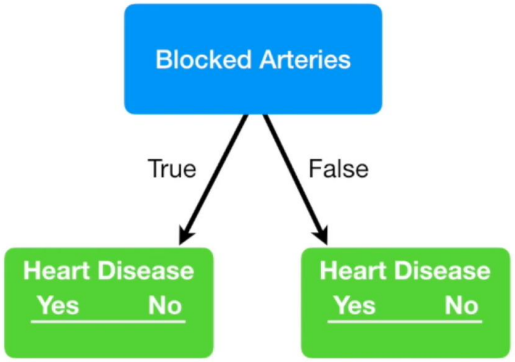
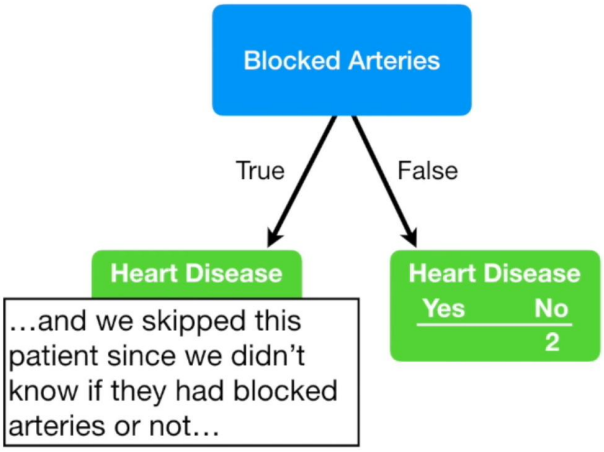
In the first video on decision trees, we calculated impurity for blocked arteries…

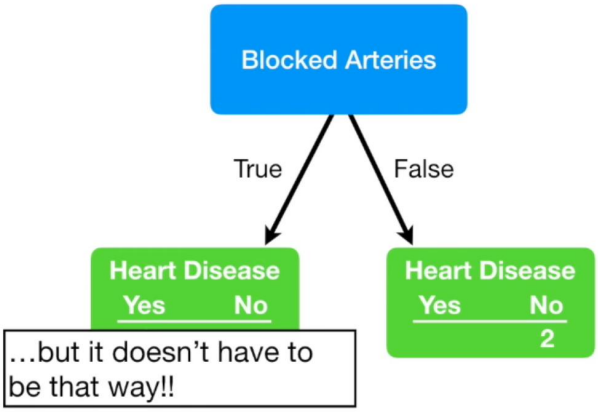
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



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…

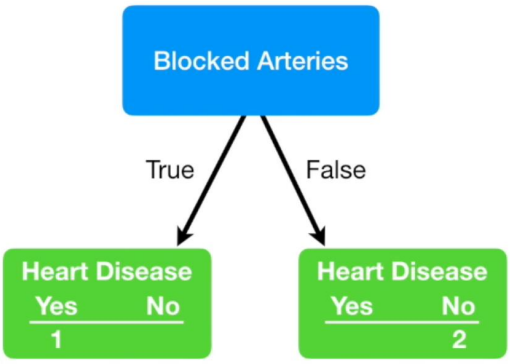


Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…

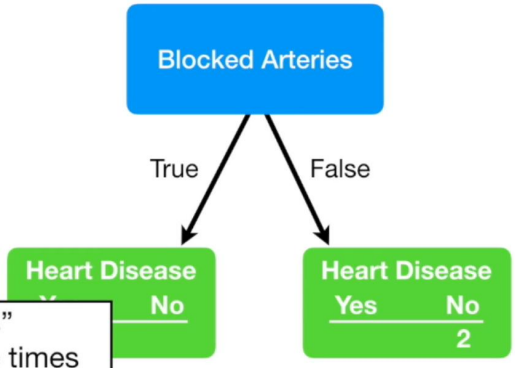


We could pick the most common option…

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…



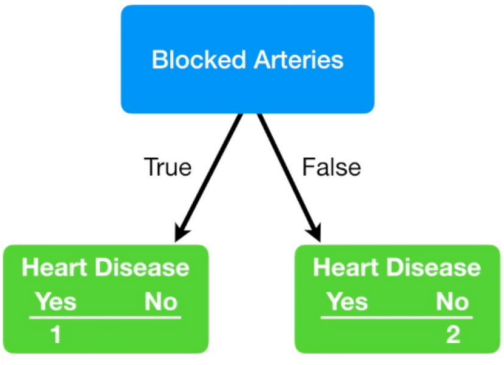
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



