Supervised Learning - Decision Trees

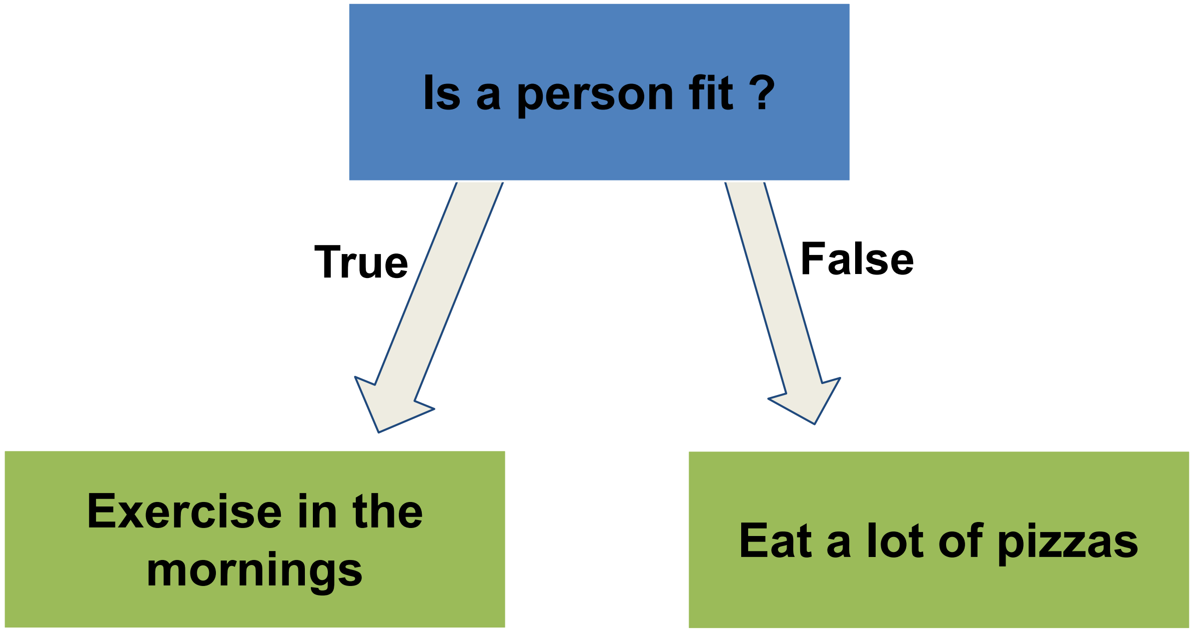
The University of Texas at Austin
College of Natural Sciences

Sep 25, 2025

Agenda

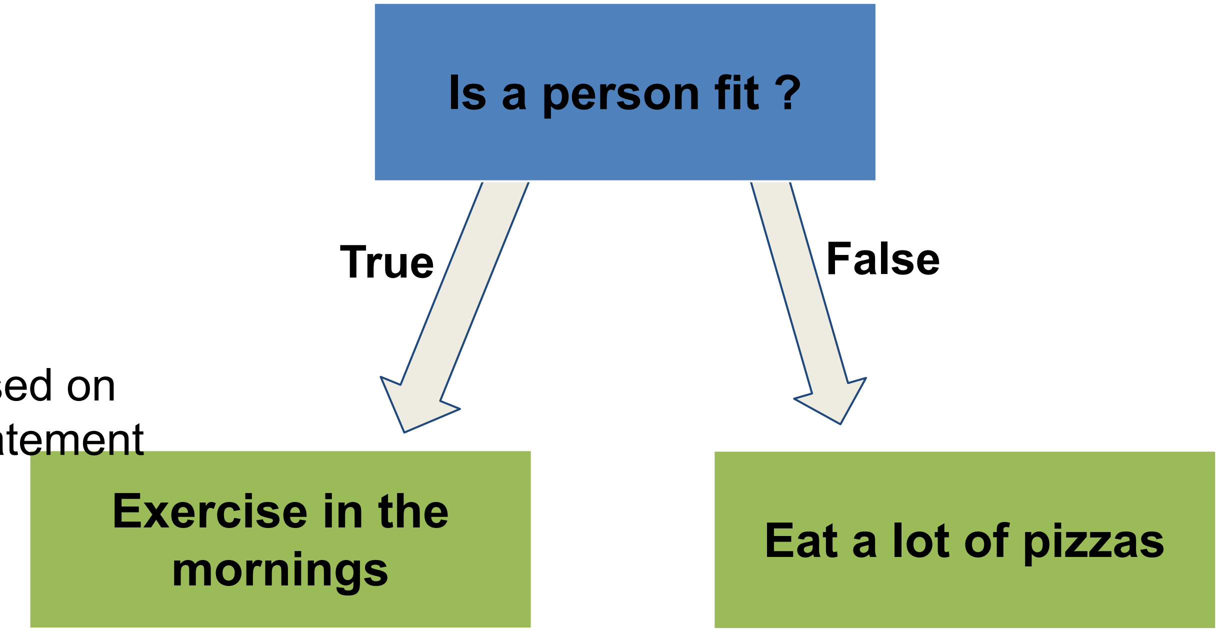
- Basic decision tree concepts
- Building a tree with Gini Impurity
- Numeric and continuous variables
- Adding branches
- Adding leaves
- Defining output values
- Using the tree
- How to prevent overfitting

Basic Decision Tree Concepts



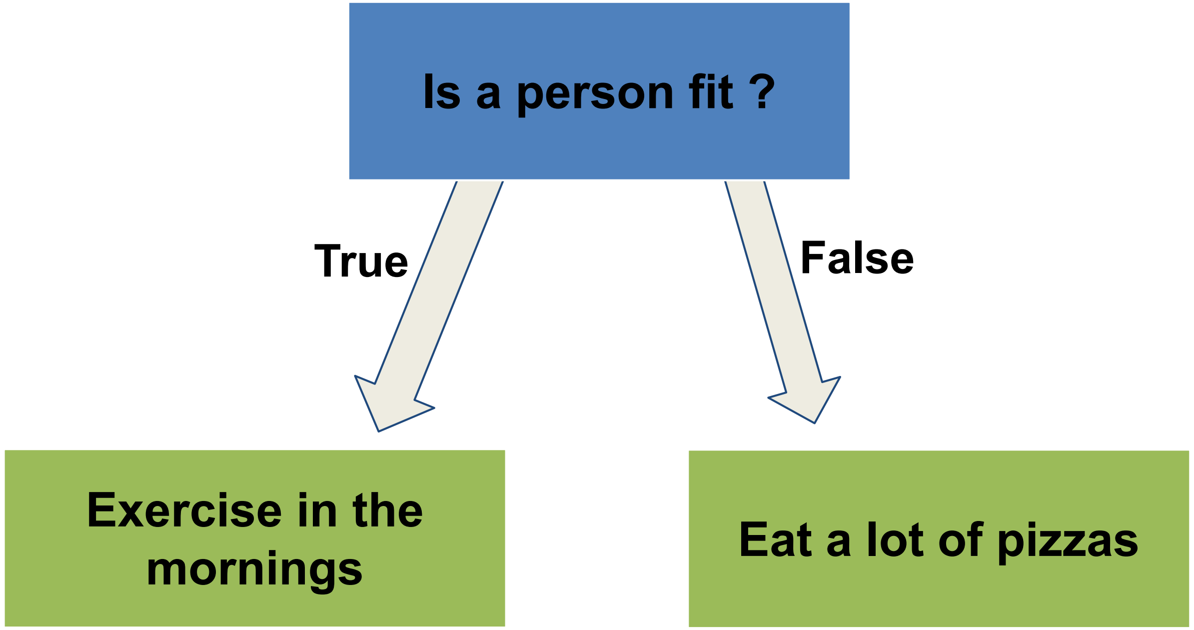
Basic Decision Tree Concepts

In general a **Decision Tree** makes a statement …



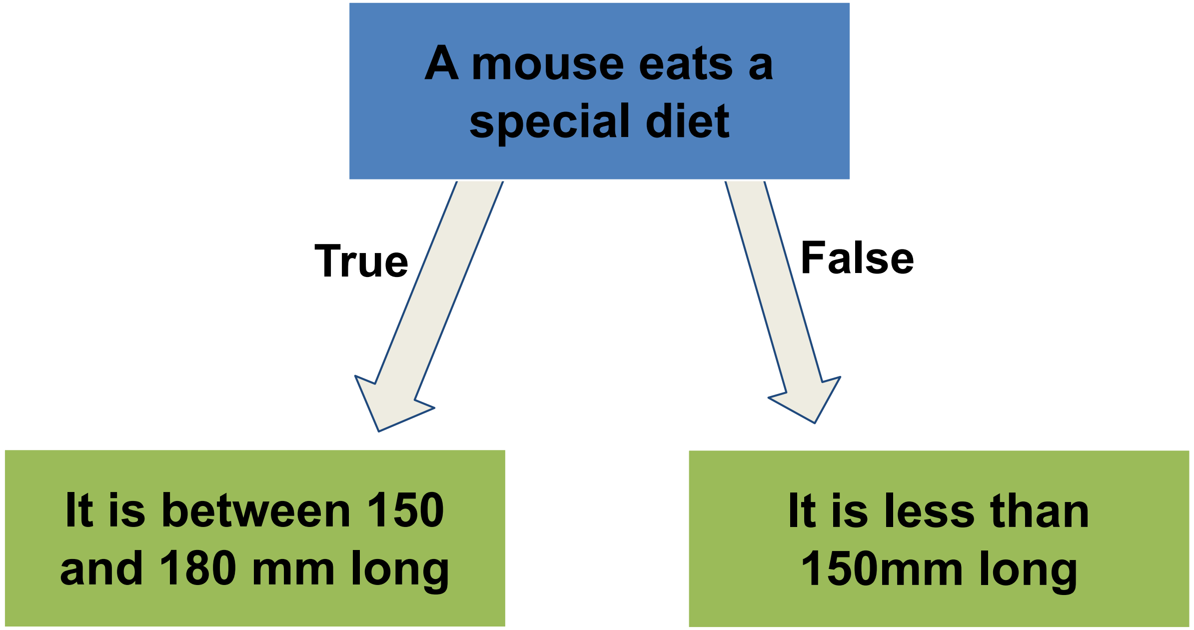
Makes a decision based on weather or not the statement is **True** or **False**

Basic Decision Tree Concepts



When a Decision Tree classifies things into categories…

Classification Tree

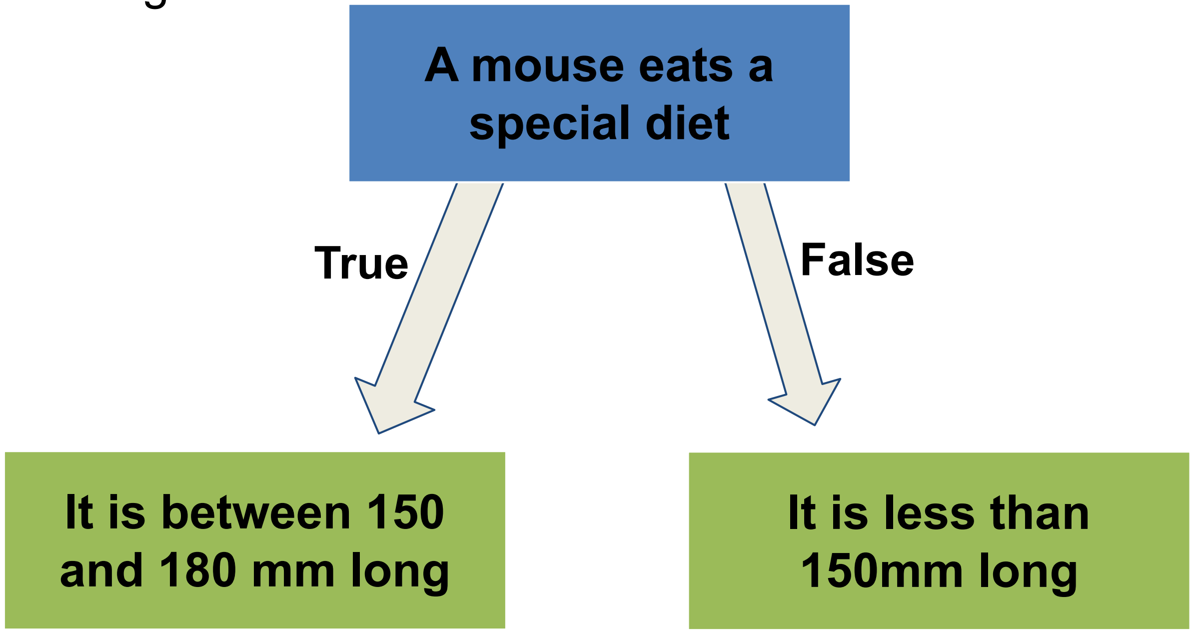


When a Decision Tree predicts numeric values…

Regression Tree

Basic Decision Tree Concepts

In this case, we are using diet…



… to predict a numeric value for mouse size

When a Decision Tree predicts numeric values…

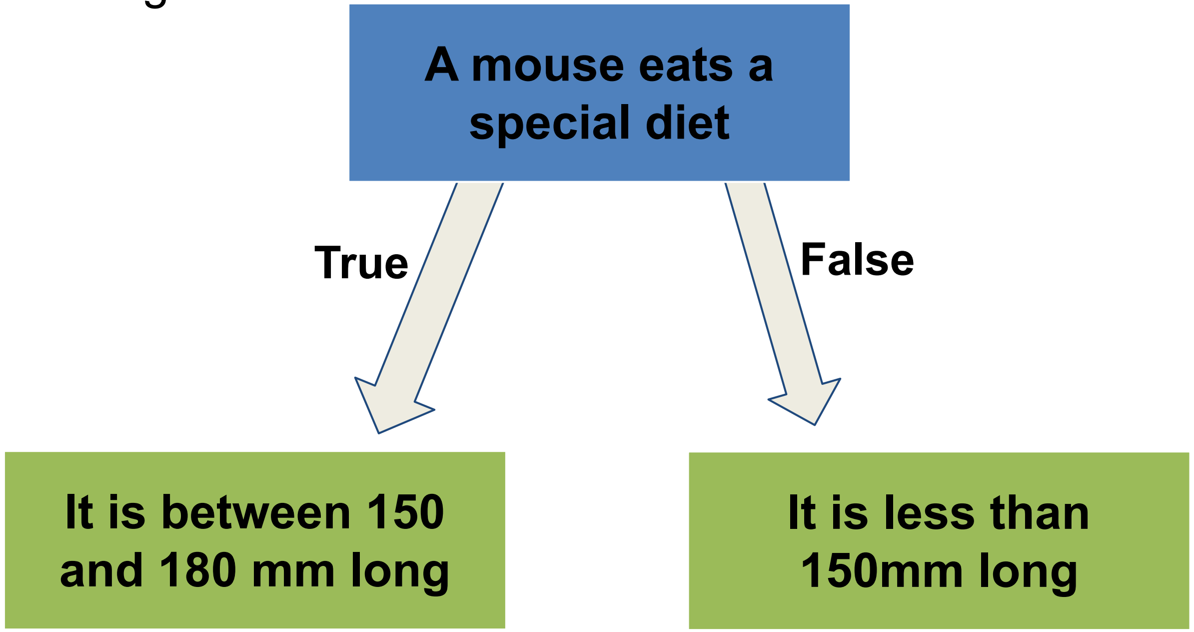
Regression Tree

ConocoPhillips

Basic Decision Tree Concepts

In this case, we are using diet…

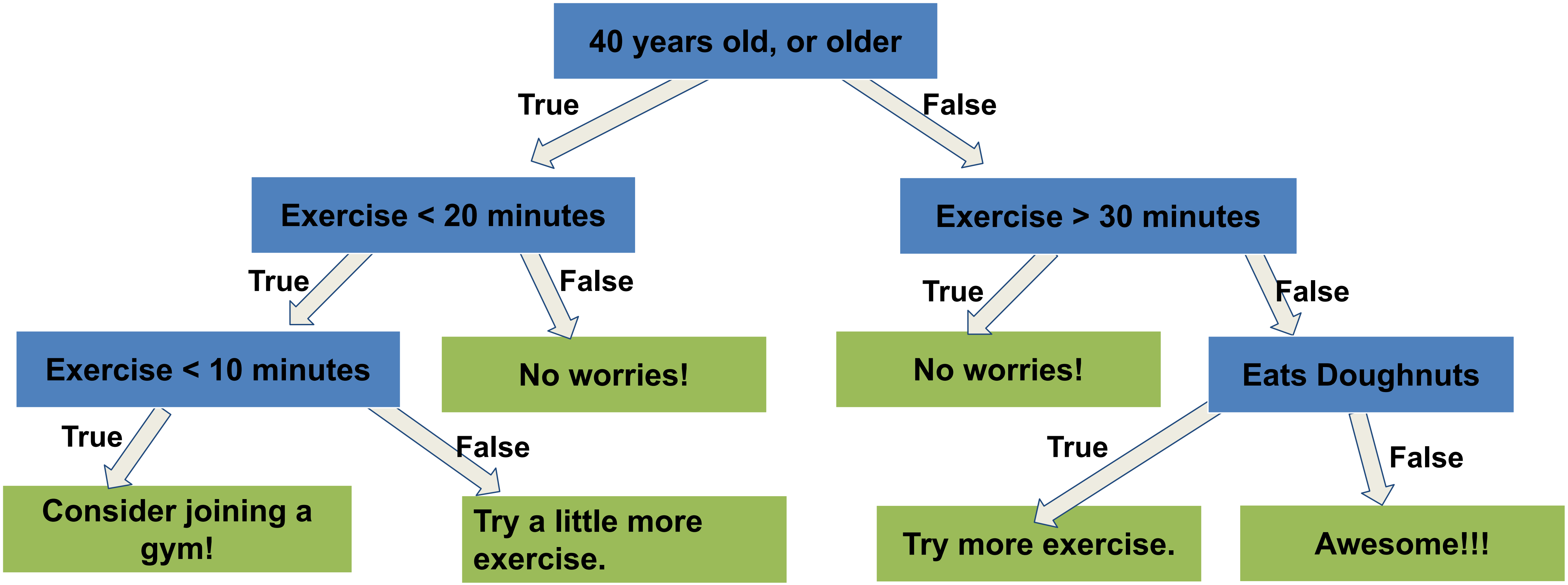
… to predict a numeric value for mouse size



