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

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

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY)

MAKEUP EXAMINATIONS, JUNE 2019

DATA WAREHOUSING AND DATA MINING [ICT 3202]

REVISED CREDIT SYSTEM

(12/06/2019)

Time: 3 Hours MAX. MARKS: 50

Instructions to Candidates:

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

1A. With a suitable example, explain the star schema for multidimensional database. **5**

1B. Explain the difference between MOLAP and ROLAP. **3**

1C. Give the difference between OLTP and data warehouse systems. **2**

2A. A media streaming website knows that 70% of its customers primarily watch on their television, 18% primarily watch on their computer, and 12% primarily watch on a mobile device. The company wonders if these percentages hold true after a recent update to the product. They take a random sample of 700 customers and obtain the results as given in Table Q.2A.

Table Q.2A

Device	Television	Computer	Mobile
Expected	70%	18%	12%
# of customers	401	197	102

They want to perform a χ^2 goodness-of-fit test to determine if these results suggest that the distribution has changed. What is the expected count of customers that watch on their computer in the sample? **5**

2B. What is data preprocessing? Explain the techniques for performing data smoothing. **3**

2C. Explain the methods for the generation of concept hierarchies for nominal data. **2**

3A. Write the pseudo code for Pincer Search algorithm. **5**

3B. Apply FP-Growth algorithm and find all the frequent itemset for the data given in Table Q.3B considering support threshold as 25%. Show all the steps.

Table Q.3B

TID	Items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

3

3C. Illustrate the advantages of using closed frequent itemset with an example. **2**

4A. Given initial seeds as X1 and X4, obtain clusters for the given dataset by applying k-means algorithm. Dataset = { X1(2,10); X2(2,5); X3(8,4); X4(9,4); X5(5,8); X6(1,2); X7(4,9) }
Also, check whether swapping the initial seeds to X2 and X5 would result in a better clustering. **5**

4B. Use Dynamic Itemset Counting to discover the frequent itemsets from the transactions below with M = 2, support threshold s=2 and confidence threshold c=60%. Show all the updates done in each database scan. The set of transactions are
T1{P, Q, S, T } , T2{P, Q, R, S, T}, T3{P, Q, R, T}, T4{P, Q, S} **3**

4C. The Probability of playing both cricket and football is 40%. The probability of playing football is 50%. There exists positive correlation between cricket and football. The Correlation measure, Lift between cricket and football is 2. Find the dependent/correlation measures all_confidence and cosine. **2**

5A. Find the root node using Information Gain as the attribute selection measure for the data given in Table Q.5A. What is the drawback of Information gain?

Table Q.5A

ID	Weather	Weekend_Job	Status	Class
1	Sunny	Yes	Rich	Cinema
2	Sunny	No	Rich	Tennis
3	Windy	Yes	Rich	Cinema
4	Rainy	Yes	Poor	Cinema
5	Rainy	No	Rich	Tennis
6	Rainy	Yes	Poor	Cinema
7	Windy	No	Poor	Cinema
8	Windy	No	Rich	Tennis
9	Windy	Yes	Rich	Cinema
10	Sunny	No	Rich	Tennis

5

5B. Write the DBSCAN algorithm. What are its advantages and disadvantages? **3**

5C. Discuss the problems faced by today's search tools in finding relevant information on the web. **2**