

Arif Romadhan

arifromadhan19@gmail.com

Classification in Decision Tree — A Step by Step CART (Classification And Regression Tree)

1. Introduction

CART (Classification And Regression Tree) is a decision tree algorithm variation, in the previous article — [The Basics of Decision Trees](#). Decision Trees is the non-parametric supervised learning approach. CART can be applied to both regression and classification problems.

As we know, data scientists often use decision trees to solve regression and classification problems and most of them use scikit-learn in decision tree implementation. [Based on documentation, scikit-learn uses an optimised version of the CART algorithm](#)

2. How Does CART Work in Classification?

[in the previous article](#) it was explained that CART uses Gini Impurity in the process of splitting the dataset into a decision tree.

Mathematically, we can write Gini Impurity as following

$$I_{Gini} = 1 - \sum_{i=1}^j p_i^2$$

$$I_{Gini} = 1 - (\text{the probability of target "No"})^2 - (\text{the probability of target "Yes"})^2$$

How does CART process the splitting of the dataset

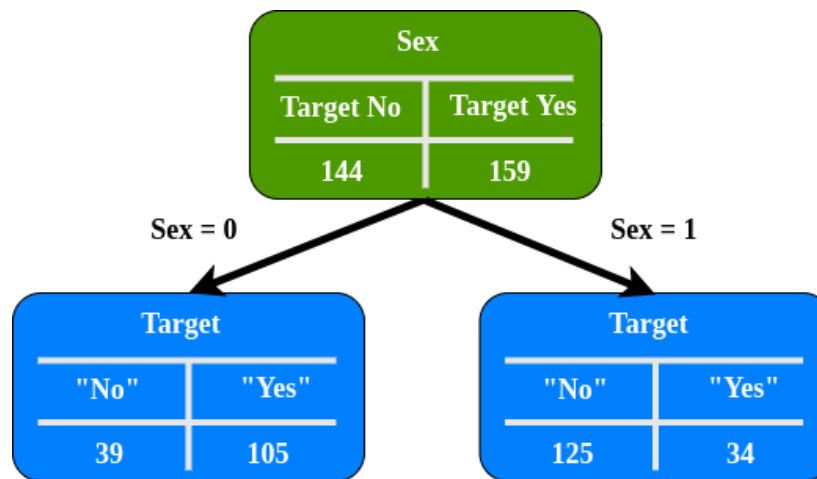
This simulation uses a Heart Disease Data set with 303 rows and has 13 attributes. Target consist 138 value 0 and 165 value 1

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	Yes
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	Yes
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	Yes
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	Yes
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	Yes

In this simulation, only use the Sex, Fbs (fasting blood sugar), Exang (exercise induced angina), and target attributes.

3.1 Classification

measure Gini Impurity in Sex



Gini Impurity - Left Node

$$I_{Left} = 1 - \left(\frac{105}{105+39}\right)^2 - \left(\frac{39}{105+39}\right)^2 = 0.395$$

Gini Impurity - Right Node

$$I_{Right} = 1 - \left(\frac{34}{34+125}\right)^2 - \left(\frac{125}{34+125}\right)^2 = 0.336$$

Total Gini Impurity - Leaf Node

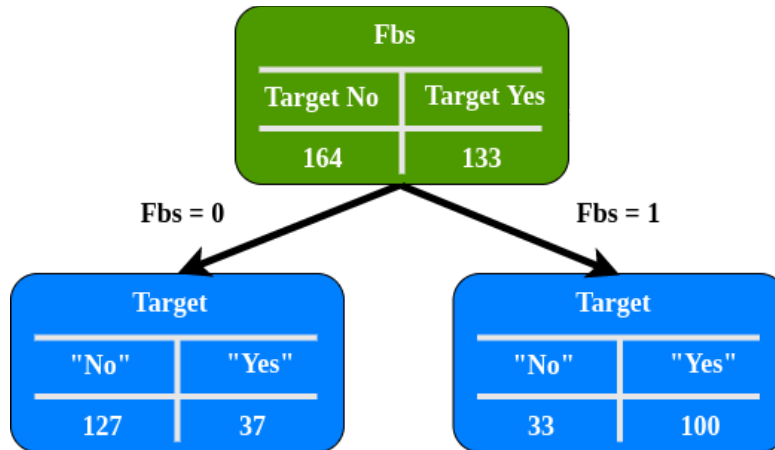
$I_{Sex} = \text{weight average of the leaf node impurities}$

