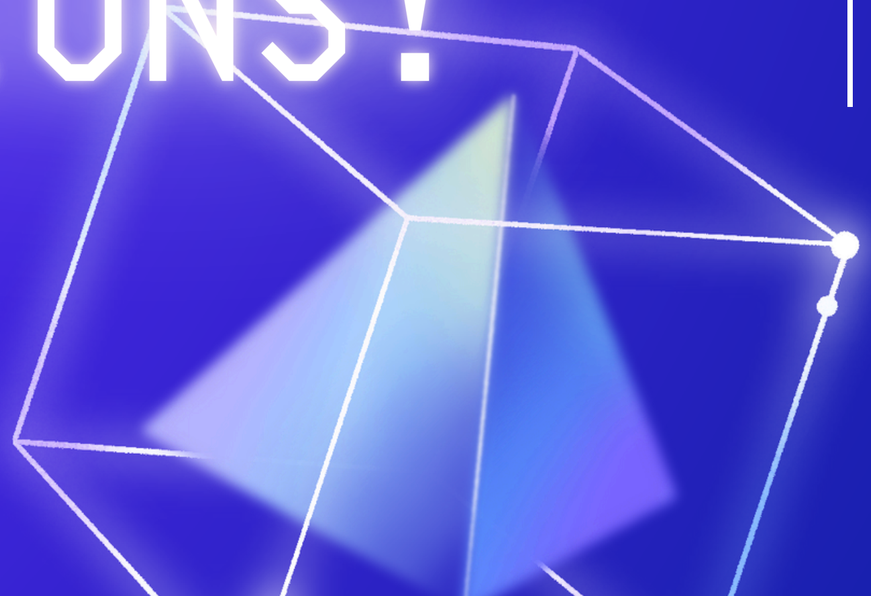
AI/MACHINE LEARNING
WORKSHOP

DAY 5: INTRODUCTION TO AI & ML: DECISION TREES

Youth Oppurtunties in Tech Inovation

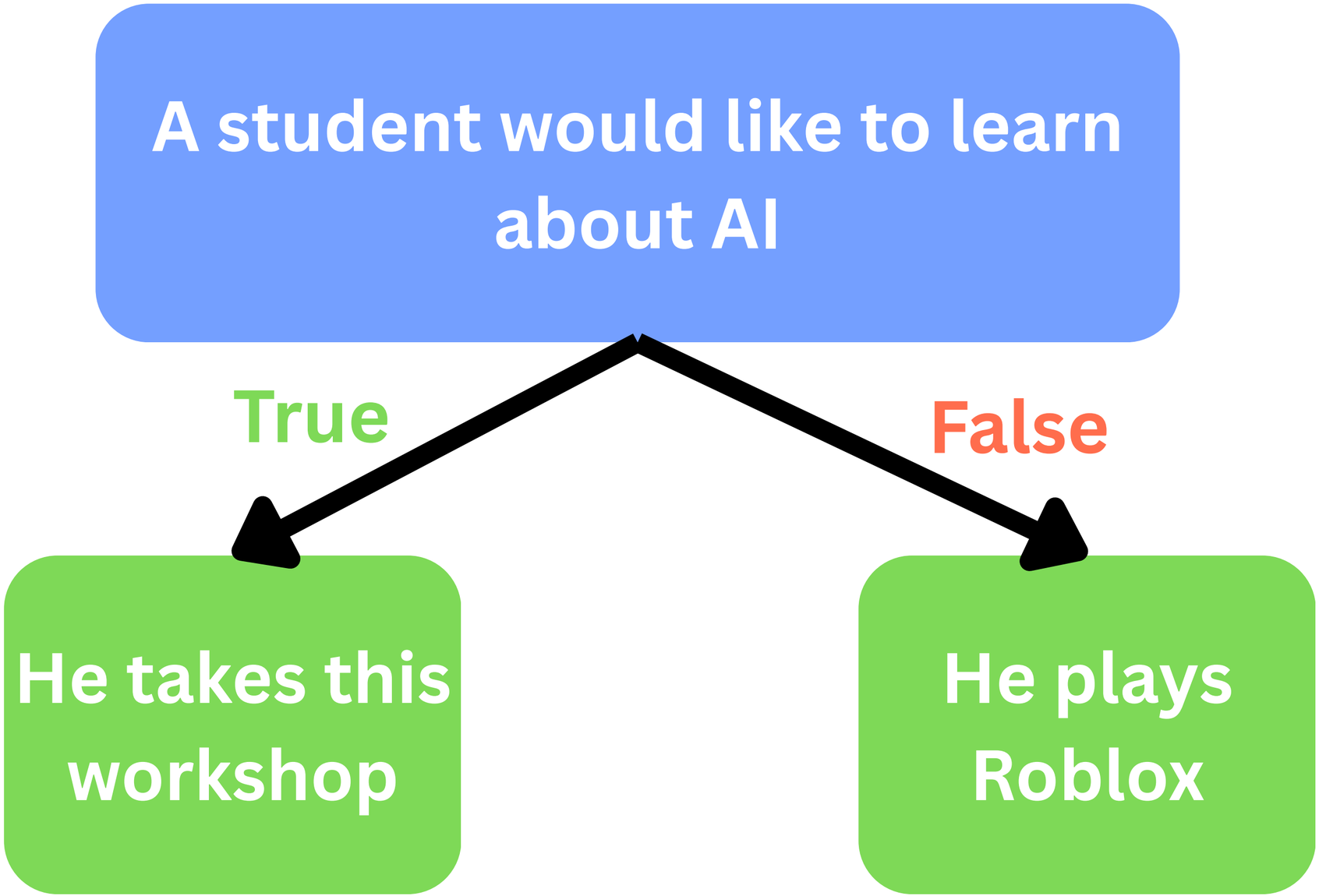
REMINDER PLEASE ASK QUESTIONS!



WHAT IS DECISION TREES:

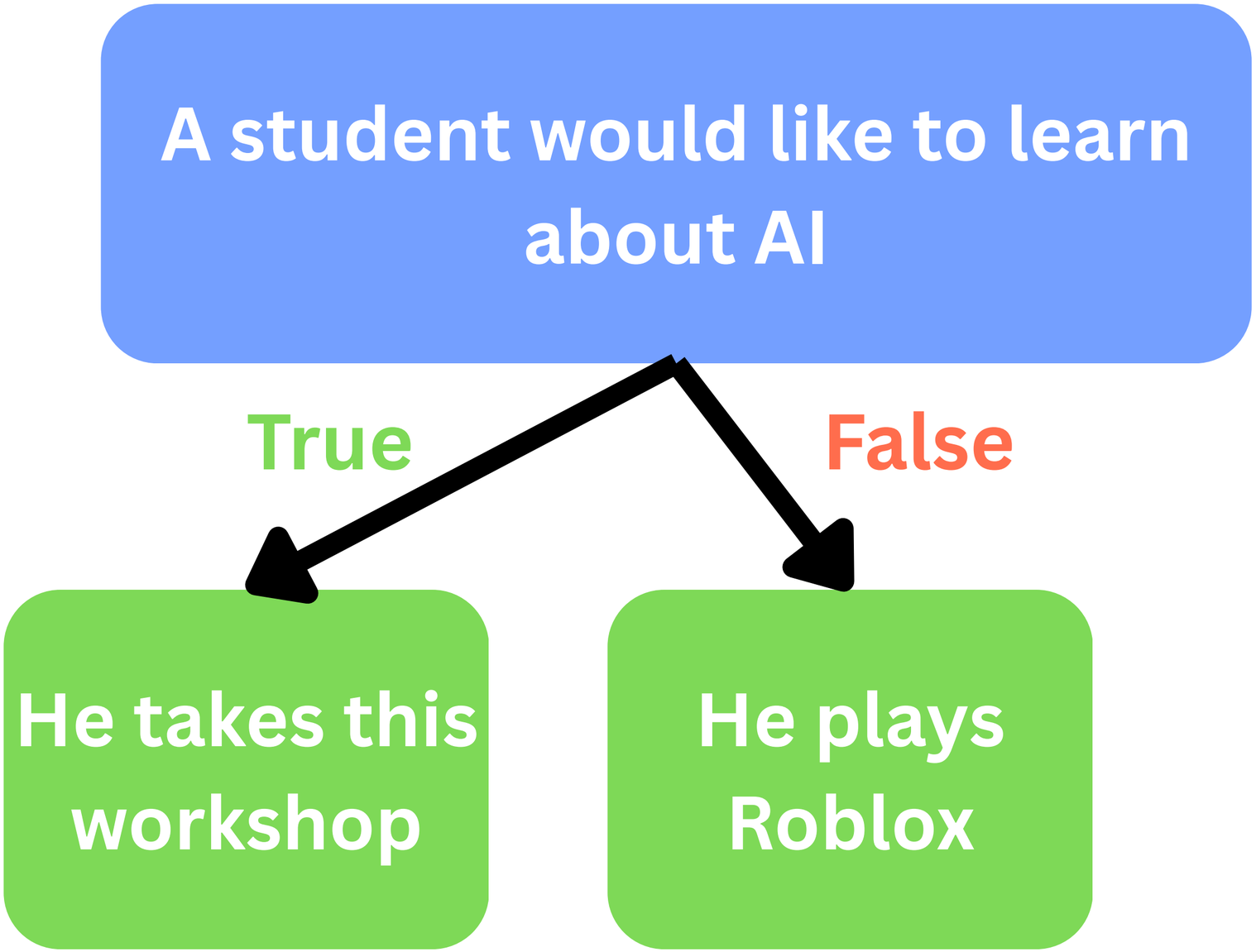
MODEL THAT MAKES PREDICTIONS BY SPLITTING DATA INTO BRANCHES BASED ON FEATURE VALUES, LEADING TO DECISION OUTCOMES AT THE LEAVES

Decision Trees

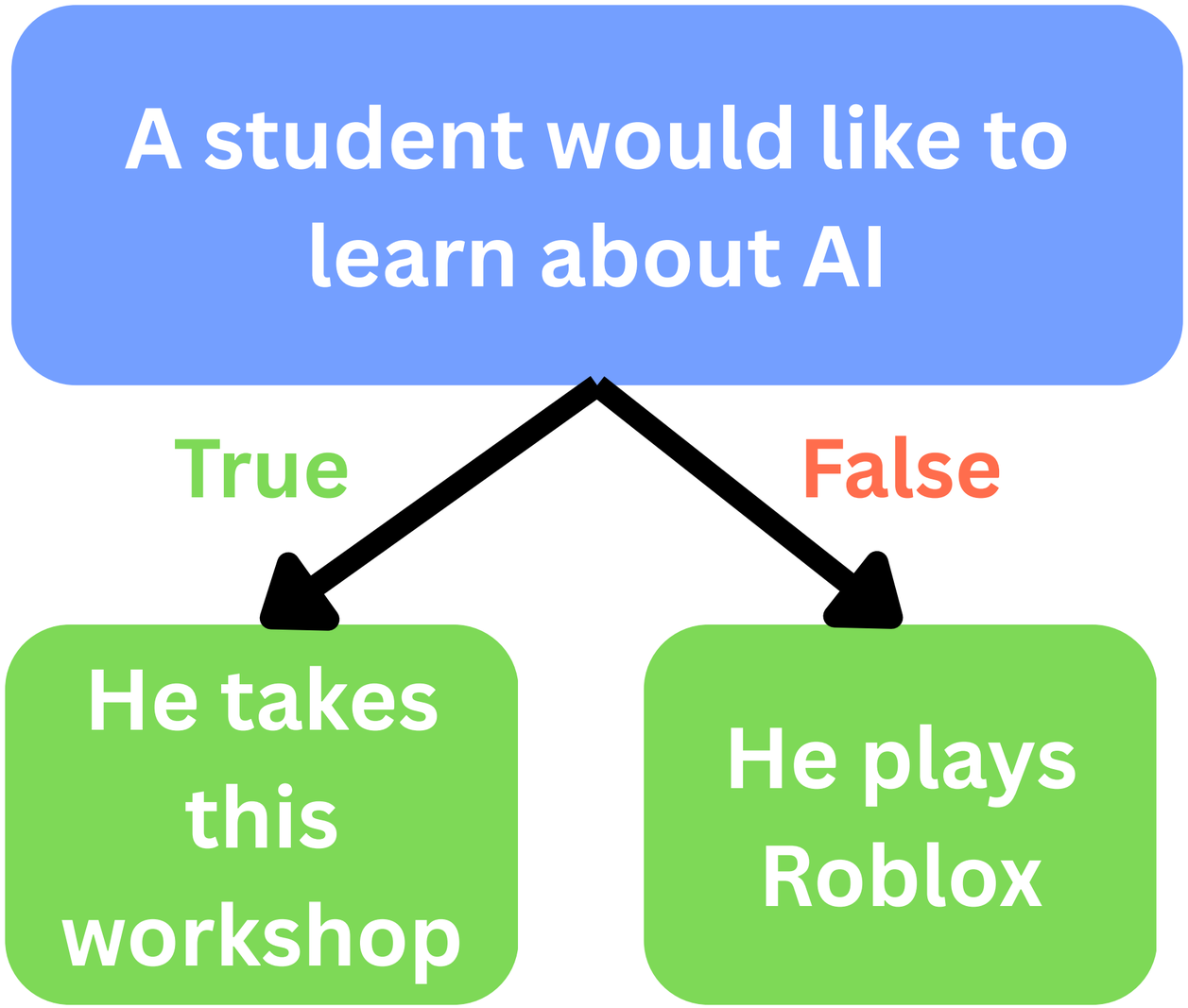


Decision Trees

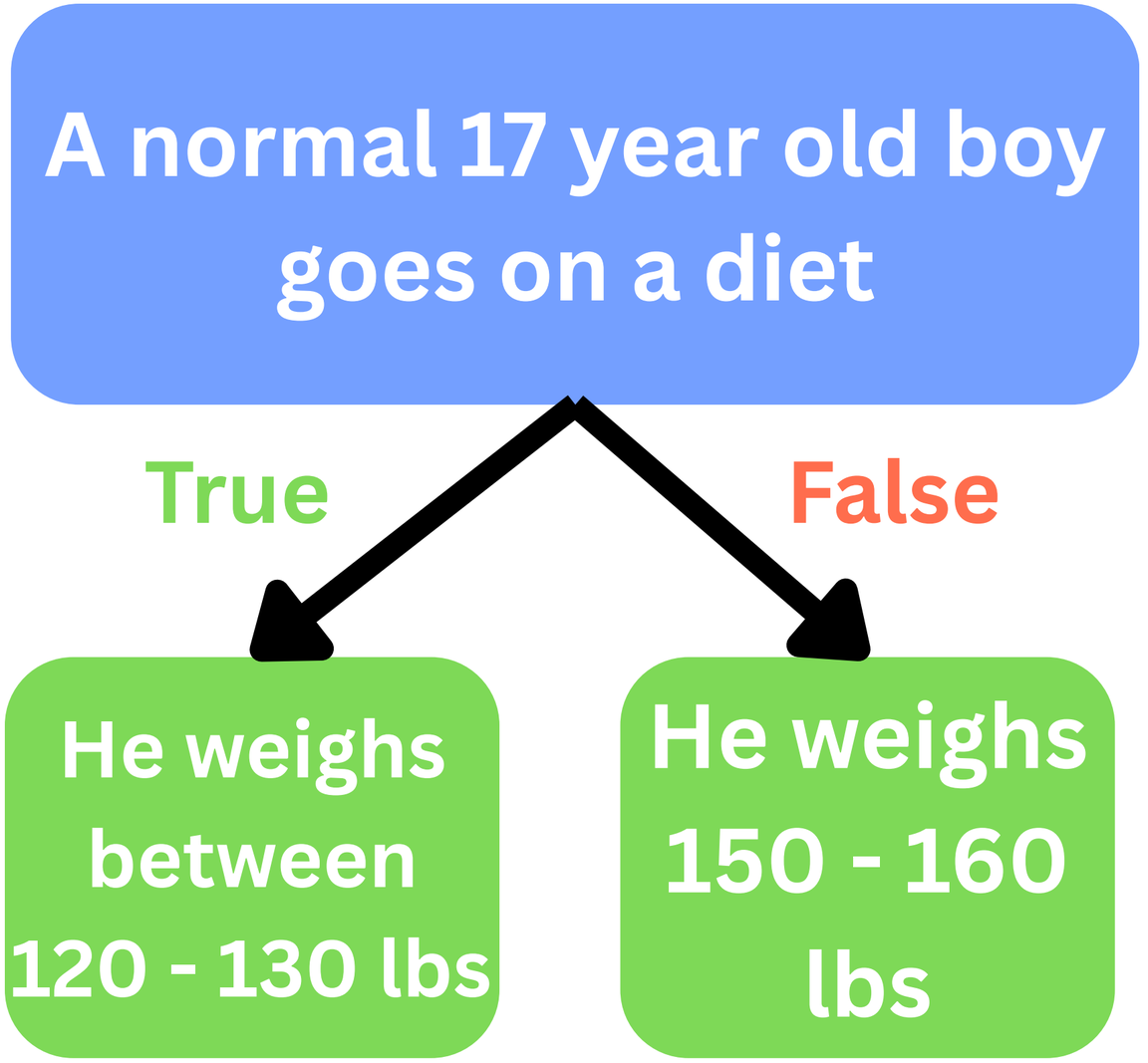
Decision trees make a decision based on weather it accepts the statement or not



Classifies into categories → Classification Trees



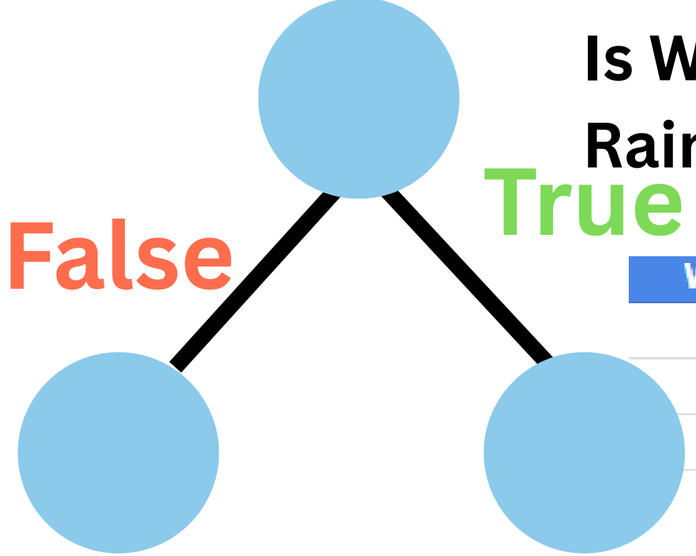
Predicts numeric values → Regression Trees



Lets use this dataset as an example

Weather	Temperature	Play Outside?
Sunny	Hot	No
Sunny	Mild	Yes
Overcast	Hot	Yes
Rainy	Mild	Yes
Rainy	Cold	No
Overcast	Cold	Yes

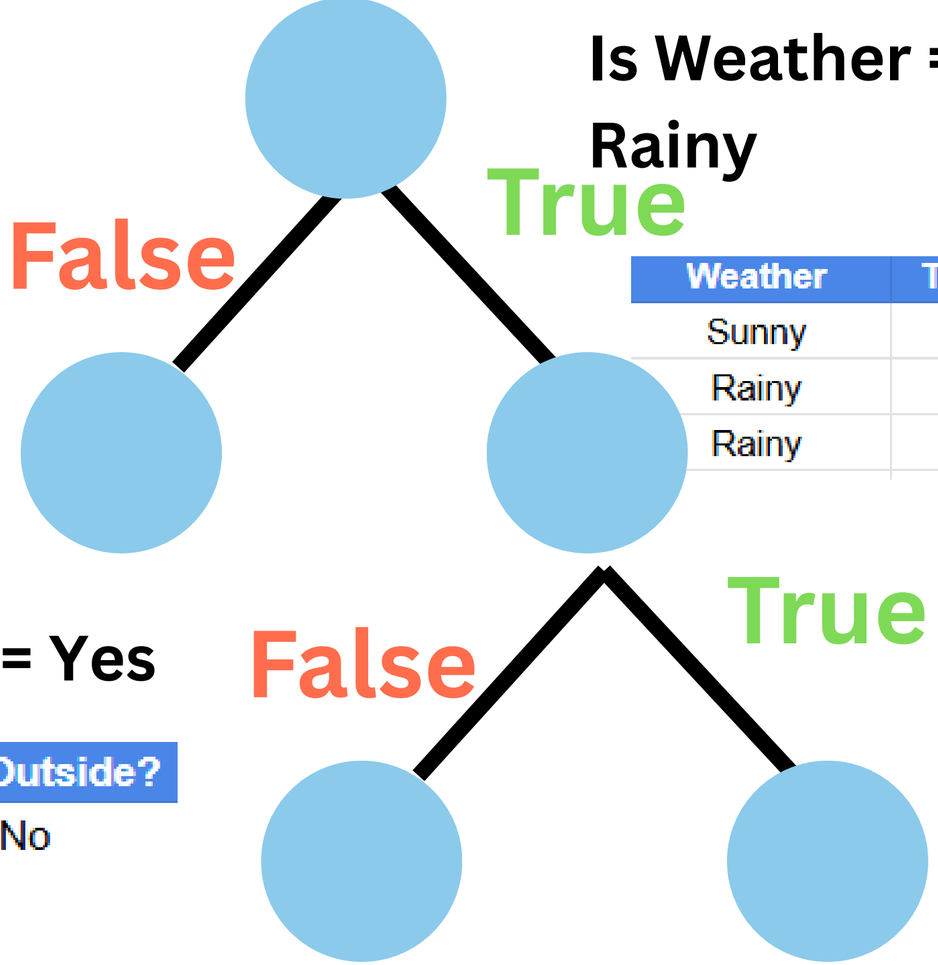
Weather	Temperature	Play Outside?
Overcast	Cold	Yes



Weather	Temperature	Play Outside?
Sunny	Mild	Yes
Rainy	Mild	Yes
Rainy	Cold	No

Weather	Temperature	Play Outside?
Sunny	Hot	No
Sunny	Mild	Yes
Overcast	Hot	Yes
Rainy	Mild	Yes
Rainy	Cold	No
Overcast	Cold	Yes

