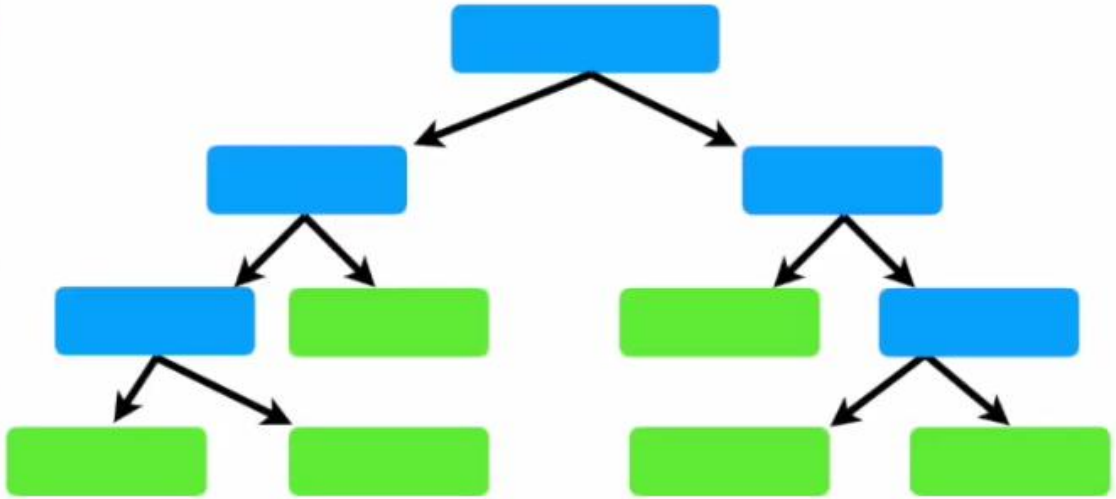
DecisionTree Classifier

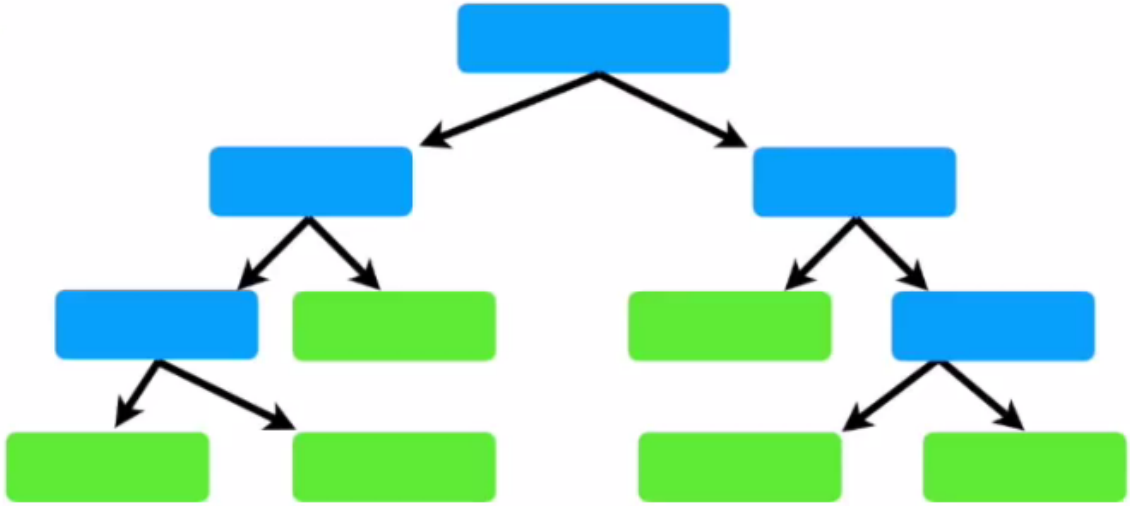
decision tree!!!

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…



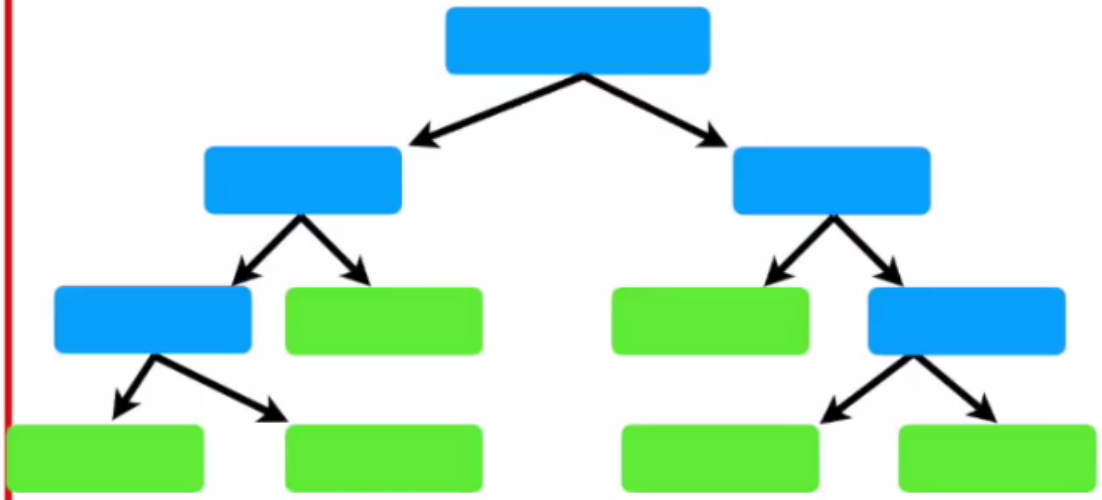
In this example, we want to create a tree that uses **chest pain**, **good blood circulation** and **blocked artery status** to predict…

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…



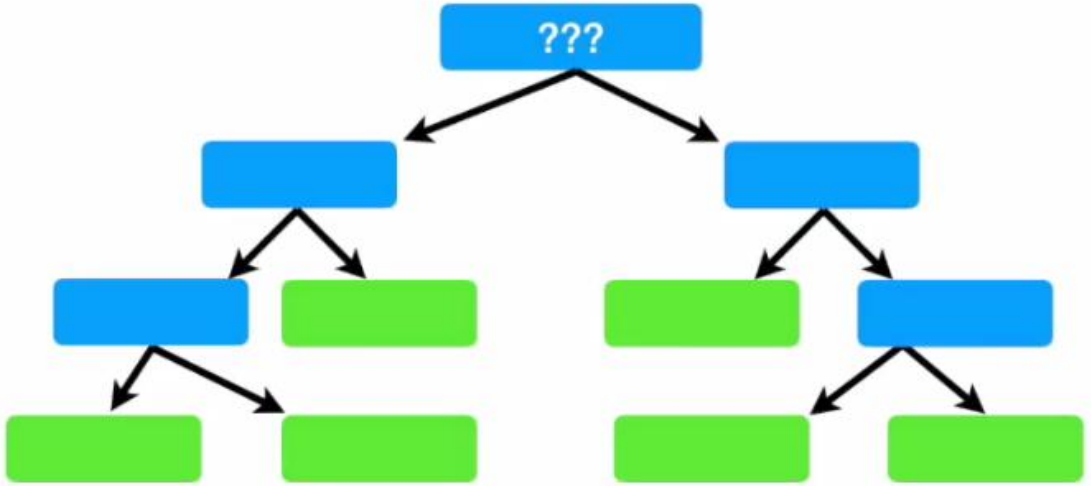
…whether or not a patient has heart disease.

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…

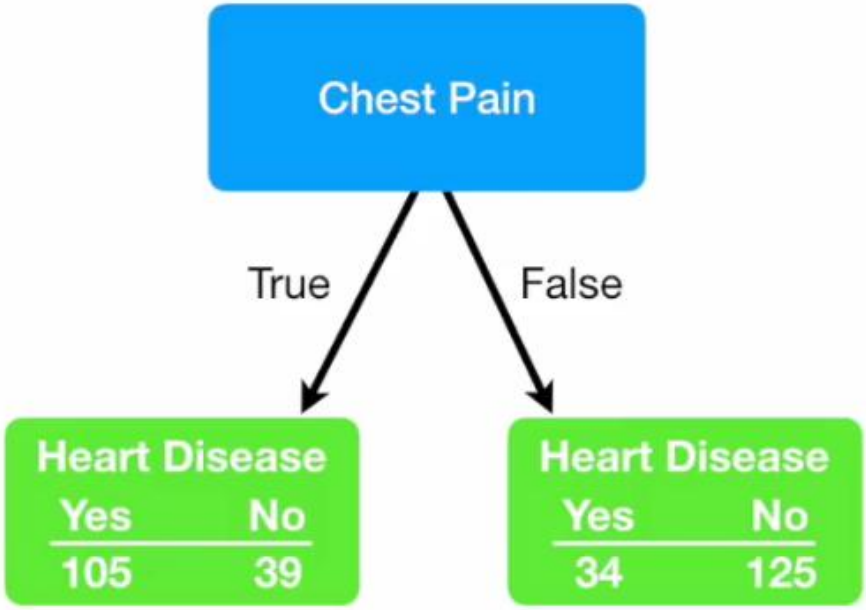
The first thing we want to know is whether **Chest Pain**, **Good Blood Circulation** or **Blocked Arteries** should be at the very top of our tree.



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…

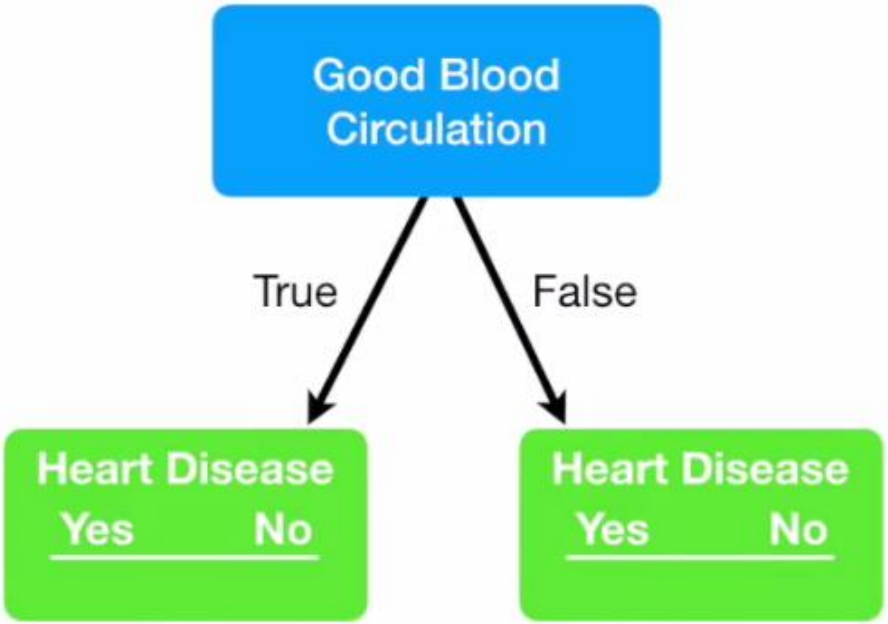
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



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

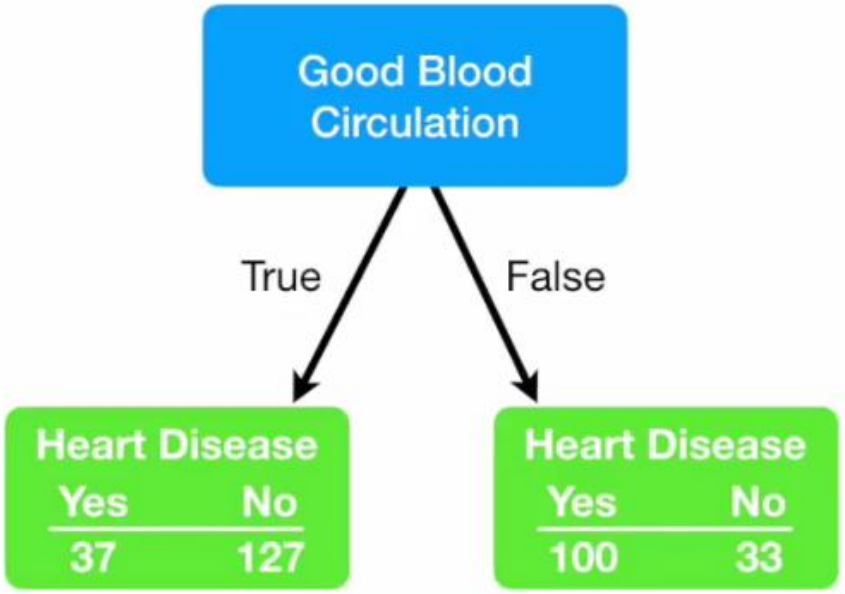
Now we do the exact same thing for **Good Blood Circulation**.

