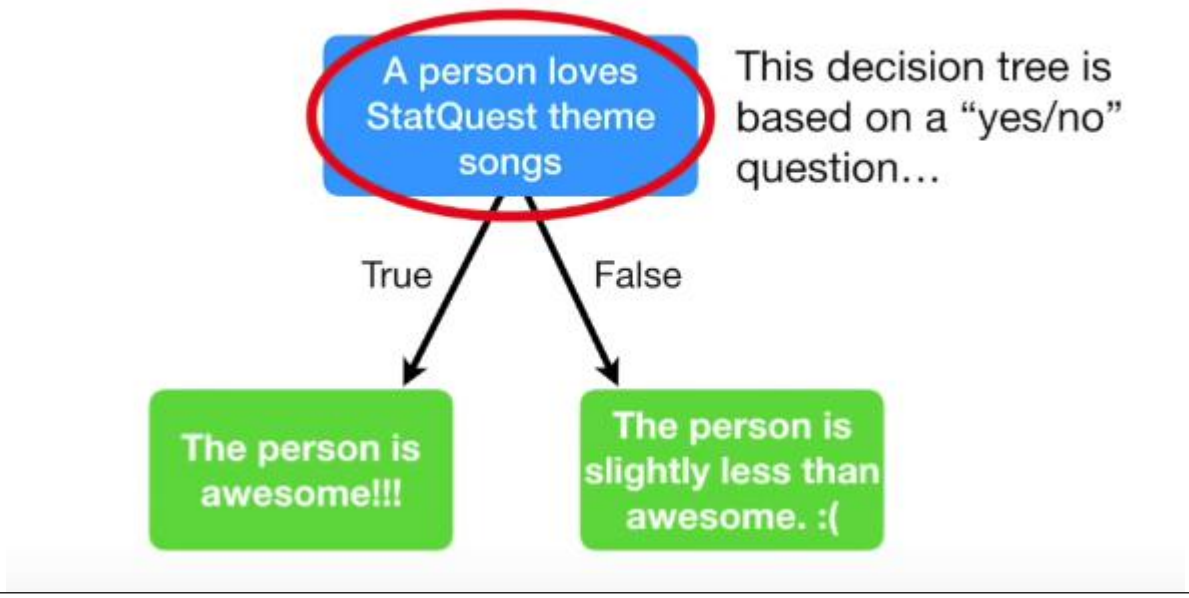
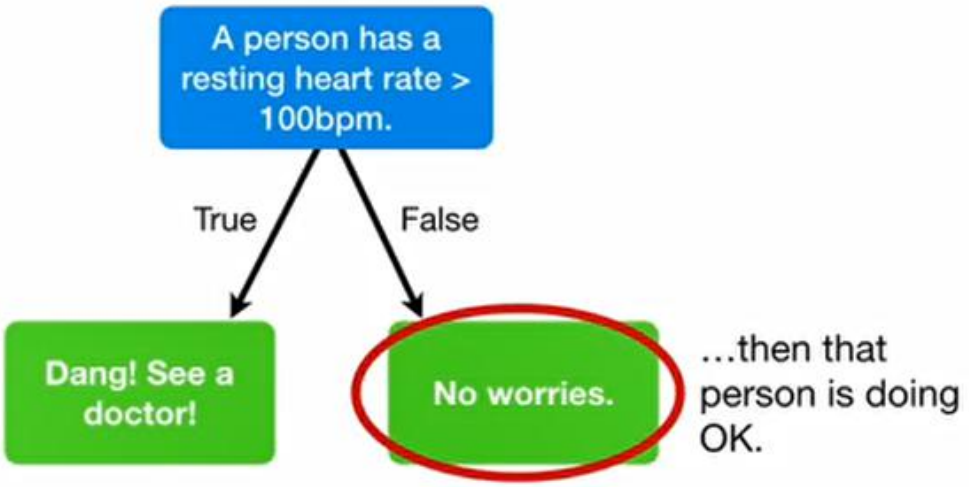
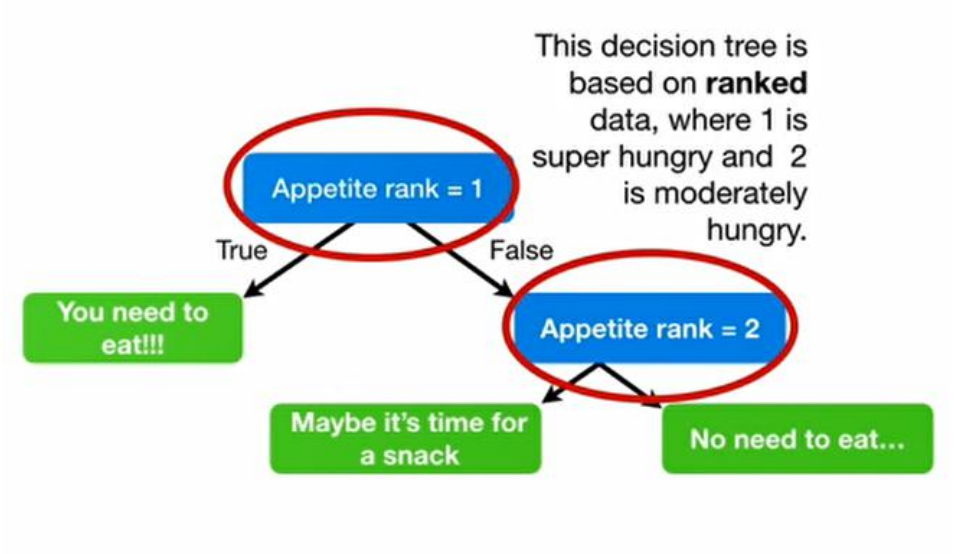
DECISION TREE

