# StatQuest学习笔记19——决策树

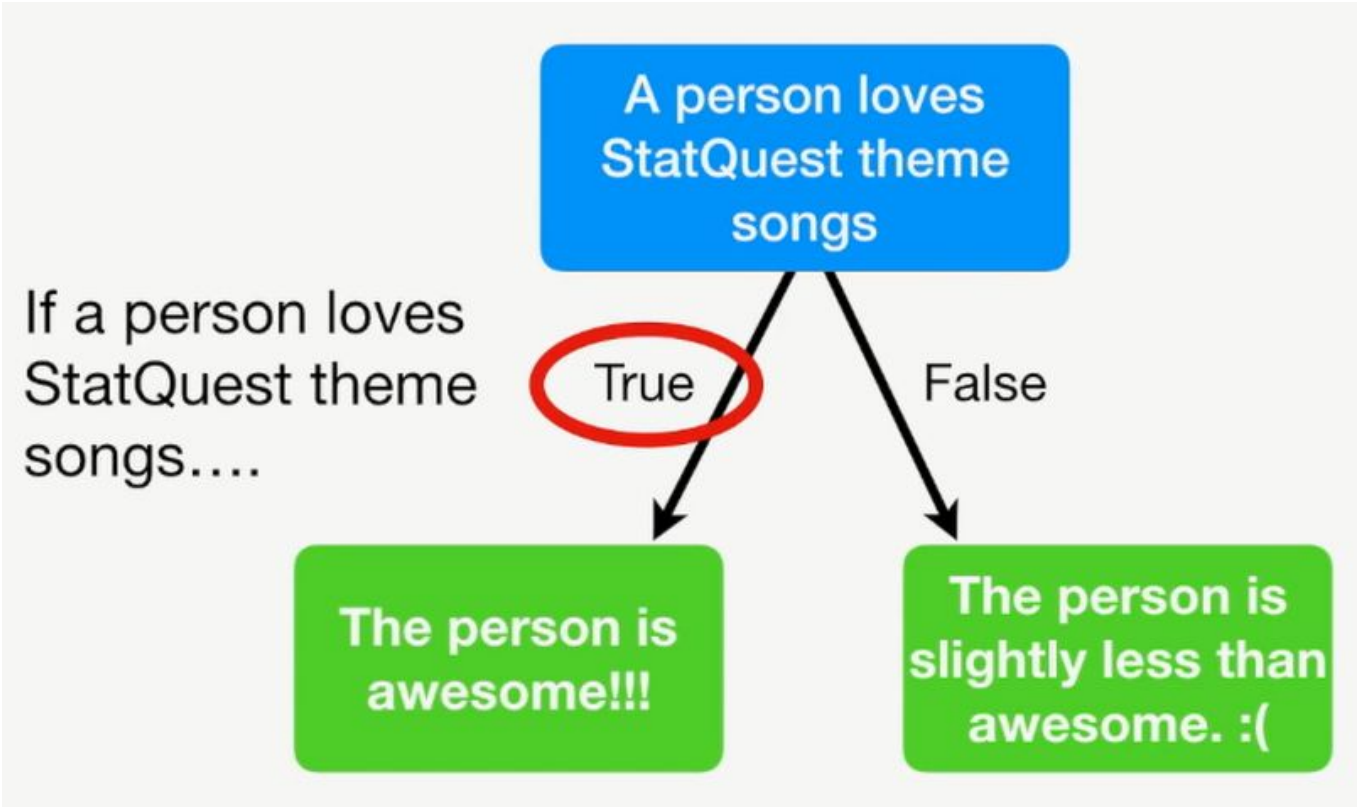
Original rvdsd 读研笔记 2018-07-11 12:00

## 前言

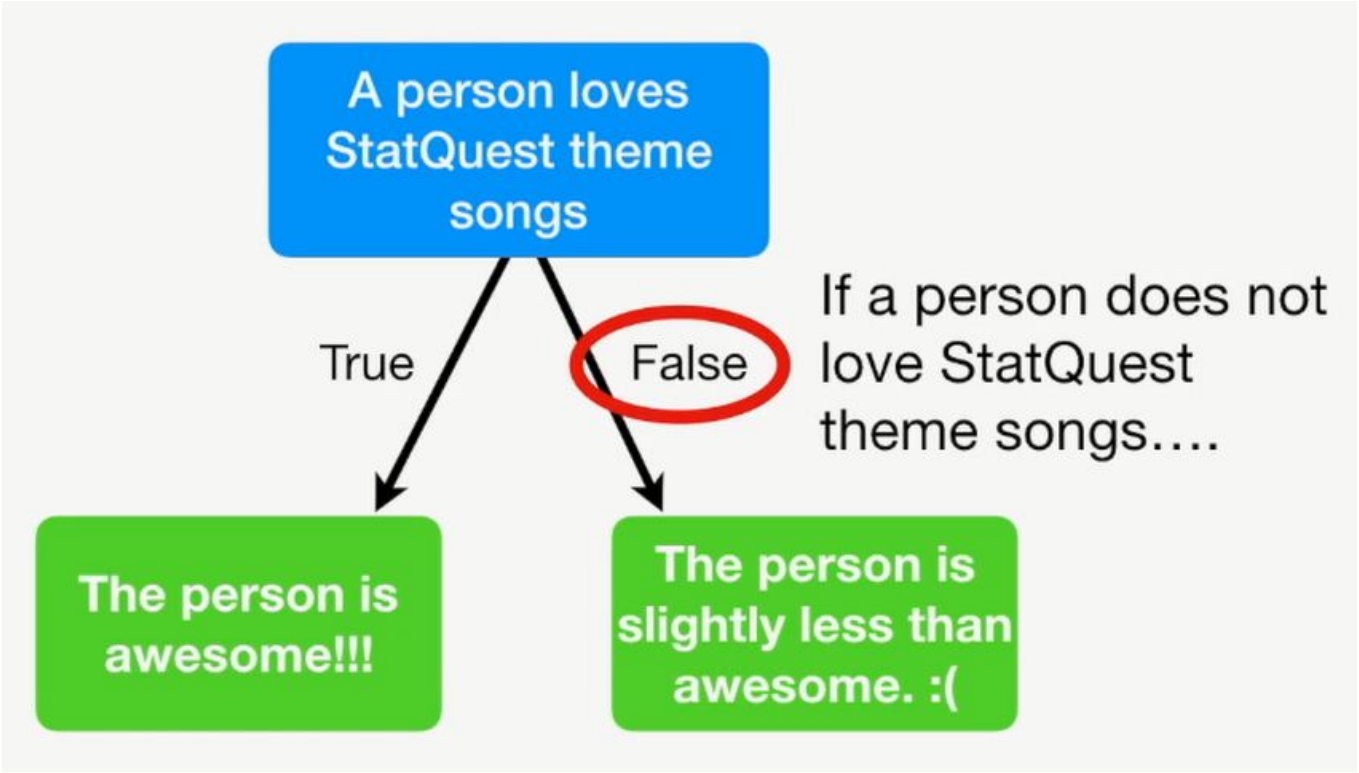
本篇笔记是StatQuet系列视频教程的第50节和第51节，主要内容是决策树（decision tree）。其中第50节的内容是决策木树的基本思想，第51节的内容是在构建决策树中，缺失值的处理。

## 决策树简单案例1

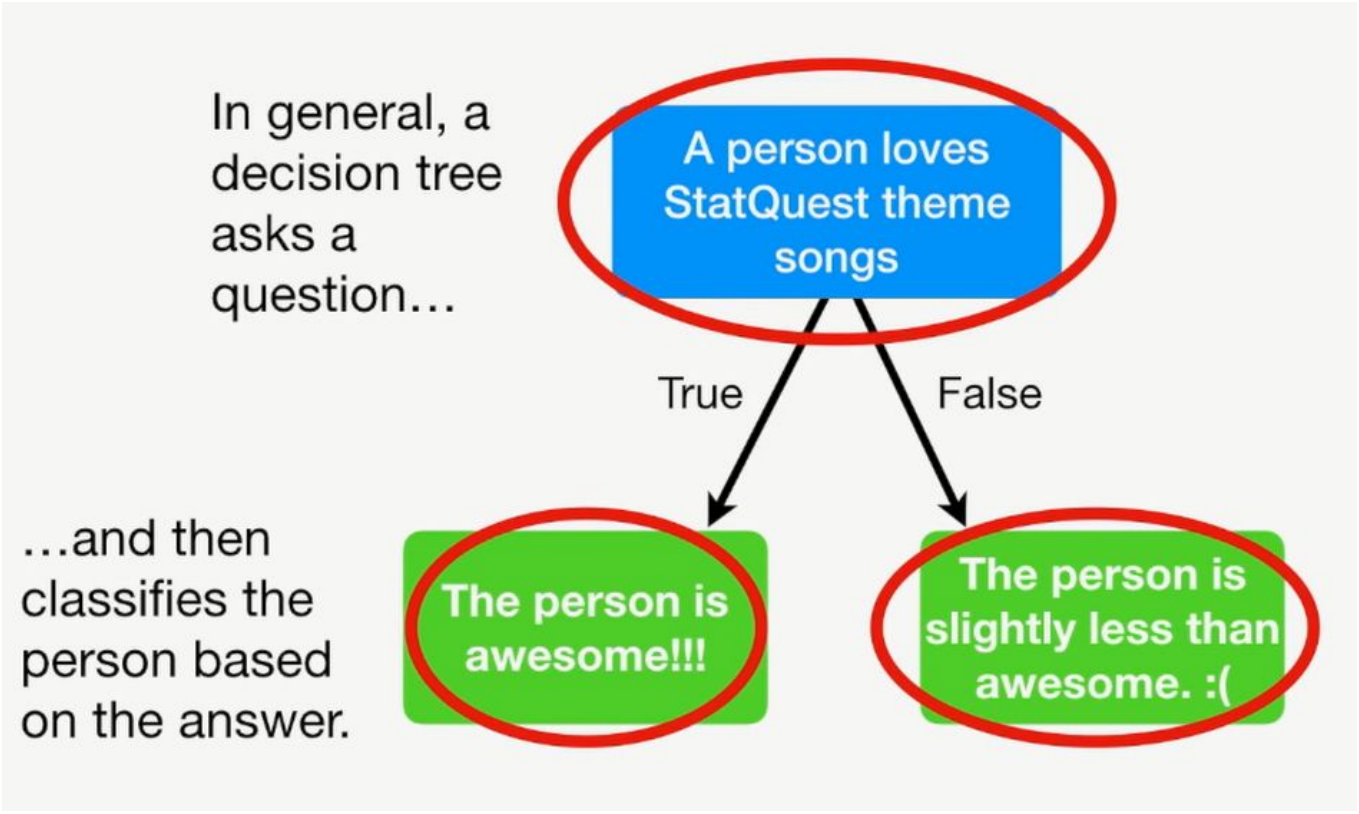
先看一下简单的决策树案例，在这个案例中，我们提出一个问题，也就是问某人，“请问，你喜欢StatQuest的主题曲吗”。回答者会回答“Yes”或“No”，根据回答者的这个回答，我们可以把回答者分为两类，如果回答Yes，则就是下面的情况，如下所示：



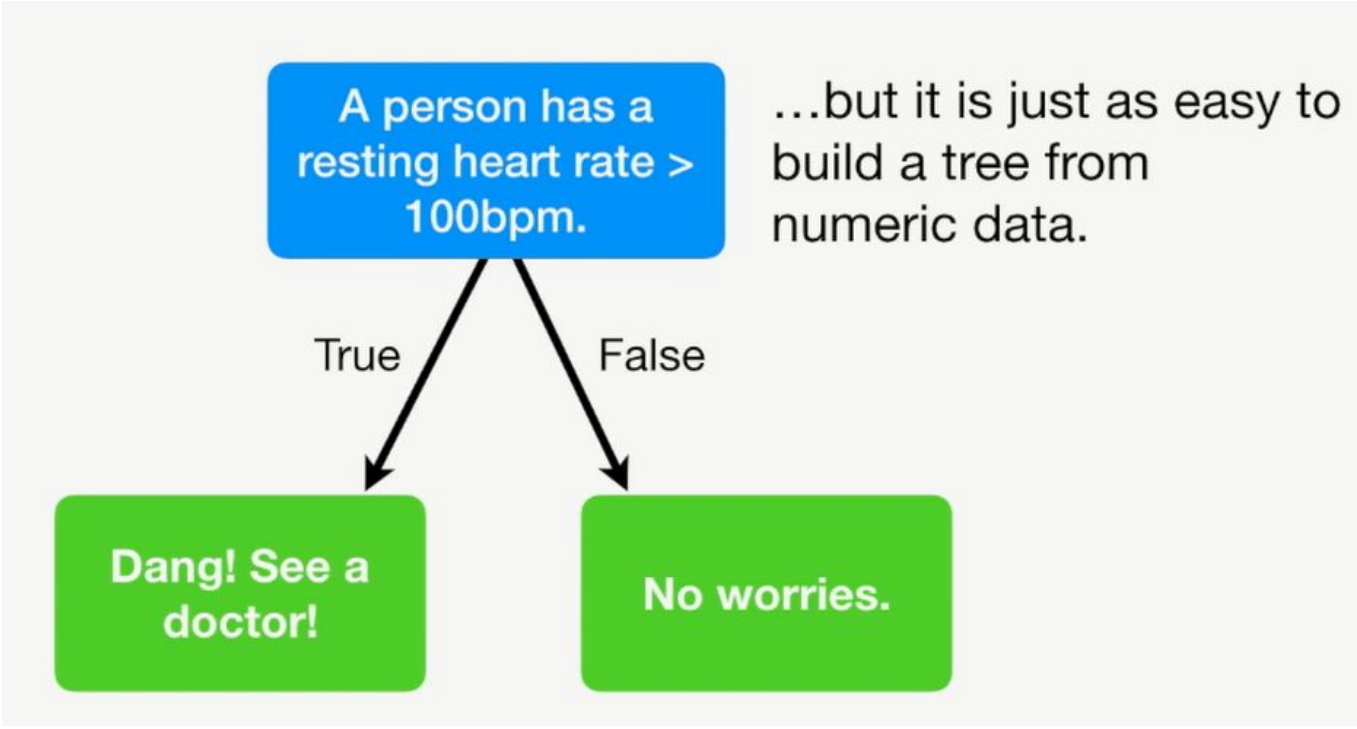

如果回答No，那么就是下面的这种情况：



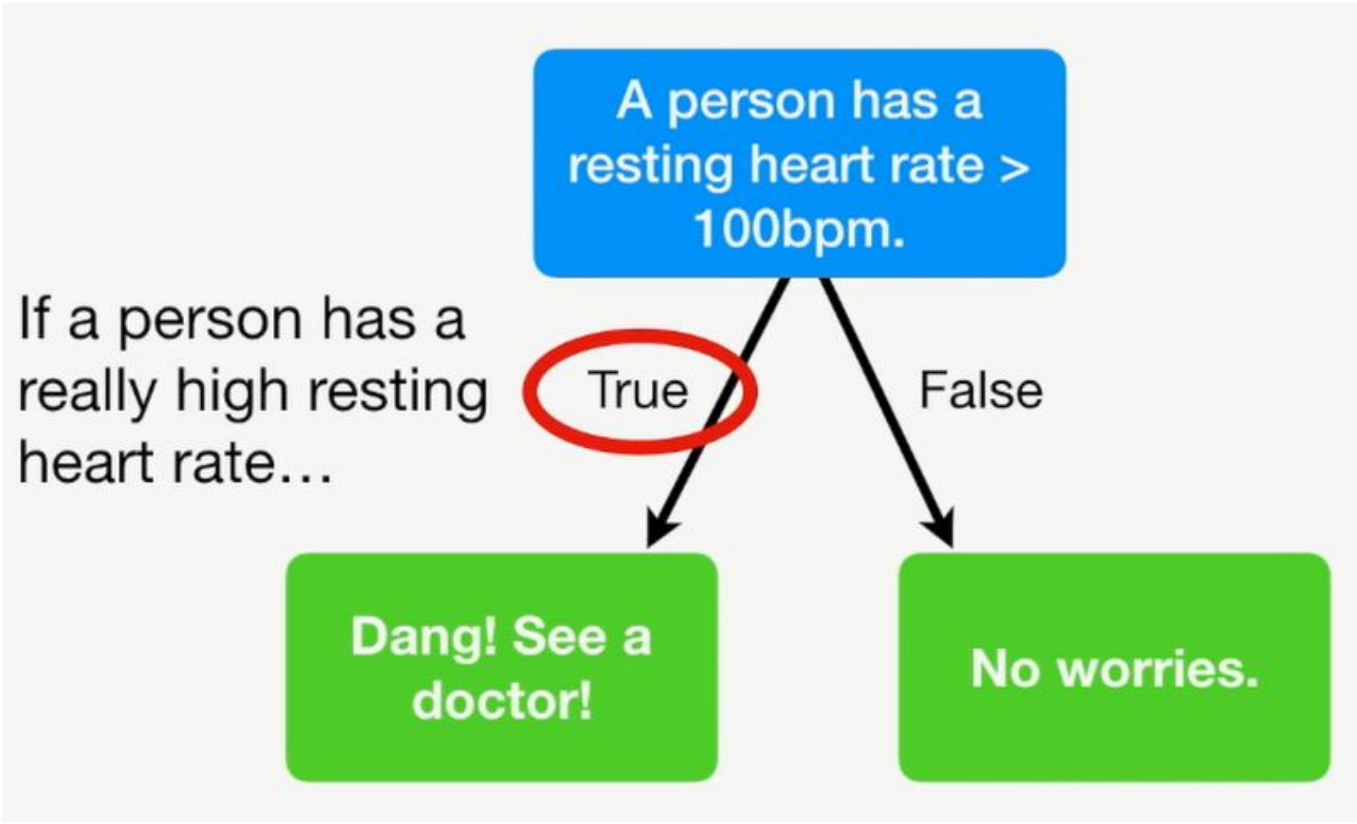

在一个常规的决策树中，通常会问一个问题，然后根据回答者的回答（就是Yes或No），就把他们进行分类，如下所示：



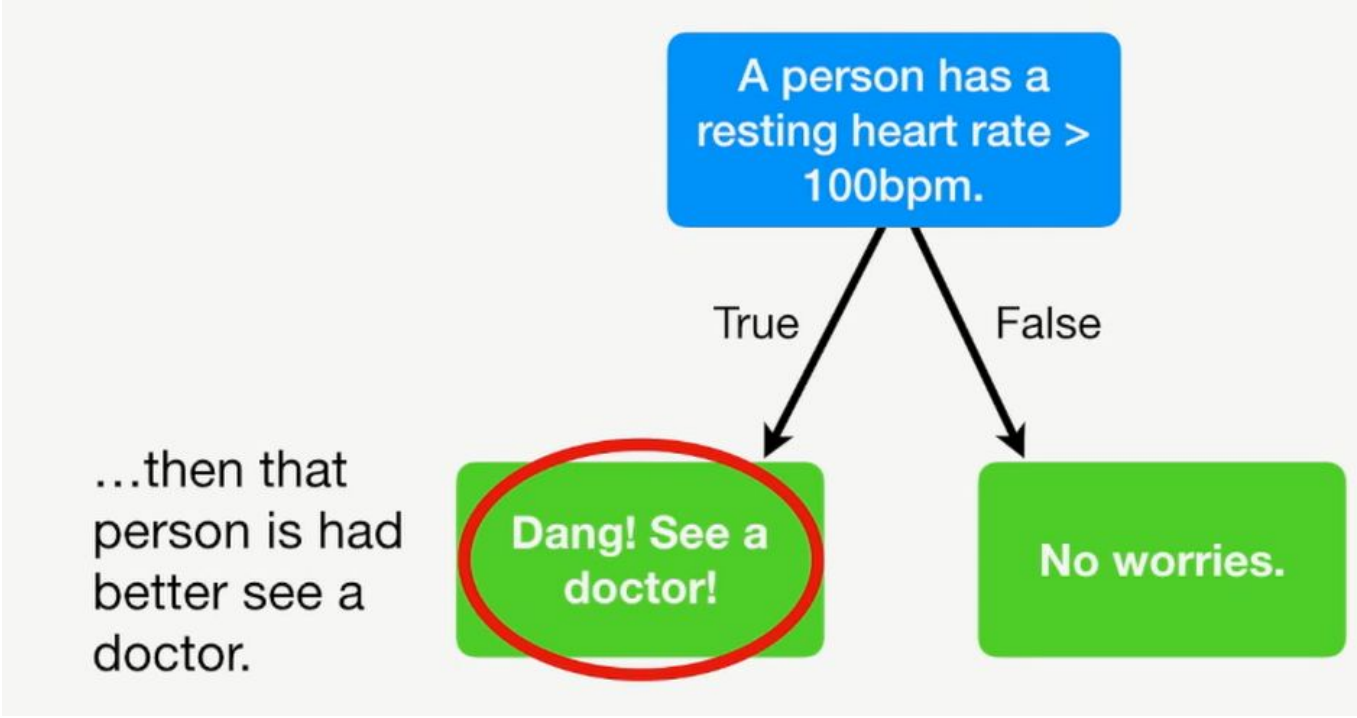

同样的，我们还可以根据一些数据来构建决策树，例如根据人的静息心率（resting heart rate）是否大于100bpm，那么就可以对他们进行分类，如下所示：



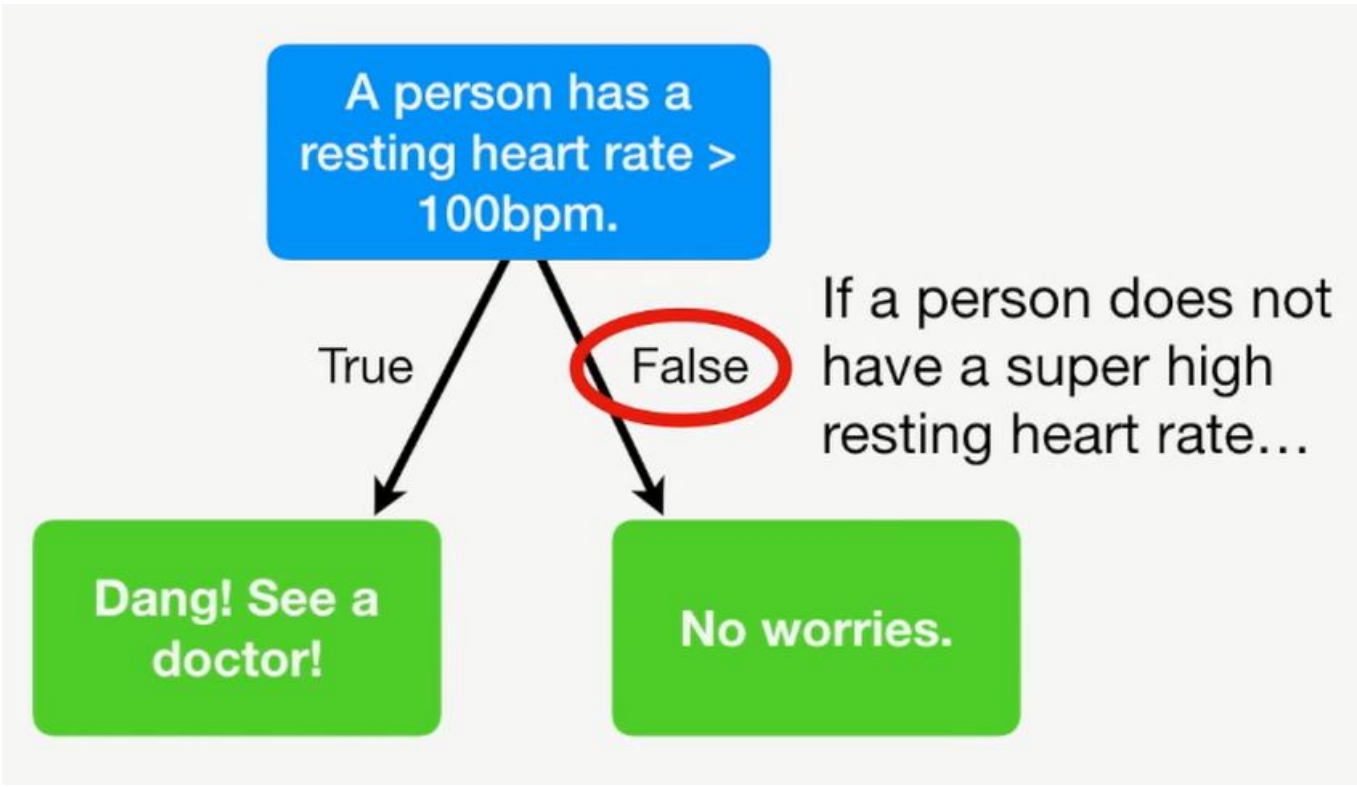

如果一个人的静息心率大于100bpm，那么就是说他的静息心率比较高，如下所示：



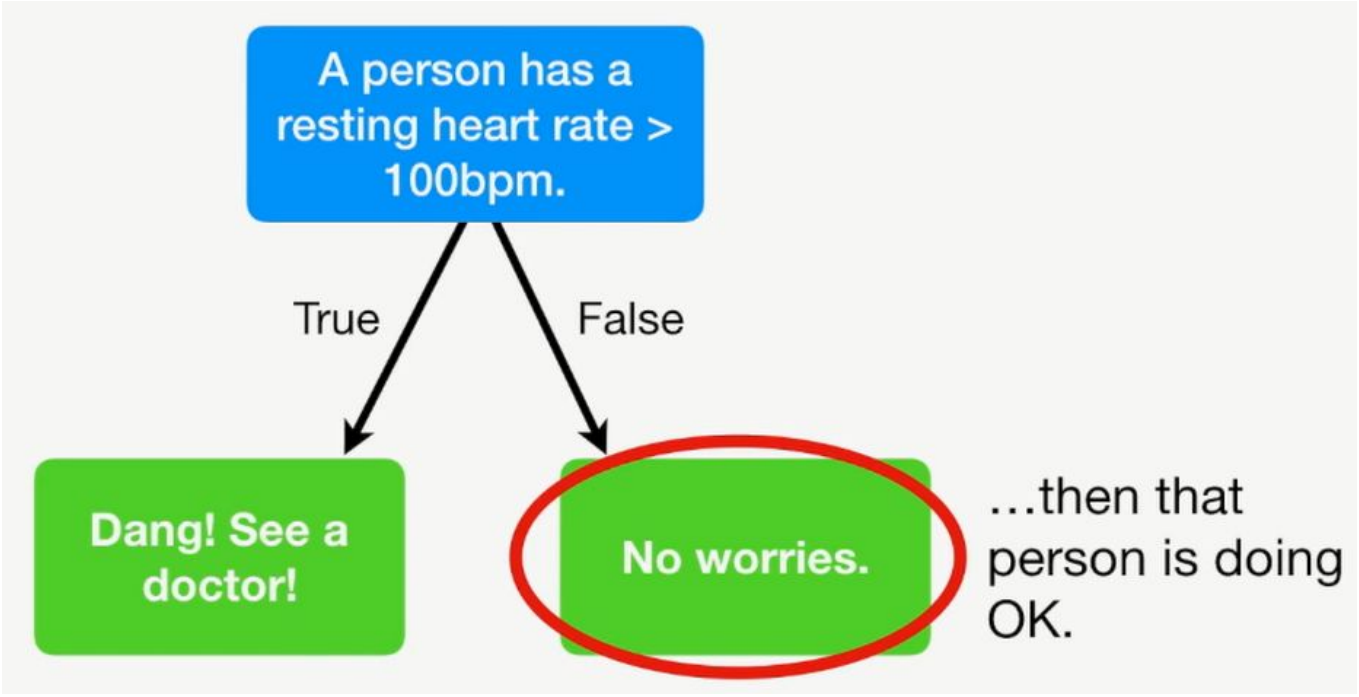

那么他就需要去看医生，如下所示：




如果一个人的静息心率并不高，如下所示：

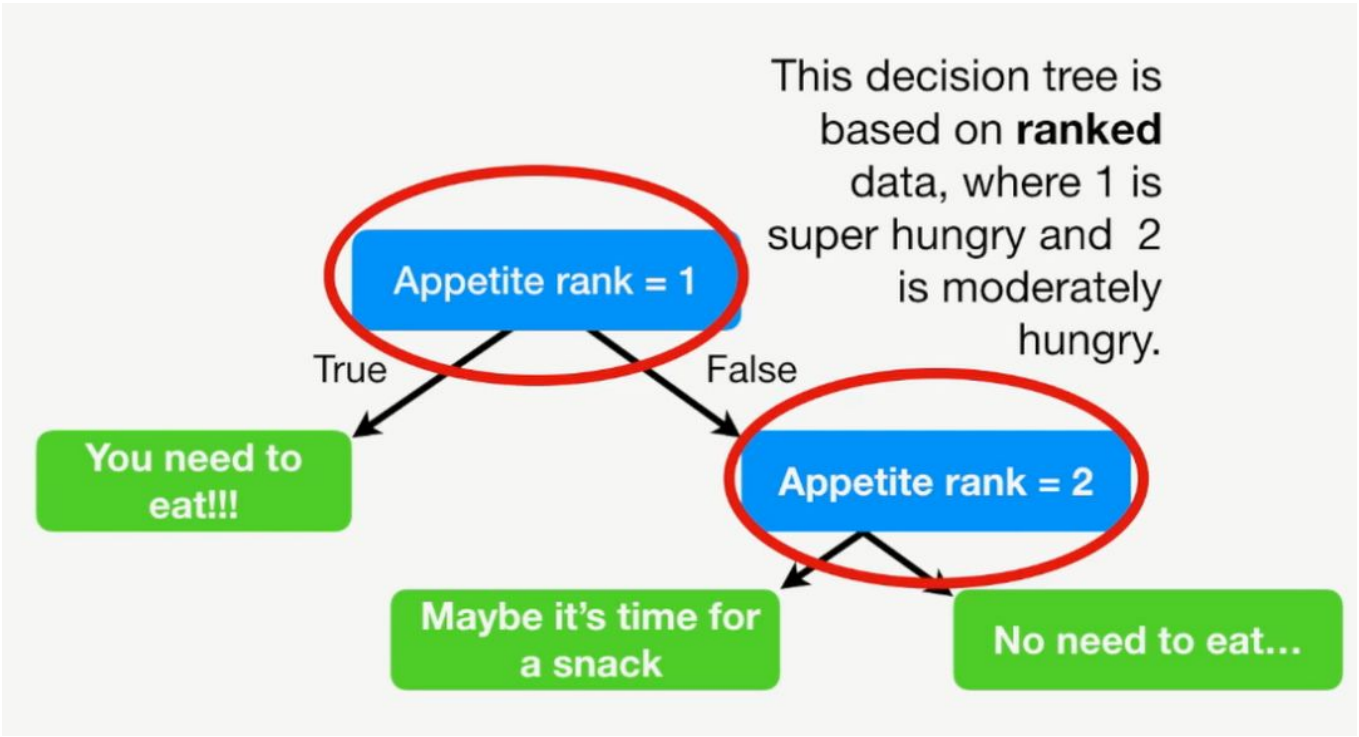



那么这个人就属于比较健康的状态了，如下所示：

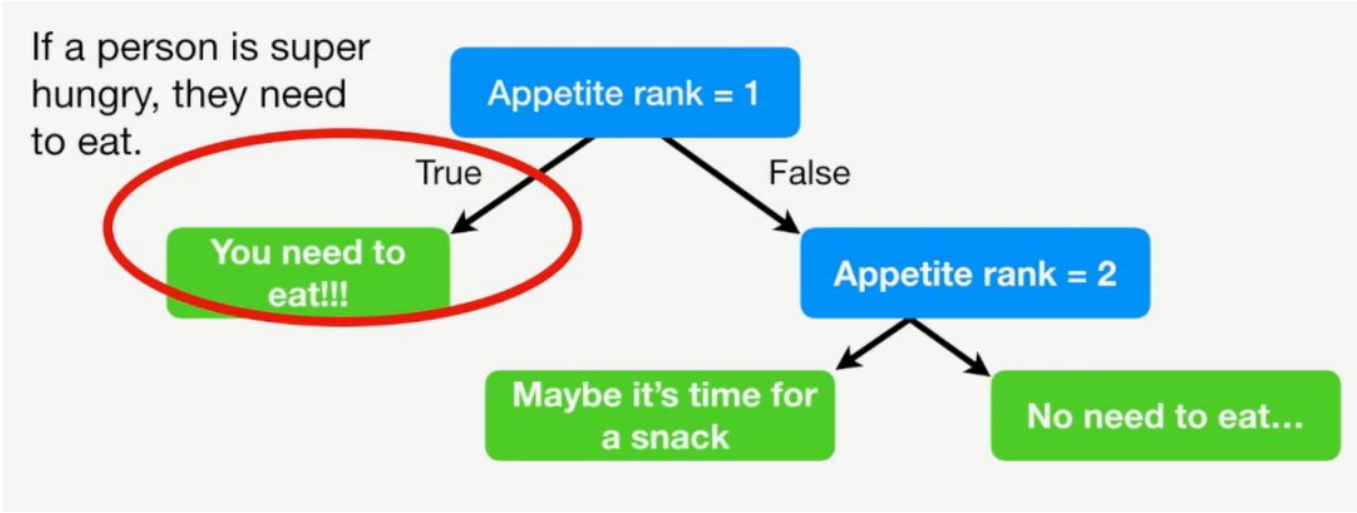



## 决策树简单案例2

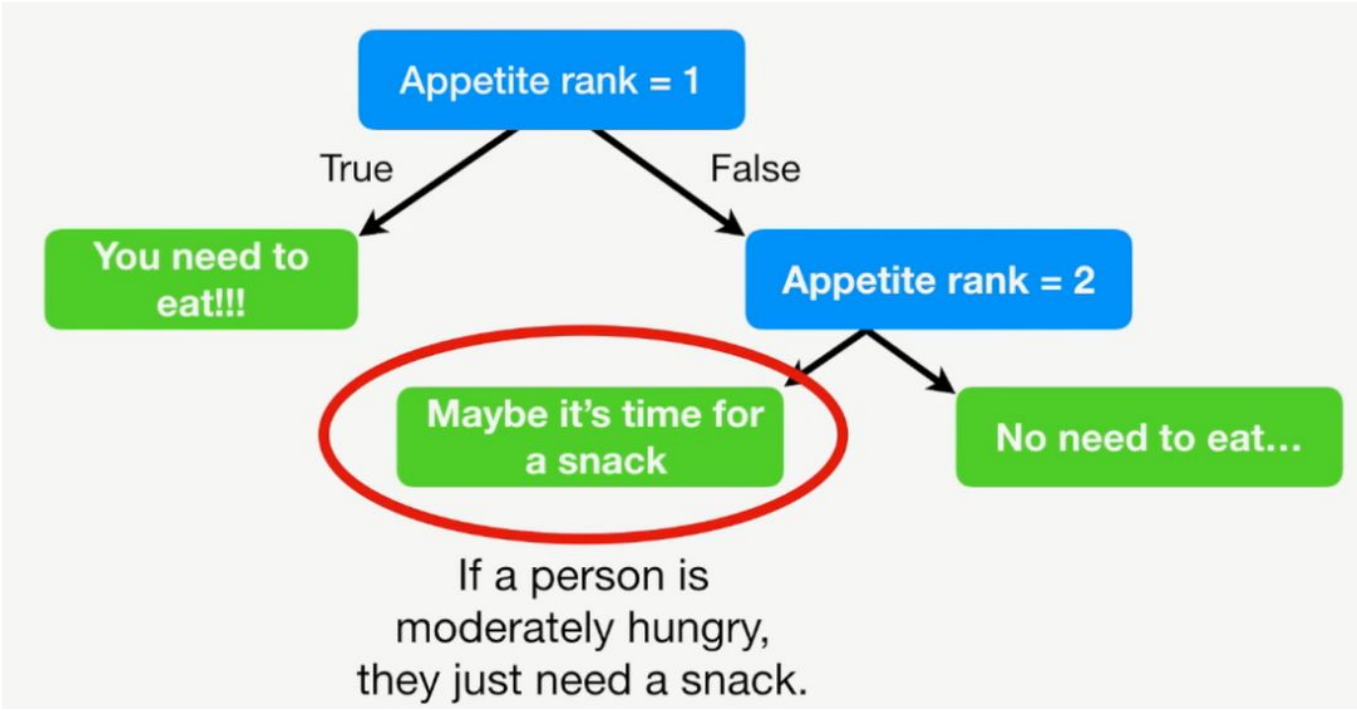
再来看一个决策树的简单案例。在这个案例中，决策树的构建的依据是一个分级数据（ranked data），其中数据`Appetite rank =1`表示`非常饿`，`Appetite rank = 2`表示`一般饿`，如下所示：



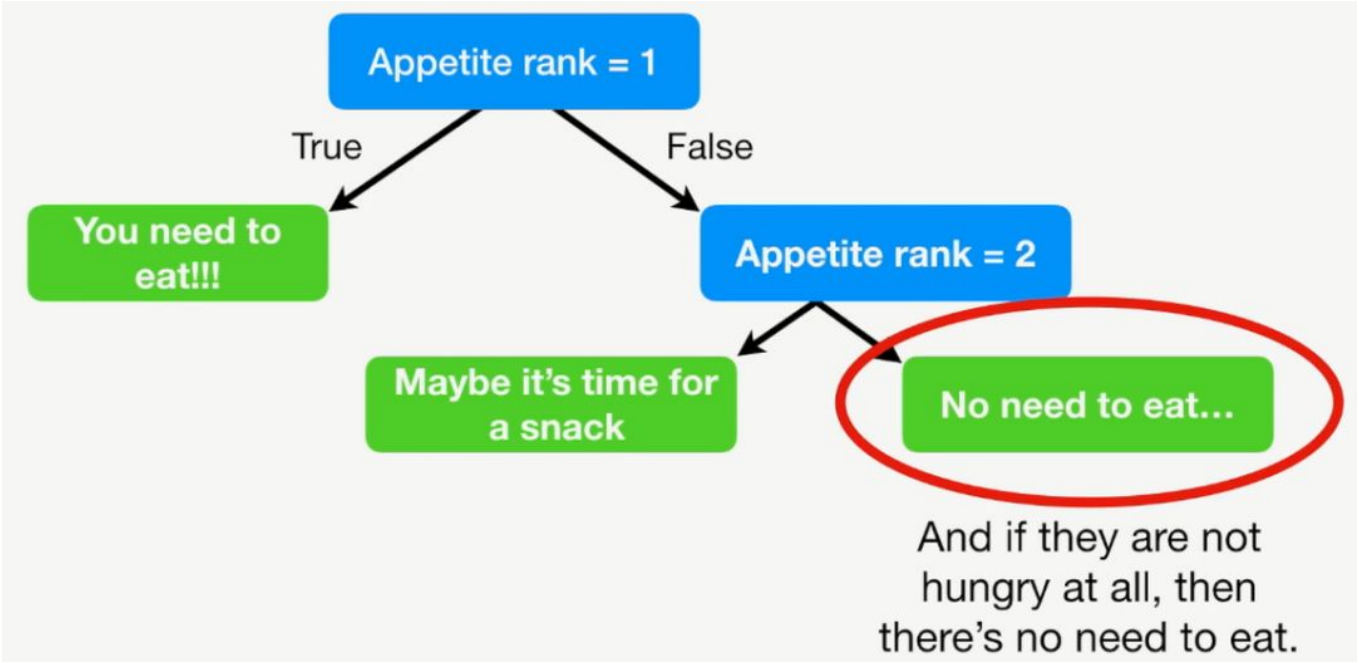

如果一个人非常饿（`Appetite rank = 1`）时，那么他就需要吃东西，如下所示：



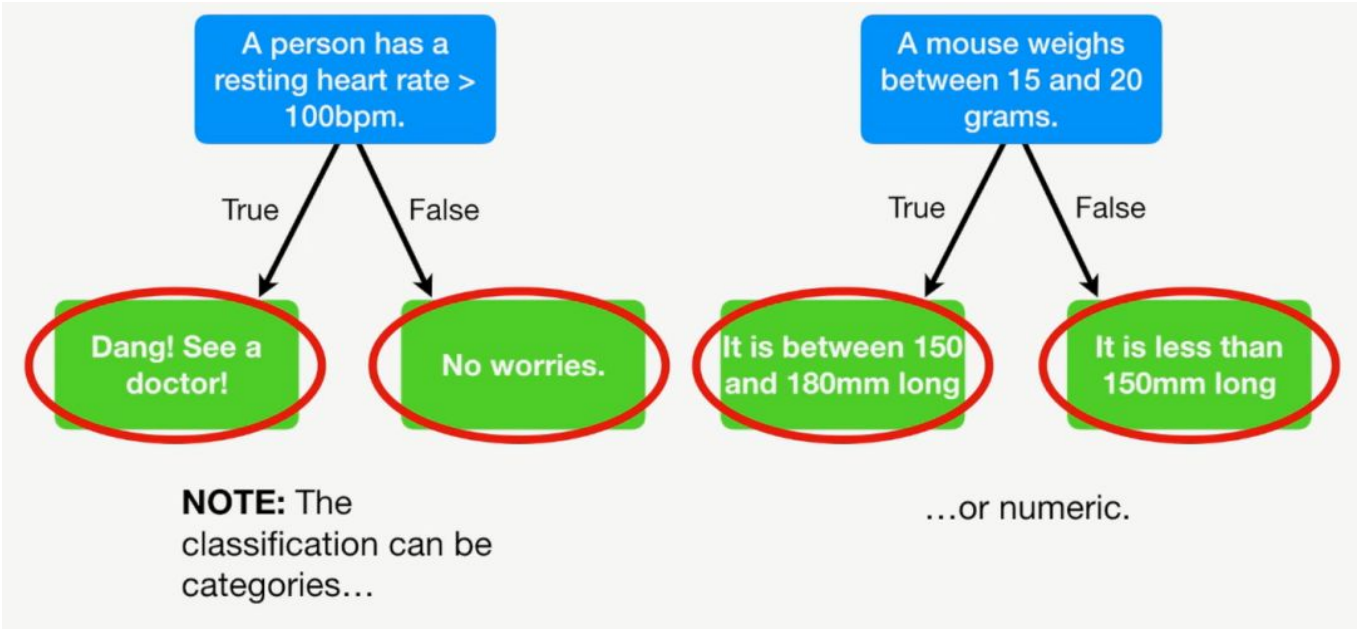

如果这个人只是处于一般饿的状态（`Appetite rank = 2`），那么他只需要只一点东西即可，如下所示：



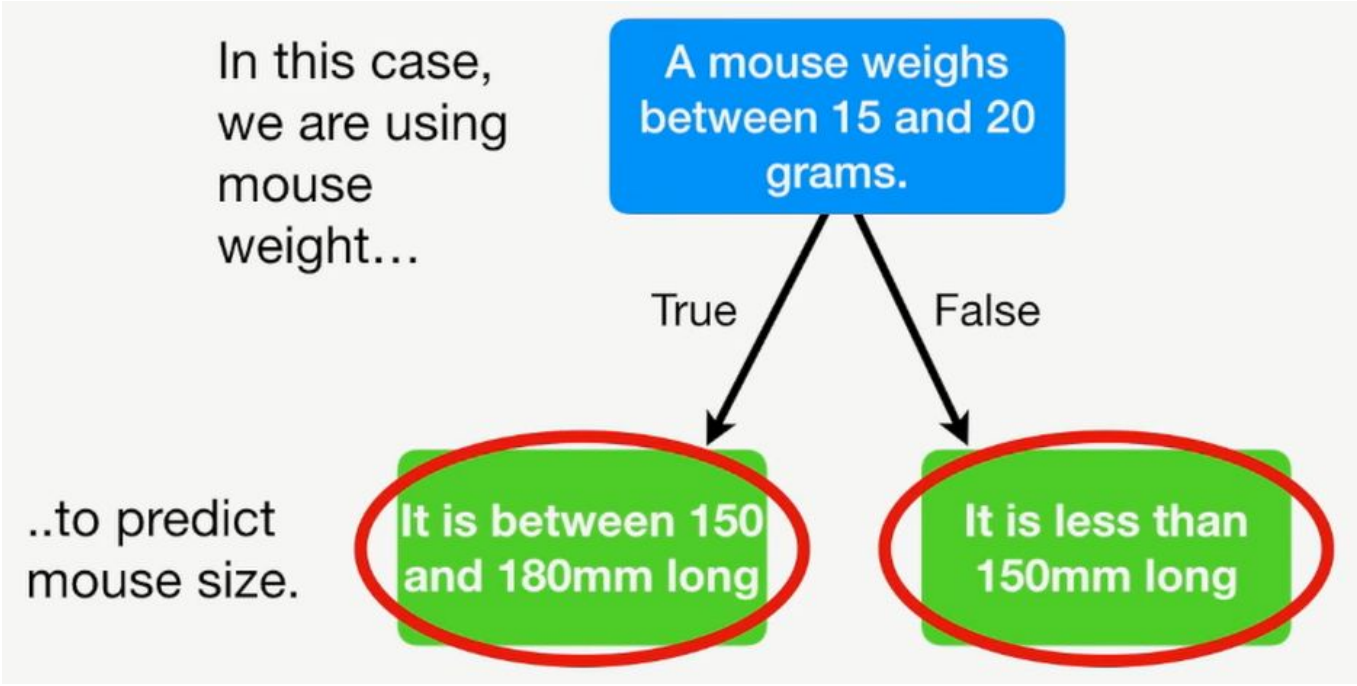

如果这个人不是特别的饿，那么他也可以不吃东西，如下所示：




需要注意的是，决策树中的分类可以是一些二分类变量，如下图左侧所示，也可以是一些连续型变量（数值型变量），如下图右侧所示：

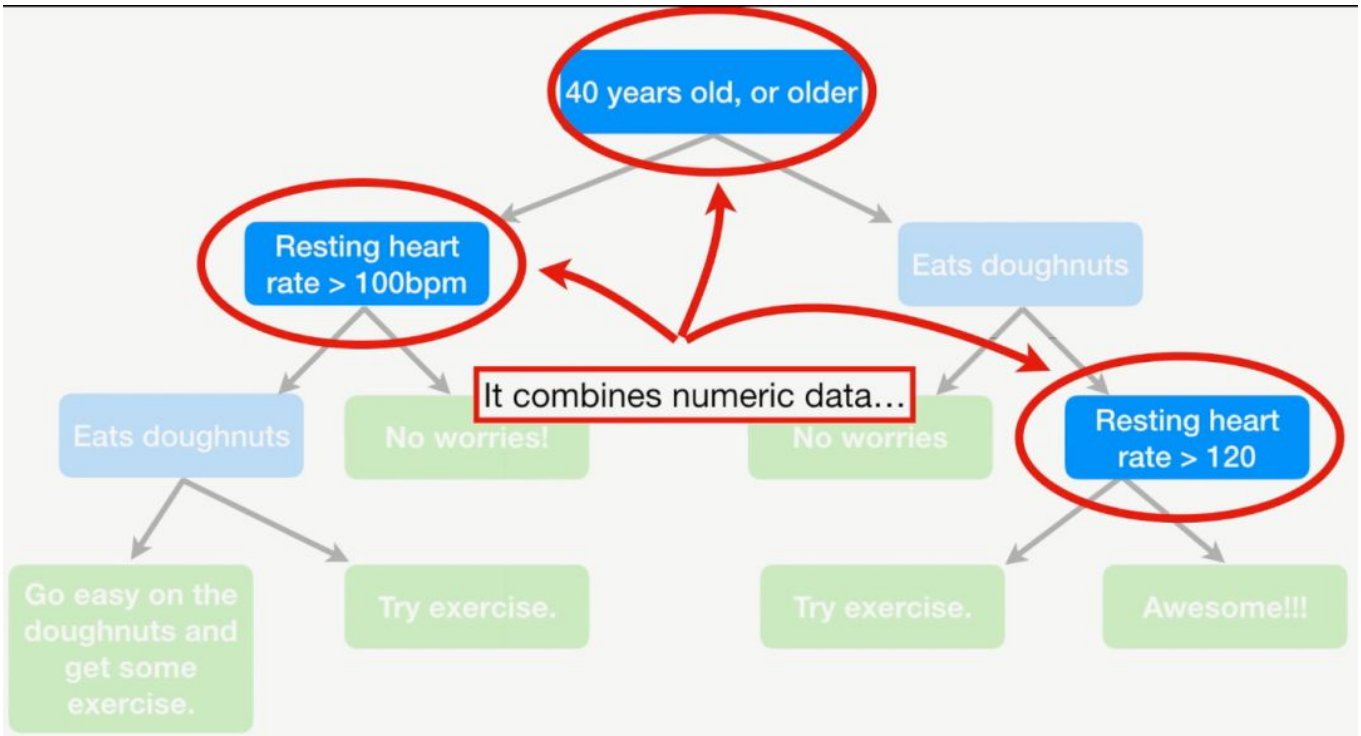



例如，上图的右侧，我们使用小鼠的体重来预测小鼠的大小（size），如下所示：

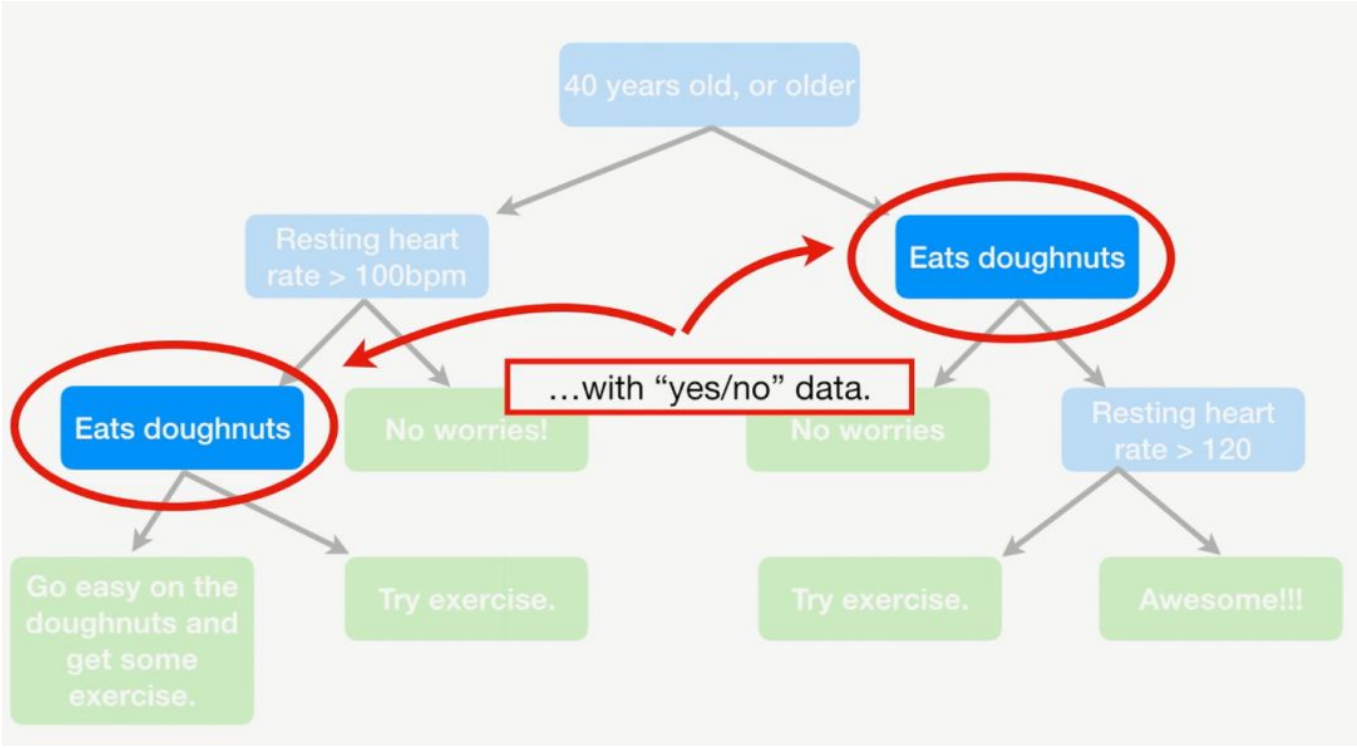



## 复杂案例

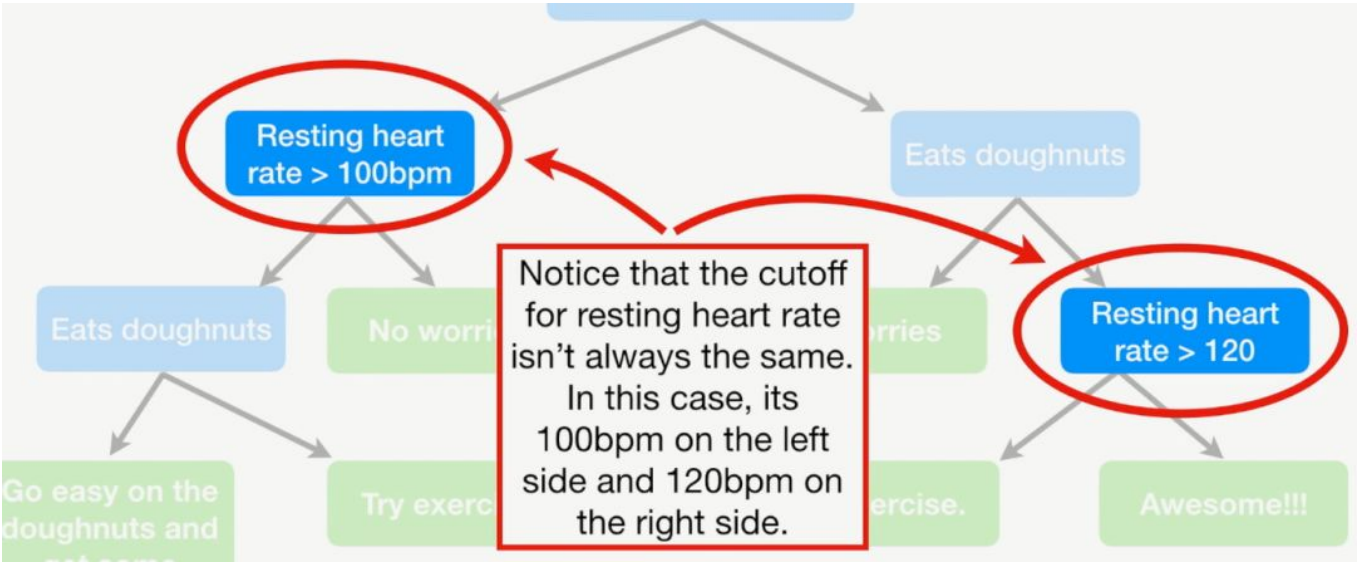
我们再看一个比较复杂的决策树案例，在这个决策树中，它混合了连续型变量和二分类变量，下图是决策树中的连续型变量，如下所示：



