



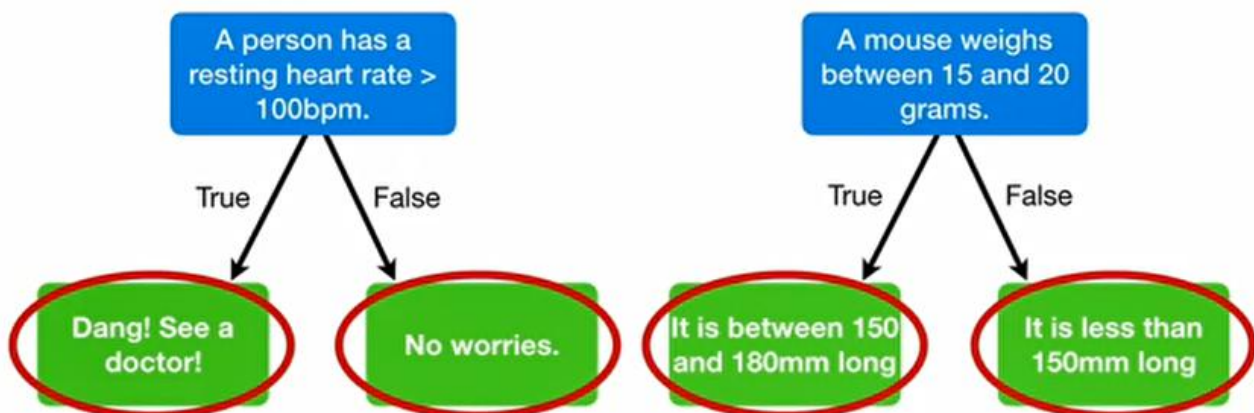
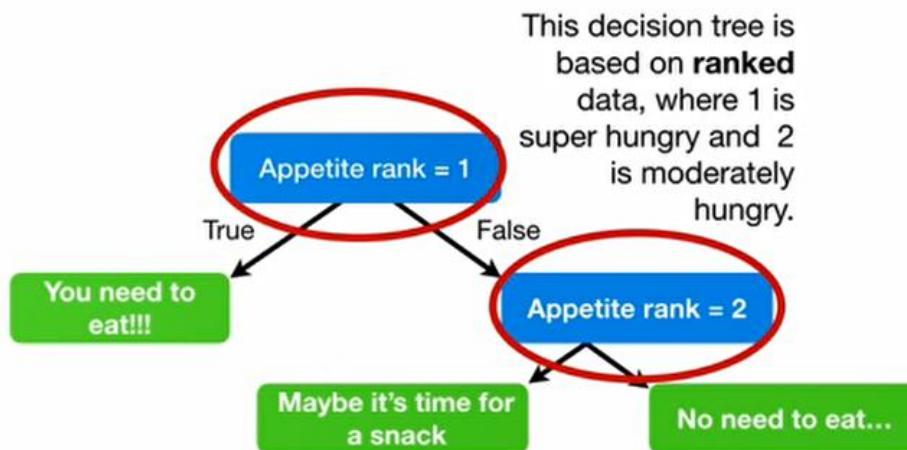
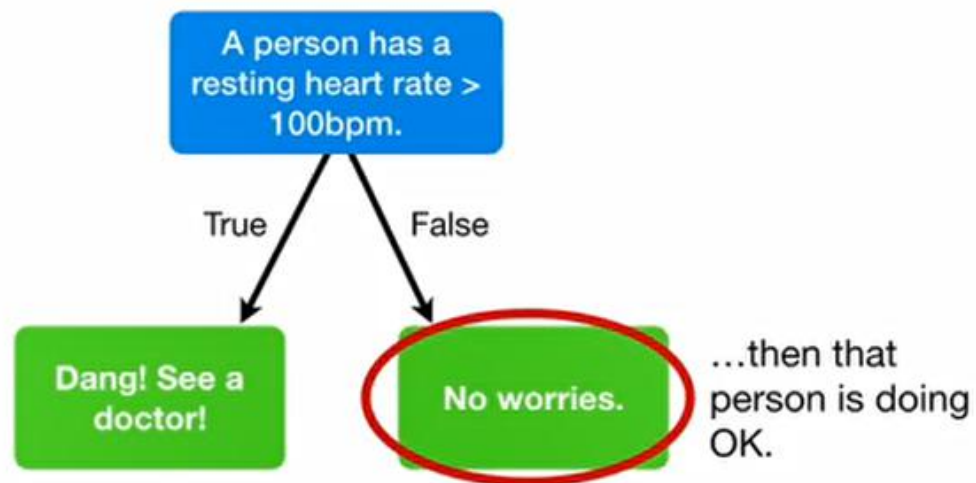
ANTICIPARE LA CRESCITA CON LE NUOVE COMPETENZE SUI BIG DATA – EDIZIONE 2

Operazione Rif. PA 2019-11596/RER “Anticipare la crescita con le nuove competenze sui Big Data - Edizione 2”, approvata dalla Regione Emilia-Romagna con DGR n° 789 del 20 maggio 2019 e co-finanziata dal Fondo Sociale Europeo PO 2014-2020

