Dr. M.G.R.
EDUCATIONAL AND RESEARCH INSTITUTE
DEEMED TO BE UNIVERSITY
University with Graded Autonomy Status
(An ISO 21001 : 2018 Certified Institution)
Periyar E.V.R. High Road, Maduravoyal, Chennai-95. Tamilnadu, India.

RECORD NOTEBOOK

BCS18L11 –DATA MINING LAB

2024–2025 (ODD SEMESTER)

DEPARTMENT

OF

COMPUTER SCIENCE AND ENGINEERING

NAME	:	SUMIT KUMAR
REGISTER NO	:	211061101448
COURSE	:	B.TECH (CSE)
YEAR/SEM/SEC	:	IV / VII/ H

BONAFIDE CERTIFICATE

Register No: 211061101448

Name of Lab: DATA MINING LAB

Department: COMPUTER SCIENCE AND ENGINEERING

Certified that, this Record note book is a **bonafide** record of work done by **SUMIT KUMAR** of **IV Year B. Tech / CSE, Sec - 'H'** in the **DATA MINING LAB** during the year **2024-2025.**

Signature of Lab-in-Charge   **Signature of Head of Dept**

Submitted for the Practical Examination held on _____

Internal Examiner   **External Examiner**

INDEX

Exp No: 1   **Date:**

INTRODUCTION TO THE WEKA MACHINE LEARNING TOOL KIT

AIM

To study WEKA tool kit.

THEORY

WEKA is open source java code created by researchers at the University of Waika to in New Zealand. It provides many different machine learning algorithms, including the following classifiers:

- Decision tree(j4.8, an extension of C4.5)
- MLP, aka multiple layer perceptron (a type of neural net)
- Naïve bayes
- Rule induction algorithm ms such as JRip
- Support vector machine
- And many more…

The GUI WEKA Chooser (class weka.gui.GUI Chooser) provides a starting point for launching Weka's main GUI applications and supporting tools. If one prefers a MDI ("multiple document interface") appearance, then this is provided by an alternative launcher called "Main" (class weka.gui.Main). The GUI Chooser consists of four buttons—one for each of the four major WEKA applications - and four menus

The button scan be used to start the following applications:

- **Explorer:** An environment for exploring data with WEKA (the rest of this documentation deals with this application in more detail).
- **Experimenter:** An environment for performing experiments and conducting statistical tests between learning schemes.
- **Knowledge Flow:** This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.

- **Simple CLI:** Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

WEKA Explorer

The user interface Section Tabs

At the very top of the window, just below the title bar, is a row of tabs. When the Explorer is first started only the first tab is active; the others are grayed out. This is because it is necessary to open (and potentially pre-process) a data set before starting to explore the data.

The tabs are as follows:

a) **Pre-process:** Choose and modify the data being acted on.
b) **Classify:** Train and test learning schemes that classify or perform regression.
c) **Cluster:** Learn clusters for the data.
d) **Associate:** Learn association rules for the data.
e) **Select attributes:** Select the most relevant attributes in the data.
f) **Visualize:** View an interactive 2D plot of the data.

Once the tabs are active, clicking on them flicks between different screens, on which the respective action scan be performed. The bottom are of the window (including the status box, the log button, and the Weka bird) stays visible regardless of which section you are in. The Explorer can be easily extended with custom tabs.

The menu consists of four sections

1. Pre-processing
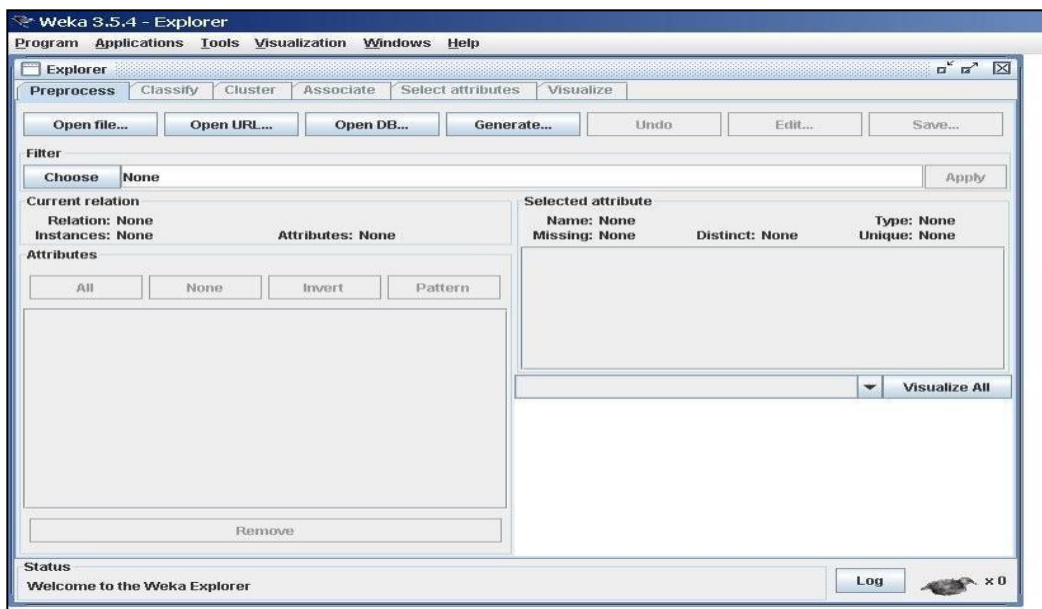
OPTION

i   *All:* All boxes are ticked.
ii *None:* All boxes are cleared (unticked).
iii *Invert:* Boxes that are ticked become unticked and vice versa.
iv *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. Once the desired attributes have been selected, they can be removed by clicking the Remove button below the list of attributes. Note that this can be undone by clicking the Undo button,which is located next to the Edit button in the top-right corner of the Pre-process panel.

LOADING DATA

The first four buttons at the top of the pre-process section enable you to load data into WEKA:

1) **Open file:** Brings up a dialog box allowing you to browse for the data file on the local file system.
2) **Open URL:** Asks for a Uniform Resource Locator address for where the data is stored.
3) **Open DB:** Reads data from a database. (Note that to make this work you might have to edit the file in weka/experiment/Database Utils.props.)
4) **Generate:** Enables you to generate artificial data from a variety of Data Generators. Using the Open file.button you can read files in a variety of formats: WEKA's ARFF format, CSV format, C4.5 format, or serialized Instances format.ARFF files typically have a .arff extension, CSV files a .csv extension, C4.5 files a.data and .names extension and serialized Instances objects a.bsi extension.

Working with Filters
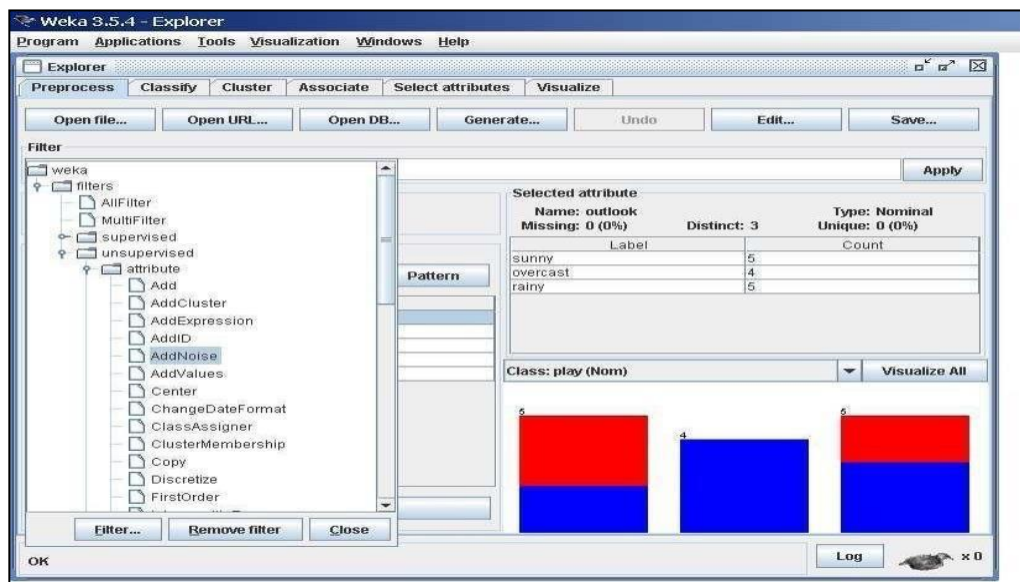
The pre-process section allows filters to be defined that transform the data in various ways. The Filter box is used to set up the filters that are required. At the left of the Filter box is a Choose button. By clicking this button it is possible to select one of the filters in WEKA. Once a filter has been selected, its name and options are shown in the field next to the Choose button. Clicking on this box with the

leftmousebuttonbringsupaGenericObjectEditordialogbox.A click with the right mouse button(or Alt+Shift+left click) brings up a menu where you can choose, either to display the properties in a Generic Object Editor dialog box, or to copy the current setup string to the clipboard.

2) Classification
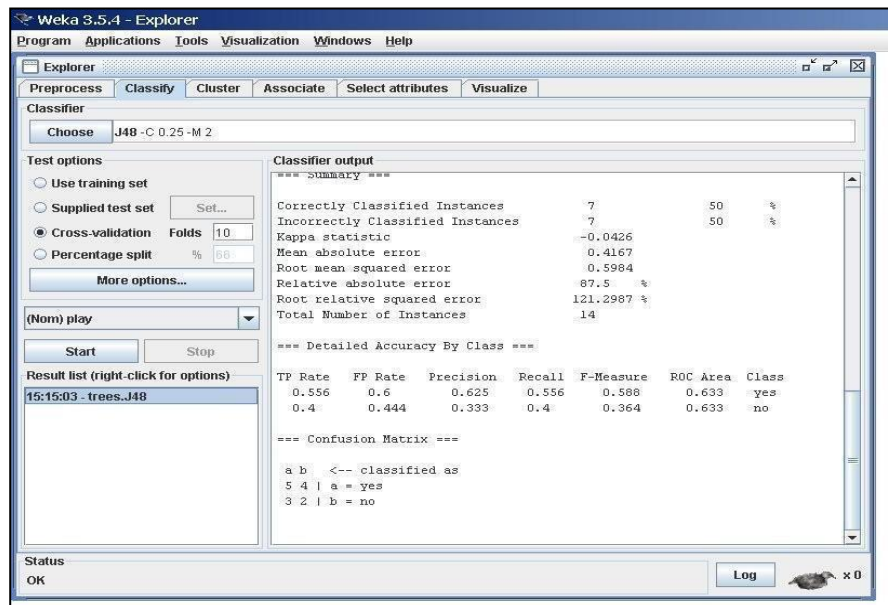
Selecting a Classifier

At the top of the classify section is the Classifier box. This box has a text field that gives the name of the currently selected classifier, and its options. Clicking on the text box with the left mouse button brings up a Generic Object Editor dialog box, just the same as for filters that you can use to configure the options of the current classifier. With a right click (or Alt+Shift+left click) you can once again copy the setup string to the clipboard or display the properties in a Generic Object Editor dialog box. The Choose button allows you to choose one of the classifiers that are available in WEKA.

3) Clustering
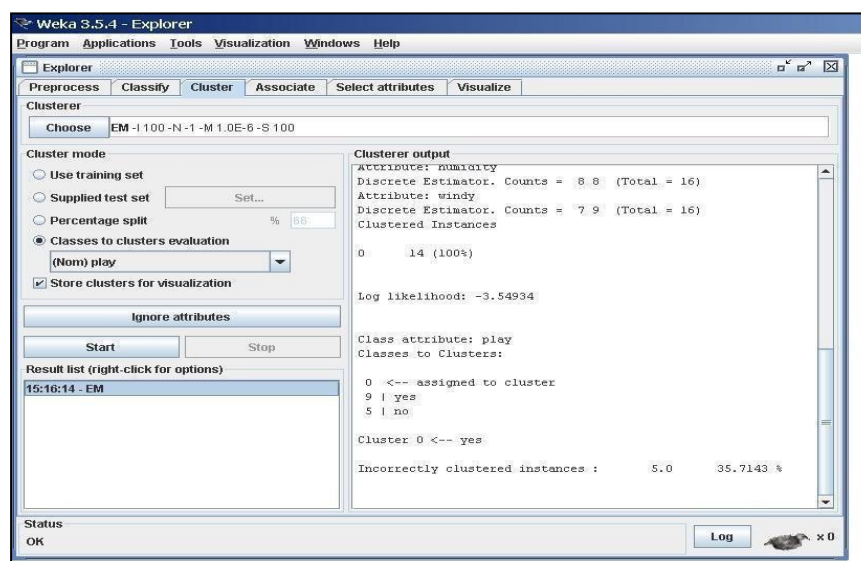
By now you will be familiar with the process of selecting and configuring objects. Clicking on the clustering scheme listed in the Cluster box at the top of the window brings up a Generic Object Editor dialog with which to choose a new clustering scheme.

Cluster Mode

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 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 on training is complete. When dealing with datasets that are so large that memory becomes a problem it may be helpful to disable this option.

4) Associating

Setting Up

This panel contains schemes for learning association rules, and the learners are chosen and configured in the same way as the clusters, filters, and classifiers in the other panels.

Learning Associations

Once appropriate parameters for the association rule learner have been set, click the Start button. When complete, right-clicking on an entry in the result list allows the results to be viewed or saved.