Weather	Temperature	Play Outside?
Overcast	Cold	Yes

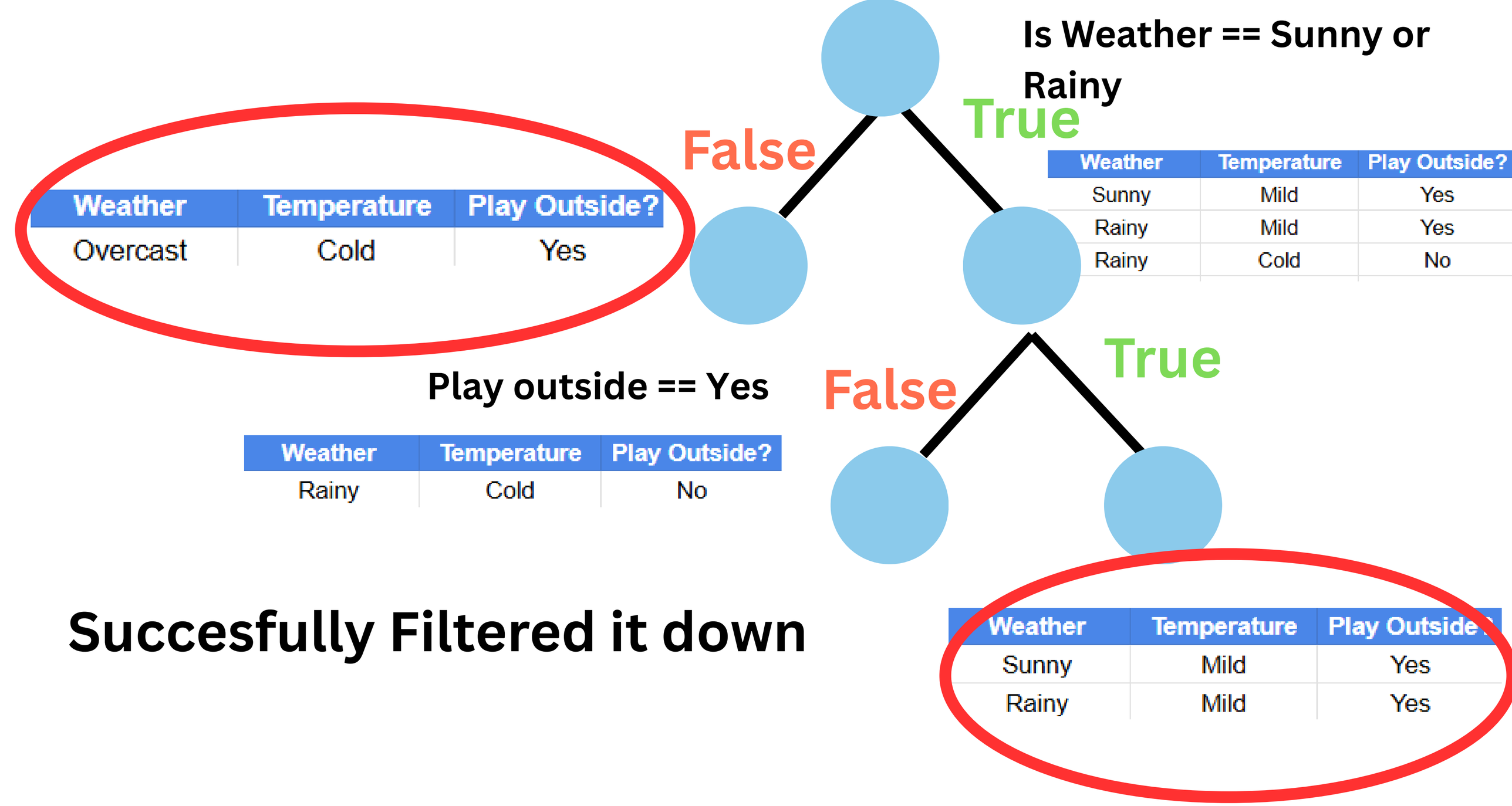


Weather	Temperature	Play Outside?
Sunny	Mild	Yes
Rainy	Mild	Yes
Rainy	Cold	No

Play outside == Yes

Weather	Temperature	Play Outside?
Rainy	Cold	No

Weather	Temperature	Play Outside?
Sunny	Mild	Yes
Rainy	Mild	Yes



Weather	Temperature	Play Outside?
Overcast	Cold	Yes

Weather	Temperature	Play Outside?
Sunny	Mild	Yes
Rainy	Mild	Yes
Rainy	Cold	No

Weather	Temperature	Play Outside?
Rainy	Cold	No

Weather	Temperature	Play Outside?
Sunny	Mild	Yes
Rainy	Mild	Yes

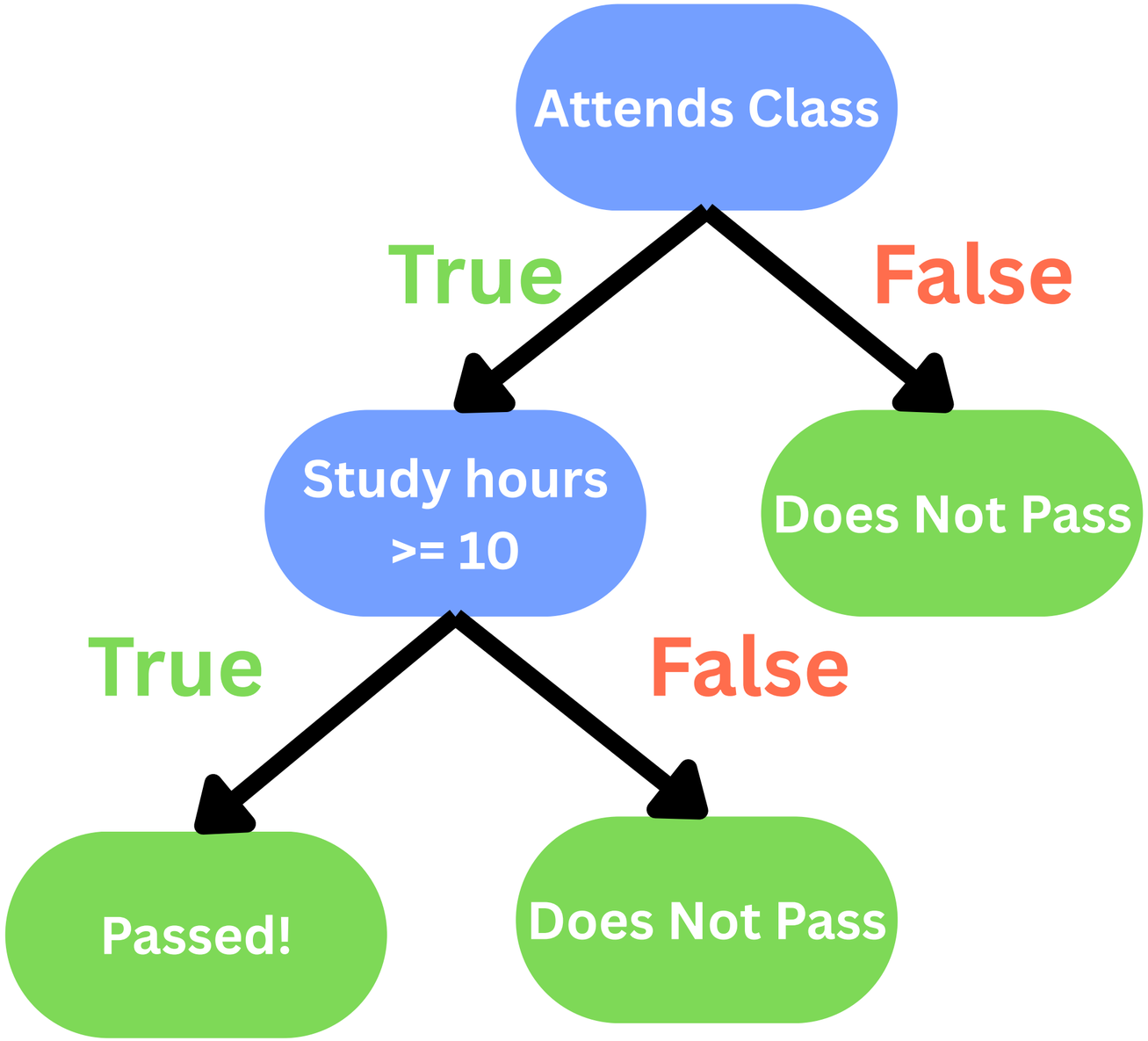
Succesfully Filtered it down

NOW LETS BUILD A TREE WITH GINI IMPURITY

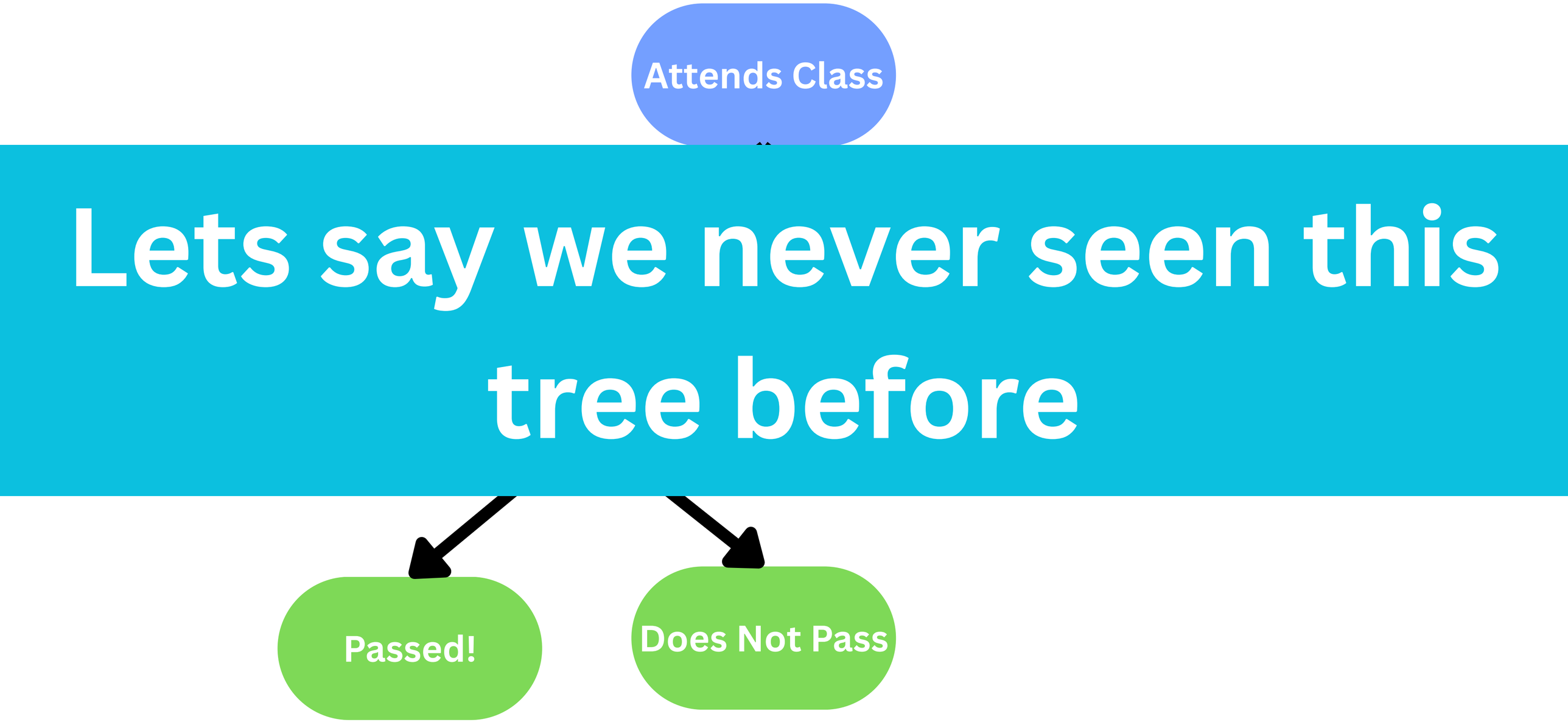
Lets use this dataset as an example

Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes

This is how our tree would look

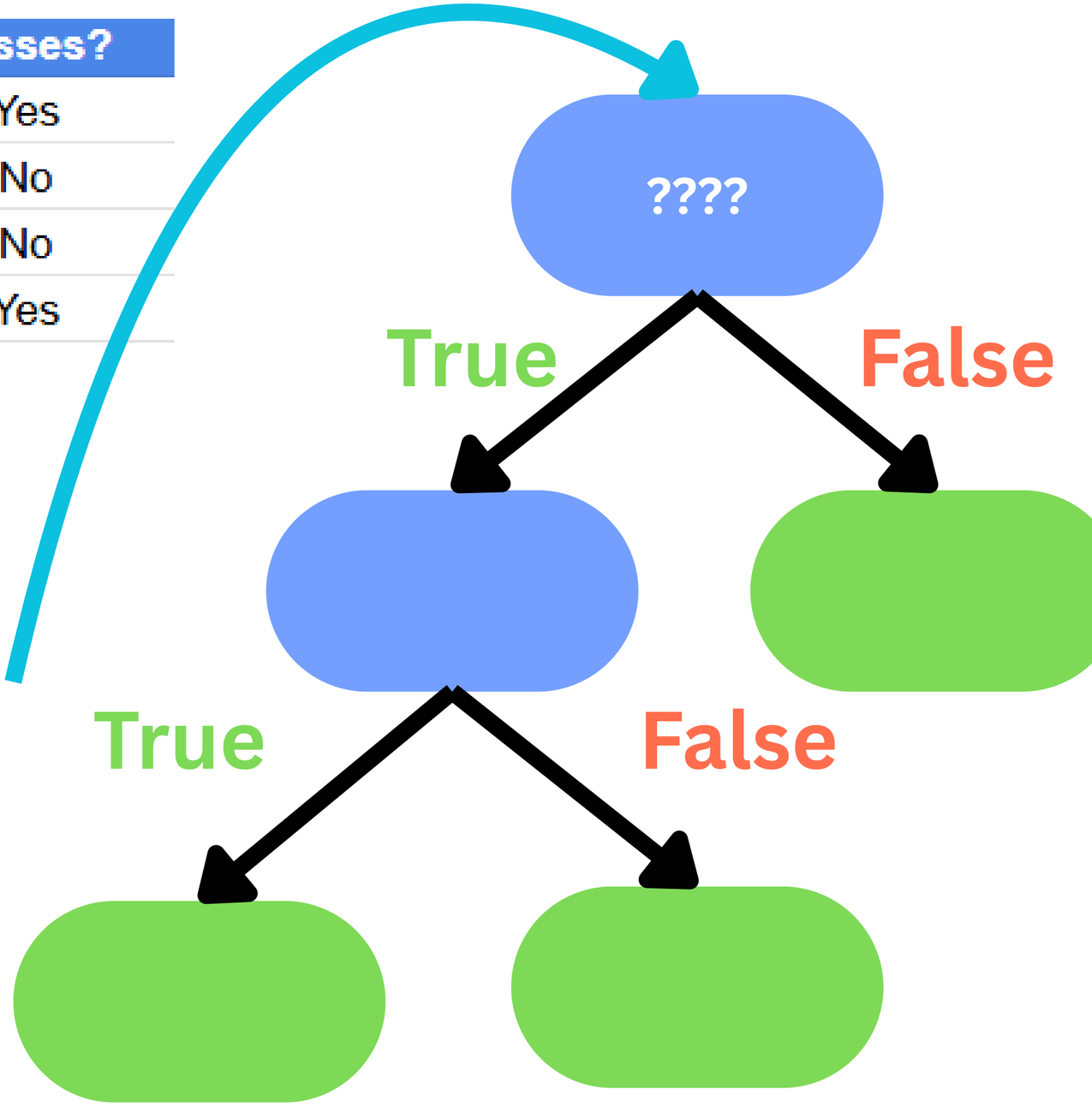


This is how our tree would look



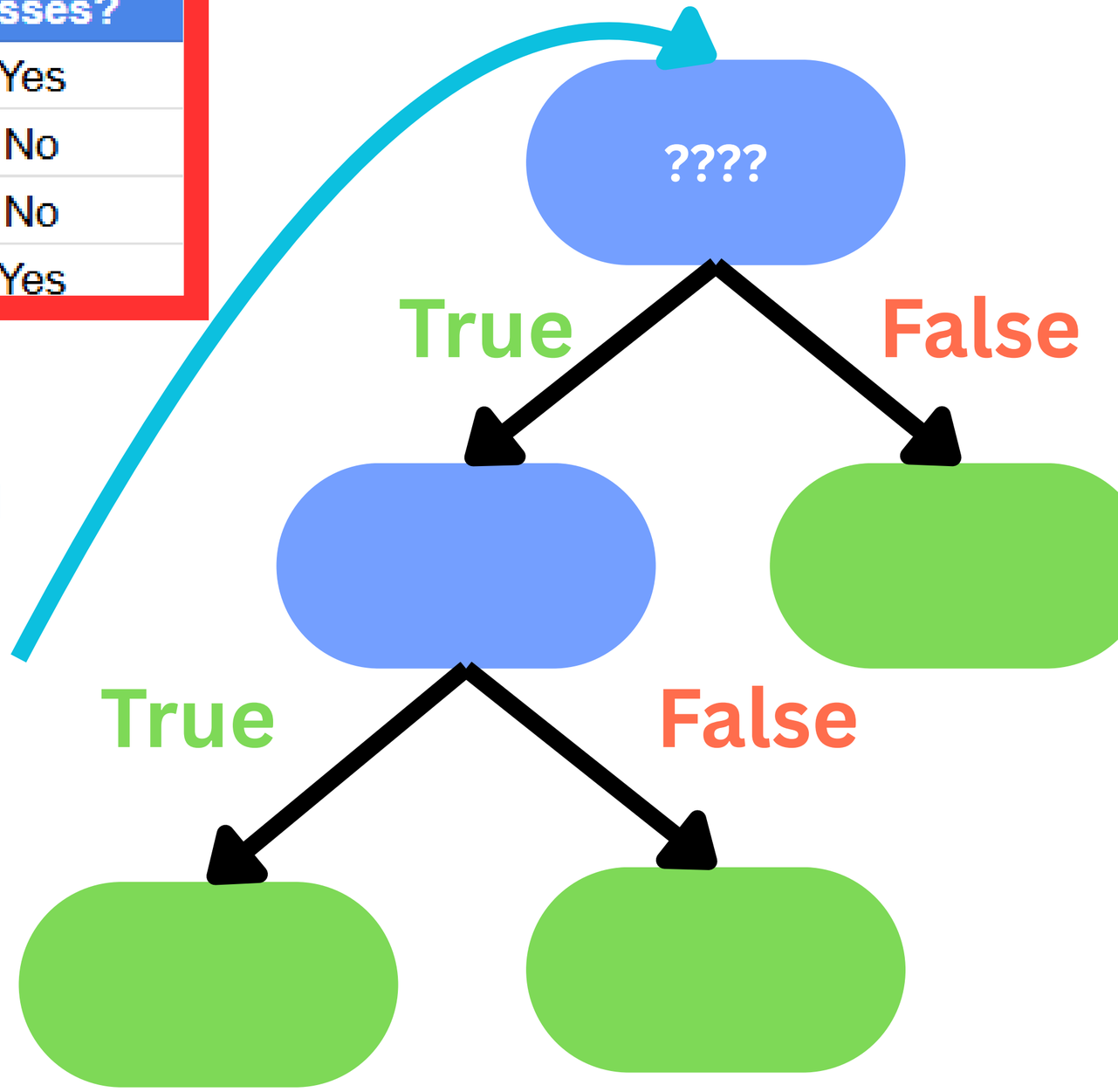
Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes

We need to know what should be the first question (Attends Class, Homework, Study Hours)



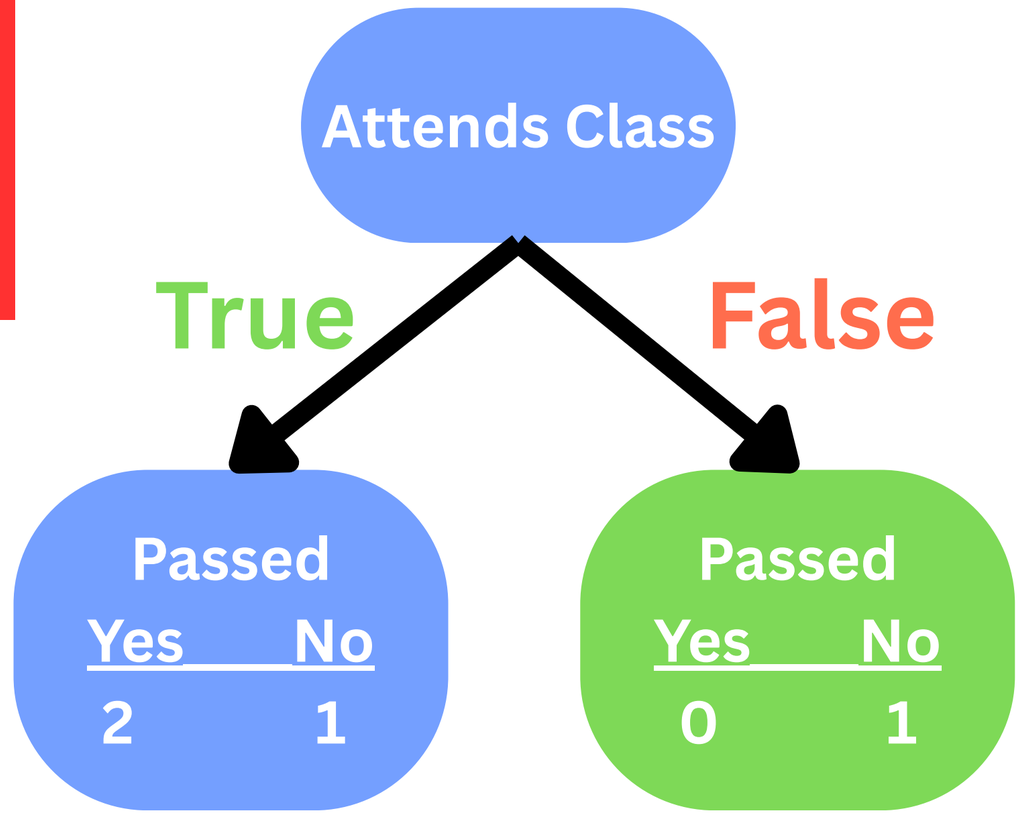
Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes

To do this we will start with attends class and see how that affects if you pass or not

