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



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



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

	gineering Knowledge: Apply the knowledge of mathematics, science, engineering
PO 1	fundamentals, and an engineering specialization to the solution of complex engineering
	problems.
	oblem Analysis: Identify, formulate, review research literature, and analyse complex
PO2	engineering problems reaching substantiated conclusions using first principles of
	mathematics, natural sciences and engineering sciences.
	sign/Development of Solutions: Design solutions for complex engineering problems and
DO 3	design system components or processes that meet the specified needs with appropriate
PO 5	consideration for the public health and safety, and the cultural, societal, and
	environmental considerations.
	nduct Investigations of Complex Problems: Use research-based knowledge and research
PO 4	methods including design of experiments, analysis and interpretation of data, and
	synthesis of the information to provide valid conclusions for complex problems.
	odern Tool Usage: Create, select, and apply appropriate techniques, resources, and
PO 5	modern engineering and IT tools including prediction and modelling to complex
	engineering activities with an understanding of the limitations.
	e Engineer and Society: Apply reasoning informed by the contextual knowledge to
PO 6	assess societal, health, safety, legal and cultural issues and the consequent responsibilities
	relevant to the professional engineering practice.
	vironment and Sustainability: Understand the impact of the professional engineering
PO 7	solutions in societal and environmental contexts, and demonstrate the knowledge of, and
	need for sustainable development.
DO 8	hics: Apply ethical principles and commit to professional ethics and responsibilities and
100	norms of the engineering practice.
PO 0	dividual and Team Work: Function effectively as an individual, and as a member or
109	leader in diverse teams, and in multidisciplinary settings.
PO 10	mmunication: Communicate effectively on complex engineering activities with the
1010	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.
	oject Management and Finance: Demonstrate knowledge and understanding of the
PO 11	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
1012	in independent and lifelong learning in the broadest context of technological change.

PROGRAMME SPECIFIC OUTCOME

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

U20ITCJ04- DATA MINING

PART A- INTRODUCTION OF THE COURSE

Course Co	de	1	12017	C.104							L	Τ		Р	С
			02011	CUUT							3	0		2	4
Course Tit	le]	DATA	MIN	ING										
Course Ca	tegory]	Profes	sional	Core	(C)					Conta	ct Hrs	5	75	
Pre-requis	te						(Co- Re	quisit	e	Nil				
Name of th	e Course	e Co	ordina	ator											
Course off	eringDep	ot./Se	chool				Ι	T / So	C						
Course Ob	jective a	nd S	Summa	ary											
• ′	Го Undei	stan	d the c	concep	ts of D	ata Mi	ining. I	Familia	arize w	vith As	sociatio	on rule	minin	ıg.	
]	Familiari	ze w	ith va	rious C	Classifi	cation	algorit	hms.							
• '	Го Undeı	stan	d the c	concep	ts of C	luster	Analys	sis.							
• ′	Го Famil	iariz	e with	Outlie	er analy	ysis tec	chnique	es and	applic	ations	of Data	ı minir	ng in d	ifferen	t
	domains.														
Course Ou	tcomes (COs	3)												
CO1	Gain kn	owle	edge a	bout th	ne conc	cepts of	f Data	Mining	z .						
CO2	Apply A	Asso	ciation	n rule n	nining	techni	que.								
CO3	Use var	ious	Classi	ificatio	n algo	rithms	•								
CO4	Gain kn	owle	edge o	n the c	oncept	ts of C	luster A	Analys	is.						
CO5	Identify	y Ou	tlier a	nalysis	techni	iques.									
CO6	Underst	and	the im	portan	ce of a	pplyin	g Data	minin	g conc	epts in	differe	ent doi	nains.		
Mapping /	Alignme	nt o	f COs	with l	PO & 1	PSO		1			-			1	
										0	-	7	1	2	3
	6	5	03	04	02	00	6	80	50	6	ō	6	SC	SC	SC
CO1															
CO2															+
CO3				<u> </u>											
CO4				<u> </u>											
CO5															
CO6															
(Tick mark	or level	of c	orrela	ntion: 1	- 3-Higł	n. 2-M	edium	. 1-Lov	w)	1	1	1	1	1	

PART B- CONTENT OF THE COURSE

UNIT I DATA WAREHOUSING, BUSINESS ANALYSIS AND ON-LINEANALYTICALPROCESSING (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

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

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

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

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.

9

9

9

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:

Create file	Course (MI)	0.000	1		1026	r.44	
Upen nie	Open Ukt	Open UB		ace	Unao	tat	5000
lter							
Choose None							Appl
utrent relation				Selected attribute			
Relation: bank-data				Name: ane		T	ane: Numeric
Instances: 600	Attrib	utes: 12		Missing: 0 (0%)	Distinct: 5	ia Uni	que: 0 (0%)
thibutes				Statistic		Value	
				Minimum		18	
All	None	Invert F	attern	Maximum		67	
				Mean		42.395	
No. Name				StdDev		14.425	
2 children 8 car 9 save_act 10 current_act 11 mortgage 12 pap				Class: pep (Nom)		87	Visualde
				59	69	57	
	Remove						

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.

00		Open	
Look In:	drsarang		
_pycache anaconda AndroidSe AndroidSe Application	e 13 tudioProjects tudioProjects copy ons	 bar blobcity bower_components Calibre Library Desktop 	 Invoke options dialog Note: Some file formats offer additional options which can be customized when invoking the options dialog.
File <u>N</u> ame: Files of <u>T</u> ype:	Arff data files (*.a Arff data files (*.a Arff data files (*.a C4.5 data files (*. C4.5 data files (*. C5V data files (*. JSON Instances file JSON Instances file libsvm data files (*.	rff) rff) rff.gz) names) data) sv) es (*.json) es (*.json.gz) *.libsvm)	

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	taset name
<pre>@attribute outlook {sunny, overcast, rainy} @attribute temperature {hot, mild, cool} @attribute humidity {high, normal} @attribute windy {TRUE, FALSE} @attribute play {yes, no}</pre>	- Attributes
<pre>@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 rainy, mild, normal, FALSE, yes overcast, mild, normal, TRUE, yes</pre>	t / Class variable
overcast, hot, normal, FALSE, yes	 Data Values

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:

6	43.			-			-			-		-				1		100	0.
	-	-	-	at the part of		-												-	123
1.6	2.2		100	1	1.54	10.0		100	121			6 E	12	210	10.0				
100	-	-	the la	and the	1.0	101.000			a lare	1.64	-	-	6 . Ind	ingent in					
	-			Contraction of the local division of the loc			-		-	-	-				-				
-	-			8	-			_		-	-								_
	4		4	A		1.00					- 44						- 4	1.1	
	34	10	140	A2 Investme															
	145	1.8	14	G.L True patients															
	43	8.0	12	A.I. my antesa															
	48	84	15	8.0 Horemone															
	1	4.4	14	REMEASURE.															
	1.4	1.1	17	DAINS BRIDE															
	- 64-	2.4	18	All reserves		÷													
	1	31	15	E210vielles															
	.14	24	10.	R.L.Weinfield		_													
	4.5	11	1.1.	611514694															
	64	- 17	1.0	ADVIDUATE.															
	- 64	34	18	T2 investme															
	4.8	- 1	1.0	ALT REPORTED															
	14		10	8.2789 coltral															
	14	1.4	- MC	KENDERSON															
	51	4.4	1.5.	0.4 million 4 million 4															
	- 54	8.9	18	0.4 Househous															
	38	- 44	14	R.B. HILLING															
	10	1.0	105	E.E.Waterbeite															
	- 11-	2.0	1.5	EXTRO-ARTINA															
	34	1.8	17	T.Directory .															
	14	3.2	130	12 Invidend															
	- 84	10.	- 1	RETER OTHER															
	- 84	8.8	1.7	AN INCLUSION															
	. 64	14	10	III in adding															
6	1.00.	15.00	_		_						18.					_	_		
ŝ																pell v	Lamp.	1	

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:

Al		
	ABB THE REPORT OF THE PARTY AND ADDRESS AND ADDRESS ADDRES ADDRESS ADDRESS ADD	
	ar potential Second and a Second a	
Notes and 1		21111

Data Text File:

×3.		-				100	Sector Sector		-					
	-		Contraction of the later		1000			-			100-11			- 7
21		1000	1 0	104(11)	1.00	-	127.1	2 3	- F	1 24	210	1.12		
-	-	-	and interior	10.00	-	-	farm in	-	-	- lond	ingeneral free			
-	-		Contraction of the local division of the loc			100	Andread of the second	-	-			_		
-				_	_	_	_				-	-	_	
				1.6.10										 1
3.6	10	14	A2 Investme											
45	1.8	14	ALC: No other											
4.1	1.2	1211	ALC: UNLASTICA.											
48	84	15	8,210,000,000											
1.1	4,4	14	REMERSION.											
14	1.8	17	Address											
-64	2.4	18.1	All reventions	514	1.1									
1	3.0	13	E210+sellos											
34	- 34	1.0	R.L.Weinfield		- 19									
67	- 10	1.1.	610514694											
14	- 370	1.0	ADMINING.											
- 64	- 14	18	11 investme											
14	- 8	14	ALT HE STATE											
10	- 5		8.0100.00004											
- 25		- M.	A LONG BATTLE											
32	- 55	- 55	0.4 (0) 440 (0)											
25	- 15	10	a la constante											
-25-	- 10	100	d links output											
22			A COLUMN A											
- 22	- 10		Thisselfor											
11	1.2	15	the second second											
22	- 12	- 20	A Constant											
1.1	11	12	and successing.											
10	14	10	TTO and the											
100	-			-	_	_	_	-				-		

