Hall Ticket No | | | | | | | | | | |

Question Paper Code: ACAC07

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal-500043, Hyderabad

B.Tech V SEMESTER END EXAMINATIONS (REGULAR/ SUPPLEMENTARY) - FEBRUARY 2024
Regulation: UG20

STATISTICAL FOUNDATIONS OF DATA SCIENCE

Time: 3 Hours **CSE (AI&ML)** **Max Marks: 70**

Answer ALL questions in Module I and II
Answer ONE out of two questions in Modules III, IV and V
All Questions Carry Equal Marks
All parts of the question must be answered in one place only

MODULE – I

1. (a) Differentiate between structured and unstructured data, and how do these differences impact the storage, processing, and analysis of data in various industries and applications? [BL: Understand| CO: 4|Marks: 7]

   (b) "Quantitative and qualitative data differ in nature and require distinct methods of data collection and analysis in research". How these differences can impact the methods of data collection and analysis. [BL: Apply| CO: 1|Marks: 7]

MODULE – II

2. (a) Illustrate the working of random forest with a neat diagram. Write the differences between feature selection and feature extraction. [BL: Understand| CO: 2|Marks: 7]

   (b) For creating a decision tree model to predict if a loan applicant will likely default, outline the approach to this task, including the steps taken to prepare the data, train the decision tree model, and evaluate its performance. Also, describe how the decision tree would be interpreted to gain insights into the factors affecting loan defaults. [BL: Apply| CO: 2|Marks: 7]

MODULE – III

3. (a) How are vectors and matrices utilized in data science applications? Elaborate with suitable examples. [BL: Understand| CO: 3|Marks: 7]

   (b) In a survey, 60% of respondents said they prefer tea over coffee. If you randomly select 10 individuals from the survey population, what is the probability that at least 7 of them prefer tea? [BL: Apply| CO: 3|Marks: 7]

4. (a) Summarize data transformation functions. Outline about probability and discuss the significance of probability in the context of statistics. [BL: Understand| CO: 4|Marks: 7]

   (b) In a dataset of student exam scores, assuming a normal distribution, what proportion of students scored above 85? The mean score is 75, with a standard deviation of 10. [BL: Apply| CO: 4|Marks: 7]

MODULE – IV

5. (a) Mention the steps involved in a hypothesis test to determine if the average test scores of two groups are significantly different. [BL: Understand| CO: 5|Marks: 7]

   (b) With a dataset comprising the heights of students in a class, elucidate the utilization of descriptive statistics to summarize and analyze the data. Detail the steps for computing the mean, median, mode, standard deviation, and range. Elucidate how these statistics would be interpreted to glean insights into the distribution of heights within the class. [BL: Apply| CO: 5|Marks: 7].

6. (a) Write a short note on scatter plot. Demonstrate different types of correlation with a suitable diagram. [BL: Understand| CO: 5|Marks: 7]

   (b) A study is conducted to estimate the average response time of a website's server to user requests. Describe the confidence interval techniques employed to determine a range of values within which the average response time will likely fall. Include the steps necessary to collect data, calculate the confidence interval, interpret the results, and effectively communicate them to stakeholders. [BL: Apply| CO: 5|Marks: 7]

MODULE – V

7. (a) Illustrate Simpson's Paradox. Give an example scenario where Simpson's Paradox could occur. Include details about the subgroups and the overall trend. [BL: Understand| CO: 6|Marks: 7]

   (b) In analyzing the distribution of student exam scores in a class, outline the use of histograms and box plots to represent this distribution visually. Detail the steps in creating these visualizations, selecting appropriate bin widths and labels for histograms, interpreting the patterns and outliers displayed in the box plot, and effectively communicating the findings to stakeholders. [BL: Apply| CO: 6|Marks: 7]

8. (a) Apply the "why/how/what" strategy of presenting data. Explain how this strategy can enhance the effectiveness of data communication ? [BL: Understand| CO: 6|Marks: 7]

   (b) Working on a project that involves analyzing the correlation between rainfall and crop yields in a farming region, describe how graphs and statistics will be used to visualize and analyze this relationship. Include the steps to select appropriate graph types, calculate relevant statistical measures, interpret the results, and communicate the findings to stakeholders. Additionally, explain using graphs and statistics to identify patterns, trends, and correlations between rainfall and crop yields. [BL: Apply| CO: 6|Marks: 7]