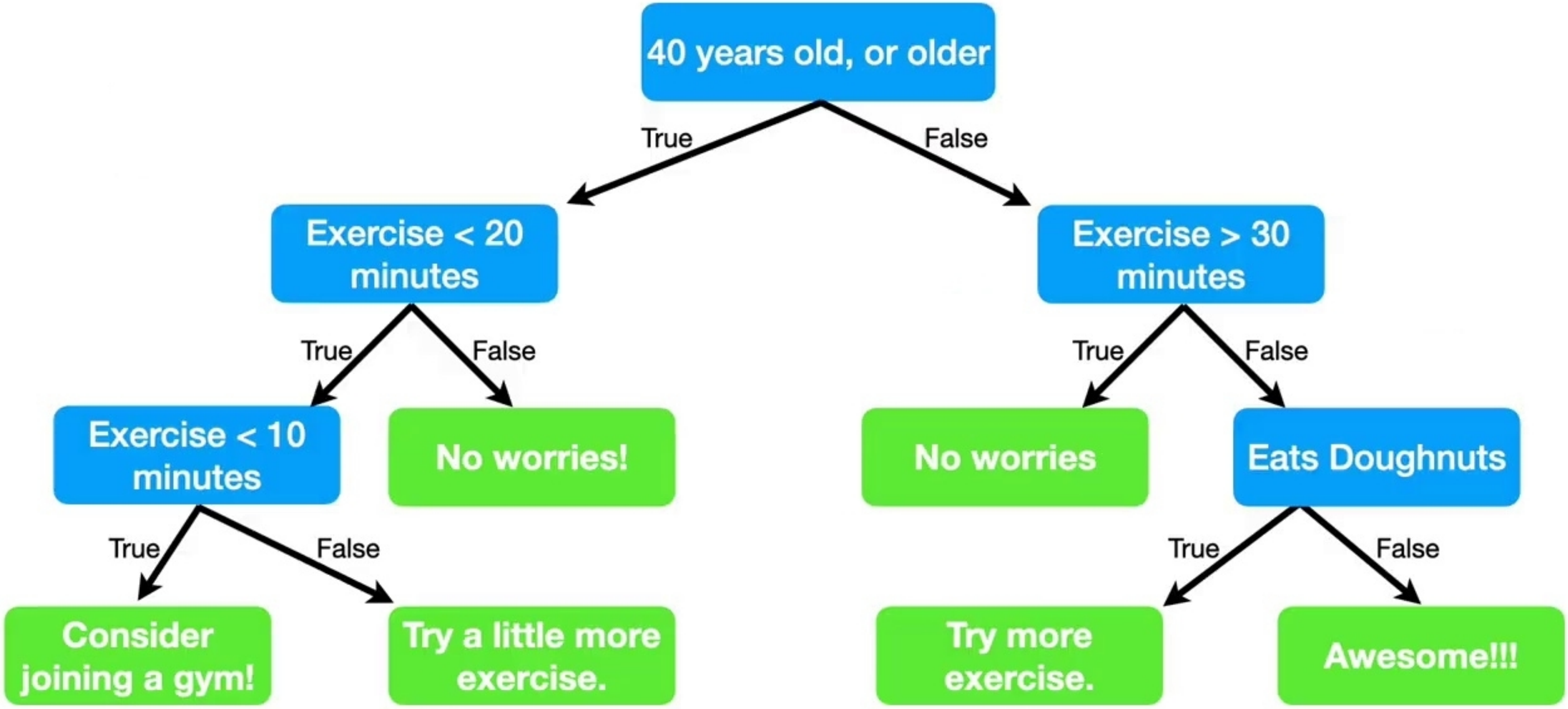
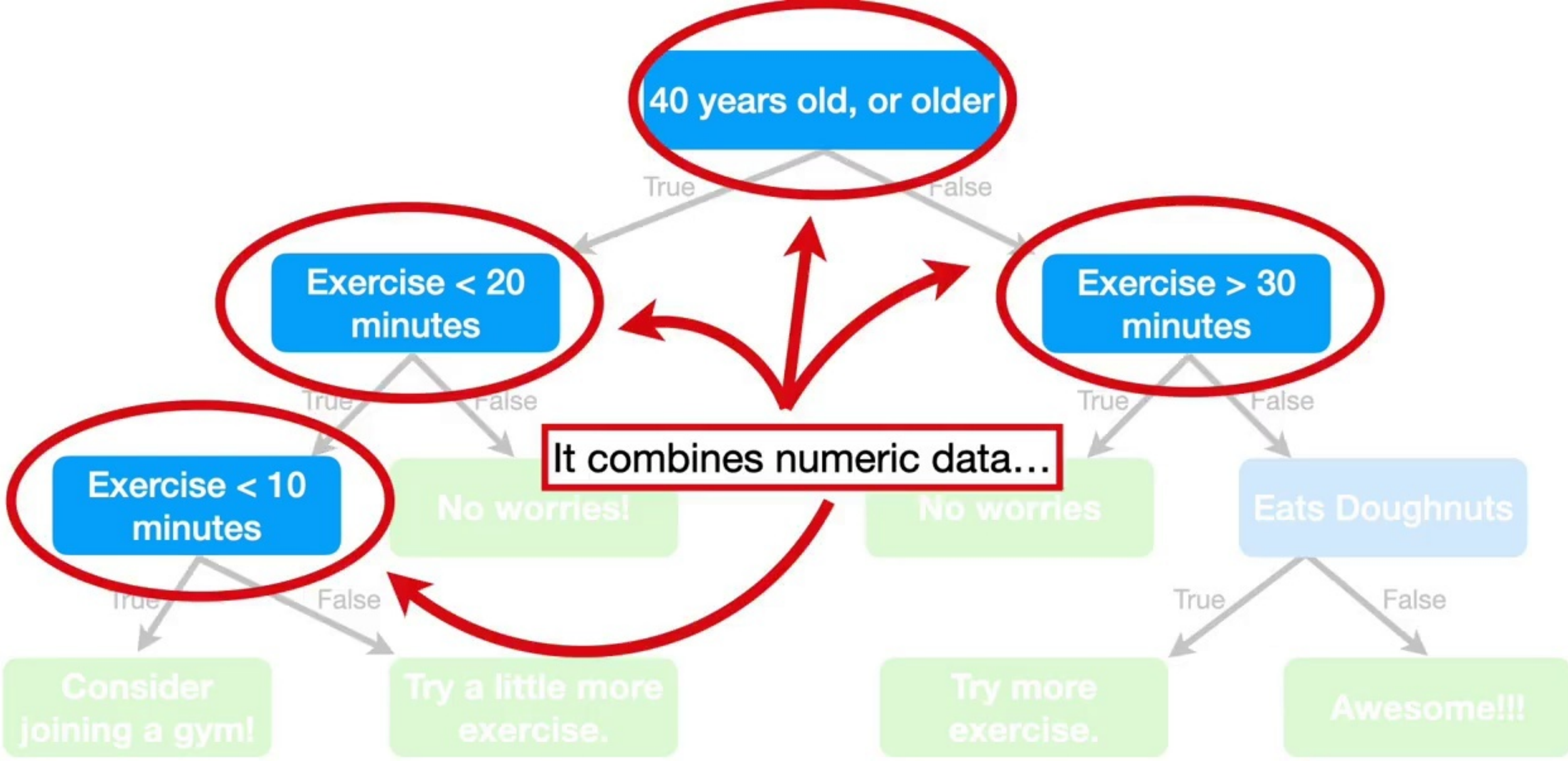
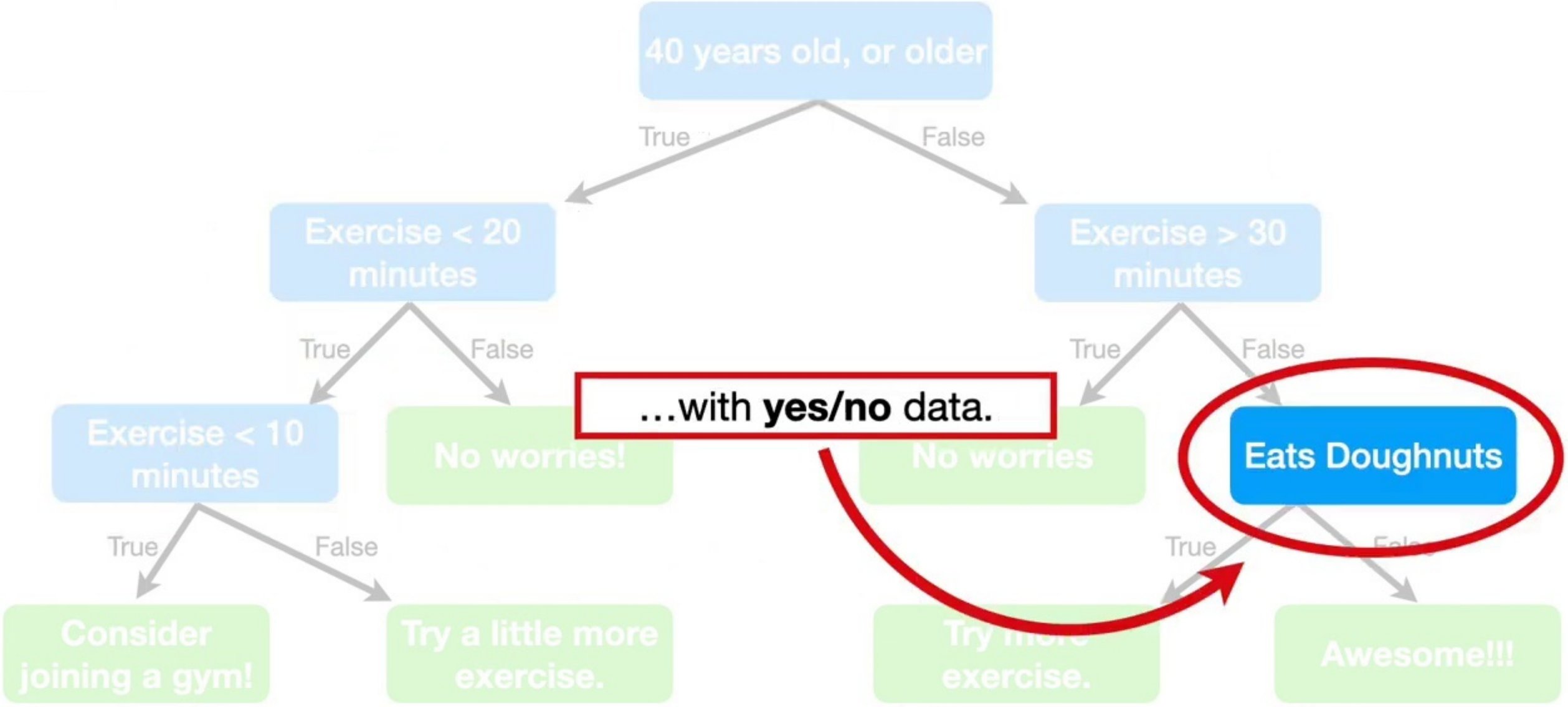
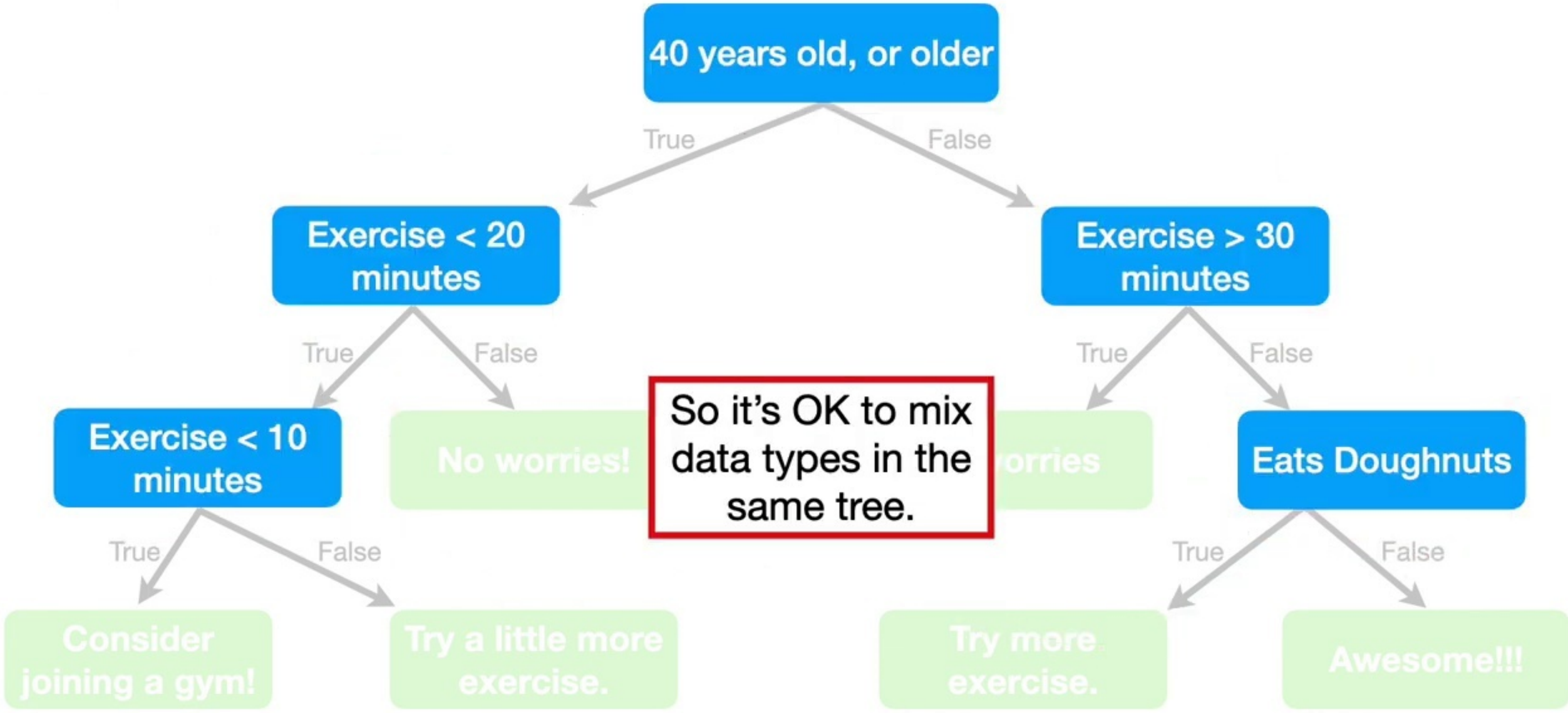
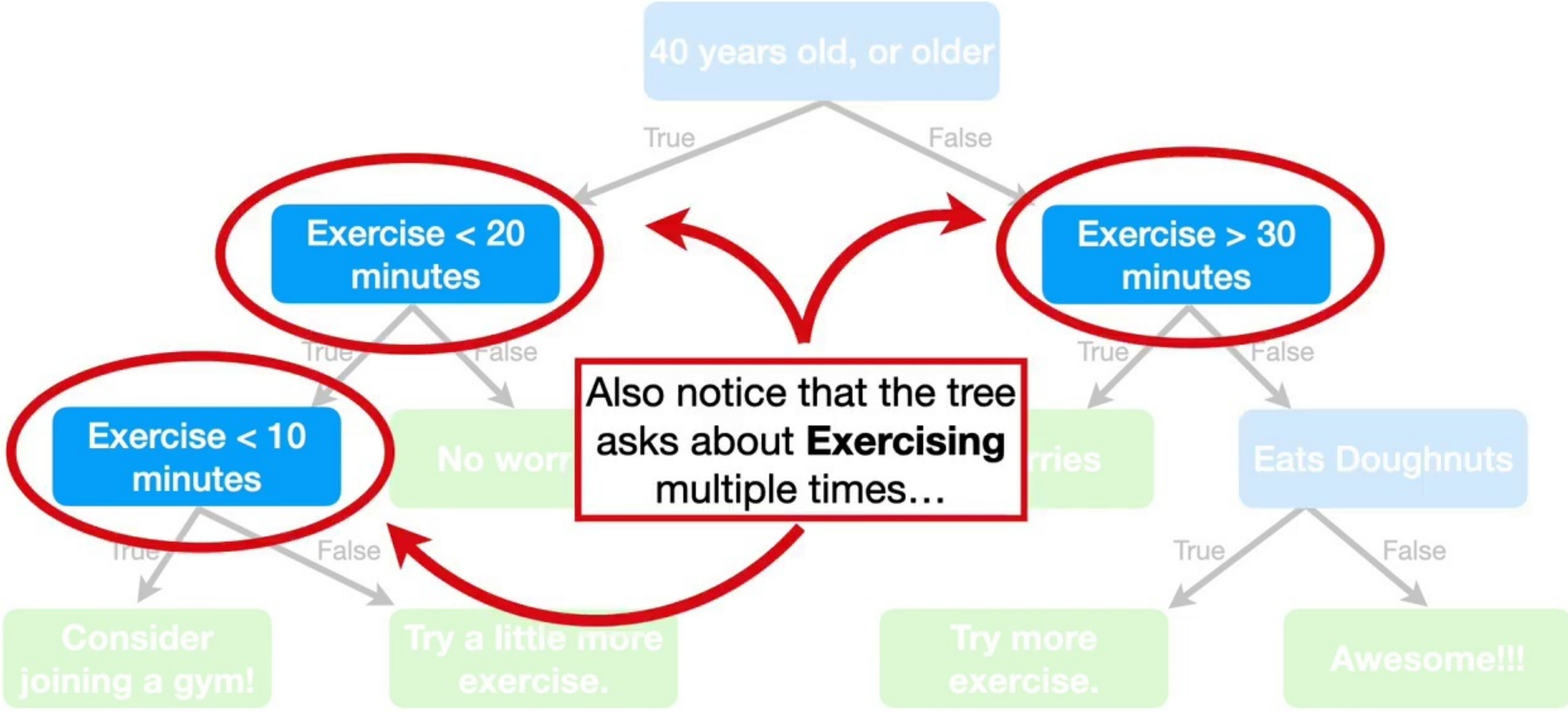
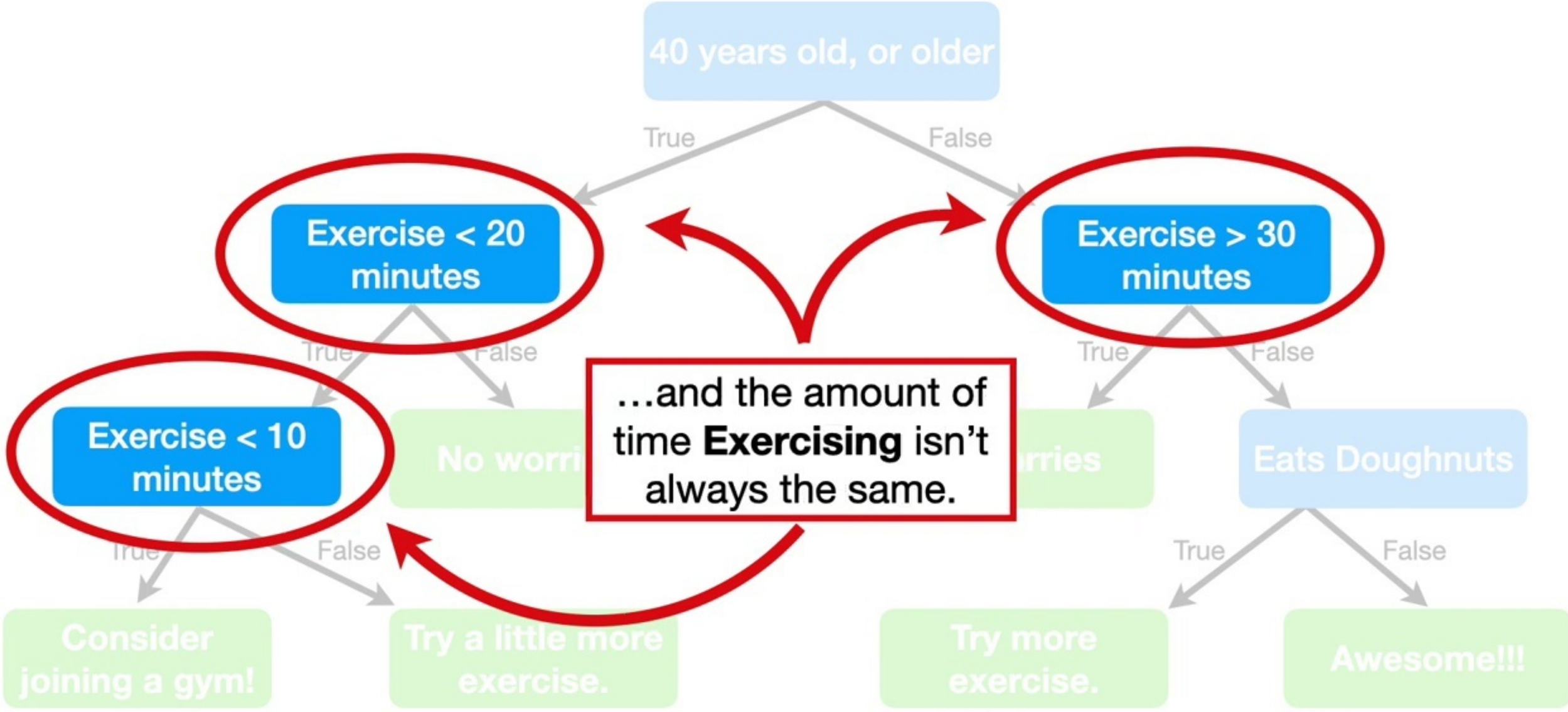

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

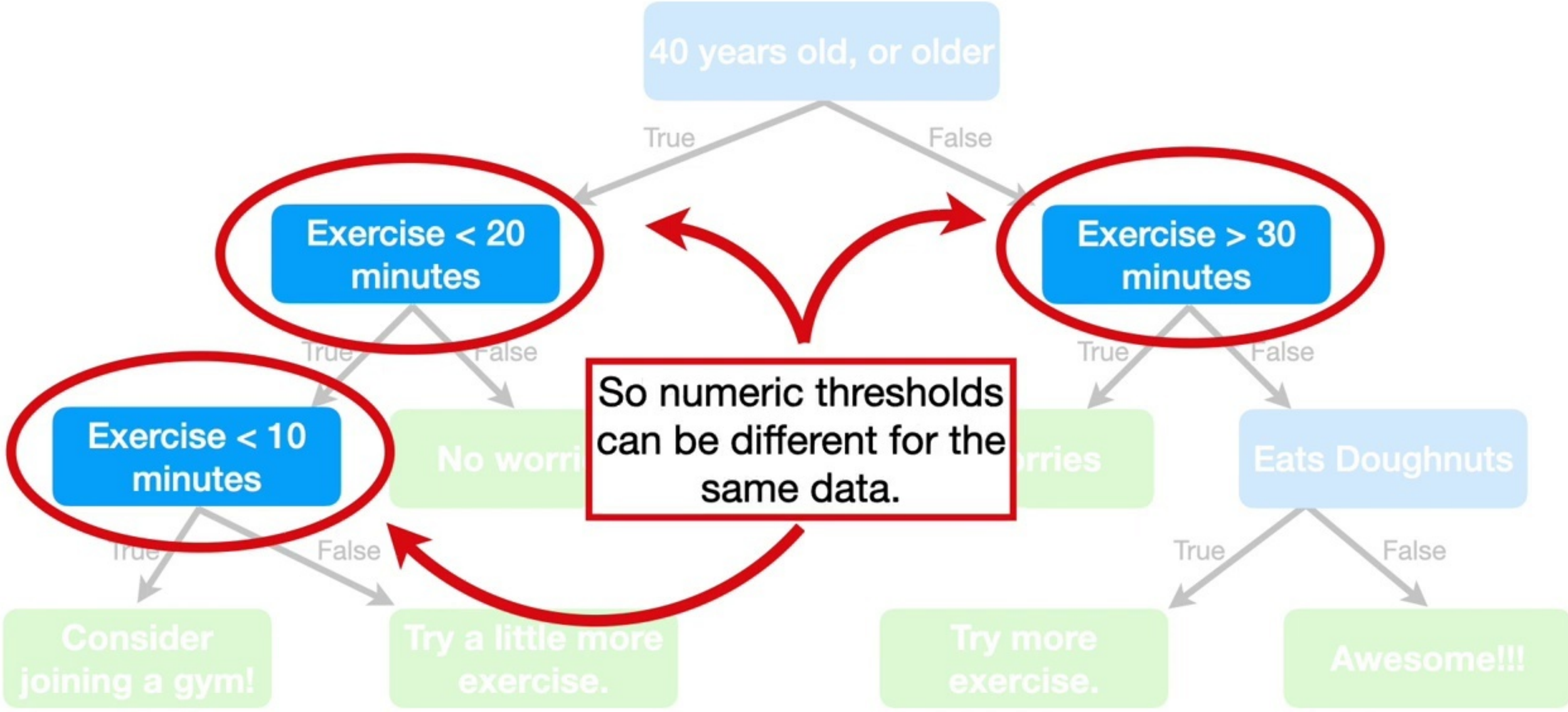

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
It combines numeric data…
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

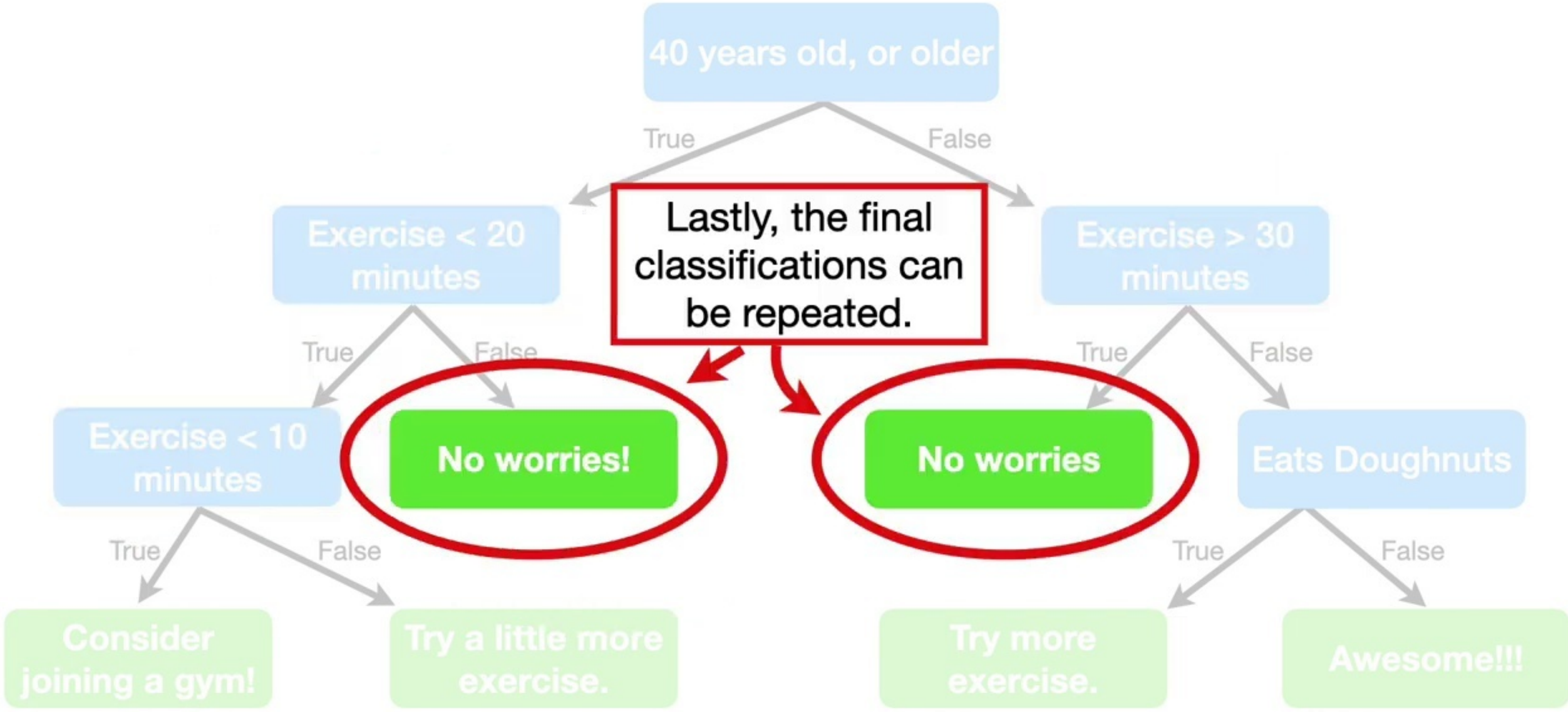

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
…with yes/no data.
True
False
True
Consider joining a gym!
Try a little more exercise.
Awesome!!!

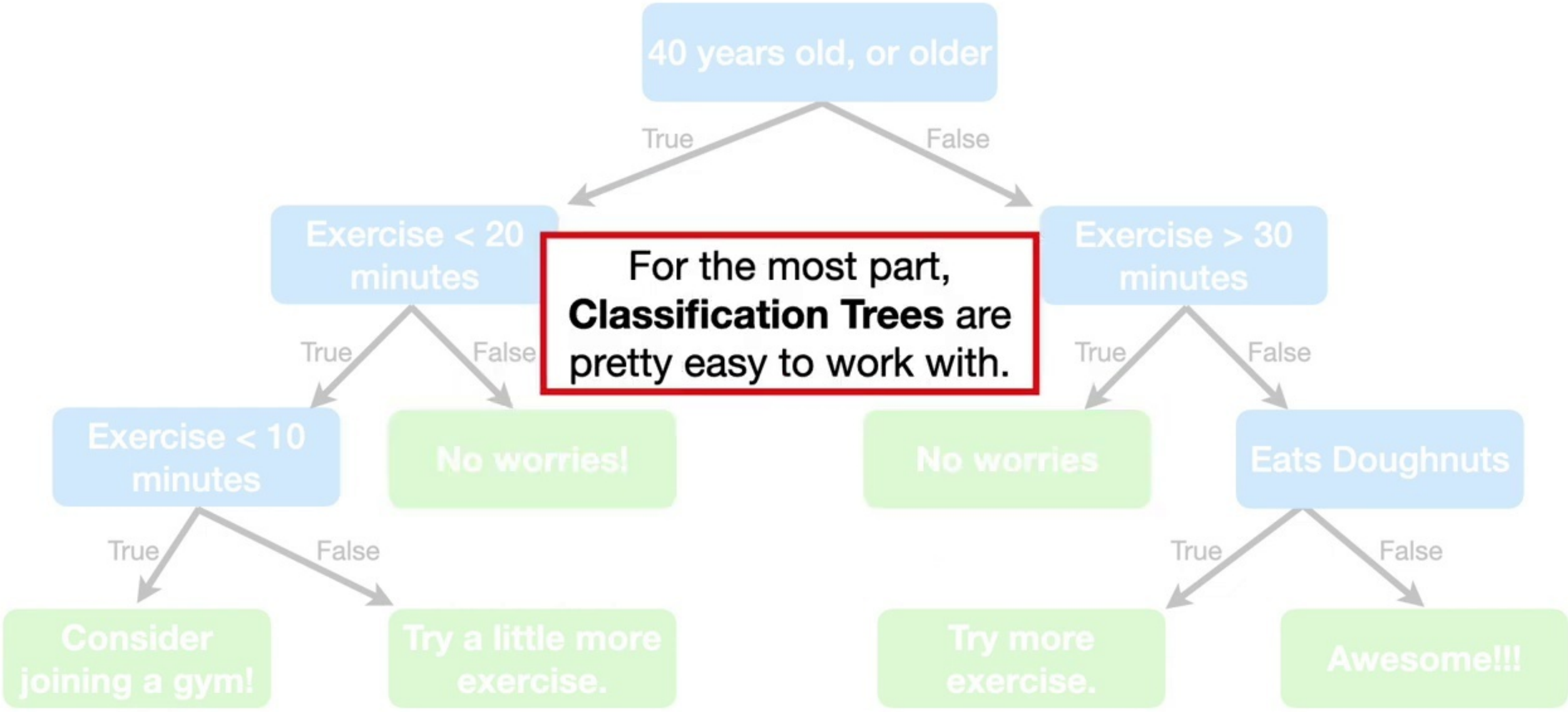

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
So it's OK to mix data types in the same tree.
orries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

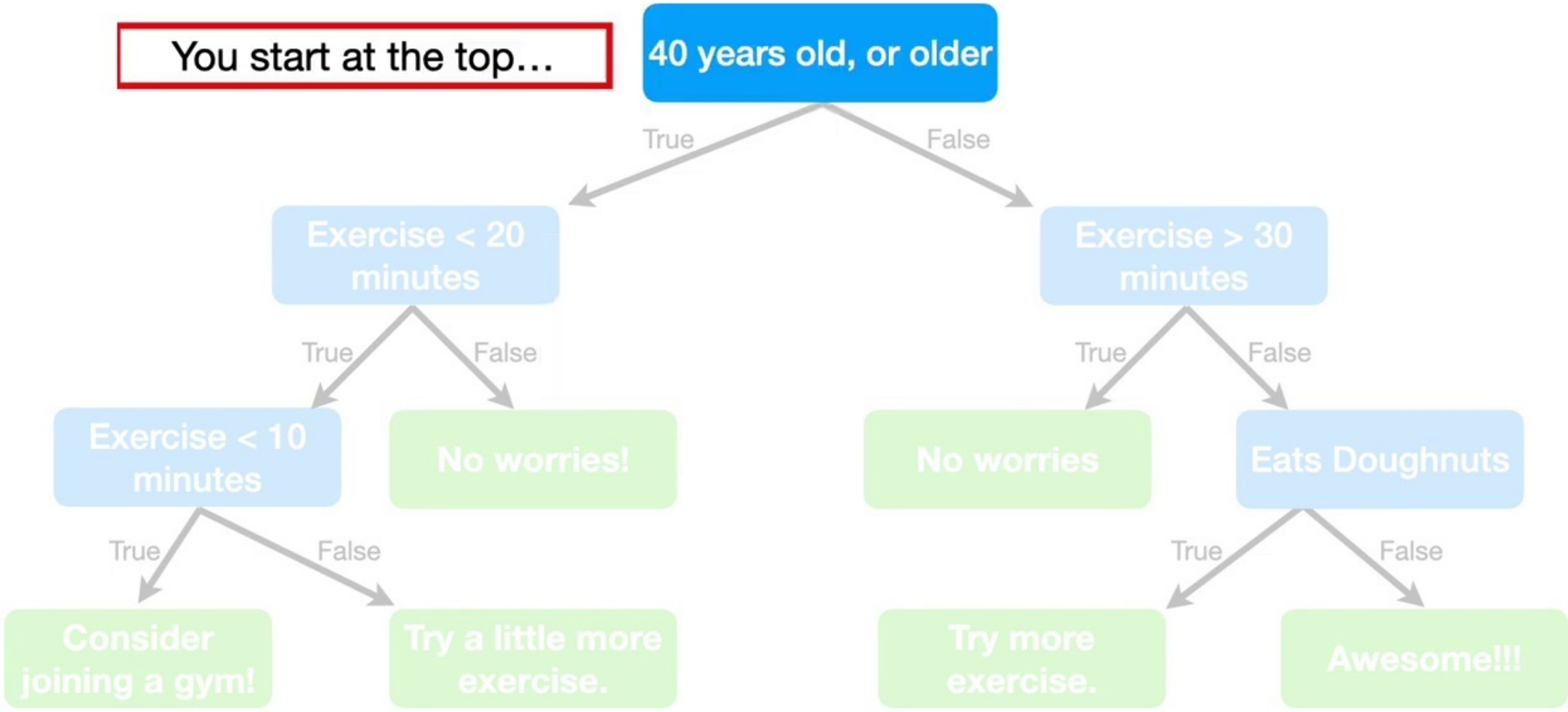

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
Eats Doughnuts
Also notice that the tree asks about **Exercising** multiple times…
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!


40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worri
rries
Eats Doughnuts
…and the amount of time **Exercising** isn't always the same.
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

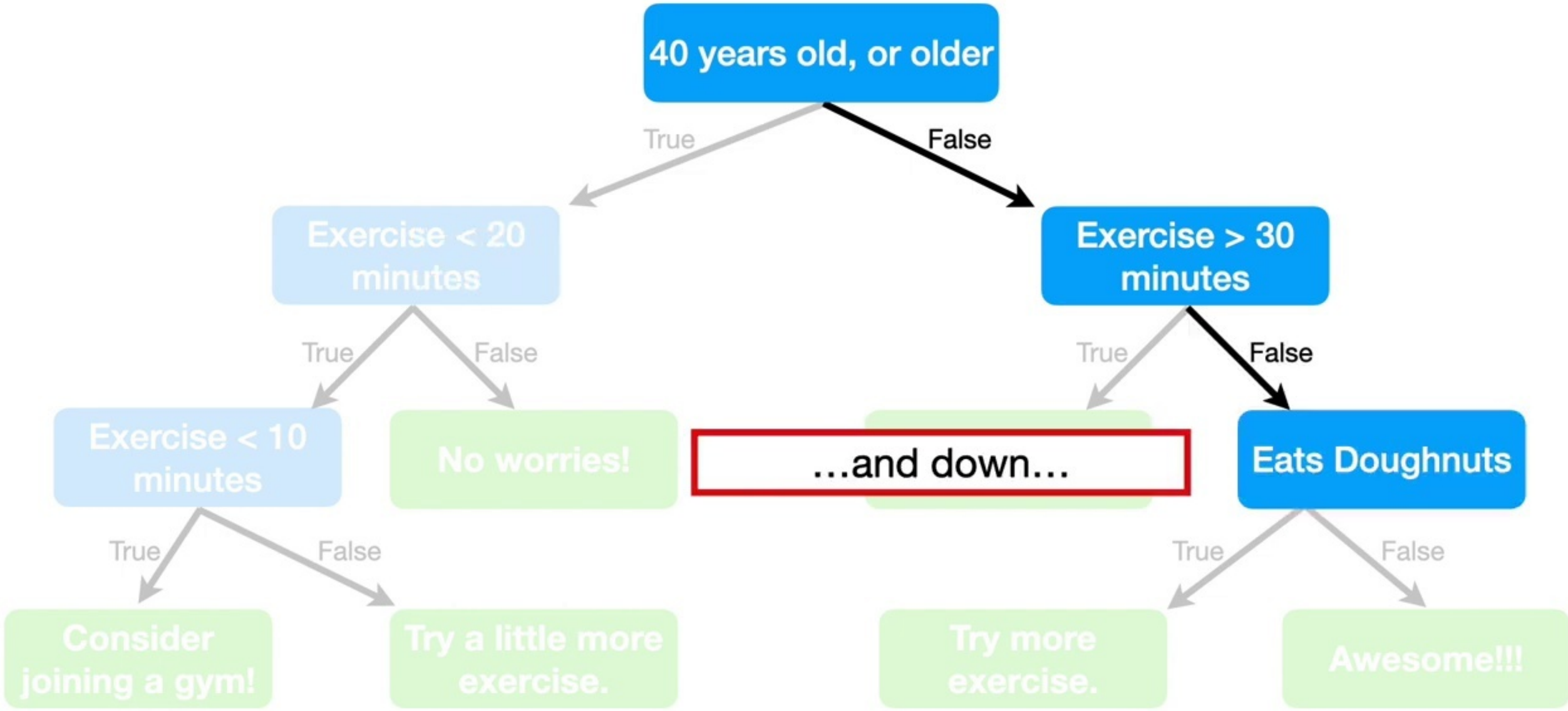

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
So numeric thresholds can be different for the same data.
Eats Doughnuts
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

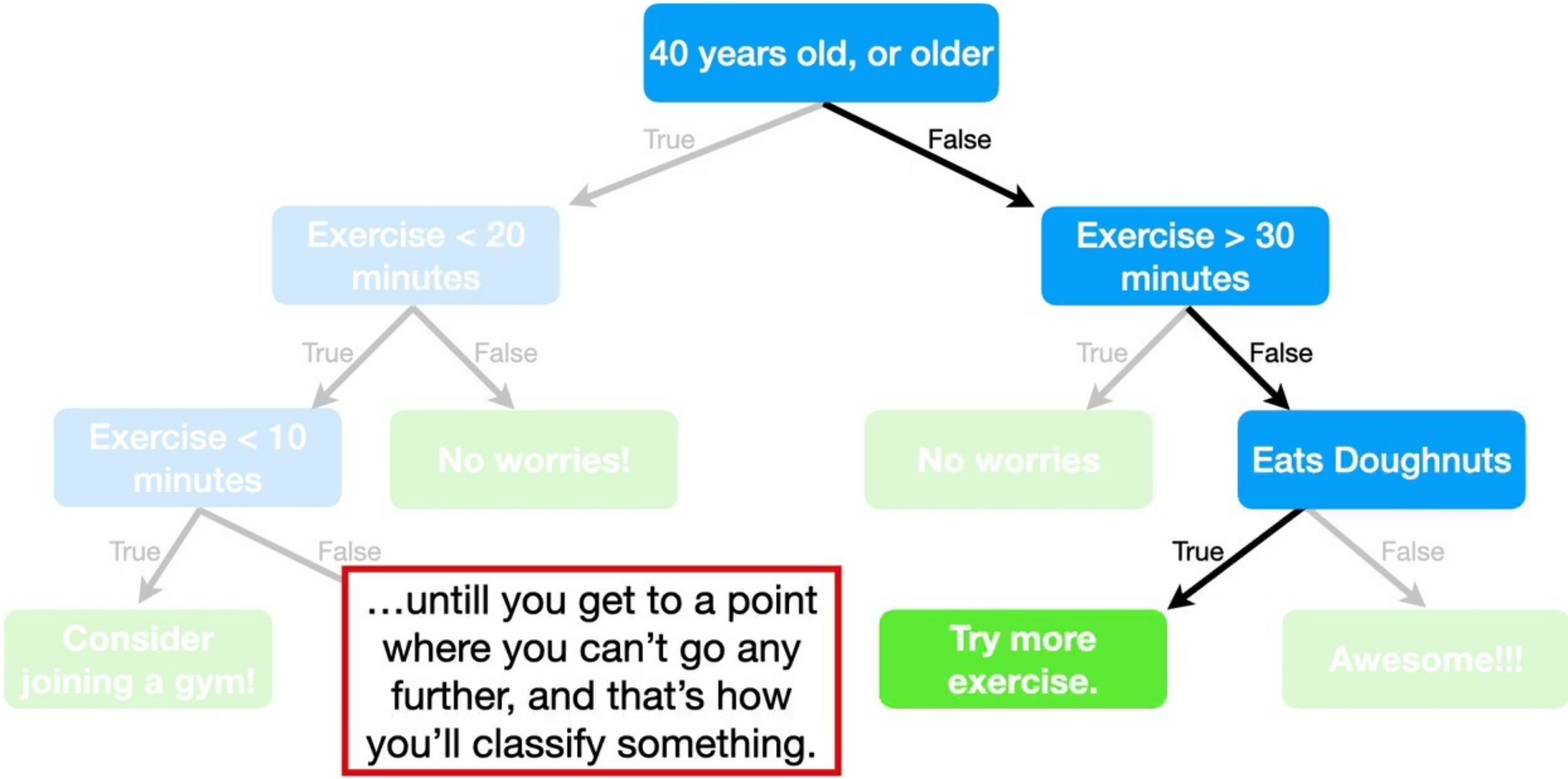

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
Lastly, the final classifications can be repeated.
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

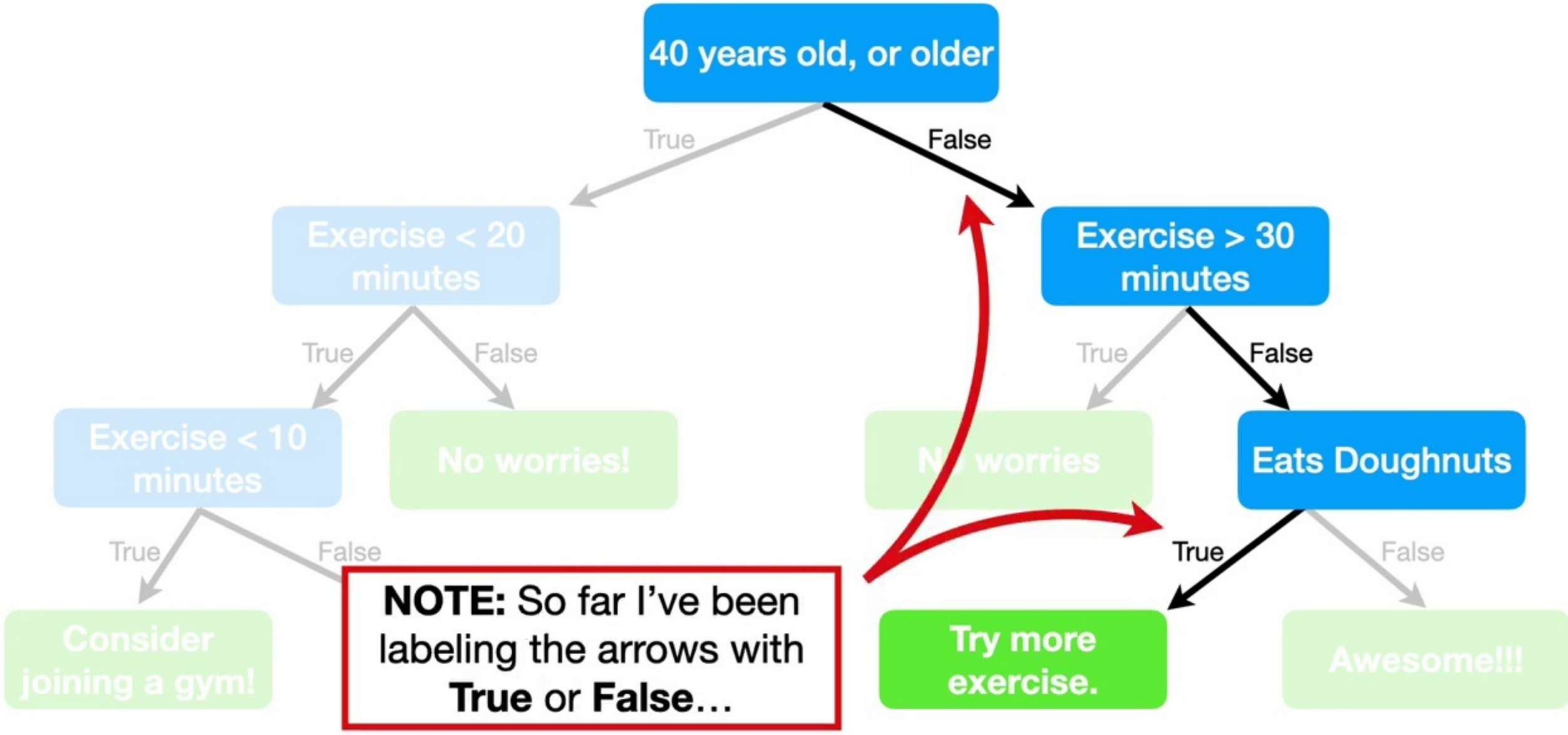

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
For the most part, **Classification Trees** are pretty easy to work with.
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

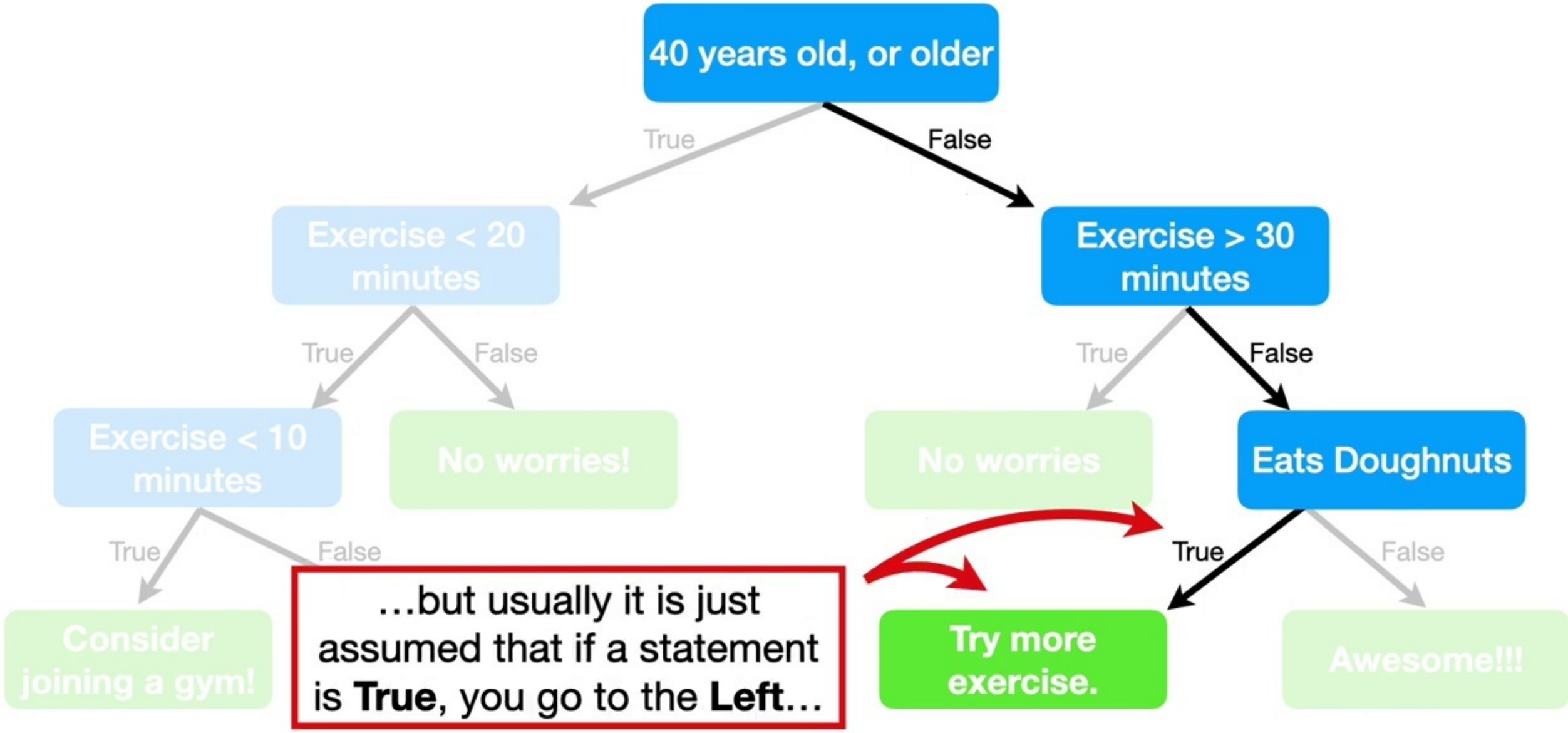

You start at the top…
40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

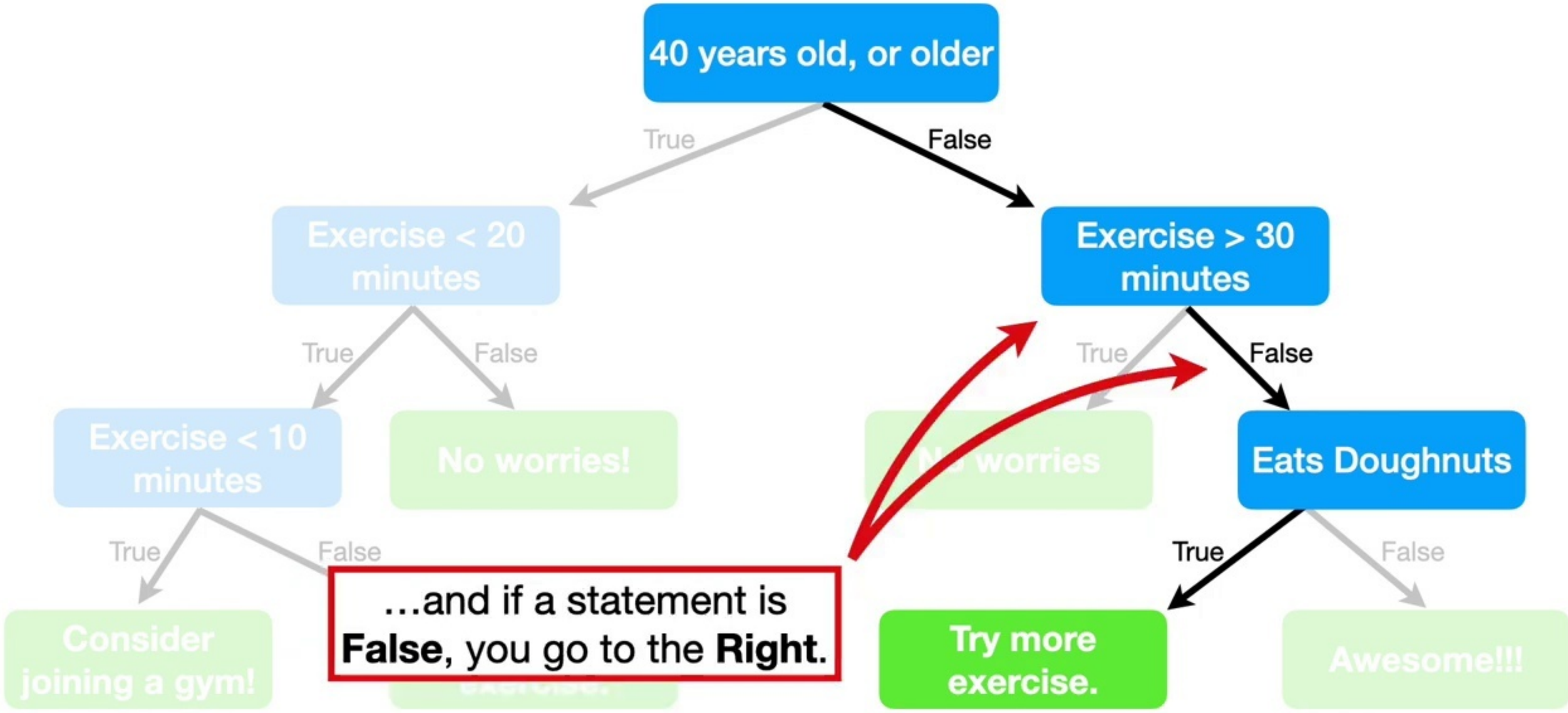

40 years old, or older
True
False
…and work your way down…
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!

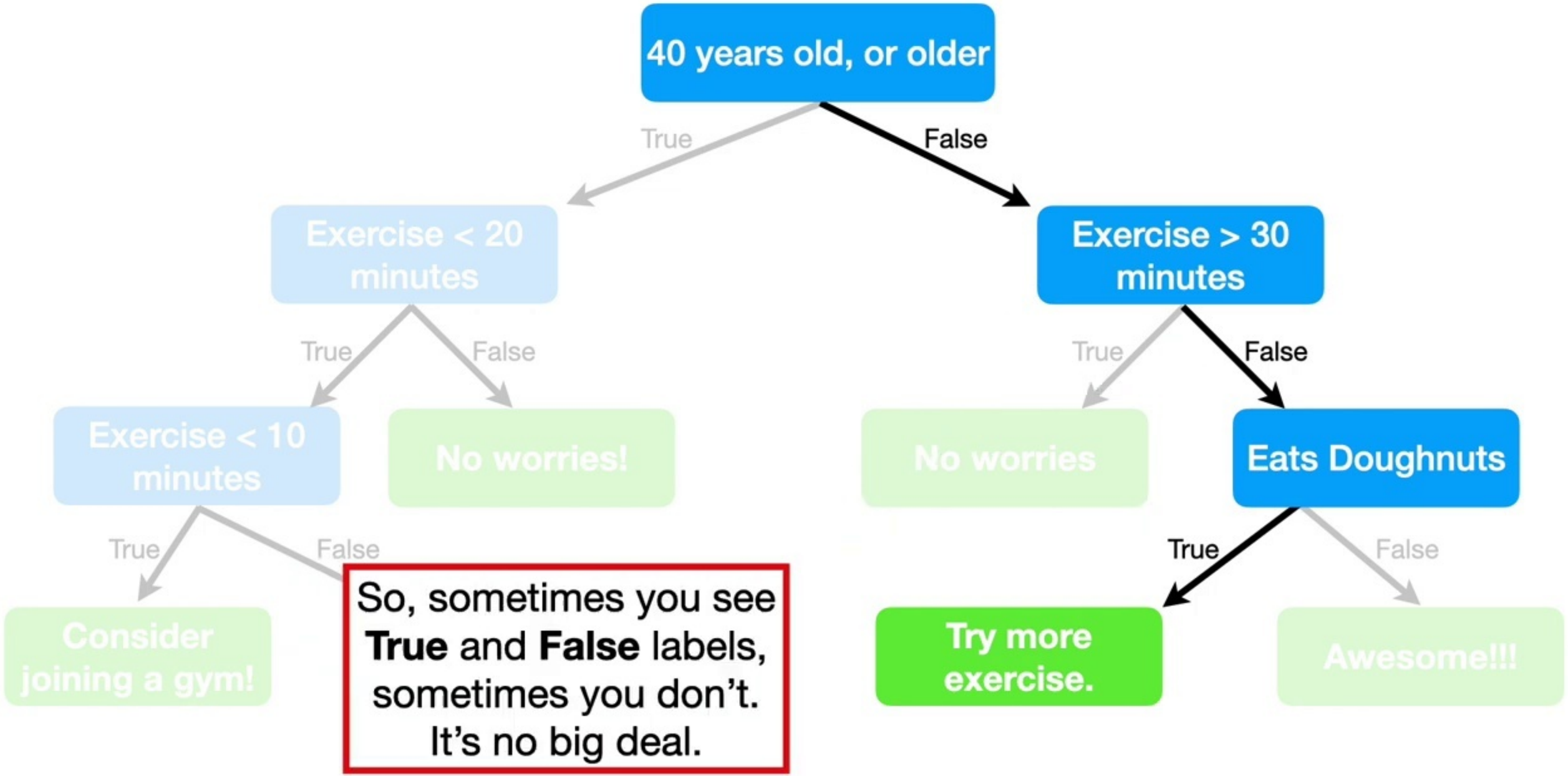

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
…and down…
Eats Doughnuts
True
False
True
False
Consider joining a gym!
Try a little more exercise.
Try more exercise.
Awesome!!!


40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
…untill you get to a point where you can't go any further, and that's how you'll classify something.
Try more exercise.
Awesome!!!

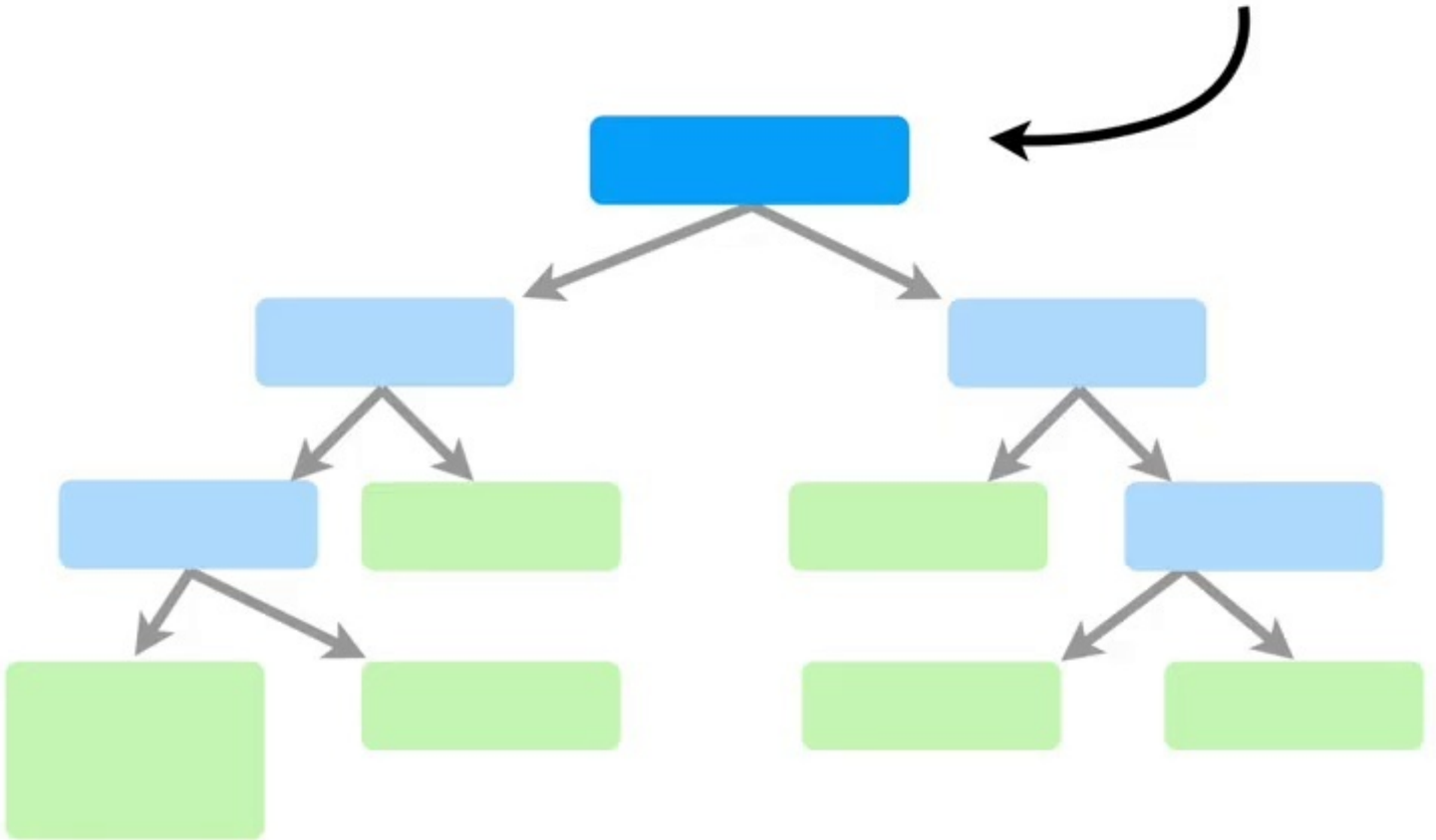

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
NOTE: So far I've been labeling the arrows with True or False…
Try more exercise.
Awesome!!!


40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
…but usually it is just assumed that if a statement is True, you go to the Left…
Try more exercise.
Awesome!!!

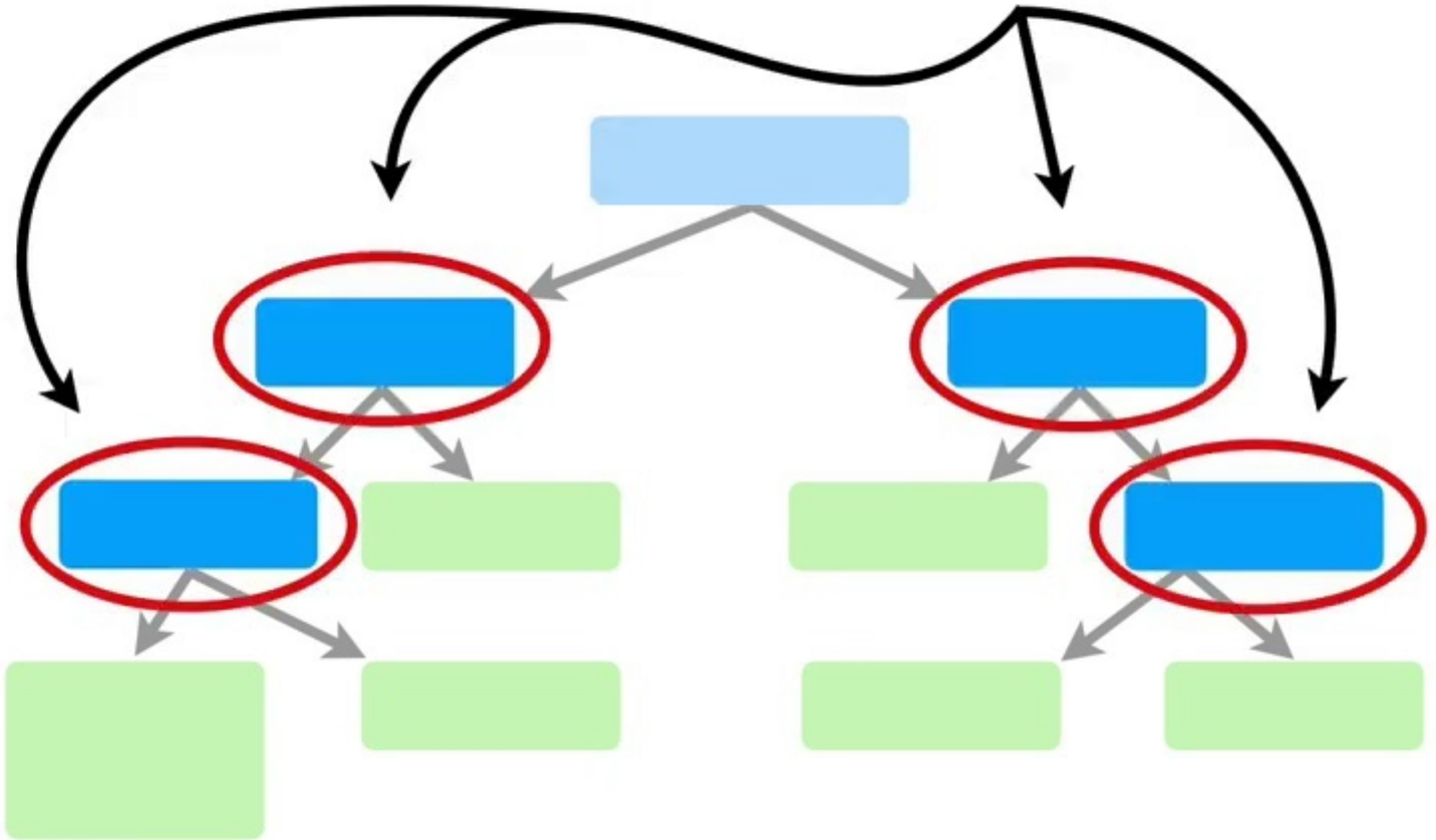

40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
…and if a statement is **False**, you go to the **Right**.
Consider joining a gym!
Try more exercise.
Awesome!!!


40 years old, or older
True
False
Exercise < 20 minutes
Exercise > 30 minutes
True
False
True
False
Exercise < 10 minutes
No worries!
No worries
Eats Doughnuts
True
False
True
False
Consider joining a gym!
So, sometimes you see True and False labels, sometimes you don't. It's no big deal.
Try more exercise.
Awesome!!!

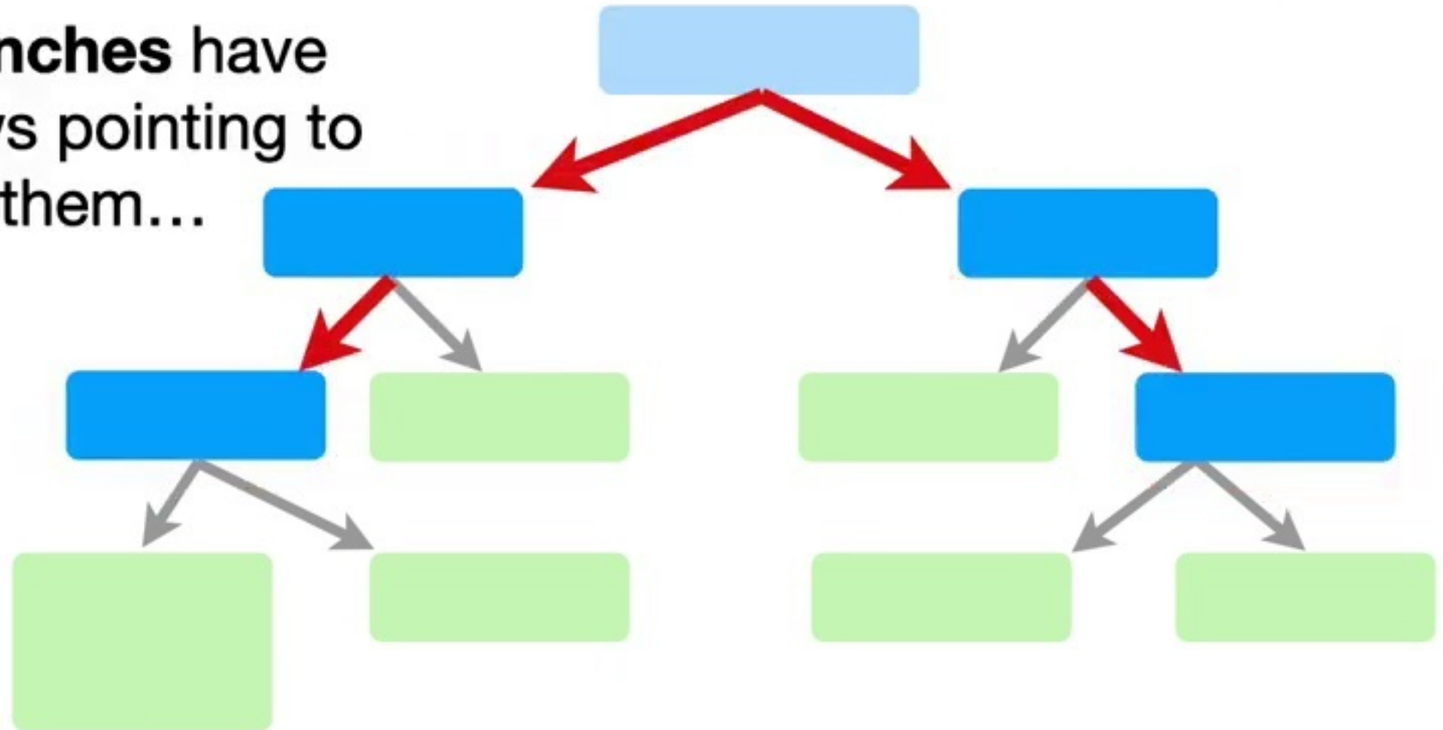
The very top of the tree is called the **Root Node** or just **The Root**.

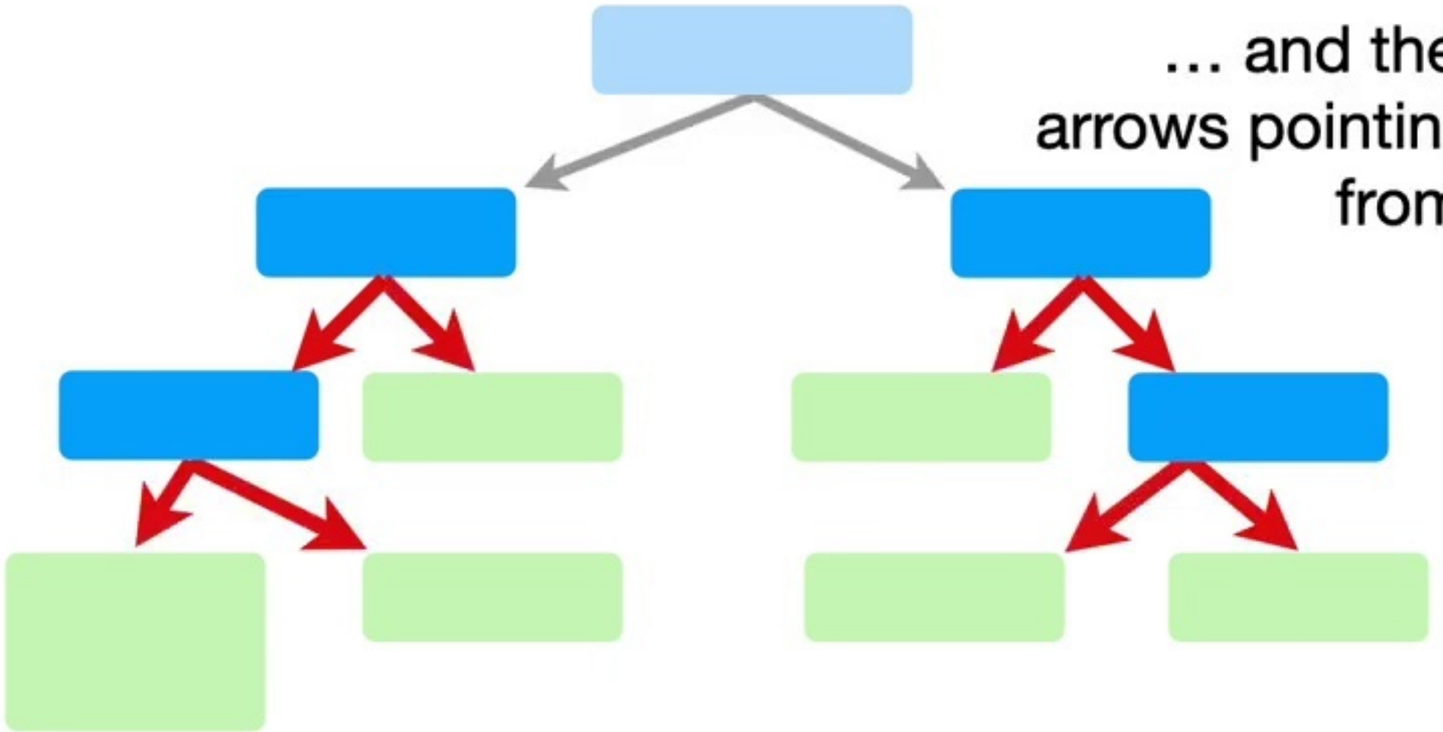


These are called **Internal Nodes**, or **Branches**.

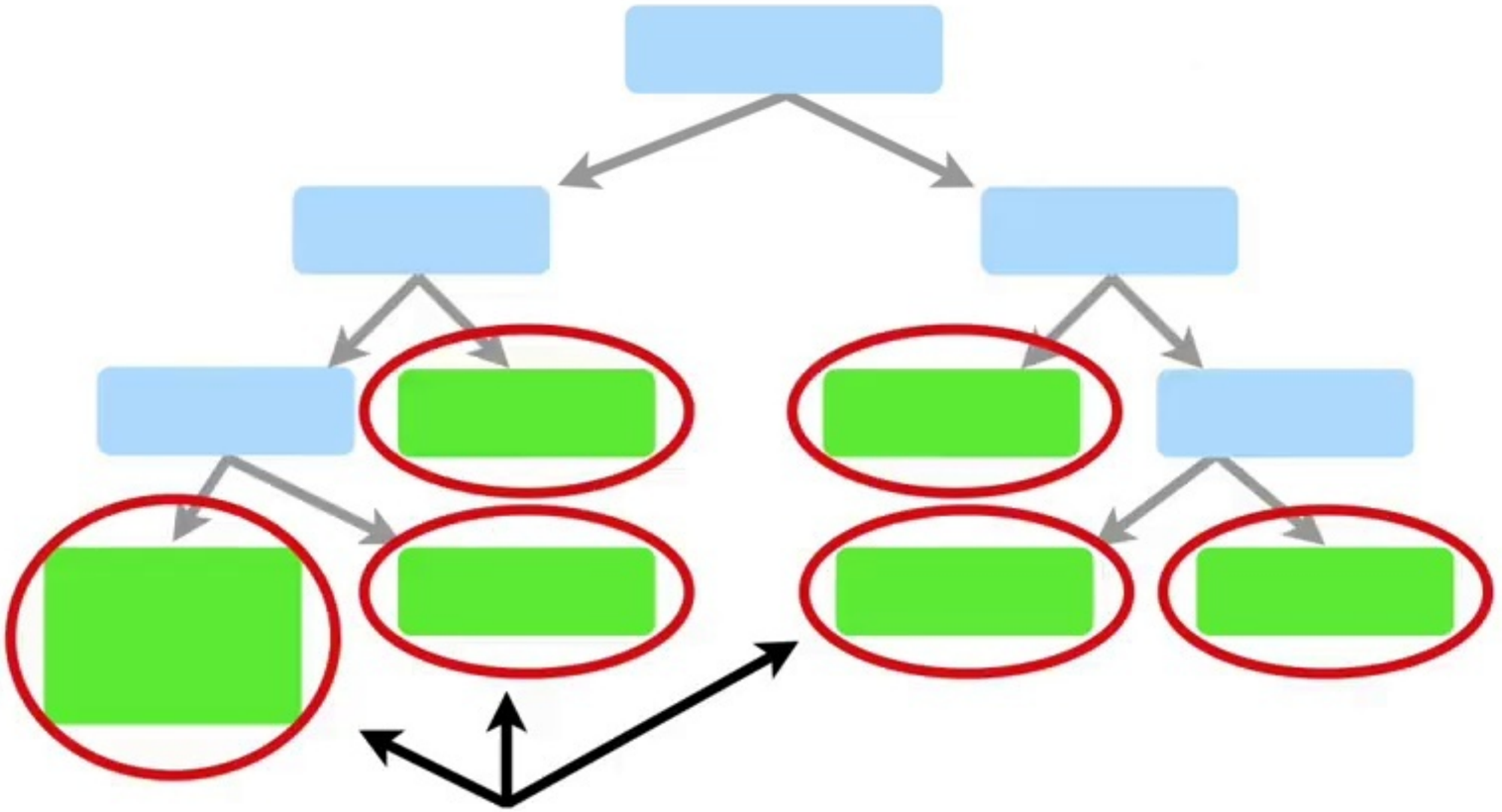


Branches have arrows pointing to them…

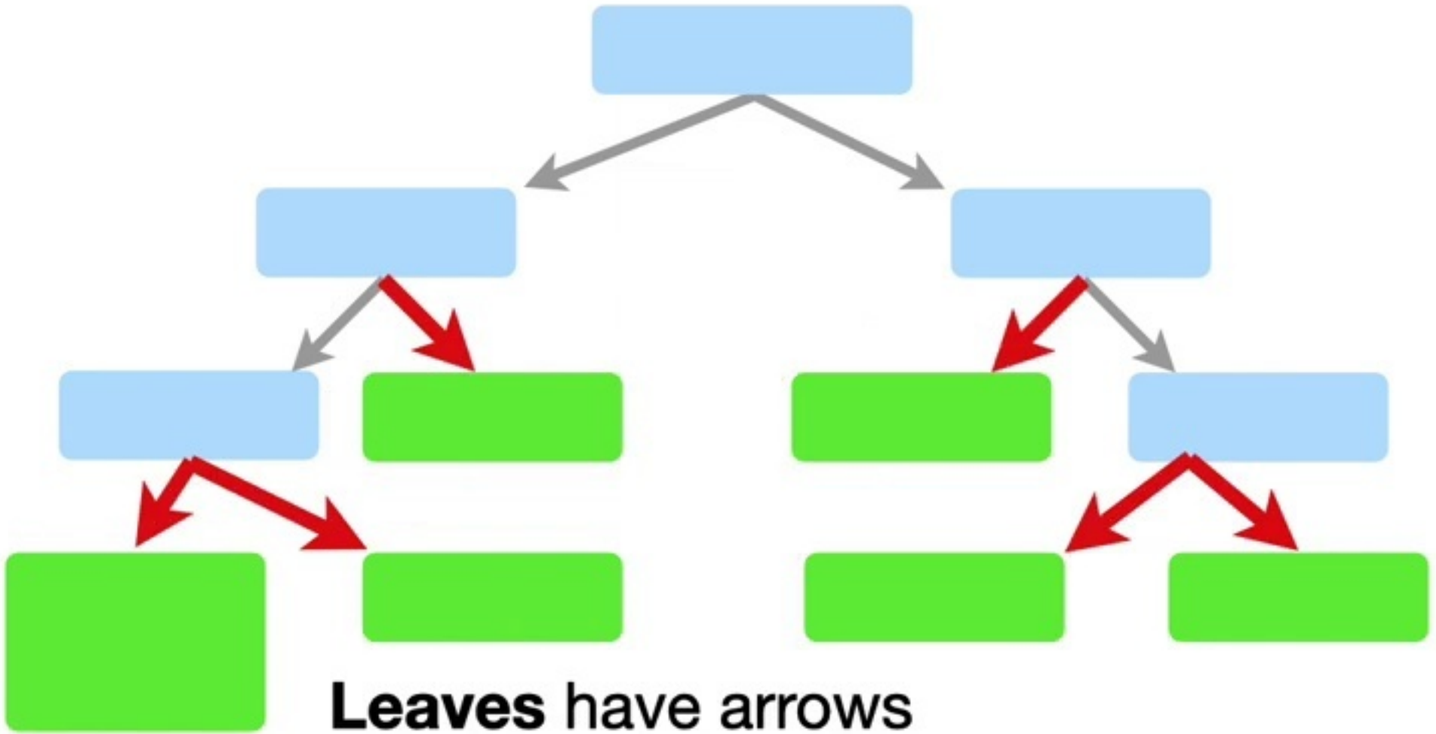




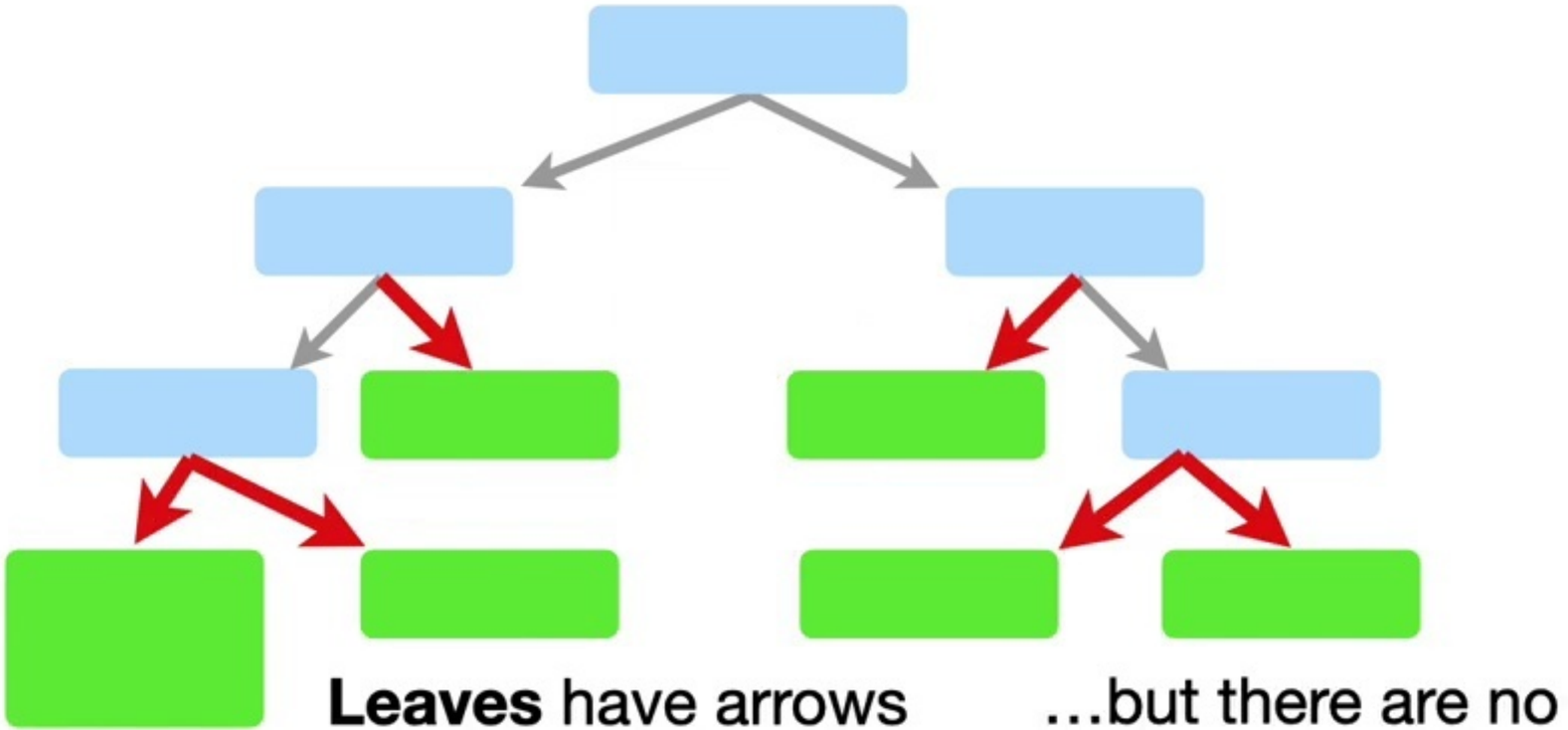
… and they have
arrows pointing away
from them.