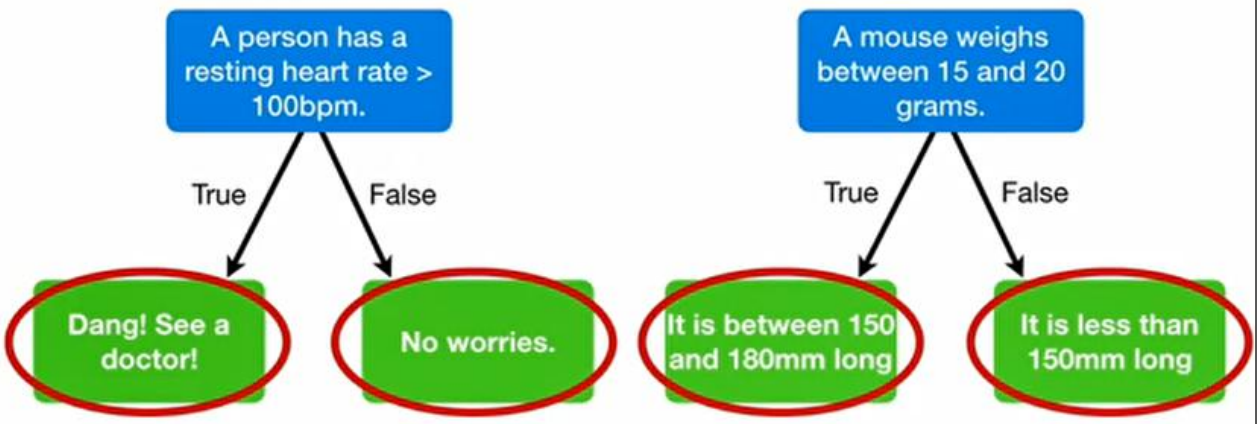
Big Data Lab
Decision tree



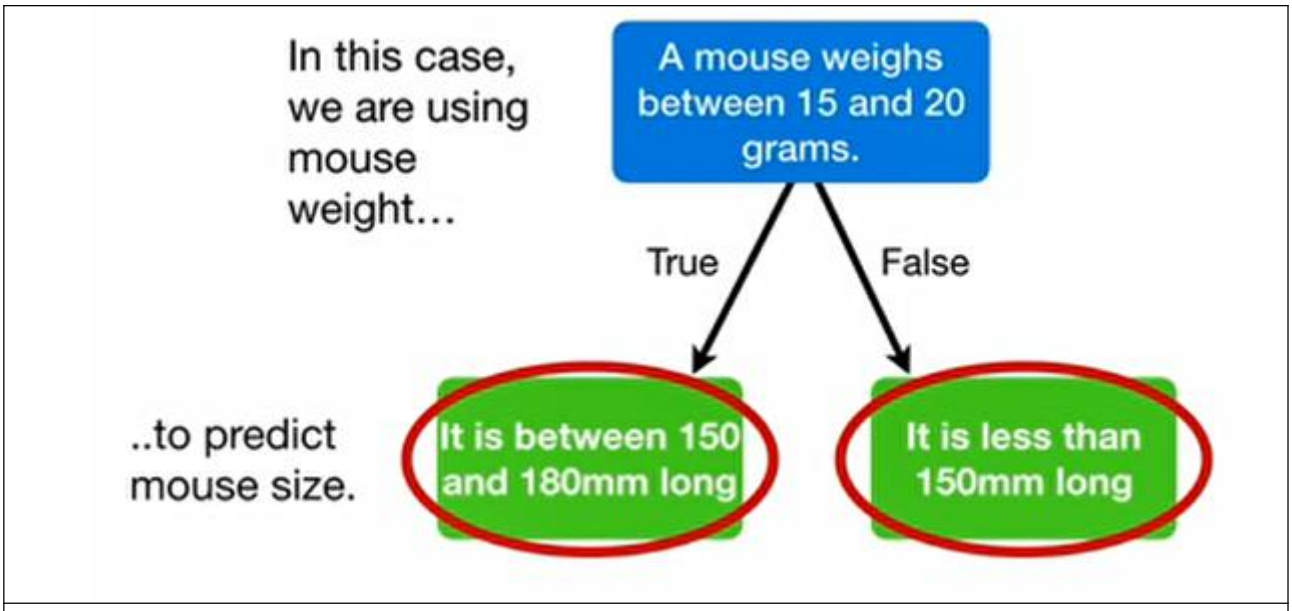
DECISION TREE

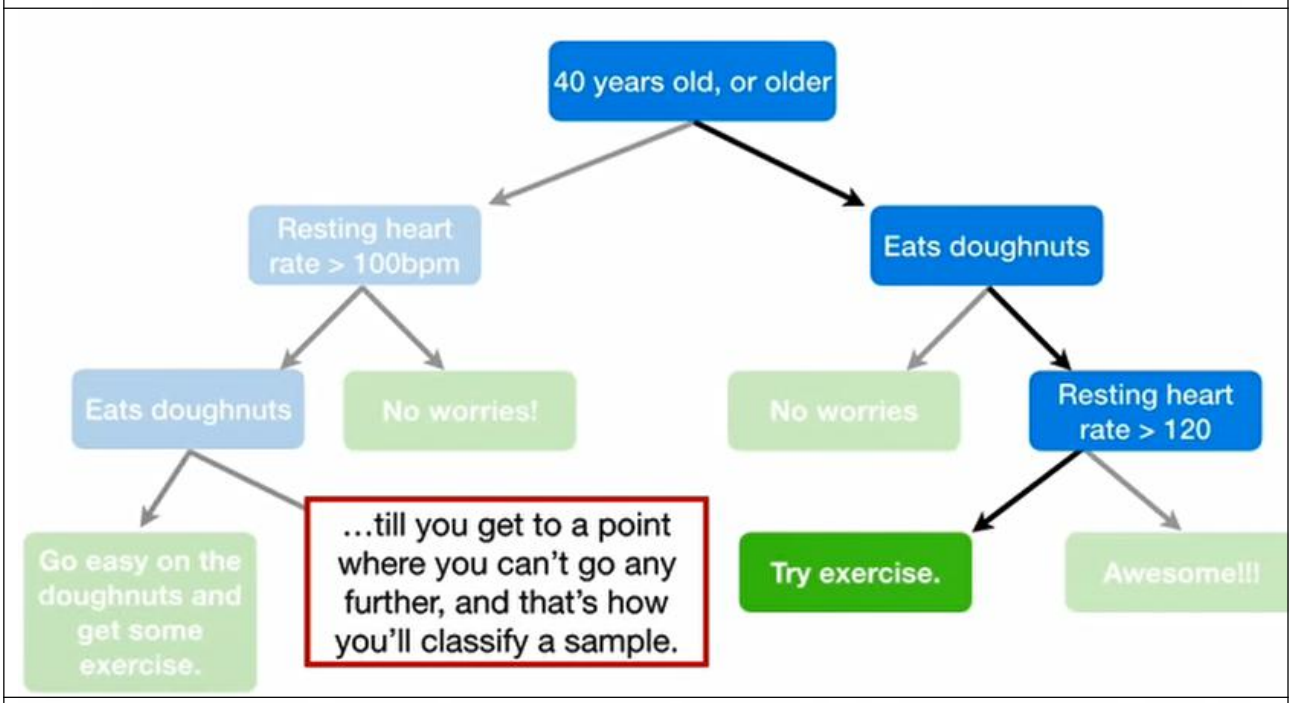


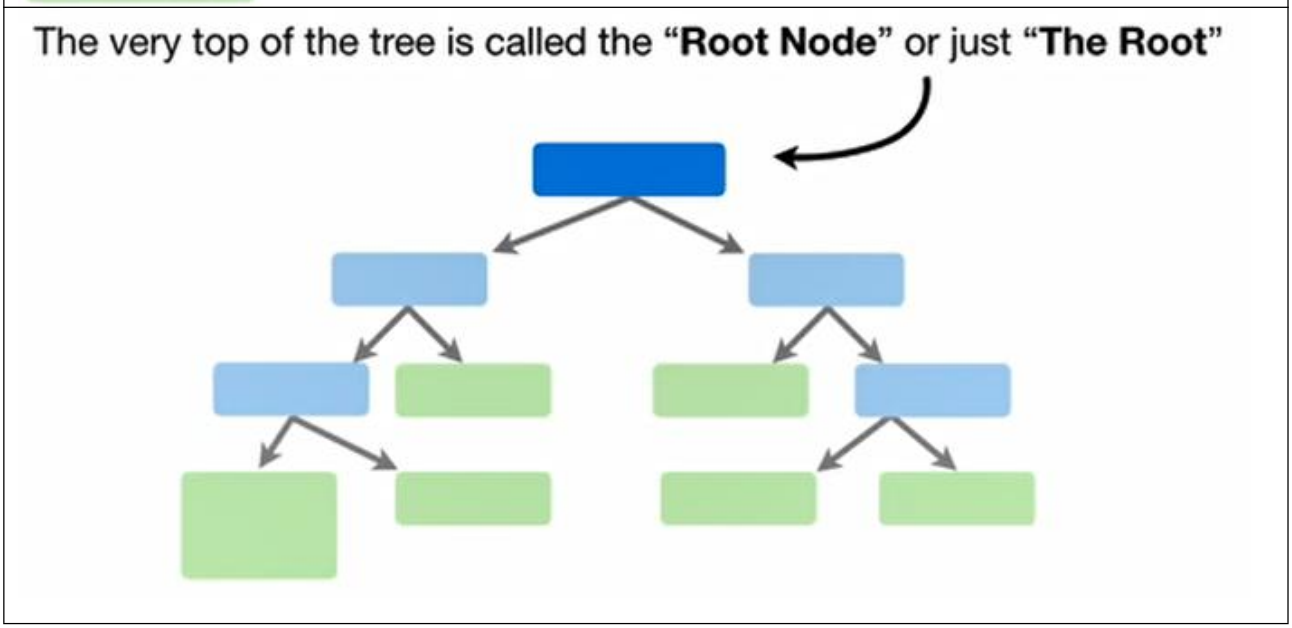




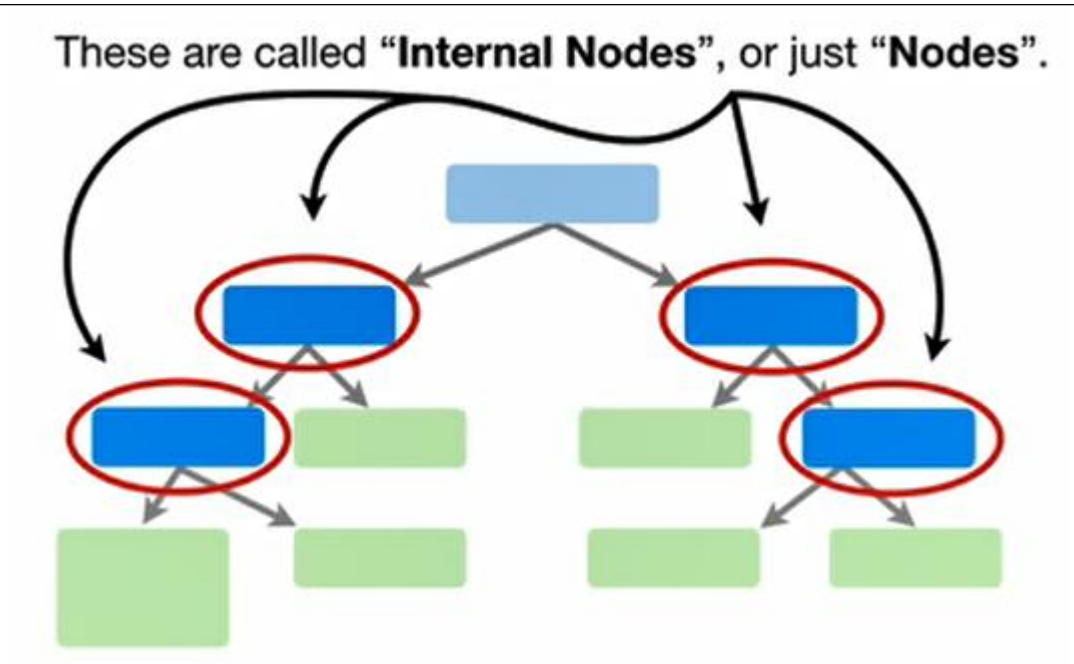
DECISION TREE

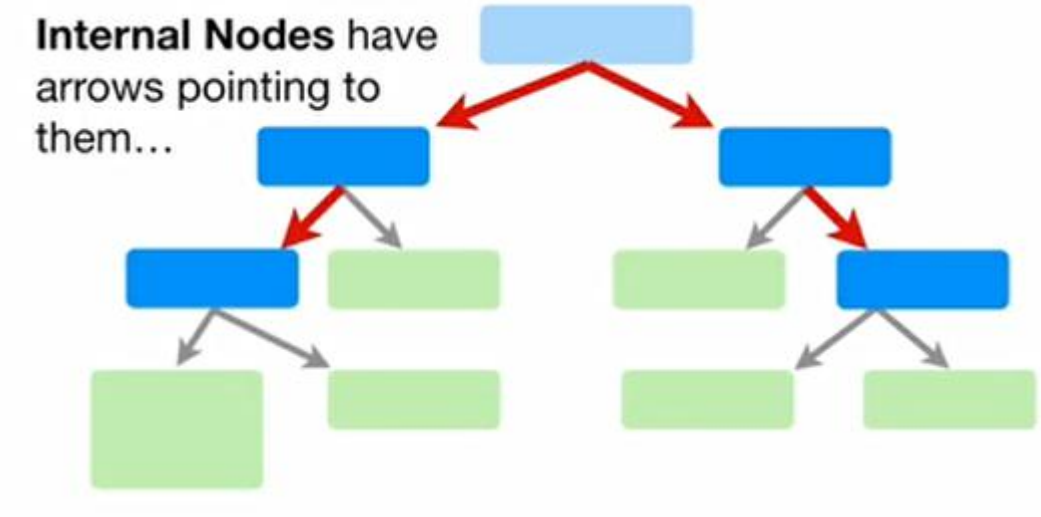


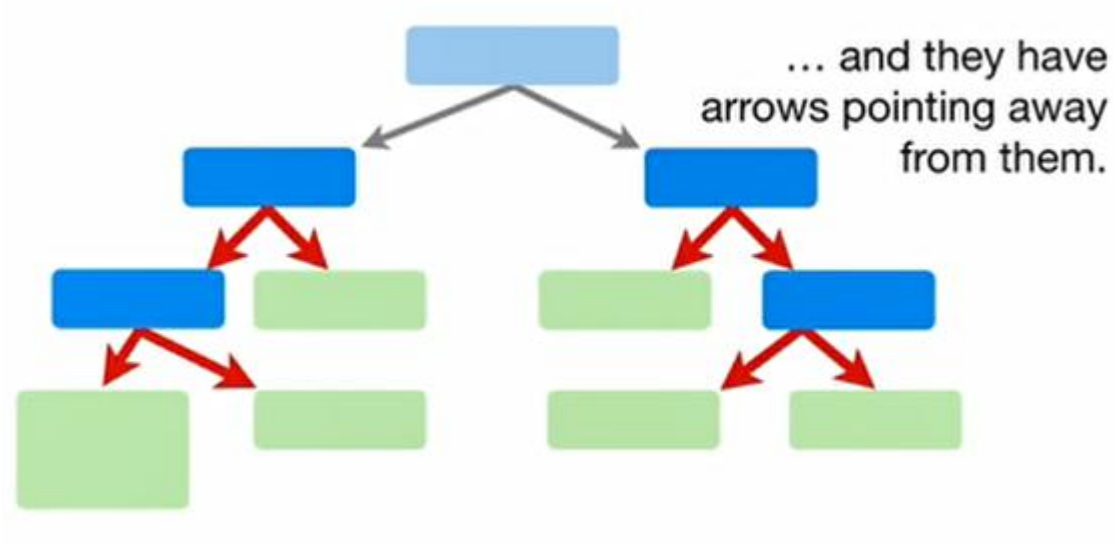




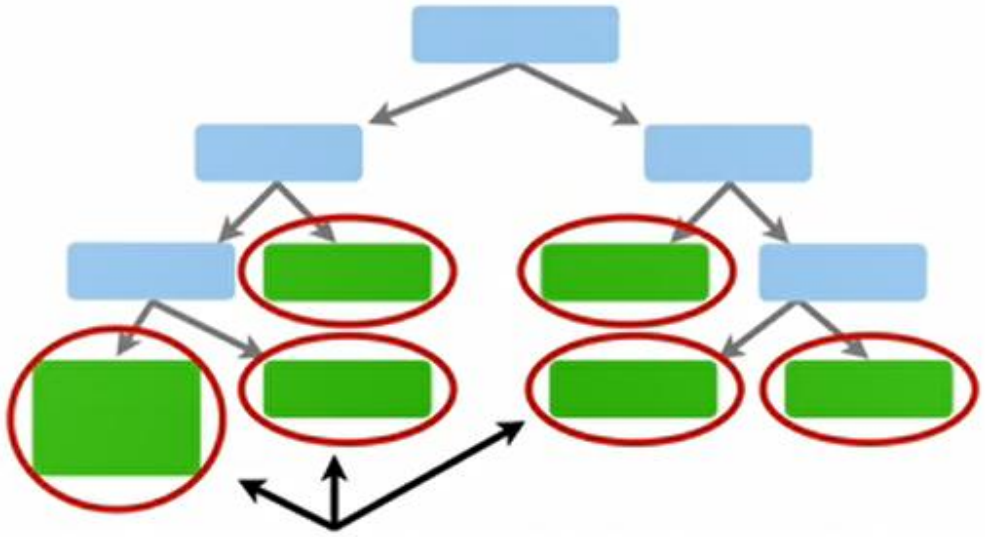
DECISION TREE



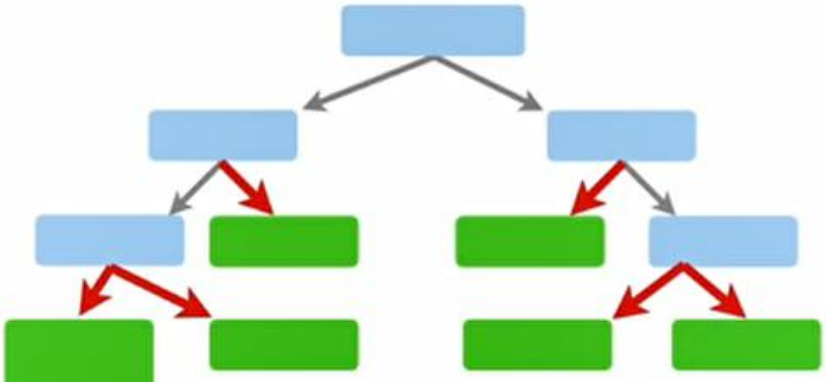




DECISION TREE



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



