51N0407 /

51N0407 /

B.TECH. V SEMESTER (NEW SCHEME)
MAIN/BACK EXAMINATION 2024-25
COMPUTER SCIENCE & ENGINEERING
5CS5-12 - Introduction to Data Science
[5AM5-12, 5IT5-12, 5CM5-12, 5CY5-12]
Common to AM, IT, CM, CY, CS

Time : 3 Hours]

[Max. Marks : 70

[Min. Passing Marks :

Instructions to Candidates :

Part-A : Short Answer Type Questions (up to 25 words) 10 × 2 = 20 marks. All 10 questions are compulsory.

Part-B : Analytical/Problem Solving questions 5 × 4 = 20 marks. Candidates have to answer 5 questions out of 7.

Part-C : Descriptive/Analytical/Problem Solving questions 3 × 10 marks = 30 marks. Candidates have to answer 3 questions out of 5.

Schematic diagrams must be shown wherever necessary. Any data you feel missing may suitably be assumed and stated clearly. Units of quantities used/ calculated must be stated clearly.

Use of the following supporting materials is permitted during examination. (Mentioned in form no. 205).

1 Nil

2 Nil

Part-A

1. Outline the key stages of the data analytics lifecycle.
2. What are the 5 V's of big data?
3. What is the need for data science in the present era ?
4. Define hypothesis testing, and write its purpose.
5. What types of errors may arise in machine learning models ?
6. How can missing data in a dataset be effectively addressed and filled ?
7. How do effective data analysis and visualization contribute to informed decision-making in various industries ?
8. Explain the Naive Bayes theorem.
9. Outline the key analyses you would perform to understand the characteristics of the data better.
10. Explore the practical applications of data science within the Agriculture sector.

Part-B

1. Explain the following :
   (i) Hypothesis Testing
   (ii) p-Value
2. Write about the Method in Descriptive Statistics and Explain the role of inferential statistics in analyzing various datasets.
3. What are the Nominal, Ordinal, Discrete, and Continuous data types ? Discuss with examples.
4. Describe Exploratory Data Analysis and its role in data science.

5. Explain Regression analysis and types of Regression Analysis in detail.

6. Explain the purpose of a Support Vector Machine in supervised learning technique.

7. Explain the random forest technique with functionality.

Part-C

1. Explain correlation. Describe the statement "correlation is not causation" with an example in detail.

2. How does the application of machine learning contribute to the field of data science, and what specific challenges and opportunities arise in integrating machine learning techniques for effective data analysis and decision-making ?

3. Provide an in-depth analysis of the phenomena of overfitting and underfitting within machine learning, including a thorough explanation supported by a specific example for each.

4. Build a decision tree model for a given dataset with 100 instances. The dataset is split into Positive (P) and Negative (N) classes. Initially, the dataset contains 60 instances of class P and 40 instances of class N. You apply a binary split based on a feature, resulting in two child nodes :

   Node 1 contains 20 instances of class P and 30 instances of class N.

   Node 2 contains 40 instances of class P and 10 instances of class N.

   Calculate the following :

   (a) The entropy of the initial dataset before the split.

   (b) The entropy of each child node after the split.

5. Detail the steps you would take to clean and prepare the data for analysis. Include how you would handle missing values, outliers, and any data transformation you might apply. Support your answer with canonical examples.

****