Daffodil International University

Faculty of Science & Information Technology

Department of Computer Science and Engineering

Mid-Semester Examination, Fall-2024

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

Level: 3 Term: 2 Batch:61

Exam Duration: 1.5 Hours **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. a) You are working on a multi-class classification problem with three classes (A, B, C). After testing your model, you get the following confusion matrix: **[5]** **CO1**

	Predicted A	Predicted B	Predicted C
Actual A	50	10	5
Actual B	8	60	12
Actual C	5	15	55

Analyze the confusion matrix and calculate the F1-Score, specificity, and as well as the error rate of the classifier.
Based on your analysis, what are the strengths and weaknesses of this model?

2. a) A dataset contains information on customer spending in a retail store, with the following features: **[5]** **CO2**

ID	Age	Income (k$)	Spending Score	Label (Class)
1	25	30	45	Low
2	35	60	80	High
3	23	45	50	Medium
4	45	70	85	High
5	30	40	55	Medium
6	55	85	90	High

Given the provided dataset, apply the KNN algorithm with k=5 to classify the new customer with the following characteristics:
Age = 50,Income = 65k,Spending Score = 70

		Calculate the distances between this new customer and each of the other customers in the dataset using the Euclidean distance formula. Assign the customer to the class based on the majority vote of the 5 nearest neighbors.		
3	*a)*	**Analyze** the impact of using the dicing operation on a large data cube that stores sales data. How would this operation help focus analysis on a specific state and product category?	[5]	CO2
4.				CO3

ID	Age	Income	Student	Credit Rating	Buys Car
1	<=30	High	No	Fair	No
2	<=30	High	No	Excellent	No
3	31-40	High	No	Fair	Yes
4	>40	Medium	No	Fair	Yes
5	>40	Low	Yes	Fair	Yes
6	>40	Low	Yes	Excellent	No
8	31-40	Low	Yes	Excellent	Yes
9	<=30	Medium	No	Fair	No
10	<=30	Low	Yes	Fair	Yes
11	>40	Medium	Yes	Fair	Yes
12	<=30	Medium	Yes	Excellent	Yes
13	31-40	Medium	No	Excellent	Yes
14	31-40	High	Yes	Fair	Yes
15	>40	Medium	No	Excellent	No

	a)	***Evaluate*** the performance of the ID3 decision tree when applied to the provided dataset. After constructing the tree, test it on unseen examples and compare the accuracy of the tree before and after pruning.	[8]
	b)	***Summarize*** the advantages and disadvantages of using the ID3 algorithm for small datasets like this one.	[2]