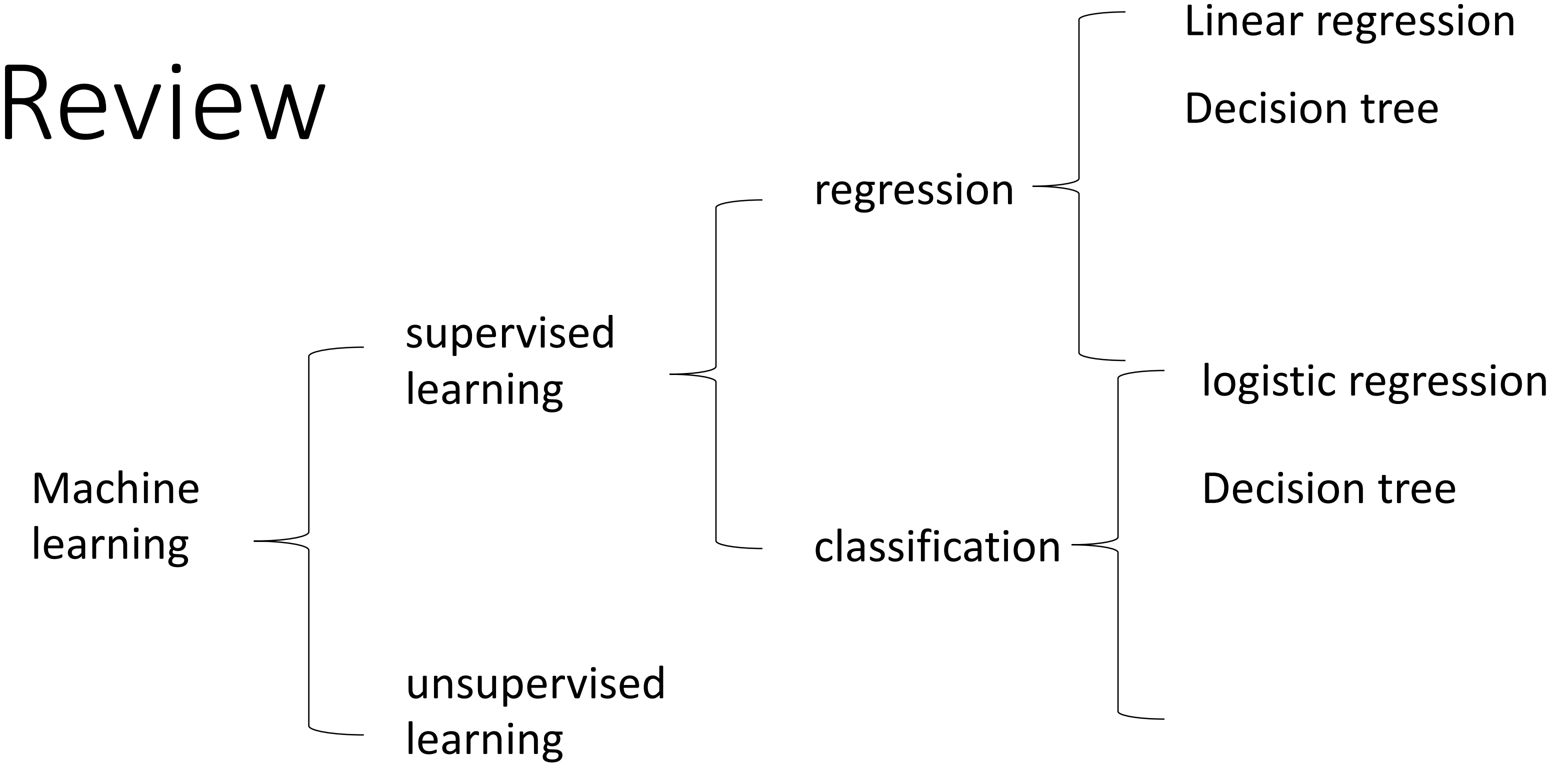
# L13 Decision Tree

Prof. Xun Jiao

# Before class

- HW3 (Mini Project) Due 3pm Mar. 7
- Mar. 7 – Mar. 14: No HW, review for Test
- Test on Mar. 14
    - Examples/Exercise in class
    - HWs

# Review



Machine learning

supervised learning

unsupervised learning

regression

classification

Linear regression

Decision tree

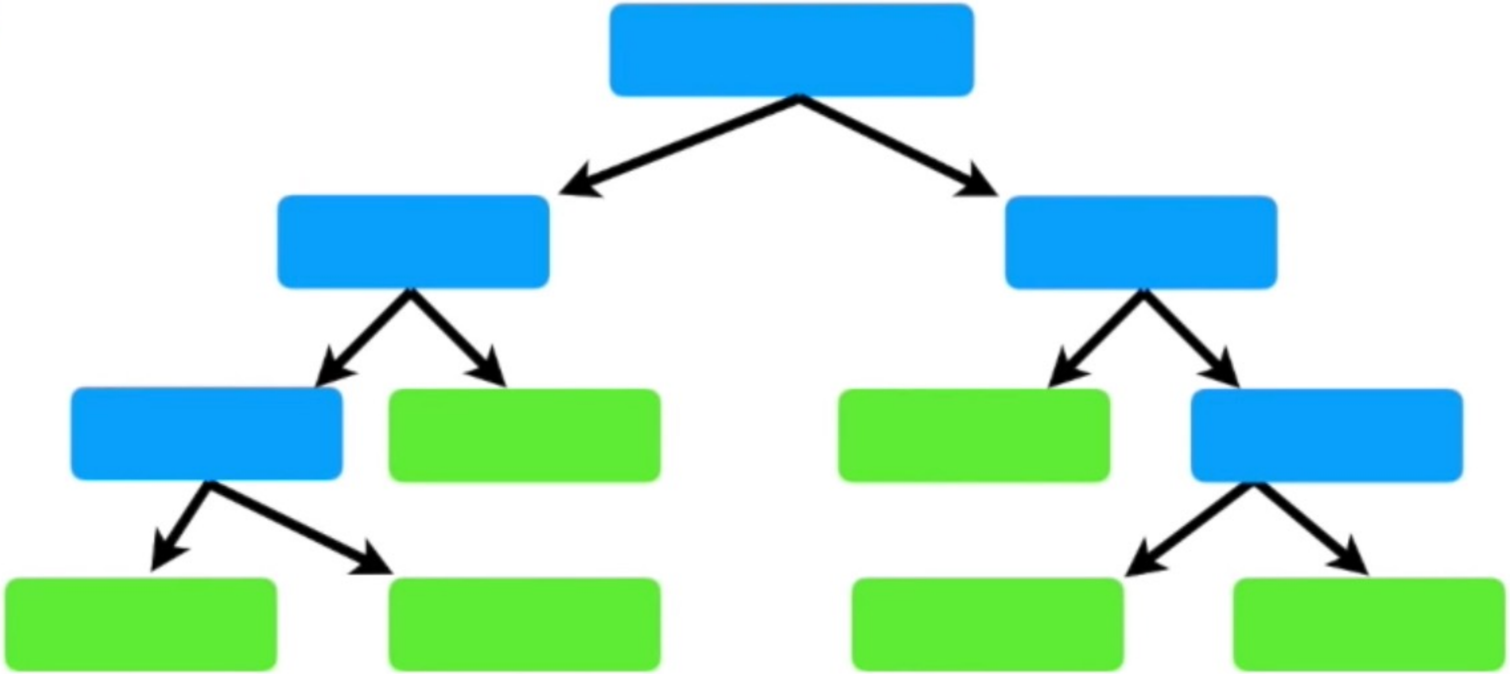
logistic regression

Decision tree

Now we are ready to talk about how to go from a raw table of data…

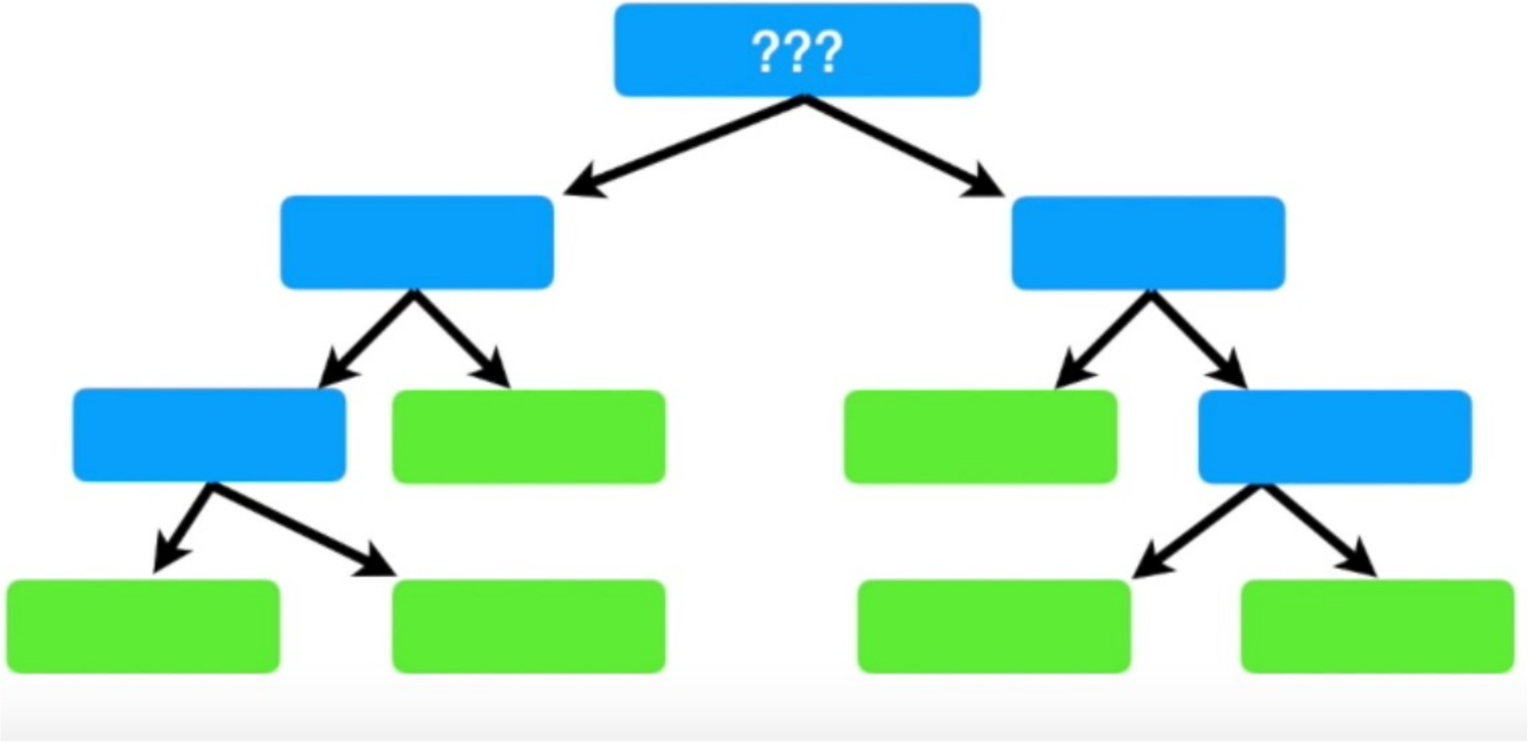
| Chest Pain | Good Blood Circulation | Blocked Arteries | Heart Disease |
|---|---|---|---|
| No | No | No | No |
| Yes | Yes | Yes | Yes |
| Yes | Yes | No | No |
| Yes | No | ??? | Yes |
| etc… | etc… | etc… | etc… |

