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

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	**gineering Knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
PO2	**oblem 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	**sign/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	**nduct 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	**odern 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	**e 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	**vironment 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	**hics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
PO 9	**dividual and Team Work:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
PO 10	**mmunication:** 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	**oject 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	**fe-long 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 offeringDept./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 varyingconfidence (%) 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:

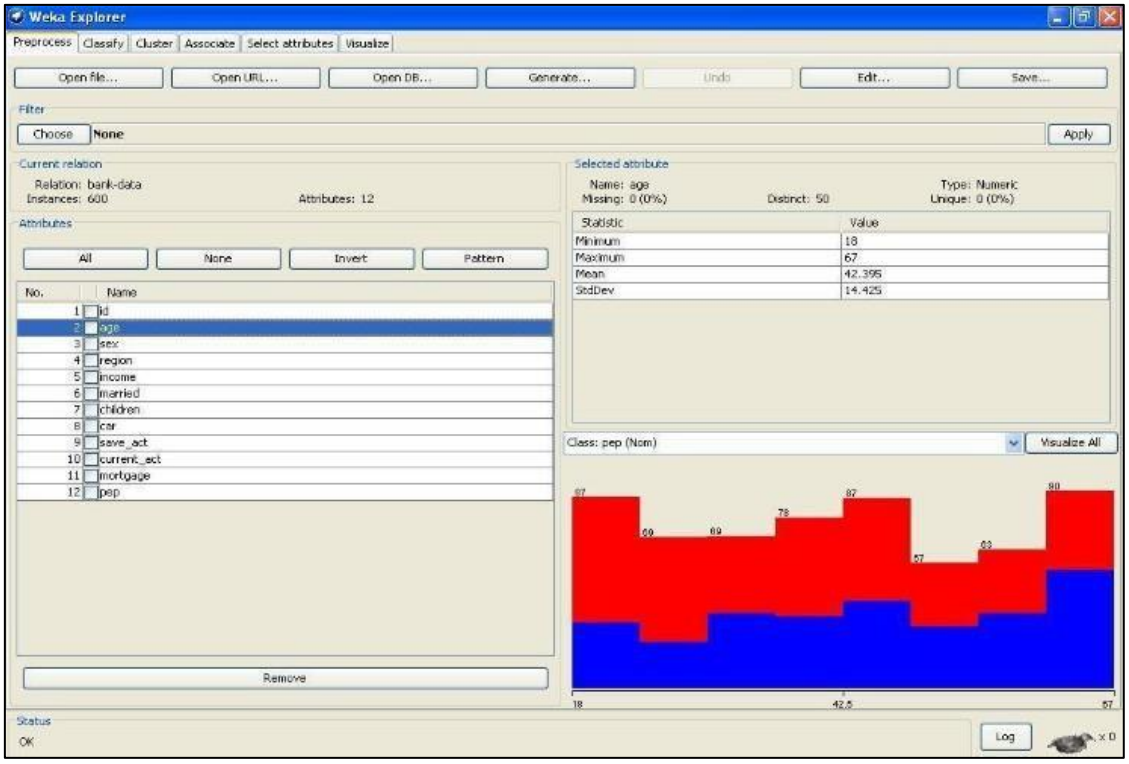
To list all the categorical (or nominal) attributes and the real-valued attributes separately.

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:

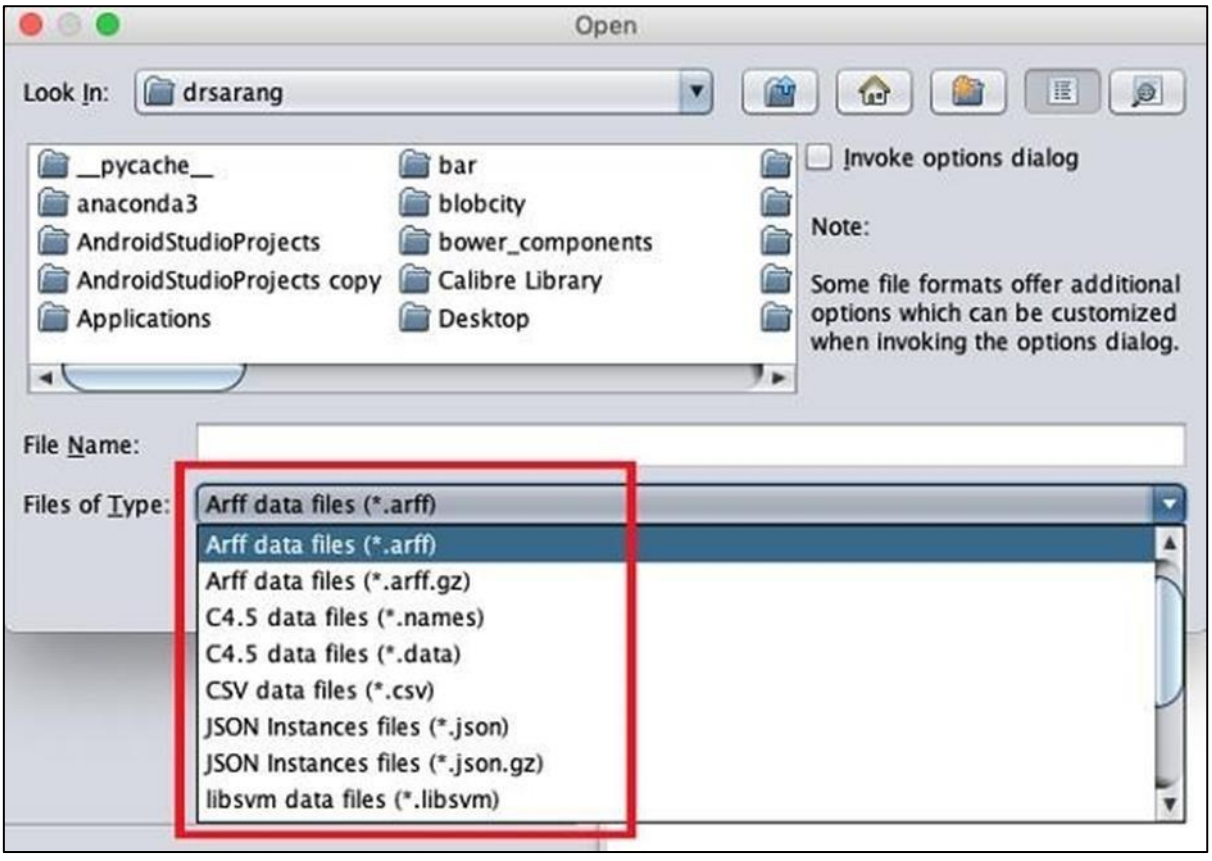
To study the file formats for the data mining.

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

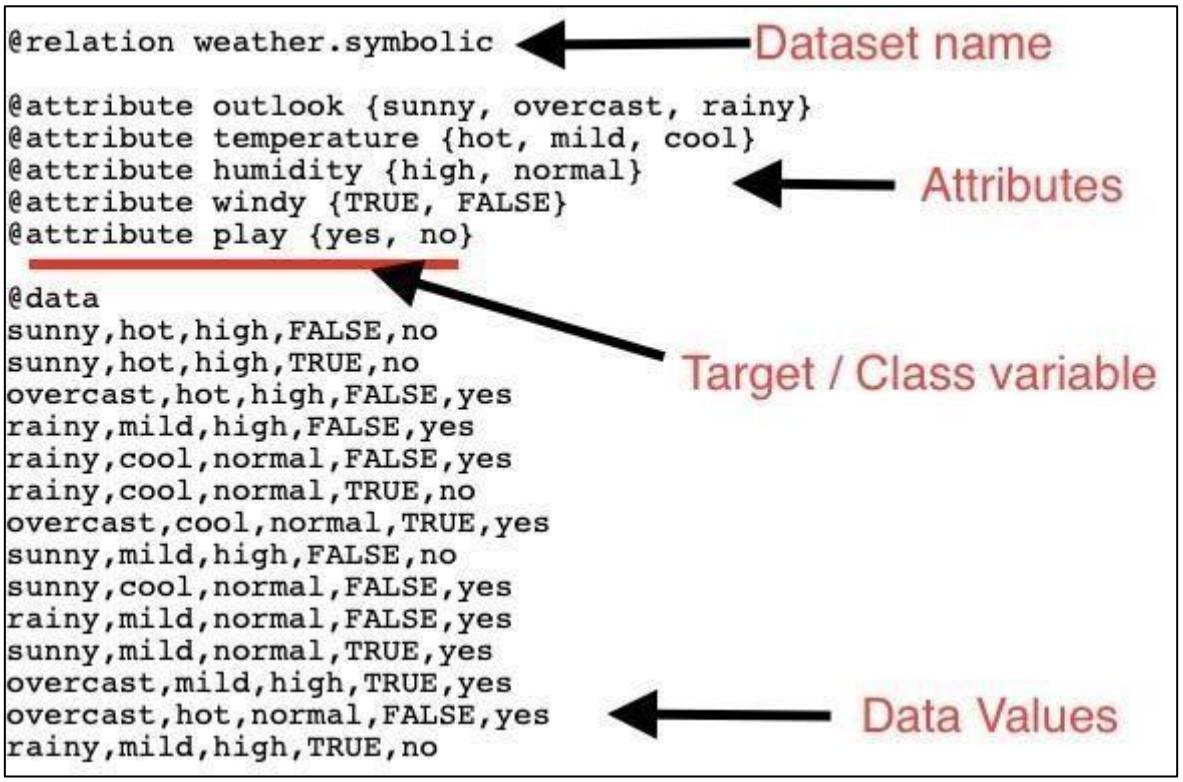
The types of files that it supports are listed in the drop-down list box at the bottom of the screen.This is shown in the screenshot given below.



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:



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:

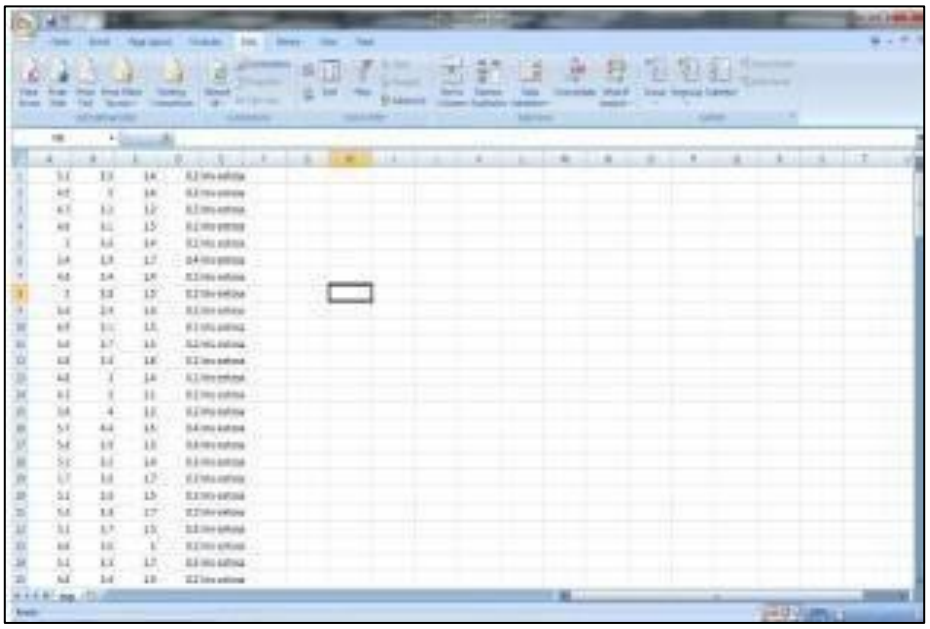
Most of the data that we have collected from public forum is in the text format that cannot be read byWeka tool. Since Weka (Data Mining tool) recognizes the data in ARFF format only we have to convert the text file into ARFF file.

ALGORITHM:

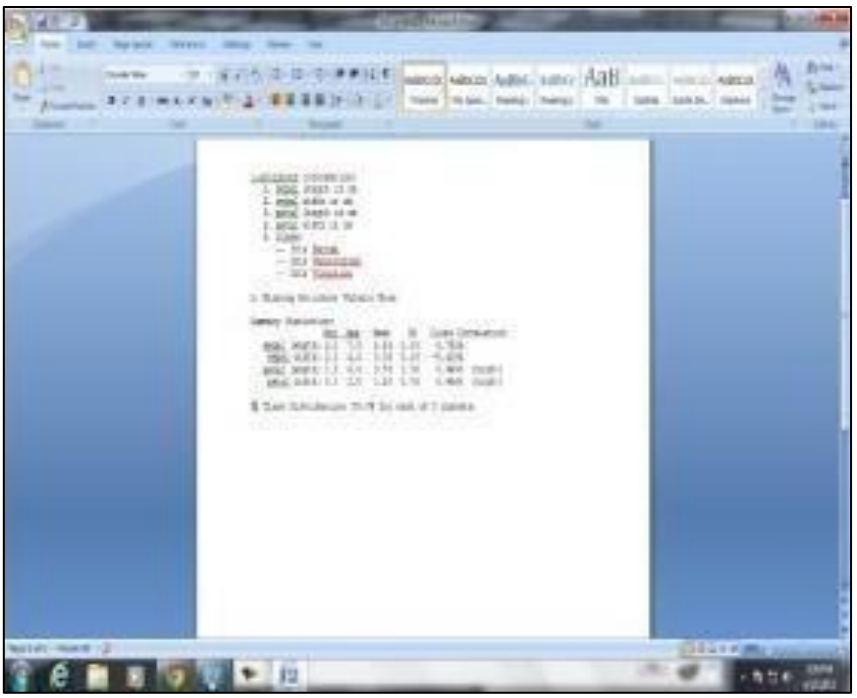
1. Download any data set from UCI data repository.
2. Open the same data file from excel. It will ask for delimiter (which produce column) in excel.
3. Add one row at the top of the data.
4. Enter header for each column.
5. Save file as .CSV (Comma Separated Values) format.
6. Open Weka tool and open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

OUTPUT:

Data Text File:



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:

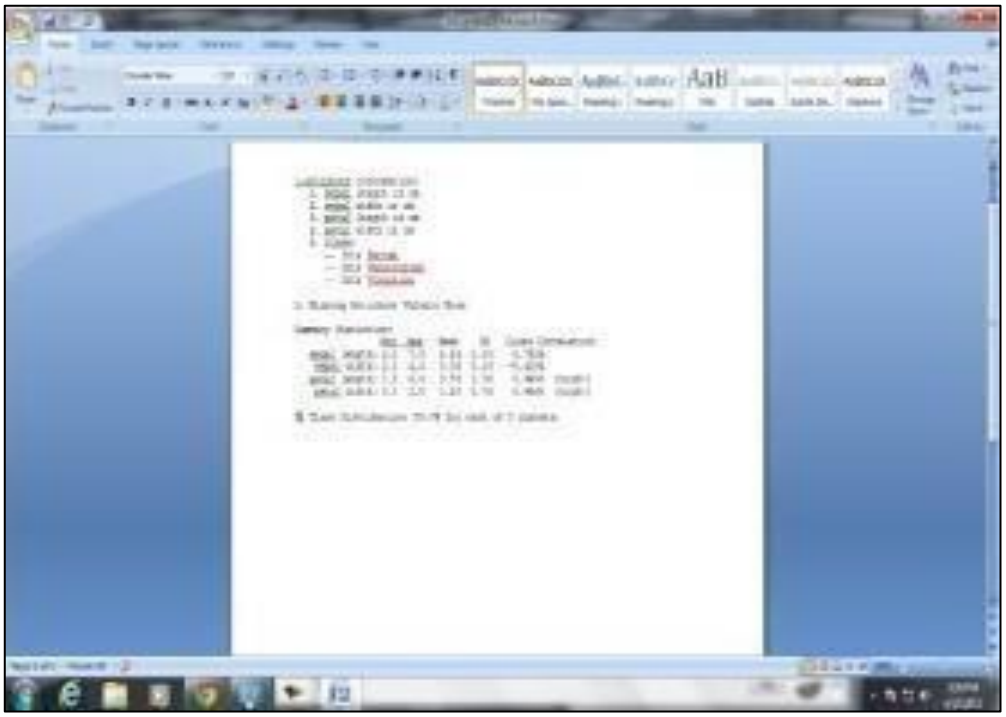
Since the data in the Weka tool is in ARFF file format we have to convert the ARFF file to text format for further processing.