Lastly, these are called **Leaf Nodes**, or just **Leaves**.



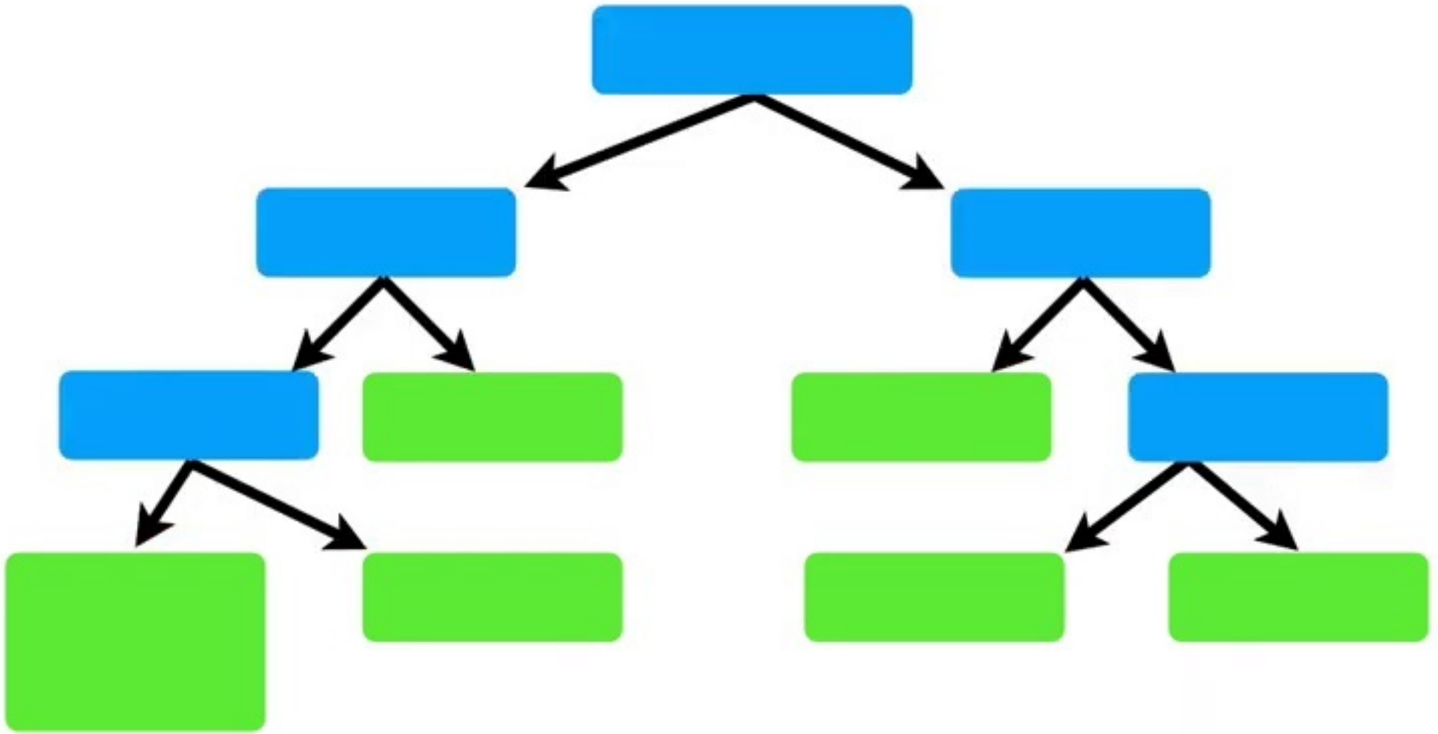
Leaves have arrows pointing to them…



Leaves have arrows pointing to them…

…but there are no arrows pointing away from them.

Now that we know how to use and interpret **Classification Trees**…



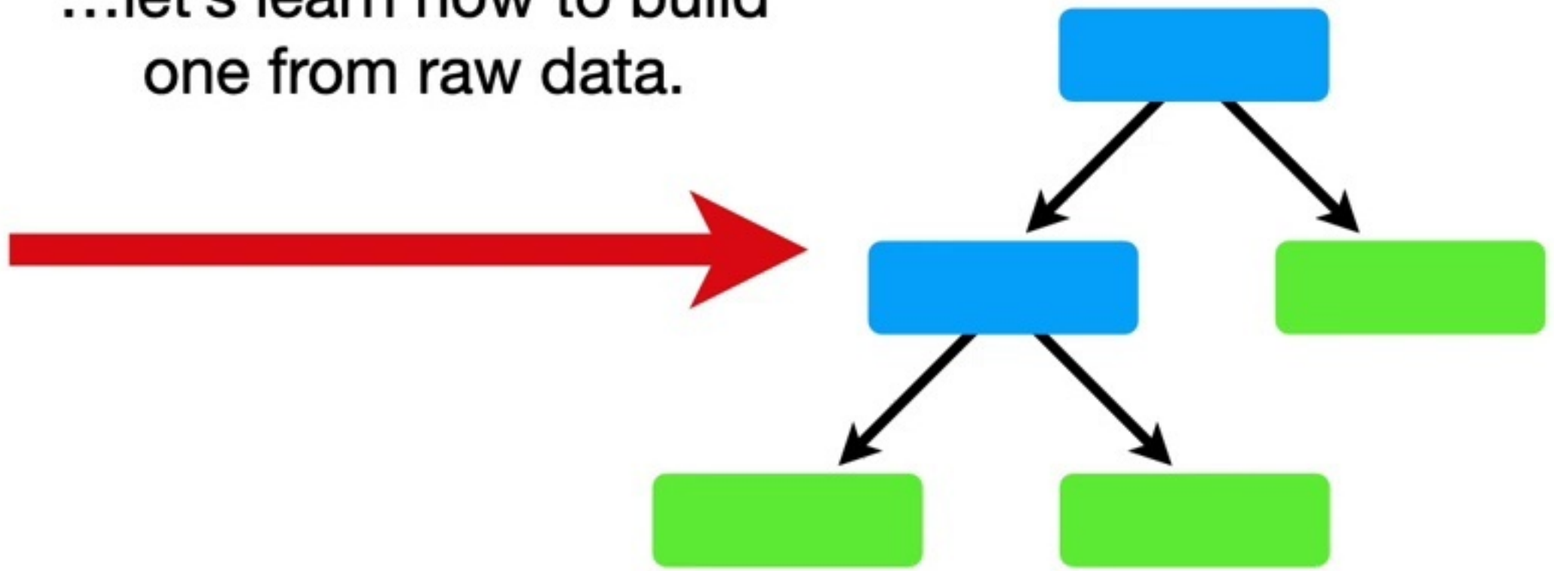
…let’s learn how to build
one from raw data.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

…let's learn how to build one from raw data.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

…let's learn how to build one from raw data.



Loves Popcorn	Loves Soda	Age	Loves Cool as Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

This data tells us whether or not someone **Loves Popcorn**…


Loves Popcorn
Loves Soda
Age
Loves Cool As Ice
Yes
Yes
7
No
Yes
No
12
No
No
Yes
18
Yes
No
Yes
35
Yes
Yes
Yes
38
Yes
Yes
No
50
No
No
No
83
No
…whether or not they **Love Soda**…

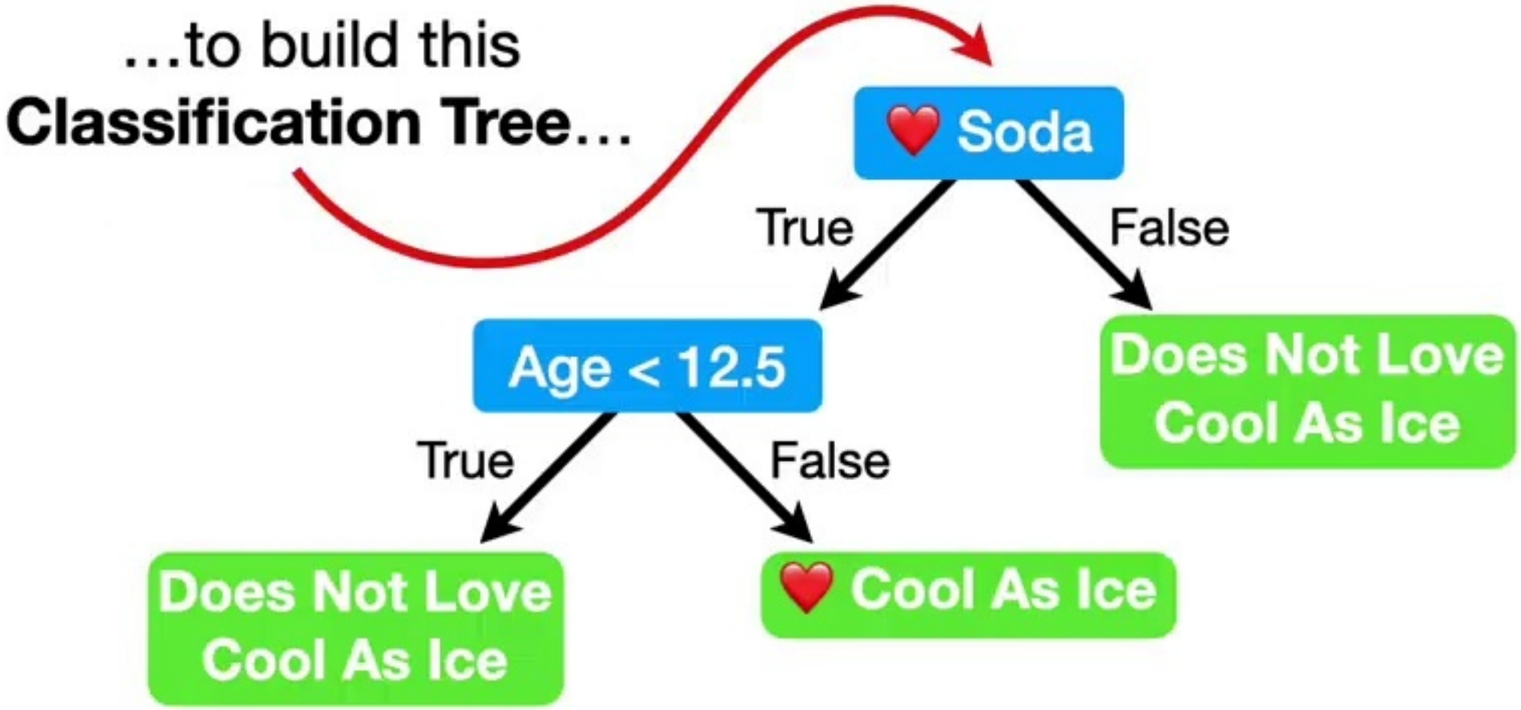
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

…and whether or not they **Love** the **1991** blockbuster, **Cool As Ice**, starring Vanilla Ice.

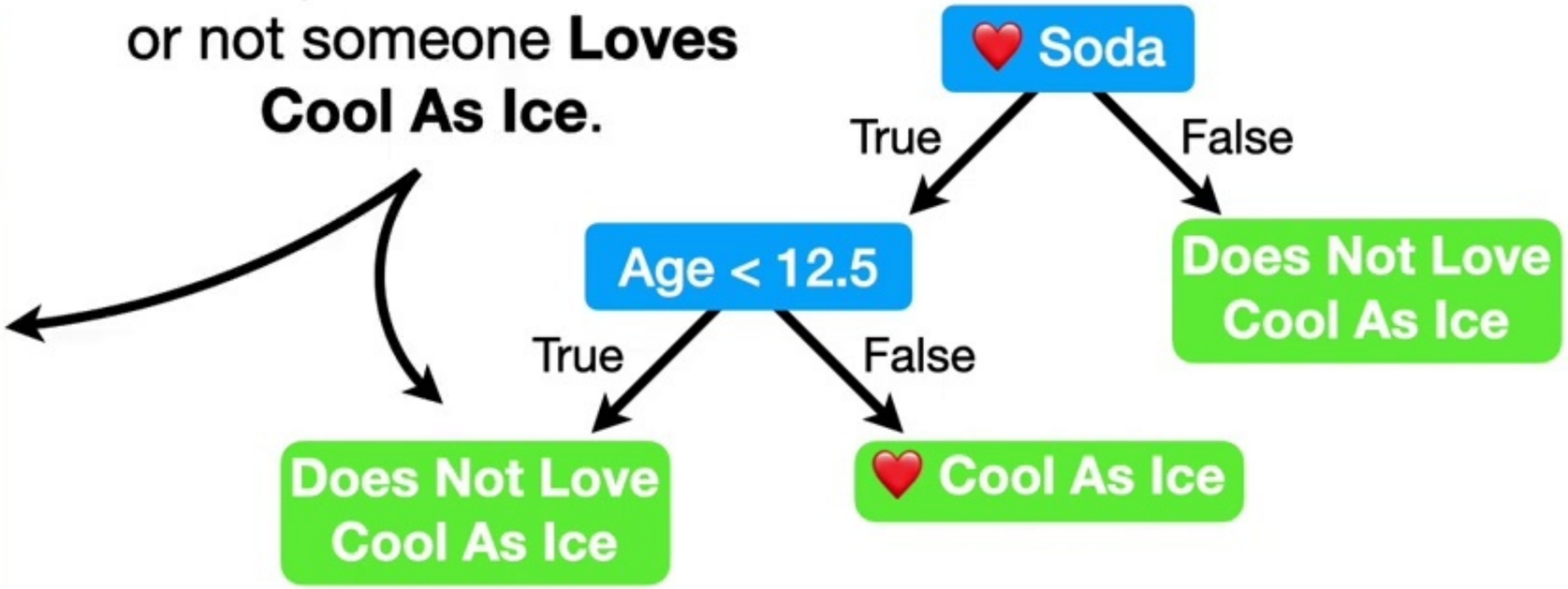
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



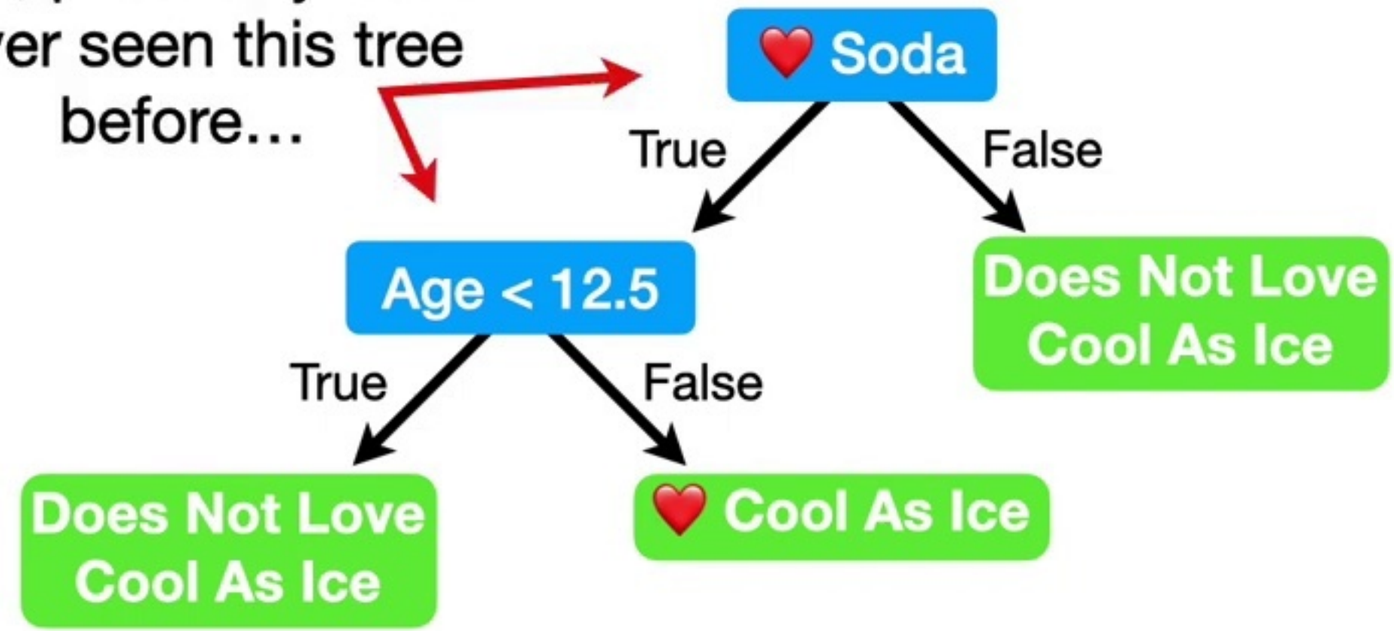
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

…that predicts whether or not someone **Loves Cool As Ice.**



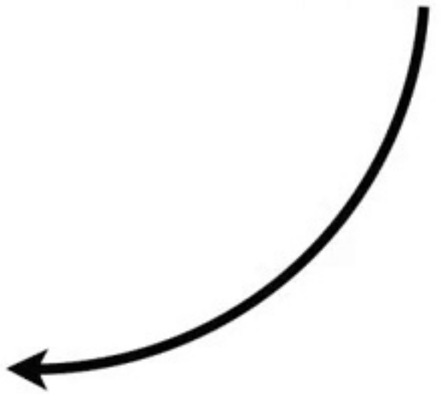
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Now, pretend you've never seen this tree before…



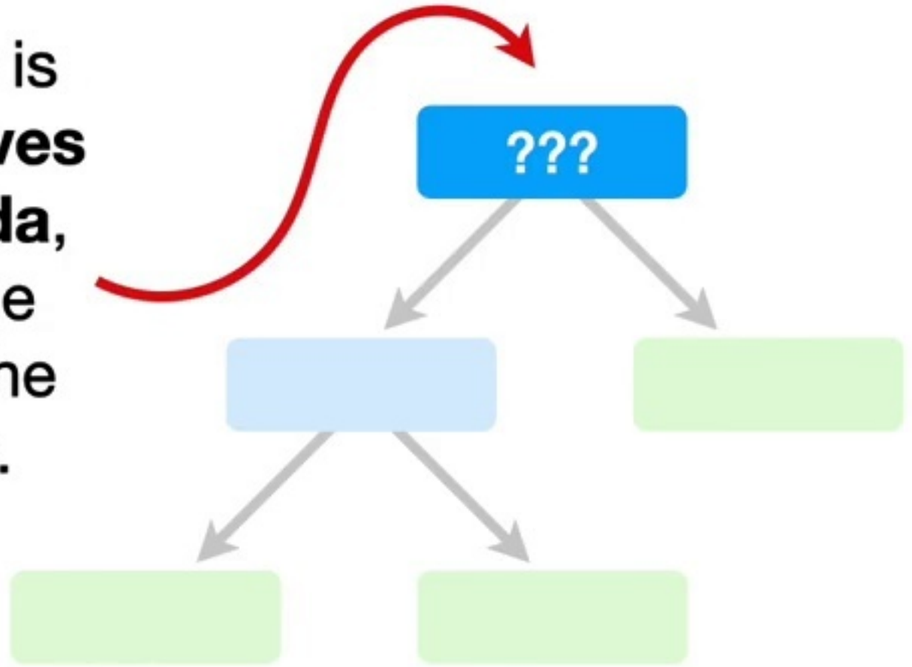
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

…and let's see how to build a tree starting with just data.



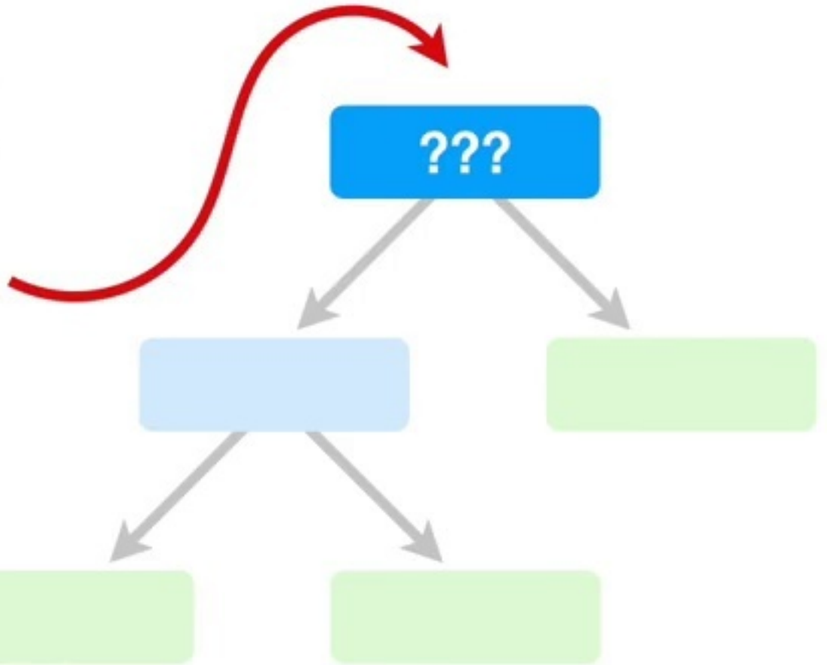
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

The first thing we do is decide is whether **Loves Popcorn**, **Loves Soda**, or **Age** should be the question we ask at the very top of the tree.



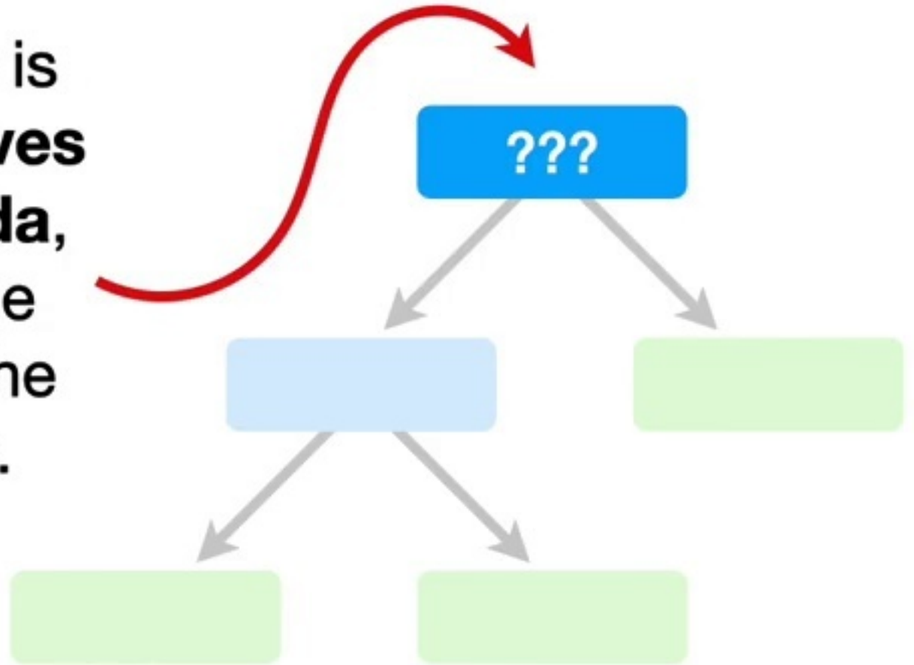
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

The first thing we do is decide is whether **Loves Popcorn**, **Loves Soda**, or **Age** should be the question we ask at the very top of the tree.



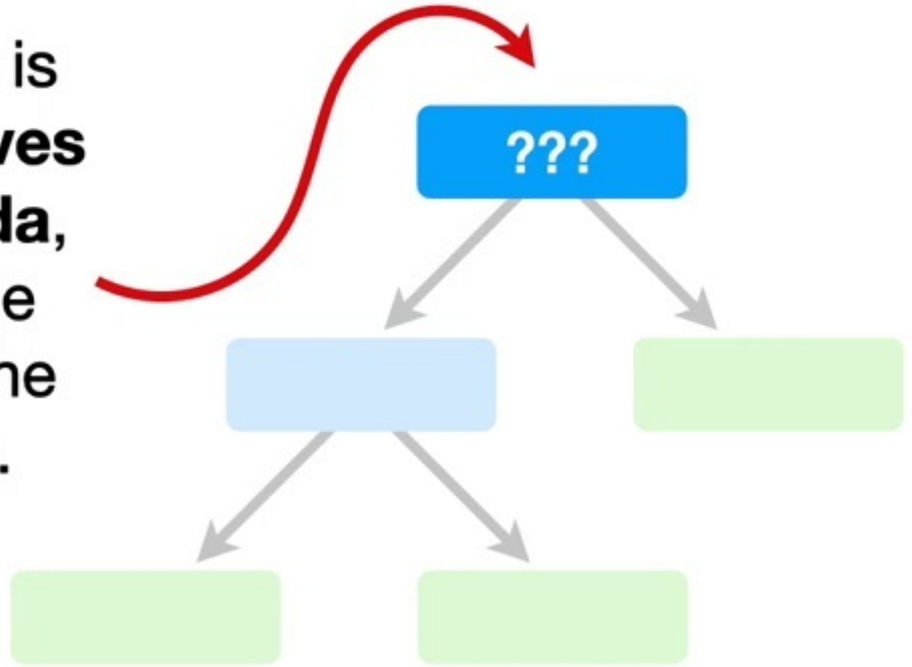
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

The first thing we do is decide is whether **Loves Popcorn**, **Loves Soda**, or **Age** should be the question we ask at the very top of the tree.



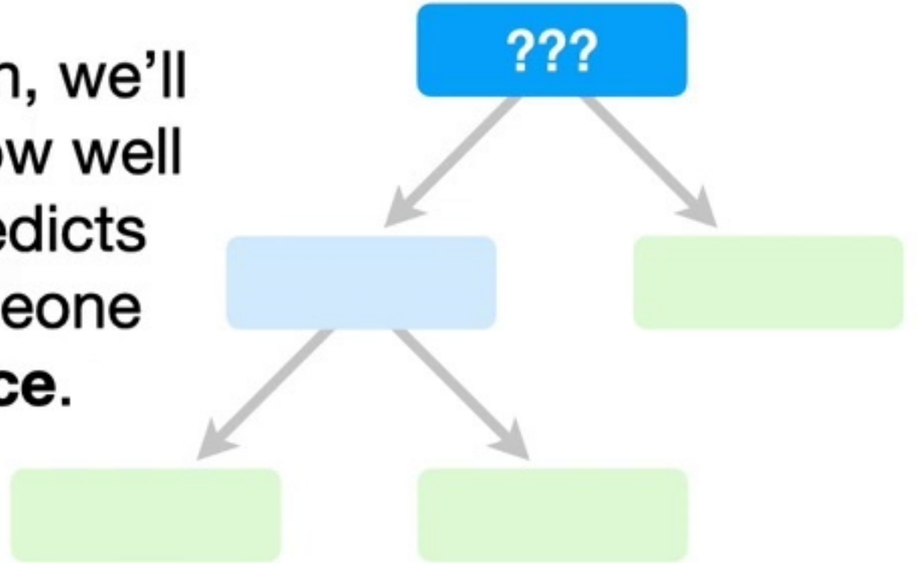
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

The first thing we do is decide is whether **Loves Popcorn**, **Loves Soda**, or **Age** should be the question we ask at the very top of the tree.

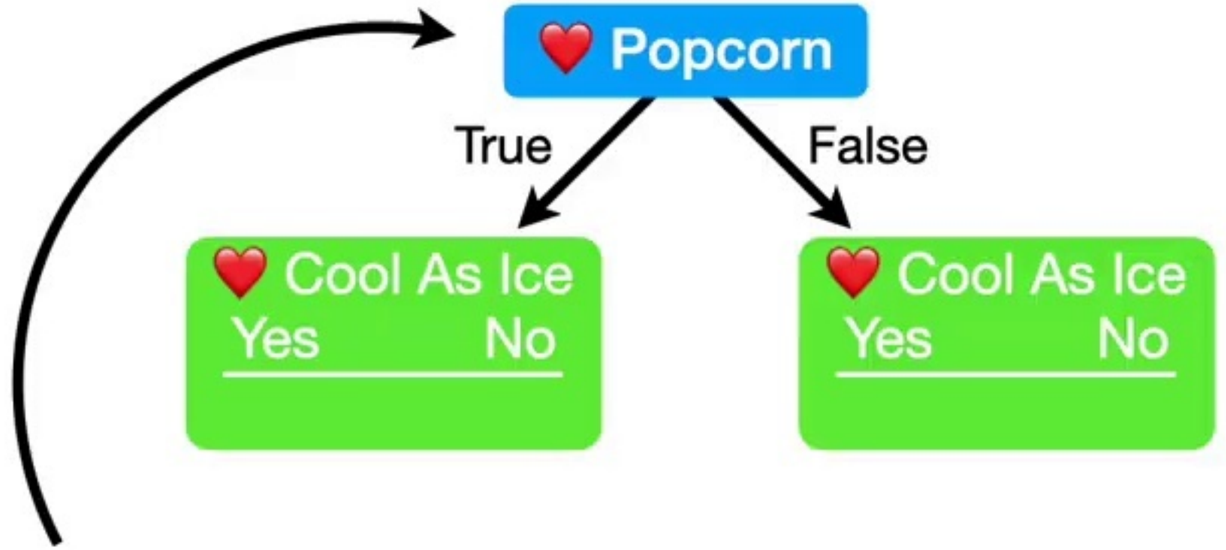


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

To make that decision, we'll start by looking at how well **Loves Popcorn** predicts whether or not someone **Loves Cool As Ice**.

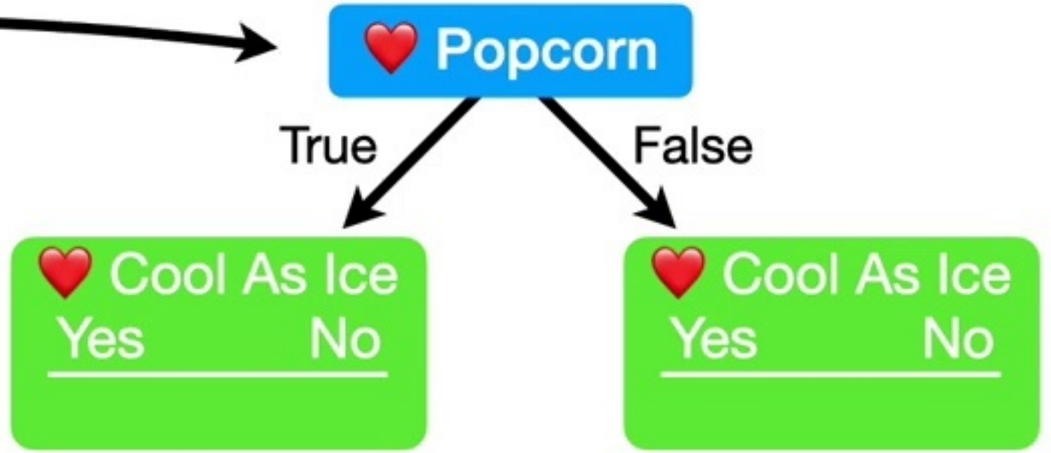


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



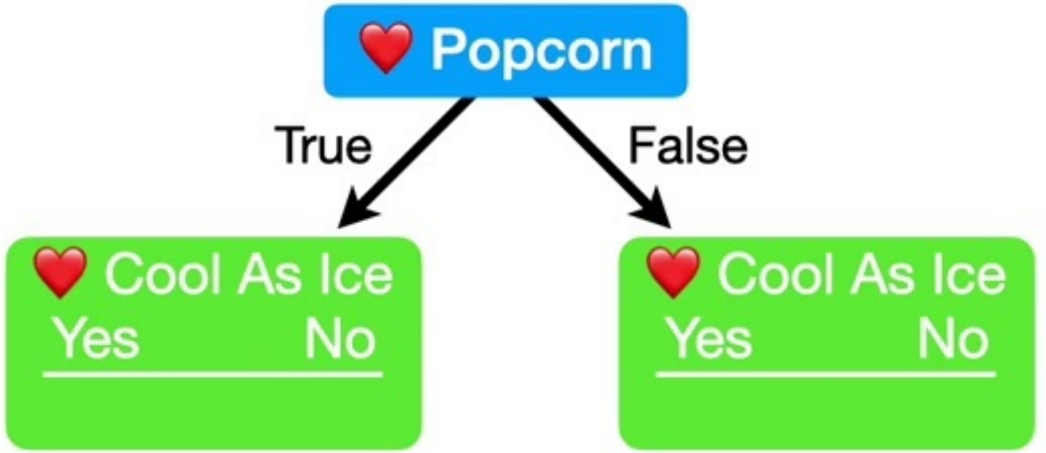
To do this, we'll make a super simple tree that only asks if someone **Loves Popcorn**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



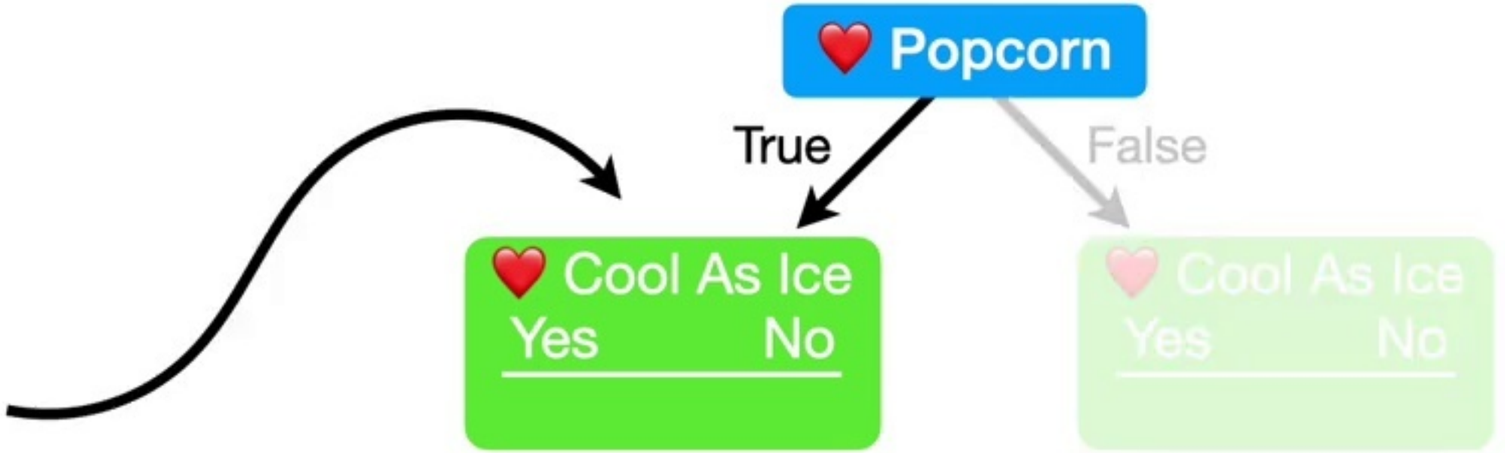
...and then we run the data down the tree.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



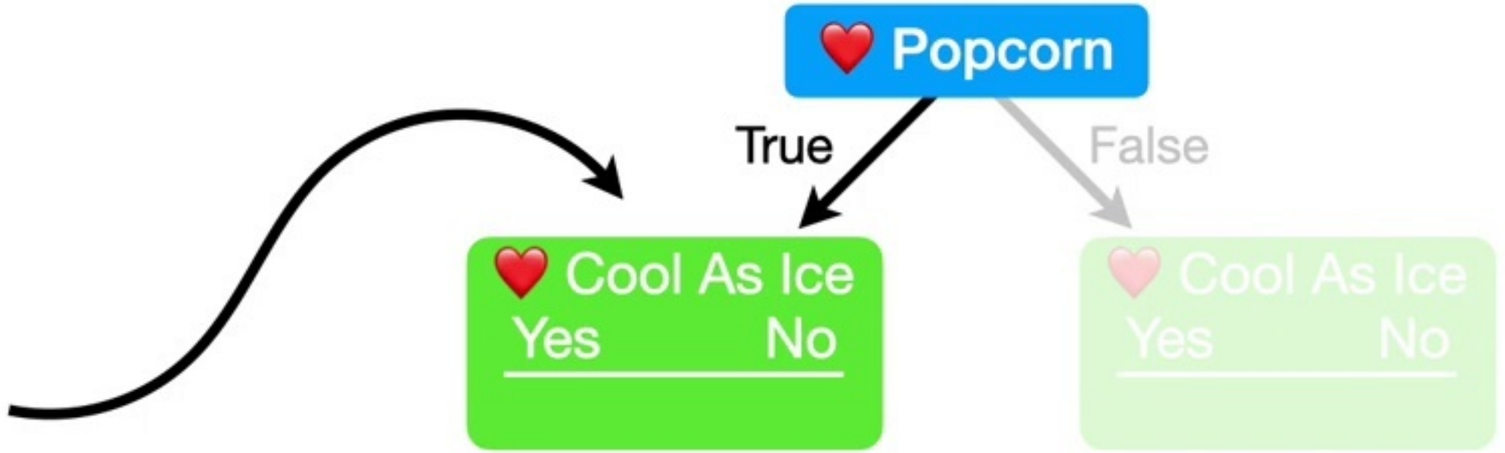
For example, the first person in the dataset **Loves Popcorn**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



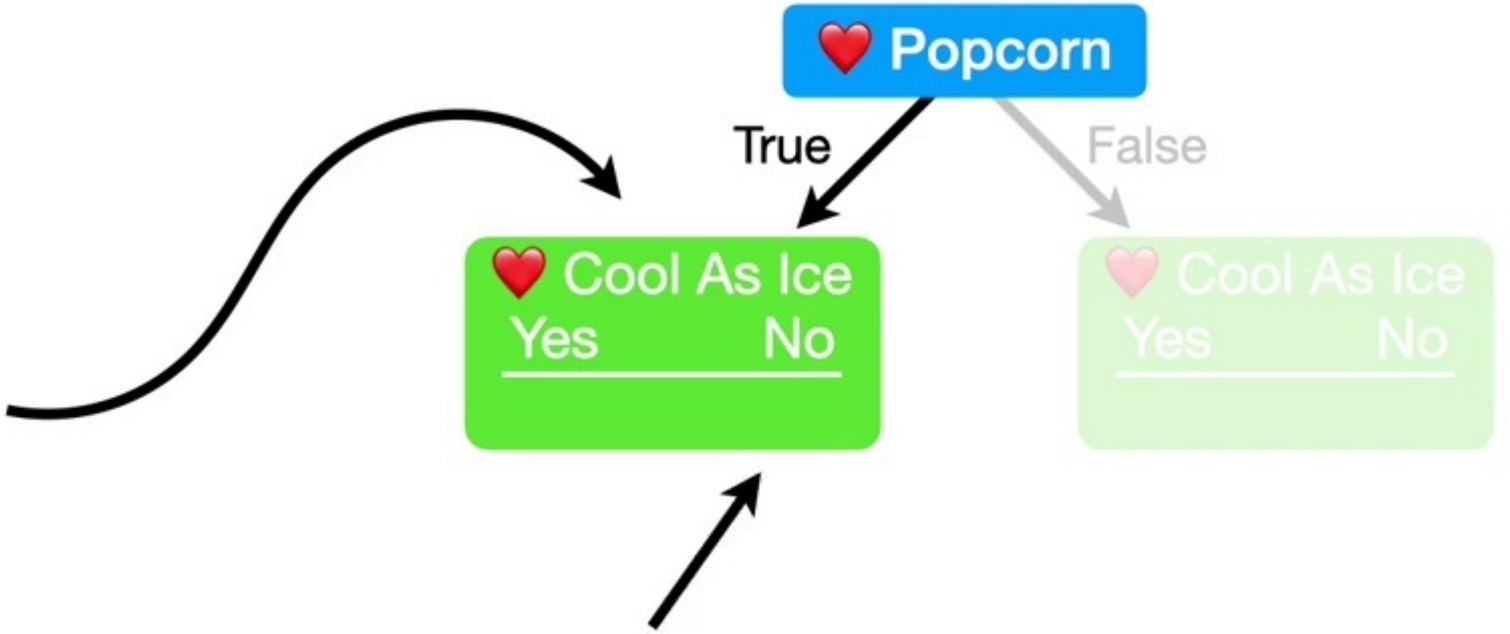
…so they go the **Leaf** on the *left*.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	N
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



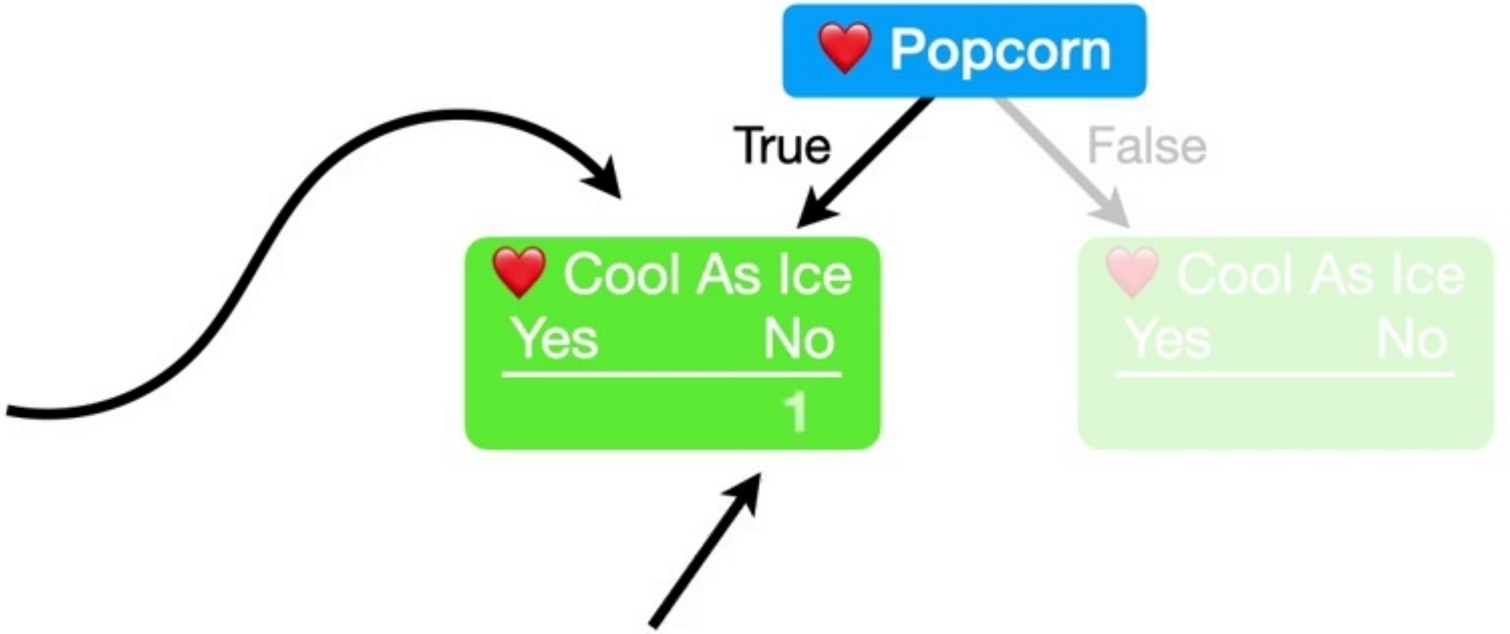
And because they *do not* **Love Cool As Ice**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



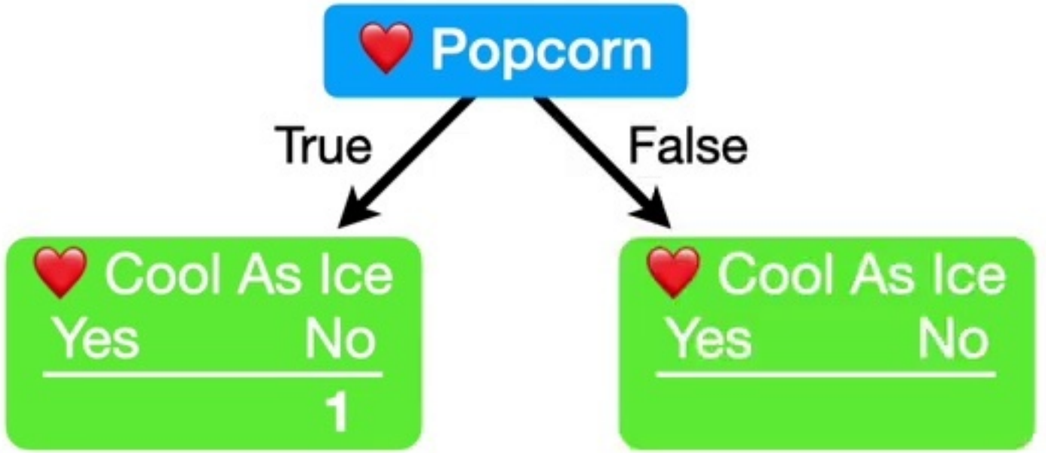
…we'll keep track of that by putting a **1** under the word **No**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



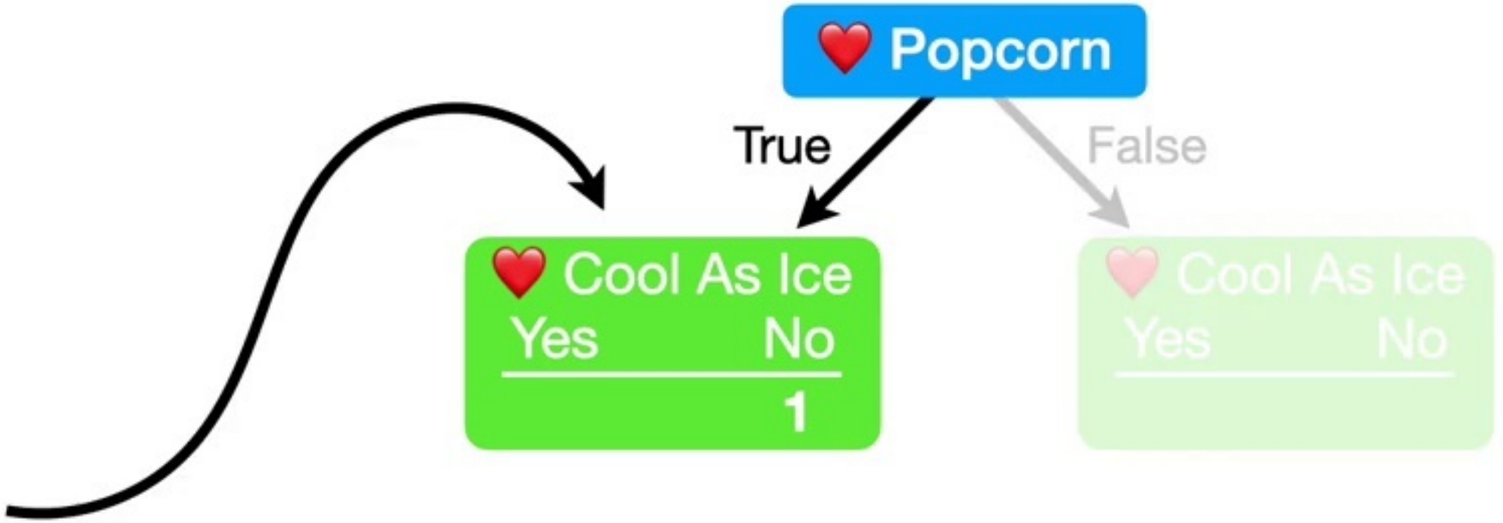
…we'll keep track of that by putting a **1** under the word **No**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
N	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



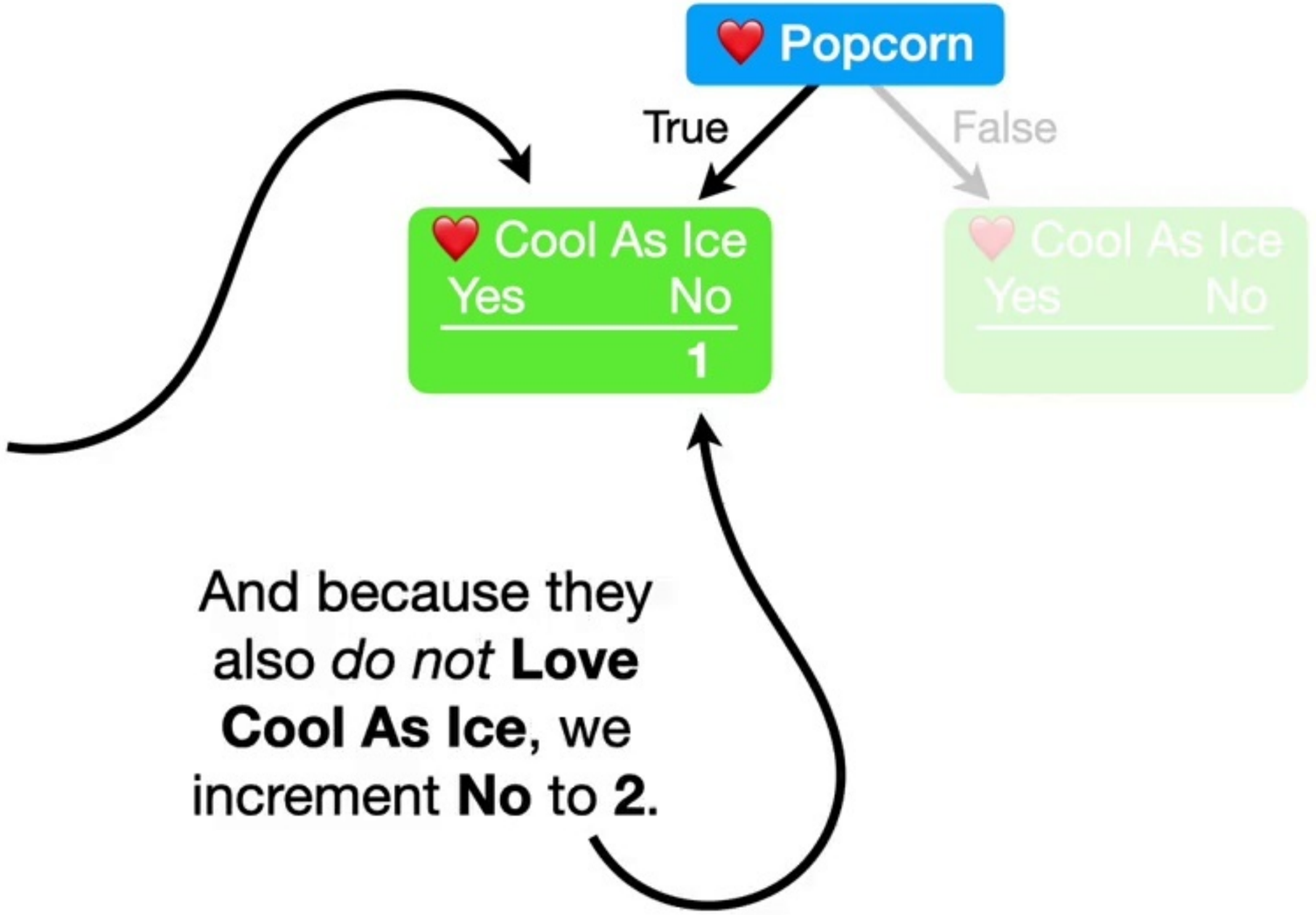
The second person in the dataset also **Loves Popcorn**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

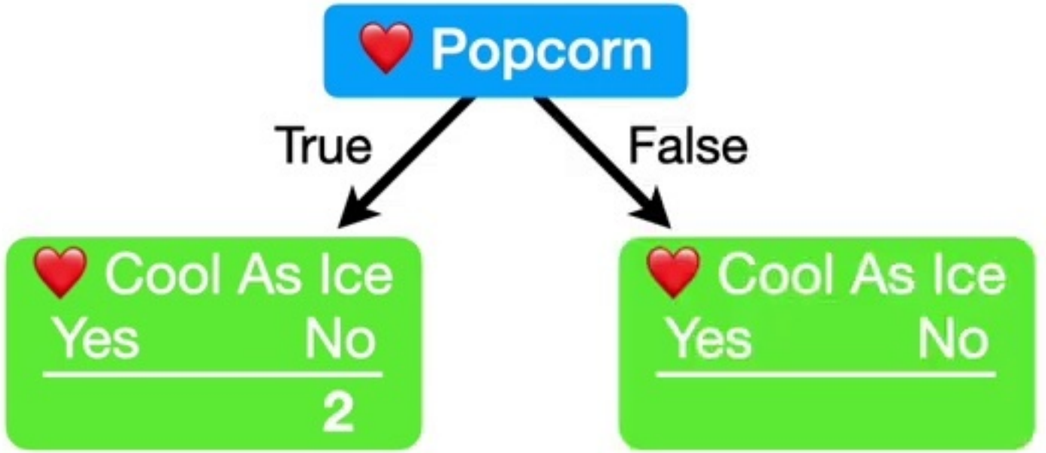


…so they also go to the **Leaf** on the *left*.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

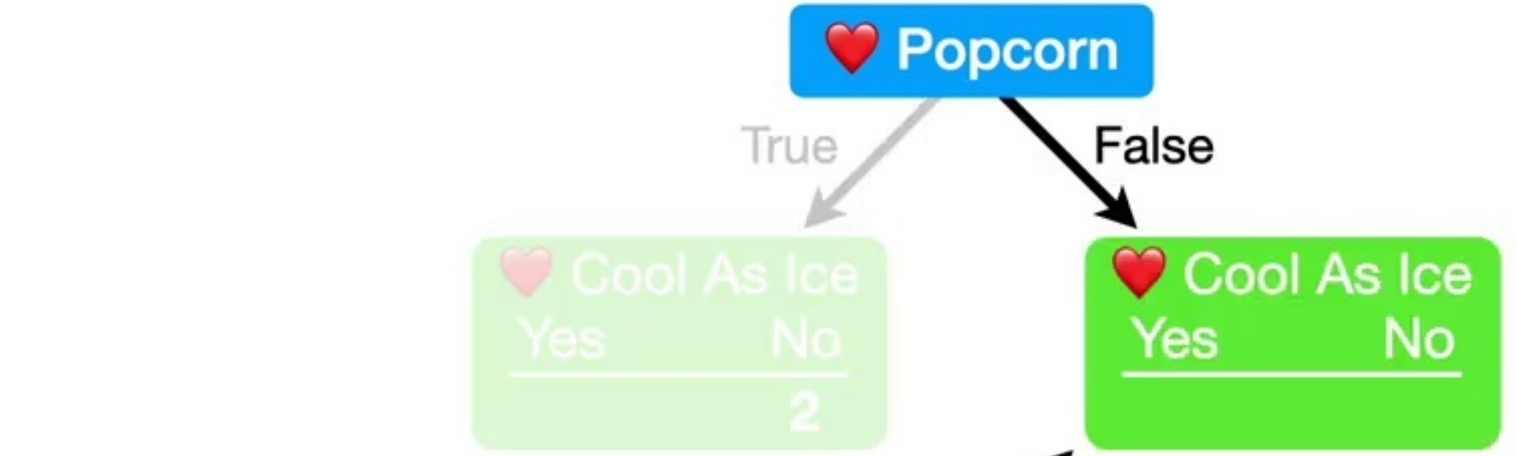


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



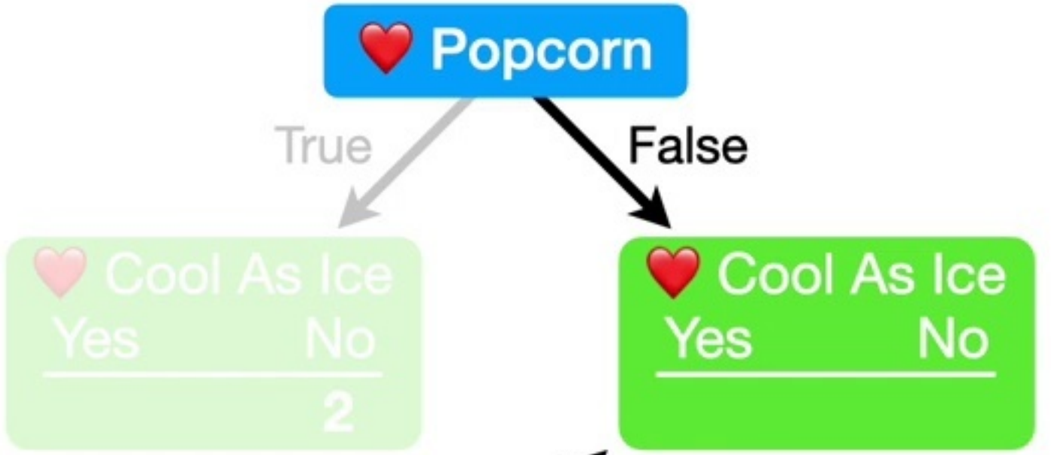
The third person *does not* **Love Popcorn**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



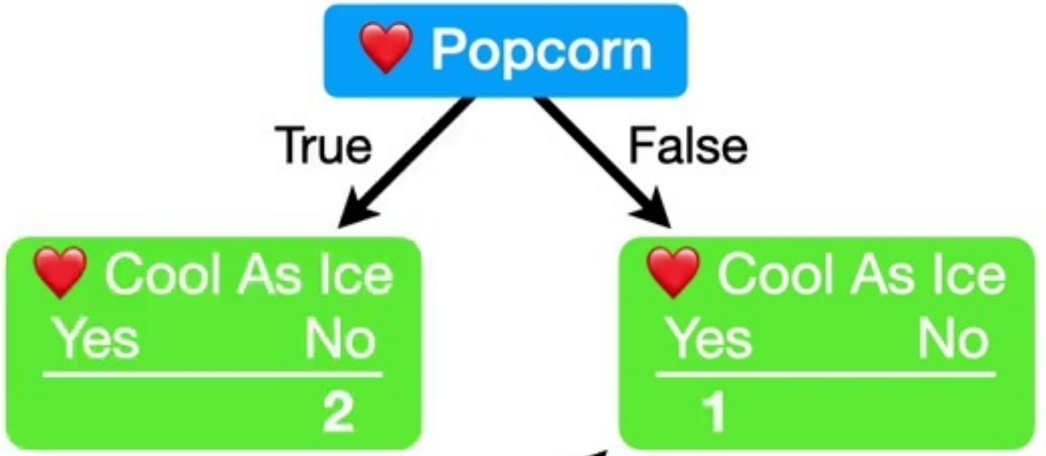
…so they go to the **Leaf** on the *right*.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



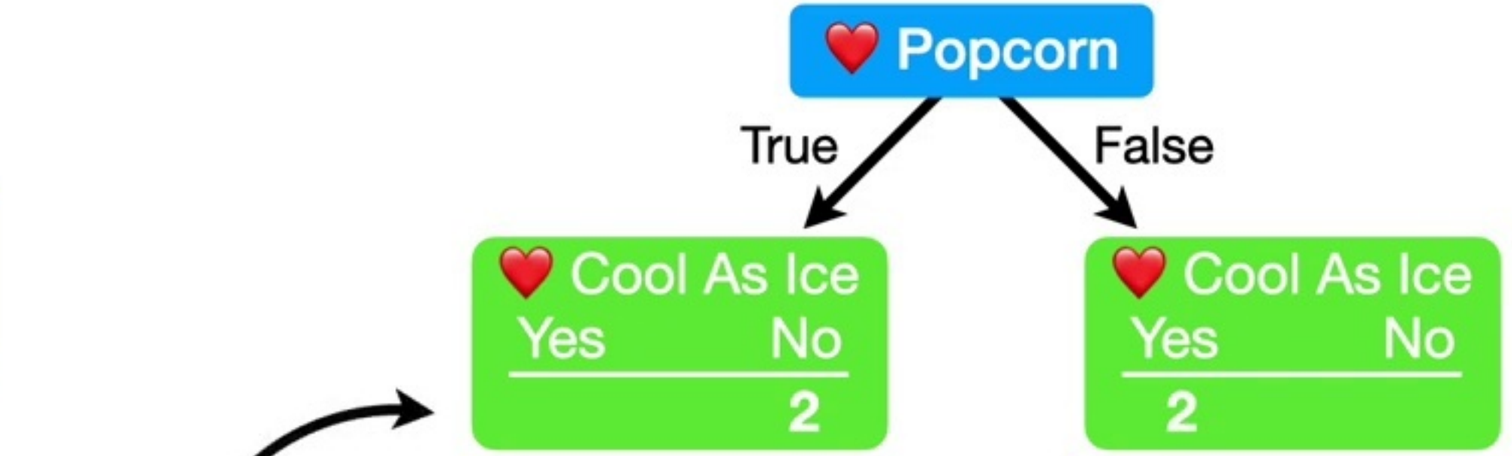
And because they **Love Cool As Ice**, we put a **1** under the word **Yes**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



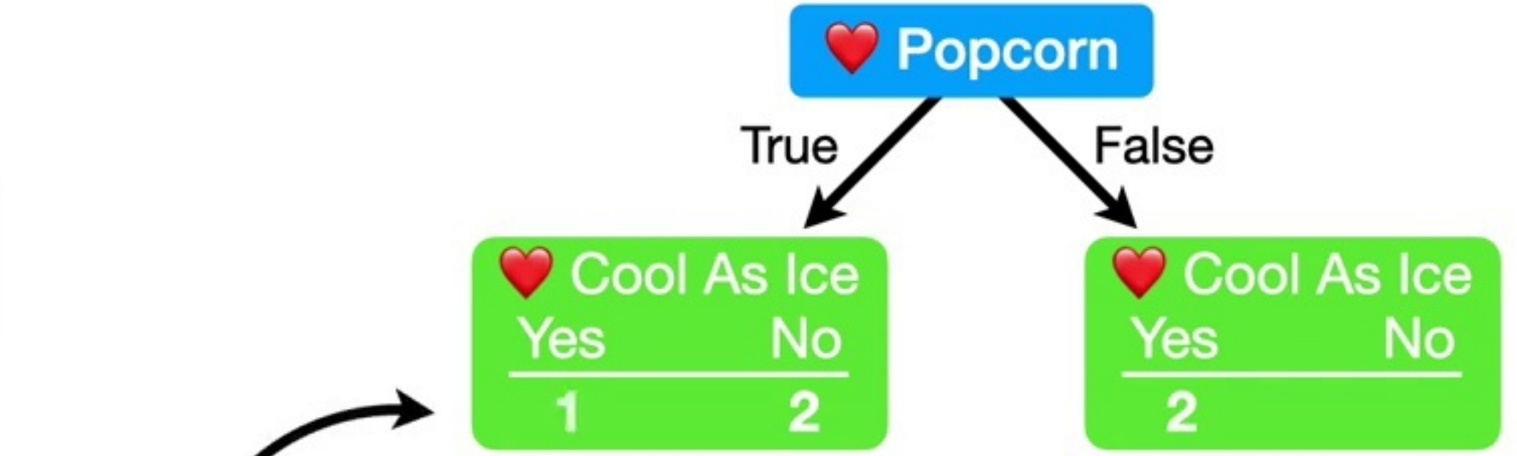
Likewise, we run the remaining rows down the tree, keeping track of whether or not each one **Loves Cool As Ice**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



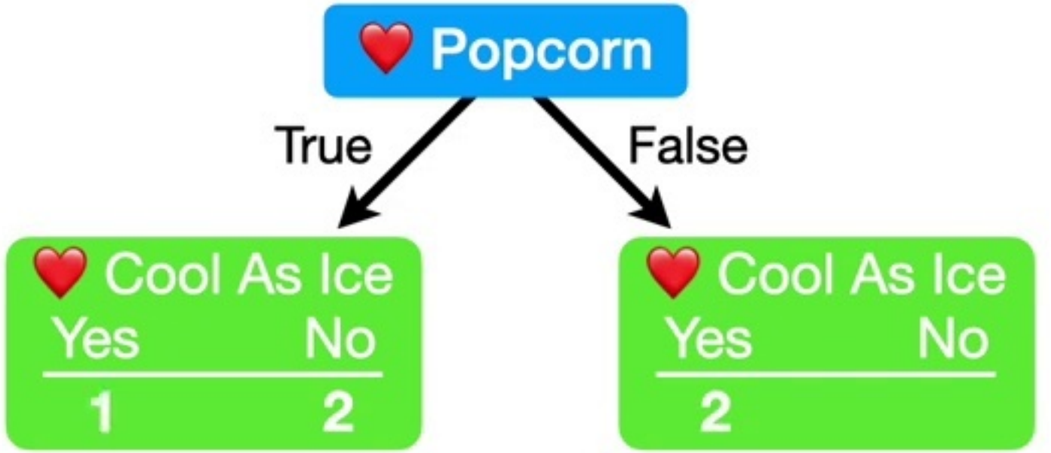
Likewise, we run the remaining rows down the tree, keeping track of whether or not each one **Loves Cool As Ice**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



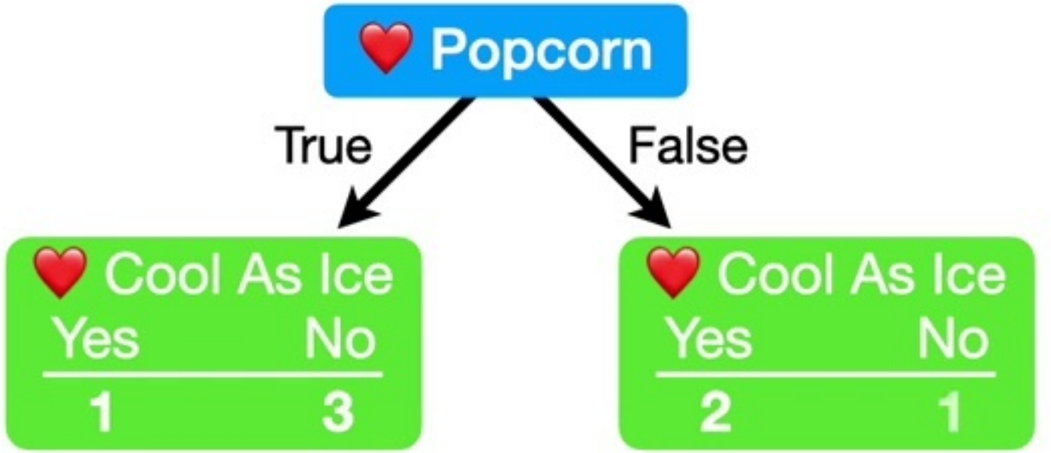
Likewise, we run the remaining rows down the tree, keeping track of whether or not each one **Loves Cool As Ice**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



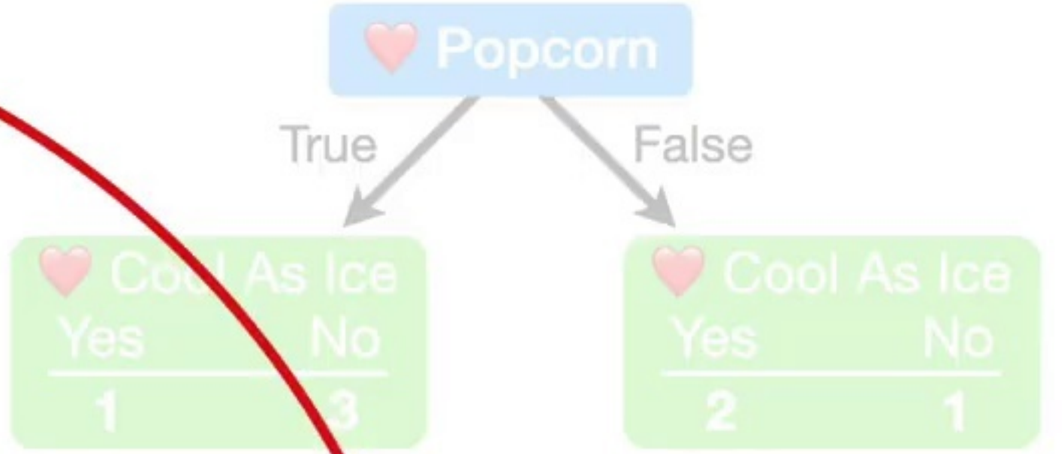
Likewise, we run the remaining rows down the tree, keeping track of whether or not each one **Loves Cool As Ice**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

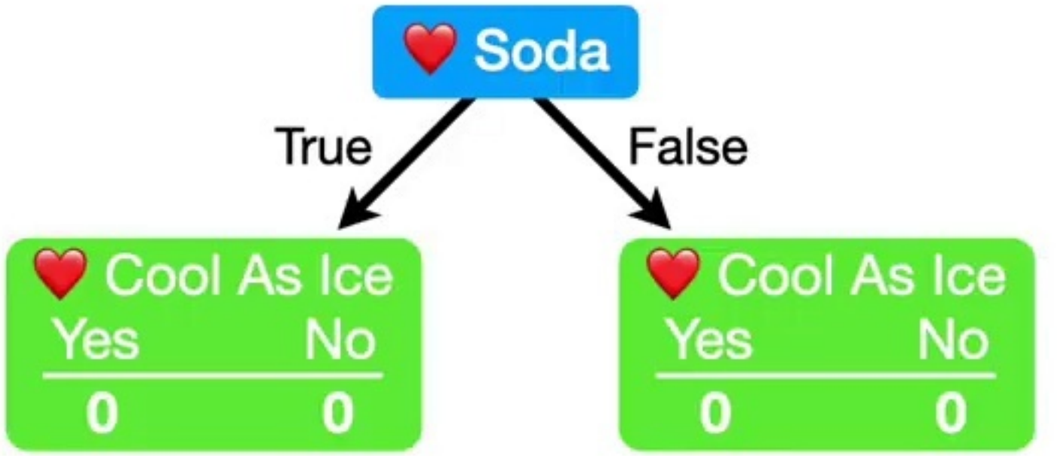


Likewise, we run the remaining rows down the tree, keeping track of whether or not each one **Loves Cool As Ice**.

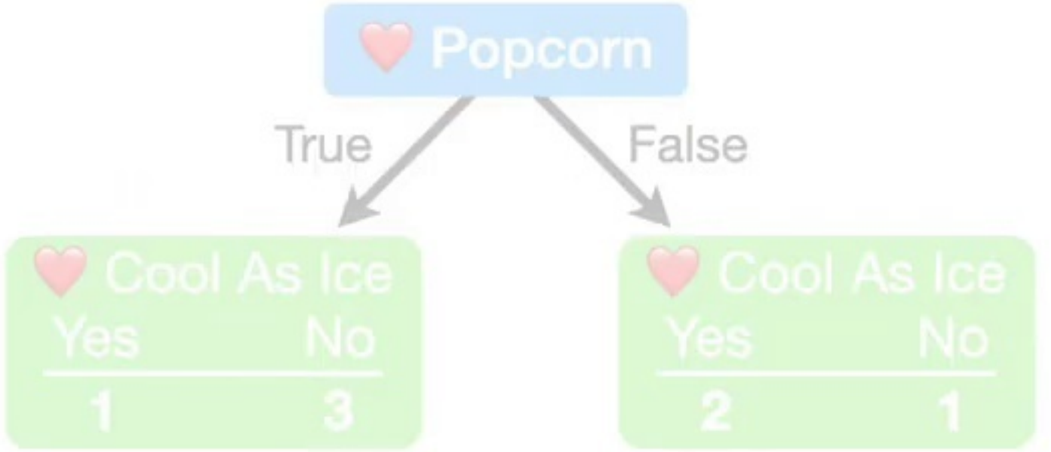
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



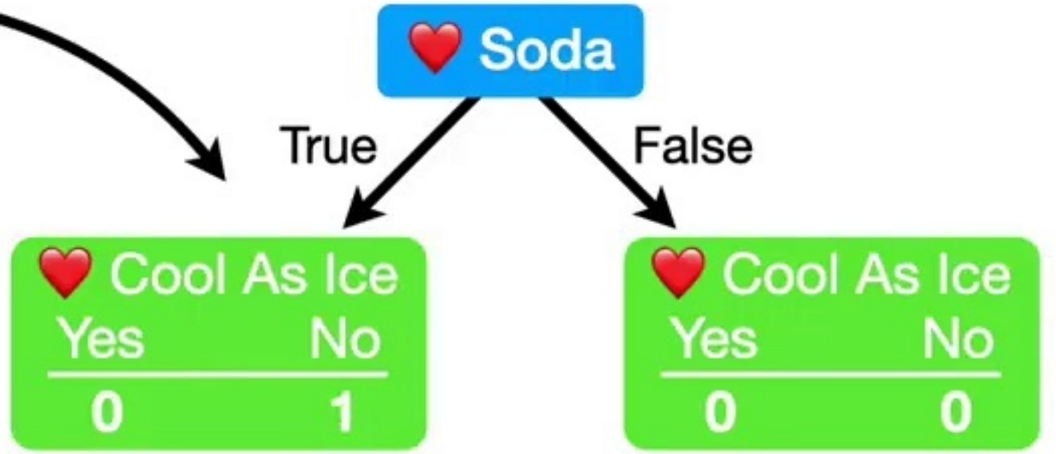
Now let's do the exact same thing for **Loves Soda**.



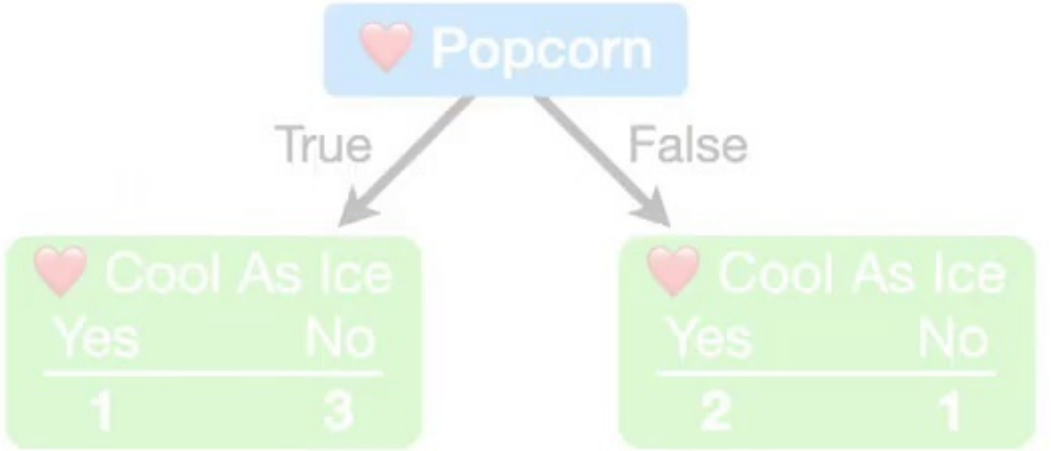
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



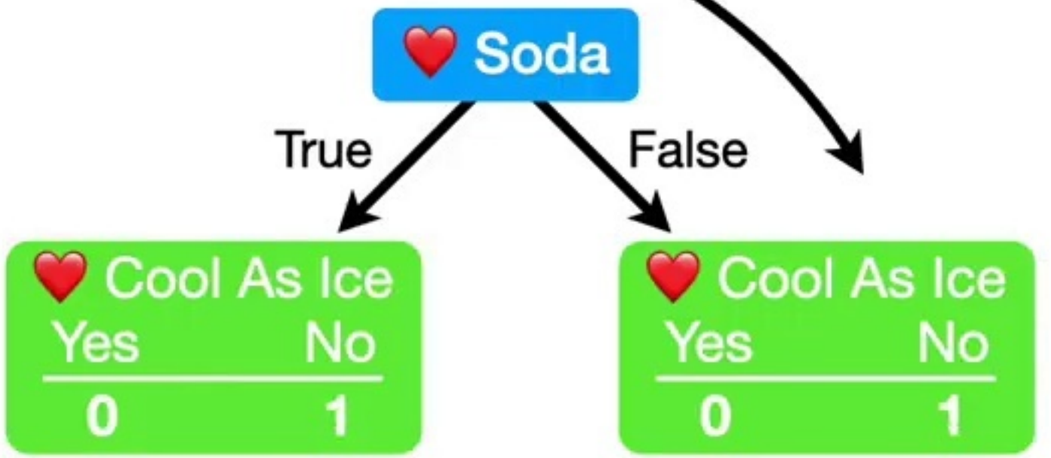
Now let's do the exact same thing for **Loves Soda**.



