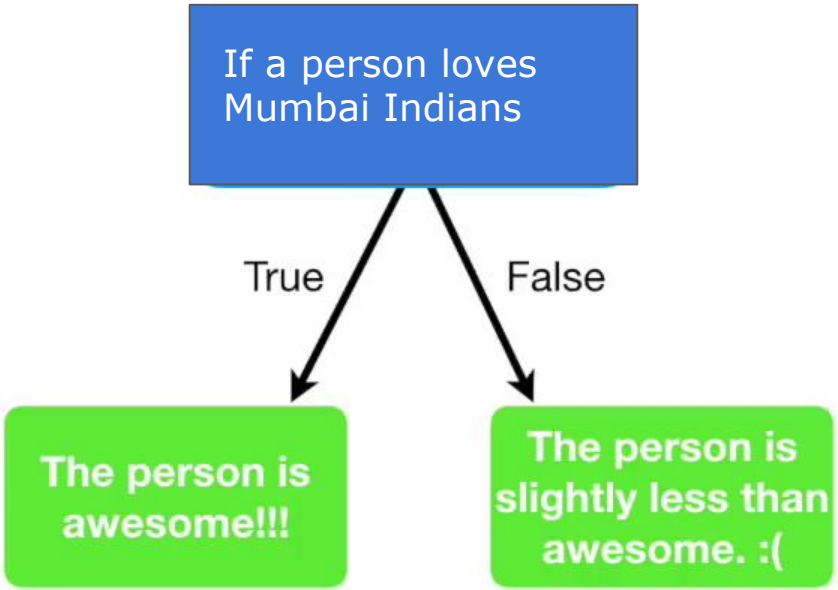
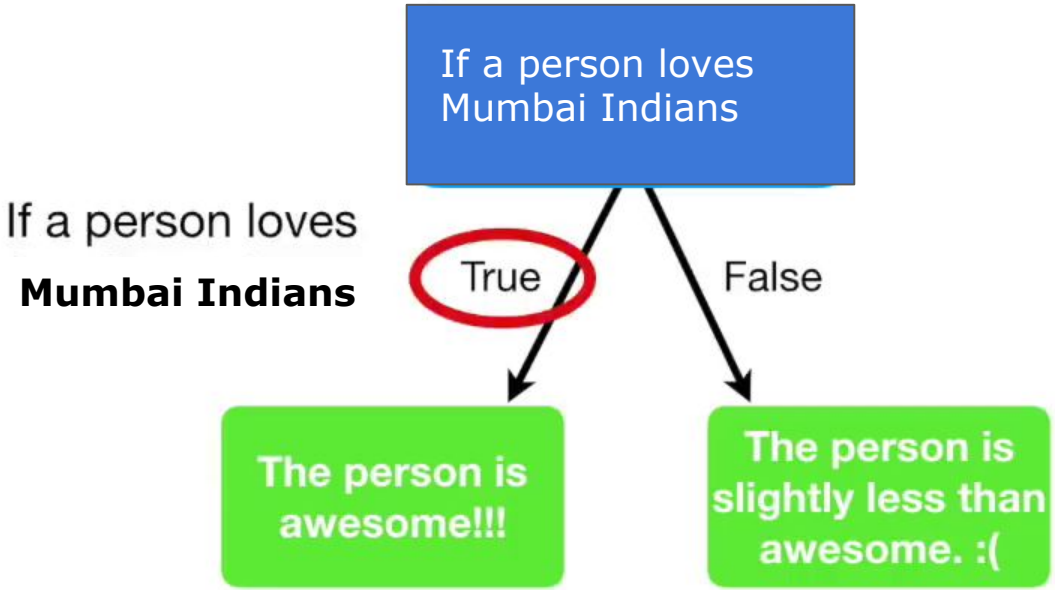
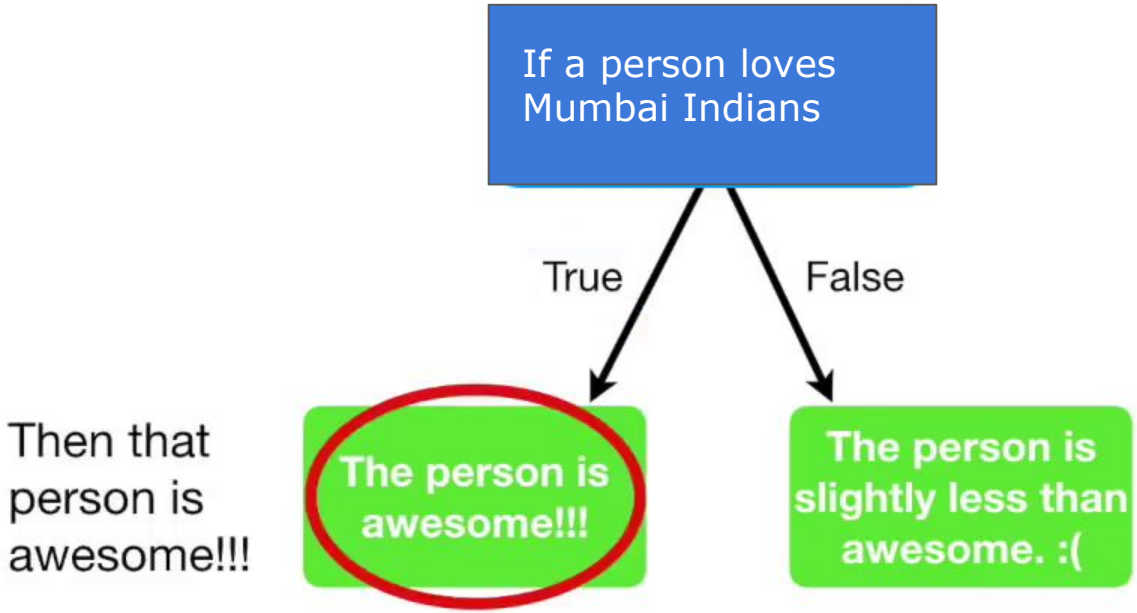
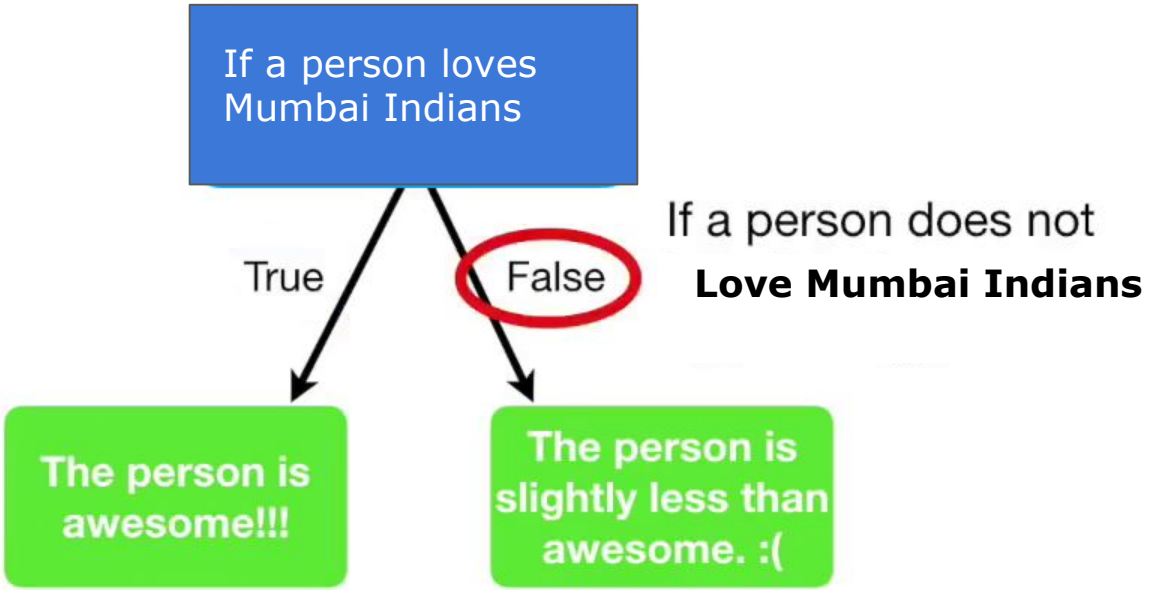
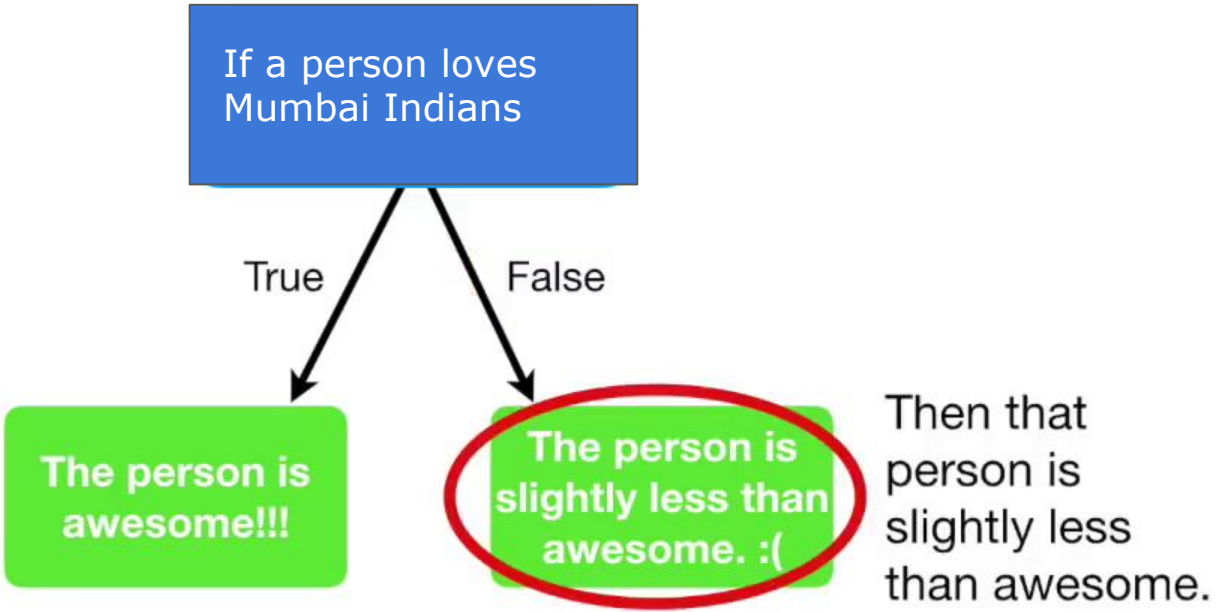
Here's a simple decision tree…



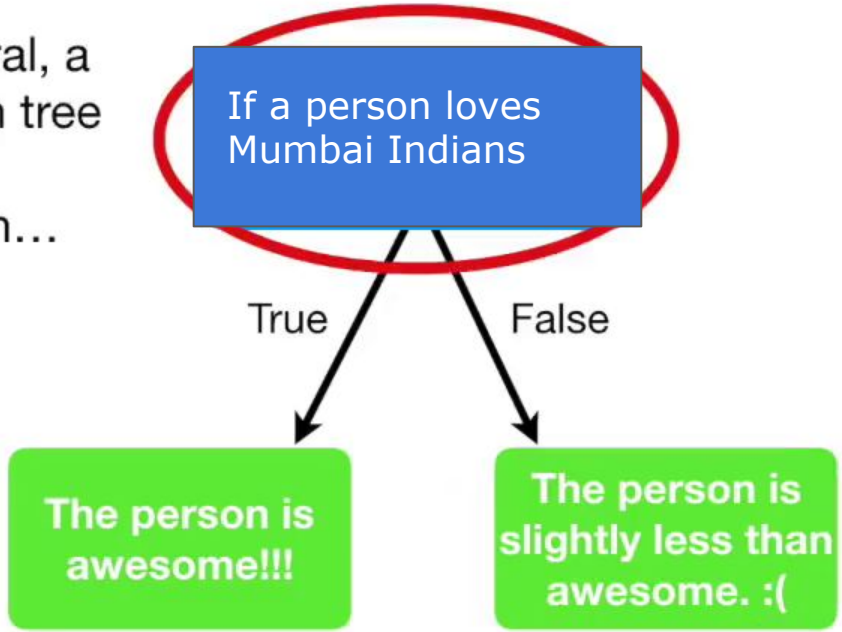

If a person loves
Mumbai Indians
If a person loves
Mumbai Indians
True
False
The person is awesome!!!
The person is slightly less than awesome. :(


If a person loves Mumbai Indians
True
False
Then that person is awesome!!!
The person is awesome!!!
The person is slightly less than awesome. :(

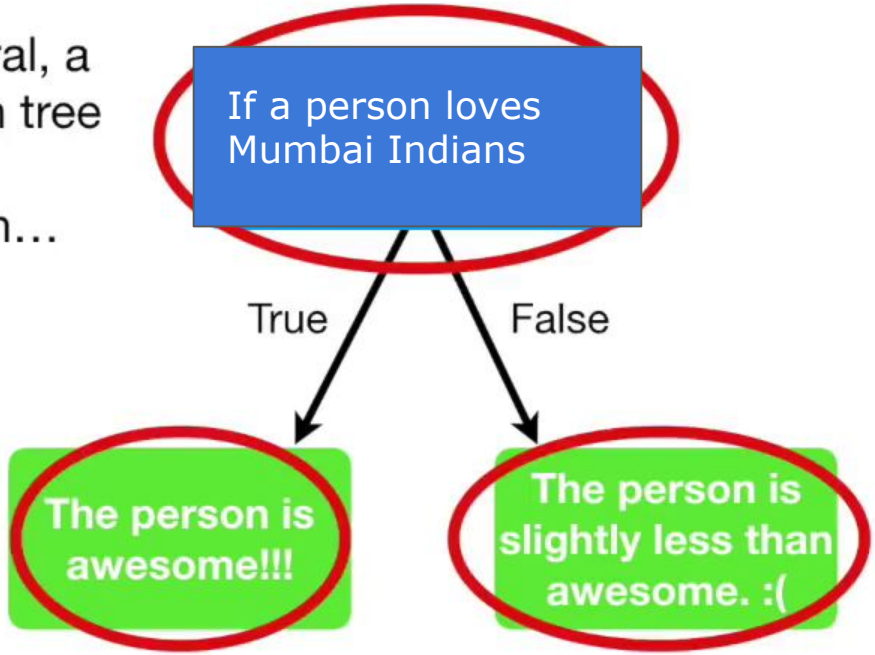

If a person loves Mumbai Indians
True
False
If a person does not
Love Mumbai Indians
The person is awesome!!!
The person is slightly less than awesome. :(


If a person loves Mumbai Indians
True
False
The person is awesome!!!
The person is slightly less than awesome. :(
Then that person is slightly less than awesome.

In general, a decision tree asks a question...

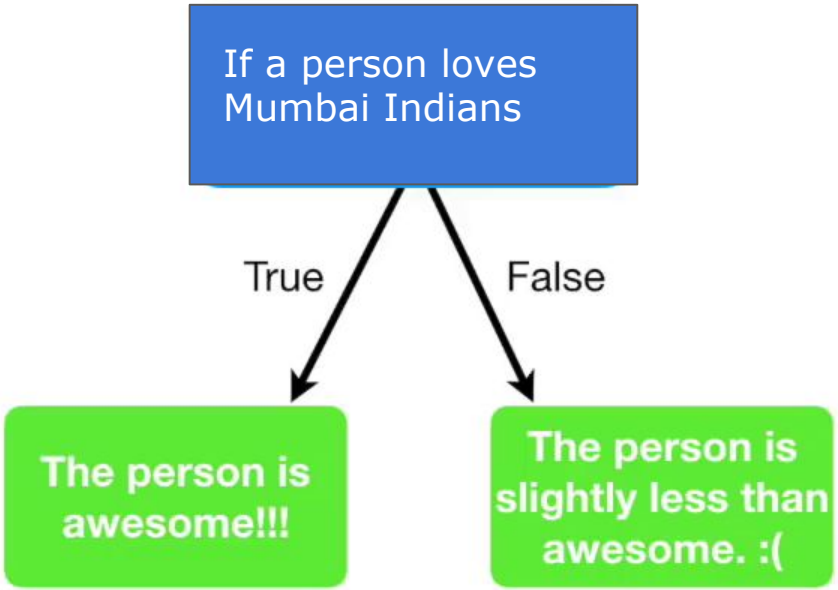


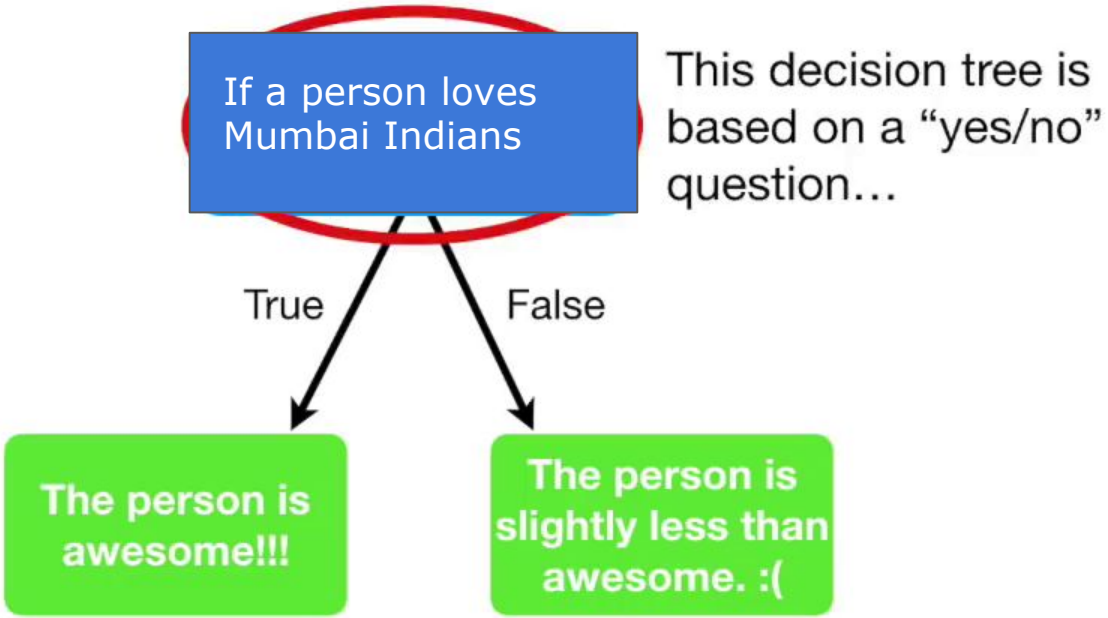
In general, a decision tree asks a question…

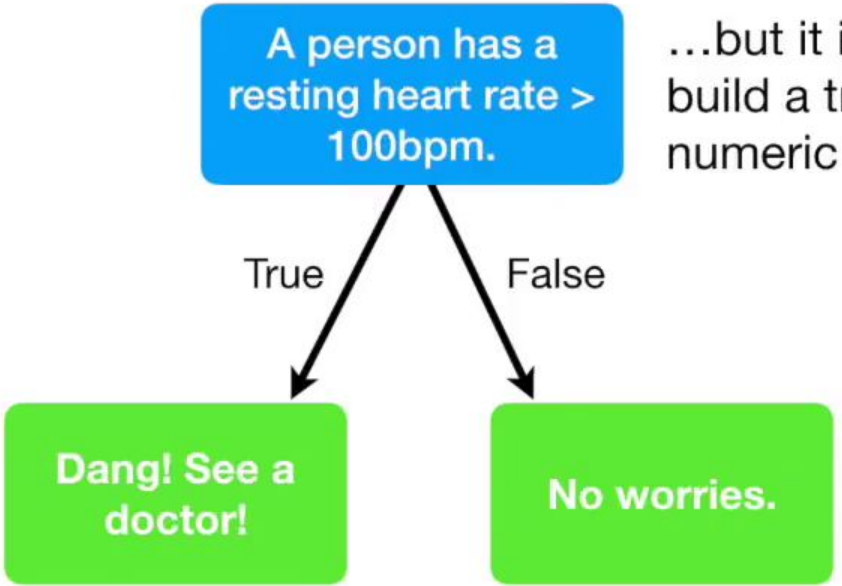



…and then classifies the person based on the answer.

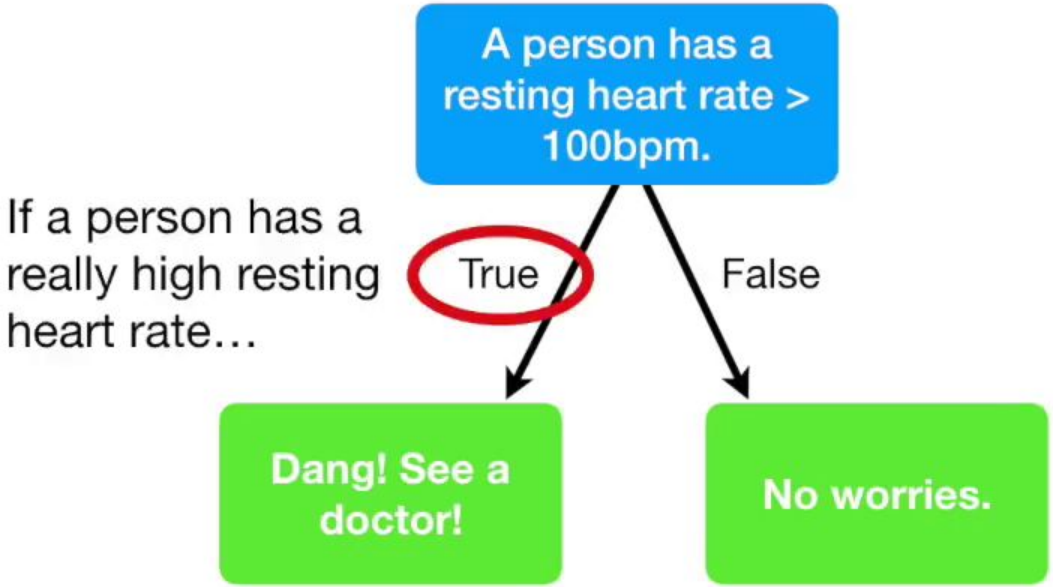
# No big deal!!!!


If a person loves Mumbai Indians
True
False
The person is awesome!!!
The person is slightly less than awesome. :(

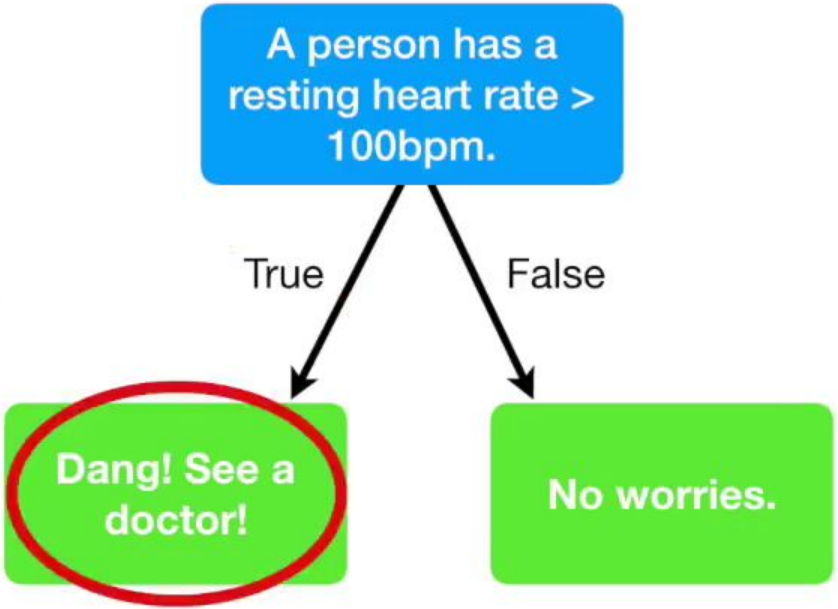

If a person loves Mumbai Indians
This decision tree is based on a “yes/no” question…
True
False
The person is awesome!!!
The person is slightly less than awesome. :(

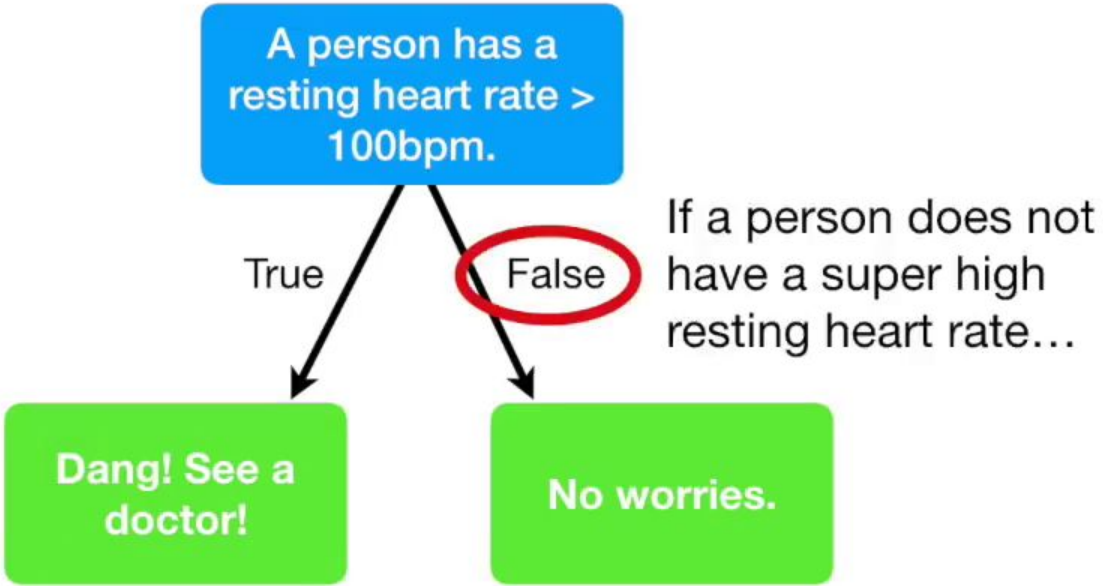


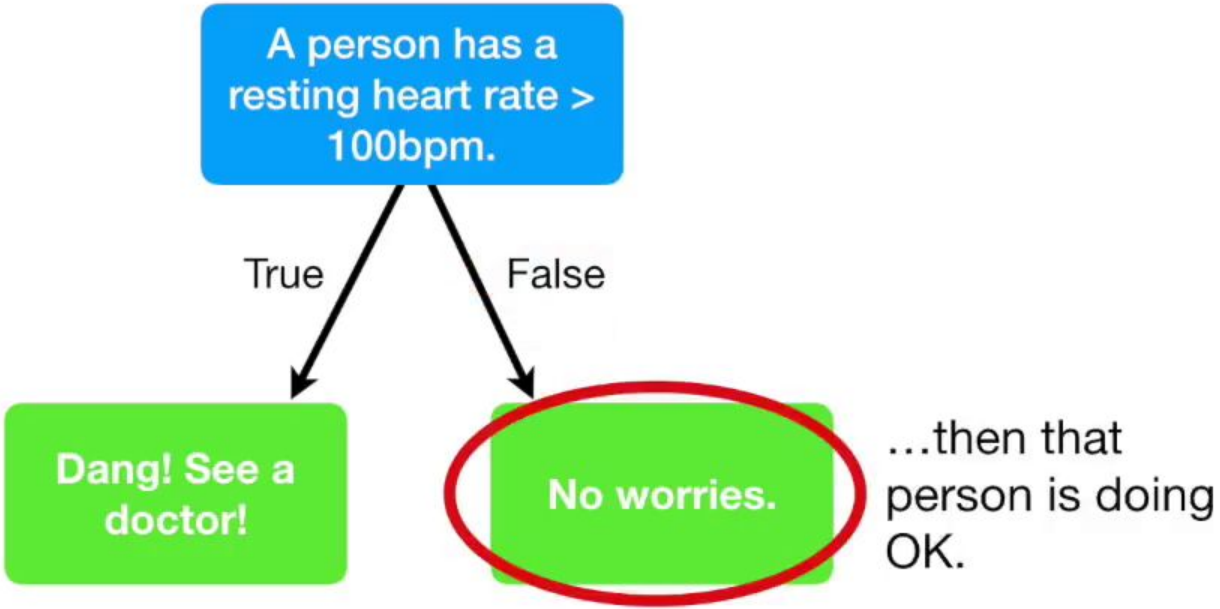

…but it is just as easy to build a tree from numeric data.


A person has a resting heart rate > 100bpm.
If a person has a really high resting heart rate…
True
False
Dang! See a doctor!
No worries.

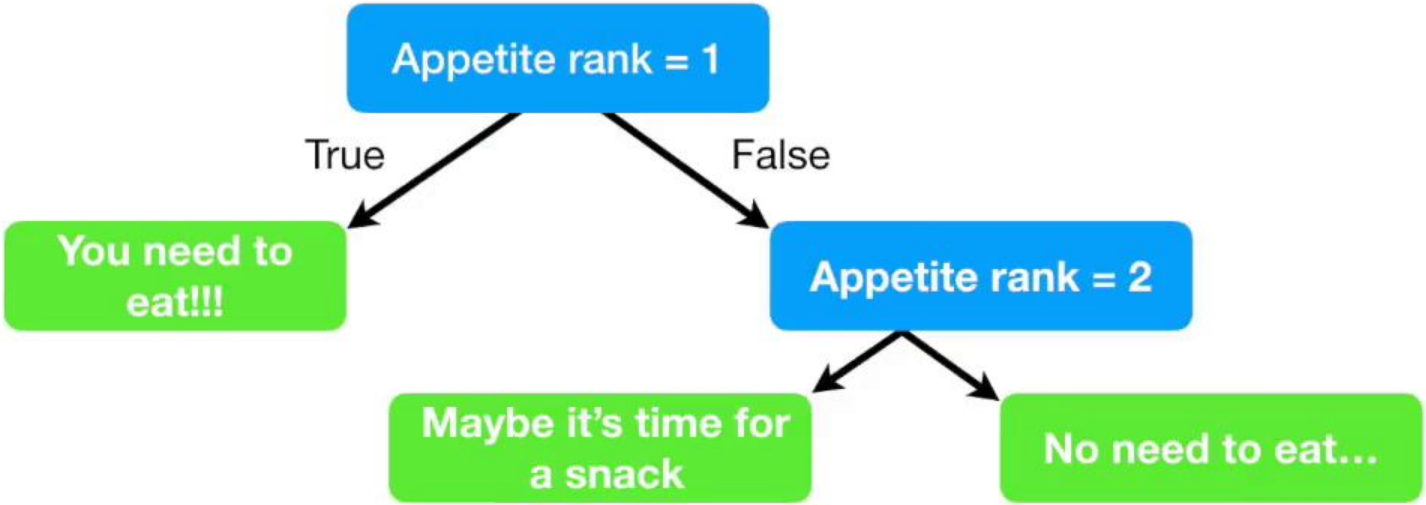
…then that person is had better see a doctor.



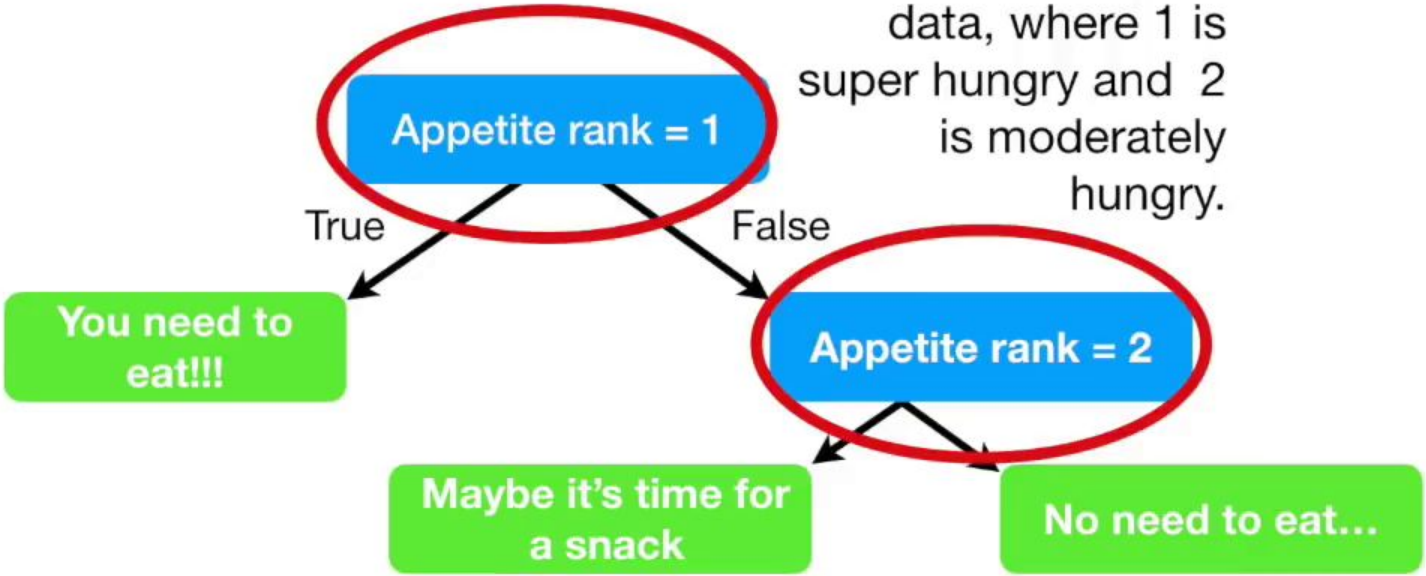

A person has a resting heart rate > 100bpm.
True
False
If a person does not have a super high resting heart rate…
Dang! See a doctor!
No worries.


A person has a resting heart rate > 100bpm.
True
False
Dang! See a doctor!
No worries.
…then that person is doing OK.

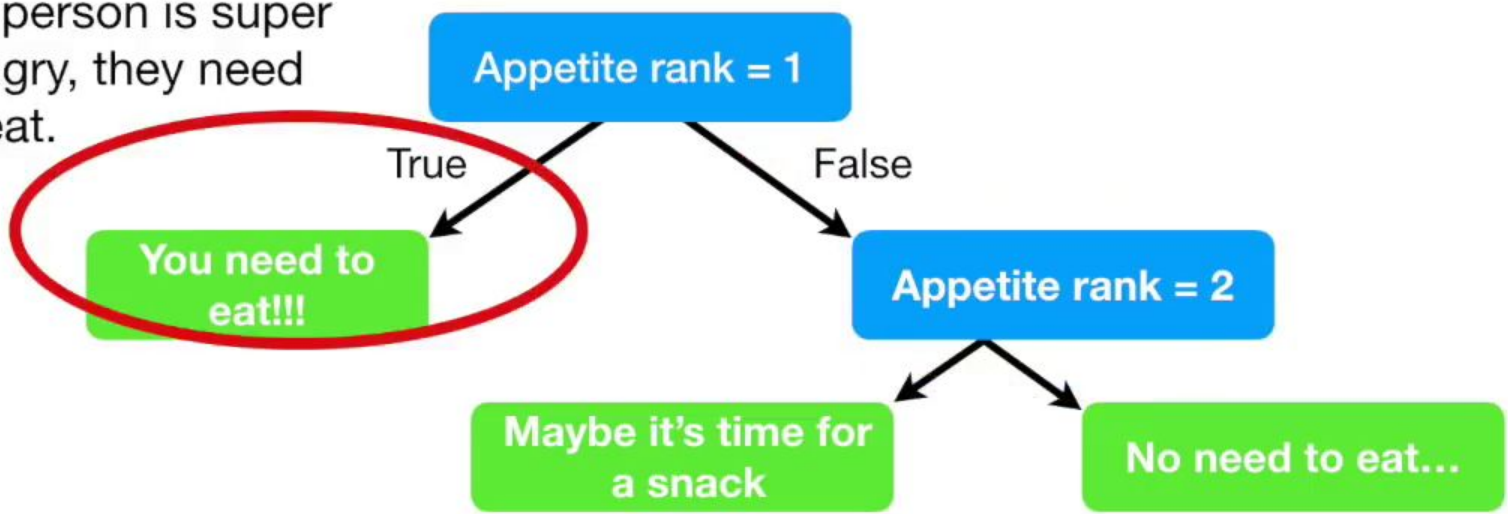
One more simple decision tree…

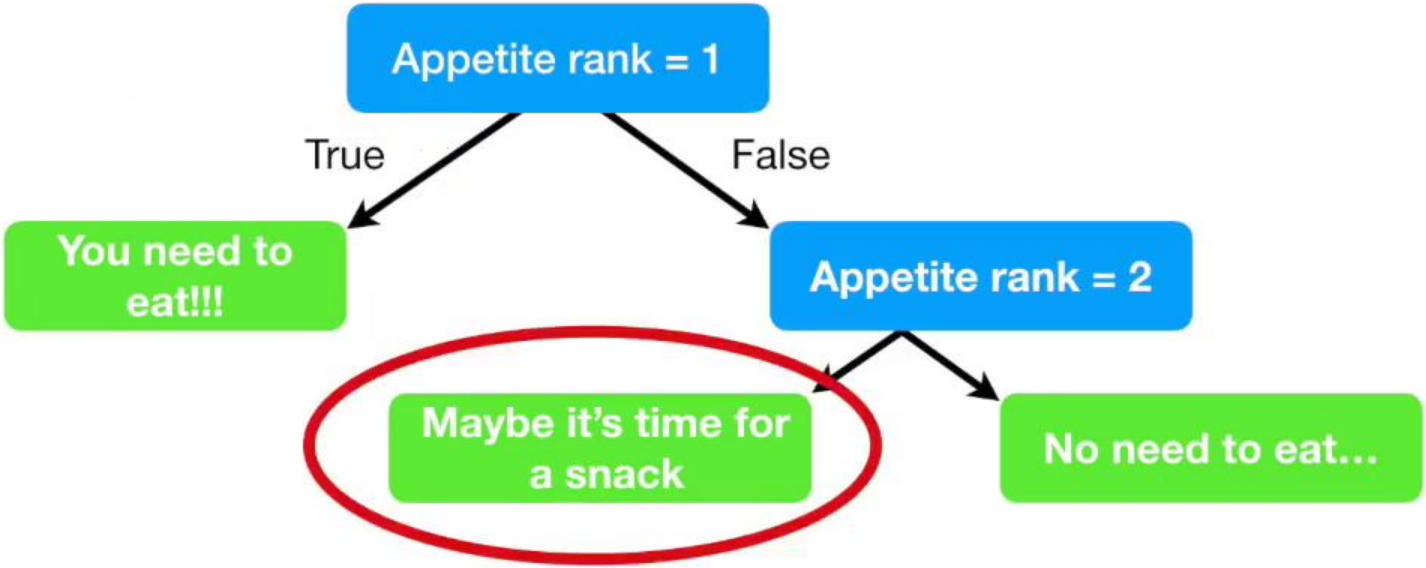


This decision tree is based on **ranked** data, where 1 is super hungry and 2 is moderately hungry.

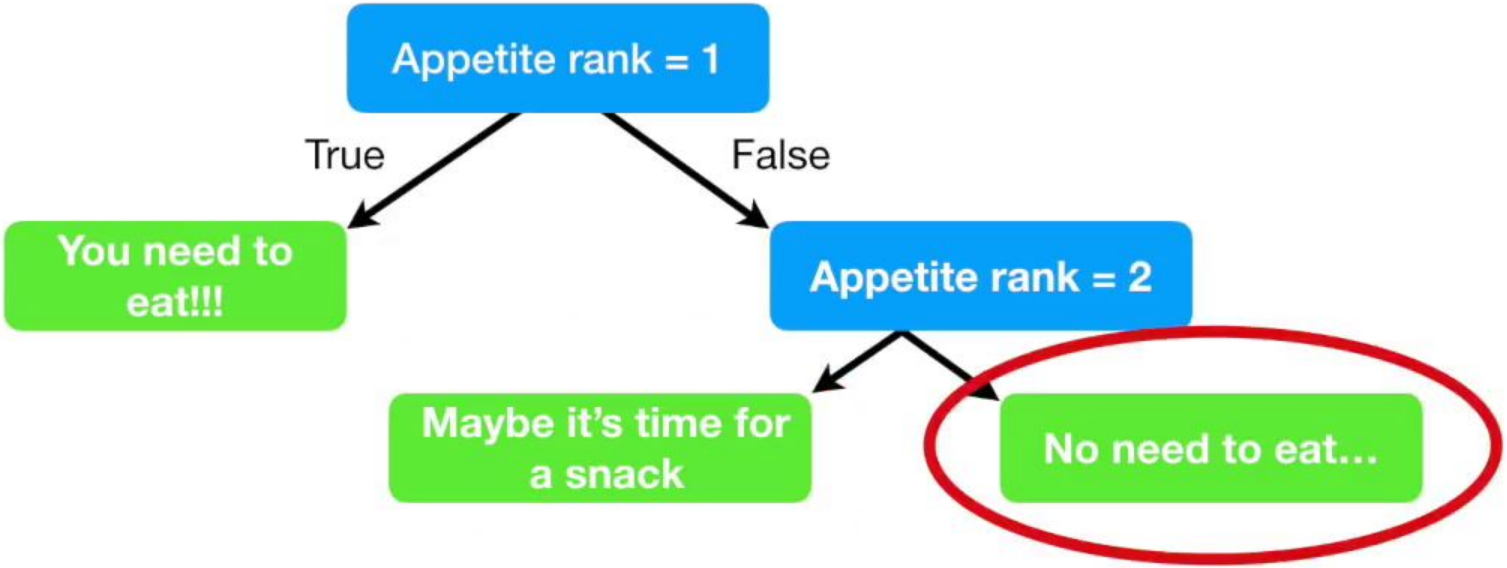


If a person is super hungry, they need to eat.

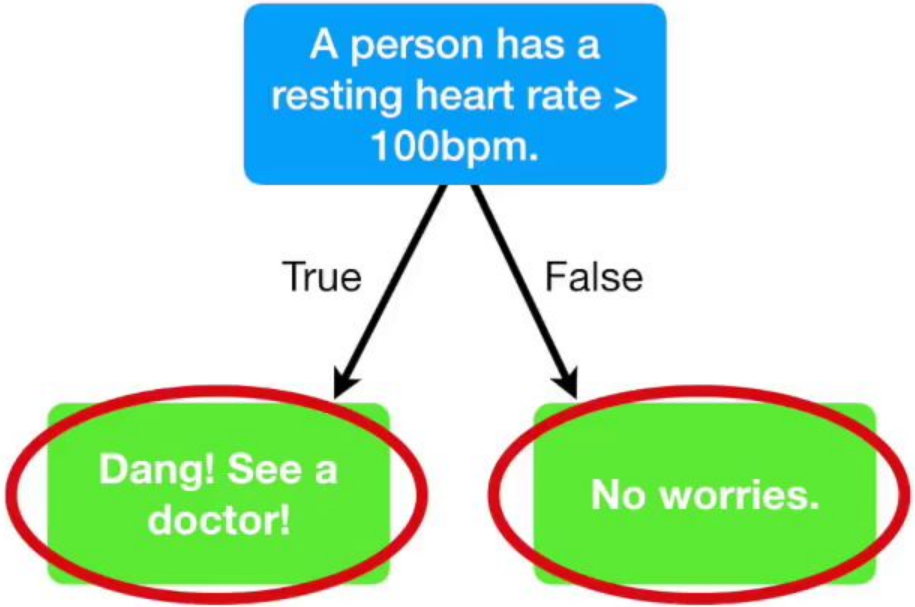




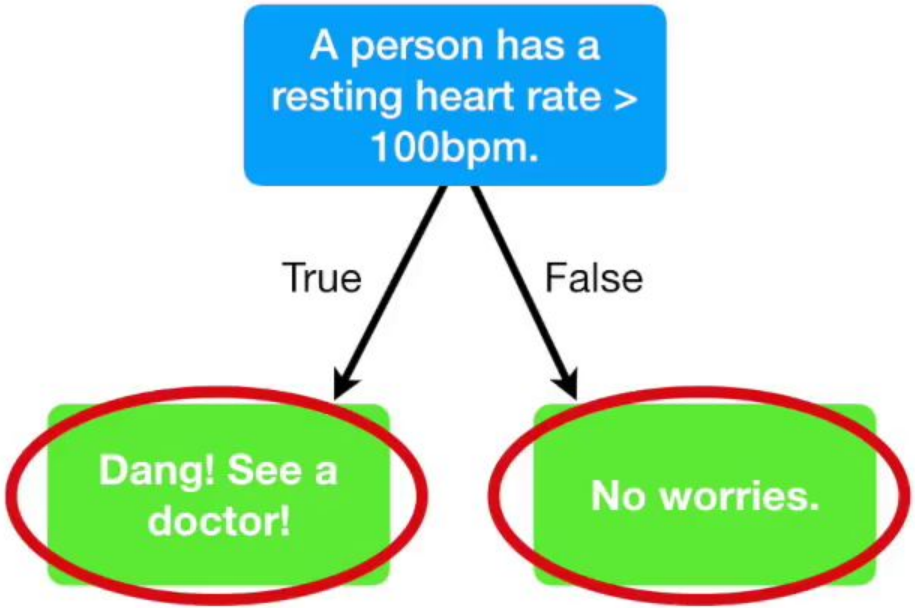

If a person is moderately hungry, they just need a snack.

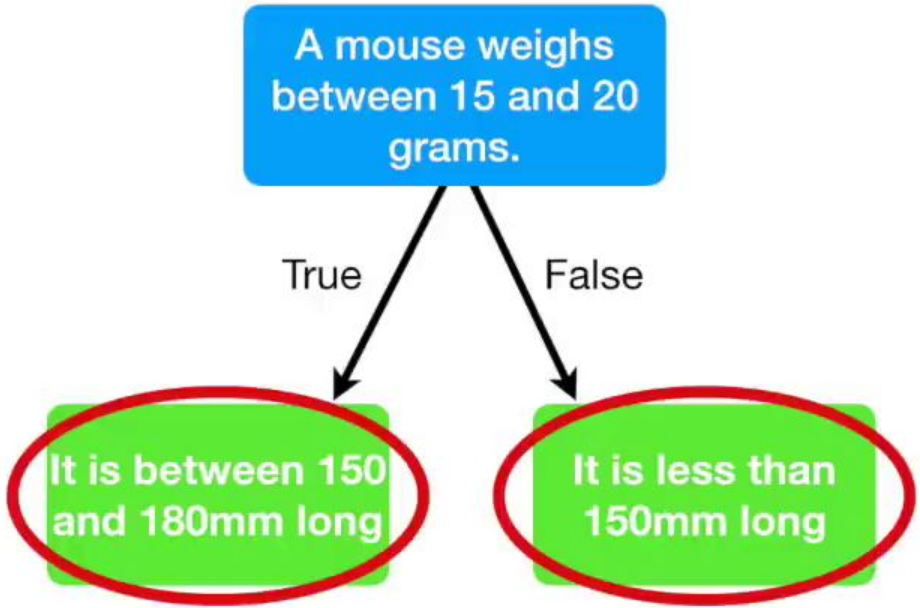



And if they are not hungry at all, then there's no need to eat.




**NOTE:** The classification can be categories…

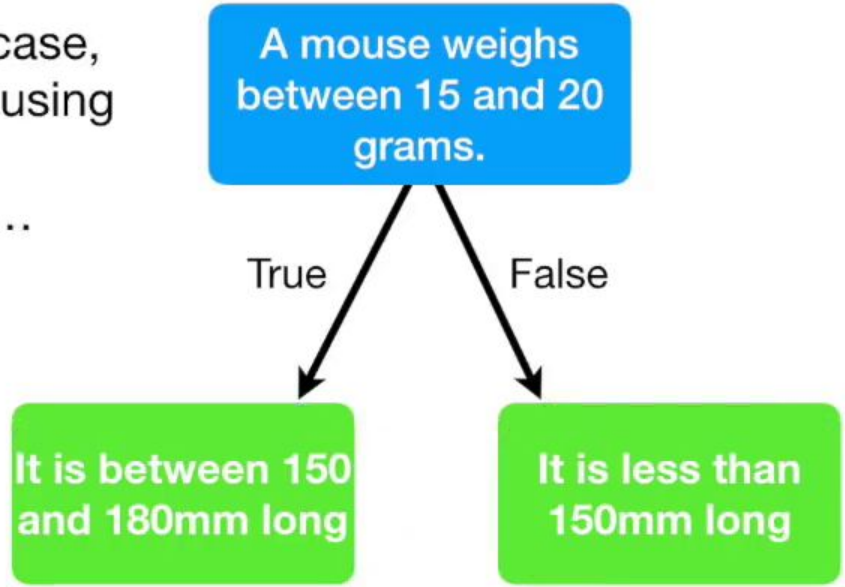


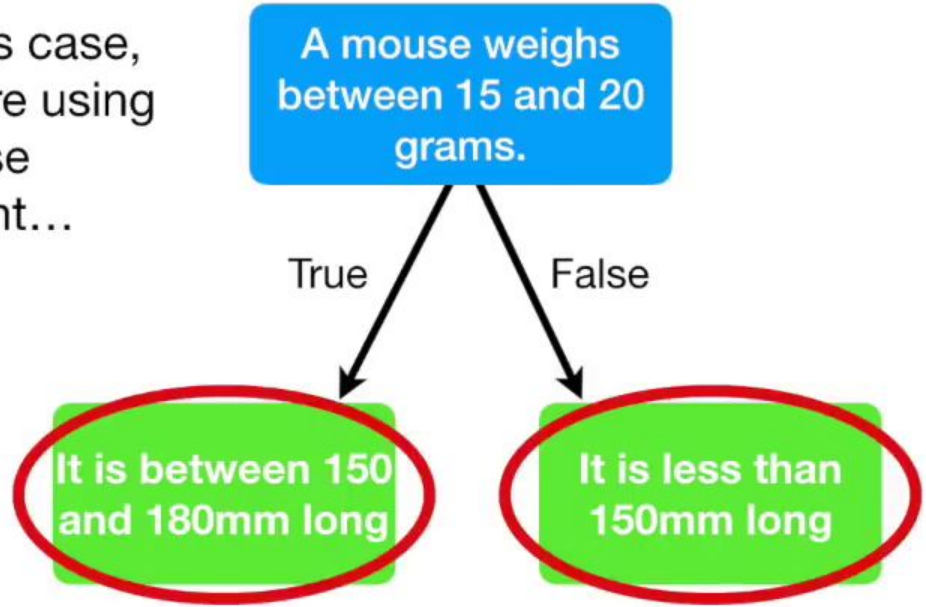

**NOTE:** The classification can be categories…

…or numeric.

