

DATA MINING LABORATORY

(V semester of B.Tech)

As per the curriculam and syllabus

Of

Bharath Institute of Higher Education & Research

**PREPARED BY
DR.YOGESH RAJ KUMAR**

NEW EDITION

Department of information technology



Bharath
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Declared as Deemed - to - be - University under section 3 of UGC Act 1956)
ACCREDITED WITH 'A' GRADE BY NAAC

DATAMINING

(V semester of B.Tech)

As per the curricullam and syllabus

Of

Bharath Institute of Higher Education & Research

**PREPARED BY
DR.YOGESH RAJ KUMAR**

Department of information technology



Bharath
INSTITUTE OF HIGHER EDUCATION AND RESEARCH
(Declared as Deemed - to - be - University under section 3 of UGC Act 1956)
ACCREDITED WITH 'A' GRADE BY NAAC

SCHOOL OF COMPUTING
DEPARTMENT OF INFORMATION TECHNOLOGY

LAB MANUAL

SUBJECT NAME: DATA MINING LABORATORY

SUBJECT CODE: U20ITCJ04

Regulation - 2020

VISION AND MISSION OF THE INSTITUTE

VISION

“Bharath Institute of Higher Education & Research (BIHER) envisions and constantly strives to provide an excellent academic and research ambience for students and members of the faculties to inherit professional competence along with human dignity and transformation of community to keep pace with the global challenges so as to achieve holistic development.”

MISSION

- To develop as a Premier University for Teaching, Learning, Research and Innovation on par with leading global universities.
- To impart education and training to students for creating a better society with ethics and morals.
- To foster an interdisciplinary approach in education, research and innovation by supporting lifelong professional development, enriching knowledge banks through scientific research, promoting best practices and innovation, industry driven and institute-oriented cooperation, globalization and international initiatives.
- To develop as a multi-dimensional institution contributing immensely to the cause of societal advancement through spread of literacy, an ambience that provides the best of international exposures, provide health care, enrich rural development and most importantly impart value-based education.
- To establish benchmark standards in professional practice in the fields of innovative and emerging areas in engineering, management, medicine, dentistry, nursing, physiotherapy and allied sciences.
- To imbibe human dignity and values through personality development and social service activities.

VISION AND MISSION OF THE DEPARTMENT

VISION

To be an excellence in education and research in Information Technology producing global scholars for improvement of the society

MISSION

- To provide sound fundamentals, and advances in Information Technology, Software Engineering, data Communications and Computer Applications by offering world class curriculum.
- To create ethically strong leaders and expert for next generation IT.
- To nurture the desire among faculty and students from across the globe to perform outstanding and impactful research for the benefit of humanity and, to achieve meritorious and significant growth.

PROGRAM EDUCATIONAL OBJECTIVES (PEO)

The Program Educational Objectives (PEOs) of Information technology are listed below: The graduate after 3-5 years of programme completion will

PEO1: PREPARATION

To provide students with sound fundamental in Mathematical, Scientific and Engineering fundamentals necessary to formulate, analyse, and comprehend the fundamental concepts essential to articulate, solve and assess engineering problems and to prepare them for research & development and higher learning.

PEO2: CORE COMPETENCE

To apply critical reasoning, quantitative, qualitative, designing and programming skills, to identify, solve problems and to analyze the experimental evaluations, and finally making appropriate decisions along with knowledge of computing principles and applications and be able to integrate this knowledge in a variety of industry and inter-disciplinary setting.

PEO3: PROFESSIONALISM

To broaden knowledge to establish themselves as creative practicing professionals, locally and globally, in fields such as design, development, problem solving to production support in software industries and R&D sectors.

PEO4: SKILL

To provide better opportunity to become a future researchers / scientist with good communication skills so that they may be both good team-members and leaders with innovative ideas for a sustainable development.

PEO5: ETHICS

To be ethically and socially responsible solution providers and entrepreneurs in Computer Science and other engineering discipline.

PROGRAMME OUTCOMES

PO 1	Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
PO2	Problem Analysis: Identify, formulate, review research literature, and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences and engineering sciences.
PO 3	Design/Development of Solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
PO 4	Conduct Investigations of Complex Problems: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions for complex problems.
PO 5	Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
PO 6	The Engineer and Society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.
PO 7	Environment and Sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
PO 8	Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
PO 9	Individual and Team Work: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO 10	Communication: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and

	write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
PO 11	Project Management and Finance: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.
PO 12	Lifelong Learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

PROGRAMME SPECIFIC OUTCOME

PSO 1	Programming Design : Design and develop algorithm for real life problems using latest technologies and solve it by using computer programming languages and database technologies .
PSO 2	IT Business Scalable Design : Analyze and recommend computing infrastructures and operations requirements and Simulate and implement information networks using configurations, algorithms, suitable protocol and security for valid and optimal connectivity.
PSO 3	Intelligent Agents Design : Design and execute projects for the development of data modeling, data analytics and knowledge representation in various domain.

U20ITCJ04- DATA MINING

PART A- INTRODUCTION OF THE COURSE

Course Code	U20ITCJ04	L	T	P	C										
		3	0	2	4										
Course Title	DATA MINING														
Course Category	Professional Core (C)	Contact Hrs		75											
Pre-requisite		Co- Requisite	Nil												
Name of the Course Coordinator															
Course offering Dept./School		IT / SoC													
Course Objective and Summary															
- To Understand the concepts of Data Mining. Familiarize with Association rule mining. Familiarize with various Classification algorithms. - To Understand the concepts of Cluster Analysis. - To Familiarize with Outlier analysis techniques and applications of Data mining in different domains.															
Course Outcomes (COs)															
CO1	Gain knowledge about the concepts of Data Mining.														
CO2	Apply Association rule mining technique.														
CO3	Use various Classification algorithms.														
CO4	Gain knowledge on the concepts of Cluster Analysis.														
CO5	Identify Outlier analysis techniques.														
CO6	Understand the importance of applying Data mining concepts in different domains.														
Mapping / Alignment of COs with PO & PSO															
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
CO1															
CO2															
CO3															
CO4															
CO5															
CO6															
(Tick mark or level of correlation: 3-High, 2-Medium, 1-Low)															

PART B- CONTENT OF THE COURSE

UNIT I DATA WAREHOUSING, BUSINESS ANALYSIS AND ON-LINE ANALYTICAL PROCESSING (OLAP) 9

Data warehousing Components – Building a Data warehouse - Mapping the Data Warehouse to a Multiprocessor Architecture - DBMS Schemas for Decision Support - Metadata - Reporting and Query tools – Applications - Tool Categories - The Need for Applications - Multidimensional Data Model - OLAP Guidelines - Multidimensional versus - Multirelational OLAP - Categories of Tools - OLAP Tools and the Internet - Integration of a Data Mining System with a Data Warehouse.

UNIT II DATA MINING – INTRODUCTION 9

Introduction to Data Mining Systems – Data and types - Types of Data - Data Mining Functionalities - Classification of Data Mining Systems - Data Mining Task Primitives - Knowledge Discovery Process - Data Objects and attribute types, Statistical description of data, Data Preprocessing – Cleaning, Integration, Extraction, Reduction, Transformation and discretization, Data Visualization, Data similarity and dissimilarity measures - Issues Data Preprocessing

UNIT III DATA MINING – FREQUENT PATTERN ANALYSIS 9

Interestingness of Patterns - Mining Frequent Patterns - Associations and Correlations – Mining Methods- Mining Various Kinds of Association Rules - Correlation Analysis - Constraint Based Association Mining.

UNIT IV CLASSIFICATION AND PREDICTION 9

Basic Concepts Decision Tree Induction - Bayesian Classification –Rule Based Classification – Classification by Back Propagation – Support Vector Machines — Associative Classification - Lazy Learners - Other Classification Methods – Prediction.

UNIT V CLUSTERING 9

Cluster Analysis - Types of Data - Categorization of Major Clustering Methods - K-means - Partitioning Methods - Hierarchical Methods - Density-Based Methods - Grid Based Methods - Model-Based Clustering Methods - Clustering High Dimensional Data - Constraint, Based Cluster Analysis - Outlier Analysis, Data Mining Applications - Case studies based on Data Mining Tool.

LAB

1. Introduction to exploratory data analysis using R
2. Demonstrate the Descriptive Statistics for a sample data like mean, median, variance and correlation etc.
3. Demonstrate Missing value analysis and different plots using sample data
4. Demonstration of apriori algorithm on various data sets with varying confidence (%) and support (%)
5. Demo on Classification Techniques using sample data Decision Tree, ID3 or CART
6. Demonstration of Clustering Techniques K-Mean and Hierarchical
7. Simulation of Page Rank Algorithm and Demonstration on Hubs and Authorities.
8. Demo on Classification Technique using KNN.
9. Demonstration on Document Similarity Techniques and measurements
10. Design and develop a recommendation engine for the given application.

LIST OF EXPERIMENTS & SCHEDULE

COURSE CODE: U20ITCJ04

COURSE TITLE: DATA MINING LAB

S.No	DATE	NAME OF THE EXPERIMENT	PAGE NO.	SIGN
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		Data pre-processing – data filters		
8		Feature selection		
9		Web mining		
10		Text mining		
11		Design of fact & dimension tables		
12		Generating graphs for star schema.		

EX.NO:1

LISTING APPLICATIONS FOR MINING

AIM:

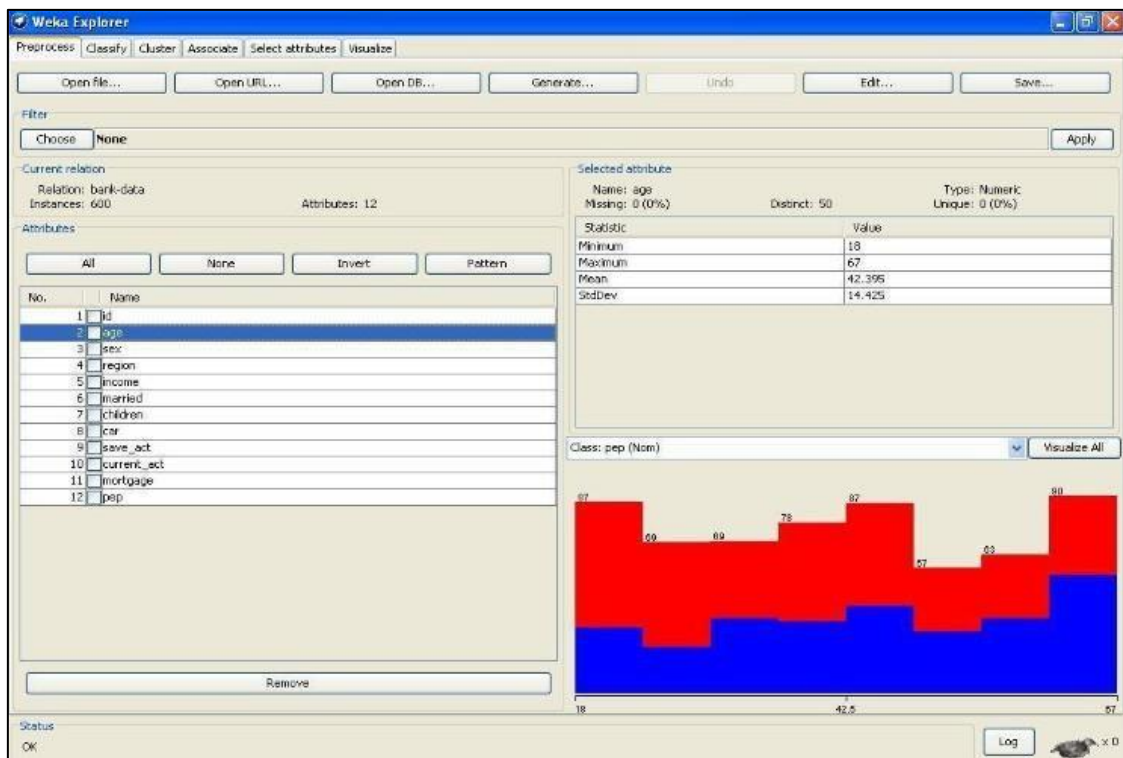
To list all the categorical (or nominal) attributes and the real-valued attributes separately.

