.

Printed Page:-4 **Subject Code:- BBT0514**

Roll. No:

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA
(An Autonomous Institute Affiliated to AKTU, Lucknow)
B.Tech
SEM: V - THEORY EXAMINATION (2025 - 2026)
Subject: Data Science

Time: 3 Hours **Max. Marks: 100**

General Instructions:

IMP: *Verify that you have received the question paper with the correct course, code, branch etc.*
1. *This Question paper comprises of* ***three Sections -A, B, & C.*** *It consists of Multiple Choice Questions (MCQ's) & Subjective type questions.*
2. *Maximum marks for each question are indicated on right -hand side of each question.*
3. *Illustrate your answers with neat sketches wherever necessary.*
4. *Assume suitable data if necessary.*
5. *Preferably, write the answers in sequential order.*
6. *No sheet should be left blank. Any written material after a blank sheet will not be evaluated/checked.*

SECTION-A 20

1. Attempt all parts:-

1-a. State which programming language is widely used in Data Science? [CO1,K1] 1

(a) HTML
(b) C++
(c) Python
(d) PHP

1-b. Select the process of cleaning and transforming data is called? [CO1,K1] 1

(a) Data entry
(b) Data mining
(c) Data preprocessing
(d) Data feeding

1-c. State an outlier is a value that? [CO2,K1] 1

(a) Significantly deviates from others
(b) Equals the mean
(c) Is missing
(d) Is categorical

1-d. State, Z-score greater than ___ indicates an outlier? [CO2,K1] 1

(a) 1
(b) 2
(c) 3
(d) 5

1-e. Identify, If regression line is Y = 2 + 3X, predicted Y for X=4 is? [CO3,K1] 1

(a) 10
(b) 11
(c) 12
(d) 14

1-f. State, Skewness > 0 indicates? [CO3,K1] 1

(a) Positive skew
(b) Negative skew
(c) Symmetric
(d) Random

1-g. Identify, A positive correlation indicates? [CO4,K1] 1

(a) One variable increases while the other decreases
(b) Both variables increase together
(c) No relationship
(d) Random relationship

1-h. State, a normal distribution is? [CO4,K1] 1

(a) Skewed left
(b) Symmetrical around mean
(c) Uniform
(d) Bi-modal

1-i. Identify, Case studies show DS helps in? [CO5,K1] 1

(a) Personalized medicine
(b) Audio processing
(c) Painting
(d) Games

1-j. State, Plot for distribution shape? [CO5,K1] 1

(a) Pie
(b) Histogram
(c) Scatter
(d) Line

2. Attempt all parts:-

2.a. State what is IoT in Data Science applications? [CO1,K1] 2

2.b. Discuss what is Principal Component Analysis (PCA), with an example? [CO2,K2] 2

2.c. Explain what does R^2 (coefficient of determination) represent? [CO3,K2] 2

2.d. Solve, following are daily sales (₹): 100, 150, 200, 150, 250 Find the mean, mode and median? [CO4,K3] 2

2.e. State what does the coefficient in logistic regression represent, with an example? [CO5,K1] 2

SECTION-B 30

3. Attempt all parts:-

3.a. Answer any one of the following:-

3.a.(i)	Explain how tools like Python, R, and Tableau are used in Data Science projects. Mention any 2 packages of Python and its role in data analysis? [CO1,K2]	6
3.a.(ii)	Describe data cleaning and preprocessing steps in the Data Science workflow? [CO1,K2]	6
3.b. Answer any one of the following:-		
3.b.(i)	Compare filter methods, wrapper methods, and embedded methods of feature selection with examples? [CO2,K4]	6
3.b.(ii)	Explain any two statistical methods for outlier detection and removal? [CO2,K2]	6
3.c. Answer any one of the following:-		
3.c.(i)	Differentiate between correlation and regression in terms of purpose and interpretation? [CO3,K4]	6
3.c.(ii)	Explain regression and step-by-step procedure for calculating simple linear regression manually? [CO3,K2]	6
3.d. Answer any one of the following:-		
3.d.(i)	Describe what is null and alternative hypothesis? How do scientist come up with accept or reject the hypothesis? [CO4,K2]	6
3.d.(ii)	Describe the process of formulating a scientific hypothesis. Explain the steps scientists follow from observation to testing with an example? [CO4,K2]	6
3.e. Answer any one of the following:-		
3.e.(i)	Discuss and write a short note on applications of logistic regression in real-world scenarios? [CO5,K2]	6
3.e.(ii)	Explain what is the purpose of using heatmaps and scatter plots in data visualization? [CO5,K2]	6

SECTION-C 50

4. Answer any one of the following:-

4-a.	Differentiate between descriptive, predictive, and prescriptive analytics? [CO1,K4]	10
4-b.	Discuss real-world case studies where predictive analytics drastically improved operational efficiency (healthcare readmission, manufacturing maintenance, retail forecasting)? [CO1,K2]	10

5. Answer any one of the following:-

5-a.	Solve, The data below represents the amount of advertising expenditure (in ₹'000) and corresponding sales revenue (in ₹'000): Advertising: 10, 15, 20, 25, 30 Sales: 25, 30, 40, 45, 50 Compute the Pearson correlation coefficient and analyze the type of correlation between advertising and sales? [CO2,K3]	10
5-b.	Solve, The data below shows the number of hours spent on physical exercise (X) and the corresponding weight loss in kg (Y) for five individuals:? [CO2,K3] (2, 1), (3, 2), (4, 3.2), (5, 3.8), (6, 4.5). a) Calculate the mean of X and Y. b) Compute the Pearson's correlation coefficient (r) between X and Y. c) Interpret the result.	10

6. Answer any one of the following:-

.

6-a.	Implement, A teacher runs a regression model to predict students' exam scores using study hours (X_1) and attendance (X_2). The model is: $Y = -92.68 + 10.97X_1 + 17.52X_2$. Predict the exam score for a student who studies 4 hours and attends 6 lectures? [CO3,K3]	10
6-b.	Solve, Using the data: X = [2,4,6,8,10], Y = [3,7,9,12,15], find the regression line of Y on X? [CO3,K3]	10

7. Answer any <u>one</u> of the following:-

7-a.	Explain the concept of p-value. How is it used in hypothesis testing? Illustrate with an example? [CO4,K2]	10
7-b.	Solve, A sample of 20 items has mean 42 units and S.D. 5 units. Test the hypothesis that it is a random sample from a normal population with mean 45 units. (Test at 5% level of significance)? [CO4,K3] t 0.025,df, 19 =2.093.	10

8. Answer any <u>one</u> of the following:-

8-a.	Explain and write a detailed note on visualization tools used in data science? [CO5,K2]	10
8-b.	Describe, if a researcher wants to predict customer churn using logistic regression. Explain the steps involved from data preprocessing to model evaluation? [CO5,K2]	10