$$I_{Sex} = \left(\frac{144}{144+159}\right) I_{Left} + \left(\frac{159}{144+159}\right) I_{Right}$$

$$I_{Sex} = \left(\frac{144}{144+159}\right) 0.29 + \left(\frac{159}{144+159}\right) 0.49$$

$$I_{Sex} = 0.364$$

measure Gini Impurity in Fbs



Gini Impurity - Left Node

$$I_{Left} = 1 - \left(\frac{127}{127+37}\right)^2 - \left(\frac{37}{127+37}\right)^2 = 0.349$$

Gini Impurity - Right Node

$$I_{Right} = 1 - \left(\frac{33}{100+33}\right)^2 - \left(\frac{100}{100+33}\right)^2 = 0.373$$

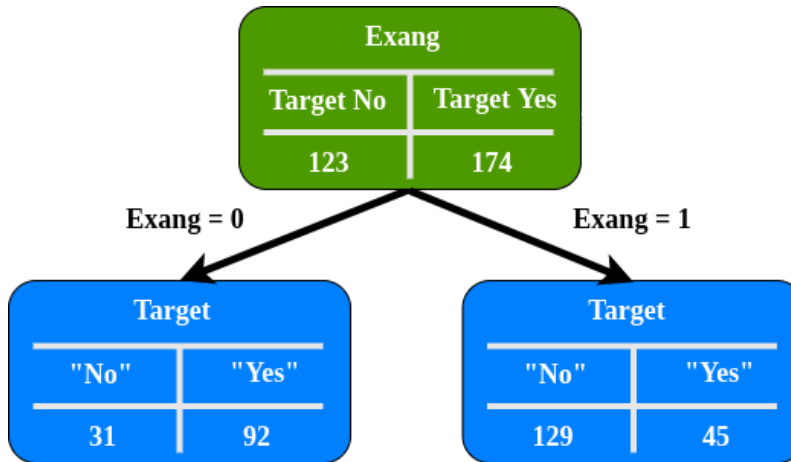
Total Gini Impurity - Leaf Node

$$I_{Fbs} = \left(\frac{164}{164+133}\right) I_{Left} + \left(\frac{133}{164+133}\right) I_{Right}$$

$$I_{Fbs} = \left(\frac{164}{164+133}\right) 0.349 + \left(\frac{133}{164+133}\right) 0.373$$

$$I_{Fbs} = 0.360$$

measure Gini Impurity in Exang



Gini Impurity - Left Node

$$I_{Left} = 1 - \left(\frac{31}{31+92}\right)^2 - \left(\frac{92}{31+92}\right)^2 = 0.377$$

Gini Impurity - Right Node

$$I_{Right} = 1 - \left(\frac{129}{129+45}\right)^2 - \left(\frac{45}{129+45}\right)^2 = 0.383$$