Prog. 4 Ed. 7 Titolo “Tecnologie & Software di Data Science



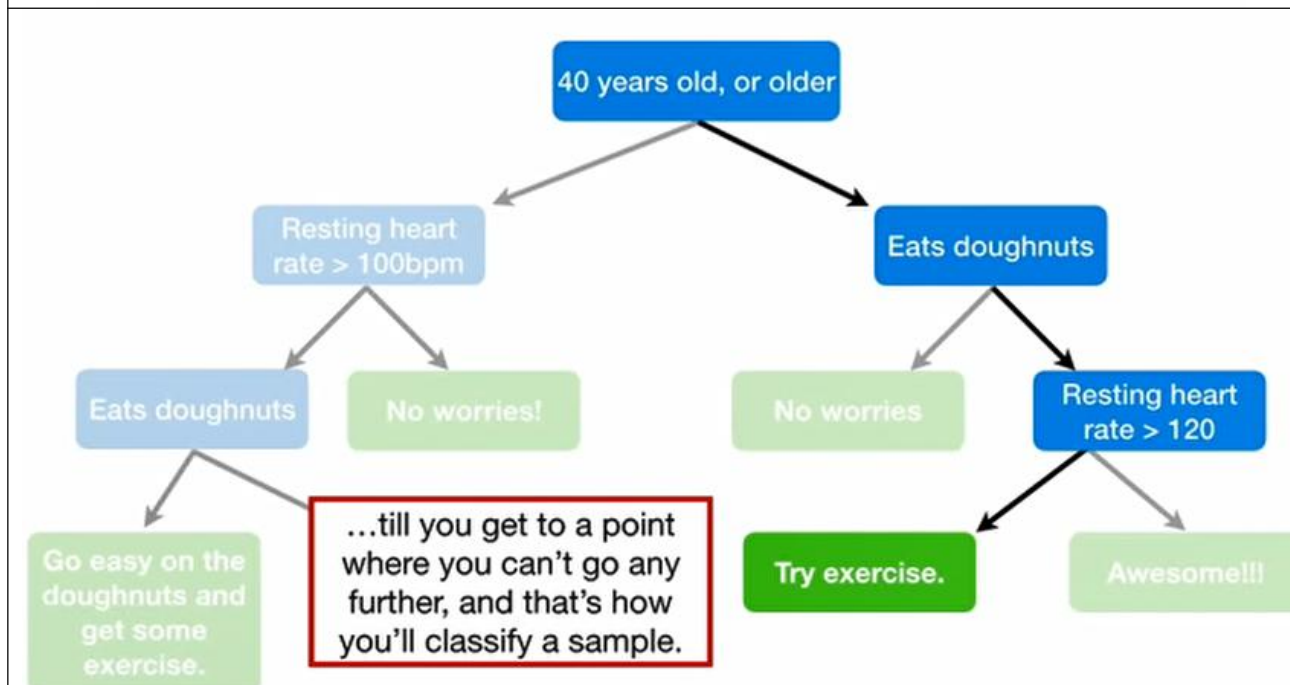
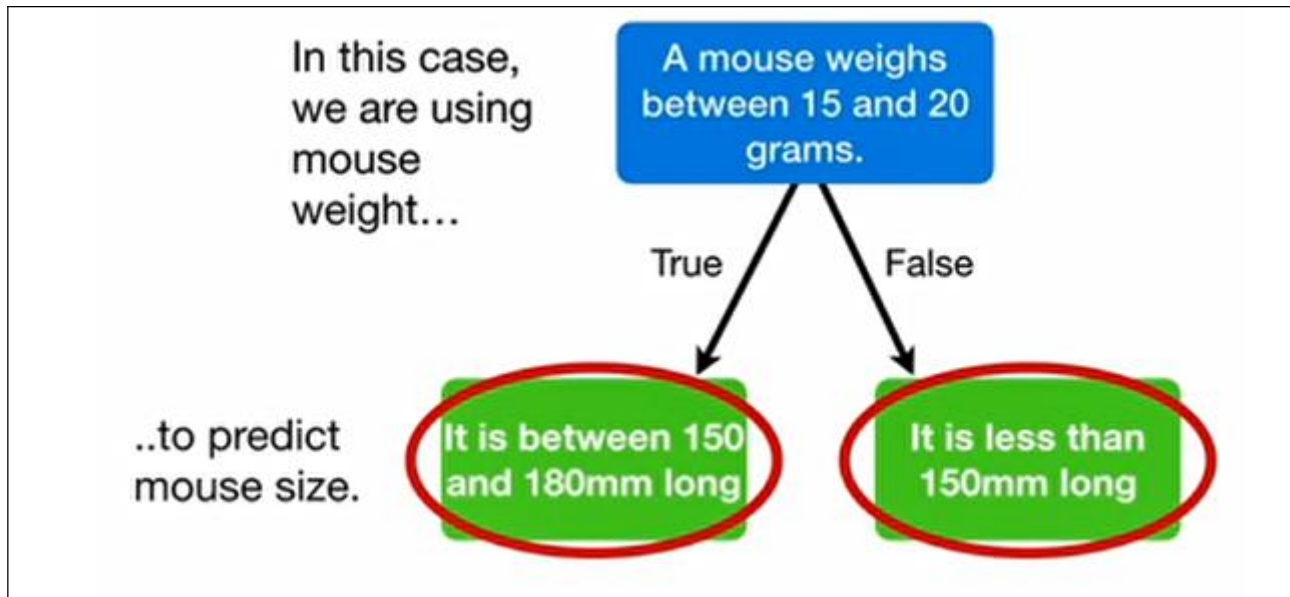
DECISION TREE



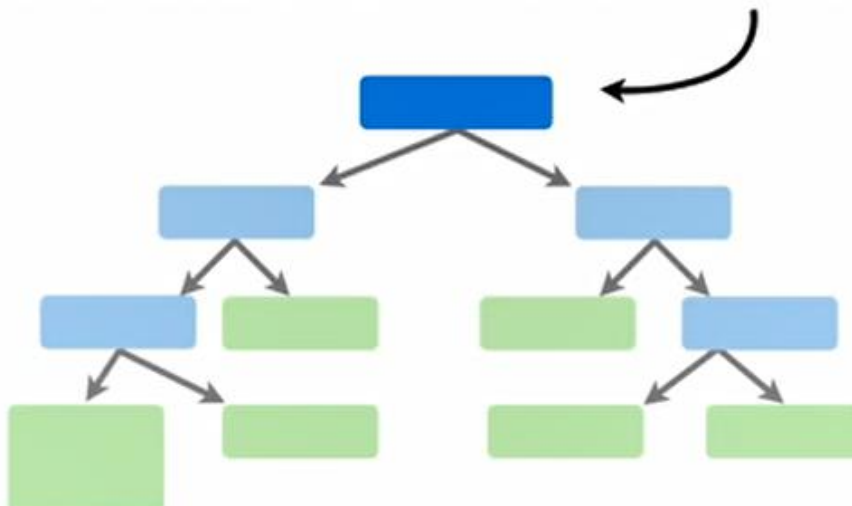
NOTE: The classification can be categories...

...or numeric.

