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:	30
		Submission Date:	
Section:	BDS-6A,B		
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

a) 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.

b) Identify the independent and dependent variables.

c) Which independent variable correlates strongly with salary (Y)?

d) Which independent variable has the weakest correlation with salary? Does this mean the variable does not affect salary? Explain.

Question 2 (15)

A fitness app wants to predict whether a person will achieve their weight loss goal based on their exercise and dietary habits. The following dataset categorizes user habits into categorical values:

User	Exercises Daily	Follows Diet Plan	Caloric Intake	Water Intake	Achieved Goal
1	Yes	No	High	Low	Yes
2	No	Yes	Medium	Medium	No
3	Yes	Yes	Low	High	Yes
4	No	No	High	Low	No
5	Yes	Yes	Low	High	Yes
6	No	Yes	Medium	Low	No
7	Yes	No	High	Low	No
8	No	Yes	Low	High	Yes
9	Yes	Yes	Low	High	Yes

a) Compute the information gain for the categorical feature "Exercises Daily?" and determine whether it is a strong predictor of weight loss success.

b) Compute the information gain for "Water Intake" and "Caloric Intake" and determine which factor is the most important in predicting weight loss.

c) Based on the dataset trends, determine the likelihood of achieving a weight loss goal for a person who does not exercise daily, consumes low calories, and drinks high amounts of water.

d) Using the feature with the highest information gain, draw the first level of a decision tree. Evaluate whether this feature is useful for predicting weight loss success.

e) After splitting by the best feature, describe the outcome distribution (Yes/No) in each resulting subgroup and analyze the patterns.