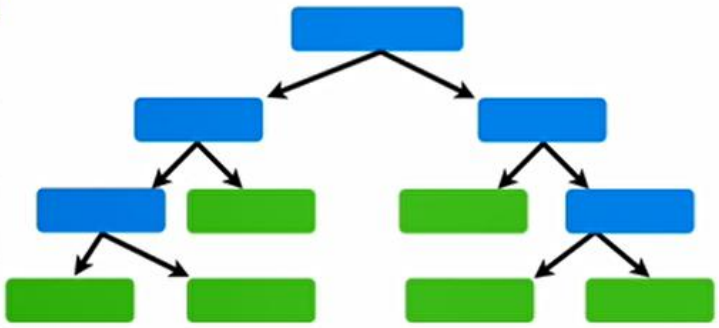
Leaf Nodes have arrows pointing to them…

…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…

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…

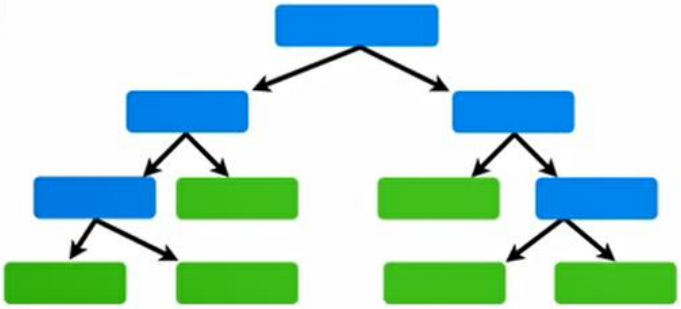
…to a decision tree!!!



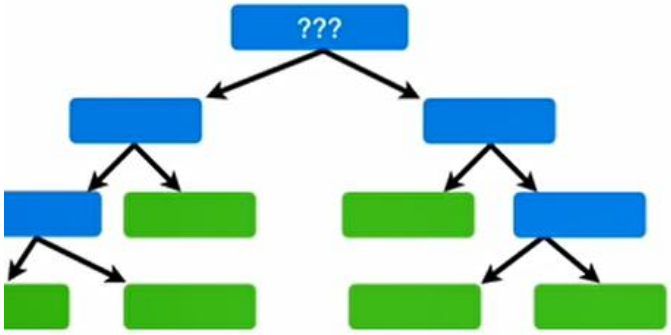
DECISION TREE

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…



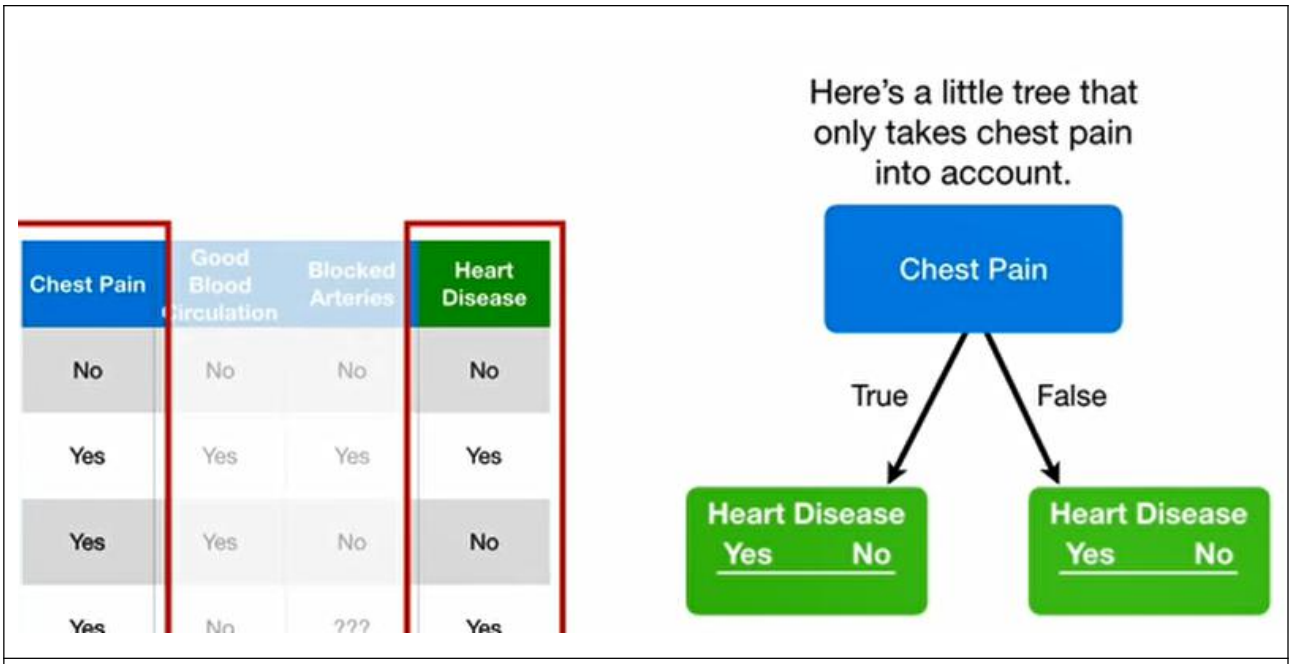
The first thing we want to know is whether **Chest Pain**, **Good Blood Circulation** or **Blocked Arteries** should be at the very top of our tree.

