Hall Ticket No □□□□□□□□□□ Question Paper Code: ACIC01

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal-500043, Hyderabad

B.Tech VI SEMESTER END EXAMINATIONS (REGULAR/SUPPLEMENTARY) - JUNE 2025

Regulation: UG-20

DATA MINING AND KNOWLEDGE DISCOVERY

Time: 3 Hours **(COMMON TO CSE | CSIT | CSE(CS))** **Max Marks: 70**

Answer ALL questions in Module I and II

Answer ONE out of two questions in Modules III, IV and V

All Questions Carry Equal Marks

All parts of the question must be answered in one place only

MODULE – I

1. (a) Describe different types of attributes that can be found in a dataset. Provide examples for each type. [BL: Understand| CO: 1|Marks: 7]

   (b) A healthcare organization wants to predict patient readmission rates. Given a dataset of patient records, describe the process of building a predictive model. Which data mining techniques would you employ? Elucidate. [BL: Apply| CO: 1|Marks: 7]

MODULE – II

2. (a) What is data cleaning? List and explain the different ways of handling missing values. [BL: Understand| CO: 2|Marks: 7]

   (b) Assume that the value of the income attribute are 2000,3000,4000,6000 and 10,000. The income has to be mapped to the range [0.0,1.0]. Do min-max normalization, z-score normalization and decimal scaling for income attribute. [BL: Apply| CO: 2|Marks: 7]

MODULE – III

3. (a) Compare and contrast three OLAP server architectures: MOLAP, ROLAP, and HOLAP. Discuss their advantages, disadvantages, and suitable use cases for each architecture. [BL: Understand| CO: 3|Marks: 7]

   (b) Suppose a data warehouse consists of three dimensions time, doctor and patient and two measures count and charge, where charge is the fee that a doctor charges a patient for a visit.
   i) Enumerate three classes of schemas that are popularly used for modelling data warehouses.
   ii) Draw a schema diagram for above data warehouse using schema classes listed in (i). [BL: Apply| CO: 3|Marks: 7]

4. (a) Outline about multidimensional data model. Explain typical OLAP operations on multidimensional data. [BL: Understand| CO: 4|Marks: 7]

   (b) Explain the importance of indexing OLAP data and describe two indexing strategies that can be used to optimize query performance in an OLAP system. [BL: Understand| CO: 4|Marks: 7]

MODULE – IV

5. (a) Write about decision tree induction with an algorithm for generating decision tree from training tuples. [BL: Understand| CO: 5|Marks: 7]

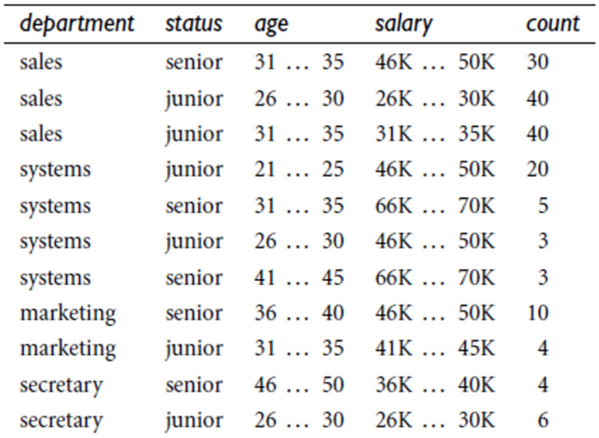
   (b) Construct FP tree and find out frequent patterns for the following transaction data given in Table 1. Assume minimum support count as 3. [BL: Apply| CO: 5|Marks: 7]

Table 1

Transaction ID A	Items
T1	{E, K M, N, O, Y}
T2	{D, E, K, N, O, Y}
T3	{A, E, K, M}
T4	{C, K, M, U, Y}
T5	{C, E, I, K, O, O}

6. (a) Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules. [BL: Understand| CO: 5|Marks: 7]

   (b) The following image shown in Figure 1 consists of training data from an employee database. The data have been generalized. For example, "31 ... 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row. [BL: Apply| CO: 5|Marks: 7]



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
systems	junior	21 ... 25	46K ... 50K	20
systems	senior	31 ... 35	66K ... 70K	5
systems	junior	26 ... 30	46K ... 50K	3
systems	senior	41 ... 45	66K ... 70K	3
marketing	senior	36 ... 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

Figure 1

Let status be the class-label attribute.

i) Design a multilayer feed-forward neural network for the given data. Label the nodes in the input and output layers.

ii) Using the multilayer feed-forward neural network obtained in (i), show the weight values after one iteration of the backpropagation algorithm, given the training instance "(sales, senior, 31 . . . 35, 46K . . . 50K)". Indicate your initial weight values and biases and the learning rate used.

MODULE – V

7. (a) Mention different types of data used for cluster analysis. List and explain the typical requirements of clustering in data mining. [BL: Understand| CO: 6|Marks: 7]

   (b) Consider the points $A_1(2, 10)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1(5, 8)$, $B_2(7, 5)$, $B_3(6, 4)$, $C_1(1, 2)$, $C_2(4, 9)$. Assume that Euclidean distance is used and the initial centers of the clusters are A1,B1 and C2. The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the k-means algorithm to show only

   i) The three cluster centers after the first round of execution.

   ii) The final three clusters. [BL: Apply| CO: 6|Marks: 7]

8. (a) How is the k-Medoids clustering method different from agglomerative and divisive clustering. Explain the method and develop an algorithm. [BL: Understand| CO: 6|Marks: 7]

   (b) Prove that in DBSCAN, the density-connectedness is an equivalence relation. [BL: Apply| CO: 6|Marks: 7]

— ∘ ∘ ◯ ∘ ∘ —