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:

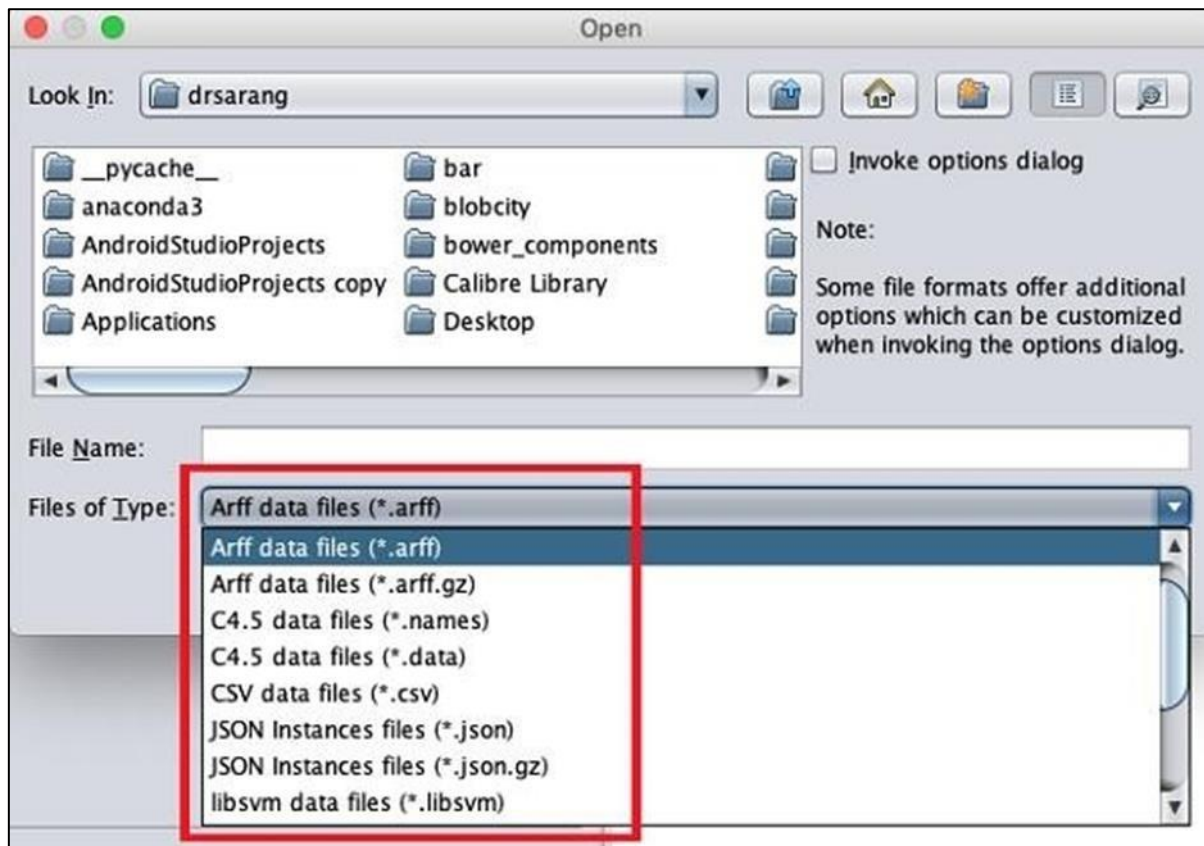
To study the file formats for the data mining.

INTRODUCTION:

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

- ARFF
- ARFF.gz
- bsi
- csv
- dat
- data
- json
- json.gz
- libsvm
- m
- names
- xrff
- 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
```

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 were 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:

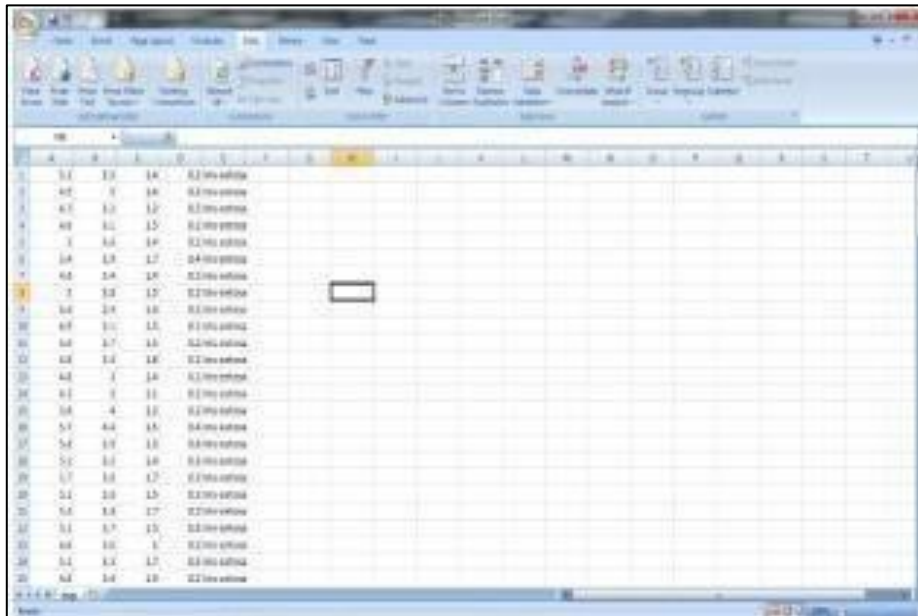
Most of the data that we have collected from public forum is in the text format that cannot be read by Weka tool. Since Weka (Data Mining tool) recognizes the data in ARFF format only we have to convert the text file into ARFF file.

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	2	3	4
1	0.1	0.2	0.3	0.4
2	0.2	0.3	0.4	0.5
3	0.3	0.4	0.5	0.6
4	0.4	0.5	0.6	0.7
5	0.5	0.6	0.7	0.8
6	0.6	0.7	0.8	0.9
7	0.7	0.8	0.9	1.0
8	0.8	0.9	1.0	1.1
9	0.9	1.0	1.1	1.2
10	1.0	1.1	1.2	1.3
11	1.1	1.2	1.3	1.4
12	1.2	1.3	1.4	1.5
13	1.3	1.4	1.5	1.6
14	1.4	1.5	1.6	1.7
15	1.5	1.6	1.7	1.8
16	1.6	1.7	1.8	1.9
17	1.7	1.8	1.9	2.0
18	1.8	1.9	2.0	2.1
19	1.9	2.0	2.1	2.2
20	2.0	2.1	2.2	2.3

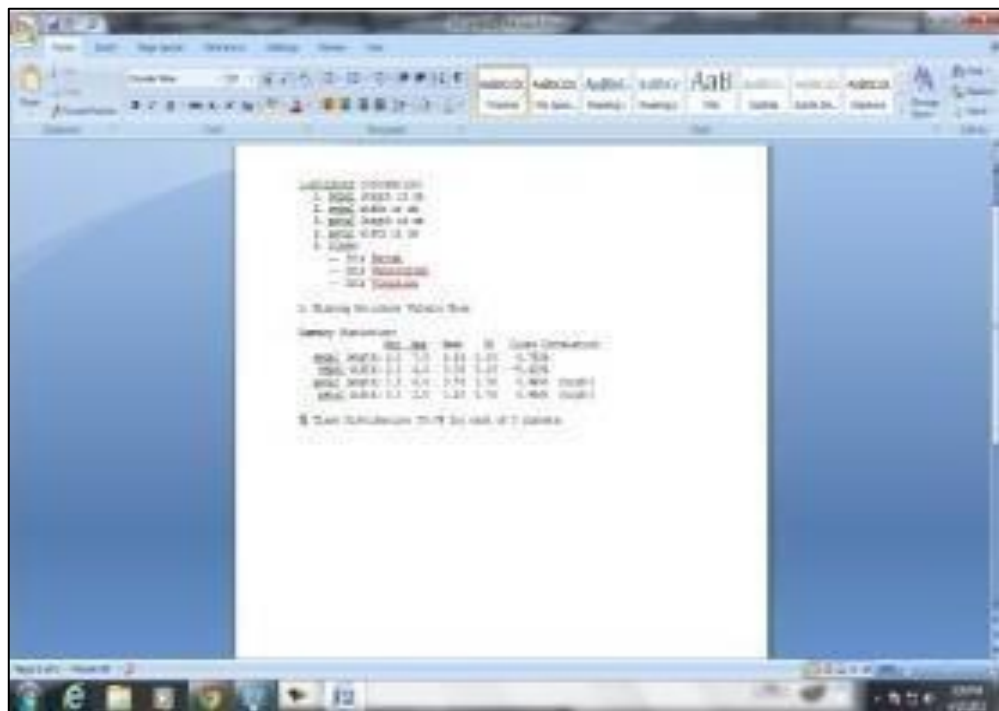
Data ARFF File:



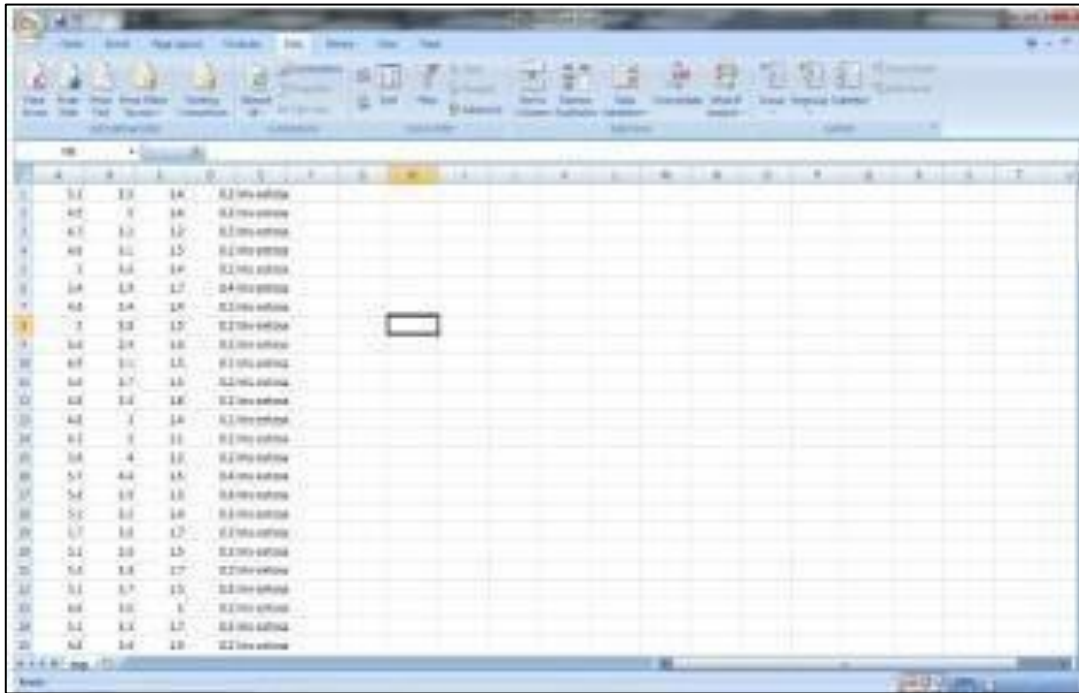
RESULT:

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

EX.NO:3b.



Data Text File:



	A	B	C	D	E
1	1	0.1	1	0.4	0.2 test instance
2	2	0.2	2	0.6	0.2 test instance
3	3	0.3	3	0.8	0.2 test instance
4	4	0.4	4	1.0	0.2 test instance
5	5	0.5	5	1.2	0.2 test instance
6	6	0.6	6	1.4	0.2 test instance
7	7	0.7	7	1.6	0.2 test instance
8	8	0.8	8	1.8	0.2 test instance
9	9	0.9	9	2.0	0.2 test instance
10	10	1.0	10	2.2	0.2 test instance
11	11	1.1	11	2.4	0.2 test instance
12	12	1.2	12	2.6	0.2 test instance
13	13	1.3	13	2.8	0.2 test instance
14	14	1.4	14	3.0	0.2 test instance
15	15	1.5	15	3.2	0.2 test instance
16	16	1.6	16	3.4	0.2 test instance
17	17	1.7	17	3.6	0.2 test instance
18	18	1.8	18	3.8	0.2 test instance
19	19	1.9	19	4.0	0.2 test instance
