional components you wish to install:

☑ Associate Files

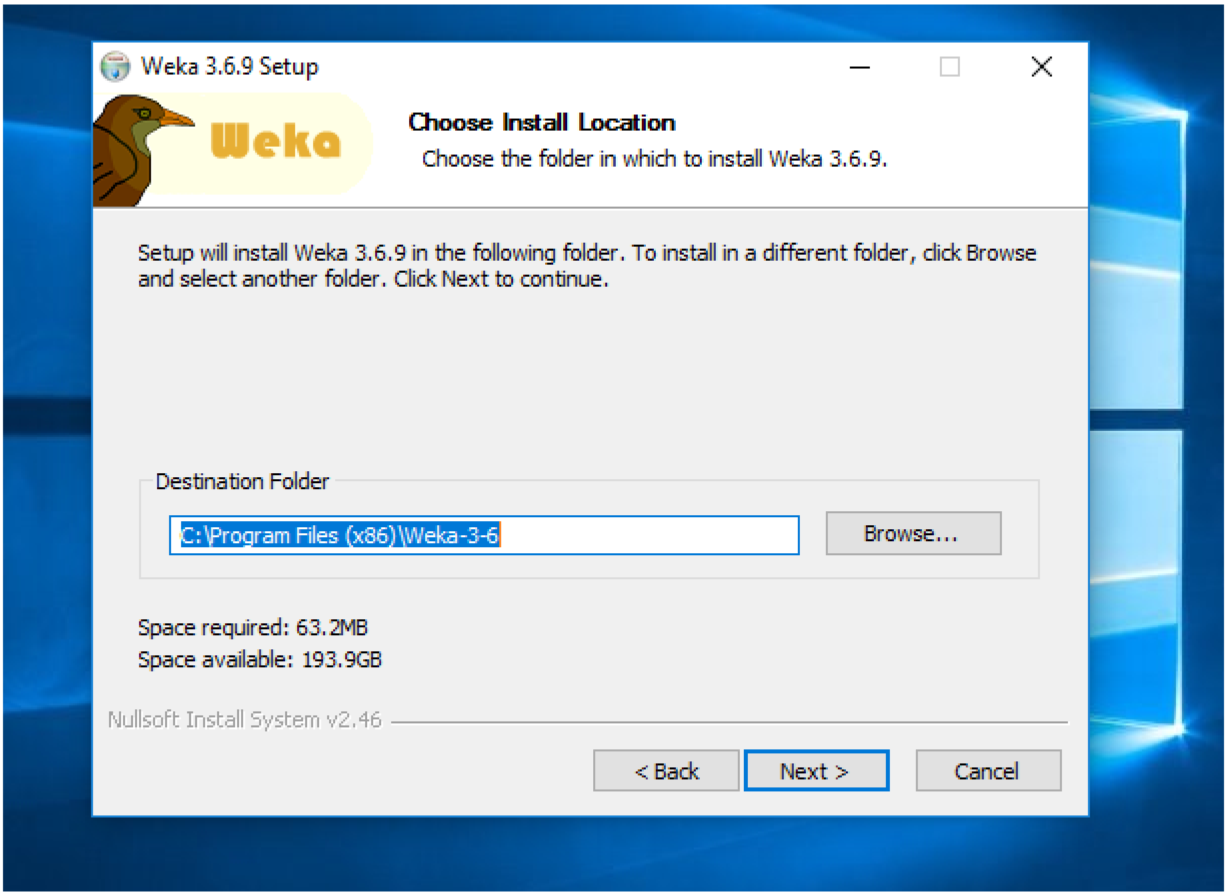
Description

Position your mouse over a component to see its description.

Space required: 63.2MB

Nullsoft Install System v2.46

Click on Next Option



Weka 3.6.9 Setup

Choose Install Location

Choose the folder in which to install Weka 3.6.9.

Setup will install Weka 3.6.9 in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder

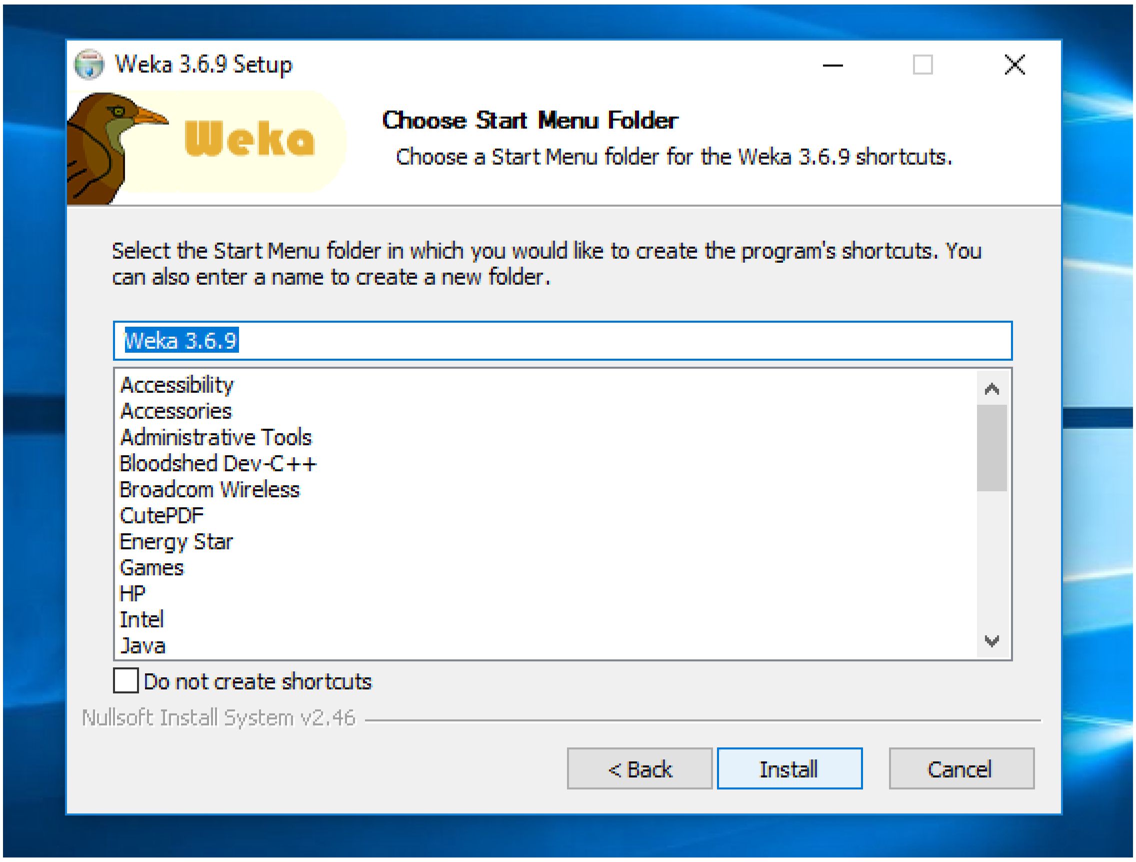
C:\Program Files (x86)\Weka-3-6

Space required: 63.2MB

Space available: 193.9GB

Nullsoft Install System v2.46

Click on Next option



Weka 3.6.9 Setup

Choose Start Menu Folder

Choose a Start Menu folder for the Weka 3.6.9 shortcuts.

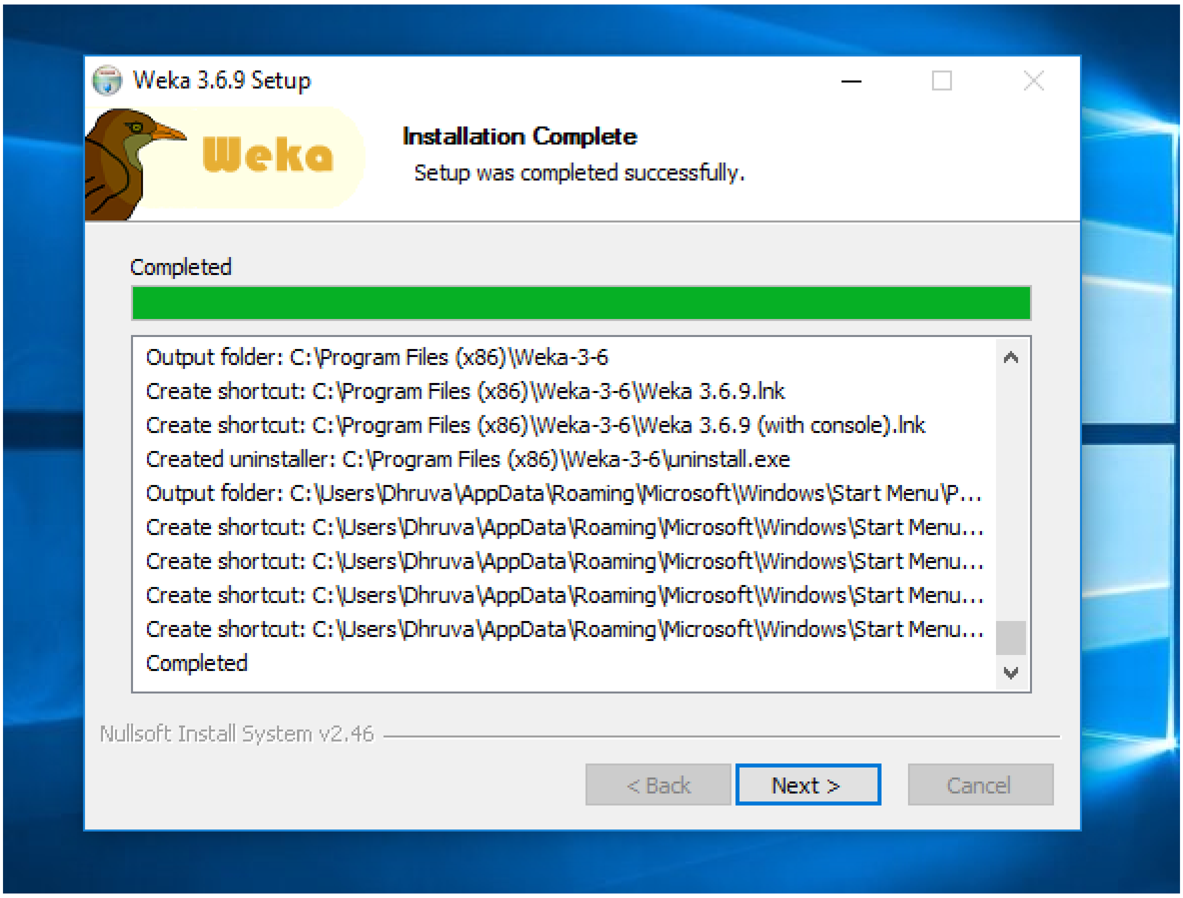
Select the Start Menu folder in which you would like to create the program's shortcuts. You can also enter a name to create a new folder.

Weka 3.6.9

- Accessibility
- Accessories
- Administrative Tools
- Bloodshed Dev-C++
- Broadcom Wireless
- CutePDF
- Energy Star
- Games
- HP
- Intel
- Java

□ Do not create shortcuts

Click on install button.It extracts all the packages



Installation Complete

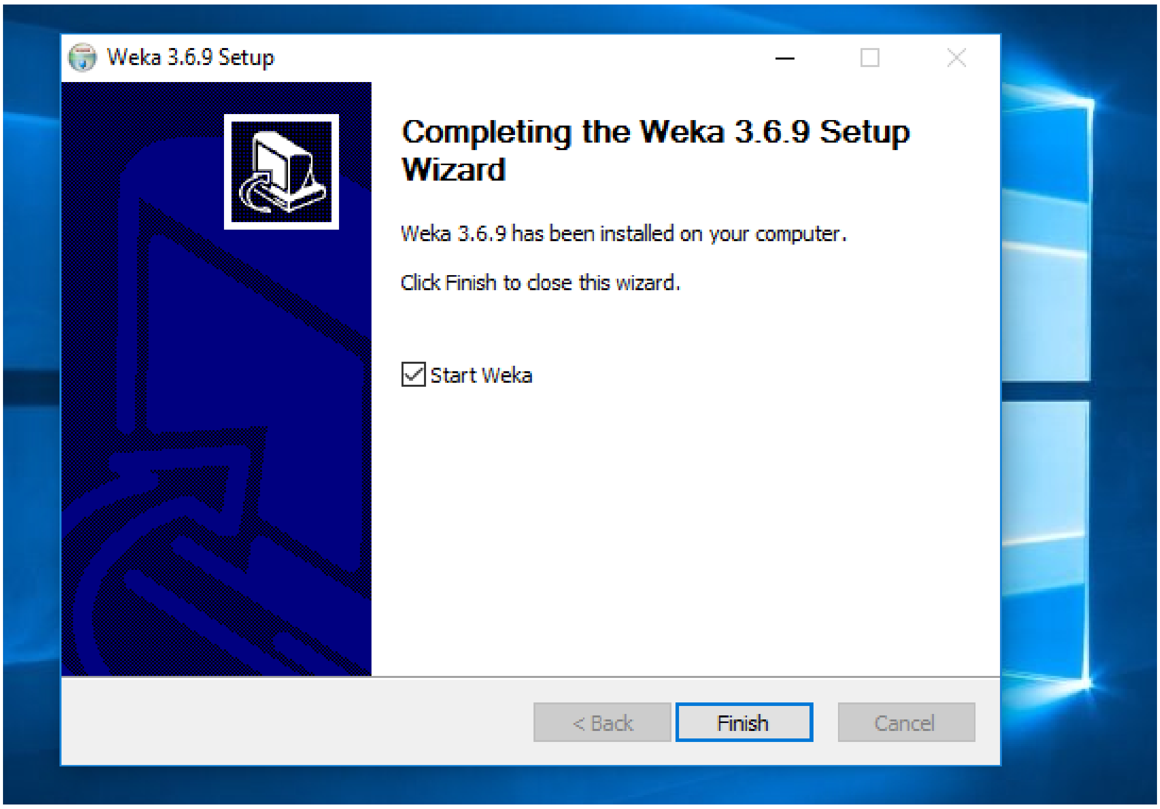
Setup was completed successfully.

Completed

```
Output folder: C:\Program Files (x86)\Weka-3-6
Create shortcut: C:\Program Files (x86)\Weka-3-6\Weka 3.6.9.lnk
Create shortcut: C:\Program Files (x86)\Weka-3-6\Weka 3.6.9 (with console).lnk
Created uninstaller: C:\Program Files (x86)\Weka-3-6\uninstall.exe
Output folder: C:\Users\Dhruva\AppData\Roaming\Microsoft\Windows\Start Menu\P...
Create shortcut: C:\Users\Dhruva\AppData\Roaming\Microsoft\Windows\Start Menu...
Create shortcut: C:\Users\Dhruva\AppData\Roaming\Microsoft\Windows\Start Menu...
Create shortcut: C:\Users\Dhruva\AppData\Roaming\Microsoft\Windows\Start Menu...
Create shortcut: C:\Users\Dhruva\AppData\Roaming\Microsoft\Windows\Start Menu...
Completed
```

Nullsoft Install System v2.46

Click on next button



Completing the Weka 3.6.9 Setup Wizard

Weka 3.6.9 has been installed on your computer.

Click Finish to close this wizard.

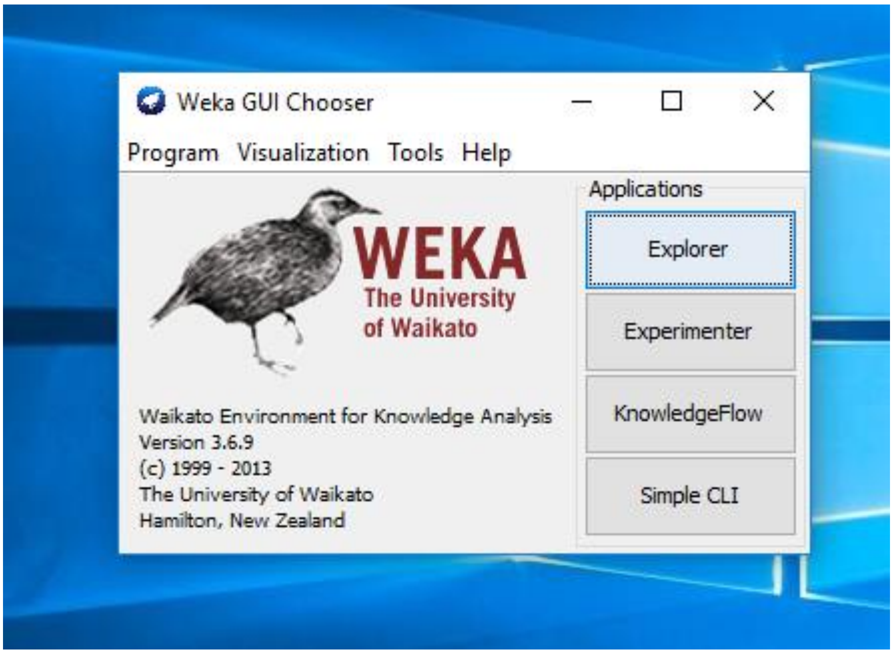
☑ Start Weka

Click on finish button

CHAPTER 2: LAUNCHING WEKA EXPLORER

2.1 Starting with Weka

Once the program has been loaded on the user's machine it is opened by navigating to the programs start option and that will depend on the user's operating system.



Waikato Environment for Knowledge Analysis
Version 3.6.9
(c) 1999 - 2013
The University of Waikato
Hamilton, New Zealand

There are four options available on this initial screen.

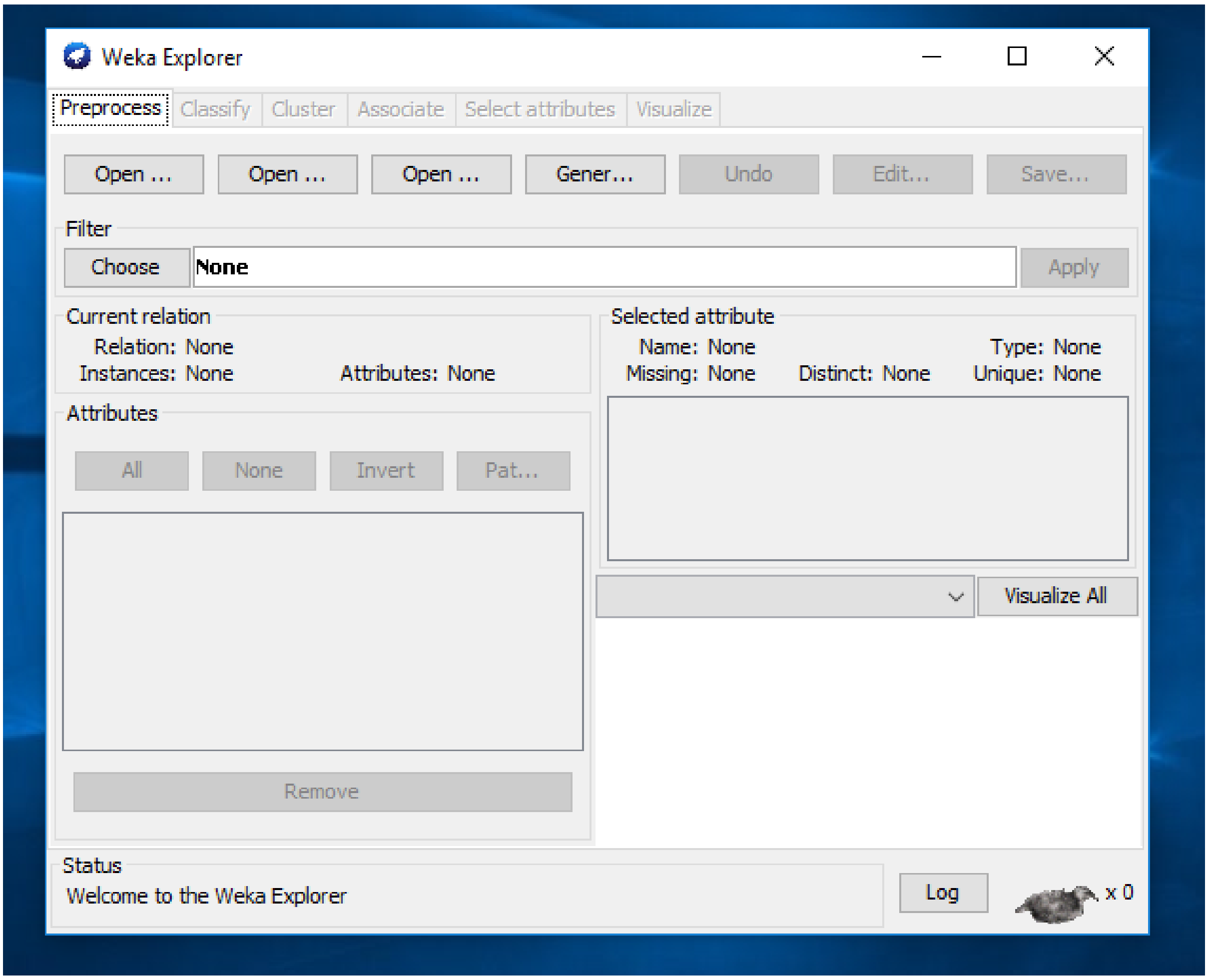
1. **Explorer-** the graphical interface used to conduct experimentation on raw data

2. **Simple CLI-** provides users without a graphic interface option the ability to execute commands from a terminal window.

3. **Experimenter-** this option allows users to conduct different experimental variations on data sets and perform statistical manipulation

4. **Knowledge Flow-**basically the same functionality as Explorer with drag and drop functionality. The advantage of this option is that it supports incremental learning from previous results.

After selecting the Explorer option the program starts and provides the user with a separate graphical interface.



Filter: None

Relation: None
Instances: None
Attributes: None

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

Welcome to the Weka Explorer

Figure: 2

Figure 2 shows the opening screen with the available options. At first there is only the option to select the Pre-process tab in the top left corner. This is due to the necessity to present the data set to the application so it can be manipulated. After the data has been pre-processed the other tabs become active for use.

There are six tabs:

1**. Pre-process-** used to choose the data file to be used by the application

2. **Classify-** used to test and train different learning schemes on the pre-processed data file under experimentation

3. **Cluster-** used to apply different tools that identify clusters within the data file

4. **Association-** used to apply different rules to the data file that identify association within the data

5. **Select attributes-**used to apply different rules to reveal changes based on selected attributes inclusion or exclusion from the experiment

6. **Visualize-** used to see what the various manipulation produced on the data set in a 2D format, in scatter plot and bar graph output.

2.2 Pre-processing:

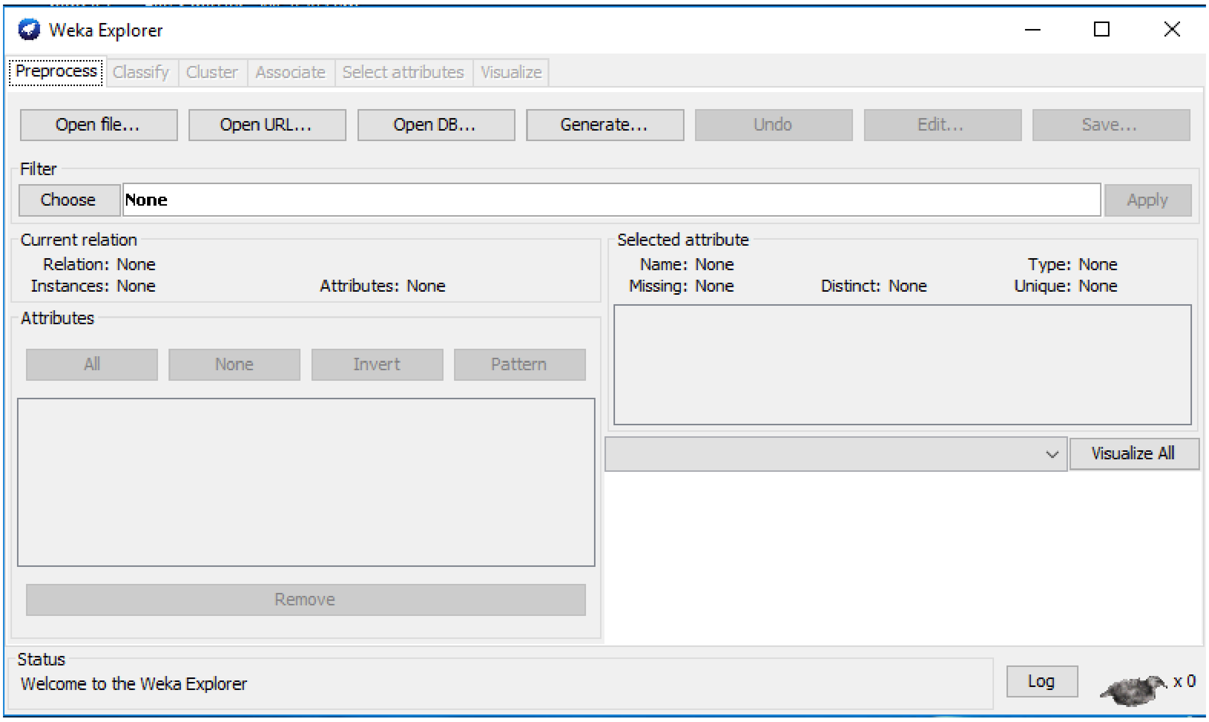
In order to experiment with the application the data set needs to be presented to WEKA in a format that the program understands. There are rules for the type of data that WEKA will accept. There are three options for presenting data into the program.

♦ **Open File-** allows for the user to select files residing on the local machine or recorded medium.

♦ **Open URL-** provides a mechanism to locate a file or data source from a different location specified by the user.

♦ **Open Database-** allows the user to retrieve files or data from a database source provided by the user.

There are restrictions on the type of data that can be accepted into the program. Originally the software was designed to import only ARFF files, other versions allow different file types such as CSV, C4.5 and serialized instance formats. The extensions for these files include .csv, .arff, .names, .bsi and .data.



Filter: None

Current relation
Relation: None
Instances: None
Attributes: None

Selected attribute
Name: None
Missing: None
Distinct: None
Type: None
Unique: None

Attributes

Status
Welcome to the Weka Explorer

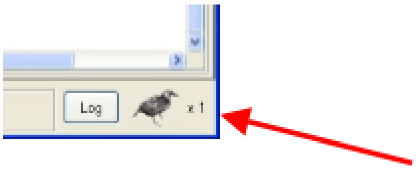
At the bottom of the window there is 'Status' box. The 'Status' box displays messages that keep you informed about what is going on. For example, when you first opened the 'Explorer', the message says, "Welcome to the Weka Explorer". When you loading "weather.arff" file, the 'Status' box displays the message "Reading from file…". Once the file is loaded, the message in the 'Status' box changes to say "OK". Right-click anywhere in 'Status box', it brings up a menu with two options:

1. **Available Memory** that displays in the log and in 'Status' box the amount of memory available to WEKA in bytes.

2. **Run garbage collector** that forces Java garbage collector to search for memory that is no longer used, free this memory up and to allow this memory for new tasks.

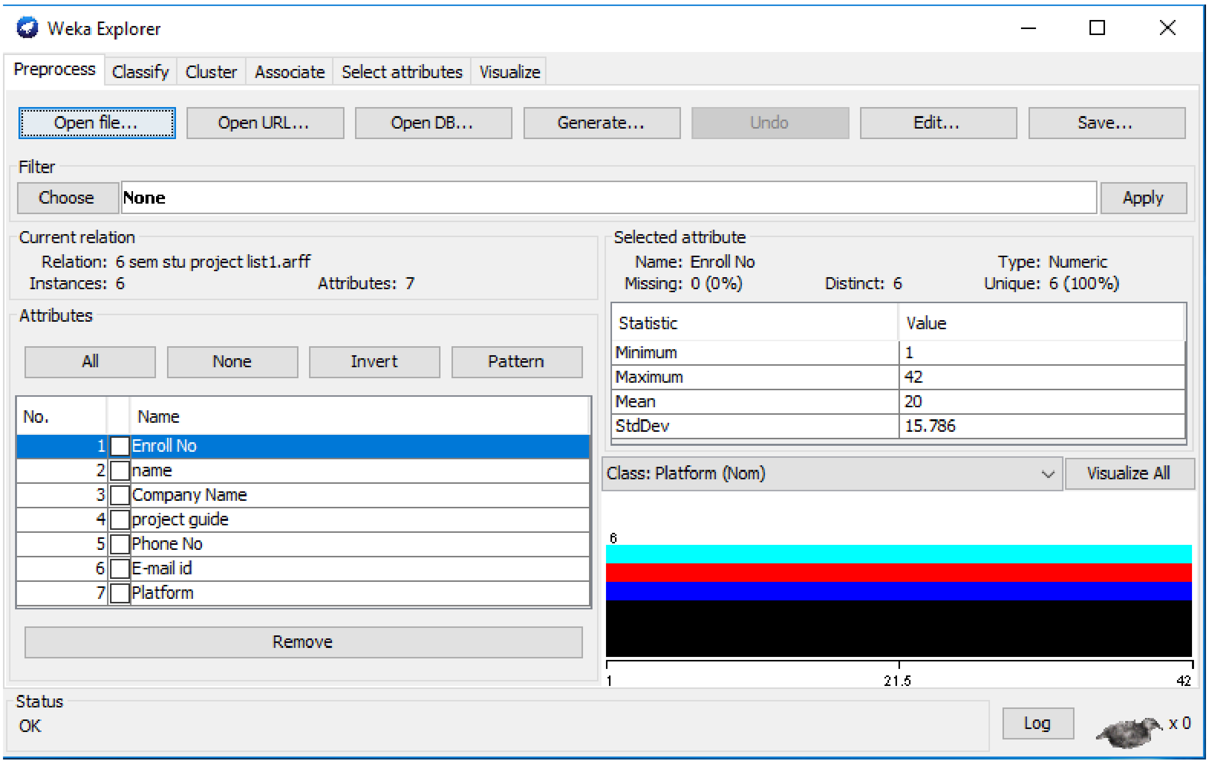
To the right of ‘Status box’ there is a ‘Log’ button that opens up the log. The log records every action in WEKA and keeps a record of what has happened. Each line of text in the log contains time of entry. For example, if the file you tried to open is not loaded, the log will have record of the problem that occurred during opening.

To the right of the ‘Log’ button there is an image of a bird. The bird is WEKA status icon. The number next to ‘X’ symbol indicates a number of concurrently running processes. When you loading a file, the bird sits down that means that there are no processes running. The number of processes besides symbol ‘X’ is zero that means that the system is idle. Later, in classification problem, when generating result look at the bird, it gets up and start moving that indicates that a process started. The number next to ‘X’ becomes 1 that means that there is one process running, in this case calculation.