…to a decision tree!!!

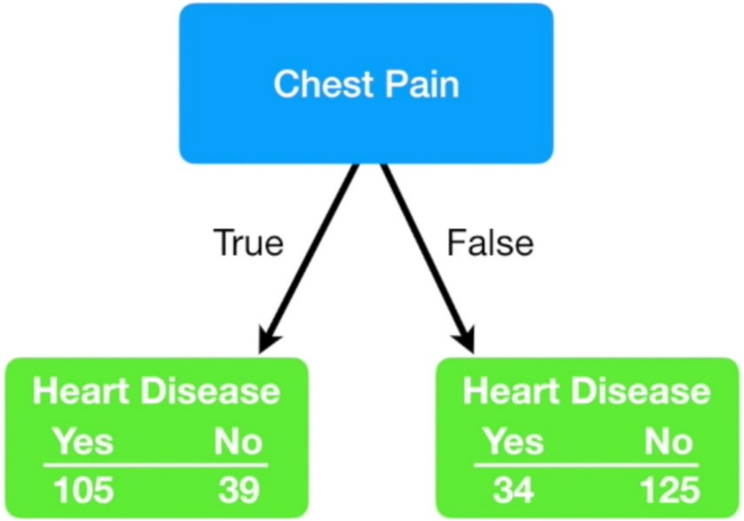


# First Step: who would be the top (root)?

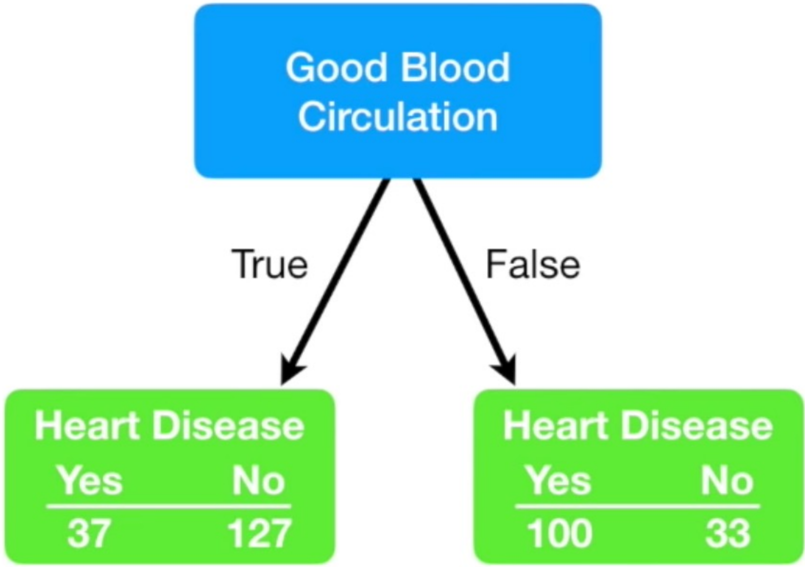
- Chest pain? Blood circulation? Blocked Arteries?
-



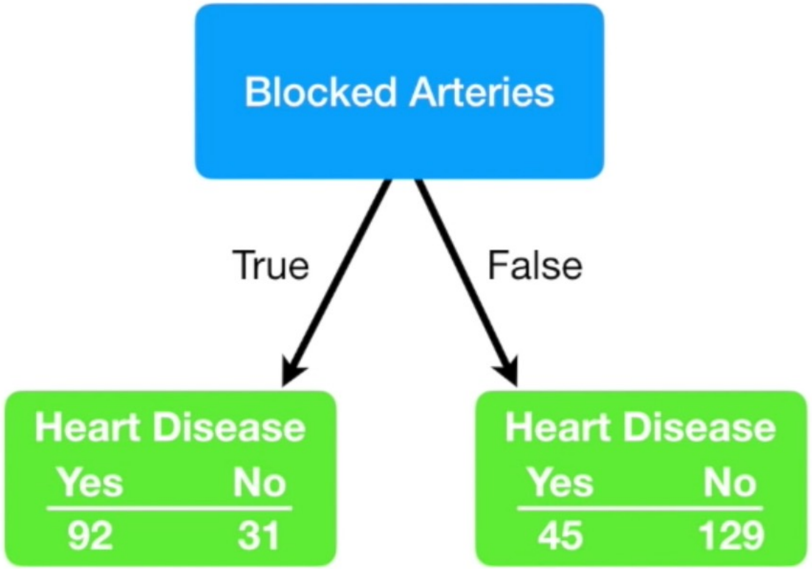
# Which one has best separation ability?



Good, but not perfect:
Can mostly separate, but still
some errors



Good, but not perfect

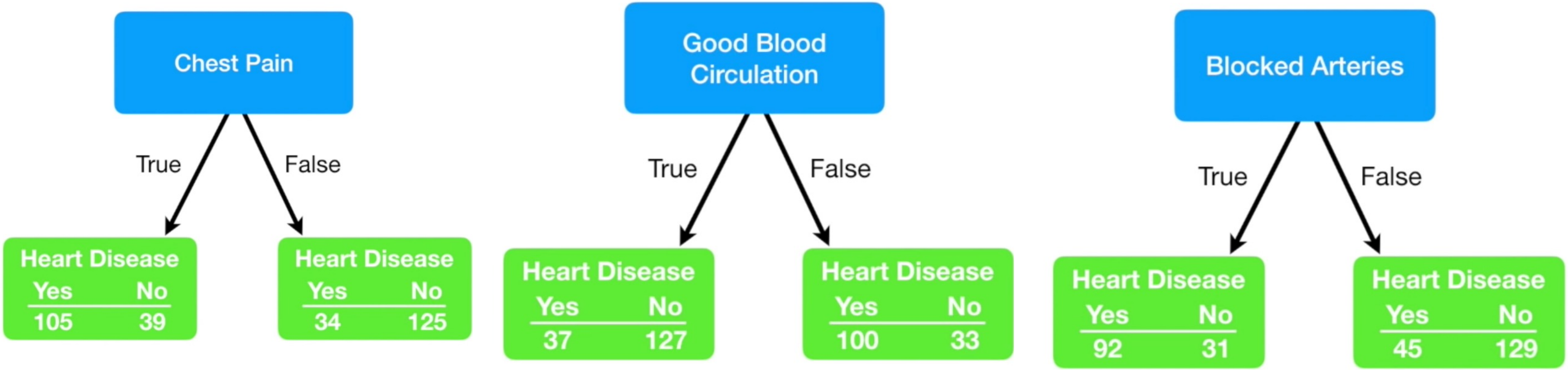


Good, but not perfect

# So, how can we quantify their separation ability?

- Impurity!!!

