ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)
ORGANISATION OF ISLAMIC COOPERATION (OIC)
Department of Computer Science and Engineering (CSE)

SEMESTER FINAL EXAMINATION   **WINTER SEMESTER, 2022-2023**
DURATION: 3 HOURS   **FULL MARKS: 150**

CSE 4739: Data Mining

Programmable calculators are not allowed. **Do not write anything on the question paper.**
Answer **all 6 (six)** questions. Figures in the right margin indicate full marks of questions whereas
corresponding CO and PO are written within parentheses.

1. a) A dataset consisting of objects A, B, C, D, E, F, and G with the distance matrix in Table 1 is   10
       given: (CO3)
                                                                                                              (PO2)

    Table 1: Distance matrix for Question 1. a)

Distance	A	B	C	D	E	F	G	H
A	0	3	3	4	9	10	8	6
B		0	3	7	8	9	8	7
C			0	6	6	6	3	6
D				0	14	15	7	7
E					0	4	2	6
F						0	4	7
G							0	6
H								0

    Assume DBSCAN to be run for this dataset with MINPOINTS=3 and epsilon=5. How many
    clusters will DBSCAN return and what are these clusters? Which objects are outliers and
    border points in the clustering result obtained earlier? Give reasons for your answers.

    b) Describe each of the following clustering algorithms in terms of the following criteria: Shapes   3 × 3
       of clusters that can be determined, input parameters that must be specified, and limitations.   (CO1)
       i. k-means   (PO1)
       ii. BIRCH
       iii. DBSCAN

    c) Clustering is recognized as an important data mining task with broad applications. Give one   3 + 3
       application example for each of the following cases: (CO1)
       i. An application that uses clustering as a major data mining function.   (PO1)
       ii. An application that uses clustering as a preprocessing tool for data preparation for other
           data mining tasks.

2. a) Show conditions under which density-based clustering is more suitable than partitioning   7
       and hierarchical clustering. Give application examples to support your argument.   (CO1)
                                                                                                              (PO1)

    b) When using an agglomerative clustering method or a divisive clustering method, a core need   10
       is to measure the distance between two clusters. Compute the distance between clusters A   (CO2)
       (1,6,2,5,3) and B (3,5,2,6,6) using four distance measures.   (PO1)

c) The following is a set of one-dimensional points: 1, 1, 2, 3, 5, 8, 13, 21, 33, 54. Perform two iterations of k-means on these points using the two initial centroids 0 and 11.

8
(CO2)
(PO1)

Table 2: Data with Ground Class and Classifier Prediction Probabilities for Question 3. a)

Tuple #	Class	Probability
1	P	0.95
2	N	0.85
3	P	0.78
4	P	0.66
5	N	0.60
6	P	0.55
7	N	0.53
8	N	0.52
9	N	0.51
10	P	0.40

3. a) The data tuples and probability value, as returned by a classifier is shown in Table 2. For threshold values 0.8, 0.6, 0.4, and 0.2 compute the values for the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). Use them to compute the true positive rate (TPR) and false positive rate (FPR) and plot the ROC curve for the data

10
(CO3)
(PO1)

b) Describe the following in short:
   i. Naive Bayes is an exceedingly simple algorithm. Name one reason why it can still be effective for something like spam classification.
   ii. The stopping criteria in the K-means algorithm.
   iii. Difference between Classification and Clustering

3 × 2
(CO1)
(PO1)

c) Show the dendrogram for agglomerative hierarchical clustering using complete linkage for the points in Table 3.

9
(CO2)
(PO1)

Table 3: Pair-wise distances between 4 points for Question 3. c)

Distance	A	B	C	D
A	0.00	0.08	0.65	0.78
B		0.00	0.39	0.56
C			0.00	0.63
D				0.00

4. a) What are the advantages and disadvantages of the FP (Frequent Pattern) growth algorithm?

8
(CO3)
(PO3)

b) Use the methods below to normalize the following group of data:
200, 300, 400, 600, 1000
   i. min-max normalization by setting $min = 0$ and $max = 1$.
   ii. z-score normalization.

2 × 4
(CO1)
(PO1)

c) Illustrate the Apriori principle. Show how the partitioning method is applied to improve the Apriori algorithm.

<div align="right">3 + 6
(CO2)
(PO2)</div>

5. a) SVM is posed as a constrained optimization problem having an objective function and a constraint to fit in the Lagrangian equation. Show how you get the objective function and the constraint.

<div align="right">9
(CO2)
(PO1)</div>

b) Assume a scenario where 300 out of 1000 of the emails are categorized as spam. The probabilities of "buy", "computer", "won", "faculty", and "meeting" to show up in a spam email are 0.23, 0.1, 0.85, 0.01, and 0.05 respectively. Suppose we have an uncategorized email that contains the words "computer", "buy", and "meeting", and does not contain "won" or "faculty". Answer the following questions:

i. Why is naive Bayesian classification called "naive"?

ii. Determine the probability that this exact set of words shows up in an email that is known to be spam.

iii. If we have determined that the probability that the exact combination of words shows up in any email is 0.02, what is the probability that an uncategorized email containing those words is spam?

<div align="right">3 x 3
(CO2)
(PO1)</div>

c) Explain in brief how Random Forest works.

<div align="right">7
(CO1)
(PO1)</div>

Table 4: Dataset for Question 6. a)

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

6. a) Draw the full decision tree that would be learned for the dataset in Table 4 using information gain. Show necessary calculation steps

<div align="right">10
(CO2)
(PO1)</div>

b) Answer the following briefly:

i. How does a Decision Tree handle continuous(numerical) features?

ii. Do we require Feature Scaling for Decision Trees?

<div align="right">2 x 4
(CO1)
(PO1)</div>

c) Why do we require Pruning in Decision Trees? Explain.

<div align="right">7
(CO1)
(PO1)</div>