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:

F	ile Ho	me Insert	Page Layo	ut For	mulas D	ata Re	view	View		
Pa	ste	Calibri B Z U - Calibri Font	• 11 • • A* A* • •		■ 副・ ● 困・ ※・	% Number	A Styles	Cells	Σ + 2√ + 	
	F5	•	- ·	f.						~
-	A	B	С	D	E	F		G	н	E
1	x	Y								
2	3	30								
3	8	57								
4	9	64				-	_			10
5	13	72					1.0			
6	3	36								
7	6	43								
8	11	. 59								
9	21	90							-	
LO	1	20								
11	16	83								
12										~

OUTPUT:

2					Veka Explorer		-		×
Preprocess	Classify	Cluster	Associate	Select attributes	Visualize				
Classifier									
Choose	Simpl	eLinear	Regressio	'n					
Test option:	5		11	Classifier output					
Use tra Supplie	aining set	Se	tores (=== Run inform	ation ===				1
O Cross-	validation Itage split	Folds 96	10 66	Scheme: Relation: Instances:	weka.classifiers.funct: linear 10	lons.SimpleLinearReg	ressi	.on	
ſ	More optio	ns		Attributes:	2 X Y				
(Num) Y			~	Test mode:	evaluate on training de	ita			
Start		Stop		=== Classifie:	model (full training a	set) ===			
Result list (r 23:12:32 - f	right-click f functions.	for option SimpleLine	earRegr	Linear regres:	ion on X				1
				3.54 * X + 23	21				
				Time taken to	build model: 0 seconds				
د			2						

9			Weka Explorer		- 🗆 🗙
Preprocess Classify (Cluster Associat	e Select attributes	Visualize		
Classifier					
Choose Simple	LinearRegressi	on			
Test options		Classifier output			
Use training set Supplied test set Cross-validation Percentage split More option	Set Folds 10 % 66 5	Linear regres 3.54 * X + 23 Time taken to	sion on X .21 build model: 0 seco	nds	^
(Num) Y	~	=== Evaluatio === Summary =	n on training set ==	-	
Start Result list (right-dick fo	Stop r options)	Correlation c Mean absolute Root mean squ Relative abso Root relative Total Number	oefficient error ared error lute error squared error of Instances	0.9721 4.5238 5.111 24.4264 % 23.4449 % 10	
					~

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 " \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 " $\mathbf{P}(\mathbf{H}|\mathbf{X})$ reflects the probability that costumer \mathbf{X} will buy a computer given that we know the costumers' age and income.

INPUT:

Pa	ste	Calibri B Z U · Font	Page Layo * 11 * A A * A * Ta	ut For E E E E Alignm	mulas D	Number	Styles	Cells	Δ 2 Σ + Δ + Δ + Δ + Δ + Δ + Δ + Δ + Editing	
	F5	•	6 .	fx						7
-	A	в	С	D	E	F		G	н	E
1	×	Y								
2	3	30								
з	8	57								
4	9	64								10
5	13	72					1.1			
6	3	36				- 1 1				
7	6	43								
8	11	59								
9	21	90								
10	1	20								
11	16	83								
12										-

OUTPUT:

9					Weka Explorer 🛛 🗕 🗖	×
Preprocess	Classify	Cluster	Associat	te Select attributes	Visualize	
Classifier						
Choose	Simpl	leLinear	Regress	łon		
Test option	s			Classifier output		
 Use tra 	aining set			=== Run inform	mation ===	^
Supplie Cross- Percen	d test set validation itage split More optic	Folds 96	10 66	Scheme: Relation: Instances: Attributes:	weka.classifiers.functions.SimpleLinearRegression linear 10 2 X Y	
(Num) Y			*	Test mode:	evaluate on training data	
Start		Stop	s	=== Classifies	r model (full training set) ===	
Result list (1 23: 12:32	right-dick f	for option SimpleLine	is) earRegr	Linear regres:	sion on X	1
¢			3	Time taken to	build model: 0 seconds	2

0	Weka Explore	Weka Explorer -						
Preprocess Classify Cluster Asso	ciate Select attributes Visualize							
Classifier								
Choose SimpleLinearRegre	ssion							
Test options	Classifier output							
Use training set Supplied test set Set Cross-validation Folds Percentage split More options	Linear regression on X 3.54 * X + 23.21 Time taken to build model:	0 seconds	^					
(Num) Y	Evaluation on training	set ===						
Start Stop	=== Summary ===							
Result list (right-click for options)	Correlation coefficient Mean absolute error Root mean squared error Relative absolute error Root relative squared error Total Number of Instances	0.9721 4.5238 5.111 24.4264 % 23.4449 % 10						

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:

										part.	Long L
ation, training activity	. fitera unico	erviced up	tribute Aa	olaceHost	ngviskas visks, fl	tere unou	anvied.a	tokuta JN	ondize G	1.8-70.0	
E D Last Ner	First Netter Norminal	City	State	Gender	Student Status	Magor	Epurity incoles	Age	SAT	Average score (grade)	The second second second second second
p.dDoeo1	SMMP01	Los An	Caller	Penale	Draduate	Dalties.	1.5	1.571	1.051	87.0	
0.030D0002	DAMAGOOD	LOE ATL.	Arizona	Penale	Underpreduate	Math	15	1.047	8.607	65.0	The state is a second sec
0.060D0E01	DOED1	Elwa	Nes Y	Male	Graduate	Math	LS.	1.381	1.505	39.0	1 N T
0.090., 00802	20802	Latin	Next Y	Penale	Graduate	Ecos	1.5	3.714	1.385	78.0	
0.121D0600	20602	Defiance	Ohio	Fencale	Graduata	Ecos	LG	3.404	1.171	62.0	
0.331	30804	TH AVH	29430	Mae	toraduate:	8001	19 act	1.333	1.001.	65.0	
0.351D0805	20805	Ones	North	Male	Graduate	Palitics	1.5	1.0	1.24	0.58	Color Eck P
0.212	DAMEO 2	Liberal	Canicke.	Festale	Undergraduate	Palitics	1.6	8.140.	1.516.	87.0	2 2 10001
0.2%Z D0804	JANED4	Marinea	Cariada	reilae	undergraduate	Math	cariada	8.0	1.487	91.0	C2600
0.272 D0805	DANEDS	New Y.	New Yes	Penale	Graduate	Math	1.5	8.714	1.721	71.0	* *
0.202. DOE06	20606	Hot C	Modist	54264	Undergraduate .	Ecol	LS.	8,0	1.45	\$2.0	and the second se
0-333 D0006	DAMAGOR	Jove	Wrights	Fende	Gradualtz	Math	L6	8-252	1.281	79.0	
0.353D0807	20807	Verna	Bugets	Pale	Graduate	Palitics	dulgeria	8.571	1.307	79,0	are (grade)
0.793 D0E08	30608	Mesiclei	Rissa	Mile	Graduate	FORICI	Rusia	1.571.	1.178	70.0	
0.424. D0807	DANED7	Drunk	New You	Penale	Undergraduate	Math	US	1,142.	1.0	82.0	67,3
0.454D0D08	DANEOR	MERC	Ltan	CALIF.	Undergraduels	Ecol	19	8.0	1.497	83.0	63.3 Bi Prest Delate Formal
0.484D0E09	DANE09	Arstein	Holand	Festale	Undergraduate	M951	Holland	1.047	1.361	75.0	28.0
0.515., D0E10	SAMETE	Mexico	Nexco	Panae .	Graduate	Politica	Mesico	3.615	1.937	. \$5.0	78.0 Safe
0.54500011	DANE 11	Graces	Verez	Febale.	Undergraduate	MATI	WETHERS	.8.0	1.941	82.0	65.0
0.575	10609	San Juar	Puerto	Mae	Critolace	F001CI	1.8	1.71%	1,602	55.0	69.0
0.60600012	DANE D	Remate	custou	renaw	Undergraduate	DCD4	1.5	1.047	1.400	17.0	96.0 K L
DiamDueip	PCh 30	PROV T	Paper F	204	Cridergradues	acos	M4.	8.04	1.191	82.0	87/0 SAT Average soor
0.000.100013	Depti 10	THE S.	P90550	resse.	or aduced	PORIES .	100	\$1220m	C B C These	- 63.0	91.0 80 2203 57
COVDDUELN	DAGE TA	Deang	Crista	renas.	Cride tractere	Part	August 1	1.0.1	8.314	74.0	21.0
0.727.130611	2.16-51	Macon .	. Svujeden	projet	Turpe Blag way	PORTO	neveden)	2.041	1.199	88.0	82.0
D.757 DOCLE	DOC 10	EHDOF.	Parecip	Plac	Creaupiz	ECO4	03	8,476	8,000	70.0	79.0 00 2221 78
nave hopes	DAME +C	Loca.	OU ab	Canala	Induced ate	Error	G	3 /000	10	64.0	300 1716 78
D.AM. DOPLA	00014	Renter .	Acare.	Male	Frank usin	Paking	Buiendos	8.571	1.000	12.0	70.0 1701 65
0.878 00015	000.0	Arres	Instance	Mala	Lindarry ad usin	Free	15	1 04T	1.497	75.0	8 1700 62
0.908	DAMEDI	ice de	Caty	Female	first ats	Enking	15	8.571	1.451	67.6	80 1577 56
0.030	0444901	Carlora	Arrena	Pende	Dedarrend ante	Math	1.6	LOIT.	1.487.	67.0	(20) 13/7 Will
0.959 DOF LD	00001	Date	New York	Make	fired.ats	Math	1.5	8.566	1.905	70.0	22A) H 1942 97
1.0D0E19	00602	Lada.	New Y	Mon	Grad ate	Ros	1.6	1.714.	1.384	3.0	18 1813 51
			1.001103			All the second	57				47.0 B 2041 71 +
											101% (c) - () - (c)
										Linto Cer Cenad	
	_	_	_	_	_	_	_	_	_	Contraction (Contraction)	
No. of Concession, Name	- House	1 m	dial.	1	All Income			-		Statement Statements	

