QP Code: RN23BTECH079 Reg. No □□□□□□□□□□ AY 23

GANDHI INSTITUTE OF ENGINEERING AND TECHNOLOGY, ODISHA, GUNUPUR (GIET UNIVERSITY)

B. Tech (Third Semester - Regular) Examinations, November – 2024

23BCSPC23001 - Introduction to Data Science

(CSE, CSE(DS))

Time: 3 hrs Maximum: 60 Marks

Answer ALL questions

(The figures in the right hand margin indicate marks)

PART – A **(2 x 5 = 10 Marks)**

Q.1. Answer ***ALL*** questions		CO #	Blooms Level
a.	Mention any four application fields in which data science can be applied.	CO1	K2
b.	What is data discretization and why is it important in data analysis?	CO2	K2
c.	Provide the general form of the equation for simple linear regression and multiple linear regression.	CO3	K2
d.	Define NULL hypothesis and Alternative hypothesis. Provide an example for each.	CO5	K2
e.	What is overfitting? How it can be avoided?	CO6	K2

PART – B **(10 x 5 = 50 Marks)**

Answer ***ALL*** the questions		Marks	CO #	Blooms Level
2. a.	Discuss the potential security risks associated with data breaches in data science.	5	CO1	K2
b.	Consider a logistics industry management system. Identify the need of data science in logistics industry management system to enhance business and management. Also describe in detail about uses of data science in logistics industry automation system.	5	CO1	K2
	(OR)			
c.	Consider a fraud detection system in banking sector that requires to implement a set of proactive measures to detect and avoid fraudulent activities and financial losses. Illustrate the different stages of data science project development with respect to the above scenario.	5	CO1	K2
d.	Describe any two roles involved in data science project development with their responsibilities.	5	CO1	K2
3.a.	Outline the steps involved in handling categorical data during the pre-processing phase.	5	CO2	K3
b.	Consider your own dataset and explain how to calculate the skewness and kurtosis values to assess the distribution of the data.	5	CO2	K3
	(OR)			
c.	Define the term simple linear regression. Evaluate the regression from the given data and evaluate the standard error	5	CO2	K3

X	1	3	10	16	26	36
Y	42	50	75	100	150	200

d.	Consider the daily temperatures (in °C) for a week are as follows: 22, 25, 23, 28, 30, 32, and 26. Find the mean, median and standard deviation.	5	CO2	K3
4.a.	Explain how does Box plot help to identify outliers. Mention the steps to handle outliers.	5	CO4	K3

b. Explain what a residual plot is, its purpose, and how it helps in diagnosing the performance of a regression model. 5 CO4 K3

(OR)

c. A researcher is studying the relationship between the number of hours spent studying and the test score of students. The following data is provided: 5 CO4 K3

Calculate the regression equation to predict the test score based on hours spent studying. What would you predict the test score to be if a student studied for 7 hours?

Hours Spent (X)	**Test Score (Y)**
2	55
3	60
5	70
6	75
8	85

d. A company is exploring the relationship between the hours of training (X) and the employee performance score (Y). After analyzing the data, the company finds that a third-degree polynomial regression fits the data better than a linear regression model. What is the general form of a third-degree polynomial regression model? Explain why a third-degree polynomial might provide a better fit than a linear regression model in this case. 5 CO4 K3

5.a. Explain the concepts of Type I and Type II errors in hypothesis testing. 5 CO5 K3

b. What do you mean by chi-squared test? The number of scooter accidents per month in a certain town was as follows: 5 CO5 K3

12, 8, 20, 2, 14, 10, 15, 6, 9, 4

Use the chi-squared test to determine if these frequencies are in agreement with the belief that accident conditions were the same during this period ($x^2_{0.06} = 16.92$)

(OR)

c. What is a heat map and explain how is it useful in correlation analysis. 5 CO5 K3

d. Consider the following exam scores of a group of students: 72,75,78,80,82,85,88,90,92,95. 5 CO5 K3

Compute the kurtosis and analyse whether the data shows a peaked or flat distribution relative to a normal distribution

6.a. What are the different classification evaluation metrics? Provide the formula to calculate it. 5 CO6 K2

b. What is cross-validation, and why is it important in model evaluation? 5 CO6 K2

(OR)

c. How does grid search help in finding the optimal hyper-parameters for a machine learning model? Describe the steps involved in performing a grid search. 5 CO6 K2

d. Describe how you would use Mean Squared Error (MSE) and Mean Absolute Error (MAE) to evaluate the performance of a regression model. 5 CO6 K2

--- End of Paper ---