Hall Ticket No | | | | | | | | | | | Question Paper Code: AITC17

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal-500043, Hyderabad

B.Tech VI SEMESTER END EXAMINATIONS (REGULAR) - JULY 2023
Regulation: UG-20
DATA WAREHOUSING AND DATA MINING
Time: 3 Hours **(INFORMATION TECHNOLOGY)** **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) What is data warehouse? With the help of a neat sketch, explain the various components of data warehousing system. [BL: Understand| CO: 1|Marks: 7]

   (b) Identify the primary key attributes of E-Commerce application database description to find fact table and dimension table and build the star and snow flake schema for the same. [BL: Apply| CO: 1|Marks: 7]

MODULE – II

2. (a) How a data mining system can be integrated with a data warehouse? Discuss with an example. [BL: Understand| CO: 2|Marks: 7]

   (b) Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 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.

   i) Use min-max normalization to transform the value 35 for age onto the range [0.0,1.0].

   ii) Z-Score normalization

   iii) Decimal scaling normalization [BL: Apply| CO: 2|Marks: 7]

MODULE – III

3. (a) Discuss an alternative method that have been developed to overcome the limitations of the apriori algorithm. [BL: Understand| CO: 3|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: 3|Marks: 7]

4. (a) Explain in detail about the algorithm for mining frequent itemsets without candidate generation. [BL: Understand| CO: 4|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%). [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 – IV

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

   (b) A training set for a Naïve Bayes classifier is shown in the following Table 2. The classifier's goal is to forecast whether or not a loan borrower will default on payments. The class label is indicated in the table's defaulted borrower column. [BL: Apply| CO: 5|Marks: 7]

Table 2

Home Owner	Material Status	Annual Income	Defaulted Borrower
Yes	Single	125K	No
No	Married	100K	No
No	Single	70K	No
Yes	Married	120K	No
No	Divorced	95K	Yes
No	Married	60K	No
Yes	Divorced	220K	No
No	Single	85K	Yes
No	Married	75K	No
No	Single	90K	Yes

Given the test record, X=(Home Owner= No, Marital Status= Married, Annual Income=120K). Calculate the probability for the following attributes

i) Home Owner= No and Defaulted Borrower = Yes

ii) Home Owner= No and Defaulted Borrower = No

iii) Marital Status= Married and Defaulted Borrower = Yes

iv) Marital Status= Married and Defaulted Borrower = No

6. (a) Discuss potential challenges and limitations associated with backpropagation for classification, such as overfitting, vanishing gradients, and the need for extensive training data. [BL: Understand| CO: 5|Marks: 7]

(b) Consider a training set that contains 60 positive examples and 100 negative examples, for each of the following candidate rule.

Rule R1: Covers 50 positive examples and 5 negative examples.

Rule R2: Covers 2 positive examples and no negative examples.

Determine which is the best and worst candidate rule according to

i) Rule accuracy

ii) Likelihood ratio statistic

iii) Laplace measure [BL: Apply| CO: 5|Marks: 7]

MODULE – V

7. (a) What is hierarchical clustering? With an example discuss dendrogram representation for hierarchical clustering of data objects. [BL: Understand| CO: 6|Marks: 7]

(b) Consider five points X_1, X_2, X_3, X_4, X_5 with following coordinates as a two dimensional sample for clustering: X_1 = (0,2.25); X_2 = (0,0.25); X_3 = (1.25,0); X_4 = (4.5, 0), X_5 = (4.5,2.5)
Implement the K-mean partitioning algorithm on the given data set.
[BL: Apply| CO: 6|Marks: 7]

8. (a) List different types of variables used in cluster analysis. Describe them with suitable examples.
[BL: Understand| CO: 6|Marks: 7]

(b) Explain the importance of feature extraction and selection in mining multimedia databases. Discuss techniques such as image and video feature extraction (e.g., color, texture, shape, motion), audio feature extraction (e.g., pitch, timbre, rhythm). [BL: Apply| CO: 6|Marks: 7]

– ∘ ∘ ◯ ∘ ∘ –