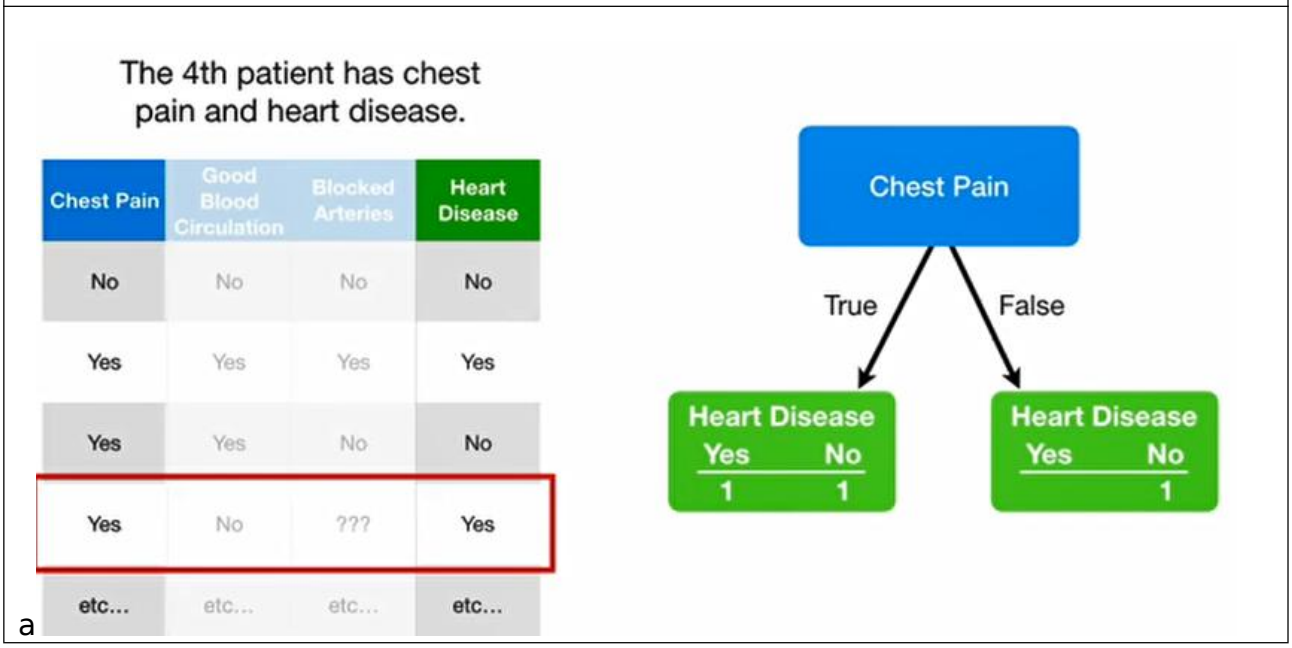


We start by looking at how well **Chest Pain** alone predicts heart disease…

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

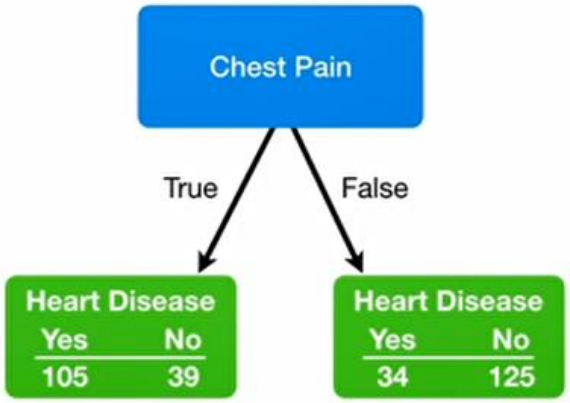




a

DECISION TREE

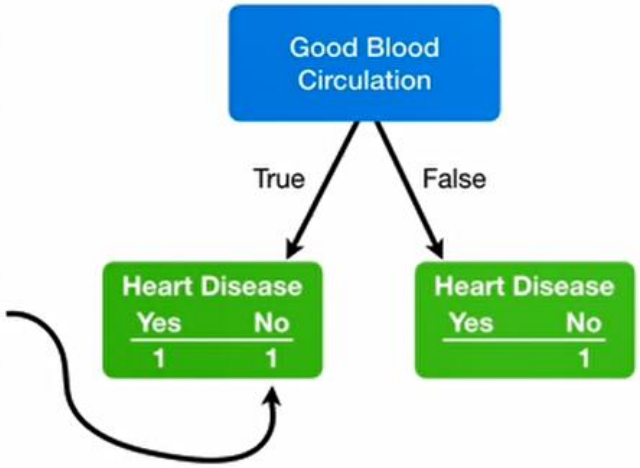
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...



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

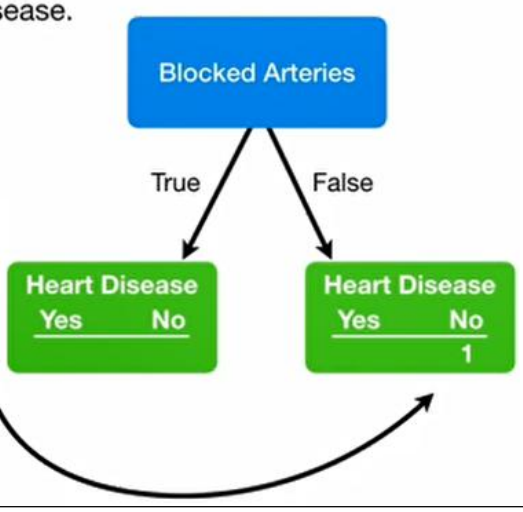
Now we do the exact same thing for **Good Blood Circulation**.

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.

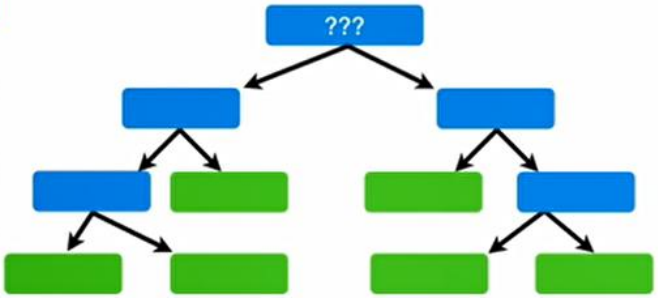
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...

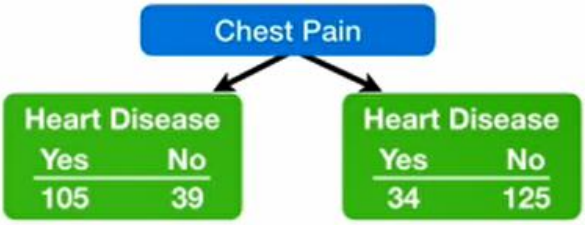


DECISION TREE

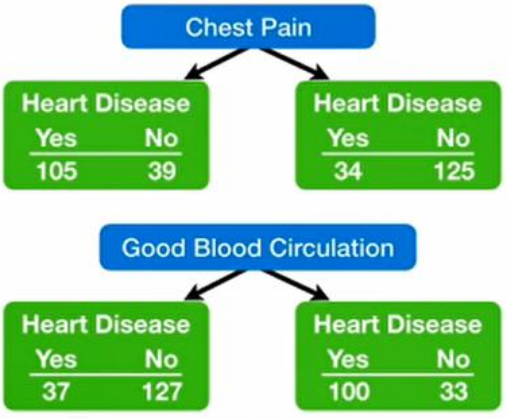
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...

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**).





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

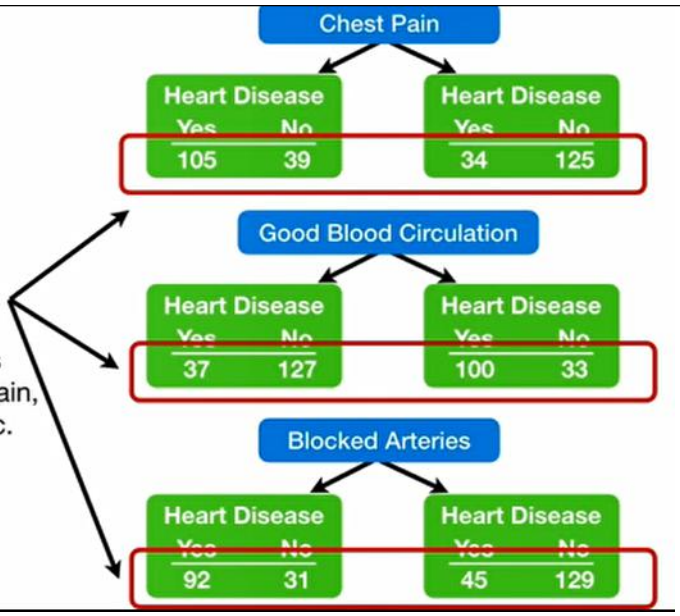


Then we looked at how well **Good Blood Circulation** separated patients with and without heart disease.

It wasn't perfect either.

