CSE 427 Fall 2025
Assignment 1 and practice problems
Submit answers to Questions 1, 3, 4, 5 and 6 as Assignment 1
Submission deadline is 25.11.2025

1. Consider the data table below and answer the following questions. Here the target column is RESULT.

STUDENT CATEGORY	DEPARTMENT	STUDY HOURS	RESULT
Junior	CSE	3	Average
Senior	EEE	6.5	Good
Senior	CSE	5	Bad
Junior	EEE	4	Average
Senior	EEE	8	Good
Junior	EEE	2.5	Bad
Junior	CSE	5	Good
Senior	CSE	7	Good
Senior	CSE	4	Good
Junior	EEE	4	Bad

a) Find which attribute would be a better candidate for the root node among STUDENT CATEGORY and DEPARTMENT.
b) Consider the better attribute (from a) as the actual root node of the classification tree. Now complete the 2nd level of the tree.

2.

Consider the data table below and answer the following questions. Here the target column is GPA.

STUDENT CATEGORY	DEPARTMENT	STUDY HOURS	GPA
Junior	CSE	3	3.1
Senior	EEE	6.5	3.6
Senior	CSE	5	2.7

Junior	EEE	4	3.3
Senior	EEE	8	3.75
Junior	EEE	2.5	2.89
Junior	CSE	5	3.69
Senior	CSE	7	3.80
Senior	CSE	4	3.55
Junior	EEE	4	2.9

a) Find which attribute would be a better candidate for the root node among STUDENT CATEGORY and DEPARTMENT.
b) Consider the better attribute (from a) as the actual root node of the regression tree. Now complete the next level only of the tree.

3.

Dataset: Whether a patient will recover from an illness within a week:

Temperature (°F)	Symptom	Medication	Gender	Recover
101	Cough	Antibiotic	Male	Yes
99	Fever	Paracetamol	Female	Yes
103	Cough	Antibiotic	Male	No
98	Headache	Paracetamol	Female	Yes
102	Fever	Antibiotic	Female	No
100	Headache	Paracetamol	Male	No
97	Fever	Paracetamol	Female	Yes
104	Cough	Antibiotic	Male	No
99	Headache	Paracetamol	Female	Yes
101	Fever	Paracetamol	Male	No

Find whether a patient having **Temperature = 99°F, Symptom = Cough, Medication = Antibiotic, and Gender = Female** will recover within a week or not, using the Naïve Bayes classifier.

4. Suppose, you are given the following patient dataset to predict Flu considering Fever, Cough, Oxygen Saturation Percentage, Smoker and Age as features. You can see that the last instance has some missing values.

Instance / Example No.	Fever	Cough	Oxygen Saturation Percentage	Smoker	Age	Flu
1	No	No	98	No	22	**No**
2	Yes	Yes	94	No	30	**Yes**
3	Yes	Yes	92	Yes	45	**Yes**
4	No	Yes	96	No	28	**No**
5	Yes	??	??	Yes	??	**Yes**

After running Random Forest on some estimators, you get the Proximity Matrix below.

	Ex1	**Ex2**	**Ex3**	**Ex4**	**Ex5**
Ex1	—	0.20	0.10	0.25	0.15
Ex2	0.20	—	0.35	0.15	0.35
Ex3	0.10	0.35	—	0.10	0.40
Ex4	0.25	0.15	0.10	—	0.10
Ex5	0.15	0.35	0.40	0.10	—

a) What was the initial guess for the missing values before applying Random Forest?
b) Using the proximity matrix, estimate the missing values.

5. Suppose, you are given the following binary classification dataset and you want to implement the AdaBoost Algorithm on it. Here, Device Color, Device Type and Usability Score are the features and Rating is the target variable.

ID	Device Color	Device Type	Usability Score	Rating
1	Red	Mobile	42	**Positive**
2	Blue	Desktop	35	**Negative**
3	Green	Mobile	35	**Positive**
4	Red	Desktop	28	**Negative**
5	Blue	Mobile	28	**Negative**
6	Green	Desktop	42	**Positive**
7	Red	Mobile	30	**Positive**
8	Blue	Desktop	30	**Negative**
9	Green	Mobile	42	**Negative**
10	Red	Desktop	35	**Positive**

a) What will be the initial weight for all the instances?
b) Using GINI Index to choose the best separator, find out the first stump. Show your work.
c) Using the stump from b), calculate the Amount of Say (α) for the stump.
d) Using the result of c), calculate the updated weights for correctly classified and incorrectly classified instances.

6. Define the cross entropy loss function for logistic regression model. Derive the weight update formula for weights from the derivative of the loss function.

7. Find the cross entropy loss for the following examples using a Z score given by Z = 0.5 X1 + 0.25 X2 + X3 + 0.75 X4 – 1.4. Assume you are using logistic regression for classification task. Now adjust the weights for a single iteration.

X_1	X_2	X_3	X_4	Y
0.56	0.60	0.38	0.50	Yes
0.90	0.80	0.13	1.00	No
1.00	0.60	0.25	0.33	Yes