Question Paper

Exam Date & Time: 14-May-2022 (10:00 AM - 01:00 PM)

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

SIXTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, MAY 2022

DATA WAREHOUSING AND DATA MINING [ICT 3253]

Marks: 50 **Duration: 180 mins.**

A

Answer all the questions. Section Duration: 180 mins

Instructions to Candidates:
Answer ALL questions
Missing data may be suitably assumed

1)

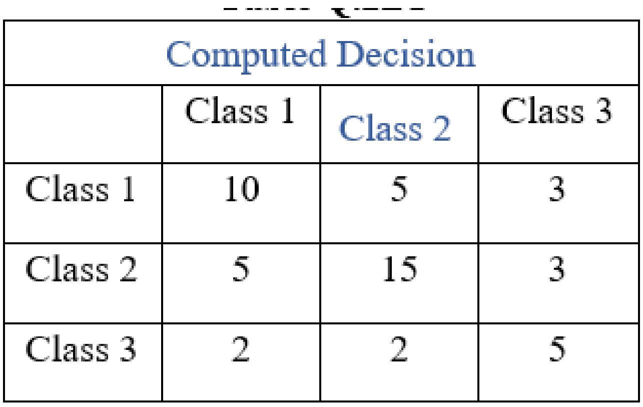
A) Consider a franchise of retail stores having the business setup only in India. The analysis requirements of the franchise include getting to know which items are purchased together by each individual consumer. They wish to know the sales figures in terms of sales amount in Rupees as well as quantity of the individual stores and also for the city, state and region in which they are located. They also wish to know how sales differ over different months, quarters and years; how sales figures change with the hour of the day - e.g., how sales of morning hours are different from sales of evening hours, etc.; how buying habits of male consumers are different from that of the female consumers; how buying habits of married consumers are different from that of the unmarried consumers; how buying habits of consumers vary with their native languages (e.g., Kannada, Telugu, Marathi, etc.). (5)

i. Design a star schema for such a data warehouse clearly identifying the fact table and dimension tables, their primary keys, and foreign keys. Also, mention which columns in the fact table represent dimensions and which ones represent measures or facts.

ii. Is there any downside with respect to star schema you have designed? How can you overcome it?

B) Use the three-class confusion matrix in Table Q.1B1 to answer questions i) through iii). Use confusion matrix for Model X and confusion matrix for Model Y in Table Q.1B2 to answer questions iv) through vi). (3)

Table Q.1B1



	Computed Decision		
	Class 1	Class 2	Class 3
Class 1	10	5	3
Class 2	5	15	3
Class 3	2	2	5

Table Q.1B2

Model X	Computed Accept	Computed Reject	**Model Y**	Computed Accept	Computed Reject
Accept	10	5	Accept	6	9
Reject	25	60	Reject	15	70

i. What percent of the instances were correctly classified?
ii. How many *class 2* instances are in the dataset?

ii. How many *class 2* instances are in the dataset?

iii. How many instances were incorrectly classified with*class 3*?

iv. How many instances were classified as an accept by Model X?

v. Compute the lift for Model Y.

vi. You will notice that the lift for both models is the same. Assume that the cost of a false reject is significantly higher than the cost of a false accept. Which model is the better choice?

C) What is the category of web mining when itis used to predict user behaviour? Explain. (2)

2) A) With an example, compare and contrast Data Matrix and Dissimilarity Matrix. Suppose that the table Q.2A records details about three students. Find the dissimilarity and Jaccard coefficient between the attributes. (5)

Table Q.2A

Name	Gender	Distinction in XII?	NRI Quota?	Knows French?	Covid vaccinated?	Manipalite?	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

B) Find the Page rank of web pages A,B,C,D as in Figure Q.2B after two iterations. Assume damping factor to be 0.85 and initial Page rank of each page to be 1/4. (3)

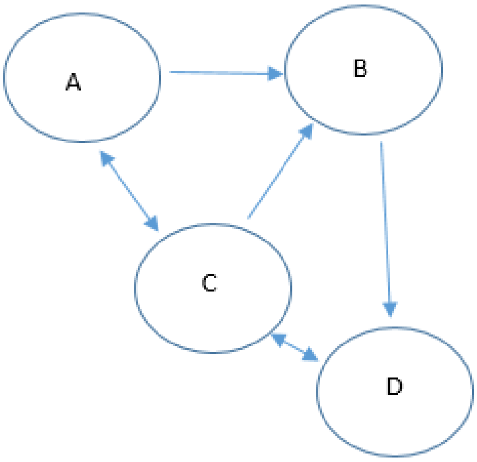
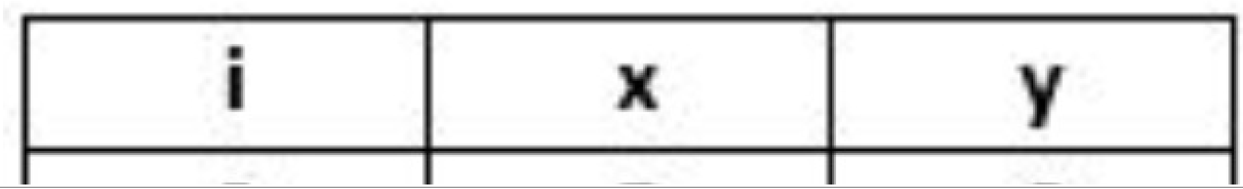


