Seat No: _____ Enrollment No: _____

PARUL UNIVERSITY

FACULTY OF ENGINEERING & TECHNOLOGY

B.Tech. Winter 2022 - 23 Examination

Semester: 7 **Date: 06/10/2022**
Subject Code: 203105431 **Time: 10:30 am to 01:00 pm**
Subject Name: Data Mining **Total Marks: 60**

Instructions:
1. All questions are compulsory.
2. Figures to the right indicate full marks.
3. Make suitable assumptions wherever necessary.
4. Start new question on new page.

Q.1 Objective Type Questions **(15)**

1. The type of relationship in star schema is _____.
   A) many-to-many. B) one-to-one. C) one-to-many. D) many-to-one.
2. The data is stored, retrieved & updated in _____.
   A) OLAP. B) OLTP. C) SMTP. D) FTP.
3. ETL stands for _____
4. The KDD Process Consists of _____ Step.
   A) five. B) two. C) three. D) four.
5. _____ is data about data.
6. Which of the following is generally used in finding hidden structure and patterns in a given unlabelled data?
   A) Supervised learning B) Unsupervised learning
   C) Reinforcement learning D) None of the above
7. The primary use of data cleaning is:
   A) Removing the noisy data B) Correction of the data inconsistencies
   C) Transformations for correcting the wrong data D) All of the above
8. KDD Stands for_____
9. Define Temporal Mining.
10. The left-hand side of an association rule is called _____.
    A) consequent. B) onset. C) antecedent. D) precedent.
11. Define Data Transformation
12. what is Attribute selection measures.
13. what is Cuboid?
14. when Discretization required.
15. Define Outlier.

Q.2 Answer the following questions. (Attempt any three) **(15)**

A) What are the reasons for missing values in real world data? Describe various methods for handling this problem.
B) explain support and confidence.
C) Define True Positive, True Negative, False Positive and False Negative
D) Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order): 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. (a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

Q.3 A) Explain data warehouse with its architecture. **(07)**

B) Trace the results of using the Apriori algorithm on the grocery store example with support threshold s=33.34% and confidence threshold c=60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets. Also indicate the association rules that are generated and highlight the strong ones, sort them by confidence **(08)**

Transaction ID	Items
T1	HotDogs, Buns, Ketchup
T2	HotDogs, Buns
T3	HotDogs, Coke, Chips
T4	Chips, Coke
T5	Chips, Ketchup
T6	HotDogs, Coke, Chips

OR

B) Suppose that a data warehouse for Parul University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and avg grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. **(08)**

(a) Draw a snowflake schema diagram for the data warehouse.

(b) Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Parul University student.

(c) If each dimension has five levels (including all), such as "student < major < status < university < all", how many cuboids will this cube contain (including the base and apex cuboids)?

Q.4 A) Explain KDD process with its neat diagram. **(07)**

OR

A) explain baye's theorem and naïve Bayesian classification. Why naïve Bayesian classification is called "naive"? **(07)**

B) What is preprocessing? Explain with its different stages. **(08)**