



VIT

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

REG.NO.:

SLOT: C2+TC2

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

Programme Name & Branch	:	B.Tech Computer Science and Engineering
Course Code & Name	:	BCSE206L – Foundations of Data Science
Faculty Name	:	Prof. Rajasekhara Babu M, Prof. Santhi K, Prof. Radhakrishnan Delhibabu, Prof. A.Devipriya
Class Number	:	VL2025260502530/ 2058/ 2532/ 2057
Date of Examination	:	18-Mar-2026 & 09:30 AM - 11:00 AM

Exam Duration: 90 minutes

Maximum Marks: 50

General instruction(s):

1. Answer All Questions
2. 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: Demonstrate proficiency in data analytics.

CO3: Apply advanced tools to work on dimensionality reduction and mathematical operations.

Q. No.	Question	Module	Marks	CO	BL
1.	Consider the following documents and the query. Construct the document-term matrix using TF-IDF weighting (stop words - are, in) and compute the cosine similarity to rank the documents in order of their relevance to the query. Show all steps clearly, including vector construction, dot product, magnitude, and cosine similarity. Document 1: "Birds fly in blue sky." Document 2: "Fish swim in clear water." Document 3: "Sky and water are blue." Query: "blue water"	4	10	CO2	4
2.	Consider a small training corpus for a sentiment analysis task: Doc 1 (Positive): "I love this curious cat", Doc 2 (Positive): "cat jumped outside", Doc 3 (Negative): "watch the bad cat", Doc 4 (Negative): "bad birds outside" Calculate the probability that the test document "curious birds" belongs to the Positive class and the Negative class using the Naïve Bayes algorithm with Laplace (add-1) smoothing. Show the prior probabilities, conditional word probabilities, and the final predicted class.	4	10	CO3	3

3.	You are given a 2-D dataset as follows; P1(2,0), P2(0,1), P3(3,1), P4(4,2), P5(5,3). Perform Principal Component Analysis (PCA) to reduce the data to 1D: (a) Standardize the dataset (show only the standardized points, not intermediate steps). (b) Compute the covariance matrix of the standardized data. (c) The eigenvalues and eigenvectors are given below. Use them to identify the principal component. → Eigenvalues: $\lambda_1 = 1.68$ $\lambda_2 = 0.32$ Eigenvectors: $v_1 = [0.89, 0.45]$, $v_2 = [-0.45, 0.89]$ (d) Project the standardized data points onto the principal component v_1.	5	10	CO3	3
4.	Let us suppose that you are provided with a movie dataset that has the following issues: Some columns contain missing values. A few entries contain outliers in the duration column. You need to check the relationship between variables such as duration and genre. As a Data Scientist, perform the following tasks using Python. Write the code snippets and briefly explain the reasoning behind each step: (1) Load the dataset and display its basic statistics using an appropriate function. (2) Demonstrate two different methods to handle missing values. (3) Detect outliers in the duration column using z-score method and show how you would handle them. (4) Visualize the relationship between duration and genre using any Python visualization tool.	5	10	CO3	4
5.	Consider the following 5 data points in 2D space: P (2,6), Q (3,4), R (4,7), S (6,2), T (7,4). (1) Compute the Euclidean distance matrix between all points. (2) Apply Agglomerative Hierarchical Clustering using the Single Linkage method. Show the sequence of cluster merges step by step until all points form a single cluster. (3) Draw a rough sketch of the dendrogram showing the merging order and distances.	5	10	CO3	3

<<<END>>>