Selecting Attributes

Searching and Evaluating

Attributes election involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction. To do this, two objects must be setup: an attribute evaluator and a search method. The evaluator determines what method is used to assign a worth to each subset of attributes. The search method determines what style of search is performed.
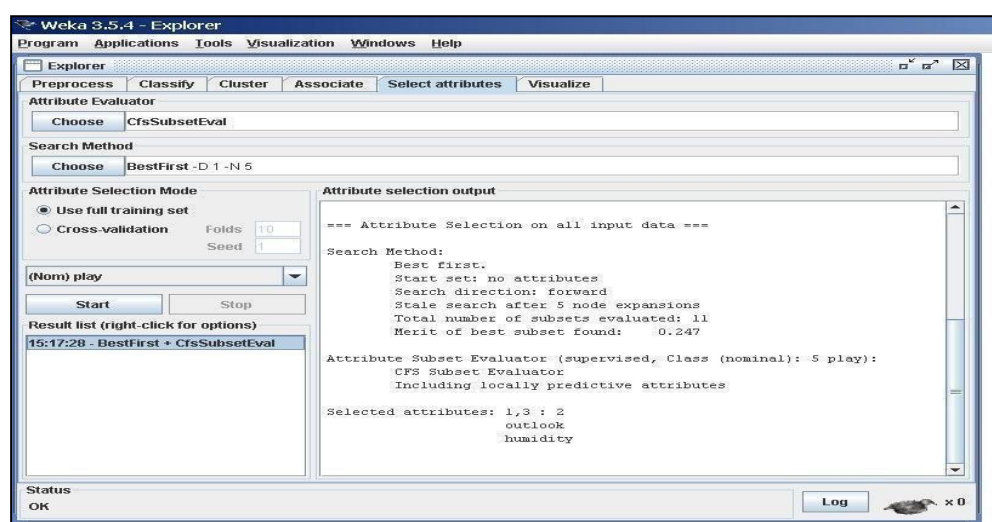
OPTIONS

The Attribute Selection Mode box has two options:

1. Useful training set: This 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. As with Classify, there is a drop-down box that can be used to specify which attribute to treat as the class.
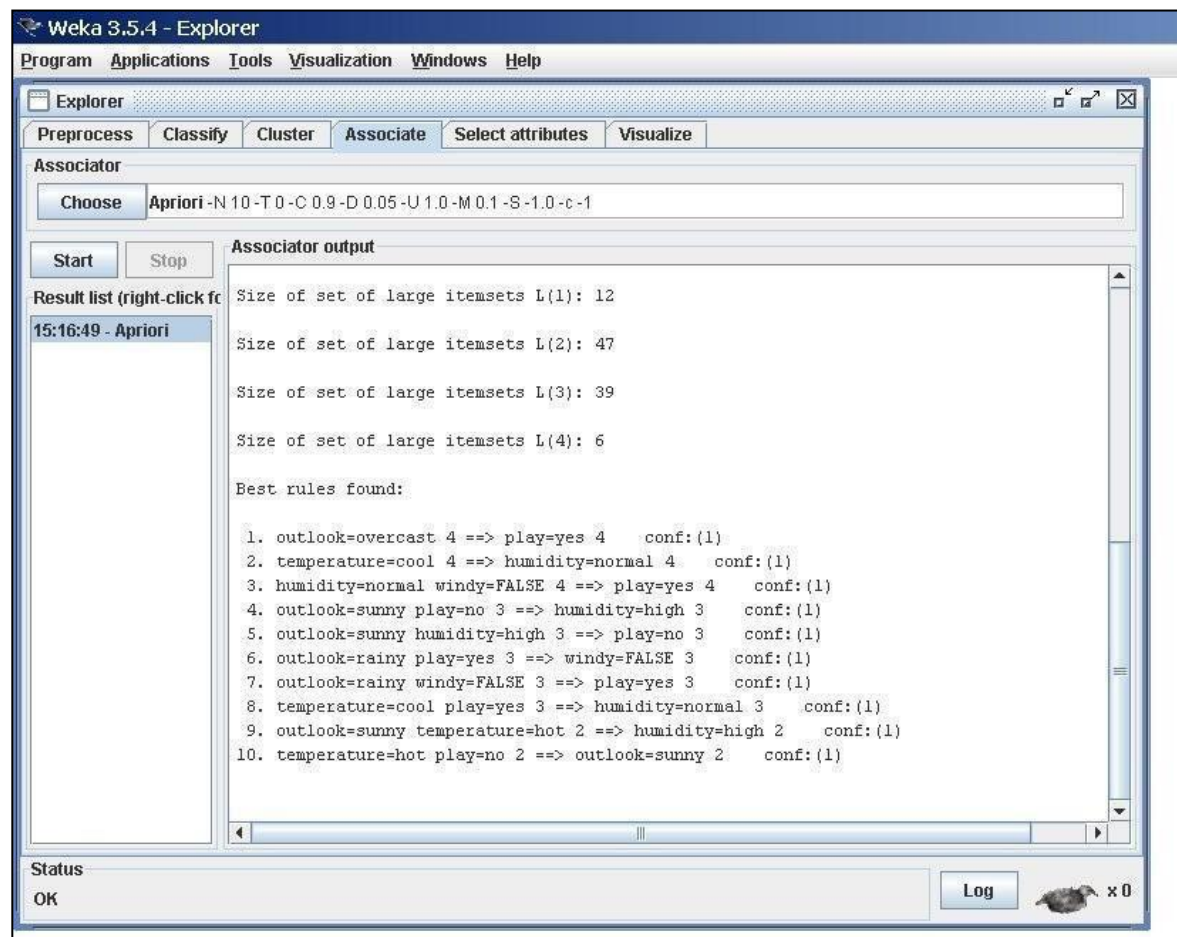
RESULT

Therefore, the WEKA tool kit is studied and output is verified successfully.

CLASSIFICATION INTRODUCTION TO EXPLORATORY DATA ANALYSIS USING RATTLE AN OPEN SOURCE TOOL(R)

AIM

To Conduct Exploratory Data Analysis Using Rattle.

THEORY

Summary

The Summary check box provides numerous measures for each variable, including, in the first instance, the minimum, maximum, median, mean, and the first and third quartiles.Generally,if the mean and median are significantly different then we would think that there are some entities with very large values in the data pulling theme an in one direction. It does not seem to be the case for Age but is for Income.

    i   Summary check box

ii   Describe check box

```
Rattle: Effective Data Mining with R: audit.csv
Project  Edit  Tools  Settings  Help                    ⓘ Rattle Version 2.2.26 togaware.com
  Execute   New   Open   Save   Export   Quit    ⦿ Two Class  ○ Unsupervised  ○ Time Series
                                                 ○ Multi Class  ○ Regression   ○ Text Miner
Data | Select | Explore | Transform | Model | Evaluate | Log |
Type: ⦿ Summary  ○ Distributions  ○ GGobi  ○ Correlation  ○ Hierarchical  ○ Principal Components
□ Use Sample  ☑ Summary  □ Describe  □ Basics  □ Kurtosis  □ Skewness  □ Show Missing

Summary of the full dataset.

The data contains 141 entities with missing values.

(Hint: 25% of values are below 1st Quartile.)

     Age            Employment          Education            Marital
Min.   :17.00   Private   :1411   HSgrad   :660   Absent              :669
1st Qu.:28.00   Consultant: 148   College  :442   Divorced            :266
Median :37.00   PSLocal   : 119   Bachelor :345   Married             :917
Mean   :38.62   SelfEmp   :  79   Master   :102   Married-spouse-absent: 22
3rd Qu.:48.00   PSState   :  72   Vocational: 86  Unmarried           : 67
Max.   :90.00   (Other)   :  71   Yr11     : 74   Widowed             : 59
                NA's      : 100   (Other)  :291
     Occupation         Income              Sex             Deductions
Executive    :289   Min.   :   146.1   Female: 632   Min.   :   0.00
Professional:247    1st Qu.: 35079.4   Male  :1368   1st Qu.:   0.00
Clerical     :232   Median : 57885.7                 Median :   0.00
Repair       :225   Mean   : 85896.3                 Mean   :  67.57
Service      :210   3rd Qu.:119643.5                 3rd Qu.:   0.00
(Other)      :696   Max.   :421362.7                 Max.   :2904.00
NA's         :101
     Hours             Accounts         Adjustment          Adjusted
Min.   : 1.00   UnitedStates:1804   Min.   : -2194   Min.   :0.0000
1st Qu.:38.00   Mexico      :  43   1st Qu.:     0   1st Qu.:0.0000
Median :40.00   Philippines :  13   Median :     0   Median :0.0000
Mean   :40.07   Vietnam     :   8   Mean   :  2006   Mean   :0.2315
3rd Qu.:45.00   China       :   7   3rd Qu.:     0   3rd Qu.:0.0000
Max.   :99.00   (Other)     :  82   Max.   :102937   Max.   :1.0000
                NA's        :  43

Generated by Rattle 2007-03-22 21:22:23 gjw
================================================================
Find:              🔍 Find  ⋈ Next
Data summary generated.
```
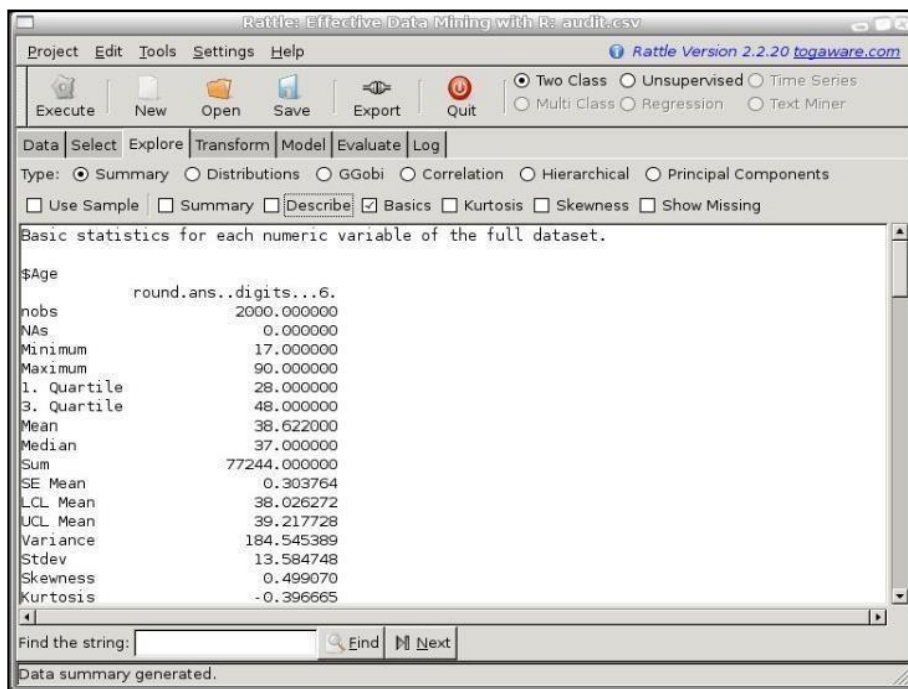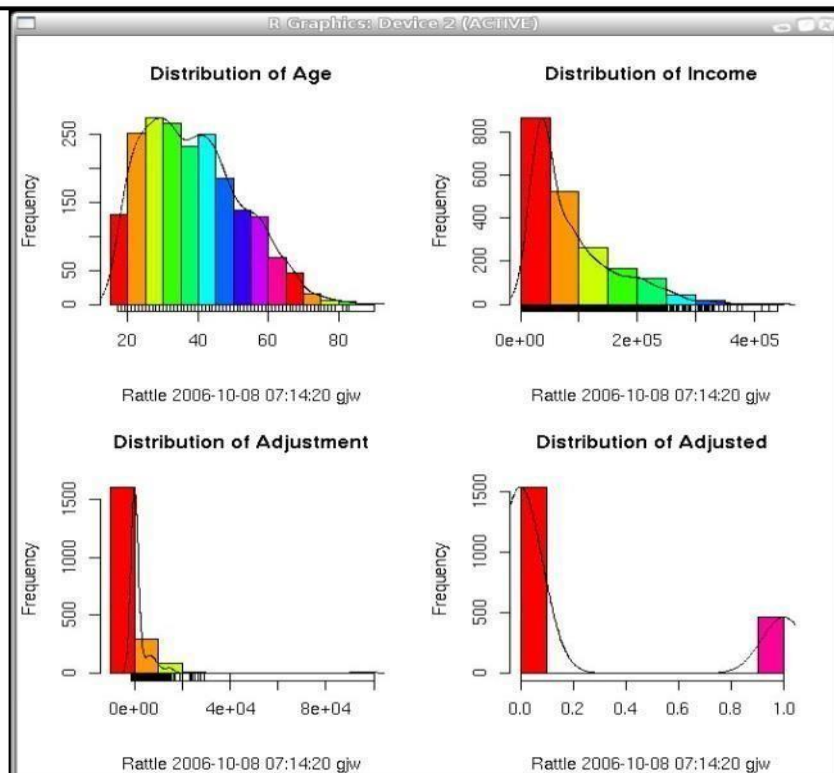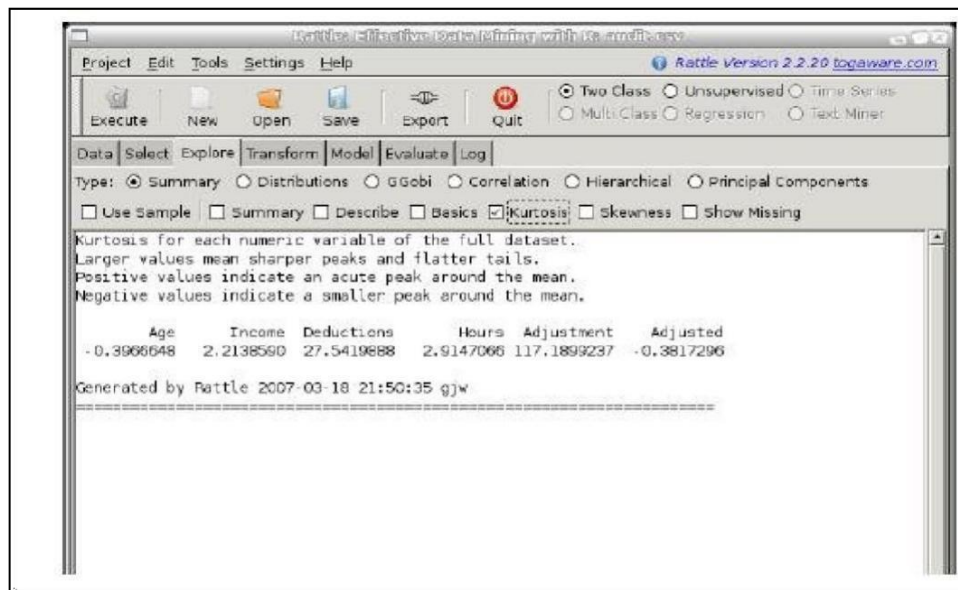
iii   Basics check box

```
Rattle: Effective Data Mining with R: audit.csv
Project  Edit  Tools  Settings  Help                    ⓘ Rattle Version 2.2.20 togaware.com
  Execute   New   Open   Save   Export   Quit    ⦿ Two Class  ○ Unsupervised  ○ Time Series
                                                 ○ Multi Class  ○ Regression   ○ Text Miner
Data | Select | Explore | Transform | Model | Evaluate | Log |
Type: ⦿ Summary  ○ Distributions  ○ GGobi  ○ Correlation  ○ Hierarchical  ○ Principal Components
□ Use Sample  □ Summary  □ Describe  ☑ Basics  □ Kurtosis  □ Skewness  □ Show Missing

Basic statistics for each numeric variable of the full dataset.

$Age
              round.ans..digits...6.
nobs              2000.000000
NAs                  0.000000
Minimum             17.000000
Maximum             90.000000
1. Quartile         28.000000
3. Quartile         48.000000
Mean                38.622000
Median              37.000000
Sum              77244.000000
SE Mean              0.303764
LCL Mean            38.026272
UCL Mean            39.217728
Variance           184.545389
Stdev               13.584748
Skewness             0.499070
Kurtosis            -0.396665

Find the string:            🔍 Find  ⋈ Next
Data summary generated.
```

iv   Kurtosis

The kurtosis is a measure of the nature of the peaks in the distribution of the data. A larger value for the kurtosis will indicate that the distribution has a sharper peak, as we can see in comparing the distributions of Income and Adjustment. A lower kurtosis indicates a smoother peak.