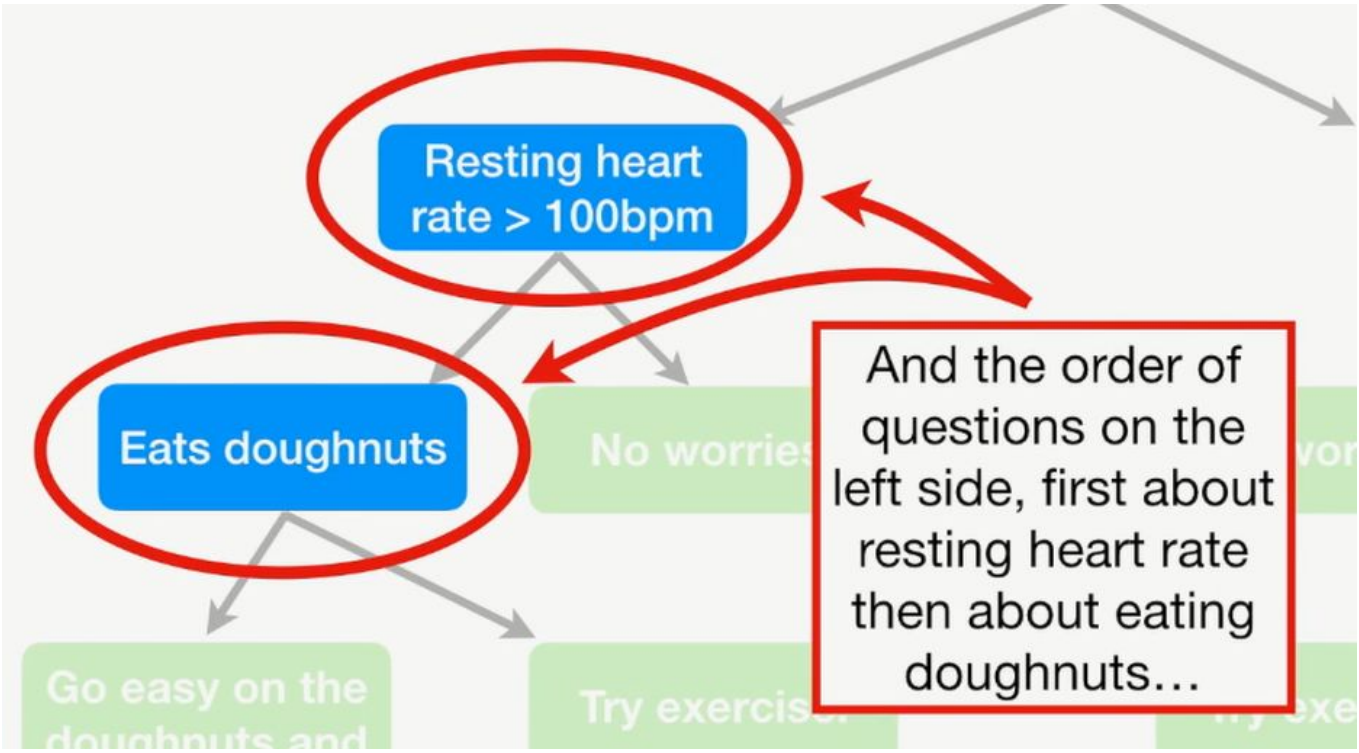

接着是二分类变量，它使用的yes/no来进行区分，如下所示：



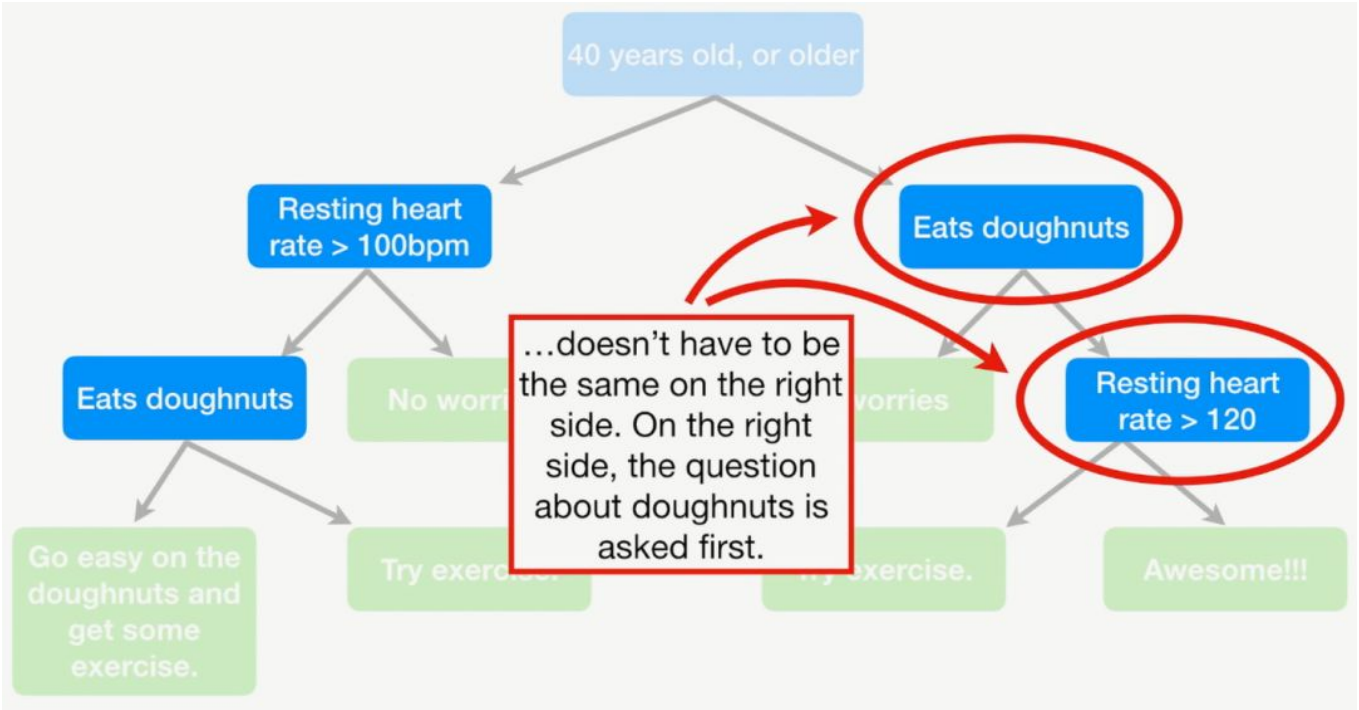

在这个决策树中，我们需要注意的是，静息心率的阈值并非都是相同的，例如在左侧，它的阈值是100bpm，而在右侧，它的阈值是120bpm，如下所示：



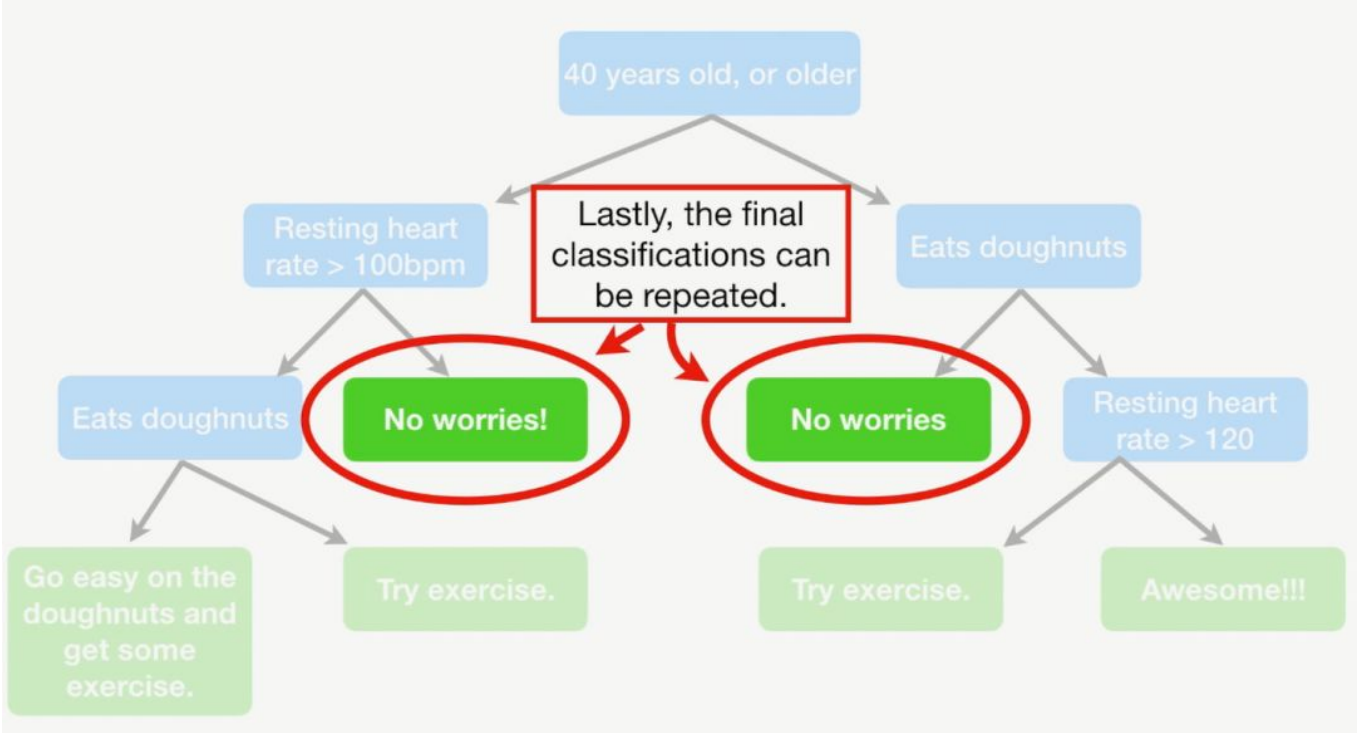

在决策树中，问题的顺序也可以不一样，例如在左侧，第一个问题是静息心率，然后为是否吃甜甜圈（doughnut），如下所示：



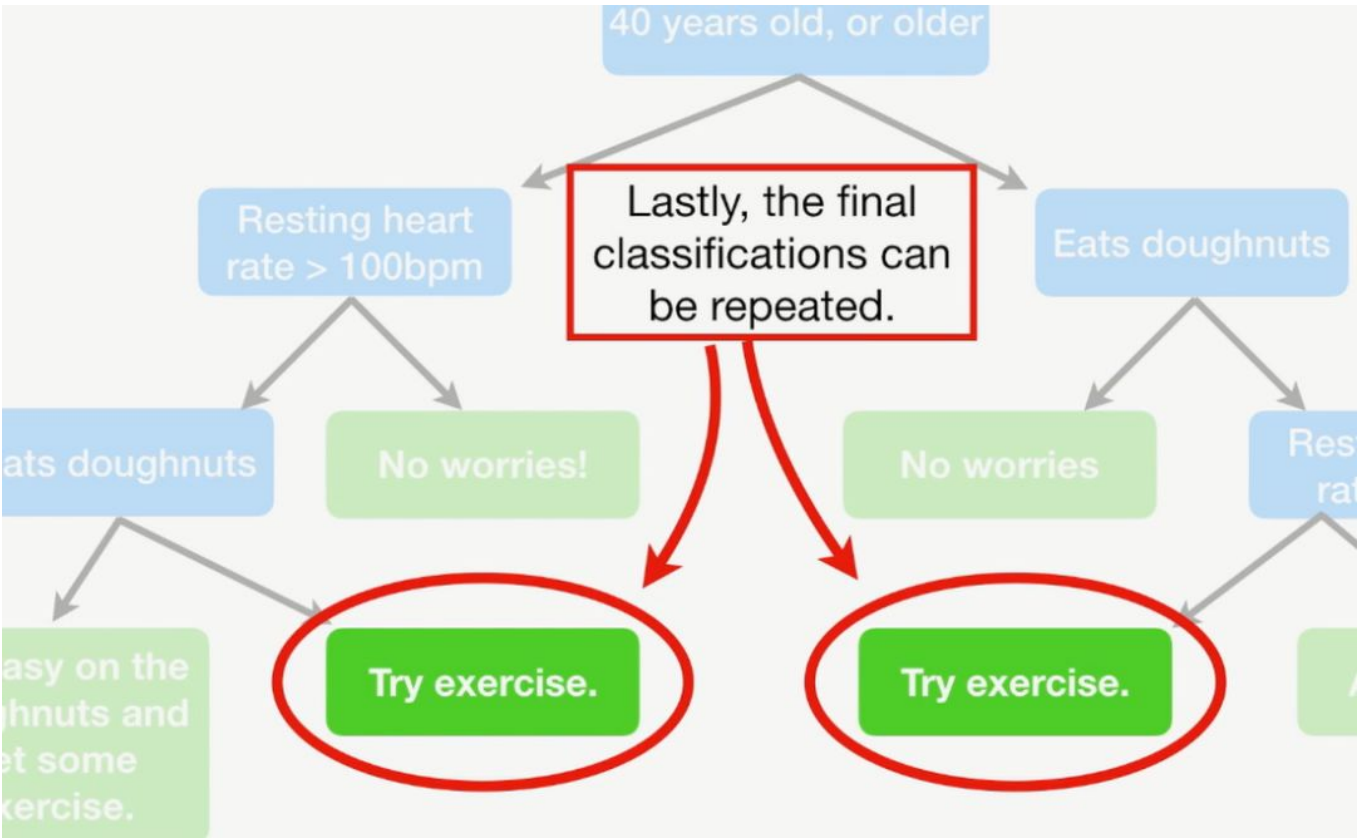

而在右侧，第一个问题则为是否吃甜甜圈，然后是静息心率，如下所示：



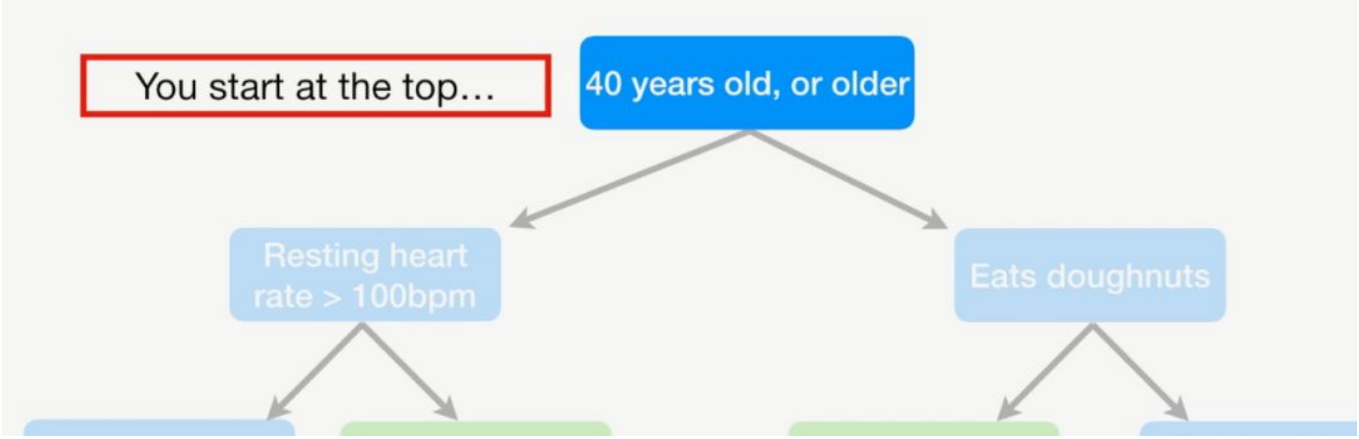

最后的归类可以是一样的，如下所示：



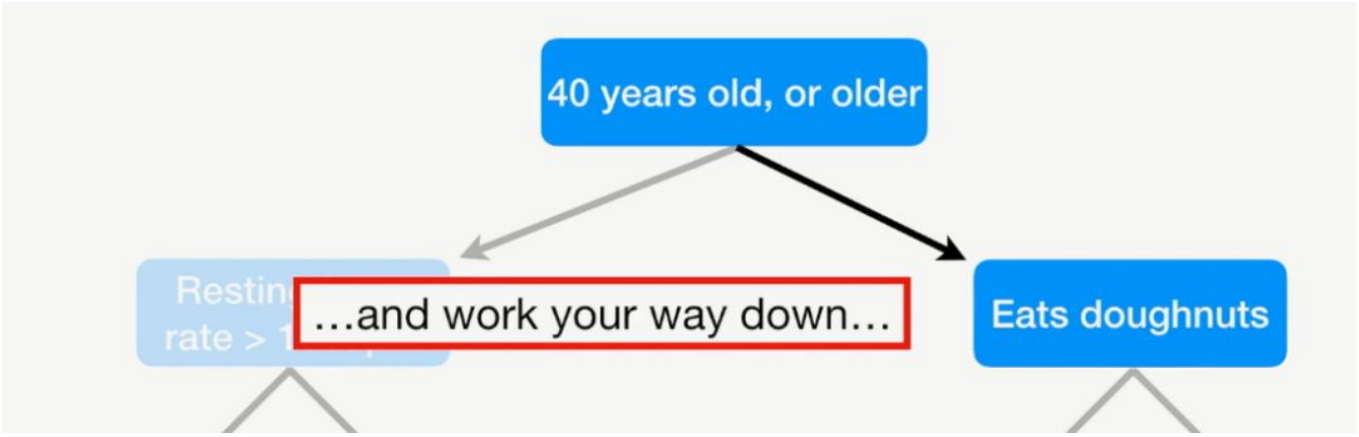

还有两个相同的分支，如下所示：



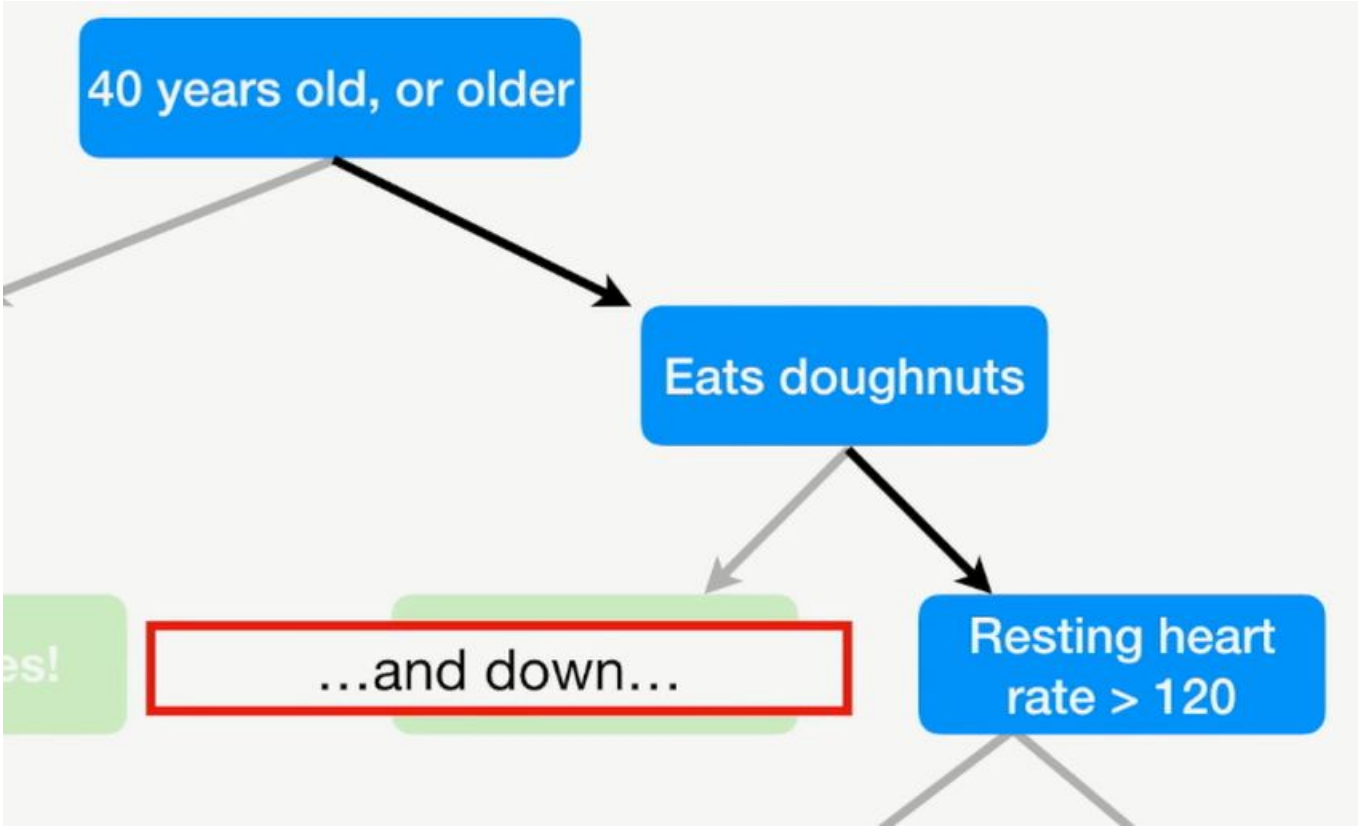

对于大多数的决策树来说，它们都很直观，从顶层开始，如下所示：



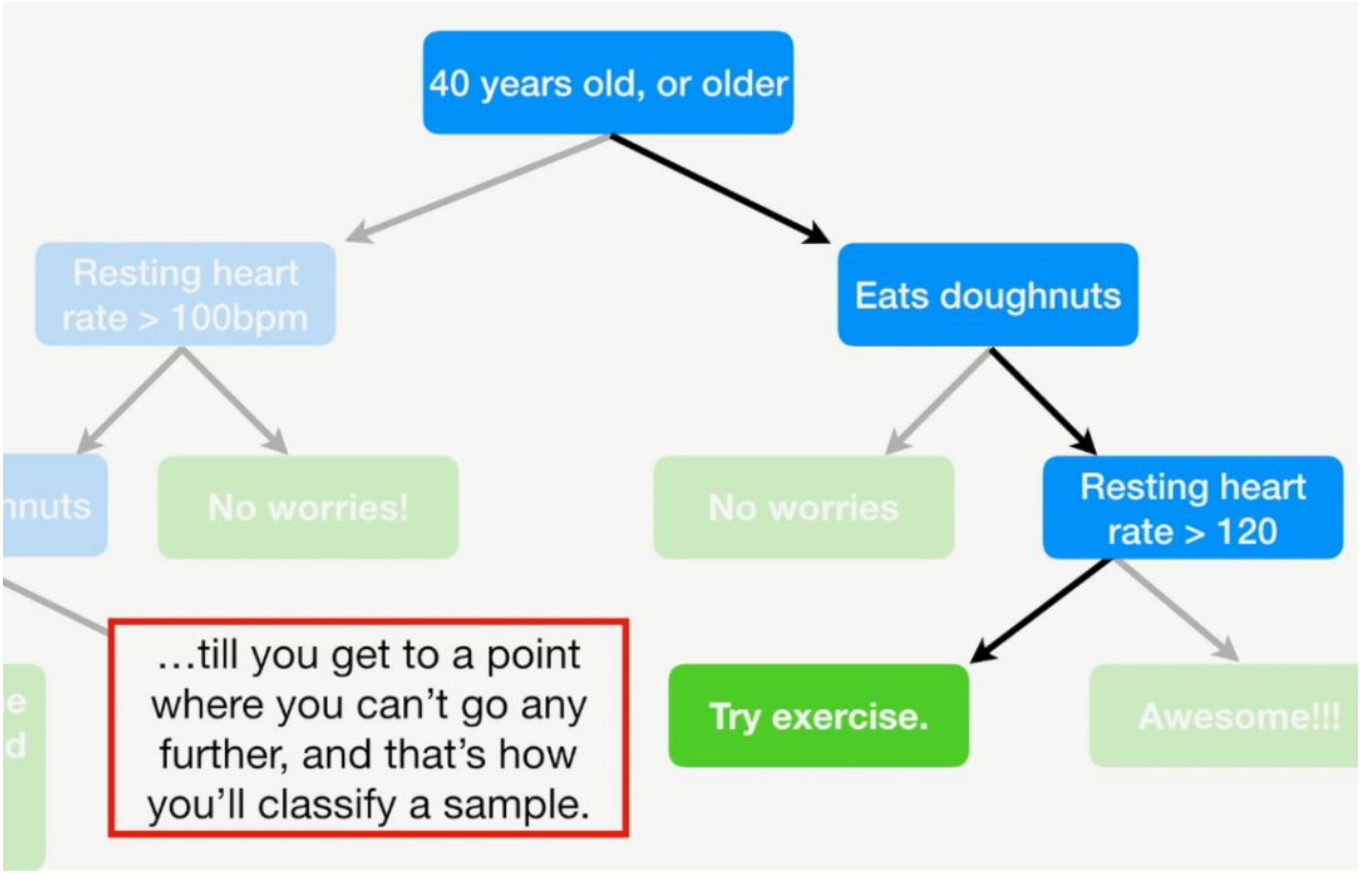

接着进入下一层，如下所示：




再下一层，如下所示：

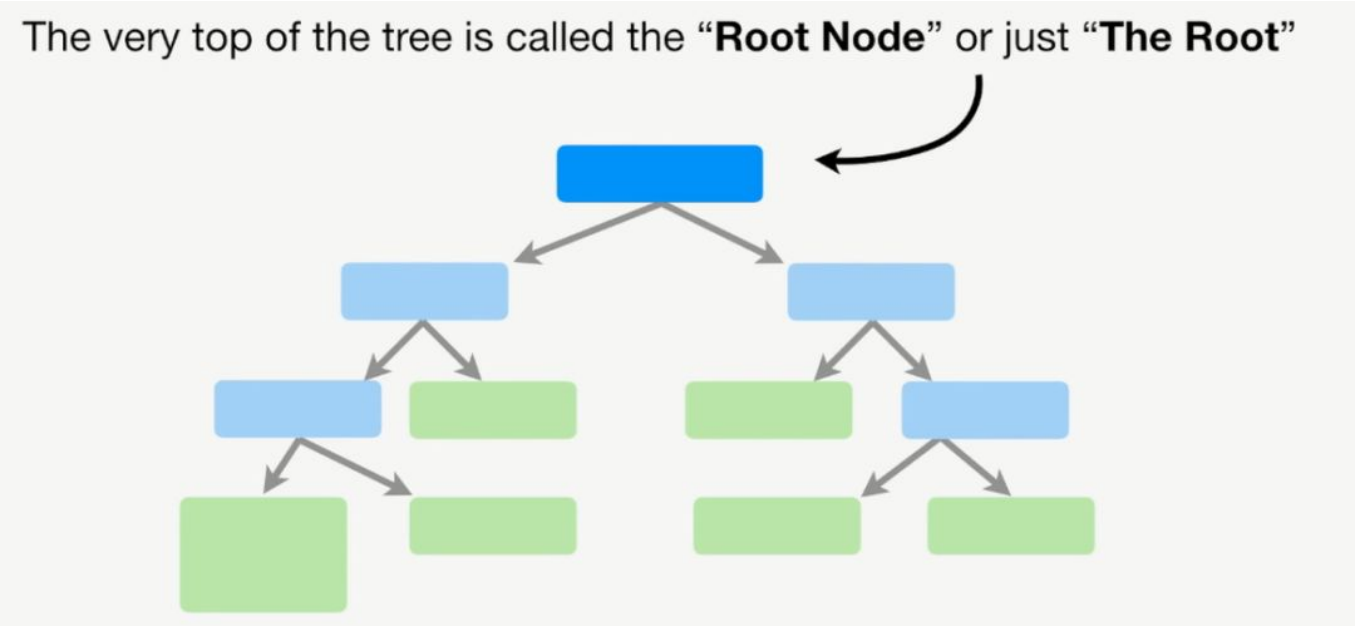



一直到最底层为止，此时就对一个样本进行了分类，如下所示：

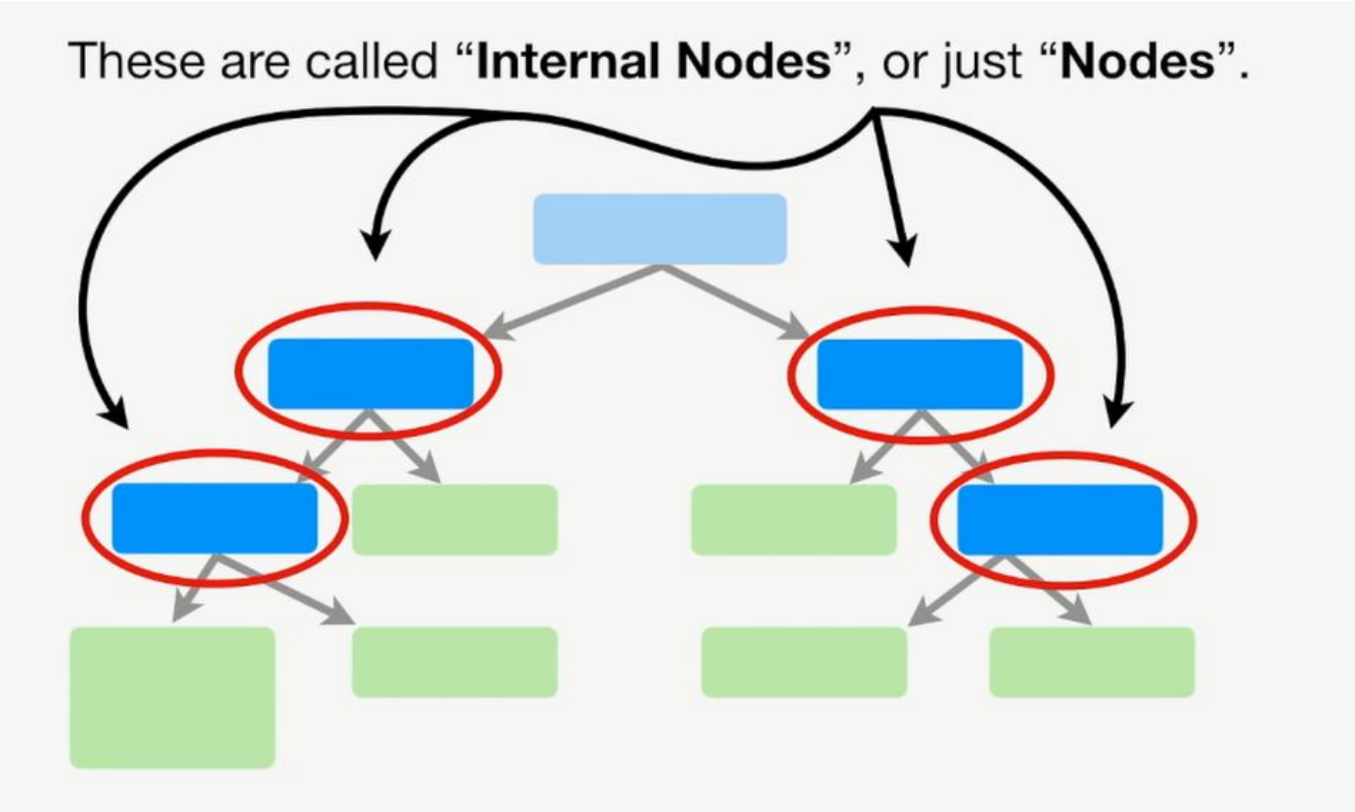


# 决策树术语

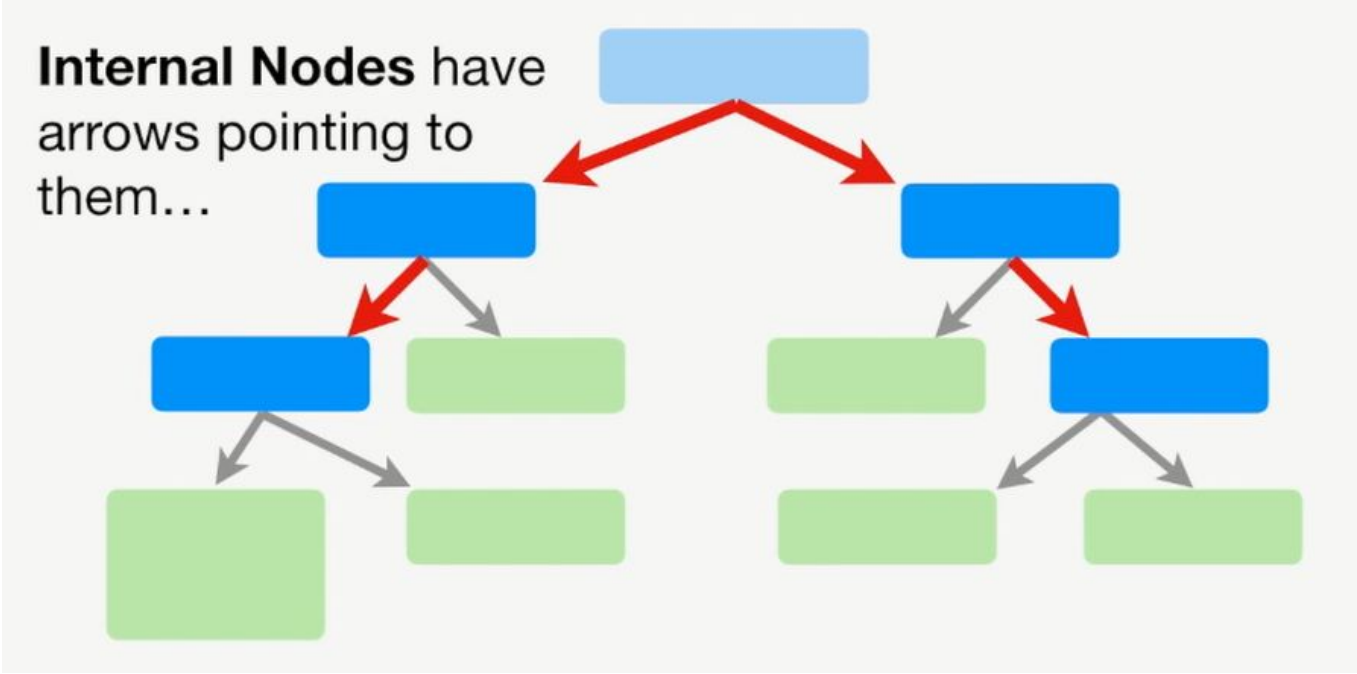
我们来看一下决策树的相关术语，我们把树最顶层的部分称为`根结点（Root Node）`或直接称为`根（The root）`。如下所示：