ALGORITHM:

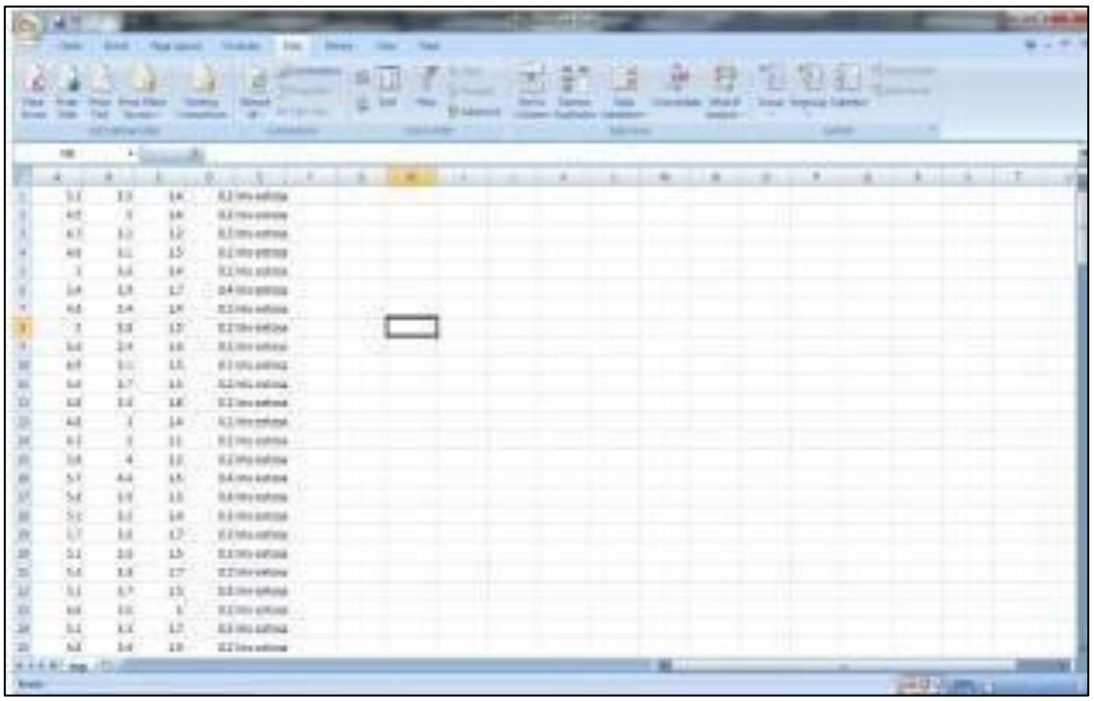
1. Open any ARFF file in Weka tool.
2. Save the file as CSV format.
3. Open the CSV file in MS-EXCEL.
4. Remove some rows and add corresponding header to the data.
5. Save it as text file with the desire delimiter.

OUTPUT:

Data ARFF File:



Data Text File:



RESULT:

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

EX. No: 4

TRAINING THE GIVEN DATASET FOR AN APPLICATION

AIM:

To apply the concept of Linear Regression for training 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 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.

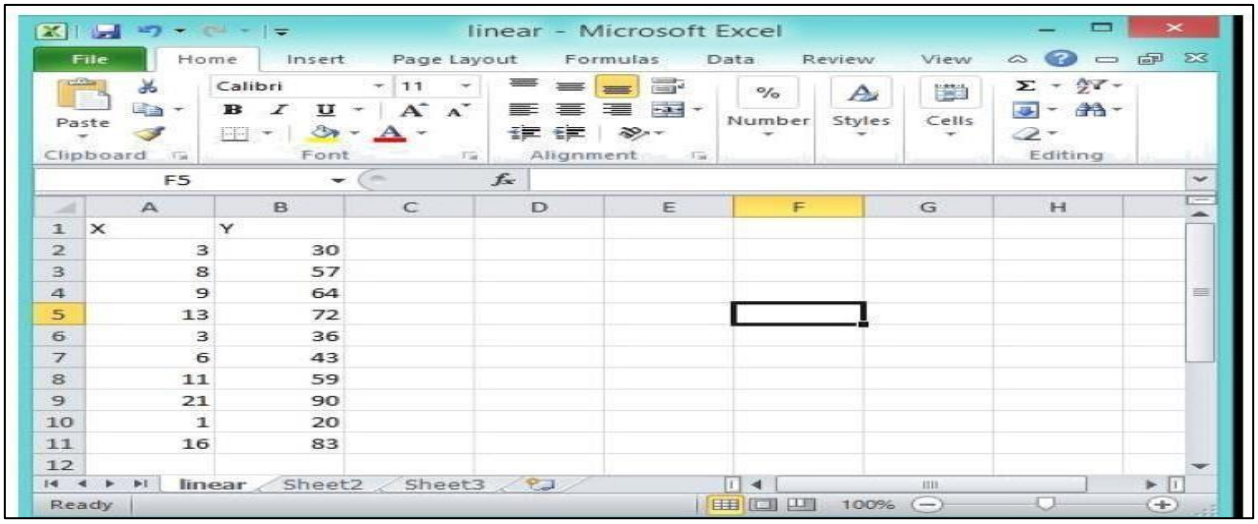
LINEAR REGRESSION:

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: Y=ax+b

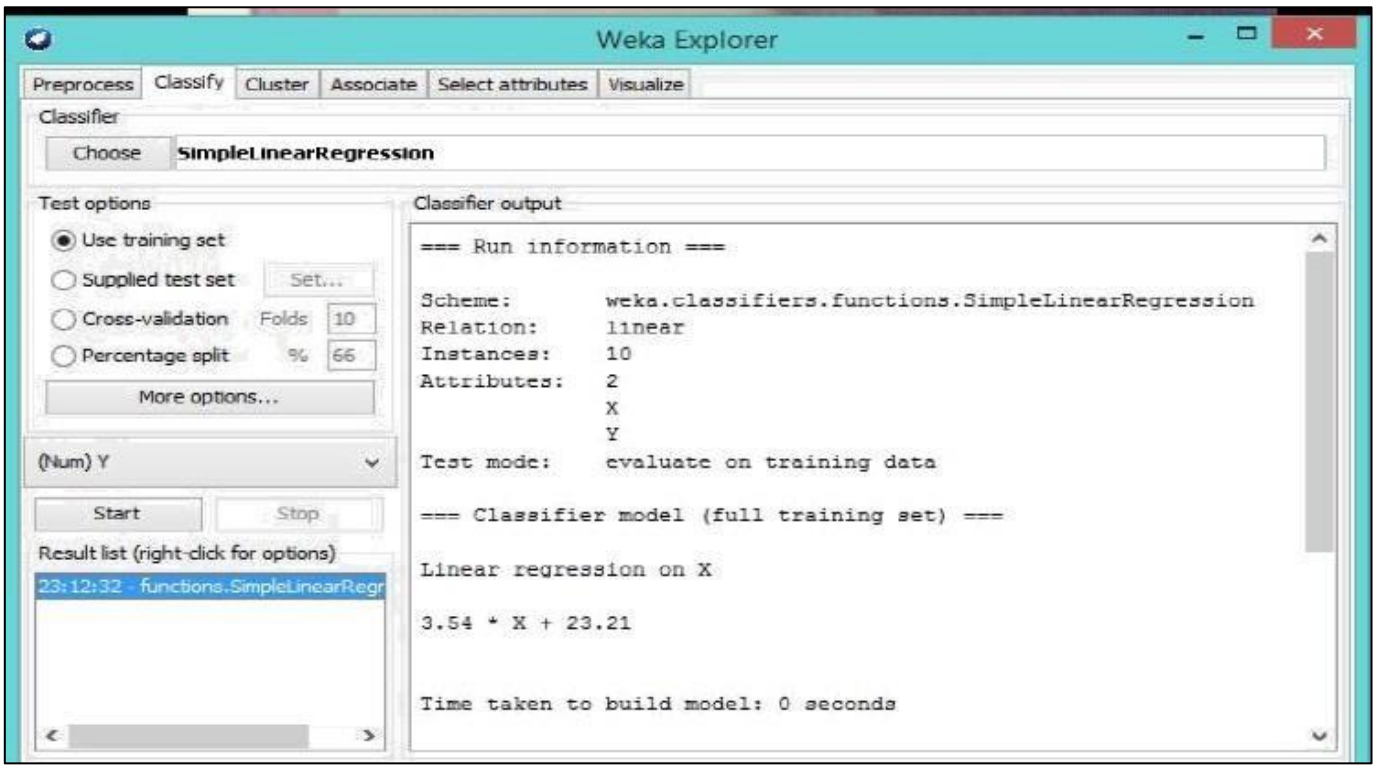
PROBLEM:

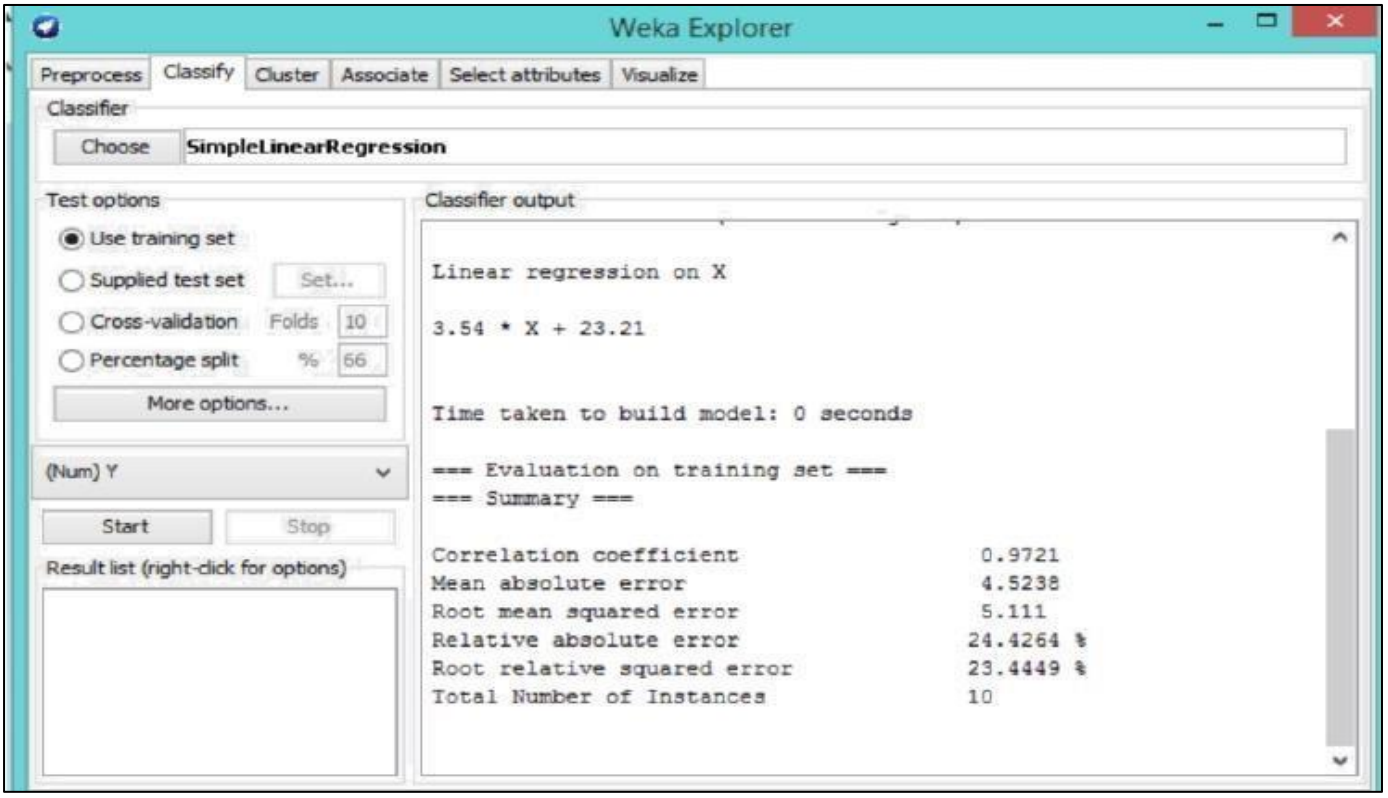
Consider the dataset below where x is the number of working experiences of a college graduate and y is the corresponding salary of the graduate. Build a regression equation and predict the salary of college graduate whose experience is 10 years.

INPUT:



OUTPUT:





RESULT:

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

EX. No: 5

TESTING THE GIVEN DATASET FOR AN APPLICATION

AIM:

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

ALGORITHM:

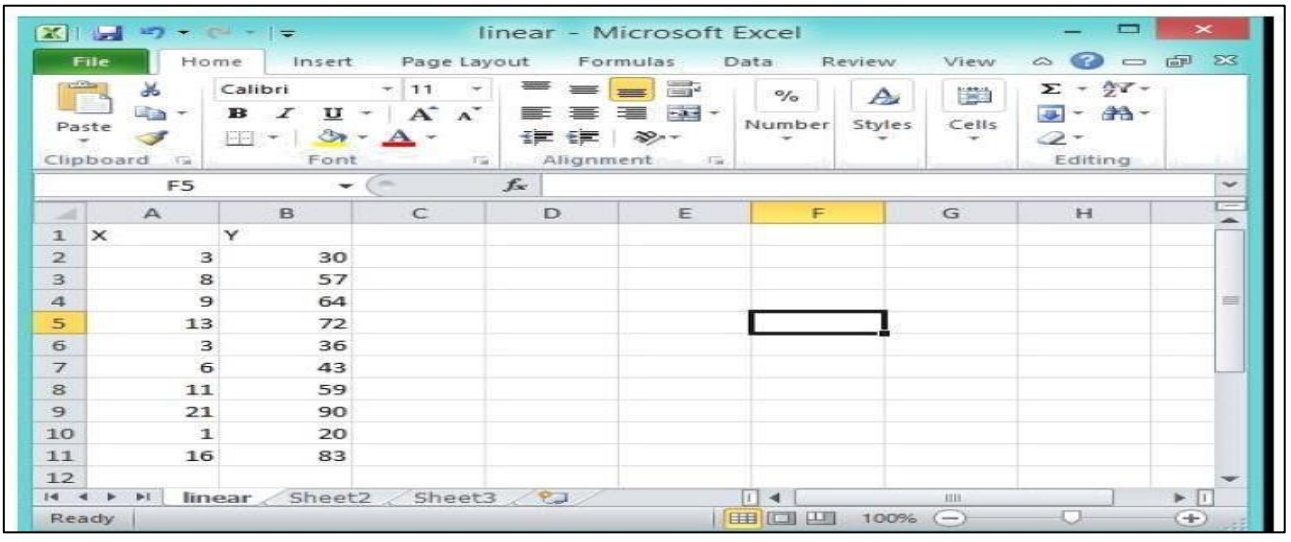
1. Open the weka tool.
2. Download a dataset by using UCI.
3. Apply replace missing values.
4. Apply normalize filter.
5. Click the Classification Tab.
6. Apply Navie Bayes Classification.
7. Find the Classified Value.
8. Note the OUTPUT.