20	20	2.0	20	4.2	0.2 test instance

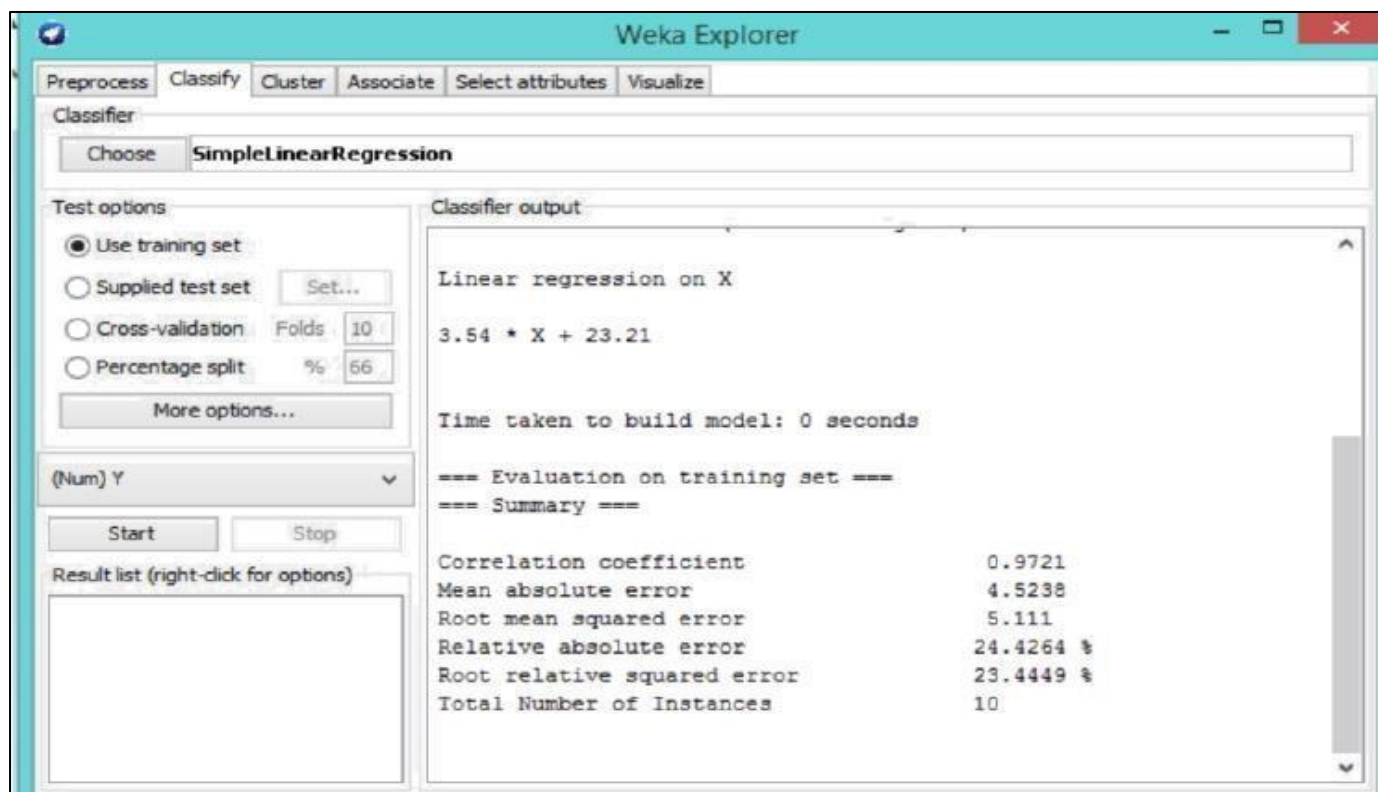
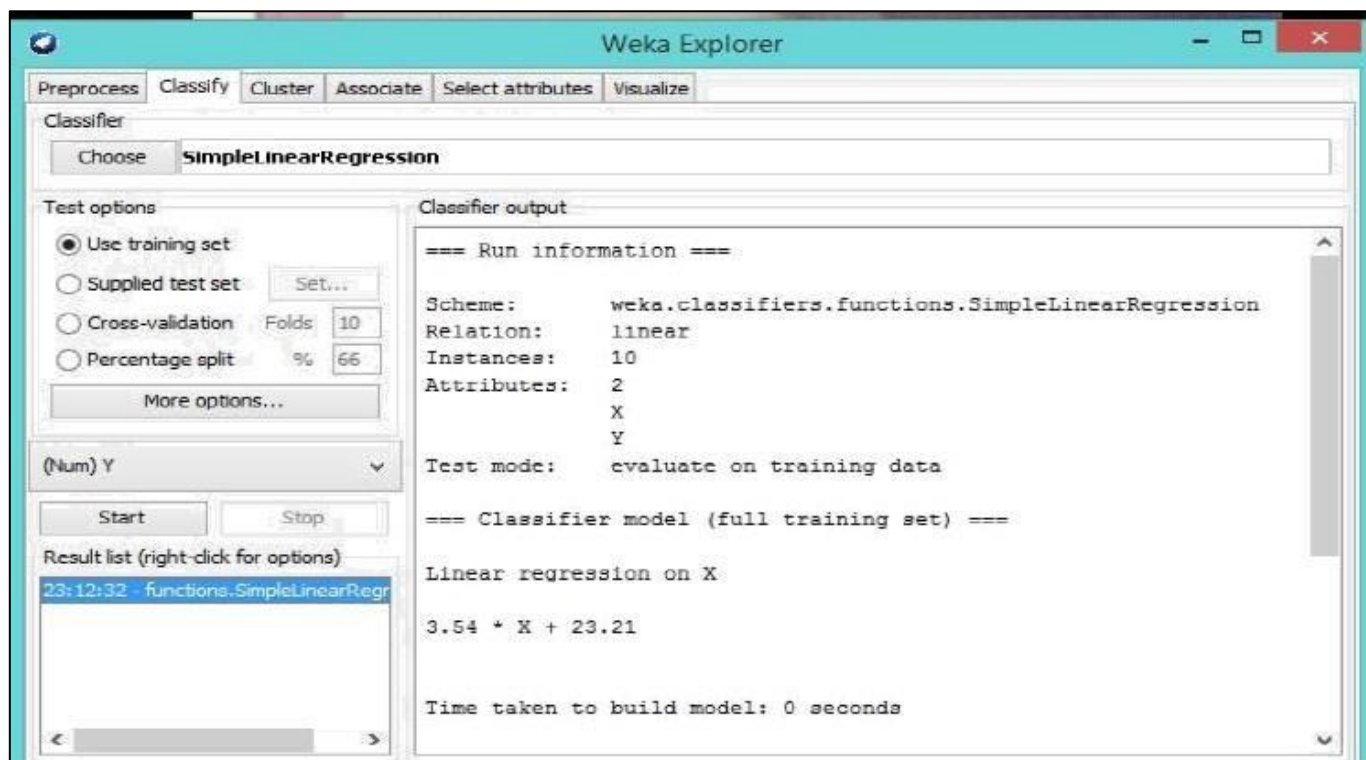
RESULT:

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

EX. No: 4

INPUT:

OUTPUT:



RESULT:

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

EX. No: 5

AIM:

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

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 Classification Tab.
6. Apply Navie Bayes Classification.
7. Find the Classified Value.
8. Note the OUTPUT.

Bayes' Theorem in the Classification Context:

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

Example: predict whether a costumer will buy a computer or not " Costumers are described by two attributes: age and income " $\mathbf{X}$ is a 35 years-old costumer with an income of 40k " $\mathbf{H}$ is the hypothesis that the costumer will buy a computer " $\mathbf{P(H|X)}$ reflects the probability that costumer $\mathbf{X}$ will buy a computer given that we know the costumers' age and income.

INPUT:

linear - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View

Clipboard Font Alignment Number Styles Cells Editing

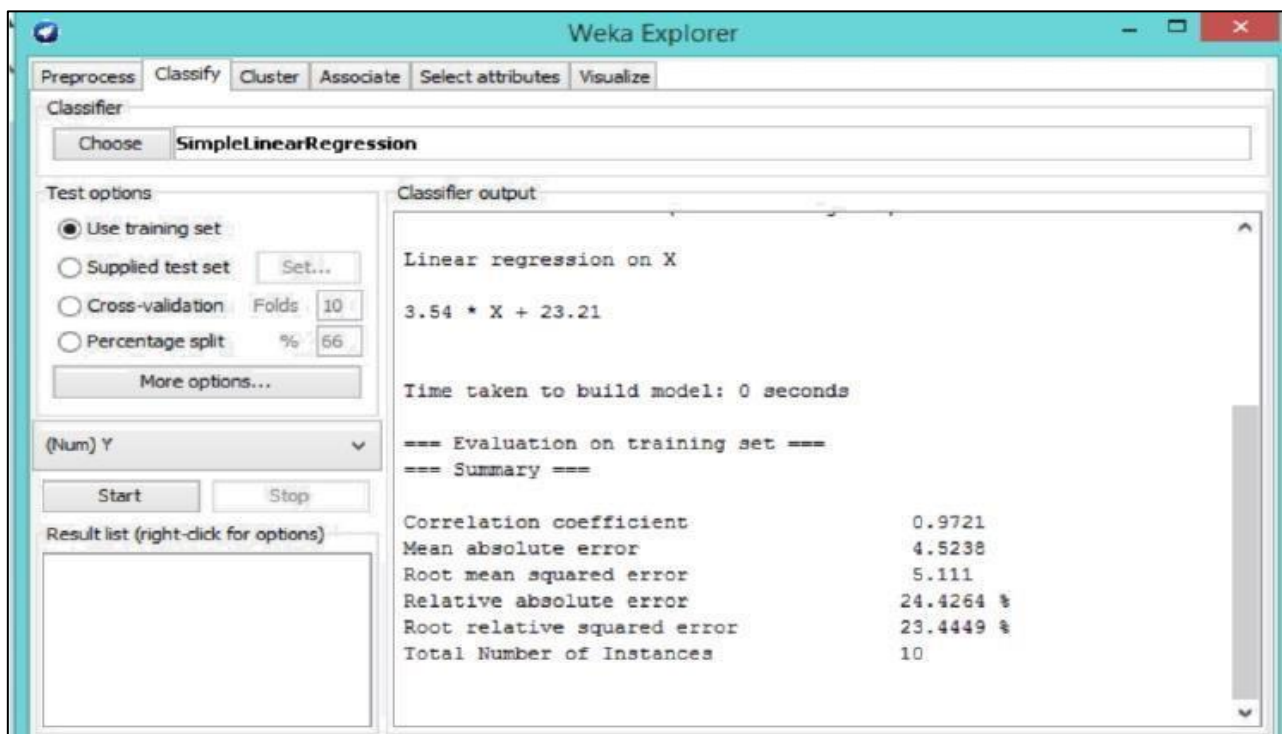
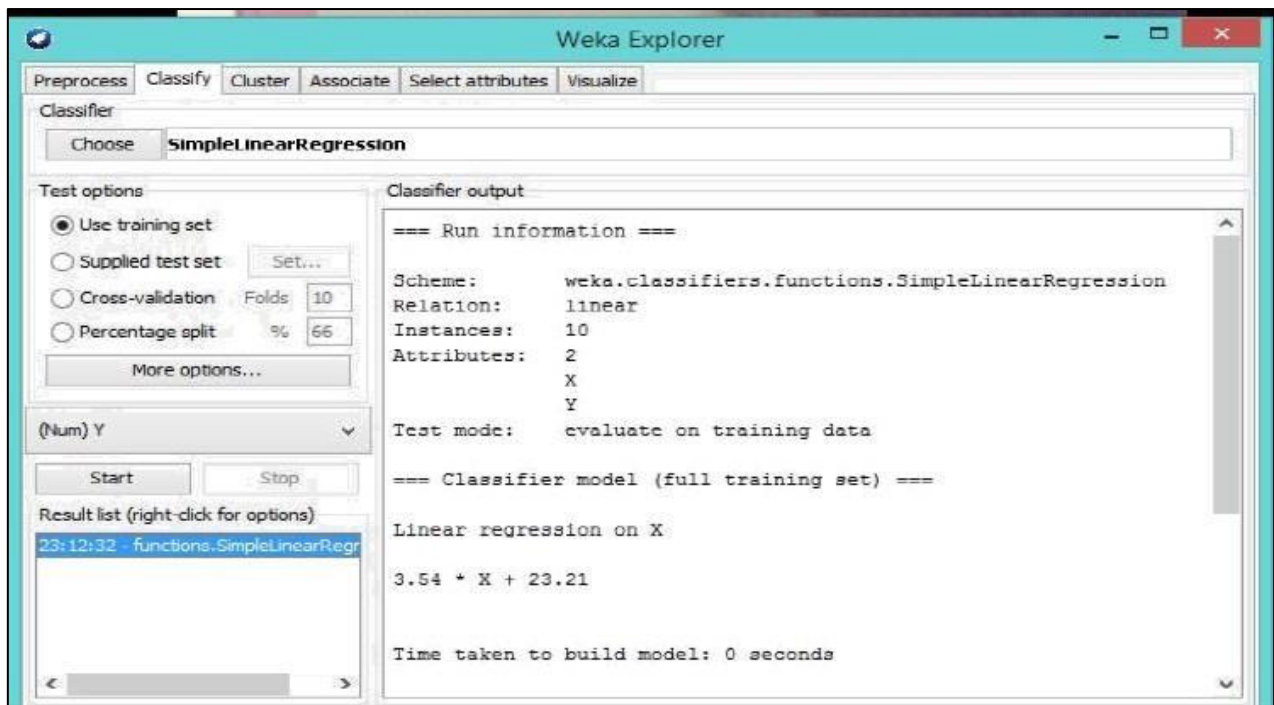
F5

	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								

linear Sheet2 Sheet3

Ready 100%

OUTPUT:



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:

FileEditViewHelp

Dataset: training set.vectors

Filter: none

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

Visualize

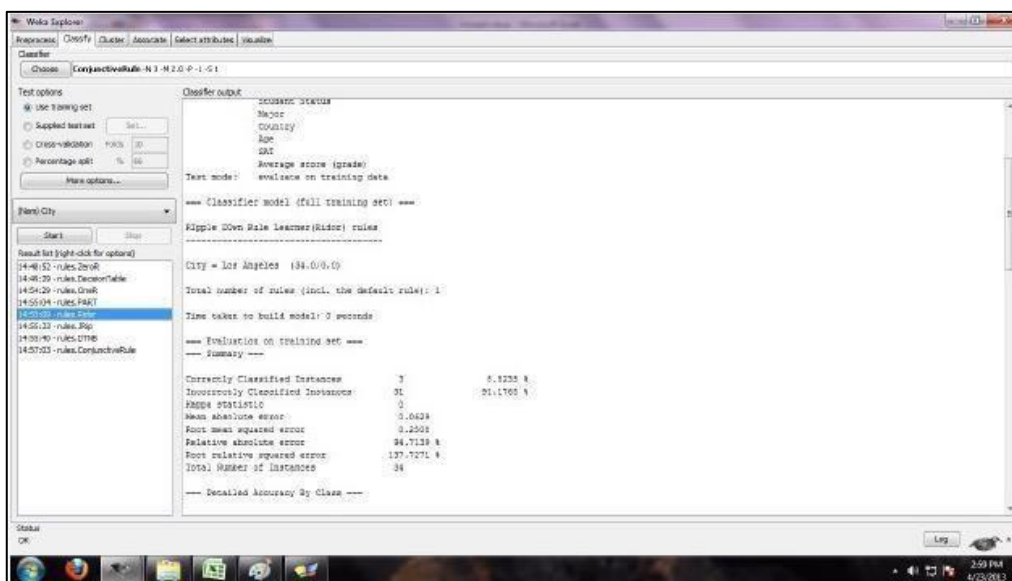
Visualize

Visualize

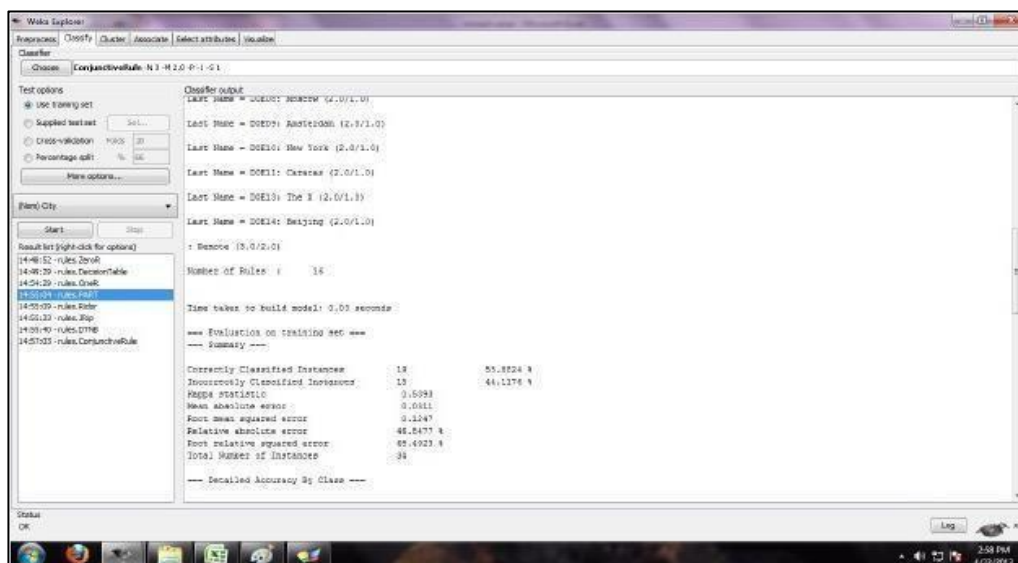
Visualize

Visualize

Ridor:



PART:



OneR:

The screenshot shows the Weka Explorer interface with the OneR classifier selected. The 'Classifier output' pane displays the following results:

Classifier output