v   Skewness

The skewness is a measure of how a symmetric a lour data is distributed. A positive skew indicates that the tail to the right is longer, and a negative skew that the tail to the left is longer.
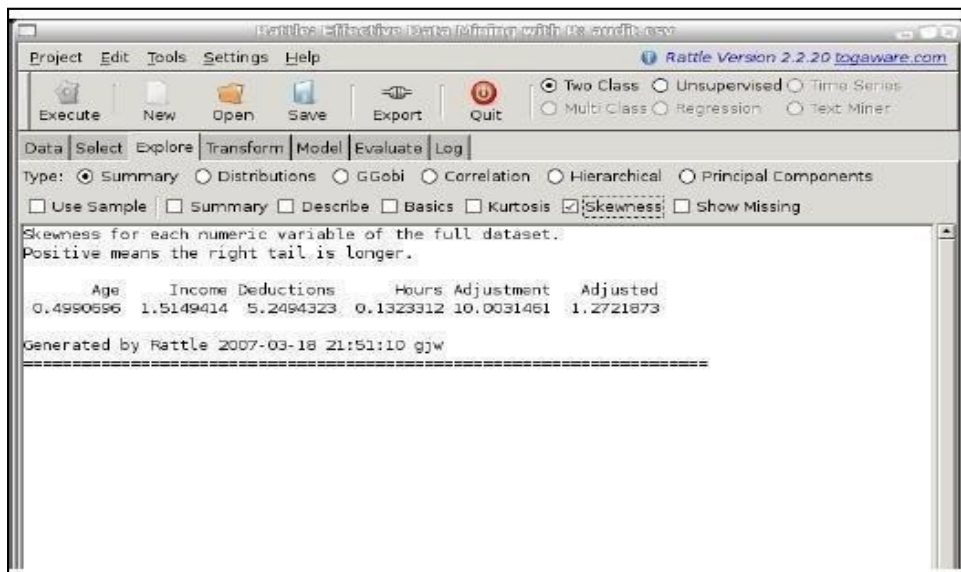
vi   Missing

Missing values present challenges to data mining. The Show Missing check button of the Summary option of the Explore tab provides a summary of missing values in our dataset. Following figure illustrates the missing value summary. Such information is useful in understanding structure in the missing values.
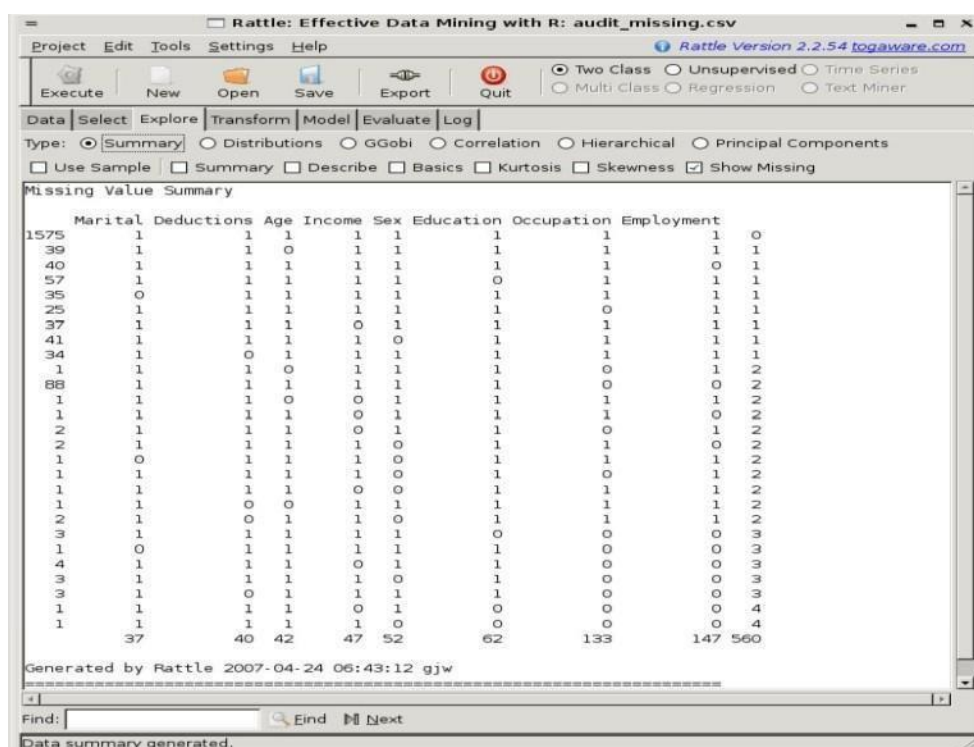
The missing value summary table is presented with the variables listed along the top. Each row corresponds to a pattern of missing values. A1 indicates a value is present, where as a 0 indicates a value is missing.

RESULT

Classification introduction to exploratory data analysis using rattle an open source tool (r) is executed and results verified.

INTRODUCTION TO REGRESSION USING RATTLE AN OPEN SOURCE TOOL (R)

AIM

To perform Correlation and linear regression analysis that are the most commonly used techniques for investigating the relationship between two quantitative variables.

THEORY

CORRELATION ANALYSIS
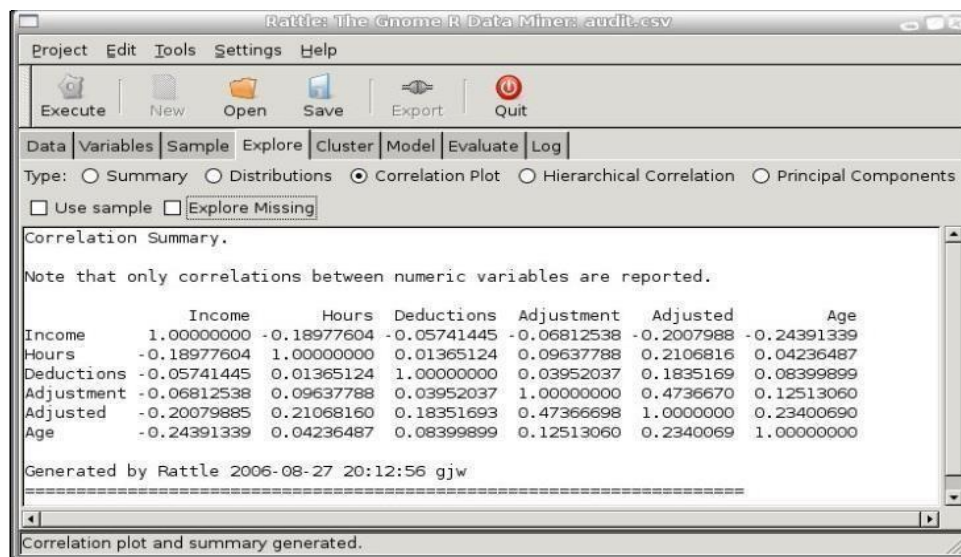
A correlation plot will display correlations between the values of variables in the dataset. In addition to the usual correlation calculated between values of different variables, the correlation between missing values can be explored by checking the Explore Missing check box. The first thing to notice for this correlation plot is that only the numeric variables appear. Rattle only computes correlations between numeric variables at this time. The second thing to note about the graphic is that it is symmetric about the diagonal. The correlation between two variables is the same, irrespective of the order in which we view the two variables. The third thing to note is that the order of the variables does not correspond to the order in the dataset, but to the order of the strength of any correlations, from the least to the greatest. This is done simply to achieve a more pleasing graphic which is easier to take in.
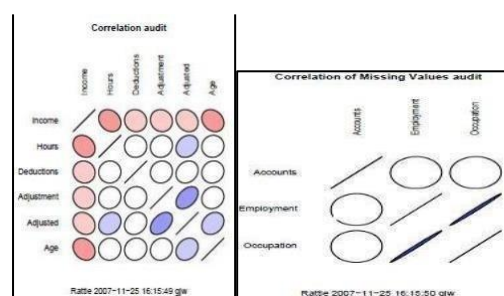
We interpret the degree of any correlation by both the shape and colour of the graphic elements. Any variable is, of course, perfectly correlated with itself, and this is reflected as the diagonal lies on the diagonal of the graphic. Where the graphic element is a perfect circle, then there is no correlation between the variables, as is the case in the correlation between Hours and Deductions although in fact there is a correlation, just a very weak one. The colours used to shade the circles give another (if perhaps redundant) clue to the strength of the correlation. The intensity of the colour is maximal fora perfect correlation, and minimal(white) if there is no correlation. Shades of red are used for negative correlations and blue for positive correlations.

By selecting the Explore Missing check box you can obtain a correlation plot that will show any correlations between the missing values of variables. This is particularly useful to understand how missing values in one variable are related to missing values in another. We notice immediately that only three variables are included in this correlation plot. Rattle has identified that the other variables in fact have no missing values, and so there is no point including them in the plot. We also notice that a categorical variable, Accounts, is included in the plot even though it was not included in the usual correlation plot. In this case we can obtain a correlation for categorical variables since we only measure missing and presence of a value, which is easily interpreted as numeric.

The graphic shows us that Employment and Occupation are highly correlated in their presence of missing values. That is, when Employment has a missing value, so does Occupation, and vice versa, at least in general. The actual correlation is 0.995 (which can be read from the Rattle text view window), which is very close to 1. On the other hand, there is no (in fact very little at 0.013) correlation between Accounts and the other two variables, with regard missing values. It is important to note that the correlations showing missing values may be based on very small samples, and this information is included in the text view of the Rattle window. For example, in this example we can see that there are only 100, 101, and 43 missing values, respectively, for each of the three variables having any missing values. This corresponds to approximately 5%, 5%, and 2% of the entities, respectively, having missing values for these variables.
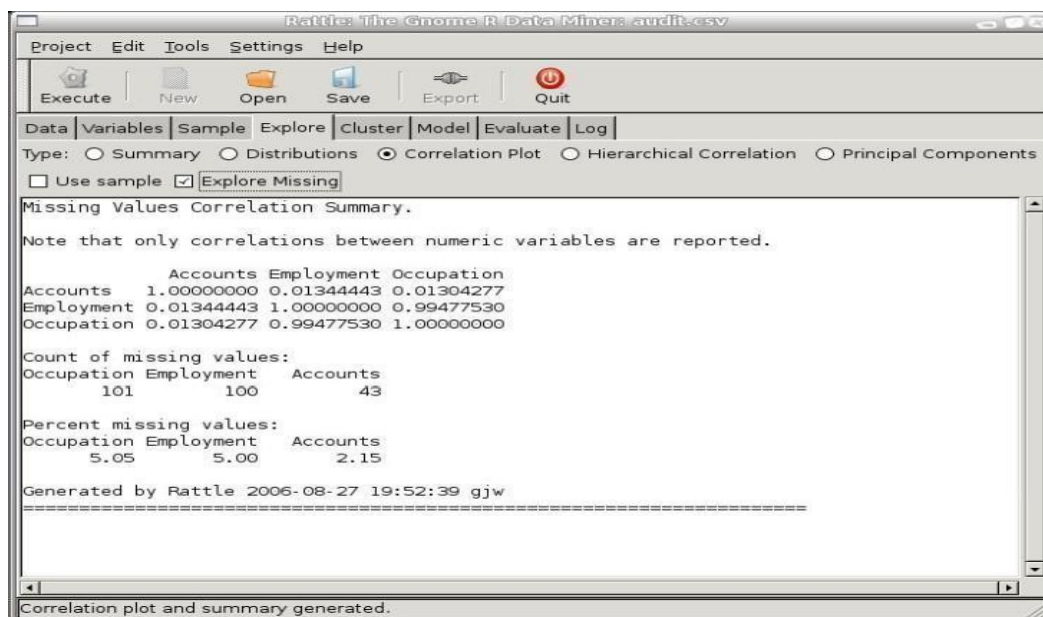
Rattle uses the default R correlation calculation known as Pearson's correlation, a common measure of correlation.

<u>Hierarchical Correlation</u>
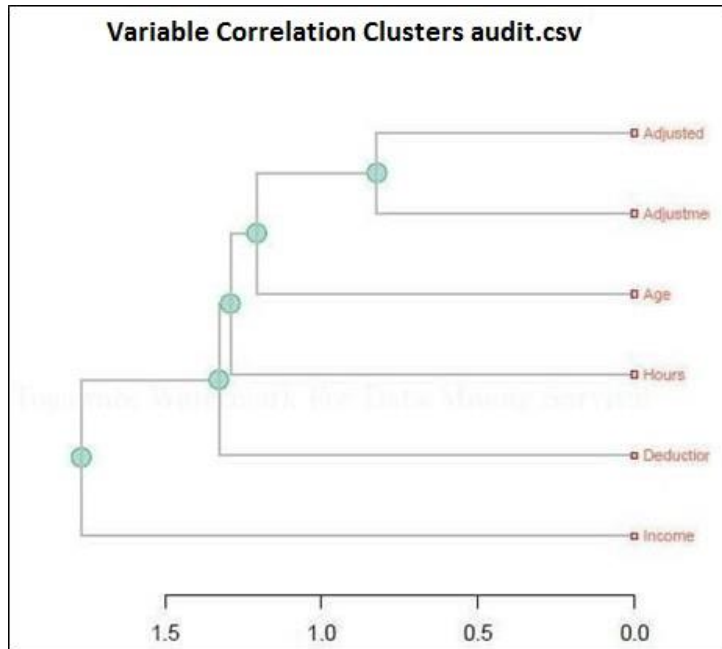
RESULT

Introduction to regression using rattle an open source tool is executed and results are verified.

Exp No: 4   **Date:**

CLASSIFICATION USING WEKA TOOL KIT – PART 1

AIM

Demonstration of classification rule process on dataset student.arff using j48 algorithm.