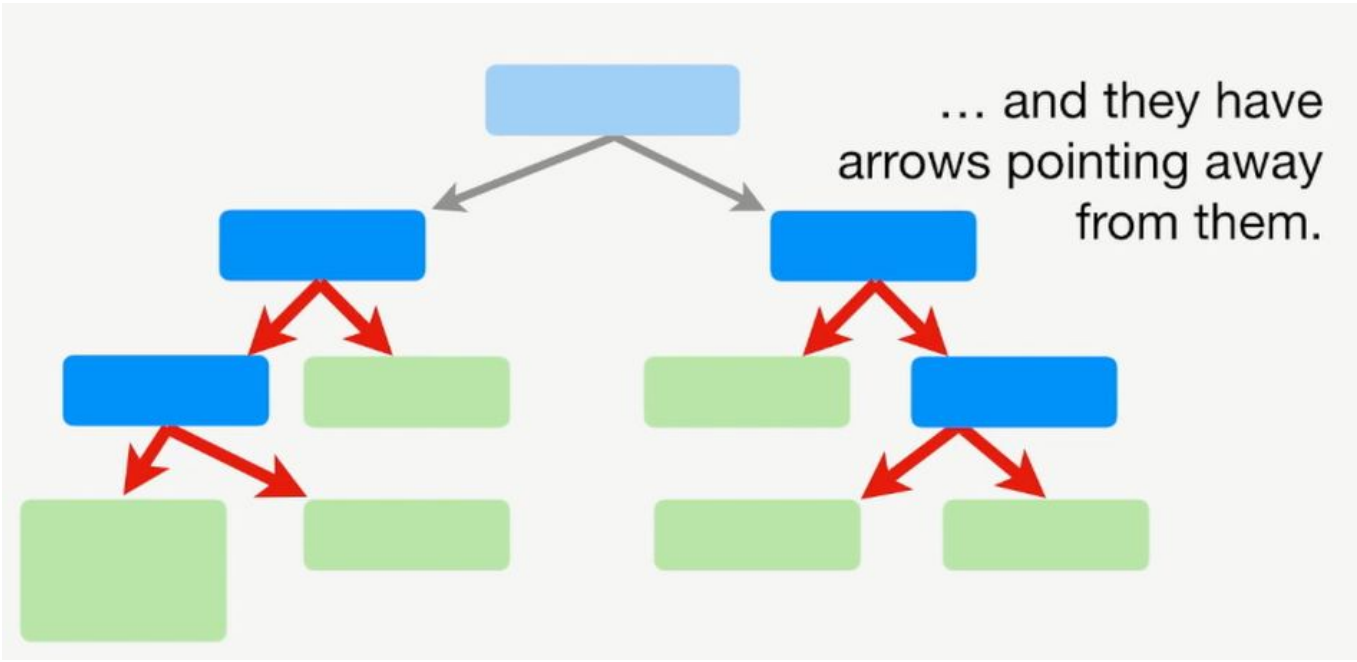

把根节点以下最底层之间的部分称为`内部节点（Internal Nodes）`或`节点（Nodes）`，如下所示：



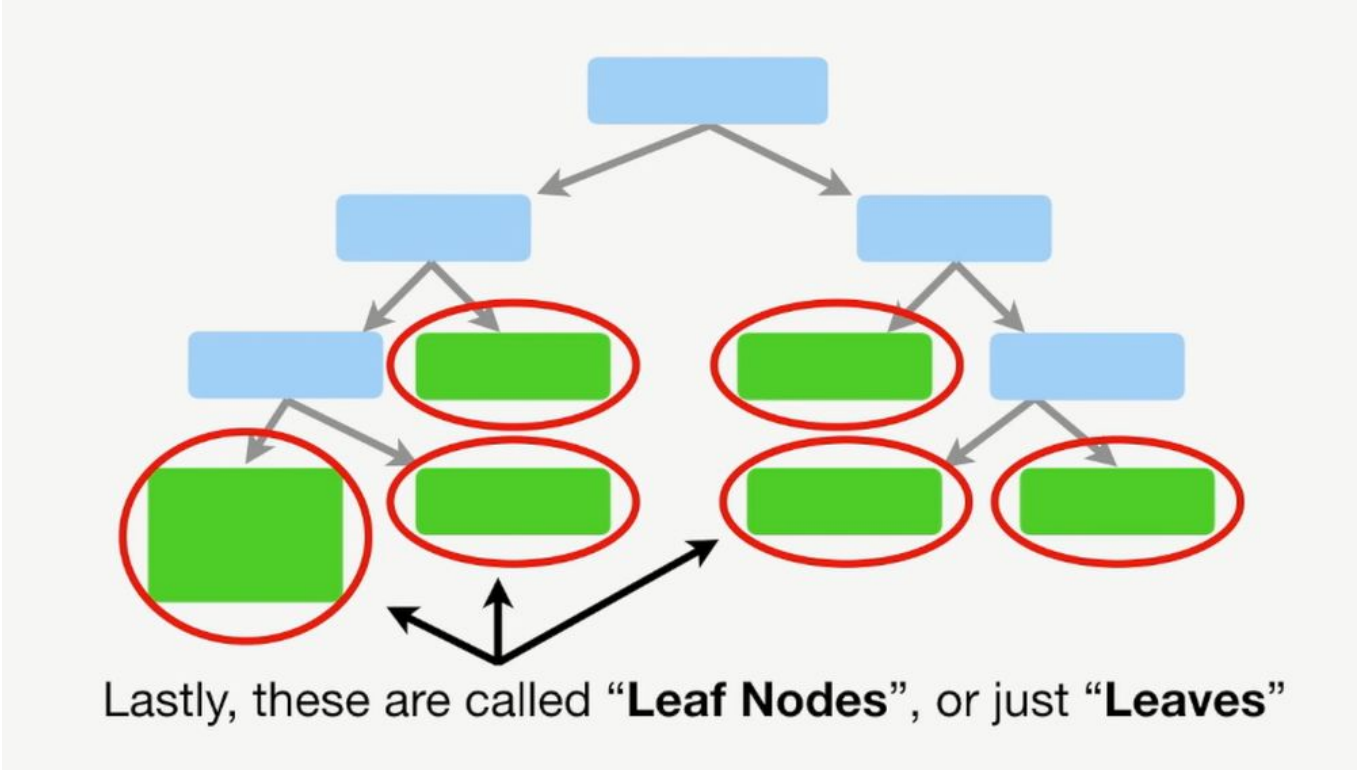
内部节点通过箭头指向下一级的内部节点，如下所示：



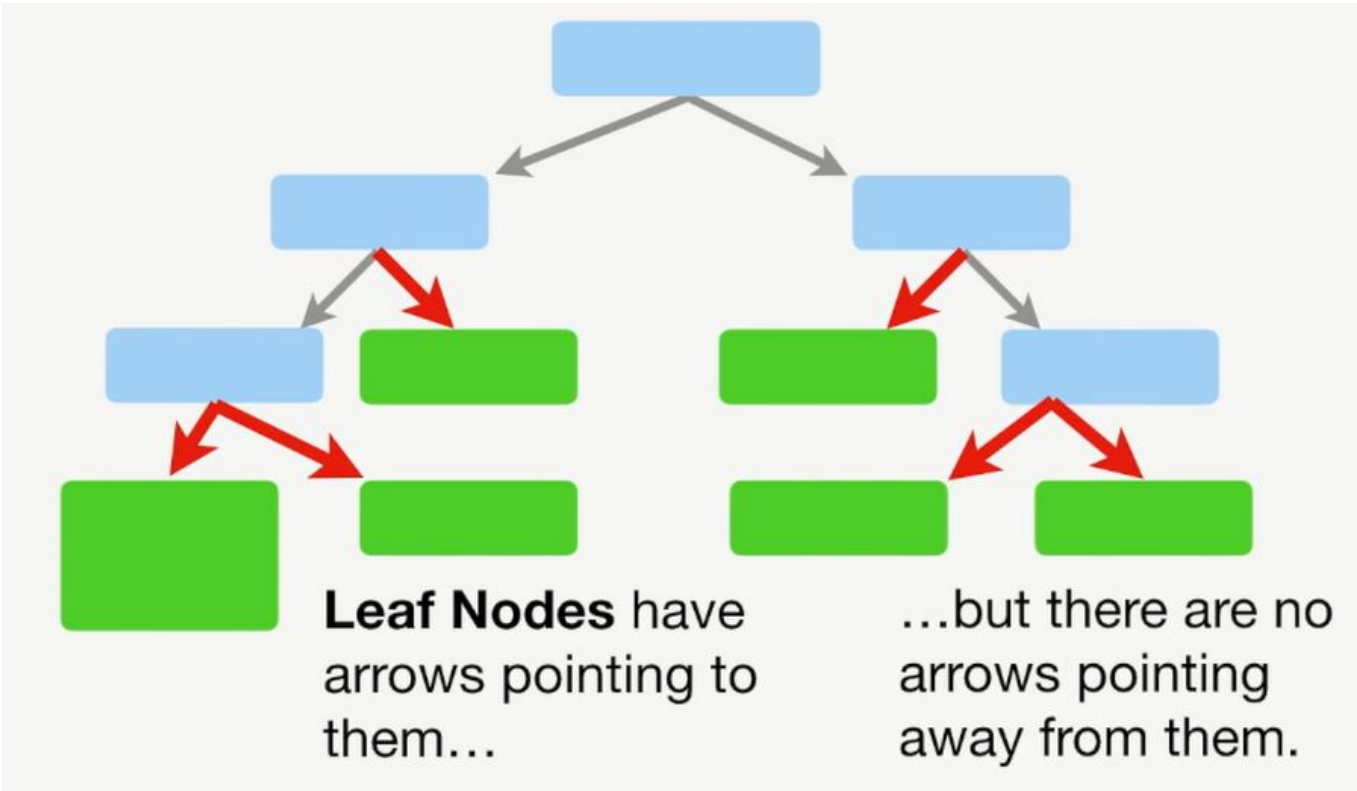

或者是指离（point away）下一层的内部节点，如下所示：




那些终点（也就是没有下一层的节点）被称为`叶子节点（Leaf Nodes）`或`叶子（Leaves）`，如下所示：

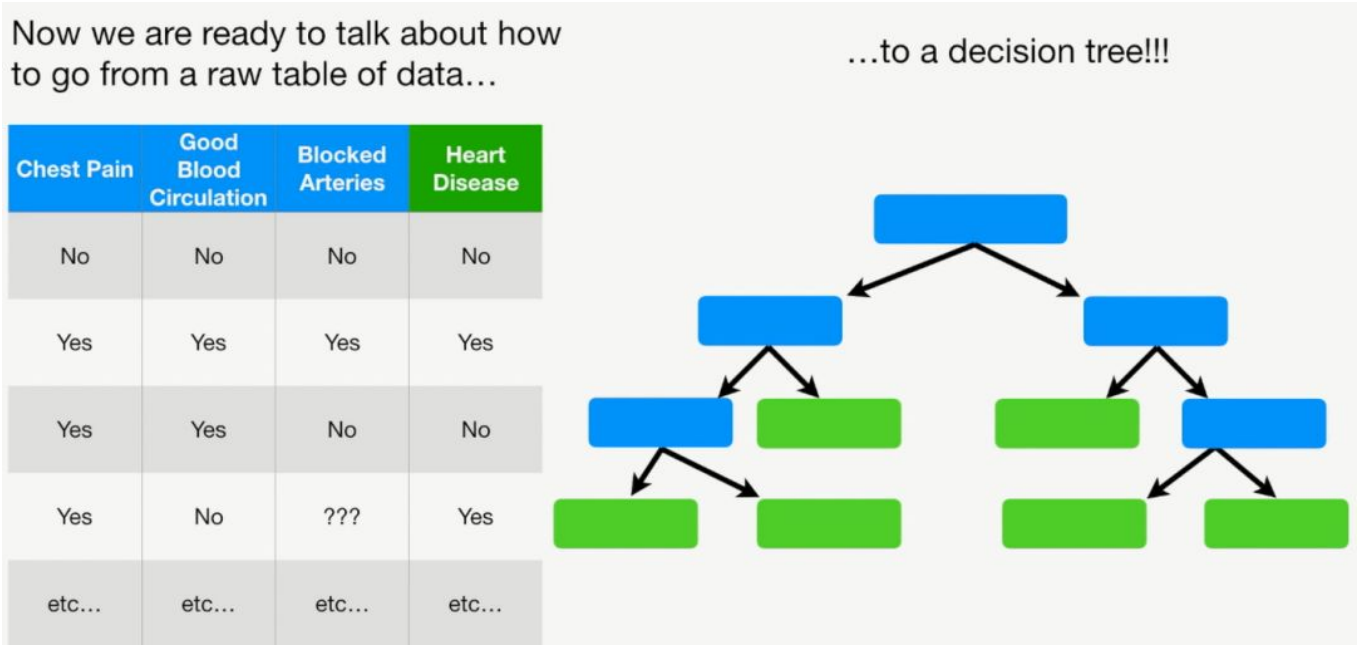



内部节点的箭头指向叶子，但叶子自身已经没有箭头指向其他地方，如下所示：

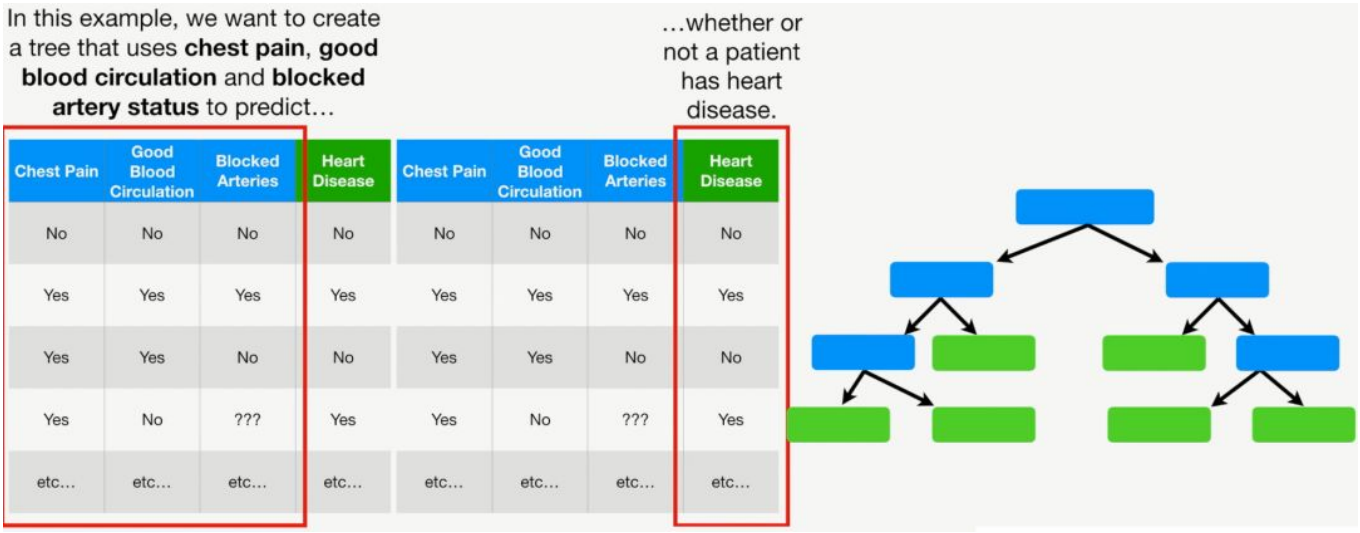



## 构建决策树

现在我们讲一下，如何利用一批数据构建决策树，如下所示：

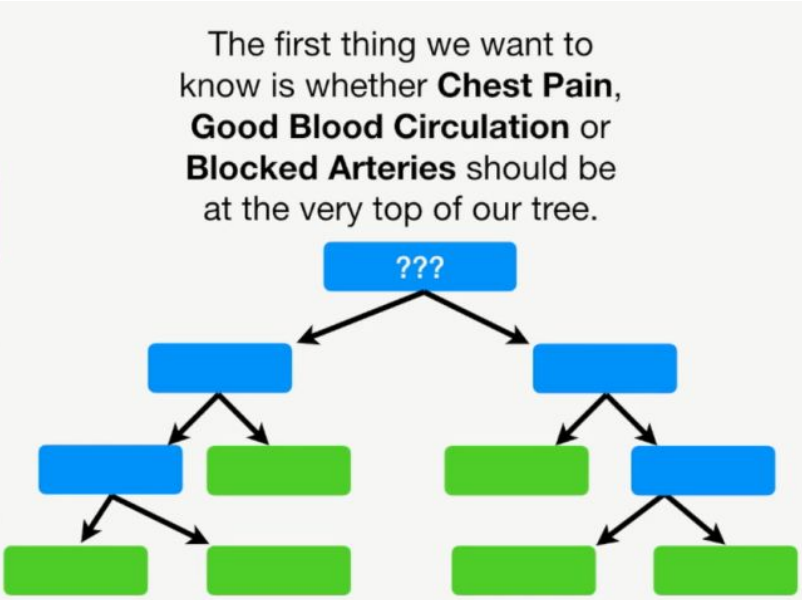



我们会利用这个表中的 `胸痛（chest pain）`、 `良好血液循环（good blood circulation）`和 `动脉阻塞状态（blocked artery status）`构建一个决策树来预测一个患者是否患有心脏病，如下所示：



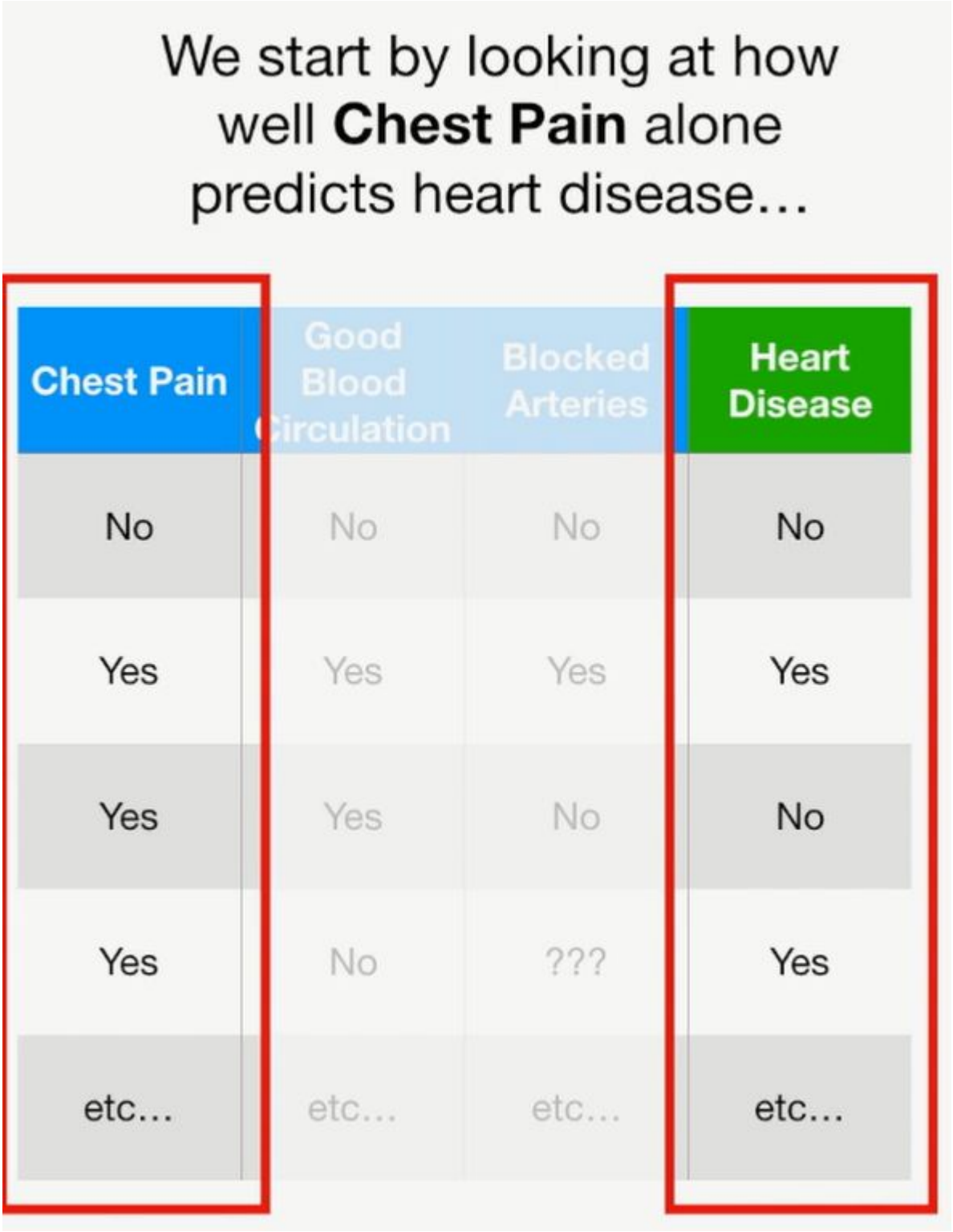

第一步我们需要知道的是，我们使用哪些变量作为根结点，是 `胸痛（chest pain）`、 `良好血液循环（good blood circulation）`，还是 `动脉阻塞状态（blocked artery status）`，如下所示：

| 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... |

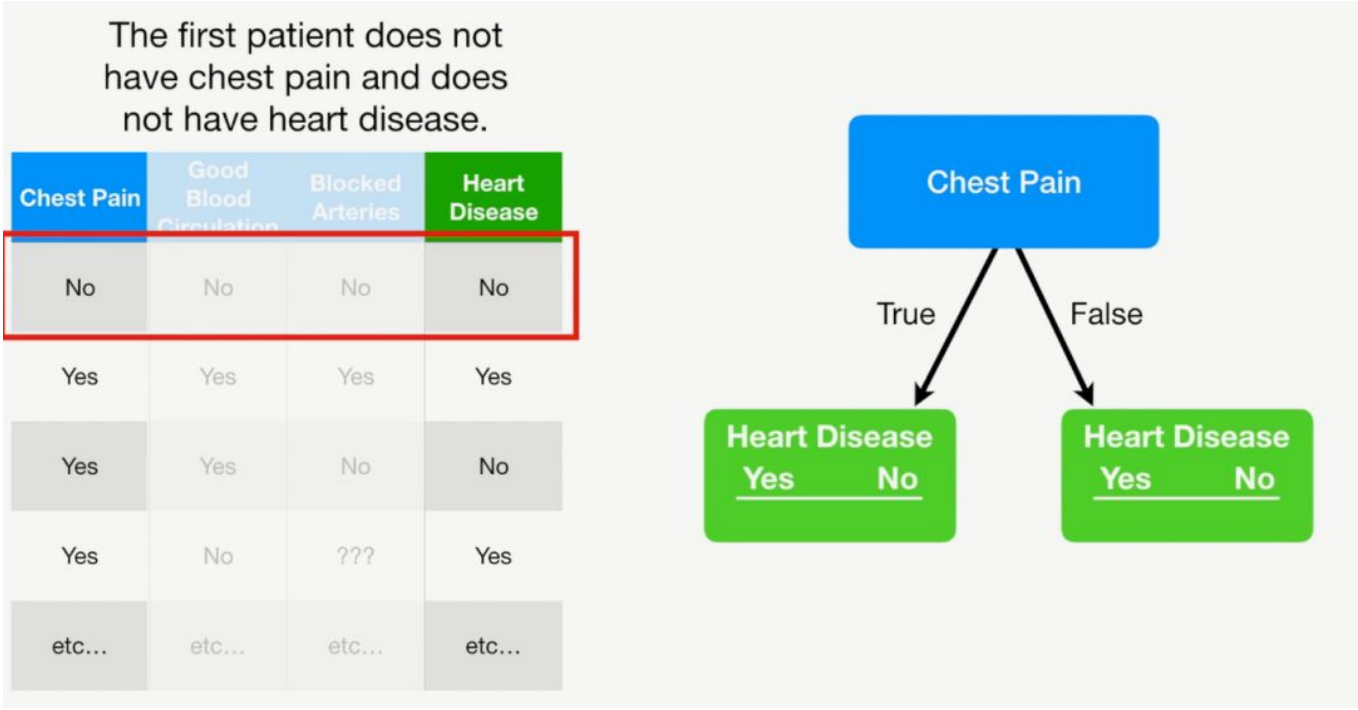



## 单一变量1构建决策树

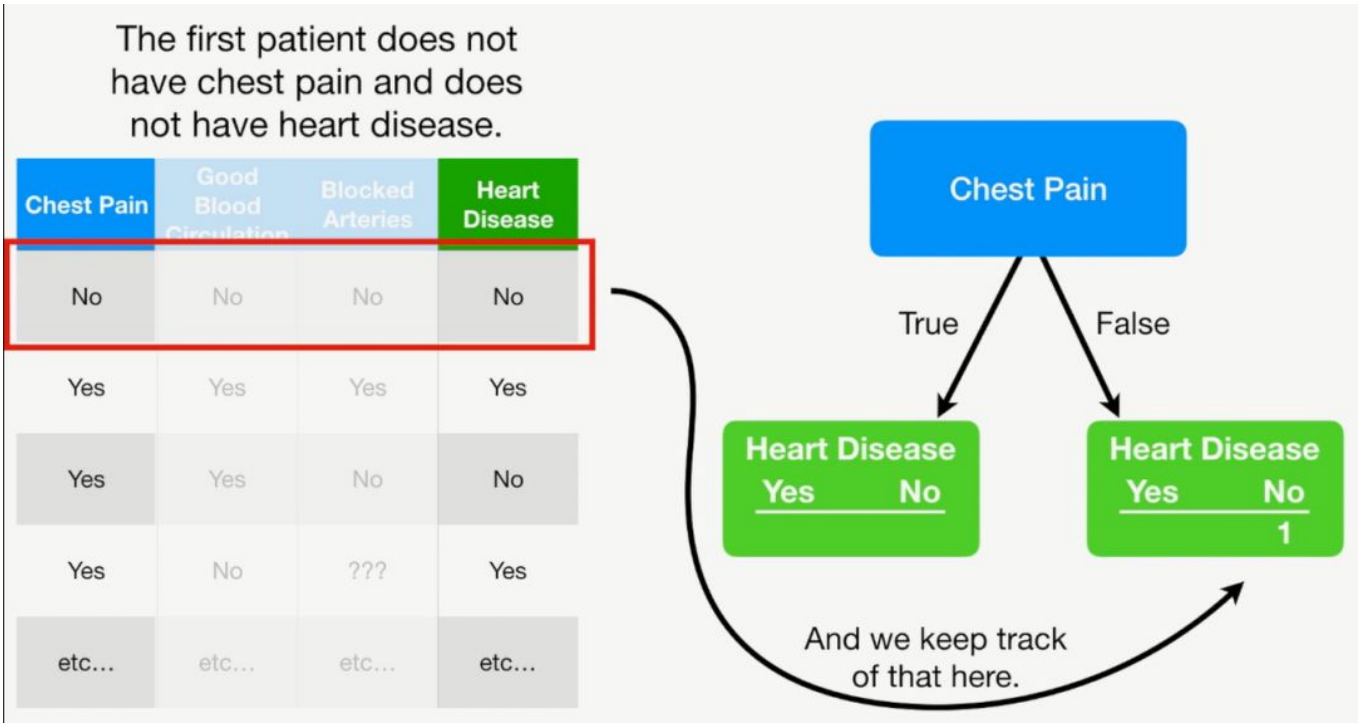
此时，我们先看一下，`胸痛（chest pain）`这一个变量预测心脏病的效果如何，如下所示：



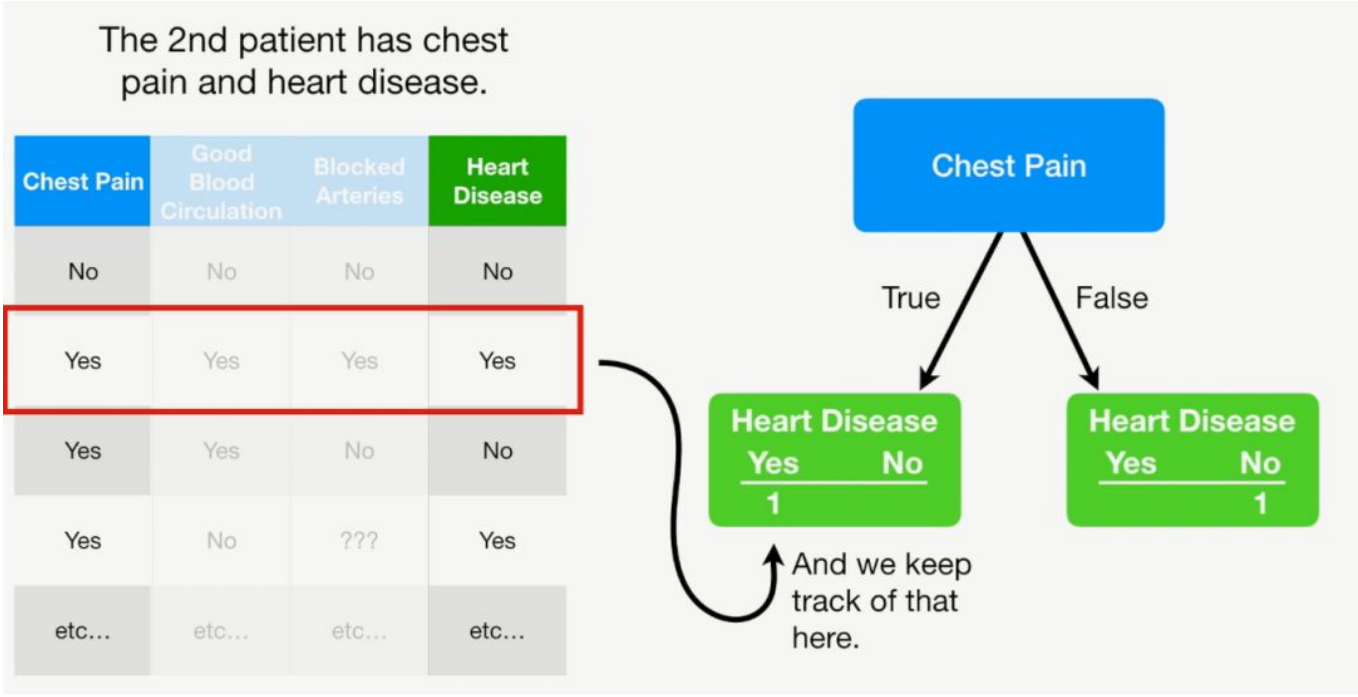

先用这一个变量（胸痛（chest pain））来构建一个小的决策树，如下所示：



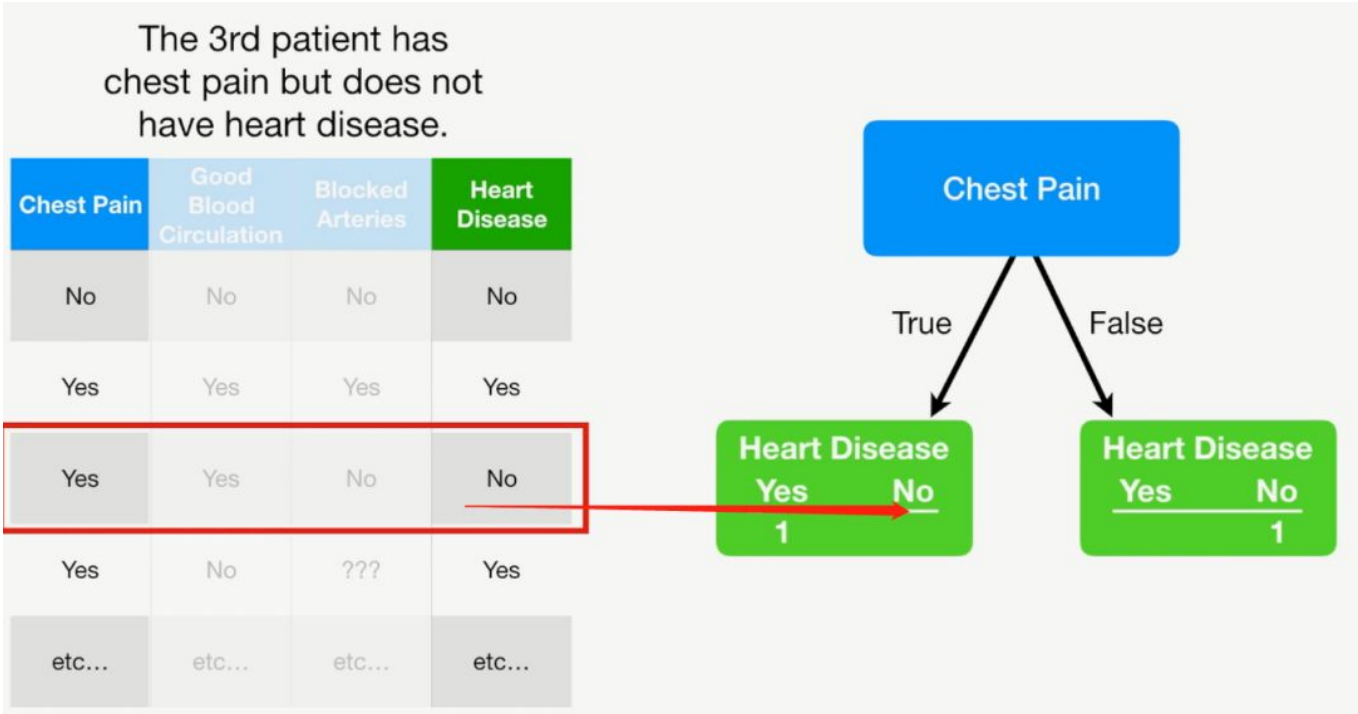

先看第一行的数据，这个患者没有出现胸痛，也没有心脏病，对应到决策树上就是右侧的`False`中的`No`，如下所示：



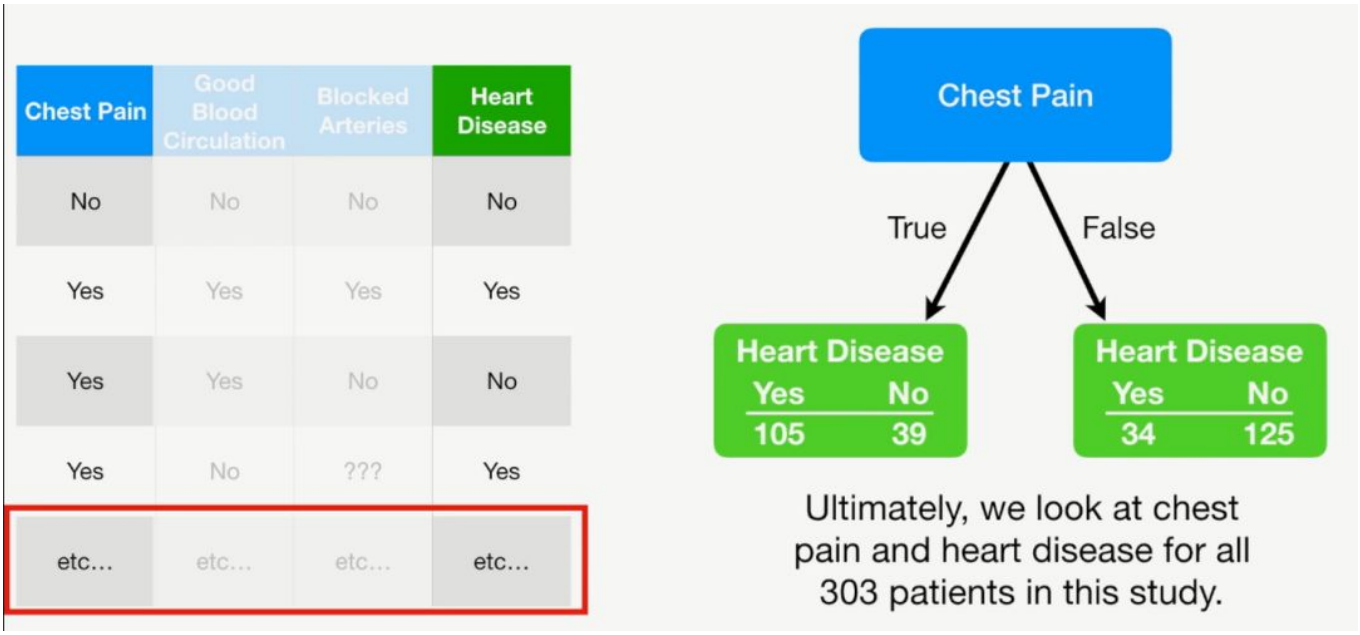

再看第2行，也就是第2个患者，他出现了胸痛，也出现了心脏病，就归于左侧的`True`的`Yes`这一类，如下所示：




再看第3行，也就是第3个患者，他出现了胸痛，但没有出现心脏病，就归到了左侧`True`的`No`这个分类，如下所示：

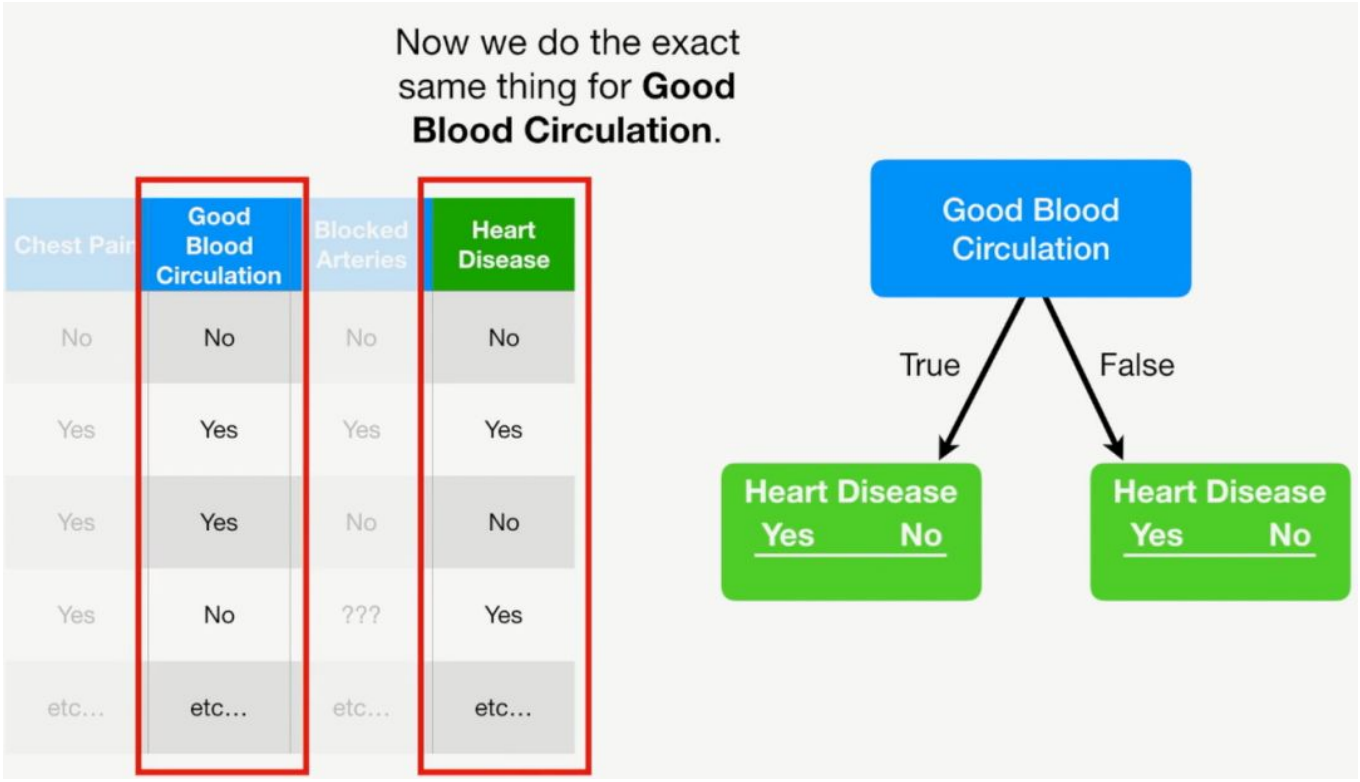



把所有的患者（303个）都统计后，会得到最终是否得心脏病的统计信息，如下所示：

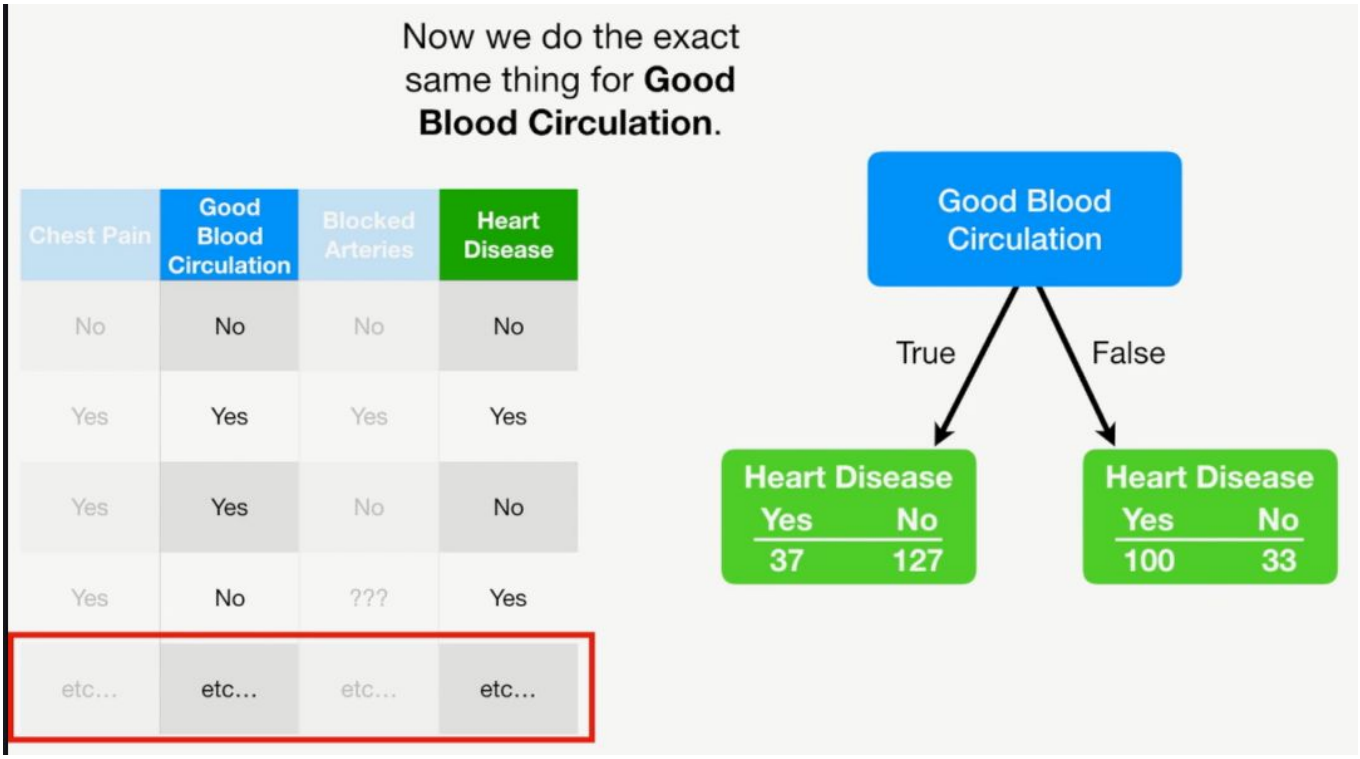



## 单一变量2构建决策树

现在我们使用`良好血液循环（good blood circulation）`这个变量来构建小的决策树，整个过程与前面的一样，如下所示：

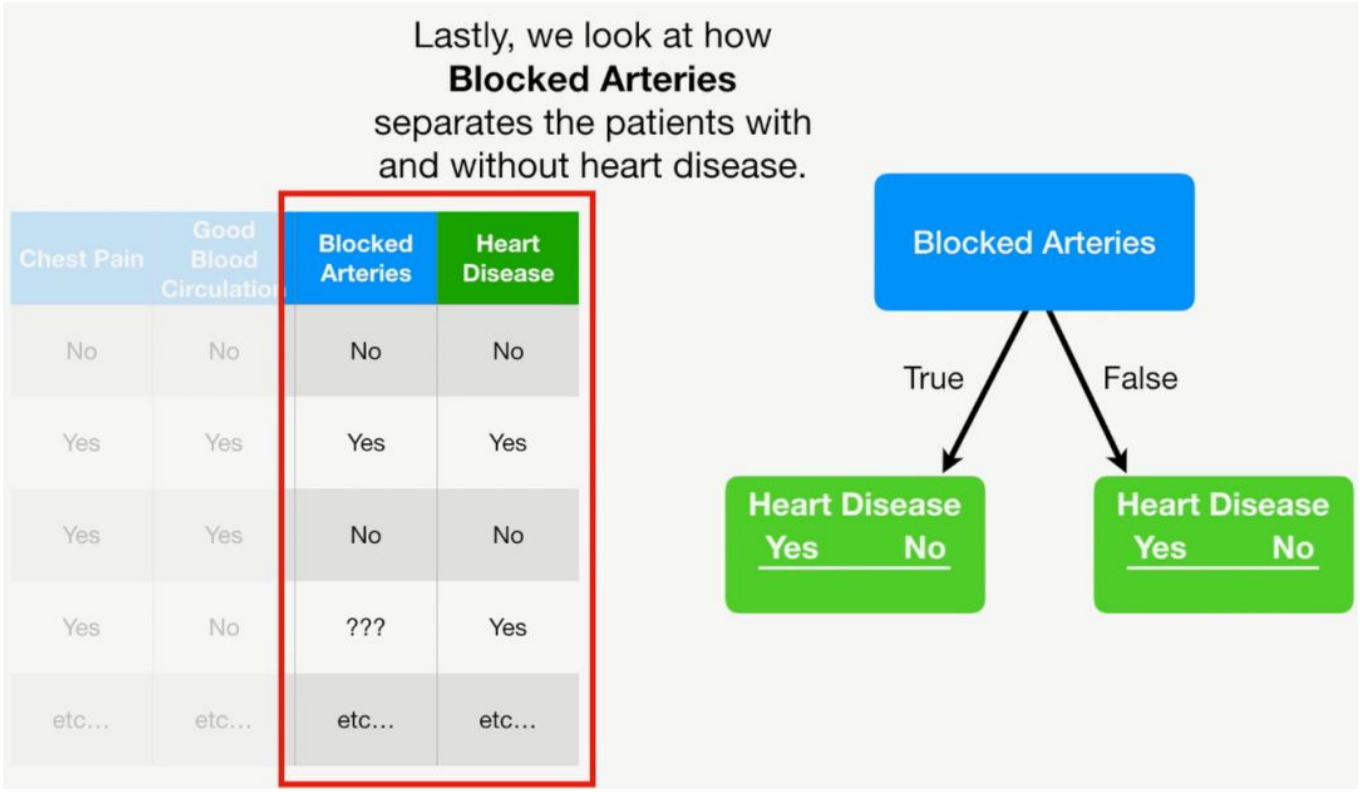



最终得到的结果如下所示：

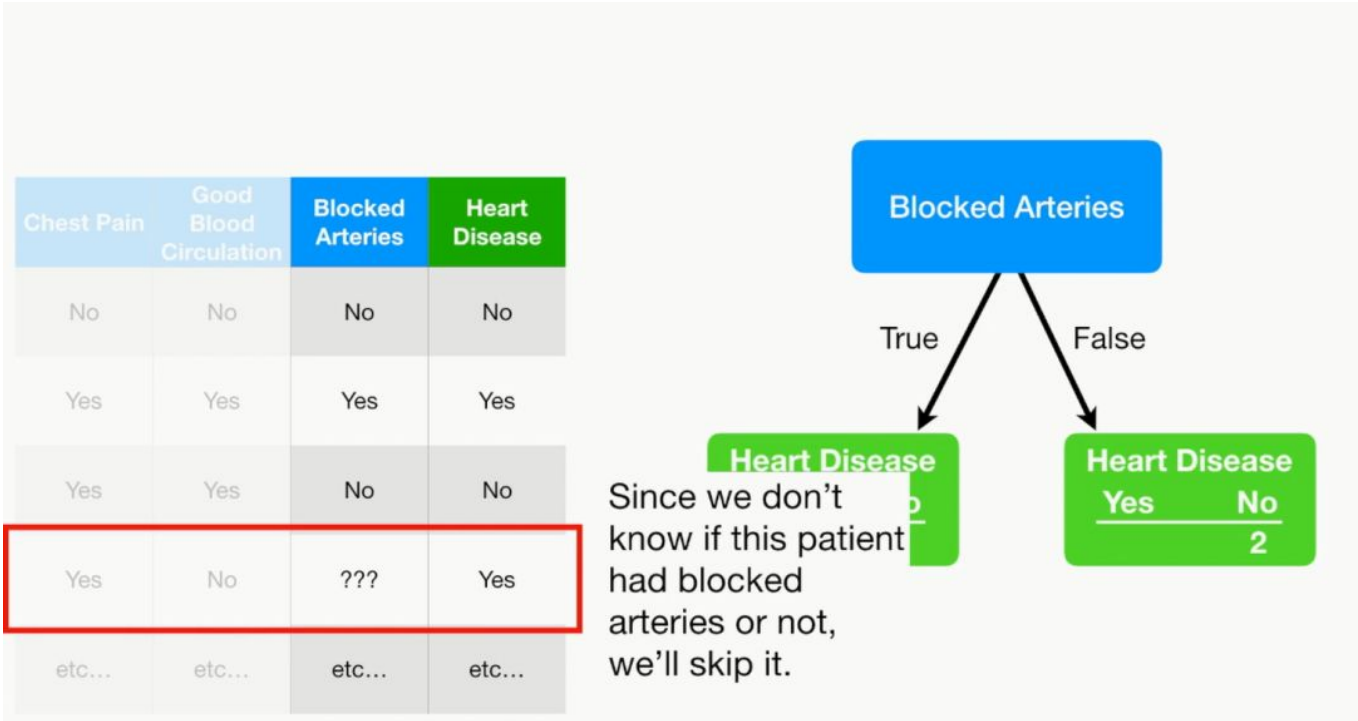



## 单一变量3构建决策树

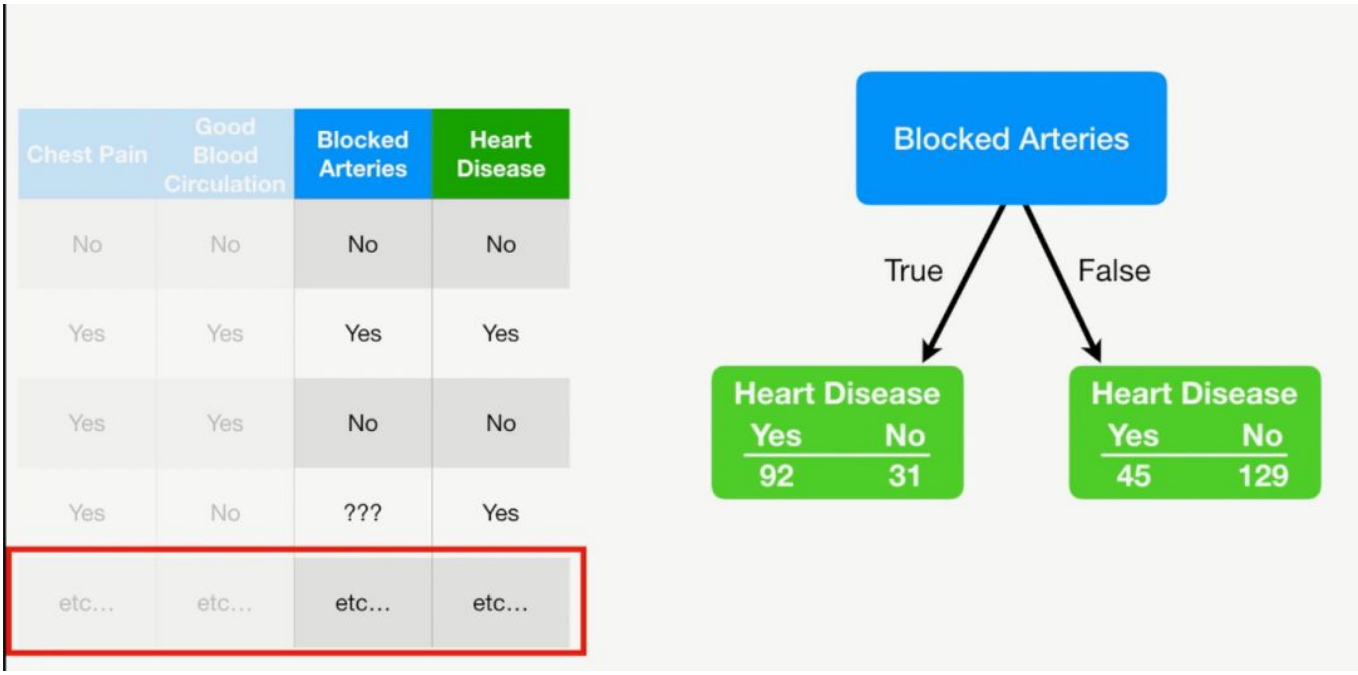
现在我们使用`动脉阻塞状态（blocked artery status）`这个变量来构建小的决策树，整个过程与前面的一样，如下所示：



