Reg. No.										

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

VI SEMESTER B.TECH. (INFORMATION TECHNOLOGY) GRADE IMPROVEMENT / MAKEUP EXAMINATIONS, AUGUST 2021

SUBJECT: DATA WAREHOUSEING & DATA MINING [ICT 3253]
REVISED CREDIT SYSTEM
(07/08/2021)

Time: 2 Hours MAX. MARKS: 40

Instructions to Candidates:
- Answer **ANY FOUR FULL** questions.
- Write the detailed steps for all the problems/algorithms.
- Missing data, if any, may be suitably assumed.

1A. Outline the possible major research challenges of multimedia data mining? How is it different from text mining? **4M**

1B. With an example, compare and contrast Data Matrix and Dissimilarity Matrix. Find the dissimilarity and Jaccard coefficient by using the Table Q.1B that has details about three students.

Table Q.1B

Name	Gender	Distinction in XII?	NRI Quota?	Knows French?	Covid vaccinated?	Place:Manipal?	Nerd?
Vansh	M	Yes	No	Yes	No	No	No
Yash	M	Yes	No	Yes	No	Yes	No
Rohan	M	Yes	Yes	No	No	No	No

6M

2A. Perform the following for the data: (22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40).

(i) smooth the data by applying "smooth by bin means" by considering bin depth as 5
(ii) normalize the data by using "min-max normalization" by setting $min = 0$ and $max = 1$
(iii) normalize the data by using "decimal scaling" method. **6M**

2B. Explain all the OLAP operations and give one example for each. **4M**

3A. Find the frequent pattern for the transactional database: T1={18, 40, 510, 527}, T2={18, 40, 179}, T3={18, 40, 510, 125}, T4={527, 740}, T5={527, 740, 795}, T6={18} by using FP-tree algorithm with minimum support count>1. Show detailed steps. **5M**

3B. Find the frequent pattern for the transactional database: T1={A, B, E, F}, T2={ A, B, E, C }, T3={ A, B, D }, T4={F, G, H}, T5={F, G}, T6={B} by using apriori algorithm with a minimum support count>1. Show detailed steps. **5M**

4A. Explain the following and give one example for each.
(i) Closed pattern
(ii) Max pattern
(iii) Frequent pattern **3M**

4B. Construct a decision tree which predicts whether a patient could get heart attack or not for the data given in Table Q.4B

Table Q.4B

Patient ID	Chest Pain	Male	Smokes?	Exercises?	Heart attack?
1	Yes	Yes	No	Yes	Yes
2	Yes	Yes	Yes	No	Yes
3	No	No	Yes	No	Yes
4	No	Yes	No	Yes	No
5	Yes	No	Yes	Yes	Yes
6	No	Yes	Yes	Yes	No

7M

5A. The contingency matrix as given in Table Q.5A shows information and reviews of 500 movies by 2 independent annotators.

Table Q.5A

		Annotator B	
		Positive	Negative
Annotator A	Positive	25	25
	Negative	100	350

a) Find the accuracy of the reviews
b) For this scenario, is accuracy the best suited performance evaluation metric? Justify.
c) If the training data is increased, will it detoriate the performance of the model? Justify.
d) If we include only those features which are highly correlated and thereby reduce the feature representation, would it improve the performance of the model? Justify. **4M**

5B. Cluster the following eight points into 3 clusters by k-medoid method: *A*1(2, 10)*; A*2(2, 5)*; A*3(8, 4)*; B*1(5,8)*; B*2(7, 5)*; B*3(6, 4)*; C*1(1, 2)*; C*2(4, 9). Consider A1, B1, C1 as the initial set of medoids. Use the Manhattan distance method to find the distance. Determine the total cost of clustering after swapping A1 with A2. Identify various 4 cases to which each point belongs to after swapping A1 with A2. **6M**

6A Discuss the various Semi-Supervised Methods of outlier detection. **5M**

6B Write a neat block diagram depicting a general information retrieval architecture and explain. **5M**