Total No. of Printed Pages: 02

PRN No.	

PAPER CODE	V313-232 (RE)

December 2023 (REEXAM)

TY (SEMESTER - I)

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

(PATTERN 2020)

Time: [2 Hrs] **[Max. Marks: 60]**

() 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 two sub questions each from each Question 1 ,2, 3,4,5,and 6 respectively**

Q. No.	Question Description	Max. Marks	CO mapped	BT Level
Q.1	a) Describe the difference between supervised and unsupervised learning and their relevance in data science projects.	[5]	1	Understand
	b) Discuss the difference between Business Intelligence (BI) and Data Science (DS) in the context of analytics	[5]	1	Understand
	c) Explain Data analytics life cycle with all phases of life cycle for any example	[5]	1	Understand
Q2	a) Determine the significance of discretization in machine learning.	[5]	2	Apply
	b) Illustrate the use of data reduction techniques in data preprocessing.	[5]	2	Apply
	c) Suppose you're working on a data science project of "Predictive Analysis for Student Placement in VIIT". How would you apply data cleaning and preprocessing methods to handle missing values and outliers in the Data Preparation phase?	[5]	2	Apply
Q3.	a) Examine the advantages of K medoids algorithm over K means algorithm	[5]	3	Analyze
	b) We have given a collection of 8 points. P1=[0.1,0.6] P2=[0.15,0.71],P3=[0.08,0.9],P4=[0.16,0.85],P5=[0.2,0.3] P6=[0.25,0.5] ,P7=[0.24,0.1] ,P8=[0.3,0.2]. Perform the k-mean clustering with initial centroids as m1=P1 =Cluster#1=C1 and m2=P8=cluster#2=C2. Answer the following. 1] Which cluster does P6 belong to? 2] What is the population of cluster around m2? 3] What is updated value of m1 and m2?	[5]	3	Analyze
	c) Apply k-medoid for the given dataset where the number of clusters will be 2.	[5]	3	Analyze

	x	y
0	5	6

<table>
<tr><td></td><td>
<table>
<tr><td>1</td><td>4</td><td>5</td></tr>
<tr><td>2</td><td>4</td><td>6</td></tr>
<tr><td>3</td><td>6</td><td>7</td></tr>
<tr><td>4</td><td>7</td><td>8</td></tr>
</table>
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
</td><td></td><td></td><td></td></tr>
<tr><td>Q.4</td><td>a) Illustrate naive Bayes Classifier algorithm</td><td>[5]</td><td>4</td><td>Apply</td></tr>
<tr><td></td><td>b) Illustrate how does the Binary tree classification determine which variable to break at the root node and which at its child nodes?</td><td>[5]</td><td>4</td><td>Apply</td></tr>
<tr><td></td><td>c)Demonstrate use of Decision tree and Naïve bayes classifier using real time application</td><td>[5]</td><td>4</td><td>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? Find out Accuracy, Precision, Recall and F1 Score for the given problem.
<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>
</td><td>[5]</td><td>5</td><td>Apply</td></tr>
<tr><td></td><td>b) During the treatment of cancer patients, the doctor needs to be very careful about which patients need to be given chemotherapy. Which evaluation metric among following should be used to decide whom to give chemotherapy? Justify your answer.
Metrics: Precision, Recall, Accuracy, F1 score.</td><td>[5]</td><td>5</td><td>Apply</td></tr>
<tr><td></td><td>c) In Spam email classifier, which of the two false predictions would you care about more:
i. Falsely classifying a non-spam email as spam
ii. Falsely claiming a spam email as non-span.
Identify the evaluation metric that can be used for evaluating this.</td><td>[5]</td><td>5</td><td>Apply</td></tr>
<tr><td>Q.6)</td><td>a) Compare and contrast the fundamental principles of simplicity and clarity in data visualization. How does each principle contribute to effective communication of information? Provide examples of visualizations that successfully apply each principle</td><td>[5]</td><td>6</td><td>Analyze</td></tr>
<tr><td></td><td>b) Evaluate the strengths and weaknesses of two different types of data visualizations (e.g., bar charts vs. pie charts) for representing categorical data</td><td>[5]</td><td>6</td><td>Analyze</td></tr>
<tr><td></td><td>c) Elaborate the effects of poor data cleaning and wrangling on quality of visual representation.</td><td>[5]</td><td>6</td><td>Analyze</td></tr>
</table>