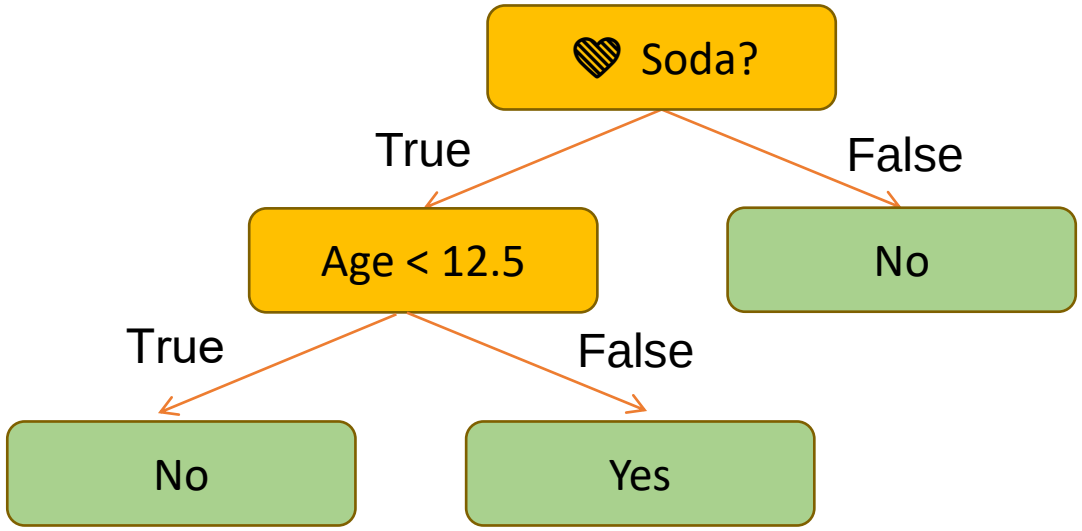
Decision Trees – Working on a Dataset

Dr. Muhammad Wasim

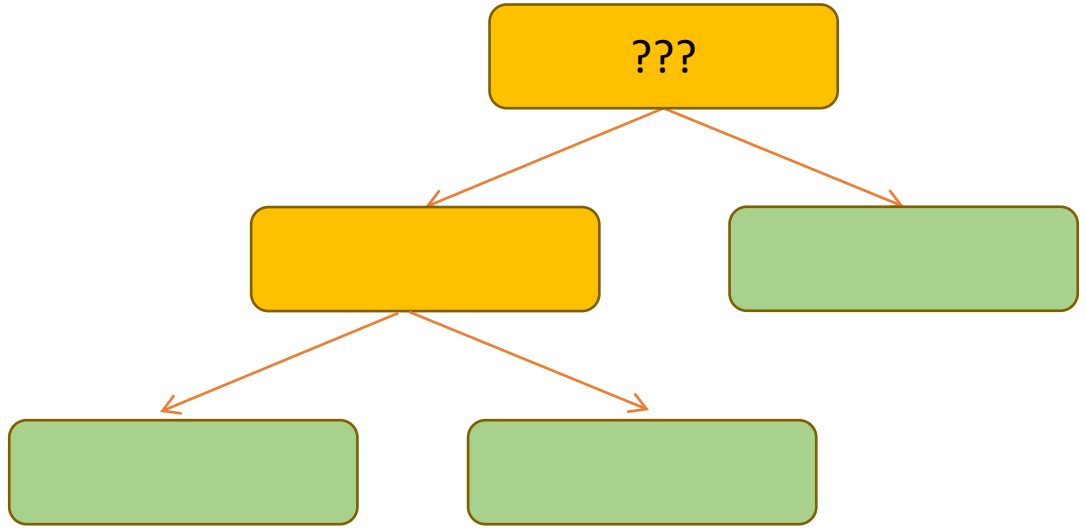
An Example Dataset

Loves Popcorn	Loves Soda	Age	Loves A Movie
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 Classification Tree for “Loves A Movie” Dataset

