Reg No.:_______________ Name:__________________________

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

MCA (Two Years) 3rd Semester (S,FE) Examination June 2025

Course Code: 20MCA201

Course Name: DATA SCIENCE AND MACHINE LEARNING

Max. Marks: 60 Duration: 3 Hours

PART A

Answer all questions, each carries3 marks. Marks

1 What is data visualization? Why is data visualization important? (3)

2 Compare and contrast classification and regression in data science. (3)

3 Explain disadvantages of K-NN classifier. (3)

4 What is meant by “learning” in the context of machine learning? (3)

5 Explain the advantages and disadvantages of decision trees. (3)

6 Define information gain. What is its use in Decision Tree? (3)

7 Explain the different types of layers in an ANN ? (3)

8 Discuss the terms hyperplanes and support vectors. (3)

9 Explain random forest. (3)

10 How performance of a machine learning model is evaluated using ROC curves? (3)

PART B

Answer any one question from each module. Each question carries 6 marks.

Module I

11 Illustrate the data science process with the help of diagram. (6)

OR

12 Explain the visualization techniques for analysing univariate and multivariate data. (6)

Module II

13 Consider the dataset given below. Using k-NN algorithm, predict the class label for the new instance with brightness=20 and saturation =35. Choose k=1 and k=3. (6)

Brightness	Saturation	Class
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red
25	80	Blue

OR

14 Given a training dataset. Predict the class of a new patient with thesymptoms Fever: Yes, Cough: No, Body Ache: Yes, Fatigue: No, using Naive Bayes classifier. (6)

Patient	Fever	Cough	Body Ache	Fatigue	Disease
1	Yes	Yes	No	Yes	Disease
2	No	Yes	Yes	No	No Disease
3	Yes	No	Yes	Yes	Disease
4	Yes	Yes	Yes	No	No Disease
5	No	Yes	No	No	No Disease
6	Yes	Yes	Yes	Yes	Disease
7	Yes	No	No	Yes	Disease
8	No	No	Yes	No	No Disease
9	Yes	Yes	No	Yes	Disease
10	No	No	No	No	No Disease

Module III

15 Consider the following dataset: (6)

ID	Age	Blood Pressure	Health Status

1	25	120	Good
2	30	130	Fair
3	35	110	Good
4	20	140	Poor
5	40	125	Fair
6	28	115	Good
7	32	135	Poor
8	38	120	Fair
9	26	110	Good
10	22	130	Poor

1) Find the entropy of the training dataset with respect to target feature ‘Health Status’

2) Calculate the information gain of age relative to these training examples.

OR

16 Use the following data to construct a linear regression model for the auto insurance premium as a function of driving experience. (6)

Driving Experience (x)	5	2	12	9	15	6	25	16
Monthly Auto Insurance Premium (y)	64	87	50	71	44	56	42	60

Module IV

17 Explain properties of Artificial Neural Networks. (6)

OR

18 Explain the methods by which a non-linearly separable data can be classified using SVM. (6)

Module V

19 Demonstrate the working of k- means clustering by considering the following data set and assume first two samples as initial centroids. (Only one iteration is enough) (6)

Sample	**X**	**Y**
1	1.0	2.0
2	1.5	1.8
3	1.3	2.3
4	3.0	3.5
5	3.5	3.8
6	3.2	4
7	2.8	3.2

OR

20 A sentiment analysis tool classifies 80 comments as positive, out of which 60 are actual positives. The tool misses 20 positive comments. The total number of comments is 150, with 100 being positive. Construct a confusion matrix and calculate precision, recall, and accuracy. (6)

****