UPES
End Semester Examination, December 2023

Course: Machine Learning
Program: B. Tech. (CSE-H+NH)-All Spec.
Course Code: CSAI3014P

Semester : 5th
Time : 03 hrs.
Max. Marks: 100

Instructions: Attempt all questions

SECTION A
(5Qx4M=20Marks)

S. No.		Marks	CO
Q 1	Suppose you have a dataset with two features, "Age" and "Income,". 	Person	Age (years)	Income (USD)	 A	25	50000 B	30	60000 C	35	75000 D	22	48000 Do the Z-score scaling for the above-mentioned features.	4	CO1
Q 2	What is exploratory data analysis? Why is it essential in machine learning?	4	CO1
Q 3	Suppose you have a binary classification problem where you are predicting whether an email is spam (1) or not spam (0). You have a dataset of four email examples, and your model produces the following predicted probabilities and actual labels. Email	Predicted Probability (P(y=1))	Actual Label (y) E1	0.9	1 E2	0.7	1 E3	0.4	0 E4	0.6	1 Calculate the average Binary Cross Entropy for this dataset.	4	CO2
Q 4	How multiclass classification is implemented in logistic regression. Use suitable example to explain it.	4	CO4
Q 5	Discuss the significance of Regularization.	4	CO4

SECTION B
(4Qx10M= 40 Marks)

Q 6	Explain gradient descent optimization using regression algorithm.	10	CO2

The data tables within Q1 and Q3:

Q1 dataset:

Person	Age (years)	Income (USD)
A	25	50000
B	30	60000
C	35	75000
D	22	48000

Q3 dataset:

Email	Predicted Probability (P(y=1))	Actual Label (y)
E1	0.9	1
E2	0.7	1
E3	0.4	0
E4	0.6	1

Q 7	You are working on a medical diagnosis model to detect a rare disease. The confusion matrix for your model's predictions is as follows: - True Positives (TP): 8 - True Negatives (TN): 120 - False Positives (FP): 4 - False Negatives (FN): 2 Calculate the following evaluation metrics based on this confusion matrix: a) Accuracy b) Precision c) Recall (Sensitivity) d) Specificity e) F1 score	10	CO2
Q 8	Define K-Nearest Neighbors (KNN) and explain how it is used for classification. Discuss the significance of the parameter 'k' in KNN.	10	CO3
Q 9	Discuss the concept of residual errors in the context of Gradient Boosting. How are these residuals used in subsequent iterations to refine the model's predictions? **OR** Explain the regularization techniques used in XG Boost to prevent overfitting. Discuss the significance of hyperparameters such as learning rate, maximum depth, and the number of trees.	10	CO4
SECTION-C **(2Qx20M=40 Marks)**			

Q 10	Explain the following terms in context of Support Vector Machine:
	   a) Soft and Hard Margin
	   b) Kernel
	   c) Hyper-plane
	   d) Marginal Distance
	   e) Support Vectors
	
	OR
	
	   a) Explain the basic principles of Principal Component Analysis (PCA) and how it is used for dimensionality reduction. Discuss its applications in feature extraction and visualization.
	   b) Compare and contrast K-means clustering with hierarchical clustering. Provide a step-by-step explanation of the K-means clustering algorithm, including how the initial centroids are chosen and how clusters are updated in each iteration.
		20	CO4

Q 11	Draw the step-by-step decision tree using the concept of entropy and information gain for the data set below. Show the intermediate	20	CO3

calculations in detail while drawing the decision tree.

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No