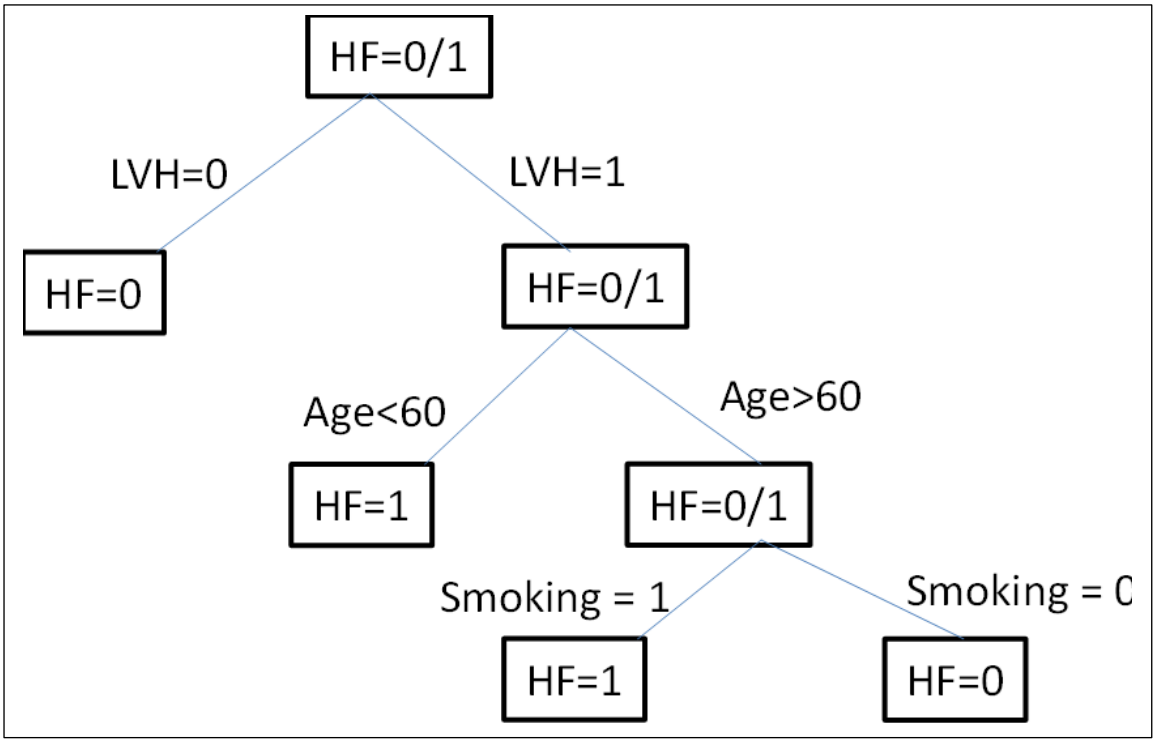
PRACTICE QUESTION FOR GINI IMPURITY

Below is a sample of a small study with 3 attributes from 4 patients. The outcome is whether the patient has Heart Failure (HF):

Left Ventricular Hypertrophy (LVH)	Age >60 years old (A)	Smoking (S)	Outcome: Heart Failure (HF)
0	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

(a) Using the dataset above, draw a decision tree with minbucket=1 which classifies HF as 0/1 using all the binary variables (LVH -> Age -> Smoking) in the stated sequence. [5 marks]

Solution:



Labels in the box should be:

1. Number and Proportion of HF = 0; and HF = 1;

(b) Is this tree optimal? Explain your intuition [3 marks]

Solution:

When split by Smoking, HF can already be differentiated completely.
Other reasonable explanations can be accepted

(c) Evaluate the GINI indexes and reduction in impurity given a split for LVH, Age and Smoking. Show your workings. [5 marks]

Solution:

Parent Node:

- 2 have HF = 1,
- 2 have HF = 0.

$p=2/4=0.5$

$\text{Gini(parent)}=1-p^2-(1-p)^2=1-(0.5)^2-(0.5)^2=1-0.25-0.25=0.5$

If Split by LVH

LVH = 0

- 1 have HF = 0
- 0 have HF = 1 –> Gini = 0

LVH =1

- 2 have HF = 1,
- 1 have HF = 0,
- Gini = 0,444

$\text{Gini(LVH)}=1/4X\ 0+3/4\ X\ 0.4444=0.3333$

$\Delta\text{Gini}=0.5-0.3333=0.1667$

If Split by Age

Smoking = 0

- 1 have HF = 0,
- 1 have HF = 1,
- Gini = 0.5

Smoking = 1

- 1 have HF = 0,
- 1 have HF = 1,
- Gini = 0

$\text{Gini(Smoking)}=2/4\cdot0.5+2/4\cdot0.5=0.5$

$\Delta\text{Gini}=0.5-0.5=0$

If Split by Smoking

Smoking = 0

- 2 have HF = 0,
- 0 have HF = 1,
- Gini = 0

Smoking = 1

- 0 have HF = 0,
- 2 have HF = 1,
- Gini = 0

$\text{Gini(Smoking)}=1/4\cdot0+2/4\cdot0=0$

$\Delta\text{Gini}=0.5-0=0.5$

Gini Formula:

Gini Index is defined as: $Gini(Y) = 1 - \sum_{j=1}^{n} p_j^2$ where p_j: Relative frequency of class *j* in *Y*

Given *Y* is split into two subsets (*D_1 and D_2)* by attribute *A,* the revised Gini Index is (weighted by the subgroup size):

$$Gini_A(Y) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

Change in Gini Index is given by:

$$\Delta Gini(A) = Gini(Y) - Gini_A(Y)$$