WALCHAND COLLEGE OF ENGINEERING, SANGLI.

(An Autonomous Institute)

Final Year B.Tech. (Information Technology)

END SEMESTER EXAMINATION NOV./DEC.-2016

DATA MINING (2IT402)

ESE

Day, Date and Time: Thursday, 01/12/2016, 03.00pm to 05.00pm

Exam Seat Number: _____

Max Marks: **50**

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions: i) All questions are compulsory. Writing question number is compulsory.
ii) Figures to the right of question text indicate full marks.
iii) Assume suitable data wherever necessary.
iv) Avoid unnecessary explanation; answers must be concise and with given question only

Text on the right of marks indicates course outcomes (only for faculty use).

			Marks	
Q1		**State 'True' or 'False'; if False state correct statement.** i.) F-test statistic is used for a global test of significance. ii.) A q-q plot is a convenient way of graphically depicting groups of numerical data through their quartiles. iii.) Multilevel association rules are mined with the help of concept hierarchy. iv.) Count matrix is used for binary splitting of numerical attributes. v.) If we have an attribute X that has distinct values for each record, then Info (X,T) is 0, thus Gain(X,T) is minimum.	5	CO1
Q2		**Write short note on following Terms (ANY 4)** i.) Web Mining Taxonomy ii.) Explain Multidimensional Analysis with example iii) Spatial Trend Analysis iv.) Trend analysis in time series data v.) Application of text mining	8	CO2
Q3	A)	State the methods to fill in the missing values for attributes in data mining process.	3	CO1
Q3	B)	Write 3-4-5 rule? Why it is used?	3	CO2
Q3	C)	What are different methods for mining multidimensional association rules ?	3	CO1
Q3	D)	Explain the terms- Guillotine cut, Overfit and Attribute selection error w.r.t. decision tree. **OR** Explain the terms- core object, directly density reachable and density connected in Density-based clustering method with figure.	3	CO2
Q4	A)	For following crosstab; calculate t-weight and d-weight to fill all values marked by shaded cells.	4	CO3

Year of passing	Students joined Higher Education			Students joined for job			All class		
	Count	t-Wt %	d-Wt %	Count	t-Wt %	d-Wt %	Count	t-Wt %	d-Wt %
2012	12		40.00	48	80.00		60	100.00	33.33
2013	8	13.33		52	86.67		60		33.33
2014	10	16.67		50		33.33	60	100.00	33.33
All Years		-	100	150	-	100	180	-	100

Q4 B) Calculate the correlation coefficient between the number of study h... sleeping hours of different students. Comment on correlation.

Number of Study hours	2	4	6	8	10
Number of Sleeping hours	10	9	8	7	6

OR

Find linear regression coefficients and predict salary for 10 Years' experience using following data.

Experience (Years)	3	8	9	13	3	6	8	21	1	16
Salary (Thousands)	30	57	64	72	36	43	59	90	20	83

Q5 A) From given data; identify the class of following test case by using Naïve Bayes Classifier. 6 CO3
Test case – {Refund = No; Marital status = Married; Income = 120 K}

Marital status	Income (K)	Home loan refund	Defaulter
Single	125	Yes	No
Married	100	No	No
Single	70	No	No
Married	120	Yes	No
Divorced	95	No	Yes
Married	60	No	No
Divorced	220	Yes	No
Single	85	No	Yes
Married	75	No	No
Single	90	No	Yes

Q5 B) Apply agglomerative hierarchical clustering algorithm with complete linkage for following data. Draw dendrogram. 6 CO3

	A	B	C	D	E
A	0	-	-	-	-
B	662	0	-	-	-
C	877	295	0	-	-
D	255	468	754	0	-
E	412	268	564	219	0
F	996	400	138	869	669

Q5 C) From given data identify the class for test case using Naïve Bayes Classifier 6 CO3
Test case- (Have legs: No; Give birth: Yes; Can fly: No; Live in water: Yes)

