**DATA MINING & WAREHOUSING**

**LABORATORY**

**COURSE CODE: E2UC406B**

Lab Manual

*for*

BACHELOR OF

Engineering & Technology

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**GALGOTIAS UNIVERSITY, GREATER NOIDA**

**UTTAR PRADESH**

**SUBMITTED BY:**

**ABHINAV KUMAR CHOUDHARY (21SCSE1011615)**

**YEAR-II**

**SEM-IV**

**SEC-6**

**SUBMITTED TO:**

**Dr. GULSHAN SHRIVASTAV**

**LIST OF EXPERIMENTS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **EXP.**  **No.** | **Date** | **TITLE OF EXPERIMENTS** | **Page No.** | **Mark (10)** | **SIGN.** |
| 1 |  | Introduction to WEKA tool | 3 |  |  |
| 2 |  | To create data set in arff file format | 13 |  |  |
| 3 |  | To Create an Employee Table using WEKA | 14 |  |  |
| 4 |  | To Create a Weather Table using WEKA | 16 |  |  |
| 5 |  | To Apply Pre-Processing techniques on Weather Table | 18 |  |  |
| 6 |  | To Apply Pre-Processing techniques on Employee Table | 23 |  |  |
| 7 |  | To Normalize Weather Table data using Knowledge Flow | 28 |  |  |
| 8 |  | To Normalize Employee Table data using Knowledge Flow | 31 |  |  |
| 9 |  | Finding Association Rules for Buying data | 34 |
| 10 |  | Finding Association Rules for Banking data | 37 |  |  |
| 11 |  | Finding Association Rules for Employee data. | 40 |
| 12 |  | To Construct Decision Tree for Weather data | 43 |  |  |
| 13 |  | To Construct Decision Tree for Customer data | 46 |  |  |
| 14 |  | To Construct Decision Tree for Location data | 49 |  |  |
| 15 |  | Implementation of cross-validation for the weather table  using the J48 algorithm | 52 |  |  |
|  |  | **TOTAL** | |  |  |
|  |  | **AVG** | |  |  |



# Opening the program

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. Figure 1 is an example of the initial opening screen on a computer with Windows XP.

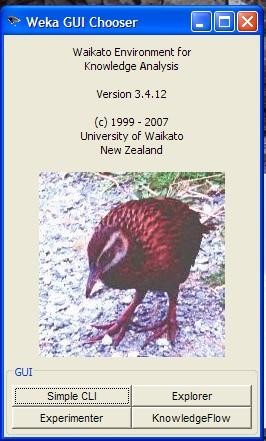


Figure 1 Chooser There are four options available on this initial screen.

* Simple CLI- provides users without a graphic interface option the ability to execute commands from a terminal window.
* Explorer- the graphical interface used to conduct experimentation on raw data
* Experimenter- this option allows users to conduct different experimental variations on data sets and perform statistical manipulation
* 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

While the options available can be useful for different applications the remaining focus of the user manual will be on the Experimenter option through the rest of the user guide.

