Reg. No.:

Name :

VIT

Vellore Institute of Technology

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

Continuous Assessment Test I – February 2024

Programme	:	B. Tech (CSE)	Semester	:	Winter 2023-24
Course Title	:	Machine Learning	Code	:	BCSE209L
			Slot	:	G1 + TG1
Faculty Name	:	Suganya G, Subbulakshmi T, Rajalakshmi R, R Jothi, Kalaipriyan, Manimegalai, Trilok Nath Pandey, D Jeya Mala	Class Nbr	:	CH2023240501634, CH2023240501626, CH2023240501605, CH2023240501617, CH2023240501619, CH2023240501620, CH2023240501622, CH2023240501630
Duration	:	1 hour 30 mins	Max. Marks	:	50

Answer all the Questions

S.No		Description	Marks
1.	a)	Assume that, you are a data scientist at a large retail company. The company has outlets all over the country and you have been given the task of analyzing the sales data for the past 5 years. Identify any suitable machine learning technique to forecast the future sales amount of the company, in the next financial year. Justify your answer. (2 marks)	[10]
	b)	If there are inconsistencies in the sales data, say 5% of total entries contain data entry errors in the fields viz., phone numbers of the customers and item code. How would you address this issue ? You have also observed that, there are few fraudulent transactions. Suggest a suitable ML approach to determine these kinds of transactions. (2 marks)	
	c)	Self-driving vehicle company X uses few sensors and collects data of the car's surrounding environment in real time. This data helps guide the car's response in different situations, whether it is a human crossing the street, a red light, or another car on the highway. Which type of ML algorithm might be employed by the company X, to guide car's real-time response based on it environment. Justify your answer. (2 marks)	
	d)	Assume that, an ML system was applied to predict the cancer patients based on few health conditions. Given below is the result of that system. Calculate the precision, recall, accuracy and false positive rate. Also, comment on the reliability of the system. (4 marks)	

Actual / Prediction	Not Cancer	Cancer
Not Cancer	50	10
Cancer	5	100

2. Assume that, in a Chemical Manufacturing Company, three different chemicals are formed with the compositions of two substances x1 and x2 in different ratios of the quantity. During further research, the company has found a new chemical P, with a composition (1.7, 1.8), and the company wants to place the new chemical under any of the suitable existing categories of chemicals if its composition is close to 4 known chemicals. As the data is huge, they want some intelligent assistance to analyse the data and provide the solution by checking the similarity between the already known chemicals and the new one. As an ML expert, your task is to apply a suitable technique to determine the category of the new chemical P. Elaborate the chosen algorithm and illustrate the steps in detail. [10]

Chemical	Substances composition	Class Label
A1	(1.5,1.5)	C2
A2	(0.3, 0.8)	C1
A3	(2.0,2.3)	C3
A4	(1.9, 1.4)	C2
A5	(0.3, 0.4)	C1
A6	(1.7, 1.5)	C2
A7	(2.0, 2.8)	C3
A8	(1.2, 1.5)	C2
A9	(1.8, 1.5)	C2
A10	(0.5, 0.1)	C1
A11	(2.4, 2.8)	C3
A12	(1.8, 1.4)	C2

3. Consider a scenario where a company wants to understand the relationship between the years of experience and salary of its employees. The company has the following data for 10 employees: **Years of Experience**: 2, 3, 4, 6, 8, 10, 11, 15, 16, 17 and **Salary** (in thousands): 30, 35, 37, 42, 45, 50, 55, 60, 65, 70 [10]
What kind of relationship exists between the above two variables? How can this information help the company in its future hiring and salary negotiation processes? Apply a suitable ML algorithm to determine the expected salary of a person, if he has 13 years of experience.

