Roll No.											

PRESIDENCY UNIVERSITY

BENGALURU

End Term Examinations –December-2025

Date: 15-12-2025 **Time:** 1.00pm to 04.00pm

School: SOCSE	**Program:** BD	
Course Code : CBD3406	**Course Name:** Introduction to Data Science and Big Data	
Semester: V	**Max Marks**: 100	**Weightage**: 50%

CO – Levels	**CO1**	**CO2**	**CO3**	**CO4**	**CO5**
Marks	**24**	**24**	**24**	**28**	

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

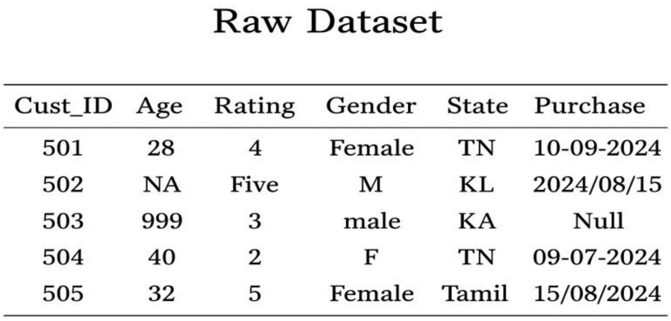
Part A

Answer ALL the Questions. Each question carries 2marks. 10Q x2M=20M

1.	Define data science	2 Marks	L1	CO1
2.	State the meaning of 'Velocity' in the 5Vs of Big Data.	2 Marks	L1	CO1
3.	Define standard deviation.	2 Marks	L2	CO2
4.	Define outliers in a dataset.	2 Marks	L2	CO2
5.	List any two commonly used distance metrics in supervised learning.	2 Marks	L2	CO3
6.	List two popular classification algorithms used in machine learning.	2 Marks	L2	CO3
7.	Illustrate the roles of the Name node	2 Marks	L2	CO4
8.	List any two sources of semi-structured data.	2 Marks	L2	CO4
9.	Define Spark DataFrame	2 Marks	L1	CO4
10.	Illustrate the difference between SQL and NoSQL databases.	2 Marks	L2	CO4

PART-B

Answer all the Questions. **Total Marks 80M**

<table>
<tr><td>11.
12.</td><td>Outline the key responsibilities in data science team
Or
Summarize the various stages of the Data Science Workflow</td><td>10 Marks</td><td>L1</td><td>CO1</td></tr>
<tr><td>13.
14.</td><td>Outline the architecture and components of Hadoop Distributed File System (HDFS) with a neat diagram
Or
Summarize 5 VS with real world example</td><td>10 Marks</td><td>L1</td><td>CO1</td></tr>
<tr><td>15.
16.</td><td>Identify data quality issues in the following dataset and write the cleaning steps to correct them.

Raw Dataset
<table>
<tr><th>Cust_ID</th><th>Age</th><th>Rating</th><th>Gender</th><th>State</th><th>Purchase</th></tr>
<tr><td>501</td><td>28</td><td>4</td><td>Female</td><td>TN</td><td>10-09-2024</td></tr>
<tr><td>502</td><td>NA</td><td>Five</td><td>M</td><td>KL</td><td>2024/08/15</td></tr>
<tr><td>503</td><td>999</td><td>3</td><td>male</td><td>KA</td><td>Null</td></tr>
<tr><td>504</td><td>40</td><td>2</td><td>F</td><td>TN</td><td>09-07-2024</td></tr>
<tr><td>505</td><td>32</td><td>5</td><td>Female</td><td>Tamil</td><td>15/08/2024</td></tr>
</table>
Or
Explain its purpose of descriptive statistics</td><td>10 Marks</td><td>L2</td><td>CO2</td></tr>
<tr><td>17.</td><td>Illustrate and show the image of Heat map for the following data
<table>
<tr><th>Age</th><th>Avg_Session _Time</th><th>Total_Purch ases</th><th>Clicks_Per_S ession</th><th>Customer _Rating</th></tr>
<tr><td>58</td><td>19.11</td><td>13</td><td>7.2</td><td>4</td></tr>
<tr><td>48</td><td>8.9</td><td>6</td><td>14.6</td><td>5</td></tr>
<tr><td>34</td><td>16.04</td><td>16</td><td>7.7</td><td>5</td></tr>
<tr><td>27</td><td>5.2</td><td>5</td><td>9</td><td>3</td></tr>
<tr><td>40</td><td>8.36</td><td>5</td><td>12.4</td><td>4</td></tr>
</table></td><td>10 Marks</td><td>L2</td><td>CO2</td></tr>
</table>

18.	**Or** Explain with chart of a histogram to show frequency distribution of sample student marks data			
19. 20.	Apply Linear Regression using a simple dataset and explain the concept of correlation between dependent and independent variables. **Or** Apply Logistic regression to sample dataset	10 Marks	L3	CO3
21. 22.	Apply True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) with example **Or** Apply a K-NN classifier with a small data set (10 Marks)	10 Marks	L3	CO3
23. 24.	Describe YARN'S primary role in managing big data workloads within the Hadoop ecosystem **Or** Describe the working of the Map phase and Reduce phase in MapReduce with word count example	10 Marks	L2	CO4
25. 26	Illustrate about Apache S**park** **Or** Explain the architecture and advantages of using Spark SQL in data science	10 Marks	L2	CO4