不过这个变量中有一个不明（也就是说除了Yes或No外，还有一个其他的结果，就是三个问号`???`，这个我们后面会继讲到如何处理），如下所示：

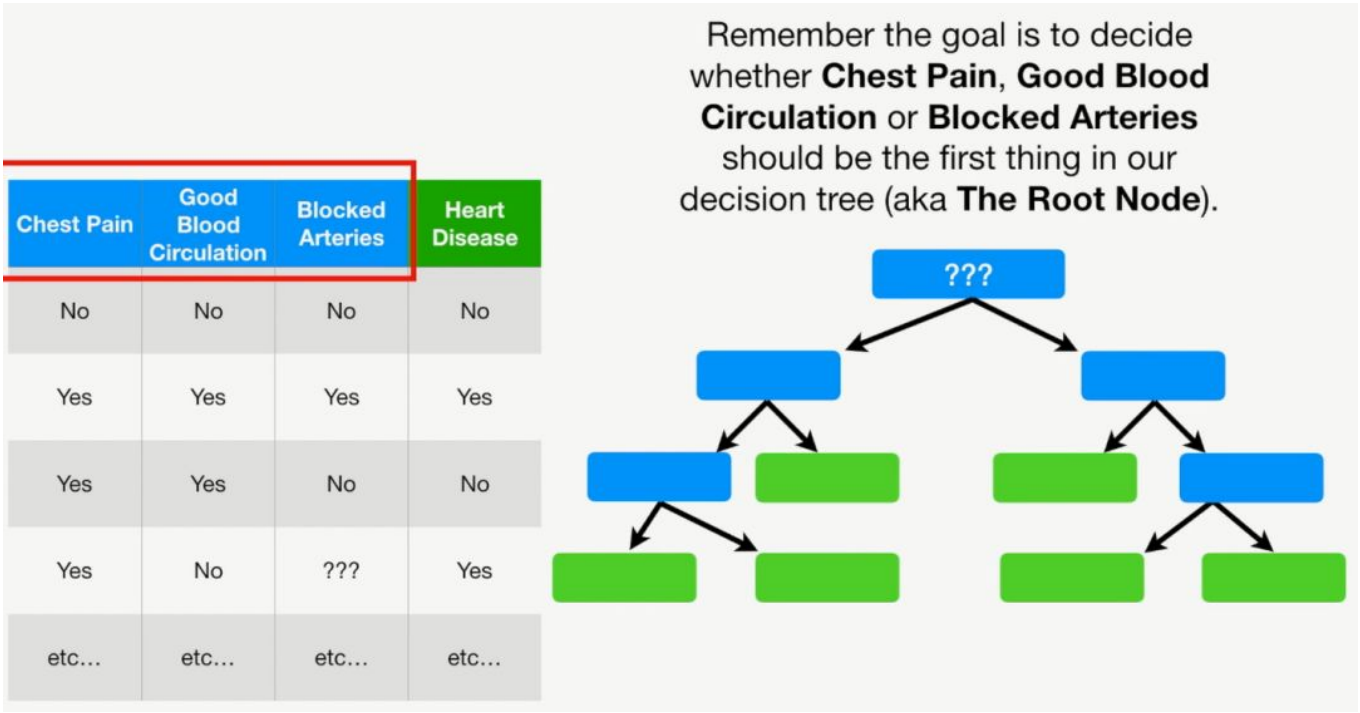



最终的结果如下所示：：

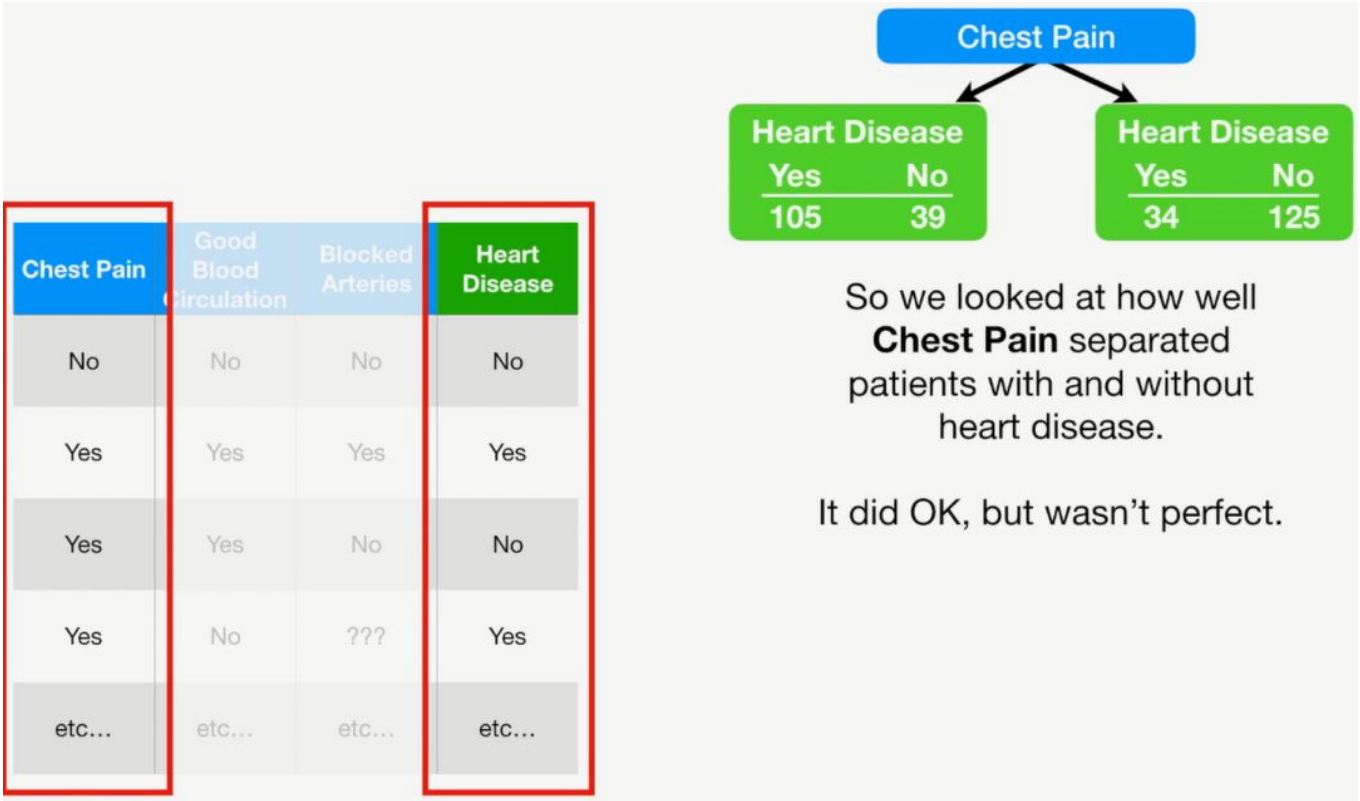



## 根节点设置

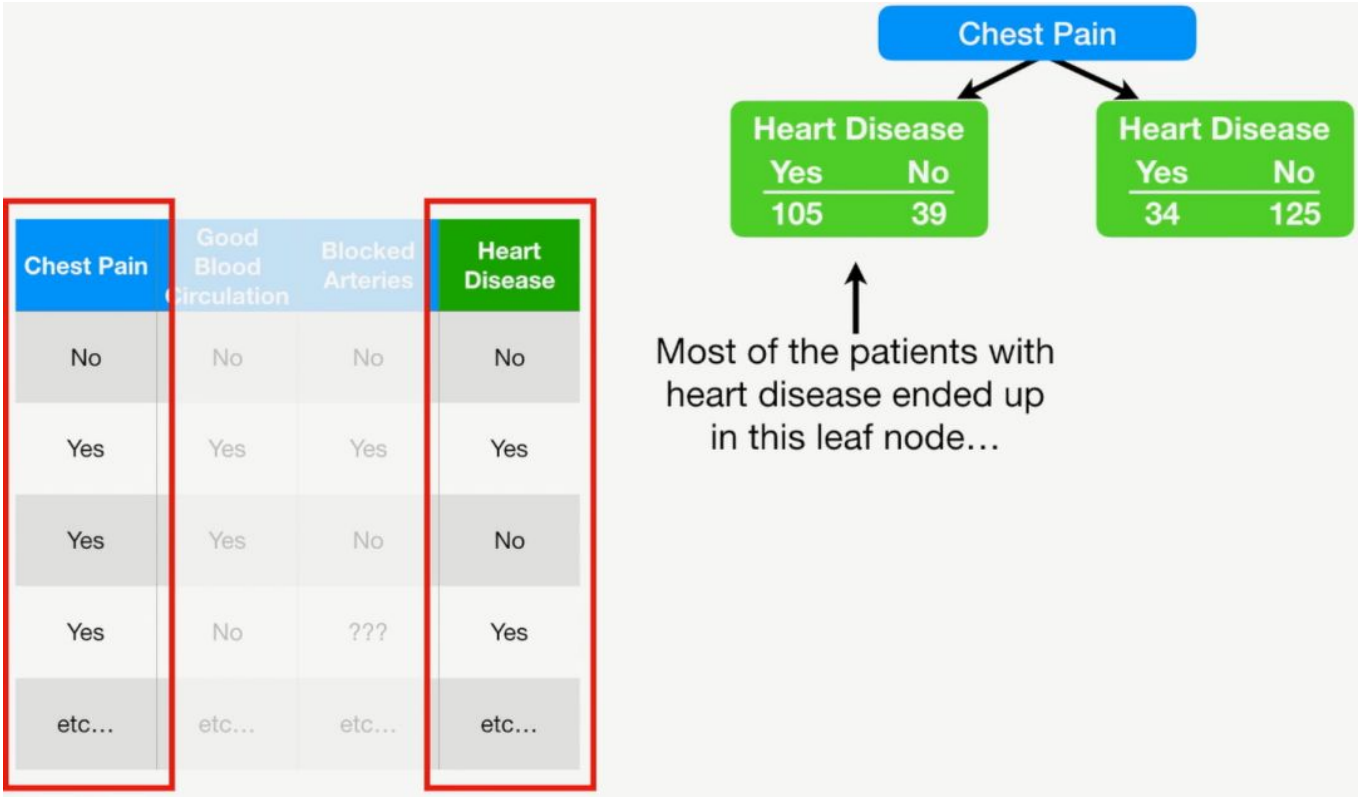
前面我们使用了三个不同的变量，即胸痛（chest pain）、 良好血液循环（good blood circulation）和 动脉阻塞状态（blocked artery status）构来构建小的决策树，我们最终的目标就是要找到使用哪个变量做为根节点，如下所示：



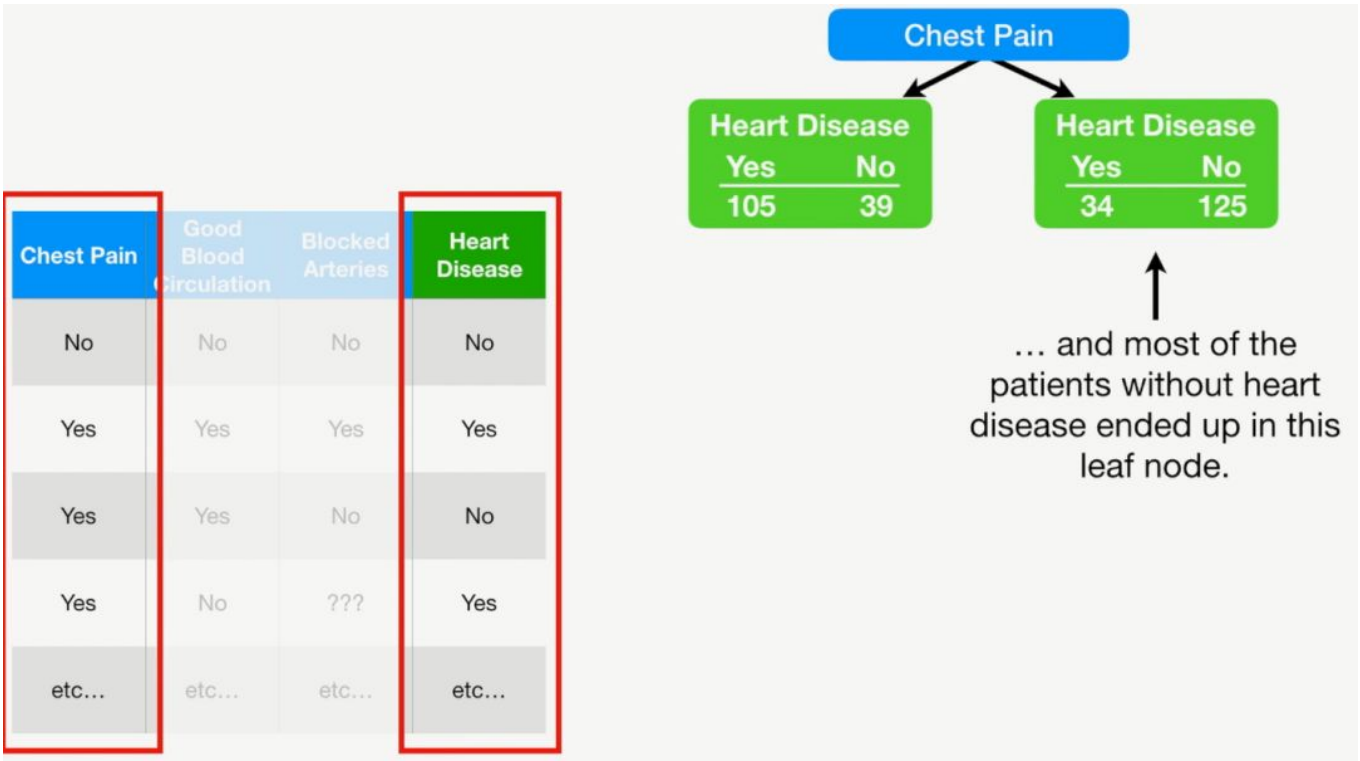

此时我们看一下胸痛（chest pain）这个变量对是否患有心脏病的区分效果，虽然能够区分，但是效果并非完美，如下所示：



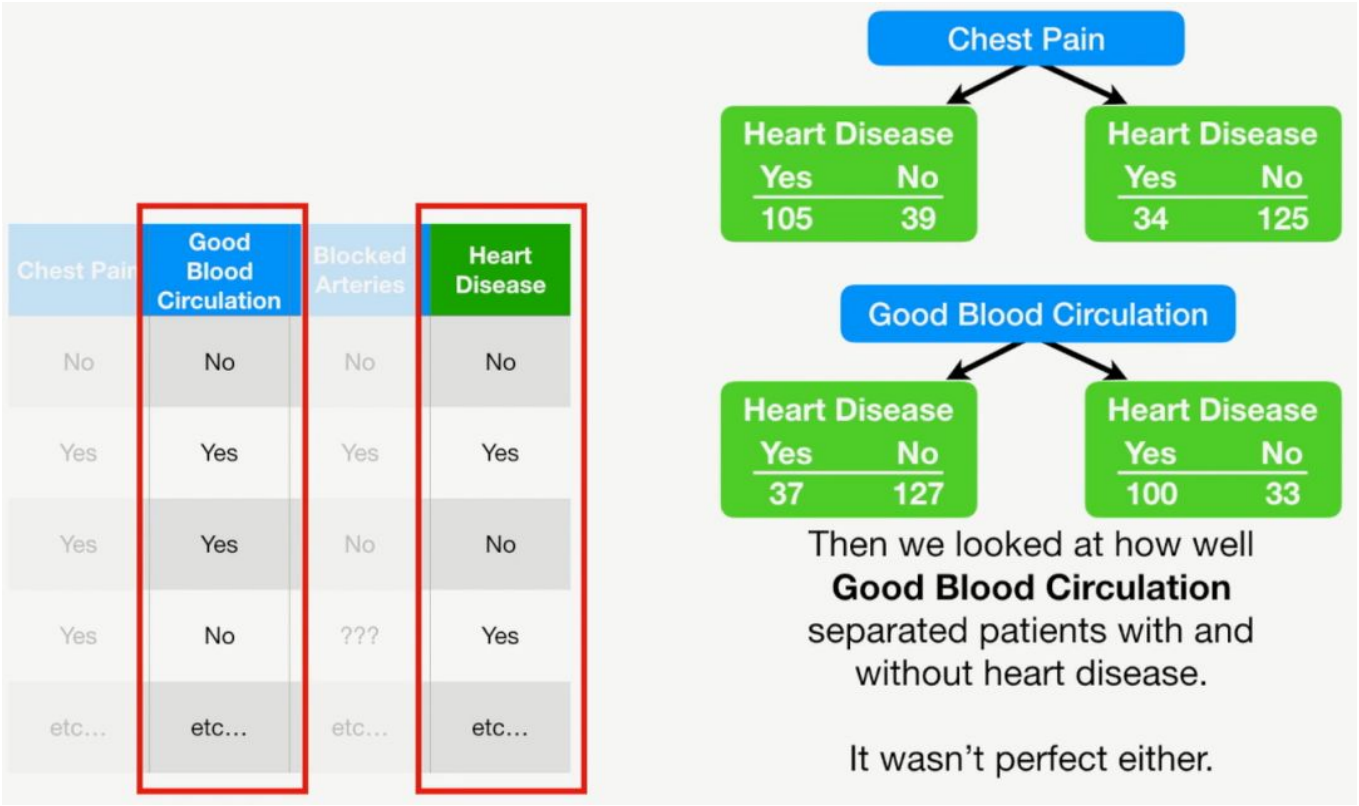
因为那多数患有心脏病的患者最终会落到 `Yes` 这个叶子节点上来，大多数没有心脏病的患者会落后右侧 `No` 这个节点上来，我们看一下患有心脏病患者的节点情况，如下所示：



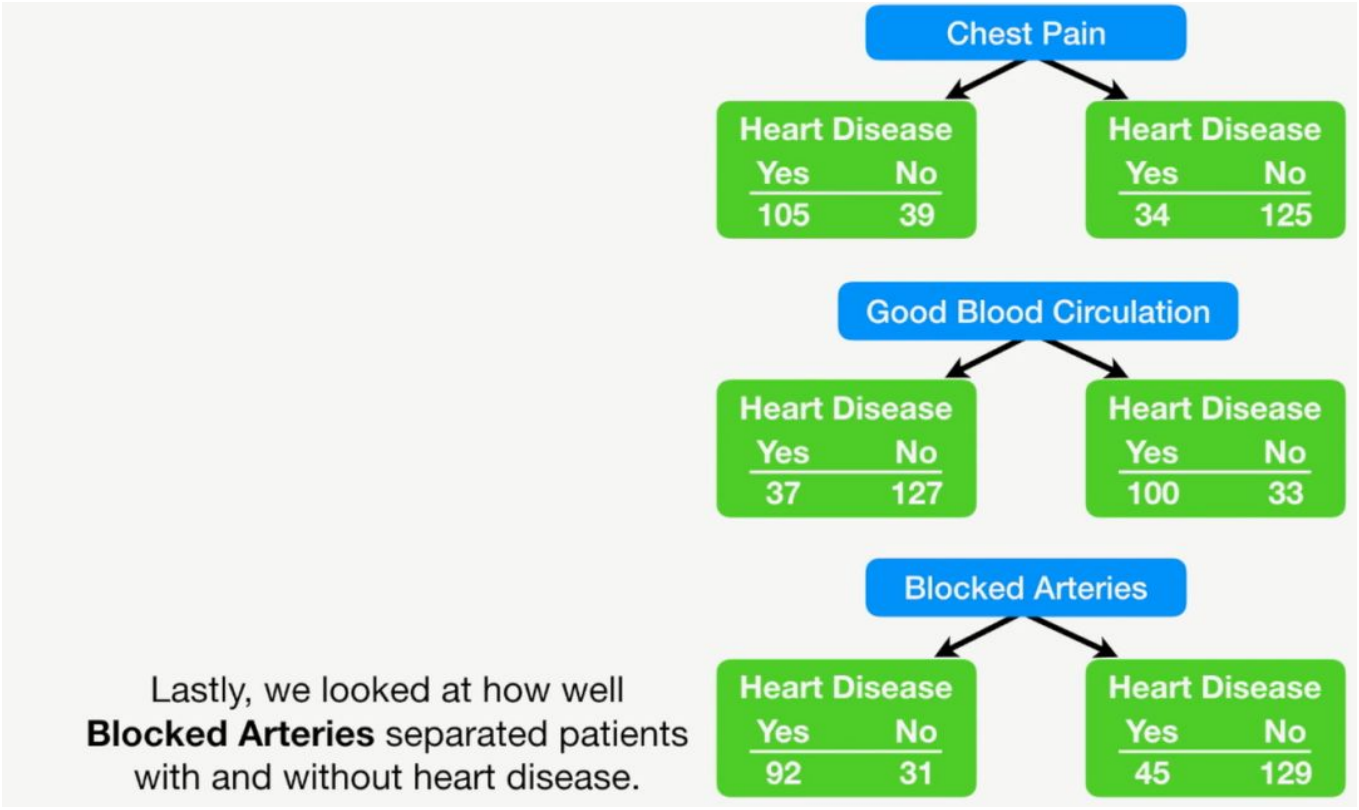

再看一下没有患心脏病的患者情况，如下所示：



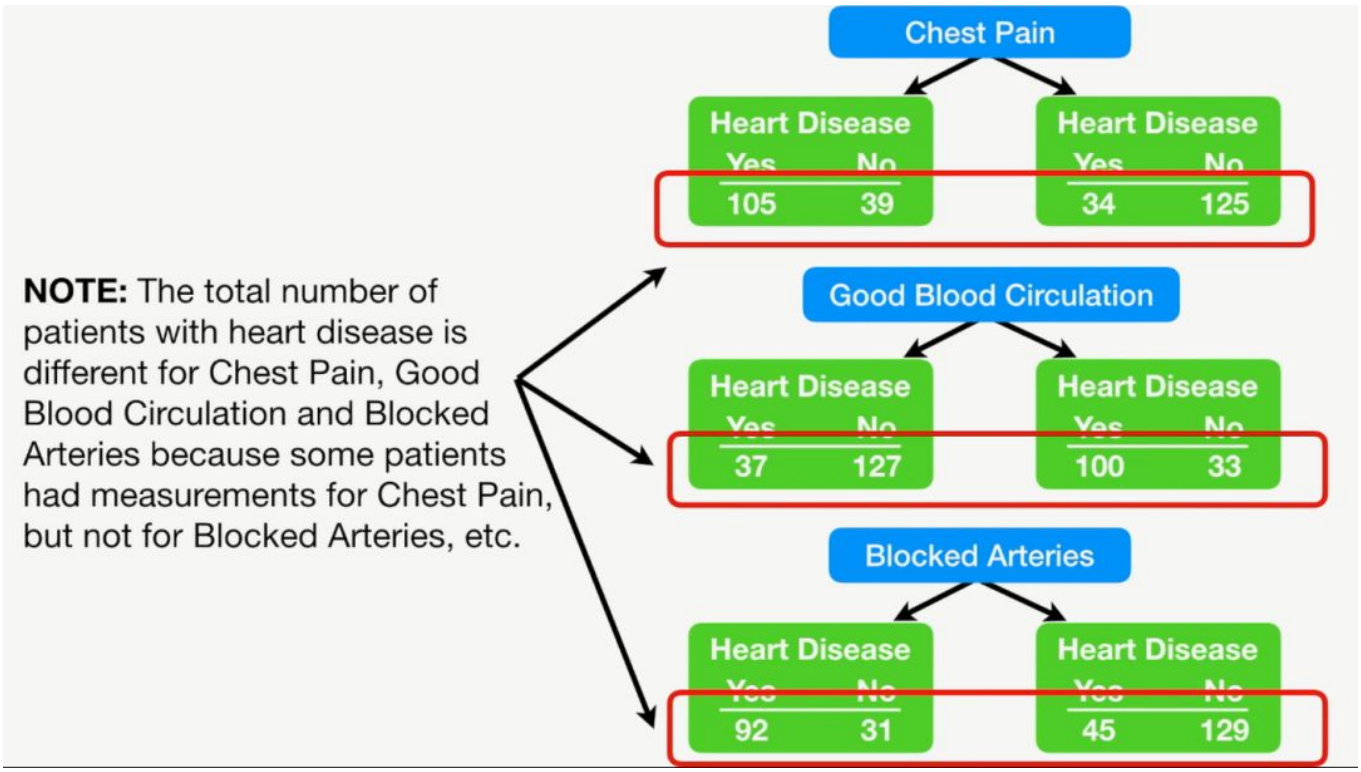
我们再看一下良好血液循环（good blood circulation）这个变量对心脏病患者的区分情况，它也不是特别的完美，如下所示：




最终，我们看一下动脉阻塞状态（blocked artery status）这个变量对患者的区分情况，如下所示：

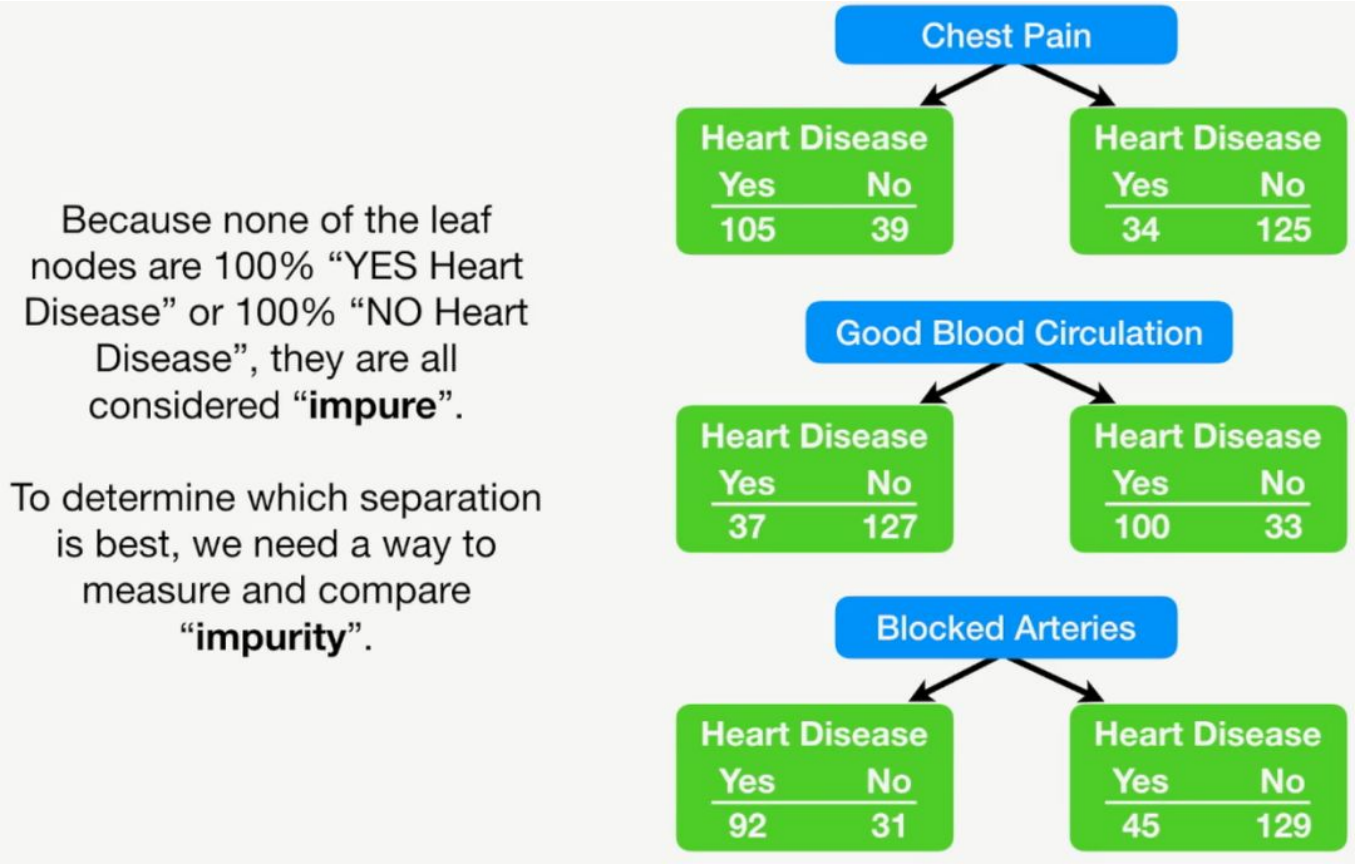



我们发现，这三个不同的变量用于区分患者是否患有心脏病的数目是不同的，如下所示：

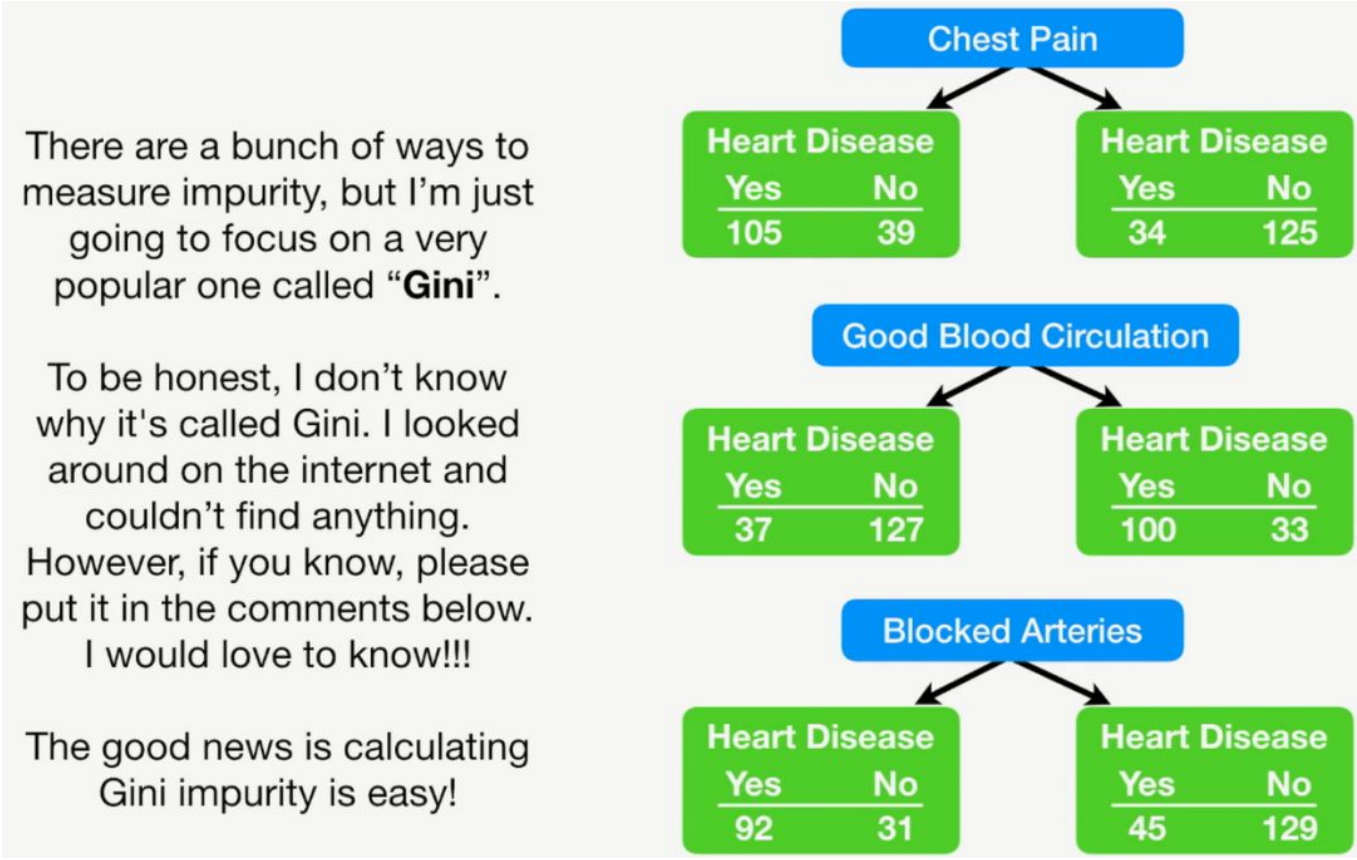



## 基尼不纯度（Gini impurity）

从中我们可以发现，没有一个叶子节点能够显示100%的患有心脏病，或者是100%没有患心脏病，因此我们称这些叶子节点不纯（`impure`），为了找到哪个区分的程度更好，我们计算并比较不纯度（impurity），如下所示：

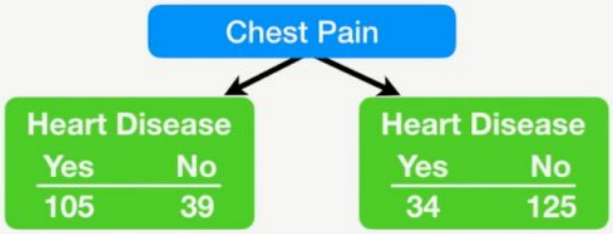



计算不纯度（impurity）的方法很多，其中最常用的方法称为基尼（Gini）不纯度，如下所示：




此时，我们计算一下胸痛（Chest pain）这个变量构建的决策树的基尼不纯度，如下所示：


对于某个叶子结点来说，它的基尼不纯度公式如下所示：

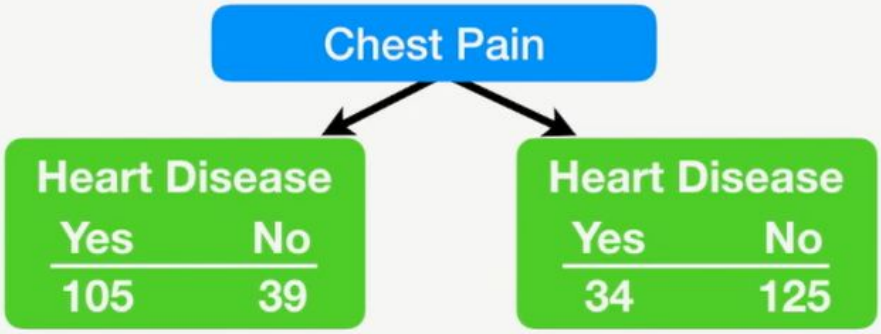



For this leaf, the Gini impurity = 1 - (the probability of "yes")$^2$ - (the probability of "no")$^2$

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

$$= 0.395$$

此时，再计算另外一个叶子结点的基尼不纯度，如下所示：

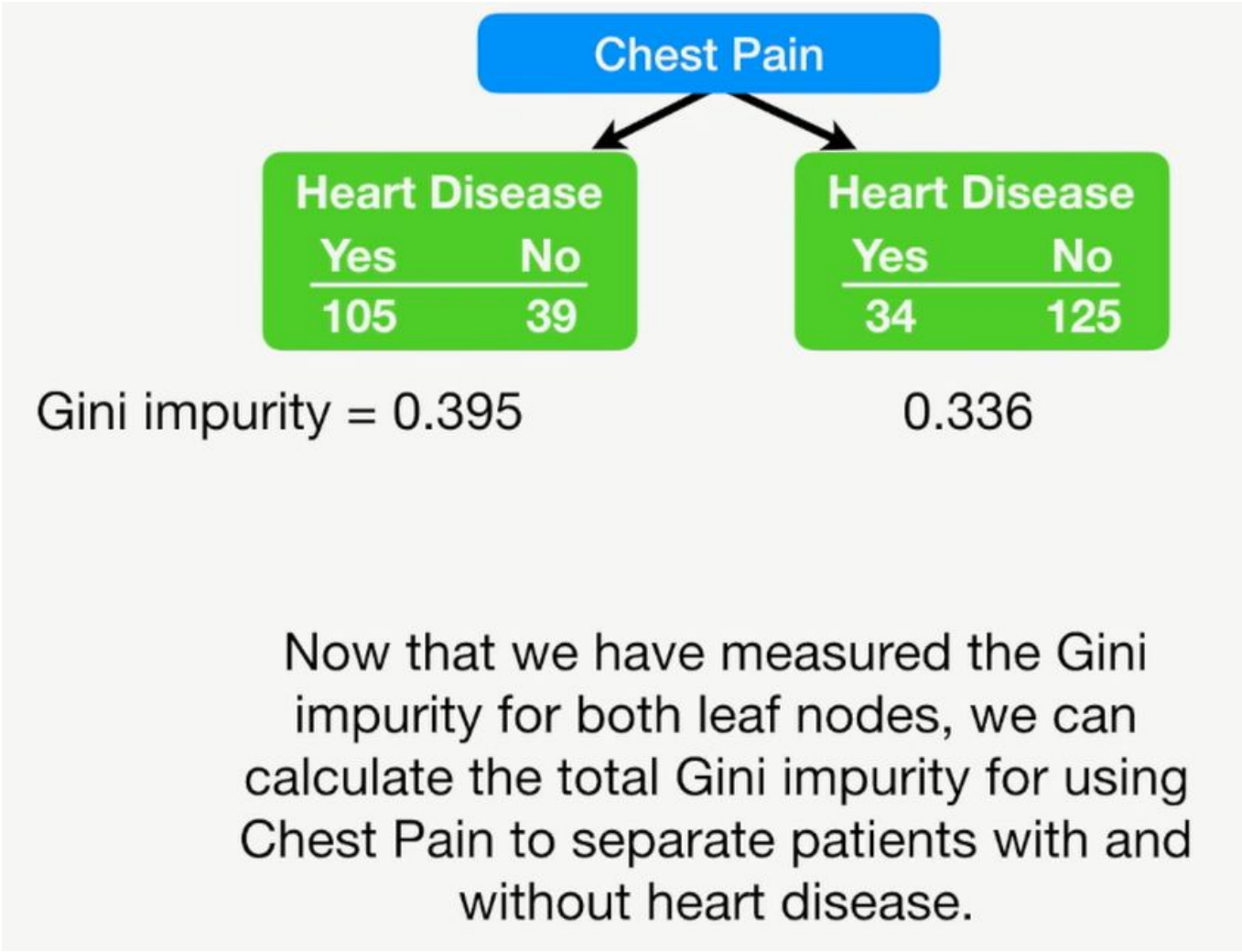



$$= 1 - (\text{the probability of "yes"})^2 - (\text{the probability of "no"})^2$$

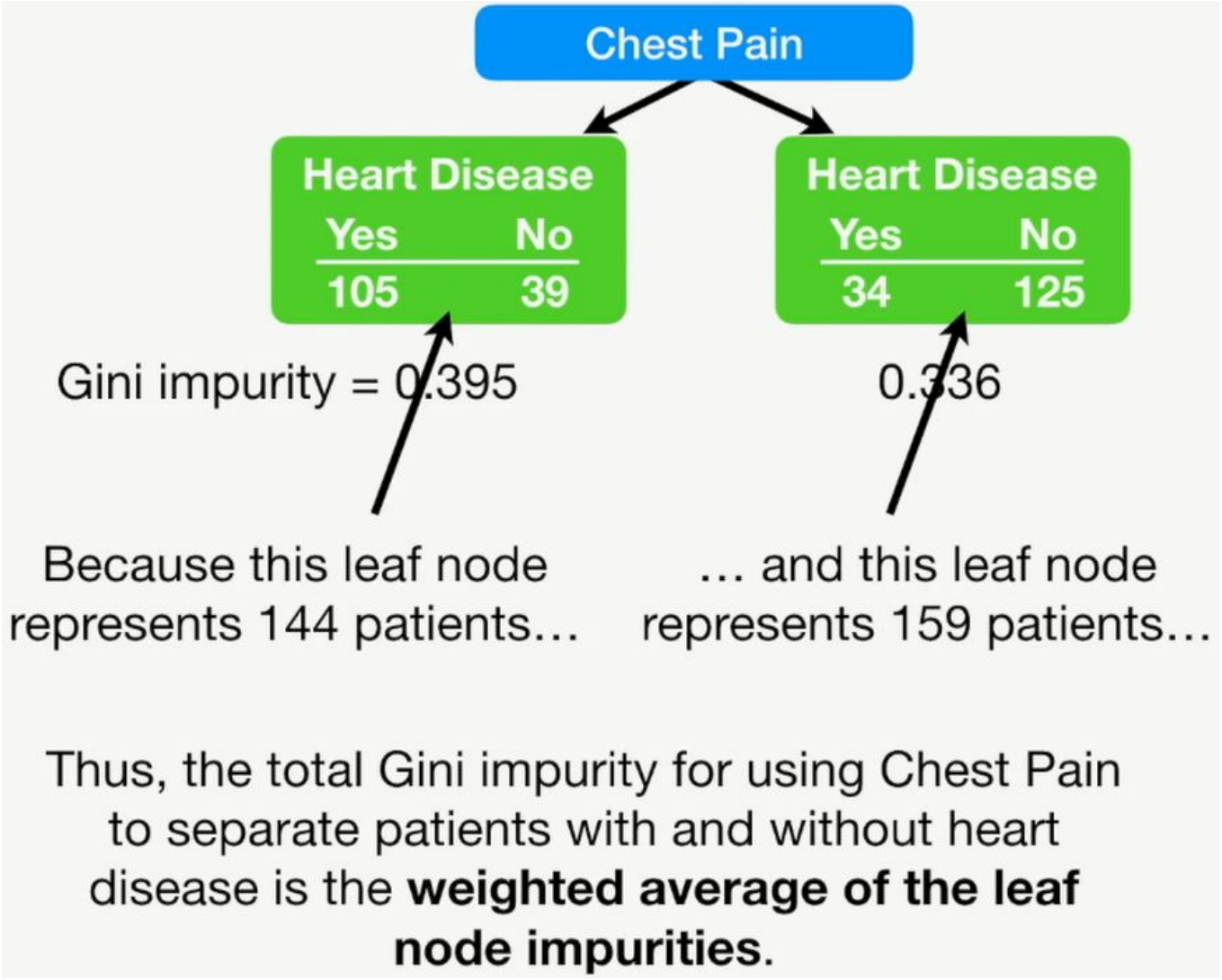
$$= 1 - \left(\frac{34}{34 + 125}\right)^2 - \left(\frac{125}{34 + 125}\right)^2$$

$$= 0.336$$

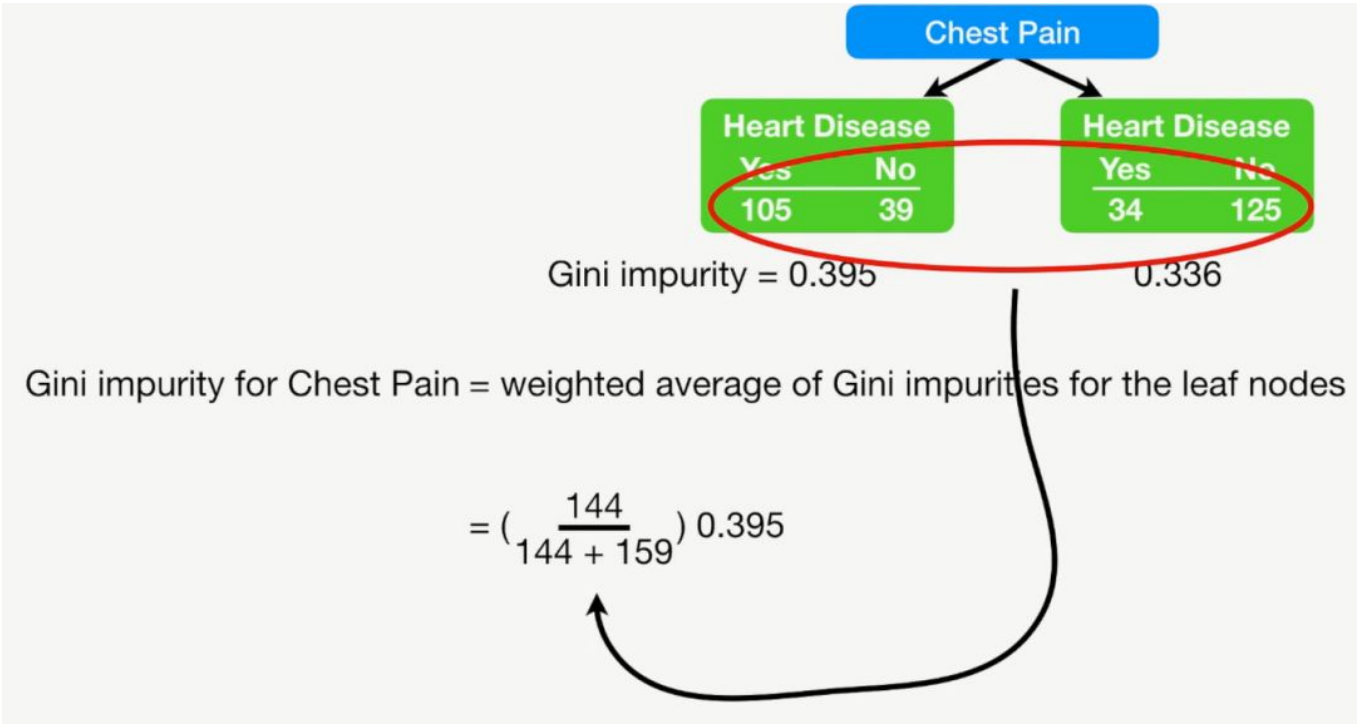
由于我们已经计算了这两个叶子节点为的基尼不纯度，那么我们就能计算使用了胸痛（chest pain）这个变量区分患者的总基尼不纯度，如下所示：



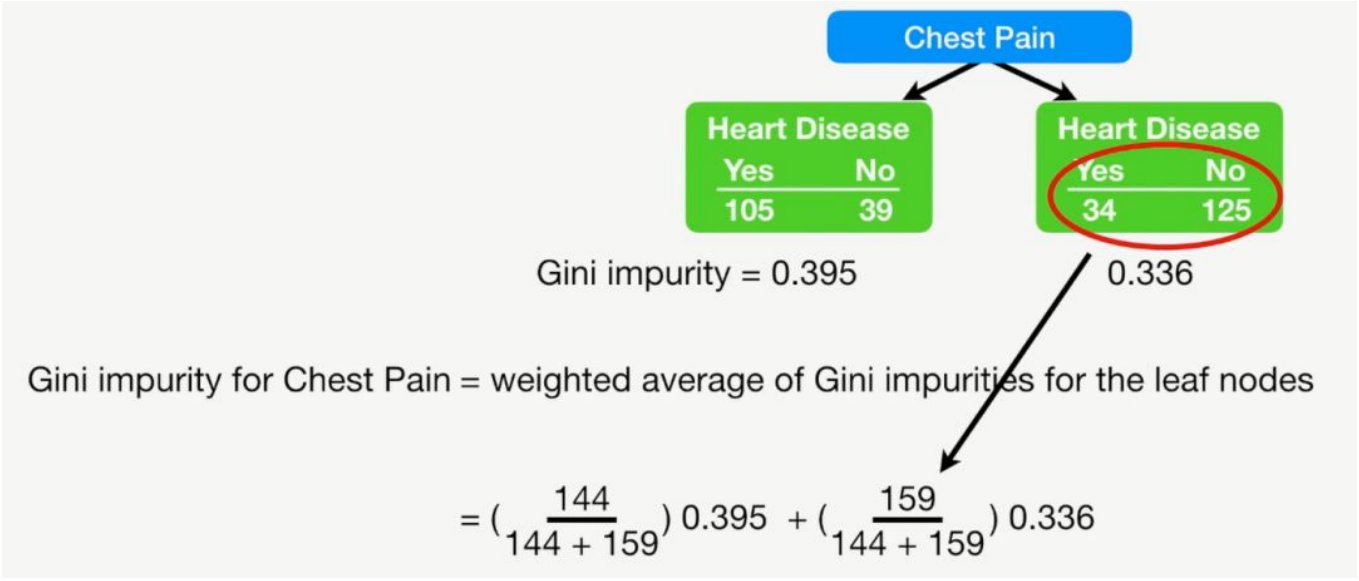

左侧的叶子节点一共有144个患者，右侧的叶子节点一共有159个患者，计算胸痛这个变量来区分患者的总的基尼不纯度的就是叶子节点不纯度加权平均数（weighted average of the leaf node impurities），如下所示：



