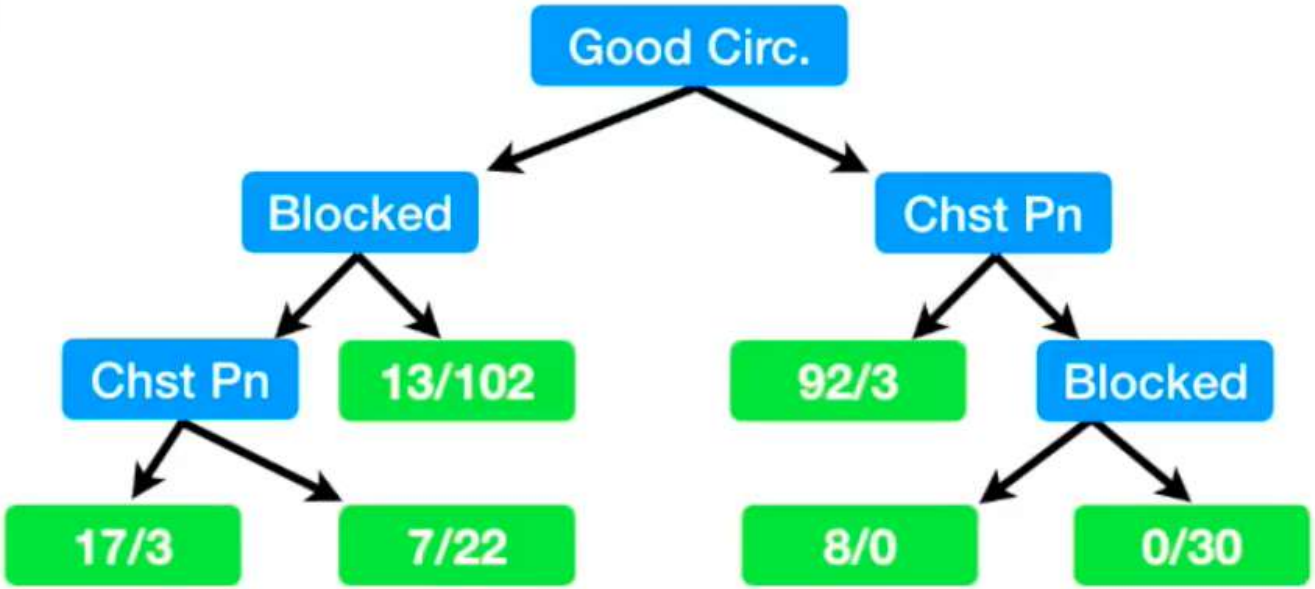
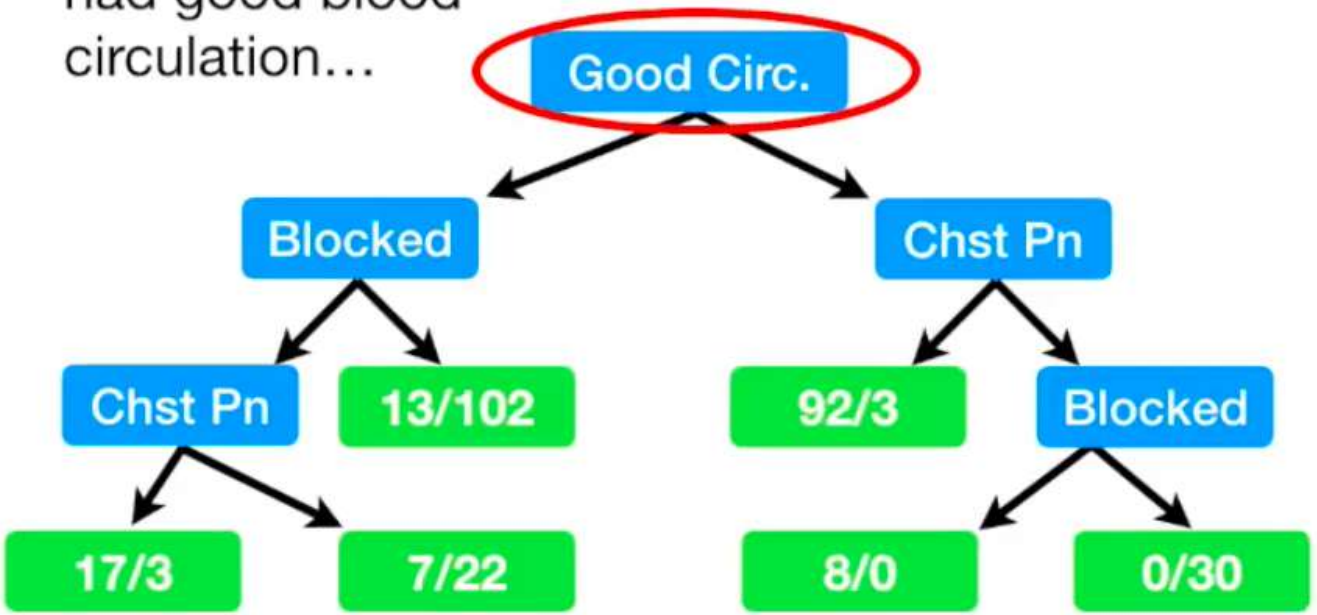
In the first StatQuest on Decision Trees, we started with a table of data…