ZeroR:



· Woka Iaplova	and the second	-	survey was been allowed which	
Respracess Crestly Cluster Associate	Galact attributed visualize			
Classifier				
Choose ConjunctiveRule N 3-5	42.0 P-1-51			
Test options © toes training set © Supplied testiset © cross-valication Parentage aptit Mess options.	Obselferouput scosence scosence scores bioger country Age SAT assertings Teart mode: evaluate on treining /	data		
Plant) City	Classifier model (fill training Ripple DOwn Rile learner(Ridor) rule	880)		
Result fait (right-cloir for optional) 14-48-52 -rules ZeroR 14-48-29 -rules, DecetorTable 14-55-29 -rules, Draik 14-55-04 -rules, PART	City = los Angeles (34.0/0.0) Total number of rules (incl. the de	failt fuls): 1		
14-55-53 - nám Franc 14-55-13 - nám Fran 14-55-53 - nám Fordina Conjunctiva 14-57-53 - nám Fordinactiva Pale	Time taken to build model: 3 pecend Evaluation on training set Summary			
	Correctly Classified Instances Invocatority Clearlind Instances PARDs Statistic Neam abalities error Palative squared error Palative shoutce error Not meak thoustone error Intel Rasker of Instances	3 31 0.0623 0.2203 0.2003 0000000000	5.1233 % 51.(1765 %	
Statul	Detailed Accuracy By Class			
OK.				Log
	🗎 🖾 🐼 💌	ALC: NO	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	 41 13 11 229 Pet #23/2013

PART:

Welca Explorer			manage in such as the		and the second
Prepracess 0/851/ Chuster Associate	- Select attributed Viousibe				
Casafer					
Groces ConjunctiveRule N 3 -	M 2.0 -P -1 -S 1				
Test options	Classifier output				
Use training set	Last same = DUEDOS SEGETE (2.0/1.0)				*
O Suppled text set	Last Make = DOED9: Amsterdam (2:3/1	-02			
Cress-validation Hoks 20	Taur Dana - Defidi Day Sore (D. 641)				
🗇 Percentage split 🐘 💷	The same - control and the project	1			
Mare optiona	Last Name = DOEL1: Cataras (2.0/1.0				
	Last Here = D0E13) The 1 (2.0/1.0)				
Plant) City					
Sket Stat	Last Mane = DOb14: Beijing (2.0/1.0	1			10
Result list (right-click for options)	: Benote (5,0/2/0)				
14HB152 - rules ZeroR	The second s				
14:45:29 - rules Decelor/Table	Nomber of Roles v 16				21
14 SERA HURS PART					
14:50:09 - rules. Rider	Time takes to build model: 0.00 sec	unda			
14:55133 - rules: Jisp					
14:57:03 - rules, ContunctiveRule	Evaluation on training set				
	a second second second				
	Correctly Classified Instances	1.0	55,8224 9		
	Incorrectly Classified Instances	1.3	44,1276 9		
	Regpa statistic	0.5093			
	Peak aperinte error	G. DALL			
	Falative absolute error	45,5477 8			
	Fort relative strared error	62.4223.4			
	lotal Number of Instances	34			
	The state of the s				
	recalled accuracy of class				
Statui	and the				
OK .					🔊
				1.1	41 87 IS 258 PM
					4/28/2013

OneR:

Weiks Explorer				
nepraceus 0%51/ Chiese Lassociate	Select attributes Vie alize			
Gioces ConjunctiveRule -N 3 -	N 2.0 P-1-S1			
Rein colors # Use Taxeny set 5 Supple faith terms Check of the set Ferries of the	Consider output JULIUM> Marrieronm JULIUM -> Marrieronm JULIUM -> Dentition JULIUM -> Dentition -> Dentition -> Dentition Homoscole -> Dentition Homosc	23 1. 0.9653 0.003 0.005 3.0053 3.005	97. 155 1 2.1422 1	
itatui	1.0			
*				

JRip:

Wika Explorer		Name of Concession, Name	
Respracesz (39917) (Chuster Associate Select attributes Visualize			
Casefor			
Choose ConjunctiveRule N 3 -H 2.0 P -1 -5 1			
Test options Dessifier output			
use having set			
C Suppled testant Sel 351			
Cress-valuation toks m Average store (grade)			
Test mode: evaluate on training	data		
Contracting and the second sec			
Mana options			
(RIF sales)			
(ven) City			
Start Stort Store a JERITA an Clause article	ana 17 0/0 10		
Result list Sight-cick for optional wy City-Los Armelas (32,0/20,0)	1038 (ALV(X-1)		
14-rRIS2-rules.3evoR			
14:48:29 -rules.DecisionTable Sumber of Rules / 2			
14-54-29 - rules, OneR 14-55-08 - o lar SePT			
14-51-59 - Tube Roby Time takes to build model: 0.01 and	and a		
and a second			
14 Stido -rules DTMB Evaluation on training set			
14:3/323 - numer Consumer version			
Former by Classified Instances	240	14 7050 8	
Incorrectly Classified Instances	23	50.2041 N	
Happe statistic	0,0663		
Mean absolute sourc	0.0623		
Root mean squared error	0.1764		
Falative absolute error	W3.7591 8		
Total Sumer of Instances	34		
Detailed Accuracy By Class			
Status			the second s
ok.			Leg 🖉 ×
	1000	the second se	2-69 PM
			* 🕫 🖬 🗤 4/3/2013

DTNB:

repracess Cosony Lauder Associate la	fect attributes Viounize			
Clearfor	1219/96/20			
Chapse ConjunctiveRule -H 2 -H 2.0	P-1-51			
Test options	Obsoffer output Sacyor Country Apt SAT Average score (grade)			
Percentage split % [66	Test mode: evaluate on training d	ata		
N.C	Cleasifier model (fall training	atb)		
(Nem) City +	Decision Table)			
Start Star	Number of training instances: 34			
Value at proportion of a contral) (H49:52 - Utel Star Velles 2009 (H49:52 - Utel Star Velles 2009 (H49:52 - Utel Star Velles (H49:52 - Ute	James et alle 7 42 by Bajority slad Boalassia (for festure election) Festure etc. 1,2,4 The takes to build model: 0.15 sec fundation training set - fundation training set	a. CV (Leave one out nite	5 	
	correctly classified Tartaces Theoremoty (classified Instances Maps statistic Nama dealares error Root mean squared error Root relative squared error Poot relative squared error Total Numer of Instances Decalled Accuracy By Class	30 8 9.6771 0.0049 0.1049 03.0493 03.0493 03.0493 8 90.2521 8 94	80.1200 * 111242 A	
Status				

TEST DATA:

Cont. a la cont		
praces Costry Cauter Assocate	extationes i locate	
Chases ConjunctiveRule N 3 -H 2	04-141	
st oplans & Use Yorking set Supplied teatment Set	Ossifer outurt Charge Charge Strate Bendez Anne	
Percentage split % (0) Mare options	accuarto actual Major Country Age SAT	
en) Oty •	Zowrage score (grade) Test mode: evaluate on training data	
and hit fydyd cad tro golaena) (#182 - 1462 2004) 39739 - nies Decremy Telde 59739 - nies Decremy Telde 59739 - nies Decremy Telde 59739 - nies Telde 57737 - ni	<pre> Classifier model (full training set) ZeroR predicts close values Tel kviv Time takes to build model: 0 seconds</pre>	
that .	T) late 19 Date Frecision Recall F-Measure ACC Area Clear	
() () () () () () () () () ()		Ligiji 🛷

ZeroR:

• Woka Esplorer	And the second se	(v v) (1) - me 3v
Prepracess Classify Chuster desact	ate Select attitutes visualize	
Classifier		
Chases ConjunctiveRale N	140.0-1-61	
Test options	Operator support Sector Sect	
Mare options	Viassifier model (Full unuming stb)	
(Nem) City	* Ripple Down Rule Issumet (Riday) onles	
start Start		
Result list (right-click for options)	City = Tel Aviv (16.0/0.0)	1
14-98:32 - rules, 20104 14-98:32 - rules, DistantonTable 14-59:29 - rules, Distanton 14-59:04 - rules, PART 14-55:09 - rules, RART 14-55:370 - rules, Richt 14-55:370 - rules, Richt	Total number of sules finds the default rule: 1 Time takes to build model: 3 seconds	
14-511-10 - rules, DTNB 14-57503 - rules, ConjunctiveRule 15-04:17 - rules, ZeroR	we related to realize the second real real and real real real real real real real real	
13:01:33 - rules River 13:02:33 - rules RMRT 13:02:148 - rules OneR 13:00:02 - rules Migo 13:00:02 - rules Migo 13:00:02 - rules DecisionTable 13:00:19 - rules DecisionTable 13:00:29 - rules ConjunctiveRule	Information Constrained and a constrained of the constraints of the co	
Status OK	lecalled Accurany By Close IF Rate IF Fate Frecision Facel F-Measure ROT Aces Clear	
🛞 🔮 💌 🛛		• 1 1 1 1 1 4/23/2013

Ridor:

	tact act-woka, film	riunspervi	ec.atvia	alqui R. artu	ceMissing/	taluan .	S. 1		1. I	N				
6	D Last Nerre Noniral	First Name Nominal	City	State Nominal	Gender Norsital	Student Status Noniral	Major Nominal	Country Nonine	Age Numeric	SAT Numeric	Average score (grade) Runaric		5dk	Save
	35.0 D0E20	20803	Tel Avrie	Ohio	Male	Graduate	Econ	LS .	37.0	1701.0	45.0			
	36.0 00021	00004	Tel Avir	Nes Y.,	Make	Graduata.	Ecos	lated	25.0	1705.0	69.0			11.2010
	37/0 D0E22	20605	Onex	N011	Mate	Graduate	Folia	US .	38.0	1577.0	95.0			
1	38.0 D0823	DANED3	Liberal	Kanasa	Mela	Undergrad.ate	Politica	1.5	21.0	3842.0	87.0			
<u> </u>	20.0 D0624	DANED+	Mantreal	Canada	Famale	Undergraduate	Mats	Carada	18.0	2013.0	11.0			uper Nameric
-	40/00/0819	OWNED!	NOV You	NEW Y	renae	2/61/06/07	Math	10	35-0	2041.0	71.0		Us	ique: 16 (300%)
-	41.0000232	pueup	Plot C	POLICE.	, Pier	Undergraduate	ECD4	10	13.0	1780.00	12.0			
+	42.000827	June 1	Line .	in course	PROLED	bras.ars	BAD:	ALM.		1813.0	14.0		Volue	
+	10/00/00/00	2000	Maria	Bugato	Mala	Canturate	Pulling .	Duyana	28/0	30.57.0	17/0		35	
	45.000630	DANEO7	Dourk	Rate V	Canal	Sindarmad rate	Math.	LC.	24.0	1222	10.0		50	
+	-managerst	TANKING.	Mente	1.Bals	Frank	Linder on ad unio	Free	1.0	18.0	3571.0	80.0		42.5	
+	47.0 DOF 52	MILEOS	Verale.	History	Penale	Lindergrad als	Math	ripland	18.0	1494.0	15.0		4,761	
ŝ.	46/10/0E33	DAVETE	Menico	Mesizo	Fessie	Fratate	Folitics	Newco	31.0	2248.0	45.0			
	49.0 DOE 34	30801	Emira	New Y.	Male	Craduate	Math	1.5	28.0	2221.0	78.0			
÷	50.0 000735	30812	Lacka	New Y	Mala	Graduate	Ecol	1.5	31.0	1715.0	10.0			
													20	
												Unit OK Cent	- 10	
		_	-	-	_	_	_	_	_	-				
					Ren	pine .	_	_		_				
														Lig all

PART:

Waka Esplorer	and an	an and the second se
vepraceux Classify Chicter Associate	Gelect attributes Visualize	
Classefler		
Choces ConjunctiveRule N 1	(10+-1-51	
Test options © Use it anwing set © Supplied taut aut	Cheodier ougut Wenner * Temais and ID (= 0.0666675 Hentenal (2.0/1.3) Grudare Granus - Granusce MM	
C Drets-vieldeton Hors 20 C Percentage cellt % 20 Mare options	10 <- 0.400001: Chama (0.01.0) Gender = Male AUU 10 <- 0.400001 likeral (0.01.0)	
Nen) Oly Sket 31as	ID << 0.7120231 Hosenow (2.0/1.0)	7
Result last (right-cick for options) 14-46:52 - rules Zervik 14-36:520 - rules Decelor/Table 14-54:20 - rules OneR 14-551-20 - rules OneR	Planter of Fules : 7	T
14:53:539 - rules, Rider 14:55:23 - rules, Jop 14:51:40 - rules, D1148 14:51:535 - rules, ConjunctiveRule	EVELUSCIOL DE L'EXILIPO 365 Souvaiv	
15 04:17 - rules, Jacob 13 04:28 - rules, Rider 13 03:24 - rules, Rider 15 04:48 - rules, OneR	Incorrectly Classified Instances 6 50 6 Incorrectly Classified Instances 5 50 8 Mages Patientic C 14335	
13 01:02 - rules, Pilp 13:05:00 - rules, Driff 15:06:19 - rules, DecisionTable 15:06:24 - rules, ConjunctiveRule	Peter mode appared server 0.1001 Relative Auxilians server 0.1000 Relative Auxilians server 25.100 % Rotor melative seaves 74.1008 % Total Number of Instances 16	
Steka	Detailed Accuracy Br Class TF Bate FF Bate Frecision Recall F-Measure ROC Area Class	
ok		1. Log
		3-56 PM

OneR:

• Weka Explorer			-	-	Street Long Tree		the Carl and Some
Prepracess (39957) Cluster Associate	elect attributes Visualse						
Classifier							
Chases ConjunctiveRule N 3 -H 2	.0. ₽ - i ≤ 1						
Test spilons W use t parmy set Status D says and the spilon test and D says added in	Conster output DELI -> Initial -> Mentreal DELI -> Mentreal DELI -> Net Youx DELI -> Net Youx DELI -> Net Youx DELI -> Net Yours DELI -> Net Mentre DELI -> Net Mentre DELI -> Net Mentre DELI -> Net Net An Her DELI -> Net Annet Mentre DELI -						-
Star1 Stap	DOEDA -> Einite						
Result list (right-click for options)	DGE35 -> Lackmenna						
144413. 1408-3044 1445-23 - Hale Consortialis 1445-29 - Hale Consortialis 1445-29 - Hale Consortialis 1455-29 - Hale Consortialis 1455-20 - Hale Consortialis 1455-20 - Hale Consortialis 1555-21 - Hale Consortialis 1557-23 - Hale Consortialis 1557-24 - Hale Consortialis	Time taker to build model: O second Evaluation of training set Evaluation of training set Evaluation of training 	IS IS I I I I I I I I I I I I I	t 1	100 0 T-Measur	a a a a a a a a a a a a a a a a a a a		Ę
Status							(
OK			-				<u>lig</u> 40° *1
🛞 🔮 💌 😭	Ø 🖬 🖽	14		1		and the second s	• 41 13 19 336 PM

JRip:

• Weka Deplorer		-	Name and Add Owner, Name	and the set of
Prepracess CASSIFI Chucker Associate Select attributes Visualize				
Disarter				
Chases ConjunctiveRule № 1 - № 2.0 -P -1 -5 1				
Test options Classifier output				
be towny or Country Age Sai Descentage and to Provide on the Descentage and to Descentage and to Descentage and to Descentage and to	0 data ug ast]			•
Perc) City • RTP rules				
Start Start - o firs-lei kriv richtigt				
Smarkin [gdr4:dd fr geron] Hamilia (hdr2: code Hamilia (hdr2: code Ha	2 14 0 0.1142 0.2411 0.2411 0.2411 0.2616 1 0.2616 1 14 14 14 14	12:5 61:3	 BOT June Class 	
Status	Contraction of the state			
UR				<u> </u>
🛞 🥹 📨 🔛 🧭 💶 🖼	1000	1	and the second se	* 48 TO 15 346 PM

DTNB:

Responses Ottoff Outse Associate Salect Caracter Channe ConjunctiveRule N.3. N.2.0 P Test options Otto W Use Tarming set	20./me Veuellan < 1		
Causer ConjunctiveRule N.1.H.2.0 P Text options Case # Use Towny set	G1 Mr compa		
Channe ConjunctiveRule N 3 H 2.0 P est options One @ Use Towng set	6 t ler outsat		
Test oplions One w use training set	ler natal		
C. Sappled better 1+1	Josephilic Linkson (1997) Josephilic Linkson (1997) Rockson	;	
State) Hanker of Environmen 18 Noralled Acourson &: Class ==== TH Bate fH Bate Frecision Recell f-Heart	re RCC Ares Class	
OK .			102 100.13
		the second se	AND DESCRIPTION OF