2.3 Loading data:

The most common and easiest way of loading data into WEKA is from ARFF file, using ‘Open file…’ button . Click on ‘Open file…’ button and choose “project details ” file from your local filesystem. Note, the data can be loaded from CSV file as well because some databases have the ability to **convert data only into CSV format.**



Filter: None

Current relation
Relation: 6 sem stu project list1.arff
Instances: 6 Attributes: 7

Selected attribute
Name: Enroll No Type: Numeric
Missing: 0 (0%) Distinct: 6 Unique: 6 (100%)

Statistic	Value
Minimum	1
Maximum	42
Mean	20
StdDev	15.786

Attributes

No.		Name
1	□	Enroll No
2	□	name
3	□	Company Name
4	□	project guide
5	□	Phone No
6	□	E-mail id
7	□	Platform

Class: Platform (Nom)

Status
OK

Once the data is loaded, WEKA recognizes attributes that are shown in the ‘Attribute’ window. Left panel of ‘Preprocess’ window shows the list of recognized attributes:

No: is a number that identifies the order of the attribute as they are in data file.

Selection tick boxes: allow you to select the attributes for working relation.

Name: is a name of an attribute as it was declared in the data file.

The ‘Current relation’ box above ‘Attribute’ box displays the base relation (table) name and the current working relation - “project details ”, the number of instances - 6 and the number of attributes – 7.

During the scan of the data, WEKA computes some basic statistics on each attribute. The following statistics are shown in ‘Selected attribute’ box on the right panel of ‘Preprocess’ window:

Name is the name of an attribute.

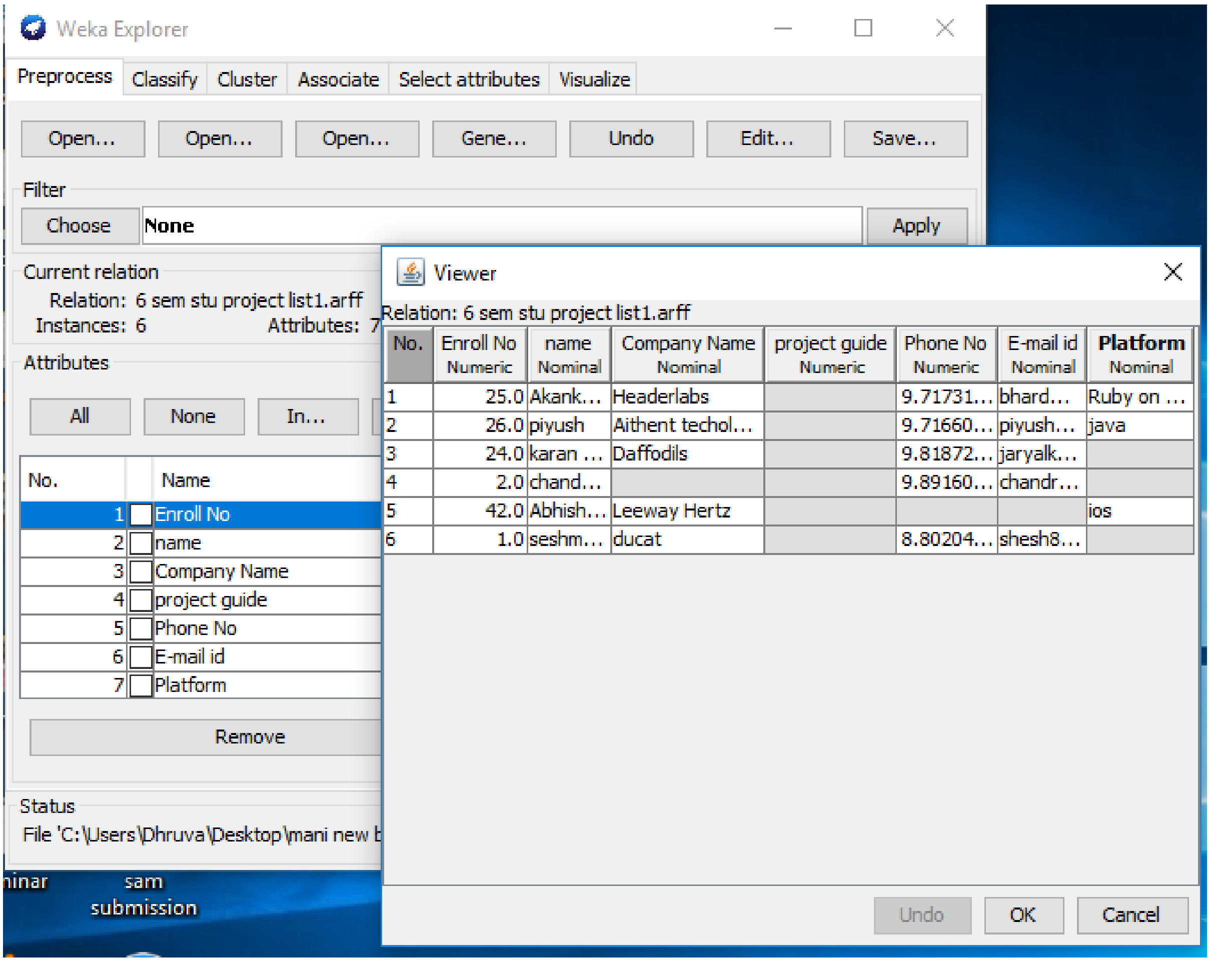
Type is most commonly Nominal or Numeric.

Missing is the number (percentage) of instances in the data for which this attribute is unspecified.

Distinct is the number of different values that the data contains for this attribute.

Unique is the number (percentage) of instances in the data having a value for this attribute that no other instances have.

Once the data is loaded into weka changes can be made to the attributes by clicking edit button shown above.



Filter: None

Current relation
Relation: 6 sem stu project list1.arff
Instances: 6 Attributes: 7

Attributes

No.		Name
1	□	Enroll No
2	□	name
3	□	Company Name
4	□	project guide
5	□	Phone No
6	□	E-mail id
7	□	Platform

Status
File 'C:\Users\Dhruva\Desktop\mani new b

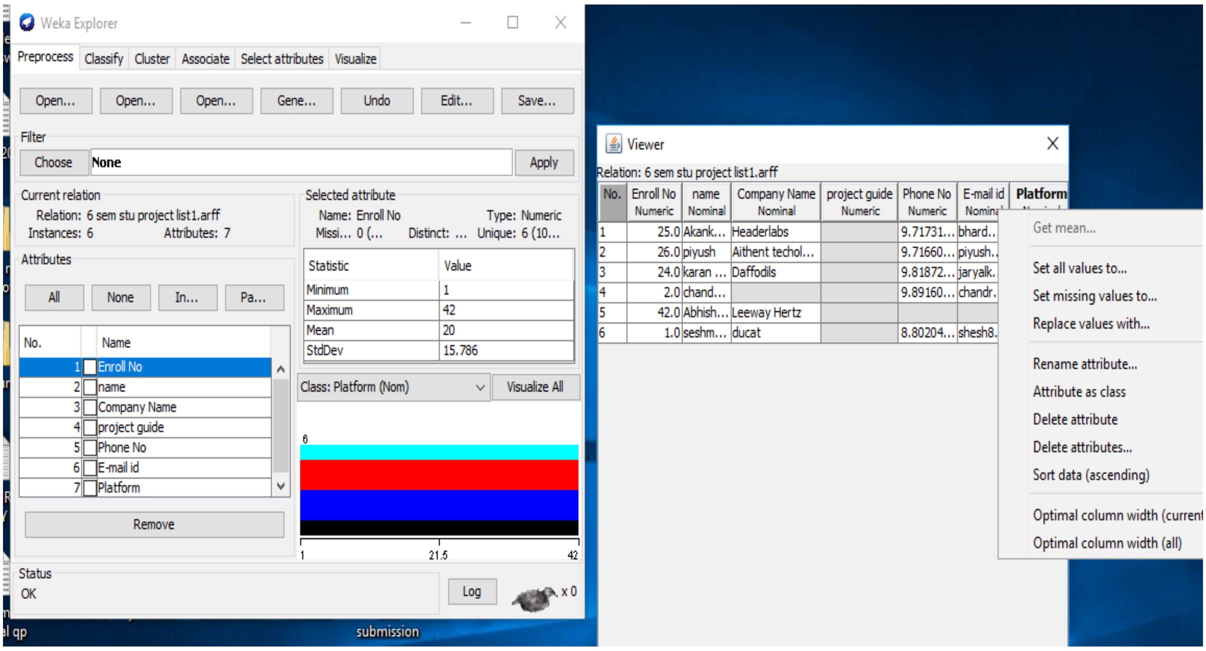
Relation: 6 sem stu project list1.arff

No.	Enroll No Numeric	name Nominal	Company Name Nominal	project guide Numeric	Phone No Numeric	E-mail id Nominal	**Platform** Nominal
1	25.0	Akank...	Headerlabs		9.71731...	bhard...	Ruby on ...
2	26.0	piyush	Aithent techol...		9.71660...	piyush...	java
3	24.0	karan ...	Daffodils		9.81872...	jaryalk...	
4	2.0	chand...			9.89160...	chandr...	
5	42.0	Abhish...	Leeway Hertz				ios
6	1.0	seshm...	ducat		8.80204...	shesh8...	

To make the changes double click on the attribute value and update the details as user required .

Different operations can be performed through edit are as follows:

1) delete the attribute
2) Replace the attribute value
3) Set all values
4) Set missing values etc.



Filter: None

Current relation
Relation: 6 sem stu project list1.arff
Instances: 6 Attributes: 7

Selected attribute
Name: Enroll No Type: Numeric
Missi... 0 (... Distinct: ... Unique: 6 (10...

Statistic	Value
Minimum	1
Maximum	42
Mean	20
StdDev	15.786

Attributes

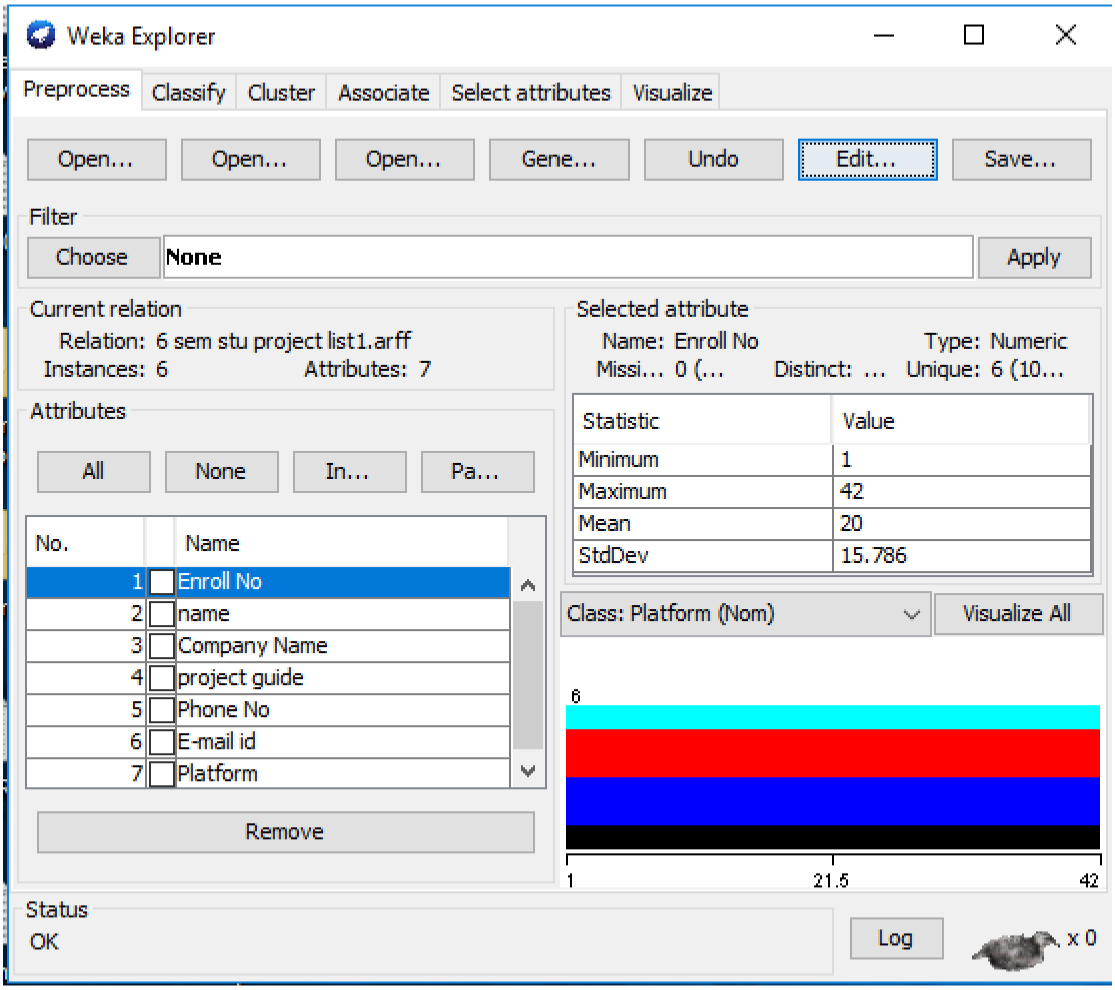
No.		Name
1	□	Enroll No
2	□	name
3	□	Company Name
4	□	project guide
5	□	Phone No
6	□	E-mail id
7	□	Platform

Status
OK

Relation: 6 sem stu project list1.arff

No.	Enroll No Numeric	name Nominal	Company Name Nominal	project guide Numeric	Phone No Numeric	E-mail id Nominal	Platform
1	25.0	Akank...	Headerlabs		9.71731...	bhard...	
2	26.0	piyush	Aithent techol...		9.71660...	piyush...	
3	24.0	karan ...	Daffodils		9.81872...	jaryalk...	
4	2.0	chand...			9.89160...	chandr...	
5	42.0	Abhish...	Leeway Hertz				
6	1.0	seshm...	ducat		8.80204...	shesh8...	

After update of values the minimum ,maximum , mean and standard deviation values gets changed.



Filter: None

Current relation
Relation: 6 sem stu project list1.arff
Instances: 6 Attributes: 7

Selected attribute
Name: Enroll No Type: Numeric
Missi... 0 (... Distinct: ... Unique: 6 (10...

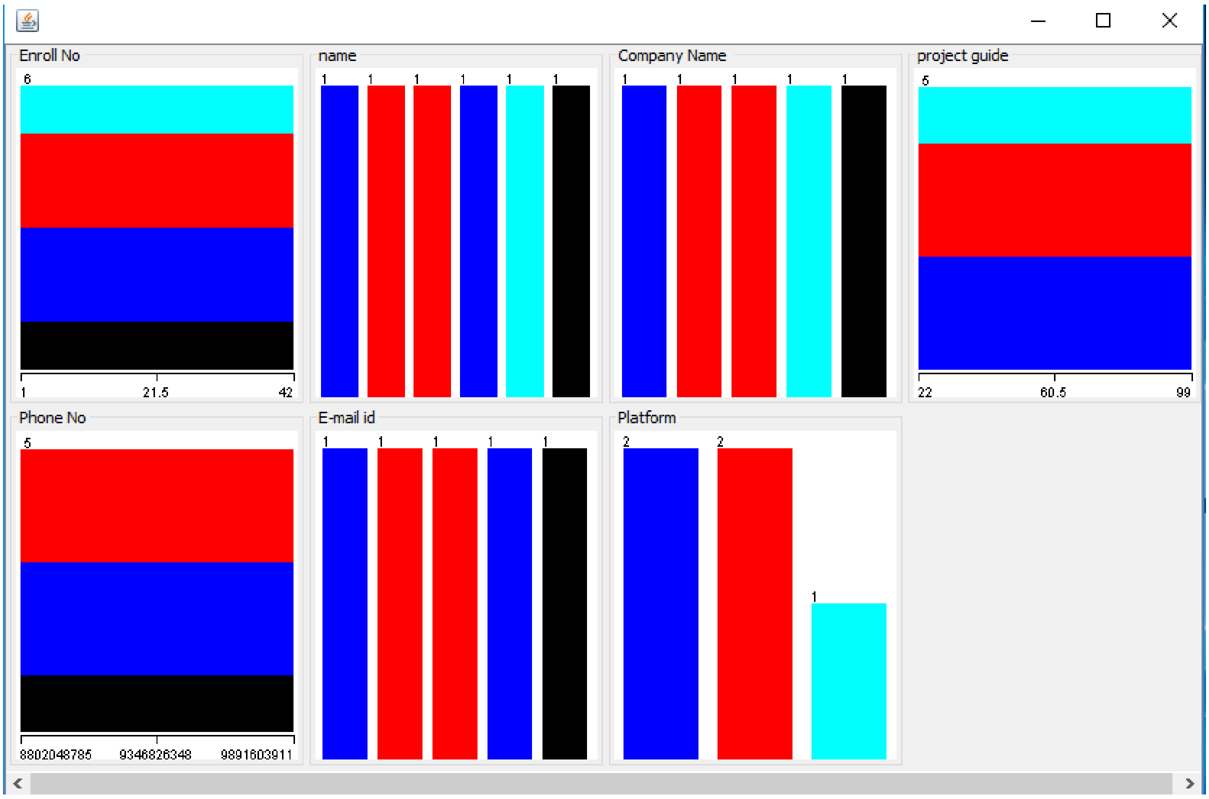
Statistic	Value
Minimum	1
Maximum	42
Mean	20
StdDev	15.786

Attributes

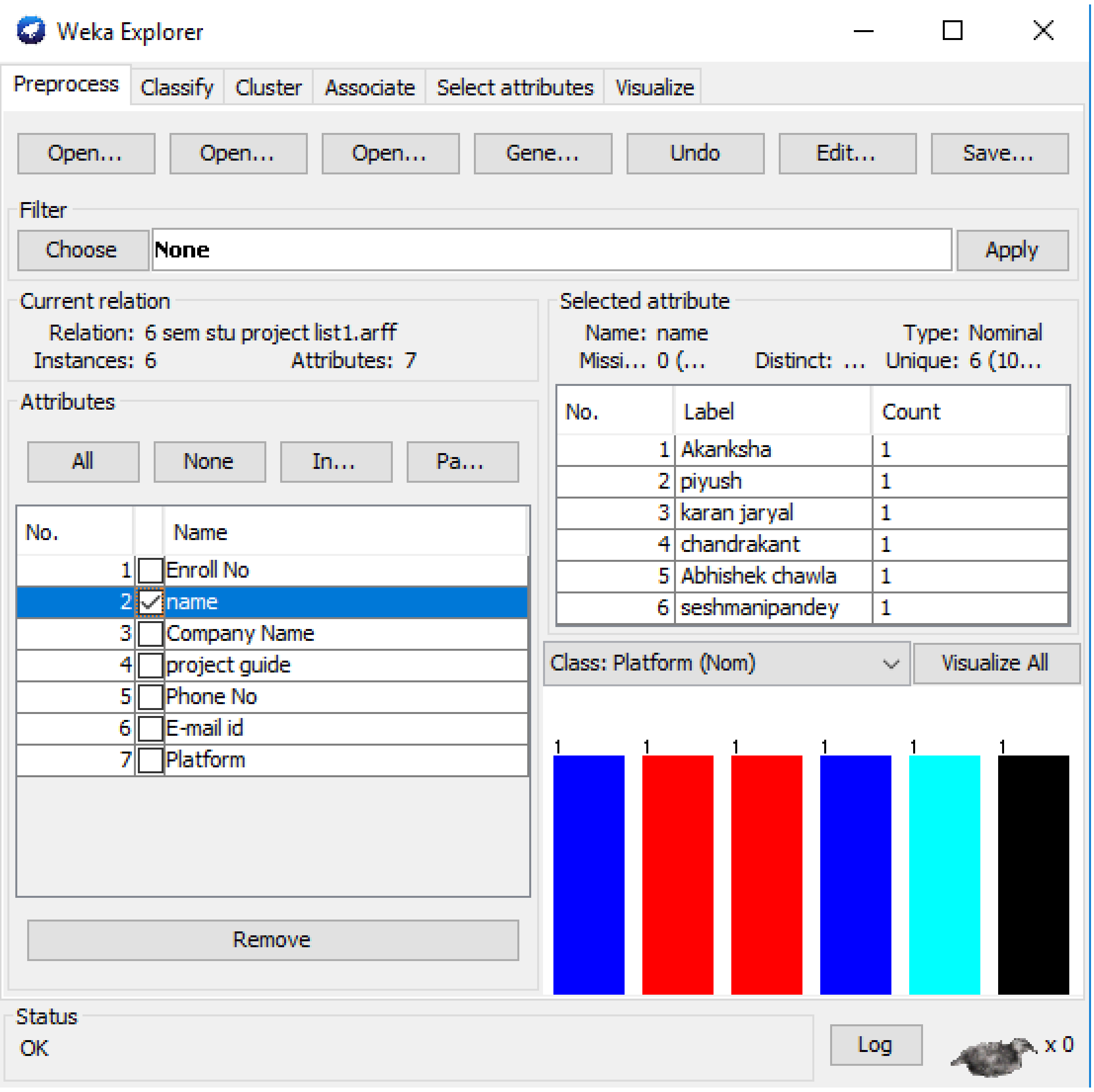
No.		Name
1	□	Enroll No
2	□	name
3	□	Company Name
4	□	project guide
5	□	Phone No
6	□	E-mail id
7	□	Platform

Status
OK

Click on visualize all



Attribute selection:



Filter: None

Current relation
Relation: 6 sem stu project list1.arff
Instances: 6 Attributes: 7

Attributes

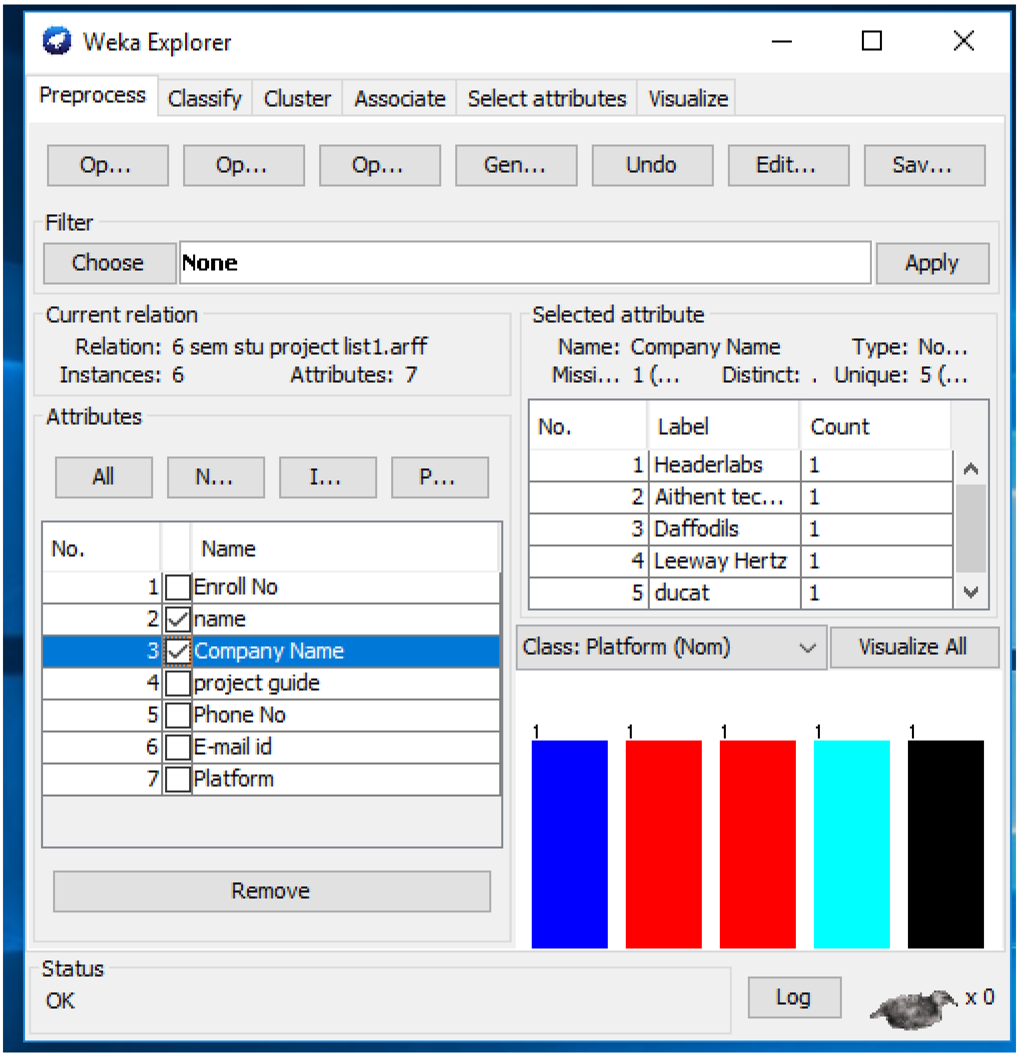
No.		Name
1	□	Enroll No
2	☑	name
3	□	Company Name
4	□	project guide
5	□	Phone No
6	□	E-mail id
7	□	Platform

Selected attribute
Name: name Type: Nominal
Missi... 0 (... Distinct: ... Unique: 6 (10...

No.	Label	Count
1	Akanksha	1
2	piyush	1
3	karan jaryal	1
4	chandrakant	1
5	Abhishek chawla	1
6	seshmanipandey	1

Class: Platform (Nom)

Status
OK



Filter: None

Current relation
Relation: 6 sem stu project list1.arff
Instances: 6 Attributes: 7

Selected attribute
Name: Company Name Type: No...
Missi... 1 (... Distinct: . Unique: 5 (...

No.	Label	Count
1	Headerlabs	1
2	Aithent tec...	1
3	Daffodils	1
4	Leeway Hertz	1
5	ducat	1

Attributes

No.		Name
1	☐	Enroll No
2	☑	name
3	☑	Company Name
4	☐	project guide
5	☐	Phone No
6	☐	E-mail id
7	☐	Platform

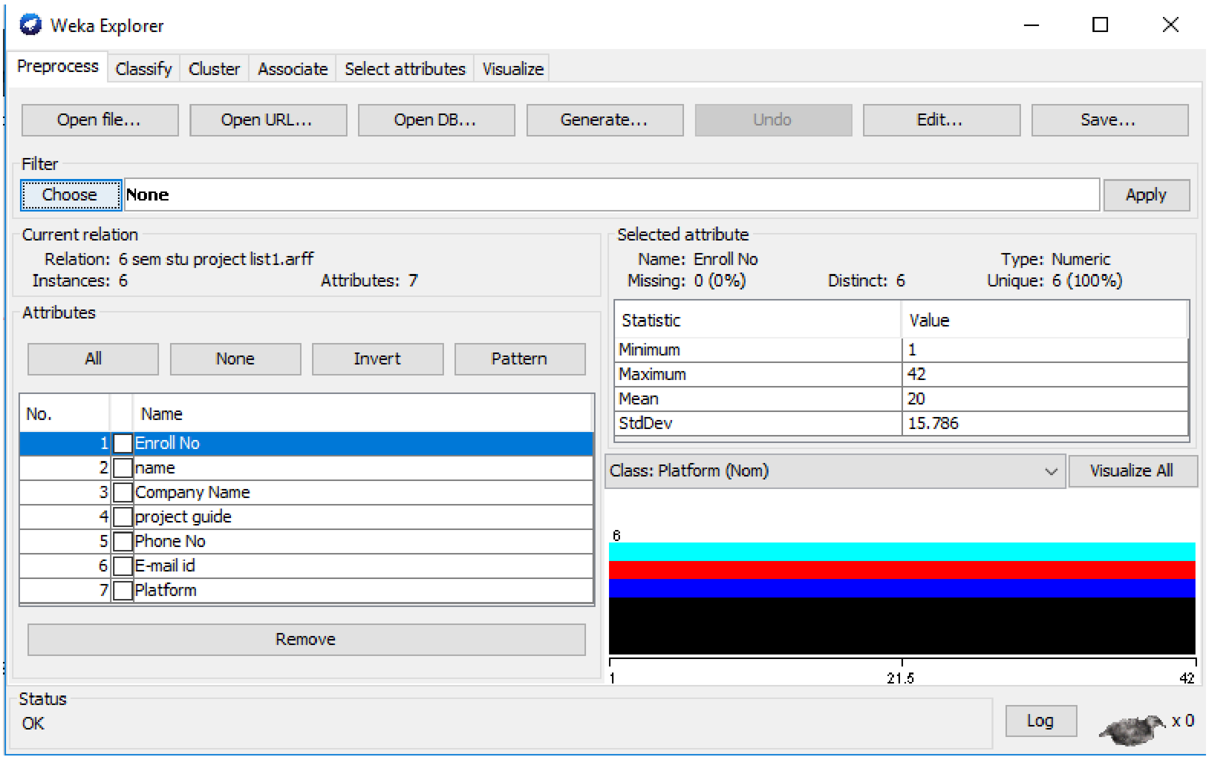
Class: Platform (Nom)

Status
OK

2.4 Setting Filters

Pre-processing tools in WEKA are called “filters”. WEKA contains filters for discretization, normalization, resampling, attribute selection, transformation and combination of attributes .Some techniques, such as association rule mining, can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes.

Using filters you can replace the discrete values to nominal values.



