

DATA WAREHOUSING AND DATA MINING LABORATORY

LIST OF EXPERIMENTS

	NAME OF THE EXPERIMENT
1.	Listing applications for mining.
2.	File format for data mining
3.	Conversion of various data files
4.	Training the given dataset for an application
5.	Testing the given dataset for an application
6.	Generating accurate models
7.	Feature selection
8.	Web mining
9.	Text mining
10.	Design of fact & dimension tables

INDEX

S.No.	NAME OF THE EXPERIMENT	Page No.
1.	Listing applications for mining.	
2.	File format for data mining	
3.	Conversion of various data files	
4.	Training the given dataset for an application	
5.	Testing the given dataset for an application	
6.	Generating accurate models	
7.	Feature selection	
8.	Web mining	
9.	Text mining	
10.	Design of fact & dimension tables	

EX.NO:1

LISTING APPLICATIONS FOR MINING

AIM:

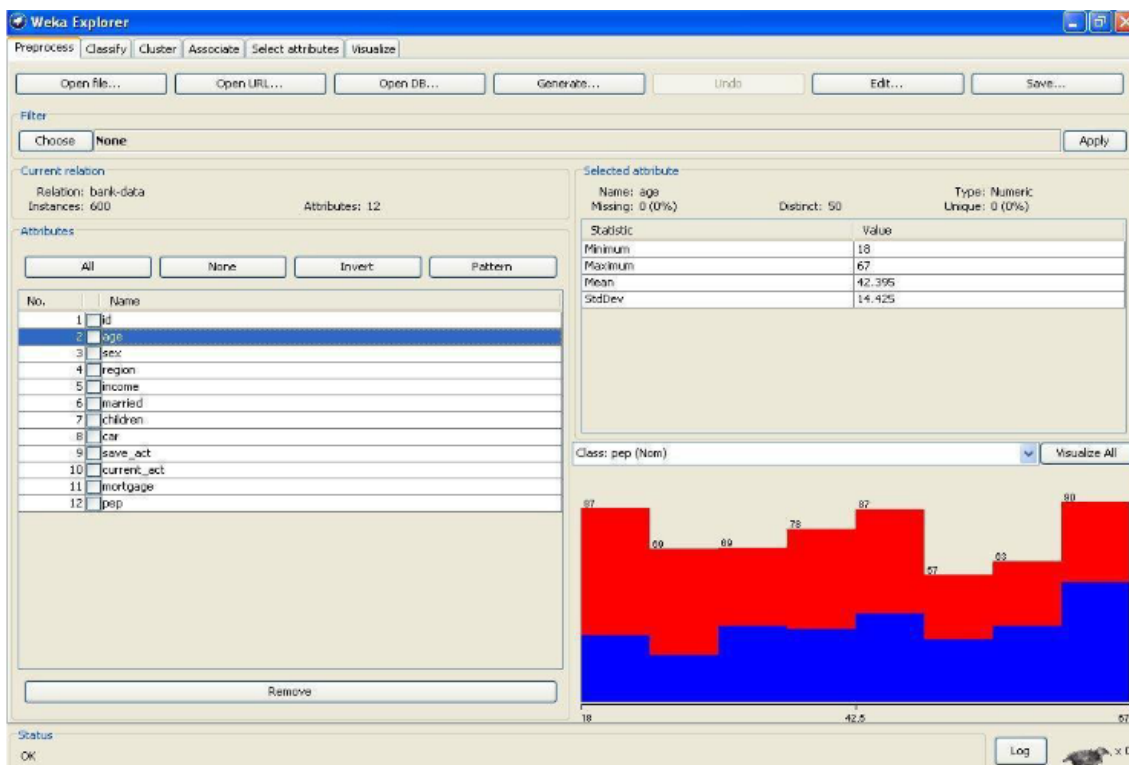
To list all the categorical (or nominal) attributes and the real-valued attributes separately.

RESOURCES: Weka mining tool1.

PROCEDURE:

- 1)Open the Weka GUI Chooser.
- 2)Select EXPLORER present in Applications.
- 3)Select Preprocess Tab.
- 4)Go to OPEN file and browse the file that is already stored in the system “bank.csv”.
- 5)Clicking on any attribute in the left panel will show the basic statistics on that selected attribute.1.4

OUTPUT:



Result:

Thus the listing applications for the data mining was studied.

EX.NO:2

FILE FORMAT FOR DATA MINING

Aim: To study the file formats for the data mining.

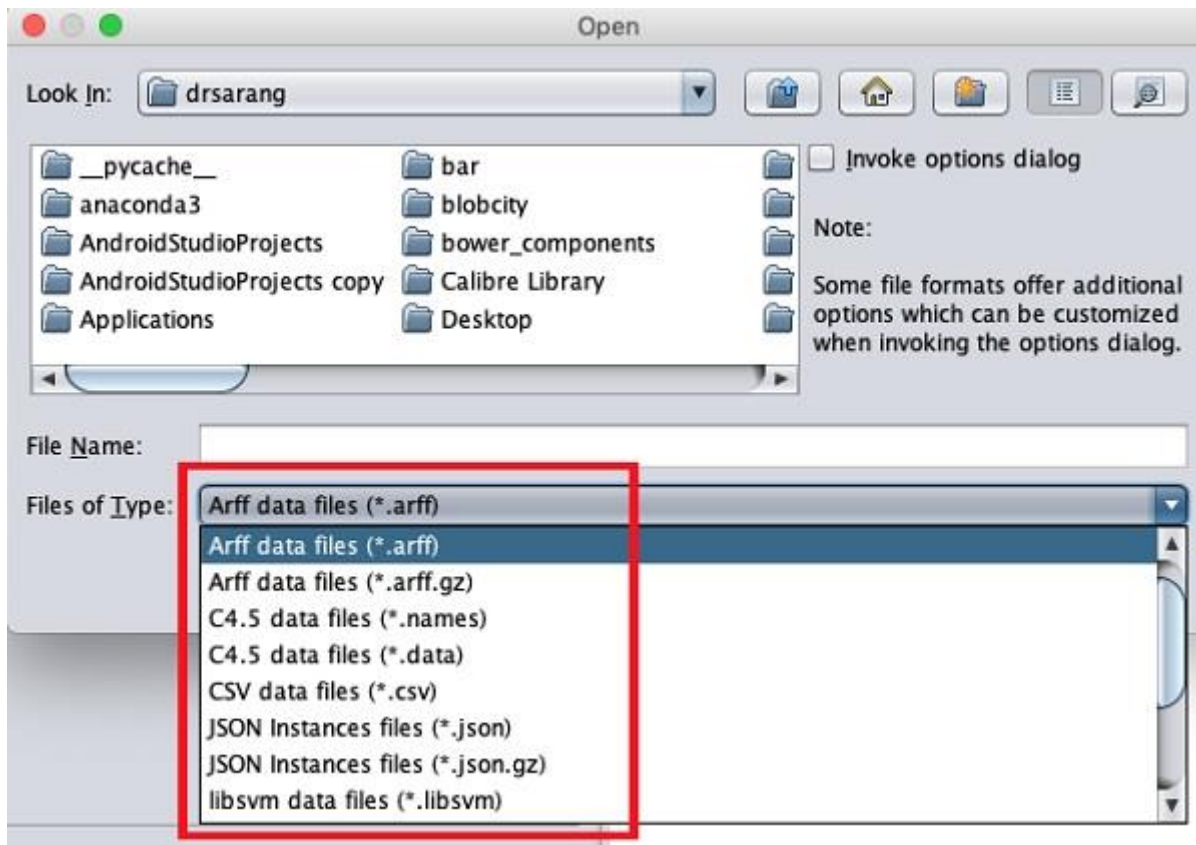
Introduction:

WEKA supports a large number of file formats for the data. The complete list of file formats are given here:

1. arff
2. arff.gz
3. bsi
4. csv
5. dat
6. data
7. json
8. json.gz
9. libsvm
10. m
11. names
12. xrff
13. xrff.gz

The types of files that it supports are listed in the drop-down list box at the bottom of the screen.

This is shown in the screenshot given below.



As you would notice it supports several formats including CSV and JSON.

The default file type is Arff.

Arff Format

An Arff file contains two sections - header and data.

The header describes the attribute types.

The data section contains a comma separated list of data.

As an example for Arff format, the Weather data file loaded from the WEKA sample databases is shown below:

```

@relation weather.symbolic
@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no

```

Diagram annotations:

- Dataset name**: Points to `@relation weather.symbolic`
- Attributes**: Points to the list of attributes: `@attribute outlook {sunny, overcast, rainy}`, `@attribute temperature {hot, mild, cool}`, `@attribute humidity {high, normal}`, `@attribute windy {TRUE, FALSE}`, and `@attribute play {yes, no}`
- Target / Class variable**: Points to the `play` attribute in the `@data` rows, which is underlined in red in the original image.
- Data Values**: Points to the individual data rows in the `@data` section.

From the screenshot, you can infer the following points –

The `@relation` tag defines the name of the database.

The `@attribute` tag defines the attributes.

The `@data` tag starts the list of data rows each containing the comma separated fields.

The attributes can take nominal values as in the case of outlook shown here –

```
@attribute outlook (sunny, overcast, rainy)
```

The attributes can take real values as in this case –

```
@attribute temperature real
```

You can also set a Target or a Class variable called play as shown here –

```
@attribute play (yes, no)
```

The Target assumes two nominal values yes or no.

Result:

Thus the different file formats for the data mining was studied.

EX.NO:3a**CONVERSION OF TEXT FILE INTO ARFF FILE****Aim:**

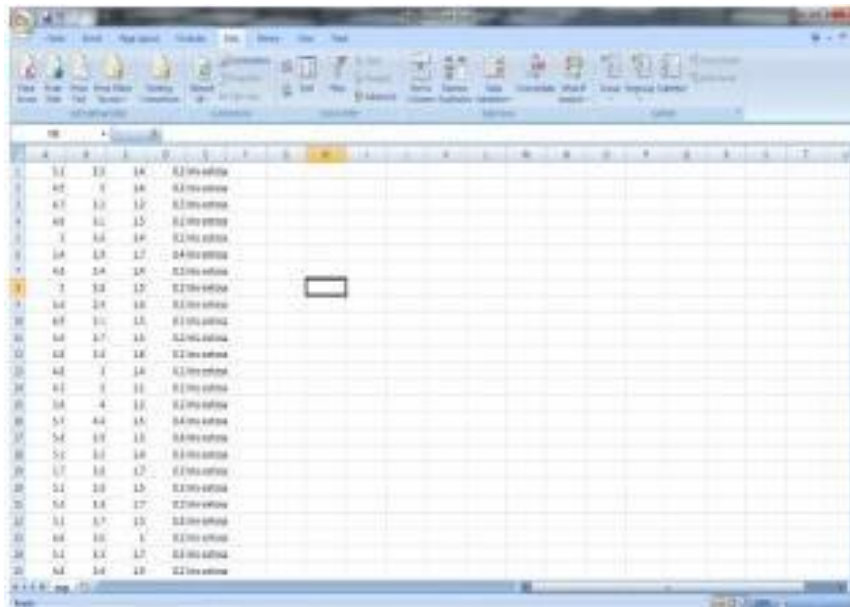
To convert a text file to ARFF(Attribute-Relation File Format) using Weka3.8.2 tool.

Objectives:

Most of the data that we have collected from public forum is in the text format that cannot be read by Weka tool. Since Weka (Data Mining tool) recognizes the data in ARFF format only we have to convert the text file into ARFF file.

Algorithm:

1. Download any data set from UCI data repository.
2. Open the same data file from excel. It will ask for delimiter (which produce column) in excel.
3. Add one row at the top of the data.
4. Enter header for each column.
5. Save file as .CSV (Comma Separated Values) format.
6. Open Weka tool and open the CSV file.
7. Save it as ARFF format.

Output:**Data Text File:**

1	5.1	3.5	14	0.1 (no setosa)
2	4.9	3	14	0.1 (no setosa)
3	4.7	3.2	13	0.1 (no setosa)
4	4.0	3.1	15	0.1 (no setosa)
5	5	3.6	14	0.1 (no setosa)
6	5.4	3.8	17	0.4 (no setosa)
7	4.6	3.4	16	0.2 (no setosa)
8	5	3.5	17	0.2 (no setosa)
9	5.4	3.9	16	0.1 (no setosa)
10	4.9	3.1	15	0.1 (no setosa)
11	5.4	3.7	15	0.2 (no setosa)
12	5.3	3.4	16	0.1 (no setosa)
13	4.8	3	14	0.1 (no setosa)
14	5.1	3	15	0.1 (no setosa)
15	5.4	4	15	0.1 (no setosa)
16	5.1	4.4	15	0.4 (no setosa)
17	5.4	3.9	15	0.4 (no setosa)
18	5.1	3.5	14	0.1 (no setosa)
19	5.7	3.8	17	0.1 (no setosa)
20	5.1	3.8	15	0.1 (no setosa)
21	5.6	3.8	17	0.2 (no setosa)
22	5.1	3.7	15	0.2 (no setosa)
23	4.9	3.6	1	0.1 (no setosa)
24	5.1	3.3	17	0.1 (no setosa)
25	4.8	3.4	18	0.2 (no setosa)

Data ARFF File:



Result:

Thus, conversion of a text file to ARFF(Attribute-Relation File Format) using Weka3.8.2 tool is implemented.