4. You are given a dataset from a Micro Biology Lab, that describes whether the substance is Poisonous or not. The characteristics of the substances viz., Colour, Toughness, Fungus and Appearance are assumed to be independent of each other. Using the given data, design an appropriate probabilistic based classifier to determine whether a substance having the following characteristics is poisonous or not. [10]
[Colour: Green; Toughness : Hard; Fungus: Yes ; Appearance = Smooth]

Colour	**Toughness**	**Fungus**	**Appearance**	**Poisonous**
Green	Hard	No	Smooth	No
Green	Hard	Yes	Smooth	No
Brown	Soft	No	Wrinkled	No

			Orange	Hard	No	Wrinkled	Yes		
			Green	Soft	Yes	Smooth	Yes		
			Green	Hard	Yes	Wrinkled	Yes		
			Orange	Hard	No	Wrinkled	Yes		
8.		Decision trees are preferred for easy interpretability. Brief the algorithm of constructing a decision tree. Also elaborate the different criteria, viz. Information gain and Gini impurity for selecting the suitable attribute and splitting the data into partitions. Write the pros and cons of tree-based algorithms. Assume the example data with at least 5 samples and 4 attributes. [Alg: 3 marks, IG and Gini : 4 marks, Pros and Cons: 3 marks]							[10]

⇔⇔⇔

Reg. No.: 21BAI1735

Name :

VIT

Vellore Institute of Technology

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

Continuous Assessment Test 2 (CAT2) – July 2023

Programme	:	**B.Tech. CSE (AI & ML)**	Semester	:	Fall Inter Sem. 2022-23
Course Code	:	**BCSE209L**	Class Nbr(s)	:	CH2022232501020/ CH2022232501015/ CH2022232501164/ CH2022232501016/ CH2022232501018
Course Title	:	**Machine Learning**			
Faculty(s)	:	**Dr S.K. Ayesha / Dr R. Bhargavi /Dr R. Jothi / Dr. D Mansoor Hussain/ Dr. P. Prakash**	Slot	:	**B2**
Time	:	**90 Minutes**	Max. Marks	:	**50**

Answer all the Questions

Q. No.	Question Text	Marks
1.	a) A fully connected neural network has the following specifications: (Multi Layer) i. 4 Inputs x_1, x_2, x_3, x_4 ii. 2 Hidden Layers iii. Hidden Layer 1 is made up of neuron with function a_j and sigmoid function b_j: $a_j = \sum_i^4 W_{i,j}^1 x_i + \beta_j^1$ and $b_j = \sigma(a_j)$ iv. Hidden Layer 2 is made up of neuron with function c_k and sigmoid function d_k: $c_k = \sum_j^4 W_{j,k}^2 b_j + \beta_k^2$ and $d_k = \sigma(c_k)$ v. The output layer is made up of $e = \sum_k^3 W_k^3 d_k + \beta^3$ and $o = \sigma(e)$ vi. Include the Weight terms in all layers for all the connections marking them as $W^m_{i,j}$ with m as their layer number, and i,j are neurons which are connected with the weight $w_{i,j}$ vii. Include the bias terms in all layers marking them as β^m with m as their layer number. Draw a clean architecture of the above specified network. Make sure you label your drawing appropriately. [04 Marks] b) What can we say about this Neural Network if we replace the activation functions with $b_j = a_j$, $d_k = c_k$ and $o = e$. Specifically what other ML method can this be compared to? Explain your answer in one or two short sentences. [03 Marks] c) Consider k-fold cross-validation. Let's consider the trade-offs of larger or smaller k (the number of folds). With a higher number of folds, the estimated error will be, on average, A. Higher. B. Lower. C. Same. D. Can't tell. Give justification for your choice in one or two sentences. [03 Marks]	10

2. You are given the following dataset and asked to create a decision tree to predict whether a given mushroom with certain features is poisonous

	Colour	Height	Stripes	Texture	Poisonous?
1	Purple	Tall	Yes	Rough	Yes
2	Purple	Tall	Yes	Smooth	Yes
3	Red	Short	Yes	Hairy	No
4	Blue	Short	No	Smooth	No
5	Blue	Short	Yes	Hairy	Yes
6	Red	Tall	No	Hairy	No
7	Blue	Tall	Yes	Smooth	Yes
8	Blue	Short	Yes	Smooth	Yes
9	Blue	Tall	No	Hairy	No
10	Blue	Short	Yes	Rough	Yes
11	Red	Short	No	Smooth	No
12	Purple	Short	No	Hairy	Yes
13	Red	Tall	Yes	Hairy	No
14	Purple	Tall	Yes	Hairy	Yes
15	Purple	Tall	No	Rough	No
16	Purple	Tall	No	Smooth	No

Identify attribute at the root node of the decision tree for the above set of data, using information gain measure for attribute selection. [14 Marks]

14

3. To track the root cause and spread of a virus, data from different patients affected by the virus is collected. The distance matrix computed using Euclidean distance for a subset of the patient dataset is given below. Draw the dendrogram for the given data and elaborate on the entire process of clustering. Apply agglomerative clustering to arrive at 2 clusters. (use single linkage). Also, indicate the clusters by listing all the cluster elements under each cluster if you cluster the same dataset into 3 clusters.

