Roll No. | | | | | | | | | | | | **Total No. of Pages : 02**

Total No. of Questions : 08

M.Tech.(Computer Science & Engineering) (Sem.–1)

DATA SCIENCE

Subject Code : MTCS-108-18

M.Code : 75158

Date of Examination : 19-01-2023

Time : 3 Hrs. **Max. Marks : 60**

INSTRUCTIONS TO CANDIDATES :

1. **Attempt any FIVE questions out of EIGHT questions.**
2. **Each question carries TWELVE marks.**

1. a) What is the curse of dimensionality?

   b) How do you decide whether your linear regression model fits the data?

2. What is the difference between squared error and absolute error? In a population of tiny birds, the diameter of the egg and the weight of the hatchling (the baby bird that hatches from the egg) follows the regression model. The summary statistics in the sample are: correlation = 0.75

	mean	**SD**
egg diameter (mm)	23	0.5
bird weight (gm)	6	0.4

   Find the regression estimate of the weight of a bird that hatches from an egg of diameter 24 mm.

3. In a population, 85% of the people are in Class A and the remaining 15% are in Class B. For people in Class A, a classifier has an accuracy of 90% (that is, among Class A people, 90% are classified as Class A and 10% as Class B). For people in Class B, the accuracy of the classifier is 98%. One person is picked at random from the population.

   What is the chance that the person is classified correctly?

4. What technique is used to predict categorical responses? What is logistic regression? State an example when you have used logistic regression recently.

5. Why data cleaning plays a vital role in analysis? Differentiate between univariate, bivariate and multivariate analysis.

6. What are the different methods of collecting large amount of Data from Social Media? What are the most popular APIs used for Data Collection? What are the different rate limitations on these APIs? How is data collected from multiple sources handled?

7. What are categorical variables? A test has a true positive rate of 100% and false positive rate of 5%, There is a population with a 1/1000 rate of having the condition the test identifies. Considering a positive test, what is the probability of having that condition?

8. In a large random sample of U.S. households, the median annual income is $54,000. This original sample is bootstrapped 5,000 times and the sample median is recorded for each of the bootstrap samples. The middle 95% interval of these values is ($53,000, $55,000).

   a) True **or** false (explain your answer): The interval ($53,000, $55,000) is an approximate -bootstrap 95% confidence interval for the median income of all the households in the sample.

   b) Pick the option that you think best completes the sentence, and explain your choice. The percent of all U.S. households with annual incomes in the range ($53,000, $55,000)

      I. Is about 95%.
      II. Is about 50%.
      III. cannot be approximated based on the information given.

NOTE : Disclosure of Identity by writing Mobile No. or Making of passing request on any page of Answer Sheet will lead to UMC against the Student.