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

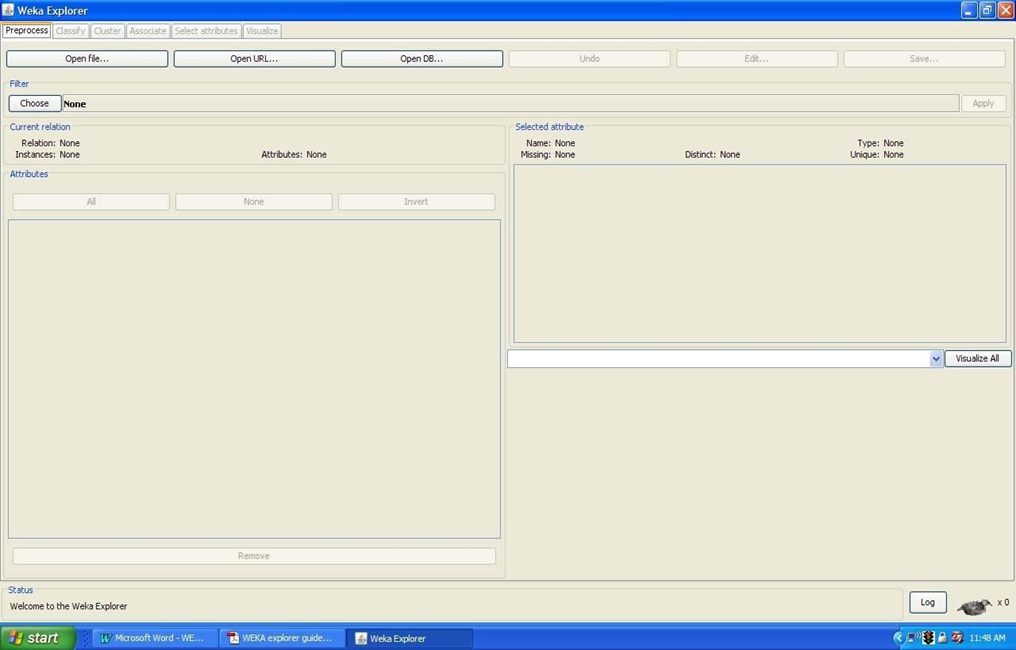
****

Figure 2

Figure 2 shows the opening screen with the available options. At first there is only the option to select the Preprocess 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 preprocessed the other tabs become active for use.

There are six tabs:

1. Preprocess- used to choose the data file to be used by the application
2. Classify- used to test and train different learning schemes on the preprocessed 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

Once the initial preprocessing of the data set has been completed the user can move between the tab options to perform changes to the experiment and view the results in real time. This provides the benefit of having the ability to move from one option to the next so that when a condition becomes exposed it can be placed in a different environment to be visually changed instantaneously.

# Preprocessing

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, newer 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. Figure 3 shows an example of selection of the file weather.arff.

