HW2: Decision Trees

1. Suppose you want to learn a decision tree from a simple dataset comprising 8 samples, each with 3 binary attributes (X_1, X_2, X_3) and a binary class label (Y).

Instance	X_1	X_2	X_3	Y
1	0	0	0	0
2	0	0	1	0
3	0	1	0	1
4	0	1	1	1
5	1	0	1	1
6	1	0	1	1
7	1	1	0	0
8	1	1	0	0

a. Compute the entropy of the class label.

$$H(Y) = -\sum_{j=1}^{k} p_j \log_2 p_j$$

$$= -\big(P(Y=1)\big)\log_2(P(Y=1))$$

$$-\big(P(Y=0)\big)\, log_2\big(P(Y=0)\big)$$

$$= -\left(\frac{4}{8}\right)\log_2\left(\frac{4}{8}\right) - \left(\frac{4}{8}\right)\log_2\left(\frac{4}{8}\right) = 1$$

b. Calculate the information gain for each of the three attributes.

$$G(Y|X) = H(Y) - H(Y|X)$$

$H(Y) = 1$

$P(X_1 = 1) = \frac{4}{8}, P(X_1 = 0) = \frac{4}{8}$

$$H(Y|X = x_i) = -\sum_{j=1}^{k} P\big(Y = y_j \big| X = x_i\big)\, log_2 P\big(Y = y_j \big| X = x_i\big)$$

$$H(Y|X_1 = 1) = -P(Y = 1|X_1 = 1)\log_2 P(Y = 1|X_1 = 1)$$

$$- P(Y = 0|X_1 = 1)\log_2 P(Y = 0|X_1 = 1)$$

$$= -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) = 1$$

$$H(Y|X_1 = 0) = -P(Y = 1|X_1 = 0)\log_2 P(Y = 1|X_1 = 0)$$

$$- P(Y = 0|X_1 = 0)\log_2 P(Y = 0|X_1 = 0)$$

$$= -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) = 1$$

$$H(Y|X) = \sum_{j=1}^{k} P\big(X = x_j\big) H\big(Y|X = x_j\big)$$

$$H(Y|X_1) = P(X_1 = 1)H(Y|X_1 = 1) + P(X_1 = 0)H(Y|X_1 = 0)$$

$$= \left(\frac{4}{8}\right)(1) + \left(\frac{4}{8}\right)(1) = 1$$

$$G(Y|X_1) = 1 - 1 = 0$$

$P(X_2 = 1) = \frac{4}{8}, P(X_2 = 0) = \frac{4}{8}$

$$H(Y|X = x_i) = -\sum_{j=1}^{k} P\big(Y = y_j\big|X = x_i\big)\, log_2\, P\big(Y = y_j\big|X = x_i\big)$$

$$\begin{aligned} H(Y|X_2 = 1) &= -P(Y = 1|X_2 = 1) \log_2 P(Y = 1|X_2 = 1) \\ &\quad - P(Y = 0|X_2 = 1) \log_2 P(Y = 0|X_2 = 1) \\ &= -\left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right) - \left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right) = 1 \end{aligned}$$

$$\begin{aligned} H(Y|X_2 = 0) &= -P(Y = 1|X_2 = 0) \log_2 P(Y = 1|X_2 = 0) \\ &\quad - P(Y = 0|X_2 = 0) \log_2 P(Y = 0|X_2 = 0) \\ &= -\left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right) - \left(\frac{2}{4}\right) \log_2 \left(\frac{2}{4}\right) = 1 \end{aligned}$$

$$H(Y|X) = \sum_{j=1}^{k} P\big(X = x_j\big) H(Y|X = x_j)$$

$$H(Y|X_2) = P(X_2 = 1)H(Y|X_2 = 1) + P(X_2 = 0)H(Y|X_2 = 0) = \left(\frac{4}{8}\right)(1) + \left(\frac{4}{8}\right)(1) = 1$$

$$G(Y|X_2) = 1 - 1 = 0$$

$P(X_3 = 1) = \frac{4}{8}, P(X_3 = 0) = \frac{4}{8}$

$$H(Y|X = x_i) = -\sum_{j=1}^{k} P\big(Y = y_j\big|X = x_i\big)\, log_2\, P\big(Y = y_j\big|X = x_i\big)$$

$$\begin{aligned} H(Y|X_3 = 1) &= -P(Y = 1|X_3 = 1) \log_2 P(Y = 1|X_3 = 1) \\ &\quad - P(Y = 0|X_3 = 1) \log_2 P(Y = 0|X_3 = 1) \\ &= -\left(\frac{3}{4}\right) \log_2 \left(\frac{3}{4}\right) - \left(\frac{1}{4}\right) \log_2 \left(\frac{1}{4}\right) = 0.811 \end{aligned}$$