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…

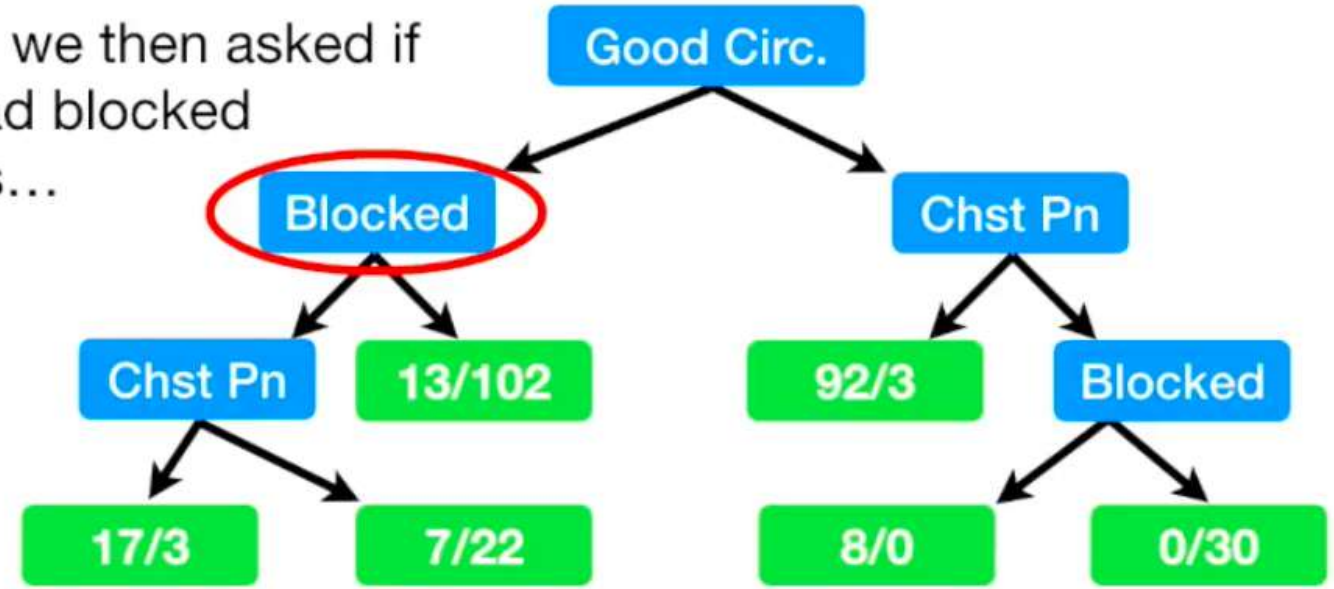
…and built a decision tree that gave us a sense of how likely a patient might have heart disease if they have other symptoms.



