Seat No. : _____

AN-107

April-2025

Int. M.Sc. (CA & IT), Sem.-VIII

Data Mining & Data Analytics

Time : 2:30 Hours] **[Max. Marks : 70**

1. (A) Answer briefly any **two** of the following : **10**
   (1) Explain the top-down approach used in Data designing. Write down the advantages and disadvantages of this approach.
   (2) What is ETL (Extract, Transform, Load) ? Explain its components and workflow with the help of a diagram.
   (3) What is Cloud Data Warehousing ? Explain the key characteristics of Cloud Data Warehousing.
   (4) Compare OLAP and OLTP systems, highlighting their key differences.

1. (B) Draw a Snowflake Schema diagram for a Sales Data Warehouse using the following fact and dimension tables : **4**
   - Fact Table :
     Sales_Fact (Sales_ID, Product_ID, Customer_ID, Date_ID, Quantity_Sold, Total_Amount)
   - Dimension Tables :
     (1) Product_Dim (Product_ID, Product_Name, Category_ID)
     (2) Category_Dim (Category_ID, Category_Name)
     (3) Customer_Dim (Customer_ID, Customer_Name, Region)
     (4) Date_Dim (Date_ID, Date, Month, Year)

2. (A) Answer any **two** of the following : **8**
   (1) List the issues in Data integration. Explain any one in detail.
   (2) Explain the following procedures for attribute subset selection :
      (a) Stepwise forward selection
      (b) Stepwise backward elimination
   (3) Define Knowledge Discovery in Databases (KDD). Describe the major steps involved in the KDD process.
   (4) The monthly spending (in Rs.) of 5 customers is given as :
      [90, 100, 110, 120]
      (a) Normalize the data using Min-Max Normalization (range 0 to 1)
      (b) Normalize the data using Z-score Normalization, given :
         - Mean (μ) = 110
         - Standard Deviation (σ) = 14

2. (B) A school administrator wants to determine whether there is an association between students' grade level and their preferred mode of learning. The data collected from a sample of students is shown below : **6**

Mode of Learning	Grade 10	Grade 12	Total
Online	25	15	40
In-Person	15	25	40

Using the Chi-Square Test for Independence, test whether there is a significant relationship between grade level and preferred mode of learning at the 0.05 significance level.

Given : At a significance level of 5%, consider the values for degrees of freedom.

Degree of Freedom	1	2	3	4
Significance level	3.841	5.991	7.815	9.0488

3. (A) Answer any **one** of the following : **6**

(1) An e-commerce store wants to analyze customer purchase patterns. The transactions below represent items frequently bought together.

Transaction ID	Items Purchased
T1	Laptop, Mouse, Headphones
T2	Mouse, Keyboard
T3	Laptop, Mouse
T4	Laptop, Keyboard, Mouse
T5	Headphones, Keyboard
T6	Laptop, Mouse, Headphones, Keyboard
T7	Laptop, Keyboard

Task :

(1) Apply the Apriori algorithm with a minimum support count = 2.

(2) Identify frequent item sets.

(3) Generate the association rules along with their confidence calculations for the following item sets :

- {Laptop, Mouse} $\rightarrow$ {Keyboard}
- {Mouse } $\rightarrow$ {Laptop}
- {Laptop} $\rightarrow$ {Keyboard}
- (Laptop, Mouse, Headphone} $\rightarrow$ {Keyboard}

Determine the strong association rules among these based on their confidence values.

(2) Given the following transaction data set, apply the FP-Growth algorithm with a minimum support count = 3

(1) Construct the FP-tree.

(2) Identify the Conditional Pattern Base.

(3) Generate the Conditional FP-tree.

(4) Generate all frequent patterns.

Transaction ID	Items Purchased
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}

3. (B) Answer any **two** of the following questions : **8**

(1) Explain the role of Association Rules in Market Basket Analysis. What do support, confidence, and lift represent ?

(2) Explain the different variations of the Apriori algorithm that aim to improve its efficiency.

(3) Differentiate between strong and weak association rules in the context of Market Basket Analysis.

4. (A) A career guidance platform wants to recommend a suitable career path to a new user based on their interests. **8**

Below is the data set showing the interest scores and corresponding career paths chosen by existing users :

User ID	Analytical Score	Creative Score	Social Score	Career Path
1	9	3	2	Data Scientist
2	3	9	5	Graphic Designer
3	8	4	3	Software Engineer
4	2	8	7	Content Writer
5	10	2	1	Research Analyst
6	4	6	8	HR Manager
7	6	5	5	? (New User)

Using the 3-Nearest Neighbor Classifier, suggest three career paths that align best with the new user's profile (User 7), based on Euclidean distance.

4. (B) Answer any **two** of the following : **6**

(1) A small restaurant keeps track of customer feedback to understand whether a customer is satisfied with their service. The data consists of two attributes : "Food Quality" and "Service", along with the decision "Satisfied" (Yes/No).

Food Quality	Service	Satisfied
Good	Fast	Yes
Good	Slow	Yes
Average	Fast	Yes
Poor	Fast	No
Poor	Slow	No

(1) Calculate the entropy of the target variable ("Satisfied").

(2) Calculate the Information Gain for each attribute (Food Quality & Service).

(3) Determine the root node and construct a decision tree.

(2) Explain the concept of Support Vector Machine (SVM) in classification. How does SVM find the optimal hyperplane for separating classes ? Illustrate with a suitable diagram.

(3) What is tree pruning in decision trees ? Performance. Differentiate between pre-pruning and post-pruning.

5. (A) You are given a data set with two features : App Usage Hours per Week and Number of In-App purchases for 5 years. **6**

Apply K-Means Clustering with K = 2 clusters and perform the first iteration, showing the updated centroids at the end.

User	App Usage hours / week	In-App purchases
U1	4	2
U2	10	5
U3	12	7
U4	3	1
U5	9	6

5. (B) Answer any **two** of the following : **8**

(1) What are the different methods of clustering in data mining ? Explain the clustering methods with suitable examples.

(2) What are outliers in data analysis ? Explain the different types of outliers with suitable examples.

(3) Explain the concept of Cluster Analysis. Describe the different types of data that can be used in clustering.