EX.NO:3b.

CONVERSION OF ARFF TO TEXT FILE

Aim:

To convert ARFF (Attribute-Relation File Format) into text file.

Objectives:

Since the data in the Weka tool is in ARFF file format we have to convert the ARFF file to text format for further processing.

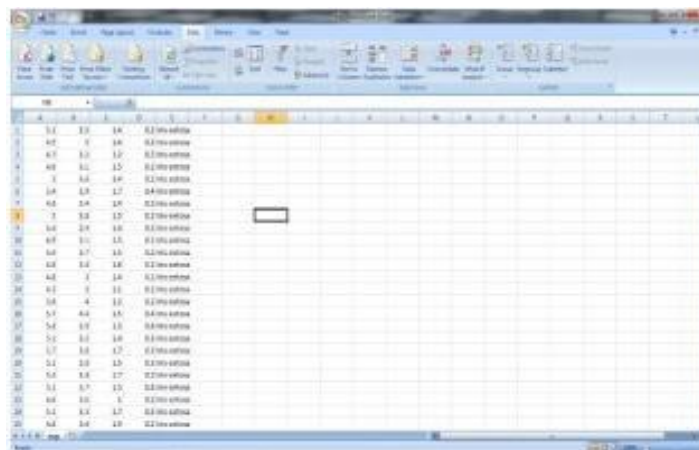
Algorithm:

1. Open any ARFF file in Weka tool.
2. Save the file as CSV format.
3. Open the CSV file in MS-EXCEL.
4. Remove some rows and add coreseponding header to the data.
5. Save it as text file with the desire delimiter.

Data ARFF File:



Data Text File:



Result: Thus conversion of ARFF (Attribute-Relation File Format) into text file is implemented.

TRAINING THE GIVEN DATASET FOR AN APPLICATION

To apply the concept of Linear Regression for training the given dataset.

1. Open the weka tool.
2. Download a dataset by using UCI.
3. Apply replace missing values.
4. Apply normalize filter.
5. Click the Classify Tab.
6. Choose the Simple Linear Regression option.
7. Select the training set of data.
8. Start the validation process.
9. Note the output.

In statistics, Linear Regression is an approach for modeling a relationship between a scalar dependent variable Y and one or more explanatory variables denoted X . the case of explanatory variable is called Simple Linear Regression.

Coefficient of Linear Regression is given by: $\mathbf{Y=ax+b}$

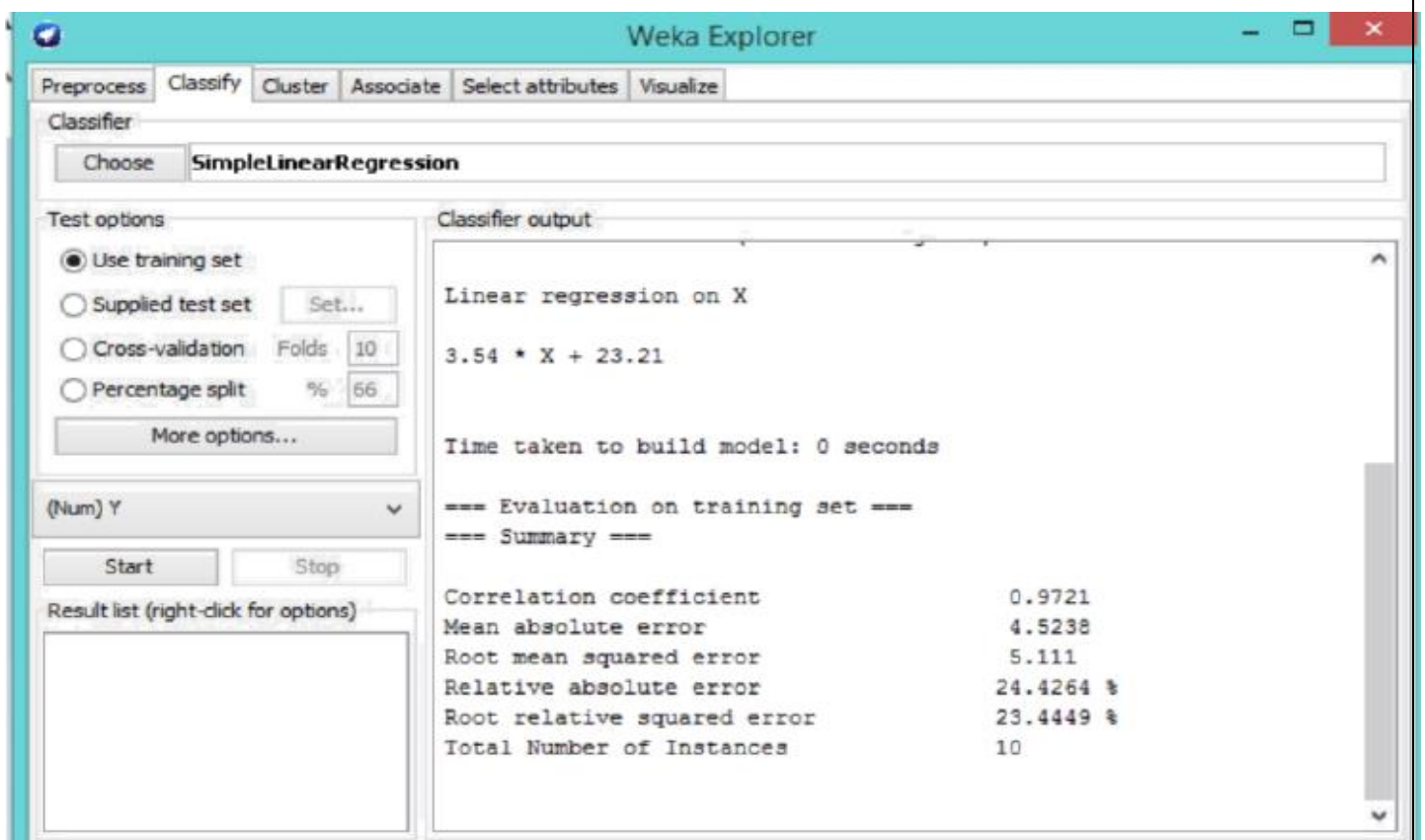
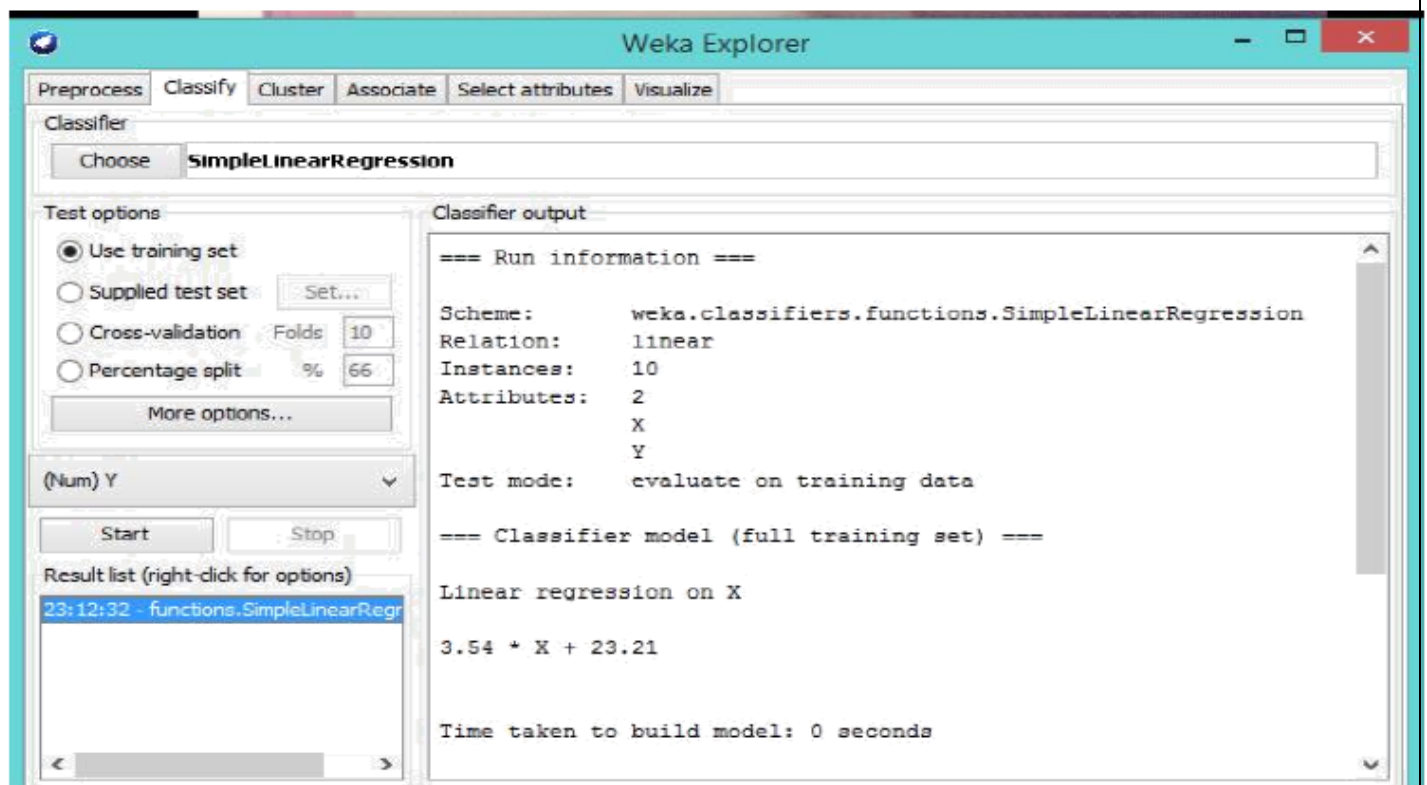
Consider the dataset below where x is the number of working experience of a college graduate and y is the corresponding salary of the graduate. Build a regression equation and predict the salary of college graduate whose experience is 10 years.

The screenshot shows the Microsoft Excel interface. The ribbon at the top includes tabs for File, Home, Insert, Page Layout, Formulas, Data, Review, and View. The Home tab is active, showing options for Clipboard, Font, Alignment, Number, Styles, Cells, and Editing. The formula bar at the top displays 'F5' and a function icon. The worksheet grid shows columns A through H and rows 1 through 12. The data in the grid is as follows:

	A	B	C	D	E	F	G	H
1	X	Y						
2		3	30					
3		8	57					
4		9	64					
5		13	72					
6		3	36					
7		6	43					
8		11	59					
9		21	90					
10		1	20					
11		16	83					
12								

The status bar at the bottom shows 'Ready', the active sheet is 'linear', and the zoom level is 100%.

Output:



Result: Thus the concept of Linear Regression for training the given dataset is applied and implemented.

TESTING THE GIVEN DATASET FOR AN APPLICATION

To apply the Navie Bayes Classification for testing the given dataset.

1. Open the weka tool.
2. Download a dataset by using UCI.
3. Apply replace missing values.
4. Apply normalize filter.
5. Click the Classification Tab.
6. Apply Navie Bayes Classification.
7. Find the Classified Value.
8. Note the output.

X is a data tuple. In Bayesian term it is considered “evidence”. H is some hypothesis that X belongs to a specified class C . $P(H|X)$ is the posterior probability of H conditioned on X .

Input Data:

[illegible]

Output data:

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section shows 'Supplied test set' is selected. The 'Classifier output' pane displays the following information:

--- Run information ---

Scheme: weka.classifiers.misc.InputMappedClassifier -I -trim -W weka.classifiers.bayes.1
Relation: nb
Instances: 14
Attributes: 5
age
income
student
credit
buys computer
Test mode: user supplied test set: 1 instances

--- Classifier model (full training set) ---

InputMappedClassifier:
Naive Bayes Classifier

Attribute	Class	
	no (0.38)	yes (0.63)
age		
youth	4.0	3.0
middle	1.0	5.0
senior	3.0	4.0
[total]	8.0	12.0
income		

The screenshot shows the Weka Explorer interface with the NaiveBayes classifier selected. The 'Test options' section shows 'Supplied test set' is selected. The 'Classifier output' pane displays the following information:

--- Run information ---

Scheme: weka.classifiers.misc.InputMappedClassifier -I -trim -W weka.classifiers.bayes.1
Relation: nb
Instances: 14
Attributes: 5
age
income
student
credit
buys computer
Test mode: user supplied test set: 1 instances

--- Classifier model (full training set) ---

InputMappedClassifier:
Naive Bayes Classifier

Attribute	Class	
	no (0.38)	yes (0.63)
age		
youth	4.0	3.0
middle	1.0	5.0
senior	3.0	4.0
[total]	8.0	12.0
income		
high	3.0	3.0
medium	3.0	5.0
low	2.0	4.0
[total]	8.0	12.0
student		
no	5.0	4.0
yes	2.0	7.0
[total]	7.0	11.0
credit		
fair	3.0	7.0
excellent	4.0	4.0
[total]	7.0	11.0

Attribute mappings:

Model attributes	Incoming attributes
(nominal) age	--> 1 (nominal) age
(nominal) income	--> 2 (nominal) income
(nominal) student	--> 3 (nominal) student
(nominal) credit	--> 4 (nominal) credit

Weka Explorer

Preprocess

Classify

Cluster

Associate

Select attributes

Visualize

Classifier

Choose

NaiveBayes

Test options

☐ Use training set

☒ Supplied test set

Set...