When a Decision Tree predicts numeric values…

Regression Tree

Basic Decision Tree Concepts

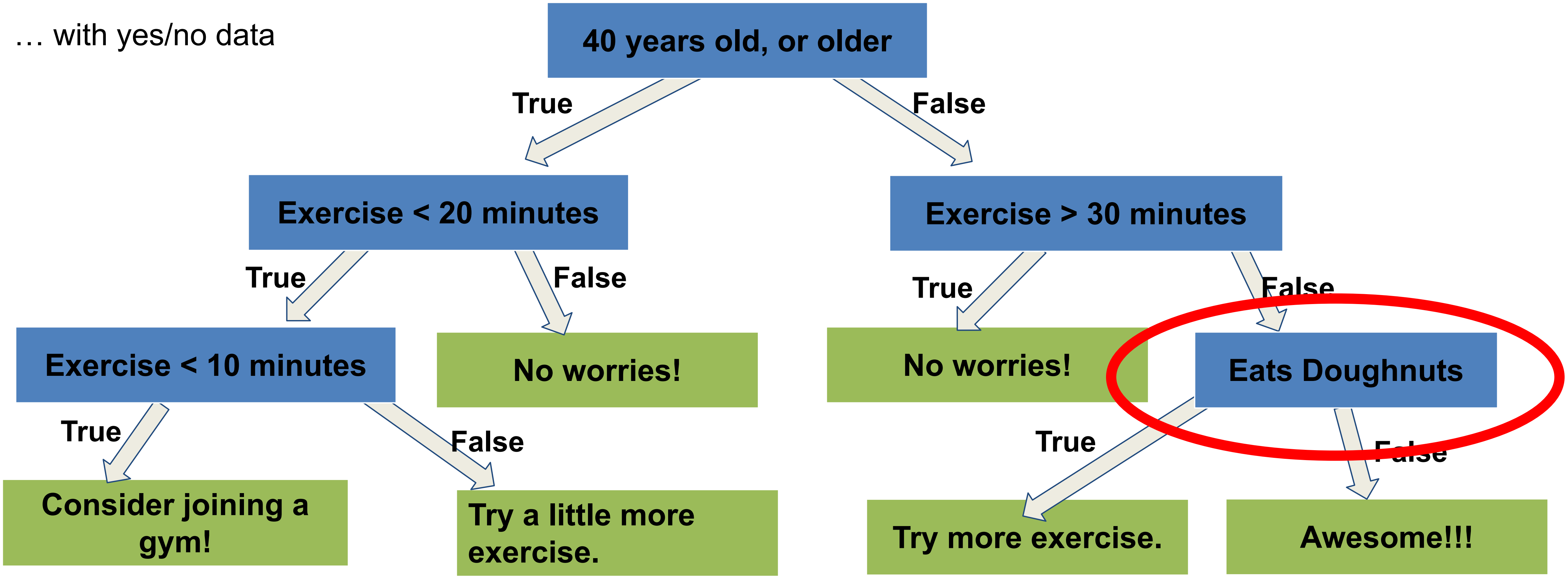


Basic Decision Tree Concepts

It combines numeric data

Basic Decision Tree Concepts

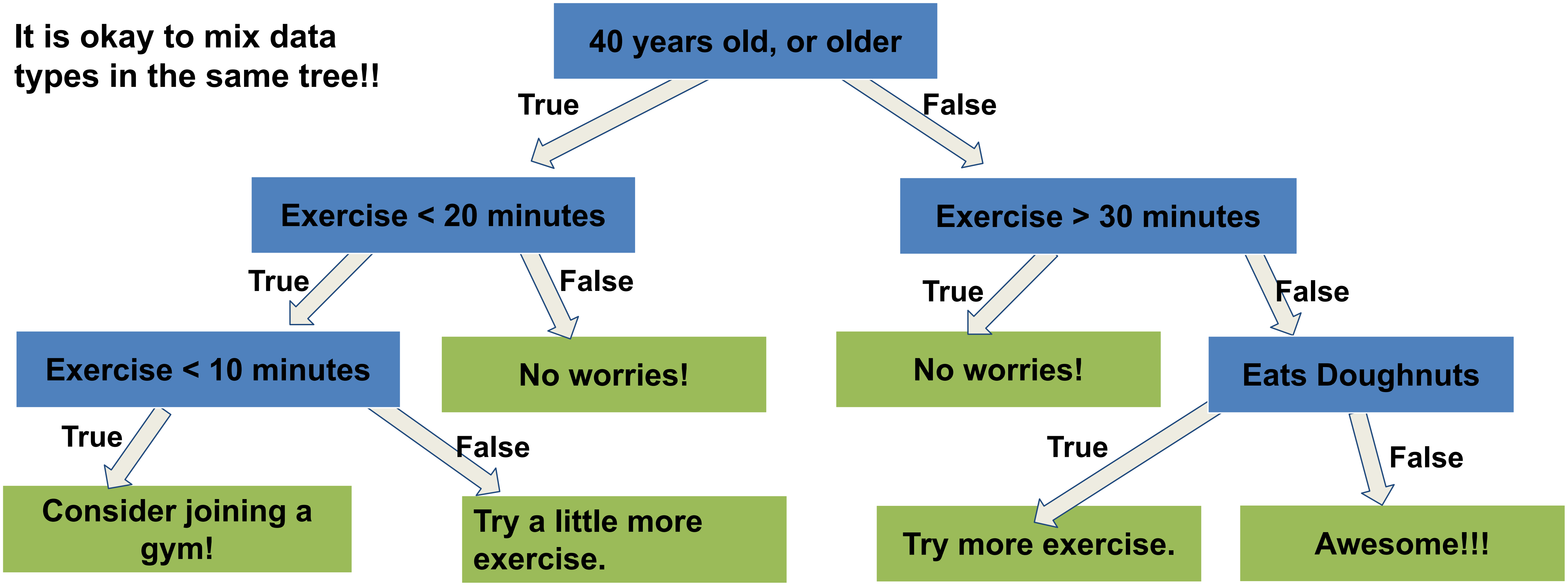
… with yes/no data



ConocoPhillips

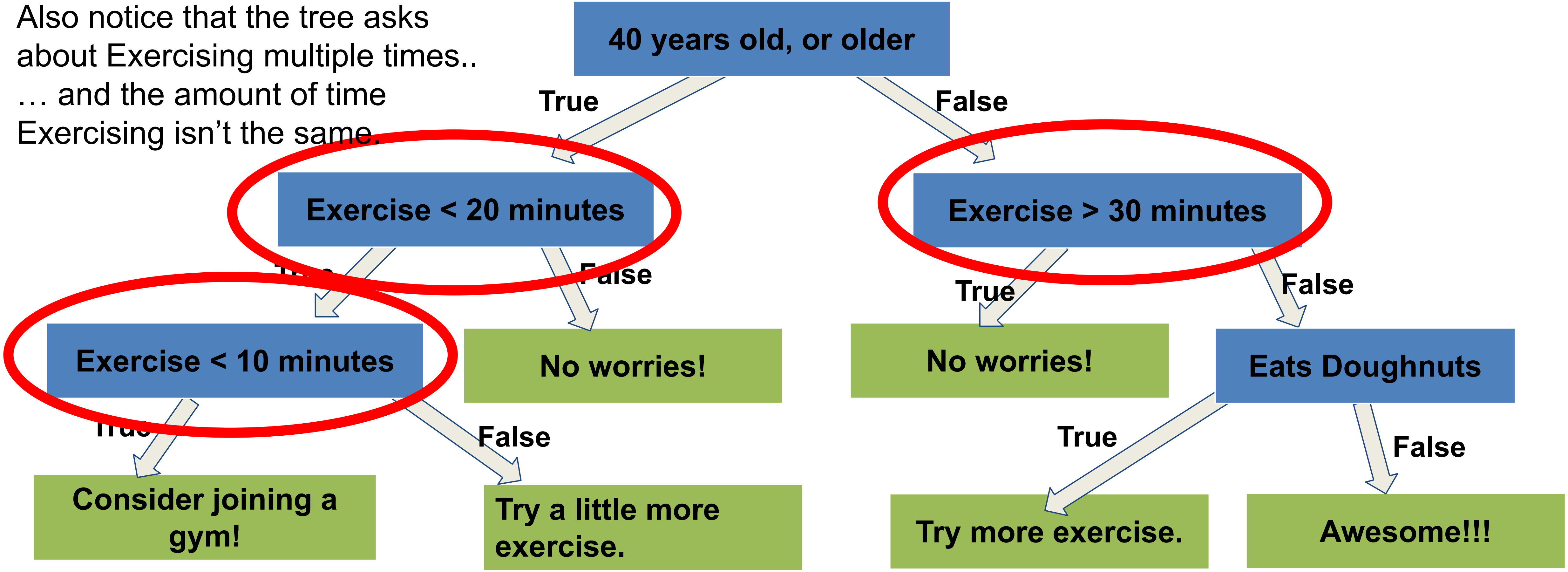
Basic Decision Tree Concepts

It is okay to mix data types in the same tree!!



Basic Decision Tree Concepts

Also notice that the tree asks about Exercising multiple times.. … and the amount of time Exercising isn't the same.

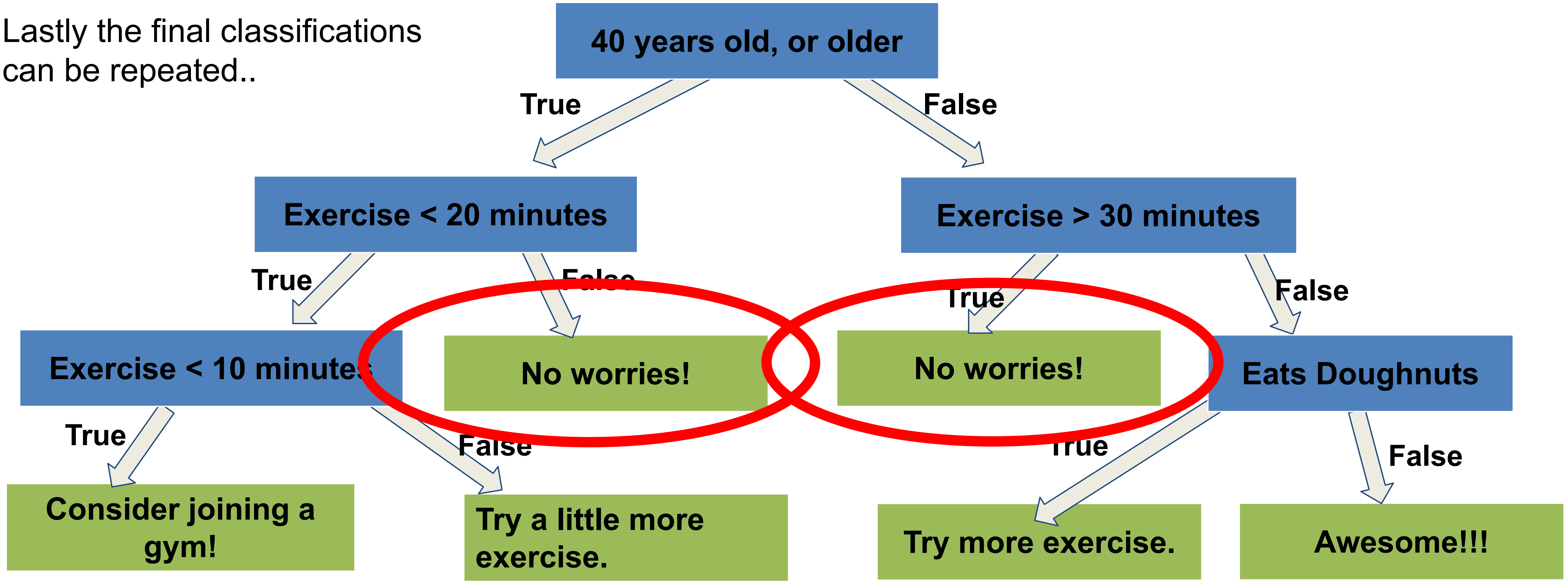


So numeric thresholds can be different for the same data.

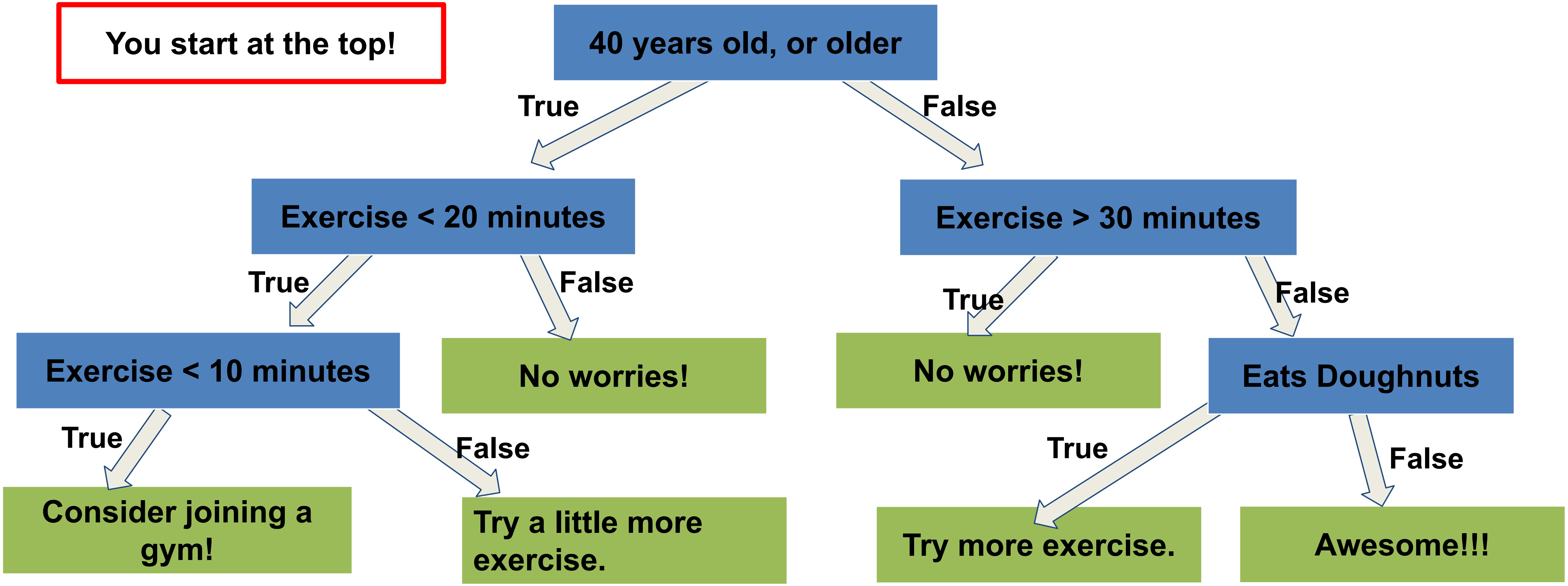
ConocoPhillips

Basic Decision Tree Concepts

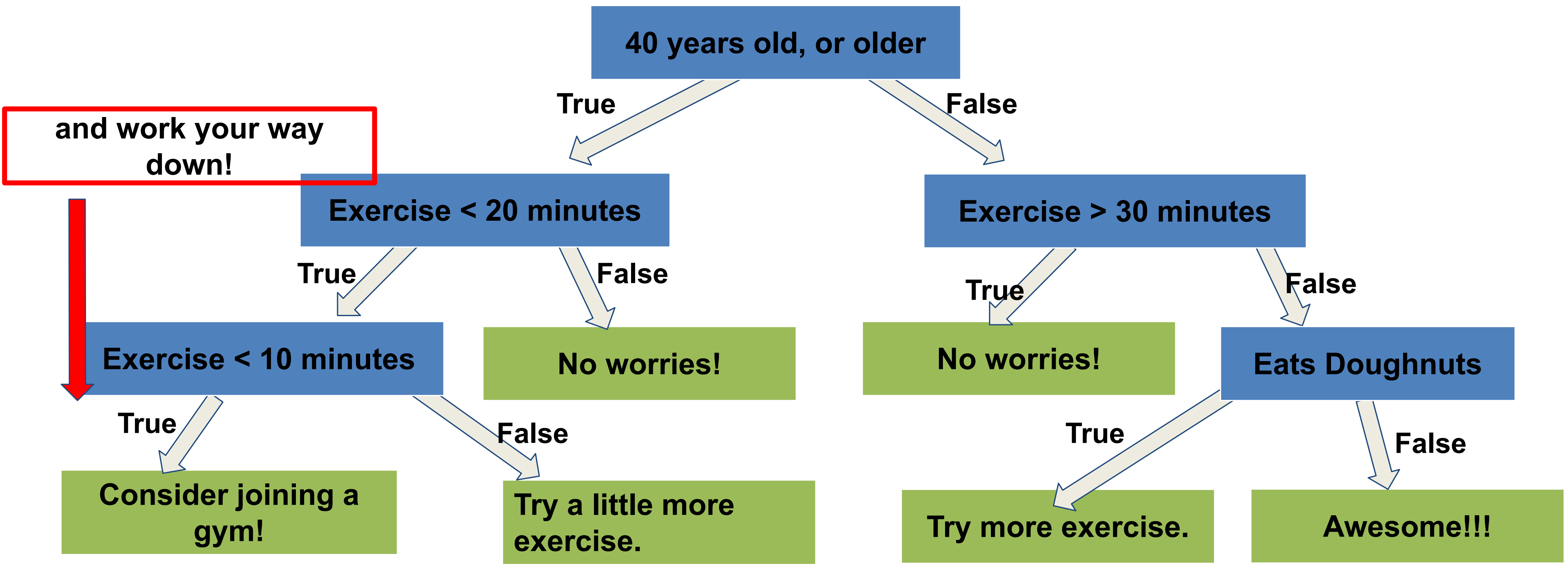
Lastly the final classifications can be repeated..