Figure Q.2B

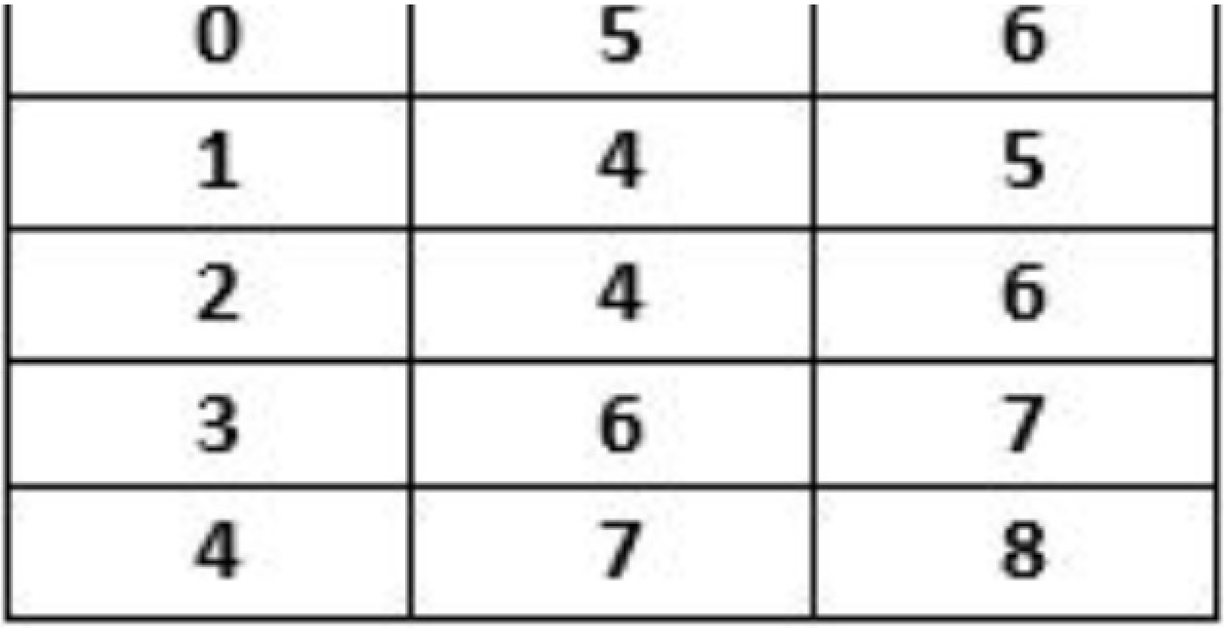
C) A classification model may change dynamically along with the changes of training data streams. This is known as concept drift. Explain why decision tree induction may not be a suitable method for such dynamically changing data sets. Is naive Bayesian a better method on such data sets? Explain your reasoning. (2)

3) A) Obtain the clusters for the dataset given in Table Q.3A. using K-medoid algorithm. Assume the initial centroids C1 and C2 as object i2 and i3 respectively and number of cluster, k=2. Shift the centroid C1 to a new data point i1 and check whether the shift will result in a better cluster. Use Manhattan distance metric for distance computation. (5)

Table Q.3A



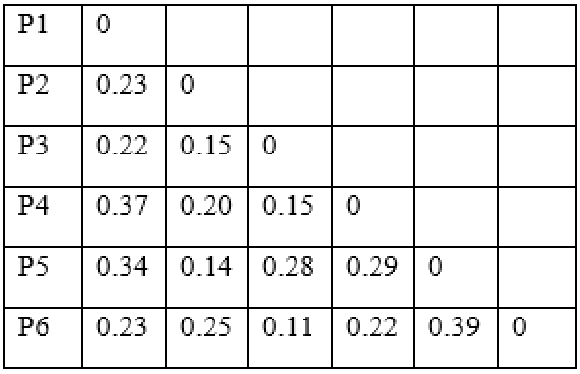
i	x	y



0	5	6
1	4	5
2	4	6
3	6	7
4	7	8

B) Obtain clusters by applying average linkage hierarchical clustering on the distance matrix given in Table Q.3B. (3)

Table Q.3B



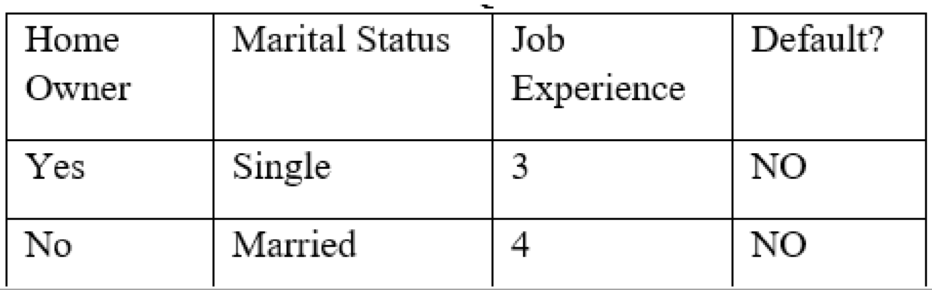
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

