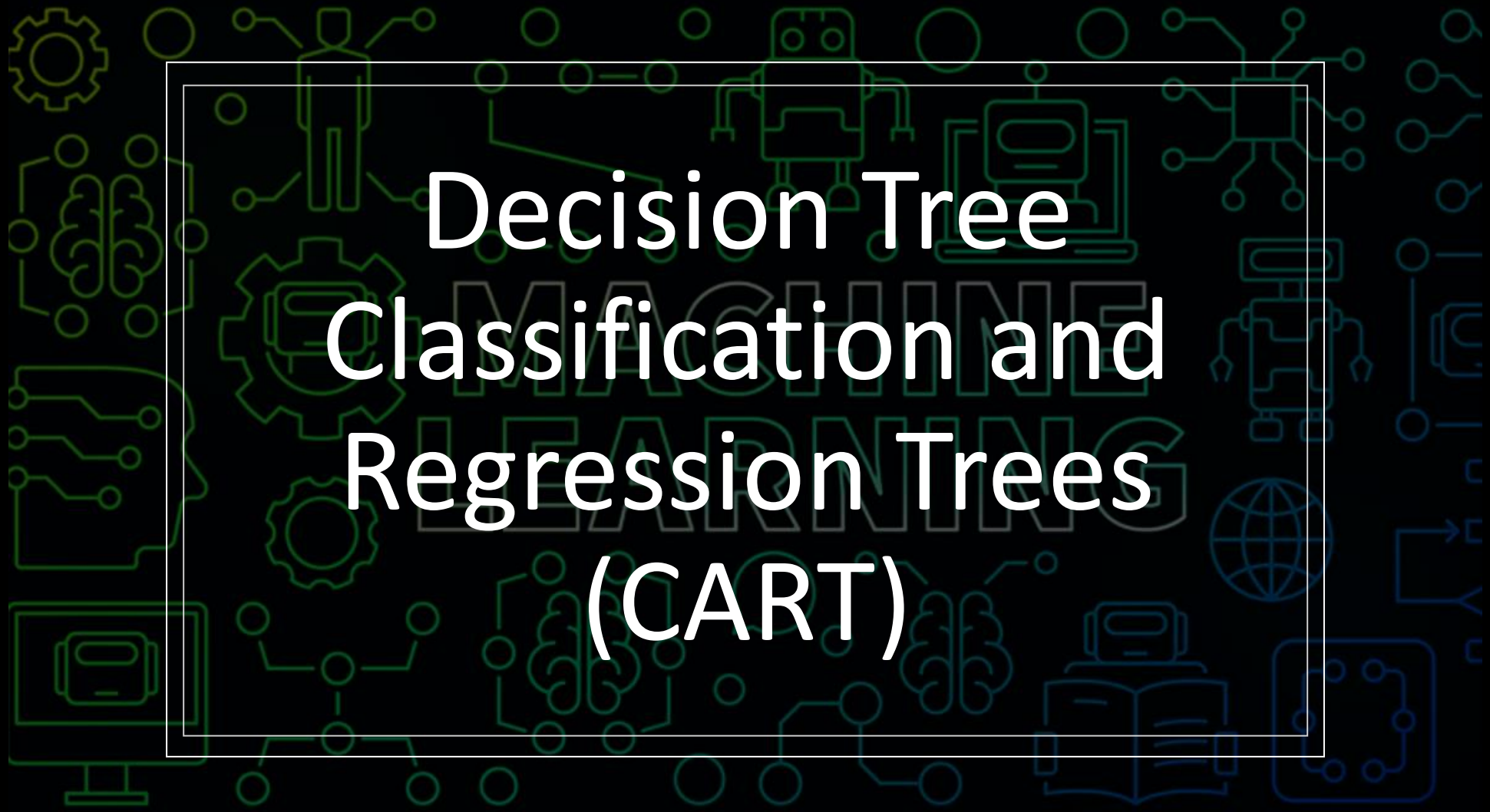
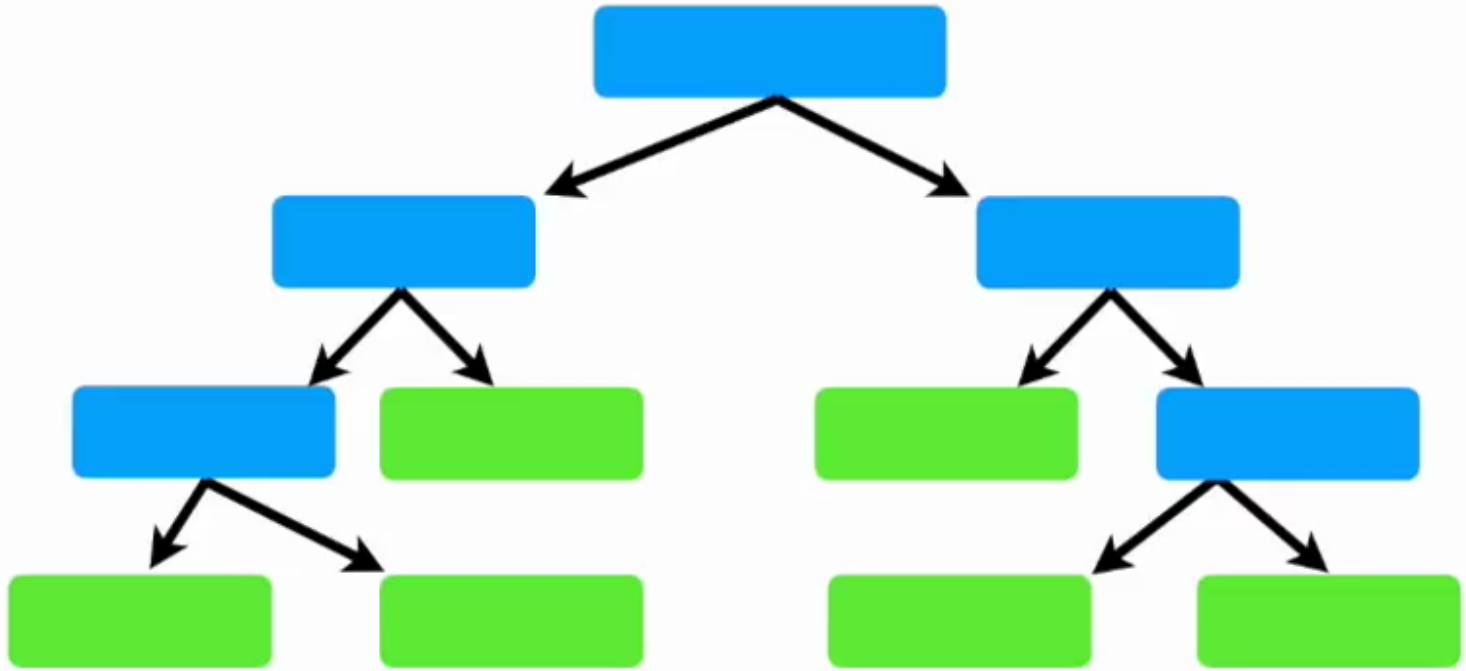


Decision Tree Classification and Regression Trees (CART)

இதுவரை – ID3 vs C4.5

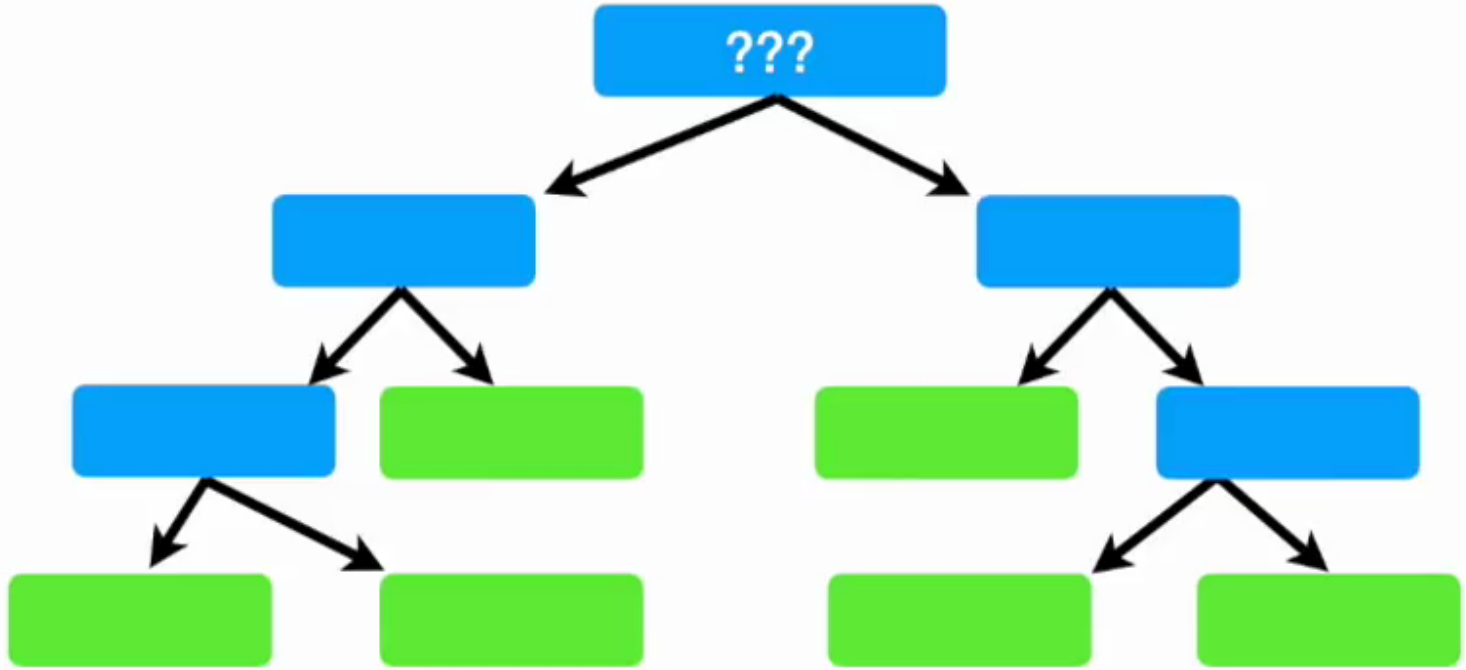
In this example, we want to create a tree that uses **chest pain**, **good blood circulation** and **blocked artery status** to predict…

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

The first thing we want to know is whether **Chest Pain**, **Good Blood Circulation** or **Blocked Arteries** should be at the very top of our tree.