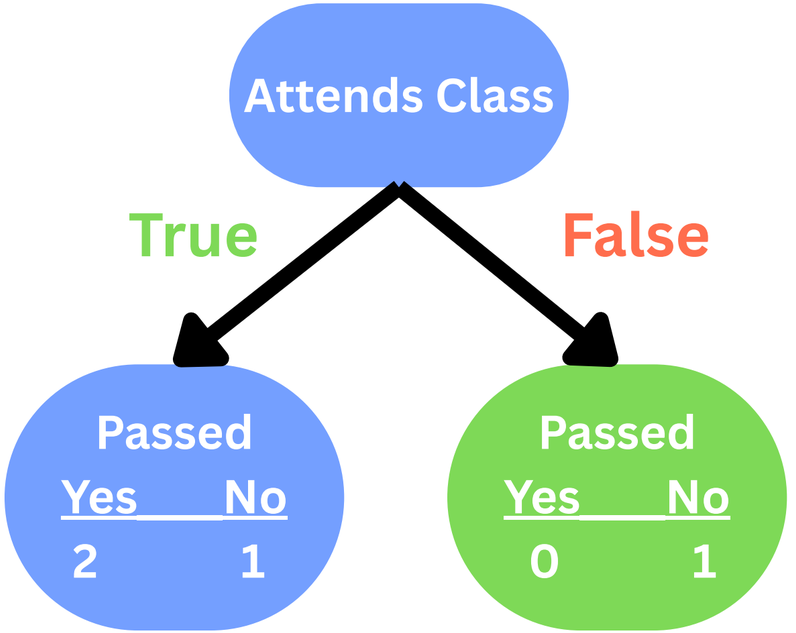


Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes

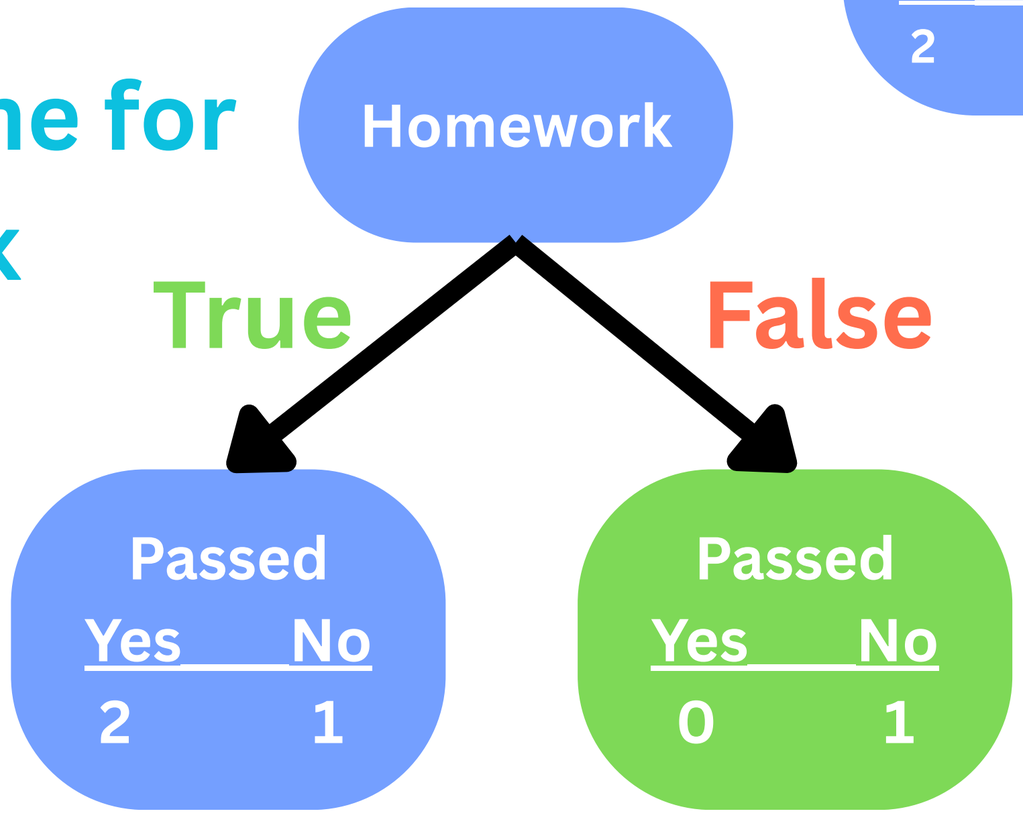
Lets make a simple tree to figure it out!



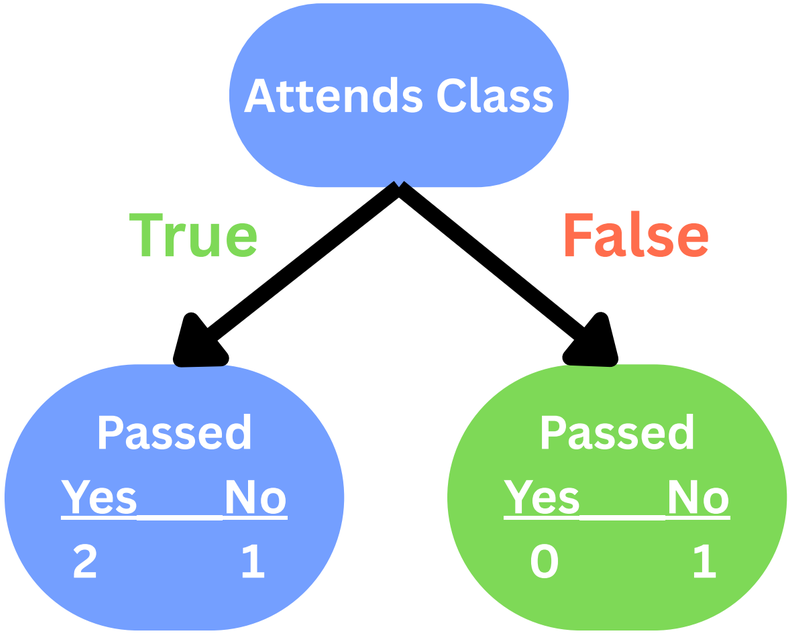
Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes



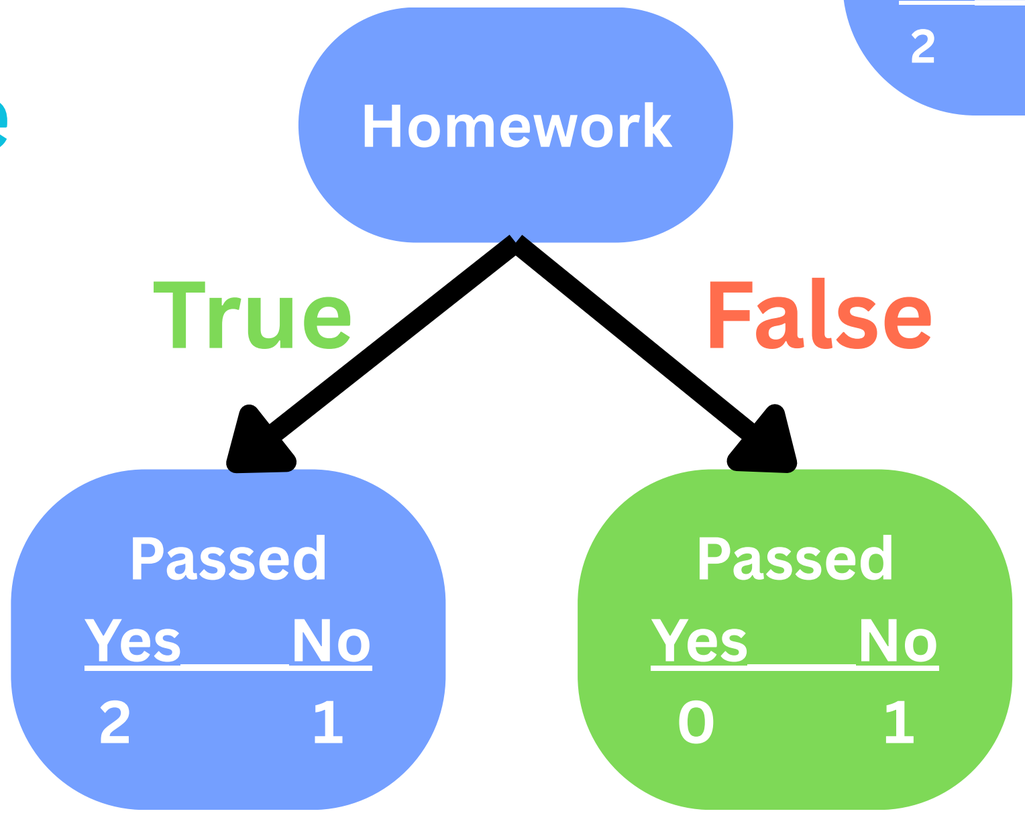
Lets do the same for Homework



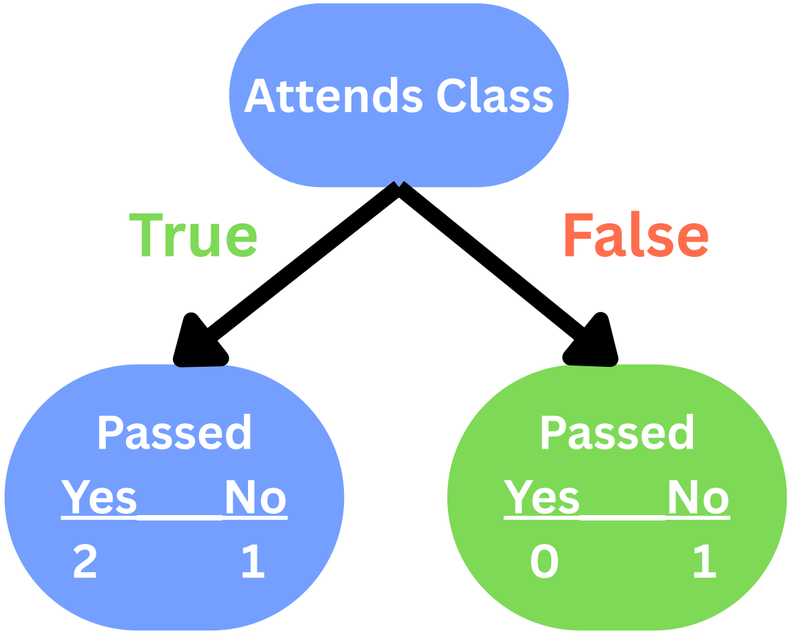
Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes



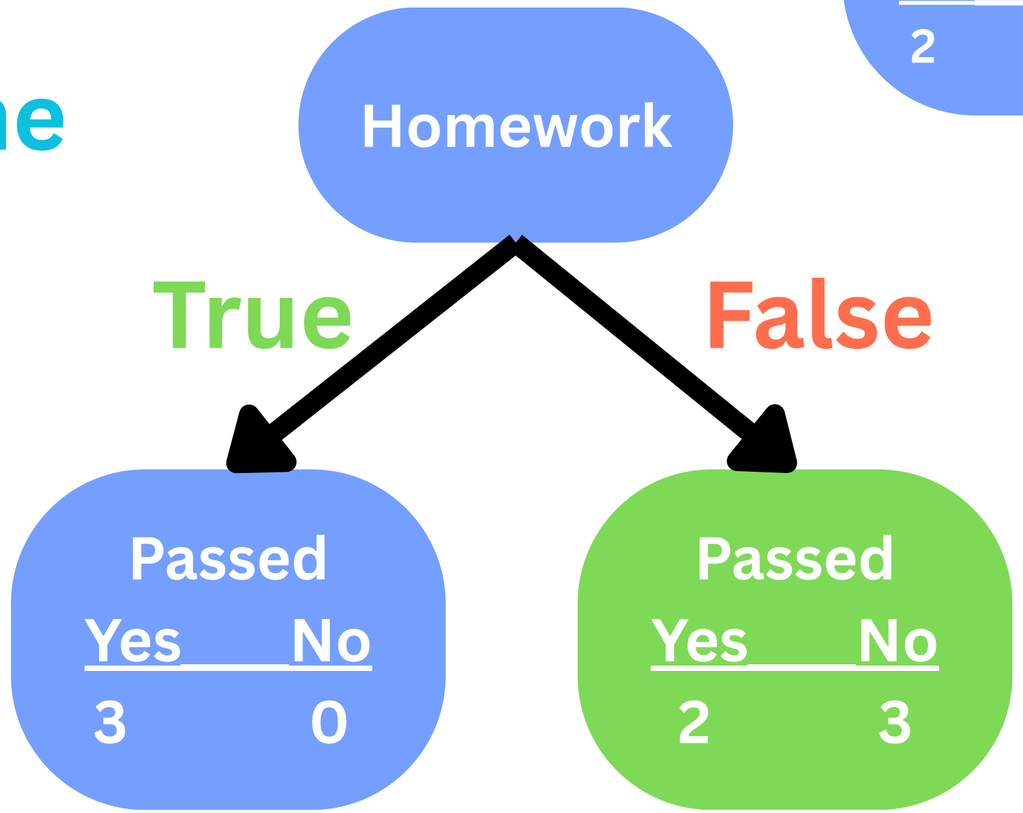
Now that we have these we notice they both are the same



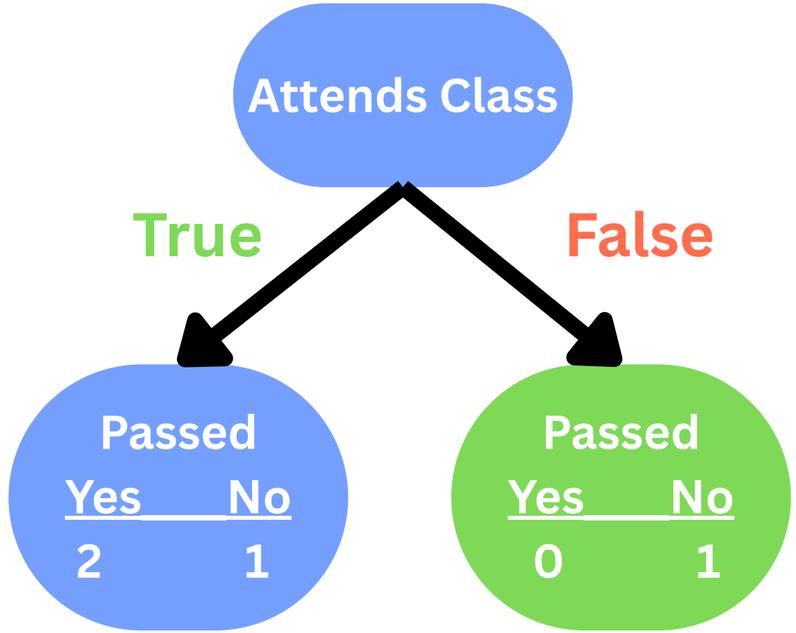
Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes



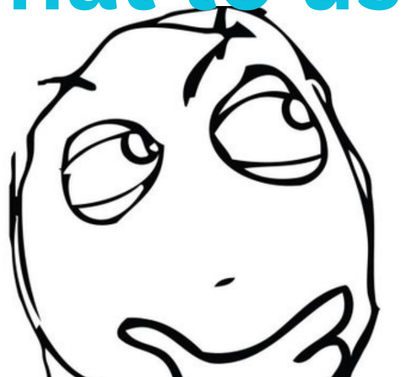
Lets just make some values up for this example

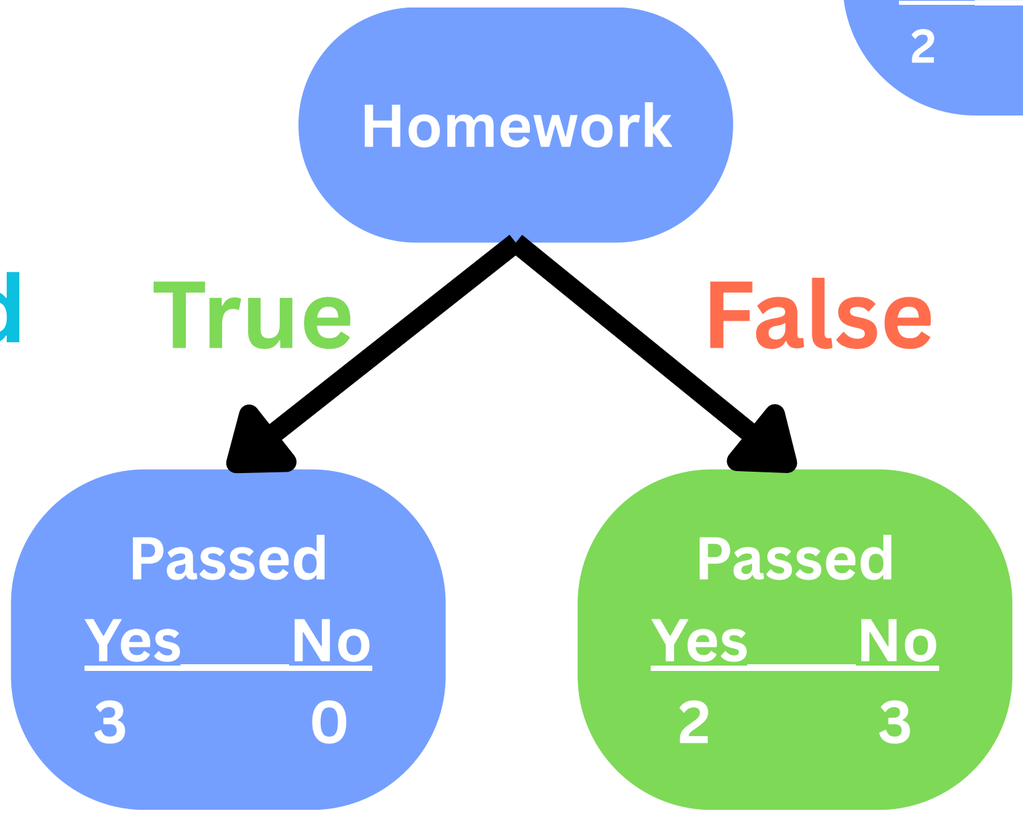


Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes



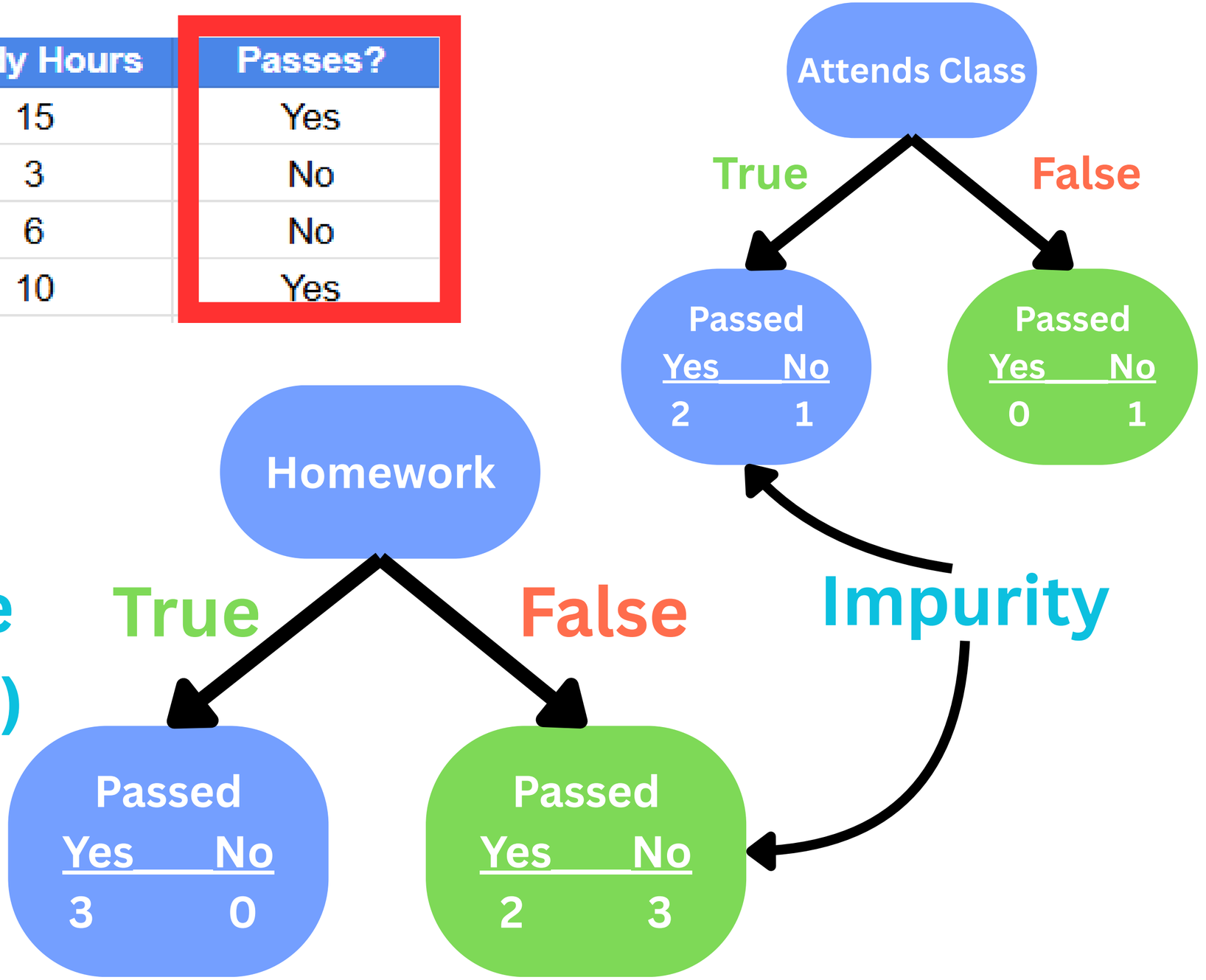
How can we quantify the difference and find what to use