	Name	Give birth	Can fly	Live in water	Have legs	Class
1	Human	Yes	No	No	Yes	Mammals
2	Python	No	No	No	No	Non-mammals
3	Salmon	No	No	Yes	No	Non-mammals
4	Whale	Yes	No	Yes	No	Mammals
5	Frog	No	No	sometimes	Yes	Non-mammals
6	Komodo	No	No	No	Yes	Non-mammals
7	Bat	Yes	Yes	No	Yes	Mammals
8	Pigeon	No	Yes	No	Yes	Non-mammals
9	Cat	Yes	No	No	Yes	Mammals
10	Leopard shark	Yes	No	Yes	No	Non-mammals
11	Turtle	No	No	sometimes	Yes	Non-mammals
12	Penguin	No	No	sometimes	Yes	Non-mammals
13	Porcupine	Yes	No	No	Yes	Mammals
14	Eel	No	No	Yes	No	Non-mammals
15	Salamander	No	No	sometimes	Yes	Non-mammals
16	Gila monster	No	No	No	Yes	Non-mammals
17	Platypus	No	No	No	Yes	Mammals
18	Owl	No	Yes	No	Yes	Non-mammals
19	Dolphin	Yes	No	Yes	No	Mammals
20	eagle	No	Yes	No	Yes	Non-mammals

WALCHAND COLLEGE OF ENGINEERING, SANGLI.

(An Autonomous Institute)

Final Year B.Tech. (Information Technology)

MAKEUP EXAMINATION SEM. I APRIL/MAY-2017

DATA MINING (2IT402)

MakeUp

Day, Date and Time: Saturday, 06/05/2017, 02.00pm to 05.00pm

Exam Seat Number: _____

Max Marks: **100**

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions: i) All questions are compulsory. Writing question number is compulsory. The answers may not be assessed if question number is not written.
ii) Figures to the right of question text indicate full marks.
iii) Assume suitable data wherever necessary.
iv) Write the answers with neat handwriting.

Text on the right of marks indicates course outcomes (only for faculty use).

			Marks	
Q1	A)	State major Tasks in Data Preprocessing.	4	CO1
Q1	B)	Define following terms for data warehouse? i) Star schema; ii) Snowflake schema; iii) Fact constellation schema.	4	CO1
Q1	C)	What is OLAP? State typical OLAP Operations.	4	CO3
Q1	D)	What is normalization? Why it is used? Use z-score normalization method to normalize the group of data: 18,22,25,42	4	CO2

Q2	A)	State the major types of concept hierarchy and explain rule-based hierarchy with example.	4	CO3
Q2	B)	Give approaches for the integration of a data mining system with a database or data warehouse system. State which approach you think is the most popular, and why.	4	CO3
Q2	C)	Give 5 point summary for following age data. Age: 23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 57, 58, 58, 60, 61.	5	CO2
Q2	D)	How we can use the information gain for identifying weakly or stromgly relevant attributes? Calculate information gain for 'Stream' attribute in following data.	6	CO2

Gender	Stream	Count
M	Art	16
F	Art	22
M	Science	18
F	Art	25
M	Art	21
F	Science	18

Target class: Passed in First class

Gender	Stream	Count
M	Art	16
F	Commerce	22
M	Commerce	18
F	Art	25
M	Science	21
F	Science	18

Contrasting class: Passed in Second class

Q3	A)	Give the classification of association rule mining based on different criteria	5	CO1
Q3	B)	State the methods to Improve Apriori's Efficiency	5	CO1
Q3	C)	find association rule with minconf=50% and minsup=40% for the sales data given below	6	CO2

Transaction ID	Item set
1	Milk, Bread, Jam
2	Bread, Butter, Juice
3	Soda, Bread, Butter
4	Bread, Juice, Soda
5	Milk, Juice

