



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(s). Kalyanaraman P, Nalini, Badrinath, Rajasekara Babu
Class Number	VL2025260502056, 2530, 2533, 2535
Date of Examination	17-Mar-2026 & 02:00 PM - 03:30 PM
Exam Duration : 90 Minutes	Maximum Marks: 50

General instruction(s):

- Answer All Questions
- Blooms Taxonomy Level[BL] (1- Remember, 2- Understand, 3- Apply, 4- Analyse, 5- Evaluate, 6- Create)
- Course Outcomes[CO]:
CO3. Apply advanced tools to work on dimensionality reduction and mathematical operations.
CO4. Handle various types of data and visualize them using through programming for knowledge representation.

Q. No.	Question	Module	Marks	CO	BL
1.	Given a query and a set of documents, apply the Vector Space Model using Term Frequency - Inverted Document Frequency method to order the documents based on cosine similarity. • Document 1: Data Mining Machine Learning data analysis • Document 2: Deep Learning Machine Learning AI industry • Document 3: AI Machine Learning techniques Big Data analysis • Query: Machine Learning industry analysis	4	10	4	4
2.	Classification: Naive Bayes Assume a small training corpus with two classes, Spam (S) and Ham (H): Doc1 (S): win free money Doc2 (H): meeting schedule confirmed Doc3 (S): free money confirmed 1. Construct the vocabulary frequency table and calculate the prior probabilities for classes S and H. (4 Marks) 2. Using the Naïve Bayes algorithm (with Laplace smoothing applied), calculate the probability that a new document containing the phrase free schedule belongs to the Spam class. (6 Marks)	4	10	4	3

3	Clustering: K-Means Suppose you have 4 data points: A(1,1), B(2,1), C(4,3), D(5,4) and you initialize $K = 2$ clusters with starting centroids at A and D. (a) Perform exactly one iteration of the K-means algorithm using Manhattan distance. Show the mathematical steps for the new cluster assignments and calculate the coordinates of the new centroids. (6 Marks) (b) Identify and explain one major limitation of the K-means algorithm when handling non-spherical data distributions or datasets with heavy noise. (4 Marks)	5	10	3	4																																																						
4.	You are provided with a Time Series Dataset tracking daily ecommerce sales. The data contains severe outliers and missing values. <table border="1"> <thead> <tr> <th>Date</th><th>Orders</th><th>Revenue (\$)</th><th>Visitors</th><th>Conversion Rate(%)</th><th>AdSpend (\$)</th></tr> </thead> <tbody> <tr><td>2024-01-01</td><td>120</td><td>2400</td><td>1500</td><td>8.0</td><td>300</td></tr> <tr><td>2024-01-02</td><td>135</td><td>2600</td><td>1600</td><td>8.4</td><td>320</td></tr> <tr><td>2024-01-03</td><td>NaN</td><td>2550</td><td>1580</td><td>8.1</td><td>310</td></tr> <tr><td>2024-01-04</td><td>140</td><td>NaN</td><td>1650</td><td>8.5</td><td>330</td></tr> <tr><td>2024-01-05</td><td>950</td><td>22000</td><td>9000</td><td>10.5</td><td>1200</td></tr> <tr><td>2024-02-02</td><td>130</td><td>2500</td><td>NaN</td><td>8.3</td><td>315</td></tr> <tr><td>2024-02-03</td><td>118</td><td>2350</td><td>1480</td><td>NaN</td><td>295</td></tr> <tr><td>2024-02-04</td><td>115</td><td>2300</td><td>1450</td><td>7.9</td><td>NaN</td></tr> </tbody> </table> (a) Detail the specific pandas and numpy functions or strategies you would utilize to programmatically identify and impute the missing data. (5 Marks) (b) Propose a step-by-step method using Python libraries to aggregate this daily data into monthly trends to perform Exploratory Data Analysis (EDA). (5 Marks)	Date	Orders	Revenue (\$)	Visitors	Conversion Rate(%)	AdSpend (\$)	2024-01-01	120	2400	1500	8.0	300	2024-01-02	135	2600	1600	8.4	320	2024-01-03	NaN	2550	1580	8.1	310	2024-01-04	140	NaN	1650	8.5	330	2024-01-05	950	22000	9000	10.5	1200	2024-02-02	130	2500	NaN	8.3	315	2024-02-03	118	2350	1480	NaN	295	2024-02-04	115	2300	1450	7.9	NaN	5	10	4	4
Date	Orders	Revenue (\$)	Visitors	Conversion Rate(%)	AdSpend (\$)																																																						
2024-01-01	120	2400	1500	8.0	300																																																						
2024-01-02	135	2600	1600	8.4	320																																																						
2024-01-03	NaN	2550	1580	8.1	310																																																						
2024-01-04	140	NaN	1650	8.5	330																																																						
2024-01-05	950	22000	9000	10.5	1200																																																						
2024-02-02	130	2500	NaN	8.3	315																																																						
2024-02-03	118	2350	1480	NaN	295																																																						
2024-02-04	115	2300	1450	7.9	NaN																																																						
5.	Dimensionality Reduction: PCA (a) Explain the fundamental mathematical objective of Principal Component Analysis (PCA) when it is applied to a high-dimensional dataset. (4 Marks) (b) Consider the following points $X_1 = (1, 2, 3, 4)$ and $X_2 = (2, 4, 6, 8)$. The eigenvectors of the covariance matrix are given as $e_1^T = [0.447 \ 0.894]$ and $e_2^T = [-0.894 \ 0.447]$ Compute the Principal Component (PC1) and PC2 using the Principal Component Algorithm. Plot the calculated PC1 axis and PC2 axis along with the original values and the reduced dimensional values in the new axis.	5	10	3	4																																																						