```

J4E101 -> Amsterdam
J4E102 -> Mexico
J4E103 -> Chicago
J4E104 -> San Juan
J4E105 -> Rome
J4E106 -> New York
J4E107 -> The B
J4E108 -> Beijing
J4E109 -> Stockholm
J4E110 -> Edinburgh
J4E111 -> Interconcrete
J4E112 -> Loco
J4E113 -> Monte Aires
J4E114 -> Aom
(18/34 instances correct)

Time taken to build model: 0 seconds

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

Correctly Classified Instances      33      97.058 %
Incorrectly Classified Instances     1      2.942 %
Mean absolute error                  0.003
Root mean squared error              0.045
Relative absolute error              0.003 %
Root relative squared error          24.756 %
Total Number of Instances           34

--- Detailed Accuracy By Class ---

```

JRip:

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The 'Classifier output' pane displays the following results:

Classifier output

```

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

=== Classifier model (full training set) ===
JRIP rules:
=====
(First Rule = J4E101) => City=Lackawanna (2.0/0.0)
=> City=Los Angeles (31.0/20.0)

Number of Rules : 2

Time taken to build model: 0.03 seconds

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

Correctly Classified Instances      5      14.705 %
Incorrectly Classified Instances    29      85.294 %
Mean absolute error                 0.043
Root mean squared error             0.174
Relative absolute error              93.759 %
Root relative squared error          96.895 %
Total Number of Instances           34

--- Detailed Accuracy By Class ---

```

DTNB:

The screenshot shows the Weka Explorer interface with the DTNB classifier selected. The 'Classifier output' pane displays the following results:

Classifier output

```

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

=== Classifier model (full training set) ===
Decision Table:
Number of training instances: 34
Number of Rules : 14
100 branches covered by majority class.
Evaluation (fpr feature selection): CV (leave one out)
Feature set: 1,3,4

Time taken to build model: 0.15 seconds

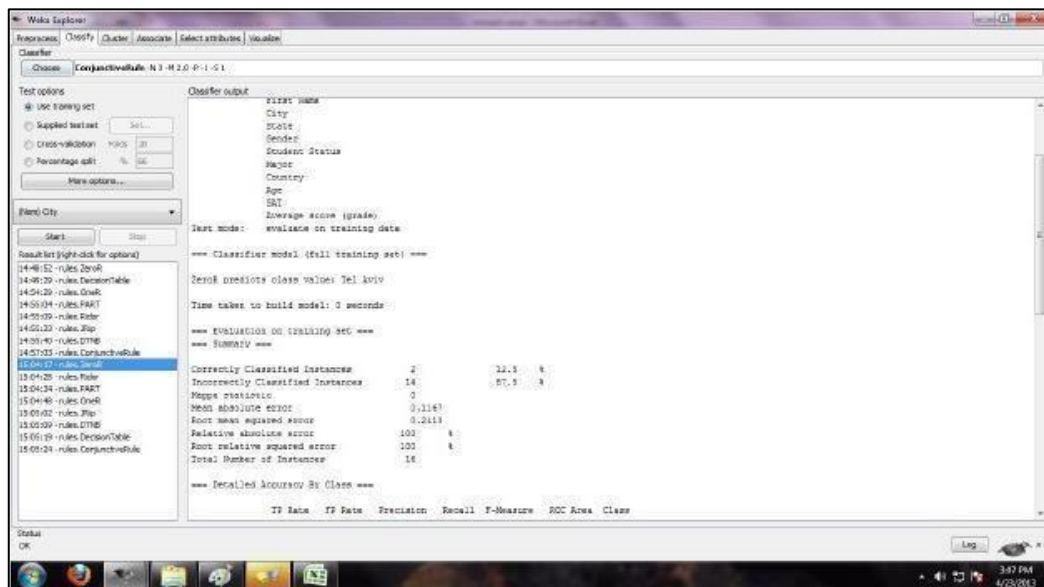
=== Evaluation on training set ===
--- Summary ---

Correctly Classified Instances      30      88.235 %
Incorrectly Classified Instances     4      11.764 %
Mean absolute error                 0.071
Root mean squared error             0.194
Relative absolute error              93.823 %
Root relative squared error          93.252 %
Total Number of Instances           34

