POKHARA UNIVERSITY

Level: Bachelor
Programme: BE
Course: Data Science and Analytics (New)

Semester: Spring

Year : 2025
Full Marks : 100
Pass Marks : 45
Time : 3 hrs.

Candidates are required to give their answers in their own words as far as practicable.

The figures in the margin indicate full marks.

Attempt all the questions.

1. a) Explain the Knowledge Discovery from Database (KDD) process with its major phases. 7

b) A numeric variable has the following values: 8

32.15, 87.49, 59.60, 41.78, 12.04, 18.67, 9.31, 190.54, 101.22, 24.88. Standardize the values.

2. a) Explain power law distribution by giving real life examples. Why are power laws called heavy-tailed distribution? 8

b) Define Pearson correlation. You are provided with the following dataset of 5 individuals, where x represents age and y represents annual salary: 7

Age	Salary in Rs
25	30000
32	45000
40	55000
48	65000
55	75000

Calculate the Pearson correlation coefficient between Age and Salary and interpret the result. Does the data suggest a strong, weak, or no linear relationship between age and salary?

3. a) What is Ordinary Least Squares (OLS) estimation, and how is the multivariate linear regression problem formulated using matrix algebra? 8

b) You are given the following data: 7

x	0.4	0.9	1.4	1.9
y	-1.8	1.2	8.5	19.1

Estimate the value of y at x = 8 using the Nadaraya-Watson regression estimator. Take bandwidth h = 0.2.

OR

Given a dataset with D observations on n numeric variables $x_1, \dots, x_n$, how do you compute the n principal components and determine the amount of variance explained by each of the components? Explain the process.

4. a) Consider a simple linear regression model: $Y=\beta 0+\beta 1X+\epsilon$, where Y is the dependent variable and X is the independent variable. Using the dataset: 8

Observation	X	Y
1	2	5
2	3	7
3	5	10
4	7	15
5	9	18

i. Formulate the matrix representation of the regression model.

ii. Using the estimated regression equation, predict the value of Y when X=6.

b) Explain R^2 and Adjusted R^2 in the context of model evaluation in detail. 7

5. a) You are building a decision tree classifier. At a certain node, you have the following data: 8

A	B	Class label
True	False	Cat
True	True	Cat
True	True	Cat
True	False	Dog
True	True	Cat
False	False	Dog

False	False	Dog
False	False	Dog
True	True	Dog
True	False	Dog

Which attribute should the decision tree use for splitting; A or B?

Justify your answer using a purity index such as Gini impurity or entropy.

b) Describe the CART algorithm for decision trees. How are splits determined in classification? 7

6. a) What is time series analysis? Explain its components in detail. 7

OR

How do you select the optimal lag length in ARIMA model? Explain in detail.

b) What conditions must be satisfied to conclude that event A causes event B? Describe. Why is Granger causality not considered true causality? 8

7. Write short notes on: **(Any two)** 2×5

a) Statistical significance

b) PCA

c) Causation vs. correlation