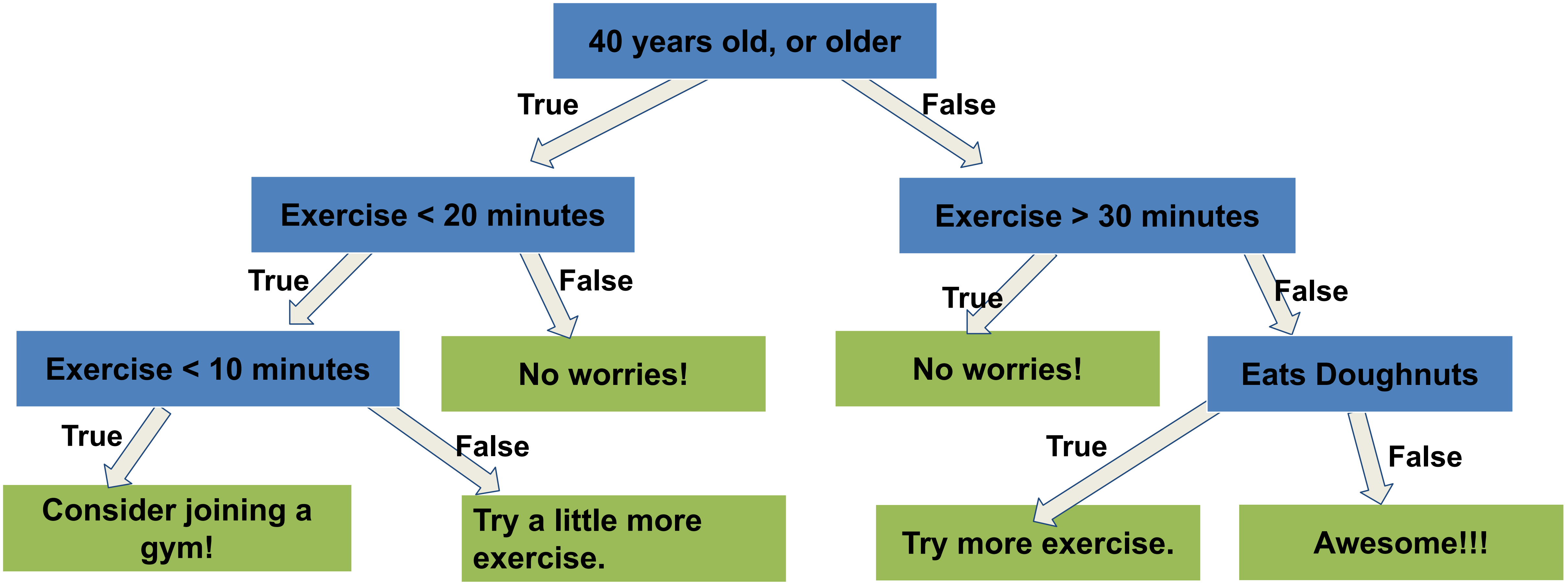
Basic Decision Tree Concepts



Basic Decision Tree Concepts



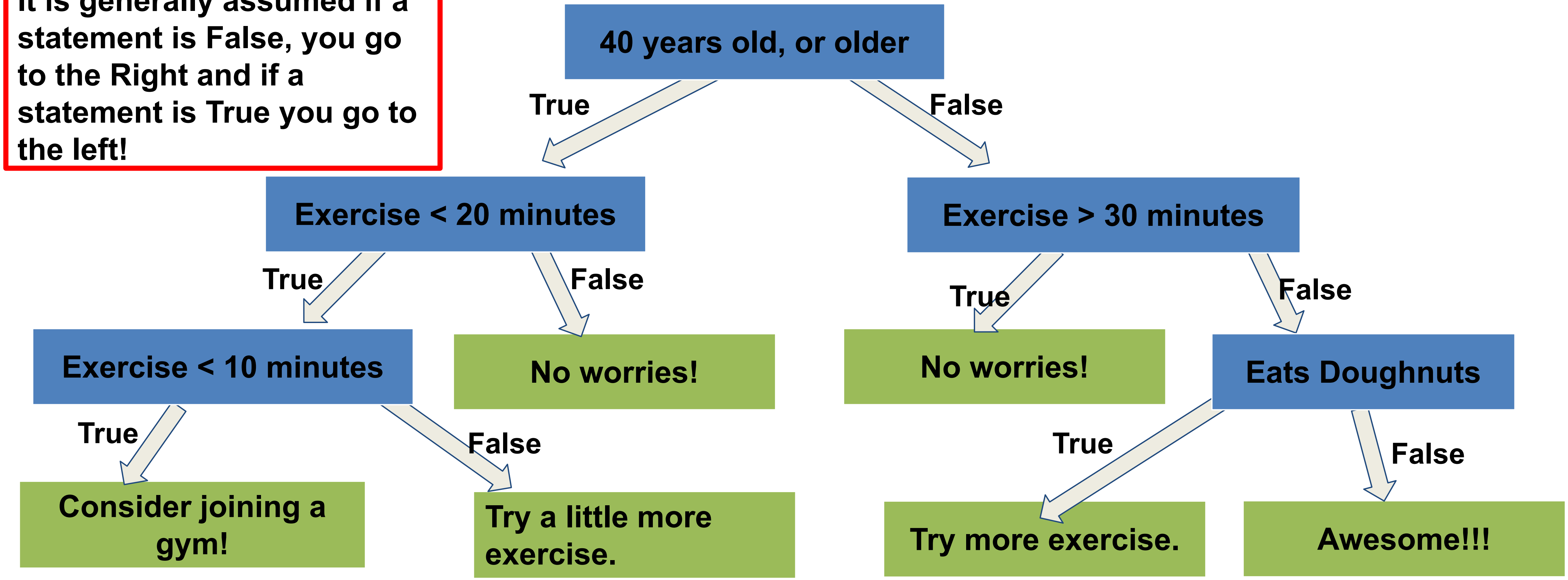
Basic Decision Tree Concepts



… until you get to a point where you can't go any further, and that's how you'll classify something!

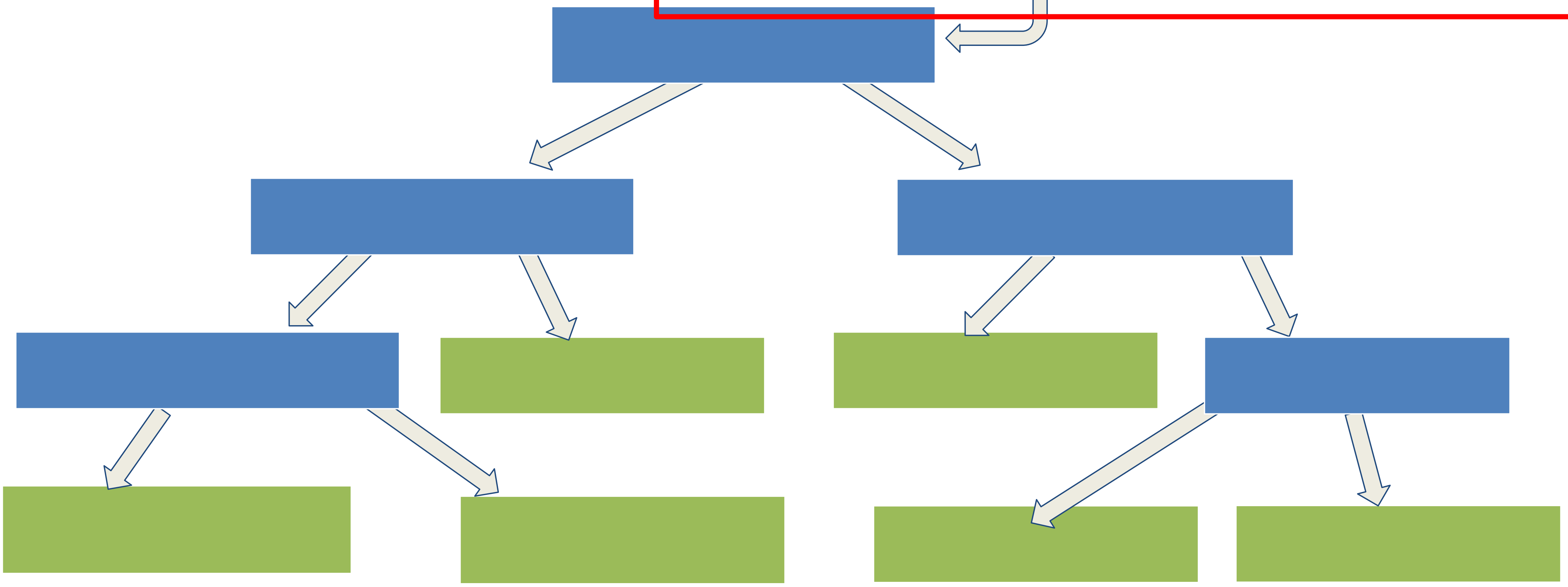
Basic Decision Tree Concepts

It is generally assumed if a statement is False, you go to the Right and if a statement is True you go to the left!



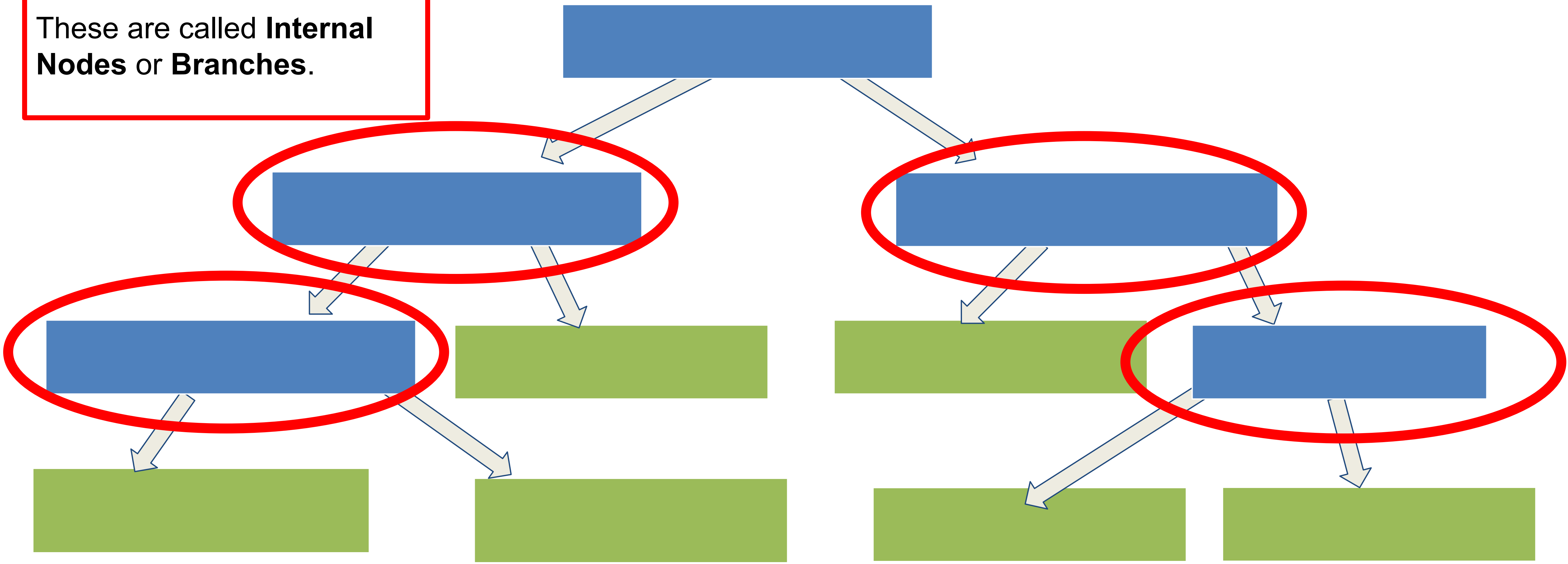
Basic Decision Tree Concepts

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



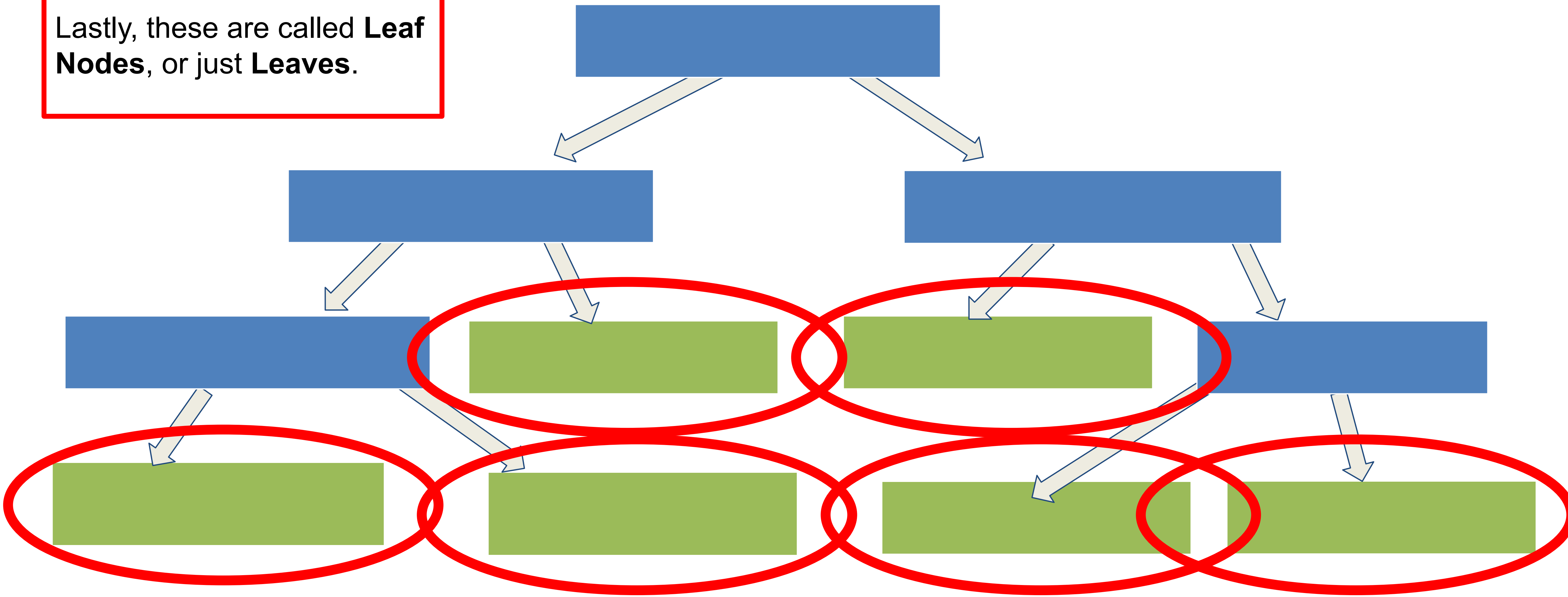
Basic Decision Tree Concepts

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



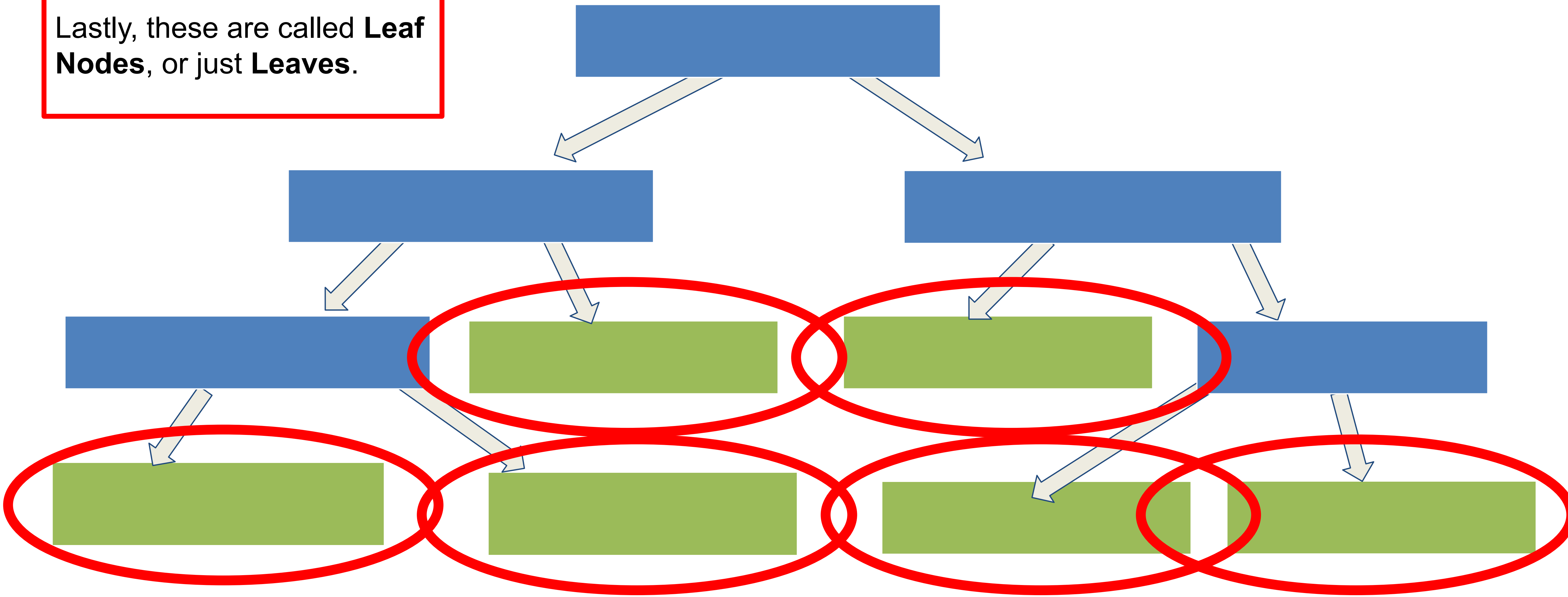
Basic Decision Tree Concepts

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



Basic Decision Tree Concepts

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



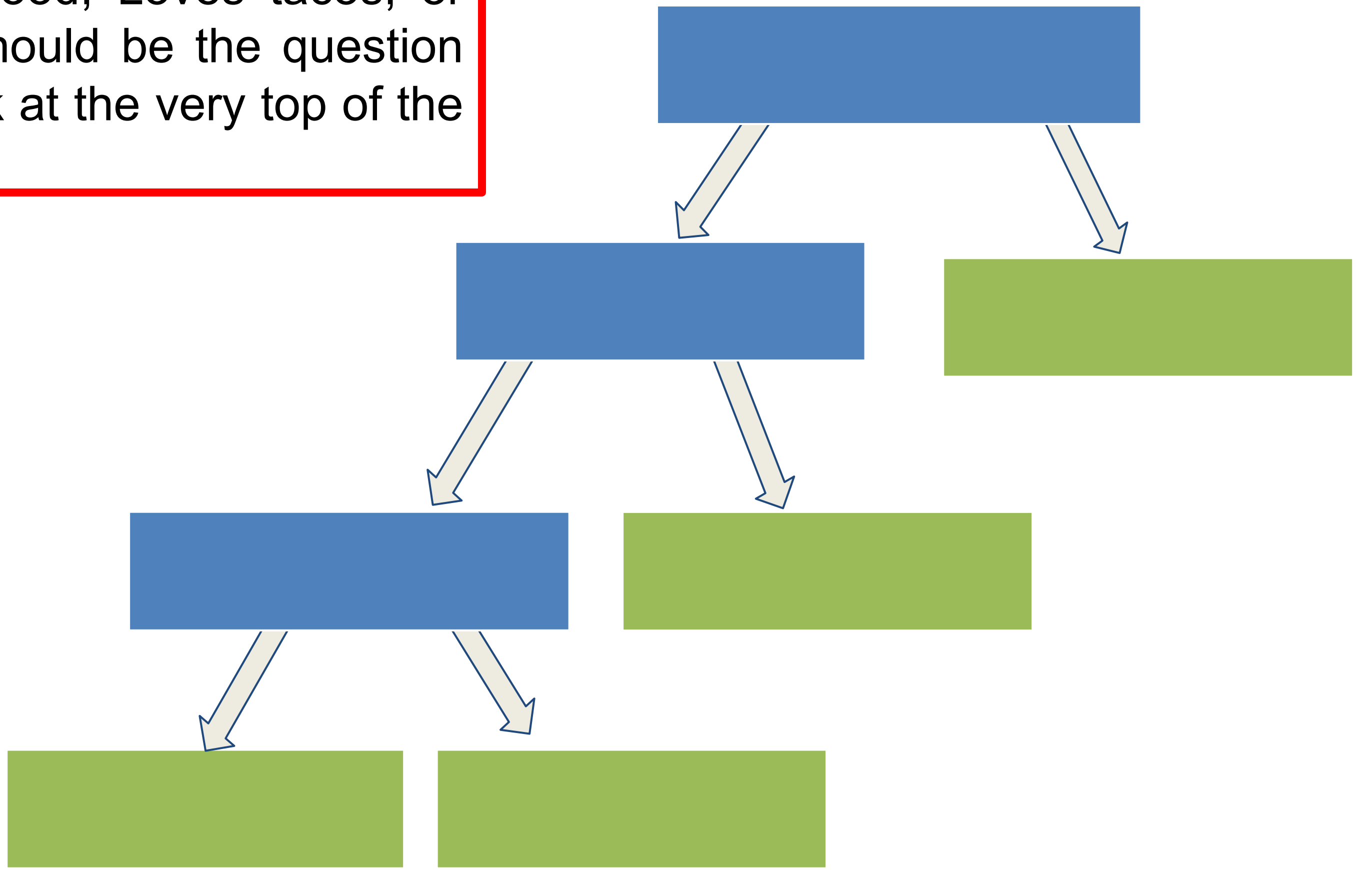
Building a tree with Gini Impurity

Loves spicy food	Loves tacos	Age	Latino
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