Q4	A)	State the criteria for comparing and evaluating classification and prediction methods.	4	CO1
Q4	B)	What is Bayesian belief network (BBN)? State the characteristics of BBN.	4	CO1

Q4 C) A training data is given as follows. Identify the attribute for best splitting at first step of the data set by using info gain and entropy. 8

Outlook	Temperature	Humidity	Windy	Class
Sunny	Hot	High	False	No-Play
Sunny	Hot	High	True	No-Play
Overcast	Hot	High	False	Play
Rain	Mild	High	False	Play
Rain	Cool	Normal	False	Play
Rain	Cool	Normal	True	No-Play
Overcast	Cool	Normal	True	Play
Sunny	Mild	High	False	No-Play
Sunny	Cool	Normal	False	Play
Rain	Mild	Normal	False	Play
Sunny	Mild	Normal	True	Play
Overcast	Mild	High	True	Play
Overcast	Hot	Normal	False	Play
Rain	Mild	High	True	No-Play

Q5 A) State the typical requirements of clustering in data mining. 4 C

Q5 B) State Major Clustering Approaches. 4 C

Q5 C) Describe each of the following clustering algorithms in terms of the following criteria: 5 C

Clustering algorithms	Shapes of clusters that can be determined	Input parameters that must be specified	Limitations
k-means			
k-medoids			
ROCK			
CHAMELEON			
DBSCAN			

Q5 D) Apply Single linkage agglomerative hierarchical clustering for following data of distances between cities. 5 CO

	BOS	NY	DC	CHI	SF	LA
BOS	0	206	429	963	3095	2979
NY	206	0	233	802	2934	2786
DC	429	233	0	671	2799	2631
CHI	963	802	671	0	2142	2054
SF	3095	2934	2799	2142	0	379
LA	2979	2786	2631	2054	379	0

Q6 A) Write a short note on - trend, cycle, seasonal and outlier in trend analysis. 5 CO3

Q6 B) Write a short note on - Subsequence Matching 5 CO3

Q6 C) Write a short note on - Types of Text Data Mining 5 CO3

WALCHAND COLLEGE OF ENGINEERING, SANGLI.

(An Autonomous Institute)

Final Year B.Tech. (Information Technology)

MAKEUP EXAMINATION: SEMESTER I MAY-2019

DATA MINING (3IT402)

4

Day, Date and Time: Saturday, 11/05/2019, 02.00pm to 05.00pm

Exam Seat Number: _____

Max Marks: **100**

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions: i) All questions are compulsory. Writing question number is compulsory. The answers may not be assessed if question number is not written. Assume suitable data wherever necessary.
ii) Figures to the right of question text indicate full marks.
iii) Mobile phones and programmable calculators are strictly prohibited.
iv) Except Exam Seat Number writing anything on question paper is not allowed.
Exchange/Sharing of stationery, calculator etc. not allowed.

Text on the right of marks indicates course outcomes (only for faculty use).

			Marks	
Q1	A)	State and define in short: major Tasks in Data Preprocessing.	5	CO1
Q1	B)	Differentiate in between OLAP and OLTP.	5	CO1
Q1	C)	Draw and explain a starnet query model with suitable data.	5	CO1

Q2	A)	What is 'concept hierarchy'? State the major types of it and explain rule-based hierarchy in detail with example	5	CO2
Q2	B)	In following data, calculate information gain for "department" attribute. How we can use the information gain for identifying weakly relevant attributes?	6	CO3

Gender	Department	Grade	Count
M	IT	B	16
F	IT	A	22
M	CSE	A	18
F	IT	A	25
M	IT	A	21
F	CSE	A	18

Gender	Department	Grade	Count
M	IT	B	16
F	ELN	C	22
M	ELN	C	18
F	IT	C	25
M	CSE	B	21
F	CSE	A	18

Q2	C)	What is a box and whisker plot? A sample of 10 boxes has these weights (in Kg): 25,28,29,29,30,34,35,35,37,38	6	CO3

Q3	A)	Explain Apriori algorithm and find association rules for following data set with minimum support=20%.	6	CO3

Tid	Items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E

Q3	B)	State the methods to Improve Apriori's Efficiency .	5	CO1
Q3	C)	In constraint based association rule mining, which constrains are used.	5	CO2

Q4	A)	State the criteria for comparing and evaluating classification and prediction methods.	4	CO1
Q4	B)	What is Bayesian belief network (BBN)? State the characteristics of BBN.	4	CO1

(P.T.O.)