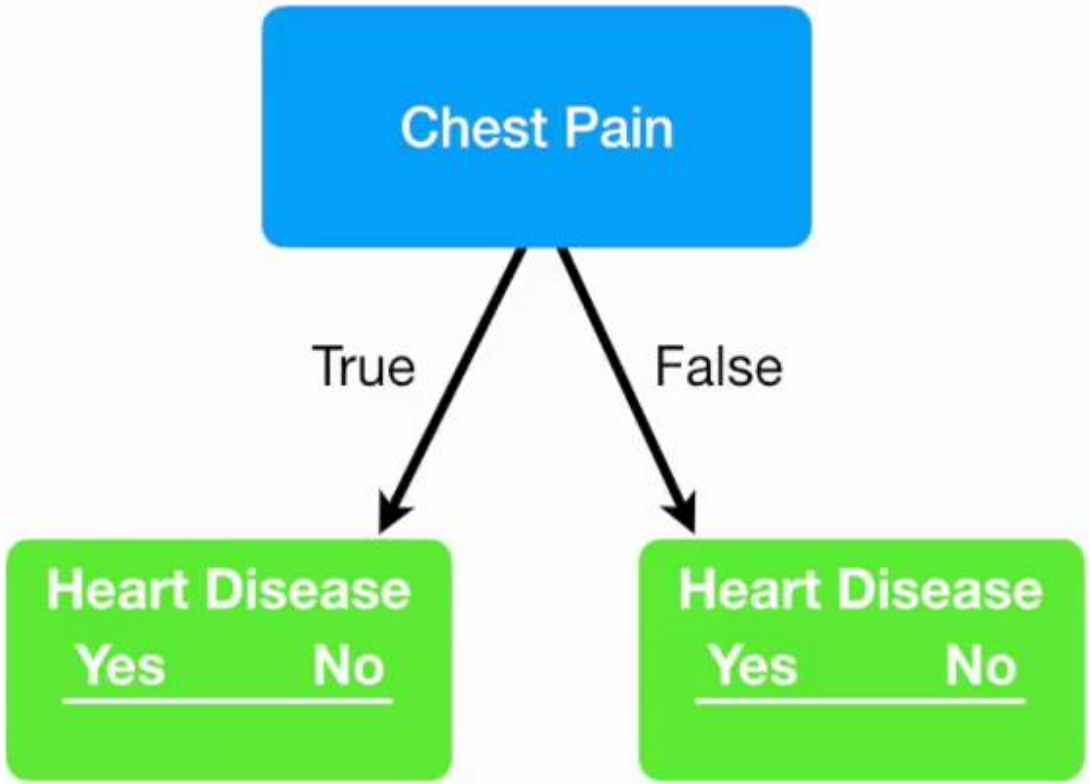


We start by looking at how well **Chest Pain** alone predicts heart disease…

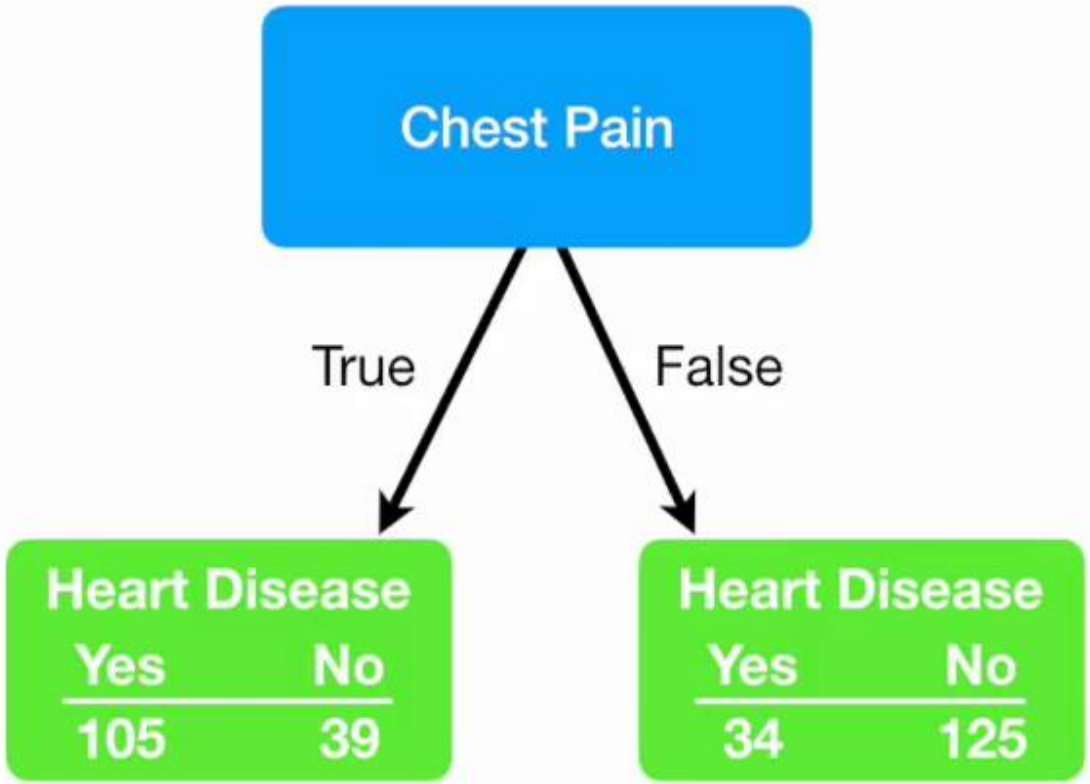
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

Here's a little tree that only takes chest pain into account.



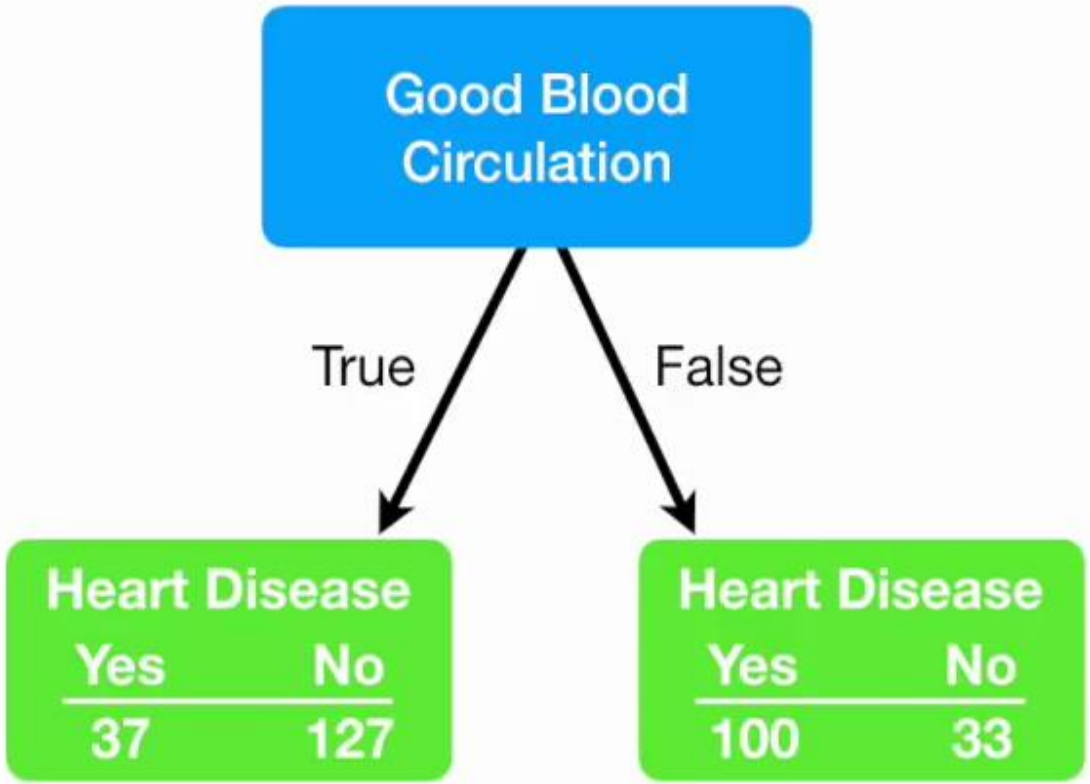
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



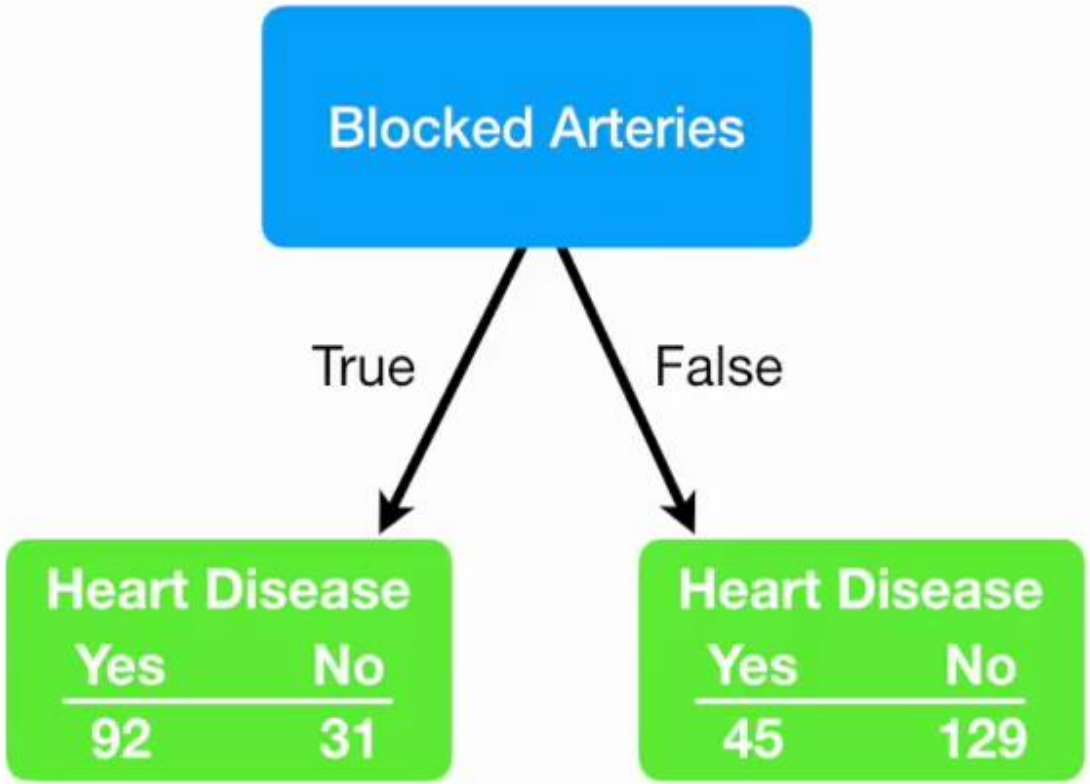
Ultimately, we look at chest pain and heart disease for all 303 patients in this study.

Now we do the exact same thing for **Good Blood Circulation**.

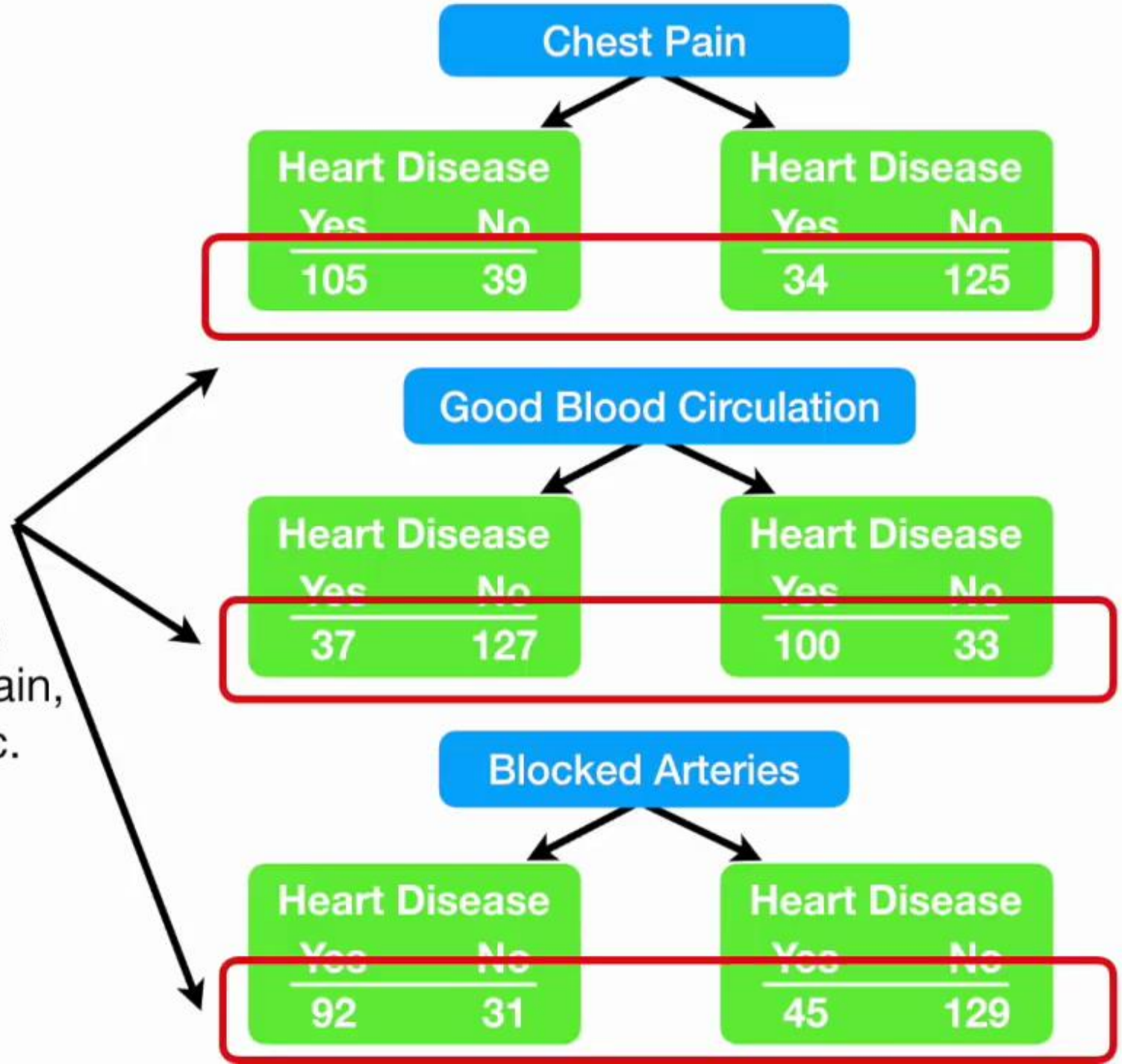
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



