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: 130006

B.E. / B.Tech. DEGREE END-SEMESTER EXAMINATIONS – NOV. / DEC. 2025

Seventh Semester

Biomedical Engineering

U19CTOE3 – FUNDAMENTALS OF DATA SCIENCE

(Common to IT and BT)

(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.	State the key characteristics of big data and outline the major factors that contributed to the evolution of data science from traditional statistical analysis methods.	2	K1	CO1
2.	Distinguish between the primary responsibilities of a data engineer and a data analyst in a typical data science team.	2	K2	CO1
3.	Compare the effectiveness of histogram-based and clustering-based approaches for data discretization. Illustrate the comparison with suitable examples.	2	K2	CO2
4.	Examine the relationship between data transformation techniques and data integration challenges. Show how attribute construction differs from attribute subset selection in the context of dimensionality reduction.	2	K2	CO2
5.	Mention the purpose of a heat map and how can it be used to visualize correlation in a dataset.	2	K2	CO3
6.	Determine the median of the following data set: 85, 90, 75, 88, 92, 80, 78.	2	K2	CO3
7.	State the purpose of sampling in a model development and validation.	2	K2	CO4
8.	Recall the definition of model validation in the context of 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.	Define the purpose of matrix plots in data visualization.	2	K1	CO5

PART – B

(5 x 13 = 65 Marks)

Q.No.		Questions	Marks	KL	CO
11.	a)	Consider an e-commerce company that wants to implement a customer behaviour analysis system to improve sales and customer retention. Design the complete data science project pipeline by identifying and explaining each stage of the data science project lifecycle for this scenario. Also, analyse the various data science roles required for this project and justify the specific responsibilities each role would have in implementing the customer behaviour analysis system. Analyse the potential data security issues that could arise during this project development and implementation and propose appropriate security measures to address each identified risk.	13	K3	CO1
		(OR)			
	b)	A traditional retail chain with 500 physical stores wants to transform into a data-driven organization to compete with online retailers. Implement a datafication strategy by identifying and categorizing different types of data sources available to this retail chain. Justify how each data source can be transformed into valuable digital information for a business intelligence. Also, design a comprehensive data science application framework that addresses at least three different business domains such as customer analytics, inventory management, sales optimization, etc., Describe the application of big data concepts within each selected domain.	13	K3	CO1
12.	a)	Design a comprehensive data collection strategy for a healthcare organization that wants to implement predictive analytics for patient care management. Demonstrate the application of different data collection methods including primary and secondary sources. Analyze the challenges in integrating heterogeneous healthcare data from electronic health records, medical devices, patient portals, and external databases. Construct a data-preprocessing pipeline that addresses data quality issues, standardization requirements, and privacy concerns. Evaluate how data transformation and reduction techniques can be applied to optimize datasets for machine learning applications while maintaining data integrity.	13	K3	CO2
		(OR)			

b) A retail company has collected customer purchase data from their online platform. The raw dataset is shown below: K3 CO2

Customer_ ID	Age	Income	Purchase_ Amount	City	Rating
C001	25	45000	1200	Mumbai	4.2
C002	-2	52000	850	Delhi	3.8
C003	35	NULL	2100	Mumbai	4.5
C004	28	38000	1650	Chennai	NULL
C005	42	67000	950	Mumbai	4.1
C006	28	38000	1650	Chennai	3.9
C007	150	55000	1400	Delhi	3.7
C008	31	41000	1100	Kolkata	4.0
C009	29	48000	NULL	Mumbai	4.3
C010	33	59000	1800	Delhi	4.6

Apply comprehensive data pre-processing techniques to clean and transform this dataset.

i. Apply appropriate techniques to handle the missing values in Income, Purchase_Amount, and Rating columns with proper justification. Remove duplicate records and calculate the final dataset size. 6

ii. Implement equal-width binning for the Income column with 3 bins and show the bin ranges and assignments. Finally, apply min-max normalization to the Purchase_Amount column and show the normalized values for all records. 7

13. a) A researcher is studying the effect of three different fertilizers on plant growth. The growth measurements (in cm) for three groups of plants treated with Fertilizer A, Fertilizer B, and Fertilizer C are: 13 K2 CO3

Fertilizer A: 6, 8, 4, 5, 3
Fertilizer B: 8, 12, 9, 11, 6
Fertilizer C: 13, 9, 11, 8, 7

Using one-way ANOVA, test if there is a significant difference in the mean growth among the three fertilizer groups at a 5% significance level. Show all calculations and state your conclusion.

(OR)

b) A class of 20 students took a statistics exam and obtained the following scores: 13 K2 CO3

85, 90, 75, 92, 88, 79, 83, 95, 87, 91, 78, 86, 89, 94, 82, 80, 84, 93, 88, 81

Calculate the mean, standard deviation, skewness, and kurtosis for this dataset. Using the scores, construct a box plot and identify any outliers. Finally, explain how each of these descriptive statistics helps in understanding the distribution and variability of the exam scores.

14.	a)	Given the following eight data points: $A1 = (2,10)$, $A2 = (2,5)$, $A3 = (8,4), A4 = (5,8)$, $A5 = (7,5)$, $A6 = (6,4)$, $A7 = (1,2)$, $A8 = (4,9)$, and starting with initial centroids at $A1, A4, A7$, execute one full 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.	13	K3	CO4
		(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)	Design a detailed process for documenting a data science project, including best practices for producing effective presentations of results. Illustrate how you would organize graphical outputs, annotate visualization for clarity, and prepare the final deliverable for a non-technical audience.	13	K3	CO5

PART – C

(1 x 15 = 15 Marks)

Q.No.		Questions	Marks	KL	CO
16.	a)	A small company's sales y (in thousands) for 6 months is recorded along with its advertising expense x (in thousands). The data is:	15	K4	CO4

Month	Advertising Expense x	Sales y
1	2	5
2	3	7
3	5	11
4	7	14
5	9	15
6	10	17

Compute the slope b_1 and intercept b_0. Predict sales when advertising expense is 8. Show all calculations step-by-step and interpret the results.

(OR)

b)	A retail company wants to use data science to keep its customers engaged and loyal. Design a detailed data science project plan that covers all the key stages from identifying the problem to developing, testing, and deploying the solution. Discuss the major data security and privacy concerns that could arise at each stage, and suggest practical ways to protect customer information. Explain how adhering these steps and security practices can help make the project both successful and trustworthy.	15	K3	CO1