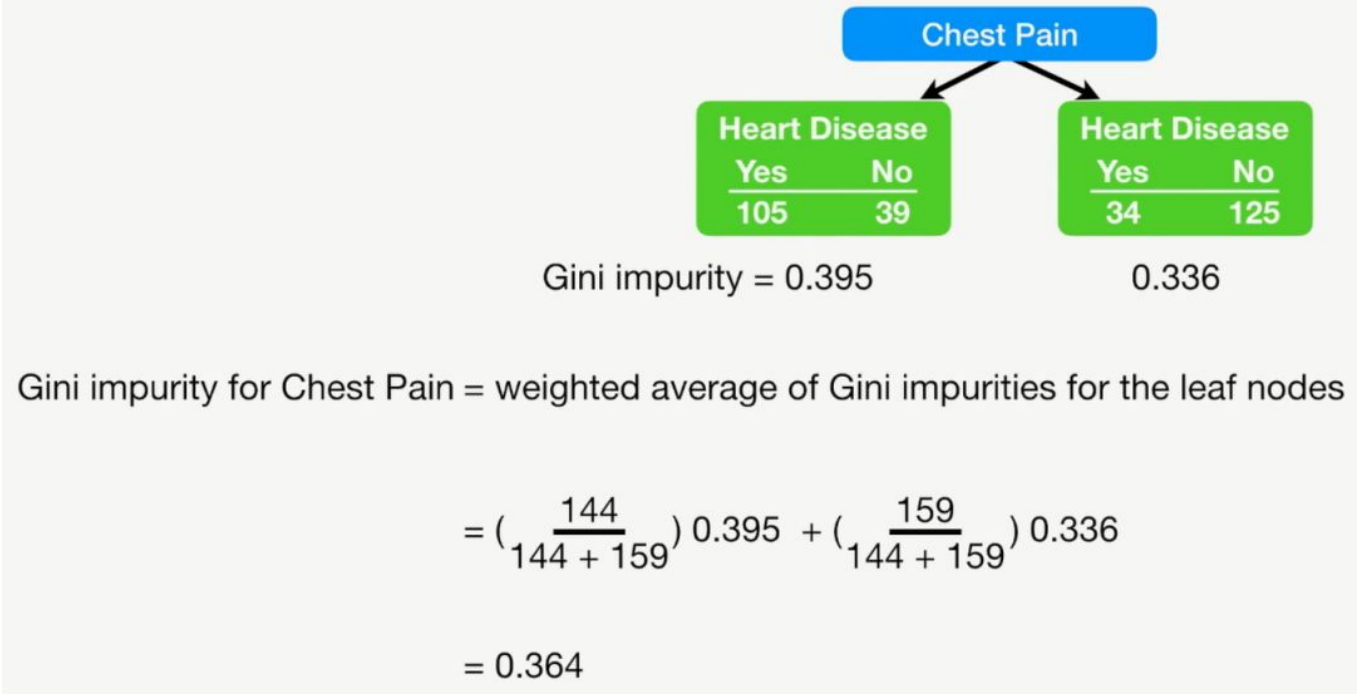

它的公式是，常规的基尼不纯度乘以某个叶子节点的基尼不纯度，如下所示：



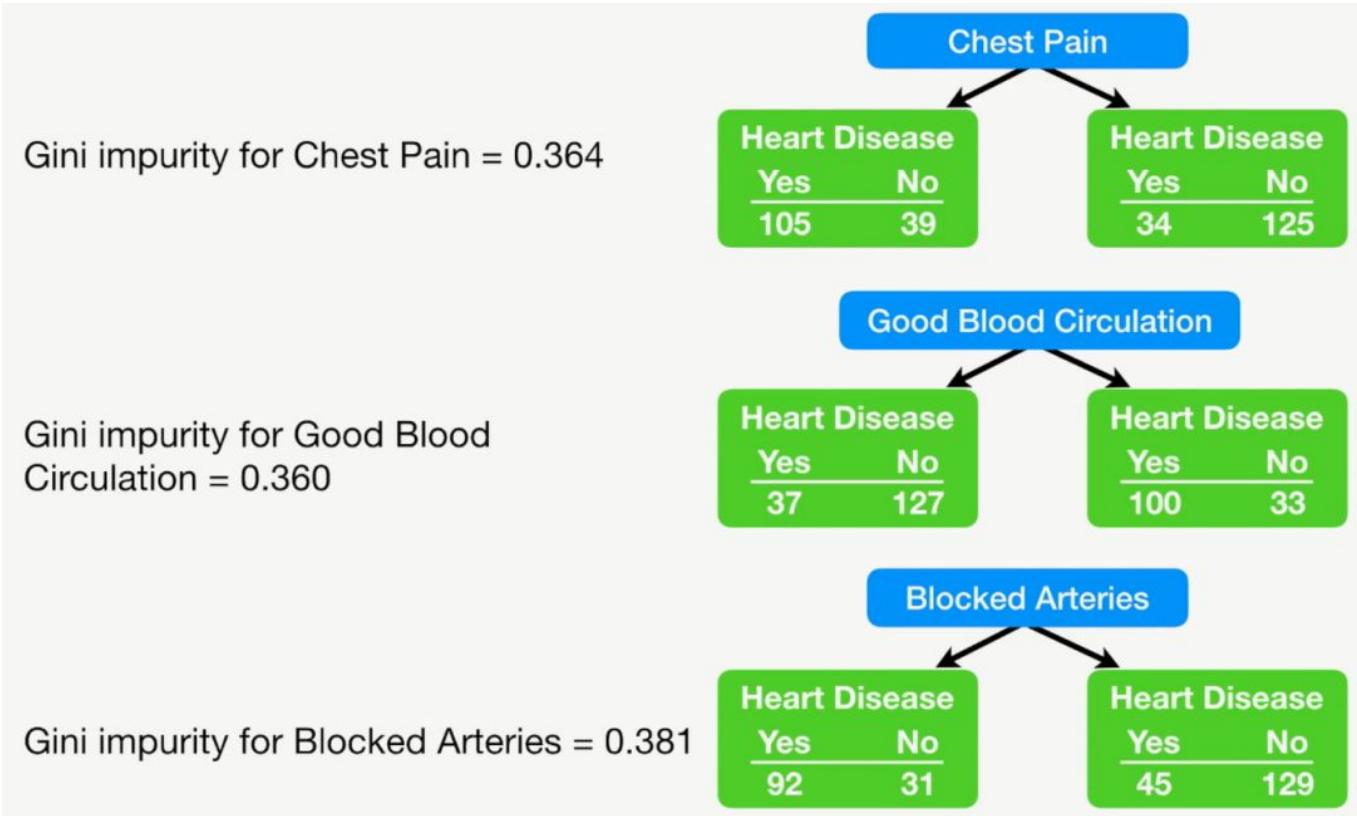

另外一部分则是如下所示：



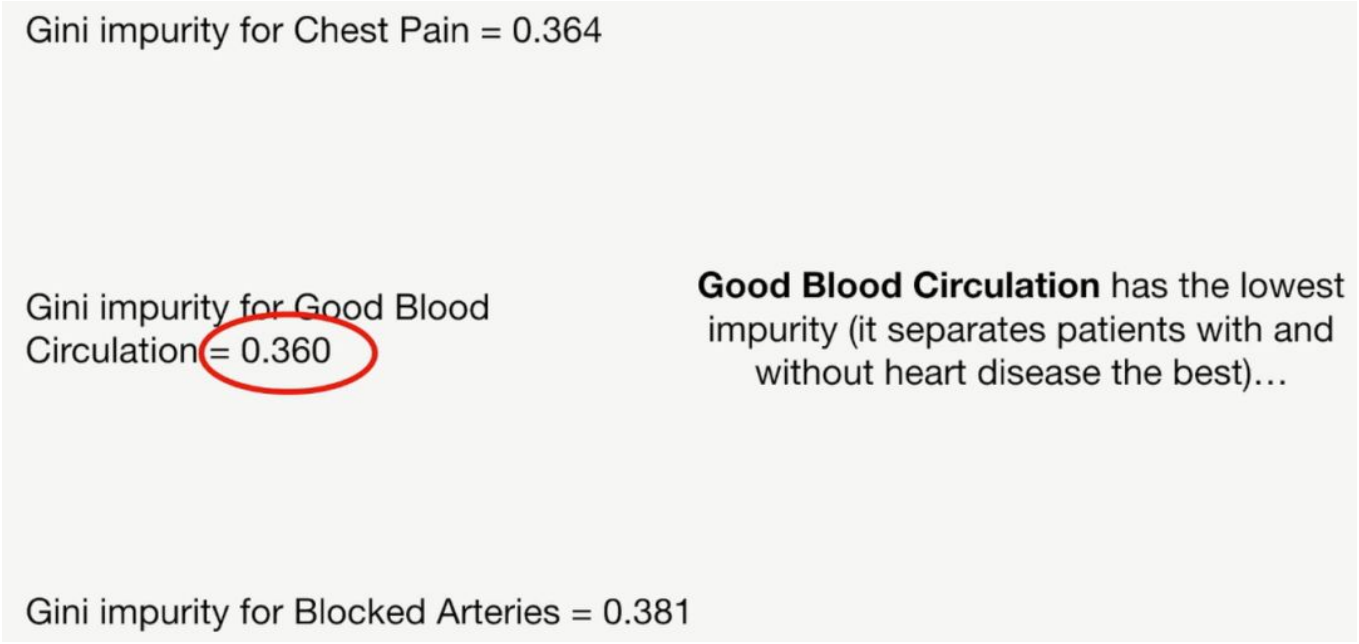

总的结果如下所示：




此时，计算出剩余两个变量的总基尼不纯度，如下所示：

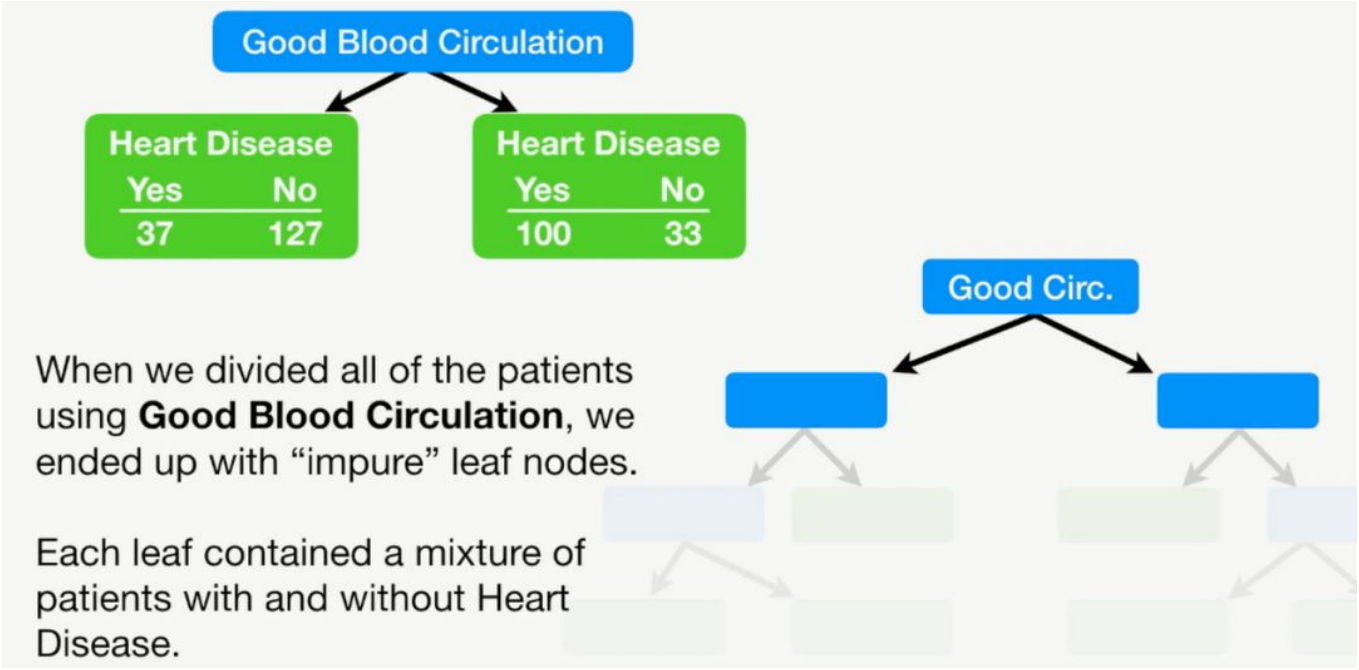



从中我们可以发现，良好血液循环（good blood circulation）这个变量的总基尼不纯度的值最低，如下所示：

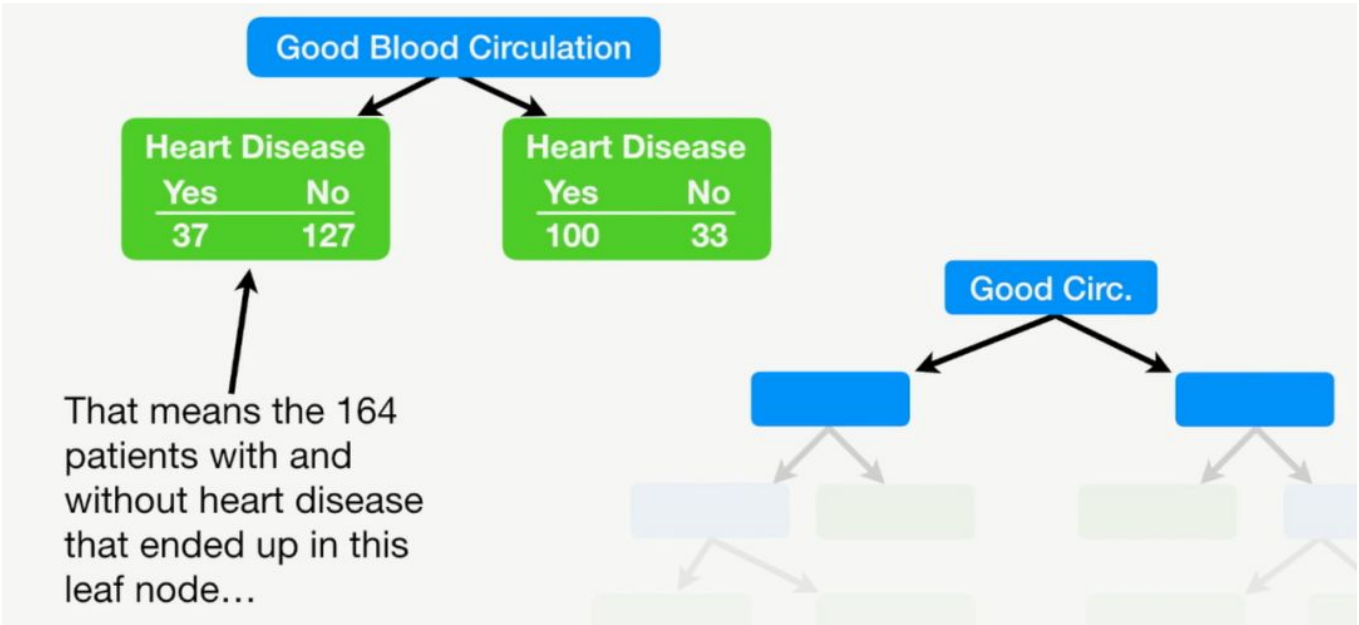



那么我们就使用它作为根节点，如下所示：

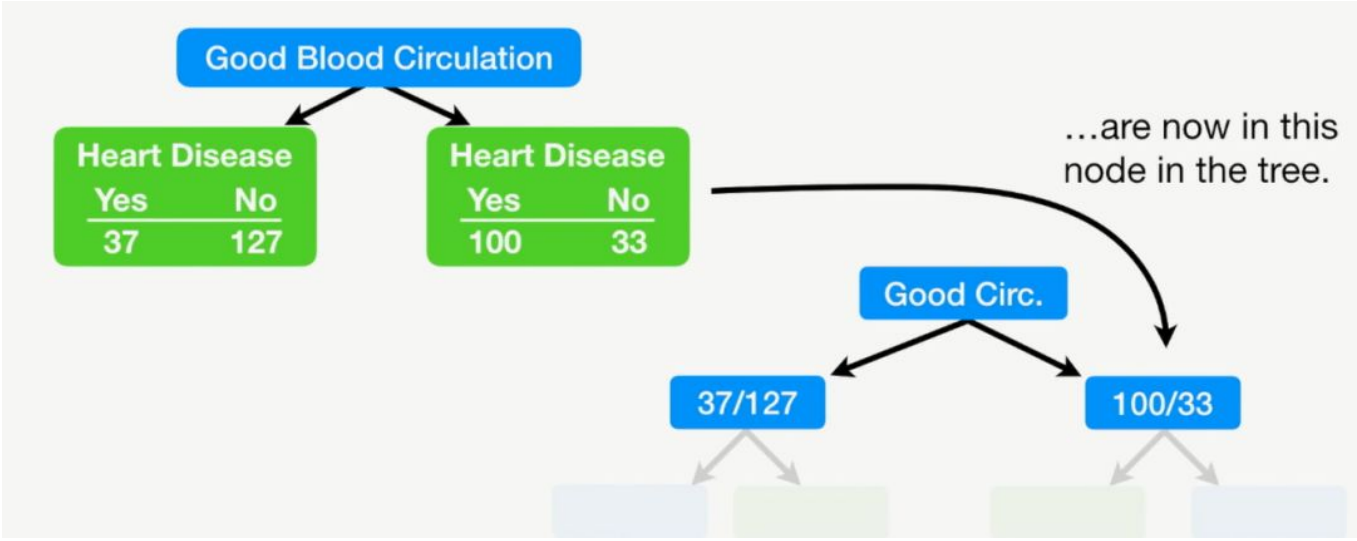

虽然我们采用了良好血液循环（good blood circulation）这个变量作为根节点，但它的叶子节点是“不纯”（impure）的，每个叶子中都含有患心脏病与不患心脏病的患者，如下所示：



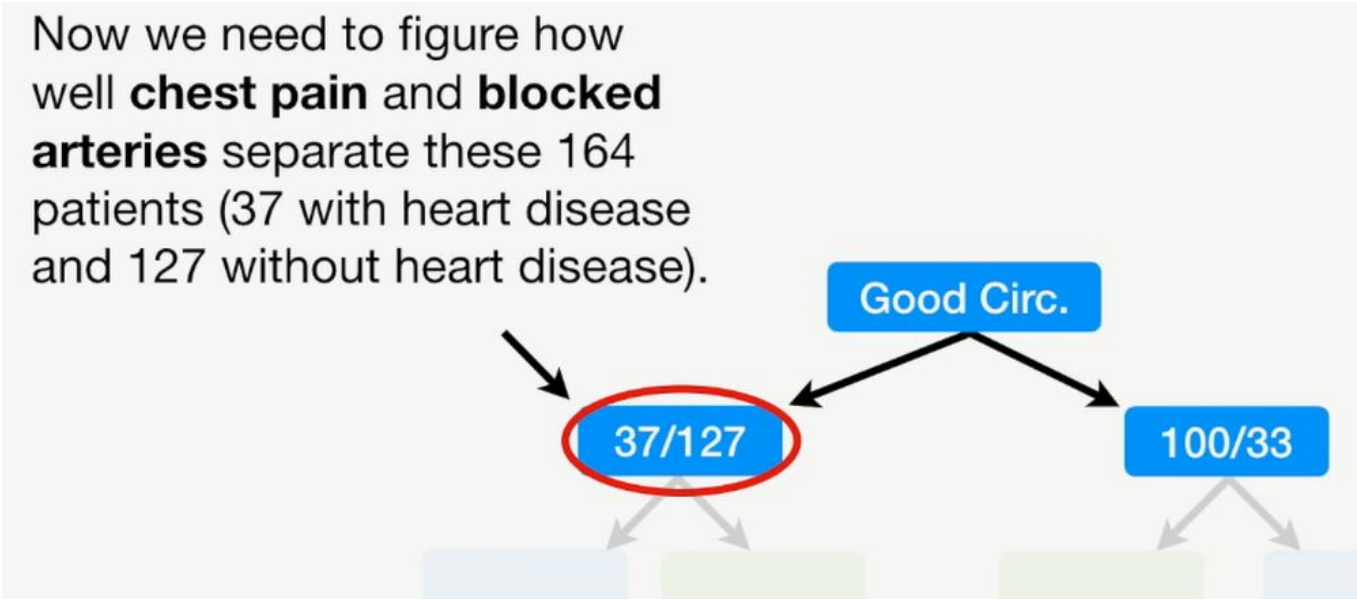

例如，在良好血液循环（good blood circulation）的决策树的叶子节点中，有164个患者，其中37个患有心脏病，127个没有心脏病，如下所示：



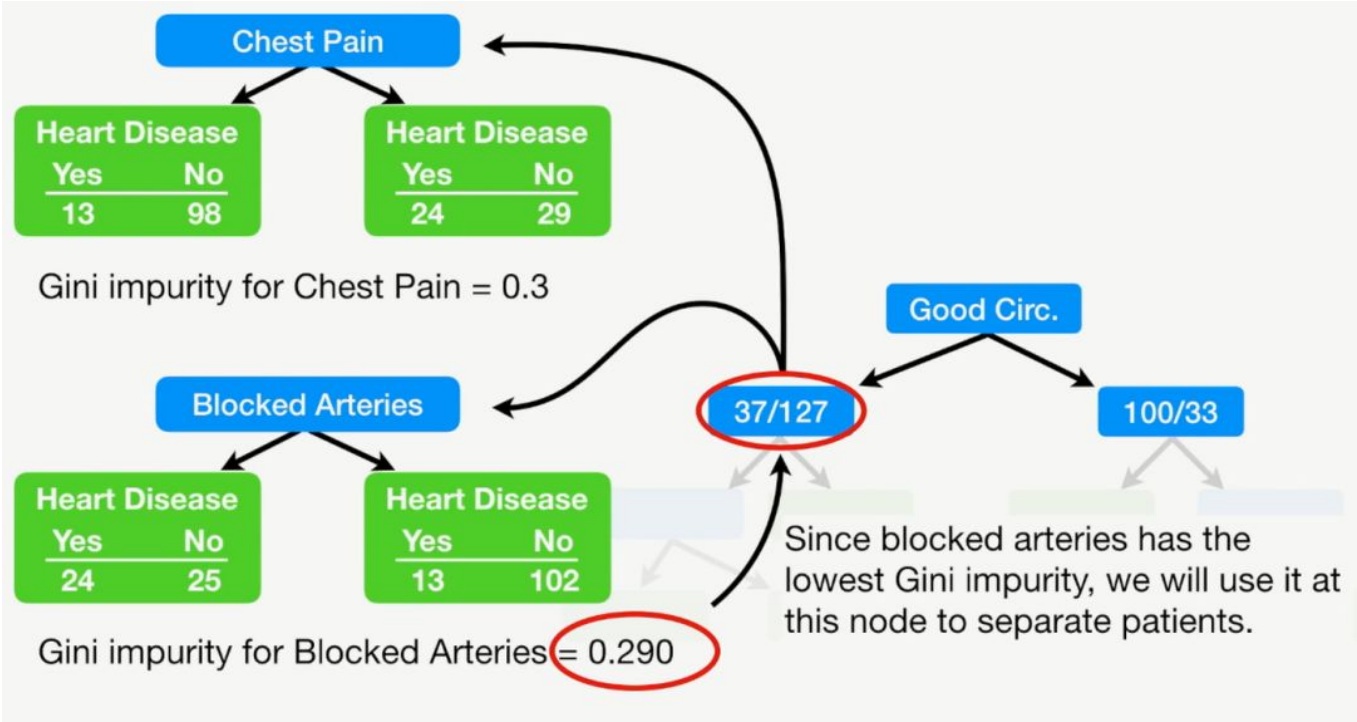

在右侧的叶子结点中，有133个患者，其中100个患有心脏病，33个不患心脏病，如下所示：



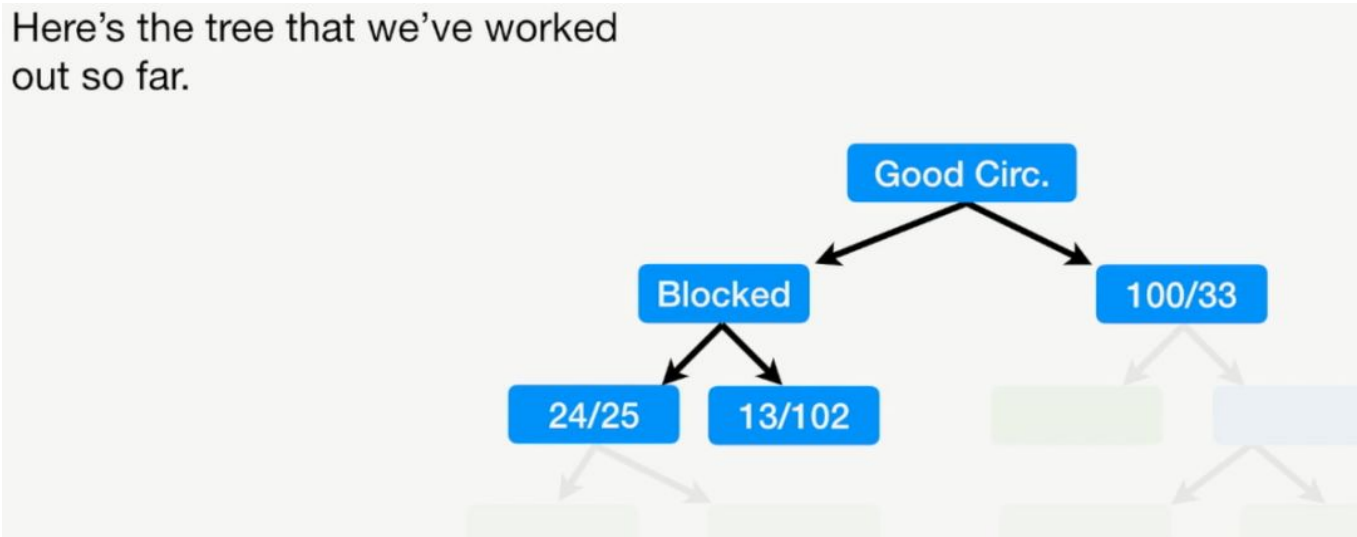

此时，我们需要计算出胸痛（chest pain）和动脉阻塞状态（blocked artery status）这两个变量用于区分左侧164个患者的情况，如下所示：



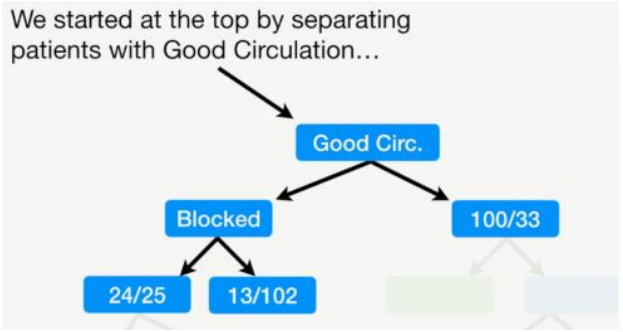

此时，我们分别计算一下胸痛（chest pain）和动脉阻塞状态（blocked artery status）在良好血液循环（good blood circulation）的决策树的左侧叶子节点下的基尼不纯度，如下所示：

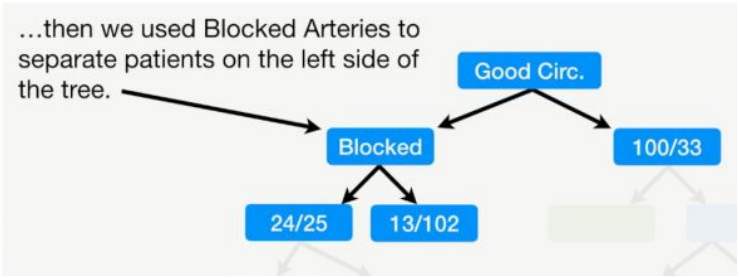



由于动脉阻塞状态（blocked artery status）在这个叶子节点的基尼不纯度数值最低，因此，我们使用这个变量来区分这些患者，如下所示：

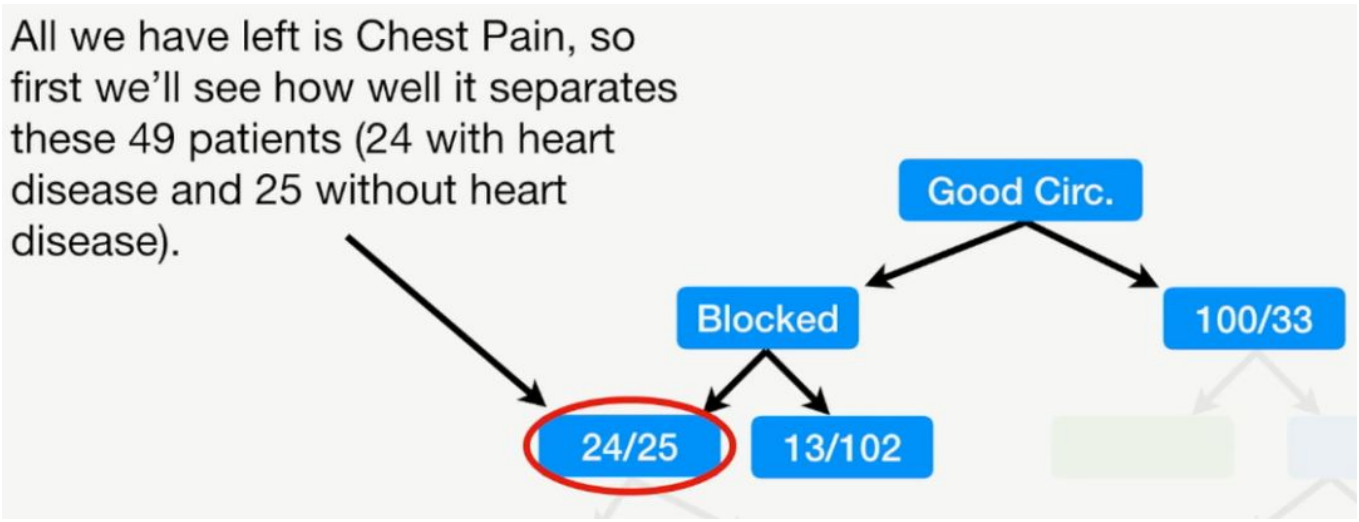



此时，我们知道了，我们构建的这个决策树，先是使用了良好血液循环（good blood circulation）作为根节点，然后使用了动脉阻塞状态（blocked artery status）作为内部节点，如下所示：

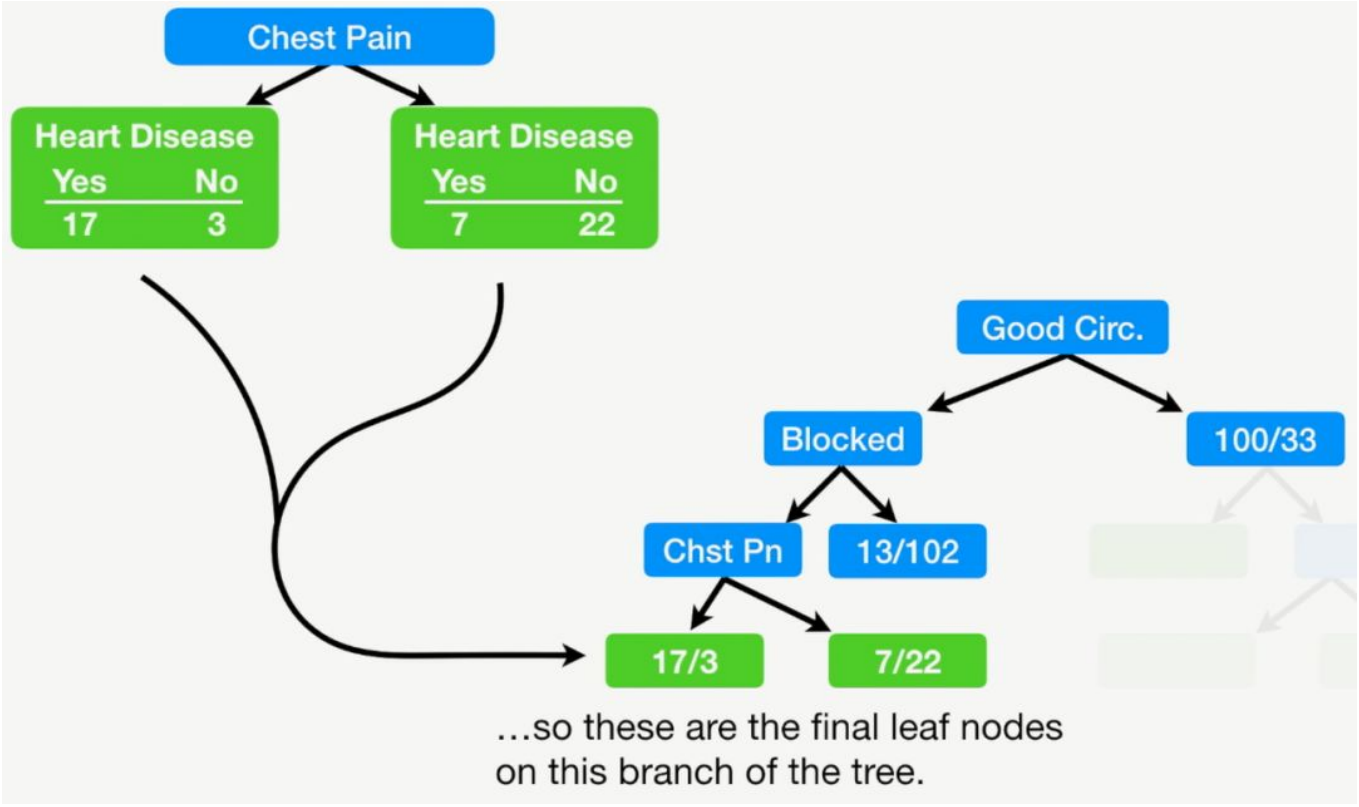


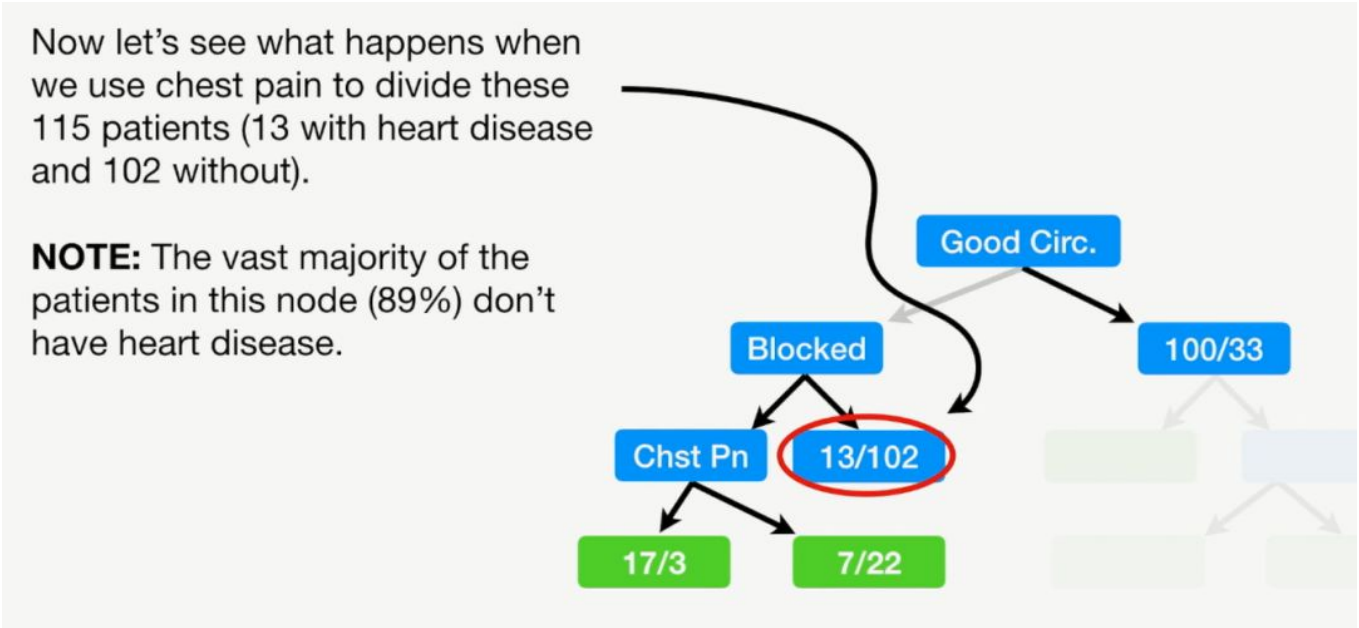

最后，我们使用剩下的变量，即胸痛（chest pain）来区分上一级的内部节点，如下所示：



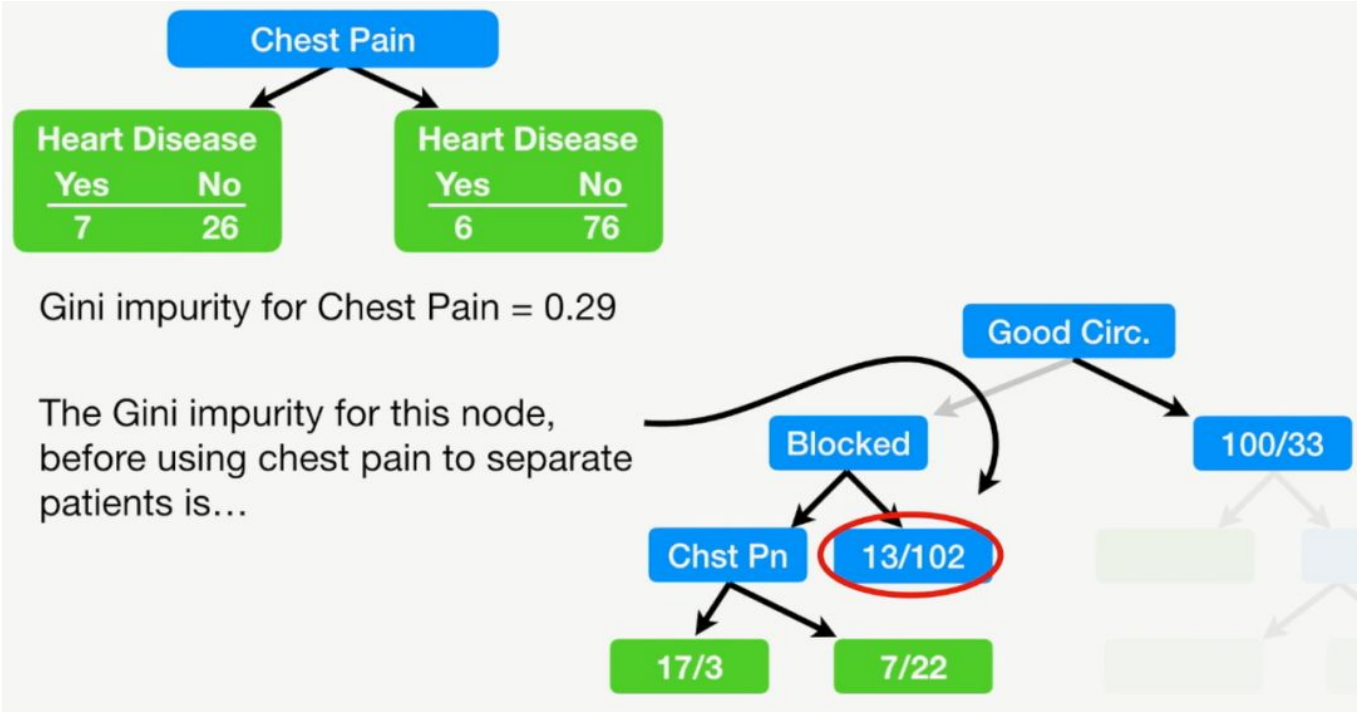

最终我们会发现，胸痛（chest pain）这个变量能够很好地区分最后一个叶子节点中的患者，如下所示：



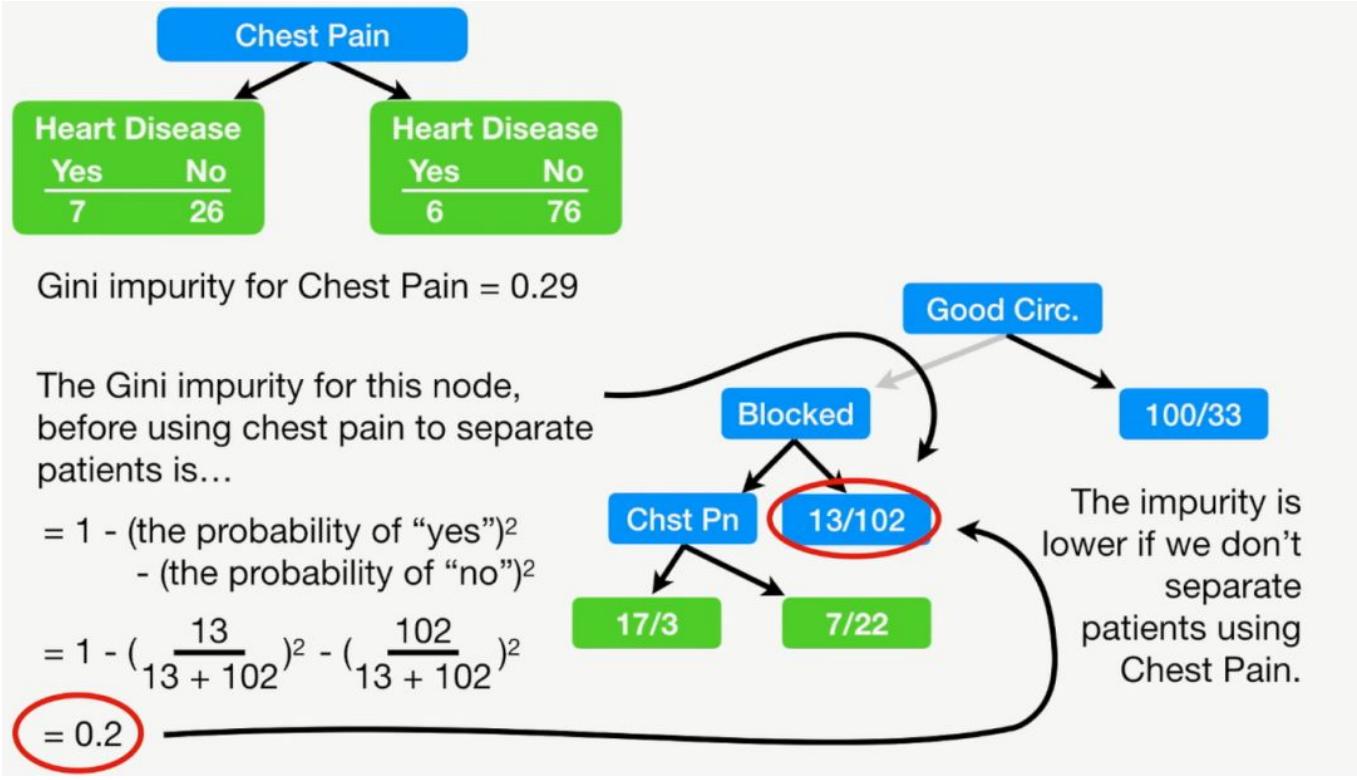

我们再看一下胸痛（chest pain）这个变量区分右侧叶子节点的情况，在这个节点中有115名患者，其中不患心脏病的患者是102人，比例为89%，如下所示：




此时把上面红圈的那个内部节点用胸痛（chest pain）这个变量进行区分，并计算它的基尼不纯度，结果为0.29，如下所示：

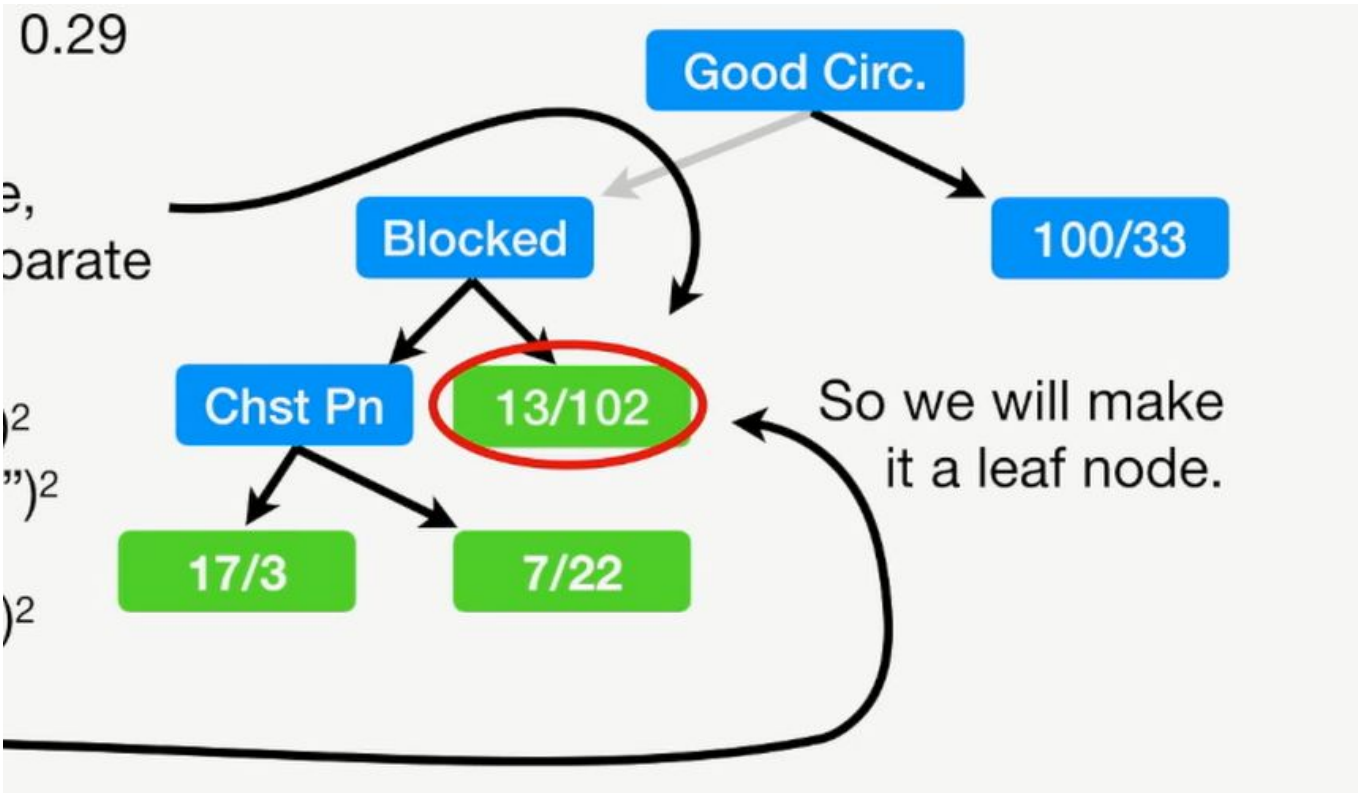



如果我们不使用胸痛（chest pain）这个变量来区分这个叶子结点，我们计算一下它的基尼不纯度，就是0.2，如下所示：

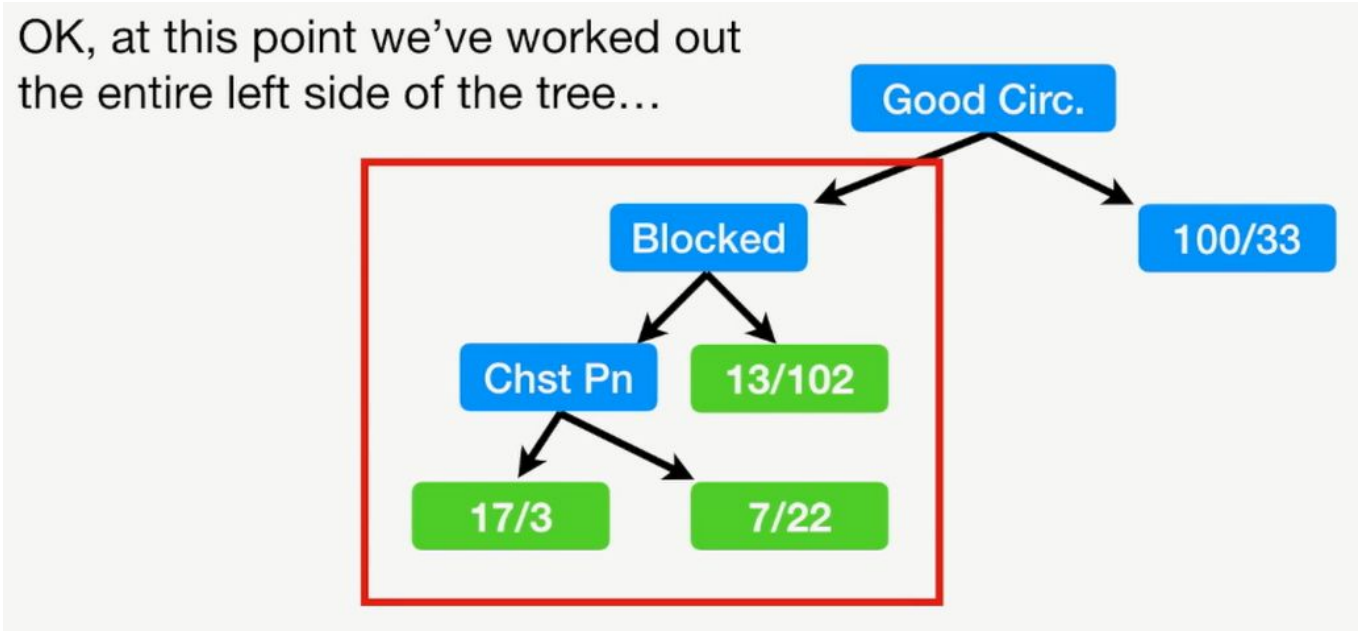



从中我们可以发现，如果我们使用了胸痛（chest pain）这个变量来区分这个叶子节点，那么它的基尼不纯度就相比没有使用变量之前上升了，因此我们并不需要胸痛这个变量来区分

