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

10IS74

Seventh Semester B.E. Degree Examination, June/July 2023
Data Warehousing and Data Mining

Time: 3 hrs. Max. Marks:100

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

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 warehouse? Explain in detail the different key features of warehouse. **(10 Marks)**

b. Explain in detail the difference between ODS and warehouse. **(05 Marks)**

c. What is Data Mart? **(05 Marks)**

2 a. Explain Codd's OLAP characteristics. **(10 Marks)**

b. Describe data cube operations, with an example. **(10 Marks)**

3 a. What is data mining? Explain various data mining task with examples. **(10 Marks)**

b. List and explain general characteristics of data sets. **(05 Marks)**

c. Distinguish between categorical and numerical attributes. **(05 Marks)**

4 a. Consider data transaction ID : **(10 Marks)**

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

Apply FP growth algorithm to find frequent itemset ending in 'e'.

b. Write a procedure in Apriori – gen function, which merges a pair of frequent item set. Explain with example. **(10 Marks)**

PART – B

5 a. Explain the various measures for selecting the best split with an example. **(06 Marks)**

b. Give the difference between rule based ordering and class based ordering scheme. **(04 Marks)**

c. Consider a training set that contain 100 +ve examples and 400 –ve examples for each of the following candidate rule. Determine which is the best and worst candidate according to
i) Rule accuracy ii) Foil Information gain.
R1 : A → + (covers 4 + ve and 1 – ve examples)
R2 : B → + (covers 30 +ve and 10 –ve examples)
R3 : C → + (covers 100 +ve and 90 –ve examples). **(10 Marks)**

6 a. Consider the following data sets for a binary classification.

Tid	Refund	Marital status	Taxable income	Class
1	Yes	Single	125 K	No
2	No	Married	100 K	No
3	No	Single	70 K	No
4	Yes	Married	120 K	No
5	No	Divorced	95 K	Yes
6	No	Married	60 K	No
7	Yes	Divorced	220 K	No
8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90 K	Yes

i) Calculate the information gain for each attribute
ii) Draw decision tree by selecting best split. **(10 Marks)**

b. What is Baye's theorem? Show how it is used for classification. **(05 Marks)**

c. What are the approaches used for extending binary classifier to handle multiclass problem? **(05 Marks)**

7 a. Briefly outline how to compute dissimilarity between object described by following types of variables in cluster :

i) Interval scaled variable ii) Binary variable. **(10 Marks)**

b. What is Clustering? Describe the following approaches to clustering method :

i) Partitioning method ii) Hierarchical methods.

Give example in each case. **(10 Marks)**

CMRIT LIBRARY
BANGALORE - 560 037

8 a. Explain web content mining and how it is used for discovering useful information from the web. **(10 Marks)**

b. Write short notes on : i) Spatial data mining ii) Text mining. **(10 Marks)**

* * * * *

VTU-05-06-2023 01:13:43pm - CR - CR - CR - CR - CR - CR - CR - CR
05-06-2023 01:34:04pm - CR - CR - CR - CR - CR - CR - CR - CR