Q4	C)	From given data ,identify the class of following test case by using Naïve Bayes Classifier. Test case - {Home loan Refund = No, Marital status= Married, Income=120K}	6	CO2

Marital status	Income (K)	Home loan refund	Defaulter
Single	125	Yes	No
Married	100	No	No
Single	70	No	No
Married	120	Yes	No
Divorced	95	No	Yes
Married	60	No	No
Divorced	220	Yes	No
Single	85	No	Yes
Married	75	No	No
Single	90	No	Yes

Q4	D)	From given data identify the class for test case using Naïve Bayes Classifier. {Have Legs= No, Give Birth= Yes, Can Fly= No, Live in water=Yes}	6	CO2

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

WALCHAND COLLEGE OF ENGINEERING, SANGLI.

(An Autonomous Institute)

ESE

Final Year B.Tech. (Information Technology)

END SEMESTER EXAMINATION: SEMESTER-I NOVEMBER-2018

DATA MINING (3IT402)

Day, Date and Time: Saturday, 24/11/2018, 10.00am to 12.00Noon

Exam Seat Number: _____

Max Marks: 50

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions: i) All questions are compulsory. Writing question number is compulsory. The answers may not be assessed if question number is not written. Assume suitable data wherever necessary.

ii) Figures to the right of question text indicate full marks.

iii) **Mobile phones and programmable calculators are strictly prohibited.**

iv) **Except Exam Seat Number writing anything on question paper is not allowed. Exchange/Sharing of stationery, calculator etc. not allowed.**

Text on the right of marks indicates course outcomes (only for faculty use).

		Marks	
Q1	Select appropriate option for following	4	CO1

1. Can decision trees be used for performing clustering?
   A. True; B. False
2. What is the minimum no. of variables/ features required to perform clustering?
   A. 0; B. 1; C. 2; D. 3
3. Which of the following can act as possible termination conditions in K-Means?
   1. For a fixed number of iterations.
   2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
   3. Centroids do not change between successive iterations.
   4. Terminate when RSS falls below a threshold.

   A. 1, 3 and 4 ; B. 1, 2 and 3; C. 1, 2 and 4 D. All of the above
4. Standardisation of features is required before training a Logistic Regression.
   A. TRUE; B. FALSE
5. Which of the following is relatively easier to estimate in time series modeling?
   A. Seasonality; B. Cyclical; C. No difference between Seasonality and Cyclical
6. Why do we prefer information gain over accuracy when splitting?
   A. Decision Tree is prone to overfit and accuracy doesn't help to generalize
   B. Information gain is more stable as compared to accuracy
   C. Information gain chooses more impactful features closer to root
   D. All of these
7. Which of the following are the disadvantage of Decision Tree algorithm?
   A. Decision tree is not easy to interpret; B. Decision tree is not a very stable algorithm
   C. Decision Tree will over fit the data easily if it perfectly memorizes it; D. Both B and C
8. What can be the maximum depth of decision tree (where k is the number of features and N is the number of samples)? Our constraint is that we are considering a binary decision tree with no duplicate rows in sample (Splitting criterion is not fixed).
   A. N; B. $N-k-1$; C. $N-1$; D. $k-1$

Q2 A)	State the criteria for comparing and evaluating classification and prediction methods.	3	CO1
Q2 B)	What is Bayesian belief network (BBN)? State the characteristics of BBN.	3	CO2

(P.T.O.)

