M.C.S.E. FIRST YEAR SECOND SEMESTER EXAMINATION-2025
(Artificial Intelligence and Data Science)
DATA ANALYSIS AND VISUALIZATION

Time: 3 hours **Full Marks: 100**

1.A. (i)Define feature engineering and explain its role in preprocessing.
(ii)Explain how proper feature engineering enhances model performance.
(iii)Discuss the role of data normalization and transformation in data preprocessing.
(iv) Explain one data transformation and one data normalization technique.

4 + 4 + 6 + 6

OR

1.B. (i)Design a case study where raw data from a real-world domain (e.g., healthcare, e-commerce, finance) is preprocessed for machine learning.
(ii) Detail the challenges in the raw dataset and describe the specific preprocessing techniques applied to resolve them.
(iii) How does incorrect or inconsistent data categorization affect inference and decision-making in data analysis? Give examples where poor data categorization led to misleading results or bias in data-driven applications.

8 + 5 + 7

2.A. (i)Discuss the steps involved in statistical hypothesis testing and their application in solving real-life problems.
(ii) Use a practical example (e.g., marketing campaign effectiveness, new drug trials, student performance) to illustrate the process from formulating hypotheses to interpreting p-values and making inferences.

10 + 10

OR

2.B. (i) A factory claims that the average lifetime of its LED bulbs is 50,000 hours. A consumer agency tests a random sample of 36 bulbs and finds a mean lifetime of 48,500 hours with a standard deviation of 3,000 hours. At a 5% significance level, test whether the bulbs last less than claimed. State the null and alternative hypotheses, perform the test, and interpret the results.
(ii) Two different diets are tested for weight loss. 10 individuals follow Diet A with a mean weight loss of 4.5 kg (SD = 1.2 kg), while 12 individuals follow Diet B with a mean weight loss of 5.3 kg (SD = 1.5 kg). At a 0.01 significance level, test if Diet B is significantly more effective than Diet A.

10 + 10

3.A. (i)You are given a dataset: X=[(2,0),(0,2),(3,1),(4,3)]. Project the data onto the first principal component.
(ii) Given: Mean of Class 1: $\mu 1$=[2, 3], Mean of Class 2: $\mu 2$=[5, 7] and within-class scatter matrix $=\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. Find the LDA projection vector that maximizes the separation between the classes.

10 + 10

OR

3.B (i)Both partitional (e.g., K-means) and agglomerative clustering methods make assumptions about the structure of data patterns and cluster shape. Critically evaluate these assumptions.

(ii)Agglomerative clustering is often more interpretable but less scalable than partitional clustering. Critically analyze the trade-offs between interpretability and computational efficiency when choosing between these methods.

5 +5

3.B(iii) Compare DB-index for clusterings {(1, 2), (3, 5, 6, 8)} and {(1, 2), (3, 5), (6, 8)} 10

4.A (i) Explain the invertibility condition of MA(1) and its importance in model estimation. Determine whether the following MA(1) process is invertible: $X_t=\mathbf{a}_t+\beta_1\mathbf{a}_{t-1}$ for $\beta_1=1.5$.
(ii) Ten successive observations on a stationary time series: 1.6, 1.8, 0.2, 0.5, 0.8, 1.5, 0.9, 1.3, 2.5, 1.3. Compute autocorrelation coefficients r_1.
(iii)For the model $(1-2.3B)(1-B)X_t = (1- 0.77B)\mathbf{a}_t$, find p, d, q and express it as ARIMA (p, d, q). Determine whether the process is stationary and invertible.

6 + 8 + 6

OR

4.B(i) You have a fitted MA(1) model: $X_t=75+\mathbf{a}_t+0.5\mathbf{a}_{t-1}$. You are given $\mathbf{a}_8=0.6$, forecast X9.
(ii) A time series: 105, 108, 111, 115, 119 shows a clear upward trend. Fit an ARIMA(1,1,0) model with $\boldsymbol{\alpha}_1=0.7$ and forecast the next value.
(iii)A sudden change in the behavior of a time series (e.g., due to a pandemic, policy shift, or economic crash) leads to structural breaks. Discuss how this affects AR, MA, and ARIMA forecasting accuracy.

7 + 7 + 6

5.A. (i) Explain how a scatter plot and correlation matrix are used to explore relationships between variables. Illustrate with an example.
(ii) Discuss the use of treemaps for visualizing hierarchical categorical variables. What kinds of patterns can be observed from such plots?
(iii) Discuss the role of aesthetics in data visualization. How do elements like color, shape, size, and layout contribute to the effectiveness of a visual presentation? Provide examples to support your answer.
(iv) Describe how multiple distributions can be compared using overlapping density plots. What are the advantages and potential issues of this approach?
(v) Explain how color, size, and area are used to represent proportions in data visualization. Give examples of effective and misleading usage.

4 × 5

OR

5.B. (i) Describe the different types of color blindness (e.g., protanopia, deuteranopia, tritanopia). How do these conditions influence how users perceive color in data visualizations?
(ii) Explain the differences between 2D and 3D data visualizations. What are the advantages and disadvantages of each in representing multidimensional data?
(iii) Discuss the importance of labeling in data visualizations. How do accurate, clear, and concise labels enhance understanding and prevent misinterpretation?
(iv) What is the principle of repetition in typography for data visualizations? How can consistent font usage and styling enhance user experience and visual hierarchy?
(v) Evaluate how whitespace around text (padding and margins) contributes to the overall readability of data visualizations. How does it support visual balance and reduce clutter?

4 × 5