



Roll No.....

End Semester Examination, December 2024
BSc (Computer Sc/Data Science)- Semester III
Fundamentals of Machine Learning
Course Code: ENSP209

Time: 2.5 hrs.

Max Marks: **50**

Note: **All Questions are compulsory.** Marks are indicated against each question.

PART-A

Attempt any 5 questions (2*5=10 marks)

Q.1 First question will consist of 7 small questions out of which student has to attempt 5 questions.

1) Unlike the conventional approach, which relies on hard-coded instructions, machine learning systems rely on __ to improve performance.	[CO-1] [L-1]
2) In the machine learning workflow, __ is the process of preparing raw data for model training.	[CO-1] [L-1]
3) In linear regression, the _____ method is used to find the line that minimizes the sum of squared differences between observed and predicted values.	[CO-2] [L-1]
4) _____ techniques like Lasso and Ridge are used in regression to prevent overfitting by penalizing large coefficients.	[CO-2] [L-1]
5) In multiple linear regression, an assumption is that there is no _____ among the independent variables.	[CO-2] [L-1]
6) In a decision tree, _____ occurs when the model learns patterns specific to the training data but does not generalize well to new data.	[CO-3] [L-1]
7) In Ensemble Learning, _____ is a method that combines models by training multiple instances on different subsets of the data.	[CO-4] [L-1]

PART-B

(Attempt all questions)

Q.2 Identify and discuss the various data quality issues in the following dataset. Suggest the solution against the identified issues [CO-1] [L-3] **10**

Dataset:

ID	Price	Bedrooms	SquareFeet	Location	BuiltYear
1	\$200,000	3	1500	New York, NY	2005

ID	Price	Bedrooms	SquareFeet	Location	BuiltYear
2	\$250,000	4	2000	San Francisco, CA	2010
3	\$NaN	3	1800	New York, NY	NaN
4	\$300,000	3	-	Boston, MA	2008
5	\$250000	4	2000	San Francisco, CA	2010
6	\$400,000	5	2500	San Francisco, CA	2012
7	\$150,000	2	1200	New York, NY	2002
8	\$99999999	4	3000	Boston, MA	2015
9	\$NaN	4	NaN	San Francisco, CA	2011
10	\$300,000	3	2000	Boston, MA	2008

OR

[CO-1] [L-3] **10**

You have the following dataset of customers with information about their age, income, and city. Perform data preprocessing by applying normalization, standardization, and one-hot encoding.

Dataset:

CustomerID	Age	Income	City
1	25	50000	New York
2	40	60000	San Francisco
3	35	55000	Los Angeles
4	50	65000	New York

1. Normalization: Apply Min-Max normalization to the Age and Income columns.
2. Standardization: Standardize the Age and Income columns.
3. One-Hot Encoding: Apply one-hot encoding to the City column

[CO-2] [L-3]

10

Q.3 A company wants to predict the sales revenue based on advertising expenditure. The following data shows the advertising expenditure (in \$1000s) and the corresponding sales revenue (in \$1000s) for different campaigns.

Advertising (\$1000s)	Sales Revenue (\$1000s)
10	25
15	30
20	40
25	45
30	50

1. Find the equation of the simple linear regression line, $y = b_0 + b_1 \cdot X$
2. Predict the sales revenue if the advertising expenditure is \$22,000.
3. Calculate the metrics like RMS, R square

[CO-2] [L-3] **10**

OR

A company wants to predict whether a customer will buy a product (1 for "Yes," 0 for

"No") based on their income level (in \$1000s). They have data on five customers as shown below:

CustomerID	Income (\$1000s)	Purchased (Y)
1	25	0
2	30	0
3	35	0
4	40	1
5	45	1

Using logistic regression, the model fitted is:

$$\hat{Y} = \frac{1}{1 + e^{-(b_0 + b_1 \times \text{Income})}}$$

where $b_0 = -10$ and $b_1 = 0.3$.

1. Calculate the probability of purchase for each customer.
2. Classify each customer based on a threshold of 0.5.
3. Calculate the accuracy of the model.

[CO-3] [L-3]

10

Q.4 A company wants to classify a new customer as either a "High-Spender" (Class 1) or "Low-Spender" (Class 0) based on their income and age.

The following data shows existing customers' income, age, and their spending class:

CustomerID	Income (\$1000s)	Age	Class (Spending)
1	40	30	0
2	50	25	0
3	60	28	1
4	70	35	1
5	80	32	1

The new customer has an income of \$65,000 and is 29 years old. Use the k-NN algorithm with k=3 to classify this new customer.

[CO- 3] [L-3] **10**

OR

A company wants to predict whether a customer will buy a new product based on two features: Age and Income. We have the following data:

Customer	Age	Income	Buys Product?
1	Young	High	No
2	Young	Low	No
3	Middle	High	Yes
4	Senior	High	Yes
5	Senior	Low	Yes
6	Senior	Low	No
7	Middle	Low	Yes
8	Young	Low	No
9	Young	High	Yes
10	Senior	High	Yes

Now, we want to predict whether a Young person with High income will buy the product or not, using the Naive Bayes algorithm.

10

Q.5. We have a dataset with four points, and we want to cluster them into two clusters (K=2) using the K-Means algorithm.
The points are:

Point	X	Y
A	2	3
B	5	8
C	1	2
D	6	9

[CO- 4] [L-2] **10****OR**

- A. What is Ensemble Learning, and how does it improve model performance compared to individual models?
- B. Explain the difference between **Bagging**, **Boosting**, and **Stacking** in ensemble learning.
- C. Describe the Random Forest algorithm and how it differs from a single Decision Tree.