Q2 C) Given all previous patient's symptoms data and diagnosis. Does the patient with following symptoms have flu? Use Naïve Bayes classifier. 5 CO3

Symptoms: Chills-Y; Runny nose- N; headache- Mild; fever-Y; Flu-?

Chills	Runny Nose	Headache	Fever	Flu
Y	N	Mild	Y	N
Y	Y	No	N	Y
Y	N	Strong	Y	Y
N	Y	Mild	Y	Y
N	N	No	N	N
N	Y	Strong	Y	Y
N	Y	Strong	N	N
Y	Y	Mild	Y	Y

Q2 D) From given data ,identify the class of following test case by using Naïve Bayes Classifier. 5 CO3

Test case - Married, 120, No

Marital status	Income (K)	Home loan refund	Defaulter
Single	125	Yes	No
Married	100	No	No
Single	70	No	No
Married	120	Yes	No
Divorced	95	No	Yes
Married	60	No	No
Divorced	220	Yes	No
Single	85	No	Yes
Married	75	No	No
Single	90	No	Yes

Q3 A) State the typical requirements of clustering in data mining. 2 CO1

Q3 B) Define outlier. State the approaches used for outlier detection. 2 CO1

Q3 C) Dissimilarities between all pairs of seven samples in given in following table. Perform agglomerative hierarchical clustering using complete linkage method and draw resulting dendogram 5 CO3

SAMPLES	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Q4 A) Give steps for performing a similarity search 2 CO1

Q4 B) Describe Web Mining Taxonomy with diagram 2 CO2

Q4 C) State some common operations on spatial databases 2 CO2

Q4 D) Draw Trend, Cycle, Seasonal Pattern and Trend with seasonal pattern for demand forecasting. 2 CO2

Q5	A)	Determine the regression equation by using the regression slope coefficient and intercept value as shown in the regression table given below.	4	CO3

X Values	Y Values
55	52
60	54
65	56
70	58
80	62

Q5	B)	State the basic principles of Attribute-Oriented Induction.	2	CO2
Q5	C)	State the methods to fill in the missing values for attributes in data mining process.	2	CO2
Q5	D)	Apply Apriori algorithm to find maximal frequent itemset from following data. Use minimum support count = 3.	5	CO3

Transaction ID	Items brought
1	Milk, Tea, Cake
2	Egg, Tea, Cold drink
3	Milk, Egg, Tea, Cold drink
4	Egg, Cold drink
5	Juice

WALCHAND COLLEGE OF ENGINEERING, SANGLI.
(An Autonomous Institute)
Final Year B.Tech. (Information Technology)
MAKEUP EXAMINATION APRIL/MAY-2018
DATA MINING (3IT402)

MakeUp

Day, Date and Time: Friday, 04/05/2018, 02.00pm to 05.00pm

Exam Seat Number: _____

Max Marks: **100**

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions: i) All questions are compulsory. Writing question number is compulsory. The answers may not be assessed if question number is not written. Assume suitable data wherever necessary.
ii) Figures to the right of question text indicate full marks.
iii) **Mobile phones and programmable calculators are strictly prohibited.**
iv) **Except Exam Seat Number writing anything on question paper is not allowed.**
Exchange/Sharing of stationery, calculator etc. not allowed.

Text on the right of marks indicates course outcomes (only for faculty use).

