INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Mid-Autumn Semester Examination 2024-25

Date of Examination: 20/09/2024 **Session (FN/AN): FN** **Duration: 2 hrs.** **Full Marks: 60**

Subject No.: ES60011 **Subject: Application of Machine Learning in Biological Systems**

Department/Center/School: School of Energy Science and Engineering

Specific charts, graph paper, log book etc., required

Special Instructions (if any): *(1) Answer all questions, (2) In case of reasonable doubt make practical assumptions and write that on your answer script, (3) The parts of each question must be answered together, (4) Calculator is allowed*

---

1. Mention whether the following statements are true or false.
   a. Missing values in a dataset can always be ignored without affecting the efficiency of a machine learning model.
   b. Random Forest can be used for classification and regression tasks.
   c. Removing outliers from a dataset always improves the performance of a ML model.
   d. Underfitting occurs when a regression model is too simple and does not capture the underlying patterns of the data.
   e. In logistic regression, the target variable is binary, and the model predicts the probability of the target being in one of two classes.
   f. In regression analysis, multicollinearity refers to the presence of highly correlated independent variables in the model.
   g. The F1 score is the arithmetic mean of precision and recall.
   h. A deeper decision tree is less likely to overfit the data.
   i. Reinforcement learning involves an agent learning to make decisions by taking actions in an environment to maximize cumulative reward.
   j. K-means clustering is a supervised learning algorithm. (1*10)

2. (a) Handle this missing data problem in Table 2a with attribute mean belonging to each class.
(b) Given the dataset in Table 2b: Apply a moving average filter with a window size of 3 to smooth the data with 2 iterations. (3+3)

Table 2a

Product	Price	Quality
A	300	Medium
B	500	Medium
C	600	High
D	250	Low
E	125	Low
F	488	?
G	760	High
H	?	High
I	900	?
J	?	Low
K	150	?
L	234	Low
M	?	Medium

Table 2b

Humidity (%)
80
83
88
78
34
80
83
84
79

20 Sept 2024, 11:04 pm

a. Calculate the total gini impurity value for Proper sleep symptom for separating the patients with or without **obesity**.

b. Evaluate which symptom can be more fitted as a root node for creating the decision tree for **obesity**. (3+4)

Table 6.

Adequate Physical Activity	Proper sleep	Healthy Diet	Obesity
Yes	No	No	Yes
Yes	Yes	No	No
No	Yes	No	No
Yes	No	Yes	Yes
No	No	Yes	No
No	Yes	No	Yes
Yes	No	Yes	Yes
Yes	Yes	No	No

7. Table 7. represents the drug effectiveness corresponding to the provided dosage.

(a) Considering the threshold of the dosage > 10, calculate the sum of squared residuals (SSR) of the dosage.

(b) Considering the threshold of the dosage > 21, calculate the SSR of the dosage. (3+3)

Table 7.

Dosage	**Drug effectiveness**
2	0
6	8
9	22
12	34
15	43
18	67
20	77
22	83
26	90
28	96
29	100

8. Take null hypothesis among input – output variables. Do a chi^2 (χ^2) correlation analysis for Table 8. Comment on the null hypothesis validity. (Consider critical value for χ^2 is 5.991, p-value= 0.05) (5)

Table 8.

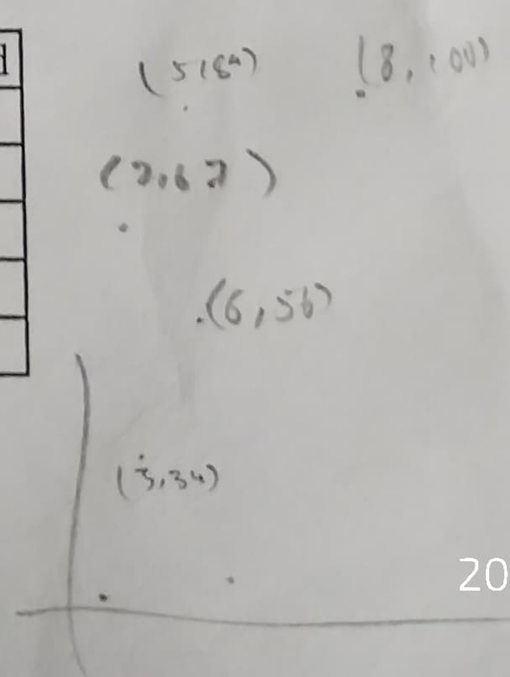
	Maths	Physics	Biology
Students passed	100	91	110
Students failed	23	31	26

9. (a) A classifier is evaluated on a test dataset of 100 instances. The classifier produced the following results:
Out of the 50 actual positive instances, the classifier correctly identified 40 as positive.
Out of the 50 actual negative instances, the classifier correctly identified 35 as negative.
Calculate its accuracy and recall value. (2+2)

(b) Find final centroids for the given dataset (Table 9.) applying K-means clustering while K=2. (4)

Table 9.