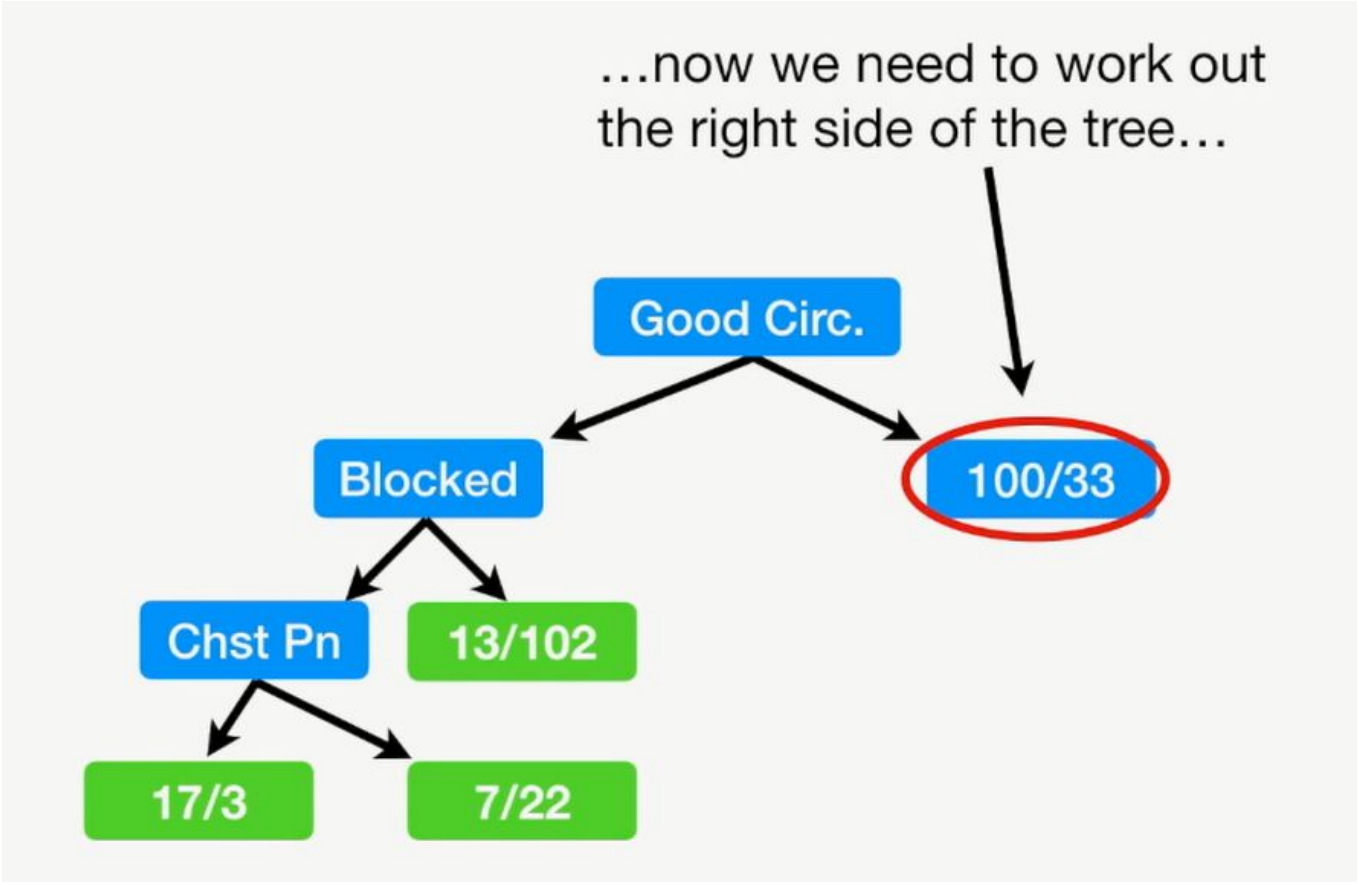
这个叶子节点，只把它当成叶子节点，而不是进一步地区分，当其当成内部节点，如下所示：




因此在决策树的左侧这些节点中，我们已经计算出了最终的结果，如下所示：




我们现在转向右侧的决策树，如下所示：

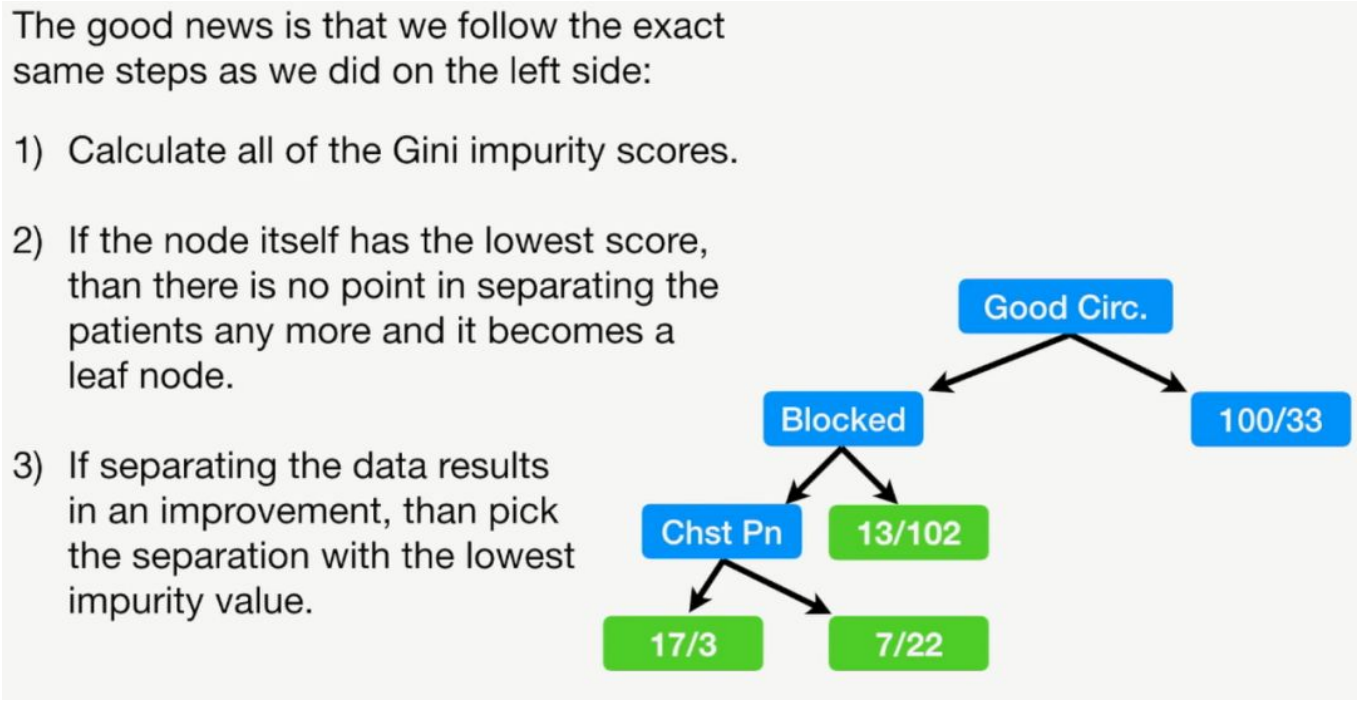



计算右侧的所有节点跟左侧的一样，流程如下所示：

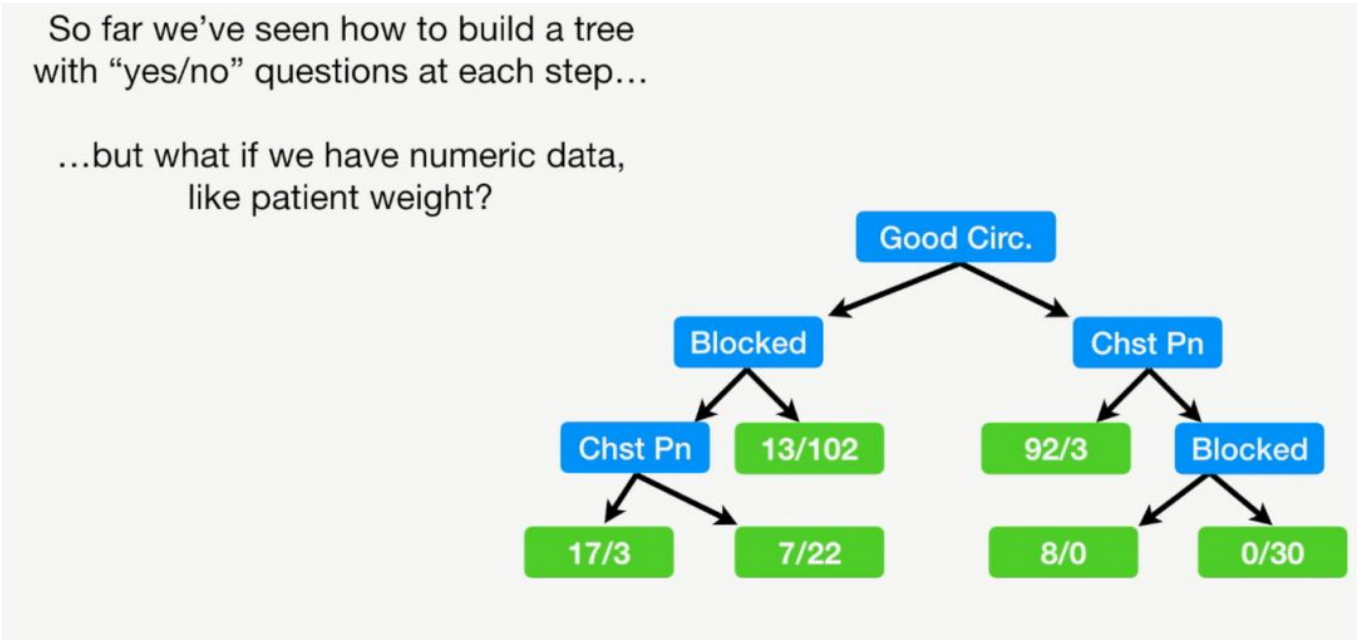
第一，计算出所有的基尼不纯度；

第二，如果节点自身的基尼不纯度比那些加了变量后进一步区分后的基尼不纯度还要低，那么这个节点就被设置为叶子节点；

第三，如果区分的数据会改善最终的分类结果，那么就选择最低不纯度来进行区分，如下所示：

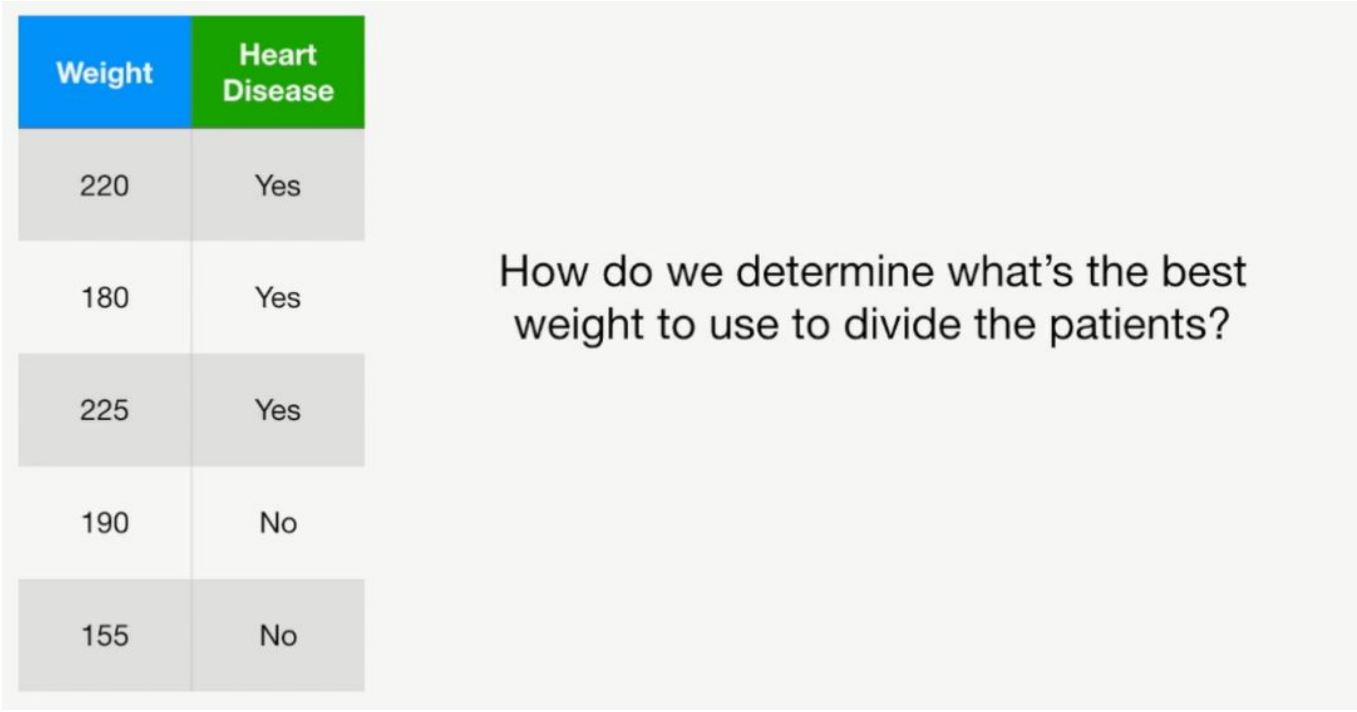



最终，我们会生成一个完整的决策树，如下所示：




## 连续型变量构建决策树

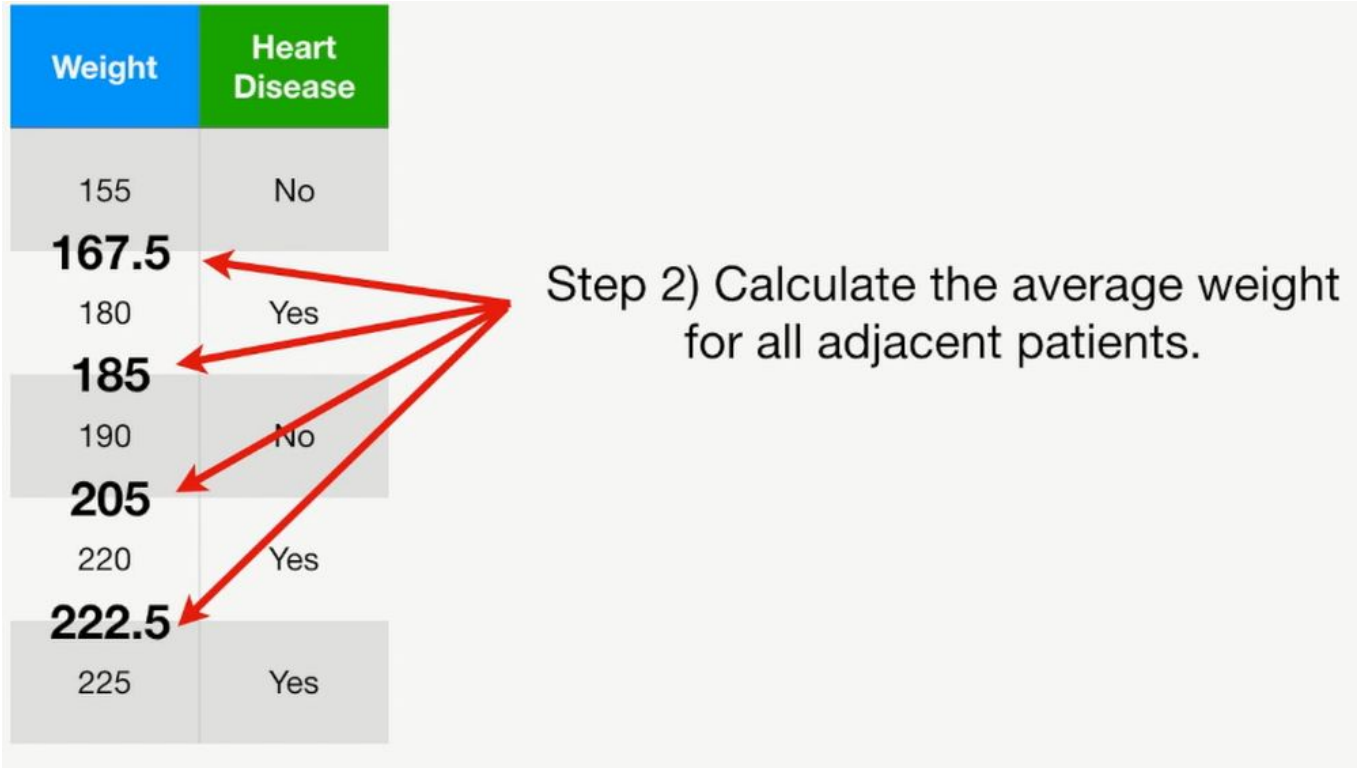
前面我们讲的都是二分类变量（yes或no）构建决策树的过程，如果我们的变量是连续型的，例如患者的体重，如何构建决策树呢，如下所示：



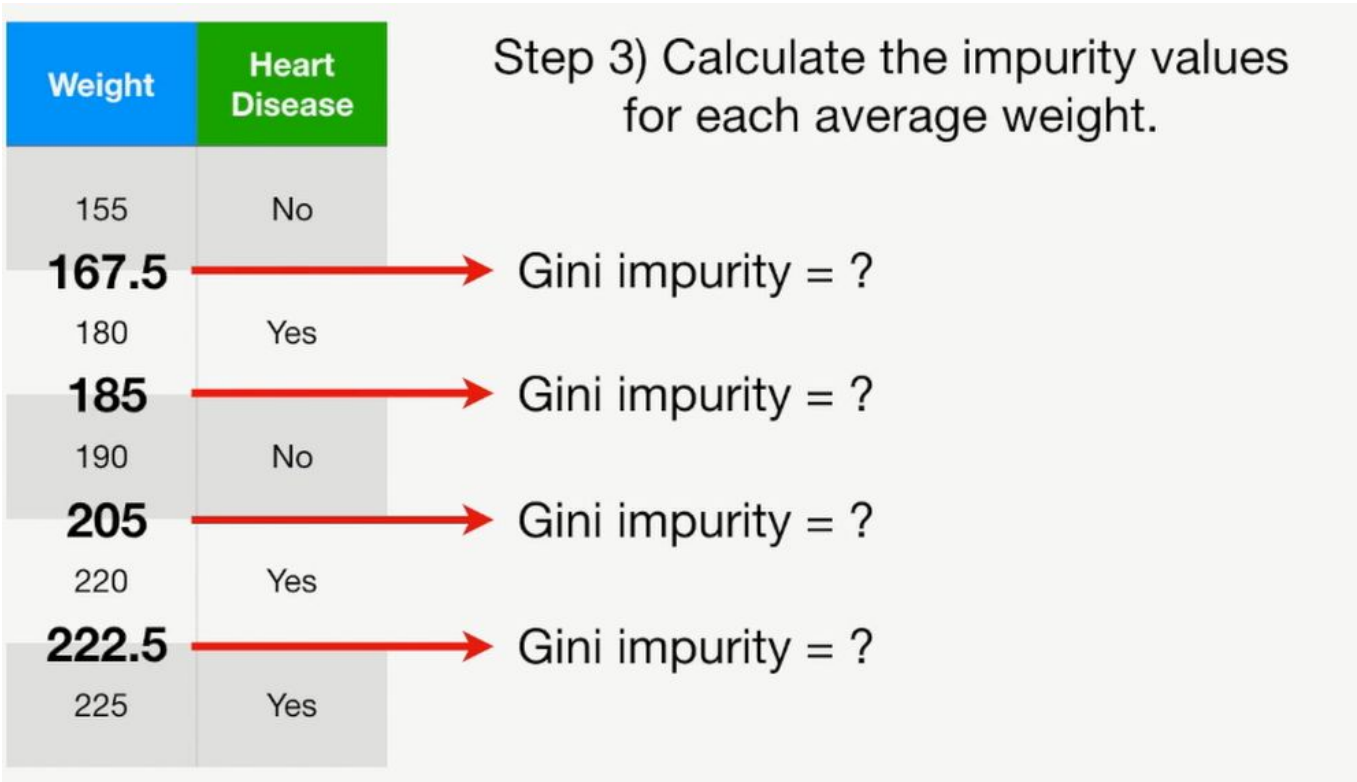

下面就讲这种变量构建决策树的流程：

第一步：对体重进行排序，顺序从小到大，如下所示：

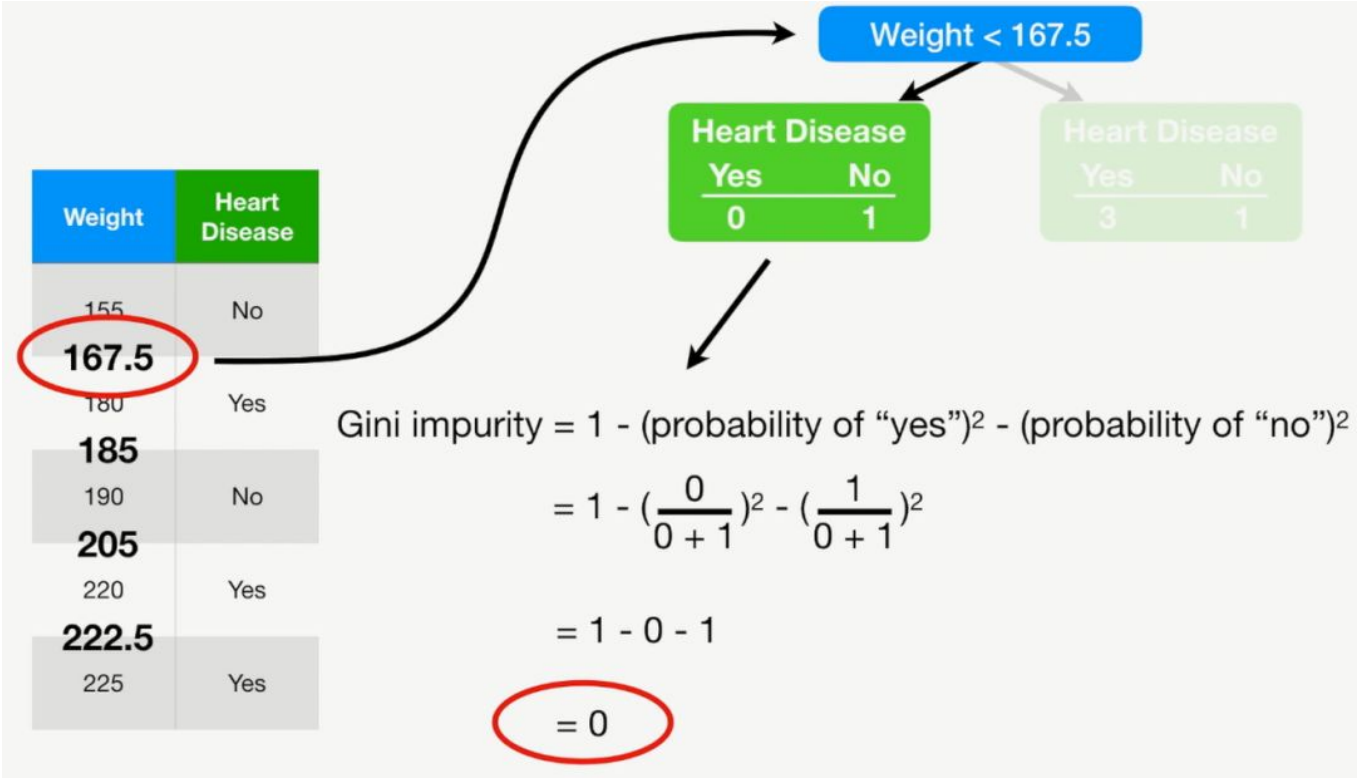

第二步：计算所有相邻患者的加权平均数，如下所示：



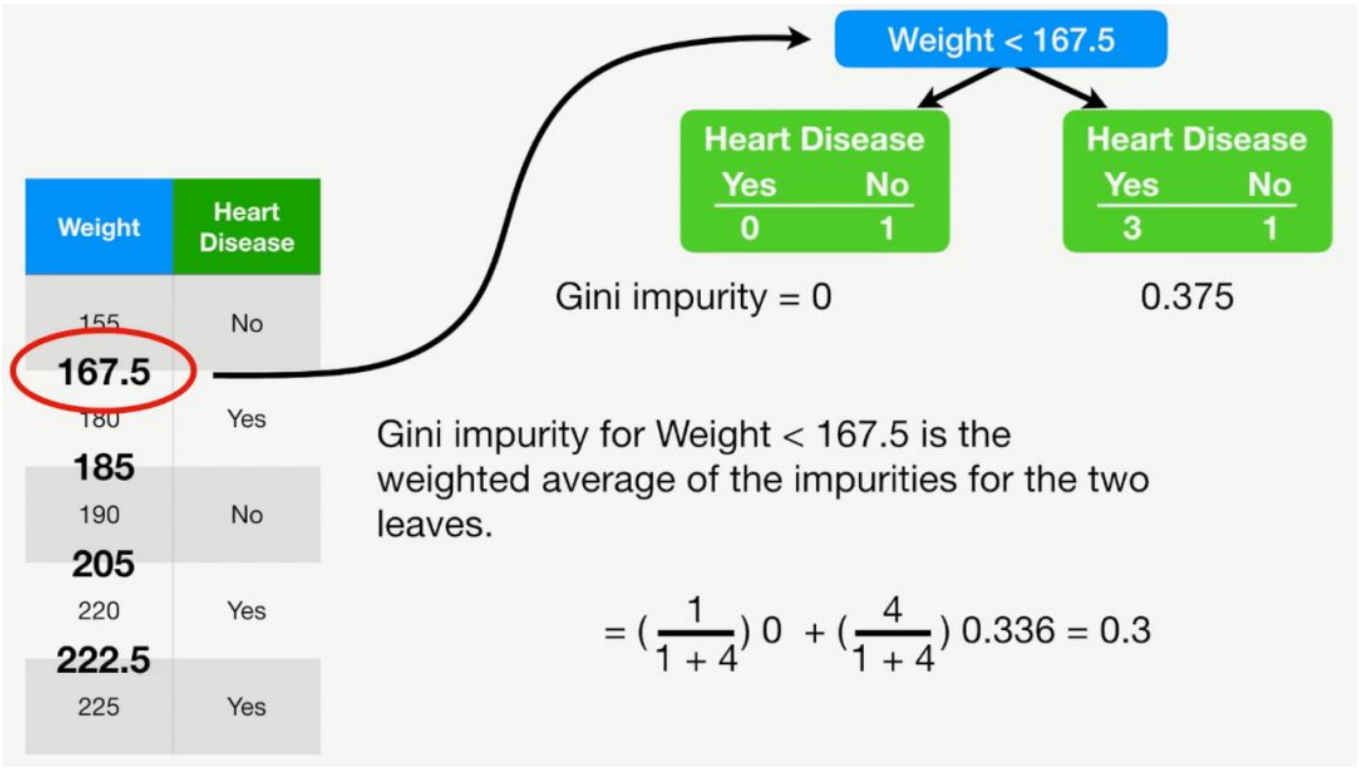

第三步：计算每个加权平均数的不纯度值，如下所示：



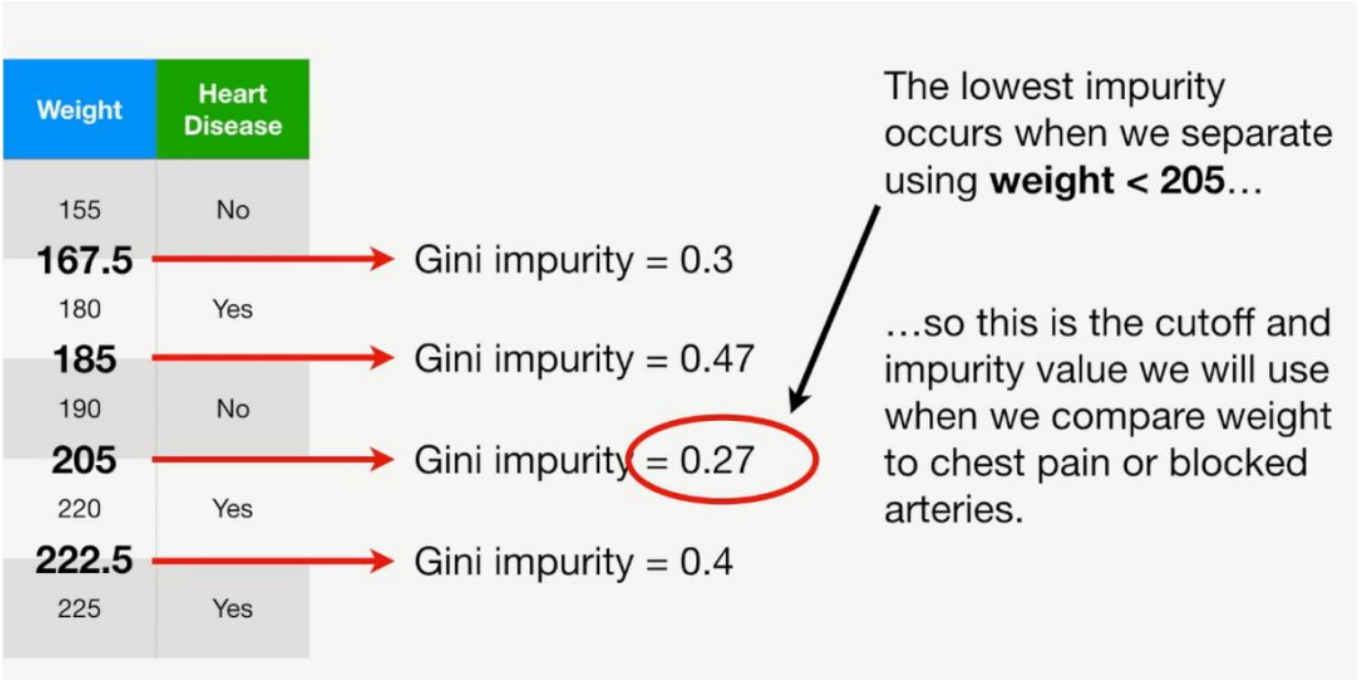

现在看一下计算流程，先看第一个167.5这个平均数的基尼不纯度，如下所示：




计算出的结果为0，再计算一下右侧的基尼不纯度，为0.375，那么计算一下总的基尼不纯度，如下所示：



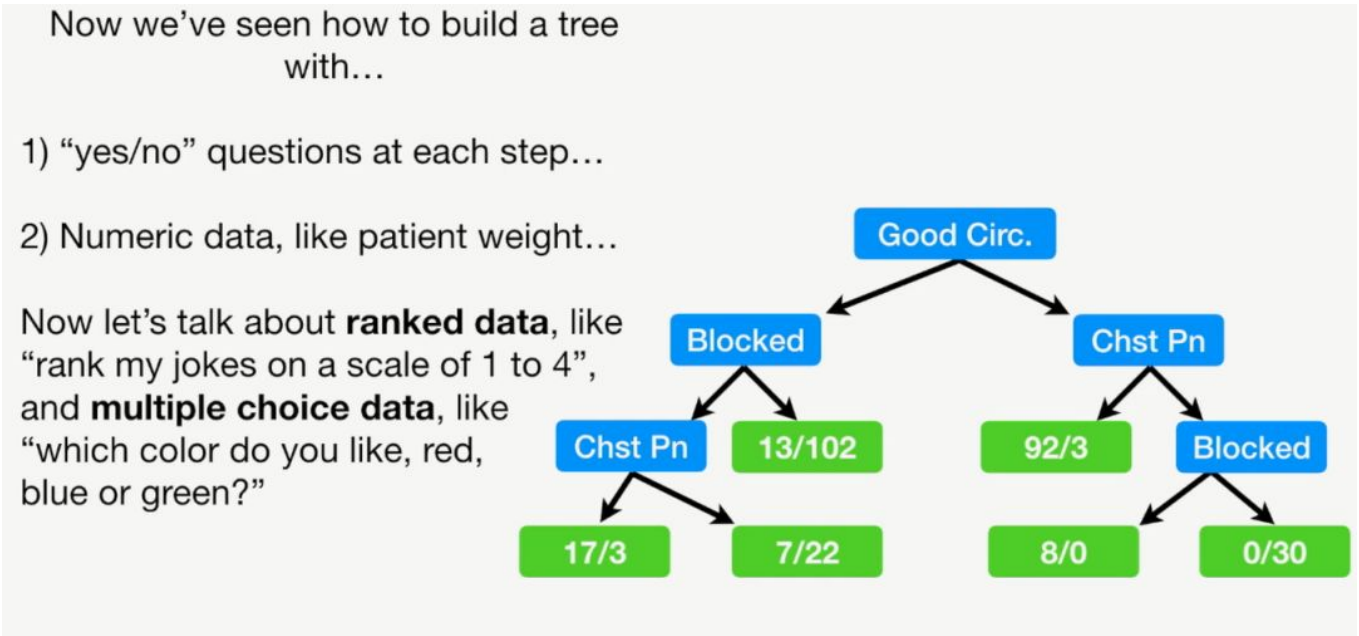

现在计算出所有的变量的基尼不纯度，我们发现205这个体重的基尼不纯度为0.27，它的值最低，因此我们会选择205作为阈值，如下所示：



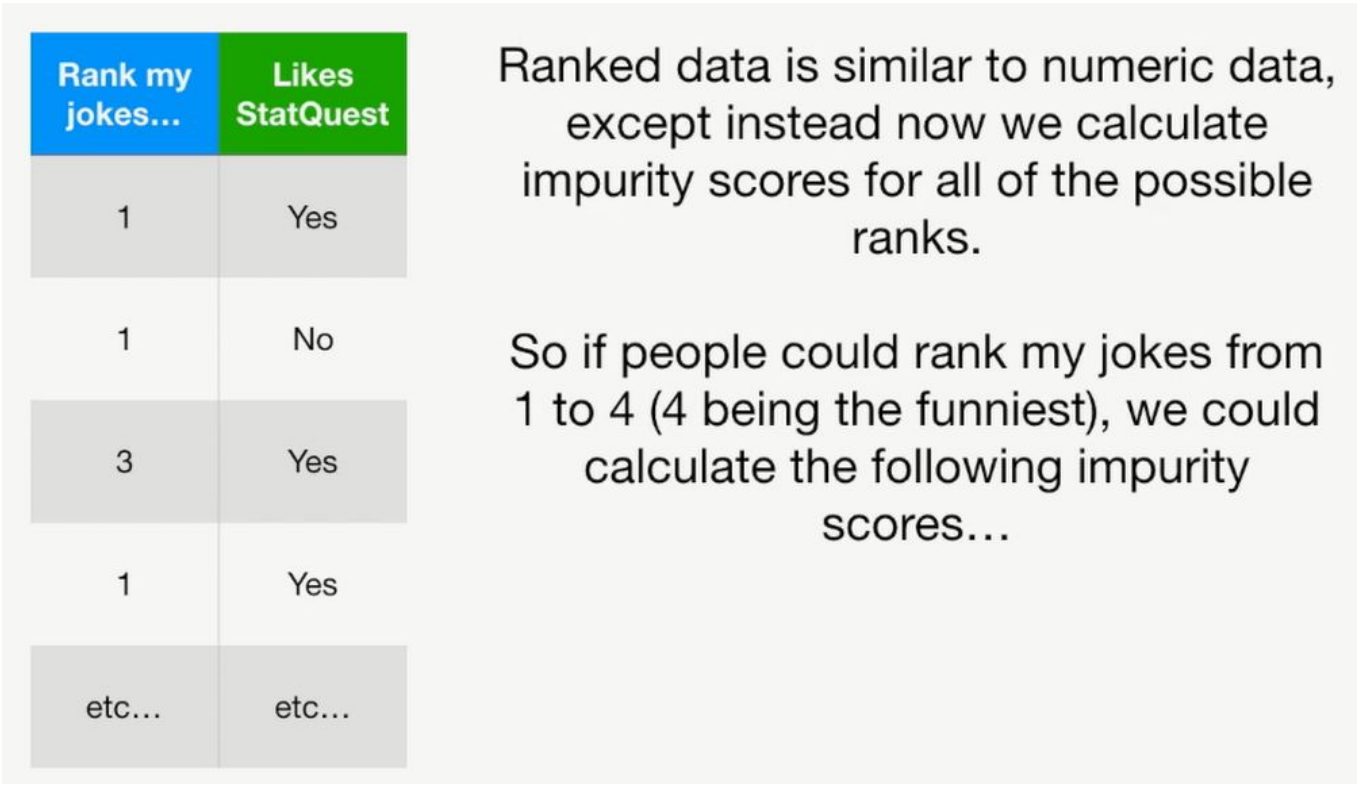

## 含有排序数据的决策树构建

我们前面讲了含有二分类变量（yes或no）与连续型变量（体重）数据的决策树构建，现在我们再谈一种含有“排序数据”（ranked data）的决策树构建，排序数据就是像这样的数据，例如“把我的笑话按照搞笑程度从1到4排列”。

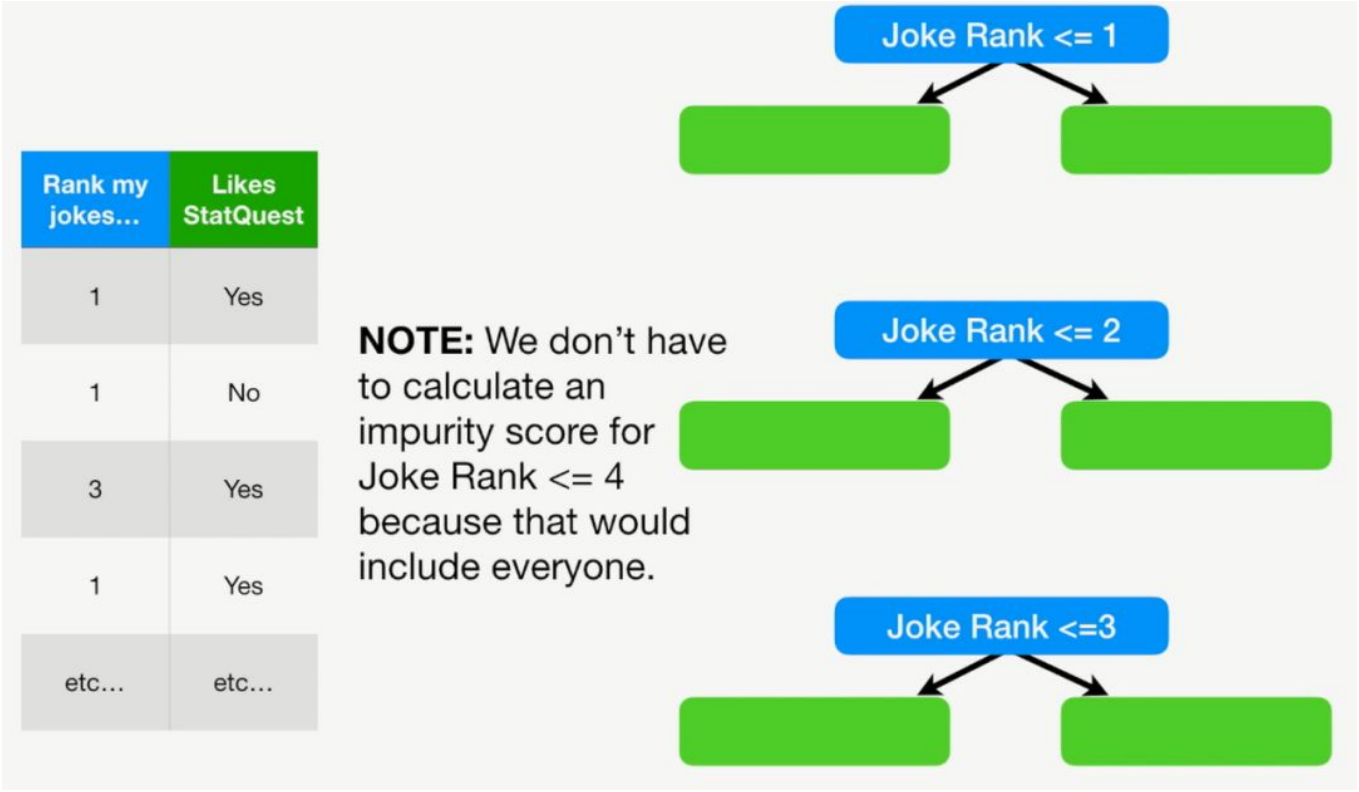
此外，我们还会涉及一种多个选择数据（multiple choice data）的决策树的构建，多个选择数据就像这样的数据，“你喜欢什么颜色，红色，绿色，还是蓝色？”如下所示：



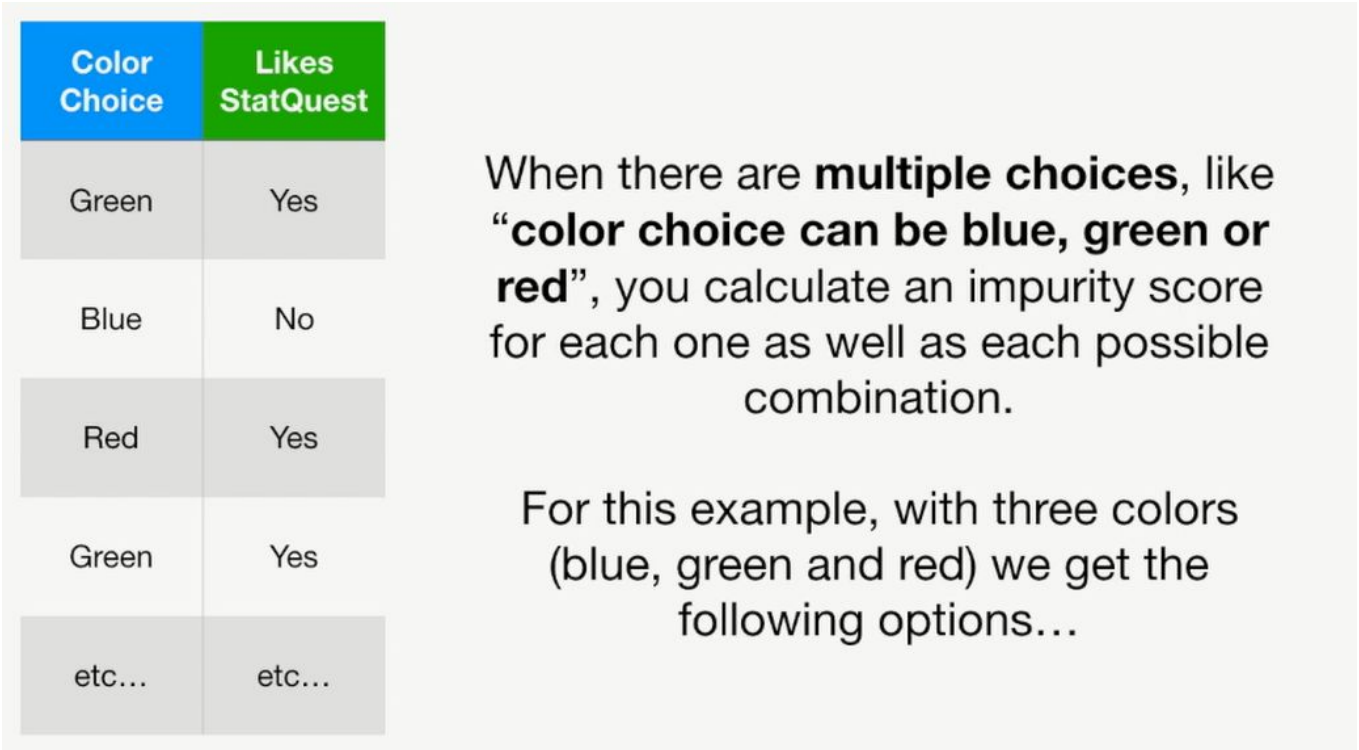
排序数据类似于连续型变量的数据（例如体重），只不过我们计算的是这些可能排序的不纯度，例如，把我的笑话按照搞笑程度进行排序，其中4是最搞笑，1不太搞笑，现在我们计算这些笑话的不纯度，如下所示：