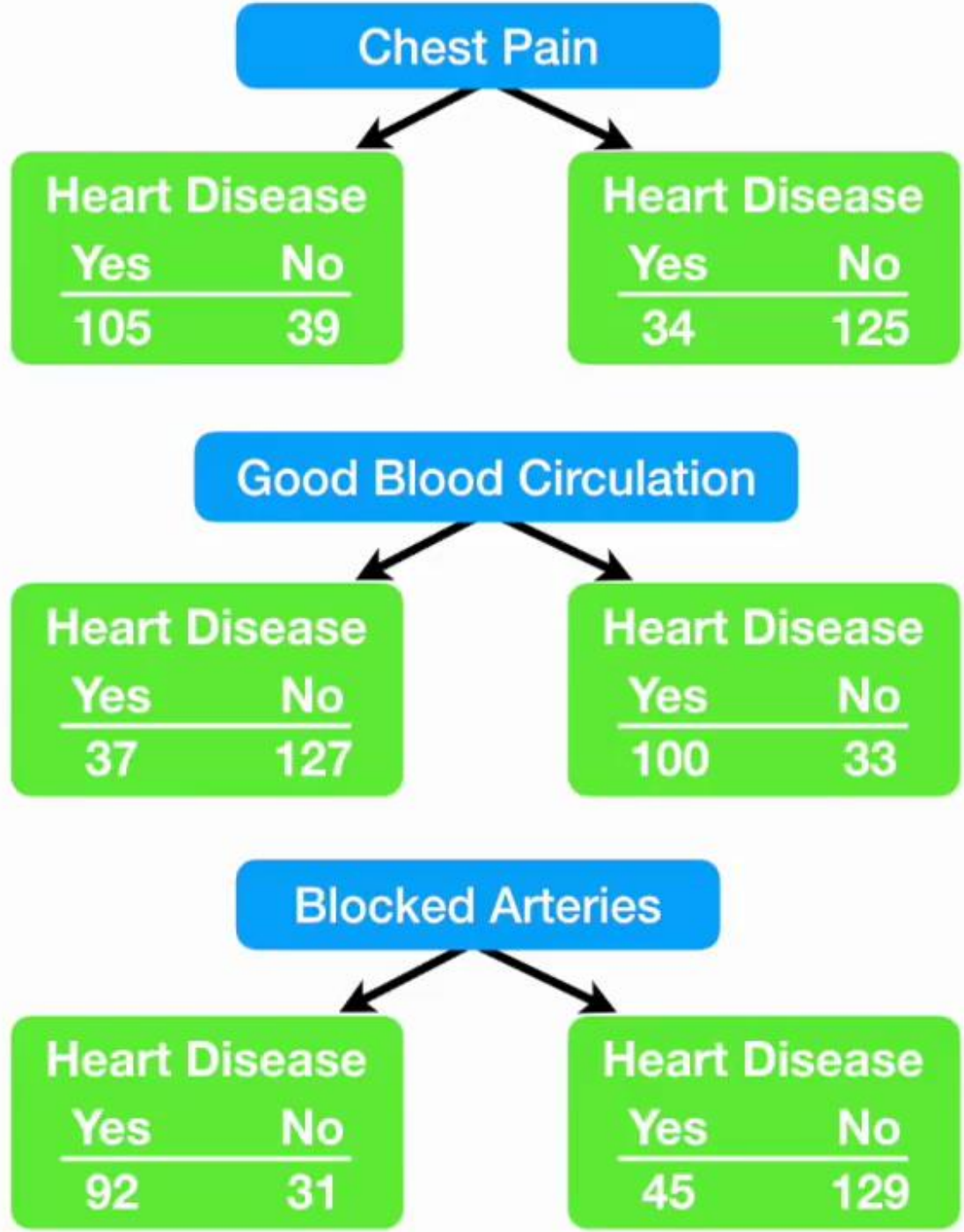
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

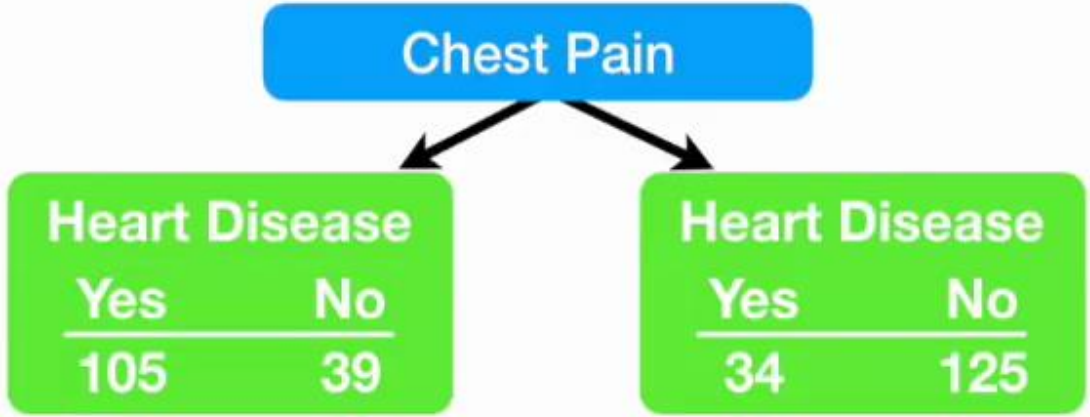


NOTE: The total number of patients with heart disease is different for Chest Pain, Good Blood Circulation and Blocked Arteries because some patients had measurements for Chest Pain, but not for Blocked Arteries, etc.



There are a bunch of ways to measure impurity, but I'm just going to focus on a very popular one called "**Gini**".

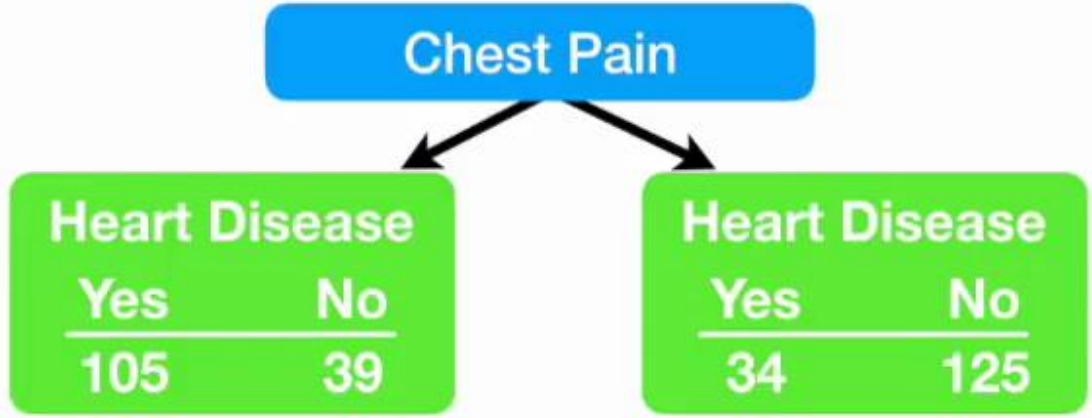




For this leaf, the Gini impurity = 1 - (the probability of "yes")^2 - (the probability of "no")^2

$$= 1 - \left(\frac{105}{105 + 39}\right)^2 - \left(\frac{39}{105 + 39}\right)^2$$

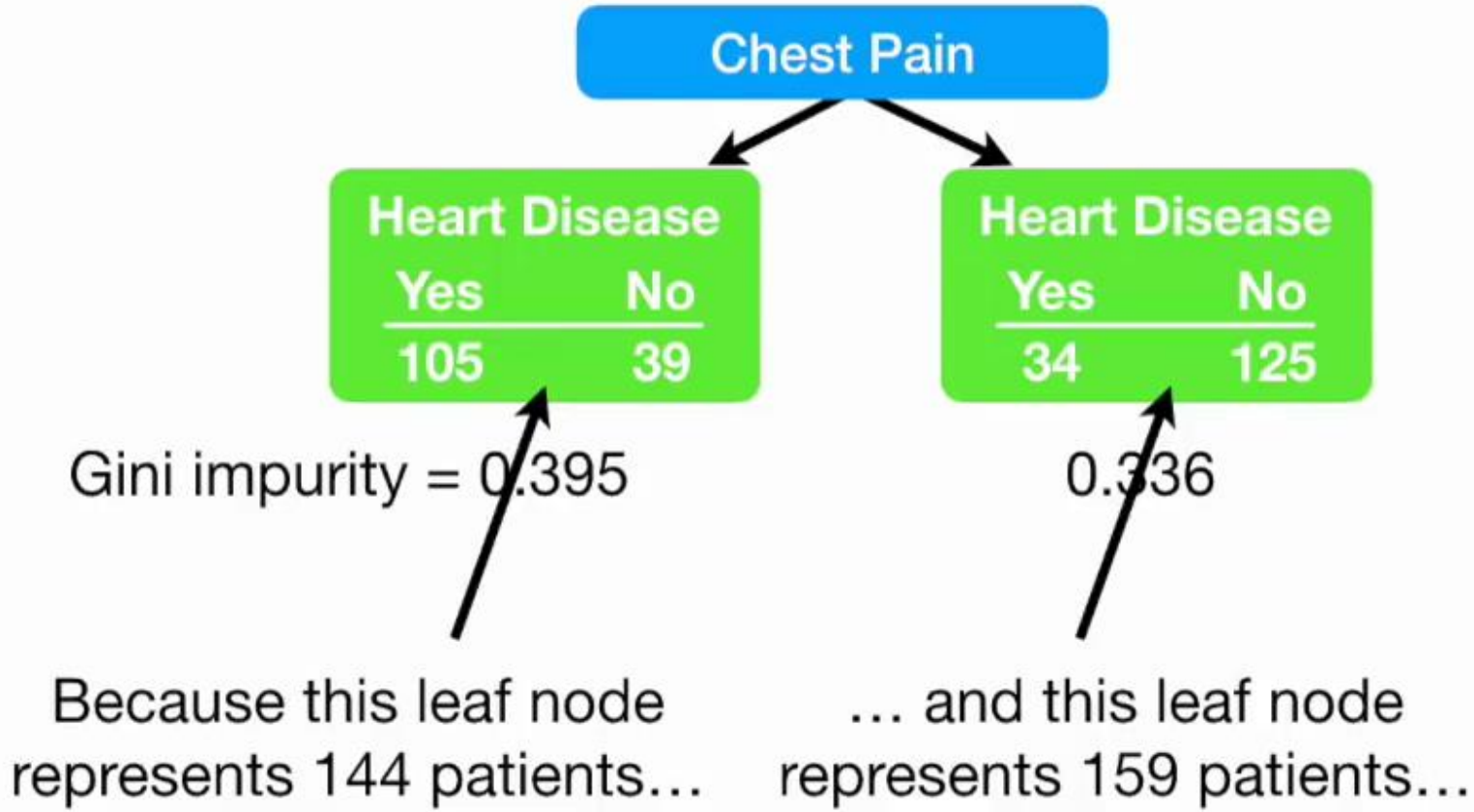
$$= 0.395$$



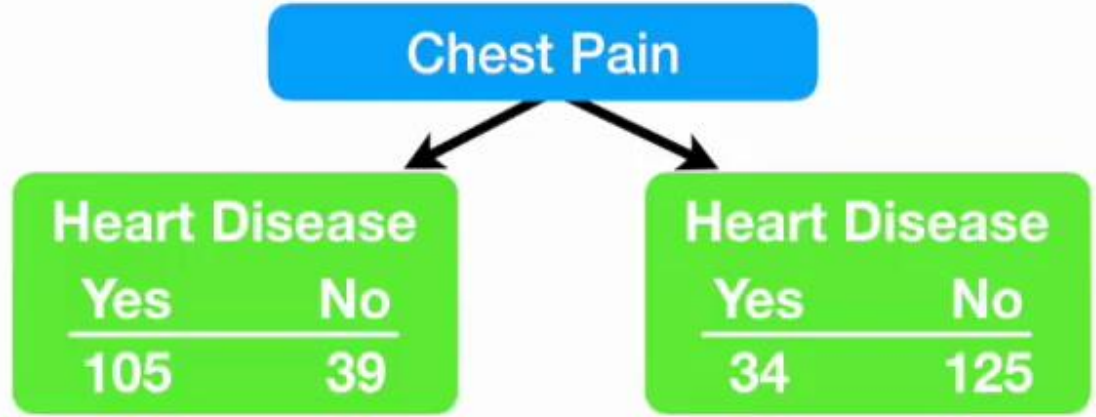
$$= 1 - (\text{the probability of “yes”})^2 - (\text{the probability of “no”})^2$$

$$= 1 - \left(\frac{34}{34 + 125}\right)^2 - \left(\frac{125}{34 + 125}\right)^2$$

$$= 0.336$$



Thus, the total Gini impurity for using Chest Pain to separate patients with and without heart disease is the **weighted average of the leaf node impurities**.