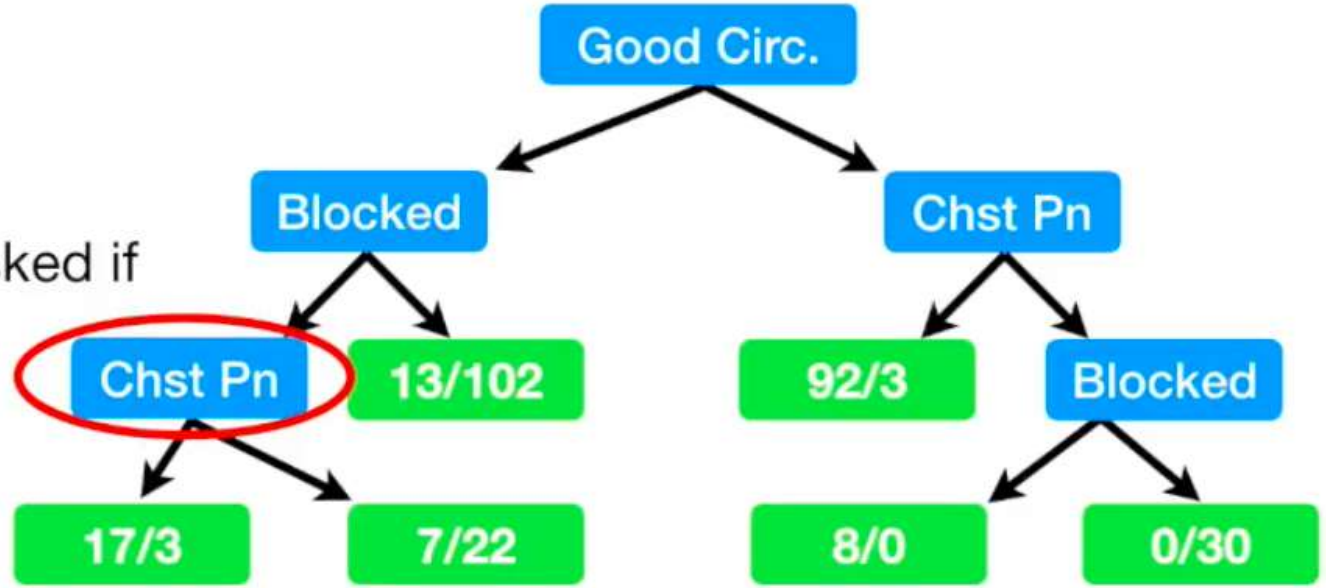
We first asked if a patient had good blood circulation…



