Practice Problem 1

1. CO6 You have been given a dataset containing 8 rows and four features ("unique_id", "colour", "size", and "tail_length"). Each row represents a rat, and you are interested in finding the value of the label "has_disease". Your goal is to build a decision tree from the table below:

unique_id	colour	size	tail_length	has_disease
#1	black	large	5.6	NO
#28	white	large	2.2	YES
#3	black	small	3.8	YES
#34	black	small	4.2	YES
#26	black	large	1.2	NO
#11	white	small	1.4	NO
#32	black	small	2.3	YES
#13	white	large	3.5	NO

a. Among the columns "colour", and "size", which would you choose as the root node of your decision tree if you used Information Gain (IG) to construct the tree? **Construct** a decision tree using these two variables.

b. Why is "unique_id" a bad choice for root node? **Explain**.

c. If you want to use "tail_length" as a node while building the decision tree, what must you do beforehand? **Explain**.

d. Given two variables X = {outcome of an unbiased dice that can be rolled to obtain an integer value between 1 and 6 with equal probability} and Y = {outcome of an unbiased coin that can be tossed with an equal probability of heads and tails}, **Identify** which is larger: entropy of X or entropy of Y?

Practice Problem 2

2. CO5 a. In the table below, you are given a dataset containing 9 rows and 3 features X_1, X_2, X_3. Y is the label. Using naive bayes classifier, **Determine** the most likely value of Y if $X_1 = 1, X_2 = a, X_3 = q$. You don't need to use any kind of smoothing or normal distribution. Just derive the probabilities from frequencies.

X_1	X_2	X_3	Y
1	a	p	0
2	b	r	1
3	b	p	1
3	c	q	1
2	c	r	0
1	b	q	1
2	a	p	0
3	a	r	1
3	b	q	0

Practice Problem 3

3. CO5

a. Suppose X is a discrete random variable whose domain is exhaustive and mutually exclusive. Now the domain of X = {A, B, C}. Assume P(A) = 0.5 and P(B) = 0.3, then **determine** (i) P(C) and (ii) P(A U B).

b. Suppose two coins are tossed simultaneously. Assume Event A = the 1st coin coming up heads and Event B = the 2nd coin coming up tails. Now **determine** the value of P(A∩B).

c.

	A		A'	
	B	B'	B	B'
C	0.1	0.2	0.2	Y
C'	X	0.1	0.1	0.1

Using the given table answer the following questions:

(i) Assume the events A and B showcase absolute independence and P(B|A) = 0.5. Now **determine** the value of X

(ii) Using the ans obtained from (I), **determine** the value of Y

(iii) Using the ans obtained from (I) and (II), **determine** the value of P(A|B∩C)

Practice Problem 4

1. CO3, CO4

a) Consider the training data below

i. Considering 'Edible' as the class, **Compute** entropy for this dataset.

ii. **Compute** information gain for:

a) Color

b) Size

iii. **Compare** between *Color* and *Size*. Which one is the better feature? Why?

Color	*Size*	*Shape*	*Edible*
Yellow	Small	Round	Yes
Yellow	Small	Round	No
Green	Small	Irregular	Yes
Green	Large	Irregular	No
Yellow	Large	Round	Yes
Yellow	Small	Round	Yes
Yellow	Small	Round	Yes
Yellow	Small	Round	Yes
Green	Small	Round	No
Yellow	Large	Round	No
Yellow	Large	Round	Yes
Yellow	Large	Round	No
Yellow	Large	Round	No
Yellow	Large	Round	No
Yellow	Small	Irregular	Yes
Yellow	Large	Irregular	Yes

Practice Problem 5

5. CO3, CO5

	Pandemic		No Pandemic		Total
	Online Class	Offline Class	Online Class	Offline Class	
Public Uni	**0.142**	**0.037**	**0.165**	**0.072**	
Private Uni	**0.103**	**0.146**	**0.217**	**0.118**	
Total					**1**

a) **Apply** Probabilistic Inference to answer the following questions (a-d) based on the given data,

i) Is Pandemic Conditionally Independent of Public Uni Given Online Class?

ii) Is Private Uni Independent of Online Class?

iii) Find the marginal probability of Offline Class.

iv) Find the value of P(Private Uni ^ Offline Class | No Pandemic)

v) **Explain** why Naive Bayes is called Naive? How can it outperform bayes theorem?

vi) Explain why you can omit the denominator while comparing two probabilistic results and make decisions using Bayes theorem?

Practice Problem 6

1. CO4

Animal	Weight (Kg)	Color	Pet?
Dog	12	Black	No
Cat	8	Orange	Yes
Dog	17	White	No
Dog	13	Orange	Yes
Rat	4	White	Yes
Rat	5	White	Yes
Dog	18	Black	No
Cat	11	Orange	No
Cat	9	Black	Yes
Rat	6	White	Yes

An animal is considered heavy if it weighs more than 10kg. Now, answer the following questions:

a) Is dog conditionally independent of heavy if the color is black? **Show** full calculation.

b) Given a heavy weighted orange cat, is it more likely to be pet or not? **Apply** naïve bayes theorem to solve it. (No need to show learning phase)