Bayes' Theorem in the Classification Context:

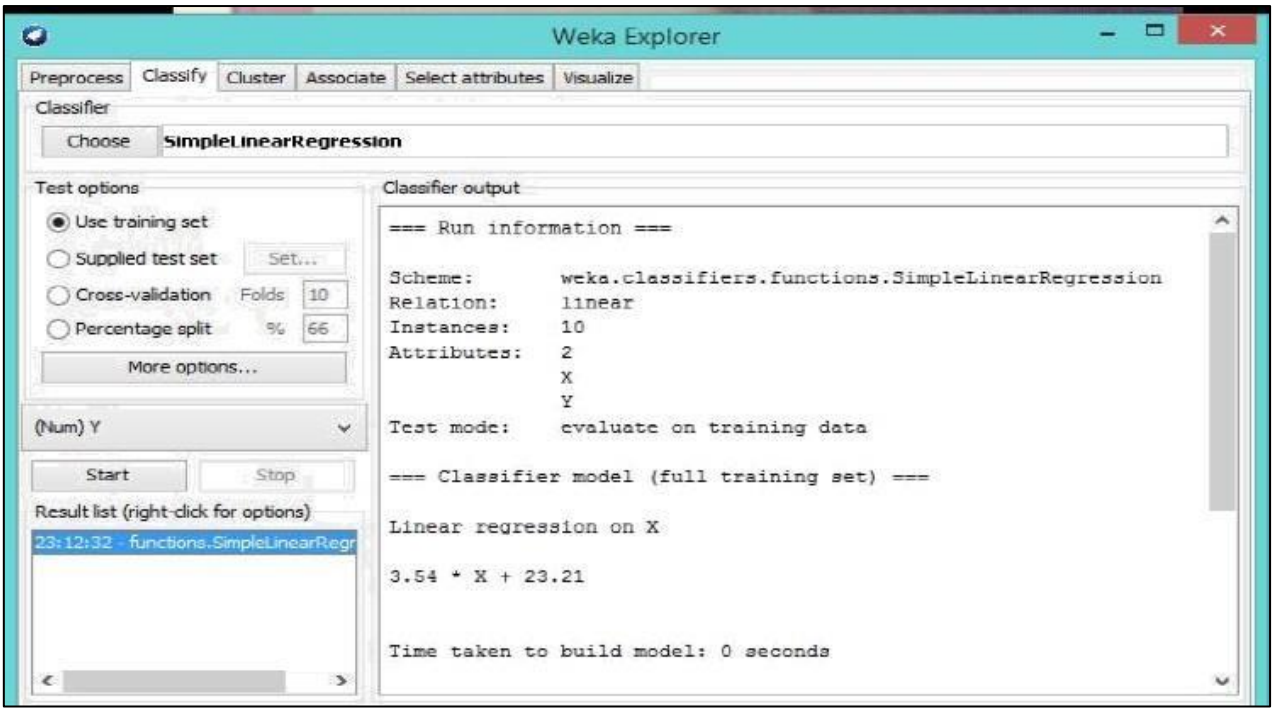
X is a data tuple. In Bayesian term it is considered "evidence". **H** is some hypothesis that **X** belongs to a specified class **C**. **P(H|X)** is the posterior probability of **H** conditioned on **X** .

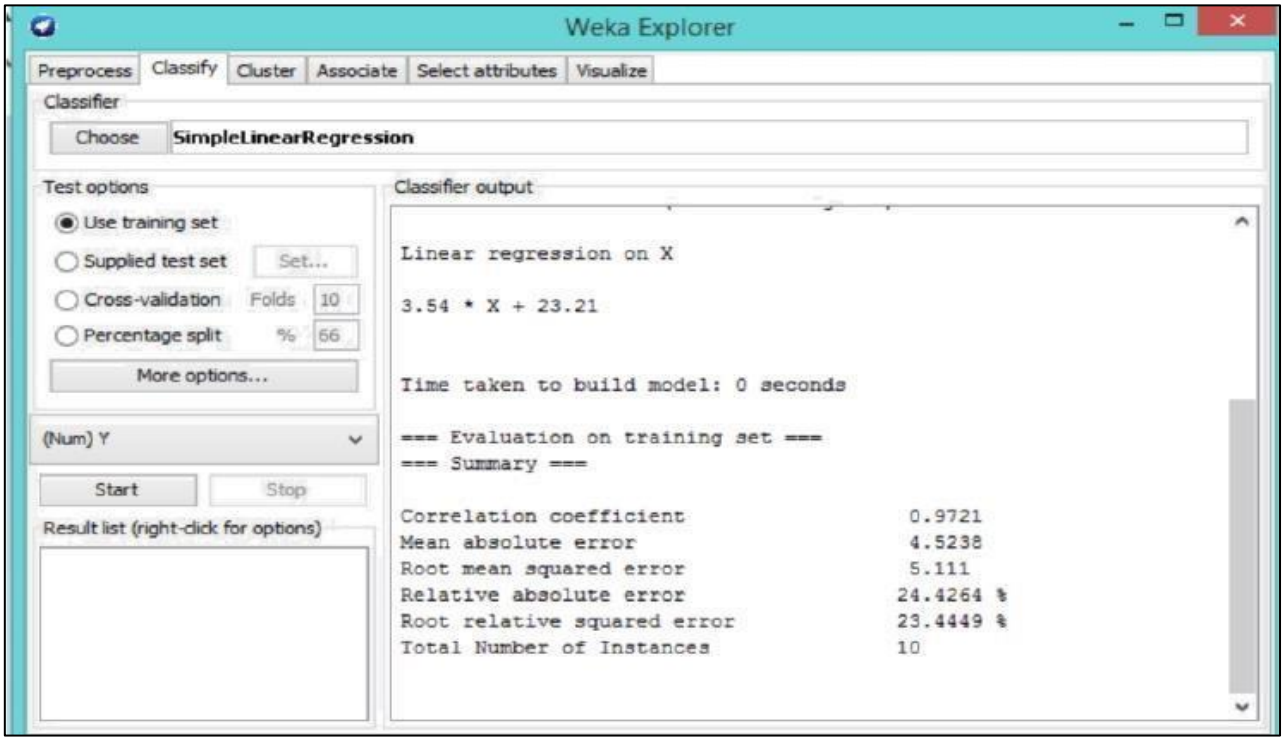
Example: predict whether a costumer will buy a computer or not " Costumers are described by two attributes: age and income " **X** is a 35 years-old costumer with an income of 40k " **H** is the hypothesis that the costumer will buy a computer " **P(H|X)** reflects the probability that costumer **X**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

INPUT:



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.

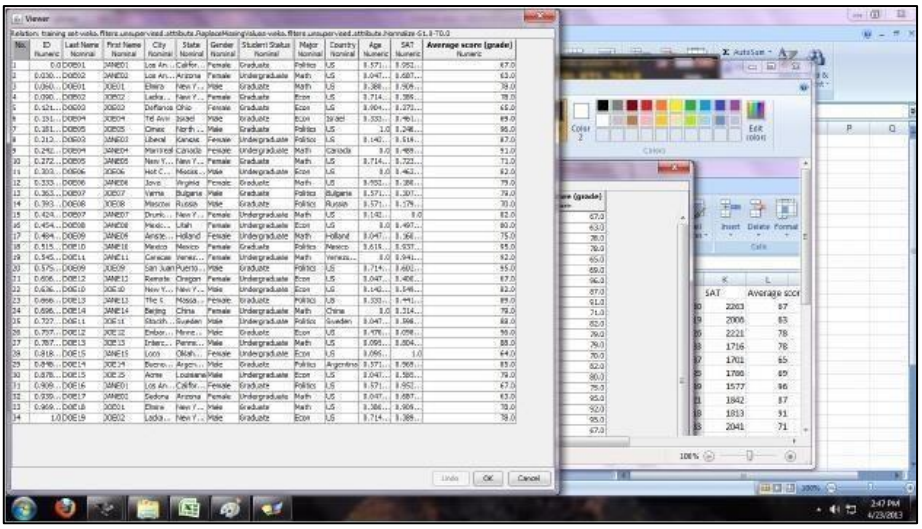
OBJECTIVES:

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

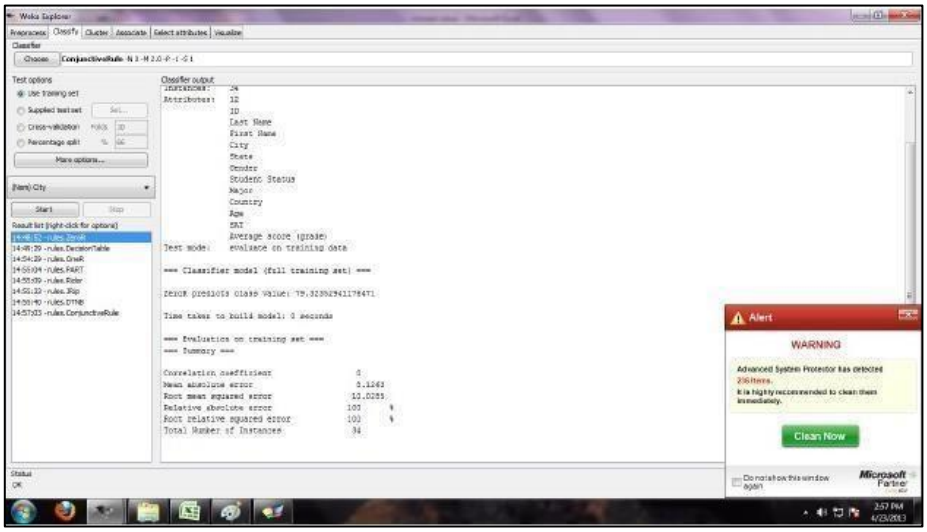
ALGORITHM:

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.

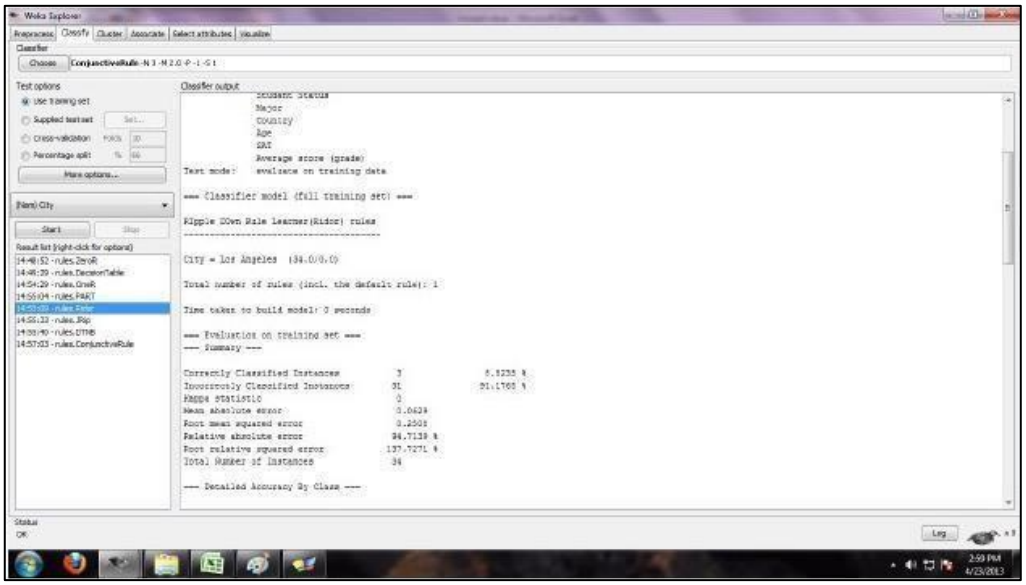
TRAINING DATA:



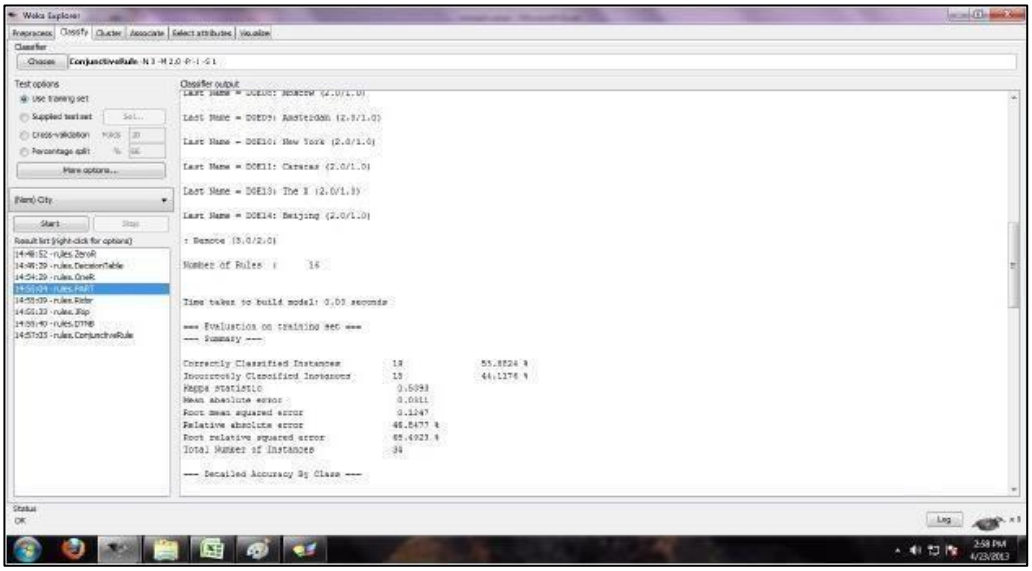
ZeroR:



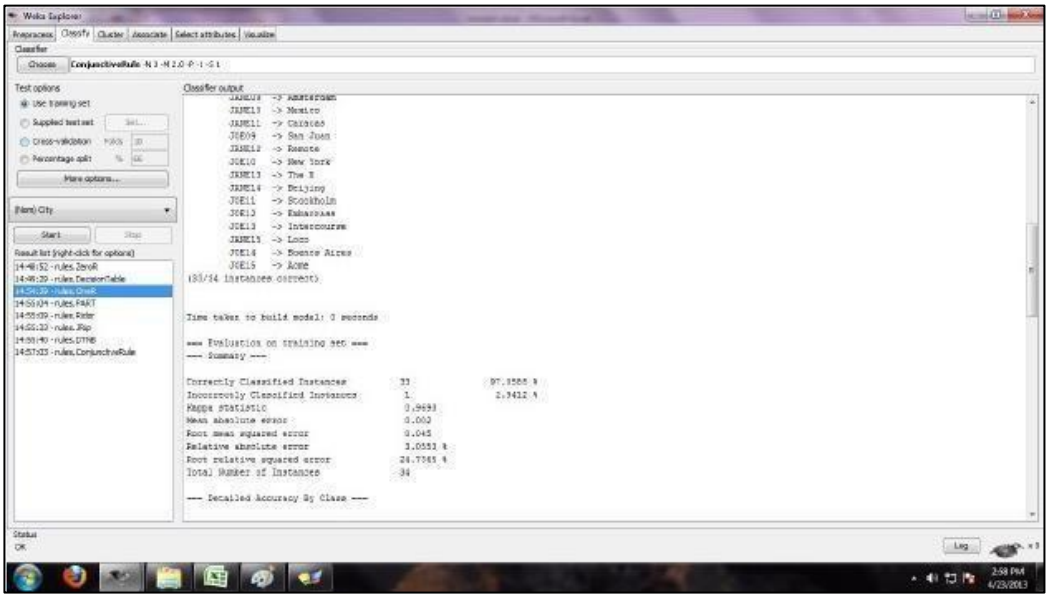
Ridor:



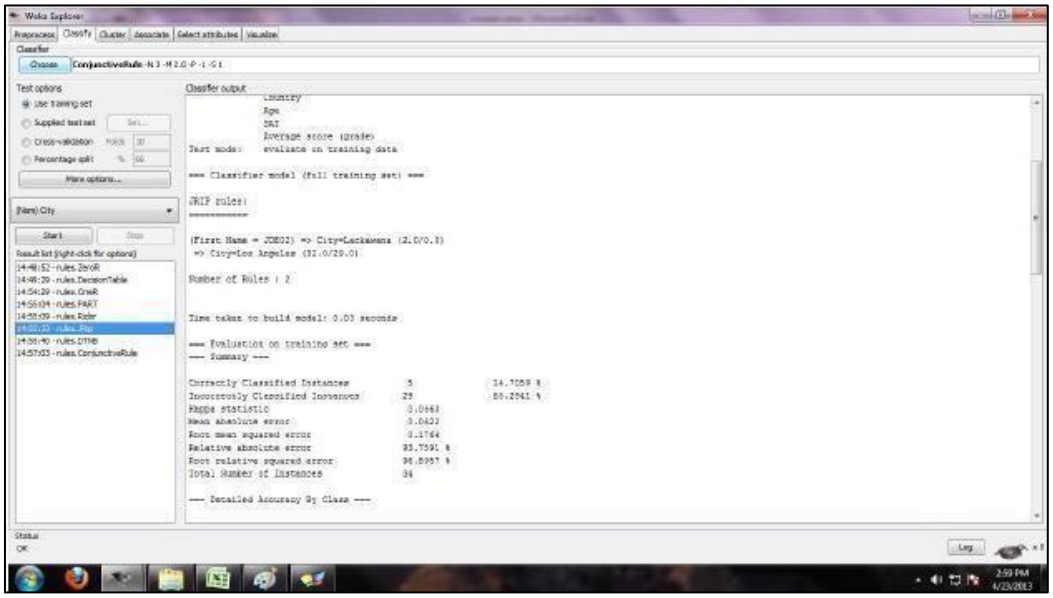
PART:



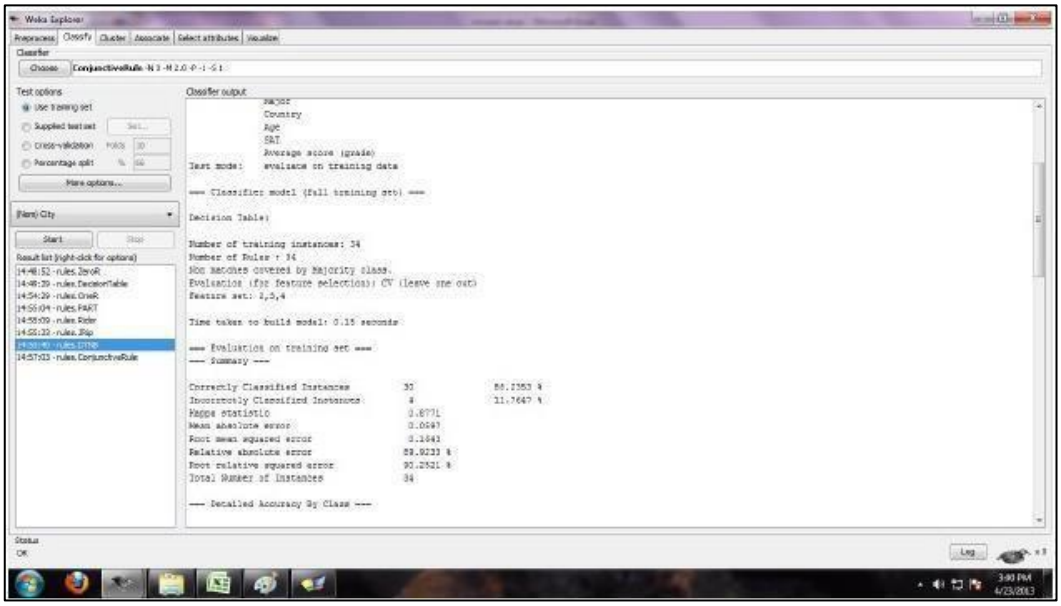
OneR:



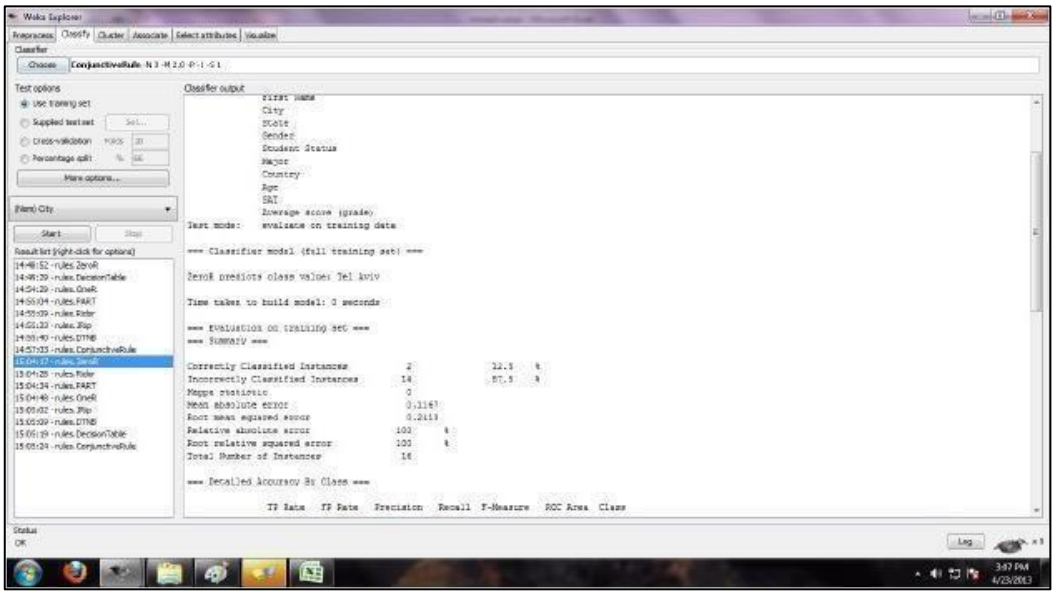
JRip:



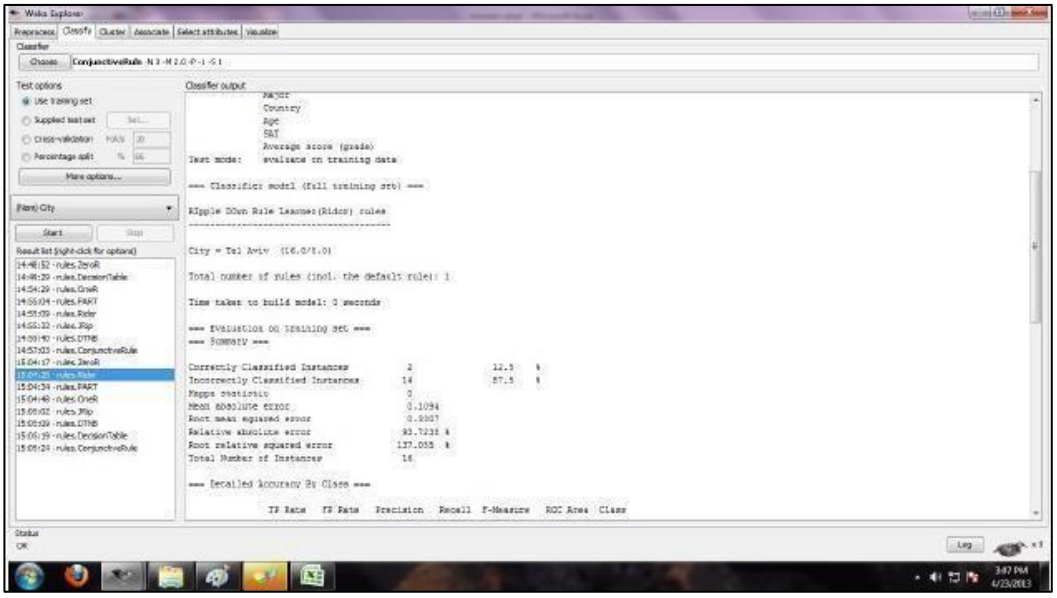
DTNB:



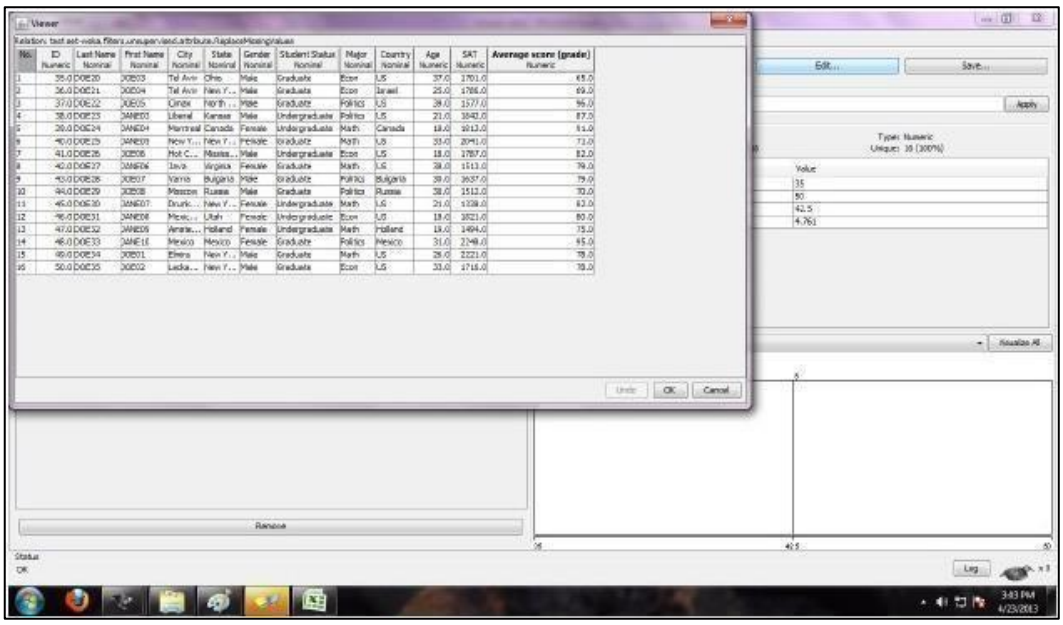
TEST DATA:



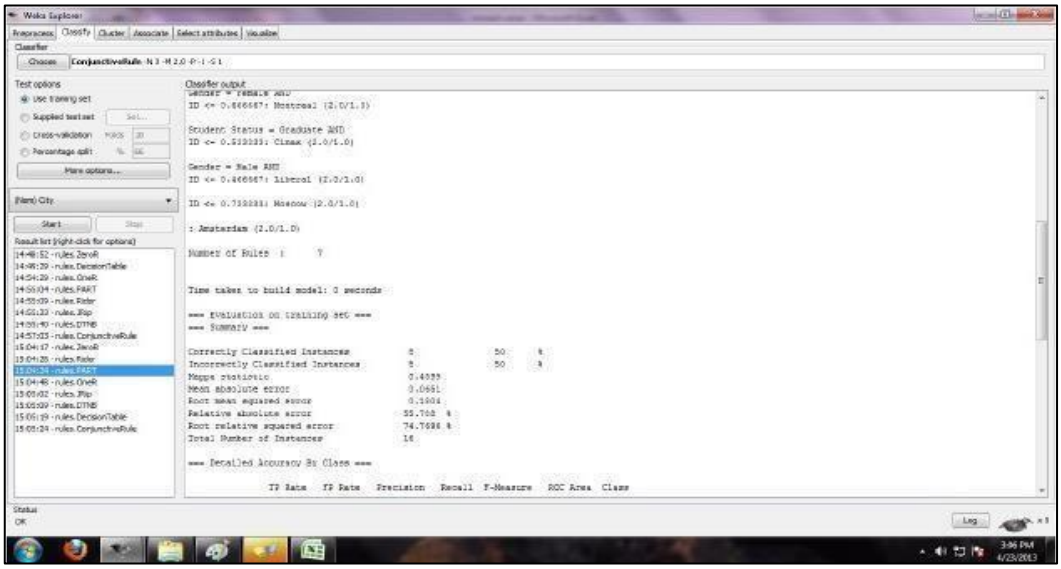
ZeroR:



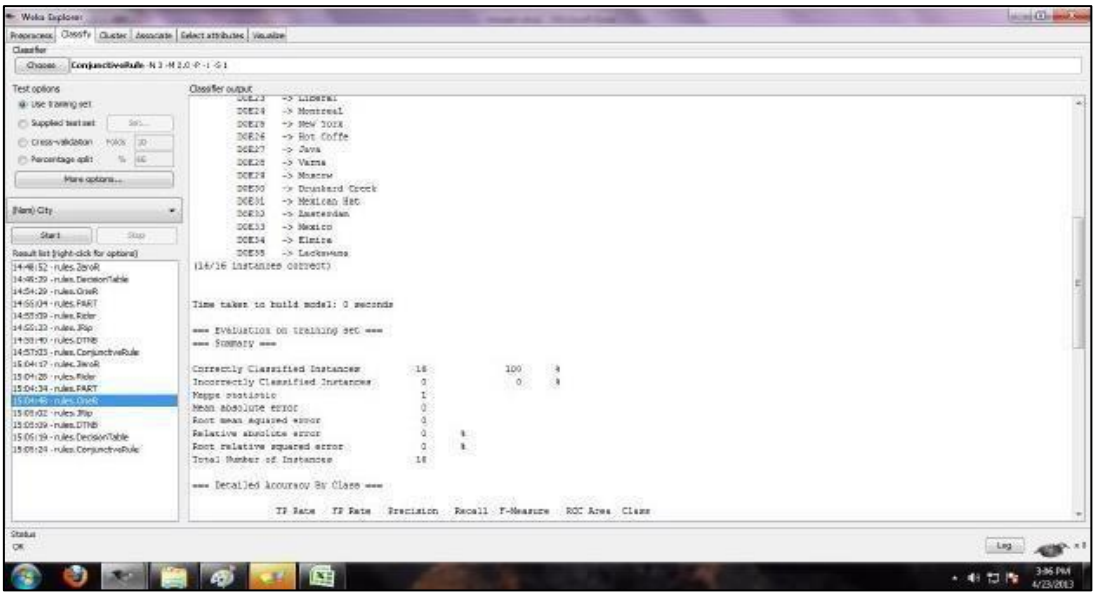
Ridor:



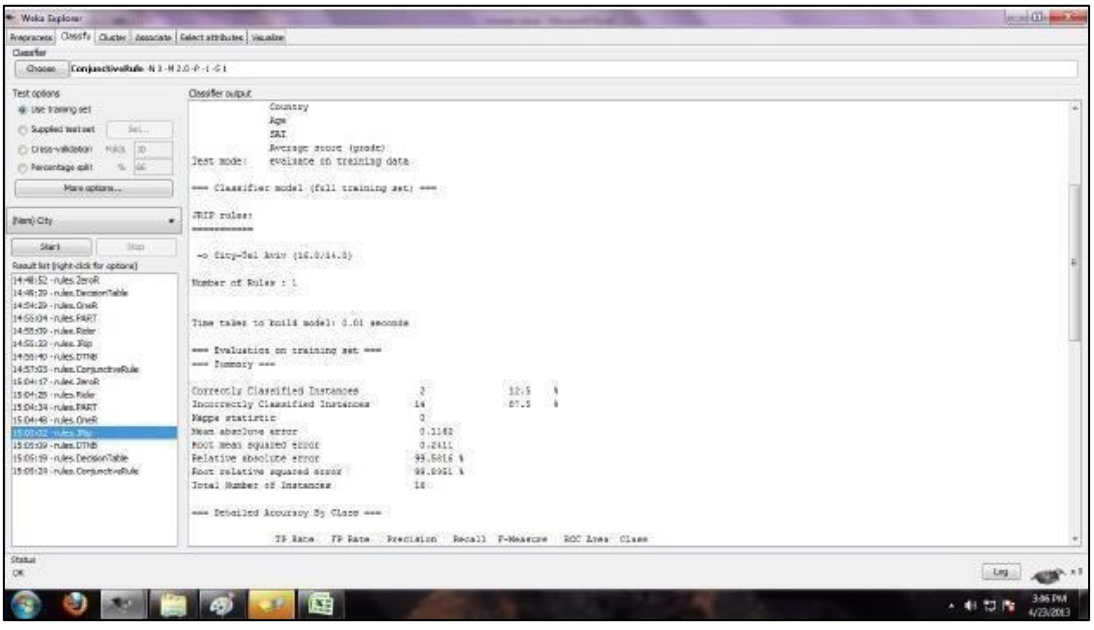
PART:



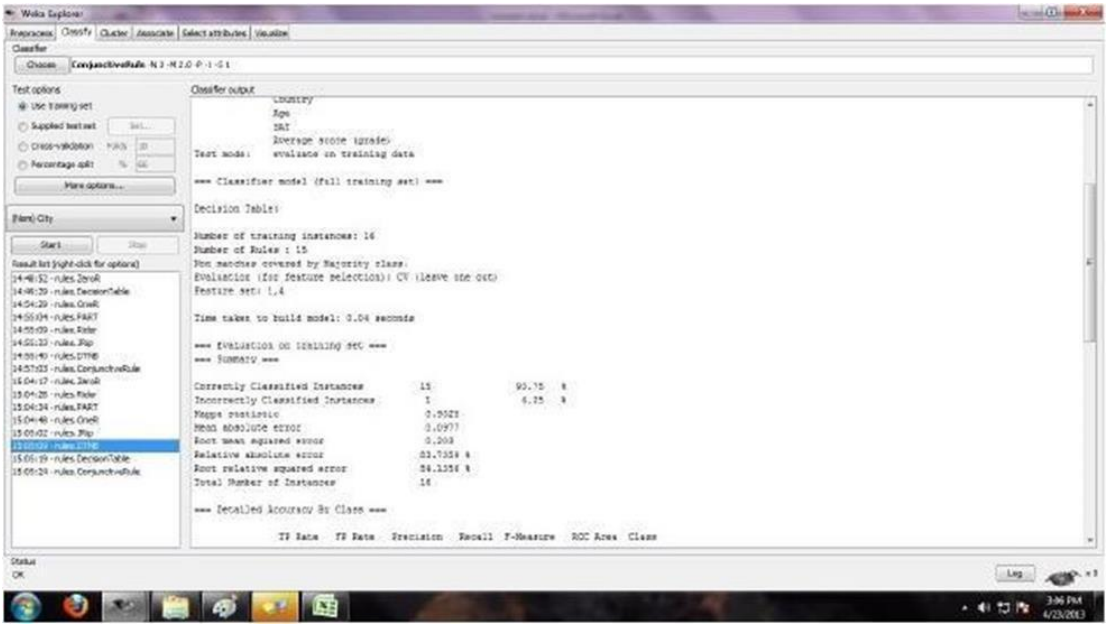
OneR:



JRip:



DTNB:



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

$$\text{Accuracy}=\sqrt{\sum (\text{old-new})^2}$$

$$\text{Percentage of accuracy}=\frac{\text{Accuracy}}{\sum \text{old value}} \times 100$$

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.

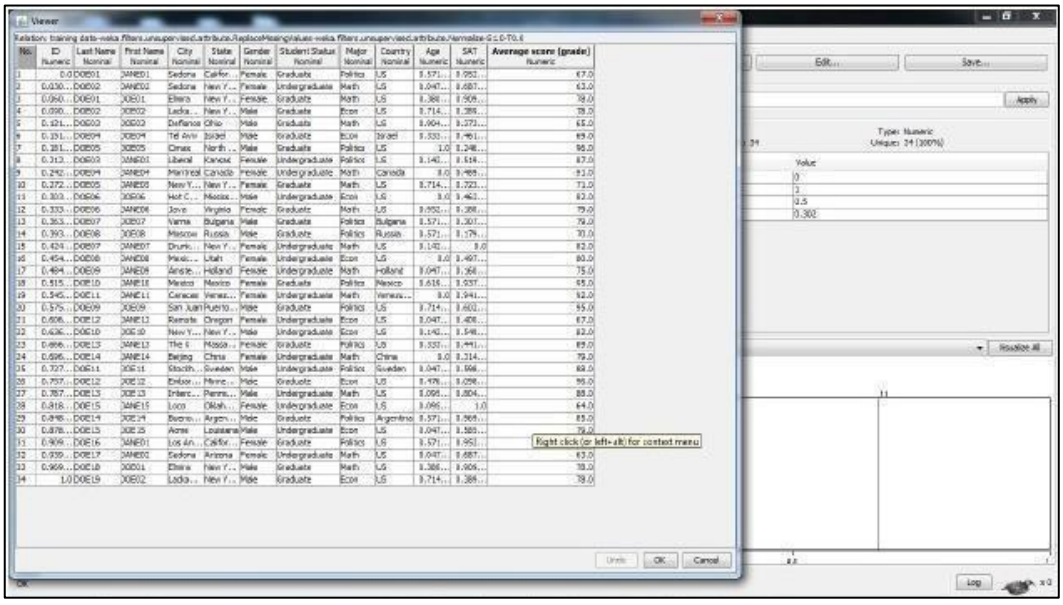
OBJECTIVES:

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

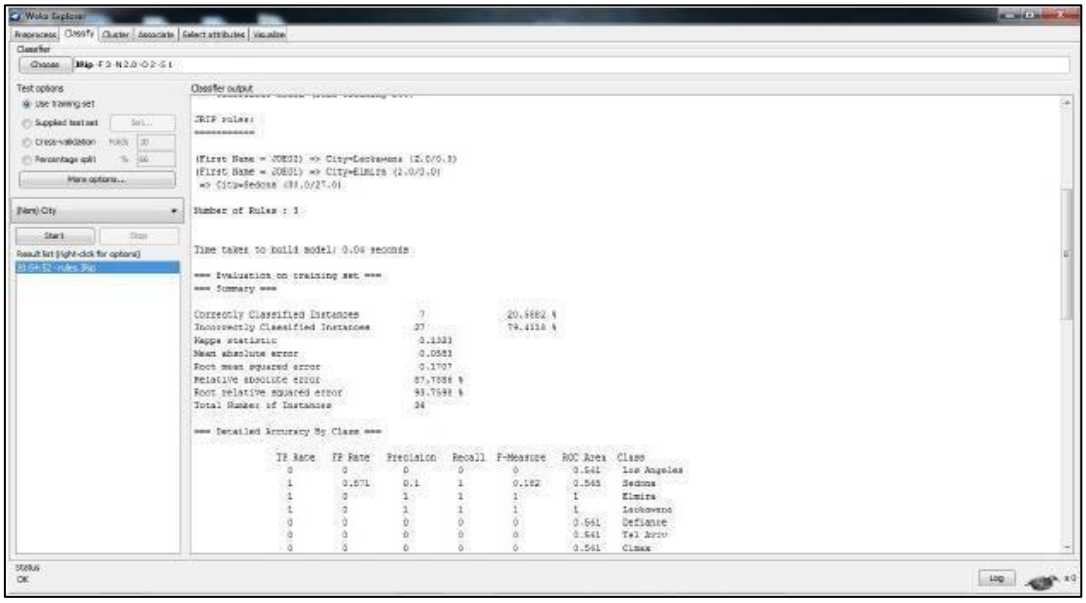
ALGORITHM:

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.

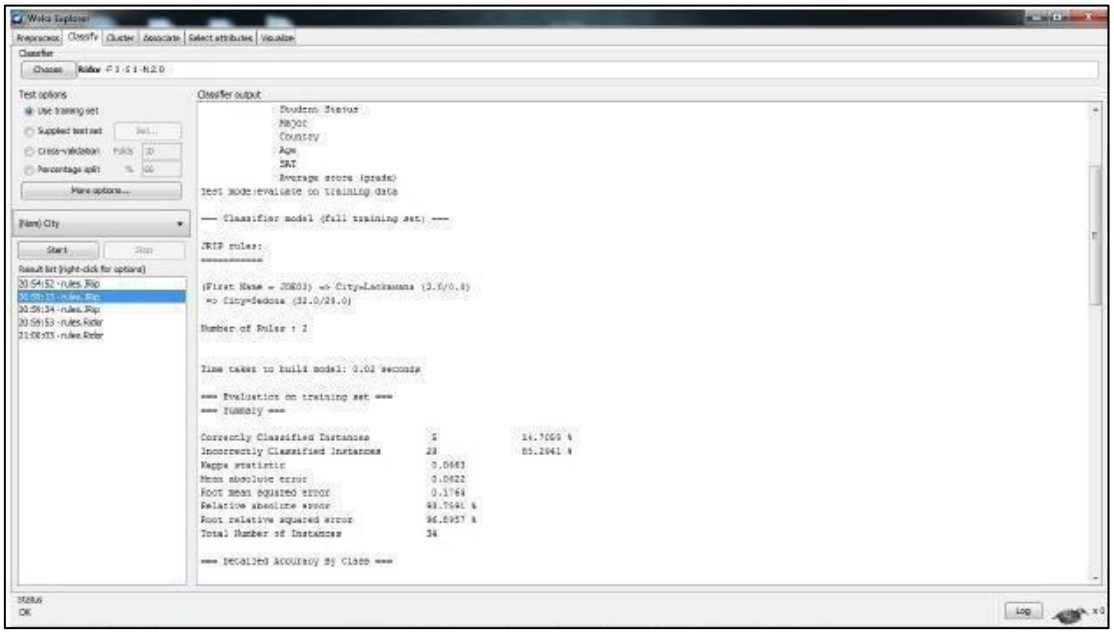
Training Data:



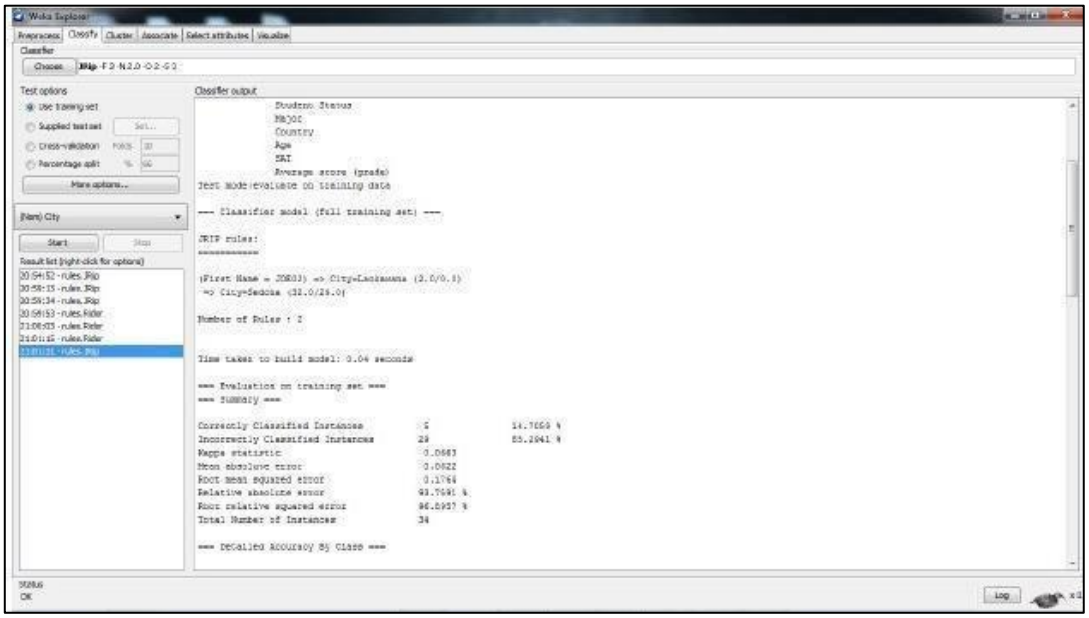
JRip(seed=1):



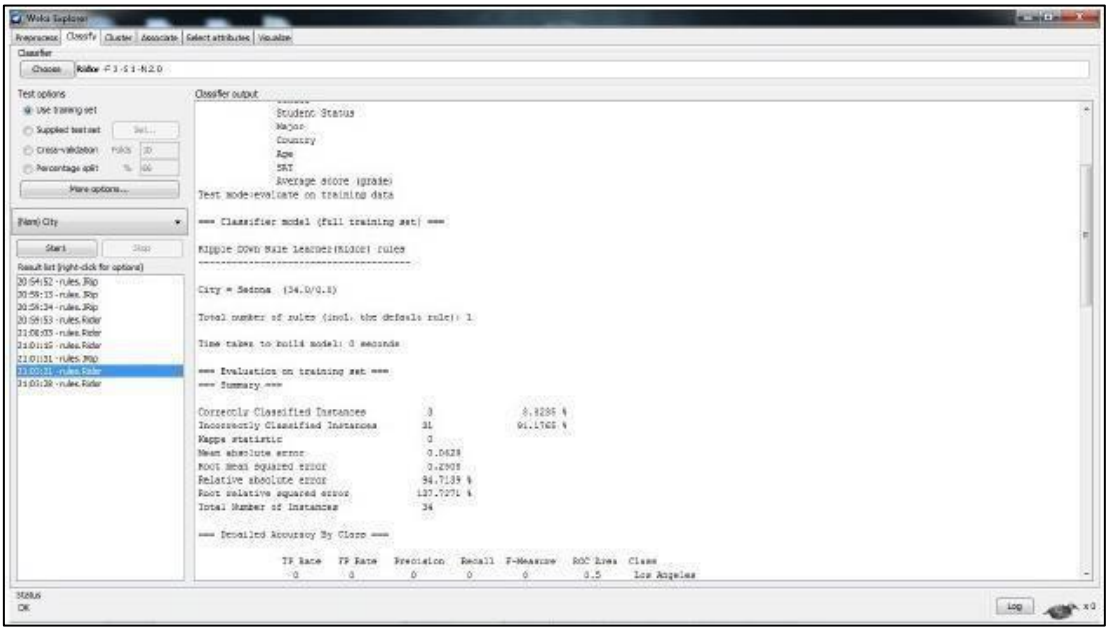
JRip(seed=2):