DECISION TREE

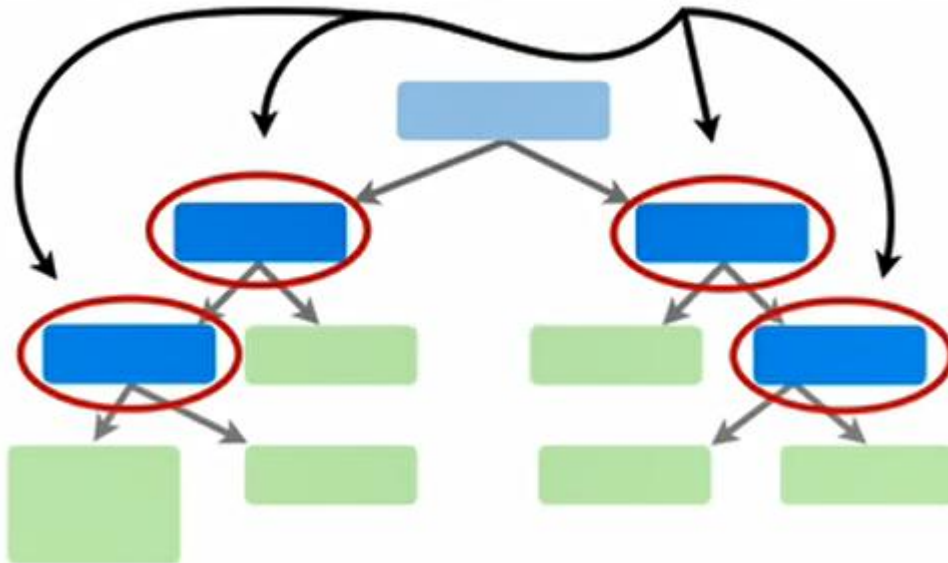


The very top of the tree is called the **“Root Node”** or just **“The Root”**

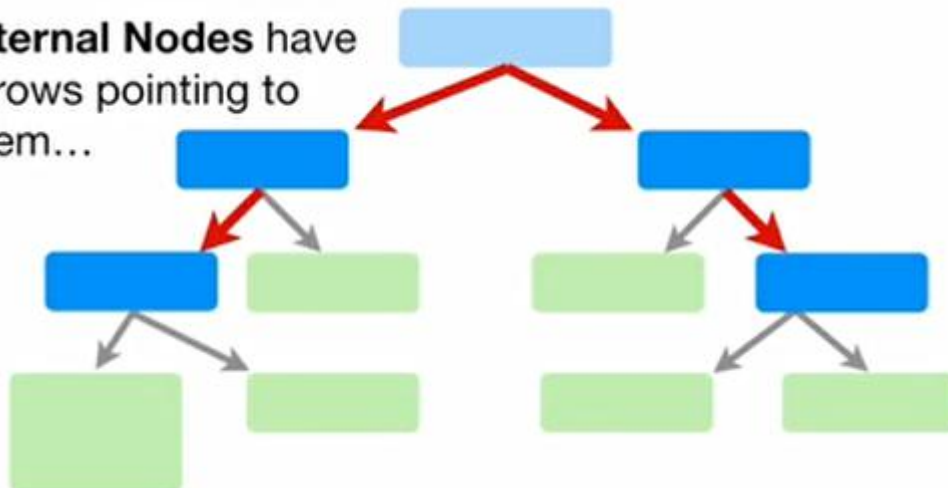


DECISION TREE

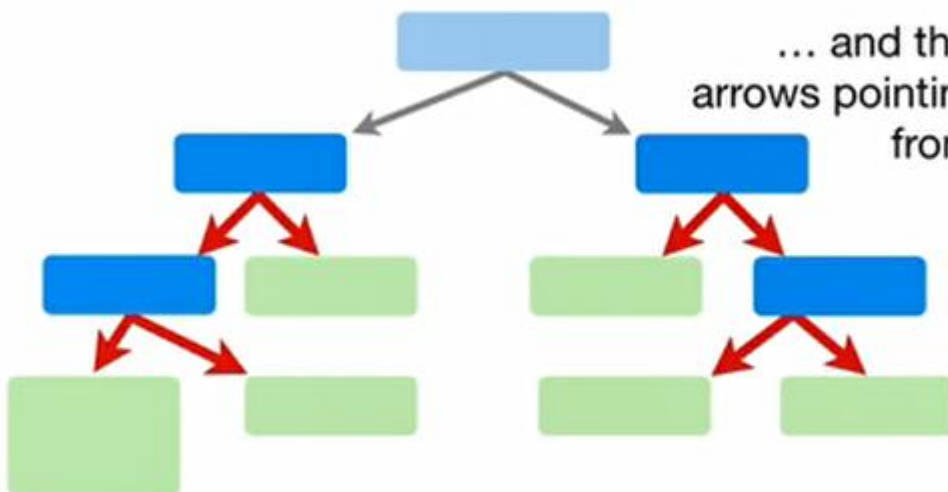
These are called “**Internal Nodes**”, or just “**Nodes**”.



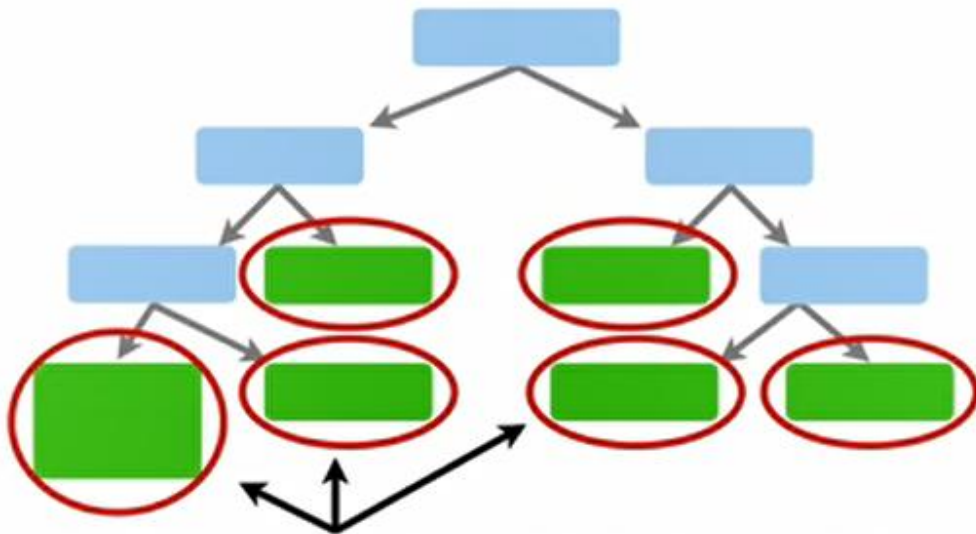
Internal Nodes have
arrows pointing
to them...



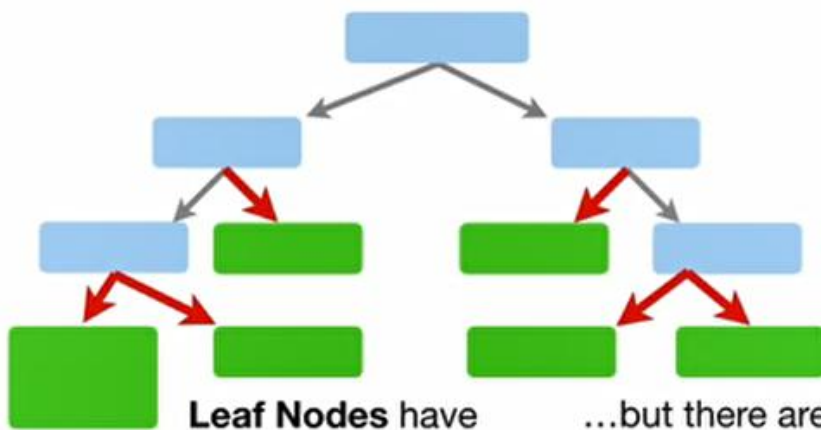
... and they have
arrows pointing away
from them.