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…



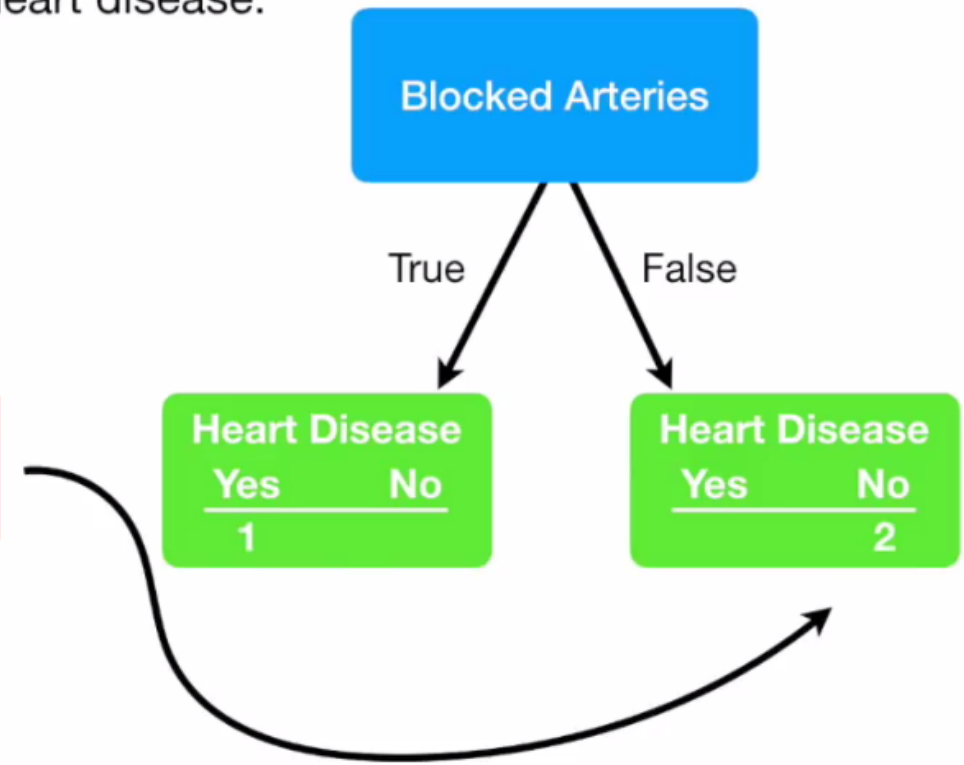
Now we do the exact same thing for **Good Blood Circulation**.

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…

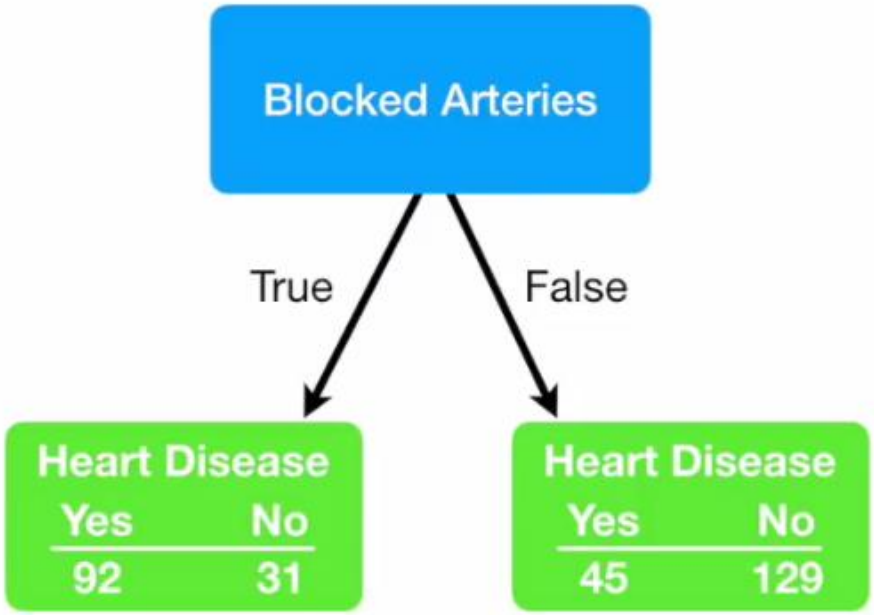


Lastly, we look at how **Blocked Arteries** separates the patients with and without heart disease.

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…

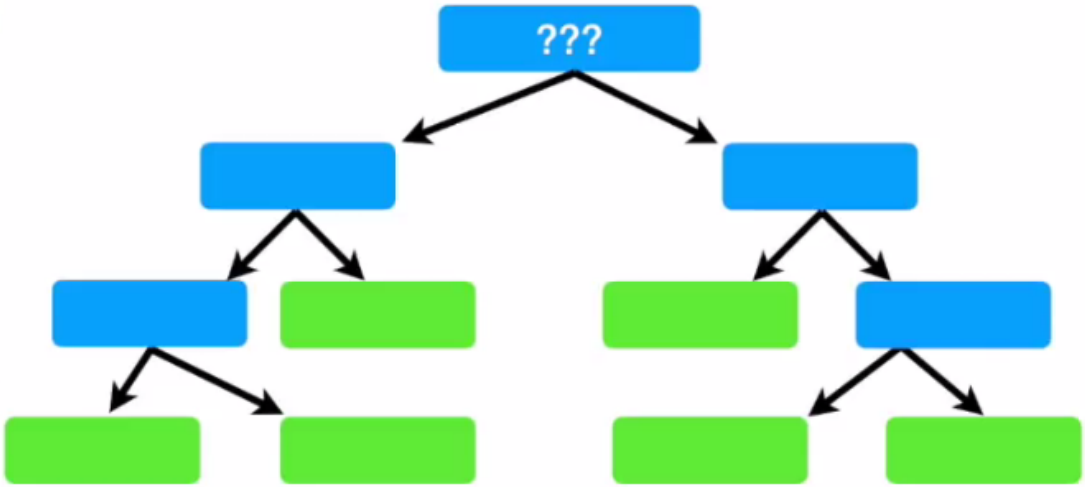


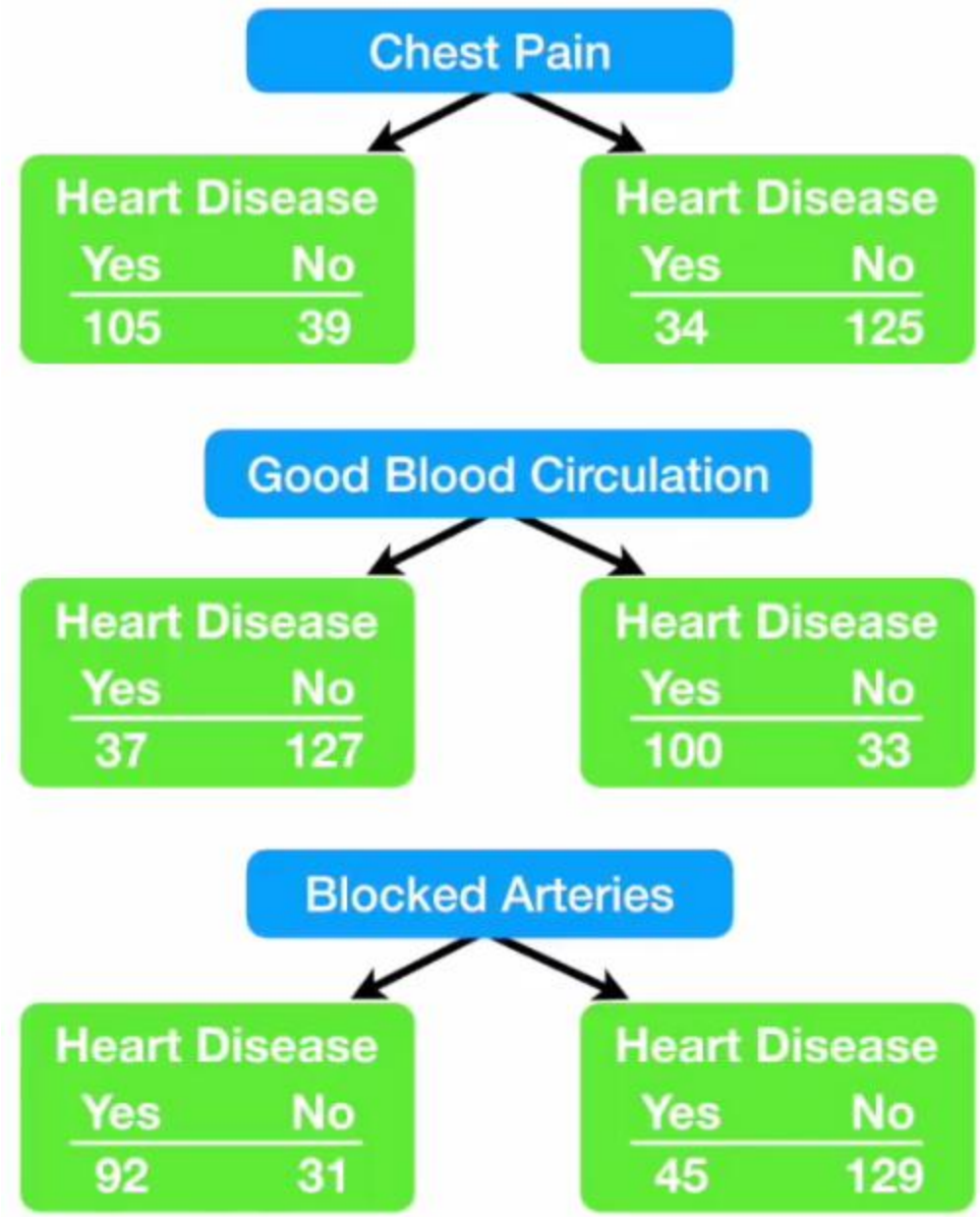
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



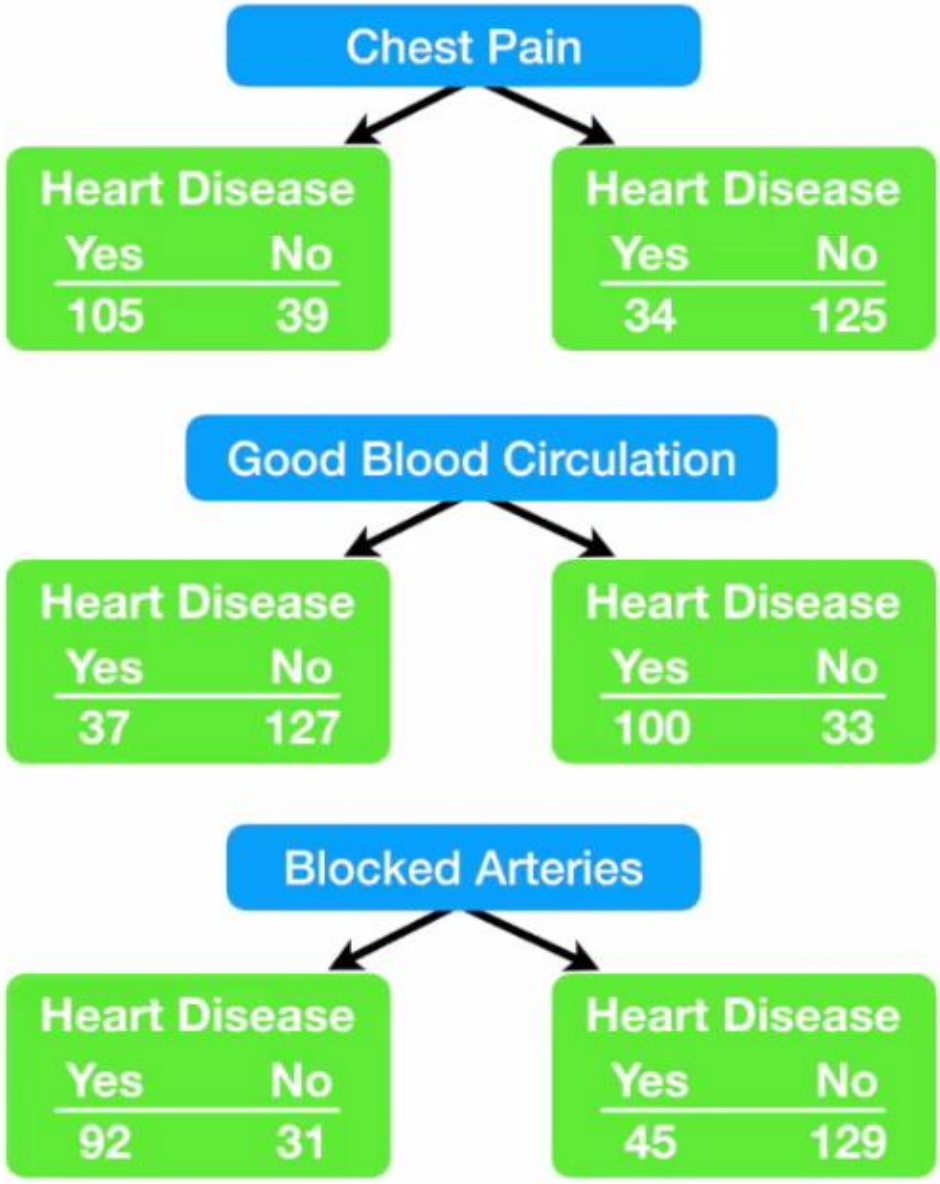
goal is to decide whether **Chest Pain**, **Good Blood Circulation** or **Blocked Arteries** should be the **The Root Node**