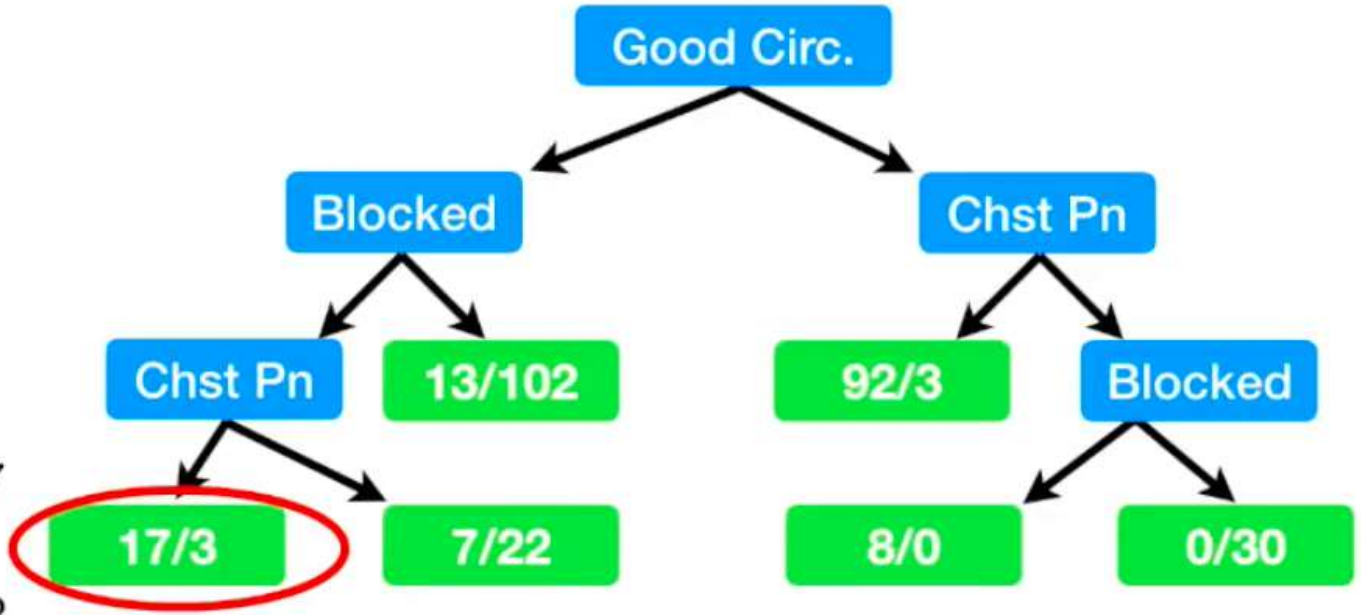
…if so, we then asked if they had blocked arteries…