THEORY

This experiment illustrates the use of j-48 classifier in WEKA. The sample data set used in this experiment is "student" data available at arff format. This document assumes that appropriate data pre processing has been performed.
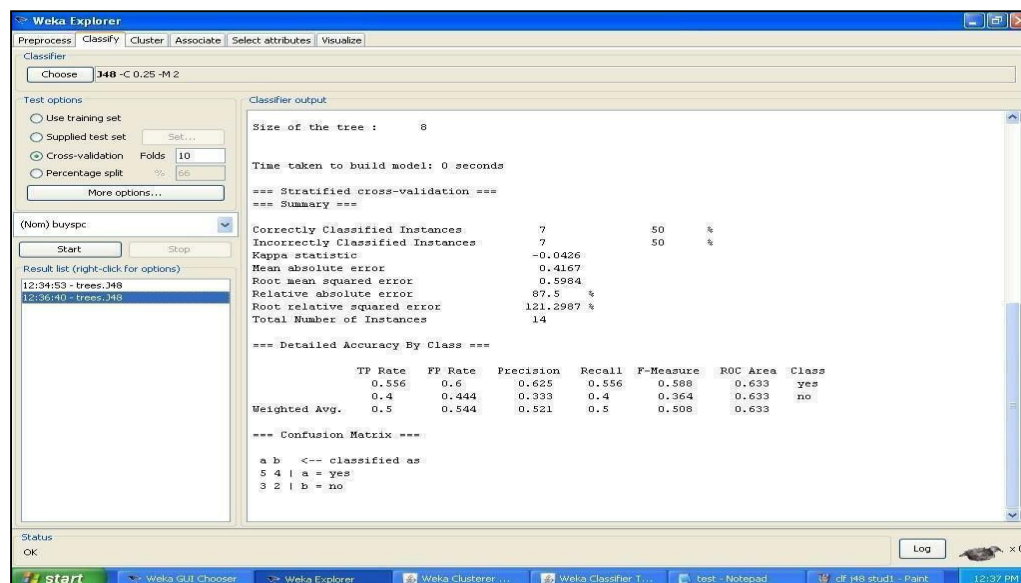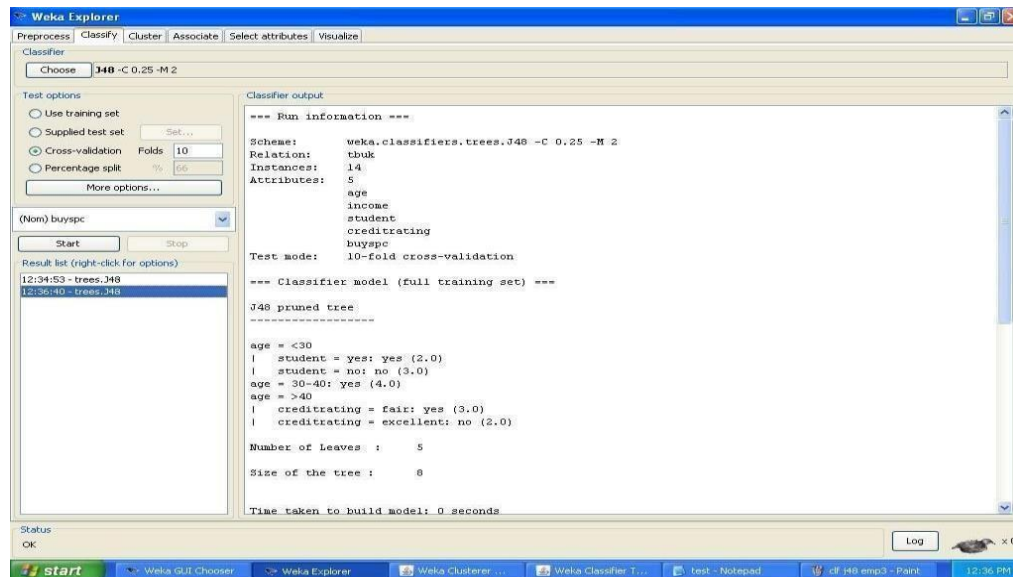
PROCEDURE

1. We begin the experiment by loading the data (student.arff) into WEKA.
2. Next we select the "classify" tab and click "choose" button to select the "j48" classifier.
3. Now we specify the various parameters. These can be specified by clicking in the text box to the right of the chose button. In this example, we accept the default values. The default version does perform some pruning but does not perform error pruning.
4. Under the "text" options in the main panel. We select the 10-fold cross validation as our evaluation approach. Since we don't have separate evaluation data set, this is necessary to get a reasonable idea of accuracy of generated model.
5. We now click "start" to generate the model. The ASCII version of the tree as well as evaluation statistic will appear in the right panel when the model construction is complete.
6. Note that the classification accuracy of model is about 69%. This indicates that we may find more work (either in pre-processing or in selecting current parameters for the classification).
7. Now WEKA also lets us a view a graphical version of the classification tree. This can be done by right clicking the last result set and selecting "visualize tree" from the pop-up menu.
8. We will use our model to classify the new instances.
9. In the main panel under "text" options click the "supplied test set" radio button and then click the "set" button. This wills pop-up a window which will allow you to open the file containing test instances.

Datasetstudent.arff

@relation student
@attribute age{<30,30-40,>40}
@attribute income {low, medium, high} @attribute student {yes, no}
@attributecredit-rating{fair,excellent} @attribute buyspc {yes, no}
@data
%
<30, high,no,fair,no
<30,high,no,excellent,no 30-40, high, no, fair, yes
>40,medium,no,fair,yes
>40,low, yes, fair,yes
>40, low, yes, excellent, no 30-40,low,yes,excellent,yes
<30,medium,no,fair,no

<30,low, yes, fair,no>40,medium,yes,fair,yes

<30, medium, yes, excellent, yes 30-40,medium,no,excellent,yes 30-40, high, yes, fair, yes

>40,medium,no,excellent,no

%

The following figures shows the classification rules that were generated when j48 algorithm is applied on the given dataset.

RESULT

Demonstration of classification rule process on dataset student.arff using j48 algorithm is executed and results are verified.

Exp No: 5 **Date:**

CLASSIFICATION USING WEKA TOOL KIT – PART 2

AIM

Demonstration of classification rule process on dataset employee.arff using id3 algorithm.

THEORY

This experiment illustrates the use of id3 classifier in WEKA. The sample data set used in this experiment is "employee" data available at arff format. This document assumes that appropriate data pre-processing has been performed.

PROCEDURE

1. We begin the experiment by loading the data (employee.arff) into WEKA.
2. Next we select the "classify" tab and click "choose" button to select the "id3" classifier.
3. Now we specify the various parameters. These can be specified by clicking in the text box to the right of the chose button. In this example, we accept the default values his default version does perform some pruning but does not perform error pruning.
4. Under the "text "options in the main panel. We select the 10-fold cross validation as our evaluation approach. Since we don't have separate evaluation data set, this is necessary to get a reasonable idea of accuracy of generated model.
5. We now click "start" to generate the model. The ASCII version of the tree as well as evaluation statistic will appear in the right panel when the model construction is complete.
6. Note that the classification accuracy of model is about 69%. This indicates that we may find more work (either in pre-processing or in selecting current parameters for the classification).
7. Now WEKA also lets us a view a graphical version of the classification tree. This can be done by right clicking the last result set and selecting "visualize tree" from the pop-up menu.
8. We will use our model to classify the new instances.
9. In the main panel under "text" options click the"supplied test set"radio button and then click the"set" button. This will show pop-up window which will allow you to open the file containing test instances.
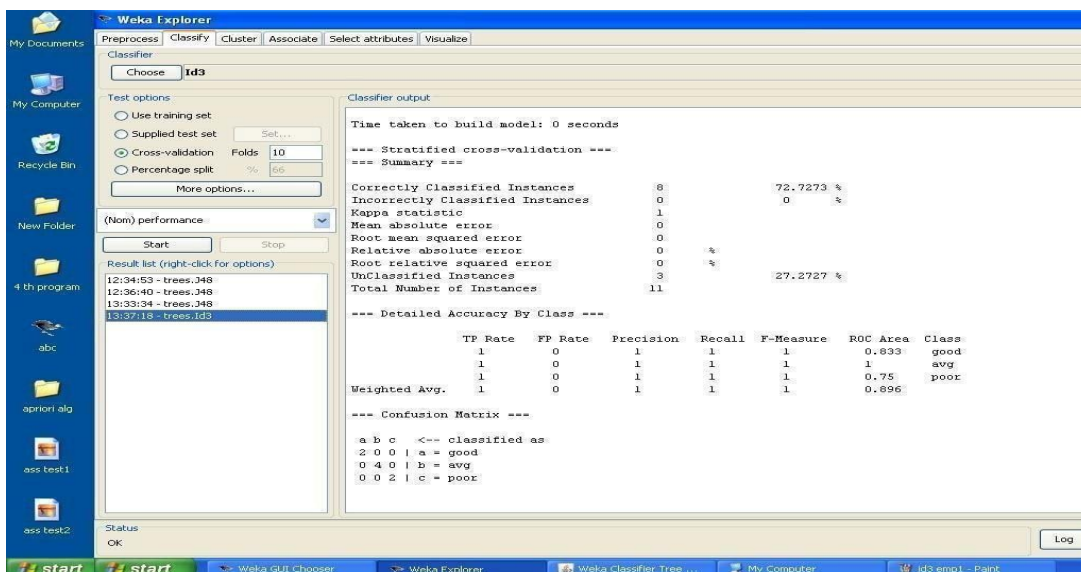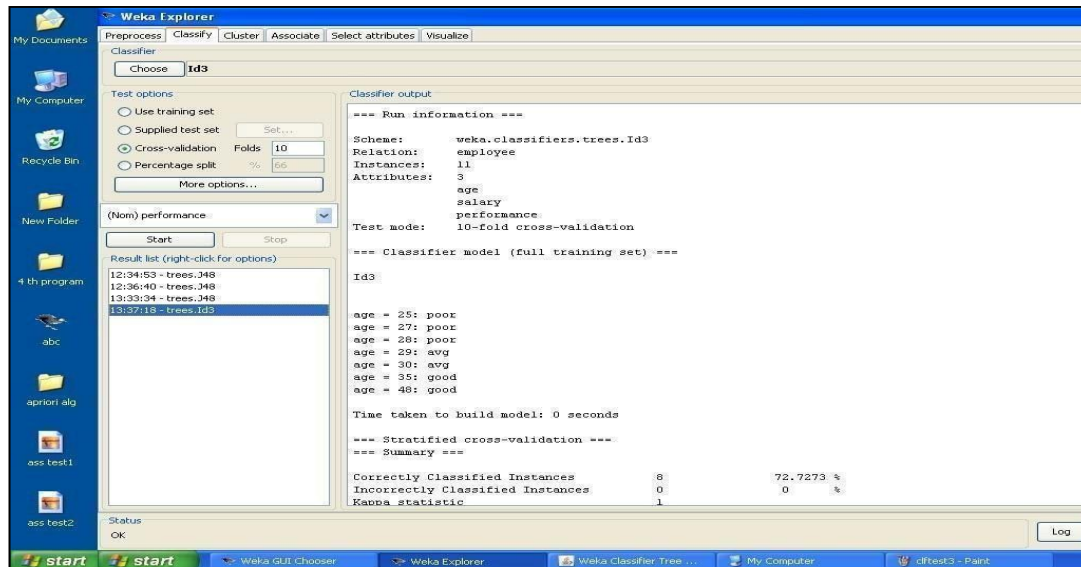
Dataset employee.arff

@relation employee
@attribute age{25,27,28,29,30,35,48}
@attribute salary{10k,15k,17k,20k,25k,30k,35k,32k}
@attribute performance {good, avg, poor}
@data
%
25,10k,poor
27,15k,poor
27,17k,poor
28,17k,poor
29,20k,avg
30,25k,avg

29,25k,avg
30,20k,avg
35,32k,good
48,34k,good
48,32k,good
%

OUTPUT:

The following figure shows the classification rules that were generated when id3 algorithm is applied on the given dataset.

RESULT

Demonstration of classification rule process on dataset employee.arff using id3 algorithm is executed and results are verified.

Exp No: 6   **Date:**

PERFORMING DATA PREPROCESSING FOR DATA MINING IN WEKA

AIM

Demonstration of pre-processing on dataset student.arff.

THEORY

This experiment illustrates some of the basic data pre-processing operations that can be performed using WEKA-Explorer. The sample dataset used for this example is the student data available in arff format.

PROCEDURE

1. Loading the data. We can load the dataset into WEKA by clicking on open button in pre-processing interface and selecting the appropriate file.

2. Once the data is loaded, weka will recognize the attributes and during the scan of the data WEKA will compute some basic strategies on each attribute. The left panel in the above figures show the list of recognized attributes while the top panel indicates the names of the base relation or table and the current working relation (which are same initially).

3. Clicking on an attribute in the left panel will show the basic statistics on the attributes for the categorical attributes the frequency of each attribute value is shown, while for continuous attributes we can obtain min, max, mean, standard deviation and deviation etc.

4. The visualization in the right button panel in the form of cross-tabulation across two attributes.

Note:we can select another attribute using the drop down list.

5. Selecting or filtering attributes

Removing an attribute-When we need to remove an attribute,we can do this by using the attribute filters in WEKA.Inthe filter model panel, click on choose button. This will show a pop up window with a list of available filters.

Scroll down the list and select the "weka.filters.unsupervised.attribute.remove"filters.

6. a)Next click the text box immediately to the right of the choose button. In the resulting dialog box enter the index of the attribute to be filtered out.

   b) Make sure that invert selection option is set to false. The click OK now in the filter box. You will see "Remove-R-7".

   c) Click the apply button to apply filter to this data. This will remove the attribute and create new working relation.

   d) Save the new working relation as an arff file by clicking save button on the top (button) panel (student.arff).

Discretization

Sometimes association rule mining can only be performed on categorical data. This requires performing discretization on numeric or continuous attributes.

In the following example let us discretize age attribute.

- Let us divide the values of age attribute into three bins (intervals).
- First load the data set into weka (student.arff)
- Select the age attribute.
- Activate filter-dialog box and select "WEKA.filters.unsupervised.attribute.discretize" from the list.
- To change the defaults for the filters, click on the box immediately to the right of the choose button.
- We enter the index for the attribute to be discretized. In this case the attribute is age. So we must enter '1' corresponding to the age attribute.
- Enter'3'as the number of bins. Leave the remaining field values as they are.
- Click OK button.
- Click apply in the filter panel. This will result in a new working relation with the selected attribute partition into 3 bins.
- Save the new working relation in a file called student-data-discretized.arff

Datesetstudent.arff

@relation student
@attribute age{<30,30-40,>40}
@attribute income {low, medium, high}
@attribute student {yes, no}
@attribute credit-rating{fair,excellent}
@attribute buyspc {yes, no}
@data
%
<30, high,no,fair,no
<30,high,no,excellent,no 30-40, high, no, fair, yes
>40,medium,no,fair,yes
>40,low, yes, fair,yes
>40, low, yes, excellent, no 30-40,low,yes,excellent,yes
<30,medium,no,fair,no
<30,low, yes, fair,no
>40,medium,yes,fair,yes
<30, medium, yes, excellent, yes 30-40,medium,no,excellent,yes 30-40, high, yes, fair, yes
>40,medium,no,excellent,no
%

The following figure shows the effect of discretization.