Patient	Days admitted	Dosage applied
Patient 1	2	67
Patient 2	6	56
Patient 3	8	100
Patient 4	3	34
Patient 5	5	89



3. You are tasked with training a multiple linear regression model using the delta rule. The model has two input variables x_1 and x_2 with corresponding weights w_1 and w_2 and target y. Initially, all weights are set to 1. The learning rate c is set to 1. Perform two iterations to the given dataset (Table 3.), updating the weights after each example using the delta rule. Evaluate the new weights after each iteration.

(3+3)

Table 3.

Input x_1	Input x_2	Target y
0.5	-0.2	1
1	1	0
-0.5	0.5	1

4. (a) Detect the outliers through IQR from the given dataset in Table 4a: [Try to detect maximum no. of outliers]

(b) Do a max-min normalisation for the given dataset in Table 4b.

Table 4a.

scores
12
56
45
67
72
99
60
75
160
250
55
71
34

Table 4b.

Patient	Dosage
Patient A	23
Patient B	45
Patient C	67
Patient D	89
Patient E	12

(3+3)

5. Consider a data set with 8 patients where 4 are diabetic and 4 are not diabetic. We have calculated the log(odds) of obesity for each candidate by fitting line X.

The log (odds) of each candidate data point for line X is as follows:

log(odds) of a = +1.6, log(odds) of b = -1.2, log(odds) of c = +2.4, log(odds) of d = -0.8, log(odds) of e = -1.6, log(odds) of f = +1, log(odds) of g = +0.6, log(odds) of h = -2.1

Next, we rotate the line (Y) and calculate the log(odds of obesity) for all the candidate data.

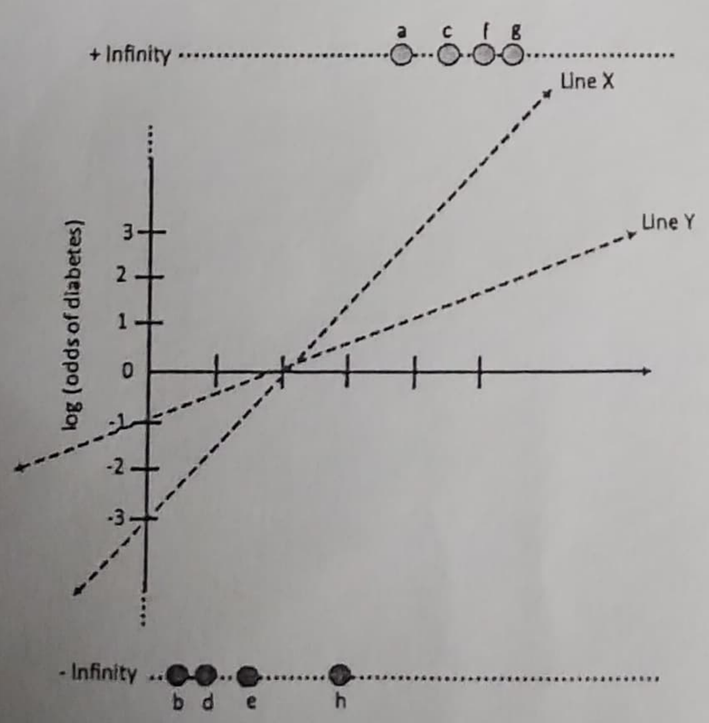
The log (odds) of each candidate data point for line Y is as follows:

log(odds) of a = +0.2, log(odds) of b = -1.5, log(odds) of c = +1.8, log(odds) of d = -1.9, log(odds) of e = -0.5, log(odds) of f = +0.3, log(odds) of g = +0.06, log(odds) of h = -1.2

a. Which line can be considered the best fitting line for the above scenario and why?

b. Calculate R^2 for line X.

(4+2)



6. Here, we have provided a dataset of 6 patients in Table 6. with **Obesity** and its symptoms as observed in these patients.