Total No. of Printed Pages: 02

PRN No.	

PAPER CODE	V313-232-ESE

December 2023 (ENDSEM) EXAM

TY (SEMESTER - I)

COURSE NAME: DATA SCIENCE AND MACHINE LEARNING **Branch: COMPUTER ENGINEERING** **COURSE CODE: CSUA31202**

(PATTERN 2020)

Time: [1Hr. 30 Min] **[Max. Marks: 40]**

() Instructions to candidates:

1) **Figures to the right indicate full marks.**
2) **Use of scientific calculator is allowed**
3) **Use suitable data wherever required**
4) **All questions are compulsory. Solve any one sub question from Question 3 and any two sub questions each from Questions 4,5 and 6 respectively.**

Q. No.	Question Description	Max. Marks	CO mapped	BT Level
Q.1	a) Explain in detail how the model building phase is built by team in data analytics life cycle?	[2]	1	Understand
Q2	a) Explain categorical data and numerical data with example	[2]	2	Understand
Q3.	a) Using k-medoid for the given dataset where the number of clusters will be 2.	[6]	3	Analyze

	x	y
0	5	6
1	4	5
2	4	6
3	6	7
4	7	8

I. Randomly select 2 medoids M1(4,6) and M2(6,7) and do cluster assignment.
II. Calculate cost.
III. Randomly select 2 medoids M1(4,5) and M2(6,7) and do cluster assignment.
IV. Calculate cost.

Analyze the two costs and decide whether the algorithm should keep running or stop.

	b)Examine the advantages of K medoids algorithm over K means algorithm.	[6]	3	Analyze

<table>
<tr><td>Q.4</td><td>a) Illustrate the concept of entropy and information gain? Explain how to calculate with suitable steps.

b) Executing a binary classification tree algorithm is a simple task. But how does tree split take place? Identify how does the tree determine which variable to break at the root node and which at its child nodes?

c) Demonstrate use of Decision tree and Naïve bayes classifier using real time application</td><td>[5]

[5]

[5]</td><td>4

4

4</td><td>Apply

Apply

Apply</td></tr>
<tr><td>Q.5</td><td>a) Calculate Accuracy, Precision, Recall and F1 Score for the following Confusion Matrix on Water Shortage in Schools; Also suggest which metric would not be a good evaluation parameter here and why?
<table><tr><th>The Confusion Matrix</th><th>Reality: 1</th><th>Reality: 0</th></tr><tr><td>Prediction: 1</td><td>75</td><td>5</td></tr><tr><td>Prediction: 0</td><td>5</td><td>15</td></tr></table>Find out Accuracy, Precision, Recall and F1 Score for the given problem.

b) During the treatment of cancer patients, the doctor needs to be very careful about which patients need to be given chemotherapy. Choose correct evaluation metric among following should be used to decide whom to give chemotherapy? Justify your answer.
Metrics: Precision, Recall, Accuracy, F1 score.

c) Choose suitable the performance metric of a recommendation system?</td><td>[5]

[5]

[5]</td><td>5

5

5</td><td>Apply

Apply

Apply</td></tr>
<tr><td>Q.6)</td><td>a) Evaluate the impact of outliers on numerical data visualizations. Using a line chart and a scatter plot as examples, discuss how outliers can affect the perception of trends and correlations.

b) Estimate the use of pivot tables in handling granularity in visual representation of data.

c) Relate the effects of poor data cleaning and wrangling on quality of visual representation.</td><td>[5]

[5]

[5]</td><td>6

6

6</td><td>Analyze

Analyze

Analyze</td></tr>
</table>