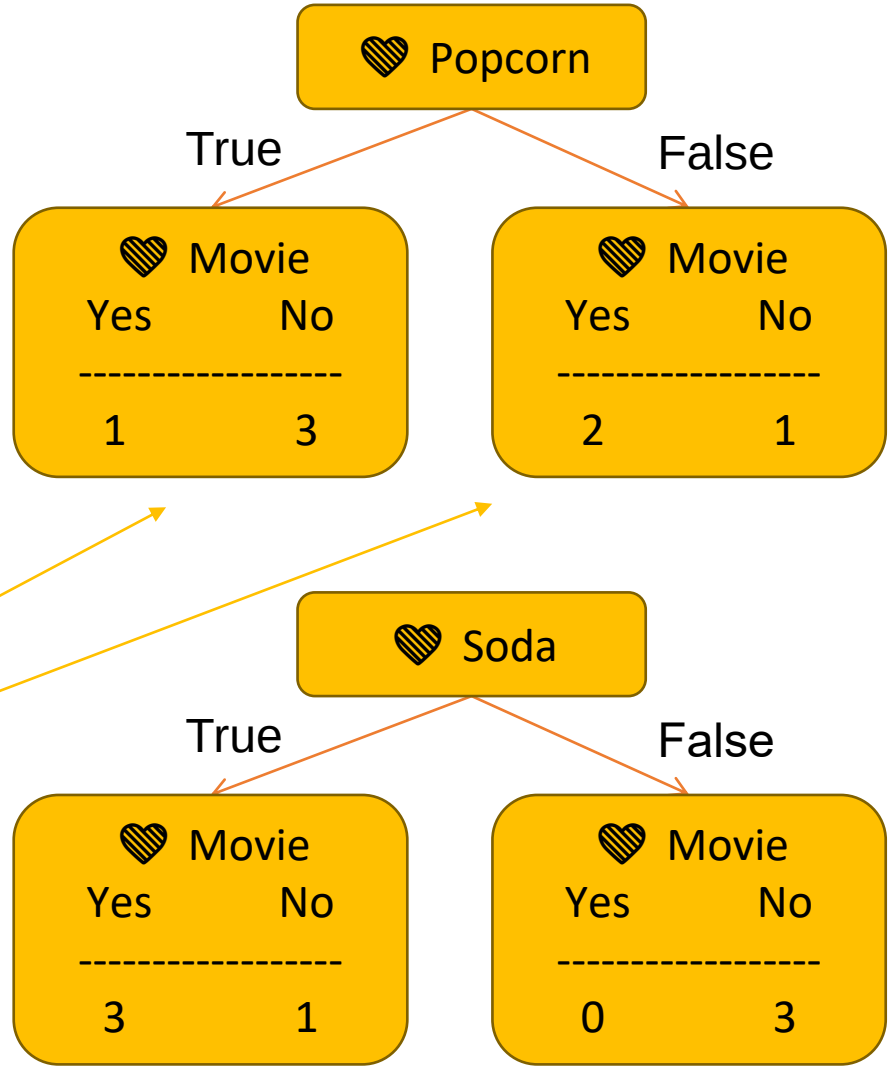
Loves Popcorn	Loves Soda	Age	Loves A Movie
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



Which feature is the most interesting question to ask to predict if a person loves A Movie?

Building a Classification Tree

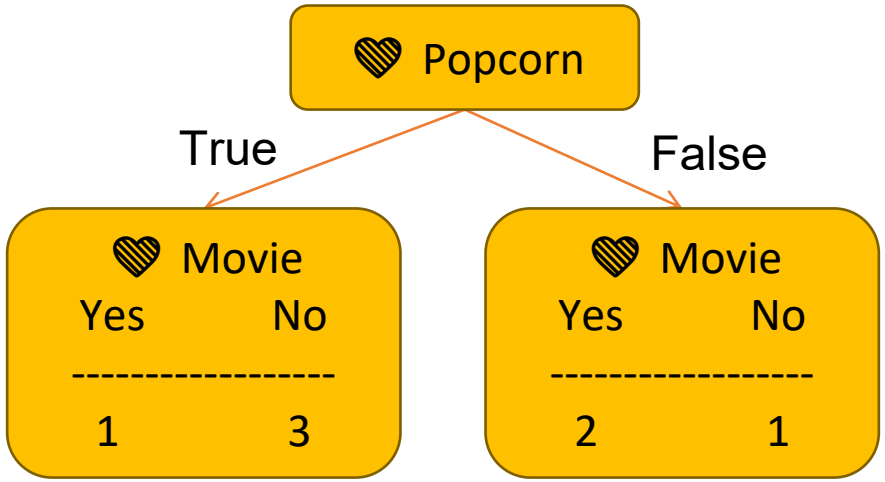
Loves Popcorn	Loves Soda	Age	Loves A Movie
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



Impure Leaves

There are several ways to quantify the impurity of the leaves (Gini Impurity, Information Gain etc.)

Building a Classification Tree



$$Gini\ Impurity\ of\ Leaf = 1 - P(Yes)^2 - P(No)^2$$

$$= 1 - \left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2 = 0.375$$

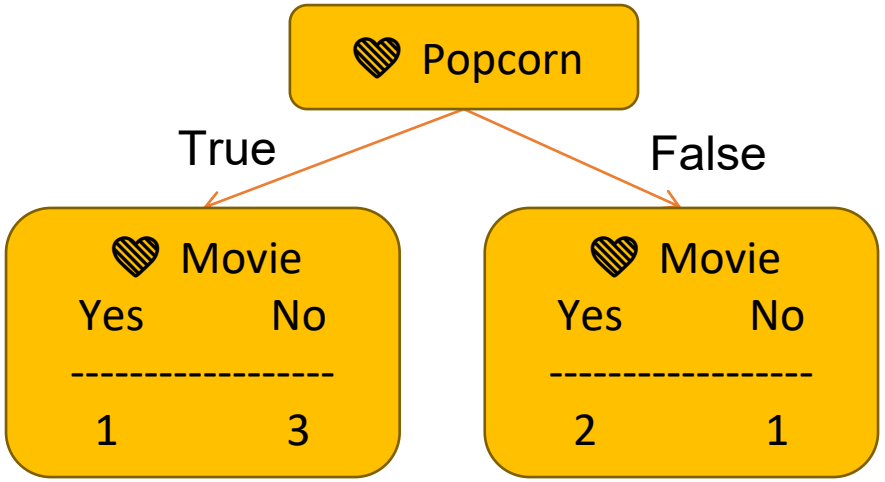
$$Gini\ Impurity\ of\ Leaf = 1 - P(Yes)^2 - P(No)^2$$