# These nodes are all IMPURE



Chest Pain

True

False

Heart Disease

| Yes | No |
|---|---|
| 105 | 39 |

Heart Disease

| Yes | No |
|---|---|
| 34 | 125 |

Good Blood Circulation

True

False

Heart Disease

| Yes | No |
|---|---|
| 37 | 127 |

Heart Disease

| Yes | No |
|---|---|
| 100 | 33 |

Blocked Arteries

True

False

Heart Disease

| Yes | No |
|---|---|
| 92 | 31 |

Heart Disease

| Yes | No |
|---|---|
| 45 | 129 |

Good, but not perfect:
Can mostly separate, but still some errors

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

Good, but not perfect

# Compute impurity

- Multiple options:
  - Gini
  - Entropy

**Gini index** (a criterion to minimize the probability of misclassification):

$$Gini = 1 - \sum_j p_j^2$$

binary class $$Gini = 1 - (p_1^2 + p_2^2)$$

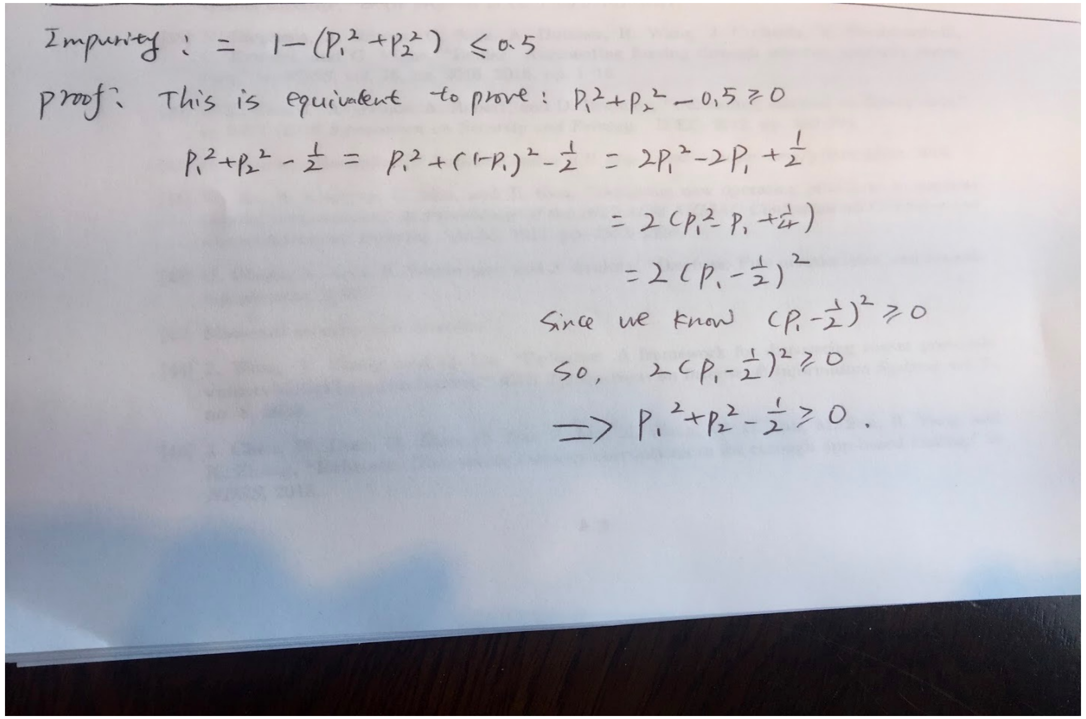
What is the range of Gini???

# What is the range of Gini???

**Gini index** (a criterion to minimize the probability of misclassification):

$$Gini = 1 - \sum_i p_j^2$$

binary class $$Gini = 1 - (p_1^2 + p_2^2)$$



Impurity : $= 1-(P_1^2+P_2^2) \leq 0.5$

proof: This is equivalent to prove: $P_1^2+P_2^2-0.5 \geqslant 0$

$$P_1^2+P_2^2-\frac{1}{2} = P_1^2+(1-P_1)^2-\frac{1}{2} = 2P_1^2-2P_1+\frac{1}{2}$$

$$= 2(P_1^2-P_1+\frac{1}{4})$$

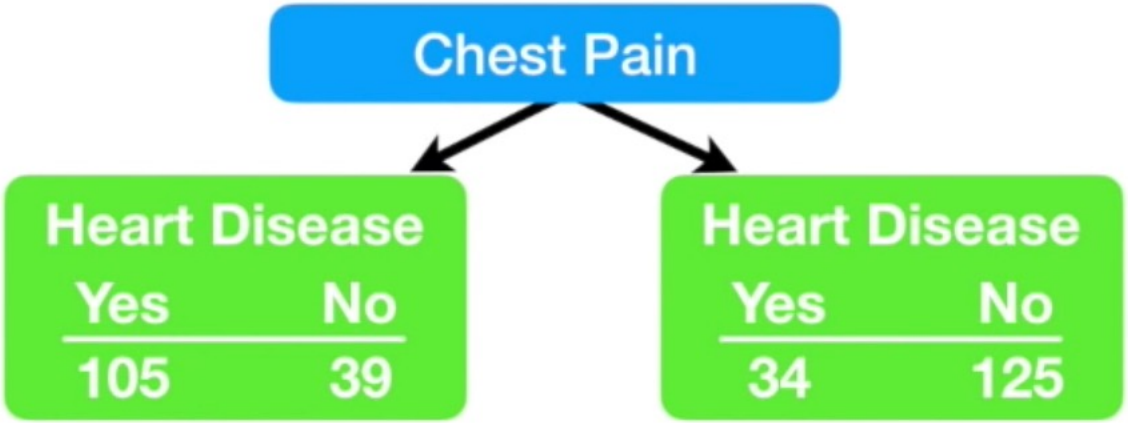
$$= 2(P_1-\frac{1}{2})^2$$

Since we know $(P_1-\frac{1}{2})^2 \geqslant 0$

So, $2(P_1-\frac{1}{2})^2 \geqslant 0$

$\Longrightarrow P_1^2+P_2^2-\frac{1}{2} \geqslant 0$.