Building a tree with Gini Impurity

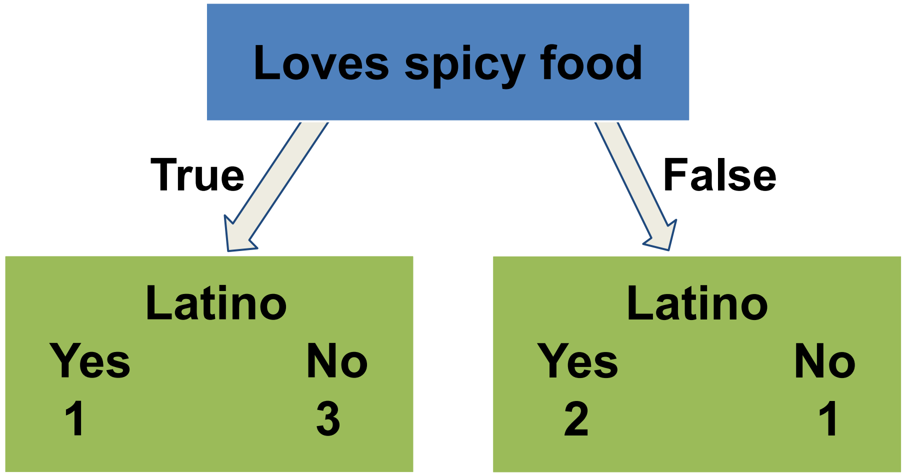
Loves spicy food	Loves tacos	Age	Latino
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

The first thing we do is decide is whether Loves spicy food, Loves tacos, or Age should be the question we ask at the very top of the tree



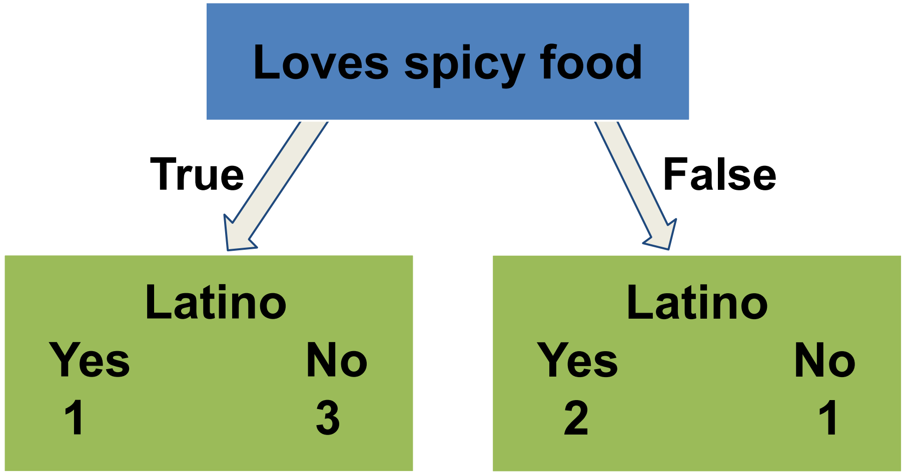
Building a tree with Gini Impurity

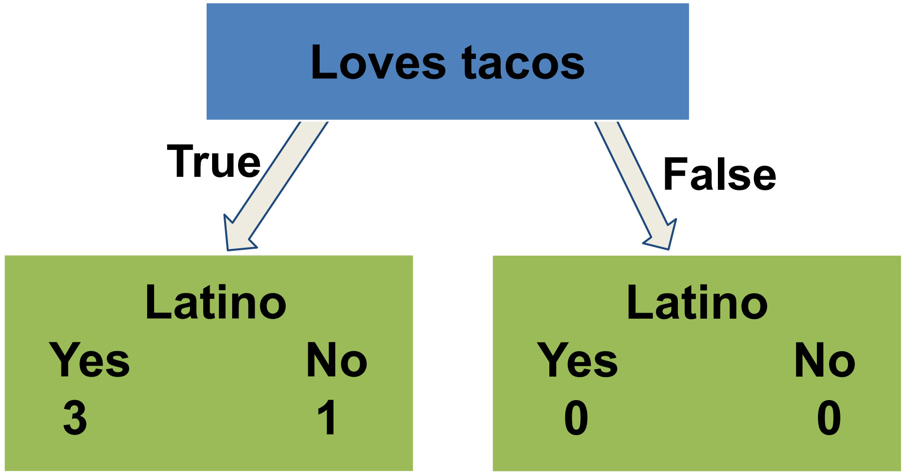
Loves spicy food	Loves tacos	Age	Latino
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



Building a tree with Gini Impurity

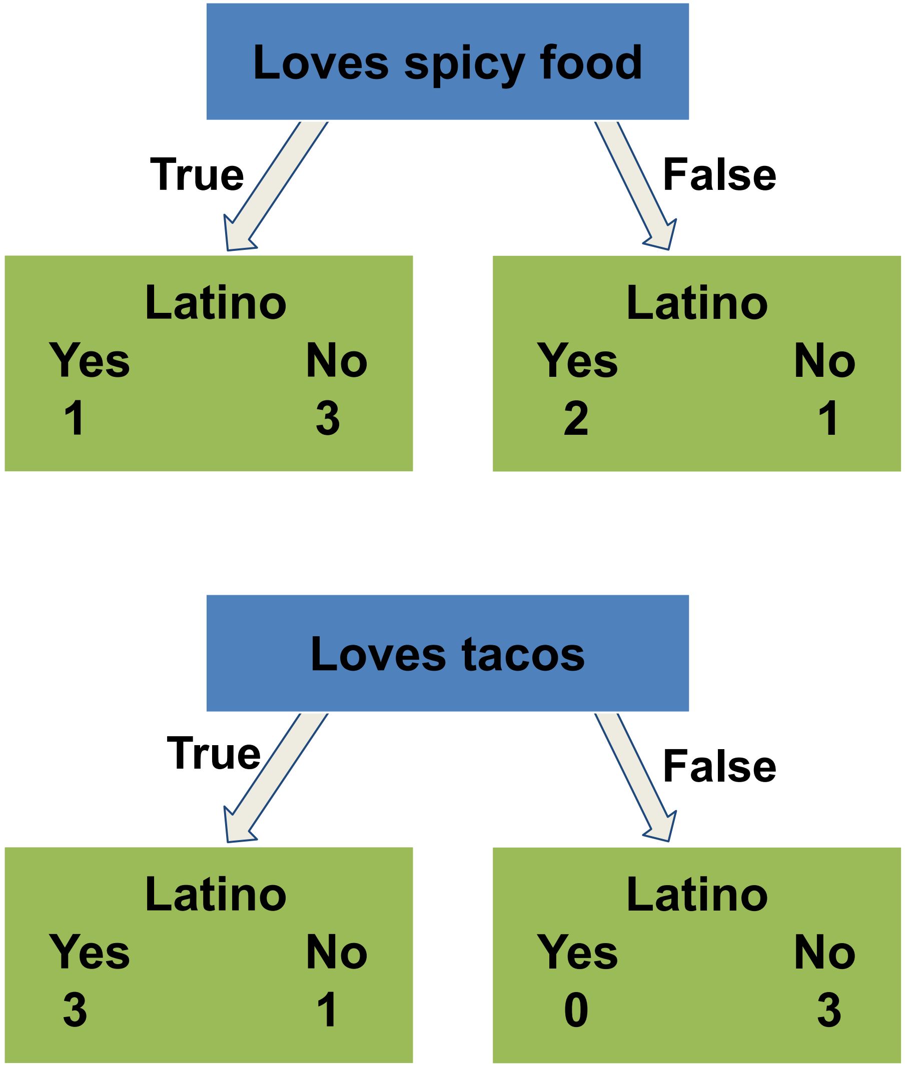
Loves spicy food	Loves tacos	Age	Latino
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No





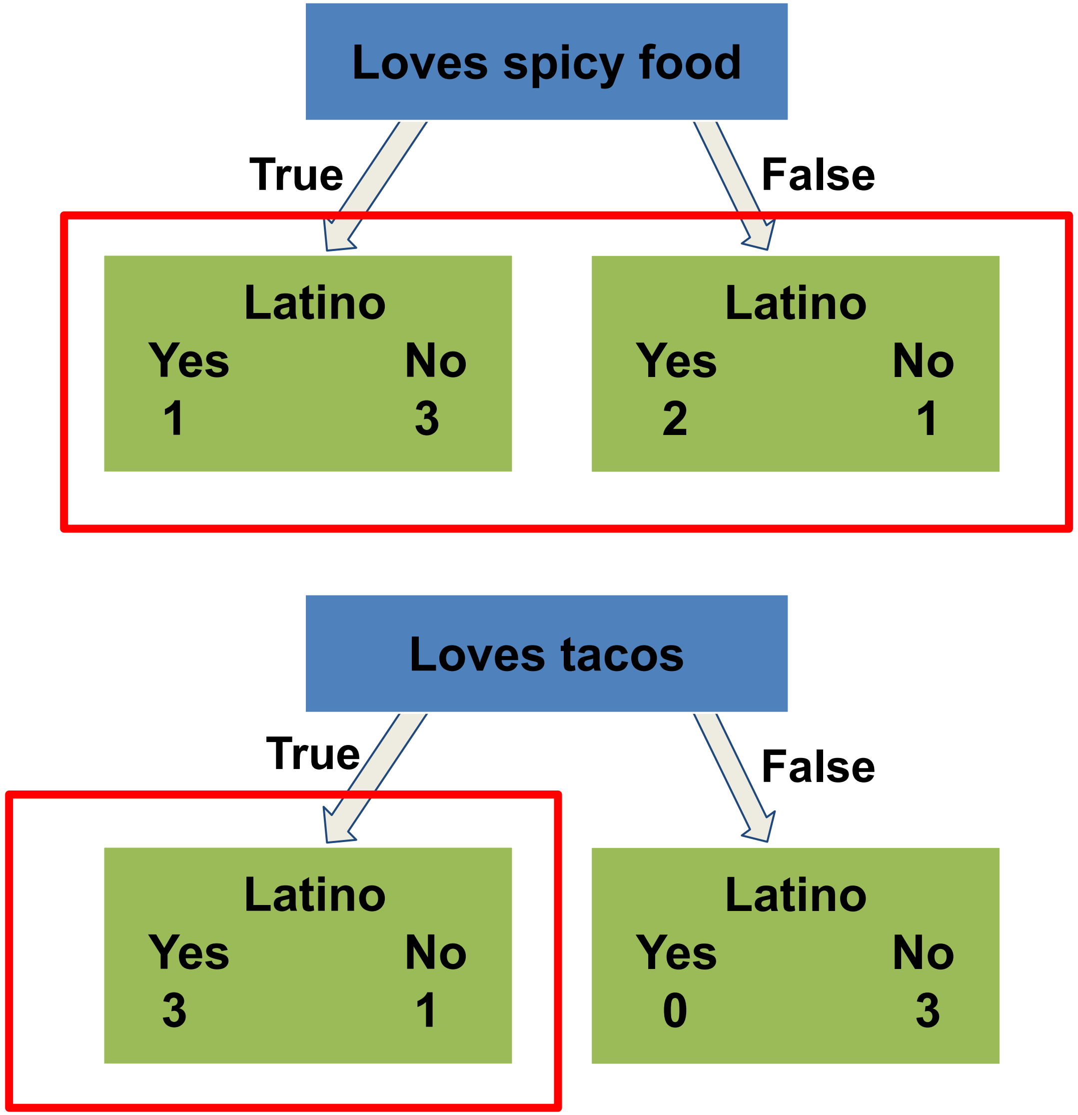
Building a tree with Gini Impurity

Looking at the two little trees we see that neither one does a predicting who will and who will not be **Latino**.



Building a tree with Gini Impurity

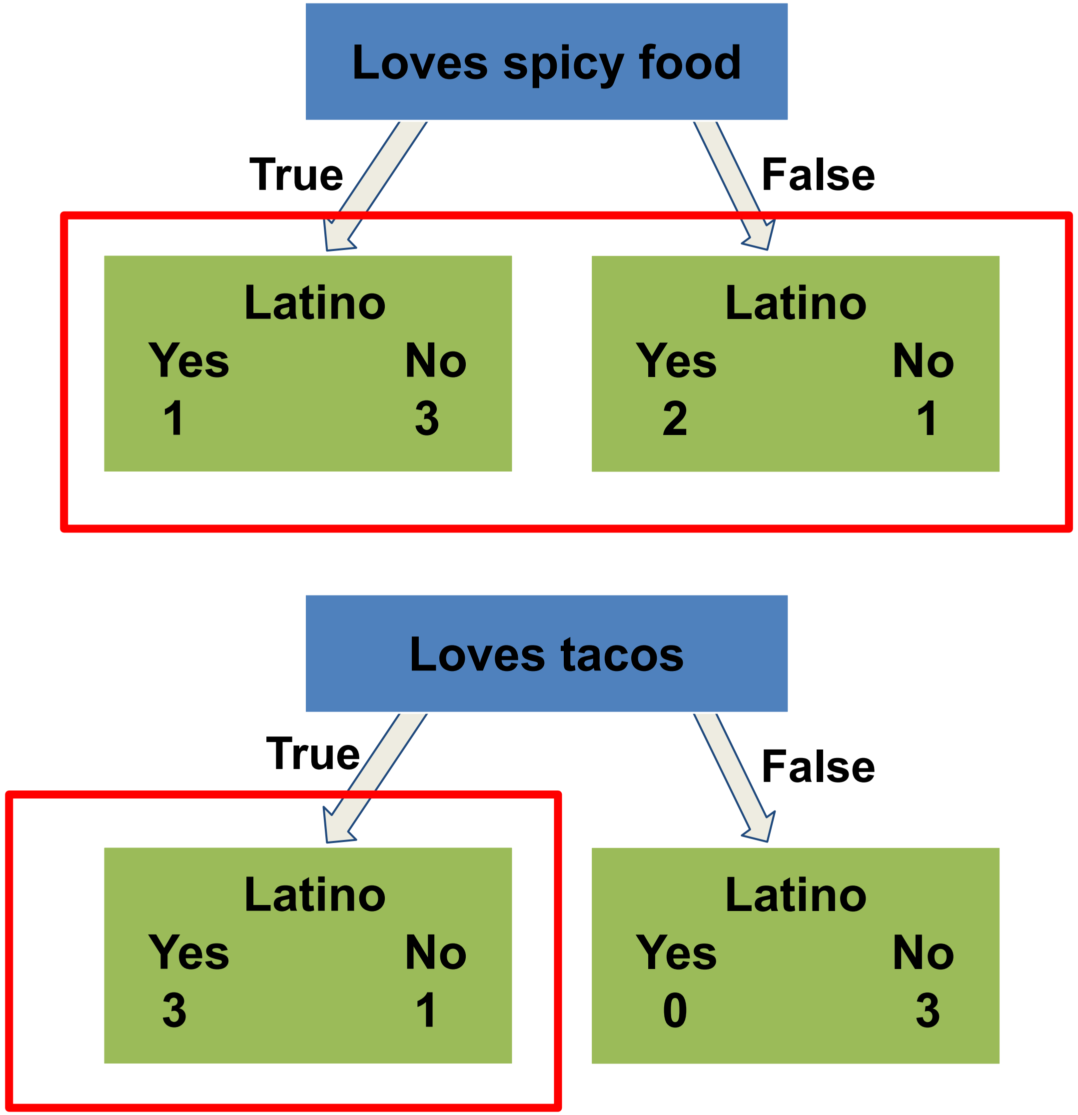
These three **Leaves** contain mixtures of people that are and are not latinos.



Building a tree with Gini Impurity

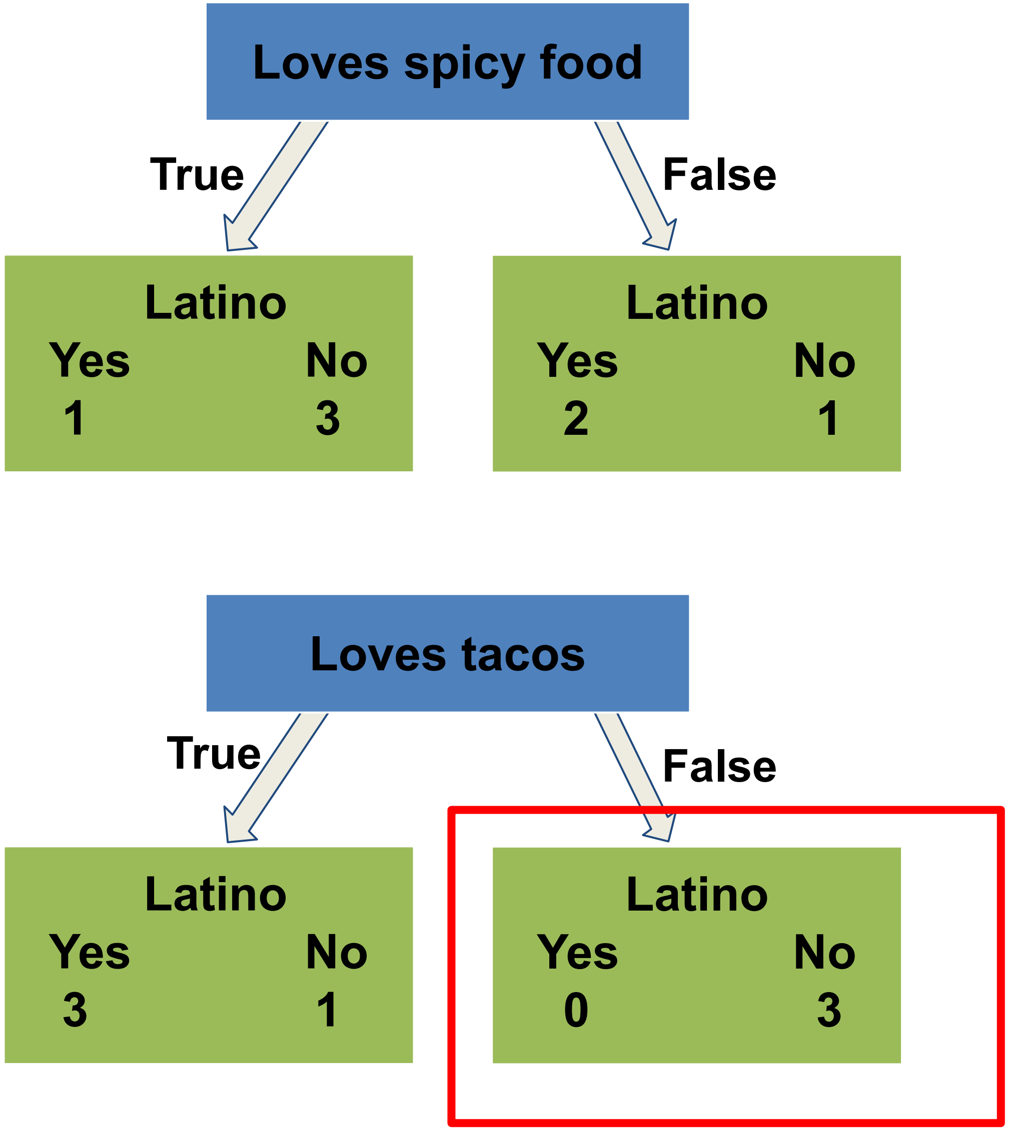
These three **Leaves** contain mixtures of people that are and are not latinos.