$$= 1 - \left(\frac{2}{2+1}\right)^2 - \left(\frac{1}{2+1}\right)^2 = 0.444$$

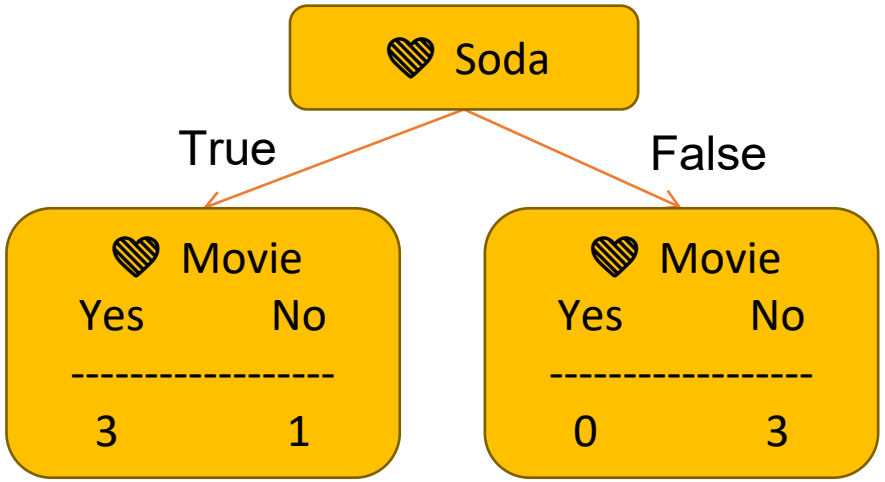
$$Total\ Gini\ Impurity = weighted\ average\ of\ Gini\ impurities\ for\ the\ leaves$$

$$= \frac{4}{4+3} \times 0.375 + \frac{3}{4+3} \times 0.444 = 0.405$$

Building a Classification Tree



$Gini\ Impurity\ for\ Love\ Popcorn = 0.405$



$Gini\ Impurity\ for\ Love\ Soda = 0.214$

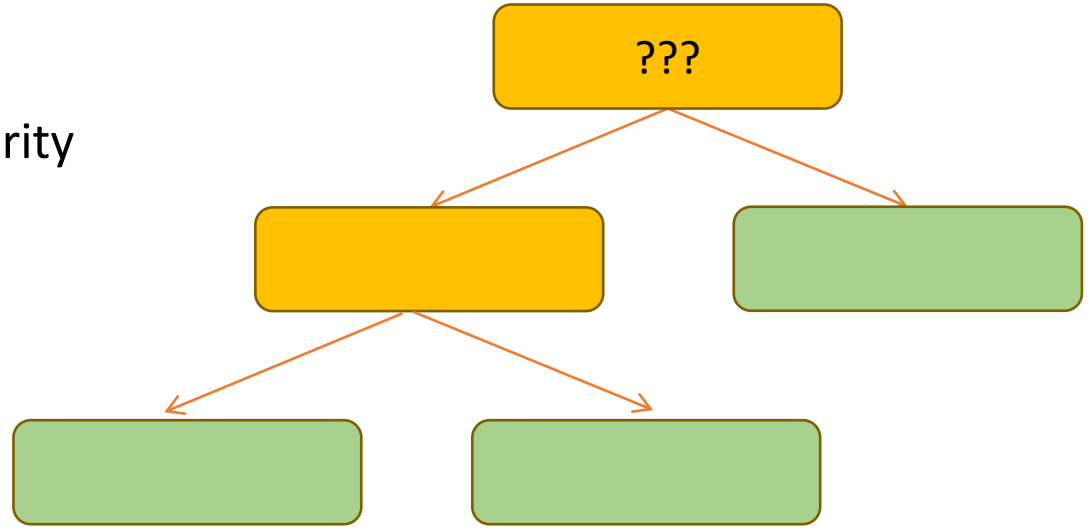
Building a Classification Tree

Loves Popcorn	Loves Soda	Age	Loves A Movie
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, 15, 26.5, 36.5, 44, 66.5

