

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 is data science and why do we need data science? | (3) |
| 2 | Explain the different types of data. | (3) |
| 3 | Explain the differences between supervised and unsupervised machine learning algorithms. | (3) |
| 4 | What are the strengths and weaknesses of K-NN algorithm | (3) |
| 5 | How to simplify a decision tree by pruning. | (3) |
| 6 | Explain the Ordinary Least Square method in regression. | (3) |
| 7 | Define activation function. Give two examples. | (3) |
| 8 | What is maximum margin hyperplane? | (3) |
| 9 | What is K-fold cross validation? | (3) |
| 10 | Explain bootstrap sampling | (3) |

PART B*Answer any one question from each module. Each question carries 6 marks.***Module I**

- | | | |
|----|---|-----|
| 11 | Explain the various processes for preparing a dataset to perform a data science task. | (6) |
|----|---|-----|

OR

- | | | |
|----|---|-----|
| 12 | The tensile strength in megapascals for 15 samples of tin were determined and found to be: 34.61, 34.57, 34.40, 34.63, 34.63, 34.51, 34.49, 34.61, 34.52, 34.55, 34.58, 34.53, 34.44, 34.48 and 34.40. Calculate the mean and standard deviation from the mean for these 15 values, correct to 4 significant figures. | (6) |
|----|---|-----|

Module II

- | | | |
|----|---|-----|
| 13 | Based on the survey conducted in an institution the students are classified based on the 2 attributes academic excellence and other achievements. Consider the data set given. Find the classification of a student with value of X is 5 and Y is 7 | (6) |
|----|---|-----|

based on the data of trained samples using KNN algorithm. Choose $k = 3$

X [Academic Excellence]	Y [Activities]	Z [Classification]
8	6	Outstanding
5	6	Good
7	3	Good
6	9	Outstanding

OR

- 14 Consider a training data set consisting of the fauna of the world. Each unit has 3 features named “Swim”, “Fly” and “Crawl”. Let the possible values of these features be as follows:

Swim - Fast, Slow, No

Fly - Long, Short, Rarely, No

Crawl - Yes, No

For simplicity, each unit is classified as “Animal”, “Bird” or “Fish”. Let the training data set be as in the table below . Use naive Bayes algorithm to classify a particular species if its features are (Slow, Rarely, No)

Sl. No.	Swim	Fly	Crawl	Class
1	Fast	No	No	Fish
2	Fast	No	Yes	Animal
3	Slow	No	No	Animal
4	Fast	No	No	Animal
5	No	Short	No	Bird
6	No	Short	No	Bird
7	No	Rarely	No	Animal
8	Slow	No	Yes	Animal
9	Slow	No	No	Fish
10	Slow	No	Yes	Fish

11	No	Long	No	Bird
12	Fast	No	No	Bird

Module III

- 15 Consider the following set of training examples:

(6)

Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

- a) Find the entropy of this collection of training examples with respect to the target function “classification”? (3 marks)
- b) Calculate the information gain of a2 relative to these training examples? (3 marks)

OR

- 16 How to estimate the parameters of a linear regression model ?

(6)

Module IV

- 17 Discuss the basic idea behind the back propagation algorithm.

(6)

OR

- 18 a) Define linearly separable dataset. Give an example each of a dataset that is linearly separable and of a dataset that is not linearly separable. (3 marks)

(6)

- b) Define kernel function. Explain the kernel trick to construct a classifier for a dataset that is not linearly separable. (3 marks)

Module V

- 19 Suppose 10000 patients get tested for flu; out of them, 9000 are actually healthy and 1000 are actually sick. For the sick people, a test was positive for 620 and

(6)

negative for 380. For the healthy people, the same test was positive for 180 and negative for 8820. Construct a confusion matrix for the data and compute the precision and recall for the data.

OR

20 Explain the concepts of bagging and boosting.

(6)