Impure leaves



Building a tree with Gini Impurity

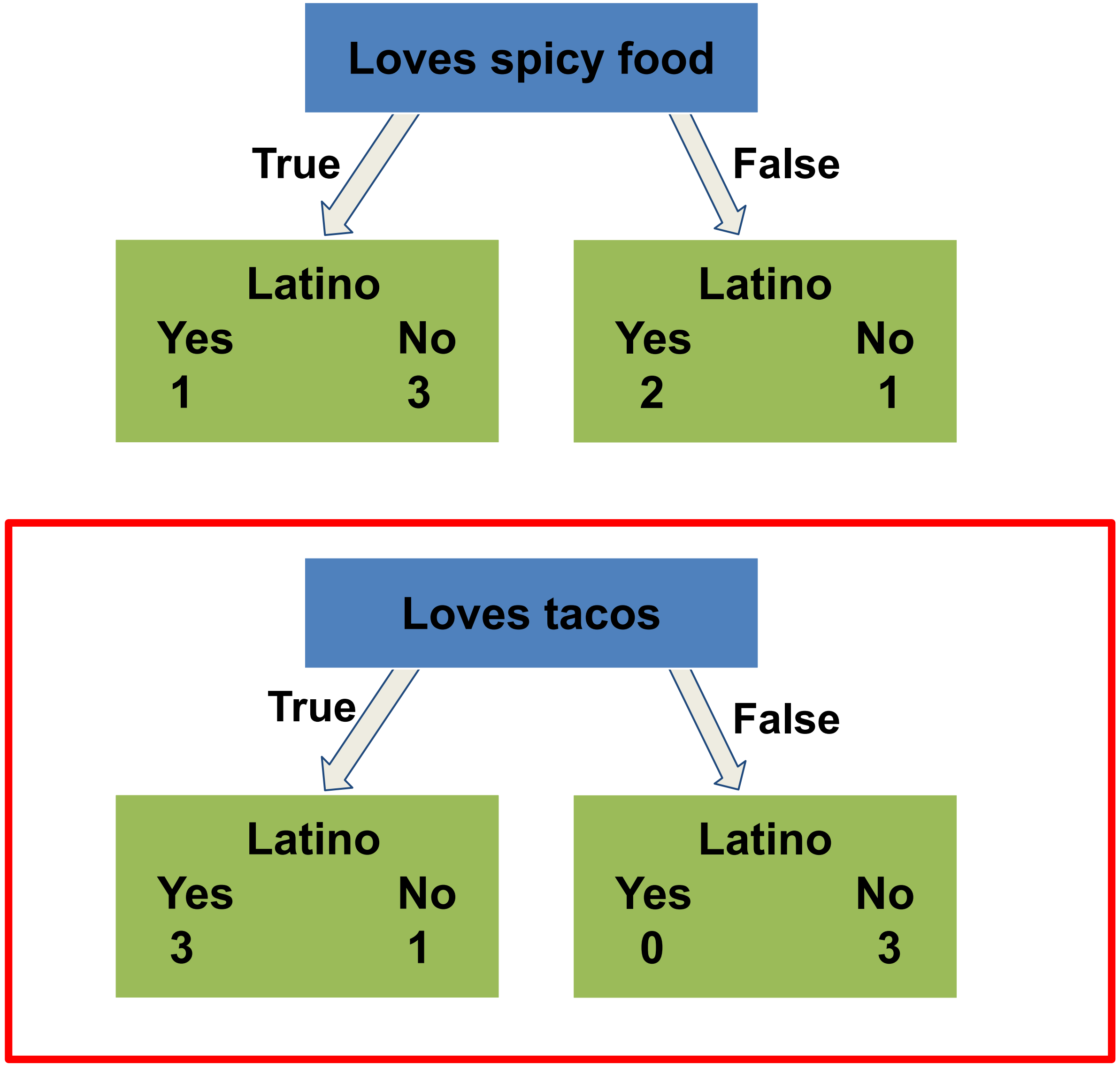
This Leaf only contain people who is not Latino



Building a tree with Gini Impurity

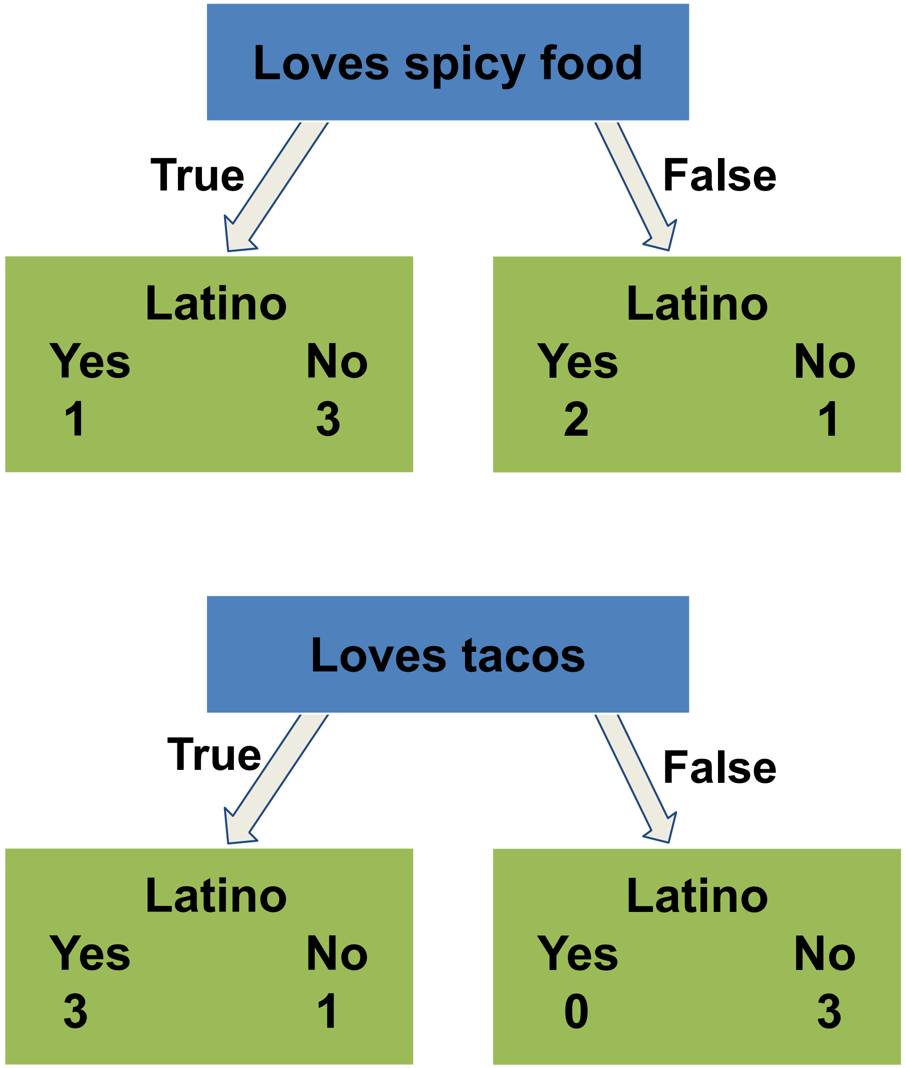
Only one impure leaf!

It seems like **Loves tacos** does a better job predicting who is Latino.



Building a tree with Gini Impurity

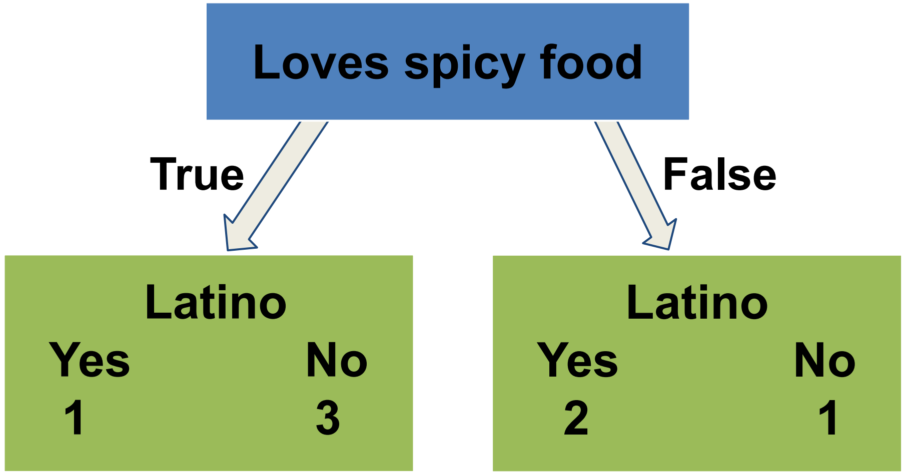
How can we quantify the **impurity** of the leaves?

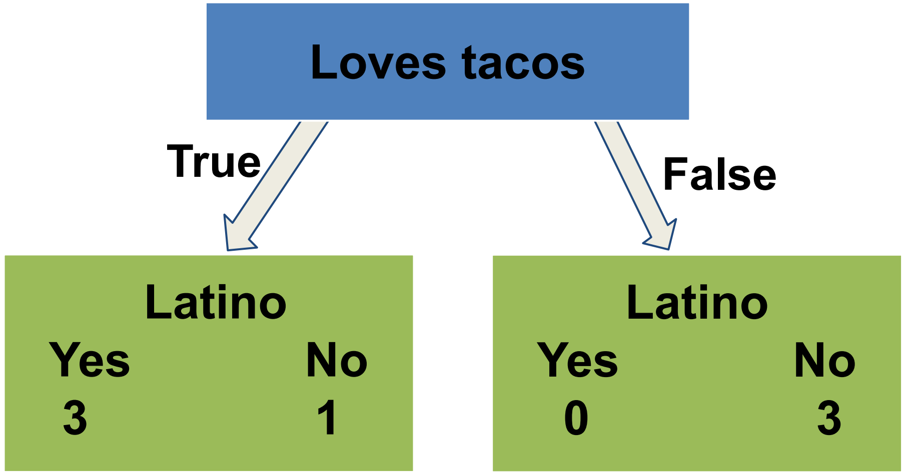


Building a tree with Gini Impurity

How can we quantify the **impurity** of the leaves?

- Gini Impurity
- Entropy
- Information Gain

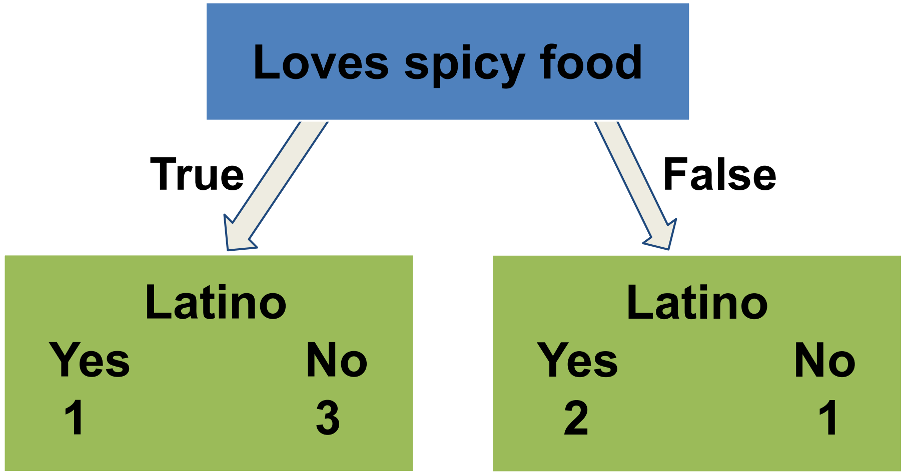




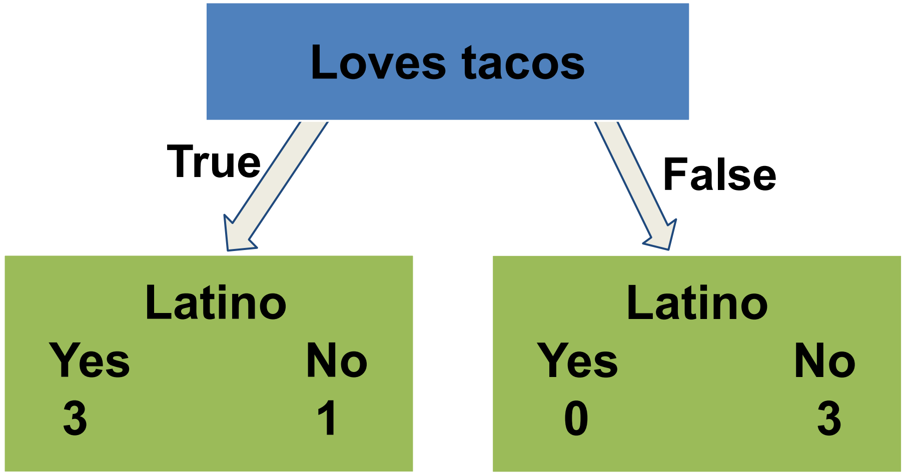
Building a tree with Gini Impurity

Gini Impurity for loves spicy food

Step 1: Calculate gini impurity for individual leaves.
Gini(tacos) = 1 - sum((precision)^2)=1-sum(75%^2)=



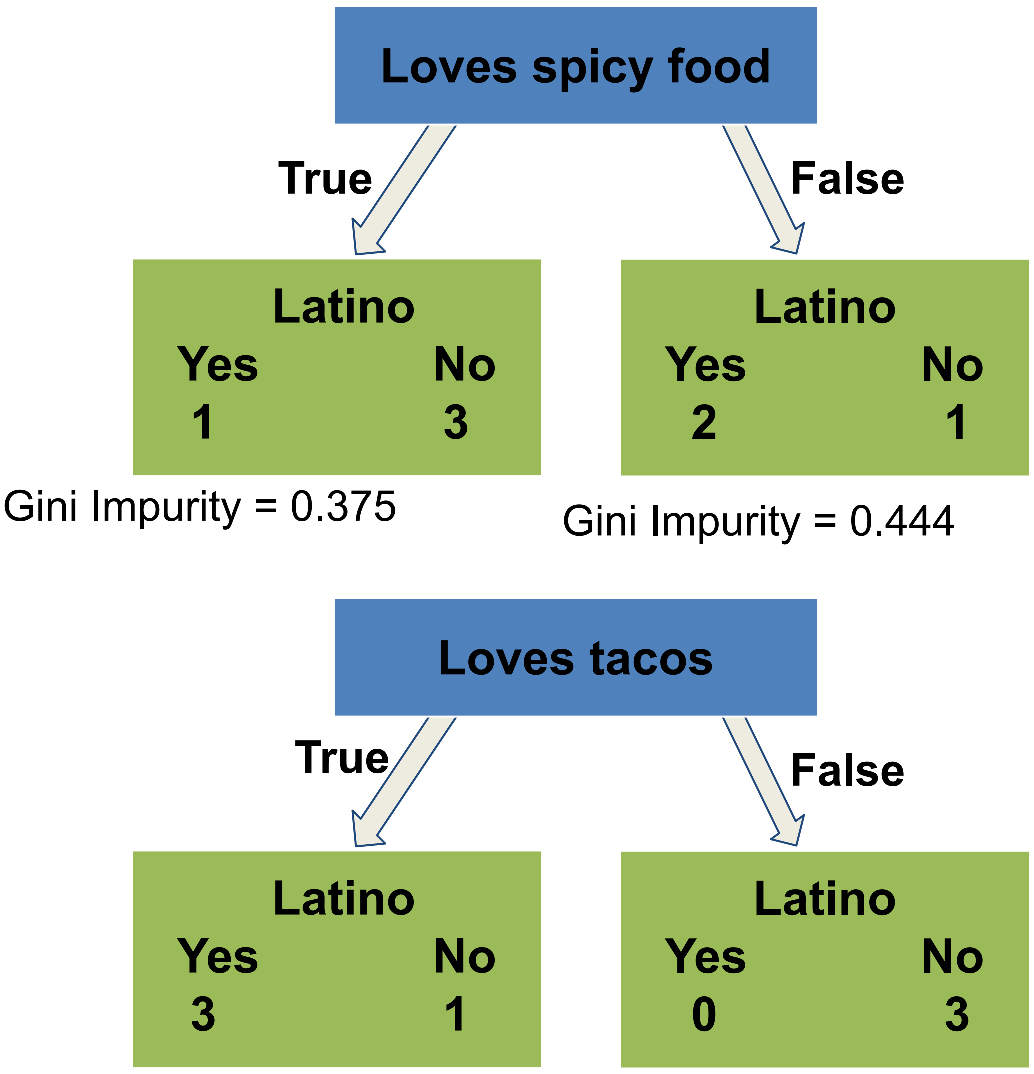
Gini Impurity = 0.375



Numeric and Continuous Variables

Step 1: Calculate gini impurity for individual leaves.
Step 2: Calculate the Total Gini Impurities

Total Gini Impurity = Weighted average of **Gini Impurities** for the **Leaves**
=(4/ (4+3))0.375 + (3/(4+3))0.444 = 0.405