Total Gini Impurity - Leaf Node

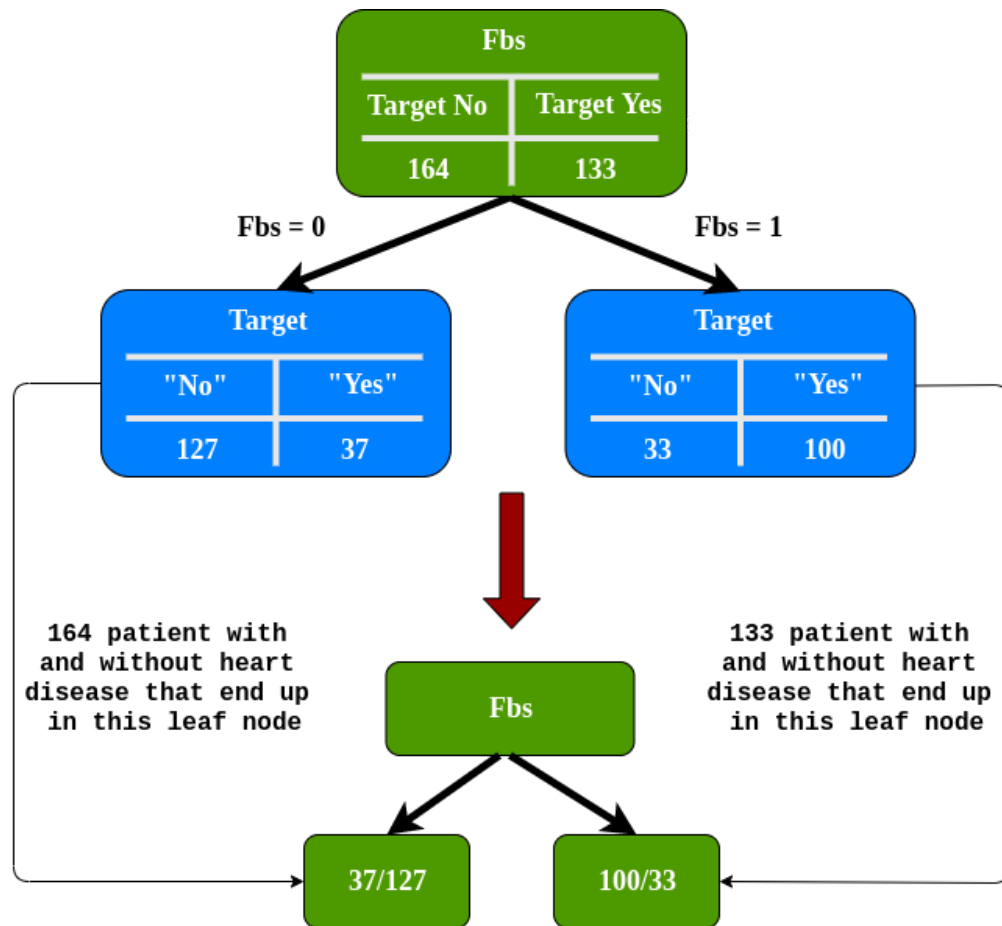
$$I_{Exang} = \left(\frac{123}{123+174}\right) I_{Left} + \left(\frac{174}{123+174}\right) I_{Right}$$

$$I_{Exang} = \left(\frac{123}{123+174}\right) 0.377 + \left(\frac{174}{123+174}\right) 0.383$$