Gini impurity = 0.395

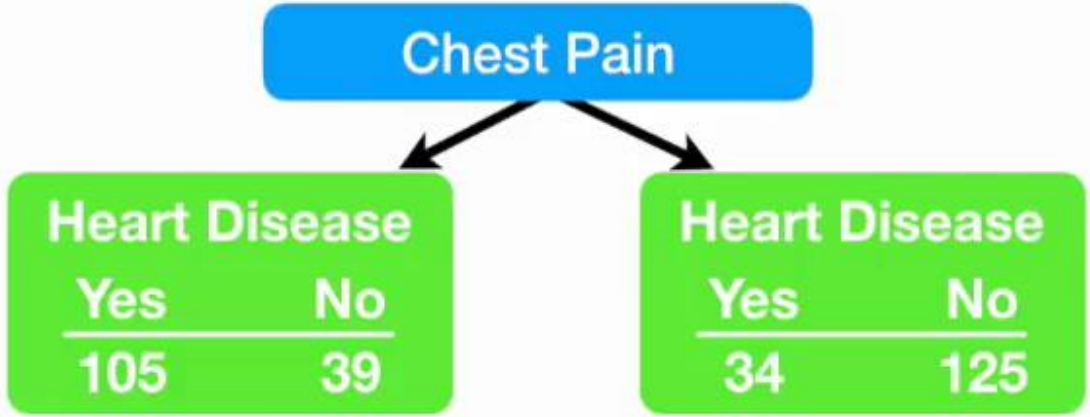
0.336

Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes

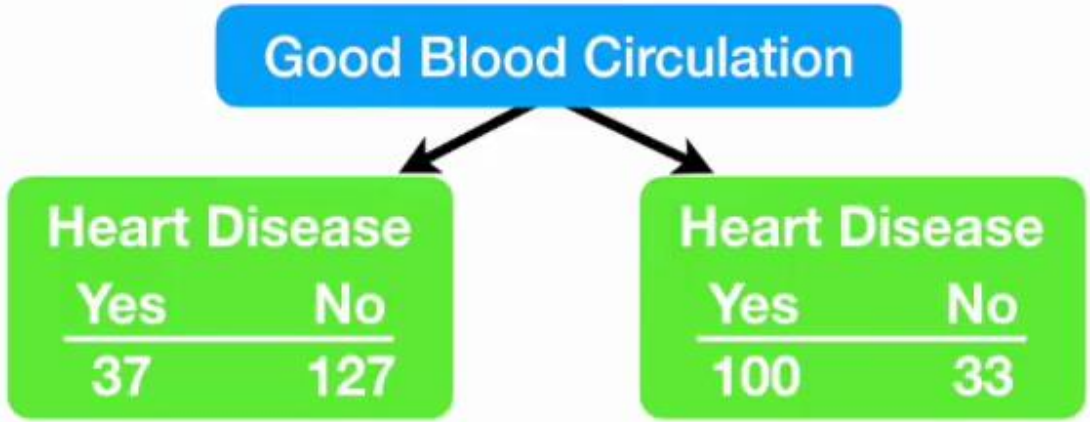
$$= \left(\frac{144}{144 + 159}\right) 0.395 + \left(\frac{159}{144 + 159}\right) 0.336$$

$$= 0.364$$

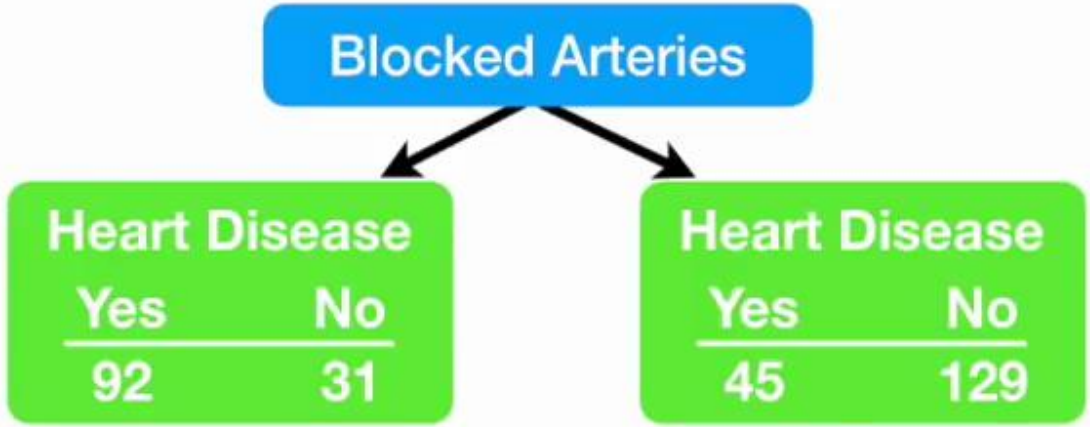
Gini impurity for Chest Pain = 0.364



Gini impurity for Good Blood Circulation = 0.360

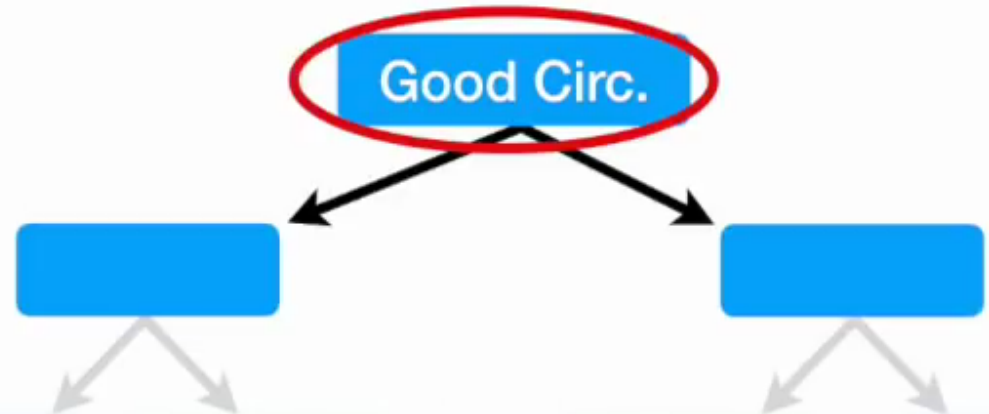


Gini impurity for Blocked Arteries = 0.381



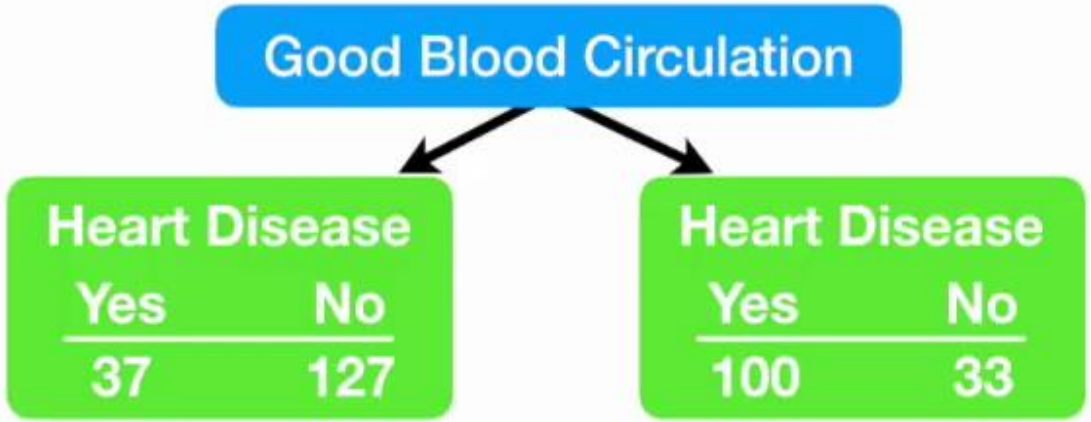
Gini impurity for Chest Pain = 0.364

…so we will use it at the root of the tree.



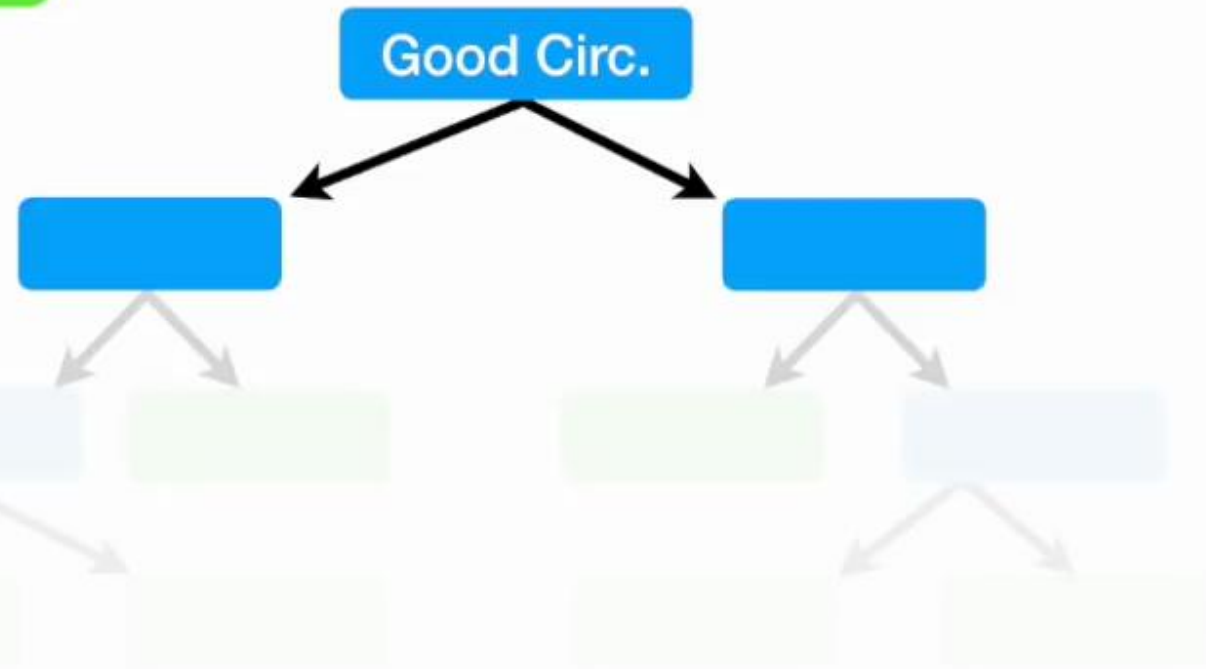
Gini impurity for Good Blood Circulation = 0.360