In this case, we are using mouse weight…

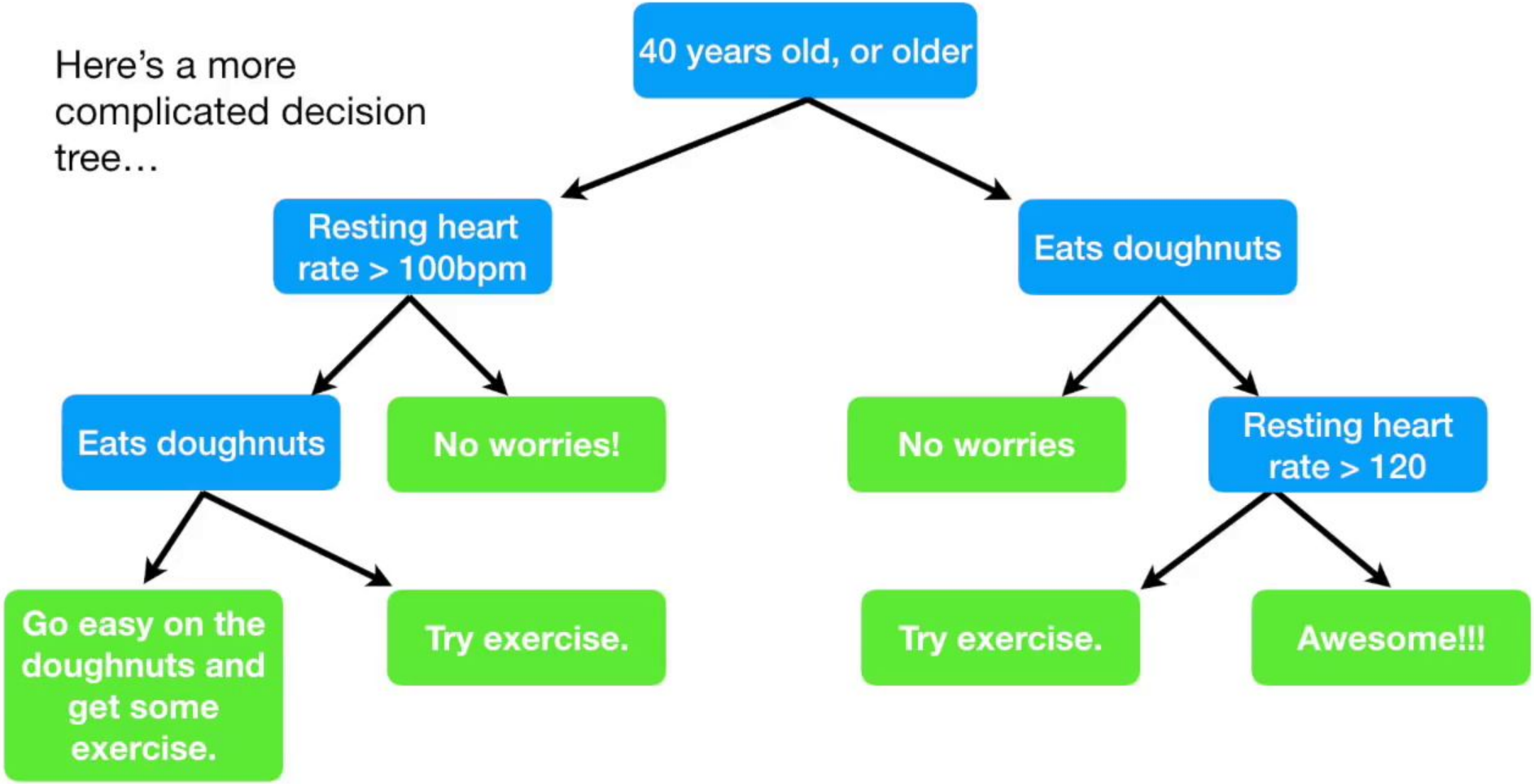


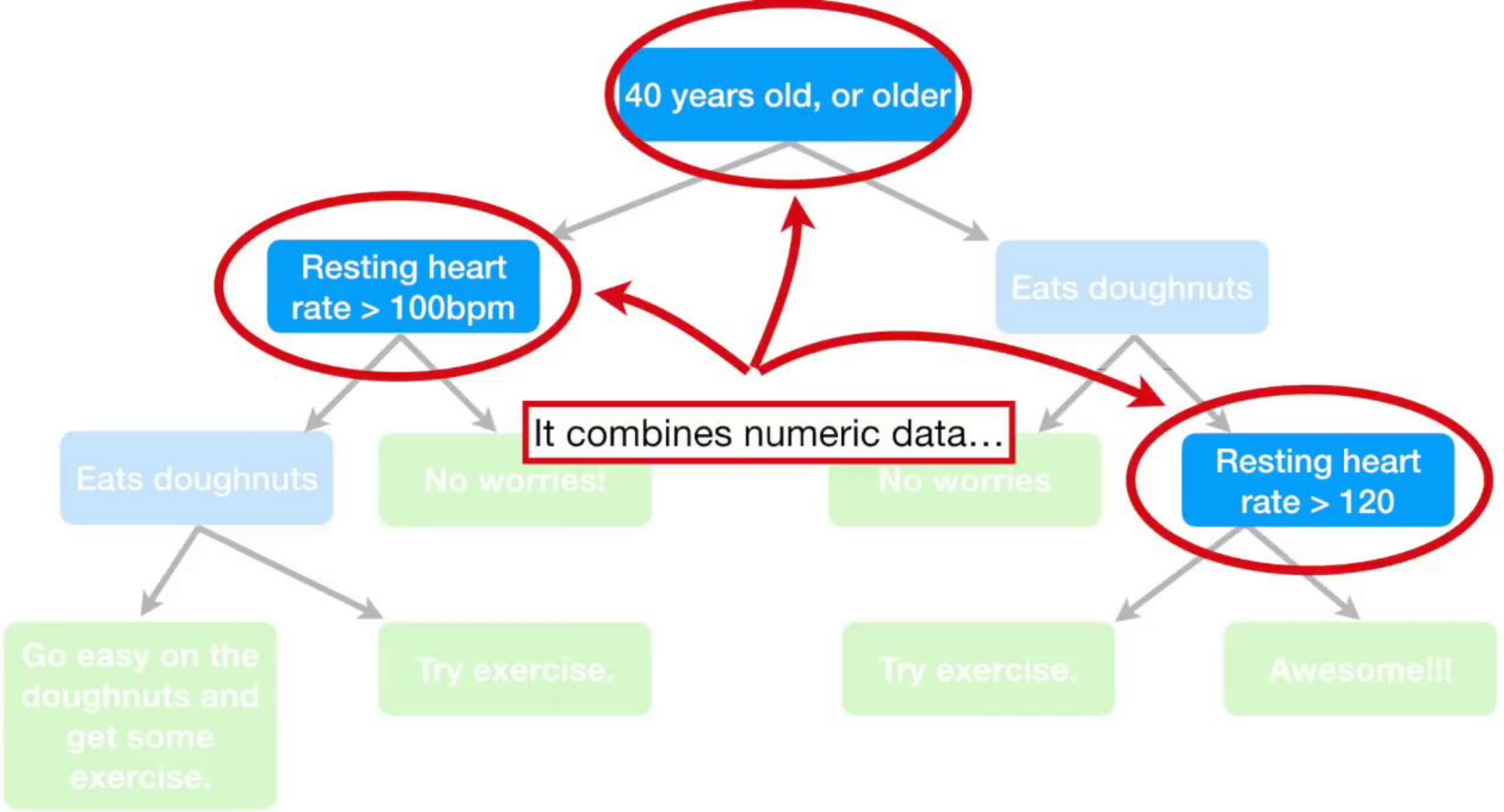
In this case, we are using mouse weight…

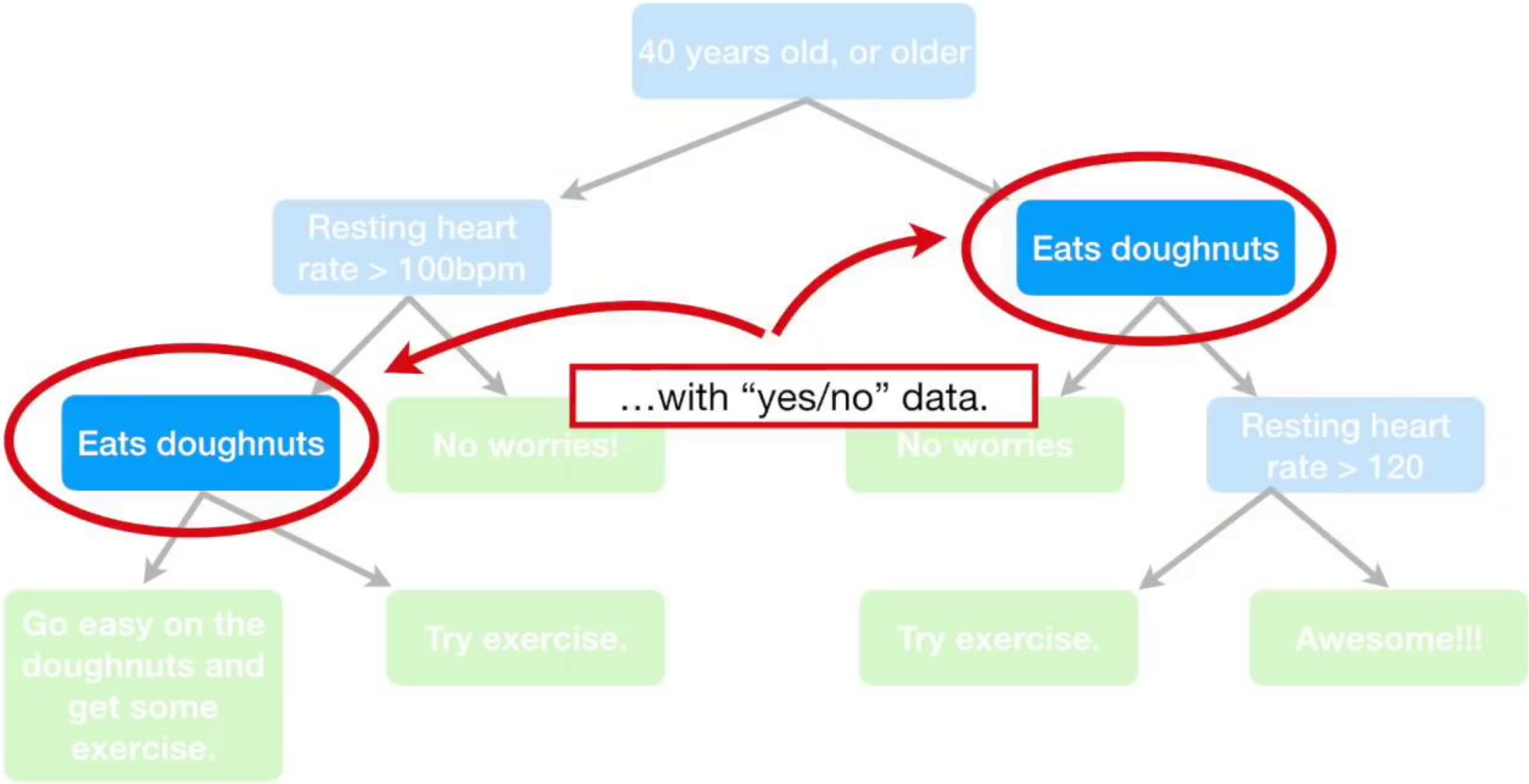


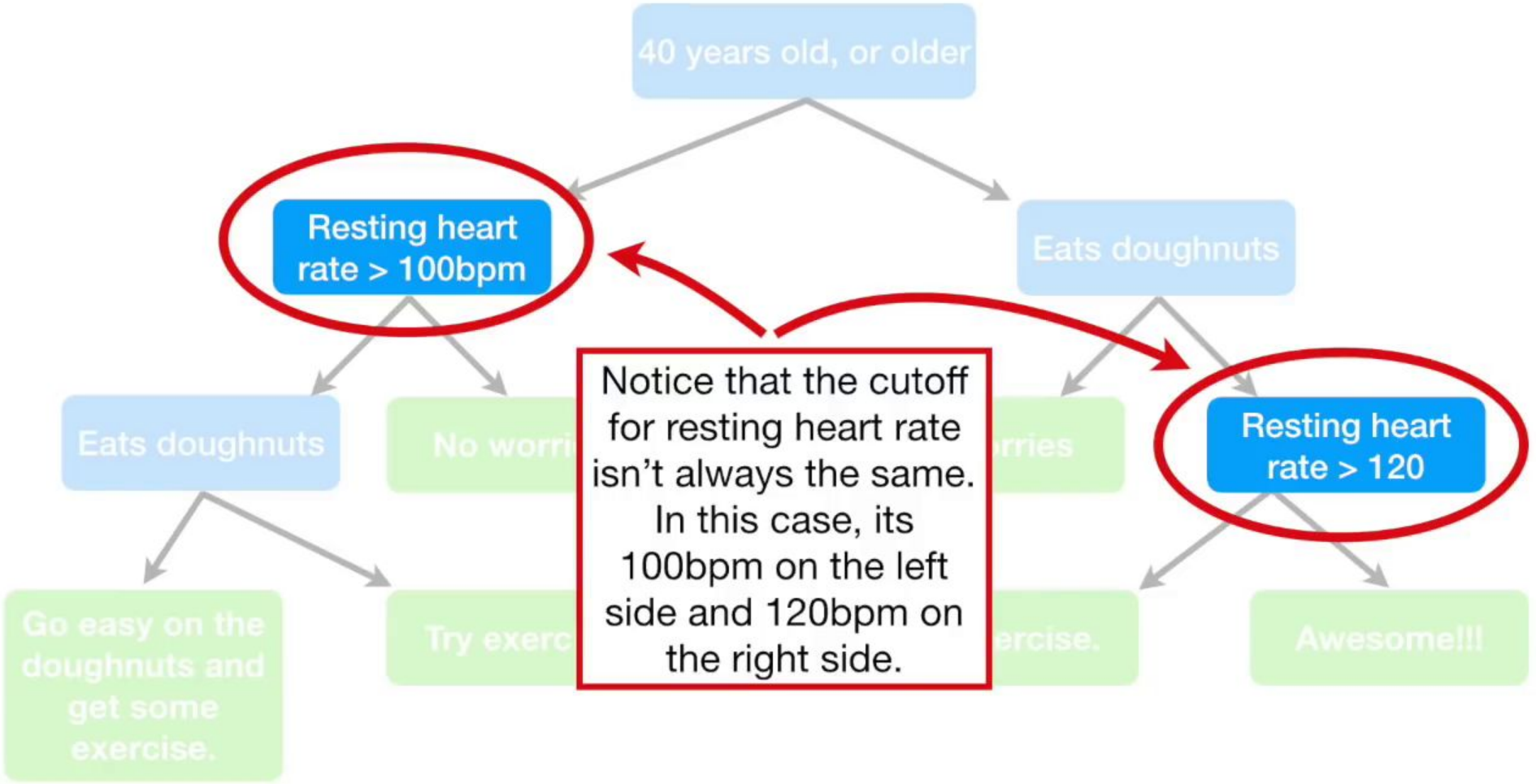

..to predict mouse size.

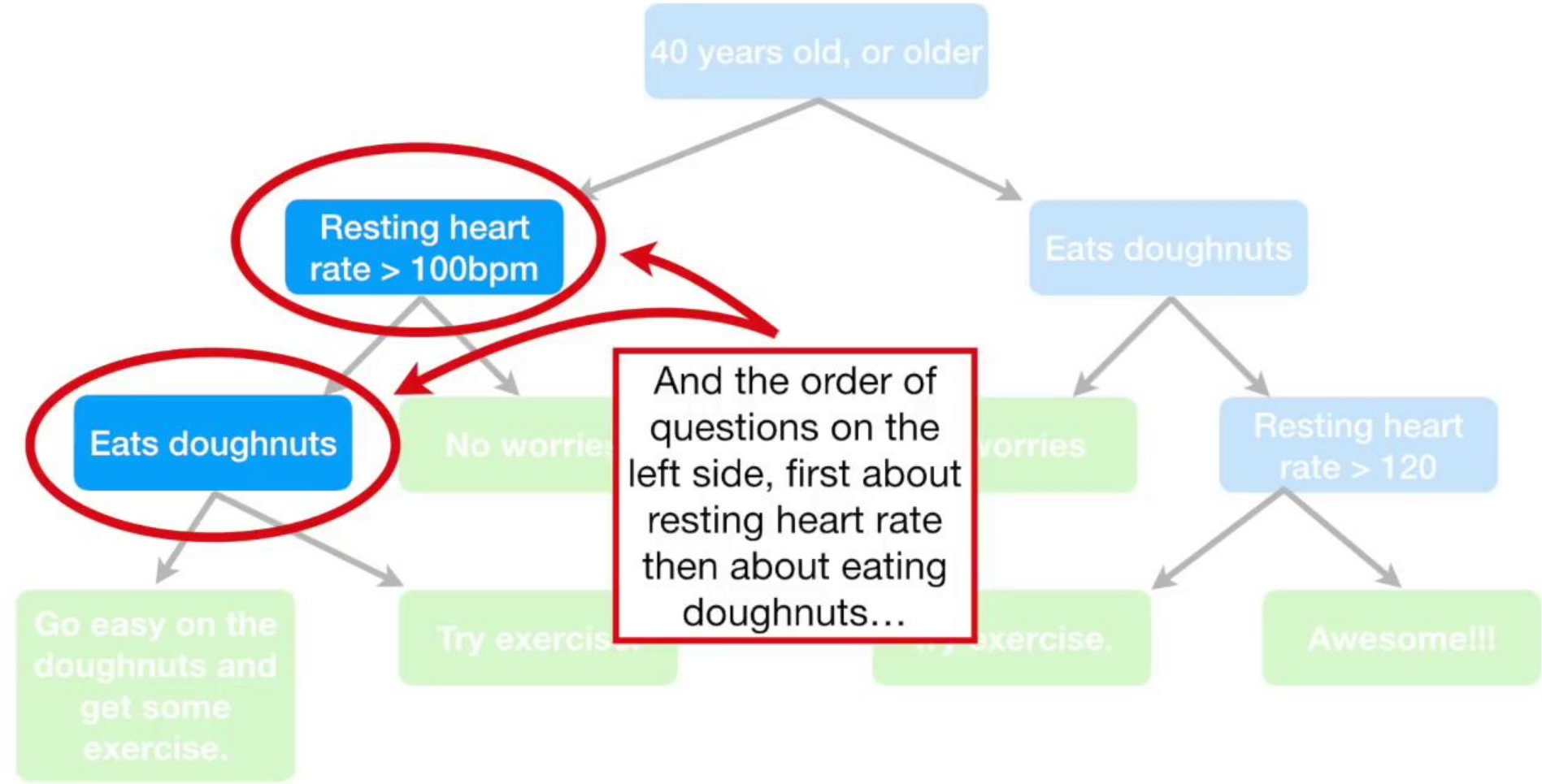
Here's a more complicated decision tree…

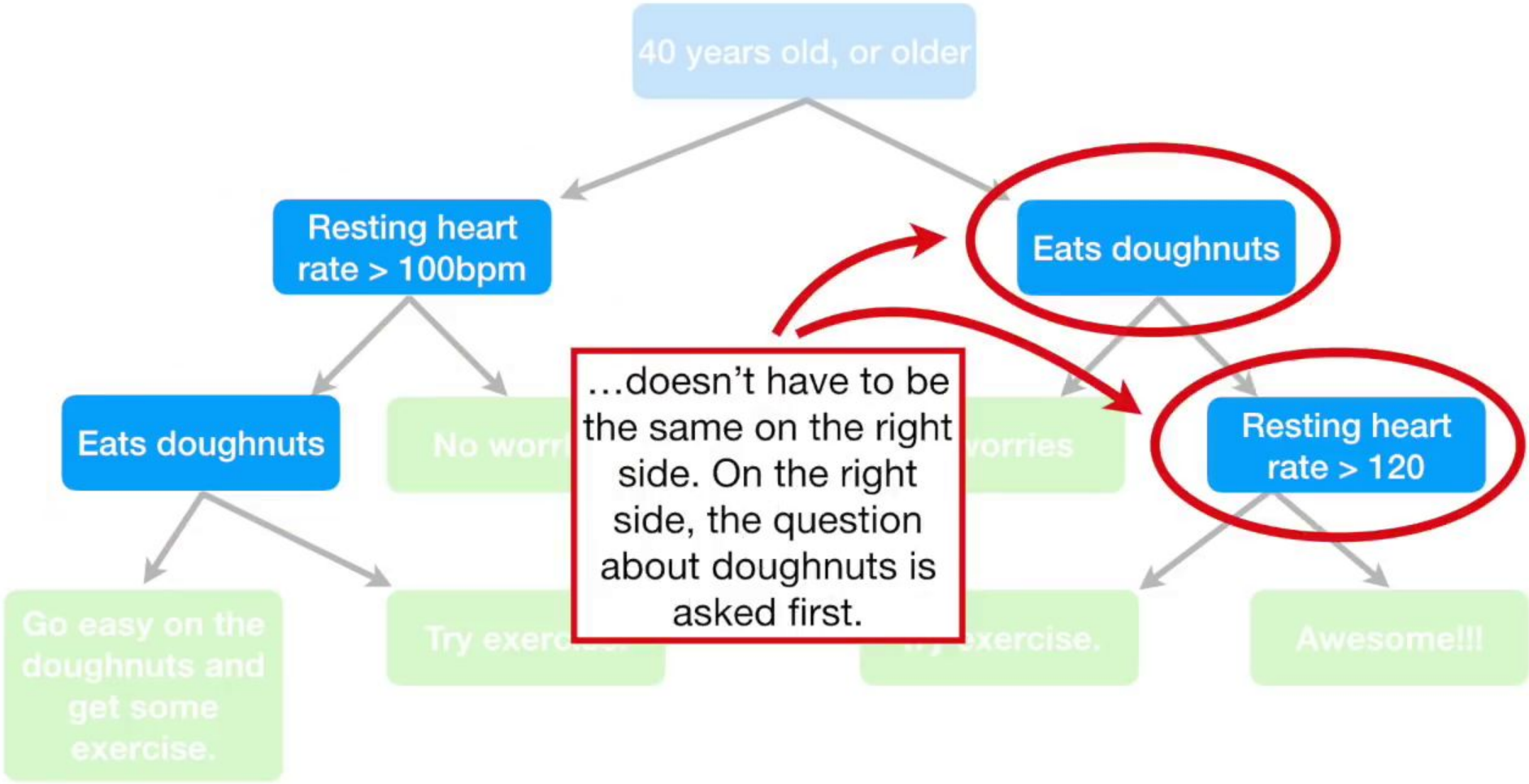


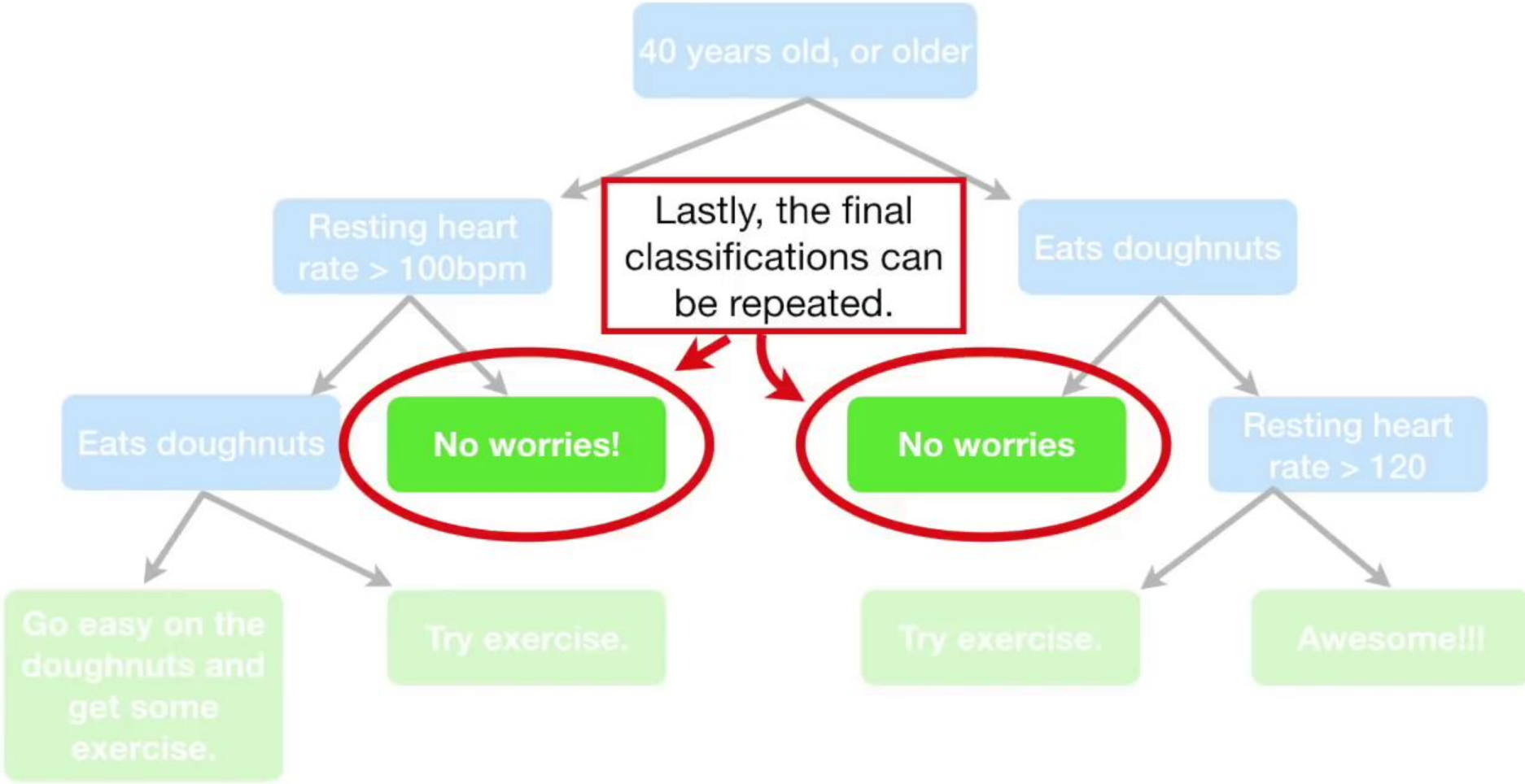

40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
It combines numeric data…
Eats doughnuts
No worries!
No worries
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Try exercise.
Try exercise.
Awesome!!!

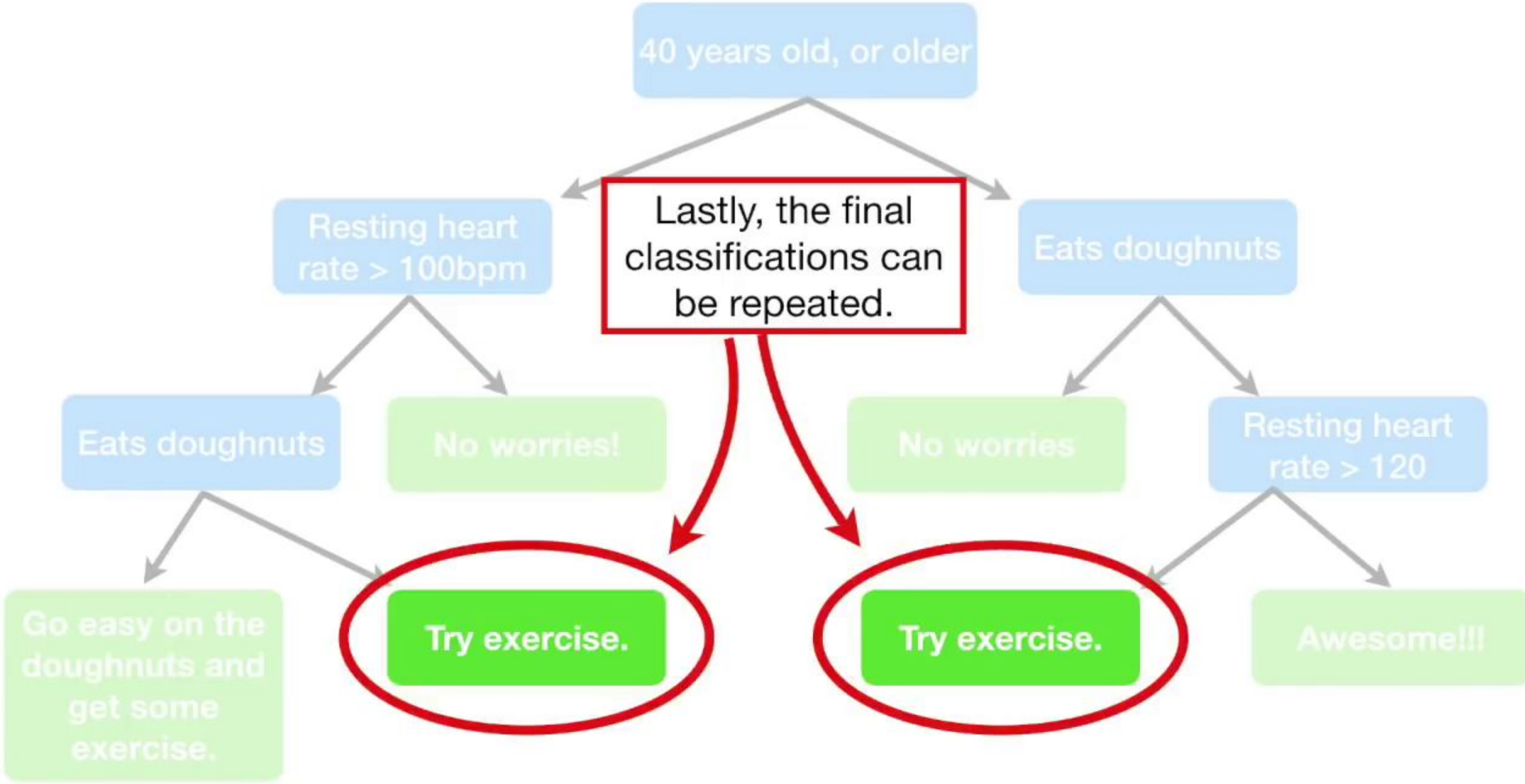

40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
Eats doughnuts
No worries!
No worries
Resting heart rate > 120
…with “yes/no” data.
Go easy on the doughnuts and get some exercise.
Try exercise.
Try exercise.
Awesome!!!

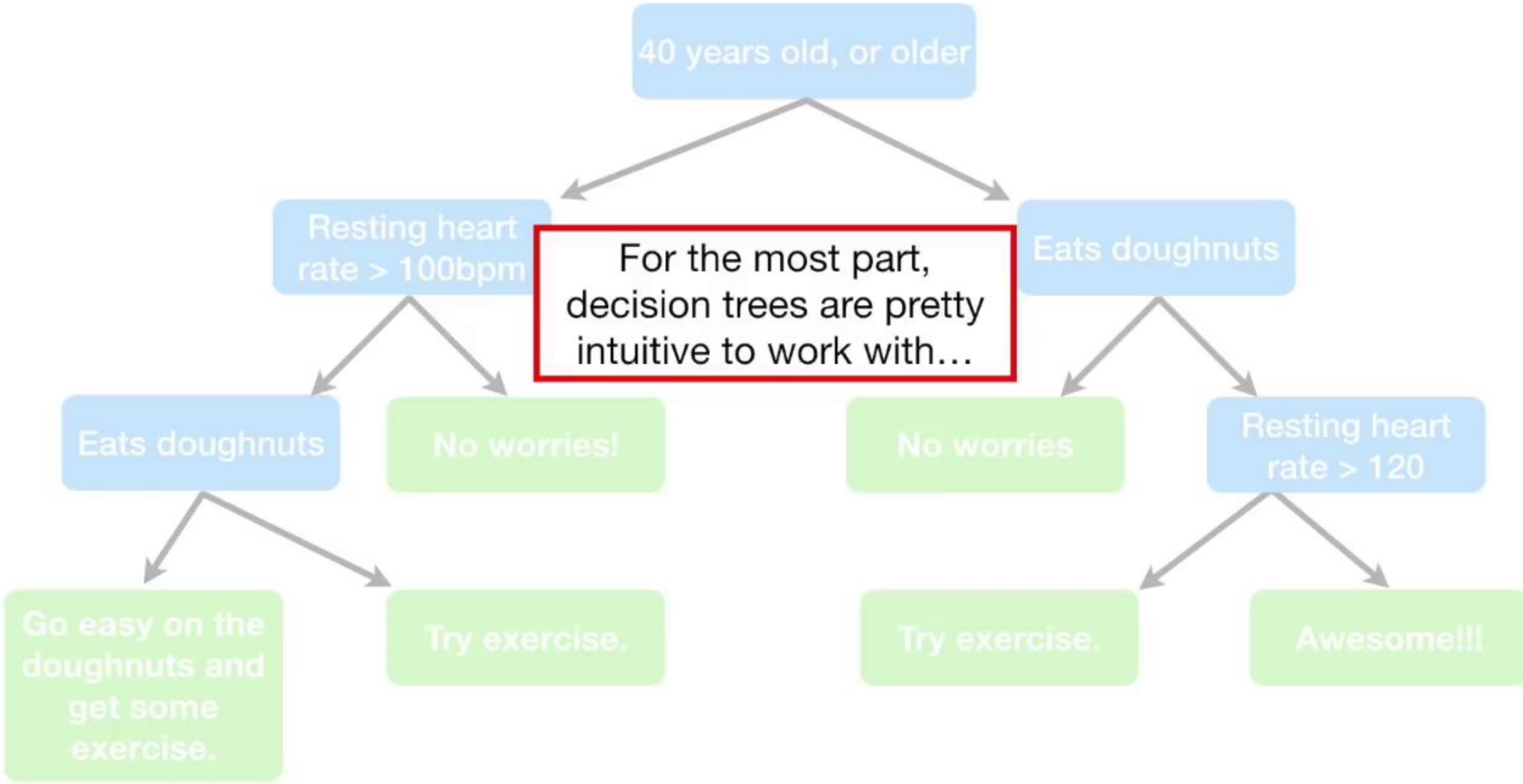

40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
Eats doughnuts
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Awesome!!!
Notice that the cutoff for resting heart rate isn't always the same. In this case, its 100bpm on the left side and 120bpm on the right side.

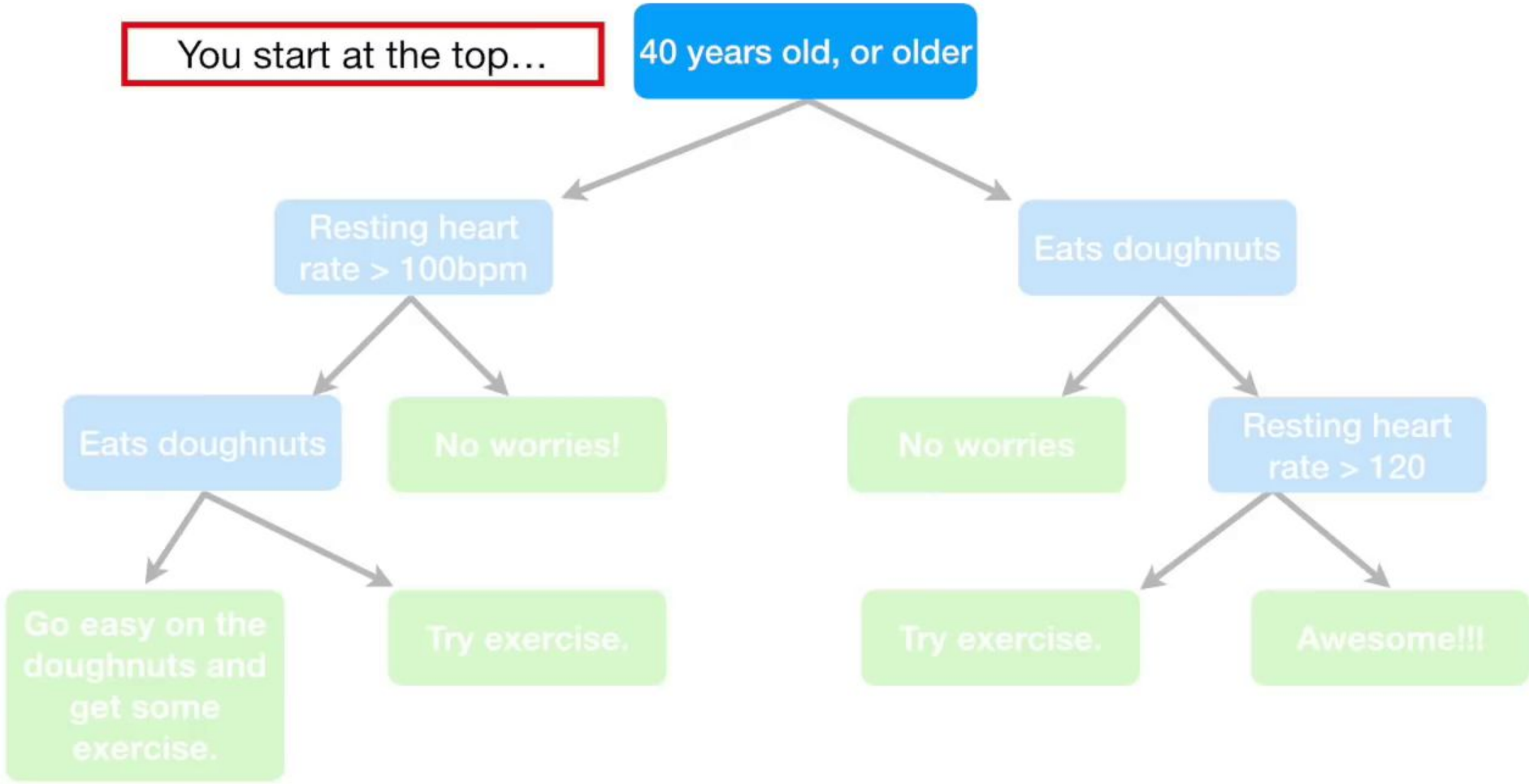

40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
Eats doughnuts
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Awesome!!!
And the order of questions on the left side, first about resting heart rate then about eating doughnuts…

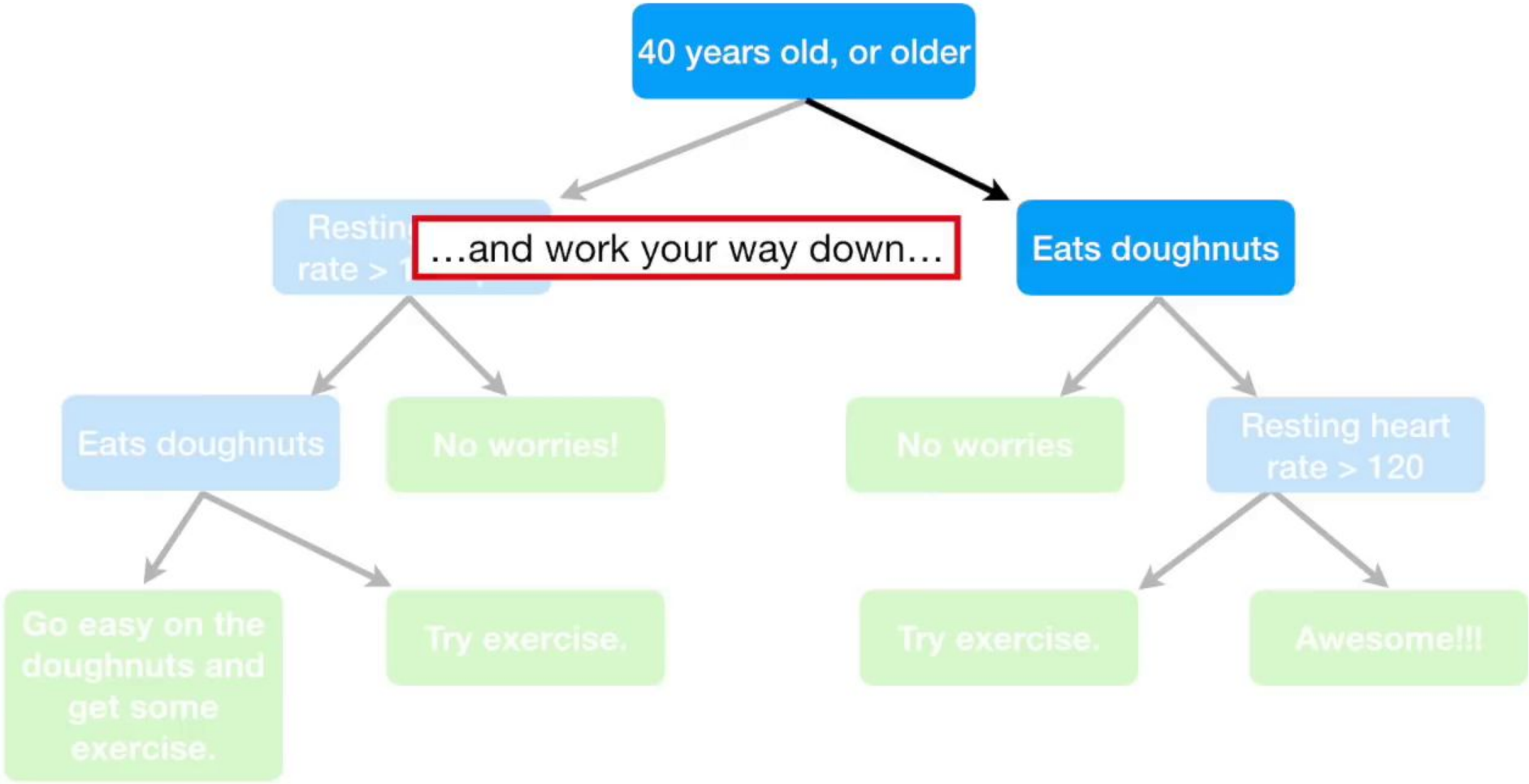

40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
Eats doughnuts
Resting heart rate > 120
…doesn't have to be the same on the right side. On the right side, the question about doughnuts is asked first.
Go easy on the doughnuts and get some exercise.
Awesome!!!

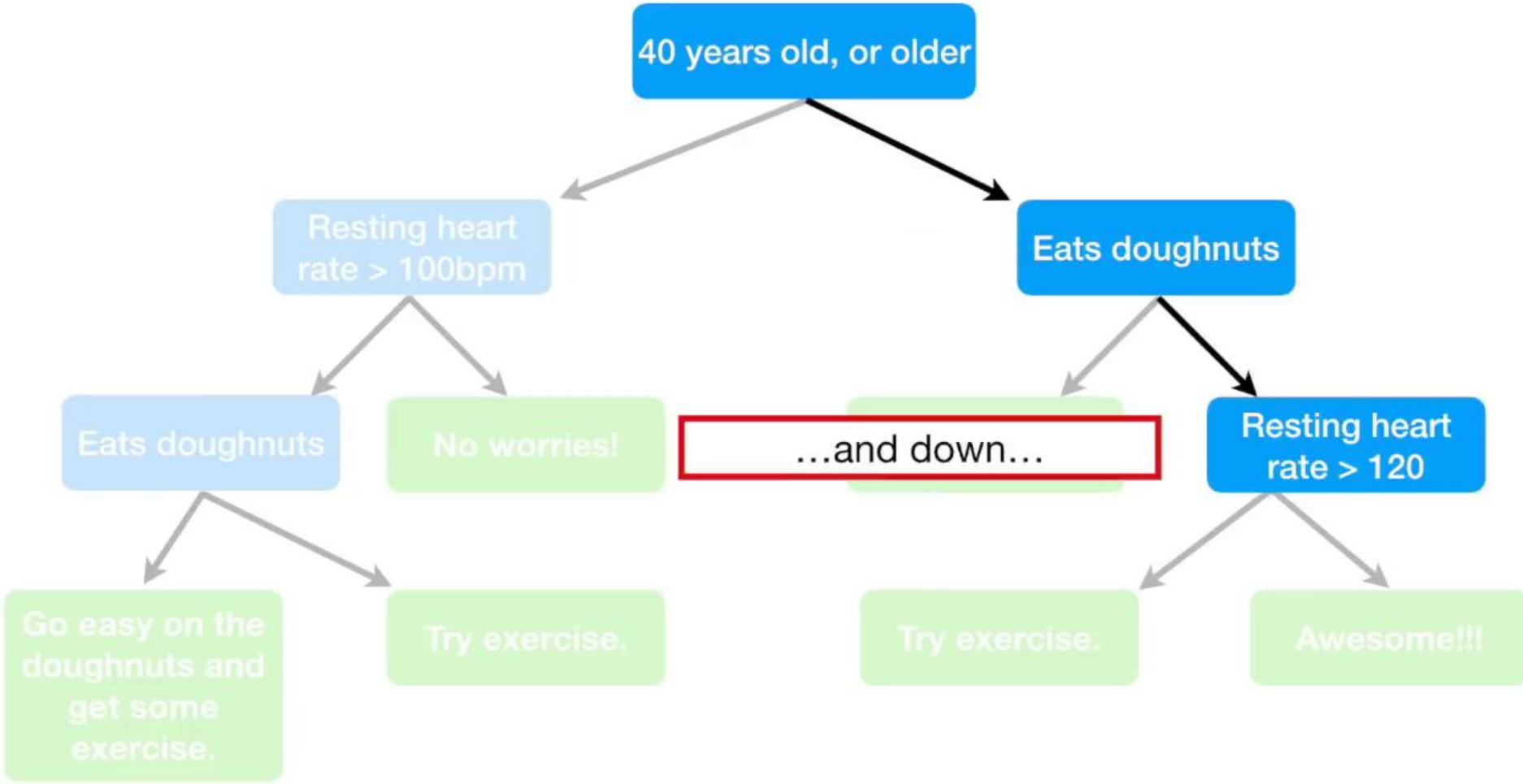

40 years old, or older
Resting heart rate > 100bpm
Lastly, the final classifications can be repeated.
Eats doughnuts
Eats doughnuts
No worries!
No worries
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Try exercise.
Try exercise.
Awesome!!!

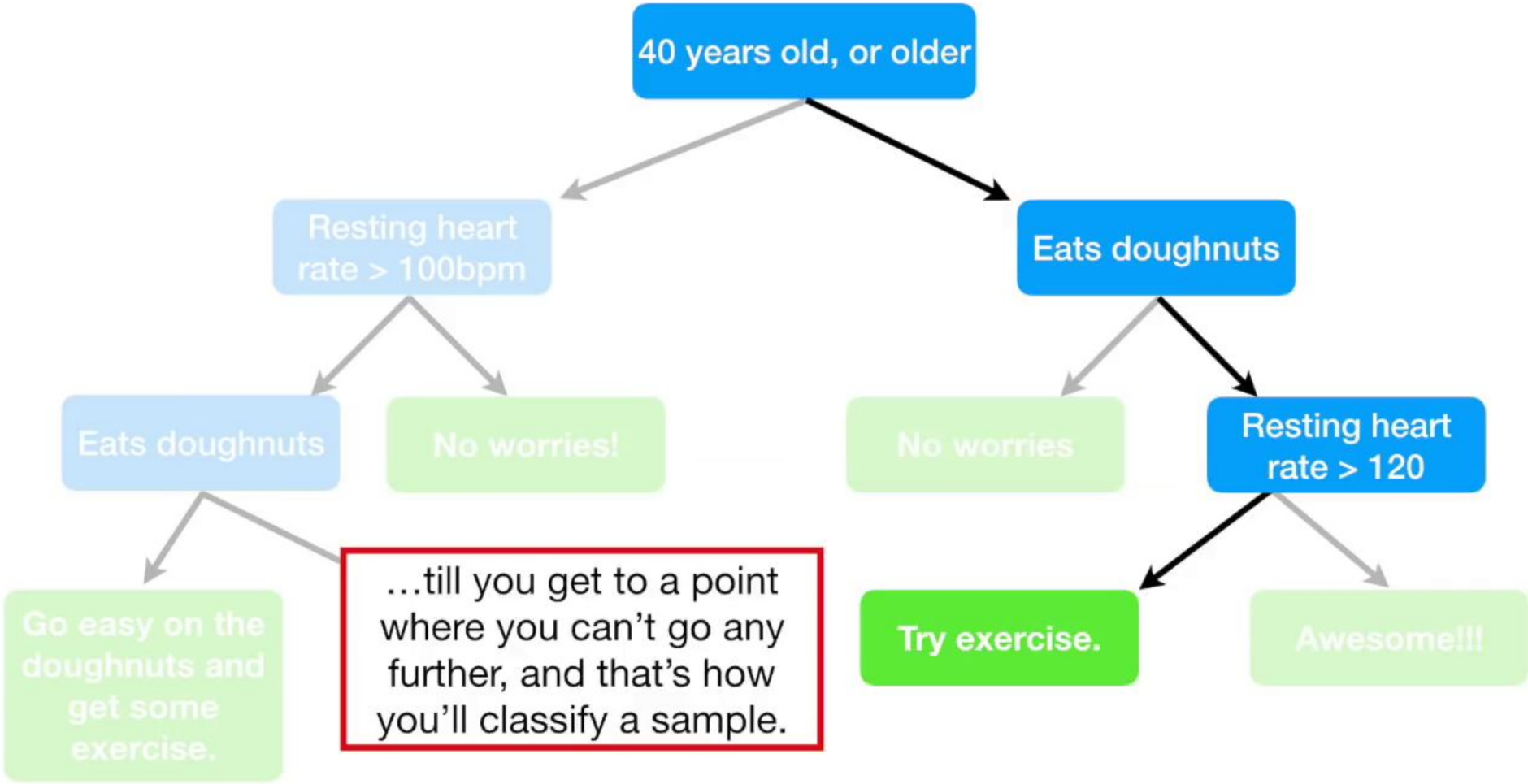

40 years old, or older
Resting heart rate > 100bpm
Lastly, the final classifications can be repeated.
Eats doughnuts
Eats doughnuts
No worries!
No worries
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Try exercise.
Try exercise.
Awesome!!!


40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
For the most part, decision trees are pretty intuitive to work with…
Eats doughnuts
No worries!
No worries
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Try exercise.
Try exercise.
Awesome!!!


