MANGALORE INSTITUTE OF TECHNOLOGY & ENGINEERING
(A Unit of Rajalaxmi Education Trust®, Mangalore)
Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE, New Delhi
Accredited by NAAC with A+ Grade & ISO 9001:2015 Certified Institution

Model Question Paper

Second Semester MCA Degree Examination

Machine Learning

Time: 3 Hours **Max. Marks: 100**

Note: 1. Answer any FIVE full questions, choosing ONE full question from each module.
2. M: Marks, L: RBT (Revised Bloom's Taxonomy) level, C: Course outcomes.

<table>
<tr><th colspan="3">Module -1</th><th>M</th><th>L</th><th>C</th></tr>
<tr><td rowspan="2">Q1</td><td>a.</td><td>A marketing team is analyzing the success of their email campaign. Apply Logistic regression to predict whether an email will be clicked (1) or not clicked (0) based on the following two features: Email Length (in characters). Time Sent (morning = 1, afternoon=0) and interpret the same.
The dataset below shows the email length, time sent, and whether the email was clicked:
<table><tr><th>Email Length</th><th>Time Sent</th><th>Clicked (1/0)</th></tr><tr><td>50</td><td>1</td><td>0</td></tr><tr><td>60</td><td>0</td><td>0</td></tr><tr><td>100</td><td>1</td><td>1</td></tr><tr><td>120</td><td>1</td><td>1</td></tr><tr><td>90</td><td>0</td><td>0</td></tr><tr><td>150</td><td>0</td><td>1</td></tr><tr><td>200</td><td>1</td><td>1</td></tr></table></td><td>10</td><td>L3</td><td>C1</td></tr>
<tr><td>b.</td><td>You are a data scientist working for an automobile company. Apply multilinear regression model to predict the car price based on these features such as engine size, horsepower, and mileage.
The dataset includes the following features:
1. Engine Size (in liters): The engine size of the car in liters.
2. Horsepower (hp): The horsepower of the car.
3. Mileage (km per liter): The fuel efficiency of the car in kilometers per liter.
4. Car Price (Target): The price of the car (in $1000).</td><td>10</td><td>L3</td><td>C1</td></tr>
</table>

Car_ID	Engine_Size (liters)	Horsepower (hp)	Mileage (km/l)	Car_Price ($1000)
C001	2.0	150	12	220
C002	1.8	120	14	190
C003	3.0	180	10	280
C004	2.5	170	11	240
C005	2.2	160	13	230
C006	1.6	110	15	180
C007	3.2	200	9	300
C008	2.7	175	11	260
C009	1.9	130	14	200
C010	2.1	155	12	215

OR

Q2	a.	You are working for a **retail company** that wants to understand the relationship between **advertising spending** and **monthly sales**. The company collects data for 10 months on how much they spend on advertising and the corresponding sales for each month. Apply simple **linear regression model** to predict sales based on advertising spending. Interpret the result.	10	L3	C1

Month	Advertising Spend (in $1000s)	Sales (in $1000s)
1	5	10
2	8	15
3	7	14
4	10	20
5	6	12
6	9	18
7	4	8
8	11	22
9	3	6
10	12	24

	b.	Construct a **credit scoring model** that can classify customers as either **good credit risk (1)** or **bad credit risk (0)** based on a given dataset. Give Illustrate how this model works for the given dataset. Explain how the model will make predictions based on the features provided. What metric(s) would you use to evaluate the performance of the model and why?	10	L3	C1

Age	Annual Income ($1000s)	Loan Amount ($1000s)	Credit History (1 = Good, 0 = Bad)	Number of Credit Cards	Credit Risk (1 = Good, 0 = Bad)
25	40	10	1	2	1
45	100	20	1	4	1
34	50	15	0	3	0
28	60	5	1	2	1
55	120	30	1	5	1
39	70	10	0	3	0
23	35	8	1	2	1
50	95	25	0	4	0
30	80	12	1	3	1
40	85	22	0	3	0

