No. of Printed Pages : 6 **RCSE–001**

Ph. D. (COMPUTER SCIENCE)

(PHDCS)

Term-End Examination

June, 2025

RCSE-001 : DATA MINING

Time : 3 Hours *Maximum Marks : 100*

Note : *(i) Question No. **1** is compulsory and carries 40 marks.*

*(ii) Answer any **three** questions from the rest.*

(iii) Use of scientific calculator is allowed.

1. (a) Explain Data Mining functionalities with examples. 5

   (b) How is data warehouse different from database ? Also, mention various characteristics of a data warehouse. 5

(c) Consider that minimum and maximum value of an attribute age are 13 and 50 respectively. Use min-max normalization to transform the age values 35 for, where the range is [0.0, 1.0]. 5

(d) What is support and confidence in Association Rule Mining ? Explain with an example. 5

(e) With the help of an example, explain star and snowflake schemas. 5

(f) What is over-fitting and how can it be avoided ? Explain. 5

(g) Describe how the first level of itemsets is generated in Apriori algorithm. 5

(h) Define the Percentage Split approach for training and testing a model. Discuss its advantages and disadvantages. 5

2. (a) Describe the challenges in data mining with respect to performance issues. 5

(b) What is the difference between the *three* main types of data warehouse usage : Information processing, Analytical processing and Data mining ? Also, discuss the motivation behind OLAP mining. 10

(c) Compare enterprise warehouse, data mart and cloud warehouse with the help of suitable examples. 5

3. (a) Given the following data (in increasing order) for an attribute 'age' :

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

Apply bin means, bin median and bin boundaries to smooth these data using a bin depth of 3. Illustrate your steps. 8

(b) Explain with the help of an example where data mining is crucial to the success of a business. What data mining functions does this business need ? Can

they be performed alternatively by data query processing or simple statistical analysis ? 12

4. A database has five transactions. Let min-supp=60% and min-conf = 80% :

TID	Items-bought
T_{100}	{K, A, D, B}
T_{200}	{D, A, C, E, B}
T_{300}	{A, C, B, E}
T_{400}	{B, D, E}
T_{500}	{A, D, B}

(a) Find all frequent itemsets *without candidate generation*. 10

(b) List all of the strong association rules (with min_supp and min_conf.) 10

5. (a) The following table consists of training data from an employee database. The data have been generalized. For a given row entry, count represents the number

of data types having the values for department, status, age and salary given below. Let salary be the class level attribute :

Department	Status	Age	Salary (₹)	Count
Sales	Senior	31... 35	46k... 50k	30
Sales	Junior	26... 30	26k... 30k	40
Sales	Junior	31... 35	31k... 35k	40
System	Junior	21... 25	46k... 50k	20
System	Senior	31... 35	66k... 70k	5
System	Junior	26... 30	46k... 50k	3
System	Senior	41... 45	66k... 70k	3
Marketing	Senior	31... 40	46k... 50k	10
Marketing	Junior	31... 35	41k... 45k	4
Secretary	Senior	46... 50	36k... 40k	4
Secretary	Junior	26... 30	26k... 30k	6

Given a data sample with the values X = {“System”, “Junior”, “26 .... 30”} for the attributes department, status and age respectively. What would be a

Naive-Bayesian classification of the salary for the sample ? Illustrate your steps. 15

(b) Define performance measuring parameters : Accuracy, Recall, Precision and F_1-measure using confusion matrix. 5

× × × × ×