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) Define OLAP. Discuss its operation in data warehousing.

(b) Explain briefly different kinds of data on which data mining is applied.

(c) Define the concept of clustering in context of data analysis.

(d) What are the causes of overfitting in classification ? Point out its potential solutions.

(e) What is the concept of frequent pattern mining in data mining ? Discuss its significance and applications.

(f) What is the purpose of FP-growth algorithm ? Discuss its advantages over Apriori algorithm.

(g) Define time series data mining. Mention the characteristics of time series data and write down the common techniques used for its analysis.

(7x4)

2. (a) Discuss the main types of clustering algorithms with their key characteristics. Show the process of any clustering algorithm with an example.

(b) Explain how data mining is applied on sequence data. Is there any specific constraint for sequence data mining algorithms ?

(9+9)

3. (a) Given a dataset of student exam scores in two subjects (Mathematics and English), perform hierarchical clustering to group students based on their performance. Apply single-linkage clustering method with Euclidean distance as the similarity measure. Show the dendrogram representing the hierarchical clustering process and discuss the resulting clusters.

Student	Mathematics Score	English Score
S1	70	75
S2	60	80
S3	85	90
S4	55	70
S5	75	95

(b) How does clustering differ from classification, and regression techniques ? Mention the key features of the above three methods.

(9+9)

4. (a) What are the applications of Data Warehousing and Data Mining ? Discuss any two applications in detail.

(b) Consider the following dataset representing information about patients and whether they have a particular medical condition (1 for positive, 0 for negative) :

Patient ID	Age	Gender	Blood Pressure	Cholesterol Level	Medical Condition
1	45	Male	High	Normal	1
2	30	Female	Normal	High	0
3	55	Male	High	High	1
4	40	Female	Normal	Normal	0
5	50	Male	High	Normal	1

Using the given dataset :

(i) Calculate the number of patients with positive medical condition.

(ii) Determine the percentage of patients with positive medical condition.

(iii) Calculate the average age of patients with negative medical condition.

(iv) Determine the most common blood pressure level among patients with Positive medical condition. **(6+12)**

5. (a) Apply Apriori algorithm to find frequent item sets in the following transaction database. Use a minimum support threshold of 2.

{1, 2, 3, 4}
{1, 2, 4}
{1, 3, 4}
{2, 3, 5}
{2, 3, 4}

(b) Write a short note on the evolutionary path of database technology that has led to the need for data warehousing and data mining. **(9+9)**

6. (a) Write down the key differences between OLAP and OLTP in view of data model, query, response time, concurrency, workload.

(b) Explain the following OLAP operations: slice, dice, roll-up (drill-up), and drill-down (drill-through). **(9+9)**

7. (a) Explain in detail the several techniques used in data preprocessing to clean transform integrate, reduce and prepare data for further analysis.

(b) Write a short note on Attribute Oriented Induction(AOI). Discuss its uses and limitations. **(9+9)**

- o 0 o -

SPACE FOR ROUGH WORK