Department of Computer Systems Engineering,
University of Engineering and Technology, Peshawar,
Pakistan

Midterm Exam (8th, Spring 2024) | Paper: Artificial Intelligence (CSE-407)
Time: 2 Hours | **Marks:** 30

Note: Attempt all questions on the answer sheet.
Any other instructions, if required.

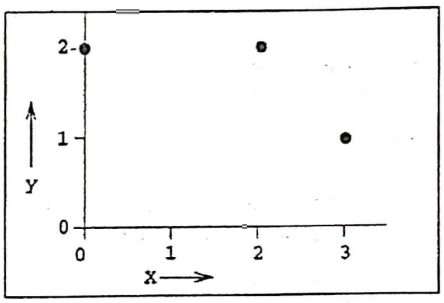
Question No. 1 (Marks=2, CLO-1)

Suppose that $X1, \ldots Xm$ are categorical input attributes and Y is categorical output attribute. Suppose we plan to learn a decision tree without pruning, using the standard algorithm.

(True or False) If Xi and Y are independent in the distribution that generated this dataset, then Xi will not appear in the decision tree. explain your answer.

Question No. 2 (Marks=5, CLO-2)

Suppose you have this data set with one real-valued input and one real-valued output:



X	Y
0	2
2	2
3	1

a. What is the R^2 value?
b. Suppose we use a trivial algorithm of predicting a constant $y = c$. What is the mean squared error in this case? (Assume c is learned from the non-left-out data points.)

Question No. 3 (Marks=5, CLO-2)

Pakistan International Airlines has developed 2 different classifiers (A and B) for the prediction whether a flight originating from Peshawar will arrive at its final destination on time or not. True or Positive here is 'On time' and it refers to the case when the flight is no more than 5 minutes late than the scheduled time. The classifiers were tested on a data-set of 500 flights, and the results are as follows:

	Actual	
	On time	Late
Classifier A, predicted on time	131	155
Classifier A, predicted late	19	195
Classifier B, predicted on time	82	72

Classifier B, predicted late	68	278

a) Construct confusion matrix for both the classifiers?
b) Which is the preferable classifier in terms of accuracy?

Question No. 4 (Marks=5, CLO-3)

Compute K Nearest Neighbor Algorithm

Hints:

a. Determine parameter K number of nearest neighbors.
b. Calculate the distance between the query-instance and all the training samples.
c) Sort the distance and determine nearest neighbors based on the K^{th} minimum distance.
d) Gather the category of the nearest neighbors.
e) Use simple majority of the category of nearest neighbors as the prediction value of the query instance.

X1= acid durability (seconds)	X2= Strength (kg/sq meter)	Y = classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with x1=3 and x2= 7. Without another expensive survey, can we guess what the classification of this new tissue is?

Question No. 5 (Marks=10, CLO-3)

A second-hand car dealer has 10 cars for sale. She decides to investigate the link between the age of the cars, x years, and the mileage, y thousand miles. The data collected from the cars are shown in the table below.

Age, x (years)	2	2.5	3	4	4.5	4.5	5	3	6	6.5
Mileage, y (thousands)	22	34	33	37	40	45	49	30	58	58

[You may assume that $\sum x = 41, \sum y = 406, \sum x^2 = 1, \sum xy = 1818.5$]

a. Find Sxx and Sxy
Hints $S_{xx} = \sum x^2 - (\sum x)^2 / n$ $\quad S_{xy} = \sum xy - (\sum x \sum y)/n$
b. Find the equation of the least squares regression line in the form $y = ax + b$. Give the values of a and b to 2 decimal places.
c. Give a practical interpretation of the slope b
d. Using your answer to part (b), find the mileage predicted by the regression line for a 5-year-old car.