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

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

Percentage of accuracy= Accuracy $\times 100$

 \sum old value

OUTPUT:

Student Details Table: Missing values

Relat	tion: weather				
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		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:

Un	d: outlook C	tapparatura	2: humiditu	A windu	Enlow
NO.	Nominal	Numeric	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	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	Old Data	Predicted data	Errors	(Error)2
Location				
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:

Maximal March 801 24MED1 802 24MED1 803 24MED1 804 2000 801 24MED1 802 24MED1 803 24MED1 804 24MED1 805 24MED1 802 24MED1 803 24MED1 804 24MED1 805 24MED1 805 24MED1	al Nonina Sedona Sedona Eltero Leda	Califor New Y	Pemale Female Cessilia	Roninal Graduate Undergraduate	Nominal Politica	Noninai	Numeric	Huneric	Average more (grace)		000	2.45		Saur
601 04NE01 802 04H203 801 00E01 802 00E02 003 00E02	Sedone Sedone Eliero Lecka	Califor New Y	Penale Penale	Graduate Undergraduate	Folitica.	1.17			Teat Inte		and the second	D00.11		
802 DWW200 801 D0801 802 D0802 700 D0802	Sedore Elera Lecka	Nevi Y.,	Fattale	Undergraduate		1.9	5.571	1.751	17.0					
601 00601 602 00602 503 00602	Eltera Ledos	Nevi Y.	Calcol No.		Batt	65	1.047	a.607	63.0					
802 20802 802 20802	Lador		renae	Graduate	Math	US .	1.381	1/908	30					100
203 20203		Dimit Y	Male	Graduate	Econ 🔿	L5	107140.	1.389	78.0					1.50
	Defiance	Ohio	Mala	Graduata	Math	LS	1.904	3.273	0.23				There is a set	
P0800 P05	Tel Avia	25/30	Mat	traduate	EC04	19 aei	\$1533	1,401	19.0		A 14		Lineary 24 (2007	10
605 30805	Ones	North	Male	Graduate	Palitica	5	1.0	1.246	0.89		1 m		Complete Children	
E03 DAMED1	Lbeck	Kandad	Feasile	jundergrad.use	Politics	1.6	1.140	1.514	17.0			Volue		
COM DANEOH	Manifea	Carada	renaie	Undergraduate.	Math	Canada	1.0	0.465	31.0			10		
605 DANEOS	New Y	Nevi Y	Pamaia.	Graduate	Math	L5	1.714	1.723	71.0			1		
506 20506	Hot C	Moting .	. Maie	Lindergraduate	Ecol :	1.9	3,6	1.61	\$2.0			0.5		
206 DANCOR	Jove	Virginia	Fende	Greduetz	Math	15	1/952.01	1.386	19.0			0.302		
207 pde07	Varna	Dupeta	Mate	Graduate	Folitica	D./pena	L37L.	1.301	79.0					
608 00608	Mescow	Rissa	Male	Graduate	Folitics	RUSSIA -	1,571	4,178.0	. 10.0					
607 DANEDT	Drunk.	New You	Pemale	Undergraduate	Math	US .	1,141	1.0	12.0					
DOD DANEDOS	Maxic	Litals	Penale.	Undergreduele	Ecpt	10	1.0	1.497	0.08					
E09 DAME09	Anste	Holand	Feilale	Undergraduate	Math	Holland	1.047	1.361	75.0					
eto pavete	Mestod	Mexico	Pernale	Gredulate	Palitez	Nepicp	5.615	1.937	\$5.0					
CLL DANCLI	Cereces	Verse	("emale	Undergraduate	Matte	Veneza	0.0	1.941	\$2.0					
609 20609	San Juar	Puerto.	Mae	Graduate	Foliation	US	1.714m	1.602	. 95.0					
eta paveta	Remote	Chegort	Penale.	Undergraduate	Econ	15	1.047	1.406	67.0					
£10 00£10	New Y	Nevi Y	Mala	Undergraduate	Ecos ·	LS .	1.146	1.541	#2.0					
ELS DAMETS	The s	Massia	Penale	staubara	PORRS	1.8	8,537	109910	19.0					· 1954
CL4 DANE14	Detting	Chris.	Famale.	Undergred.ete	Math	China	5.0	0.314	79.0					
611 00E11	Stacih.	. Sueden	Male	Lindergrad.une	Politics	Sueden	1.047	3.596	0.68					
E12 20E12	Enlast	Mine	Mec	Oreducitz	Ecot .	US .	1.426.0	3.056	95.0					
613 DOE13	Interc	Penns	Male	Undergraduate.	Math	LS .	1.095	1.804	0.85				11	
ELS DANE19	1.000	iORah	Fensle	Underproducte	Econ	1.5	1.095	1.0	64.0					
P1300 P13	DIO DI	Argen	Mee	Graduatz	Politica	Argentrio	1.571	8.565	85.0					
c13 00E15	Acres	Louisers	Male	Undergreduate	B:pt	1.5	1.047	1.505	79.0					
EL6 DANED1	Los An	Cafor	Feisale	Graduate	FOR SIG	1.8	1.571	1.95	Right click (dr	left+alt) for context menu				
E17 DAMERG	Sedona	Arizona	Pemale	Undergraduate	Math	US	1,040	0.687.1	63,0					
		Dave Y	Make	Graduate	Rati	LS.	1.386	1.905	18.0					
C10 00001	CORN.								100 T					
	02 Doesent 04 Doesent 05 Janetzo 06 Janetzo 07 Doesent 08 DOEsent 08 DOEsent 08 DOEsent 08 DOEsent 08 DOEsent 09 DOEsent 10 Martin 11 Martin 12 Martin 13 Martin 13 Martin 13 DOEsent 14 Martin 15 Martin	Dist Subset Labor 99 Subset Parrito 90 Subset Parrito 90 Subset Parrito 90 Subset Next 91 Subset Next 92 Subset Next 93 Subset Next 94 Subset Next 95 Subset Next 96 Subset Next 96 Subset Next 97 Subset Next 98 Subset Next 99 Subset Next 90 Subset Next 90 Subset Next 90 Subset Next	Display Likewith Context Display Material Material Material Display Material Material Material <td>00 Medical Liberal Director Ferridor 00 Medical Liberal Director Ferridor 00 Medical Director Construction 00 Medical Director Construction 00 Medical Director 00 Medical Director</td> <td>00 Marcel Laboral Function Marcel Product Pro</td> <td>00 Millioni Jakradi Naroka, Franka Parkagoskano Parkagoskano</td> <td>District Liberal Incode Feature Index products Patter Li 00 StateCol Herrike Z. State Index products Patter Patter</td> <td>District Jahred Descence Finance Finance Descence Finance Finance Descence Finance Finance Descence Finance Finance Descence Descence Finance Descence Descence</td> <td>03 Medicii Jacrat Service Serv</td> <td>03 Medial Randa Koncel Fester Marke Indergraduate Fields 6 1.142. 1.141.</td> <td>03 Medial Jacked Activation Price Price</td> <td>00 Marcoll Labord Former Marcoll Former Former<td>00 Marcial Mar</td><td>000 0.00001 0</td></td>	00 Medical Liberal Director Ferridor 00 Medical Liberal Director Ferridor 00 Medical Director Construction 00 Medical Director Construction 00 Medical Director 00 Medical Director	00 Marcel Laboral Function Marcel Product Pro	00 Millioni Jakradi Naroka, Franka Parkagoskano Parkagoskano	District Liberal Incode Feature Index products Patter Li 00 StateCol Herrike Z. State Index products Patter Patter	District Jahred Descence Finance Finance Descence Finance Finance Descence Finance Finance Descence Finance Finance Descence Descence Finance Descence Descence	03 Medicii Jacrat Service Serv	03 Medial Randa Koncel Fester Marke Indergraduate Fields 6 1.142. 1.141.	03 Medial Jacked Activation Price Price	00 Marcoll Labord Former Marcoll Former Former <td>00 Marcial Mar</td> <td>000 0.00001 0</td>	00 Marcial Mar	000 0.00001 0

JRip(seed=1):

Carefa Characteria	Colors and the literature									
Canadian Control Line ter Accord	See Later to Control									
Chases 384p F3 N2.0 02-5										
Test options	Ownite output									
 Use training set 	the second second large second and				2					
C Supplied test ant Sel	JRIP pulses									
Cress-validation Folds III		AND 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1								
C Percentage galt 5 50	(First Hame = XOED1) => City=Lockswama (2.0/0.1) (First Hame = XOED1) => City=Elmins (2.0/0.0)									
Allow and parts										
Par optara	 Sitz=Sedons (31,0/27.0). 				-					
(Hern) Oty	· Number of Rules : 1									
Start Stat										
Result list (right-click for optional)	Time takes to build model: 0.04 seconds									
In Fill Roman State										
	see Suppary see									
	Correctly Classified Instances 7	20.568Z	5							
	Kappe statistic 0.	19.4118								
	Nean absolute error 0.	0583								
	Roch mean squared error 0.	1707								
	Pelative absolute arror 87.	1006 0								
	Total Sumer of Instance 36	(011 #								
	Detailed Accuracy Sy Class									
	TE Bace FF Pate Preciaio	Recall F-Measure	ROC Area Class							
	a a p	0 0	0.541 los Angeles							
	1 0.671 0.1	1 0.182	0.563 Sedona							
	1 0 1	1 1	I Elmira							
		à à	0.541 Defiance							
	0 0 0	0 0	d Sel Tel Ario							
	2.0 2.0 2.0	A	A 843 AL 414							