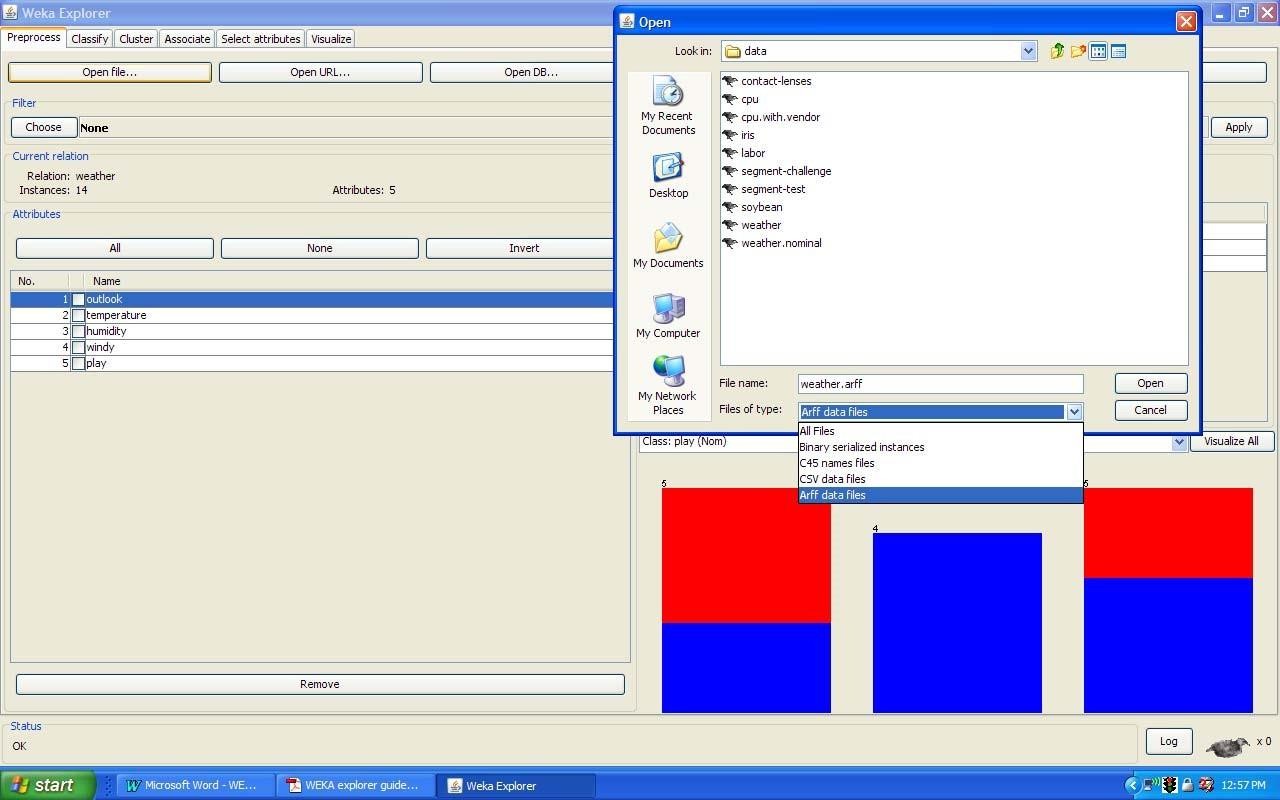


Figure 3

Once the initial data has been selected and loaded the user can select options for refining the experimental data. The options in the preprocess window include selection of optional filters to apply and the user can select or remove different attributes of the data set as necessary to identify specific information. The ability to pick from the available attributes allows users to separate different parts of the data set for clarity in the experimentation. The user can modify the attribute selection and change the relationship among the different attributes by deselecting different choices from the original data set. There are many different filtering options available within the preprocessing window and the user can select the different options based on need and type of data present.

# Classify

The user has the option of applying many different algorithms to the data set that would in theory produce a representation of the information used to make observation easier. It is