DECISION TREE



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



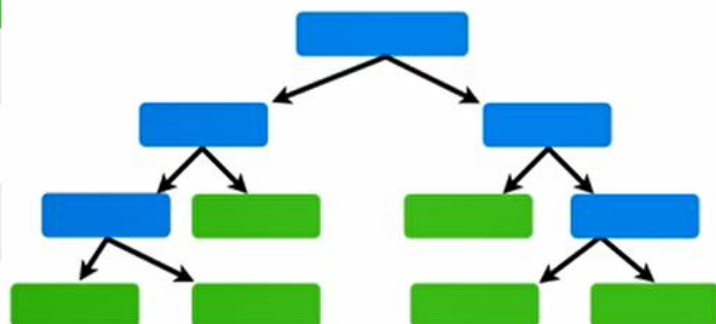
Leaf Nodes have arrows pointing to them...

...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...

...to a decision tree!!!

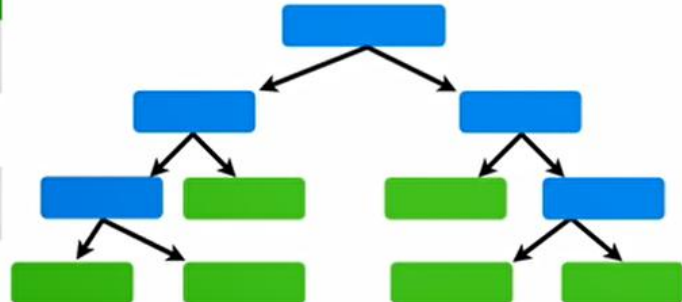
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...



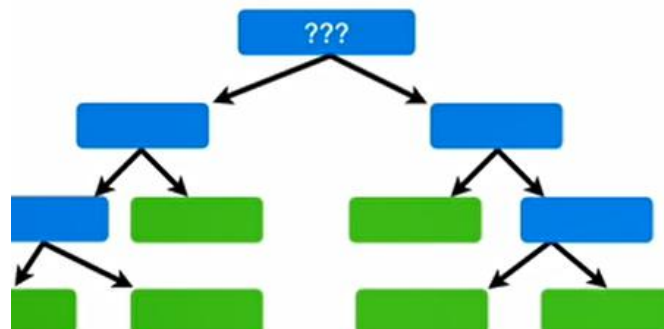
DECISION TREE

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...



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.



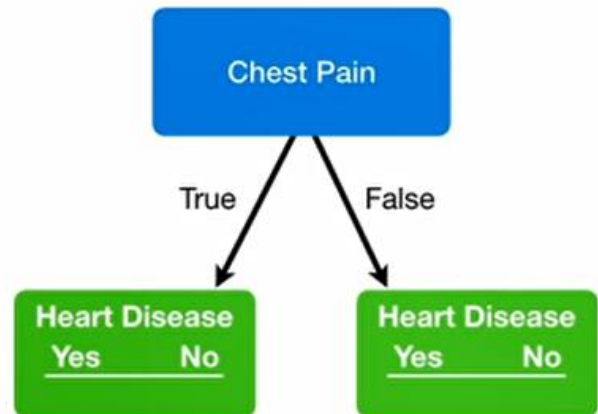
We start by looking at how well **Chest Pain** alone predicts heart disease...

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...

DECISION TREE

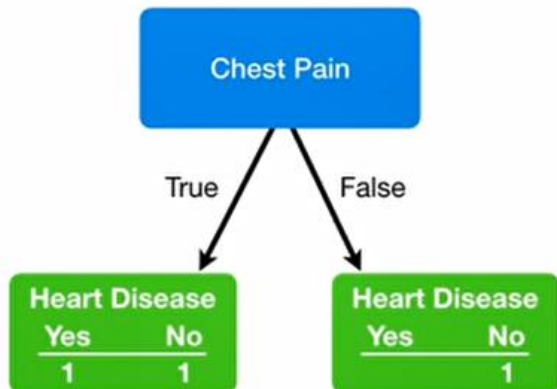
Here's a little tree that only takes chest pain into account.

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



The 4th patient has chest pain and heart disease.

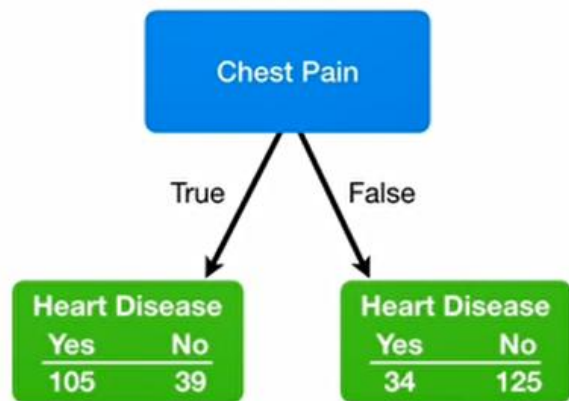
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...



a

DECISION TREE

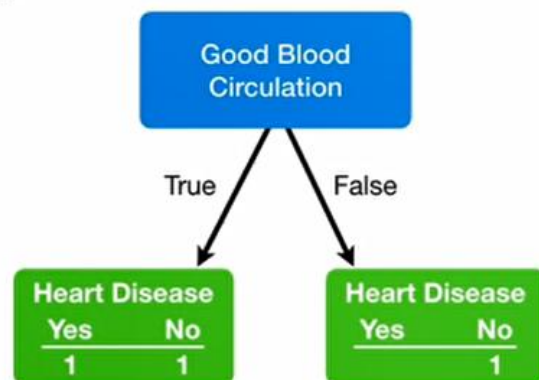
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...



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

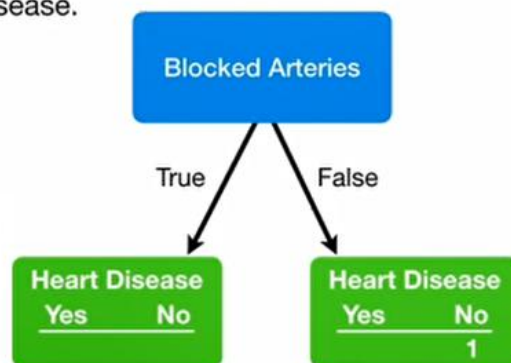
Now we do the exact same thing for **Good Blood Circulation**.