Gini impurity

0.429
0.343
0.476
0.476
0.343
0.429

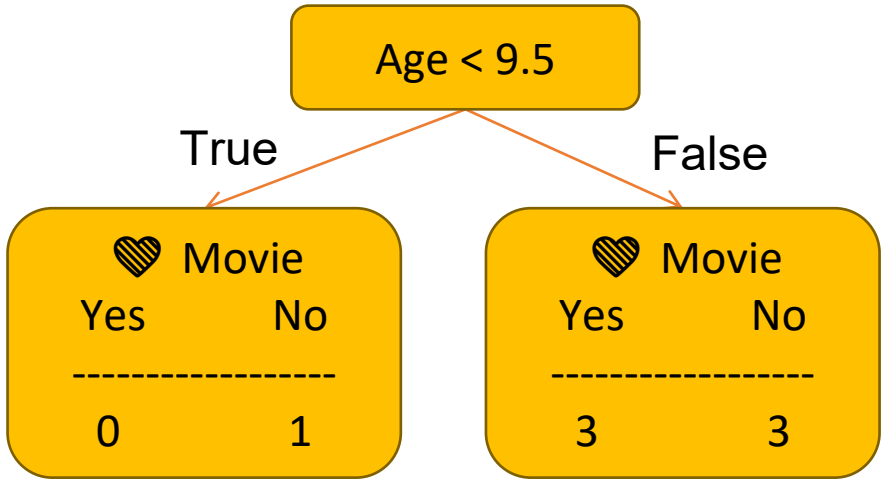


Now, we need to calculate the gini impurity for age!

$Gini\ Impurity\ for\ Love\ Popcorn = 0.405$

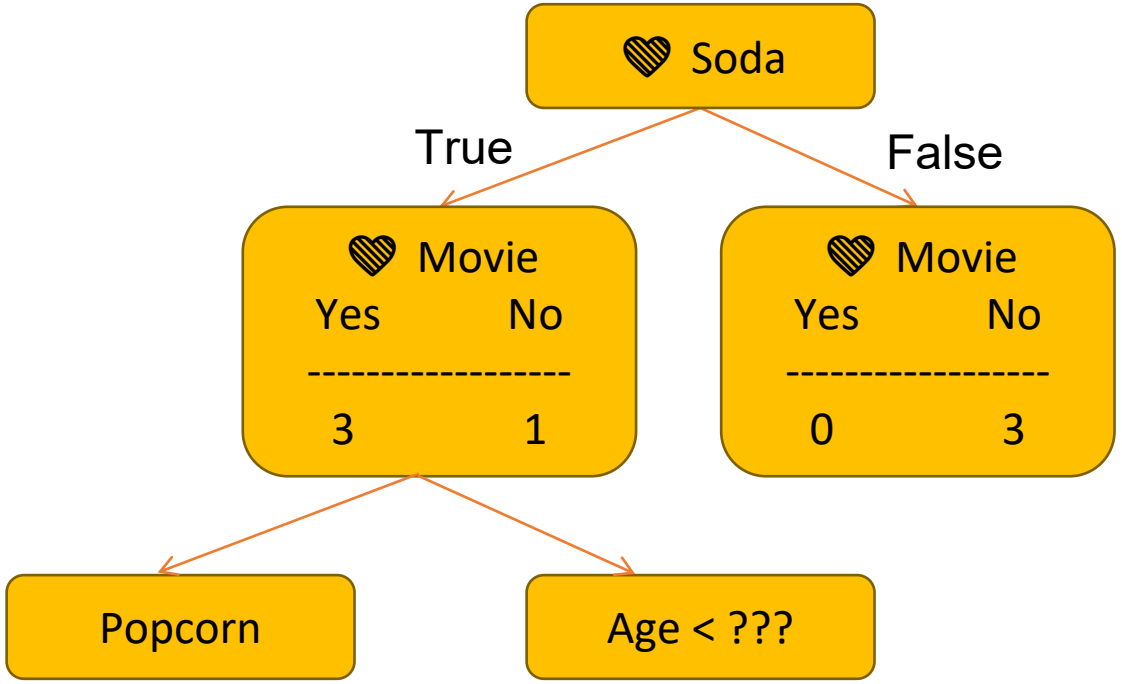
$Gini\ Impurity\ for\ Love\ Soda = 0.214$