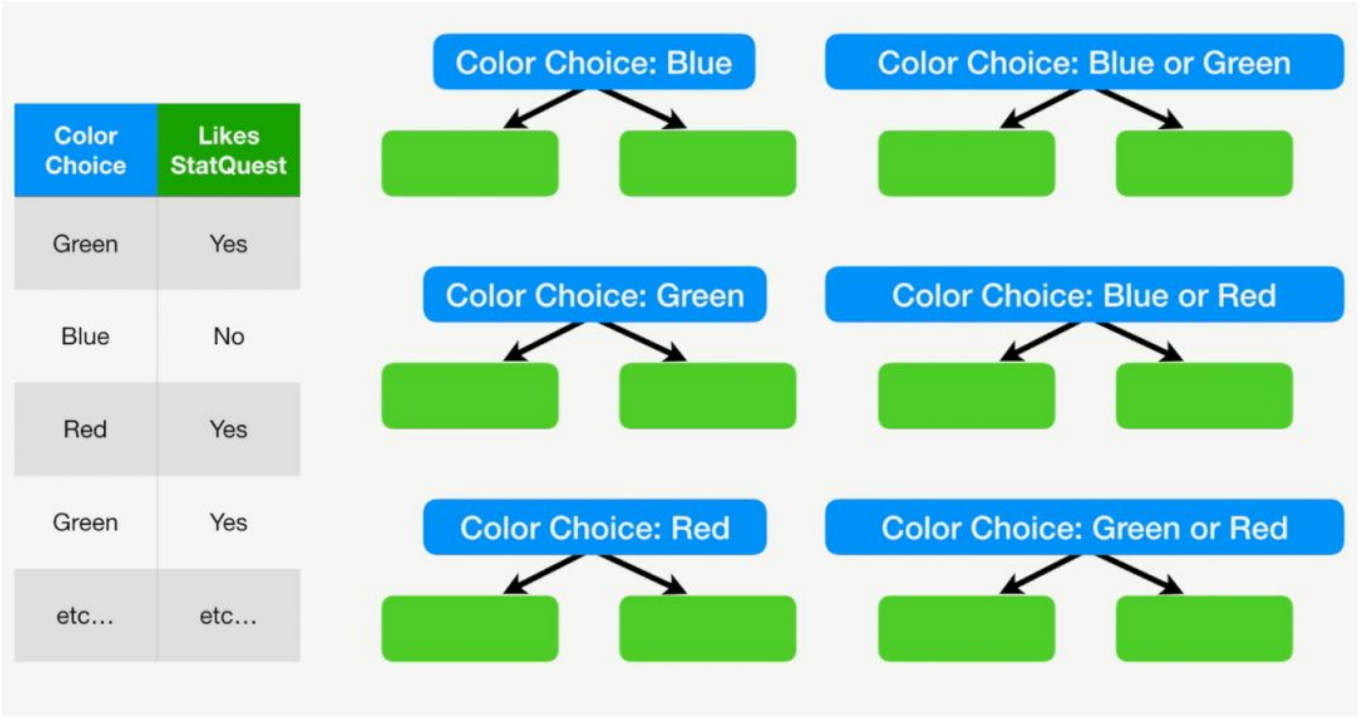

在计算这几个数据的不纯度时，我们并不计算`Joke Rank <=4`这种情况，因此它基本上包括了所有的情况，如下所示：



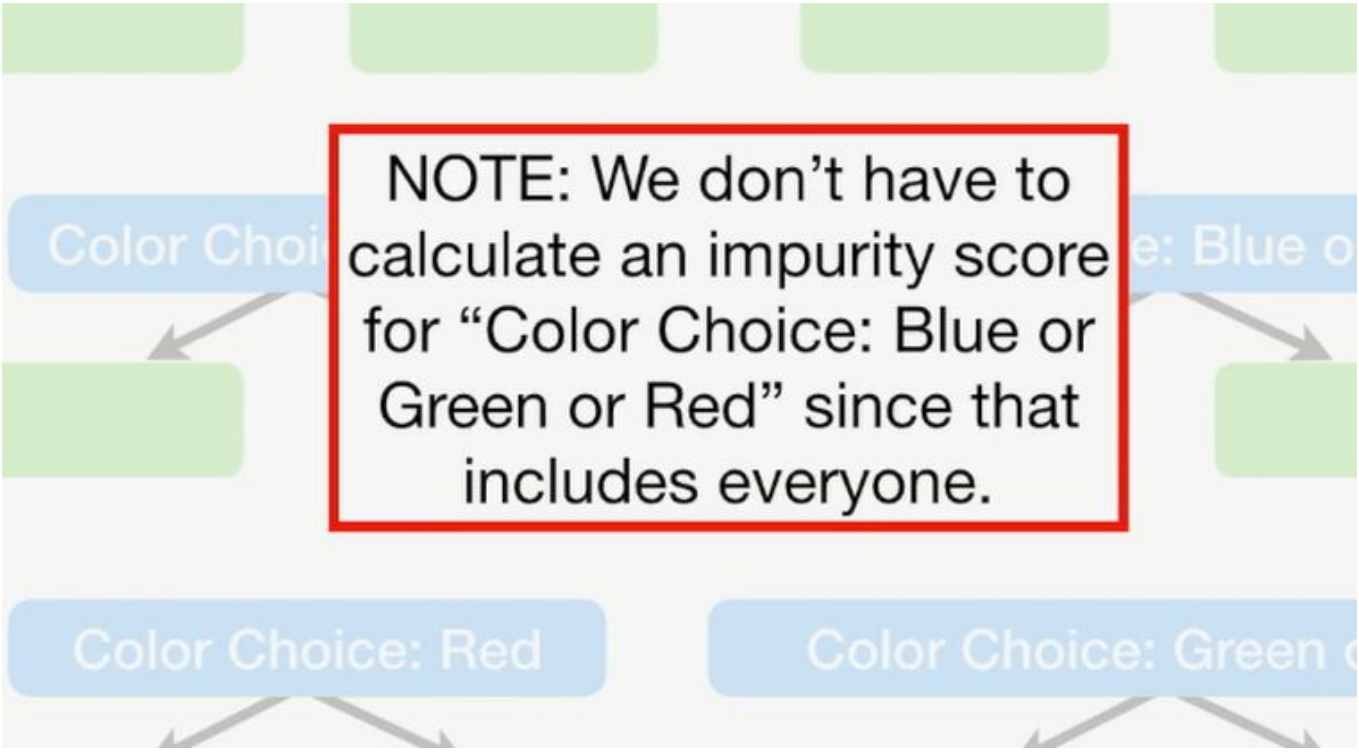

现在我们看一下含有多个选择的数据，例如你喜欢什么颜色，颜色可以选择蓝色，绿色或红色，此时我们会计算每个组合的不纯度，在这个案例中，我们只有这三个种颜色，如下所示：




它的计算过程如下所示：

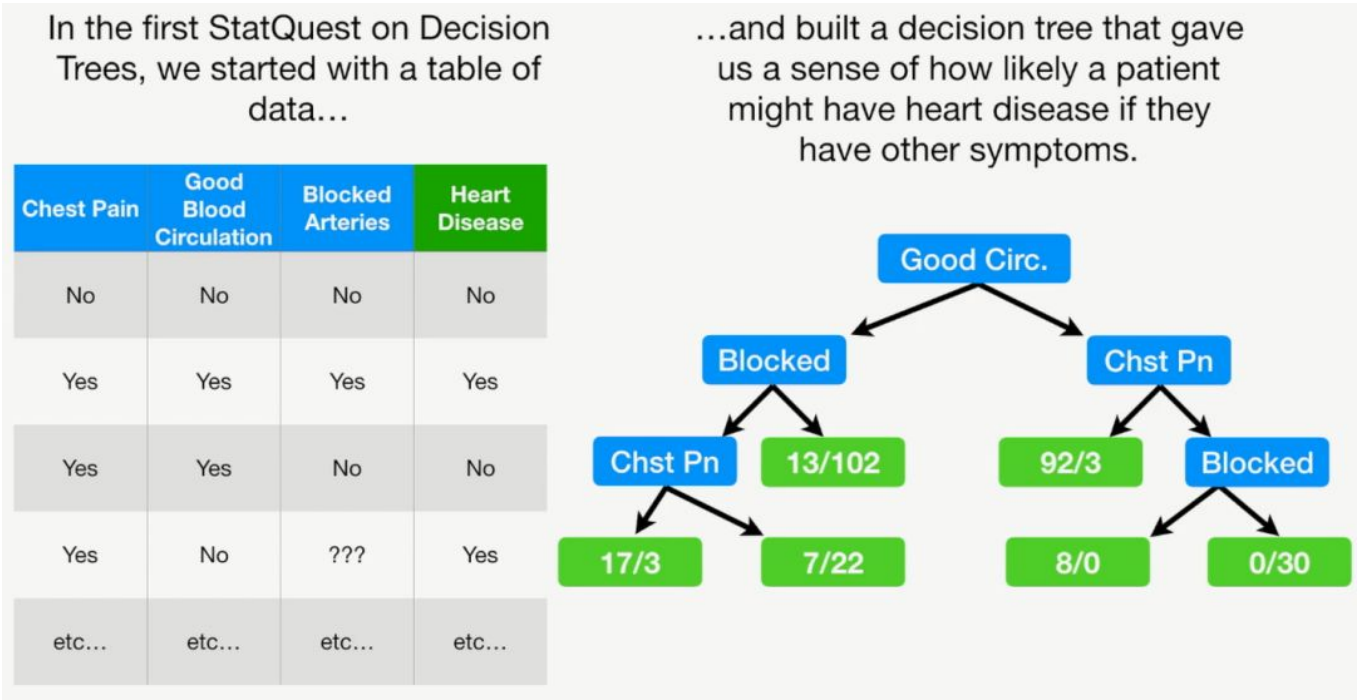



在这个案例中，我们并不过计算“蓝色或绿色或红色”这种情况，因此这种情况包含了所有的情况，如下所示：

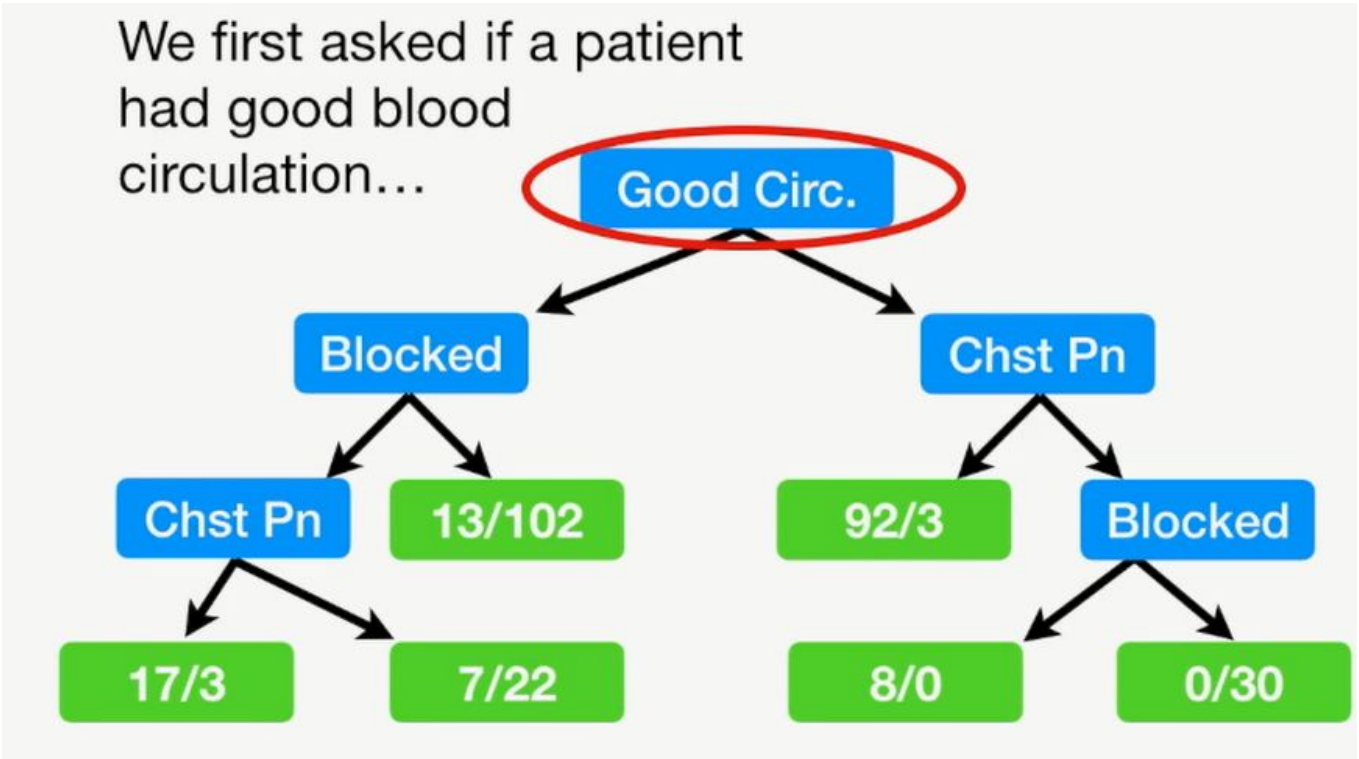



## 决策树的过拟合

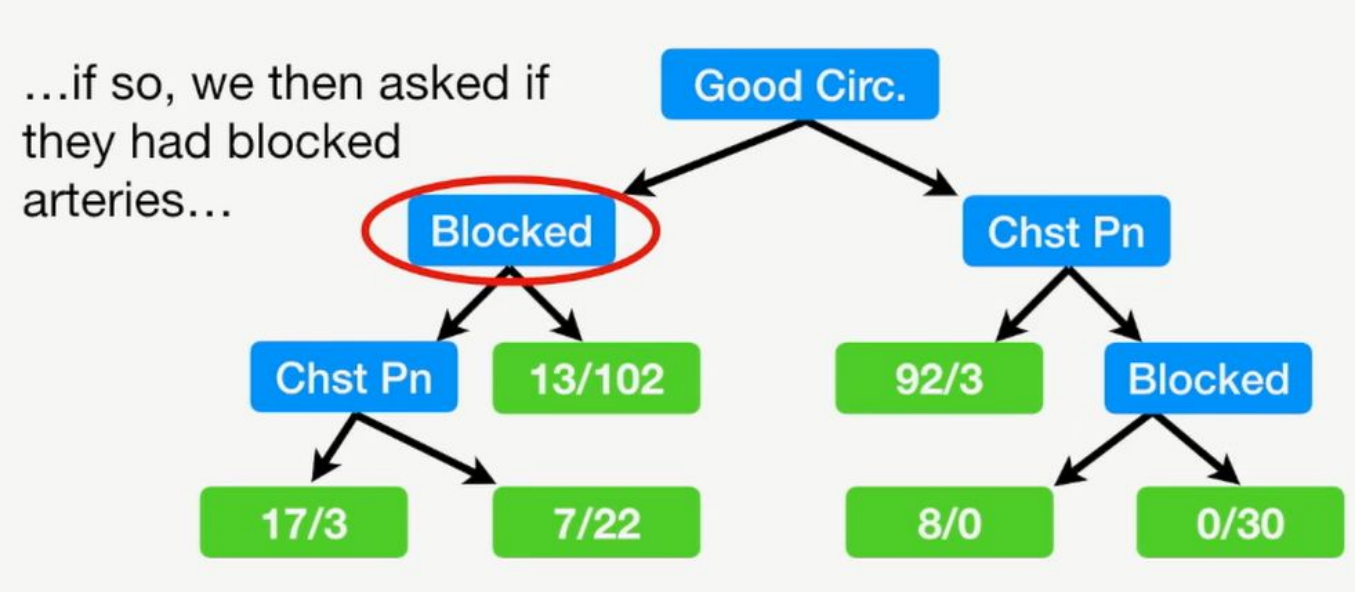
在前面的案例中，我们使用了一组数据来构建决策树，这组数据中有三个变量，分别为胸痛（chest pain）、良好血液循环（good blood circulation）和动脉阻塞状态（blocked artery status），如下图左图所示，下图右图是构建好的决策树，如下所示：



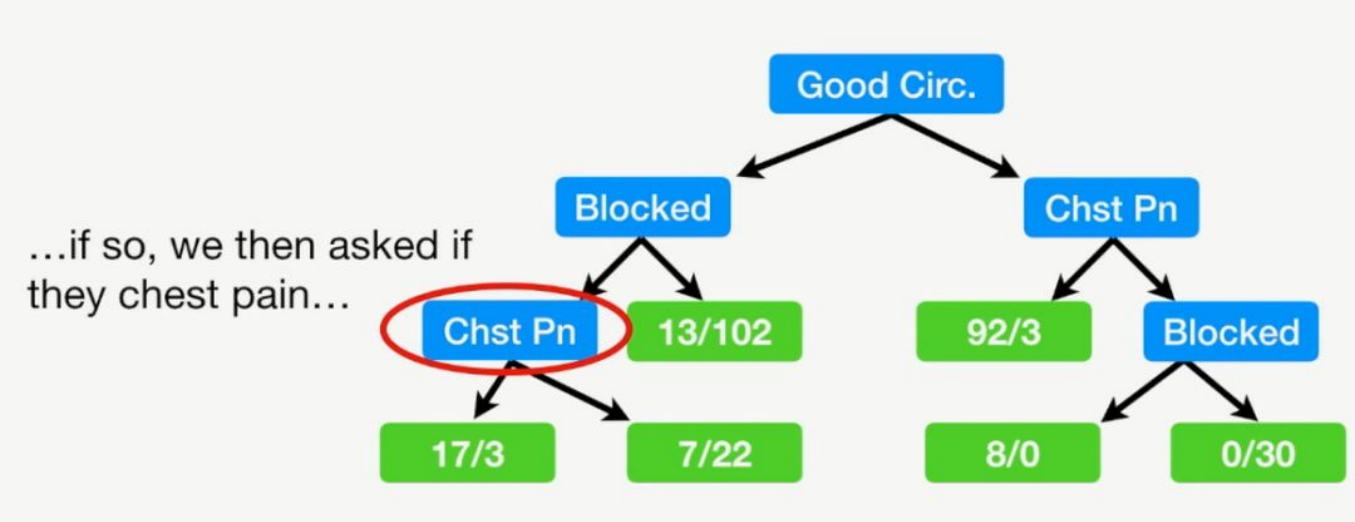

当我们构建了决策树后，我们就可以寻问患者第一个问题了，你的血液循环是否良好，如下所示：



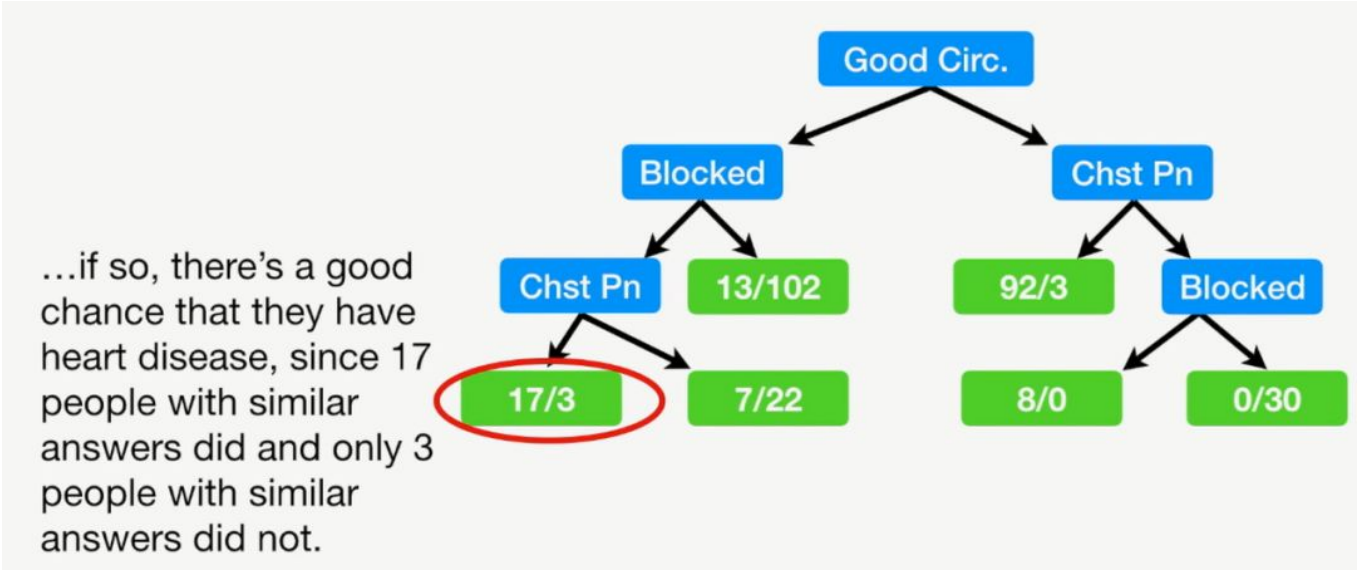
接着，询问是否出现了动脉阻塞，如下所示：



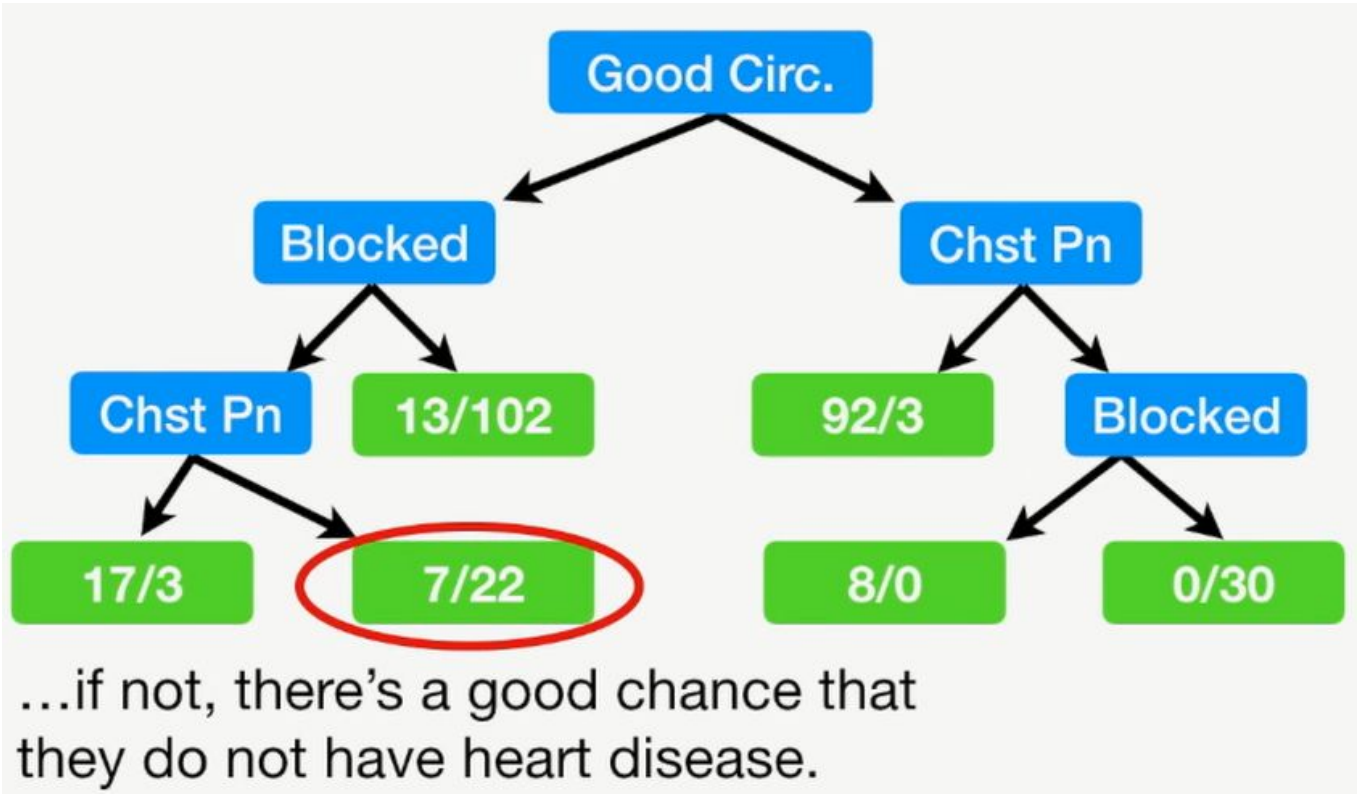

然后再询问，是否胸痛，如下所示：




如果都回答是，那么这个患者就有很大的可能性患了心脏病，因为17个类似的回答的患者都患有心脏病，仅有3个类似回答的患者没有患心脏病，如下所示：

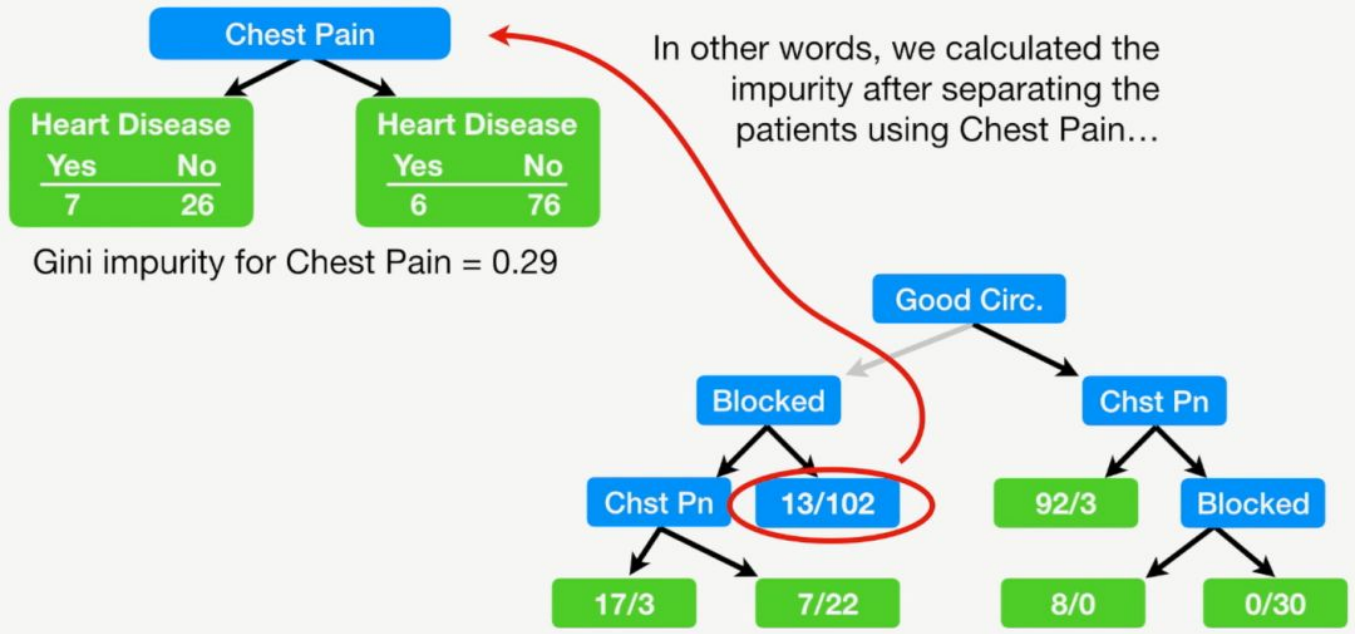



如果在是否出现胸痛这个问题时，回答是否，那么这个患者就不太可能患有心脏病，如下所示：



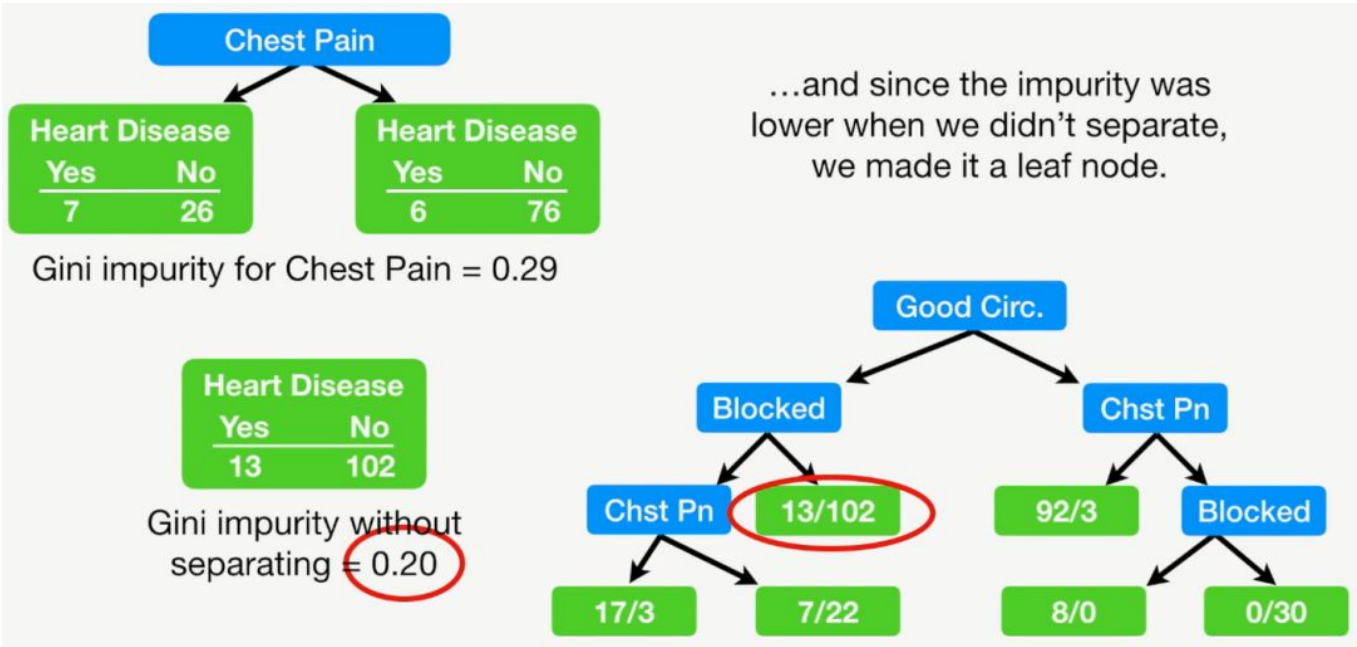

我们再往后退一层，如果一个患者回答说他的血液循环良好，并且没有出现动脉阻塞，此时我们就不需要再问他是否出现胸痛了，因为在这一层中计算所得的不纯度更低，如下所示：


换句话讲，当我们在这个节点中使用了胸痛这个变量进行区分后，它的不纯度得分为0.29，如下所示：

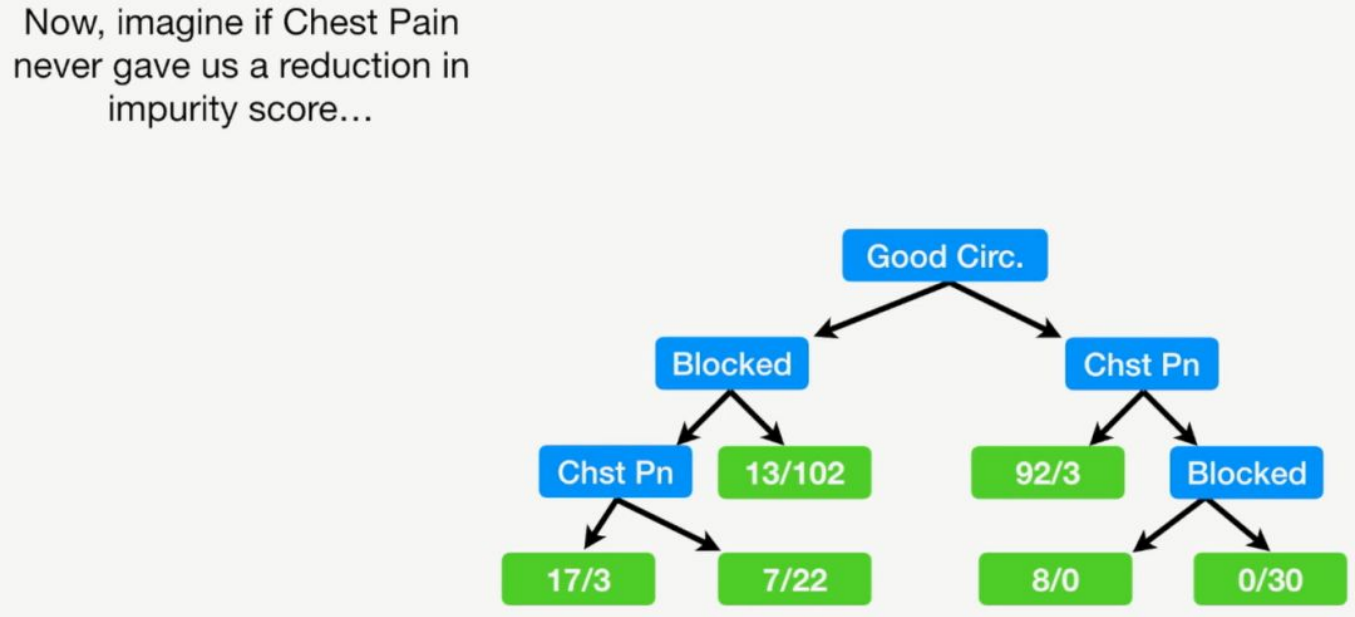



如果不用胸痛这个变量进行区分，它的不纯度得分只有0.20，如下所示：

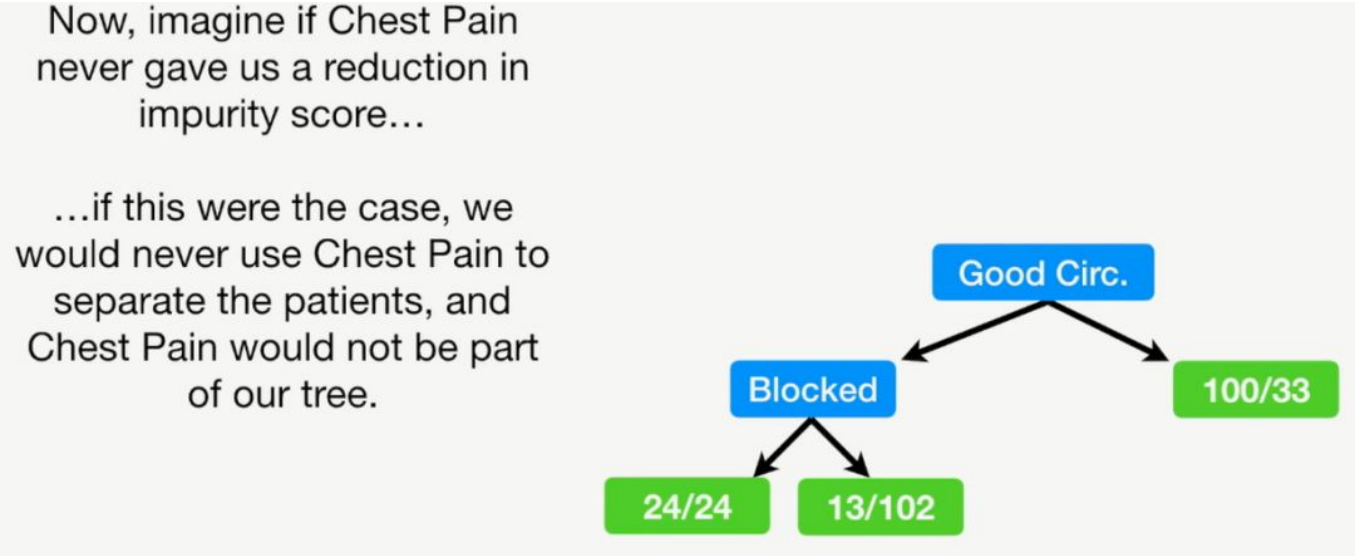

由于我们不用胸痛这个变量进行区分时，它的不纯度得分更低，因此它就是一个叶子节点（也就是说不要进行进一步的区分），如下所示：



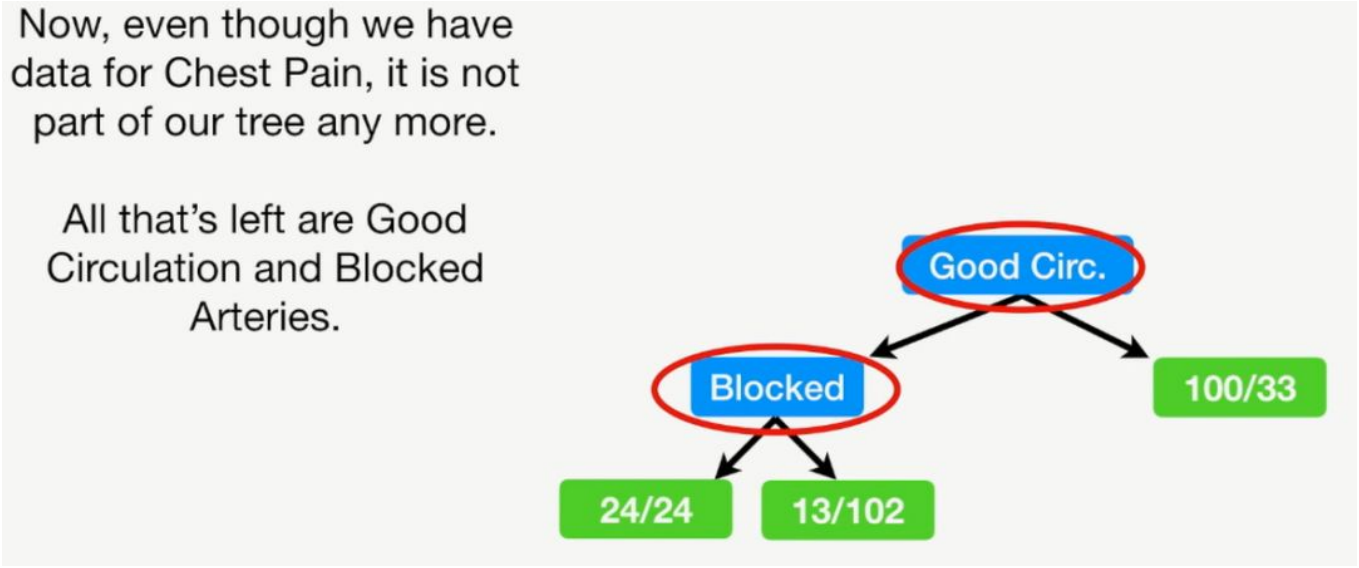

此时，我们试想一下，如果胸痛这个变量从来就不能给出一个更低的不纯度，如下所示：



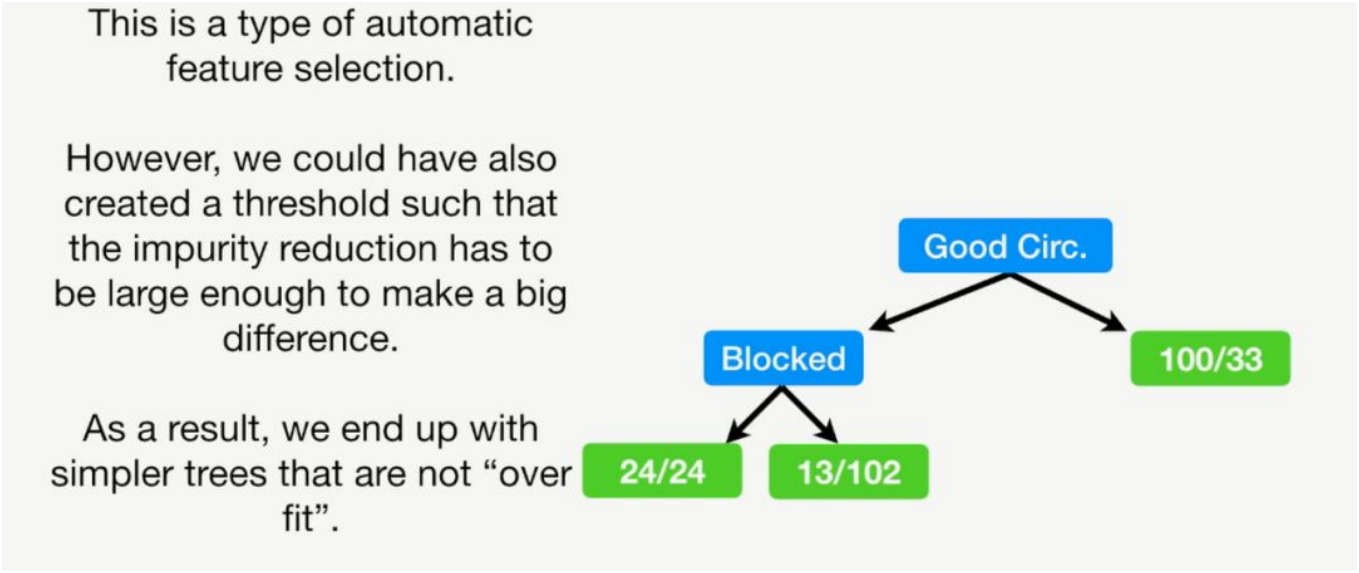

在这种情况下，我们就不要用胸痛这个变量来区分患者了，我们就要把这个变量从决策树中除去，如下所示：



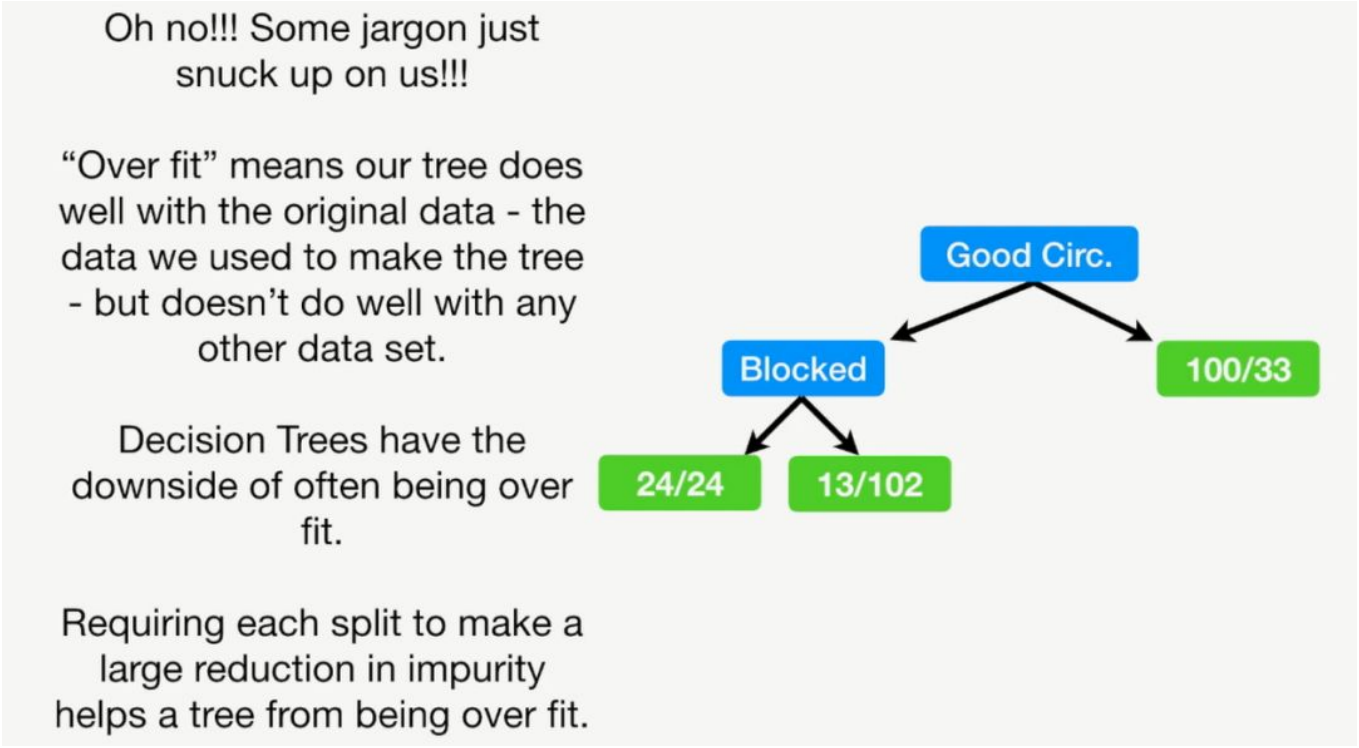

即使我们有胸痛这个变量，也不要添加到决策树上，此时决策树上只剩了血液循环和动脉阻塞这两个变量，如下所示：



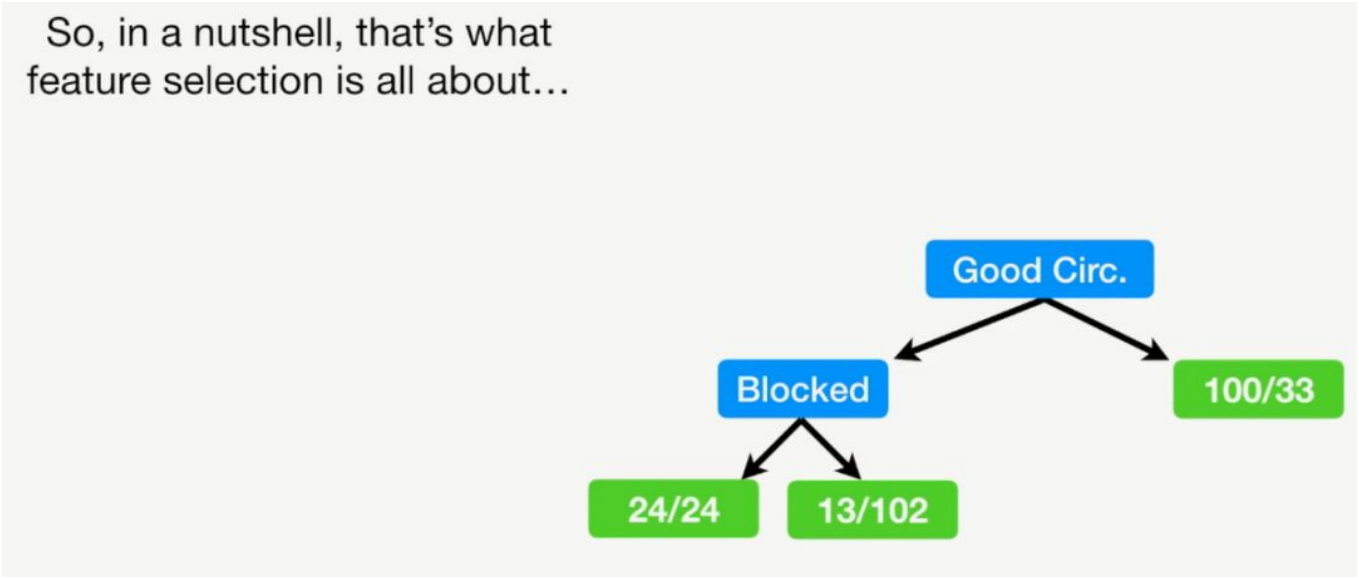

这就是一种自动特征选择，但是，我们还是需要设定一个阈值，从而使不纯度降低到足以能够产生很强区分的程度，最终我们会构建出一个更加简洁的决策树，并且这个决策树不出现“过拟合”，如下所示：




这个提到了一个术语，“过拟合”（over fit），过拟合指的是我们构建的决策树与原始数据（就是构建决策树的数据）匹配得很好，但是这个决策树却与其他的数据（我们拿到一个新的数据，放到这个决策树中）无法很好的匹配，决策树如果构建不好，很容易出现的毛病就是“过拟合”，如果要解决这个问题，就需要大幅度地降低决策树的不纯度，从而避免决策树过拟合，如下所示：

