BENNETT UNIVERSITY
THE TIMES GROUP

Enrolment No: _____ Name of Student: _____

Department/ School: _____

SUPPLEMENTARY EXAMINATION, ODD SEMESTER FEBRUARY 2025

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

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	1,3,8	1,3,6	1,3	1,3,8	1,3,6	1,3,7	1,3,6
BTL	1,2	1,2,3	2,3,4	3,4	1,2	1,2,3	2,3,4	1,2

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) A retail company wants to segment its customers based on their purchasing behavior to offer personalized recommendations. They have a large dataset containing transaction histories but no predefined labels for customer groups. What type of learning is helpful in this scenario and why? Name a suitable algorithm and explain how it would work. **[3 Marks]**

A2) A bank has implemented a machine-learning model to detect fraudulent transactions. Assume fraudulent transactions as a positive test case, represented by 1. The model was tested on 10 transactions, and the actual and predicted labels are as follows:

Transaction	Actual Label	Predicted Label
1	1	1
2	0	0
3	1	0
4	0	1
5	1	1
6	0	0
7	1	1
8	0	0
9	1	0
10	0	0

Find the following performance measures from the given table **[2+1+1+1+1+1=7 Marks]**

a) TP, TN, FP, FN
b) Construct the Confusion Matrix
c) Accuracy
d) Precision
e) Recall
f) F1-Score

A3) A dataset contains the following five training points with two features (X1, X2) and their respective class labels (A or B):

Point	X1	X2	Class
P1	2	4	A
P2	4	6	A
P3	4	2	B
P4	6	4	B
P5	6	6	A

A test point P6 (X1 = 5, X2 = 5) needs to be classified using the KNN algorithm with K = 3. **[4+2=6 Marks]**

a) Compute the Euclidean distance between P6 and all training points.
b) Identify the three nearest neighbors and determine the class label of P6.

A4) Answer the following questions to demonstrate your understanding: **[1+1+1+1=4 Marks]**

a) Difference between Supervised and Unsupervised Learning
b) Difference between Clustering and Classification.
c) Difference between Underfitting and Overfitting.
d) Write short notes on the Logistic Regression Model.

BENNETT UNIVERSITY
THE TIMES GROUP

SECTION B

Max Marks:20

B1) What is a Support Vector Machine (SVM)? Explain the terminologies of SVM. **[2 Marks]**

B2) Apply K-means clustering to group the given data into the two clusters. Assume the initial cluster centroid as (185,72) and (170,56). **[6 Marks]**

Sample No	X	Y
1	185	72
2	170	56
3	168	60
4	179	68
5	182	72
6	188	77

B3) A company wants to predict whether a new customer will purchase a product based on past customer data. The following dataset is given with attributes (Age, Income, Purchase):

Customer	Age	Income	Purchase (Yes/No)
C1	Young	High	No
C2	Young	Medium	No
C3	Middle	High	Yes
C4	Senior	Medium	Yes
C5	Young	Low	Yes
C6	Middle	Low	No
C7	Senior	Medium	Yes
C8	Young	Medium	No

A new customer (Age = Middle, Income = Medium) needs to be classified using the Naïve Bayes classifier. **[6 Marks]**

B4) A company wants to predict whether a customer will buy a product based on the features Age and Income. The following dataset is provided:

Customer	Age	Income	Purchase (Yes/No)
C1	30	High	No
C2	35	Medium	Yes
C3	40	High	Yes
C4	50	Low	No
C5	45	Medium	Yes
C6	28	Low	No
C7	33	High	Yes
C8	25	Medium	No

Using the decision tree algorithm, calculate the information gain for the Age and Income attributes, and determine which attribute should be used as the root node for the decision tree. **[1+2+2+1=6 Marks]**

a) Calculate the entropy of the entire dataset.
b) Calculate the entropy for both Age and Income attributes.
c) Compute the information gain for both attributes.
d) Determine which attribute to choose as the root node based on the highest information gain.