You start at the top…
40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
Eats doughnuts
No worries!
No worries
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Try exercise.
Try exercise.
Awesome!!!

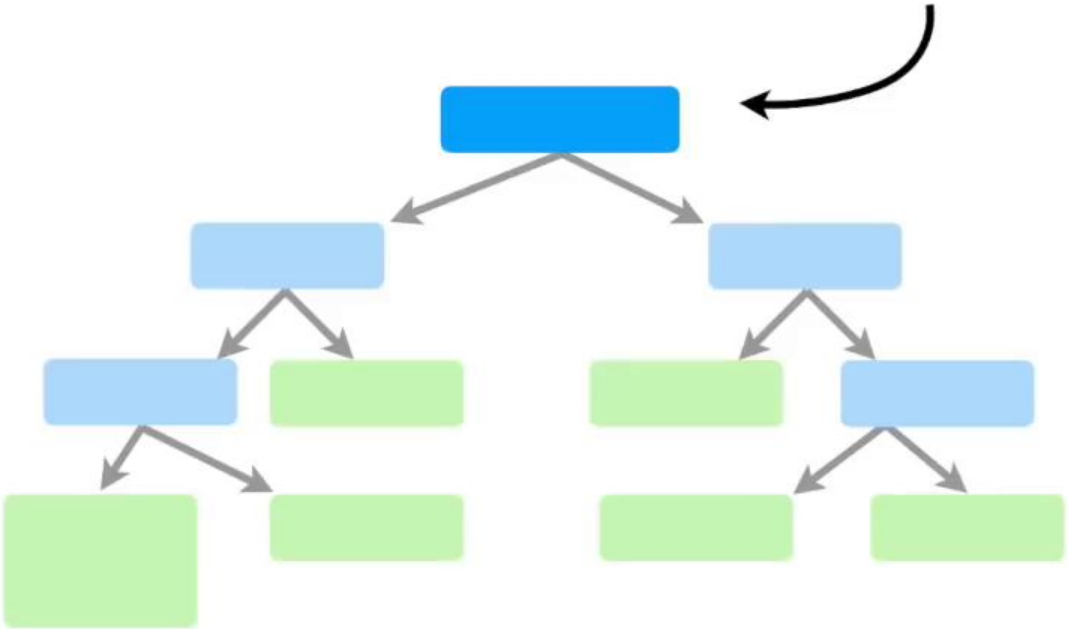

40 years old, or older
…and work your way down…
Eats doughnuts
Eats doughnuts
No worries!
No worries
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Try exercise.
Try exercise.
Awesome!!!


40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
Eats doughnuts
No worries!
…and down…
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
Try exercise.
Try exercise.
Awesome!!!

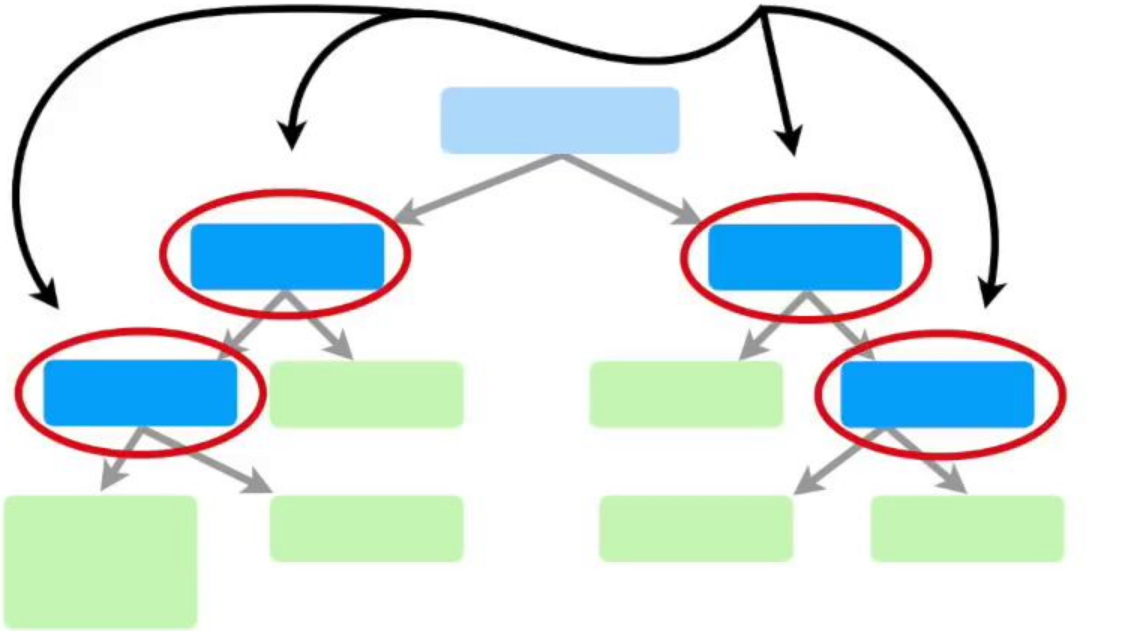

40 years old, or older
Resting heart rate > 100bpm
Eats doughnuts
Eats doughnuts
No worries!
No worries
Resting heart rate > 120
Go easy on the doughnuts and get some exercise.
…till you get to a point where you can't go any further, and that's how you'll classify a sample.
Try exercise.
Awesome!!!

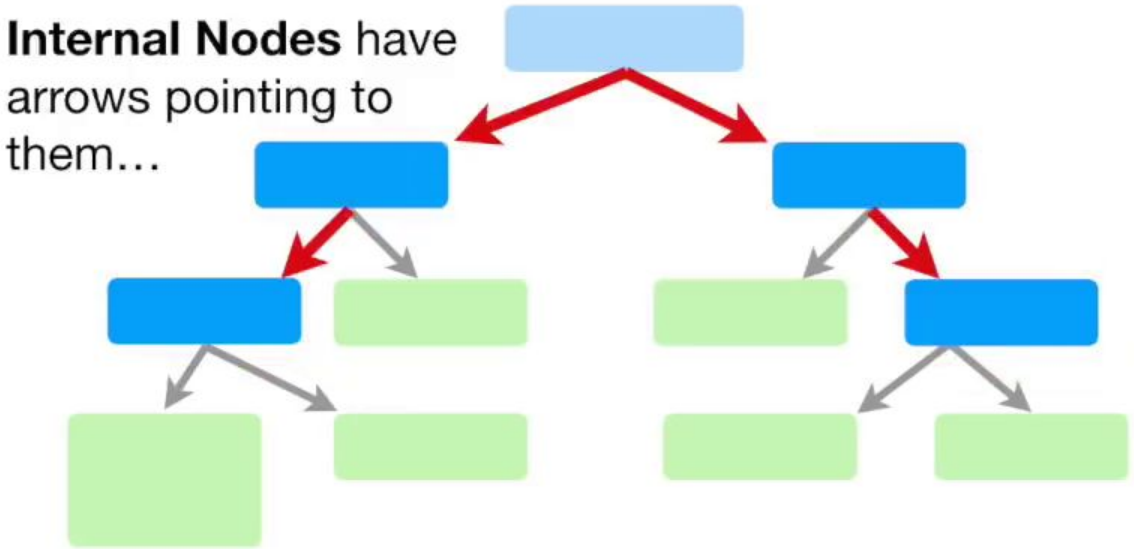
# Jargon Alert!!!

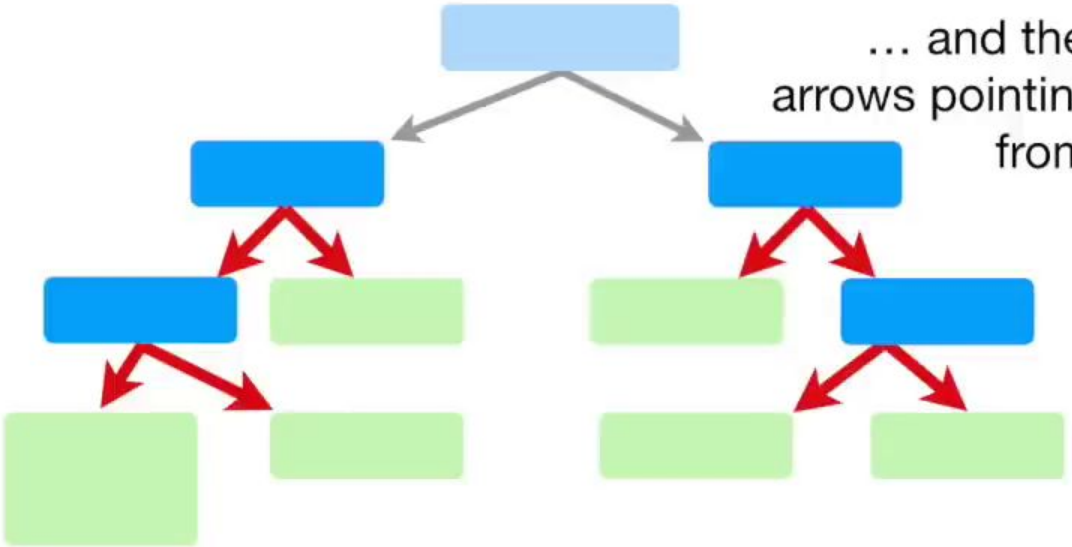
The very top of the tree is called the “**Root Node**” or just “**The Root**”



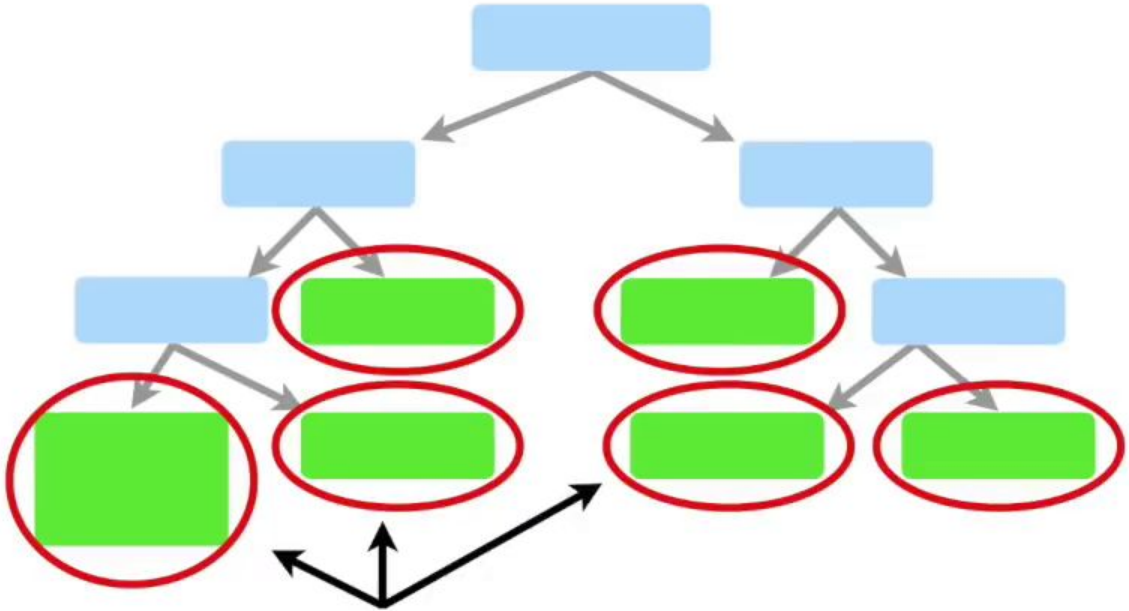
These are called “**Internal Nodes**”, or just “**Nodes**”.



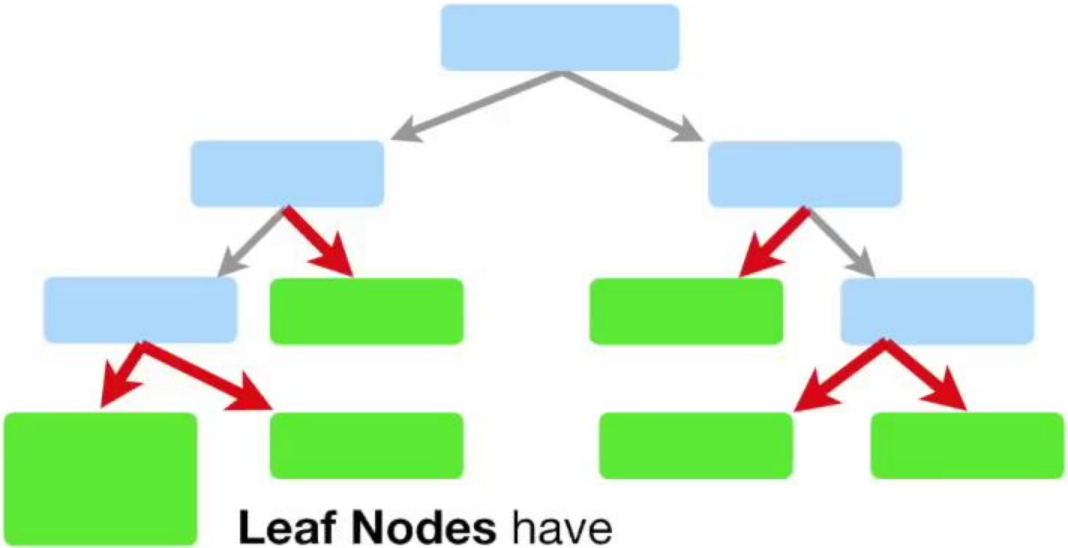

**Internal Nodes** have arrows pointing to them…



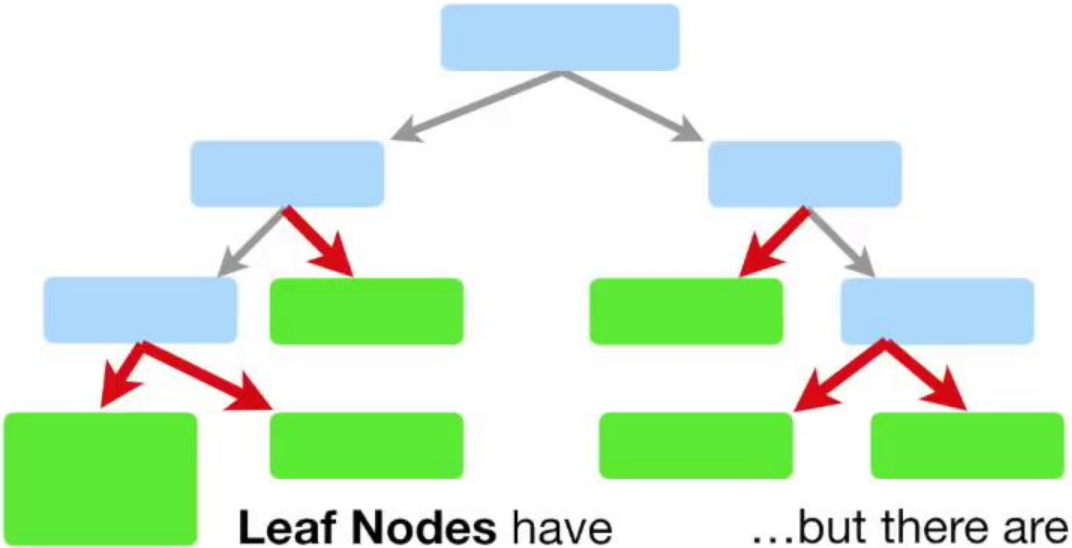
… and they have
arrows pointing away
from them.



Lastly, these are called “**Leaf Nodes**”, or just “**Leaves**”



**Leaf Nodes** have arrows pointing to them…



**Leaf Nodes** have arrows pointing to them…

…but there are no arrows pointing away from them.

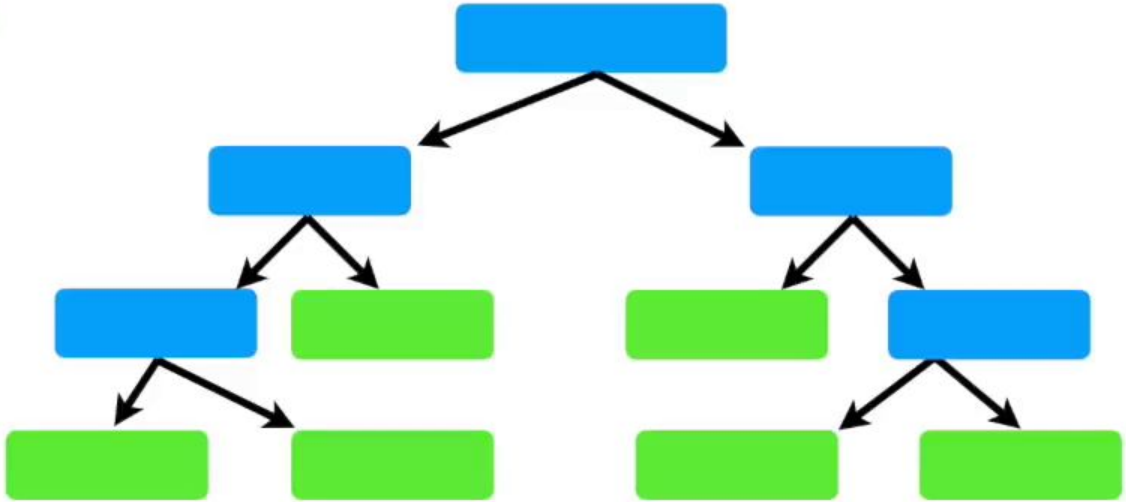
Now we are ready to talk about how to go from a raw table of data…

| 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 are ready to talk about how to go from a raw table of data…

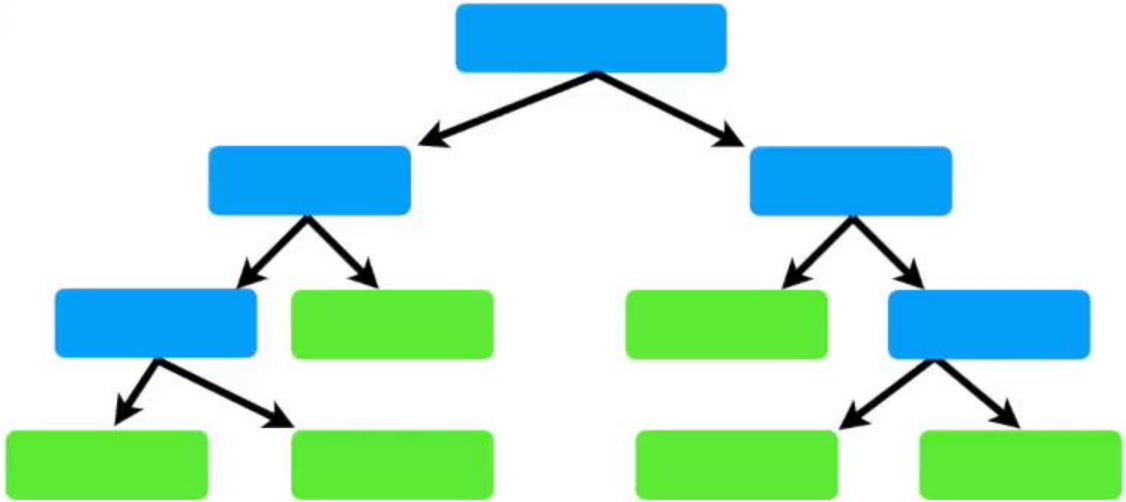
| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |

…to a decision tree!!!



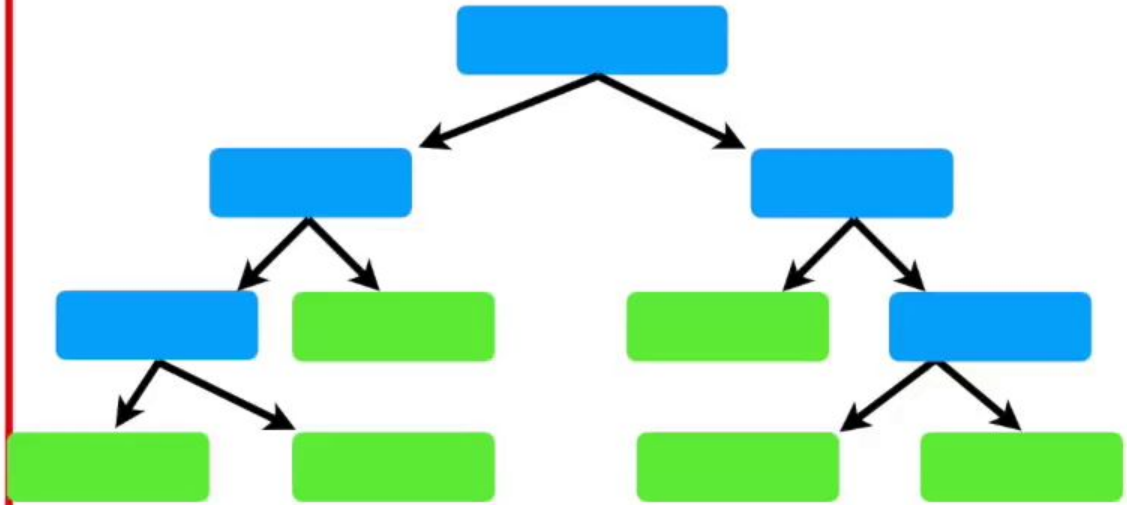
In this example, we want to create a tree that uses **chest pain**, **good blood circulation** and **blocked artery status** to predict…

| 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… |



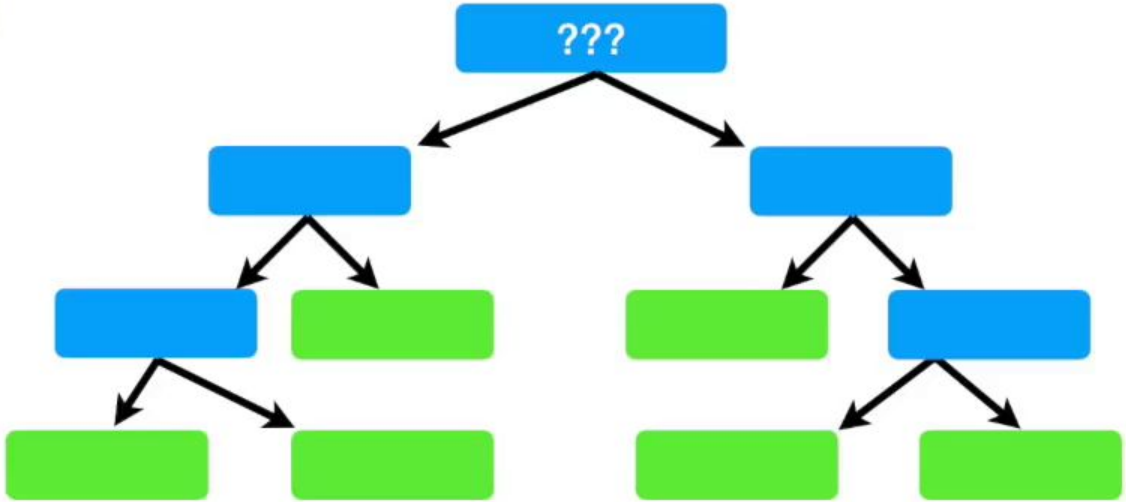
…whether or not a patient has heart disease.

| 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.

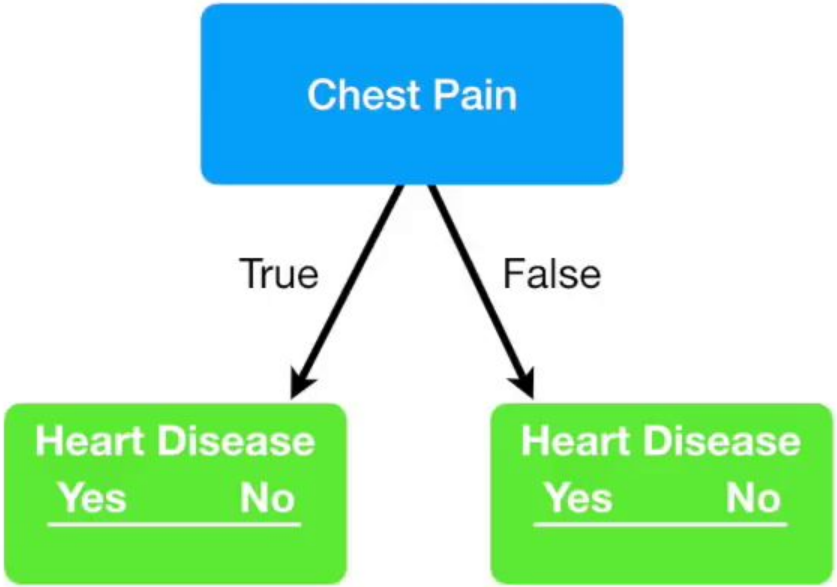


We start by looking at how well **Chest Pain** alone predicts heart disease…

| 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… |

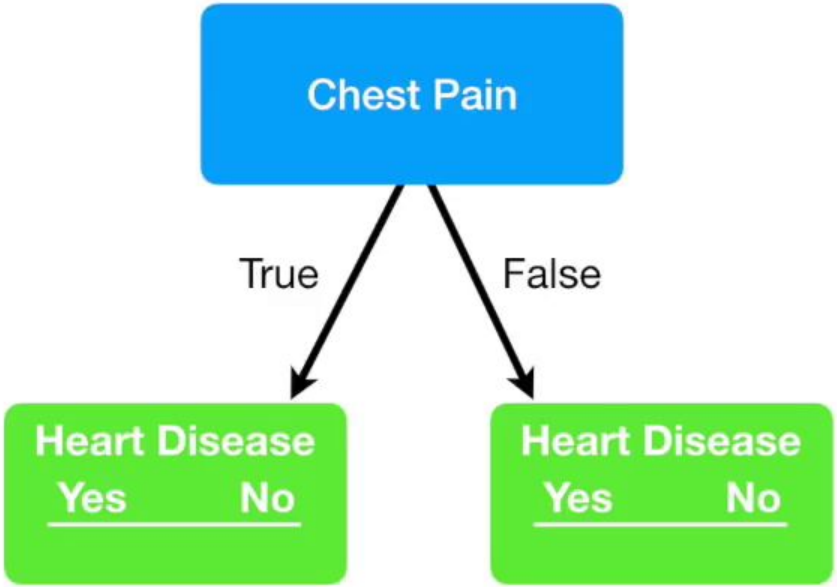
| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |

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



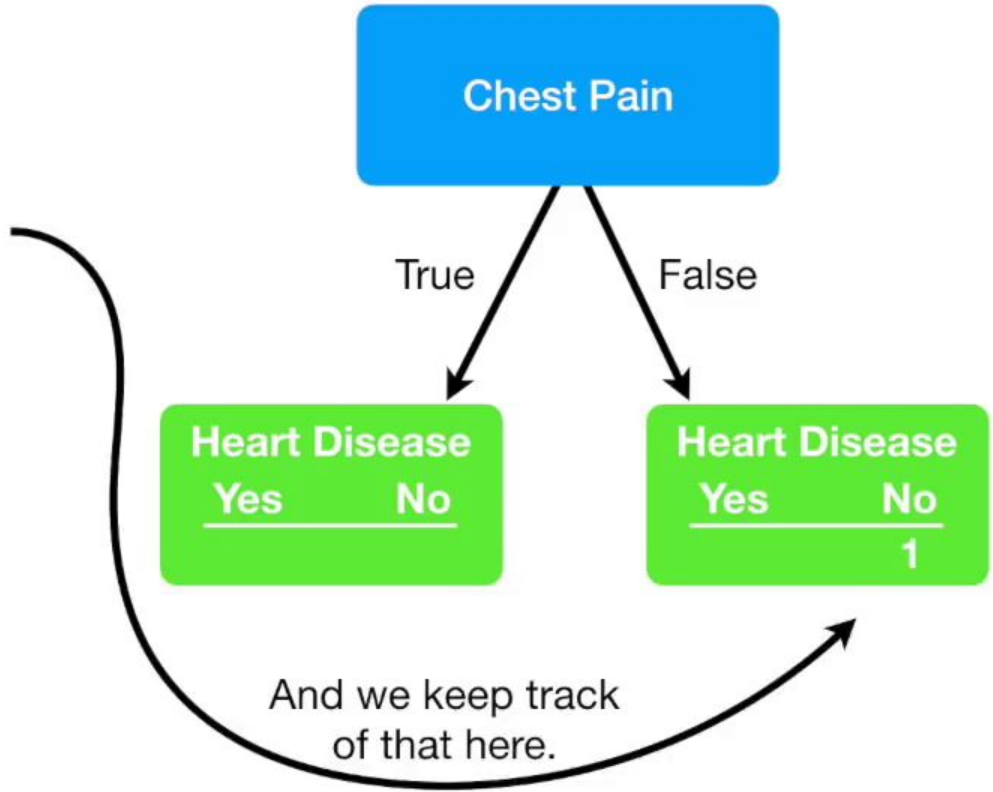
The first patient does not have chest pain and does not have heart disease.

| 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… |



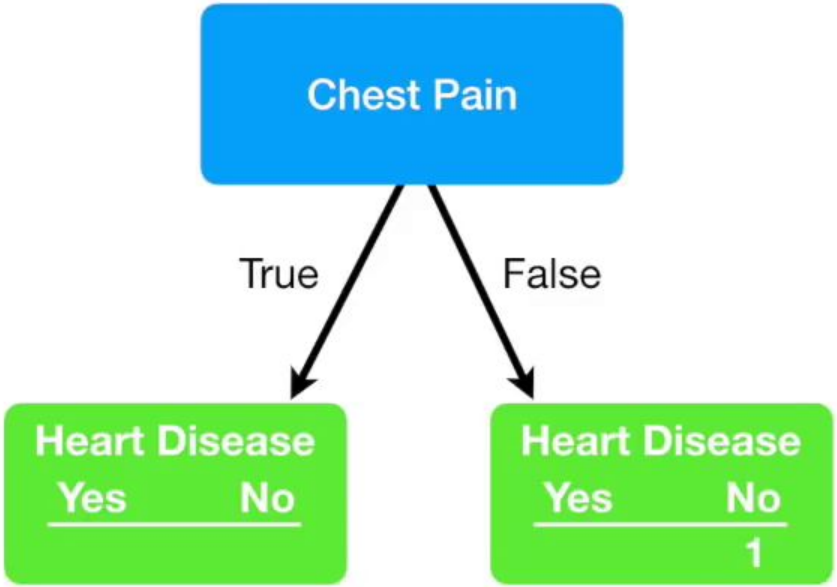
The first patient does not have chest pain and does not have heart disease.

| 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… |



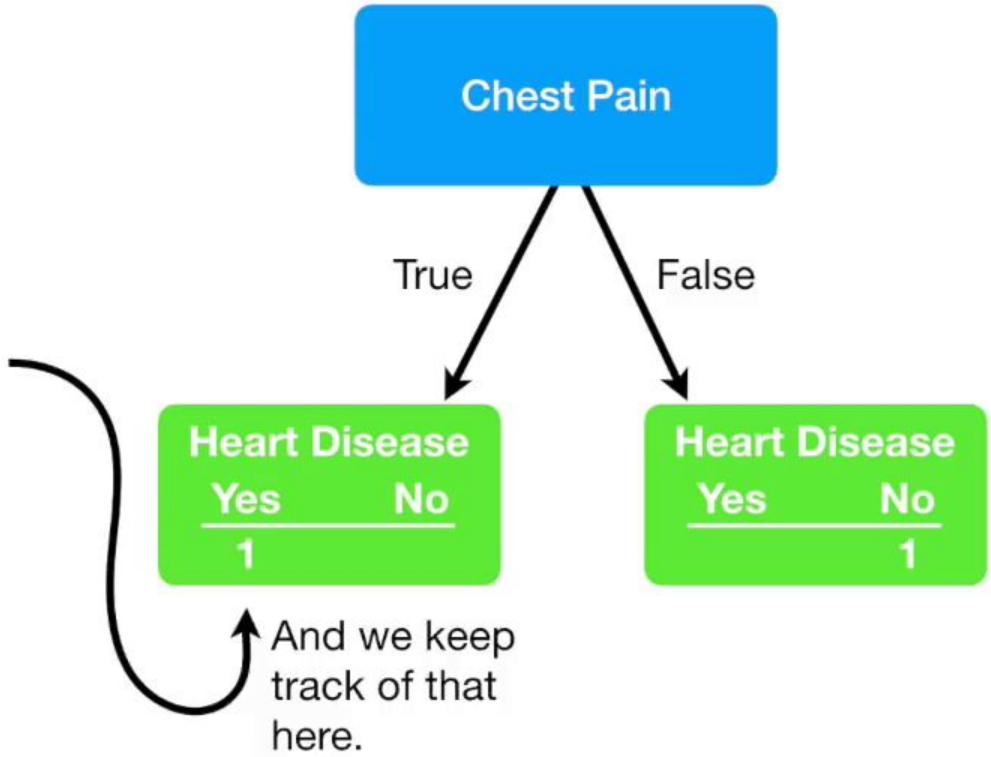
The 2nd patient has chest pain and heart disease.

| 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… |



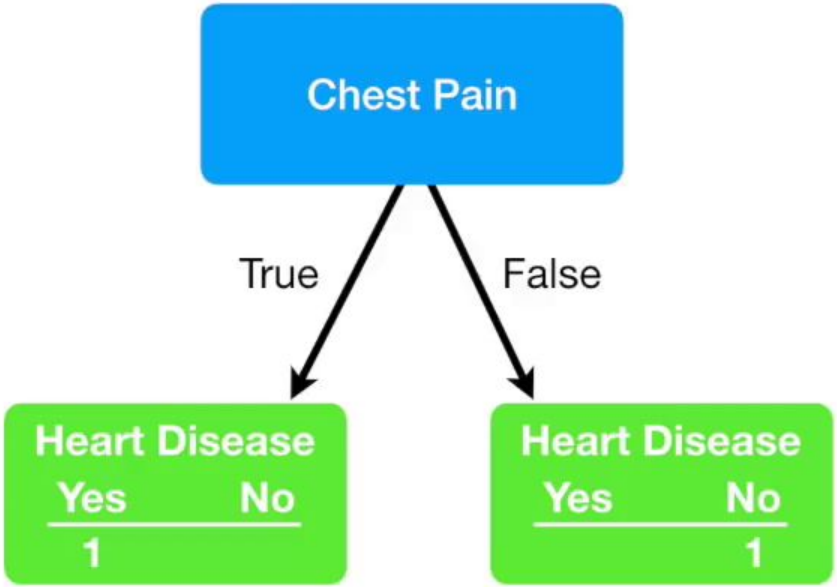
The 2nd patient has chest pain and heart disease.

| 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… |



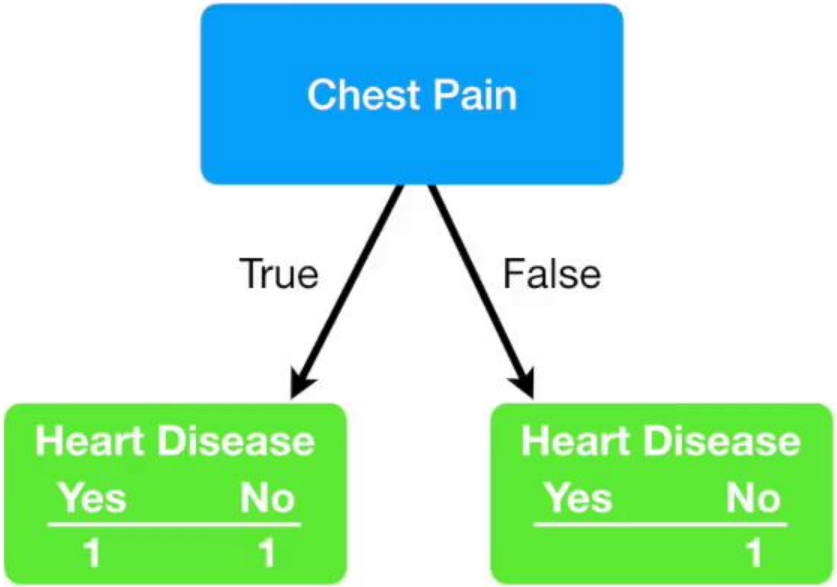
The 3rd patient has chest pain but does not have heart disease.

| 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… |

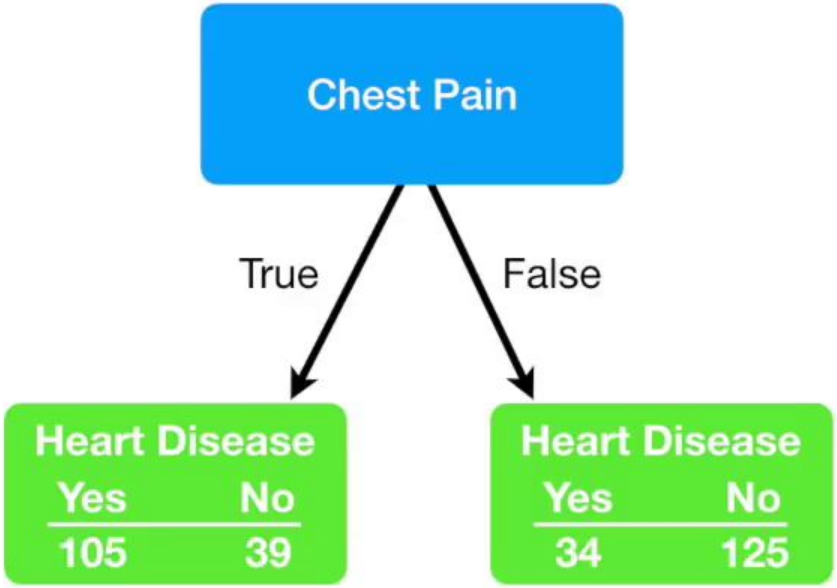


The 4th patient has chest pain and heart disease.

| 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… |



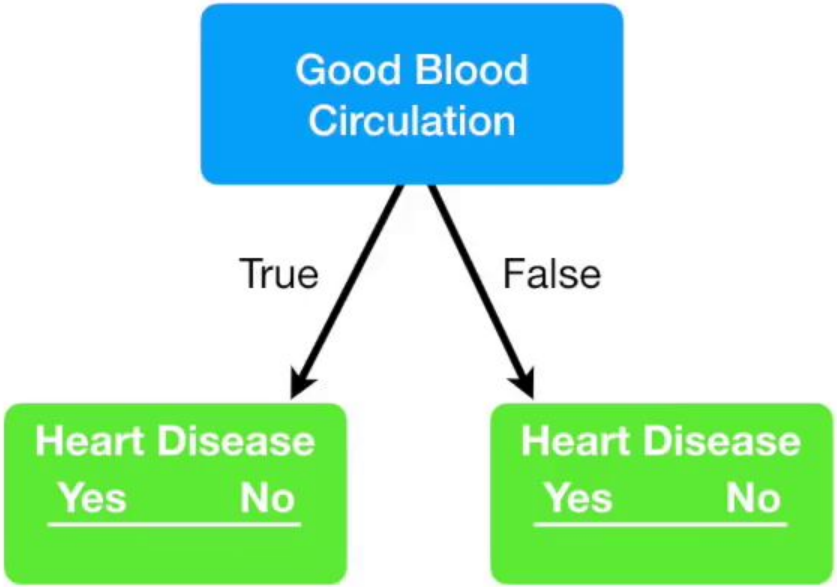
| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |




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

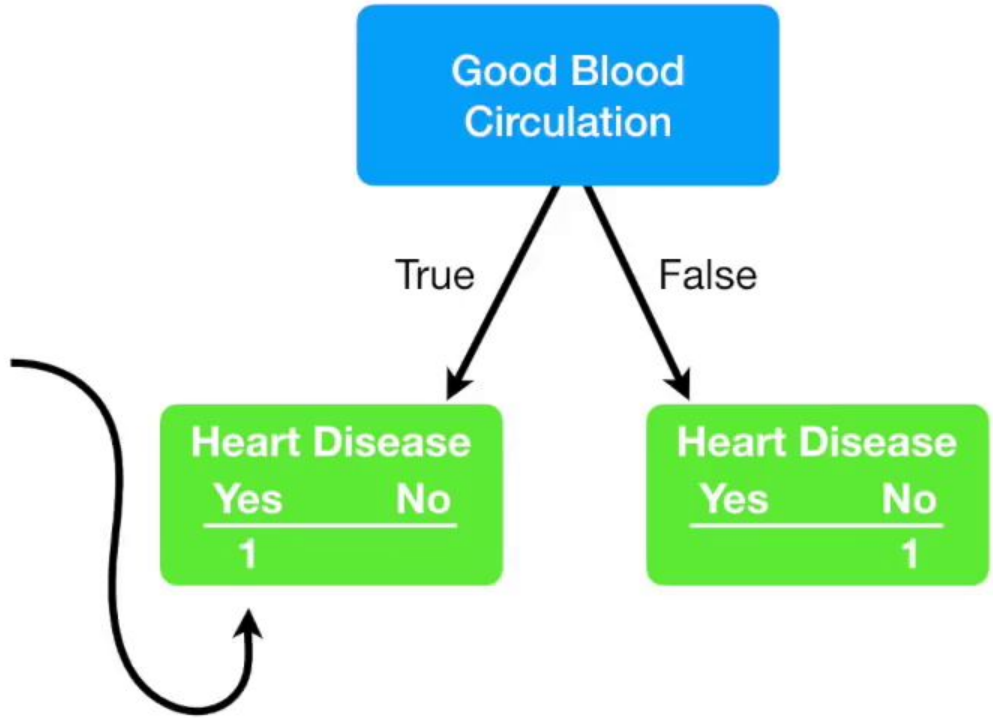
Now we do the exact same thing for **Good Blood Circulation**.

| 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… |



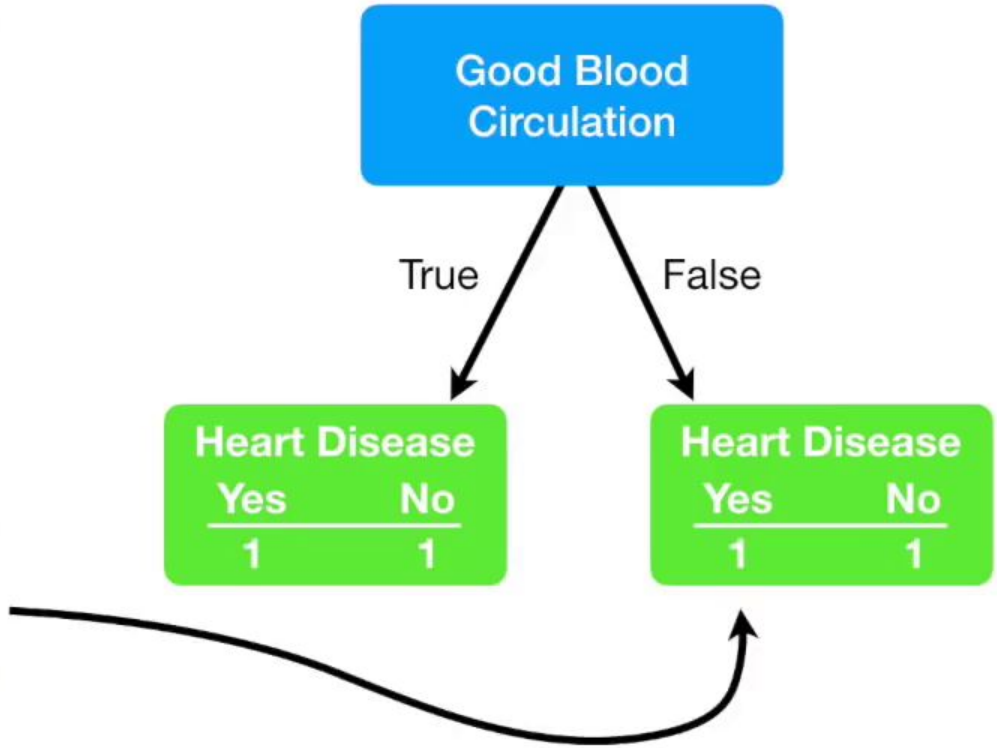
Now we do the exact same thing for **Good Blood Circulation**.

| 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… |



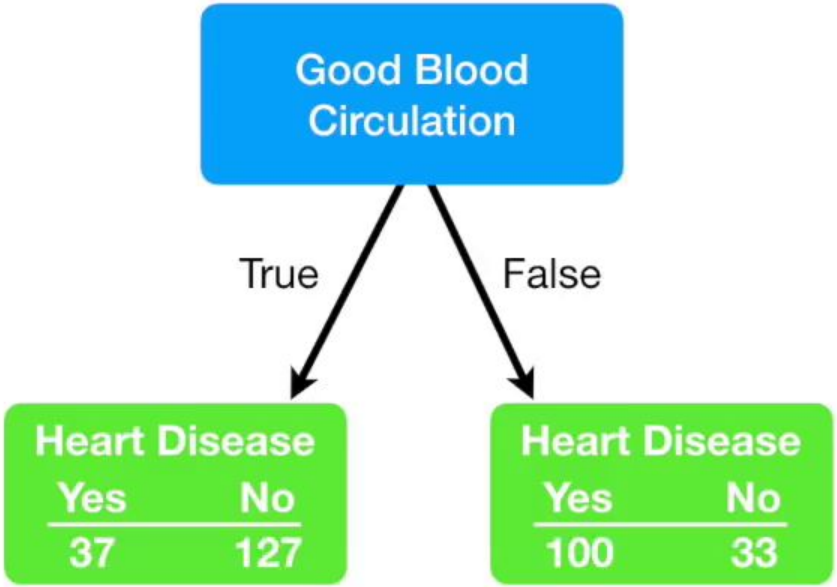
Now we do the exact same thing for **Good Blood Circulation**.

| 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… |



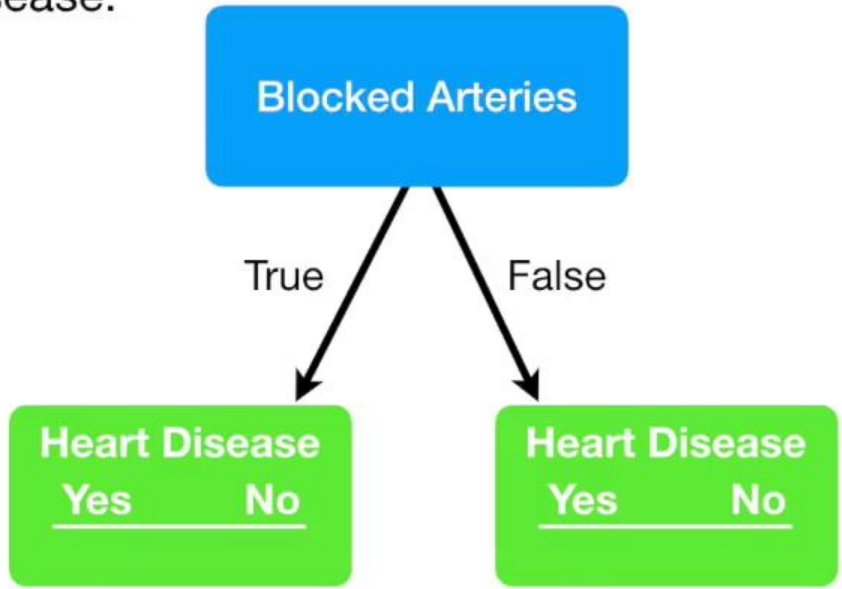
Now we do the exact same thing for **Good Blood Circulation**.

| 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… |



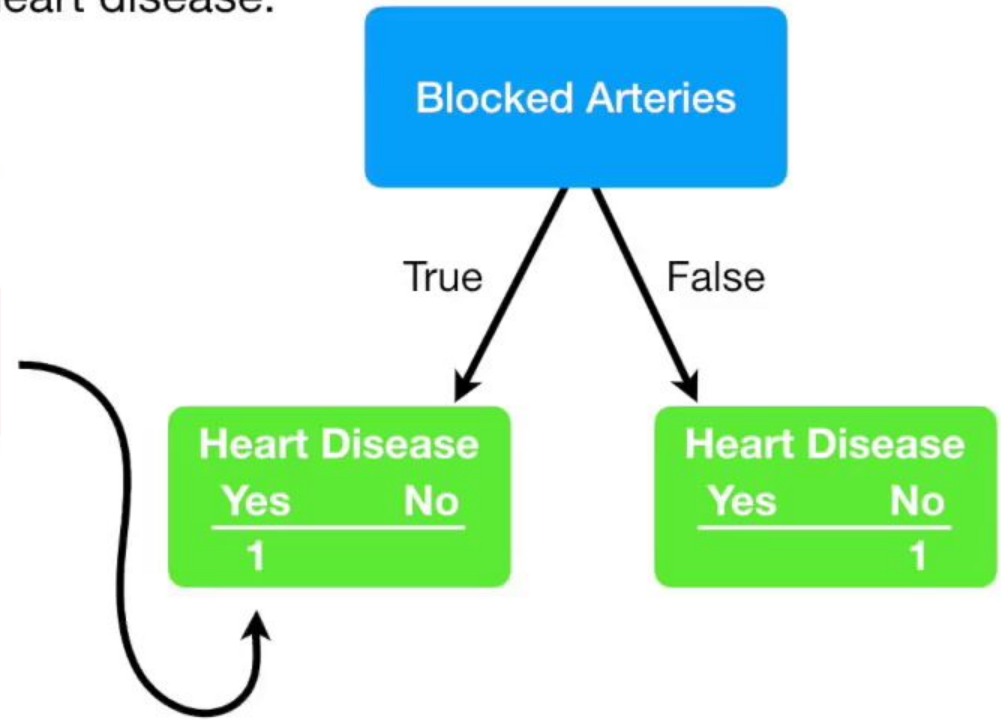
Lastly, we look at how
**Blocked Arteries**
separates the patients with
and without heart disease.

| 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… |



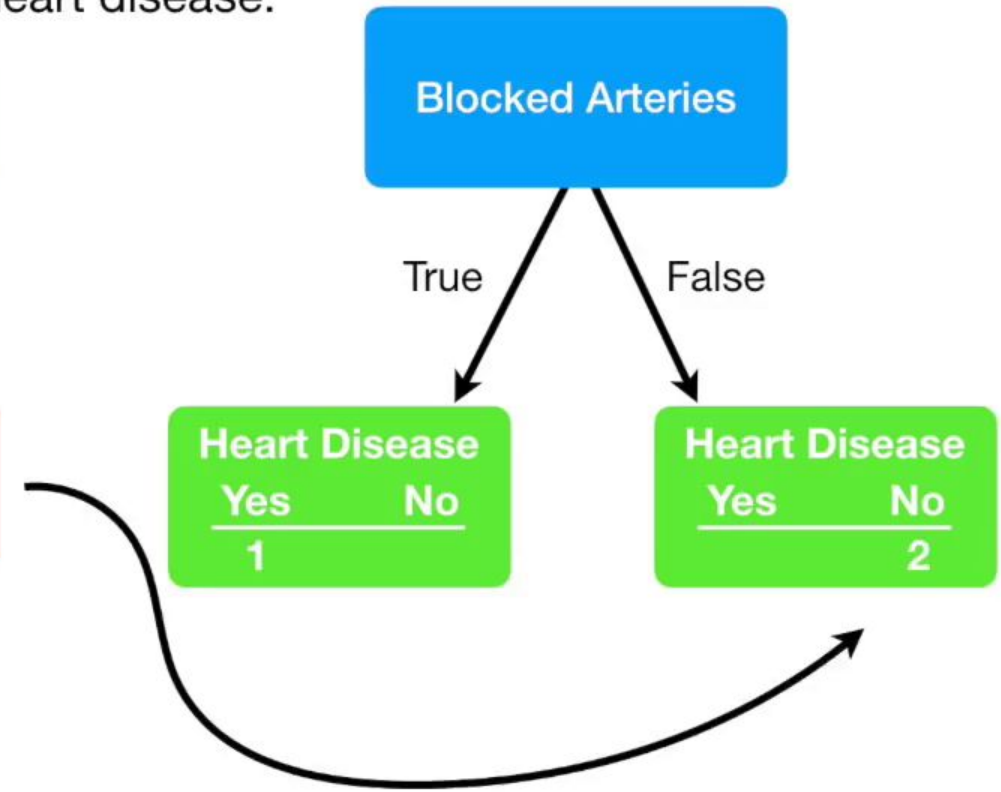
Lastly, we look at how
**Blocked Arteries**
separates the patients with
and without heart disease.

| 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… |

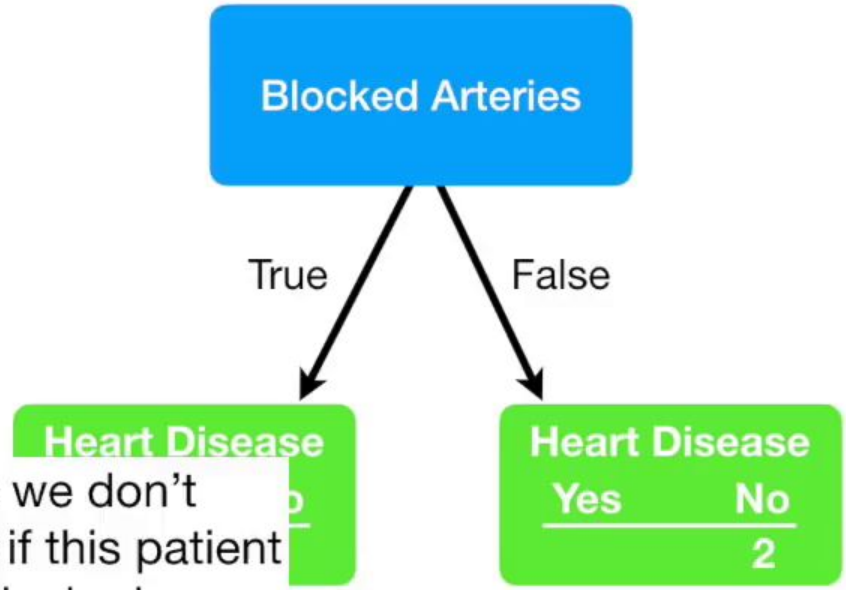


Lastly, we look at how
**Blocked Arteries**
separates the patients with
and without heart disease.

| 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… |

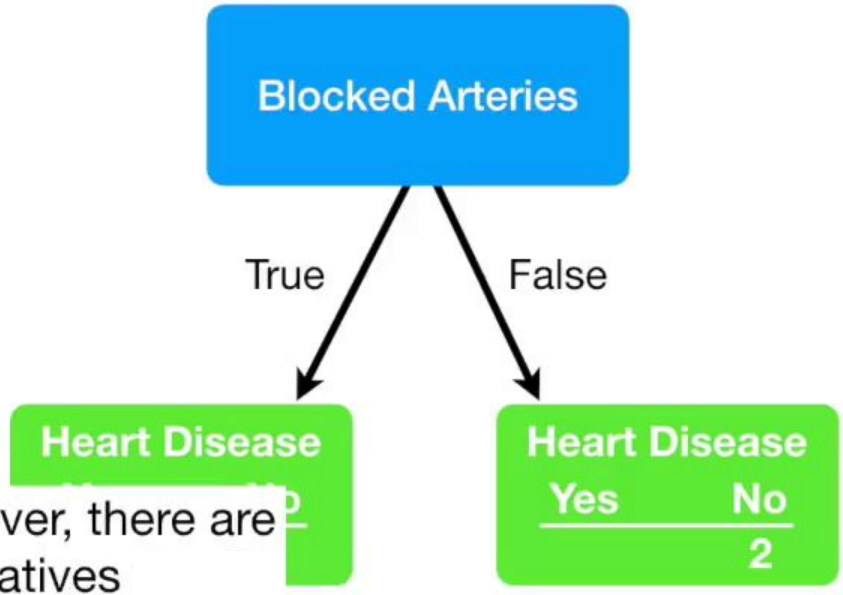


| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |




Since we don't know if this patient had blocked arteries or not, we'll skip it.