# Compute Gini



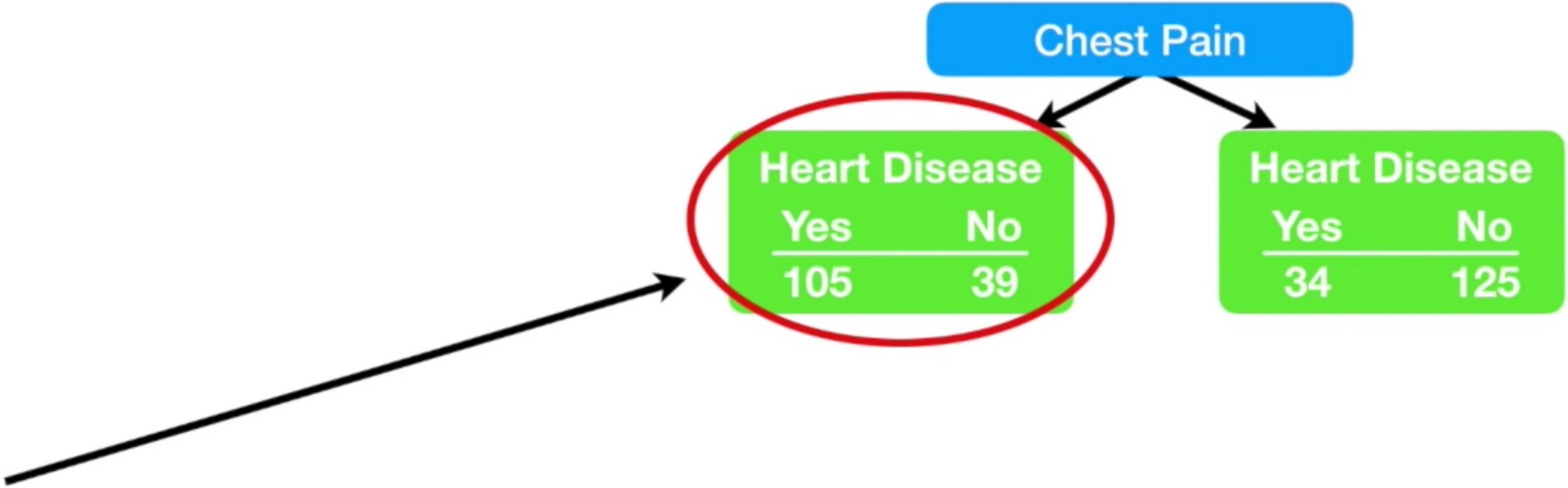
Chest Pain

Heart Disease

| Yes | No |
|---|---|
| 105 | 39 |

Heart Disease

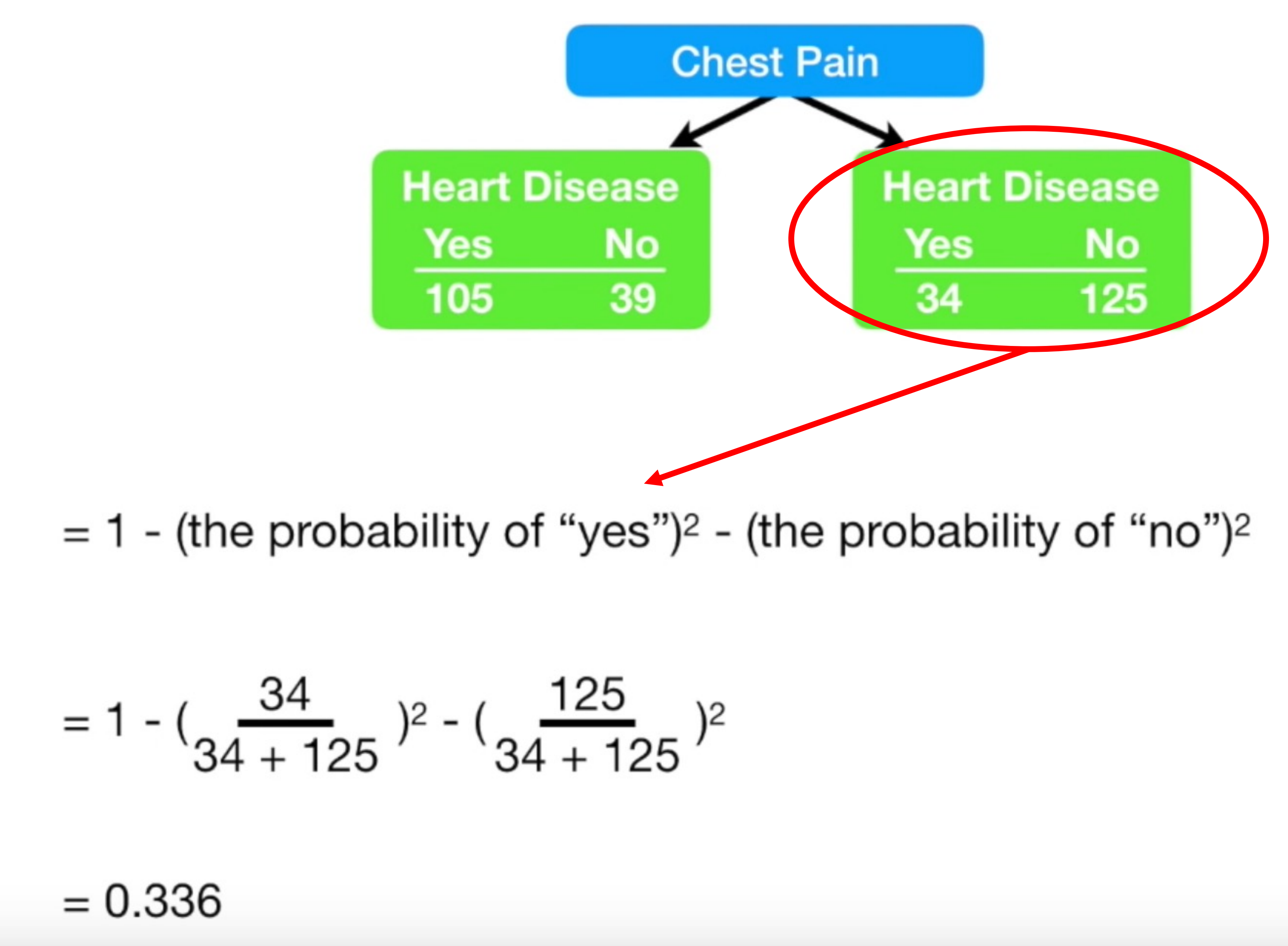
| Yes | No |
|---|---|
| 34 | 125 |



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

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

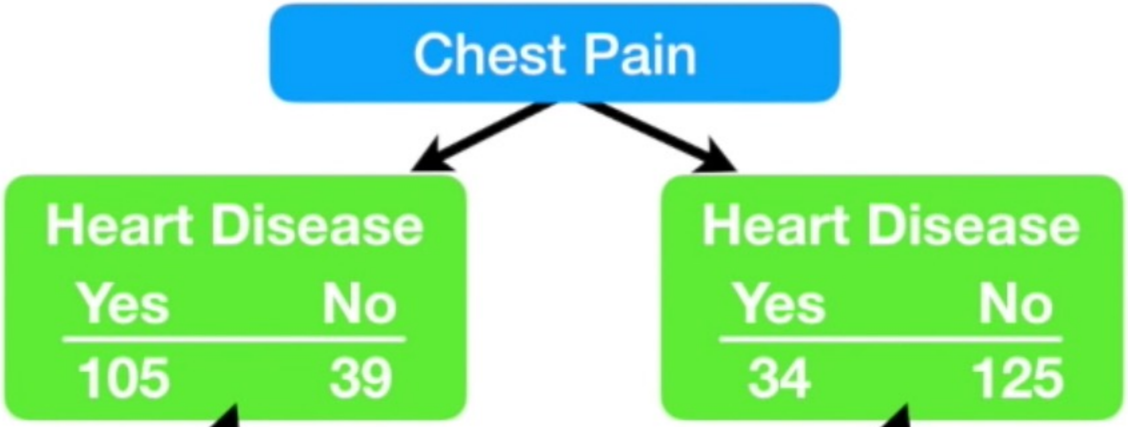
$$= 0.395$$



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

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

$$= 0.336$$



Chest Pain

| Heart Disease | |
|---|---|
| Yes | No |
| 105 | 39 |

| Heart Disease | |
|---|---|
| Yes | No |
| 34 | 125 |

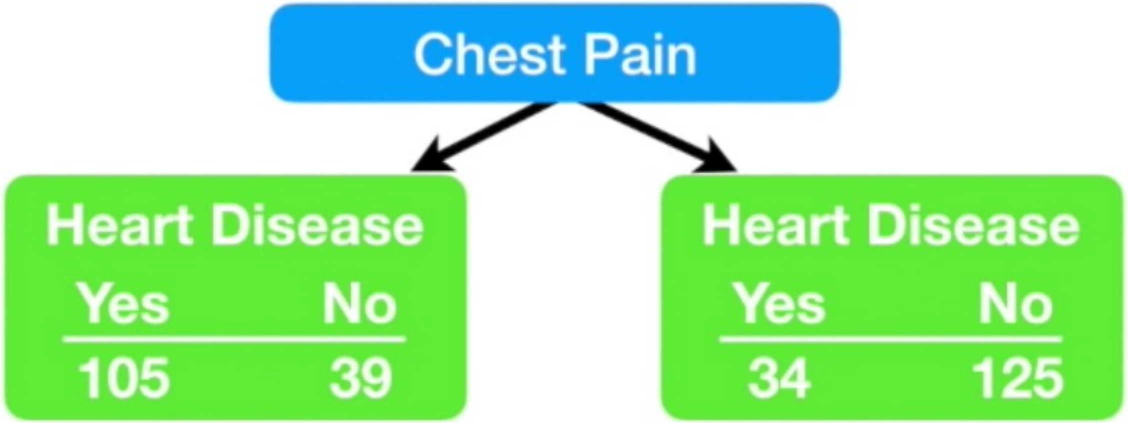
Gini impurity = 0.395

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**.



