



Name:

Enrolment No:

UPES

End Semester Examination, December 2023

Course: Introduction to Data Science

Program: Master of Computer Application

Course Code: CSDS 7001

Instructions: Attempt all the questions. All questions are compulsory.

Semester: III

Time : 03 hrs.

Max. Marks: 100

SECTION A
(5Qx4M=20Marks)

S. No.		Marks	CO
Q 1	Define the data analytics and list the types of data analytics.	4	CO1
Q 2	Explain the data challenges in data science.	4	CO2
Q 3	Explain discretization and binarization for numerical data.	4	CO3
Q 4	Explain the feature learning for machine learning.	4	CO4
Q 5	Define cluster analysis. Explain it with examples.	4	CO5

SECTION B
(4Qx10M= 40 Marks)

Q 6	The data on number of patients attending a hospital in a month are given below:						10	CO2			
	No. of patients	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50			50 - 60		
	No. of days attending the hospital	2	6	9	7	4			2		
Determine the average number of patients attending the hospital in a day.											
Q 7	A random variable X has the following probability distribution:								10	CO2	
	$X = x$	0	1	2	3	4	5	6			7
	$P(X = x)$	0	k	$2k$	$2k$	$3k$	k^2	$2k^2$			$7k^2 + k$
Determine the value of (i) k , (ii) mean, and (iii) standard deviation.											
Q 8	Define the partition methods. Explain the K- means clustering algorithm with an example.						10	CO5			

Q 9	<p>Draw the histogram and frequency curve for the following data:</p> <table border="1" data-bbox="233 258 1291 506"> <tr> <td>Profit range ('000 Rs.)</td> <td>10-14</td> <td>15-19</td> <td>20-24</td> <td>25-29</td> <td>30-34</td> <td>35-39</td> <td>40-44</td> <td>45-49</td> <td>50-54</td> </tr> <tr> <td>Number of companies</td> <td>3</td> <td>7</td> <td>18</td> <td>25</td> <td>20</td> <td>12</td> <td>6</td> <td>5</td> <td>2</td> </tr> </table> <p style="text-align: center;">OR</p> <p>120 students at the college were asked to opt for different work experience. The details of these options are as follows:</p> <table border="1" data-bbox="233 726 1282 884"> <tr> <td>Areas of work experience</td> <td>Photography</td> <td>Clay modelling</td> <td>Kitchen gardening</td> <td>Doll making</td> <td>Book binding</td> </tr> <tr> <td>No. of students</td> <td>6</td> <td>30</td> <td>48</td> <td>12</td> <td>24</td> </tr> </table> <p>Represent the above data through a Pie diagram.</p>	Profit range ('000 Rs.)	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	Number of companies	3	7	18	25	20	12	6	5	2	Areas of work experience	Photography	Clay modelling	Kitchen gardening	Doll making	Book binding	No. of students	6	30	48	12	24	10	CO3
Profit range ('000 Rs.)	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54																										
Number of companies	3	7	18	25	20	12	6	5	2																										
Areas of work experience	Photography	Clay modelling	Kitchen gardening	Doll making	Book binding																														
No. of students	6	30	48	12	24																														
SECTION-C (2Qx20M=40 Marks)																																			
Q 10A	Explain the following terms: <i>(i)</i> types of data and datasets, <i>(ii)</i> data quality, and <i>(iii)</i> data issues, and <i>(iv)</i> data models.	10	CO1																																
Q 10B	Define the data wrangling. Write a short note on data cleaning, data aggregation, and data sampling.	10	CO1																																
Q 11	Define Principal Component Analysis (PCA). Explain all the steps involved in PCA with appropriate example.	20	CO4																																
OR																																			
Q 11A	Define the Decision Tree Algorithm. List down the attribute selection measures used by the ID3 algorithm to construct a decision tree.	10																																	
Q 11B	Write the advantages and disadvantages of the decision trees.	10																																	