C) In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. (2)

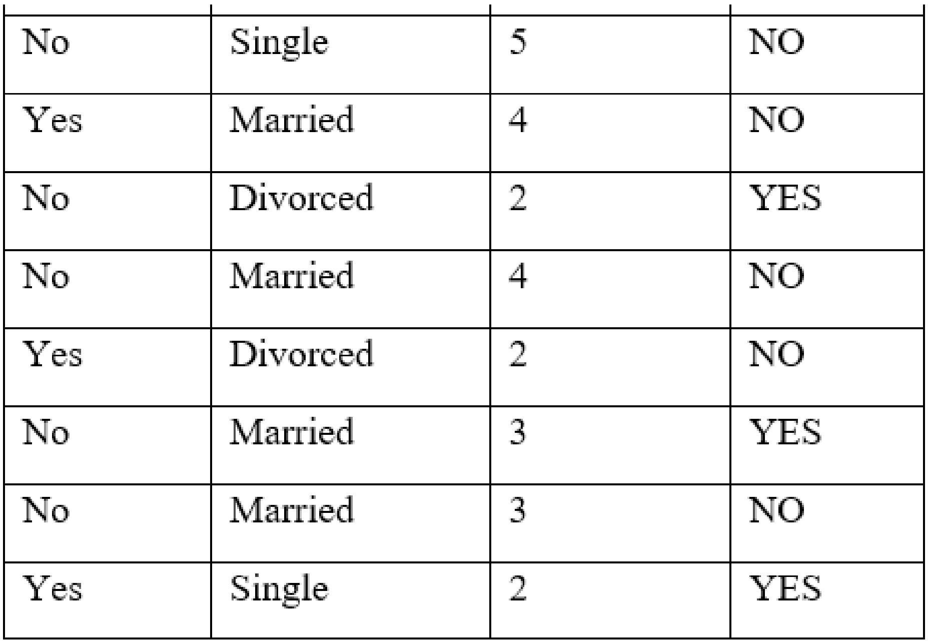
4) A) Consider a Table Q.4A of tuples which tells whether a person will default his loan or not. Predict using naïve bayes classification whether Mr.Sodhi Jr. would default his loan if he doesn't own a house and is married with a job experience of 3years. (5)

Also show, how the Laplacian correction is used to avoid computing probability values of zero?

Table Q.4A



Home Owner	Marital Status	Job Experience	Default?
Yes	Single	3	NO
No	Married	4	NO



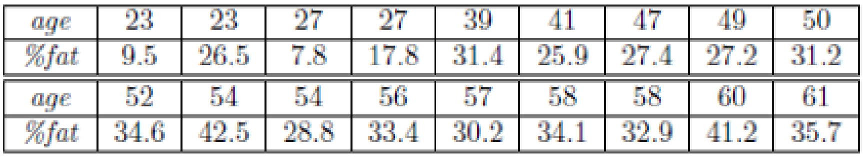
No	Single	5	NO
Yes	Married	4	NO
No	Divorced	2	YES
No	Married	4	NO
Yes	Divorced	2	NO
No	Married	3	YES
No	Married	3	NO
Yes	Single	2	YES

B) Design and explain a typical framework for data warehousing in business. (3)

C) How does the epsilon value affect the DBSCAN Clustering Algorithm? (2)

5) Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the results as in Table Q.5A . (5)

A) **Table Q.5A**



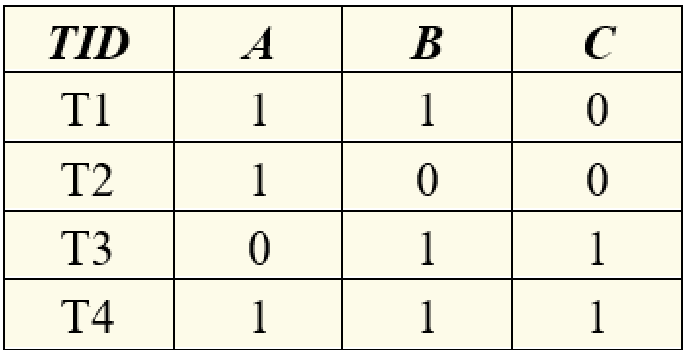
age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2
age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

i. Calculate the mean, median and standard deviation of age and %fat

ii. Draw the boxplots for age and %fat

B) Given min_supp = 25% and M = 2, apply DIC algorithm on the dataset given in Table Q.5B to obtain the frequent patterns. (3)

Table Q.5B



TID	*A*	*B*	*C*
T1	1	1	0
T2	1	0	0
T3	0	1	1
T4	1	1	1

C) Every one of your neighbours moving out of the neighbourhood on the same day is an example of which type of outlier? Defend your answer. (2)

-----End-----