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…



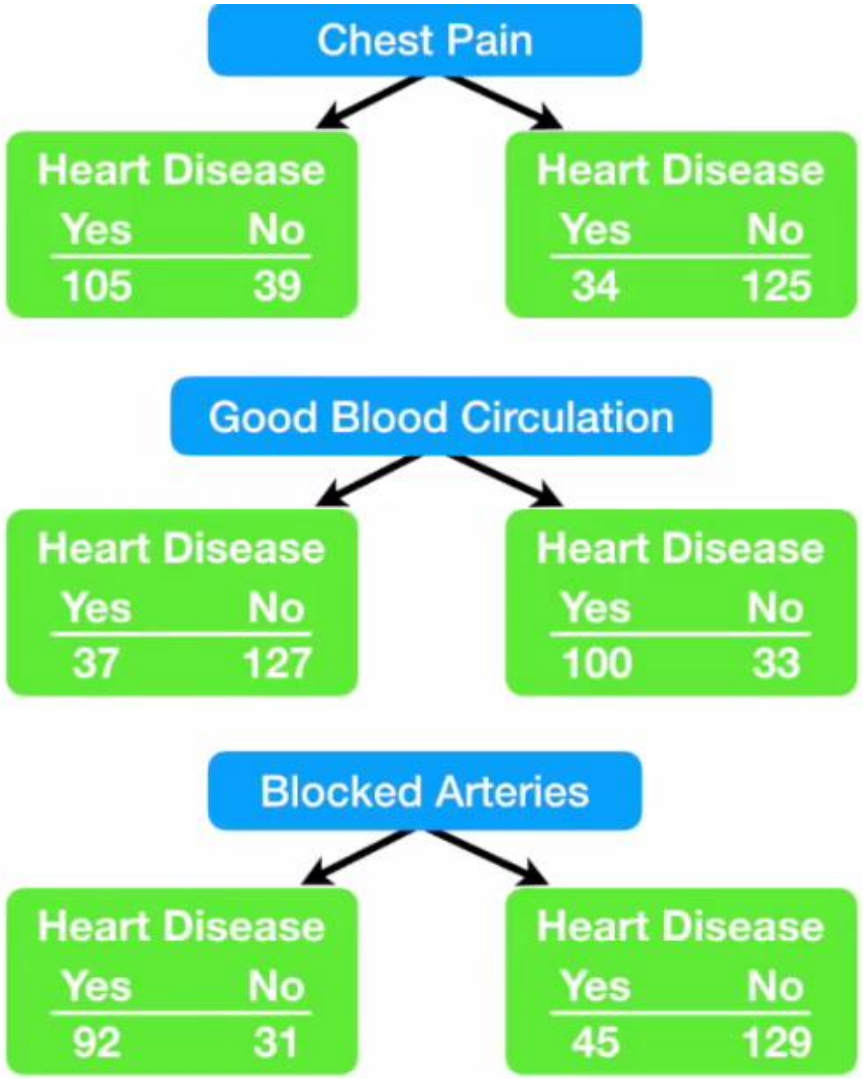

Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
Good Blood Circulation
Heart Disease
Yes No
37 127
Heart Disease
Yes No
100 33
Blocked Arteries
Heart Disease
Yes No
92 31
Heart Disease
Yes No
45 129

Because none of the leaf nodes are 100% “YES Heart Disease” or 100% “NO Heart Disease”, they are all considered “**impure**”.

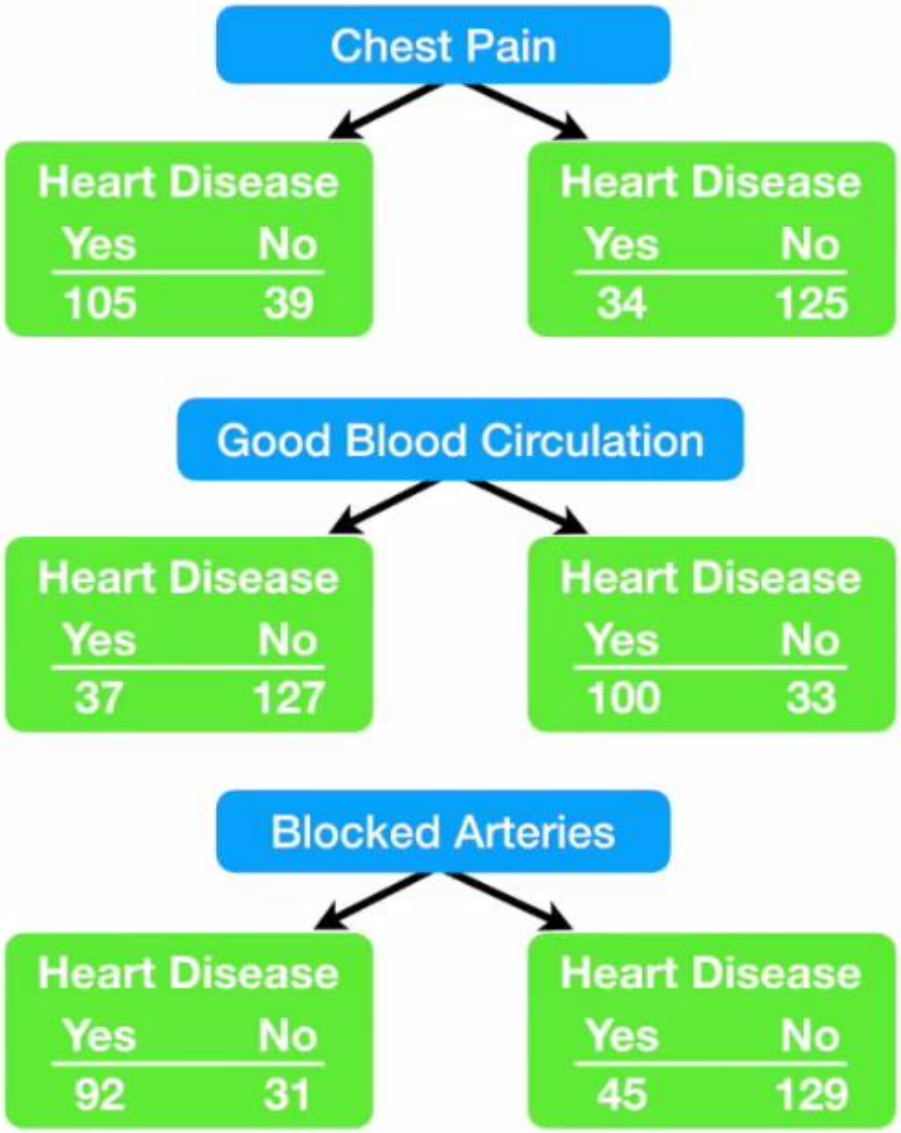


Because none of the leaf nodes are 100% “YES Heart Disease” or 100% “NO Heart Disease”, they are all considered “**impure**”.

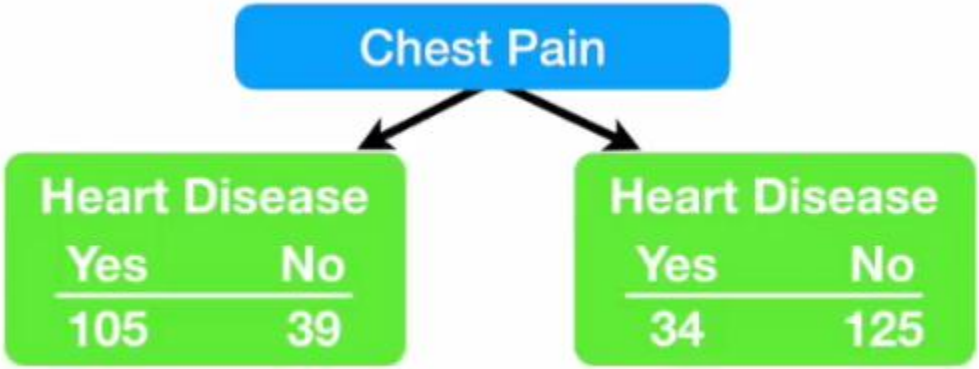
To determine which separation is best, we need a way to measure and compare “**impurity**”.

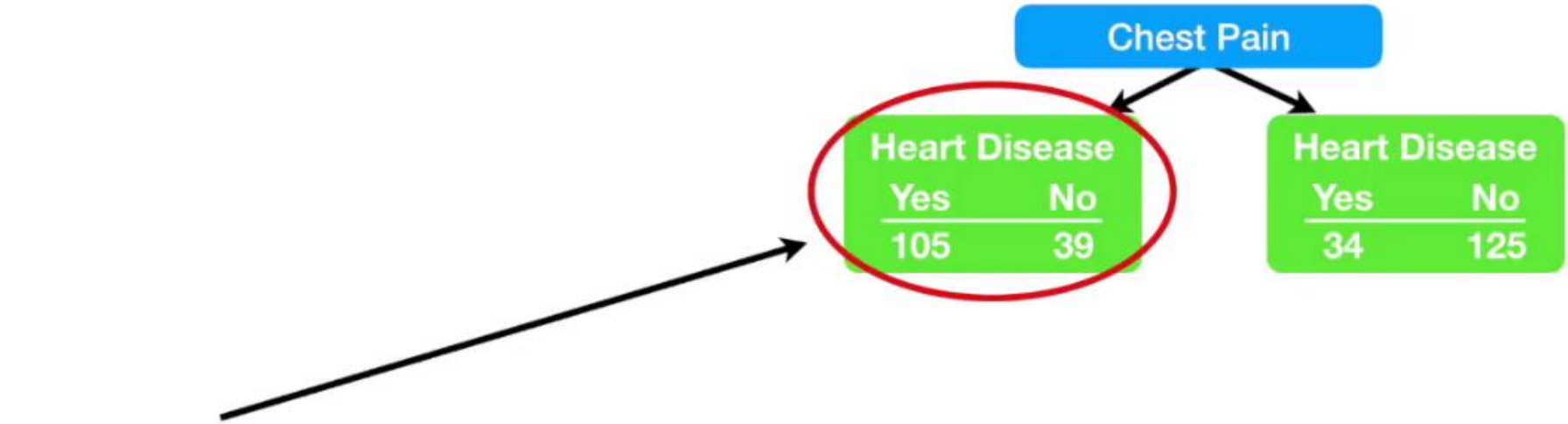


There are a bunch of ways to measure impurity, but I'm just going to focus on a very popular one called "**Gini**".

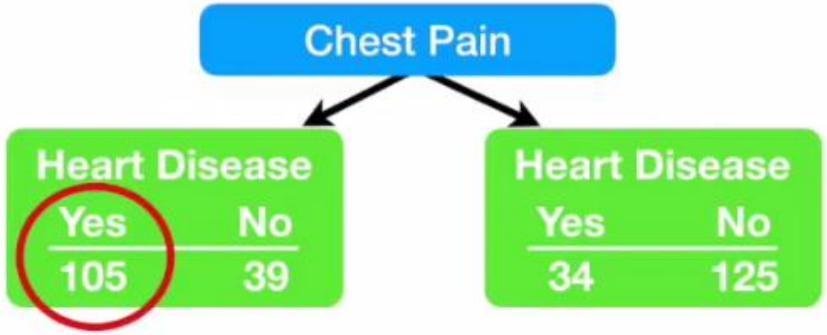


Let's start by calculating Gini impurity for Chest Pain…



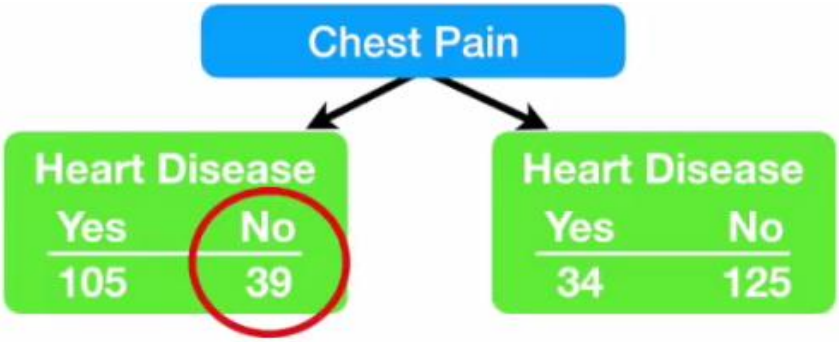


For this leaf, the Gini impurity = 1 - (the probability of “yes”)^2 - (the probability of “no”)^2



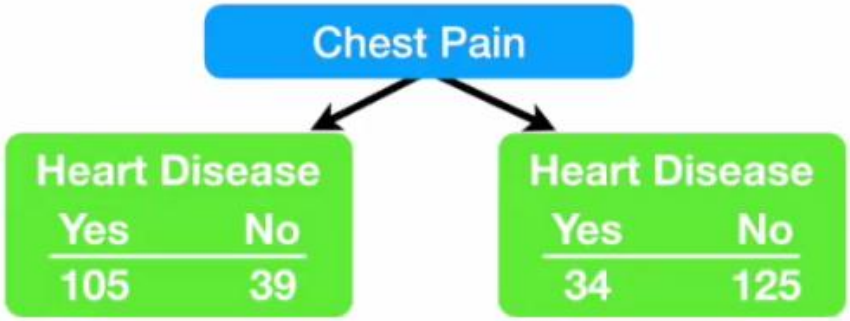
For this leaf, the Gini impurity = 1 - (the probability of "yes")^2 - (the probability of "no")^2

