SIES Graduate School of Technology (Autonomous) NAAC A+ (Affiliated to University of Mumbai)	***End Semester Examination (R-24) F 2026*** ***Answer Key with marking scheme***
Branch: CE	Course: Statistical Foundation for Data Science
Year/ Semester: 2/4	Course code: MDMC4021
Time: 03 hours	Marks: 80

		M
Q. 1	Attempt any FOUR. (All questions carry equal marks)	
A.	Find the **four fundamental subspaces** of the matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 1 & 0 \end{pmatrix}$ **@ 1. Column Space span** $\left\{\begin{bmatrix}1\\2\\1\end{bmatrix}, \begin{bmatrix}2\\4\\1\end{bmatrix}\right\}$ **2. Row Space: span**$\{[1 \quad 0 \quad -1], [0 \quad 1 \quad 1]\}$ **3. Null Space span** $\left\{\begin{bmatrix}1\\-1\\1\end{bmatrix}\right\}$ 4. **Left Null Space span** $\left\{\begin{bmatrix}-2\\1\\0\end{bmatrix}\right\}$	
B.	A market research study was conducted to examine whether there is an association between customer service rating and salary level of customers. A sample of 100 customers was categorized into three service rating levels (Excellent, Good, Poor) and three salary groups (Low, Medium, High). The observed frequencies are shown in the contingency table below: Service Low Medium High Total Excellent 9 10 7 26 Good 11 9 31 51 Poor 12 8 3 23 Total 32 27 41 100 **@ χ2≈18.65, χcritical2=9.488, 18.65>9.488 So, Reject H_0**	
C.	Differentiate between Quantitative data vs. Qualitative data **@ Any 5 Correct Comparison Point 5 Marks**	
D.	Why is Exploratory Data Analysis important? Explain with suitable points. **@ Excpected Points: Definition of EDA, Understanding dataset / structure, Handling missing values or outlier, Understanding distribution / relationships, Role in model selection / improving accuracy each carry 1 marks.**	
E.	Using Newton's Method, find the root of $f(x) = e^x - 2x^2$ between 0 & 100, using Newton-Raphson method to an accuracy of 5 digits. **@Root: 1.4880**	
F.	Describe the concept of Data Reduction. How is Factor Analysis used for reducing data? Give an example.	

@Expected Points: Definition of Data Reduction, Purpose/importance (simplify, reduce size), Definition of Factor Analysis, Explanation of how it reduces variables, Suitable example.

Q.2 Attempt any FOUR. (All questions carry equal marks)

A. The following table shows a 2D dataset: Perform Principal Component Analysis (PCA) to reduce the dataset from 2D to 1D

X	Y
2	11
3	14
7	26

@

$$Z \cdot u_1 = \begin{bmatrix} -2 & -1 & 3 \\ -6 & -3 & 9 \end{bmatrix} \cdot \frac{1}{\sqrt{10}} \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Final Result:

$$\boxed{\frac{1}{\sqrt{10}}[-20,\ -10,\ 30]}$$

B. A random variable X has the following probability function:

X	0	1	2	3	4	5	6	7
$P(X)$	0	K	$2K$	$2K$	$3K$	K^2	$2K^2$	$7K^2 + K$

Find:

(i) K

(ii) Evaluate $P(X < 6)$, $P(X \geq 6)$ and $P(0 < X < 5)$

(iii) Determine the distribution function of X

(iv) $P(1.5 < X < 4.5 \mid X > 2)$

(v) Mean, Variance and Standard deviation

(vi) The smallest value of n for which $P(X \leq n) > \frac{1}{2}$

@

1. K=1/10

2. $P(X < 6) = 0.81$, $P(X \geq 6) = 0.19$ and $P(0 < X < 5) = 0.8$

4. $P(1.5 < X < 4.5 \mid X > 2) = 5/7$

5. Mean = 3.66, Variance = 3.38 and Standard deviation = 1.84

6. n=4

3.