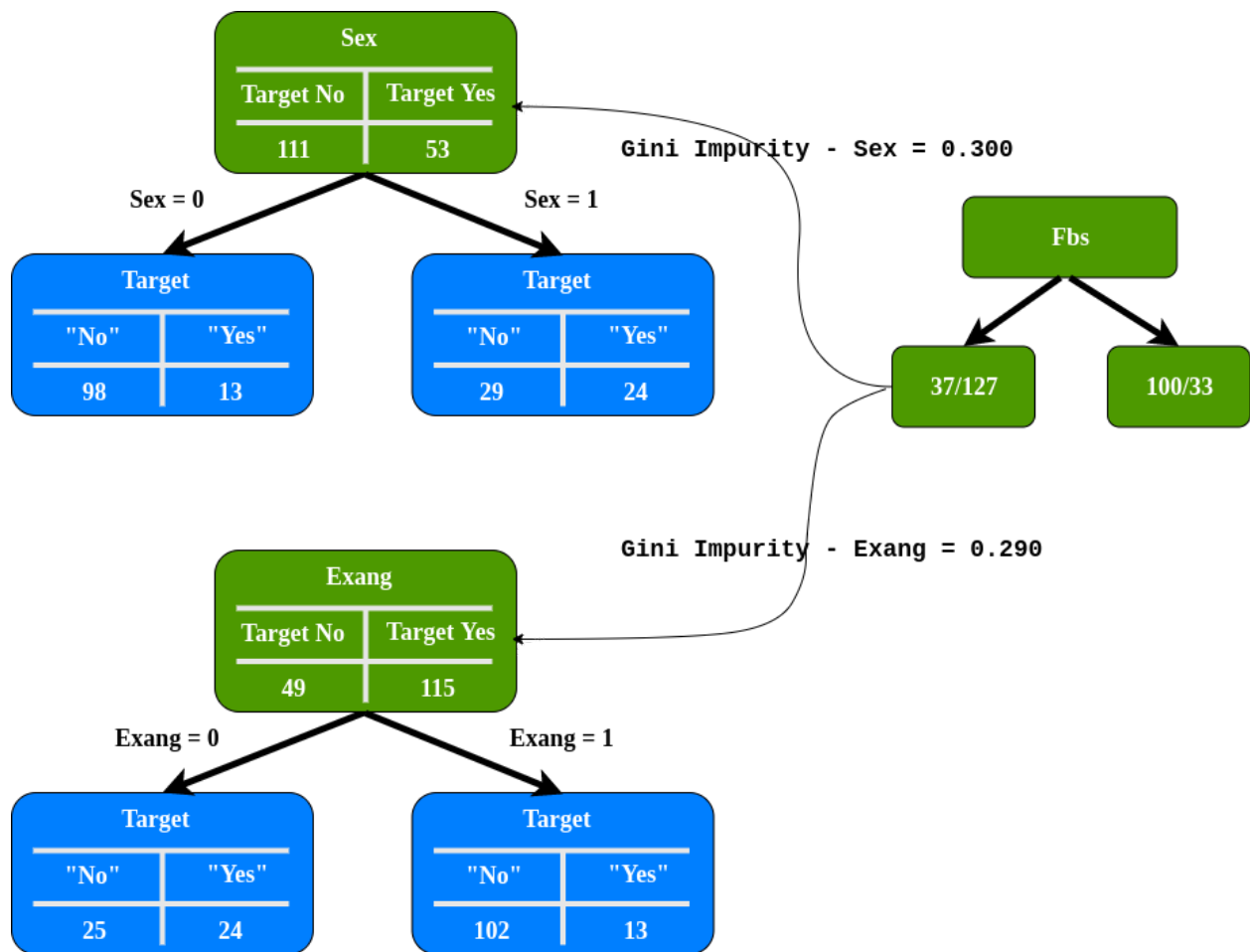
$$I_{Exang} = 0.381$$

Fbs (fasting blood sugar) has the lowest Gini Impurity, so well use it at the Root Node

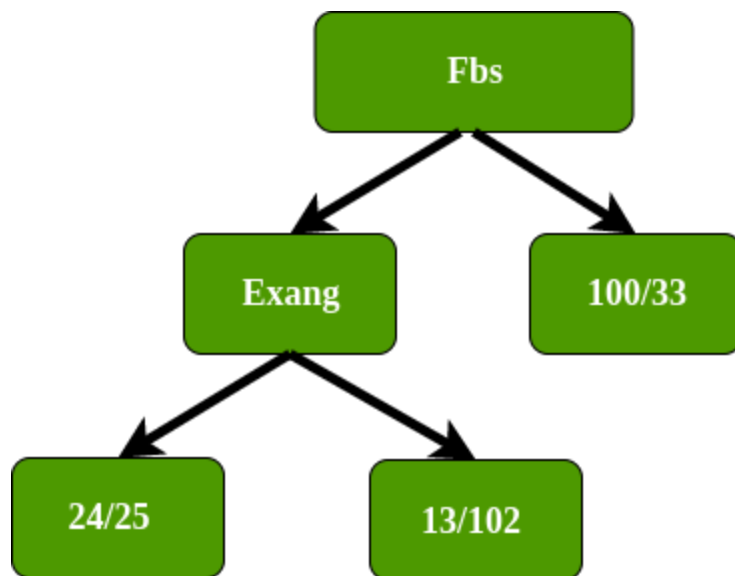
As we know, we have Fbs as Root Node, when we divide all of the patients using Fbs (fasting blood sugar), we end up with "Impure" leaf nodes. Each leaf contained with and without Heart Disease.



