Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

School of Computer Science & Engineering

Mid - Term Examinations - November 2024

Semester: V | **Date**: 05-11-2024

Course Code: CSE2028 | **Time**: 02:00pm – 03:30pm

Course Name: STATISTICAL FOUNDATION FOR DATA SCIENCE | **Max Marks**: 50

Program: B.TECH(CSD) | **Weightage**: 25%

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

Part A

Answer ALL the Questions. Each question carries 2marks. 5Qx2M=10M

1	Define the term 'explanatory variable' and provide an example.	2 Marks	L1	CO1
2	xplain the difference between a sample and a population in statistics.	2 Marks	L2	CO1
3	What is the role of randomization in data collection?	2 Marks	L1	CO1
4	What is kernel ridge regression, and how does it extend ridge regression to handle non-linear data?	2 Marks	L1	CO2
5	Which regression method, L1 or L2, is better suited for feature selection, and why?	2 Marks	L1	CO2

Part B

Answer ALL Questions. Each question carries 10 marks. 4QX10M=40M

6	**6a.**	What is frequency distribution, and why is it important?	2 Marks	L1	CO1
	6b.	Explain how a histogram helps in visualizing frequency distribution.	3 Marks	L2	CO1
	6c.	Apply the concept of frequency distribution by constructing a histogram using a given dataset. Analyze the data trends observed in the histogram	5 Marks	L3	CO1

Or

7	**7a.**	Define quartiles and explain how they are calculated.	2 Marks	L1	CO1
	7b.	Explain the significance of percentiles and how they help in data interpretation	3 Marks	L2	CO1
	7c.	Apply the concept of box plots to visualize data distribution, and explain how the median, quartiles, and outliers are represented in a box plot	5Marks	L3	CO1
8	**8a.**	What is polynomial regression? How does it extend linear regression to capture non-linear relationships?	2 Marks	L1	CO2
	8b.	Explain when polynomial regression should be used instead of linear regression. What challenges does it introduce?	3 Marks	L2	CO2
	8c.	Apply polynomial regression to a dataset with a non-linear trend (e.g., fitting a quadratic model to a dataset). Evaluate its performance compared to linear regression.	5 Marks	L3	CO2

Or

9	**9a.**	Define multiple linear regression and provide an example where it can be applied.	2Marks	L1	CO2
	9b.	Describe the difference between simple linear regression and multiple linear regression. Why is multicollinearity a concern in multiple regression models?	3 Marks	L2	CO2
	9c.	Apply multiple linear regression to a dataset with more than two predictors. Analyze the importance of each predictor using p-values and the overall model performance using adjusted R-squared.	5 Marks	L3	CO2
10	**10a.**	What is inferential statistics, and how does it differ from descriptive statistics?	2 Marks	L1	CO1
	10b.	Explain the concept of a confidence interval and how it is used in statistical inference.	3 Marks	L2	CO1
	10c.	Apply the concept of confidence intervals to estimate a population parameter from a sample dataset. Discuss how confidence levels and sample size impact the width of the interval.	5 Marks	L3	CO2

Or

11	**11a.**	What is Data Preprocessing in Data Science?	2 Marks	L1	CO1
	11b.	Difference Between Data Cleaning and Data Transformation	3 Marks	L2	CO1
	11c.	Imagine you are analyzing a customer dataset where 20% of the data contains missing values for income. What steps would you take to preprocess this data, and what strategies would you use for handling missing values?	5Marks	L3	CO1
12	**12a.**	Define ridge regression. How does it differ from ordinary least squares	2 Marks	L1	CO2
	12b.	Explain the purpose of regularization in regression models. Why is ridge regression useful in preventing overfitting?	3 Marks	L2	CO2
	12c.	Apply ridge regression to a dataset with multicollinearity issues. Analyze the effect of the regularization parameter (λ) on model performance and compare it with OLS regression	5 Marks	L3	CO2
		Or			
13	**13a.**	Define Lasso (L1) and Ridge (L2) regularization.	2 Marks	L1	CO2
	13b.	Explain the key differences between L1 and L2 regularization in terms of model performance and feature selection.	3 Marks	L2	CO2
	13c.	Apply both Lasso and Ridge regression to a dataset and compare their effects on model performance and feature selection. Use cross-validation to determine the best regularization parameter.	5 Marks	L3	CO2