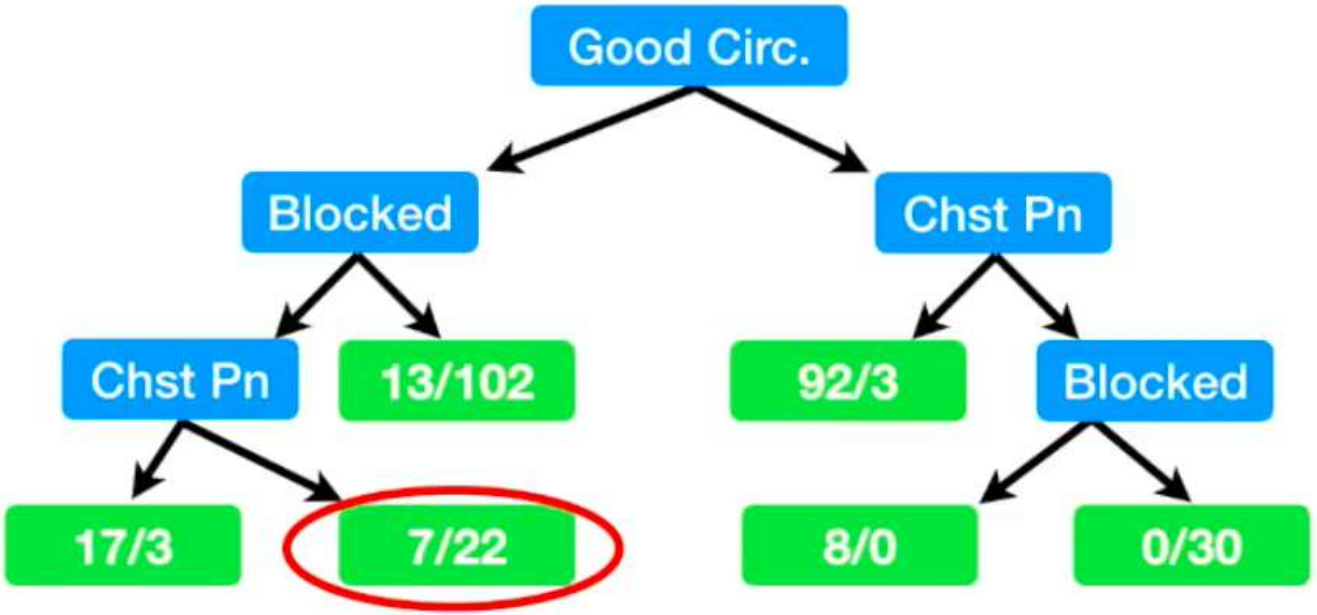


…if so, we then asked if they chest pain…



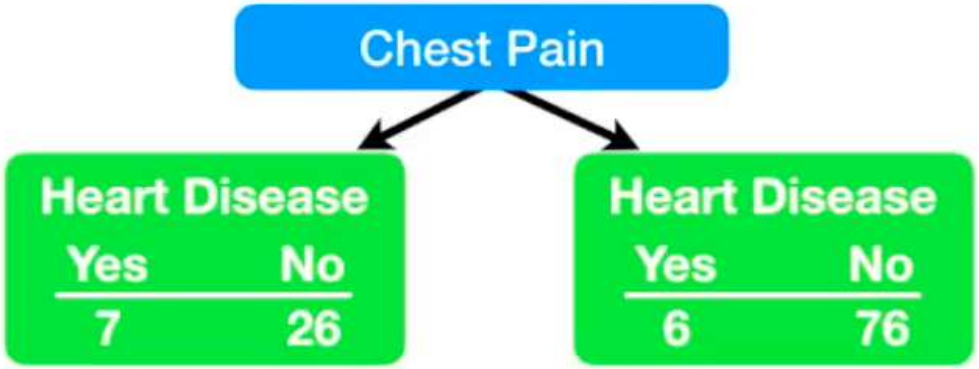
…if so, there's a good chance that they have heart disease, since 17 people with similar answers did and only 3 people with similar answers did not.





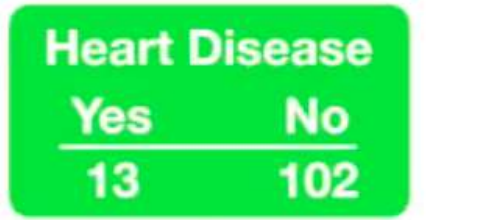
…if not, there's a good chance that they do not have heart disease.

However, remember that if someone had good circulation and did not not have blocked arteries, we did not ask about chest pain because there was less impurity in our results if we didn't.

