No. of Printed Pages : **2**

Sl. No.

BE6-R4 : DATA WAREHOUSING AND DATA MINING

NOTE :

1. **Answer question 1 and any FOUR questions from 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 data warehousing and data mining.
(b) What are the key features of a data warehouse ? Explain.
(c) Mention any four reasons to perform data pre-processing.
(d) What are four basic analytical operations that can be performed on an OLAP cube ?
(e) What is the significance of data discretization in data mining ? Differentiate between top down and bottom up discretization processes.
(f) Differentiate between descriptive and predictive data mining.
(g) Define data cube approach for data characterization. What are its advantages ? **(7x4)**

2. (a) Compare OLAP and OLTP.
(b) Explain top-down approach to design a data warehouse with a suitable diagram. Discuss its advantages and disadvantages.
(c) Explain fact constellation schema. **(7+7+4)**

3. (a) What are the characteristics of a star schema ? Compare it with snowflake schema.
(b) Schema of a database university_database is as follows :
Student (name, sno, status, branch, cgpa, birth_date, birth_place, address)
Course (cno, title, department)
Grading (sno, cno, faculty, semester, grade)
Use the database to write DMQL queries for the following.
(i) Find the general characteristics of the graduate students in computer science in relevance to attributes cgpa, birth_place and address, for the students born in India. Make sure that any tuple taking less than 5% of the total count should not be included in final result.
(ii) Find the discriminant features to compare graduate students versus undergraduate students in computer science in relevance to attributes cgpa, birth_place and address, for the students born in India.
(iii) Classify students according to their cgpa's and find their classification rules for those in computer science branch and born in India with the attributes birth_place and address in consideration.
(iv) Find strong association relationships for those students who are in computer science branch and born in India, in relevance to the attributes cgpa, birth_place and address. Keep support and confidence thresholds to 5% and 70% accordingly.
(c) What is the role of concept hierarchy ? Provide DMQL syntax for the following.
(i) To specify concept hierarchy at the schema level.
(ii) To specify concept hierarchy by set grouping.
(iii) To insert a term as a sub-ordinate concept of a super-ordinate one in a hierarchy or delete a term from it. **(10+4+4)**

4. (a) What are different types of OLAP systems ? Discuss hierarchical structure of OLAP. What are its advantages and disadvantages ?

(b) Consider the example of sales of four companies C1, C2, C3 & C4 per quarter on the basis of product category (Men's, Women's, Electronics and Home). Out of the four companies, two companies C1 and C2 are from India and C3 and C4 are from America. Apply four basic OLAP operations and explain the obtained results. **(10+8)**

5. (a) Discuss the basic principles and approach of attribute oriented induction based data generalization.

(b) What is the significance of multilevel association rules in data mining ? Discuss different approaches for multilevel association rule mining.

(c) Differentiate between FP growth and Apriori algorithm used for mining frequent items. **(7+7+4)**

6. (a) Discuss decision tree algorithm with an example. What are the advantages of using decision tree for supervised learning ?

(b) Explain step by step process of K-Nearest Neighbour algorithm. The table below shows classification label Y (Good/ bad) of 4 items based on 2 features X1 and X2. Apply KNN algorithm (K = 3) to find out the classification label of a new item with X1 = 3 and X2 = 7.

X1	X2	Y
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

(c) Differentiate between Linear and Non-linear regression. **(6+8+4)**

7. (a) What are the advantages of using hierarchical clustering ? Discuss any two approaches used for hierarchical clustering.

(b) What is web mining ? Briefly explain the following :

(i) Web usage mining

(ii) Web content mining

(c) What are outliers ? How does outlier affect K-means clustering ? **(8+6+4)**

- o 0 o -