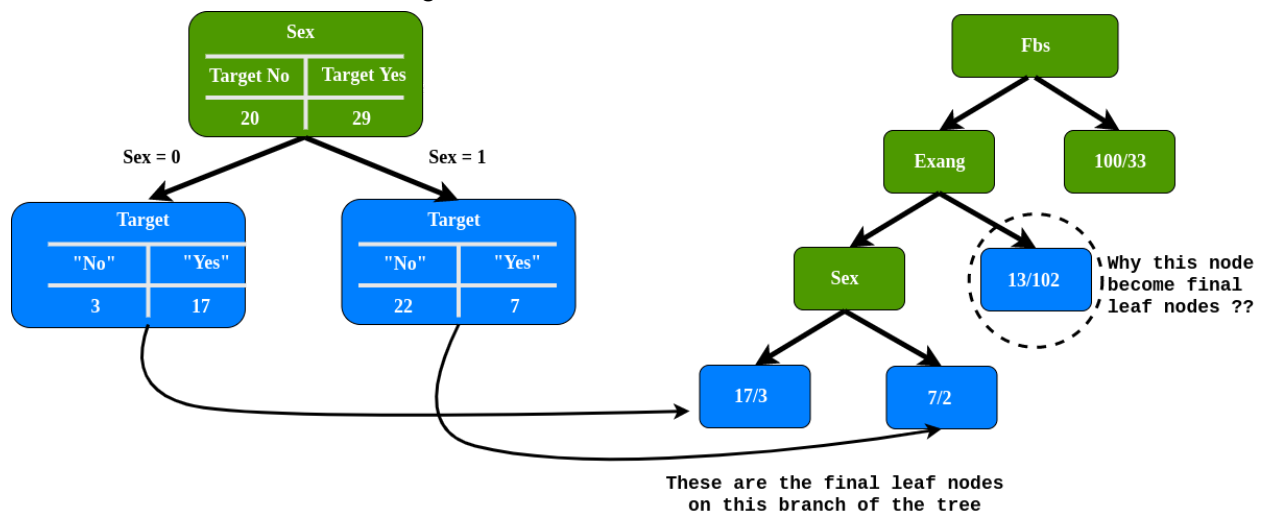
we need to figure how well sex and exang separate these patient in left node of Fbs



Exang (exercise induced angina) has the lowest Gini Impurity, we will use it at this node to separate patients.



in the left node of Exang (exercise induced angina), how well it separates these 49 patients (24 with heart diseases and 25 without heart disease). Since only the attribute sex is left, we put sex attribute in the left node of Exang

