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: 60018

B.E. / B.Tech. DEGREE END-SEMESTER EXAMINATIONS – NOV. / DEC. 2025
Fifth Semester
Computer Science and Engineering
U23ITOE3 - DATA SCIENCE AND ANALYTICS
(Common to EEE, ECE, BME & CST)
(Regulation 2023)

Time: Three Hours | Maximum: 100 Marks

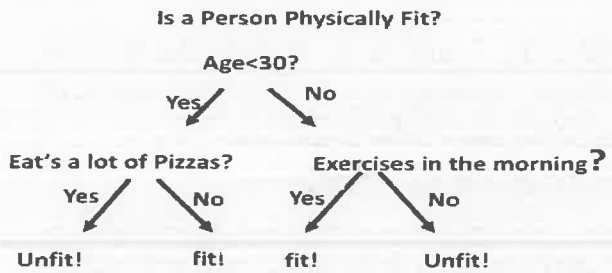
Answer ALL the questions

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.	Give an example of a dataset with a non-Gaussian distribution.	2	K1	CO1
2.	What is brushing and linking in exploratory data analysis?	2	K2	CO1
3.	Illustrate all possible decisions that can be made by the following decision tree.	2	K1	CO2



4.	Find the sample mean and median value for the best actress Oscar winner data set: 34 34 26 37 42 41 35 31 41 33 30 74 33 49 38 61 21 41 26 80 43 29 33 35 45 49 39 34 26 25 35 33.	2	K2	CO2
5.	Differentiate between traditional programming and ML in solving data problems.	2	K1	CO3
6.	Illustrate how clustering can be applied in customer segmentation.	2	K2	CO3
7.	Define length-squared sampling.	2	K1	CO4
8.	A data stream contains values [a, b, a, c, a, b]. Find the number of distinct elements.	2	K2	CO4

9.	Differentiate between structured and unstructured text data.	2	K1	CO5
10.	Show the Bag of Words representation of the sentence: "Machine learning is fun."	2	K2	CO5

PART – B

(5 x 13 = 65 Marks)

Q.No.		Questions	Marks	KL	CO
11.	a)	Explain in detail about exploratory data analysis in dataset analysis and knowledge discovery process also describe how are missing and nullified data attributes handled and modified during the preprocessing stage?	13	K1	CO1
		(OR)			
	b)	Explain the importance of visualization in EDA using real-world examples.	13	K1	CO1
12.	a)	Discuss in detail the role of machine learning in building recommendation systems with examples from Netflix, Amazon, or Spotify.	13	K1	CO2
		(OR)			
	b)	Explain the significance of model training, validation, and testing phases in the machine learning-based data science process.	13	K1	CO2
13.	a)	A dataset of 6 points in 2D space is given as: (2,3),(3,4),(5,6),(8,8),(9,9),(10,10)(2, 3), (3, 4), (5, 6), (8, 8), (9, 9), (10, 10)(2,3),(3,4),(5,6),(8,8),(9,9),(10,10) You are required to cluster the data into k = 2 clusters using the K-Means algorithm.		K2	CO3
		i. Initialize centroids at (2, 3) and (8, 8). Perform the first two iterations of K-Means, showing step-by-step calculations of distances, assignments, and centroid updates.	5		
		ii. Construct the final clusters after iteration 2.	4		
		iii. Critically analyze whether this clustering is meaningful in terms of intra-cluster similarity and inter-cluster separation.	4		
		(OR)			
	b)	Given 8 data points along a line: (1), (2), (2.1), (2.2), (8), (8.1), (8.2), (20)(1), (2), (2.1), (2.2), (8), (8.1), (8.2), (20)(1), (2), (2.1), (2.2), (8), (8.1), (8.2), (20) Parameters: Eps = 0.5, MinPts = 2		K2	CO3
		i. Identify the core, border, and noise points.	5		
		ii. Form the clusters step by step using DBSCAN rules.	4		
		iii. Illustrate the final clusters and discuss why DBSCAN is better than K-Means for this dataset.	4		

14.	a)	Describe a stream-processing method to track both distinct elements (F0) and frequent elements simultaneously and justify your design choices.	13	K2	CO4
		(OR)			
	b)	Explain the trade-off between accuracy and computational cost in matrix multiplication using sampling methods.	13	K2	CO4
15.	a)	Illustrate the process of extracting keywords from 50 customer reviews using a suitable text mining technique.	13	K1	CO5
		(OR)			
	b)	Describe a preprocessing pipeline for Reddit comments that includes tokenization, stop-word removal, stemming/lemmatization, and vectorization.	13	K1	CO5

PART – C

(1 x 15 = 15 Marks)

Q.No.		Questions	Marks	KL	CO
16.	a)	A fitness company wants to predict whether a user is at risk of high blood pressure based on daily activity data collected via a wearable device. The dataset for 10 users is given below:	15	K3	CO5

User ID	Age (Years)	Steps per Day	Average Heart Rate (bpm)	Hours of Sleep	High BP Risk (Yes=1, No=0)
1	45	5000	85	6	1
2	34	8000	70	7	0
3	50	4000	90	5	1
4	28	10000	65	8	0
5	60	3000	95	6	1
6	38	7000	75	7	0
7	42	4500	88	6	1
8	30	9000	68	7	0
9	55	3500	92	5	1
10	36	7500	72	7	0

i. Construct a Decision Tree to classify high BP risk. Use Entropy/Information Gain as the splitting criterion. Show step-by-step calculation for the first split.

ii. Draw the decision tree after two levels of splitting.

iii. Predict the risk for a new user: Age = 48, Steps per Day = 4200, Heart Rate = 87 bpm, Sleep = 6 hours.

iv. Critically analyze the strengths and limitations of using a decision tree in real-time health monitoring applications.

(OR)

b) A company wants to classify customer feedback into Positive (1) and Negative (0). The dataset contains 5 sample feedbacks: 15 K3 CO4

Feedback ID	Text	Label
1	"The product is amazing and very useful"	1
2	"I hated the delivery, it was too late"	0
3	"Amazing experience, product works great"	1
4	"Delivery was delayed and product broken"	0
5	"I love this product, very satisfied"	1

i. Preprocess the text:
   - Remove stop words (e.g., "is", "the", "was", "and", "very") and Convert all words to lowercase

ii. Construct the Bag of Words (BoW) matrix for these 5 feedbacks.

iii. Apply Stemming using Porter Stemmer and show the transformed words for each feedback.

iv. Apply Lemmatization and show the transformed words for each feedback.

v. Compare Stemming vs Lemmatization for this dataset and discuss which is better for text classification.

vi. Suggest how the BoW matrix could be used as input for a machine learning classifier to predict feedback sentiment.