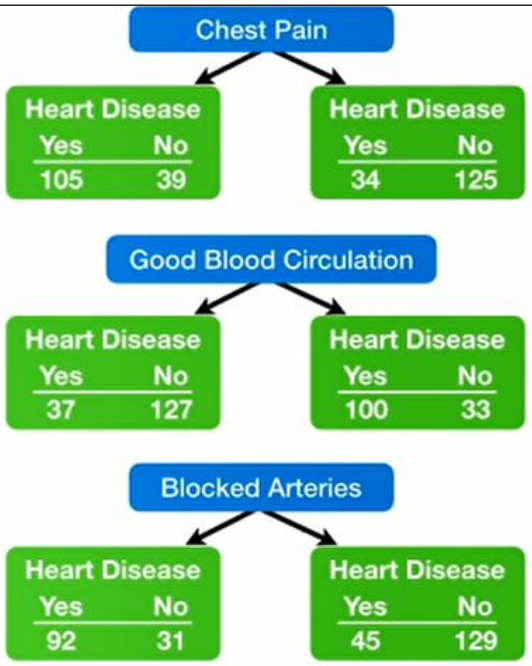
DECISION TREE

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.

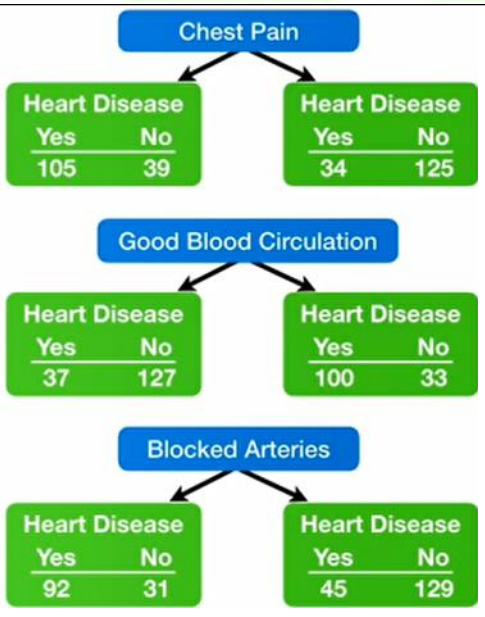


Because none of the leaf nodes are 100% "YES Heart Disease" or 100% "NO Heart Disease", they are all considered "**impure**".

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**".

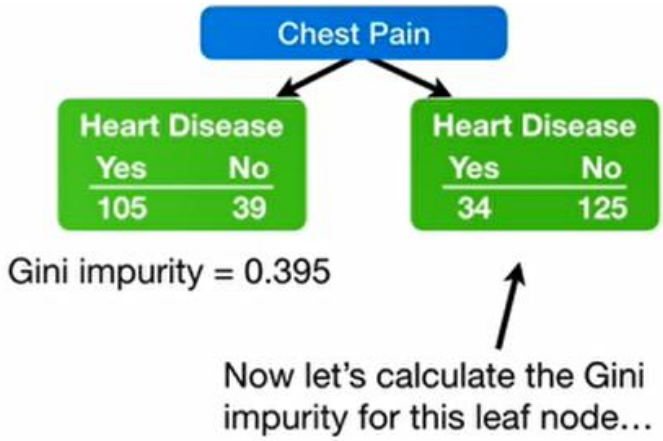


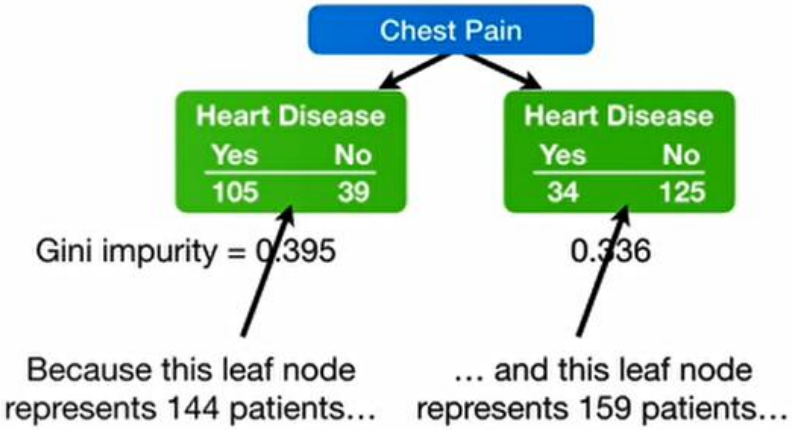
DECISION TREE

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

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

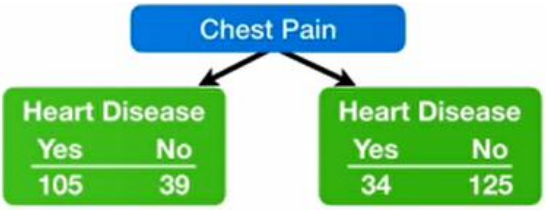
$$= 0.395$$





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



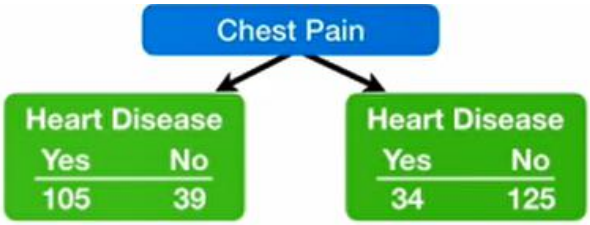
Gini impurity = 0.395 0.336

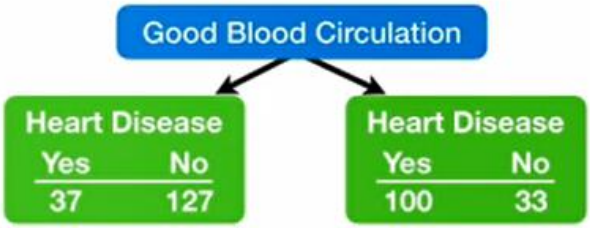
Gini impurity for Chest Pain = weighted average of Gini impurities for the leaf nodes

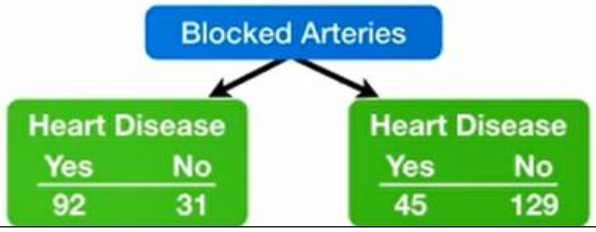
$$= \left(\frac{144}{144 + 159}\right) 0.395 + \left(\frac{159}{144 + 159}\right) 0.336$$

