Sub Code: ECS-103

Roll No.

B. Tech. (4 Credits)
Year: II Semester: IV
Minor Test Examination Even Semester 2025-2026
Subject Name: Introduction to Data Science

Time: 2 Hr. **Max Marks: 20**

Note: Attempt all questions. Answer to the point. Irrelevant answers will attract negative marks.

Q1	**Attempt any two parts of the following. (From Units I and II)**	**Marks**	**CO**	**BL**	**PO**
a)	What do we understand by the term data? Explain the end-to-end process involved in extracting useful insights from data. Illustrate the process with a real-world application.	4	CO1	L4	PO2
b)	List any three major sectors where Data Science is widely used, with a suitable example. Differentiate between structured, semi-structured, and unstructured data in brief.	2+2	CO2	L2	PO1
c)	6, 12, 18, 25, 31, 44, 52, 60, 72, 85, 240 (i) Find the first quartile (Q_1), median (Q_2), and third quartile (Q_3). (ii) Compute the Interquartile Range (IQR). (iii) Determine whether the dataset contains any outliers using the 1.5 × IQR rule	2+1+1	CO2	L3	PO1
Q2.	**Attempt any two parts of the following. (Unit I only)**				
a)	Outline the societal impact of Data Science by highlighting two benefits and one ethical issue.	3	CO1	L3	PO6
b)	Examine how Data Science affects society at large. Briefly distinguish among different categories of data.	3	CO1	L4	PO8
(c)	The data collection contains 275 observations, each with 12 measured attributes and 3 derived attributes obtained through feature engineering. The dataset is stored in a tabular format where each row corresponds to an observation, and each column corresponds to an attribute. (i) Determine the effective dimensionality of the dataset. (ii) Specify the order of the data matrix used to represent the dataset. (iii) Classify the dataset as structured, semi-structured, or unstructured, and justify your classification	3	CO1	L4	PO1
Q3.	**Attempt any two parts of the following. (Unit II only)**				
a)	For the given Distribution 3, 7, 14, 18, 22, 27, 29, 33, 38, 46, 52 (a) Apply equal-width binning by dividing the data into 4 bins. (b) Clearly specify the bin range (width) and assign each data point to its appropriate bin. (c) Perform bin smoothing using bin means and write the final smoothed dataset after binning.	3	CO2	L4	PO2
b)	Discuss any two quantitative techniques used in EDA in detail.	3	CO2	L3	PO4
c)	Explain three data cleaning techniques for missing values with suitable examples	3	CO2	L6	PO3

****

ECS-103

ROLL NO. | | | | | | | | | | |

B. TECH.

YEAR: 2nd SEMESTER: IVth

MAJOR EXAMINATION: 2025-26

Introduction to Data Science

MAX MARKS: 50

TIME: 3 Hrs.

Note: Attempt ALL questions.

			Marks	COs	BL	PO	PI Code
Q1.		Attempt any ***Five*** parts of the following. (Unit I and Unit II)					
	(a).	Define Data Science and state its primary objective in modern data-driven environments.	2	1	1	1	1.1.1
	(b).	Differentiate between structured, semi-structured, and unstructured data with suitable examples.	2	1	2	1	1.1.1
	(c).	What is Exploratory Data Analysis (EDA)?	2	2	1	1	1.2.1
	(d).	A dataset contains 150 observations with 8 original features and 4 newly generated features. (i) Determine the dimensionality of the dataset. (ii) Represent the order of the data matrix.	2	2	3	1	1.2.1
	(e).	State two principles of effective data visualization.	2	1	1	1	1.3.1
	(f).	Given the dataset: 10, 15, 20, 25, 30, 35, 40, 45 Calculate Q_1, Q_2, Q_3 and compute the Interquartile Range (IQR).	2	2	3	6	6.1.1
	(g).	Examine the impact of data science on society by discussing two advantages and one ethical concern.	2	1	4	1	1.3.1
Q2.		Attempt any ***Two*** parts of the following. (Unit III)					
	(a).	In what ways can poor feature generation or selection negatively impact model performance? Explain with examples.	5	3	2	1	1.2.1
	(b).	Differentiate between feature generation and feature selection with suitable examples.	5	3	4	2	2.1.1
	(c).	I. A dataset contains **8 samples** with **2 features (X and Y)** as given below. Standardize the dataset. II. Compute the **covariance matrix** of the features. III. Find the **eigenvalues and eigenvectors** of the covariance matrix. IV. Determine the **principal components**.	5	3	3	3	3.1.2

Sample	X	Y
1	2	3
2	3	5
3	4	7
4	5	9
5	6	11
6	7	13

		7	8	15					
		8	9	17					
Q3.		Attempt any ***Two*** parts of the following. (Unit III)							
	(a).	Explain 3 major feature selection algorithms in detail.			5	3	2	1	1.2.1
	(b).	What are overfitting, underfitting, and an optimal model in data science? Explain each concept clearly with suitable details.			5	3	4	2	2.2.1
	(c).	How does creating new features differ from reducing existing features? Discuss their effects on model complexity and performance			5	3	3	3	3.2.1
Q4.		Attempt any ***Two*** parts of the following. (Unit IV)							
	(a).	How can data visualization transform raw data into meaningful information? Explain with suitable examples.			5	4	2	1	1.3.1
	(b).	What factors should be considered while designing a clear and effective visualization? Discuss their significance.			5	4	2	2	2.1.2
	(c).	Construct a Q–Q (quantile–quantile) plot to visualize the distribution of the following dataset and assess its conformity with a normal distribution: Data: 5, 7, 9, 10, 12, 15, 18, 20, 22, 25 I. Generate a Q–Q plot as a data visualization technique to compare the sample data with a normal distribution. II. Interpret the visualization to evaluate how closely the data follows a normal distribution. III. Explain in brief how the Q–Q plot helps in understanding the distribution pattern compared to numerical measures alone.			3+1+1	4	4	2	2.2.2
Q5.		Attempt any ***Two*** parts of the following. (Unit IV)							
	(a).	Evaluate how storytelling techniques influence the interpretation of data and improve decision-making in data visualization.			5	4	4	10	10.1.1
	(b).	A dataset showing the distribution of marks obtained by students in an examination is to be analyzed using a histogram. The grouped data is provided below: I. Construct a **histogram** to visually represent the data distribution. II. Analyze the histogram to identify **patterns, trends, and concentration of values** in the dataset. III. Discuss how the visualization highlights any **skewness or spread** in the data			5 (2+1.5+1.5)	4	3	3	3.1.2

Marks Range	**Frequency**
0 – 10	3
10 – 20	5
20 – 30	9
30 – 40	12
40 – 50	8
50 – 60	3

	(c).	Compare different data visualization tools based on their features, usability, and typical applications, with suitable examples.	5	4	2	10	1.3.1