Reg No.:_______________ 0520MCA201122201 Name:___________________________

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Third Semester MCA (Two Years) Degree (R, S) Examination December/January 2023-24

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	What are the different data types used in data science?	(3)
2	Explain two data science tasks with example.	(3)
3	Illustrate the learning process by machine.	(3)
4	Can Naive Bayes algorithm handle Numerical continuous variables? Justify your answer.	(3)
5	What is pruning and why it is important in decision trees?	(3)
6	What is rule based classification and how is this technique used to derive classification rules from decision tree.	(3)
7	How do artificial neural networks model the human brain?	(3)
8	What is the significance of maximum margin hyper plane in support vector machine?	(3)
9	How are clusters formed in k-means clustering?	(3)
10	Explain about K-fold cross validation.	(3)

PART B

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

Module I

11	Describe data science process with a diagram.	(6)
	OR	
12	Explain the different 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 height=172 cm and weight =57 kg. Choose k=1 and k=3 (6)

Height (cm)	Weight (kg)	Class
167	51	Underweight
182	62	Normal
176	69	Normal
173	64	Normal
172	65	Normal
174	56	Underweight
169	58	Normal
173	57	Normal
170	55	Normal
169	53	Underweight

OR

14 Given a training dataset. Predict the Species type of new instance with Colour=Brown, Legs=2,Height=Tall, Smelly=No using Naive bayes classifier (6)

Colour	Legs	Height	Smelly	Species
White	3	Short	Yes	M
Brown	2	Tall	No	M
Brown	3	Short	Yes	M
White	3	Short	Yes	M
Brown	2	Short	No	H
White	2	Tall	No	H
White	2	Tall	No	H
White	2	Tall	Yes	H

Module III

15 Consider the training dataset. (6)

Outlook	Temp	Humidity	Wind	Play Tennis(Target Feature)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes

1)Find the entropy of the training dataset with respect to target feature 'Play Tennis'

2)Choose the best attribute to split the training dataset

OR

16 (6)

X	20	30	40	50	60	70
Y	58	54	50	46	42	38

Obtain a linear regression model from the data given in the table above. Assume that 'X' is the independent variable

Module IV

17 Explain the working of back propagation algorithm. What is the importance of Gradient descent Optimizer in it? (6)

OR

18 Explain how Support Vector Machine classifier can be used with non linearly separable data (6)

Module V

19 A search engine returns 40 records out of which only 20 are relevant. It fails to return 40 additional relevant records and there are 100 records in the database. Construct a confusion matrix and find out the precision and recall score for the search. (6)

OR

20 Suppose you are working on spam detection system. Assume ‘Spam’ is the positive class and ‘Not Spam’ is the negative class. A Test dataset contains 1000 e-mails,90% of these are ‘Not Spam’ and 10% are ‘Spam’. If the classifier always predicts as “Not Spam’, Calculate accuracy, Precision and Recall of classifier (6)

****