Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF INFORMATION SCIENCE
END TERM EXAMINATION - JAN 2023

Semester : Semester III - 2021
Course Code : CSA2019
Course Name : Sem III - CSA2019 - R Programming for Datascience
Program : B.Sc. Data Science

Date : 6-JAN-2023
Time : 9.30AM - 12.30PM
Max Marks : 100
Weightage : 50%

Instructions:
(i) Read all questions carefully and answer accordingly.
(ii) Question paper consists of 3 parts.
(iii) Scientific and non-programmable calculator are permitted.

PART A

ANSWER ALL THE FOLLOWING QUESTIONS **10 X 2 = 20M**

1. Mention the drawbacks of data science
(CO1) [Knowledge]

2. Define exploratory data analysis.
(CO2) [Knowledge]

3. Define covariation among variables with an example.
(CO2) [Knowledge]

4. Explain gather and spread functions
(CO1) [Knowledge]

5. List any two advantages of Decision trees
(CO4) [Knowledge]

6. Write a short note on principal component analysis algorithm
(CO4) [Knowledge]

7. write a function to plot scatterplot on income versus happiness on the dataset named "Income_data"
(CO2) [Knowledge]

8. Define KNN algorithm
(CO4) [Knowledge]

9. Mention the ways to deal with missing values in data analysis
(CO2) [Knowledge]

10. Outline the three rules which make the dataset tidy.
(CO2) [Knowledge]

PART B

ANSWER ALL THE FOLLOWING QUESTIONS **5 X 10 = 50M**

11. Write a program in R
a. to find factorial of a given number
b. to check whether a given number is prime or not

(CO1) [Comprehension]

12. Explain the following dplyr functons with syntax and example
a. filter() b. summarise() c.rename() d. select() e. distinct()

(CO2) [Comprehension]

13. Discuss the assumptions of simple linear regression with an example.

(CO3) [Comprehension]

14. Classification and regression are two categories of supervised machine learning algorithms. Explain in brief the techniques in each category.

(CO4) [Comprehension]

15. Discuss K Nearest Neighbour algorithm with its features and an example

(CO4) [Comprehension]

PART C

ANSWER ALL THE FOLLOWING QUESTIONS **2 X 15 = 30M**

16. Consider the data set "heart_data" available in the R environment. It gives a relation between the input variables biking, smoking and output variable heart disease. Explain the steps to build and evaluate a multi linear regression model in R to find how heart disease is affected by the given input variables.

(CO3,CO4) [Application]

17. Write programs in R for the following questions
a. Mohan wants to purchase a Digital-SLR camera, but he is confused with many options available in the market. Software is required to recommend Mohan to select the camera based on Pixel size (PZ) and Price (P). Three cameras are available in the market, CANON, NIKON, and SONY. Write a program in R with the following requirements.
Requirements:
i. Read the pixel (PZ) and P from the user.
ii. The CANON camera is recommended if Megapixel is in the range 35-40 or price starts from ₹45,000.
iii. The NIKON camera is recommended if Megapixel is in the range 25-34 or price starts from ₹30,000.
iv. Otherwise SONY is recommended.
v. Print the correct recommendation.

b. create a vector with 6 numbers and find the maximum and minimum element of the vector. Also find the sum of the elements.

(CO1) [Application]