Module- 2

Q3	a.	Apply Gaussian density function to height data for adult men in a specific country. Explain how you would estimate the mean and variance of the Gaussian distribution using Maximum Likelihood Estimation (MLE). Discuss the approach for assessing the goodness-of-fit for the Gaussian model and how you would interpret the results.	10	L3	C2
	b.	You are developing a classification model to predict whether a patient has a specific disease based on medical test results. Apply Bayesian Decision Theory to this medical classification problem and explain how you would use Bayes' Theorem to calculate the posterior probability of the disease given the test results and make a classification decision. What steps would you follow to ensure accurate predictions?	10	L3	C2

Height (cm)
172
168
181
175
169
177
173
178
182
174

Patient_ID	Test_Result_1	Test_Result_2	Test_Result_3	Has_Disease (1 = Yes, 0 = No)
1	0.85	0.75	0.90	1
2	0.40	0.35	0.50	0
3	0.95	0.80	0.88	1
4	0.30	0.25	0.40	0
5	0.70	0.60	0.80	1
6	0.45	0.50	0.55	0

- **Test_Result_1, Test_Result_2, Test_Result_3**: Represent the outcomes of different medical tests (e.g., blood pressure, cholesterol levels, or other diagnostic measures).
- **Has_Disease**: A binary label indicating whether the patient has the disease (1 = Yes, 0 = No).

OR

Q4 a. Make use of the given dataset which containing information about customers, their credit scores, and their historical default rates to assess the expected loss from high-risk and low-risk customers. Explain the metrics you would use to measure expected loss for each class and how you would leverage this information for informed business decisions. (10, L3, C2)

Customer ID	Credit Score	Default Rate	Risk Category
1	750	0.01	Low Risk
2	620	0.15	High Risk
3	680	0.05	Low Risk
4	550	0.25	High Risk
5	700	0.07	Low Risk
6	580	0.20	High Risk
7	720	0.03	Low Risk
8	600	0.18	High Risk
9	690	0.08	Low Risk
10	540	0.30	High Risk

	b.	Make use of multi-class classification algorithm to classify cars into categories such as "Sedan," "SUV," "Truck," "Coupe," and "Convertible" based on features like engine size, weight, and fuel efficiency. Apply the process of data preparation, model training, and performance evaluation. Additionally, explain how you would address missing or inconsistent feature values in the dataset.	10	L3	C2

Car ID	Car Type	Engine Size (L)	Weight (kg)	Fuel Efficiency (mpg)
1	Sedan	2.0	1400	30
2	SUV	3.5	2000	22
3	Truck	5.0	2500	15
4	Coupe	2.5	1300	28
5	Convertible	3.0	1500	25
6	Sedan	1.8	1350	32
7	SUV	4.0	2100	20
8	Truck	6.0	2700	12
9	Coupe	2.2	1250	29
10	Convertible	3.5	1550	23

Module – 3

Q5	a.	Apply histogram estimator in the distribution of customer ages from a recent survey dataset given. Explain how you would determine the appropriate bin width and number of bins for the histogram. Discuss how you would interpret the histogram and utilize it to understand the distribution of customer ages.	10	L3	C3

CustomerID	Age	Gender	Income	SubscriptionType
1	22	Female	30000	Basic
2	34	Male	45000	Premium
3	29	Female	38000	Basic
4	45	Male	70000	Premium
5	31	Female	52000	Standard
6	27	Male	32000	Basic
7	40	Female	60000	Premium
8	55	Male	80000	Standard
9	33	Female	48000	Basic
10	50	Male	75000	Premium

	b.	You are working for a fitness centre, and your task is to classify new members into two categories based on their fitness level: **Fit (1)** or **Not Fit (0)**. The dataset contains the following features: 1. **Age**: The age of the member (in years). 2. **Daily Activity**: The average daily activity level in minutes. 3. **BMI**: The Body Mass Index (BMI) of the member. 4. **Fitness Level (Target)**: Whether the member is fit (1) or not fit (0). **Apply the k-NN Algorithm**: Implement the k-NN algorithm on the dataset to classify the fitness level of a new member with the following information: • **Age**: 32	10	L3	C3

- **Daily Activity**: 50 minutes
- **BMI**: 25.5

Use **k=3** for the classification. What is the predicted fitness level for this new member?