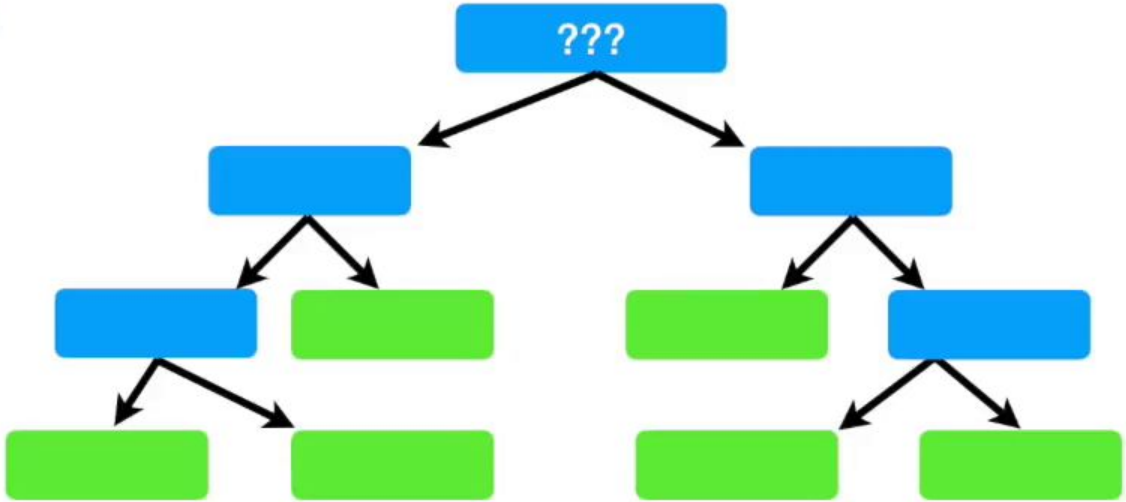
| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |



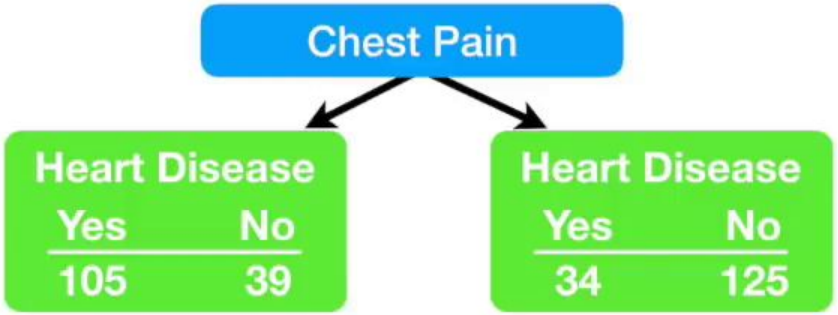

However, there are alternatives

| 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… |

Remember the goal is to decide whether **Chest Pain**, **Good Blood Circulation** or **Blocked Arteries** should be the first thing in our decision tree (aka **The Root Node**).

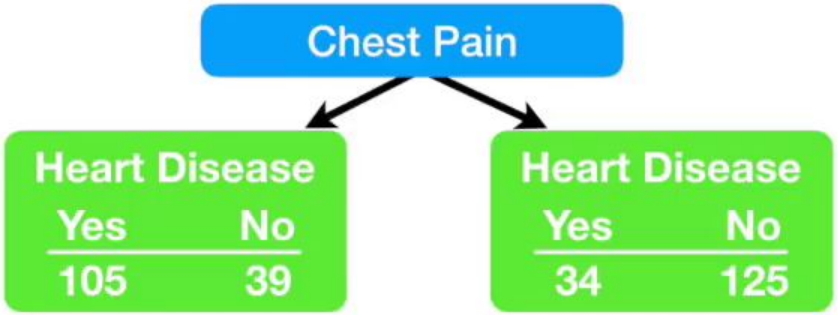


| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |




So we looked at how well **Chest Pain** separated patients with and without heart disease.

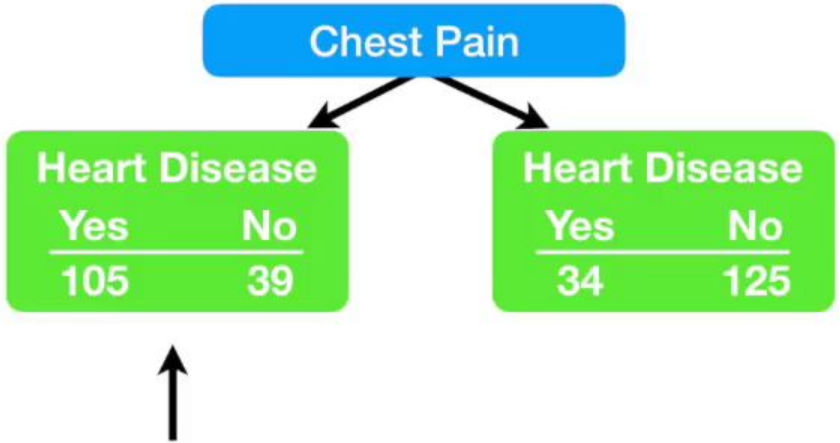
| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |



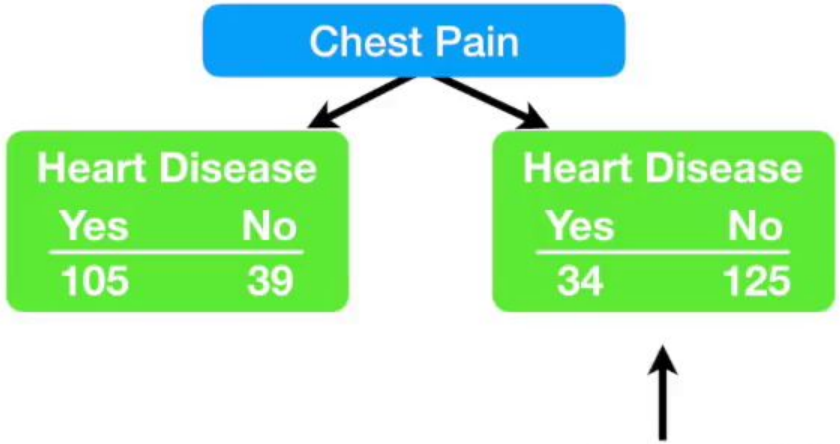

So we looked at how well **Chest Pain** separated patients with and without heart disease.

It did OK, but wasn't perfect.

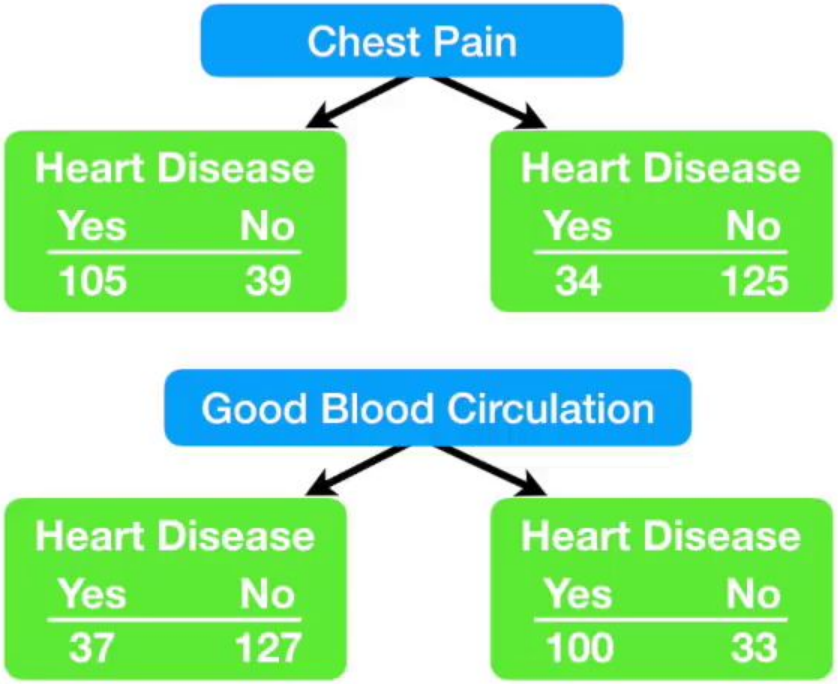
| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |



| 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… |

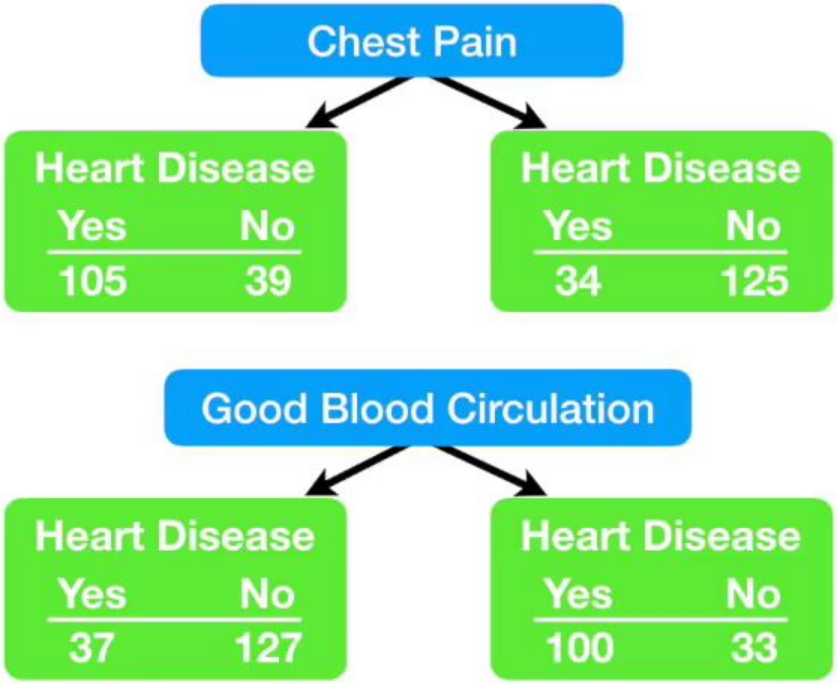


| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |



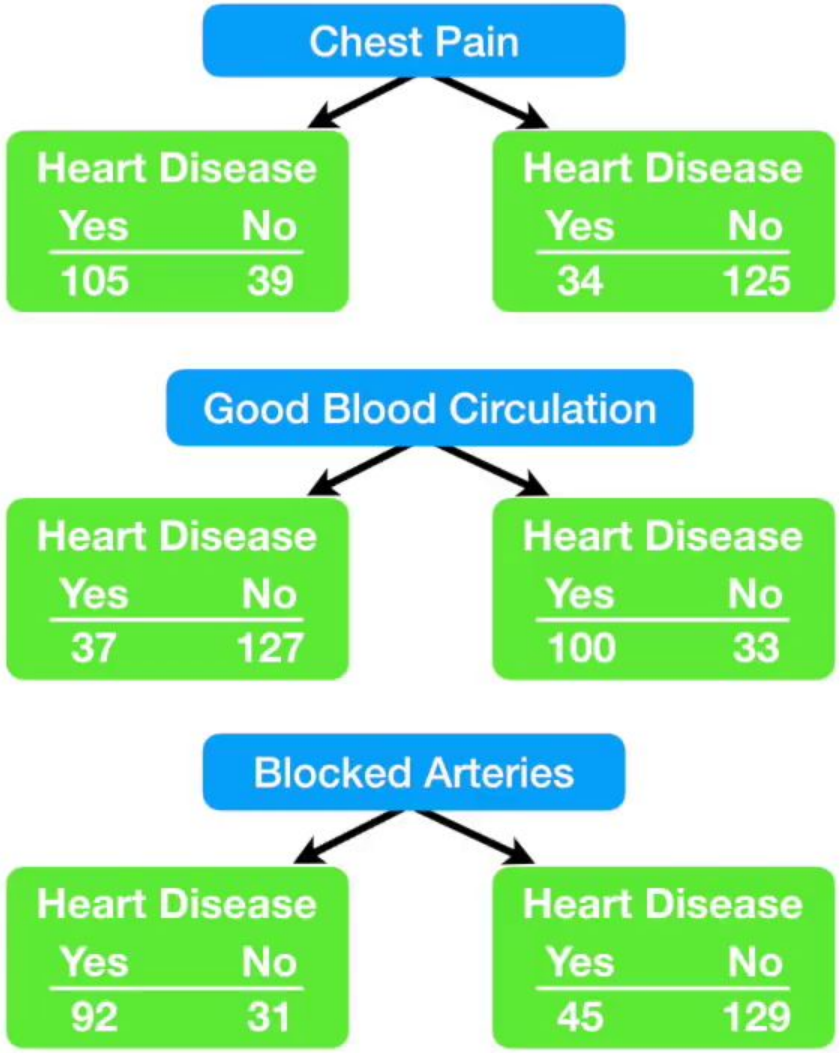

Then we looked at how well **Good Blood Circulation** separated patients with and without heart disease.

| 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… |



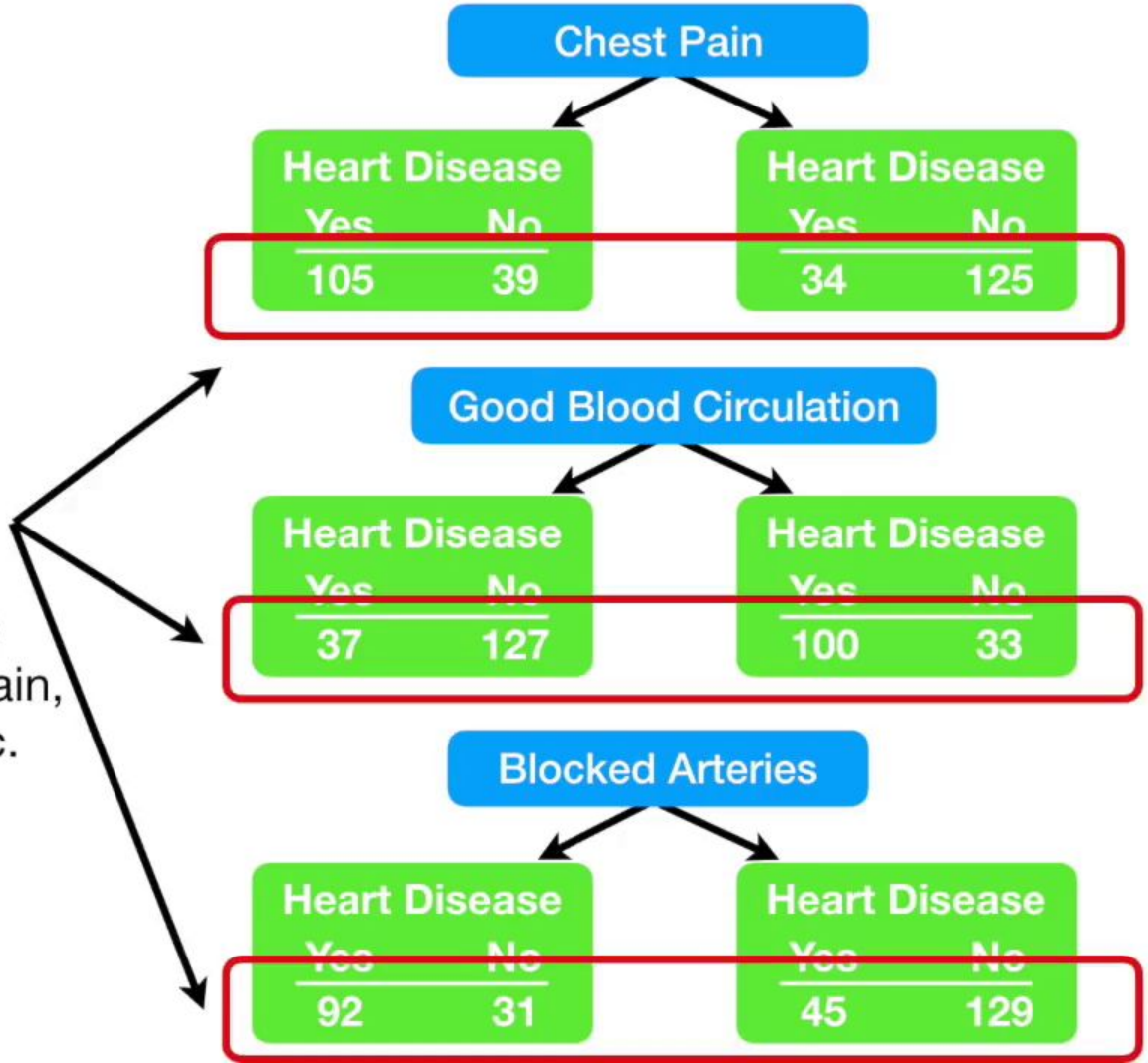

Then we looked at how well **Good Blood Circulation** separated patients with and without heart disease.

It wasn't perfect either.

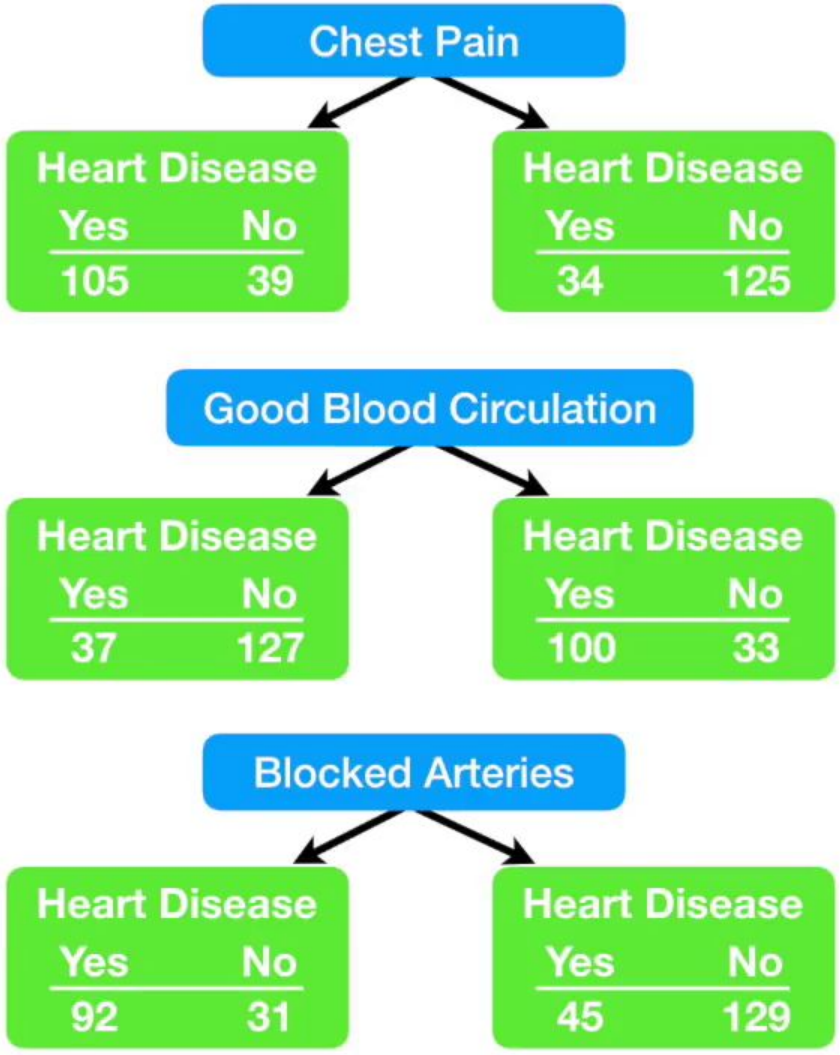



Lastly, we looked at how well **Blocked Arteries** separated patients with and without heart disease.

**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.

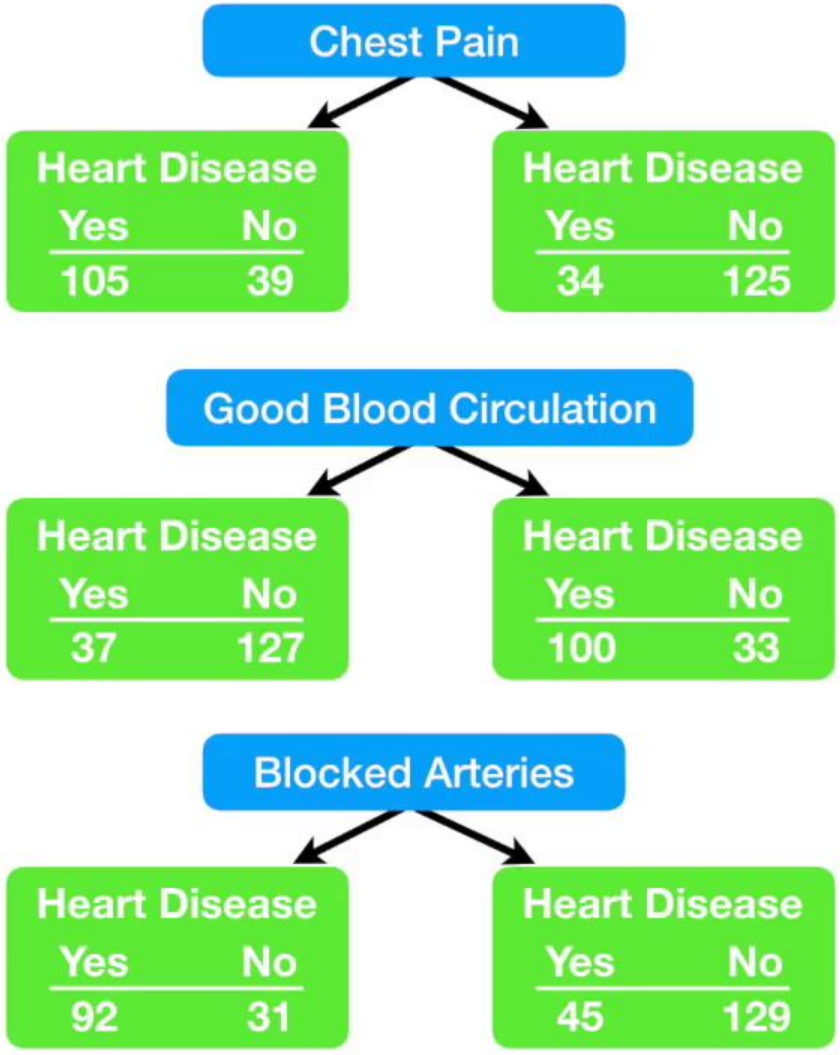


Because none of the leaf nodes are 100% “YES Heart Disease” or 100% “NO Heart Disease”, they are all considered “**impure**”.

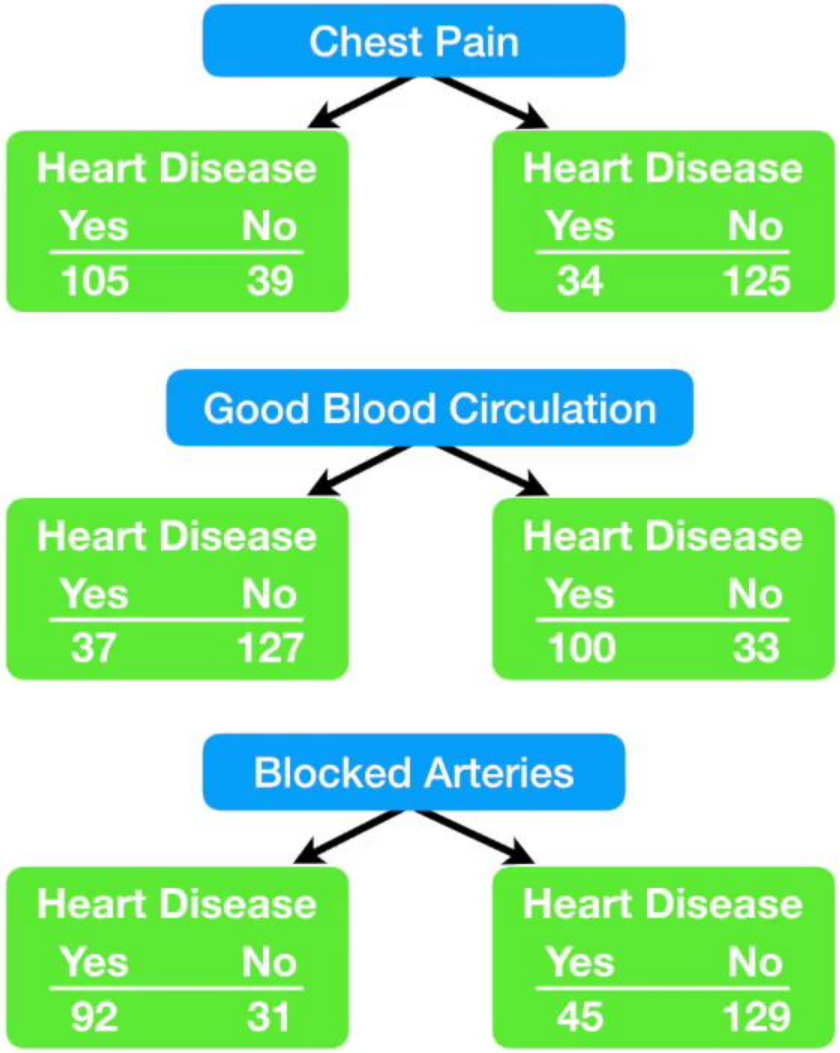


Because none of the leaf nodes are 100% “YES Heart Disease” or 100% “NO Heart Disease”, they are all considered “**impure**”.

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

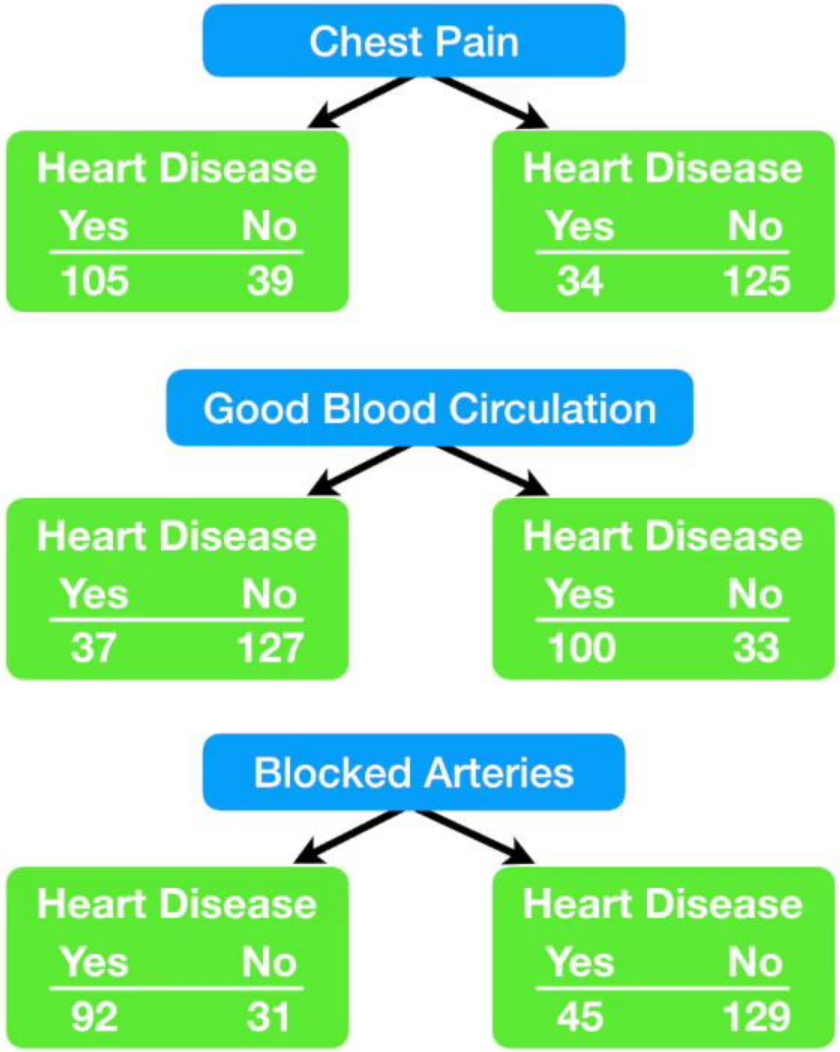


There are a bunch of ways to measure impurity, but I'm just going to focus on a very popular one called "**Gini**".

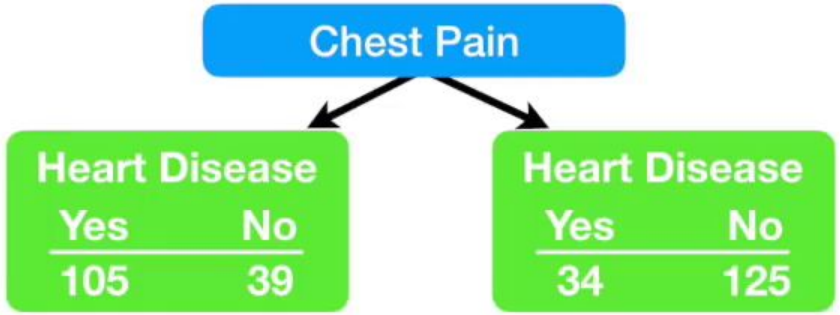


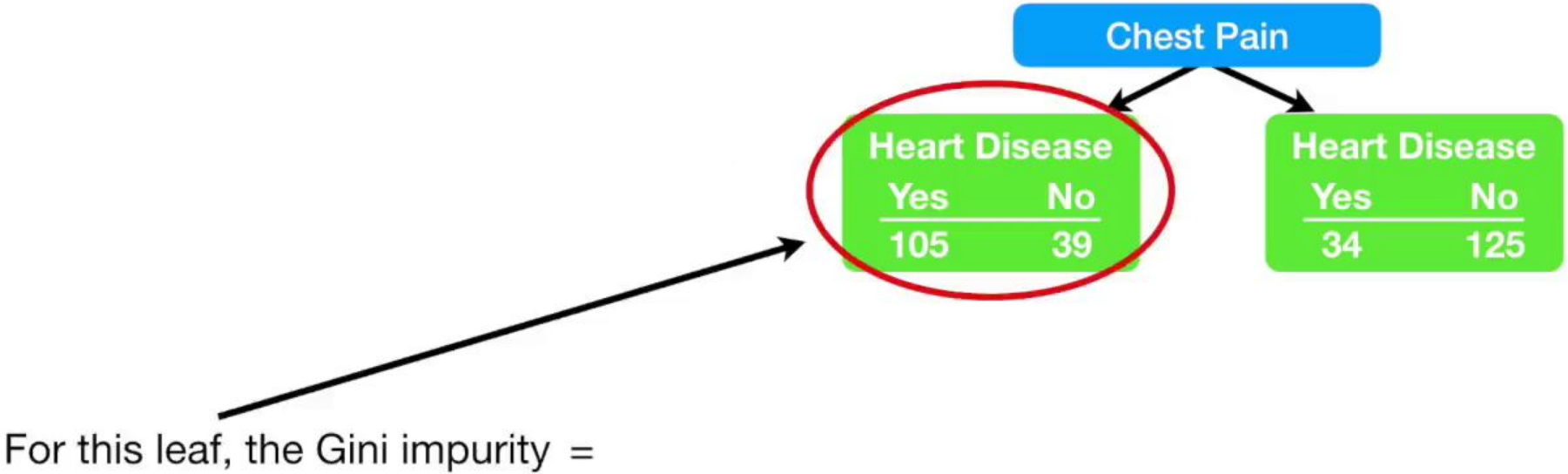
There are a bunch of ways to measure impurity, but I'm just going to focus on a very popular one called "**Gini**".

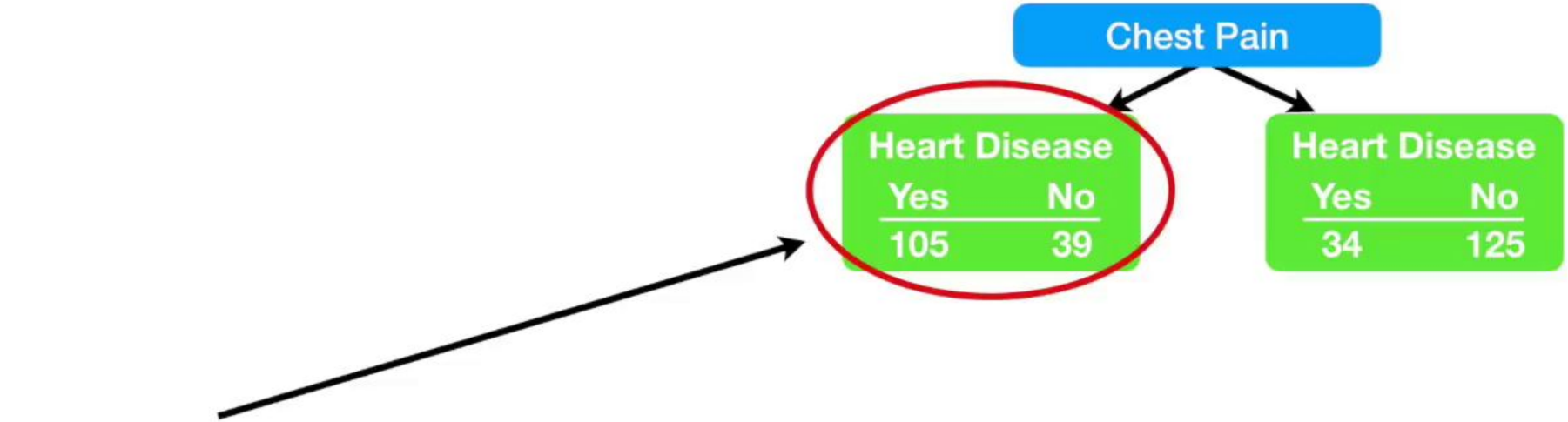
The good news is calculating Gini impurity is easy!



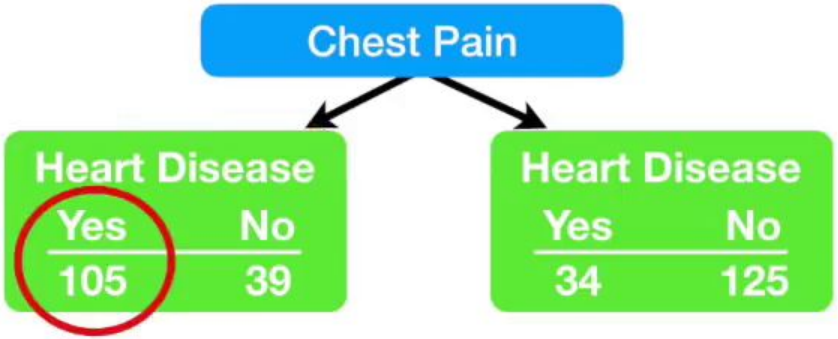
Let's start by calculating Gini impurity for Chest Pain…




Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
For this leaf, the Gini impurity =

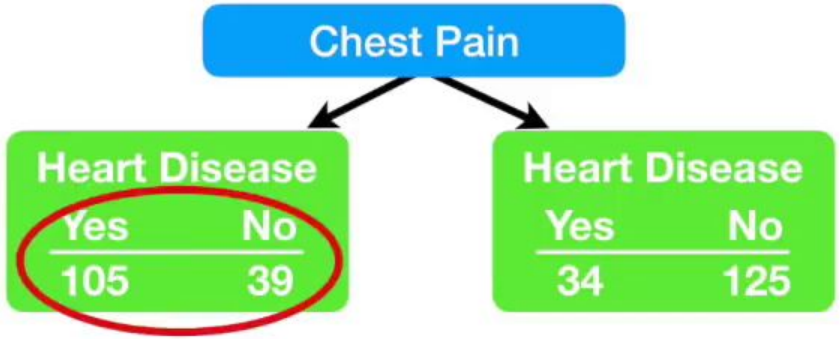



For this leaf, the Gini impurity = 1 - (the probability of “yes”)$^2$ - (the probability of “no”)$^2$



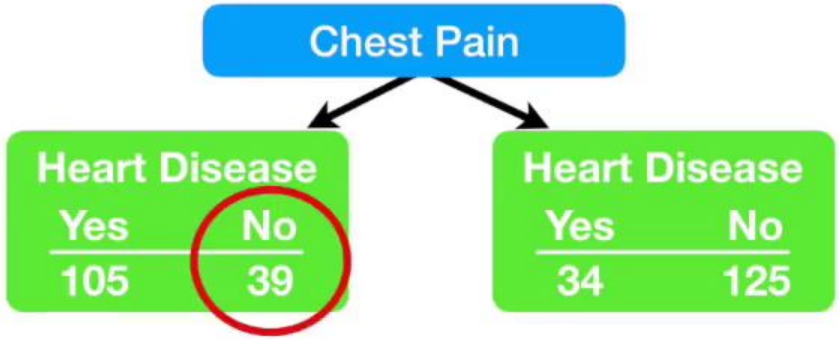

For this leaf, the Gini impurity = 1 - (the probability of "yes")$^2$ - (the probability of "no")$^2$

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



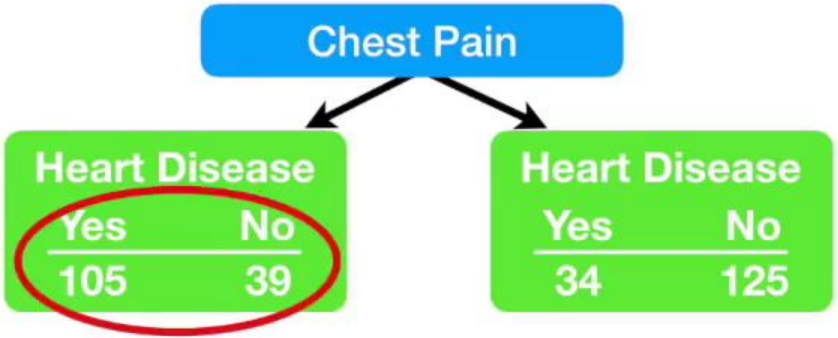

For this leaf, the Gini impurity = 1 - (the probability of "yes")$^2$ - (the probability of "no")$^2$

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



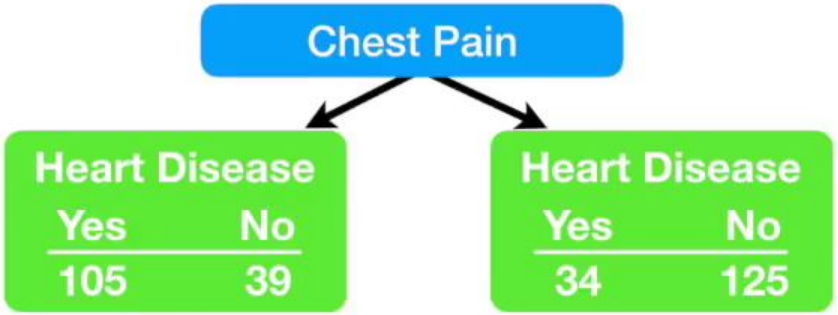

For this leaf, the Gini impurity = 1 - (the probability of "yes")$^2$ - (the probability of "no")$^2$

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




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

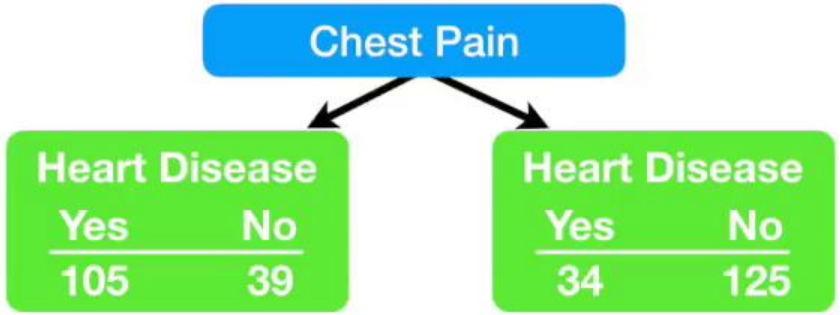
$$= 1 - \left(\frac{105}{105 + 39}\right)^2 - \left(\frac{39}{105 + 39}\right)^2$$



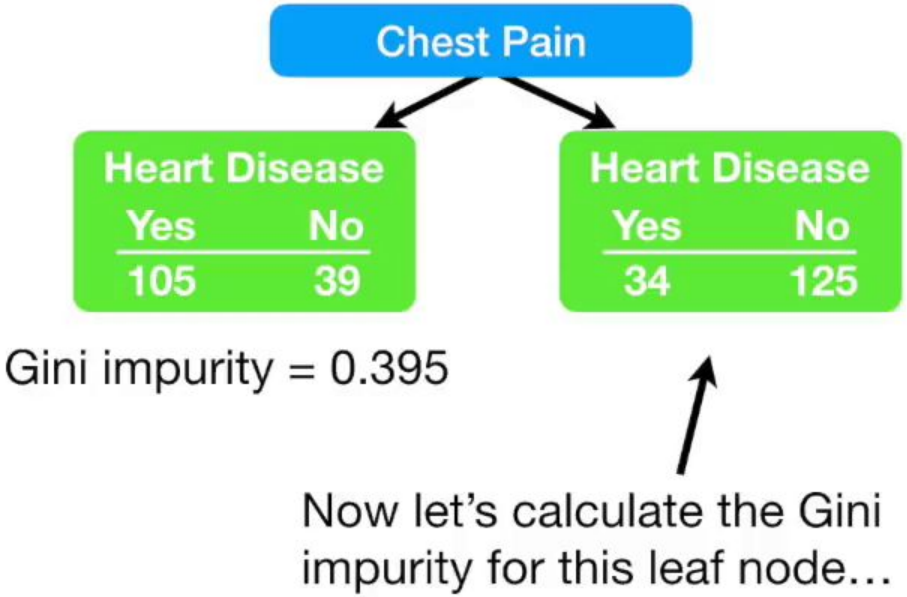

For this leaf, the Gini impurity = 1 - (the probability of "yes")$^2$ - (the probability of "no")$^2$

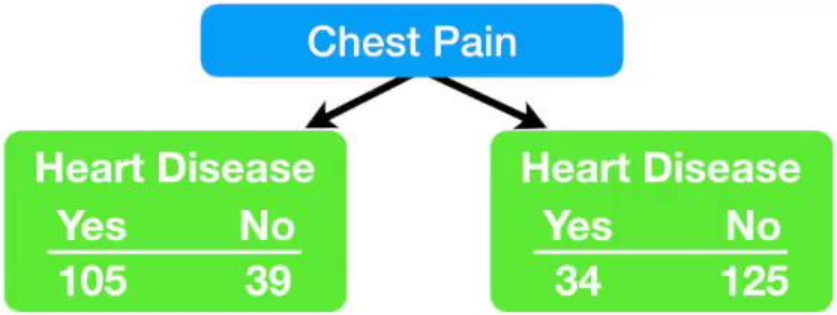
$$= 1 - \left(\frac{105}{105 + 39}\right)^2 - \left(\frac{39}{105 + 39}\right)^2$$

$$= 0.395$$

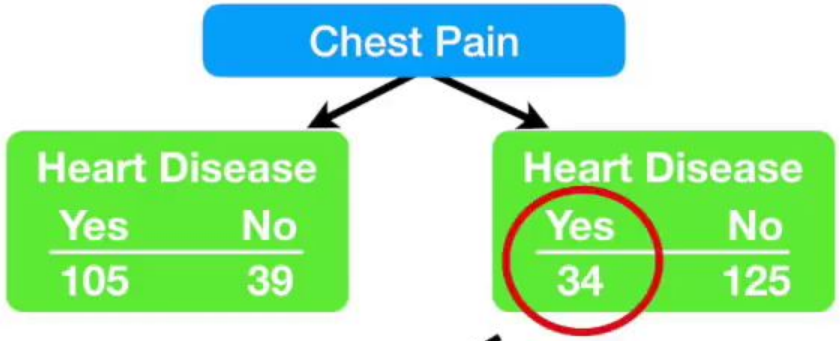



Gini impurity = 0.395


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…

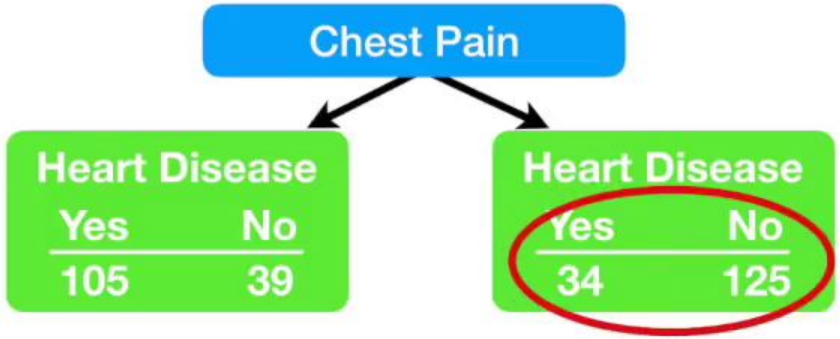



= 1 - (the probability of “yes”)$^2$ - (the probability of “no”)$^2$



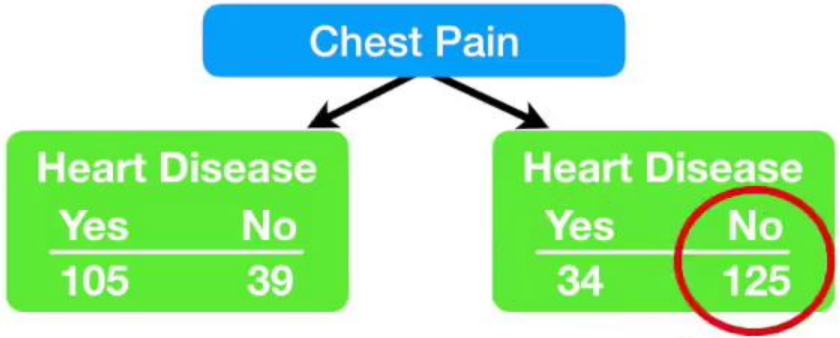

$$= 1 - (\text{the probability of “yes”})^2 - (\text{the probability of “no”})^2$$

