**DECISION TREE - GINI INDEX**

1. Case study

| Chest Pain | Weight | Heart Disease (Target) |
|---|---|---|
| Yes | 50 | 1 |
| No | 40 | 1 |
| Yes | 50 | 0 |
| Yes | 40 | 1 |
| No | 50 | 0 |

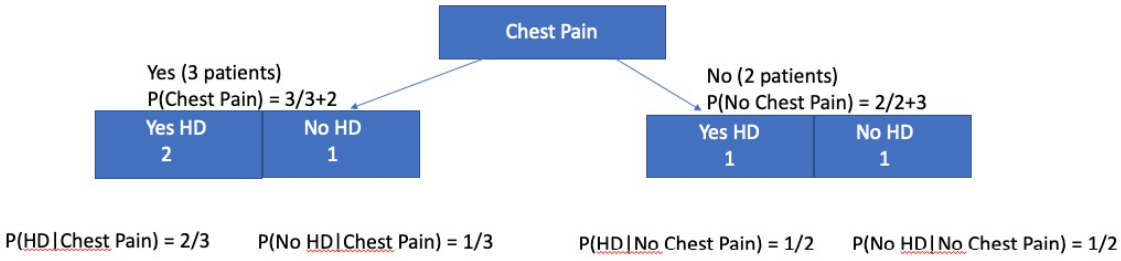
2. At each node, we will see how well these candidates (Chest Pain, Weight) separate patients with heart disease (HD) and patients with no heart disease.

Question: How to know which candidate we will choose to do root node. If so, what splitting value we will choose ?

3. Solution

Step 1: Split data at each node

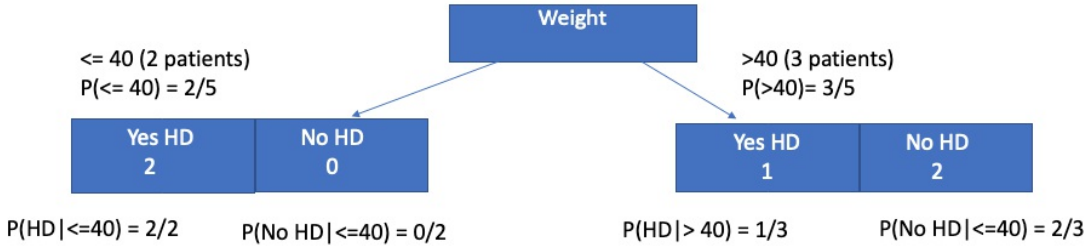
Because Chest Pain only has 2 levels (Yes/No), we split into 2 parts: Yes and No



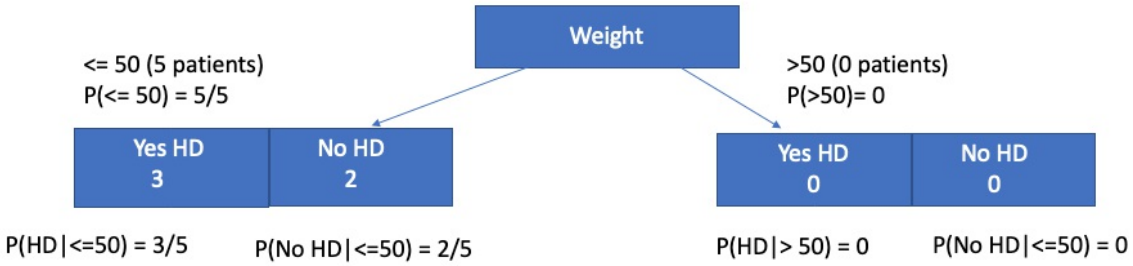

Weight

Because weight is numeric, we will use all of distinct values as splitting value. At each distinct value, we will compare <= 40; <= 50

Weight = 40




Weight = 50




Step 2: Evaluate splits
There are 3 populuar ways
- Use Logworth
- Use Gini impurity
- Use Entropy

We use these measurements to calculate for each node at different thresold, then compare among these nodes to select the best.

