Roll No……..

Dr B R Ambedkar National Institute of Technology, Jalandhar

B Tech 3rd Semester (Information Technology)

ITPC–208, Data Mining and Warehousing
Mid-Semester Examination, March-2024

Duration: 02 Hours Max. Marks: 30 Date: 20 March 2024

Marks Distribution & Mapping of Questions with Course Outcomes (COs)				
Question Number	1	2	3	4
Max. Marks	8	8	6	8
CO No.	1	2	3	4
Cognitive Level	R	U	An	Ap
**Section/Chapter/Unit	-	-	-	-

Note:

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

1. Answer any four of the following: [2 X 4 = 8 Marks]

A) Name any five Data Mining function.
B) Define Data Objects with example.
C) Differentiate Discrete and Continuous Attribute.
D) Explain properties of Normal Distribution Curve.
E) Boxplot often tell more than Histogram. (True or False). Justify.
F) Name the parameters that represents five number summary to measure dispersion of data.

2. Answer any four of the following: [2 x 4 = 8 marks]
a) Explain the major tasks performed during Data Preprocessing in brief and suggest some ways to handle noisy data.

b) Suppose two stocks A and B have the following values in one week: (2, 5), (3, 8), (5, 10), (4, 11), (6, 14). If the stocks are affected by the same industry trends, will their prices rise or fall together? Explain in detail.
c) Explain the necessity for data reduction. Suggest some data reduction strategies.
d) Ellyse Perry has scored the following runs (4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34) in the last 12 matches of the Women's Premier League while representing the Royal Challenger Bangalore Women's Team. The coach has now instructed the team analyst (you) to apply a discretization method for data (last 12 match scores) smoothing.
e) The salary of Data Analyst at reputed multinational company falls in between $12000 and $98000 per year with mean salary of $54000 and standard deviation of $16,000. Apply min-max normalization and Z-score normalization for the salary $73,600.

3) Answer the following: [6 Marks]
As the Chief Election Commissioner, your mission is to categorize various cities into three distinct clusters using K-means clustering. The cities (represented as Cartesian points) are follows: Patiala (2,10); Varanasi (2,6); Hyderabad (11,11); Jalandhar (6,9); Gorakhpur (6,4); Prayagraj (1,2); Phagwara (5,10); Amritsar (4,9); Amravati (10,12); Patna (7,5); Bengaluru (9,11); Ayodhya (4,6); Kartarpur (3,10); Ludhiana (3,8); Raipur (6,11). Let us suppose that each cluster will have the same polling date. Answer the following:
A) Whether Kartarpur and Ayodhya will indeed share a same polling date?
B) Name the cities that will have the same polling date as Varanasi?
C) Name the cities having same polling date as Hyderabad?

4) Answer any two of the following: [4 + 4 = 8 Marks]
a) Consider a dataset containing information about IPL teams (KKR, MI, CSK, LSG) and their performance in previous seasons, including factors such as total wins, batting average, bowling average, and current form. Using a decision tree classifier, predict the team that is most likely to win IPL 2024 (The root node attribute will decide the winner. For example, if batting average/total win/current form becomes the root node then team having highest probabilities corresponding to this attribute will become the winner. In case of

bowling average becoming root node, team having lowest bowling average will become the winner) The performance details (probabilities) are as follows:

Team	Total Wins	Batting Average	Bowling Average	Current Form
KKR	0.25	0.20	0.15	0.30
MI	0.30	0.25	0.20	0.35
CSK	0.28	0.22	0.18	0.32
LSG	0.17	0.33	0.47	0.03

B) For the following table details, calculate the following:
i) Entropy for each Attribute (Gender, Shirt Size, Car Type) ii) Gini for each Attribute (Car type, Shirt Size, Gender)

Cust ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

C) For the following table, specify which attribute between a1 and a2 is best split attribute and why? Also, calculate the values of Gini index for a1 and a2; and information gain for a1 and a2.

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−