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



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

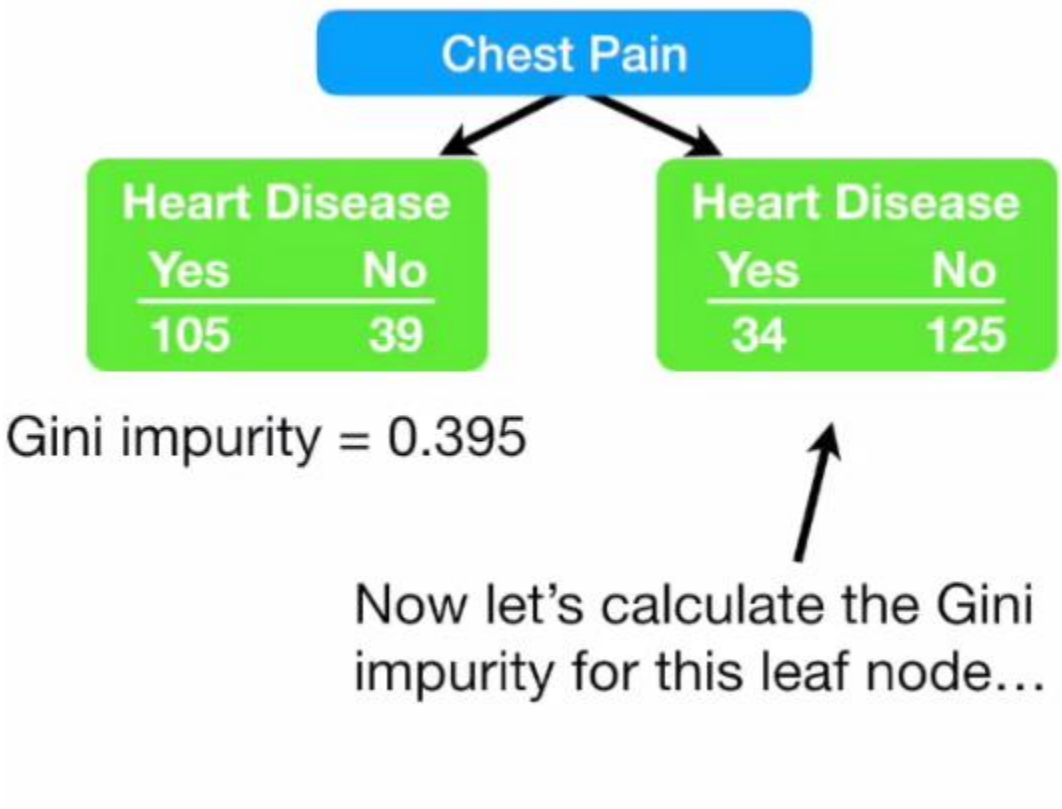
$$= 1 - \left(\frac{105}{105 + 39}\right)^2 - \left(\frac{39}{105 + 39}\right)^2$$

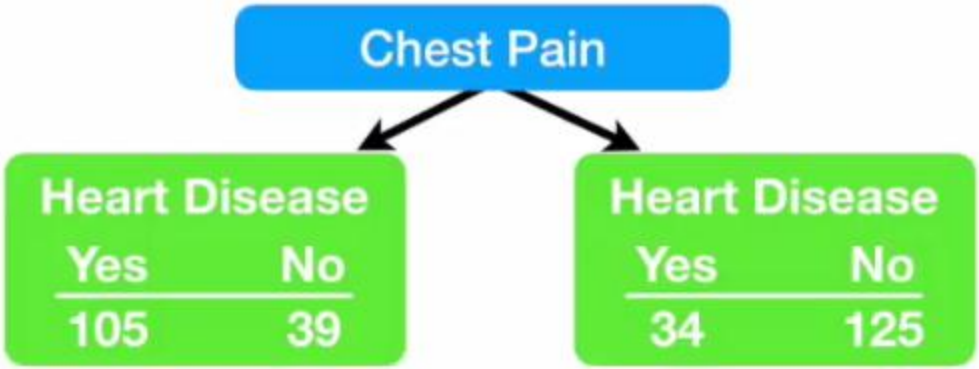


For this leaf, the Gini impurity = 1 - (the probability of "yes")^2 - (the probability of "no")^2

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

$$= 0.395$$

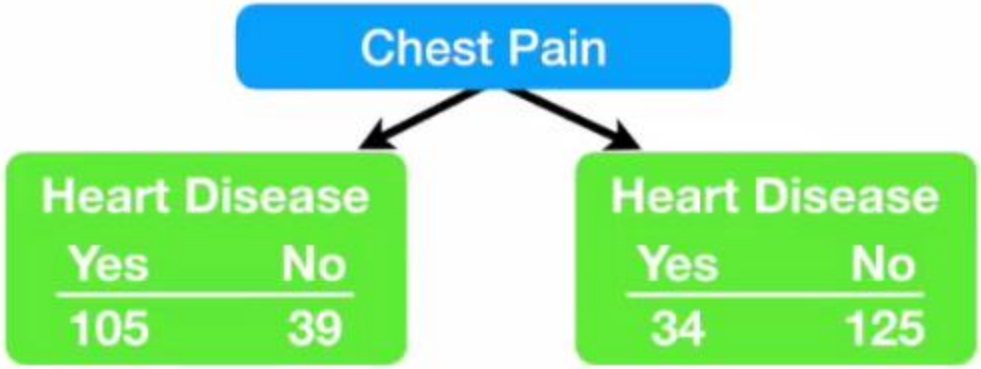

Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Gini impurity = 0.395
Now let's calculate the Gini impurity for this leaf node…



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

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

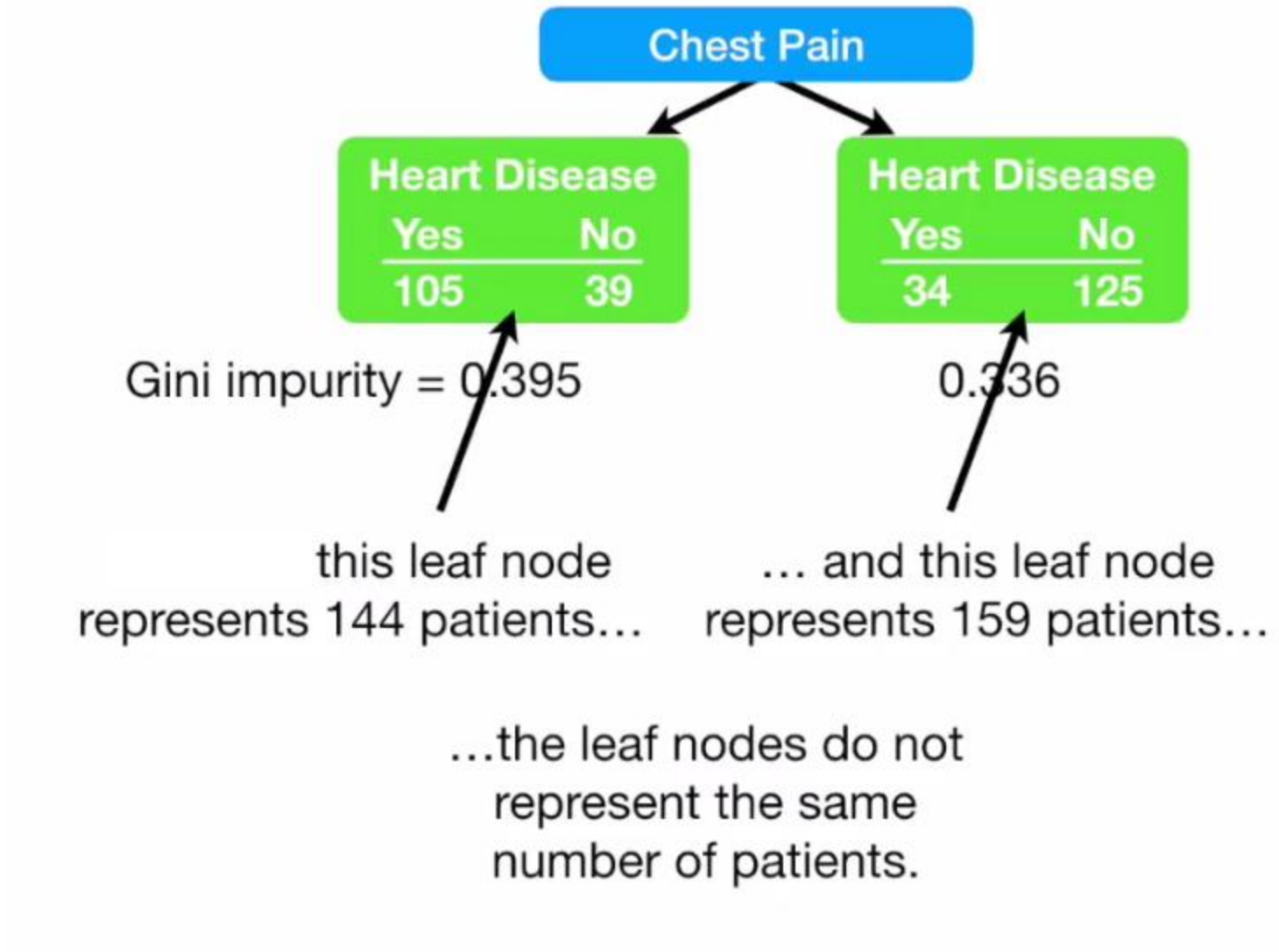
$= 0.336$

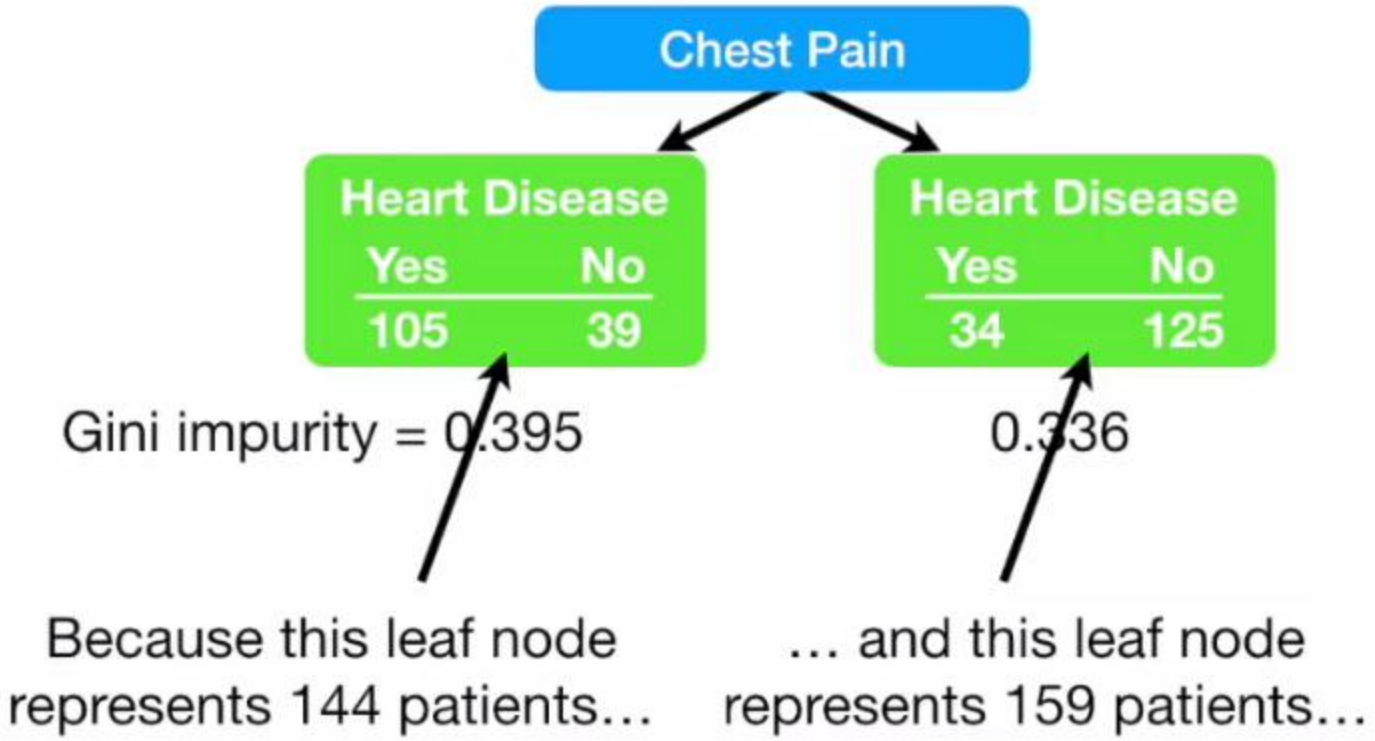


Gini impurity = 0.395

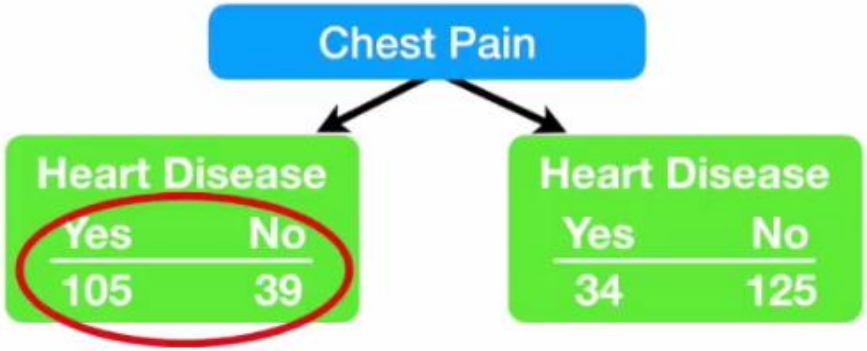
0.336

Now that we have measured the Gini impurity for both leaf nodes, we can calculate the total Gini impurity for using Chest Pain to separate patients with and without heart disease.


Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Gini impurity = 0.395
0.336
this leaf node
represents 144 patients…
… and this leaf node
represents 159 patients…
…the leaf nodes do not
represent the same
number of patients.



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**.