(Mod 4)

	1	2	3	4	5	6	7	8
1	0	9	3	6	11	5	13	17
2		0	7	5	10	4	26	9
3			0	9	2	15	18	16
4				0	8	22	25	12
5					0	5	7	9
6						0	6	14
7							0	23
8								0

14

4. Consider a multiclassification model which can classify the images of a Bus, Car, Van, and Auto. The following table gives the ground truth and the predictions by the model.

Sl. No	Actual	Predicted
1	Bus	Bus
2	Car	Auto
3	Van	Van
4	Car	Car
5	Car	Auto
6	Bus	Auto
7	Auto	Auto
8	Car	Bus
9	Bus	Bus
10	Car	Car
11	Van	Bus
12	Van	Van
13	Bus	Van
14	Car	Car

i. Write the confusion matrix for the model. [2 Marks]

ii. Compute F1 score for the individual classes and the overall F1 score for the model. [10 Marks]

12

Reg. No.:

Name :

VIT

Vellore Institute of Technology

Continuous Assessment Test II – April 2024

Programme	:	B. Tech (CSE)	Semester	:	Winter 2023-24
Course Title	:	Machine Learning	Code	:	BCSE209L
			Slot	:	G1 + TG1
Faculty Name	:	Dr. Suganya G Dr. Subbulakshmi T Dr. Rajalakshmi R Dr. R Jothi Dr. Kalaipriyan Dr. Manimegalai Dr. Trilok Nath Pandey Dr. D. Jeya Mala	Class Nbr	:	CH2023240501634, CH2023240501626, CH2023240501605, CH2023240501617, CH2023240501619, CH2023240501620, CH2023240501622, CH2023240501630
Duration	:	1 hour 30 mins	Max. Marks	:	50

Answer all the Questions

S.No		Description	Marks
1		You are given with n features of a dataset that are un-labelled in nature. And on interpretation of the collected data you have figured out that may be helpful, if you could cluster them first to interpret the relationship among them. To begin with, you have taken 2 features from that data and tried to project in a 1-D space to have a clear idea, as you know reducing the dimensions to get the principal components will help to understand the data better. Apply this technique on the below data and elucidate the steps in detail along with the identified component and the projected data onto the 1-D space.	15

F1	F2
8	4
14	7
11	3
5	1
9	2

2		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 taxonomy clustering that uses Euclidian distance method and complete linkage method for the computation. Display the dendrogram and show the cluster assignments for the three clusters.	15
3		Based on the provided scenarios, choose the most suitable ensemble model for each of the following scenarios: • 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.	10
4		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 for this complex data. Provide examples of real-world applications where this technique has demonstrated effectiveness in solving classification problems.	10

(4-she)

⇔⇔⇔

VIT®

Vellore Institute of Technology

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

Reg. No. : 21BCE1808

Max: 84

Final Assessment Test (FAT) - May 2024

Programme	**B.Tech.**	Semester	**WINTER SEMESTER 2023 - 24**
Course Title	**MACHINE LEARNING**	Course Code	**BCSE209L**
Faculty Name	**Prof. TRILOK NATH PANDEY**	Slot	**G1+TG1**
		Class Nbr	**CH2023240501622**
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.

(2)

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]**

(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>.

03. Consider the dataset mentioned below. **[10]**

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.

05. 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]**

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

⇦⇨⇦

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of the UGC Act, 1956)

Reg. No. : 22BAI1429

Final Assessment Test (FAT) - May 2024

Programme	**B.Tech.**	Semester	**WINTER SEMESTER 2023 - 24**
Course Title	**MACHINE LEARNING**	Course Code	**BCSE209L**
Faculty Name	**Prof. Mohan R**	Slot	**C1+TC1**
		Class Nbr	**CH2023240503074**
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

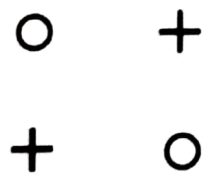
Answer all questions (5 X 10 Marks = 50 Marks)

01. A food delivery company which performed well for 2 years is not able to perform well in recent months as they are not able to deliver food on time. **[10]**

As a result, their customers are unhappy. So, the company has requested you to help them in improving their business.