$$\begin{aligned} H(Y|X_3 = 0) &= -P(Y = 1|X_3 = 0) \log_2 P(Y = 1|X_3 = 0) \\ &\quad - P(Y = 0|X_3 = 0) \log_2 P(Y = 0|X_3 = 0) \\ &= -\left(\frac{1}{4}\right) \log_2 \left(\frac{1}{4}\right) - \left(\frac{3}{4}\right) \log_2 \left(\frac{3}{4}\right) = 0.811 \end{aligned}$$

$$H(Y|X) = \sum_{j=1}^{k} P\big(X = x_j\big) H(Y|X = x_j)$$

$$\begin{aligned} H(Y|X_3) &= P(X_3 = 1)H(Y|X_3 = 1) + P(X_3 = 0)H(Y|X_3 = 0) \\ &= \left(\frac{4}{8}\right)(0.811) + \left(\frac{4}{8}\right)(0.811) = 0.811 \end{aligned}$$

$$G(Y|X_3) = 1 - 0.811 = 0.189$$

c. Which attribute should be selected for the root of the decision tree? Why?

i. The attribute X_3 should be selected for the root of the decision tree because its information gain is the greatest out of the three attributes.

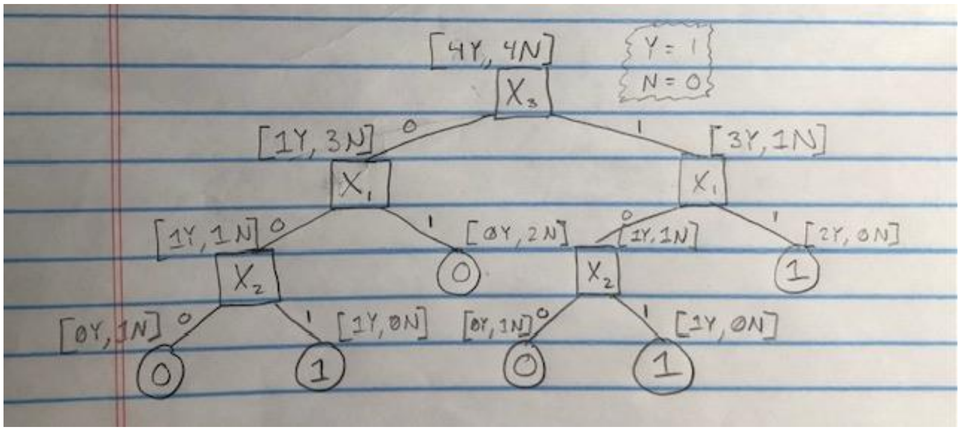
d. After the root node is selected, the entire tree can be learned by recursively splitting the data into subgroups, finding the next best attribute to split on, dividing the subgroup into smaller groups, and so on. How do you know when to stop growing the tree (i.e. what are the stopping criteria)?

i. The subset of training examples have the same output.

ii. The subset of training examples have the same values for all input attributes.

iii. There are no training examples for a specific leaf node.

e. By manually running the algorithm described above, draw the resulting tree.



f. Compute the training error.

i. $\frac{0}{8}$ training samples were incorrect; therefore, the training error is 0.

g. Now suppose you are presented new instances for which the class Y is unknown. Use your decision tree to predict the label of each instance listed below.

Instance	X_1	X_2	X_3	Y
9	1	1	1	?
10	1	0	0	?
11	0	1	1	?

