BE6-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) What is a decision tree? What do the internal nodes and leaves of the decision tree represent?

b) What is a dendrogram? What do the leaves and root of the dendrogram represent?

c) Give one example each of numeric, ordinal, nominal and binary attribute from a health care database.

d) What is the difference between the OLAP and OLTP?

e) What is the difference between test and training accuracy of a classifier?

f) What is ROLAP?

g) What is the difference between gini index and entropy?

(7x4)

2.

a) What are the differences between operational database systems and data warehouses?

b) A data-warehouse for a university consists of four dimensions – student, course, semester, instructor. Two measure are maintained – count and average-grade. Average grade is the average grade for a course, semester, instructor at the lowest level, count is the number of students. Draw a star schema for the data-warehouse.

(8+10)

3.

a) How is a data mart related to data warehouse?

b) What is a data cube? How many cuboids are there in an n-dimensional cube?

c) What are are two main methods of indexing OLAP data?

(6+6+6)

4.

a) Write algorithm for K-Nearest Neighbor. Why is it called a Lazy Classifier?

b) Differentiate between Gain Ratio and Information Gain. Compute both values for the given data set.

Data	low	medium	low	low	high	high	high	medium	medium	low
Class	blue	blue	blue	blue	red	red	red	red	blue	blue

(9+9)

5.

a) How Crossover and Mutation is performed in Genetic Algorithm? Explain with example.

b) What is the connection between computations of Minkwoski, Euclidean and Manhanttan distance? Compute Euclidean distance between each pair of the points given below, and show in form of a distance matrix

A(2, 5), B(3, 2), C(7,2), D(6,2), E(1, 1)

(9+9)

6.

a) Write an algorithm for finding association rules from a set **F** of frequent item sets along with their respective support values. The rules should satisfy minimum confidence criterion *minc*.

b) What are the different methods to test the performance of a classifier?

(9+9)

7. Write short notes on **any three** of the following:

a) Text Mining

b) Spatio-Temporal Mining

c) Web usage mining

d) Hypothesis Testing

e) Decision Trees for Classification

(3x6)