BE6-R4: DATA WAREHOUSE & 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 entropy? How is it related to information gain?

b) What is a concept hierarchy? Explain with the help of an example.

c) Formally define association rule mining problem.

d) How can you select which modelling technique to use for data mining?

e) What is a confusion matrix? Explain with the help of an example.

f) Differentiate between document classification and document clustering analysis.

g) Explain attribute oriented induction.

(7x4)

2.

a) Given the following confusion matrix for a classifier, find out its accuracy, recall and precision of "True" class.

	Predicted True	Predicted False
Actual True	25	75
Actual False	40	60

b) Give expressions with notations for Gain Ratio and Information Gain. What is the advantage of Gain Ratio over Information Gain?

c) Write an algorithm for K-nearest neighbor classification.

(6+6+6)

3.

a) Write K means algorithm for clustering. Derive the expression for its computational complexity.

b) What is normalization? Given the following data values, apply z-score normalization and min-max normalization to transform the values.

10, 16, 18, 15, 11, 16

(9+9)

4.

a) What is Naïve Bayes Classifier? What is the weakness of the assumption in the method?

b) What is bagging? How does it improve performance?

c) How is similarity search carried out in multimedia data?

(9+4+5)

5.

a) Given the following dataset, apply Apriori algorithm to find the frequent item sets using minimum support = 2.

Transcation id: items
1 : A,B,C
2 : B,C
3 : C,D,E
4 : C,D,E,F
5 : A,B
6 : B
7 : A,D

b) What are multi-level association rules? Explain with the help of an example.

(10+8)

6.

a) How would you compute the dissimilarity/distance between objects with i) asymmetric binary variables ii) nominal variables iii) interval scaled variables.

b) Generate all rules with two consequents from the following set of frequent item sets.
{abcd, abc, abd, bcd, ab, ac, ad, bc, bd, cd, a, b, c, d}

(9+9)

7. Write short notes on:

a) ROLAP and MOLAP

b) Star Schema and Snowflake Schema

c) Temporal and Spatial Databases

(6+6+6)