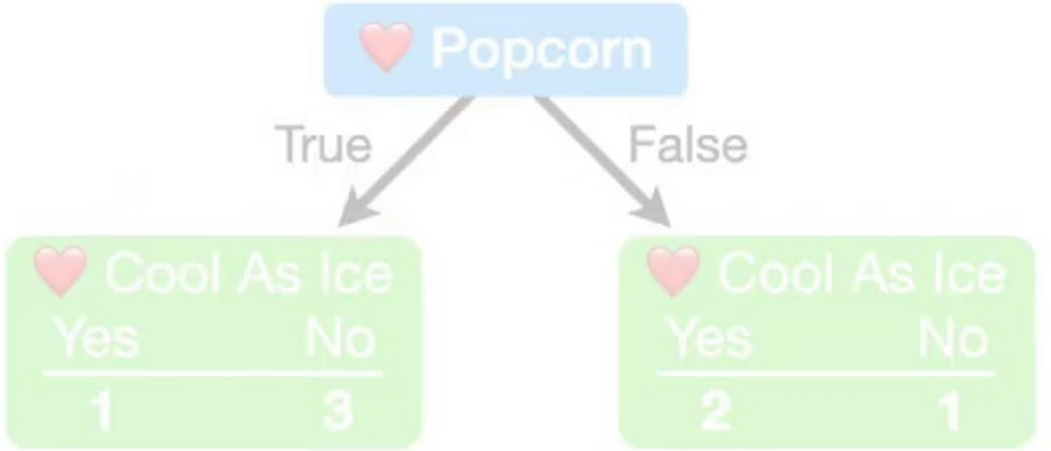
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



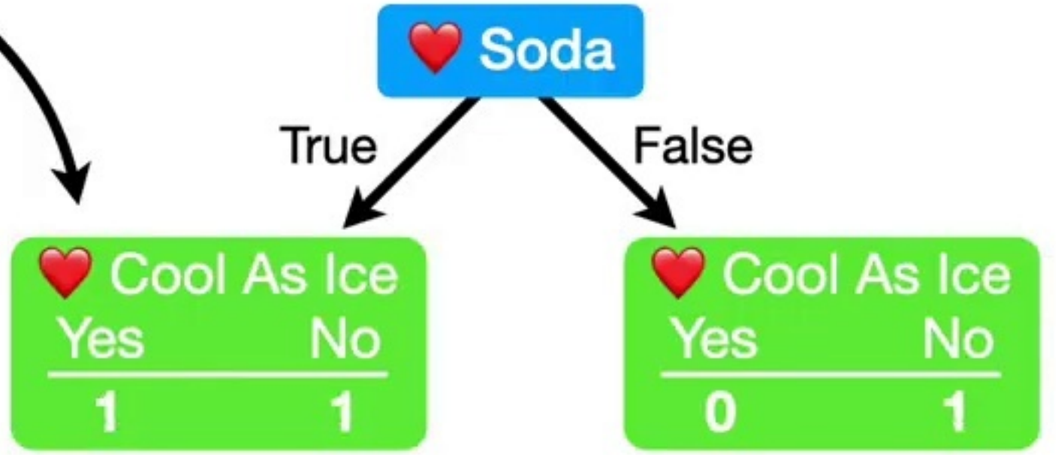
Now let's do the exact same thing for **Loves Soda**.



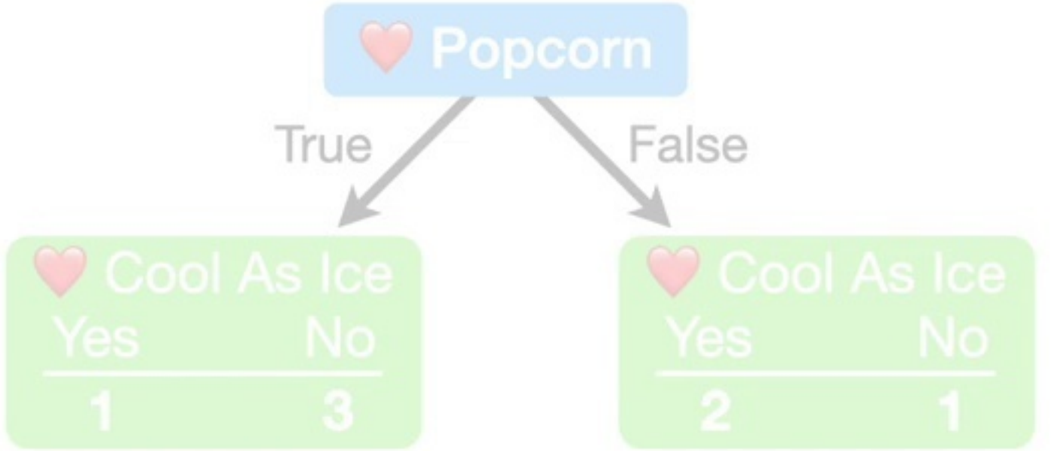
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



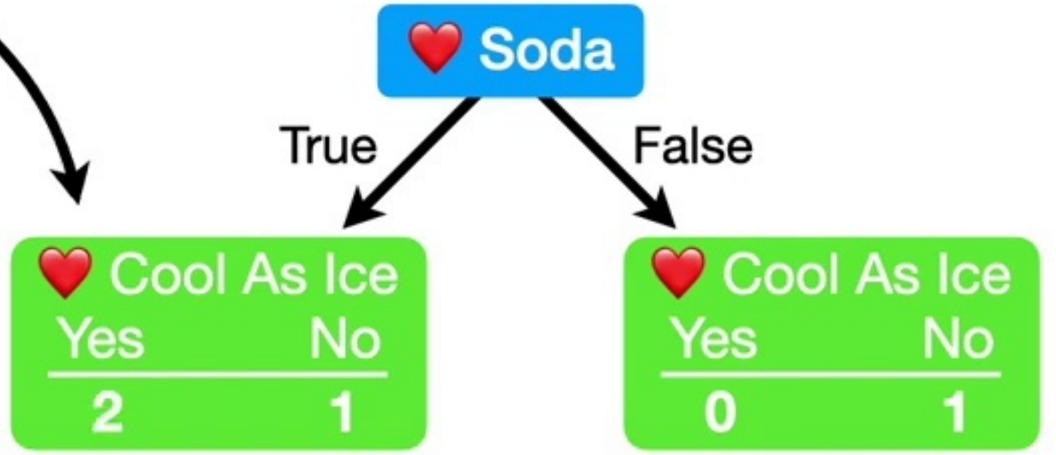
Now let's do the exact same thing for **Loves Soda**.



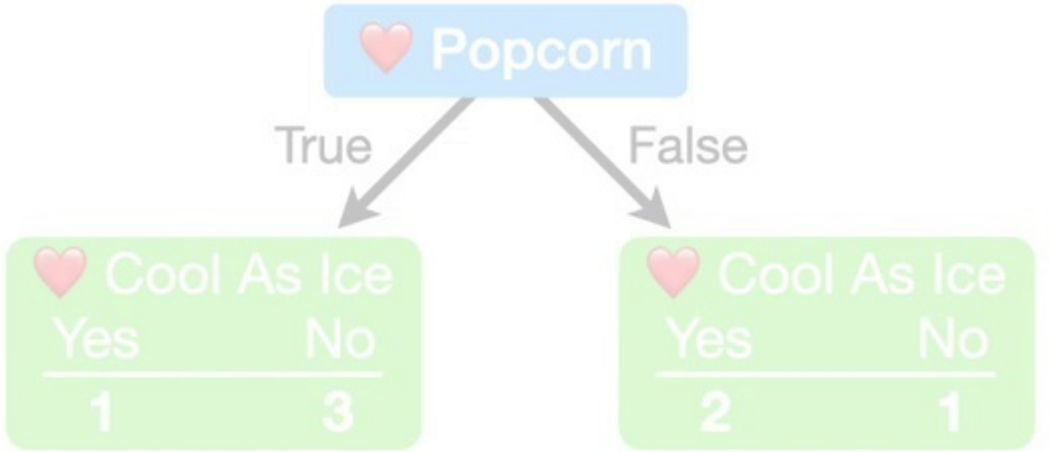
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



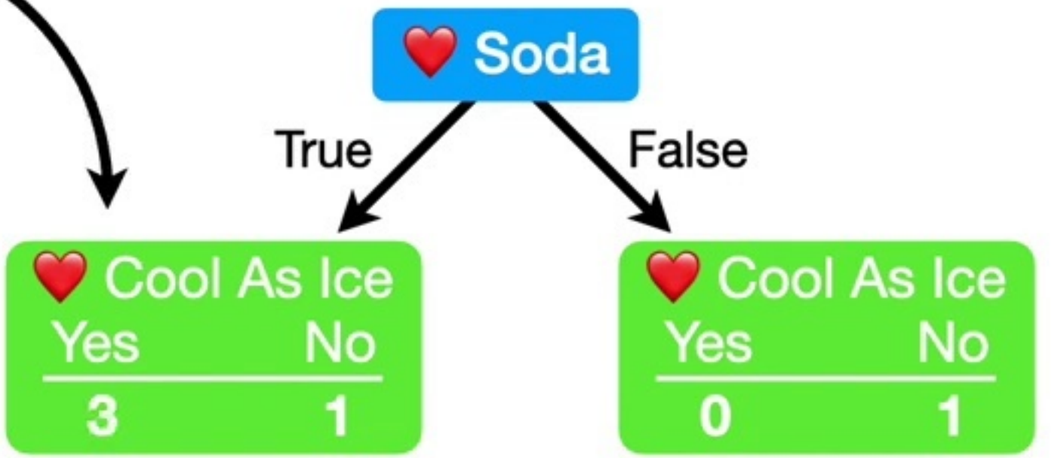
Now let's do the exact same thing for **Loves Soda**.



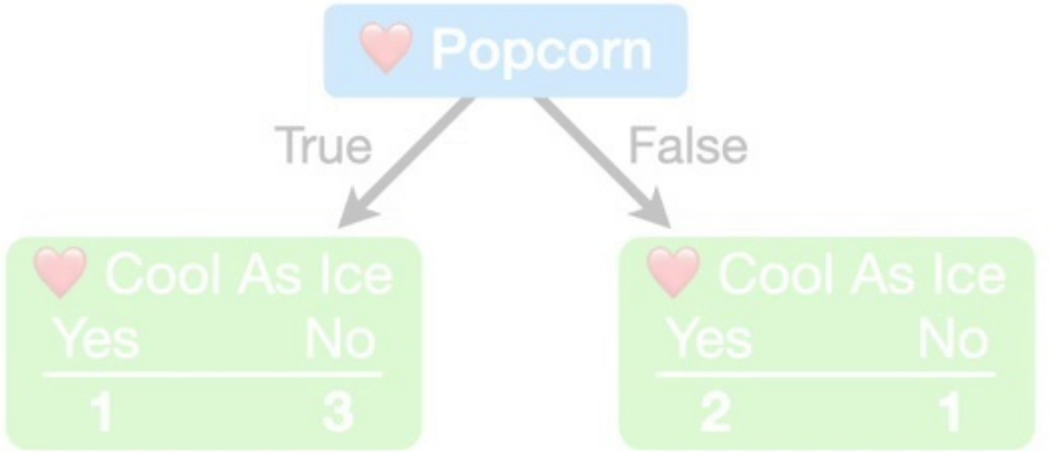
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



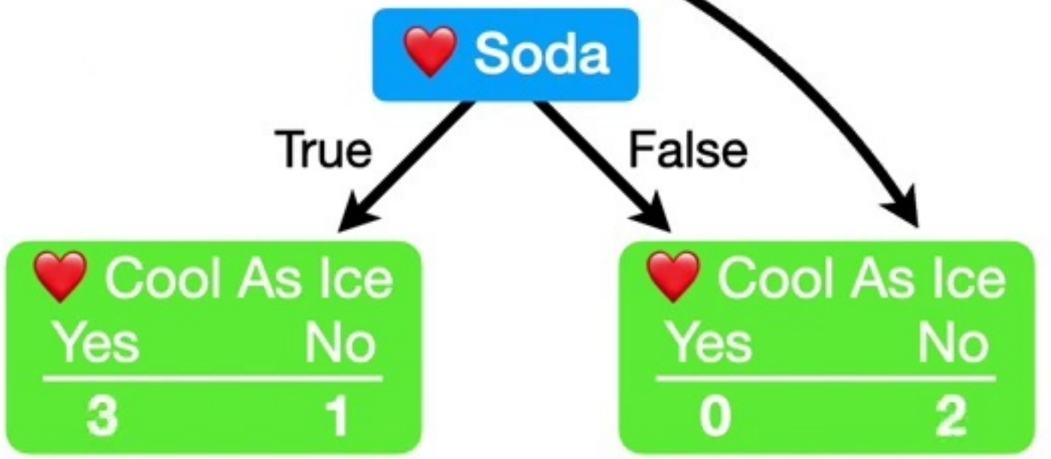
Now let's do the exact same thing for **Loves Soda**.



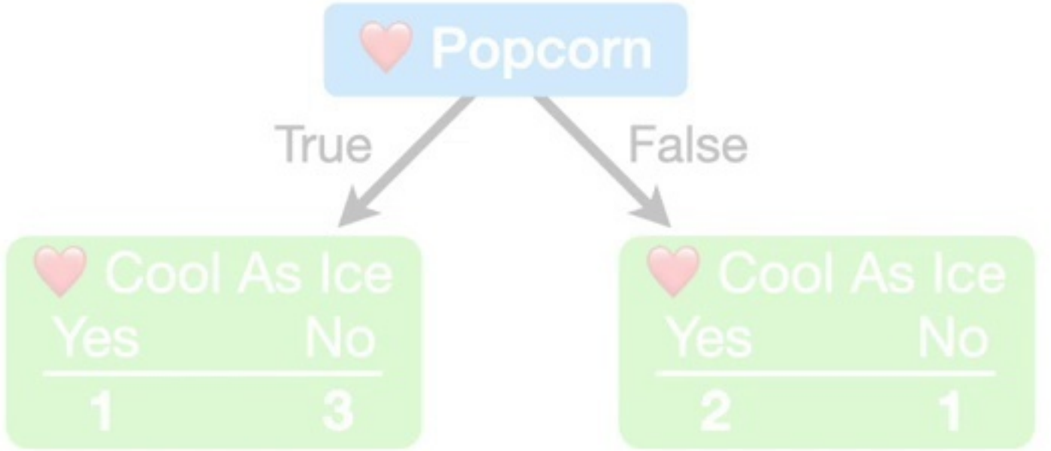
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



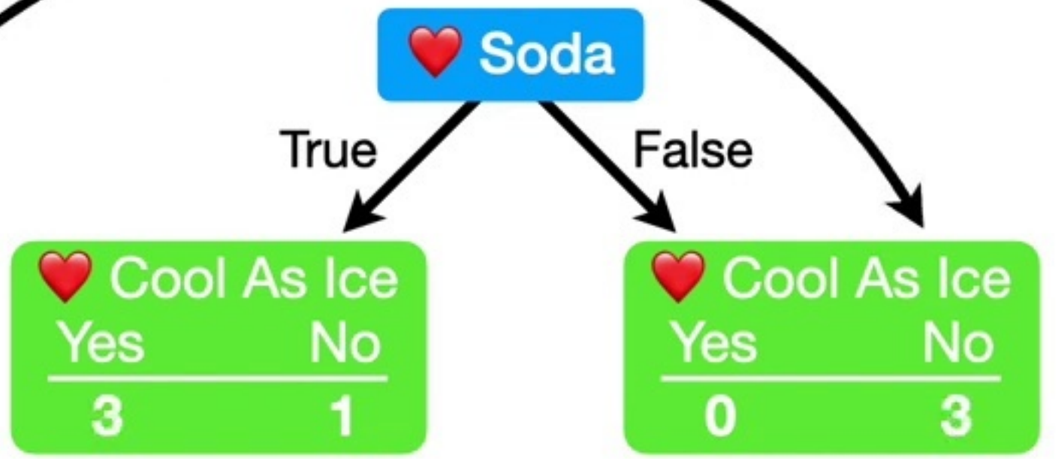
Now let's do the exact same thing for **Loves Soda**.



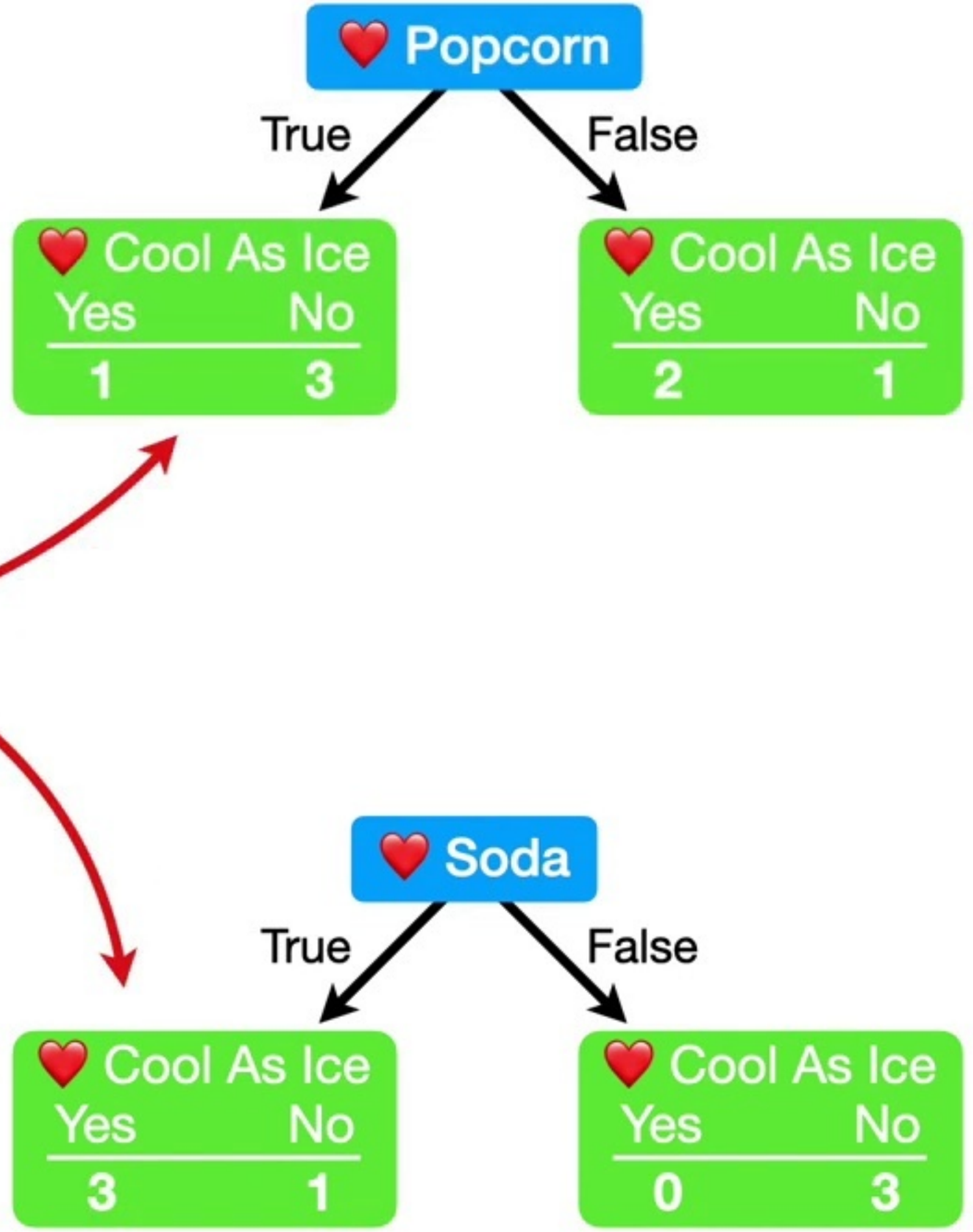
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

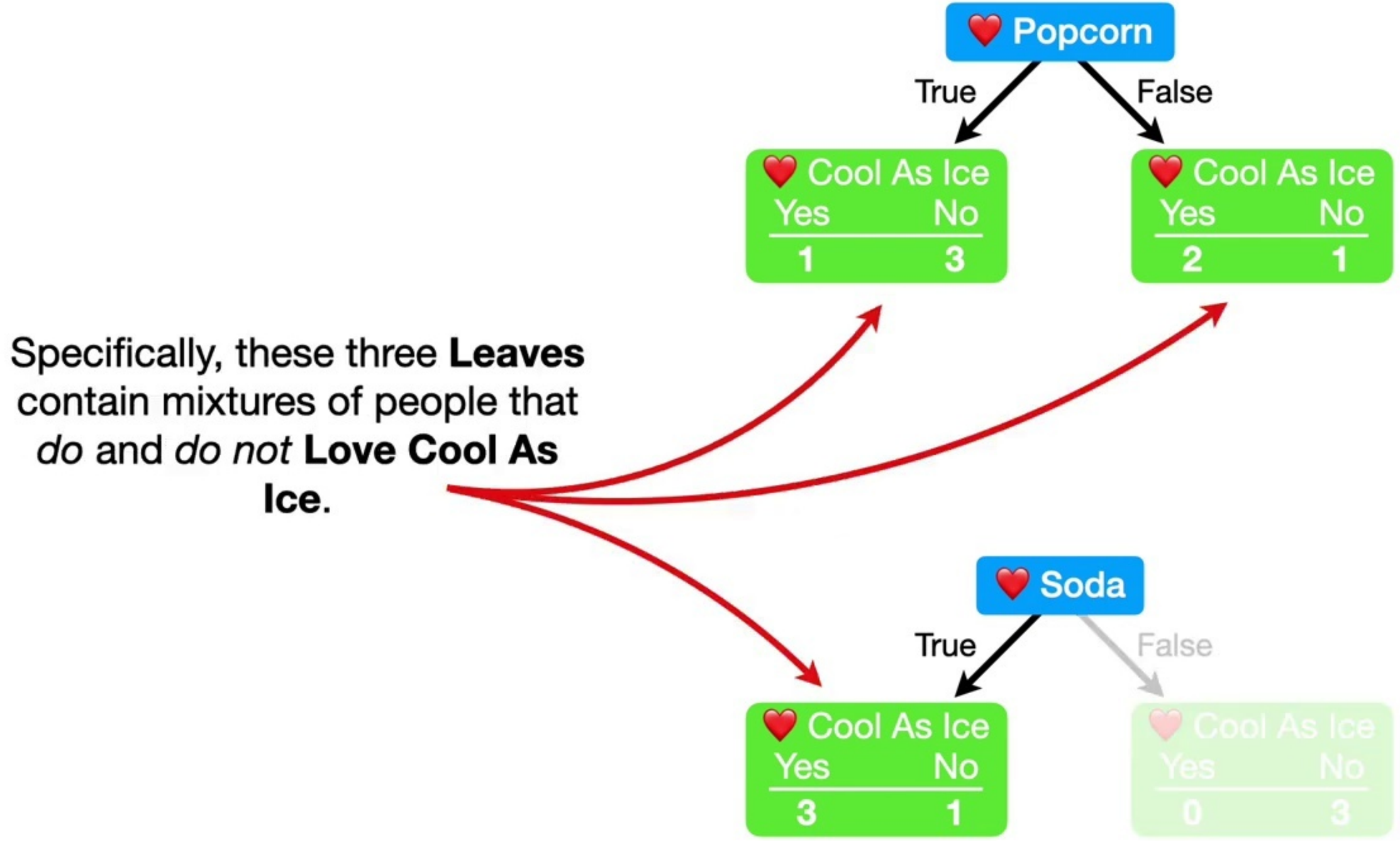


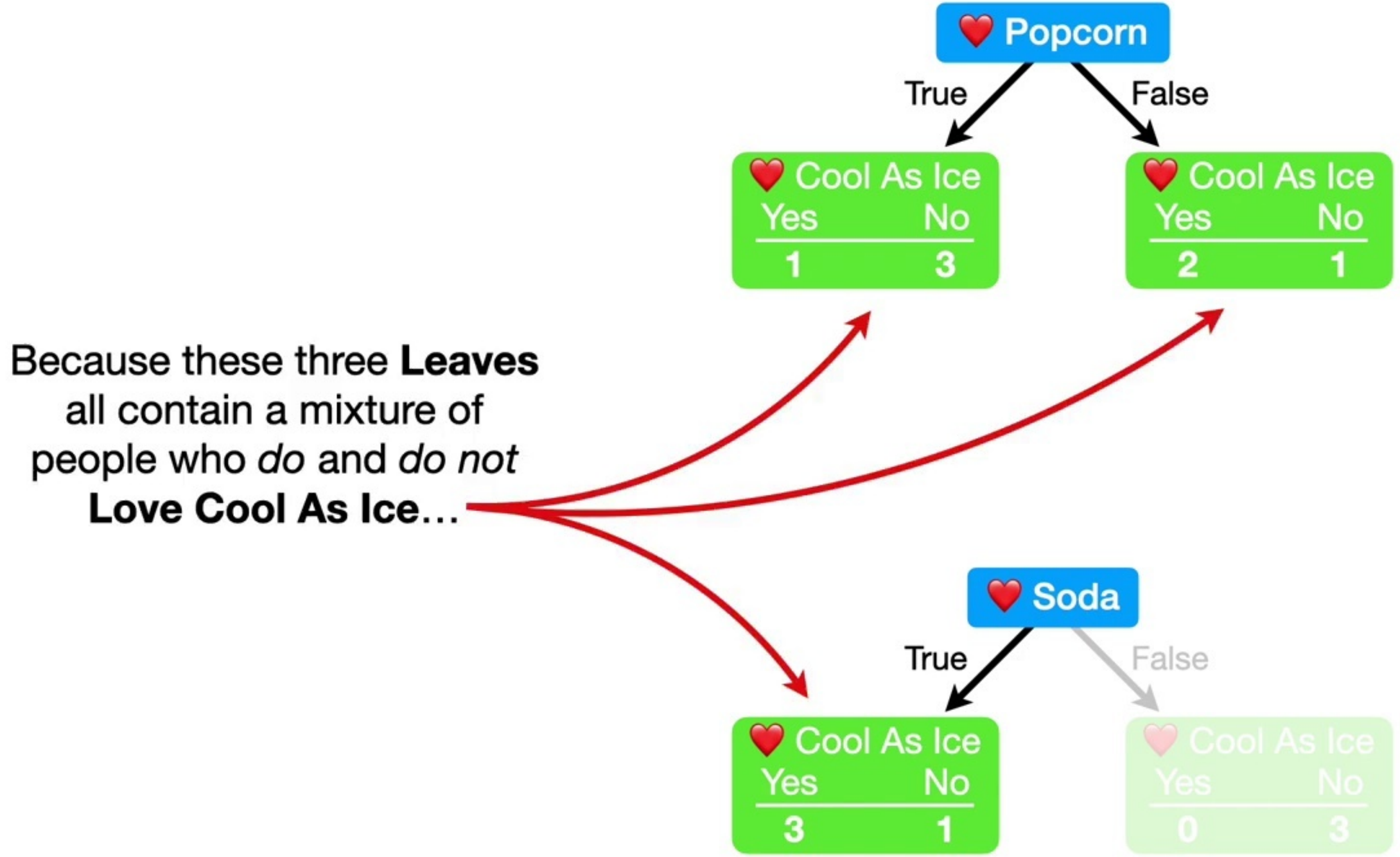
Now let's do the exact same thing for **Loves Soda**.

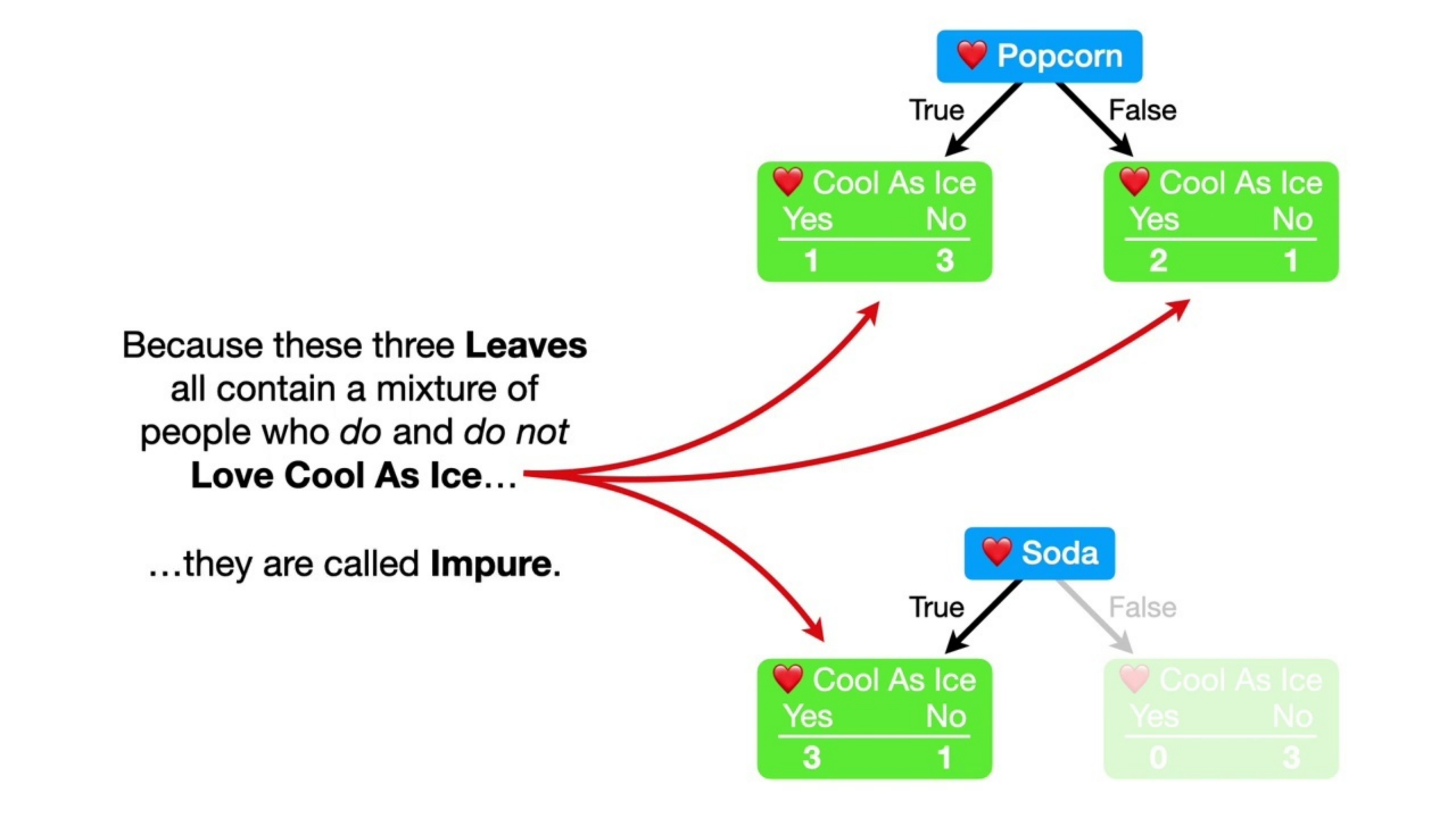


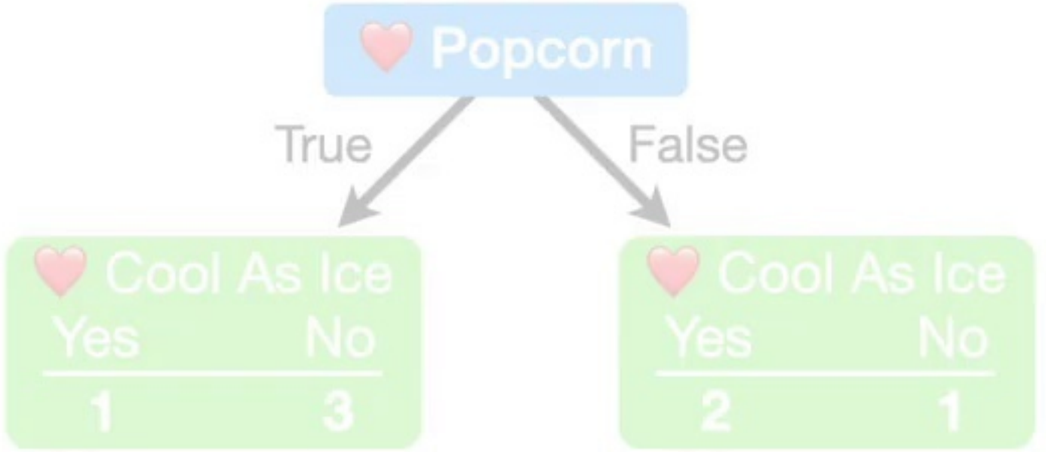
Looking at the two little trees, we see that neither one does a perfect job predicting who *will* and who *will not* **Love Cool As Ice**.



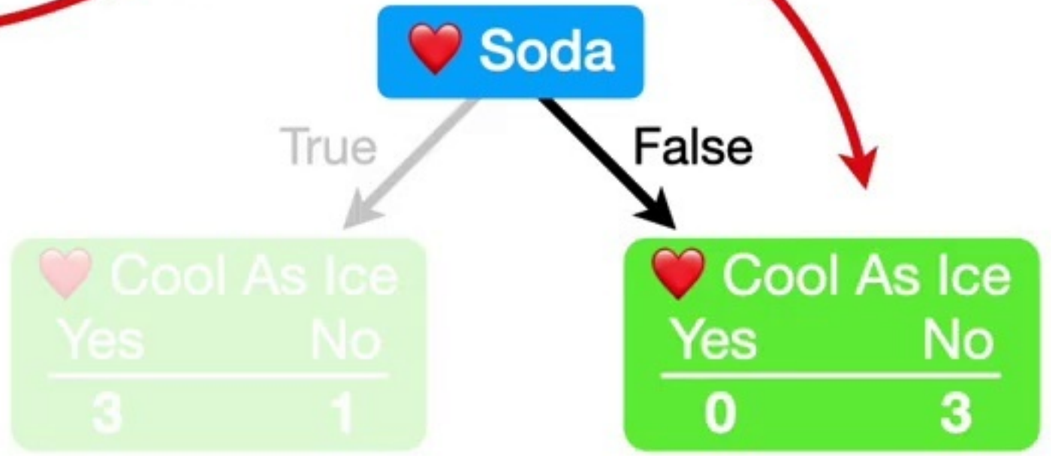

Popcorn
True
False
Cool As Ice
Yes No
1 3
Cool As Ice
Yes No
2 1
Specifically, these three **Leaves** contain mixtures of people that *do* and *do not* **Love Cool As Ice**.
Soda
True
False
Cool As Ice
Yes No
3 1
Cool As Ice
Yes No
0 3

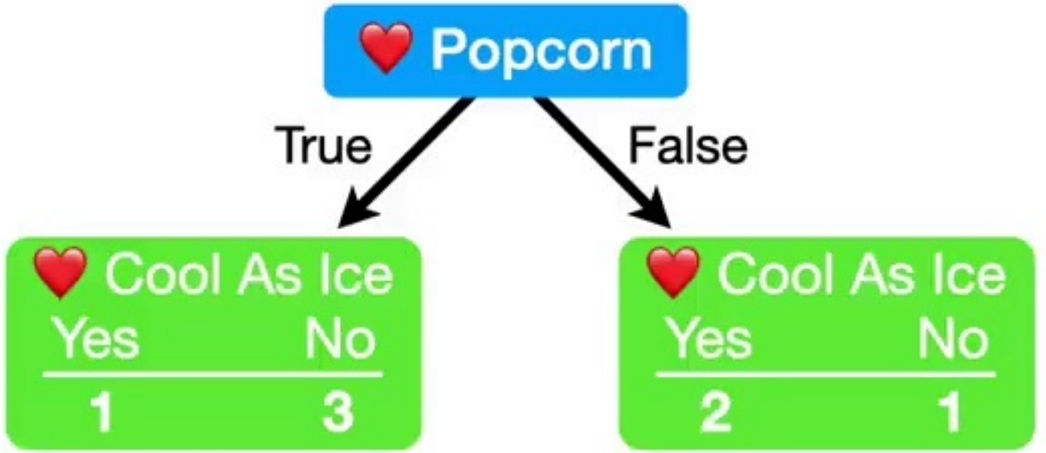

Popcorn
True
False
Cool As Ice
Yes No
1 3
Cool As Ice
Yes No
2 1
Because these three **Leaves** all contain a mixture of people who *do* and *do not* **Love Cool As Ice**…
Soda
True
False
Cool As Ice
Yes No
3 1
Cool As Ice
Yes No
0 3


Popcorn
True
False
Cool As Ice
Yes No
1 3
Cool As Ice
Yes No
2 1
Because these three **Leaves** all contain a mixture of people who *do* and *do not* **Love Cool As Ice**…
…they are called **Impure**.
Soda
True
False
Cool As Ice
Yes No
3 1
Cool As Ice
Yes No
0 3

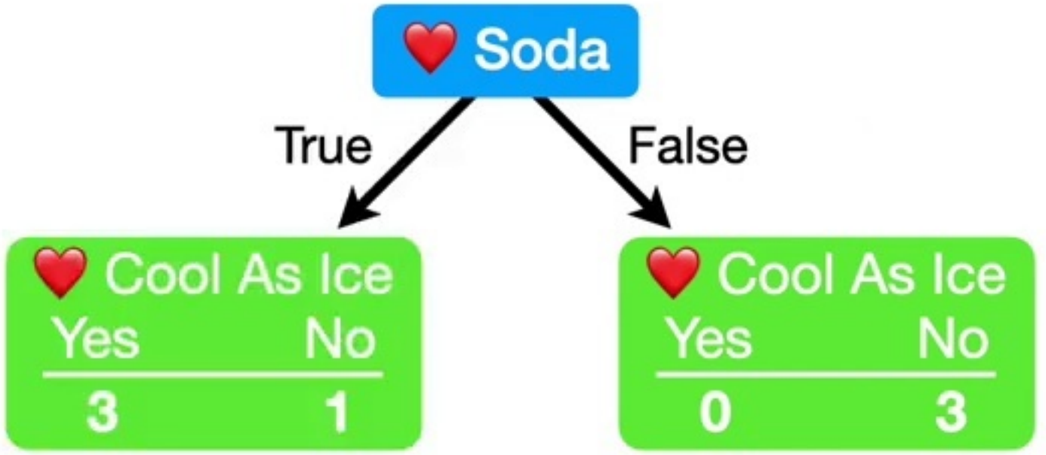


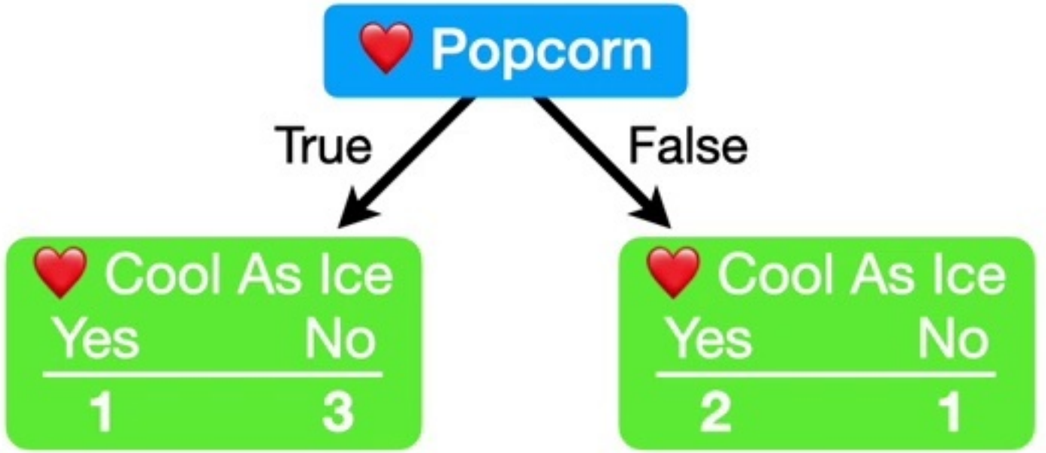
In contrast, this **Leaf** only contains people who *do not* **Love Cool As Ice**.



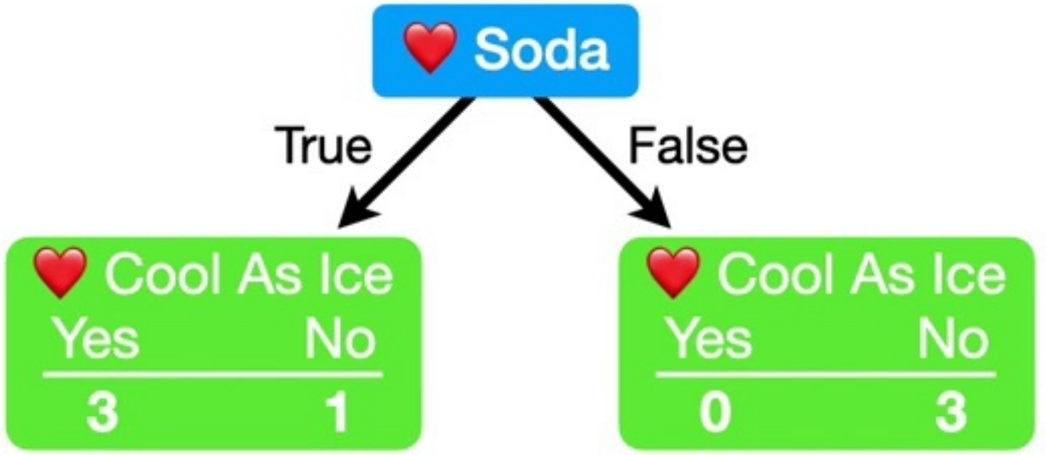


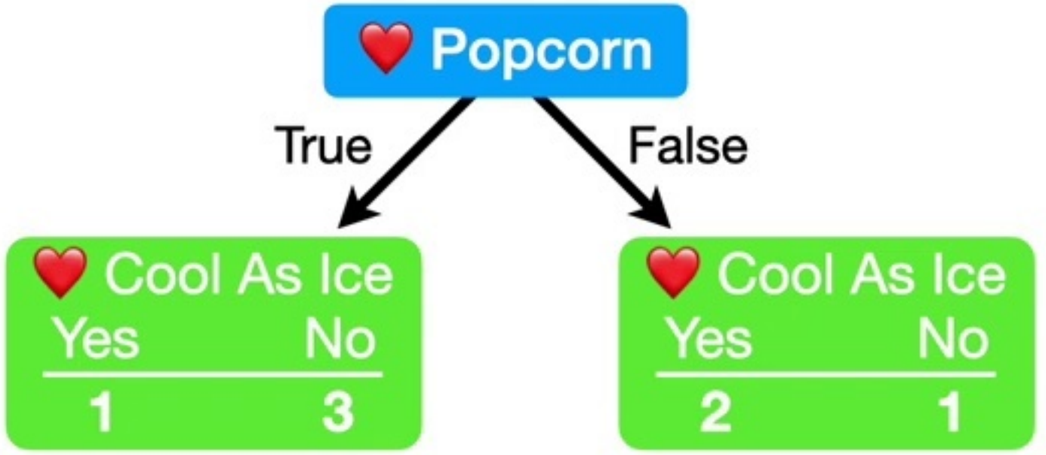
Because both **Leaves** in the **Loves Popcorn** tree are **Impure**…



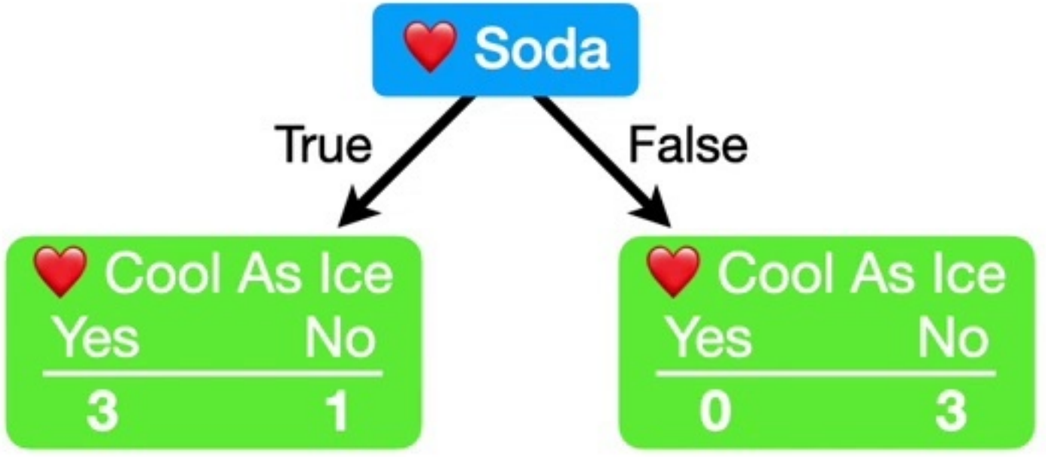


…and only one **Leaf** in the **Loves Soda** tree is **Impure**…

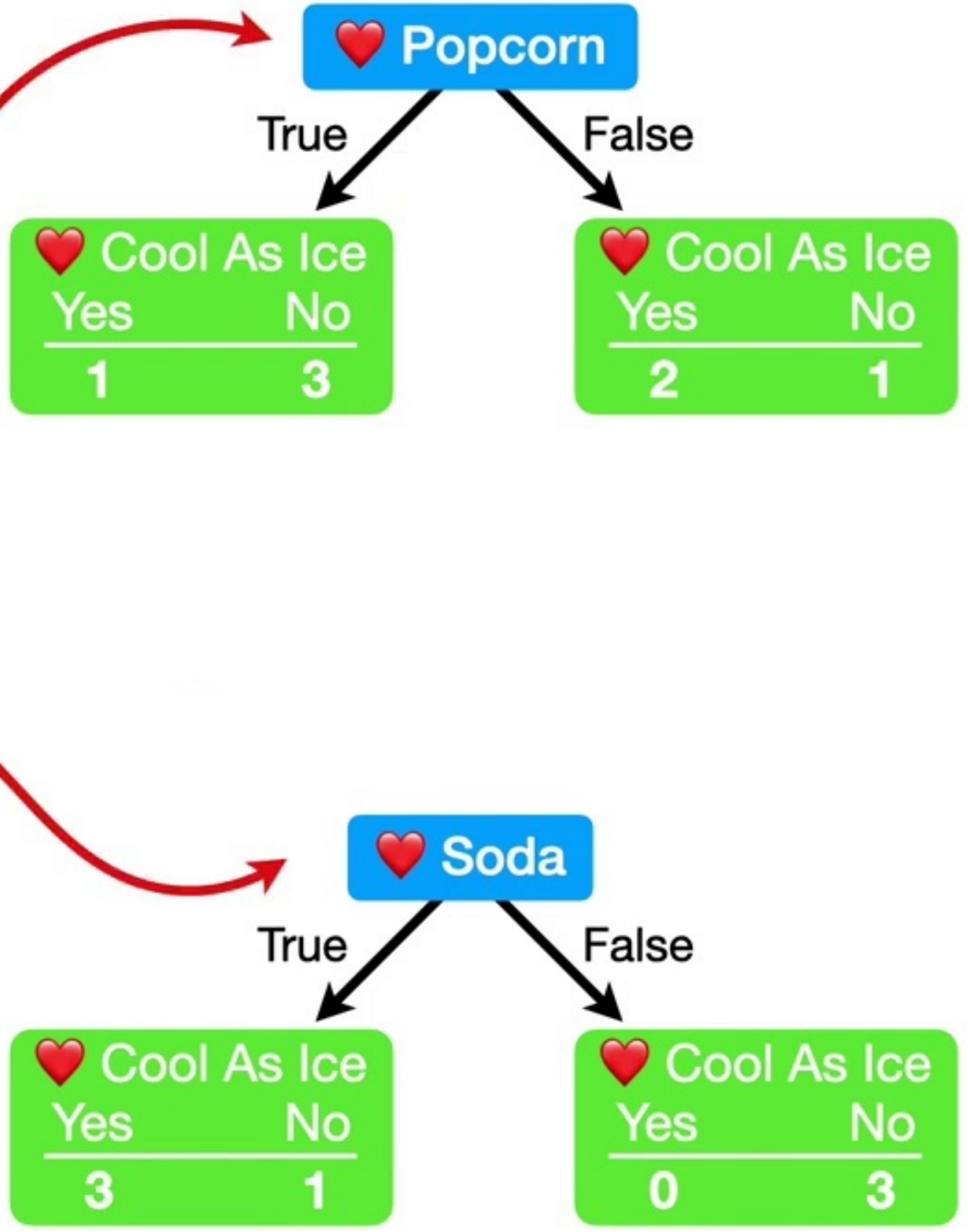


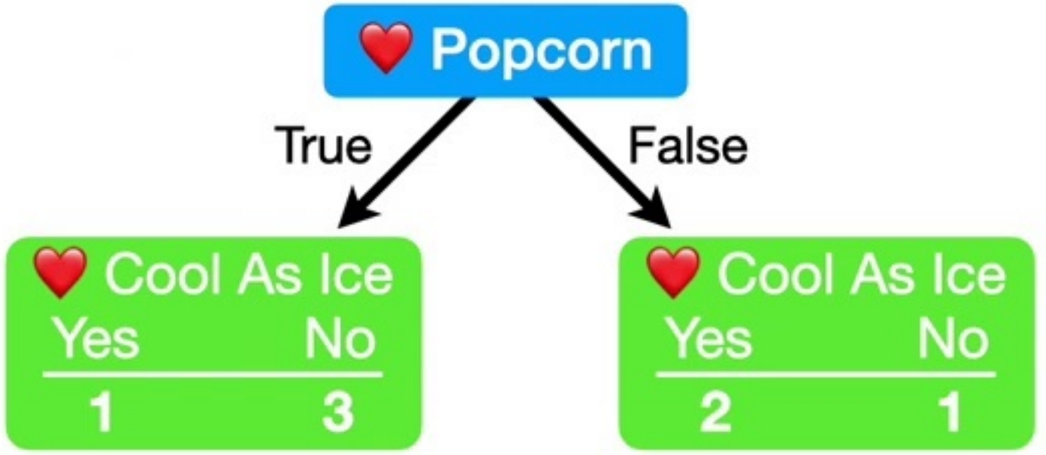


…it seems like **Loves Soda** does a better job predicting who *will* and *will not* **Love Cool As Ice**…

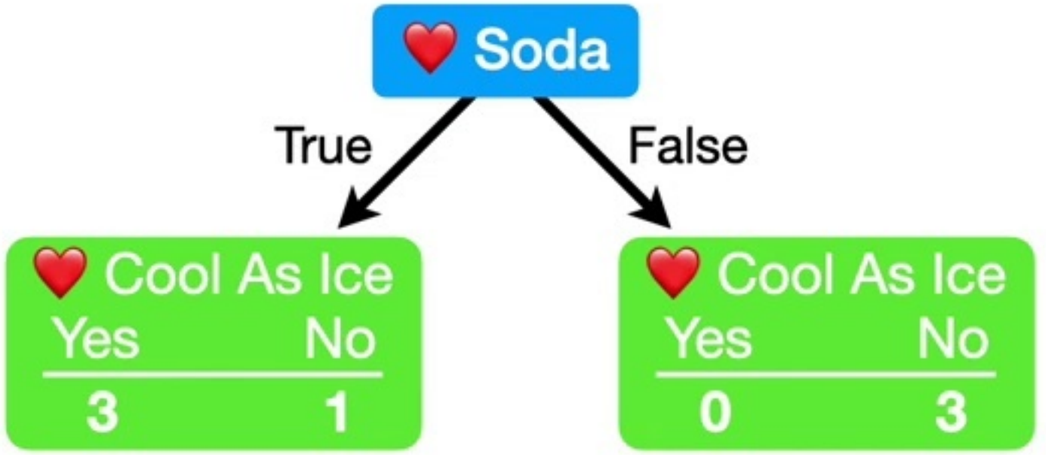


…but it would be nice if we could quantify the differences between **Loves Popcorn** and **Loves Soda**.

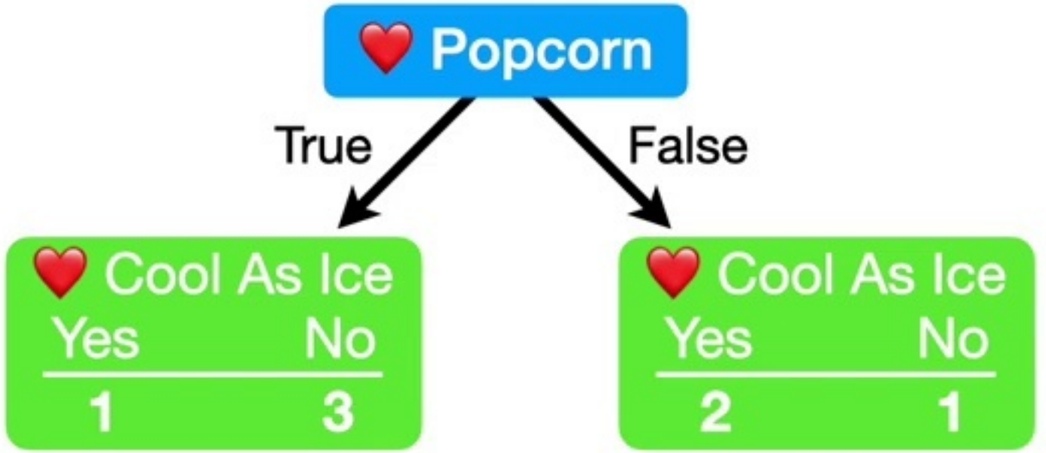


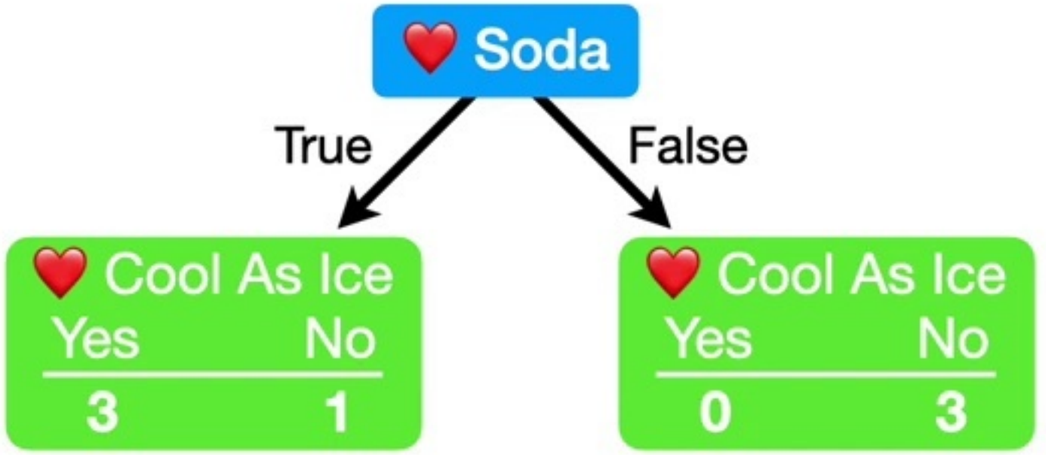


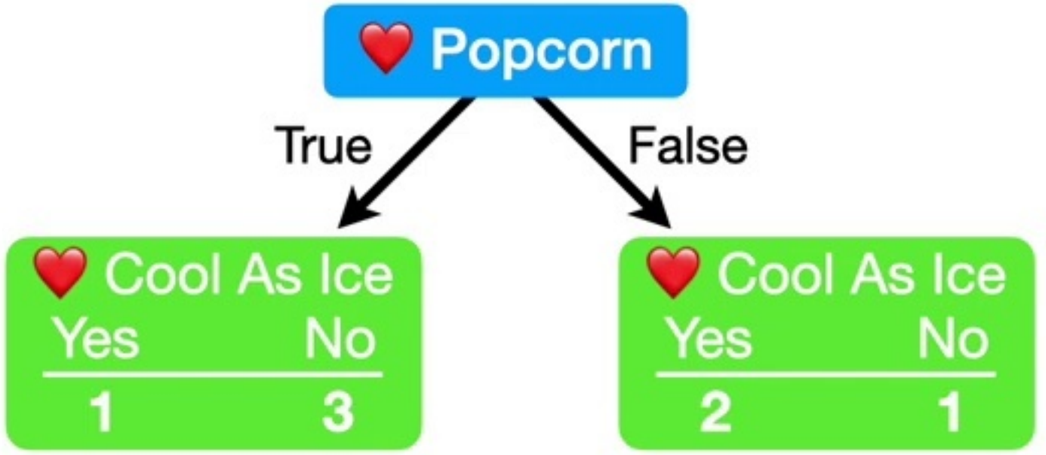
The good news is that there are several ways to quantify the **Impurity** of the leaves.



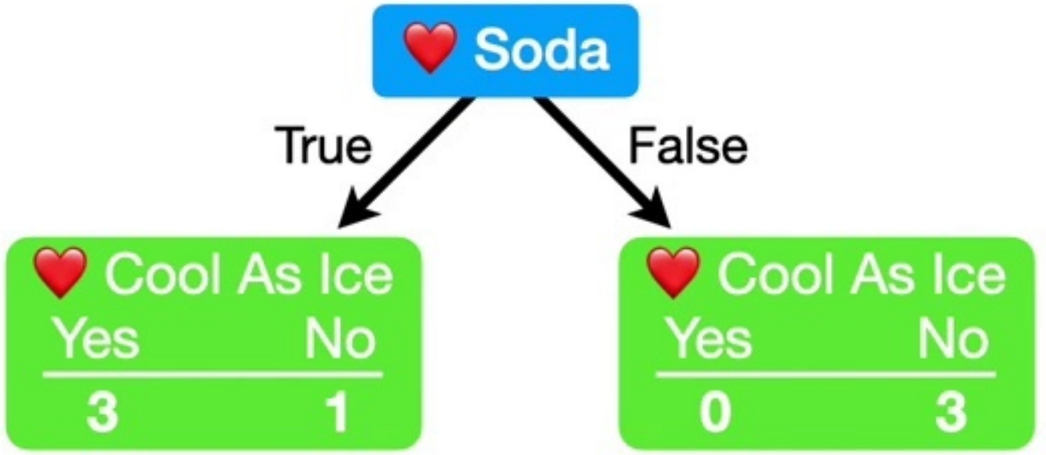
One of the most popular methods is called **Gini Impurity**, but there are also fancy sounding methods like **Entropy** and **Information Gain**.





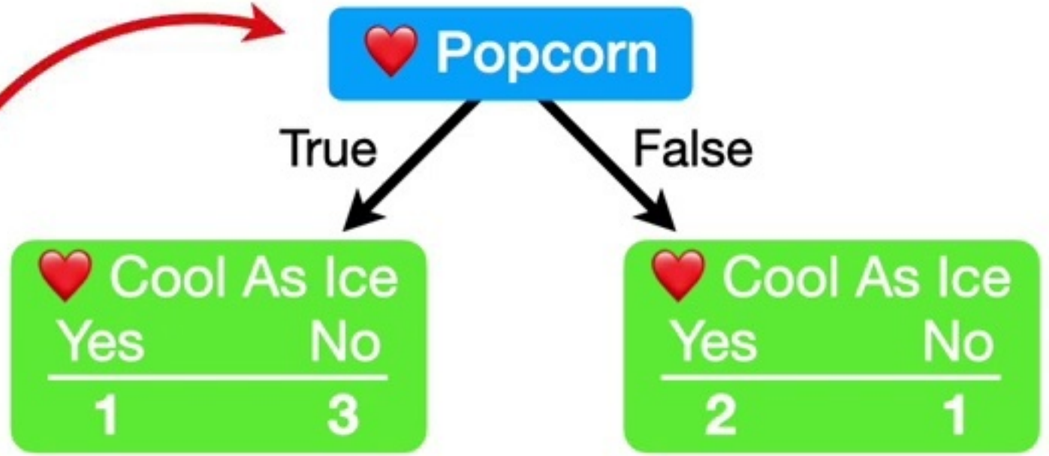


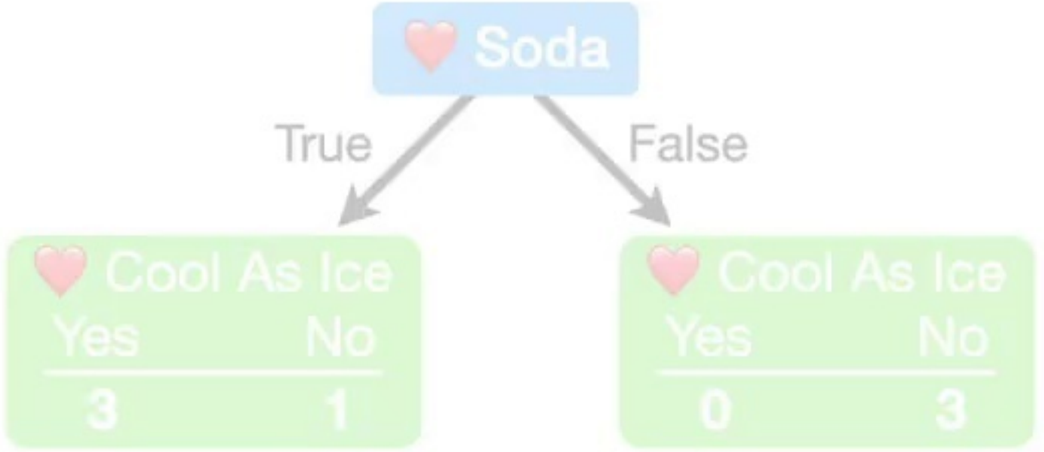
However, numerically, the methods are all quite similar*, so we will focus on **Gini Impurity** since, not only is it very popular, I think it is the most straightforward.

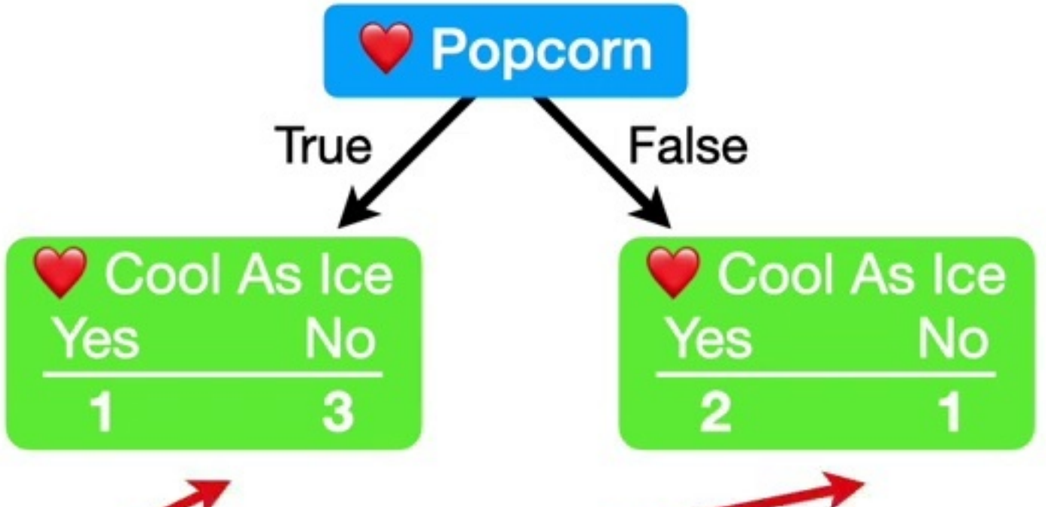


* For details, see page 321 of the Introduction to Statistical Learning in R.

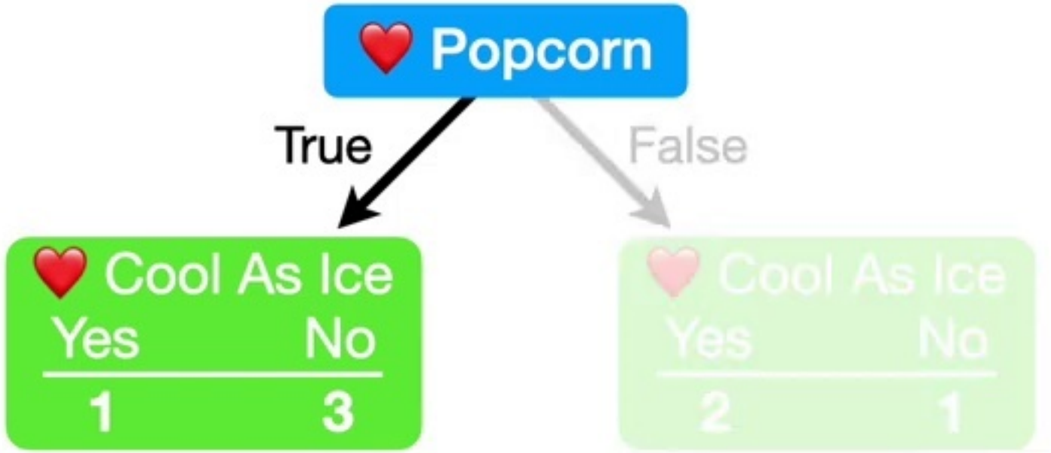
So let's start by calculating the **Gini Impurity** for **Loves Popcorn**.





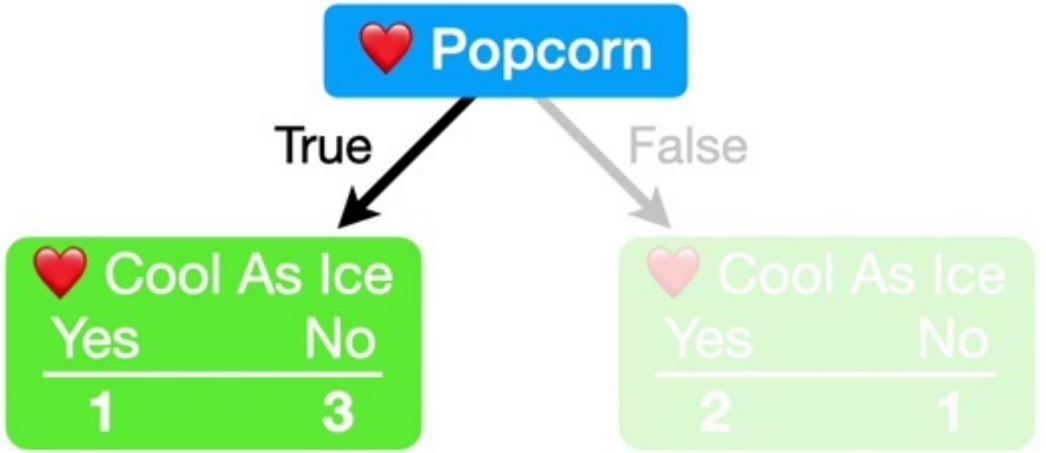


To calculate the **Gini Impurity** for **Loves Popcorn**, we start by calculating the **Gini Impurity** for the individual **Leaves**.



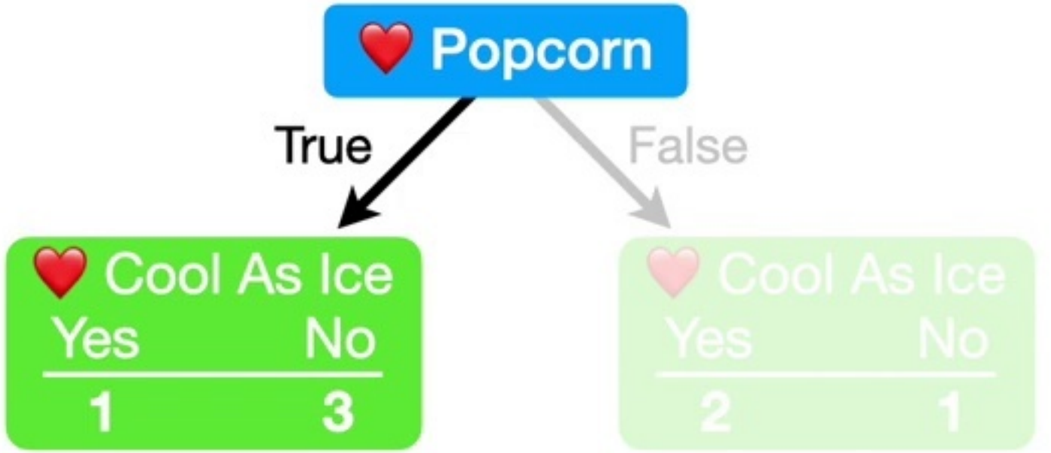
The **Gini Impurity** for the **Leaf** on the left is…

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2



The **Gini Impurity** for the **Leaf** on the left is…

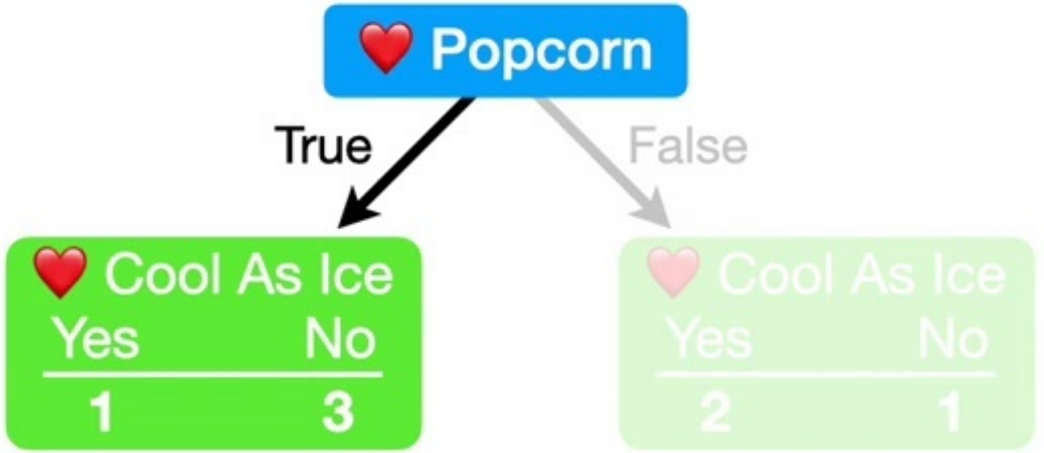
Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2



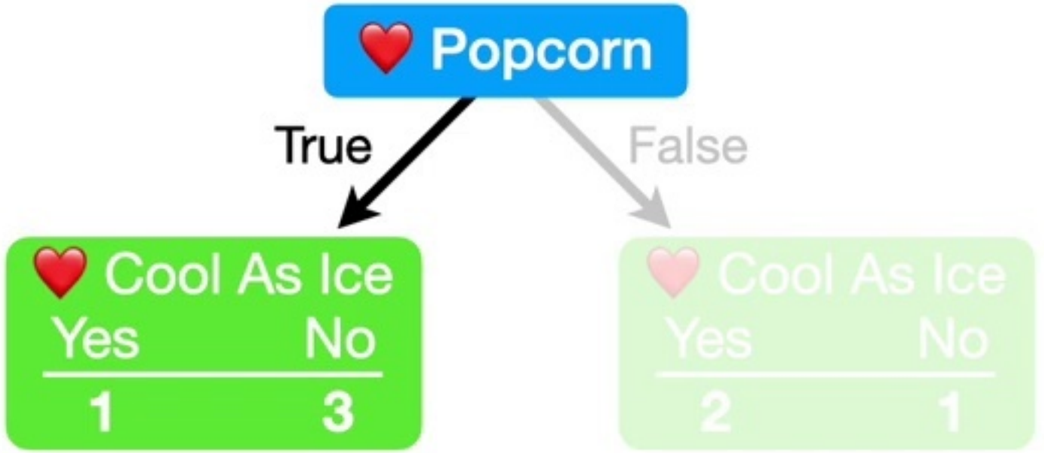
The **Gini Impurity** for the **Leaf** on the left is…

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

The **Gini Impurity** for the **Leaf** on the left is…



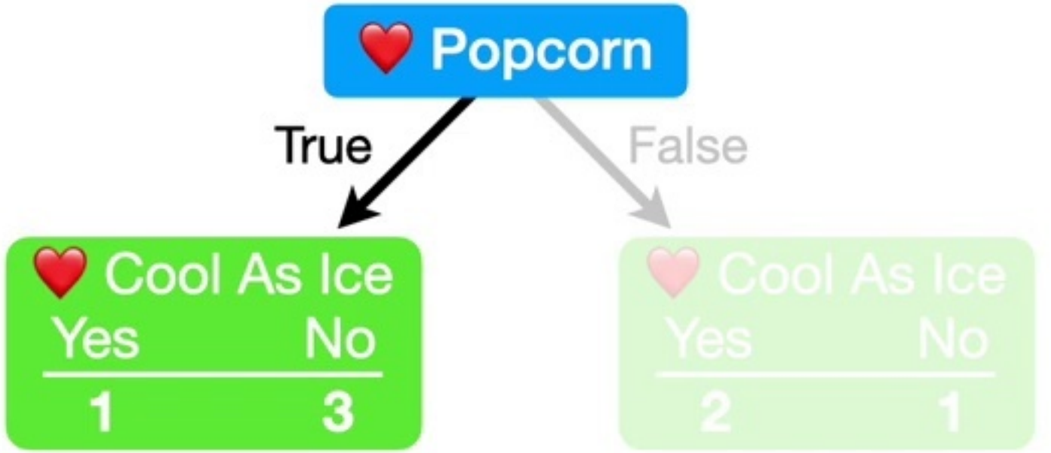
Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

= 1

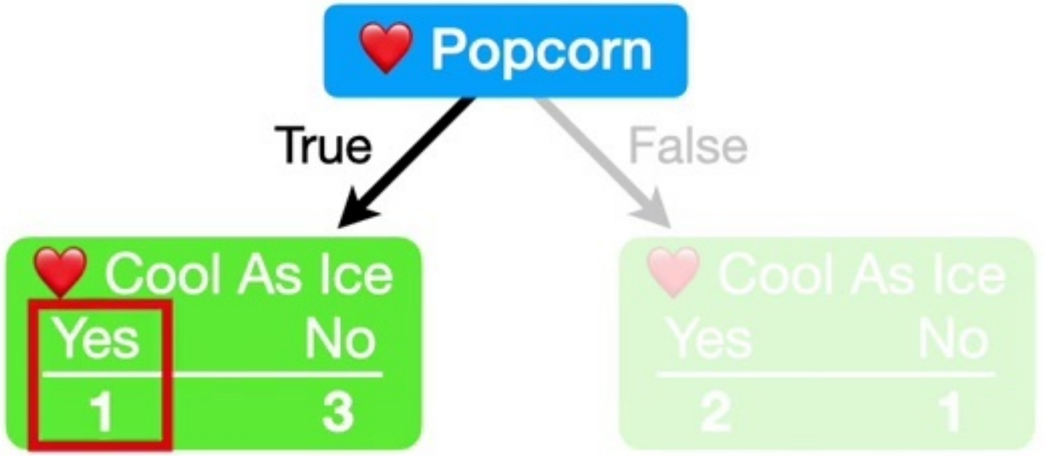
So we start out with **1**…



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

= 1

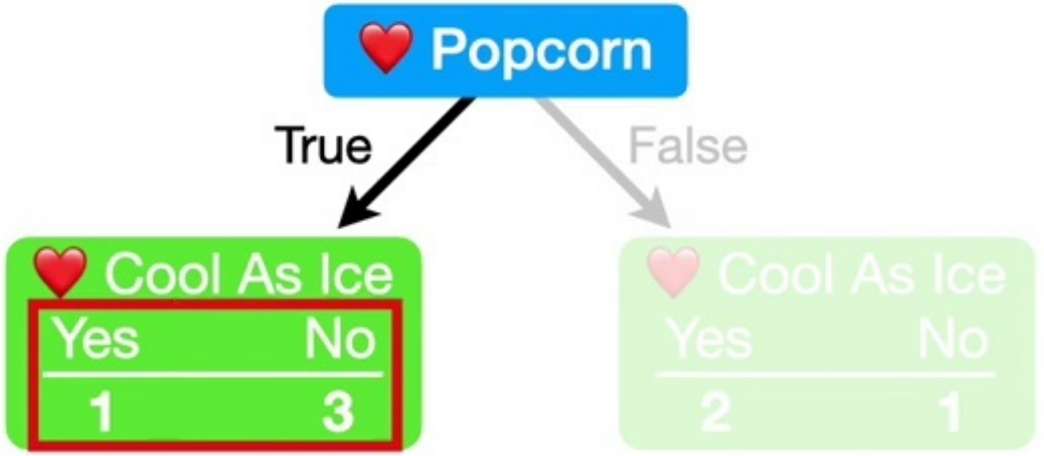
…then we subtract the squared probability of someone in this **Leaf Loving Cool As Ice**…



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2$$

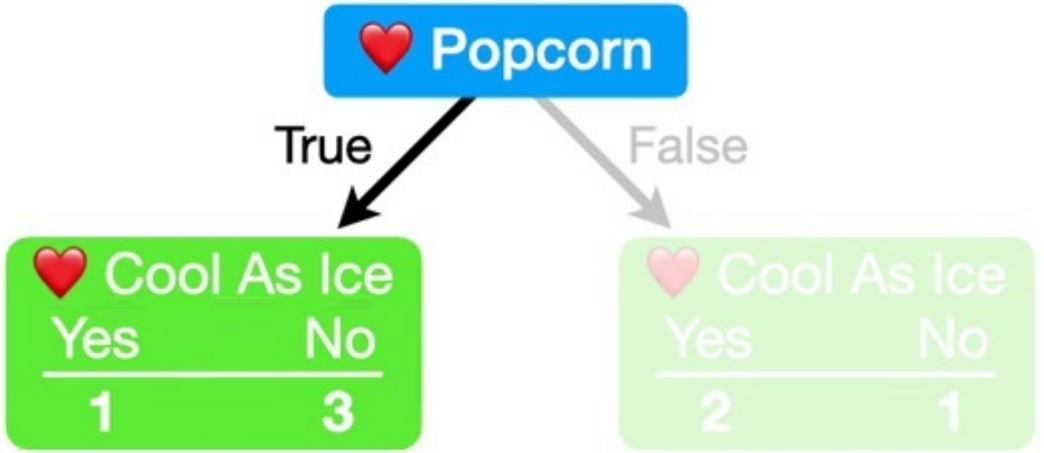
…which is **1**, the number of people in the **Leaf** who **Loved Cool As Ice**…



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2$$

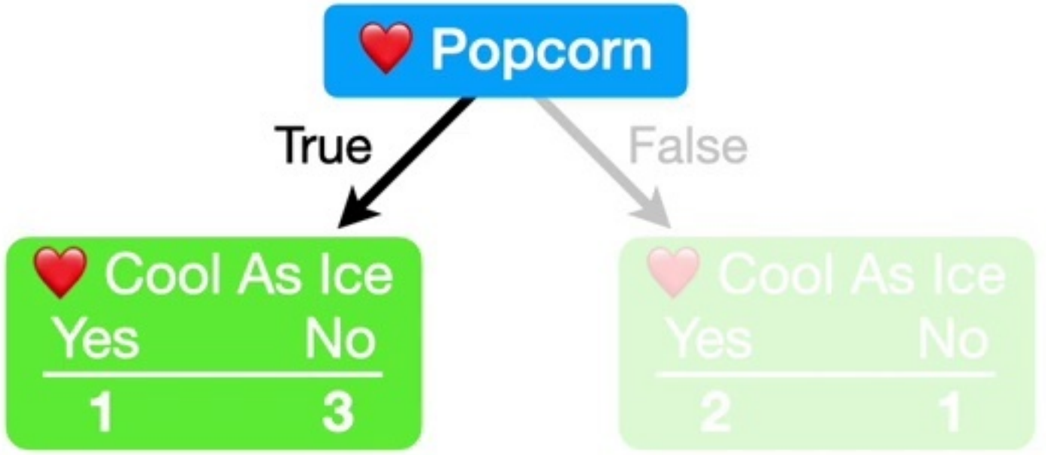
…divided by the total number of people in the **Leaf**, **4**…



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2$$

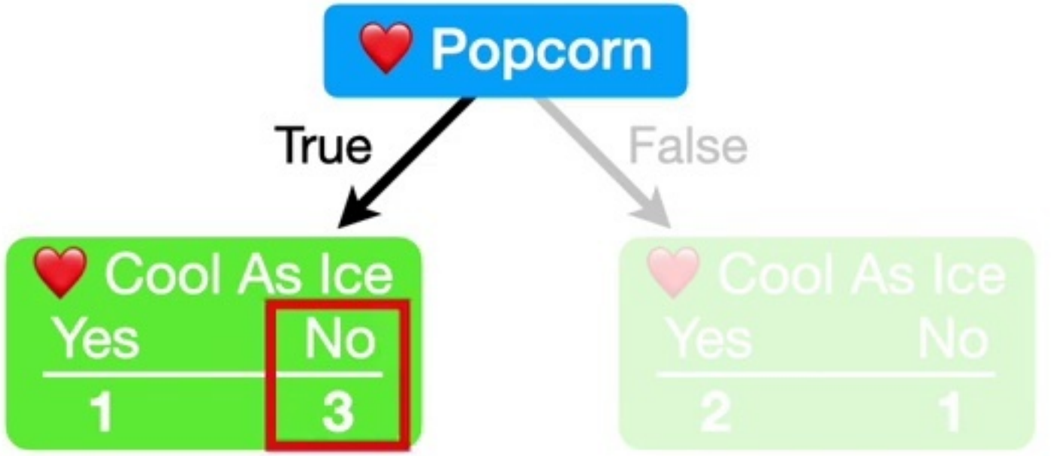
…squared.



