Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

B.Tech Degree S3 (S) Examinations May 2026 (2024 Scheme)

Course Code: PBADT304

Course Name: INTRODUCTION TO DATA SCIENCE

Max. Marks: 40 Duration: 2 hours 30 minutes

PART A

(Answer all questions. Each question carries 2 marks)

		CO	Marks
1	Explain the role of accurate data on patient history and treatment outcomes in identifying the causes of patient readmissions and improving healthcare strategies.	CO1	(2)
2	Compare Machine Learning (ML) and Deep Learning (DL) in terms of their purpose, methods, and complexity. Provide examples for each.	CO1	(2)
3	List any four descriptive statistics used to summarize data.	CO2	(2)
4	A dataset shows the scores of 6 students in 3 different tests. Construct a matrix to represent the dataset.	CO2	(2)
5	Given the sorted dataset: {10, 15, 18, 22, 25, 30, 34, 38, 40, 45, 50, 55} Divide the data into three equal-frequency bins and perform smoothing using bin means and bin boundaries.	CO3	(2)
6	Compare two classifiers using their ROC curves. How do you decide which model is better?	CO3	(2)
7	A data scientist generates new features such as price per square foot and distance to the city center, while eliminating irrelevant features like house color to enhance model performance. Identify the process being applied.	CO4	(2)
8	Explain the basic idea of filter and wrapper methods in feature selection.	CO4	(2)

PART B

(Answer any one full question from each module, each question carries 6 marks)

Module -1

9	A logistics company faces high operational costs due to duplicate data storage and inefficient data usage. Explain the data science process that can help reduce redundancy and improve efficiency.	CO1	(6)
10	A banking system collects the following data: • Transaction records stored in tables • Customer emails and complaint messages • JSON files containing account activity logs Classify each type of data as structured, unstructured, or semi-structured. Also explain why identifying data types is important before analysis.	CO1	(6)

Module -2

11	Given the matrix $A=\begin{matrix}2 & 0\\ 0 & 1\end{matrix}$ Perform the singular value decomposition of the matrix.	CO2	(6)
12	A retail company collects large amounts of customer data, including age, income, spending score, purchase frequency, website activity, and product preferences. The dataset has many features, making it complex and difficult to visualize and analyze. As a data analyst, you decide to apply Principal Component Analysis (PCA) to simplify the dataset. • Outline the key steps involved in PCA • Explain how PCA helps in reducing dimensionality • Describe how important information is preserved while eliminating less significant features	CO2	(6)

Module -3

13	A hospital tests a model to predict disease with results TP=120, TN=300,FP=40, FN=40 apply suitable classification evaluation metrics to	CO3	(6)

assess the model's performance and explain their significance in medical diagnosis.

14	A bank develops a model to detect fraudulent transactions. The model is tested on 200 transactions. • The model predicted 80 transactions as fraud, out of which 60 were actually fraud. • The model predicted 120 transactions as non-fraud, but 20 of them were actually fraud. • Construct the confusion matrix. • Calculate:Accuracy,Precision,Recall,F1-score	CO3	(6)

Module -4

15	A hospital aims to predict whether a patient is at high risk of heart disease based on age, blood pressure, cholesterol levels, and lifestyle habits. Suggest an appropriate machine learning technique and explain why it is suitable.	CO4	(6)
16	Explain the construction and functioning of a Decision Tree algorithm and describe the role of splitting criteria like Entropy and Gini Index in selecting the best splits.	CO4	(6)

***