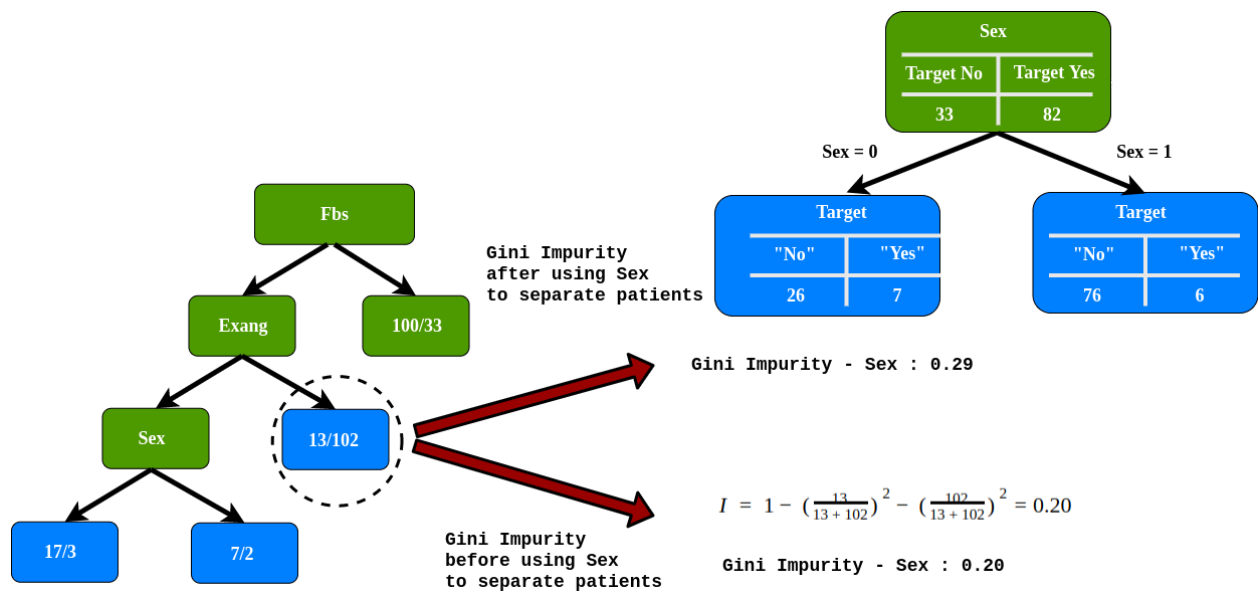


As we can see, we have final leaf nodes on this branch, but why is the leaf node circled including the final node?

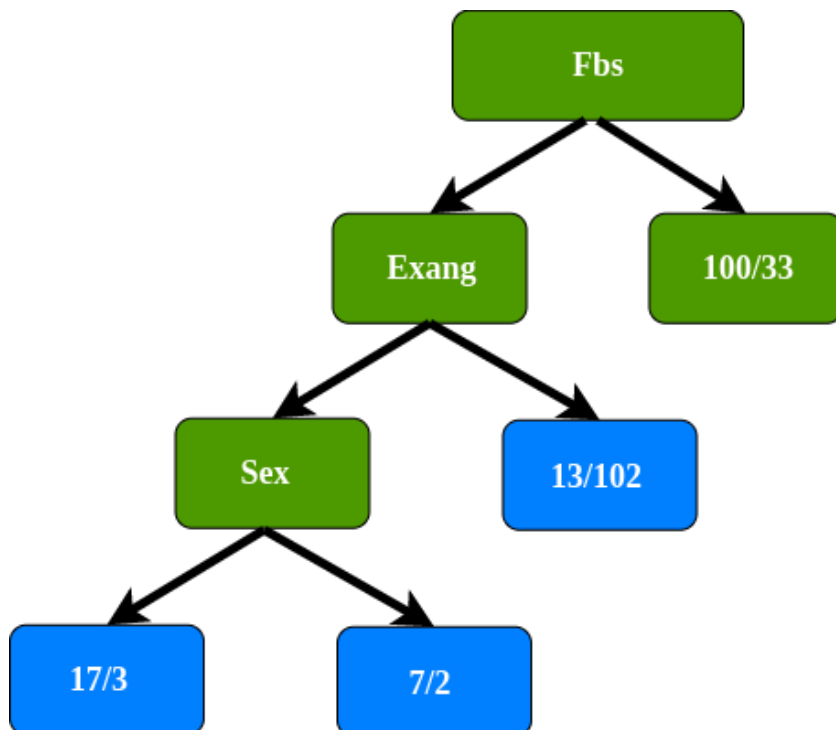
Note : the leaf node circled, 89% don't have heart diseases

Do these new leaves separate patients better than what we had before ?