☐ Cross-validation

Folds

10

☐ Percentage split

%

66

More options...

(Nom) buys computer

Start

Stop

Result list (right-click for options)

17:44:02 - misc.InputMappedClassifier

Classifier output

Time taken to build model: 0 seconds

Evaluation on test set

Summary

Correctly Classified Instances

1

100

%

Incorrectly Classified Instances

0

0

%

Kappa statistic

1

Mean absolute error

0.1404

Root mean squared error

0.1404

Relative absolute error

37.4302

%

Root relative squared error

37.4302

%

Coverage of cases (0.95 level)

100

%

Mean rel. region size (0.95 level)

100

%

Total Number of Instances

1

Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0	?	no
1	1	0	1	1	1	?	yes
Weighted Avg.	1	0	1	1	1	0	

Confusion Matrix

a b

<-- classified as

0 0 | a = no

0 1 | b = yes

Result:

Thus the Navie Bayes Classification for testing the given dataset is implemented.

EX. No: 6

GENERATE ACCURATE MODEL

Aim:

To find the good result (by improving the performance) using the training set and testing data set for numerical values.

Objectives:

To develop training and testing data using numerical data set in order to get accurate model for classification.

ALGORITHM:

1. Download any data set.
2. Save the file with .ARFF format.
3. Apply 'Replace Missing Values' filter.
4. Normalize the values by applying normalize filter.
5. Go to unsupervised instance remove percentage
6. Right click on that (show properties) option then select 70% true and save it as training.arff
7. Select the original data set then right click on show properties then give 70% false and save it as testing.arff
8. Select classification and apply various algorithms.

TRAINING DATA:

The screenshot displays the Weka software interface. On the left, the 'Viewer' window shows a dataset with columns: ID, Last Name, First Name, City, State, Gender, Student Status, Major, Country, Age, SAT, and Average score (grade). The data is sorted by SAT score in descending order. On the right, a scatter plot window is visible, showing a plot of SAT score versus Average score (grade). The plot includes a color-coded legend and a 'Show' button. The main window shows a list of data points with their corresponding SAT and Average score values.

ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)
0.0	DOE01	JANE01	Los An...	Califor...	Female	Graduate	Politics	US	19.571...	1.952...	67.0
0.036...	DOE02	JANE02	Los An...	Arizona	Female	Undergraduate	Math	US	19.047...	1.687...	63.0
0.060...	DOE01	JOE01	El Paso	New Y...	Male	Graduate	Math	US	19.368...	1.909...	79.0
0.090...	DOE02	JOE02	Lubbock	New Y...	Female	Graduate	Econ	US	19.714...	1.985...	78.0
0.121...	DOE03	JOE03	Dallas	Ohio	Female	Graduate	Econ	US	19.904...	1.272...	66.0
0.131...	DOE04	JOE04	Tel Aviv	Israel	Male	Graduate	Econ	Israel	19.333...	1.461...	89.0
0.181...	DOE05	JOE05	Cheng	Florida	Male	Graduate	Politics	US	1.0	1.246...	95.0
0.212...	DOE03	JANE03	Liberal	Canada	Female	Undergraduate	Politics	US	19.142...	1.614...	87.0
0.242...	DOE04	JANE04	Montreal	Canada	Female	Undergraduate	Math	Canada	19.0	1.489...	91.0
0.272...	DOE05	JANE05	New Y...	New Y...	Female	Graduate	Math	US	19.714...	1.722...	71.0
0.302...	DOE06	JOE06	Hot C...	Mexico	Male	Undergraduate	Econ	US	19.463...	1.0	62.0
0.333...	DOE06	JANE06	Jove	Virginia	Female	Graduate	Math	US	19.952...	1.386...	79.0
0.363...	DOE07	JOE07	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	19.571...	1.307...	79.0
0.393...	DOE08	JOE08	Moscow	Russia	Male	Graduate	Politics	Russia	19.571...	1.579...	70.0
0.424...	DOE07	JANE07	Dunk...	New Y...	Female	Undergraduate	Math	US	19.142...	1.0	82.0
0.454...	DOE08	JANE08	Medic...	Utah	Female	Undergraduate	Econ	US	19.0	1.497...	80.0
0.484...	DOE09	JANE09	Ancst...	Holland	Female	Undergraduate	Math	Holland	19.047...	1.566...	75.0
0.515...	DOE10	JANE10	Mexico	Mexico	Female	Graduate	Politics	Mexico	19.615...	1.537...	65.0
0.545...	DOE11	JANE11	Caracas	Venez...	Female	Undergraduate	Math	Venez...	19.0	1.941...	92.0
0.575...	DOE09	JOE09	San Juan	Puerto...	Male	Graduate	Politics	US	19.194...	1.682...	95.0
0.606...	DOE12	JANE12	Romato	Chong	Female	Undergraduate	Econ	US	19.047...	1.486...	87.0
0.636...	DOE10	JOE10	New Y...	New Y...	Male	Undergraduate	Econ	US	19.142...	1.546...	82.0
0.666...	DOE13	JANE13	Thic C	Moscow	Female	Graduate	Politics	US	19.333...	1.441...	89.0
0.696...	DOE14	JANE14	Beijing	China	Female	Undergraduate	Math	China	19.0	1.514...	79.0
0.727...	DOE11	JOE11	Stockh...	Sweden	Male	Undergraduate	Politics	Sweden	19.047...	1.566...	88.0
0.757...	DOE12	JOE12	Ember...	Minnesota	Male	Graduate	Econ	US	19.196...	1.096...	90.0
0.787...	DOE13	JOE13	Inter...	Penns...	Male	Undergraduate	Math	US	19.092...	1.504...	88.0
0.818...	DOE15	JANE15	Loom	OKlah...	Female	Undergraduate	Econ	US	19.066...	1.0	64.0
0.848...	DOE14	JOE14	Buenos	Argen...	Male	Graduate	Politics	Argentina	19.571...	1.905...	83.0
0.878...	DOE15	JOE15	Adms	Louisiana	Male	Undergraduate	Econ	US	19.047...	1.585...	79.0
0.908...	DOE16	JANE16	Los An...	Califor...	Female	Graduate	Politics	US	19.571...	1.952...	67.0
0.938...	DOE17	JANE17	Sedona	Arizona	Female	Undergraduate	Math	US	19.047...	1.687...	93.0
0.969...	DOE18	JOE18	El Paso	New Y...	Male	Graduate	Math	US	19.368...	1.909...	79.0
1.0	DOE19	JOE19	Lubbock	New Y...	Male	Graduate	Econ	US	19.714...	1.386...	78.0

ZeroR:

The screenshot shows the Weka Explorer interface with the ZeroR classifier selected. The 'Test options' section on the left has 'Use training set' selected. The 'Classifier output' section on the right displays the following information:

Classifier output
Instances: 34
Attributes:
ID
Last Name
First Name
City
State
Gender
Student Status
Major
Country
Age
SAT
Average score (grade)
Test mode: evaluate on training data

==== Classifier model (full training set) ====
ZeroR predictions class values: 78,32382941176471
Time taken to build model: 0 seconds

==== Evaluation on training set ====
==== Summary ====

Correlation coefficient	0
Mean absolute error	0.1263
Root mean squared error	10.0285
Relative absolute error	100 %
Root relative squared error	100 %
Total Number of Instances	34

An 'Alert' window is visible on the right side of the screen, displaying a 'WARNING' message: 'Advanced System Protector has detected 256 items. It is highly recommended to clean them immediately.' with a 'Clean Now' button.

Ridor:

The screenshot shows the Weka Explorer interface with the Ridor classifier selected. The 'Test options' section on the left has 'Use training set' selected. The 'Classifier output' section on the right displays the following information:

Classifier output
Attributes:
Student Status
Major
Country
Age
SAT
Average score (grade)
Test mode: evaluate on training data

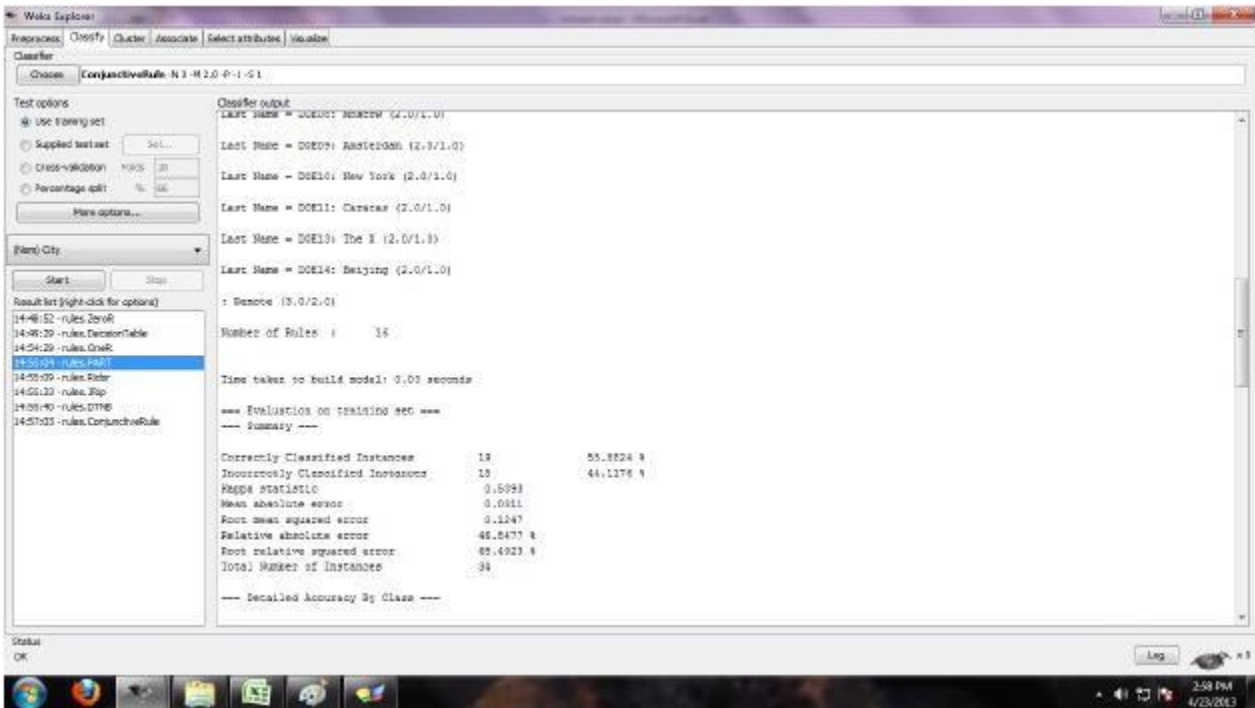
==== Classifier model (full training set) ====
Ripple Down Rule Learner(Ridor) rules
City = Los Angeles (34.0/0.0)
Total number of rules (incl. the default rule): 1
Time taken to build model: 0 seconds

==== Evaluation on training set ====
==== Summary ====

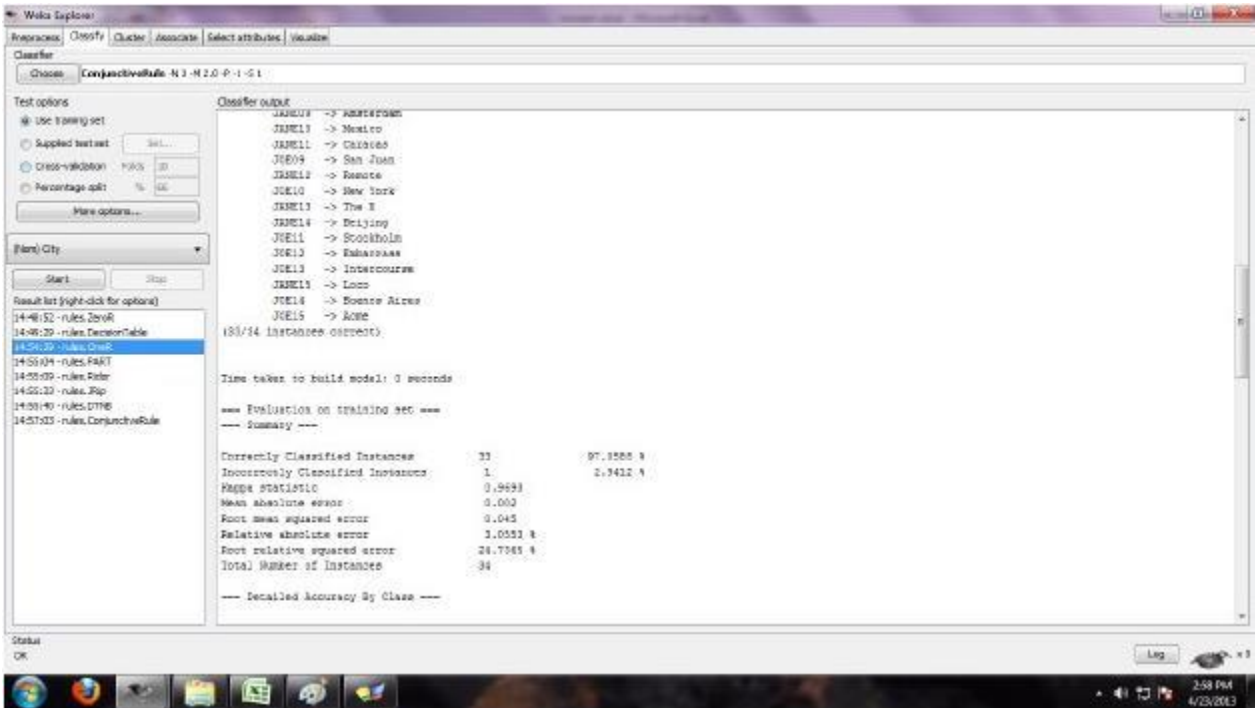
Correctly Classified Instances	3	8.8235 %
Incorrectly Classified Instances	31	91.1765 %
Ridge Statistic	0	
Mean absolute error	0.0424	
Root mean squared error	0.2501	
Relative absolute error	94.7139 %	
Root relative squared error	137.7271 %	
Total Number of Instances	34	

