Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING
END TERM EXAMINATION - JAN 2024

Semester : Semester VII - 2020
Course Code : CSE3133
Course Name : Predictive Analytics for Big Data
Program : B.Tech. Computer Science and Engineering

Date : 03-JAN-2024
Time : 9:30AM - 12:30 PM
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.
(iv) Do not write any information on the question paper other than Roll Number.

PART A

ANSWER ALL THE QUESTIONS **5 X 2M = 10M**

1. Why do industries use big data? (CO1) [Knowledge]
2. What are the six steps of data sourcing? (CO2) [Knowledge]
3. Name the Apache Hadoop key components. (CO4) [Knowledge]
4. Find out the RDD operations. (CO4) [Knowledge]
5. Label the interfaces of Spark SQL and its interaction with Spark. (CO4) [Knowledge]

PART B

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

6. Compare the structured, semi-structured, and unstructured data. (CO1) [Comprehension]

7. Assume, in a dataset, we have age as a continuous variable and good or bad loan as a categorical target variable. We might be interested in finding the logical separations needed to create bins for different age groups. First, we create a larger number of age groups and then calculate the WOE for each group. If there is a monotonic trend of WOE values (either descending or ascending), then we can confirm that our bins have a general trend. If it's not monotonic, then we need to compress the bins to form new groups and show the WOE values and information values.

Age Group	Good (1)	Bad (0)
18-35	2000	400
36-55	3402	201
56-70	1900	92

(CO2) [Comprehension]

8. Find a linear regression equation for the following two sets of data:

X	2	4	6	8
Y	3	7	5	10

(CO3) [Comprehension]

9. Demonstrate job scheduling with its various types

(CO4) [Comprehension]

10. Explain discretized stream processing with an example.

(CO4) [Comprehension]

PART C

ANSWER ALL THE QUESTIONS **2 X 20M = 40M**

11. Suppose that the data mining task is to cluster points into three clusters. Where the points are A1 (3,11) A2 (3,6).,A3(9,5), B1(6,9), B2(8,6). B3 (7,5), C1 (2,3), and C2 (5,10). Solve the distance function, which is Euclidean distance. Initially, we assign A1, B1, and C1 as the centres of each cluster, respectively.

(CO4) [Application]

12. Solve the logical computations with perceptrons and create logical gates with the AND gate. The data is given as W1 =1.2, W2 = 0.6, threshold = 1, and learning rate n = 0.5.

(CO3) [Application]