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...



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

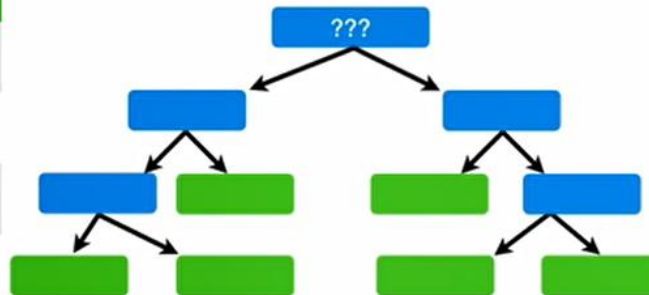
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...



DECISION TREE

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...

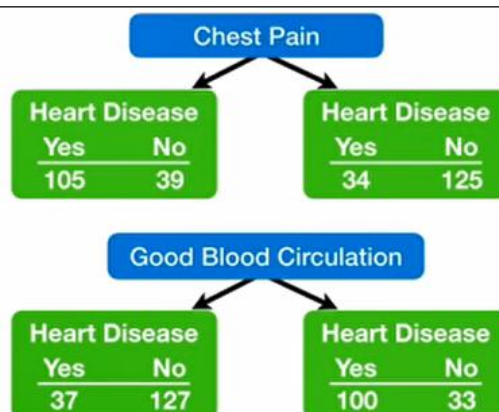


Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes



Most of the patients with heart disease ended up in this leaf 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...

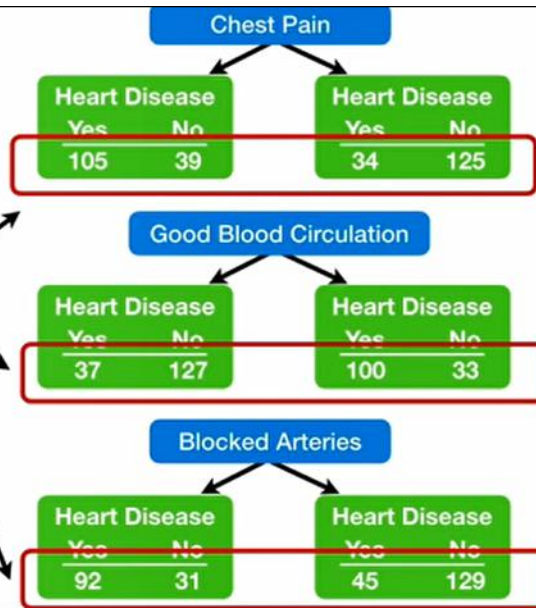


Then we looked at how well **Good Blood Circulation** separated patients with and without heart disease.

It wasn't perfect either.

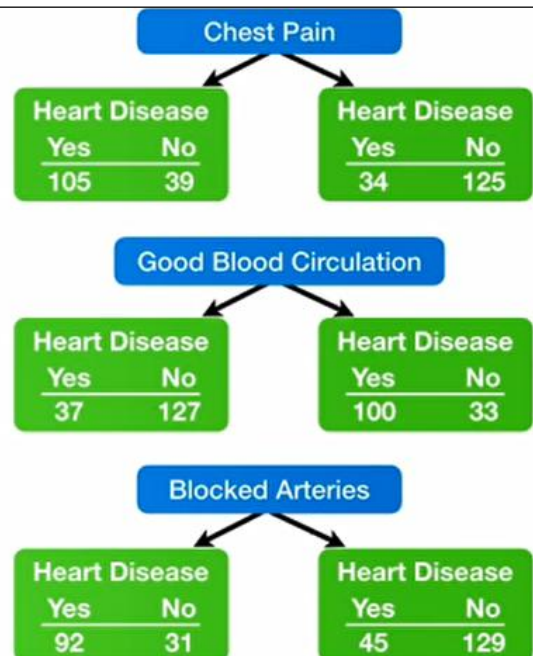
DECISION TREE

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.

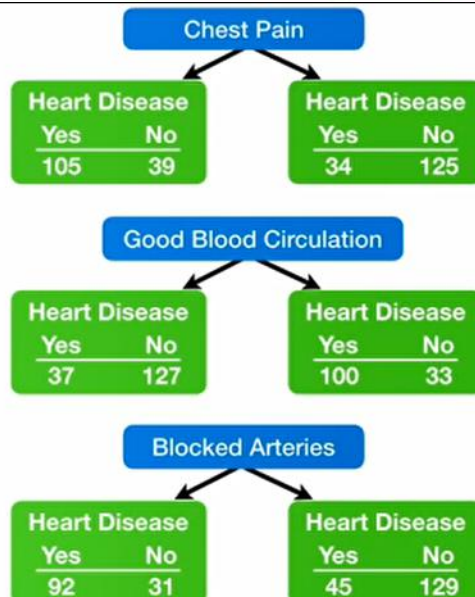


Because none of the leaf nodes are 100% "YES Heart Disease" or 100% "NO Heart Disease", they are all considered "impure".

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".



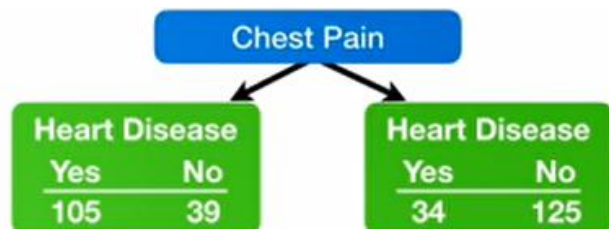
DECISION TREE



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

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

$$= 0.395$$



Gini impurity = 0.395

Now let's calculate the Gini impurity for this leaf node...



Gini impurity = 0.395

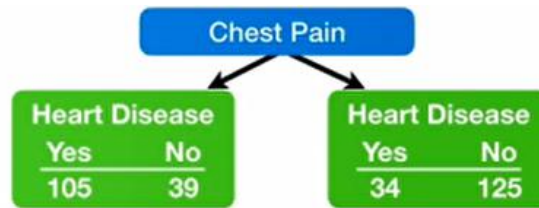
0.336

Because this leaf node represents 144 patients...

... and this leaf node represents 159 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**.