i. Instance 9: $Y = 1$

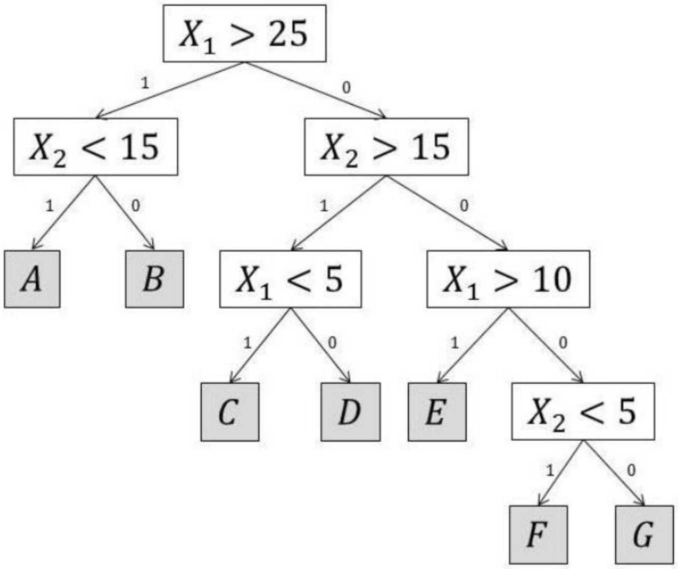
ii. Instance 10: $Y = 0$

iii. Instance 11: $Y = 1$

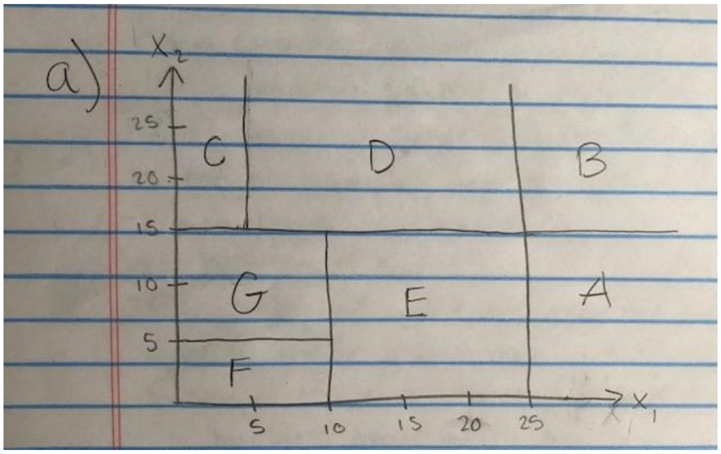
h. Do you have any basis on which to evaluate if the tree is overfitting? Why or why not? How might you combat overfitting in a decision tree?

i. I do not have any basis on which to evaluate if the tree is overfitting because I do not have the testing accuracy. I have the predicted testing labels, but I do not have what these labels actually are so I am not able to evaluate if the tree is overfitting. You can combat overfitting in a decision tree by either pruning the branches after building the full tree. You can also combat overfitting by setting aside part of the data for validation and leave the rest for training. You then generate several trees with different numbers of nodes, find the validation error for each tree, and use the tree that gives the lowest validation error.

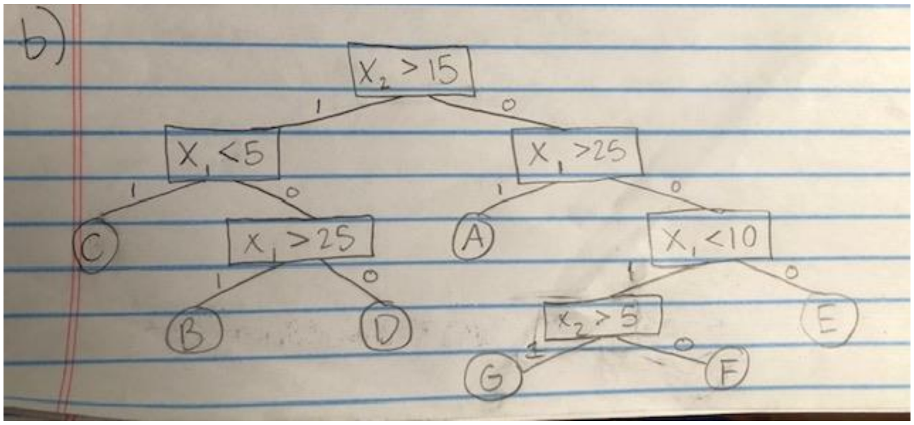
2. Consider the following decision tree.