Gini impurity for Blocked Arteries = 0.381

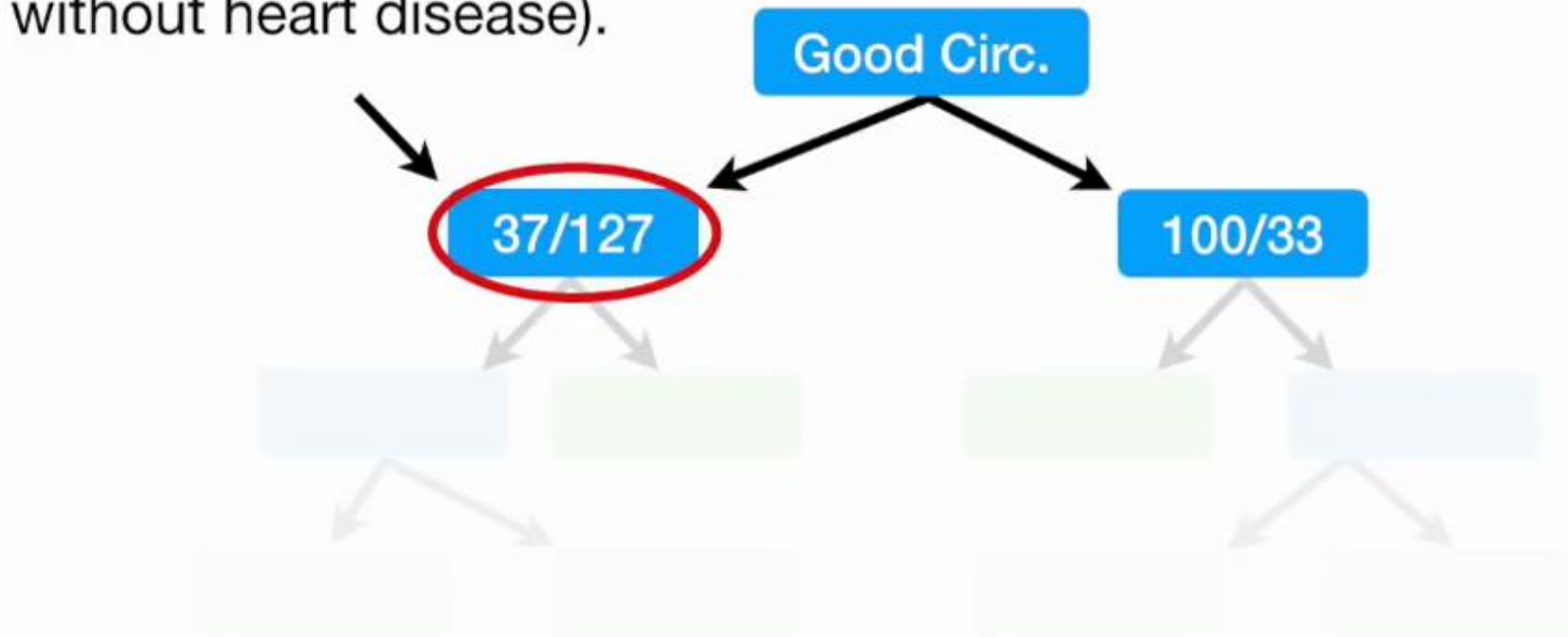


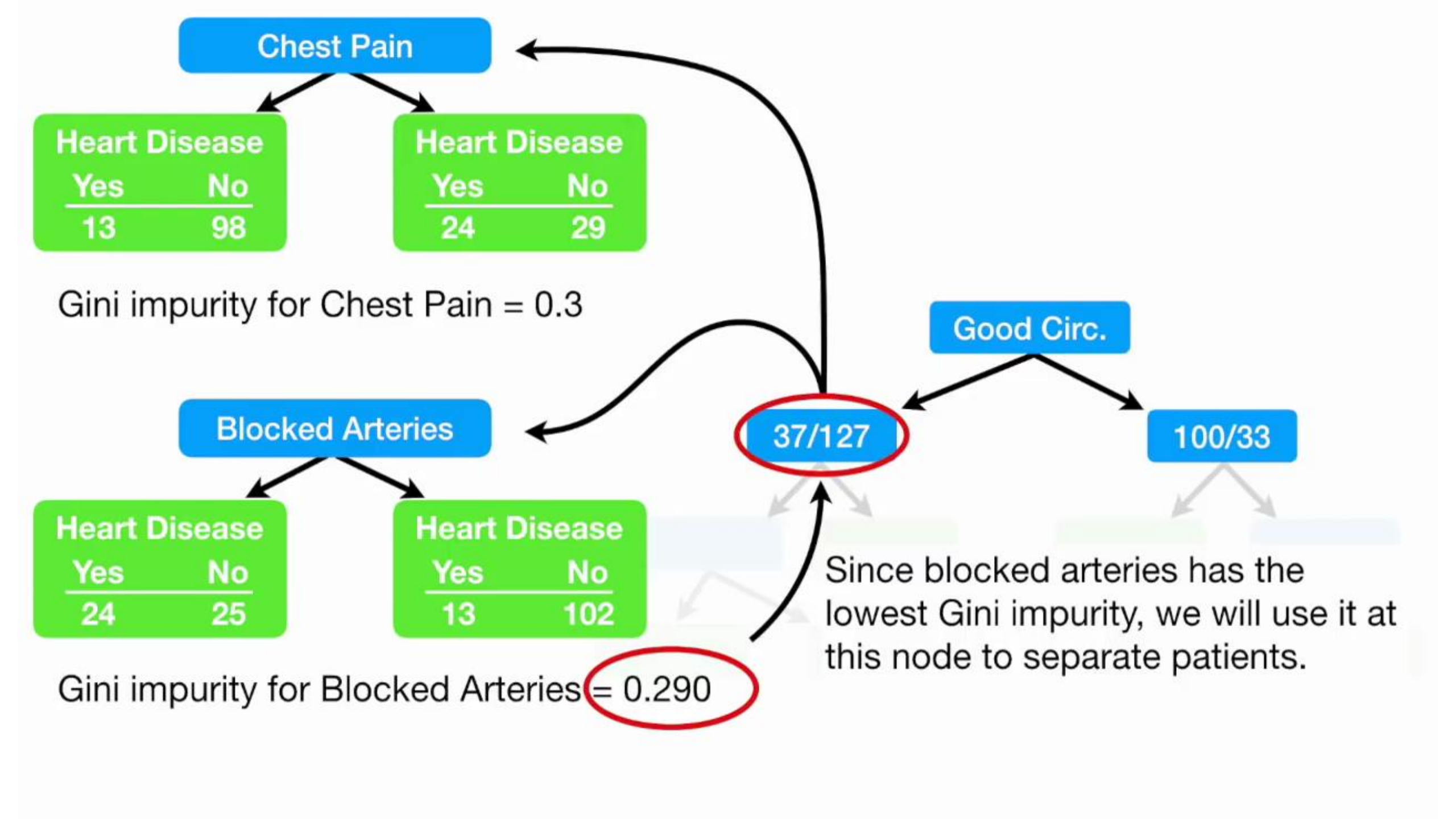
When we divided all of the patients using **Good Blood Circulation**, we ended up with “impure” leaf nodes.

Each leaf contained a mixture of patients with and without Heart Disease.



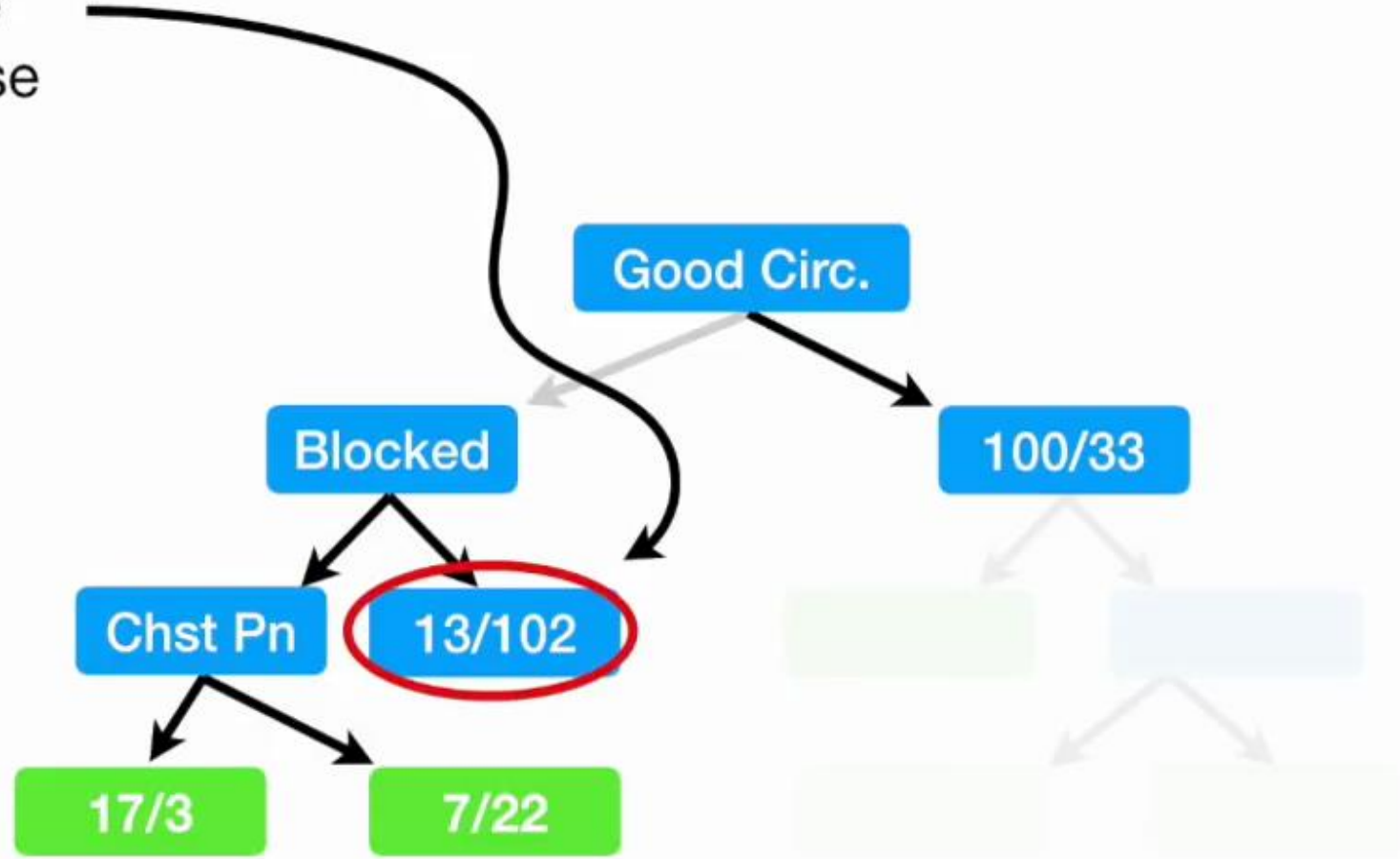
Now we need to figure how well **chest pain** and **blocked arteries** separate these 164 patients (37 with heart disease and 127 without heart disease).

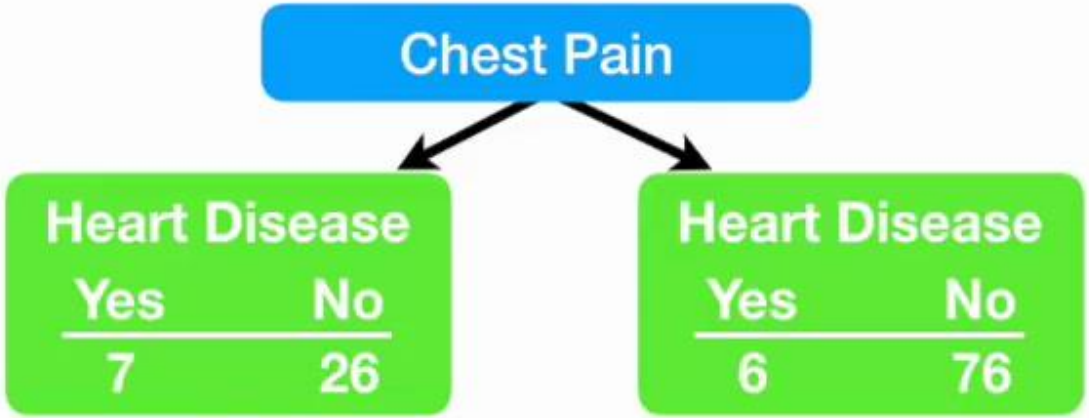



Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Gini impurity for Chest Pain = 0.3
Good Circ.
Blocked Arteries
37/127
100/33
Heart Disease
Yes No
24 25
Heart Disease
Yes No
13 102
Since blocked arteries has the lowest Gini impurity, we will use it at this node to separate patients.
Gini impurity for Blocked Arteries = 0.290

Now let's see what happens when we use chest pain to divide these 115 patients (13 with heart disease and 102 without).

NOTE: The vast majority of the patients in this node (89%) don't have heart disease.





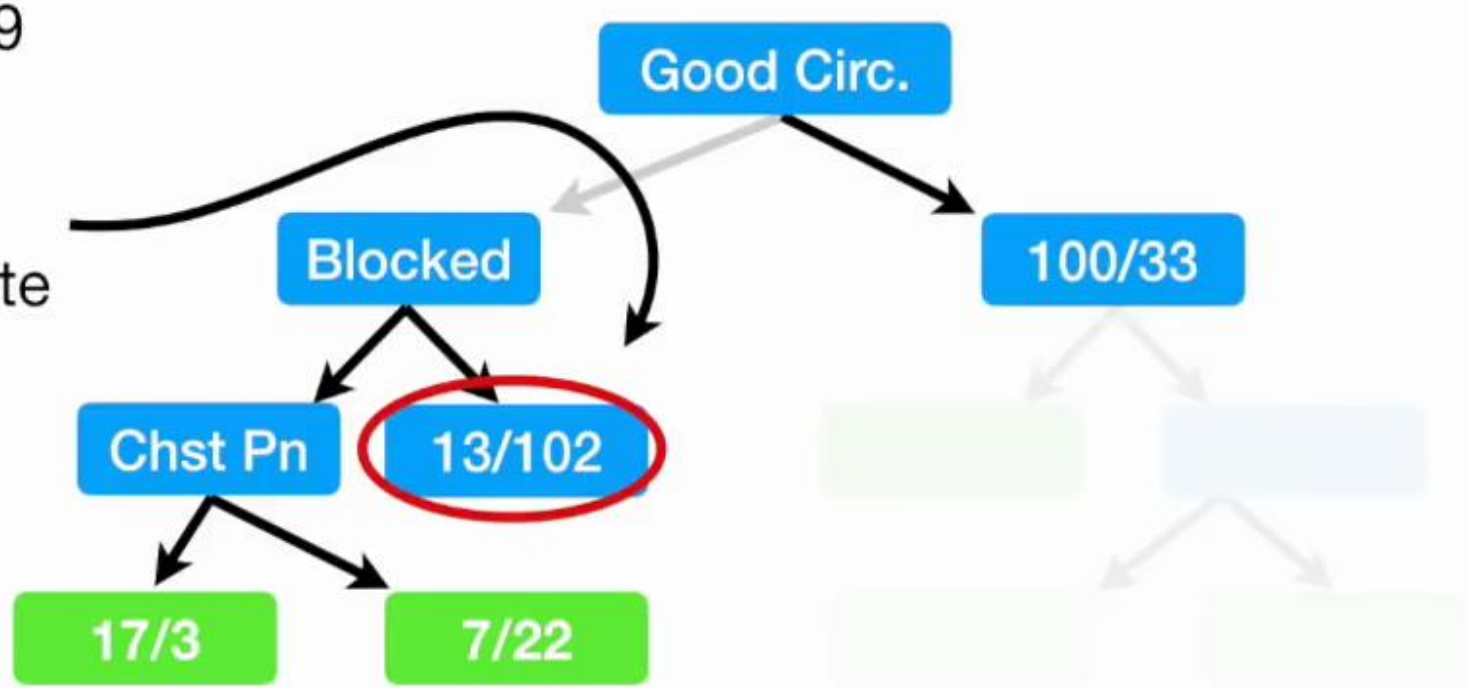
Gini impurity for Chest Pain = 0.29

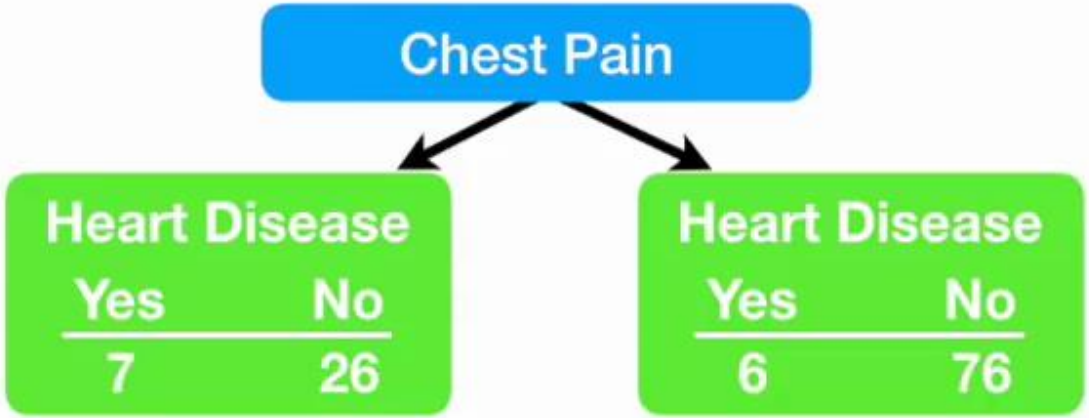
The Gini impurity for this node, before using chest pain to separate patients is…

