Institute of Information Technology
Jahangirnagar University
Professional Masters in IT (PMIT) Final Examination, Summer 2023
Course Code: PMIT-6307, Course Title: Data Mining and Knowledge Discovery

Total Marks: 60 Time: 3 Hours

Answer any **five (5)** of the following questions. The figures at the right indicate the marks.

1. a) Define data mining. Describe the steps involved in data mining when viewed as a process of knowledge discovery. 4

   b) Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied. 4

   c) What is the curse of dimensionality? Illustrate the steps of **Principal Components Analysis (PCA)**. 4

2. a) The following attributes are measured for members of a herd of Asian elephants: weight, height, tusk length, trunk length, and ear area. Based on these measurements, what sort of proximity measure would you use to compare or group these elephants? Justify your answer. 3

   b) Given two vectors, x [20, 2, 40, 10] and y [18, 1, 34, 8]; calculate the following distances: 4
      i. Manhattan distance;
      ii. Minkowski distance using p=3;

   c) Imagine that you're studying the relationship between newborns' weight (W) and length (L). You have the following weights (in kg) and lengths (in cm) of the 10 babies born last month at a local hospital. Calculate the **Pearson correlation coefficient** for the given data. Also, interpret the result. 5

W	3.63	3.02	3.82	3.42	3.59	2.87	3.0	3.46	3.36	3.30
L	53.1	49.7	48.4	54.2	54.9	43.7	47.2	45.2	54.4	50.4

3. a) Consider the following output for a dog classification problem generated by a classification algorithm. Compute accuracy and F_score for this algorithm. 3

Actual	Dog	Cat	Dog	Dog	Cat	Dog	Dog	Dog	Cat	Cat
Predicted	Dog	Dog	Dog	Dog	Cat	Dog	Cat	Dog	Cat	Dog

   b) Generate a decision tree from the following training dataset considering the **ID3** algorithm: 9

Home Owner	Marital Status	Rich	Cheat
Yes	S	Yes	No
No	M	Yes	No
No	S	No	No
Yes	M	Yes	No
No	D	Yes	Yes
No	M	No	No
Yes	D	Yes	No
No	S	Yes	Yes
No	M	No	No
No	S	Yes	Yes

4. a) Illustrate the process of the **3-nearest neighbor's** algorithm with an example. 3

b) Consider a situation where you have 1000 fruits which are either 'banana' or 'apple' or 'other'. The possible attributes of any fruit are: [Long, Sweet, and Yellow]. The training dataset and summary of the training dataset will look like Table 1 & Table 2. Now, consider a case where you're given that a fruit is long, sweet, and yellow; and you need to predict the type of that fruit using the **Naive Bayes** classifier. 6

Table 1: Training Dataset

Long	Sweet	Yellow	Fruit
0	0	1	Apple
1	0	1	Banana
0	1	0	Apple
1	1	1	Other
..	..	..	..

Table 2: Summary

Type	Long	Not Long	Sweet	Not sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Apple	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	**500**	**500**	**650**	**350**	**800**	**200**	**1000**

c) Consider a training set that contains 100 positive examples and 400 negative examples. For each of the following candidate rules, determine the best and worst candidate rule according to rule accuracy. 3

i. R1: A→+(covers 4 positive and 1 negative examples);
ii. R2: B→+(covers 30 positive and 10 negative examples);
iii. R3: C→+(covers 100 positive and 90 negative examples);

5. a) An intrusion detection dataset consists of 10,000 normal and 10 anomalous sequences. Describe the class imbalance problem for this dataset. How do you mitigate this problem? 4

b) Demonstrate the working process of the **Support Vector Machine (SVM)** algorithm. 4

c) Draw a basic architecture of an **Artificial Neural Network (ANN)**. Write down the steps involved in the process of training an ANN. 4

6. a) Choosing the optimal number of clusters is a big task in the K-means clustering algorithm. How can you choose the value of K? 3

b) You have to cluster the following eight points into three clusters: A1 (2, 10), A2 (2, 5), A3 (8, 4), A4 (5, 8), A5 (7, 5), A6 (6, 4), A7 (1, 2), A8 (4, 9). Initial cluster centers are A1, A4, and A7. Use the **K-means** Algorithm to find the three cluster centers after the second iteration. 5

c) Consider the age of six persons attending a program: 12, 16, 19, 36, 21, and 37. Apply the **Agglomerative Clustering** algorithm with the simple difference of age to build the hierarchical clustering dendrogram of the persons. 4

7. a) Given a set of transactions, T contains 20 data items. To find all the rules having greater support and confidence than the corresponding thresholds, a **brute-force approach** for mining association rules cannot be used. Why? 3

b) Why do many association rule mining algorithms decompose the problem into two major subtasks by decoupling the support and confidence requirements? Briefly describe the two-step approach with an example. 3

c) Suppose we have the following dataset having various transactions. From this dataset, we need to find the frequent itemsets and generate the association rules using the **Apriori** algorithm for a minimum support of 50% and minimum confidence of 70%. 6

Transaction	List of Items
T1	A, B, C
T2	B, C, D
T3	D, E
T4	A, B, D
T5	A, B, C, E
T6	A, B, C, D