			Marks	
Q1	A)	State the stages/processes in data mining.	4	CO1
Q1	B)	What is data transformation? .	4	CO1
Q1	C)	State and explain major tasks in data preprocessing.	4	CO1
Q1	D)	What you mean by binning? State different binning methods and smooth following data by these methods. Data- 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34	4	CO2
Q1	E)	Consider the set of age values – 13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70. Transform the age value 35 by using Z-score normalization.	4	CO2

Q2	A)	Explain the various Interesting measures required in the discovery of patterns.	5	CO1
Q2	B)	Find mean, mode, median, first quartile, third quartile and draw box plot for following data. 13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35,35,35,36,40,45,46,52,70.	5	CO2
Q2	C)	In following data, calculate information gain for "department" attribute. How we can use the information gain for identifying weakly relevant attributes?	6	CO3

Gender	Department	Grade	Count
M	IT	B	16
F	IT	A	22
M	CSE	A	18
F	IT	A	25
M	IT	A	21
F	CSE	A	18

Target class: Campus Recruited

Gender	Department	Grade	Count
M	IT	B	16
F	ELN	C	22
M	ELN	C	18
F	IT	C	25
M	CSE	B	21
F	CSE	A	18

Contrasting class: Not Recruited

Q3	A)	Give the classification of association rule mining based on different criteria.	3	CO1
Q3	B)	How does ARCS (Association Rules Clustering System) work? Explain the steps in ARCS. Give the limitations of ARCS.	5	CO2
Q3	C)	For following transactional data draw FP-tree with minimum support count=2.	6	CO3

TID	1	2	3	4	5	6	7	8	9
List of Items	I1, I2, I5	I2, I4	I2, I3	I1, I2, I4	I1, I3	I2, I3	I1, I3	I1, I2 ,I3, I5	I1, I2, I3

(P.T.O.)

Q4	A)	State the advantages and disadvantages of Decision Tree classification	4	C
Q4	B)	A training data is given as follows. Identify the attribute for best splitting at first step of the data set by using info gain and entropy.	7	C

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	NoPlay
sunny	hot	high	true	NoPlay
overcast	hot	high	false	Play
rain	mild	high	false	Play
rain	cool	normal	false	Play
rain	cool	normal	true	NoPlay
overcast	cool	normal	true	Play
sunny	mild	normal	false	NoPlay
sunny	cool	high	false	Play
rain	mild	normal	false	Play
sunny	mild	normal	true	Play
overcast	mild	normal	true	Play
overcast	hot	high	false	Play
rain	mild	normal	true	NoPlay

Q4	C)	From given data, identify the class of following test case by using Naïve Bayes Classifier. Test case – $X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$	7	CO

Marital status	Income (K)	Home loan refund	Defaulter
Single	125	Yes	No
Married	100	No	No
Single	70	No	No
Married	120	Yes	No
Divorced	95	No	Yes
Married	60	No	No
Divorced	220	Yes	No
Single	85	No	Yes
Married	75	No	No
Single	90	No	Yes

Q5	A)	State the typical requirements of clustering in data mining.	4	CO1
Q5	B)	Explain Hierarchical Clustering method in detail. What is Dendogram?	4	CO1
Q5	C)	Define outlier. Explain the approaches of outlier detection.	4	CO3
Q5	D)	Give steps for K-means clustering algorithm. State Strength and Weakness.	4	CO2

Q6	A)	Short note on - Precision and recall for Text Retrieval.	4	CO3
Q6	B)	Short note on - Spatial Classification and Spatial Trend Analysis.	4	CO3
Q6	C)	Short note on - Similarity search in image data.	4	CO3
Q6	D)	Short note on - Steps for performing a similarity search.	4	CO3

WALCHAND COLLEGE OF ENGINEERING, SANGLI.

(An Autonomous Institute)

ESE

Final Year B.Tech. (Information Technology)

END SEMESTER EXAMINATION SEM. I NOVEMBER-2017

DATA MINING (3IT402)

Exam Seat Number: _____

Day, Date and Time: Wednesday, 22/11/2017, 03.00pm to 05.00pm

Max Marks: **50**

IMP: Verify that you have received question paper with correct course, code, branch etc.

Instructions: i) All questions are compulsory. The answers may not be assessed if question number is not written.
ii) Attempt all questions in ORDER.
iii) Figures to the right of question text indicate full marks.
iv) Assume suitable data wherever necessary, Write the answers with neat handwriting.
v) Only FX82 series non programmable Calculator is allowed.