= 1 - (the probability of "yes")^2
- (the probability of "no")^2

$$= 1 - \left(\frac{13}{13 + 102}\right)^2 - \left(\frac{102}{13 + 102}\right)^2$$

= 0.2





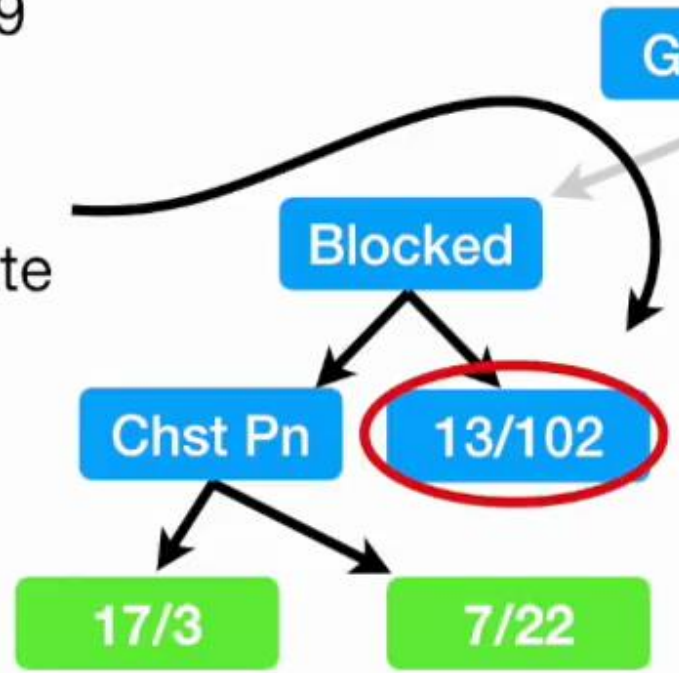
Gini impurity for Chest Pain = 0.29

The Gini impurity for this node, before using chest pain to separate patients is…

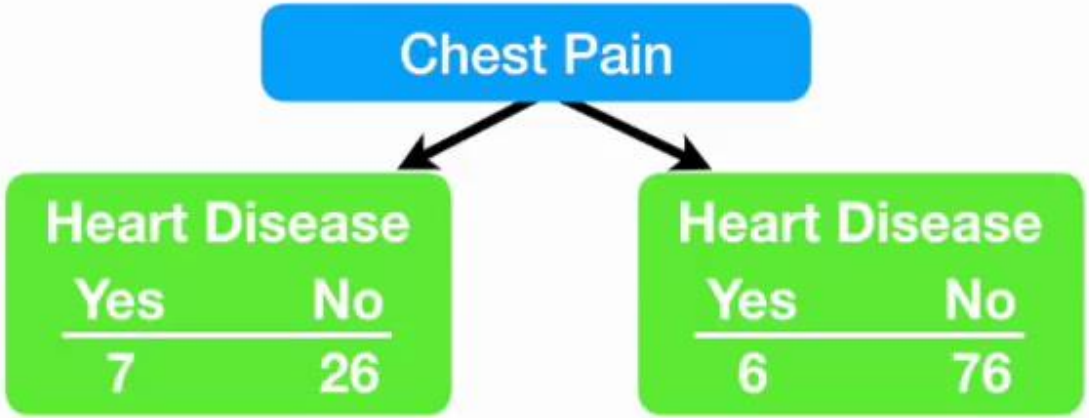
= 1 - (the probability of "yes")^2
- (the probability of "no")^2

$$= 1 - \left(\frac{13}{13 + 102}\right)^2 - \left(\frac{102}{13 + 102}\right)^2$$

= 0.2



The impurity is lower if we don't separate patients using Chest Pain.



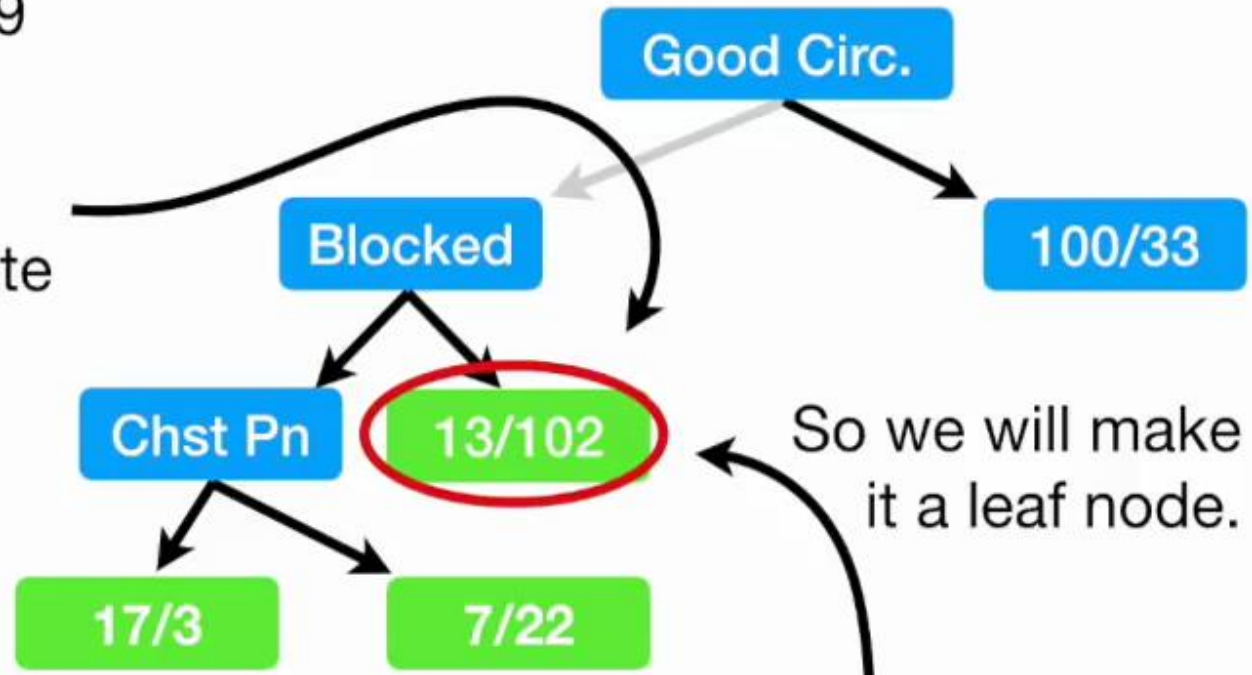
Gini impurity for Chest Pain = 0.29

The Gini impurity for this node, before using chest pain to separate patients is…

= 1 - (the probability of "yes")^2
- (the probability of "no")^2

$$= 1 - \left(\frac{13}{13 + 102}\right)^2 - \left(\frac{102}{13 + 102}\right)^2$$

= 0.2



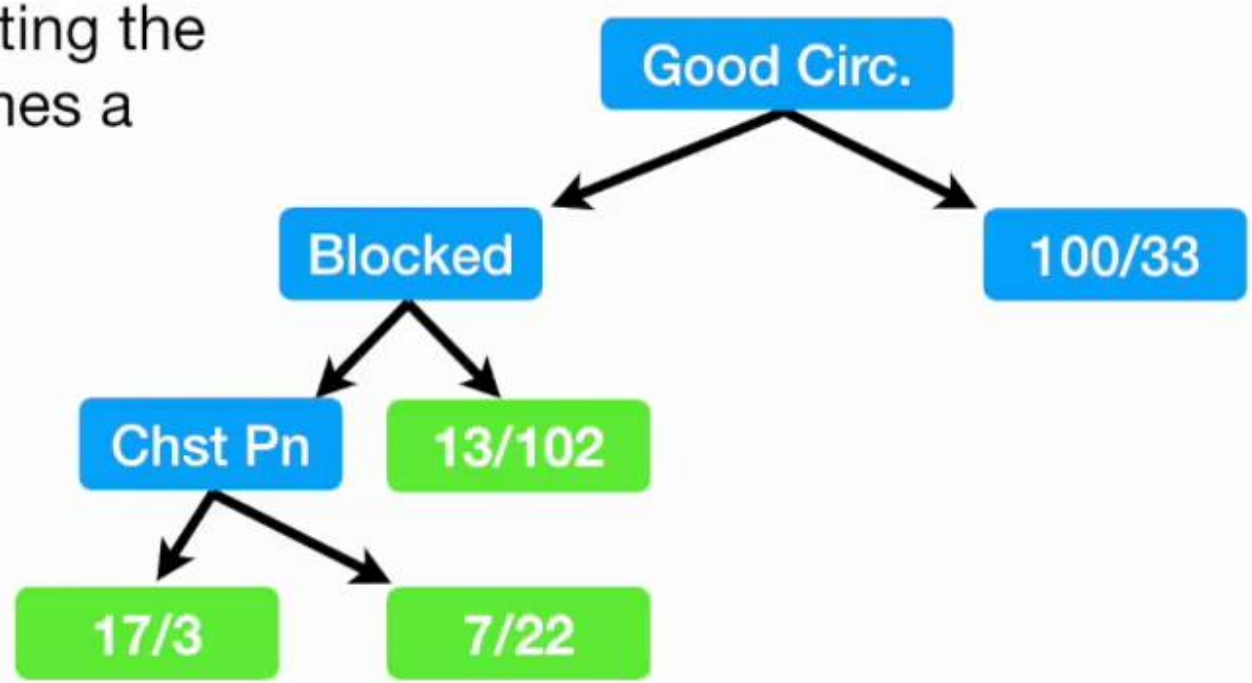
So we will make it a leaf node.

The good news is that we follow the exact same steps as we did on the left side:

1) Calculate all of the Gini impurity scores.

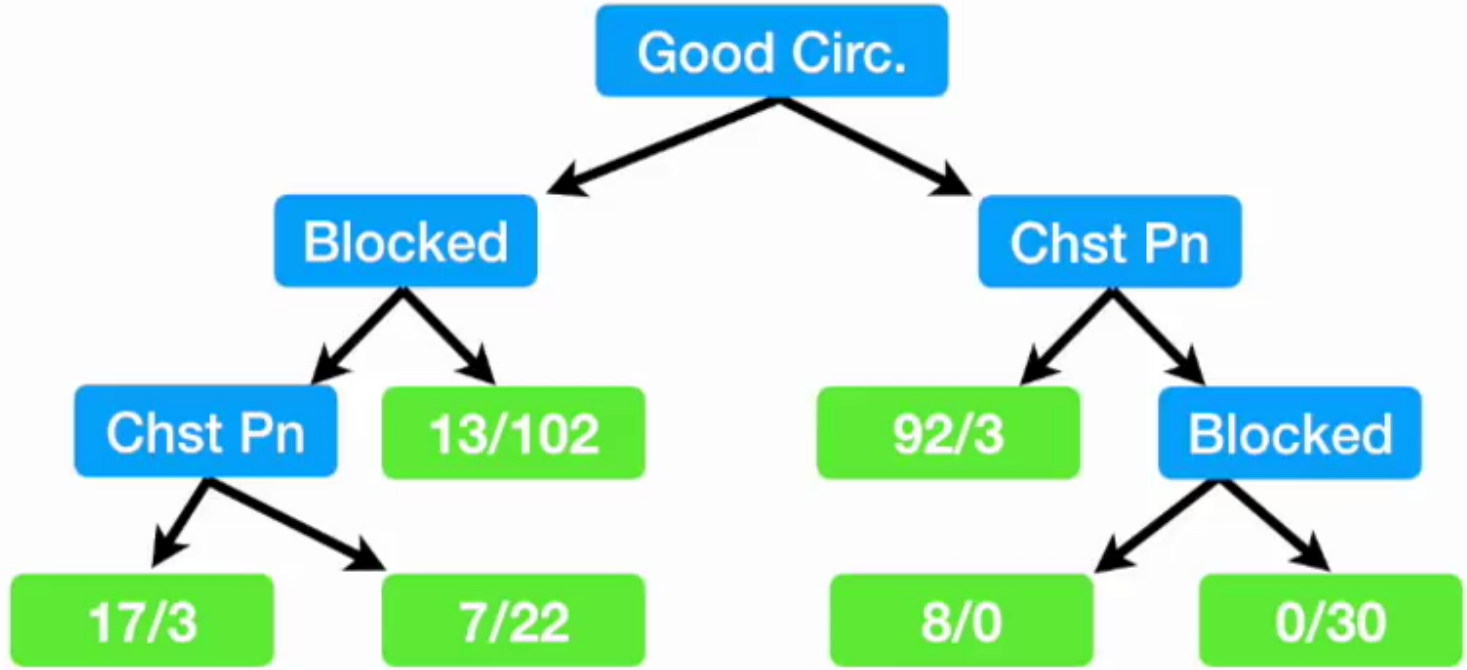
2) If the node itself has the lowest score, than there is no point in separating the patients any more and it becomes a leaf node.

3) If separating the data results in an improvement, than pick the separation with the lowest impurity value.



So far we've seen how to build a tree with "yes/no" questions at each step…

…but what if we have numeric data, like patient weight?