$Gini\ Impurity\ for\ Age < 15 = 0.343$



Building a Classification Tree

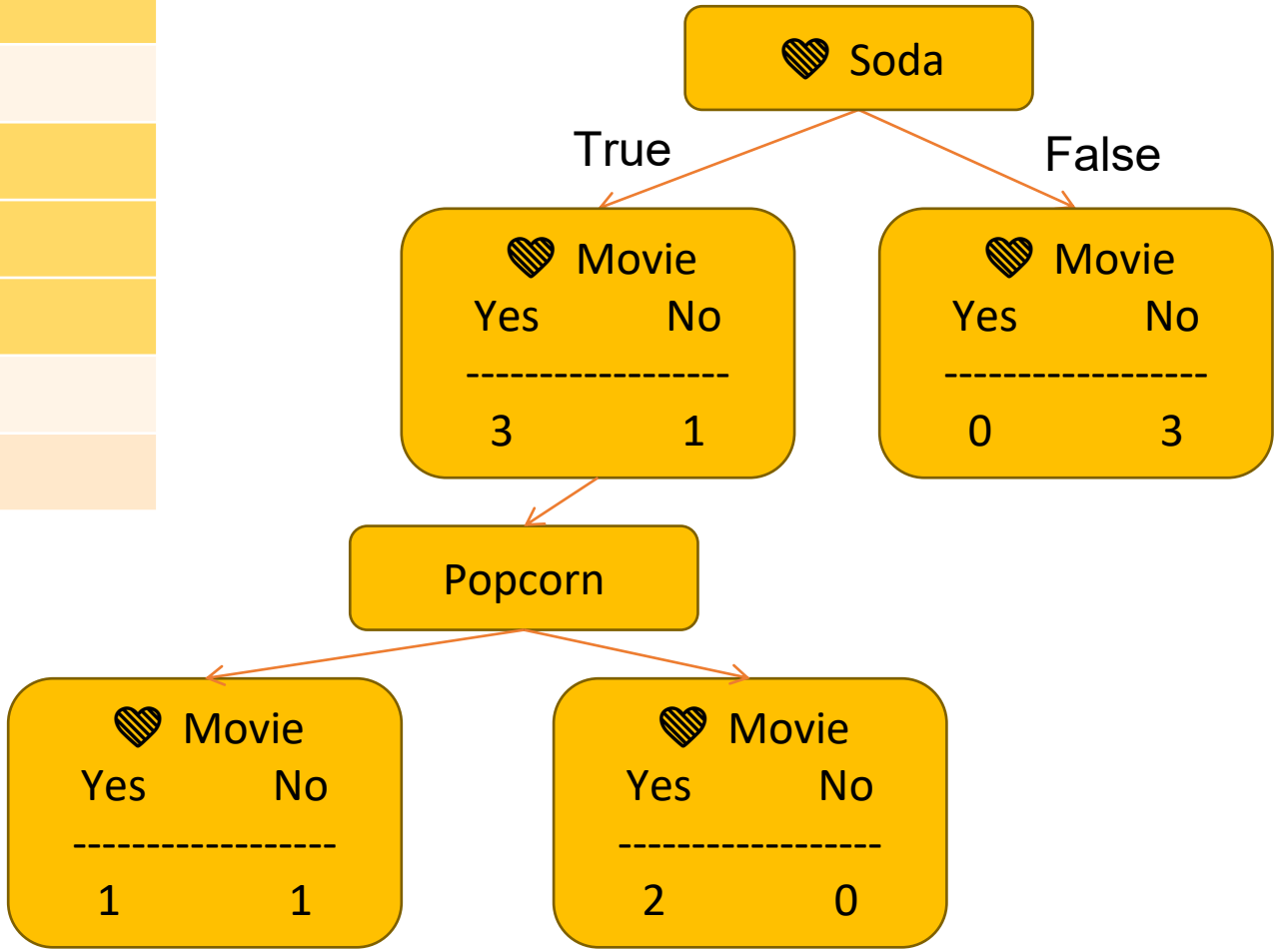
Loves Popcorn	Loves Soda	Age	Loves A Movie
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



Building a Classification Tree

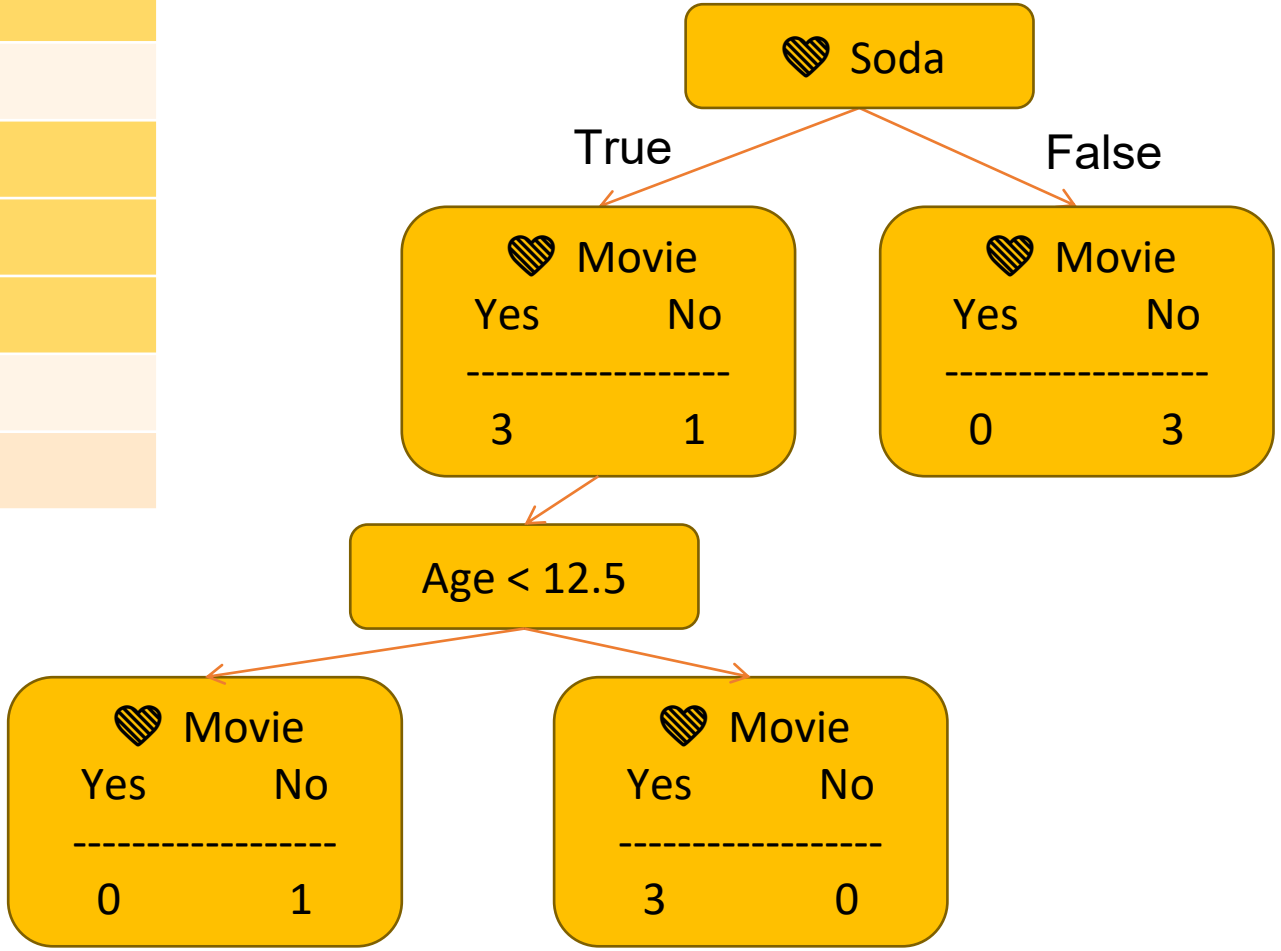
Loves Popcorn	Loves Soda	Age	Loves A Movie
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