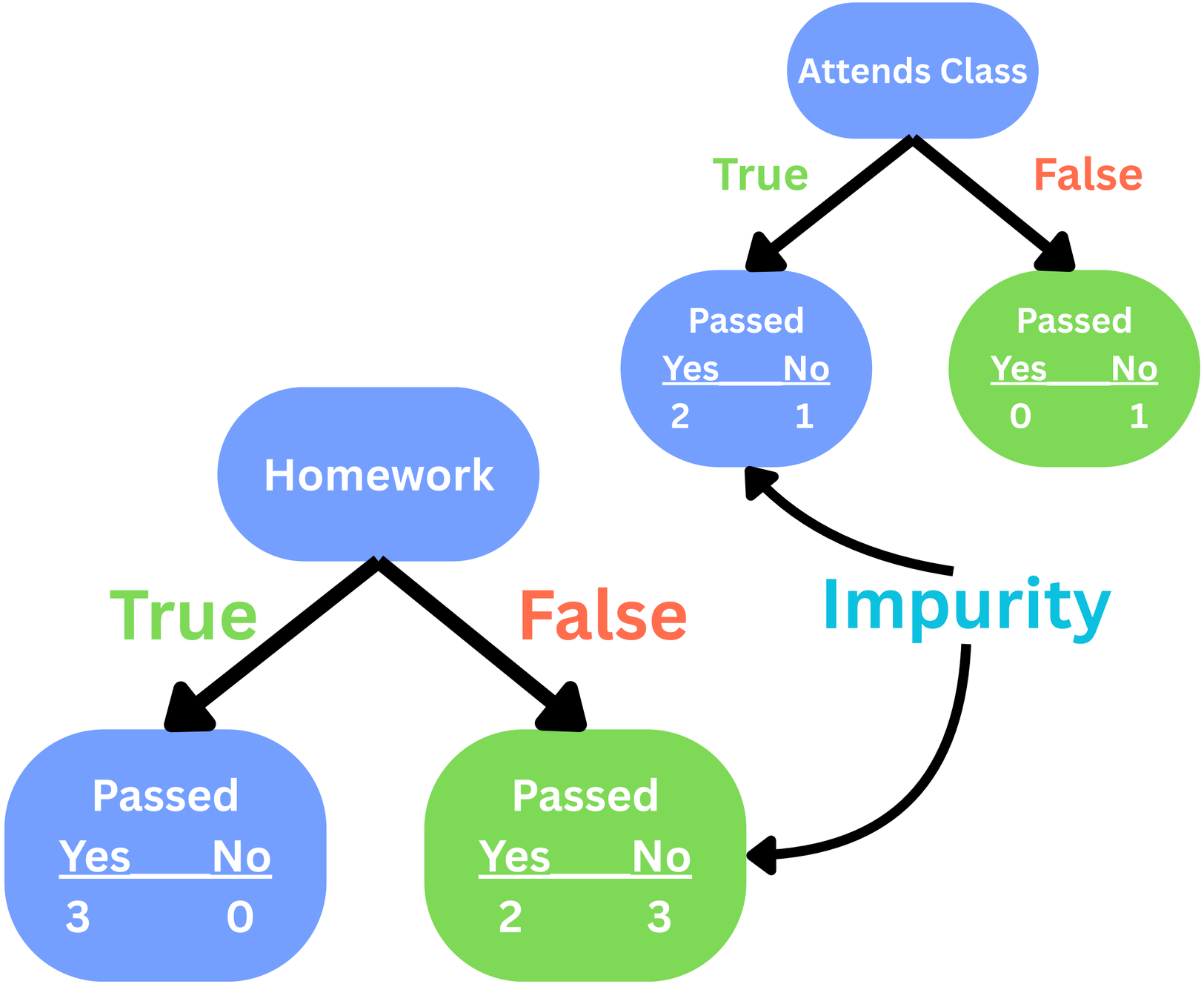


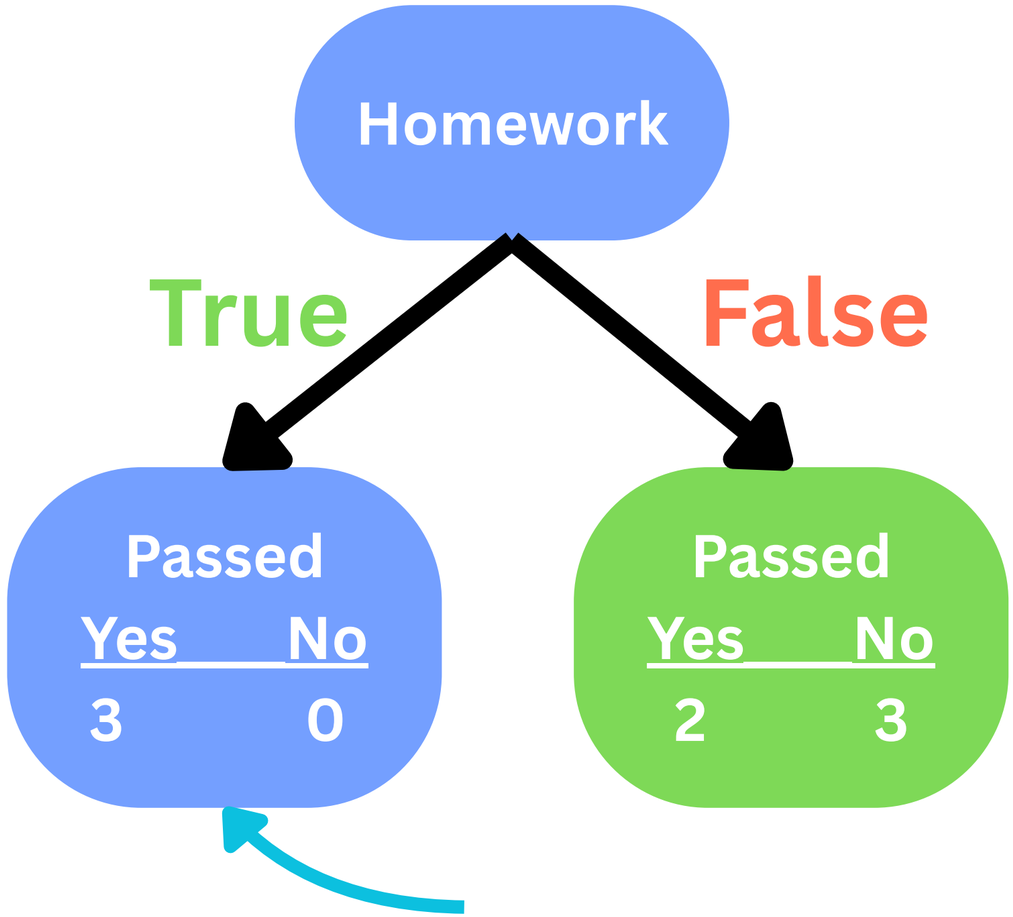
Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes

There are many ways to calculate the impurity of the leaves (final nodes)

