EXAM NO.______________

GANPAT UNIVERSITY

M.TECH SEM: II (COMPUTER ENGINEERING/INFORMATION TECHNOLOGY)

REGULAR EXAMINATION APRIL-JUNE 2017

3CE202/3IT202: DATA MINING & DATA WAREHOUSING

MAX. TIME: 3 HRS **MAX. MARKS: 60**

Instructions: (1) This Question paper has two sections. Attempt each section in separate answer book.
(2) Figures on right indicate full marks.
(3) Be precise and to the point in answering the descriptive questions.
(4) Assume data, if necessary.

SECTION – I

Q-1. **[A]** Explain KDD process with example. **[4]**

[B] What is data transformation? Explain the techniques for data transformation in brief. Given one-dimensional data set X = {-5.0, 23.0, 17.6, 7.23, 1.11}, normalize the data set using min-max normalization on interval [0, 1]. **[6]**

OR

Q-1. **[A]** Explain architecture of data mining system. **[3]**

[B] What is noisy data? Apply applicable two smoothing techniques to handle noise in a given attribute values 13, 15, 15, 16, 19, 20, 20, 21, 22, 22, 25, 25, 26, 27, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46. **[3]**

[C] Discuss the importance of Correlation Analysis. According to the survey done in a company of 2000 employees, following readings on individuals’ interest were observed. State the correlation between reader & writer using chi-square analysis for the given data: **[4]**

	Reader	Not a reader
Writer	800	700
Not a writer	400	100

Q-2. **[A]** Differentiate OLAP and OLTP. **[4]**

[B] Define: Lattice of cuboid. The measures are displayed about sales data (number of units sold). Draw 4-D data cube of sales data. Apply any four types of OLAP operations on it. **[6]**

Location= “Mehsana”					Location=”Ahmedabad”				
	Item					Item			
Time	Comp.	CPU	CD	Fax	Time	Comp.	CPU	CD	Fax
T1	201	132	123	233	T1	238	189	143	239
T2	520	124	435	142	T2	508	193	434	542
T3	234	322	532	144	T3	238	392	532	544
T4	433	433	534	233	T4	439	493	544	223

Location= “Bhuj”					Location=”Gandhidham”				
	Item					Item			
Time	Comp.	CPU	CD	Fax	Time	Comp.	CPU	CD	Fax
T1	224	135	163	273	T1	270	137	173	237
T2	250	143	474	172	T2	505	125	437	147
T3	215	336	562	174	T3	237	327	572	147
T4	243	233	574	273	T4	437	437	537	273

OR

Q-2. **[A]** Which methods are used to improve the efficiency of apriori algorithm? Explain it in brief. Apply any one method on dataset given in Table 1. [5]

[B] Define: Sequential frequent pattern. Given minimum support as 30% and minimum confidence as 80%. Find out strong Association Rules from given dataset using Apriori algorithm for given Table 1. [5]

Table 1	
T_ID	Item_Name
I1	Bread, Butter, Cake, Pastry, Jam, Milk
I2	Milk, Butter, Noodles, Jam
I3	Jam, Tea, Noodles, Paneer, Bread
I4	Noodles, Paneer, Butter, Milk
I5	Tea, Paneer, Jam, Chocolate, Noodles
I6	Cake, Pastry, Paneer, Butter, Bread, Milk
I7	Milk Powder, Bread, Tea, Noodles, Milk
I8	Chocolate, Paneer, Bread
I9	Cheese, Jam, Chocolate, Butter, Bread
I10	Paneer. Bread, Chocolate, Noodles, Butter

Q-3. **[A]** Explain different types of schemas for multidimensional database with example. [4]

[B] Define following terms with suitable example: [6]

(i) Data characterization

(ii) Data discrimination

(iii) Multilevel association rule

SECTION – II

Q-4. **[A]** Find SMC coefficient and rao's coefficient for given 6-dimensional categorical samples P = (A, B, A, B, A, A) and Q = (B, B, A, B, B, A). [4]

[B] Discuss every terms of DBSCAN clustering method with suitable example. [6]

OR

Q-4. **[A]** What is rule based classification? How to check the percentage of correctly classified tuples? Explain it by suitable example. [4]

[B] Given the samples X1 = {1, 0}, X2 = {0, 1}, X3 = {2, 1}, and X4 = {3, 3}. Suppose that samples are randomly clustered into two clusters C1 = {X1, X3} and C2 = {X2, X4}. Apply k-means partitioning algorithm up to 2nd iteration and discuss what changes occurs in 1st and 2nd iteration in clusters. [6]

Q-5. **[A]** Which are typical requirements of clustering in data mining? [3]

[B] Train the given dataset using ID3 algorithm and generate decision tree. [7]

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

OR

Q-5. **[A]** Find the dissimilarity between person A and B for given binary variables. [3]

Person	Gender	Fever	Fatigue	Headache	Test-1	Test-2	Test-3	Test-4
A	M	Y	N	N	N	P	P	N
B	F	Y	Y	Y	P	N	P	N

[B] Define: Prediction. Explain baye's theorem. [7]

Age	Income	Student	Credit_Rating	Class: Buys Laptop
>30	Medium	No	Excellent	No
<=20	High	No	Fair	No
21..30	High	Yes	Fair	Yes
<=20	High	No	Excellent	No
21..30	Medium	No	Excellent	Yes
21..30	High	No	Fair	Yes
<=20	Medium	Yes	Excellent	Yes
>30	Medium	No	Fair	Yes
>30	Medium	Yes	Fair	Yes
>30	Low	Yes	Fair	Yes
<=20	Low	Yes	Fair	Yes
>30	Low	Yes	Excellent	No
21..30	Low	Yes	Excellent	Yes
<=20	Medium	No	Fair	No

Predict a class label of an unknown tuple X= {Age ='<=20' , Income= 'Medium', Student='Yes', Credit_Rating='Fair'}

Q-6. [A] Given a set of 5-dimensional categorical samples: A = (1, 0, 1, 1, 0), B = (1, 1, 0, 1, 0), C = (0, 0, 1, 1, 0), D = (0, 1, 0, 1, 0), E = (1, 0, 1, 0, 1), F = (0, 1, 1, 0, 0). [6]

Use similarity measure based on jaccard's coefficient. Apply agglomerative hierarchical clustering and plot the dendrogram.

[B] Explain web mining briefly. [4]

----------END OF PAPER----------