difficult to identify which of the options would provide the best output for the experiment. The best approach is to independently apply a mixture of the available choices and see what yields something close to the desired results. The Classify tab is where the user

selects the classifier choices. Figure 4 shows some of the categories.

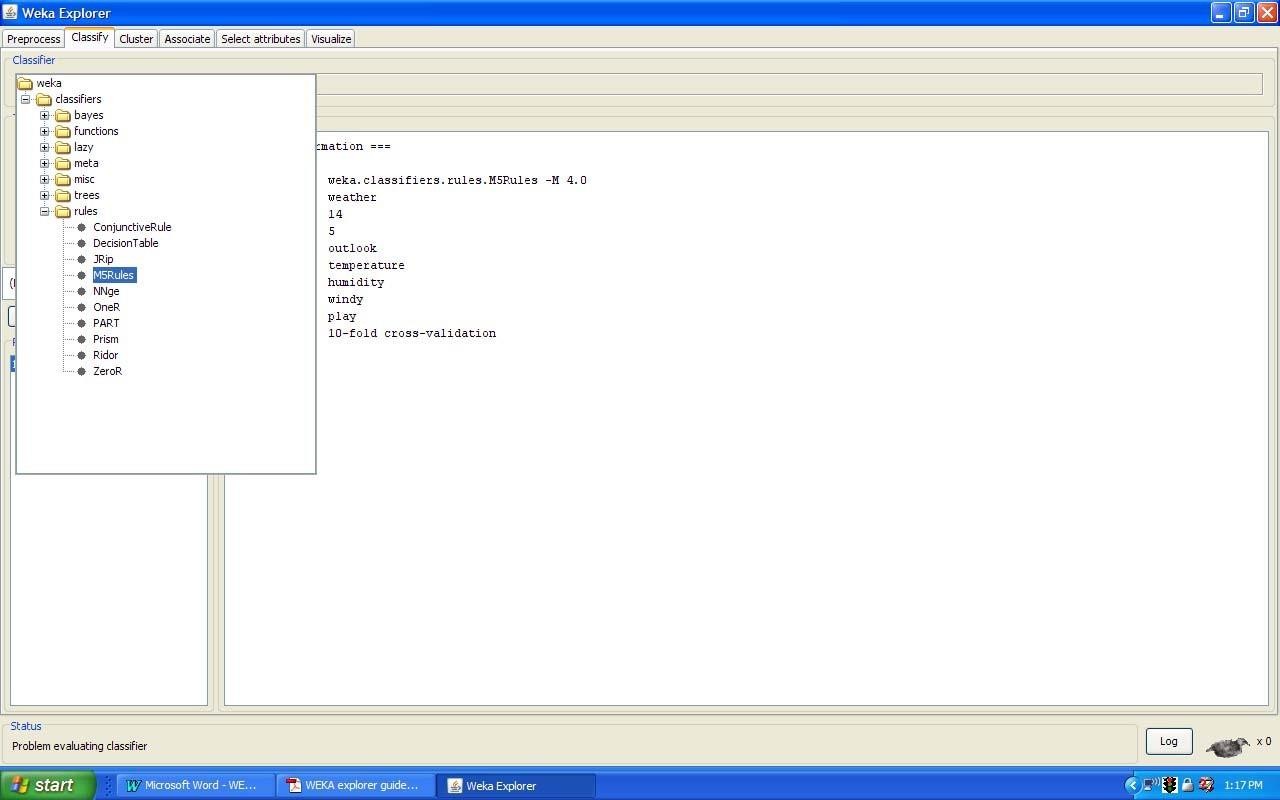


Figure 4

Again there are several options to be selected inside of the classify tab. Test option gives the user the choice of using four different test mode scenarios on the data set:

1. Use training set
2. Supplied training set
3. Cross validation
4. Split percentage

There is the option of applying any or all of the modes to produce results that can be compared by the user. Additionally inside the test options toolbox there is a dropdown menu so the user can select various items to apply that depending on the choice can provide output options such as saving the results to file or specifying the random seed value to be applied for the classification.

