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

SAINTGITS COLLEGE OF ENGINEERING (AUTONOMOUS)

(AFFILIATED TO APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY, THIRUVANANTHAPURAM)

SIXTH SEMESTER B.TECH DEGREE EXAMINATION (S), AUGUST 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. Define hypothesis testing. What are its uses?
3. Is variable selection necessary in data analytics? Justify your answer.
4. What are the different data sources?
5. Compare supervised and unsupervised learning.
6. Explain Linear regression.
7. What are the key roles for the new big data ecosystem?
8. What is Hadoop Distributed File System (HDFS)?
9. What is the use of generic function in R programming?
10. How dirty data can be detected in the data exploration phase?

PART B

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

MODULE I

11. a) With necessary examples explain Measure of central tendency and measure of dispersion. (12)

    b) What is the use of Grouped data? (2)

OR

12. a) What is the need of probability distribution in data analytics? Explain different types of probability distributions. (12)

    b) What is Inductive statistics? (2)

MODULE II

13. a) With neat sketches and necessary examples, explain Data analytics Lifecycle. (10)

    b) How missing values are handled in data analytics. (4)

OR

14. a) Given the following data, use PCA to reduce the dimension from 2 to 1

Feature	Example 1	Example 2	Example 3	Example 4
x	4	8	13	7
y	11	4	5	14

(12)

b) Differentiate between predictive and descriptive analytics. (2)

MODULE III

15. a) Explain Naive Bayes algorithm. Use the following dataset to find *"if weather is sunny, then the player should play or not".*

Outlook		Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

(12)

b) What is the use of KNN algorithm? (2)

OR

16. a) A database has 9 transactions, the list of items purchased by a customer. Find all the frequent items using Apriori algorithm, where the minimum_support_count = 2. (12)

TID	List_of_items_IDs
T100	I_1,I_2,I_5
T200	I_2,I_4
T300	I_2,I_3
T400	I_1,I_2,I_4
T500	I_1,I_3
T600	I_2,I_3
T700	I_1,I_3
T800	I_1,I_2, I_3,I_5
T900	I_1,I_2,I_3

b) Differentiate between classification and clustering. (2)

MODULE IV

17. a) With a neat sketch explain current analytical architecture. (8)

b) Explain developing and executing a Hadoop MapReduce Program. (6)

OR

18. a) Explain credit risk modelling with suitable examples. (10)

b) Explain the Hadoop Ecosystem. (4)

MODULE V

19. a) What is the need of R programming? With necessary examples explain different attribute and Data Types in R programming. (12)

b) What is the difference between type I error and type II error? (2)

OR

20. a) Explain different Statistical Methods for Evaluation. (12)

b) What you meant by Exploratory Data Analysis? (2)

****