National University of Computer and Emerging Sciences, Lahore Campus

Course:	Data Mining	Course Code:	DS3002
Program:	BS (Data Science)	Semester:	Fall 2025
Date:		Total Marks:	50
Section:	BDS-6A,B	Submission Date:	
Assignment:	1		

*

No late submissions are allowed.

Show all steps and calculations in your work to earn full points. Partial solutions will not be accepted.

If you encounter any issues or have questions, **reach out early** for clarification.

Plagiarism is strictly prohibited.

Question 1 (15)

A research study analyzes multiple factors' impact on students' exam scores. The following sample data represents the hours a student studied, practiced questions attempted, and slept, and their corresponding exam scores.

Study Hours	Practice Questions Attempted	Sleep Hours	Exam Marks
1.5	5	8	45
3.5	15	6	60
7	40	5	88
5.5	30	7	76
9.5	50	4	98
6.5	35	6	83
8	45	5	91

a) Identify the independent variables (X1,X2,X3) and the dependent variable (Y).

b) Compute the Pearson correlation coefficient for each independent variable concerning the dependent variable

c) Determine which factor has the strongest correlation with the dependent variable.

d) Interpret the result and explain whether an increase in study hours impacts exam performance.

e) If a student studied for 6 hours, attempted 25 practice questions, and slept 7 hours, estimate their expected exam score based on correlation trends. Does it align with the overall trend in the dataset?

Question 2 (15)

Given the following dataset, determine which feature is most useful for predicting the outcome:

Day	Geographic Region	Temperature	Humidity	Wind	Outcome
D1	A	Hot	High	Weak	No
D2	A	Hot	High	Strong	No
D3	B	Hot	High	Weak	Yes
D4	C	Mild	High	Weak	Yes
D5	C	Cool	Normal	Weak	Yes
D6	C	Cool	Normal	Strong	No
D7	B	Cool	Normal	Strong	Yes
D8	A	Mild	High	Weak	No
D9	A	Cool	Normal	Weak	Yes
D10	C	Mild	Normal	Weak	Yes
D11	A	Mild	Normal	Strong	Yes
D12	B	Mild	High	Strong	Yes
D13	B	Hot	Normal	Weak	Yes
D14	C	Mild	High	Strong	No

a) Calculate the entropy for the dataset based on the Outcome.

b) Compute the information gain for each feature (Geographic Region, Temperature, Humidity, and Wind) with respect to the Outcome.

c) Based on the information gain, identify the best feature for splitting the dataset and explain why it is the most useful feature for predicting the Outcome.

d) After the first split, identify the next best feature (from the remaining ones) for further splitting the dataset and explain why it is a good choice.