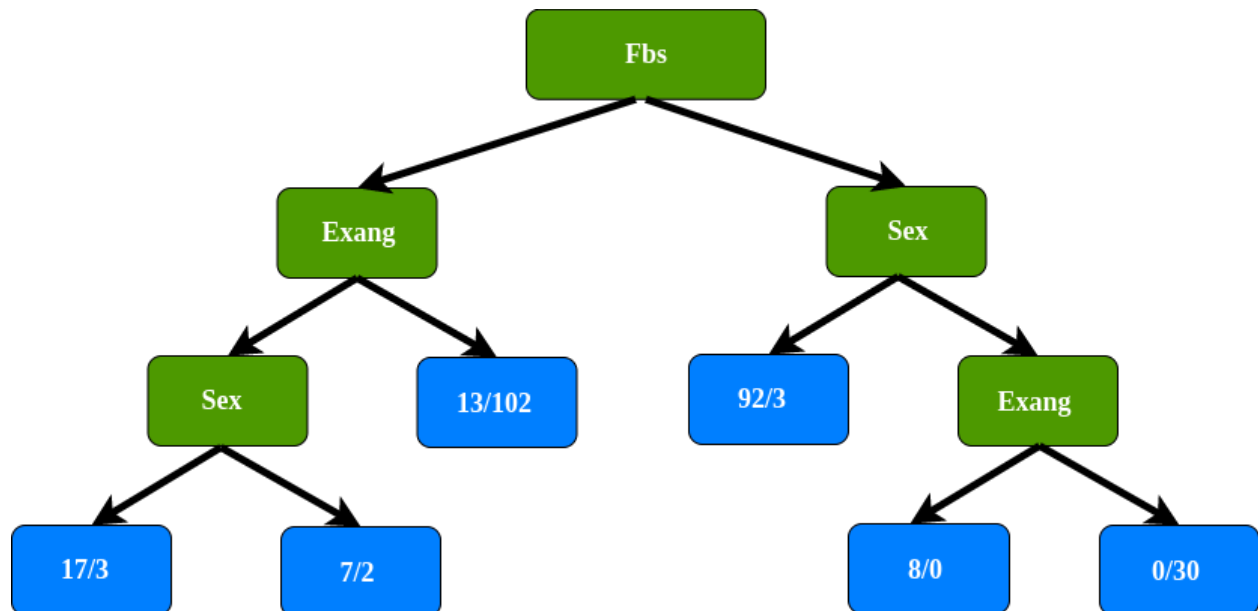
In order to answer those question, we must compare Gini Impurity using attribute sex and Gini Impurity before using attribute sex to separate patients.



The Gini Impurity before using sex to separate patients is lowest, so we don't separate this node using Sex. The final leaf node on this branch of tree :



Do the same thing on the right branch, so the end result of a tree in this case is



Main point when process the splitting of the dataset


1. calculate all of the Gini impurity score
2. compare the Gini impurity score, after n before using new attribute to separate data. If the node itself has the lowest score, than there is no point in separating the data
3. If separating the data result in an improvement, than pick the separation with the lowest impurity score

How to calculate Gini Impurity in continuous data?

