Enrolment No: _____   Name of Student: _____

Department/ School: _____

END-TERM EXAMINATION, ODD SEMESTER DECEMBER 2024

COURSE CODE: CSET211	**MAX. DURATION 2 HRS**
COURSE NAME: Statistical Machine Learning	**TOTAL MARKS 40**
PROGRAM: BTECH	

Mapping of Questions to Course and Program Outcomes

Q.No.	A1	A2	A3	A4	B1	B2	B3	B4
CO	1	1	1	2	2	3	3	3
PO	1,3	1,3,12	1,3,12	1,3,12	1,3,12	1,3,12	1,3,12	1,3,12
BTL	1,2	1,2,3	2,3,4	3,4	2,3	1,2,3	2,3,4	1

GENERAL INSTRUCTIONS: -

1. Do not write anything on the question paper except name, enrolment number and department/school.

2. Carrying mobile phones, smartwatches and any other non-permissible materials in the examination hall is an act of UFM.

COURSE INSTRUCTIONS:

a) A scientific calculator is permissible in the examination hall.

b) All questions are compulsory

SECTION A

Max Marks:20

A1) Identify a suitable machine learning task, i.e., Classification, Regression, or Clustering, for the following applications. Explain why?   **[1+1+1=3 Marks]**

    a)   Gold Price prediction
    b)   Prediction of Heart Disease
    c)   Grouping of News

A2) The confusion matrix for the case study: whether the person is going to buy a car or not, is given below   **[1+1+1+1+1=5 Marks]**

2792

n=1000	Predicted: Buy	Predicted: Not Buy
Actual: Buy	1150	27
Actual: Not Buy	35	1580

Find the following performance measures from the given confusion matrix

    a) TP, TN, FP, FN
    b) Accuracy
    c) Precision
    d) Recall
    e) F1-Score

A3) A retail company wants to segment its customers based on their purchasing habits to enhance targeted marketing. Each customer is described by two attributes: Monthly Purchase Frequency (MPF) and Average Purchase Value (APV). Customers are categorized into two groups: Group A (frequent buyers) and Group B (occasional buyers). A new customer, represented by the data point P = (6, 6), needs to be classified into one of these groups using the K-Nearest Neighbors (KNN) algorithm with k = 3. **[6 Marks]**

MPF	APV	Target
3	4	Group A
5	5	Group A
2	6	Group A
8	8	Group B
7	7	Group B
9	5	Group B

A4) Consider the training data in the following table, where Season and Wind are Independent attributes and Visit is a dependent attribute. In the table, the Season attribute has values "R" (for Rainy) or "S" (for Sunny) or "O" (for Overcast), Wind has values "S" (for strong) or "W" (for weak), and Visit has two class labels "Yes" or "No". Identify the class label for the day (Season =R, Wind=S), according to naïve Bayesian classification **[6 Marks]**

Season	Wind	Visit
R	S	Yes
S	W	Yes
O	S	Yes
O	S	Yes
S	W	No
R	S	Yes
S	W	Yes
O	W	Yes
R	W	No
S	S	No
S	S	Yes
R	S	No
O	W	Yes
O	S	No

SECTION B

Max Marks:20

[2 Marks]

B1) a) Discuss K-mean clustering.

b) A school is analyzing the study habits of its students to provide personalized learning resources. The school has recorded the average daily study hours of 10 students over a month. They aim to group students into two clusters: light learners and dedicated learners. Use K Means clustering to group the given data {2, 4, 7, 10, 12, 3, 20, 30, 11, 35} into the two clusters. Assume cluster centroid as m1=2 and m2=4.

[6 Marks]

B2) Consider the following dataset that contains 3 features (Outlook, Temp, and Wind) and 1 Dependent variable (Play Cricket) that contains 2 output classes. Identify the best attribute to select the root node for the decision tree using Information gain

[6 Marks]

Outlook	Temp	Wind	Play Cricket
Sunny	Hot	Weak	No
Sunny	Hot	Strong	No
Rain	Cool	Weak	Yes
Sunny	Cool	Strong	Yes

B3) Explain the following terminologies

[1.5+1.5+1.5+1.5 = 6 Marks]

a) PCA
b) Underfitting vs Overfitting
c) SVM
d) Bagging Vs Boosting

-ALL THE BEST-