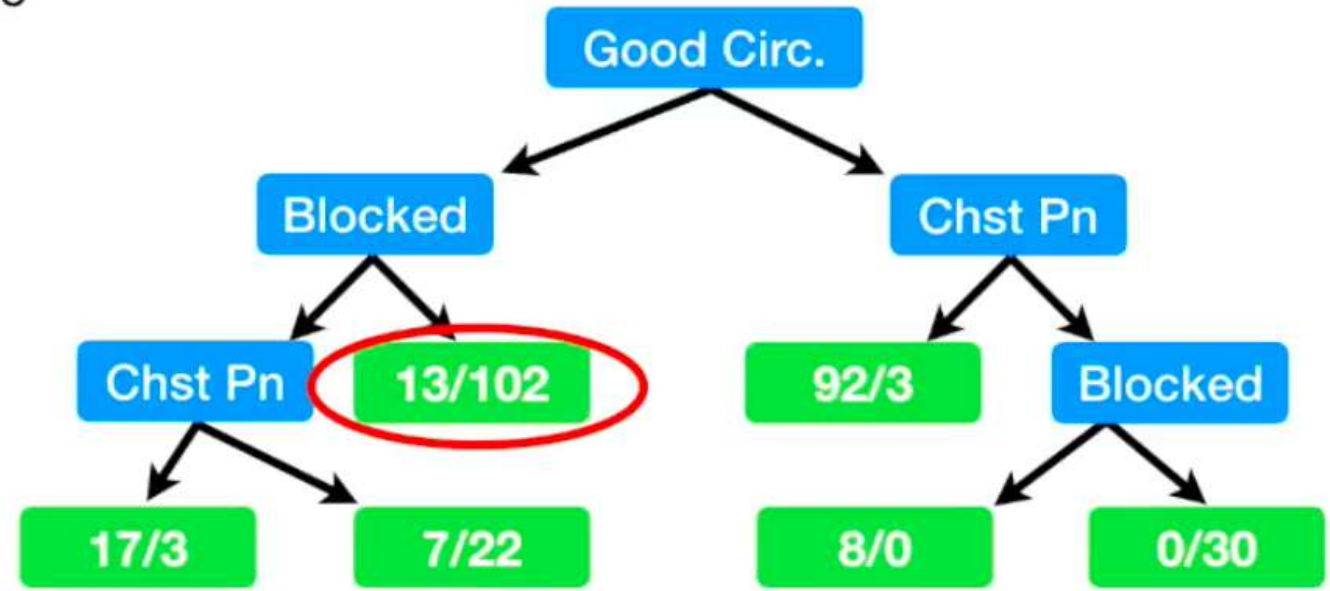


Gini impurity for Chest Pain = 0.29

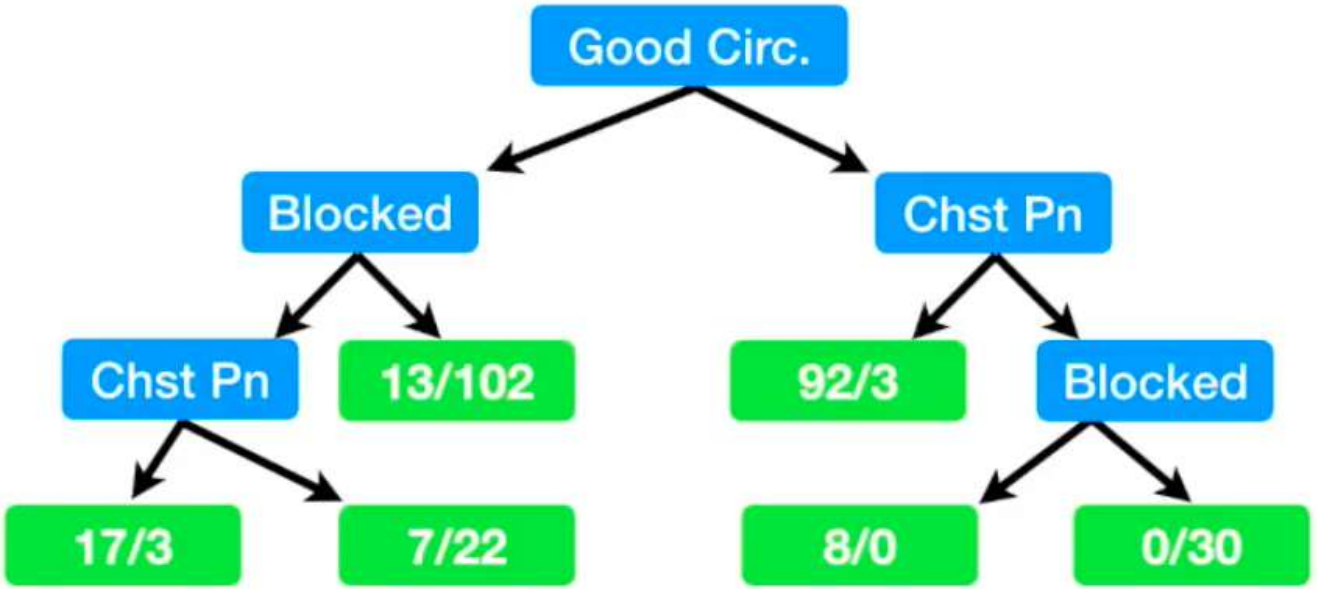
…and since the impurity was lower when we didn't separate, we made it a leaf node.



