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

REG.NO.:

SLOT: F1+TF1

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

Programme Name & Branch : **B.Tech – COMPUTER SCIENCE AND ENGINEERING**
Course Code and Course Name : **BCSE334L - Predictive Analytics**
Faculty Name(s) : ANURADHA J, VIJAYASHERLY V, SANTHI K, THIRUNAVUKKARASAN M, UMA PRIYA D, SAYAN SIKDER.d
Class Number(s) : VL2025260101658, VL2025260101648, VL2025260101661, VL2025260101653, VL2025260101650, VL2025260101656
Date of Examination : **22.08.2025**
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:**

CO1- Understand the importance of predictive analytics and processing of data for analysis. **CO2** - Describe different types of predictive models.

Q. No	Question	M	CO	BL
1.	1. A hospital is trying to decide whether to invest in a new AI-based patient monitoring system. They have data on patient outcomes, current monitoring efficiency, and projected costs. a. Which model type (predictive, descriptive, or decision) should be used, and justify your choice. b. Explain how an analytical technique (such as regression analysis, classification, or optimization) could be applied to reach a decision.	10	1	2

1. Hospital Decision on AI-Based Patient Monitoring System

a. Model Type: Decision Model

Recommended Model Type: Decision Model

Justification:

The hospital is trying to make a choice—whether to invest in a new AI-based patient monitoring system. This scenario goes beyond just understanding data patterns (descriptive) or forecasting outcomes (predictive). It involves evaluating alternatives, trade-offs, and outcomes, considering multiple factors such as cost, efficiency, and patient outcomes.

Justification points with example- 2 marks

b. Analytical Technique: Optimization

Recommended Technique: Optimization

Optimization can be used to determine the best possible decision regarding the investment by analyzing constraints and objectives.

Justification points with example- 2 marks

2. Given the dataset below, which contains the test scores of five students in various subjects, use the K-Nearest Neighbours method to impute the missing values. Perform the calculations and use k=3 for the imputation. How would you calculate the missing value for "Student 1 - English" and Student 2 – Science marks using KNN imputation?

Student	Math	Science	English	History	Art
1	85	78	---	92	88
2	72	---	85	80	75
3	90	88	92	----	85
4	65	70	75	78	----
5	----	85	80	89	82