x	F(x) = P(X ≤ x)
x < 1	0
1	0.1
2	0.3
3	0.5
4	0.8
5	0.81
6	0.83
7	1

C.	Explain Bracketing Methods for solving nonlinear equations. Using the **Bisection Method** and **False Position Method**, find the root of the equation $f(x) = x^3 - 4x - 9$in the interval [2, 3] up to three iterations. Show all calculations. **@Root lies approximately in [2.6,2.7]**
D.	Describe the steps involved in preparing a dataset for machine learning, covering data cleaning, missing value treatment, feature engineering, visualization, analysis, and model selection. **@Expected Points: Definition / Introduction of data preparation, Data Cleaning (any 2 valid points explained), Handling Missing Values (methods like deletion/imputation), Feature Engineering (creation, transformation, encoding), Data Visualization (purpose + examples of plots), Data Analysis / EDA (relationships, correlation, insights), Model Selection (choosing appropriate ML algorithm).**
E.	Compute the Linear Discriminant projection for the following two-dimensional dataset. Samples for class 1: X1= (x1, x2) = {(4,2), (2,4), (2,3), (3,6), (4,4)} Samples for class 2: X2= (x1, x2) = {(9,10), (6,8), (9,5), (8,7), (10,8)} @
F.	Explain how a dataset is understood using various plots and graphs, and how insights from analysis help in drawing conclusions and selecting an appropriate machine learning model. **@Expected Points: Introduction: Role of visualization/EDA in understanding dataset, Explanation of Histogram (distribution of data), Explanation of Scatter Plot (relationship between variables), Explanation of Box Plot (outliers, spread), Explanation of Bar Chart / Pie Chart (categorical comparison), Insights from analysis (patterns, trends, correlations), Detection of issues (outliers, missing values, imbalance), Drawing conclusions (feature importance, data behavior), Model selection based on insights (linear vs non-linear, classification/regression), Overall clarity / conclusion linking EDA to ML performance.**
Q.3	Attempt any FOUR. (All questions carry equal marks)
A.	Compute the Singular Value Decomposition (SVD) of the matrix $A = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix}$ @ $A = P\Sigma_A Q^T = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \frac{1}{\sqrt{6}} \begin{bmatrix} 2 & -1 & 1 \\ 0 & \sqrt{3} & \sqrt{3} \\ -\sqrt{2} & -\sqrt{2} & \sqrt{2} \end{bmatrix}$
B.	On a final examination in Mathematics, the **mean score** was **72** and the **standard deviation** was **15**. Determine the **standard score (z-score)** of a student receiving: **(a)** 60 marks **(b)** 93 marks **(c)** 72 marks @ • **Z = Below average** • $z = 1.4$**: Above average** • $z = 0$**: Exactly average**

X1	4	2	2	3	4	9	6	9	8	10
X2	2	4	3	6	4	10	8	5	7	8
1st LD	4.46	3.48	3.06	5.2	5.3	12.35	8.8	10.2	10.19	12.42

C.	Define underdetermined and overdetermined systems of linear equations. For the given system, identify its type and justify: $\begin{cases} x + y + z = 2 \\ 2x + 3y + z = 5 \end{cases}$ **@n>m⇒3>2; System is Underdetermined**	
D.	Describe any five data visualization techniques from the following with diagrams and use-cases: Bar graph, Pie chart, Histogram, Scatter plot, Time-series graph, Dot plot. **@** **Bar Graph → 2 marks** **Pie Chart → 2 marks** **Histogram → 2 marks** **Scatter Plot → 2 marks** **Time Series Graph → 2 marks**	
E.	Classify data into quantitative and qualitative types, and explain each subtype with examples. **@ Expected Points: Quantitative data definition, Discrete + Continuous explanation with examples, Qualitative data definition, Nominal + Ordinal + Binary explanation with examples.**	
F.	Apply Penalty Function Method to minimize $f(x) = x^2 + 2x$ subject to x >= 1.Formulate the penalty function and solve. **@Optimal solution: x=1 & Minimum value: f(x)=3**	