The classifiers in WEKA have been developed to train the data set to produce output that has been classified based on the characteristics of the last attribute in the data set. For a specific attribute to be used the option must be selected by the user in the options menu before testing is performed. Finally the results have been calculated and they are shown in the text box on the lower right. They can be saved in a file and later retrieved for comparison at a later time or viewed within the window after changes and different results have been derived.

# Cluster

The Cluster tab opens the process that is used to identify commonalties or clusters of occurrences within the data set and produce information for the user to analyze. There are a few options within the cluster window that are similar to those described in the classifier tab. They are use training set, supplied test set, percentage split. The fourth option is classes to cluster evaluation, which compares how well the data compares with a pre-assigned class within the data. While in cluster mode users have the option of ignoring some of the attributes from the data set.

This can be useful if there are specific attributes causing the results to be out of range or for large data sets. Figure 5 shows the Cluster window and some of its options.

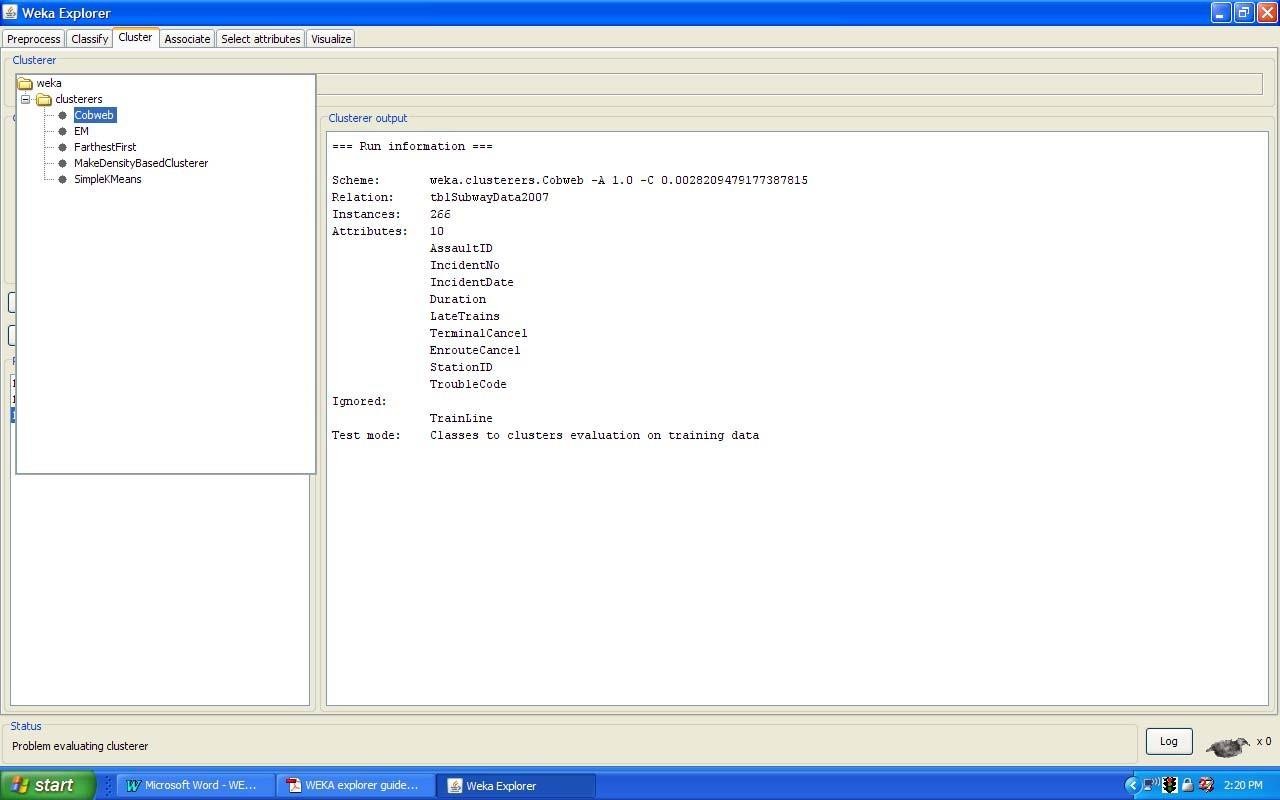


Figure 5

# Associate

The associate tab opens a window to select the options for associations within the data set. The user selects one of the choices and presses start to yield the results. There are few options for this window and they are shown in Figure 6 below.

