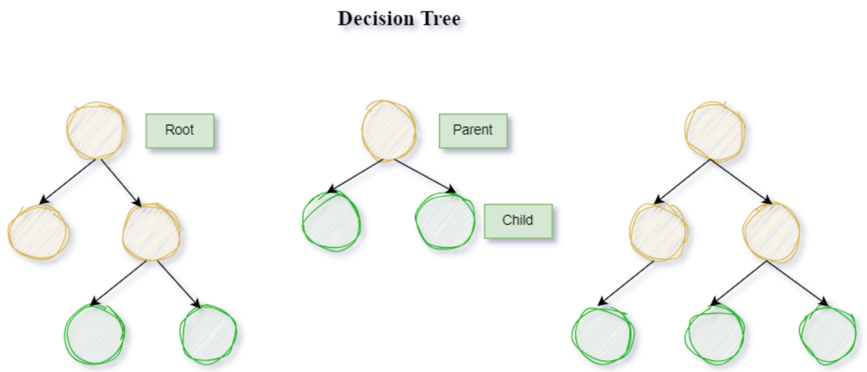
Decision Tree

$$Decision\ Tree$$

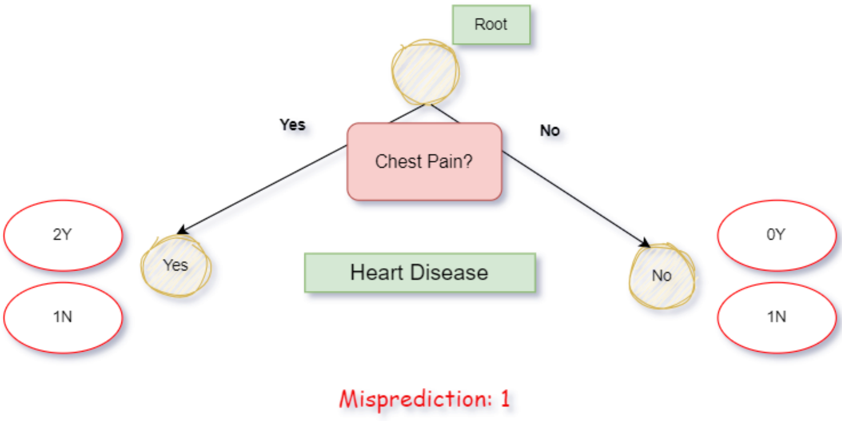


Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	Yes	Yes

- Based on the table on the left, we will decide whether a patient has heart disease or not.

Steps

- First, we will create the decision tree
- Here, for the first table, we have 3 features. **F1, F2, F3**



- We consider the feature **"chest pain"** and we have 1 misclassification.
- We need to select the feature where we have the least number of misclassifications.
- We have equations to find this.

Impurity

- We calculate the impurity for each feature by creating a tree like the above and selecting the feature that has the **lowest impurity**.

Different Impurities

- Gini

$$1 - \sum P(i)^2$$

- Entropy

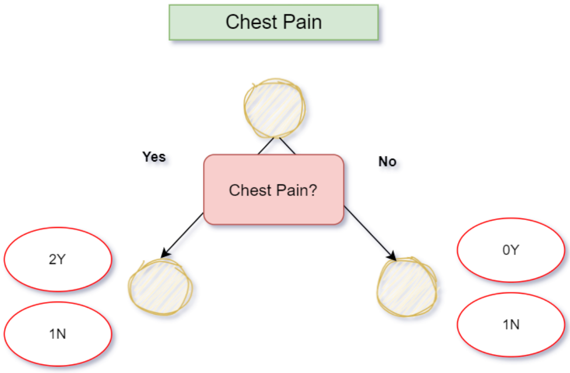
$$-\sum P_i \, log_2(p_i)$$

For example, Impurities for the features are as follows:

F1	F2	F3
.2	.8	.4

- As F1 has the minimum impurity, we will start the tree based on F1, then we calculate impurities again, find the minimum, expand the tree for the later features and so on.

Calculating Impurity using Gini



$$Gini = 1 - \sum P(i)2$$

Left Node

$$P_{Yes} = \frac{2}{3} \quad P_{No} = \frac{1}{3}$$
$$Gini = 1 - (0.66)^2 - (0.33)^2$$
$$= 0.455$$

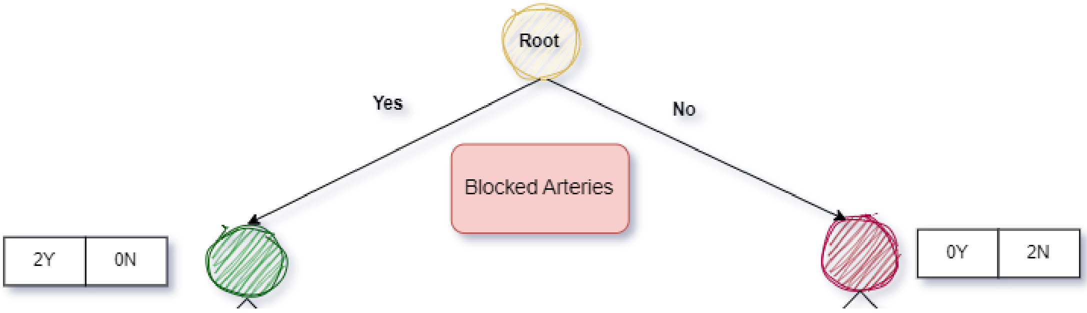
Right Node

$$P_{Yes} = \frac{0}{1} \quad P_{No} = \frac{1}{1}$$
$$Gini = 1 - 0^2 - 1^2$$
$$= 0$$