Chest Pain

| Heart Disease |  |
|---|---|
| Yes | No |
| 105 | 39 |

| Heart Disease |  |
|---|---|
| Yes | No |
| 34 | 125 |

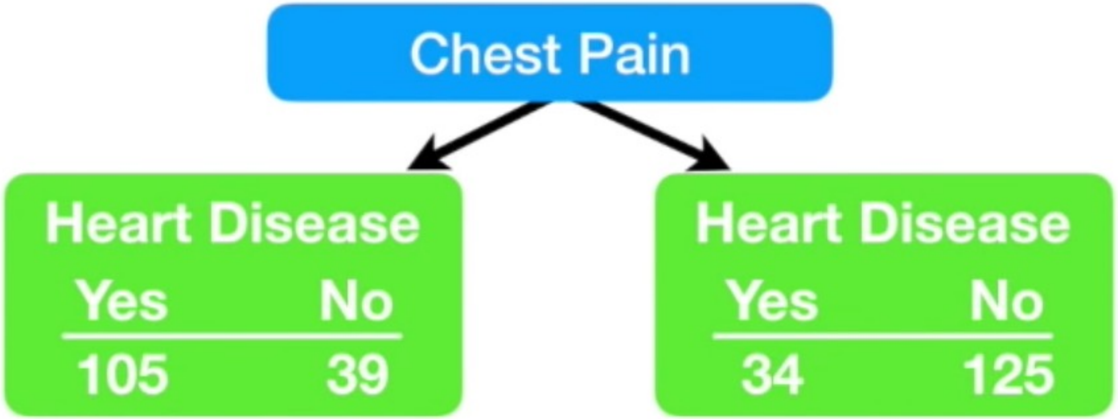
Gini impurity = 0.395    0.336

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

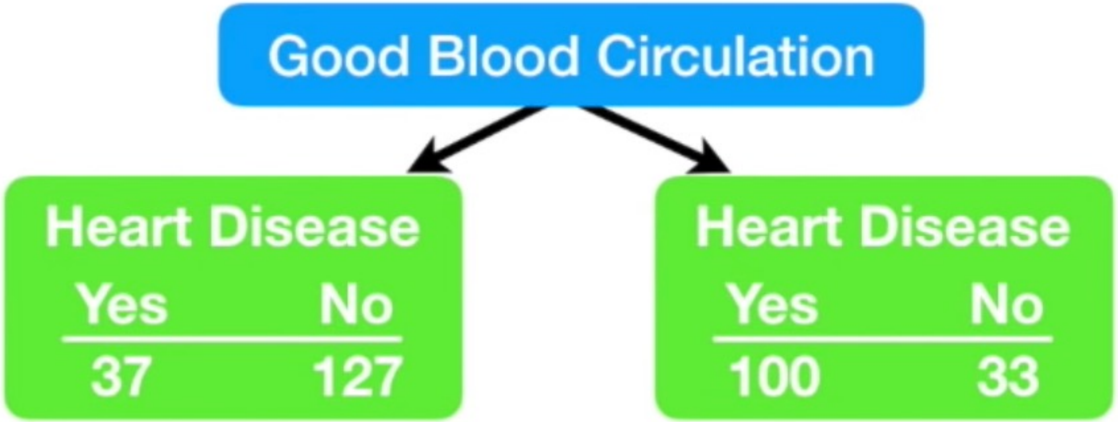


Chest Pain

Heart Disease
Yes No
105 39

Heart Disease
Yes No
34 125

Gini impurity for Good Blood Circulation = 0.360

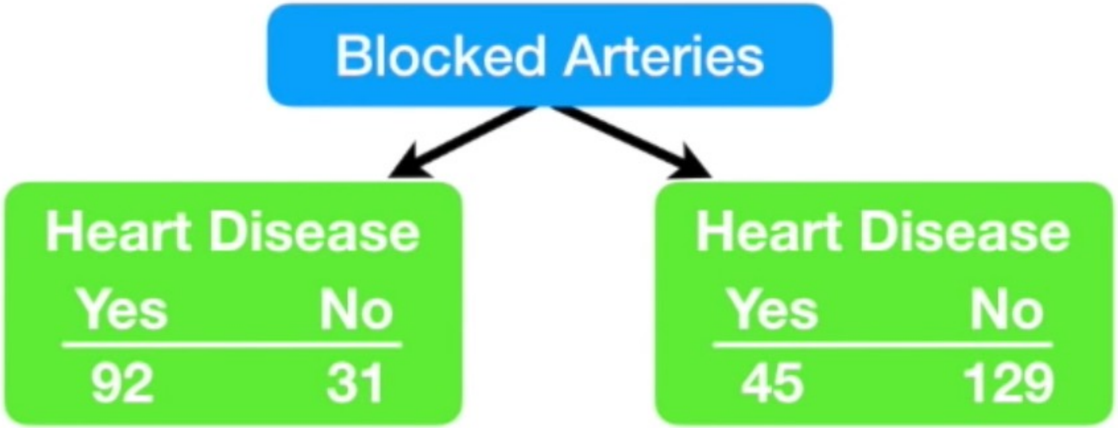


Good Blood Circulation

Heart Disease
Yes No
37 127

Heart Disease
Yes No
100 33

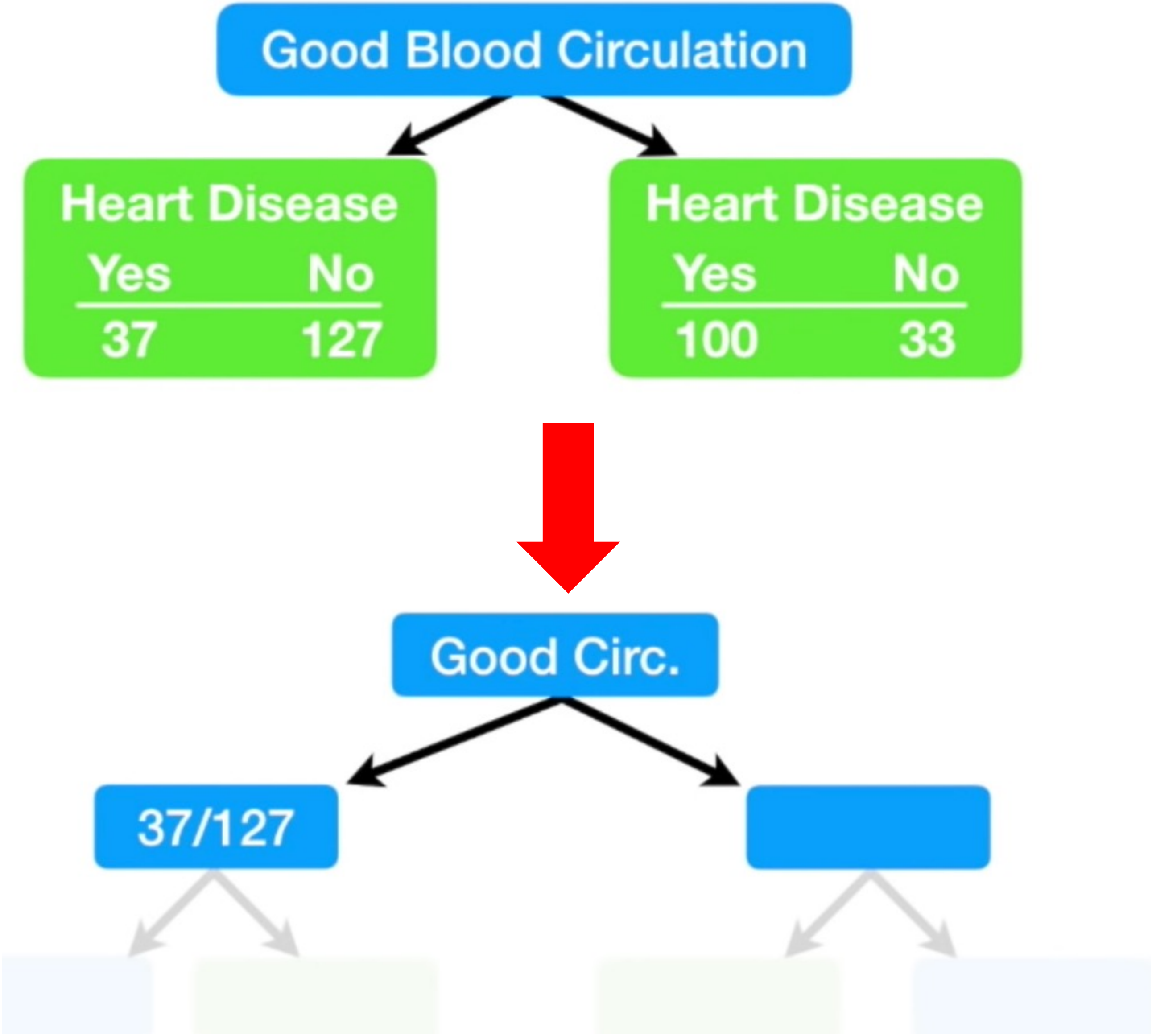
Gini impurity for Blocked Arteries = 0.381



Blocked Arteries

Heart Disease
Yes No
92 31

Heart Disease
Yes No
45 129



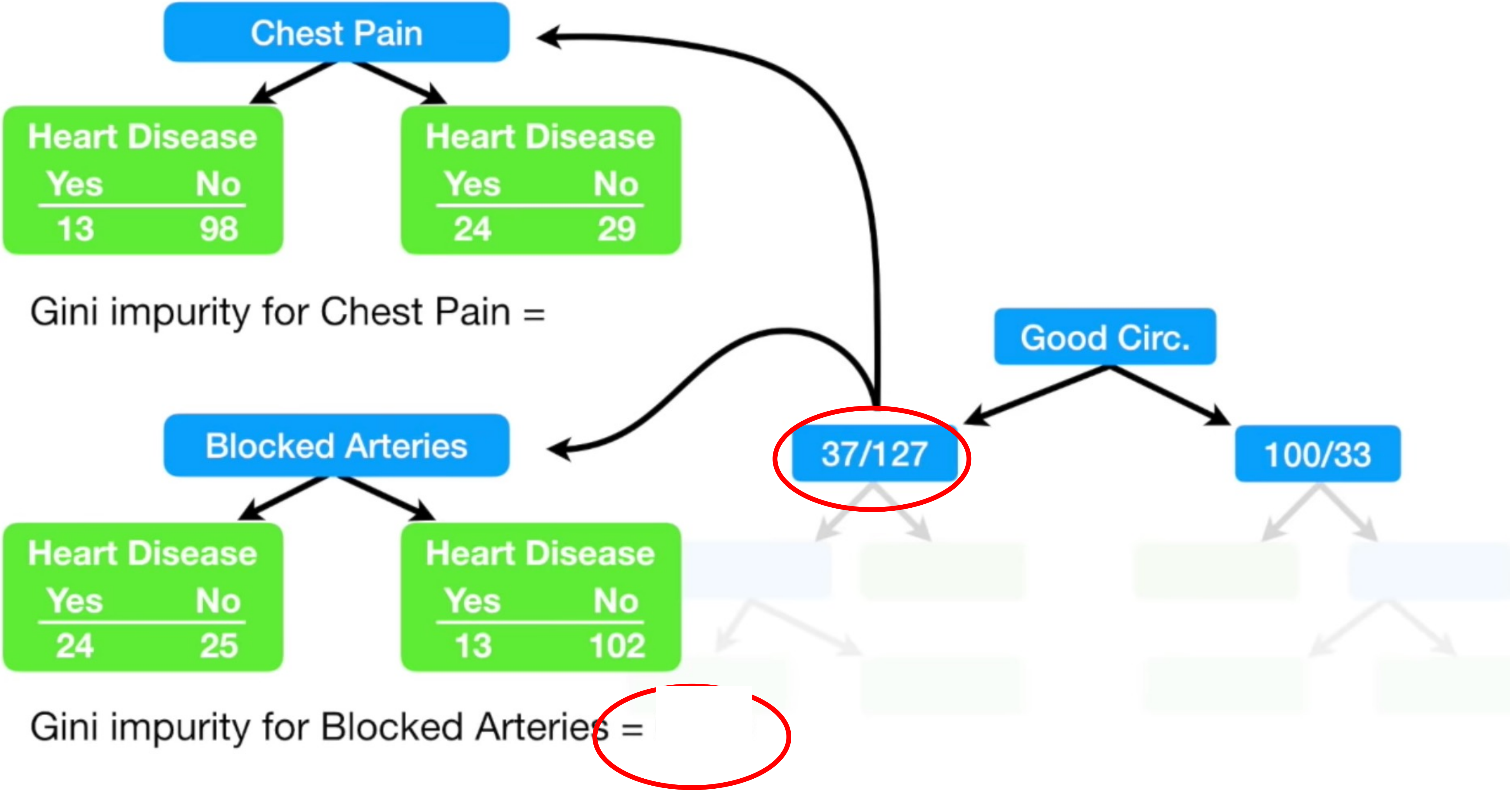
Good Blood Circulation

| Heart Disease | |
|---|---|
| Yes | No |
| 37 | 127 |

| Heart Disease | |
|---|---|
| Yes | No |
| 100 | 33 |

Good Circ.