DECISION TREE



Gini impurity = 0.395

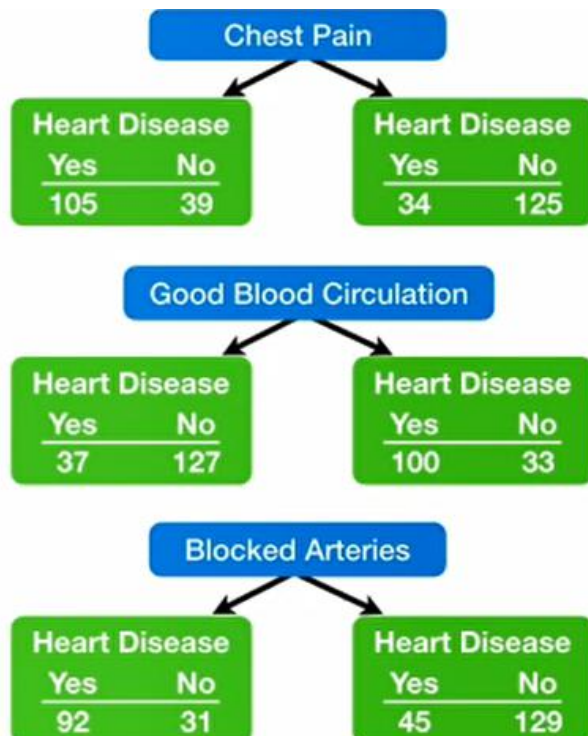
0.336

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

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

$$= 0.364$$

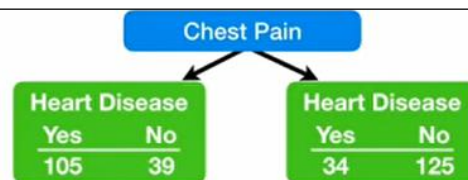
Gini impurity for Chest Pain = 0.364



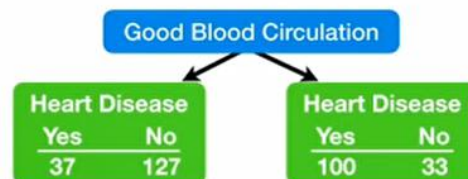
DECISION TREE

-..... SOLVE the Problem ... 10 min.

Gini impurity for Chest Pain = 0.364



Gini impurity for Good Blood Circulation = 0.360



Gini impurity for Blocked Arteries = 0.381



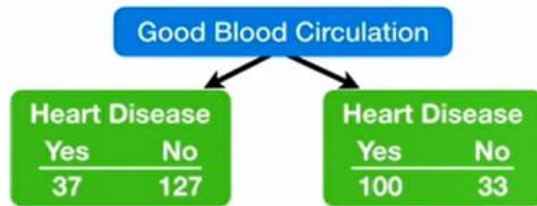
Gini impurity for Chest Pain = 0.364

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)...

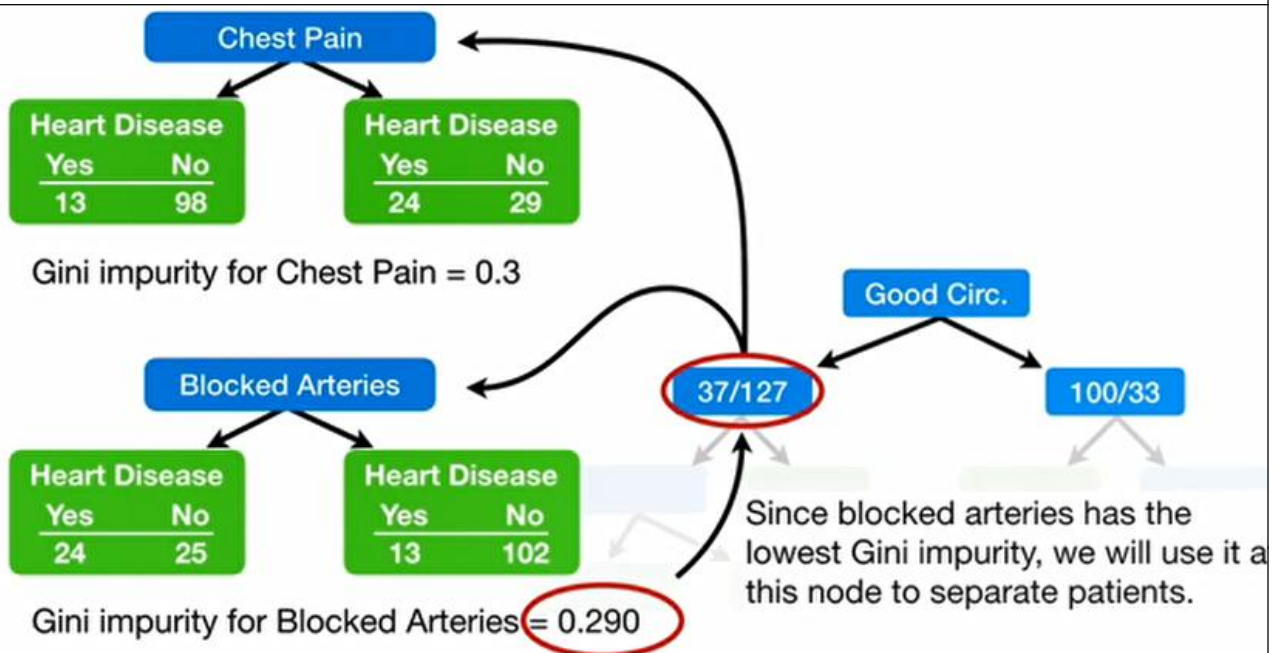
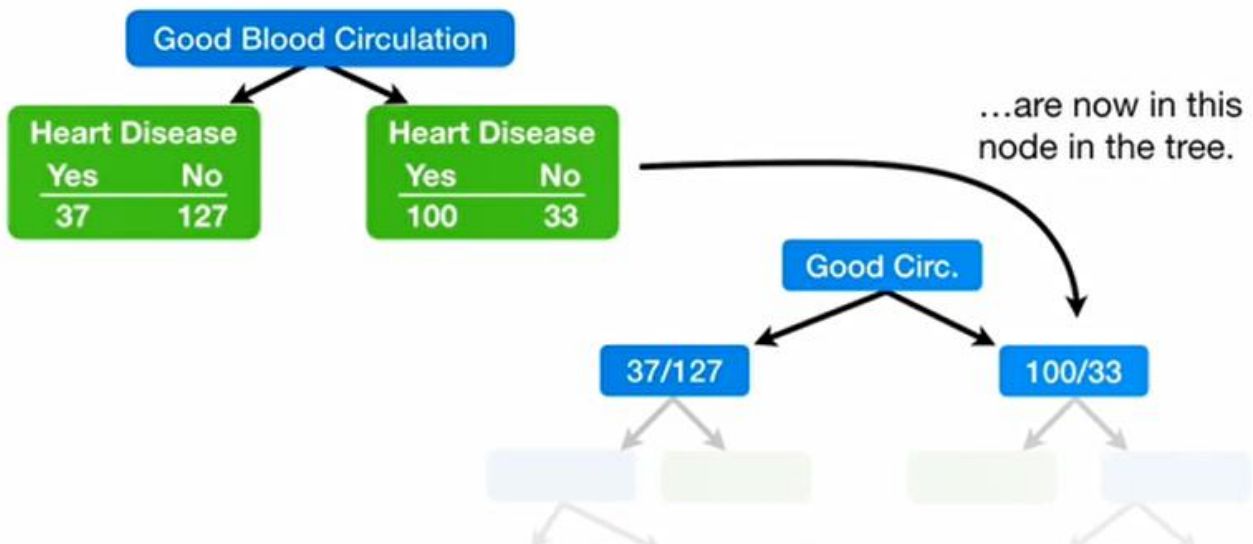
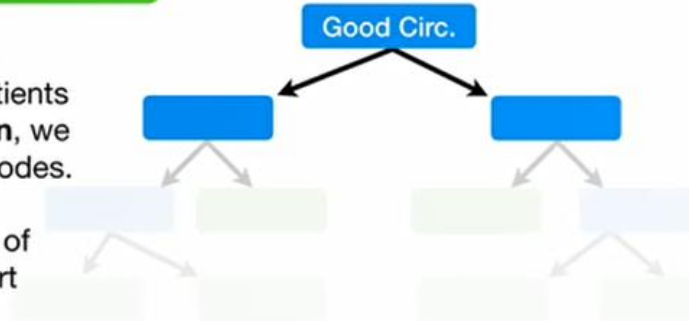
Gini impurity for Blocked Arteries = 0.381