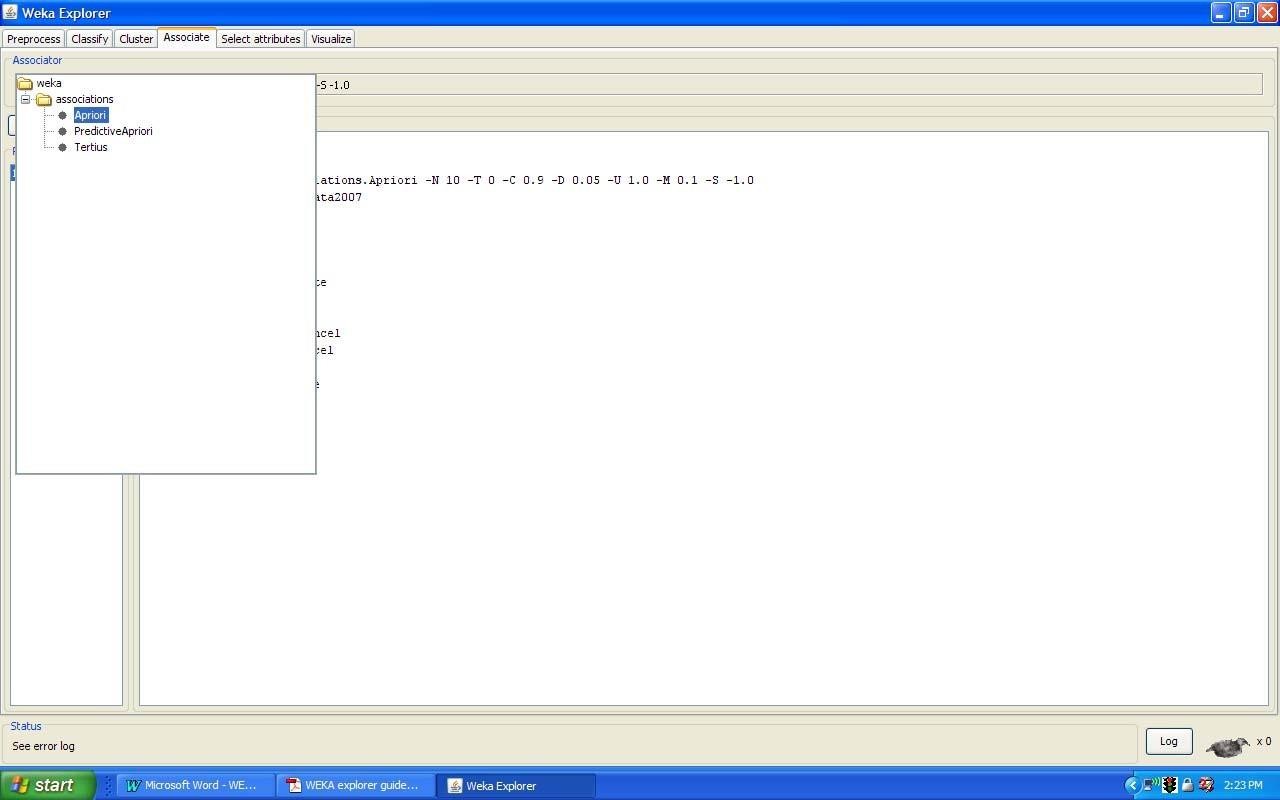


Figure 6

# Select Attributes

The next tab is used to select the specific attributes used for the calculation process. By default all of the available attributes are used in the evaluation of the data set. If the use wanted to exclude certain categories of the data they would deselect those specific choices from the list in the cluster window. This is useful if some of the attributes are of a different form such as alphanumeric data that could alter the results. The software searches through the selected attributes to decide which of them will best fit the desired calculation. To perform this, the user has to select two options, an attribute evaluator and a search method. Once this is done the program evaluates the data based on the sub set of the attributes then

performs the necessary search for commonality with the date. Figure 7 shows the opinions of attribute evaluation. Figure 8 shows the options for the search method.

