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. 2024
Seventh Semester
Computer Science and Technology
U19CTV38 – DATA SCIENCE TECHNIQUES
(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.	What is the big data ecosystem, and how does it relate to data science?	2	K1	CO1
2.	Define normal distribution and z-scores.	2	K1	CO1
3.	List some applications of machine learning in data science.	2	K2	CO2
4.	What are the common problems in handling large data?	2	K1	CO1
5.	How is regression useful in data analysis?	2	K2	CO2
6.	Define the standard error of estimate.	2	K1	CO1
7.	What does a p-value signify in statistical testing?	2	K1	CO1
8.	Describe the sampling distribution of the t-statistic.	2	K2	CO2
9.	What are the main features of Tableau?	2	K1	CO1
10.	Name some popular data science tools.	2	K1	CO1

PART – B

(5 x 13 = 65 Marks)

Q.No.		Questions	Marks	KL	CO
11.	a)	Describe the intricate steps involved in the data science process, emphasizing the significance of each stage. How do these steps interconnect to ensure the accuracy and reliability of data-driven insights?	13	K3	CO3
		(OR)			
	b)	Given an extensive dataset of millions of e-commerce transactions, discuss the challenges and best practices for summarizing and visualizing this data. How can the company effectively convey trends and insights while ensuring clarity and avoiding information overload for stakeholders?	13	K4	CO4
12.	a)	Discuss the role and importance of data preprocessing in machine learning, particularly when handling large datasets. How does effective data preprocessing contribute to the quality and performance of machine learning models, and why is it considered a crucial step in the data science workflow?	13	K3	CO3
		(OR)			
	b)	Elucidate the advanced programming techniques and strategies employed in the efficient management of large-scale datasets. Analyze how these methodologies contribute to optimizing computational performance and resource utilization, ensuring robust handling and processing of big data in a high-performance environment.	13	K4	CO4
13.	a)	An online retail company wants to analyze the relationship between the amount spent by customers and their frequency of purchases. The dataset includes customer purchase amounts and the number of transactions made by each customer over the past year. To identify trends and understand the correlation between spending and purchasing frequency, the company decides to use scatter plots. Discuss the application of scatter plots in visualizing the relationship between customer spending and purchase frequency in this case study. How can scatter plots be interpreted to uncover trends and correlations in the data, and what insights can they provide to guide strategic business decisions?	13	K5	CO4
		(OR)			
	b)	Analyze the concept of the standard error of estimate. How does it reflect the accuracy of predictions in regression analysis?	13	K3	CO3
14.	a)	Compare and contrast the t-test for two independent samples with the t-test for two related samples. In what situations would each be used?	13	K3	CO3

(OR)

	b)	A company has launched three different marketing campaigns across three distinct regions to evaluate which campaign yields the highest increase in sales. The sales data from each region for the past six months are collected. To determine if there are significant differences in the average sales increases among the three campaigns, the company decides to use Analysis of Variance (ANOVA). Explain the steps involved in conducting an Analysis of Variance (ANOVA) to evaluate the effectiveness of the three marketing campaigns in this case study. How does ANOVA help in comparing the means of sales increases across the different campaigns, and what insights can be derived from the analysis?	13	K3	CO4
15.	a)	Analyze the capabilities of Apache Flink in handling real-time data processing. How does it compare to other big data tools?	13	K3	CO3

(OR)

	b)	Discuss how TensorFlow is used in machine learning and deep learning applications. What are its advantages?	13	K3	CO3

PART – C

(1 x 15 = 15 Marks)

Q.No.		Questions	Marks	KL	CO
16.	a)	A manufacturing company wants to implement a predictive maintenance system to reduce downtime and improve efficiency. The company has collected large amounts of sensor data from its machines and needs to build a machine learning model to predict equipment failures before they occur. • Outline the data science process you would follow to develop the predictive maintenance model, including data collection, preprocessing, model training, and evaluation. • Discuss the challenges associated with handling large volumes of sensor data and the general techniques you would employ to manage these challenges. • Propose a machine learning approach using a suitable algorithm for this problem. Justify your choice and explain how you would validate the model's performance.	15	K5	CO4

(OR)

b)	A market research firm is analyzing the relationship between advertising expenditure and product sales for a client. The firm has collected data on various factors, including the amount spent on different types of advertising (e.g., online, TV, print) and the corresponding sales figures. • Using correlation and regression analysis, describe how you would assess the relationship between advertising expenditure and sales. Explain the significance of the correlation coefficient and the regression line in this context. • Develop a multiple regression model to predict sales based on different types of advertising expenditures. Interpret the coefficients and the R-squared value of your model. • Discuss the potential pitfalls of using regression analysis in this scenario, including issues of multicollinearity and regression towards the mean. How would you address these issues?	15	K5	CO4