Filter: None

Current relation
Relation: 6 sem stu project list1.arff
Instances: 6 Attributes: 7

Attributes

No.		Name
1	□	Enroll No
2	□	name
3	□	Company Name
4	□	project guide
5	□	Phone No
6	□	E-mail id
7	□	Platform

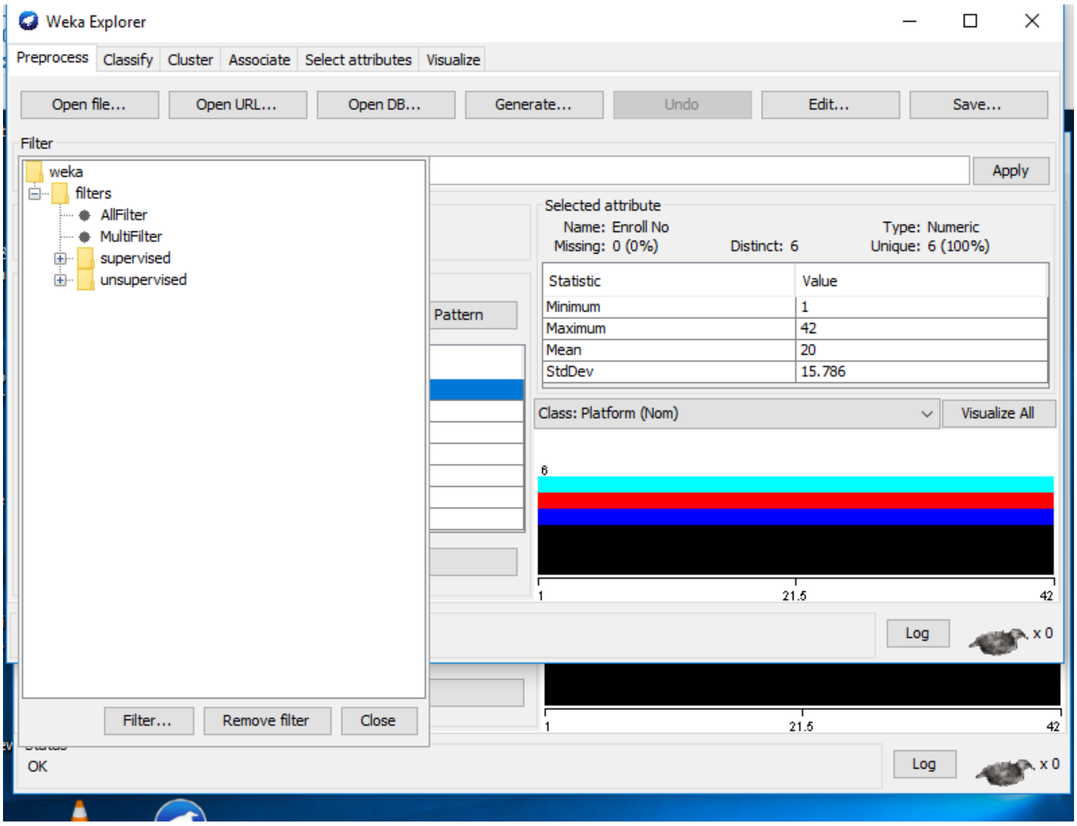
Selected attribute
Name: Enroll No Type: Numeric
Missing: 0 (0%) Distinct: 6 Unique: 6 (100%)

Statistic	Value
Minimum	1
Maximum	42
Mean	20
StdDev	15.786

Class: Platform (Nom)

Status
OK

This will show pull-down menu with a list of available filters. Select Supervised Æ Attribute Æ Discretize and click on ‘Apply’ button. The filter will convert Numeric values into Nominal



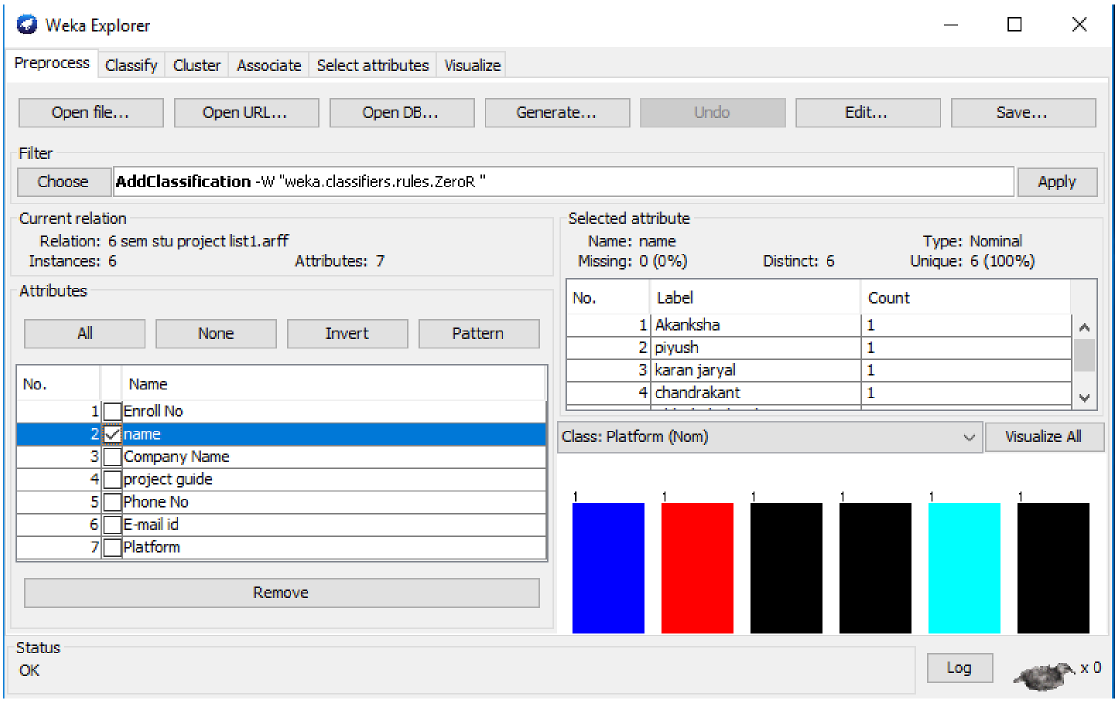
Selected attribute
Name: Enroll No Type: Numeric
Missing: 0 (0%) Distinct: 6 Unique: 6 (100%)

Statistic	Value
Minimum	1
Maximum	42
Mean	20
StdDev	15.786

Class: Platform (Nom)

OK

When filter is chosen, the fields in the window changes to reflect available options.



Filter

AddClassification -W "weka.classifiers.rules.ZeroR "

Current relation

Relation: 6 sem stu project list1.arff
Instances: 6 Attributes: 7

Selected attribute

Name: name Type: Nominal
Missing: 0 (0%) Distinct: 6 Unique: 6 (100%)

No.	Label	Count
1	Akanksha	1
2	piyush	1
3	karan jaryal	1
4	chandrakant	1

Attributes

No.		Name
1	□	Enroll No
2	☑	name
3	□	Company Name
4	□	project guide
5	□	Phone No
6	□	E-mail id
7	□	Platform

Class: Platform (Nom)

Status
OK

As you can see, there is no change in the value Outlook. Select value Temperature, look at the ‘Selected attribute’ box, the ‘Type’ field shows that the attribute type has changed from Numeric to Nominal. The list has changed as well: instead of statistical values there is count of instances.

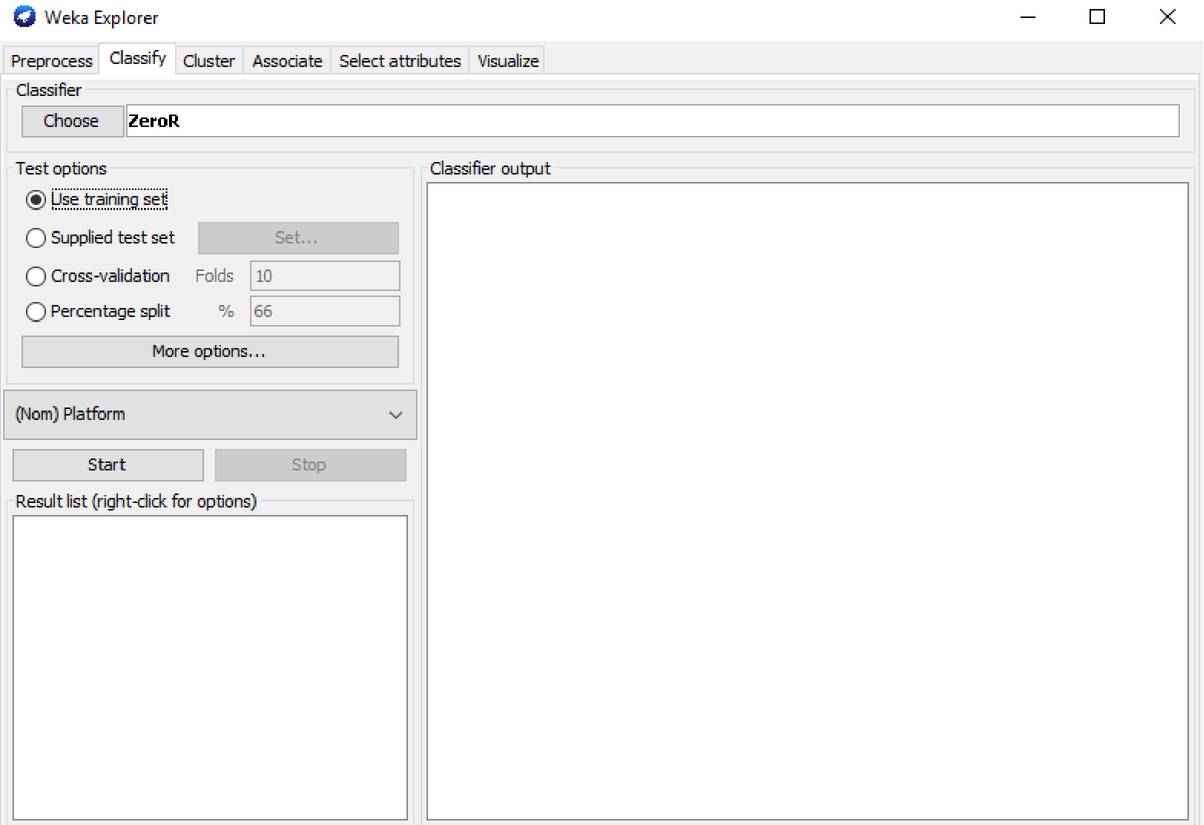
CHAPTER 3: CLASSIFIERS

3.1 Building “Classifiers” :

Classifiers in WEKA are the models for predicting nominal or numeric quantities. The learning schemes available in WEKA include decision trees and lists, instance-based classifiers, support vector machines, multi-layer perceptrons, logistic regression, and bayes’ nets. “Meta”classifiers include bagging, boosting, stacking, error-correcting output codes, and locally weighted learning.

Once you have your data set loaded, all the tabs are available to you. Click on the ‘Classify’ tab.

‘Classify’ window comes up on the screen.

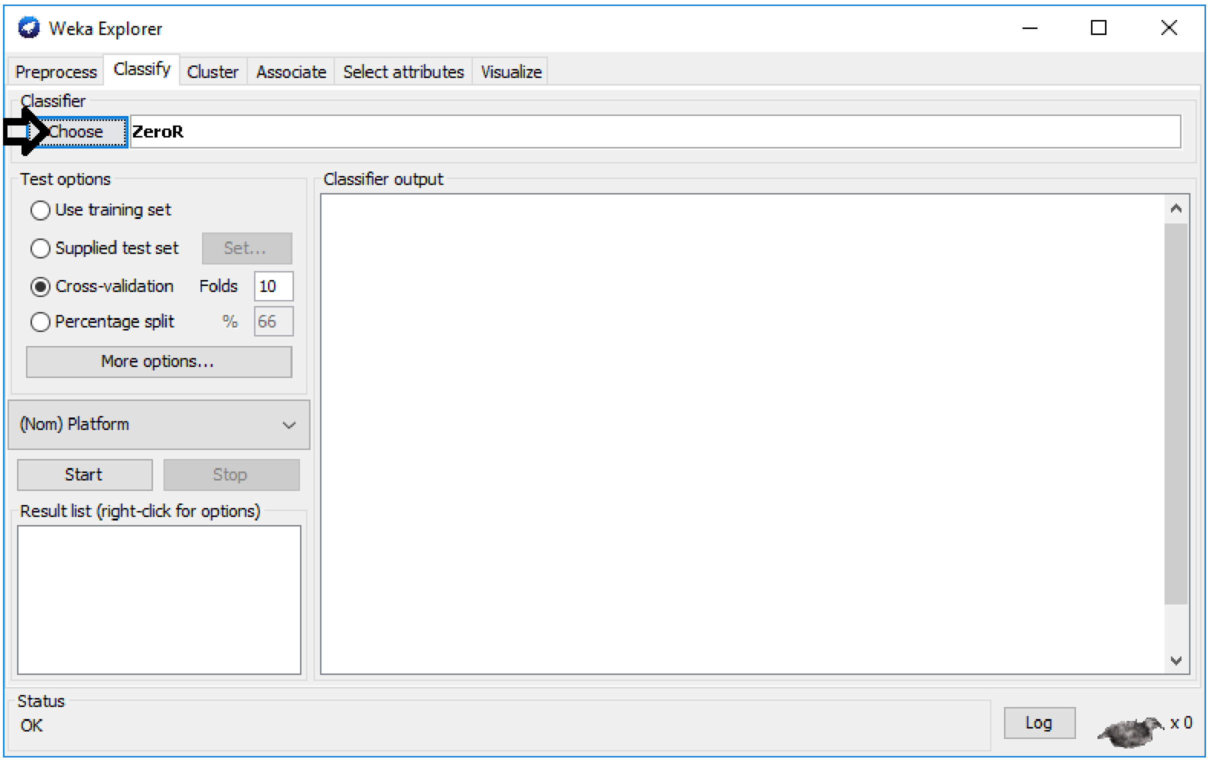


Classifier: ZeroR

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

(Nom) Platform

Now you can start analyzing the data using the provided algorithms. In this exercise you will analyze the data.



3.2 Setting Test Options:

Before you run the classification algorithm, you need to set test options. Set test options in the 'Test options' box. The test options that available are:

1. Use training set. Evaluates the classifier on how well it predicts the class of the instances it was trained on.

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

3. Cross-validation. Evaluates the classifier by cross-validation, using the number of folds that are entered in the 'Folds' text field.

4. Percentage split. Evaluates the classifier 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.

In the 'Classifier evaluation options' make sure that the following options are checked

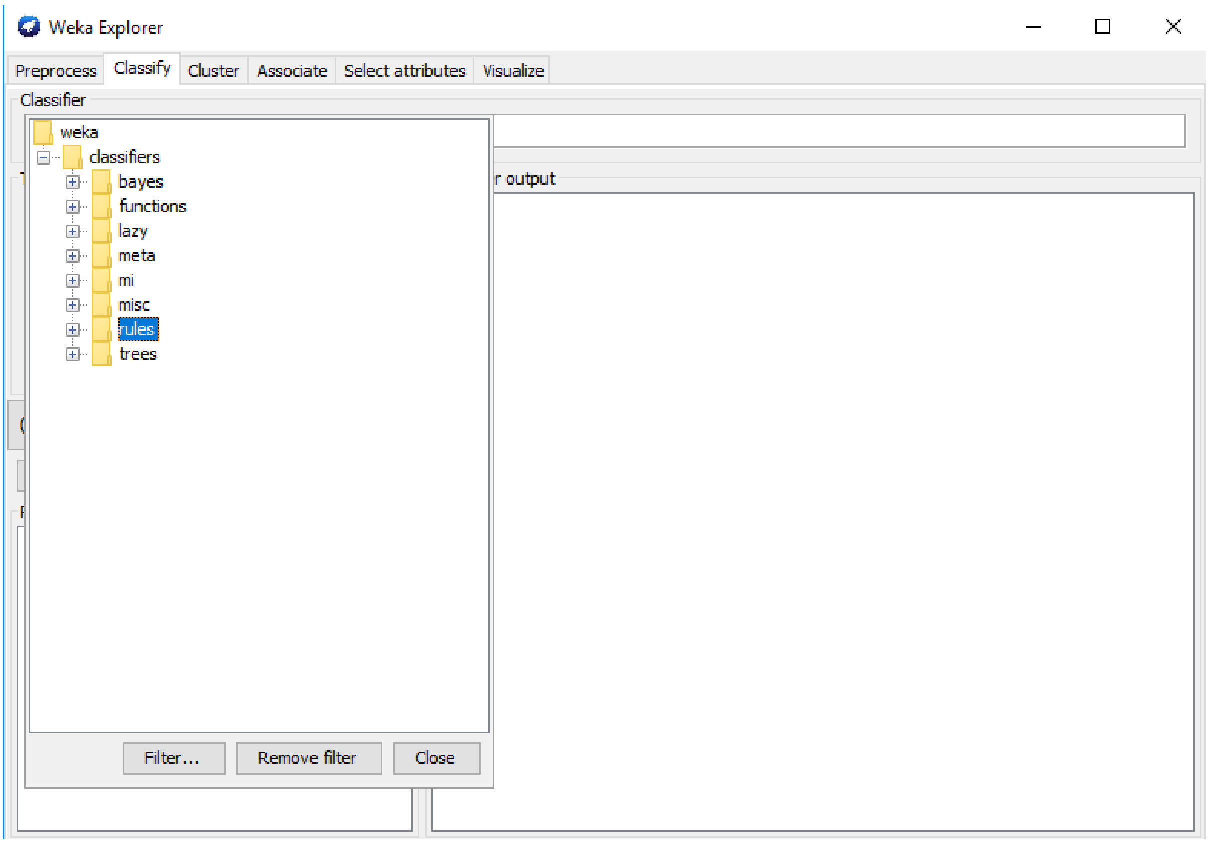
1. **Output model.** The output is the classification model on the full training set, so that it can be viewed, visualized, etc.

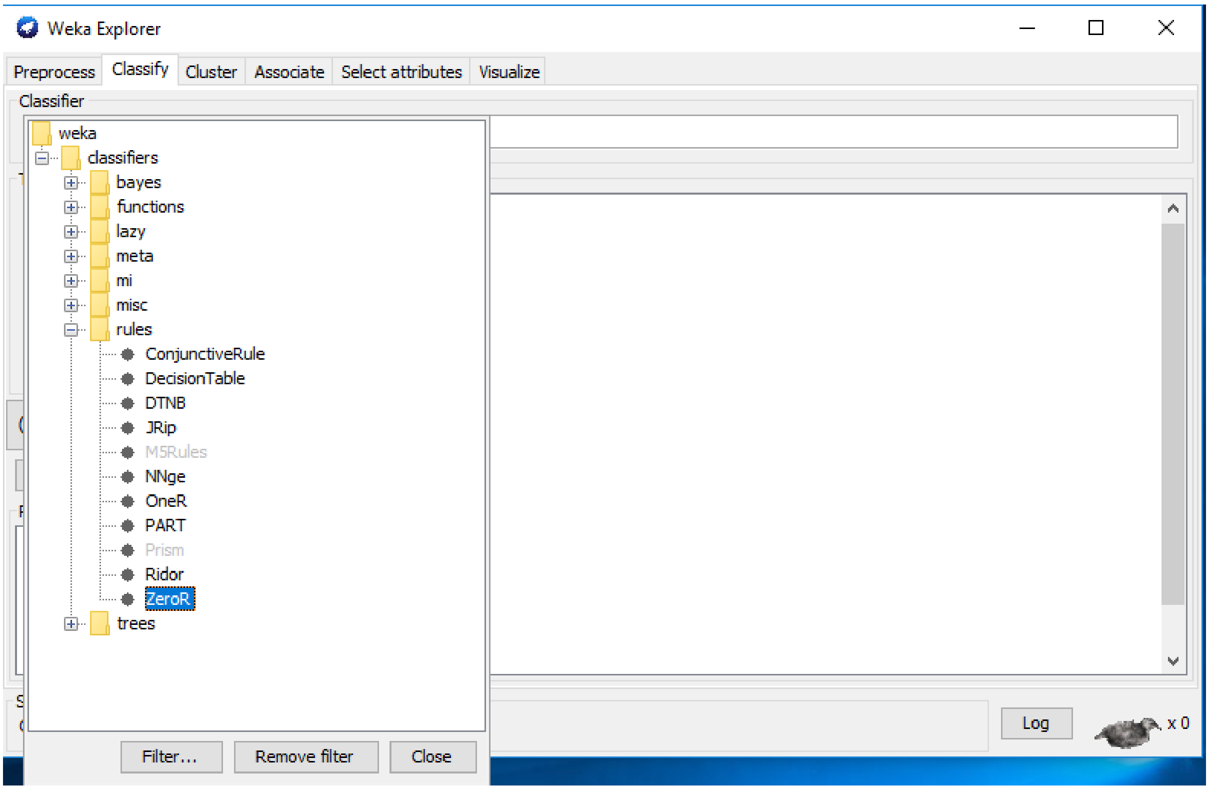
2. **Output per-class stats**. The precision/recall and true/false statistics for each class output.

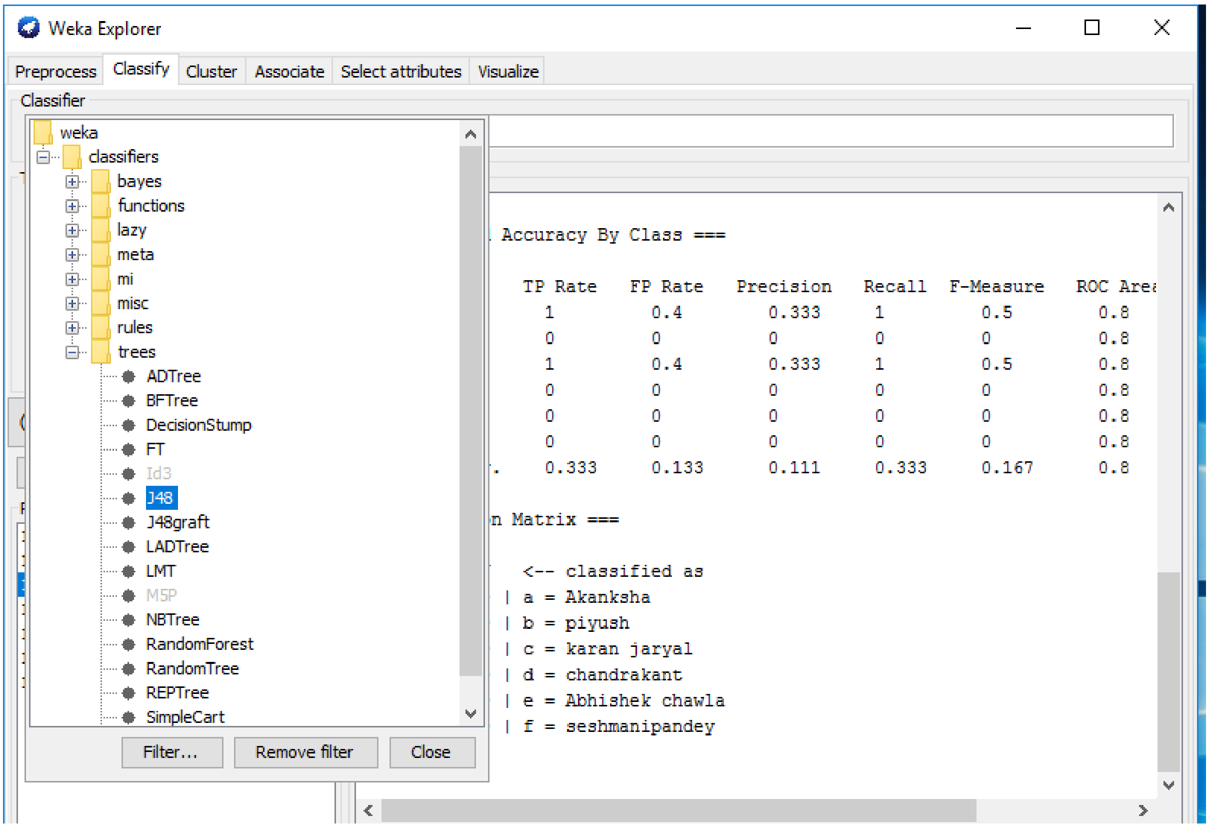
3. **Output confusion matrix.** The confusion matrix of the classifier's predictions is included in the output.

4. **Store predictions for visualization.** The classifier's predictions are remembered so that they can be visualized.

5. **Set 'Random seed for Xval / % Split' to 1.** This specifies the random seed used when randomizing the data before it is divided up for evaluation purposes



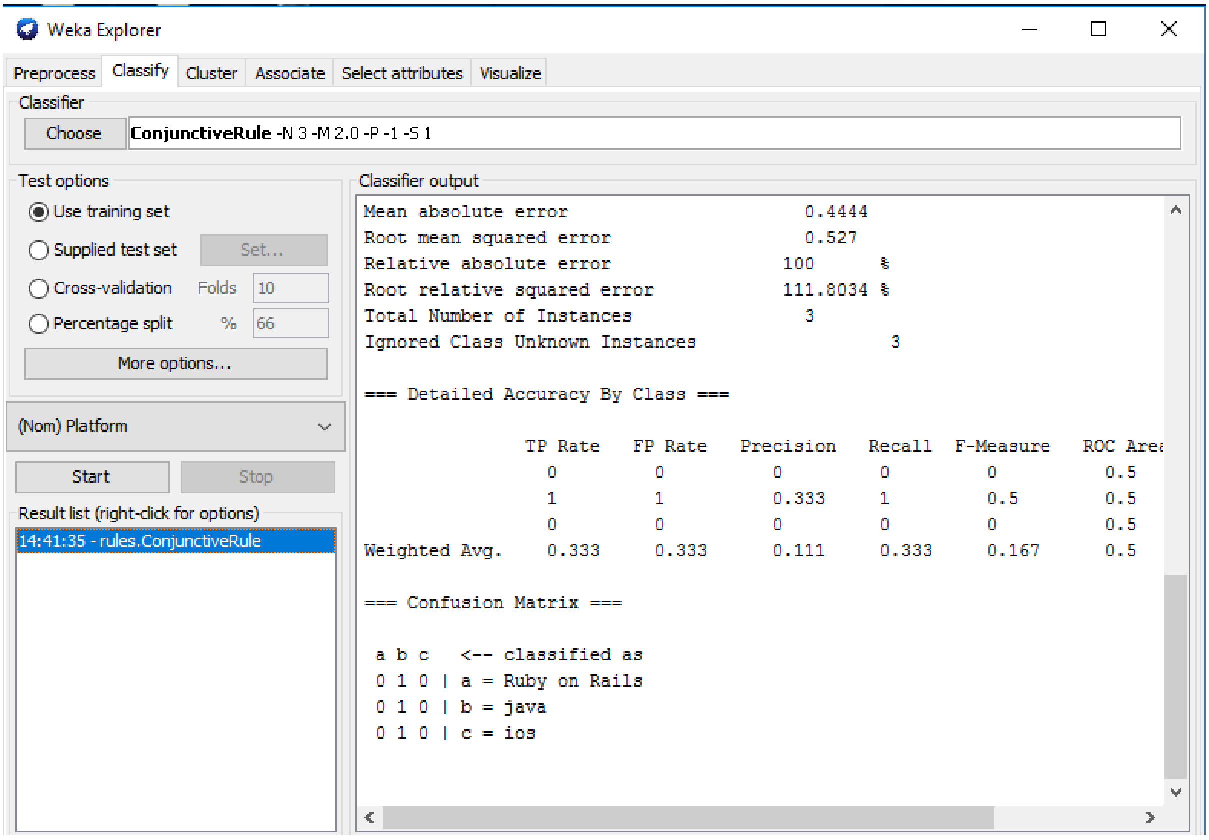




```
 Accuracy By Class ===

   TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Are
     1         0.4       0.333      1         0.5        0.8
     0         0         0          0         0          0.8
     1         0.4       0.333      1         0.5        0.8
     0         0         0          0         0          0.8
     0         0         0          0         0          0.8
     0         0         0          0         0          0.8
.    0.333     0.133     0.111      0.333     0.167      0.8

n Matrix ===

   <-- classified as
 | a = Akanksha
 | b = piyush
 | c = karan jaryal
 | d = chandrakant
 | e = Abhishek chawla
 | f = seshmanipandey
```

Once the options have been specified, you can run the classification algorithm. Click on ‘Start’ button to start the learning process. You can stop learning process at any time by clicking on ‘Stop’ button

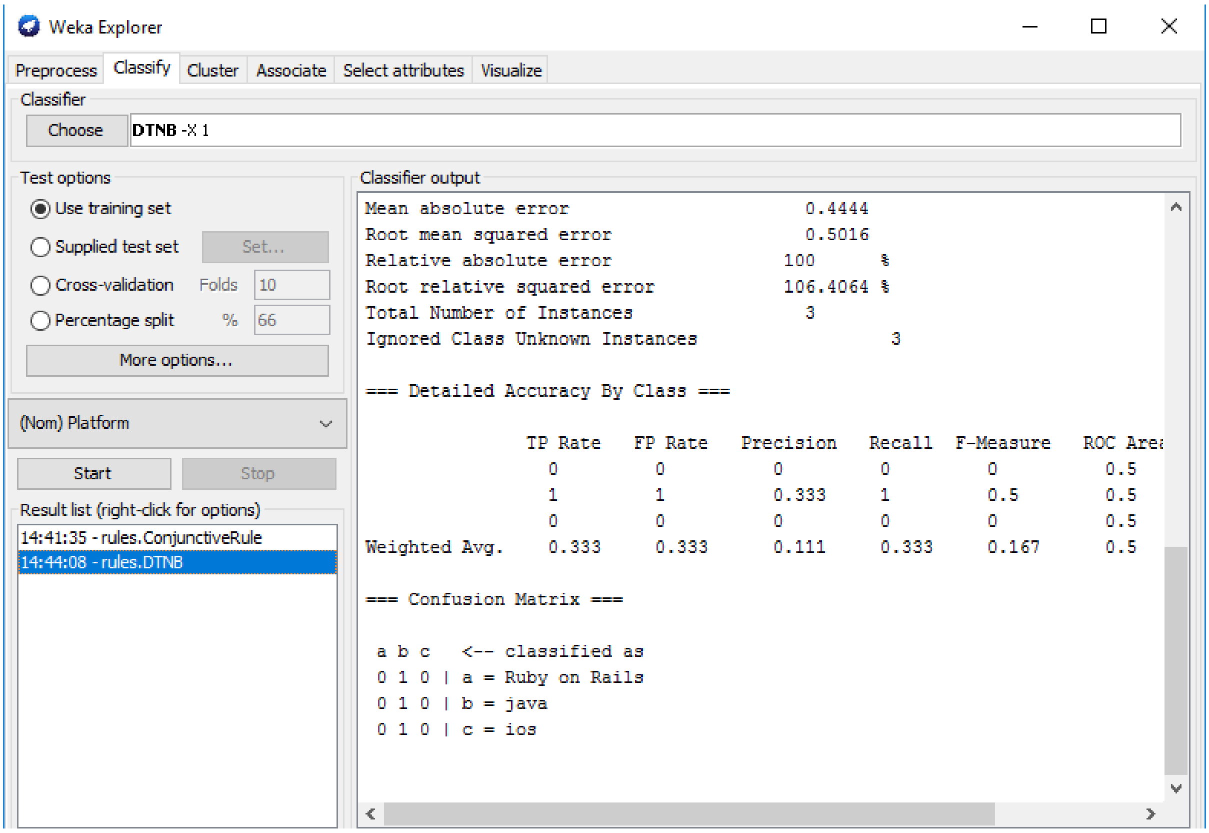
When training set is complete, the ‘Classifier’ output area on the right panel of ‘Classify’ window is filled with text describing the results of training and testing. A new entry appears in the ‘Result list’ box on the left panel of ‘Classify’ window.