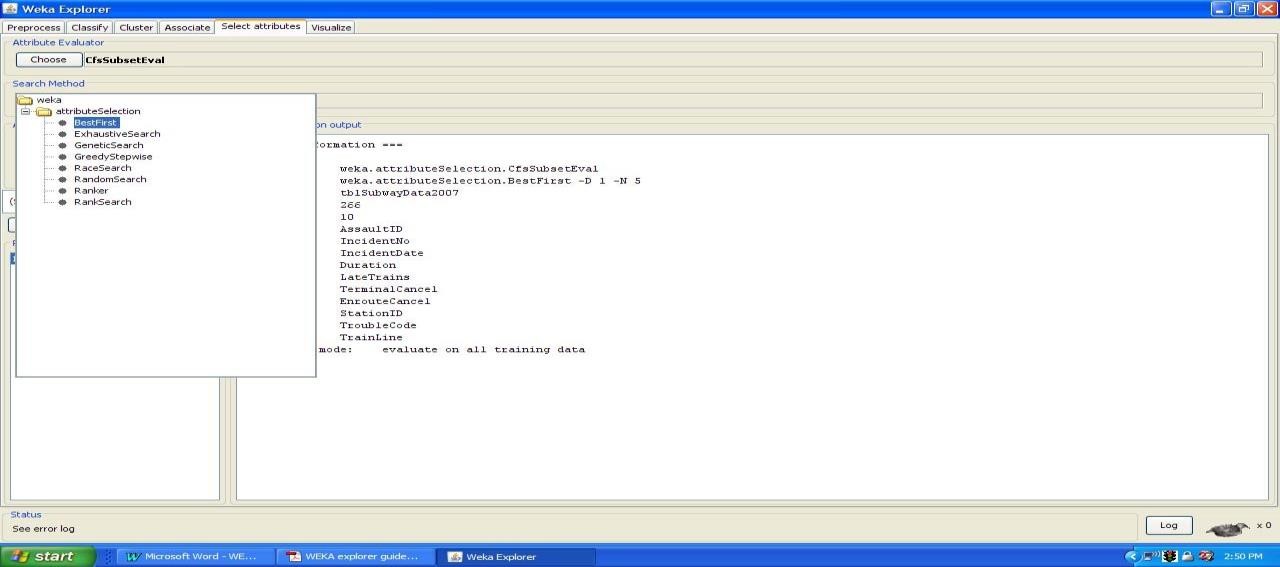


Figure 7

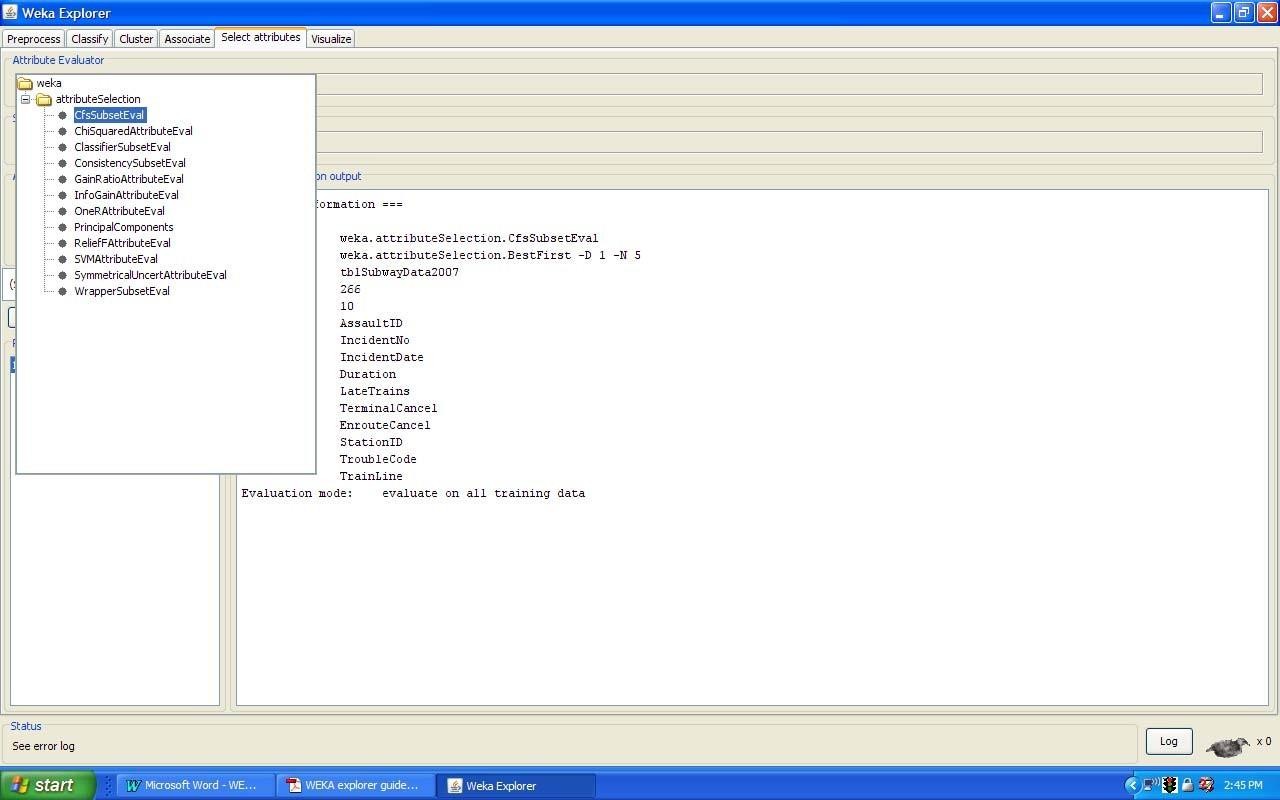


Figure 8

**Visualization**

The last tab in the window is the visualization tab. Within the program calculations and comparisons have occurred on the data set. Selections of attributes and methods of manipulation have been chosen. The final piece of the puzzle is looking at the information that has been derived throughout the process. The user can now actually see the fruit of their efforts in a two-dimensional representation of the information. The first screen that the user sees when they select the visualization option is a matrix of plots representing the different attributes within the data set plotted against the other attributes. If necessary, there is a scroll bar to view all of the produced plots. The user can select a specific plot from the matrix to view its contents for analyzation. A grid pattern of the plots allows the user to select the attribute positioning to their liking and for better understanding. Once a specific plot has been selected the user can change the attributes from one view to another providing flexibility. Figure 9 shows the plot matrix view.