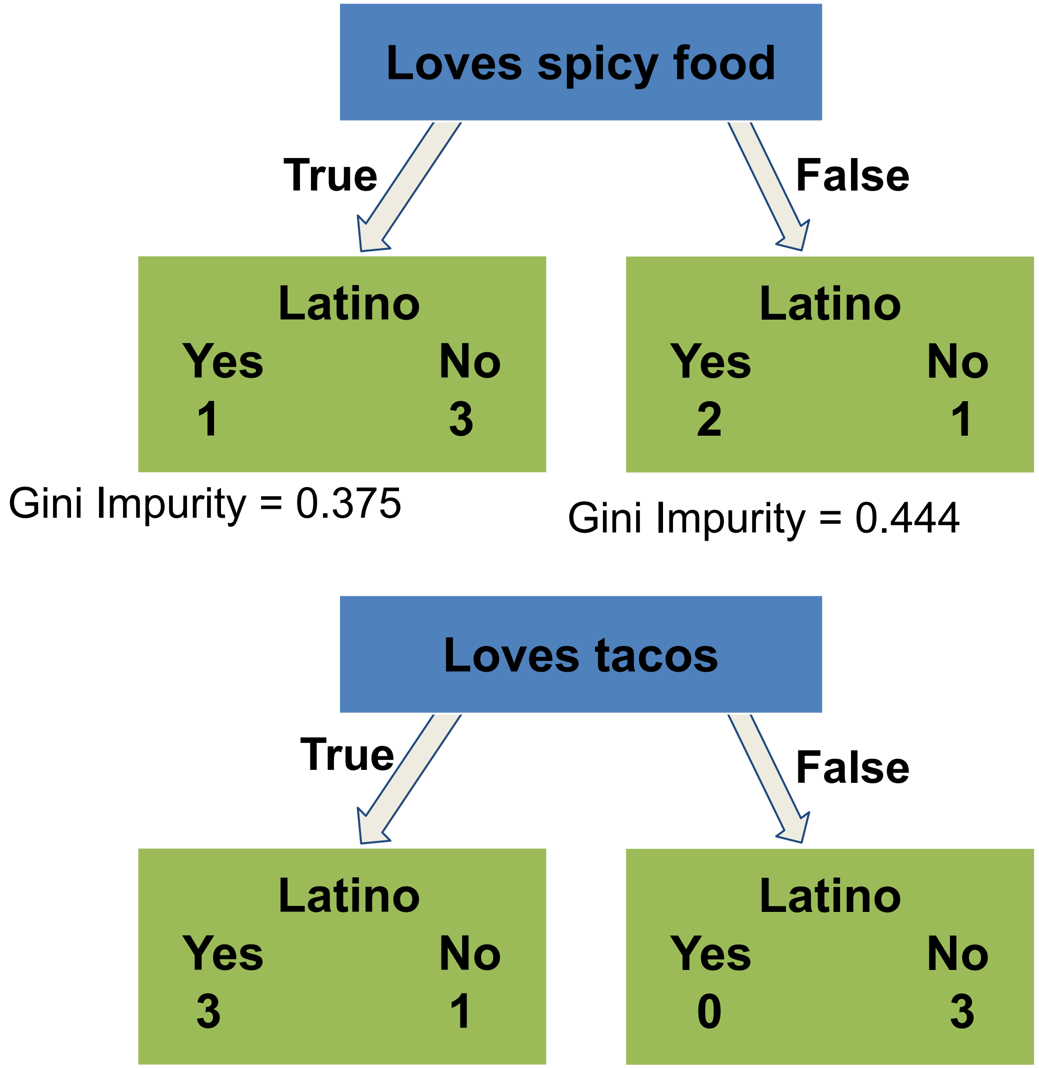


ConocoPhillips

Numeric and Continuous Variables

Step 1: Calculate gini impurity for individual leaves.
Step 2: Calculate the Total Gini Impurities

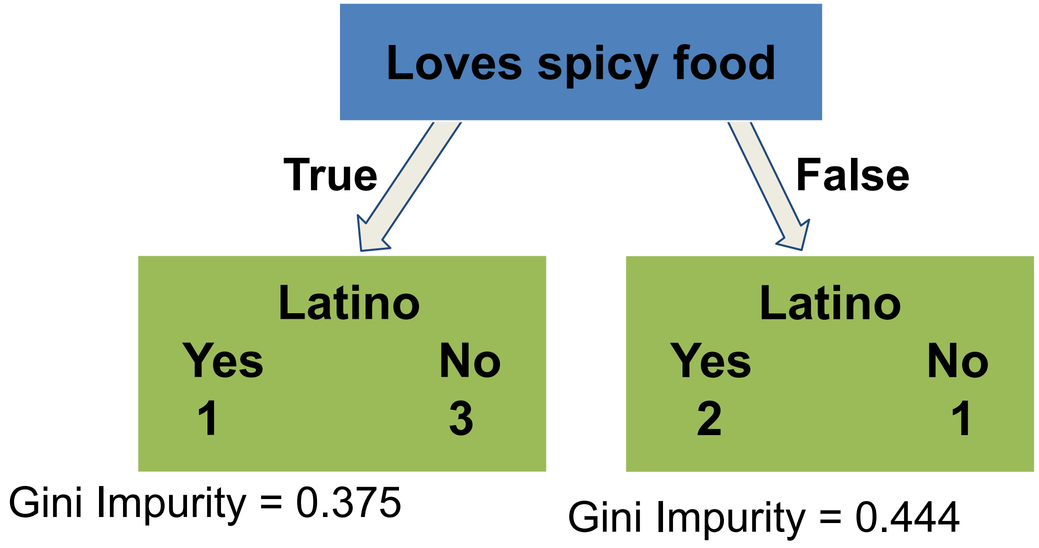
Gini Impurity for loves spicy food = 0.405



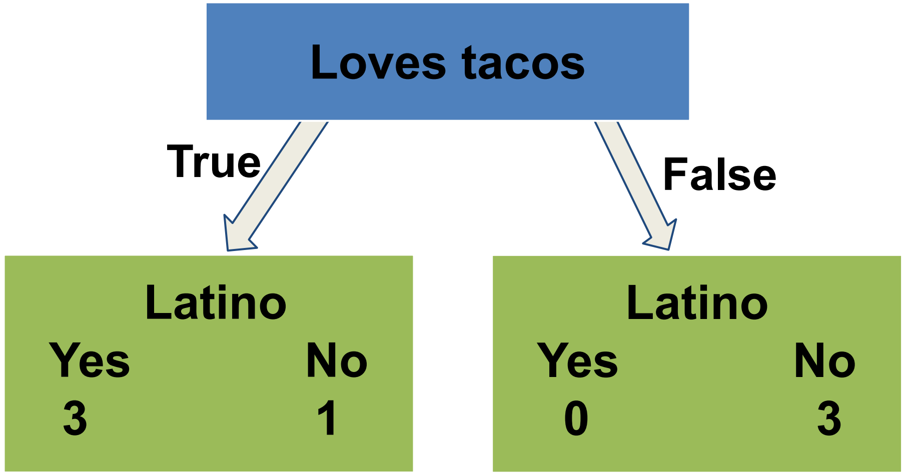
Numeric and Continuous Variables

Step 1: Calculate gini impurity for individual leaves.
Step 2: Calculate the Total Gini Impurities

Gini Impurity for loves spicy food = 0.405



Gini Impurity for loves tacos = 0.214



Numeric and Continuous Variables

Categorical Variable

Loves spicy food	Loves tacos	Age	Latino
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

Step 1: The first thing we do is sort the rows by **Age,** from lowest value to highest value.

Numeric and Continuous Variables

Categorical Variable

Loves spicy food	Loves tacos	Age	Latino
Yes	Yes	7	No
		9.5	
Yes	No	12	No
		15	
No	Yes	18	Yes
		26.5	
No	Yes	35	Yes
		36.5	
Yes	Yes	38	Yes
		44	
Yes	No	50	No
		66.5	
No	No	83	No

Step 1: The first thing we do is sort the rows by **Age,** from lowest value to highest value.
Step 2: Then we calculate the average **Age** for all adjacent people.

Numeric and Continuous Variables

Categorical Variable

Loves spicy food	Loves tacos	Age	Latino
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

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

Step 1: The first thing we do is sort the rows by **Age,** from lowest value to highest value.
Step 2: Then we calculate the average **Age** for all adjacent people.
Step 3: Calculate the Gini Impurity for each average age.

Numeric and Continuous Variables

Categorical Variable

Loves spicy food	Loves tacos	Age	Latino
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

9.5 → Gini Impurity = 0.429

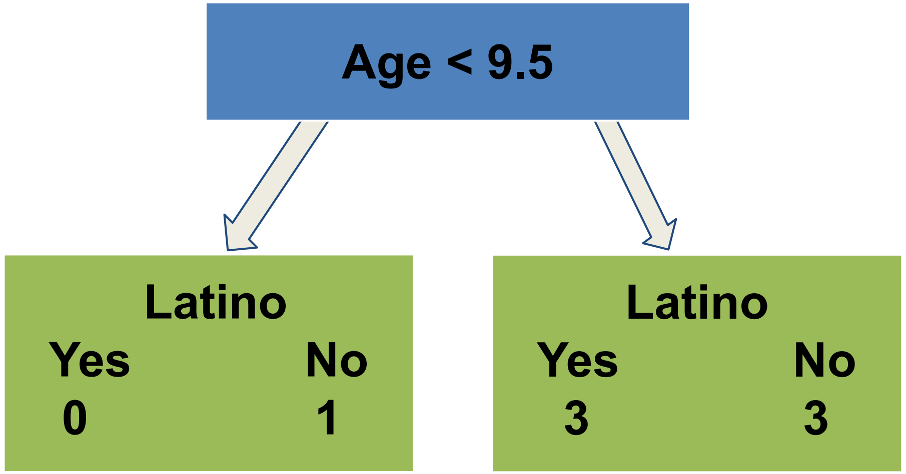
15 → Gini Impurity = 0.343

26.5 → Gini Impurity = 0.476

36.5 → Gini Impurity = 0.476

44 → Gini Impurity = 0.343

66.5 → Gini Impurity = 0.429



Gini Impurity = 0

Gini Impurity = 0.5

Numeric and Continuous Variables

Categorical Variable

Loves spicy food	Loves tacos	Age	Latino
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

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

Age < 9.5

Latino	
Yes	**No**
0	**1**

Gini Impurity = 0

Latino	
Yes	**No**
3	**3**

Gini Impurity = 0.5

Total Gini Impurity = (1/(1+6))*0 + (1/(1+6))*0.5 = 0.429

Numeric and Continuous Variables

Continuous Variable

Loves spicy food	Loves tacos	Age	Latino
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

- 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

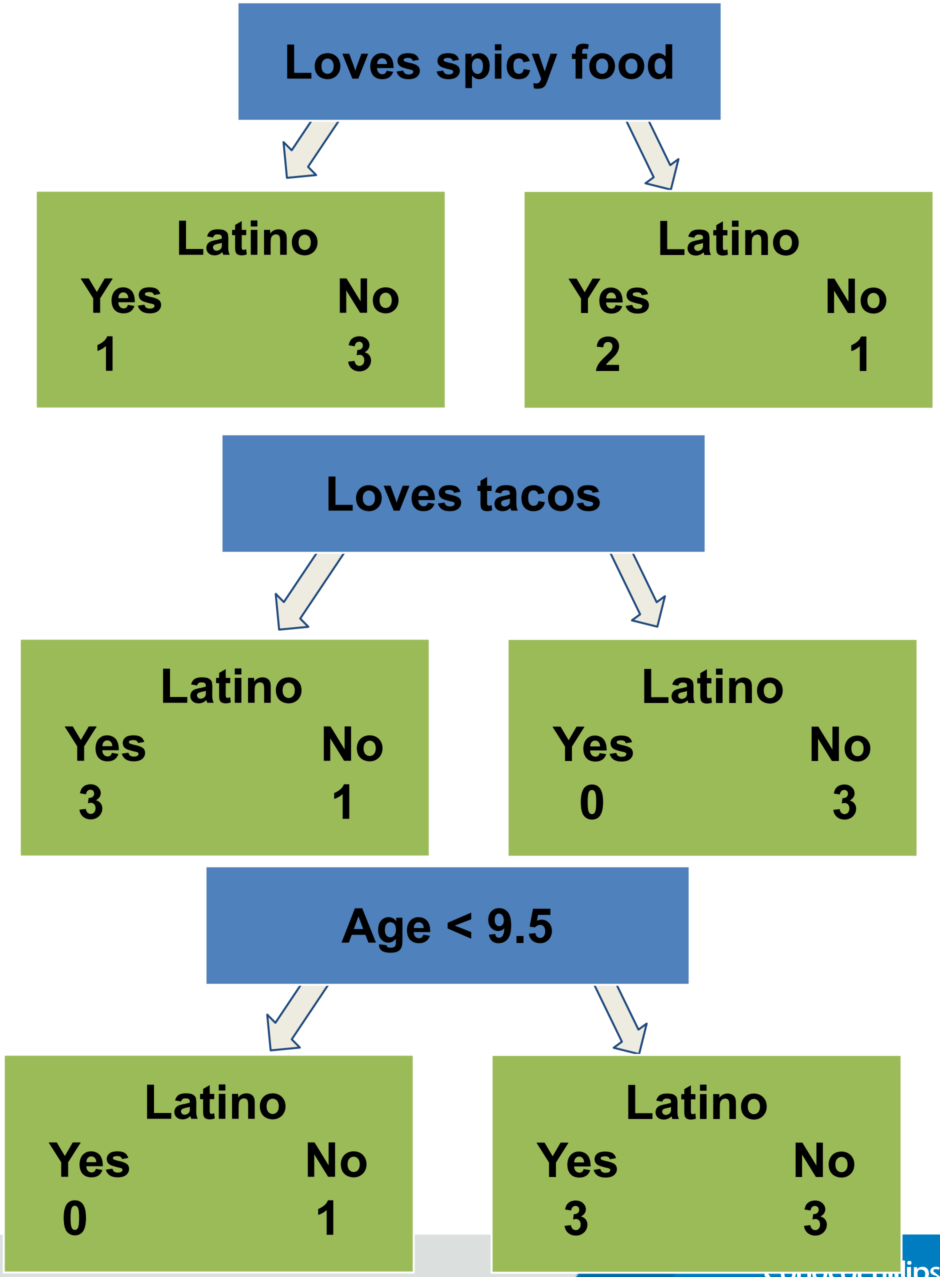
Numeric and Continuous Variables

Gini Impurity for loves spicy food = 0.405

… we know that its Leaves have the lowest Impurity…

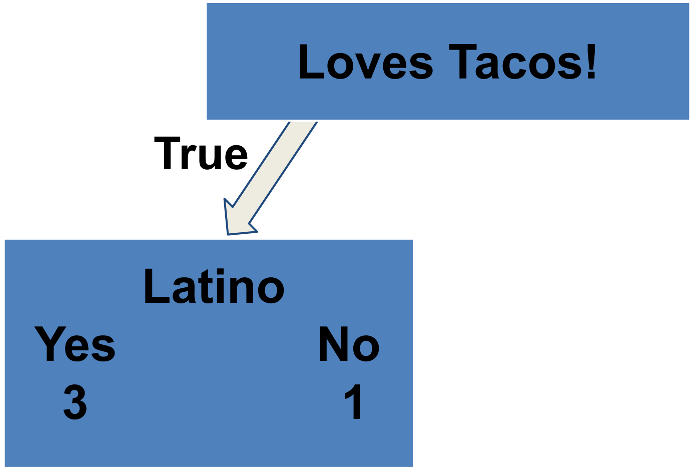
Gini Impurity for loves tacos = 0.214

Gini Impurity for ages < 15 = 0.343



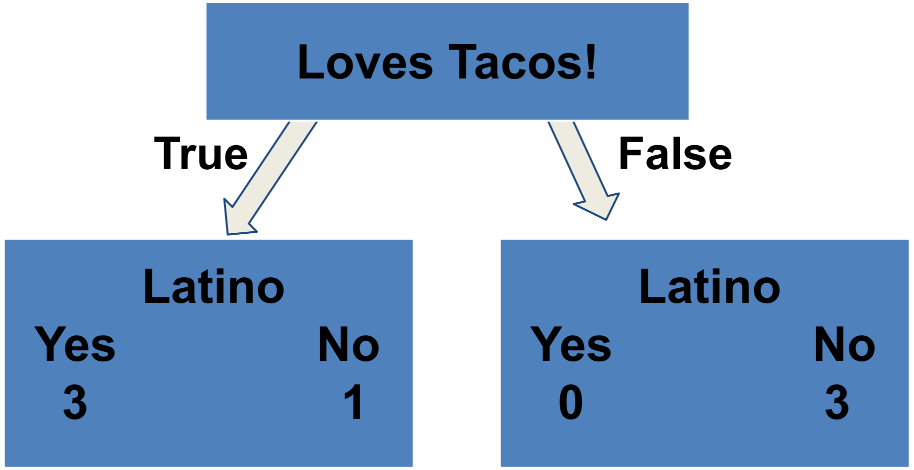
Building a tree with Gini Impurity

Loves spicy food	Loves tacos	Age	Latino
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



Adding Branches

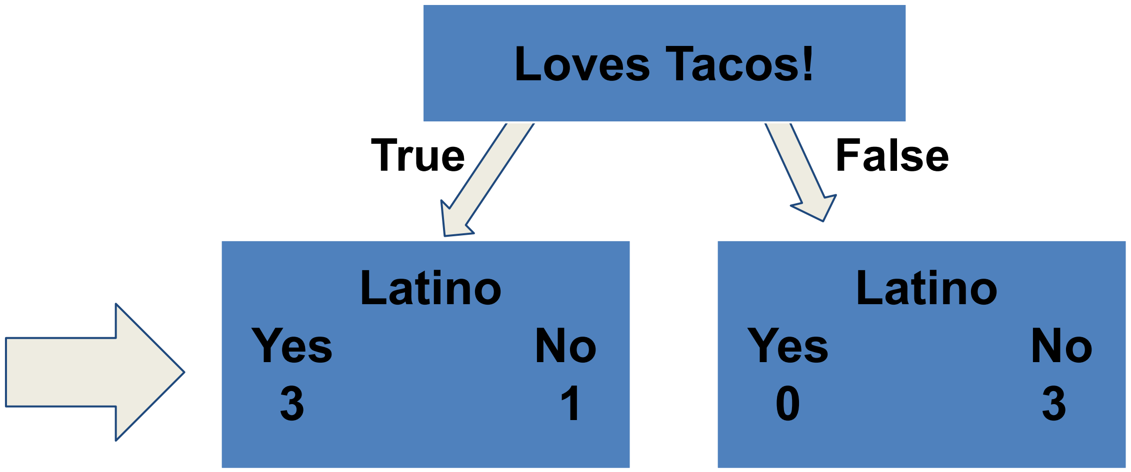
Loves spicy food	Loves tacos	Age	Latino
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



Adding Branches

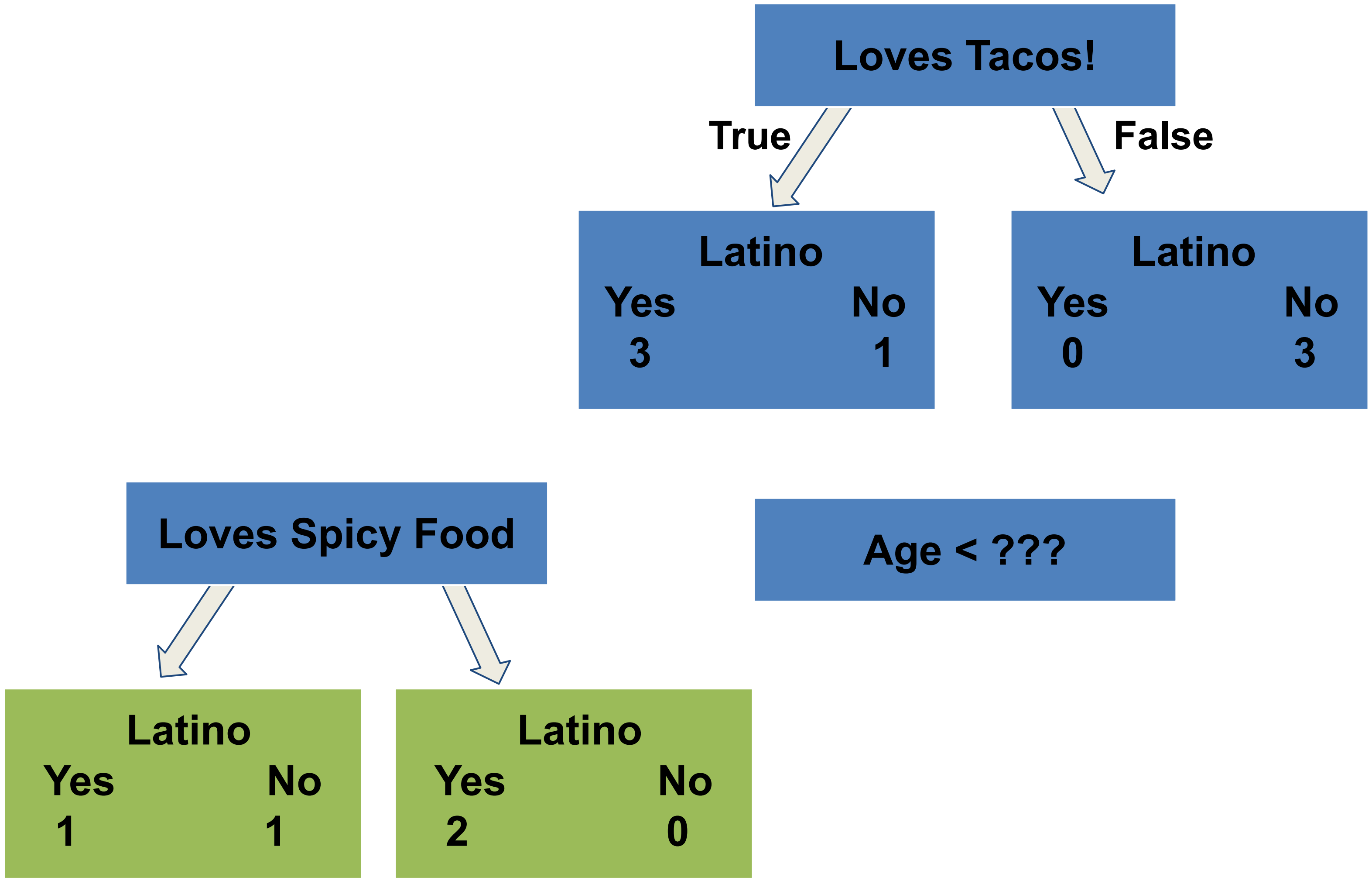
Loves spicy food	Loves tacos	Age	Latino
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Can we reduce the impurity by splitting the people that are Loves tacos based on Loves spicy food and age?