Gini impurity without separating = 0.20

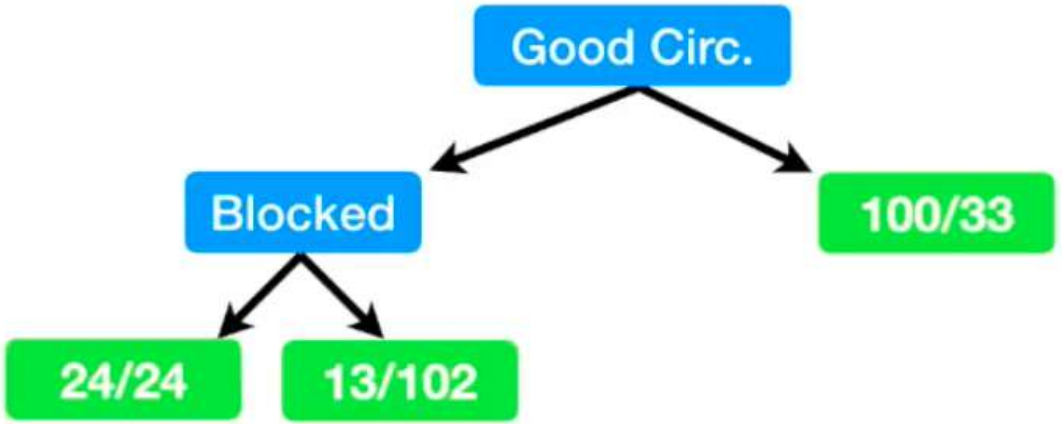


Now, imagine if Chest Pain never gave us a reduction in impurity score…

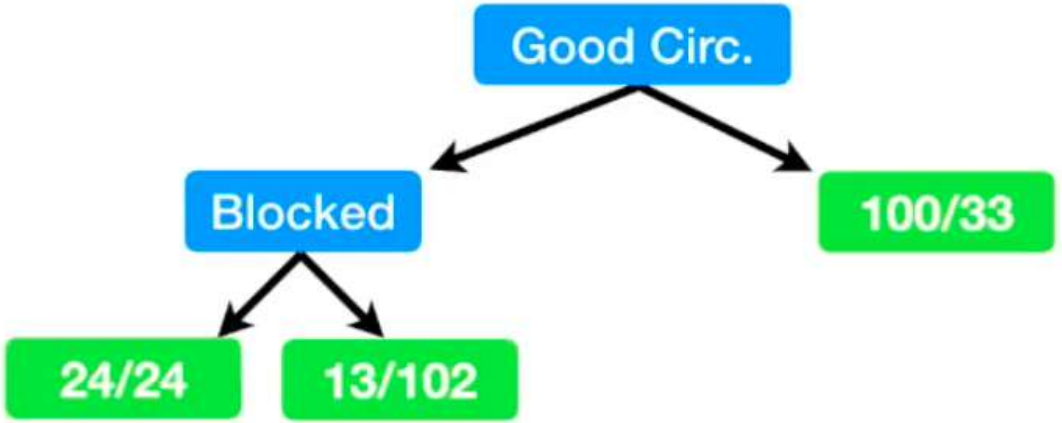


Now, imagine if Chest Pain never gave us a reduction in impurity score…

…if this were the case, we would never use Chest Pain to separate the patients, and Chest Pain would not be part of our tree.

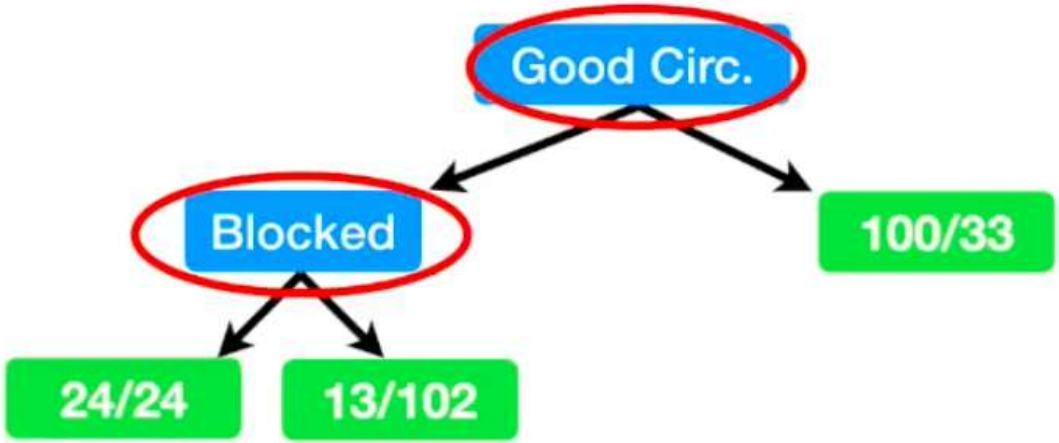


Now, even though we have data for Chest Pain, it is not part of our tree any more.