OUTPUT:
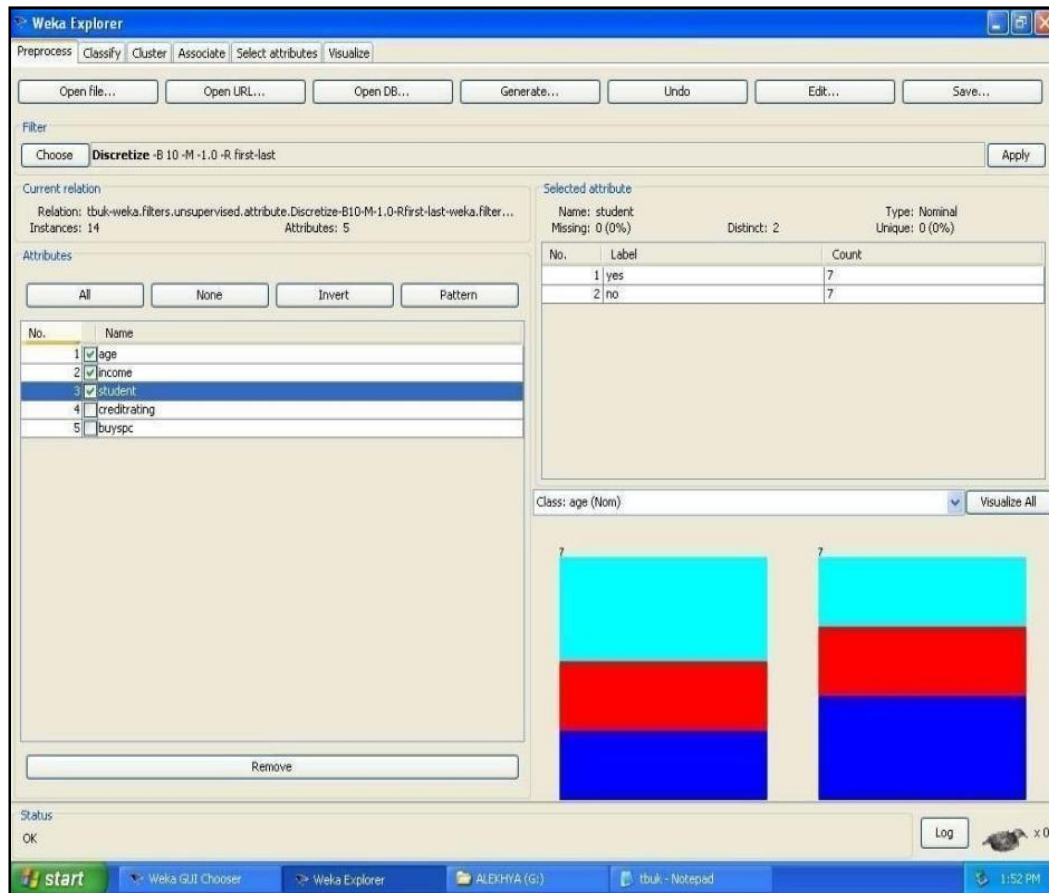
RESULT

Performing data pre-processing for data mining in WEKA is executed and results are verified.

Exp No: 7   **Date:**

PERFORMING CLUSTERING IN WEKA

AIM

Demonstration of clustering rule process on dataset student.arff using simple k-means.

THEORY

This experiment illustrates the use of simple k-mean clustering with Weka explorer. The sample data set used for this example is based on the student data available in ARFF format. This document assumes that appropriate pre-processing has been performed. Thisstudentdatasetincludes14instances.

Steps involved in this Experiment

Step1) Run the Weka explorer and load the data file student.arff in pre-processing interface.

Step2) In order to perform clustering select the 'cluster' tab in the explorer and click on the choose button.Thisstepresultsinadropdownlistofavailableclusteringalgorithms.

Step3) In this case we select 'simple k-means'.

Step4) Next click in text button to the right of the choose button to get popup window shown in the screenshots. In this window we enter six on the number of clusters and we leave the value of the seed on as it is. The seed value is used in generating a random number which is used formating the internal assignments of instances of clusters.

Step5) Once of the option have been specified. We run the clustering algorithm there we must make sure that they are in the 'cluster mode' panel. The use of training set option is selected and then we click 'start' button. This process and resulting window are shown in the following screenshots.

Step6) The result window shows the centroid of each cluster as well as statistics on the number and the percent of instances assigned to different clusters. Here clusters centroid are means vectors for each clusters. This clusters can be used to characterized the cluster.

Step7) Another way of understanding characteristics of each cluster through visualization ,we can do this, try right clicking the result set on the result. List panel and selecting the visualize cluster assignments.

Interpretation of the above visualization

From the above visualization, we can understand the distribution of age and instance number in each cluster. For instance, for each cluster is dominated by age. In this case by changing the color dimension to other attributes we can see their distribution with in each of the cluster.

Step8) We can assure that resulting dataset which included each instance along with its assign cluster. To do so we click the save button in the visualization window and save the result student k-mean. The top portion of this file is shown in the figure.
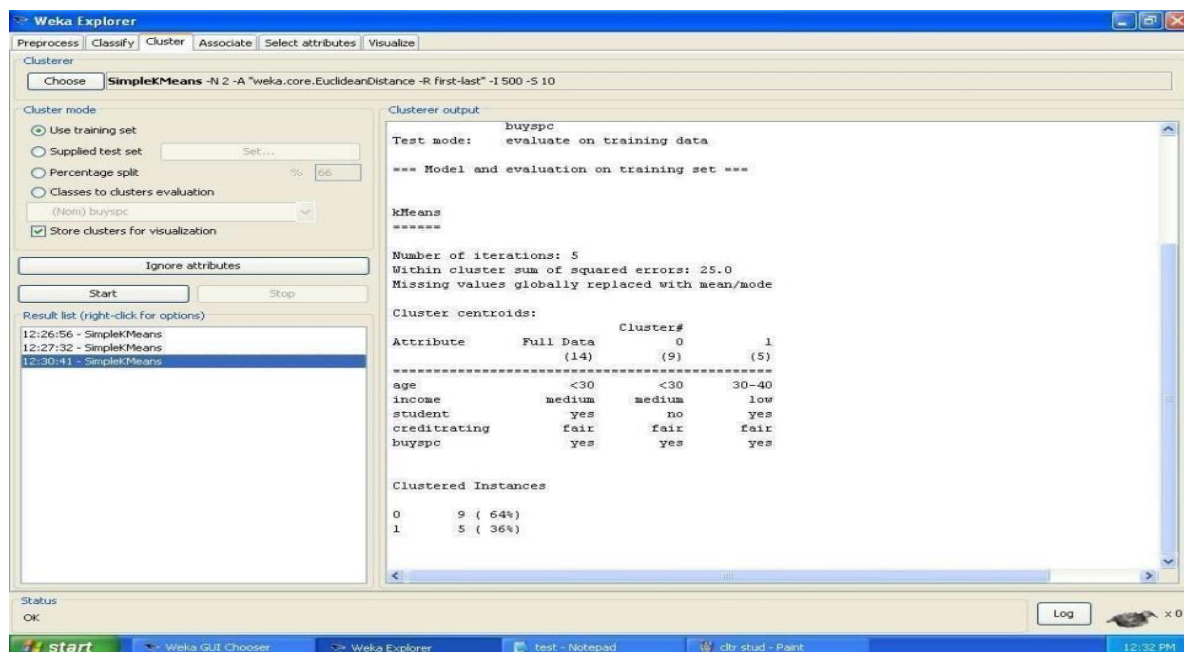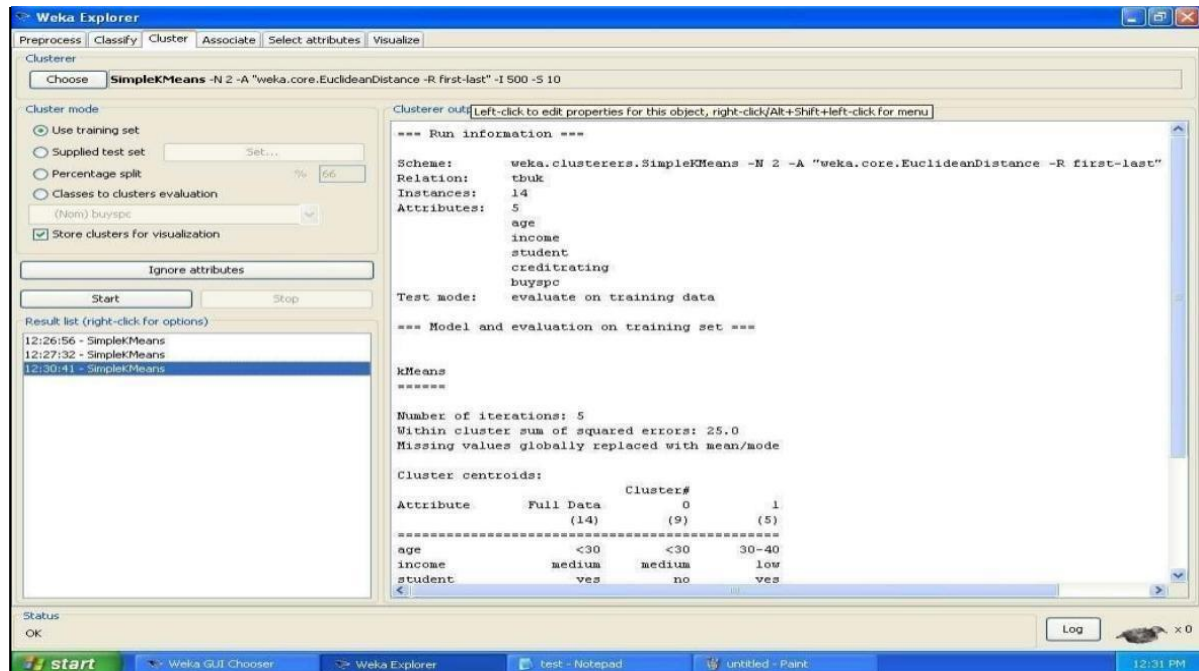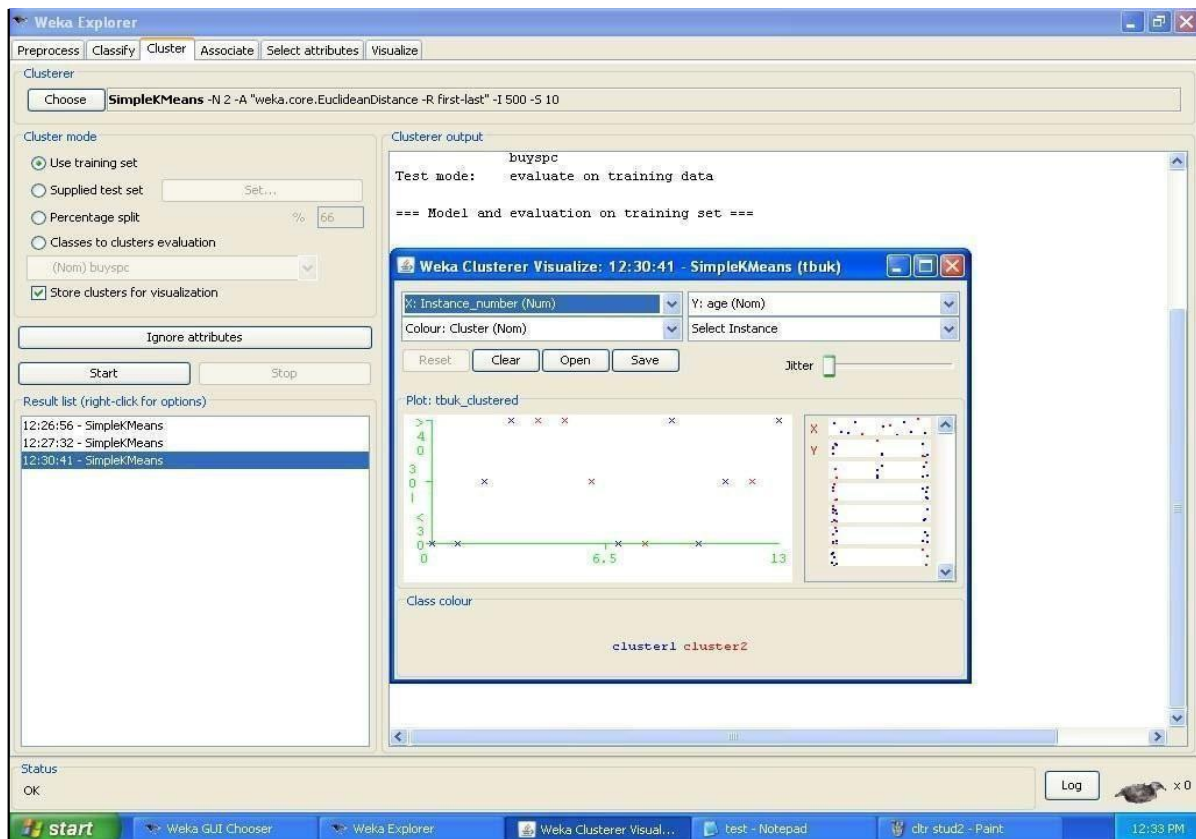
Dataset student.arff

@relation student
@attribute age{<30,30,40,>40}
@attribute income{low,medium,high}
@attribute student {yes,no} @attribute credit-rating
{fair,excellent}@attribute buyspc
{yes,no} @data
%

<30, high, no, fair, no
<30, high, no, excellent,
no30-40, high, no, fair, yes
>40, medium, no, fair, yes
>40, low, yes, fair, yes
>40, low, yes,excellent, no
30-40, low, yes, excellent, yes
<30, medium, no, fair, no
<30, low, yes, fair, no
>40, medium, yes, fair, yes
<30, medium, yes, excellent, yes 30-40, medium, no, excellent, yes30-40, high, yes, fair, yes
>40, medium, no, excellent, no
%

OUTPUT:

The following screenshot shows the clustering rules that were generated when simple k-means algorithm is applied on the given dataset.

RESULT

Thus, demonstration of clustering rule process on dataset student.arff using simple k-means has been executed and the output is verified successfully.

