Reg No.:_____ Name:_____

0520MCA201122202

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Third Semester MCA (Two Years) (S, FE) Examination May 2024

Course Code: 20MCA201

Course Name: DATA SCIENCE AND MACHINE LEARNING

Max. Marks: 60 Duration: 3 Hours

PART A

Answer all questions, each carries 3 marks. Marks

1	Describe any three objectives of data exploration.	(3)
2	What are outliers? Does it always occur due to erroneous data capture? Explain.	(3)
3	State and explain Bayesian theorem for classification.	(3)
4	Why k-NN algorithm is called a lazy learner? Discuss.	(3)
5	Explain One Rule(1R) algorithm.	(3)
6	Differentiate between regression and classification.	(3)
7	What is the role of activation function in neural networks? Explain the behaviour of any one non linear activation function.	(3)
8	What is meant by ‘Kernel Trick’ in SVM? Explain with an example.	(3)
9	What is k-fold cross validation? Discuss.	(3)
10	Distinguish between bagging and boosting.	(3)

PART B

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

Module I

11	Explain the different steps involved in preparing the dataset to make it suitable for a data science task.	(6)

OR

12	Discuss the commonly used metrics for measuring the central tendency and spread in a dataset.	(6)

Module II

13 Discuss various ways for preparing data to use with k-NN. (6)

OR

14 Explain the significance of Laplace estimator in bayesian classification. Explain the different ways to prepare numeric features in naive Bayes algorithm? (6)

Module III

15 Draw a decision tree (upto level1 only) for the following data. The target attribute is **'Job offer** ' (6)

Sl No	CGPA	Interactiveness	Practical Knowledge	Communication Skills	Job Offer
1	>=9	Yes	Very Good	Good	Yes
2	>=8	No	Good	Moderate	Yes
3	>=9	No	Average	Poor	No
4	< 8	No	Average	Good	No
5	>=8	Yes	Good	Moderate	Yes
6	>=9	Yes	Good	Moderate	Yes
7	< 8	Yes	Good	Poor	No
8	>=9	No	Very Good	Good	Yes
9	>=8	Yes	Good	Good	Yes
10	>=8	Yes	Average	Good	Yes

OR

16 Determine the regression equation by finding the regression slope coefficient and the intercept value using the following data. (6)

X	55	60	64	71	80
Y	52	54	57	68	75

Module IV

17 Explain the different types of neural network topologies. (6)

OR

18 Describe the key concepts of SVM? Explain the significance of soft margin hyperplane and explain how they are computed. (6)

Module V

19 Discuss K-Means Clustering Algorithm. Mention any one criteria for choosing the value of 'K'. (6)

OR

20 Assume that, out of 1000 patients get tested for covid, 850 are actually healthy and 150 are actually sick. For the sick people, a test was positive for 120 and negative for 30. For the healthy people, the same test was positive for 75 and negative for the rest. Construct a confusion matrix and compute the accuracy, precision and recall for the data (6)

****