JRip(seed=2):

Woka Exploser	et attributes Visiality			
Department Control (Laugher 2003-0416 Meet	tatholdes volatie			
Chapes Rider #1.51-820				
Test options Oc	sssfer output			
Vice taxing set Supplier textual Vice validation Vice validation	Dealers Starter Major Country Age SRT Average store (preds) est mode evaluate on training date			
Plant) City +	- Classifier model (fill training	(286		
Start Start <th< td=""><td><pre>KHF mile: ************************************</pre></td><td>na (3.5/0.8) nie</td><td></td><td></td></th<>	<pre>KHF mile: ************************************</pre>	na (3.5/0.8) nie		
60 33 90 90 90 90 90 90 90 90 90 90 90 90 90	orrectly Classified Tartance poprestly Classified Instances apps resistic an absolute teror of Med Systeb teror Blatics absolute synce of relative appsed error tel Number of Instances == Petalled Acoustoy Sy Class ===	5 28 0.0663 0.0622 0.1764 98.7661 k 96.6957 k 34	14.7009 4 85.2041 4	
K28.6 DK				[10] and

JRip(seed=3):

Castler				
Choose 3Nip -F 2 N 2.0 -0 2 -5 3				
Test options	Classifier output			
🔹 Use training set	Student Status			
C Suppled test and Set	Majoc			
C many statements and a large	Country			
C Crico valuedori rolla- In	Page FET			
C Parcentage aplit % 50	Reversos score (prada)			
Mare options	feet mode evaluate on training data			
(Nam) City	 Classifier model (full training 	44C)		
C data 1	JRIF miles:			
Tests				
Resoult liet (right-click for options)				
20 (54) 52 - rules JRID	(First Hane = 20803) -> City-Lankana	ma (2.0/0.1)		
00 Set 15 - Polen Dop	-o City-Sedone (32.0/25.0)			
30 (58) (53 - rules, Ridlar				
71:00:03 - rules. Rieler	Number of Fuller (2			
31.01115 - rules, Rider				
Dimital-ides Mu	Time takes to build model: 0.04 seco			
	www Evaluation on training set week			
	ees Simulary ees			
	Providents Providence Providence	32	11 117 A	
	Correctly Classified Instances		14.1000 9	
	Sama statistic	1 0561	03.0346.4	
	Nton abdalung enror	0.0522		
	Root mean squared error	0.1764		
	Relative shaolute erour	41.7691 4.		
	For relative squared error	96.0957 %		
	Total Number of Instances	34		
	men pergried scontack of crasp men			

Ridor(seed=1):

tarus tore igrāđej reining data uli training sat/ ===	
einideel cules	
8) incl. the defauls rule() 1 al: 0 seconds nog get ===	
rtance 3 8,225 4 Tabutance 3 0,1765 4 0,0028 0,0028 0,2091 54,7235 4 120,7091 4 120,700 4 10	
	aven 1947137 1 aven 1977071 1 av 34 y Claro FF Barn Frenision Recall F-Measure RoC Line Class 0 0 0 0.5 Los Angeles

Ridor(seed=2):



Test Data:

llo ll							- on magner	CREAT STR	August and and	STATE OF LEV	196.8				
-	D Last Nore	First Name	City	State	Gender	Student Status	Major	Country	Age	SAT	Average score [gnade]		20	11	- Inc.
-	0.000020	30903	Tel data	Chie.	Main	Deskale	Freis	16	1.904	1.108	18.0		COLI		3940111
	0.056 DOE21	20204	Tel Avia	Neg Y.	Male	Graduate	Ecot	breat	1.333	8.490	0.63				
	0.133. DOE22	00605	Cinte	horth	Fernal	Erat ate	FORTICS	18	1.0	1.361	6 .0				L. Art
	0.200623	DAMPEG 7	Liberal	Karman	Penale	Graduate	Politica	15	1.147	0.951	87.0				
	0.36600524	DANED+	Mantheal	Canada	Female	Undergraduate	Econ	Cartada	1.0	1.521	83.0				
	0.333D0E29	DAMEOS	New You	New Y	Penale	traduate	Math	L8	8.714	8.772	71.0			Type: Maneric	
	0.400825	20206	Hot C	Massa.	Male	Undergraduate	Econ	LS	1.0	1.493	82.0			Unique: 10 (300%)	
	D.466DOE27	DAMETON	Lava	Virginia	Fecule	Graduate	Nati	1.6	1.952	8.:90	79.0				
	0.533D0E28	30807	Nama	Bulgaria	Mate	biaduate .	POIDS	Bucario	1.571	1.321	79.0				
	0.6 D0E29	DOBOB -	Mascow	Runnie	Mala	Graduate	Politica	Runnia	1.571	1.191	10.0				
517	0.666	DANEO?	Donk	New Y.	Festale	Undergraduate'	Math	1.6	1.140	2.0	#2.0				
2	0.733 DOC 31	DANEOR	Mexico	Utah-	Penale	Undergraduate	Duon	2.6	1.0	8.536	0.08		17		
21	0.600832	DANEOS	Anete	Holand	Pareale	Undergraduate	Hats	Holland	1,047	8.371	15.0		2		
80	0.866 DOE33	DANE 18	Mickico .	Mexico	Fenale	Graduate	FORTICS	Mexico.	1.615	1.0	45.0				
1	0.933. D0E34	30801	Elmiro	New Y.	Male	Graduate	Math .	US	5.380,	1.978	78.0				
1.5	1.000035	20202	Lada	New Y	Male	Graduate	Ecos	1.5	8.714	8.415	18.0				
													-		• Noval
			_	_	_		_	_	_	_		Under DE Dercel	J		
											201				
					Pare	and a second									

Ridor(seed=2):

Woka Taplater	_						and the second distance of the second distanc	The second s	a
repracess Grooty Cluster Associate	Select attributes Visualize								
Center									
Chapes Ridler F 1 - S 1 - N 2 D									
Test options	Classifier output								
use taiwig set	SAT								
C Suppled test set	Jourses acces (prede) Teom moderevolumite on teninicy date								
Cress-validation Folds In									
() Percentage split 15, 06	flassifler model (full training set)								
Nina codona									
rar sport a.c.	Fipple DOwn Rale 1	serner (Ridor)	rules						
(Nem) Oty	r.								
	City = Sedone (34	1.0/0.85							
Start Start	Total number of m	tas direct the	defects rate	19212					
Result list [right-cick for options]		the family of		1.1.1					
20.54152 -rules, pap 20.55:15 -rules, Fap	TIME CARES 10 DULL	3 NO3E1: 0 80	101035						
20.55:34 - rules, JRp	- Industion of								
20 (59) 53 - rules, Richar 21-08 - 13 - rules, Richar	Jumary	second and							
1101:15 - rules. Rider									
21.01/31 - rules 240	Control Classified Intensors 3 0.1239 9								
En al 11.20 - n des Factor	Kappa statistic	ried mistance	a		2111,00	80 C			
	Mean absolute erro	14 C	0.04	2.8					
	Root mean squared	acror	0.25	01					
	Root relative equa	error red error	197.72	71. 6					
	Total Runker of In	staddep	34						
	Detailed Accur	acy By Class	<u></u>						
	17	Rate IT Ball	. Precision	Recall	r-measure	NOC Area	Clapp		
	0	0	0	0	0	0.6	Los Angeles		
		1	0.026	1	0.162	0.5	Sectoral		
	0.0	4	p	0	0	0.5	Lationente		
		a	0	p	0	0.0	Defigure		

