VIT®

Vellore Institute of Technology

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

Reg. No. :

Final Assessment Test (FAT) - May 2024

Programme	**B.Tech.**	Semester	**WINTER SEMESTER 2023 - 24**
Course Title	**MACHINE LEARNING**	Course Code	**BCSE209L**
Faculty Name	**Prof. Dr.T Kalaipriyan**	Slot	**G1+TG1**
		Class Nbr	**CH2023240501619**
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. A dermatology clinic wants to implement a medical diagnosis system to assist dermatologists in the diagnosis of skin lesions as benign (non-cancerous) or malignant (cancerous). The goal is to provide accurate and timely diagnoses to improve patient outcomes and reduce the need for invasive procedures. The clinic collects a large dataset of clinical attributes describing the skin lesion such as lesion size, color, texture, etc., and patient demographics (name, date of birth, address, phone number, email address, gender) along with corresponding biopsy results: benign or malignant. [10]

Suppose you have been deployed as a machine learning engineer to help the dermatologists by developing a suitable machine learning model for predicting whether skin lesion is cancerous or not.

i. What machine learning paradigm you choose for the above task? Justify. (2 marks)
ii. Give a sample dataset to be used by the model that you choose in the above question. Clearly mention independent and dependent variables in the dataset. (3 marks)
iii. Explain briefly any two preprocessing activities to be applied on the dataset in order to improve performance of the model. (3 marks)
iv. Give any two metrics to evaluate whether the model that you have built is performing well or not. (2 marks)

02. The below dataset mentioned in Table 1 is designed to aid in predicting the type of insect based on a set of characteristic features such as the Antennae_length, Eye_color, Hair_length and Presence_of_Wings. [10]

Table 1

Antennae_length	Eye_color	Hair_length	Presence_of_Wings	Type
Low	Blue	Short	Yes	Type-1
High	Brown	Long	No	Type-2
Low	Blue	Long	Yes	Type-2
Medium	Brown	Short	No	Type-1
High	Yellow	Long	Yes	Type-1
High	Brown	Long	Yes	Type-1
Low	Yellow	Short	No	Type-2
Medium	Blue	Long	Yes	Type-2
High	Brown	Short	No	Type-1
Medium	Blue	Long	No	Type-2

Apply a suitable probabilistic classifier using the above training dataset to predict the type of insect given the features < Antennae_length = High, Eye_color =Yellow, Hair_length=Short, Wings=yes>. [10]

03. Consider the dataset mentioned below.

Class (+1) : (2,2), (1,3), (4,2), (3,3)

Class (-1) : (-1,-1), (0,0), (-4,-1) , (-2,-2)

Suppose that you are tasked with constructing a maximal margin classifier for the binary classification problem mentioned above.

i. Identify the key data points from the given set of eight that are essential for defining the decision boundary. Provide a justification for your selection. (2 marks)

ii. Plot the data points and establish the decision boundary. Ensure to demonstrate step by step computations involved in determining the decision boundary. (6 marks)

iii. Compute the training error of your model. (2 marks)

04. A military intelligence agency is developing a vehicle maintenance plan to efficiently manage their fleet. They intend to categorize vehicles into distinct groups. This segregation is based on historical data regarding military vehicle operations, encompassing factors such as battery strength, distance travelled and fuel consumption. Below is a sample dataset mentioned in Table 2 containing this information for a subset of vehicles. [10]

Table 2

Vehicle Number	Battery strength	Distance Travelled (km)	Fuel Consumption (lr)
A	6	10	10
B	4	5	8
C	8	15	12
D	7	12	9
E	5	9	7
F	8	20	15

Apply a suitable clustering algorithm that merges the most similar pair of vehicle clusters at a time on the given dataset in order to partition the vehicles into distinct groups. When you are merging the two clusters, distance between these two clusters is determined by a pair of vehicles (one in each cluster) that are closest to one another. Compute each step of the algorithm, showcasing the calculations involved, and draw the final dendrogram tree.