- Now we will find the weighted sum of the left and right gini

$$\begin{aligned} Gini &= W_L * Gini(left) + W_R * Gini(right) \\ &= \frac{3}{4} * 0.455 + \frac{1}{4} * 0 \\ &= 0.34125 \end{aligned}$$

Calculating Gini for feature “Blocked Arteries”



Left Node

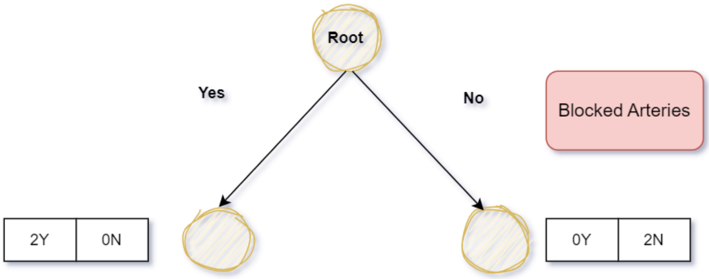
$$P_{Yes} = \frac{2}{2} \quad P_{No} = \frac{0}{2}$$
$$\begin{aligned} Gini &= 1 - (1)^2 - (0)^2 \\ &= 0 \end{aligned}$$

Right Node

$$P_{Yes} = \frac{0}{2} \quad P_{No} = \frac{2}{2}$$
$$\begin{aligned} Gini &= 1 - 0^2 - 1^2 \\ &= 0 \end{aligned}$$

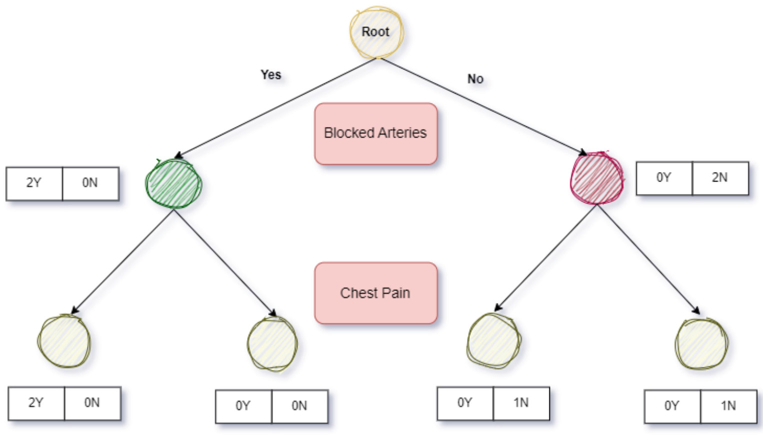
$$\begin{aligned} Gini &= W_L * Gini(left) + W_R * Gini(right) \\ &= \frac{2}{4} * 0 + \frac{2}{4} * 0 \\ &= 0 \end{aligned}$$

- The same way we will do for the **Good Blood Circulation** feature.
- As the the **Blocked Arteries** feature has the minimum Gini, we will split starting it.



- Now, for further split, we will use a concept called **information gain.**

- We will check, **based on blocked arteries values, what are the values of chest pain(we are taking chest pain as the next level of the tree).**



- Now we will traverse for a decision.
- For test data, we will use this tree to reach to the leaf nodes, where the **prediction stays.**
- In this data, the Gini values are nice, i.e. ideal situation. Not every time this can be the case. Then there might be misclassifications.

Information Gain

$$IG = Gini(Parent) - Gini(Child)$$

If IG is very close to 0, we don't have to split. If it's much greater than 0. then we split.

$$Decision\,Tree\ for\ numerical\ features$$

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

- Sort the values

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

- Then find the average of each two corresponding rows.

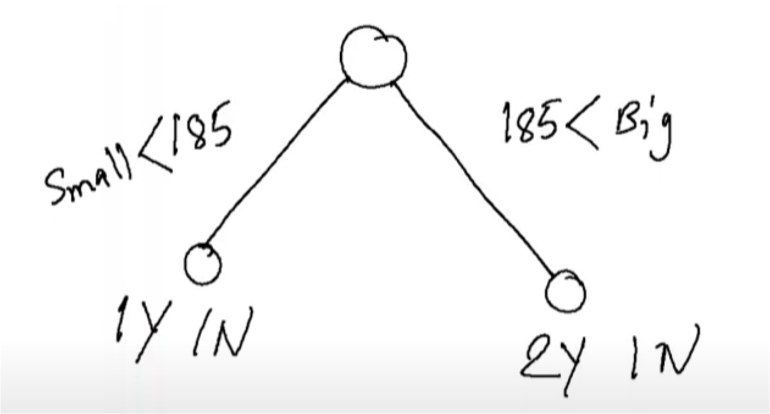
⇒ 167.5

⇒ 185

⇒ 205

⇒ 222.5

- Now, based on these avg values, we will build the decision tree.



- Then the same steps as previously discussed.