Practice Problem 7

2. CO5

Alpha	Beta	Y
Yes	Yes	No
Yes	Yes	No
No	Yes	Yes
No	No	Yes
No	No	Yes
Yes	No	No
Yes	Yes	No
Yes	Yes	Yes
No	No	Yes
No	No	Yes

a) Assume this is a binary classification problem where Y is the output label and Alpha, Beta are the input features. Now if you were asked to **Apply** ID3, then find out which of the two given features would be more appropriate as the root node of the equivalent decision tree.

b) Suppose you are given 2 scenarios involving two coins, A and B. In the first scenario, you flip coin A five times. The observed outcomes of coin A are H, T, T, T, T where H = heads and T = tails. Now in the 2nd scenario, you flip coin B five times as well. The first 3 outcomes of coin B are H, T, and T. Now mathematically **Solve** what should be the outcomes of the 4th and the 5th flip for coin B such that the second scenario would showcase a higher amount of uncertainty than the first.

c) Suppose a 3rd feature called Gamma is added in the given table. This feature is a continuous variable. In this scenario, is the feature fit enough for the ID3 algorithm? If not, then **Explain** what kind of changes to the feature you propose.

Practice Problem 8

3. CO4 Covid-19 tests all over the world aren't 100% accurate. A patient is actually positive in 85% of the cases when the test comes out positive. And a person is actually positive in 10% of the cases when the test comes out negative.

a) Of all the people who tested for Covid-19, 70% of them actually had the disease. If 1000 people participated in the tests, **Calculate** the probability of a person's test results being positive.

b) Ignore all the information given in part a. In addition, you are given that, of all the people who tested for Covid-19, 70% of them came positive. If a random person is chosen who is actually a Covid-positive, which one is more likely? Did they come out positive or negative in the tests? **Explain** mathematically.

Practice Problem 9

1. CO5 Answer the following questions based on the given Joint probability matrix.

	Male	**Female**	**Total**
Football	0.24	0.15	0.39
Rugby	0.2	0.05	0.25
Other	0.1	0.26	0.36
	0.54	0.46	1

a. **Estimate** the probability of someone playing Rugby if they are male.
b. **Estimate** the probability of being female if someone plays Football.
c. If you pick a person randomly, **estimate** the probability of that person playing one of Football or other games.
d. **Infer** whether playing rugby depends on females.

2. CO4

Practice Problem 10

3. CO6

Sl	**X1**	**X2**	**X3**	**Y**
1	Group 1	Positive	Confirm	Yes
2	Group 1	Positive	Confirm	Yes
3	Group 2	Positive	Confirm	No
4	Group 2	Negative	Deny	No
5	Group 2	Negative	Deny	No
6	Group 1	Negative	Confirm	Yes
7	Group 1	Positive	Confirm	Yes
8	Group 1	Positive	Deny	No
9	Group 2	Negative	Deny	No
10	Group 2	Negative	Confirm	No

a. Assume "Y" is the output whose value depends on the input features "**X1**", "**X2**", and "**X3**". **Find** out the root node using ID3 decision tree algorithm.
b. Suppose two other continuous input features "**X4**" and "**X5**" were added to this dataset. In this scenario, would ID3 still be suitable for this classification task? Briefly **explain** your views.
c. A study was conducted among 15 participants (10 Male, 5 Female) to assess the chances of being a smoker based on gender. It was seen that among males, the smoker to non-smoker ratio was 70:30, and it was 20:80 for females. **Find** *Entropy(Smoking|Gender = male).*

Practice Problem 11

5. CO5 Throughout the whole semester, you've been trying to get hold of your CSE422 instructor. But you never seem to find him in his office. To take matters into your own hands, you decide to install a pressure sensor under his chair and a motion sensor inside his room. These sensors will give you information of the following form.

- The pressure sensor tells you if someone is sitting on the chair (C = 1) or not (C = 0).
- The motion sensor tells you if someone is moving inside the room (M = 1) or not (M = 0).

You want to use these sensor readings to predict whether your instructor is in the room (Y = 1) or not (Y = 0).

You've also done some detective work to obtain the following logs over ten days.

Day	1	2	3	4	5	6	7	8	9	10
C	1	0	0	1	0	1	0	1	0	1
M	0	1	0	0	0	0	0	0	1	1
Y	0	0	0	1	1	1	1	1	1	1

You have used these logs to train a Naive Bayes classifier that predicts Y. The parameters of the model are shown in the following conditional probability tables. Some parameters are missing but can be calculated using the existing parameters.

Y	$P(Y)$
0	q
1	$1-q$

C	Y	$P(C \mid Y)$
0	0	p_1
0	1	
1	0	
1	1	p_2

M	Y	$P(M \mid Y)$
0	0	
0	1	
1	0	p_3
1	1	p_4

Now solve the following problems.

a. Using the conditional probability tables above, **find** P (C = 1, M = 0, Y = 1) in terms of q, p_2, and p_4. Consider that M and C are conditionally independent given Y.

b. Using the conditional probability tables above, **find** P (C = 1, M = 0, Y = 0) in terms of q, p_1, and p_3. Consider that M and C are conditionally independent given Y.

c. Using the conditional probability tables above, **find** P (Y = 1 | C = 1, M = 0) in terms of q, p_1, p_2, p_3, and p_4. Your answers to (a) and (b) should help.

d. **Solve** the values of q, p_1, p_2, p_3, and p_4.

e. Using your classifier, **predict** whether or not your instructor is in his office if C = 1 and M = 0 on day 11. Again, your answers to the previous questions should help