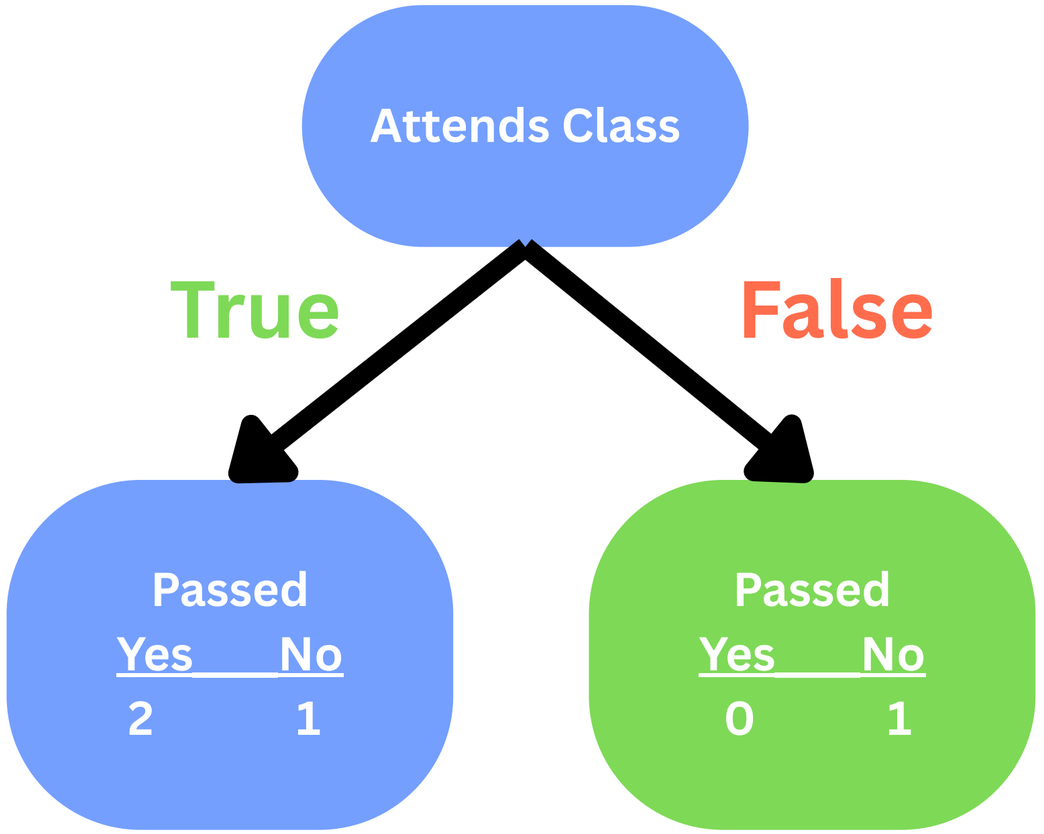


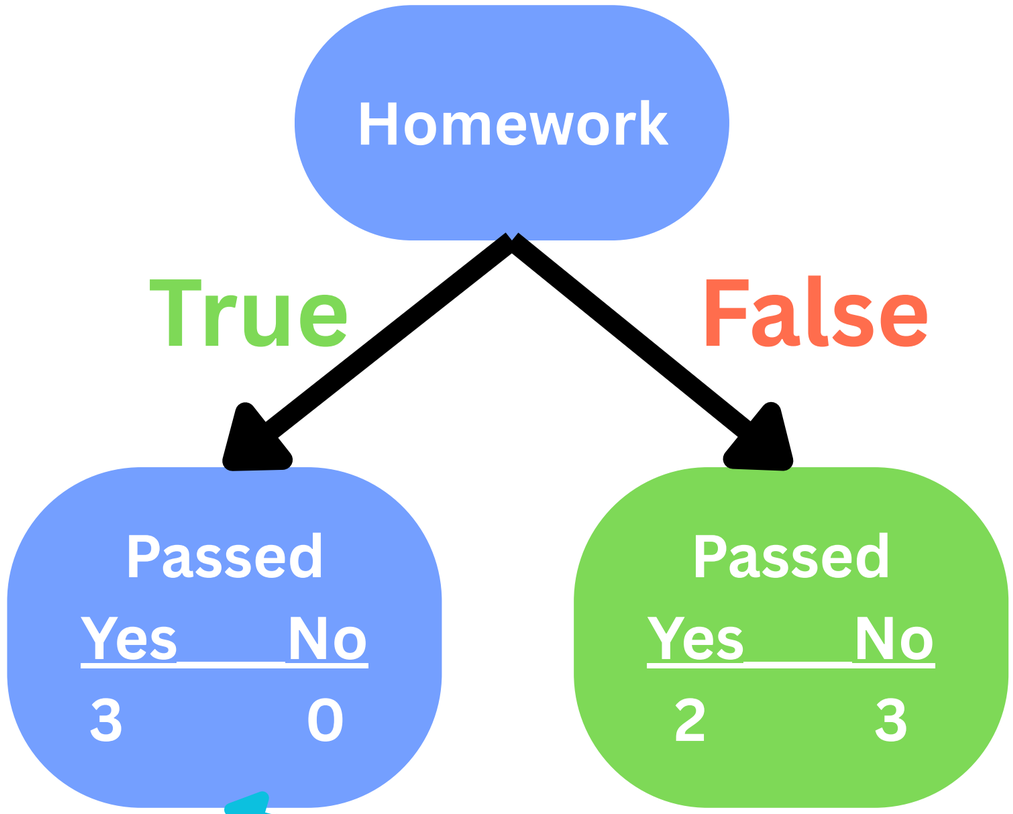
Lets use Gini Impurity Method

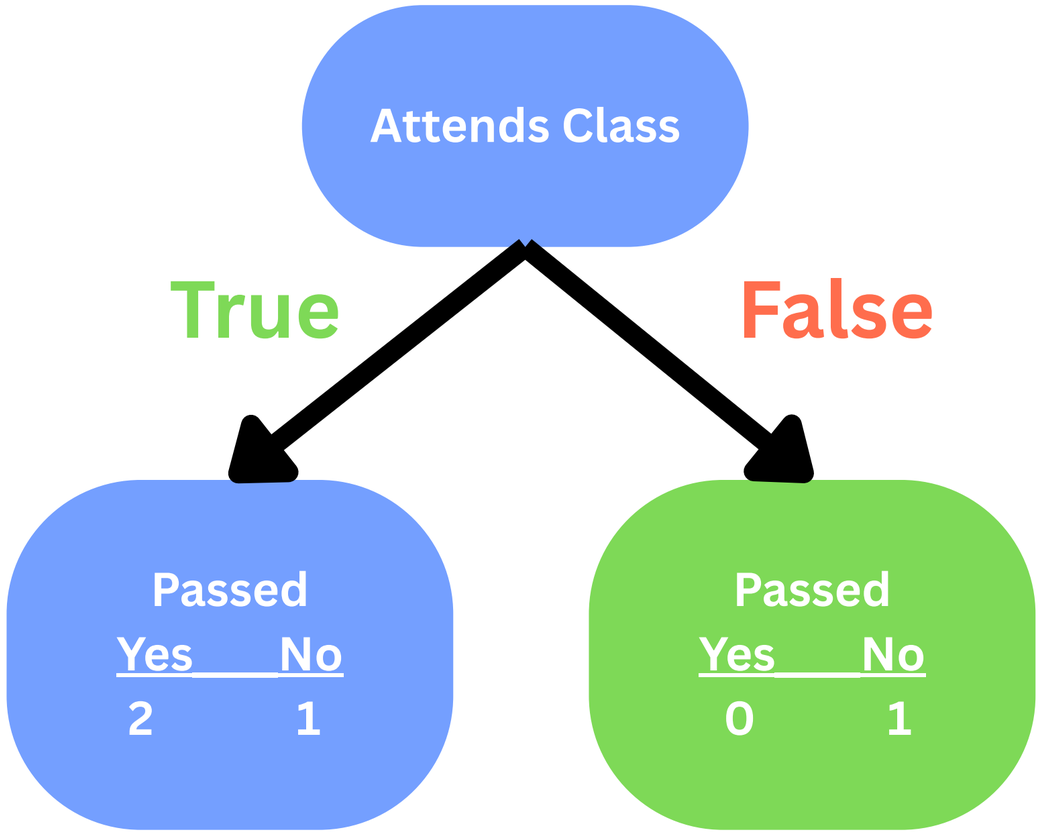




Lets start with the left leaf

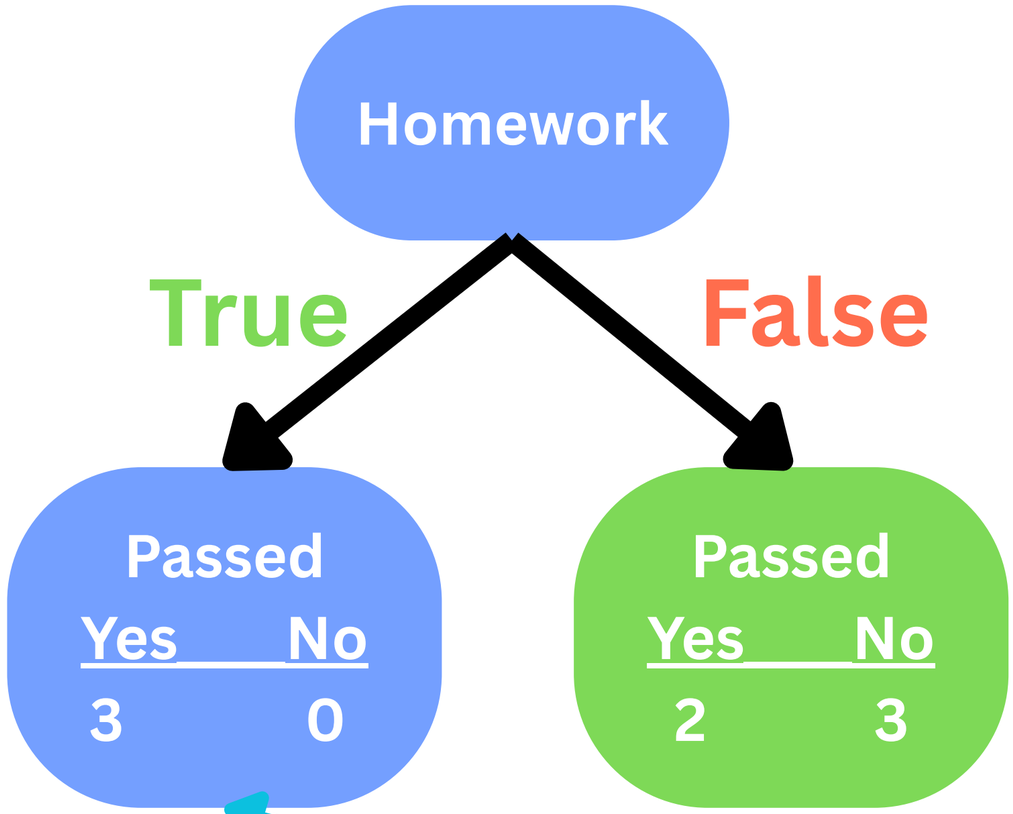


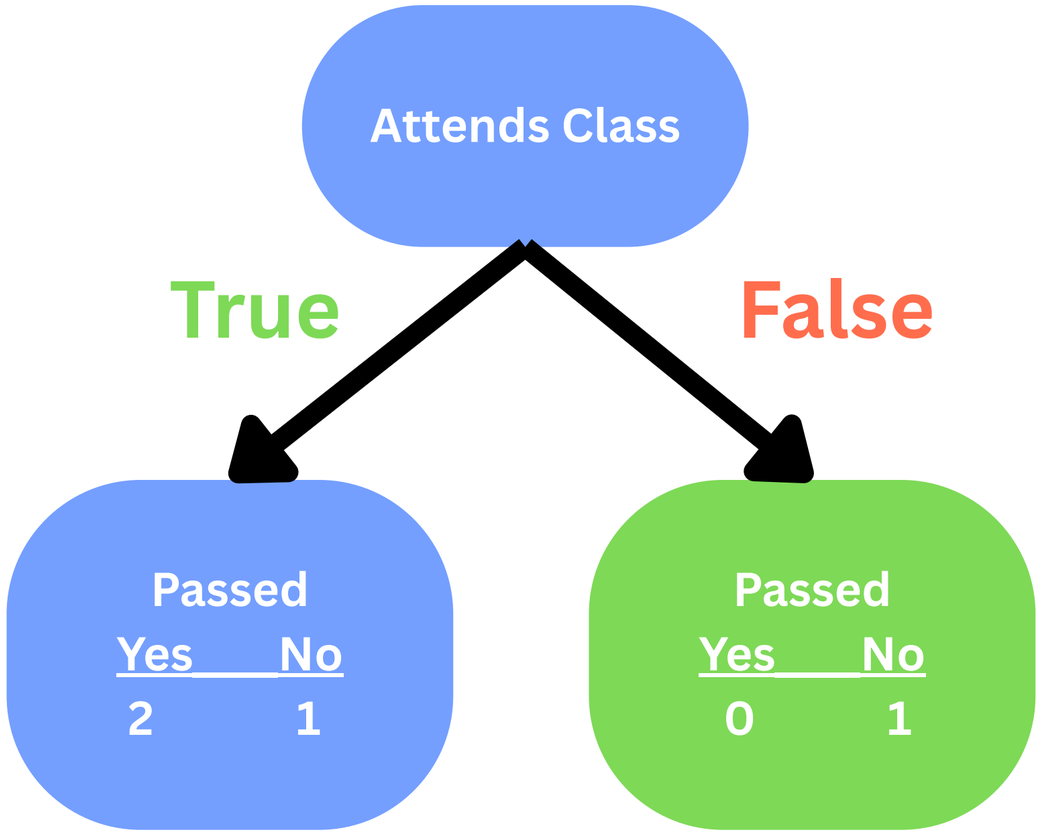




Gini Impurity for a leaf = 1 - (P(Yes))^2 - (P(No))^2

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

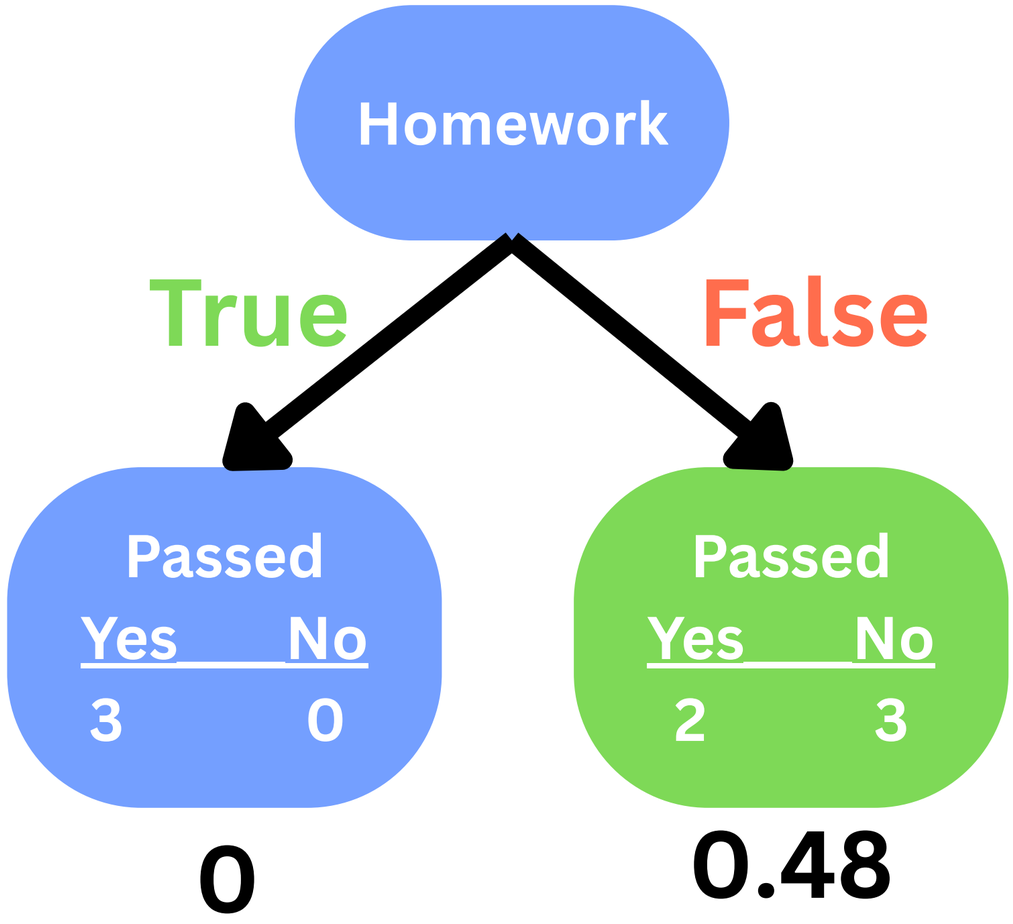


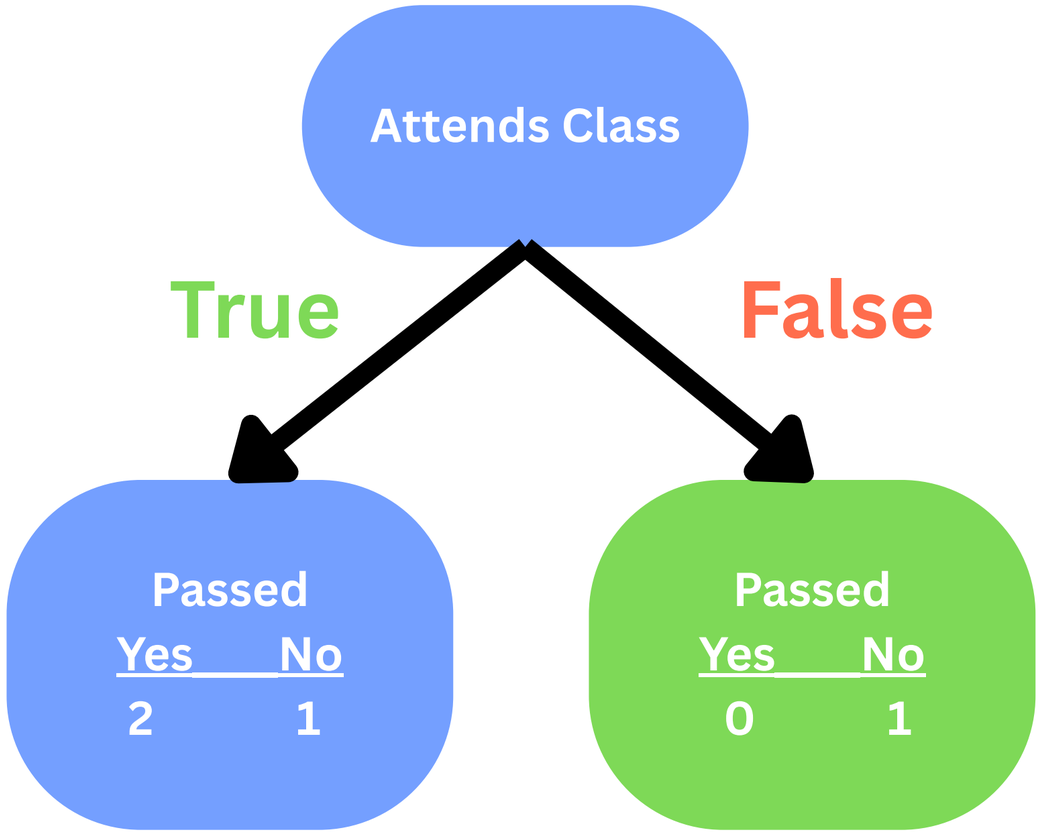


Gini Impurity for a leaf = $1 - (P(Yes))^2 - (P(No))^2$

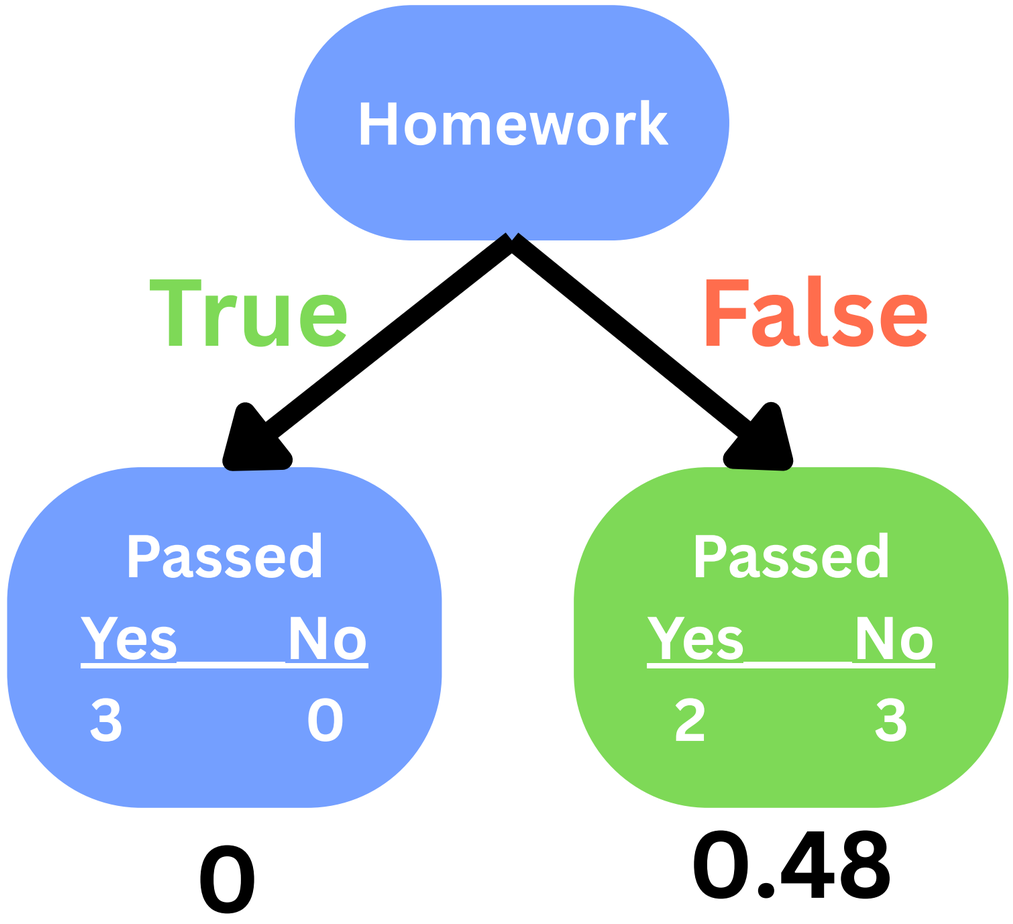
$$1 - (3/(3+0))^2 - (0/(3+0))^2$$

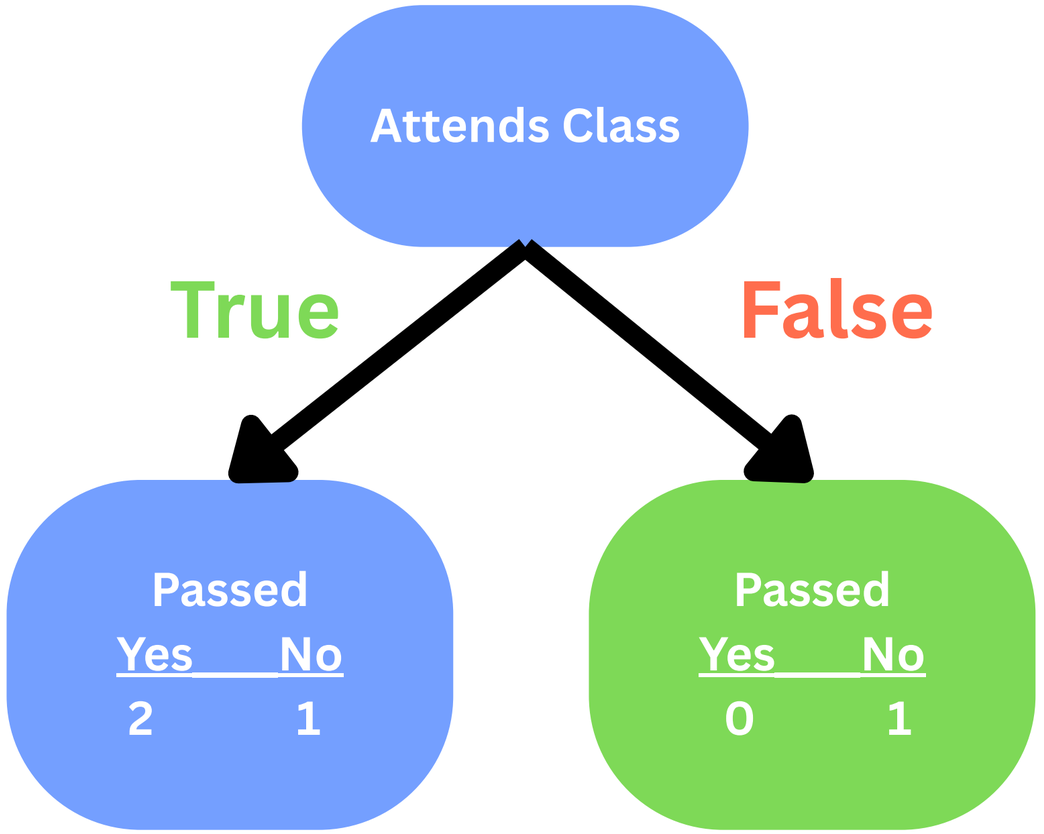
Since this is pure we get a Gini Impurity of 0



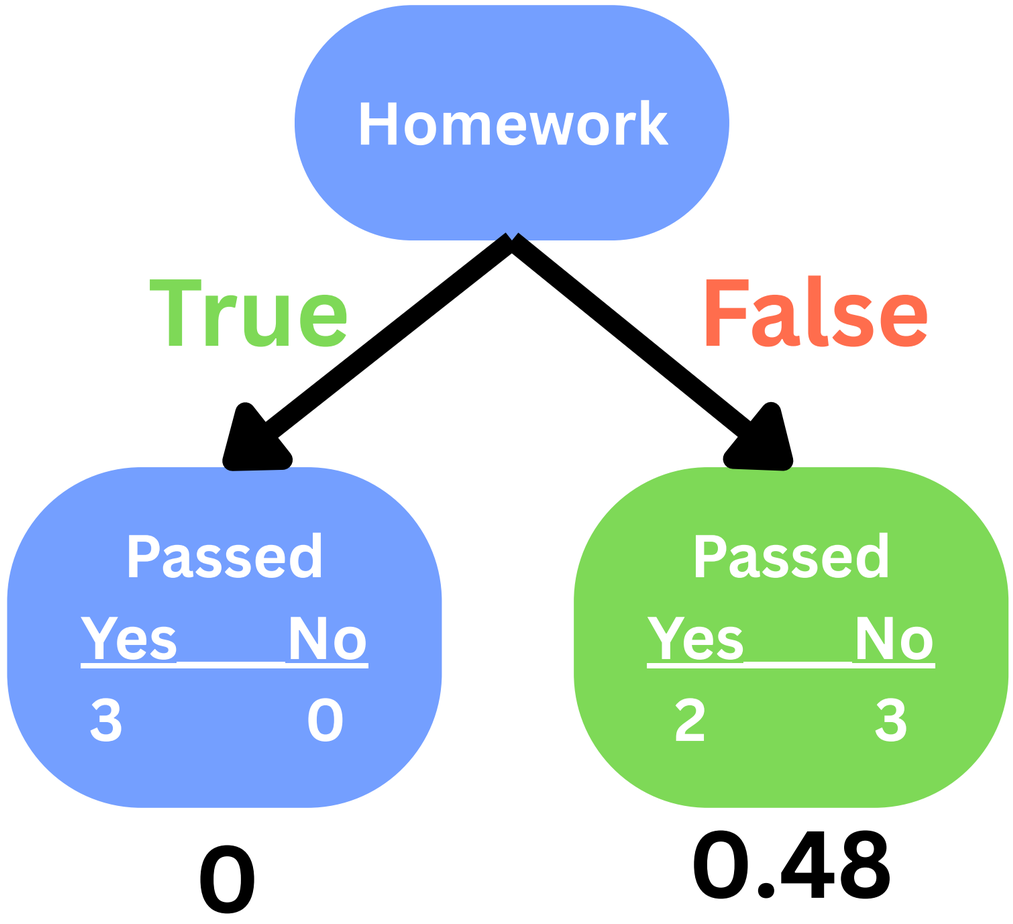


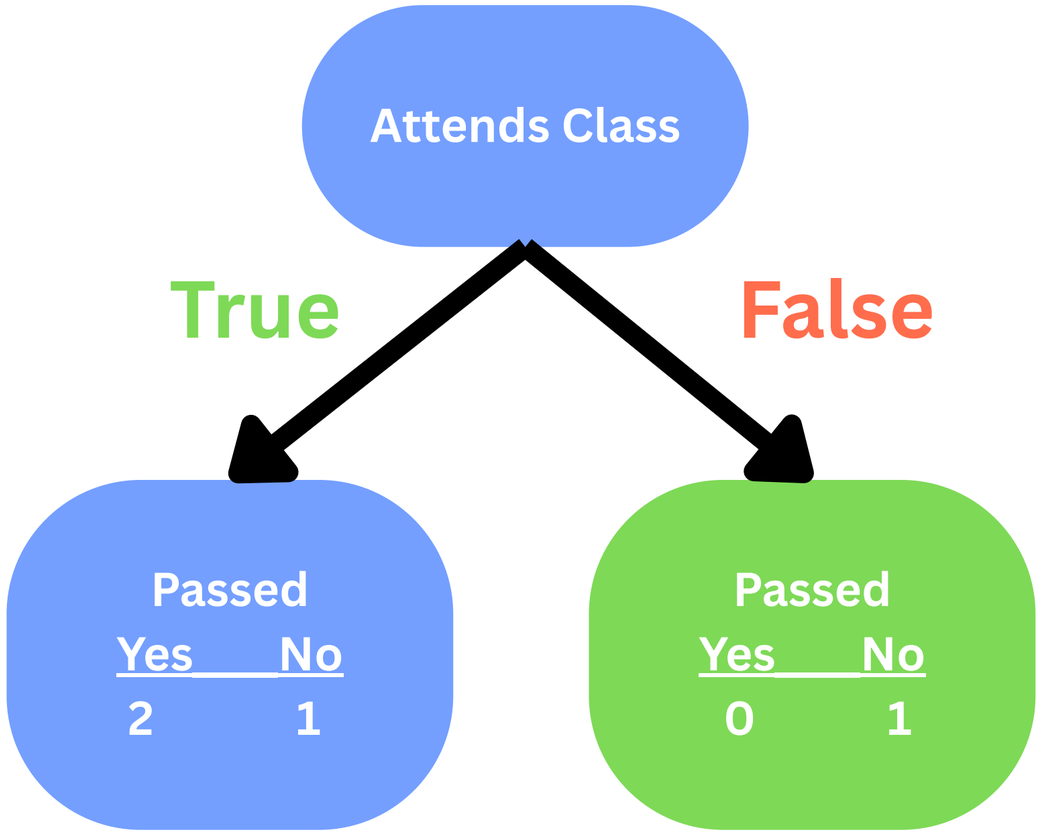
Since the number of people in both leaves are different (3 and 5 respectively) we need to get the weighted average to get the true Gini Impurity



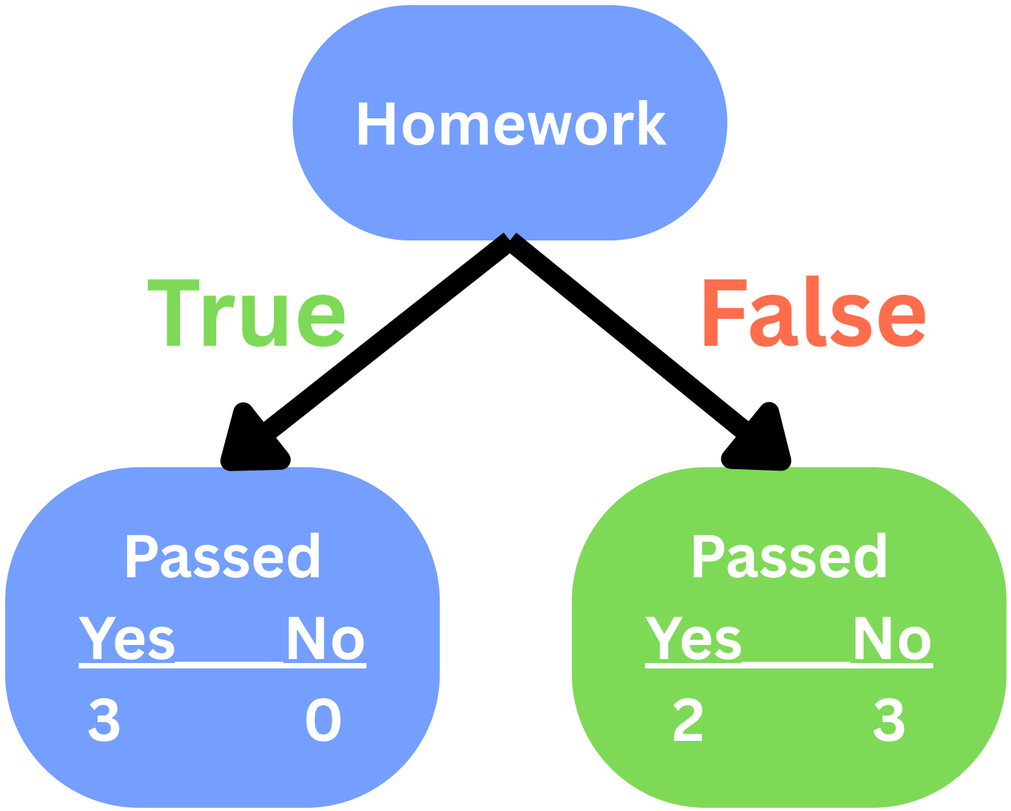


Since the number of people in both leaves are different (3 and 5 respectively) we need to get the weighted avaerage to get the total Gini Impurity





The weight is equal to the number of people in that leaf divided by total number of people in both leaves