Exp No: 8   **Date:**

<p style="text-align:center">ASSOCIATION RULE ANALYSIS IN WEKA</p>

AIM

Demonstration of association rule process on dataset contact_lenses.arff using apriori algorithm.

THEORY

This experiment illustrates some of the basic elements of association rule mining using WEKA. The sample data set used for this example is contactlenses.arff.

**Steps involved in this Experiment**

Step1) Open the data file in Weka Explorer. It is presumed that the required data fields have been discretized. In this example it is age attribute.

Step2) Clickingontheassociatetabwillbringuptheinterfaceforassociationrulealgorithm.

Step3) Will use apriori algorithm. This is the default algorithm.

Step4) In order to change the parameters for the run (example support, confidence etc) we click on the text box immediately to the right of the choose button.

Dataset contact_lenses.arff

Sumit Kumar   27   211061101448
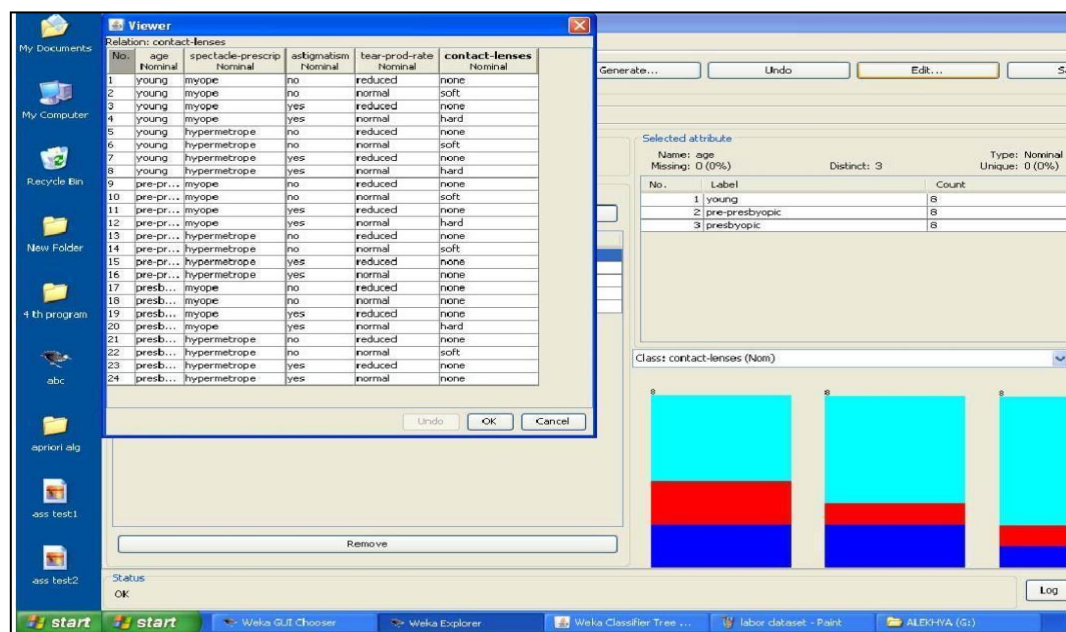
The following screenshot shows the association rules that were generated when apriori algorithm is applied on the given dataset.
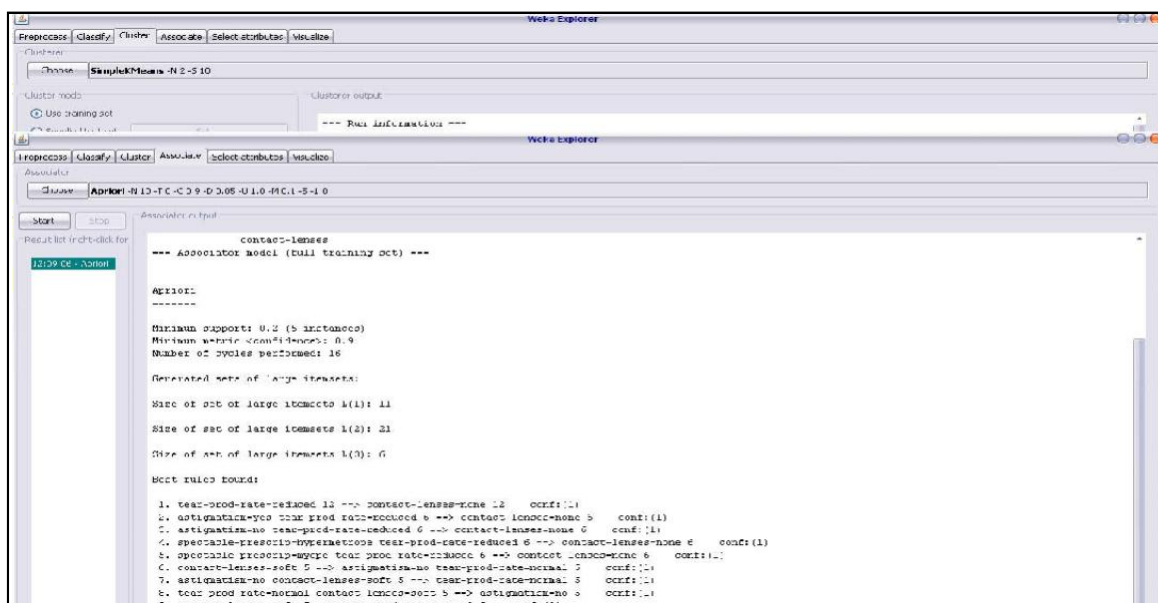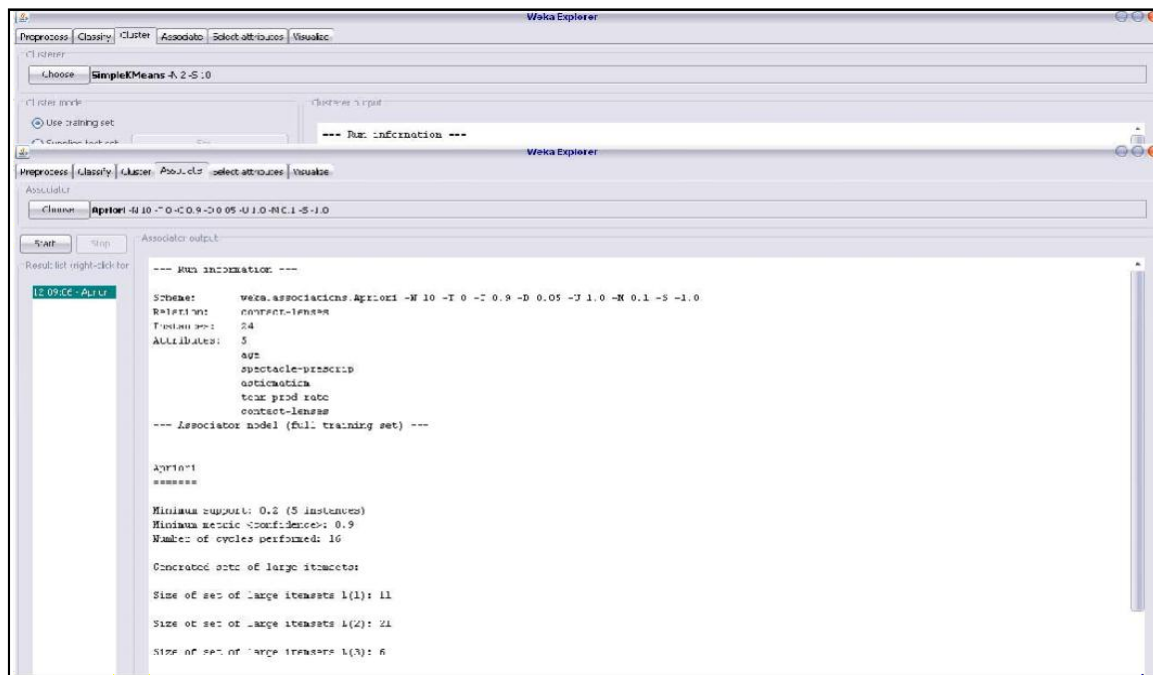
RESULT

Thus, demonstration of association rule process on dataset contact_lenses.arff using apriori algorithm has been executed and the output is verified successfully.

Exp No: 9 **Date:**

DATA MINING CASE STUDY USING THE CRISP-DM STANDARD

AIM

CRISP- DM CASE STUDY – Early Prediction Of Student Success: Mining Students Enrolment Data.

THEORY

This case study explores the socio-demographic variables (age, gender, ethnicity, education, work status, and disability) and study environment (course programme and course block), that may influence persistence or dropout of students at the Open Poly technic of New Zealand. This case study examines to what extent these factors, i.e. enrolment data help us in pre-identifying successful and unsuccessful students. The data stored in the Open Polytechnic student management system covers over 450 students who enrolled to *71150* Information Systems course was used to perform a quantitative analysis of study outcome. Based on a data mining techniques (such as feature selection and classification trees), the most important factors for student success and a profile of the typical successful and unsuccessful students are identified. The empirical results show the following:

  i   The most important factors separating successful from unsuccessful students are:ethnicity, course programme and course block;

  ii   Among classification tree growing methods Classification and Regression Tree(CART) was the most successful in growing the tree with an overall percentage of correctclassificationof60.5%;and

  iii   Both the risk estimated by the cross-validation and the gain diagram suggests that all trees, based only on enrolment data are not quite good in separating successful from unsuccessful students. The implications of these results for academic and administrative staff are discussed.

INTRODUCTION

Increasing student retention or persistence is a long term goal in all academic institutions. The consequences of student attrition are significant for students, academic and administrative staff. The importance of this issue for students is obvious: school leavers are more like lyto earnless than those who graduated. Since one of the criteria for government funding in the tertiary education environment in New Zealand is the level of retention rate, both academic and administrative staff are under pressure to come up with strategies that could increase retention rates on their courses and programmes. The most vulnerable students to low student retention at all institutions of higher education are the first-year students, who are at greatest risk of dropping out in the first term or semester of study or not completing their programme/degree on time.Therefore most retention studies address the retention of first-year students.Consequently, the early identification of vulnerable students who are prone to drop their courses is crucial for the success of any retention strategy. This would allow educational institutions to undertake timely and pro-active measures. Once identified, these 'at-risk' students can be then targeted with academic and administrative support to increase their chance of staying on the course.

The main objective of this study is to explore factors that may impact the study out come in the Information Systems course at the Open Polytechnic. The Information Systems course is a core course for those majoring in IT and for most students an entry point, i.e.the first course they are taking with the Open Polytechnic. This issue have not been examined so far for Open Polytechnic and this case study attempts to fill the gap**.**

More specifically the enrolment data were used to achieve the following objectives:

- Build models for early prediction of study out come using the student enrolment data
- Evaluate the models using cross-validation and misclassification errors to decide which model out performs other models in term of classification accuracy
- Present results which can be easily understood by the users (students, academic and administrative staff)

At the time of enrolment in the Open Poly technic of New Zealand, the only information. i.e. variables we have about students are those contained in their enrolment forms. The question we are trying to address in this case study is whether we can use the enrolment data alone to predict study outcome for newly enrolled student.

Framework for Data Mining Process

Framework for data mining applications is based on the CRISP-DM Model created by a consortium of NCR, SPSS, and Daimler-Benz companies. The modified version of the CRISP- DM model is presented on Figure 1, following the project through the general lifecycle from business and data understanding, data preparation, modelling, evaluation and deployment. The feedback from deployment to data and business understanding illustrates the iterative nature of a data mining process.
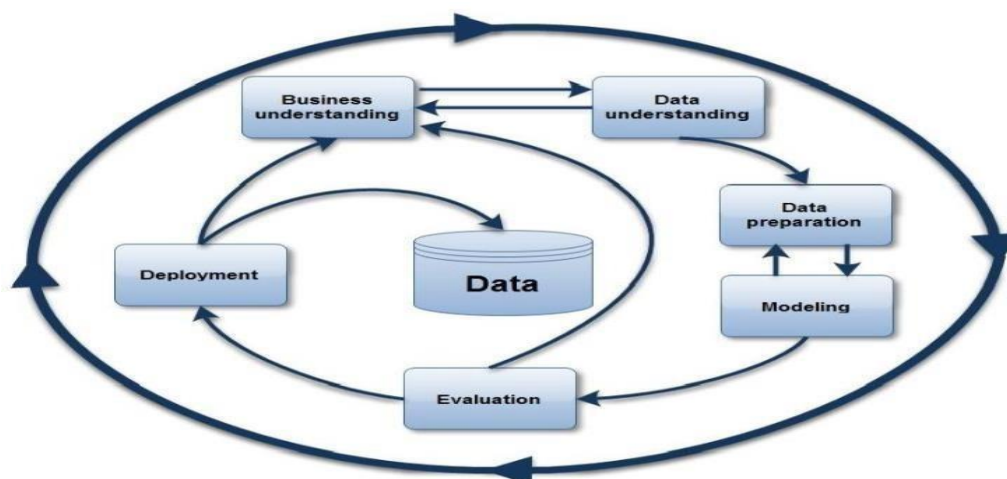

Figure1: Modified CRISP-DMModelVersion1

Business Understanding

The business understanding phase begins with the setting up of goals for the data mining project. In this paper that would be an increasing understanding of the pre-enrollment factors that may prevent students from successfully complete the course. Because we are planning to increase completion rate on the Information Systems course understanding its students population and patterns in the pre-enrollment data becomes necessary before we start developing a predictive model. In this phase we are coming up with the following questions: what is the profile of a student who successfully completes this course? Can the successful vs. unsuccessful student be distinguished in terms of demographic features (such as gender, age or ethnic origin) or study environment (such as course programme, faculty or course block)?

Depending on the answers to these questions we consider the methods and approaches that can be adopted to increase the completion rate.

Data Understanding

The scope of our research in terms of data used is limited by the data available in the Open Polytechnic Student Management System (known as Integrator) and the enrolment form used for collecting data from newly enrolled students. It is important to have a full understanding of the nature of the data and how it was collected and entered before proceeding further. In this phase an initial data exploration using a pivot table was also conducted to get some insight in the data. The enrolment form asks students to enter the following information: demographic (gender, date of birth, ethnic origins, disability, and work status), academic (secondary school qualification, course programme, course faculty, and course block), and contact details. Once the data from the enrolment form is entered in to Integrator and the Enrolment Section process the application, the enrolment date is recorded and student becomes enrolled on the course.