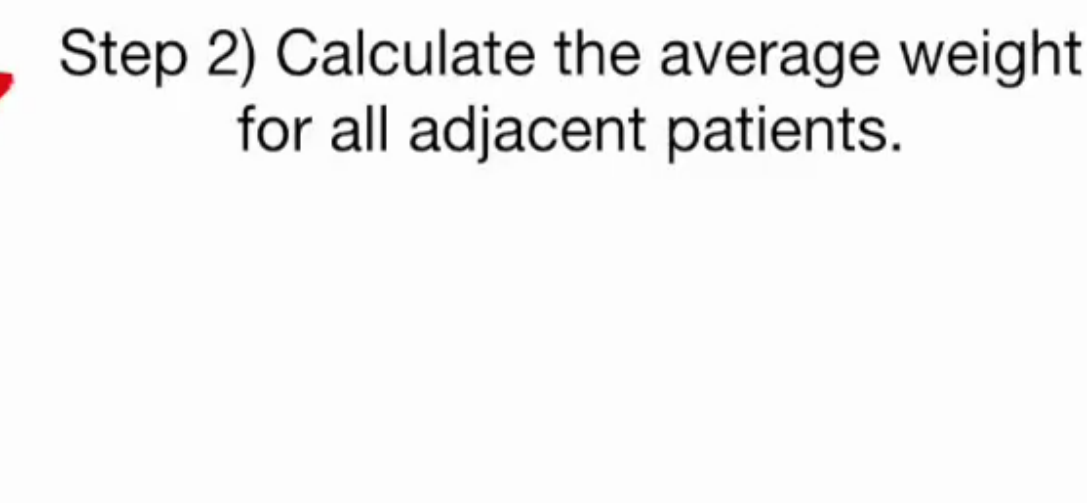
Weight	Heart Disease
220	Yes
180	Yes
225	Yes
190	No
155	No

How do we determine what's the best weight to use to divide the patients?

	Weight	Heart Disease
Lowest	155	No
	180	Yes
	190	No
	220	Yes
Highest	225	Yes

Step 1) Sort the patients by weight, lowest to highest.

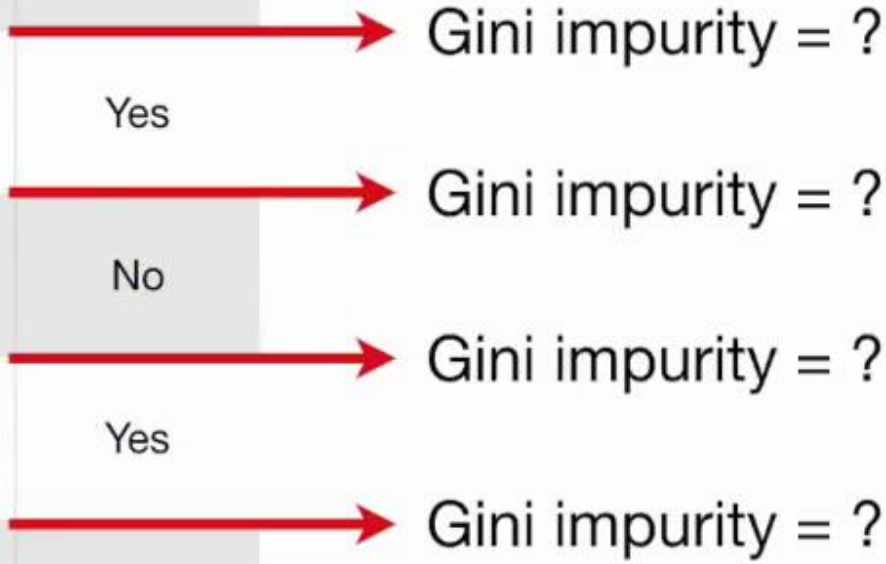
Weight	Heart Disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes

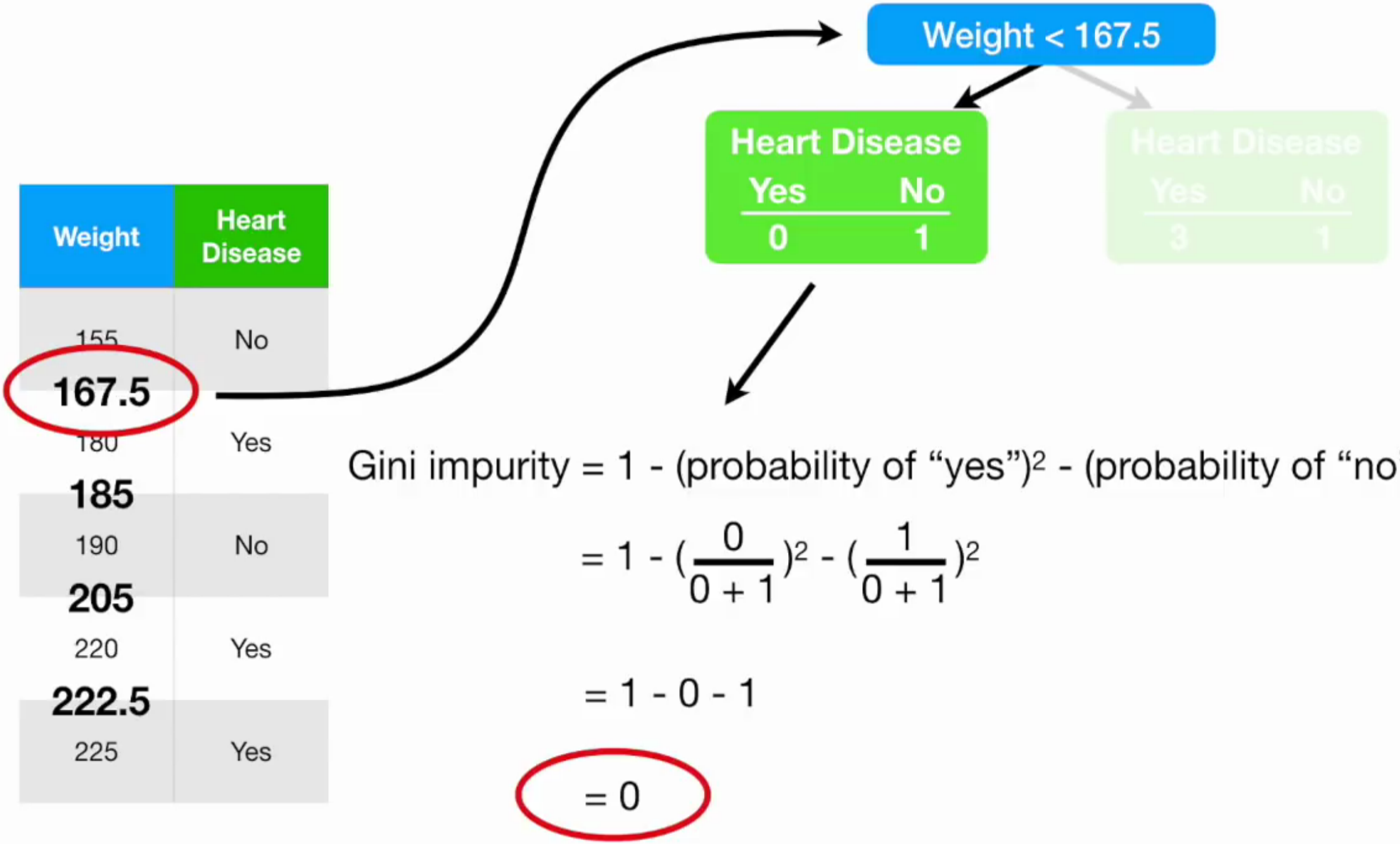


Step 2) Calculate the average weight for all adjacent patients.

Weight	Heart Disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes

Step 3) Calculate the impurity values for each average weight.




Weight < 167.5
Heart Disease
Yes No
0 1
Heart Disease
Yes No
3 1
Weight
Heart Disease
155
No
167.5
180
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Gini impurity = 1 - (probability of “yes”)² - (probability of “no”)²
= 1 - (0 / 0 + 1)² - (1 / 0 + 1)²
= 1 - 0 - 1
= 0

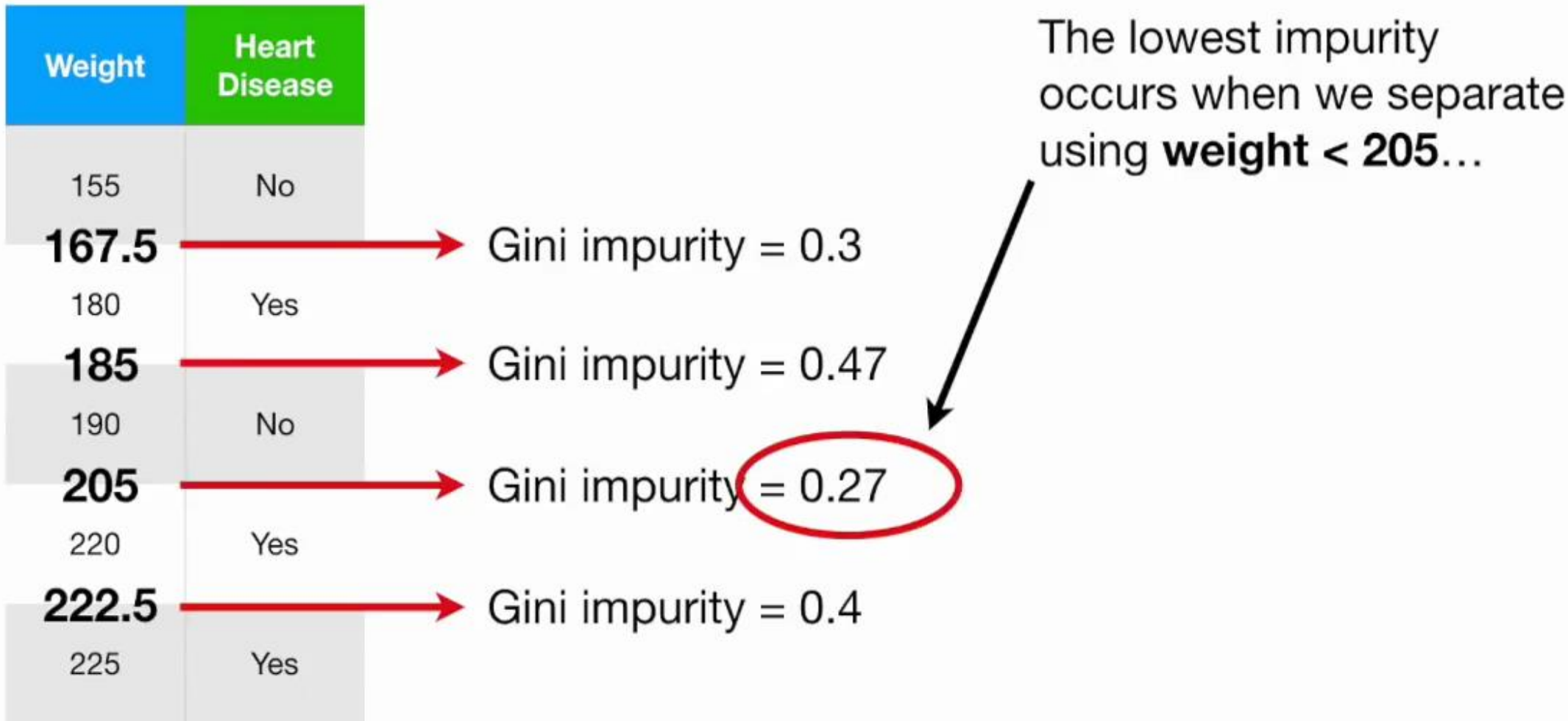
Weight	Heart Disease
155	No
167.5	
180	Yes
185	
190	No
205	
220	Yes
222.5	
225	Yes

Gini impurity = 0

0.375

Gini impurity for Weight < 167.5 is the weighted average of the impurities for the two leaves.

$$= \left(\frac{1}{1+4}\right) 0 \; + \left(\frac{4}{1+4}\right) 0.336 = 0.3$$

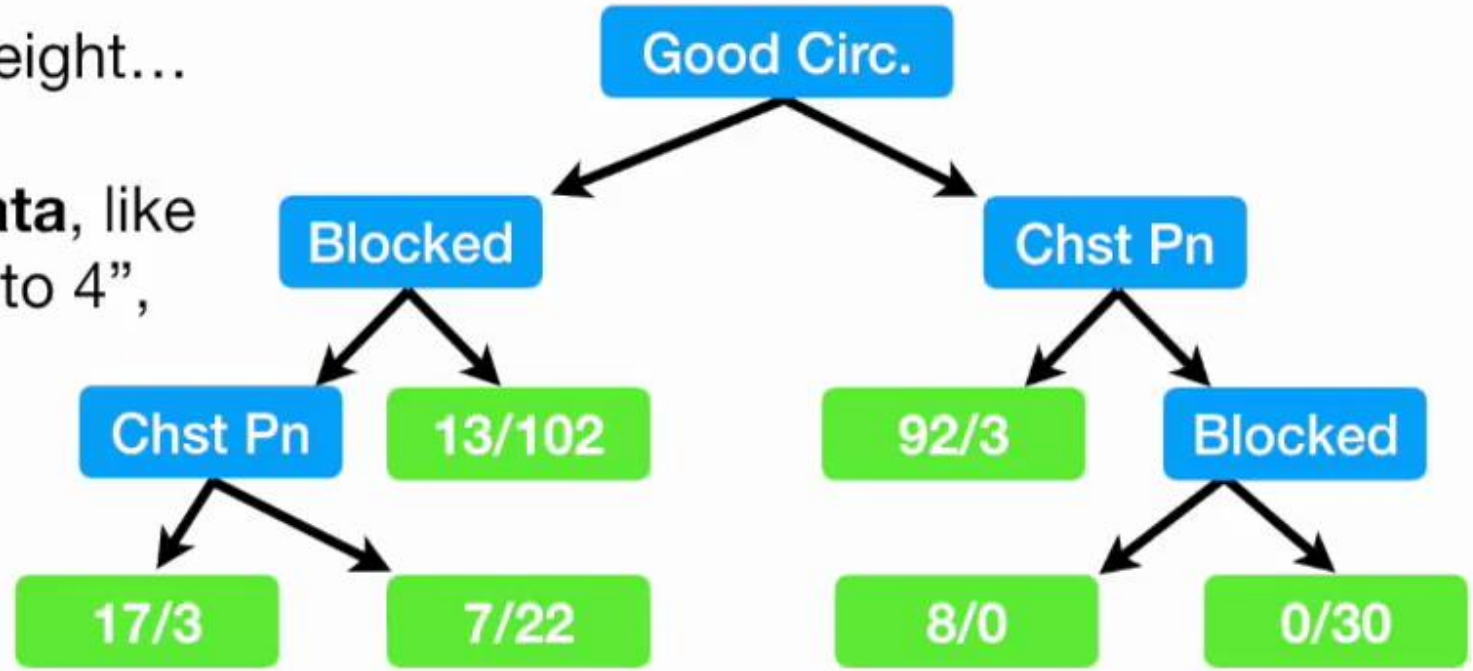

Weight
Heart Disease
155
No
167.5
Gini impurity = 0.3
180
Yes
185
Gini impurity = 0.47
190
No
205
Gini impurity = 0.27
220
Yes
222.5
Gini impurity = 0.4
225
Yes
The lowest impurity occurs when we separate using **weight < 205**…

