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

10IS74

Seventh Semester B.E. Degree Examination, Dec.2019/Jan.2020
Data Warehousing and Data Mining

Time: 3 hrs. Max. Marks:100

Note: 1. ***Answer any FIVE full questions, selecting atleast TWO questions from each part.***
2. ***Draw diagram wherever necessary.***

Important Note : 1. On completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages.
2. Any revealing of identification, appeal to evaluator and /or equations written eg, 42+8 = 50, will be treated as malpractice.

PART - A

1 a. What is Data Warehousing and Data Mining? **(03 Marks)**
b. List the major steps involved in the ETL process. **(07 Marks)**
c. List the guidelines for implementing a Data Warehouse. **(10 Marks)**

2 a. What is OLAP? List the characteristics of OLAP systems. **(06 Marks)**
b. Define dimension , members , measure and fact table with an example. **(06 Marks)**
c. Describe the operations, roll – up , drill – down , slice and dice and pivot. **(08 Marks)**

3 a. Explain different data mining tasks with diagram and examples. **(10 Marks)**
b. For the following vectors X and Y calculate the indicated similarity
X = (2, -1, 0, 2, 0, -3) ; Y = (-1, 1, -1, 0, 0, -1).
1) Cosine 2) Correlation. **(06 Marks)**
c. List any three application areas of data mining. **(04 Marks)**

4 a. Explain the Apriori algorithm for frequent itemset generation, with an example. **(08 Marks)**
b. Consider the following transaction data set to construct FP tree. Apply FP – growth algorithm to find frequent itemsets ending in e, show the steps separately. **(12 Marks)**

Transaction ID	Items
1	{a, b}
2	{b, c, d}
3	{a, c, d, e}
4	{a, d, e}
5	{a, b, c}
6	{a, b, c, d}
7	{a}
8	{a, b, c}
9	{a, b, d}
10	{b, c, e}

PART - B

5 a. Consider the training example shown in the table for a binary classification problem.
1) Compute the Gini index for customer id attribute. **(10 Marks)**
2) Compute the Gini index for shirt size attribute using binary split (grouping does not violate order property)

Customer Id	1	2	3	4	5	6	7	8	9	10
Shirt size	small	medium	medium	large	Extra large	Extra large	Small	small	Medium	large
Class	C_0	C_0	C_0	C_0	C_0	C_1	C_1	C_1	C_1	C_1

b. Explain Direct methods for Rule Extraction. **(10 Marks)**

6 a. Explain how bootstrapping , bagging and boosting improve the accuracy of classification. **(06 Marks)**

b. What is Bayes theorem and show how it is used at the basis of the Naïve Bayes method. **(08 Marks)**

c. Explain Evaluation criteria for classification methods. **(06 Marks)**

7 a. Explain desirable features of a Cluster analysis. **(06 Marks)**

b. Describe the K – means algorithm and discuss its strengths and weakness. **(07 Marks)**

c. Explain Divisive Hierarchical method and discuss its advantages and disadvantages. **(07 Marks)**

8 a. What is Web data mining? List the major differences between Conventional searching and Web searching. **(06 Marks)**

b. Explain the different dimensions in a spatial data mining. **(07 Marks)**

c. Explain the different text mining approaches. **(07 Marks)**

*****