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2$$

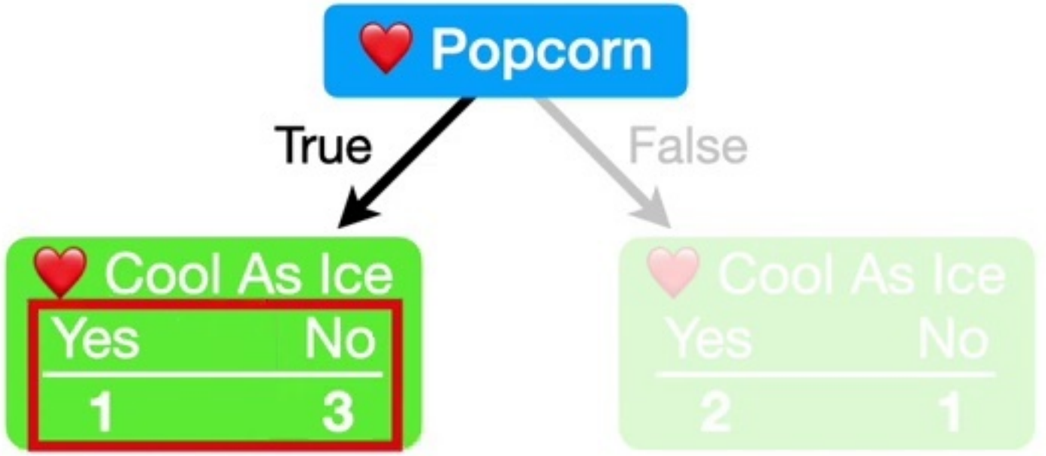
Lastly, we subtract the squared probability of someone in this **Leaf** *not* **Loving Cool As Ice**...



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

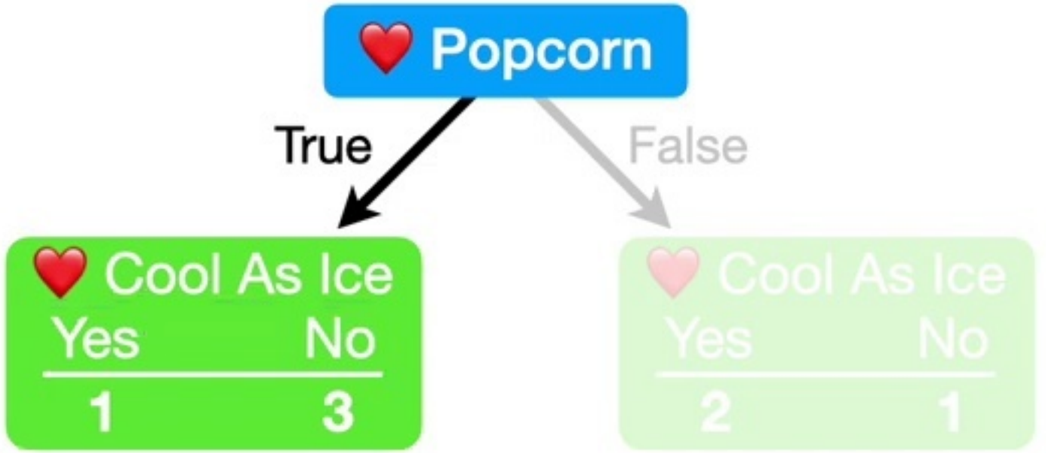
…which is **3**, the number of people in the **Leaf** who *did not* **Love Cool As Ice**…



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

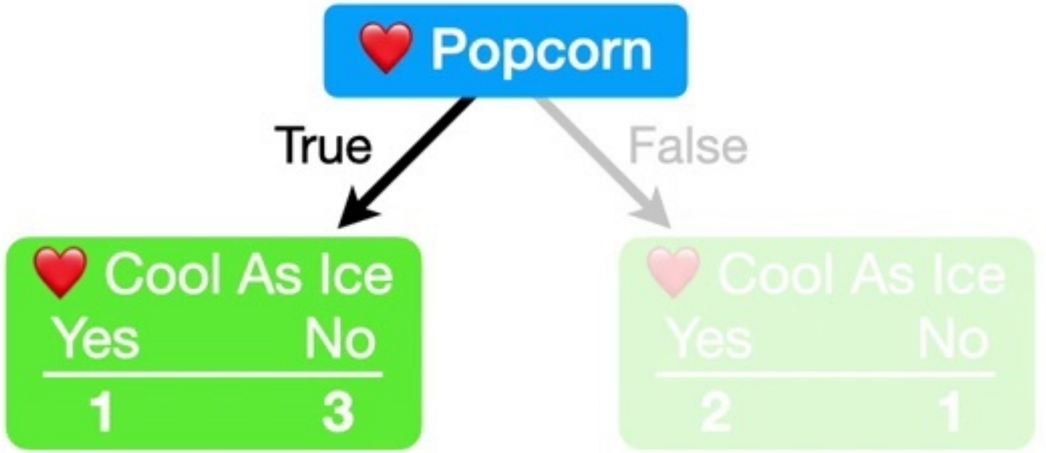
…divided by the total number of people in the **Leaf**…



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

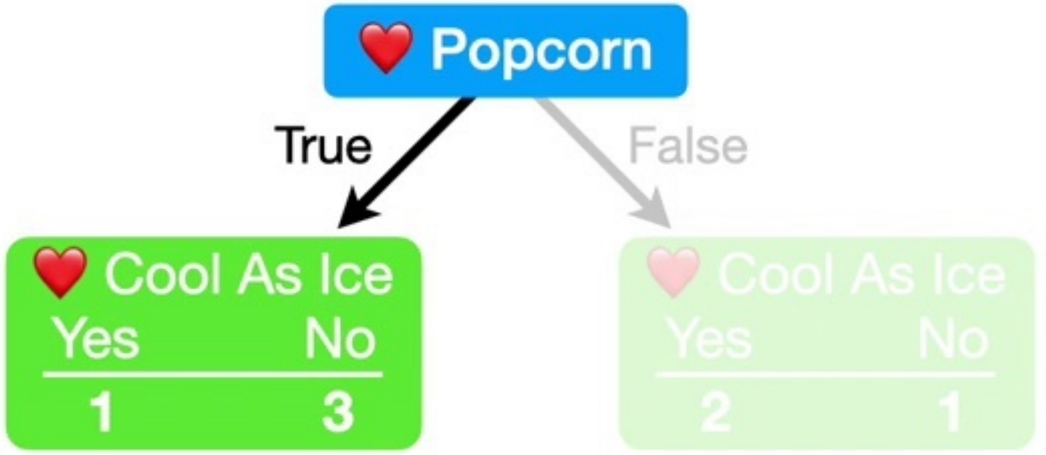
…squared.



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

And when we do the math, we get **0.375**.

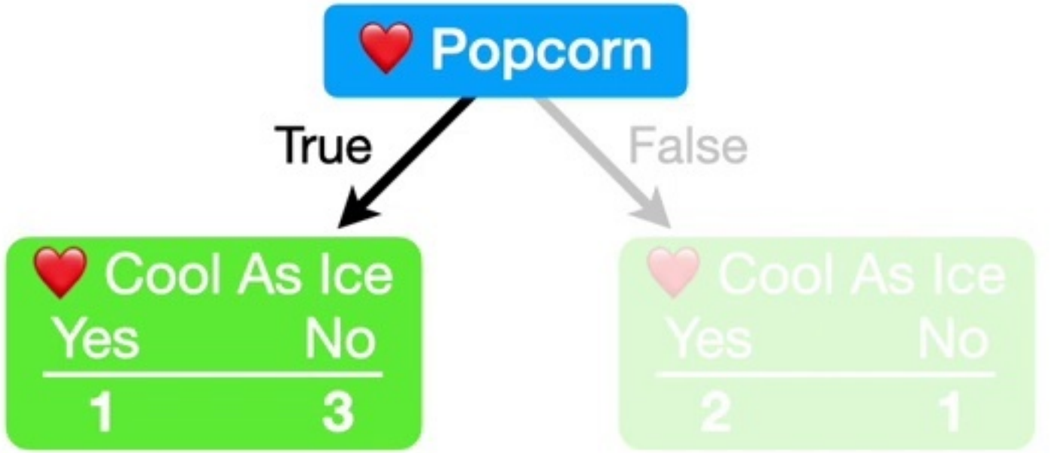


Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

$= 0.375$

And when we do the math, we get **0.375**.



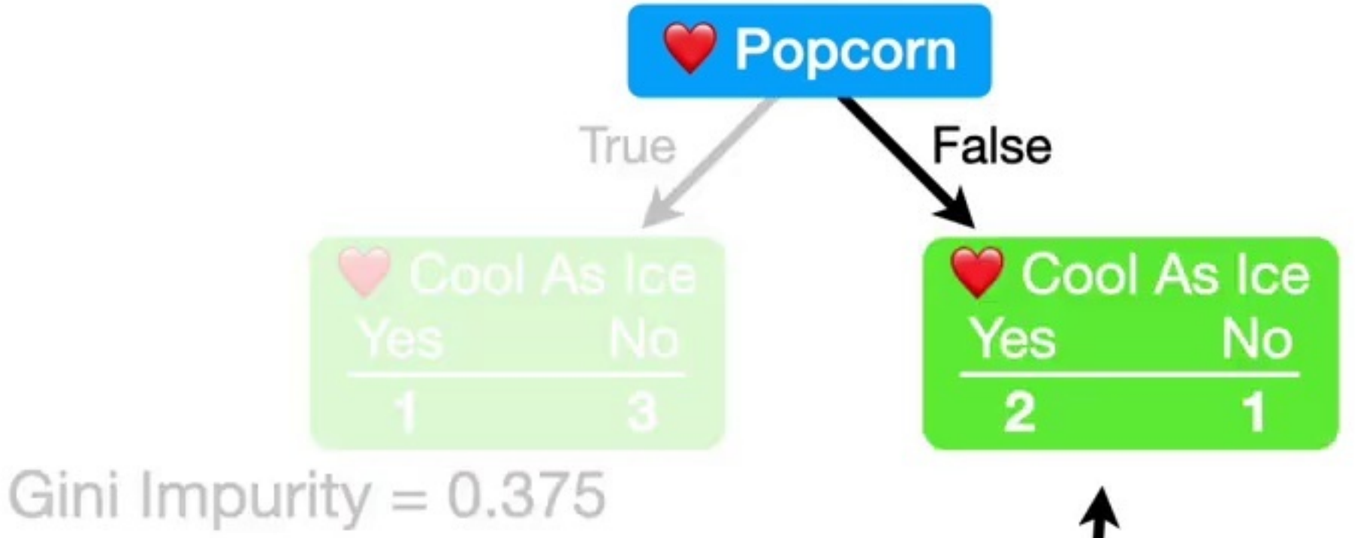
Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

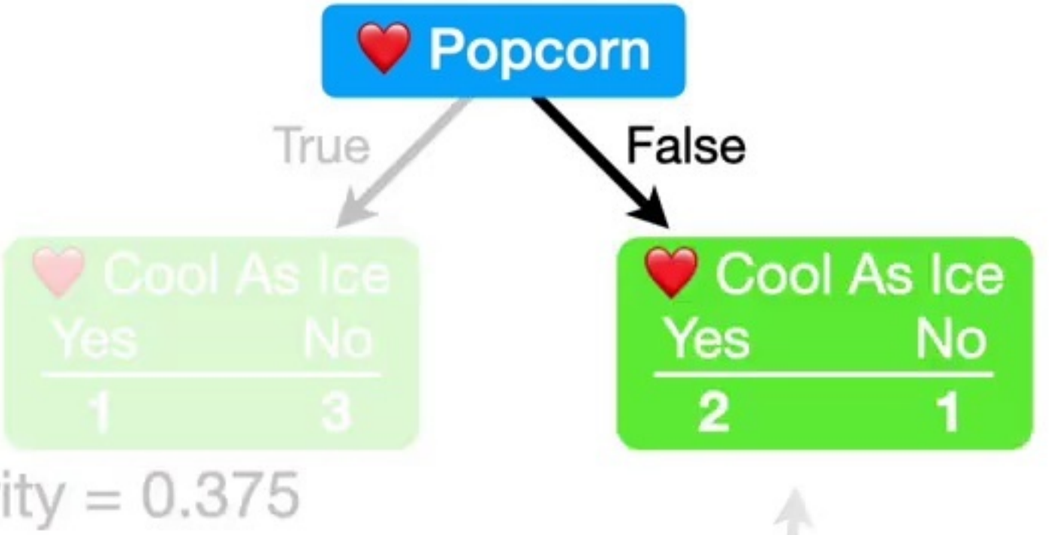
$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2$$

$$= 0.375$$

So let's put **0.375** under the **Leaf** on the left so we don't forget it.

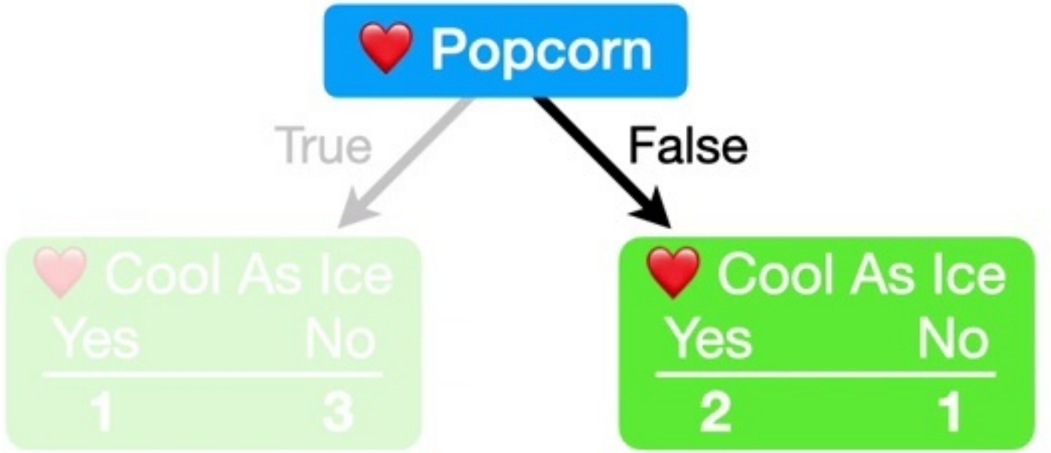


Now let's calculate the **Gini Impurity** for the **Leaf** on the right.



Gini Impurity for a Leaf = 1 - (the probability of “Yes”)^2 - (the probability of “No”)^2

Now let’s calculate the **Gini Impurity** for the **Leaf** on the right.

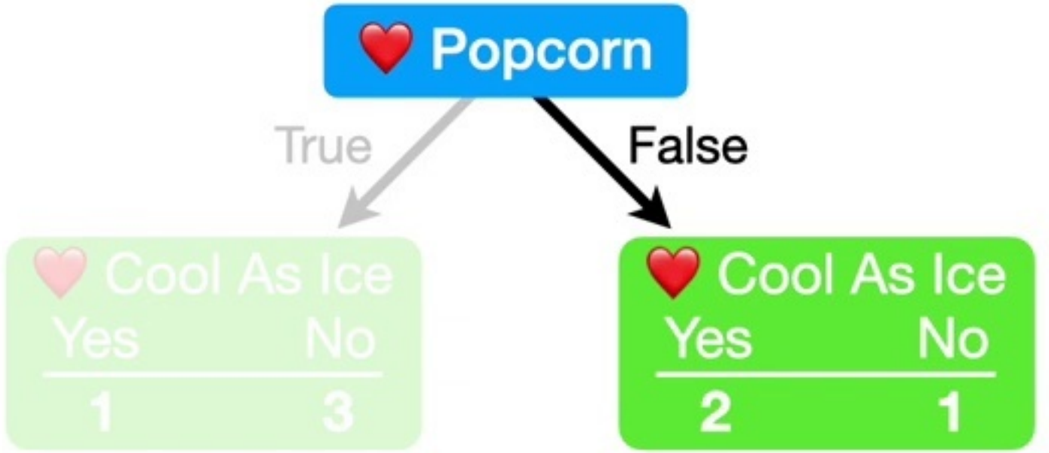


Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

= 1

Just like before, we start out with **1**…

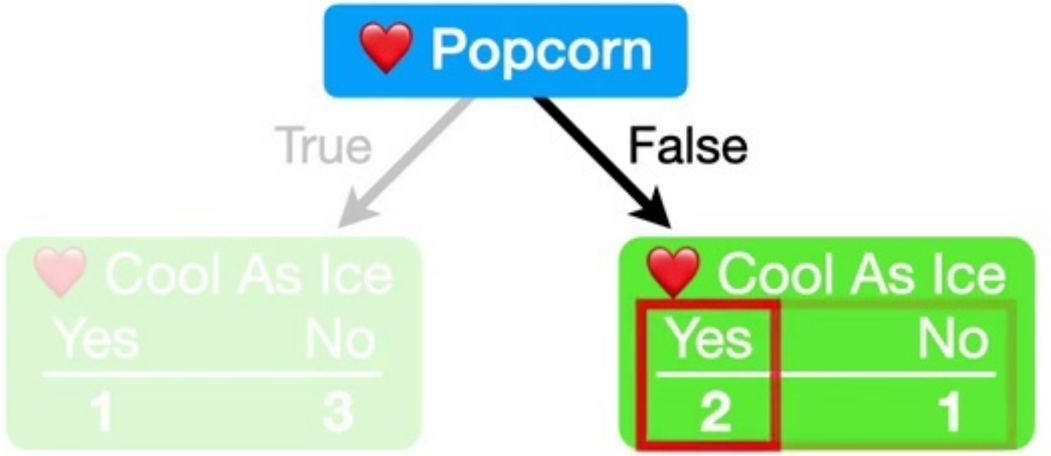


Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

= 1

…then we subtract the squared probability of someone in this **Leaf Loving Cool As Ice…**

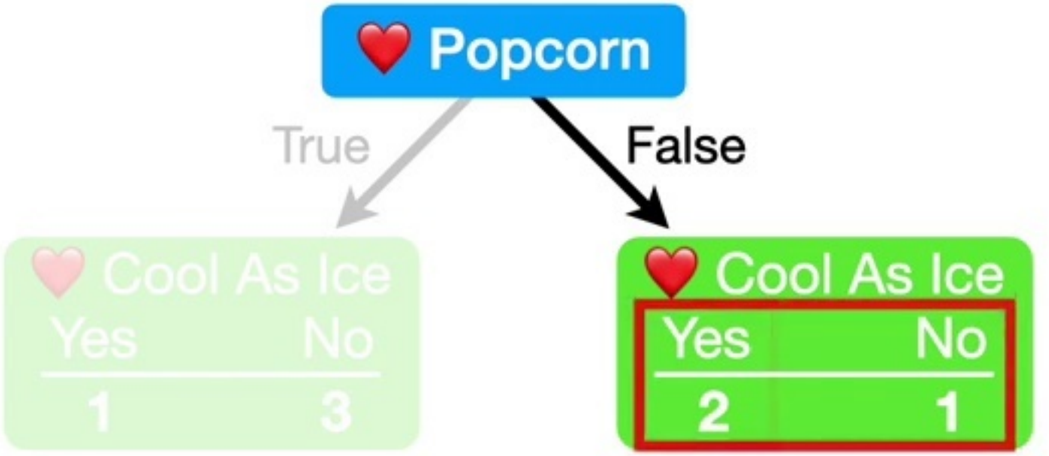


Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

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

…then we subtract the squared probability of someone in this **Leaf Loving Cool As Ice**…

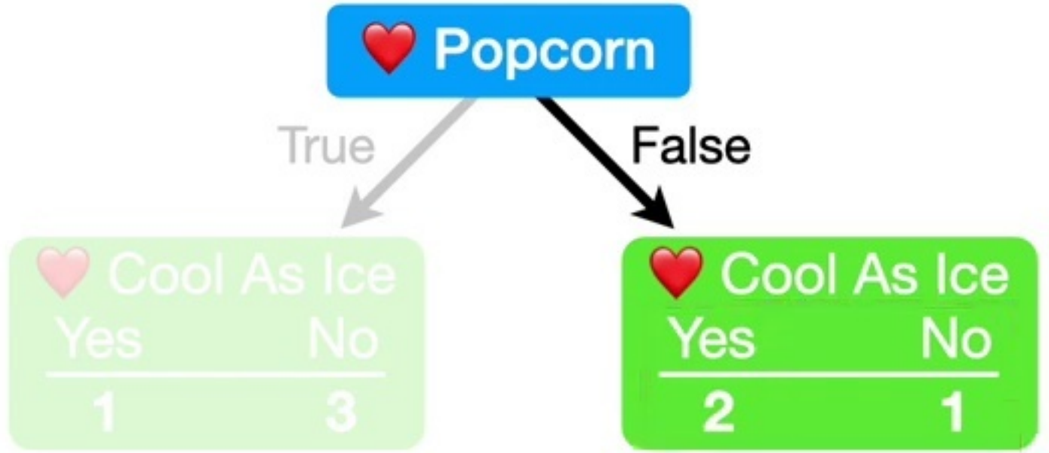


Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

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

…then we subtract the squared probability of someone in this **Leaf Loving Cool As Ice**…

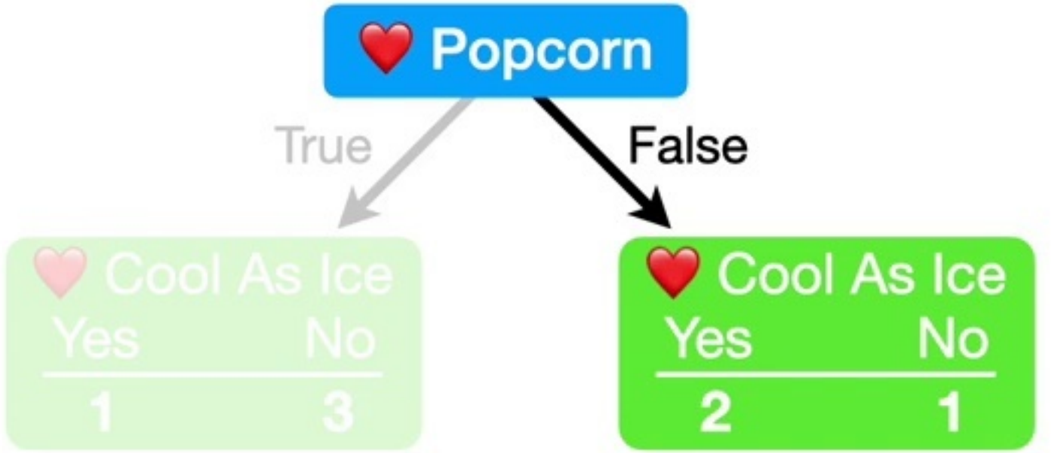


Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

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

…then we subtract the squared probability of someone in this **Leaf Loving Cool As Ice**…

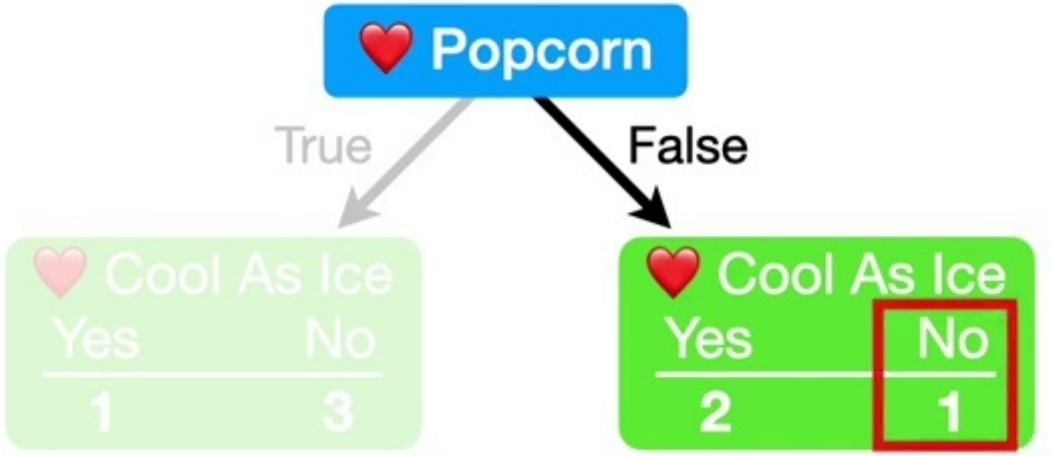


Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

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

…and the squared probability of someone in this **Leaf** *not* **Loving Cool As Ice.**

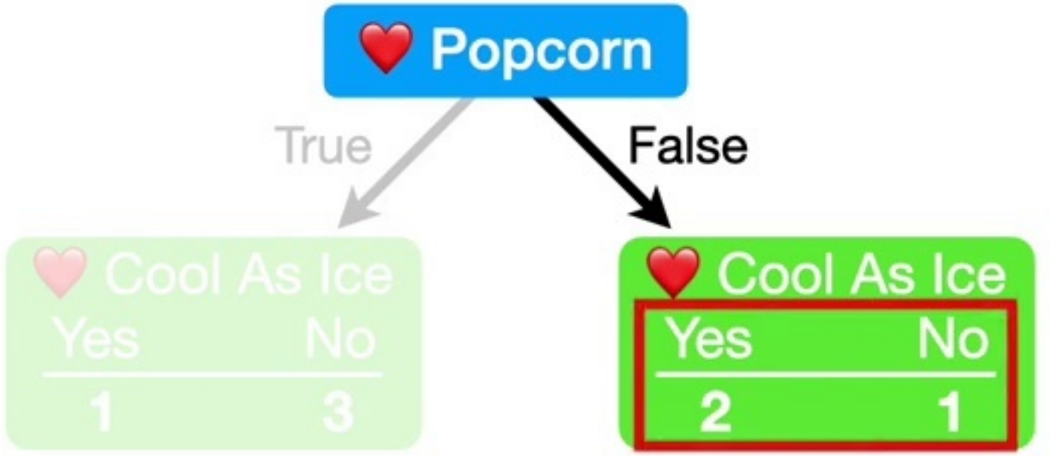


Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

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

…and the squared probability of someone in this **Leaf** *not* **Loving Cool As Ice**.

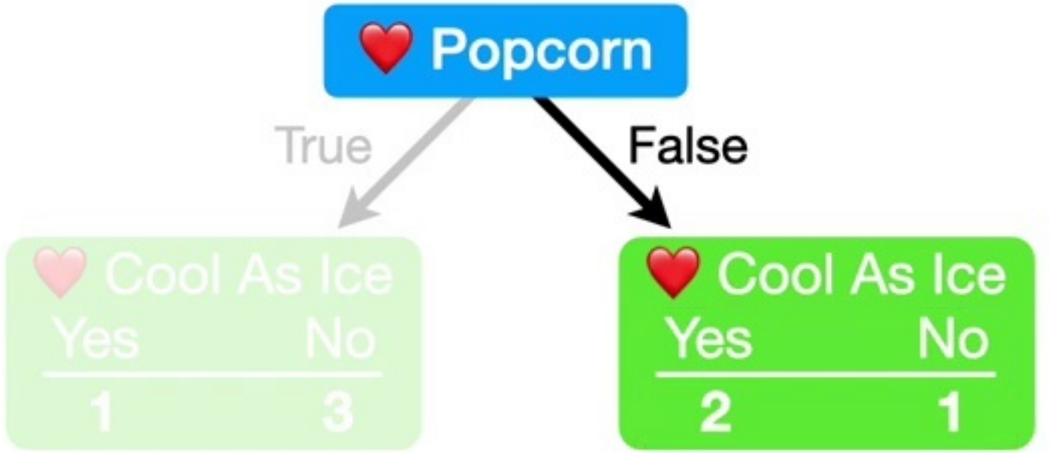


Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

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

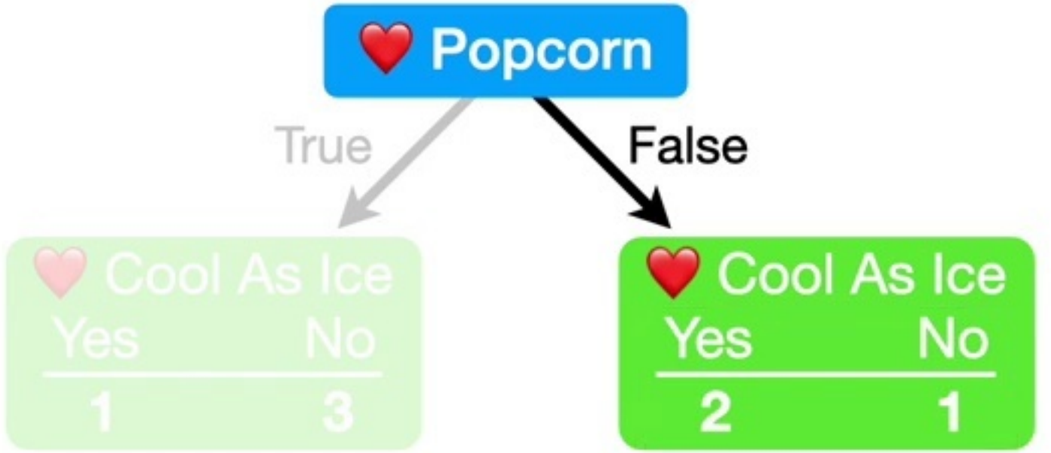
…and the squared probability of someone in this **Leaf** *not* **Loving Cool As Ice.**



Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

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

And when we do the math we get **0.444**.



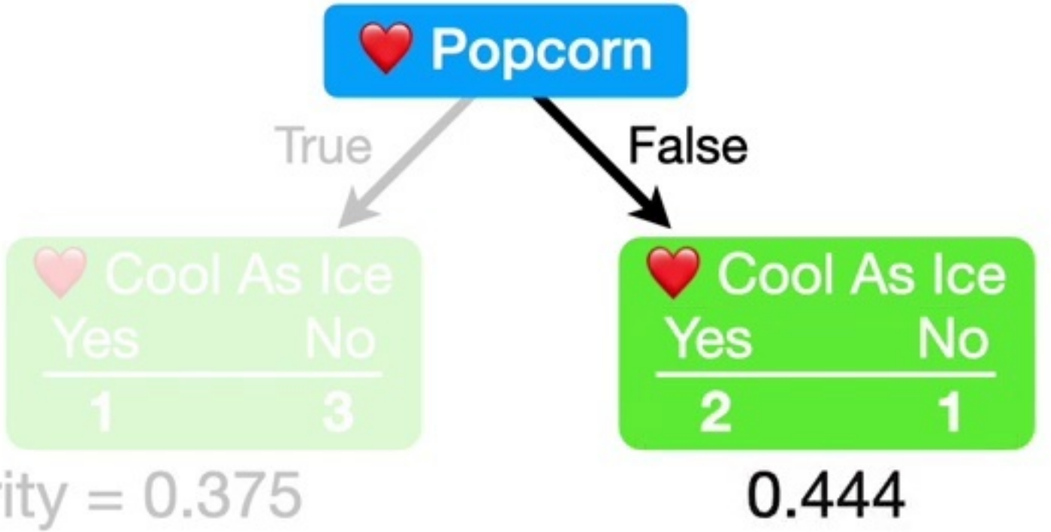
Gini Impurity = 0.375

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

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

= 0.444

And when we do the math we get **0.444**.



Gini Impurity = 0.375

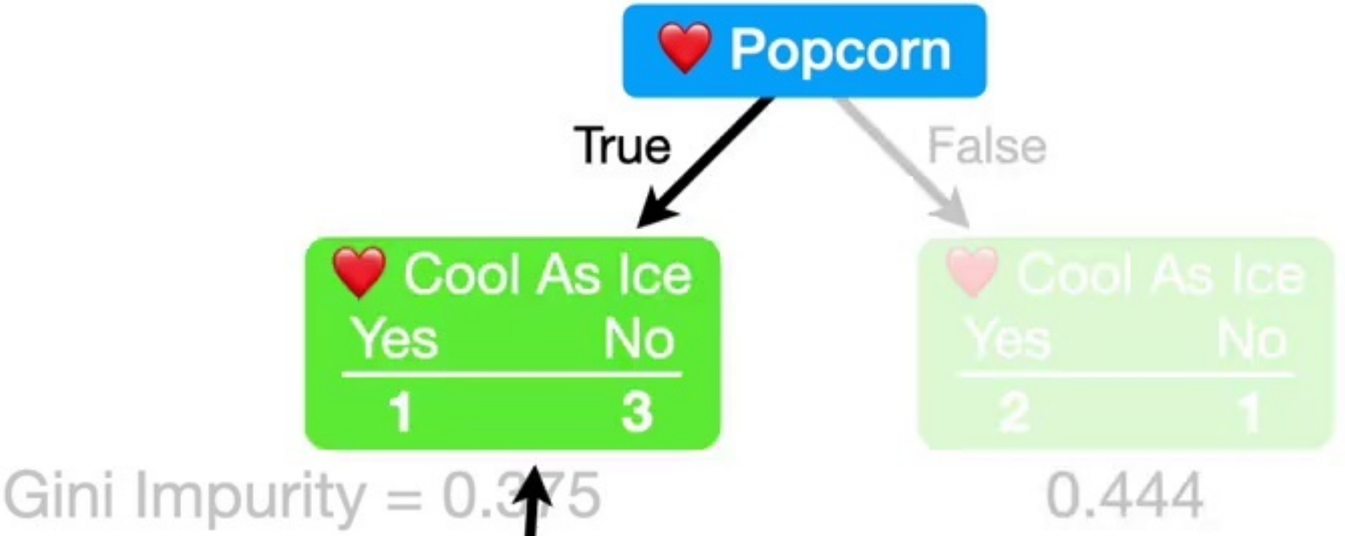
0.444

Gini Impurity for a Leaf = 1 - (the probability of "Yes")^2 - (the probability of "No")^2

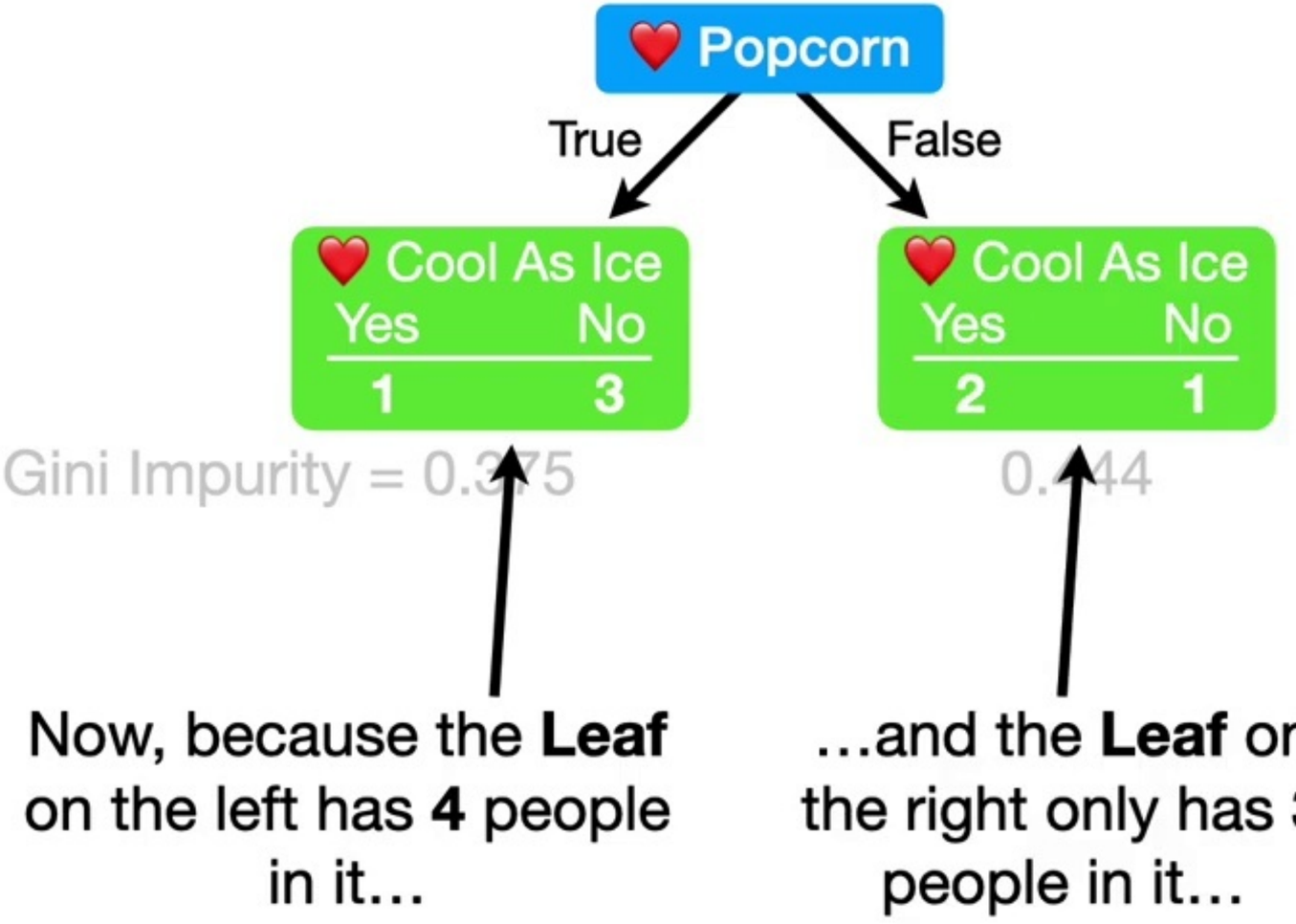
$$= 1 - \left(\frac{2}{2+1}\right)^2 - \left(\frac{1}{2+1}\right)^2$$

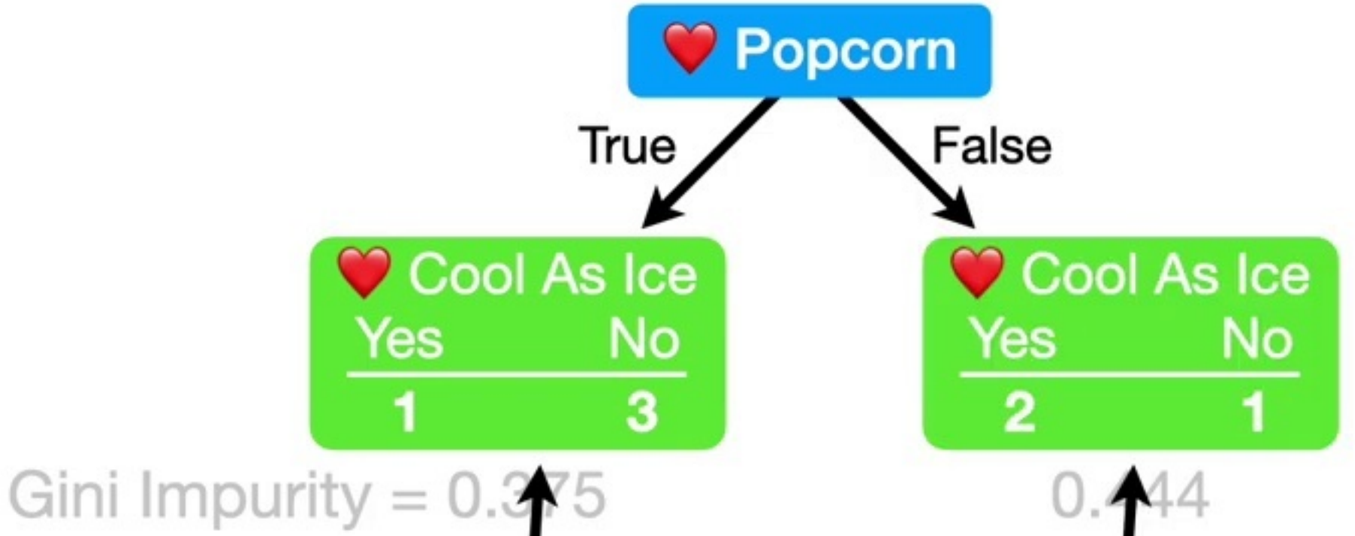
= 0.444

And when we do the math we get **0.444**.



Now, because the **Leaf** on the left has **4** people in it…

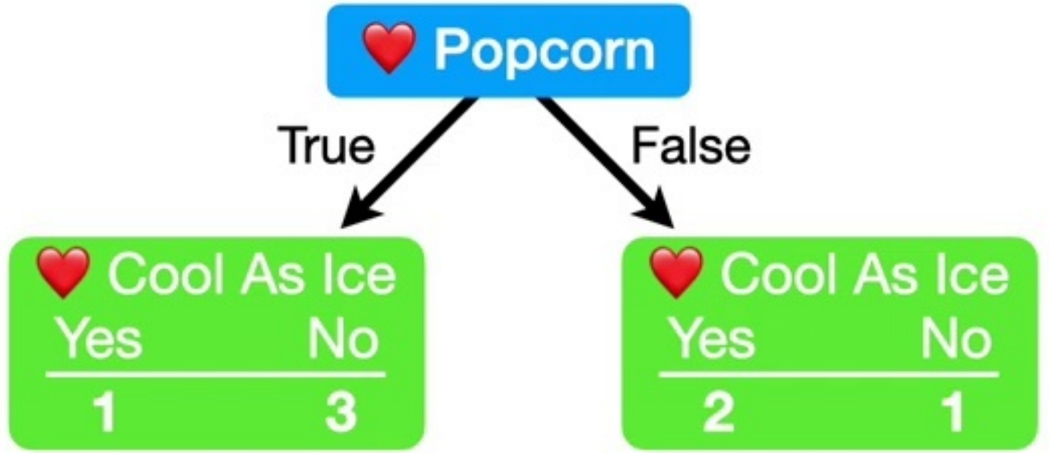

Popcorn
True
False
Cool As Ice
Yes
No
1
3
Cool As Ice
Yes
No
2
1
Gini Impurity = 0.375
0.444
Now, because the **Leaf** on the left has **4** people in it…
…and the **Leaf** on the right only has **3** people in it…



Now, because the **Leaf** on the left has **4** people in it…

…and the **Leaf** on the right only has **3** people in it…

…the **Leaves** do not represent the same number of people.

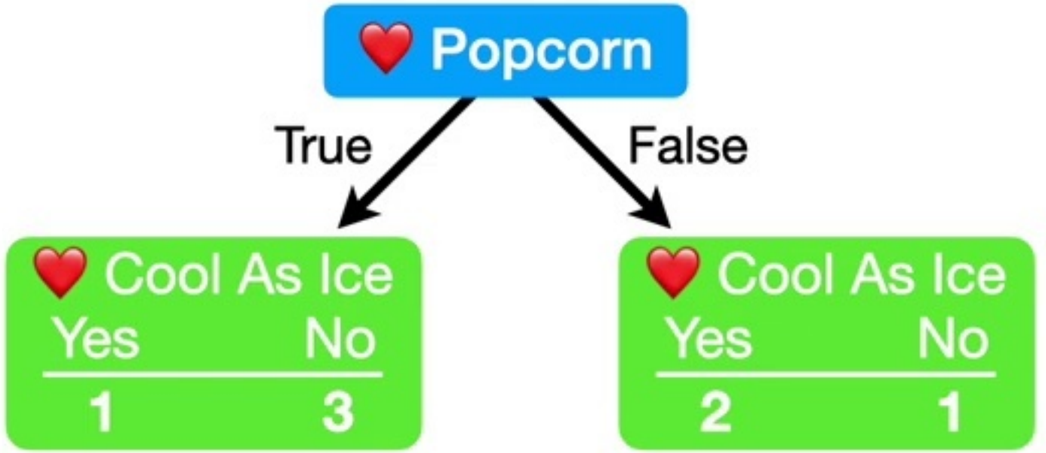


Gini Impurity = 0.375

0.444

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

Thus, the total **Gini Impurity** is the **Weighted Average** of the **Leaf Impurities**.



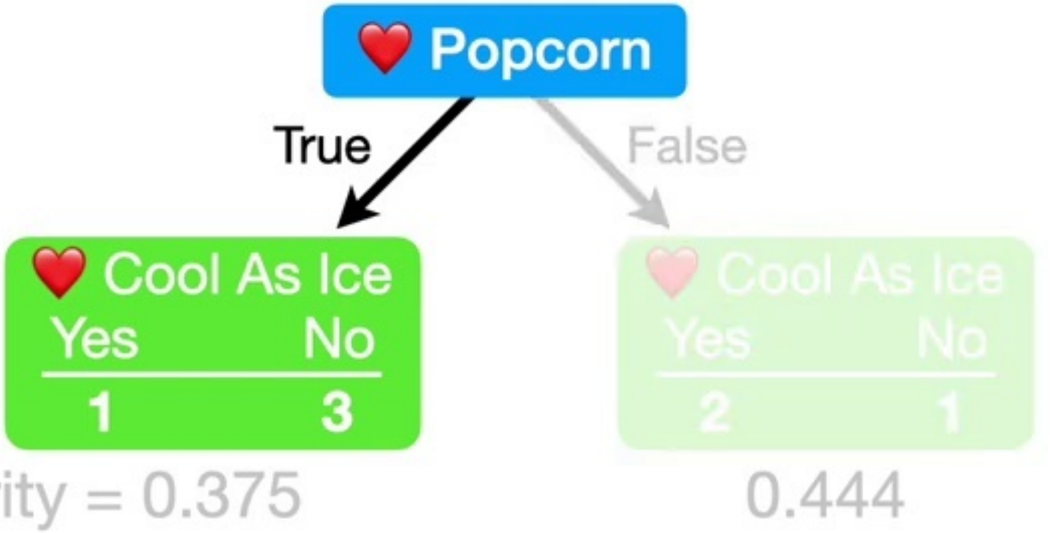
Gini Impurity = 0.375

0.444

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

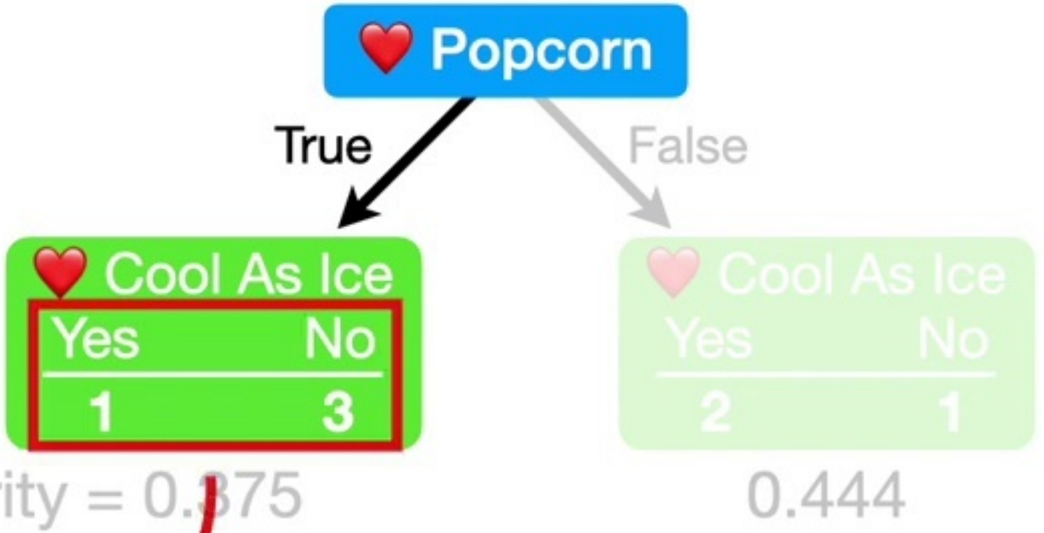
Thus, the total **Gini Impurity** is the **Weighted Average** of the **Leaf Impurities**.

We start by calculating the weight for the **Leaf** on the left.



Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

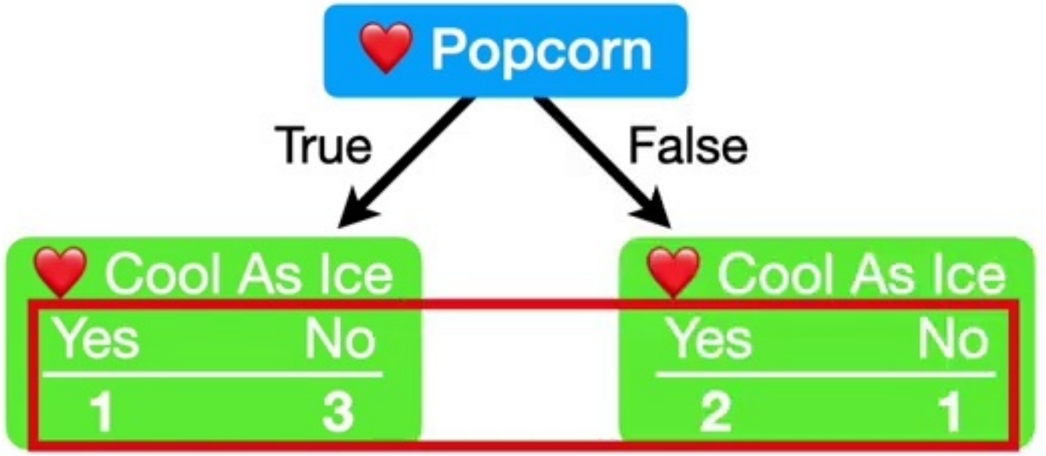
The weight for the left **Leaf** is the total number of people in the **Leaf, 4**…



Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

$$= \left(\frac{4}{4+3}\right)$$

…divided by the total number of people in both **Leaves**, **7**.



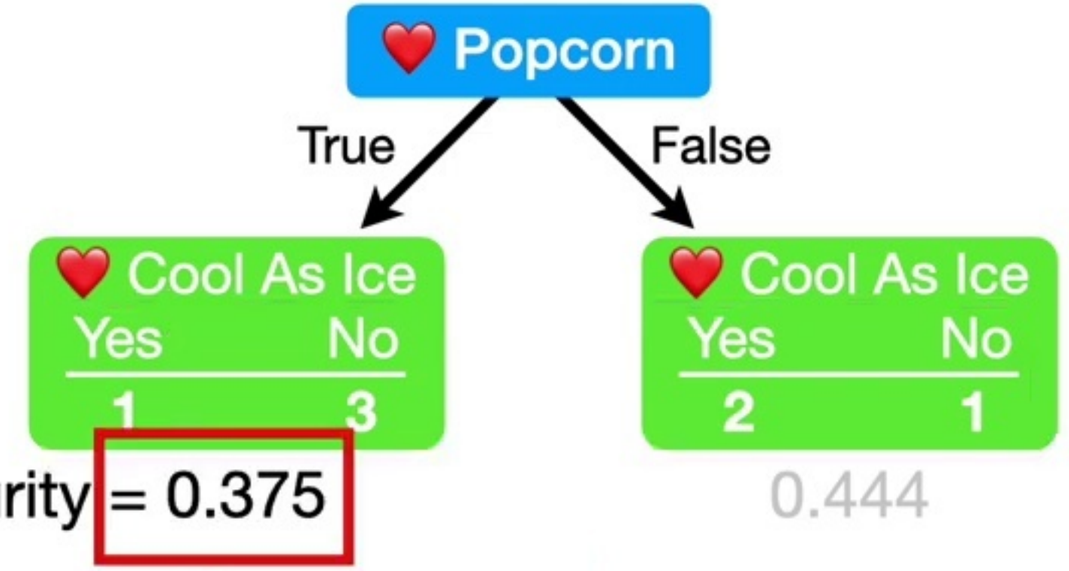
Gini Impurity = 0.375

0.444

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

$$= \left(\frac{4}{4+3} \right)$$

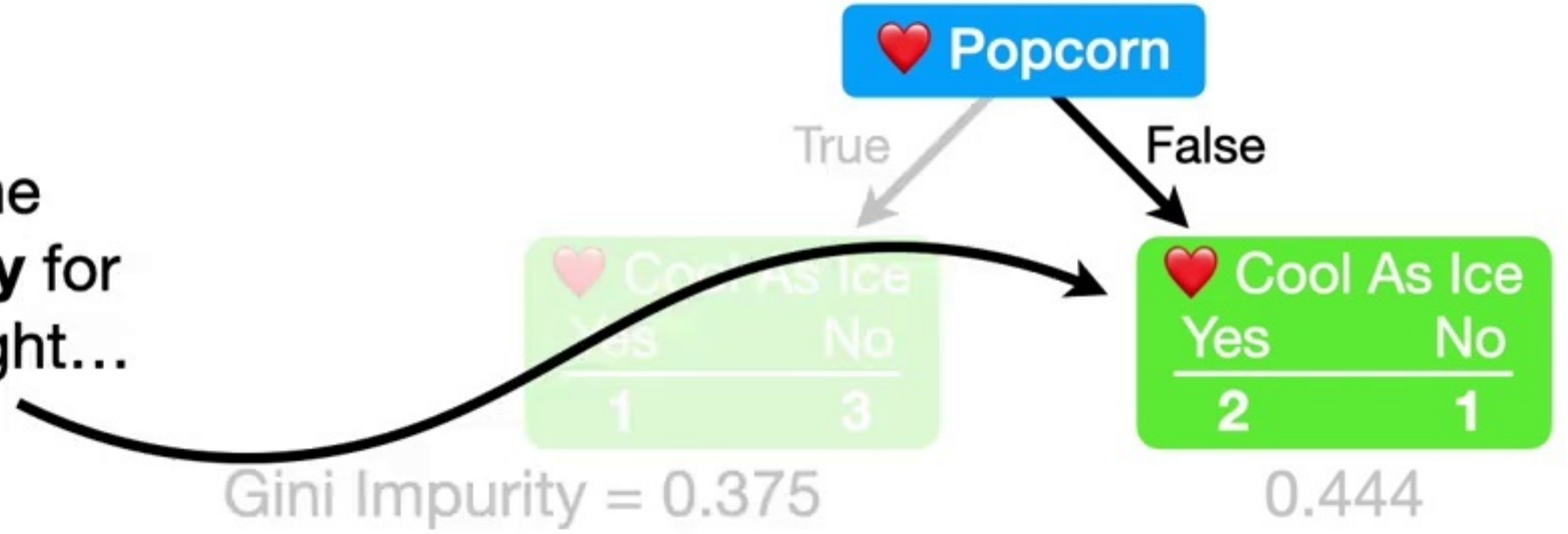
Then we multiply that weight by its associated **Gini Impurity, 0.375**.



Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

$$= \left(\frac{4}{4+3}\right) 0.375$$

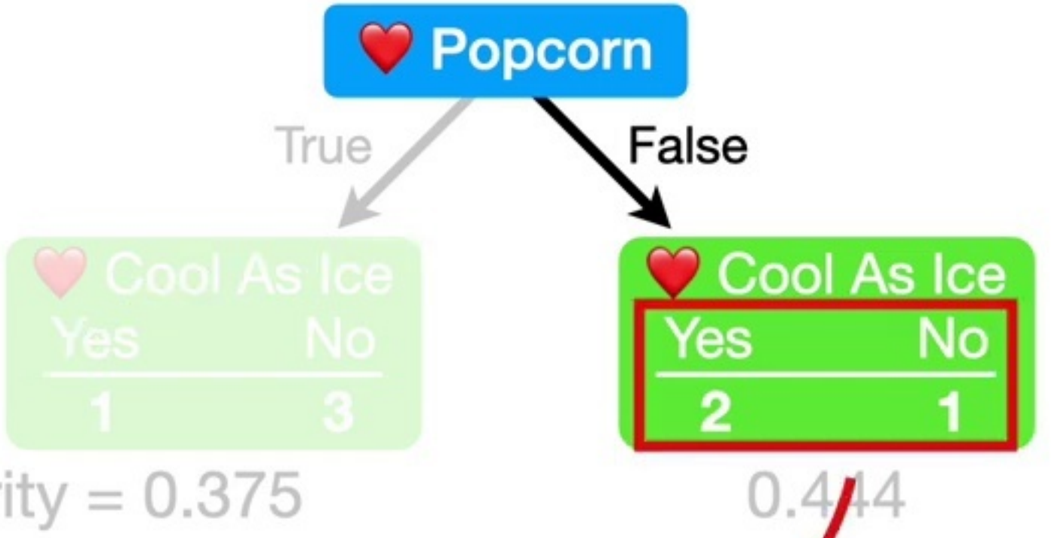
Now we add the weighted **Impurity** for the **Leaf** on the right…



Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

$$= \left(\frac{4}{4+3}\right) 0.375$$

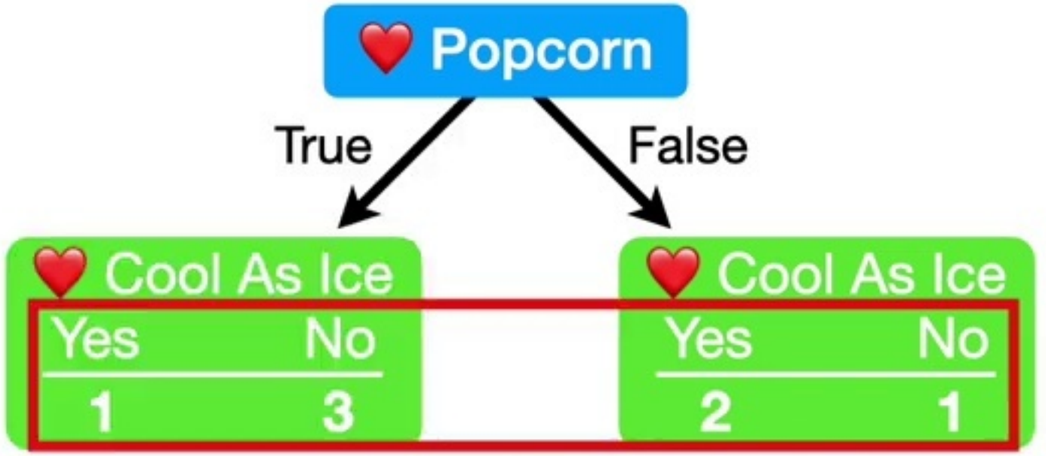
…which is the total number of people in the **Leaf, 3**…



Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

$$= \left(\frac{4}{4+3}\right) 0.375 + \left(\frac{3}{4+3}\right)$$

…divided by the total number of people in both **Leaves, 7**…



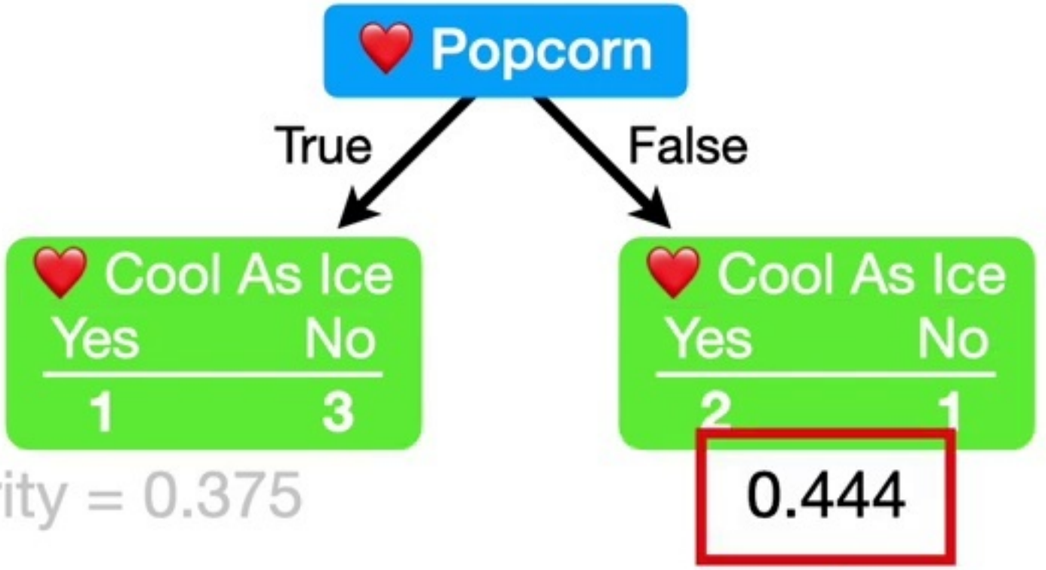
Gini Impurity = 0.375

0.444

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

$$= \left(\frac{4}{4+3}\right) 0.375 + \left(\frac{3}{4+3}\right)$$

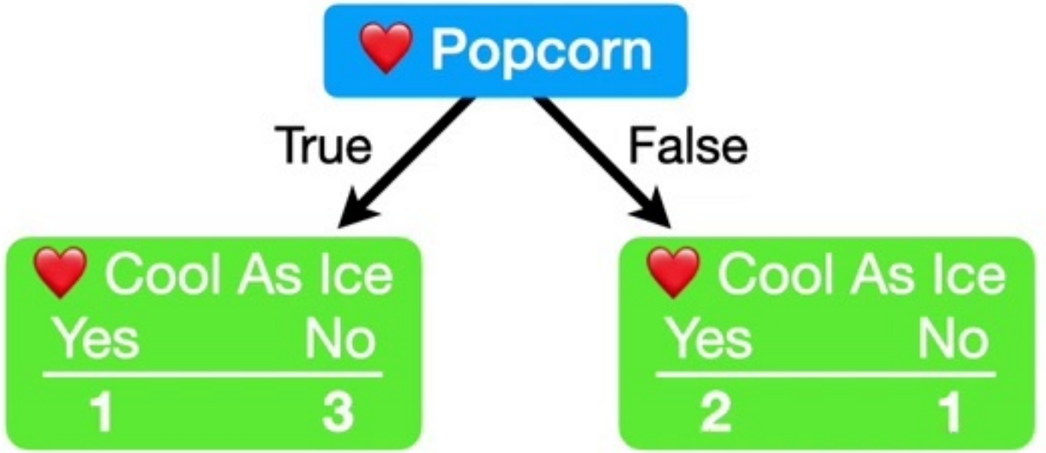
…times the associated **Gini Impurity, 0.444**.



Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

$$= \left(\frac{4}{4+3}\right) 0.375 + \left(\frac{3}{4+3}\right) 0.444$$

And when we do the math, we get **0.405**.

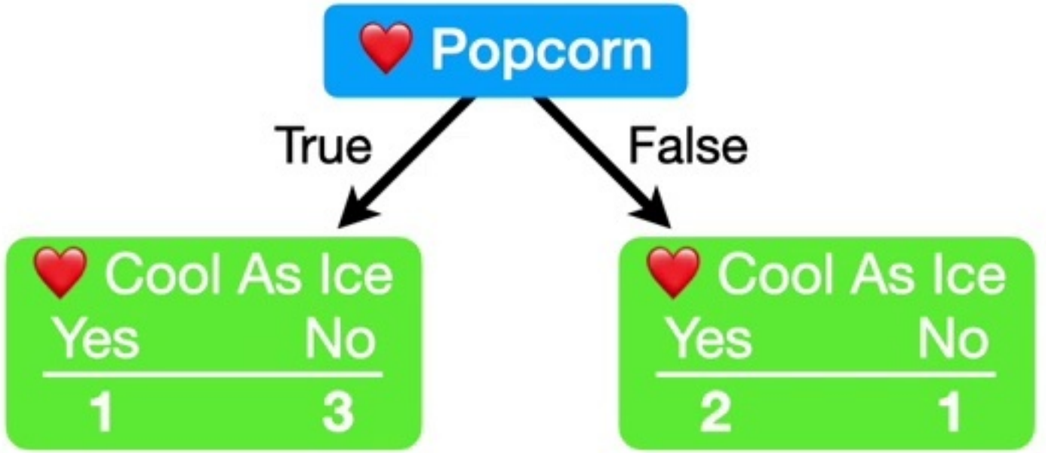


Gini Impurity = 0.375

0.444

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

$$= \left(\frac{4}{4+3}\right) 0.375 + \left(\frac{3}{4+3}\right) 0.444$$



Gini Impurity = 0.375

0.444

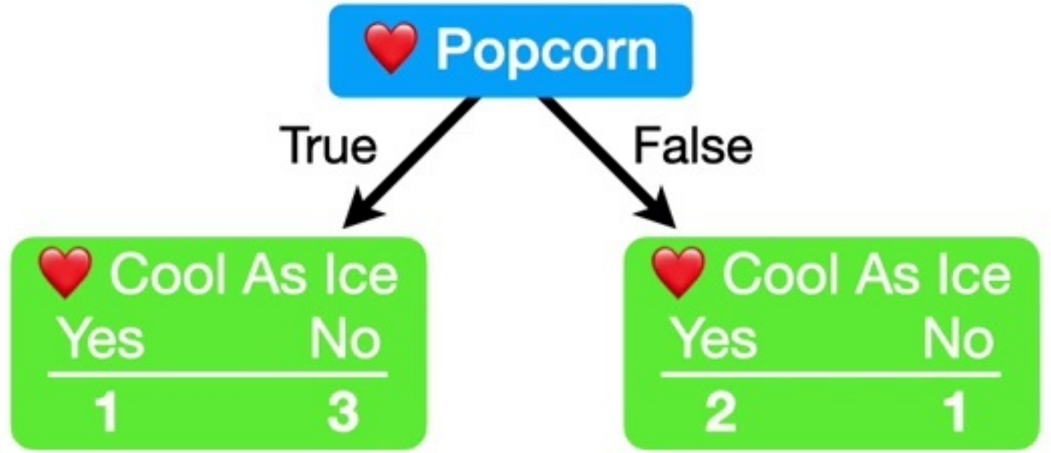
And when we do the math, we get **0.405**.

Total **Gini Impurity** = weighted average of **Gini Impurities** for the **Leaves**

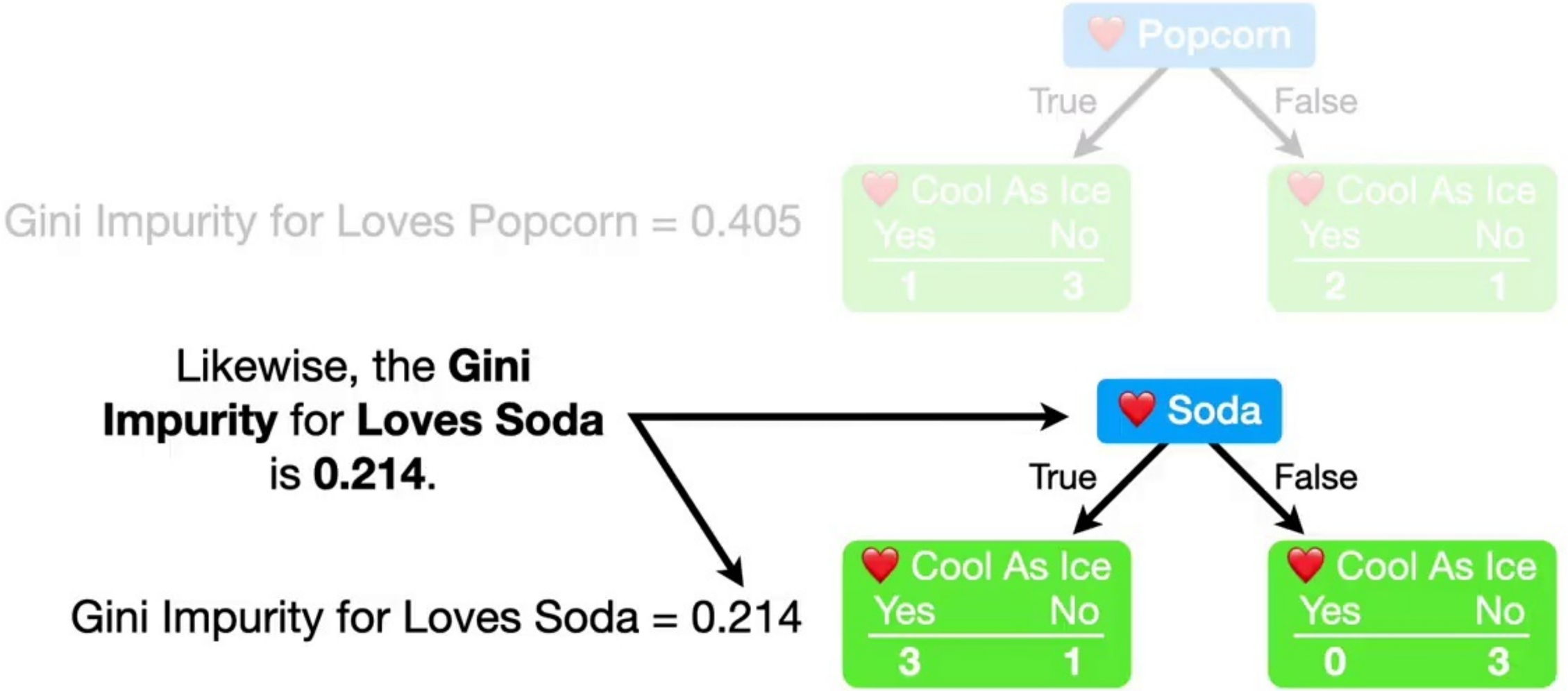
$$= \left(\frac{4}{4+3}\right) 0.375 + \left(\frac{3}{4+3}\right) 0.444$$

$$= 0.405$$

Gini Impurity for Loves Popcorn = 0.405

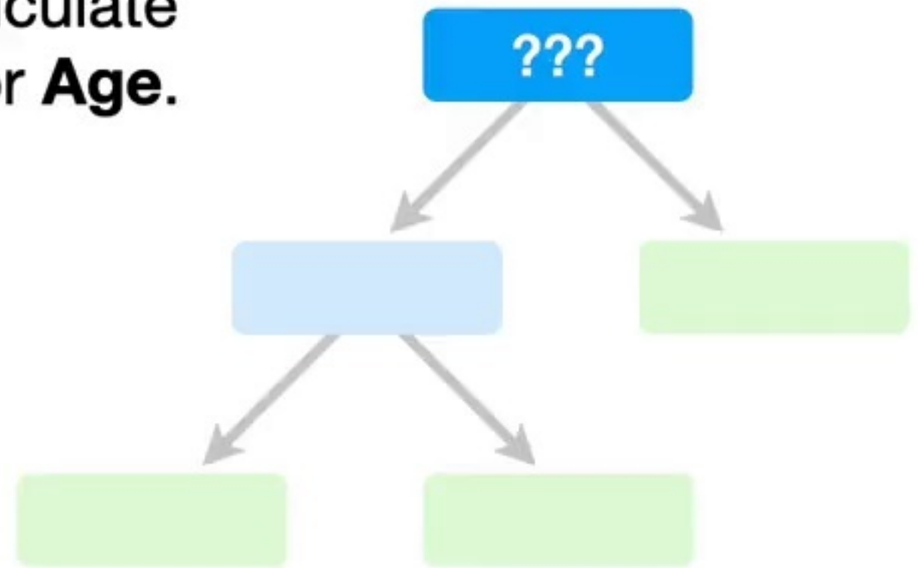


So the **Gini Impurity** for **Loves Popcorn** is **0.405**.


Popcorn
True
False
Cool As Ice
Yes No
1 3
Cool As Ice
Yes No
2 1
Gini Impurity for Loves Popcorn = 0.405
Likewise, the Gini Impurity for Loves Soda is 0.214.
Soda
True
False
Cool As Ice
Yes No
3 1
Cool As Ice
Yes No
0 3
Gini Impurity for Loves Soda = 0.214

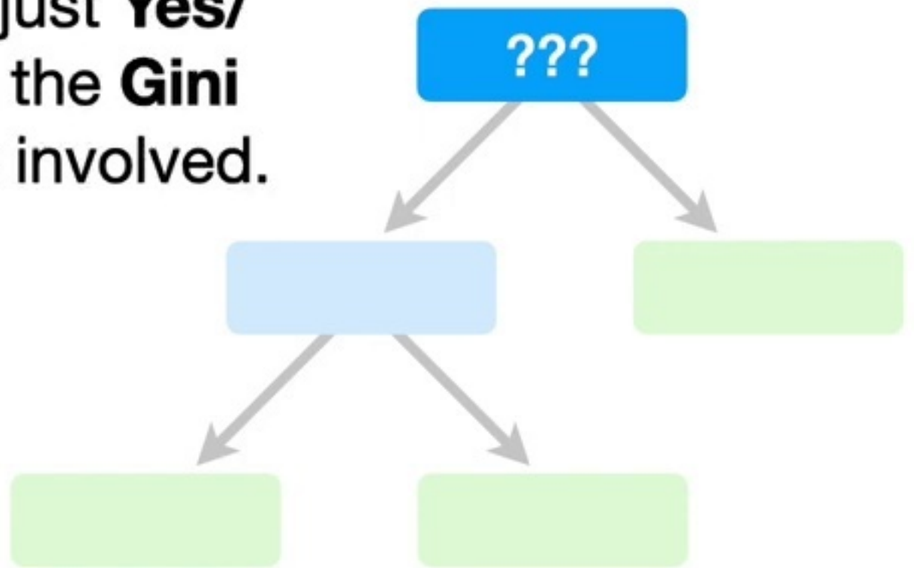
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Now we need to calculate the **Gini Impurity** for **Age**.



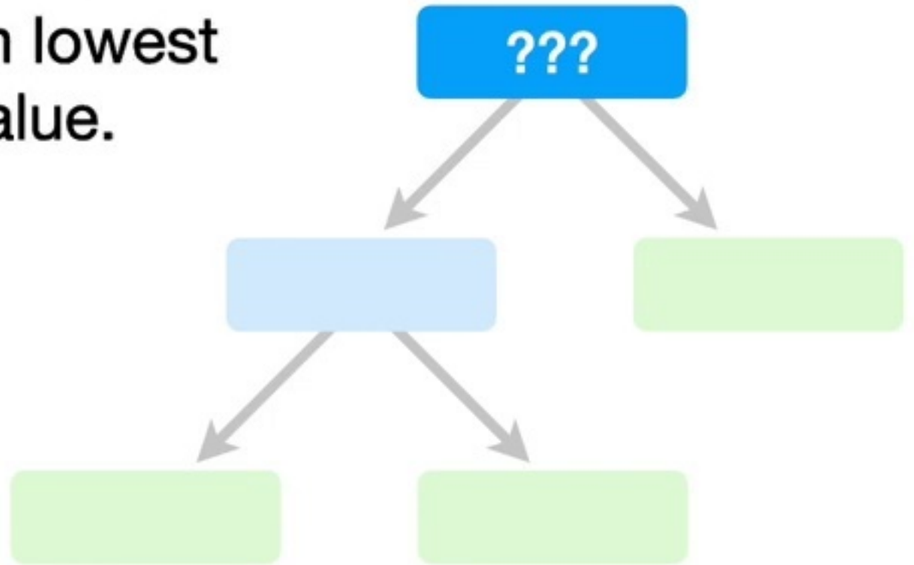
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

However, because **Age** contains numeric data, and not just **Yes/No** values, calculating the **Gini Impurity** is a little more involved.



Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

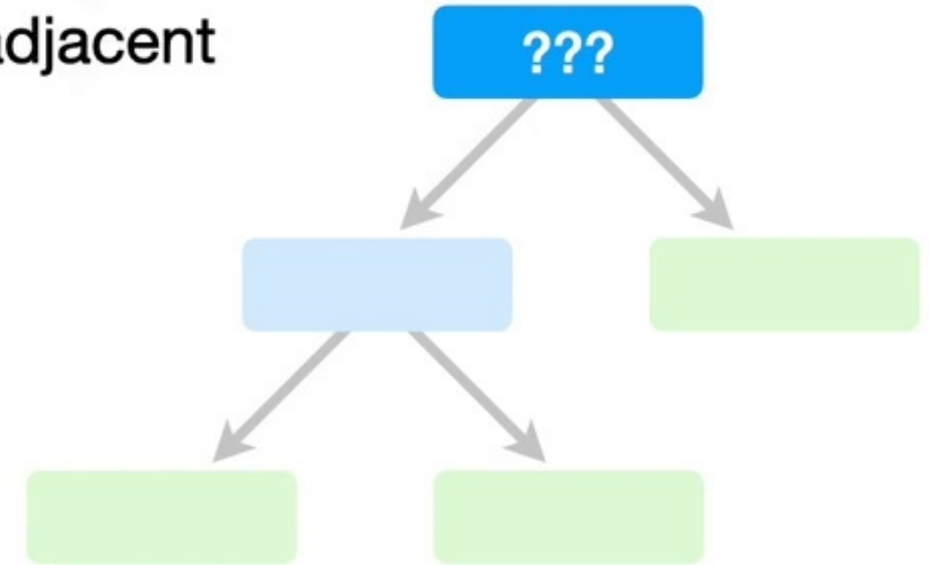
The first thing we do is sort the rows by **Age**, from lowest value to highest value.



Then we calculate the average **Age** for all adjacent people.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5

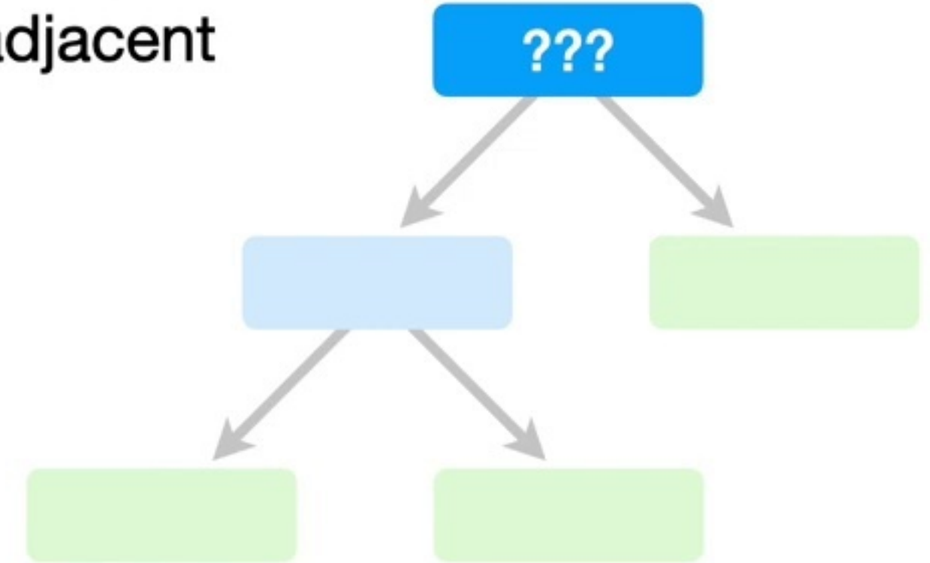


Then we calculate the average **Age** for all adjacent people.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

0.5

15



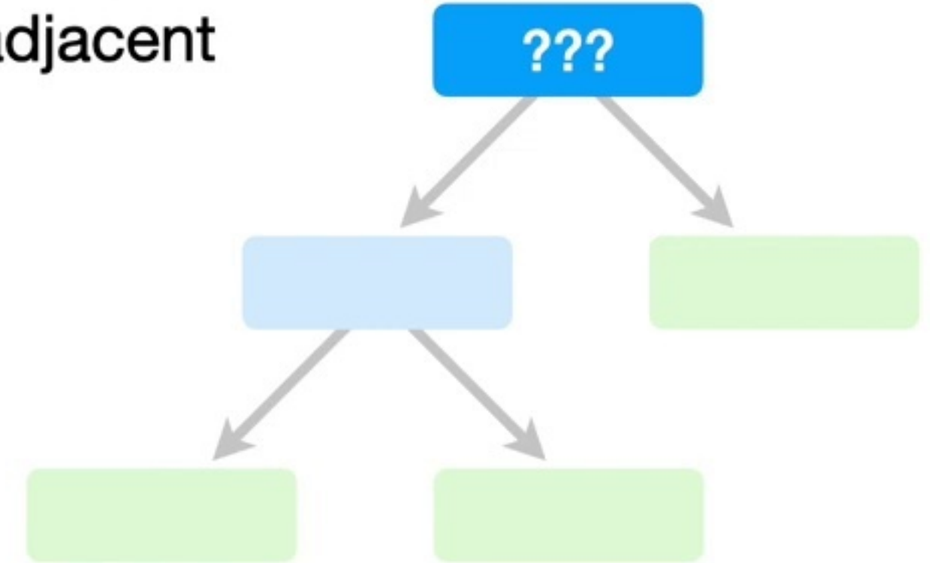
Then we calculate the average **Age** for all adjacent people.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5

15

26.5



Then we calculate the average **Age** for all adjacent people.

