Printed Page:- Subject Code:- ABT0514

Roll. No:

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA

(An Autonomous Institute Affiliated to AKTU, Lucknow)

B.Tech.

SEM: V - THEORY EXAMINATION (2022 - 2023)

Subject: Data Science

Time: 3 Hours Max. Marks: 100

General Instructions:

IMP: Verify that you have received the question paper with the correct course, code, branch etc.

1. This Question paper comprises of three Sections -A, B, & C. It consists of Multiple Choice Questions (MCQ's) & Subjective type questions.

2. Maximum marks for each question are indicated on right -hand side of each question.

3. Illustrate your answers with neat sketches wherever necessary.

4. Assume suitable data if necessary.

5. Preferably, write the answers in sequential order.

6. No sheet should be left blank. Any written material after a blank sheet will not be evaluated/checked.

SECTION A 20

1. Attempt all parts:-

1-a. Which of the following uses data on some object to predict values for other object (CO1) 1

(a) Inferential

(b) Exploratory

(c) Predictive

(d) None of the mentioned

1-b. Which of the following focuses on the discovery of (previously) unknown properties on the data (CO1) 1

(a) Data mining

(b) Big Data

(c) Data wrangling

(d) Machine Learning

1-c. Which is FALSE regarding regression? (CO2) 1

(a) It may be used for interpretation

(b) It is used for prediction

(c) It discovers causal relationships

(d) It relates inputs to outputs

1-d. What are outliers? (CO2) 1

(a) It is the main trend of our dataset

(b) Extreme datapoints in our dataset

(c) It is a regression technique

(d) Values that are correlated to eachother

1-e. Which one is the property of correlation? (CO3) 1

(a) $(-1 \leq r \leq +1)$

(b) r=0 represents no linear relationship between the two variables

(c) Correlation is unit free

(d) All of the above

1-f. The Correlation coefficient is independent of change of: (CO3) 1

(a) Scale

(b) Origin

(c) Both origin and scale

(d) neither origin nor scale

1-g. The algebraic sum of the deviations from mean is: (CO4) 1

(a) Maximum

(b) Minimum

(c) Zero

(d) None of the above

1-h. The average of the sum of squares of the deviations about mean is called… (CO4) 1

(a) Standard Deviation

(b) Variance

(c) Absolute Deviation

(d) Mean Deviation

1-i. In a logistic regression model, the decision boundary can be (CO5) 1

(a) linear

(b) non-linear

(c) both (A) and (B)

(d) none of these

1-j. What's the cost function of the logistic regression? (CO5) 1

(a) Sigmoid function

(b) Logistic Function

(c) both (A) and (B)

(d) none of these

2. Attempt all parts:-

2.a.	Compare and Contrast Business Analyst and Data Analyst? (CO1)	2
2.b.	What is data sampling? (CO2)	2
2.c.	What is the difference between a sample and a population? (CO3)	2
2.d.	What is meant by dependent and independent variables? (CO4)	2
2.e.	What varieties of logistic regression are there? (CO5)	2

SECTION B 30

3. Answer any five of the following:-

3-a.	What are the most important qualities of a data scientist? (CO1)	6
3-b.	What are the various statistical techniques employed in data science? (CO1)	6
3-c.	What are the goals of data cleaning process? (CO2)	6
3-d.	How would you deal with Outliers in your dataset? (CO2)	6
3.e.	What exactly is a hypothesis? How do scientists generate hypotheses? (CO3)	6
3.f.	What do you mean by central tendency? Describe the methods of measuring the central tendency? (CO4)	6
3.g.	How is multinomial logistic regression implemented? (CO5)	6

SECTION C 50

4. Answer any one of the following:-

4-a.	Describe the word Datafication and explain how it can be used for insurance and banking? (CO1)	10
4-b.	Differentiate between Business Intelligence and Data science with example? (CO1)	10

5. Answer any one of the following:-

5-a.	What are the definitions of Graph Data Science and its applications? (CO2)	10
5-b.	What do you mean by high dimensional data and what is dimensionality? (CO2)	10

6. Answer any one of the following:-

6-a.	A random sample of 200 measurements from a large population gave a mean value of 50 and S.D. of 9. Determine 95% confidence interval for mean of population. (CO3)	10
6-b.	The height of 8 males participating in an athletic championship are found to be 175,168,165,170,167,160,173 and 168 cm. Can we conclude that the average height is greater than 165 cm? (Test at 5% level of significance) (CO3)	10

7. Answer any one of the following:-

7-a.	The two regression lines are 3X+2Y=26 and 6X+3Y=31. Find the correlation coefficient. (CO4)	10
7-b.	Write regression equations of Y on X for the following data - (CO4) X:- 45 48 50 55 65 70 75 72 80 85 Y:- 25 30 35 30 40 50 45 55 60 65	10

8. Answer any one of the following:-

8	List the five main uses of machine learning and elaborate on the distinctions between classification and regression. (CO5)	10
8	Why is Mean Square Error (MSE) ineligible for use as a cost function in Logistic Regression? (CO5)	10