$$Gini\ Impurity\ for\ Popcorn = 0.25$$



Building a Classification Tree

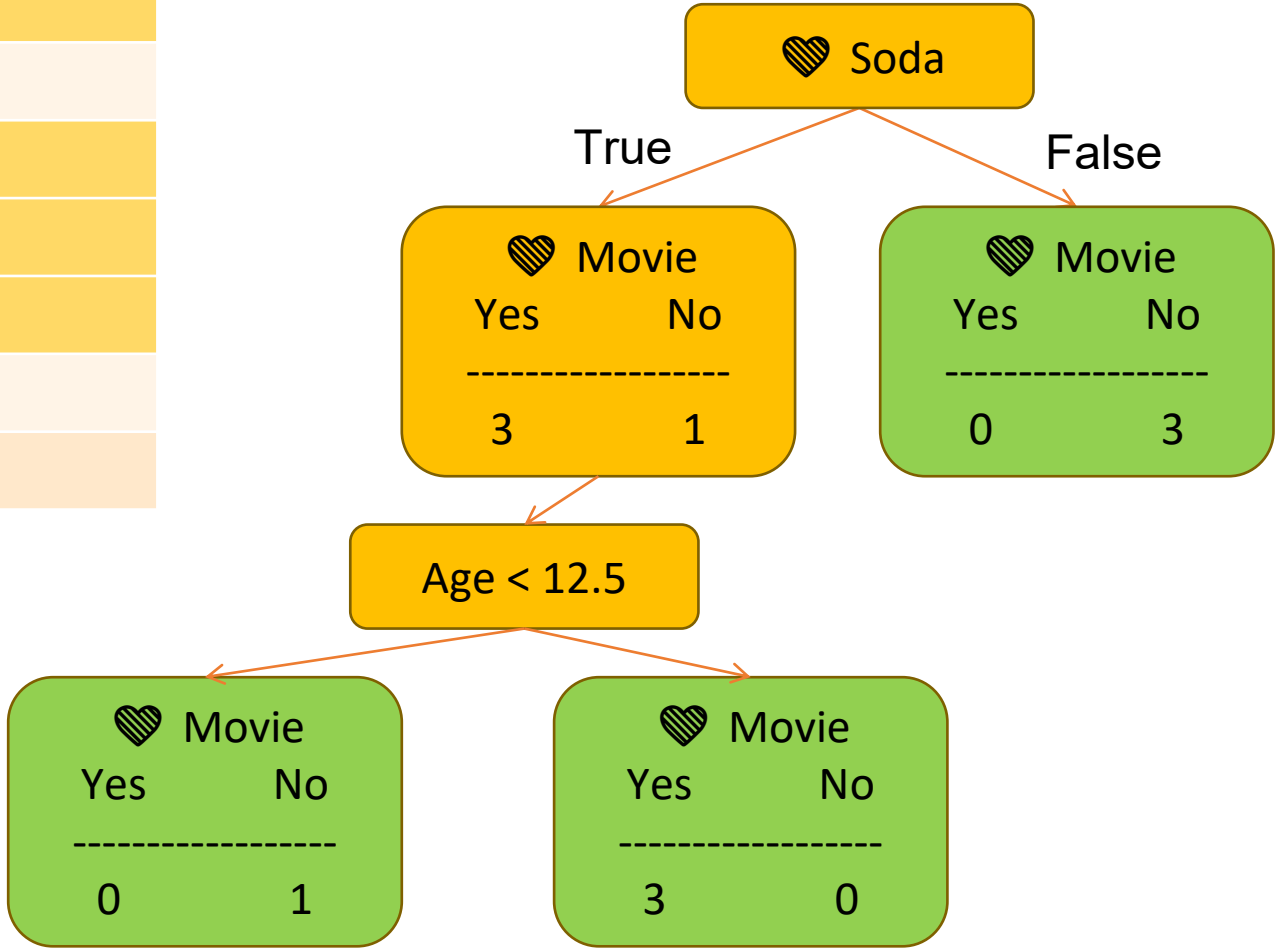
Loves Popcorn	Loves Soda	Age	Loves A Movie
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No