such as weight which is one of the attributes to determine heart disease

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

Step 1 : Order data by ascending

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

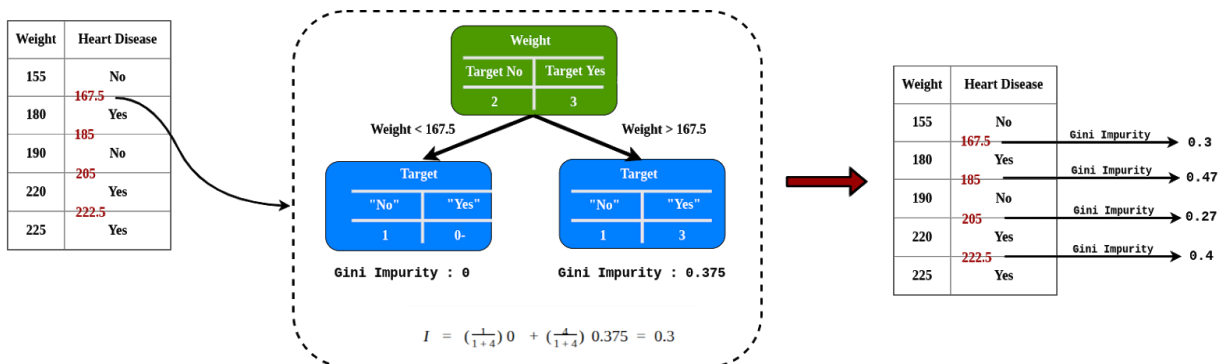
Step 2 : Calculate the average weight

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

167.5
185
205
222.5

Calculate the average weight

Step 3 : Calculate Gini Impurity values for each average weight



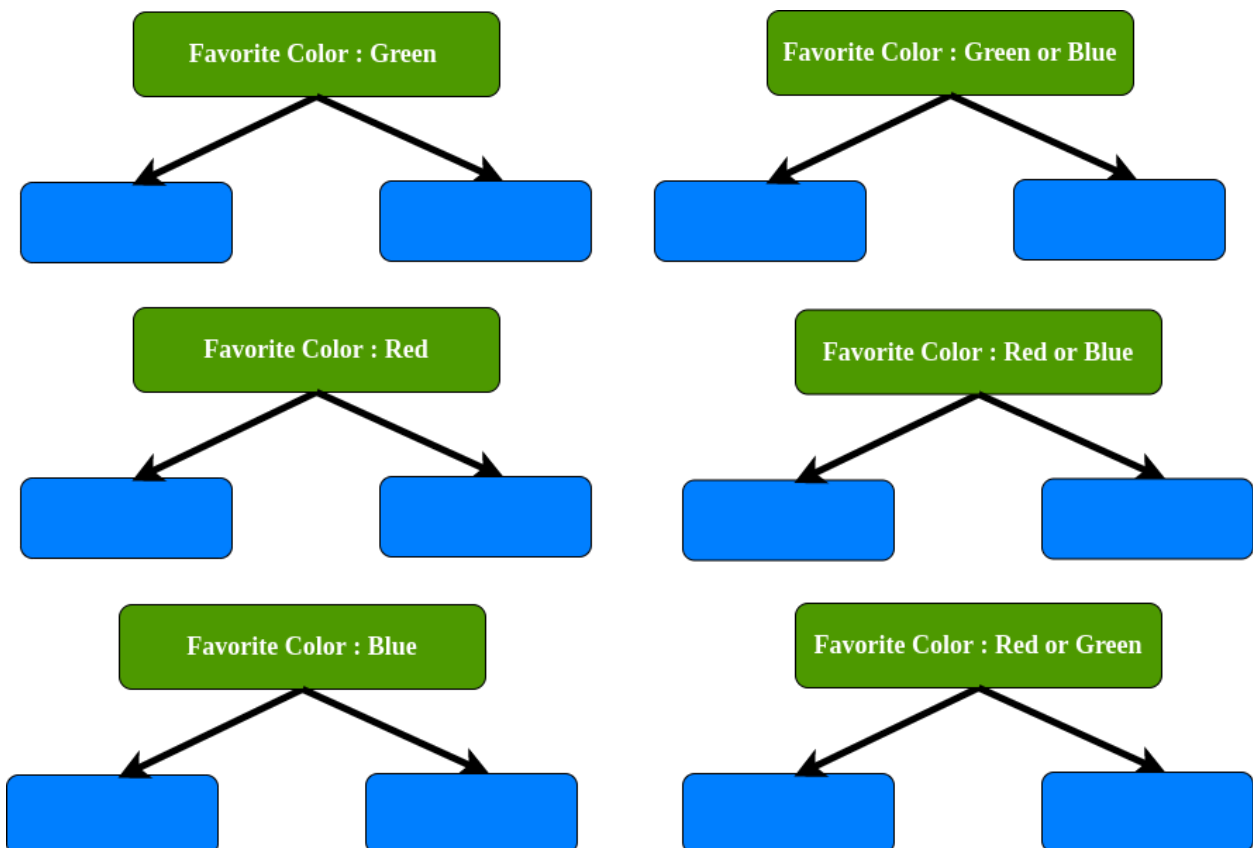
The lowest Gini Impurity is **Weight < 205**, this is the cutoff and impurity value if used when we compare with another attribute

How to calculate Gini Impurity in categorical data?

we have a favorite color attribute to determine a person's gender

Favorite Color	Sex
Green	Male
Red	Female
Blue	Male
Green	Male
etc	etc

In order to know Gini Impurity this attribute, calculate an impurity score for each one as well as each possible combination



Now, we have possible combination and we find out the lowest Gini Impurity to determine cutoff and impurity value

About Me

I'm a Data Scientist, Focus on Machine Learning and Deep Learning. You can reach me from [Medium](#), [Linkedin](#) and [Github](#).

Reference

1. <https://gdcoder.com/decision-tree-regressor-explained-in-depth/>
2. [Introduction to Statistical Learning](#)
3. Raschka, Sebastian. Python Machine Learning
4. https://en.wikipedia.org/wiki/Decision_tree_learning
5. Bonaccorso, Giuseppe. Machine Learning Algorithm
6. <https://www.youtube.com/watch?v=7VeUPuFGJHk&t=911s>