CSE

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: 50028

B.E. / B.Tech. DEGREE END-SEMESTER EXAMINATIONS – NOV. / DEC. 2024
Fifth Semester
Computer Science and Engineering
U19CSV32 – DATA SCIENCE AND ANALYTICS
(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 the significance of data science and write two benefits of using it.	2	K1	CO1
2.	List the steps for retrieving and cleansing data in a data science project.	2	K2	CO1
3.	Calculate the mean and variance for the following dataset: {12, 15, 14, 10, 13}.	2	K3	CO2
4.	Consider an experiment to analyse the average time spent on a website. A sample of 50 users are taken, and the mean time spent is observed to be 20 minutes with a standard deviation of 4 minutes. Calculate the 95% confidence interval for the true mean time spent on the website.	2	K3	CO2
5.	What is linear regression model?	2	K3	CO3
6.	Summarize the benefits of fuzzy decision tree.	2	K2	CO3
7.	Differentiate stream data and batch data.	2	K2	CO4
8.	How you would count distinct elements in a stream?	2	K1	CO4
9.	What are the different types of data that can be visualized, and why is choosing the appropriate visual technique important?	2	K1	CO5
10.	Tell the significance of Egonet in Social Network Analysis.	2	K1	CO5

PART – B

(5 x 13 = 65 Marks)

Q.No.		Questions	Marks	KL	CO
11.	a)	Explain the Data Science process in detail, explaining each step from defining the problem to building a model. Use a practical example such as predicting house prices based on features like size, number of rooms, and neighborhood quality. Explain how to transform raw data into a useful predictive model.	13	K2	CO1
		(OR)			
	b)	A dataset contains missing values in 10% of its entries. Explain how you would handle missing data and clean it before applying a machine learning model. Perform calculations to compute the missing values using the mean for the following sample data: {5, 9, -, 6, 8}.	13	K2	CO1
12.	a)	Given a web dataset, describe the steps to analyze the data using statistical inference. Explain the Challenges of Conventional Systems.	13	K2	CO2
		(OR)			
	b)	Explain the resampling and its role in reducing prediction error. Demonstrate the calculation of prediction error using cross-validation for a regression model.	13	K2	CO2
13.	a)	Organize the concept of Principal Component Analysis with its use in data analysis.	13	K3	CO3
		(OR)			
	b)	Explain how support vector machines (SVM) are applied to classification problems and write in detail about the kernel methods.	13	K3	CO3
14.	a)	Analyze the key issues in data stream query processing and how stream queries differ from traditional queries.	13	K4	CO4
		(OR)			
	b)	Categorize the various applications of Real-Time Analytics Platform (RTAP) in data stream mining? Discuss with relevant examples.	13	K4	CO4
15.	a)	Visualize the following dataset using a bar chart and pie chart: {A: 25%, B: 35%, C: 15%, D: 25%}. Discuss in detail about how different visualization techniques provide insights into the data.	13	K2	CO5
		(OR)			

b) Explain the concept of social network analysis and how collective inferencing is used in Egonets. | 13 | K2 | CO5

PART – C

(1 x 15 = 15 Marks)

Q.No.		Questions	Marks	KL	CO
16.	a)	Consider a dataset of stock market prices for real-time analysis. Describe the steps to implement a stock market prediction model using data stream mining techniques. Include data cleansing, querying, and prediction. Given historical price data, compute the moving average and exponential smoothing for stock prices over the last 5 days: {100, 102, 104, 98, 101}.	15	K4	CO4
		(OR)			
	b)	A telecom company is facing a high customer churn rate and wants to predict which customers are likely to churn based on various customer attributes. Consider the following sample dataset:	15	K4	CO3

CustomerID	MonthlyCharges	TotalUsage	ContractType	CustomerServiceCalls	Churn
C001	50	20	Monthly	1	No
C002	80	50	Yearly	3	No
C003	60	30	Monthly	5	Yes
C004	100	70	Monthly	7	Yes
C005	40	10	Yearly	2	No
...	...	...	...	...	...

The dataset includes the following attributes:

- **CustomerID**: Unique identification number for the customer.
- **MonthlyCharges**: The customer's monthly bill amount.
- **TotalUsage**: Total data usage in GB.
- **ContractType**: Type of contract (Monthly, Yearly, etc.).
- **CustomerServiceCalls**: The number of customer service calls made by the customer.
- **Churn**: A binary variable indicating whether the customer churned (Yes/No).

Using this dataset, answer the following questions:

How to perform data exploration to understand the distribution of attributes like MonthlyCharges, TotalUsage, and CustomerServiceCalls? Explain the usage of Regression Modeling, Multivariate Analysis and Neural Networks for the above scenario.