==== Detailed Accuracy By Class ====

PART:



OneR:



JRip:

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The 'Test options' section on the left shows 'Use training set' selected. The 'Classifier output' section on the right displays the following information:

Classifier output:

```

country
Age
SAT
Average score (grade)
Test mode:  evaluation on training data
=== Classifier model (full training set) ===
JRIP rules:
=====
(First Name = JER02) => City=Lackawanna (2.0/0.0)
=> City=Los Angeles (31.0/29.0)

Number of Rules : 2

Time taken to build model: 0.03 seconds

=== Evaluation on training set ===
--- Summary ---

Correctly Classified Instances      3          14.7059 %
Incorrectly Classified Instances    29          85.2941 %
Kappa statistic                    0.0669
Mean absolute error                0.0623
Root mean squared error            0.1764
Relative absolute error             93.7591 %
Root relative squared error         98.8957 %
Total Number of Instances          34

--- Detailed Accuracy By Class ---

```

The 'Result list' on the left shows a list of classifiers, with 'JRip' selected.

DTNB:

The screenshot shows the Weka Explorer interface with the DTNB classifier selected. The 'Test options' section on the left shows 'Use training set' selected. The 'Classifier output' section on the right displays the following information:

Classifier output:

```

major
Country
Age
SAT
Average score (grade)
Test mode:  evaluation on training data
=== Classifier model (full training set) ===
Decision Table:

Number of training instances: 34
Number of Rules : 34
100 instances covered by Majority class.
Evaluation (for feature selection): CV (leave one out)
Feature set: 1,3,4

Time taken to build model: 0.15 seconds

=== Evaluation on training set ===
--- Summary ---

Correctly Classified Instances      30          88.2353 %
Incorrectly Classified Instances     4          11.7647 %
Kappa statistic                    0.8771
Mean absolute error                0.0197
Root mean squared error            0.1442
Relative absolute error             29.9233 %
Root relative squared error         90.2521 %
Total Number of Instances          34

--- Detailed Accuracy By Class ---

```

The 'Result list' on the left shows a list of classifiers, with 'DTNB' selected.

TEST DATA:

Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier: **ConjunctiveRule - N 3 - M 2.0 - P 1 - G 1**

Test options

- ☒ Use training set
- ☐ Supplied test set: **Set...**
- ☐ Cross-validation: **Folds 10**
- ☐ Percentage split: **% 66**

More options...

(Name) City

Start Stop

Result list (right-click for options)

- 14:46:52 - rules.ZeroR
- 14:46:59 - rules.DecisionTable
- 14:54:29 - rules.OneR
- 14:55:14 - rules.PART
- 14:55:39 - rules.Rider
- 14:55:53 - rules.Zip
- 14:56:40 - rules.DTMB
- 14:57:03 - rules.ConjunctiveRule
- 15:00:17 - rules.ZeroR**
- 15:04:20 - rules.PART
- 15:04:46 - rules.OneR
- 15:05:02 - rules.Zip
- 15:05:09 - rules.DTMB
- 15:05:19 - rules.DecisionTable
- 15:05:24 - rules.ConjunctiveRule

Classifier output

first: name
City
State
Gender
Student Status
Major
Country
Age
SAT
Average score (grade)

Test mode: evaluation on training data

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

ZeroR predicts class values: Tel Aviv

Time taken to build model: 0 seconds

==== EVALUATION ON TRAINING SET ===

==== SUMMARY ===

Correctly Classified Instances	2	12.5	%
Incorrectly Classified Instances	14	87.5	%
Mispe percentage	0		
Mean absolute error	0.1167		
Root mean squared error	0.3413		
Relative absolute error	100	%	
Root relative squared error	100	%	
Total Number of Instances	16		

==== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
--	---------	---------	-----------	--------	-----------	----------	-------

Status: OK

Log

ZeroR:

Viewer

Viewer: test set with filters: unapplied, attributes: All, plot: Histogram

No	ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)
	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Numeric	Numeric
1	35.0	DOE20	DOE03	Tel Aviv	Ohio	Male	Graduate	Econ	US	37.0	1701.0	65.0
2	36.0	DOE21	DOE04	Tel Aviv	New Y...	Male	Graduate	Econ	Israel	25.0	1786.0	69.0
3	37.0	DOE22	DOE05	London	North...	Male	Graduate	Politics	US	38.0	1577.0	96.0
4	38.0	DOE23	DOE06	Liberal	Kansas	Male	Undergraduate	Politics	US	21.0	1640.0	87.0
5	39.0	DOE24	DOE07	Montreal	Canada	Female	Undergraduate	Math	Canada	18.0	1613.0	81.0
6	40.0	DOE25	DOE08	New Y...	New Y...	Female	Graduate	Math	US	35.0	2091.0	71.0
7	41.0	DOE26	DOE09	Hot C...	Norfolk...	Male	Undergraduate	Econ	US	18.0	1787.0	82.0
8	42.0	DOE27	DOE10	Davis	Virginia	Female	Graduate	Math	US	38.0	1513.0	79.0
9	43.0	DOE28	DOE11	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	39.0	1637.0	79.0
10	44.0	DOE29	DOE12	Moscow	Russia	Male	Graduate	Politics	Russia	38.0	1513.0	70.0
11	45.0	DOE30	DOE13	Durham	New Y...	Female	Undergraduate	Math	US	21.0	1528.0	83.0
12	46.0	DOE31	DOE14	McKen...	Utah	Female	Undergraduate	Econ	US	19.0	1621.0	80.0
13	47.0	DOE32	DOE15	Amster...	Holland	Female	Undergraduate	Math	Holland	18.0	1494.0	75.0
14	48.0	DOE33	DOE16	Mexico	Mexico	Female	Graduate	Politics	Mexico	31.0	2148.0	85.0
15	49.0	DOE34	DOE17	Elmira	New Y...	Male	Graduate	Math	US	28.0	2021.0	79.0
16	50.0	DOE35	DOE18	Lacka...	New Y...	Male	Graduate	Econ	US	33.0	1718.0	79.0

Edit... Save... Apply

Type: Numeric
Unique: 10 (100%)

Value:
35
80
42.5
4.761

Visualize All

Remove

Status: OK

Log

Ridor:

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The 'ConjunctiveRule: N3-M2.0-P-1-G1' classifier is chosen. The 'Test options' section shows 'Use training set' selected. The 'Classifier output' section displays the following results:

City = Tel Aviv (16.0/1.0)

Total number of rules (incl. the default rule): 1

Time taken to build model: 0 seconds

==== EVALUATION ON TRAINING SET ====

==== SUMMARY ====

Metric	Value	Class
Correctly Classified Instances	2	12.5 %
Incorrectly Classified Instances	14	87.5 %
Mappa statistic	0	
Mean absolute error	0.1094	
Root mean squared error	0.3307	
Relative absolute error	93.7235 %	
Root relative squared error	137.055 %	
Total Number of Instances	16	

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class

PART:

The screenshot shows the Weka Explorer interface with the 'Classifier' tab selected. The 'ConjunctiveRule: N3-M2.0-P-1-G1' classifier is chosen. The 'Test options' section shows 'Use training set' selected. The 'Classifier output' section displays the following results:

Student Status = Graduate AND
ID <= 0.466667: Montreal (2.0/1.0)

Student Status = Graduate AND
ID <= 0.512222: Cinnak (1.0/1.0)

Gender = Male AND
ID <= 0.466667: Liberal (2.0/1.0)

ID <= 0.712222: Moscow (2.0/1.0)

: Amsterdam (2.0/1.0)

Number of Rules: 1

Time taken to build model: 0 seconds

==== EVALUATION ON TRAINING SET ====

==== SUMMARY ====

Metric	Value	Class
Correctly Classified Instances	0	50 %
Incorrectly Classified Instances	0	50 %
Mappa statistic	0.4039	
Mean absolute error	0.0461	
Root mean squared error	0.1604	
Relative absolute error	55.765 %	
Root relative squared error	74.7698 %	
Total Number of Instances	16	

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class

OneR:

The screenshot shows the Weka Explorer interface with the OneR classifier selected. The 'Test options' section on the left has 'Use training set' selected. The 'Classifier output' pane on the right displays the results of the classification.

Classifier output:

```
OneR -> L1DTree
OneR -> Montreal
OneR -> New York
OneR -> Hot Coffee
OneR -> Java
OneR -> Varna
OneR -> Moscow
OneR -> Disneyland
OneR -> Mexico City
OneR -> Mexico
OneR -> Elmer
OneR -> Lockdowns
(16/16 instances correct)
```

Time taken to build model: 0 seconds

==== EVALUATION ON TRAINING SET ====

==== Summary ====

Metric	Value
Correctly Classified Instances	16
Incorrectly Classified Instances	0
Weighted Average	1.0
Mean Absolute Error	0
Root Mean Squared Error	0
Relative Absolute Error	0
Root Relative Squared Error	0
Total Number of Instances	16

==== Detailed Accuracy By Class ====

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Class						

JRip:

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The 'Test options' section on the left has 'Use training set' selected. The 'Classifier output' pane on the right displays the results of the classification.

Classifier output:

```
Country
Age
SAT
Average score (grade)
Test mode:
evaluate on training data
```

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

JRIP rules:

```
-----
-> City=San Jose (16.0/14.0)
```

Number of Rules: 1

Time taken to build model: 0.01 seconds

==== Evaluation on training set ====

==== Summary ====

Metric	Value
Correctly Classified Instances	2
Incorrectly Classified Instances	14
Weighted Average	0.125
Mean Absolute Error	0.2188
Root Mean Squared Error	0.2411
Relative Absolute Error	99.5816 %
Root Relative Squared Error	99.0951 %
Total Number of Instances	16

==== Detailed Accuracy By Class ====

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Class						

DTNB:

The screenshot shows the Weka Explorer window with the 'Classifier' tab selected. The 'ConjunctiveRule' classifier is chosen, and the 'Test options' are set to 'Use training set'. The 'Classifier output' pane displays the following information:

Classifier output

Country
Age
SAT
Average score (grades)
Test mode: evaluate on training data

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

Decision table:

Number of training instances: 16
Number of Rules : 15
Test machines covered by Majority class:
Evaluation (for feature selection): CV (leave one out)
Feature set: 1,4
Time taken to build model: 0.04 seconds

=== EVALUATION ON TRAINING SET ===

=== SUMMARY ===

Correctly Classified Instances	15	93.75 %
Incorrectly Classified Instances	1	6.25 %
Magnitude statistic	0.5625	
Mean absolute error	0.09375	
Root mean squared error	0.208	
Relative absolute error	0.7359 %	
Root relative squared error	0.4135 %	
Total Number of Instances	16	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
---------	---------	-----------	--------	-----------	----------	-------

The 'Result list (right-click for options)' on the left shows the following entries:

- 14:46:52 - rules.ZeroR
- 14:46:59 - rules.DecisionTable
- 14:54:29 - rules.ZeroR
- 14:55:04 - rules.PART
- 14:55:09 - rules.Rider
- 14:55:12 - rules.Rip
- 14:55:40 - rules.DTNB
- 14:57:23 - rules.ConjunctiveRule
- 15:04:17 - rules.ZeroR
- 15:04:25 - rules.Rider
- 15:04:34 - rules.PART
- 15:04:46 - rules.ZeroR
- 15:05:01 - rules.Rip
- 15:05:19 - rules.DecisionTable
- 15:05:24 - rules.ConjunctiveRule

The 'Status' bar at the bottom indicates 'OK'.

Result :

Thus, the good result (by improving the performance) using the training set and testing data set for numerical values is found out.

EX. No: 7

FEATURE SELECTION

AIM:

To find the good results by feature selection.