37/127



Chest Pain

| Heart Disease | |
|---|---|
| Yes | No |
| 13 | 98 |

| Heart Disease | |
|---|---|
| Yes | No |
| 24 | 29 |

Gini impurity for Chest Pain =

Blocked Arteries

| Heart Disease | |
|---|---|
| Yes | No |
| 24 | 25 |

| Heart Disease | |
|---|---|
| Yes | No |
| 13 | 102 |

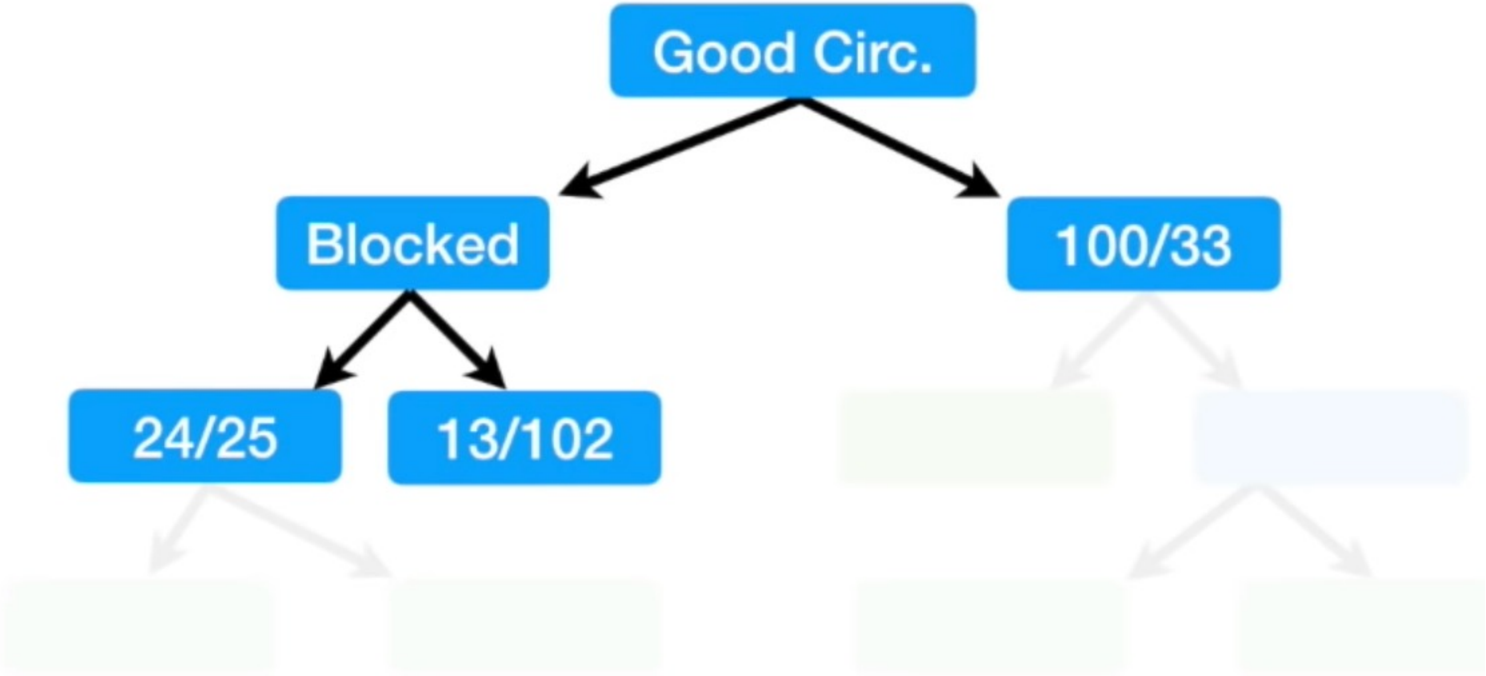
Gini impurity for Blocked Arteries =

Good Circ.

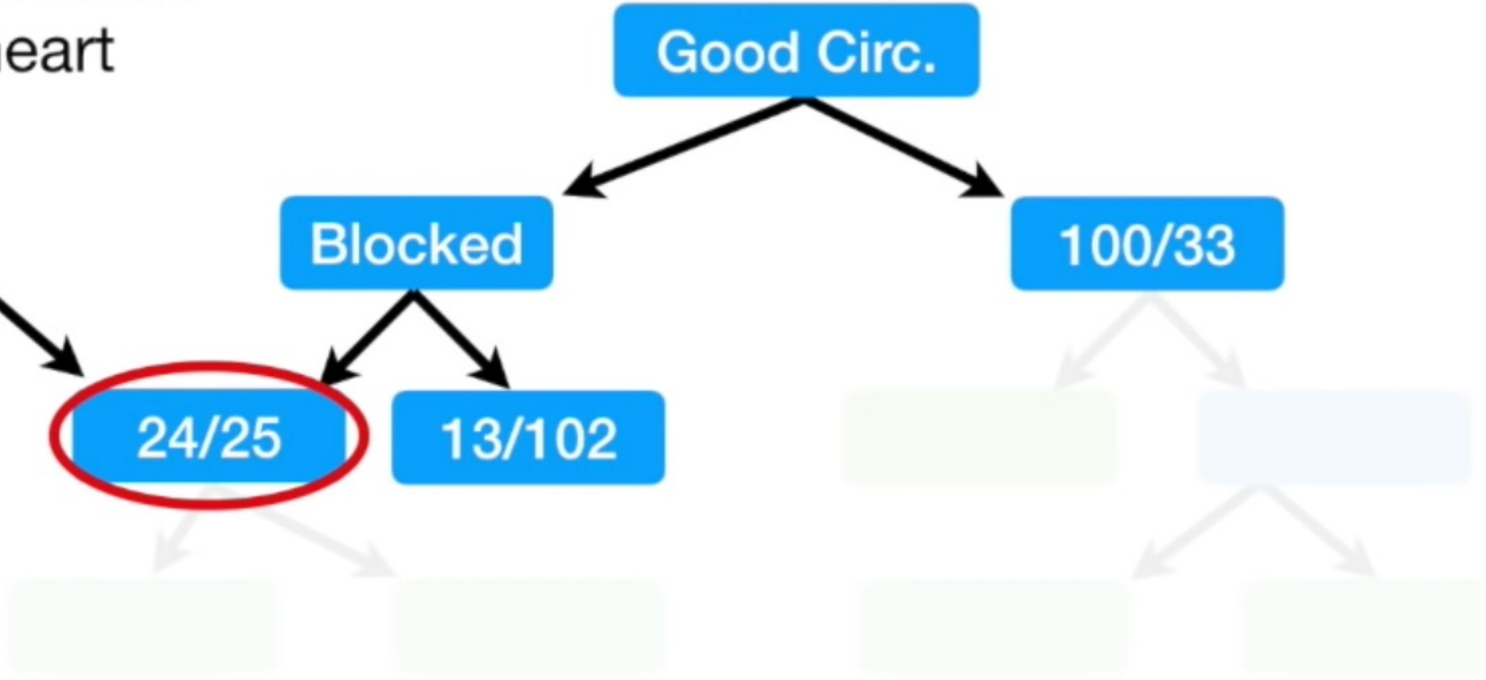
37/127

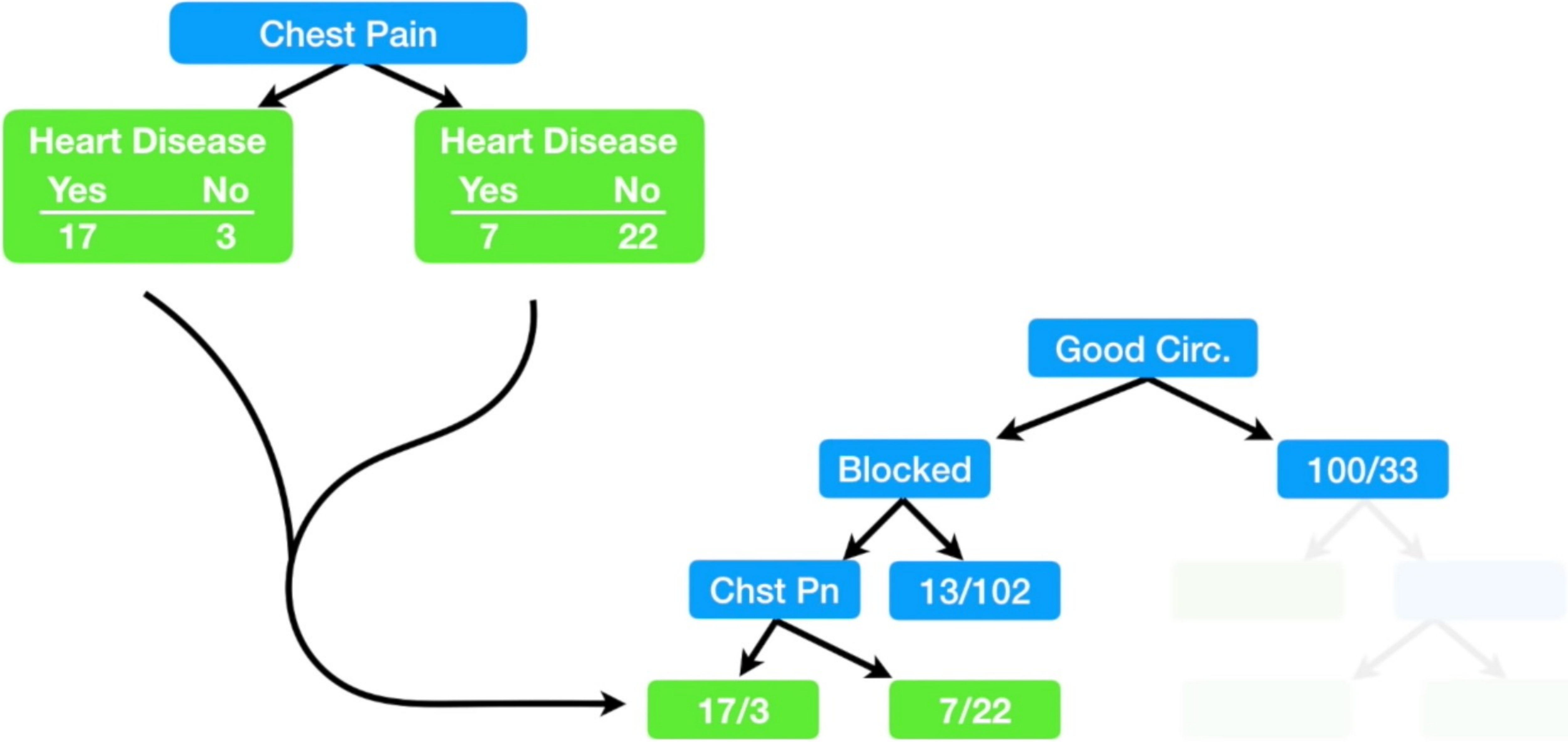
100/33

Here's the tree that we've worked out so far.

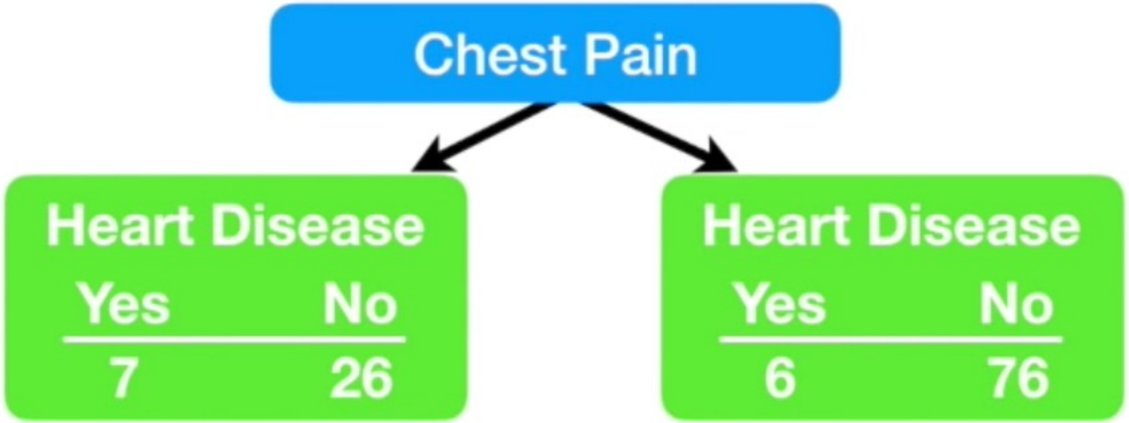


All we have left is Chest Pain, so first we'll see how well it separates these 49 patients (24 with heart disease and 25 without heart disease).





