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

b) Differentiate between Classification and Regression with the help of an example.

c) What is k-fold cross-validation?

d) List and explain methods to handle missing data?

e) What are outliers? How are they useful in data analysis?

f) What is Principle component analysis? How it is used for dimension reduction?

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

(7x4)

2.

a) Explain snowflake schema with the help of an example.

b) Differentiate between ROLAP and MOLAP server architecture with help of neat diagram.

(9+9)

3.

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

b) Write an algorithm for K-Nearest neighbor classification?

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

(6+6+6)

4. Consider the following data set shown in table where bank officials need to build decision tree to classify bank loan applications by assigning applications to one of the three risk classes that is A, B, C

Owns Home	Married	Gender	Employed	Credit Rating	Risk Class
YES	YES	MALE	YES	A	B
NO	NO	FEMALE	YES	A	A
YES	YES	FEMALE	YES	B	C
YES	NO	MALE	NO	B	B
NO	YES	FEMALE	YES	B	C
NO	NO	FEMALE	YES	B	A
NO	NO	MALE	NO	B	B
YES	NO	FEMALE	YES	A	A
NO	YES	FEMALE	YES	A	C
YES	YES	FEMALE	YES	A	C

Draw the decision tree for the above table using ID3 algorithm?

(18)

5. Consider the following data set:

Students	Age	Mark1	Mark2	Mark3
S1	18	73	75	57
S2	18	79	85	75
S3	23	70	70	52
S4	20	55	55	55
S5	22	85	86	87
S6	19	91	90	89
S7	20	70	65	60
S8	21	53	56	59
S9	19	82	82	60
S10	47	75	76	77

Given K=3, using K-Mean Clustering algorithm find the clusters?

(18)

6.

a) Given the set F = {a(100), b(75), c(50), d(25), bc(50), bd(20), bcd(5)} of frequent item-sets with the support counts in bracket. List all association rules with one item of LHS that can be generated from F. Compute confidence of each rule.

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

(10+8)

7. Write Short-notes on

a) Web Data Mining

b) Text Mining

c) Neural Networks

(6+6+6)