Roll No……..

Dr B R Ambedkar National Institute of Technology, Jalandhar

B Tech IV Semester (Information Technology)

ITPC–208, Data Mining and Warehousing,
End-Semester Examination, May-2024

Duration: 03 Hours Max. Marks: 50 Date: 22 May 2024

Marks Distribution & Mapping of Questions with Course Outcomes (COs)					
Question Number	1	2	3	4	5
Max. Marks	10	10	10	10	10
CO No.	1	2	3	4	5
Cognitive Level	R	U	An	Ap	Ap
**Section/Chapter/Unit	-	-	-	-	

Note:

1. Attempt all the questions. Answer the subparts of a question together.
2. Non-programmable calculator is allowed.

1. Answer the following: [1 + 3+ 3 +3 = 10 Marks]

a) Explain properties of Normal Distribution Curve.

b) Describe online analytical processing (OLAP) servers, their purpose, and their types in brief.

c) Outline the OLAP operations with real world application scenarios.

d) State the difficulties in data warehouse tuning and detail the usage of different tuning queries with examples.

2. Answer the following: **[3+ 7 = 10 Marks]**

a) Outline the Knowledge discovery (KDD) process and describe the KDD process from typical ML and statistics community viewpoint.

b) In order to create a beverage, salt and sugar are dissolved in water in a specific ratio. Based on these two ingredients, the beverage is classified into two categories: 'Good' or 'Bad'. The dataset for such scenario is given below:

Beverage Id	Salt	Sweet	Result
1	10	10	Bad
2	10	7	Bad
3	6	7	Good
4	4	8	Good
5	3	6	Good

You are required to utilize the K-Nearest Neighbors classifier to determine the category of a beverage where the ratio of salt to sugar is 5 and 7, respectively. Assume K = 4 while determining the category.

3) Answer the following**: [3 + 7 = 10 Marks]**

a) Define Attributes and overview different attribute types in brief with examples.

b) Implement K-Medoids clustering by initially selecting points D2 and D8 as the medoids, and calculating the absolute error. Utilize the Manhattan distance metric to measure the dissimilarity between objects. After forming the initial clusters, replace the medoid D8 with point D7. Now, Recompute the absolute error with this swapped medoid, and determine whether choosing D7 instead of D8 as the medoid results is a better choice or not with brief justification? for the following data points with K=2:

Data Points	X	Y
D1	1	5
D2	2	3
D3	2	7
D4	3	6

D5	5	1
D6	5	3
D7	6	2
D8	6	3
D9	7	4
D10	6	5

4) Answer the following: [3 + 7 = 10 Marks]

Imagine you're hired for a CTC of 56 LPA for the role of data analyst working for a popular ride-sharing company. Your company is involved in collecting vast amounts of data on pick-up and drop-off locations of passengers throughout Delhi-NCR region with following locations: A(3, 7), B(4, 6), C(5, 5), D(6, 4), E(7, 3), F(6, 2), G(7, 2) and H(8, 4).

a) State the reason that traditional clustering approaches like manually defining cluster boundaries or using distance-based algorithms like K-Means be inadequate for analyzing the distribution of pick-up and drop-off locations in a ride-sharing scenario? Explain the challenges and limitations of such methods when dealing with irregular cluster shapes, varying densities, and the presence of noise or outliers in the data.

b) Now, you are tasked with the goal to identify high-density areas (core points) where there is a high demand for rides and (border points) that would be areas with moderate demand, as well as areas with low demand (outlier). You decided to apply DBSCAN algorithm for the task considering Eps (ε) = 2.5 and MinPts = 3. Also, briefly state how adjusting epsilon (ε) and MinPts parameters affects the clustering results in terms of capturing high-density, moderate-density, and sparse regions.

5) Answer the following: [10 Marks]

Suppose, impressed by the diverse skills and talent of the 120 students of your batch, the HoD of Department set a 28 crores cumulative CTC target for your batch and directs the TPO to categorize companies as Product-based and Service-based, identifying required skills for each. The TPO believed understanding the relationship between skills and the likelihood of job offers from specific company categories could provide valuable insights. Consequently, TPO has collected data on the previous year's placement records with following insights: a) 120 students had programming skills but only 40 students having programming skills were placed in Product-based companies. b) 120 students had both data analysis skills and domain knowledge out of which 80 students having both data analysis skills, and domain knowledge, were placed in service-based companies; for the attributes: Student ID, Programming skills (Yes/No), Data analysis skills (Yes/No), Communication skills (Yes/No), Domain knowledge (Yes/No), Placed company category (Product-based/Service-based), Package (Product-based (32 lakhs) and Service-based (19 lakhs)). Now, answer the following:

a) If a student has programming skills, what is the probability (confidence) that they will be placed in a Product-based company? Also, if a student has Data analysis skills and Domain knowledge, what is the probability (confidence) that they will be placed in a Service-based company?

b) Based on the analysis of support and confidence, identify the most crucial skills required for placement in Product-based and Service-based companies.

c) Assume you possess the following skills: Programming Skills (Yes), Data Analysis Skills (No), Communication Skills (Yes), and Domain Knowledge (No). Now, based on the insights gained from the analysis, predict your potential package (32 lakhs or 19 lakhs) corresponding to whether you are placed in a Product-based company or a Service-based company.