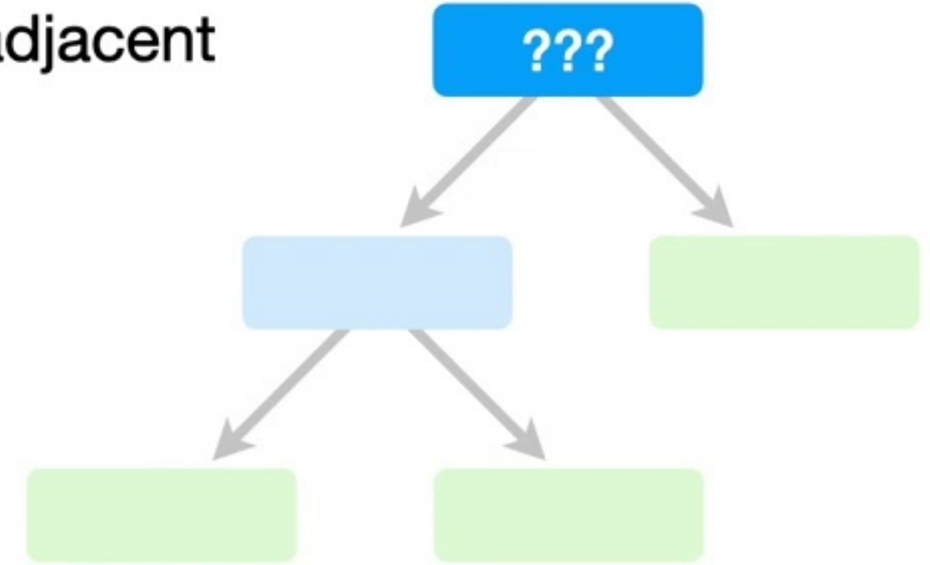
Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5

15

26.5

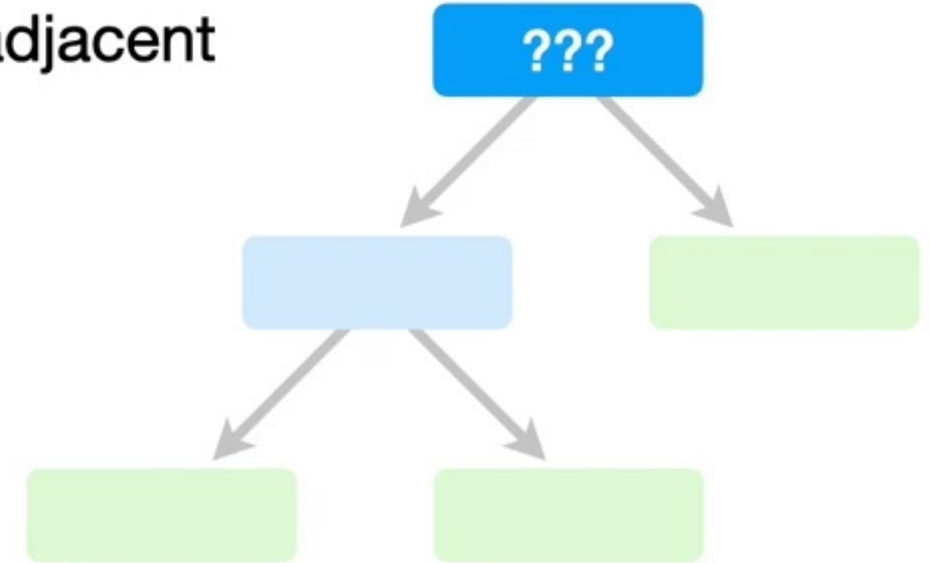
36.5



Then we calculate the average **Age** for all adjacent people.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

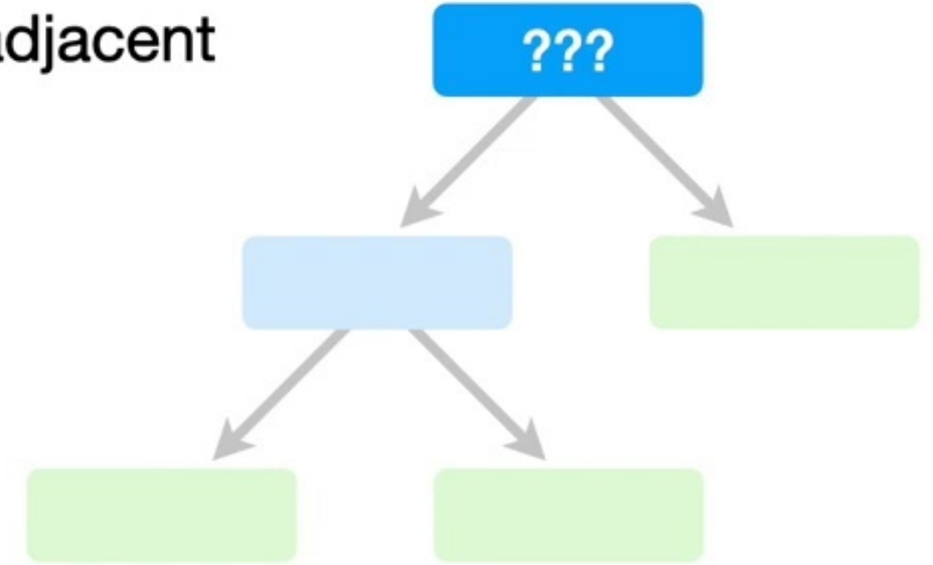
9.5
15
26.5
44



Then we calculate the average **Age** for all adjacent people.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5

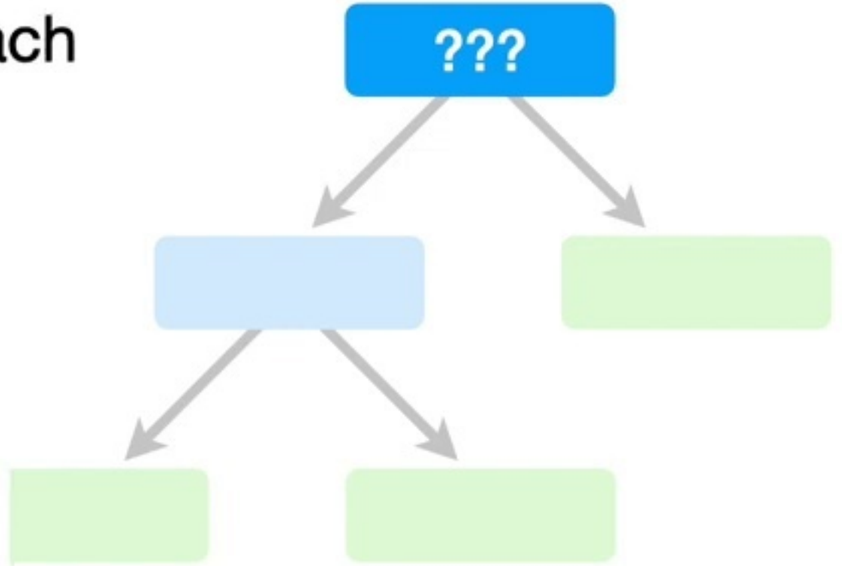


Lastly, we calculate the **Gini Impurity** values for each average age.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5

Gini Impurity = 0.429



Lastly, we calculate the **Gini Impurity** values for each average age.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5 → Gini Impurity = 0.429

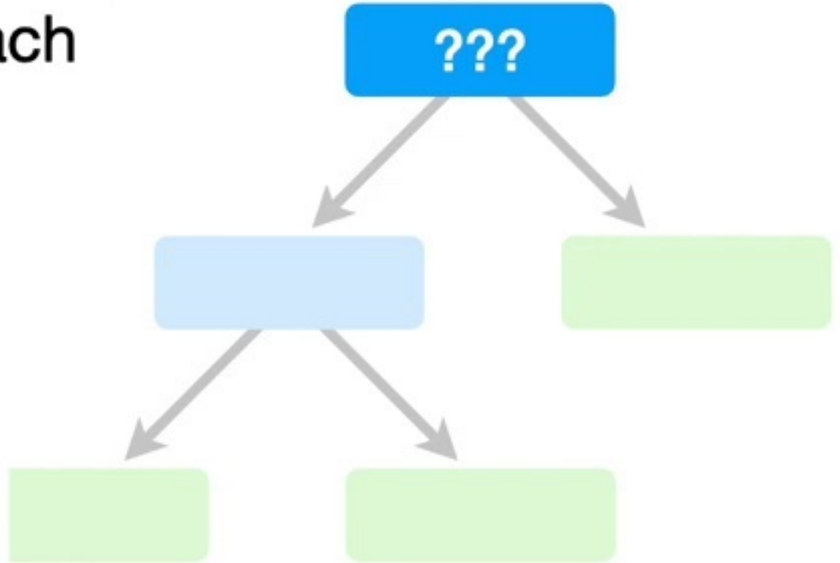
15 → Gini Impurity = 0.343

26.5

36.5

44

66.5



Lastly, we calculate the **Gini Impurity** values for each average age.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5 → Gini Impurity = 0.429

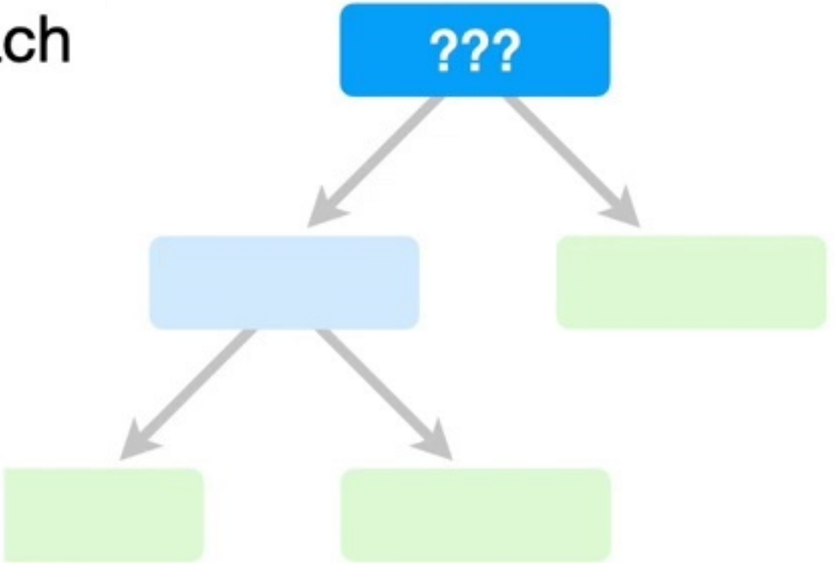
15 → Gini Impurity = 0.343

26.5 → Gini Impurity = 0.476

36.5

44

66.5



Lastly, we calculate the **Gini Impurity** values for each average age.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5 → Gini Impurity = 0.429

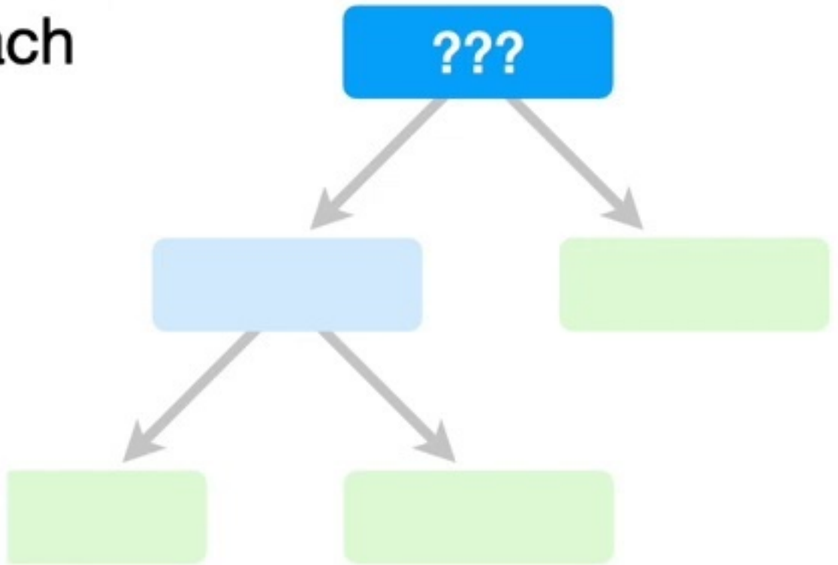
15 → Gini Impurity = 0.343

26.5 → Gini Impurity = 0.476

36.5 → Gini Impurity = 0.476

44

66.5



Lastly, we calculate the **Gini Impurity** values for each average age.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5 → Gini Impurity = 0.429

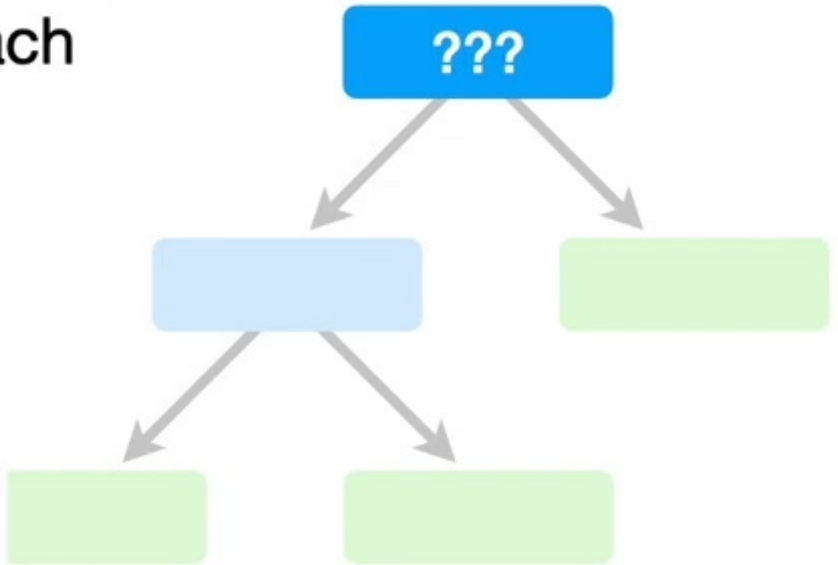
15 → Gini Impurity = 0.343

26.5 → Gini Impurity = 0.476

36.5 → Gini Impurity = 0.476

44 → Gini Impurity = 0.343

66.5



Lastly, we calculate the **Gini Impurity** values for each average age.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5 → Gini Impurity = 0.429

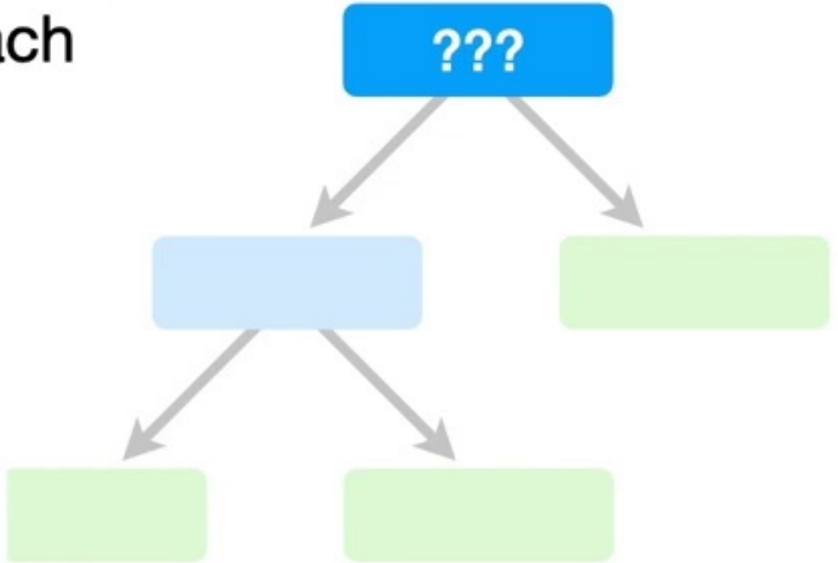
15 → Gini Impurity = 0.343

26.5 → Gini Impurity = 0.476

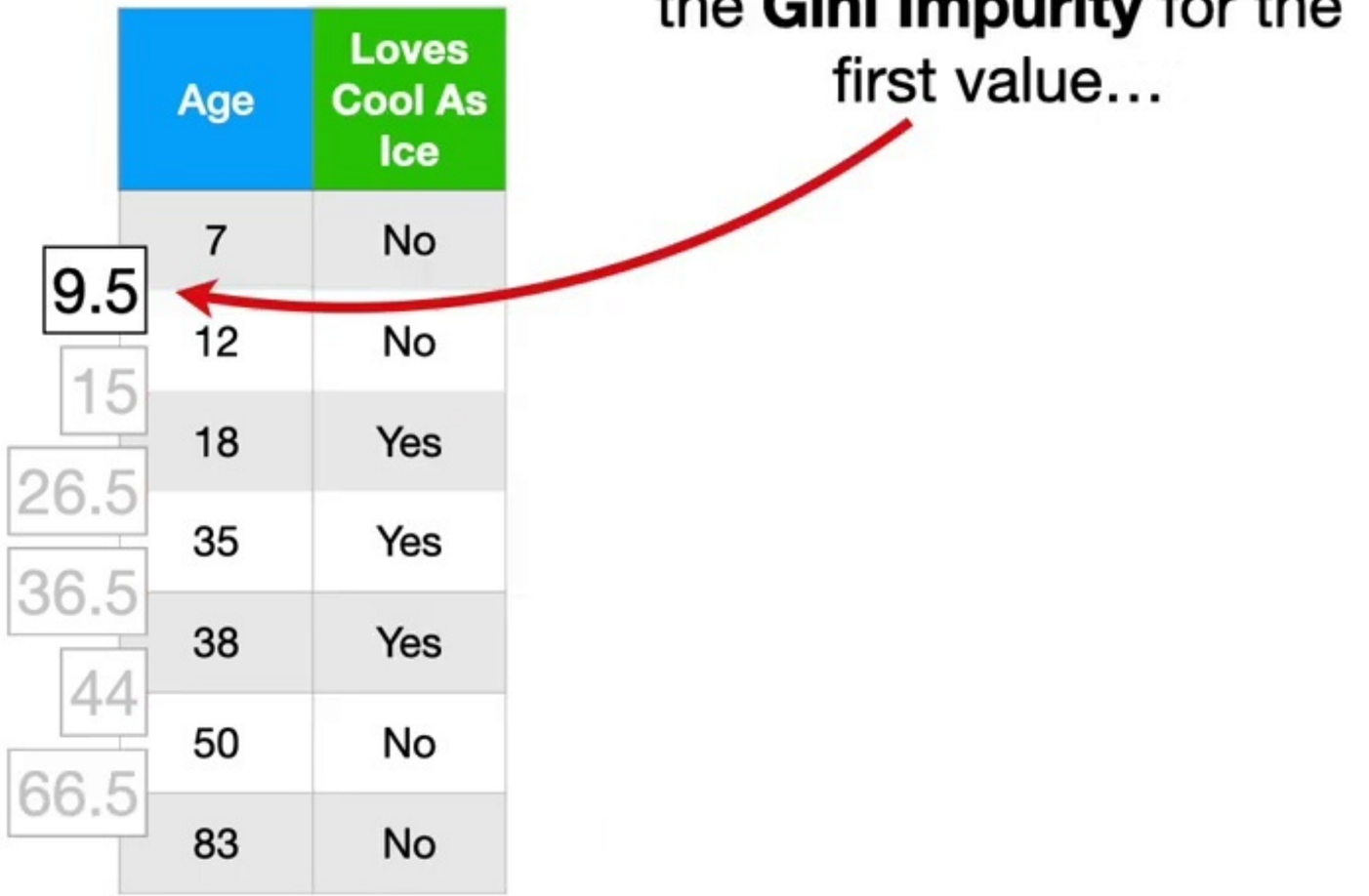
36.5 → Gini Impurity = 0.476

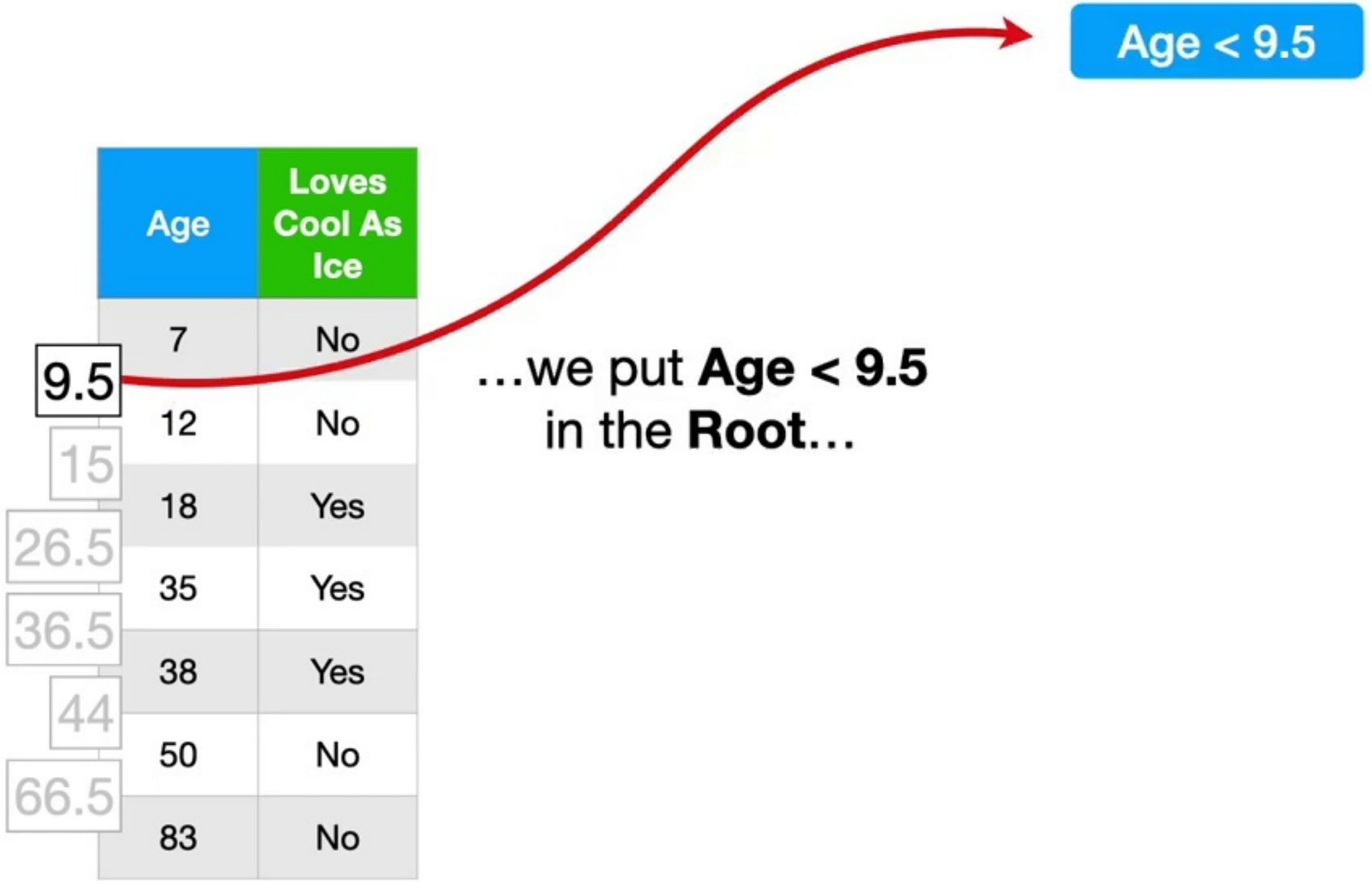
44 → Gini Impurity = 0.343

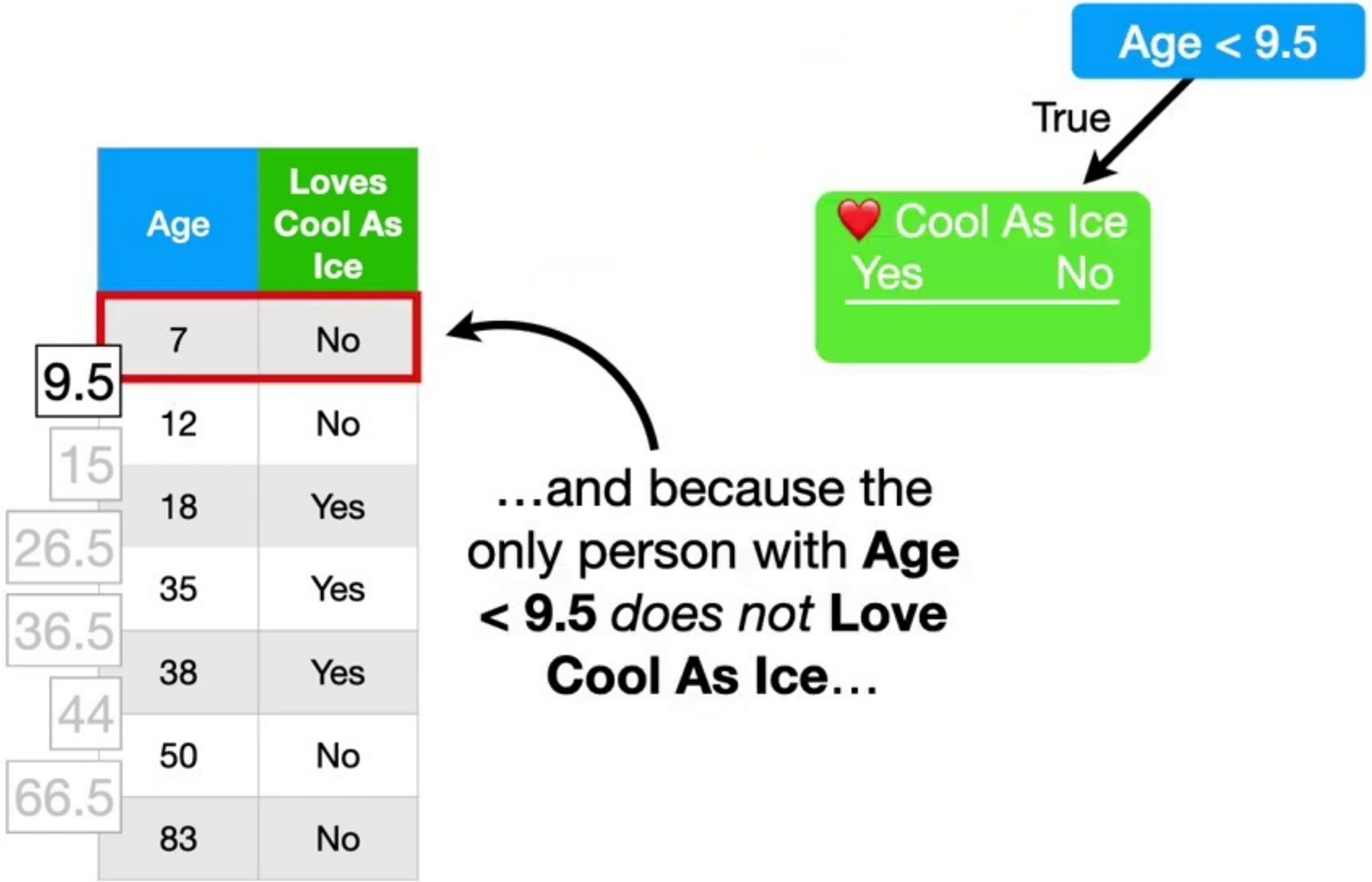
66.5 → Gini Impurity = 0.429



For example, to calculate the **Gini Impurity** for the first value…




Age < 9.5
Age
Loves Cool As Ice
7
No
9.5
12
No
15
18
Yes
26.5
35
Yes
36.5
38
Yes
44
50
No
66.5
83
No
…we put **Age < 9.5** in the **Root**…


Age < 9.5
True
❤️ Cool As Ice
Yes
No
Age
Loves Cool As Ice
7
No
12
No
18
Yes
35
Yes
38
Yes
50
No
83
No
9.5
15
26.5
36.5
44
66.5
…and because the only person with **Age < 9.5** *does not* **Love Cool As Ice**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5

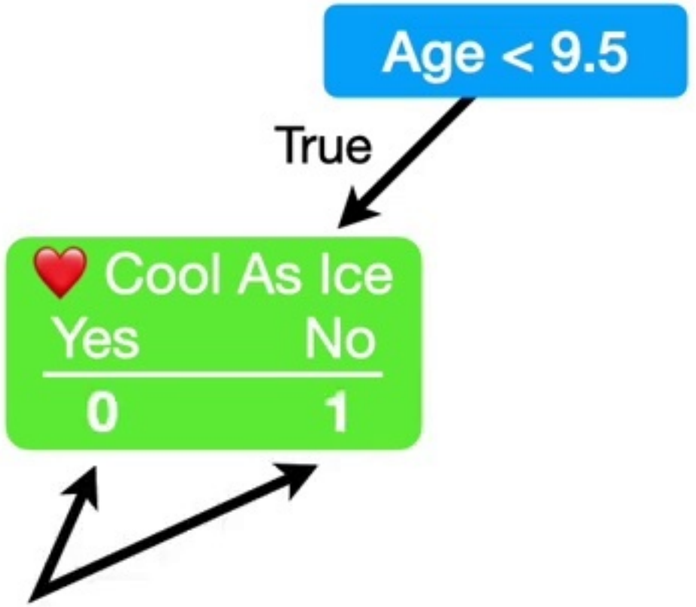
15

26.5

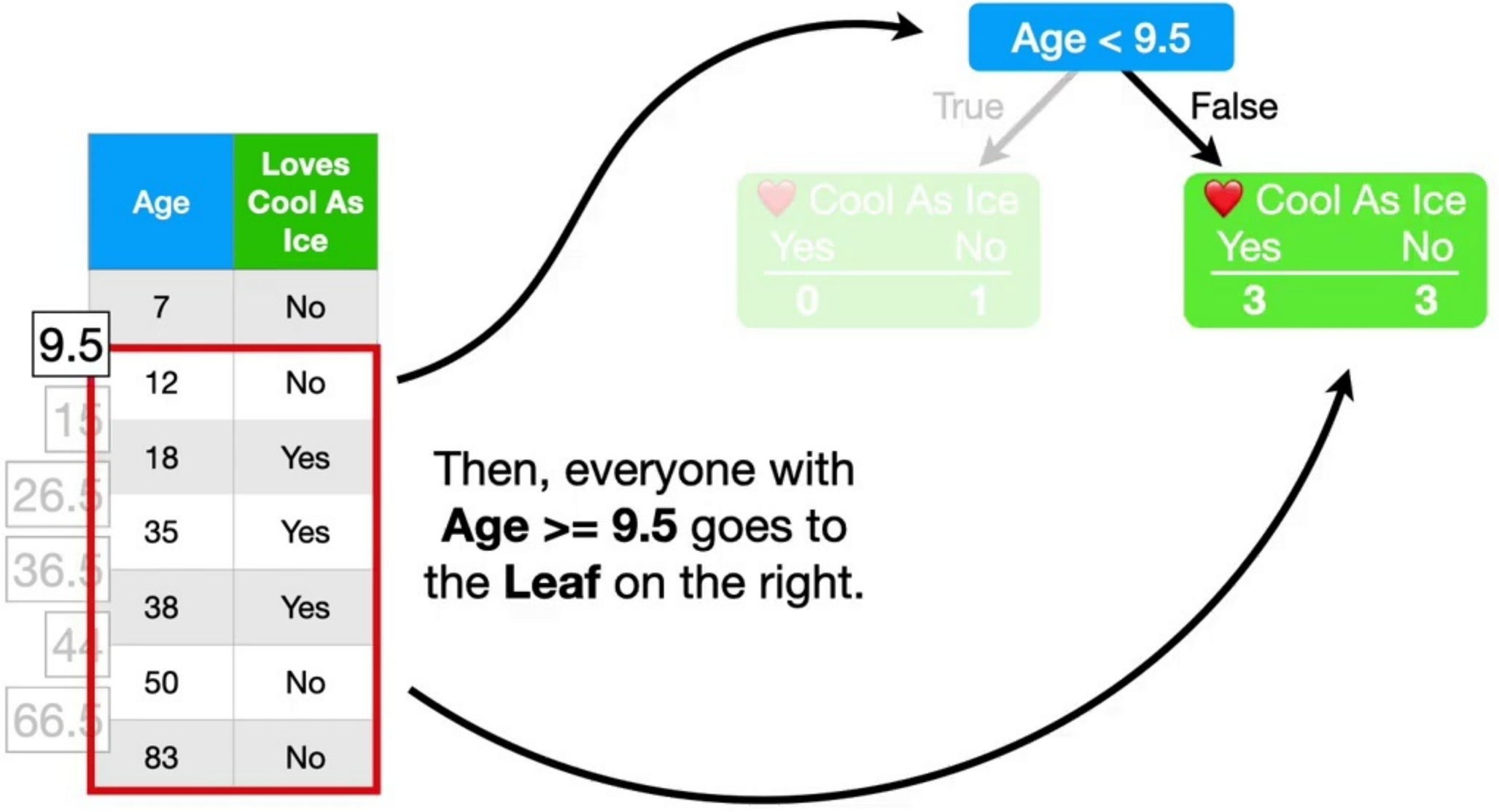
36.5

44

66.5

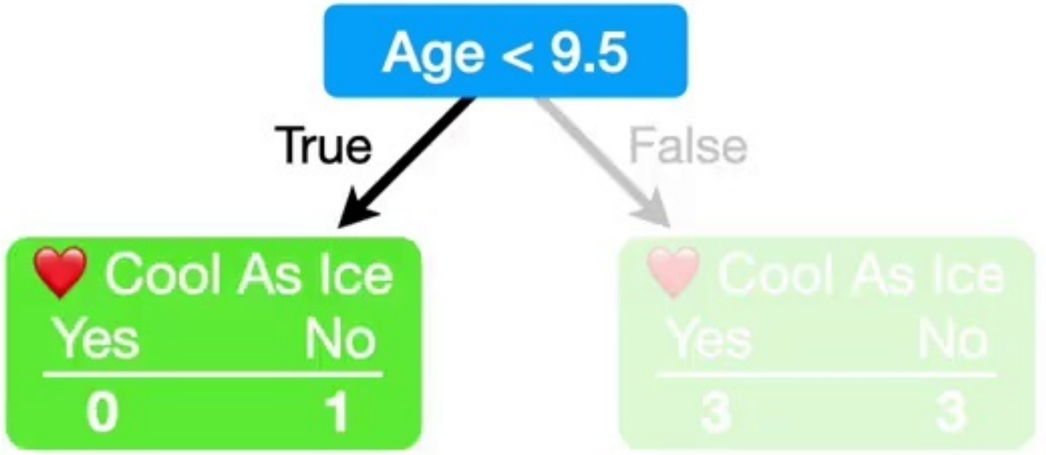


…we put a **0** under **Yes** and a **1** under **No**.


Age
Loves Cool As Ice
7
No
9.5
12
No
15
18
Yes
26.5
35
Yes
36.5
38
Yes
44
50
No
66.5
83
No
Then, everyone with
Age >= 9.5 goes to
the Leaf on the right.
Age < 9.5
True
False
Cool As Ice
Yes
No
0
1
Cool As Ice
Yes
No
3
3

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5

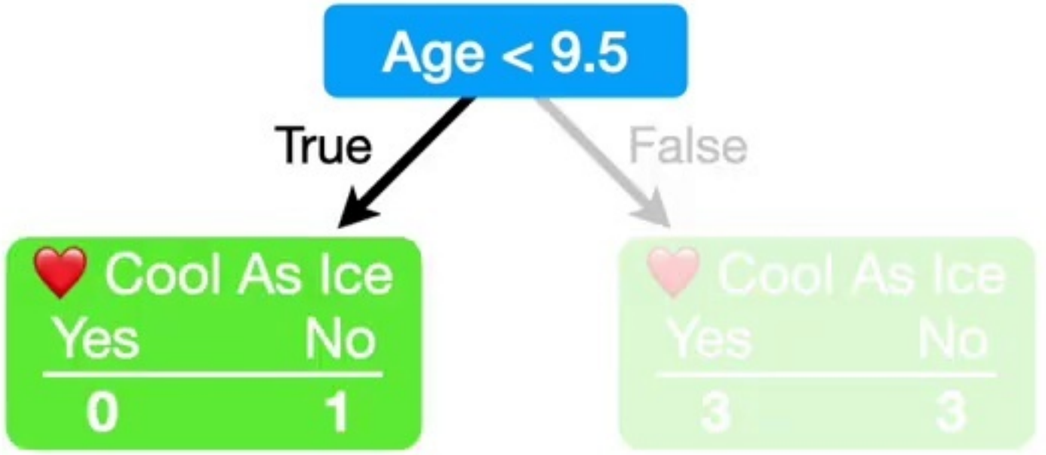


Impurity = 1 - (probability of “Yes”)^2 - (probability of “No”)^2

Now we calculate the **Gini Impurity** for the **Leaf** on the left…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



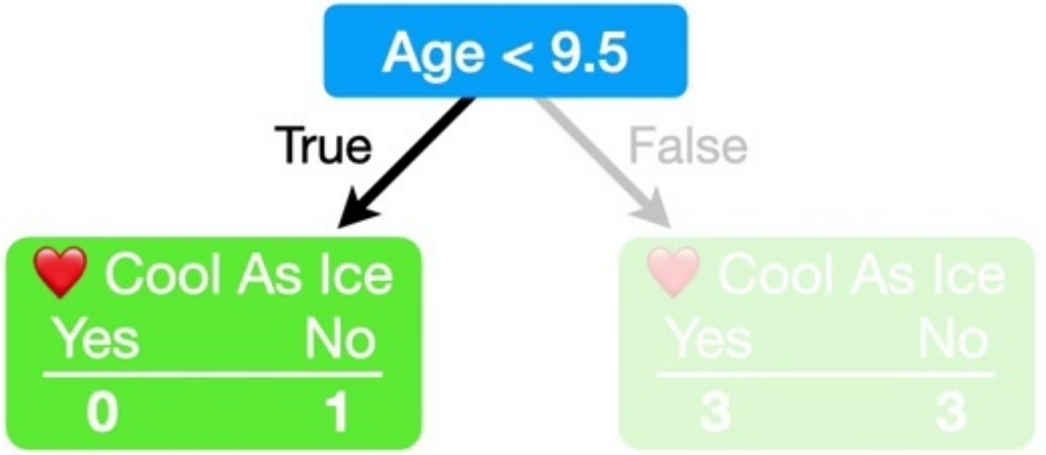
Impurity = 1 - (probability of "Yes")^2 - (probability of "No")^2

= 1

Now we calculate the **Gini Impurity** for the **Leaf** on the left…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



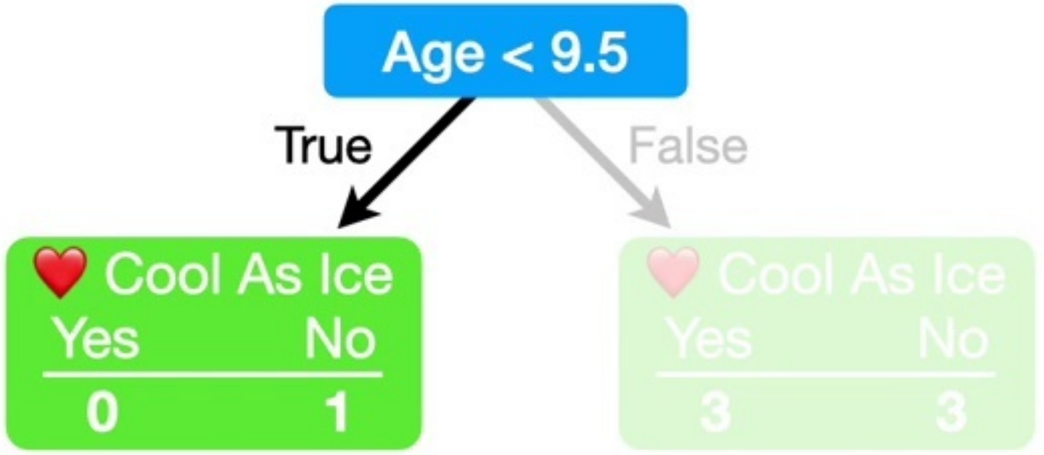
Impurity = 1 - (probability of “Yes”)^2 - (probability of “No”)^2

= 1

Now we calculate the **Gini Impurity** for the **Leaf** on the left…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



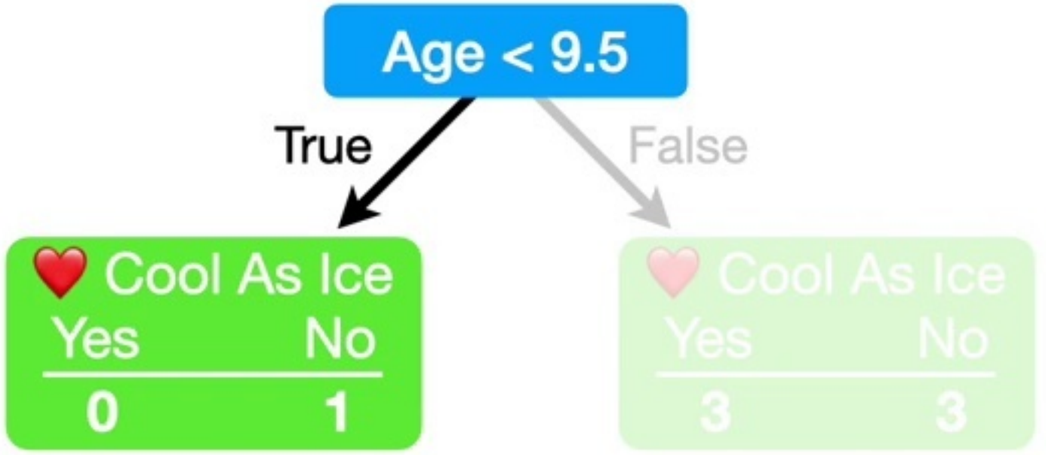
Impurity = 1 - (probability of “Yes”)^2 - (probability of “No”)^2

$$= 1 - \left(\frac{0}{0+1}\right)^2$$

Now we calculate the **Gini Impurity** for the **Leaf** on the left…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



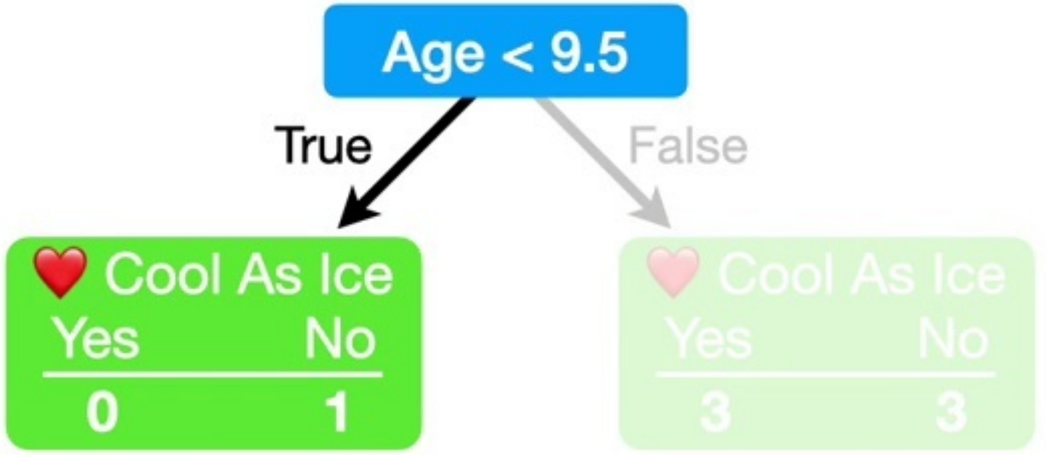
$$\text{Impurity} = 1 - (\text{probability of “Yes”})^2 - (\text{probability of “No”})^2$$

$$= 1 - \left(\frac{0}{0+1}\right)^2$$

Now we calculate the **Gini Impurity** for the **Leaf** on the left…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



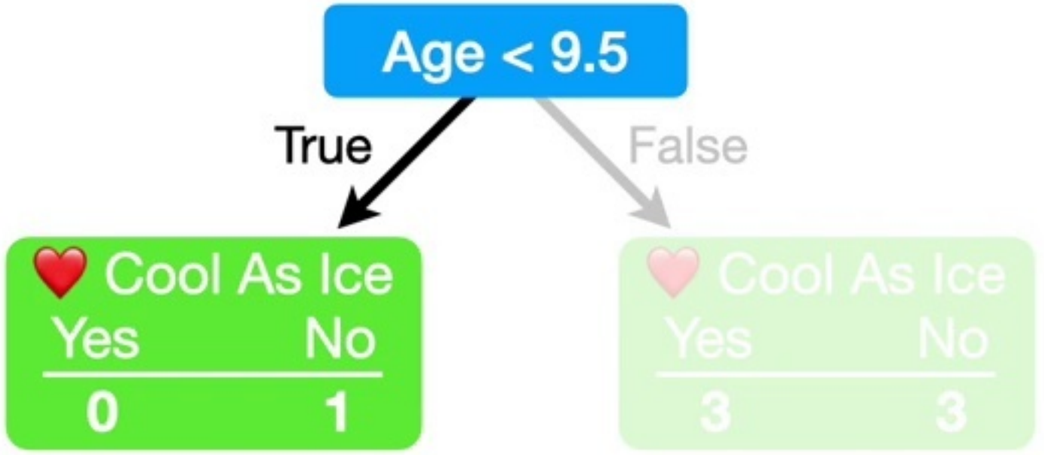
Impurity = 1 - (probability of "Yes")^2 - (probability of "No")^2

$$= 1 - \left(\frac{0}{0 + 1}\right)^2$$

Now we calculate the **Gini Impurity** for the **Leaf** on the left…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



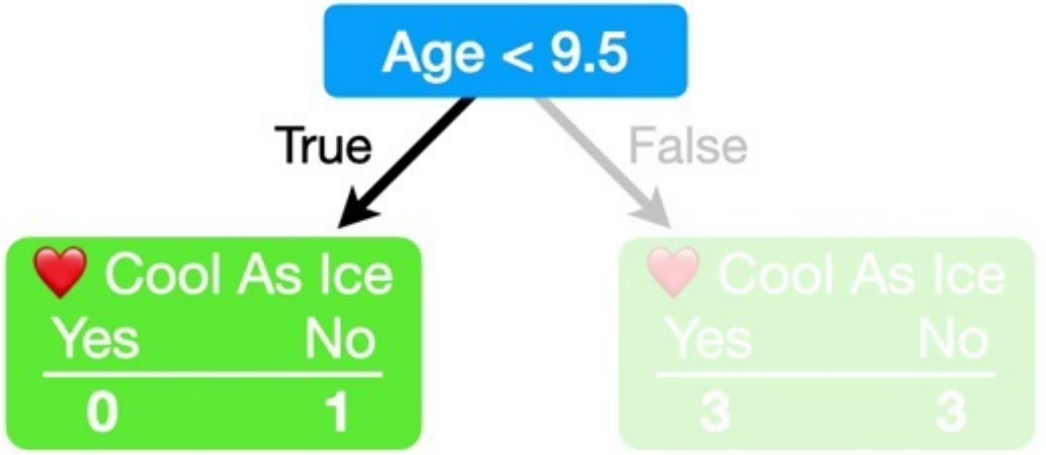
$$\text{Impurity} = 1 - (\text{probability of “Yes”})^2 - (\text{probability of “No”})^2$$

$$= 1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2$$

Now we calculate the **Gini Impurity** for the **Leaf** on the left…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



Impurity = 1 - (probability of “Yes”)^2 - (probability of “No”)^2

$$= 1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2$$

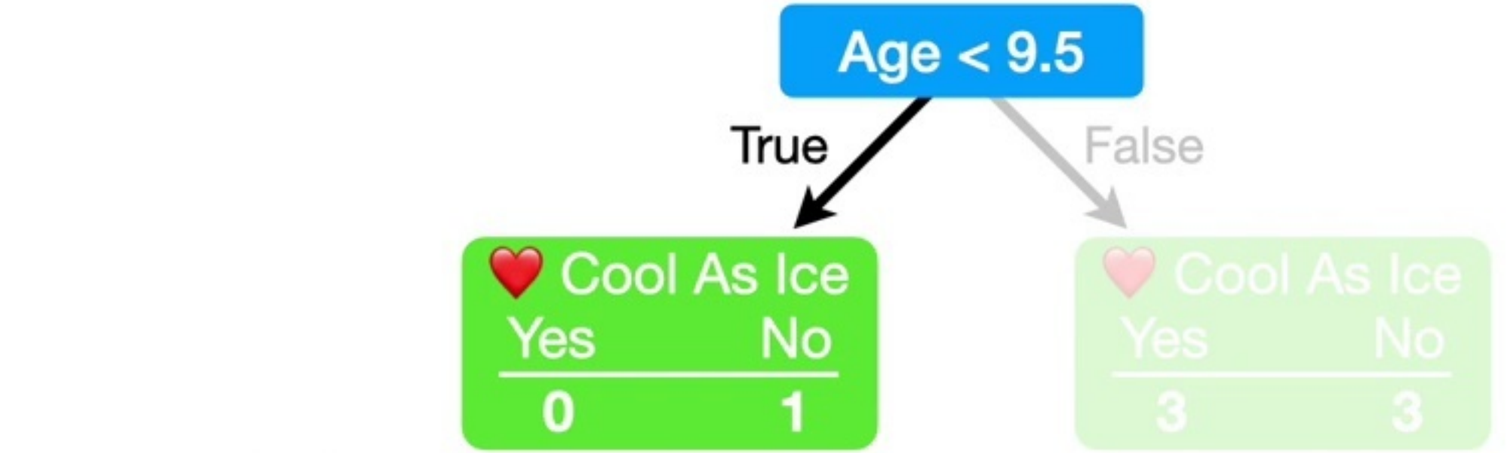
$= 0$

Now we calculate the Gini Impurity for the Leaf on the left…

…and get **0**.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



Gini Impurity = 0

Impurity = 1 - (probability of "Yes")^2 - (probability of "No")^2

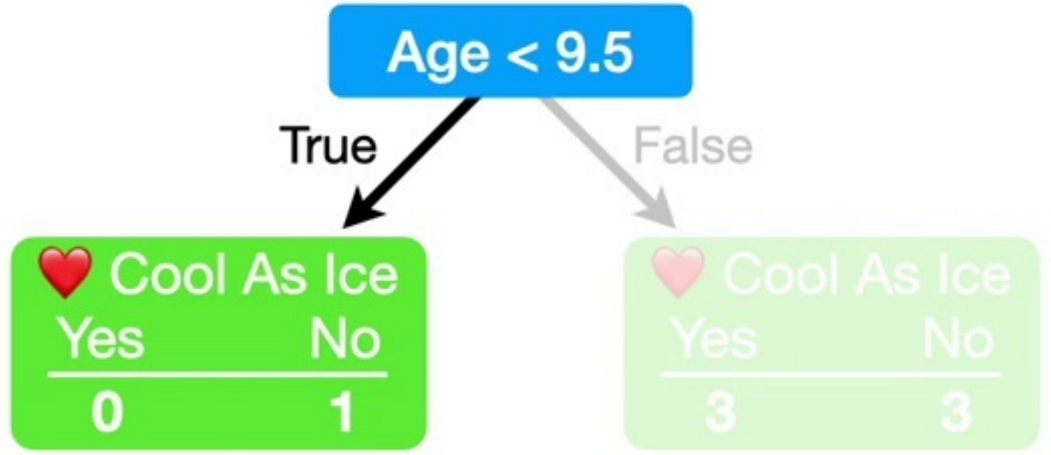
$$= 1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2$$

= 0

…and get **0**.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



Gini Impurity = 0

Impurity = 1 - (probability of "Yes")^2 - (probability of "No")^2

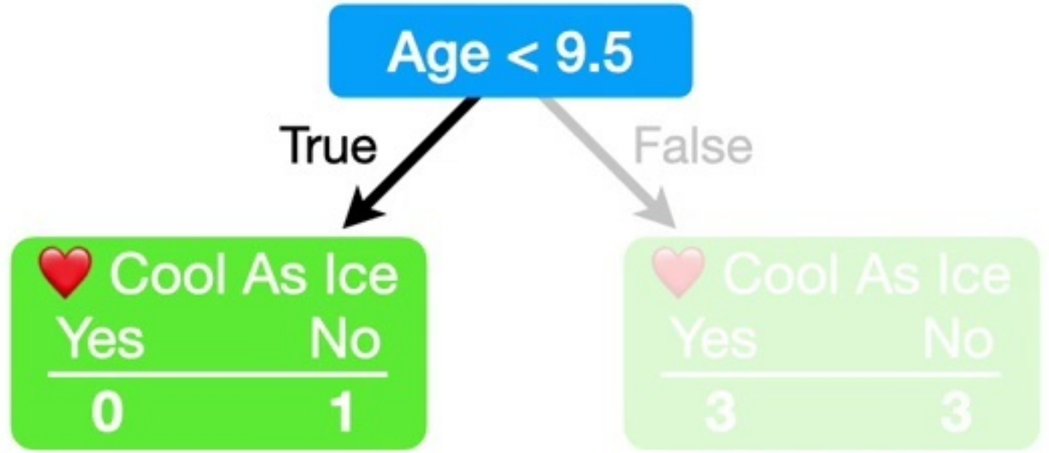
$$= 1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2$$

$$= 0$$

And this makes sense because every single person in this **Leaf** *does not* **Love Cool As Ice**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



Gini Impurity = 0

Impurity = 1 - (probability of "Yes")^2 - (probability of "No")^2

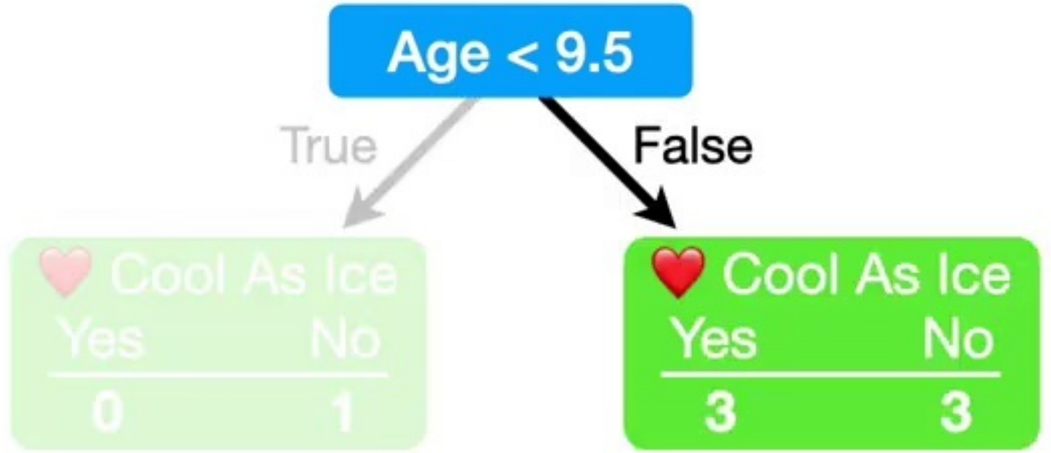
$$= 1 - \left(\frac{0}{0+1}\right)^2 - \left(\frac{1}{0+1}\right)^2$$

$$= 0$$

…so there is no **Impurity**.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5



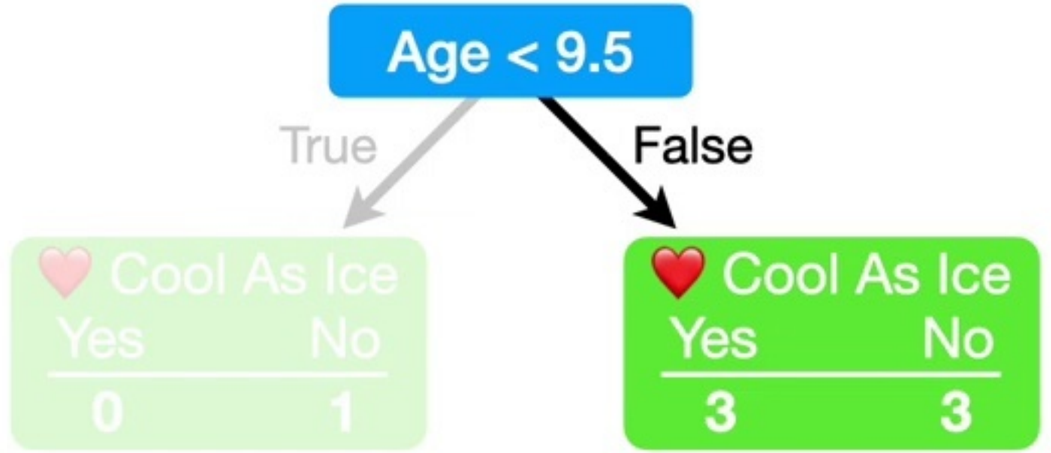
Gini Impurity = 0

Impurity = 1 - (probability of “Yes”)^2 - (probability of “No”)^2

Then we calculate the **Gini Impurity** for the **Leaf** on the right…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5



Gini Impurity = 0

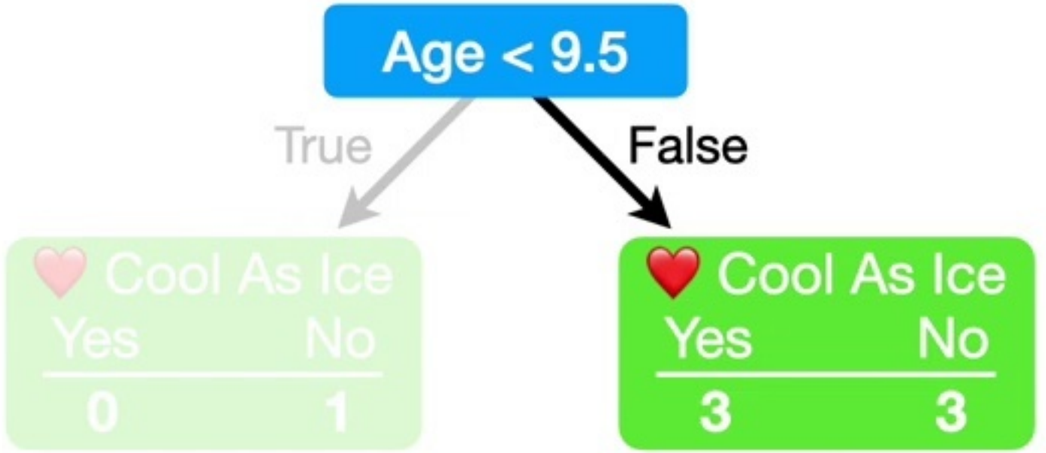
Impurity = 1 - (probability of “Yes”)^2 - (probability of “No”)^2

= 1

Then we calculate the **Gini Impurity** for the **Leaf** on the right…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



Gini Impurity = 0

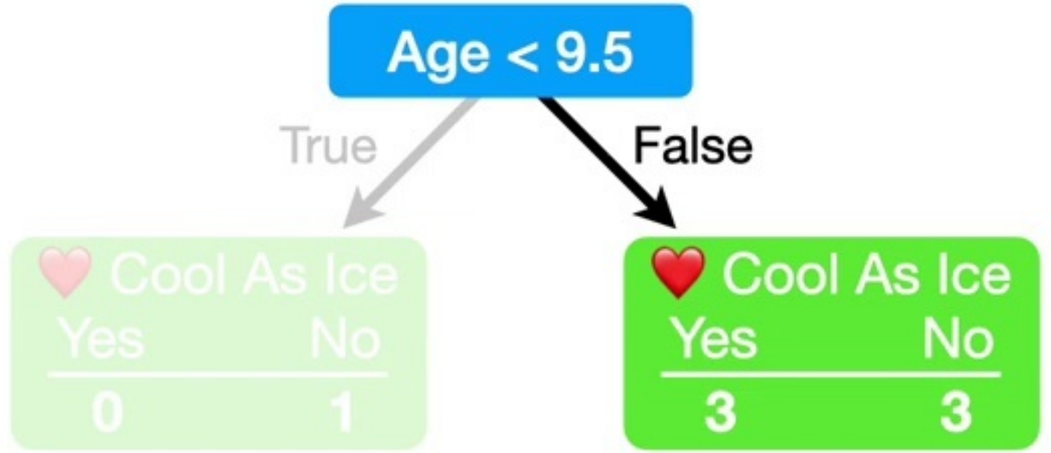
$$\text{Impurity} = 1 - (\text{probability of “Yes”})^2 - (\text{probability of “No”})^2$$

$$= 1 - \left(\frac{3}{3+3}\right)^2$$

Then we calculate the **Gini Impurity** for the **Leaf** on the right…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5

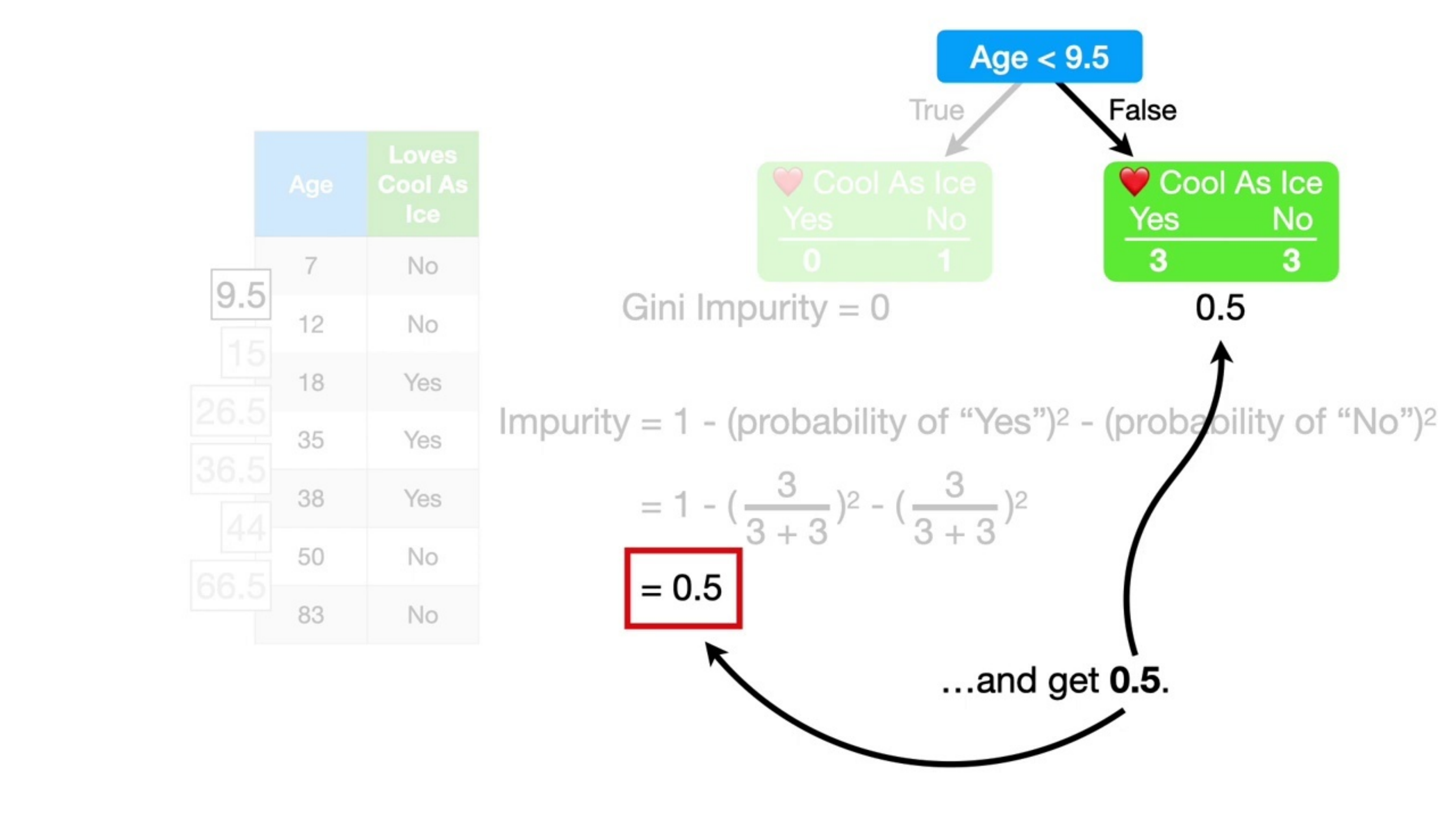


Gini Impurity = 0

$$\text{Impurity} = 1 - (\text{probability of “Yes”})^2 - (\text{probability of “No”})^2$$

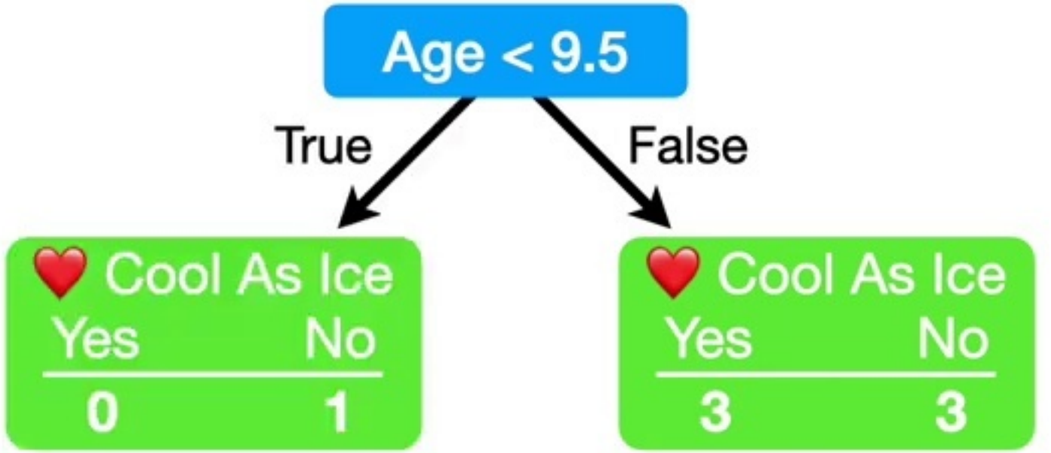
$$= 1 - \left(\frac{3}{3+3}\right)^2 - \left(\frac{3}{3+3}\right)^2$$

Then we calculate the **Gini Impurity** for the **Leaf** on the right…


Age
Loves Cool As Ice
7
No
12
No
18
Yes
35
Yes
38
Yes
50
No
83
No
9.5
15
26.5
36.5
44
66.5
Age < 9.5
True
False
Cool As Ice
Yes
No
0
1
Cool As Ice
Yes
No
3
3
Gini Impurity = 0
0.5
Impurity = 1 - (probability of “Yes”)² - (probability of “No”)²
= 1 - (3 / 3 + 3)² - (3 / 3 + 3)²
= 0.5
…and get 0.5.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5



Gini Impurity = 0

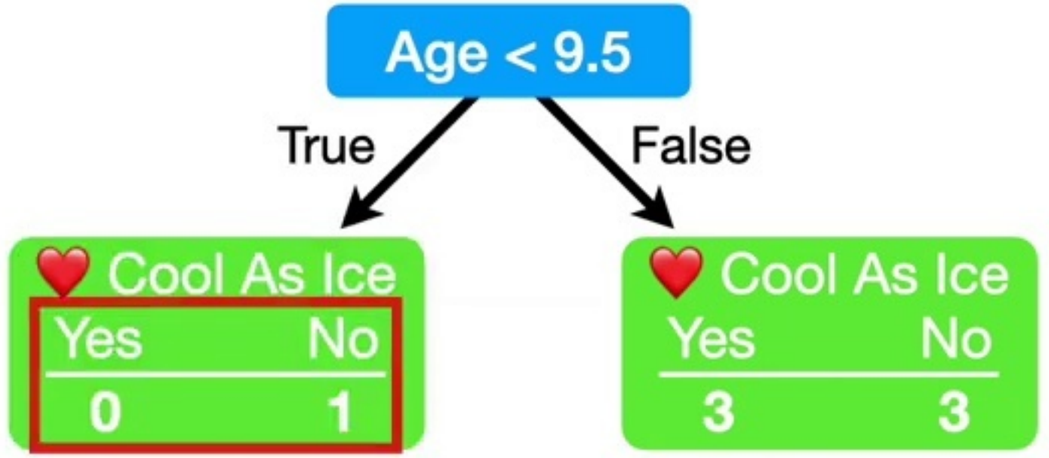
0.5

Total Gini Impurity =

Now we calculate the **Weighted Average** of the two **Impurities** to get the **Total Gini Impurity**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5



Gini Impurity = 0

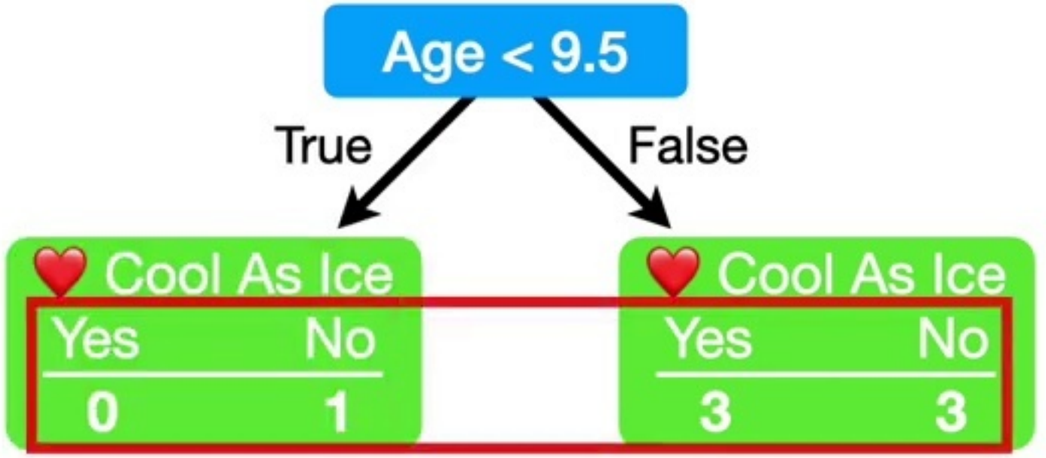
0.5

$$\text{Total Gini Impurity} = \left(\frac{1}{1+6}\right)$$

Now we calculate the **Weighted Average** of the two **Impurities** to get the **Total Gini Impurity**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



Gini Impurity = 0

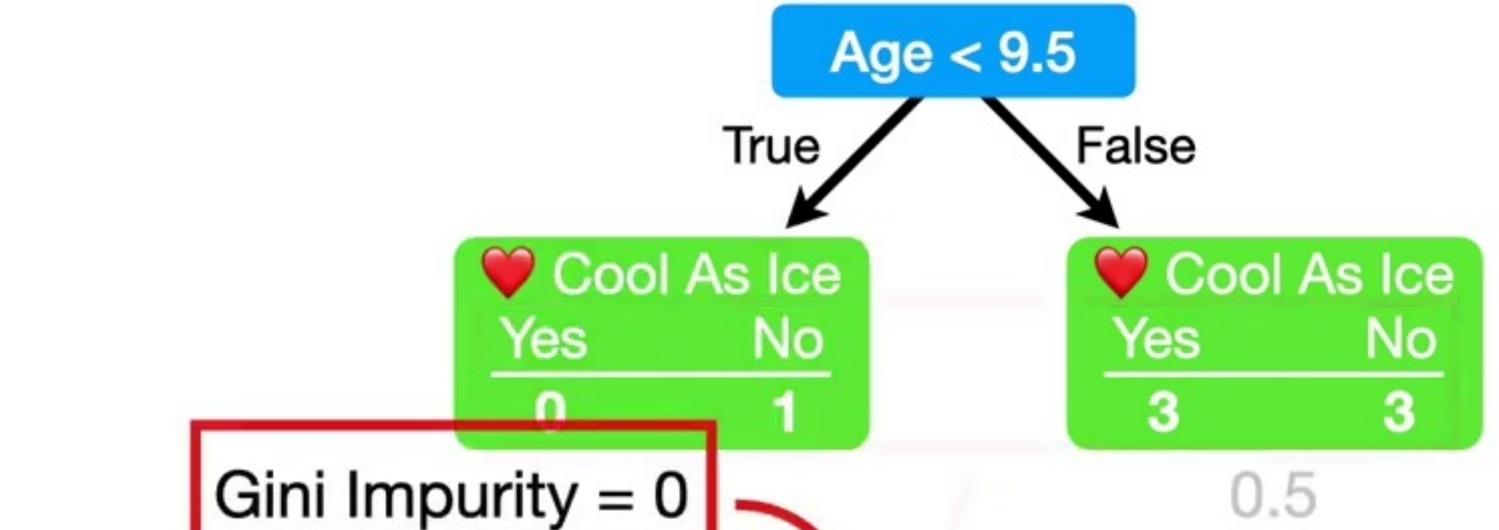
0.5

$$\text{Total Gini Impurity} = \left(\frac{1}{1+6}\right)$$

Now we calculate the **Weighted Average** of the two **Impurities** to get the **Total Gini Impurity**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5



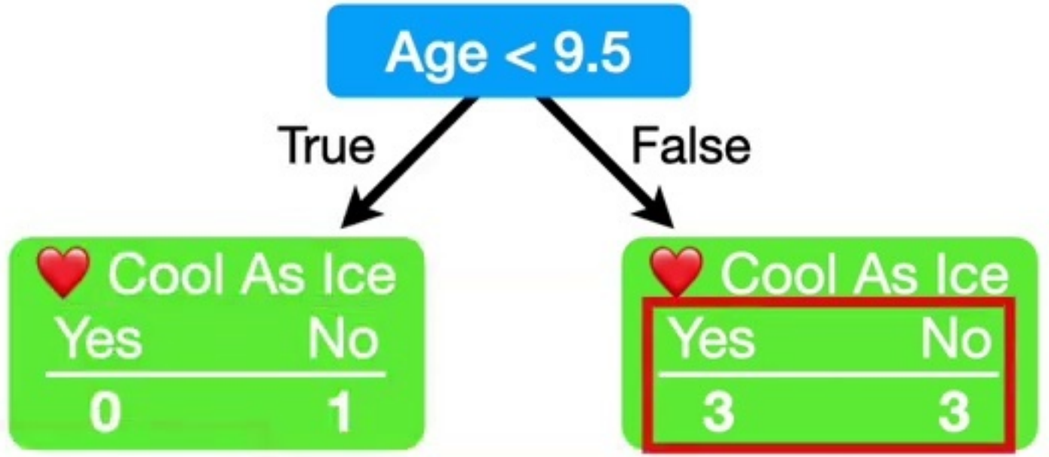
Gini Impurity = 0

$$\text{Total Gini Impurity} = \left(\frac{1}{1+6}\right) 0$$

Now we calculate the **Weighted Average** of the two **Impurities** to get the **Total Gini Impurity**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5



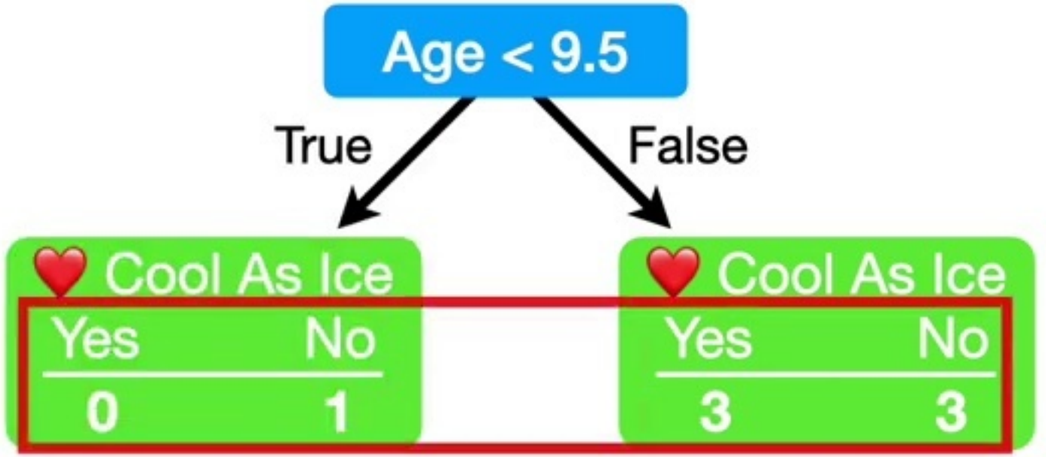
Gini Impurity = 0

$$\text{Total Gini Impurity} = \left(\frac{1}{1+6}\right) 0 + \left(\frac{6}{1+6}\right)$$

Now we calculate the **Weighted Average** of the two **Impurities** to get the **Total Gini Impurity**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5



Gini Impurity = 0

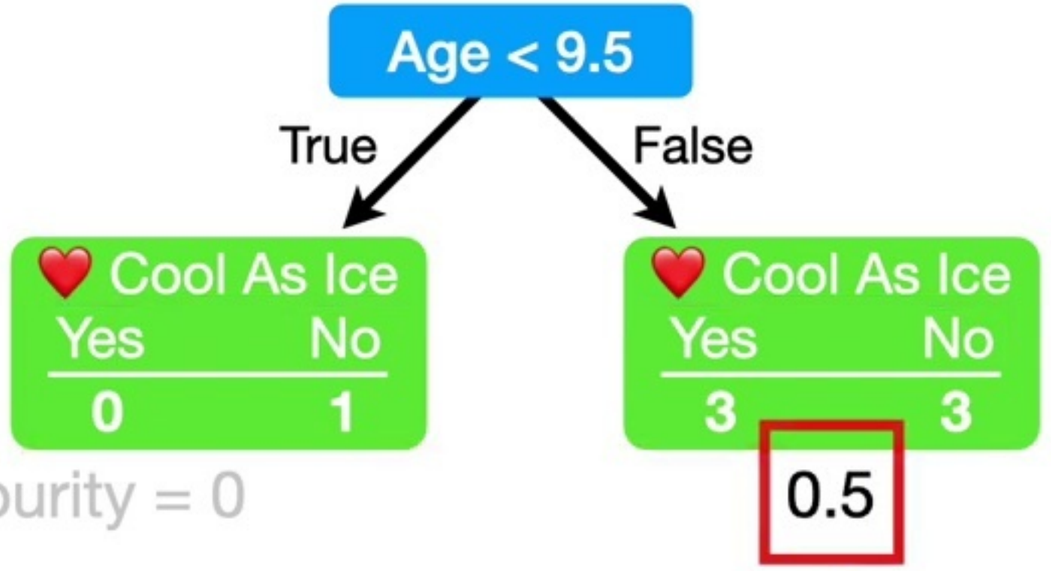
0.5

$$\text{Total Gini Impurity} = \left(\frac{1}{1+6}\right) 0 + \left(\frac{6}{1+6}\right)$$

Now we calculate the **Weighted Average** of the two **Impurities** to get the **Total Gini Impurity**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5

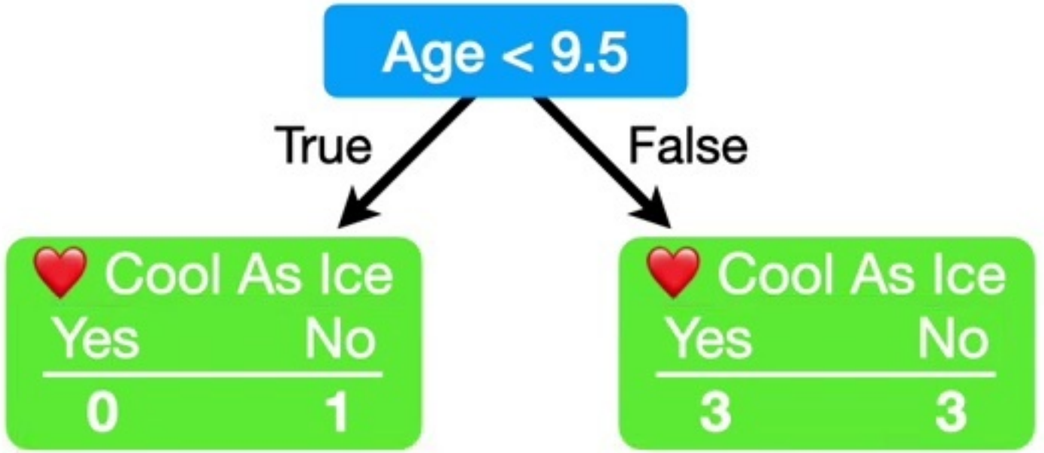


$$\text{Total Gini Impurity} = \left(\frac{1}{1+6}\right) 0 + \left(\frac{6}{1+6}\right) 0.5$$

Now we calculate the **Weighted Average** of the two **Impurities** to get the **Total Gini Impurity**…

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5

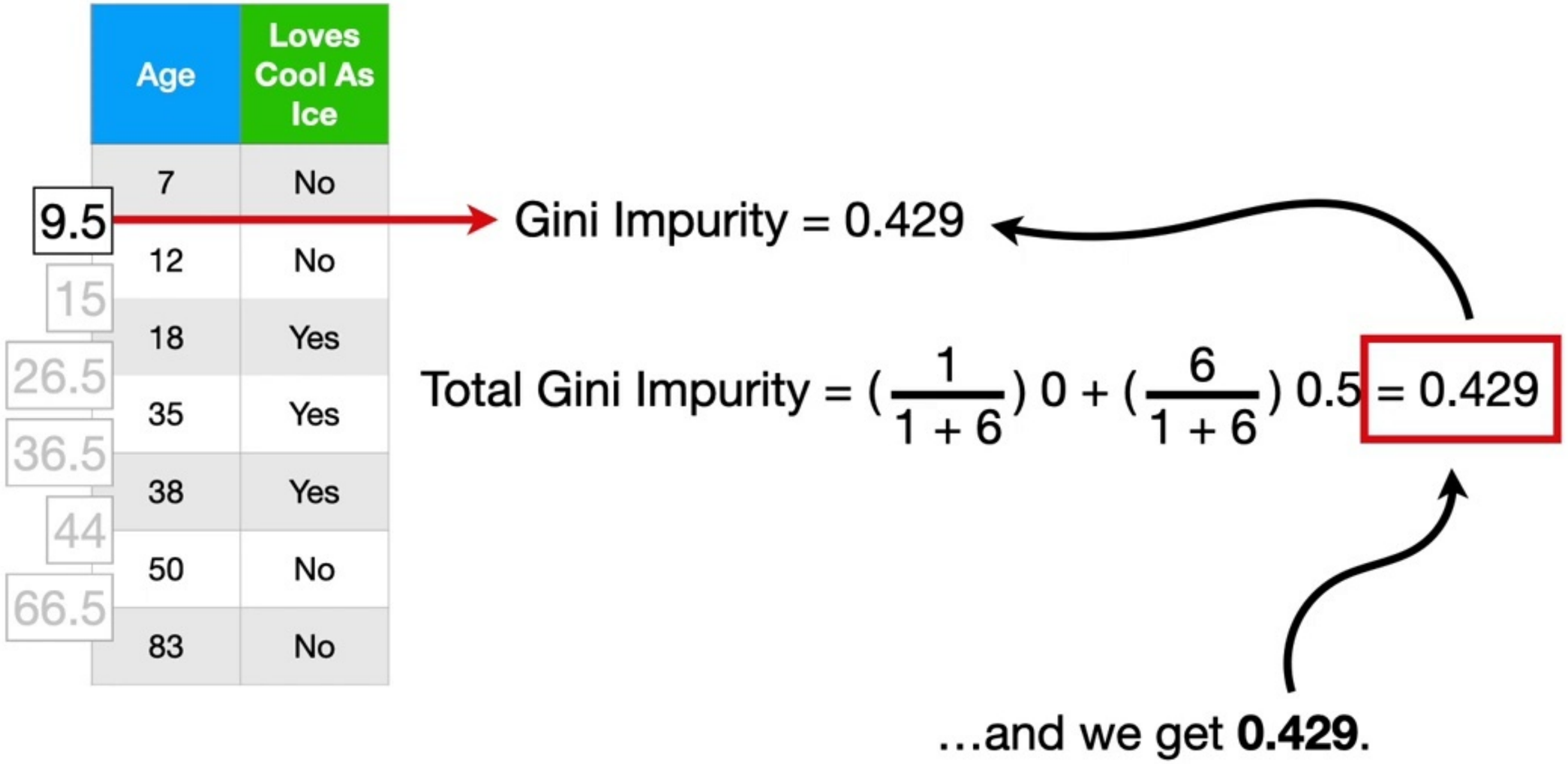


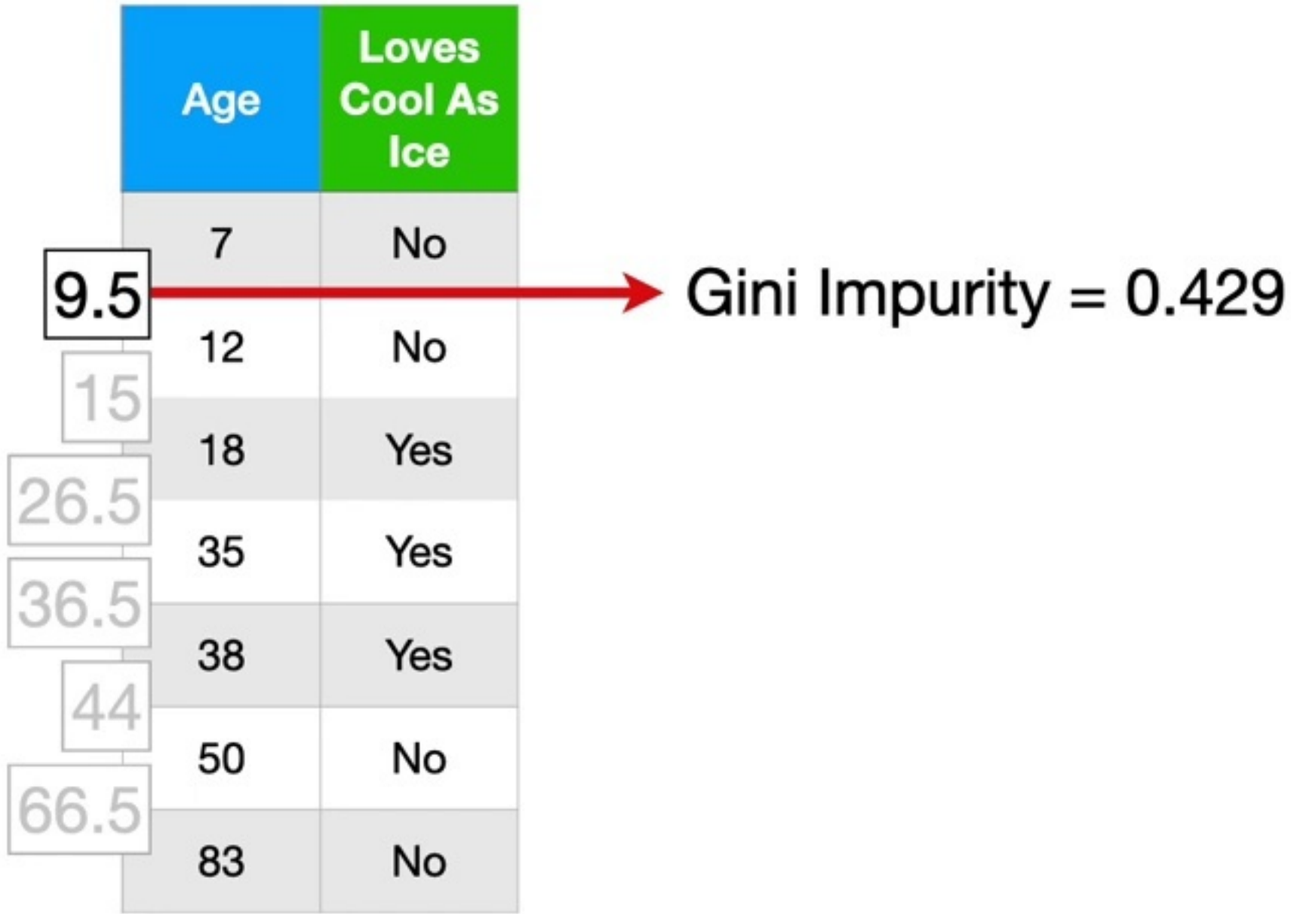
Gini Impurity = 0

0.5

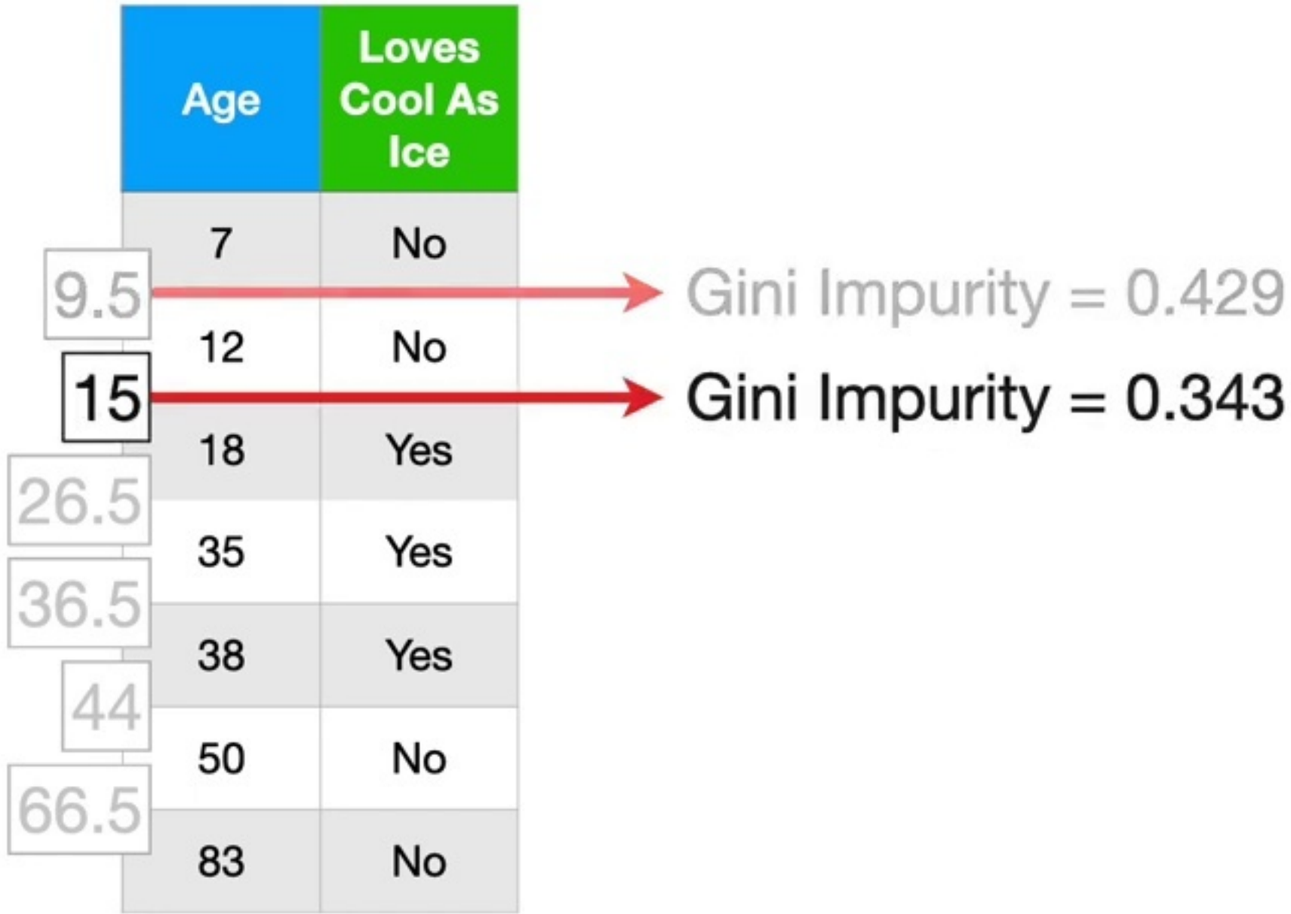
$$\text{Total Gini Impurity} = \left(\frac{1}{1+6}\right) 0 + \left(\frac{6}{1+6}\right) 0.5 = 0.429$$

…and we get **0.429**.