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



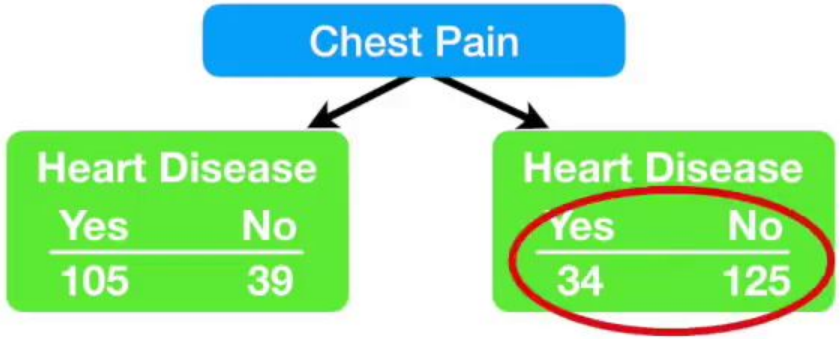

$$= 1 - (\text{the probability of “yes”})^2 - (\text{the probability of “no”})^2$$

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



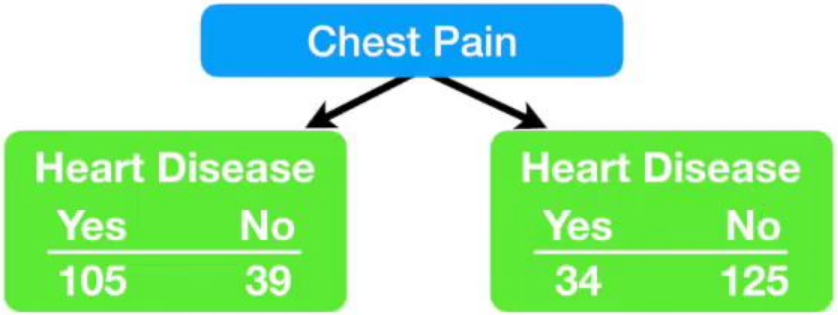

$$= 1 - (\text{the probability of “yes”})^2 - (\text{the probability of “no”})^2$$

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




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

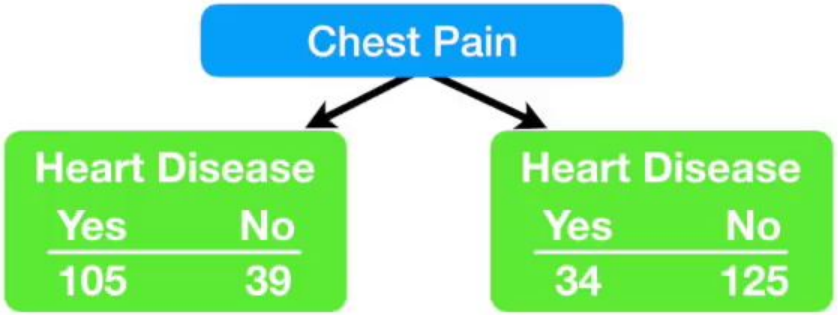
$$= 1 - \left(\frac{34}{34 + 125}\right)^2 - \left(\frac{125}{34 + 125}\right)^2$$




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

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

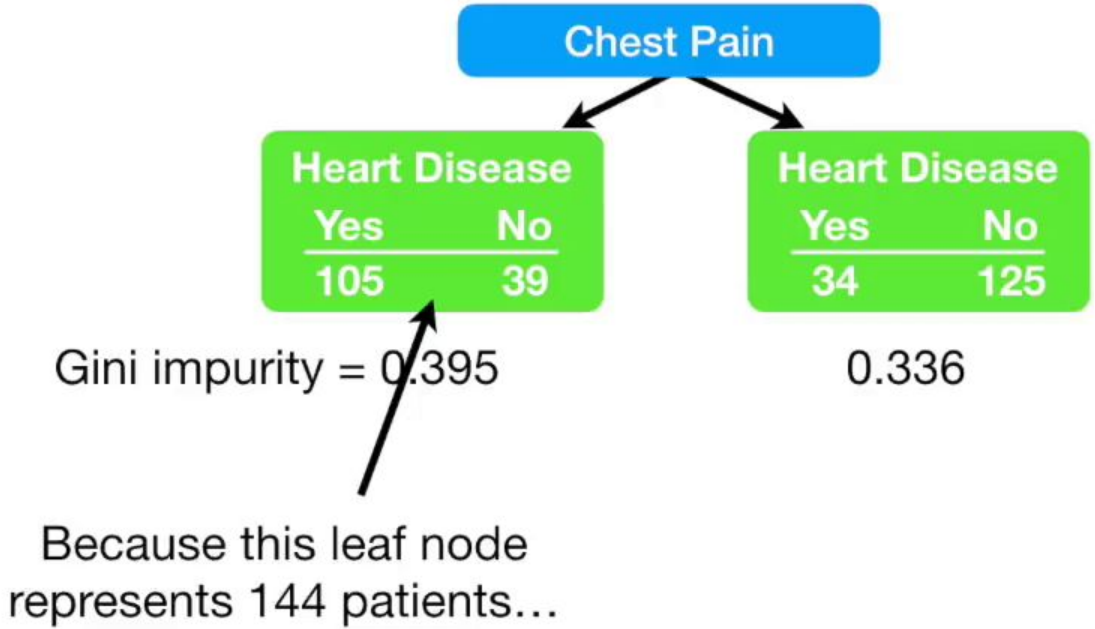
$$= 0.336$$

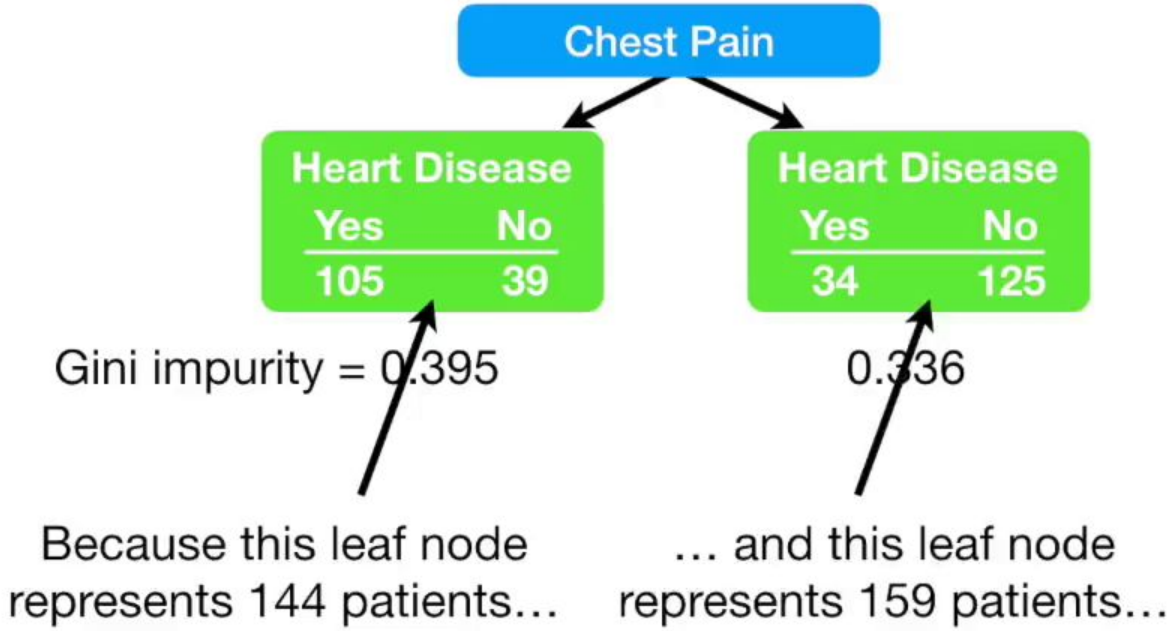


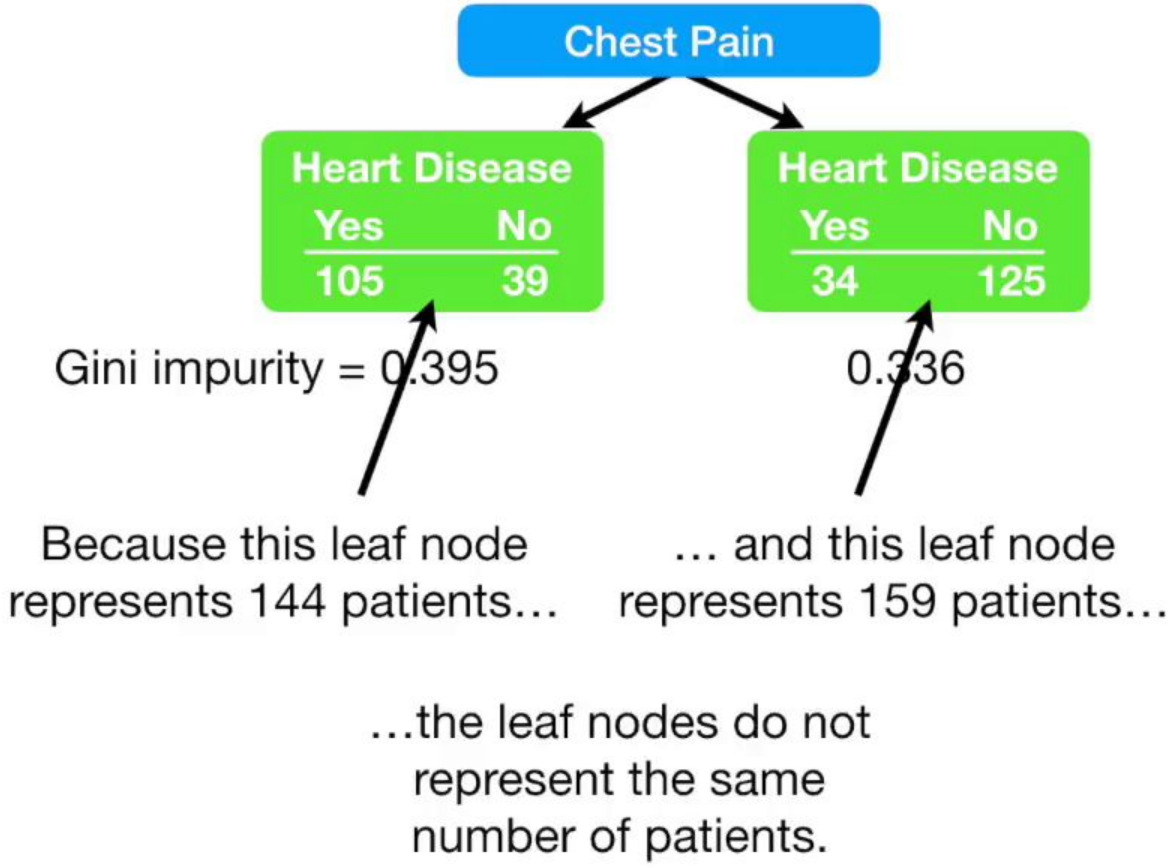

Gini impurity = 0.395

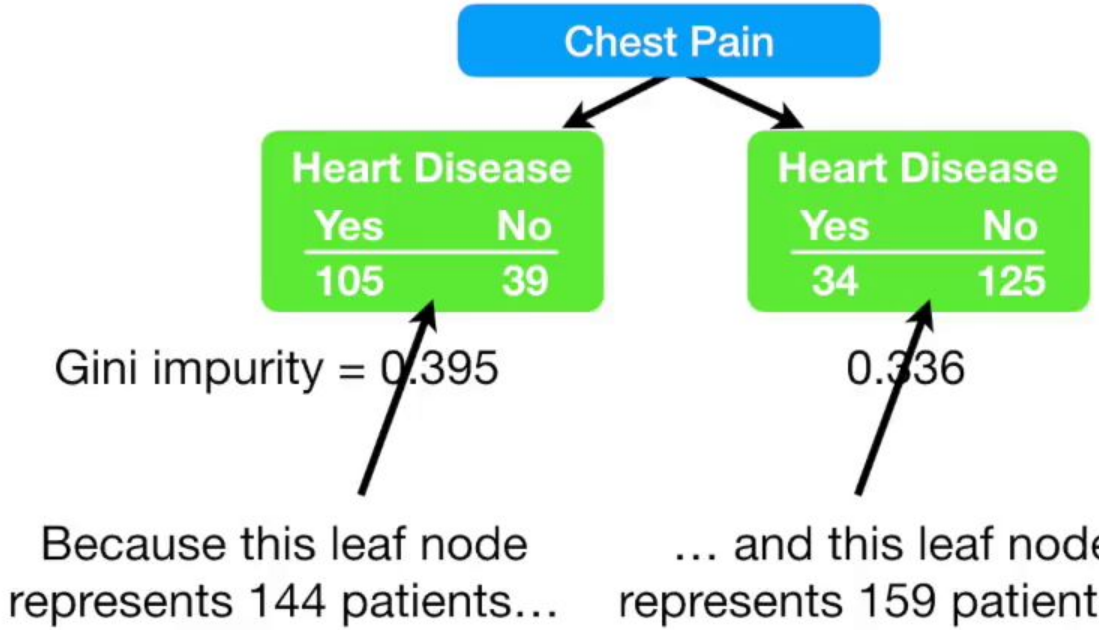
0.336

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.

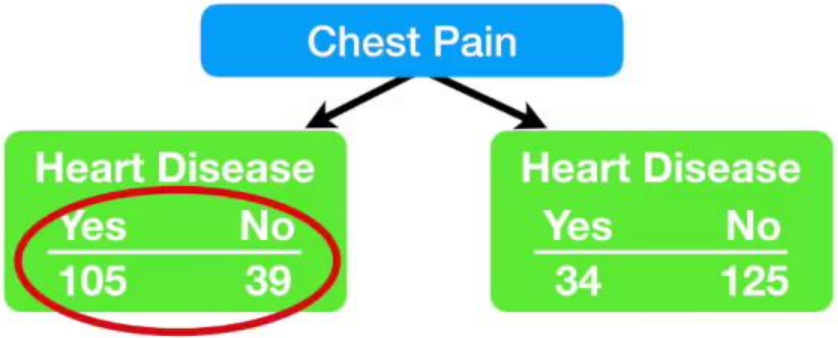

Chest Pain
Heart Disease
Yes No
105 39
Heart Disease
Yes No
34 125
Gini impurity = 0.395
0.336
Because this leaf node
represents 144 patients…


Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Gini impurity = 0.395
0.336
Because this leaf node represents 144 patients…
… and this leaf node represents 159 patients…


Chest Pain
Heart Disease
Yes
No
105
39
Heart Disease
Yes
No
34
125
Gini impurity = 0.395
0.336
Because 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**.

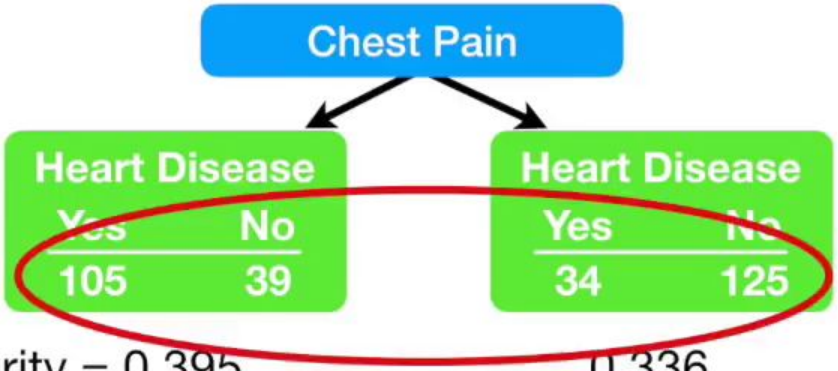



Gini impurity = 0.395

0.336

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

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

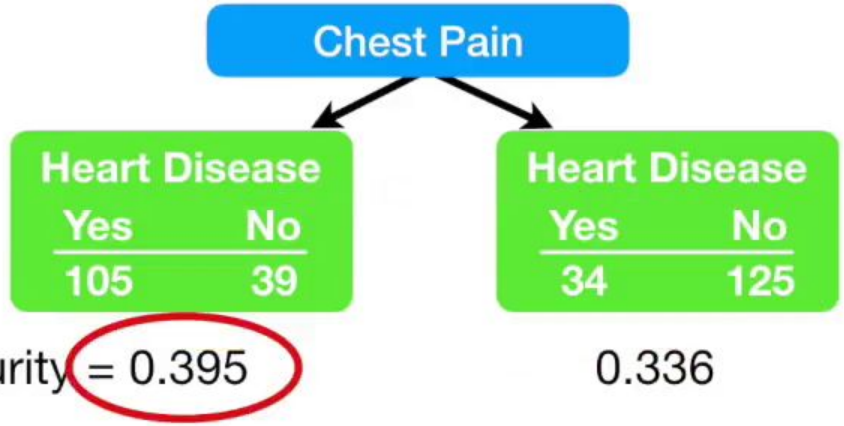



Gini impurity = 0.395

0.336

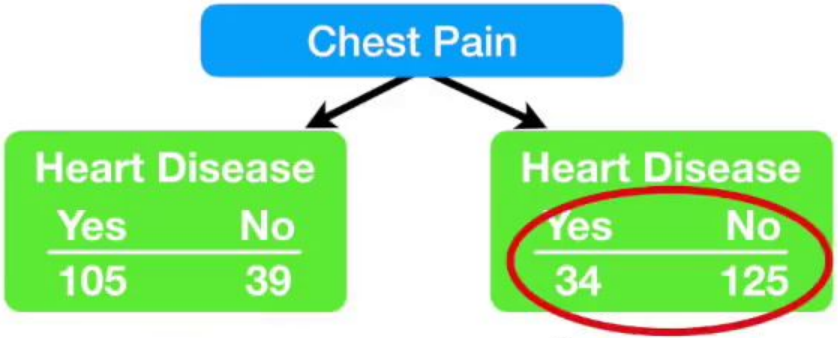
Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes

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




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

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

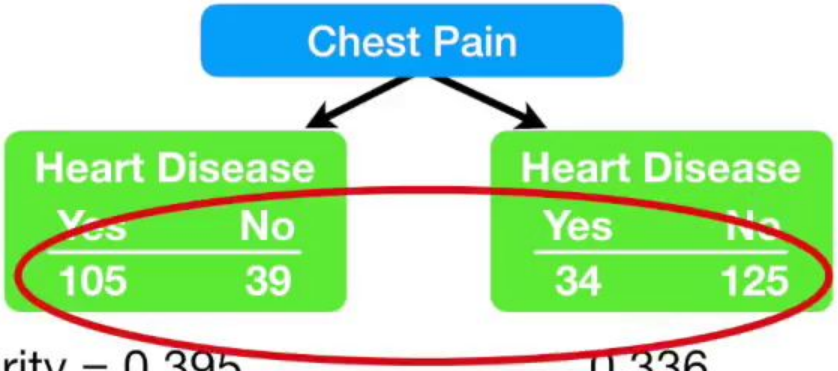



Gini impurity = 0.395

0.336

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

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

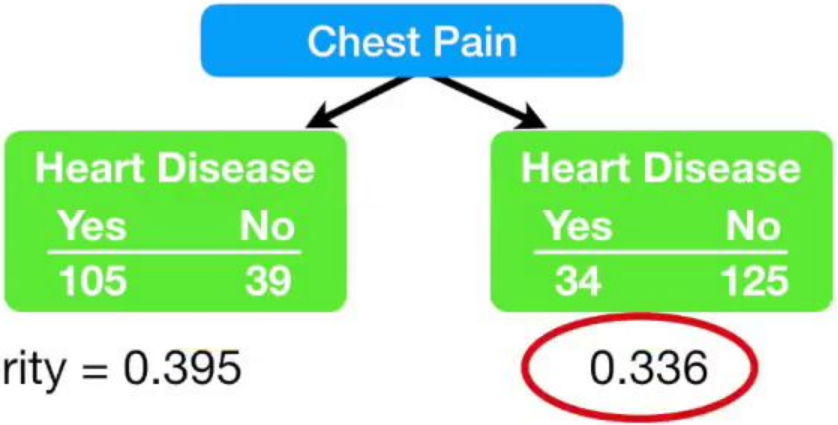



Gini impurity = 0.395

0.336

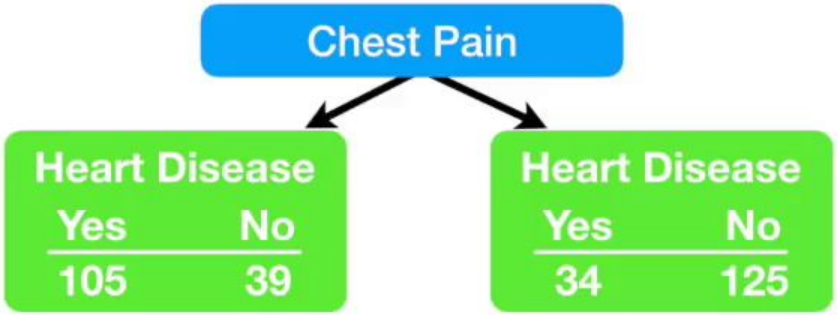
Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes

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




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

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




Gini impurity = 0.395

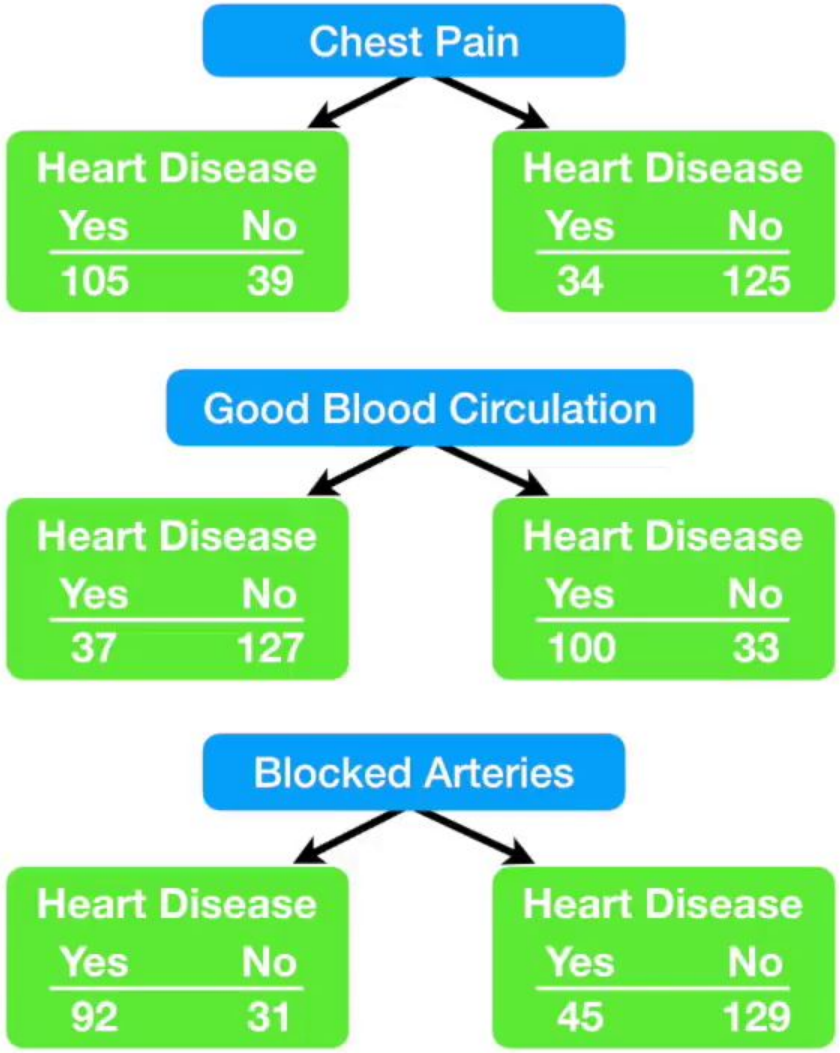
0.336

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

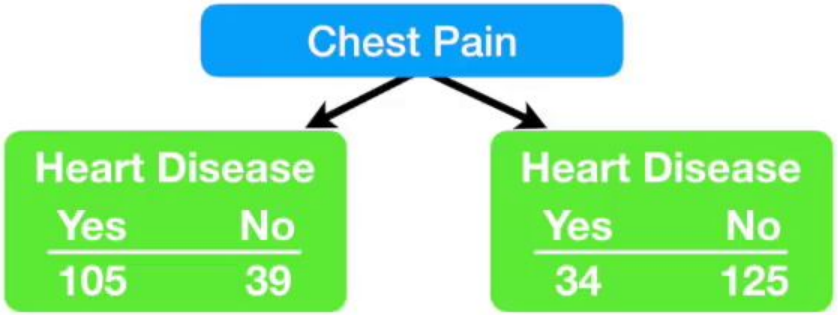
$$= \left(\frac{144}{144 + 159}\right) 0.395 + \left(\frac{159}{144 + 159}\right) 0.336$$

$$= 0.364$$

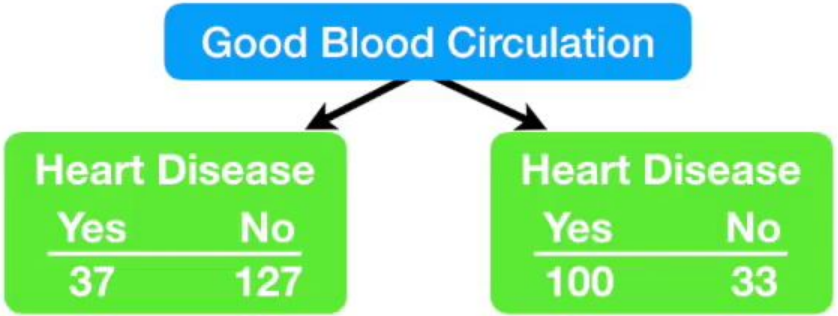
Gini impurity for Chest Pain = 0.364

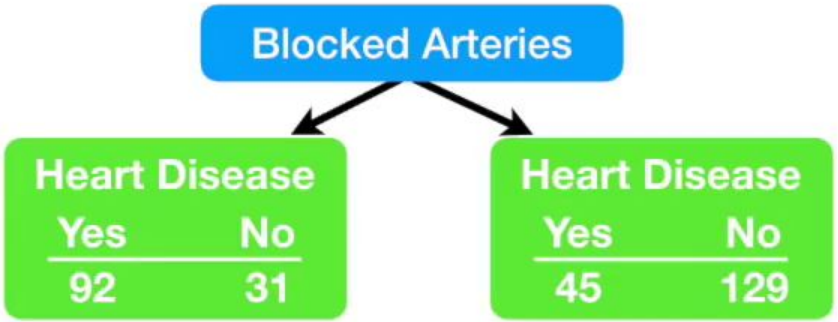


Gini impurity for Chest Pain = 0.364

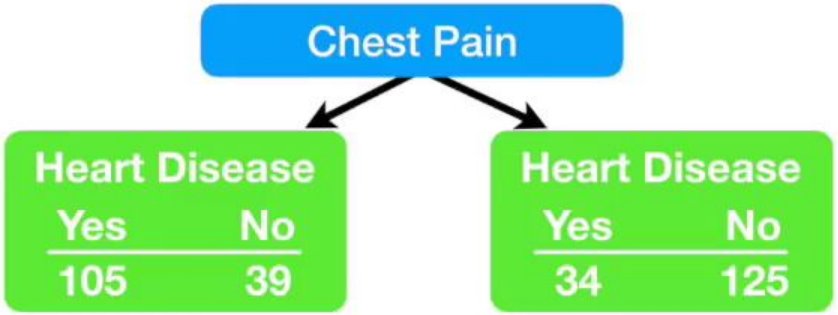



Gini impurity for Good Blood Circulation = 0.360

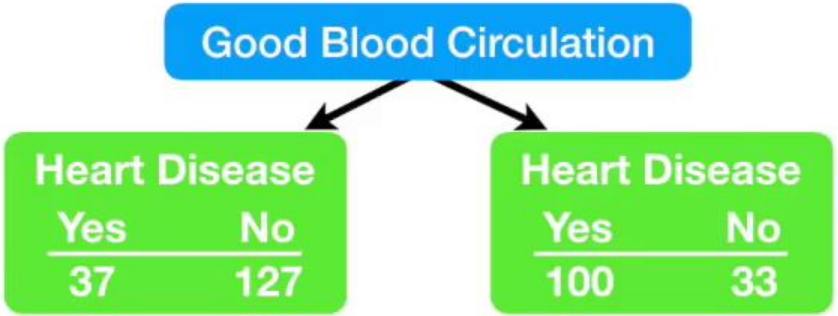


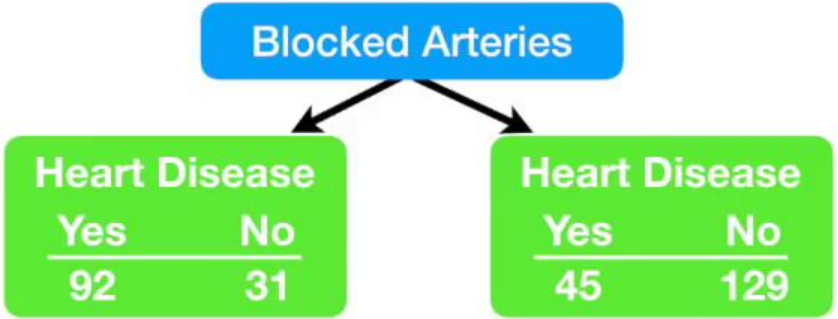
Gini impurity for Chest Pain = 0.364




Gini impurity for Good Blood Circulation = 0.360




Gini impurity for Blocked Arteries = 0.381



Gini impurity for Chest Pain = 0.364

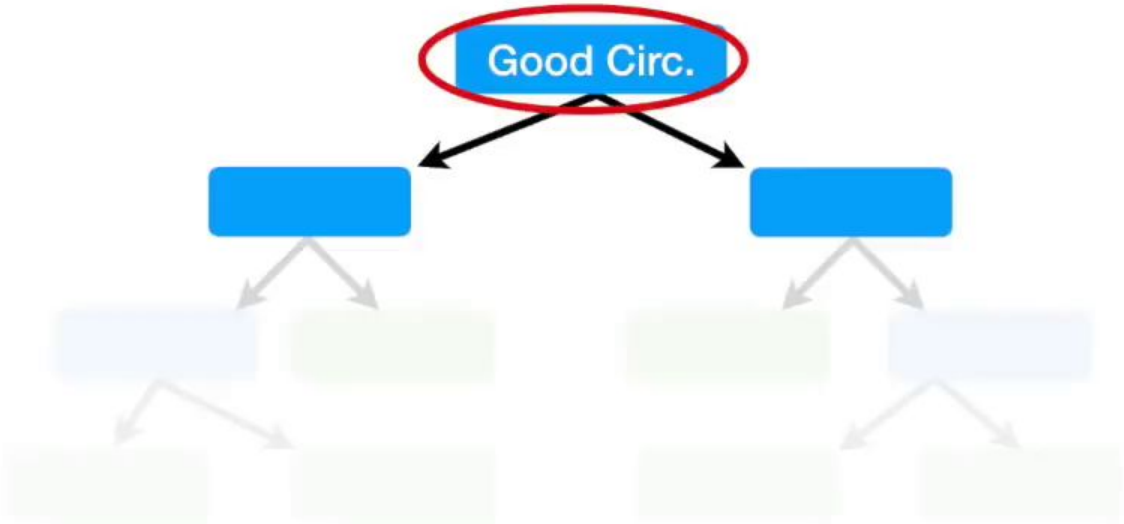
Gini impurity for Good Blood Circulation = 0.360

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

Gini impurity for Blocked Arteries = 0.381

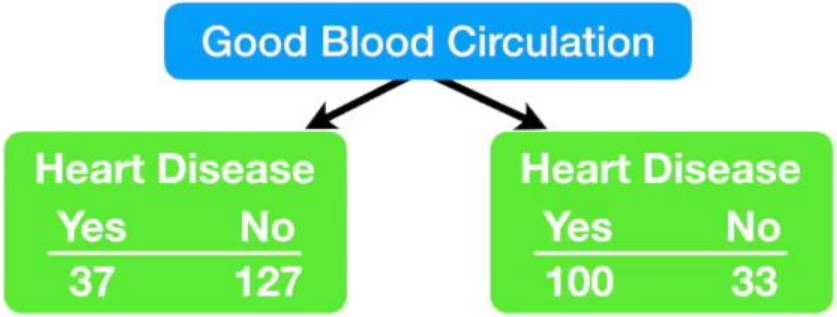
Gini impurity for Chest Pain = 0.364

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

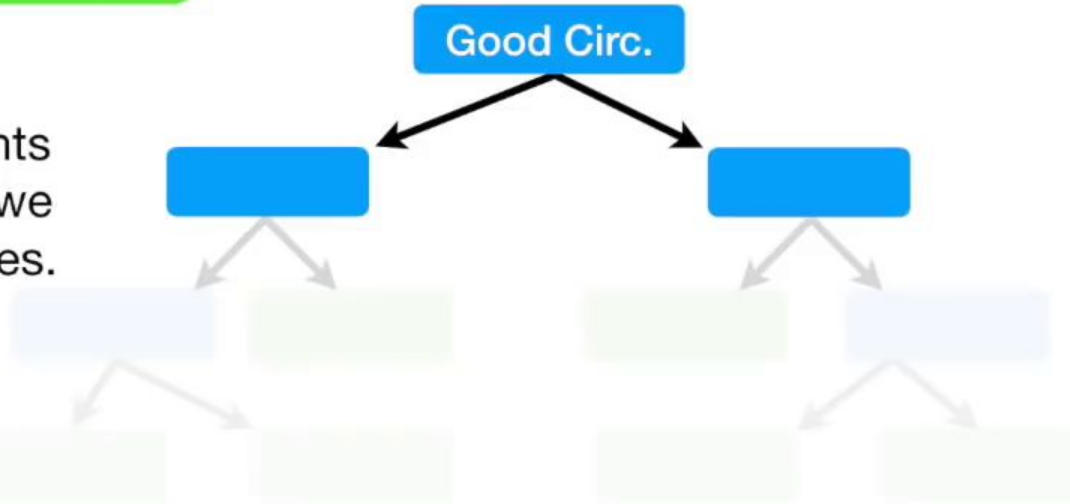


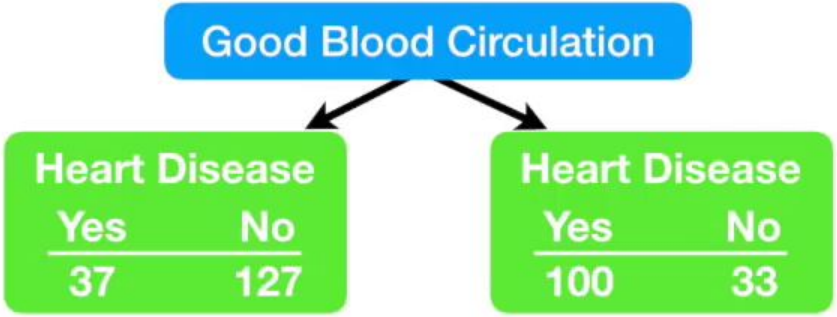

Gini impurity for Good Blood Circulation = 0.360

Gini impurity for Blocked Arteries = 0.381



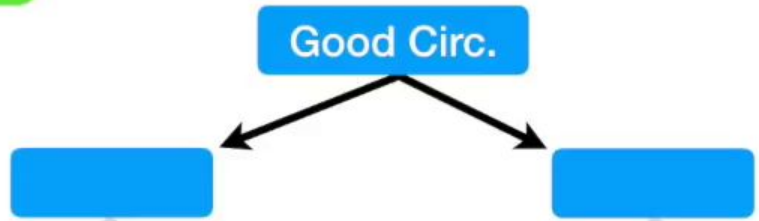

When we divided all of the patients using **Good Blood Circulation**, we ended up with “impure” leaf nodes.

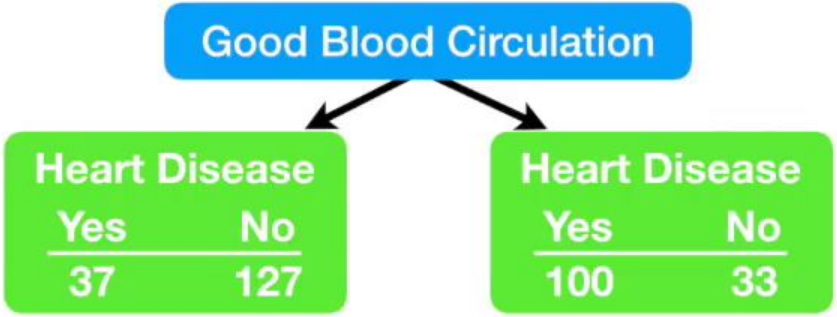




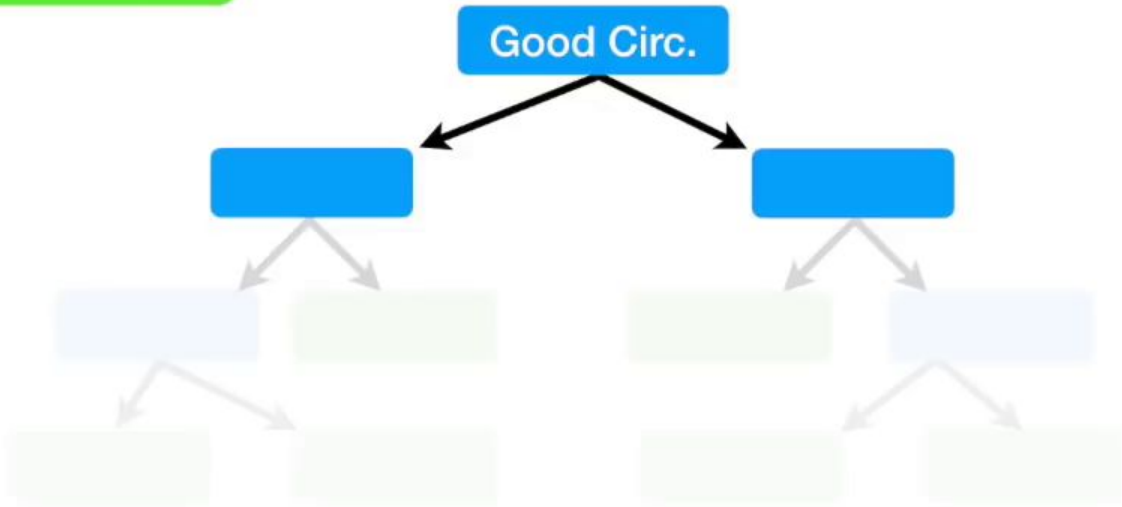

When we divided all of the patients using **Good Blood Circulation**, we ended up with “impure” leaf nodes.

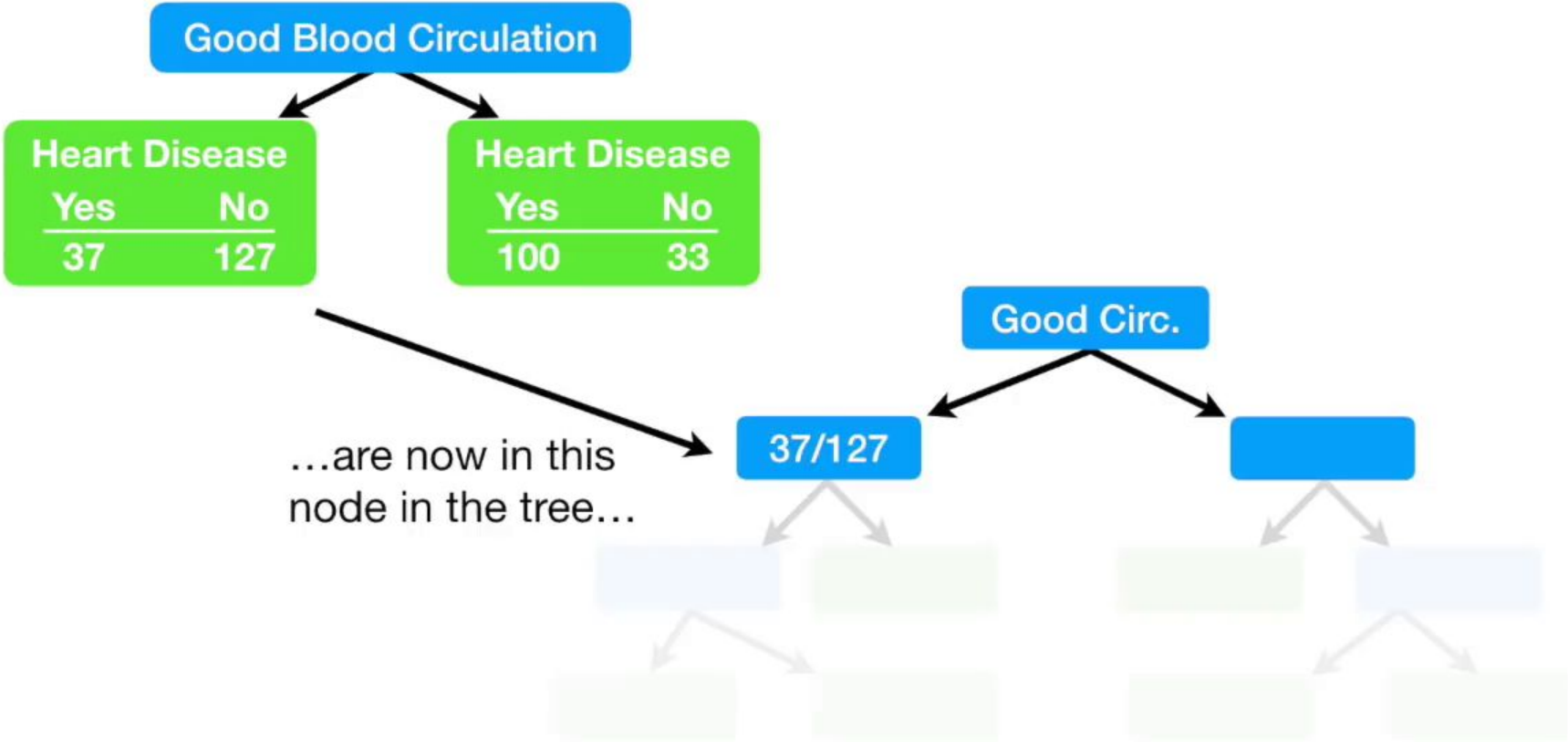
Each leaf contained a mixture of patients with and without Heart Disease.