If, overall, “yes” occurred more times than “no”, we could put “yes” here…

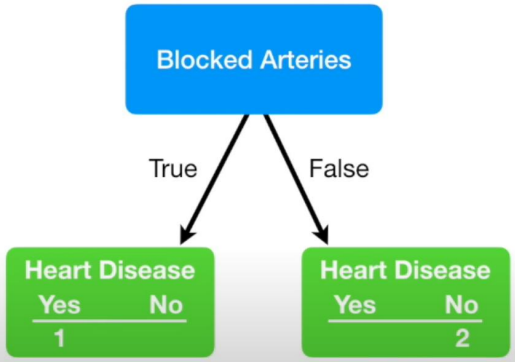
Alternatively, we could find another column that has the highest correlation with blocked arteries and use that as a guide.

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



In this case, Chest Pain and Blocked Arteries are often very similar.

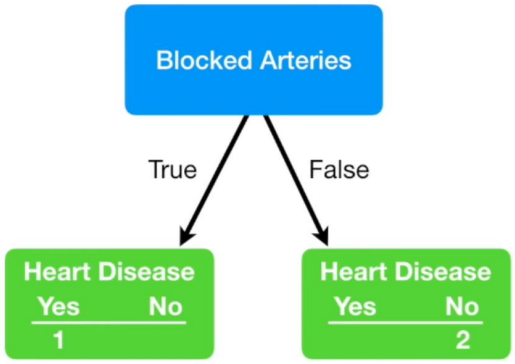
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
No	Yes	No	No
Yes	No	???	Yes
etc…	etc…	etc…	etc…



In this case, Chest Pain and Blocked Arteries are often very similar.

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

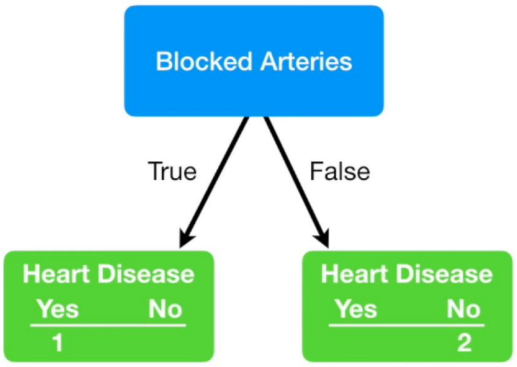
Since Chest Pain is "Yes"…



In this case, Chest Pain and Blocked Arteries are often very similar.

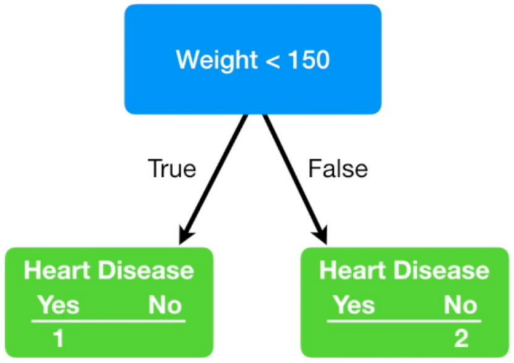
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
No	Yes	No	
Yes	No	**YES**	
etc…	etc…	etc…	

We'll make Blocked Arteries "Yes" as well.



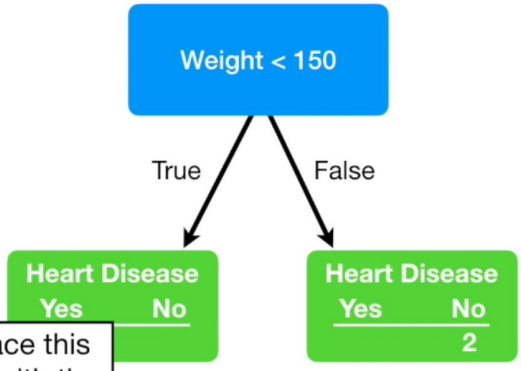
Now imagine we had weight data instead of Blocked Artery data…