Example: Gini index for Chest Pain (2 levels: Yes & No)

$$\text{GINI} = 1 - \sum (P)^2 = 1 - (P(\text{Yes})^2 + P(\text{No})^2)$$

At each node, compute GINI leftside and Gini rightside, then compute final GINI index

Chest Pain

$$\text{Gini left} = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.44$$

$$\text{Gini right} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

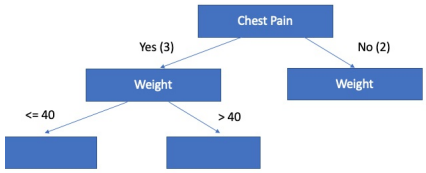
$$\text{Gini final} = P(\text{left side})\ \text{Gini Left} - P(\text{right side})\ \text{Gini right}$$

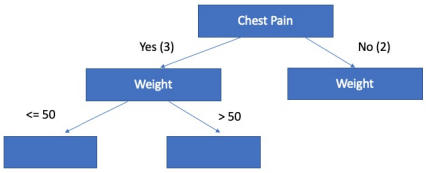
$$= \left(\frac{3}{5}\right) 0.44 + \left(\frac{2}{5}\right) 0.5 = 0.47$$

Similarly, we can compute Gini final at splitting value 40 and 50

Step 3: Compare Gini final among candidates to find the LOWEST final Gini. Example, Gini final of Chest Pain is the lowest, mean that although Chest Pain does not create pure leaves, it creates the least impurity.

Step 4: Let's say Chest Pain has the lowest Gini final, then it is root node.
Now, we have to decide what splitting value (40 or 50) we should put on the left side of Chest Pain.





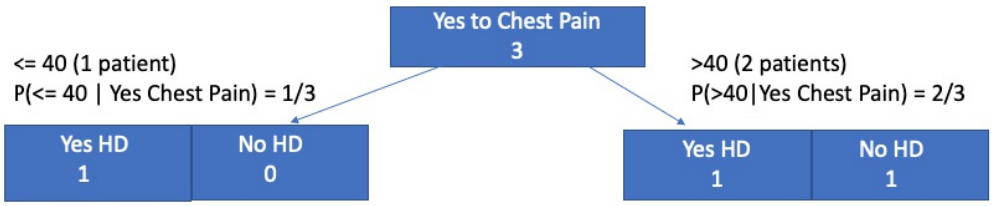

- Compute Gini final at splitting value = 40, given Yes to Chest pain
- Compute Gini final at spliting value = 50, given Yes to Chest pain
- Let's say we have extra candidate: Obese (with 2 levels), we also will compute Gini final of Obese as we did with Chest Pain

=> Choose splitting value with the lowest Gini final to do child node of left side of Chest Pain

Similarly, compute Gini final at splitting value = 40; 50 and Obese, given No to Chest pain

=> Choose splitting value with the lowest Gini final to do child node of right side of Chest Pain

Example: Compute Gini final at splitting value = 40 on the left side of Chest Pain




P(HD|<=40 & Yes Chest Pain) = 1/1 = 1

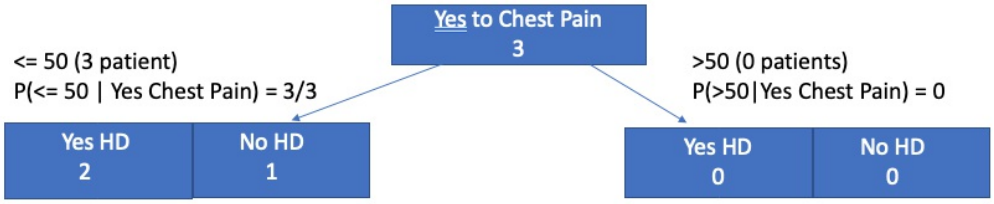
P(HD|>40 & No Chest Pain) = 1/2

$$\text{Gini left} = 1 - (1)^2 - 0 = 1$$

$$\text{Gini right} = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

$$\text{Gini final} = \left(\frac{1}{3}\right) 1 + \left(\frac{2}{3}\right) 0.5 = 0.67$$

Example: Compute Gini final at splitting value = 50 on the left side of Chest Pain

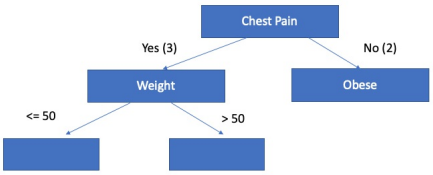



P(HD|<=50 & Yes Chest Pain) = 2/3

P(No HD|<=50 & Yes Chest Pain) = 1/3

$$\text{Gini final} = 1\left(1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right) + 0 = 0.44$$

Let's say Gini final of splitting value = 50 is the lowest, given Yes to Chest Pain, then the tree improves like this




Step 5: We keep splitting data into smaller until
- Reach min number of samples in a node
- Gini impurity does not improve
- Reach maximum depth allowed for tree