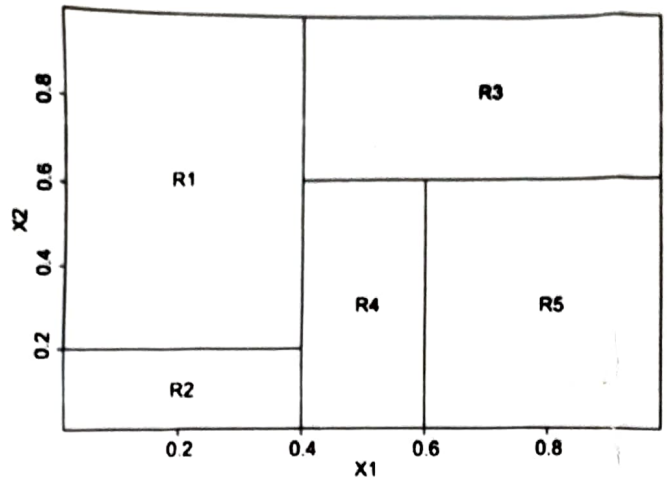
As a machine learning expert, provide them with any 5 major solutions along with corresponding ML algorithms that could be applied for each of the scenario to help them recover from this situation?

02. i) Consider the following dataset represented as data points: **[10]**



Discuss how classifiers like Logistic regression and 3-NN can classify the train data? Is it possible for the above list of classifiers to achieve zero training error on this data set. Justify your answer. (5 Marks)

ii) Draw a decision tree having zero training error for the given scenario having 5 regions R1, R2, R3, R4, R5 (which are to be the leaf nodes in the DT) using the appropriate attribute ranges. (5 marks)



03. [10]

No	Petal Length	Petal Width	class
1.	4.8	1.8	Iris-virginica
2	3.7	1.2	Iris-setosa
3	4.5	1.4	Iris-setosa
4	3.9	1.2	Iris-setosa
5	4.1	1.3	Iris-setosa
6	3.6	1.6	Iris-setosa
7	5.5	1.8	Iris-virginica
8	3.5	1.3	Iris-setosa

i) Analyze the given dataset and check whether any ML classifier model applied on it can produce consistent level of accuracy. Justify your answer. (3 marks)
ii) Recommend a suitable mechanism to get the best possible accuracy for the model. Show the step-by-step process of applying the recommended mechanism to solve the problem in the above given data set. (7 marks)

04. A telecom company is willing to share the following details with you, Data were provided for 1,00,000 customers with last 6 months of service history [10]
– Having churners and non-churners
– Type and price of current handset
– Date of last handset change/upgrade
– Total revenue
– Call behaviour statistics (type, number, duration, totals, etc.)
– Demographic information
based on the above information, answer the following question.
Assume that your ML algorithm has achieved 70 % accuracy even after trying with various hyper parameter tuning, but you are required to achieve a minimum 80% accuracy.
i) Recommend possible mechanisms that can improvise on the accuracy to 80%. (3 marks)
ii) For the above scenario show the working of any one recommended approach with proper illustration. (7 marks)

05. i) Justify the need for generating optimal set of actions for a given goal based agent. (3 marks) [10]
ii) Illustrate with a suitable example for at least 2 iterations an off-policy based approach to achieve the optimal path with max rewards (7 marks)

Section - II

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

06. A company has 6 equipments and each fitted with 3 sensors. They have collected equipment failure details based on data referencing observations of past machine runs and failures as shown in the table below. Use this information to answer below questions. [15]

Table: Machine Failure Status data

Machine Id	Sensor1	Sensor 2	…	Sensor 300	Machine Failure
M1	0.310	.399	…	0.333	No
…	…	…	…	…	…
M500	0.748	0.329	…	0.938	Yes

Assume that there are equal number of class labels in the above data and are linearly separable.

i) Draw a Neural Network architecture with corresponding activation function to classify the data. Also, mention clearly the steps on how the NN you suggested learns from its erroneous predictions. (7.5 marks)

ii) Provide detailed description of a suitable classifier that can provide a separator which can maximize the distance of the closest datapoint from the decision line / hyperplane. (7.5 marks)