Data Presentation

In this phase the data are put into a form suitable for the modelling phase. If required some selected variables are combined, transformed or used to create an variables. For example, enrolment date and the course block start date were used to generate a variable labelled as "early enrolment". Any data excluded from the data set is documented and their removal explained. For example, only a few students enrolled to Bachelor of Arts programme were on the Information Systems course. They were removed from further analysis for the reason explained later in Data and Methodology section. Data are cleaned for any duplication of records. For example, in case of Information Systems course, the course code changed in the past. If student enrolled in the time when the change in the course code happened and then re-enrolled on the same course, two records exist in the data set for the same student and the same course, but under two different course codes. In this case data for this student were merged into one, single record. The dependent variable "study outcome" with two possible outcomes (labelled as Pass and Fail) indicates whether students successfully completed the course or failed the course due to voluntary transfer / withdrawal or academic withdrawal or simply because not fulfilling course pass requirements.

Modelling

In this phase we choose and ran models on the training data set. Then we decided whether a suitable model for the data set was found that was acceptable from both analytical and managerial standpoint. In this phase we decided to use classification tree models with four different tree growing criteria.

Evaluation

The final models from the previous phase are then applied on a testing, i.e. a validation data set with the aim of assessing their predictive accuracy and consistency with the results obtained for the training dataset. This phase involves an iterative process of fitting different versions of models to training and testing data set, each time evaluating their predictive performance.

Deployment

Once we decided on the final model we can apply it to current data not used during the modelling and evaluation phase. This process is known as scoring. The model results are used to address the issues identified in the business understanding phase. The results should be presented in the user friendly format and prepared for use by administrative staff. The final model should retain the highest predictive accuracy and if it is to be continuously used it should be regularly updated, particularly if some organizational changes occur or if new factors are brought in. For example, if the new information about financial support is added to the student record or enrolment form, then the model should consider the new factor that might be relevant for study outcome.

DATA AND METHODOLOGY

Data Preparation

Variables definition and their domains are presented in Table 1.From the initial dataset all students granted cross-credit or credit were excluded because they didn't actually study this course. The courses they have previously completed were recognized and credited for this course. We also removed 6 students enrolled under Bachelor of Arts program. Such as mall number of students in this category do not allow generalization of the results for this particular category. The total number of data was reduced to 453. We needed to clarify the definition of study outcome that we used in our analysis. We considered only two possible outcomes, labelled as: Pass and Fail. Students labeled Pass successfully completed the course. Students labelled Fail transferred or withdrew from the course voluntarily or academic registry withdrew them for not completing the in-course assessments. Those students who stayed on the course until the end of the course but scored less than the course pass mark were also allocated into this category. Because of the data set size (only 453 students) numeric continuous variable such as age was converted into a categorical variable with only three age groups. In Ethnicity we combined Maori and Pacific Island students because of two reasons: they were found to be no different in preliminary bi-variant analysis and combined together constitute a small proportion of the data (less than 10%). Combining them into one ethnic group helps with model parsimony. The Secondary school variable combines all students with no secondary school up to NCEA Level2 on the New Zealand National Qualifications Framework in to one group.

Though the software applications skills (spreadsheet and database in particular) are very important on this course we do not take into account the different skill levels present on the Information Systems course. Since the students enrolling on this course have different backgrounds and levels of interest in computing, we would expect that skill level has also a significant impact on the study outcome. However, the information about their Office skills level is not available in the moment of enrolment, so not included this factor in our analysis.

Variables	Description(Domain)
Student Demographics	
Gender	Student gender (binary: female, male)
Age	Students age (numeric: 1- under 30, 2- 30 to 40 or 3 – over 40)
Ethnicity	Student's ethnic group(nominal: Pakeha, Maori & Pacific Islanders or Others)
Disability	Student has a disability(binary: yes or no)
Secondary school	Student's highest level of achievement from a secondary school on the New Zealand National Qualifications Framework(nominal:No secondary qualification,NCEA1,NCEA2, University entrance,NCEA3,Overseas or Other)
Work status	Student is working(binary:yes or no)
Early enrolment	Student enrolled for the first time in the course before start of the course (binary:yes or no)
Study Environment	
Course programme	Programme(nominal: Bachelor of Business or Bachelor of Applied Science)
Course Block	Semester in which a course is offered(Semester1, Semester2 or Semester3)
Study outcome	Study outcome (nominal: Pass – successful completion, Fail – unsuccessful completion includes also withdrawals,academic Withdrawals and transfers)

Table1:Description of variables and their domains

Methodology

Three types of data mining approaches were conducted in this study. The first approach is descriptive which is concerned with the nature of the dataset such as the frequency table and the relationship between the attributes obtained using cross tabulation analysis (contingency tables). In addition, feature selection is conducted to determine the importance of the prediction variables for modeling study outcome. The third type of data mining approach, i.e. predictive data mining is conducted by using four different classification trees. Finally, a comparison between these classification tree models was conducted to determine the best model for the dataset.

We decided to use the classification tree models because of some advantages they may have over traditional statistical models such as logistic regression and discriminant analysis traditionally used in retention studies. First, they can handle a large number of predictor variables, far more than the logistic regression and discriminant analysis would allow. Secondly, the classification tree models are non-parametric and can capture online relationships and complex interactions between predictors and dependent variable.

Results and Discussion

Before growing the classification trees we summarized the variables by categories and by study outcome, i.e. whether students passed or failed the course. Feature selection was used to rank the variables by their importance for further analysis. Finally, the classification tree results for four different growing methods are presented.

Summary Statistics

As part of the data understanding phase we carried out the cross-tabulation for each variable and the study outcome after preparing and cleaning the data. The Table 2 reports the results. Based on the results shown majority of Information Systems students are female (over 63%). However, percentage of female students who successfully complete the course are higher (65%) which suggests that female students are more likely to pass the course than their male counterpart. When it comes to age over 68% of students are above 30 with the age group between 30 and 40 being majority.This age group is also more likely to fail the course because their percentage of students who failed the course in this age group (39.7%) is higher than their overall participation in the student population(38.6%).

Variable	Domain	Count	Total	Pass	Fail
Gender	Female	286	63.1	65	61.5
	Male	167	36.9	35	38.5
Age	Under 30	136	30	30.4	29.7
	Btw 30 & 40	175	38.6	37.4	39.7
	Above 40	142	31.3	32.2	30.5
Disability	Yes	19	4.2	3.3	5
	No	434	95.8	96.7	95
Ethnicity	Pakeha	318	70.2	75.7	65.3
	Maori& Pacific Islanders	41	9.1	2.8	14.6
	Others	94	20.8	21.5	20.1

Disability was shown to be a disadvantage for Information Systems students. Students with it are more likely to fail than those without it. There are huge differences in percentage of students who successfully completed the course depending on their ethnic origin. Though Maori and Pacific Islanders make 9.1% of all students on this course their participation is significantly lower in the"Pass"sub population (i.e.2.8%) and higher in the "Fail" sub population (14.6%). Based on these results we can say that students with this ethnic origin are identified as students"at-risk".Further methods of data mining will confirm this statement.

A substantial number of students (over 40%) don't have secondary school qualification higher than NCEA Level 2 on the New Zealand National Qualification Framework and they are more vulnerable than the other two categories in this variable. Over three-fourths of Information Systems students are working and studying at the same time. Though the difference between those who work and those who
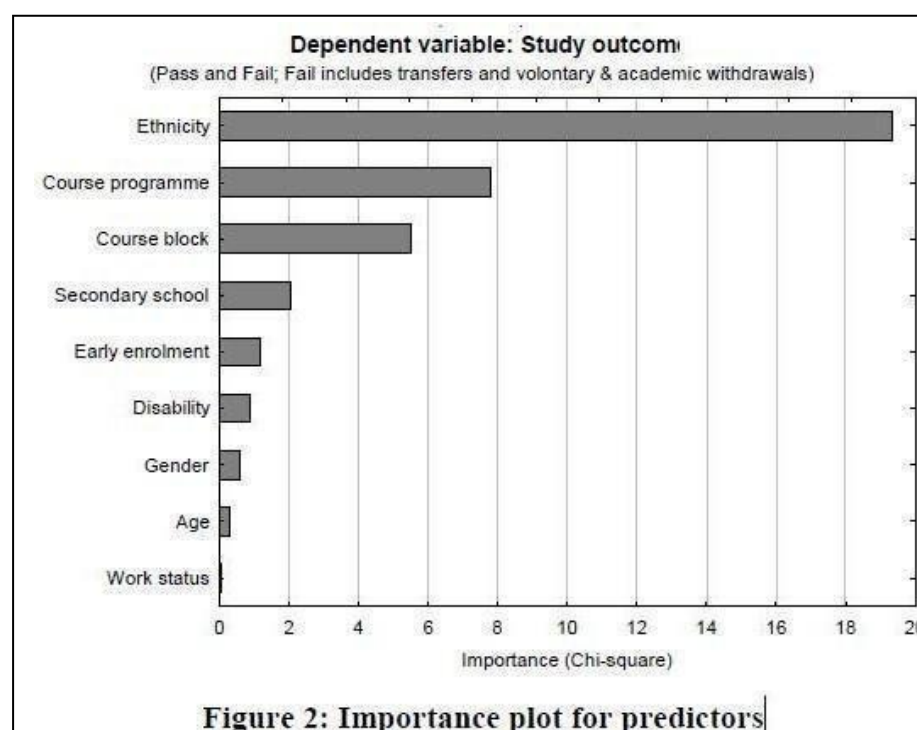
do not is not statistically significant, it is interesting to note that the students who are working are more likely to pass the course than t hose not working.

We are using early enrolment as a proxy for motivation and good time management skills. Students who are motivated and are planning their study in advance will also enroll well before the enrolment closing date. The opposite category (late comers) participates with30% in the total number of students, but these students are more likely to fail the course. Their participation in the "Fail" sub population increased from 30% to 32.2%.One third of students on this course enrolled on the Bachelor of Applied Sciences program. They are more likely to fail the course when compared with students enrolled on the Bachelor of Business program. Finally, students studying this course in the summer semester are more likely to fail than those studying in the second and first semester.

Feature Section

The number of predictor variables is not so large and we don't have to select the subset of variables for further analysis which is the main purpose of applying feature selection to data. However, feature selection could be also used as a pre-processor for predictive data mining to rank predictors according to the strength of their relationship with dependent or outcome variable. During the feature selection process no specific form of relationship, neither linear nor non-linear,is assumed. The outcome of the feature selection would be a rank list of predictors according to their importance for further analysis of the dependent variable with the other methods for regression and classification.

Results of feature selection are presented in Figure 2 on the importance plot and also in Table 3. The top three predictors for the study outcome are: ethnic origin of students,course programme they are enrolled on and course block, i.e. semester in which they are study.

Dependent variable: Study outcome
(Pass and Fail; Fail includes transfers and volontary & academic withdrawals)

Figure 2: Importance plot for predictors

Variable	Chi-square	P-value
Ethnicity	19.35	0.00006
Course Programme	7.8	0.00523
Course Block	5.51	0.06354
Secondary School	2.06	0.35748
Early enrolment	1.16	0.28131

Table3:Best predictors for dependent variable

From Table 3, *P*-values from the last column only the first three chi-square values are significant at 10% level. Though the results of the feature selection suggested continuing analysis with only the subset of predictors, which includes ethnicity, course programme and course block, we have included all available predictors in our classification tree analysis. We follow an advice given in Luan & Zhao(2006) who suggested that even though some variables may have little significance to the overall prediction outcome, they can be essential to a specific record.

Classification Trees

The objective of an analysis based on a classification tree is to identify factors that contribute the most to separation of successful from unsuccessful students. When the classification tree is formed we can calculate the probability of each student being successful. Once the classification tree is formed, it could be used in the new data set to predict the study outcome for newly en-rolled students. The classification trees for study outcome are given in Figure 2 and 4. In each tree node the number of successful students (4thline, last column) and unsuccessful students (3rdline, last column) is given, as well as the percentages for each category (2nd column) and relative and absolute size of the node(5th line). The variable names above the nodes are the predictors that provided the best split for the node according to the classification and regression tree-style exhaustive search for uni-variate splits method. This method looks at all possible splits for each predictor variable at each node. The search stops when the split with the largest improvement in goodness of fit, based on the Gini measure of node impurity, is found. Immediately above the nodes are categories which describe these nodes. Note that all available predictor variables in the dataset were included in the classification tree analysis in spite their insignificance detected in the feature selection section.

CHAID, exhaustive CHAID and QUEST

Three classification tree growing methods, namely: CHAID, exhaustive CHAID and QUEST generated exactly the same tree structure presented in Figure 3. It shows that only2 variables were used to construct the tree: (1) ethnicity and (2) course program. All the other student demographics variables were used but not included in the final model. We could change the stopping criteria to allow further growing of the tree, but that would result in nodes with just a few students. In the most extreme case we can continue splitting the tree until we create a terminal node for every student. However, we would get a model, i.e. classification tree that fits data better, but with more likely poor performance when used on a new data set. This phenomenon is known as over fitting the tree. The largest successful group (i.e. students who successfully completed the course) consists of 274(60.5%) students (Node 3). Ethnic origin of students in this group is either Pakeha or other ethnic groups (excluding Maori and

Pacific Islands students). Students in this group opted for the Bachelor of Business programme. The largest unsuccessful group (i.e. students who were unsuccessful) contains 138 students (30.5% of all participants) (Node 4). They belong to either Pakeha or other ethnic groups (excluding Maori and Pacific Islands students). The next largest group considered also as unsuccessful students, contains 41, i.e. 9.1% of all students, where 75.4% of the mare unsuccessful (Node1). They are described as Maori and Pacific Islands students.
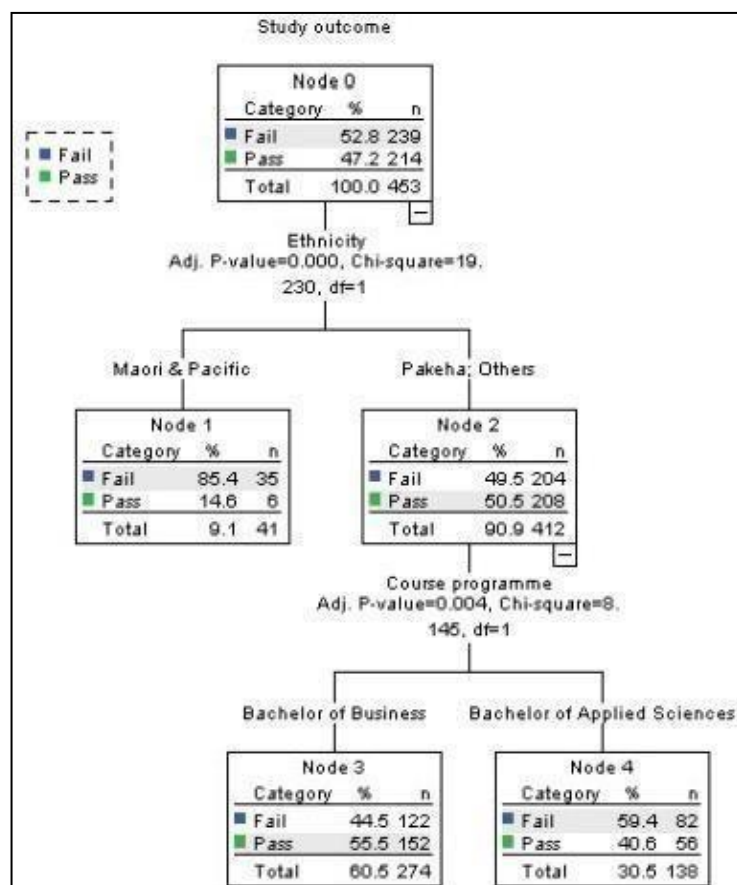

Figure 3: CHAID, exhaustive CHAID and QUEST classification tree

The overall percentage of correct classification for the study outcome is only 59.4% (Table 4). This percentage of correct classification was achieved with 2 variables only. The cross-validation estimate of the risk is 0.406 indicates that the category predicted by the model (successful or unsuccessful student) is wrong for 40.6% of the cases. So the risk of misclassifying a student is approximately 41%. This result is consistent with the results in the CHAID classification matrix (Table 4). The Overall percentage shows that the model only classified correctly 59% of students.