0

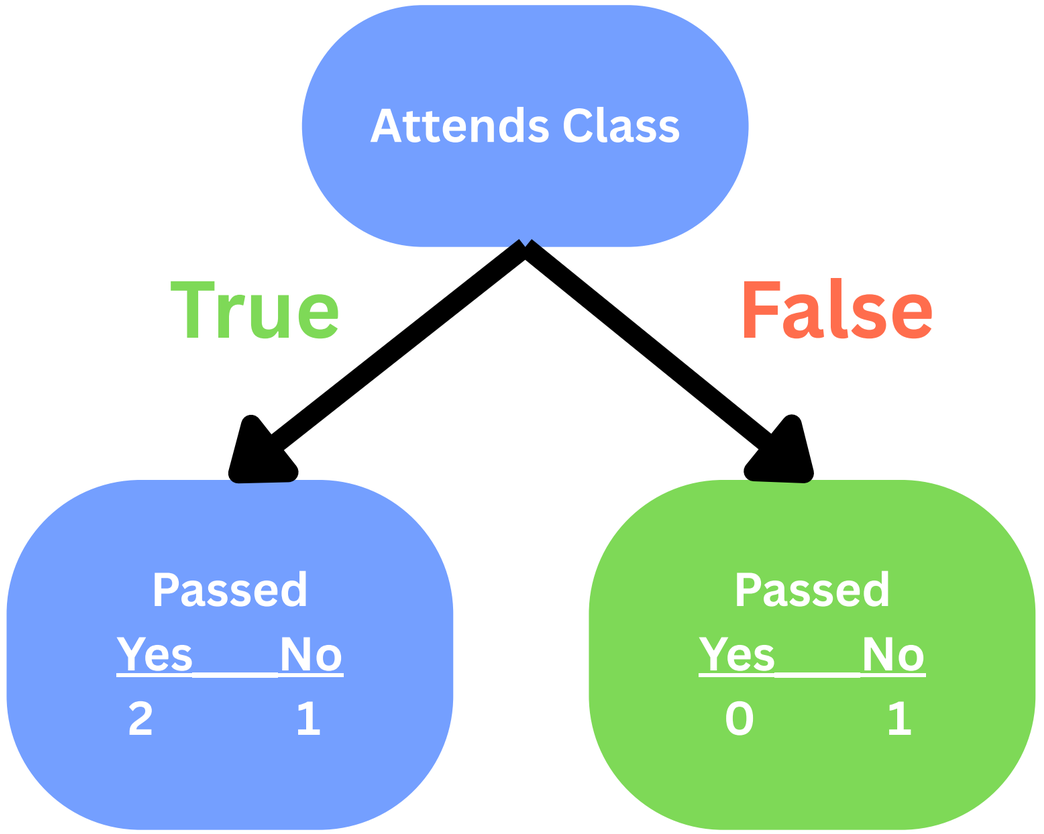
3/(3+5) =

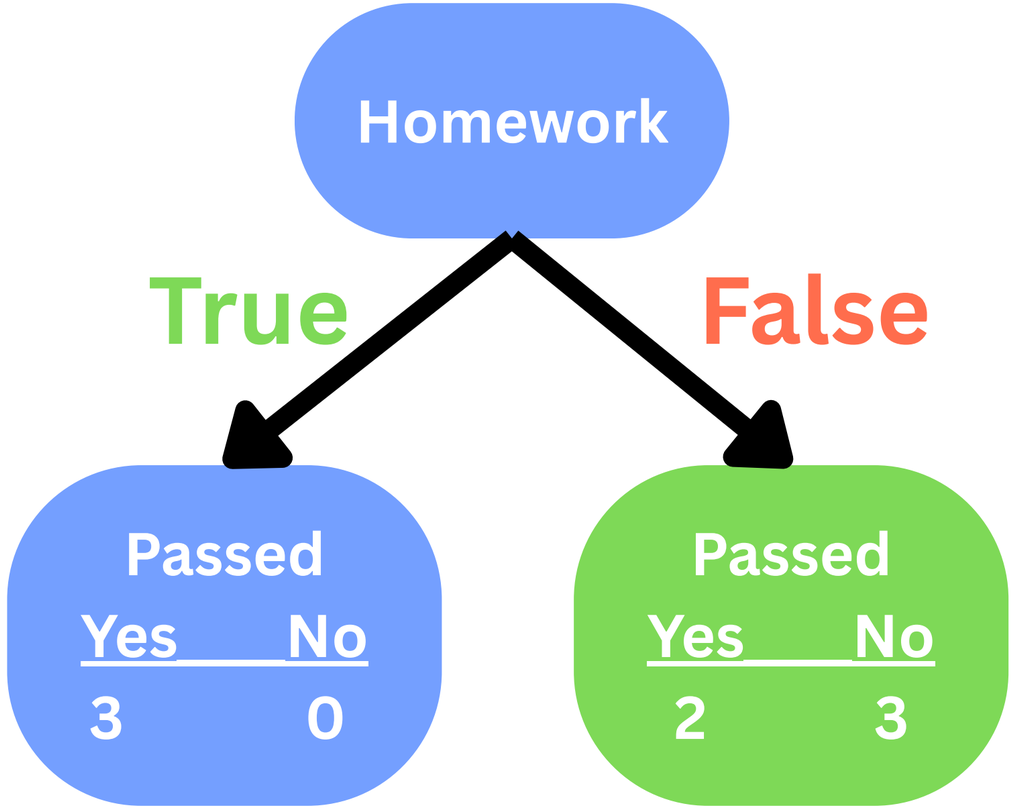
3/8

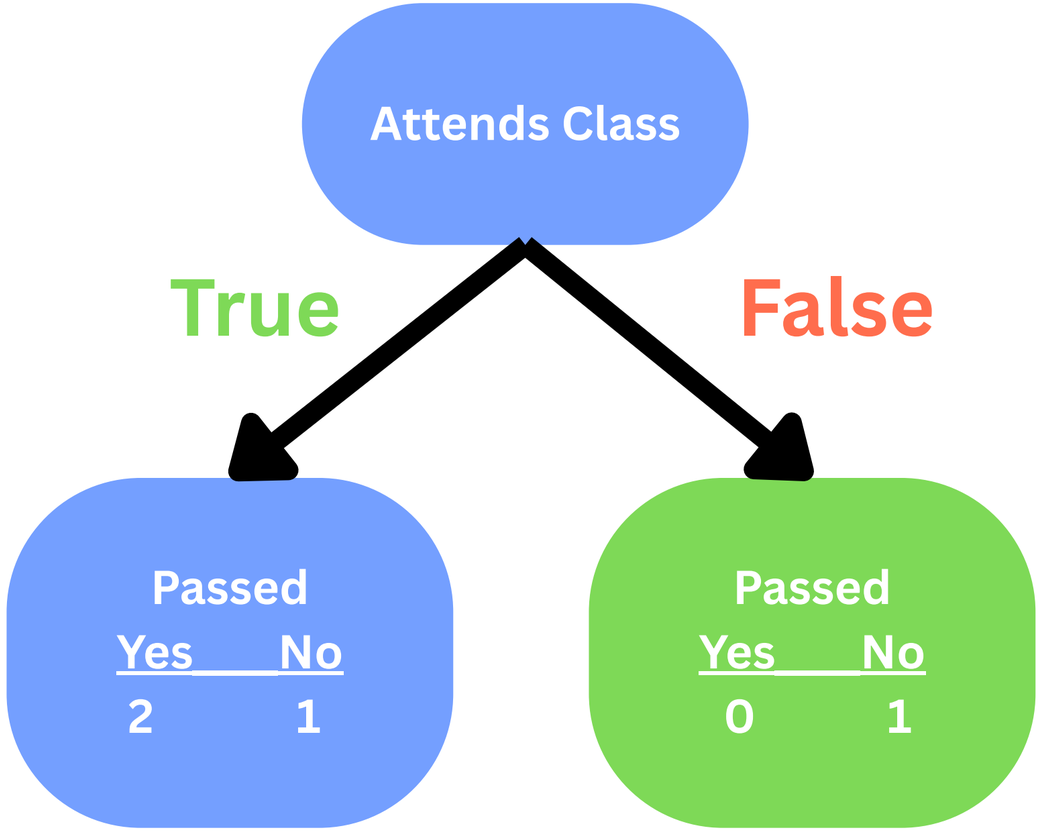
0.48

5/(3+5) =

5/8







0

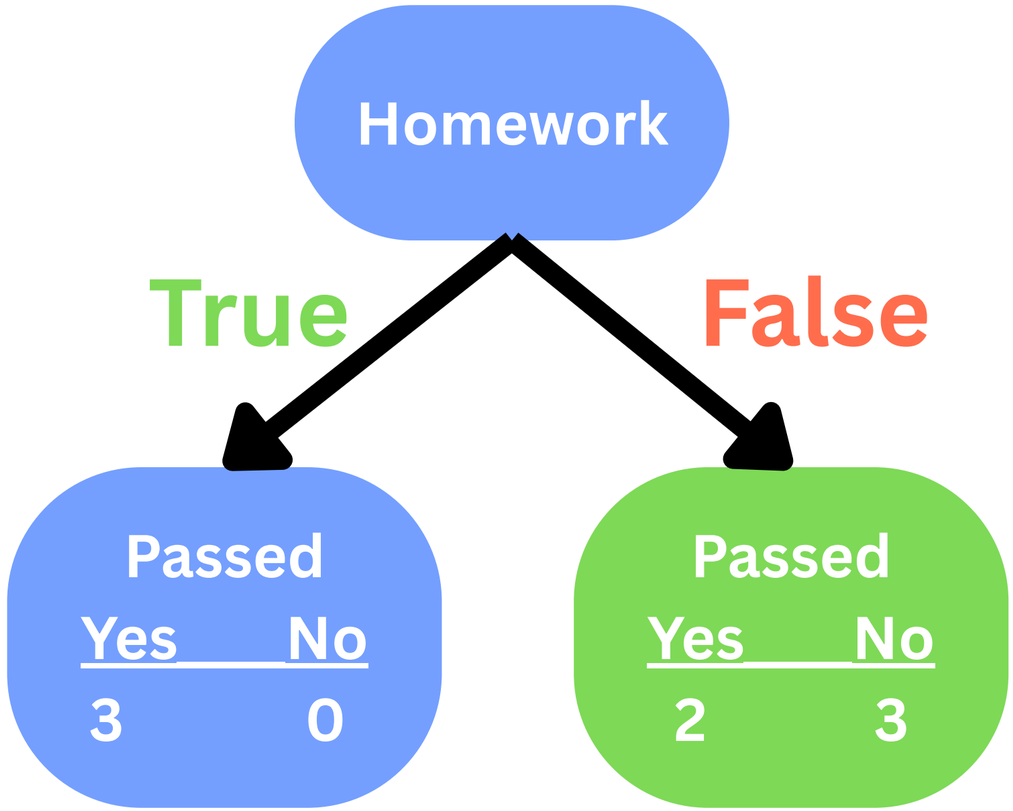
0.48

3/8*0 = 0

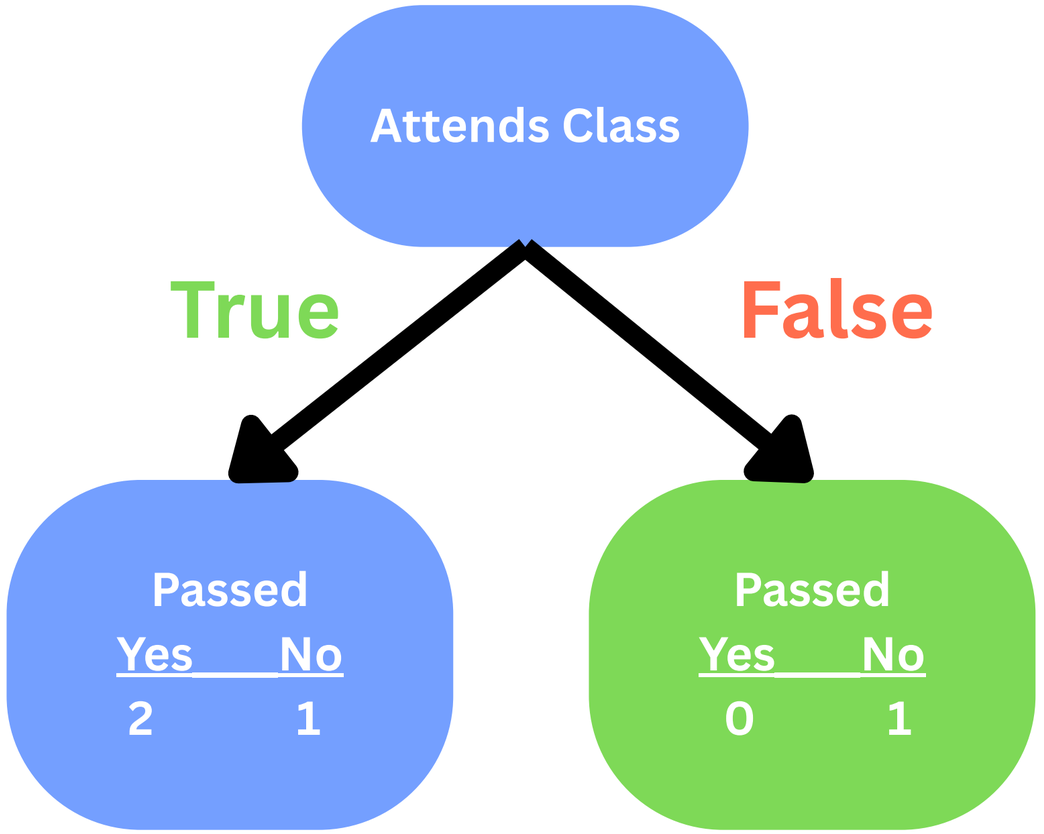
5/8*0.48 = 0.3

0 + 0.3 = 0.3 = total Gini Impurity

total = weight(Gini Impurity)



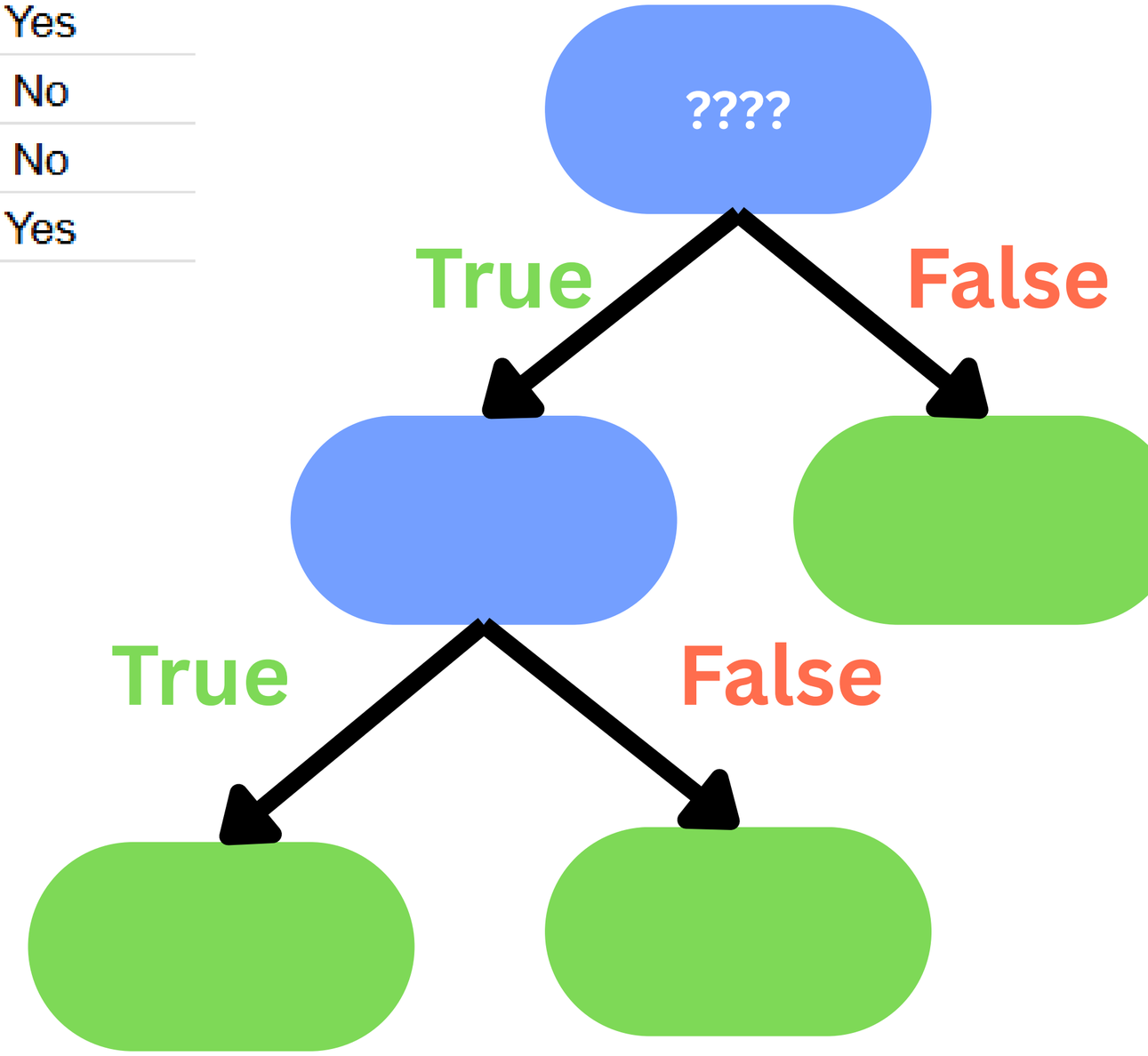
0.3 = Gini Impurity



0.333 = Gini Impurity

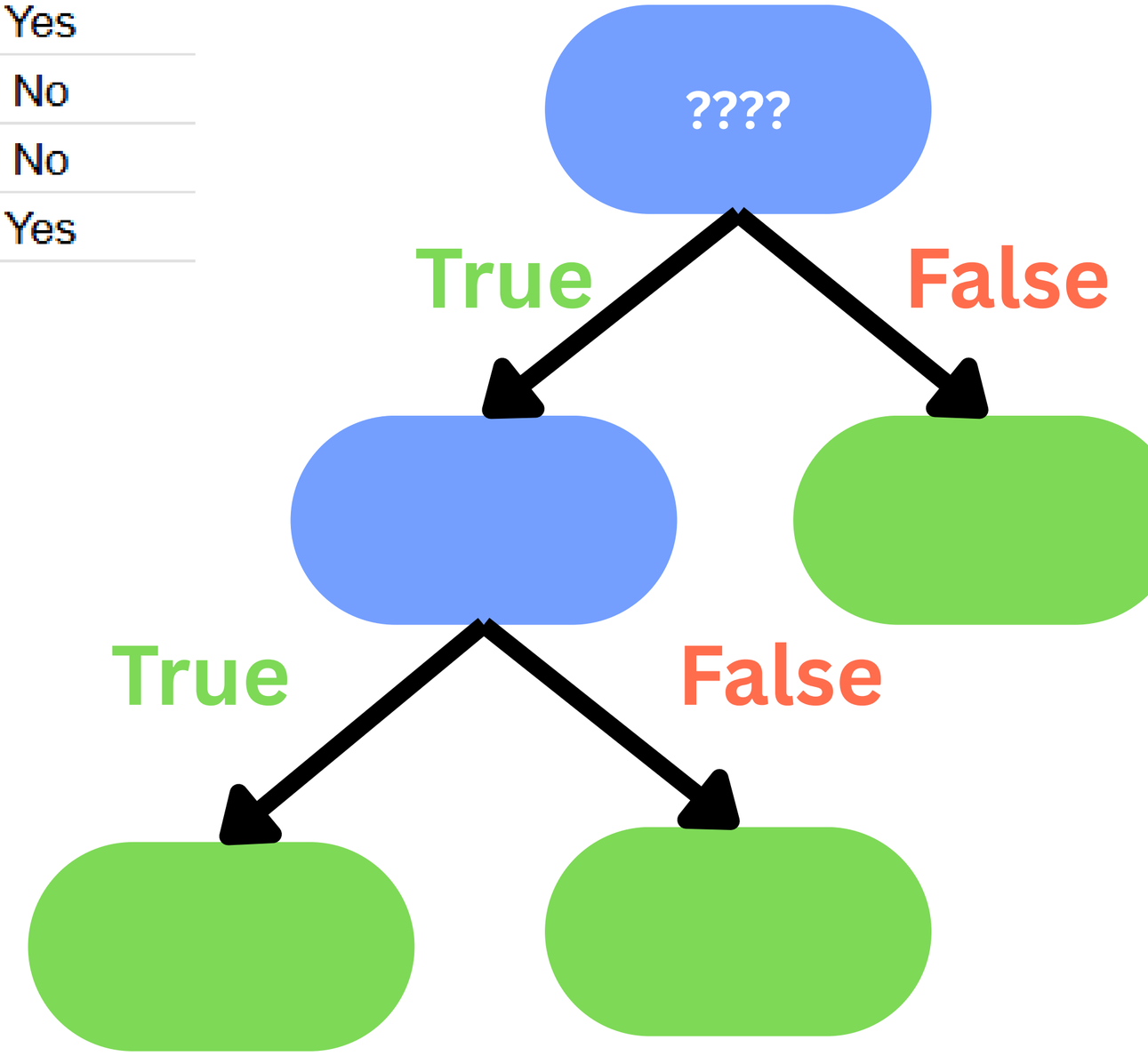
Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes

Now we need to do hours
Since hours is numeric we use another process



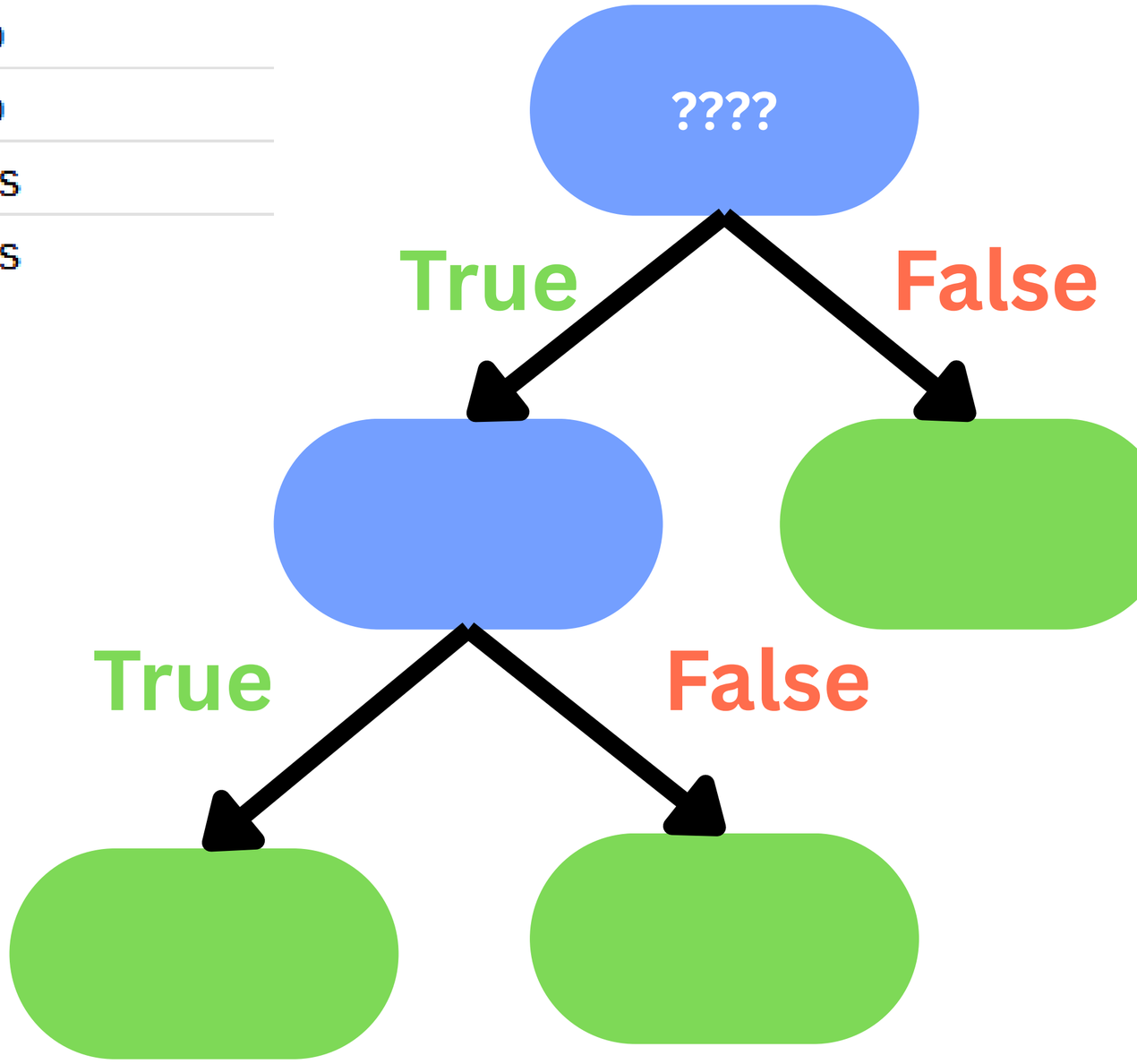
Attends Class	Homework	Study Hours	Passes?
Yes	Yes	15	Yes
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes

Now we need to do hours
Since hours is numeric we use another process



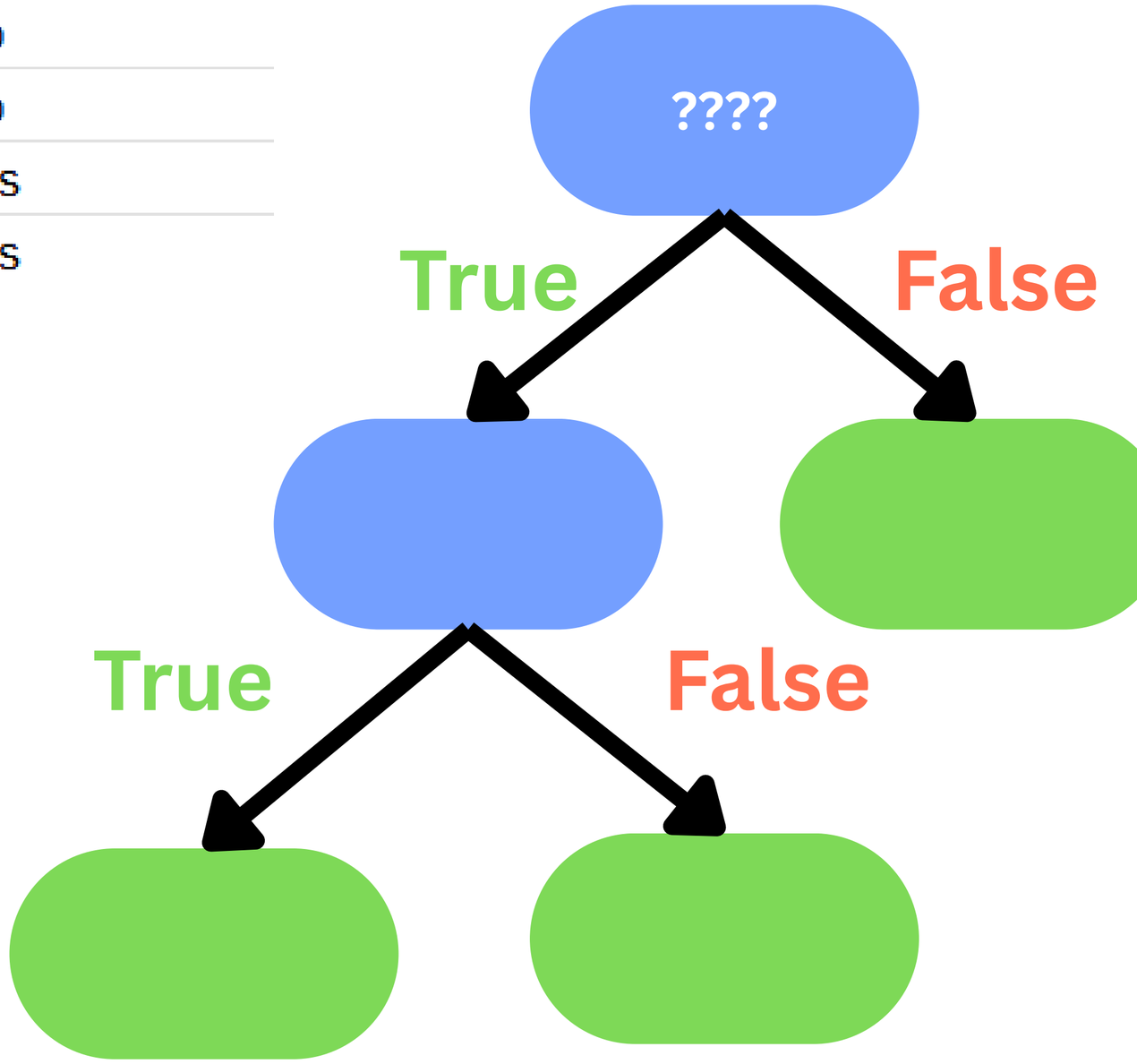
Attends Class	Homework	Study Hours	Passes?
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes
Yes	Yes	15	Yes

