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: 130013

B.E. / B.Tech. DEGREE END-SEMESTER EXAMINATIONS – NOV. / DEC. 2024
Fifth Semester
Computer Science and Engineering
U19CTOE3 – FUNDAMENTALS OF DATA SCIENCE
(Common to IT & 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.	Which two crucial steps constitute the data science process, and what makes them significant?	2	K1	CO1
2.	In the context of datafication, how are the responsibilities of a data scientist and data engineer different?	2	K2	CO2
3.	Describe a technique for handling missing data during the data cleaning process with an example.	2	K3	CO2
4.	Examine how the performance of machine learning models can be affected by data discretization. Provide two potential effects it may have on model accuracy.	2	K4	CO3
5.	Given a dataset with the following values: 12, 15, 20, 25, and 28, calculate the mean and standard deviation. Use these values to evaluate the spread of the data around the mean.	2	K3	CO5
6.	In order to comprehend the correlations between variables in a dataset, what is the aim of a heat map and how is it used?	2	K1	CO1
7.	Examine two important aspects that need to be taken into account when selecting and assessing models for a particular dataset. How do these variables affect the performance of the model?	2	K3	CO3
8.	What is the main reason for data science model validation?	2	K1	CO1
9.	Examine the differences between a box plot and a histogram in terms of how well they visualize the distribution of a single numerical variable. When would each kind of plot be more advantageous?	2	K2	CO2

10.	What are the essential components of an effective documentation that outline a machine learning project's data pretreatment pipeline?	2	K3	CO3

PART – B

(5 x 13 = 65 Marks)

Q.No.		Questions	Marks	KL	CO
11.	a)	You are leading a data science project for a retail company aiming to improve customer segmentation. The project team includes the following roles: Data Scientist, Data Engineer, Data Analyst, and Business Analyst. Describe the specific responsibilities of each role in this project. Explain how these roles will collaborate to achieve the project goals, and identify any potential challenges in coordination between them.	13	K3	CO2
		(OR)			
	b)	Discuss various data security issues in a data science project for a healthcare organization that deals with sensitive patient information. Your plan should include details about Data Access Controls and Compliance and Legal Considerations.	13	K3	CO2
12.	a)	Analyze the following aspects of data collection strategies for a project aimed at improving the recommendation system for an e-commerce website: • Sampling Methods: Compare the effectiveness of random sampling versus stratified sampling in collecting user data for training the recommendation system. Discuss the impact of each method on the representativeness and accuracy of the recommendations. • Data Integration: Assess the challenges and benefits of integrating data from multiple sources, such as user interactions on the website, purchase history, and external social media data. How does data integration influence the performance of the recommendation system?	13	K4	CO4
		(OR)			
	b)	Conduct a comparative analysis of SVD and PCA based on the following aspects: • Applications and Use Cases: Analyze the primary applications of SVD and PCA. Explain how each technique is used in practical scenarios, such as in recommendation systems or data visualization. Provide examples to illustrate their use. • Advantages and Limitations: Evaluate the advantages and limitations of using SVD versus PCA for dimensionality reduction. Discuss the computational efficiency, interpretability of results, and any constraints associated with each method.	13	K4	CO4

13. a) Evaluate the effectiveness of different EDA techniques in the context of analyzing a dataset from an e-commerce platform, focusing on the following aspects: 13 K3 CO3

- Descriptive Statistics vs. Data Visualization:
- Handling Missing Data and Outliers

(OR)

b) Evaluate the effectiveness of Correlation Statistics versus ANOVA in the context of analyzing marketing data from an e-commerce platform, focusing on the following aspects: 13 K3 CO3

- Purpose and Application
- Interpretation of Results

14. a) Consider a dataset with the following five 2-dimensional data points representing customer purchase behaviours: 13 K5 CO4

- Data Points:
  1. (2, 3)
  2. (3, 3)
  3. (6, 5)
  4. (8, 8)
  5. (7, 7)

You are to perform k-means clustering with k=2.

(OR)

b) Consider the following dataset, which contains performance ratings for two different marketing strategies evaluated by several analysts. The dataset provides the ratings given by each analyst for Strategy A and Strategy B. 13 K5 CO4

Analyst	Rating for Strategy A (Xi)	Rating for Strategy B (Yi)
1	4	3
2	2	4
3	3	2
4	5	5
5	1	3
6	3	1

Answer the following questions with respect to above data.

- Find the values of B0 and B1 with respect to Linear Regression Model which best its the given data.
- Compute the slope and intercept for the linear regression model that predicts the rating for Strategy B based on the rating for Strategy A.
- Write the equation of the regression line that best fits the data provided.
- If a new analyst rates Strategy A with a score of 3, predict the rating for Strategy B given by this analyst. Show the calculation.

15.	a)	Evaluate how you would tailor a presentation on employee performance metrics for HR managers, team leaders, and senior executives. Discuss the data visualizations you would use to communicate key performance indicators and how you would structure the presentation to highlight the impact on employee performance. Additionally, outline strategies for handling audience questions and feedback effectively.	13	K3	CO5
		(OR)			
	b)	Evaluate how you would use graphical analysis to explore relationships between purchase amount, purchase frequency, and demographic variables (age and income) in a customer dataset. Discuss the specific graphical techniques you would employ and how you would interpret and present these insights to stakeholders, including marketing teams and executives.	13	K3	CO5

PART – C

(1 x 15 = 15 Marks)

Q.No.		Questions	Marks	KL	CO
16.	a)	You have been assigned a complex data science project that will involve an e-commerce platform dataset analysis. Customer demographics, past purchases, and product ratings are all included in the dataset. Gaining practical insights to enhance customer engagement and sales tactics is your aim. Give an outline of the procedures you would use to gather and prepare the data. Add in the handling of missing values, feature selection, and data normalization. Describe how you would guarantee the accuracy and consistency of the data used for analysis. In order to comprehend the dataset, talk about the methods you might employ for exploratory data analysis.	15	K6	CO4
		(OR)			
	b)	You are tasked with conducting a comprehensive project involving the analysis of a city's traffic management dataset. The dataset includes information on vehicle types, traffic density, road conditions, and accident reports. Your goal is to derive actionable insights to improve traffic flow and reduce accidents in the city. Describe the steps you would take to collect and preprocess the data. Include considerations for handling missing values, data normalization, and feature selection. Explain how you would ensure the quality and reliability of the data for analysis. Also discuss the techniques you would use for exploratory data analysis to understand the dataset. What types of visualizations and statistical methods would you employ to uncover patterns, trends, and relationships in the data? Provide examples of how these methods can help in forming initial hypotheses.	15	K6	CO4