Member_ID	Age	Daily_Activity (min)	BMI	Fitness_Level (Target)
M001	25	60	22.5	1
M002	34	30	28.0	0
M003	45	40	26.5	0
M004	23	80	21.0	1
M005	36	20	30.0	0
M006	29	50	23.5	1
M007	40	35	27.0	0
M008	27	75	22.0	1
M009	38	25	29.0	0
M010	26	65	24.0	1

OR

Q6

a. Construct a univariate decision tree to classify individuals as "Tall" or "Short" based on a single feature, such as height. Explain how you would determine the threshold value for splitting and describe how you would interpret the resulting decision tree. — 10 | L3 | C3

Person ID	Height (cm)	Class (Tall/Short)
1	160	Short
2	155	Short
3	170	Tall
4	180	Tall
5	165	Short
6	175	Tall
7	158	Short
8	182	Tall
9	150	Short
10	178	Tall

b. Using regression tree predict product sales based on features such as based on various features like **Advertising Spend**: Amount spent on marketing (in $),**Product Price**: Price of the product (in $),**Customer Rating**: Average customer rating of the product (out of 5),**Discount Offered**: Percentage discount offered on the product. Explain how the structure of the regression tree reveals insights into the impact of each feature on sales figures and discuss the potential benefits of these insights for strategic planning, — 10 | L3 | C3

Product_ID	Advertising_Spend ($)	Product_Price ($)	Customer_Rating	Discount (%)	Sales (Units)
P001	5000	60	4.5	10%	200
P002	3000	40	4.0	15%	150
P003	7000	80	5.0	5%	250
P004	2000	35	3.5	20%	120
P005	4500	55	4.2	12%	180
P006	6000	65	4.8	8%	230

Module – 4

Q7	a.	Make use of Multilayer Perceptron(MLP) to train the machine learning model to predict hous prices using various features (square footage, number of rooms, etc.) using the given dataset c housing prices. Explain how the MLP works to learn the relationship between input feature and output predictions. Use examples from the dataset to explain how the model generalizes t unseen data.	10	L3	C4
	b.	You are provided with a binary classification dataset (determining whether a customer will purchase a product based on features). Apply perceptron to train the machine learning model to classify the data. Explain how a perceptron learns to separate the two classes and how the weight updates occur during training and the effectiveness of the perceptron for linearly separable and non-linearly separable data.	10	L3	C4

Square Footage	Number of Rooms	Age (Years)	Location Score	Price (in $)
1200	3	10	7	250,000
1500	4	5	8	350,000
900	2	20	5	150,000
1800	4	8	9	450,000
2000	5	3	10	500,000
850	2	25	6	130,000
1350	3	12	6	300,000
1600	4	7	7	400,000
1100	3	15	5	220,000
1750	4	10	8	380,000

Income (in $)	Age	Will Purchase (0/1)
25,000	22	0
50,000	35	1
15,000	18	0
60,000	45	1
28,000	23	0
70,000	40	1
32,000	28	0
85,000	50	1
22,000	20	0
65,000	42	1

OR

Q8 a. Apply backpropagation algorithm to train the model using the given dataset of flower species. Interpret how backpropagation algorithm works. Explain the convergence behaviour of the backpropagation on the dataset, discussing how learning rate and network architecture affect the training process. — 10 | L3 | C4

Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
4.7	3.2	1.3	0.2	Setosa
4.6	3.1	1.5	0.2	Setosa
5.0	3.6	1.4	0.2	Setosa
7.0	3.2	4.7	1.4	Versicolor
6.4	3.2	4.5	1.5	Versicolor
6.9	3.1	4.9	1.5	Versicolor
5.5	2.3	4.0	1.3	Versicolor
6.5	2.8	4.6	1.5	Versicolor
6.3	3.3	6.0	2.5	Virginica
5.8	2.7	5.1	1.9	Virginica
7.1	3.0	5.9	2.1	Virginica
6.3	2.9	5.6	1.8	Virginica
6.5	3.0	↓	2.2	Virginica

b. Apply the perceptron algorithm to predict whether a new advertising campaign will be **successful** or **unsuccessful** based on two factors:
1. **Advertising Budget** (in thousands of dollars).
2. **Audience Reach** (in thousands of people).