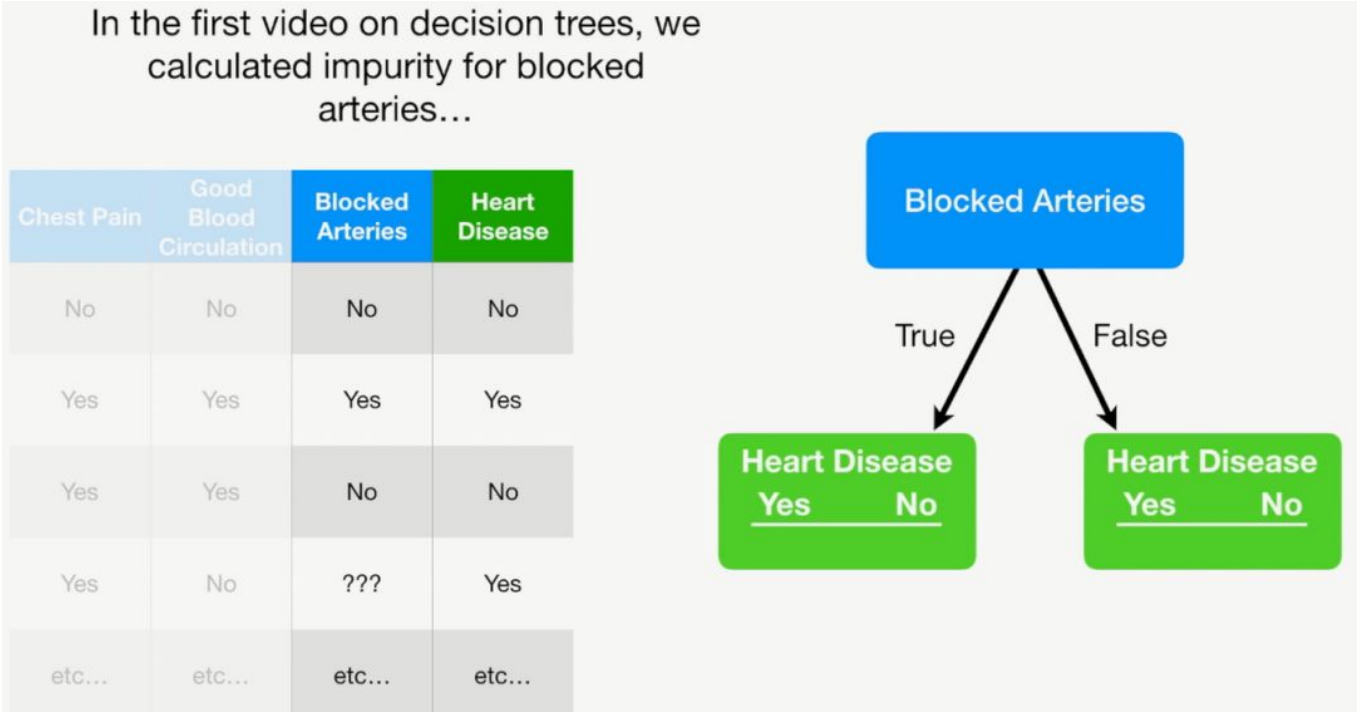



因此，简单来说，这就是特征选择要做的内容，如下所示：

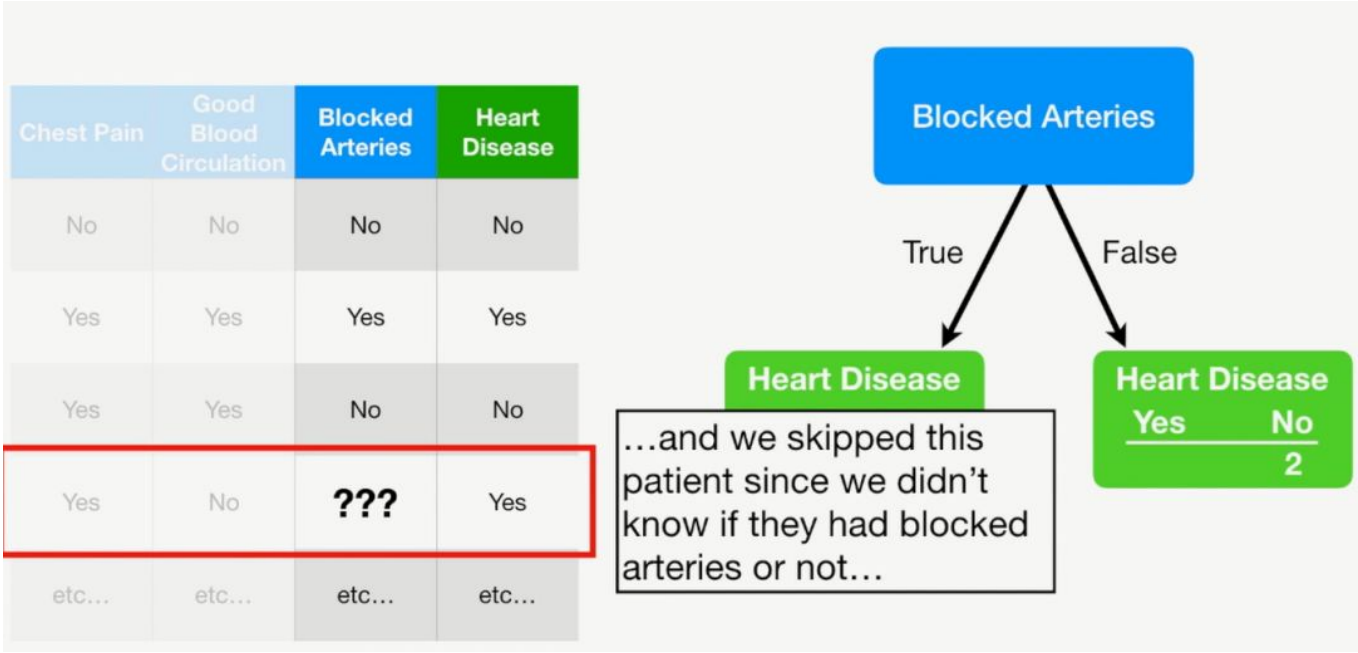



## 缺失值的处理

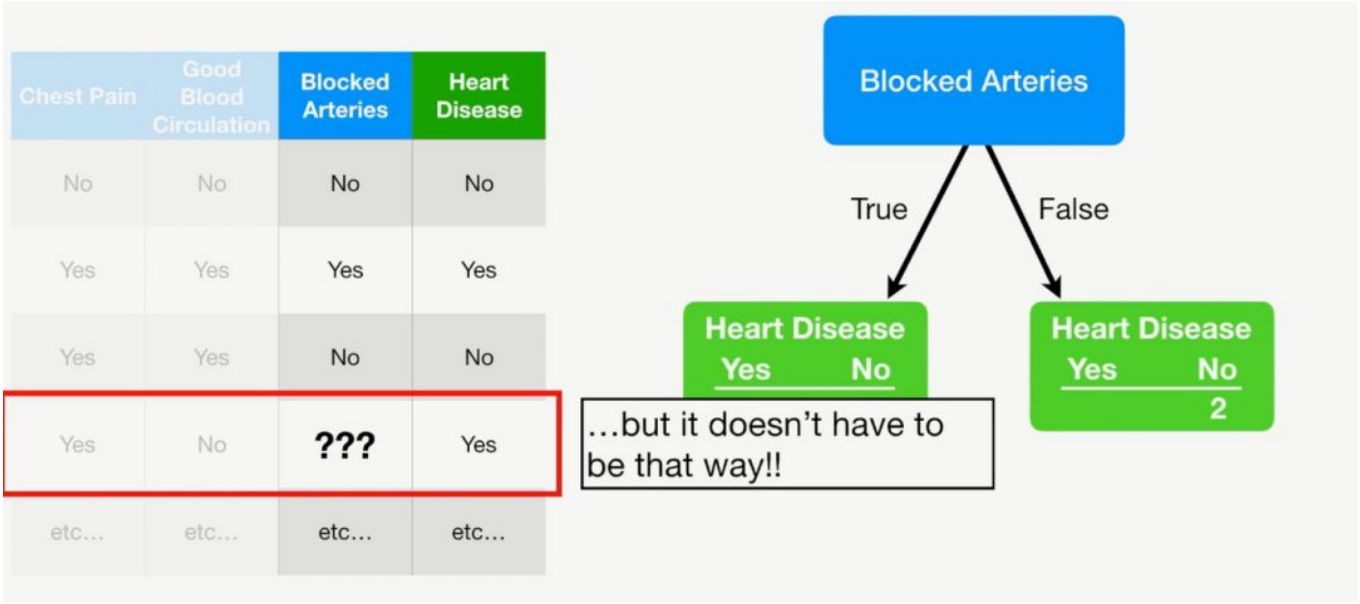
在前面的案例中，我们看到了动脉阻塞这个变量中有个缺失值，如下所示：



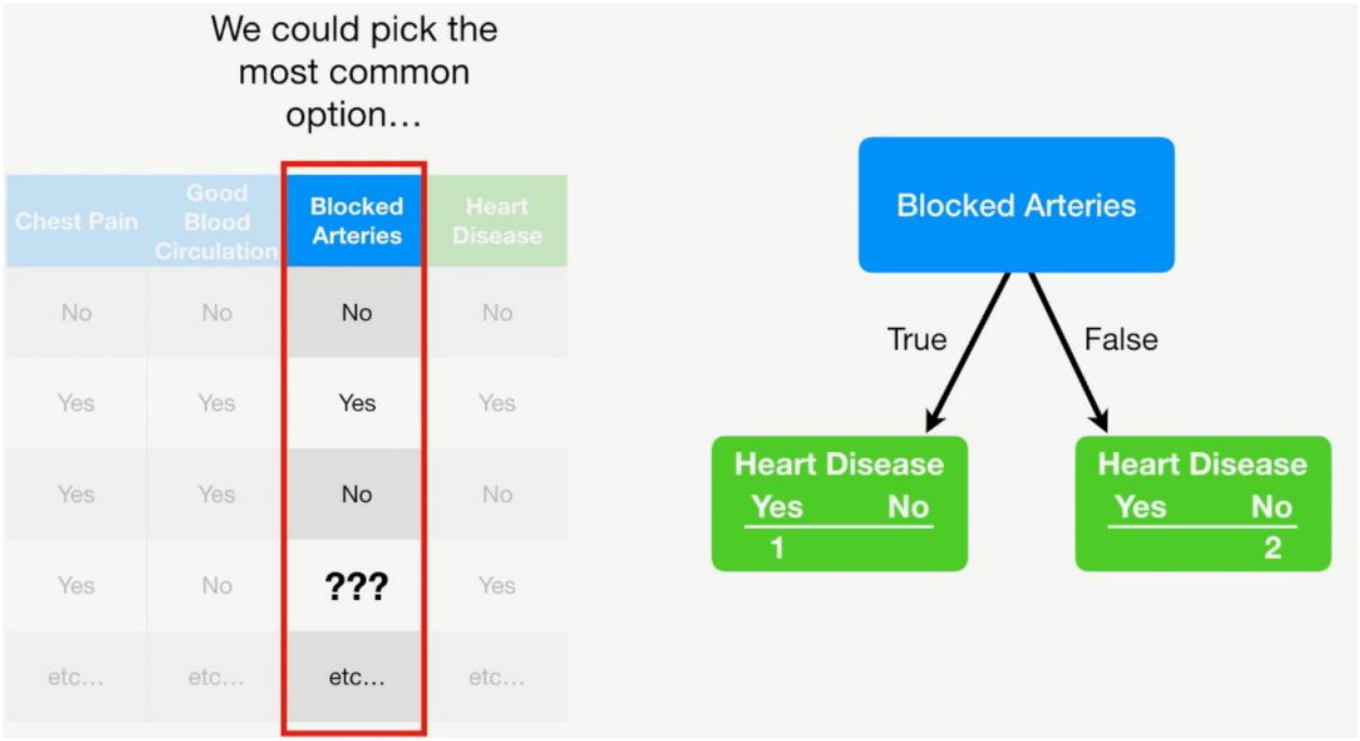

这个缺失值是在第4行，也就是第4个患者中，当我们统计到第4个患者时，就会遇到，此时我们就会跳过这个患者，因为我们不知道这个缺失值原来是Yes，还是No，如下所示：



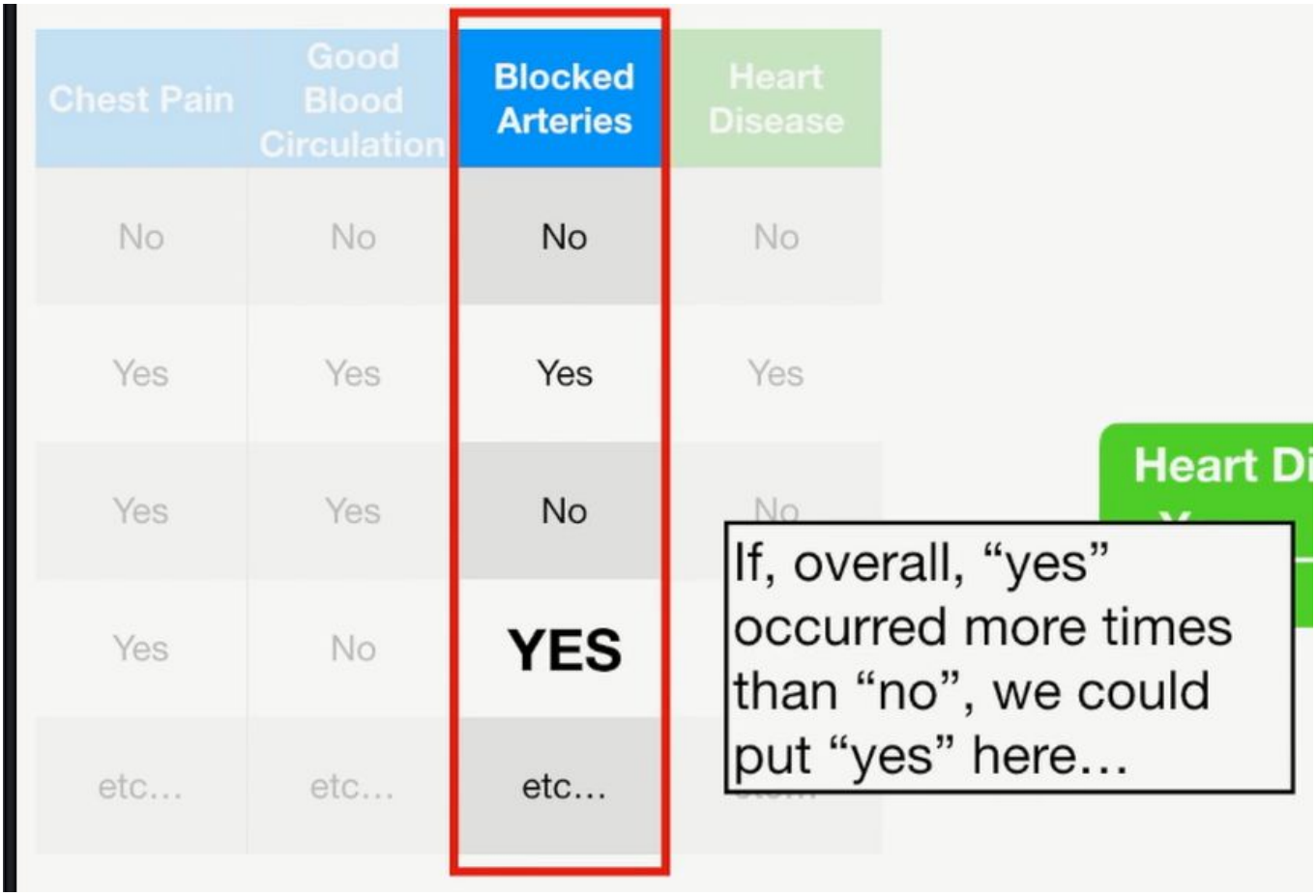

但是，并非只有一种跳过这种处理方式，如下所示：



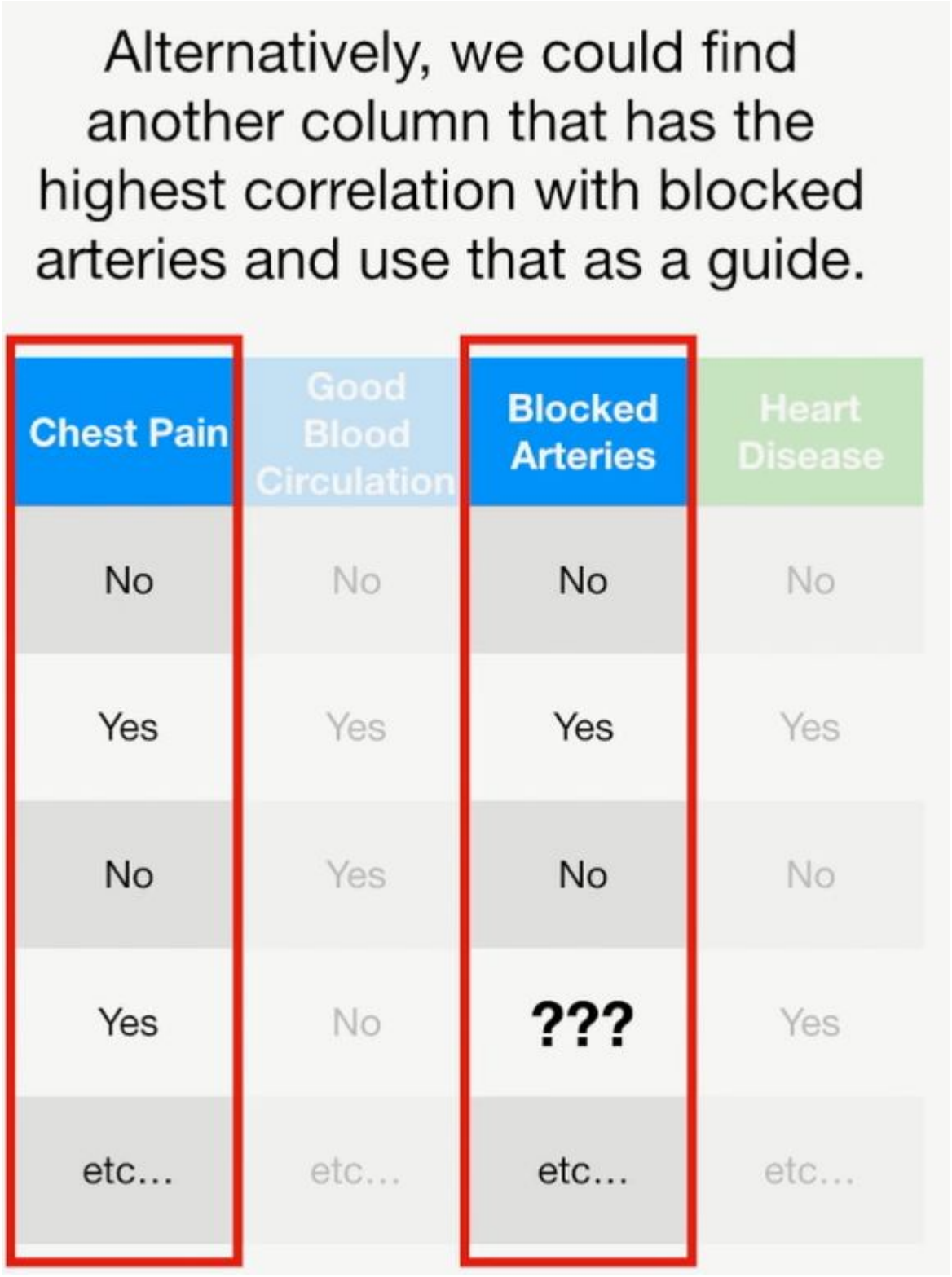

我们可以在这个变量中，找到哪个变量（yes或no）出现的频率最高，就选它即可，如下所示：



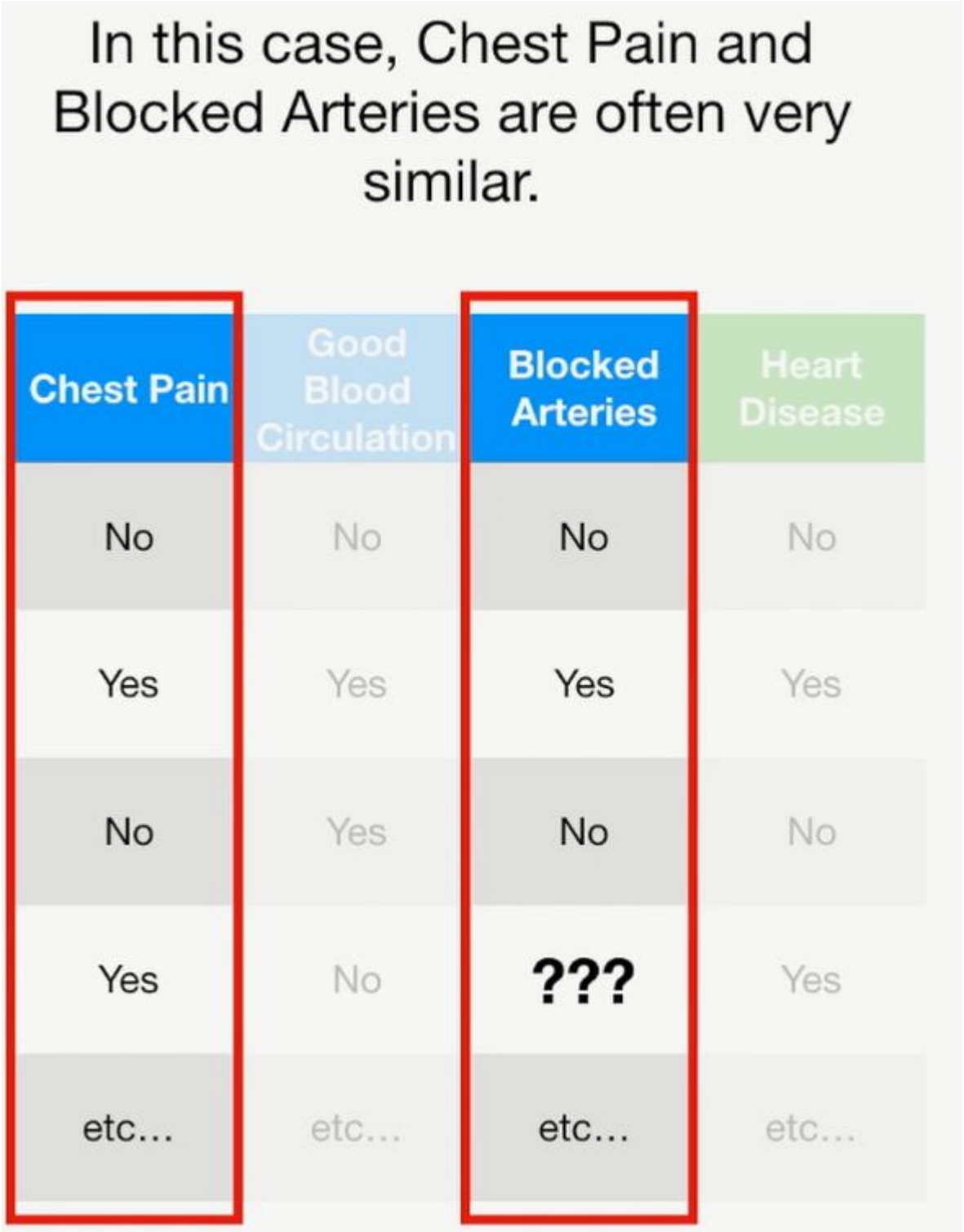

如果在动脉阻塞这个变量中，yes出现的频率更高，那么我们就把这个缺失值当作yes，如下所示：



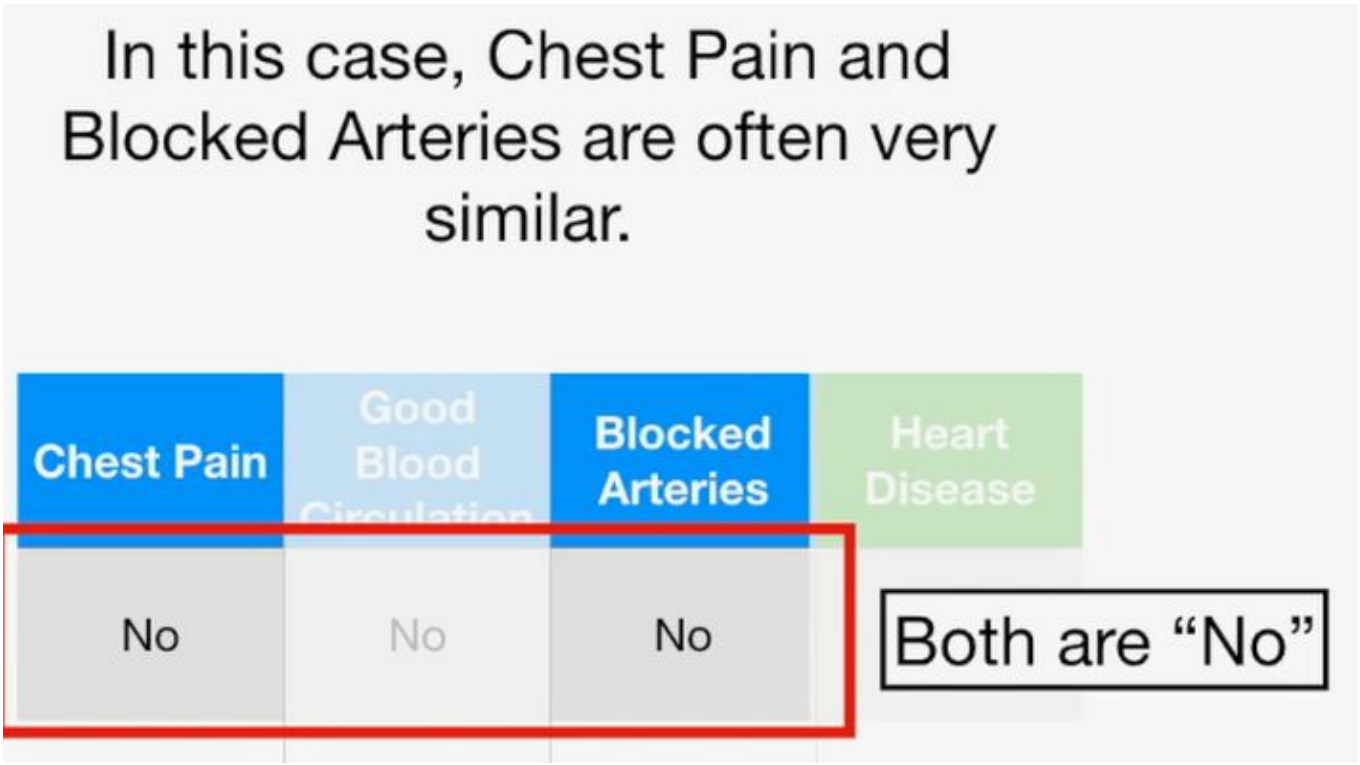

此外，我们还可以找到另外一列（例如第1列的胸痛）与这一列（动脉阻塞）关系，这种关系可以为我们提供指导，如下所示：




在这个案例中，我们发现，胸痛与动脉阻塞通常是相似的，如下所示：




例如，第1行，两个变量都是No，如下所示：

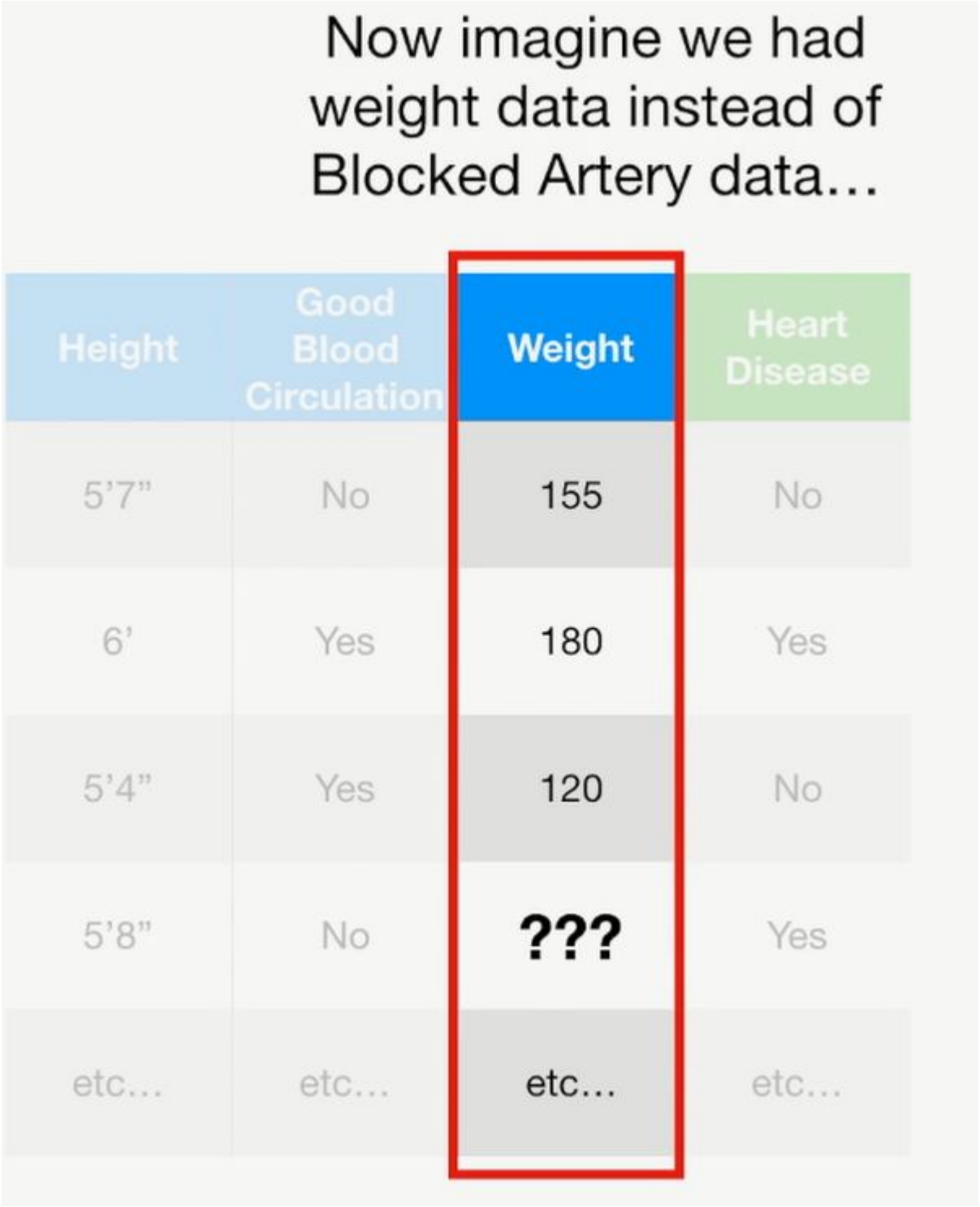



第2行，两个变量都是Yes，如下所示：

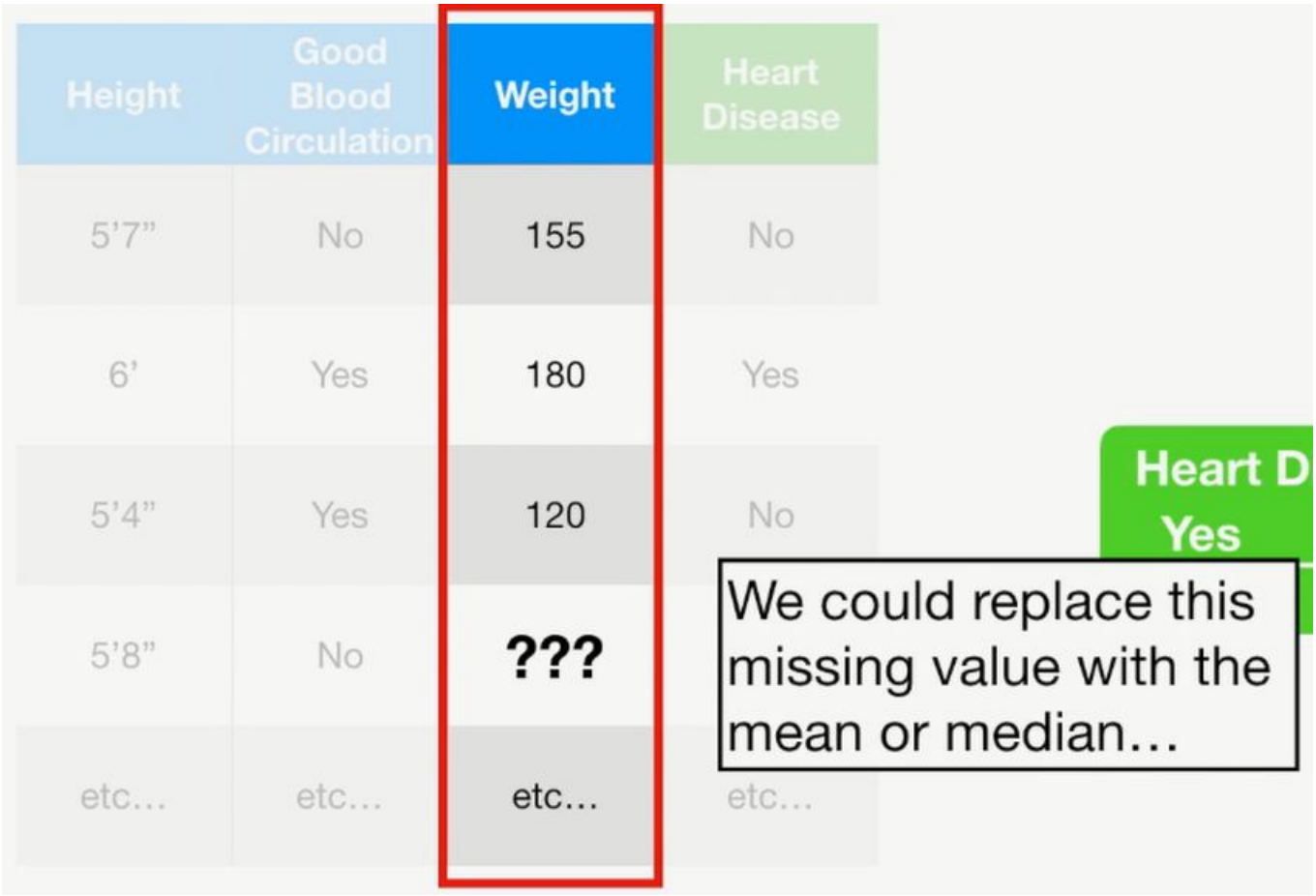

第3行，两个变量都是No，如下所示：

因此，我们就可以在这个缺失值上填上Yes，如下所示：

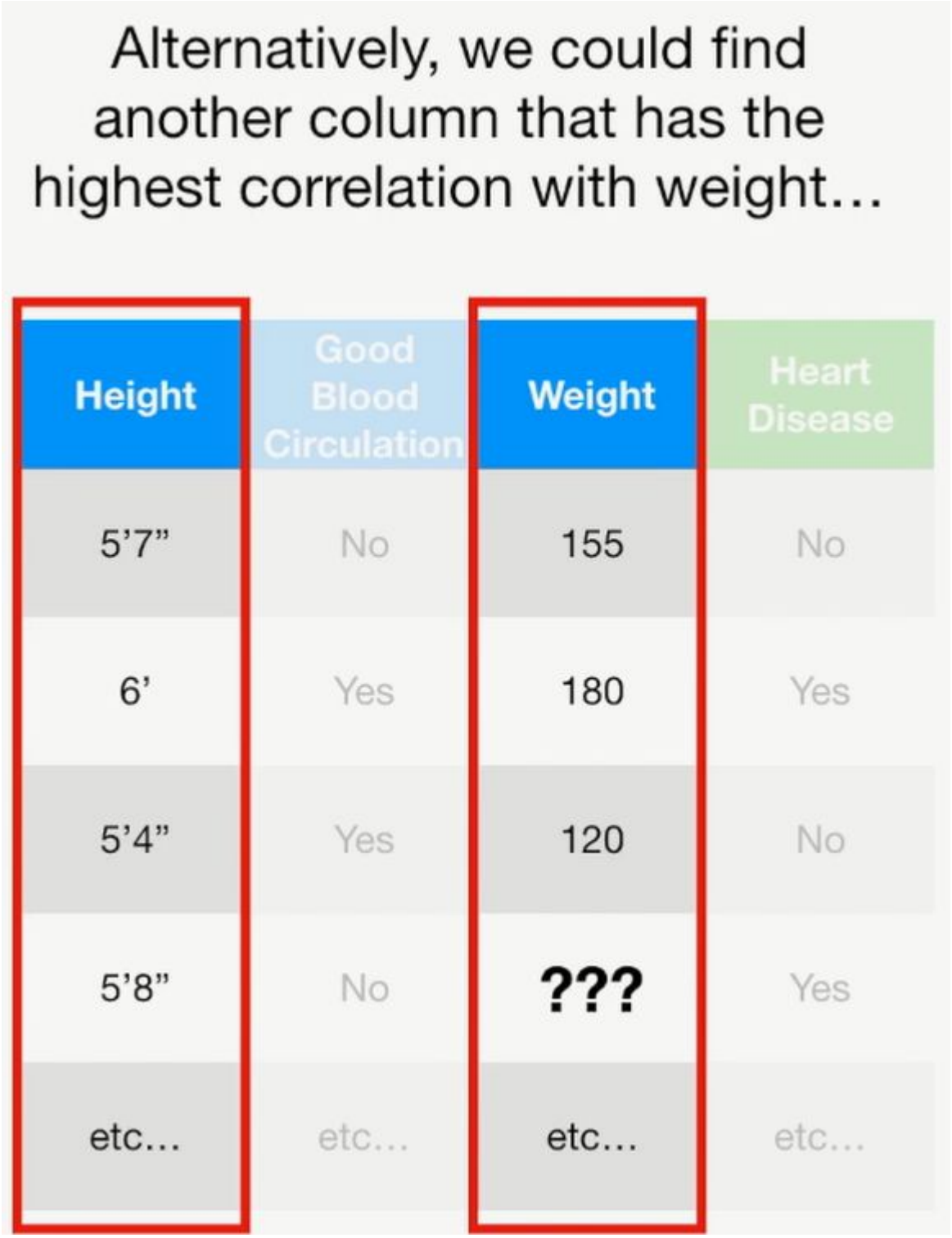

我们再考虑一种情况，假如我们此时没有动脉阻塞这个变量，这一处的变量用体重进行了替换，如下所示：



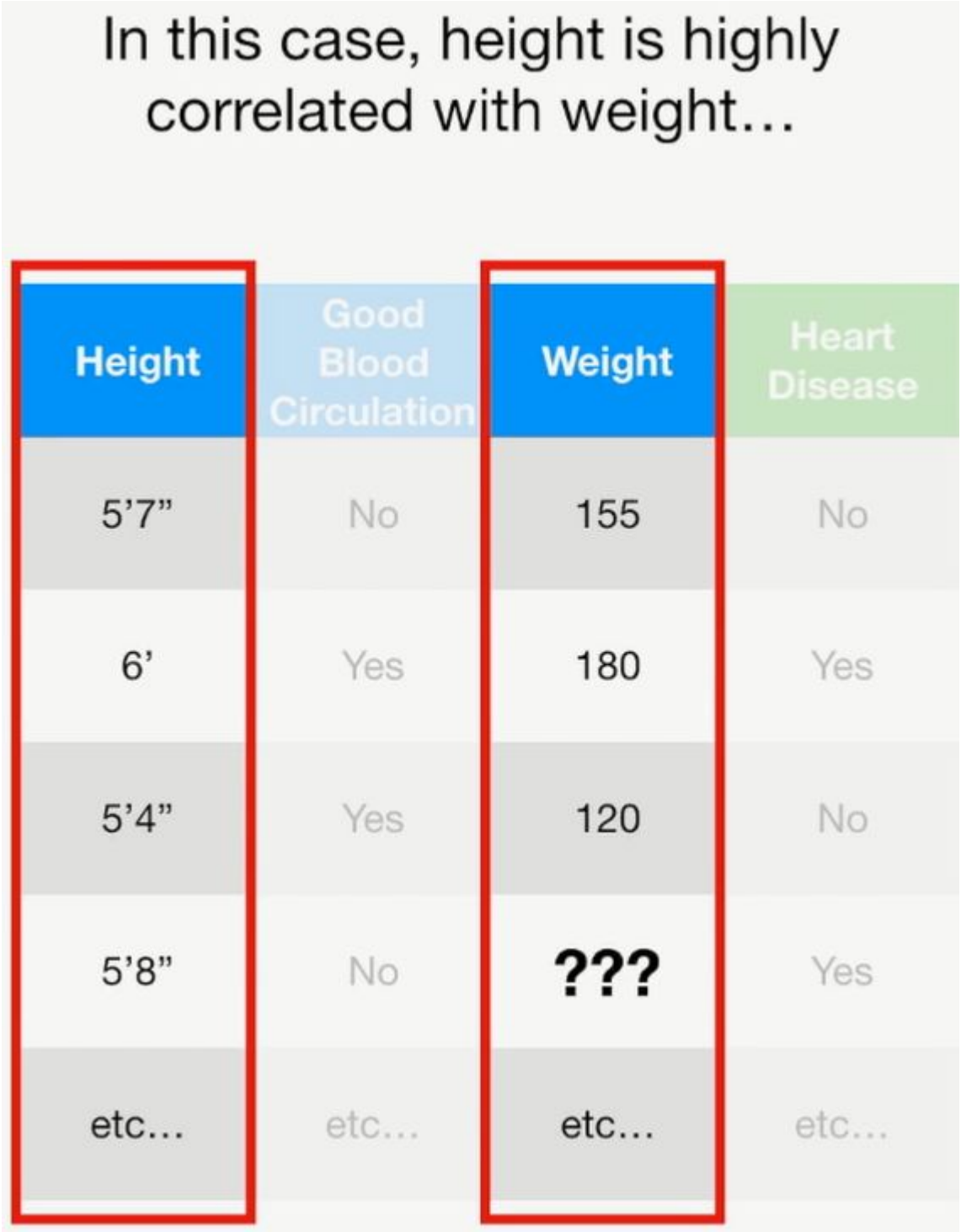

我们此时，可以使用体重这个变量的均值或中位数来替换这个缺失值，如下所示：



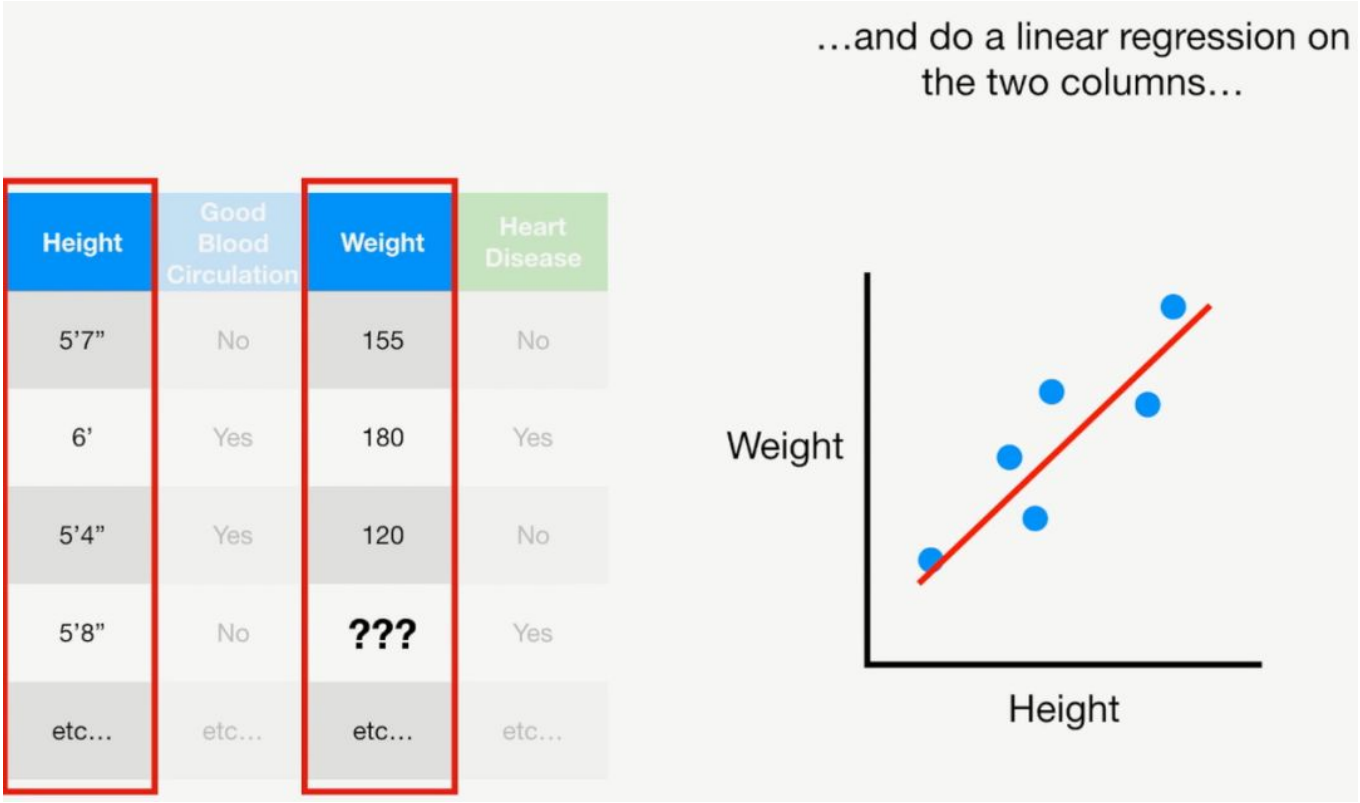

此外，我们还可以找到第1列（身高）与第3列（体重）之间的关系，用于确定体重的缺失值，如下所示：



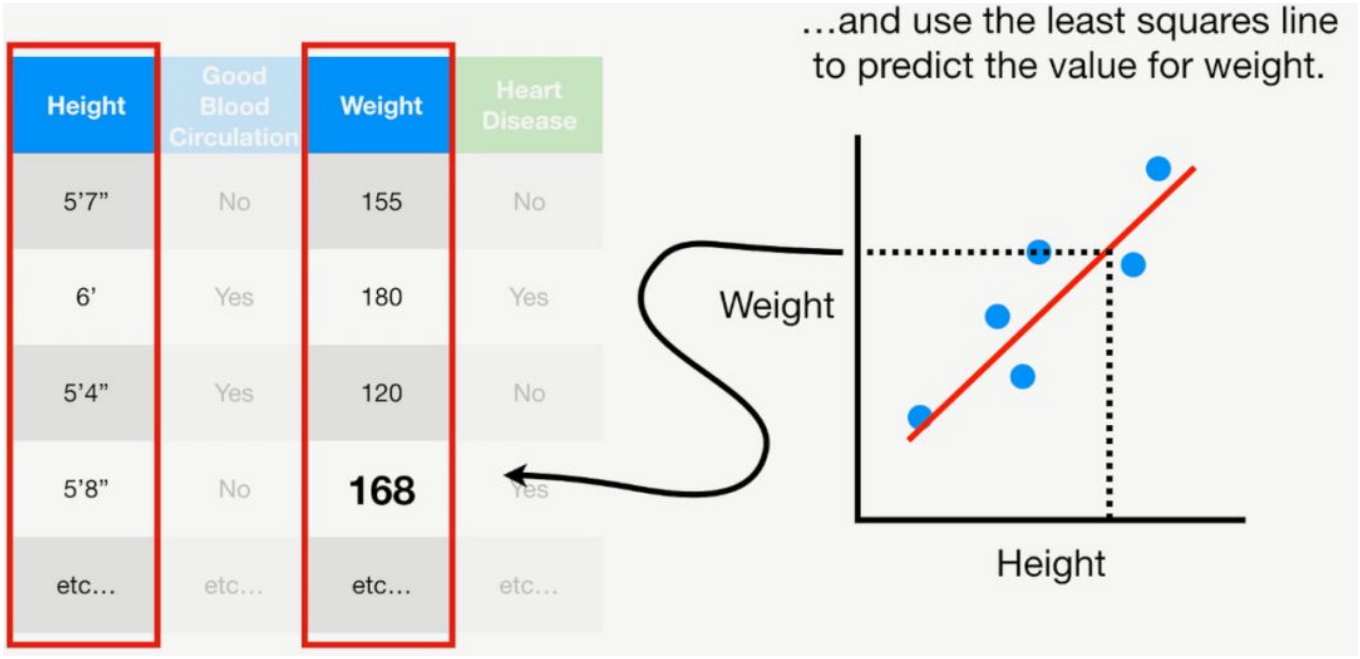

在这个案例中，我们发现身高与体重高度相关，如下所示：




由于它们都是连续型变量，因此我们可以做一下线性回归，求出这个值，如下所示：


…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…
Weight
Height

最终的结果如下所示：


…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 168
etc… etc… etc… etc…
Weight
Height