No. of Printed Pages : 4

Sl. No.

C5-R4 : DATA WAREHOUSING AND DATA MINING

DURATION : 03 Hours **MAXIMUM MARKS : 100**

Roll No. : | | | | | | |

Answer Sheet No. : | | | | | | |

Name of Candidate : _____ ; **Signature of Candidate :** _____

INSTRUCTIONS FOR CANDIDATES :

- Carefully read the instructions given on Question Paper, Answer Sheet.
- Question Paper is in English language. Candidate has to answer in English Language only.
- **Question paper contains Seven questions. The Question No. 1 is compulsory. Attempt any FOUR Questions from Question No. 2 to 7.**
- **Parts of the same question should be answered together and in the same sequence.**
- **Questions are** to be answered in the **ANSWER SHEET** only, supplied with the Question Paper.
- **Candidate cannot leave the examination hall/ room without signing on the attendance sheet and handing over his/her Answer Sheet to the Invigilator. Failing in doing so, will amount to disqualification of Candidate in this Module/Paper.**
- After receiving the instruction to open the booklet and before answering the questions, the candidate should ensure that the Question Booklet is complete in all respects.

DO NOT OPEN THE QUESTION BOOKLET UNTIL YOU ARE TOLD TO DO SO.

1. (a) Explain the difference between data mining and traditional statistical analysis.

(b) Explain the concept of data cube in OLAP.

(c) Given a transaction database with the following transactions :

{1, 2, 3}

{1, 2, 4}

{1, 3, 4}

{2, 3, 5}

Calculate the support and confidence for the rule {1, 3} => {4}.

(d) Explain the concept of data warehousing. Discuss its importance and benefits in the context of Business Intelligence.

(e) Explain the concept of Stream Data Mining. Discuss its characteristics, challenges, and applications.

(f) Explain techniques for cleaning the data in data mining.

(g) Discuss the concept of aggregation and summarization in data reduction. **(7x4)**

2. (a) Explain the challenges associated with data preprocessing in data mining.

(b) Discuss the challenges of data mining. How these challenges can be handled ? Explain by giving suitable justification. **(9+9)**

3. (a) Write down the process of ETL (Extract, Transform, Load) in the context of data warehousing. How ETL is implemented for any data mining application ?

(b) Discuss the importance and challenges of ETL that need to be implemented on data mining applications. **(10+8)**

4. (a) Explain the concept of clustering and its importance in data analysis. What are the clustering methods we have for data partition ?

(b) Describe the K-means clustering algorithm in detail. Discuss the steps involved in the K-means algorithm, including initialization, assignment, update, and convergence criteria. Give a suitable example showing the working of K-Means algorithm. **(9+9)**

5. (a) Consider a dataset containing the following points in two-dimensional space : {(2, 3), (4, 5), (3, 6), (6, 2), (8, 7), (9, 5)}. Apply the K-means algorithm with K=2 clusters. Show the initial centroids, assign points to clusters, update centroids, and iterate until convergence. Provide the final clusters and centroids.

(b) What are the key characteristics of a Data Warehouse ? Explain the characteristics in detail with its importance. **(10+8)**

6. (a) Define classification in the context of machine learning. Discuss its significance and applications.

(b) Explain the difference between binary and multiclass classification with suitable examples. **(9+9)**

7. (a) Consider the following dataset representing information about customers and whether they purchased a product (1 for purchased, 0 for not purchased) :

Customer ID	Age Group	Income Group	Gender	Region	Product Purchased
1	Young	High	Female	Urban	1
2	Adult	Low	Male	Rural	0
3	Adult	High	Female	Urban	1
4	Senior	Medium	Male	Rural	0
5	Young	Low	Male	Urban	1

Using the given dataset:

(i) Calculate the number of customers who purchased the product.

(ii) Determine the percentage of customers who purchased the product.

(iii) Calculate the average income of customers who purchased the product.

(iv) Determine the most common age group among customers who purchased the product.

(b) Given the following set of transactions :

{bread, milk, eggs}

{coffee, bread}

{bread, eggs}

{coffee, bread, eggs}

Find the frequent itemsets using the FP-growth algorithm. **(12+6)**

- o O o -

SPACE FOR ROUGH WORK