Reg No.:_____ Name:_____

0220MCA201122102

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

Third Semester MCA (Two Year) Regular and Supplementary Examination December 2022

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	Explain the objectives of data exploration.	(3)
2	With proper diagram explain holdout method.	(3)
3	How do you find the best splitting attribute in decision tree?	(3)
4	Explain methods to prepare data for use with k-NN.	(3)
5	Explain maximum margin hyperplane for linear separable data	(3)
6	Explain Laplace estimator with the help of an example.	(3)
7	Explain two types of data visualization techniques with one example each.	(3)
8	What are the strengths and weaknesses of training neural networks with backpropagation?	(3)
9	With the help of diagram explain ensembles.	(3)
10	Explain correlation in regression.	(3)

PART B

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

Module I

11	With the help of diagram explain data science process.	(6)

OR

12	Explain all the associated fields of data science in detail.	(6)

Module II

13	Consider the given dataset. Apply Naïve Bayes algorithm and predict that if a fruit has following properties, then which fruit it is. Fruit = (Yellow, Sweet, Long)	(6)

Fruit	Yellow	Sweet	Long	Total
Mango	350	450	0	650
Banana	400	300	350	400
Others	50	850	400	150
Total	800	850	400	1200

OR

14 Given the following dataset. Identify the T-Shirt Size of Tom having height 161 cm and weight 61kg using k-NN algorithm. (Choose k as 3) (6)

Height (cm)	Weight (kg)	T-Shirt Size
158	58	Medium
158	59	Medium
158	63	Medium
160	59	Medium
160	60	Medium
163	60	Medium
163	61	Medium
160	64	Large
163	64	Large
165	61	Large
165	62	Large
165	65	Large
168	62	Large
168	63	Large
168	66	Large
170	63	Large
170	64	Large
170	68	Large

Module III

15 Obtain a linear regression for the given dataset – (6)

Temp (x)	Attendance (y)
64	17
75	27
68	15
73	24
78	39
82	44
76	30
85	48
71	19
88	47

OR

16 For the following dataset, find the first splitting attribute of the decision tree to (6)

predict the diagnosis of a patient using ID3 algorithm.

Sore Throat	Fever	Swollen Glands	Congestion	Head Ache	Diagnosis
Yes	Yes	Yes	Yes	Yes	Strep Throat
No	No	No	Yes	Yes	Allergy
Yes	Yes	No	Yes	No	Cold
Yes	No	Yes	No	No	Strep Throat
No	Yes	No	Yes	No	Cold
No	No	No	Yes	No	Allergy
No	No	Yes	No	No	Strep Throat
Yes	No	No	Yes	Yes	Allergy
No	Yes	No	Yes	Yes	Cold
Yes	Yes	No	Yes	Yes	Cold

Module IV

17 Explain the characteristics need to define artificial neural networks. (6)

OR

18 Explain how SVM can be used for non-linearly separable data with the help of an example. (6)

Module V

19 Find the two clusters after one epoch for the given dataset using the k-means algorithm and Euclidean distance – (6)

No.	Height (H)	Weight (W)
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77
7	180	71
8	180	70
9	183	84
10	180	88
11	180	67
12	177	76

OR

20	There is total 165 patients, out of them, 60 are actually healthy and 105 are actually sick. For the sick people, a test was positive for 100 and negative for 5 patients. For the healthy people, the same test was positive for 10 and negative for 50. Construct a confusion matrix for the given data and compute the sensitivity and F-measure.	(6)

****