Observed	Fail	Pass	Percent correct
Fail	117	122	49.00%
Pass	62	152	71.00%
Overall Percentage	65.4	55.5	59.4%

With large numbers of false positives (122) and few false negatives (62), the CHAID model is in itself poor at identifying an unsuccessful student (positive predictive value is only 49%). It will, however, pick up 65.4% of all unsuccessful students (the sensitivity). The predictive values, which take into account the prevalence of failing the course, are generally more important in determining the usefulness of a prediction model. The negative predictive value was of more concern to the course because the objective was to minimize the probability of being in error when deciding that a student is not at risk for not completing the course. However the CHAID model, as a classification tool, will pick-up with high probability successful students (negative predictive valueis71%) and correctly identifies 55.5% of those who pass the course (the specificity).

The classification matrix indicates also another problem with the model. For unsuccessful students it predicts failure for only 49% of them, which means that 51% of unsuccessful students are inaccurately classified with the successful students. Practical consequence of this misclassification is that these students would not received additional learning support provided to the students "at risk", simply because they will be classified by the model among successful students. This feature of the model is more critical than misclassification of the successful students among unsuccessful students(29% of successful students belong to this category). In this case these students may receive additional learning support even though they don't need it. One option to increase percentage of correctly classified unsuccessful students is to change the misclassification cost matrix. With this option there is always a trade-off between increasing the percentage of correct classification of unsuccessful students and decreasing percentage of correct classification for successful students as well as decreasing the percentage of overall correct classification. In this case the increased cost for misclassification of unsuccessful to the successful group of students decreased significantly both remaining percentage of correct classification which was not compensated in equivalent increase in the initial 49%.

CART

Figure 5 shows the CART classification tree for study outcome. It shows that only three variables were used to construct the tree: (1) ethnicity, (2) course programme and (3) course block. The largest successful group (i.e. students who successfully completed the course) consists of 215(47.5%) students (Node 5). The ethnic origin of students in this group is either Pakehao rather ethnic groups (excluding Maori and Pacific Islands students). Students in this group enrolled on the Bachelor of Business programme in either Semester1 or Semester2. The large stun successful group (i.e. students who were unsuccessful) contains 138 students (30.5% of all participants)(Node 4). They belong to either Pakeha or other ethnic groups (excluding Maori and Pacific Islands students). The next largest group considered also as unsuccessful students, contains 41, i.e 9.1% of all students, where 75.4% of them are unsuccessful (Node 1). They are described as Maori and Pacific Islands students.

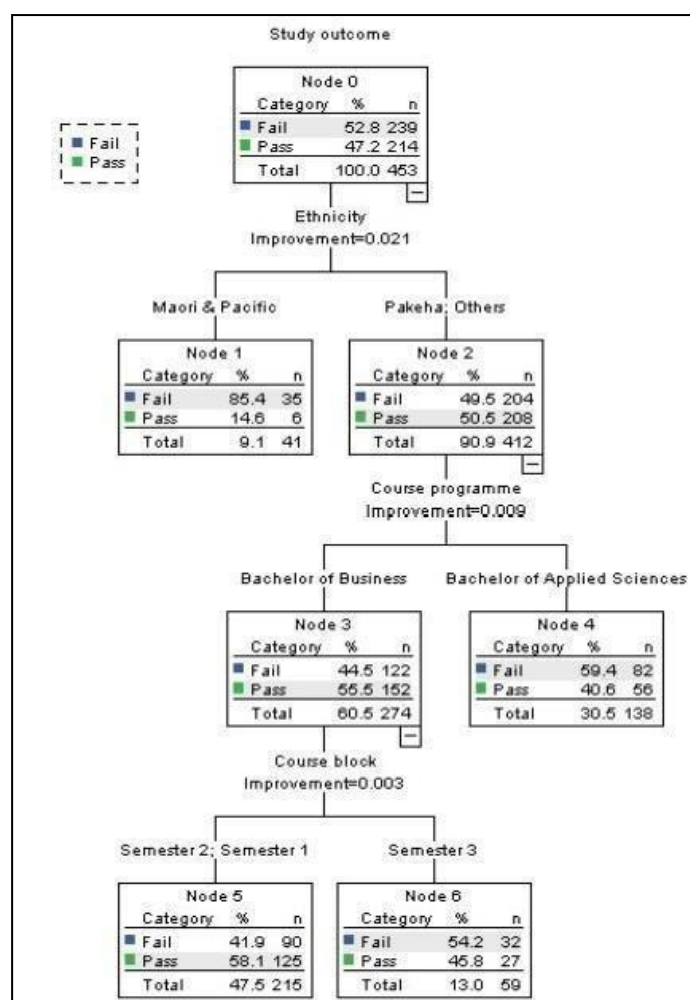| Observed	Fail	Pass	Percent correct
Fail	149	90	62.30%
Pass	89	125	58.40%
Overall Percentage	62.60%	58.10%	60.50%

Figure5: CART classification tree

be cross-validation estimate of the risk is 0.446 indicates that the category predicted by the model (successful or unsuccessful student) is wrong for 44.6% of the cases. The CART classification matrix (Table 5) shows that model classify correctly 61% of students. This is slight increase in comparison to the CHAID model. The numbers of false positives (90) for the CART model decreases and therefore increasing the positive predictive value to 62.3%. In other words it will work better than the CHAID model at identifying an unsuccessful student. The price paid for increasing accuracy is reflected in decreasing sensitivity. The CART model will pick up 62.6% of all unsuccessful students (CHAID model 65.4%). At the same time the specificity will increase to 58.1% (CHAID model55.5%).

Concluding Remarks this study examines the background information from enrolment data that impacts upon the study outcome of Information Systems students at the Open Polytechnic. Based on results from feature selection (Figure 2 and Table 3), the CHAID tree presented in Table 4 and Figures 3 and 4, the CART trees presented in Table 5 and Figures 5 and 6 it was found that the most important factors that help separate successful from unsuccessful students are ethnicity, course programme and course block. Demographic data such as gender and age though significantly related to the study outcome, according to the feature selection result, were not used in the classification trees. Unfortunately the classification accuracy from the classification trees was not very high. In the case of the CHAID tree

the overall classification accuracy was 59.4% and in the case of the CART tree slightly higher at 60.5%. This would suggest that the background information (gender, age, ethnicity, disability, secondary school, work status, and early enrolment) gathered during the enrolment process, does not contain sufficient information for an accurately separation of successful and unsuccessful students. Classifying students based on pre-enrollment information and the rules presented for each node would allow the administrative and academic staff to identify students who would be "at risk" of dropping the course even before they start with their study. Then the student support systems, such as orientation, advising, and mentoring programs, could be used to positively impact the academic successes of such students.

This study is limited in three main ways that future research can perhaps address. Firstly, this research is based on background information only. Leaving out other important factors (academic achievement, number of courses completed, motivation, financial aids,etc.) that may affect study outcome, could distort results obtained with classification trees. For example, including the assignment mark after the submission of the first course assignment would probably improve predictive accuracy of the models. To improve the model, more attributes could be included to obtain prediction models with lower misclassification errors. However, the model in this case would not be a tool for pre-enrolment, i.e. early identification of 'at-risk' students. Secondly, we used a dichotomous variable for the study outcome with only two categories: pass and fail. However, splitting the fail category into those who stayed on the course but eventually failed the course and those who voluntary transfer or were withdrawn from the course would probably provide better profiling for each of the three categories of study outcomes. The only problem we might have with the three categories for study outcome is a low prediction accurate as a result of relatively small dataset for the course. Thirdly, from a methodological point of view an alternative to a classification tree should be considered. The prime candidates to be used with this data set are logistic regression and neural networks.

RESULT

Analysis done on the above case study, examined the background information from enrolment data that impacts upon the study outcome of Information Systems students at the Open Polytechnic.

DATA MINING CASE STUDY USING THE CRISP-DM STANDARD

AIM

CRISP- DM CASE STUDY – Analysing Automobile Warranty Claims: Example Of The CRISP– DM Industry Standard Process In Action Quality assurance continues to be apriority for automobile manufacturers, including Daimler Chrysler. Jochen Hipp of the University of Tubingen, Germany, and Guido Lindner of Daimler Chrysler AG, Germany, investigated patterns in the warranty claims for Daimler Chrysler automobiles.

1. Business Understanding Phase

Daimler Chrysler's objectives are to reduce costs associated with warranty claims and improve customer satisfaction. Through conversations with plant engineers, who are the technical experts in vehicle manufacturing, the researchers are able to formulate specific business problems, such as the following:

Are there inter dependencies among warranty claims?
Are past warranty claims associated with similar claims in the future?
Is there an association between a certain type of claim and a particular garage?
The plan is to apply appropriate data mining techniques to try to uncover these and other possible associations.

2. Data Understanding Phase

The researchers make use of Daimler Chrysler's Quality Information System (QUIS), which contains information on over 7 million vehicles and is about 40 gigabytes in size. QUIS contains production details about how and where a particular vehicle was constructed, including an average of 30 or more sales codes for each vehicle. QUIS also includes warranty claim information, which the garage supplies, in the form of one of more than 5000 possible potential causes. There searchers stressed the fact that the database was entirely unintelligible to domain non-experts: "So experts from different departments had to be located and consulted; in brief a task that turned out to be rather costly." They emphasis that analysts should not underestimate the importance, difficulty and potential cost of this early phase of the data mining process, and that shortcuts here may lead to expensive reiterations of the process downstream.

3. Data Preparation Phase

The researchers found that although relational, the QUIS database had limited SQL access. They needed to select the cases and variables of interest manually, and then manually derive new variables that could be used for the modelling phase. For example, the variable number of days from selling date until first claim had to be derived from the appropriate date attributes.

They then turned to proprietary data mining software, which had been used at Daimler Chrysler on earlier projects. Here they ran into a common road block –that the data format requirements varied from algorithm to algorithm. The result was further exhaustive pre-processing of the data, to transform the attributes into a form usable for model algorithms. The researchers mention that the data preparation phase took much longer than they had planned.

4. Modelling Phase

Since the overall business problem from phase 1 was to investigate dependence among the warranty claims, the researchers chose to apply the following techniques: (1) Bayesian networks and (2) Association rules. Bayesian networks model uncertainty by explicitly representing the conditional dependencies among various components, thus providing a graphical visualisation of the dependency relationships among the components. As such, Bayesian networks represent a natural choice for modelling dependence among warranty claims. Association rules are also a natural way to investigate dependence among warranty claims since the confidence measure represents a type of conditional probability, similar to Bayesian networks.

The details of the results are confidential, but we can get a general idea of the type of dependencies uncovered by the models. One insight the researchers uncovered was that a particular combination of construction specifications doubles the probability of encountering an automobile electrical cable problem. Daimler Chrysler engineers have begun to investigate how this combination of factors can result in an increase in cable problems

The researchers investigated whether certain garages had more warranty claims of a certain type than did other garages. Their association rule results showed that, indeed, the confidence levels for the rule "If garage *X*, then cable problem," varied considerably from garage to garage. They state that further in vestigation is warranted to reveal the reasons for the disparity.

5. Evaluation Phase

The researchers were disappointed that the support for sequential-type association rules was relatively small, thus precluding generalisation of the results, in their opinion. Overall, in fact, the researchers state: "In fact, we did not find any rule that our domain experts would judge as interesting, at least at first sight." According to this criterion, then, the models were found to be lacking in effectiveness and to fall short of the objectives set for the min the business under standing phase. To account for this, the researchers point to the "legacy" structure of the database, for which automobile parts were categorised by garages and factories for historic or technical reasons and not designed for data mining. They suggest adapting and redesigning the database to make it more amenable to knowledge discovery.

6. Deployment Phase

The researchers have identified the foregoing project as a pilot project, and as such, do not intend to deploy any large-scale models from this first iteration. After the pilot project, however, they have applied the lesson learned from this project, with the goal of integrating the methods with the existing information technology environment at Daimler Chrysler. To further support the original goal of lowering claims costs, they intend to develop an intranet offering mining capability of QUIS for all corporate employees.

DISCUSSION: Lessons drawn from this case study.

First, the general impression one draws is that uncovering hidden nuggets of knowledge in databases is a rocky road. In nearly every phase, the researchers ran into unexpected road blocks and difficulties. This tells us that actually applying data mining for the first time in a company requires asking people to do something new and different, which is not always welcome. Therefore, if they expect results, corporate management must be 100% supportive of new data mining initiatives. Another lesson to draw

is that intense human participation and supervision is required at every stage of the data mining process. For example, the algorithms require specific data formats, which may require substantial pre-processing. Regardless of what some software vendor advertisements may claim, you can't just purchase some data mining software, install it, sit back and watch it solve all your problems. Data mining is not magic. Without skilled human supervision, blind use of data mining software will only provide you with the wrong answer to the wrong question applied to the wrong type of data. The wrong analysis is worse than no analysis, since it leads to policy recommendations that will probably turn out to be expensive failures. Finally, from this case study we can draw the lesson that there is no guarantee of positive results when mining data for actionable knowledge, any more than when one is mining for gold. Data mining is not a panacea for solving business problems. But used properly, by people who understand the models involved, the data requirements and the overall project objectives, data mining can indeed provide actionable and highly profit able results.

RESULT

CRISP-DM CASE STUDY - Analysing Automobile Warranty Claims: Example Of The CRISP–DM Industry Standard Process In action is done and output verified.