Register No.: ..... Name: .....

SAINTGITS COLLEGE OF ENGINEERING (AUTONOMOUS)

(AFFILIATED TO APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY, THIRUVANANTHAPURAM)

SIXTH SEMESTER B.TECH DEGREE EXAMINATION (R), MAY 2023
COMPUTER SCIENCE AND ENGINEERING
(2020 SCHEME)

Course Code : 20CST322

Course Name: Data Analytics

Max. Marks : 100 **Duration: 3 Hours**

PART A

(Answer all questions. Each question carries 3 marks)

1. A fair six-sided die is rolled twice. What is the probability that the sum of the two numbers rolled is at least 8?
2. What is the mean, median, and mode of the following data set: 2, 4, 6, 8, 8, 10?
3. What are the different types of data analysis?
4. Classify the stages in the process of Intelligent Data Analysis.
5. What is the difference between supervised and unsupervised machine learning?
6. Explain Linear regression.
7. Enumerate the terms a. OLAP b. OLTP c. RTAP
8. How can you overwrite the replication factors in HDFS?
9. What are the different forms of data types and how to test the data type in R? Give one example for each.
10. How can you load a .csv file in R?

PART B

(Answer one full question from each module, each question carries 14 marks)

MODULE I

11. a) With necessary examples explain measure of dispersion. (6)
    b) What is the interquartile range (IQR)? How is it calculated? (8)

OR

12. a) What is correlation in statistics? How is it calculated? (8)
    b) What is the difference between a confidence interval and a prediction interval, and when should you use each one? (6)

MODULE II

13. a) Explain the stages of Data Analytics Life Cycle? (6)

b) A company produces light bulbs that have a mean life of 500 hours and a standard deviation of 20 hours. Assuming a normal distribution, what is the probability that a randomly selected bulb will last between 470 and 520 hours? Also, calculate the 95% confidence interval for the mean life of the light bulbs. (8)

OR

14. a) What are some common techniques for dimensionality reduction, and how do they work? Provide an example of a real-world scenario where dimensionality reduction could be useful. (7)

b) Discuss the advantages and disadvantages of random sampling and stratified sampling. In what situations would you prefer one over the other? Provide an example to support your answer. (7)

MODULE III

15. a) Describe about training and testing data more clearly with an example? (7)

b) How is KNN different from K-means clustering Justify your answer with proper example? (7)

OR

16. a) Illustrate agglomerative hierarchical clustering with an example. (7)

b) Determine the most important documents for supervised learning. (7)

MODULE IV

17. a) How would you assess the difficulties faced by conventional systems? (8)

b) What are the differences that separate out big data architecture from the traditional one? (6)

OR

18. a) Explain in detail the Ecosystem of the Hadoop Framework. (7)

b) Explain how MapReduce collaborates with the Hadoop Software stack in the Data Analytics process. (7)

MODULE V

19. a) Elaborate the following R objects. (8)
a) vector b) data frame c) matrix d) list

b) How to get system date in R? Generate sequence of previous and coming 10 dates from today in R. (6)

OR

20. a) What are the common graphical techniques used in exploratory data analysis, and how can they be used to uncover patterns or anomalies in the data? (7)

b) What are the different types of statistical tests used for evaluating the performance of a model, and how are they applied? Can you provide an example of a situation where you would use each type of test? (7)

****