Table Q.2B

Student Id *(Identifier)*	Subject Name *(Categorial)*	Grade Of Student *(Ordinal)*	Marks *(ratio_scaled)*
1	DBMS	excellent	76
2	DAA	fair	28
3	DBMS	Good	49
4	DWDM	fair	93
5	DAA	excellent	67

2C. Define the following terms with respect to DBSCAN algorithm by giving an example for each. 3
(i) Core object (iii) Directly density reachable object
(ii) Density reachable object (iv) Density connected object 2

3A. Find all frequent itemsets by using the Pincer-Search algorithm for the transactions given in Table Q.3A assuming minimum support >=2

Table Q.3A

TID	ITEMS
1	A,B,C,D,E,F
2	A,B,C,G
3	A,B,D,H
4	B,C,D,E,K
5	A,B,C

5

3B. Apply the page rank algorithm by constructing a graph based on the following information: A→B, A→D, A→C, B→C, B→A, C→D, D→C, D→A. Assume the initial page rank as 1 and perform 2 iterations. 3

3C. Explain the concept of web content mining. 2

4A. Construct a decision tree for the training data set given in Table Q.4A by using information gain to decide the splitting attributes.

Table Q. 4A

Tid	Refund	Marital Status	Cheat
1	Yes	Single	No
2	No	Married	No
3	No	Single	Yes
4	Yes	Married	No
5	No	Divorced	Yes
6	No	Married	No
7	Yes	Divorced	No
8	No	Single	Yes
9	No	Married	No
10	No	Single	Yes

5

4B. What is the advantage of gain ratio over information gain? Explain with an example. 3

4C. Briefly describe the application of data mining in retail industry. 2

5A. Define temporal mining. Apply the temporal apriori algorithm for the data given in Table Q.5A by considering the minimum support as 3 and the temporal support as 2.

Table Q.5A

TID	ITEMS
1	1,3,4
2	2,3,5
3	1,2,3,5
4	2,5

5

5B. Explain the termination conditions for the recursive partitioning of Decision Tree induction algorithm. 3

5C. Briefly describe any 4 types of temporal data. 2

6A. Write the apriori algorithm. Given a database of four transactions (minimum support >= 2): T1={A,B,C,D,E}, T2={A,B,D,J}, T3={B,F,K,S}, T4={D,G,H,P}, show the major steps to find the frequent patterns using Apriori algorithm. 5

6B. Table Q.6B shows the percentage of boys and girls who got into trouble in school. Use chi-squared statistics to determine whether getting into trouble is correlated with gender assuming the chi-squared critical value as 2.71 at the significance level of 0.10.

Table Q.6B

	Got in trouble	No trouble
Boys	64	17
Girls	73	38

3

6C. Draw a neat diagram depicting data mining as a step in the process of knowledge discovery. 2

Reg. No. | | | | | | | | | | |

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

A Constituent Institution of Manipal University

VI SEMESTER B.TECH. (COMPUTER AND COMMUNICATION ENGINEERING) END SEMESTER EXAMINATIONS, APRIL/MAY 2017

SUBJECT: DATA MINING AND PREDICTIVE ANALYSIS [ICT 354]

REVISED CREDIT SYSTEM
(25/04/2017)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer ALL questions.
- Write the detailed steps for all the problems.
- Missing data, if any, may be suitably assumed.

1A. A hospital tested the age and body-fat data for 18 randomly selected adults with the result as given in Table Q.1A. Calculate the correlation coefficient (Pearson's product moment coefficient) to determine whether these two variables are positively or negatively correlated. 5

Table Q.1A

Age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
Age	52	54	54	56	57	58	58	60	61
% fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

1B. Construct a Pattern Count tree for the transaction database given in Table Q.1B. 3

Table Q.1B

Transaction ID	Items purchased
1	1,5,6,8
2	2,4,8
3	4,5,7
4	2,3
5	5,6,7
6	2,3,4
7	2,6,7,9
8	5

Transaction ID	Items purchased
9	8
10	3,5,7
11	3,5,7
12	5,6,8
13	2,4,6,7
14	1,3,5,7
15	2,3,9

1C. Explain the difference and similarity between discrimination and classification. 2

2A. Cluster the following points into 2 clusters by using k-means method with manhattan as a distance measure. X1(1,3), X2(3,4), X3(3,6), X4(3,8), X5(4,5), X6(4,7), X7(5,1), X8(5,5), X9(7,3), X10(7,5), X11(8,5), X12(9,4). Assume X1 and X12 as the initial cluster centers. 5

2B. Consider the sample data givn in Table Q.2B. Apply the euclidian distance to find the dissimilarity matrix for categorical, ordinal and ratio_scaled variables individually.