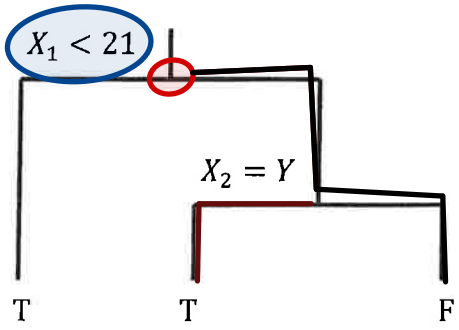
40.

You are given the following classification decision tree and data set:



i	X_1	X_2	Y
1	12	Y	T
2	23	N	F
3	4	Y	F
4	32	Y	F
5	22	N	T
6	30	Y	T
7	18	N	T

Determine the relationship between the classification error rate, the Gini index, and the cross-entropy, summed across all nodes.

A. cross-entropy > Gini index > classification error rate
B. cross-entropy > Gini index = classification error rate
C. classification error rate > Gini index > cross-entropy
D. Gini index > cross-entropy > classification error rate
E. The answer is not given by (A), (B), (C), or (D).

→: Caveat: They explicitly say: "summed across all nodes" which is different from computing a weighted average!

→: For $X_1 < 21$, we have observations $i = 1, 3, 7$
and they have $Y_1 = T,\ Y_3 = F,\ Y_7 = T$

⇒ From the tree, we know that the classification @ that node is T

⇒ The classification error is $\frac{1}{3}$ ✓

For $X_1 \geq 21$ and $X_2 = Y$, observations $i = 4, 6$
are in that terminal node w/ $Y_4 = F,\ Y_6 = T$

From the tree, the classification @ that node is T

⇒ The classification error is $\frac{1}{2}$ ✓

For $X_1 \geq 21$ and $X_2 = N$, observations $i = 2, 5$
are in that terminal node w/ $Y_2 = F,\ Y_5 = T$

In the tree, that node is F

⇒ The classification error is $\frac{1}{2}$

The overall classification error: $\frac{1}{3} + \frac{1}{2} + \frac{1}{2} = \frac{4}{3} = \frac{12}{9}$

At the 1st node, the Gini index: $\frac{1}{3}\left(\frac{2}{3}\right) + \frac{2}{3}\left(\frac{1}{3}\right) = \frac{4}{9}$

At the 2nd node, —11— : $2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}$

At the 3rd node, —11— : the same $\frac{1}{2}$

⇒ The Total Gini index: $\frac{4}{9} + \frac{1}{2} + \frac{1}{2} = \frac{13}{9}$

The cross entropy @ 1st node: $-\frac{1}{3}\ln\left(\frac{1}{3}\right) - \frac{2}{3}\ln\left(\frac{2}{3}\right)$

@ 2nd and 3rd nodes: $-\frac{1}{2}\ln\left(\frac{1}{2}\right) - \frac{1}{2}\ln\left(\frac{1}{2}\right)$

⇒ The Total Cross-Entropy: 2.022809

□