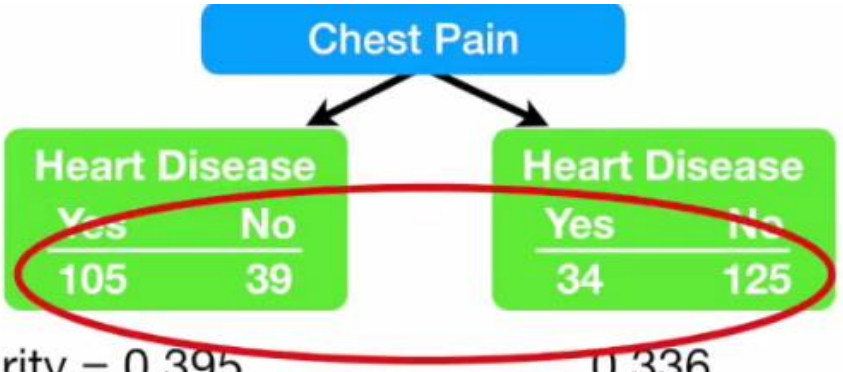


Gini impurity = 0.395

0.336

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

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

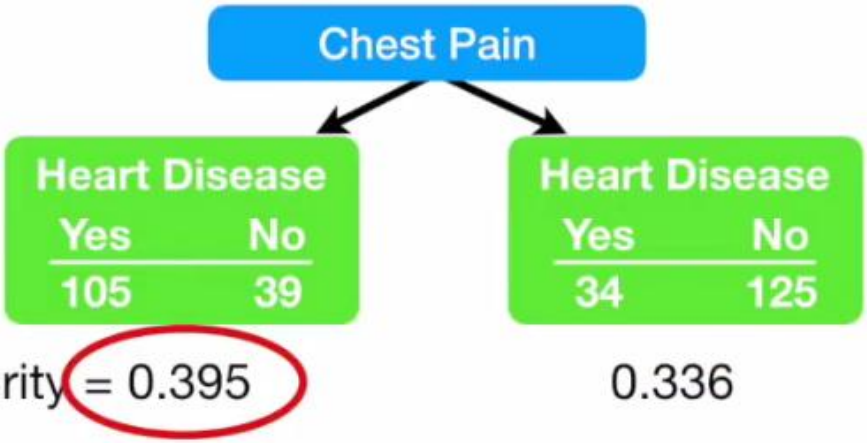


Gini impurity = 0.395

0.336

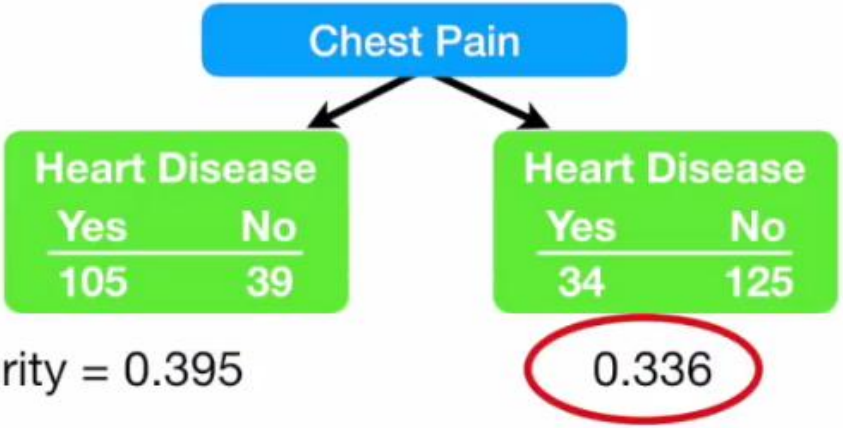
Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes

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



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

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

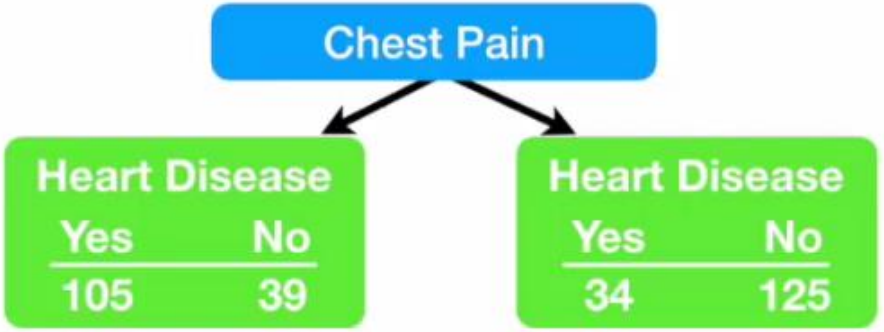


Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes

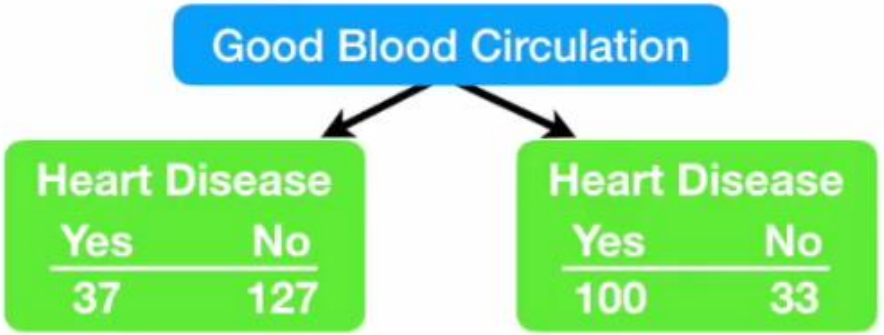
$$= \left(\frac{144}{144 + 159}\right) 0.395 + \left(\frac{159}{144 + 159}\right) 0.336$$

$$= 0.364$$

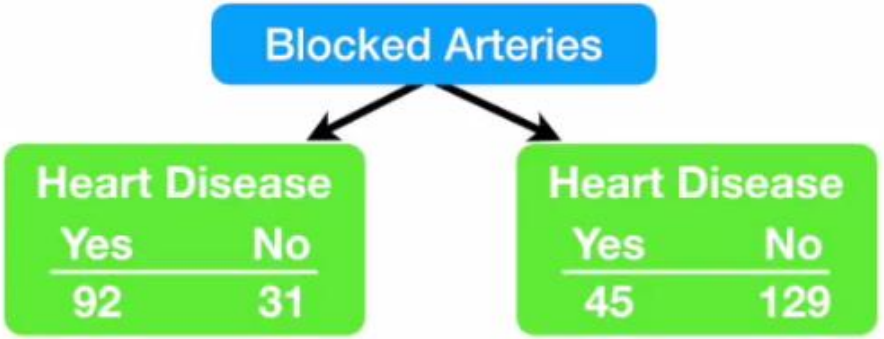
Gini impurity for Chest Pain = 0.364



Gini impurity for Good Blood Circulation = 0.360



Gini impurity for Blocked Arteries = 0.381



Gini impurity for Chest Pain = 0.364

Gini impurity for Good Blood Circulation = 0.360

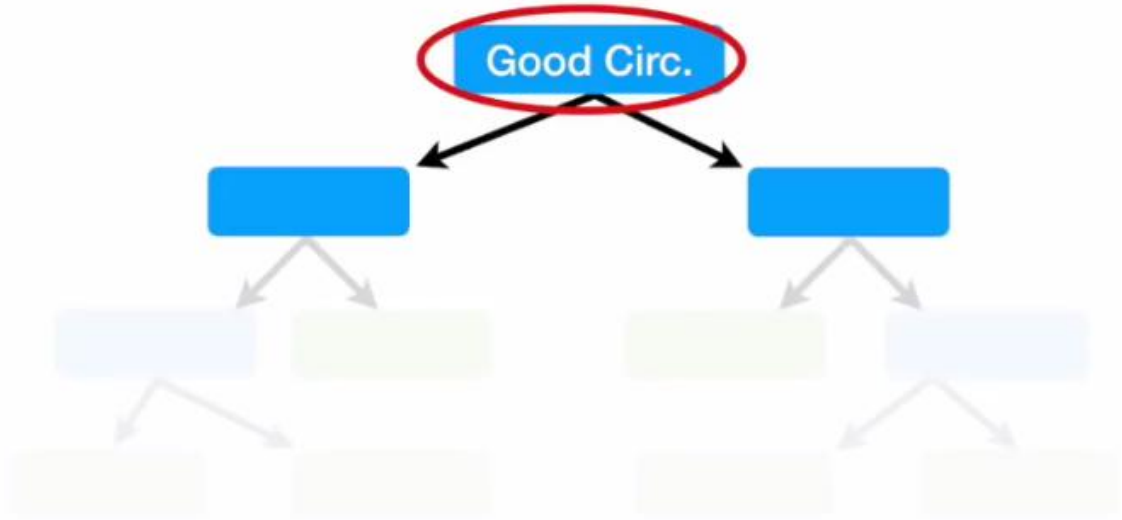
Good Blood Circulation has the lowest impurity (it separates patients with and without heart disease the best)…

Gini impurity for Blocked Arteries = 0.381

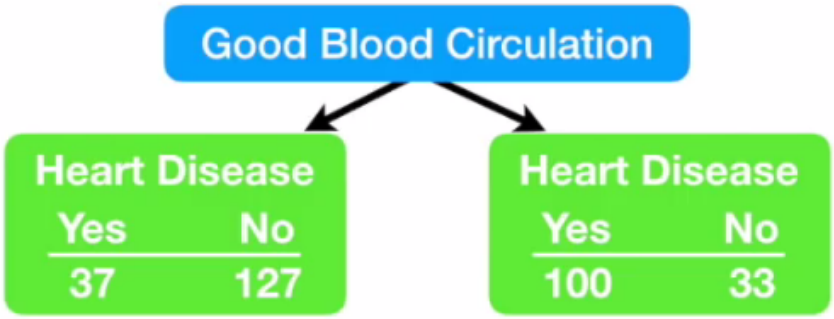
Gini impurity for Chest Pain = 0.364

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

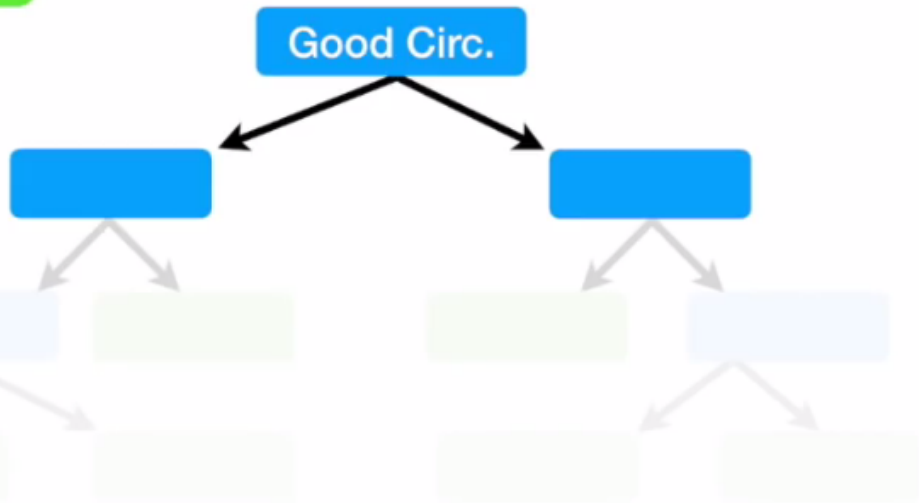
Gini impurity for Good Blood Circulation = 0.360