Text on the right of marks indicates course outcomes (only for faculty use).

			Marks	
Q1	A)	State ***True*** or ***False***. If '***False***' give correct statement. 1. Brute-force approach in association rule mining is $R = 3^m + 2^{m+1} - 1$ 2. In Multidimensional association rule analysis; quantitative attributes have implicit ordering among numeric value. 3. FP tree reduces memory consumption cost. 4. Data generalization is an essential operation in attribute oriented induction 5. Snowflake schema: A fact table in the middle connected to a set of dimension tables. 6. OLAP is used for day to day operations. 7. In DBSCAN; MinPts must be ≥ 5 8. K-medoid algorithm uses imaginary cluster center.	4	CO3
Q1	B)	Fill in the blanks. 1. In _____ histogram, the width of each bucket range is uniform. *a. Equiwidth; b. Equidepth; c. V-optimal; d. MaxDiff* 2. Which is not correct for STING: statistical information grid clustering approach. *a. Uses Multi-resolution approach; b. Stores statistical info in each cell* *c. Uses bottom-up approach; d.Each layer stores confidence interval for each cell* 3. The data are consolidated into forms appropriate for mining is called *a. Data reduction; b. Data redundancy; c. Data cleaning; d. Data transformation* 4. In _____ plot; graphs the quantiles of 1 univariate distribution against the corresponding quantiles of other *a. Quantile; b. Quantile- quantile; c. Loess curve; d. Both a and b* 5. Apriori algorithm uses _____ *a. Upward closure property; b. Downward closure property; c. None of all above* 6. The best splitter is determined as the attribute which has _____ gini value *a. Smallest; b. Greatest; c. Close to 1* 7. For binary split of numerical and categorical attributes we use _____ and _____ respectively *a. Class histogram and count matrix; b. Count matrix and class histogram* *c. Class matrix and histogram count; d. Histogram count and class matrix* 8. In most of partitioning clustering methods; objects are based on _____ *a. No of clusters; b. No of objects in each class;* *c. Distance between objects; d. Distance between clusters*	4	CO3

Q2	Answer following in brief. A) What are the stages/processes of Data Mining. Write them in order. B) State 3-4-5 rule. C) Assume suitable FP-tree and show conditional pattern for an item. D) State possible ways of integrating or coupling a data mining system with database/datawarehouse.	8	CO3

Q3	A)	What are the criterias to compare and evaluate classification and prediction methods.	2	CO2
Q3	B)	What are the major difficulties which arise while constructing a decision tree?	2	CO2

(P.T.O.)

Q3	C)	In the table below, the Xi column shows scores on the aptitude test and Yi column shows statistics grades. If a student made an 80 on the aptitude test, what statistics grade would be? Use Linear Regression.	3	CO1

Student	x_i	y_i
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

Q3	D)	From given data identify the class for test case using Naïve Bayes Classifier Test case- Have legs: NO, Give birth: YES; Can fly: YES; Live in water: YES.	5	CO1

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Q4	A)	Define Eps, MinPts, density-reachable and density-connected and give the steps to carryout DBSCAN algorithm.	5	CO2
Q4	B)	State various types of constrents for constraint based clustering.	3	CO2
Q4	C)	Carryout Aglomerative hierarchical clustering method to cluster following data. Use Complete Linkage approach. Draw resulting dendrogram.	5	CO1

Dist	A	B	C	D	E	F
A	0	662	877	255	412	996
B	662	0	295	468	268	400
C	877	295	0	754	564	138
D	255	468	754	0	219	869
E	412	268	564	219	0	669
F	996	400	138	869	669	0

Q5		Write a short note on ***Any Three*** **with diagram.** A) Trend, cycle, seasonal in Time series data mining. B) Multidimensional view of database. C) Web mining taxonomy. D) Local and global trend in Spatial data mining.	9	CO2