ConjunctiveRule -N 3 -M 2.0 -P -1 -S 1

14:41:35 - rules.ConjunctiveRule

```
Mean absolute error                      0.4444
Root mean squared error                  0.527
Relative absolute error                100      %
Root relative squared error            111.8034 %
Total Number of Instances                3
Ignored Class Unknown Instances                  3

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area
                 0        0        0         0        0          0.5
                 1        1        0.333     1        0.5        0.5
                 0        0        0         0        0          0.5
Weighted Avg.    0.333    0.333    0.111     0.333    0.167      0.5

=== Confusion Matrix ===

 a b c   <-- classified as
 0 1 0 | a = Ruby on Rails
 0 1 0 | b = java
 0 1 0 | c = ios
```



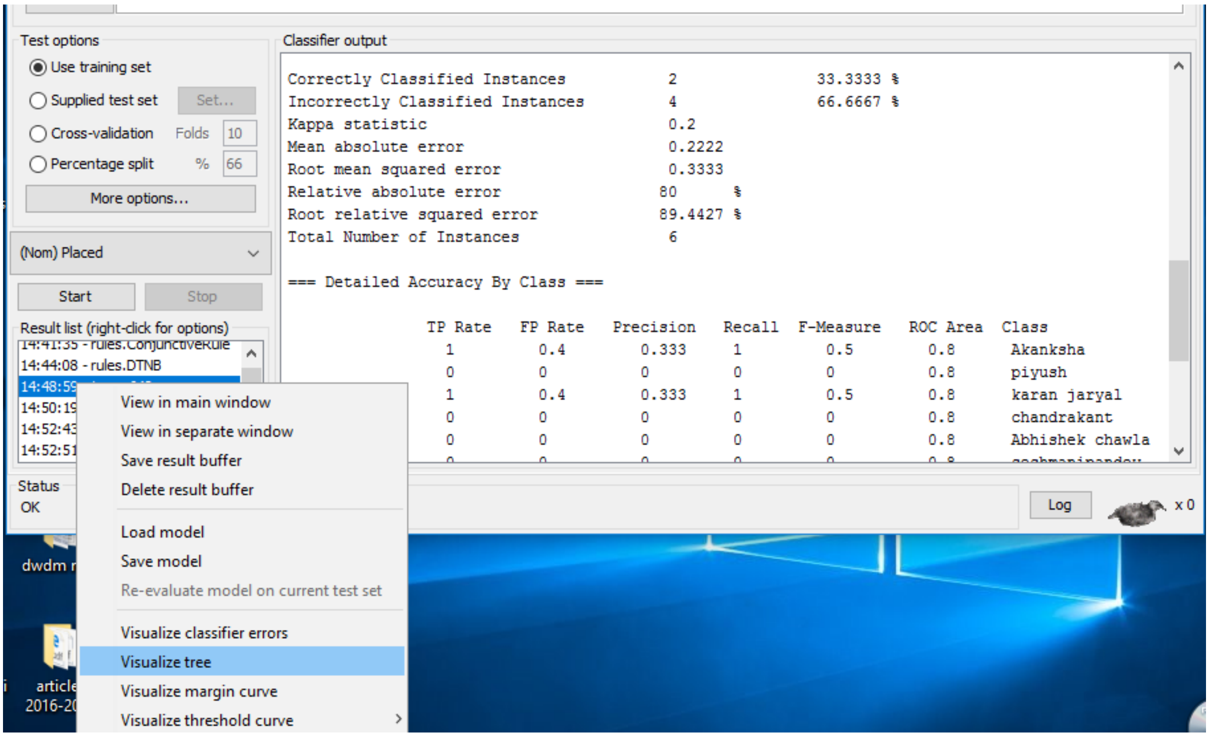
Classifier: DTNB -X 1

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

(Nom) Platform

Result list (right-click for options)
14:41:35 - rules.ConjunctiveRule
14:44:08 - rules.DTNB

```
Mean absolute error                      0.4444
Root mean squared error                  0.5016
Relative absolute error                100      %
Root relative squared error            106.4064 %
Total Number of Instances                3
Ignored Class Unknown Instances                  3

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area
                 0         0         0         0         0          0.5
                 1         1         0.333     1         0.5        0.5
                 0         0         0         0         0          0.5
Weighted Avg.    0.333     0.333     0.111     0.333     0.167      0.5

=== Confusion Matrix ===

 a b c   <-- classified as
 0 1 0 | a = Ruby on Rails
 0 1 0 | b = java
 0 1 0 | c = ios
```

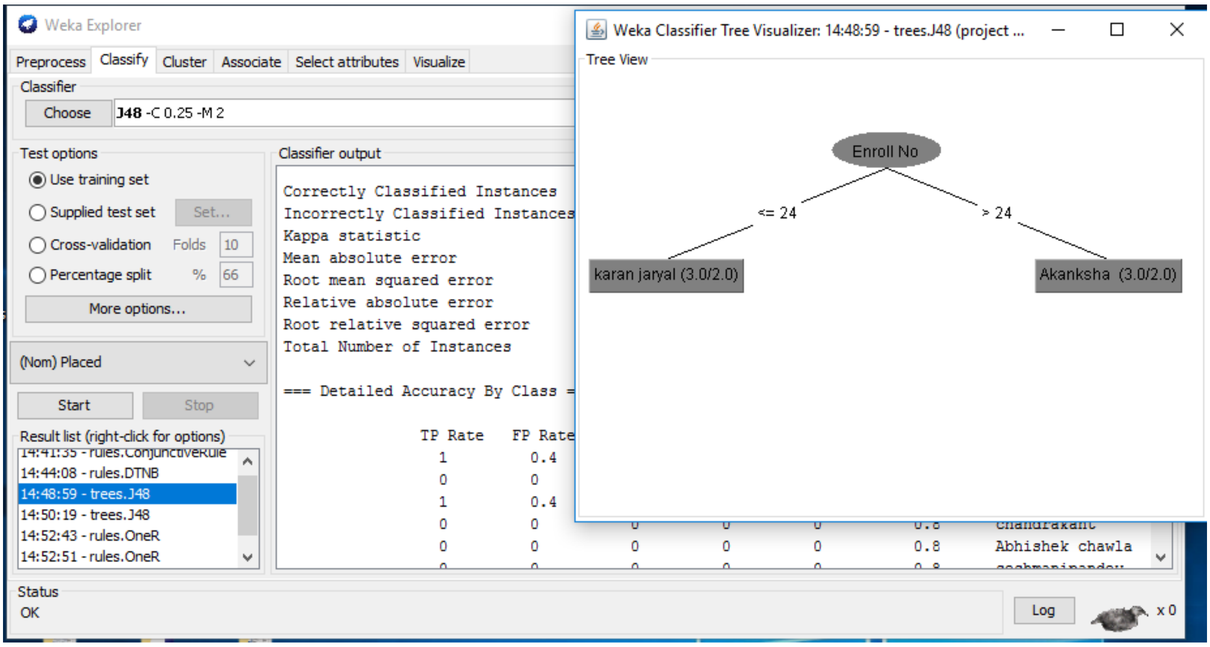


Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

(Nom) Placed

Result list (right-click for options)
14:41:35 - rules.ConjunctiveRule
14:44:08 - rules.DTNB
14:48:59
14:50:19
14:52:43
14:52:51

```
Correctly Classified Instances           2               33.3333 %
Incorrectly Classified Instances         4               66.6667 %
Kappa statistic                          0.2
Mean absolute error                      0.2222
Root mean squared error                  0.3333
Relative absolute error                 80      %
Root relative squared error             89.4427 %
Total Number of Instances                6

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.4       0.333     1         0.5        0.8       Akanksha
                 0         0         0         0         0          0.8       piyush
                 1         0.4       0.333     1         0.5        0.8       karan jaryal
                 0         0         0         0         0          0.8       chandrakant
                 0         0         0         0         0          0.8       Abhishek chawla
```

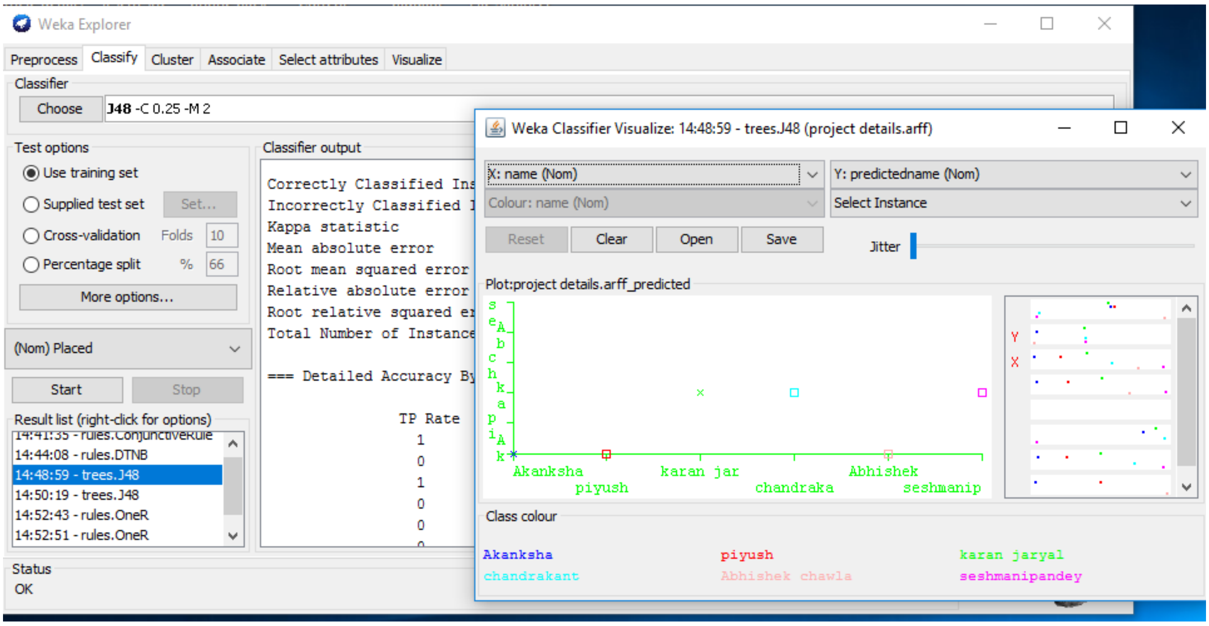


J48 -C 0.25 -M 2

```
14:41:35 - rules.ConjunctiveRule
14:44:08 - rules.DTNB
14:48:59 - trees.J48
14:50:19 - trees.J48
14:52:43 - rules.OneR
14:52:51 - rules.OneR
```

Classifier output

```
Correctly Classified Instances
Incorrectly Classified Instances
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared error
Total Number of Instances

=== Detailed Accuracy By Class =

              TP Rate   FP Rate
                1         0.4
                0         0
                1         0.4
                0         0                                      0.8       chandrakant
                0         0         0        0        0          0.8       Abhishek chawla
                0         0         0        0        0          0.8       seshmanipandey
```



J48 -C 0.25 -M 2

```
14:41:35 - rules.ConjunctiveRule
14:44:08 - rules.DTNB
14:48:59 - trees.J48
14:50:19 - trees.J48
14:52:43 - rules.OneR
14:52:51 - rules.OneR
```

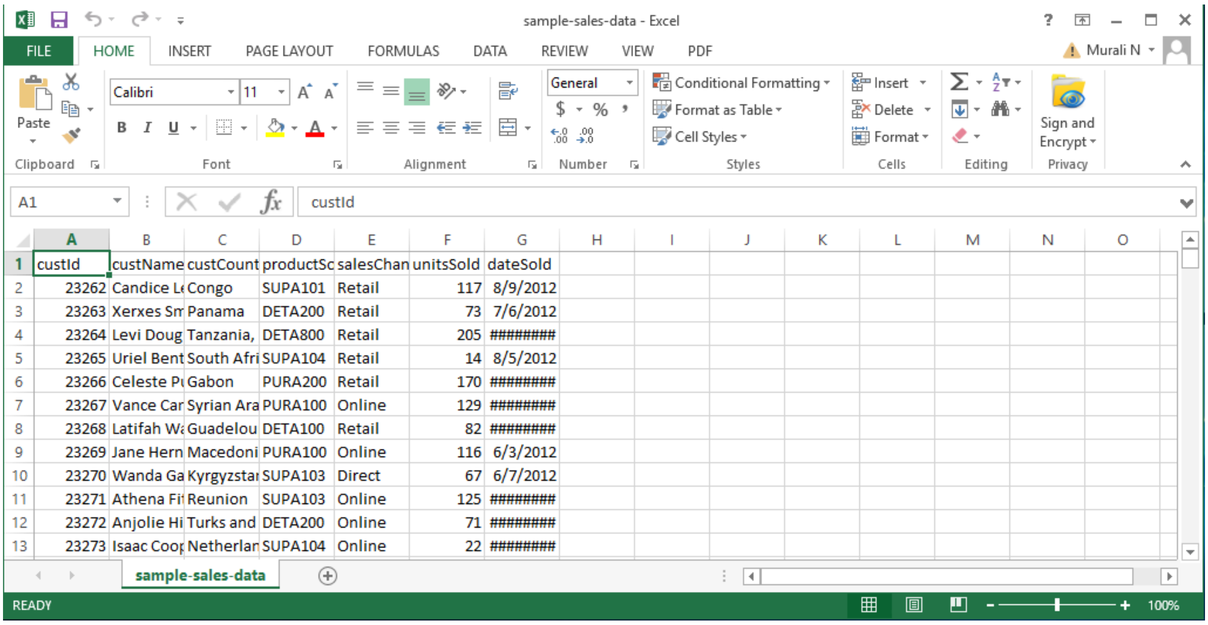
Classifier output

```
Correctly Classified Ins
Incorrectly Classified I
Kappa statistic
Mean absolute error
Root mean squared error
Relative absolute error
Root relative squared er
Total Number of Instance

=== Detailed Accuracy By

              TP Rate
                1
                0
                1
                0
                0
                0
```

CHAPTER 4: CLUSTERING

4.1 Clustering Data:

WEKA contains “clusterers” for finding groups of similar instances in a dataset. The clustering schemes available in WEKA are k-Means, EM, Cobweb, X-means, Farthest First. Clusters can be visualized and compared to “true” clusters (if given). Evaluation is based on log likelihood if clustering scheme produces a probability distribution.



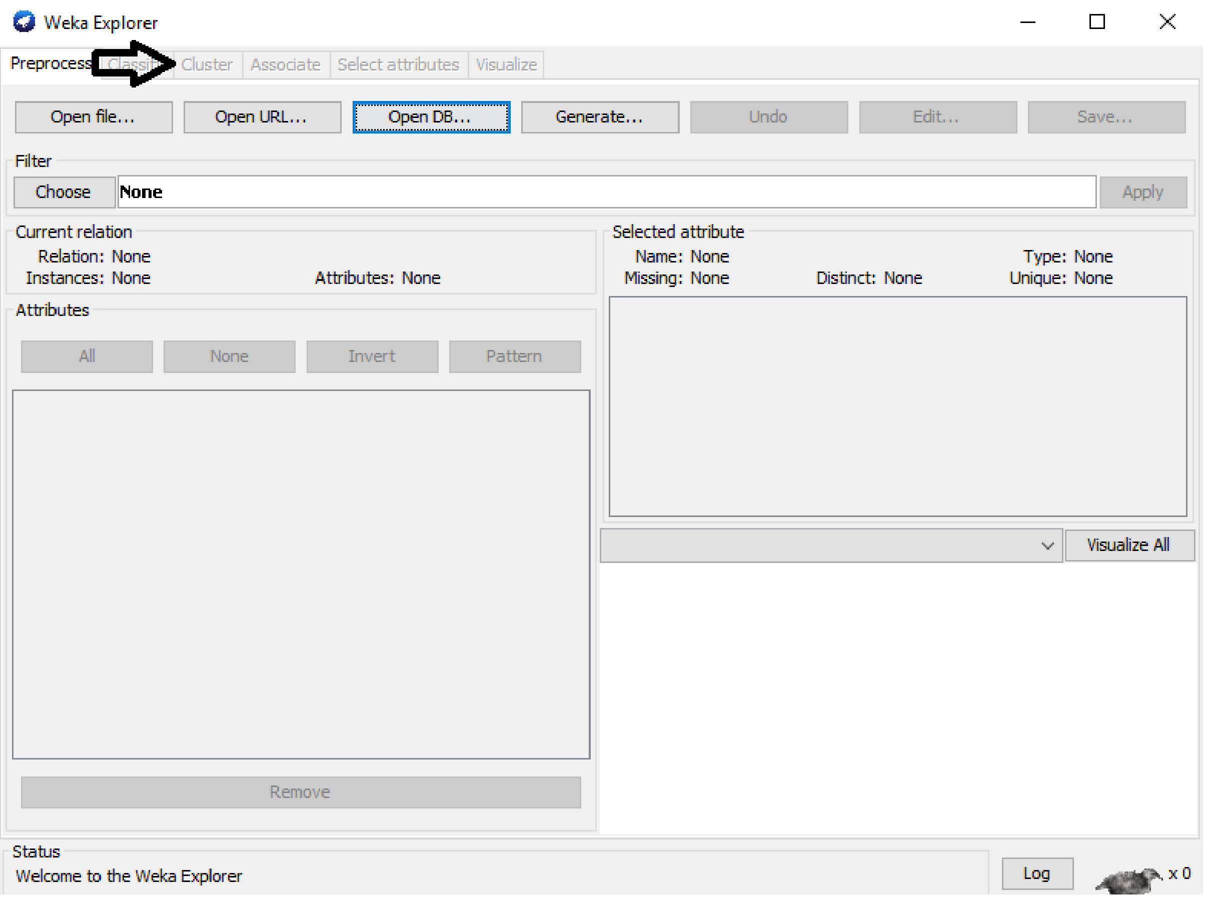
custId

	A	B	C	D	E	F	G
1	custId	custName	custCount	productSc	salesChan	unitsSold	dateSold
2	23262	Candice Lı	Congo	SUPA101	Retail	117	8/9/2012
3	23263	Xerxes Sm	Panama	DETA200	Retail	73	7/6/2012
4	23264	Levi Doug	Tanzania,	DETA800	Retail	205	#########
5	23265	Uriel Bent	South Afri	SUPA104	Retail	14	8/5/2012
6	23266	Celeste Pı	Gabon	PURA200	Retail	170	#########
7	23267	Vance Car	Syrian Ara	PURA100	Online	129	#########
8	23268	Latifah Wa	Guadelou	DETA100	Retail	82	#########
9	23269	Jane Hern	Macedoni	PURA100	Online	116	6/3/2012
10	23270	Wanda Ga	Kyrgyzstaı	SUPA103	Direct	67	6/7/2012
11	23271	Athena Fi	Reunion	SUPA103	Online	125	#########
12	23272	Anjolie Hi	Turks and	DETA200	Online	71	#########
13	23273	Isaac Coo	Netherlan	SUPA104	Online	22	#########

An international online catalog company wishes to group its customers based on common features. Company management does not have any predefined labels for these groups. Based on the outcome of the grouping, they will target marketing and advertising campaigns to the different groups. The information they have about the customers includes customer ID ,Customer Name , customer count,ProductSold, Sales Channel,Units Sold,Date Sold.

For our exercise we will use a part of the database for customers in US. Depending on the type of products sold , not all attributes are important. For example, suppose the to know the det

In ‘Preprocess’ window click on ‘Open file…’ button and select “customers.csv” file. Click ‘Cluster’ tab at the top of WEKA Explorer window.

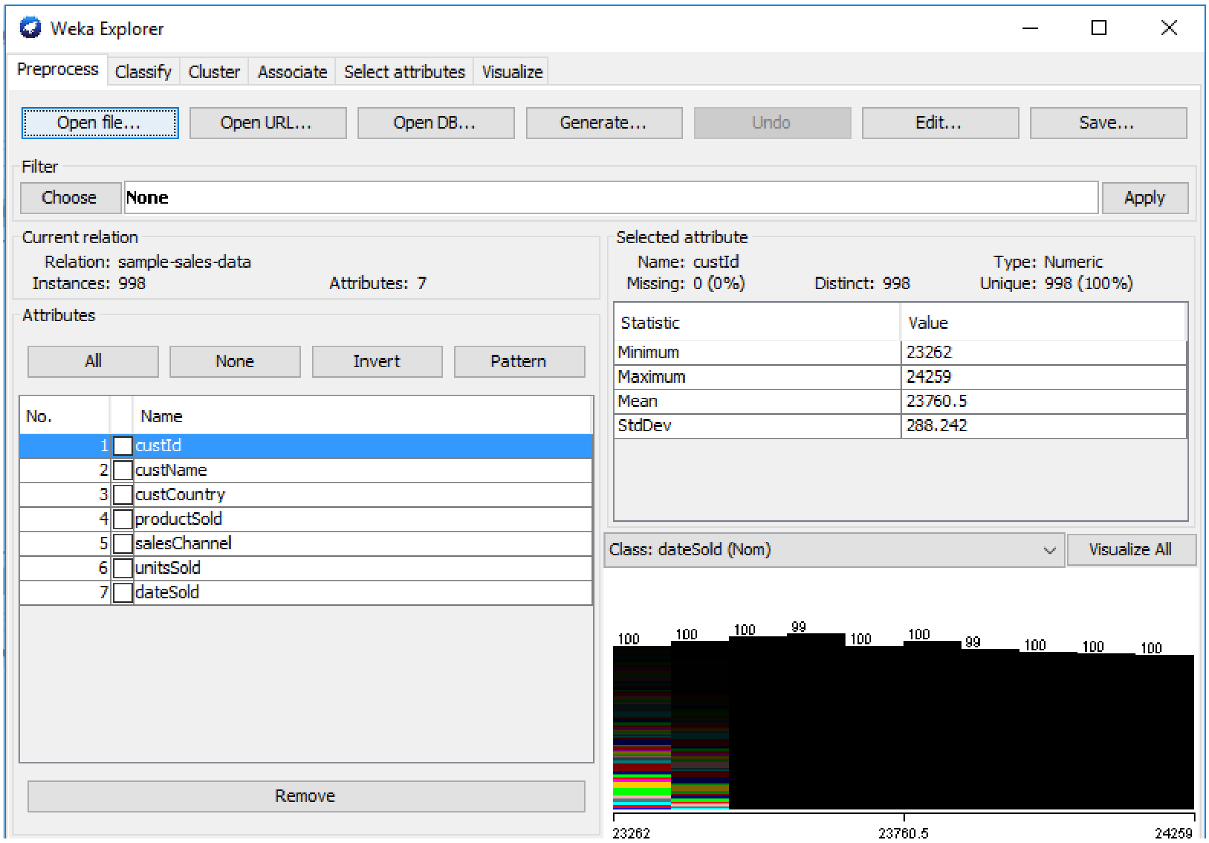


Filter: None

Current relation
Relation: None
Instances: None Attributes: None

Selected attribute
Name: None Type: None
Missing: None Distinct: None Unique: None

Status
Welcome to the Weka Explorer



Filter: None

Current relation
Relation: sample-sales-data
Instances: 998 Attributes: 7

Attributes

No.		Name
1	□	custId
2	□	custName
3	□	custCountry
4	□	productSold
5	□	salesChannel
6	□	unitsSold
7	□	dateSold

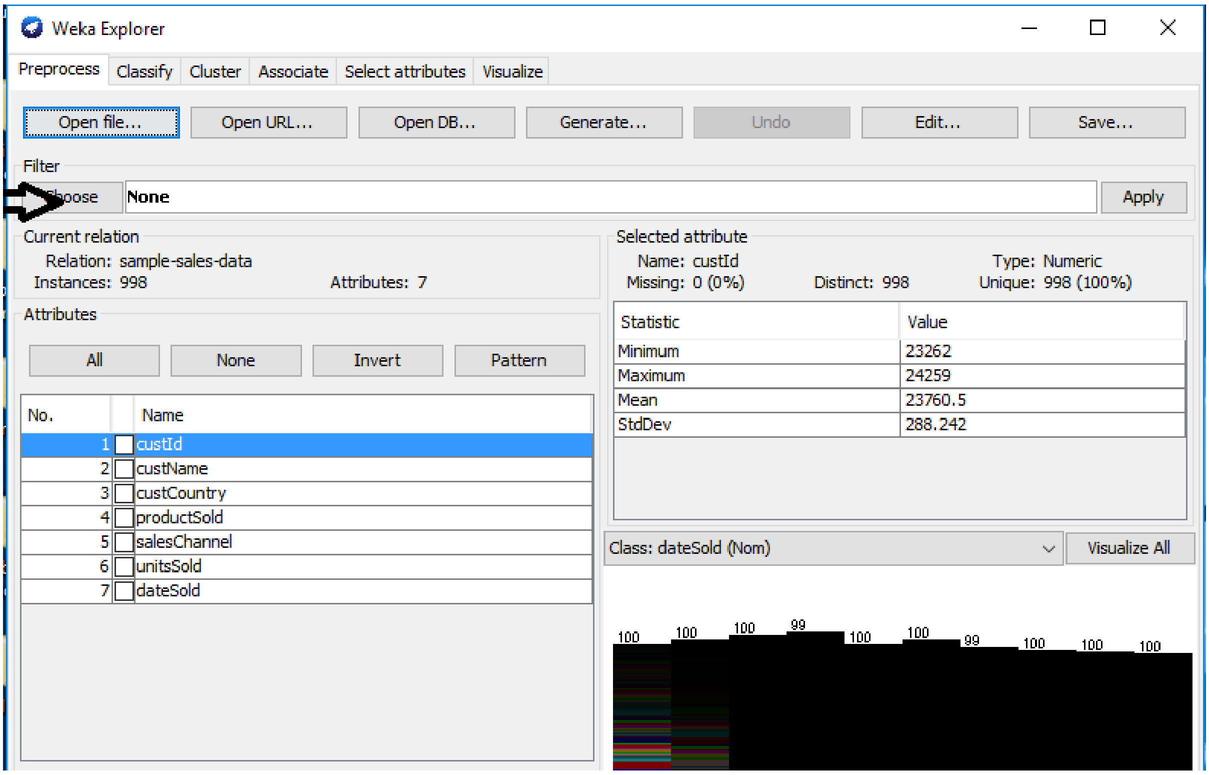
Selected attribute
Name: custId Type: Numeric
Missing: 0 (0%) Distinct: 998 Unique: 998 (100%)

Statistic	Value
Minimum	23262
Maximum	24259
Mean	23760.5
StdDev	288.242

Class: dateSold (Nom)

4.2 Choosing Clustering Scheme:

In the ‘Clusterer’ box click on ‘Choose’ button. In pull-down menu select WEKA Æ Clusterers, and select the cluster scheme ‘SimpleKMeans’. Some implementations of K-means only allow numerical values for attributes; therefore, we do not need to use a filter.