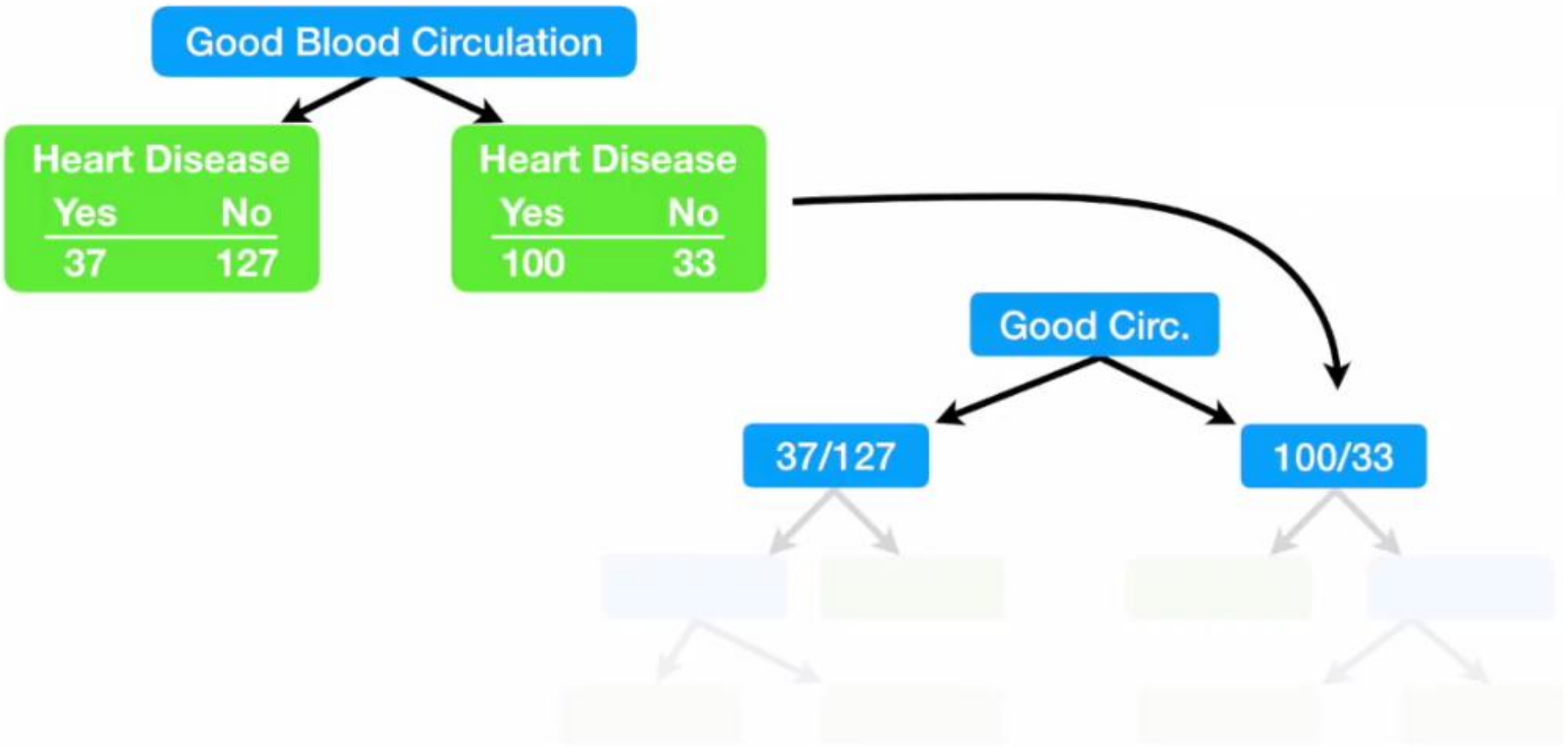


Gini impurity for Blocked Arteries = 0.381

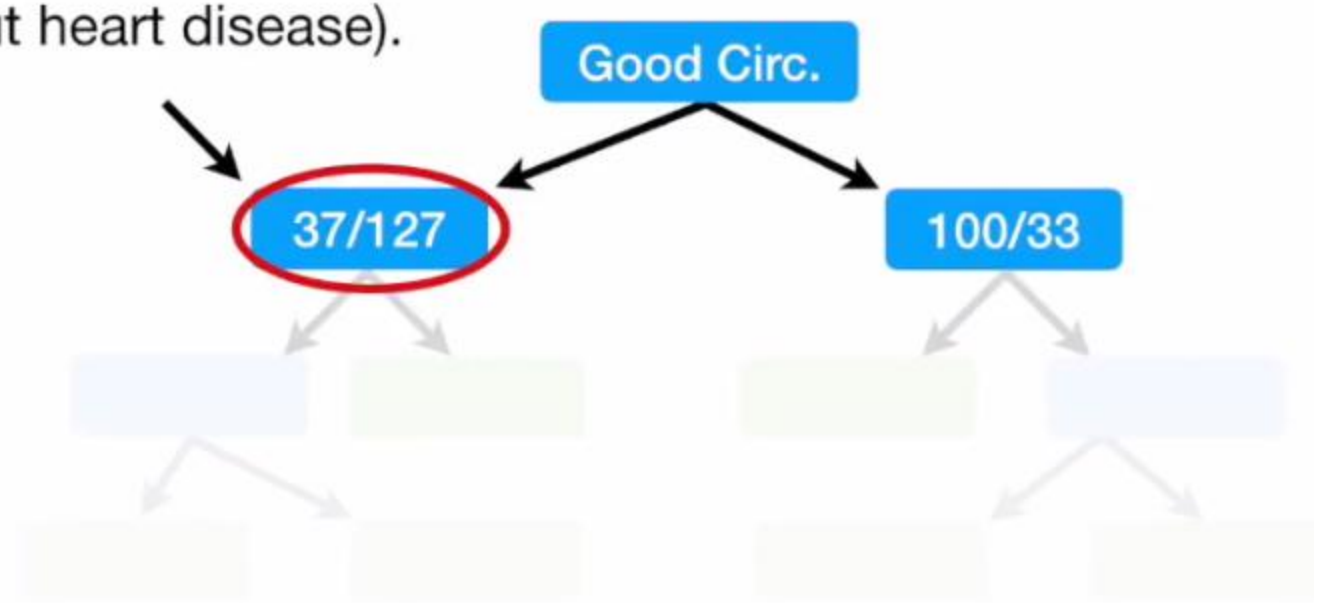


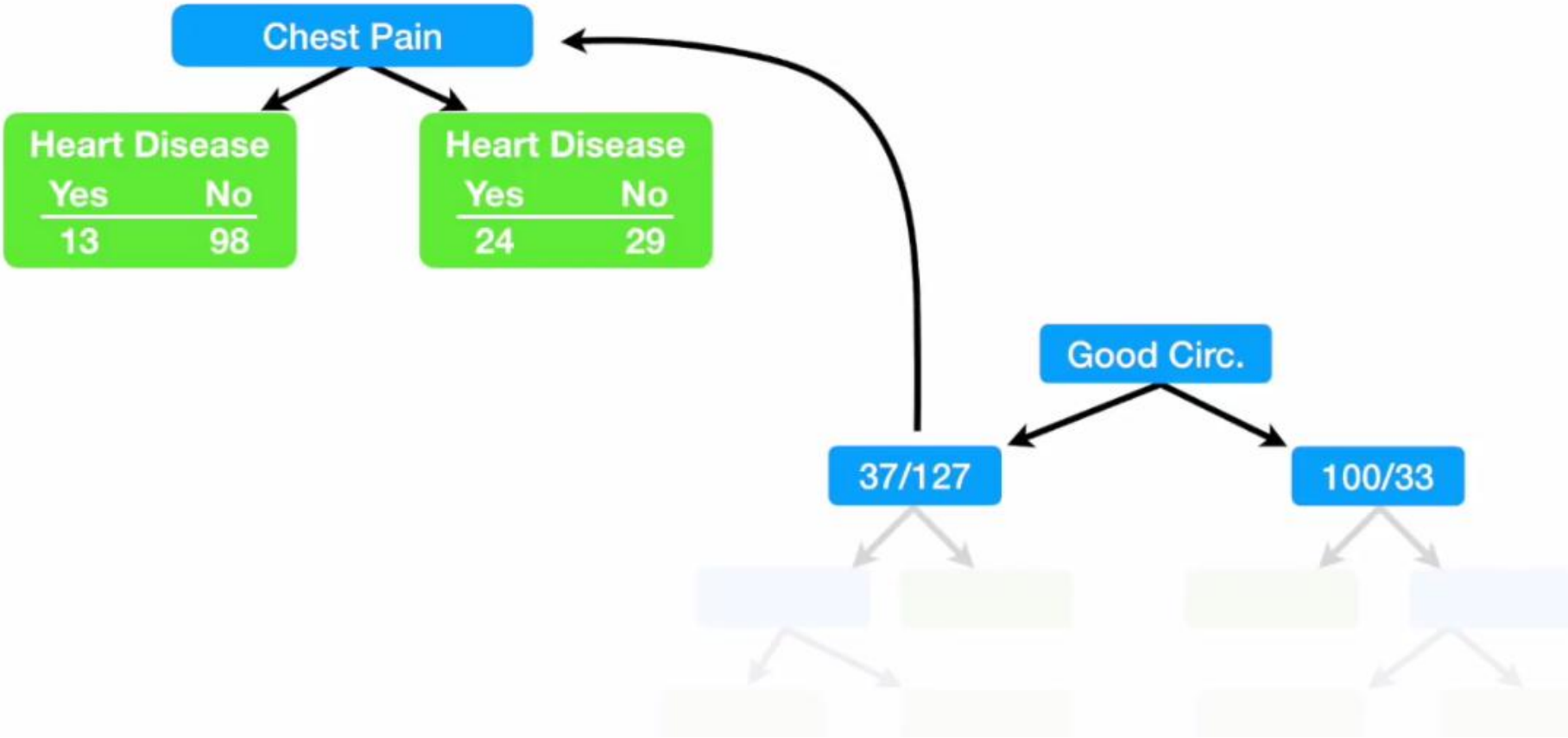
When we divided all of the patients using **Good Blood Circulation**, we ended up with "impure" leaf nodes.

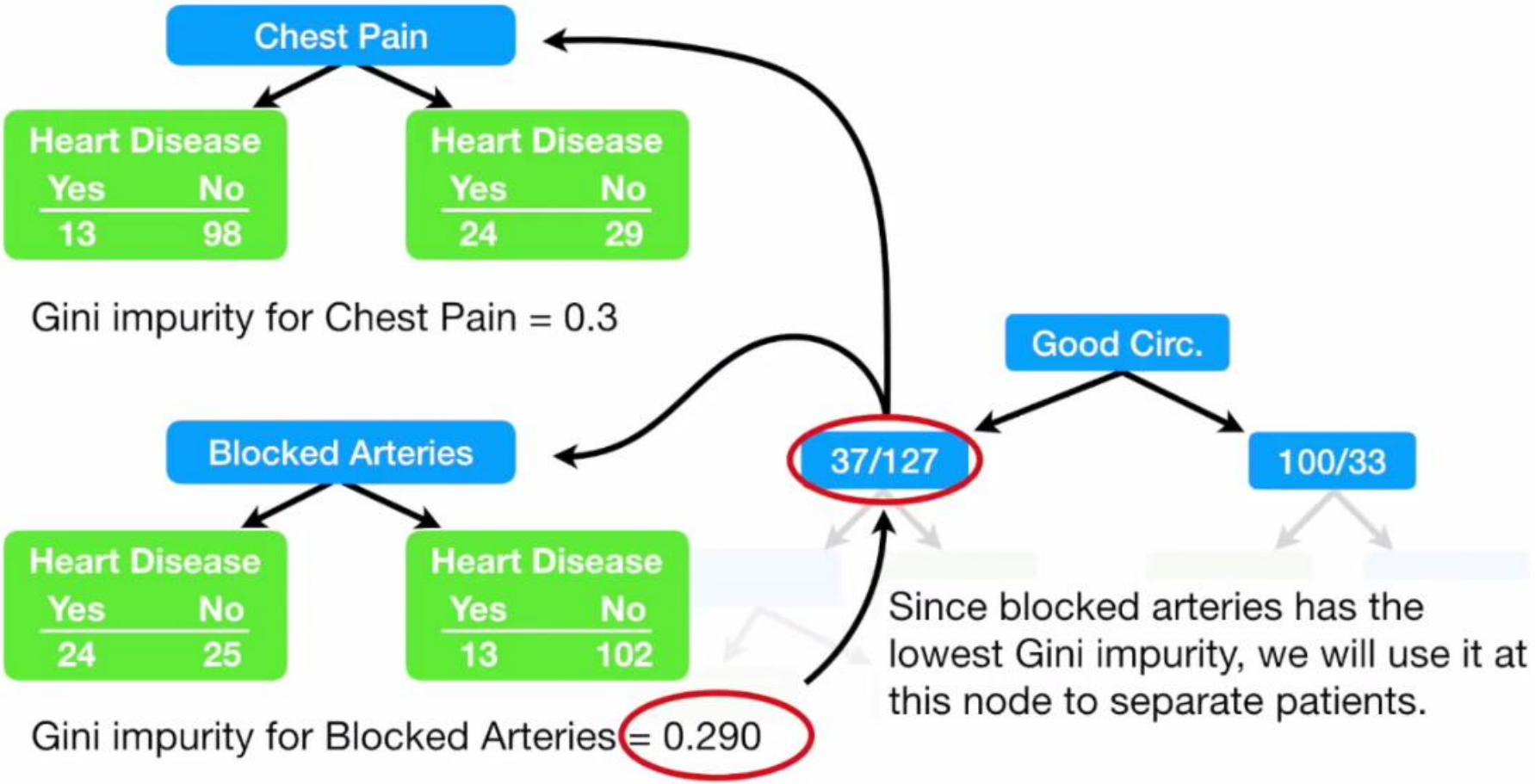



Good Blood Circulation
Heart Disease
Yes
No
37
127
Heart Disease
Yes
No
100
33
Good Circ.
37/127
100/33

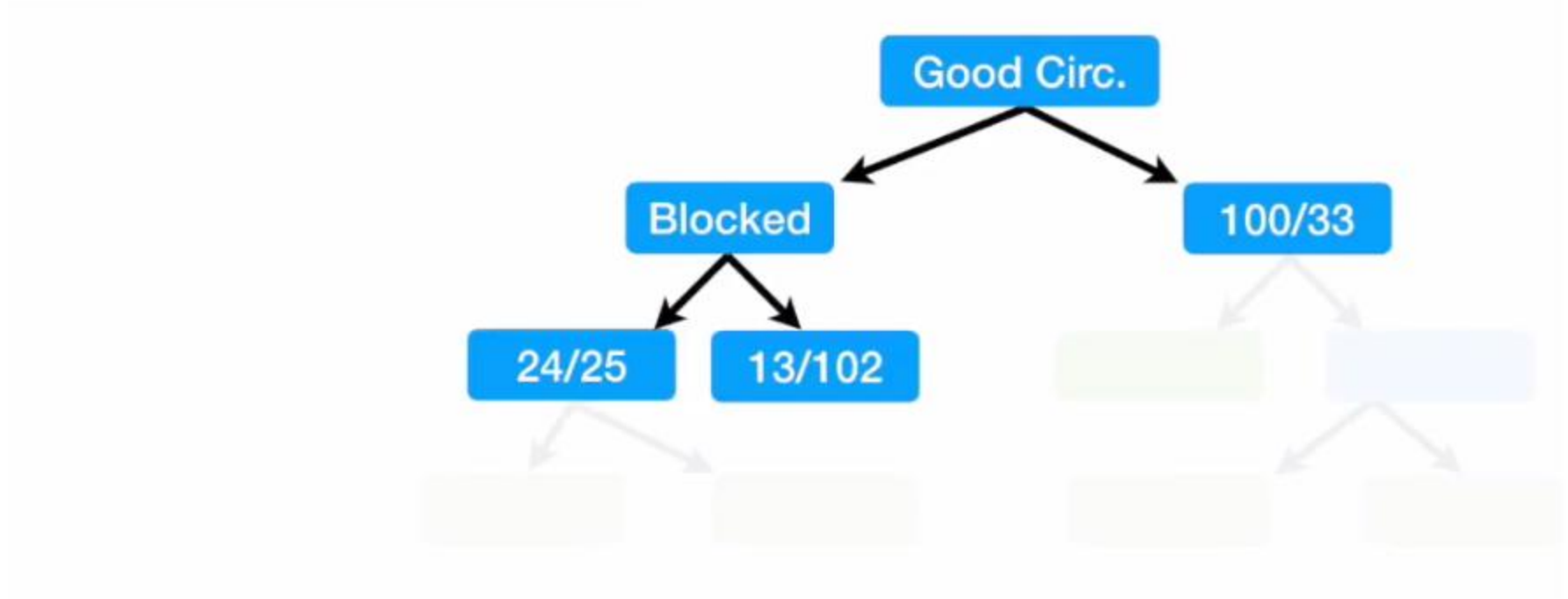
Now we need to figure how well **chest pain** and **blocked arteries** separate these 164 patients (37 with heart disease and 127 without heart disease).




