REG.NO.: [redacted]

VIT
Vellore Institute of Technology

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2024-2025

SLOT:G1+TG1

Programme Name & Branch	**:M.Tech CSE Integrated**
Course Code and Course Name	**:MDI3006 and Advanced Data Analytics**
Faculty Name(s)	**: Dr.Chellatamilan T, Dr.Ebenezer Juliet S**
Class Number(s)	**: VL2024250502876, VL2024250502880**
Date of Examination	**: 22-03-2025**
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

CO2 - Understand the advantages and limitations of the algorithms and their potential applications

CO3- Design experiments for evaluation and analyze the results to test the effectiveness of individual components of an algorithm

CO4- To explore the fundamental concepts of big data analytics

<table>
<tr><th>Q. No</th><th>Question</th><th>M</th><th>CO</th><th>BL</th></tr>
<tr><td>1.</td><td>A Navie Bayes model is employed to predict the weather of the day as either sunny or rainy using the following weather dataset. For this prediction, prepare frequency table, find the prior probability of both classes and likelihood probability of individual keyword in the given query. Then predict the weather of the day using MAP inference, if the query to the model is " It is hot but dark clouds are in sky ". Consider only the keywords such as hot, dark, clouds and sky. Use the following formula to find likelihood probability of individual key word and to avoid zero frequency problem.
$P(x_i \mid y) = (\text{count}(x_i,y) + \alpha) / (\text{count}(y) + \alpha^*N)$, where x_i -refer to individual keywords in the query, N-total number of unique words in the dataset and assume $\alpha=1$.
<table>
<tr><th>Sl.No.</th><th>Sentence(X)</th><th>Label(y)</th></tr>
<tr><td>1</td><td>The sun is bright today</td><td>Sunny</td></tr>
<tr><td>2</td><td>It is hot and sunny day</td><td>Sunny</td></tr>
<tr><td>3</td><td>It is raining heavily outside</td><td>Rainy</td></tr>
<tr><td>4</td><td>Dark clouds bring heavy rain</td><td>Rainy</td></tr>
<tr><td>5</td><td>Sunny weather makes me happy</td><td>Sunny</td></tr>
<tr><td>6</td><td>Rainy days are cold and wet</td><td>Rainy</td></tr>
</table></td><td>10</td><td>CO2</td><td>3</td></tr>
<tr><td>2.</td><td>Explore the significance of regularization in machine learning, and how does it contribute to model performance? Using a numerical example, calculate the objective function values for a typical linear model as a dictionary learning model by incorporating the following:
L1 norm regularization with $\lambda=0.1$
L2 norm regularization with $\lambda=0.02$</td><td>10</td><td>CO3</td><td>3</td></tr>
</table>

REG.NO.:

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

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
WINTER SEMESTER 2024-2025

SLOT:G1+TG1

3.	A supervised dictionary learning model is used for fingerprint-based authentication in a binary classification setting. During the training phase, the model constructs a comprehensive dictionary of basic image patches extracted from fingerprint images and learns the corresponding sparse coefficients from the training samples. Given the training dataset X, their corresponding class labels Y, and the learned dictionary D, classify the test data [1,2] using supervised dictionary learning and also explore the steps for classification. $X = \begin{bmatrix} 2 & 3 \\ 3 & 2 \\ -2 & -3 \\ -3 & -2 \end{bmatrix}$ $Y = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$ $D = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$	10	CO3	3
4.	a) Consider a Bloom filter with three hash functions and a bit array of length 15. The three hash functions h1, h2, and h3 produce the following values for the first three items "apple"," banana", "orange" in the stream. Insert them into the Bloom filter and show the status of Bloom filter after inserting three items. Items | h1 | h2 | h3 "apple" | 3 | 12 | 11 "banana" | 11 | 1 | 9 "orange" | 8 | 3 | 13 Suppose the fourth item in the stream is "mango" and produces the hash values of 3, 12, and 9 for the hash functions h1, h2, and h3 respectively. What will exist() function return if the other three items have already been inserted? State the reason of getting the answer.	5	CO4	3
	b) A Bloom Filter is a space-efficient probabilistic data structure used to test whether an element is definitely not in a stream or might be in a stream. Even though it performs quick lookups with minimal memory it allows false positives in query answering. Determine the false positive rate and explain how the false positives rate can be reduced with an example.	5		
5.	a) Consider a window size of 16. Apply DGIM algorithm for the bucket formation and show the division of stream into buckets for the following binary stream "1011 1010 1011 1001 ". Assume the bucket formation starts from right to left. Use DGIM algorithm to estimate the number of 1's for the last 12 positions of the stream. Also illustrate how the bucket will be modified for the following cases and perform merging and removal of buckets if necessary. i) when 1 enters ii)then 0 0 0 enters iii)then 1 enters	6	CO4	4
	b) The rainfall data for 6 weeks in city "ABC" is given in mm X={12,15,10,20,18,25} and decay rate is λ=0.3. Compute the smoothed rainfall rate over the past 6 weeks using Decaying Window with a normalized weighted sum.	4		

****