We first need to sort this data by study hours (low to high)



Attends Class	Homework	Study Hours		Passes?
No	Yes	3	2.5	No
Yes	No	6	8.0	No
Yes	Yes	10	12.5	Yes
Yes	Yes	15		Yes

Then calculate the average age for all adgacent values

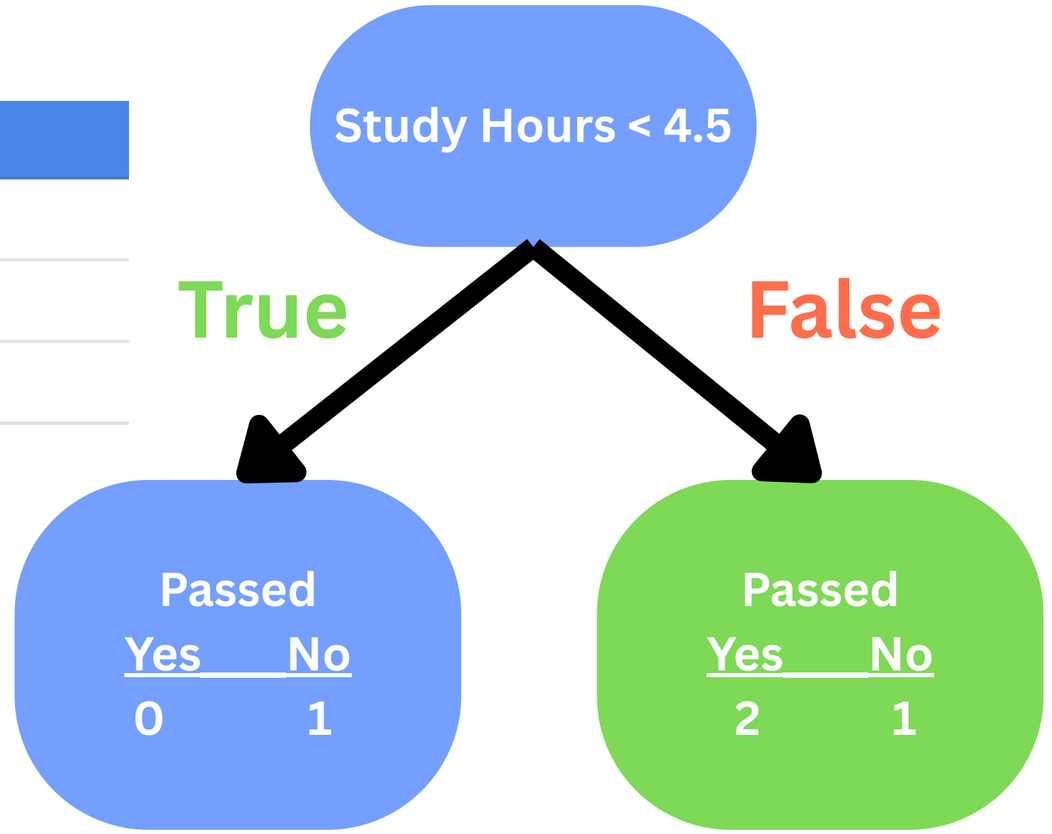


Attends Class	Homework	Study Hours	Passes?
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes
Yes	Yes	15	Yes

4.5

8.0

12.5



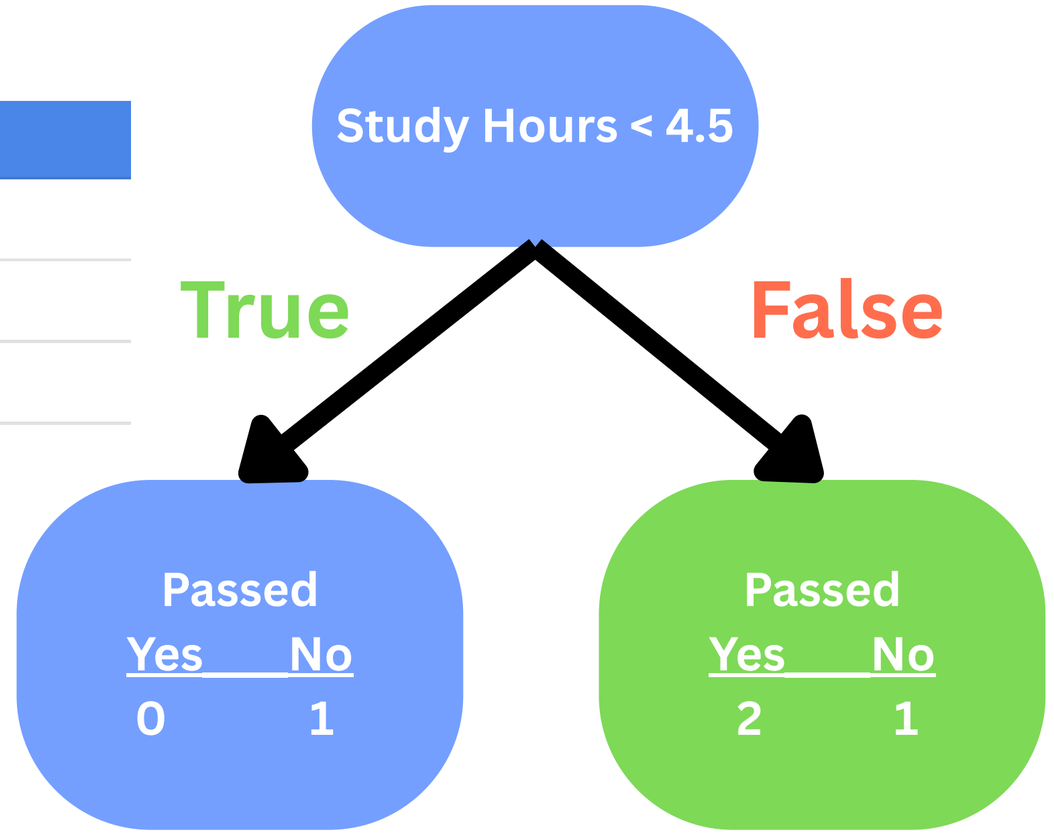
Now we calculate the Gini Impurity - Lets create our Tree

Attends Class	Homework	Study Hours	Passes?
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes
Yes	Yes	15	Yes

4.5

8.0

12.5



Now we do same steps to get Gini Impurity

Gini Impurity for a leaf = $1 - (P(Yes))^2 - (P(No))^2$

$$1 - (0/(0+1))^2 - (1/(0+1))^2$$

Since this is pure we get a Gini Impurity of 0

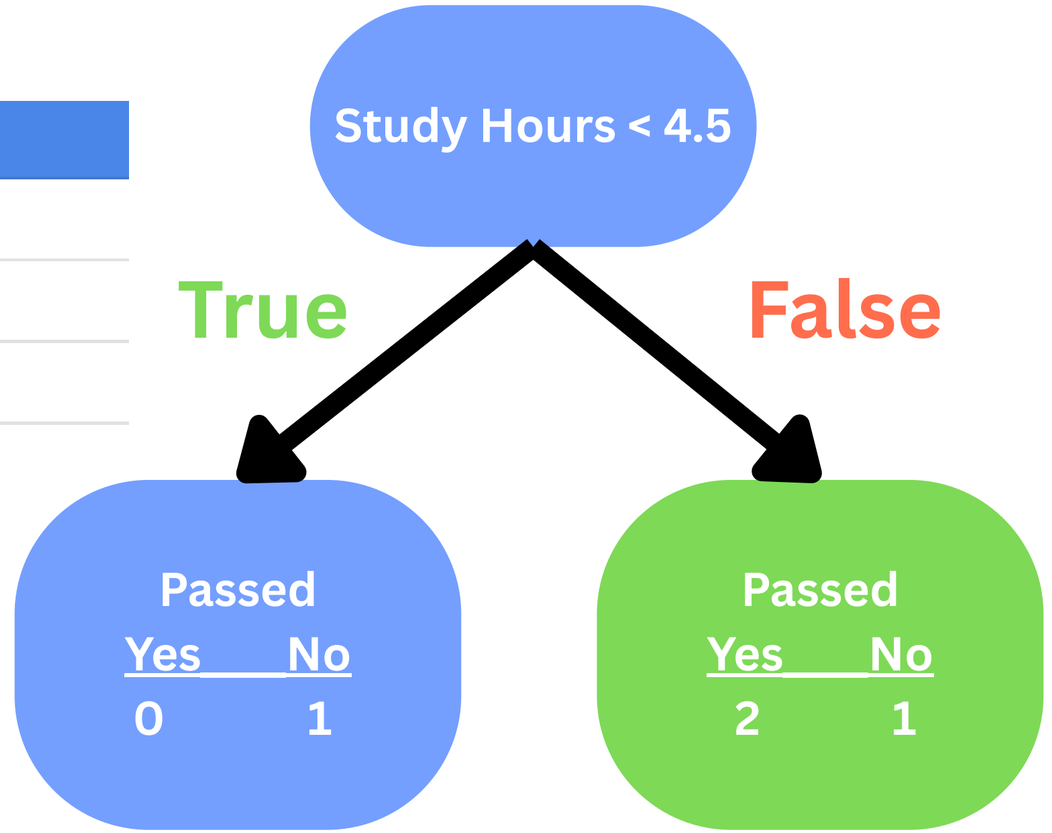
Attends Class	Homework	Study Hours	Passes?
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes
Yes	Yes	15	Yes

4.5

8.0

12.5

Lets get the total



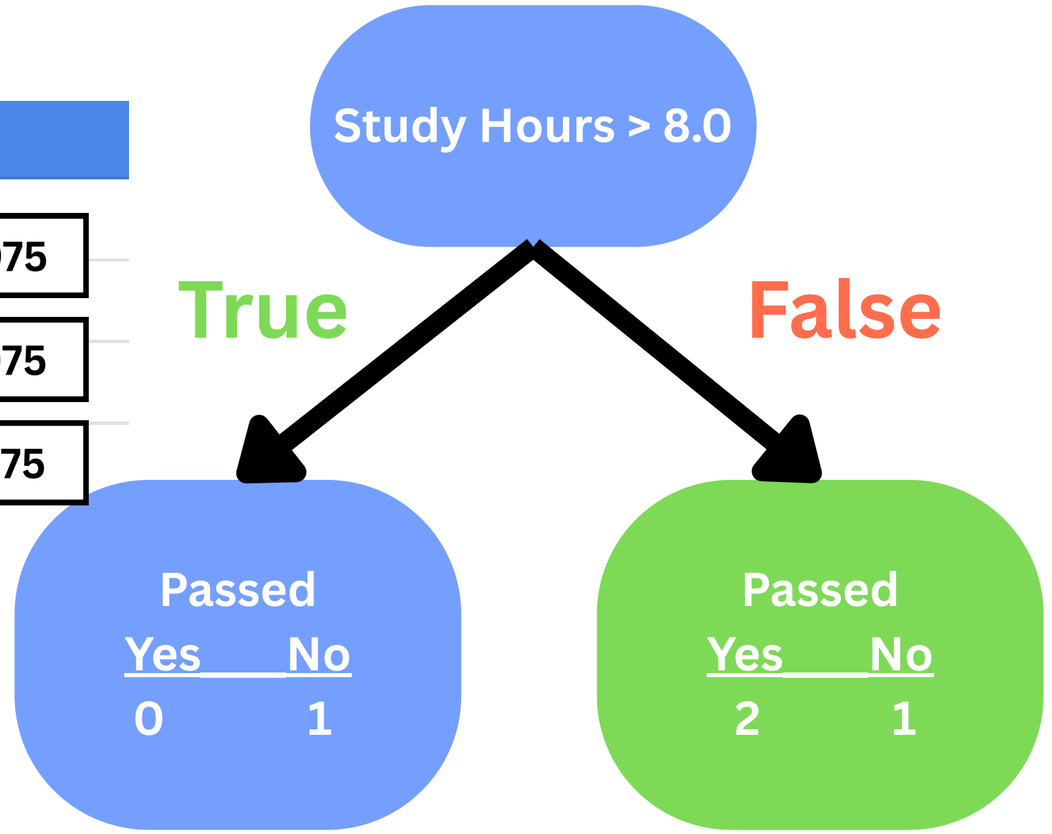
Total Gini Impurity = 1/4*0 + 3/4*0.333 = 0.24975

Attends Class	Homework	Study Hours	Passes?
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes
Yes	Yes	15	Yes

4.5 → 0.24975

8.0 → 0.14975

12.5 → 0.21975



Now repeat for all - I just made up the rest of the value for demonstration

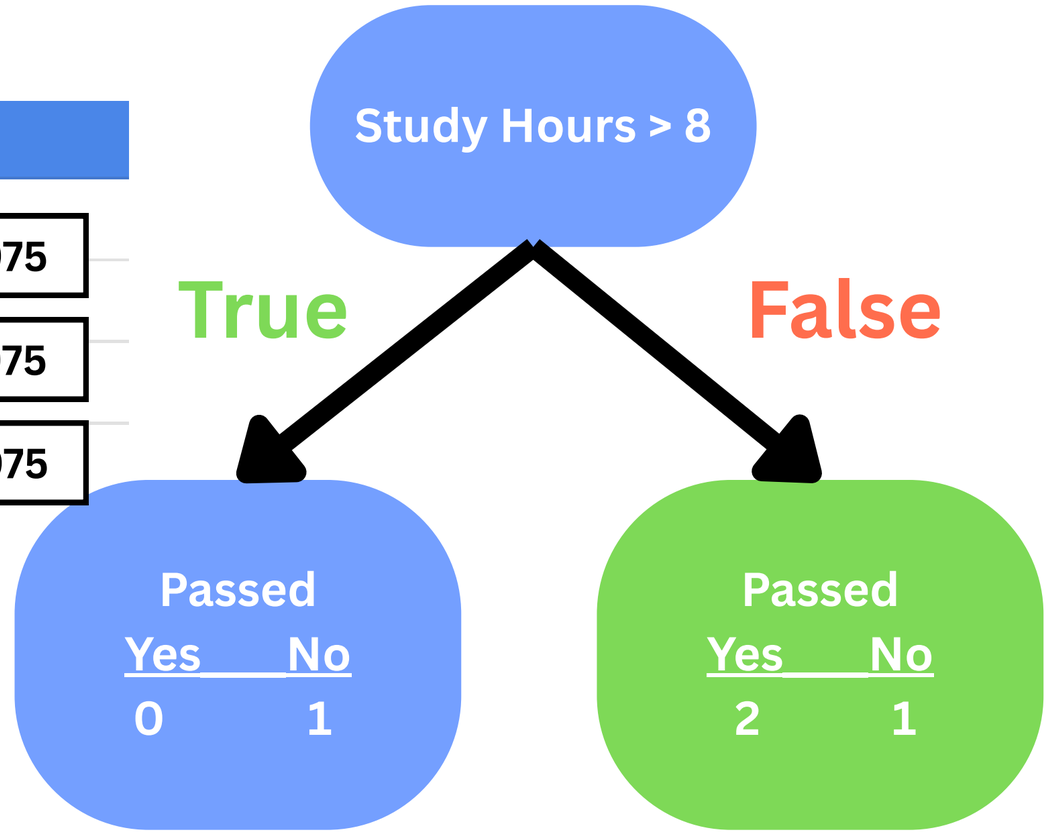
0.14975 = Gini Impurity

Attends Class	Homework	Study Hours	Passes?
No	Yes	3	No
Yes	No	6	No
Yes	Yes	10	Yes
Yes	Yes	15	Yes

4.5 → 0.24975

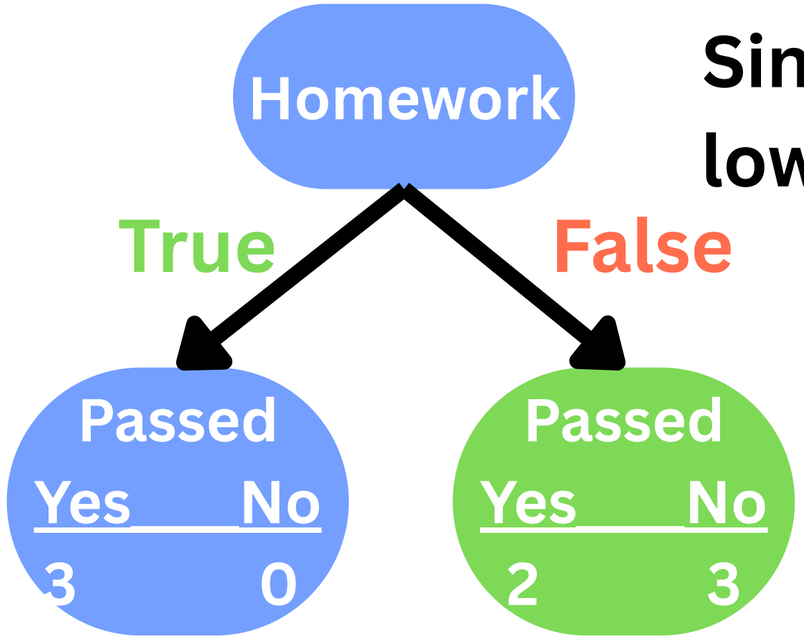
8.0 → 0.14975

12.5 → 0.34975



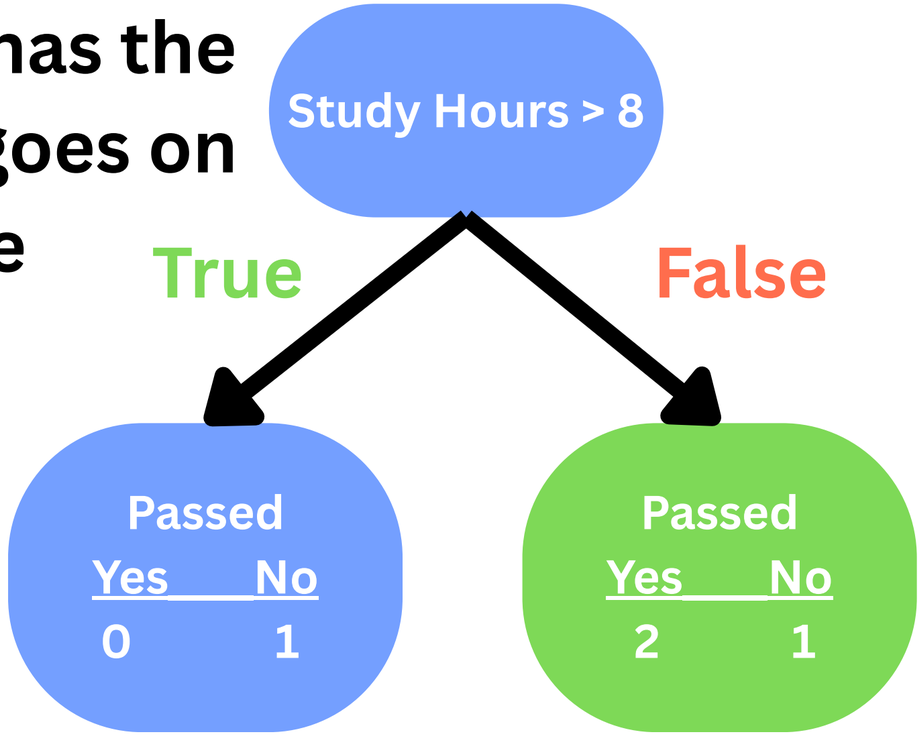
We now take the lowest impurity score and use the value corresponding to it so in this case 8 here and set it on our final tree

0.14975 = Gini Impurity

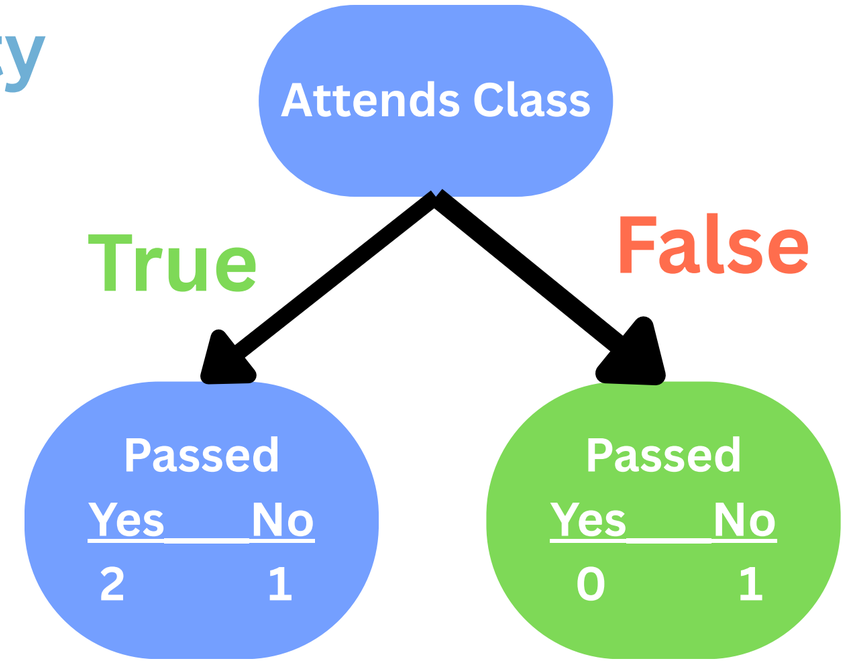


0.3 = Gini Impurity

Since now study hours has the lowest Gini Impurity it goes on the top of the tree

