USN | | | | | | | | | | |

10IS74

Seventh Semester B.E. Degree Examination, Jan./Feb. 2021
Data Warehousing and Data Mining

Time: 3 hrs. Max. Marks:100

Note: ***Answer any FIVE full questions, selecting atleast TWO questions from each part.***

PART – A

1 a. List five differences between an OLTP and Data ware house. **(10 Marks)**

b. List and explain the major steps involved in the ETL process. **(10 Marks)**

2 a. Explain the FASMI characteristics of OLAP. **(08 Marks)**

b. What is a Data Cube? Explain the different operations performed on Data Cube. **(12 Marks)**

3 a. Distinguish between : i) Random Sampling and Stratified Sampling.
ii) Jaccard coefficient and SMC iii) Discretization and Binarization. **(12 Marks)**

b. Consider the following two binary vectors :
X = (1, 1, 0, 1, 0, 1) ; Y = (1, 1, 1, 0, 0, 1).
Find i) Cosine similarity ii) Correlation similarity. **(08 Marks)**

4 a. Explain various alternative methods for generating frequent item sets. **(10 Marks)**

b. Consider the data set shown in table below Q4(b) :
i) Compute the support for itemsets {e} , {b, d} and {b, d, e} by treating each transaction ID as a market basket.
ii) Use results in part i) to compute the confidence for the association rules {b, d} → {e} & {e} → {b, d}.

Table Q4(b)

Customer ID	Transaction ID	Item Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}
2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

(10 Marks)

PART – B

5 a. Explain the characteristics of nearest neighbour classifier **(10 Marks)**

b. Consider the training examples shown in Table Q5(b) for a binary classification problem.
i) What is the entropy of this collection of training examples with respect to the positive class?
ii) What are the information gains of a_1 and a_2?

Table Q5(b)

Instance	a_1	a_2	a_3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

(10 Marks)

6 a. List and explain different techniques for improving the accuracy of classification results. **(10 Marks)**

b. Discuss the other evaluation criteria for classification methods? **(10 Marks)**

7 a. What is Clustering? How is it different than supervised classification? **(08 Marks)**

b. Describe two hierarchical clustering techniques. **(08 Marks)**

c. List one major difficulty with K – Means Algorithm. **(04 Marks)**

8 Write short notes on :

a. Web content mining.

b. Text mining.

c. Text clustering.

d. Mining spatial data bases. **(20 Marks)**

*****