No of Pages :2 Course Code :18XT83

Roll No:
(To be filled in by the candidate)

PSG COLLEGE OF TECHNOLOGY, COIMBATORE 641 004

SEMESTER EXAMINATIONS, AUGUST / SEPTEMBER 2023

MSc - THEORETICAL COMPUTER SCIENCE Semester : 8

18XT83 DATA MINING

Time : 3 Hours **Maximum Marks :100**

INSTRUCTIONS:

1. Answer **ALL** questions. Each question carries 20 Marks.
2. Subdivision (a) carries 3 marks each, subdivision (b) carries 7 marks each and subdivision (c) carries 10 marks each.

1. a) For each of the following description mention the type of attribute you will use and justify the same.
   - Exam result of students and analyze those students who got pass marks
   - Type of Pizza available
   - Grades which can be obtained by a student in semester exam.

   b) i) What is symmetric binary attribute? How it differs from asymmetric binary attribute? [3]

   ii) How will you compute Euclidean and Manhattan distance between two objects? [4]

   c) Explain the computation procedure of finding dissimilarity between objects described by nominal attributes and asymmetric binary attributes.

2. a) Write the meaning of the following rule.

   Age(X,"20..29")^income(X,"40000..49000") => buys(X,"computer")

   [support = 2% , confidence = 60%"]

   b) i) The data for the attribute age is given as

   13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45,46, 52, 70.

   Smooth the above data by using smoothing by bin with bin depth of 3. [3]

   ii) What do you understand by dispersion of data? Shown below are the sales data.

   27, 39, 3, 15, 43, 27, 19, 54, 65, 23, 45, 16

   - Find the five number summary
   - Draw box plot based on five number summary [4]

   c) Explain the different methods of handling missing values with suitable examples.

3. a) What are the measures used for data quality? Outline each with an example.

   b) i) Explain Snowflake Schema with an example. [3]

   ii) What is Outlier? How will you find? [4]

c) Use the methods below to normalize the following group of data:

200, 300, 400, 600, 1000

- min-max normalization by setting min = 0 and max = 1
- z-score normalization
- normalization by decimal scaling

4. a) What is sequence data? Give an example.

b) i) What is sequential pattern mining? [3]

ii) What is stream data processing? Outline any one method for stream data processing. [4]

c) Explain Generalized Sequential Pattern (GSP) mining with an example.

5. a) Give any two applications of Data Mining.

b) i) What is 0.632 bootstrap method for evaluating classifier accuracy? [3]

ii) What is k-fold cross validation? Mention its purpose. [4]

c) Outline the steps involved in ID3 algorithm.

/END/

FD/RL