$$= 0.364$$

Gini impurity for Chest Pain = 0.364



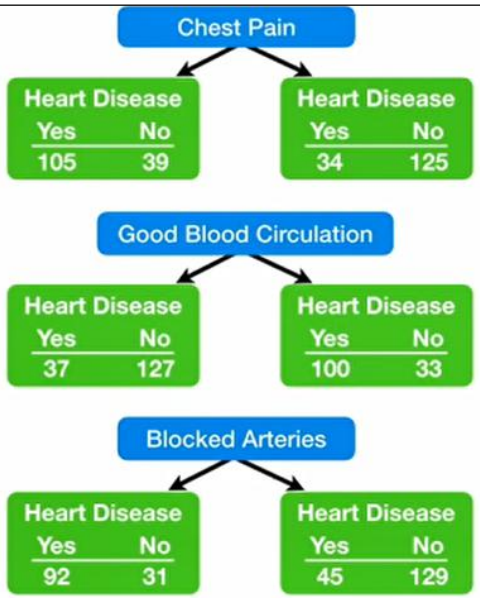




DECISION TREE

-..... 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

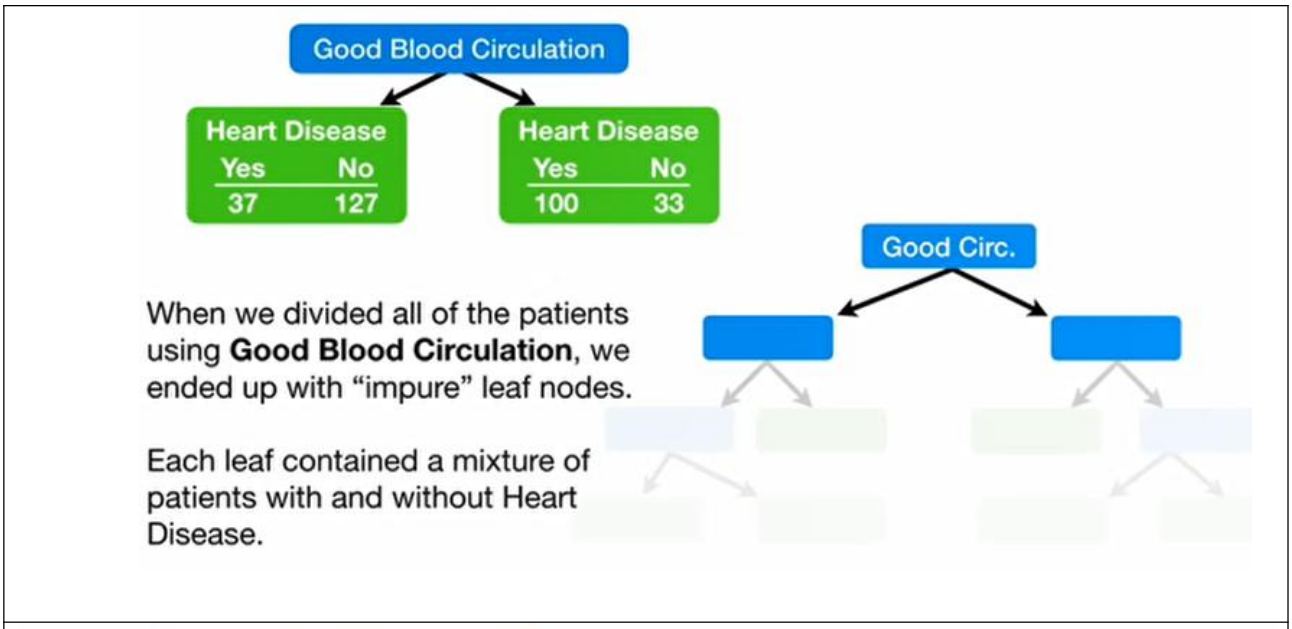
Gini impurity for Chest Pain = 0.364

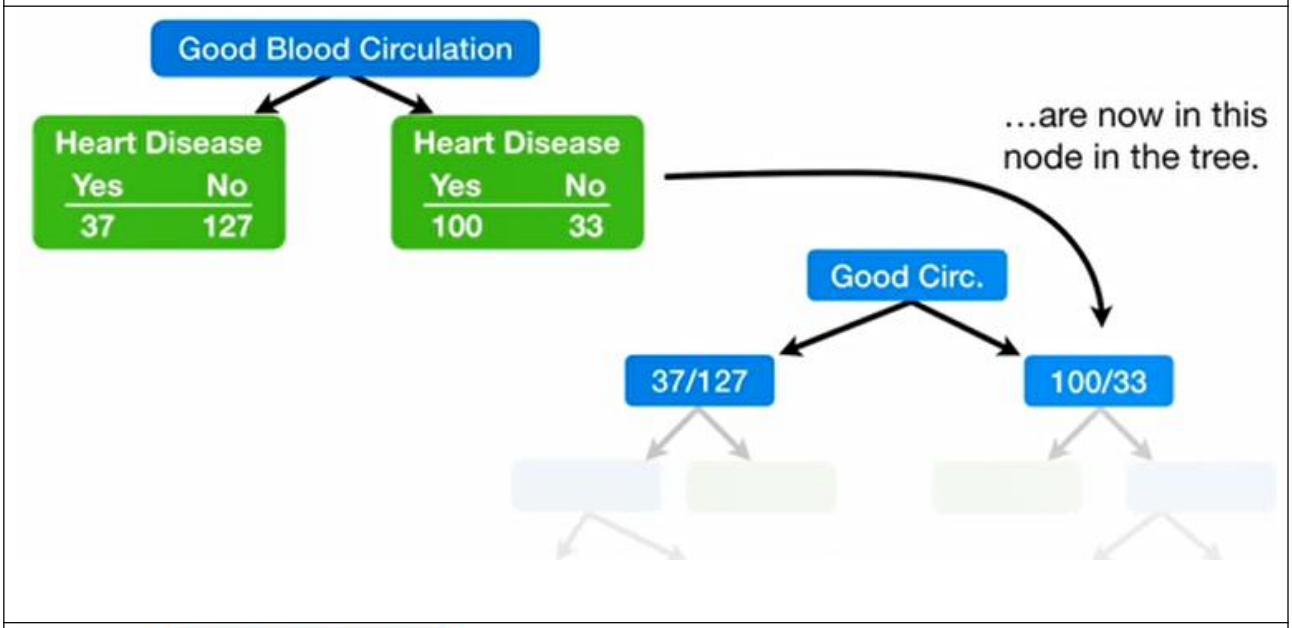
Gini impurity for Good Blood Circulation = 0.360

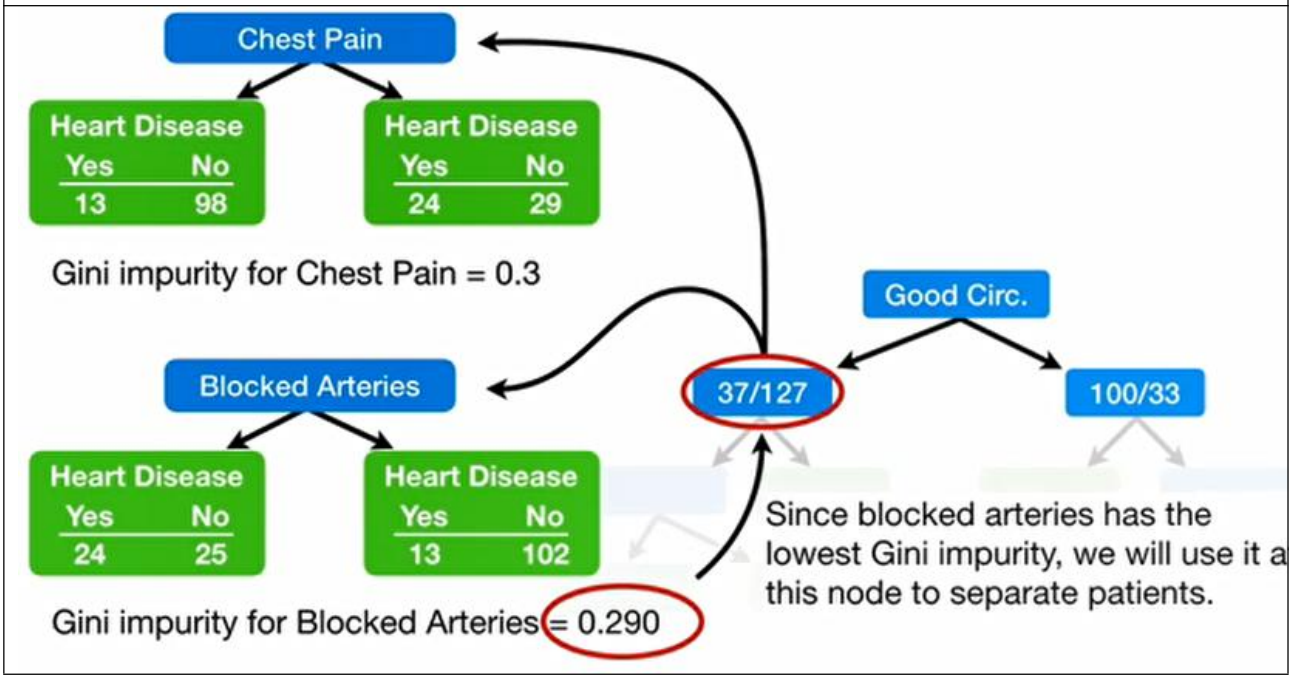
Good Blood Circulation has the lowest impurity (it separates patients with and without heart disease the best)...

Gini impurity for Blocked Arteries = 0.381

DECISION TREE

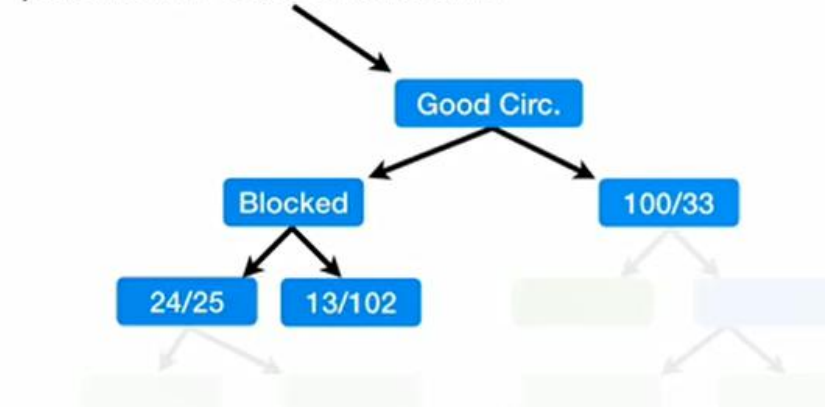




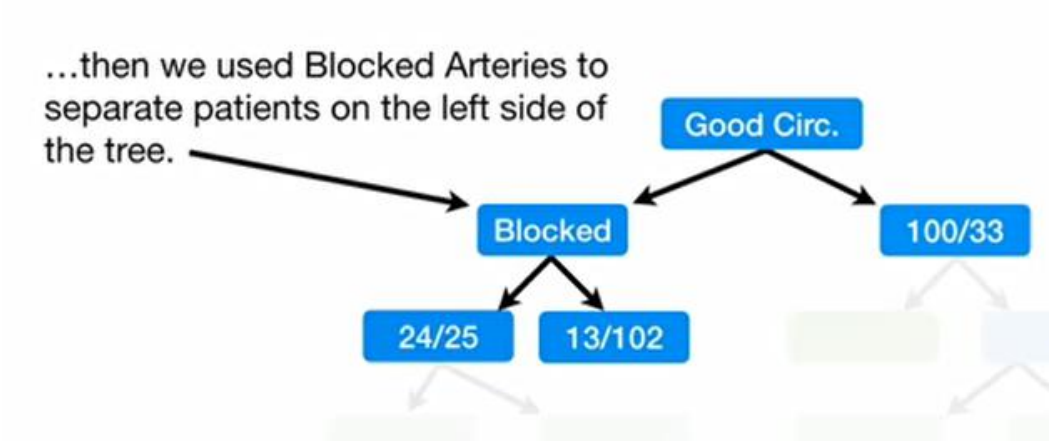


DECISION TREE

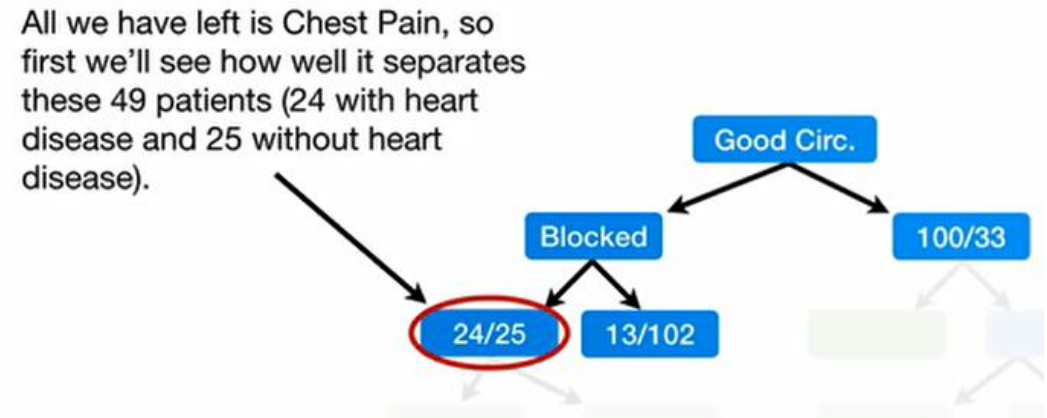
We started at the top by separating patients with Good Circulation…



