Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhaniyam Sankhya 12 of 2011)
University under section 2(f) of UGC Act, 1956

End Term (Odd) Semester Examination November 2025

Roll no...............................

Name of the Course and semester: BCA
Name of the Paper: Introduction to Data Science
Paper Code: TBD 303

Time: 3 hour

Maximum Marks: 100

Note:
- *(i)* All the questions are compulsory.
- *(ii)* Answer any two sub questions from a, b and c in each main question.
- *(iii)* Total marks for each question is 20 (twenty).
- *(iv)* Each sub-question carries 10 marks.

Q1. CO1   (2X10=20 Marks)
a. List and explain 5 applications of data science in real world.
b. Describe the concept of Hadoop and its components.
c. Discuss the various security related issues associated with data. Explain the various facets of Data science.

Q2. CO2   (2X10=20 Marks)
a. Illustrate the role of Exploratory Data Analysis (EDA) in the Data Science Process. Mention 5 techniques used in EDA.
b. Explain the need of cleaning the data. Discuss various techniques used for cleaning a data.
c. Describe the steps involved in retrieving, cleaning, and integrating data in a data science project.

Q3. CO3   (2X10=20 Marks)
a. You are working as a data scientist and got a project to analyse the data of a school. You are given marks of 10 students: 50, 60, 70, 65, 95, 80, 85, 65, 50, 55.
   i. Compute mean, median, mode of the given data.
   ii. Show how these stats can help to identify week students.
b. **i.** Describe how Z-Score is used for detecting the outliers?
   ii. A logistics company tracks delivery times and finds they follow a normal distribution, with Mean delivery time = 48 minutes, Standard deviation = 6 minutes. Calculate the Z-score for a delivery completed in 60 minutes. Show the scenario graphically using appropriate curve.
c. i. Compare and contrast the following types of t-tests with suitable examples of when each should be applied:
      I. One-sample t-test
      II. Independent two-sample t-test
      III. Paired t-test
   ii. Differentiate between the chi-square test for independence and the chi-square goodness-of-fit test with appropriate real-life examples.

Q4. CO4   (2X10=20 Marks)
a. A data scientist must create a dashboard summarizing customer purchasing behavior.
   The dataset contains numerical variables (age, spending score), categorical variables (gender, region), and time-based sales data.
   i. Analyze the difference between exploratory and explanatory visualizations in this context.
   ii. Analyze which visualization techniques (line chart, bar chart, scatter plot, heatmap, box plot, pie chart) are most appropriate for each variable type.
b. A retail company wants to analyze correlations between variables like price, discount, sales quantity, and customer rating.
   i. Analyze and recommend whether a **heatmap, pair plot, or scatter plot matrix** is more suitable and

End Term (Odd) Semester Examination November 2025

explain why mentioning the advantage and disadvantage of each.

ii. Explain how a correlation matrix can be misleading and how visualization resolves these issues.

iii. Plot a boxplot and show how boxplot is used for detecting outliers.

c. Describe the following with diagram and use case of each in real lifer scenario:

   i. BoxPlot

   ii. Pie Chart

   iii. Donut Chart

   iv. Pair Plot

   v. Radar Chart

Q5. CO5, CO6   (2X10=20 Marks)

a. Differentiate between supervised and unsupervised learning with the help of an algorithm of each.

b. Given a dataset containing points: (2,3), (1,2), (2,4), (7,8), (9,10) with labels as 1, 1, 0, 0, 1. A new point (4,6) is obtained. Apply kNN to find the label of a new point with k = 3.

c. A machine learning model is used to detect whether an email is *spam* or *not spam*. After testing the model on 200 emails, the following results were obtained: True Positives (TP) = 48, True Negatives (TN) = 120, False Positives (FP) = 12, False Negatives (FN) = 20. Calculate the following:

i. Accuracy

ii. Precision

iii. Recall

iv. F1-Score

Note For the question paper setters:

- Question paper should cover all the COs of the course.
- Please specify COs against each question.