Seat No.: _____ Enrolment No. _____

NATIONAL FORENSIC SCIENCES UNIVERSITY

M.Tech. Artificial Intelligence and Data Science

Semester – I – January - 2024

Date: 17/01/2024

Subject Code: CTMTAIDS SI P4

Subject Name: Fundamentals of Data Science and Machine Learning

Total Marks: 100

Time: 11:00 AM to 2:00 PM

Instructions:

1. **Write down each question on separate page.**
2. **Attempt all questions.**
3. **Make suitable assumptions wherever necessary.**
4. **Figures to the right indicate full marks.**

			Marks
Q.1	(a)	Both k-mean and k-medoids algorithm can perform effective clustering. a. Illustrate the strength and weakness of k-means in comparison medoids b. Illustrate the strength and weakness AGNES clustering schemes	**08**
	(b)	Determine the Jaccard distance, cosine similarity, and Euclidean distance between the vectors x= (1,1,1,1) and y= (2,2,2,2). Provide definitions for each of these distance measures.	**08**
	(c)	Explain the frequent itemset generation and rule generation in Apriori algorithm.	**08**
		OR	
	(c)	Visualization additionally gives you sense of data distribution and relationships among variables. Explain different ways of visualizing data.	
Q.2	(a)	Consider the following dataset for k=2. (3,9), (5,4), (8,7), (9,3), (10,6), (6,8), (4,5). (7,9), (11,5), (12,6).	**08**
	(b)	As a quality control manager in a chocolate factory with three production lines as A, B, and C, instruct the team to construct a decision tree to identify which line is producing counterfeit chocolates and whether they are slightly heavier or lighter than the genuine ones.	**08**
	(c)	Elaborate on the various categories of regression techniques, providing insights into the distinct characteristics of each.	**08**
		OR	
	(c)	How do 'Navies' contribute to Naive Bayesian Classifiers, and what are the strengths associated with employing the Naive Bayes Classifier?	

Q.3 **(a)** Provide an explanation on effectively visualizing hierarchical data with negative values. 08

(b) Given an animal database X: 08

TID	Items
001	Cat, Dog, Frog, Goat
002	Ant, Bat, Cat, Dog
003	Ant, Cat, Dog, Frog
004	Cat, Dog, Elephant, Goat, Ant
005	Ant, Dog, Frog, Bat
006	Bat, Cat, Goat
007	Dog, Frog, Goat
008	Ant, Bat, Goat

Using the threshold values support = 25% and confidence = 60%. find:

i. All the frequent itemsets in database.
ii. Strong association rules for database.

(c) Discuss the challenges associated with the fundamental k-Nearest Neighbour algorithm. 08

OR

(c) Provide concise real-world examples illustrating the applications of clustering, classification, and association rule mining in various contexts.

Q.4 **(a)** How should missing data be addressed in the data cleaning process? 04

(b) Consider the following set of training examples: 05

Instances	Classification	A1	A2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

a. What is the entropy of this collection of training examples with respect to the target function classification?
b. What is the information gain of a2 relative to these training examples?

(c) How can datasets be summarized and what are the various methods employed for this purpose? 05

OR

(c) For different types of data, calculation of different correlation coefficients is known. Provide a brief with example for each type of data following correlation coefficient are applicable.

a. Charles' Spearman's correlation coefficient (rr).
b. Karl Pearson's coefficient of correlation ($rr*$).
c. Chi-square coefficient of correlation ($\chi\chi 2$).

Q.5 (a) Discuss about GIS Data Visualizations. 04

(b) Consider the following data for price attribute: 05
{4,8,9,15,21,21,24,25,26,28,29,34}. Partition the same into bins using:

a. Equi-depth binning.
b. Smoothing by bin means.
c. Smoothing by bin boundaries.

(c) Explain how web social networks can be extracted and analyzed 05

OR

(c) How to handle the training examples with missing attribute values and differing costs in a decision tree learning

--- Best **Of Luck** ---