…so these are the final leaf nodes on this branch of the tree.



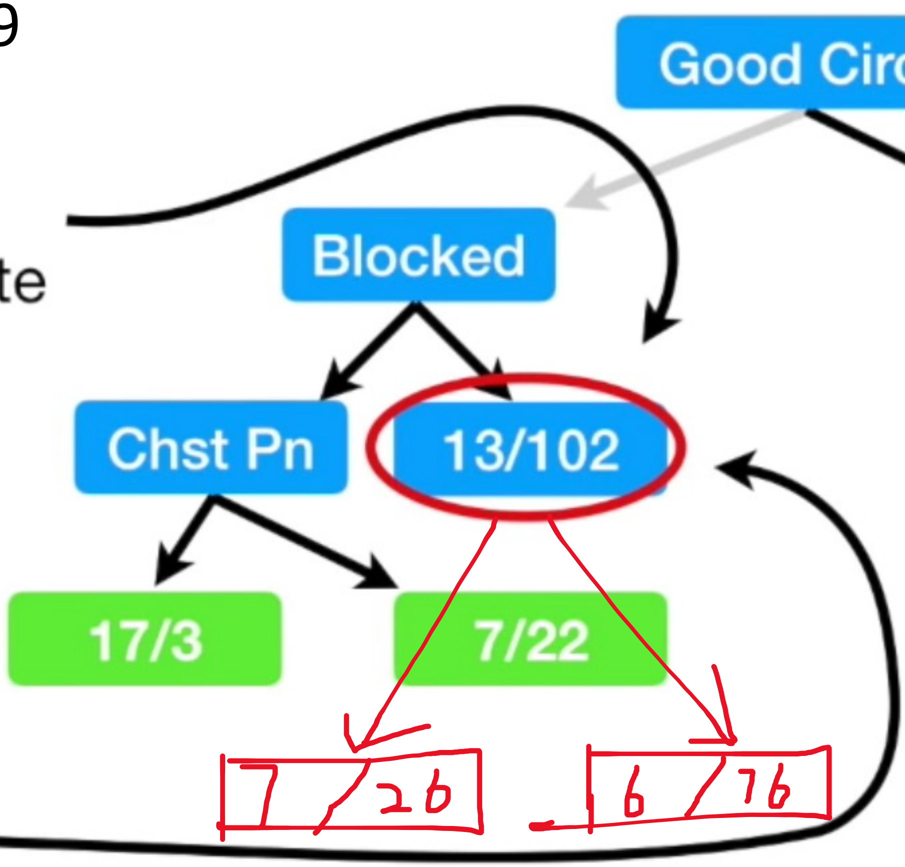
Gini impurity for Chest Pain = 0.19

The Gini impurity for this node, before using chest pain to separate patients is…

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

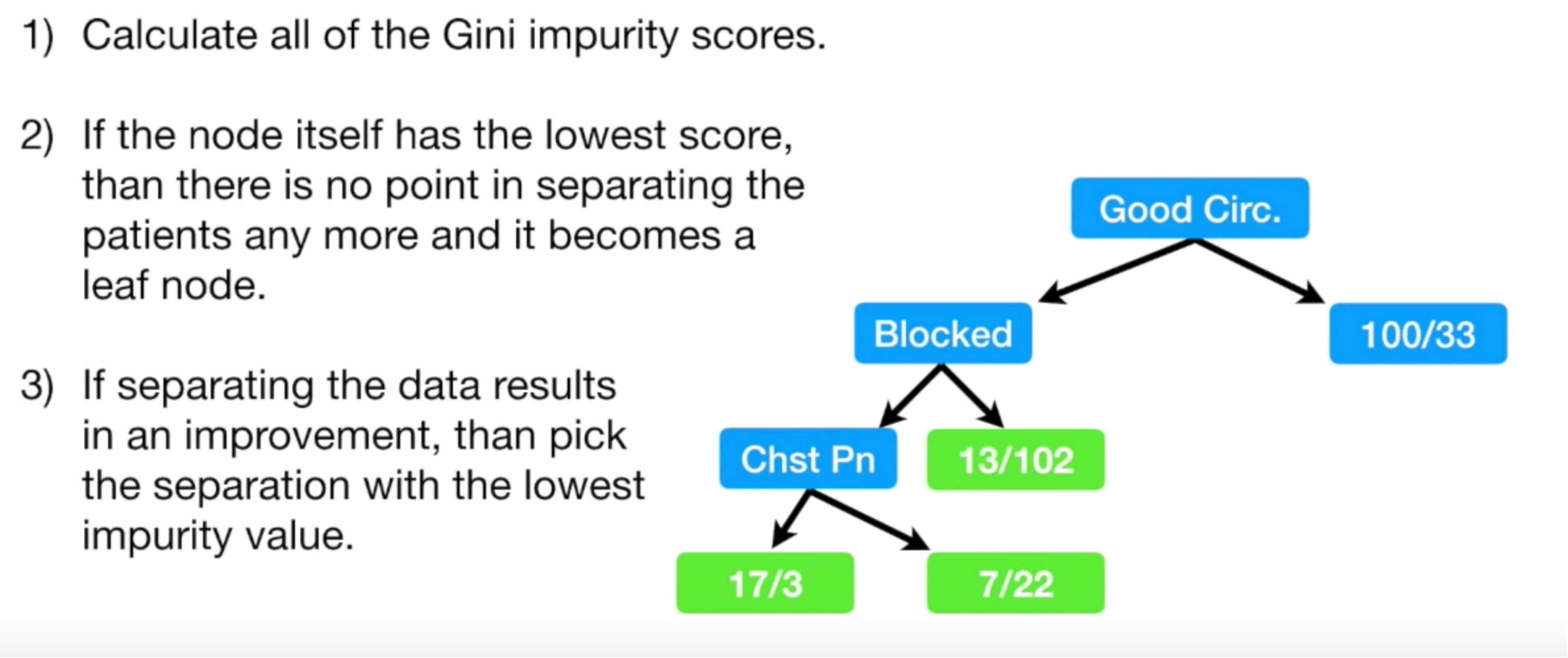
$$= 1 - \left(\frac{13}{13 + 102}\right)^2 - \left(\frac{102}{13 + 102}\right)^2$$

$$= 0.2$$



# Right side

- Repeat the steps for calculating the left sides.



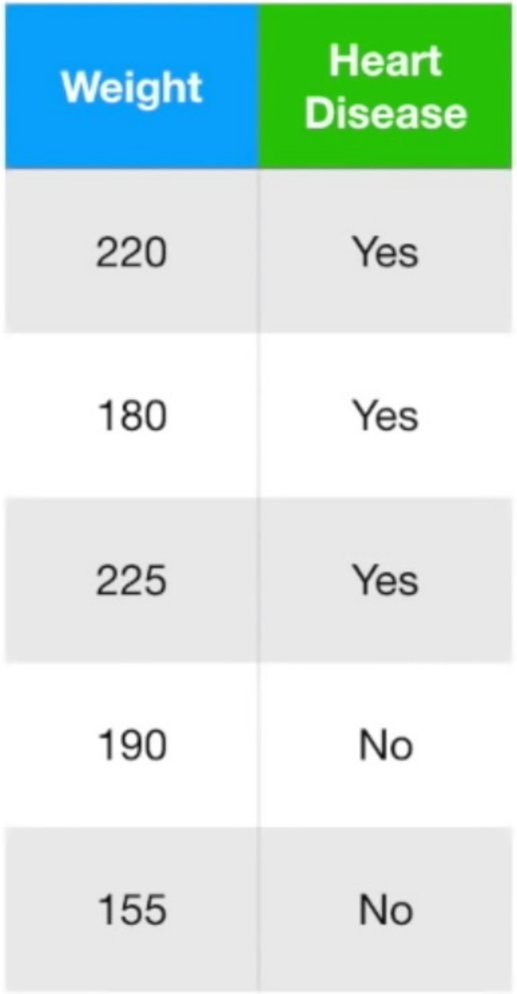
1) Calculate all of the Gini impurity scores.

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.

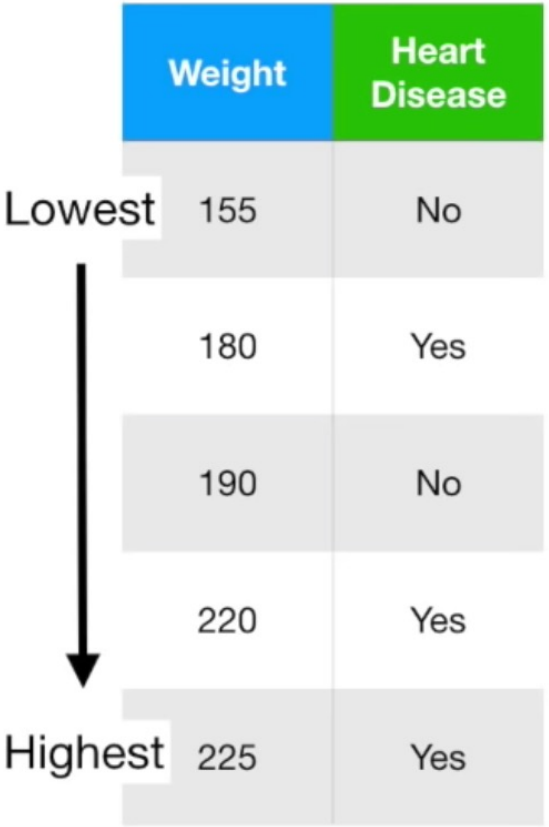
# Now we have seen YES/NO-based decision tree

- What about numerical-based DT?