You are given with a dataset mentioned in Table 3 for disease severity classification, focusing on two symptoms. However, there is a concern regarding the impact of one of the symptoms, which might be misleading for classification purposes. To address this issue and improve the accuracy of classification, the dimensionality of the dataset needs to be reduced. Apply a suitable algorithm to extract principal components from the given data and project the given data onto the 1-D space with the identified component. Elucidate the steps in detail [10]

Table 3

Symptom1	Symptom2
1	5
4	2
3	6
7	1

06. An automotive assembly plant is facing challenges in accurately diagnosing faults in its production line, leading to increased downtime, waste, and operational costs. The plant produces various automotive components, and identifying faults early in the manufacturing process is crucial to ensure product quality and prevent defective products from reaching customers. [10]

10]

As a data scientist working in the plant, you have developed a classification model for detecting the manufacturing faults in automotive components based on sensor data collected from the production line. The dataset of sensor data contains instances of both faulty and non-faulty components. However, due to the rarity of faults compared to normal operation, the majority of instances are non-faulty (negative class) and only a small percentage of instances are faulty (positive class).

i. What are the effects of nature of the dataset considered in this scenario on performance of the model that you have developed? (2 marks)

ii. How will you address this issue? Explain with suitable algorithm. (3 marks)

iii. Suppose you have eradicated the identified effect using suitable algorithm in part ii of this question. Now, you observe that your model overfits the data. Briefly explain any two methods to reduce overfitting on the current scenario. (5 marks)

07. Imagine a company that operates a fleet of autonomous delivery robots to transport packages from a distribution center to various destinations within a city. The company aims to optimize the routing and scheduling of these robots to minimize delivery times and energy consumption, while maximizing customer satisfaction. The environment consists of a city map with roads, intersections, buildings, and delivery destinations. Each delivery destination has a specific delivery time window and package size. The delivery robots have limited battery capacity and must recharge at designated charging stations periodically. The environment also includes dynamic factors such as traffic conditions, pedestrian traffic, and weather conditions. [10]

Propose and elucidate a framework, based on the given scenario, that is capable of autonomously learning the most efficient delivery routes and scheduling policies for a fleet of delivery robots. The proposed framework should encompass a policy, a reward function, a value function, and a model of the environment.

Section - II

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

08. Consider the following dataset that is used for predicting whether a fruit is edible or not based on its features namely shape, colour, odour and texture. [15]

Table 4

Shape	Colour	Odour	Texture	Edible
Irregular	White	No	Smooth	Yes
Regular	Yellow	No	Non-smooth	Yes
Regular	White	No	Non-smooth	Yes
Regular	White	Mild	Smooth	Yes
Irregular	Yellow	Mild	Smooth	Yes
Regular	Yellow	Mild	Smooth	No
Regular	Green	Mild	Non-smooth	No
Irregular	Yellow	Mild	Smooth	No
Irregular	Yellow	Strong	Non-smooth	No
Irregular	White	Strong	Smooth	No
Regular	White	Strong	Non-smooth	?

i. Apply a tree based classifier for predicting edibility of the fruit using a suitable algorithm that uses information gain based on entropy. Identify the best feature for splitting at root level of the decision tree using the training samples given in the Table 4. (5 marks)

ii. Build a decision tree model for the above dataset and draw the complete decision tree. You need to show all the calculations in obtaining the tree. Also predict whether a test fruit sample <shape =Regular, Colour=White, Odour=Strong, Texture=Non-smooth> is edible or not. (10 marks).

09. xTel, a manufacturing firm, plans to implement intelligent solutions for their IoT infrastructure. They started collecting various information from devices including fwd_pkts_sec (number of packets received per second in forward flow), bwd_pkts_sec (number of packets received per second in backward flow), ratio (ratio of downloads and uploads per second) along with the type of network behavior (normal/abnormal) in that second. [15]

The data is analyzed to follow a non-linear pattern. Given the values of features in a second 't' as below (Table 5), simulate the working of a multi-layer perceptron model (one forward pass and one backpropagation).

Assume,

- Learning rate as 0.1
- All weights as 0.5
- Bias as 0
- One hidden state with two nodes

Table 5

fwd_pkts_per_sec (1000k)	bwd_pkts_per_sec (1000k)	down_up_ratio	Network Behaviour
1	1	0.1	0