VIT®

Vellore Institute of Technology

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

Reg. No. : 23MCA0173

Final Assessment Test (FAT) - May 2024

Programme	M.C.A.	Semester	WINTER SEMESTER 2023 - 24
Course Title	MACHINE LEARNING	Course Code	PMCA507L
Faculty Name	Prof. Tulasi Prasad Sariki	Slot	B1+TB1
		Class Nbr	CH2023240501384
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.

Answer all questions (10 X 10 Marks = 100 Marks)

01. Elucidate the ethical and practical considerations associated with five key issues in machine learning: bias and fairness, privacy concerns, accountability and transparency, data quality and quantity, and model robustness and security. Provide detailed explanations for each issue and consideration, highlighting their significance in the development and deployment of machine learning systems. [10]

02. Mr. John initially learned a simple linear regression model, assuming X_1 as a feature and X_2 as a predictor. However, he later realized that the target variable is y. Now, he seeks assistance in converting the regression model into a classification model using logistic regression. Help Mr. John by completing the above steps to convert his regression model into a classification model for predicting y. [10]

Dataset:

X_1	2.5	3	3.2	3.5	4	4.5	5	5.5	5.8	6
X_2	3.5	4	3.8	4.5	5	5.5	6	6.5	6.2	6.7
y	0	0	0	0	0	1	1	1	1	1

03. Design a Decision Tree to predict the number of participants based on the given dataset. [10]

Day	Weather	Temperature	Humidity	Wind	Participants
1	Cloudy	Warm	Dry	Breezy	25
2	Sunny	Warm	Dry	Windy	30
3	Overcast	Warm	Moist	Breezy	46
4	Rainy	Moderate	Moist	Breezy	45
5	Rainy	Cool	Humid	Breezy	52
6	Snowy	Cool	Humid	Windy	23
7	Overcast	Cool	Humid	Windy	43
8	Sunny	Moderate	Dry	Breezy	35
9	Sunny	Cool	Humid	Breezy	38
10	Rainy	Moderate	Humid	Breezy	46
11	Sunny	Moderate	Humid	Windy	48
12	Overcast	Moderate	Dry	Windy	52
13	Overcast	Warm	Humid	Breezy	44
14	Rainy	Moderate	Moist	Windy	30

04. Given a dataset exhibiting high dimensionality and non-linear separability, analyze an advanced machine learning method designed to discover an optimal classification boundary for classification. Detail how this technique addresses the complexities arising from high-dimensional feature sets and intricate decision boundaries. Provide examples of real-world applications where this technique has demonstrated effectiveness in solving classification problems. [10]

05. Given the provided scenarios, select the ensemble model that best fits each of the following situations: [10]

- Customer Churn Prediction
- Credit Risk Assessment
- Medical Diagnosis

Justify your selection for each scenario based on the dataset characteristics, model requirements, and potential benefits of the chosen ensemble model.

06. Consider the data set which contains the coordinates of points in a two-dimensional plane: A: (1,1), B:(1,2), C:(2,2), D:(5,4), E:(6,4), F:(5,5), G:(5,1), H:(6,1), I:(6,2). Segment them using hierarchical clustering that uses Manhattan distance method and single linkage method for the computation. Display the dendrogram and show the cluster assignments for the three clusters. [10]

07. A database contains five transactions. The minimum support threshold is ***60%***, and the minimum confidence threshold is ***80%***. [10]

TID	Items Bought
T100	M, O, N, K, E, Y
T200	D, O, N, K, E, Y
T300	M, O, N, E, Y
T400	M, U, C, K, Y
T500	C, O, O, K, I, E

(a) Using the Apriori algorithm, find all frequent itemset. [7 Marks]
(b) List all strong association rules. [3 Marks]

08. Describe a real-world scenario where a large dataset with numerous features needs to be analyzed. Suggest a method to condense the dataset's dimensions while retaining crucial information. Discuss the advantages of this approach and provide examples of its applications in identifying patterns or anomalies within the data. [10]

09. In a dynamic environment where an agent interacts with its surroundings to achieve certain objectives, propose an advanced learning technique that enables the agent to make optimal decisions over time. Explain how this technique incorporates feedback from the environment to adjust the agent's actions and improve its performance. Provide examples of real-world applications where this approach has been successfully utilized to solve complex decision-making problems. [10]

10. Consider a scenario where you have developed a machine-learning model to classify emails as spam or not spam. After training your model, you evaluated its performance on a test set and obtained the following results: [10]

- The model correctly identified 150 spam emails out of 170 actual spam emails.
- However, it incorrectly classified 30 non-spam emails as spam.

- The model failed to identify 20 spam emails, classifying them as non-spam.
- It correctly classified 850 non-spam emails out of 880 actual non-spam emails.

Based on the provided use case, calculate the following:

a. Confusion matrix [2 Marks]

b. Accuracy [2 Marks]

c. Precision [2 Marks]

d. Recall [2 Marks]

e. F1-score [2 Marks]