Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – May / June 2026

Date: 29- 05-2026 **Time:**09:30am – 12:30pm

School: SOCSE	**Program:** B.tech	
Course Code : CSD2002	**Course Name:** Introduction to Data Science	
Semester: IV	**Max Marks**:100	**Weightage**: 50%

CO - Levels	CO1	CO2	CO3	CO4	CO5
Marks	12	12	26	24	26

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. **10Q x 2M=20M**

1.	**Define** Data Science.	2 Marks	L1	CO1
2.	**Why** is Google Colab preferred over a local Jupyter Notebook? List any two advantages of Google Colab.	2 Marks	L1	CO2
3.	**What** is a Series and what is a DataFrame in Pandas? Relate their basic differences.	2 Marks	L1	CO3
4.	**Show** how to display only petal_length and petal_width from the Iris dataset.	2 Marks	L1	CO3
5.	**Name** any two techniques used to handle outliers in a dataset.	2 Marks	L1	CO3
6.	**What** are the basic principles of effective data visualization?	2 Marks	L1	CO4
7.	**Define** a Histogram? What does it represent?	2 Marks	L1	CO4
8.	**Define** accuracy in model evaluation.	2 Marks	L1	CO5
9.	**List** the types of outcomes in a confusion matrix.	2 Marks	L1	CO5
10.	**Define** precision and recall.	2 Marks	L1	CO5

Part B

Answer the Questions. **Total Marks 80M**

11.	**a.**	**Explain** Pandas library and its importance in data analysis. (student data)	**20 Marks**	**L2**	**CO1**
	b.	**Interpret** a case scenario of Social Media and E-Commerce Analytics and explain how data science is used to analyze user behavior, personalize recommendations, and improve sales and engagement.		**L2**	**CO1**
		Or			
12.	**a.**	**Explain** the Data Science Lifecycle in detail with a neat diagram.	**20 Marks**	**L2**	**CO2**
	b.	**Compare** and contrast Jupyter Notebook, Anaconda, and Google Colab as environments for data analysis.		**L2**	**CO2**

13.	**a.**	**Apply** data cleaning techniques on the Titanic dataset to handle missing values and duplicate records.	**20 Marks**	**L3**	**CO3**
	b.	**Identify** what outliers are in a dataset and explain how they can be detected using the Interquartile Range (IQR) method.		**L3**	**CO3**
		Or			
14.	**a.**	**Develop** and utilize Feature Engineering techniques to improve model performance by creating new features from raw data.	**20 Marks**	**L3**	**CO3**
	b.	**Utilize** Pandas operations to analyze the Iris dataset and extract meaningful insights		**L3**	**CO3**

15.	**a.**	**Apply** a Line Chart to analyze trends in data. Identify suitable variables and explain how trends can be interpreted with an example.	**20 Marks**	**L3**	**CO4**
	b.	**Utilize** KDE plots to understand the distribution of data. Select an appropriate dataset and explain how the distribution can be analyzed.		**L3**	**CO4**
		Or			
16.	**a.**	**Construct** a Bar Chart to visualize student marks in 5 subjects. Organize the dataset and explain how comparison can be made effectively.	**20 Marks**	**L3**	**CO4**
	b.	**Build** a dashboard using Streamlit and Plotly for the Titanic dataset. Make use of Scatter Plot and Pie Chart to visualize the data and explain their purpose.		**L3**	**CO4**

17.	**a.**	**Compare and contrast** t-Test and p-value in hypothesis testing. Inspect their relationships and draw a conclusion about their role in decision-making.	**20 Marks**	**L4**	**CO5**
	b.	**Examine** logistic regression and classify its role in classification problems. Discover patterns and infer how it differs from linear regression.		**L4**	**CO5**
Or					
18.	**a.**	**Analyze** model evaluation metrics. Categorize accuracy, precision, recall, and confusion matrix, and examine their relationships in evaluating model performance.	**20 Marks**	**L4**	**CO5**
	b.	**Inspect** a confusion matrix and distinguish between true positives, false positives, true negatives, and false negatives. Simplify the interpretation of results.		**L4**	**CO5**