Chest Pain
Heart Disease
Yes
No
13
98
Heart Disease
Yes
No
24
29
Good Circ.
37/127
100/33

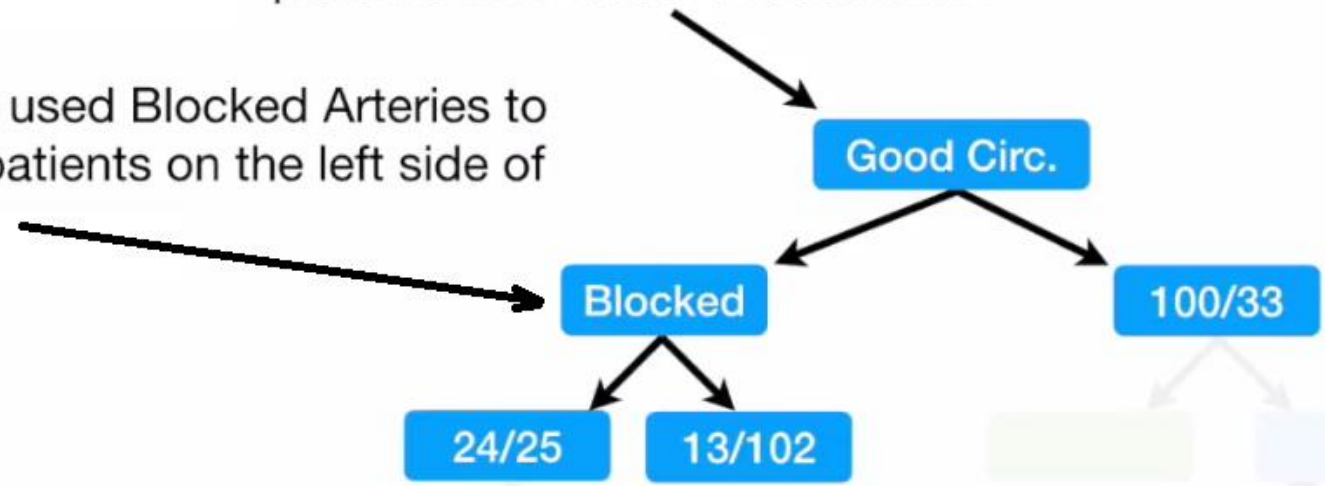

Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Gini impurity for Chest Pain = 0.3
Blocked Arteries
Heart Disease
Yes No
24 25
Heart Disease
Yes No
13 102
Gini impurity for Blocked Arteries = 0.290
Good Circ.
37/127
100/33
Since blocked arteries has the lowest Gini impurity, we will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

So our Tree Looks Like this Now !!

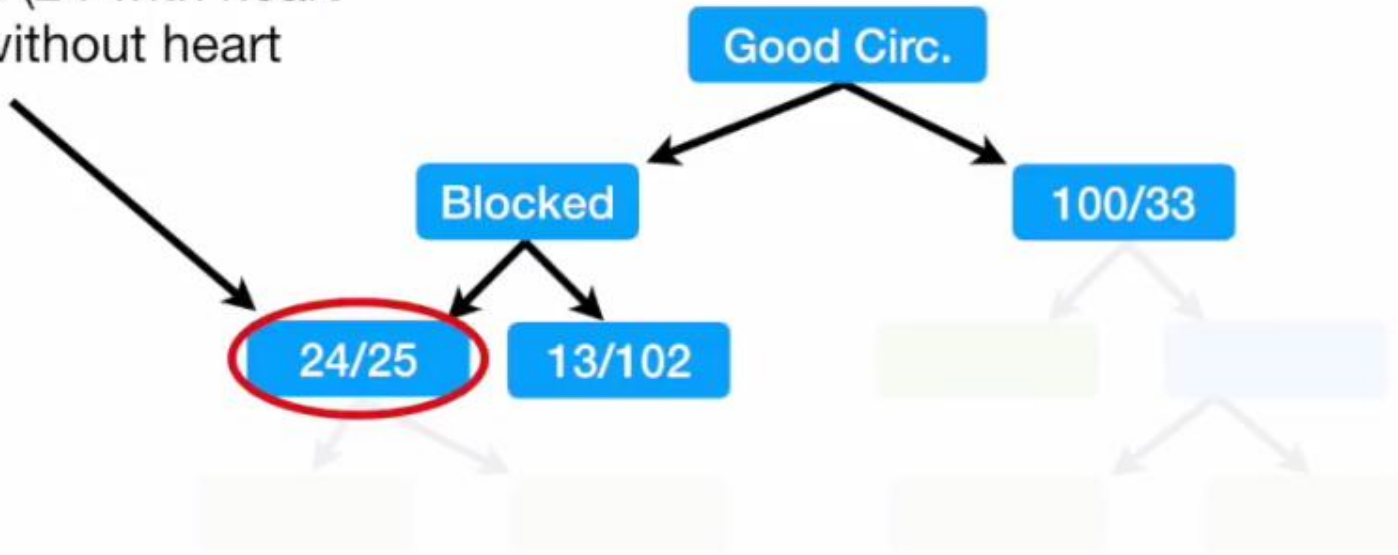


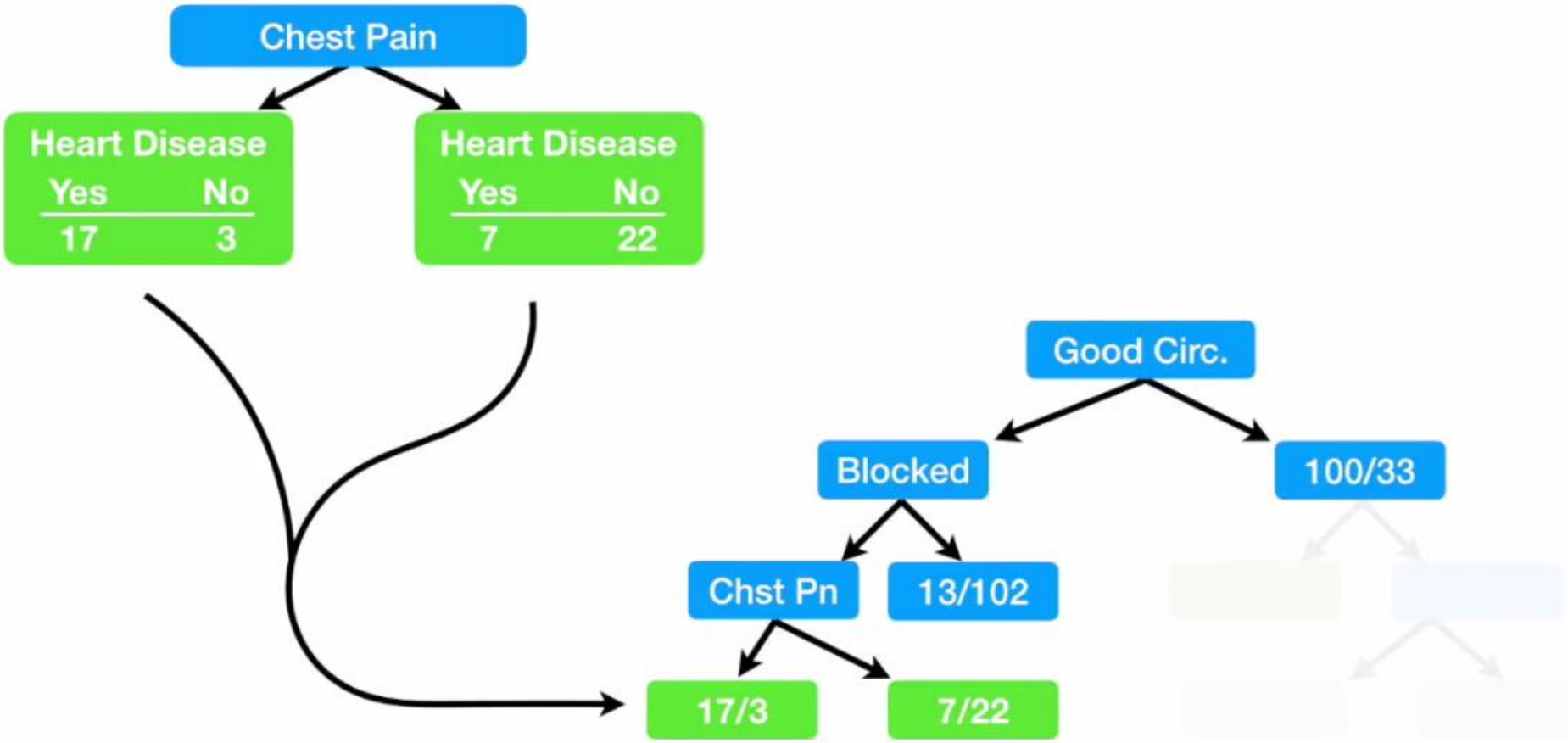
We started at the top by separating patients with Good Circulation…

…then we used Blocked Arteries to separate patients on the left side of the tree.



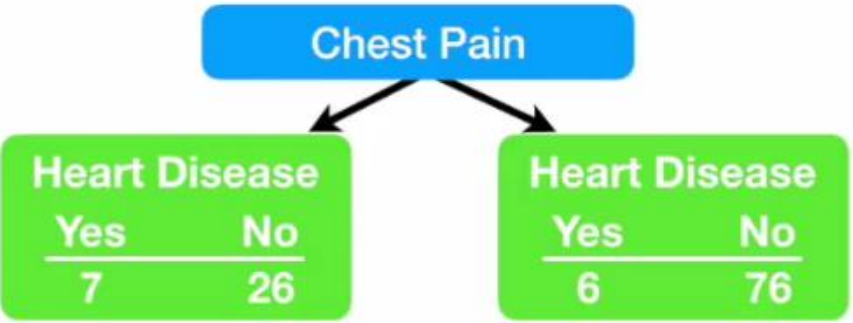
All we have left is Chest Pain, so first we'll see how well it separates these 49 patients (24 with heart disease and 25 without heart disease).



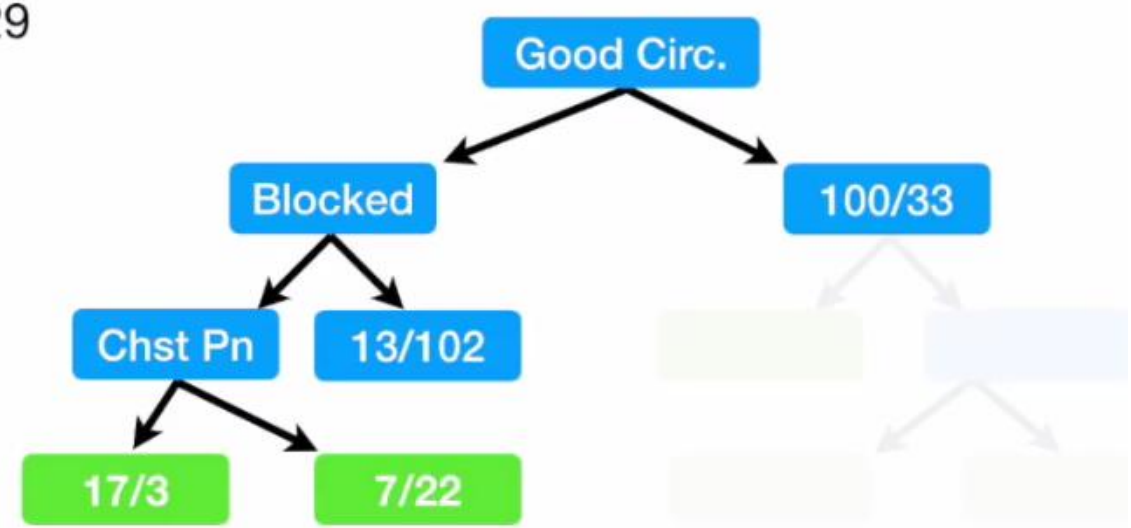


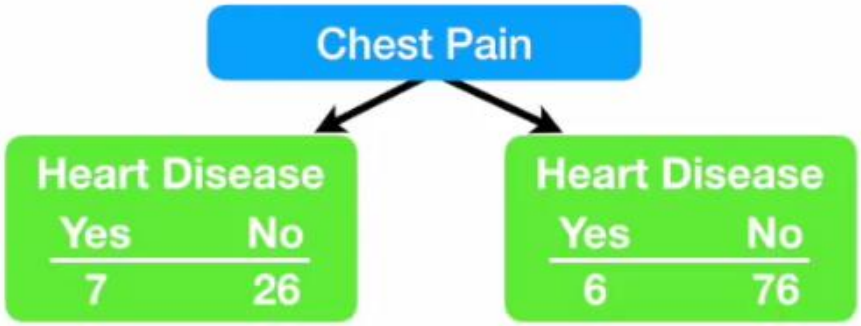
…so these are the final leaf nodes on this branch of the tree.