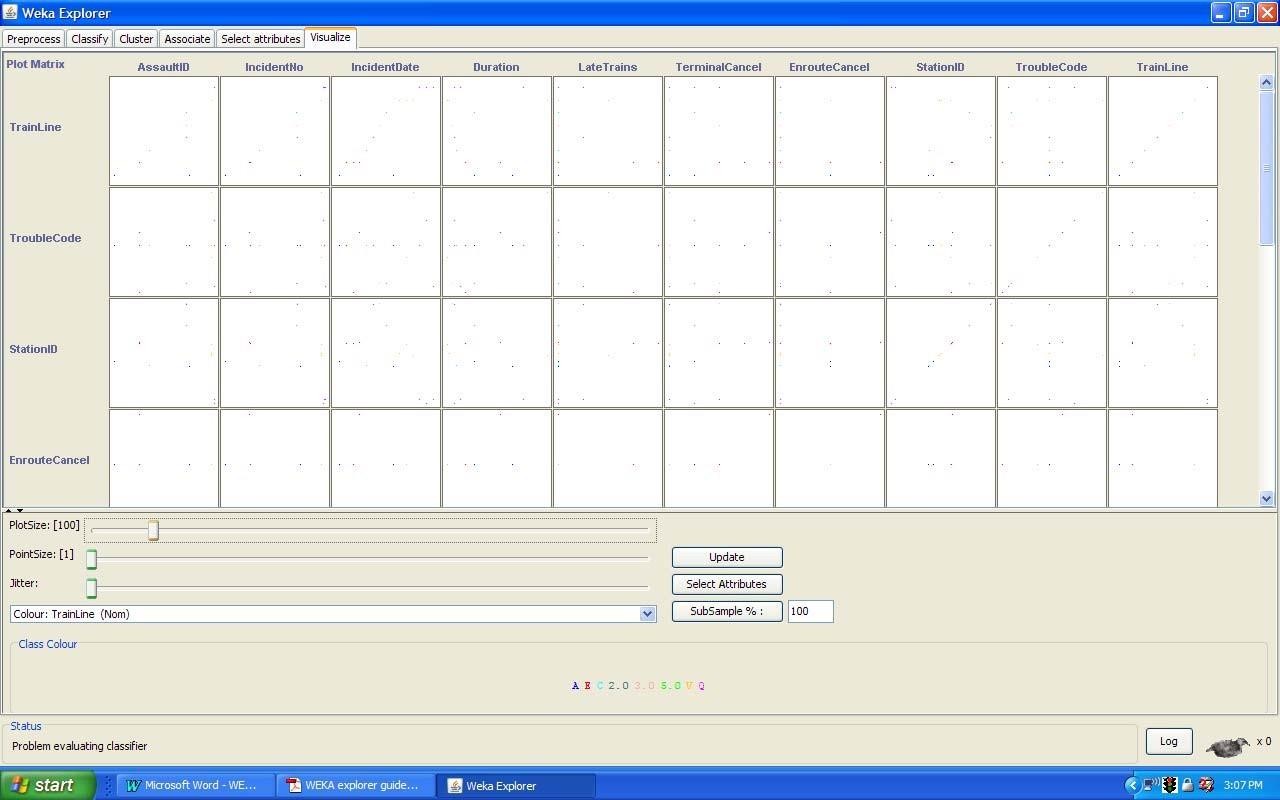


Figure 9

The scatter plot matrix gives the user a visual representation of the manipulated data sets for selection and analysis. The choices are the attributes across the top and the same from top to bottom giving the user easy access to pick the area of interest. Clicking on a plot brings up a separate window of the selected scatter plot. The user can then look at a visualization of the data of the attributes selected and select areas of the scatter plot with a selection window or by clicking on the points within the plot to identify the point’s specific information. Figure 10 shows the scatter plot for two attributes and the points derived from the data set. There are a few options to view the plot that could be helpful to the user. It is formatted similar to an X/Y graph yet it can show any of the attribute classes that appear on the main scatter plot matrix. This is handy when the scale of the attribute is unable to be ascertained in one axis over the other. Within the plot the points can be adjusted by utilizing a feature called jitter. This option moves the individual points so that in the event of close data points users can reveal hidden multiple occurrences within the initial plot. Figure 11 shows an example of this point selection and the results the user sees.

# Result:

This program has been successfully executed.

**Experiment-2**

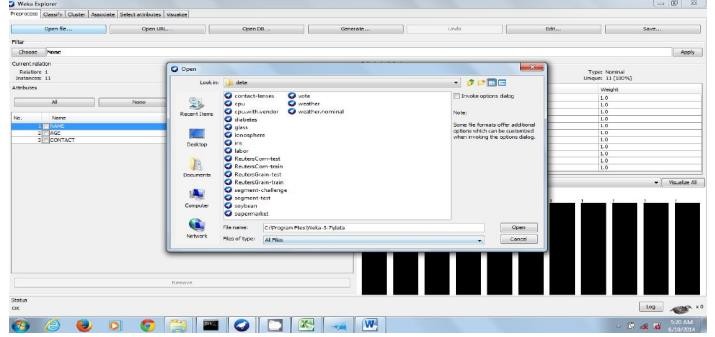
**Aim:** To create data set in arff file format

Procedure: Open MS Excel.

**Steps :**

Create a new worksheet with respective headings and data. Save the file with .csv extension

Open Preprocessor tab in explorer

Click open the file button and browse the file to open. Load the desired .csv file using open File tab.

Click SAVE shown… dialog box opens save with extension as .arff

# Result:

This program has been successfully executed.

**EXPERIMENT NO: 3**

**Aim:**

To Create an Employee Table using WEKA

# Description:

We need to create an Employee Table with training data set which includes attributes like name, id, salary, experience, gender, phone number.

# Procedure:

**Steps:**

1. Open Start  Programs  Accessories  Notepad
2. Type the following training data set with the help of Notepad for Employee Table.

@relation employee @attribute name {x,y,z,a,b} @attribute id numeric

@attribute salary {low,medium,high} @attribute exp numeric

@attribute gender {male,female} @attribute phone numeric

@data

x,101,low,2,male,250311 y,102,high,3,female,251665 z,103,medium,1,male,240238 a,104,low,5,female,200200 b,105,high,2,male,240240

1. After that the file is saved with **.arff** file format.
2. Minimize the arff file and then open Start  Programs  weka-3-4.
3. Click on **weka-3-4**, then Weka dialog box is displayed on the screen.
4. In that dialog box there are four modes, click on **explorer**.
5. Explorer shows many options. In that click on **‘open file’** and select the arff file
6. Click on **edit button** which shows employee table on weka.