Height	Good Blood Circulation	Weight	Heart Disease
5'7"	No	155	No
6'	Yes	180	Yes
5'4"	Yes	120	No
5'8"	No	**???**	Yes
etc…	etc…	etc…	etc…



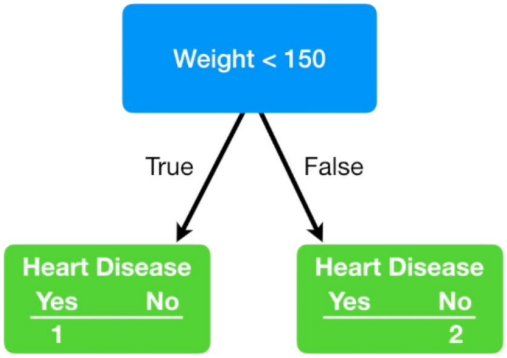
Height	Good Blood Circulation	Weight	Heart Disease
5'7"	No	155	No
6'	Yes	180	Yes
5'4"	Yes	120	No
5'8"	No	**???**	
etc…	etc…	etc…	

We could replace this missing value with the mean or median…



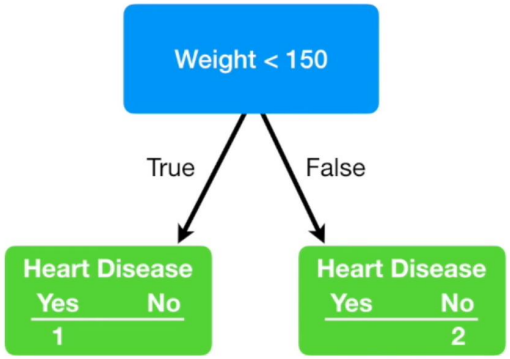
Alternatively, we could find another column that has the highest correlation with weight…

Height	Good Blood Circulation	Weight	Heart Disease
5’7”	No	155	No
6’	Yes	180	Yes
5’4”	Yes	120	No
5’8”	No	???	Yes
etc…	etc…	etc…	etc…



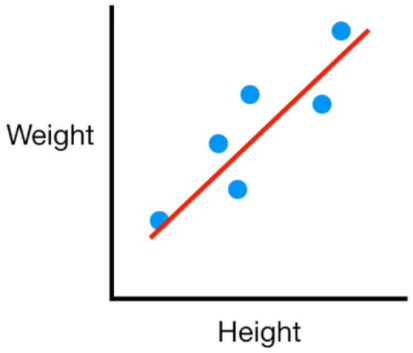
In this case, height is highly correlated with weight…

Height	Good Blood Circulation	Weight	Heart Disease
5’7”	No	155	No
6’	Yes	180	Yes
5’4”	Yes	120	No
5’8”	No	???	Yes
etc…	etc…	etc…	etc…



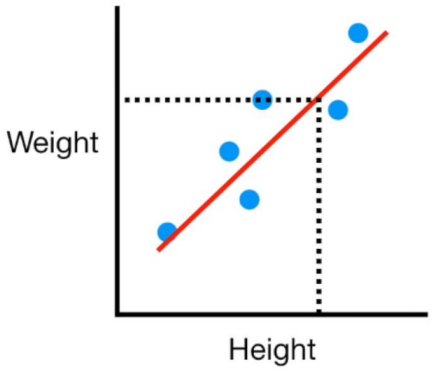
…and do a linear regression on the two columns…

Height	Good Blood Circulation	Weight	Heart Disease
5’7”	No	155	No
6’	Yes	180	Yes
5’4”	Yes	120	No
5’8”	No	???	Yes
etc…	etc…	etc…	etc…



…and use the least squares line to predict the value for weight.

Height	Good Blood Circulation	Weight	Heart Disease
5’7”	No	155	No
6’	Yes	180	Yes
5’4”	Yes	120	No
5’8”	No	???	Yes
etc…	etc…	etc…	etc…



So, you can see that if we’re missing some data, there are a lot of ways to guess at what it might be.

Height	Good Blood Circulation	Weight	Heart Disease
5’7”	No	155	No
6’	Yes	180	Yes
5’4”	Yes	120	No
5’8”	No	168	Yes
etc…	etc…	etc…	etc…