The dataset is as follows: — 10 | L3 | C4

Advertising Budget (X1)	Audience Reach (X2)	Success (Y)
10	50	Unsuccessful (-1)
20	60	Unsuccessful (-1)
30	80	Successful (+1)
40	90	Successful (+1)
50	100	Successful (+1)

Module – 5

Q9	a.	Apply value iteration to determine the optimal dispatch policy to solve a Markov Decision Process (MDP) problem for a taxi dispatch service using the provided dataset of trips and rewards based on trip distances and times. Describe the value iteration algorithm and its application in this scenario.	10	L3	C5
	b.	In an online recommendation system, you have a dataset of user interactions with different products. Identify the optimal recommendation policy using policy iteration. Explain the policy iteration algorithm and its role in finding optimal policies in reinforcement learning. Apply policy iteration to the dataset and analyze the resulting optimal policy for making product recommendations.	10	L3	C5

Trip ID	Start Location	End Location	Trip Distance (miles)	Trip Time (minutes)	Reward (fare $)	Dispatch State	Action (Dispatch/No Dispatch)
1	Downtown	Airport	12.5	30	25	Available	Dispatch
2	Uptown	Suburbs	8.0	20	18	Busy	No Dispatch
3	Airport	Downtown	12.5	32	26	Available	Dispatch
4	Suburbs	Uptown	9.5	22	20	Busy	No Dispatch
5	Downtown	Suburbs	6.5	18	15	Available	Dispatch
6	Suburbs	Airport	15.0	40	30	Available	Dispatch
7	Uptown	Downtown	7.0	17	16	Busy	No Dispatch
8	Downtown	Uptown	5.5	15	12	Available	Dispatch

		<table><tr><th>User ID</th><th>Product ID</th><th>Interaction (e.g., Rating, Click)</th><th>Reward (e.g., Purchase, Like)</th></tr><tr><td>1</td><td>A</td><td>5</td><td>1</td></tr><tr><td>1</td><td>B</td><td>3</td><td>0</td></tr><tr><td>1</td><td>C</td><td>4</td><td>1</td></tr><tr><td>2</td><td>A</td><td>2</td><td>0</td></tr><tr><td>2</td><td>B</td><td>5</td><td>1</td></tr><tr><td>2</td><td>C</td><td>1</td><td>0</td></tr><tr><td>3</td><td>A</td><td>4</td><td>1</td></tr><tr><td>3</td><td>B</td><td>2</td><td>0</td></tr><tr><td>3</td><td>C</td><td>5</td><td>1</td></tr></table>			
		OR			
Q10	a.	You are given a dataset representing a simple game environment where an agent moves between different states and earns rewards based on its actions. Utilize temporal difference (TD) learning to help the agent learn the best strategy. Describe the TD learning algorithm and how it updates the value function during the learning process. Apply TD learning to the game environment and discuss how well the agent learns to maximize rewards over time. <table><tr><th>State</th><th>Action</th><th>Next State</th><th>Reward</th></tr><tr><td>S1</td><td>A1</td><td>S2</td><td>1</td></tr><tr><td>S1</td><td>A2</td><td>S3</td><td>0</td></tr><tr><td>S2</td><td>A1</td><td>S3</td><td>1</td></tr><tr><td>S2</td><td>A2</td><td>S1</td><td>-1</td></tr><tr><td>S3</td><td>A1</td><td>S1</td><td>0</td></tr><tr><td>S3</td><td>A2</td><td>S2</td><td>1</td></tr></table> **Description:** • **State**: The current state of the agent. • **Action**: The action taken by the agent. • **Next State**: The state the agent transitions to after taking the action. • **Reward**: The reward received by the agent for taking the action.	10	L3	C5
	b.	You are working as a data scientist at a retail company that wants to predict customer churn. You have a dataset containing information about customer transactions, demographics, and whether they churned (left the service) or not. The company wants to build a machine learning model to classify customers into two categories: churned (1) or not churned (0).Apply **K-fold cross-validation** to evaluate the model's performance using 5-fold cross validation for the given dataset.Discuss the model performance	10	L3	C5

Customer_ID	Age	Annual_Income	Total_Purchases	Churn
C001	34	45,000	12	0
C002	55	72,000	25	1
C003	23	36,000	8	0
C004	45	58,000	18	1
C005	36	42,000	10	0
C006	60	80,000	30	1
C007	28	39,000	9	0
C008	50	65,000	22	1
C009	32	43,000	11	0
C010	42	70,000	24	1

-*****-