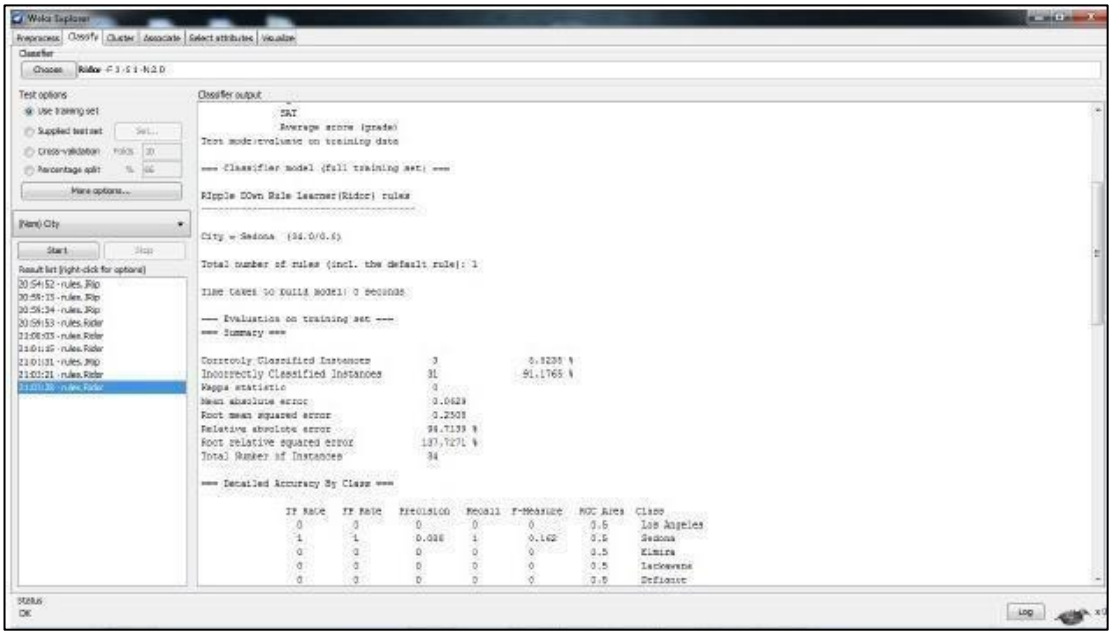
JRip(seed=3):



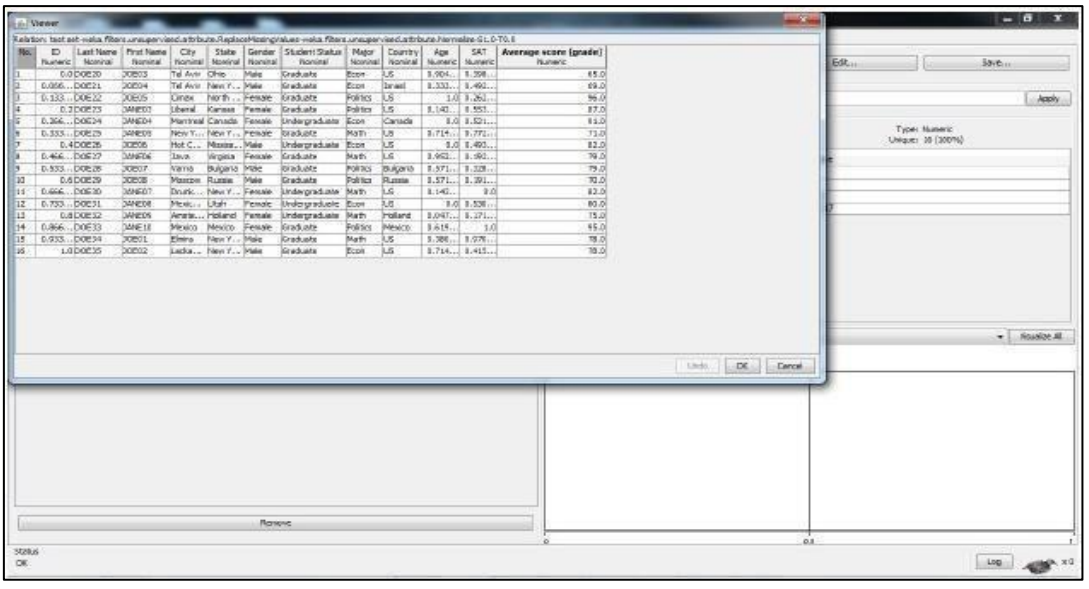
Ridor(seed=1):



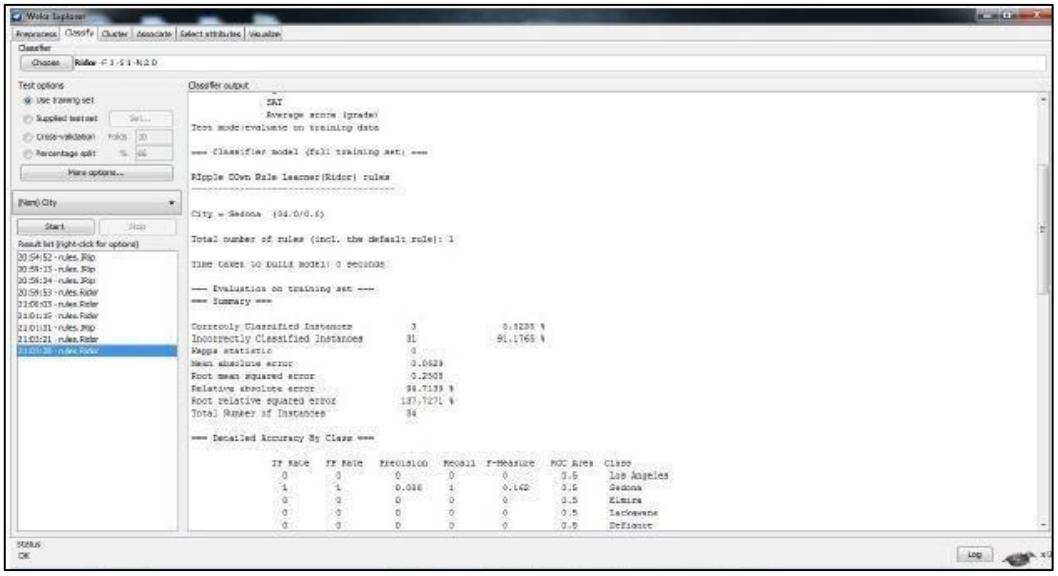
Ridor(seed=2):



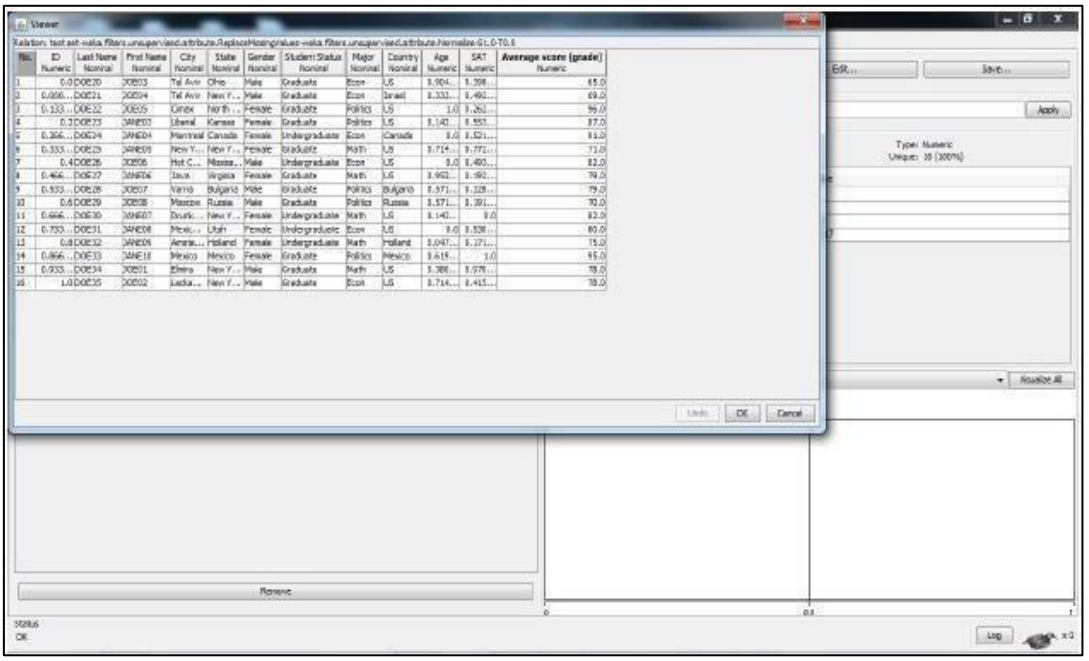
Test Data:



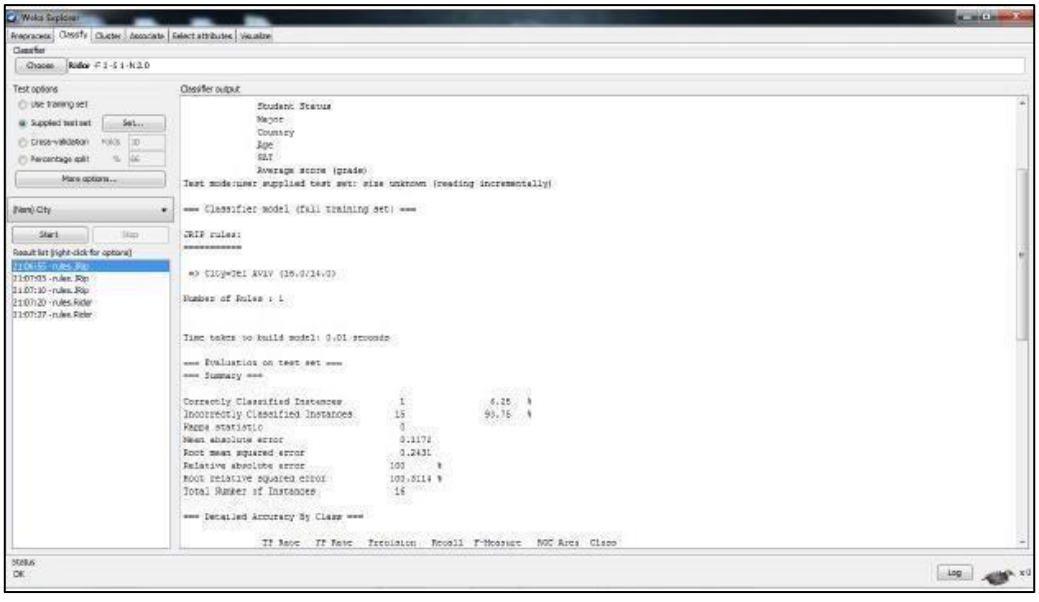
Ridor(seed=2):



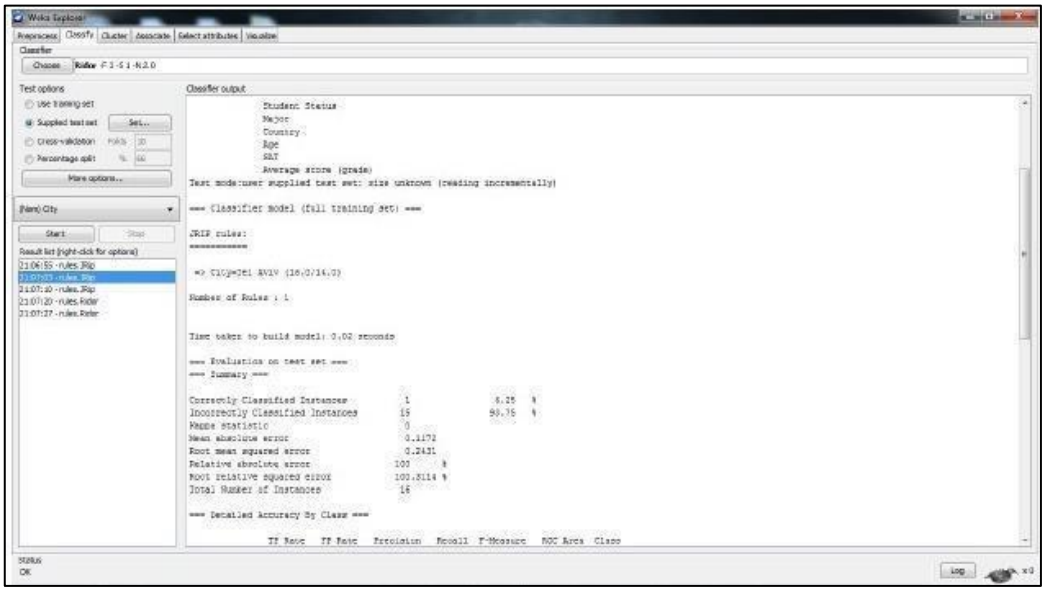
Test Data:



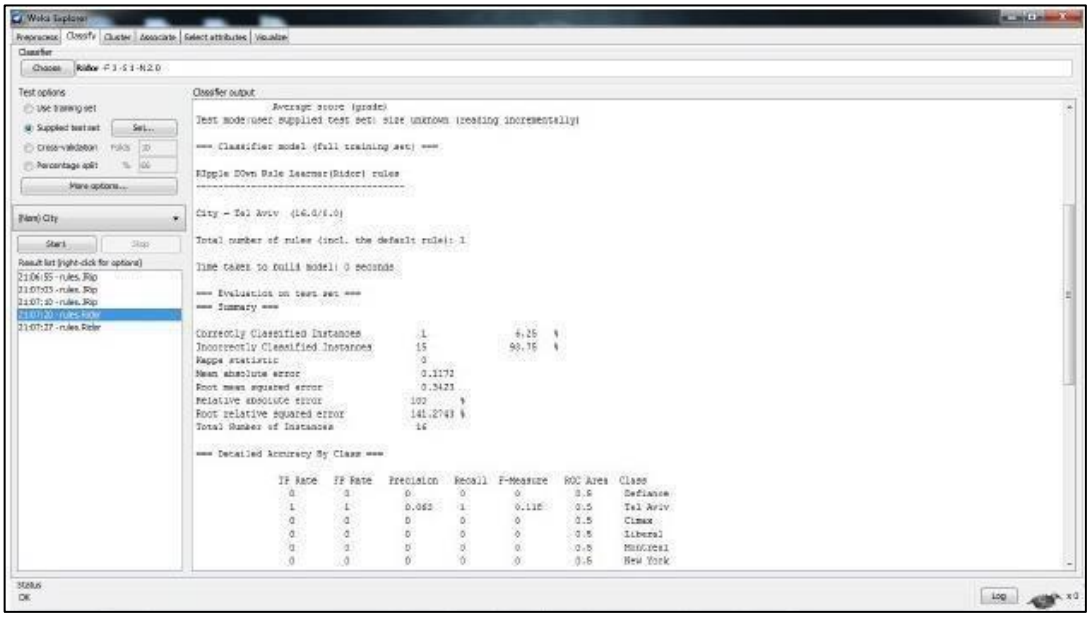
JRip(seed=1):



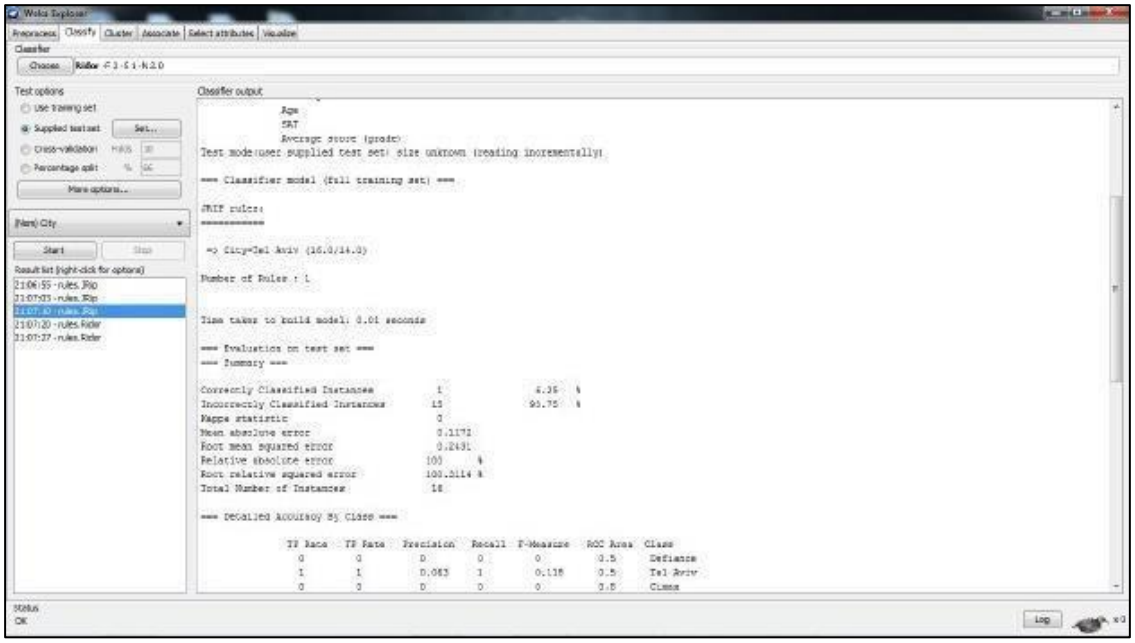
JRip(seed=2):