$$d_{15} = \sqrt{\frac{5}{4} \times [(85-78)^2 + (92-89)^2 + (88-82)^2]}$$

$$= \sqrt{\frac{5}{4}(49 + 9 + 36)}$$

$$= 10.85$$

Since $k=3$ we take the least distant 3 points

Least = {Student 2, Student 3, Student 5}

$$= \{92, 80, 85\}$$

$$= 85.66$$

=> Imputed value for Student 1 - English = 85.66

$$d_{21} = \sqrt{\frac{5}{3} \times [(85-72)^2 + (92-80)^2 + (88-75)^2]}$$

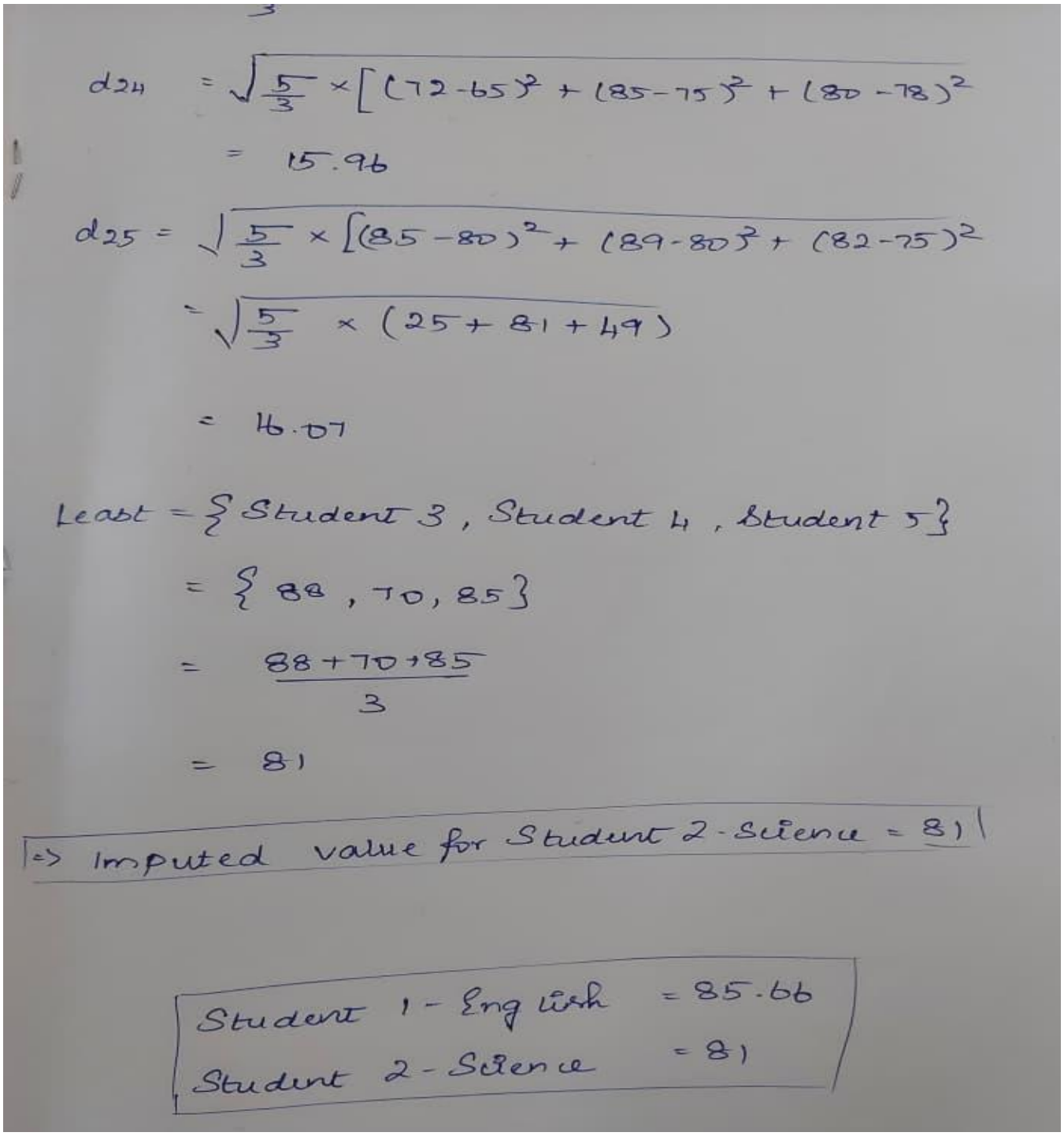
$$= \sqrt{\frac{5}{3} \times (49 + 144 + 169)}$$

$$= 28.34$$

$$d_{23} = \sqrt{\frac{5}{3} \times [(90-72)^2 + (92-85)^2 + (85-75)^2}$$

$$= \sqrt{\frac{5}{3} \times (324 + 49 + 100)}$$

$$= 28.07$$



ANSWER

1. Student 1 – English: 85.67
2. Student 2 – Science: 81.0

3. Given the 2D dataset: P1=(8,6), P2=(−3,4) , P3=(0,−5), P4(1,1) & P5(-7,-24). Apply the spatial sign transformation to each point, showing all intermediate steps. Then, plot both the original points and their transformed versions on the same 2D graph to illustrate the effect.

Spatial Sign Transformation

The spatial sign transformation maps each point (x, y) to a unit vector in the same direction:

S(x, y) = (x / sqrt(x^2 + y^2), y / sqrt(x^2 + y^2)).

Step 1: Calculations

P1 = (8, 6)

Length = sqrt(8^2 + 6^2) = sqrt(100) = 10.
Divide coordinates by 10 → Spatial sign = (0.8, 0.6).

P2 = (−3, 4)

Length = sqrt((-3)^2 + 4^2) = sqrt(25) = 5.
Divide coordinates by 5 → Spatial sign = (−0.6, 0.8).

P3 = (0, −5)

Length = sqrt(0^2 + (−5)^2) = sqrt(25) = 5.
Divide coordinates by 5 → Spatial sign = (0, −1).

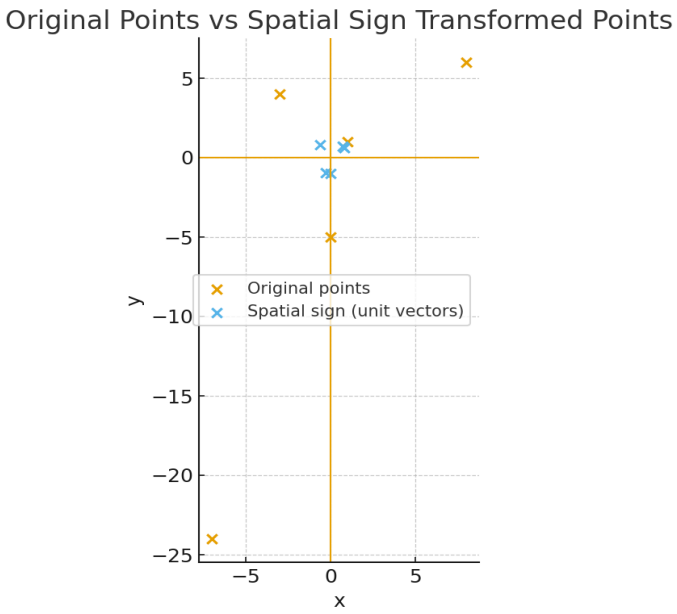
P4 = (1, 1)

Length = sqrt(1^2 + 1^2) = sqrt(2).
Divide coordinates by sqrt(2) → Spatial sign ≈ (0.707, 0.707).