DECISION TREE



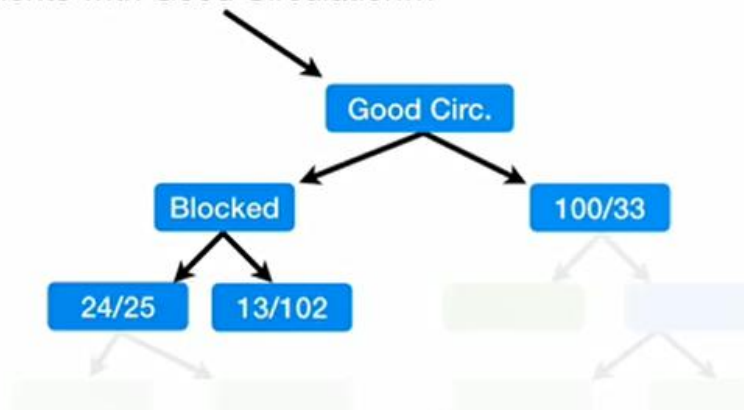
When we divided all of the patients using **Good Blood Circulation**, we ended up with “impure” leaf nodes.

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

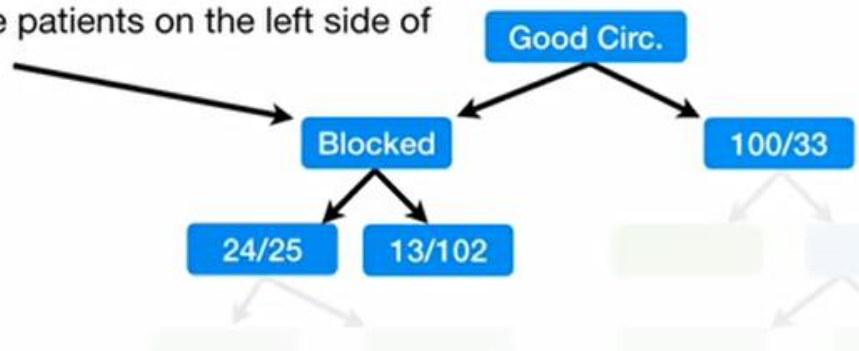


DECISION TREE

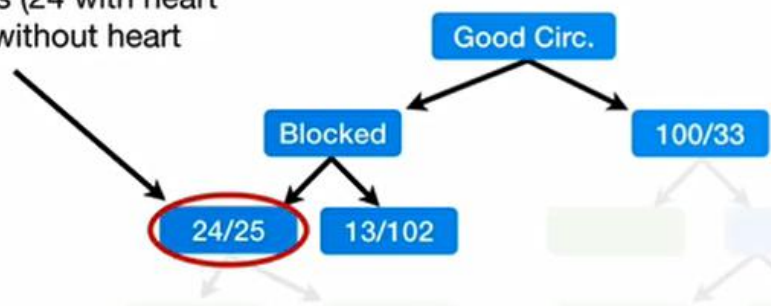
We started at the top by separating patients with Good Circulation...



