Total No. of Questions: 6

Total No. of Printed Pages: 2

PRN No.	

PAPER CODE	U314-232 (ESE)

(AY:2024-25) December 2024 (ENDSEM) EXAM

TY B.TECH (SEMESTER - I)

COURSE NAME: DATA SCIENCE AND MACHINE LEARNING **Branch: COMPUTER ENGINEERING** **COURSE CODE: CSUA31202**

(T.Y B.Tech PATTERN 2020)

Time: [1Hr 30 Min] **[Max. Marks: 40]**

Instructions to candidates:

1) **Figures to the right indicate full marks. Use of scientific calculators is allowed**
2) **Use suitable data wherever required.**
3) **All questions are compulsory. Solve any two sub questions each from Questions 1 and 2.**
4) **Solve any one sub question (2 marks) from Questions 3 ,4 ,5 and 6 and sub question of 4 marks is compulsory from questions 3,4,5,and 6.**

Q. No.	Question Description	Max. Marks	CO mapped	BT Level
Q1.	a) Identify challenges data scientists face during the data science process, particularly in terms of data quality and availability.	[4]	[1]	Understand
	b) Discuss steps you take in the Discovery and Data Preparation phases of the Data Analytic Life Cycle a fraud detection system for a banking application.	[4]	[1]	Understand
	c) Discuss Model Planning and Model Building phases to create a predictive maintenance system in a manufacturing plant that is experiencing unexpected machine failures with sensor data collected over the past year.	[4]	[1]	Understand
Q2.	a) Decide between removing rows with missing values or applying imputation methods for a customer churn dataset with missing demographic information. Discuss the impact on your predictive model.	[4]	[2]	Analyze
	b) Use clustering algorithms to identify and eliminate noisy outliers from a customer purchase dataset.	[4]	[2]	Analyze
	c) Investigate data cleaning and preprocessing methods to handle missing values and outliers in the Data Preparation phase on a data science project of "Predictive Analysis for Student Placement in VIIT"	[4]	[2]	Analyze
Q3.	a) Discuss the elbow method to determine the optimal number of clusters in K-Means. OR	[2]	[3]	Understand
	b) Discuss the criteria to choose a medoid within a cluster in k-medoid clustering.	[2]	[3]	Understand
	c) Given the following 5 data points and assuming K=2, assign the points to clusters and compute the final centroids after two iterations. Data points: (1,1), (2,1), (4,3), (5,4), (8,8) Initial centroids: C1 = (1,1), C2 = (5,4).	[4]	[3]	Apply

Q4.	a) Examine the effect of outliers on Naive Bayes? OR b) Examine the significance of "entropy" in the context of building a decision tree? c) Consider the following data set. **Confident** / **Studied** / **Sick** / **Result** Yes / No / No / Fail Yes / No / Yes / Pass No / Yes / Yes / Fail No / Yes / No / Pass Yes / Yes / Yes / Pass Find out whether the object with the attribute **Confident = Yes, Studies=Yes, Sick = No** will Fail or Pass using Bayesian classification.	[2] [2] [4]	[4] [4] [4]	Analyze Analyze Apply
Q5.	a) Analyze the limitations of using accuracy for classifier performance evaluation, especially in imbalanced datasets. OR b) Analyze the model's performance with a precision of 0.85 and a recall of 0.70. c) Interpret a ROC curve and the AUC (Area Under the Curve) for Spam Detection System.	[2] [2] [4]	[5] [5] [5]	Analyze Analyze Analyze
Q6.	a) Identify two effective visualization techniques for non-numerical data on any real-time application. OR b) Identify two visualization techniques for presenting numerical data on any real-time application. c) Demonstrate a scenario where data visualization significantly impacted decision-making in a business context. Describe the visualization type used and the outcome of its application.	[2] [2] [4]	[6] [6] [6]	Analyze Analyze Apply