Filter: None

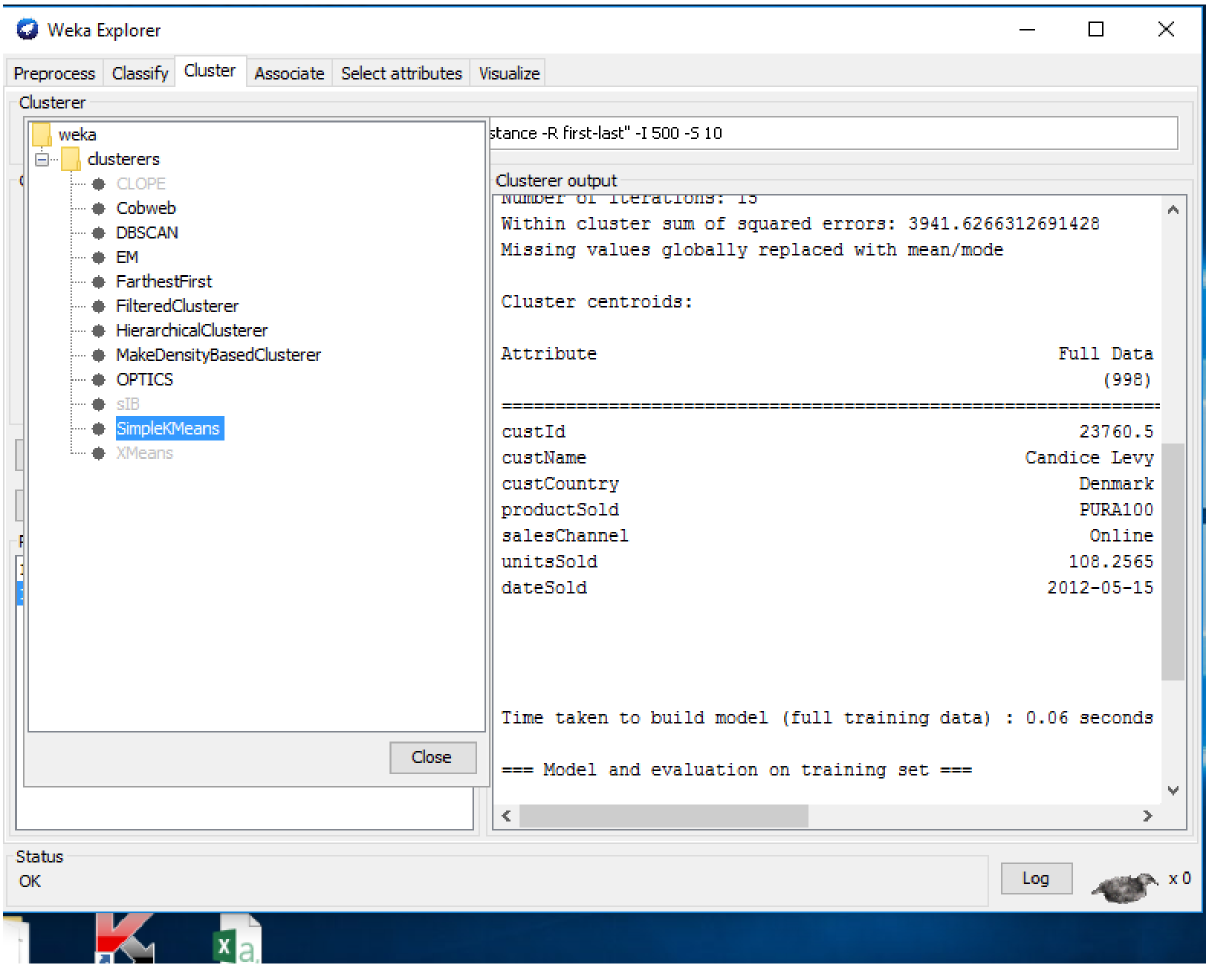
Relation: sample-sales-data
Instances: 998
Attributes: 7

No.		Name
1	□	custId
2	□	custName
3	□	custCountry
4	□	productSold
5	□	salesChannel
6	□	unitsSold
7	□	dateSold

Name: custId
Type: Numeric
Missing: 0 (0%)
Distinct: 998
Unique: 998 (100%)

Statistic	Value
Minimum	23262
Maximum	24259
Mean	23760.5
StdDev	288.242

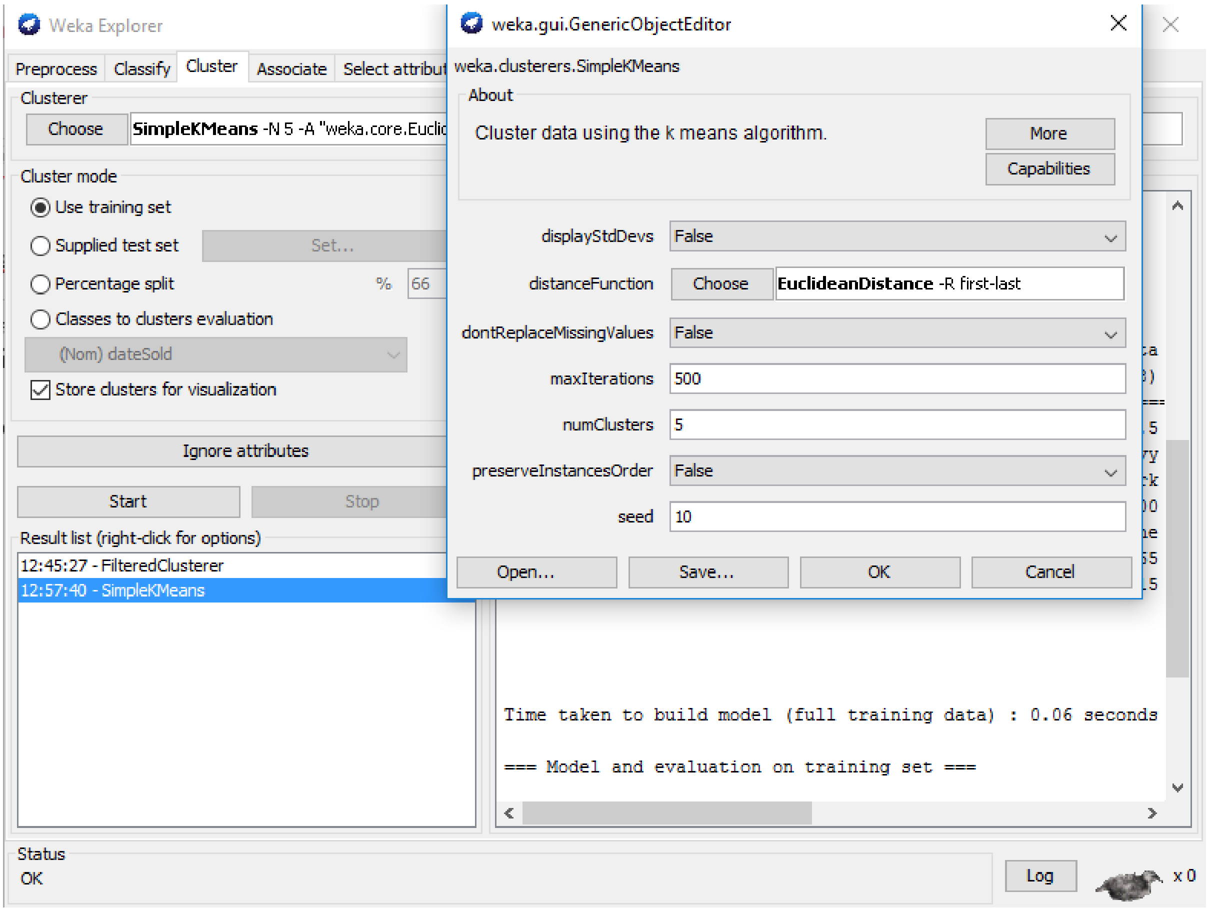
Class: dateSold (Nom)



```
stance -R first-last" -I 500 -S 10
```

```
Number of iterations: 15
Within cluster sum of squared errors: 3941.626631269142 8
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute                                           Full Data
                                                        (998)
=============================================================
custId                                                23760.5
custName                                         Candice Levy
custCountry                                           Denmark
productSold                                           PURA100
salesChannel                                           Online
unitsSold                                            108.2565
dateSold                                           2012-05-15



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

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

Once the clustering algorithm is chosen, right-click on the algorithm, “weak.gui.GenericObjectEditor” comes up to the screen. Set the value in “numClusters” box to 5 (instead of default 2) because you have five clusters in your .arff file. Leave the value of ‘seed’ as is. The seed value is used in generating a random number, which is used for making the initial assignment of instances to clusters. Note that, in general, K-means is quite sensitive to how clusters are initially assigned. Thus, it is often necessary to try different values and evaluate the results.



Clusterer: SimpleKMeans -N 5 -A "weka.core.Euclic

Cluster mode

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

Result list (right-click for options)
12:45:27 - FilteredClusterer
12:57:40 - SimpleKMeans

weka.clusterers.SimpleKMeans

About: Cluster data using the k means algorithm.

displayStdDevs: False
distanceFunction: EuclideanDistance -R first-last
dontReplaceMissingValues: False
maxIterations: 500
numClusters: 5
preserveInstancesOrder: False
seed: 10

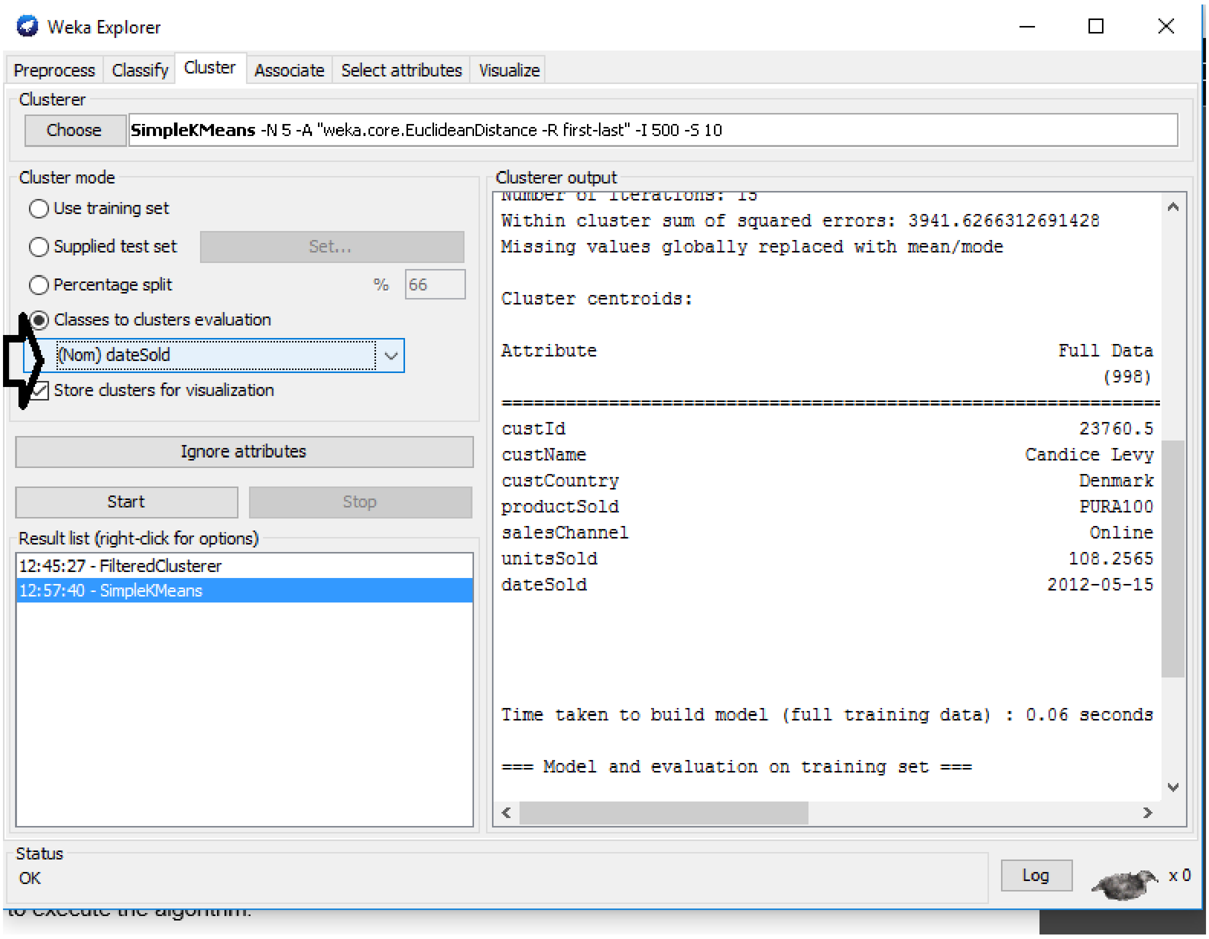
```
Time taken to build model (full training data) : 0.06 seconds

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

Status
OK

4.3 Setting Test Options:

Before you run the clustering algorithm, you need to choose ‘Cluster mode’. Click on ‘Classes to cluster evaluation’ radio-button in ‘Cluster mode’ box and select in the pull-down box below.

Once the options have been specified, you can run the clustering algorithm. Click on the ‘Start’ button to execute the algorithm.



Clusterer: SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode
- ○ Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ● Classes to clusters evaluation
- (Nom) dateSold
- ☑ Store clusters for visualization

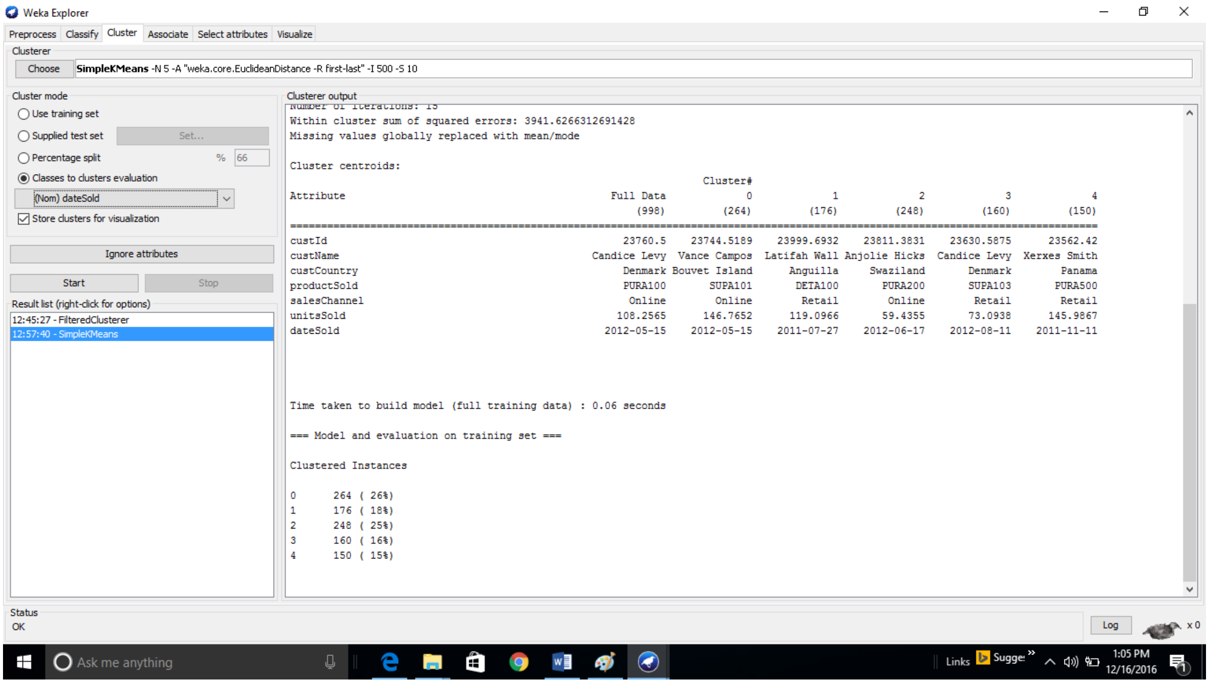
Result list (right-click for options)
- 12:45:27 - FilteredClusterer
- 12:57:40 - SimpleKMeans

Clusterer output

```
Number of iterations: 15
Within cluster sum of squared errors: 3941.6266312691428
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute                                    Full Data
                                                 (998)
==============================================================
custId                                         23760.5
custName                                  Candice Levy
custCountry                                    Denmark
productSold                                    PURA100
salesChannel                                    Online
unitsSold                                     108.2565
dateSold                                    2012-05-15




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

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

Status
OK

to execute the algorithm.



Clusterer: SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode
- ○ Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ● Classes to clusters evaluation
- (Nom) dateSold
- ☑ Store clusters for visualization

Result list (right-click for options)
- 12:45:27 - FilteredClusterer
- 12:57:40 - SimpleKMeans

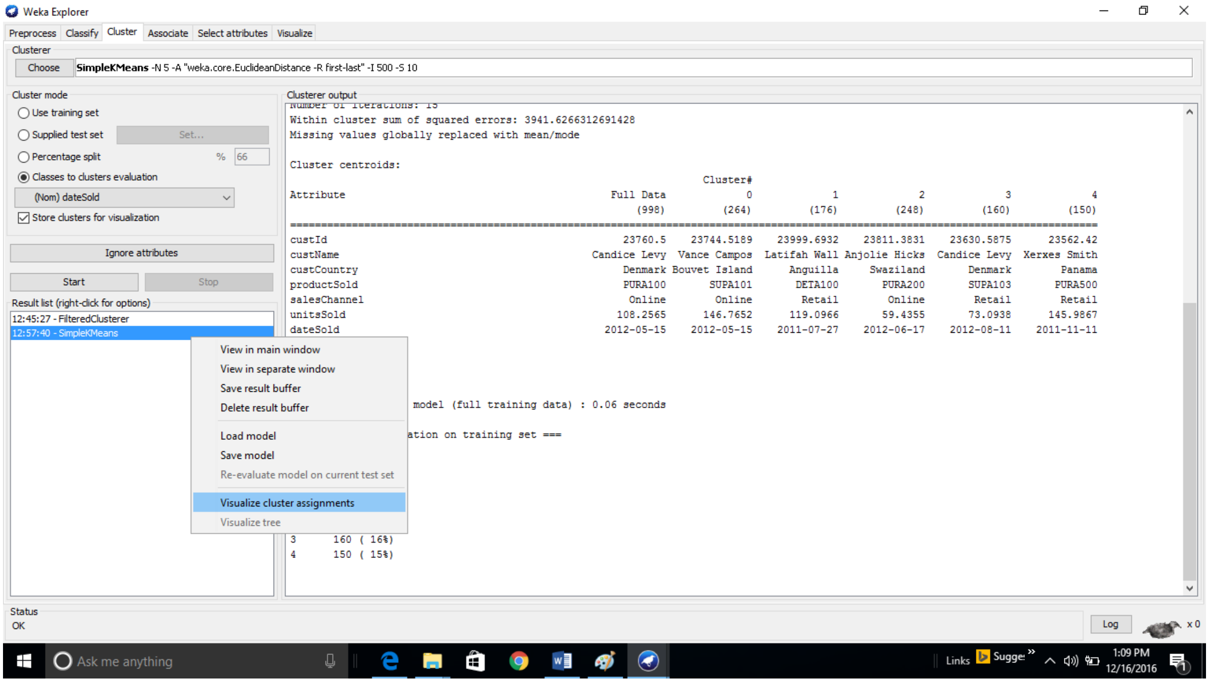
Clusterer output

```
Number of iterations: 15
Within cluster sum of squared errors: 3941.6266312691428
Missing values globally replaced with mean/mode

Cluster centroids:
                                                  Cluster#
Attribute                 Full Data            0            1            2            3            4
                              (998)        (264)        (176)        (248)        (160)        (150)
=====================================================================================================
custId                      23760.5   23744.5189   23999.6932   23811.3831   23630.5875     23562.42
custName               Candice Levy Vance Campos Latifah Wall Anjolie Hicks Candice Levy Xerxes Smith
custCountry                 Denmark Bouvet Island    Anguilla    Swaziland      Denmark       Panama
productSold                 PURA100      SUPA101      DETA100      PURA200      SUPA103      PURA500
salesChannel                 Online       Online       Retail       Online       Retail       Retail
unitsSold                  108.2565     146.7652     119.0966      59.4355      73.0938     145.9867
dateSold                 2012-05-15   2012-05-15   2011-07-27   2012-06-17   2012-08-11   2011-11-11




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

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

Clustered Instances

0       264 ( 26%)
1       176 ( 18%)
2       248 ( 25%)
3       160 ( 16%)
4       150 ( 15%)
```

Status
OK

4.4 Visualization of Results

Another way of representation of results of clustering is through visualization. Right-click on the entry in the ‘Result list’ and select ‘Visualize cluster assignments’ in the pull-down window.



Clusterer: SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

Cluster mode:
- ○ Use training set
- ○ Supplied test set
- ○ Percentage split % 66
- ● Classes to clusters evaluation
- (Nom) dateSold
- ☑ Store clusters for visualization

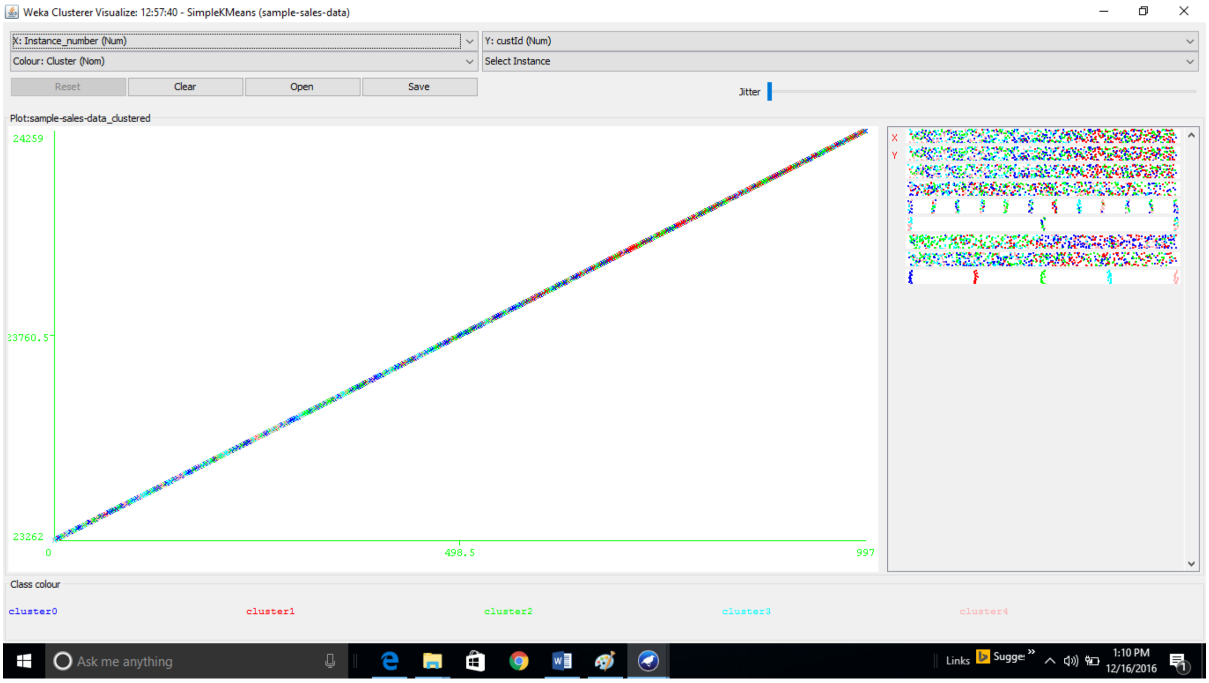
Result list (right-click for options):
- 12:45:27 - FilteredClusterer
- 12:57:40 - SimpleKMeans

Clusterer output:

```
Number of iterations: 15
Within cluster sum of squared errors: 3941.6266312691428
Missing values globally replaced with mean/mode

Cluster centroids:
                                          Cluster#
Attribute                    Full Data           0           1           2           3           4
                                 (998)       (264)       (176)       (248)       (160)       (150)
==================================================================================================
custId                         23760.5  23744.5189  23999.6932  23811.3831  23630.5875    23562.42
custName                   Candice Levy Vance Campos Latifah Wall Anjolie Hicks Candice Levy Xerxes Smith
custCountry                    Denmark Bouvet Island    Anguilla   Swaziland     Denmark      Panama
productSold                    PURA100     SUPA101     DETA100     PURA200     SUPA103     PURA500
salesChannel                    Online      Online      Retail      Online      Retail      Retail
unitsSold                     108.2565    146.7652    119.0966     59.4355     73.0938    145.9867
dateSold                    2012-05-15  2012-05-15  2011-07-27  2012-06-17  2012-08-11  2011-11-11



