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 data analyst at FAST University is investigating the relationship between the annual salaries of AI professors (Y, in thousand dollars) and three academic performance indicators:

- Research Quality Score (X_1) – an index measuring research quality based on peer reviews.
- Years of Teaching Experience (X_2) – total years of teaching at the university level.
- Publication Impact Index (X_3) – a metric evaluating the impact of research publications based on citations and journal rankings.

X_1 (Quality Score)	X_2 (Experience in Years)	X_3 (Publication Impact)	Y (Salary in $1000s)
5.1	8	6.5	45.2
6.3	15	7.8	52.4
4.7	5	5.9	38.1
7.2	12	8.2	55.6
5.8	10	7.1	48.3

Using this dataset, calculate the Pearson correlation matrix between all variables (X_1, X_2, X_3, Y) and present your results as a correlation matrix table.

Question 2 (10)

Answer the questions based on the results from Q1

a) Identify the independent and dependent variables.
b) Which independent variable correlates strongly with salary (Y)?
c) What does this suggest about salary trends at FAST University?
d) Which independent variable has the weakest correlation with salary? Does this mean the variable does not affect salary? Explain.
e) A professor claims that teaching experience (X_2) is the best salary predictor. Based on your correlation matrix, do you agree? Why or why not?
f) What does the correlation matrix suggest about the relationship between Publication Impact (X_3) and salary?
g) Is there a strong correlation between Research Quality Score (X_1) and Publication Impact (X_3)?
h) If a professor's publication impact index increases significantly, how would it likely affect their salary? What does this tell you about how these two factors relate?

Question 3 (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.

Question 4 (10)

Based on the feature selected in Question 3 (the one with the highest information gain), perform the following tasks:

a) Using the feature with the highest information gain from Question 3, draw the first level of the decision tree. Based on this feature, determine if it is useful for predicting the outcome.
b) 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.
c) After splitting by the chosen feature, describe the outcome distribution (Yes/No) in each resulting sub-group.
d) Which feature, Temperature or Wind, is more important for predicting the outcome? Justify your answer and explain which feature you would choose as the root feature in the decision tree.
e) If a new feature, Dietary Habits, were added to the dataset, how would you assess its usefulness for the decision tree?