BH/D/TY

Reg. No: [redacted]

VIT
Vellore Institute of Technology

Final Assessment Test – April 2025

Course: **MDI3003 - Advanced Predictive Analytics**

Class NBR(s): **2376/2378/2380**

Slot: **C2+TC2**

Time: **Three Hours**

Max. Marks: **100**

- ➢ **KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE**
- ➢ **DON'T WRITE ANYTHING ON THE QUESTION PAPER**

Answer ALL Questions

(10 X 10 = 100 Marks)

1. A healthcare provider wants to predict which patients are at risk of chronic diseases for early intervention. What challenges might they face in building an accurate predictive model, and how can they overcome them? Explain.

2. A supermarket chain wants to develop a machine learning model to predict daily sales for each store based on factors like **historical sales, promotions, holidays, weather conditions, and competitor pricing.** Explain how you would approach this problem from data collection to model evaluation.

3. Consider the following data about customer spending behaviour of a retail company.

Customer ID	101	102	103	104	105
Age (years)	25	34	45	23	31
Annual Income ($1000s)	40	60	80	30	50
Spending Score (1-100)	70	50	40	85	65
Online Shopping Hours per Week	5	7	10	3	6
Total Purchases per Year	15	18	20	12	17

a) Investigate whether there is a meaningful connection between a customer's **income level and spending habits.** Use appropriate visual and statistical methods to support your findings.

b) Considering multiple factors together, analyze how **income, spending score, and total purchases** interact. Identify any significant patterns.

4. A retail company wants to develop a machine learning model based on the following data to predict whether a customer will make a purchase (1) or not purchase (0) based on two key features such as **time spent on the website (in minutes)** and **number of product views.** Design a neural network using a 2-2-2 architecture. Given the following weight matrices and input values, compute the output of the network using the **Sigmoid activation function** in the hidden and output layers.

Input Layer: $X=[x_1,x_2]=[0.2,0.5]$

Weights for Hidden Layer (W_1): $\begin{bmatrix} 0.4 & 0.3 \\ 0.6 & 0.8 \end{bmatrix}$

Bias for Hidden Layer (B_1): $[0.1, 0.2]$

Weights for Output Layer (W_2): $\begin{bmatrix} 0.5 & 0.7 \\ 0.2 & 0.4 \end{bmatrix}$

Bias for Hidden Layer (B_2): $[0.3, 0.5]$

5. a. Define Weak Learners in the context of Boosting and explain their role in ensemble learning. [4]

b. Why are Decision Trees commonly used as Base Learners in Boosting? Explain with at least one real-world example where Boosting with Decision Trees is applied. [6]

6. Describe the process of selecting the best Auto Regressive Integrated Moving Average (ARIMA) model for time series forecasting. How do you determine the values of p, d, and q using Autocorrelation Functions (ACF) and Partial Autocorrelation Function (PACF) plots?

7. Consider the sample dataset given below and write a Python program to classify tweets as positive (0) or negative (1) using Naïve Bayes Classifier.

ID	Tweet	Sentiment
1	I love this product! It's amazing	1
2	Worst experience ever! I hate it	0
3	Very satisfied with the service	1
4	I am extremely disappointed	0
5	Great customer support! Fast and helpful	1
6	This is the worst thing I have ever bought	0
7	I am so happy with my purchase!	1
8	Terrible! I will never buy from here again!	0
9	Best decision ever! Totally worth it	1
10	Not good at all, waste of money!	0

Preprocess the dataset and predict the test tweet "**The product is fantastic! I love it**" as positive or negative.

8. How can Random Forest be used for both classification and regression tasks? Illustrate the RF algorithm with sample real-time examples.

9.a) A bank wants to develop a system to detect fraudulent transactions. How can a decision tree be used to classify transactions as fraudulent or legitimate? Discuss the key factors considered and the advantages of using decision trees for this problem.

OR

9.b) A hospital is using two machine learning models, Model A and Model B, to predict whether a patient has cancer (1) or not (0). Model A correctly identifies 40 cancer patients but misclassifies 20 healthy patients as having cancer. It also misses 10 actual cancer cases but correctly identifies 130 healthy patients. Model B correctly identifies 35 cancer patients and misclassifies 10 healthy patients as having cancer. It misses 15 cancer cases but correctly identifies 135 healthy patients. Calculate accuracy, precision, recall, and F1-score for both models. Which model is better for cancer detection? Explain your answer, considering the importance of correctly identifying cancer cases.

10.a) Compare PCA with feature selection techniques. How does PCA differ from selecting the most relevant features directly? In what scenarios should PCA be preferred?

OR

10.b) A store has recorded its monthly sales (in $1000s) for six months as follows: **X = [120, 150, 90, 200, 170, 130]**. Since the sales values vary significantly, it becomes challenging to compare them directly. To make the data easier to analyze, apply various transformation techniques such as scaling the values between 0 and 1, standardize the data to have a mean of 0, apply a log transformation to reduce large differences, and normalize the data using Box-Cox transformation.

⇔⇔⇔ **BH/D/TY** ⇔⇔⇔