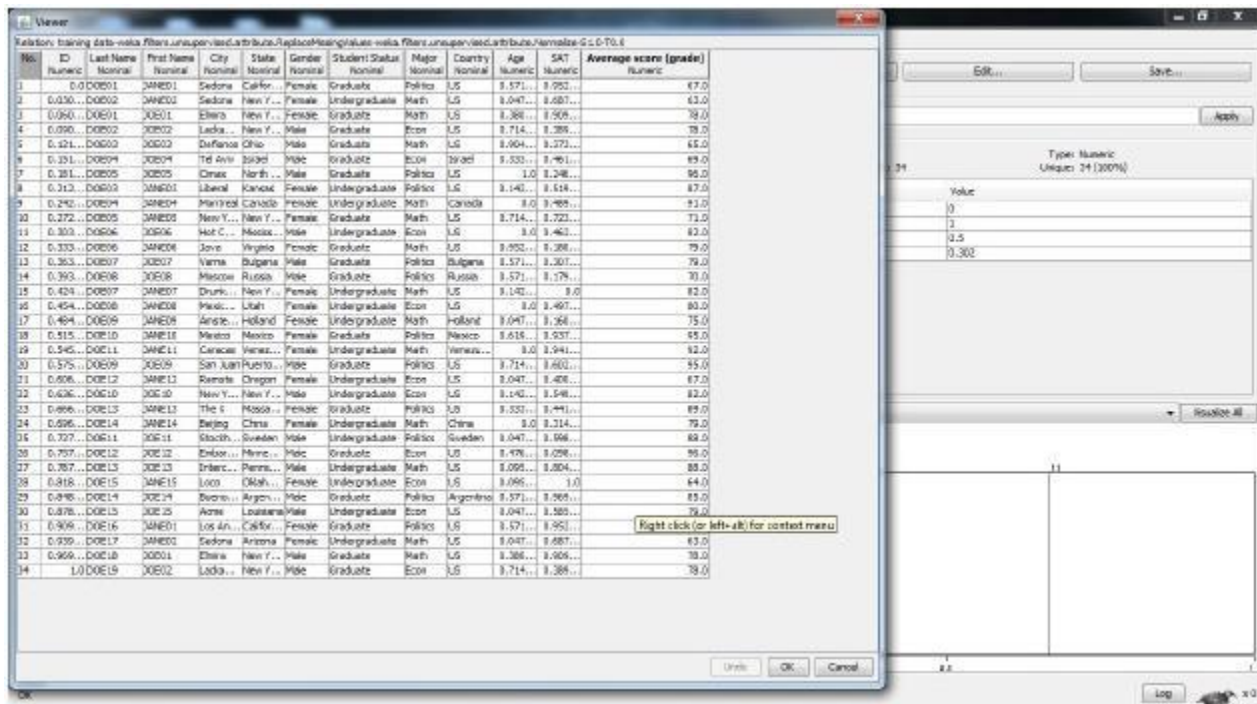
OBJECTIVES:

Any classifier/model has internal feature, those feature gives more accurate and optimal result.

ALGORITHM:

1. Download any dataset with nominal values.
2. Save it as text.arff.
3. Split it into training and testing data set.
4. Go to unsupervised instance remove percentage.
5. Right click on that show properties then select 70% true and save it as training.arff
6. Right click on that show properties then select 70% false and save it as testing.arff using original data set.
7. Open the parameter for classifying.
8. Fix the set of changing values.
9. Look at the performance.
10. Go to step 3 until the expected values of maximum value is reached.

Training Data:



The screenshot shows the Weka Explorer interface. The main window displays a dataset with 14 instances. The columns are: ID, Last Name, First Name, City, State, Gender, Student Status, Major, Country, Age, SAT, and Average score (grade). The 'Average score (grade)' column is highlighted, and a properties dialog is open for it. The dialog shows the attribute is of type 'Numeric' and has 14 unique values (100%). The 'Value' field is set to 0, and the 'Uniqueness' is 1.0. The 'Right click (or left+alt) for context menu' tooltip is visible over the 'Average score (grade)' column.

ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)
1	DOE01	JANE01	Sedona	Calif...	Female	Graduate	Politics	US	19.571...	19.952...	67.0
2	DOE02	JANE02	Sedona	New Y...	Female	Undergraduate	Math	US	19.047...	19.687...	63.0
3	DOE03	JANE03	El Paso	New Y...	Female	Graduate	Math	US	19.388...	19.508...	78.0
4	DOE04	JANE04	Lubbock	New Y...	Male	Graduate	Econ	US	19.714...	19.388...	78.0
5	DOE05	JANE05	Durham	Ohio	Male	Graduate	Math	US	19.904...	19.772...	65.0
6	DOE06	JANE06	Telluride	Utah	Male	Graduate	Econ	Utah	19.533...	19.461...	69.0
7	DOE07	JANE07	Omaha	Nebraska	Male	Graduate	Politics	US	19.0...	19.548...	86.0
8	DOE08	JANE08	Liberal	Kansas	Female	Undergraduate	Politics	US	19.142...	19.514...	67.0
9	DOE09	JANE09	Montreal	Canada	Female	Undergraduate	Math	Canada	19.0...	19.488...	71.0
10	DOE10	JANE10	New York	New York	Female	Graduate	Math	US	19.714...	19.722...	71.0
11	DOE11	JANE11	Hot Springs	Missouri	Male	Undergraduate	Econ	US	19.0...	19.462...	82.0
12	DOE12	JANE12	Jaysville	Virginia	Female	Graduate	Math	US	19.952...	19.388...	79.0
13	DOE13	JANE13	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	19.571...	19.307...	79.0
14	DOE14	JANE14	Moscow	Russia	Male	Graduate	Politics	Russia	19.571...	19.178...	70.0
15	DOE15	JANE15	Drunk	New York	Female	Undergraduate	Math	US	19.142...	19.0...	82.0
16	DOE16	JANE16	Utah	Utah	Female	Undergraduate	Econ	US	19.0...	19.467...	80.0
17	DOE17	JANE17	Amsterdam	Holland	Female	Undergraduate	Math	Holland	19.047...	19.366...	75.0
18	DOE18	JANE18	Mexico	Mexico	Female	Graduate	Politics	Mexico	19.618...	19.537...	69.0
19	DOE19	JANE19	Genoa	Italy	Female	Undergraduate	Math	Italy	19.0...	19.544...	82.0
20	DOE20	JANE20	San Juan	Puerto Rico	Male	Graduate	Politics	US	19.714...	19.602...	85.0
21	DOE21	JANE21	Ramona	Oregon	Female	Undergraduate	Econ	US	19.047...	19.408...	67.0
22	DOE22	JANE22	New York	New York	Male	Undergraduate	Econ	US	19.142...	19.548...	82.0
23	DOE23	JANE23	The City	Moscow	Female	Graduate	Politics	US	19.533...	19.441...	69.0
24	DOE24	JANE24	Beijing	China	Female	Undergraduate	Math	China	19.0...	19.314...	79.0
25	DOE25	JANE25	Stockholm	Sweden	Male	Undergraduate	Politics	Sweden	19.047...	19.566...	88.0
26	DOE26	JANE26	Emeryville	California	Male	Graduate	Econ	US	19.476...	19.098...	86.0
27	DOE27	JANE27	Interlaken	Switzerland	Male	Undergraduate	Math	US	19.092...	19.804...	88.0
28	DOE28	JANE28	Locust	Oklahoma	Female	Undergraduate	Econ	US	19.047...	19.0...	64.0
29	DOE29	JANE29	Buenos Aires	Argentina	Male	Graduate	Politics	Argentina	19.571...	19.668...	85.0
30	DOE30	JANE30	Alma	Louisiana	Male	Undergraduate	Econ	US	19.047...	19.565...	86.0
31	DOE31	JANE31	Los Angeles	California	Female	Graduate	Politics	US	19.571...	19.952...	67.0
32	DOE32	JANE32	Sedona	Arizona	Female	Undergraduate	Math	US	19.047...	19.687...	63.0
33	DOE33	JANE33	El Paso	New York	Male	Graduate	Math	US	19.388...	19.508...	78.0
34	DOE34	JANE34	Lubbock	New York	Male	Graduate	Econ	US	19.714...	19.388...	78.0

JRip(seed=1):

Weka Explorer

Reprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier: **JRip** F 2 N 2.0 O 2 S 1

Test options:

- ☒ Use training set
- ☐ Supplied test set
- ☐ Cross-validation: folds: 10
- ☐ Percentage split: % 100

File: City

Start | Stop

Result list (right-click for options):

- 20:54:52 - rules.Rip

Classifier output:

JRip rules:

```

=====
(First Rule = 10001) => City=Lackawanna (2.0/0.0)
(First Rule = 10001) => City=Elmira (2.0/0.0)
=> City=Sedona (33.0/27.0)
=====

```

Number of Rules: 1

Time takes to build model: 0.04 seconds

=== Evaluation on training set ===

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	7	20.5882 %
Incorrectly Classified Instances	27	79.4118 %
Kappa statistic	0.1323	
Mean absolute error	0.0503	
Root mean squared error	0.1707	
Relative absolute error	57.7056 %	
Root relative squared error	93.7598 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0.561	Los Angeles
1	0.871	0.1	1	0.182	0.563	Sedona
1	0	1	1	1	1	Elmira
1	0	1	1	1	1	Lackawanna
0	0	0	0	0	0.561	Defiance
0	0	0	0	0	0.561	Tal Arzu
0	0	0	0	0	0.561	Climax

Status: OK

JRip(seed=2):

Weka Explorer

Reprocess | Classify | Cluster | Associate | Select attributes | Visualize

Classifier: **JRip** F 3 S 1 N 2.0

Test options:

- ☒ Use training set
- ☐ Supplied test set
- ☐ Cross-validation: folds: 10
- ☐ Percentage split: % 100

File: City

Start | Stop

Result list (right-click for options):

- 20:54:52 - rules.Rip
- 20:55:15 - rules.Rip
- 20:55:34 - rules.Rip
- 20:55:53 - rules.Rip
- 21:00:03 - rules.Rip

Classifier output:

JRip rules:

```

=====
(First Rule = 20001) => City=Lackawanna (3.0/0.0)
=> City=Sedona (31.0/29.0)
=====

```

Number of Rules: 2

Time takes to build model: 0.02 seconds

=== Evaluation on training set ===

=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	5	14.7059 %
Incorrectly Classified Instances	29	85.2941 %
Kappa statistic	0.0663	
Mean absolute error	0.0622	
Root mean squared error	0.1764	
Relative absolute error	60.7561 %	
Root relative squared error	96.6957 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0.561	Los Angeles
1	0.871	0.1	1	0.182	0.563	Sedona
1	0	1	1	1	1	Elmira
1	0	1	1	1	1	Lackawanna
0	0	0	0	0	0.561	Defiance
0	0	0	0	0	0.561	Tal Arzu
0	0	0	0	0	0.561	Climax

Status: OK

JRip(seed=3):

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The 'Test options' section is configured with 'Use training set' selected. The 'Classifier output' pane displays the following information:

Student Status
Major
Country
Age
SAT
Average score (grade)
Test mode: evaluate on training data

--- Classifier model (full training set) ---

JRIP rules:
=====

(First Name = JORDI) => City=LosAngeles (3.0/0.0)
=> City=Seattle (32.0/29.0)

Number of Rules : 2

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===
=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	3	14.2857 %
Incorrectly Classified Instances	18	85.7143 %
Kappa statistic	0.0683	
Mean absolute error	0.0622	
Root Mean Squared Error	0.1764	
Relative absolute error	91.7491 %	
Root relative squared error	96.8957 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

STATUS: OK

Ridor(seed=1):

The screenshot shows the Weka Explorer interface with the Ridor classifier selected. The 'Test options' section is configured with 'Use training set' selected. The 'Classifier output' pane displays the following information:

Student Status
Major
Country
Age
SAT
Average score (grade)
Test mode: evaluate on training data

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

RIDGE DOWN STATE LEARNER(RIDGE) rules

City = Seattle (34.0/0.0)