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



Gini impurity for Chest Pain = 0.29





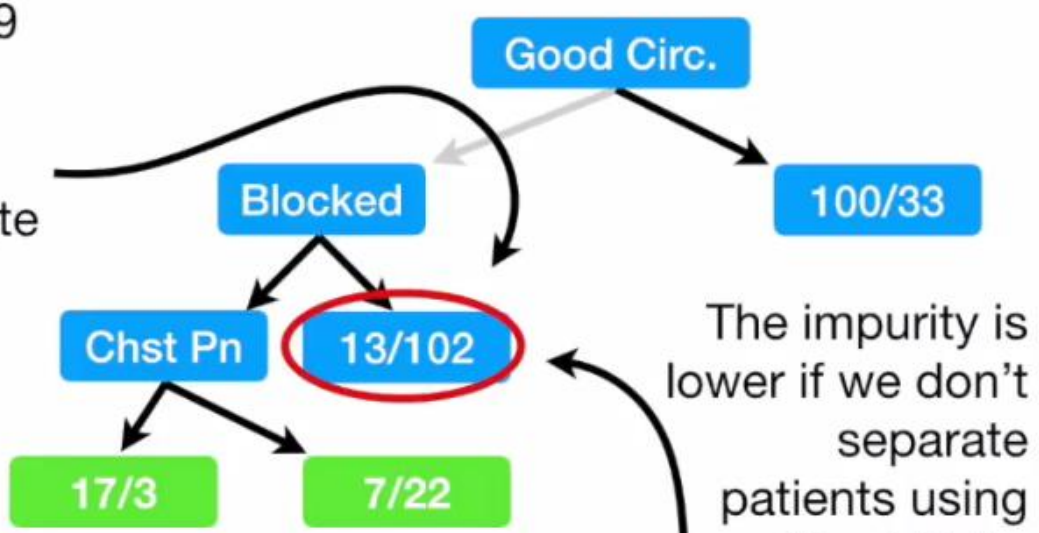
Gini impurity for Chest Pain = 0.29

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

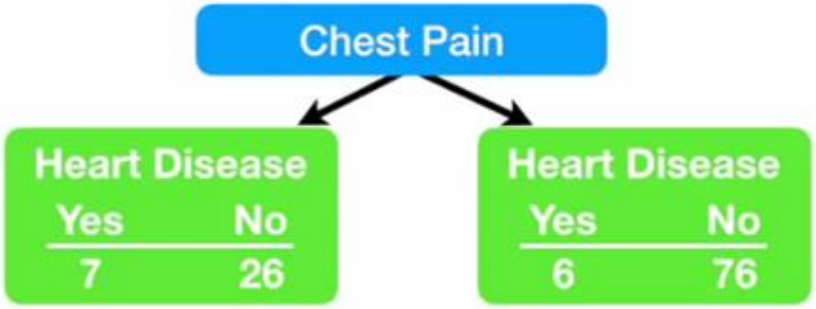
= 1 - (the probability of "yes")^2
- (the probability of "no")^2

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

= 0.2



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



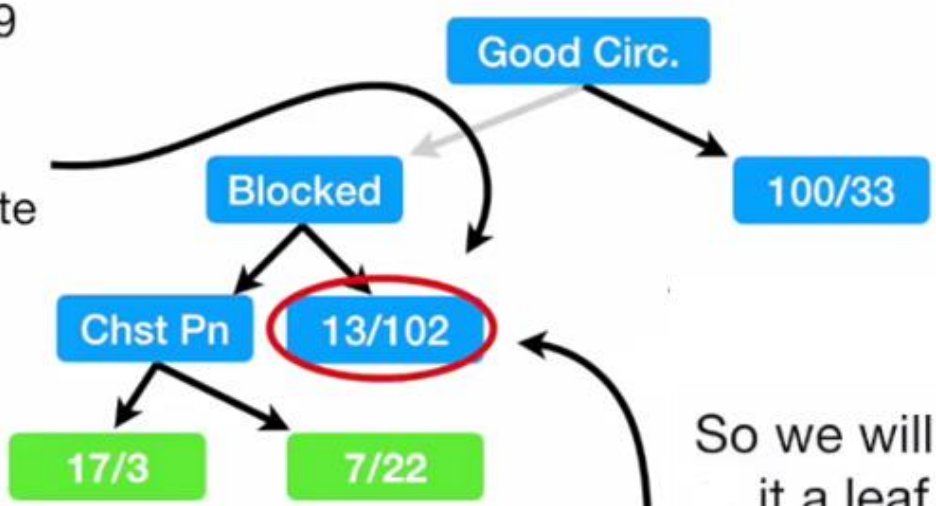
Gini impurity for Chest Pain = 0.29

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

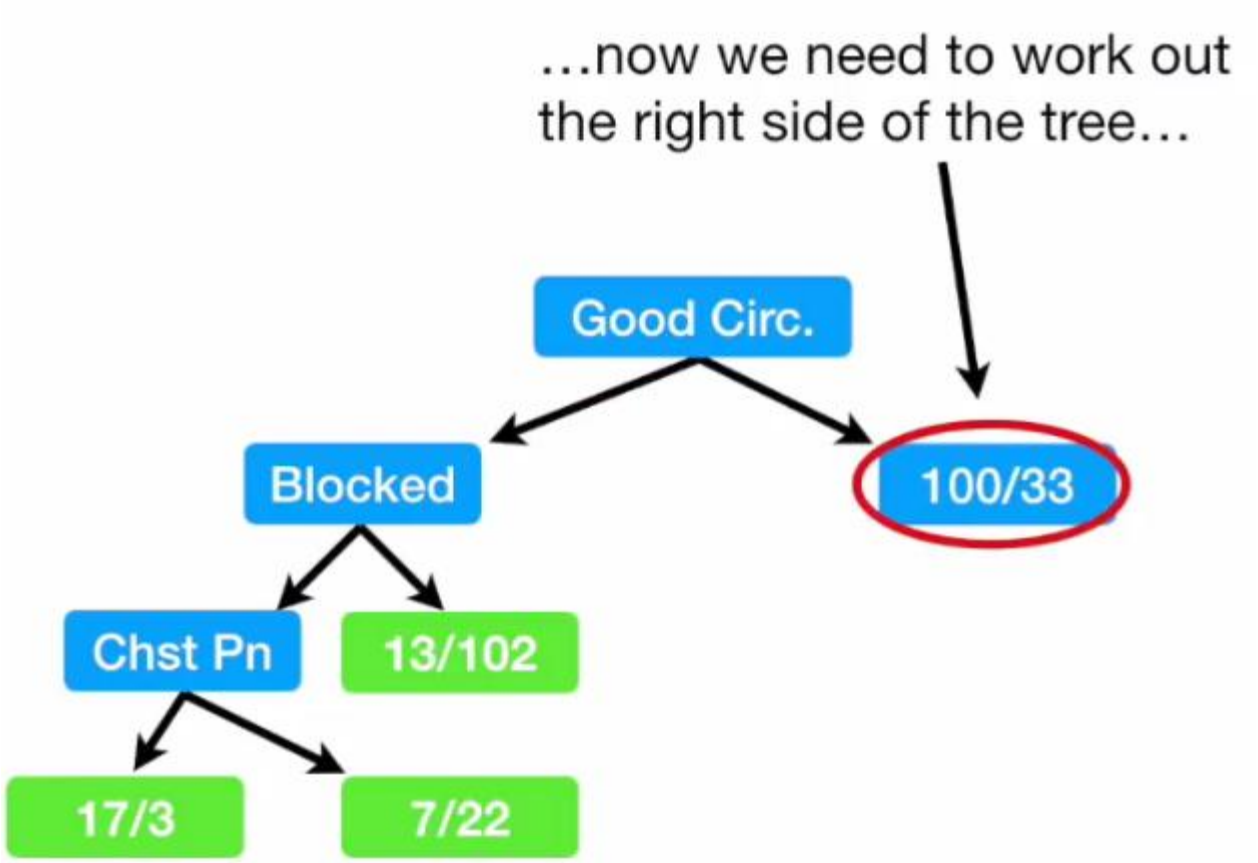
= 1 - (the probability of "yes")^2
- (the probability of "no")^2

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

= 0.2



So we will make it a leaf node.

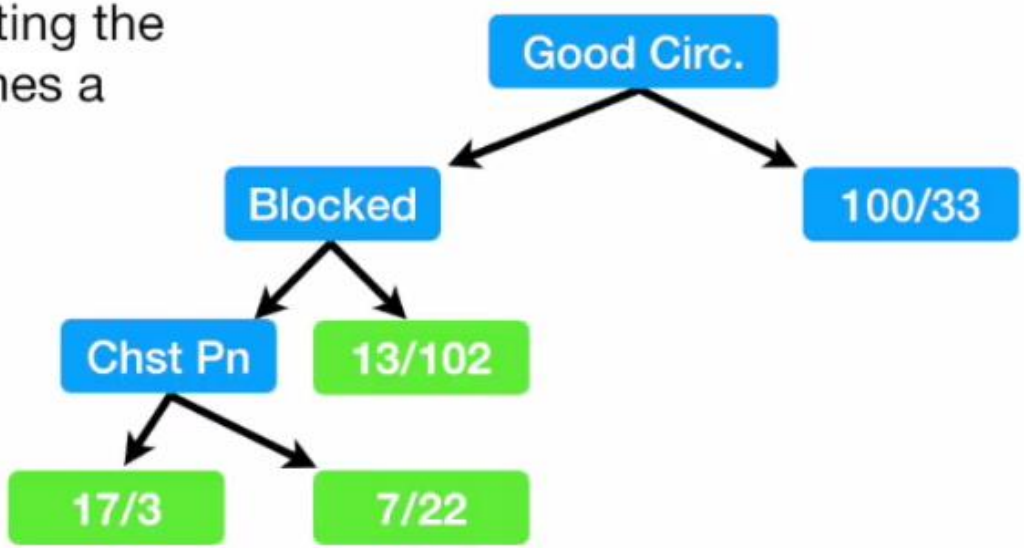

…now we need to work out
the right side of the tree…
Good Circ.
Blocked
100/33
Chst Pn
13/102
17/3
7/22

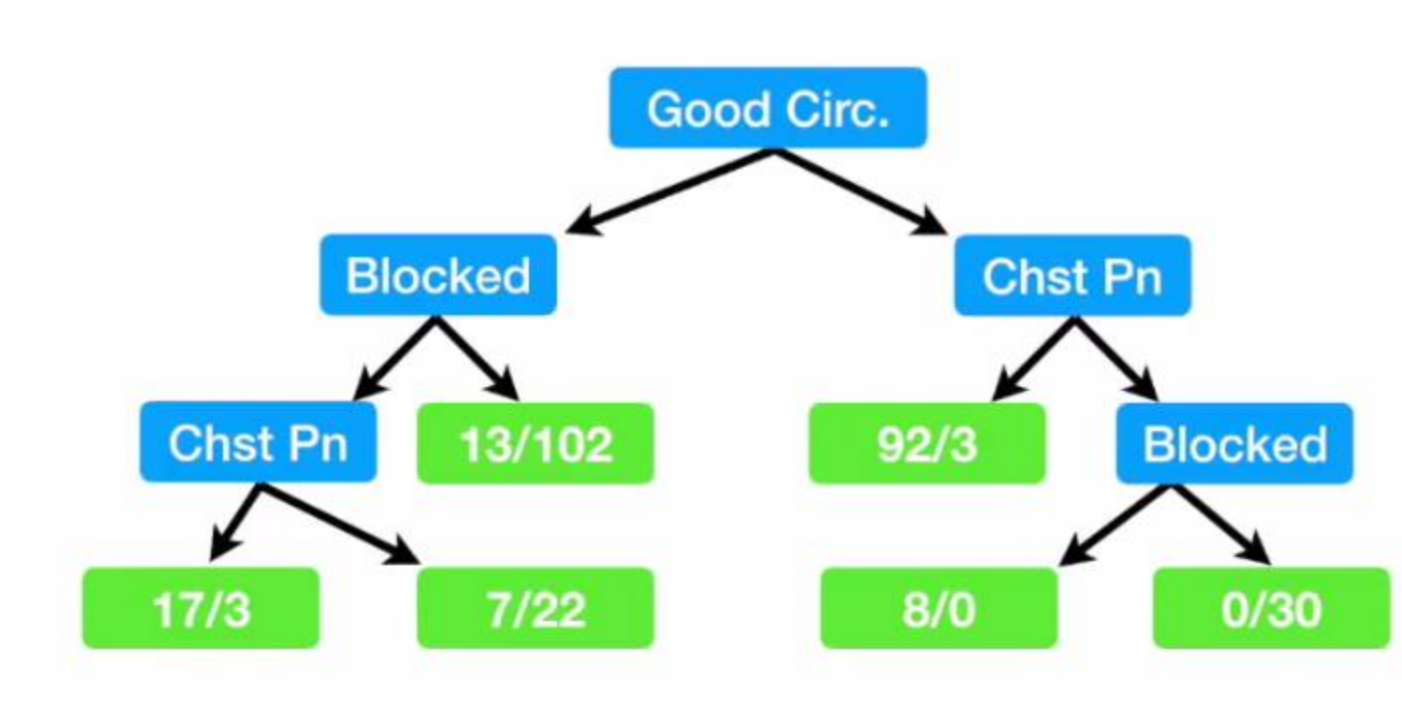
we follow the exact
same steps as we did on the left side:

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.




Good Circ.
Blocked
Chst Pn
Chst Pn
13/102
92/3
Blocked
17/3
7/22
8/0
0/30