07. a) Consider the table Machine failure status data. You are asked to perform unsupervised learning on this data set. In this context, identify the problem posed by the dataset. Explain briefly a possible solution to overcome the problem posed by the data.(3 marks) [15]

Table: Machine Failure Status data

Machine Id	Sensor1	Sensor 2	…	Sensor 300	Machine Failure
M1	0.310	.399	…	0.333	No
…	…	…	…	…	…
M500	0.748	0.329	…	0.938	Yes

b) Assume that you have been assigned a task of grouping similar customers from the table computer purchase. For this given task, identify the suitable grouping algorithm and justify your choice. (3 marks)

c) Consider the customer samples ID1, ID4, ID12 as 3 initial points for grouping. Apply the algorithm identified in (b) and show step by step procedure to obtain the groups formed upto 2 iterations.(9 marks)

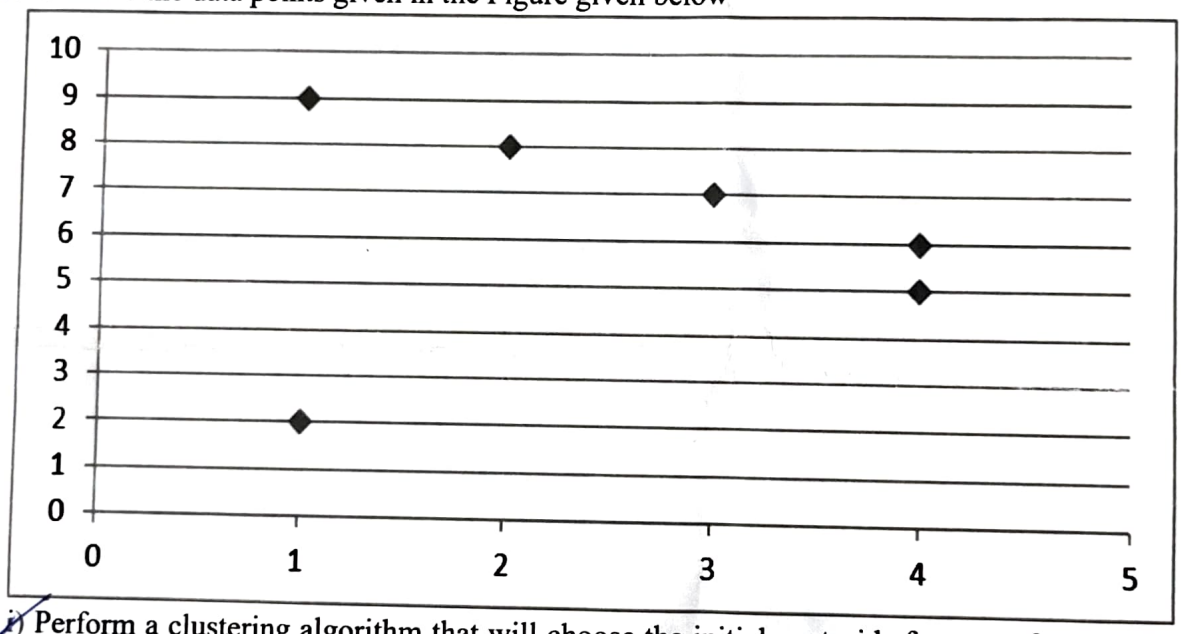
Table: Computer purchase

ID	AGE	INCOME	STUDENT	CREDIT	BUYS
1	< 31	high	no	bad	no
2	< 31	high	no	good	no
3	31 – 40	high	no	bad	yes
4	> 40	med	no	bad	yes
5	> 40	low	yes	bad	yes
6	> 40	low	yes	good	no
7	31 – 40	low	yes	good	yes
8	< 31	med	no	bad	no
9	< 31	low	yes	good	yes
10	> 40	med	yes	bad	yes
11	< 31	med	yes	good	yes
12	31 – 40	med	no	good	yes
13	31 – 40	high	yes	bad	yes
14	> 40	med	no	good	no

Section - III

Answer all questions (1 X 20 Marks = 20 Marks)

08. Consider the data points given in the Figure given below **[20]**



i) Perform a clustering algorithm that will choose the initial centroids far away from each other. (10 marks)

ii) Use Self organizing map with the above data and compare the SOM output with that of clustering output and provide the inference. (10 marks)

Note - Do the above computations for 1 iteration.

⇔⇔⇔