Total number of rules (incl. the default rule): 1

Time taken to build model: 0 seconds

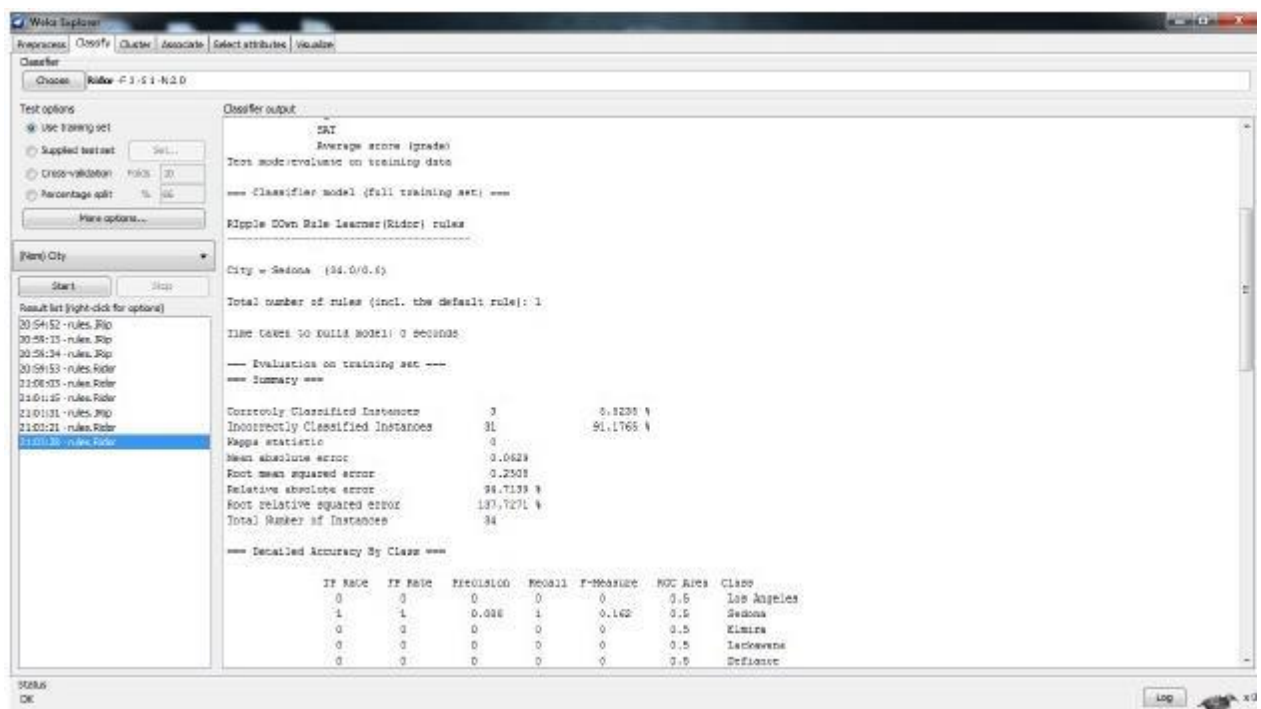
=== Evaluation on training set ===
=== Summary ===

Metric	Value	Percentage
Correctly Classified Instances	3	8.8235 %
Incorrectly Classified Instances	31	91.1765 %
Kappa statistic	0	
Mean absolute error	0.0628	
Root Mean Squared Error	0.2308	
Relative absolute error	94.7139 %	
Root relative squared error	137.7271 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0.5	Los Angeles

STATUS: OK

Ridor(seed=2):

Test Data:

Relation: last:st:vals:filter.unsupervised.attribute.ReplaceMissing/vals:vals:filter.unsupervised.attribute.Normalize-G1:0-T0:8

No.	ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)
	Numeric	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Numeric	Numeric	Numeric
1	0.0	Doe20	Jane20	Tel Aviv	Ohio	Male	Graduate	Econ	US	1.904...	1.398...	65.0
2	0.956...	Doe21	Jane21	Tel Aviv	New Y...	Male	Graduate	Econ	Israel	1.332...	1.464...	69.0
3	0.333...	Doe22	Jane22	Urbek	New Y...	Female	Graduate	Politics	US	1.0	1.261...	96.0
4	0.3	Doe23	Jane23	Urbek	France	Female	Graduate	Politics	US	1.140...	1.553...	87.0
5	0.366...	Doe24	Jane24	Montreal	Canada	Female	Undergraduate	Econ	Canada	8.0	1.521...	81.0
6	0.333...	Doe25	Jane25	New Y...	New Y...	Female	Graduate	Math	US	1.714...	1.775...	73.0
7	0.400	Doe26	Jane26	Hot C...	Massa...	Male	Undergraduate	Econ	US	8.0	1.462...	82.0
8	0.466...	Doe27	Jane27	Dava	Virginia	Female	Graduate	Math	US	1.962...	1.962...	76.0
9	0.533...	Doe28	Jane28	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	1.571...	1.328...	79.0
10	0.6	Doe29	Jane29	Moscow	Russia	Male	Graduate	Politics	Russia	1.571...	1.391...	70.0
11	0.666...	Doe30	Jane30	Druski...	New Y...	Female	Undergraduate	Math	US	1.140...	1.0	82.0
12	0.753...	Doe31	Jane31	Mexico	Utah	Female	Undergraduate	Econ	US	8.0	1.526...	80.0
13	0.8	Doe32	Jane32	Americ...	Holland	Female	Undergraduate	Math	Holland	1.047...	1.271...	75.0
14	0.866...	Doe33	Jane33	Mexico	Mexico	Female	Graduate	Politics	Mexico	1.615...	1.0	65.0
15	0.933...	Doe34	Jane34	El Paso	New Y...	Male	Graduate	Math	US	1.386...	1.978...	78.0
16	1.0	Doe35	Jane35	Lacka...	New Y...	Male	Graduate	Econ	US	1.714...	1.415...	78.0

Links OK Cancel

JRip(seed=1):

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier: **RuleF 1-51-N2.0**

Test options:

- ☐ Use training set
- ☒ Supplied test set **Set...**
- ☐ Cross-validation: Folds: 10
- ☐ Percentage split: %: 66
- More options...**

File(s): City

Start **Stop**

Result list (right-click for options):

- 11:04:55 - rules.Rip**
- 11:07:05 - rules.Rip
- 11:07:20 - rules.Rip
- 11:07:20 - rules.Rip
- 11:07:27 - rules.Rip

Classifier output:

Student Status
Major
Country
Age
SAT
Average score (grade)

Test mode: user supplied test set: size unknown (reading incrementally)

=== Classifier Model (full training set) ===

JRIP rules:
=====

=> C10:=Jel AND V (16.0/14.0)

Number of Rules : 1

Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	1	6.25 %
Incorrectly Classified Instances	15	93.75 %
Kappa statistic	0	
Mean absolute error	0.1172	
Root mean squared error	0.2431	
Relative absolute error	100 %	
Root relative squared error	100.8114 %	
Total Number of Instances	16	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class

STATUS: OK

Log

JRip(seed=2):

Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

Classifier: **RuleF 1-51-N2.0**

Test options:

- ☐ Use training set
- ☒ Supplied test set **Set...**
- ☐ Cross-validation: Folds: 10
- ☐ Percentage split: %: 66
- More options...**

File(s): City

Start **Stop**

Result list (right-click for options):

- 11:04:55 - rules.Rip**
- 11:07:05 - rules.Rip
- 11:07:20 - rules.Rip
- 11:07:20 - rules.Rip
- 11:07:27 - rules.Rip

Classifier output:

Student Status
Major
Country
Age
SAT
Average score (grade)

Test mode: user supplied test set: size unknown (reading incrementally)

=== Classifier Model (full training set) ===

JRIP rules:
=====

=> C10:=Jel AND V (16.0/14.0)

Number of Rules : 1

Time taken to build model: 0.02 seconds

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	1	6.25 %
Incorrectly Classified Instances	15	93.75 %
Kappa statistic	0	
Mean absolute error	0.1172	
Root mean squared error	0.2431	
Relative absolute error	100 %	
Root relative squared error	100.8114 %	
Total Number of Instances	16	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class

STATUS: OK

Log

JRip(seed=3):

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The 'Test options' section on the left indicates that a 'Supplied test set' is used. The 'Classifier output' pane on the right displays the following information:

Age
SAT
Average score (grades)
Test mode: user supplied test set; size unknown (reading incrementally)

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

JRIP rules:
=====

=> City=Tel Aviv (16.0/14.0)

Number of Rules : 1

Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Total Number of Instances
1	15	0	0.1172	0.2431	100	100.0114	16

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0.5	Defiance
1	1	0.063	1	0.118	0.5	Tel Aviv
0	0	0	0	0	0.5	Class

Ridor(seed=1):

The screenshot shows the Weka Explorer interface with the Ridor classifier selected. The 'Test options' section on the left indicates that a 'Supplied test set' is used. The 'Classifier output' pane on the right displays the following information:

Average score (grades)
Test mode: user supplied test set; size unknown (reading incrementally)

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

Ripple Down Rule Learner(Ridor) rules
=====

City = Tel Aviv (16.0/14.0)

Total number of rules (incl. the default rule): 1

Time taken to build model: 0 seconds

=== Evaluation on test set ===

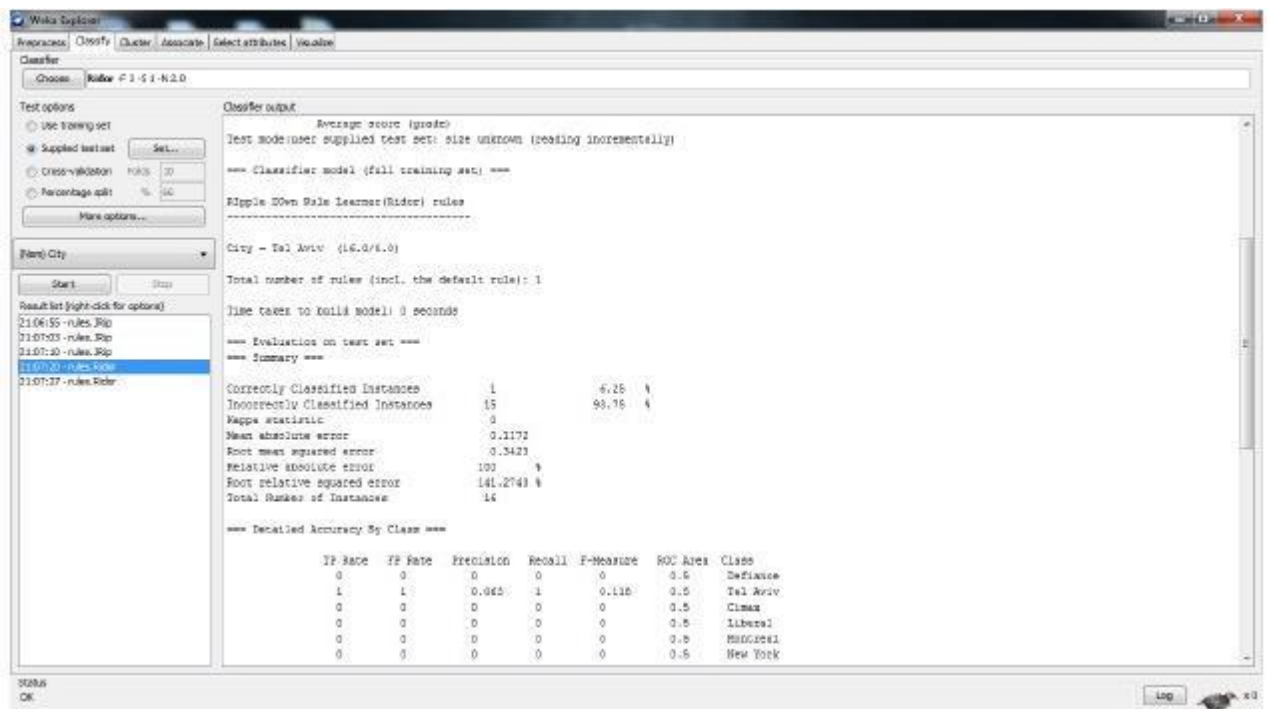
=== Summary ===

Correctly Classified Instances	Incorrectly Classified Instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error	Total Number of Instances
1	15	0	0.1172	0.2423	100	141.2743	16

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0.5	Defiance
1	1	0.062	1	0.118	0.5	Tel Aviv
0	0	0	0	0	0.5	Class
0	0	0	0	0	0.5	Liberal
0	0	0	0	0	0.5	Misc/Great
0	0	0	0	0	0.5	New York

Ridor(seed=2):



Training Data Set Performance:

TRAINING SET		
CLASSIFIER	PARAMETER SETTING	PERFORMANCE
JRip	Seed=1	Root Mean Squared Error=0.1707 Mean Absolute Error=0.0583
JRip	Seed =2	Root Mean Squared Error=0.1764 Mean Absolute Error=0.0622
JRip	Seed =3	Root Mean Squared Error=0.1764 Mean Absolute Error=0.0622
Ridor	Seed =1	Root Mean Squared Error=0.2508 Mean Absolute Error=0.0629
Ridor	Seed=2	Root Mean Squared Error=0.2508 Mean Absolute Error=0.0629

Testing Data set Performance:

TEST SET		
CLASSIFIER	PARAMETER SETTING	PERFORMANCE
<u>JRip</u>	Seed=1	Root Mean Squared Error=0.2431 Mean Absolute Error=0.1172
<u>JRip</u>	Seed =2	Root Mean Squared Error=0.2431 Mean Absolute Error=0.1172
<u>JRip</u>	Seed =3	Root Mean Squared Error=0.2431 Mean Absolute Error=0.1172
<u>Ridor</u>	Seed =1	Root Mean Squared Error=0.3423 Mean Absolute Error=0.1172
<u>Ridor</u>	Seed=2	Root Mean Squared Error=0.3423 Mean Absolute Error=0.1172

Comparison between training and testing data set:

TRAINING		
JRip	Seed=1	Root Mean Squared Error=0.1707 Mean Absolute Error=0.0583
Ridor	Seed =1	Root Mean Squared Error=0.2508 Mean Absolute Error=0.0629

TEST		
JRip	Seed=1	Root Mean Squared Error=0.2431 Mean Absolute Error=0.1172
Rider	Seed =1	Root Mean Squared Error=0.3423 Mean Absolute Error=0.1172

Result:

Thus the good results by feature selection were found.