P5 = (−7, −24)

Length = sqrt((−7)^2 + (−24)^2) = sqrt(625) = 25.
Divide coordinates by 25 → Spatial sign = (−0.28, −0.96).

Step 2: Interpretation
The original points may lie far from the origin, but after the spatial sign transformation, they are projected onto the unit circle. This keeps only the direction of each point from the origin, while discarding its distance. For instance, P3 (0, −5) and P5 (−7, −24) both point downward, so their transformed versions lie close together on the unit circle.



4. Consider the given Healthcare Dataset, the ratings are on a scale from 0-5

Patient	Blood Test	MRI Scan	Physiotherapy	Surgery	Diet Consultation
P1	5	3	0	0	4
P2	4	0	0	2	3
P3	0	4	5	3	0
P4	0	0	4	4	0
P5	3	4	0	0	5

Scale: 0 = service not taken, 5 = highest usage.
Using the dataset above, perform the following similarities.

1. Calculate the similarity between Patient(1,2) and Patient(2,3) using Cosine Similarity and Jaccard Similarity.
2. Calculate the similarity between MRI and Physiotherapy using Pearson Correlation Coefficient.

Healthcare Dataset Similarity Analysis
The dataset contains patient ratings (0–5) for healthcare services. Here we calculate patient-to-patient similarities using Cosine and Jaccard similarity, and service-to-service similarity using the Pearson Correlation Coefficient.
1. Patient-to-Patient Similarities
a) Between Patient 1 and Patient 2
Cosine Similarity ≈ 0.84 (high similarity in usage pattern).
Jaccard Similarity = 0.50 (half of their chosen services overlap).
b) Between Patient 2 and Patient 3
Cosine Similarity ≈ 0.16 (low similarity, very different usage patterns).
Jaccard Similarity = 0.20 (only one common service).
2. Attribute-to-Attribute Similarity
Between MRI Scan and Physiotherapy:
Pearson Correlation Coefficient ≈ 0.0098.
This is close to 0, indicating almost no linear relationship

5. A retail company wants to segment its customers based on Annual Income (in thousands ₹) and Spending Score (1–100). The marketing team has provided the following data for six customers:

Customer	**Annual_Income (₹000s)**	**Spending_Score**
A	15	39
B	80	81
C	25	20
D	70	78
E	18	45
F	65	80
G	80	20

The team has decided to use k-means clustering with k = 2 and the following randomly selected initial centroids: C1 = (15, 39) & C2 = (80, 81)

a. Using the Euclidean distance formula, what is the distance of each customer from C1 and C2?
b. Based on the cluster memberships, what are the coordinates of the new centroids after one iteration?
c. Identify the errors using K-means algorithm.

K-Means Clustering (k = 2): Customer Segmentation
Features: Annual Income (₹000s), Spending Score (1–100).
Initial centroids: C1 = (15, 39), C2 = (80, 81).
a) Euclidean Distances from Initial Centroids

Customer	Annual_Income	Spending_Score	Distance to C1	Distance to C2	Assigned Cluster
A	15	39	0.000	77.389	C1
B	80	81	77.389	0.000	C2
C	25	20	21.471	82.134	C1
D	70	78	67.424	10.440	C2
E	18	45	6.708	71.694	C1
F	65	80	64.661	15.033	C2
G	80	20	67.720	61.000	C2

b) New Centroids After One Iteration
C1 members: A, C, E
C2 members: B, D, F, G
New C1 = (19.33, 34.67)
New C2 = (73.75, 64.75)
c) Common Errors / Pitfalls in K-Means Application
Sensitivity to initial centroids: different random starts can yield different clusters.
Scale dependence: features must be normalized if measured on different scales.
Hard assignments: points near a decision boundary are forced into one cluster.
Non-spherical clusters: K-means assumes roughly spherical, equal-variance clusters.
Outlier sensitivity: outliers can drag centroids away from dense regions.
Imbalanced cluster sizes: K-means does not enforce balanced membership.