              model (full training data) : 0.06 seconds

             ation on training set ===




3       160 ( 16%)
4       150 ( 15%)
```

Status: OK

This brings up the ‘Weka Clusterer Visualize’ window.



X: Instance_number (Num)

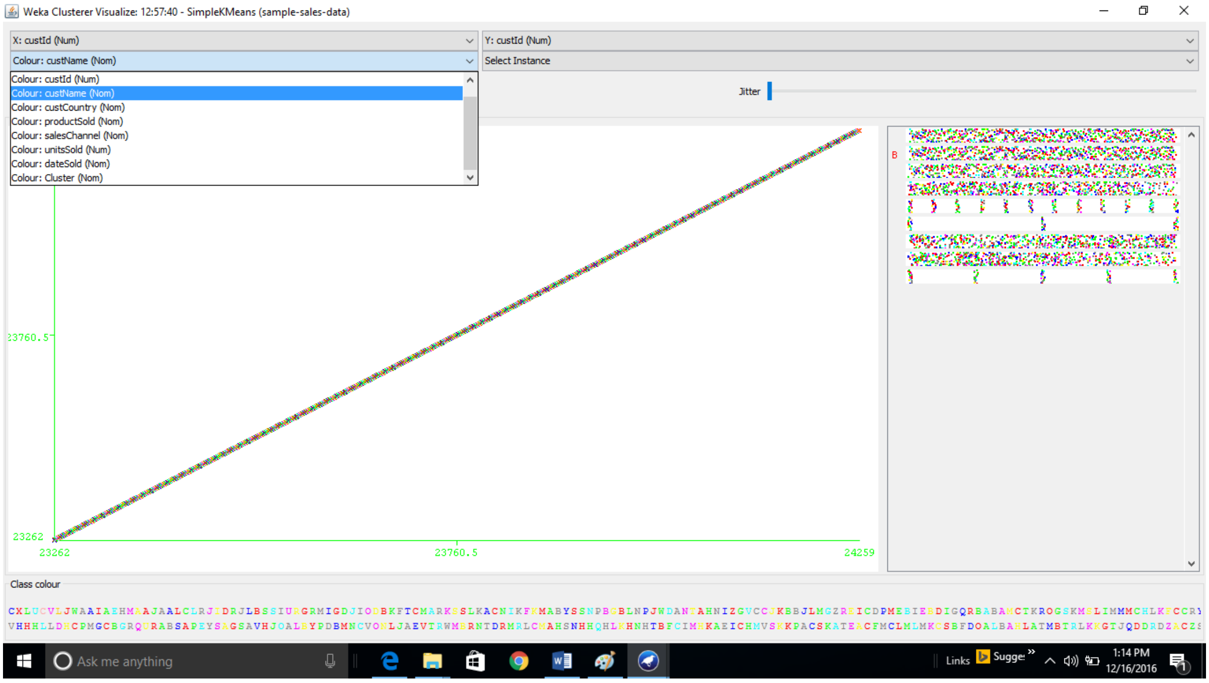
Y: custId (Num)

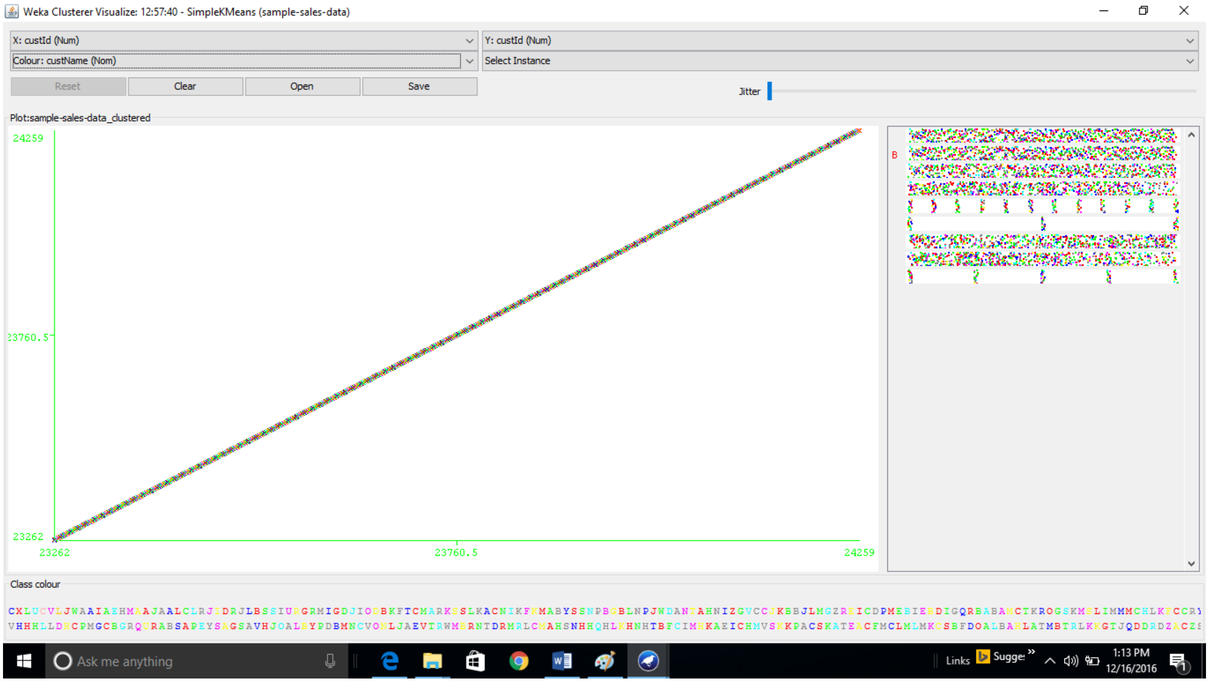
Colour: Cluster (Nom)

Select Instance

Plot:sample-sales-data_clustered

Class colour: cluster0 cluster1 cluster2 cluster3 cluster4


Weka Clusterer Visualize: 12:57:40 - SimpleKMeans (sample-sales-data)
X: custId (Num)
Y: custId (Num)
Colour: custName (Nom)
Select Instance
Colour: custId (Num)
Colour: custName (Nom)
Colour: custCountry (Nom)
Colour: productSold (Nom)
Colour: salesChannel (Nom)
Colour: unitsSold (Num)
Colour: dateSold (Nom)
Colour: Cluster (Nom)
Jitter
23760.5
23262
23262
23760.5
24259
B
Class colour
CXLUCVLJWAAIAEHMOAJAALCLRJ DRJLBSSIURGRMIGDJIODBKFTCMARKOSLKACNIKFOMABYSSNPBOBLNPJWDANOAHNIZGVCCOKBBJLMGZRXICDPMEBIENDIGQRBABAOCTKROGSKMOLIMMMCHLKOCCRY
VHHHLLDOCPMGCBGRQ RABSAPEYSAGSAVHJOALOYPDBMNCVOOLJAEVTRWMGRNTDRMRLCOAHSNHHQHLOHNHTBFCIMXKAEICHMVSOKPACSKATEOCFMCLMLMKOSBFDOALBAOLATMBTRLKOGTJQDDRDZACZS
Ask me anything
Links
Sugge
1:14 PM
12/16/2016

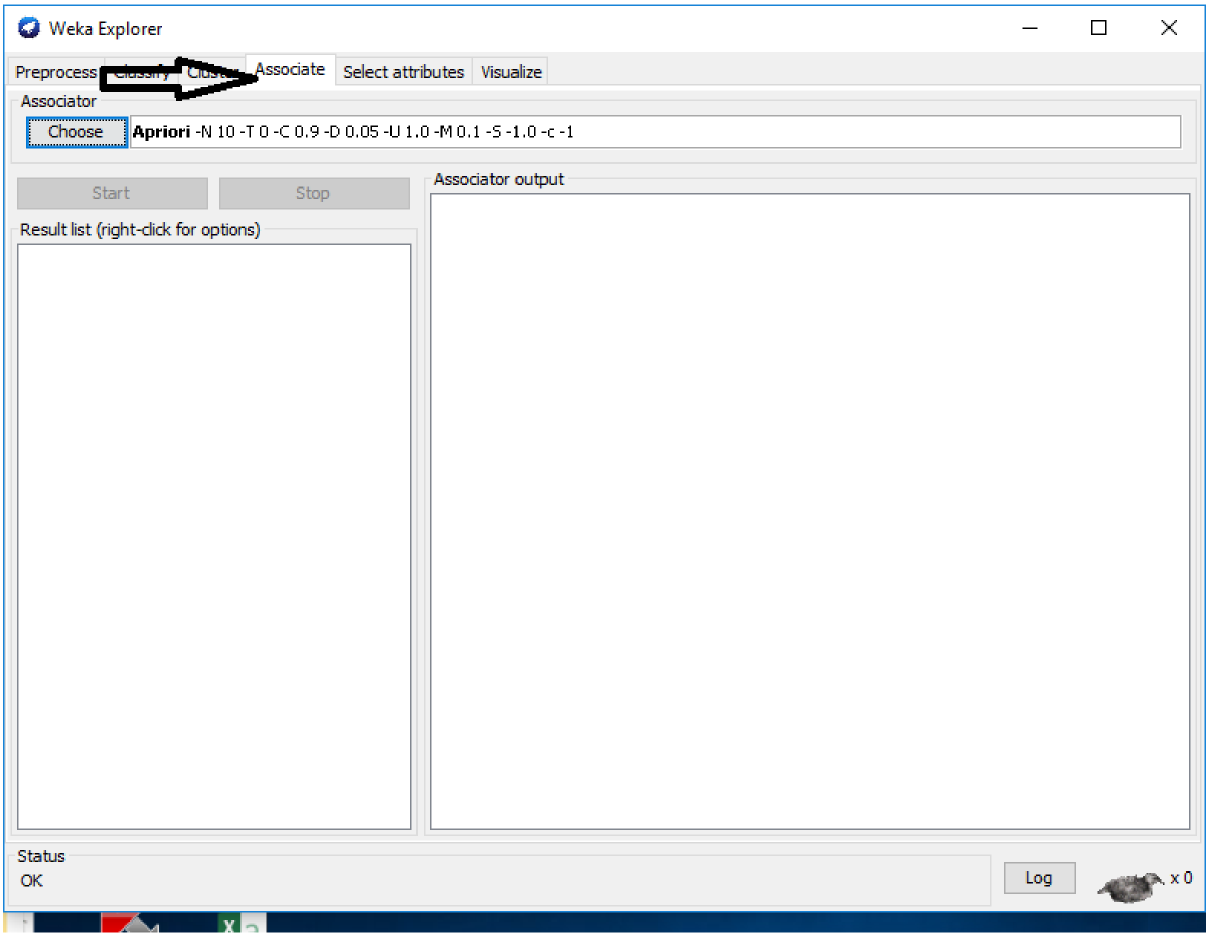

Weka Clusterer Visualize: 12:57:40 - SimpleKMeans (sample-sales-data)
X: custId (Num)
Y: custId (Num)
Colour: custName (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:sample-sales-data_clustered
24259
23760.5
23262
23262
23760.5
24259
B
Class colour
CXLUCVLJWAAIAEHMOAJAALCLRJ DRJLBSSIURGRMIGDJIODBKFTCMARKOSLKACNIKFOMABYSSNPBOBLNPJWDANOAHNIZGVCCOKBBJLMGZRXICDPMEBIENDIGQRBABAOCTKROGSKMOLIMMMCHLKOCCRY
VHHHLLDOCPMGCBGRQ RABSAPEYSAGSAVHJOALOYPDBMNCVOOLJAEVTRWMGRNTDRMRLCOAHSNHHQHLOHNHTBFCIMXKAEICHMVSOKPACSKATEOCFMCLMLMKOSBFDOALBAOLATMBTRLKOGTJQDDRDZACZS
Ask me anything
Links
Sugge
1:13 PM
12/16/2016

On the ‘Weka Clusterer Visualize’ window, beneath the X-axis selector there is a dropdown list, ‘Colour’, for choosing the color scheme. This allows you to choose the color of points based on the attribute selected. Below the plot area, there is a legend that describes what values the colors correspond to.

CHAPTER 5: ASSOCIATION

5.1 Finding Associations

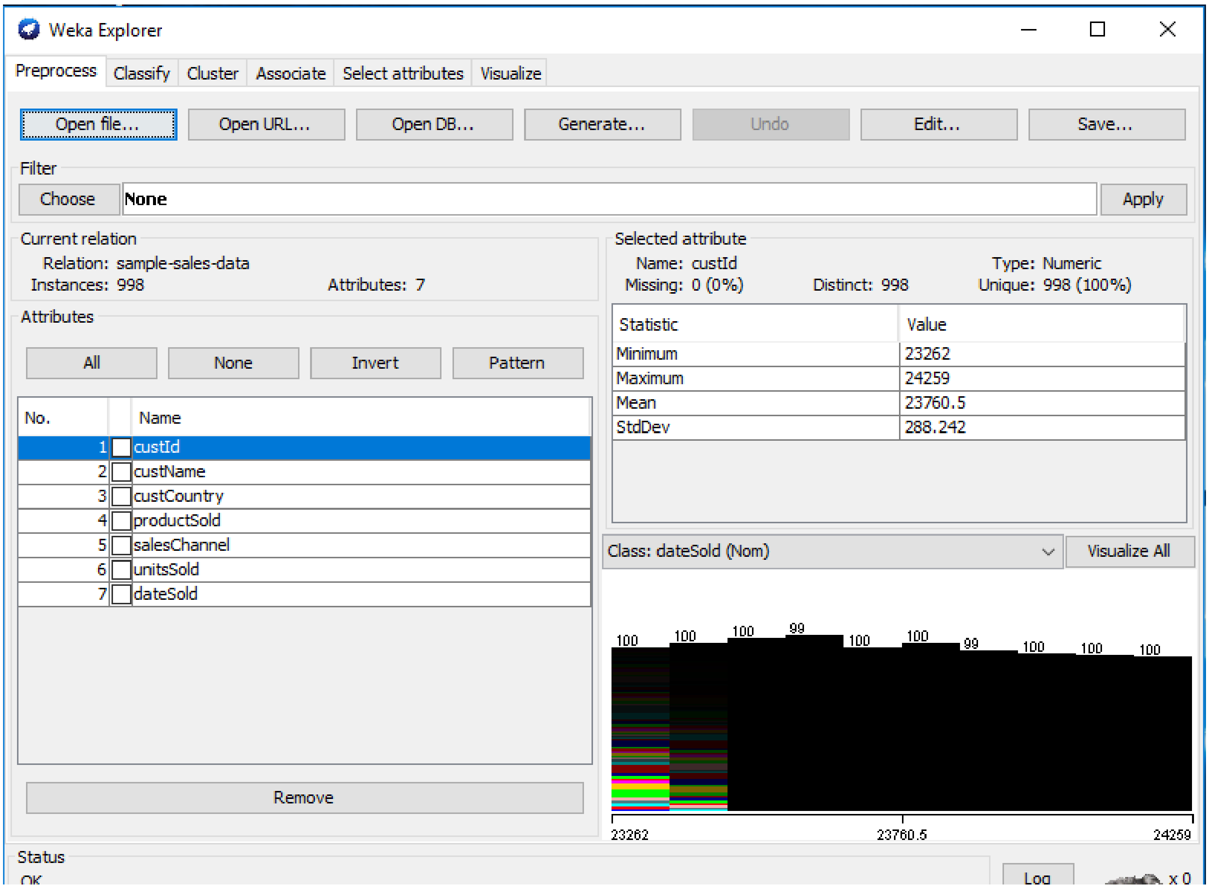
WEKA contains an implementation of the Apriori algorithm for learning association rules. This is the only currently available scheme for learning associations in WEKA. It works only with discrete data and will identify statistical dependencies between groups of attributes. Apriori can compute all rules that have a given minimum support and exceed a given confidence.



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

OK

For this exercise you will use sales data from the “sales-sample-data.csv” file.



Filter: None

Current relation
Relation: sample-sales-data
Instances: 998
Attributes: 7

Selected attribute
Name: custId
Type: Numeric
Missing: 0 (0%)
Distinct: 998
Unique: 998 (100%)

Statistic	Value
Minimum	23262
Maximum	24259
Mean	23760.5
StdDev	288.242

Attributes

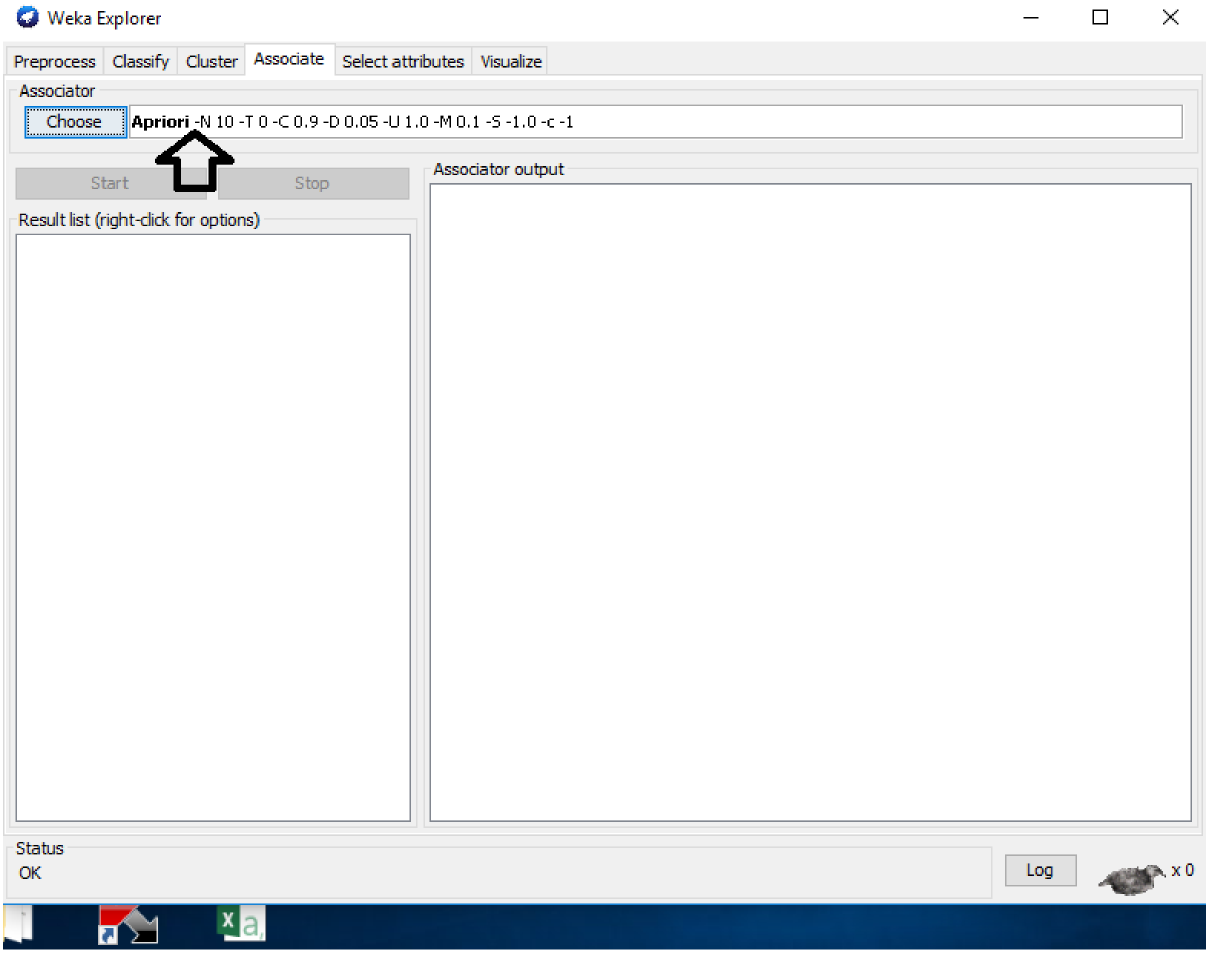
No.		Name
1	□	custId
2	□	custName
3	□	custCountry
4	□	productSold
5	□	salesChannel
6	□	unitsSold
7	□	dateSold

Class: dateSold (Nom)

Status
OK

5.2 Setting Test Options

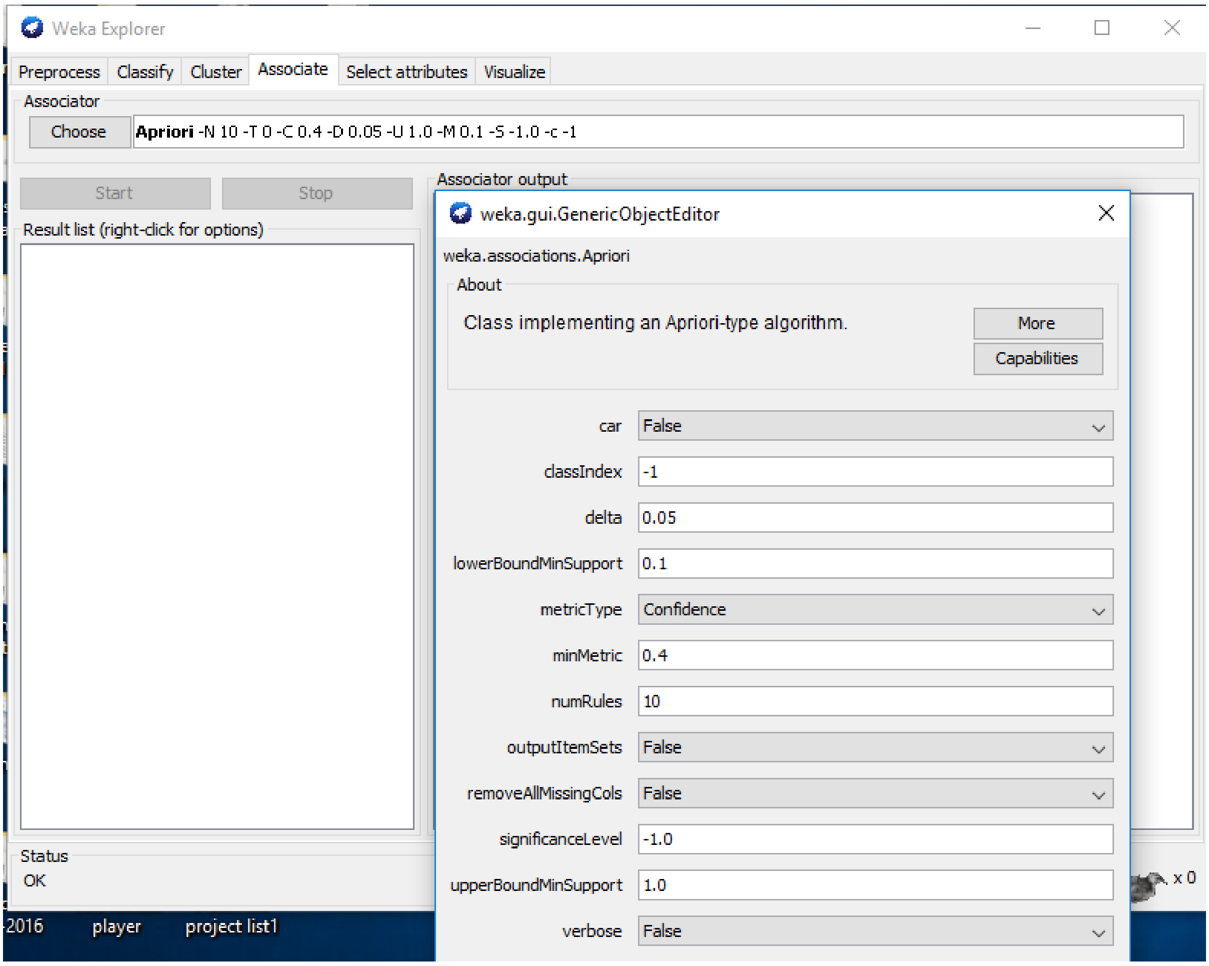
Check the text field in the ‘Associator’ box at the top of the window. As you can see, there are no other associators to choose and no extra options for testing the learning scheme



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

OK

Right-click on the ‘Associator’ box, and click on show properties,‘GenericObjectEditor’ appears on your screen. In the dialog box, change the value in ‘minMetric’ to 0.4 for confidence = 40%. Make sure that the default value of rules is set to 100. The upper bound for minimum support ‘upperBoundMinSupport’ should be set to 1.0 (100%) and ‘lowerBoundMinSupport’ to 0.1. Apriori in WEKA starts with the upper bound support and incrementally decreases support (by delta increments, which by default is set to 0.05 or 5%). The algorithm halts when either the specified number of rules is generated, or the lower bound for minimum support is reached. The ‘significanceLevel’ testing option is only applicable in the case of confidence and is (-1.0) by default (not used).



Apriori -N 10 -T 0 -C 0.4 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

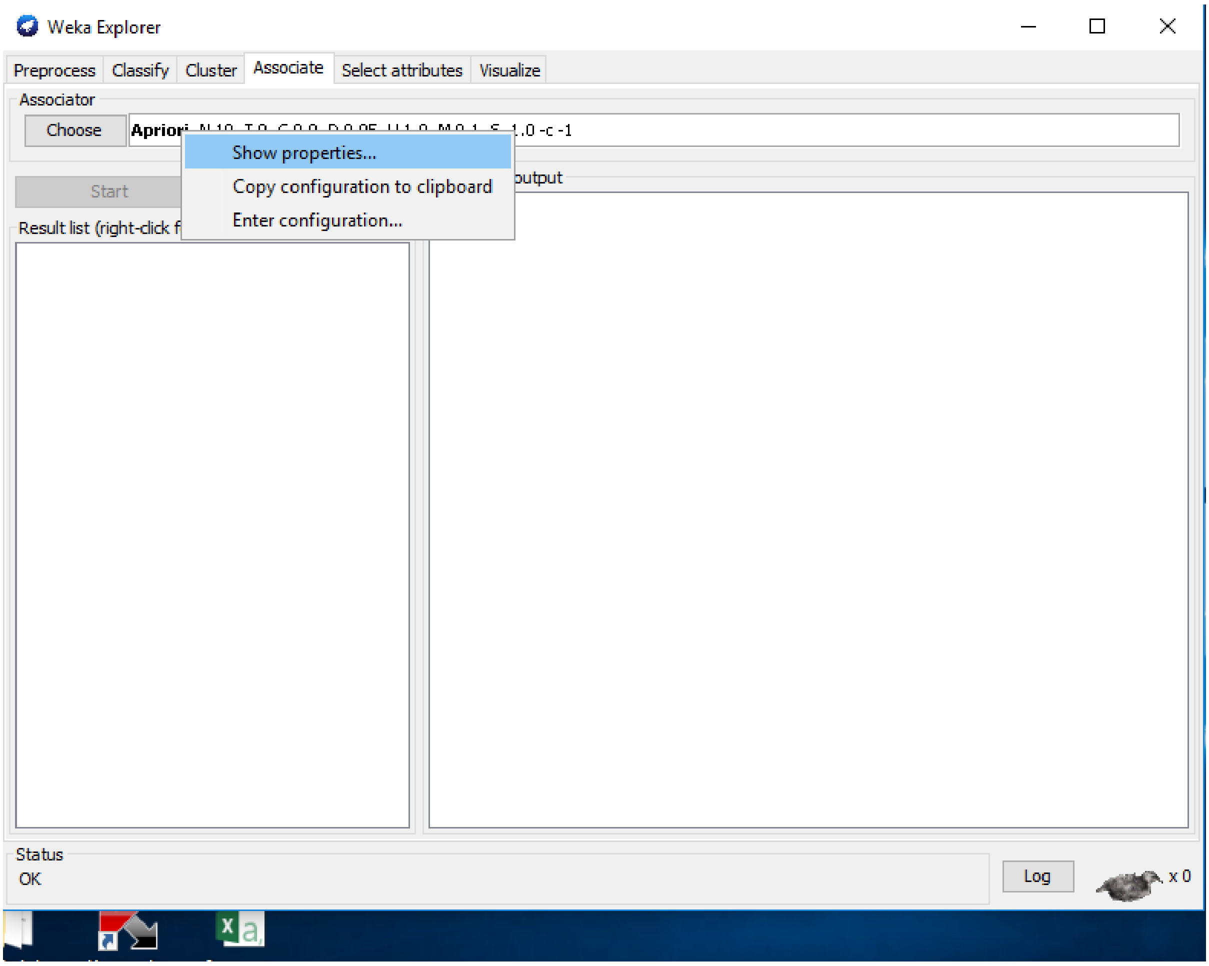
weka.associations.Apriori

About

Class implementing an Apriori-type algorithm.

car	False
classIndex	-1
delta	0.05
lowerBoundMinSupport	0.1
metricType	Confidence
minMetric	0.4
numRules	10
outputItemSets	False
removeAllMissingCols	False
significanceLevel	-1.0
upperBoundMinSupport	1.0
verbose	False

Once the options have been specified, you can run Apriori algorithm. Click on the ‘Start’ button to execute the algorithm.



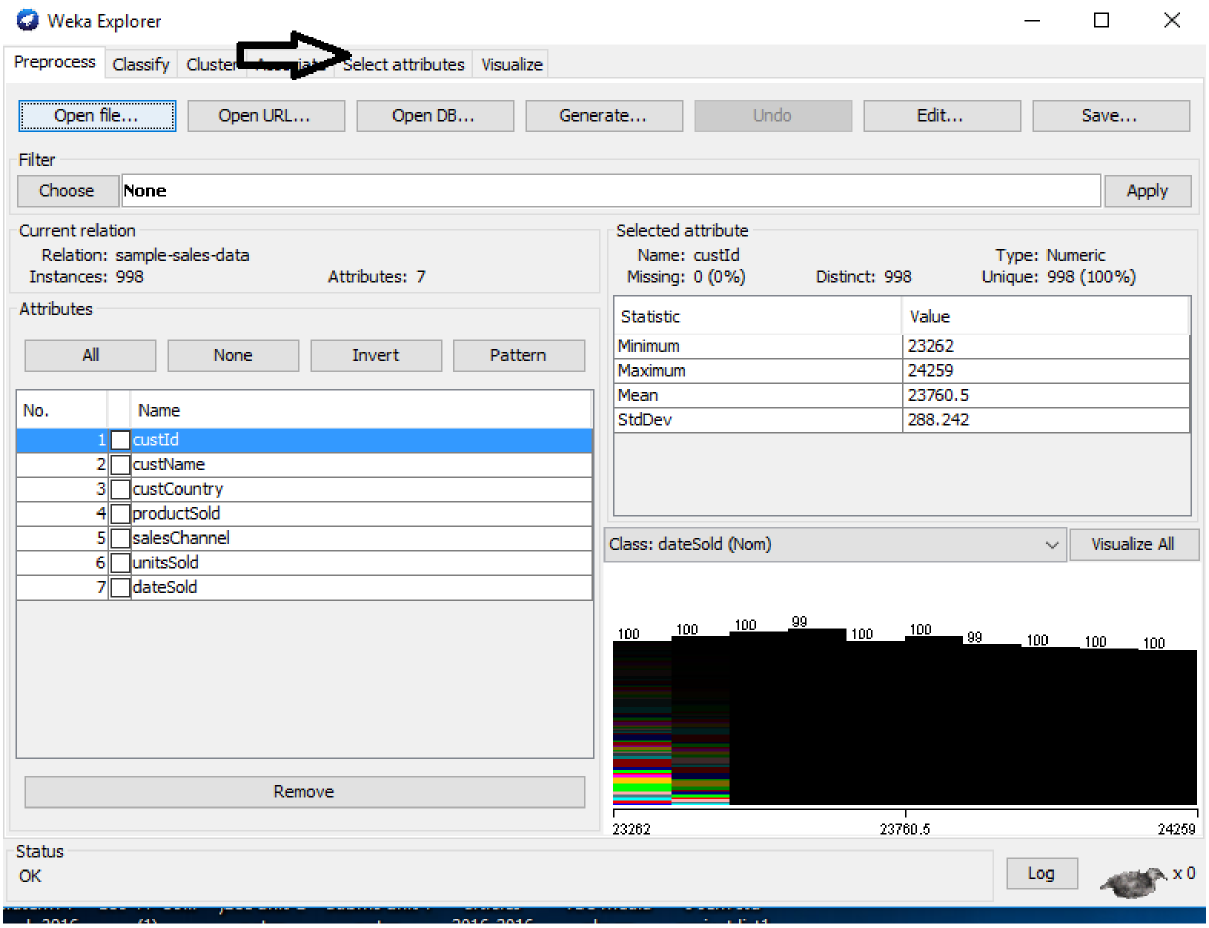
Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

CHAPTER 6: ATTRIBUTE SELECTION

6.1 Introduction:

Attribute selection searches through all possible combinations of attributes in the data and finds which subset of attributes works best for prediction. Attribute selection methods contain two parts: a search method such as best-first, forward selection, random, exhaustive, genetic algorithm, ranking, and an evaluation method such as correlation-based, wrapper, information gain, chi-squared. Attribute selection mechanism is very flexible - WEKA allows (almost) arbitrary combinations of the two methods.

To begin an attribute selection, click 'Select attributes' tab.



Filter: None

Current relation
Relation: sample-sales-data
Instances: 998 Attributes: 7

No.		Name
1	□	custId
2	□	custName
3	□	custCountry
4	□	productSold
5	□	salesChannel
6	□	unitsSold
7	□	dateSold

Selected attribute
Name: custId Type: Numeric
Missing: 0 (0%) Distinct: 998 Unique: 998 (100%)

Statistic	Value
Minimum	23262
Maximum	24259
Mean	23760.5
StdDev	288.242

Class: dateSold (Nom)

Status
OK

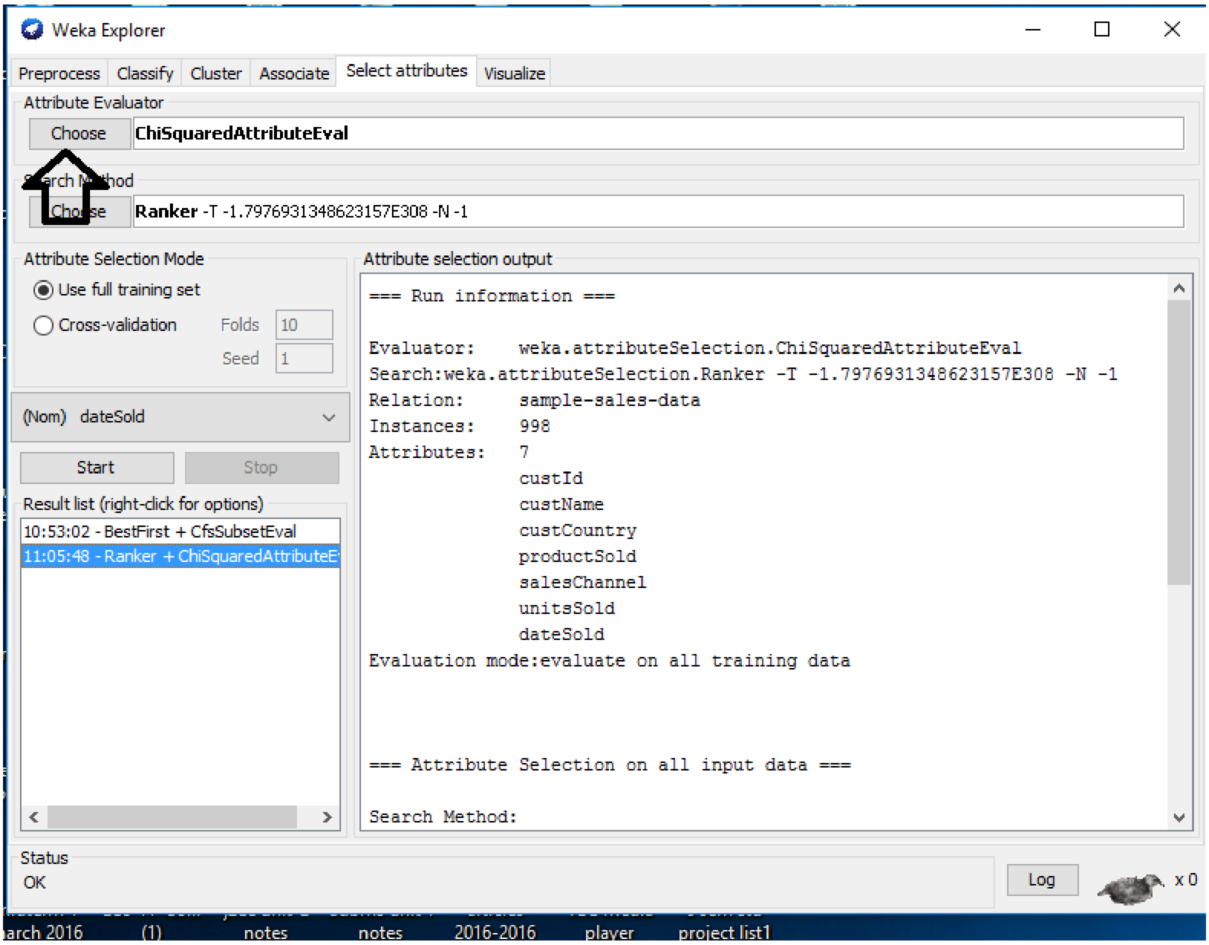
6.2 Selecting Options

To search through all possible combinations of attributes in the data and find which subset of attributes works best for prediction, make sure that you set up attribute evaluator to 'CfsSubsetEval' and a search method to 'BestFirst'. The evaluator will determine what method to use to assign a worth to each subset of attributes. The search method will determine what style of search to perform. The options that you can set for selection in the 'Attribute Selection Mode' box are :

1. **Use full training set**. The worth of the attribute subset is determined using the full set of training data.

2. Cross-validation. The worth of the attribute subset is determined by a process of cross-validation. The 'Fold' and 'Seed' fields set the number of folds to use and the random seed used when shuffling the data.

Specify which attribute to treat as the class in the drop-down box below the test options. Once all the test options are set, you can start the attribute selection process by clicking on 'Start' button.



Attribute Evaluator: ChiSquaredAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

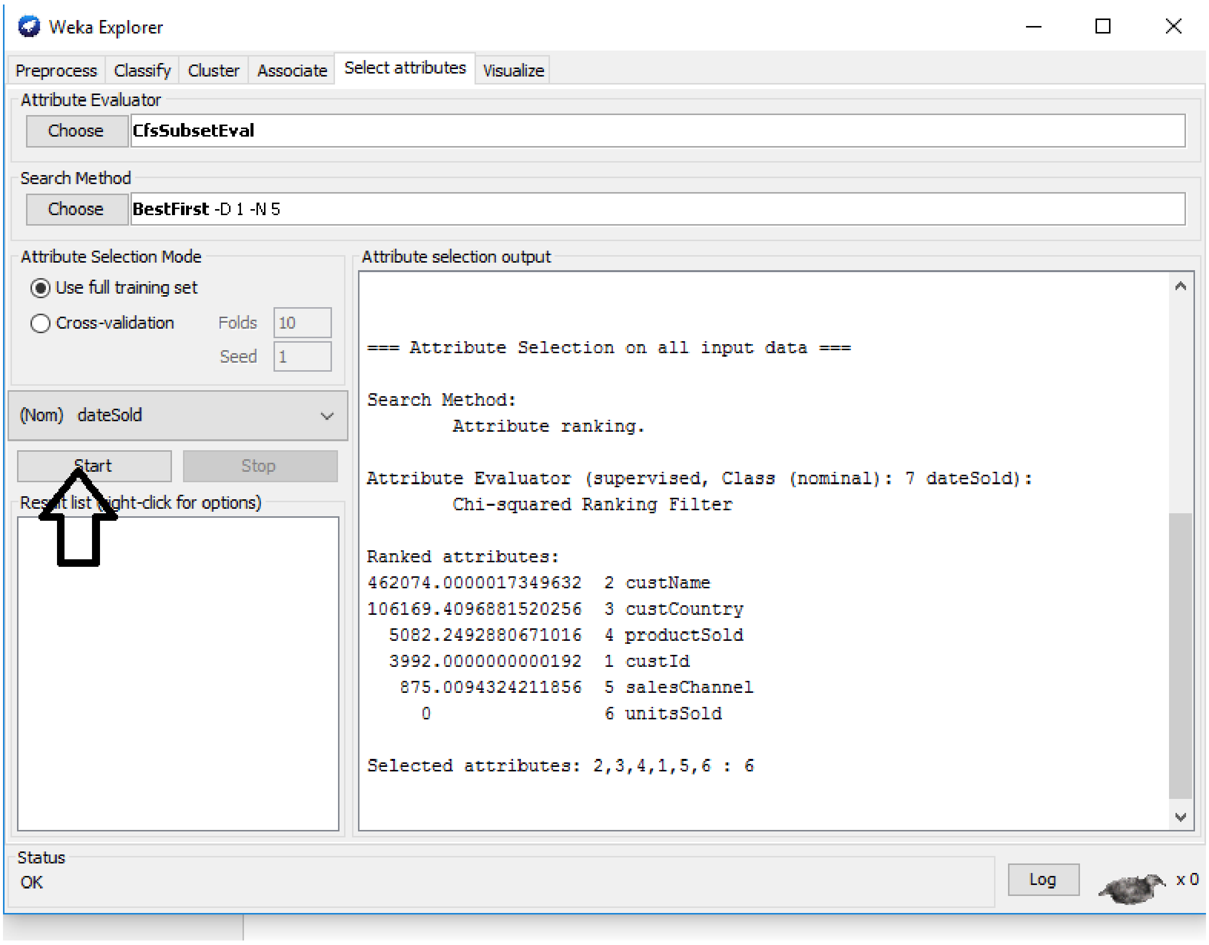
● Use full training set

○ Cross-validation Folds 10 Seed 1

(Nom) dateSold

10:53:02 - BestFirst + CfsSubsetEval
11:05:48 - Ranker + ChiSquaredAttributeE

