Daffodil International University

Faculty of Science & Information Technology

Department of Computer Science & Engineering

Mid Examination, Spring 2025

Course Code: CSE325, Course Title: Data Mining and Machine Learning

Level: 3 Term: 2 Batch: 62

Time: 01:30 Hrs **Marks: 25**

Answer ALL Questions

[The figures in the right margin indicate the full marks and corresponding course outcomes. All portions of each question must be answered sequentially.]

1. Predict whether a student will Pass an exam based on Hours Studied and Attendance. **[10]** **CO2**

Example	Hours Studied	Attendance	Pass
1	High	Good	Yes
2	High	Good	Yes
3	High	Poor	Yes
4	High	Poor	No
5	Low	Good	Yes
6	Low	Good	No
7	Low	Poor	No
8	Low	Poor	No

(i) Calculate the overall entropy of the dataset for the Pass attribute.
(ii) For each attribute (Hours Studied and Attendance):
- Calculate the entropy of each subset.
- Compute the weighted average entropy for the attribute.

(iii) Identify which attribute gives the highest information gain and explain why it should be chosen as the root of the decision tree.
(iv) Draw the final decision tree using the selected attribute, showing the branches and final classifications.

2. A hospital has developed an automated system to detect diabetic retinopathy from patient images. The system was tested on 800 images. Out of these: (a) 250 images are known to be positive for diabetic retinopathy, (b) The system correctly identified 200 positive cases, and (c) It incorrectly flagged 70 healthy images as positive. Now answer the below questions. **[5]** **CO1**

(i) Construct the confusion matrix for this binary classification problem.
(ii) Calculate the precision and recall for the positive (disease) class.

3.	A retail company wants to predict whether a customer will make a purchase ("Yes" or "No") based on two features: **Age** and **Annual Income** (in thousands). Using K=3, classify a new customer with **Age = 33** and **Annual Income = 52** using the following dataset: **(i)** Calculate the Euclidean distances between the new customer (33,52) and each customer in the dataset. **(ii)** Identify the three nearest neighbors and determine the predicted class by majority vote.	[5]	CO2
4.	Consider the following dataset with two features, X_1 and X_2: Using K=2 and the following initial centroids: (a) Centroid 1: Data Point 1 (1.0,2.0) and (b) Centroid 2: Data Point 4 (8.0,8.0). Now answer the below questions. **(i)** Compute the Euclidean distance between each data point and both centroids. Then, assign each data point to the nearest centroid. **(ii)** Calculate the new centroids for each cluster by taking the average of the X_1 and X_2 values of the data points assigned to that cluster.	[5]	CO2

Customer ID	Age	Annual Income (k)	Purchase?
1	25	40	Yes
2	30	50	Yes
3	22	20	No
4	35	60	Yes
5	40	70	No
6	28	30	No
7	32	55	Yes
8	45	80	No

Data Point	X_1	X_2
1	1.0	2.0
2	1.5	1.8
3	5.0	8.0
4	8.0	8.0
5	1.0	0.6
6	9.0	11.0