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

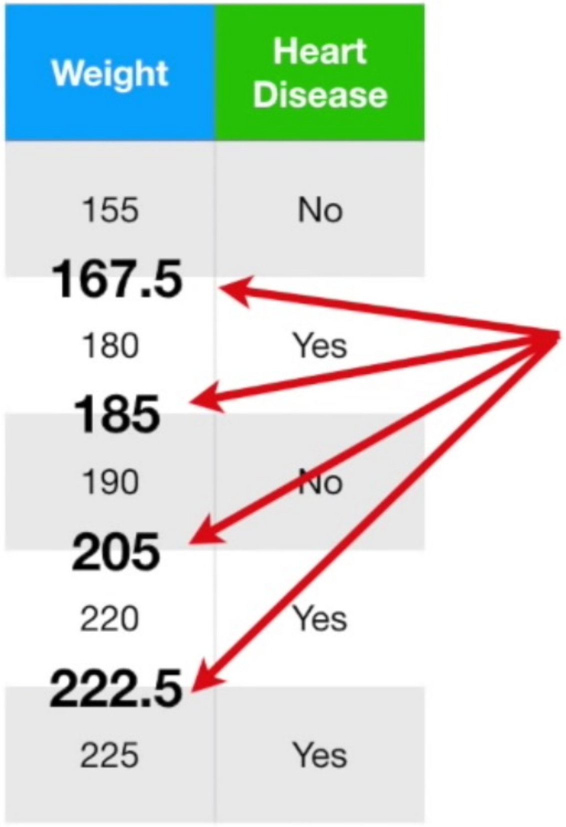
# Step 1



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

Lowest
Highest

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



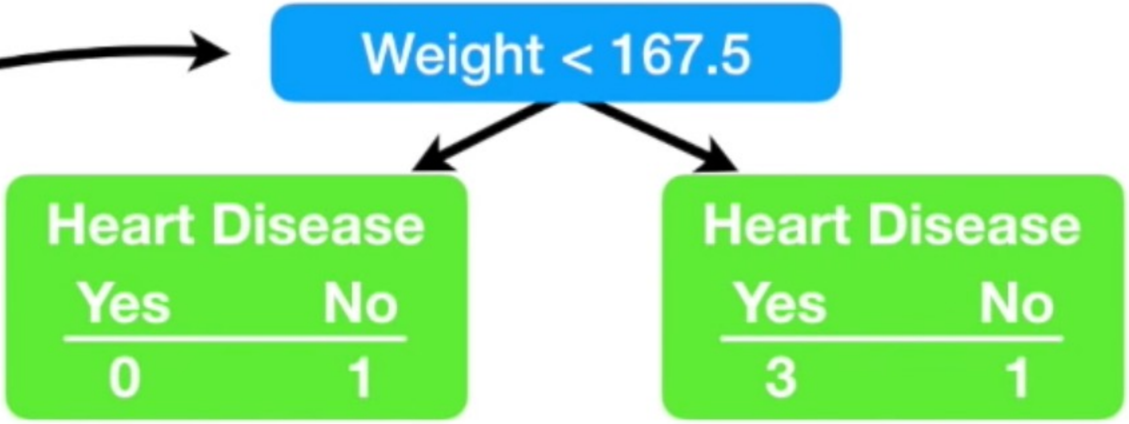
| Weight | Heart Disease |
|---|---|
| 155 | No |
| **167.5** | |
| 180 | Yes |
| **185** | |
| 190 | No |
| **205** | |
| 220 | Yes |
| **222.5** | |
| 225 | Yes |

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

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

| Weight | Heart Disease | |
|---|---|---|
| 155 | No | |
| **167.5** | | → Gini impurity = ? |
| 180 | Yes | |
| **185** | | → Gini impurity = ? |
| 190 | No | |
| **205** | | → Gini impurity = ? |
| 220 | Yes | |
| **222.5** | | → Gini impurity = ? |
| 225 | Yes | |

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



Weight < 167.5

| Heart Disease Yes | Heart Disease No |
|---|---|
| 0 | 1 |

| Heart Disease Yes | Heart Disease No |
|---|---|
| 3 | 1 |

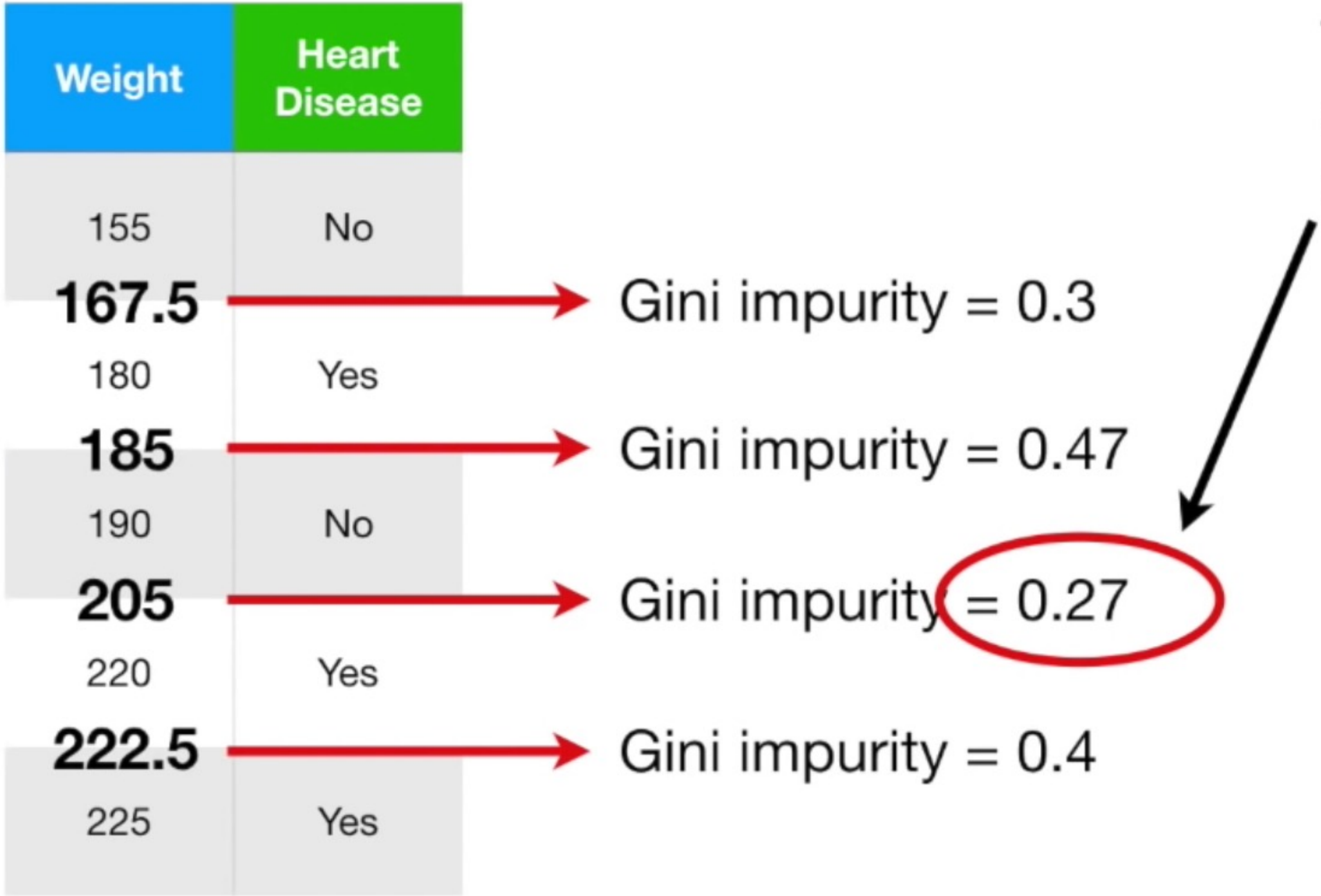
Gini impurity = 0

0.375

Gini impurity for Weight < 167.5 is the weighted average of the impurities for the two leaves.

$$= \qquad \frac{\quad}{0.375}$$

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



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

**167.5** → Gini impurity = 0.3

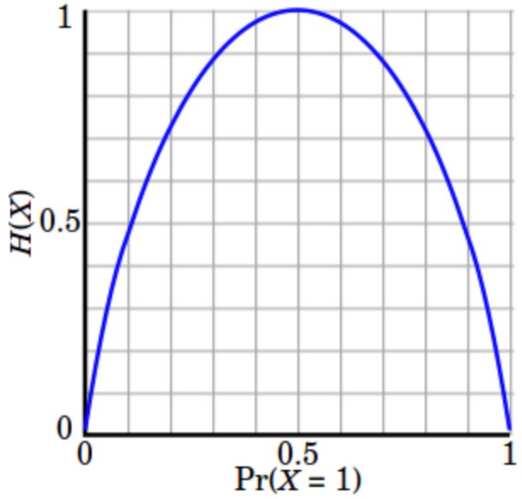
**185** → Gini impurity = 0.47

**205** → Gini impurity = 0.27

**222.5** → Gini impurity = 0.4

# Other options: Entropy

$$Entropy = -\sum p_j \log_2 p_j$$



1. Entropy of a group in which all examples belong to the same class:

$$entropy = -1 \log_2 1 = 0$$

This is not a good set for training.

2. entropy of a group with 50% in either class:

$$entropy = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

# Which one is used in practice?

- https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier
  - Look at **criterion** : *string, optional (default=”gini”)*
    - The function to measure the quality of a split. Supported criteria are “gini” for the Gini impurity and “entropy” for the information gain.