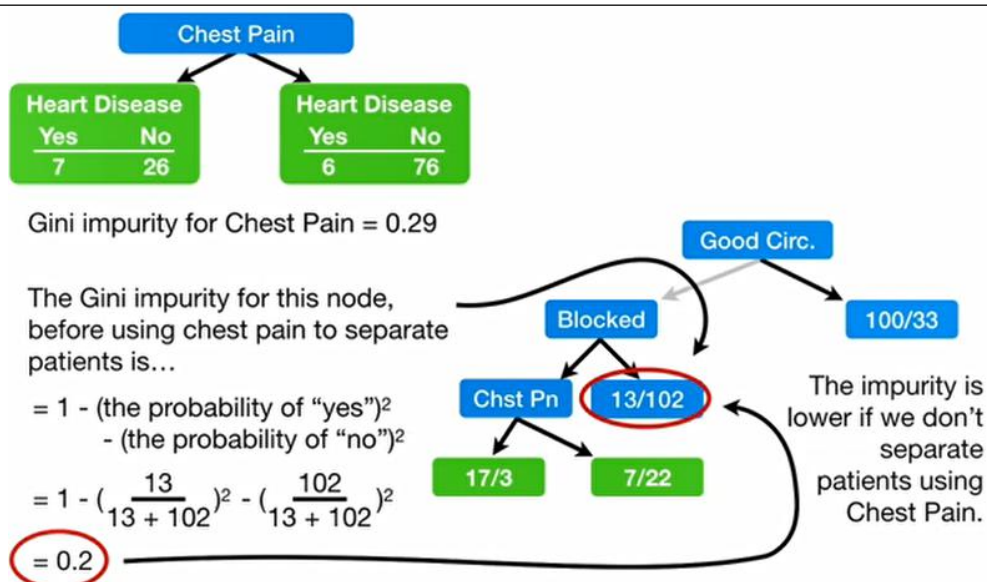
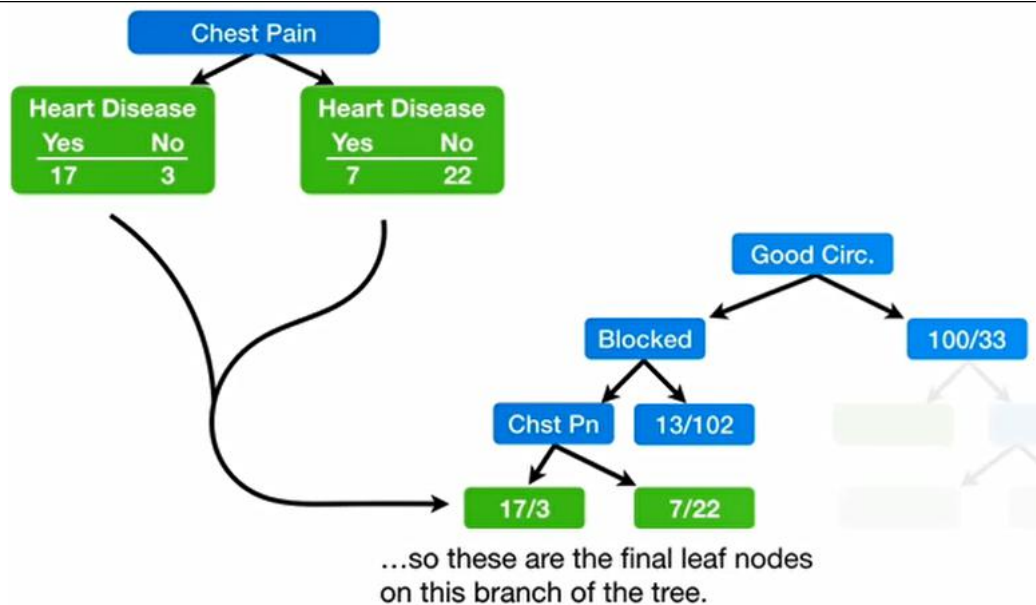
...then we used Blocked Arteries to separate patients on the left side of the tree.



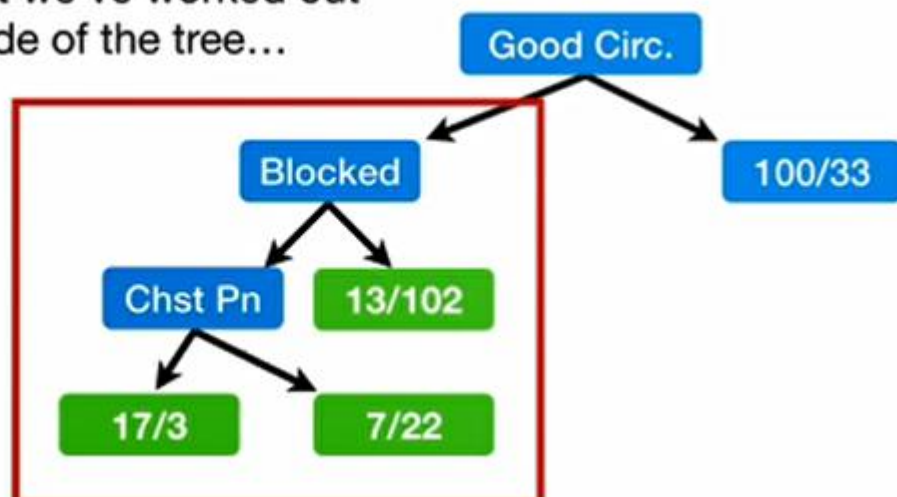
All we have left is Chest Pain, so first we'll see how well it separates these 49 patients (24 with heart disease and 25 without heart disease).



DECISION TREE



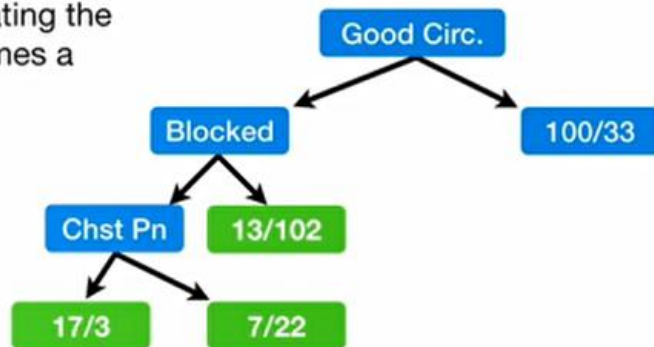
OK, at this point we've worked out the entire left side of the tree...



DECISION TREE

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

- 1) Calculate all of the Gini impurity scores.
- 2) If the node itself has the lowest score, than there is no point in separating the patients any more and it becomes a leaf node.
- 3) If separating the data results in an improvement, than pick the separation with the lowest impurity value.



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

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




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

Imagine if this were our data...

DECISION TREE

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



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

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



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

DECISION TREE

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

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

→ Gini impurity = ?

→ Gini impurity = ?

→ Gini impurity = ?

→ Gini impurity = ?

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

Gini impurity = 0

Weight < 167.5

Heart Disease	
Yes	No
0	1

Heart Disease	
Yes	No
3	1

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

→ Gini impurity = 0.47

→ Gini impurity = 0.27

→ Gini impurity = 0.4

The lowest impurity occurs when we separate using **weight < 205...**

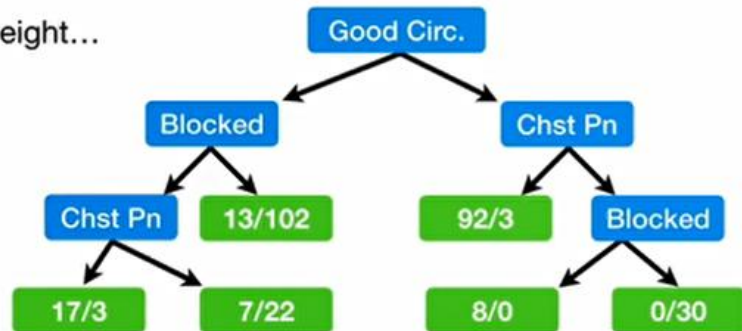
...so this is the cutoff and impurity value we will use when we compare weight to chest pain or blocked arteries.

DECISION TREE

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

1) "yes/no" questions at each step...

2) Numeric data, like patient weight...



OVERFITTING..... PRUNE!