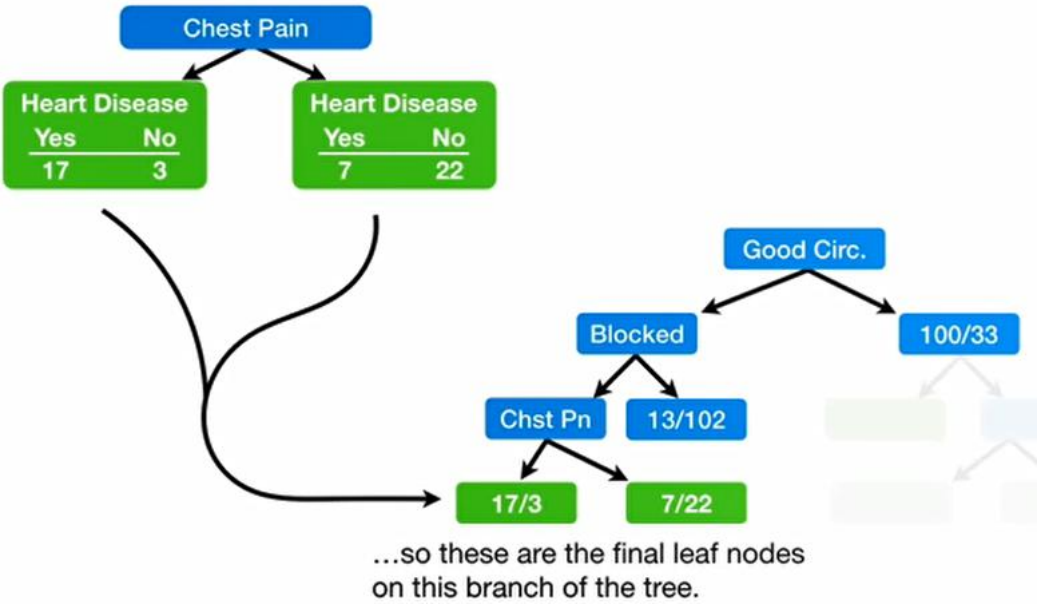
…then we used Blocked Arteries to separate patients on the left side of the tree.

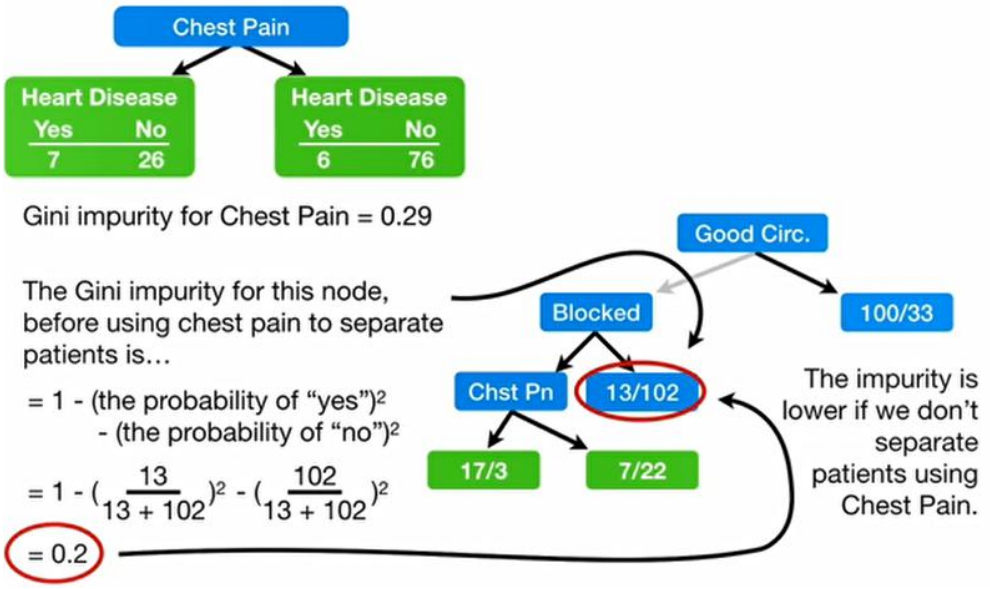


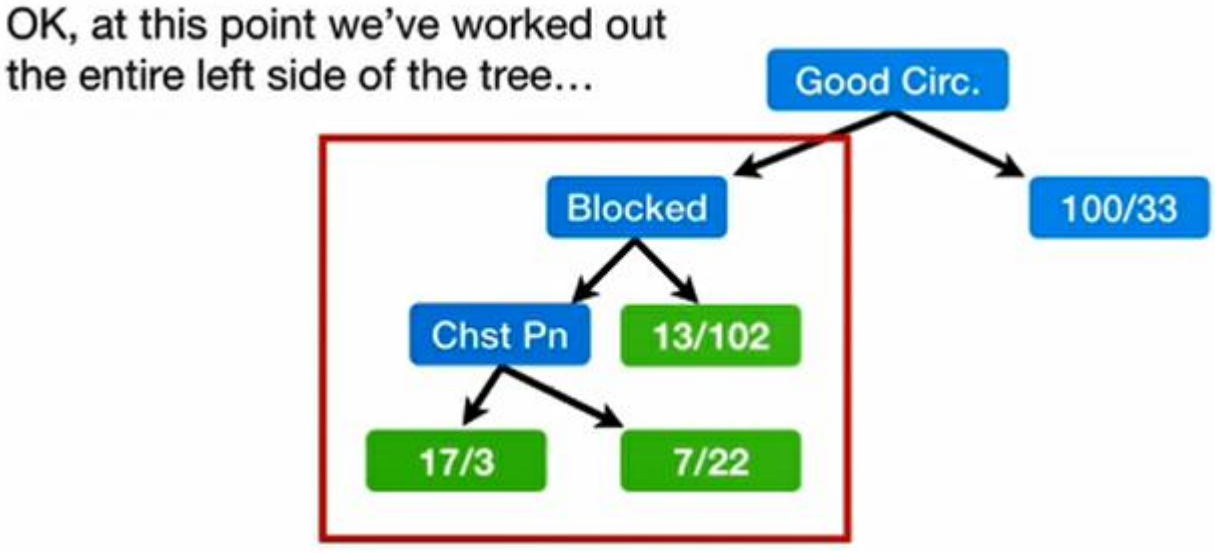
All we have left is Chest Pain, so first we'll see how well it separates these 49 patients (24 with heart disease and 25 without heart disease).



DECISION TREE







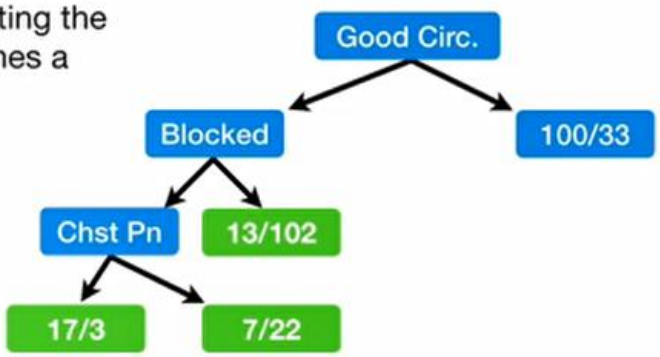
DECISION TREE

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.

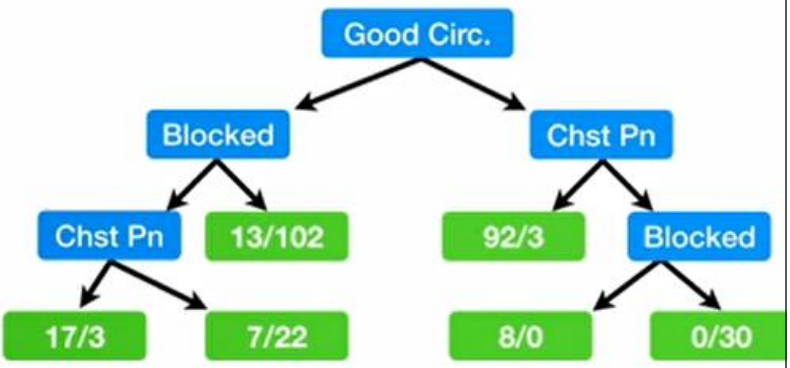
2) If the node itself has the lowest score, than there is no point in separating the patients any more and it becomes a leaf node.

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…

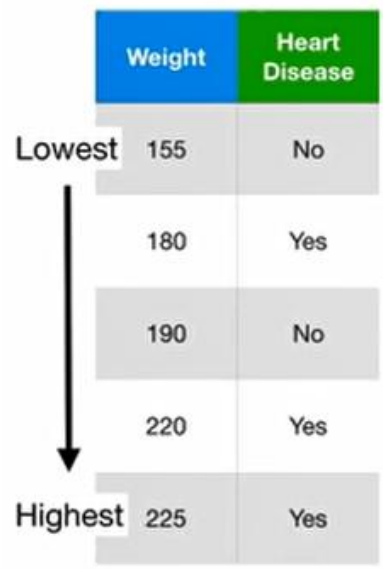
…but what if we have numeric data, like patient weight?