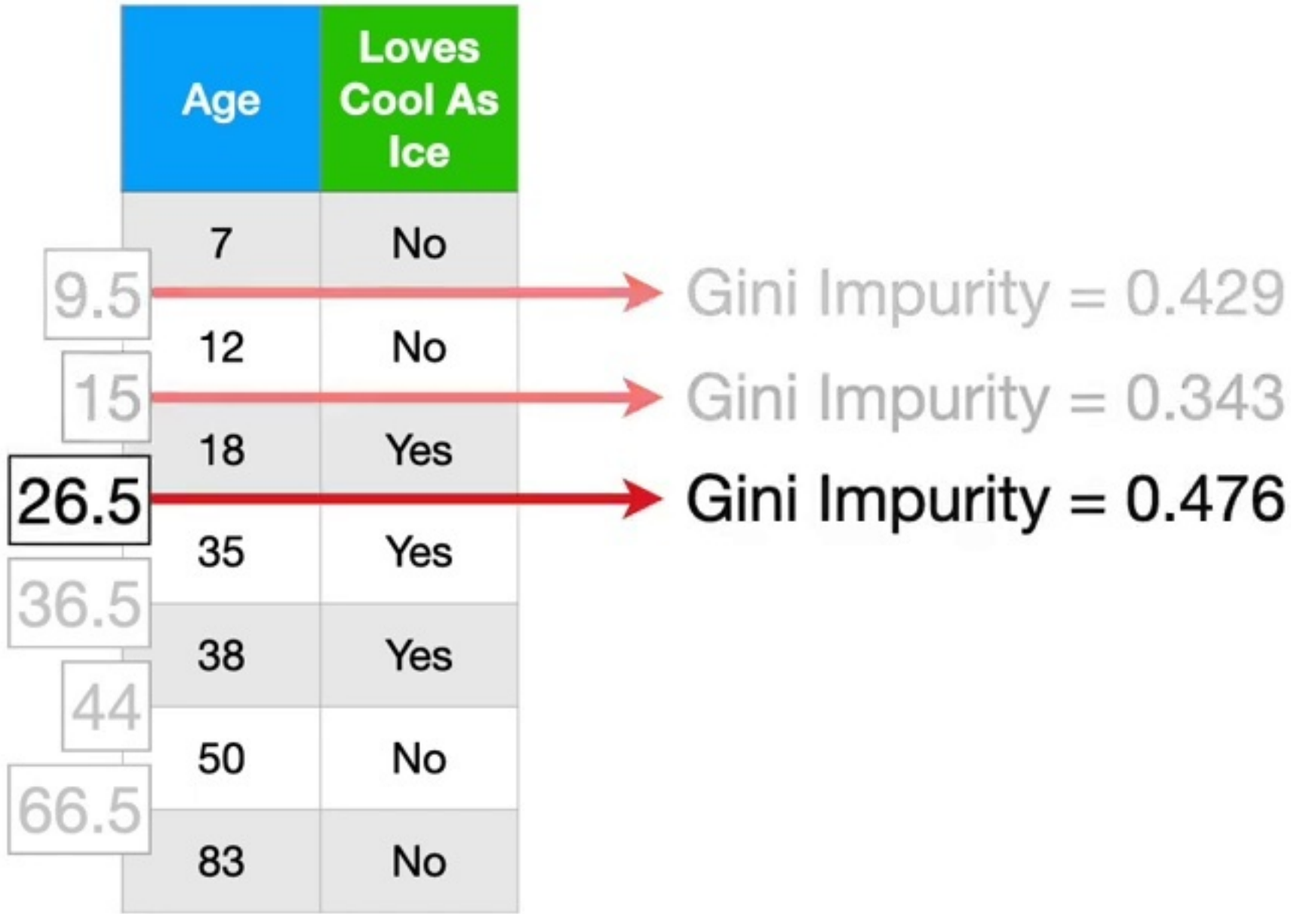

Age
Loves Cool As Ice
7
No
12
No
18
Yes
35
Yes
38
Yes
50
No
83
No
9.5
15
26.5
36.5
44
66.5
Gini Impurity = 0.429
Total Gini Impurity = (1 / 1 + 6) 0 + (6 / 1 + 6) 0.5 = 0.429
…and we get 0.429.



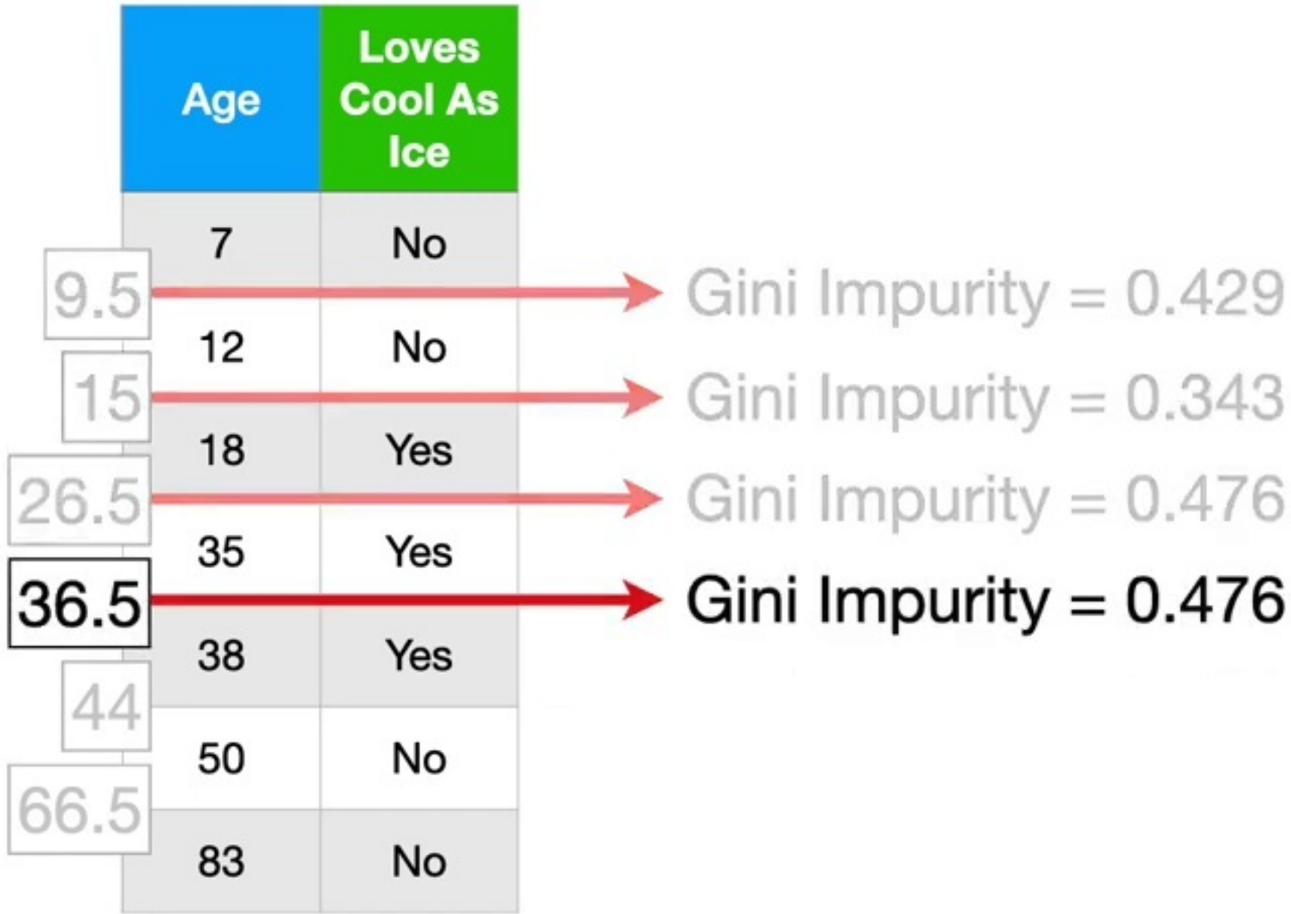
Likewise, we calculate the **Gini Impurities** for all of the other candidate values.



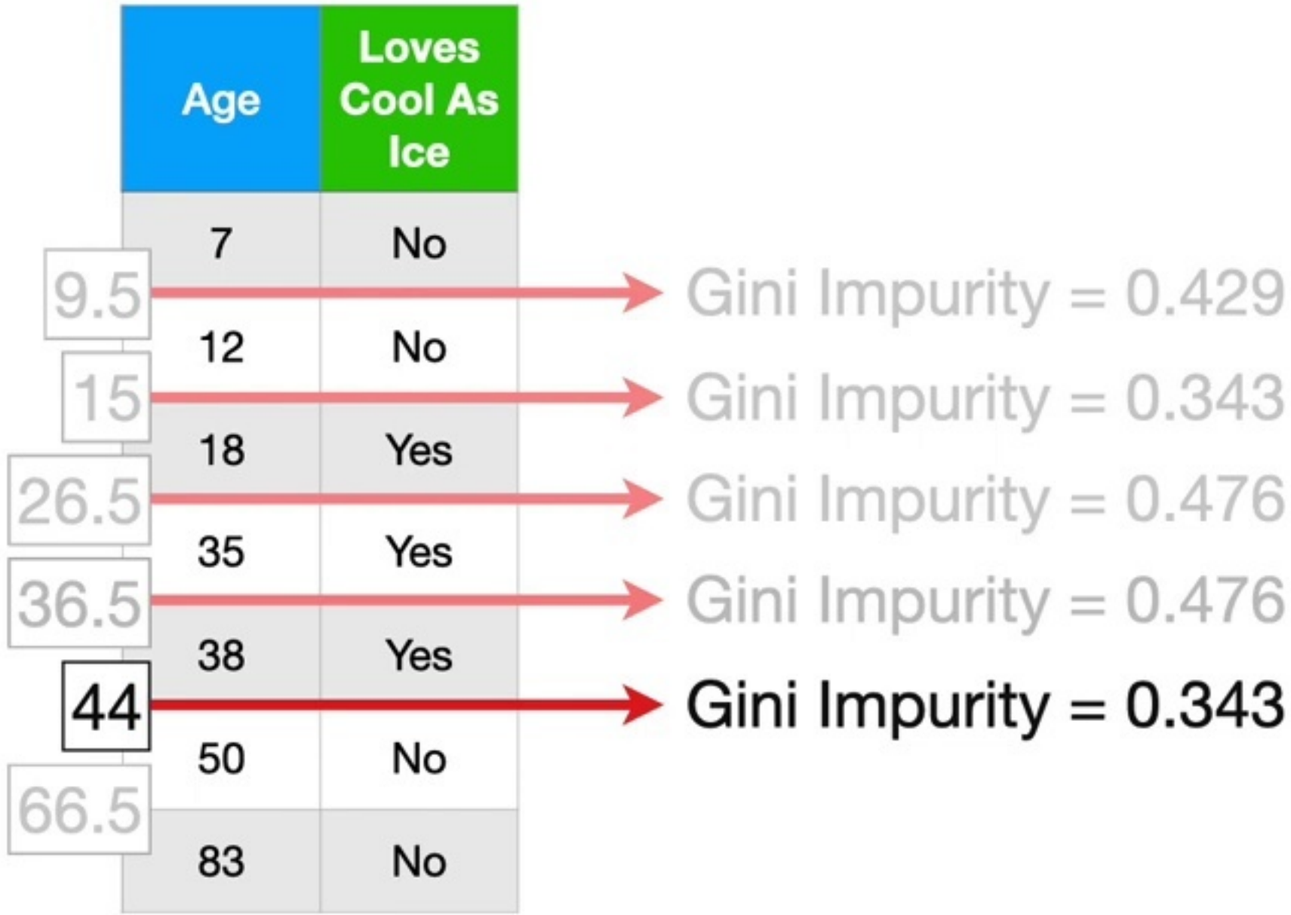
Likewise, we calculate the **Gini Impurities** for all of the other candidate values.



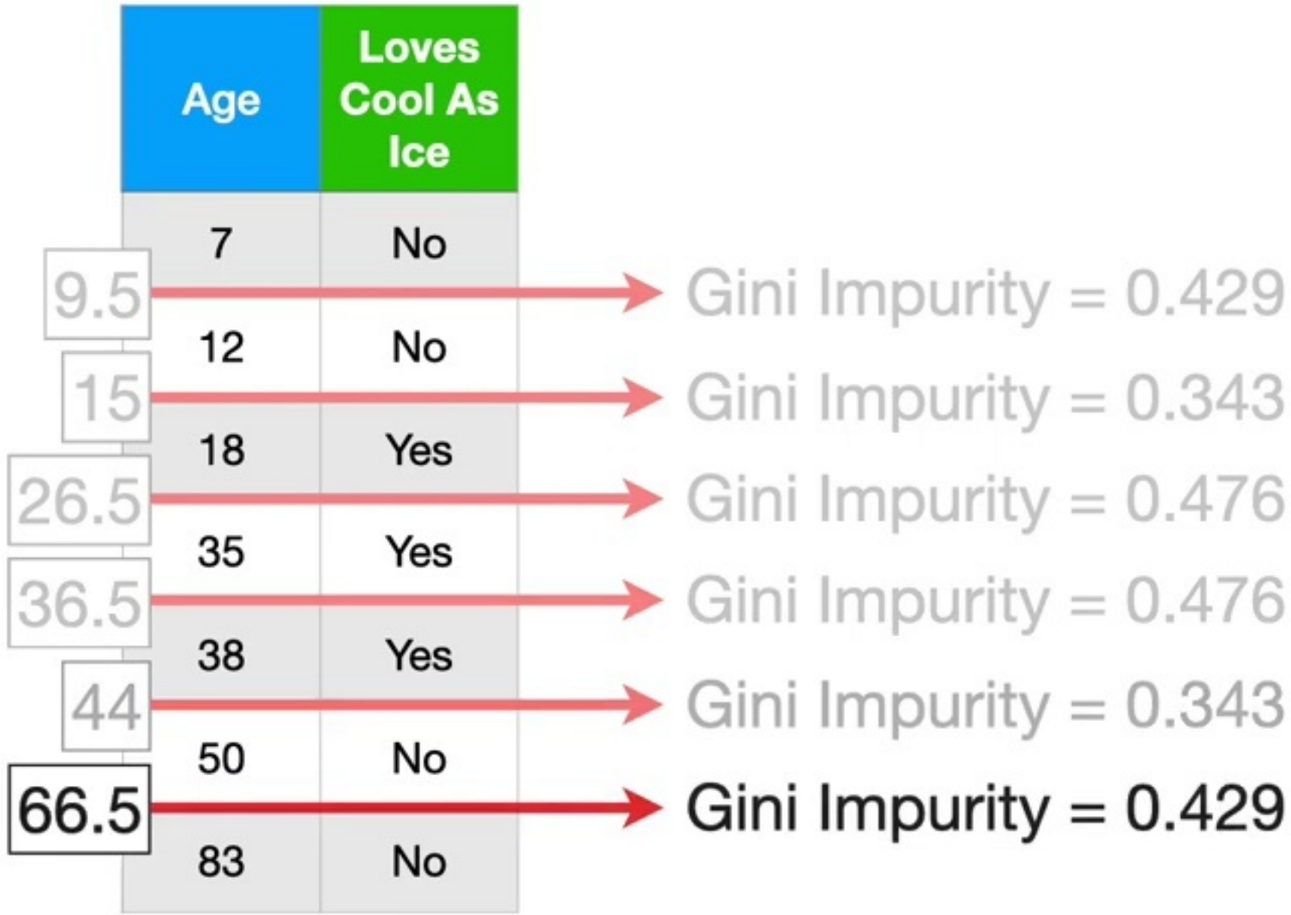
Likewise, we calculate the **Gini Impurities** for all of the other candidate values.



Likewise, we calculate the **Gini Impurities** for all of the other candidate values.

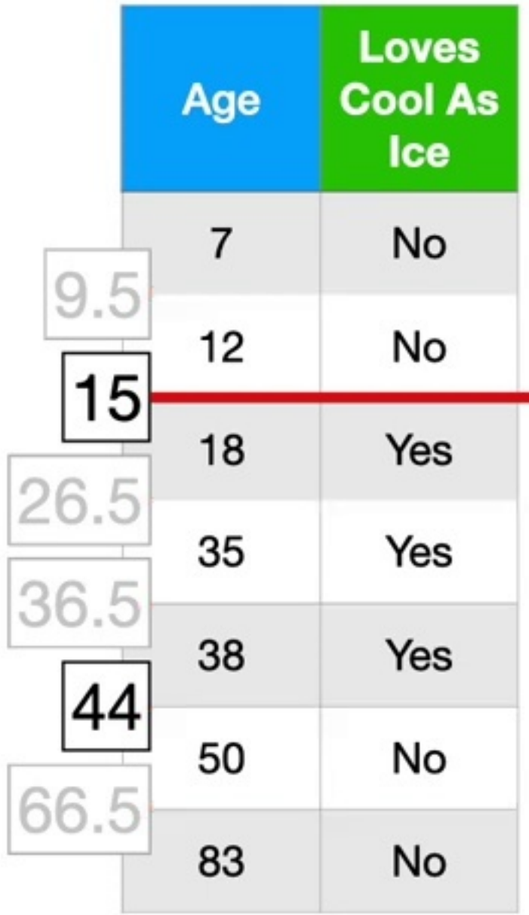


Likewise, we calculate the **Gini Impurities** for all of the other candidate values.



Likewise, we calculate the **Gini Impurities** for all of the other candidate values.


Age
Loves Cool As Ice
7
No
12
No
18
Yes
35
Yes
38
Yes
50
No
83
No
9.5
15
26.5
36.5
44
66.5
Gini Impurity = 0.429
Gini Impurity = 0.343
Gini Impurity = 0.476
Gini Impurity = 0.476
Gini Impurity = 0.343
Gini Impurity = 0.429
These two candidate thresholds, 15 and 44, are tied for the lowest Impurity, 0.343…



Gini Impurity = 0.343

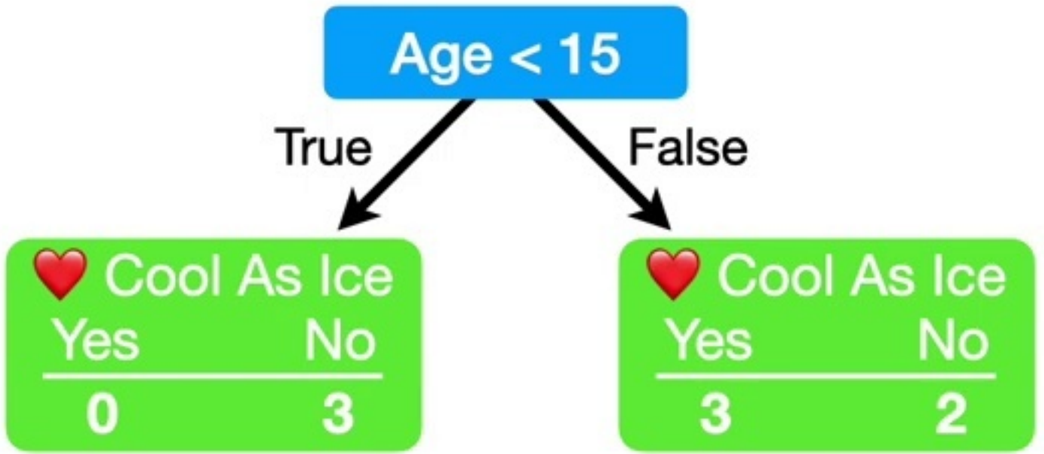
…so we can pick either one. In this case, we'll pick **15**.

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5, 15, 26.5, 36.5, 44, 66.5

Gini Impurity = 0.343

…so we can pick either one. In this case, we'll pick **15**.

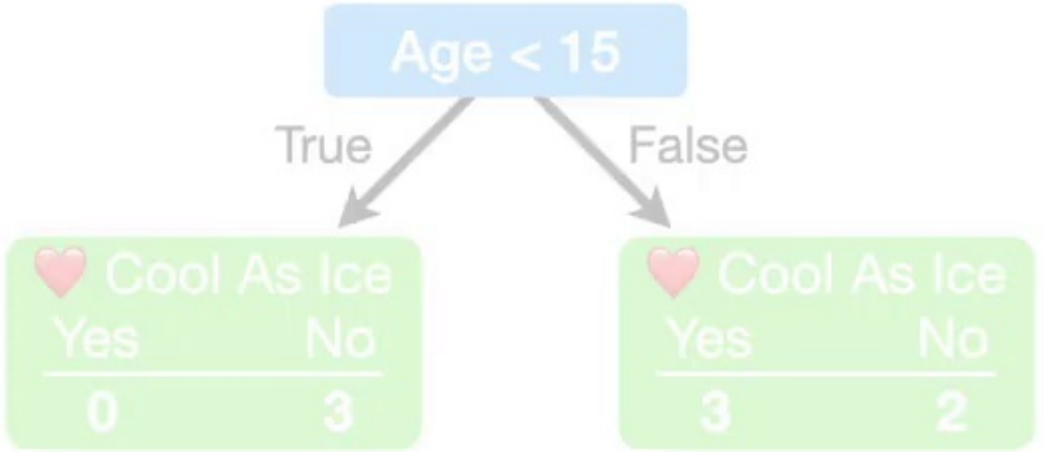


Gini Impurity for Age < 15 = 0.343

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

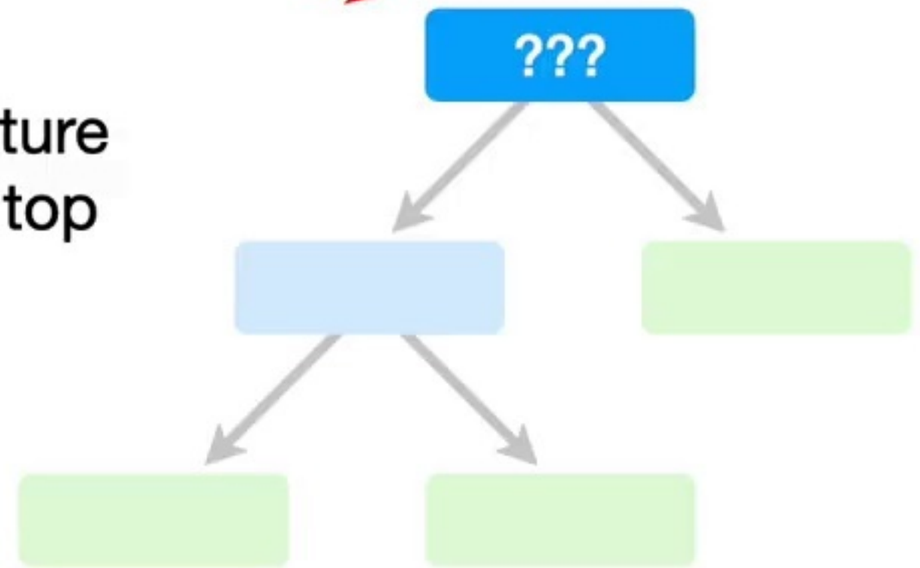
However, remember that we are comparing **Gini Impurity** values for **Age**, **Loves Popcorn** and **Loves Soda**...

Gini Impurity for Age < 15 = 0.343

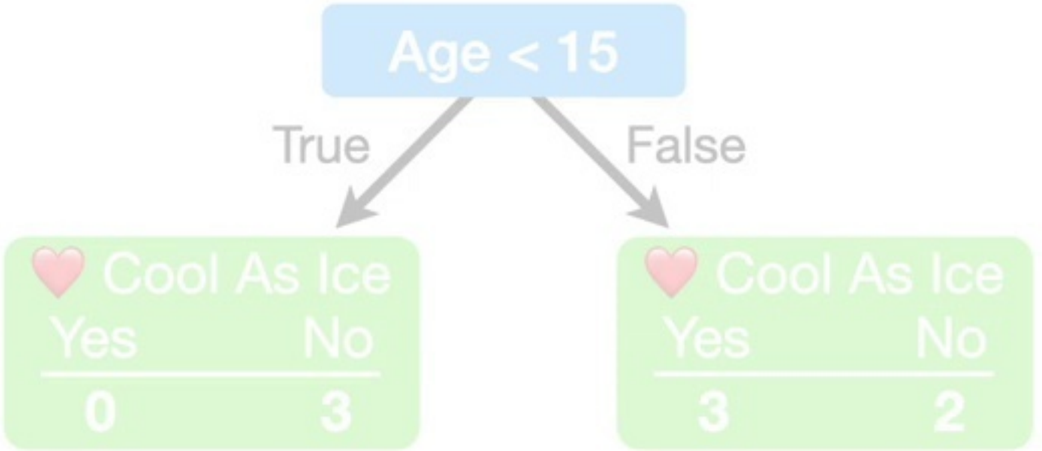


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

…to decide which feature should be at the very top of the tree.

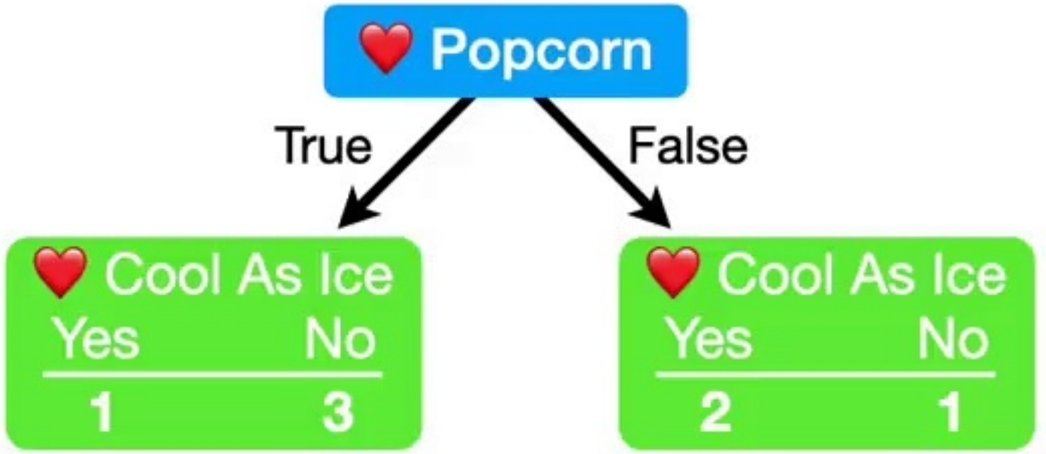


Gini Impurity for Age < 15 = 0.343

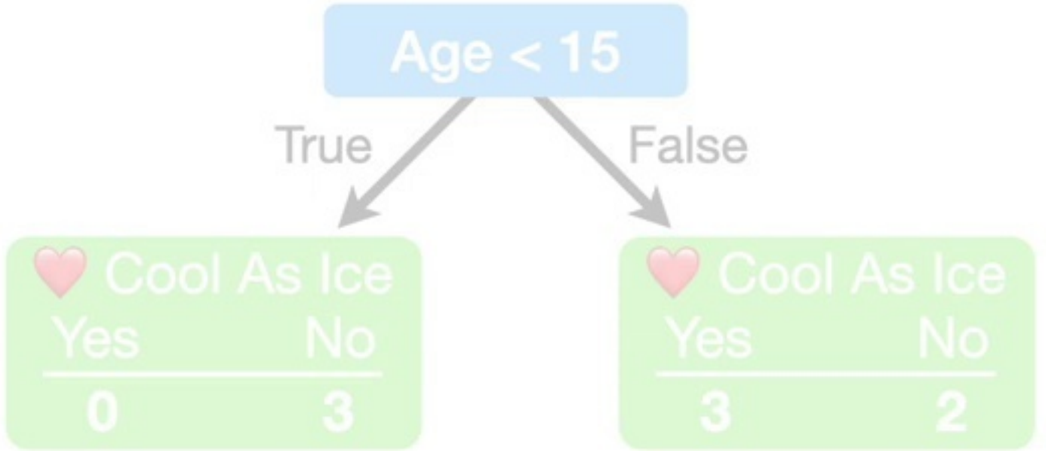


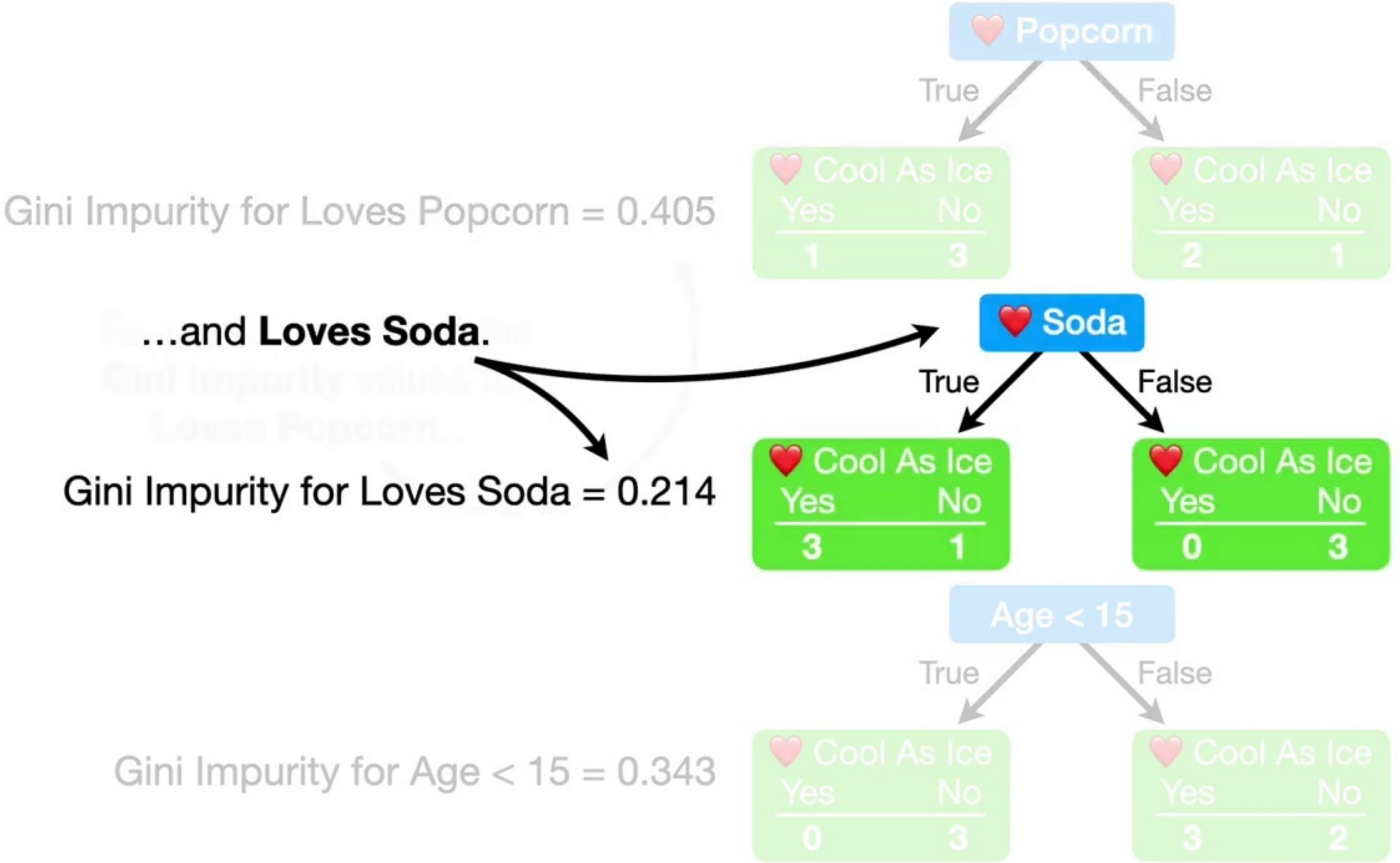
Gini Impurity for Loves Popcorn = 0.405

Earlier we calculated the **Gini Impurity** values for **Loves Popcorn**…



Gini Impurity for Age < 15 = 0.343



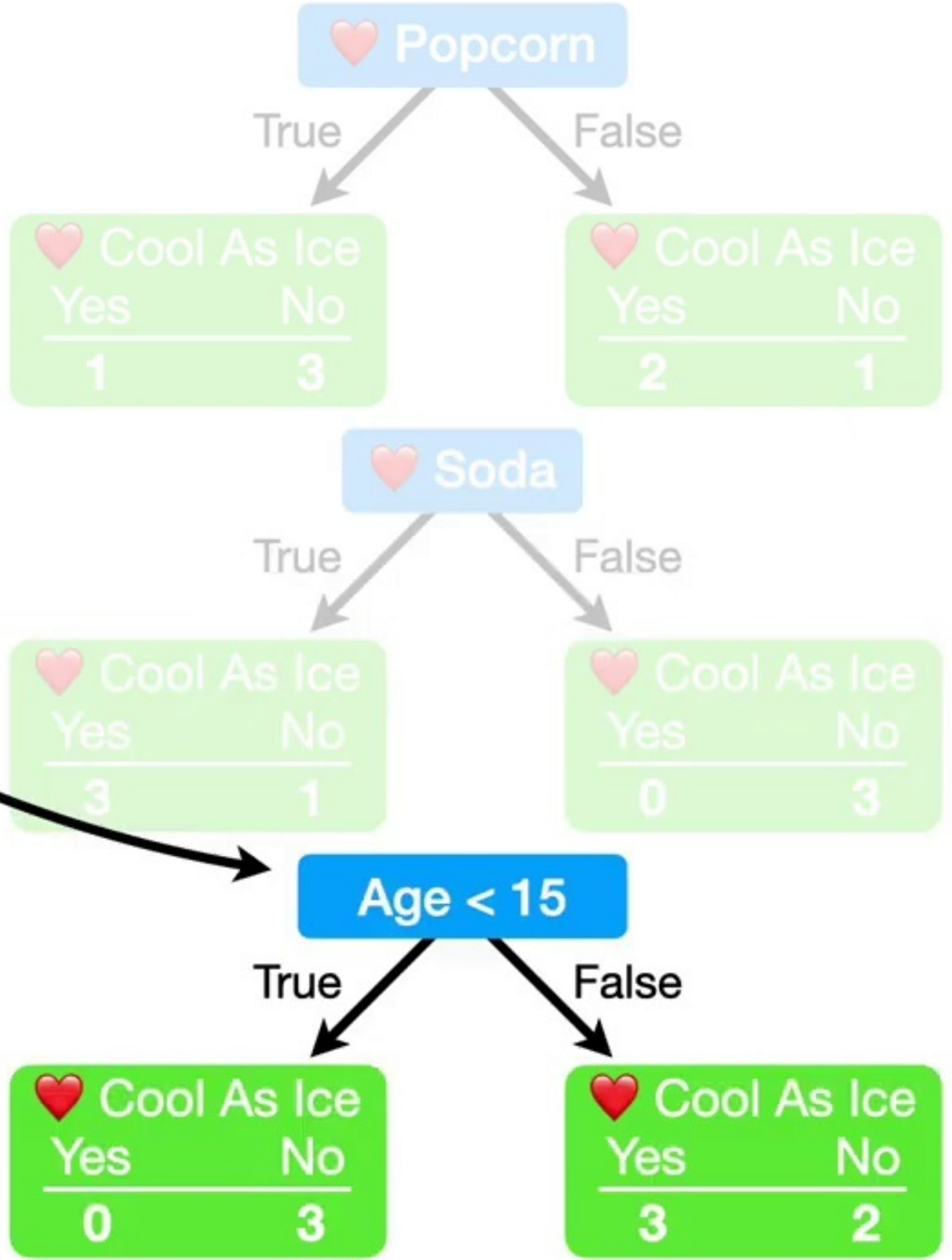

Popcorn
True
False
Cool As Ice
Yes No
1 3
Cool As Ice
Yes No
2 1
Gini Impurity for Loves Popcorn = 0.405
…and Loves Soda.
Soda
True
False
Cool As Ice
Yes No
3 1
Cool As Ice
Yes No
0 3
Gini Impurity for Loves Soda = 0.214
Age < 15
True
False
Cool As Ice
Yes No
0 3
Cool As Ice
Yes No
3 2
Gini Impurity for Age < 15 = 0.343

Gini Impurity for Loves Popcorn = 0.405

And now we have the **Gini Impurity** for **Age.**

Gini Impurity for Loves Soda = 0.214

Gini Impurity for Age < 15 = 0.343

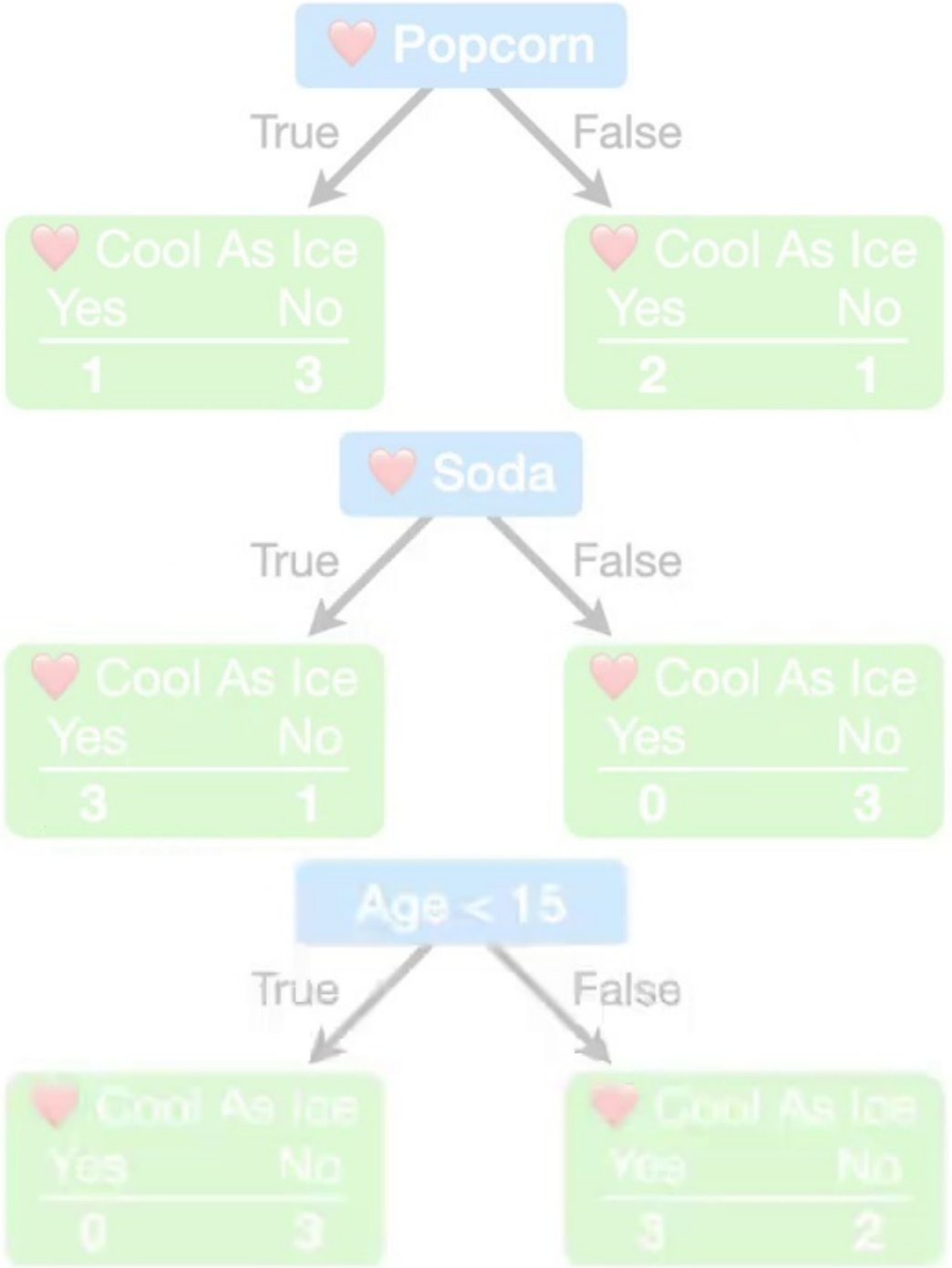


Gini Impurity for Loves Popcorn = 0.405

And because **Loves Soda** has the lowest the **Gini Impurity** overall…

Gini Impurity for Loves Soda = 0.214

Gini Impurity for Age < 15 = 0.343

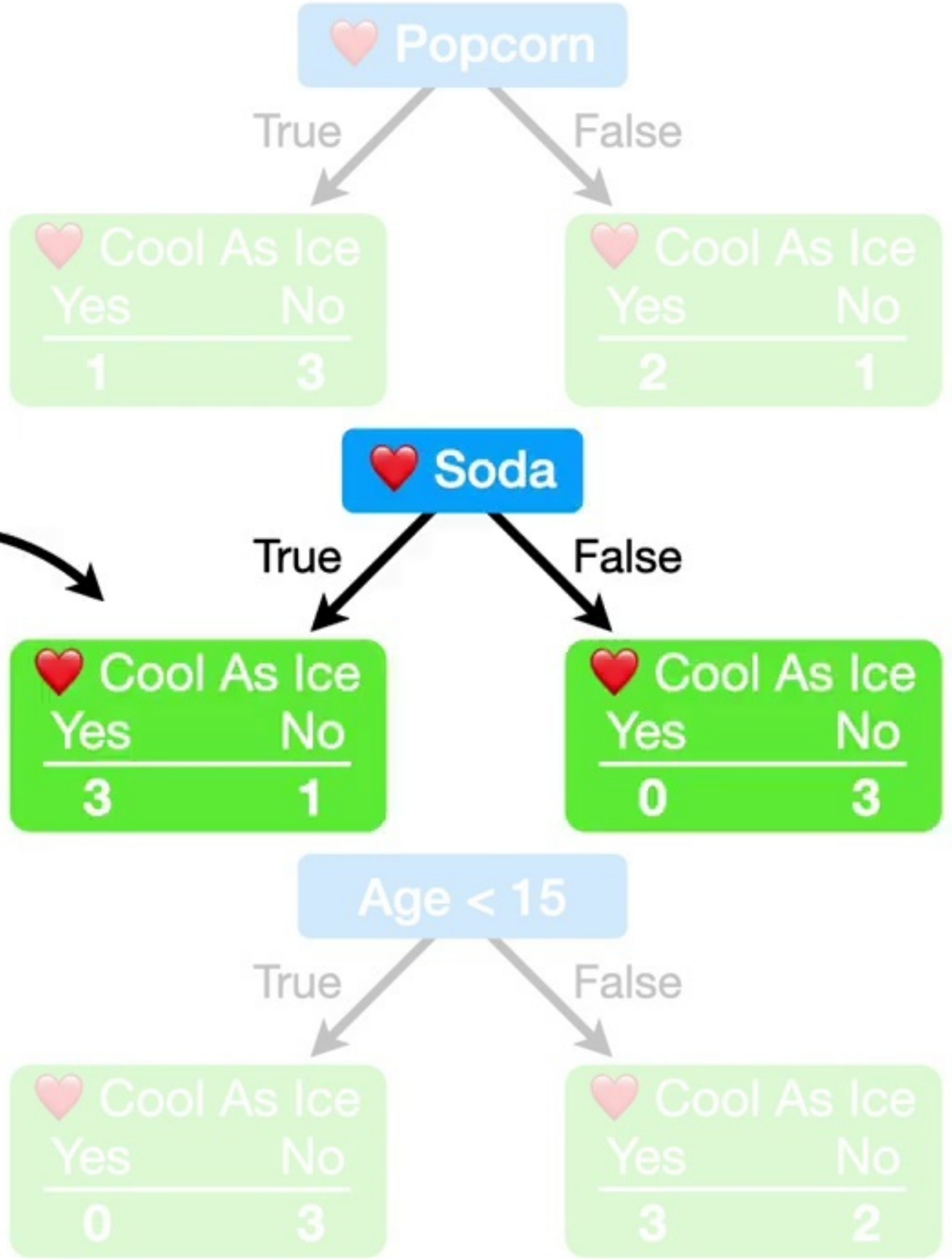


Gini Impurity for Loves Popcorn = 0.405

…we know that its **Leaves** have the lowest **Impurity**…

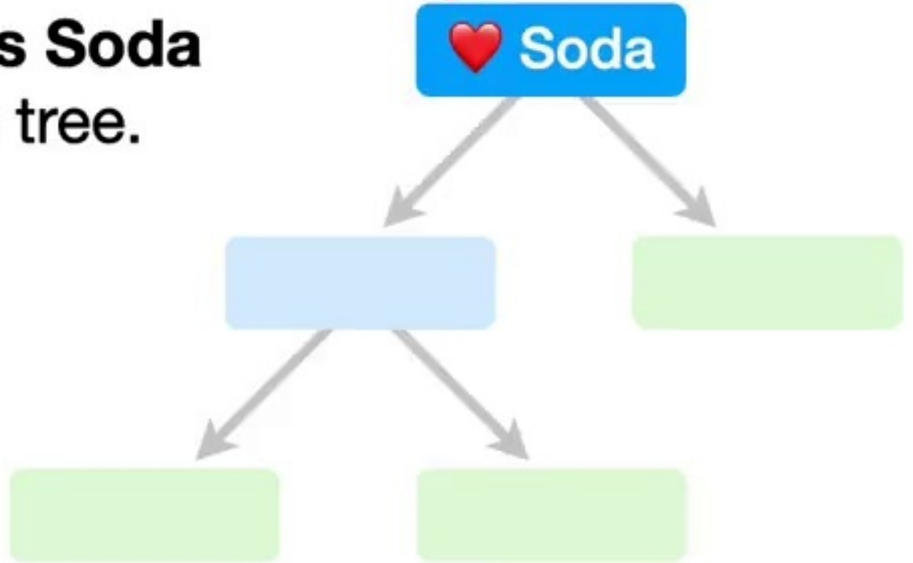
Gini Impurity for Loves Soda = 0.214

Gini Impurity for Age < 15 = 0.343



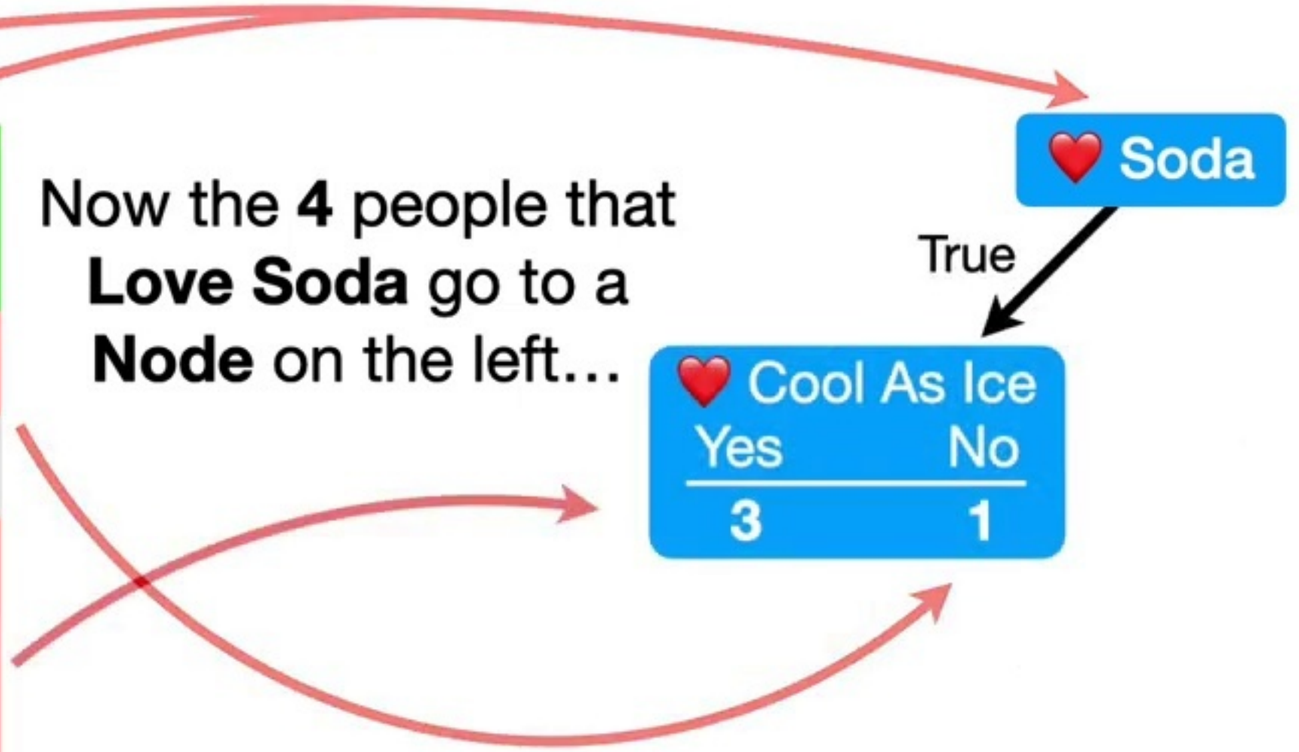
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

…so we put **Loves Soda** at the top of the tree.



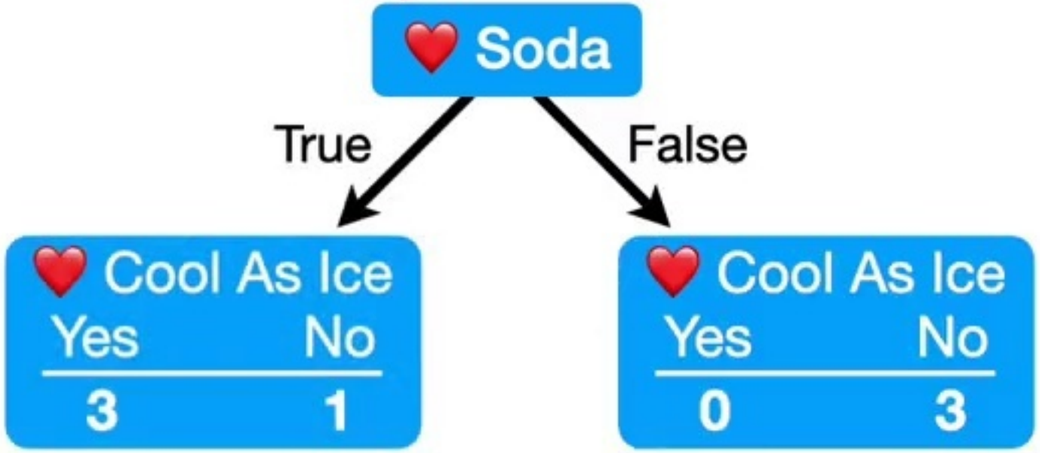
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Now the **4** people that **Love Soda** go to a **Node** on the left…

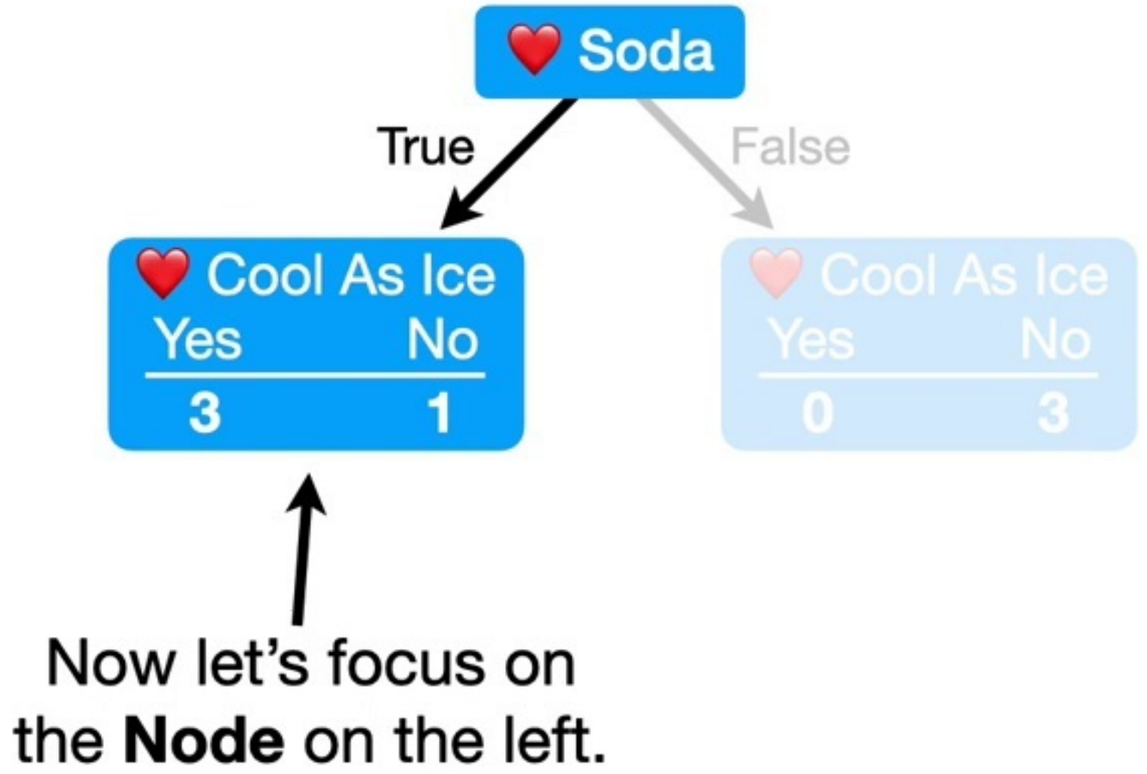


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

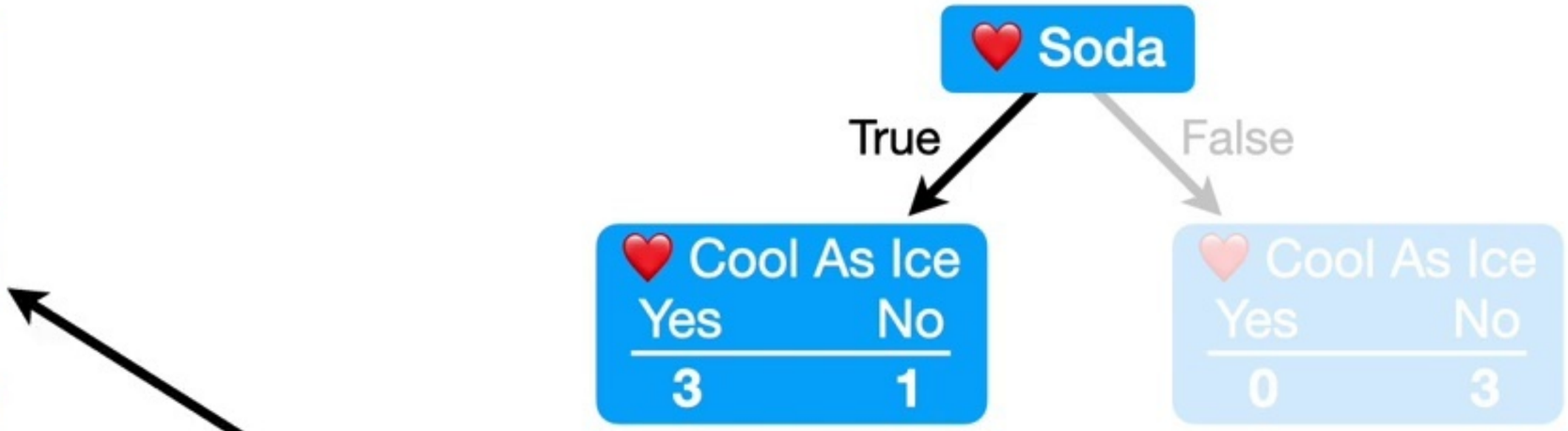
…and the people that *do not* **Love Soda** go to a **Node** on the right.



Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

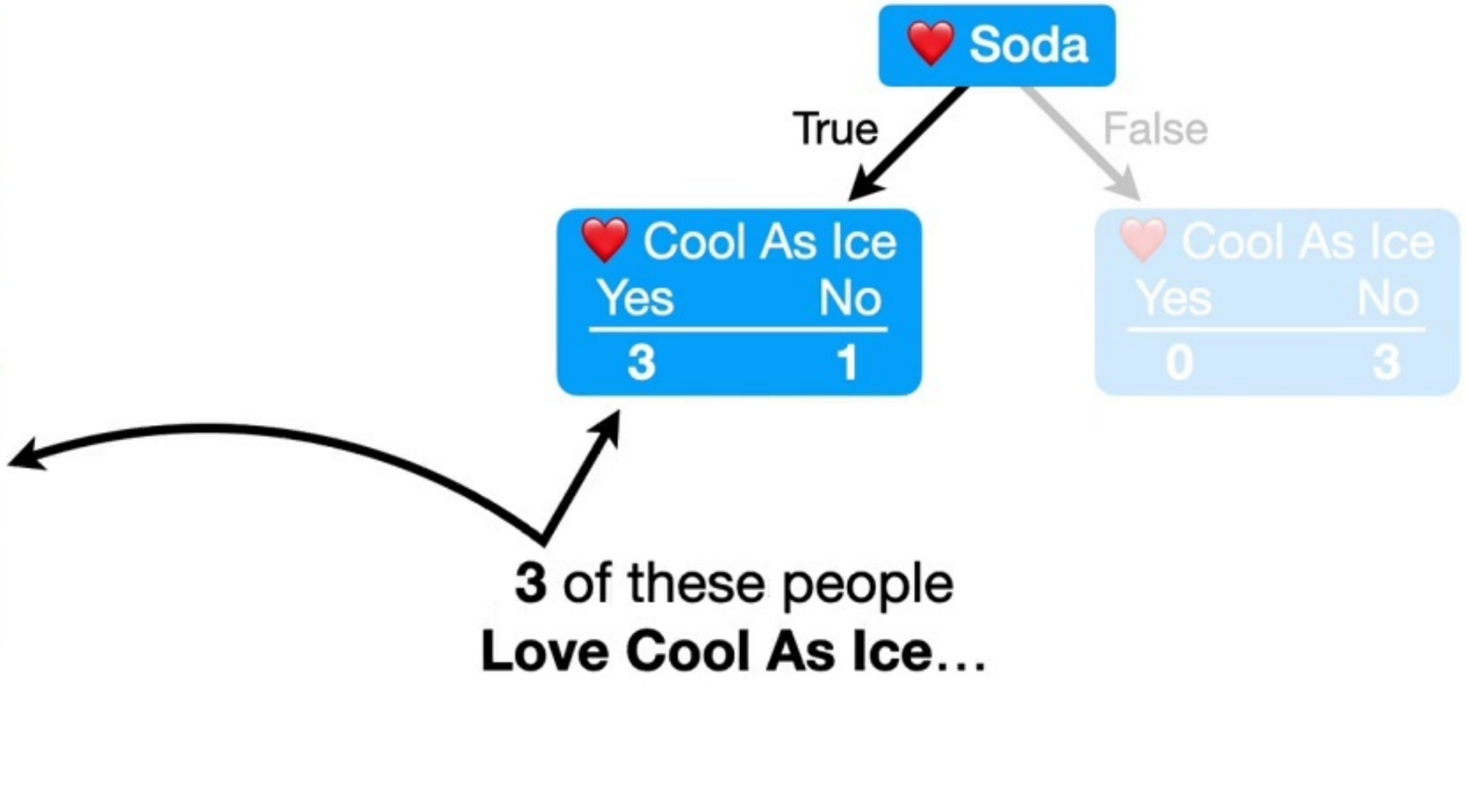


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No

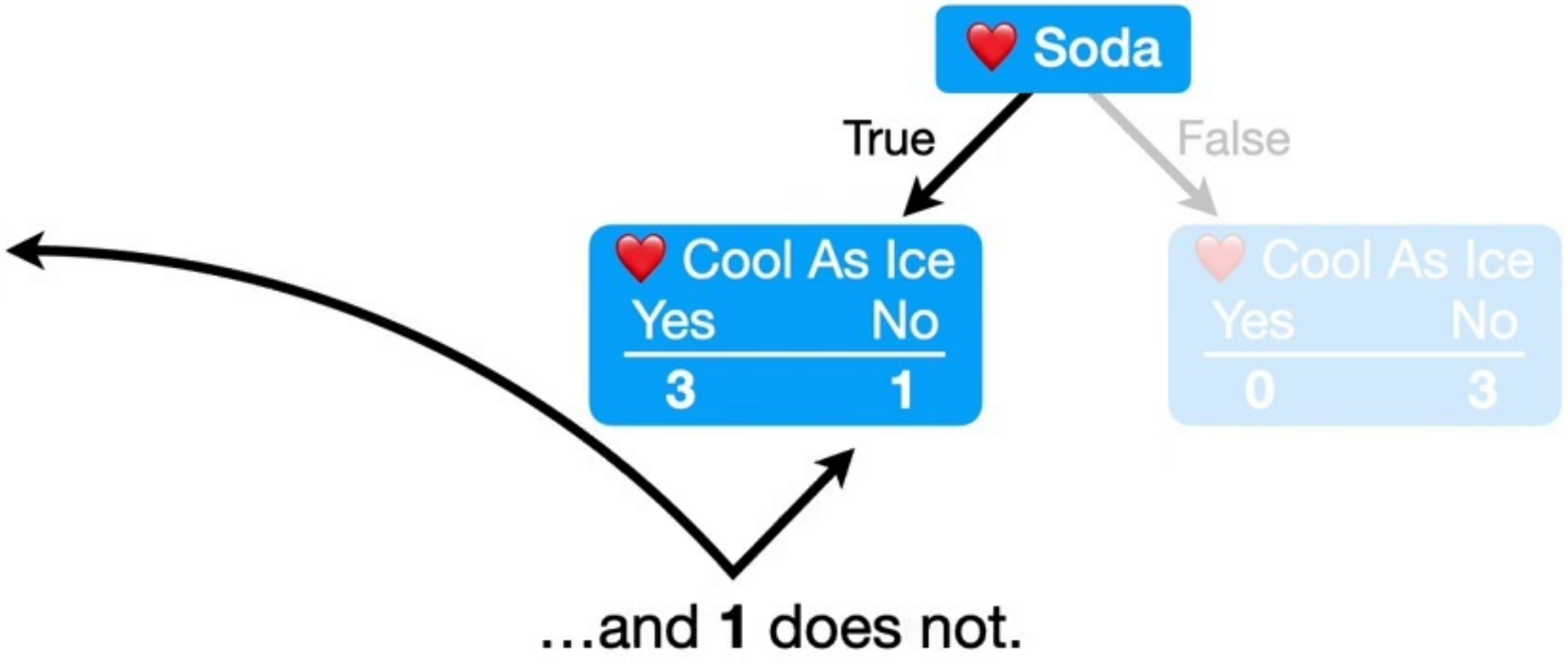


All **4** people that **Love Soda** are in this **Node.**

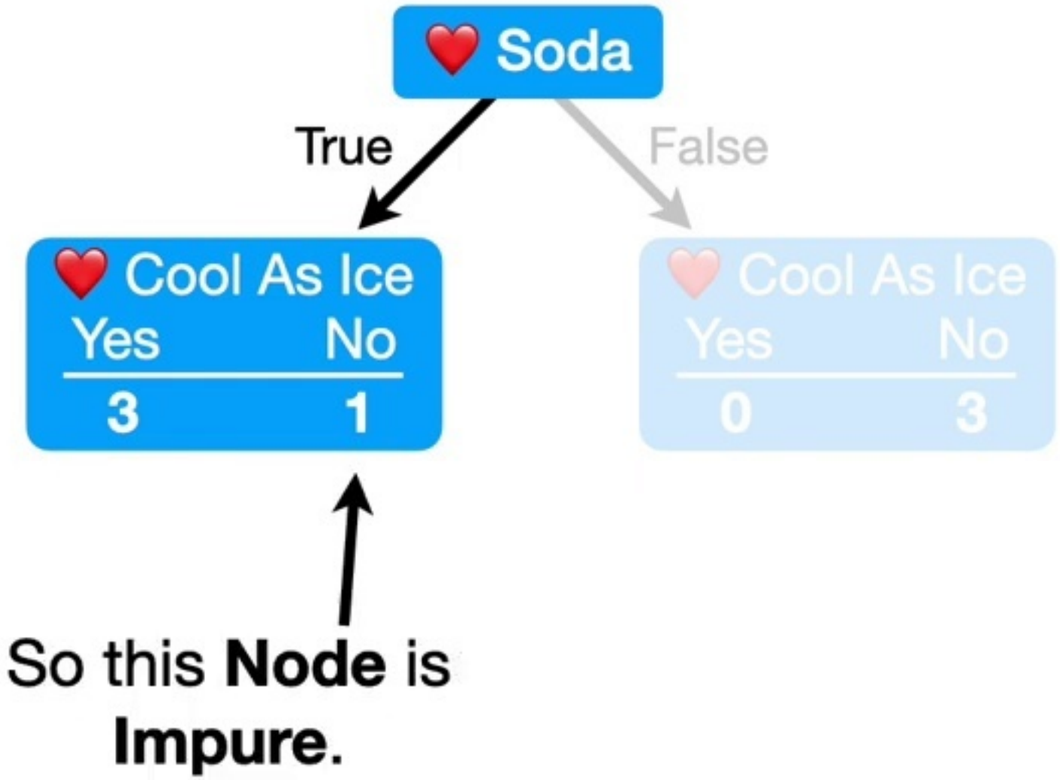
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



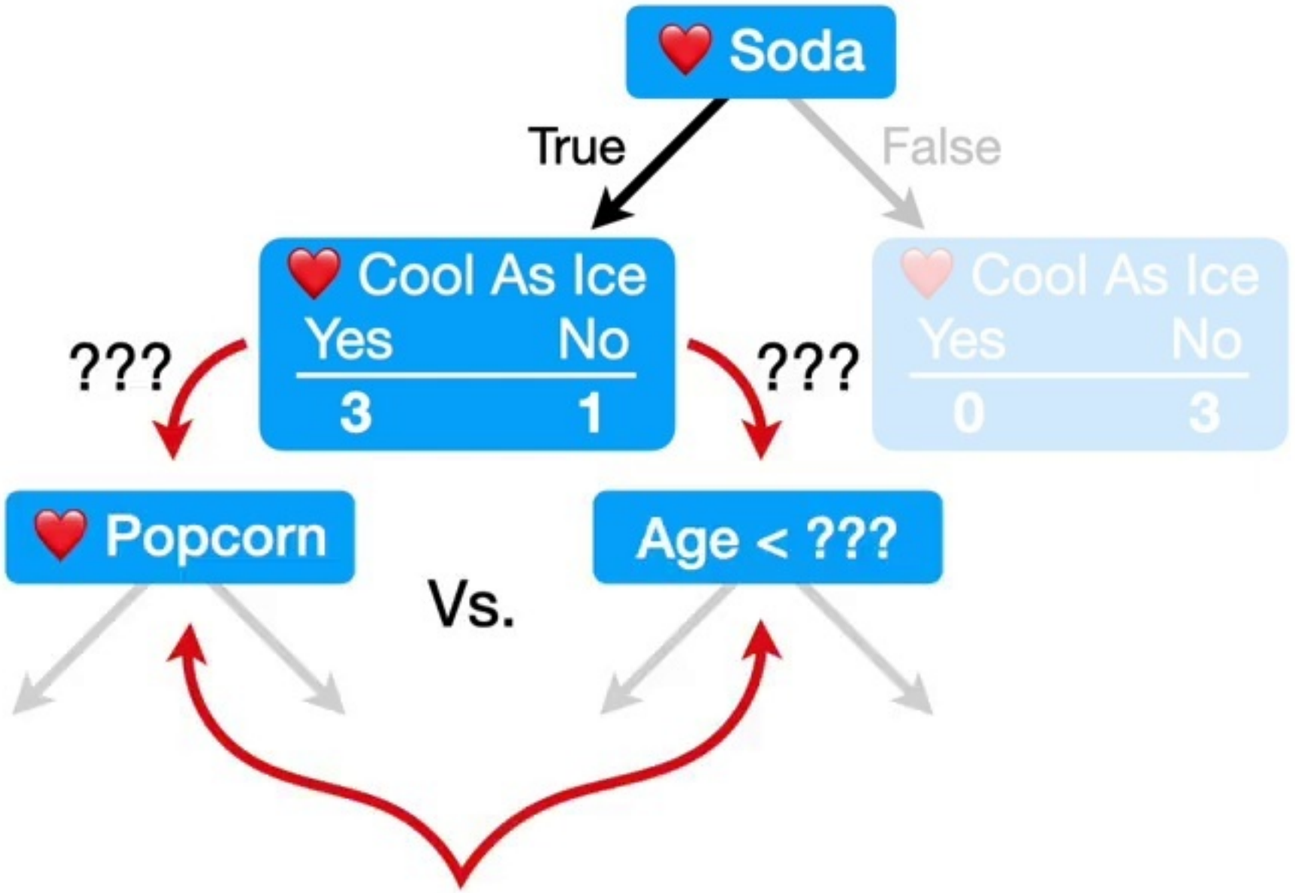
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No

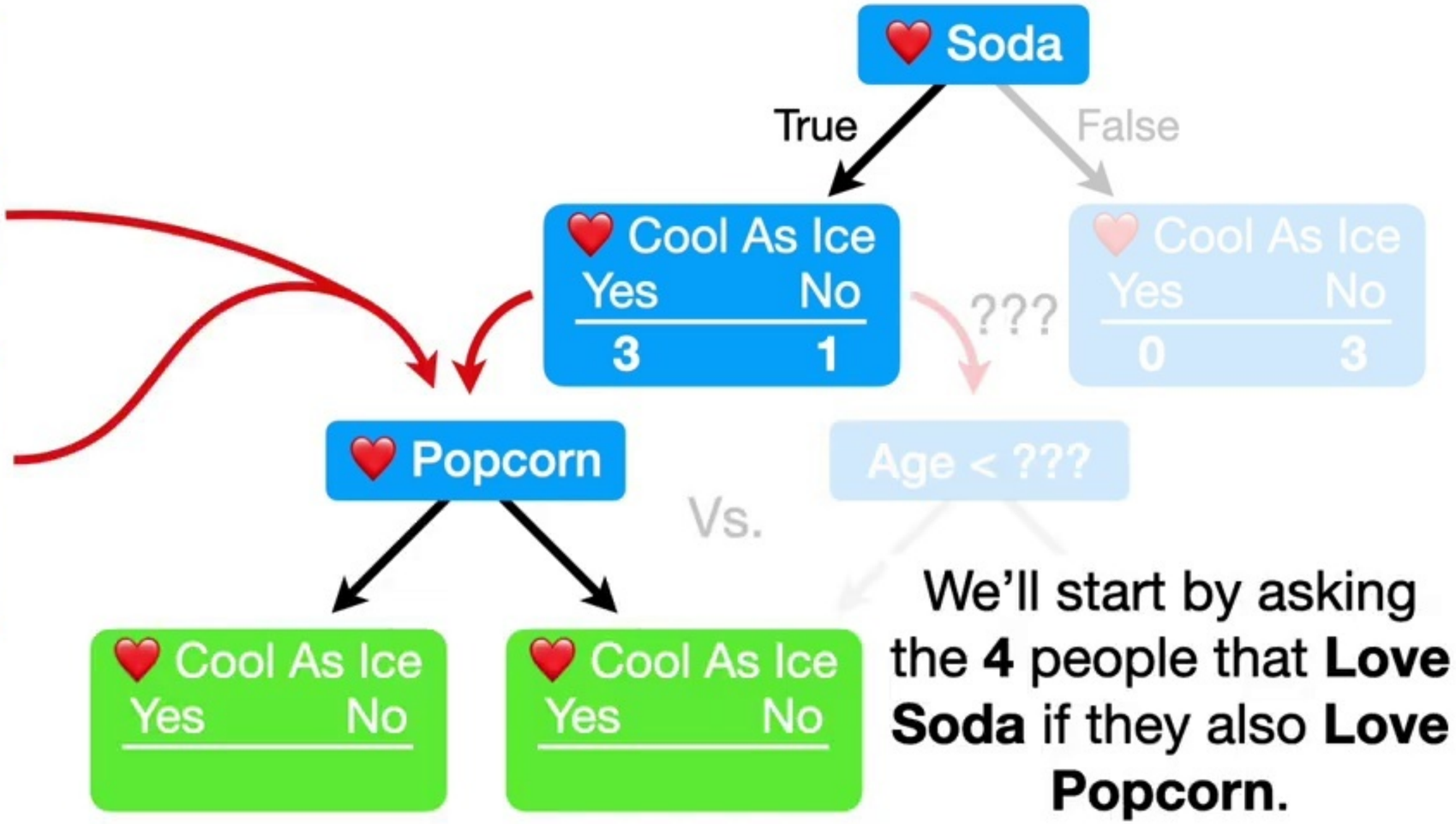


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No

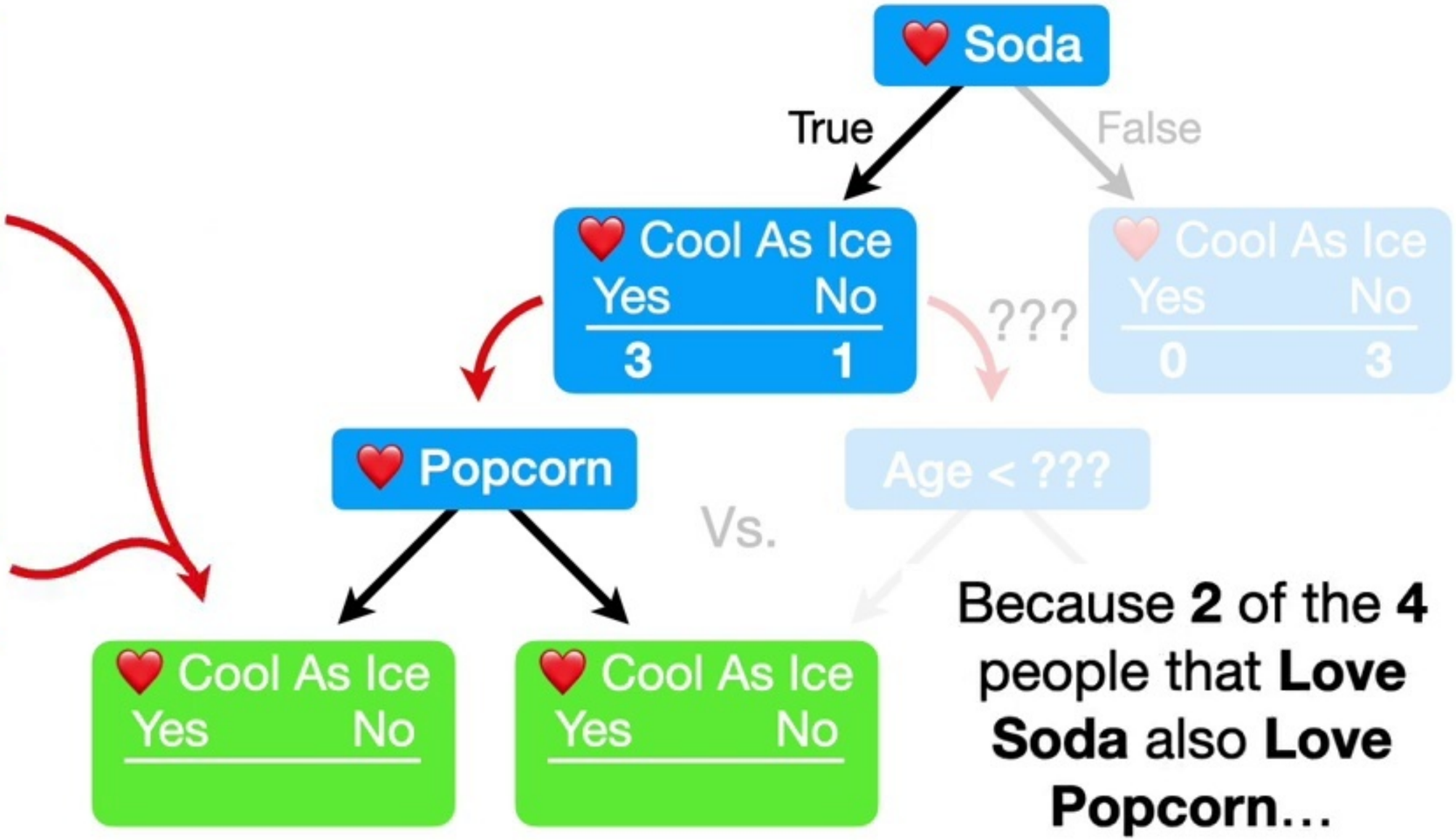


So let's see if we can reduce the **Impurity** by splitting the people that **Love Soda** based on **Loves Popcorn** *or* **Age**.