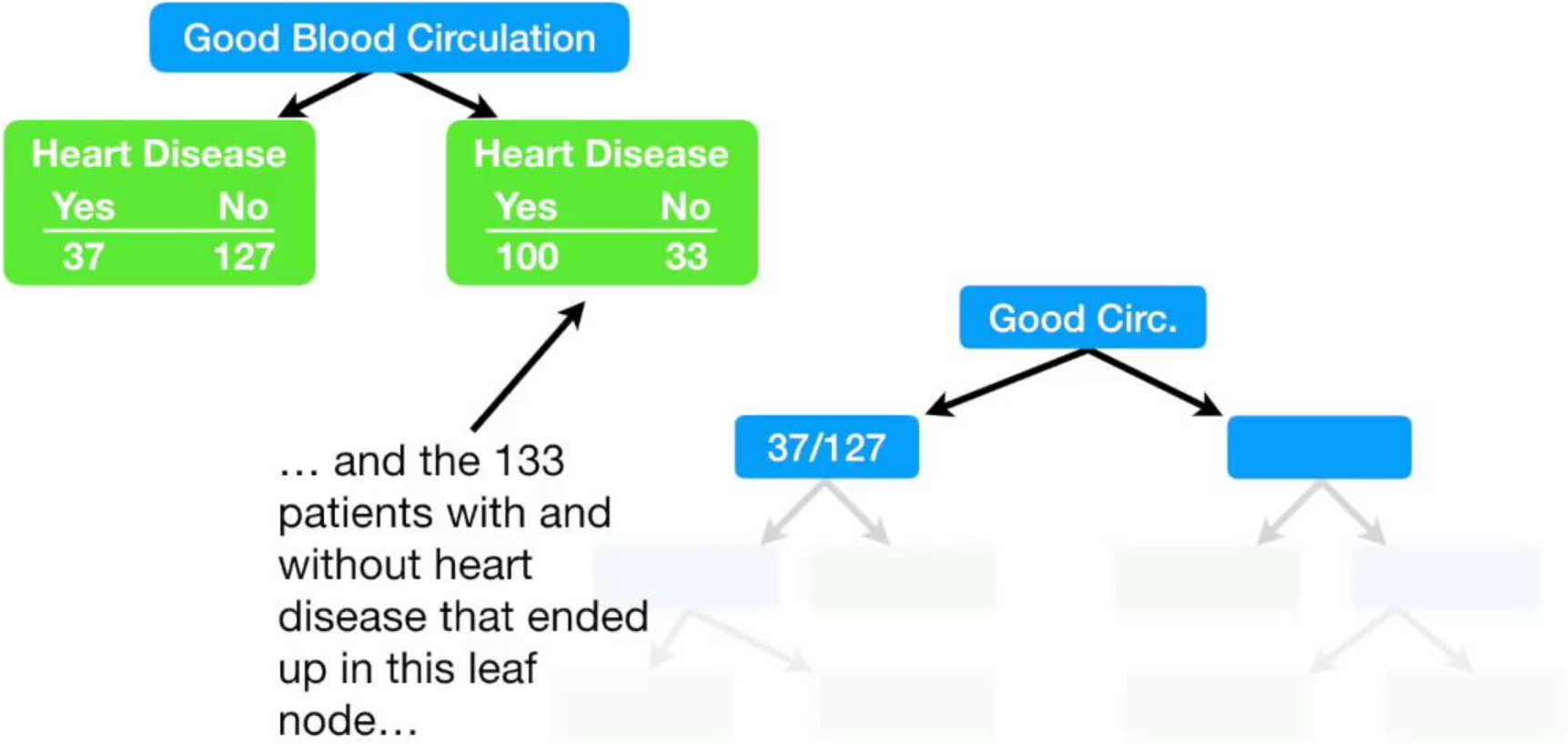


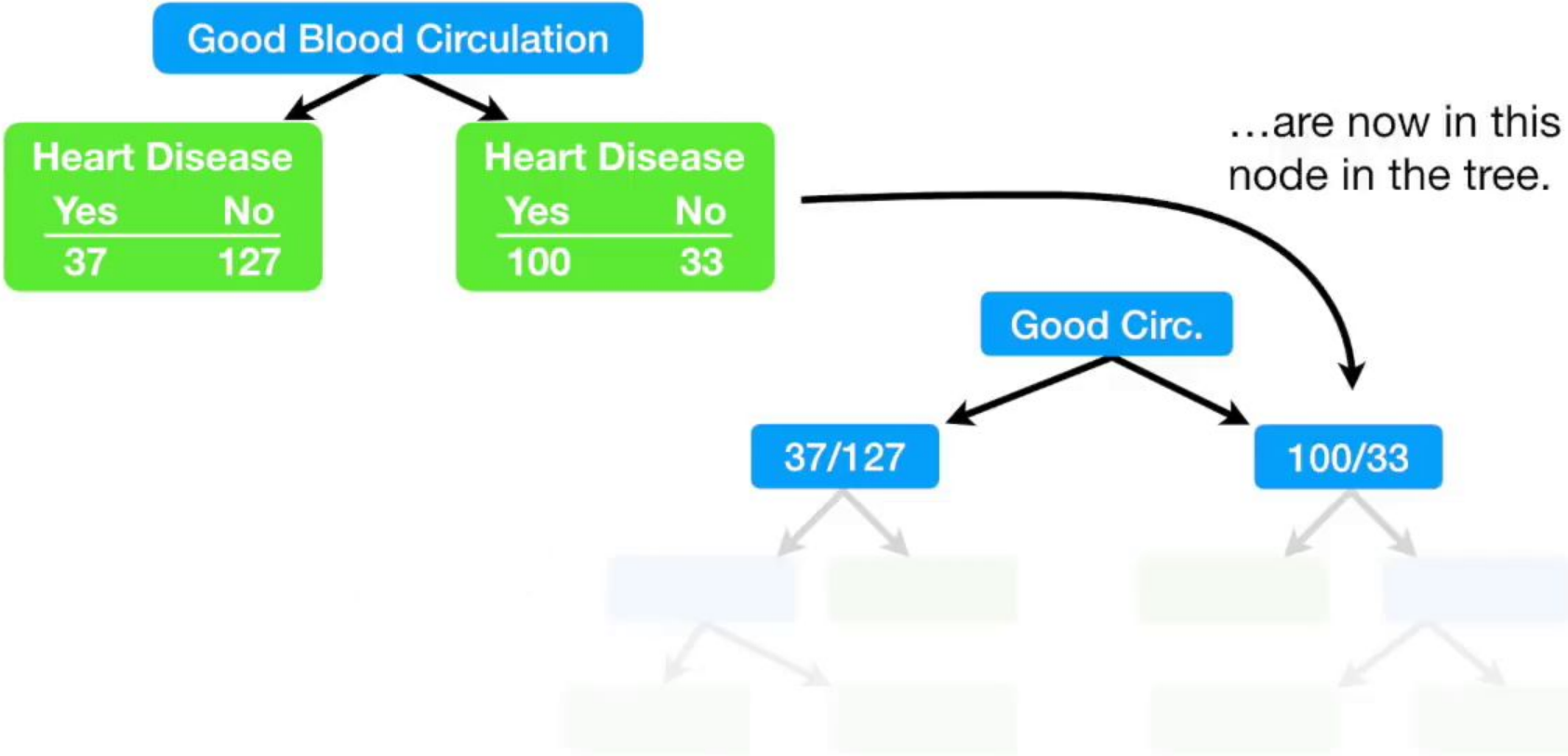



That means the 164 patients with and without heart disease that ended up in this leaf node…

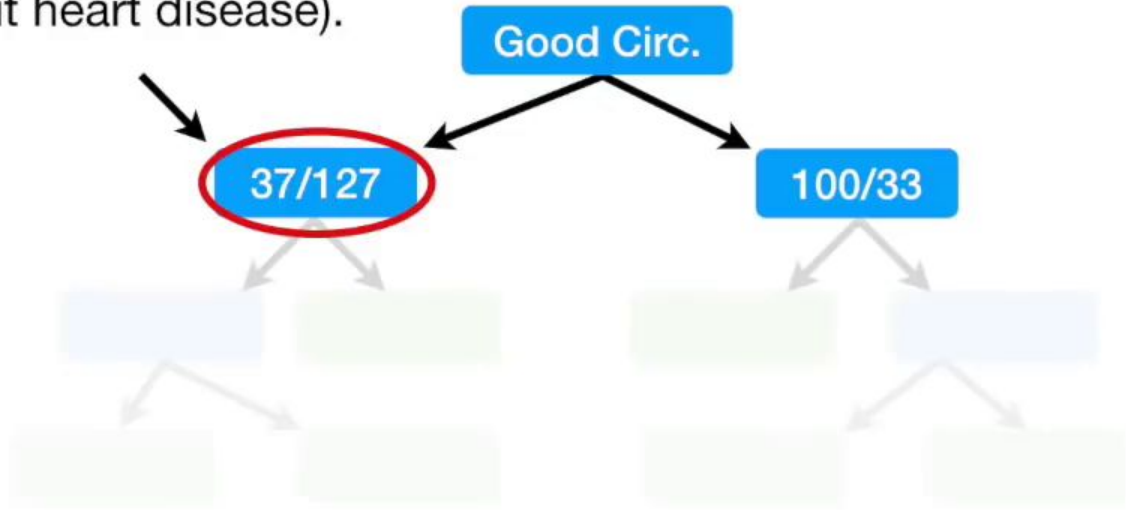


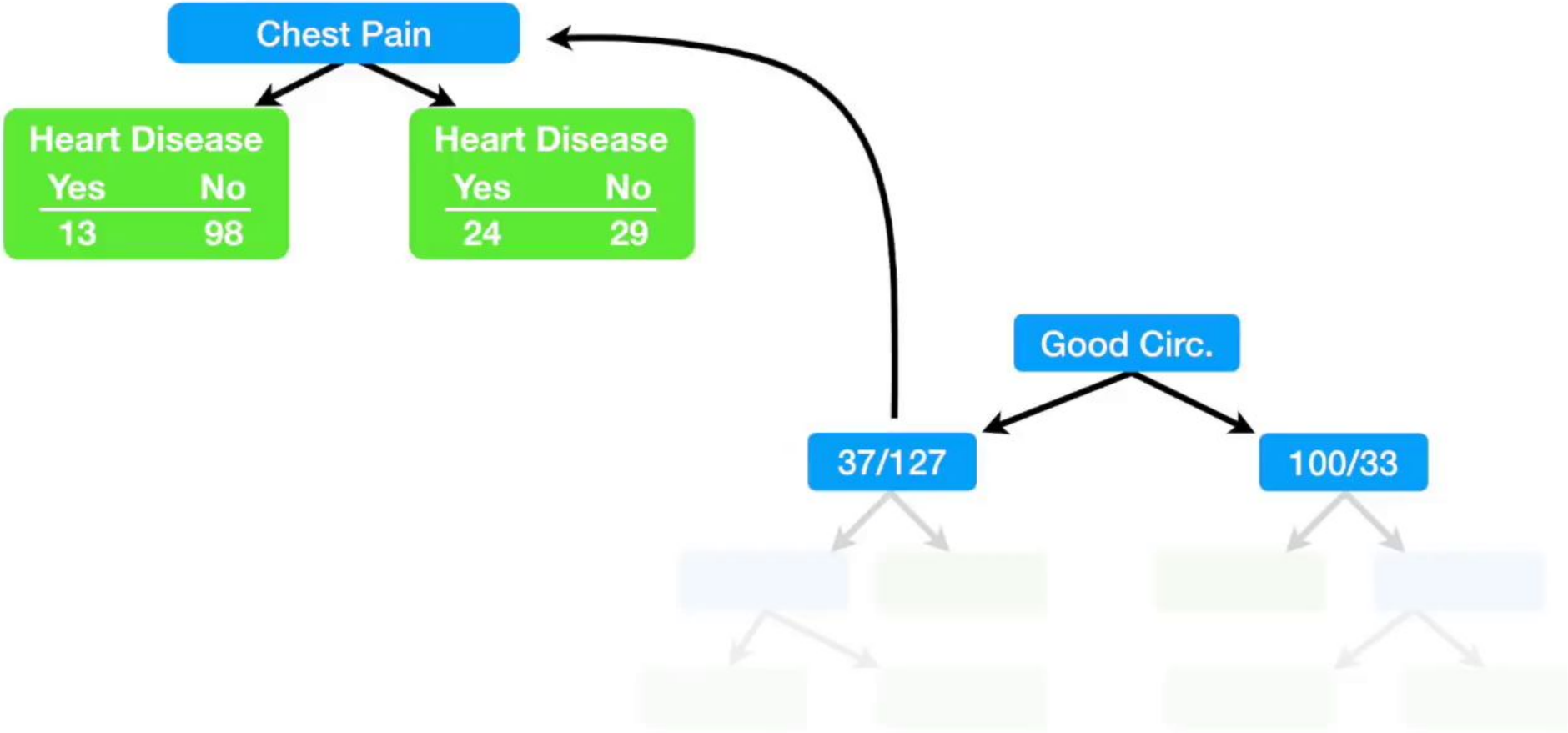

Good Blood Circulation
Heart Disease
Yes
No
37
127
Heart Disease
Yes
No
100
33
…are now in this
node in the tree…
Good Circ.
37/127

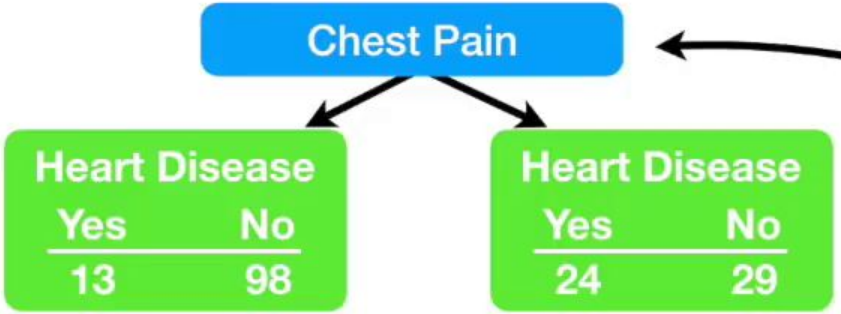

Good Blood Circulation
Heart Disease
Yes
No
37
127
Heart Disease
Yes
No
100
33
… and the 133 patients with and without heart disease that ended up in this leaf node…
Good Circ.
37/127


Good Blood Circulation
Heart Disease
Yes
No
37
127
Heart Disease
Yes
No
100
33
…are now in this node in the tree.
Good Circ.
37/127
100/33

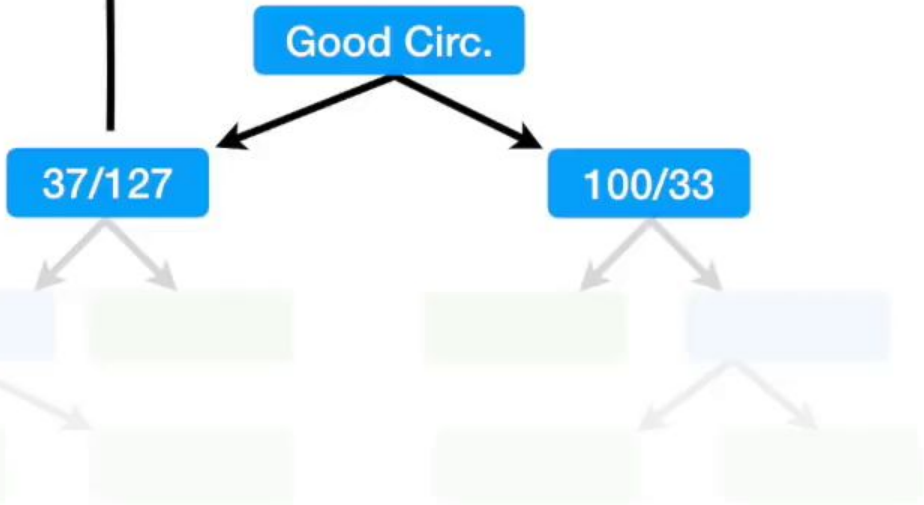
Now we need to figure how well **chest pain** and **blocked arteries** separate these 164 patients (37 with heart disease and 127 without heart disease).

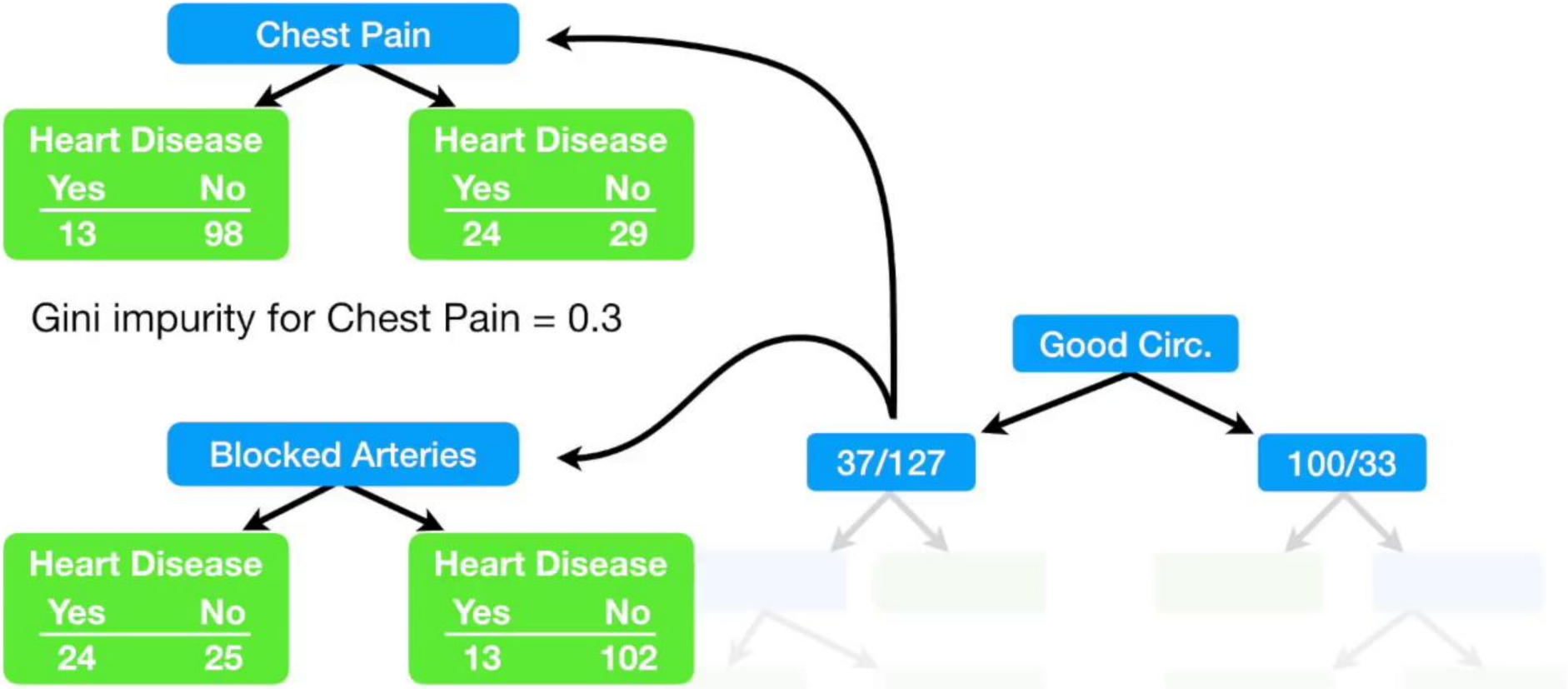


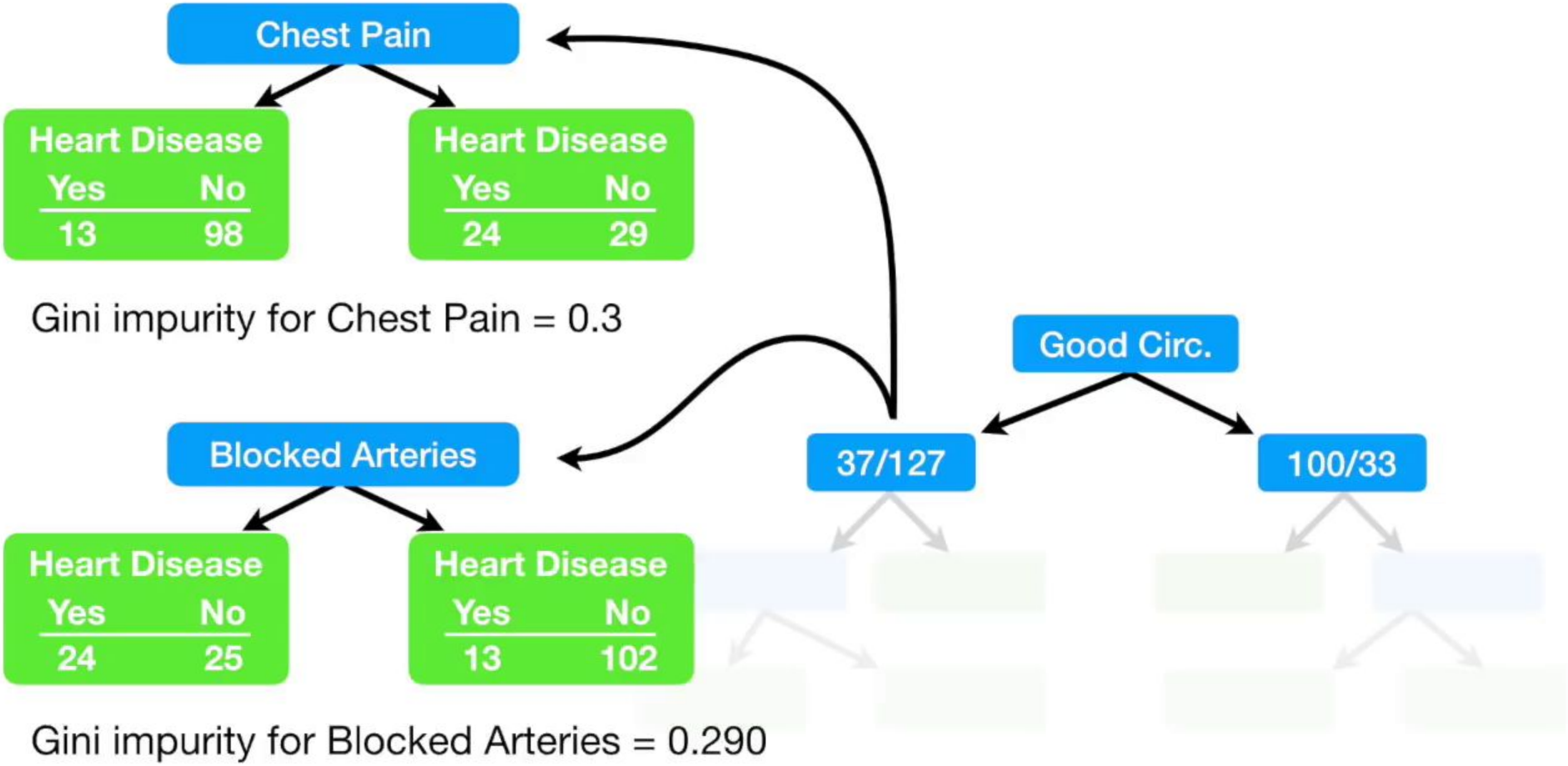

Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Good Circ.
37/127
100/33

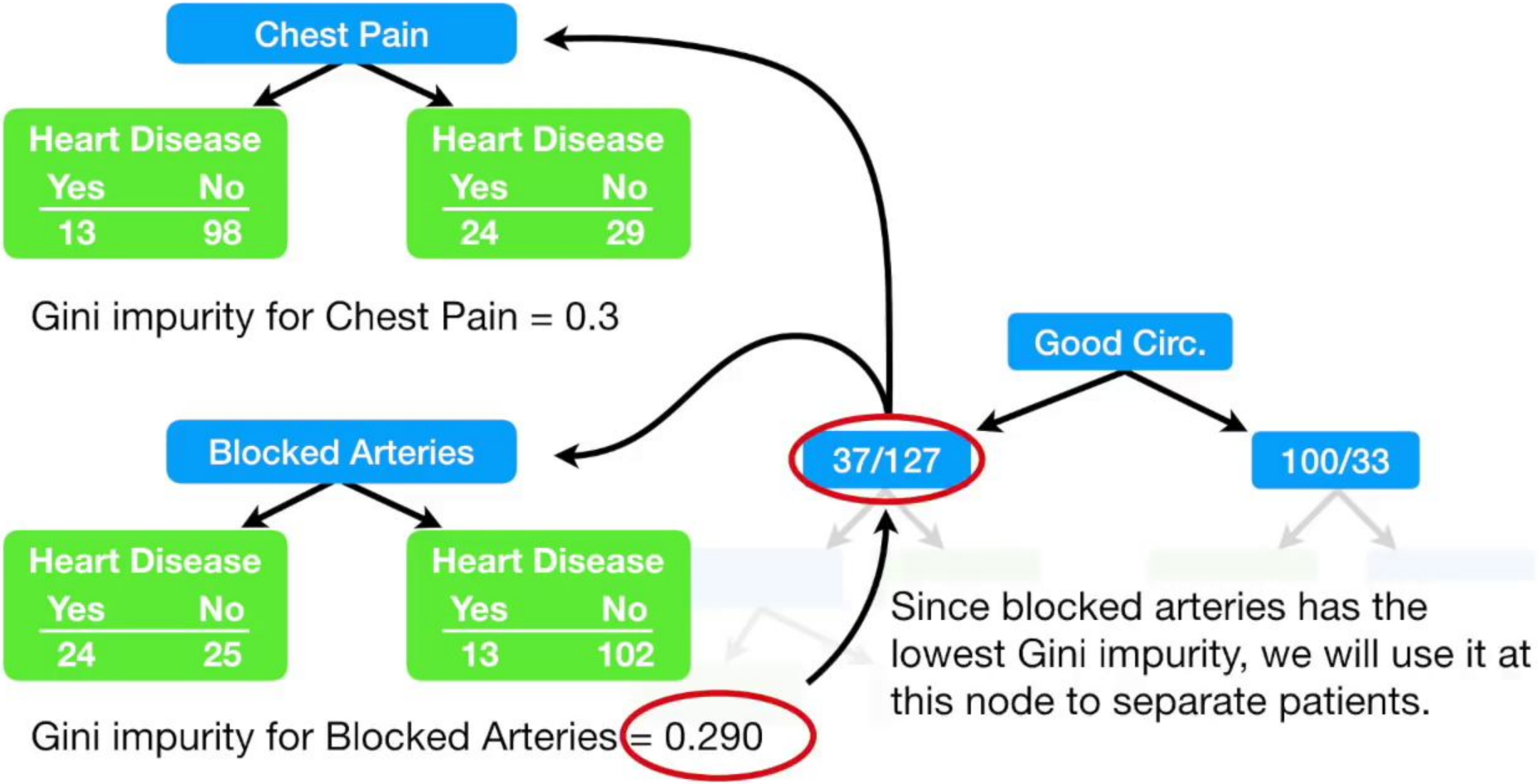



Gini impurity for Chest Pain = 0.3

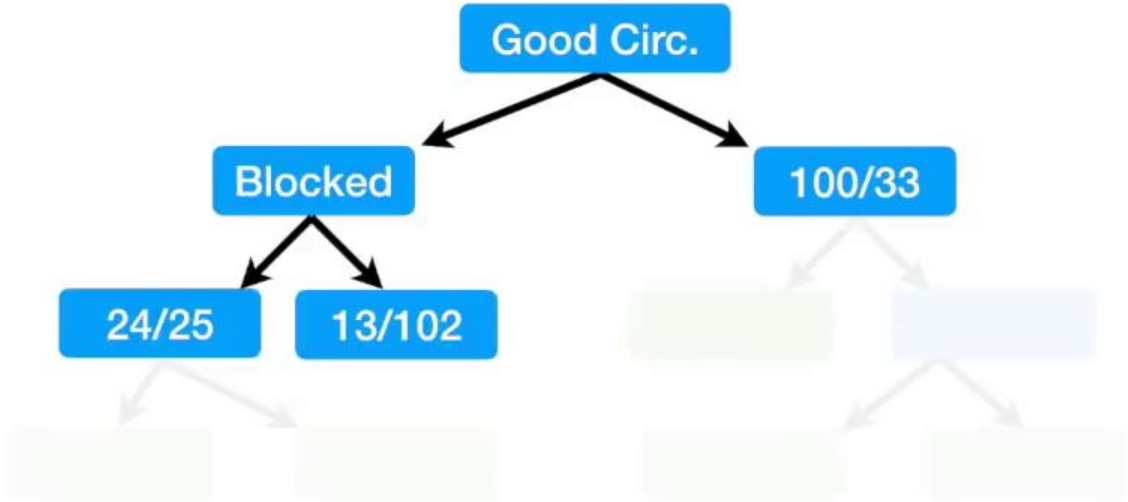



Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Gini impurity for Chest Pain = 0.3
Blocked Arteries
Heart Disease
Yes No
24 25
Heart Disease
Yes No
13 102
Good Circ.
37/127
100/33

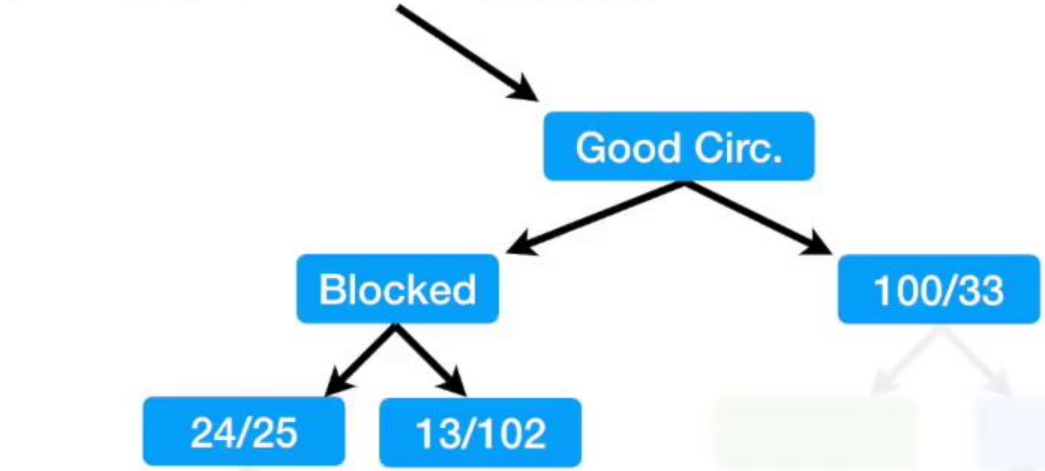

Chest Pain
Heart Disease
Yes No
13 98
Heart Disease
Yes No
24 29
Gini impurity for Chest Pain = 0.3
Blocked Arteries
Heart Disease
Yes No
24 25
Heart Disease
Yes No
13 102
Gini impurity for Blocked Arteries = 0.290
Good Circ.
37/127
100/33


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.

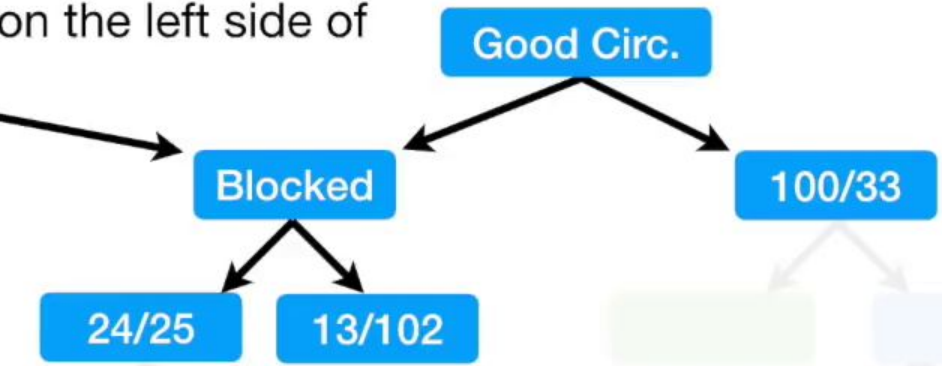
Here's the tree that we've worked out so far.



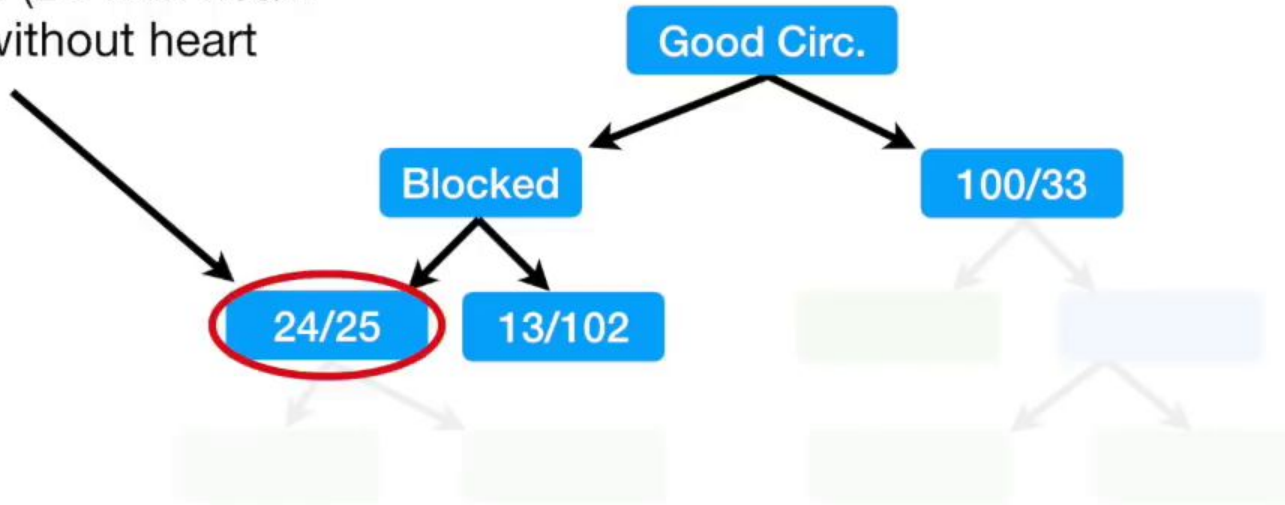
We started at the top by separating patients with Good Circulation…

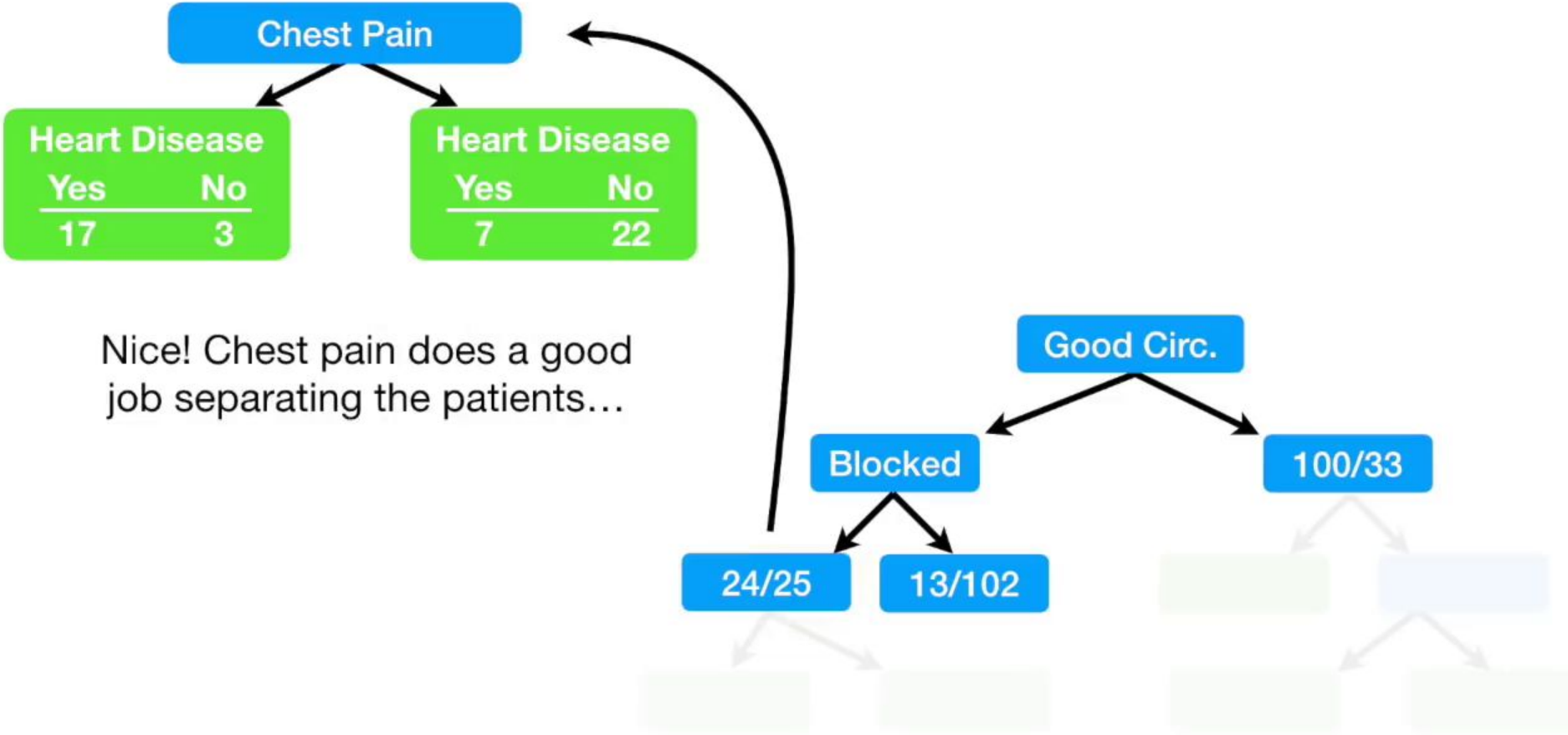


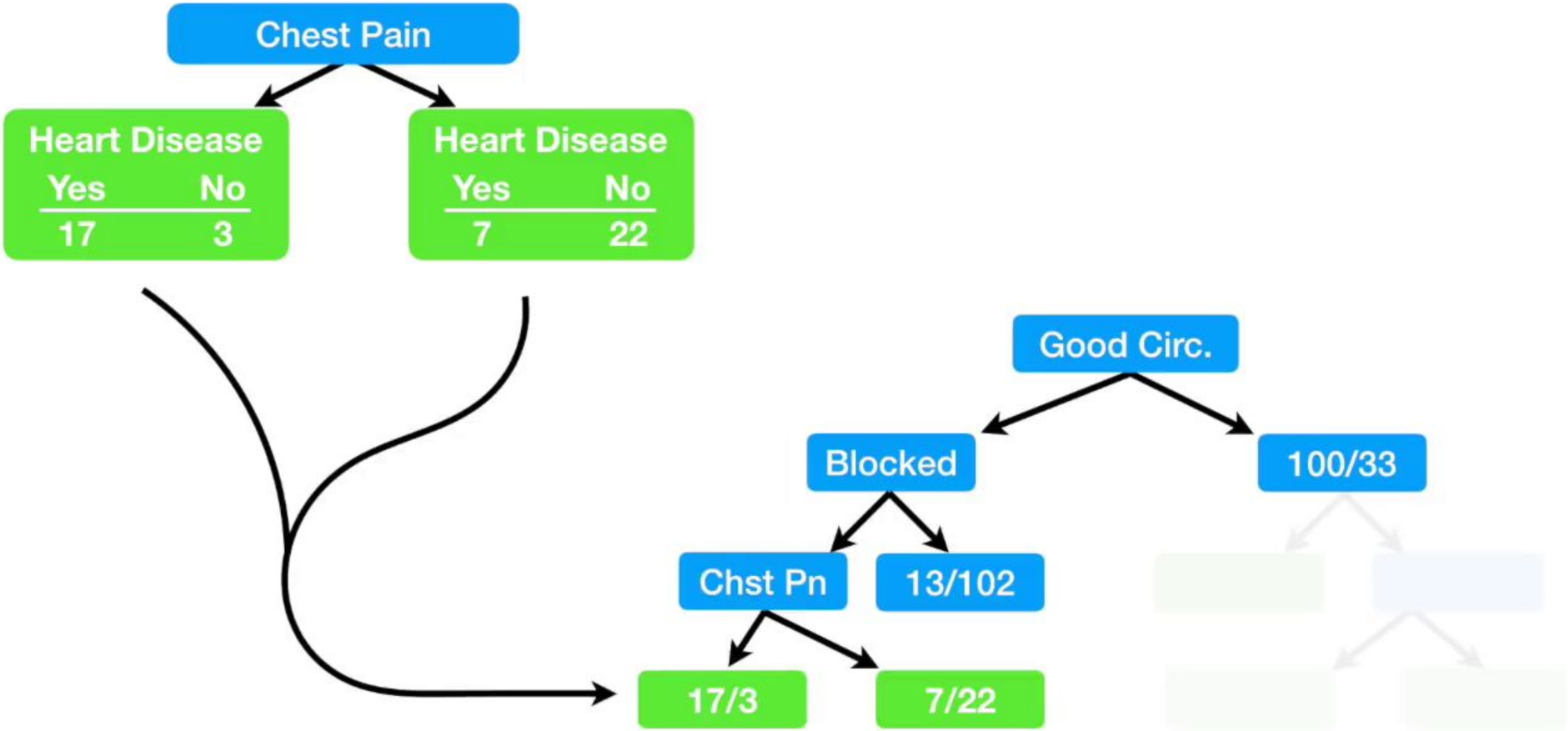
…then we used Blocked Arteries to separate patients on the left side of the tree.



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).

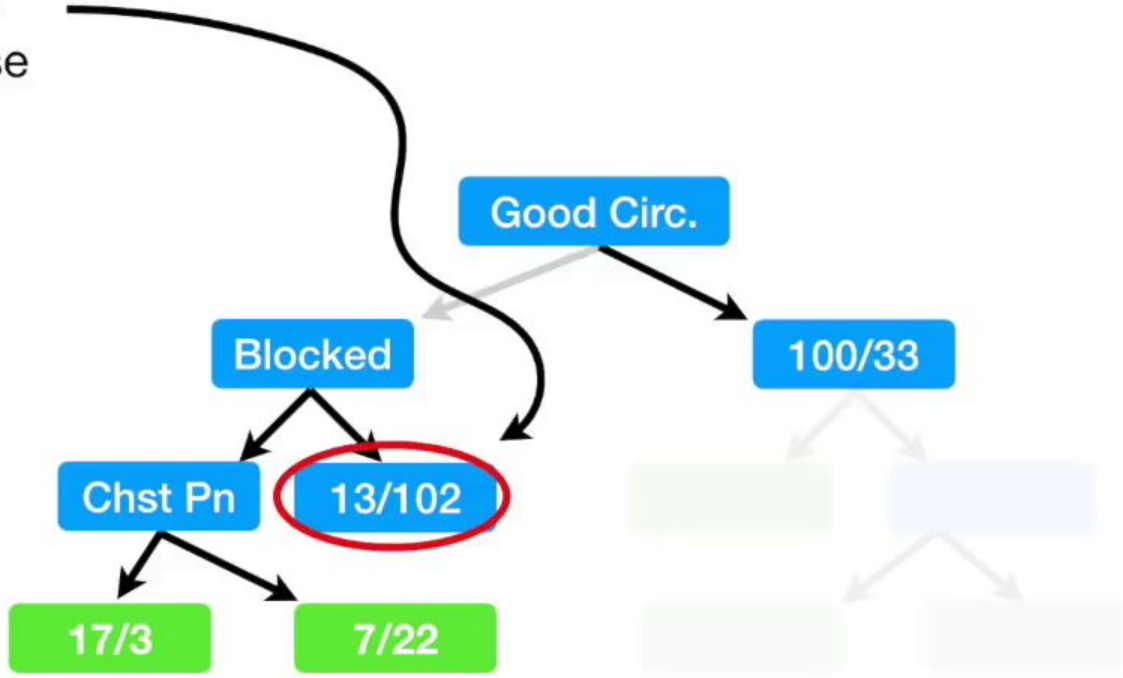



Chest Pain
Heart Disease
Yes No
17 3
Heart Disease
Yes No
7 22
Nice! Chest pain does a good job separating the patients…
Good Circ.
Blocked
100/33
24/25
13/102



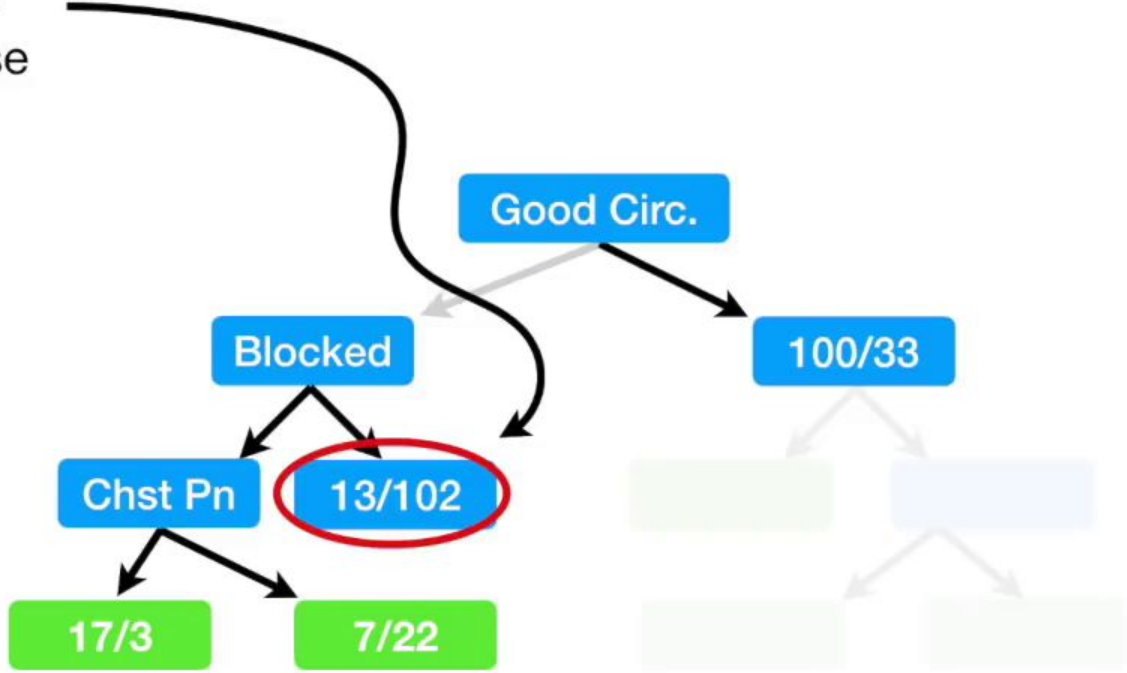

…so these are the final leaf nodes on this branch of the tree.

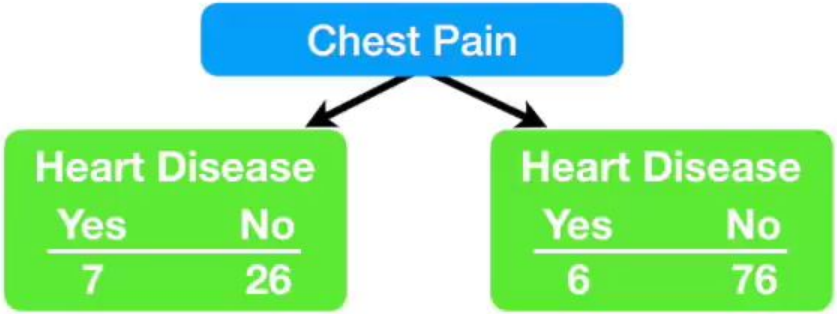
Now let's see what happens when we use chest pain to divide these 115 patients (13 with heart disease and 102 without).

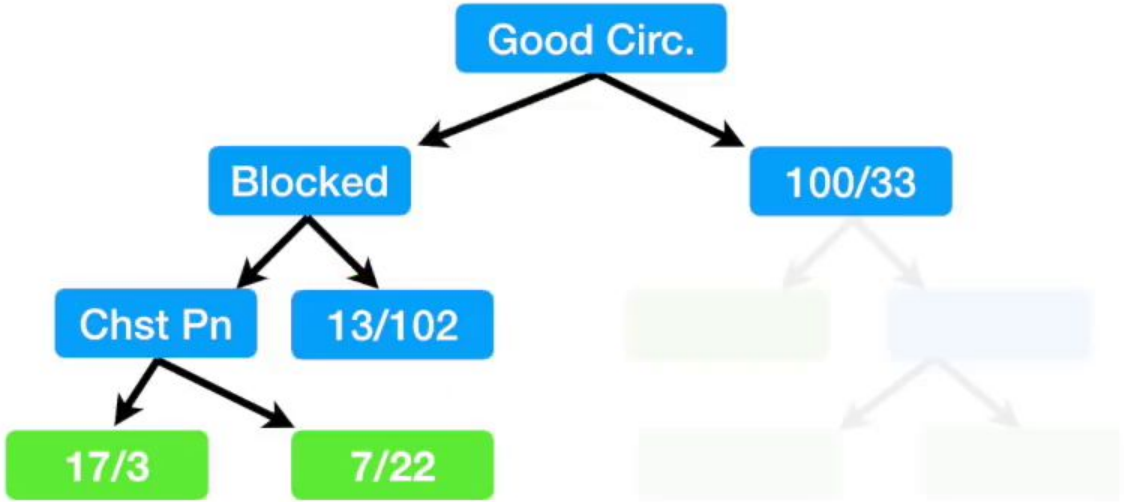


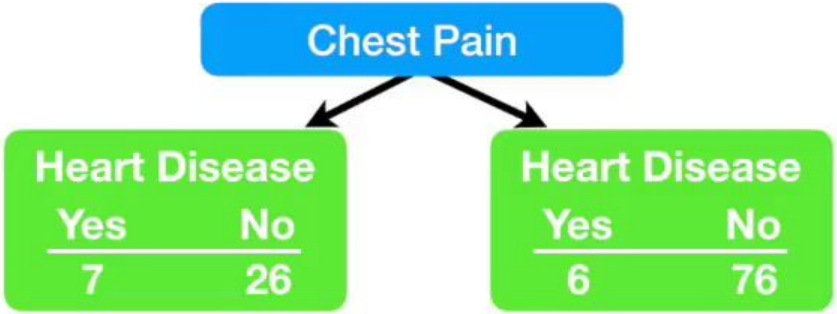
Now let's see what happens when we use chest pain to divide these 115 patients (13 with heart disease and 102 without).

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

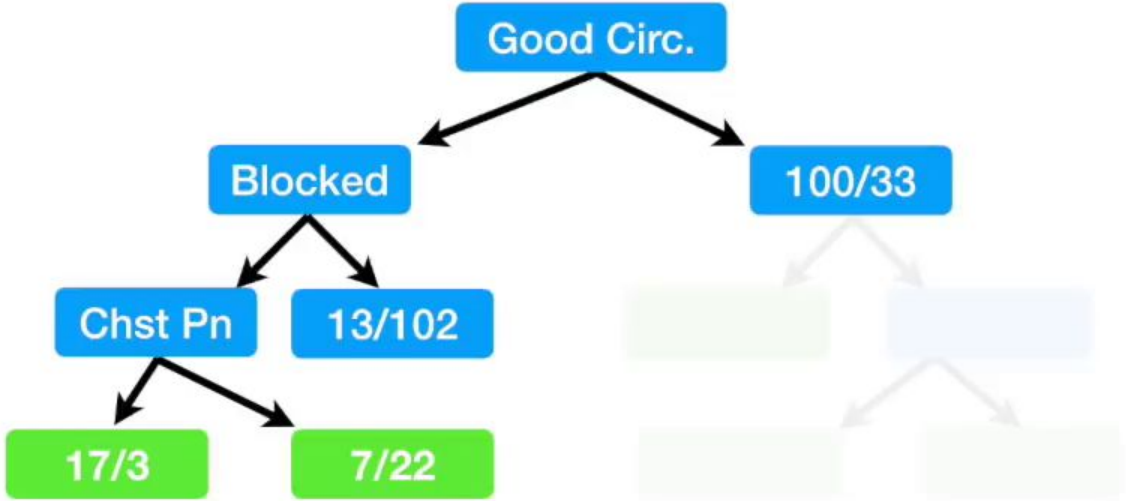


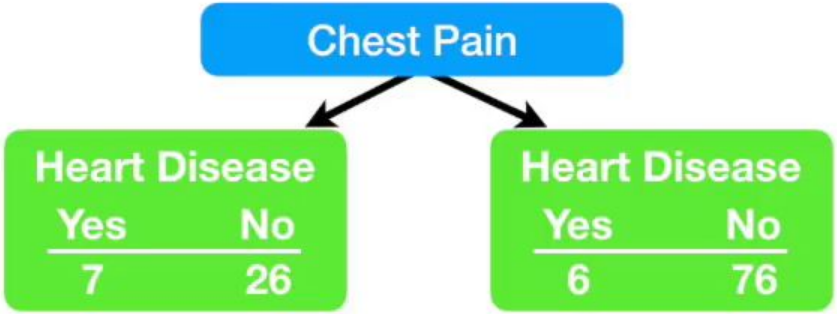

Chest Pain
Heart Disease
Yes No
7 26
Heart Disease
Yes No
6 76


Good Circ.
Blocked
100/33
Chst Pn
13/102
17/3
7/22

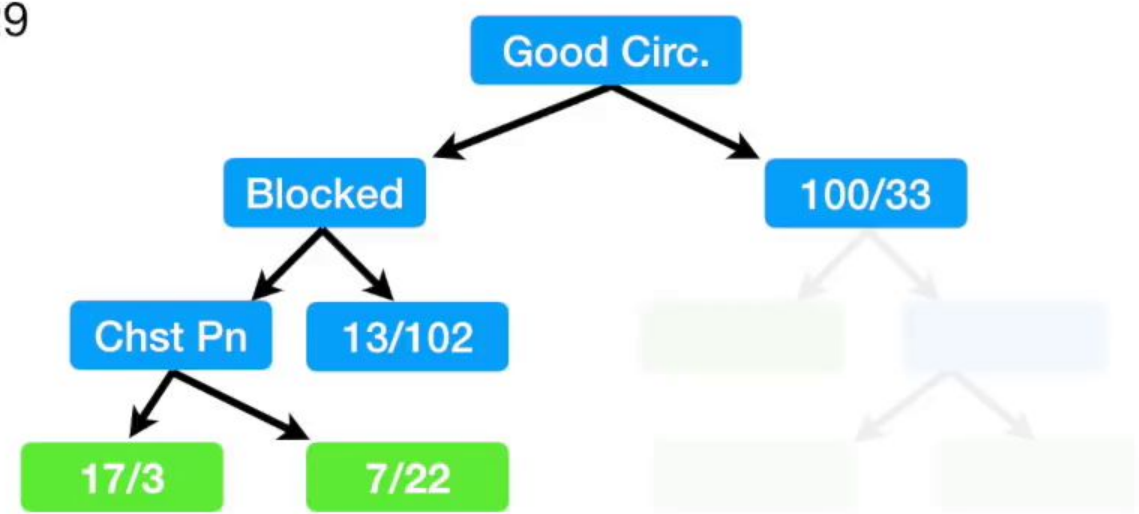


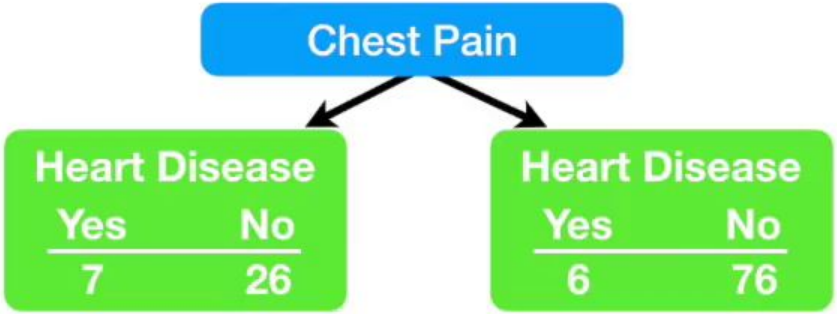

Do these new leaves separate patients better than what we had before?