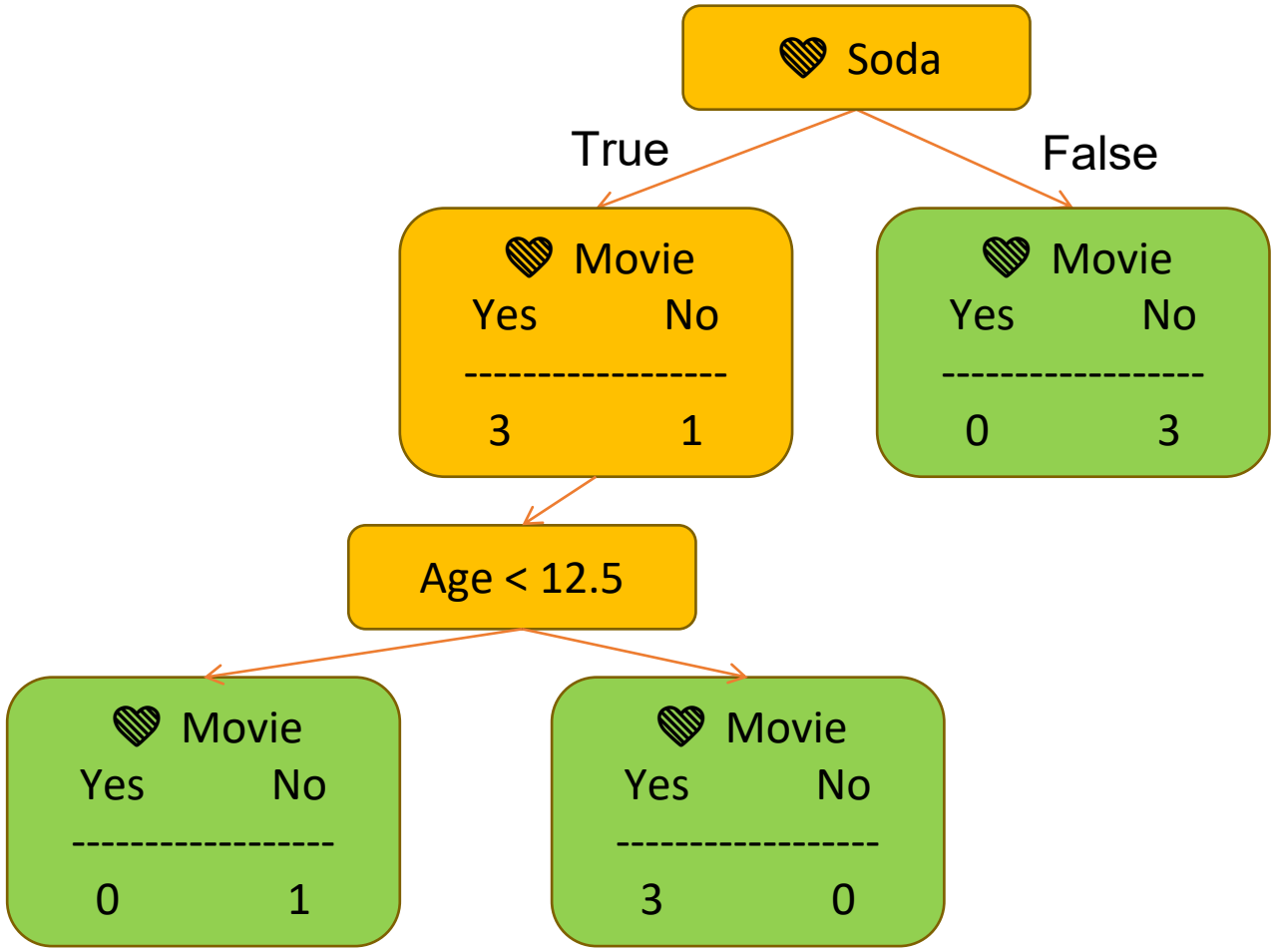
$$Gini\ Impurity\ for\ Age < 12.5 = 0$$

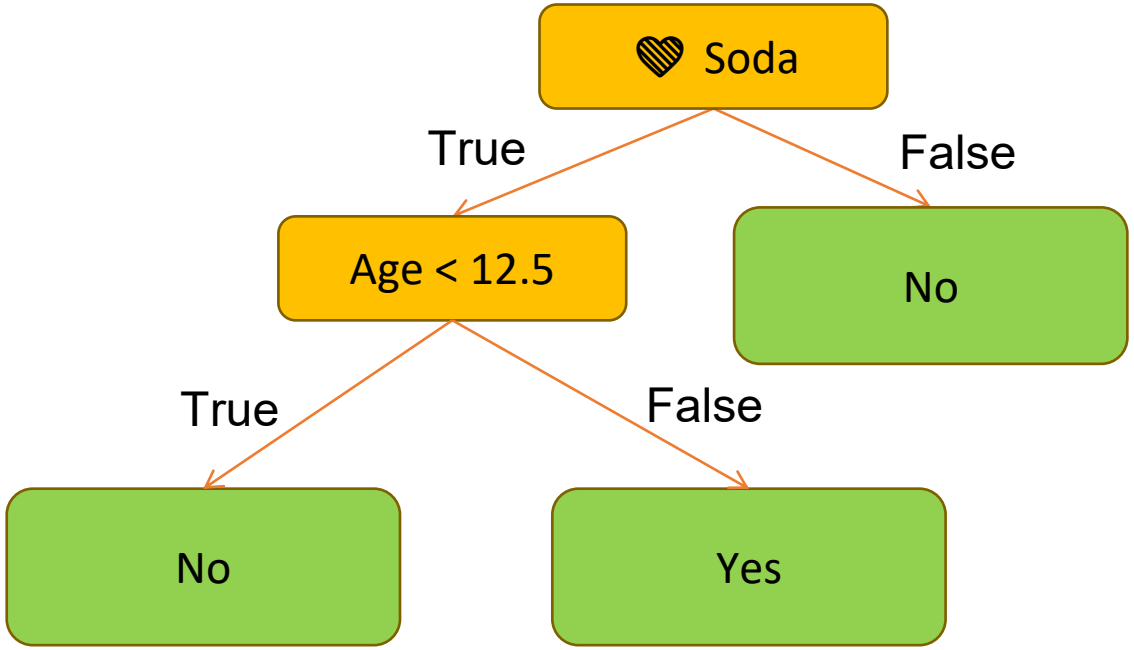
Building a Classification Tree

Loves Popcorn	Loves Soda	Age	Loves A Movie
