Reg. No. □□□□□□□□□

MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
(A constituent unit of MAHE, Manipal)

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY)

END SEMESTER EXAMINATIONS, APRIL/MAY 2019

DATA WAREHOUSING AND DATA MINING [ICT 3202]

REVISED CREDIT SYSTEM
(27/04/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. Explain the following techniques for performing data reduction:
 i. Attribute subset selection
 ii. Histograms **5**

1B. The set of data from a sample of 11 items is given in Table Q.1B.

Table Q.1B

X	11	3	17	10	14	6	13	12	7	4	16
Y	22	6	34	20	28	12	26	24	14	8	32

By considering,

$$Z_1 = X - Y$$
$$Z_2 = 2X + Y$$

Calculate the covariance and correlation between Z_1 and Z_2 **3**

1C. State any four different types of attributes. Give an example for each. **2**

2A. Give the definition of data warehousing. With a schematic diagram, explain the working of general data warehousing architecture. **5**

2B. For the transaction data set given in Table Q.2B, construct the FP – tree by considering the reversed ordering scheme.

Table Q.2B

TID	Items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{a, c, e}

min-sup = 2

3

2C. What is ETL? Briefly explain the major steps involved in ETL process. **2**

3A. Given a data set = { T1:A, B, C, D, E, F; T2:A, B, C, G; T3:A, B, D, H; T4: B, C, D, E, I; T5:A, B, C; T6:D, E, F, I }. Find the maximal frequent set using Pincer-Search algorithm by considering the support count as 2. Show all the steps. **5**

3B. What is data mining? Explain the process of knowledge discovery in databases (KDD) with a neat diagram. **3**

3C. What is data cleaning? How are missing values handled in the preprocessing stage? 2

4A. Find the root node using Gini Index as the attribute selection measure for the data given in Table Q.4A.

Table Q.4A

ID	Age	Has_Job	Credit_Rating	Class
1	Young	False	Fair	No
2	Young	False	Good	No
3	Young	True	Good	Yes
4	Young	True	Fair	Yes
5	Young	False	Fair	No
6	Middle	False	Fair	No
7	Middle	False	Good	No
8	Middle	True	Good	Yes
9	Middle	False	Excellent	Yes
10	Middle	False	Excellent	Yes
11	Young	False	Excellent	Yes
12	Young	False	Good	Yes
13	Old	True	Good	Yes
14	Old	True	Excellent	Yes
15	Old	False	Fair	No

5

4B. Consider a transactional dataset in which the item 'Pen' occurs in 3000 transactions, item 'Eraser' occurs in 2000 transactions and both the items together are present in 1700 transactions. There are 700 transactions in which the items 'Pen' and 'Eraser' are not purchased. Calculate the following:

i. Lift ii. All_confidence iii. max_confidence iv. Cosine

Which of the above measures are not null-invariant? Justify. 3

4C. Explain the following with example.

i. Intradimensional association rule

ii. Interdimensional association rule

iii. Hybrid – dimensional association rule 2

5A. Apply the Partitioning Around medoids algorithm on the below mentioned data points and obtain two clusters. Let B and H be the initial cluster medoids.

A(2, 10), B(2, 5), C(8, 4), D(5, 8), E(7, 5), F(6, 4), G(1, 2), H(4, 9)

Verify whether swapping the centroid from H to E would result in better clustering? 5

5B. Find the dissimilarity matrix for the dataset given in Table Q.5B.

Table Q.5B

Object id	Band (Nominal)	Position (Ordinal)	Salary (Numeric)
1	Red	Senior	50000
2	Green	Junior	12000
3	Blue	Mid	30000
4	Green	Senior	45000

3

5C. Explain the following terms with an example.

i. Entropy

ii. Gain Ratio 2