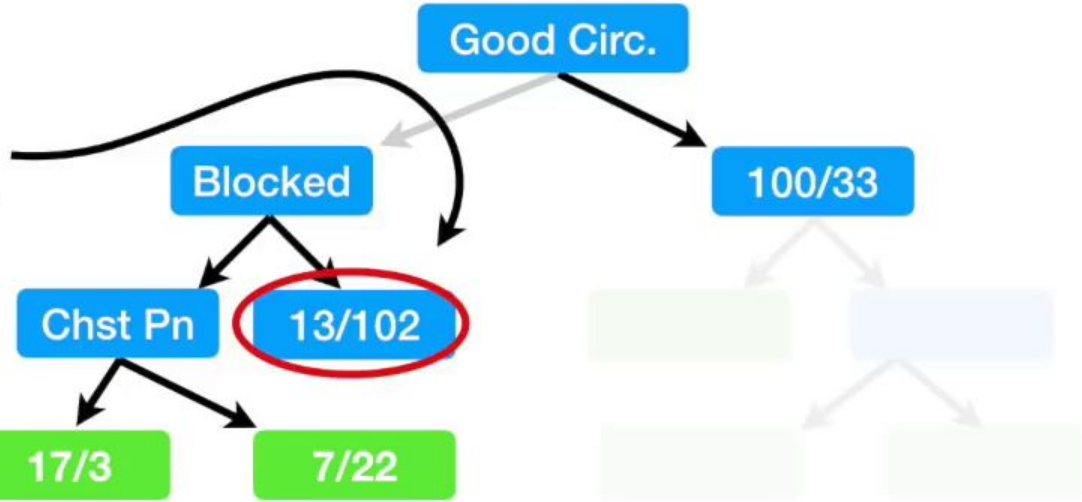

Gini impurity for Chest Pain = 0.29

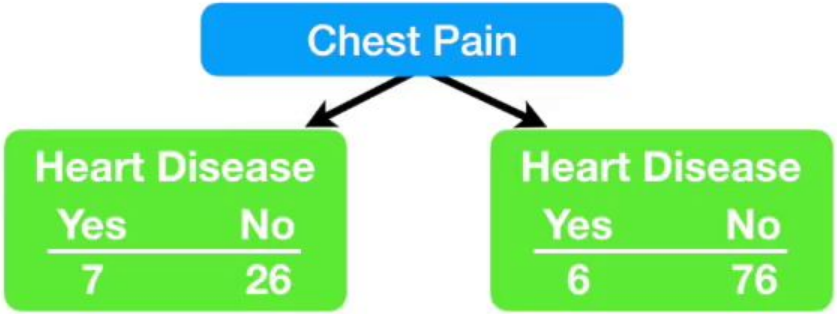





Gini impurity for Chest Pain = 0.29

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



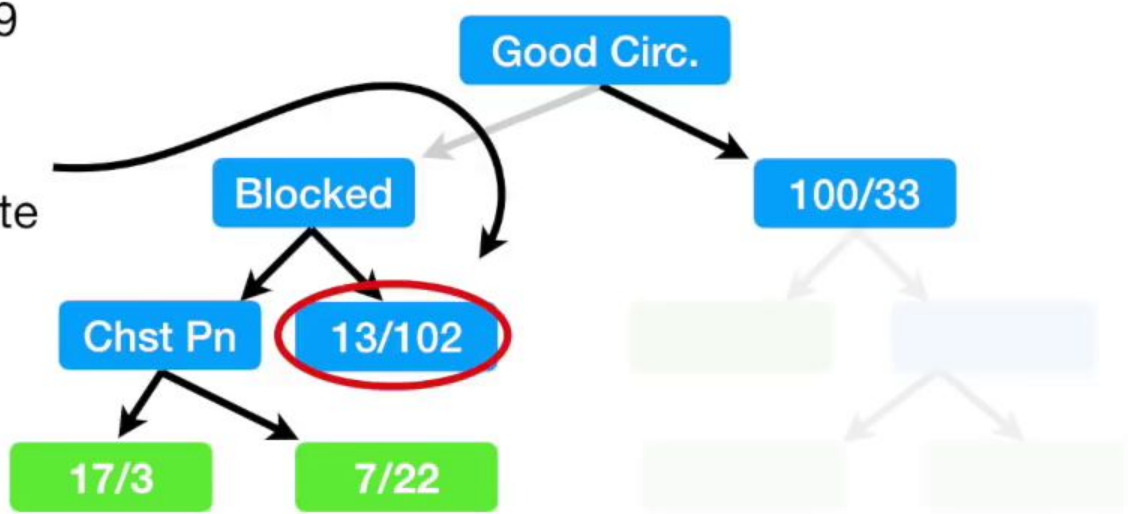


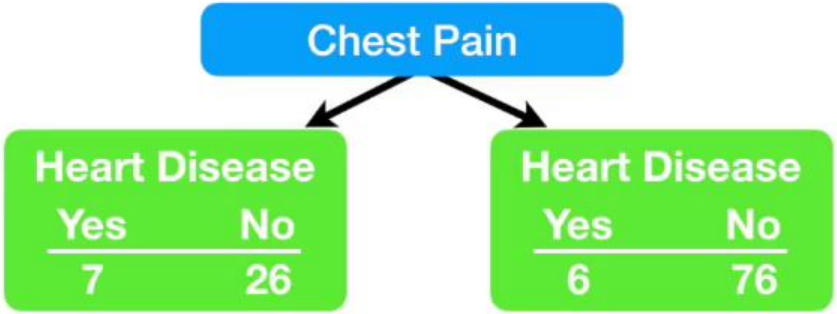

Gini impurity for Chest Pain = 0.29

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

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

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





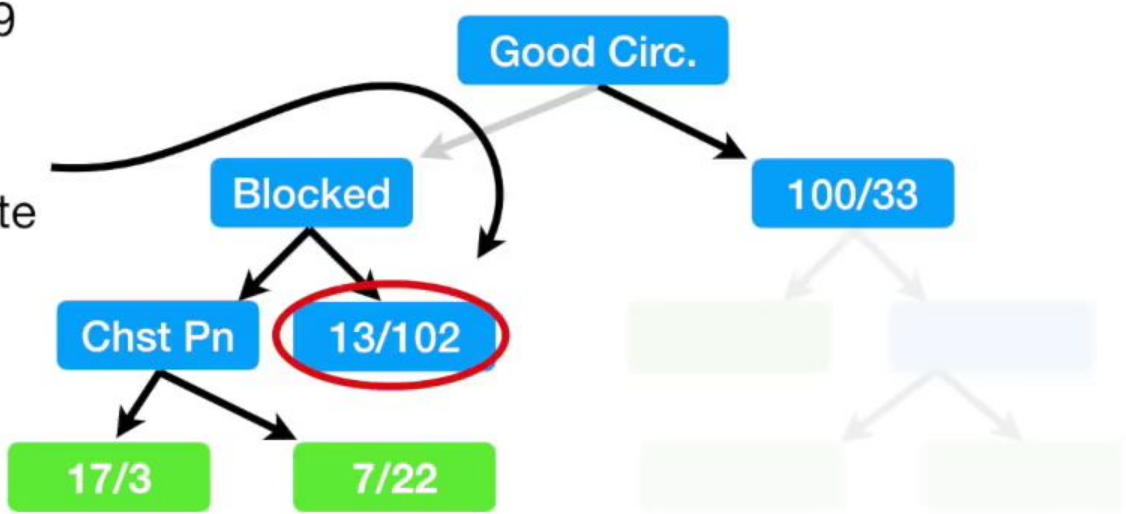

Gini impurity for Chest Pain = 0.29

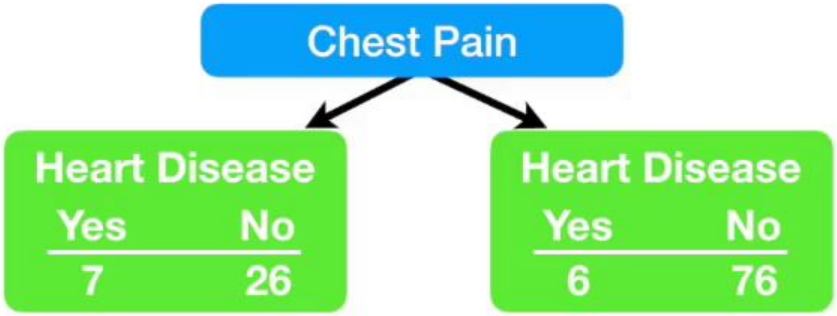
The Gini impurity for this node, before using chest pain to separate patients is…

= 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





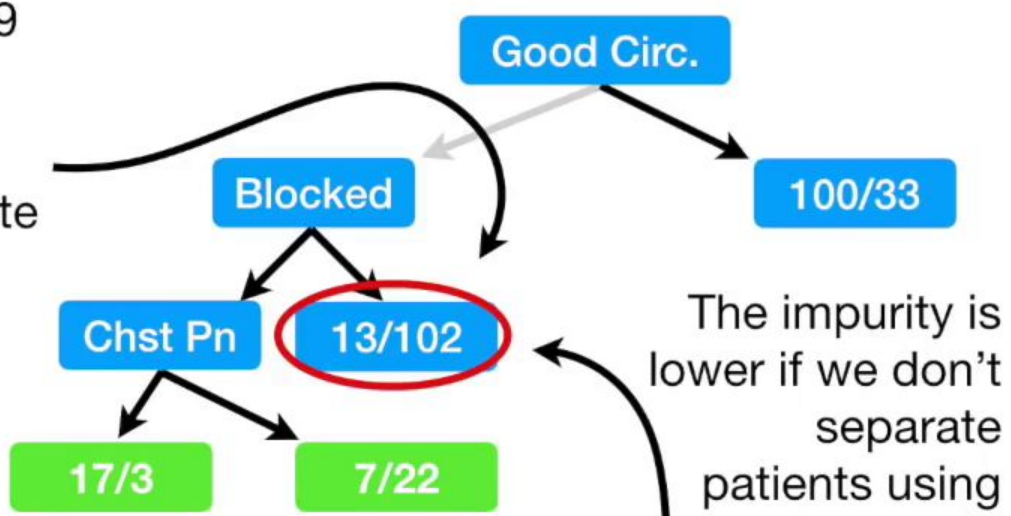

Gini impurity for Chest Pain = 0.29

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

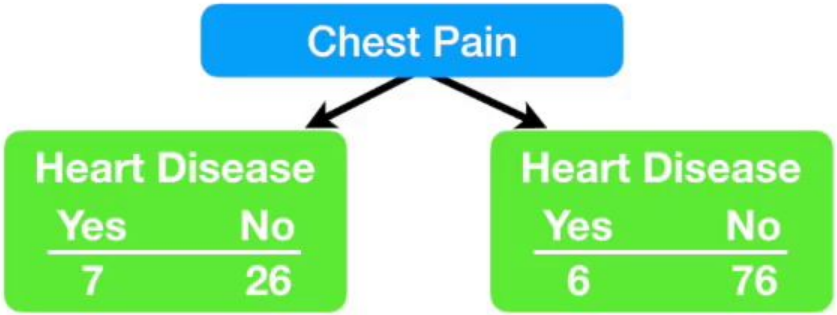
= 1 - (the probability of "yes")$^2$
- (the probability of "no")$^2$

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

= 0.2




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



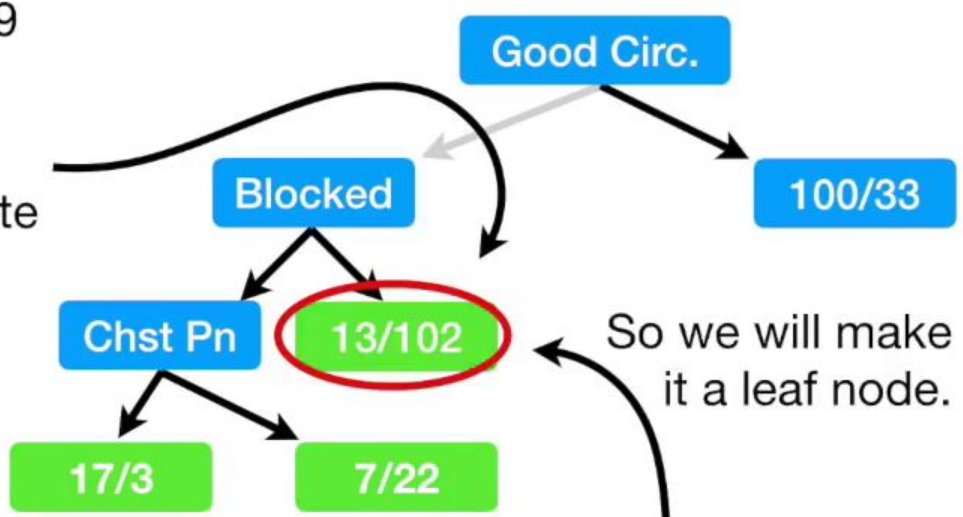

Gini impurity for Chest Pain = 0.29

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

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

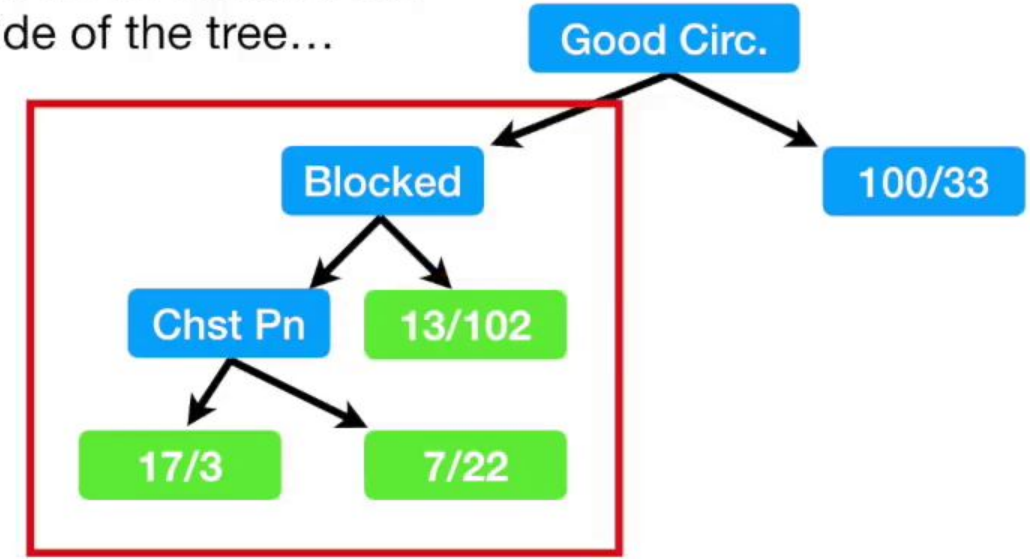
$$= 1 - \left(\frac{13}{13 + 102}\right)^2 - \left(\frac{102}{13 + 102}\right)^2$$

= 0.2

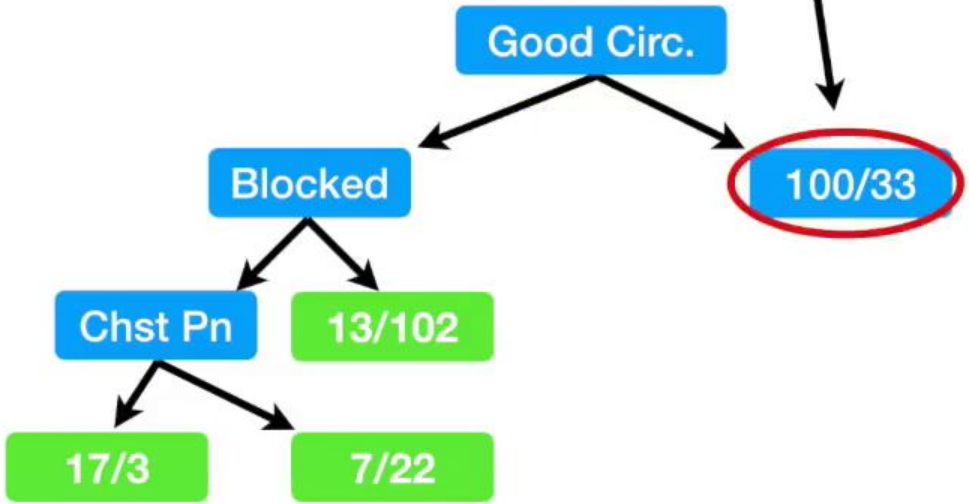



So we will make it a leaf node.

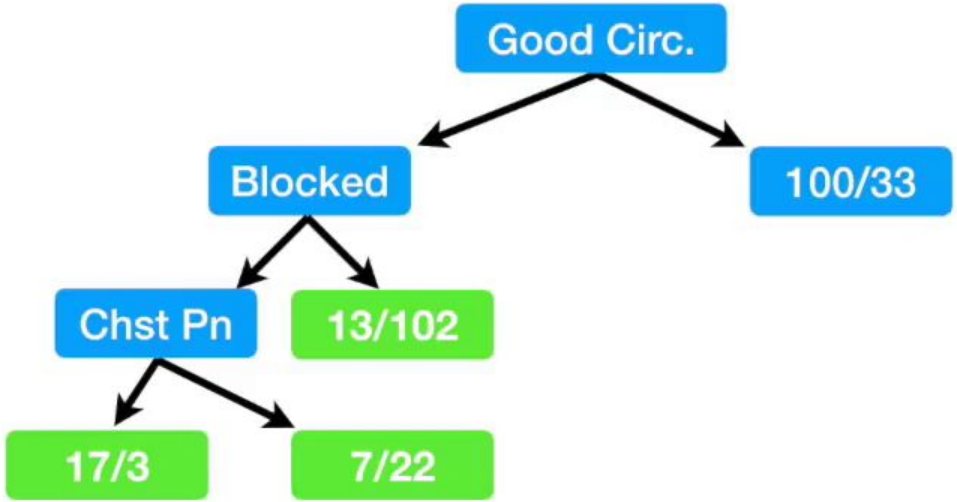
OK, at this point we've worked out the entire left side of the tree…



…now we need to work out the right side of the tree…

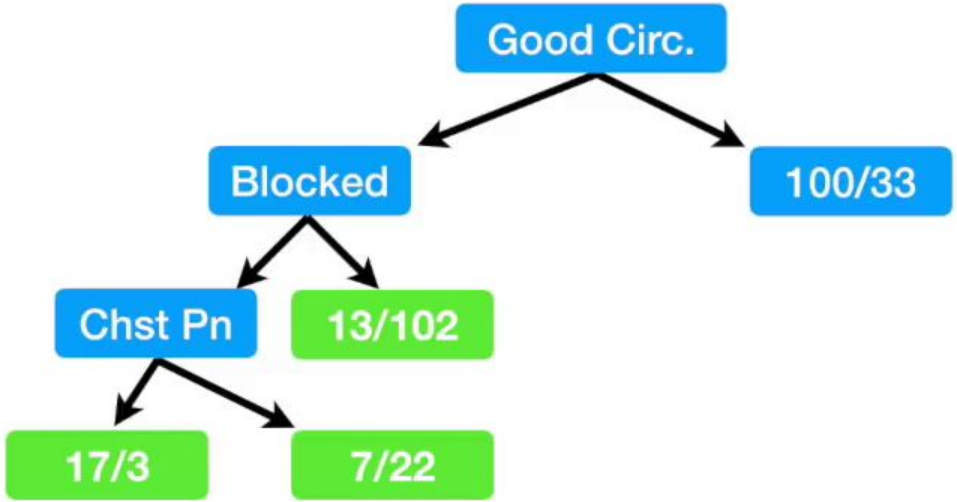


The good news is that we follow the exact same steps as we did on the left side:



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

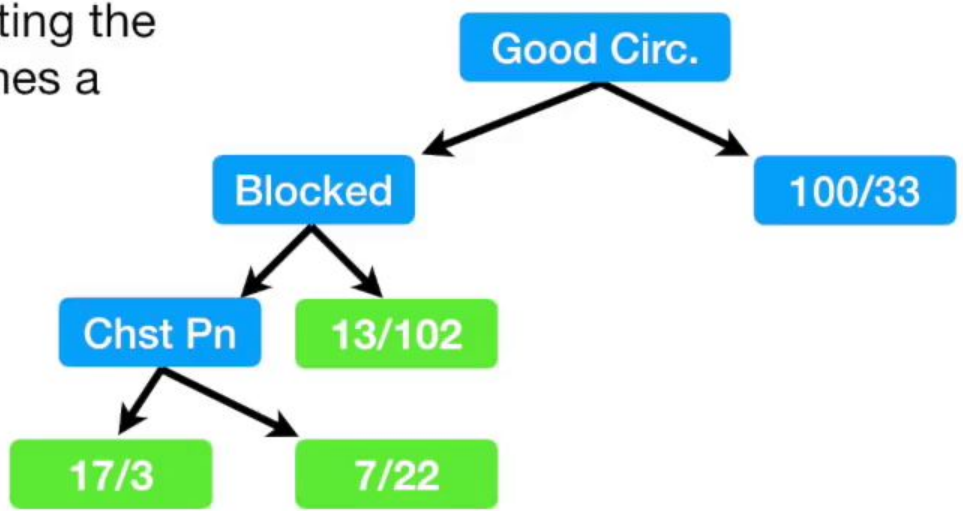
1) Calculate all of the Gini impurity scores.



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.

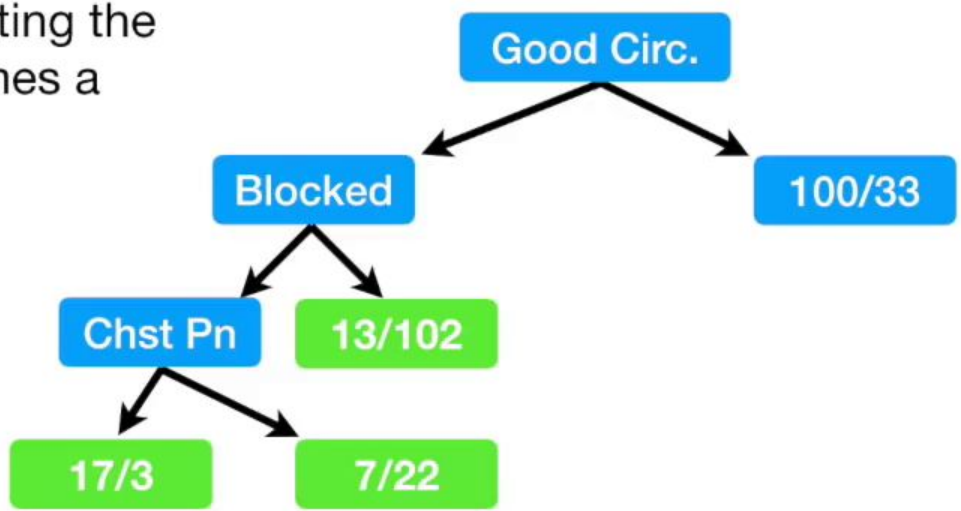


The good news is that we follow the exact same steps as we did on the left side:

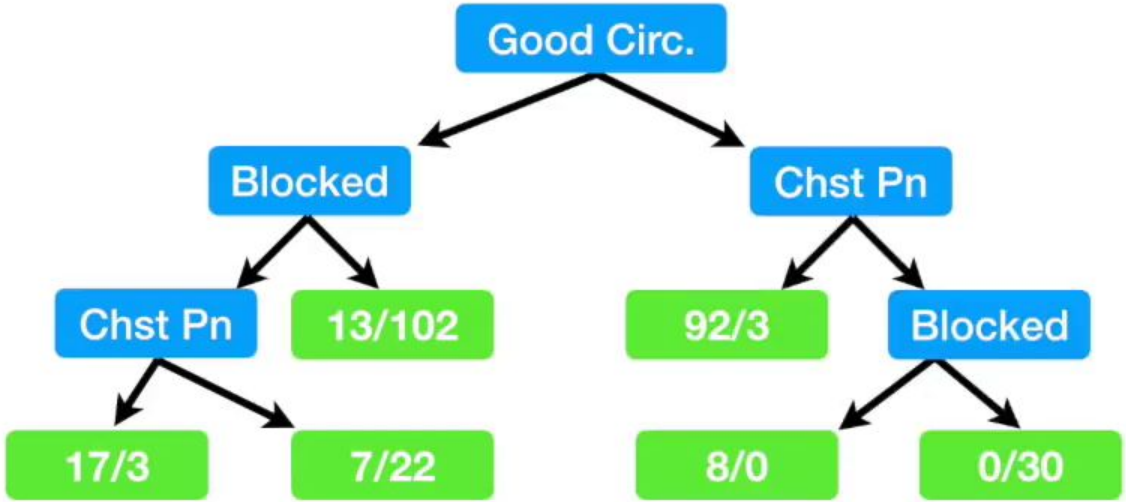
1) Calculate all of the Gini impurity scores.

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.

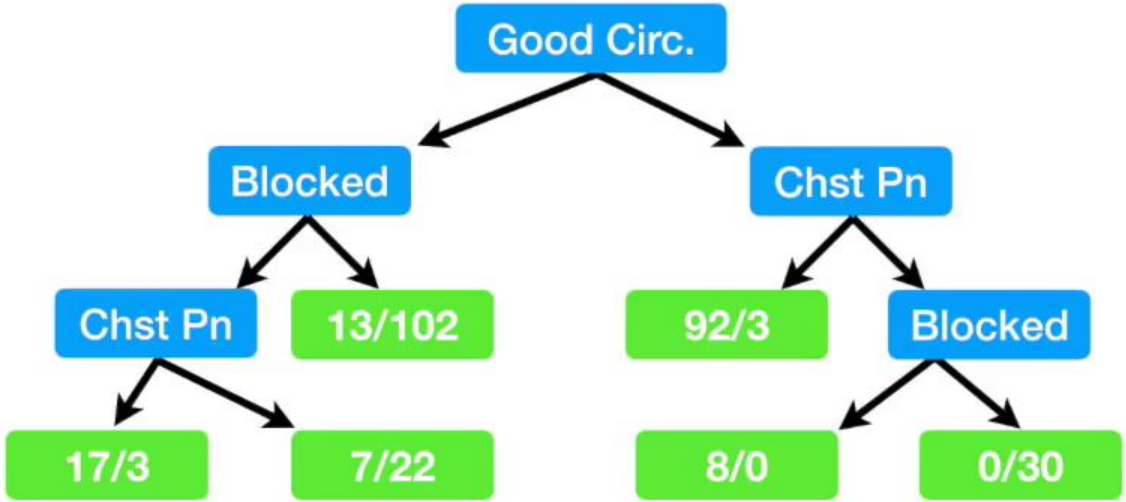
3) If separating the data results in an improvement, than pick the separation with the lowest impurity value.



Hooray!!! We made a decision tree!!!

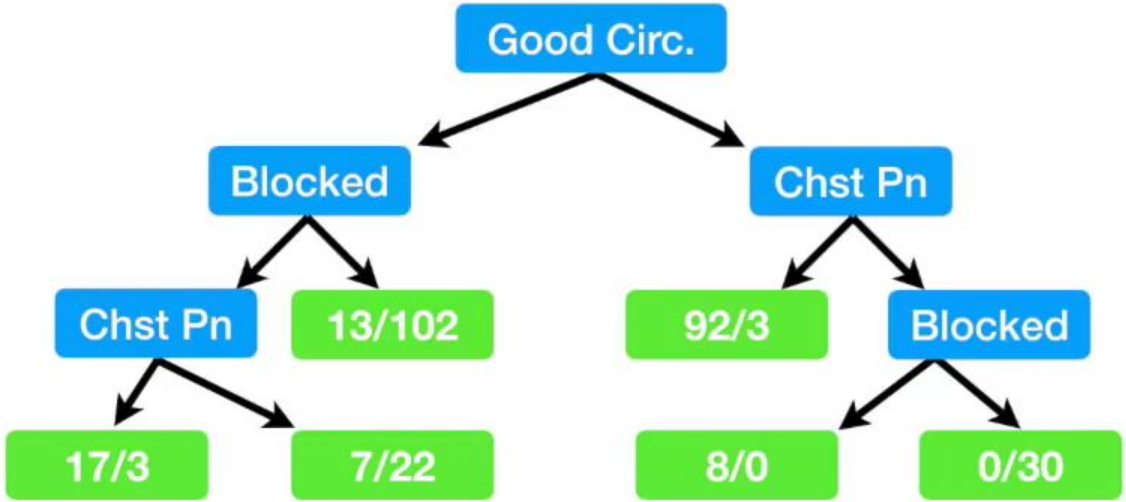


So far we've seen how to build a tree with "yes/no" questions at each step…



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?

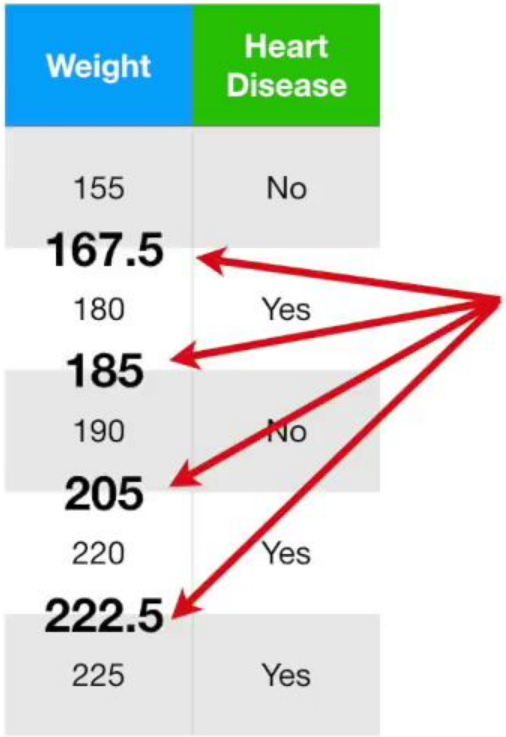


| Weight | Heart Disease |
| --- | --- |
| 220 | Yes |
| 180 | Yes |
| 225 | Yes |
| 190 | No |
| 155 | No |

| 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?


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




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 |

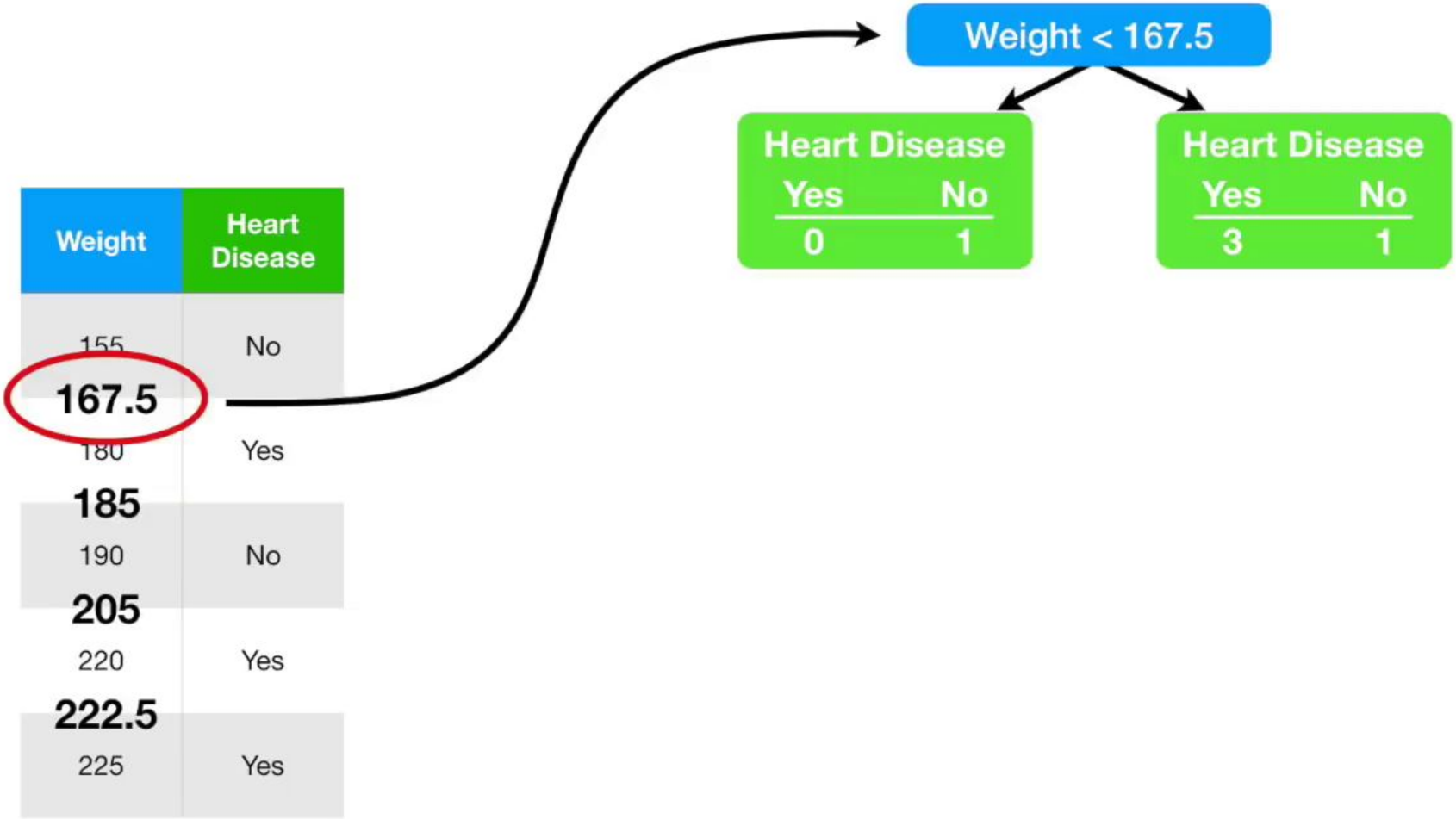
Step 3) Calculate the impurity values for each average weight.

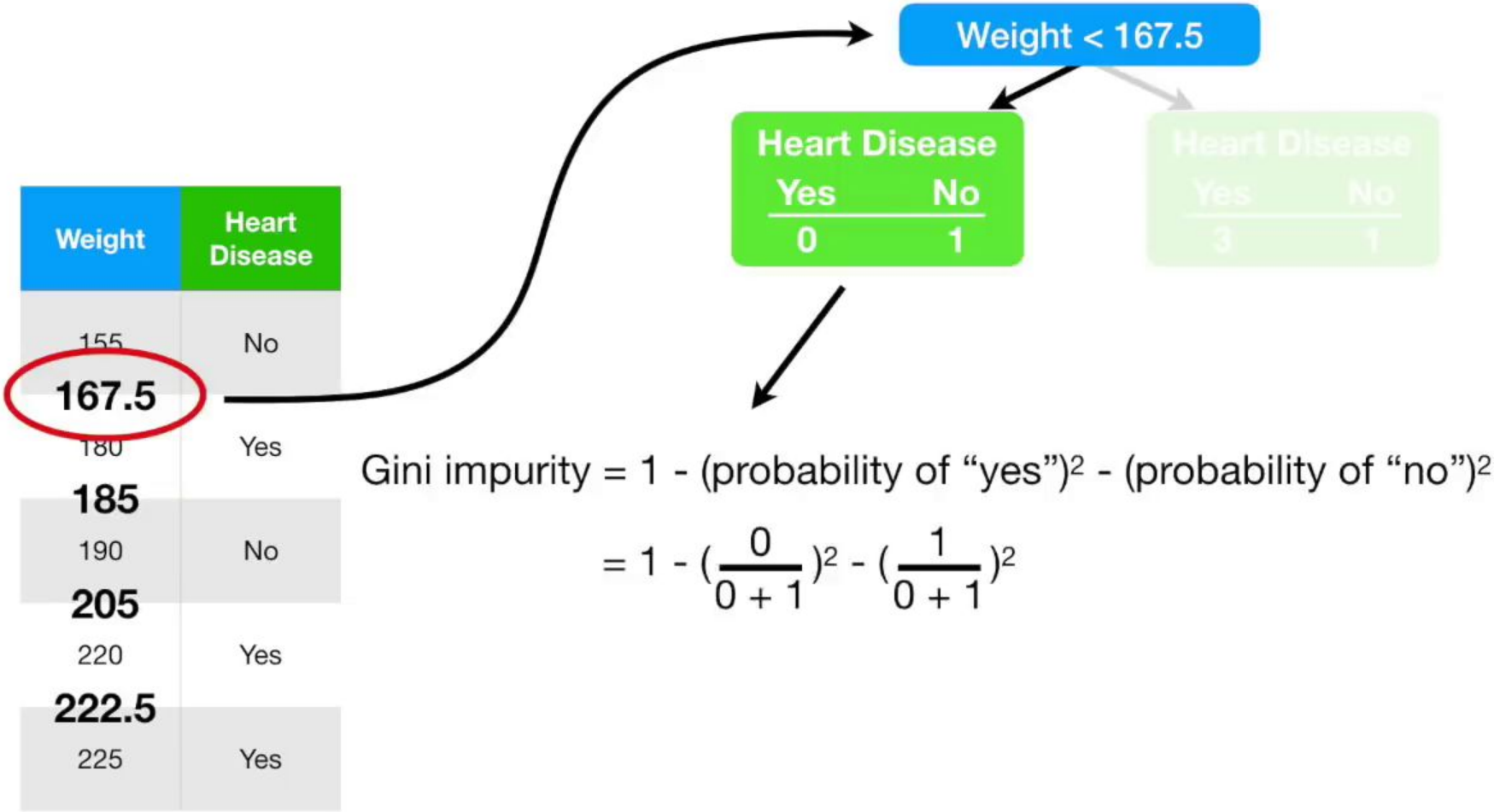
**167.5** → Gini impurity = ?

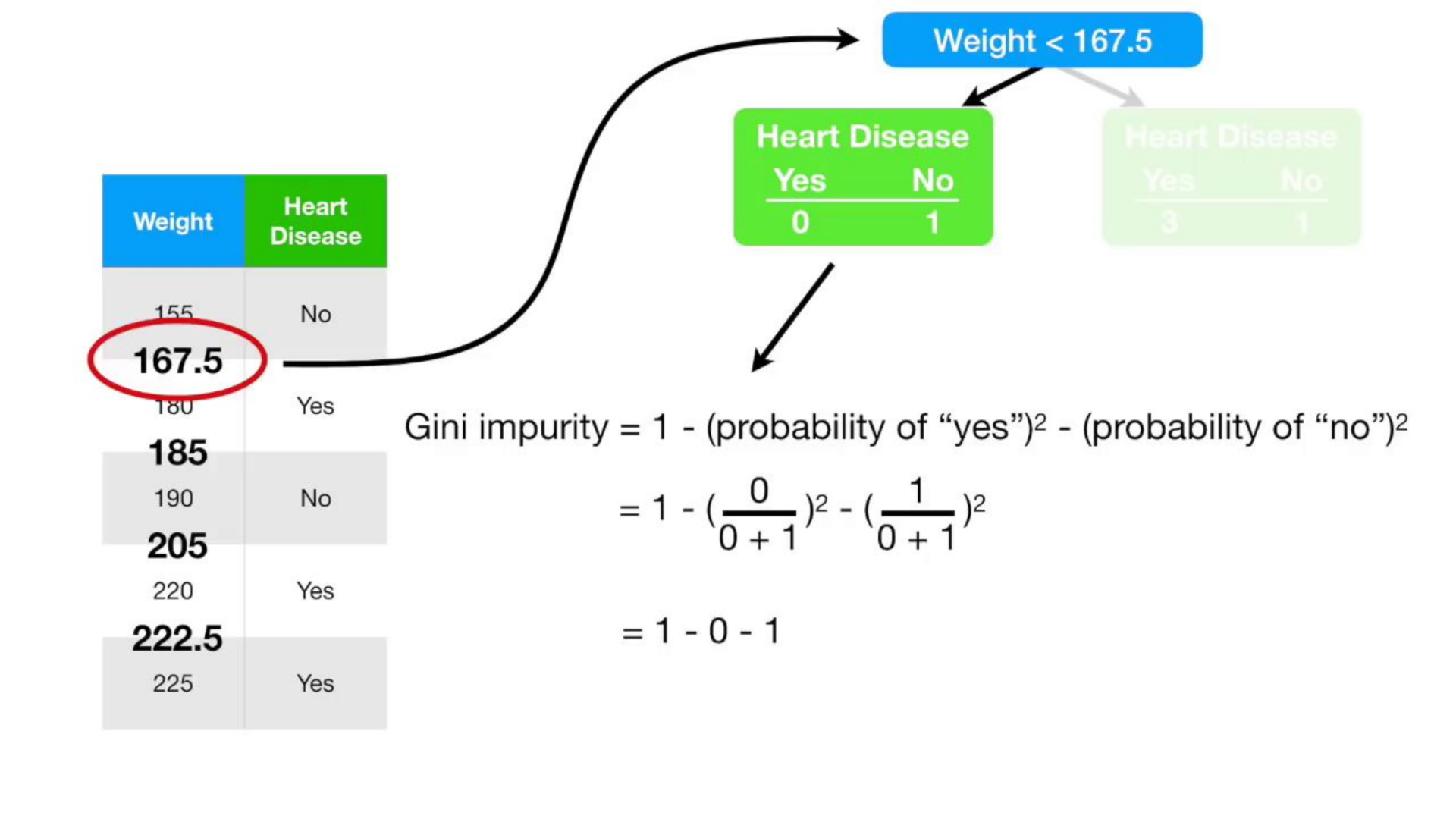
**185** → Gini impurity = ?

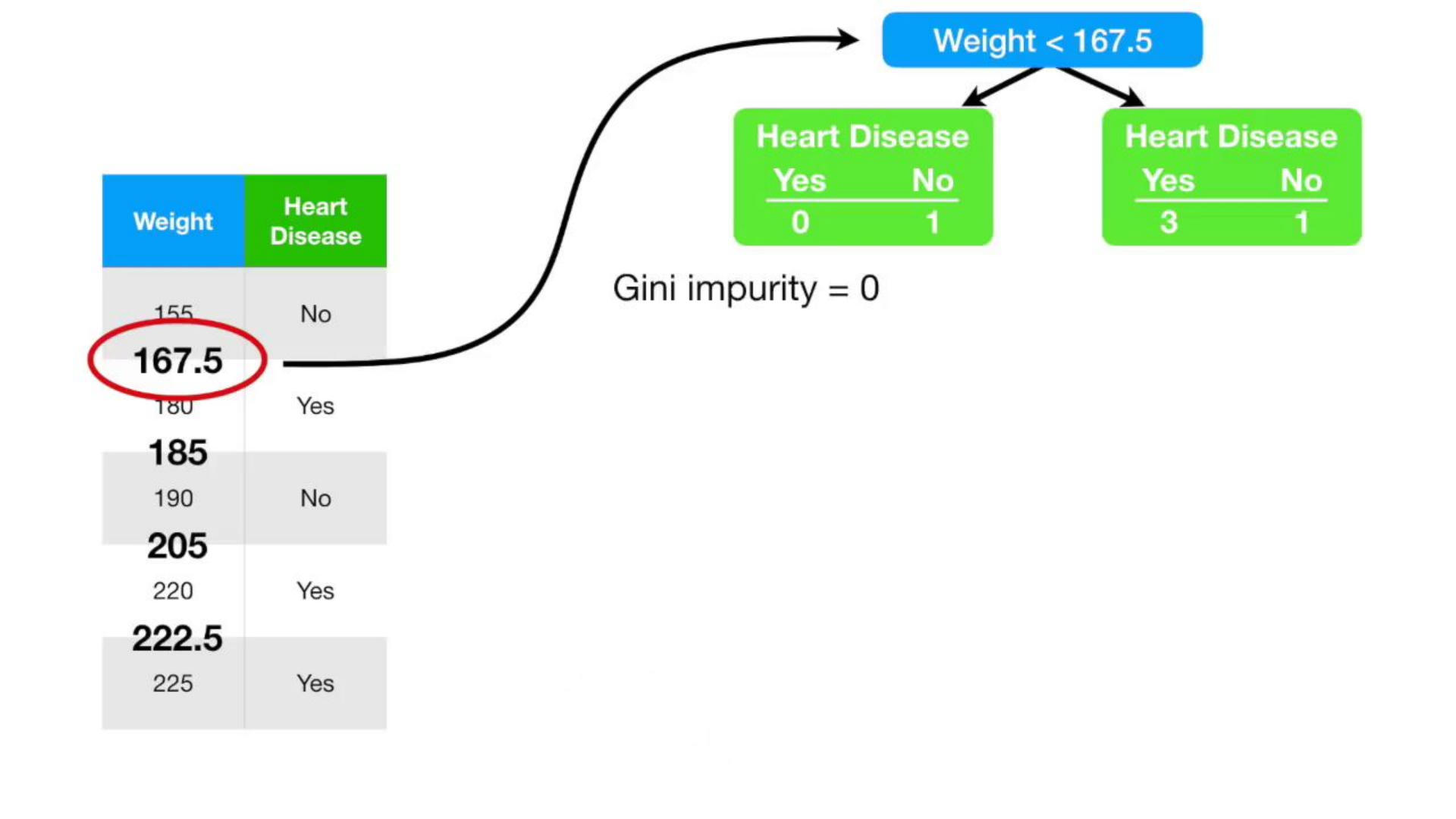
**205** → Gini impurity = ?

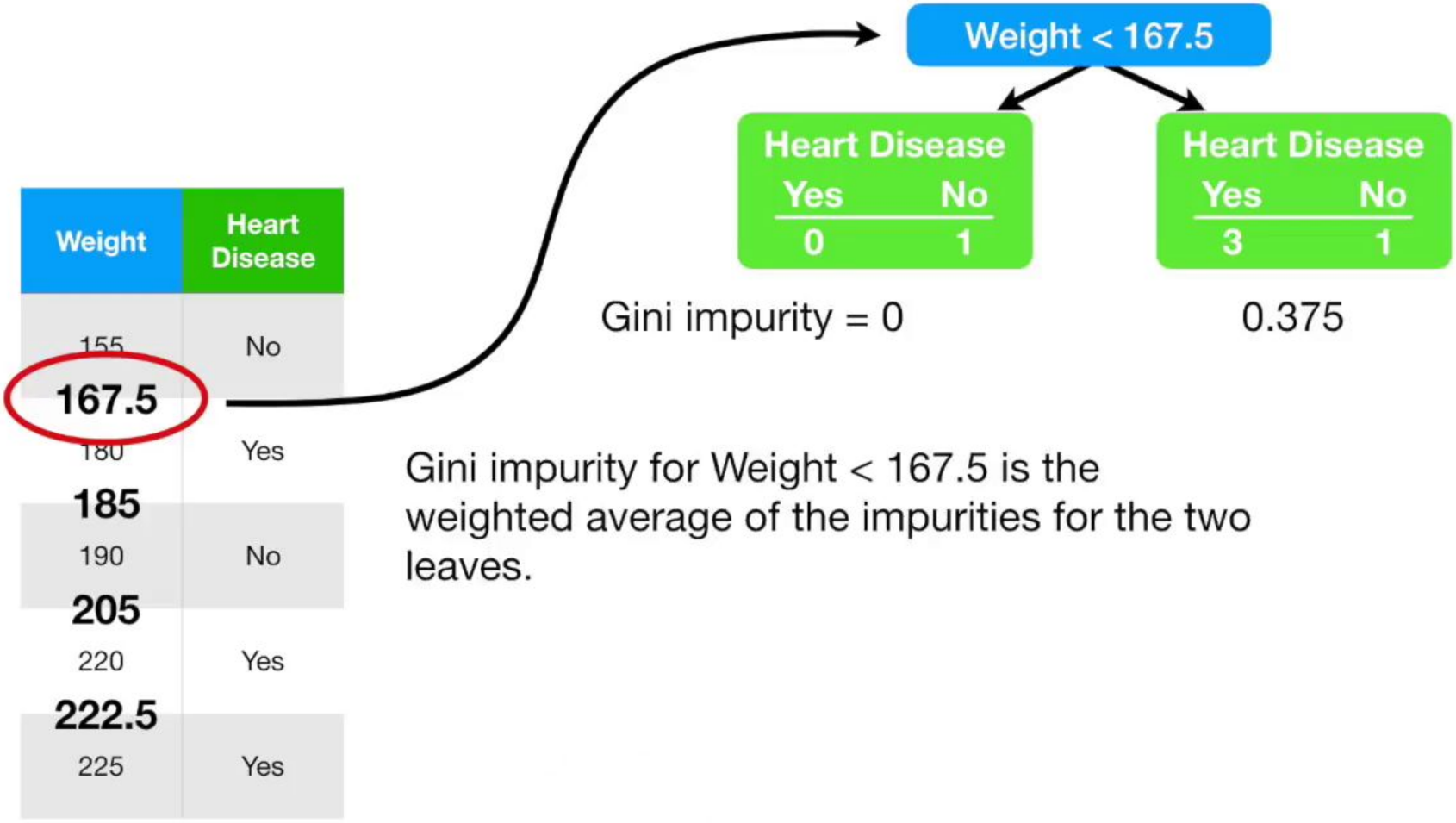
**222.5** → Gini impurity = ?

