No. of Printed Pages : **2**

Sl. No.

C5-R4 : DATA WAREHOUSING AND DATA MINING

NOTE :

1. **Answer question 1 and any FOUR from questions 2 to 7.**
2. **Parts of the same question should be answered together and in the same sequence.**

Time: 3 Hours Total Marks: 100

1. (a) Differentiate between Multidimensional Online Analytical Processing (MOLAP) and Relational Online Analytical Processing (ROLAP).

(b) "Data mining is the process of discovering interesting patterns and knowledge from large amounts of data", Justify. Enlist the major issues in data mining.

(c) Explain following statement: "Data smoothing techniques are used to eliminate "noise" and extract real trends and patterns". Explain data smoothing techniques to remove the noise.

(d) Why is tree pruning useful in decision tree induction ? What is a drawback of using a separate set of tuples to evaluate pruning?

(e) Write a short note on data mining task primitives.

(f) What is Data Mart with respect to data warehouse and data mining ?

(g) What is an Iceberg Query ? Can we use it in market basket analysis ? **(7x4)**

2. (a) List and explain in brief the Online Analytical Processing (OLAP) operations in multidimensional data model.

(b) Describe the steps involved in data mining when viewed as a process of knowledge discovery. **(9+9)**

3. (a) Frequent patterns are item sets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. Enlist various frequent item-set mining methods and explain any one method in detail.

(b) Briefly describe genetic algorithm. Also enlist the advantages of the same. **(9+9)**

4. (a) List strengths and weakness of neural network as classifier.

(b) Describe various methods which evaluate the accuracy of a classifier or a predictor.

(c) Explain Apriori Algorithm with example. **(6+6+6)**

5. (a) Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. Draw a snowflake schema diagram for the data warehouse.

(b) Discuss about Hierarchical clustering and Density based clustering.

(c) What is Time-series Database? How to characterize the time-series data using trend analysis? **(6+6+6)**

6. (a) What are multidimensional Association Rules? Explain in brief.

(b) What is classification? Compare the advantages and disadvantages of eager classification vs. lazy classification. Discuss K- Nearest-neighbor classifier. **(9+9)**

7. (a) Data reduction is a part of Data pre-processing. Explain the strategies involved in data reduction.

(b) A data warehouse is a collection of information and data derived from operational systems and external data sources. Explain 3-tier Data Warehouse Architecture. **(9+9)**

- o O o -