```
=== Run information ===

Evaluator:    weka.attributeSelection.ChiSquaredAttributeEval
Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:     sample-sales-data
Instances:    998
Attributes:   7
              custId
              custName
              custCountry
              productSold
              salesChannel
              unitsSold
              dateSold
Evaluation mode:evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
```



Attribute Evaluator: CfsSubsetEval

Search Method: BestFirst -D 1 -N 5

Attribute Selection Mode:
- ● Use full training set
- ○ Cross-validation Folds 10 Seed 1

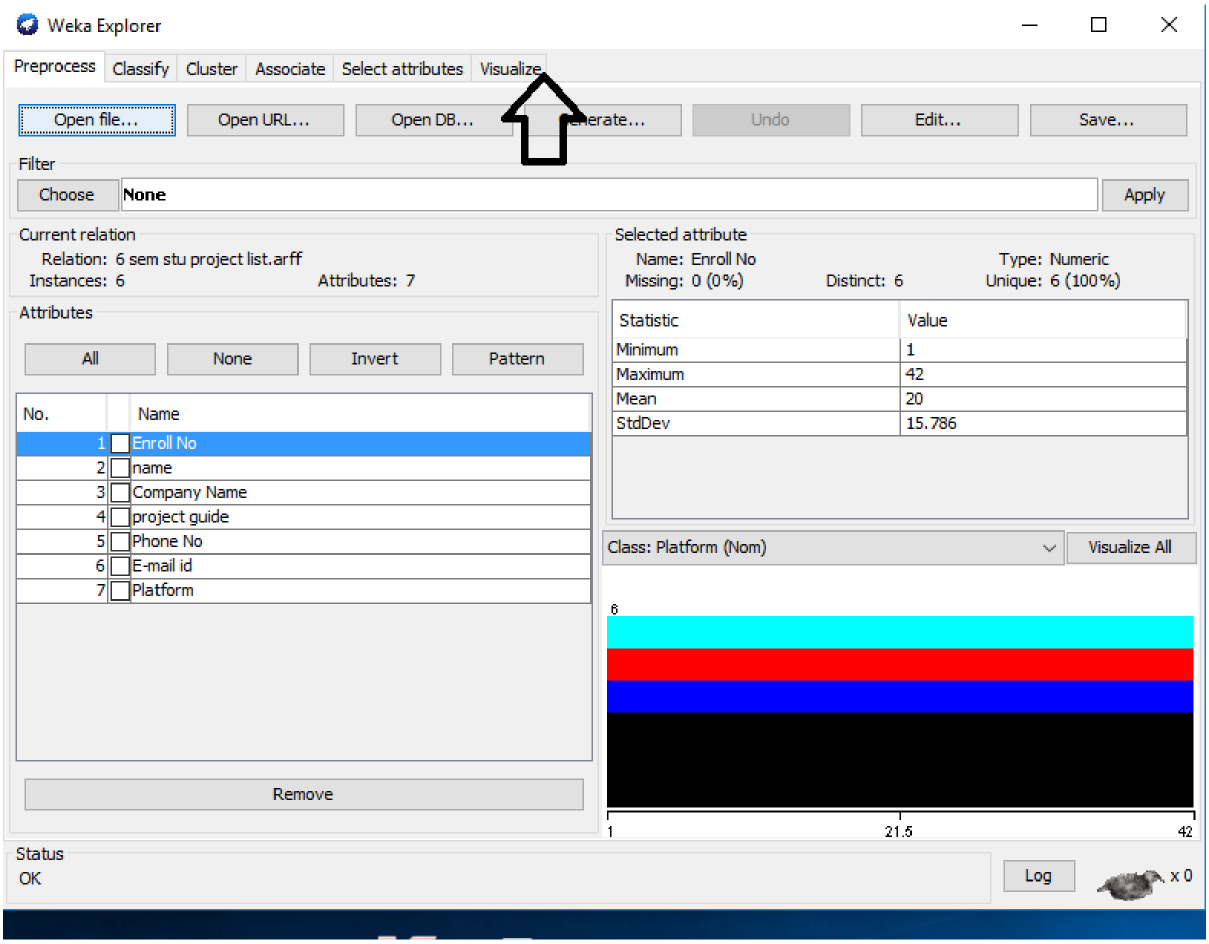
(Nom) dateSold

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 7 dateSold):
        Chi-squared Ranking Filter

Ranked attributes:
462074.0000017349632  2 custName
106169.4096881520256  3 custCountry
  5082.2492880671016  4 productSold
  3992.0000000000192  1 custId
   875.0094324211856  5 salesChannel
     0                6 unitsSold

Selected attributes: 2,3,4,1,5,6 : 6
```

CHAPTER 7: DATA VISUALIZATION

7.1 Introduction:

WEKA's visualization allows you to visualize a 2-D plot of the current working relation. Visualization is very useful in practice, it helps to determine difficulty of the learning problem. WEKA can visualize single attributes (1-d) and pairs of attributes (2-d), rotate 3-d visualizations (Xgobi-style). WEKA has "Jitter" option to deal with nominal attributes and to detect "hidden" data points.



Filter: None

Current relation
Relation: 6 sem stu project list.arff
Instances: 6 Attributes: 7

Attributes

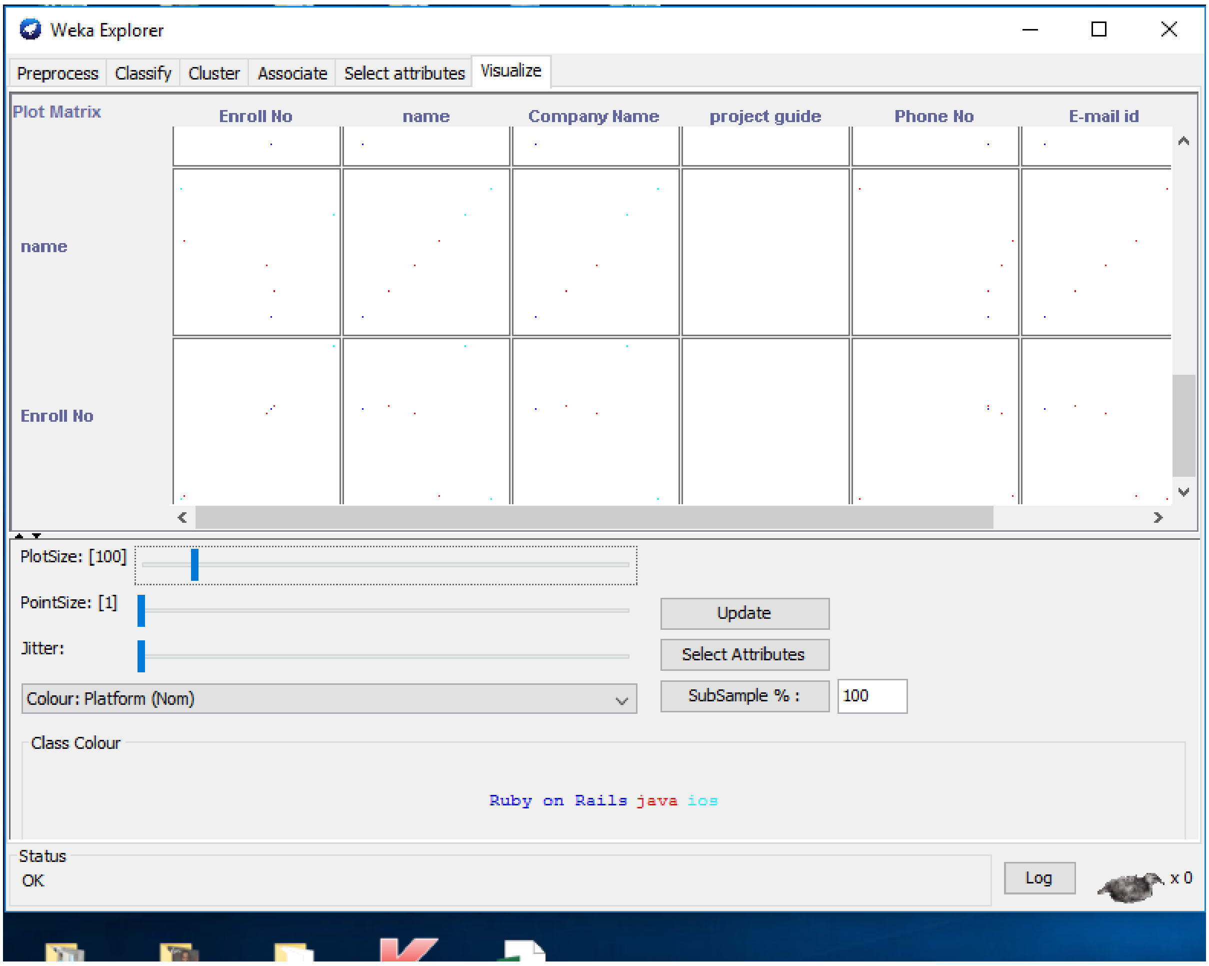
No.		Name
1	□	Enroll No
2	□	name
3	□	Company Name
4	□	project guide
5	□	Phone No
6	□	E-mail id
7	□	Platform

Selected attribute
Name: Enroll No Type: Numeric
Missing: 0 (0%) Distinct: 6 Unique: 6 (100%)

Statistic	Value
Minimum	1
Maximum	42
Mean	20
StdDev	15.786

Class: Platform (Nom)

Status
OK



PlotSize: [100]

PointSize: [1]

Jitter:

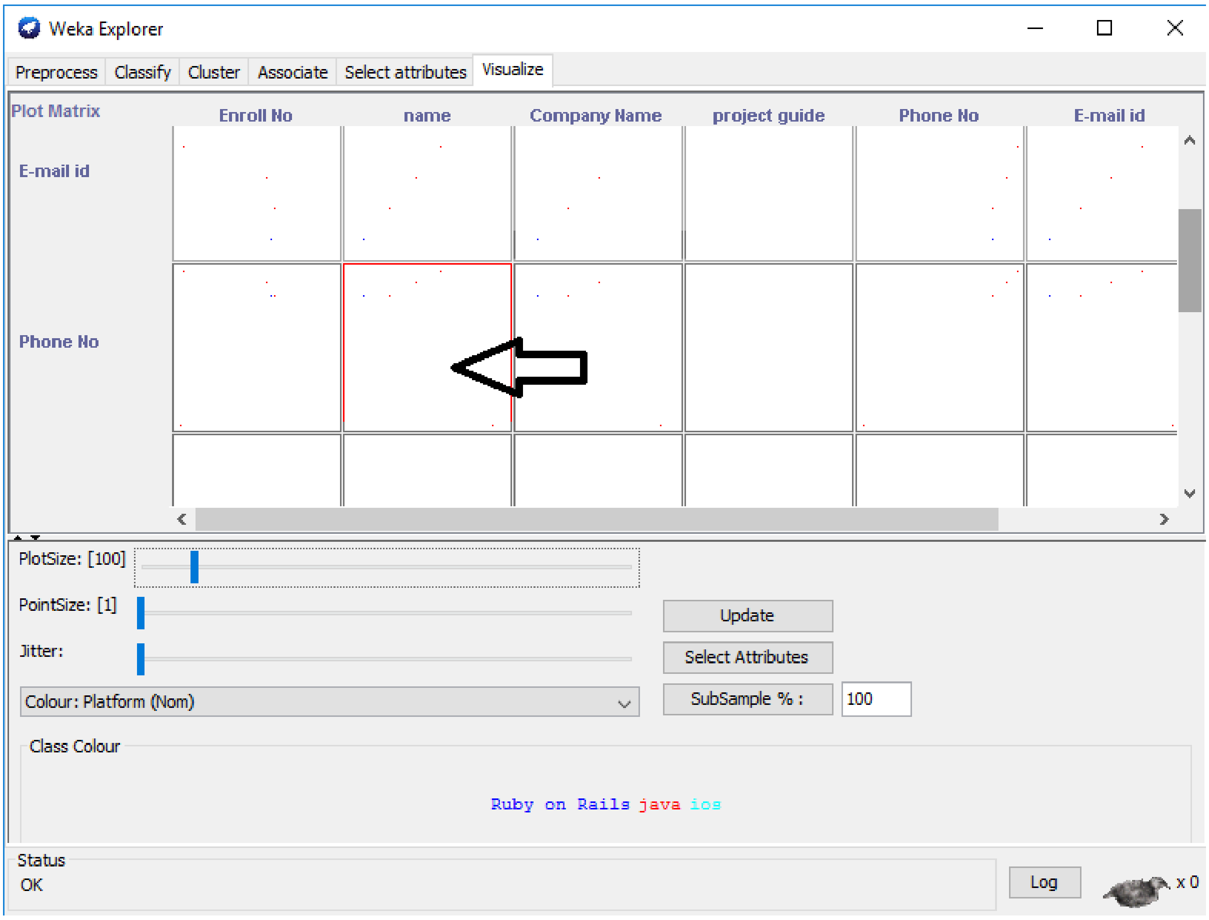
Colour: Platform (Nom)

SubSample % : 100

Status

OK

Select a square that corresponds to the attributes you would like to visualize. For example, let’s choose ‘outlook’ for X – axis and ‘play’ for Y – axis. Click anywhere inside the square that corresponds to ‘play on the left and ‘outlook’ at the top.



PlotSize: [100]

PointSize: [1]

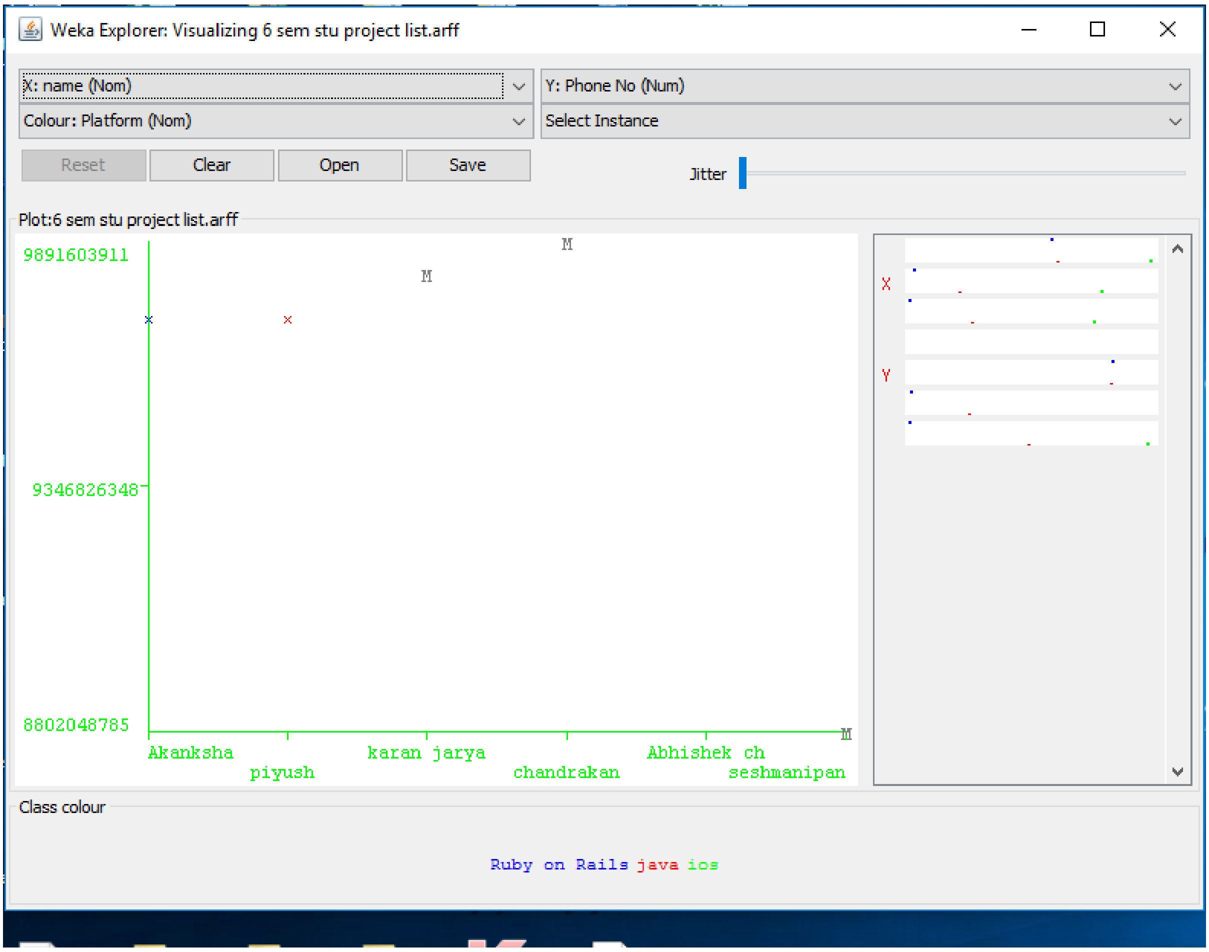
Jitter:

Colour: Platform (Nom)

SubSample % : 100

Status: OK

A ‘Visualizing r’ window appears on the screen

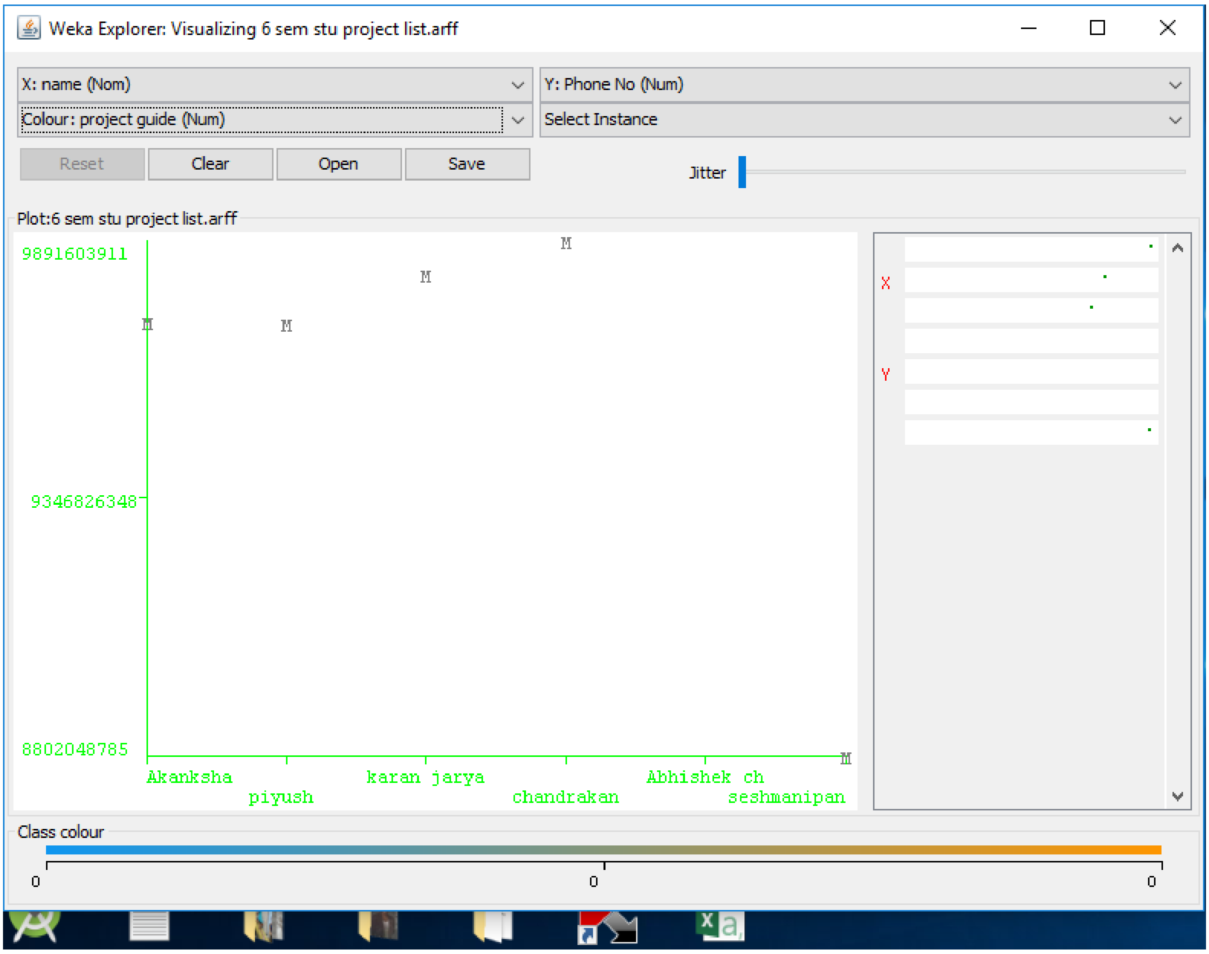


7.2 Changing the View

In the visualization window, beneath the X-axis selector there is a drop-down list, 'Colour', for choosing the color scheme. This allows you to choose the color of points based on the attribute selected.

Below the plot area, there is a legend that describes what values the colors correspond to. In your example, red represents 'no', while blue represents 'yes'. For better visibility you should change the color of label 'yes'. Left-click on 'yes' in the 'Class colour' box and select lighter color from the color palette. To the right of the plot area there are series of horizontal strips.

Each strip represents an attribute, and the dots within it show the distribution values of the attribute. You can choose what axes are used in the main graph by clicking on these strips (left-click changes X-axis, rightclick changes Y-axis). The software sets X - axis to 'Outlook' attribute and Y - axis to 'Play'. The instances are spread out in the plot area and concentration points are not visible. Keep sliding 'Jitter', a random displacement given to all points in the plot, to the right, until you can spot concentration points.



Weka Explorer: Visualizing 6 sem stu project list.arff

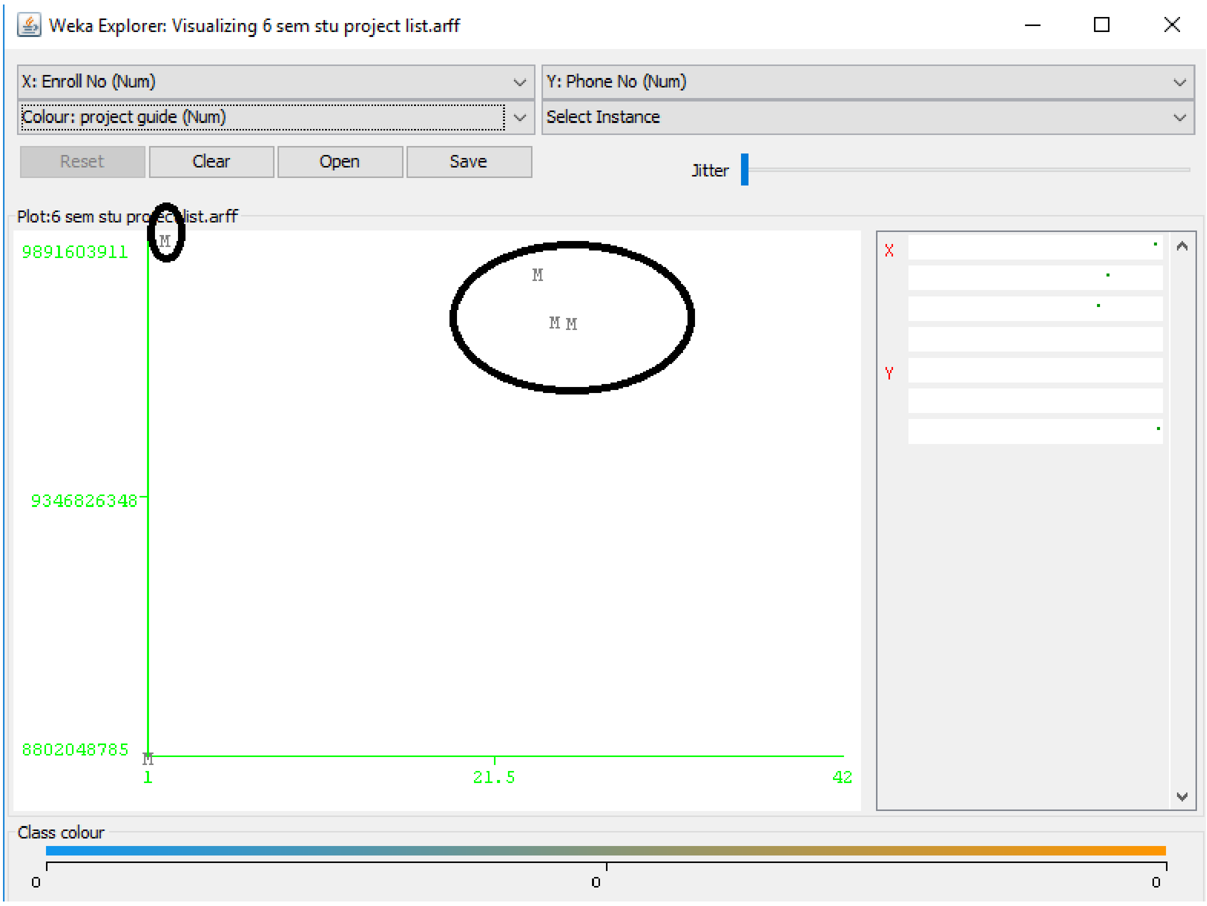
X: name (Nom)

Y: Phone No (Num)

Colour: project guide (Num)

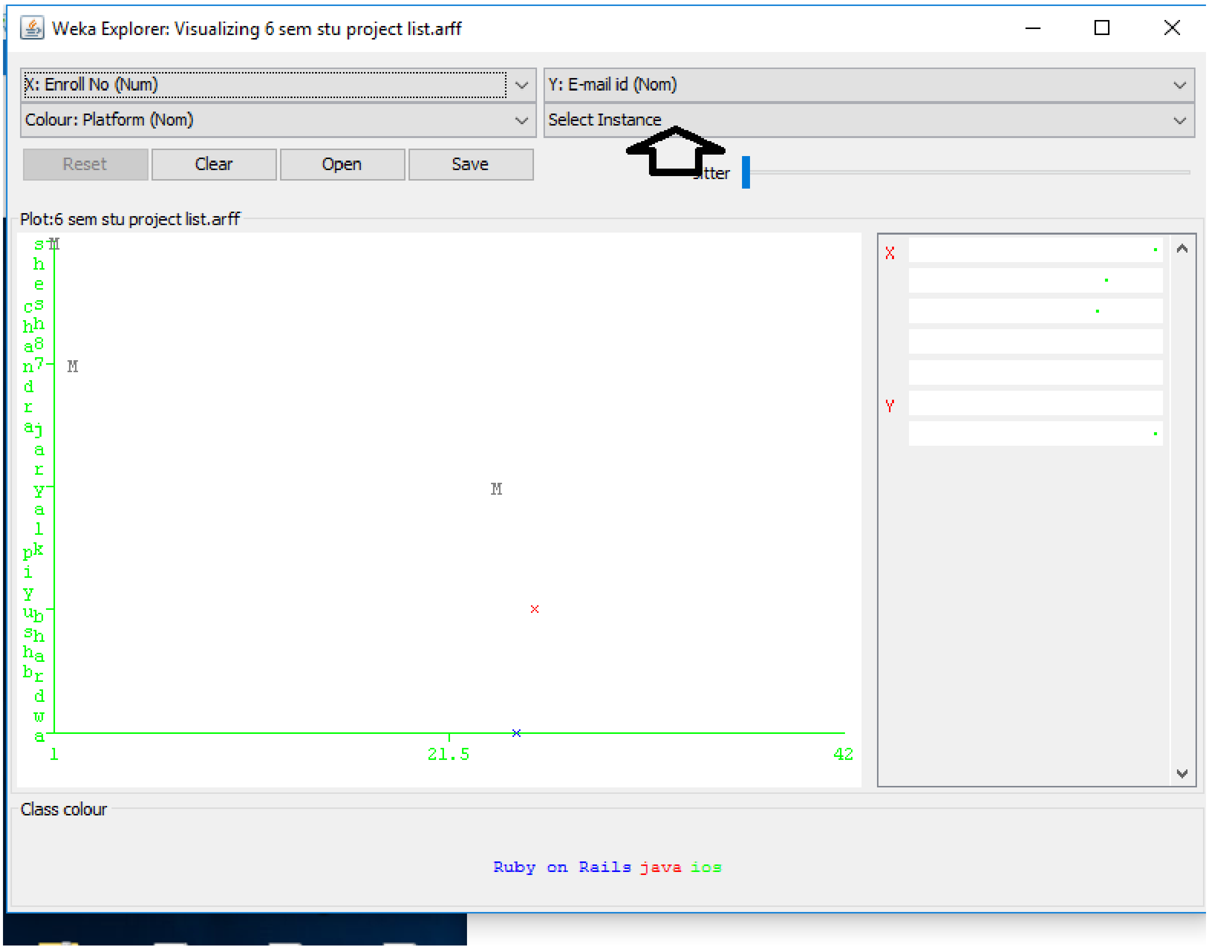
Select Instance

Plot:6 sem stu project list.arff



7.3 Selecting Instances:

Sometimes it is helpful to select a subset of the data using visualization tool. A special case is the ‘User Classifier’, which lets you to build your own classifier by interactively selecting instances. Below the Y – axis there is a drop-down list that allows you to choose a selection method. A group of points on the graph can be selected in four ways.