Adding Branches

Loves spicy food	Loves tacos	Age	Latino
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



Total Gini Impurity for this split is 0.25

Adding Branches

Loves spicy food	Loves tacos	Age	Latino
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 Tacos!

True / **False**

Latino	
Yes	**No**
3	**1**

Latino	
Yes	**No**
0	**3**

Loves Spicy Food

Gini = 0.25

Vs.

Age < ???

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

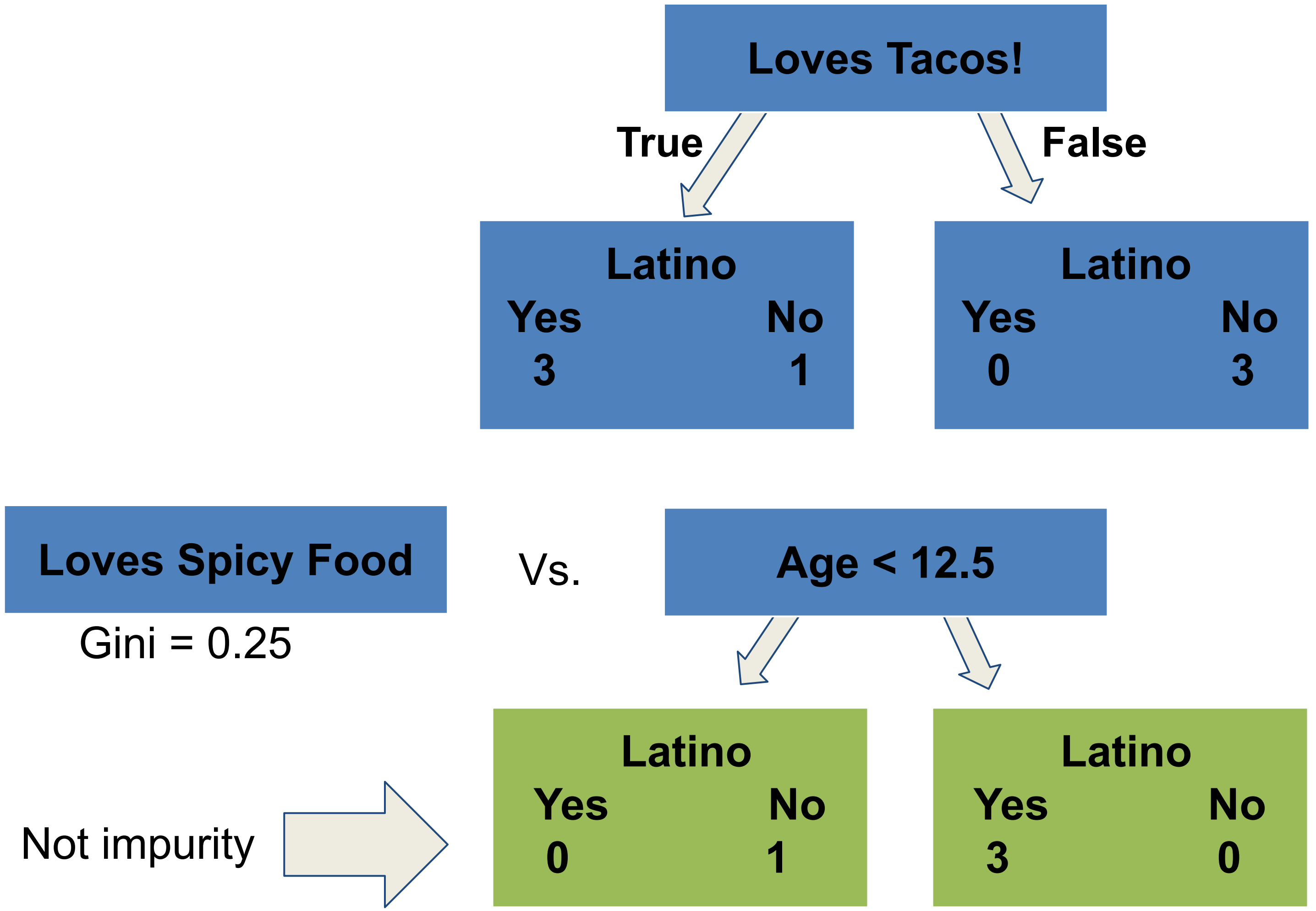
Adding Branches

Loves spicy food	Loves tacos	Age	Latino
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



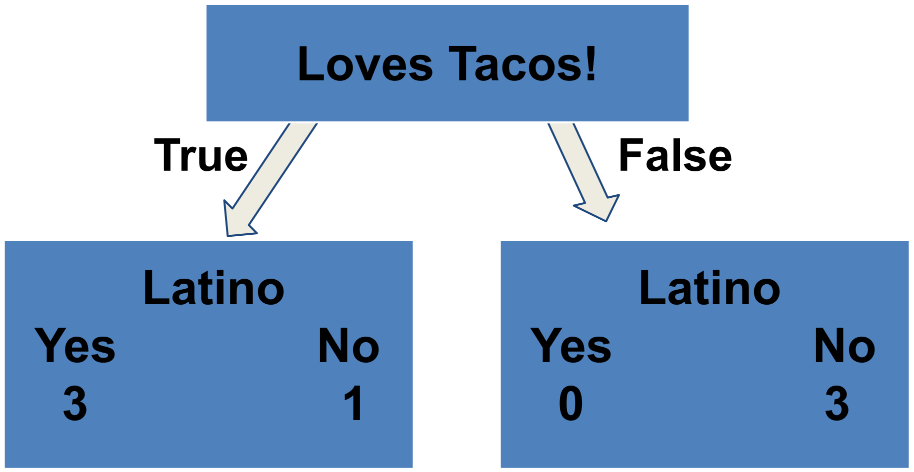
Adding Branches

Loves spicy food	Loves tacos	Age	Latino
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 Spicy Food

Gini = 0.25

Vs.

Age < 12.5

Gini = 0

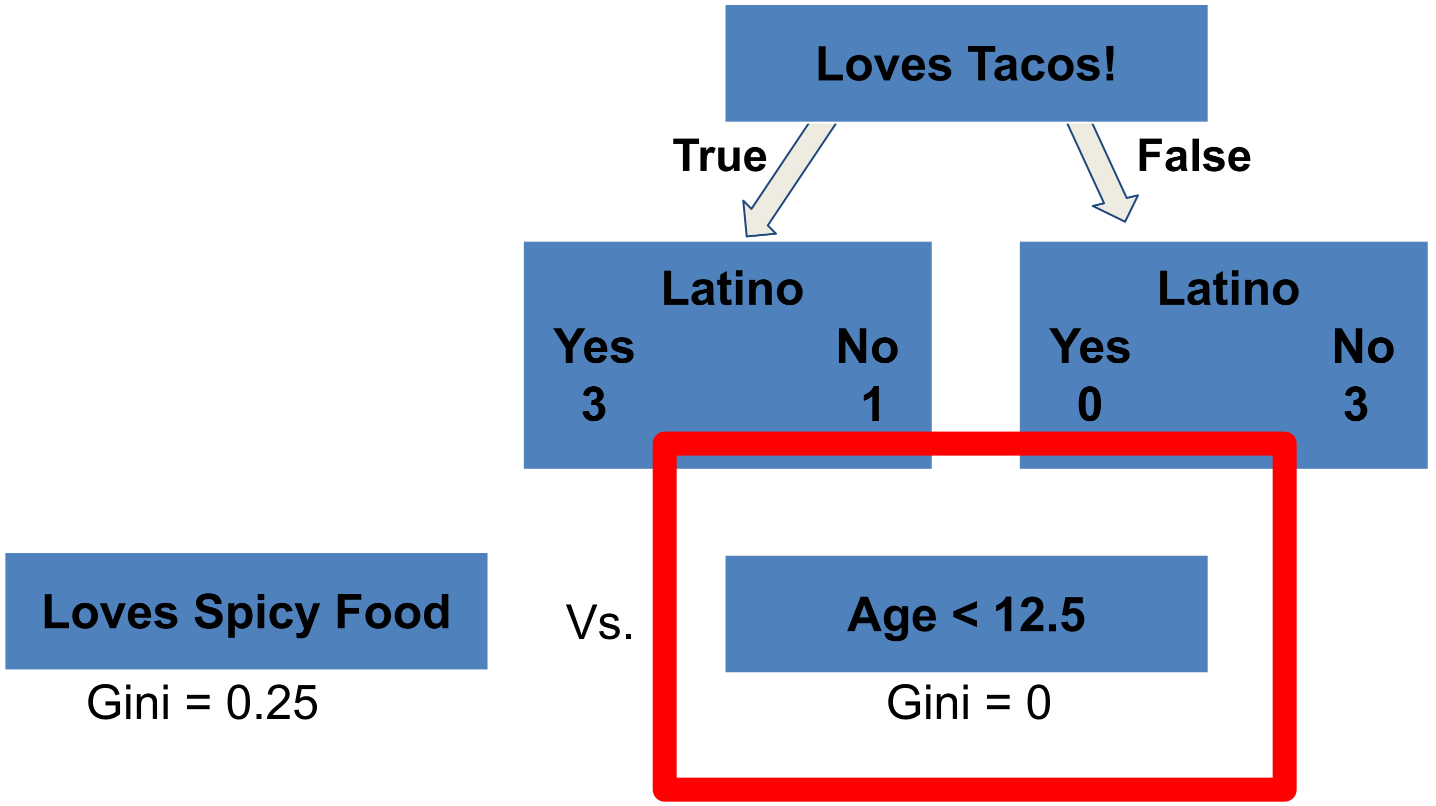
Adding Branches

Loves spicy food	Loves tacos	Age	Latino
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



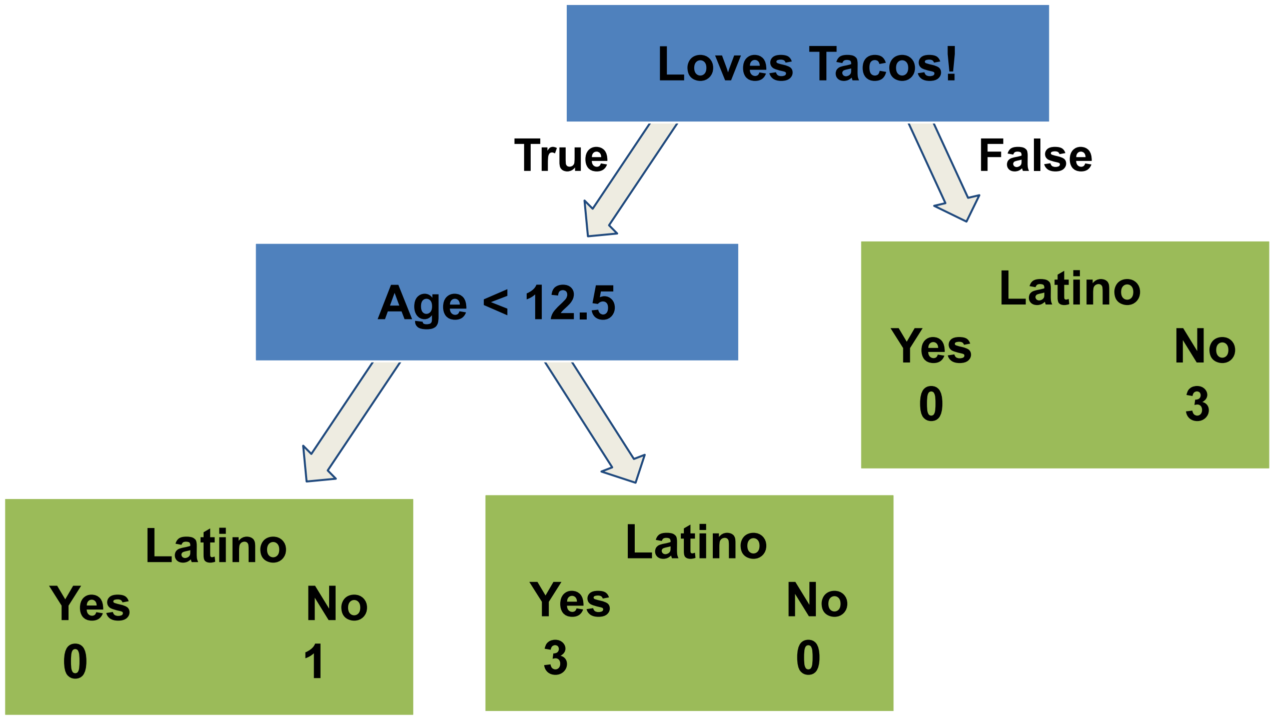
Adding Branches

Loves spicy food	Loves tacos	Age	Latino
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

