M339G : April 7th, 2025.

9. A classification tree is being constructed to predict if an insurance policy will lapse. A random sample of 100 policies contains 30 that lapsed. You are considering two splits:

Split 1: One node has 20 observations with 12 lapses and one node has 80 observations with 18 lapses.

Split 2: One node has 10 observations with 8 lapses and one node has 90 observations with 22 lapses.

The total Gini index after a split is the weighted average of the Gini index at each node, with the weights proportional to the number of observations in each node.
The total entropy after a split is the weighted average of the entropy at each node, with the weights proportional to the number of observations in each node.

Determine which of the following statements is/are true?

I. Split 1 is preferred based on the total Gini index.

II. Split 1 is preferred based on the total entropy.

III. Split 1 is preferred based on having fewer classification errors.

(A) I only

(B) II only

(C) III only

(D) I, II, and III

(E) The correct answer is not given by (A), (B), (C), or (D).

→: **Focus on the Gini index:**

Split 1. In the 1^{st} node, the majority are lapses

=> the entire 1^{st} node goes to lapses.

=> $\frac{12}{20}$ will be (properly) classified as lapses

and $\frac{8}{20}$ will not be

$$\Rightarrow GI = \frac{12}{20}\left(1-\frac{12}{20}\right)+\frac{8}{20}\left(1-\frac{8}{20}\right) = 2\cdot 0.6\cdot 0.4 = 0.48$$

In the 2^{nd} node, the majority are non-lapses

=> the entire 2^{nd} node goes to nonlapses.

=> $\frac{62}{80}$ will be (properly) classified as non-lapses

and $\frac{18}{80}$ will not be

$$\Rightarrow GI = 2\cdot\frac{62}{80}\cdot\frac{18}{80} = 0.34875$$

Altogether, for Split 1: $0.2\cdot 0.48 + 0.8\cdot 0.34875 = 0.375$

Split 2. $\boxed{1^{st}\text{ node}}$ $2\cdot 0.8\cdot 0.2 = 0.32$

$\boxed{2^{nd}\text{ node}}$ $2\cdot\frac{68}{90}\cdot\frac{22}{90} = 0.3693827$

Total: $0.1\cdot 0.32 + 0.9\cdot$ $= 0.3644$

$\boxed{\text{For the Gini Index, Split 2 is preferred.}}$

Focus on Cross-Entropy:

Split 1: $\boxed{1^{st}\text{ node}}$ $-\left(0.6\cdot\ln(0.6)+0.4\cdot\ln(0.4)\right) =$

$= \boxed{0.6730117}$

$\boxed{2^{nd}\text{ node}}$ $-\left(\frac{62}{80}\ln\left(\frac{62}{80}\right)+\frac{18}{80}\ln\left(\frac{18}{80}\right)\right) =$

$= 0.5331638$

Total: $0.2\cdot 0.673017 + 0.8\cdot 0.5331638 = 0.561334$

Split 2:

$$-\left(0.1\left(0.8\ln(0.8)+0.2\ln(0.2)\right)+\right.$$
$$\left.+0.9\left(\frac{68}{90}\ln\left(\frac{68}{90}\right)+\frac{22}{90}\ln\left(\frac{22}{90}\right)\right)\right)=$$
$$=\underline{0.5505744}$$

$\Rightarrow$ Split 2 is again preferred.

Focus on the total misclassification:

Split 1. $\frac{8+18}{100} = 0.26$

Split 2. $\frac{2+22}{100} = 0.24$

$\Rightarrow$ Split 2 is preferred

$\square$

Note on weighting in the classification error case.

$$E_m = 1-\max_k \hat{p}_{mk} = 1-\max_k \frac{n_{mk}}{n_m}$$

w/ n_m... total # of cases in the terminal node m

and

n_{mk}... total # of cases of class k in the terminal node m

$$E = \sum_m \frac{n_m}{n} E_m = \sum_m \frac{n_m}{n}\left(1-\max_k \frac{n_{mk}}{n_m}\right)$$
$$= \sum_m \frac{n_m - \max_k n_{mk}}{n}$$

$\square$