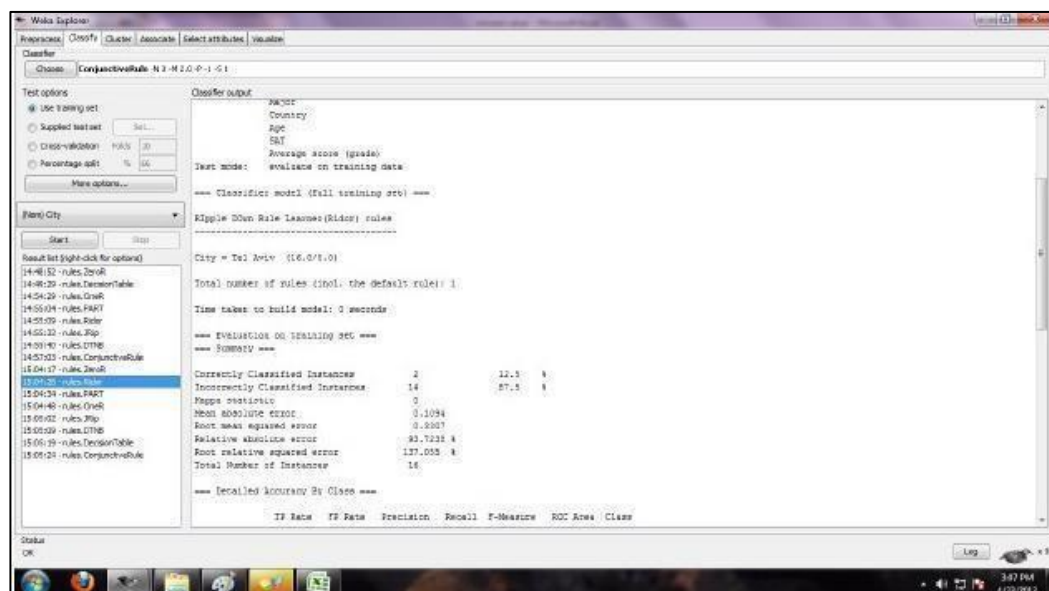
--- Detailed Accuracy By Class ---

```

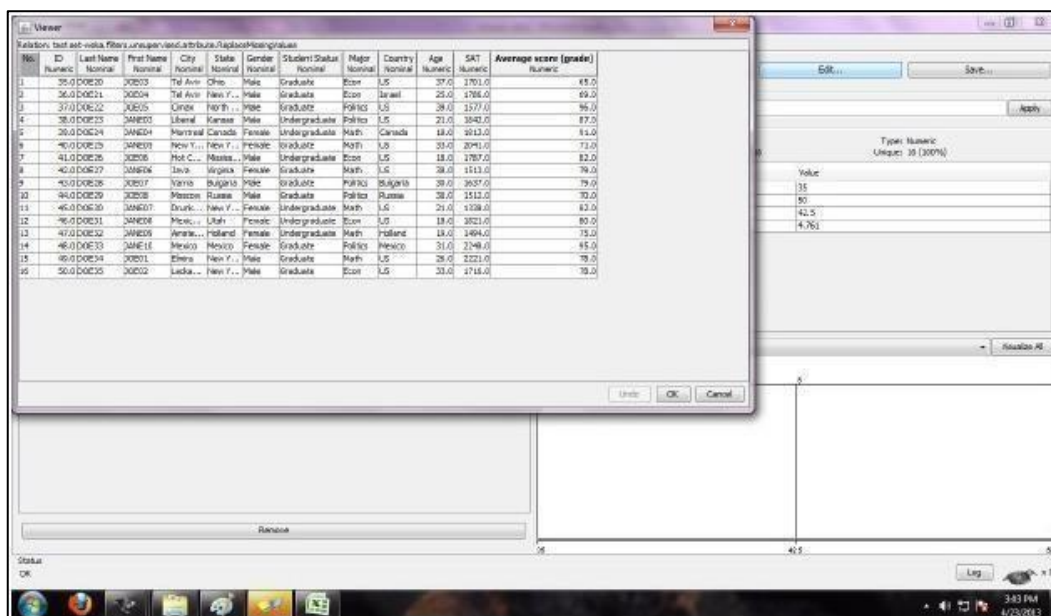

TEST DATA:



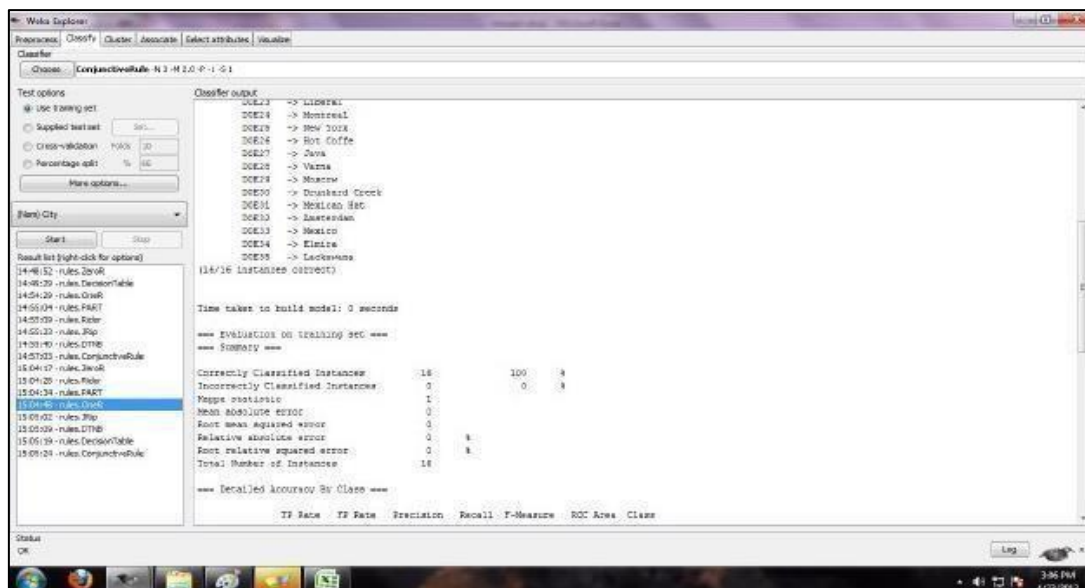
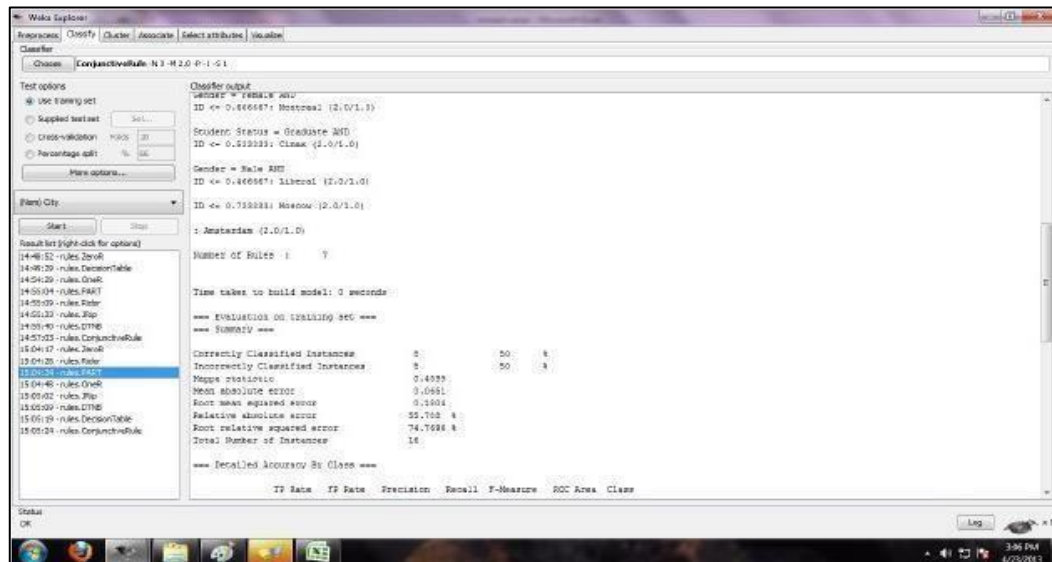
ZeroR:



Ridor:



OneR:



Weka Explorer

Instances: 100000 | Class: Country | Summary: 100000 | Visualize

Classifier

Choose: ConjunctiveRule N 3-M 2.0-P 1-L 1

Test options

☒ Use training set

☐ Supplied test set

☐ Cross-validation

☐ Percentage split

More options...

Classifier output

Country

Age

Sex

Average score (grades)

Test mode: evidence in training data

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

J48 rules:

====

-> Country = {Age = 34.0, Sex = F, Average score (grades) = 66.0} => evidence in training data

Number of Rules : 1

Time taken to build model: 0.01 seconds

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

==== Summary ====

Correctly Classified Instances	Incorrectly Classified Instances	Weighted Average
10	4	0.7143
Mean absolute error		0.2857
Root mean squared error		0.5345
Relative absolute error		99.5016 %
Root relative squared error		99.5016 %
Total Number of Instances		14

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

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.7143	0.2857	0.7143	0.7143	0.7143	0.7143	Country

Results list (right-click for options)

14:46:52 - rules.J48

14:46:52 - rules.DecisionTable

14:50:20 - rules.C4.5

14:50:20 - rules.PART

14:50:20 - rules.Rider

14:50:20 - rules.Rip

14:50:20 - rules.DT48

14:50:20 - rules.ConjunctiveRule

15:04:17 - rules.J48

15:04:20 - rules.Rule

15:04:20 - rules.PART

15:04:20 - rules.C4.5

15:04:20 - rules.DT48

15:04:20 - rules.DecisionTable

15:04:20 - rules.ConjunctiveRule

Status

OK

Log

DTNB:

The screenshot shows the Weka Explorer application window. The 'Classifier' tab is selected, and the 'ConjunctionalRule_N2_H2O_P1_G1' classifier is chosen. The 'Test options' section shows 'Use training set' selected. The 'Classifier output' section displays the following information:

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

The 'Decision Tables' section shows the following results:

=== Evaluation on testing set ===			
=== SUMMARY ===			
Correctly Classified Instances	15	93.75 %	
Incorrectly Classified Instances	1	6.25 %	
Weighted Average	0.9523		
Weighted Average Error	0.0477		
Root Mean Squared Error	0.203		
Relative Absolute Error	0.7359		
Root Relative Squared Error	0.4356		
Total Number of Instances	16		

The 'Detailed Accuracy by Class' section shows the following table:

TP Rate	TS Rate	Precision	Recall	F-Measure	ROC Area	Class
0.9375	0.9375	0.9375	0.9375	0.9375	0.9375	Class 0
0.0625	0.0625	0.0625	0.0625	0.0625	0.0625	Class 1

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
DATA PRE-PROCESSING – DATA FILTERS

AIM:

To perform the data pre-processing by applying filter.

OBJECTIVES:

The data collected from public forums have plenty of noise or missing data. Weka provides filter to replace the missing values and to remove the noisy data. So that the result will be more accurate.

ALGORITHM:

1. Download a complete data set (numeric) from UCI.
2. Open the data set in Weka tool.
3. Save the data set with missing values.
4. Apply replace missing value filter.
5. Calculate the accuracy using the formula

$$\text{Accuracy} = \sqrt{\sum (\text{old} - \text{new})^2}$$
$$\text{Percentage of accuracy} = \frac{\text{Accuracy}}{\sum \text{old value}} \times 100$$

OUTPUT:

Student Details Table: Missing values

Viewer					
Relation: weather					
No.	1: outlook Nominal	2: temperature Numeric	3: humidity Numeric	4: windy Nominal	5: play Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy		96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0		TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny			FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast			TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy		91.0	TRUE	no

Student Details Table: Replace Missing values:

Viewer					
Relation: weather-weka.filters.unsupervised.attribute.Replace					
No.	1: outlook	2: temperature	3: humidity	4: windy	5: play
	Nominal	Numeric	Numeric	Nominal	Nominal
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	74.8	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	83.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	74.8	83.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	74.8	83.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	74.8	91.0	TRUE	no

CALCULATION:

Data Location	Old Data	Predicted data	Errors	(Error) ²
J2				
J4				
J6				

RESULT:

Thus, the data pre-processing by applying filter is performed.

EX. No: 8

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:

Weka Explorer

Relation: training data - weka.filters.unsupervised.attribute.ReplaceMissingValues - weka.filters.unsupervised.attribute.Normalize 0.0-10.0

No.	ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grades)
		Normal	Normal	Normal	Normal	Normal	Normal	Normal	Normal		Normal	Normal
1	0.00001	DAVE01	Sedona	Califor	Female	Graduate	Politics	US	1.571...	1.991...	67.0	
2	0.00002	DAVE02	Sedona	New Y...	Female	Undergraduate	Math	US	1.040...	1.687...	62.0	
3	0.00003	DAVE03	Clark	New Y...	Female	Graduate	Math	US	1.368...	1.509...	78.0	
4	0.00004	DAVE04	Lacka...	New Y...	Male	Graduate	Econ	US	1.714...	1.389...	78.0	
5	0.00005	DAVE05	Defiance	Ohio	Male	Graduate	Math	US	1.604...	1.373...	65.0	
6	0.00006	DAVE06	The New	Spain	Male	Graduate	Econ	Spain	1.533...	1.761...	69.0	
7	0.00007	DAVE07	Omaha	Nebr	Male	Graduate	Politics	US	1.0...	1.546...	69.0	
8	0.00008	DAVE08	Liberal	Kansas	Female	Undergraduate	Politics	US	1.140...	1.514...	67.0	
9	0.00009	DAVE09	Marina	Canada	Female	Undergraduate	Math	Canada	1.0...	1.469...	61.0	
10	0.00010	DAVE10	New Y...	New Y...	Female	Graduate	Math	US	1.714...	1.723...	71.0	
11	0.00011	DAVE11	Hat C...	Massa...	Male	Undergraduate	Econ	US	1.0...	1.462...	62.0	
12	0.00012	DAVE12	Joyce	Virginia	Female	Graduate	Politics	US	1.052...	1.308...	79.0	
13	0.00013	DAVE13	Varna	Bulgaria	Male	Graduate	Politics	Bulgaria	1.571...	1.307...	79.0	
14	0.00014	DAVE14	Moscow	Russia	Male	Graduate	Politics	Russia	1.571...	1.179...	70.0	
15	0.00015	DAVE15	Drum...	New Y...	Female	Undergraduate	Math	US	1.140...	1.4...	62.0	
16	0.00016	DAVE16	Archie...	Utah	Female	Undergraduate	Econ	US	1.0...	1.407...	60.0	
17	0.00017	DAVE17	Archie...	Holland	Female	Undergraduate	Math	Holland	1.047...	1.366...	75.0	
18	0.00018	DAVE18	Metro	Mexico	Female	Graduate	Politics	Mexico	1.614...	1.537...	65.0	
19	0.00019	DAVE19	Caracas	Venez...	Female	Undergraduate	Math	Venezue...	1.0...	1.944...	62.0	
20	0.00020	DAVE20	San Juan	Puerto...	Male	Graduate	Politics	US	1.714...	1.602...	65.0	
21	0.00021	DAVE21	Reno	Oregon	Female	Undergraduate	Econ	US	1.047...	1.406...	67.0	
22	0.00022	DAVE22	New Y...	New Y...	Male	Undergraduate	Econ	US	1.140...	1.546...	62.0	
23	0.00023	DAVE23	The S...	Rosco...	Female	Graduate	Politics	US	1.533...	1.441...	69.0	
24	0.00024	DAVE24	Beijing	China	Female	Undergraduate	Math	China	1.0...	1.214...	79.0	
25	0.00025	DAVE25	Stockh...	Sweden	Male	Undergraduate	Politics	Sweden	1.047...	1.566...	68.0	
26	0.00026	DAVE26	Emilia...	France	Male	Graduate	Econ	US	1.476...	1.098...	69.0	
27	0.00027	DAVE27	Inter...	Permi...	Male	Undergraduate	Math	US	1.099...	1.804...	69.0	
28	0.00028	DAVE28	Utop	Chad	Female	Undergraduate	Econ	US	1.099...	1.0...	64.0	
29	0.00029	DAVE29	Buenos	Argen...	Male	Graduate	Politics	Argentina	1.571...	1.569...	65.0	
30	0.00030	DAVE30	Arnie	Louisiana	Male	Undergraduate	Econ	US	1.047...	1.585...	79.0	
31	0.00031	DAVE31	Los An...	Califor	Female	Graduate	Politics	US	1.571...	1.951...	67.0	
32	0.00032	DAVE32	Sedona	Arizon	Female	Undergraduate	Math	US	1.040...	1.687...	62.0	
33	0.00033	DAVE33	Clark	New Y...	Male	Graduate	Math	US	1.368...	1.509...	78.0	
34	0.00034	DAVE34	Lacka...	New Y...	Male	Graduate	Econ	US	1.714...	1.389...	78.0	

Filter...

Save...

Apply

Type: Numeric

Unique: 14 (100%)

Value

0

1

0.5

0.300

Right click (or left-click) for context menu

Undo

OK

Cancel

JRip(seed=1):

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The 'Classifier output' tab is active, displaying the following information:

Classifier output

JRIP rules:

```
(First Name = JOSE) => City=Lackawanna (2.0/0.0)
(First Name = JOSE) => City=Elmira (2.0/0.0)
=> City=Scranton (31.0/27.0)
```

Number of Rules : 3

Time taken to build model: 0.04 seconds

=== Evaluation on training set ===

Summary

Metric	Value	Percentage
Correctly Classified Instances	7	20.5682 %
Incorrectly Classified Instances	27	79.4318 %
Kappa statistic	0.1333	
Mean absolute error	0.0583	
Root mean squared error	0.1707	
Relative absolute error	87.7088 %	
Root relative squared error	93.7588 %	
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	0.541	Los Angeles
1	0.671	0.1	1	0.182	0.563	Scranton
1	0	1	1	1	1	Elmira
1	0	1	1	1	1	Lackawanna
0	0	0	0	0	0.561	Default
0	0	0	0	0	0.541	Pal Alto
0	0	0	0	0	0.561	Class

JRip(seed=2):

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The 'Classifier output' tab is active, displaying the following information:

Classifier output

JRIP rules:

```
(First Name = JOSE) => City=Lackawanna (2.0/0.0)
=> City=Scranton (31.0/29.0)
```

Number of Rules : 2

Time taken 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	0.0422	
Mean absolute error	0.1764	
Root mean squared error	0.1764	
Relative absolute error	93.7088 %	
Root relative squared error	96.0937 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

JRip(seed=3):

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The 'Classifier output' tab is active, displaying the following information:

Classifier output

JRIP rules:

```
(First Name = JOSE) => City=Lackawanna (2.0/0.0)
=> City=Scranton (31.0/29.0)
```

Number of Rules : 2

Time taken to build model: 0.04 seconds

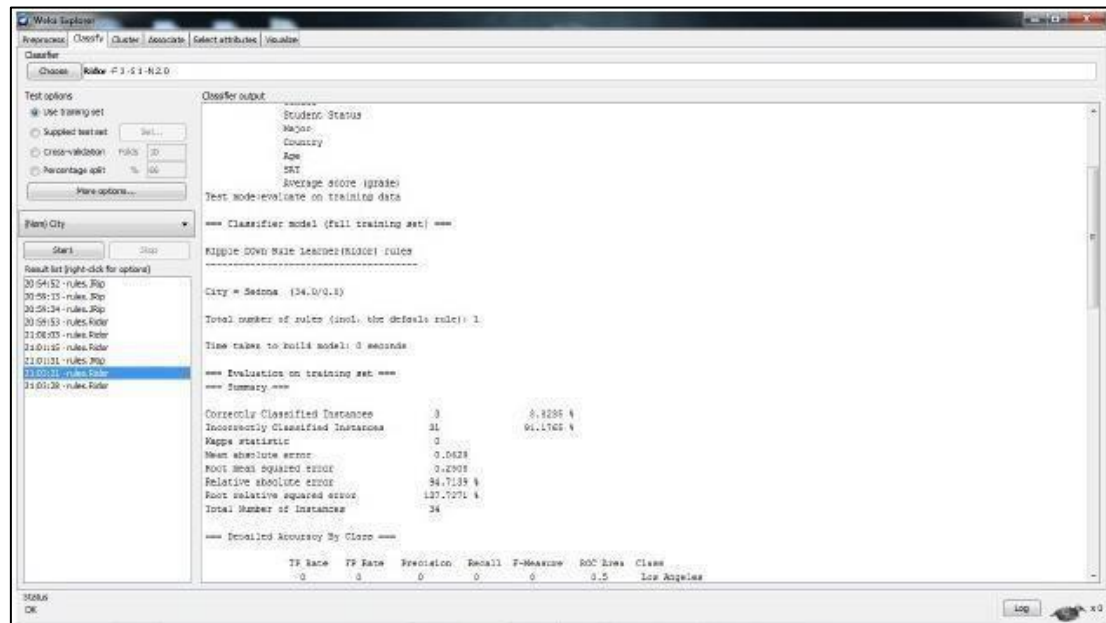
=== Evaluation on training set ===

Summary

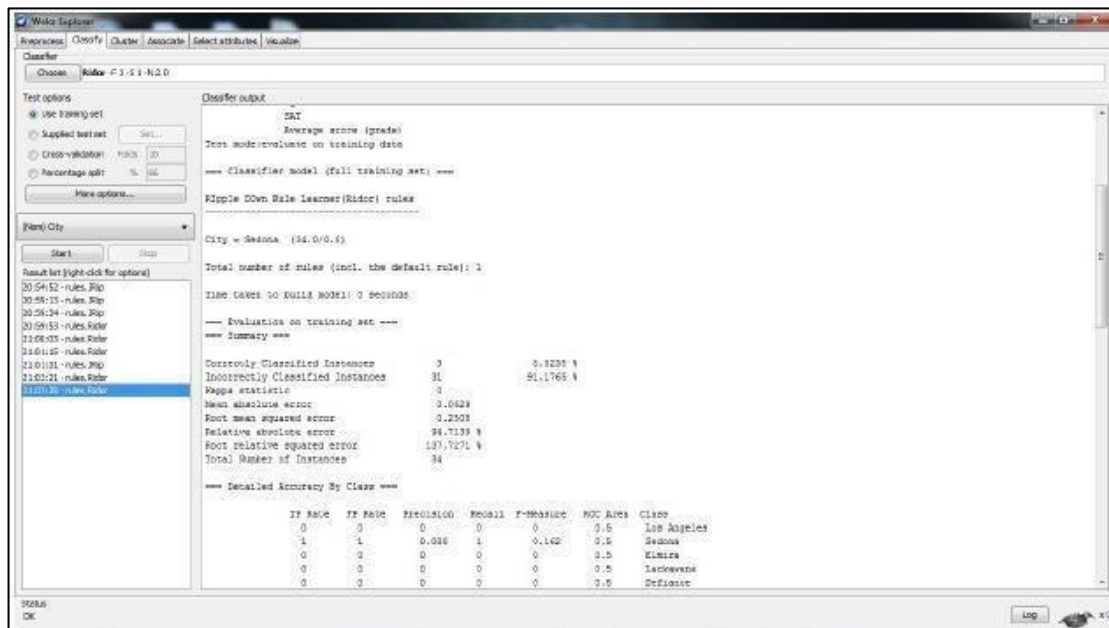
Metric	Value	Percentage
Correctly Classified Instances	5	14.7059 %
Incorrectly Classified Instances	29	85.2941 %
Kappa statistic	0.0422	
Mean absolute error	0.1764	
Root mean squared error	0.1764	
Relative absolute error	93.7088 %	
Root relative squared error	96.0937 %	
Total Number of Instances	34	

=== Detailed Accuracy By Class ===

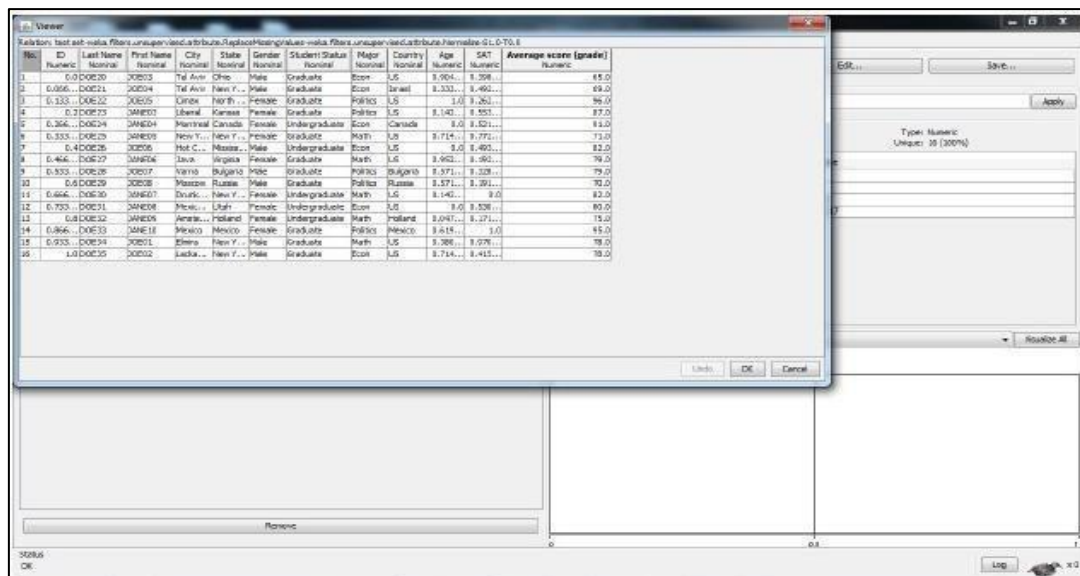
Ridor(seed=1):



Ridor(seed=2):



Test Data:



Ridor(seed=2):

Weka Explorer
 Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
 Classifier

Test options
☒ Use training set
☐ Supplied test set
☐ Cross-validation
☐ Percentage split

Classifier output
 DAT
 Average score (grade)
 Test mode: evaluate on training data
 --- Classifier model (full training set) ---
 Ripple Down Rule Learner (Ridor) rules
 City ~ Seasons (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
 Incorrectly Classified Instances: 31
 Kappa statistic: 0
 Mean absolute error: 0.0429
 Root mean squared error: 0.2003
 Relative absolute error: 94.7133 %
 Root relative squared error: 137.1271 %
 Total Number of Instances: 34

--- Detailed Accuracy By Class ---

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0	0.5	Los Angeles
1	1	1	0.000	1	0.142	0.5	Seasons
2	0	0	0	0	0	0.5	Kimira
3	0	0	0	0	0	0.5	Lacrosse
4	0	0	0	0	0	0.5	Defensive

Test Data:

Viewer
 Relations: test-set ~ data, file: unsupervised, attributes: ReplaceMissingValues ~ data, file: unsupervised, attributes: Normalize G1-G70, 6

No.	ID	Last Name	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score (grade)
1	0.0000	DOE10	JOHN	Tel Aviv	Chm	Male	Graduate	Econ	US	1.904	1.358	49.0
2	0.0000	DOE11	JOHN	Tel Aviv	Chm	Male	Graduate	Econ	Israel	1.933	1.460	49.0
3	0.133	DOE12	JANE	London	North	Female	Graduate	Politics	US	1.0	1.261	56.0
4	0.1000	DOE13	JANE	London	North	Female	Graduate	Politics	US	1.140	1.551	57.0
5	0.364	DOE14	JANE	Montreal	Canada	Female	Undergraduate	Econ	Canada	1.0	1.021	61.0
6	0.333	DOE15	JANE	New York	NY	Female	Graduate	Math	US	1.714	1.771	71.0
7	0.4000	DOE16	JOHN	Hot Springs	Ark	Male	Undergraduate	Econ	US	1.0	1.401	62.0
8	0.464	DOE17	JANE	Java	Virginia	Female	Graduate	Math	US	1.062	1.561	70.0
9	0.933	DOE18	JOHN	Varia	Bulgaria	Male	Graduate	Politics	Bulgaria	1.571	1.338	79.0
10	0.6000	DOE19	JOHN	Moscow	Russia	Male	Graduate	Politics	Russia	1.571	1.501	70.0
11	0.664	DOE20	JOHN	Druck	New York	Female	Undergraduate	Math	US	1.142	1.0	82.0
12	0.750	DOE21	JANE	New York	NY	Female	Undergraduate	Econ	US	1.0	1.538	80.0
13	0.8000	DOE22	JANE	Warsaw	Poland	Female	Undergraduate	Math	Poland	1.047	1.271	75.0
14	0.864	DOE23	JANE	Mexico	Mexico	Female	Graduate	Politics	Mexico	1.615	1.0	45.0
15	0.933	DOE24	JOHN	Elmira	New York	Male	Graduate	Math	US	1.286	1.076	79.0
16	1.0000	DOE25	JANE	Jacka	New York	Male	Graduate	Econ	US	1.714	1.415	78.0

JRip(seed=1):

Weka Explorer
 Preprocess | Classify | Cluster | Associate | Select attributes | Visualize
 Classifier

Test options
☒ Use training set
☐ Supplied test set
☐ Cross-validation
☐ Percentage split

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:
 =====
 => (City=Tel Aviv (35.0/24.0))
 Number of Rules: 1

Time taken to build model: 0.01 seconds

--- Evaluation on test set ---
 --- Summary ---
 Correctly Classified Instances: 1
 Incorrectly Classified Instances: 15
 Kappa statistic: 0
 Mean absolute error: 0.1171
 Root mean squared error: 0.2431
 Relative absolute error: 100 %
 Root relative squared error: 100.9114 %
 Total Number of Instances: 16

--- Detailed Accuracy By Class ---

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0	0	0	0	0	0	0.5	Los Angeles
1	1	1	0.000	1	0.142	0.5	Seasons
2	0	0	0	0	0	0.5	Kimira
3	0	0	0	0	0	0.5	Lacrosse
4	0	0	0	0	0	0.5	Defensive

JRip(seed=2):

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The classifier output pane displays the following information:

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:
=====

=> City=Tel Aviv (16.0/14.0)

Number of Rules : 1

Time taken to build model: 0.02 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	0.062	1	0.118	0.5	Tel Aviv
0	0	0	0	0	0.5	Classx
0	0	0	0	0	0.5	Liberal
0	0	0	0	0	0.5	Moscow
0	0	0	0	0	0.5	New York

JRip(seed=3):

The screenshot shows the Weka Explorer interface with the JRip classifier selected. The classifier output pane displays the following information:

Average score (grade):

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

== Classifier model (full training set) ==

Single Down Rule Learner(Ridor) rules

=====

City = Tel Aviv (14.0/1.0)

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	0.2431	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	Classx
0	0	0	0	0	0.5	Liberal
0	0	0	0	0	0.5	Moscow
0	0	0	0	0	0.5	New York

Ridor(seed=1):

The screenshot shows the Weka Explorer interface with the Ridor classifier selected. The classifier output pane displays the following information:

Age:
SAT:
Average score (grade):

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 ===

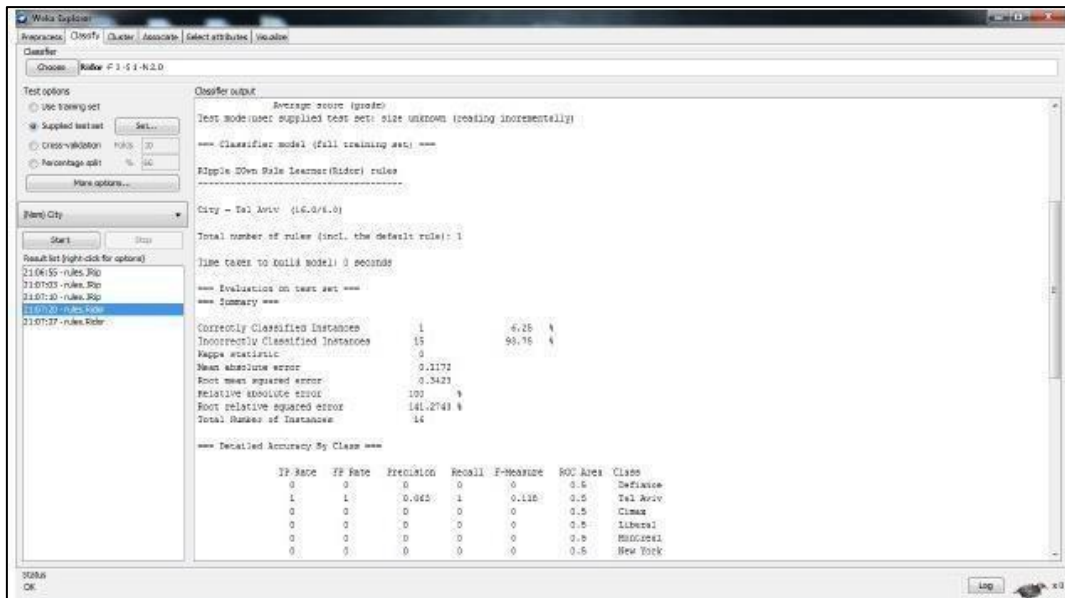
=== Summary ===

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	0.062	1	0.118	0.5	Tel Aviv
0	0	0	0	0	0.5	Classx

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
JRip	Seed=1	Root Mean Squared Error=0.2431 Mean Absolute Error=0.1172
JRip	Seed =2	Root Mean Squared Error=0.2431 Mean Absolute Error=0.1172
JRip	Seed =3	Root Mean Squared Error=0.2431 Mean Absolute Error=0.1172
Ridor	Seed =1	Root Mean Squared Error=0.3423 Mean Absolute Error=0.1172
Ridor	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: 9