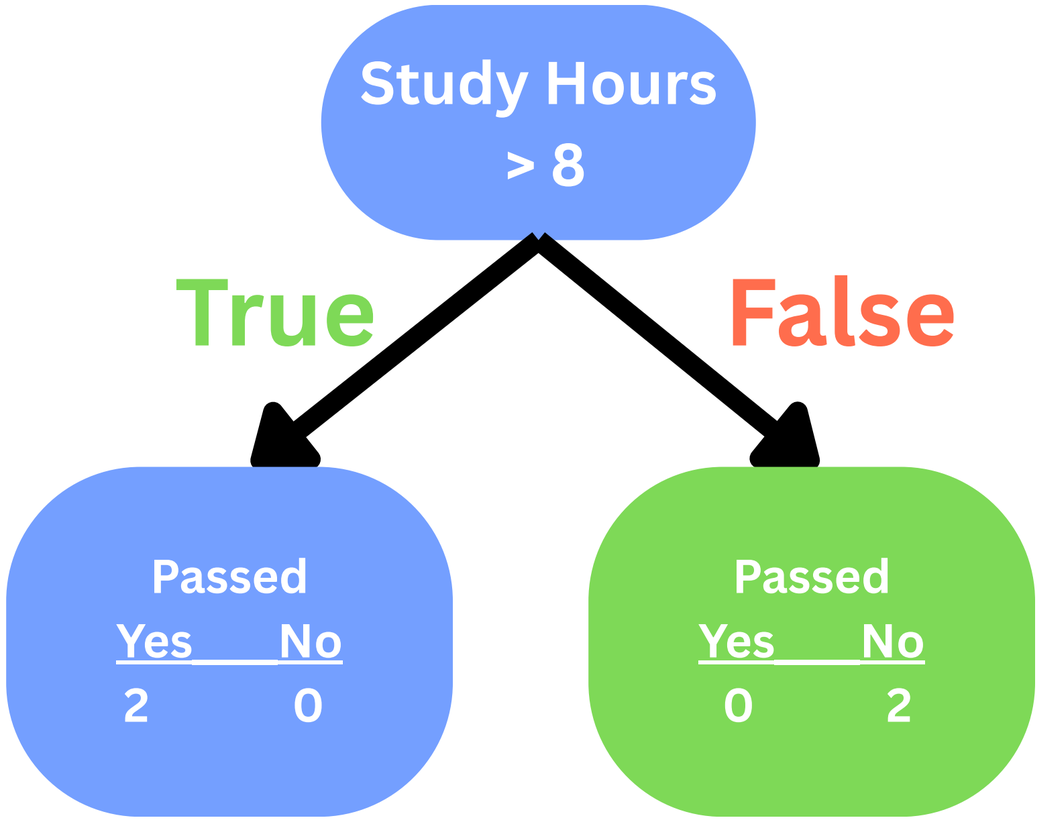


0.14975 = Gini Impurity

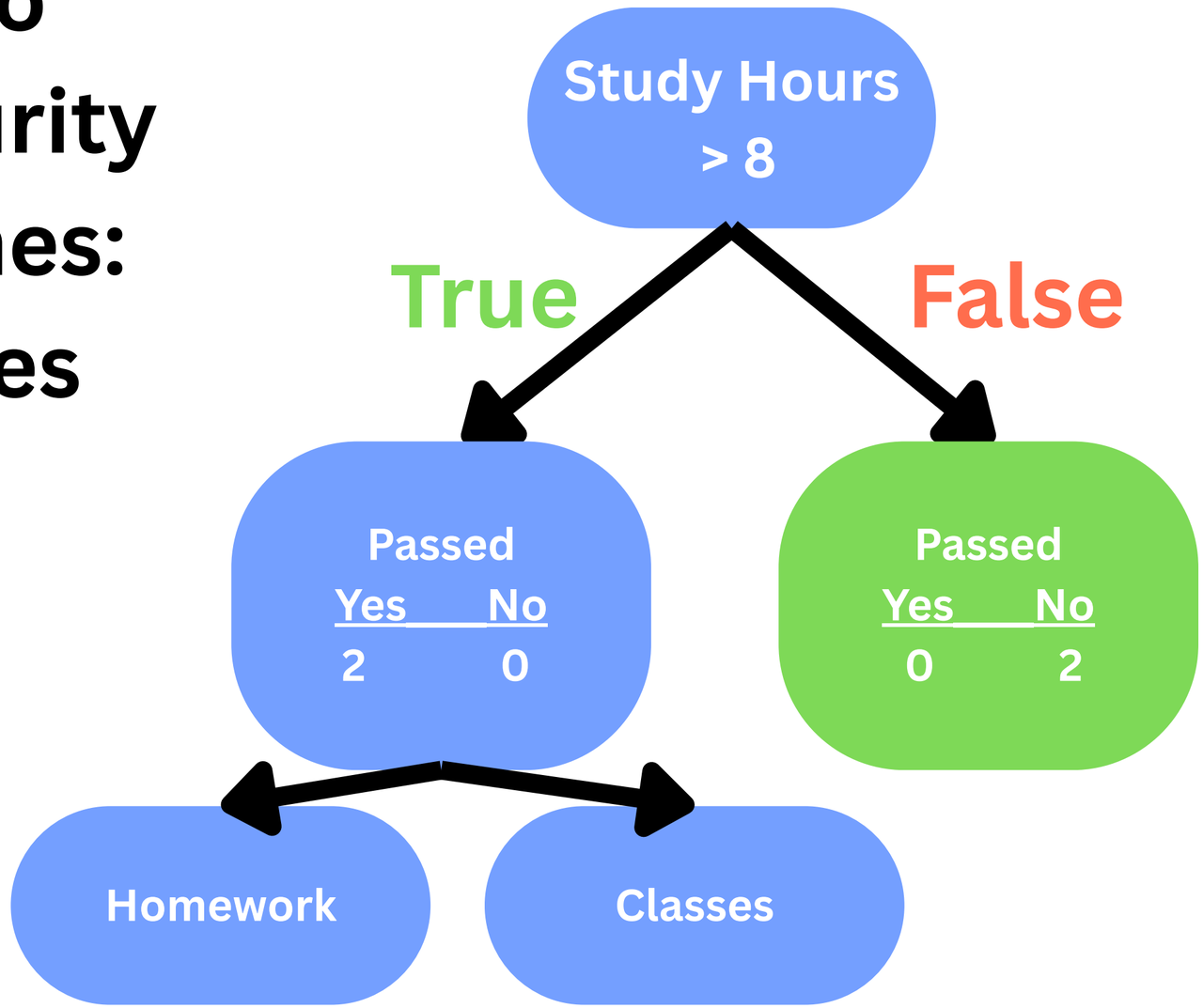


0.333 = Gini Impurity

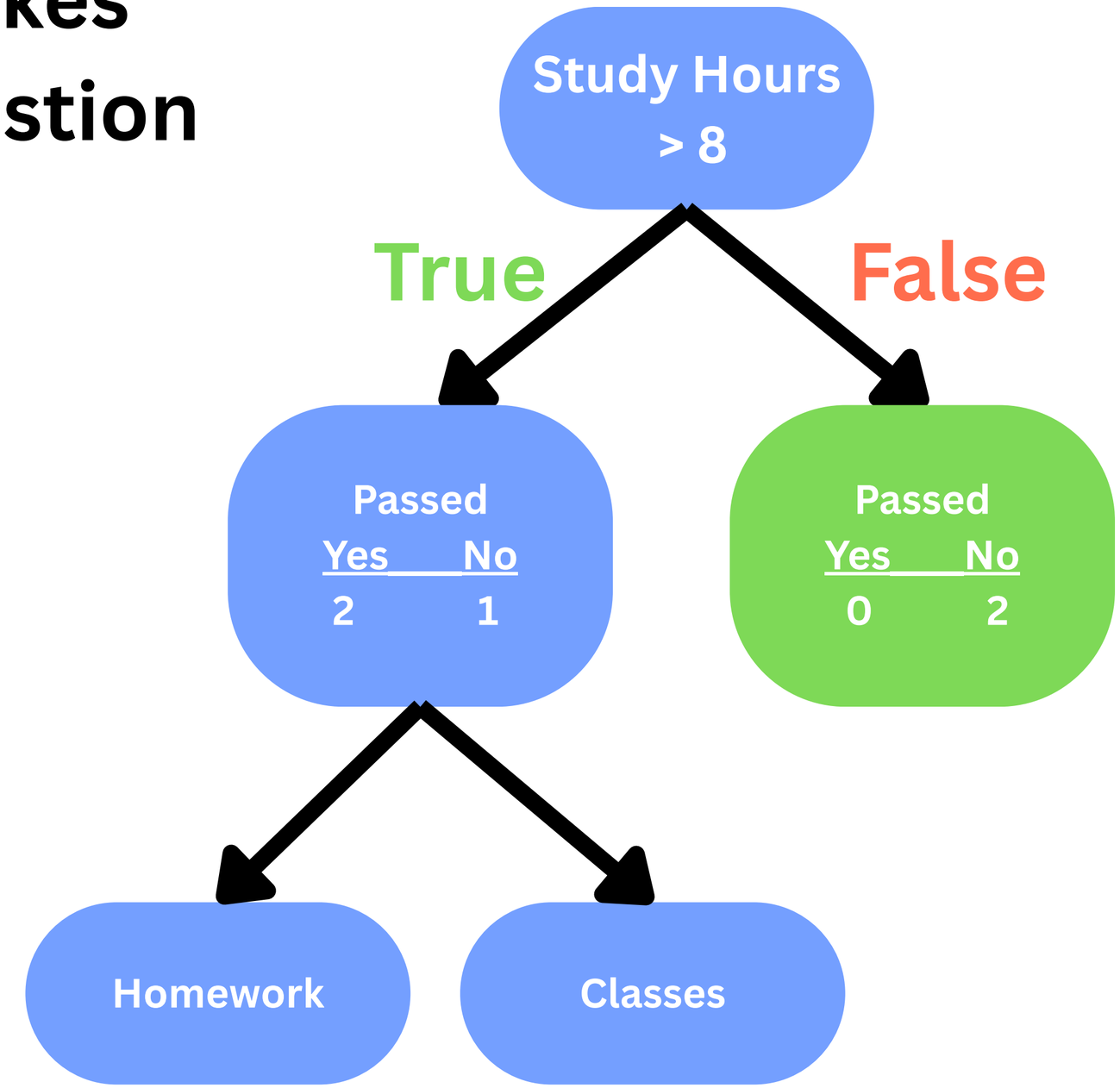
Now that study hours is up we need to figure out our next question or branch



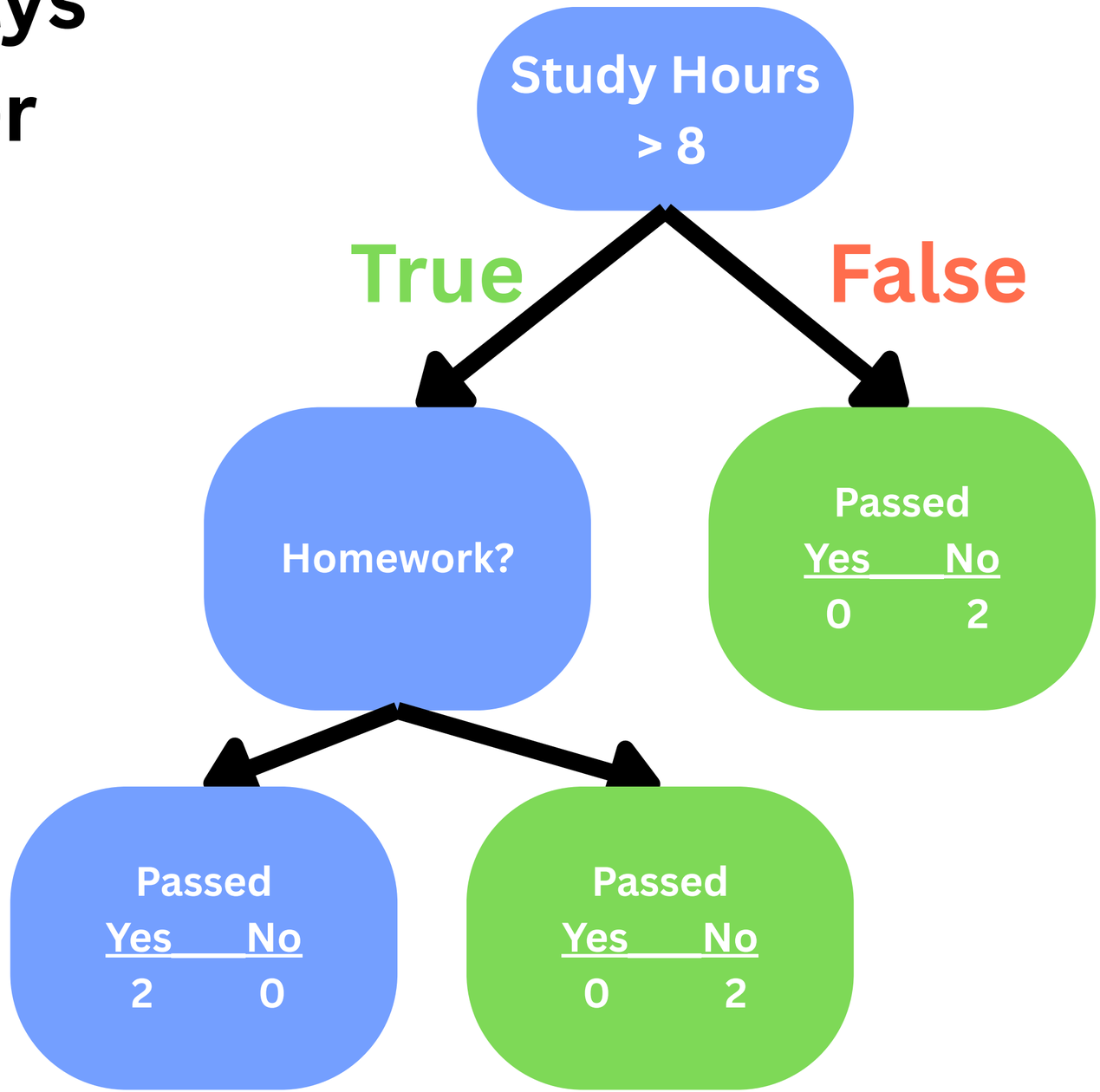
To do that we need to calculate the Gini Impurity again for the 2 outcomes: Homework and Classes



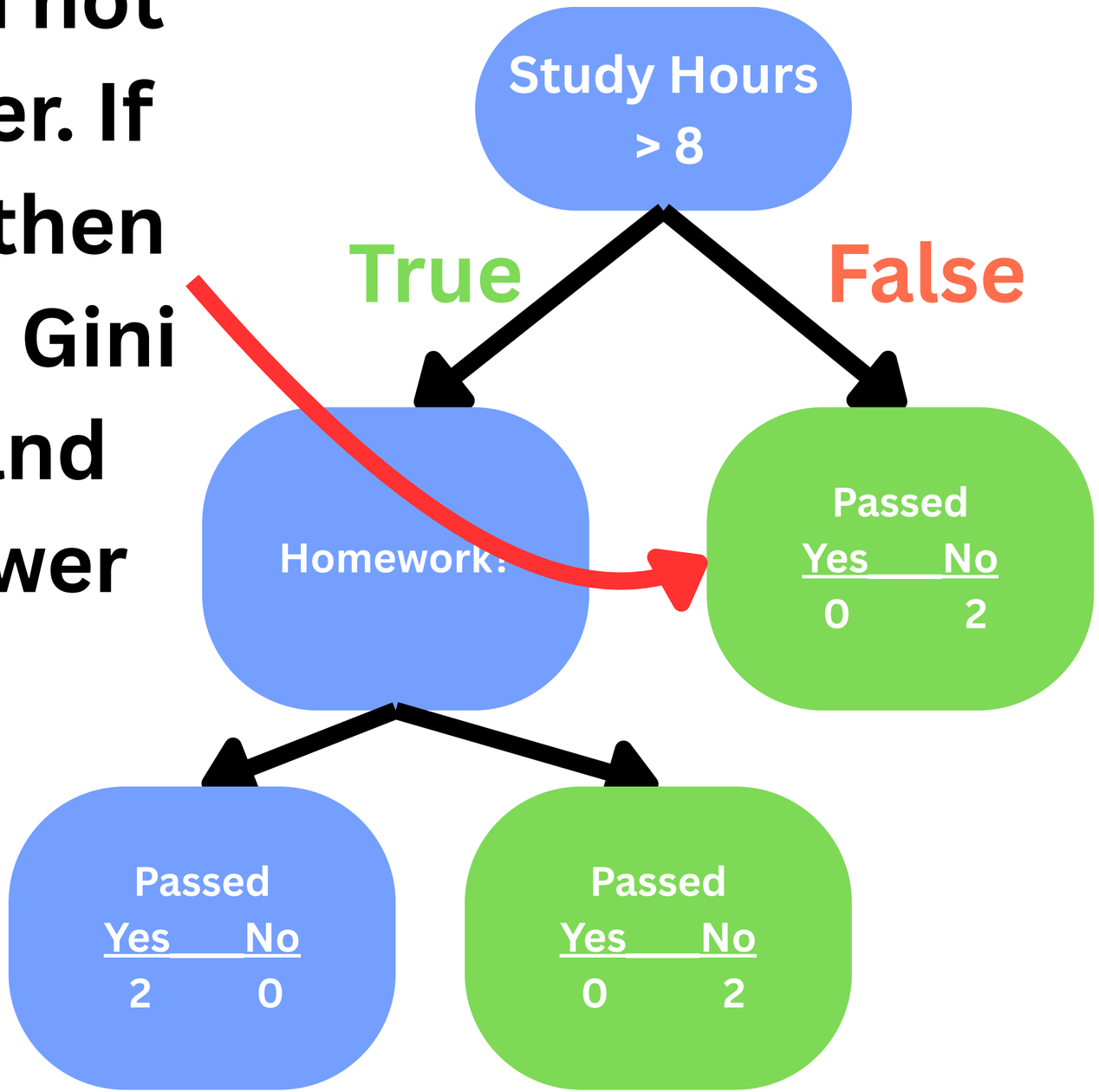
Which ever is lower takes that position as the question



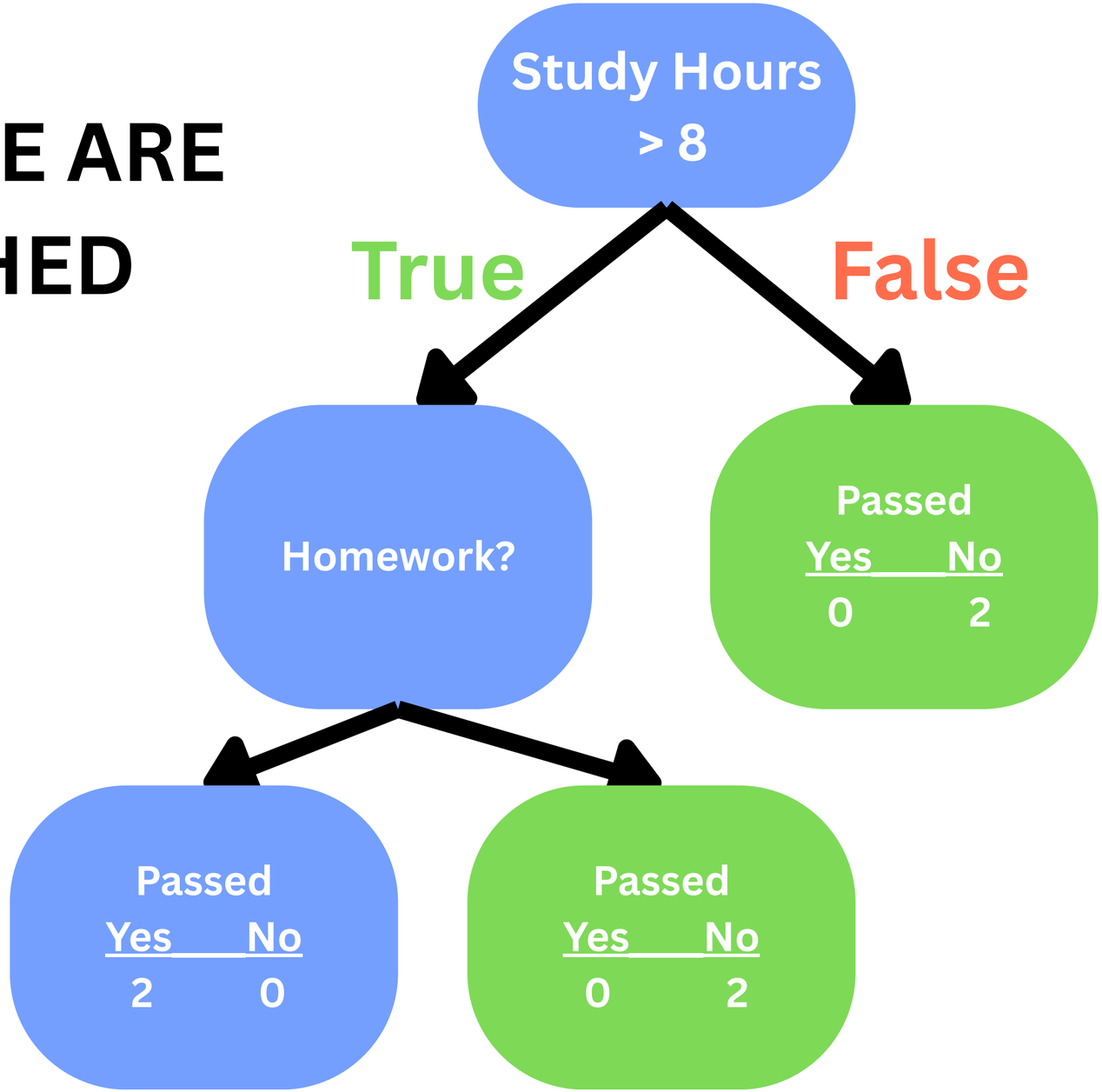
In this example lets says Homework takes over



Since this is pure we can not break it down any further. If for example it was pure then we need to calculate the Gini Impurity of the cases and decide which case is lower



TA-DA WE ARE FINISHED



QUESTIONS AND FINAL THOUGHTS!

THANK YOU!