QP Code: S253A045 Reg. No ☐☐☐☐☐☐☐☐☐☐ AY 23

Gandhi Institute of Engineering and Technology University, Odisha, Gunupur (GIET University)

B. Tech (Third Semester – Regular & Supplementary) Examinations, November 2025

23BCSPE23011/23BCD23011 – 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.	Differentiate between labeled and unlabeled datasets in Machine Learning.	CO1	K1
b.	Define Supervised Learning and Unsupervised Learning in Machine Learning.	CO1	K2
c.	Calculate the variance for the following data:	CO3	K3
d.	Differentiate NULL hypothesis and Alternative hypothesis with a suitable example.	CO1	K2
e.	Differentiate under fitting and overfitting with an example.	CO2	K4

Data for c.:

X	10	20	30	40	50	60	70
Y	1	3	13	17	27	36	38

PART – B (10 x 5 = 50 Marks)

Answer ***ALL*** the questions — Marks / CO # / Blooms Level

2. a. Differentiate between primary data and secondary data. Give suitable examples for each. — 5 CO1 K2

b. Explain any five major applications of Data Science in real-world scenarios. — 5 CO2 K3

(OR)

c. Explain the life cycle of Data Science with neat diagram? — 5 CO4 K2

d. Find the outlier in the given grouped data using Z-Score formula. — 5 CO3 K3

Score Range	30-39	40-49	50-59	60-69	70-79
Frequency	3	9	2	1	20

3.a. What is Data Reduction? Explain its importance and describe any three common techniques used for reducing data size with suitable examples. — 5 CO3 K3

b. Define Skewness in statistics and explain its importance in data analysis. Using Karl Pearson’s Coefficient of Skewness, calculate the skewness for the following data and interpret the result. — 5 CO1 K1

Class Interval	0-5	5-10	10-15	15-20	20-25	25-30
Frequency (f)	6	9	16	17	7	5

(OR)

c. The marks of 50 students in a test are given in the grouped form below: — 5 CO2 K3

Score Range	0-10	10-20	20-30	30-40	40-50	50-60
Frequency	3	5	12	15	10	5

Calculate the Arithmetic Mean of the distribution by the Direct method**,** Assumed-mean method**,** and Step-deviation method, and show that all three give the same result.

d. Explain Boxplot. How are boxplots used for outlier detection? A dataset representing the test scores of a group of students: — 5 CO3 K4

Data (test scores): 30,50,80,85,99,111,121,121,132,136,190,220

Construct a boxplot for the above given test scores.

4.a. Recall WMA(Weighted Moving Average)? The following dataset shows the monthly sales (in units) of a company for the last 6 months: — 5 CO2 K3

Month	Jan	Feb	Mar	Apr	May	Jun
Sales	110	125	140	150	165	190

Using a 3-month weighted moving average to forecast sales for July, where the weights for the most recent three months are 0.5 (Jun), 0.3 (May), and 0.2 (Apr) respectively.

b. Define Kurtosis and explain its importance in statistics. Using the following class intervals and frequency data, calculate the Coefficient of Kurtosis and interpret the result. 5 CO2 K3

Class Interval	0-5	5-10	10-15	15-20	20-25	25-30
Frequency (f)	5	9	12	15	6	3

(OR)

c. Explain various data visualization graphs used in Data Science. Discuss the purpose of each with suitable examples. 5 CO3 K3

d. Define simple linear regression. From the data below, fit the least-squares regression line of Y on X, Compute the Standard Error of Estimate, and predict Y when $X = 15$. 5 CO3 K4

Experience (Years) (X)	2	4	6	8	10
Salary (₹ in thousands) (Y)	30	45	55	65	85

5.a. Define multiple linear regression and explain its purpose in data analysis. Given the following dataset showing hours of study (X_1), attendance percentage (X_2), and exam scores (Y) for six students, develop a multiple linear regression equation to predict exam scores. 5 CO3 K4

Student	**Study Hours (X_1)**	**Attendance (%) (X_2)**	**Exam Score (Y)**
1	2	60	50
2	3	65	55
3	5	70	65
4	6	80	75
5	8	85	85

Also, estimate the predicted exam score for a student who studies **9** hours and has 90% attendance.

b. Define ANOVA and explain its purpose. A company wants to compare the productivity of employees trained under three different programs — X, Y, and Z. The units produced by 4 employees under each program are as follows: 5 CO4 K4
X: [20,25,30,35], Y: [15,20,25,30], Z: [50,60,70,80]
At a 5% level of significance, test whether there is a significant difference in productivity among the three training programs.(Given: $F_{0.05},2,9 = 3.89$))

(OR)

c. Define polynomial regression and explain how it differs from simple linear regression.
Given the following dataset showing the relationship between years of experience (X) and annual salary (Y) of employees, develop a polynomial regression model to predict the salary for an employee with 12 years of experience. 5 CO3 K4

Employee	**Experience (Years) (X)**	**Salary (₹ in thousands) (Y)**
1	2	30
2	4	45
3	6	55
4	8	65
5	10	85
6	11	95

d. The following dataset shows the number of study hours**,** sleep hours, and the corresponding exam scores of 6 students:

Student	**Study Hours**	**Sleep Hours**	**Exam Score**
1	2	8	50
2	3	7	60
3	4	6	70
4	5	6	75
5	6	5	80

Find the correlation between the three subjects and identify which pair of subjects shows the strongest correlation and visualize through a heat map.

5 CO1 K2

6.a Explain T-Test and various types of T-Test? A nutritionist claims that the average daily protein intake of adults in a city is 60 grams. A researcher selects a sample of 25 adults and finds a sample mean of 57 grams with a sample standard deviation of 5 grams**.** Using a two-tailed t-test at the 5% significance level with T_tab = 2.064 (for 24 degrees of freedom), is there sufficient evidence to reject the nutritionist's claim? 5 CO4 K5

b. A survey is conducted to see whether gender affects preference for different chocolate flavors among college students. A sample of 100 students provides the following data:

Gender	**Dark**	**Milk**	**White**	**Caramel**	**Total**
Male	15	20	10	5	50
Female	10	25	15	0	50
Total	25	45	25	5	100

Using a chi-square test for independence at the 5% significance level, and χ^2_tab = 7.815 (for 3 degrees of freedom), test whether chocolate flavor preference is independent of gender**.**

5 CO4 K2

(OR)

c. The following table shows the actual and predicted values of house prices (in lakhs) obtained from a regression model.

Observation	**1**	**2**	**3**	**4**	**5**
Actual (Y)	50	60	70	80	90
Predicted ($\hat{Y}$)	48	65	68	75	88

Calculate the following regression performance metrics: 5 CO4 K3

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score

Interpret the results — which metric indicates the best model performance?

d. A binary classification model is tested on a dataset of 100 samples. The confusion matrix is given below:

	Predicted Positive	Predicted Negative
Actual Positive	40	10
Actual Negative	20	30

5 CO4 K3

Calculate the following metrics using the given data:

- Accuracy, Precision, Recall (Sensitivity), F1-Score

Also, interpret which metric gives a better understanding of model performance when false negatives are more critical.

--- End of Paper ---