UPES

End Semester Examination, May 2023

Course: Machine Learning   Semester: IV
Program: B.Tech(CSE/AIML)   Time : 03 hrs.
Course Code:CSAI 2001   Max. Marks: 100

Instructions: All questions are compulsory

SECTION A
(5Qx4M=20Marks)

S. No.		Marks	CO
Q 1	Compare linear and logistic regression with suitable examples.	4	CO1
Q 2	Why data pre-processing is important? List some of the data preprocessing methods.	4	CO1
Q 3	Which machine learning model is a more interpretable decision tree or Artificial Neural Network? Justify your answer.	4	CO2
Q 4	Differentiate between Bagging and Boosting techniques. Why ensemble learning is used?	4	CO2
Q 5	Refer to the following diagrams. How SVM will classify such types of cases. How maximum margin can be achieved in SVM?	4	CO3
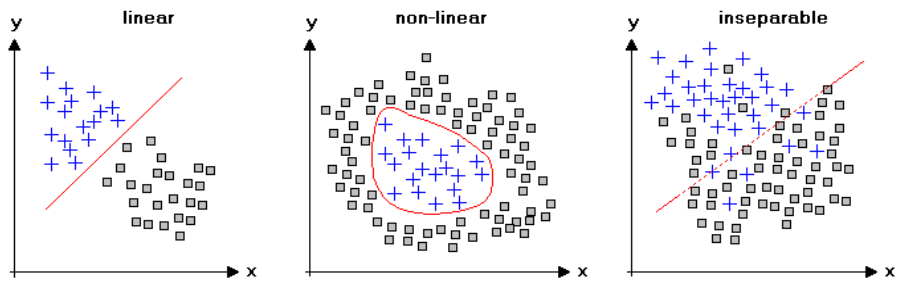
SECTION B
(4Qx10M= 40 Marks)

Q 6.	Classification and prediction methods can be compared and evaluated based on the following criteria : • Accuracy • Speed • Robustness • Scalability • Interpretability Discuss each criterion mentioned above in brief with a suitable example.	10	CO3

Q 7.	The following table shows the midterm and final exam grades obtained for the Machine Learning course. Use the method of least squares to find an equation for the prediction of a student's final exam grades based on the student's mid-term grade in the course. Also, predict the final exam grade of a student who received an 86 on the mid term.	10	CO2

x	72	50	81	74	94	86	59	83	65	33	88	81
y	84	63	77	78	90	75	49	79	77	52	74	90

(Assume x as the Midterm exam and y as the Final exam)

Q 8.	Write an algorithm for k-nearest neighbor classification given k and n, the number of attributes describing each tuple. List its merits and demerits.	10	CO4
Q 9.	Suppose a root node is created using training data D of size 14, where there are nine tuples belonging to the class **Buy_Computer= Yes** and five tuples belonging to **Buys_Computer= No**. Compute the Gini index to verify the purity of the node.Also discuss the following in the context of a decision tree: **Entropy** **Gini index** **Information gain.** **OR** Compare Clusteing and Classification in brief with appropriate examples. Given two objects represented by tuples (22,1,42,10) and (20,0,36,8): - Compute the Euclidean distance between the two objects - Compute the Manhatten distance between the two objects.	10	CO5

<div align="center">

SECTION-C
(2Qx20M=40 Marks)

</div>

Q 10		20	CO4

input the data

create scatter plot

linear relationship ? N → try to linearize ? Y → apply transformation

Y (from linear relationship)

calculate regression

N (from try to linearize) → try other than straight line model

check residuals

Referring to the above block diagram was used to design a possible linear regression model and compute the model accuracy by checking residuals. In this context summarize the all steps starting from the significance of scatter plot, transformation, other possible models, computing best-fit line, and all possible error checks (in linear regression) such as
- Mean absolute error
- Mean squared error
with their mathematical formulas.

Q 11	Design an artificial neural network model (ANN) where the input layer has four nodes, one hidden layer has three nodes and the output layer has one node. Also include activation function, bias, and weights at the appropriate place of your model. Draw a sketch of a biological neuron and compare it with the ANN model. In the context of your ANN model discuss the following points: • Types of activation function • Bias • Weights • Feedforward neural network. How ANN can be used for a classification problem? Suppose any ANN model has produced the following classification results : _(table below)_ Design a confusion matrix and compute the accuracy, precision, recall, and F-score of the above ANN model output. **OR** Clustering has been popularly recognized as an important data mining task with broad applications. Give one application example for each of the following cases : • An application that takes clustering as a major data mining function • An application that takes clustering as a preprocessing tool for data preparation for other data mining tasks	**20**	**CO5**

Actual	1	1	0	0	1	0
Predicted	0	1	1	0	0	1