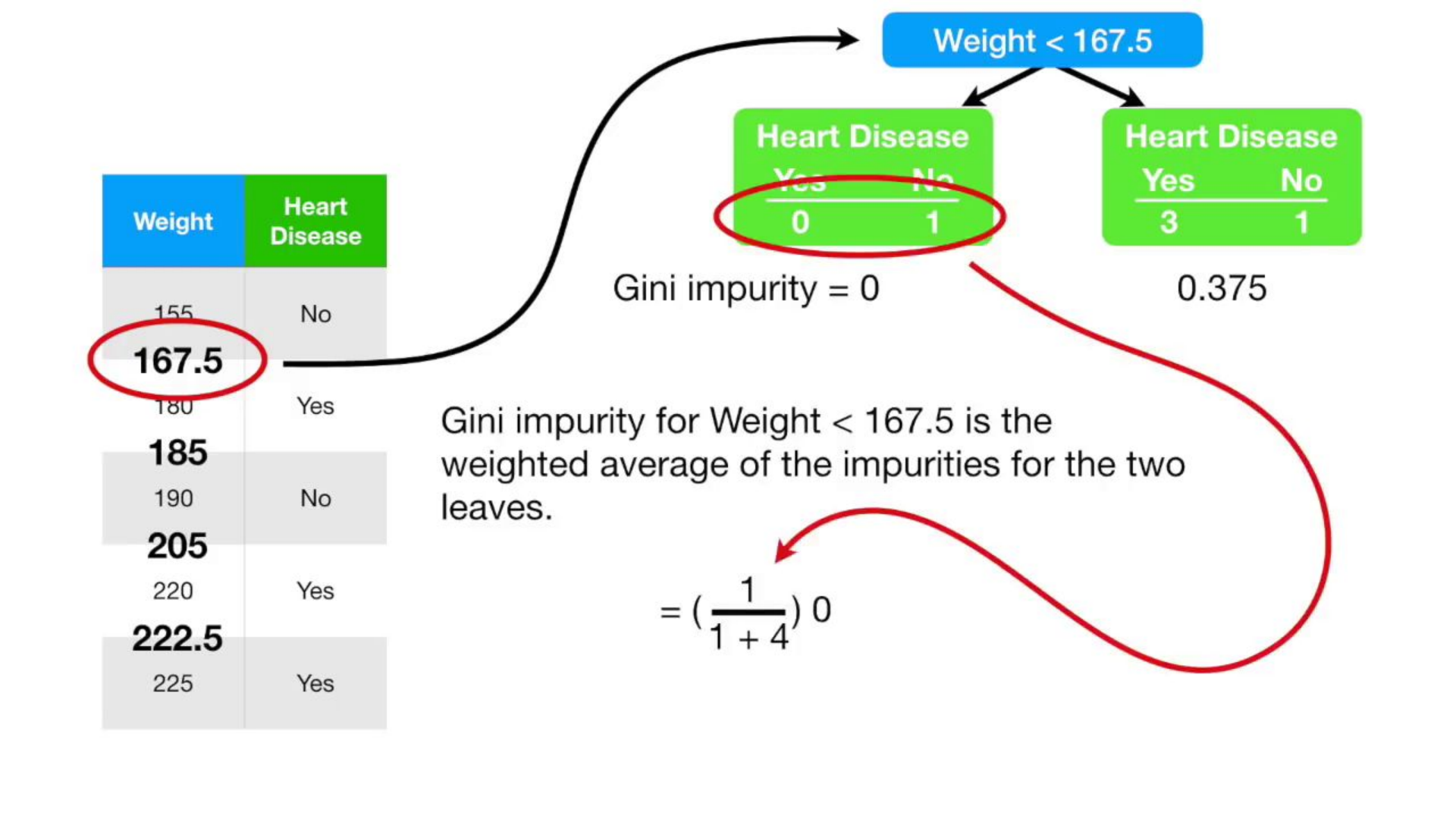

Weight
Heart Disease
155
No
167.5
180
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Weight < 167.5
Heart Disease
Yes
No
0
1
Heart Disease
Yes
No
3
1

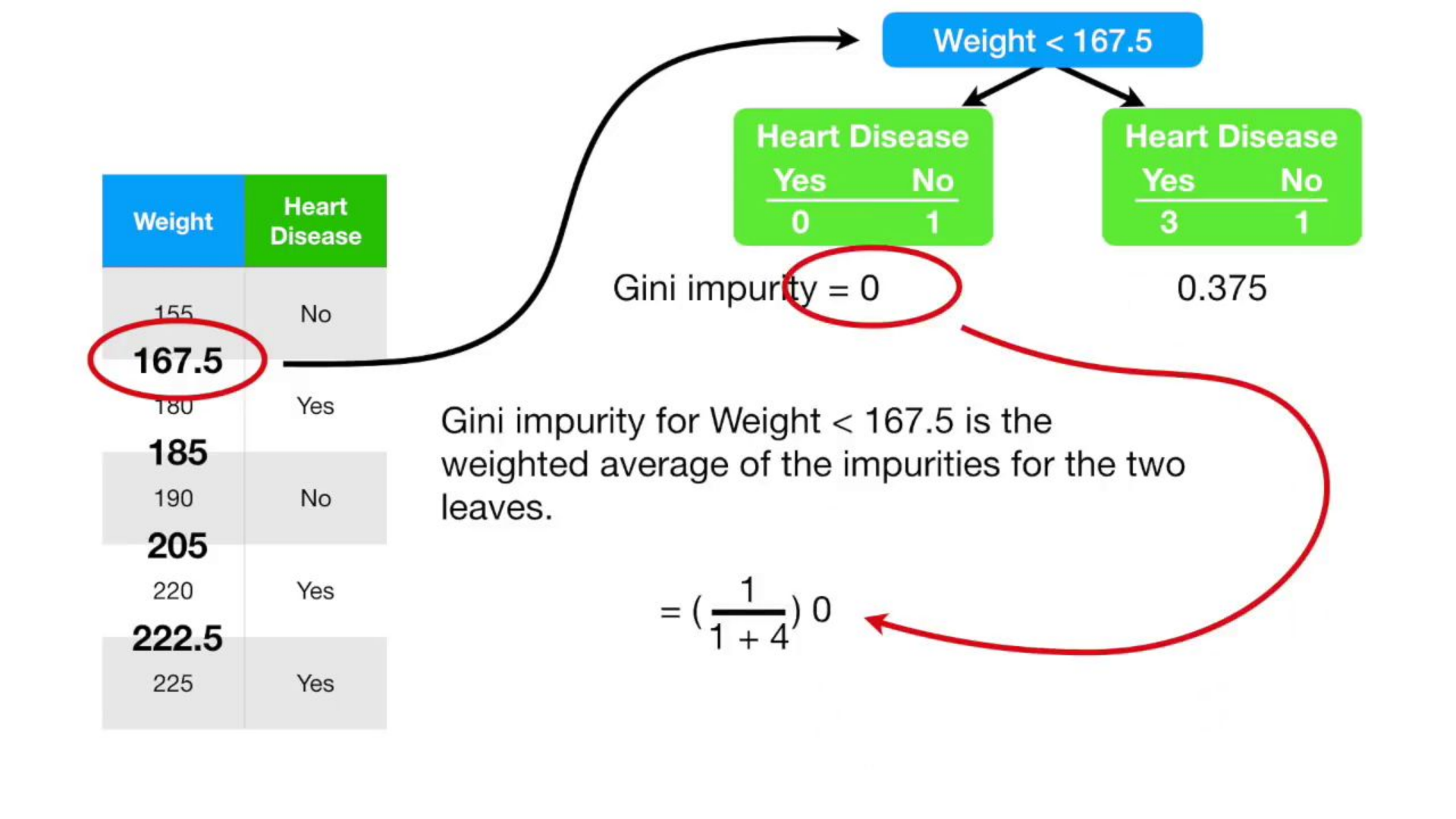

Weight < 167.5
Heart Disease
Yes No
0 1
Heart Disease
Yes No
3 1
Weight
Heart Disease
155
No
167.5
180
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Gini impurity = 1 - (probability of “yes”)² - (probability of “no”)²
= 1 - (0 / 0 + 1)² - (1 / 0 + 1)²


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


Weight < 167.5
Heart Disease
Yes No
0 1
Heart Disease
Yes No
3 1
Gini impurity = 0
Weight
Heart Disease
155
No
167.5
180
Yes
185
190
No
205
220
Yes
222.5
225
Yes


Weight < 167.5
Heart Disease
Yes No
0 1
Heart Disease
Yes No
3 1
Gini impurity = 0
0.375
Weight
Heart Disease
155
No
167.5
180
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Gini impurity for Weight < 167.5 is the weighted average of the impurities for the two leaves.


Weight
Heart Disease
155
No
167.5
180
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Weight < 167.5
Heart Disease
Yes
No
0
1
Heart Disease
Yes
No
3
1
Gini impurity = 0
0.375
Gini impurity for Weight < 167.5 is the weighted average of the impurities for the two leaves.
$= (\frac{1}{1+4})\,0$


Weight < 167.5
Heart Disease
Yes No
0 1
Heart Disease
Yes No
3 1
Gini impurity = 0
0.375
Weight
Heart Disease
155
No
167.5
180
Yes
185
190
No
205
220
Yes
222.5
225
Yes
Gini impurity for Weight < 167.5 is the weighted average of the impurities for the two leaves.
$= (\frac{1}{1+4}) 0$

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


Gini impurity = 0

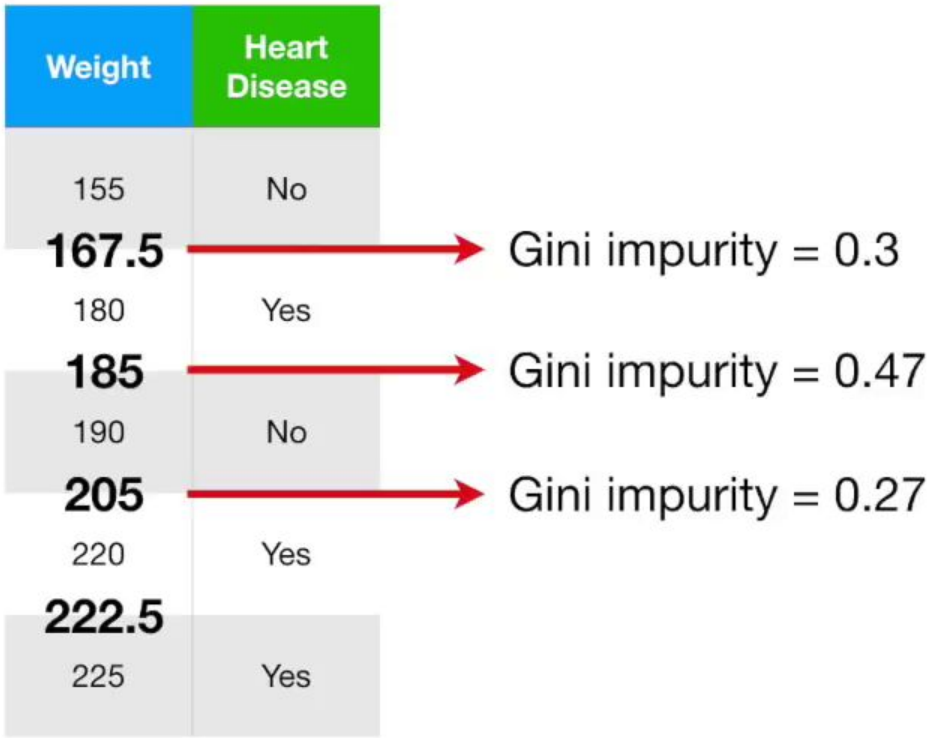
0.375

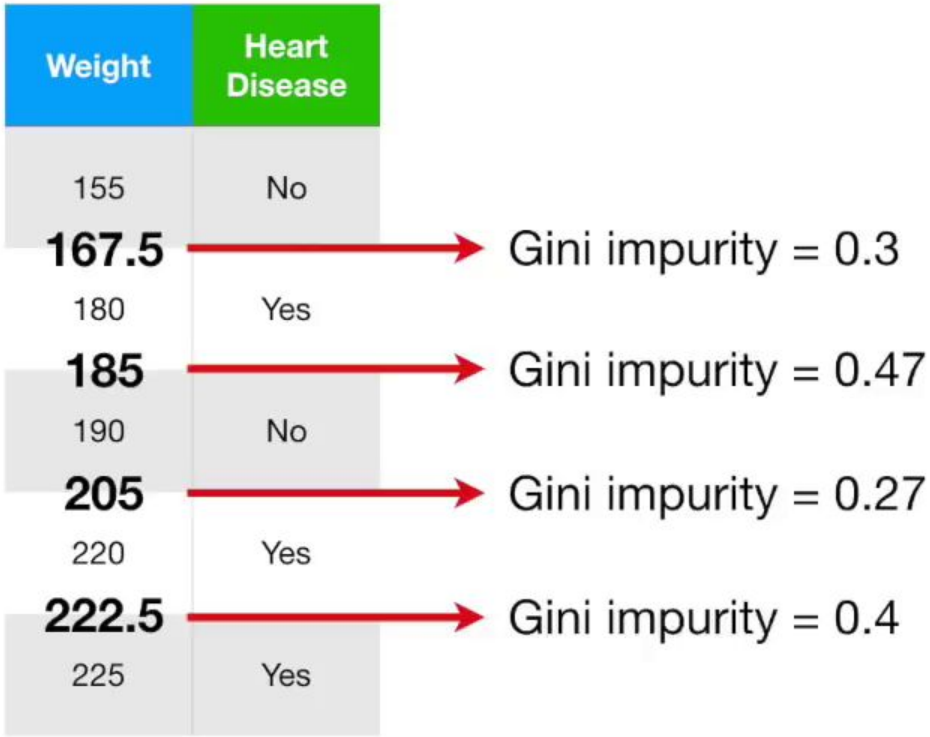
Gini impurity for Weight < 167.5 is the weighted average of the impurities for the two leaves.

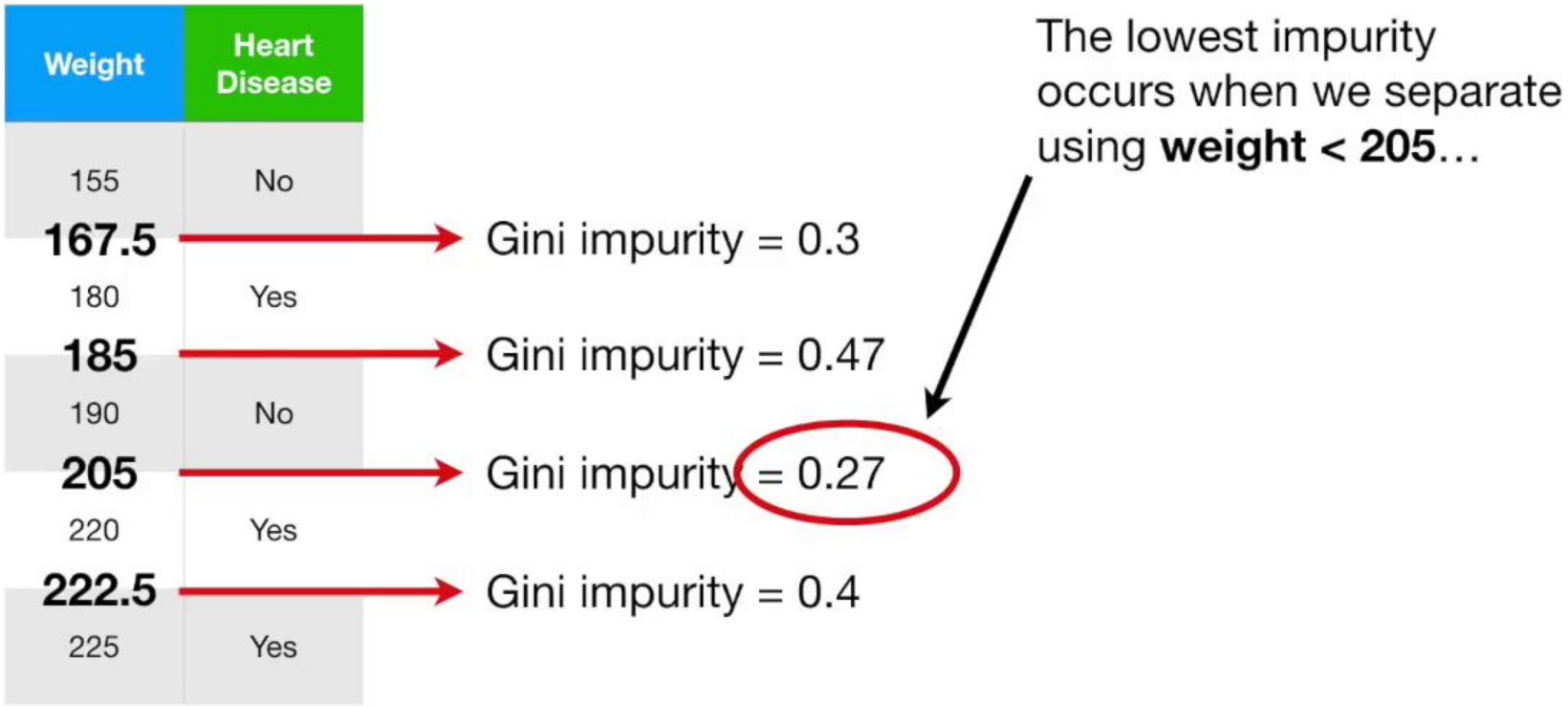
$$= \left(\frac{1}{1+4}\right) 0 + \left(\frac{4}{1+4}\right) 0.336 = 0.3$$

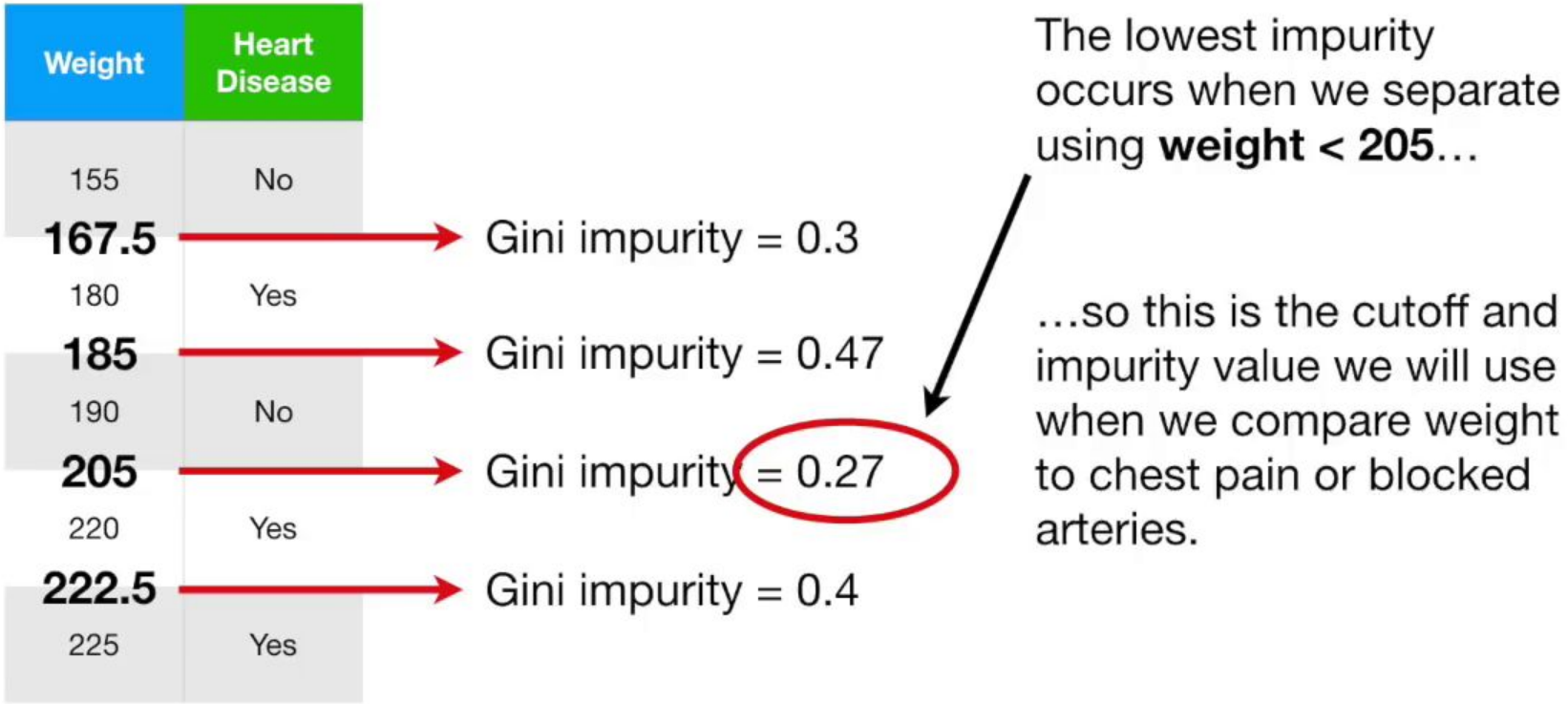
| Weight | Heart Disease |
| --- | --- |
| 155 | No |
| **167.5** | |
| 180 | Yes |
| **185** | |
| 190 | No |
| **205** | |
| 220 | Yes |
| **222.5** | |
| 225 | Yes |

Gini impurity = 0.3

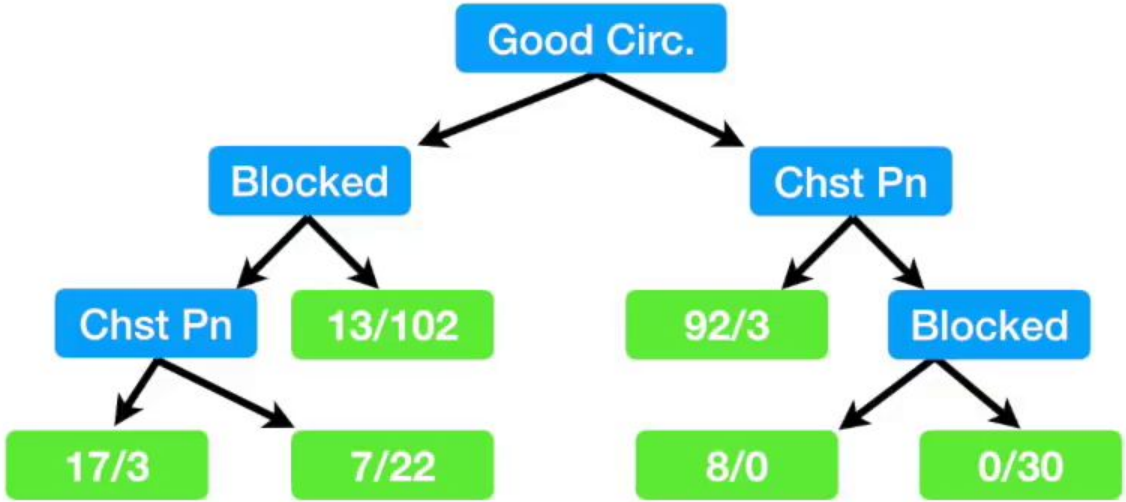

Weight
Heart Disease
155
No
167.5
Gini impurity = 0.3
180
Yes
185
Gini impurity = 0.47
190
No
205
Gini impurity = 0.27
220
Yes
222.5
225
Yes


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


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

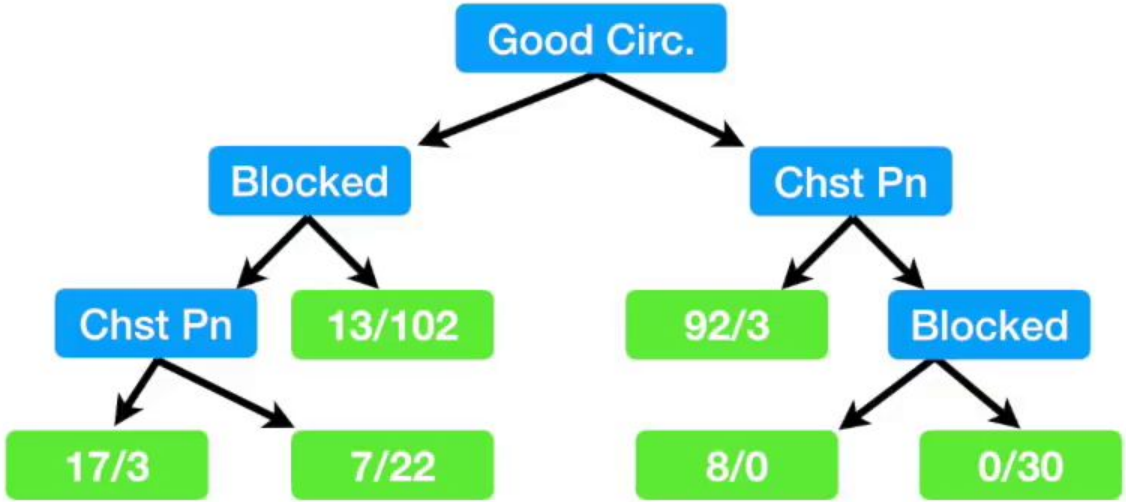

Weight
Heart Disease
155
No
167.5
Gini impurity = 0.3
180
Yes
185
Gini impurity = 0.47
190
No
205
Gini impurity = 0.27
220
Yes
222.5
Gini impurity = 0.4
225
Yes
The lowest impurity occurs when we separate using **weight < 205**…
…so this is the cutoff and impurity value we will use when we compare weight to chest pain or blocked arteries.

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



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

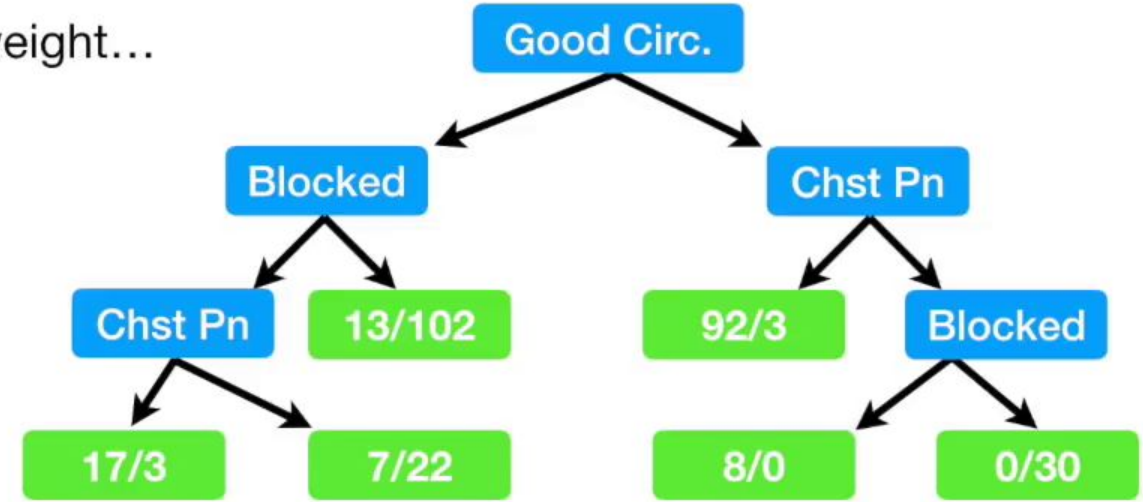
1) "yes/no" questions at each step…



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

1) "yes/no" questions at each step…

2) Numeric data, like patient weight…

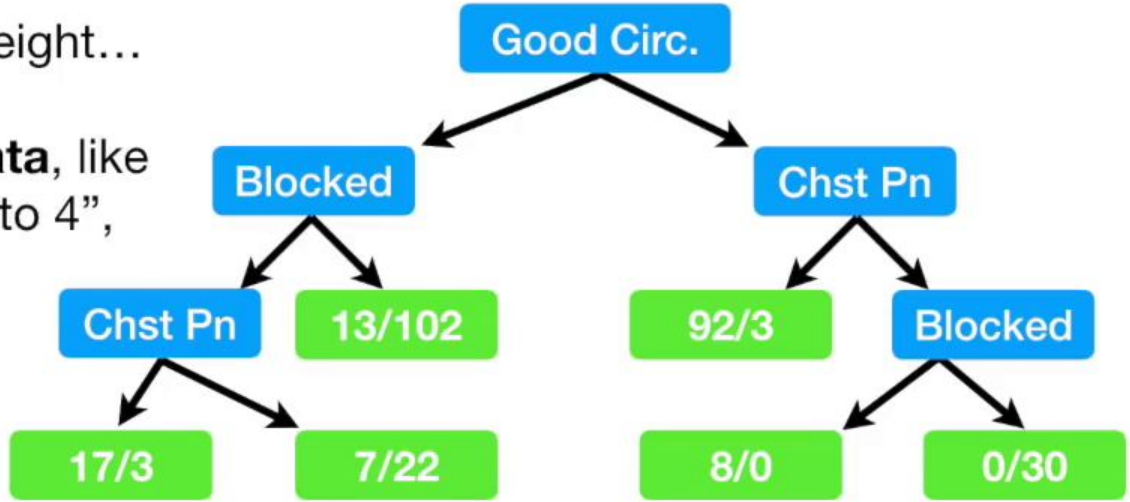


Now we've seen how to build a tree with…

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… |

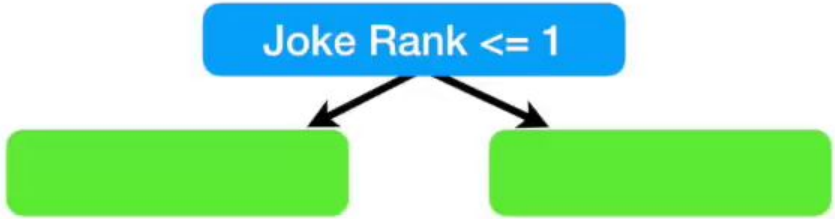
Ranked data is similar to numeric data, except instead now we calculate impurity scores for all of the possible ranks.

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

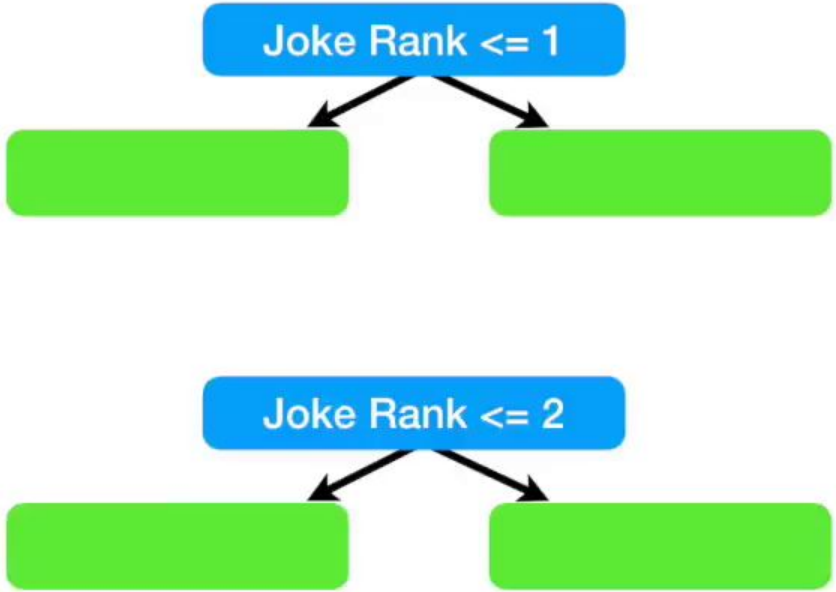
Ranked data is similar to numeric data, except instead now we calculate impurity scores for all of the possible ranks.

So if people could rank my jokes from 1 to 4 (4 being the funniest), we could calculate the following impurity scores…

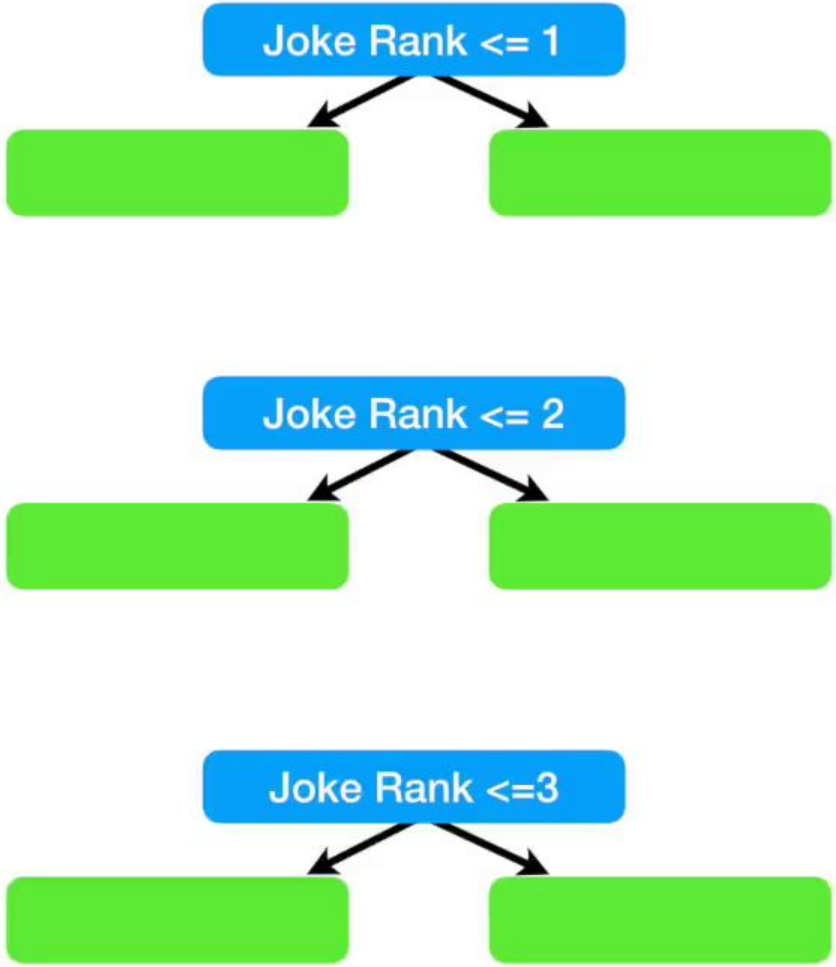
| Rank my jokes… | Likes StatQuest |
|---|---|
| 1 | Yes |
| 1 | No |
| 3 | Yes |
| 1 | Yes |
| etc… | etc… |



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

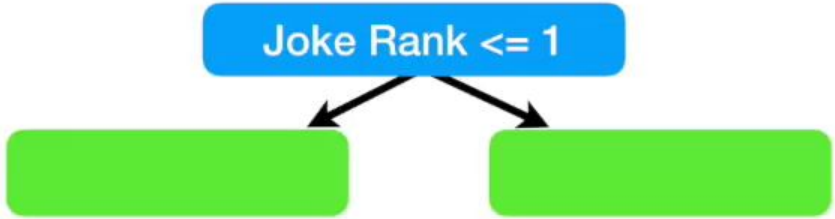


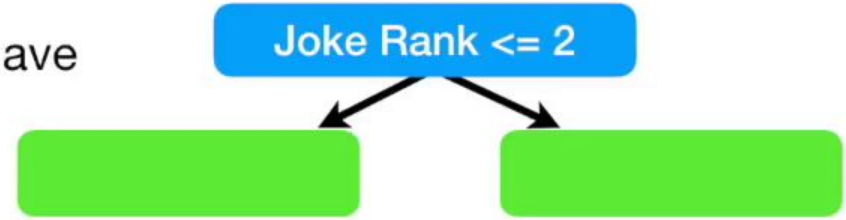
| Rank my jokes… | Likes StatQuest |
|---|---|
| 1 | Yes |
| 1 | No |
| 3 | Yes |
| 1 | Yes |
| etc… | etc… |

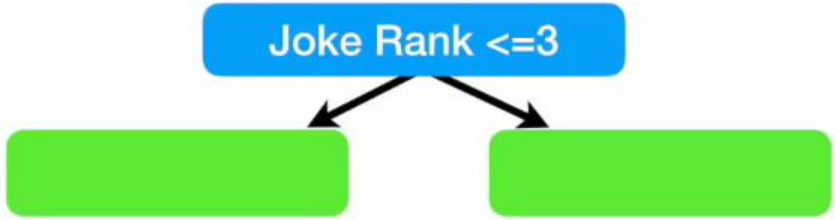


| Rank my jokes… | Likes StatQuest |
|---|---|
| 1 | Yes |
| 1 | No |
| 3 | Yes |
| 1 | Yes |
| etc… | etc… |

**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… |

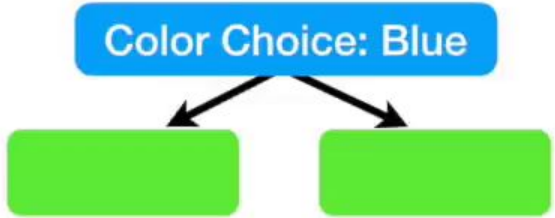
When there are **multiple choices**, like "**color choice can be blue, green or red**", you calculate an impurity score for each one as well as each possible combination.

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

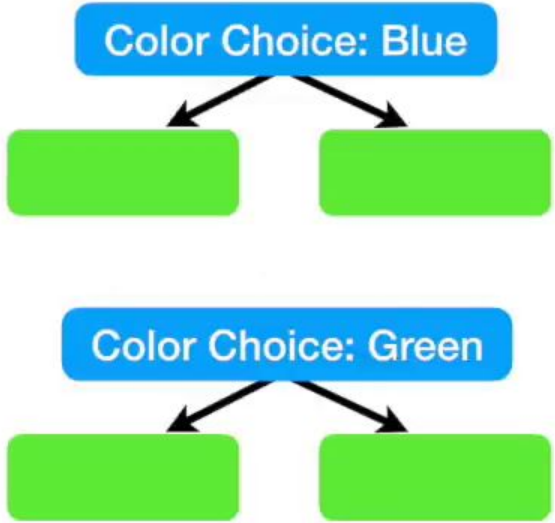
When there are **multiple choices**, like "**color choice can be blue, green or red**", you calculate an impurity score for each one as well as each possible combination.

For this example, with three colors (blue, green and red) we get the following options…

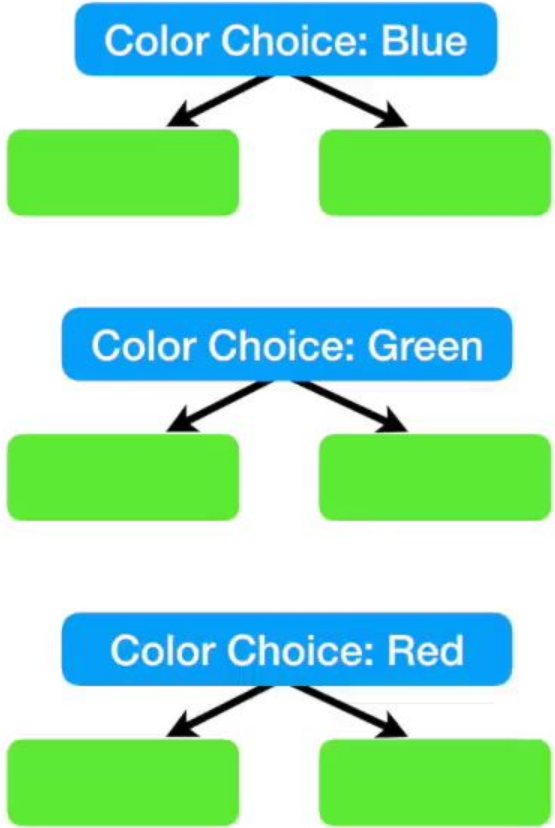
| Color Choice | Likes StatQuest |
|---|---|
| Green | Yes |
| Blue | No |
| Red | Yes |
| Green | Yes |
| etc… | etc… |



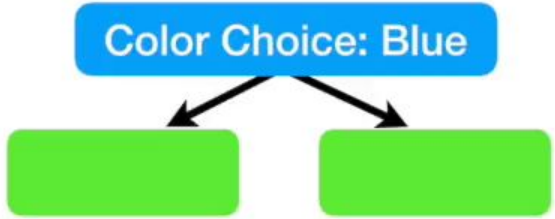
| Color Choice | Likes StatQuest |
|---|---|
| Green | Yes |
| Blue | No |
| Red | Yes |
| Green | Yes |
| etc… | etc… |

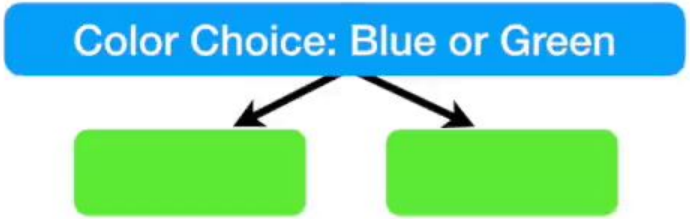


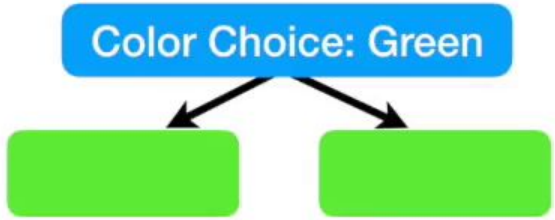
| Color Choice | Likes StatQuest |
|---|---|
| Green | Yes |
| Blue | No |
| Red | Yes |
| Green | Yes |
| etc… | etc… |

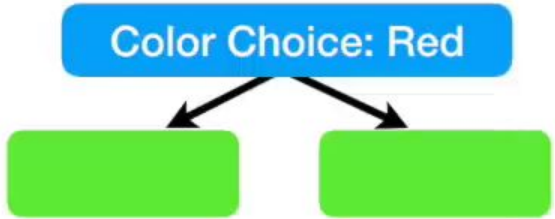


| Color Choice | Likes StatQuest |
|---|---|
| Green | Yes |
| Blue | No |
| Red | Yes |
| Green | Yes |
| etc… | etc… |

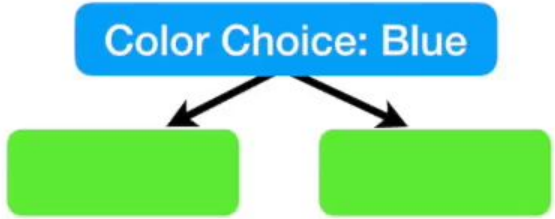


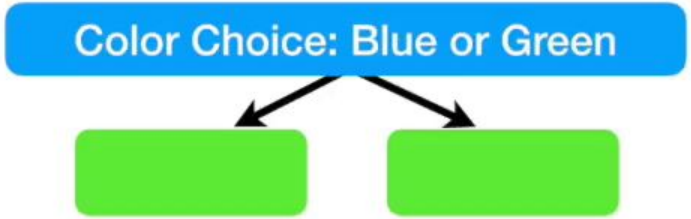



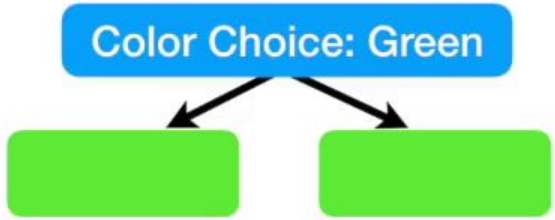



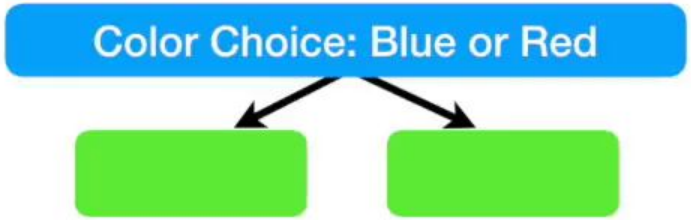



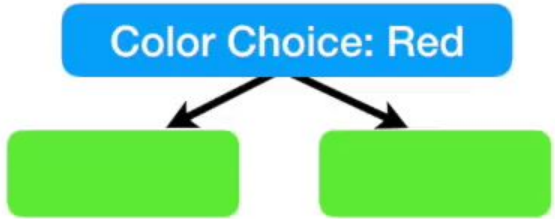
| Color Choice | Likes StatQuest |
|---|---|
| Green | Yes |
| Blue | No |
| Red | Yes |
| Green | Yes |
| etc… | etc… |



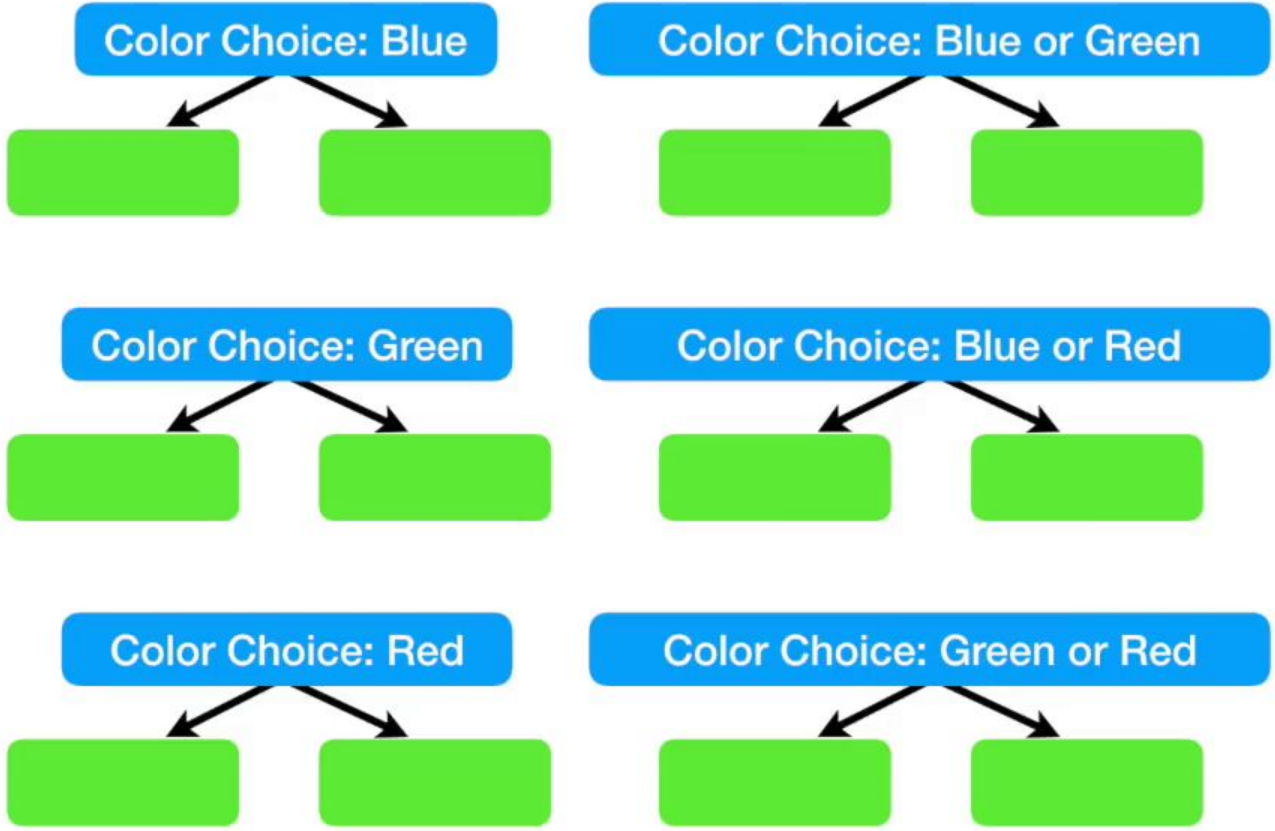



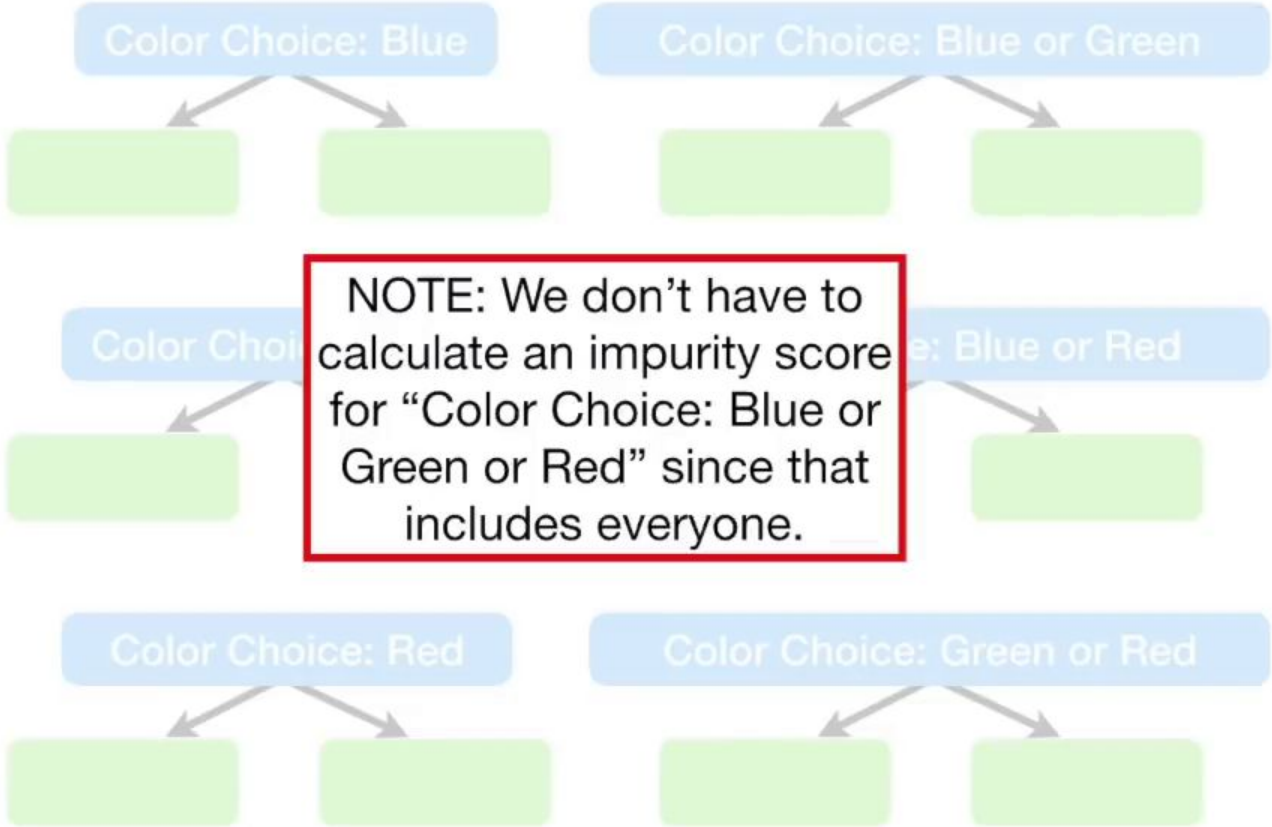



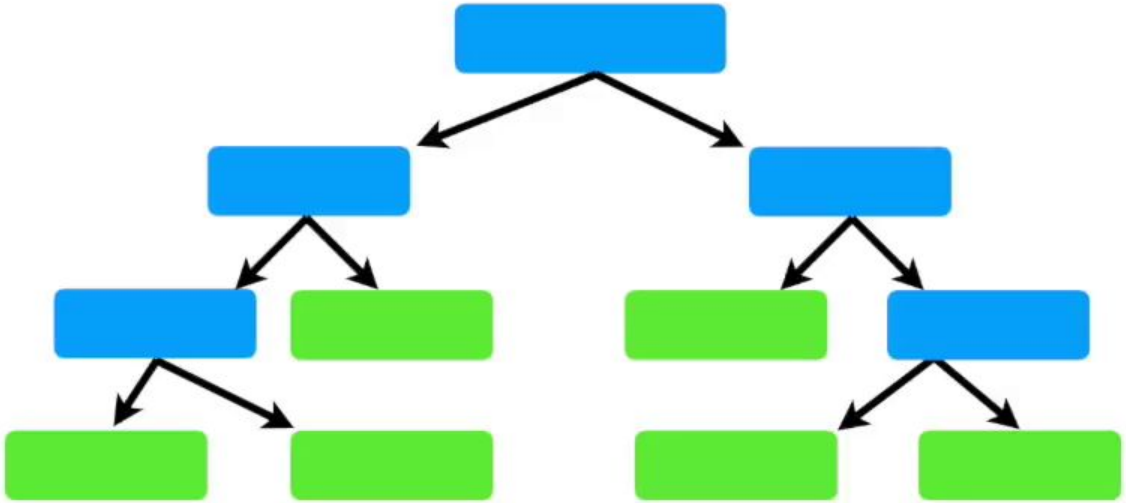
| Color Choice | Likes StatQuest |
|---|---|
| Green | Yes |
| Blue | No |
| Red | Yes |
| Green | Yes |
| etc… | etc… |



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



BAM!!! Now we know how to make a use decision trees!!!



# The End!!!