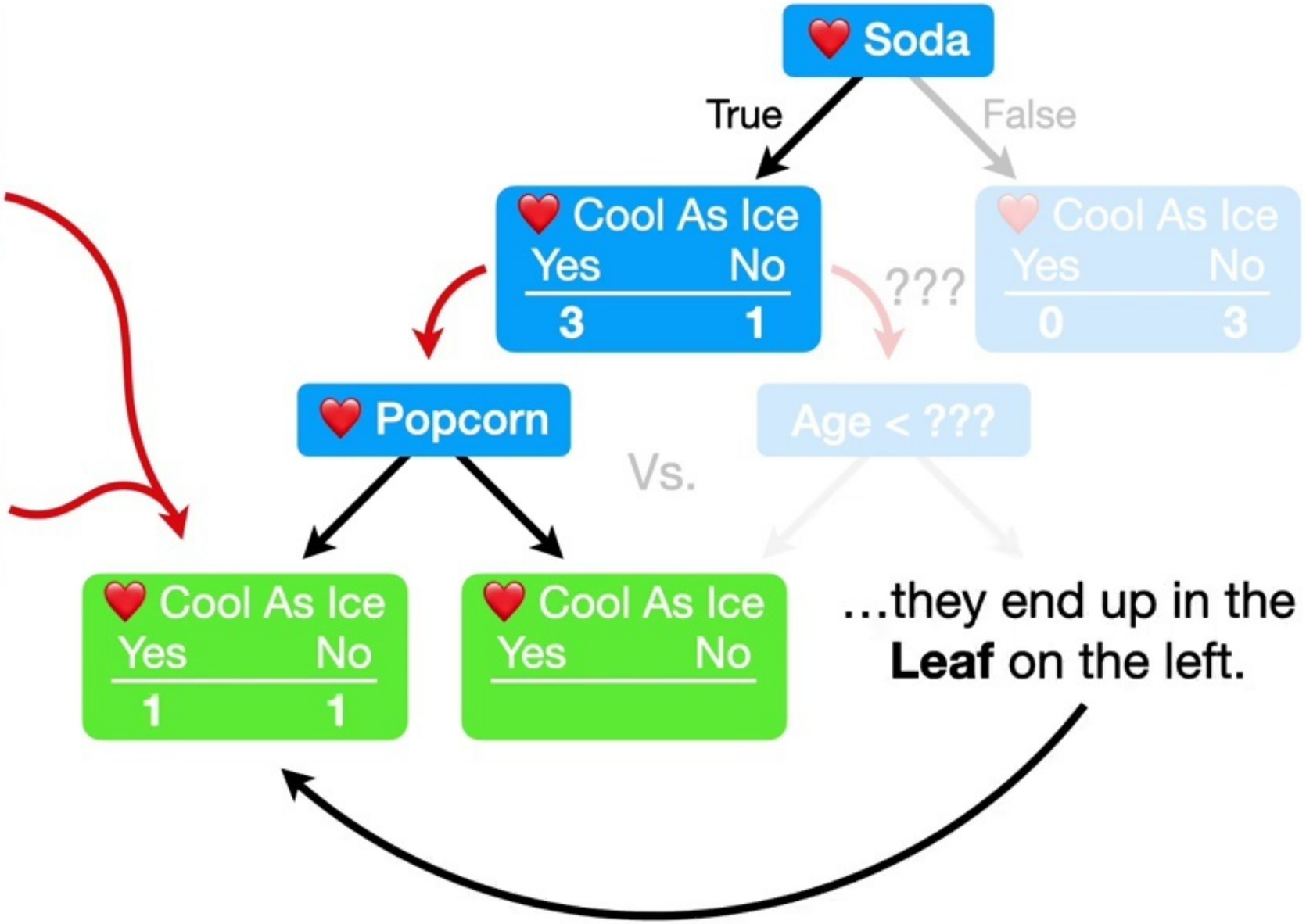
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



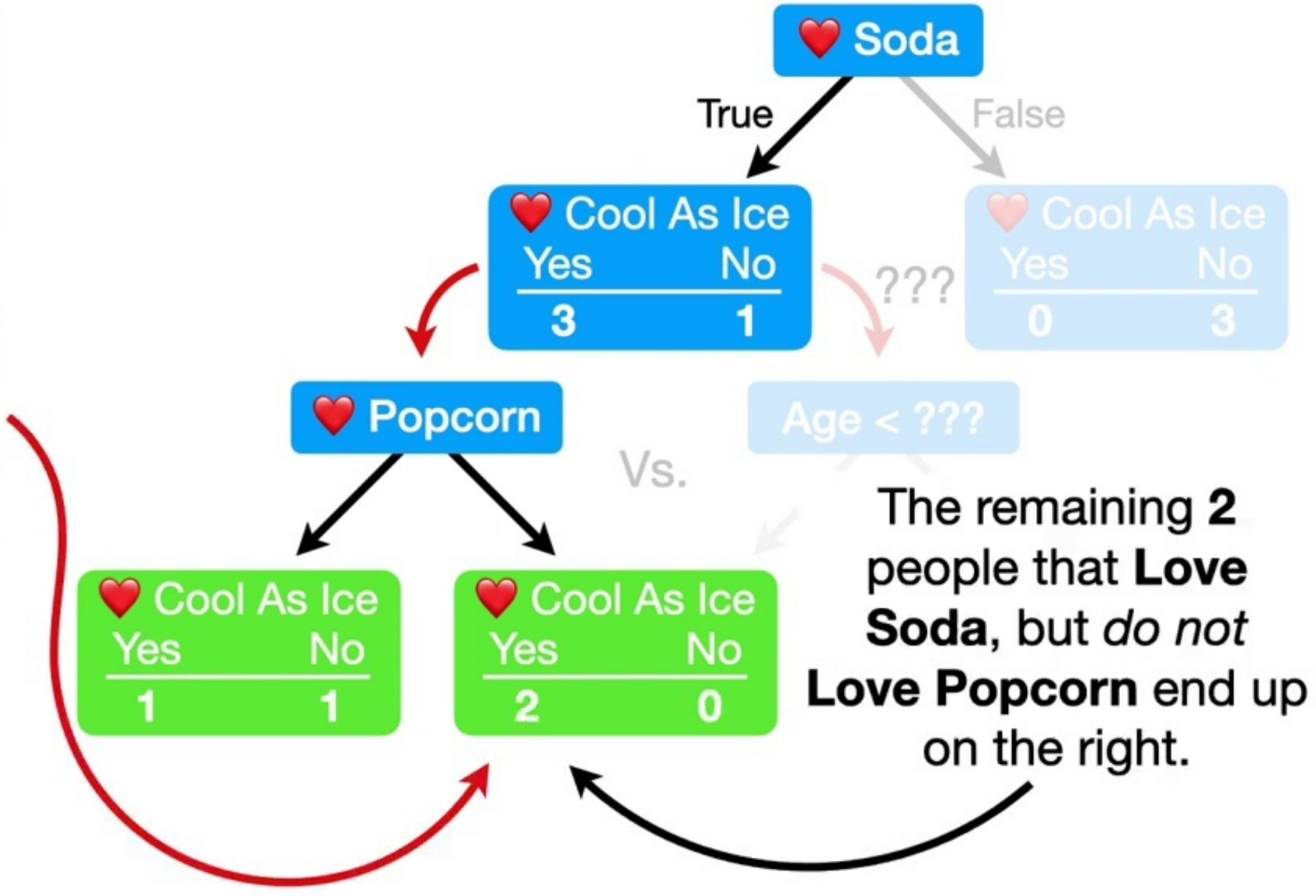
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



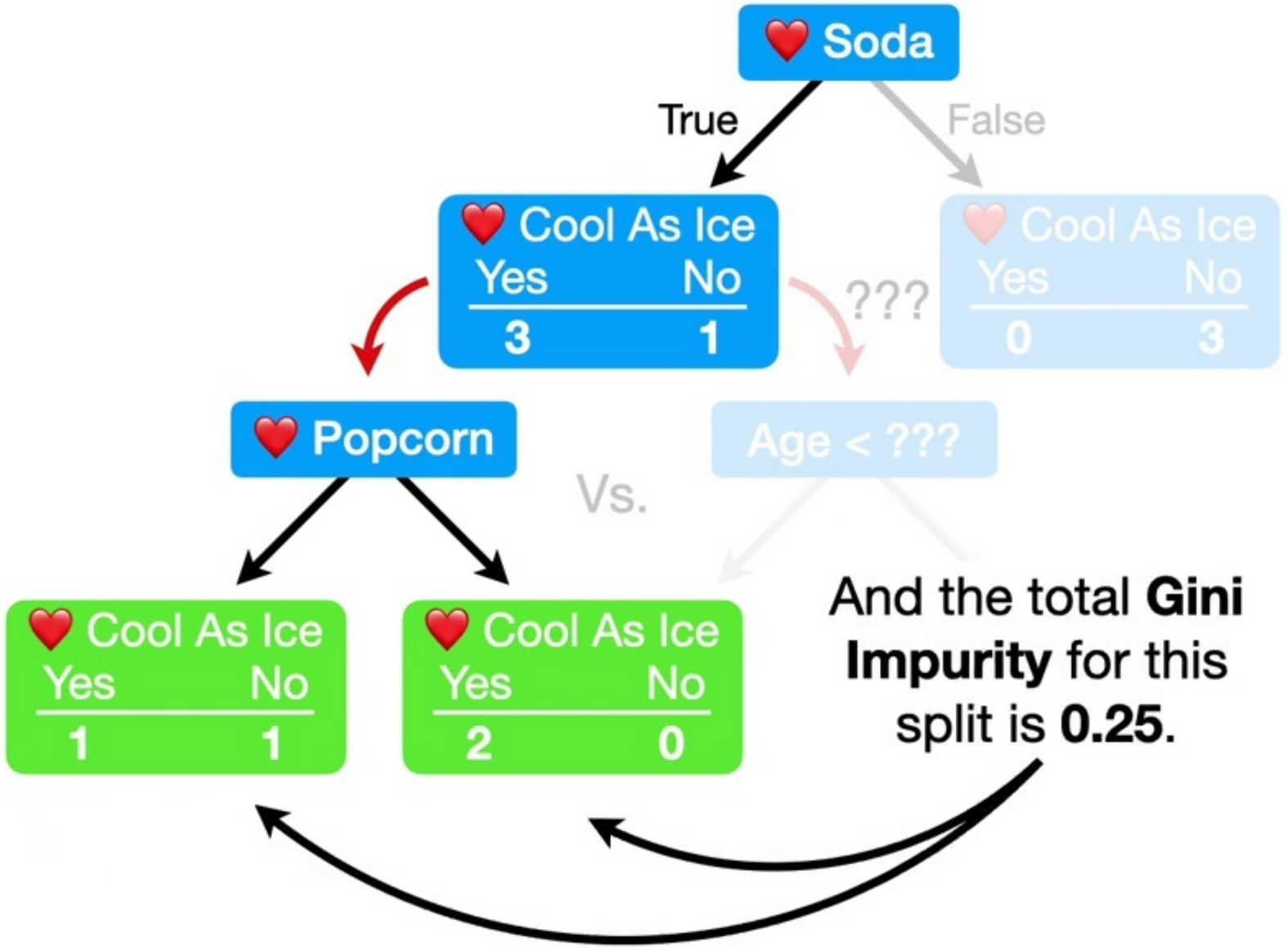
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



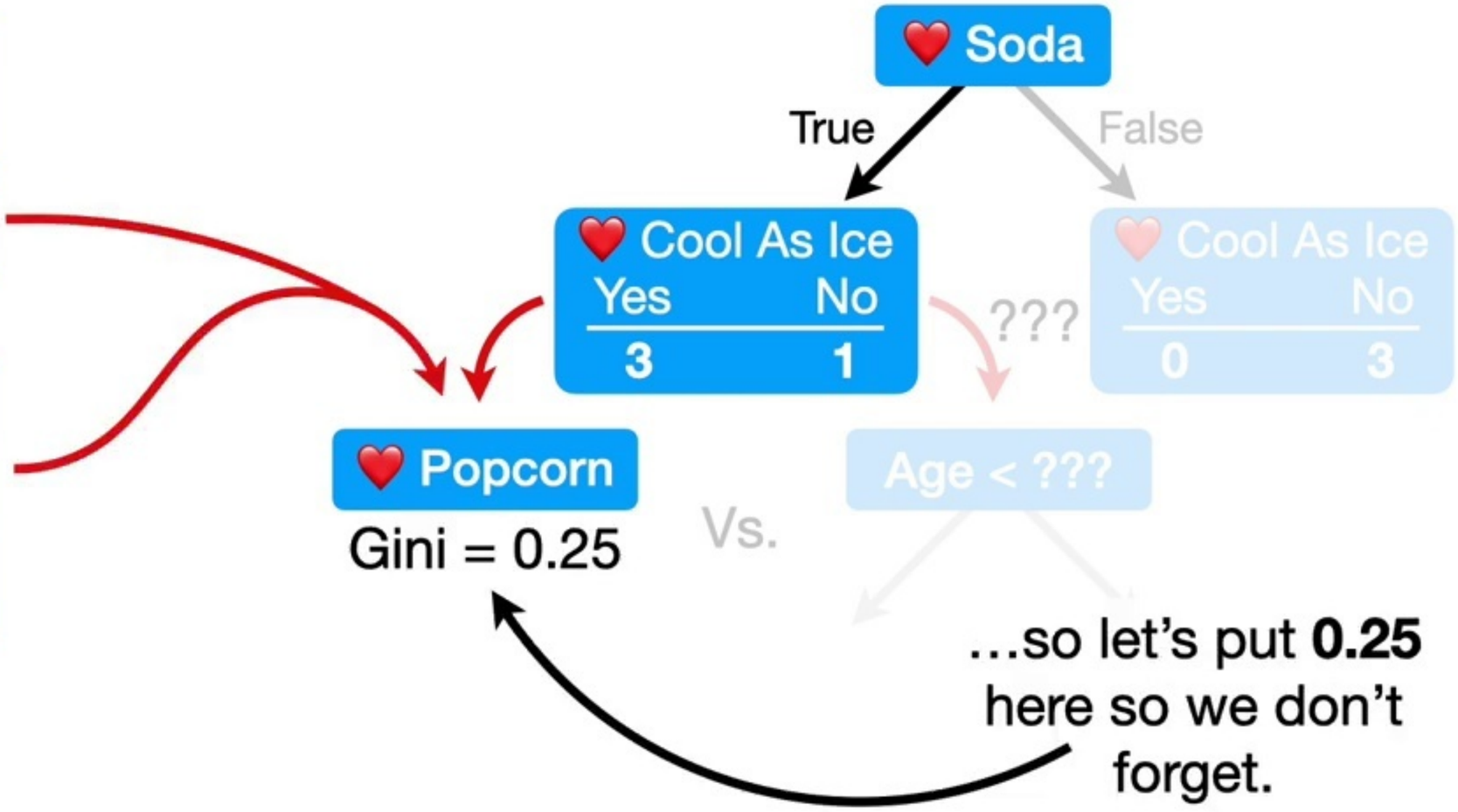
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



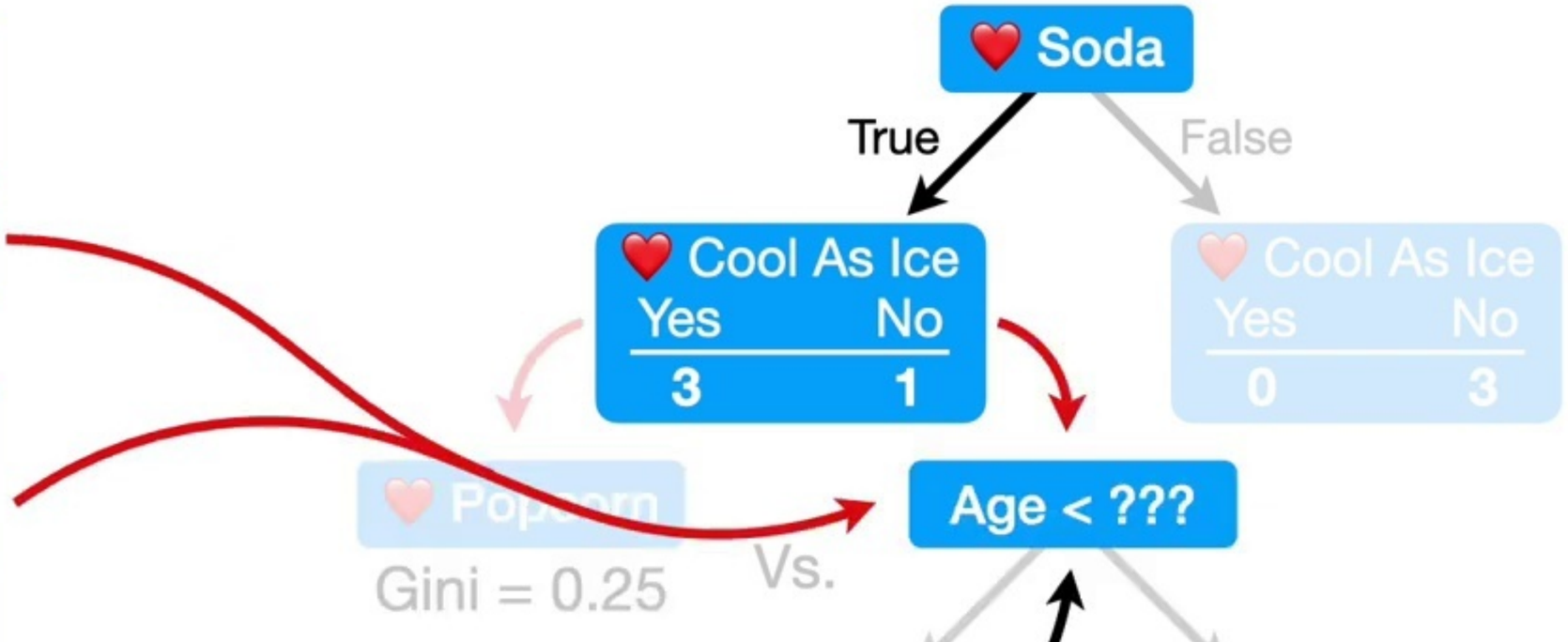
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No

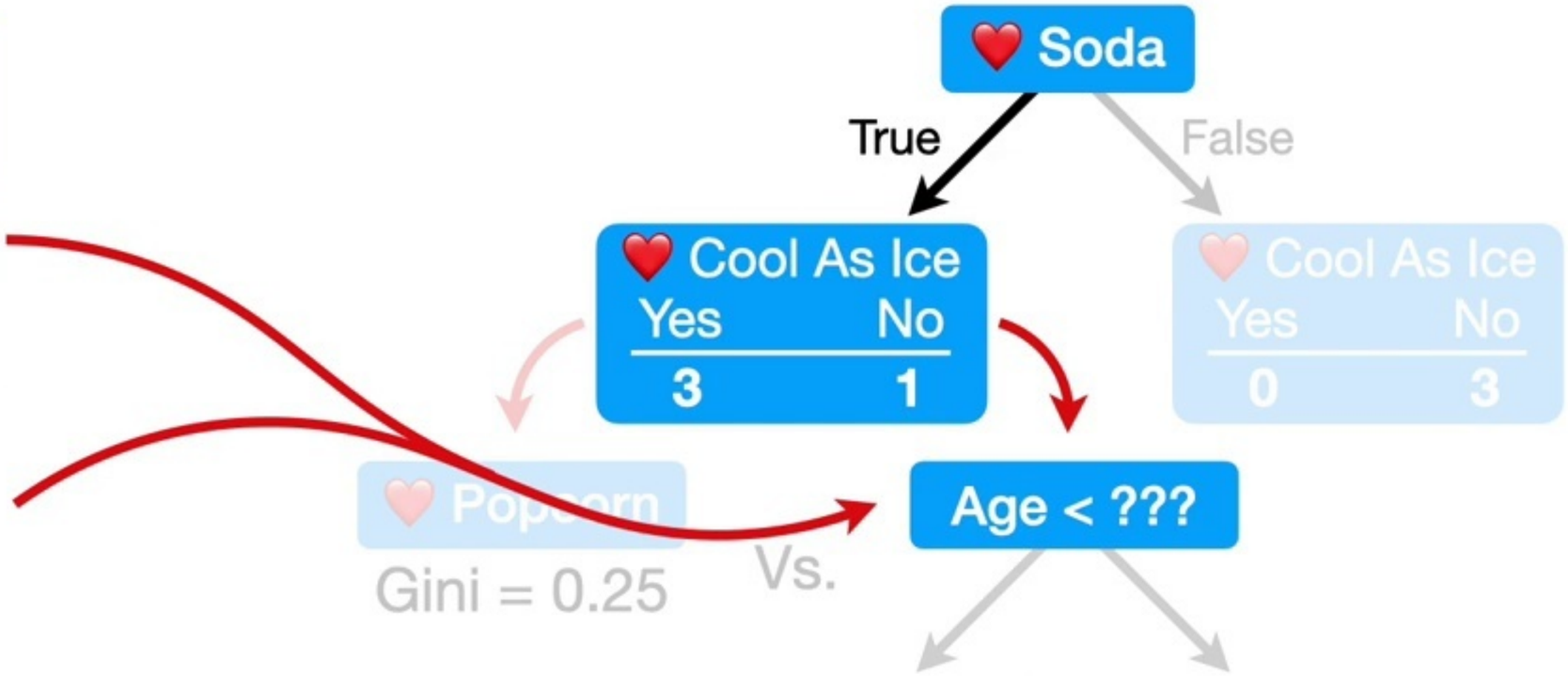


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



Now we test different **Age** thresholds, just like before…

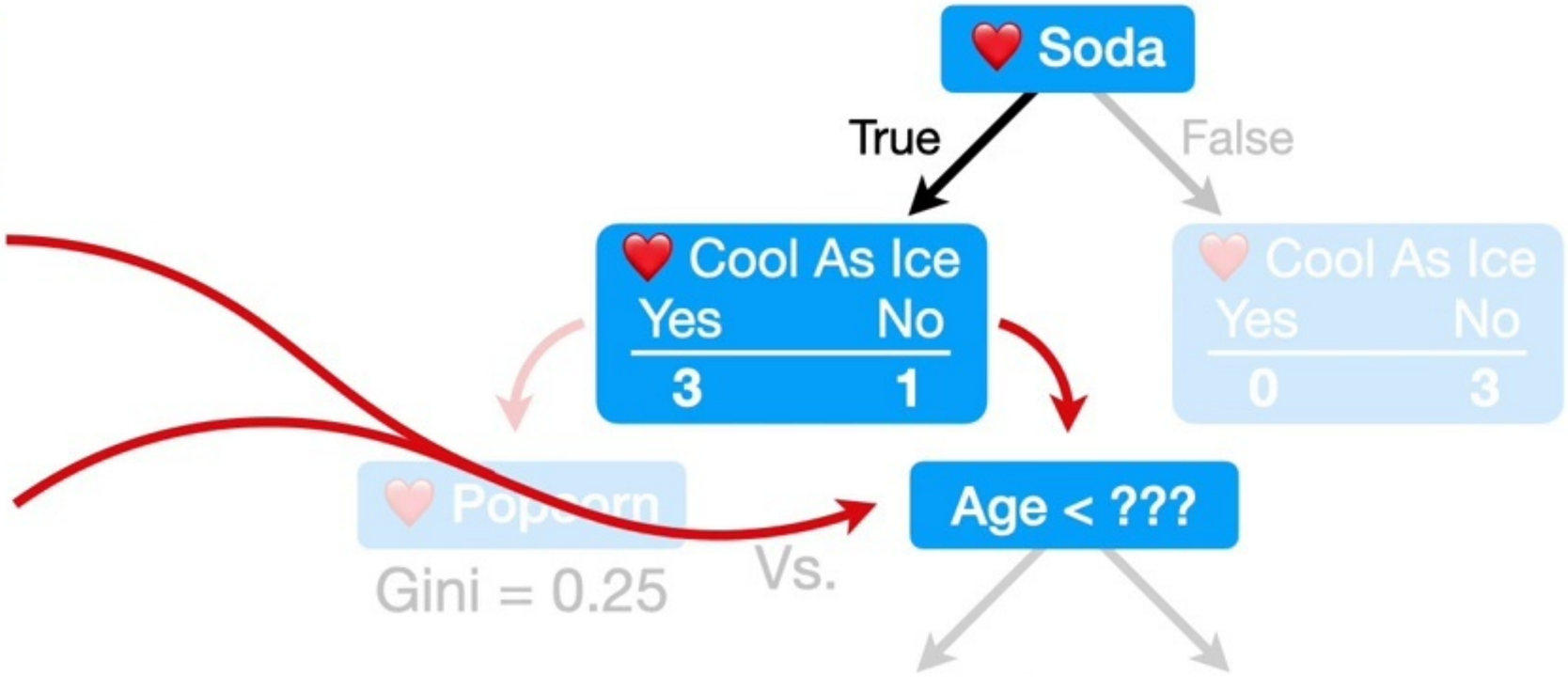
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



…only this time we only consider the **Ages** of people who **Love Soda**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No

12.5

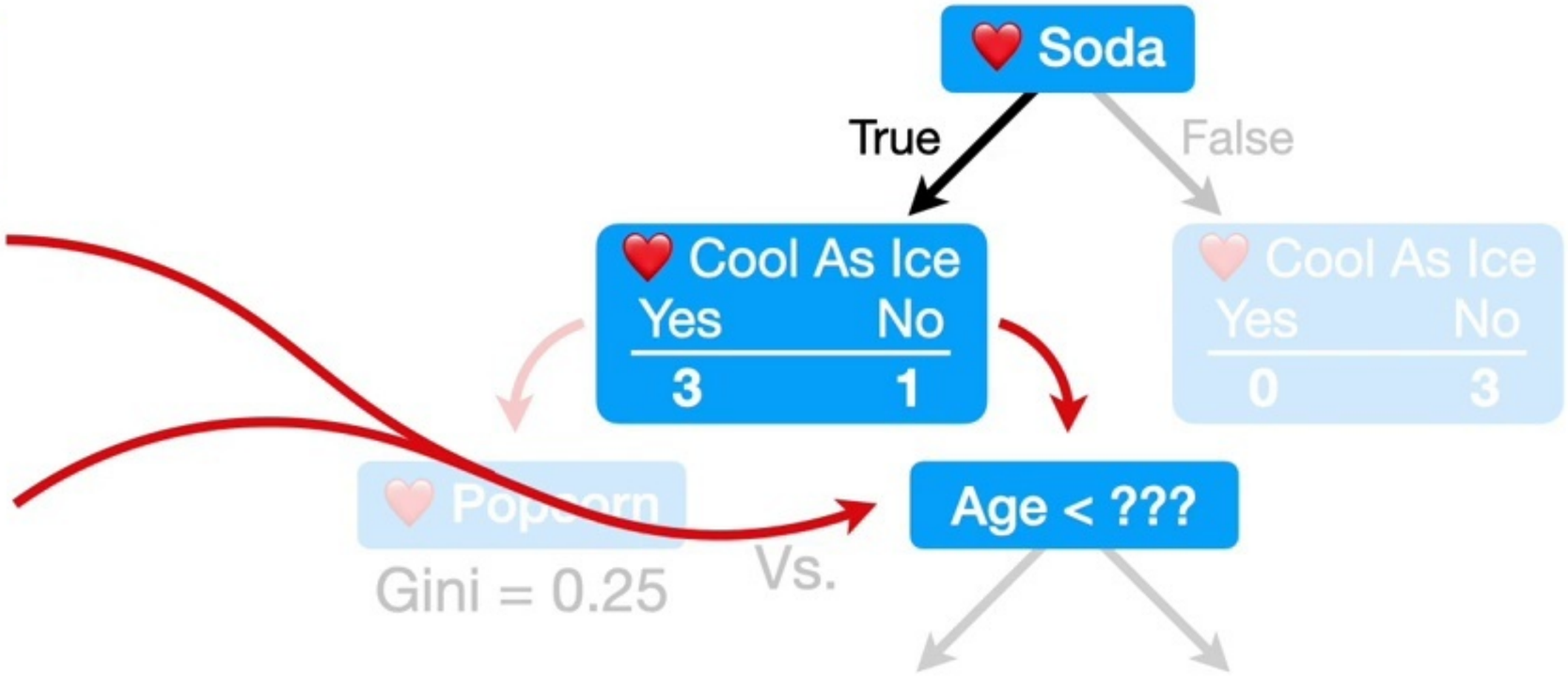


…only this time we only consider the **Ages** of people who **Love Soda**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

12.5

26.5



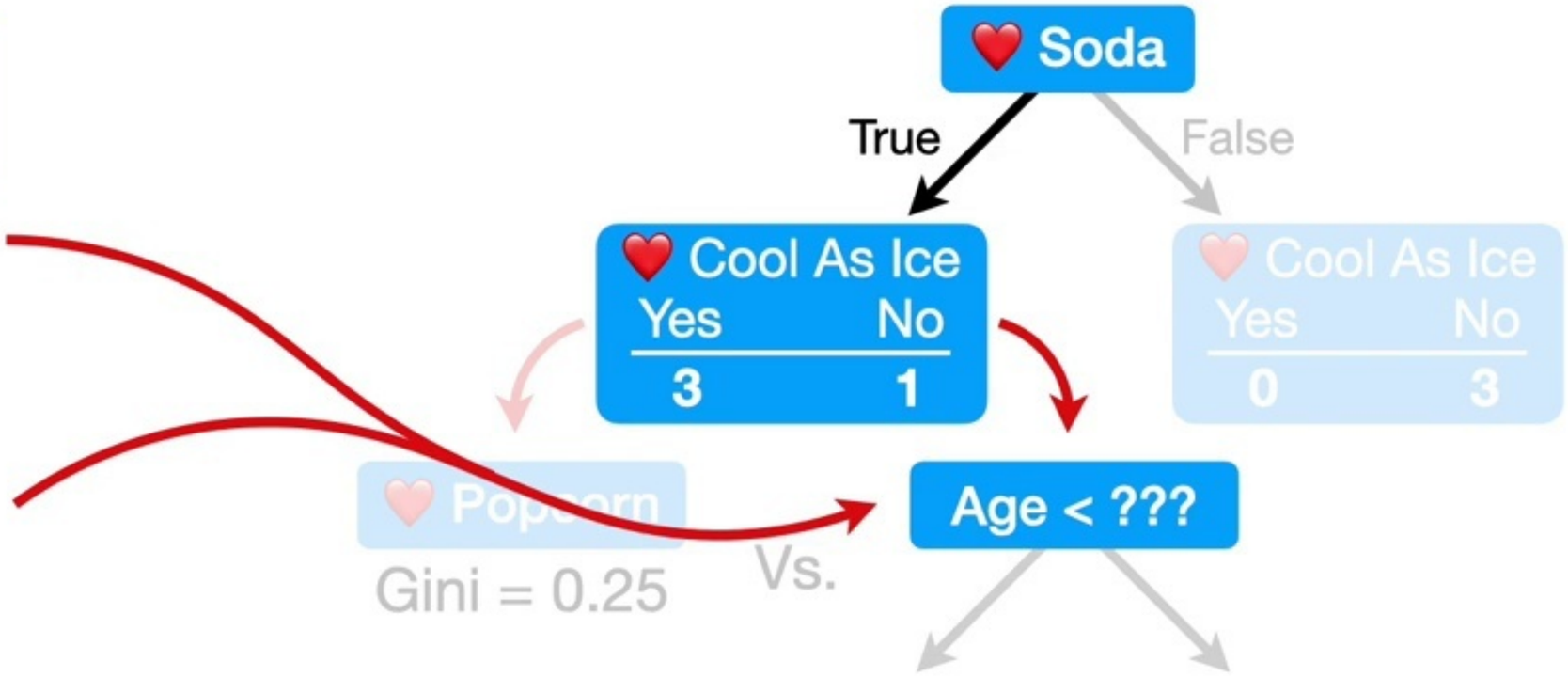
…only this time we only consider the **Ages** of people who **Love Soda**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

12.5

26.5

36.5



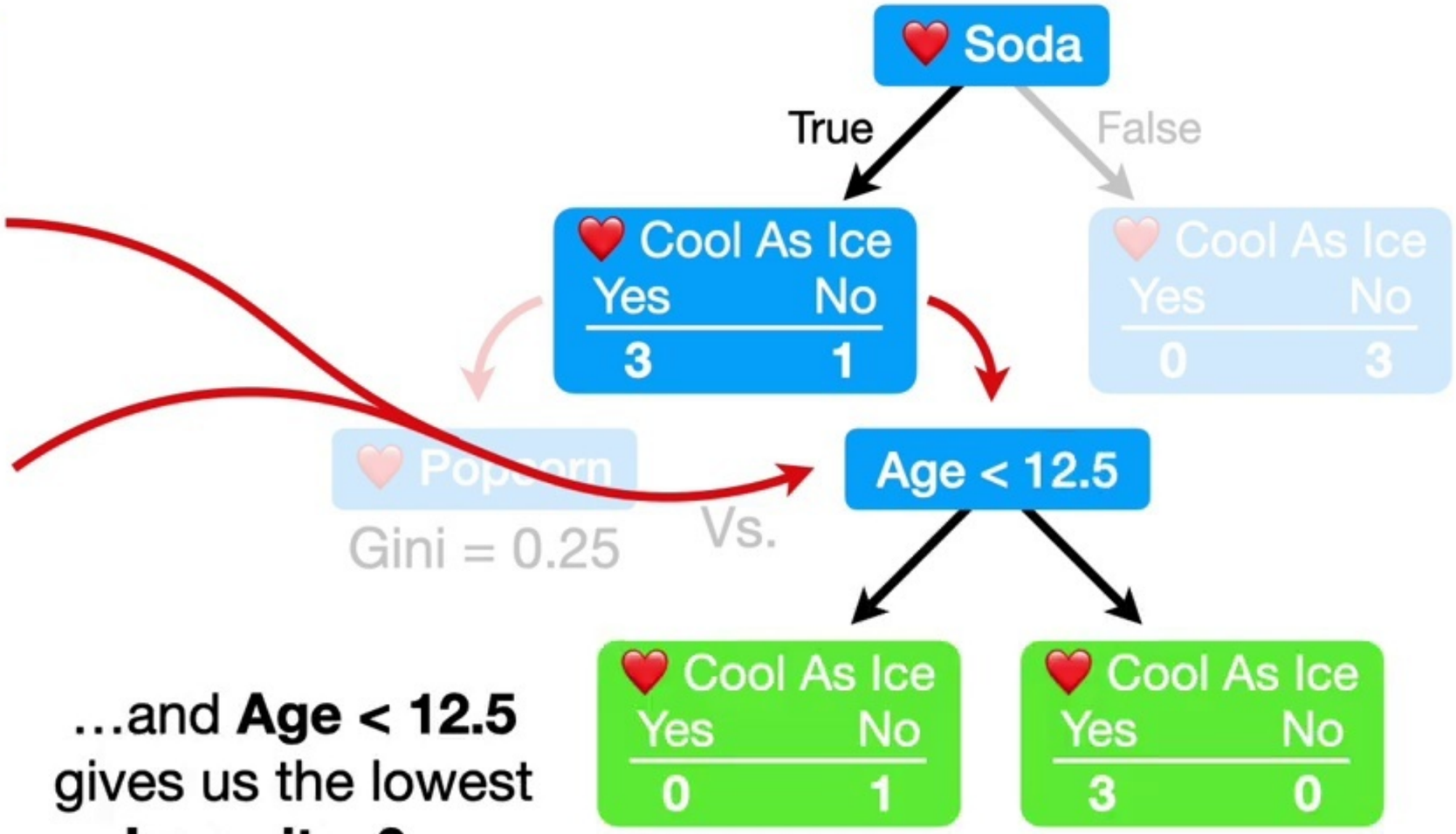
…only this time we only consider the **Ages** of people who **Love Soda**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

12.5

26.5

36.5



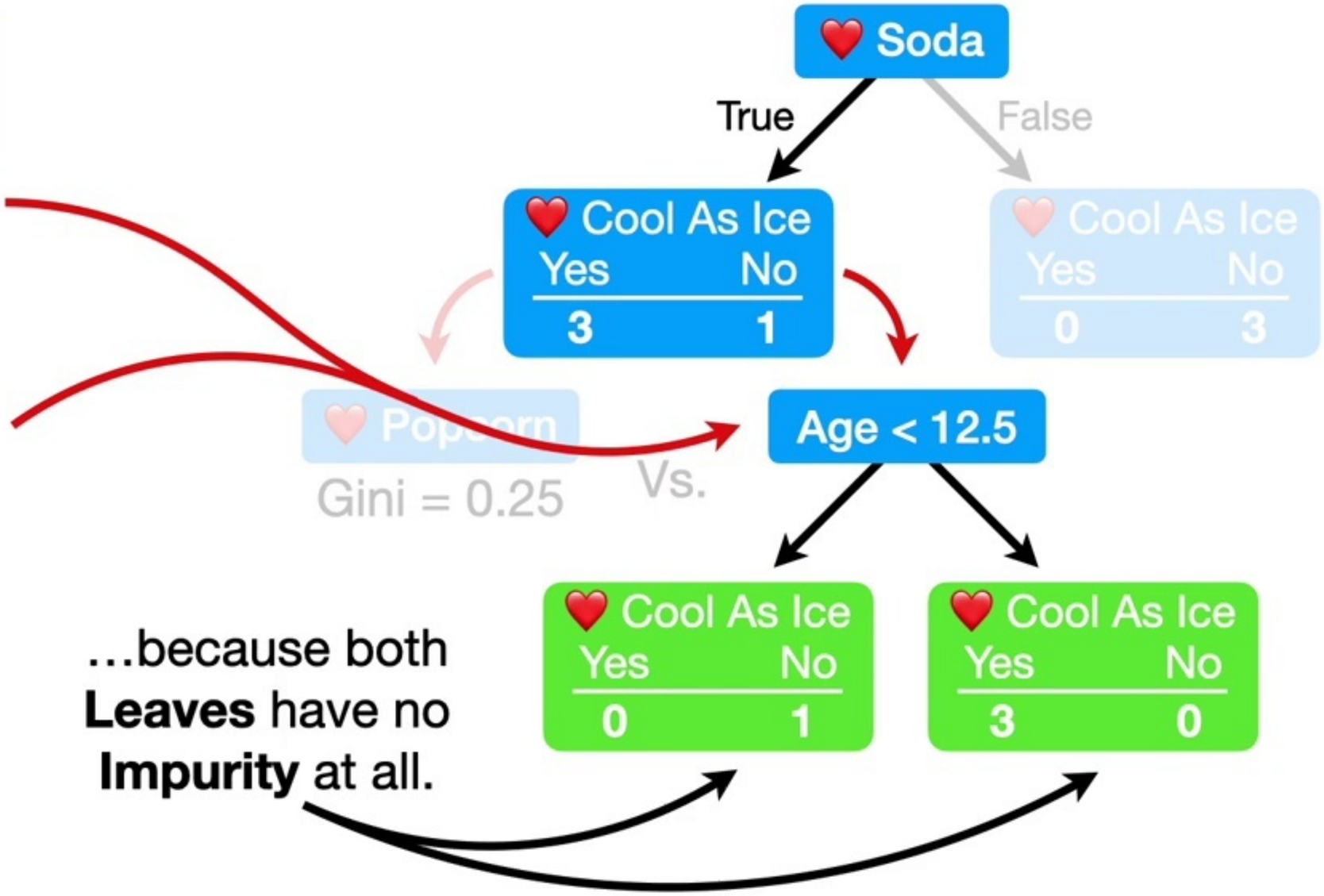
…and **Age < 12.5** gives us the lowest **Impurity**, **0**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

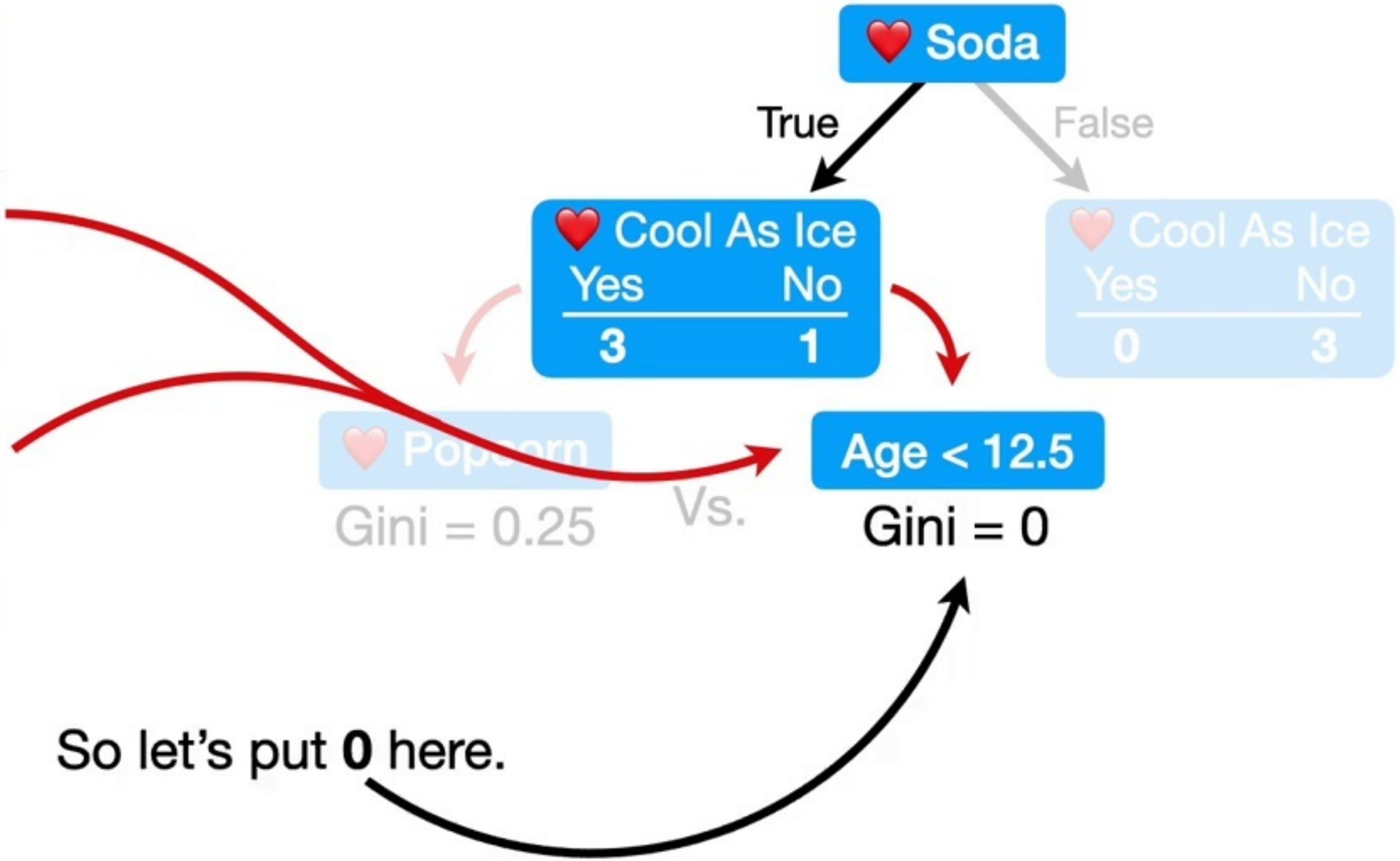
12.5

26.5

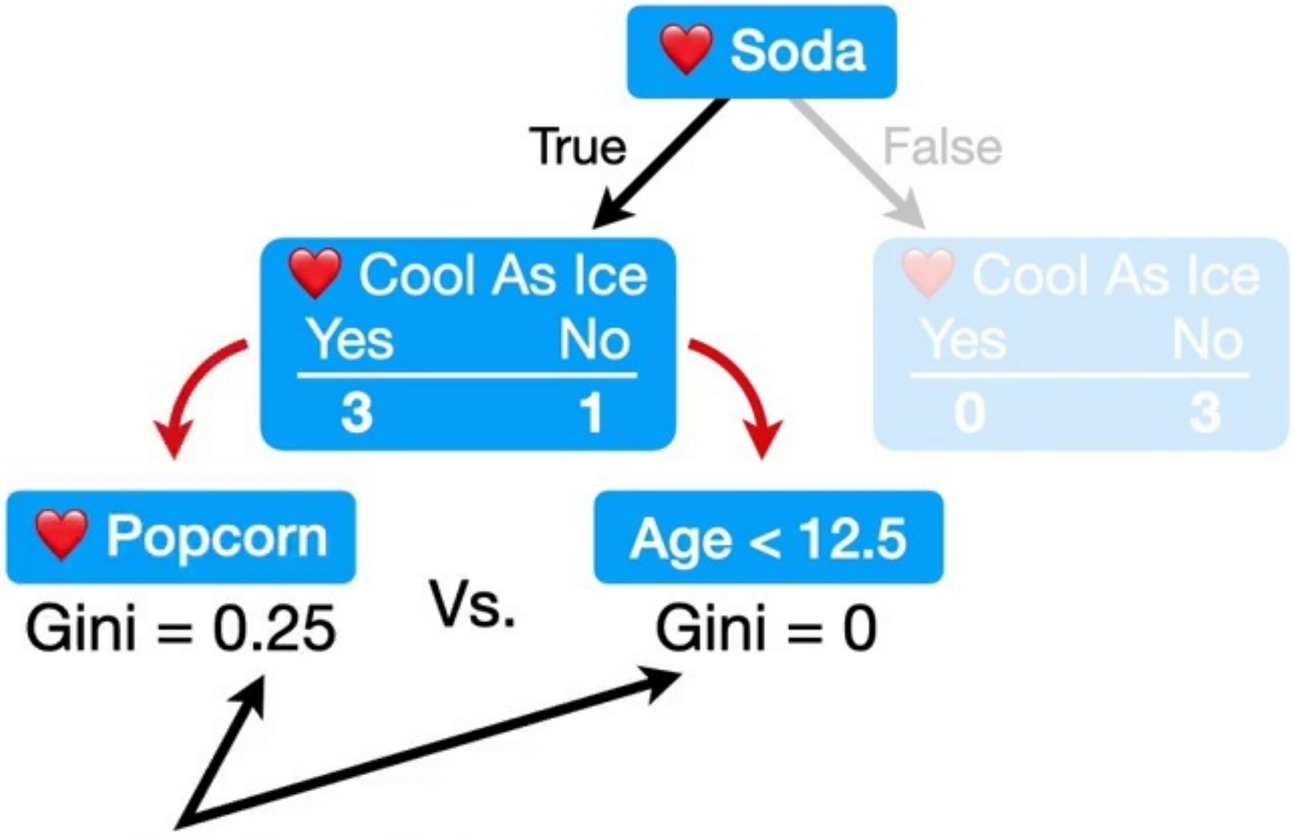
36.5



Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No

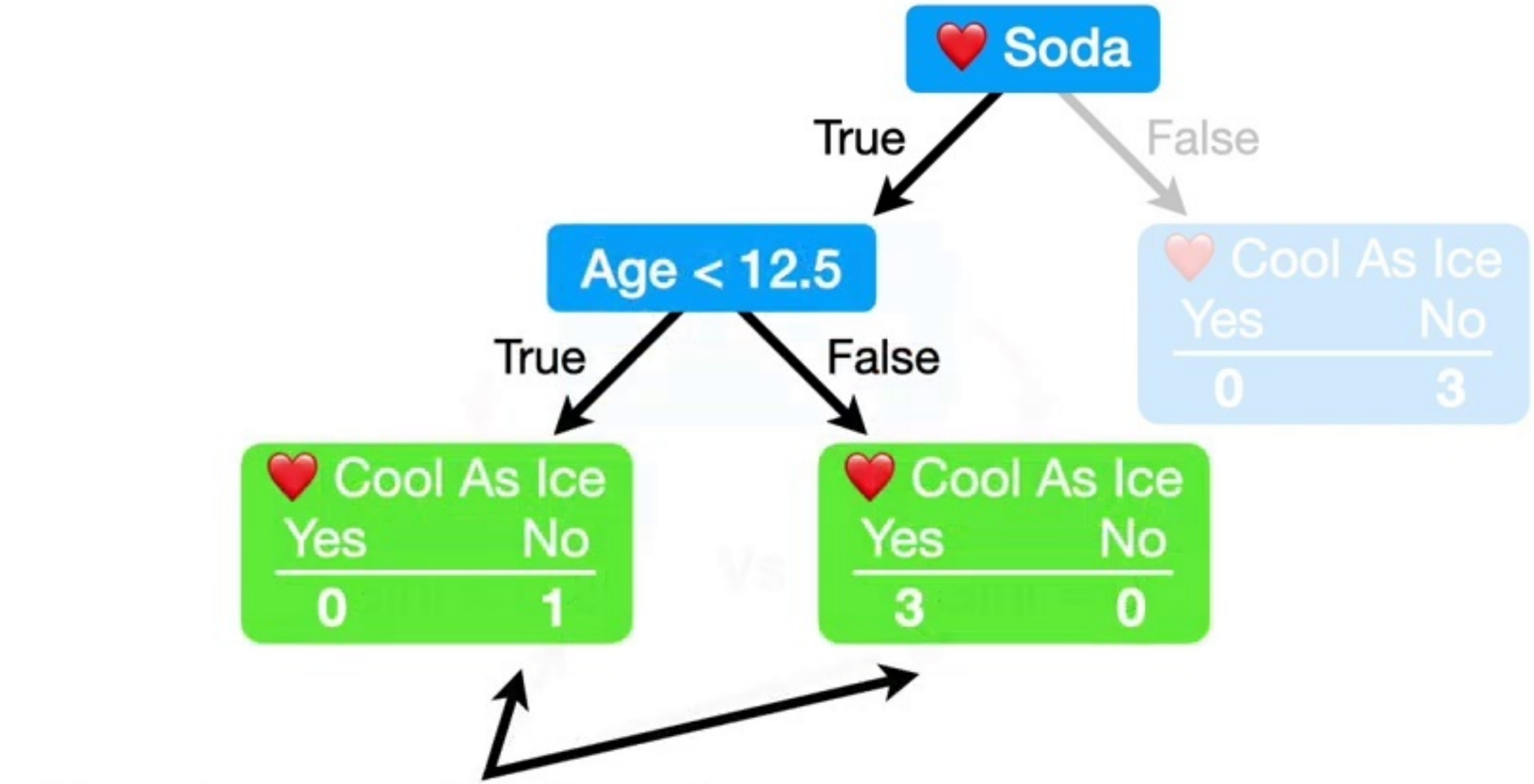


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



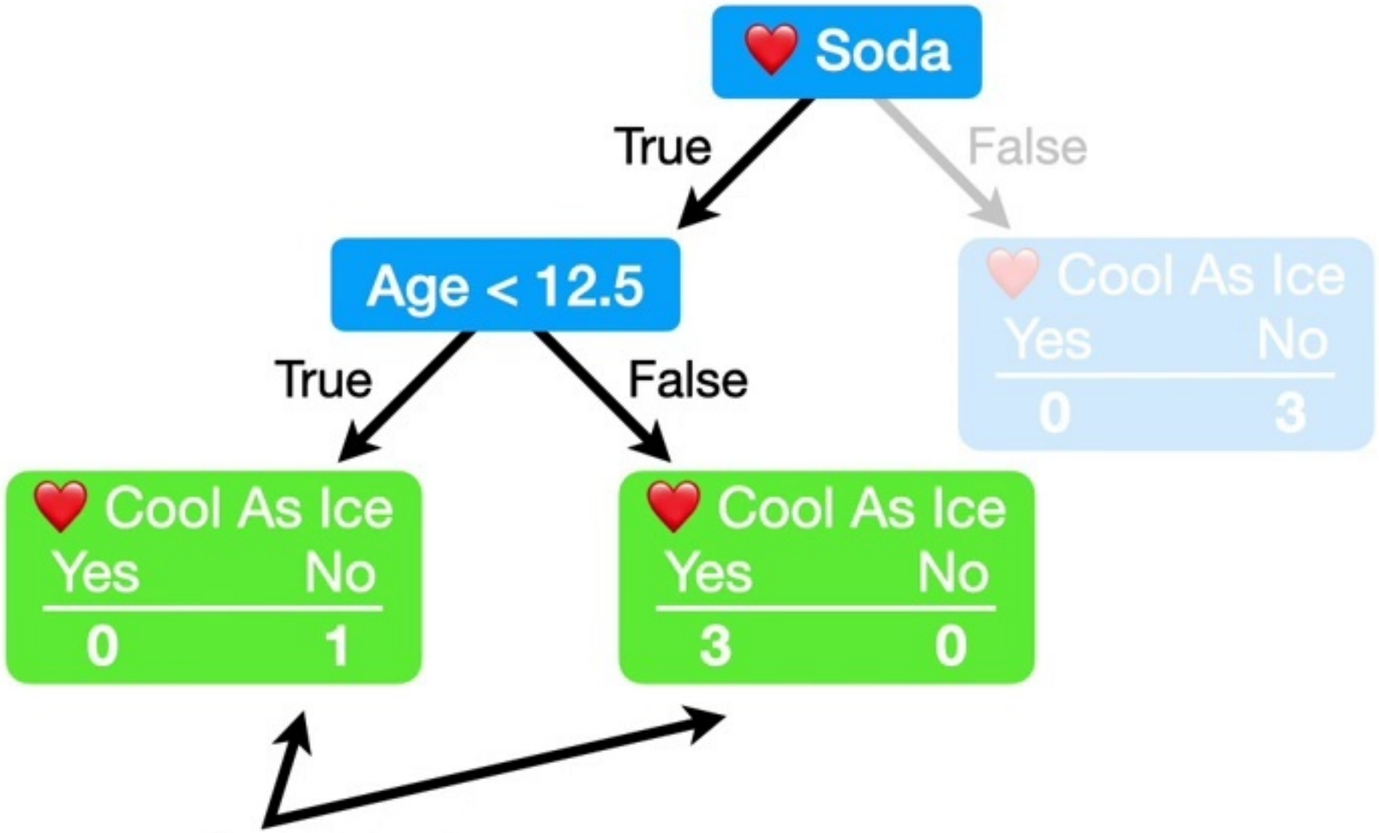
Now, because **0** is less than **0.25**, we will use **Age < 12.5** to split this **Node** into **Leaves**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



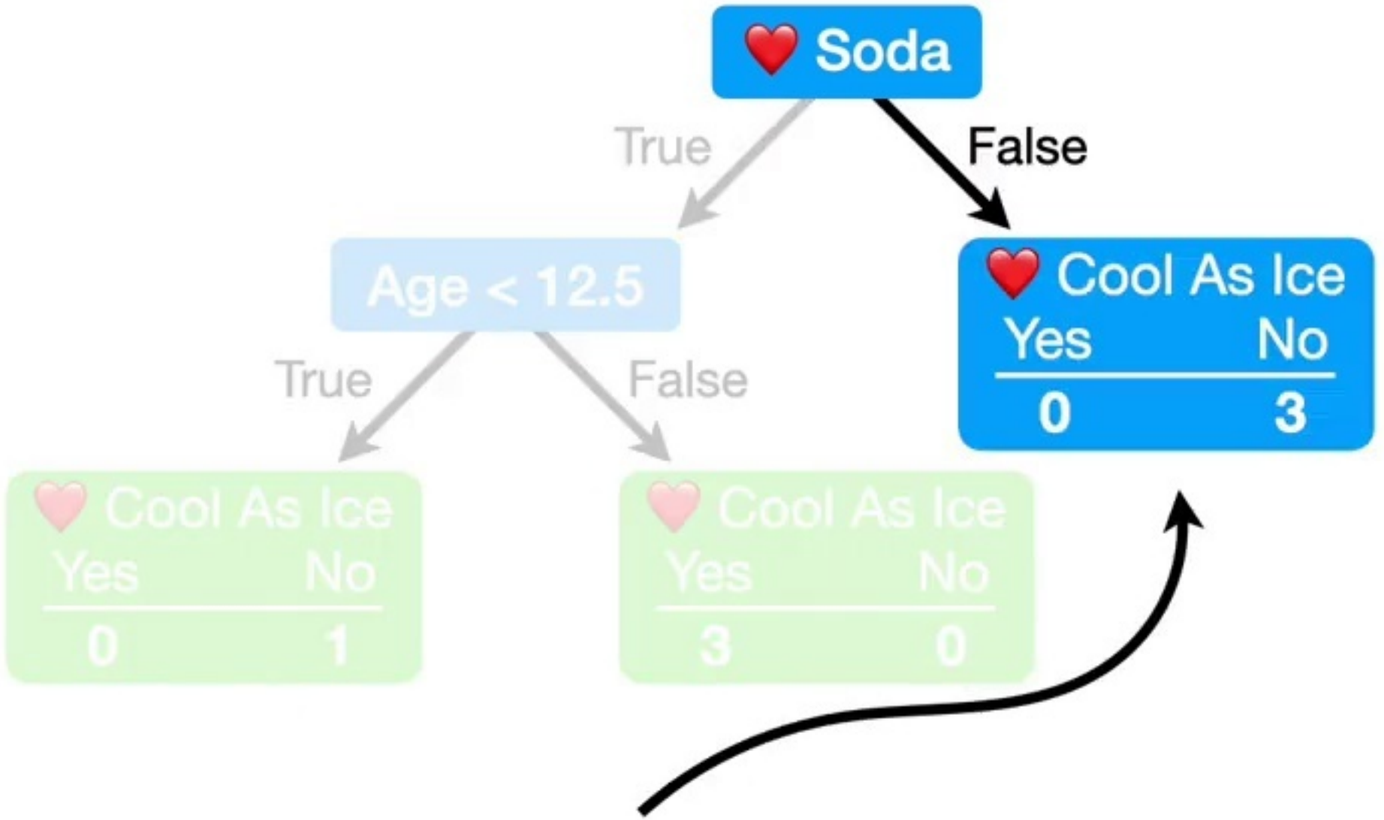
Now, because **0** is less than **0.25**, we will use **Age < 12.5** to split this **Node** into **Leaves**.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	**Yes**	7	No
Yes	No	12	No
No	**Yes**	18	Yes
No	**Yes**	35	Yes
Yes	**Yes**	38	Yes
Yes	No	50	No
No	No	83	No



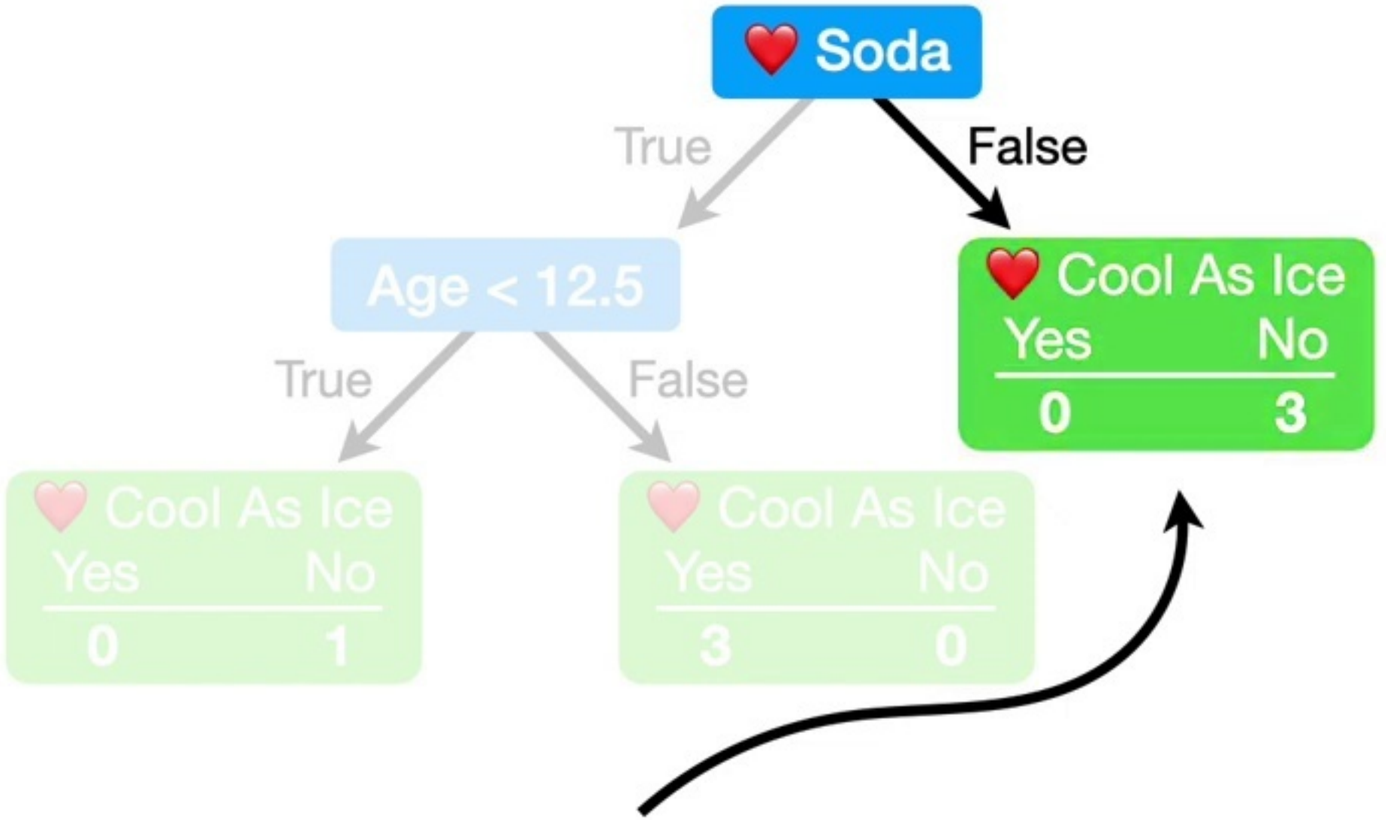
NOTE: These are **Leaves** because there is no reason to continue splitting these people into smaller groups.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	**No**	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	**No**	50	No
No	**No**	83	No



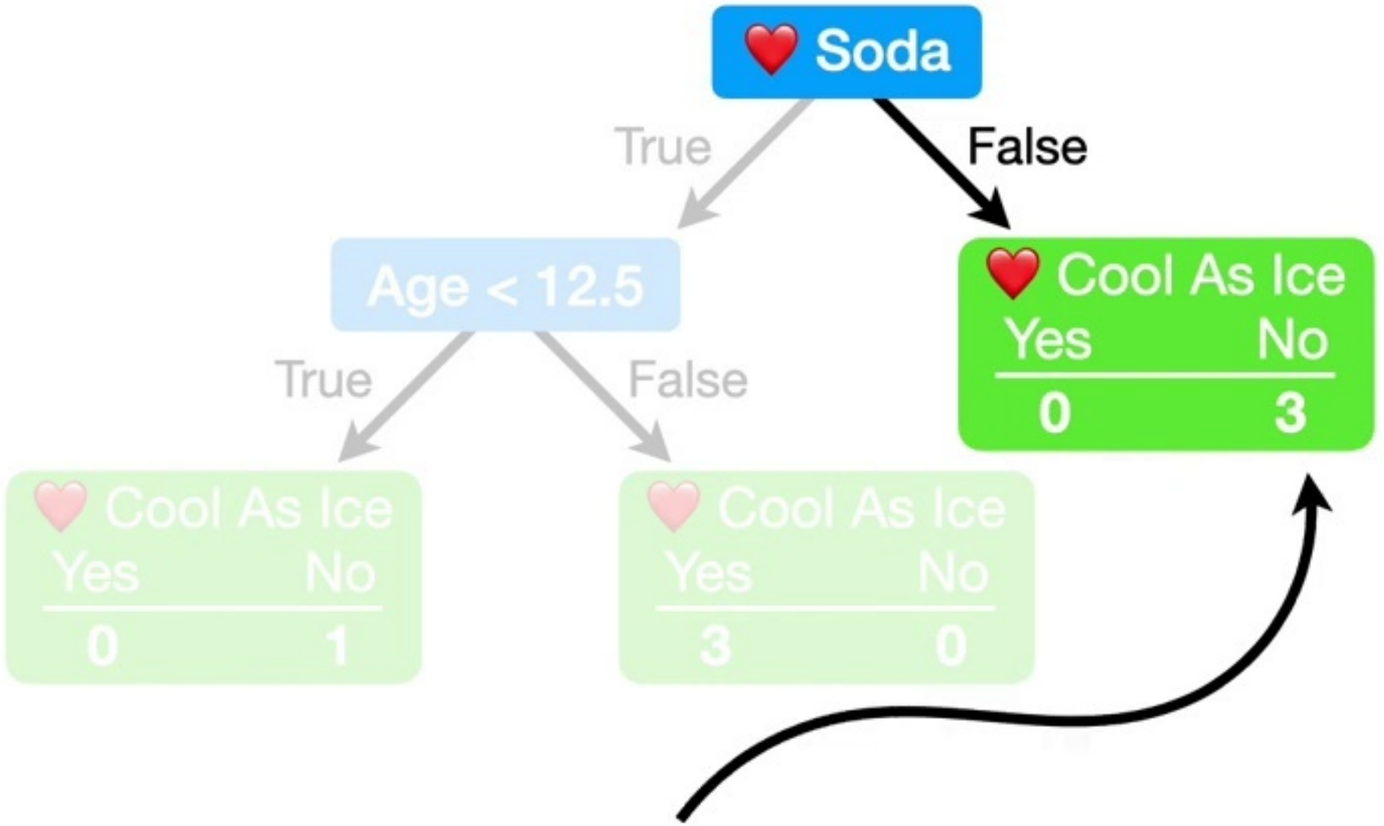
Likewise, this **Node**, consisting of the **3** people who *do not* **Love Soda**, is also a **Leaf**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	**No**	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	**No**	50	No
No	**No**	83	No



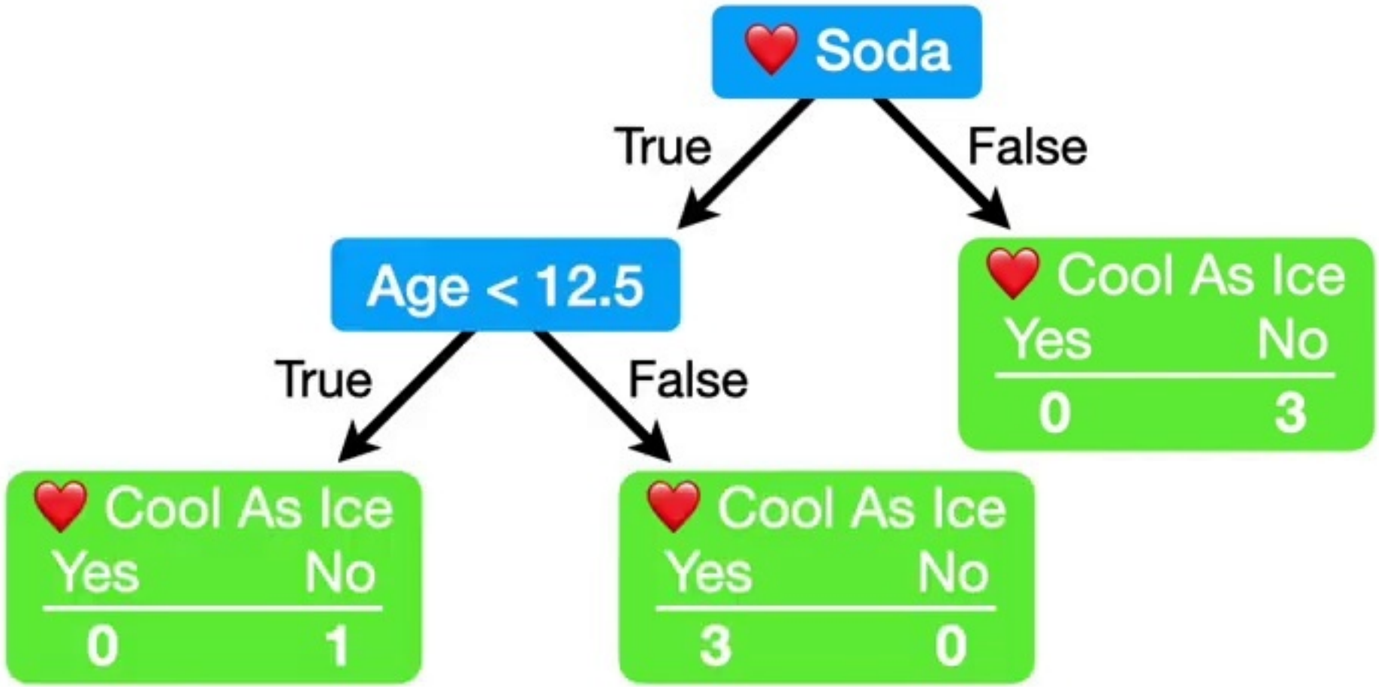
Likewise, this **Node**, consisting of the **3** people who *do not* **Love Soda**, is also a **Leaf**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	**No**	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	**No**	50	No
No	**No**	83	No

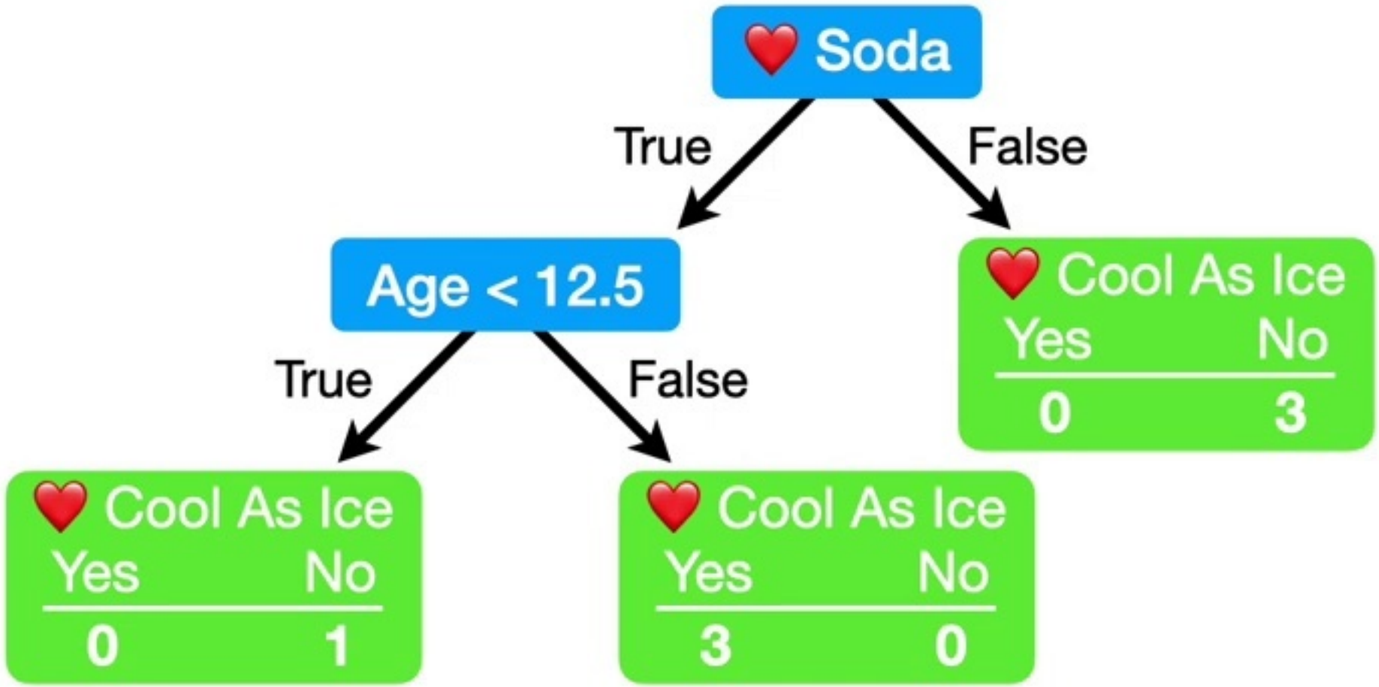


…because there is no reason to continue splitting these people into smaller groups.

