VIT
Vellore Institute of Technology

Continuous Assessment Test (CAT) – II – MARCH 2025

Programme	:	B.Tech. CSE	Semester	:	Winter Semester 2024-25
Course Code	:	BCSE206L	Class Nbr	:	CH2024250502049 CH2024250502551 CH2024250502047
Course Title	:	FOUNDATIONS OF DATA SCIENCE			
Faculty	:	Dr. ARTHI M Dr. D. SARANYARAJ Dr. DOMINIC SAVION M	Slot	:	C1+TC1
Time	:	90 minutes	Max. Marks	:	50

Answer all the Questions

Q. No.	Sub-division	Question Text	Marks
1		A manufacturing company wants to reduce machine breakdowns by analyzing sensor data from its production line. The goal is to implement a predictive maintenance system that forecasts failures before they happen, preventing downtime and reducing maintenance costs. There are few challenges they face which are listed below, i. The company has sensor data from different machines, but some old machines do not have IoT-enabled sensors. Analyse how this missing data in the discovery phase is handled? [2] ii. The sensor readings are recorded at different time intervals and contain noise. Perform a method to maintain data consistency and remove irrelevant fluctuations? [2] iii. There is no clear threshold that separates normal machine behaviour from potential failures. Describe an approach to design an effective predictive model? [2] iv. The factory workers are not data experts. Suggest a way to present your findings and predictions that they can easily understand and use for decision-making? [2] v. outline a step-by-step approach to designing an effective predictive maintenance system. Discuss the techniques you would use to address each challenge. [2]	10
2.		Farmer Joe operates a small stall at the local farmers' market, where he primarily sells apples and bananas. To help his young son, Timmy, learn about these fruits, Joe wants to create a simple sorting system. He has observed that apples are typically red or green and roundish in shape, while bananas are usually yellow and long or curved. One morning, Timmy finds a yellow, round fruit. He is unsure if it's an apple or a banana. Joe wants to use his notebook data to help Timmy identify it. Using the data Joe collected, Predict whether the yellow and round fruit Timmy found is an apple or a banana, which includes color, shape, and fruit type, [10]	10

Fruit	Color	Shape
Apple	Red	Round
Apple	Red	Roundish
Banana	Yellow	Long
Banana	Yellow	Curved
Apple	Green	Round
Banana	Yellow	Long
Apple	Red	Roundish
Banana	Yellow	Curved

3		A university research team is analyzing feedback from students about different courses. They aim to group similar feedback into clusters based on topic (e.g., Teaching Quality, Course Content, Lab Facilities) while also assessing whether the feedback is positive, negative, or neutral. Perform NLP techniques to the following sentences	
		Review 1: "The professor explains concepts very clearly and makes learning engaging."	
		Review 2: "The lab equipment is outdated and needs urgent upgrades."	
		Review 3: "The syllabus is well-structured, but the assignments are too difficult."	
		The dataset consisting of raw student feedback, containing typos, informal language, and varying sentence structures. How would you preprocess the text to ensure meaningful sentiment words?	10
		Use a lexicon-based sentiment analysis approach where:	
		o Positive words: ["great" (+1), "excellent" (+1), "engaging" (+1), "clearly" (+1), "well-structured" (+1)] [3] o Negative words: ["outdated" (-1), "boring" (-1), "hard" (-1), "urgent" (-1), "difficult" (-1)] [3] o Compute sentiment scores for each review and classify them as Positive, Negative, or Neutral. [4]	
4		A medical diagnostic center wants to classify patients based on their blood sugar levels to identify potential risk groups. The center collected 9 patient records, each containing the fasting blood sugar (FBS) level and the HbA1c percentage (a long-term glucose indicator). The data is	

Patient	FBS Level (mg/dL)	HbA1c (%)
P1	85	4.8
P2	100	5.2
P3	115	5.7
P4	130	6.0
P5	145	6.3
P6	160	6.9
P7	175	7.3
P8	190	7.6
P9	205	8.0

shown below:

20

i. Determine the optimal number of clusters (K) using the Elbow Method by computing the within-cluster sum of squares (WCSS) for K=1 to K=5 and plotting the WCSS curve with three categories: Low Risk, Moderate Risk, and High Risk, based on their FBS and HbA1c levels. [5]

ii. Apply K-Means clustering on the given dataset using the identified K. Assign each patient to a cluster. [5]

iii. If a new patient with FBS = 125 mg/dL and HbA1c = 5.9% visits the centre, predict their risk category using the cluster centroids obtained. [5]

iv. Using a suitable Python library plot the Elbow Curve based on the given features. [5]