EX. No: 8

Web Mining

Aim:

To apply the web mining technique clustering algorithm for the given dataset.

Introduction to Web Mining:

Web mining is an application of data mining techniques to find information patterns from the web data. Web mining helps to improve the power of web search engine by identifying the web pages and classifying the web documents. Web mining is very useful to e-commerce websites and e-services.

Web Content Mining :

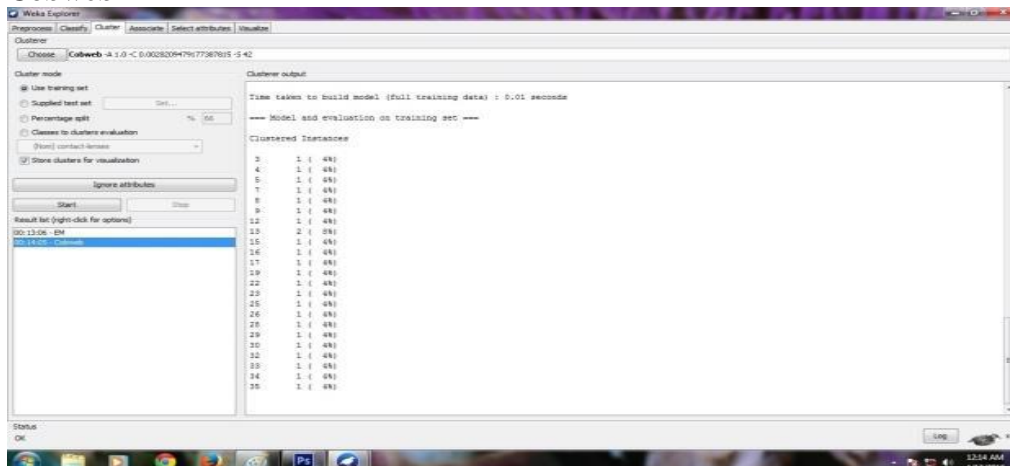
Web content mining can be used for mining of useful data, information and knowledge from web page content. Web structure mining helps to find useful knowledge or information pattern from the structure of hyperlinks. Due to heterogeneity and absence of structure in web data, automated discovery of new knowledge pattern can be challenging to some extent. Web content mining performs scanning and mining of the text, images and groups of web pages according to the content of the input (query), by displaying the list in search engines. For example: If an user wants to search for a particular book, then search engine provides the list of suggestions.

ALGORITHM:

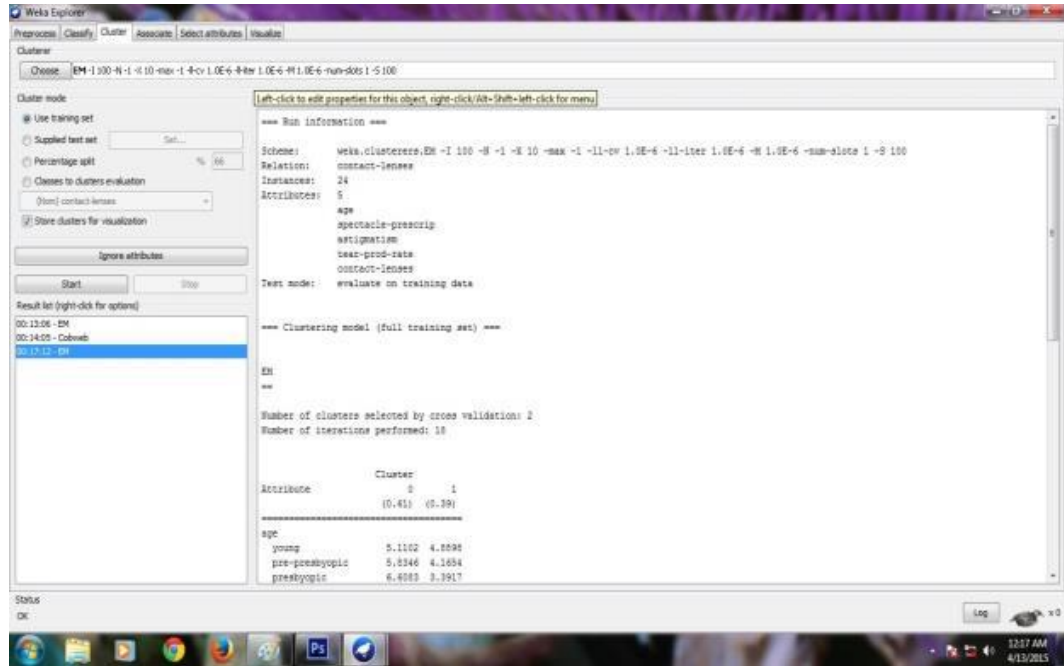
1. Open the weka tool.
2. Download a dataset by using UCI.
3. Apply replace missing values.
4. Apply normalize filter.
5. Click the cluster tab.
6. Apply all algorithms one by one.
7. Find the no of clusters that are formed
8. Note the output.

Output:

Cobweb



EM



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Cluster: Choose EM -I 100 -N -1 -G 10 -max -1 -4 -cv 1.0E-6 -iter 1.0E-6 -M 1.0E-6 -num-slots 1 -S 100

Cluster mode

- ☒ Use training set
- ☐ Supplied test set
- ☐ Percentage split
- ☐ Classes to clusters evaluation
- ☐ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

- 00:13:06 - EM
- 00:14:05 - Cobweb
- 00:15:12 - EM

Left-click to edit properties for this object, right-click/Alt-Shift-left-click for menu

Run information

```

Scheme: weka.clusterers.EM -I 100 -N -1 -G 10 -max -1 -4 -cv 1.0E-6 -iter 1.0E-6 -M 1.0E-6 -num-slots 1 -S 100
Relation: contact-lenses
Instances: 24
Attributes: 5
age
spectacle-prescrip
astigmatism
near-vision
contact-lenses
Test mode: evaluate on training data
  
```

Clustering model (full training set)

EM

Number of clusters selected by cross validation: 2
Number of iterations performed: 10

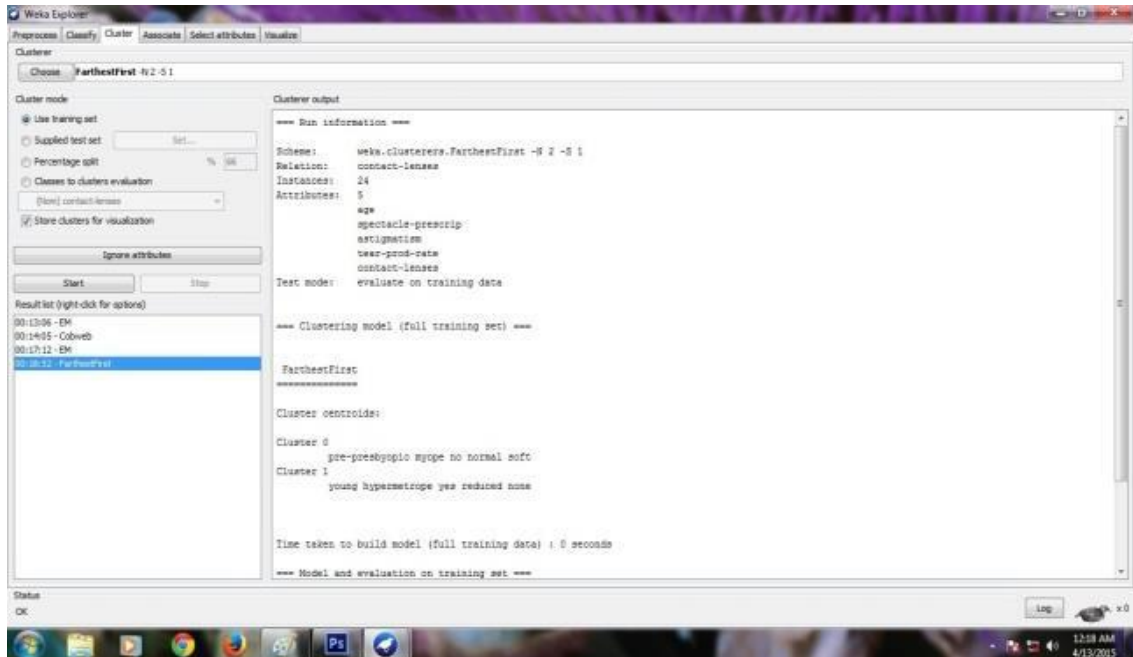
Attribute	Cluster 0	Cluster 1
age		
young	5.1102	4.6898
pre-presbyopic	5.8346	4.1654
presbyopic	4.4083	3.5917

Status: OK

Log

12:17 AM 4/13/2015

Farthest First



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Cluster: Choose FarthestFirst -N 2 -S 1

Cluster mode

- ☒ Use training set
- ☐ Supplied test set
- ☐ Percentage split
- ☐ Classes to clusters evaluation
- ☐ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

- 00:13:06 - EM
- 00:14:05 - Cobweb
- 00:15:12 - EM
- 00:16:12 - FarthestFirst

Run information

```

Scheme: weka.clusterers.FarthestFirst -N 2 -S 1
Relation: contact-lenses
Instances: 24
Attributes: 5
age
spectacle-prescrip
astigmatism
near-vision
contact-lenses
Test mode: evaluate on training data
  
```

Clustering model (full training set)

FarthestFirst

Cluster centroids:

Cluster 0
pre-presbyopic myope no normal soft

Cluster 1
young hypermetrope yes reduced none

Time taken to build model (full training data) : 0 seconds

Model and evaluation on training set

Status: OK

Log

12:18 AM 4/13/2015

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Clusterer
Choose: **SimpleKMeans** #2 -A "weka.core.EuclideanDistance-R test-latt" -S 500 -num-sets : 5 10

Cluster mode
☐ Use training set
☐ Supplied test set [Set...]
☒ Percentage split % 66
☐ Classes to clusters evaluation
 (None) contact-lenses +
☒ Store clusters for visualization

Ignore attributes
Start Stop

Result list (right-click for options)

00:13:56 - EM
00:14:05 - Cobweb
00:17:12 - EM
00:18:52 - PartitionFirst
00:21:29 - FilteredClusterer
01:25:30 - HierarchicalClusterer
01:25:35 - MaleDensityBasedClusterer
01:25:38 - SimpleKMeans

Cluster output

```

Relation:   contact-lenses
Instances:  24
Attributes: age
            spectacle-prescrip
            astigmatism
            tear-prod-rate
            contact-lenses
Test mode:  evaluate on training data

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

Cluster 0
{(((((2.0;1,2.0;1);0,2.0;1);0,{(2.0;1,2.0;1);0,2.0;1);0,{(2.0;1,2.0;1);0,2.0;1);0,{(2.0;1,2.0;1);0,2.0;1);0},{(2.0;1,2.0;1);0,2.0;1);0},{(2.0;1,2.0;1);0,2.0;1);0})}

Cluster 1
{((1.0;1,1.0;1);0,1.0;1);0,1.0;1)}

Time taken to build model (full training data) : 0.07 seconds

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

Clustered Instances

0      20 ( 83%)
1       4 ( 17%)
  
```

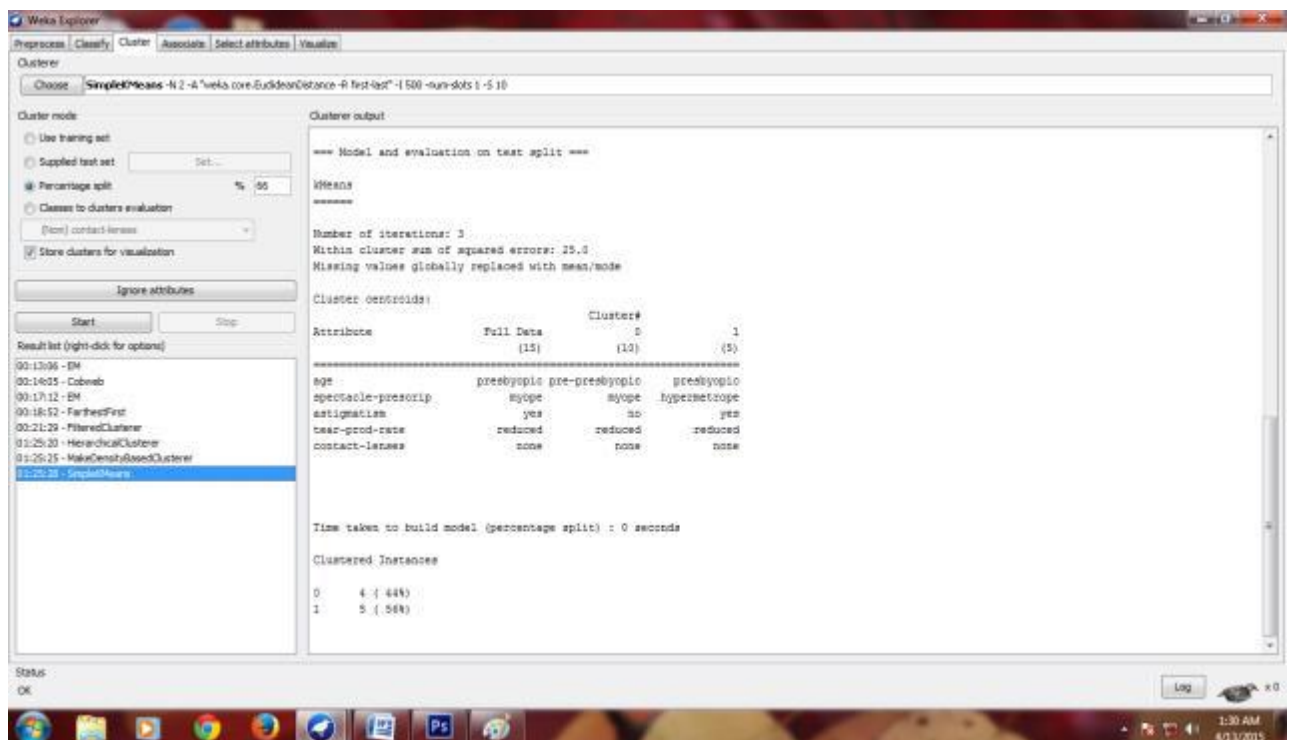
Status OK

The screenshot shows the Weka Explorer application window. The 'Cluster' tab is selected in the top menu. The 'SimpleKMeans' algorithm is chosen for clustering. The 'Cluster mode' section has 'Use training set' selected. The 'Cluster output' section displays the following information:

- Discrete Estimator. Counts = 7 5 (Total = 12)
- Attribute: rear-prod-rate
- Discrete Estimator. Counts = 7 5 (Total = 12)
- Attribute: contact-lenses
- Discrete Estimator. Counts = 2 3 5 (Total = 12)
- Cluster: 1 Prior probability: 0.3529
- Attribute: age
- Discrete Estimator. Counts = 2 1 5 (Total = 8)
- Attribute: spectacle-prescrip
- Discrete Estimator. Counts = 2 5 (Total = 7)
- Attribute: astigmatism
- Discrete Estimator. Counts = 2 5 (Total = 7)
- Attribute: tear-prod-rate
- Discrete Estimator. Counts = 4 8 (Total = 7)
- Attribute: contact-lenses
- Discrete Estimator. Counts = 1 2 5 (Total = 8)
- Time taken to build model (percentage split) : 0 seconds
- Clustered Instances
- 0 6 (67%)
- 1 5 (53%)
- Log likelihood: -4.55299

The 'Result list' on the left shows the sequence of operations performed, with 'SimpleKMeans' highlighted.

Simple KMeans:



Result:

Thus the web mining technique clustering algorithm for the given dataset is implemented.

Aim:

To find association between data and to find the frequent item set for text mining.

Text Data Mining

Text data mining can be described as the process of extracting essential data from standard language text. All the data that we generate via text messages, documents, emails, files are written in common language text. Text mining is primarily used to draw useful insights or patterns from such data. The purchasing of one product when another product is purchased represents an association rule. Association rules are frequently used by retail store to assist in marketing, advertising, floor placement, and inventory control. Association rules are used to show the relationship between data items.

Keyword-based Association Analysis in text mining:

It collects sets of keywords or terms that often happen together and afterward discover the association relationship among them. First, it preprocesses the text data by parsing, stemming, removing stop words, etc. Once it pre-processed the data, then it induces association mining algorithms. Here, human effort is not required, so the number of unwanted results and the execution time is reduced.

ALGORITHM:

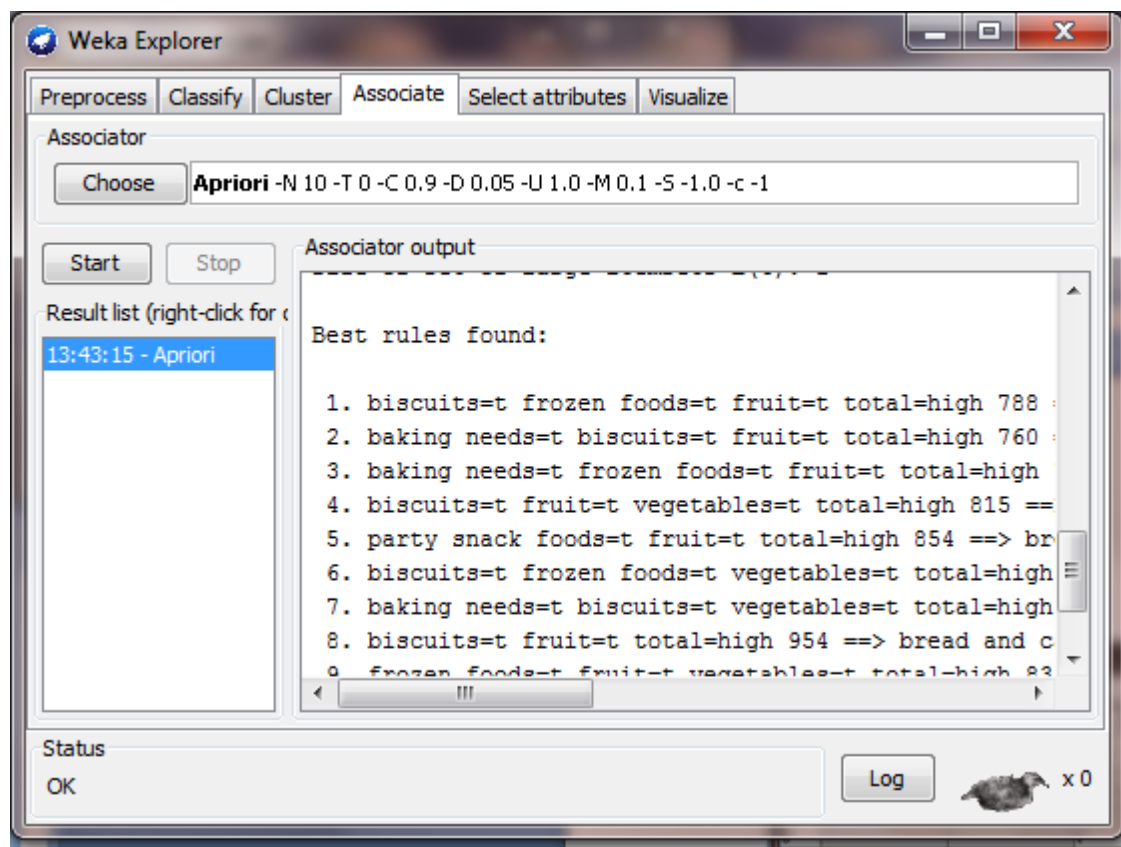
1. Open dataset
2. Select associate
3. Choose different algorithm for association
4. Observe the performance
5. Select the association rule with the maximum confidence rule.

INPUT:

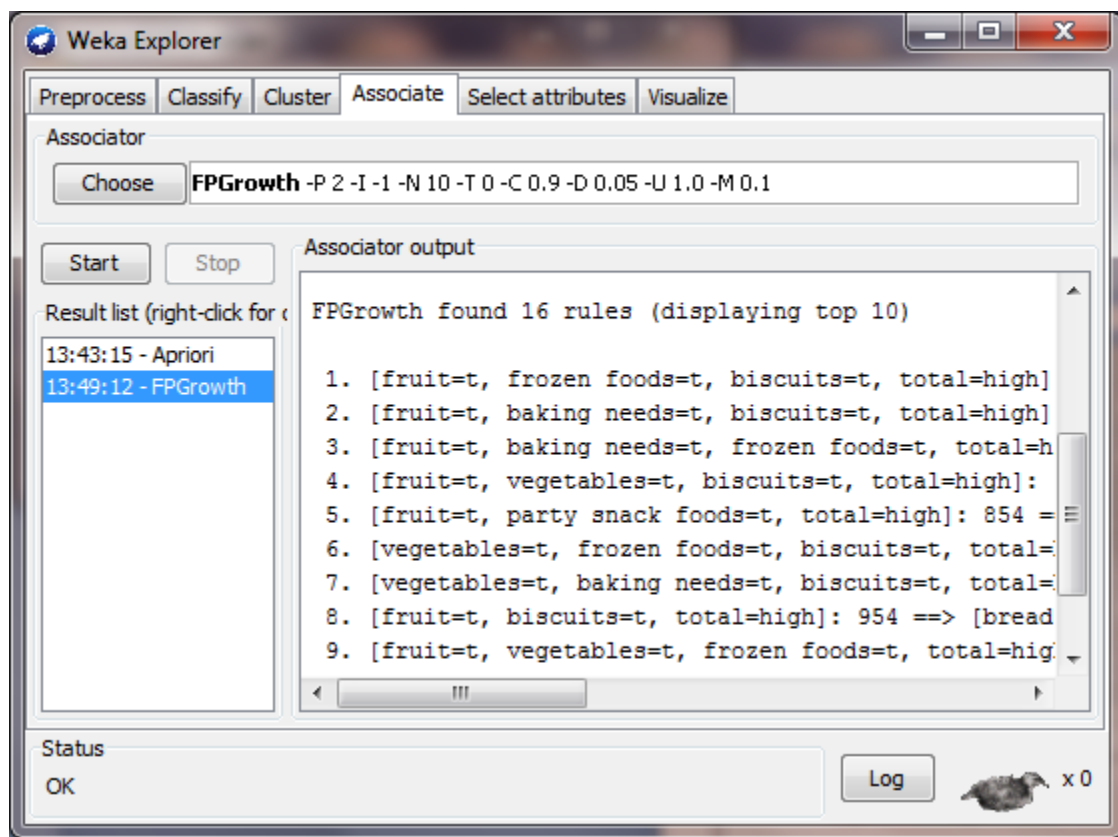
SuperMarket data set

No.	1: department1 Nominal	2: department2 Nominal	3: department3 Nominal	4: department4 Nominal	5: department5 Nominal
1					
2	t				
3					
4	t				
5					
6			t		
7	t				
8					
9	t		t		
10					
11					
12	t				
13	t	t			
14					
15					
16	t				t
17					
18	t		t		
19	t				
20	t				
21		t			t
22	t	t			
23					

OUTPUT: Apriori Algorithm



FP-Growth Algorithm:



Result:

Thus association between data and to find the frequent item set for text mining was found.

Aim:

To design fact and dimension tables.

Fact Table :

A fact table is used in the dimensional model in data warehouse design. A fact table is found at the center of a star schema or snowflake schema surrounded by dimension tables. A fact table consists of facts of a particular business process e.g., sales revenue by month by product. Facts are also known as measurements or metrics. A fact table record captures a measurement or a metric.

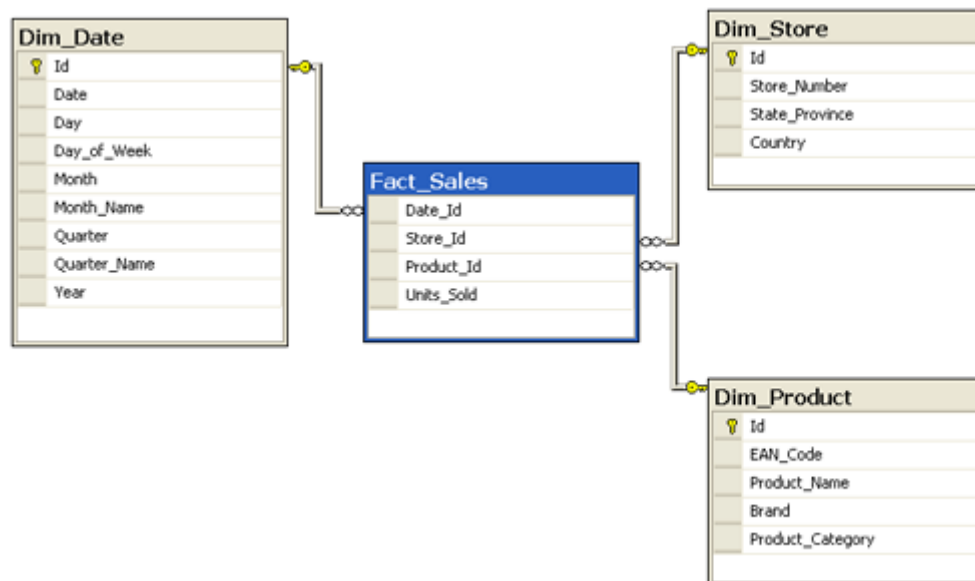
Designing fact table steps

Here is overview of four steps to designing a fact table:

1. **Choosing business process to model** – The first step is to decide what business process to model by gathering and understanding business needs and available data
2. **Declare the grain** – by declaring a grain means describing exactly what a fact table record represents
3. **Choose the dimensions** – once grain of fact table is stated clearly, it is time to determine dimensions for the fact table.
4. **Identify facts** – identify carefully which facts will appear in the fact table.

Fact table FACT_SALES that has a grain which gives us a number of units sold by date, by store and by product.

All other tables such as DIM_DATE, DIM_STORE and DIM_PRODUCT are dimensions tables. This schema is known as the star schema.



Result: Thus design fact and dimension tables are created.