VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of the UGC Act, 1956)

Reg. No. : 21BCE8062

Final Assessment Test (FAT) - May 2024

Programme	**B.Tech.**	Semester	**WINTER SEMESTER 2023 - 24**
Course Title	**MACHINE LEARNING**	Course Code	**BCSE209L**
Faculty Name	**Prof. R Jothi**	Slot	G2+TG2
		Class Nbr	**CH2023240501618**
Time	**3 Hours**	Max. Marks	**100**

General Instructions:

- Write only Register Number in the Question Paper where space is provided (right-side at the top) & do not write any other details.

Section - I

Answer all questions (7 X 10 Marks = 70 Marks)

01. Assume we have a set of data collected from patients who have visited a famous health care unit in city ABC, during the year 2023. For each patient a set of features like temperature, height, weight, blood sugar, etc. had been collected along with the corresponding disease they have been diagnosed for. In addition to these clinical features fitness features like, amount of physical activity, calorie in take, etc., had also been collected. [10]

i) Suppose you would like to use the above dataset to decide whether a new visiting patient has diabetes, heart disease, or Alzheimer (a patient can have one or more of these diseases). Which type of machine learning model you choose? Will you use a separate model for each of the diseases or a single model for all the diseases? Justify. (5 marks)

ii) Which type of machine learning model will be suitable in preparing a customized nutrition recommendation system for the patients? Briefly explain functioning of model. (5 marks)

02. A government hospital conducted a study to analyze the effectiveness of their facilities by the public. For this, they have formed a team that recorded the equipment/lab utilization on a particular month. The details of patient count and lab utilization (in Minutes, per day) are given below. [10]

Patient count	Lab utilization (in minutes)
100	90
204	130
300	150
40	40
50	65
250	140
125	100
90	70

i) Find the correlation coefficient between the patient count and lab utilization time. Comment on the above data in terms of effective lab utilization. (5 marks)

ii) Obtain the regression equation to predict the lab utilization in terms of patient count. Using the obtained regression equation predict the number of hours the lab might be utilized when there is a heavy crowd with 600 patients in a day. (5 marks)

03. Suppose you are working on a dataset to predict customer churn for a telecom company and collected data on four features namely, "Monthly Charges" , "Tenure", "Internet Service" and "Contract". [10]

Monthly Charges	Tenure	Internet Service	Contract	Churn
65	12	Fiber optic	Month-to-month	Yes
90	24	DSL	One year	No
85	8	Fiber optic	Month-to-month	Yes
75	14	None	Month-to-month	No
110	30	Fiber optic	Two years	No
50	5	DSL	Month-to-month	Yes
100	20	Fiber optic	One year	Yes
70	10	None	Month-to-month	No

i) Construct a decision tree to predict whether a customer will churn or not based on these features using the CART algorithm.(4 marks)

ii) Show the splitting criteria at each node and the resulting leaf nodes.(4 marks)

iii) Using the constructed decision tree predict whether a customer will churn or not based on the features: Monthly Charges: 75, tenure: 15 months, Internet Service: Fiber Opticand Contract: One year (2 marks)

04. "ABC" is a research agency, which collected data points (shown below) from the survey conducted to analyze the winning possibility of XYZ party in an election. After the perusal of the data, ABC found that they follow a non-linear pattern. They need the expert assistance to classify the data with larger margin to have minimum classification error, assume you are the expert and help them by answering their following queries. [10]

Data set:

Positive: (1,1),(2,2)
Negative: (4,4), (-1,-1)

i) Which algorithm is more suitable to classify? (2 marks)

ii) How do you classify them using your model? Explain the steps in detail with respect to the data points provided above. (8 marks)

05. Imagine that you are working as a data analyst for a sports university. The university maintains player statistics with features such as points scored, assists, rebounds, minutes played, player position, etc. Assume that you are given the goal to identify distinct player profiles, to organize training programs to maximize team performance. Identify the proper mechanism to arrive at appropriate groups to float training based on the sample data given below. Assume 3 groups with data points 1,3,6 as initial centroids. [10]

Points Scored	Minutes Played	Player Position	Assists
5	30	1	2
6	32	2	1
3	35	3	2
2	45	1	2
5	50	2	2
6	33	2	1
7	29	2	1
5	33	1	1

06. a) A company wants to develop a new system to group the customer preferences based on their purchase behaviour. The data they have collected includes information about customers' purchases of various products over a period of time. However, the data is incomplete owing to the reason of missing entries for some customers and products. [10]

i) How the Expectation-Maximization algorithm can be applied to handle the missing data in this scenario? (2 marks)

ii) Describe the iterative steps involved in the EM algorithm and how they would be implemented to cluster customers based on their purchasing behavior.
(3 marks)

b) Consider a binary classification scenario for weather forecasting where the two classes are "Rainy" and "Not Rainy," Use the following confusion matrix derived based on these two classes to calculate the metrics such as accuracy, precision, recall, specificity and F1-score. (5 marks)

		Actual	
		Rainy	Not Rainy
Predicted	Rainy	40	10
	Not Rainy	5	45

07. a) Mr. Sam is working on a project to develop an anomaly detection system for spotting fraudulent transactions in a financial dataset. He had decided to use a one-class SVM for this assignment. Try to project the suggestions for the following queries from the perception of Mr. Sam. (5 marks) [10]

i) Explain how a one-class SVM works in the context of anomaly detection.

ii) Given the highly imbalanced nature of the dataset (very few fraudulent transactions compared to legit ones), how would you evaluate the performance of the one-class SVM?

iii) List out few potential limitations or challenges of using a one-class SVM in this scenario. How do you address them?

b) Imagine you are organizing a trekking expedition to summit a challenging mountain peak. As the team leader, you are considering the integration of reinforcement learning techniques to assist trekkers in navigation, resource management, and safety protocols during the ascent of the mountain peak. Discus the necessary steps to be followed in training the system with this scenario. (5 marks)

Section - II

Answer all questions (2 X 15 Marks = 30 Marks)

08. a) Assume that you are an expert in stock market analysis, you should develop a predictive model for forecasting stock prices in a highly volatile market. You suspect that the traditional regression models struggle to capture the complex relationships and ensued sudden fluctuations in stock prices. Explain how to handle this by exploring the boosting algorithms to improve the predictive-accuracy of the ML model. (5 marks) [15]

b) You have a collection of weak classifiers, each based on the rating and review of a product, along with their corresponding recommendations. The table below shows the initial classification results, but some samples were misclassified by the previous weak classifier. Your task is to enhance the classification accuracy by addressing the misclassified samples using two decision stumps for all attributes, perform this classification for one cycle. (10 marks)

Rating	Review	Product Recommendation
> =4.5	Good	Yes
< 4.5	Bad	No
>=4.5	Good	No
<4.5	Bad	Yes
<4.5	Good	Yes

09. Given a collection of points in a spatial dataset, the task is to cluster together points that are closely situated, with a condition that at least three neighboring points (MinPts=3) must exist within a radius of 2 units (ε=2). B. Identify and mark outlier points that reside in sparse areas, where their nearest neighbors are too far away. [15]

Points	X	Y
P1	8	4
P2	7	5
P3	2	5
P4	6	4
P5	1	2
P6	2	10
P7	4	9