Roll No. | | | | | | | | | | | | |

Total No. of Pages : 02

Total No. of Questions : 8

M.Tech.(Computer Science & Engineering) E-I (2018 Batch) (Sem.–1)

DATA SCIENCE

Subject Code : MTCS-108-18

M.Code : 75158

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. Differentiate between Data Science, Machine Learning and Al. Python or R - Which one would you prefer for text analytics?

2. Researchers-studying health insurance in the United States have gathered data on whether or not people are insured. There are several thousand people in the study. The table insured contains one row for each person. The table has three columns in the following order: the column Name contains the person's name; Zip Code contains the zip code of the person's home address; and Insured is a 0/1 variable where 1 means "insured" and 0 means “not insured”. The table states consists of one row for each zip code in the United States. The first column is labeled Zip Code and contains the zip code; the second column is labeled State and contains the name of the state (such as California, or New York) in which that zip code is located. Write Python code in each of the following parts. You can use multiple lines of code. The last line of your code should evaluate to the element described in the question.

   (a) the proportion of insured people in the study

   (b) a state that has the largest number of insured people among the all states represented in the study

3. A data science class has 500 students. As part of an assignment, each student tests the fairness of a coin using data from his/her own set of tosses of the coin. All 500 students test the same coin, and they all test the same pair of hypotheses:

   Null: The coin is fair. Alternative: The coin is not fair.

All of the students use the 5% cutoff for the P-value. You can assume that all the students perform the same test based on the same large number of tosses.

Suppose that, unknown to the students, the coin is fair. About how many students will conclude that the coin is not fair?

Pick one option and justify your choice :

a) No students

b) 5 students

c) 10 students

d) 25 students

e) 250 students

4. a) What is Regularization and what kind of problems does regularization solve?

   b) What is multicollinearity and how you can overcome it?

5. 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?

6. a) What is power analysis? What is K-means? How can you select K for K-means?

   b) What is Collaborative filtering? What is the difference between Cluster and Systematic Sampling?

7. a) What is Machine Learning? Can you use machine learning for time series analysis?

   b) Write a function that takes in two sorted lists and outputs a sorted list that is their union.

8. What are the different types of data Visualizations? Also explain different types of Data Encodings and Retinal Variables.

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.