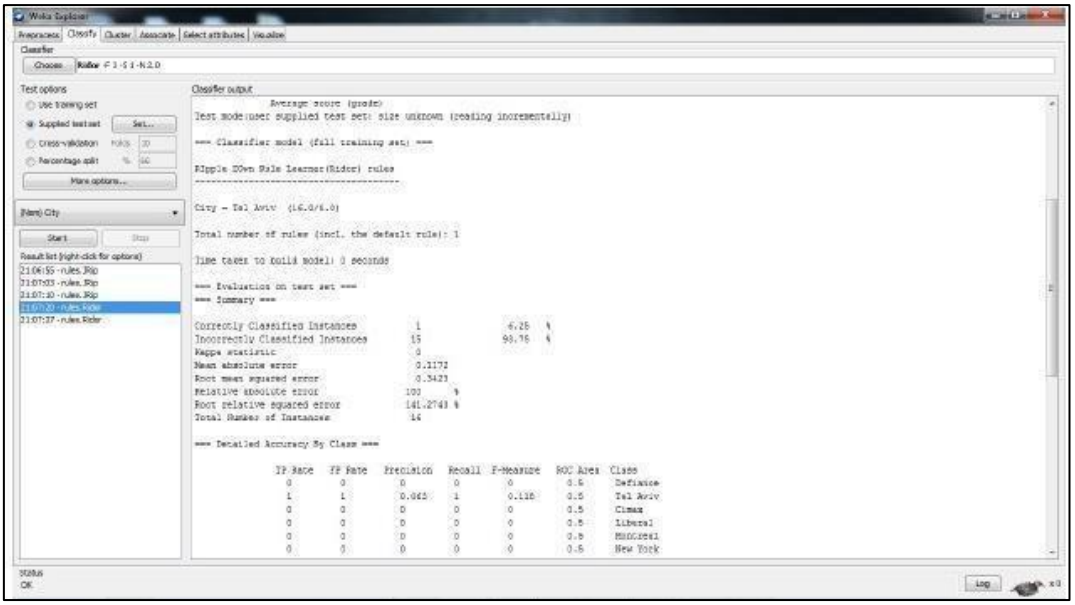
JRip(seed=3):



Ridor(seed=1):



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:

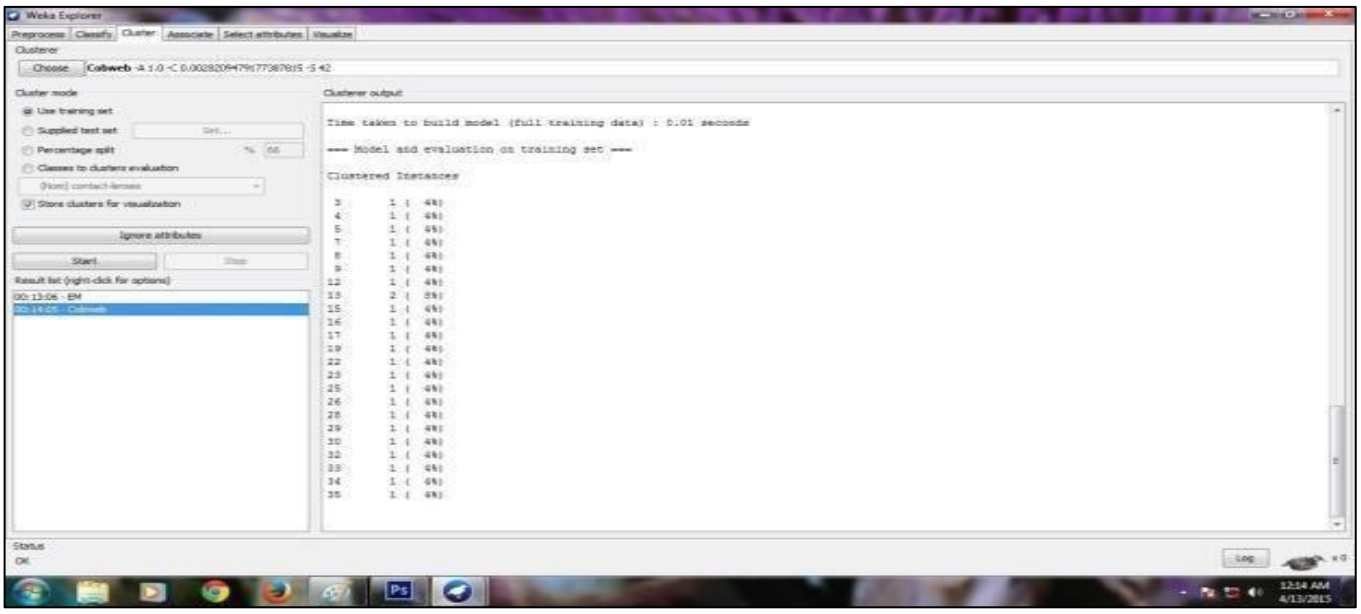
Web content mining can be used for mining of useful data, information and knowledge from web page content. Web structure mining helps to find useful knowledge or information pattern from the structure of hyperlinks. Due to heterogeneity and absence of structure in web data, automated discovery of new 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:

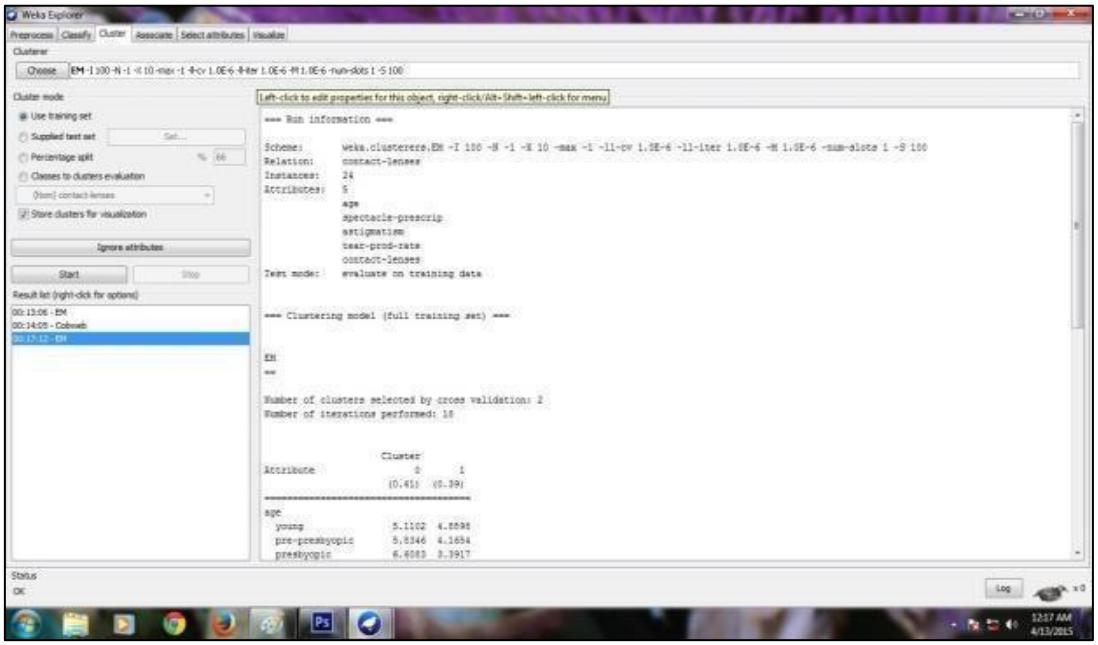
1. Open the weka tool.
2. Download a dataset by using UCI.
3. Apply replace missing values.
4. Apply normalize filter.
5. Click the cluster tab.
6. Apply all ALGORITHMs one by one.
7. Find the no of clusters that are formed
8. Note the OUTPUT.

OUTPUT:

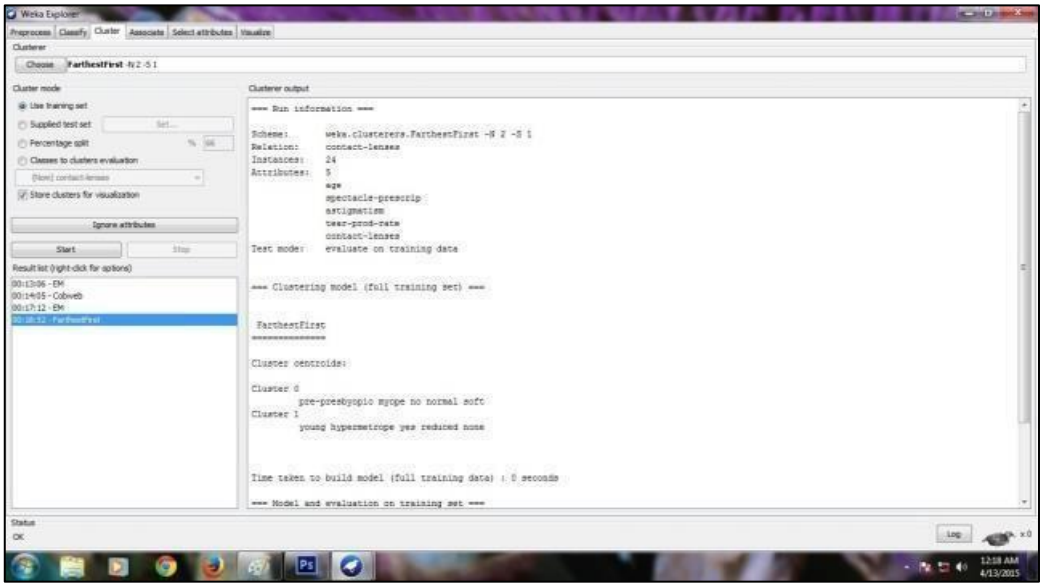
Cobweb



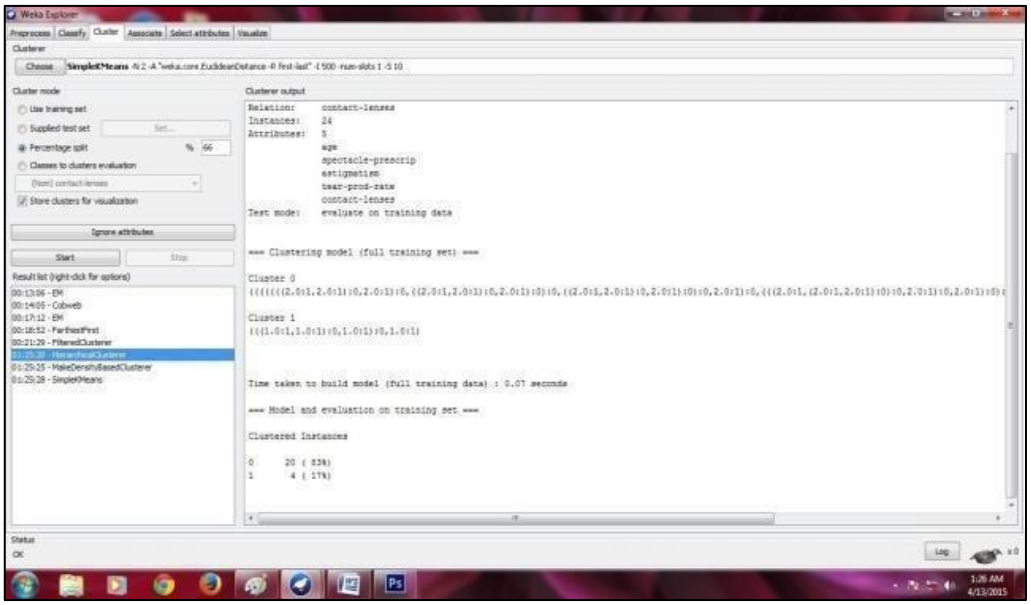
EM



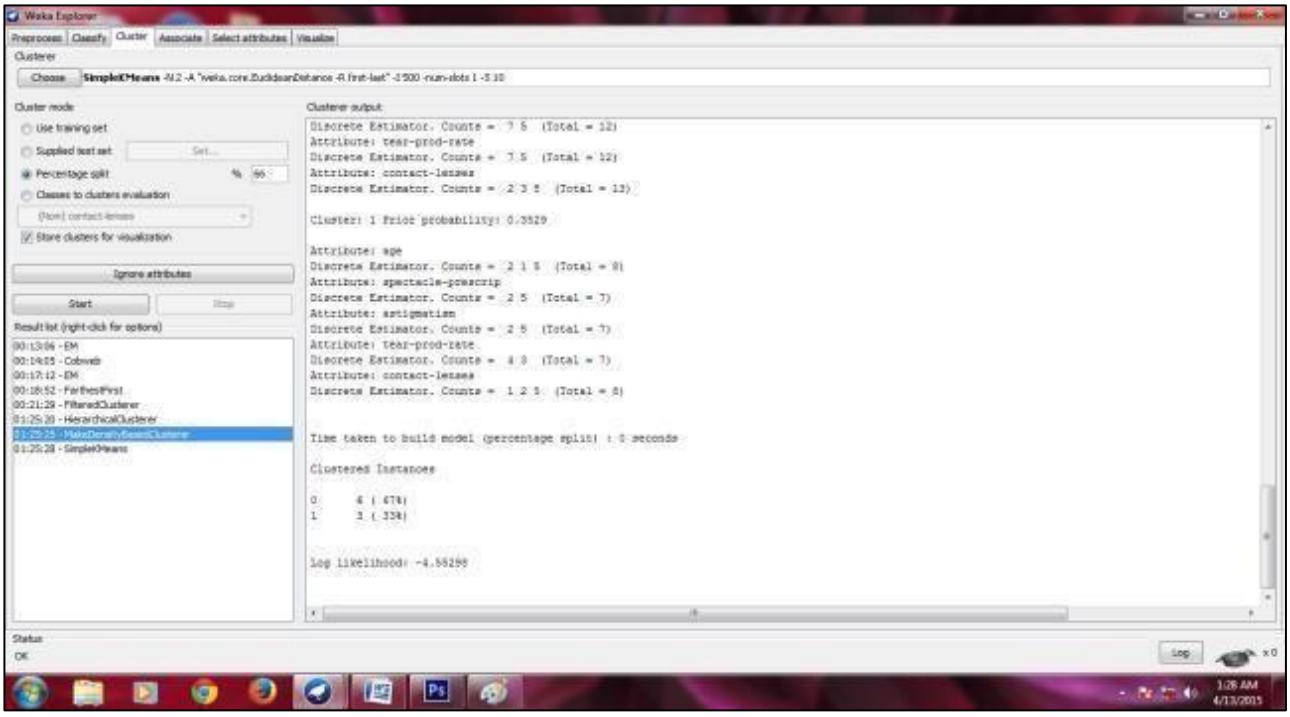
Farthest First:



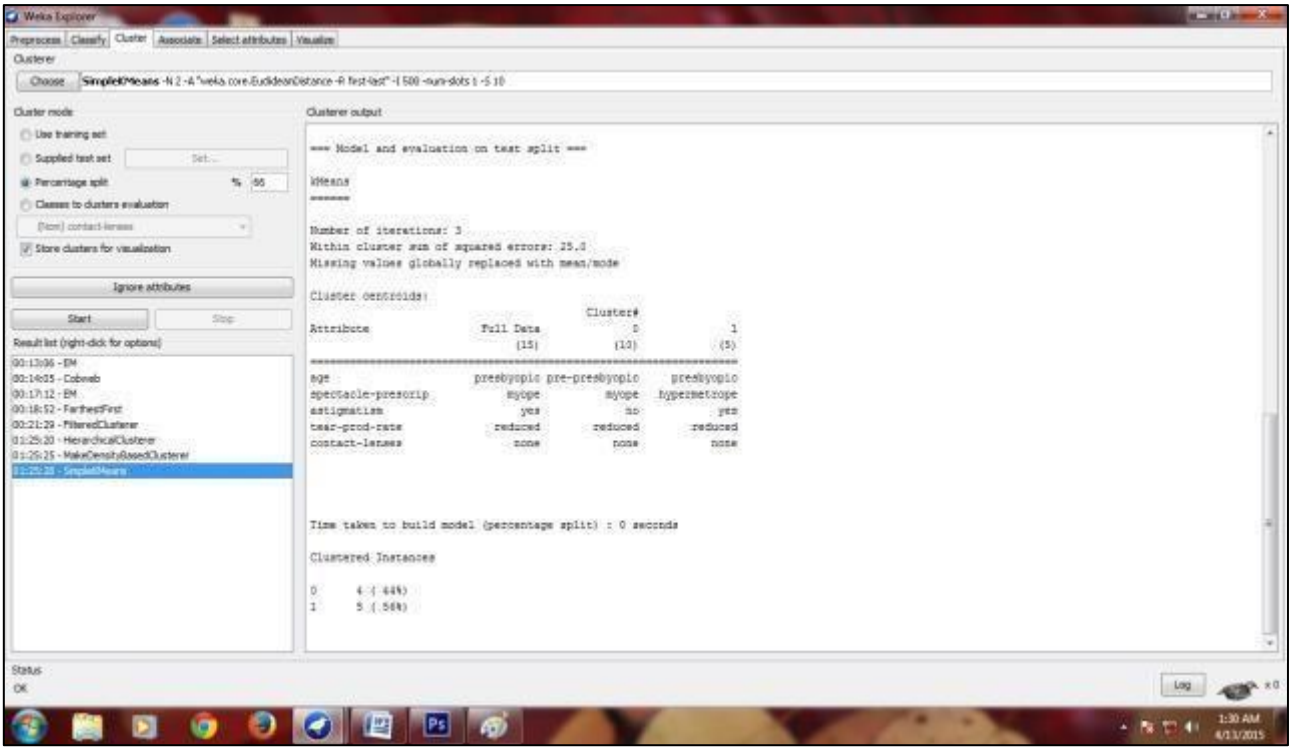
Filtered Cluster:



Hierarchical Cluster



Simple KMeans:



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:

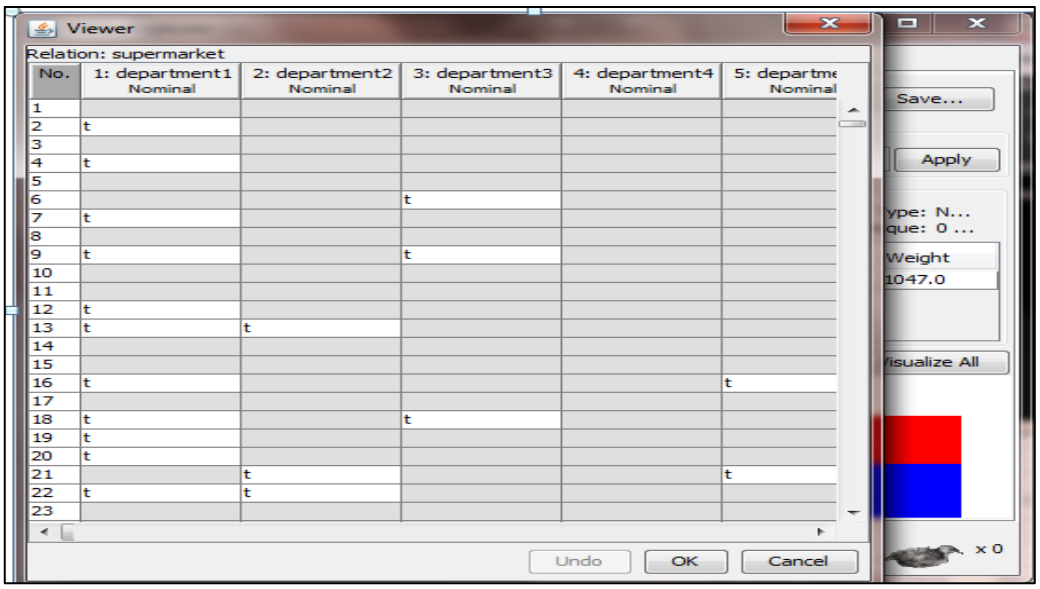
It collects sets of keywords or terms that often happen together and afterward discover the association relationship among them. First, it preprocesses the text data by parsing, stemming, removing stop words, etc. Once it pre-processed the data, then it induces association mining ALGORITHMs. Here, human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

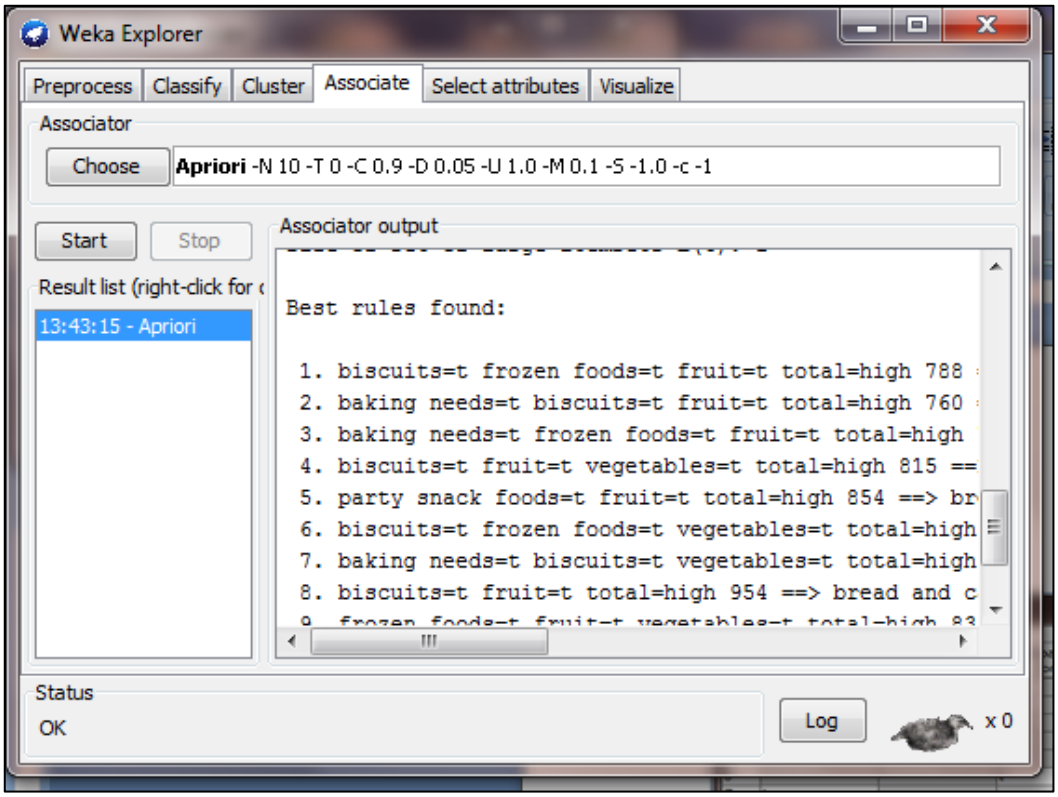
INPUT:

Super Market data set

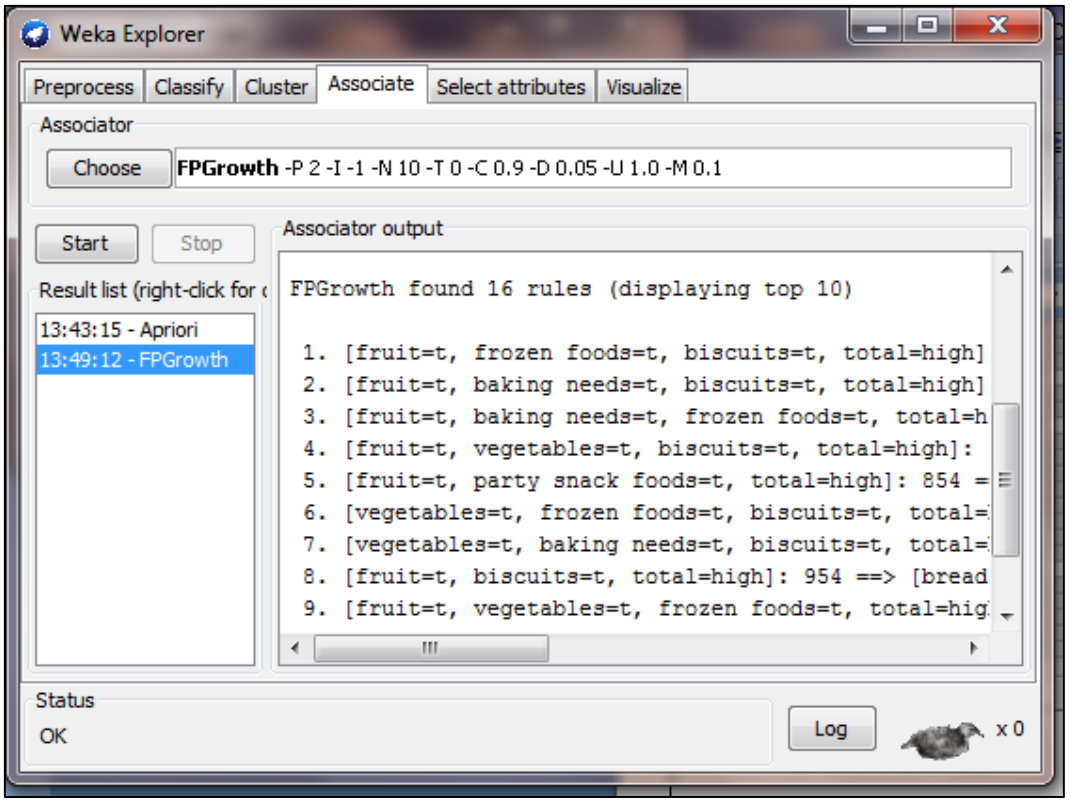


OUTPUT:

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.

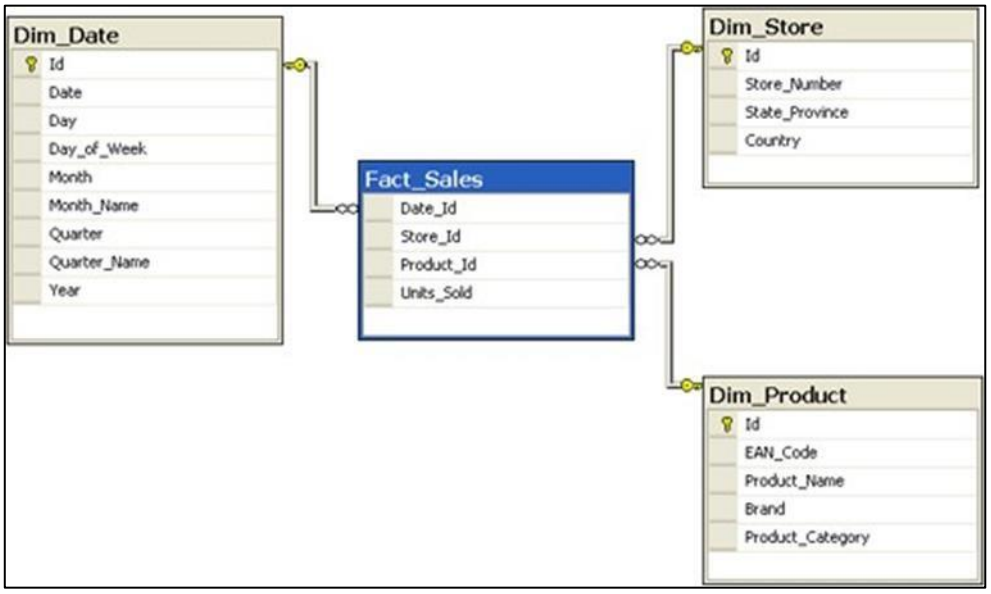
Designing fact table steps:

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.

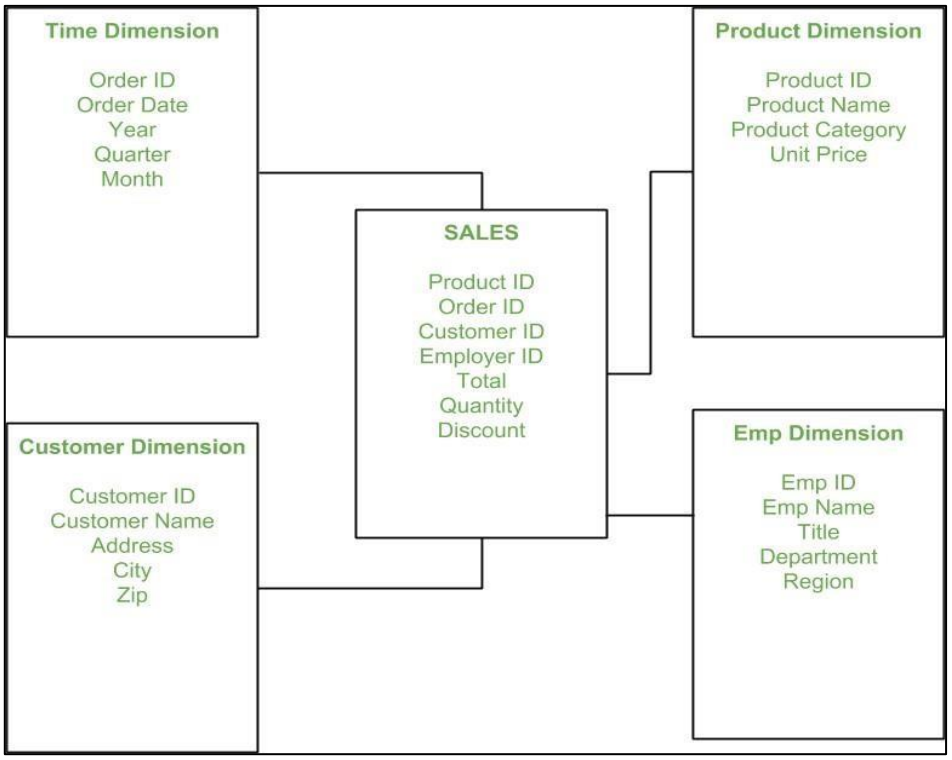
EX. No: 12
GENERATING GRAPHS FOR STAR SCHEMA

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.