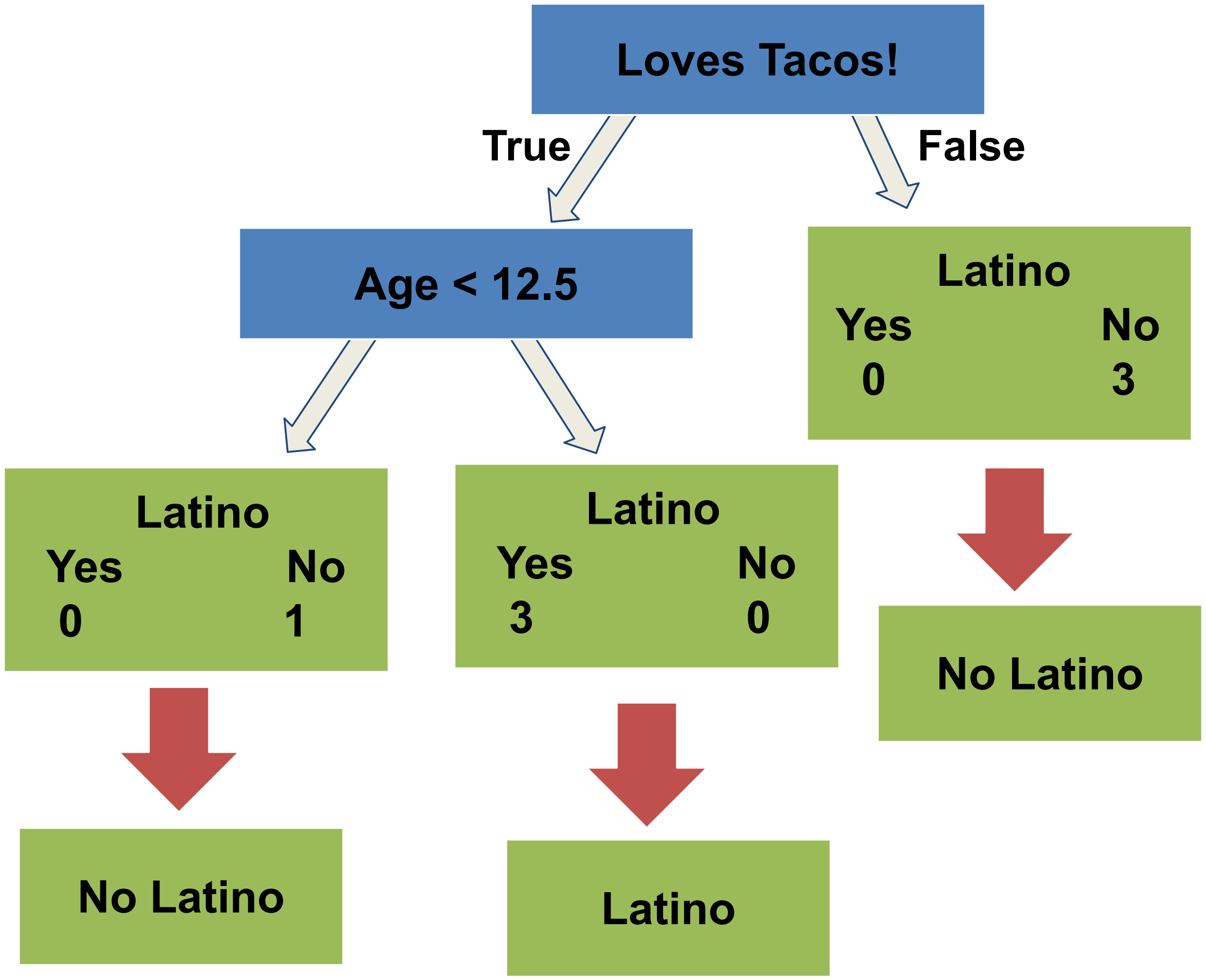
12.5

26.5

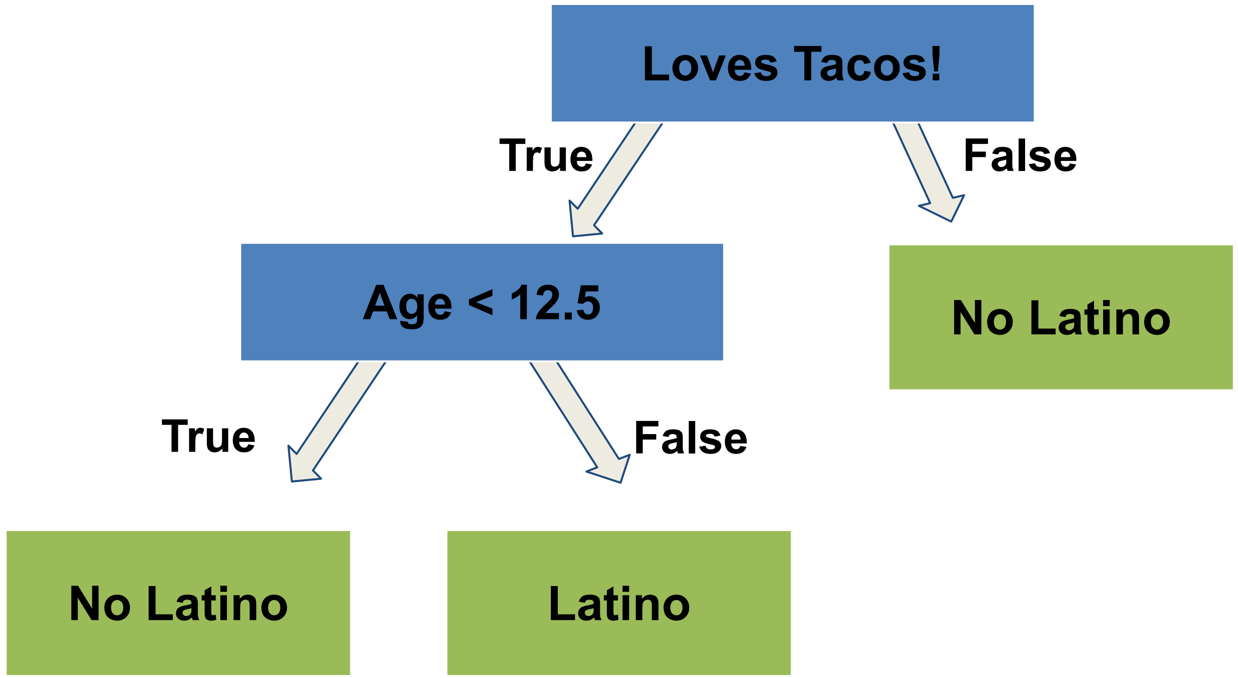
36.5



Output Values



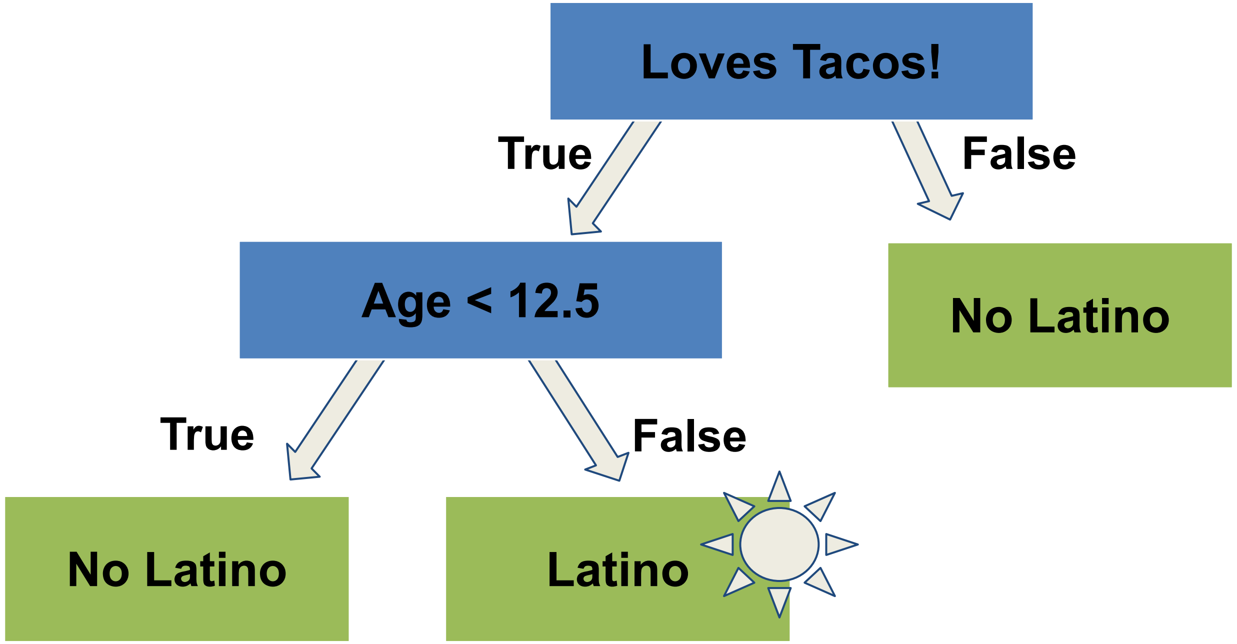
Output Values



Using the Tree

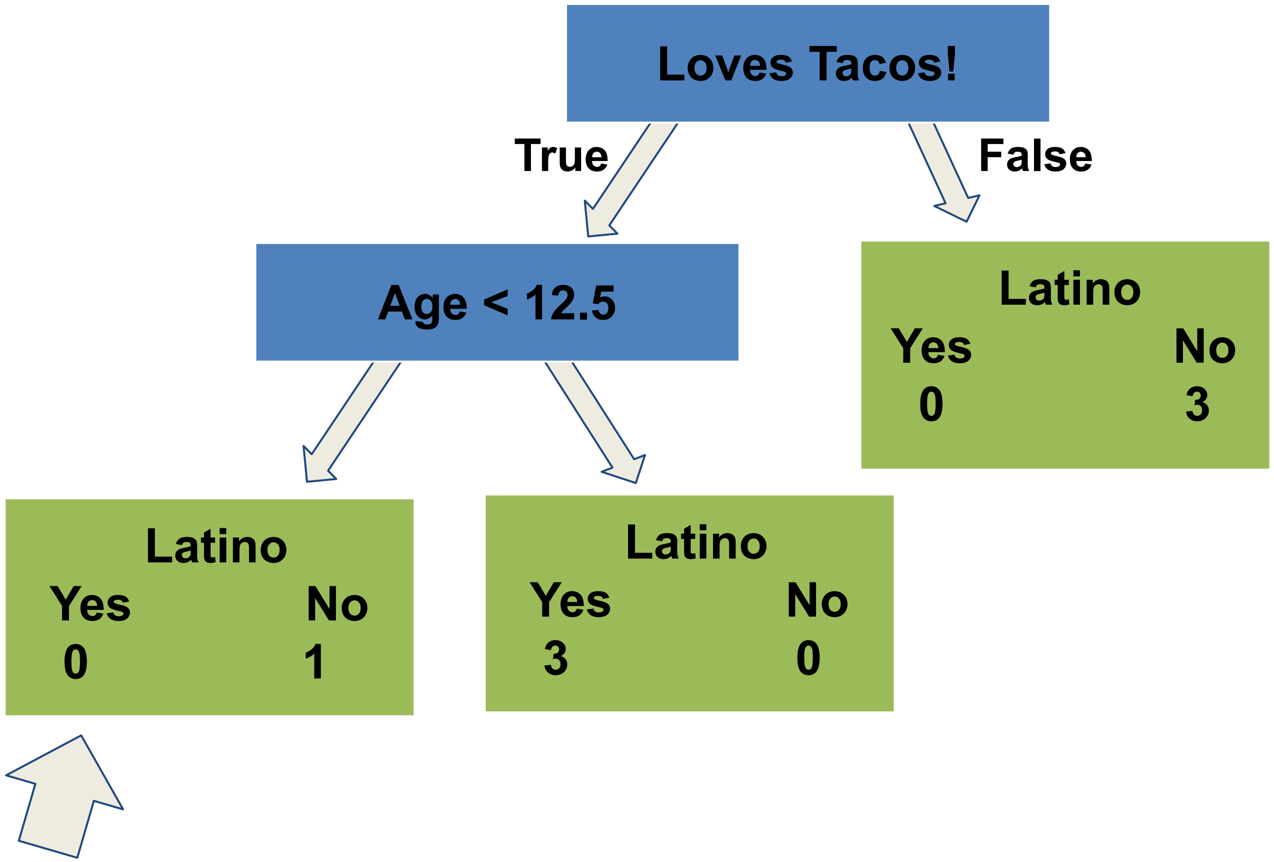
Loves spicy food	Loves tacos	Age	Latino
Yes	Yes	15	???

We want to predict if they are Latino



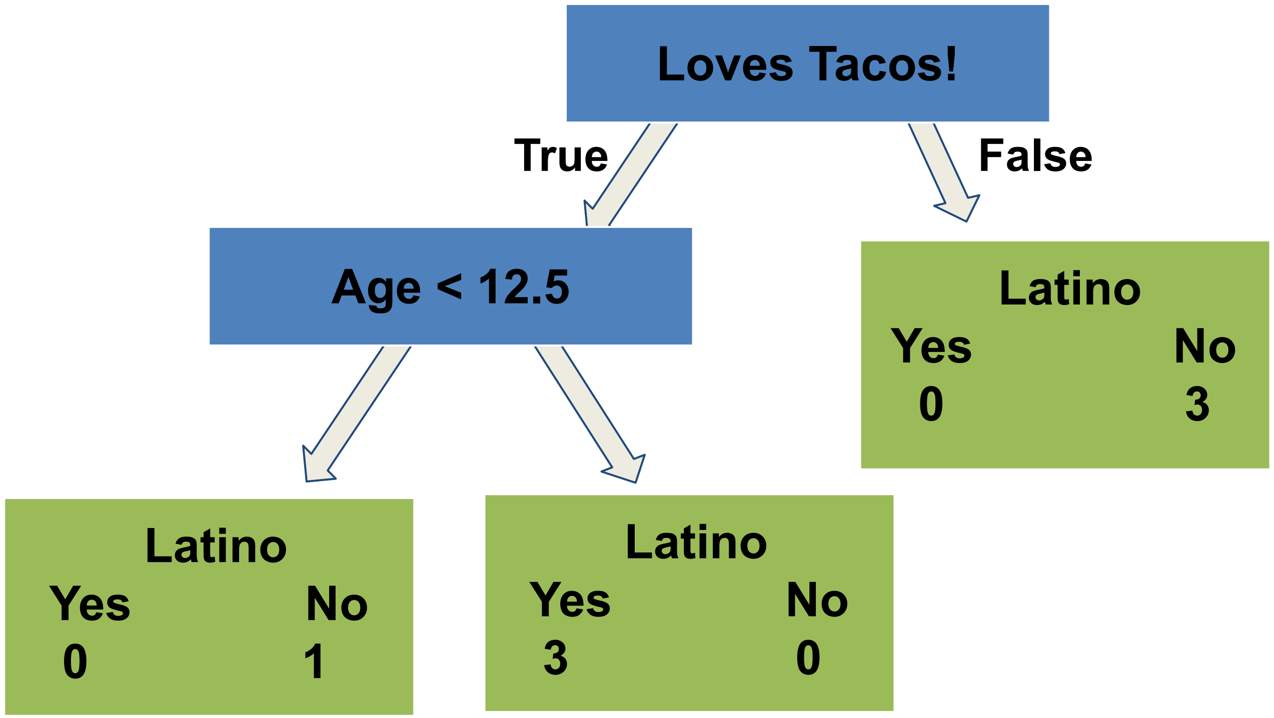
Prevent Overfitting

Loves spicy food	Loves tacos	Age	Latino
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 is hard to have confidence that it will do a great job making predictions with future data.

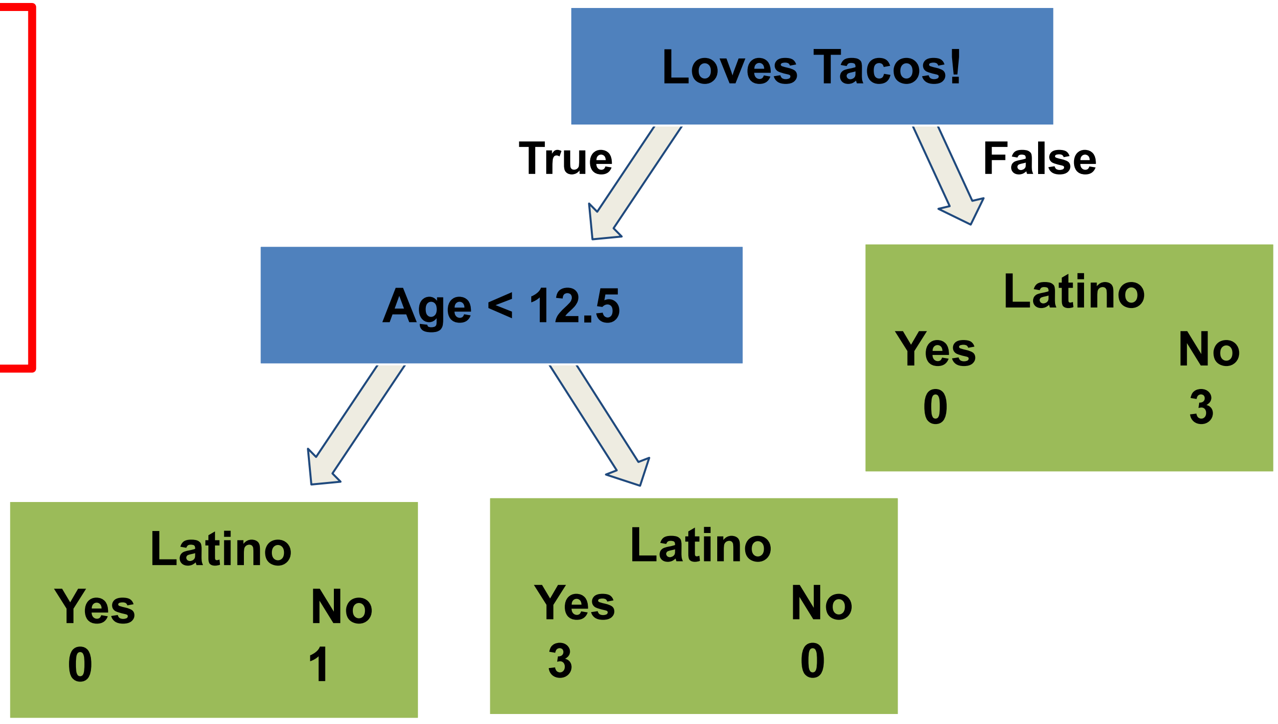
Prevent Overfitting



Possible overfit the data!

Prevent Overfitting

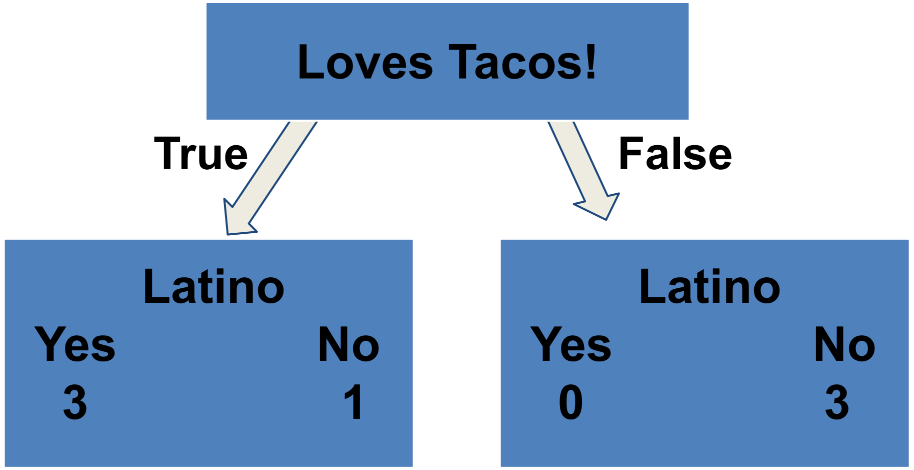
Method 1: Pruning!



Possible overfit the data!

Prevent Overfitting

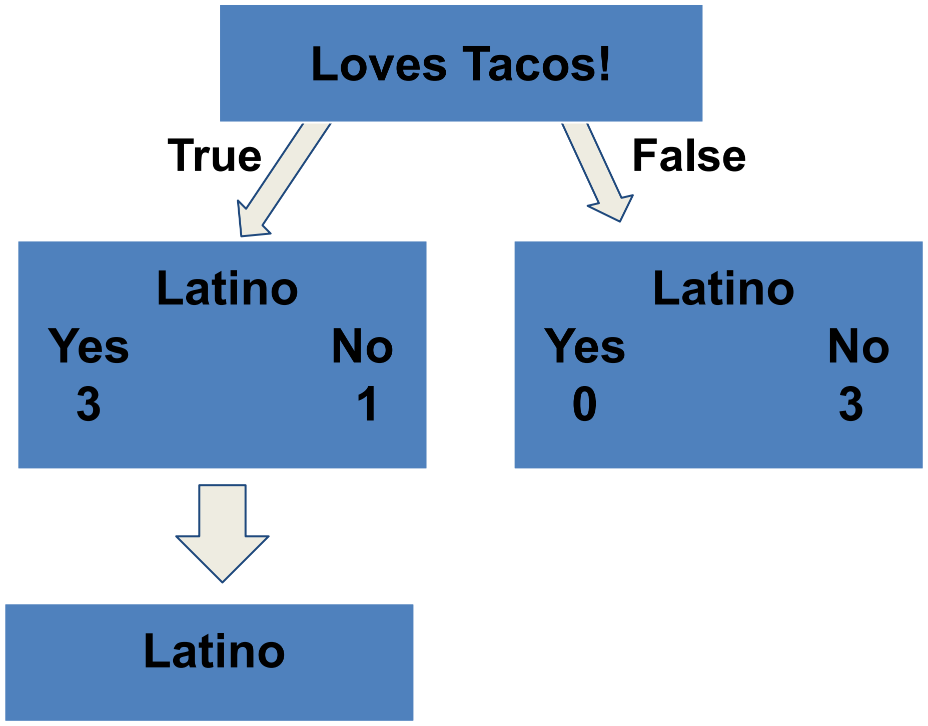
Method 2: Put limits on how the trees grow.
Example: By requiring 3 or more people per leaf.



- Select this leaf!
- Uncertainty : We can say that 75% of the people in the Leaf is Latino
- Even the Leaf is Impure we need an output value to make a classification…

Prevent Overfitting

Method 2: Put limits on how the trees grow.
Example: By requiring 3 or more people per leaf.



- Select this leaf!
- Uncertainty : We can say that 75% of the people in the Leaf is Latino
- Even the Leaf is Impure we need an output value to make a classification…