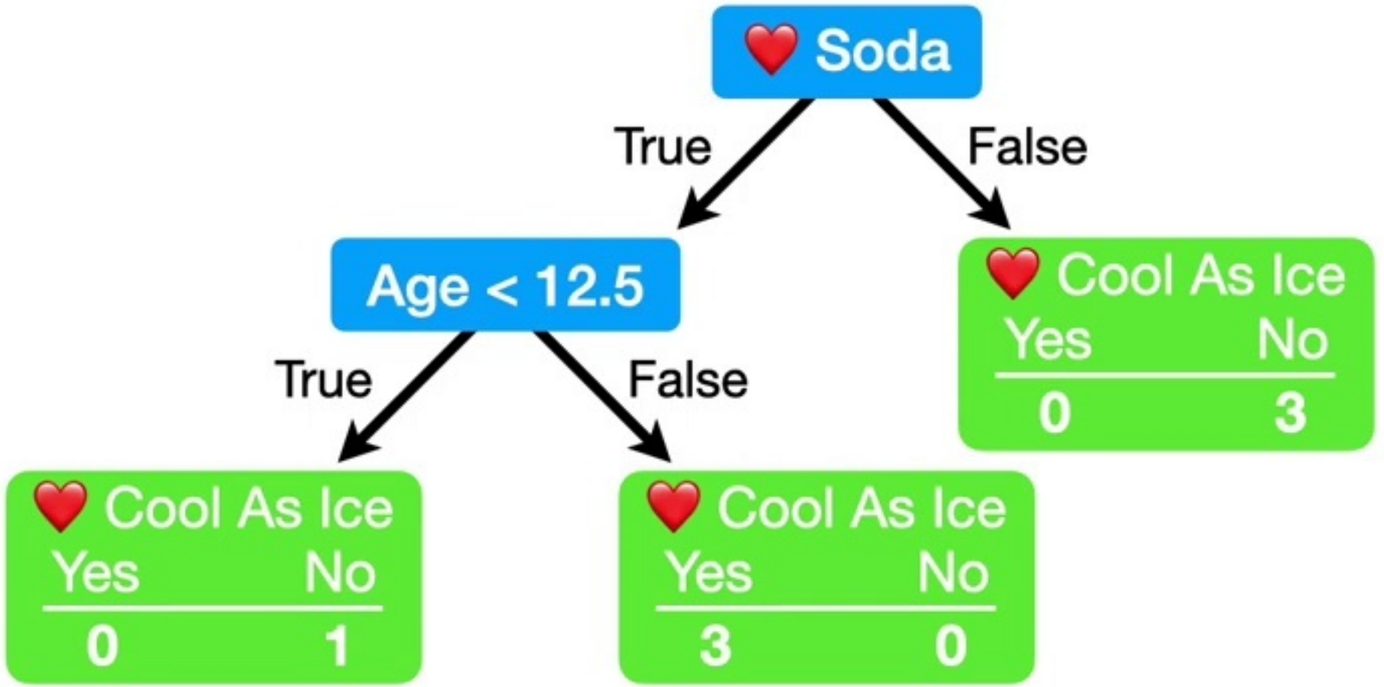
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



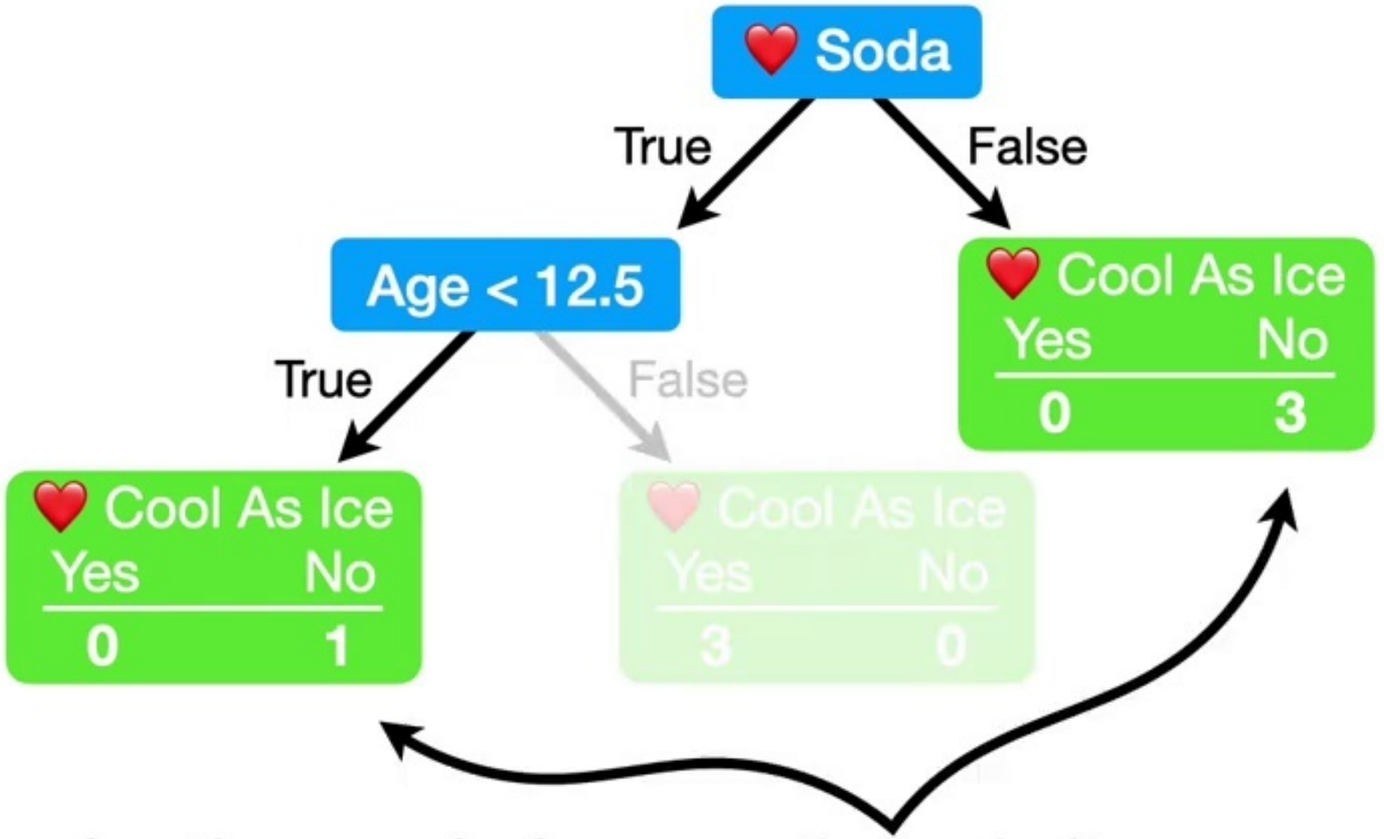
Now there is just one last thing we need to do before we are done building this tree.



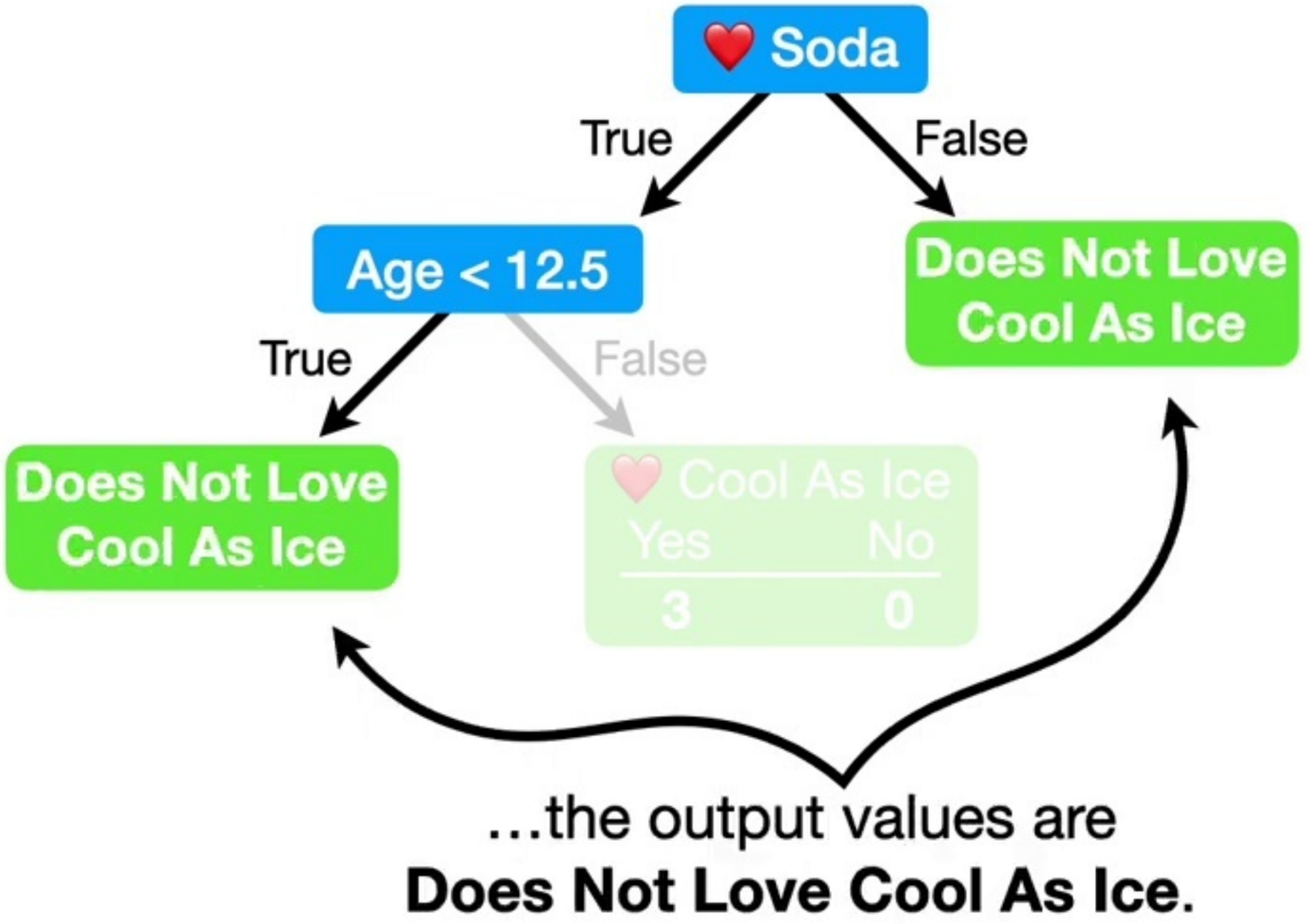
We need to assign output values for each **Leaf**.

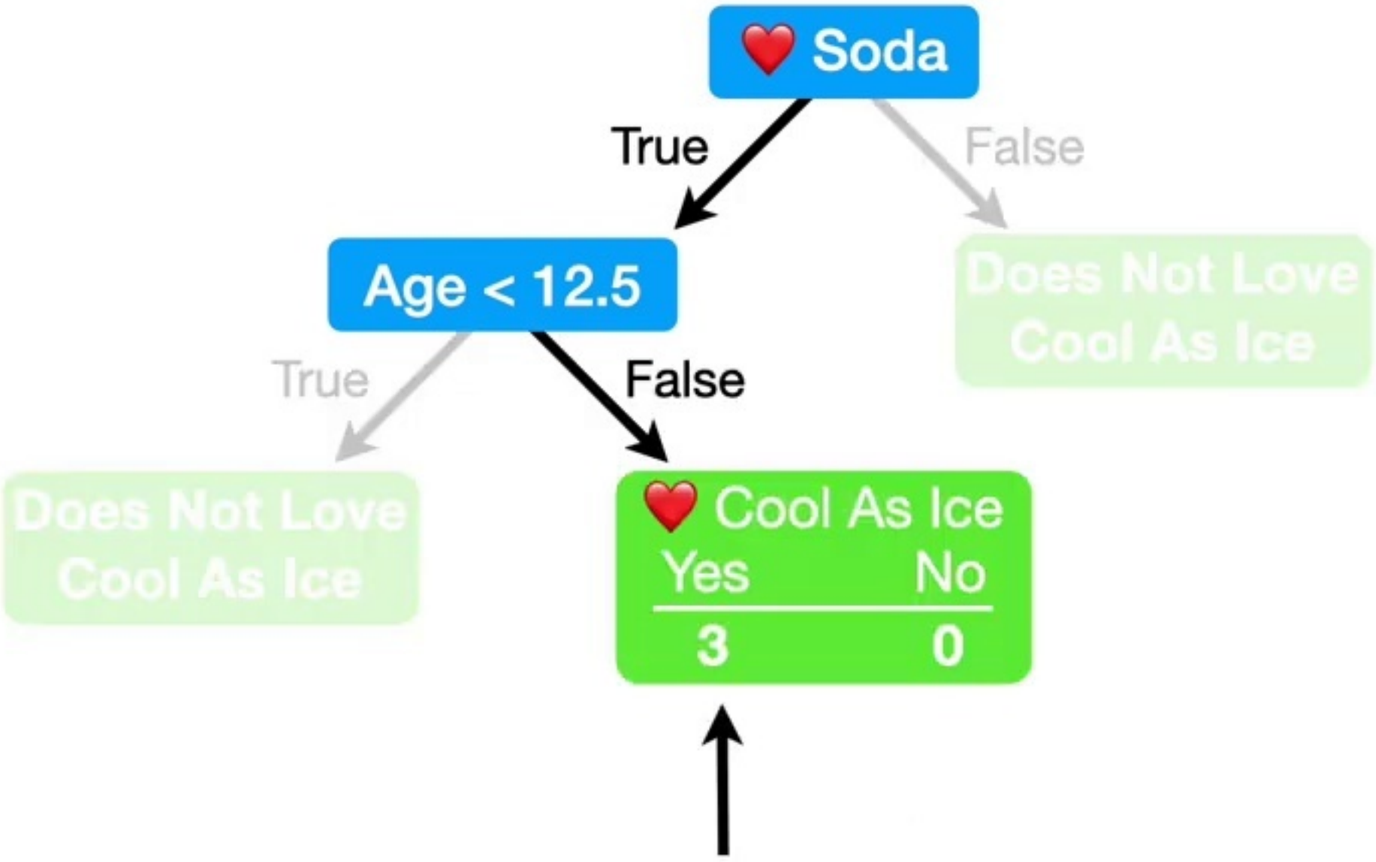


Generally speaking, the output of a **Leaf** is whatever category that has the most votes.

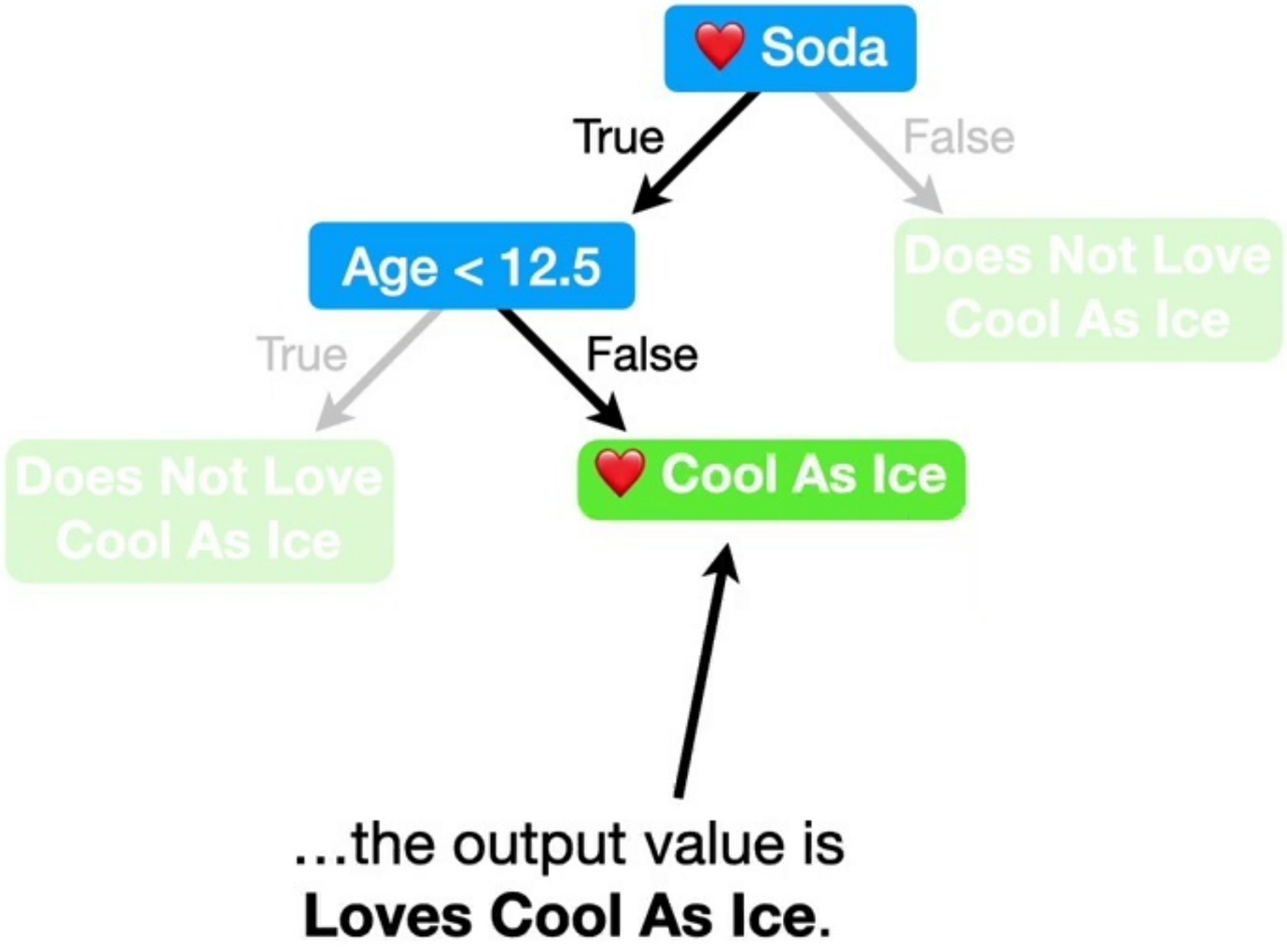


In other words, because the majority of the people in these **Leaves** *do not* **Love Cool As Ice**…

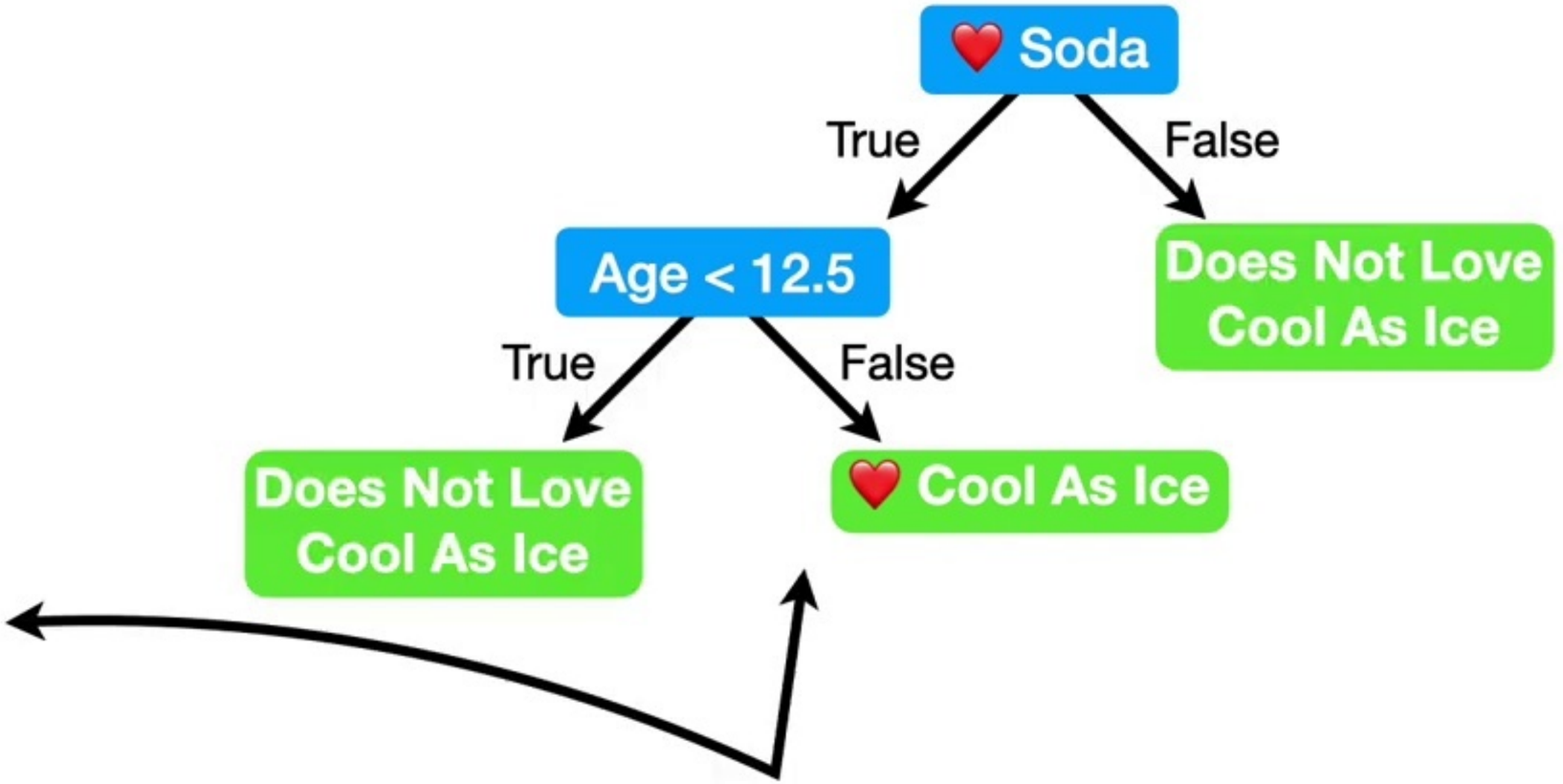

❤️ Soda
True
False
Age < 12.5
Does Not Love Cool As Ice
True
False
Does Not Love Cool As Ice
❤️ Cool As Ice
Yes
No
3
0
…the output values are
Does Not Love Cool As Ice.



And because the majority of the people in this **Leaf Love Cool As Ice**…


❤️ Soda
True
False
Age < 12.5
Does Not Love Cool As Ice
True
False
Does Not Love Cool As Ice
❤️ Cool As Ice
…the output value is
Loves Cool As Ice.

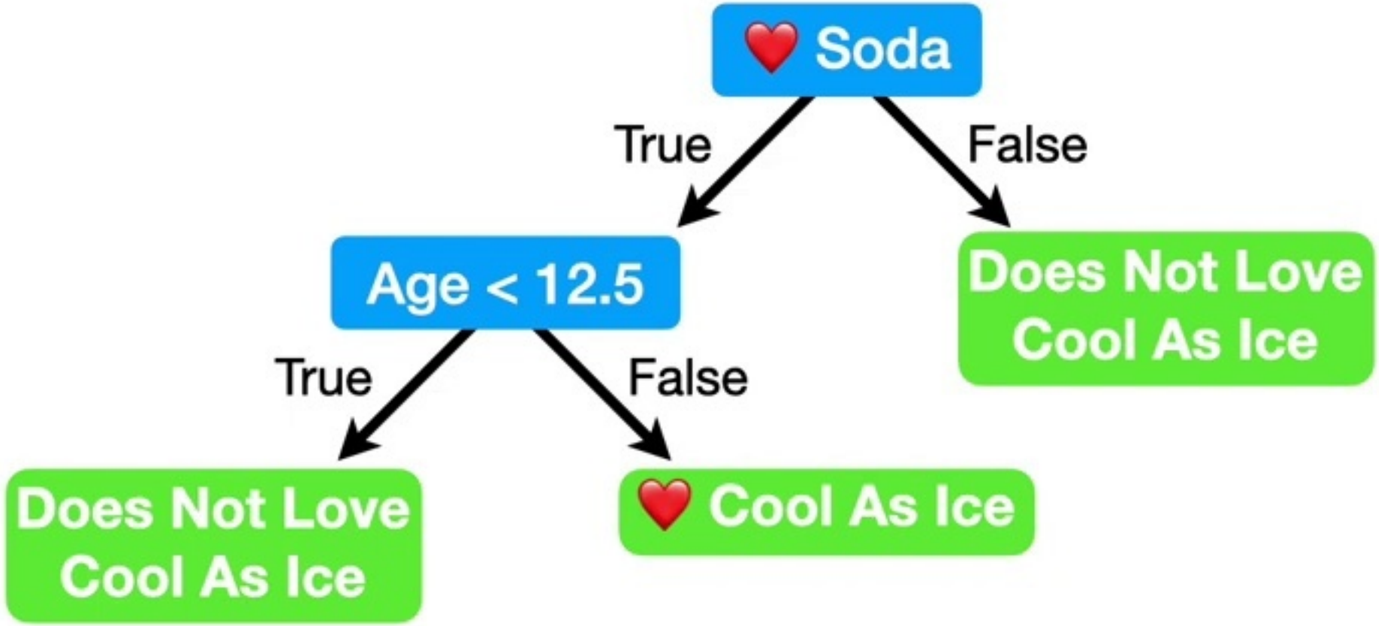
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



Hooray!!!
We finished building a **Tree** from this data.

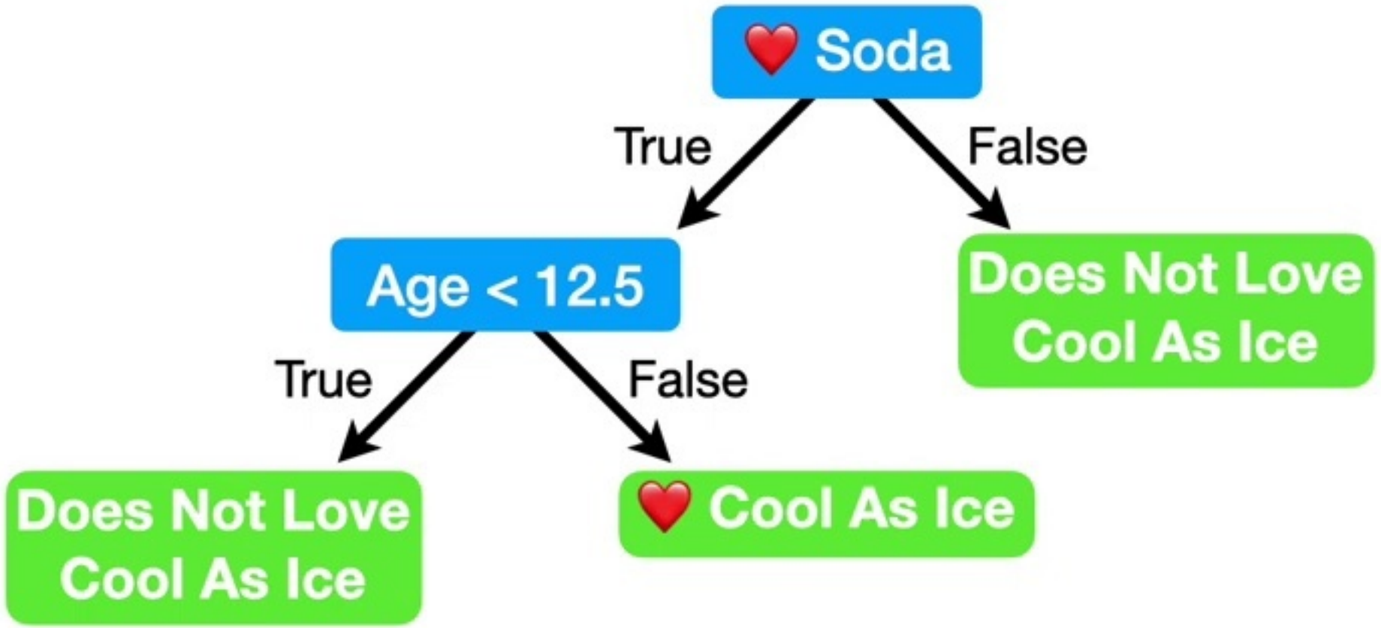
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	15	???

Now, if someone new comes along…

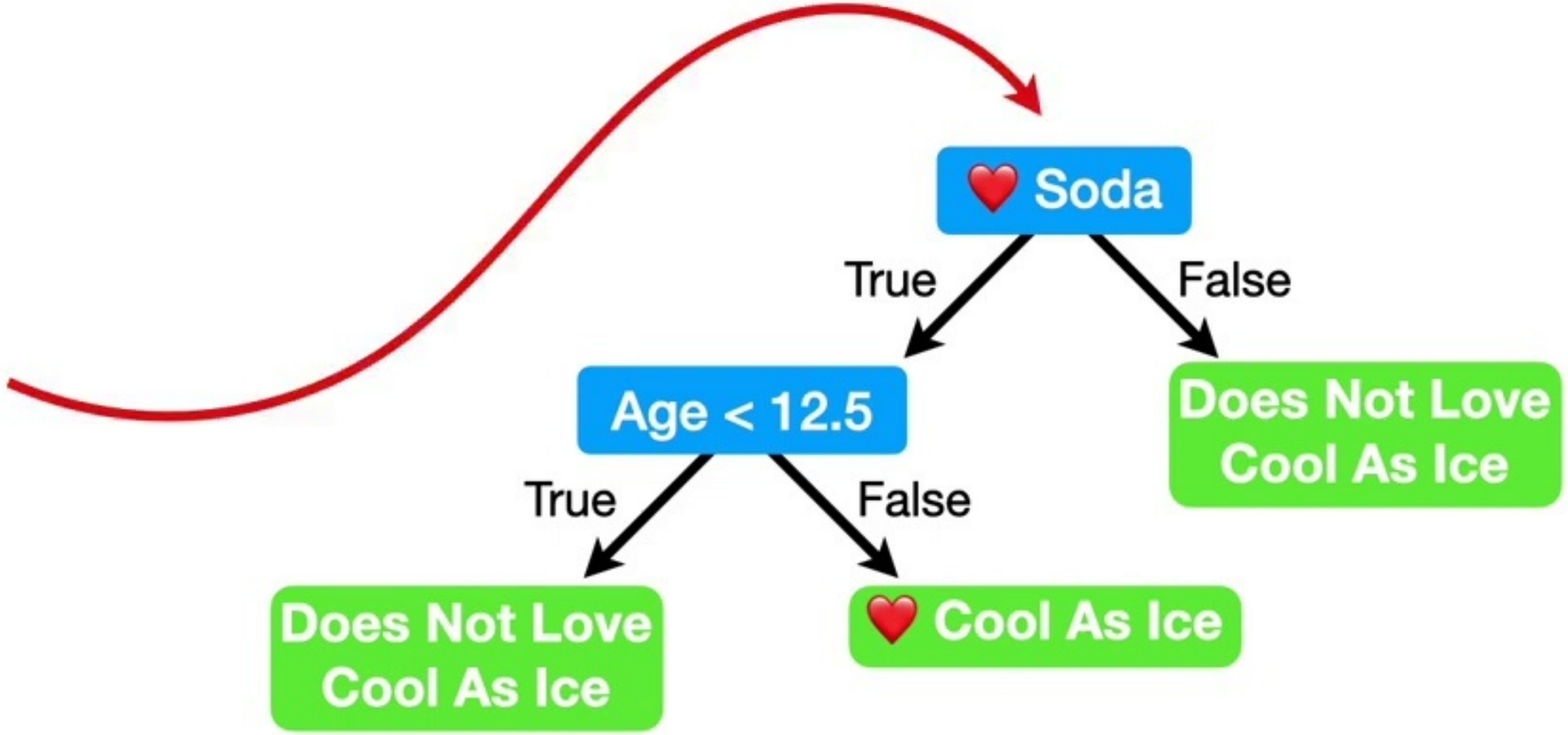


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	15	???

…and we want to predict if they will **Love Cool As Ice**…

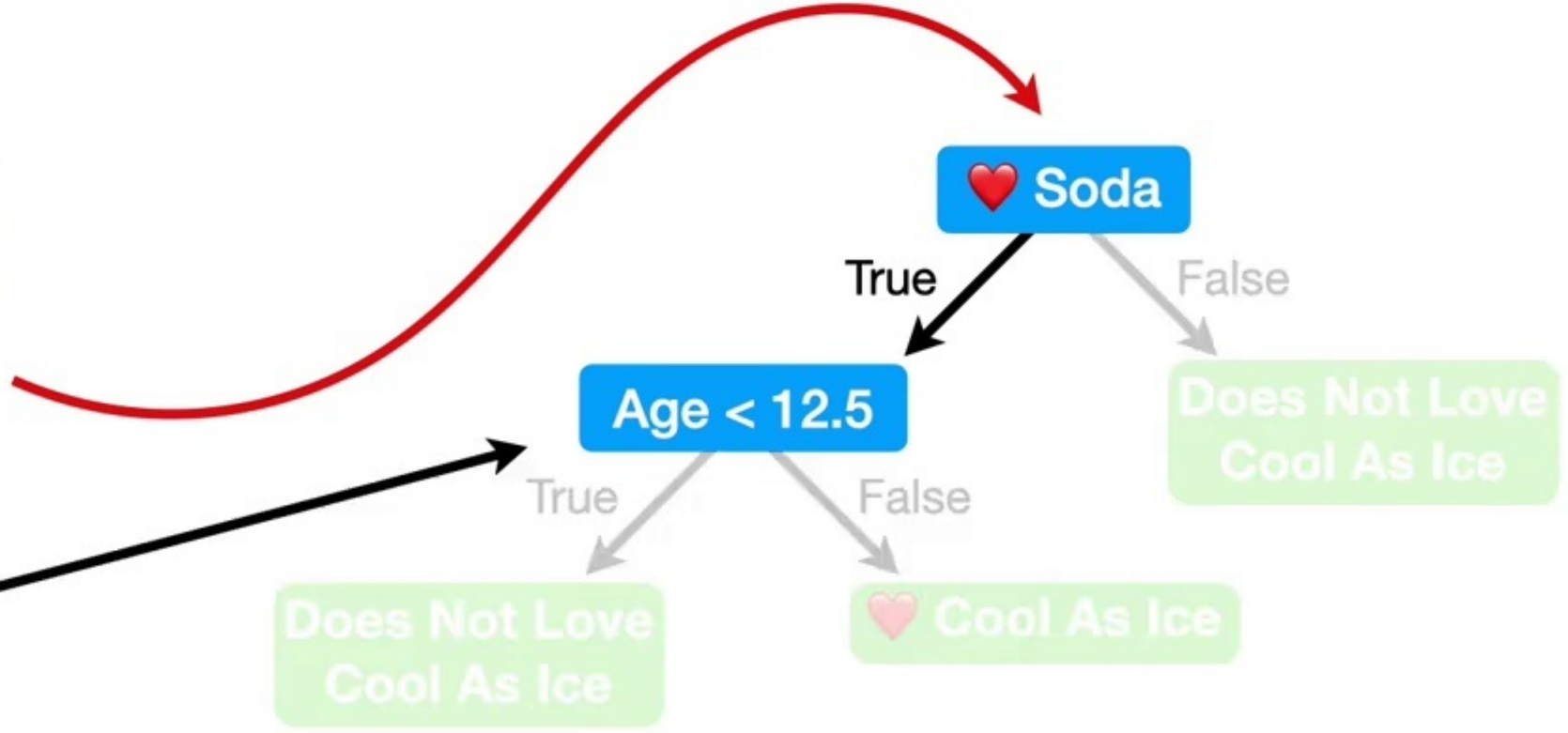


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	15	???



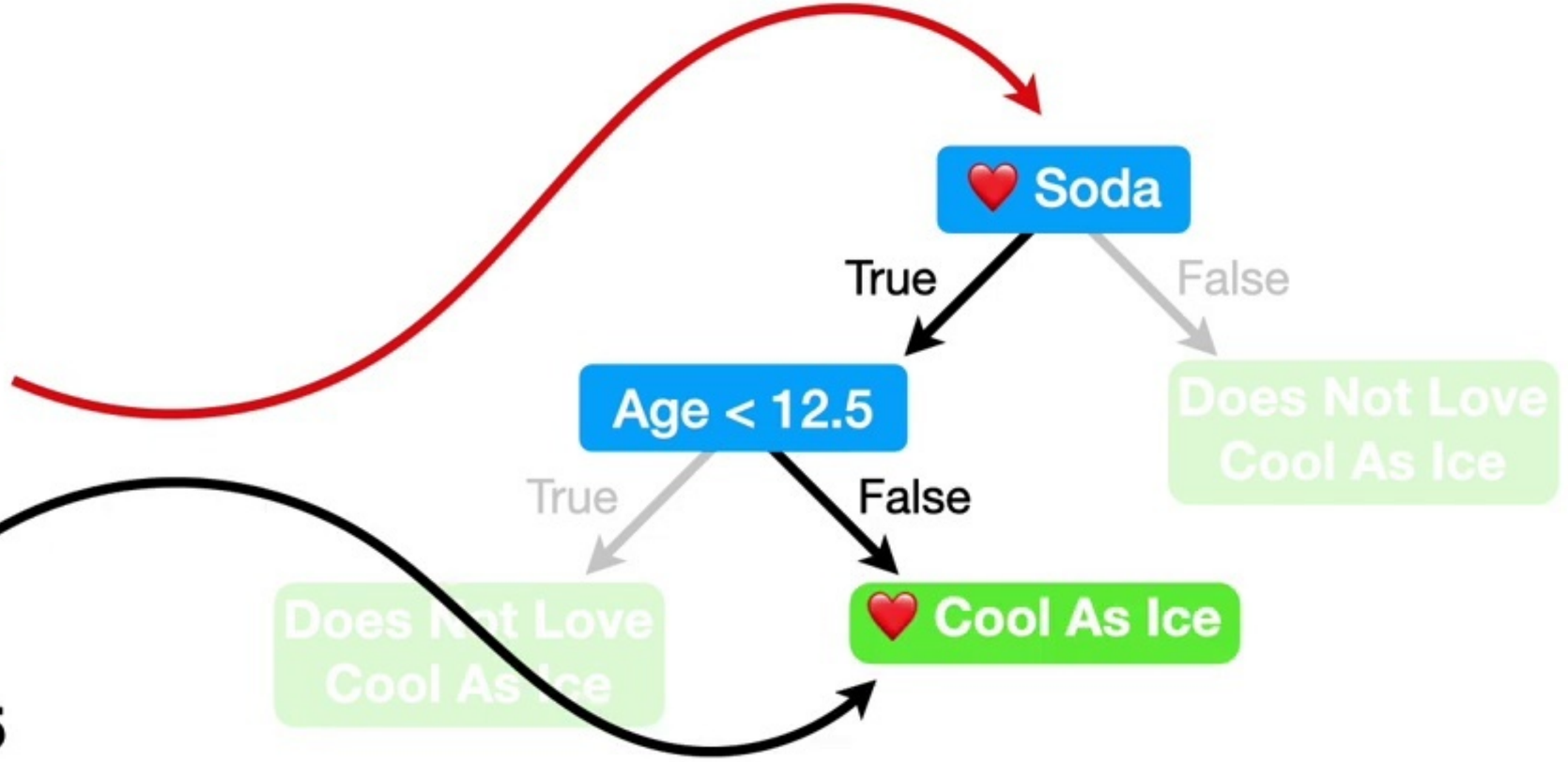
…then we run the data down our tree.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	15	???



And because they **Love Soda**, they go to the left…

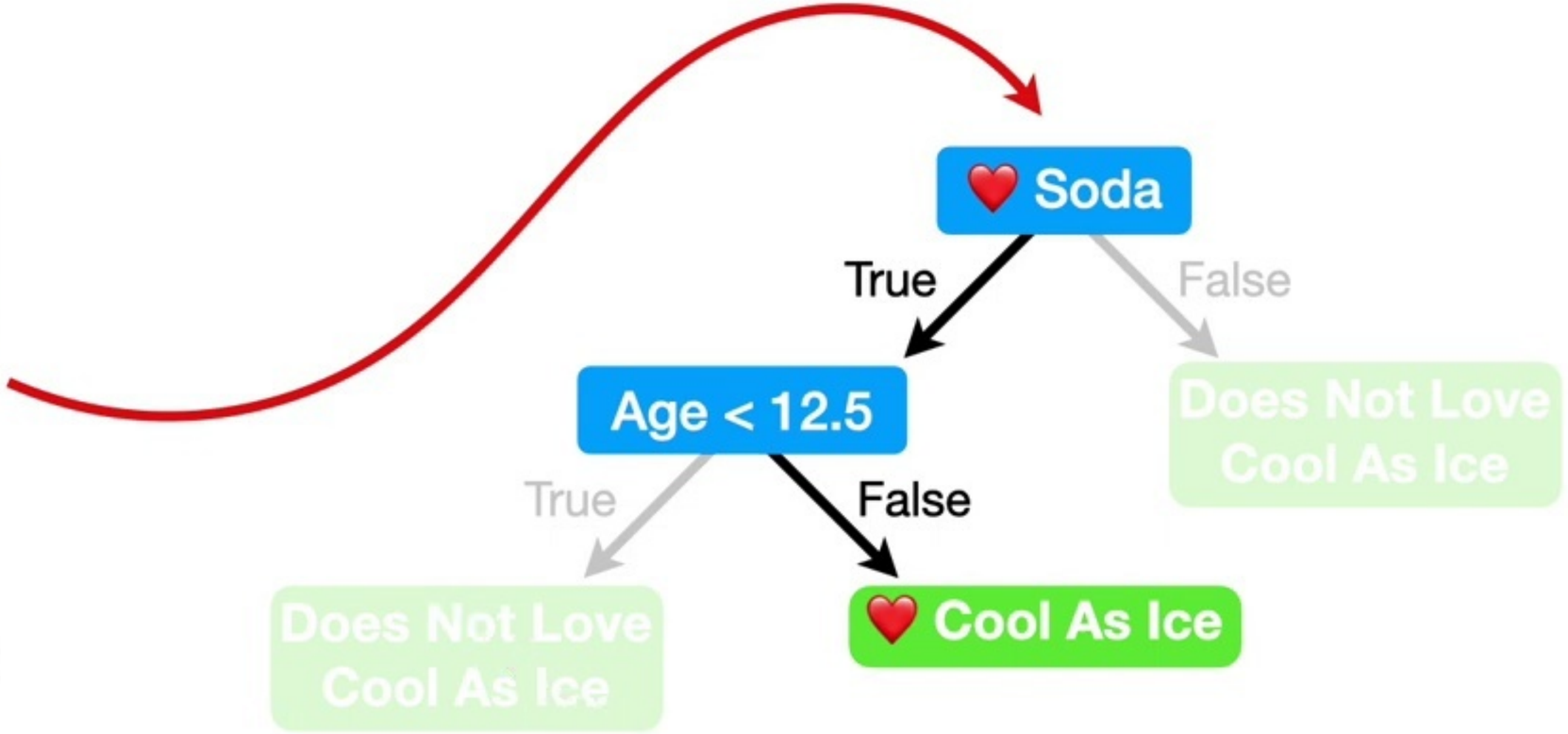
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	15	???



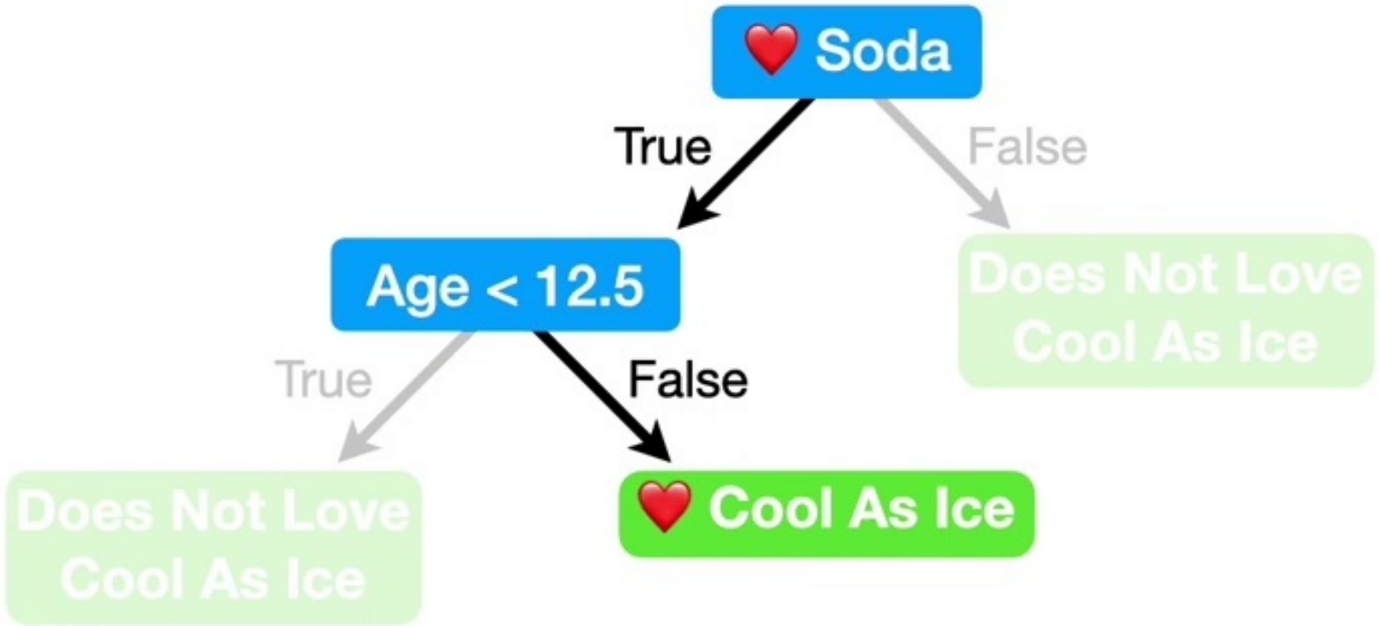
…and because they are **15**, so **Age < 12.5** is **False**, they end up in this **Leaf**…

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	15	**YES!!!**

…and we predict that they will **Love Cool As Ice!!!**

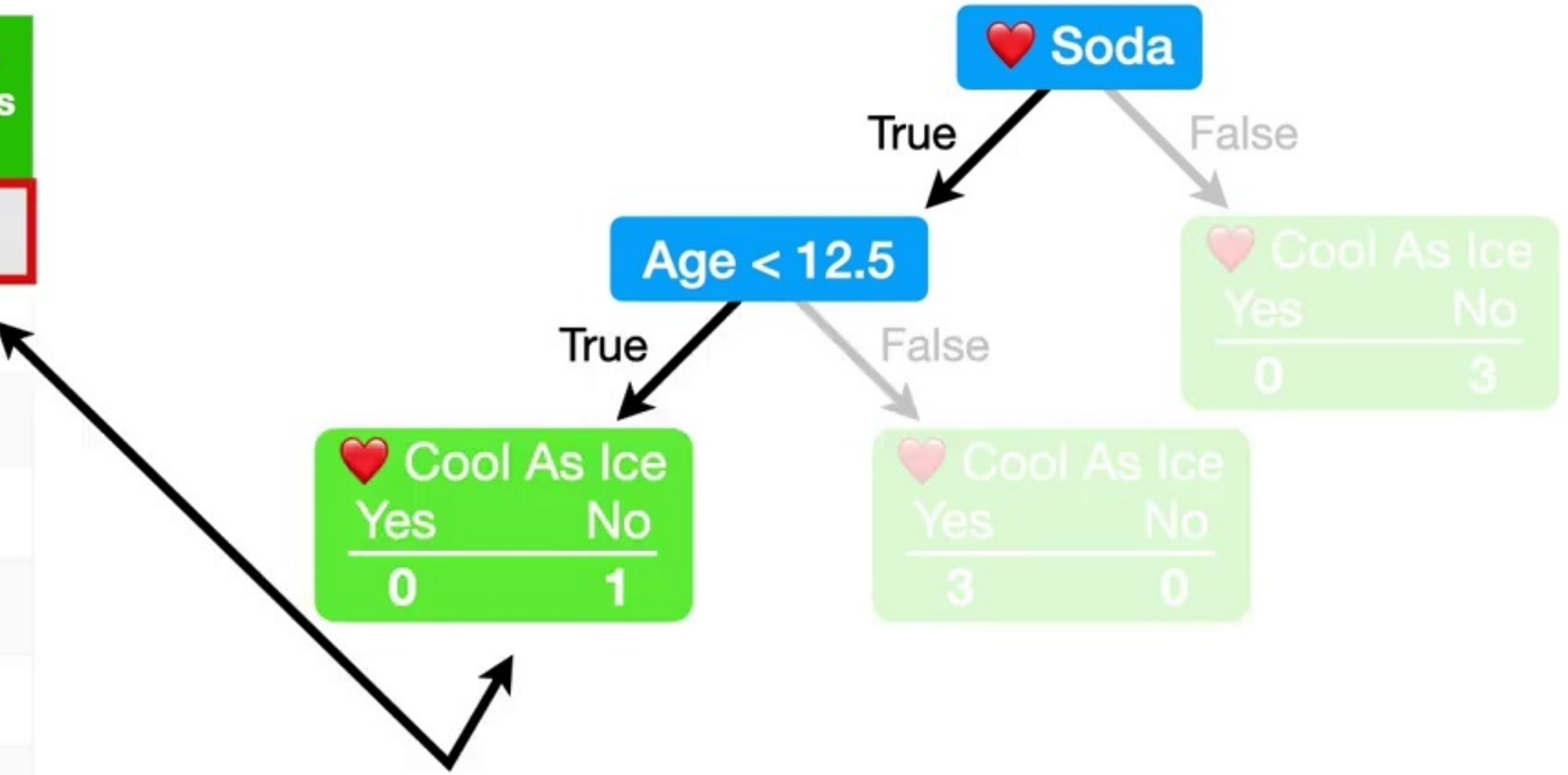


Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	15	**YES!!!**



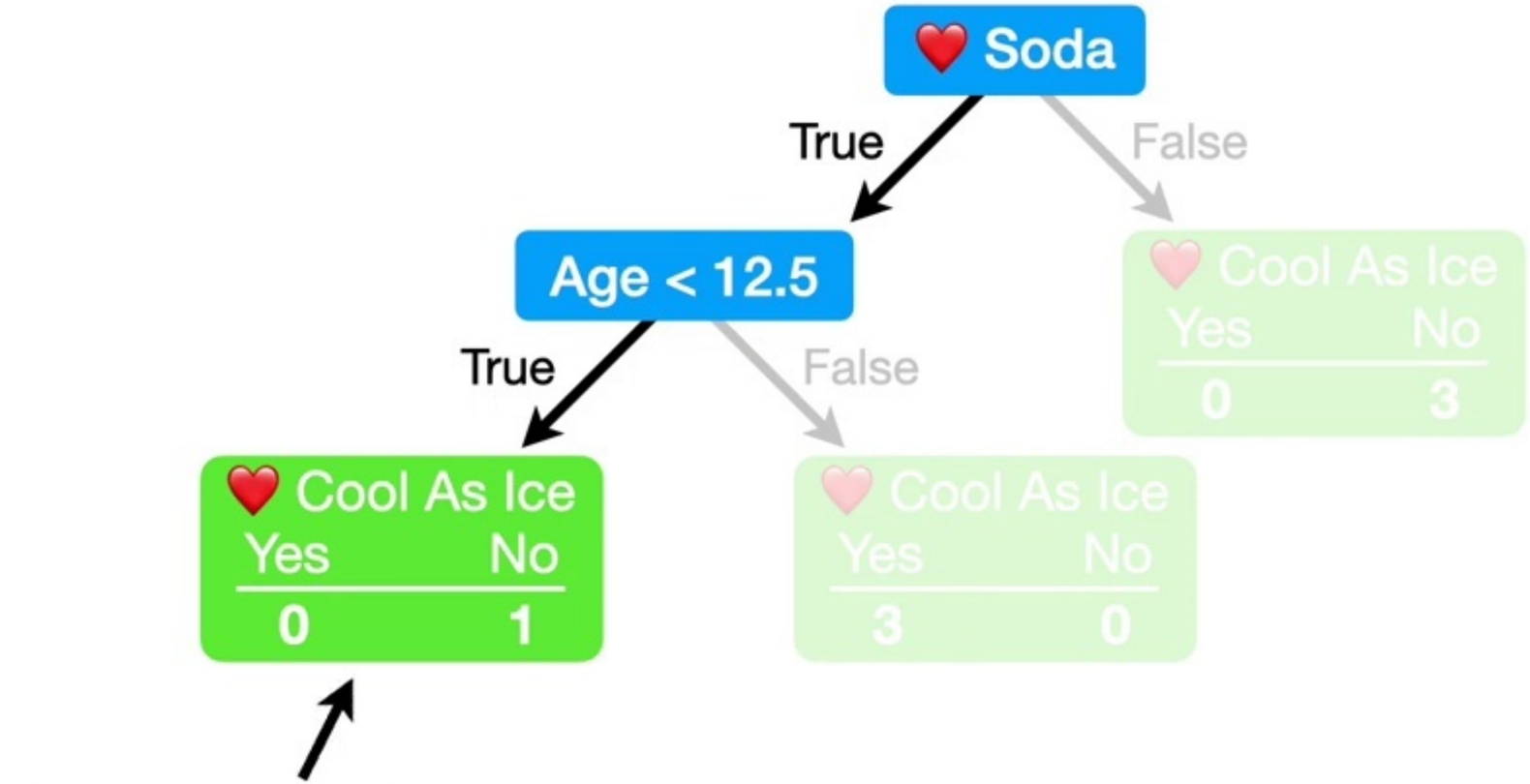
OK, now that we understand the **Main Ideas** of how to build and use **Classification Trees**, let's discuss one technical detail.

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

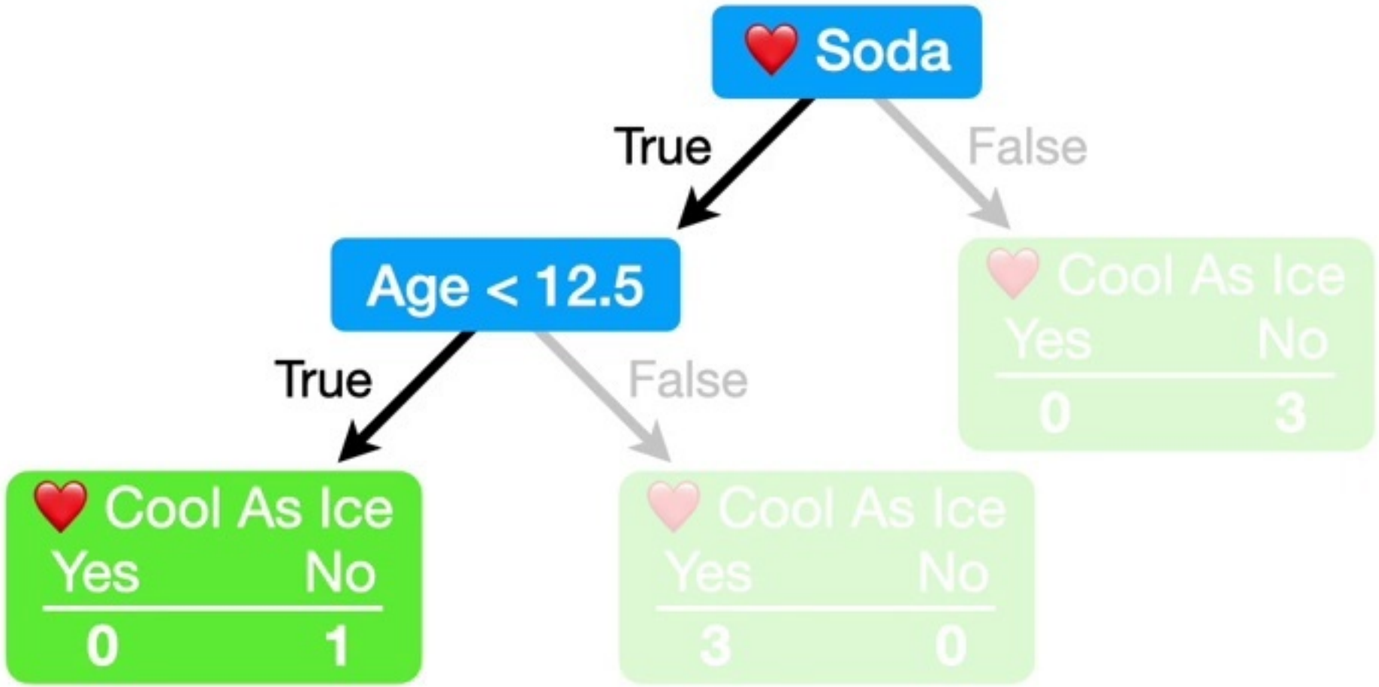


Remember, when we built this tree, only one person in the original dataset made it to this **Leaf**.

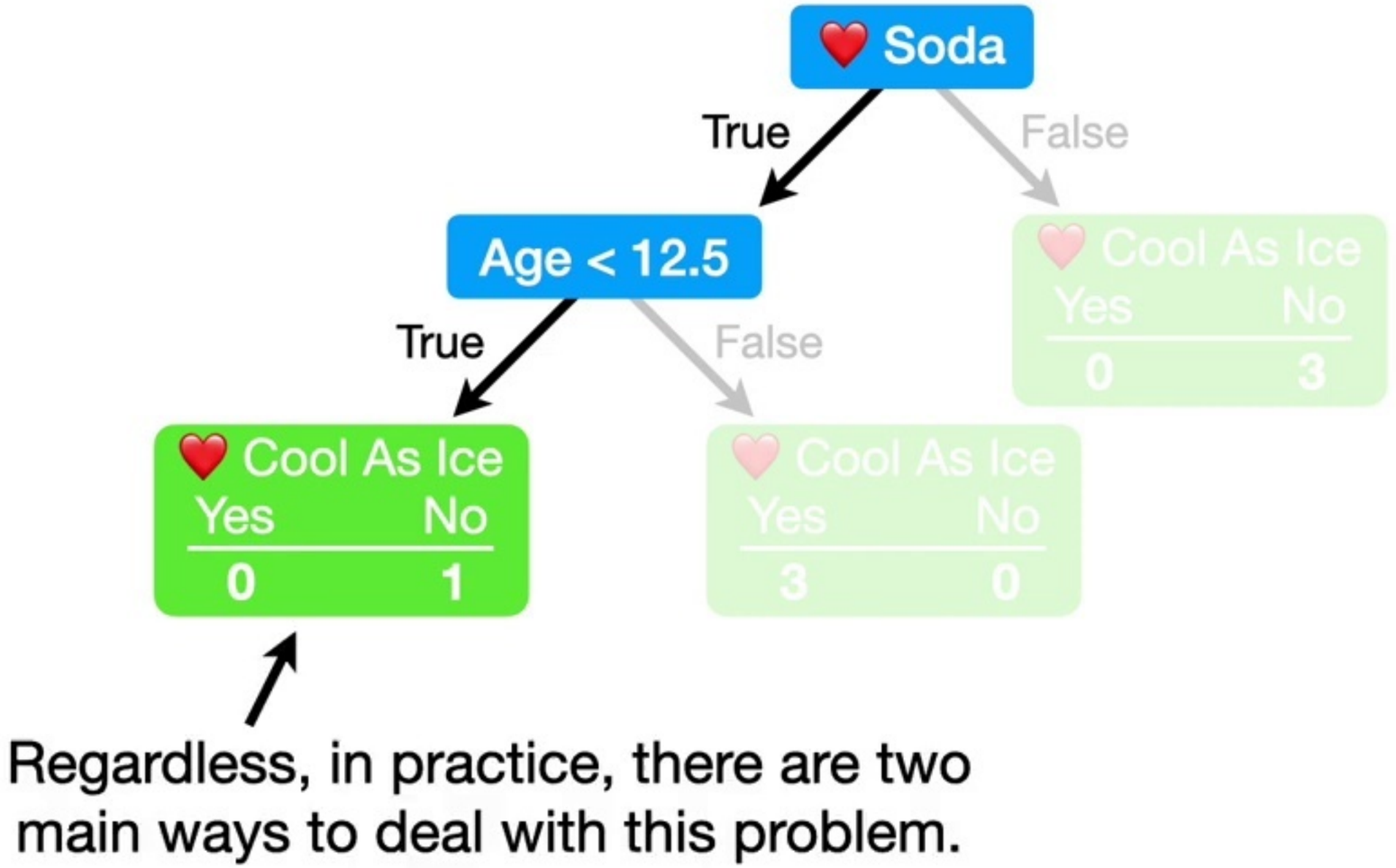
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

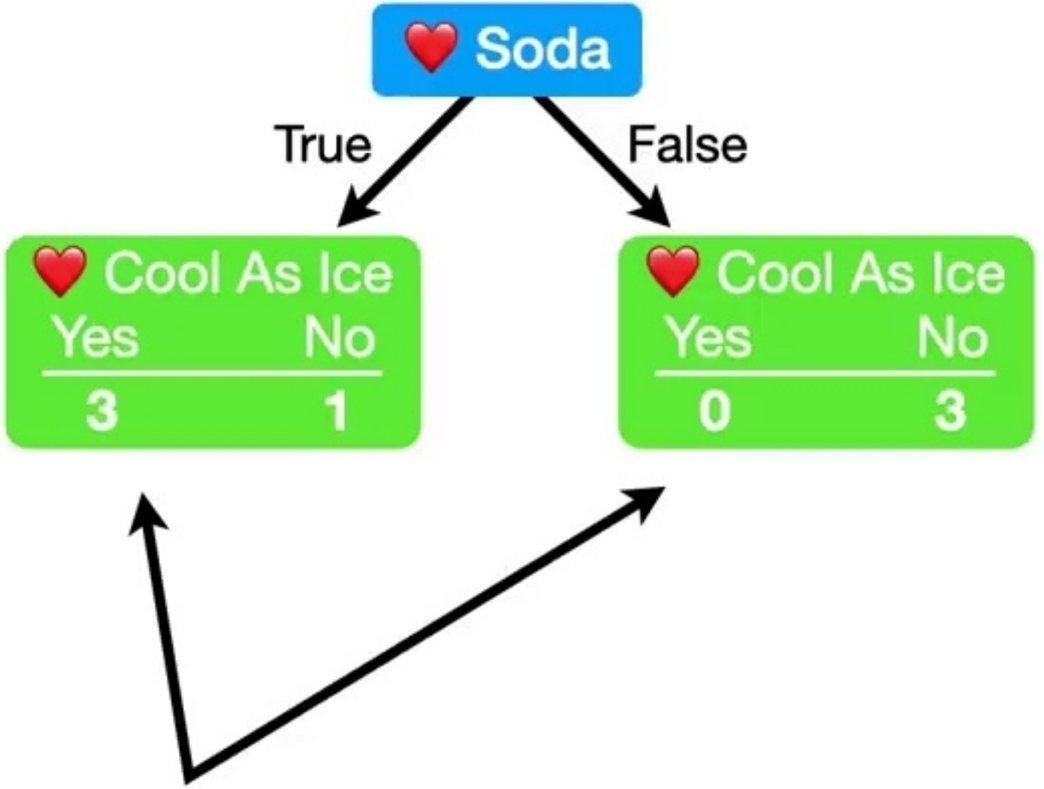


Because so few people made it to this **Leaf**, it's hard to have confidence that it will do a great job making predictions with future data.

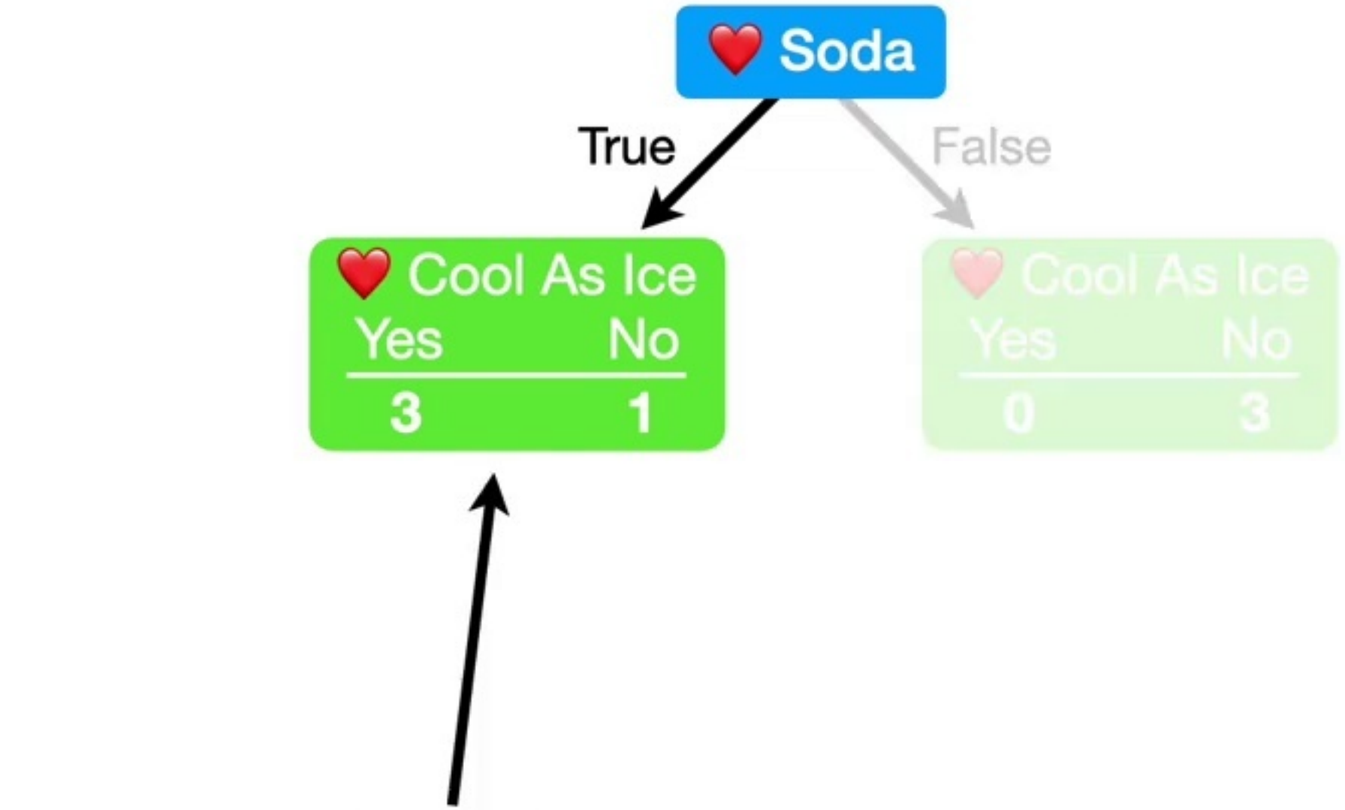


And it's possible that we have **Overfit** the data.

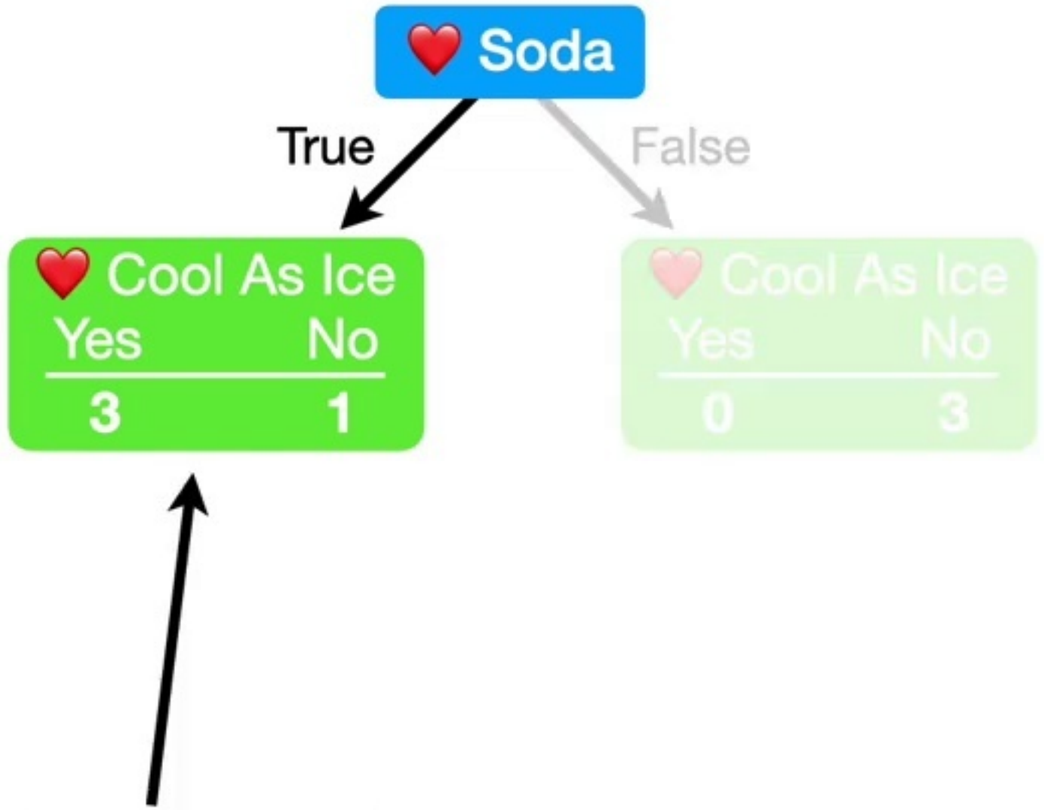

Soda
True
False
Age < 12.5
True
False
Cool As Ice
Yes No
0 3
Cool As Ice
Yes No
0 1
Cool As Ice
Yes No
3 0
Regardless, in practice, there are two main ways to deal with this problem.



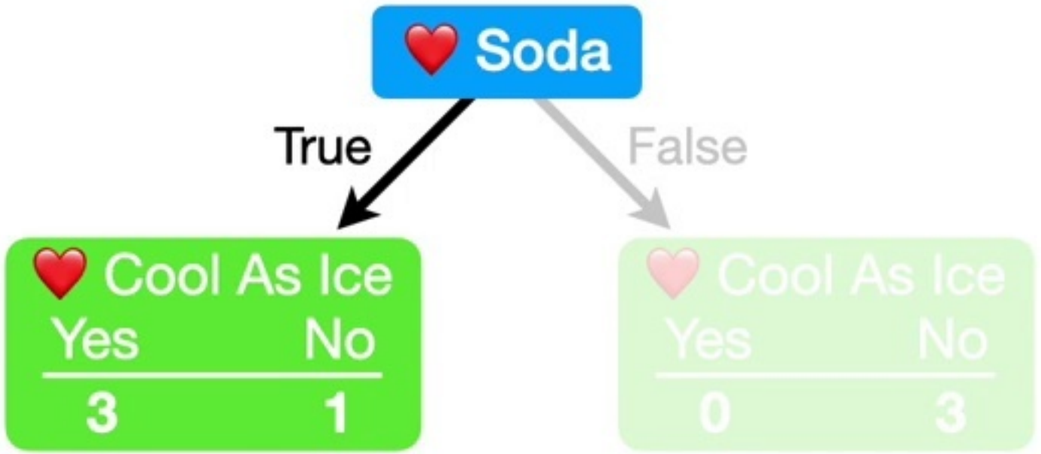
limits on how trees grow, for example, by requiring **3** or more people per leaf.



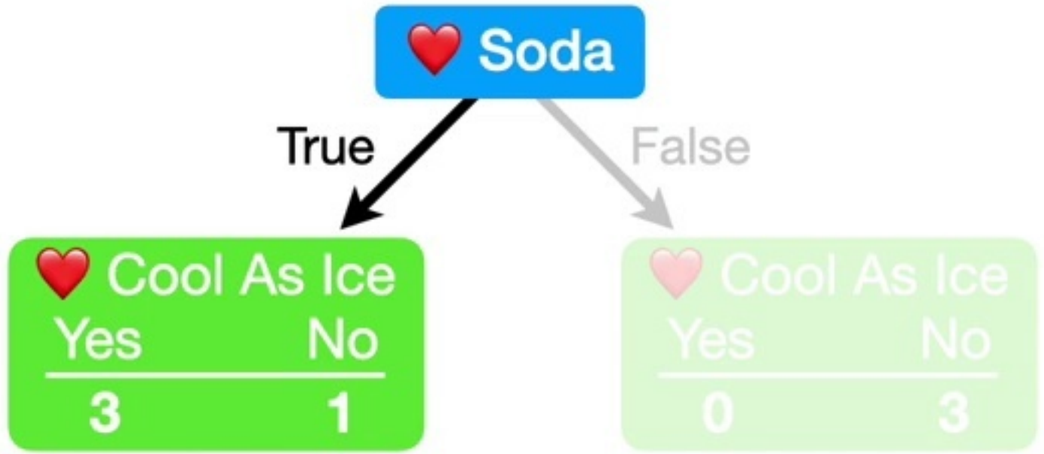
Now we end up with an
Impure Leaf…



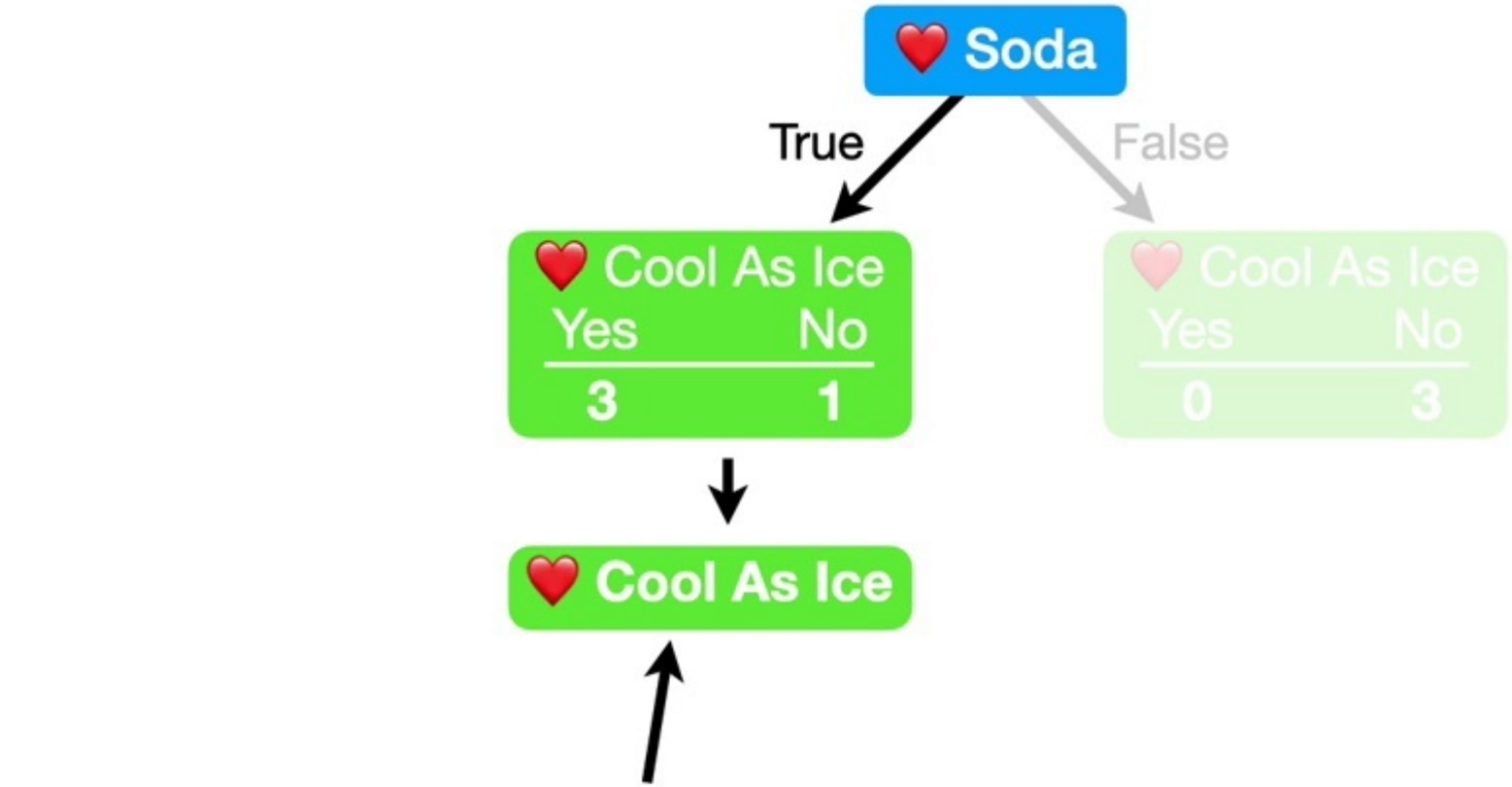
…but also a better sense of the accuracy of our prediction, because we know that only **75%** of the people in the **Leaf Loved Cool As Ice**.



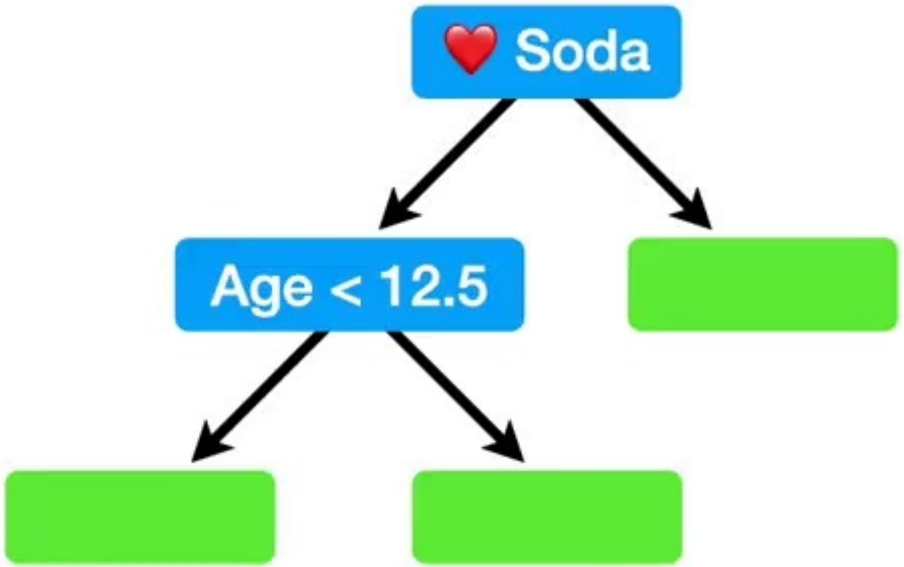
NOTE: Even when a **Leaf** is **Impure** we still need an output value to make a classification…



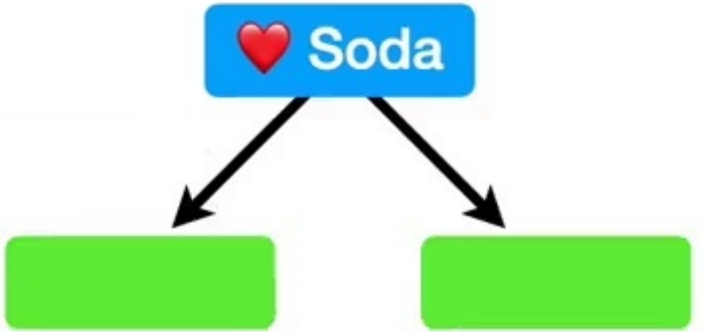
…and since most of the people in this leaf **Love Cool As Ice**, that will be the output value.



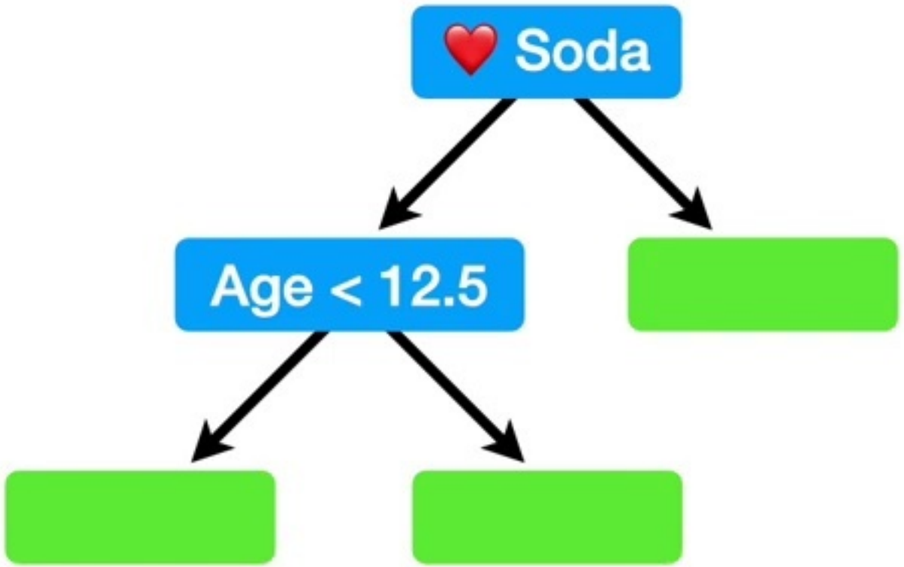
…and since most of the people in this leaf **Love Cool As Ice**, that will be the output value.



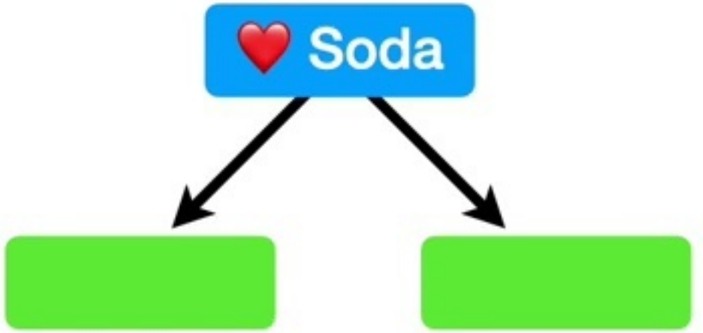
Vs.



ALSO NOTE: When we build a tree, we don't know in advance if it is better to require **3** people per leaf or some other number…



Vs.



…so we test different values with something called **Cross Validation** and pick the one that works best.