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

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal-500043, Hyderabad

B.Tech VI SEMESTER END EXAMINATIONS (REGULAR) - JULY 2023
Regulation: UG-20
DATA MINING AND KNOWLEDGE DISCOVERY

Time: 3 Hours **(COMMON TO CSE | CSE(CS) | CSIT)** **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) How would you design a data visualization to illustrate user engagement patterns on a social media platform? Explain. [BL: Understand| CO: 1|Marks: 7]

   (b) Create a detailed analysis of the interplay between objective analysis and subjective interpretation in the knowledge discovery process, with regard to a shoe company analyzing customer feedback data to improve its product. [BL: Apply| CO: 1|Marks: 7]

MODULE – II

2. (a) Describe the steps involved in data preprocessing in detail. Explain each step's significance in preparing data for analysis. [BL: Understand| CO: 2|Marks: 7]

   (b) Identify the possible advantages that data discretization could have for a telecom provider looking at call records. Write a detailed note about it. [BL: Apply| CO: 2|Marks: 7]

MODULE – III

3. (a) What is a data warehouse? With the help of a neat sketch, demonstrate the various components in a data warehousing system. [BL: Understand| CO: 3|Marks: 7]

   (b) Design a data warehouse model for an enterprise that needs to integrate data from multiple sources, ensure scalability, and maintain data consistency and quality. Outline the components and processes involved in the data warehouse model, and mention any considerations or trade-offs to be aware of during implementation. [BL: Apply| CO: 3|Marks: 7]

4. (a) Mention the different schemas for the data multidimensional data models. Describe it in detail. [BL: Understand| CO: 4|Marks: 7]

   (b) Design a snowflake schema for a data warehouse to address challenges related to dimensional hierarchies, scalability, and data integrity. Outline the components and relationships within the schema and mention any considerations or trade-offs to be aware of during implementation. [BL: Apply| CO: 4|Marks: 7]

MODULE – IV

5. (a) Write about classification. With a neat figure, explain the general approach for solving classification model. [BL: Understand| CO: 5|Marks: 7]

(b) A grocery store wants to analyze their sales data to identify frequent itemsets and association rules using the Apriori algorithm. However, they are struggling to implement the algorithm effectively due to the large size of their dataset. Propose a problem-solving approach to address this issue and improve the efficiency of applying the Apriori algorithm for the grocery store.

[BL: Apply| CO: 5|Marks: 7]

6. (a) Illustrate the algorithm of decision tree induction with suitable example.

[BL: Understand| CO: 5|Marks: 7]

(b) Find the frequent itemsets and generate association rules on the following transaction given in Table 1. Assume that minimum support threshold (s = 33.33%) and minimum confident threshold (c = 60%). Using Apriori algorithm. [BL: Apply| CO: 4|Marks: 7]

Table 1

Transaction ID	Items
T1	Hot Dogs, Buns, Ketchup
T2	Hot Dogs, Buns
T3	Hot Dogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	Hot Dogs, Coke, Chips

MODULE – V

7. (a) Summarize about clustering. Describe different clustering methods with suitable examples.

[BL: Understand| CO: 6|Marks: 7]

(b) A retail company wants to use K-means clustering with the partitioning method to segment their customer base. However, they are unsure about determining the optimal number of clusters (K) and initializing the cluster centroids. Propose a problem-solving approach to address these challenges and improve the K-means clustering process for the retail company.

[BL: Apply| CO: 6|Marks: 7]

8. (a) Classify various types of sequence data. Explain them with a suitable examples.

[BL: Understand| CO: 6|Marks: 7]

(b) A research institute is working on mining complex types of data that involve symbolic sequences, such as DNA sequences or natural language text. They are facing difficulties in effectively analyzing and extracting meaningful patterns from these symbolic sequences. Propose a problem-solving approach to address these challenges and improve the mining process for the research institute.

[BL: Apply| CO: 6|Marks: 7]