Graphic Era
HILL UNIVERSITY
Established by an Act of the State Legislature of Uttarakhand (Adhiniyam Sankhya 12 of 2011)
University under section 2(f) of UGC Act, 1956

End Term (Odd) Semester Examination December 2024

Roll no. 2294038

Name of the Course and semester: B. Tech (CSE) 5th
Name of the Paper: Machine Learning
Paper Code: TCS-509
Time: 3-hour

Maximum Marks: 100

Note:

(i) All the questions are compulsory.
(ii) Answer any two sub questions from a, b and c in each main question.
(iii) Total marks for each question is 20 (twenty).
(iv) Each sub-question carries 10 marks.

Q1. (2X10=20 Marks) (CO1)

a. Consider the following dataset representing the Names ages of participants in a survey:

Name	Ram	Raj	Rai	Robin	Aman	Niki	Atul	Ajay	Ali	Bob	Tom	Tonny
Age	25	32	29	34	41	28	35	30	37	45	38	40

Tasks:

1. Calculate the Mean, Median, and Mode of the dataset.
2. Identify and comment on any outliers.

b. Define machine learning and describe its three main approaches: Supervised, Unsupervised, and Reinforcement Learning. Provide examples of real-world applications for each approach.

c. Compute the following for the below mentioned Dataset

Age	Frequency
0 – 5	5
6-10	7
11-15	4
16-20	6

1. Range of the data.
2. Average Deviation
3. Absolute Deviation
4. Squared Deviation
5. Standard Deviation

Q2. (2X10=20 Marks)(CO2)

a. Consider the following dataset with missing values

A	B	C	D
7.0	40.0	NaN	0.02
4.0	40.0	NaN	NaN
8.0	30.0	NaN	-0.53
NaN	20.0	5.0	-0.11
7.0	NaN	5.0	0.22

Write a python code to handle the missing values using the following techniques individually. Also display the content of dataset after each operation.

1. Drop missing values
2. Fill missing values with the mean
3. Forward fill
4. Interpolate missing values

b. Discuss the role of outliers in statistical data analysis. How can outliers affect measures such as mean and standard deviation, and what methods can be used to handle them?

c. Define Exploratory Data Analysis (EDA) and explain its importance in the data analysis process. Describe the key steps involved in performing EDA.

Q3. (2X10=20 Marks)(CO3)

a. Explain the following data types with an example of each.

1. Numerical Data
2. Discrete Data
3. Continuous Data
4. Categorical data

b. Consider the following dataset

X	Y
1	2
2	4
3	5
4	4
5	5

Use linear regression to determine the equation of the best-fitting line. Specifically, calculate:

1. The slope (m) of the line.
2. The y-intercept (b) of the line.

c. .Consider the following dataset

x_1	x_2	Y
6	5	1
46	11	0
14	14	1
46	6	0

and apply the logistic regression using gradient descent. The initial values of the weights are:

$w_0 = 1$
$w_1 = 1$
$w_2 = 1$

The learning rate (α) is 0.5
Perform the first iteration of gradient descent and calculate the updated values of the weights w_0, w_1, and w_2.

End Term (Odd) Semester Examination December 2024

Q4. (2X10=20 Marks) (CO4)

a. Consider the following dataset with five data points (A, B, C, D, E) in a 2D space

Data Point	X	Y
A	1	2
B	2	1
C	4	5
D	7	8
E	8	7

Perform the following tasks.

1. Draw the dendrogram based on the hierarchical clustering process.
2. Interpret the dendrogram and determine the number of clusters if the distance threshold is set to d = 5.

b. Consider the following dataset with six data points in a 2D space:

Data Point	X	Y
A	1.0	1.0
B	2.0	2.0
C	3.0	3.0
D	8.0	8.0
E	8.5	8.0
F	9.0	9.0

Using the DBSCAN algorithm with the following parameters:
$\varepsilon = 2.5$ (neighborhood radius)
Minimum points (MinPts) = 2

Perform the following tasks:
1. Identify core points, border points, and noise points.
2. Perform clustering and assign clusters to the points.

c. You are given the following 6 data points in a 2D space:

Data Point	X	Y
A	6	3
B	2	8
C	1	2
D	7	9
E	4	5
F	3	4

Perform k-means clustering with k = 2. Use the following initial cluster centroids: Centroid 1: (2, 3), Centroid 2: (6, 8) and recompute the centroids after the assignment.

End Term (Odd) Semester Examination December 2024

Q5. (2X10=20 Marks) (CO5, CO6)

a. Consider the following dataset generate a correlation matrix and determine the features having high correlation.

X	Y	Z
2	2	6
3	4	5
4	6	4
5	8	3
6	10	2

b. Define following.

1) k-Fold Cross-Validation
2) Precision
3) Recall
4) F1-Score
5) Accuracy

c. A classification model has the following confusion matrix for a test dataset:

	Predicted Positive	Predicted Negative
Actual Positive	50	10
Actual Negative	5	35

Calculate the model's accuracy, precision, recall, and F1-score.