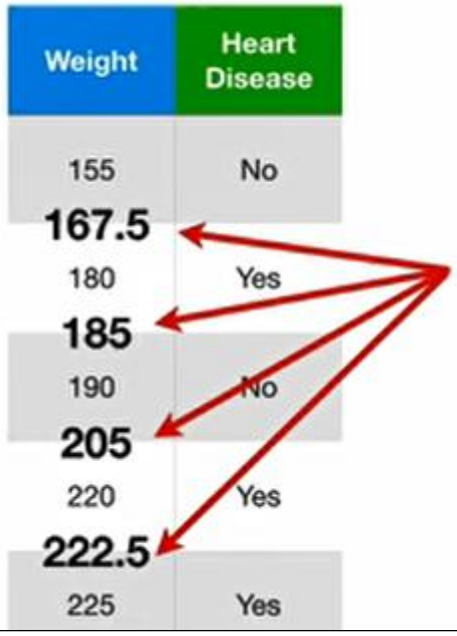
Weight	Heart Disease
220	Yes
180	Yes
225	Yes
190	No
155	No

Imagine if this were our data…

DECISION TREE

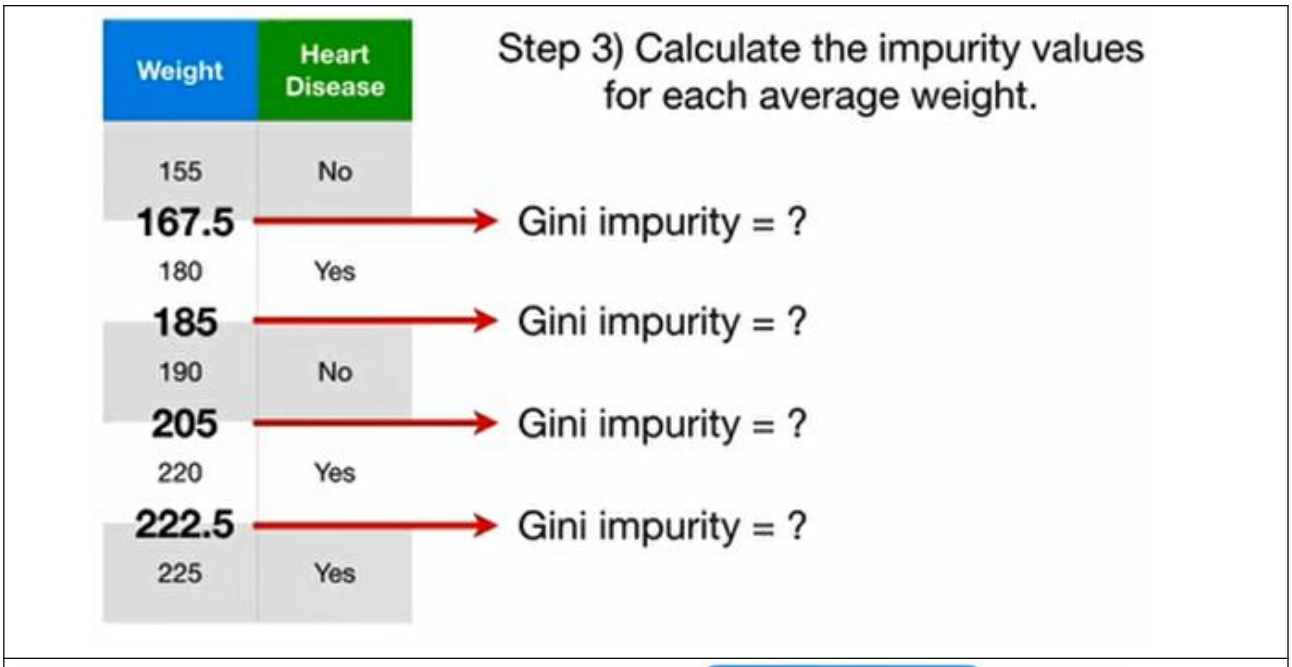


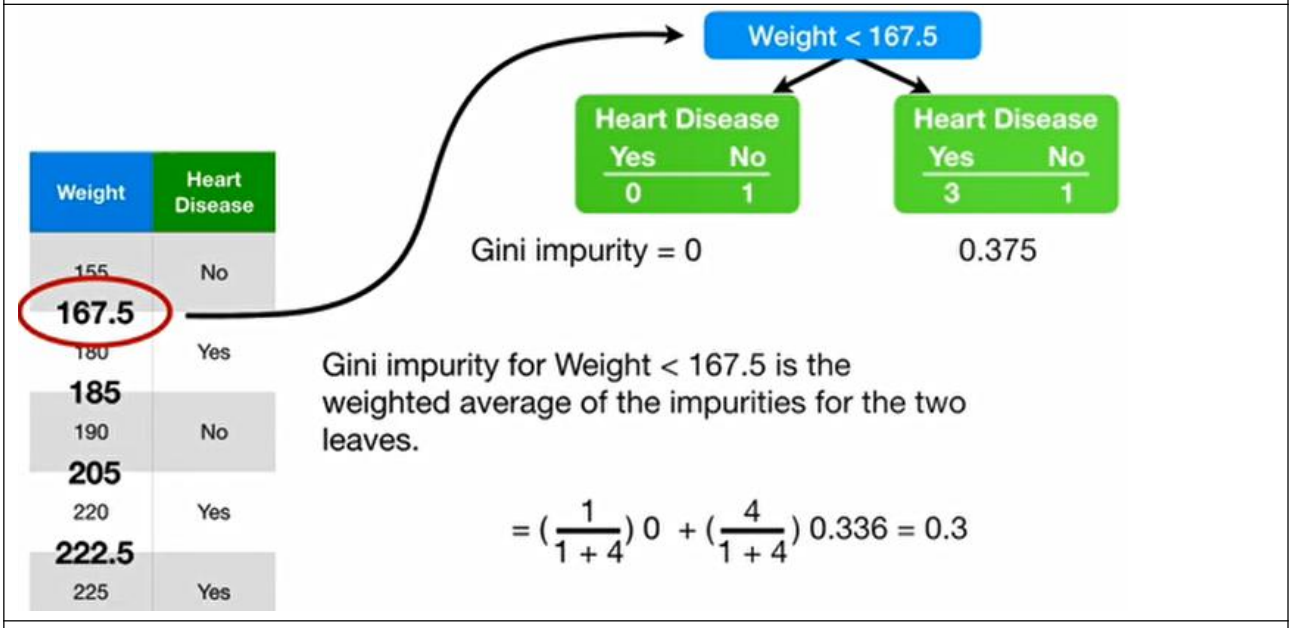
Step 1) Sort the patients by weight, lowest to highest.

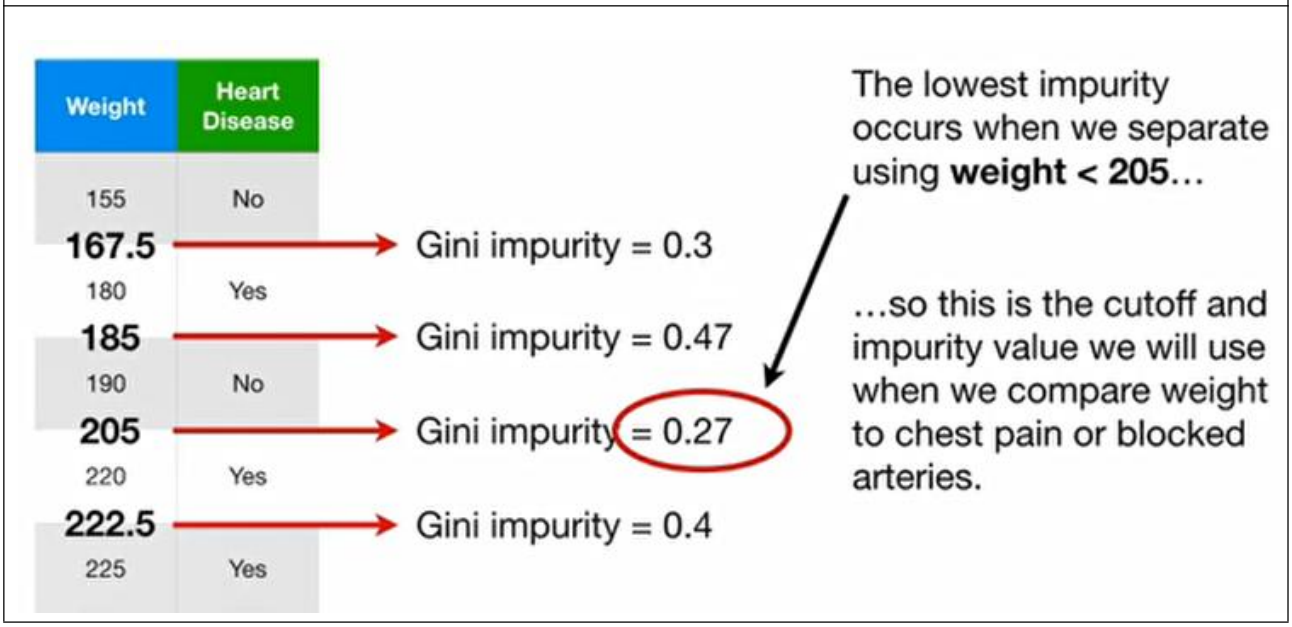


Step 2) Calculate the average weight for all adjacent patients.

DECISION TREE





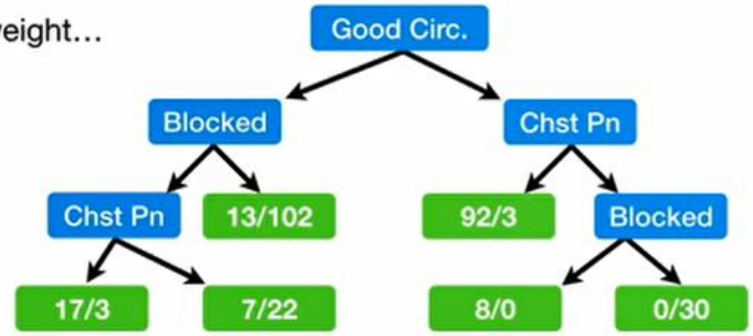


DECISION TREE

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!