Reg. No.										

MANIPAL INSTITUTE OF TECHNOLOGY

MANIPAL

(A constituent unit of MAHE, Manipal)

VI SEMESTER B.TECH. (COMPUTER AND COMMUNICATION ENGINEERING)

MAKE-UP EXAMINATIONS, JUNE 2019

SUBJECT: DATA MINING AND PREDICTIVE ANALYSIS [ICT 3252]

REVISED CREDIT SYSTEM

(19/06/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

- Answer **ALL** the questions.
- Missing data if any, may be suitably assumed.

1A. i. Write the pseudo code for Partition association rule mining algorithm. **5**

ii. Find the frequent patterns for the set of transactions using Partition association algorithm. Assume minimum support count as 33% and number of partitions =3.

T1: {1,3,5,6} T2: {2,4,7} T3: {1,4,5} T4: {1,2,3,5} T5: {1,3,5} T6: {1,2,3,4}

1B. Find the Pearson's product moment coefficient for the data given in Table Q.1B. **3**

Table Q.1B.

Hemoglobin (hb) level	Packed cell Volumes(PCV)
15.5	0.450
13.6	0.420
13.5	0.440
13.0	0.395
13.3	0.395
12.4	0.370
11.1	0.390
13.1	0.400
16.1	0.445
16.4	0.470
13.4	0.390
13.2	0.400
14.3	0.420
16.1	0.450

1C. Explain any one algorithm used to model web topology. **2**

2A. Construct a Frequent Pattern Tree and generate the frequent patterns from the following set of transactions assuming the minimum support count as 3. **5**

T1:{A,B,D,E,F} T2: {B,D} T3:{A,C,E} T4:{A,B,C,D} T5:{D,E,F} T6:{A,E,F}

2B. Explain the GSP algorithm used for solving sequence mining problem. **3**

2C. Referring to question Q.1A., generate association rules and find strong association rules from the transaction data assuming confidence =60%. **2**

3A. i. How k-medoids algorithm is better than k-means algorithm? Explain. **5**

ii. Consider the dataset given in Table Q.3A. Let O3(3, 8) and O5(7, 4) be the initial cluster medoids. Apply k-medoid clustering algorithm and check whether swapping the centroid from O5 to O6 would result in better clustering?

Table Q.3A.

Point	X axis	Y axis
O1	6	7
O2	6	2
O3	8	3
O4	5	8
O5	4	7
O6	7	4
O7	2	6
O8	3	7

3B. i. Mention the various ways in which correlation between attributes can be obtained. **3**

ii. A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency Table Q.3B. Is there a gender gap? Do the men's voting preferences differ significantly from the women's preferences? The $\chi 2$ value at 0.005 significance level is 10.597.

Table Q.3B.

	Rep	Dem	Ind
Male	200	150	50
Female	250	300	50

3C. Consider the contingency table given in Table Q.3C. Find the pattern interestingness measures using the following metrics: **2**

i. Kulc ii. Cosine

Table Q.3C.

	Milk	$\overline{Milk}$
coffee	10000	1000
$\overline{coffee}$	1000	100000

4A. i. Consider the data given in Table Q.4A. and compute the root node using Information Gain as the attribute selection measure. **5**

ii. What is the drawback of using Information Gain as Attribute selection measure in Decision Tree Induction algorithm?

Table Q.4A.

Eggs	Pouch	Flies	Feathers	Class
Yes	No	Yes	Yes	Bird
No	No	No	No	Mammal
Yes	Yes	No	No	Marsupial
Yes	No	No	Yes	Bird
No	Yes	No	No	Marsupial
No	Yes	No	No	Marsupial
Yes	No	Yes	Yes	Bird
Yes	No	Yes	Yes	Bird
Yes	No	No	Yes	Bird
Yes	No	No	No	Mammal
No	Yes	No	No	Marsupial
No	Yes	No	No	Marsupial

4B. Write the pseudo code for Pincer-Search association rule mining. **3**

4C. With the help of suitable diagram, explain knowledge discovery process. **2**

5A. Find the frequent itemsets using dynamic itemset counting algorithm for the dataset given in Table Q.5A. Assume the minimum support count as 2 and number of partitions as 2. **5**

Table Q.5A.

Transaction Id	Items
T1	1, 5,7,9
T2	2,3,4,5
T3	1,5,6,9
T4	2,3
T5	1,9
T6	2,3,5,9

5B. Compute the clusters using Agglomerative hierarchical clustering for the data given in Table Q.5B. Also, draw the dendrogram. **3**

Table Q.5B.

Object id	X1	X2
A	1	1
B	1	0
C	0	2
D	2	4
E	3	5

5C. Explain various issues in data mining. **2**