Reg.No.:												

VIVEKANANDHA COLLEGE OF ENGINEERING FOR WOMEN
[AUTONOMOUS INSTITUTION AFFILIATED TO ANNA UNIVERSITY, CHENNAI]
Elayampalayam – 637 205, Tiruchengode, Namakkal Dt., Tamil Nadu.

Question Paper Code: 130005

B.E. / B.Tech. DEGREE END-SEMESTER EXAMINATIONS – NOV. / DEC. 2025
Fifth Semester
Computer Science and Technology
U23CT511 – FOUNDATIONS OF DATA SCIENCE
(Regulation 2019)

Time: Three Hours | Maximum: 100 Marks

Answer ALL the questions

Knowledge Levels (KL)	K1 – Remembering	K3 – Applying	K5 - Evaluating
	K2 – Understanding	K4 – Analyzing	K6 - Creating

PART – A

(10 x 2 = 20 Marks)

Q.No.	Questions	Marks	KL	CO
1.	Mention two applications of data science in various fields.	2	K1	CO1
2.	List the stages in data science project.	2	K1	CO1
3.	Mention the importance of data reduction in data science projects.	2	K2	CO2
4.	State the relationship between data transformation techniques and data integration challenges.	2	K2	CO2
5.	Describe the purpose of a pivot table.	2	K1	CO3
6.	Write a short note on binomial distribution.	2	K1	CO3
7.	State the purpose of sampling in model development and validation.	2	K2	CO4
8.	Outline the measures to handle underfitting and overfitting in machine learning.	2	K1	CO4
9.	State two common graphics parameters that can be customized in plot functions and their effects.	2	K2	CO5
10.	Identify one common technique used to visualize relationships among multiple variables in a dataset and briefly describe how it helps in understanding multivariate data.	2	K2	CO5

PART – B

(5 x 13 = 65 Marks)

Q.No.		Questions	Marks	KL	CO
11.	a)	A transportation company plans to introduce a Smart Fleet Management System to optimize route planning and maintenance schedules.		K3	CO1
		i. Design the Data Science Project Pipeline covering the main lifecycle stages relevant to this system.	5		
		ii. Discuss the key data specialist roles required (e.g., Data Engineer, Data Scientist, ML Engineer) and explain their responsibilities.	5		
		iii. Identify potential data security challenges and suggest mitigation strategies to protect sensitive vehicle and driver data.	3		
		(OR)			
	b)	A chain of fitness centers is transitioning to a data-driven approach to personalize member workout plans and improve engagement.		K3	CO1
		i. Propose a Datafication Strategy that categorizes new and existing data sources, such as membership data, wearable devices, and social media feedback.	5		
		ii. Develop a Data Science Framework addressing key domains including user profiling, activity prediction, and personalized recommendations.	5		
		iii. Explain how Big Data principles (e.g., Volume, Variety, Velocity, Veracity, Value) support these applications and enable scalable insights.	3		
12.	a)	A telecommunications company wants to predict customer churn using data from call records, customer service interactions, and billing information.		K3	CO2
		i. Propose a comprehensive data collection plan that includes primary and secondary data sources.	4		
		ii. Analyze the challenges of integrating these heterogeneous sources and how you would address them.	4		
		iii. Design a preprocessing pipeline to handle missing values, duplicates, and data transformation while maintaining data quality. Outline key steps and justify choices.	5		
		(OR)			

b) A real estate company has a dataset containing property listings with missing values, outliers, and inconsistent formats for price and area. The dataset includes: 13 K3 CO2

Property_ID	Area (sq.ft)	Price (in Lakhs)	Location	Year_Built	Rooms	Age (Years)
P001	1200	50	Chennai	2005	3	15
P002	950	NULL	Coimbatore	2010	2	10
P003	1500	70	NULL	2000	4	20
P004	800	40	Chennai	2015	2	-5
P005	NULL	55	Bangalore	2012	3	12

Apply appropriate data preprocessing techniques to clean and prepare this data for analysis. Identify and handle missing and invalid values, detect outliers, and perform data normalization.

13. a) A food company tests three different packaging materials (A, B, and C) to see which preserves freshness best. Five samples for each material are tested for days until spoilage: 13 K2 CO3

Material A: 12, 15, 14, 13, 16
Material B: 10, 9, 8, 7, 12
Material C: 14, 15, 16, 14, 17

Use one-way ANOVA at 5% significance level to determine if the mean shelf life differs by packaging material. Show all calculations and interpret the results.

(OR)

b) A class of 25 students took an aptitude test with scores recorded as follows: 13 K2 CO3

58, 62, 70, 45, 55, 66, 72, 61, 49, 68, 73, 64, 60, 59, 71, 55, 58, 65, 67, 56, 69, 74

Calculate the mean, standard deviation, skewness, and kurtosis of the scores. Construct a box plot to identify any potential outliers. Explain how these descriptive statistics help in understanding the students' performance distribution.

14. a) A logistics company wants to group 8 delivery locations based on their coordinates for optimized route planning. The coordinates (x, y) of the locations are: 13 K3 CO4

L1 = (3, 5), L2 = (4, 7), L3 = (3, 8), L4 = (8, 2),
L5 = (7, 3), L6 = (8, 4), L7 = (1, 1), L8 = (2, 0).

Starting with initial cluster centroids at L1, L4, and L7, perform one complete iteration of the K-Means clustering algorithm using Euclidean distance. Assignment of data points to clusters, calculate new centroids after the iteration. Explain whether further iterations are necessary for convergence.

(OR)

	b)	Explain the logistic regression model and its applications in binary classification problems. How the logistic (sigmoid) function transforms the linear combination of input features to a probability. Describe the assumptions underlying logistic regression and the methods used for estimating model parameters. Finally, outline how model performance is evaluated in logistic regression using metrics such as accuracy, precision, recall, and the F1 score.	13	K3	CO4
15.	a)	Given a dataset and results from a data science model, demonstrate how to create multiple plots in one window using matrix plots and plot functions. Explain how to customize graphics parameters to improve visualization aesthetics, export graphs in different formats, and ensure reproducibility of graphical results.	13	K3	CO5
		(OR)			
	b)	Describe the key steps and best practices for documenting a data science project effectively. Explain how to organize and present graphical outputs clearly, including annotating visualizations for better understanding. Discuss how you would tailor the final project deliverable to suit a non-technical audience, ensuring clarity and engagement.	13	K3	CO5

PART – C

(1 x 15 = 15 Marks)

Q.No.		Questions	Marks	KL	CO
16.	a)	A retail company has gathered monthly sales data and corresponding advertising expenses for the past 8 months as shown:	15	K4	CO4

Month	Advertising Expense (in $1000)	Sales (in $10000)
1	10	50
2	15	60
3	20	75
4	25	80
5	30	95
6	35	100
7	40	110
8	45	120

Compute the slope b_1 and intercept b_0. Predict sales when advertising expense is $28,000. Show all calculations step-by-step and interpret the results.

(OR)

	b)	A retail company wants to use data science to improve customer retention. Propose a data science project plan highlighting key lifecycle stages essential for this goal. Emphasize the important data security issues that must be addressed to protect customer data during the project. Explain how these stages and security measures contribute to the project's success.	15	K3	CO1