Weka Explorer: Visualizing 6 sem stu project list.arff

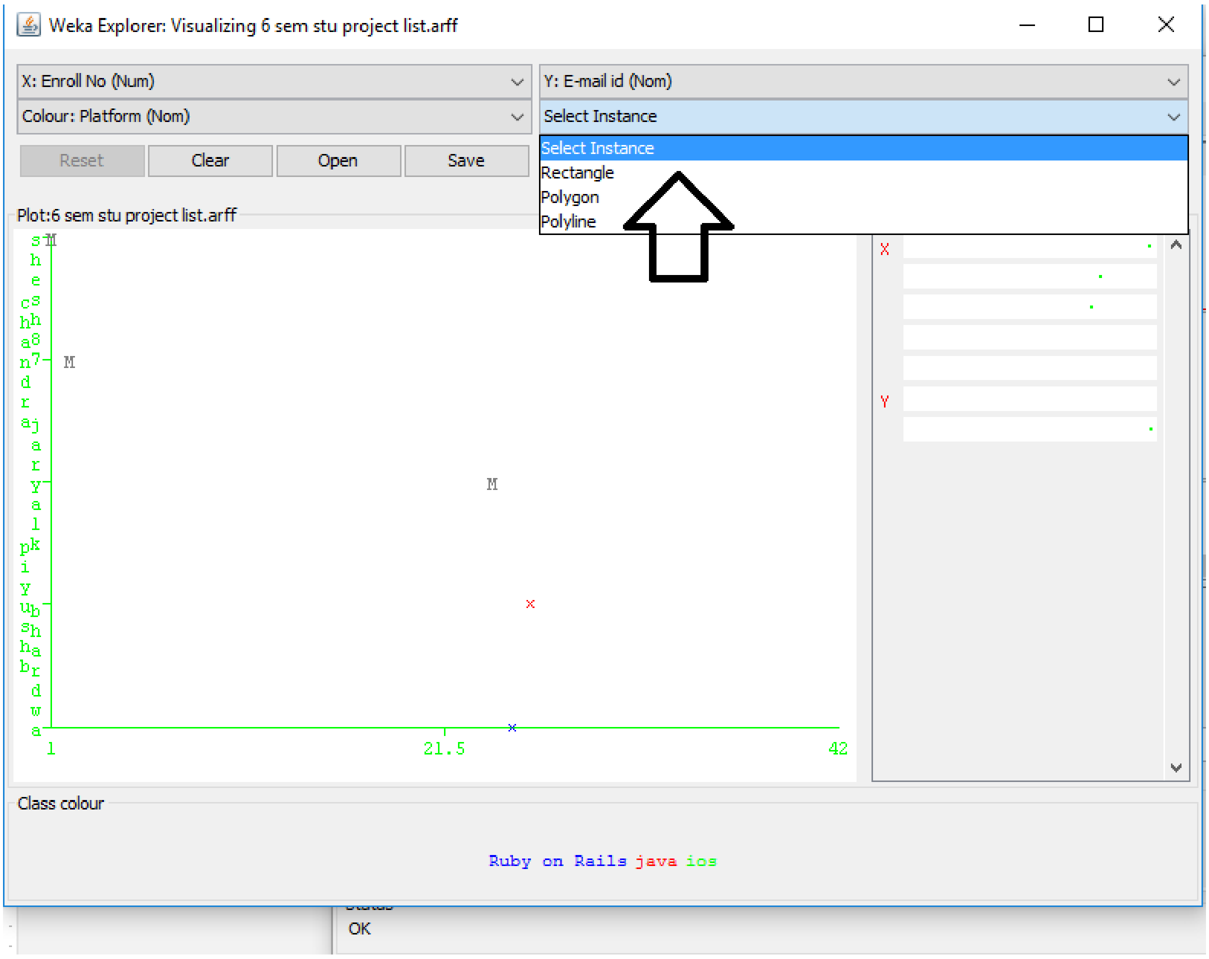
X: Enroll No (Num)

Y: E-mail id (Nom)

Colour: Platform (Nom)

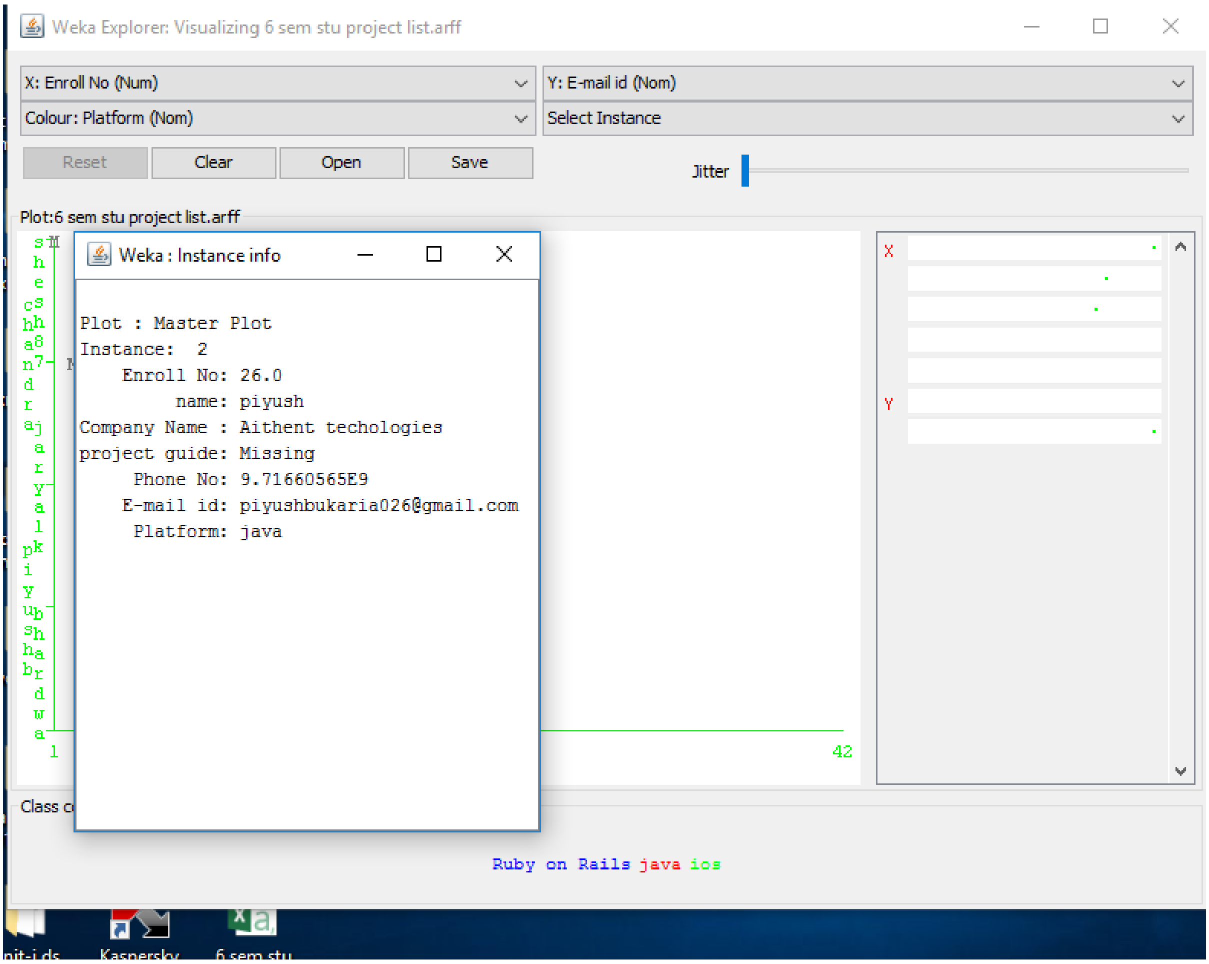
Plot:6 sem stu project list.arff

Ruby on Rails java ios



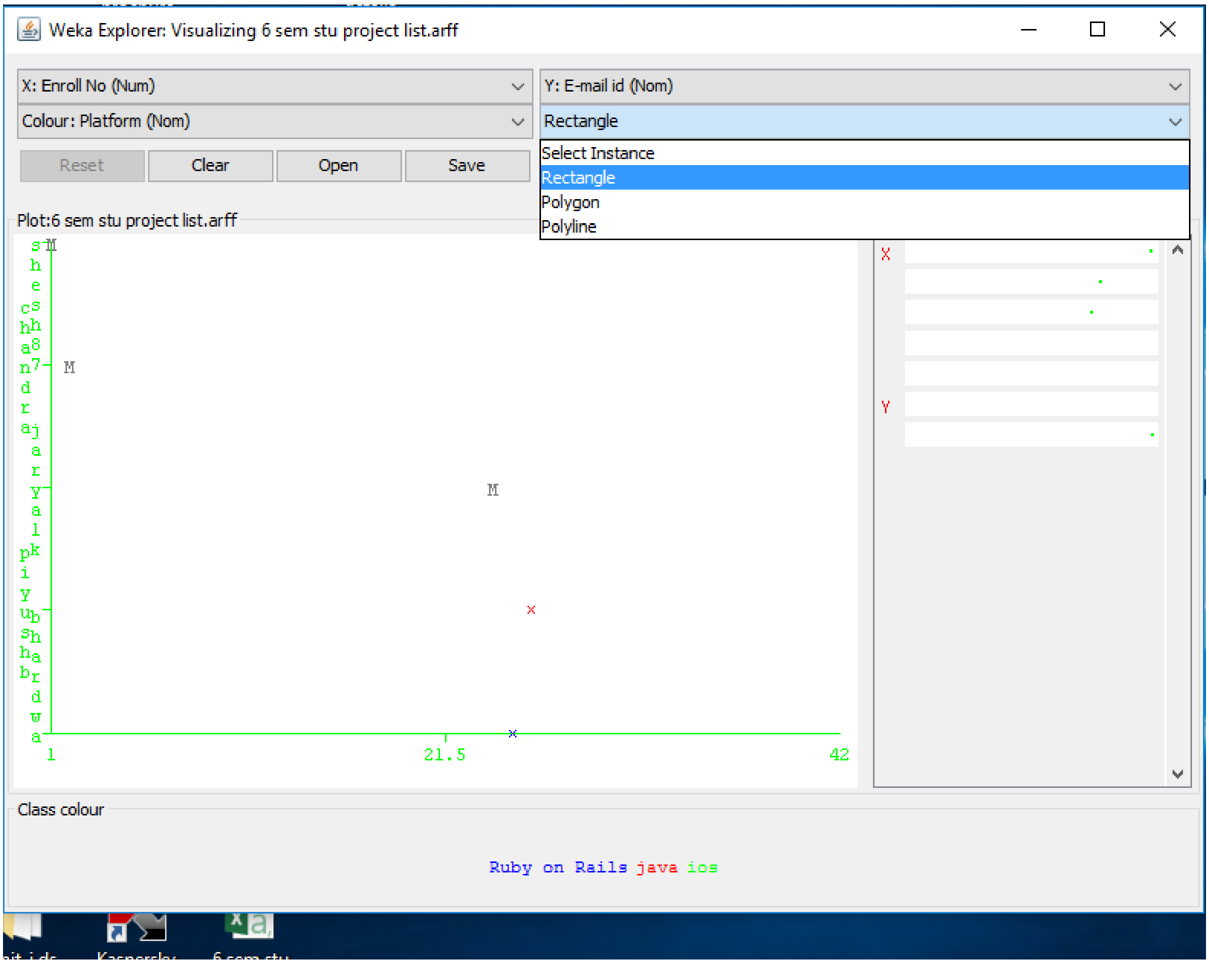
1. Select Instance

Click on an individual data point. It brings up a window listing attributes of the point. If more than one point will appear at the same location, more than one set of attributes will be shown.



```
Plot : Master Plot
Instance:  2
    Enroll No: 26.0
         name: piyush
Company Name : Aithent techologies
project guide: Missing
     Phone No: 9.71660565E9
    E-mail id: piyushbukaria026@gmail.com
     Platform: java
```

2. Rectangle: You can create a rectangle by dragging it around the points.



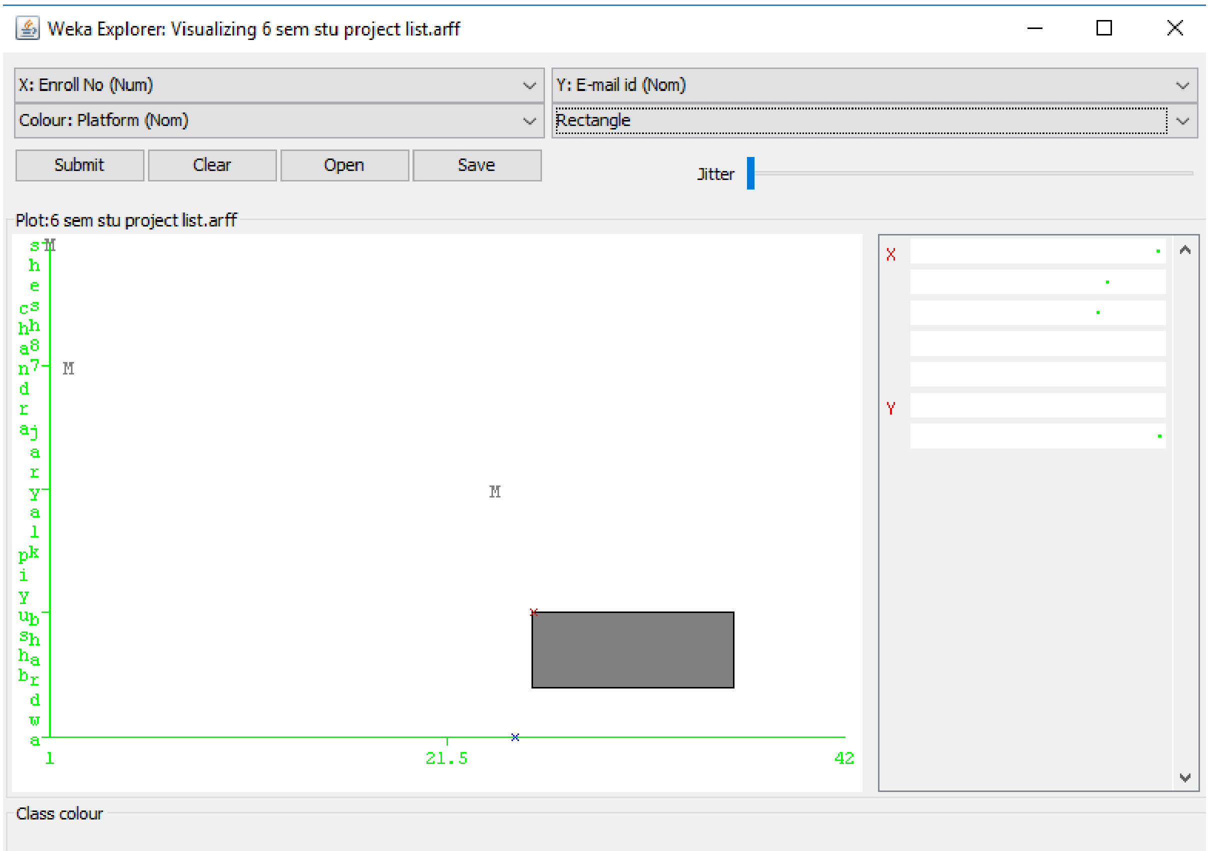
Weka Explorer: Visualizing 6 sem stu project list.arff

X: Enroll No (Num)

Y: E-mail id (Nom)

Colour: Platform (Nom)

Rectangle



Weka Explorer: Visualizing 6 sem stu project list.arff

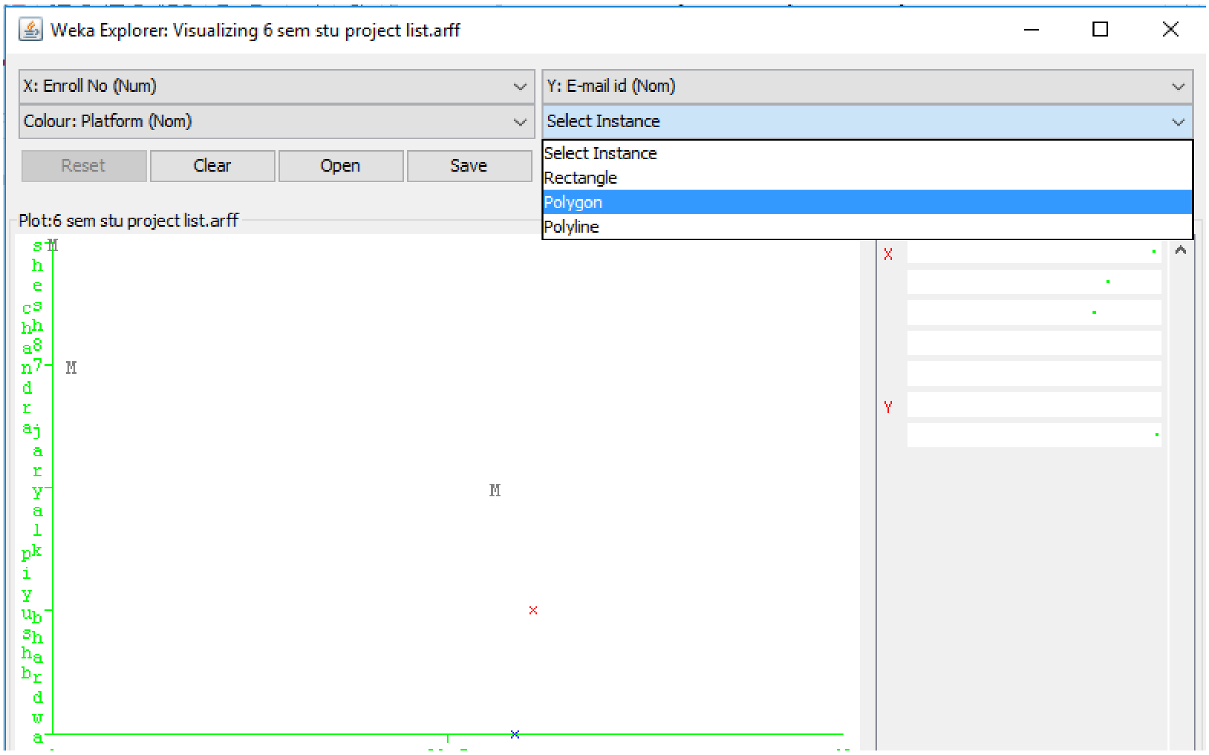
X: Enroll No (Num)

Y: E-mail id (Nom)

Colour: Platform (Nom)

Rectangle

3. Polygon: You can select several points by building a free-form polygon. Left-click on the graph to add vertices to the polygon and right-click to complete it.



Weka Explorer: Visualizing 6 sem stu project list.arff

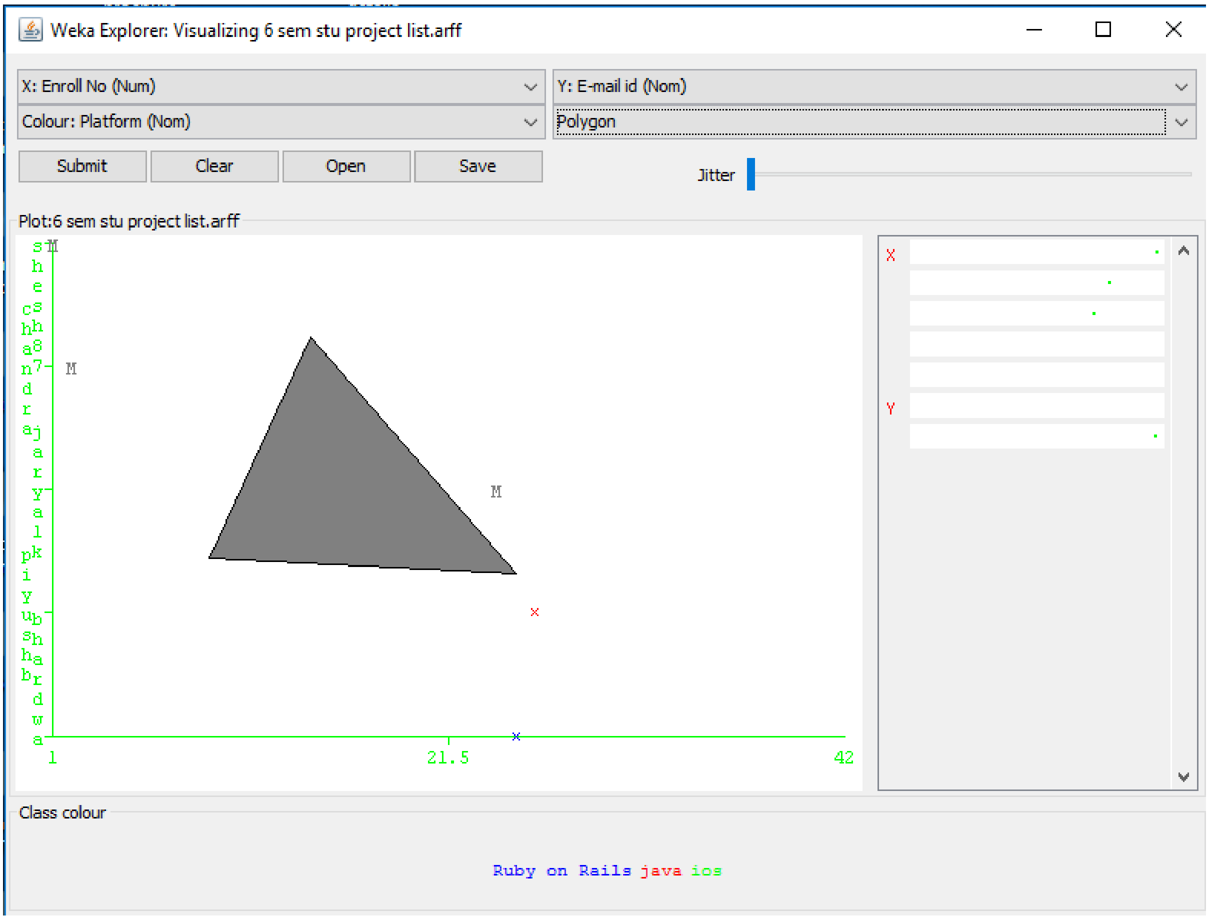
X: Enroll No (Num)

Y: E-mail id (Nom)

Colour: Platform (Nom)

Select Instance

Plot:6 sem stu project list.arff



Weka Explorer: Visualizing 6 sem stu project list.arff

X: Enroll No (Num)

Y: E-mail id (Nom)

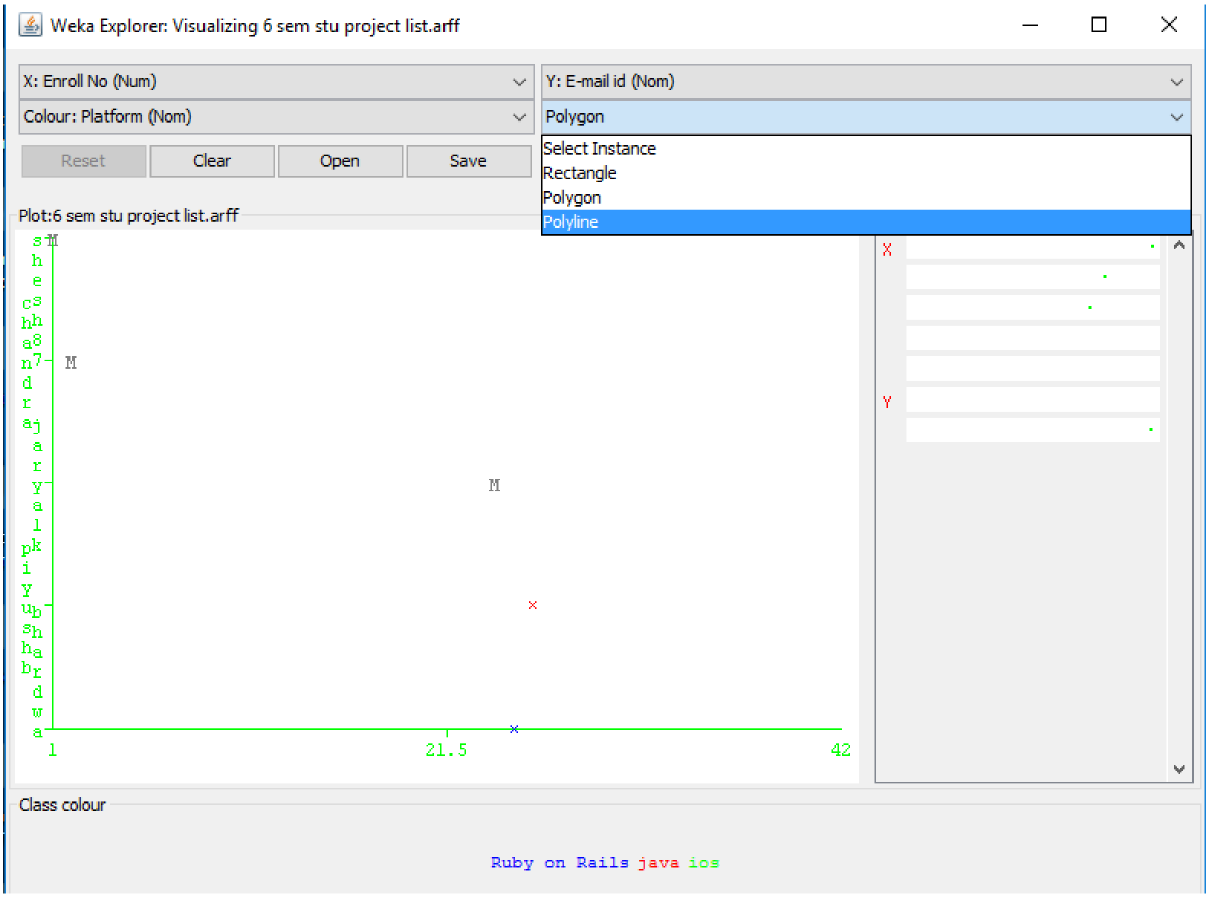
Colour: Platform (Nom)

Polygon

Plot:6 sem stu project list.arff

Class colour

4. Polyline: To distinguish the points on one side from the once on another, you can build a polyline. Left-click on the graph to add vertices to the polyline and right-click to finish.



Weka Explorer: Visualizing 6 sem stu project list.arff

X: Enroll No (Num)

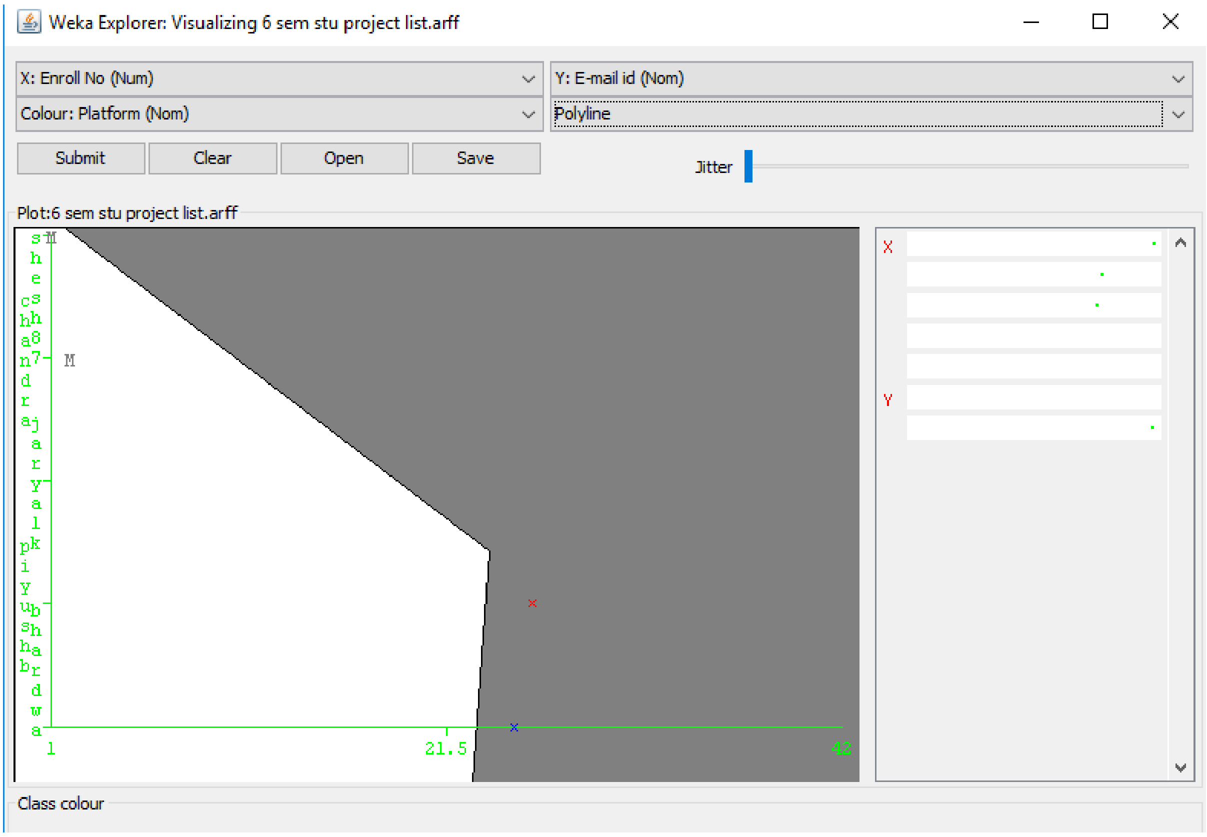
Y: E-mail id (Nom)

Colour: Platform (Nom)

Polygon

Plot:6 sem stu project list.arff

Class colour



Weka Explorer: Visualizing 6 sem stu project list.arff

X: Enroll No (Num)

Y: E-mail id (Nom)

Colour: Platform (Nom)

Polyline

Plot:6 sem stu project list.arff

Class colour

8. Conclusion:

This concludes WEKA Explorer Tutorial. You have covered a lot of material since the Tutorial Introduction. There is a lot more to learn about WEKA than what you have covered in these seven exercises. But you have already learned enough to be able to analyze your data using preprocessing, classification, clustering, and association rule tools. You have learned how to visualize the result and select attributes. This knowledge will prove invaluable to you. If you plan to do any complicated data analysis, which require software flexibility.

9. References:

1. Witten, E. Frank, Data Mining, Practical Machine Learning Tools and Techniques with Java Implementation, Morgan Kaufmann Publishers, 2000.

2. R. Kirkby, WEKA Explorer User Guide for version 3-3-4, University of Weikato, 2002. 3. Weka Machine Learning Project, http://www.cs.waikato.ac.nz/~ml/index.html.

4. E.Frank, Machine Learning With WEKA, University of Waikato, New Zealand.

5. B. Mobasher, Data Preparation and Mining with WEKA, http://maya.cs.depaul.edu/~classes/ect584/WEKA/association_rules.html, DePaul University, 2003.

6. M. H. Dunham, Data Mining, Introductory and Advanced Topics, Prentice Hall, 2002.

7. WEKA Tutorial: • Machine Learning with WEKA

8. WEKA Data Mining Book: • Ian H. Witten and Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)

9 • WEKA Wiki: http://weka.sourceforge.net/wiki/index.php/Main_Page