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

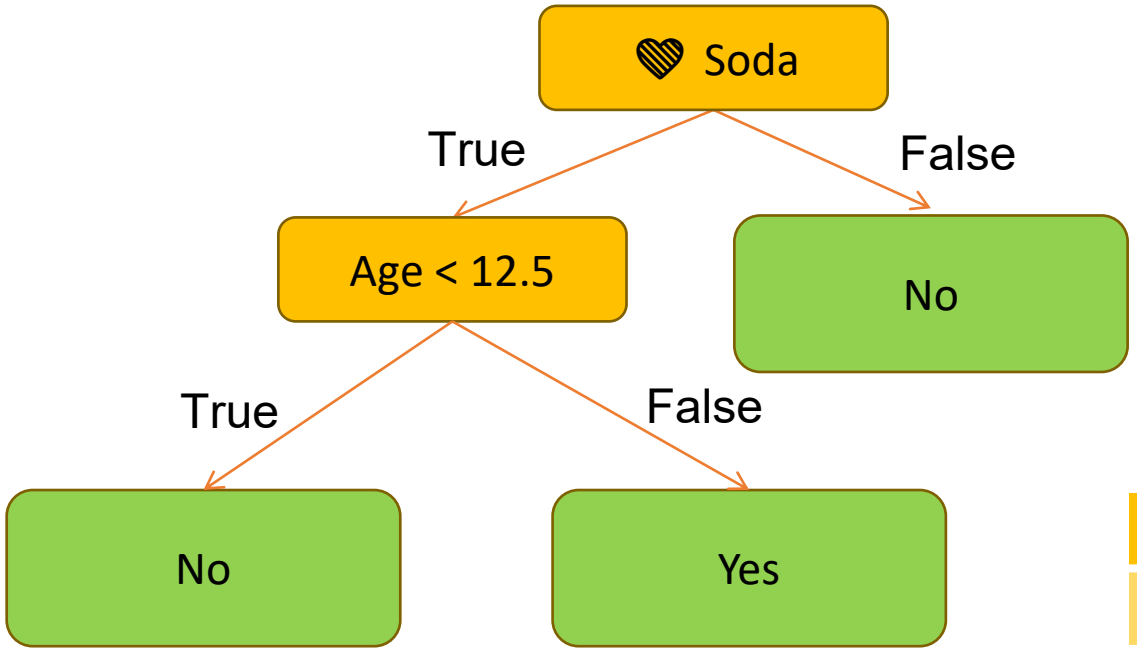


Building a Classification Tree





Building a Classification Tree



Loves Popcorn	Loves Soda	Age	Loves A Movie
Yes	Yes	15	