Printed Page:- 04

Subject Code:- ABT0514N

Roll. No:

NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA

(An Autonomous Institute Affiliated to AKTU, Lucknow)

B.Tech

SEM: V - THEORY EXAMINATION (2024- 2025)

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 is not a step in data analysis? (CO1, K1) 1

(a) Obtain the data

(b) Clean the data

(c) EDA

(d) None of the mentioned

1-b. Which of the following is correct skills for a Data Scientist? (CO1, K1) 1

(a) Probability & Statistics

(b) Machine Learning / Deep Learning

(c) Data Wrangling

(d) All of the above

1-c. Which of the following is the most important language for Data Science? (CO2, K2) 1

(a) Java

(b) Ruby

(c) R

(d) None of the mentioned

1-d. What does "mode" mean? (CO2, K2) 1

(a) average

(b) middle

(c) Most frequent Number

(d) variation

1-e. Identify the correct one. (CO3, K3) 1

(a) Standard deviation=√Variance

(b) Variance=(Standard deviation)^2

(c) mean=(∑x)/n

(d) All of the above

1-f. Predict which one is the types of regression (CO3, K2) 1

(a) Linear Regression

(b) Non-Linear Regression

(c) Multiple-Linear Regression

(d) All of the above

1-g. Predict the shape of the Normal Curve is _____ (CO4, K3) 1

(a) Bell Shaped

(b) Flat

(c) Circular

(d) Spiked

1-h. Determine the option, "The average of the sum of squares of the deviations about mean is called"? (CO4, K3) 1

(a) Standard Deviation

(b) Variance

(c) Absolute Deviation

(d) Mean Deviation

1-i. Logistic regression is used to determine? (CO5, K3) 1

(a) classification

(b) regression

(c) clustering

(d) All of these

1-j. In a logistic regression model, the decision boundary can be determined by (CO5, K3) 1

(a) linear

(b) non-linear

(c) both (A) and (B)

(d) none of these

2. Attempt all parts:-

2.a. Identify the two differences between structured and unstructured data? (CO1, K1) 2

2.b. Explain briefly different types of data present in data science? (CO2, K2) 2

2.c. Describe dependent and independent variables? (CO3, K2) 2

2.d.	Illustrate the importance of Critical Value Parameter? (CO4, K3)	2
2.e.	Express the meaning of Logistic Regression? (CO5, K3)	2

SECTION-B 30

3. Answer any five of the following:-

3-a.	Briefly describe the data science lifecycle model? (CO1, K1)	6
3-b.	Describe the various statistical techniques employed in data science? (CO1, K1)	6
3-c.	Describe Data Preprocessing? Explain preprocessing steps do you know? (CO2, K2)	6
3-d.	Explain Dimensionality reduction? (CO2, K2)	6
3.e.	Discuss Regression Analysis. Explain the difference between Correlation and Regression. (CO3, K2)	6
3.f.	Determine why do we test the null hypothesis to establish an effect instead of the alternative hypothesis? (CO4, K3)	6
3.g.	Determine how would you assess whether the assumptions of Logistic Regression hold true before applying the model? (CO5, K3)	6

SECTION-C 50

4. Answer any one of the following:-

4-a.	Define Data Science. Explain its key components and describe how Data Science combines techniques from statistics and computer science. (CO1, K1)	10
4-b.	Describe the changes that data science is making to the healthcare sector. (CO1, K1)	10

5. Answer any one of the following:-

5-a.	Give a brief explanation of the various forms of data and their sources. (CO2, K2)	10
5-b.	Explain Graph Data Science and what are the applications of Graph Data Science? (CO2, K2)	10

6. Answer any one of the following:-

6-a.	The two regression lines are 3X+2Y=26 and 6X+3Y=31. Predict the correlation coefficient. (CO3, K3)	10
6-b.	Define Coefficient of Variation. Also compute the Median of following data Age Group:- 0-20 20-40 40-60 60-80 80-100 Frequency:- 15 32 54 30 19 (CO3, K3)	10

7. Answer any one of the following:-

7-a.	You are given a sample of 20 items with a mean of 42 units and a standard deviation of 5 units. Using a significance level of 5%, perform a hypothesis test to determine whether the sample could have come from a population with a mean of 45 units. (CO4, K3)	10
7-b.	Two samples are given: one of size 10 with a mean of 20.3 and a standard deviation of 3.5, and the other of size 14 with a mean of 18.6 and a standard deviation of 5.2. At a 5% significance level, test whether the means of the two populations are equal. (CO4, K3)	10

8. Answer any <u>one</u> of the following:-

8-a.	Using your knowledge of machine learning, apply the five main uses of machine learning to a real-world scenario of your choice. Then, explain how you would choose between classification and regression techniques. (CO5, K3)	10
8-b.	Express the Case Studies of Data Science in Biotechnology with helps of some examples? (CO5, K3)	10