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:

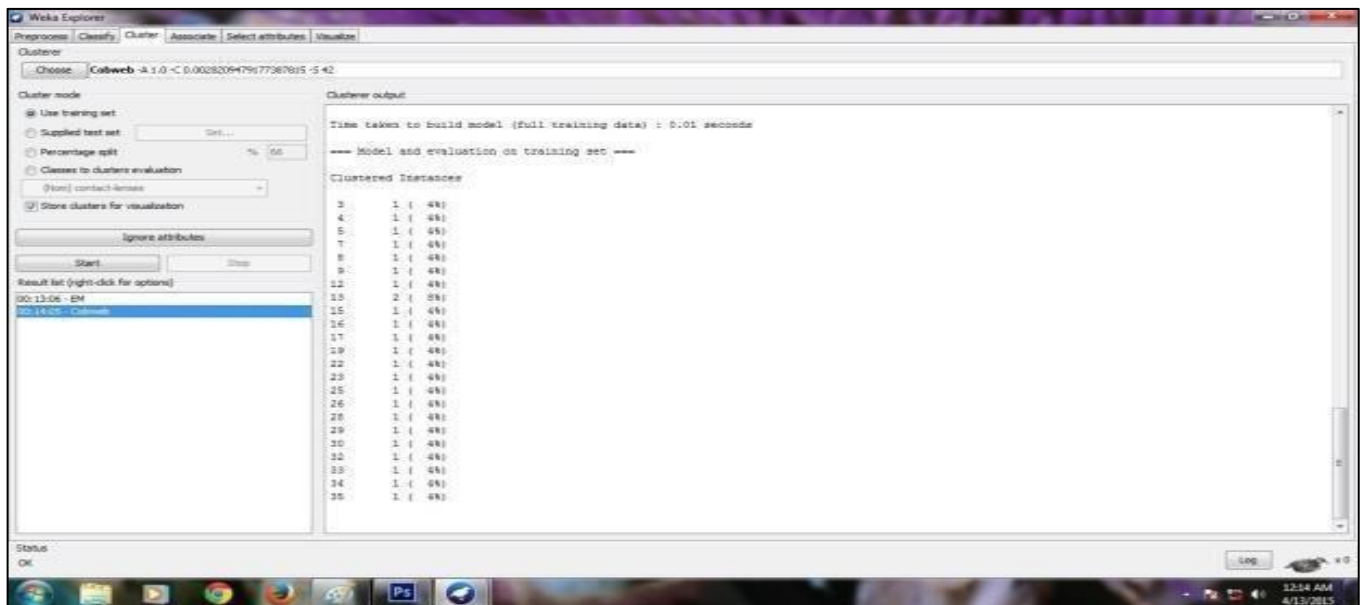
Web content mining can be used for mining of useful data, information and knowledge from web page content. Web structure mining helps to find useful knowledge or information pattern from the structure of hyperlinks. Due to heterogeneity and absence of structure in web data, automated discovery of new 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 a user wants to search for a particular book, then search engine provides the list of suggestions.

ALGORITHM:

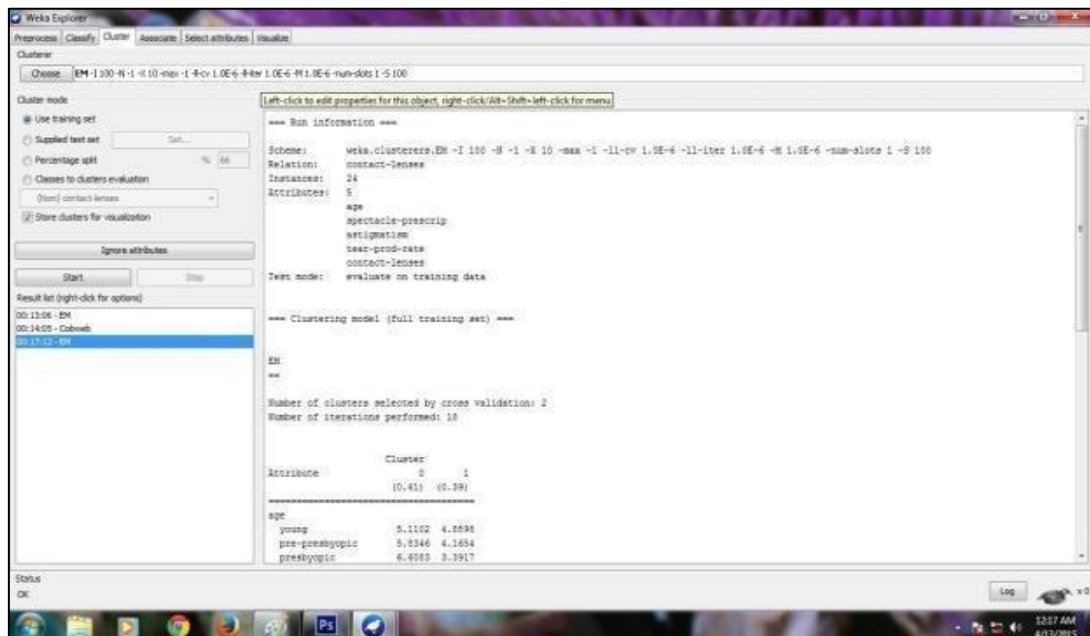
1. Open the weka tool.
2. Download a dataset by using UCI.
3. Apply replace missing values.
4. Apply normalize filter.
5. Click the cluster tab.
6. Apply all ALGORITHMS one by one.
7. Find the no of clusters that are formed
8. Note the OUTPUT.

OUTPUT:

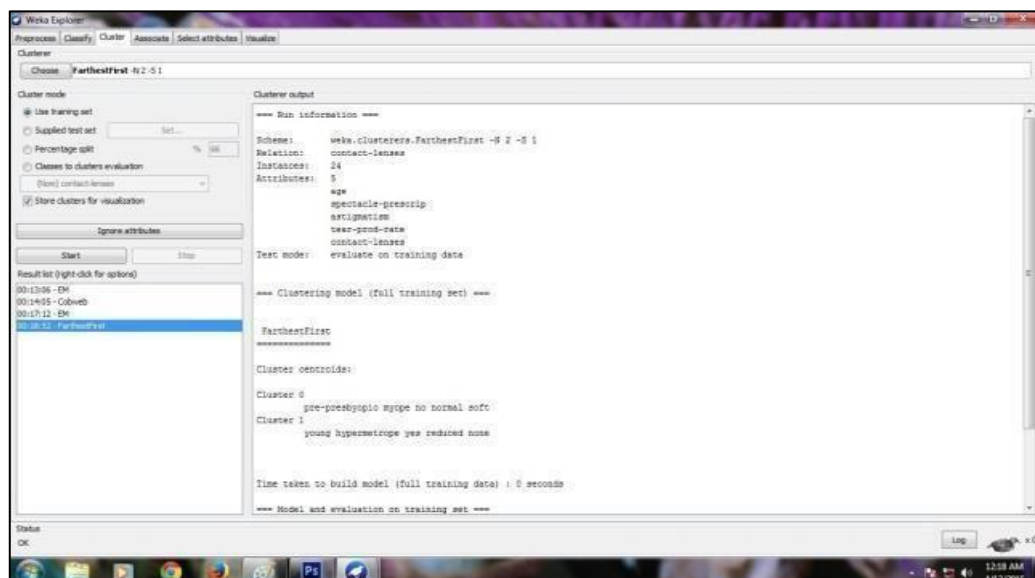
Cobweb



EM



Farthest First:



Filtered Cluster:

The screenshot shows the Weka Explorer interface with the SimpleKMeans clustering algorithm applied to a filtered dataset. The 'Cluster mode' section on the left shows 'Percentage split' at 66%. The 'Cluster output' section on the right displays the following information:

Relation: contact-lenses
Instances: 24
Attributes: 5
age
spectacle-prescrip
astigmatism
tear-prod-rate
contact-lenses
Test model: evaluate on training data

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

Clustered Instances

Cluster	Instances
0	20 (83%)
1	4 (17%)

Hierarchical Cluster

The screenshot shows the Weka Explorer interface with the HierarchicalCluster clustering algorithm applied to a filtered dataset. The 'Cluster mode' section on the left shows 'Percentage split' at 66%. The 'Cluster output' section on the right displays the following information:

Discrete Estimator. Counts = 3 5 (Total = 12)
Attribute: tear-prod-rate
Discrete Estimator. Counts = 3 5 (Total = 12)
Attribute: contact-lenses
Discrete Estimator. Counts = 2 3 1 (Total = 13)
Cluster: 1 Prior probabilities: 0.3519

Attribute: age
Discrete Estimator. Counts = 2 1 1 (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 3 (Total = 7)
Attribute: contact-lenses
Discrete Estimator. Counts = 1 2 1 (Total = 8)

Time taken to build model (percentage split) : 0 seconds

Clustered Instances

Cluster	Instances
0	4 (17%)
1	3 (13%)

log likelihood: -4.55298

Simple KMeans:

The screenshot shows the Weka Explorer interface with the Simple KMeans algorithm applied to a dataset. The 'Clusterer' tab is selected, and the 'SimpleKMeans' algorithm is chosen. The 'Cluster mode' section shows 'Percentage split' set to 55%. The 'Cluster output' section displays the following information:

=== Model and evaluation on test split ===
KMeans
=====

Number of iterations: 3
Within cluster sum of squared errors: 25.0
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (15)	Cluster# 0 (10)	Cluster# 1 (5)
age		presbyopic	presbyopic
spectacle-prescrip		myope	hypermetrope
astigmatism		yes	no
tear-prod-rate		reduced	reduced
contact-lenses		none	none

Time taken to build model (percentage split) : 0 seconds

Clustered Instances

Cluster#	Count	Percentage
0	4	44%
1	5	56%

The status bar at the bottom shows 'OK' and a 'Log' button.

RESULT:

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

EX. No: 10

TEXT MINING

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:

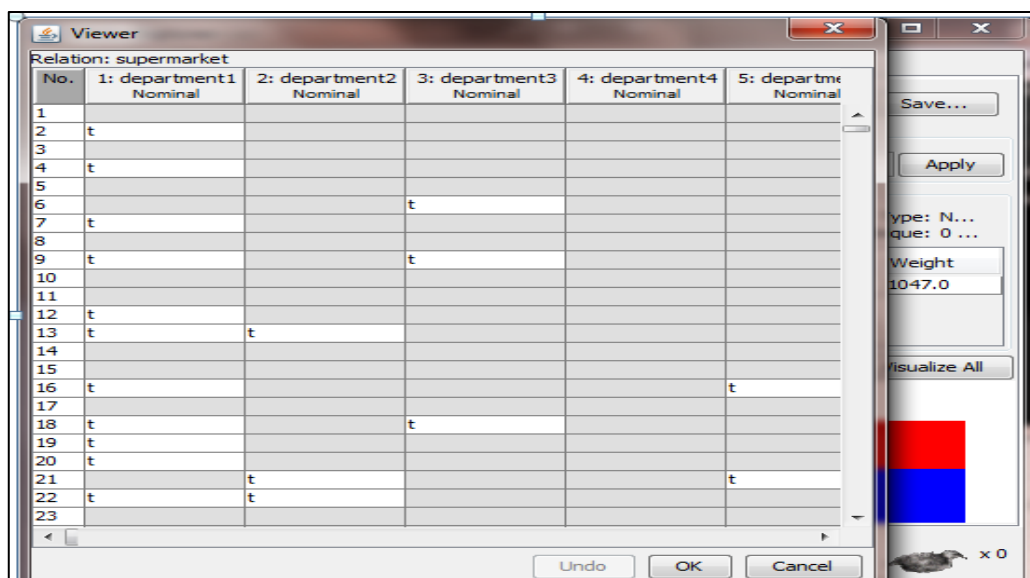
It collects sets of keywords or terms that often happen together and afterward discover the association relationship among them. First, it preprocesses the text data by parsing, stemming, removing stop words, etc. Once it pre-processed the data, then it induces association mining ALGORITHMs. Here, human 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:

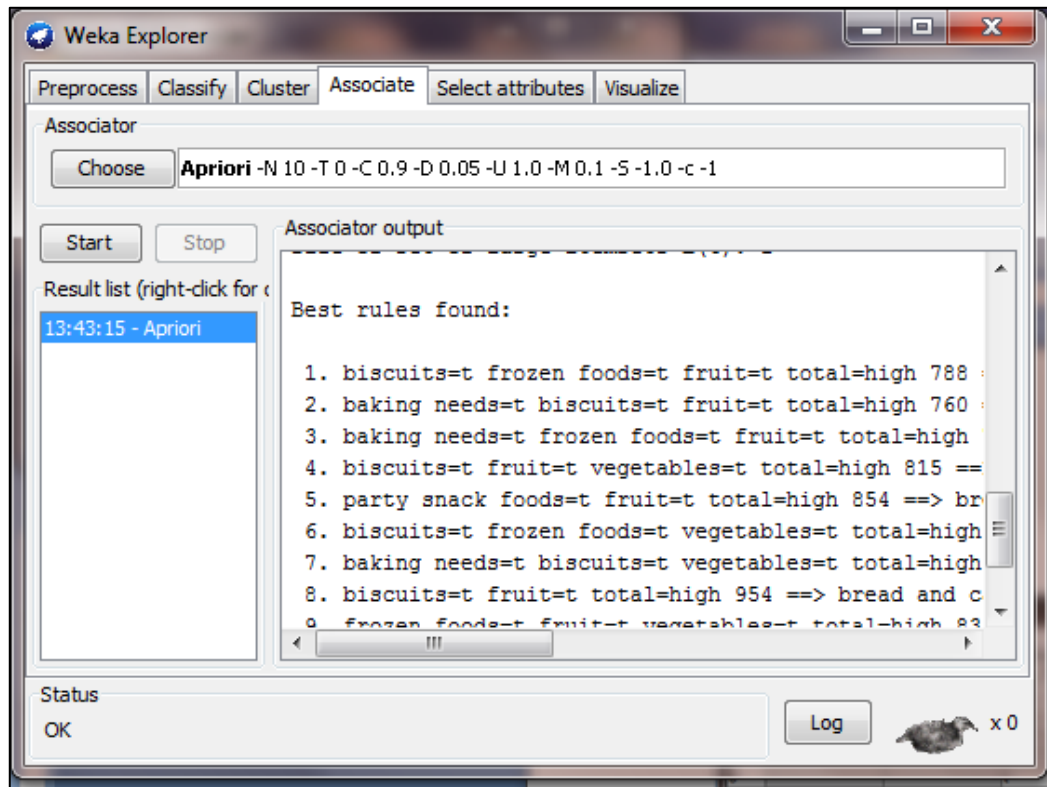
Super Market data set



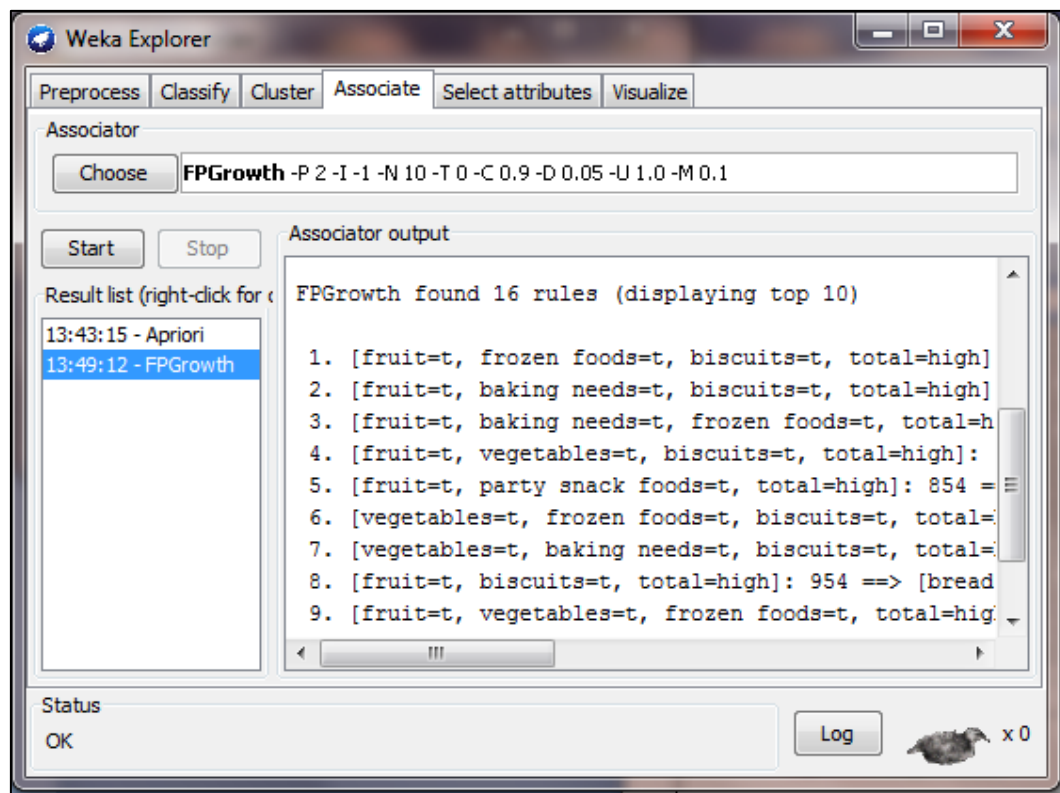
No.	1: department1 Nominal	2: department2 Nominal	3: department3 Nominal	4: department4 Nominal
1				
2	t			
3				
4	t			
5				
6			t	
7	t			
8				
9	t		t	
10				
11				
12	t			
13	t	t		
14				
15				
16	t			t
17				
18	t		t	
19	t			
20	t			
21		t		t
22	t	t		
23				

OUTPUT:

Apriori Algorithm:



FP-Growth Algorithm:



RESULT:

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

Ex. No: 11

DESIGN OF FACT AND DIMENSION TABLES

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.

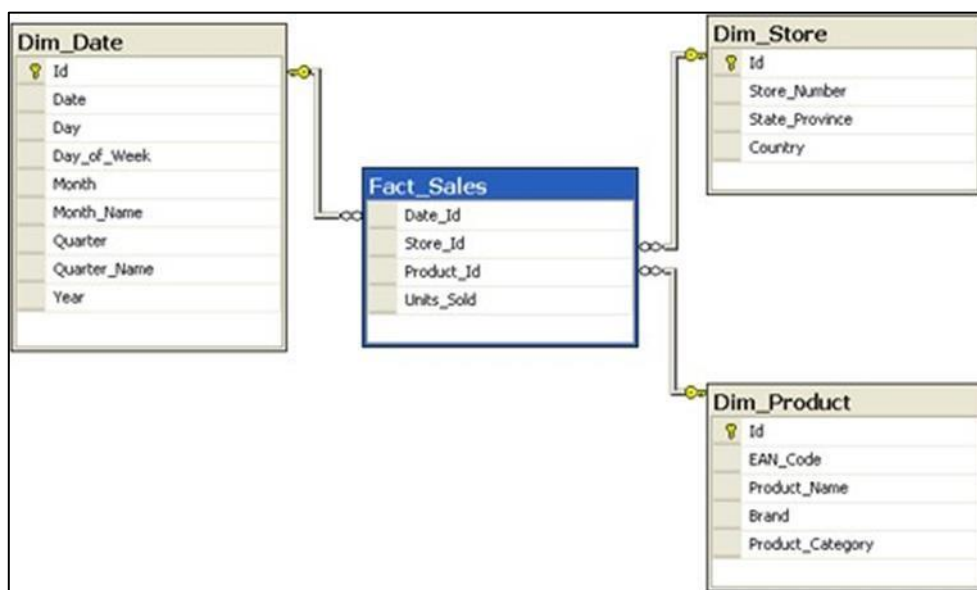
Designing fact table steps:

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.

EX. No: 12

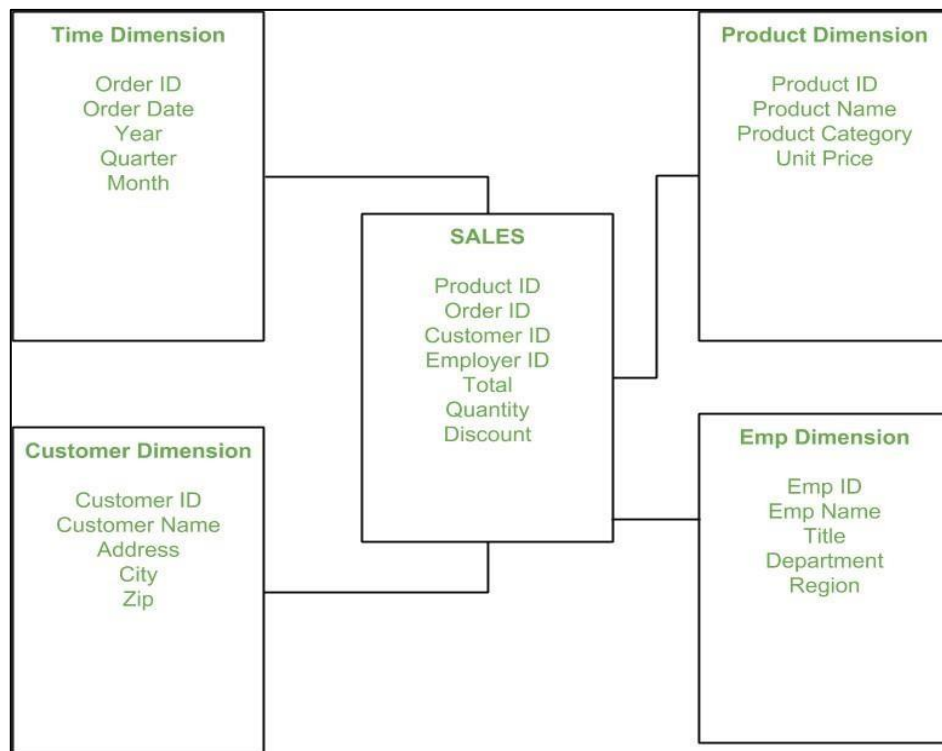
GENERATING GRAPHS FOR STAR SCHEMA

AIM:

To generate graphs for star schema.

INTRODUCTION:

Star schema is the fundamental schema among the data mart schema and it is simplest. This schema is widely used to develop or build a data warehouse and dimensional data marts. It includes one or more fact tables indexing any number of dimensional tables. The star schema is a necessary case of the snowflake schema. It is also efficient for handling basic queries. It is said to be star as its physical model resembles to the star shape having a fact table at its center and the dimension tables at its peripheral representing the star's points.



In the above demonstration, **SALES** is a fact table having attributes i.e. (Product ID, Order ID, Customer ID, Employer ID, Total, Quantity, Discount) which references to the dimension tables. Employee dimension table contains the attributes: Emp ID, Emp Name, Title, Department and Region. Product dimension table contains the attributes: Product ID, Product Name, Product Category, Unit Price. Customer dimension table contains the attributes: Customer ID, Customer Name, Address, City, Zip. Time dimension table contains the attributes: Order ID, Order Date, Year, Quarter, Month.

In Star Schema, Business process data, that holds the quantitative data about a business is distributed in fact tables, and dimensions which are descriptive characteristics related to fact data. Sales price, sale quantity, distant, speed, weight, and weight measurements are few examples of fact data in star schema. Often, A Star Schema having multiple dimensions is termed as Centipede Schema. It is easy to handle a star schema which has dimensions of few attributes.

RESULT:

Thus, the graphs for star schema are generated.