Now, even though we have data for Chest Pain, it is not part of our tree any more.

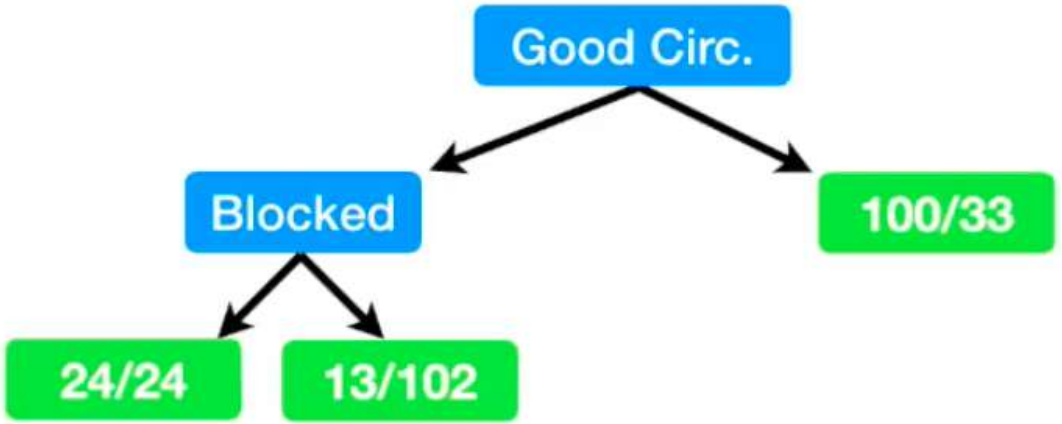
All that's left are Good Circulation and Blocked Arteries.



This is a type of automatic feature selection.

However, we could have also created a threshold such that the impurity reduction has to be large enough to make a big difference.

As a result, we end up with simpler trees that are not “over fit”.

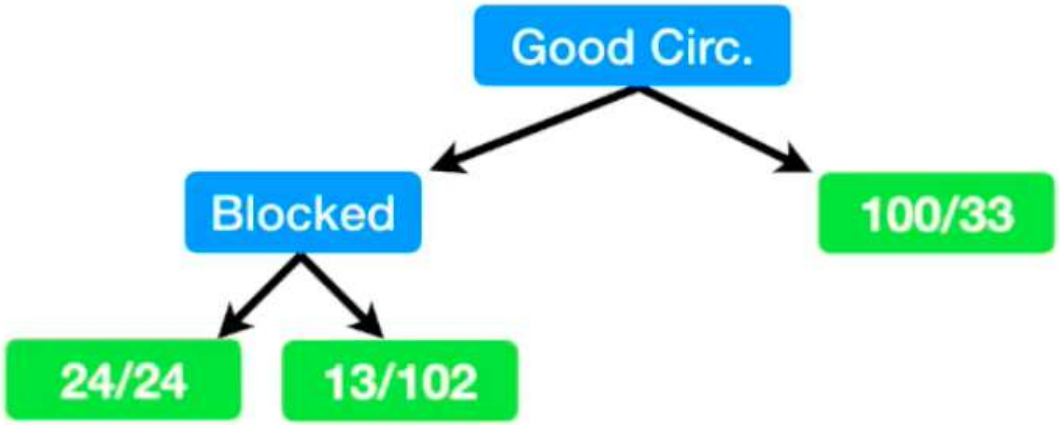


Oh no!!! Some jargon just snuck up on us!!!

"Over fit" means our tree does well with the original data - the data we used to make the tree - but doesn't do well with any other data set.

Decision Trees have the downside of often being over fit.

Requiring each split to make a large reduction in impurity helps a tree from being over fit.



So, in a nutshell, that's what feature selection is all about…