



VIT

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

REG.NO.: 21M100285

SLOT: E1+TE1

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - I
WINTER SEMESTER 2025-2026

Programme Name & Branch : M.Tech. (Integrated) CSE (Data Science)
Course Code and Course Name : MDI4001 - Machine Learning for Data Science
Faculty Name(s) : Dr. AARTHY S.L, Dr. JEEVANANTHAM A.K, Dr. MADHURI W, Dr. SAIRABANU J
Class Number(s) : 1539, 1540, 1541, 1542
Date of Examination : 31-01-2026
Exam Duration : 90 minutes
Maximum Marks: 50

General instruction(s):

- Answer All Questions
- M - Max mark; CO - Course Outcome; BL - Blooms Taxonomy Level (1 - Remember, 2 - Understand, 3 - Apply, 4 - Analyse, 5 - Evaluate, 6 - Create)
- Course Outcomes
 1. Understanding the nuances of an ML sequence
 2. Derive an understanding of a Model's deficiency
 3. Gaining knowledge of mathematical concepts involved in Gradient Descent

Q. No	Question	Module	Marks	CO	BL
1.	A data science team is developing a machine learning model to predict house prices using a dataset that contains several numerical features (such as area, number of rooms, and age of the building) along with some noisy and weakly relevant attributes. During experimentation, the team observes that a low-degree polynomial underfits the data, while a high-degree polynomial performs extremely well on the training set but inconsistently on unseen data. As a machine learning engineer, explain what is hyperparameter and how hyperparameter tuning is to be carried out for selecting a generalized model in this scenario.	1	10	1	2
2.	A machine learning team is developing a model to classify emails as spam or not spam using a large dataset. They experiment with two different models and obtain the following observations: • Model A (Simple Linear Classifier): Training error = 22%, Validation error = 24% Estimated bias (error due to simplifying assumptions) = High Estimated variance (sensitivity to data fluctuations) = Low • Model B (Highly Complex Model): Training error = 1%, Validation error = 15% Estimated bias = Low Estimated variance = High Based on this scenario, analyze the two models and explain which model is underfitting, which model is overfitting, and why? In your answer, clearly relate training and validation errors to the concepts of bias and variance, explain how model complexity influences the bias-variance trade-off, and justify which type of model should be selected to achieve good generalization on unseen email data.	1	10	1	2



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - I
WINTER SEMESTER 2025-2026

3.	An automated academic analytics system is developed to predict whether a student will fail an examination (Positive) or pass (Negative) based on attendance, internal assessments, and assignment scores. Two classification models, Model A and Model B, are evaluated using the same test dataset of 200 students, and the following observations are obtained: • Model A Out of 80 actual failed students, the model correctly predicts 72 students as failed. Out of 120 actual passed students, the model incorrectly predicts 48 students as failed. • Model B Out of 80 actual failed students, the model correctly predicts 52 students as failed. Out of 120 actual passed students, the model incorrectly predicts 12 students as failed. Construct the confusion matrix for both models. Compute the following performance measures for each model: Accuracy, Precision, Recall (Sensitivity), F1-score. If the institution's primary objective is to identify as many at-risk students as possible for timely academic intervention, which model should be selected	2	10	3	3															
4.	StarEstate wants to predict the price of houses based on the number of bedrooms and age of the house. They have collected the following data for 4 houses: <table border="1"><thead><tr><th>Bedrooms</th><th>Age of House (in years)</th><th>Price (in Lakhs)</th></tr></thead><tbody><tr><td>2</td><td>10</td><td>25</td></tr><tr><td>3</td><td>5</td><td>35</td></tr><tr><td>4</td><td>8</td><td>40</td></tr><tr><td>3</td><td>12</td><td>30</td></tr></tbody></table> Using Multiple Linear Regression, predict the price of a house with 3 bedrooms and 7 years of age.	Bedrooms	Age of House (in years)	Price (in Lakhs)	2	10	25	3	5	35	4	8	40	3	12	30	2	10	3	3
Bedrooms	Age of House (in years)	Price (in Lakhs)																		
2	10	25																		
3	5	35																		
4	8	40																		
3	12	30																		
5.	Consider the following points in a 2D plane: • Positive class (+1): (3,1), (3,-1), (6,1), (6,-1) • Negative class (-1): (1,0), (0,1), (0,-1), (-1,0) Using the Support Vector Machine (SVM) algorithm, find the equation of the optimal separating hyperplane, identify the support vectors and schematically illustrate the hyperplane along with the margins and support vectors.	3	10	2	4															

Formula Sheet:

Multiple Linear Regression by Matrix Method: $\beta = (X^T X)^{-1} X^T Y$