Test Data:

	on: test est-weige file	in unsupervi	inter a trib	uter.Repla	ceMesine/	alues waka filter	Luneaper	vied.atri	auto hierre	also GLO	T0.8					
10.	D Last Nere Numeric Nominal	Pirut Name Norvital	City	State	Gender	Student Status Rominal	Major	Country	Age Numeric	SAT Numeric	Average score (grade)		ER.	1	iave	_
2	0.000020	20803	Tel Avtir	Ohio	Male	Craduate	Econ.	LS.	1.904	1.298	65.0					
	0.056D0021	20004	Tel Avril	Nex Y.	Hele	Graduata	Ecot	brasi	8.333	1.491	69.0					
	0-133 - D0E22	20E05	Cinex.	North	Fenale	Graduate	Folitics .	US	1.0	1.261.	96.0					1.0
	0.7 DOE75	DANE 07	Liberal	Karman	Penale.	Graduate	Politica	1.5	1.141	0.553	87.0					
	0.36600624	DANED+	Mantreal	Canada	Female	Undergraduate	Econ	Cartada	1,0	1.521	0.18			142112	an each	
	0.333D0E29	DANEOS	NOV You	New You	renae	Stackard	Math	UB	\$1714.S	9.772.0	31.0			Linear .	Manueric M. Commit	
	0.400826	202506	Hot C	Massa.	. Maie	Undergraduate	Econ	1.5	1.0	1.493	12.0			- and an	in the state	
	D.466DOE27	DANEDE	lava	Virgitia	Feixale	Graduate	Nati	LS	1.952	3.:92	79.0		1			
	0.533. DOE25	30807	Vanió	Bulgaria	Mate	biaduate	Politics	Bulgaria	8.571.	1.321.	79.0					
	0.6 D0829	DOBOB	Mascow	Ruma	Male	Graduate	Palitics	Ruma	1.571	1.391	70.0					
	D.666	D6NE07.	Dnik	New Y	Festale	Undergraduate	Math	1.9	1.140	8.0	#2.0					
	0.733D0E31	DANK08	Mestin	Utah-	Fende	Undergraduele	Duon .	2.6	1.0	1.536	0.08		0			
	0.800832	DAVEOS	Anata	holand	Tenale	Undergraduate	Path	rolard	1,047	8.37L.	15.0					
ļ	0.866 00E33	PARETR	946,600	146000	resse	Fragrage	F083CS	PRENCO:	1,619.1	3.0	45.0					
ŝ	0/933D0E34	20801	20mmo	NOW Y	2494	Graduate	Path	1.5	\$1380,	1.976	78.0					
	Lupueto	pueue	1403	D90917	New C	Incesting.	ECD4	pa .	8.716	1.415	(L. 10)					
															•	Roue
												tinto DE Derei	Ŀ		•]	fox
							_			_		Tarta DE Deroi	J		•]	Rout
			_	_	_			_	_			Units DE Devel			•]	Road
			_		_	_	_			_		Units DE Devel	J		•	Rout
												Units DE Devel	J		•	Rous
												Link, DL Deal	J		•	ficus
					Row	pre						Link, DL Deal			•	foat

JRip(seed=1):

Preprincess Crestly Cluster Associate Cluster Crosse Ridler # 1-5 1-N 2.0	elect attRutes. vecalize	
Chases Factor #1 - 5 1 + 5 2.0 Text colors Use Transverget Use Transverget Supplied text set D Text colors Use Transverget D Text colors Use Transverget	Consider pages Touchest Taisues Major country Jos Sait Joverage and untre (grade) Test modelines applied test, set: six untreen (reading incrementally) == Classifier model (full Itaining set) == TRIE rules: == Test rules: == Test rules : == Test rule	
	Foot max squared stor: 0.2401 Palative should arrow 100 % Moul milaive squared choor 100.0118 % Total Ranker of Instances 16 === Dennied Armiracy By Class === II have IF Fame Freculation Frecold F-Measure MoC Area Class	

Weka Espices		
Resonances Classify Cluster Association	e Geschatzbutes Vacable	
Classfer	M The county burg to a volume of the second se	
Ghapes Ridor € 1-5 1-8.2.0		
Test options	Oberfer subut	
Use homig set	Divident Status	
@ Suppled test set	3 Major	
Distance and the line	2 Country	
	Roje com	
D Aercentage spit 1 1	Bail	
Mare options	Test mode turer supplied test set: site unknown (reading incrementally)	
(Han) City	 end (lassifier model (full training set) end 	
Start Star	SRIF miss:	
Date & Bet Statut and the entire all		
01.06165 - n.dec Title		
1101513 crules Pap	 CLCP=CE1 &V1V (16(0/16.0) 	
11:07:30 -rules. 3Rp	Sumber of Rules 1 1	
23.07.27 - rules kilow		
	Time taken to build model: 0.02 seconds	
	www. Evaluation on test set and	
	lumary	
	Correctly Classified Datances 1 8.25 %	
	Desired and and a second	
	Seen subsidie error 0.1172	
	Root mean squared error 0.2431	
	Felative absolute error 100 #	
	Noot FLATIVE SQUARED SIDON 100,3114 %	
	10191 Water of Tildforden Te	
	letailed Accuracy By Class	
	To Basing 75 Taking Restal and Property Following SNT Long Classes	

JRip(seed=3):

landfor												
Chapes Rider # 3 - 5 1 - N 2 0												
est options	Ossister output											
🗇 USE BIBHNIG HET	Beerage score (grade)											
Suppled test set	les mode une spipile tes set univer une and intraction intractionally											
Cress-validation rokis (2)												
Nercentage epit 10 00												
Minus configures	Elpple Down Ba	Ripple Down Male Learner(Rider) reles										
Part your a												
iani) Oty	City - Tal Avi	v (16.0/	10.1									
	Table and											
Stert, Star	Ther becks a	e intes d	ince. une	Selarit true								
eault list (right-click for options)	line takes to	lime takes to build model: 0 securids										
2106/55-048.800												
1:07:10 - rules, JRp	Eveluation	DI DEFT.	PET NON									
10/120 / rules Futur	and Sumary of	2										
1:07:27 - rules Rider	Correctly Clas	eified In	stances	1		6.25	4					
	Incorrectly CL	ensified .	Instances	15		98,75						
	Rappe statistic 0											
	Seen shablute error 0.1172											
	Foot mean aggated struct 0.3423											
	Fort relative	Prelactive approach approx										
	Total Sunker o	f Instance	12	16								
		manarad										
	men Detailed A	centracy S	Class and									
		IF Race	FF Fate	Precision	Recall	F-Measure	ROC Area	Class				
		4	4	0	0	0	0.5	Sefiance				
		1	4	0.053	4	0.116	0.5	Isl Aviv				
		0	4	0	0	0	0.5	(Lines)				
		a	4	0	3	0	0.5	MIGTORIA				
				0	10	0	0.5	Real Variate				

Ridor(seed=1):

Wolca Taplosar		Constant Summer Sum							
repracess 0/9517/ Cluster Associate	a Galert attituise Viscalae								
Janifer									
Chapes Rador + 3 - 5 1 - N 2.0									
Test options	Ossife output								
 Use training set 	Aga .	1							
@ Supplied test set	557								
Criss-validation Hillis III	Average source (prace)								
C Percentage split 15. 56									
Mare options	Classifier model (full training set)								
		10							
Nen) Oty .									
Start	-> Ciry-Gal Aviv (16.0/18.0)								
Result list (right-click for options)	Purber of Poles : L								
2108/55 - rules 340 11/07/05 - rules 340									
LET.D MAR St.									
21/07/20 - rules Rider	The takes to being model, o.o. seconds								
1230(227-FORE REEF	Evaluation on test set								
	Justary								
	Convently Classified Instances 1 5.35 4								
	Incorrectly Classified Instances 15 91.75 4								
	Vappe statistic 0								
	Wean absolute error 0,1171								
	Felative theorem terror 101 4								
	Root relative squared error 103.3114 8								
	Total Wanher of Instances 16								
	www. Period Leng Variation No. Cleane www.								
	second when the second s								
	TE Mana TE Fata Frecision Recall T-Measure ROC Area Class								
	0 0 0 0 0 0 0 15 Definite T 1 0.051 0 015 Definite								
	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	-							
Data									
)K		Lop							
110									

🖕 Woka Explorer								4.364				
Prepraces Otesty Ductor Assoc	ate Select attributes Visualize				-							
Geater												
Choose Rador € 1-5 1-N.2.0												
Test options	Classifier output											
O Use horizing set	Average	score (gra)	te)									
a Succied text set	Test modernser supplie	Test mode user supplied test set: size unknown (reading inorementally)										
C route abilition with In	and Charifter motel 4	new Classifier model (full training and and										
C Cress viewood i roks in	_											
Cuercesde ens	Ripple Divn Sale Learn	er (Rider) :	niles									
Mare aptions			20052									
B1-12-00-	City - Tal Aviv (16.0	16.01										
Send City		Tradition of the second s										
Start	Total number of rules	Total number of rules (incl. the default rule): 1										
Result list (right-click for options)	Time caper to notili no	Time taxes to build model) 0 econds										
21.061\$5 - rules JRip												
21:07:03 -rules. Rip 21:07:10 - order. Rip	Evaluation on test	ant must										
110/120 - Mes Roll	eme Summary eme											
21:07:27 - rules. Fisher	Correctly Classifies D	istances	- E		6.25	4						
	Incorrectly Classified	Instances	1.5		98,75	4						
	Kappa statistic		4									
	Root mean senated error	3	0.34	23								
	Relative appointe erro	2	100									
	Root relative equared	rear	141.27	43.9								
	Total Number of Instan	346	16									
	Detailed Accuracy	Sy Class										
	100.000					w.201100.000	CARACTER IN CONTRACTOR					
	IF SACE	dr Fate	Precision	Recall	F-Measure	NUC AREA	CLISSS					
	ĩ	1	0.065	1	0.116	0.5	Tel Aviv					
	a	a	D	D	0	0.5	Climate					
	a	α	D	D	0	0.5	Liberal					
	0	d a	0	0	0	0.6	Man Week					

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							
<u>Ridor</u>	Seed =1	Root Mean Squared Error=0.2508 Mean Absolute Error=0.0629							
<u>Ridor</u>	Seed=2	Root Mean Squared Error=0.2508 Mean Absolute Error=0.0629							

Testing Data set Performance:

	TE	ST 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

	TRAINING				
JRip	Seed=1	Root Mean Squared Error=0.1707			
		Mean Absolute Error=0.0583			
<u>Ridor</u>	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

Weks Explorer	
Preprocess Cleasify Charler Associate Select attribute	a Venaster
Ousterer	
Choose Cobweb -4 1.0 < 0.002820947917738781	1542
Cluster mode	Clusterer output
@ Use baining set	
C Suppled test set	Time taken to build model (full training data) : 0.01 seconds
Percentage split % 66	Model and evaluation on training set
Cleases to clusters evaluation	Contered Testacces
(Nors) contact-lenses +	
Store clusters for visualization	3 1 44
	N 1 1 441
apore attroutes	II T 1.(44)
Sart. Des	5 1 (44) 5 1 (44)
Result lat (right-club for options)	12 1.(49)
	$ \begin{array}{cccccccccccccccccccccccccccccccccccc$
Status OK	
😨 🚞 🗵 🧔 🕑	- N: 12 41 423/215

EM

Cluster mode	Left-click to edit properties for this object, right-click/Alt-Shift-left-click for menu					
(a) Extering pet () Supplied text set () Sec () Percentage path () Clonest to clusters evolution () Oral control lenses () Store clusters for visualization () Store set this tex	<pre>weaka.clusterers.EM = 7 100 -M -1 -K 10 -max -1 -11-cv 1.0E-6 -11-iter 1.0E-6 -M 1.0E-6 -max-alors 1 -9 100 Emplation: contract-lenses Instances: 24 Intributes: 5 app apotalls-greecrip aptotalls-greecrip aptotalls-greecrip attapmatism tage-products</pre>					
	outrat-leases					
Start store	Test mode: evaluate on training data					
00:13:08 - 5M 00:14:03 - Cobreek 20:15:12 - 01						
	Number of clusters meleoted by cross validation: 2 Wanter of iterations performed: 10 Cluster Attribute 0 1					
	(0.45) (0.39)					
	age 5.1102 4.8598 pre-penahyopin 5.8346 4.1654					

Farthest First:

Choose FarthestFirst 4/2 51		
Cluster mode	Clusterer output	
Luter nool Luter nool Luter nool Luter nool Luter could be test set Proceed te	Tables: webs.clusterers.fathestFlust -# 2 -5 1 Bactaion: contact-immas Bactaion: 24 Attributes: 5 setupation setupation temperature Test moder evaluate on training data Clustering model (full training bet)	
	satuserinist Cluster d pre-presbygio myrpe no normal soft Cluster d young hypermetroge yes reduced none	
	Time taken to build model (full training data) : 0 ercosds Nodel and evaluation on training set	
lata OK		Log

Filtered Cluster:

Weka Explorer							
Preprocess Classify Outlier Associate Select attroute	a Vinaire						
Outerer							
Choose SimpleKMeans N2 A 'weba.com Euclid	arCistance -9 Fest-last" 4 500 -rue-sides 1 -5 10						
Cluster mode	Custerer subjut						
C Lise training set	Relation context-lenses +						
Suppled test set Set-	Instances 24 Attributes 5 age						
Percentage solit % 66							
Classes to clusters evaluation	apertacle-prescrip						
(New) contact-levent +	tear-prod-size						
[V] Store clusters for visualization	outstollinge						
	lest model evaluate on training data						
Ignore attributes	www Clustering model (full training met) www. Cluster 0						
Start İtra							
Result list (right-dick for options)							
00112365-EM 0011405-Cobweb 0011712-EM 0012512-Farther#hot 0021238-PhenedCostene	<pre>(((((((2,01,2,01))0,2,01))0,2,01)00,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,2,01)00,0,0,01)00,0,000,000,000,000,000,0</pre>						
81:25:25 - MakeDenshi/SasedClusterer 81:25:28 - Singlet/Means	Time taken to build model (full training data) : 0.07 seconds						
	Model and evaluation on training det						
	Clustered Instances						
	0 20 (534) 1 4 (179)						
Datus OK	······································						
😨 🗒 関 🎯 🙆	- Part () 100 -						

Hierarchical Cluster

Waka Esplorer		and Calendary States, Stat
Preprocess Gassify Custer Associate Select attribut	zee Venales	
Oustere		
Choose SimpletCHevens -N.2 -A "weka core.Buda	dear/Debance 4 first-last" -1 500 mun-alote 1 -5 10	
Ouster mode	Clusterer sulput:	
O Use training set	Discrete Estimator. Counte = 7 5 (Total = 12)	
C Supplied text set	Attribute: tear-prod-rate	
# Percentage galit %	Attribute: contect-leases	
Casses to clusters evaluation	Discrete Estimator. Counts = (2 3 5 (Total = 13)	
(Novi) contact-lenses +	Cluster: 1 Prior mediability: 0.3879	
V Store dusters for visualization	construct a server produced provide of the	
	Attribute: age	
Igrore attributes	Attribute: stectacle-prescrip	
Chart IIma	Discrete Estimator. Counts = 2 5 (Total = 7)	
The first shift or second	Attribute: antigmation	
Network (reproduction for bollonia)	Discrete Estimator, Counts = 2 5 (Total = 7)	
00:D415 - Cobweb	Disorete Estimator. Counts = 4 8 (Total = 7)	
00:17:12 - EM	Attribute: contact-lenses	
00:18:52 - Farthesener 00:21:29 - PiteredCusterer	Discrets Estimator. Counts = 1 2 5. (Total = 0)	
01/25/20 - HerardskalClusterer	19 3 37 19 9 9 3 A	
01-25-25 MakeDensitySener/Clusterer 01-25-28 - Genela/Maanse	Time taken to build model (percentage split) : 0 seconds	
	Closteret Instances	
	0 6 (678)	
	1 1 (334)	
	A DECEMBER OF	1
	log likelihood: -4.55298	
	- rC +	
Outer	C.KORI (III)	
DK.		100
		- 108 AM

Simple KMeans:

recent Classify Custor Associate Selectory forer Decore Serveret Means -N 2 -4 Server Associated	t affributes Vaualize							
terer Transe - SimpletMeans -N 2 -A Suela m								
Change Scroplet/Means -N 2 - 6 Stretch to								
a sea parte de la constante de	ore-EuclideanDistance -R first-last" -1 500 -num	Helots 1 -5 10						
ater mode	Clusterer output							
Use training set	S I NOR STRATEGIC							
Suppled test set	www Model and evaluat	weew Model and evaluation on test split week						
Percentage split	6 65 kHenna							
Classes to dusters evaluation								
(New) contact lenses	Bunber of iterations:	3						
Store clusters for vacalization	Within cluster sum of	f squared errors:	15.0					
	Hissing values global	ily replaced with :	seat/mode					
Ignore attributes	Cluster centrolds;							
Start Stop		222403	Cluster#	10 A				
aultilet (right-dick for options)	ATTRIDUCE	(15)	(12)	(5)				
13036 - EM								
19:05 - Cobreb	BOST AND	presbyopic pr	e-presbyopto	pre-sbyopio				
18:52 - FarthestFirst	astignation	763	35	Ats				
21:29 - FilteredChaterer	tear-prod-rate	reduced	reduced	reduced				
25:20 - Herarchca/Custerer 25:15 - Maka/assitu/Lucat/Duraterer	contact-lenses	DODE	none	DODE				
City 28 - Separate and a second second								
	Time define the build a							
	line takes to putie a	sodel (percentage	sbric) : 0 80	contra				
	Clustered Instances							
	*** ********							
	0 4 6 6690							
	1. 1.000							
4.6								
						Log		
		-	-	-	-			

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

Relatio	on: supermarket					
No.	1: department1 Nominal	2: department2 Nominal	3: department3 Nominal	4: department4 Nominal	5: departme Nominal	5340
1					<u> </u>	Jave
2	t					
3	-					Apply
4	t					IC Apply
5						
7	+		t			ype: N
8						que: 0
9	t		t			Weight
10	-		-			to 47.0
11						1047.0
12	t					
13	t	t				
14						
15						Visualize All
16	t				t	
17						
18	t •		τ			
20	t +					
21	•	+			+	
22	t	t				
23	-	-				
•					•	

OUTPUT:

Apriori Algorithm:



FP-Growth Algorithm:

🥥 Weka Explorer
Preprocess Classify Cluster Associate Select attributes Visualize Associator Choose FPGrowth -P 2 -I -1 -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1
Start Stop Associator output
Result list (right-dick for (13:43:15 - Apriori 13:43:15 - Apriori 13:43:12 - FPGrowth 1. [fruit=t, frozen foods=t, biscuits=t, total=high] 2. [fruit=t, baking needs=t, biscuits=t, total=high] 3. [fruit=t, vegetables=t, biscuits=t, total=high]: 5. [fruit=t, party snack foods=t, total=high]: 854 = 6. [vegetables=t, frozen foods=t, biscuits=t, total= 7. [vegetables=t, baking needs=t, biscuits=t, total= 8. [fruit=t, biscuits=t, total=high]: 954 ==> [bread 9. [fruit=t, vegetables=t, frozen foods=t, total=higf]: 4. III.
Status OK Log x 0

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: 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.