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

End Term (Even) Semester Examination May-June 2025

Roll no.....

Name of the Program and semester: MCA 4
Name of the Course: Data Science using R
Course Code: TMC 401

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. (2X10=20 Marks) CO1

a. Define data science. List the stages in the data science process.
b. Compare and contrast supervised and unsupervised learning.
c. Explain the relationship between data preparation and model accuracy.

Q2. (2X10=20 Marks) CO5

a. Define descriptive statistics. Interpret a boxplot and describe its components.
b. Describe the characteristics of a well-prepared dataset.
c. What is correlation? Create a heatmap to show variable correlations in R.

Q3. (2X10=20 Marks) CO4

a. what is decision tree? Describe how rule induction works in classification. Explain any decision tree algorithm (C4.5, ID3, CART).
b. Analyze confusion matrix results from classification. Give all the evaluation metrices achieved from confusion matrix
c. What is K-means Clustering algorithm. How would you define number of Clusters and cetroids.

Q4. (2X10=20 Marks) CO2, CO3

a. Define linear and logistic regression. Apply logistic regression on binary data in R.
b. What is Association rule? Explain support, confidence, and lift in rules mining.
c. Apply feature scaling in regression modeling. Describe how regression is used in real-world scenarios.

Q5. (2X10=20 Marks) CO6

a. Define text mining and time series.
b. Explain the process of tokenization and stemming. Apply TF-IDF on a text dataset.
c. Explain autocorrelation and lag in time series.