Now we've seen how to build a tree with…

1) "yes/no" questions at each step…

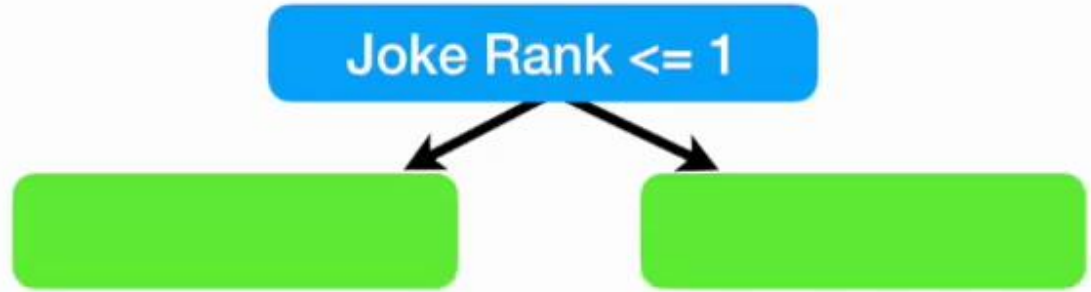
2) Numeric data, like patient weight…

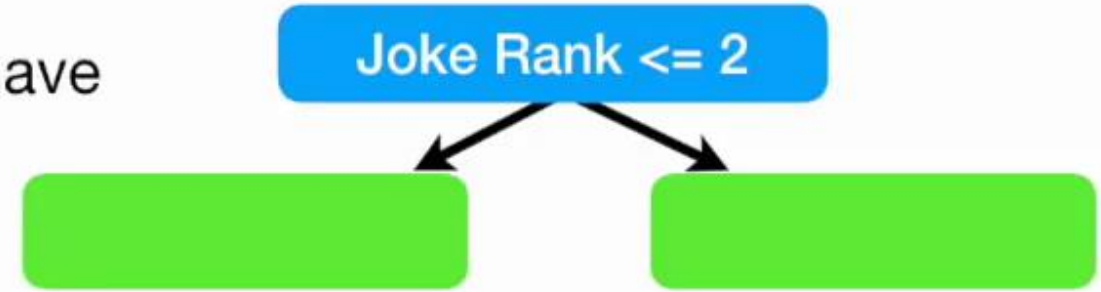
Now let's talk about **ranked data**, like "rank my jokes on a scale of 1 to 4", and **multiple choice data**, like "which color do you like, red, blue or green?"

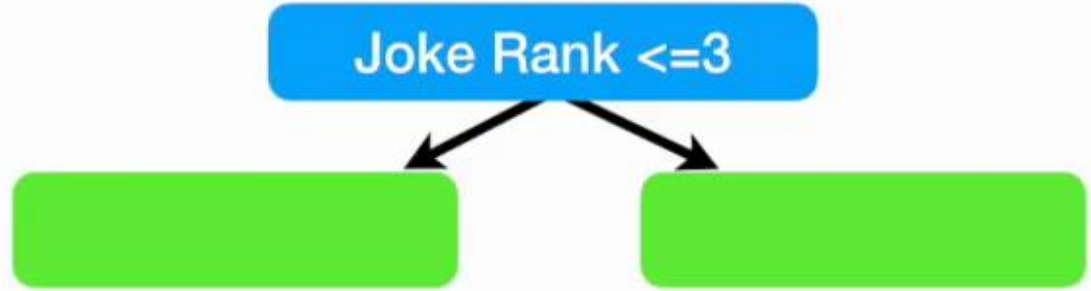


Rank my jokes…	Likes StatQuest
1	Yes
1	No
3	Yes
1	Yes
etc…	etc…

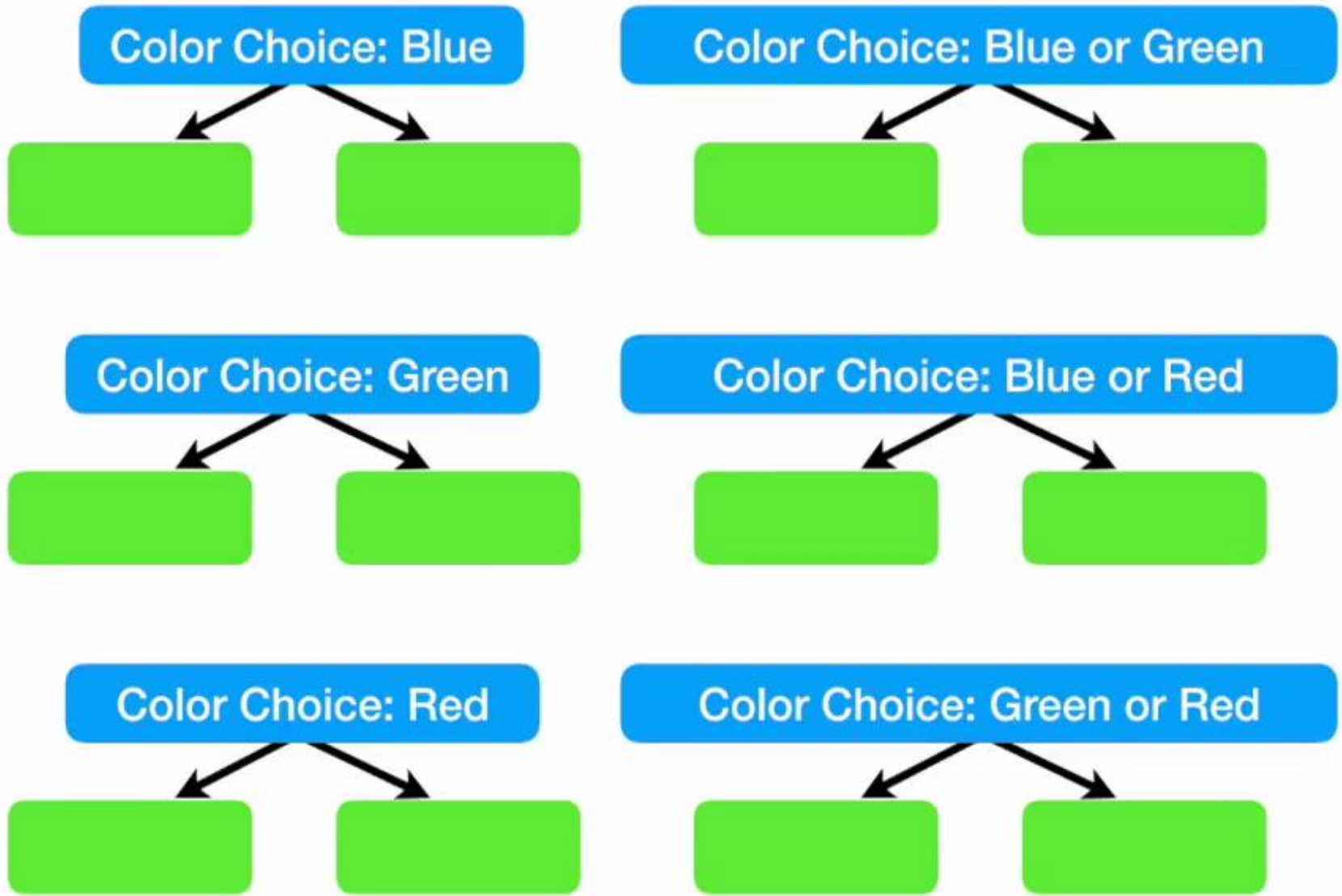
NOTE: We don't have to calculate an impurity score for Joke Rank <= 4 because that would include everyone.



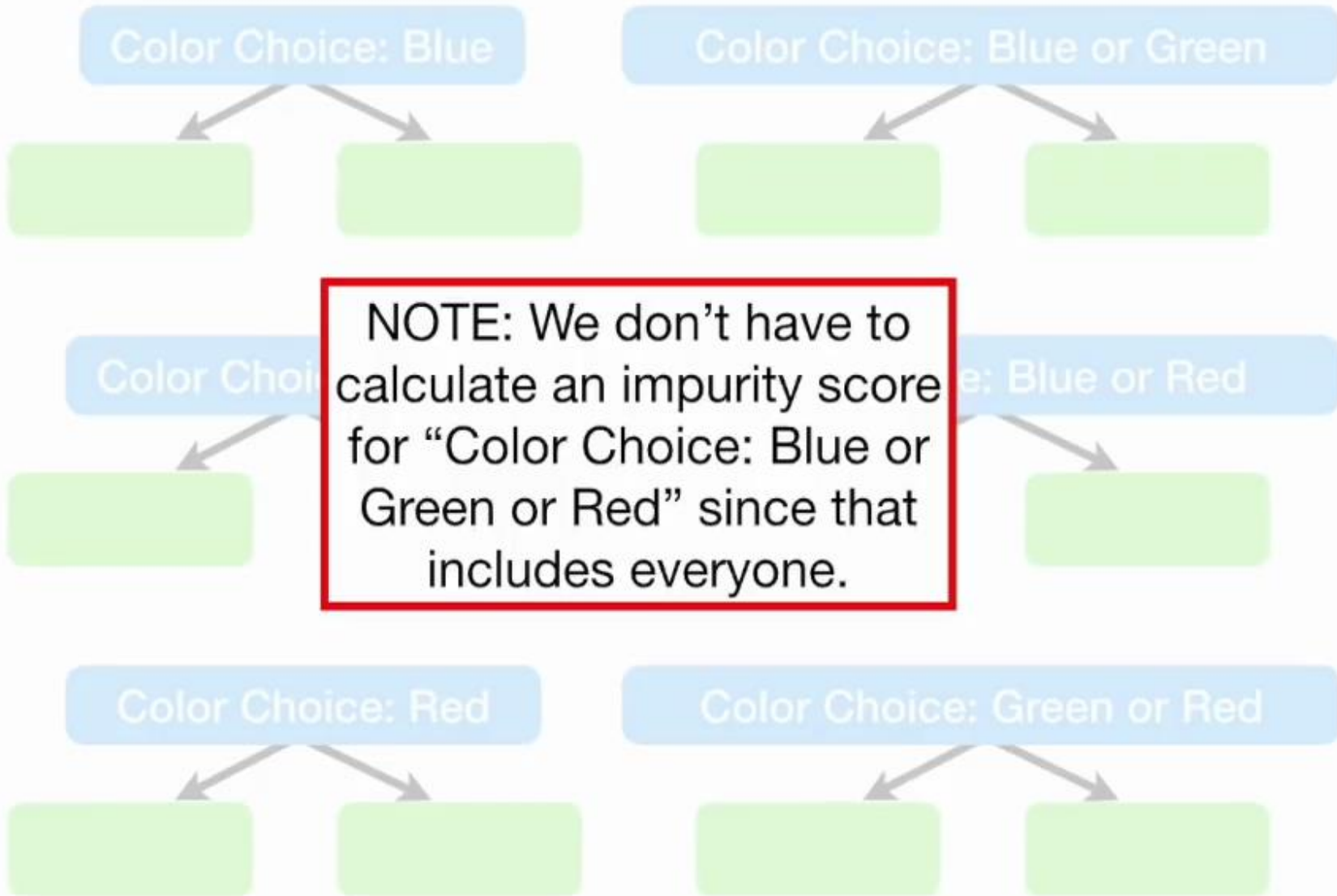




Color Choice	Likes StatQuest
Green	Yes
Blue	No
Red	Yes
Green	Yes
etc…	etc…



Color Choice	Likes StatQuest
Green	Yes
Blue	No
Red	Yes
Green	Yes
etc…	etc…




மரம் வளர்ப்போம் மழை பெறுவோம்
மண் வளத்தை காப்போம்