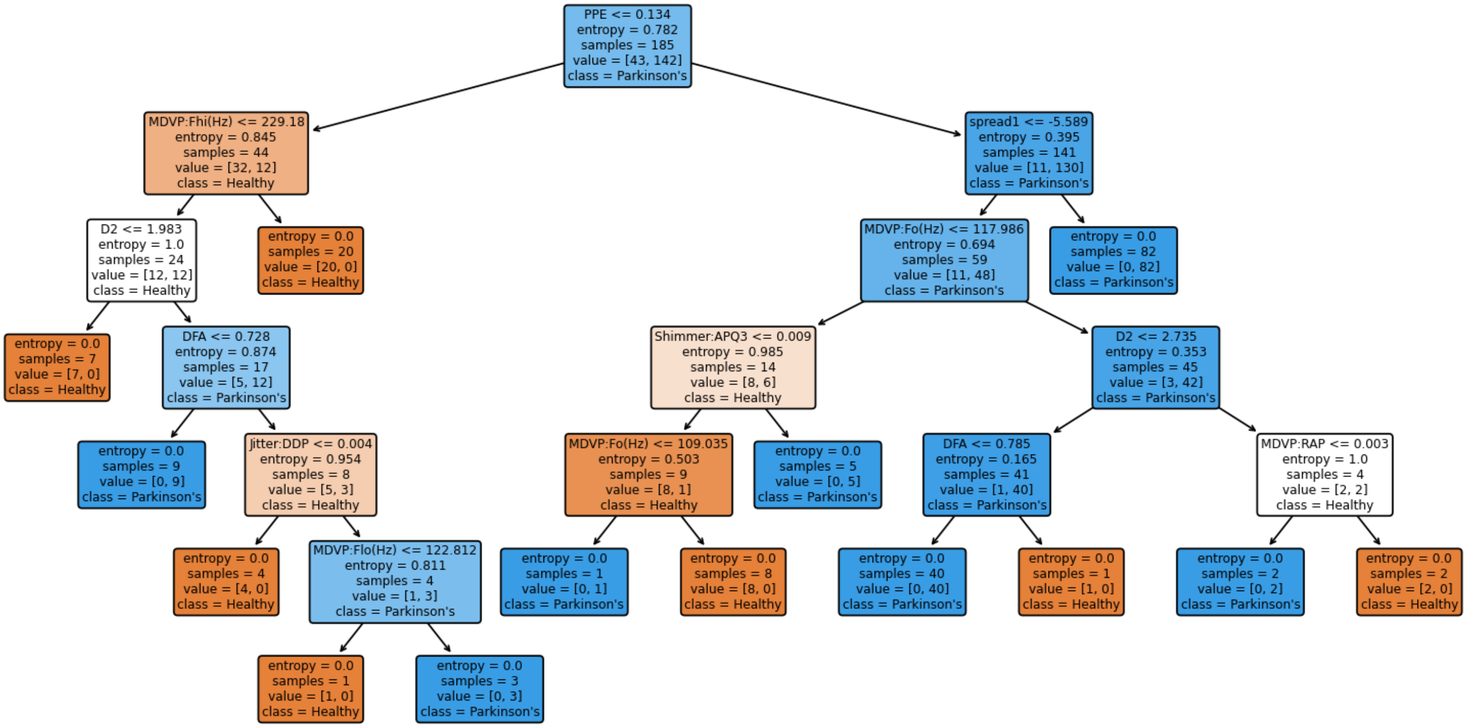
a. Draw the decision boundaries defined by this tree in 2D space ($X1$,$X2$). Each leaf is labeled with a letter. Write this letter in the corresponding region of the instance space.



b. Draw another decision tree that is syntactically different from the one shown above but defines the same decision boundaries.



3. Python Program
   a. What does the decision tree look like? Show a visualization of the full learned tree using `matplotlib` and `plot_tree`. In your visualization, include feature names, class names, filled nodes, and rounded nodes.



   b. What is the depth of the tree? Consider the depth to be the maximum number of <u>decision nodes</u> (do not count leaf nodes) that could be used to classify a single example.
      i. The depth of the tree is 6.

c. How many leaf nodes are there?

i. There are 14 leaf nodes.

d. What is the name of the attribute used to split the data at the root node of the tree?

i. The name of the attribute used to split the data at the root node is PPE.

e. Compute the information gain for the attribute listed above by hand.

$$G(Y|X = PPE) = H(Y) - H(Y|X = PPE)$$

$$H(Y) = H[1^-, 1^+] = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

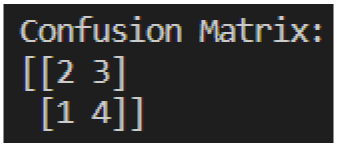
$$\begin{aligned} H(Y|X = PPE) \\ &= \left(\frac{44}{185}\right)(-\left(\frac{32}{44}\right)\log_2\left(\frac{32}{44}\right) \\ &- \left(\frac{12}{44}\right)\log_2\left(\frac{12}{44}\right)) + \left(\frac{141}{185}\right)(-\left(\frac{11}{141}\right)\log_2\left(\frac{11}{141}\right) \\ &- \left(\frac{130}{141}\right)\log_2\left(\frac{130}{141}\right)) = 0.502 \end{aligned}$$

$$G(Y|X = PPE) = 1 - 0.502 = 0.498$$

f. How many of the 22 attributes are used in the tree? What does this mean about the other attributes?

i. 10 of the 22 attributes are used in the tree. This means that the other attributes do not provide us with the highest information gain ever in the tree. This means that the attributes not included in the tree never provide us with more information than all the other attributes at any node in the tree.

g. Consider the confusion matrix for this tree on the test data. Does the tree more often misclassify people who are healthy or people who have Parkinson's? Practically speaking, which bias would you rather have (misclassifying healthy or sick people)? Explain your reasoning.



```
Confusion Matrix:
[[2 3]
 [1 4]]
```

i. The tree more often misclassifies people who have Parkinson's, misclassifying it 3 out of 7 instances, or 42.9% of the time, while the tree misclassifies people who are healthy 1 out of 3 instances, or 33% of the time. Practically speaking, I would rather be misclassified as sick because I would personally not want to be sick in the first place and it's more comfortable being healthy. I am opposed to being told I'm healthy when I'm actually sick because it's less comfortable being sick.

h. Does your decision tree overfit your data? How do you know?

i. Yes, the decision tree overfits the data because there is a training accuracy of 100% while the testing accuracy is 60%.