Decision Trees

- Used for classification and Regression as well.

Attributes

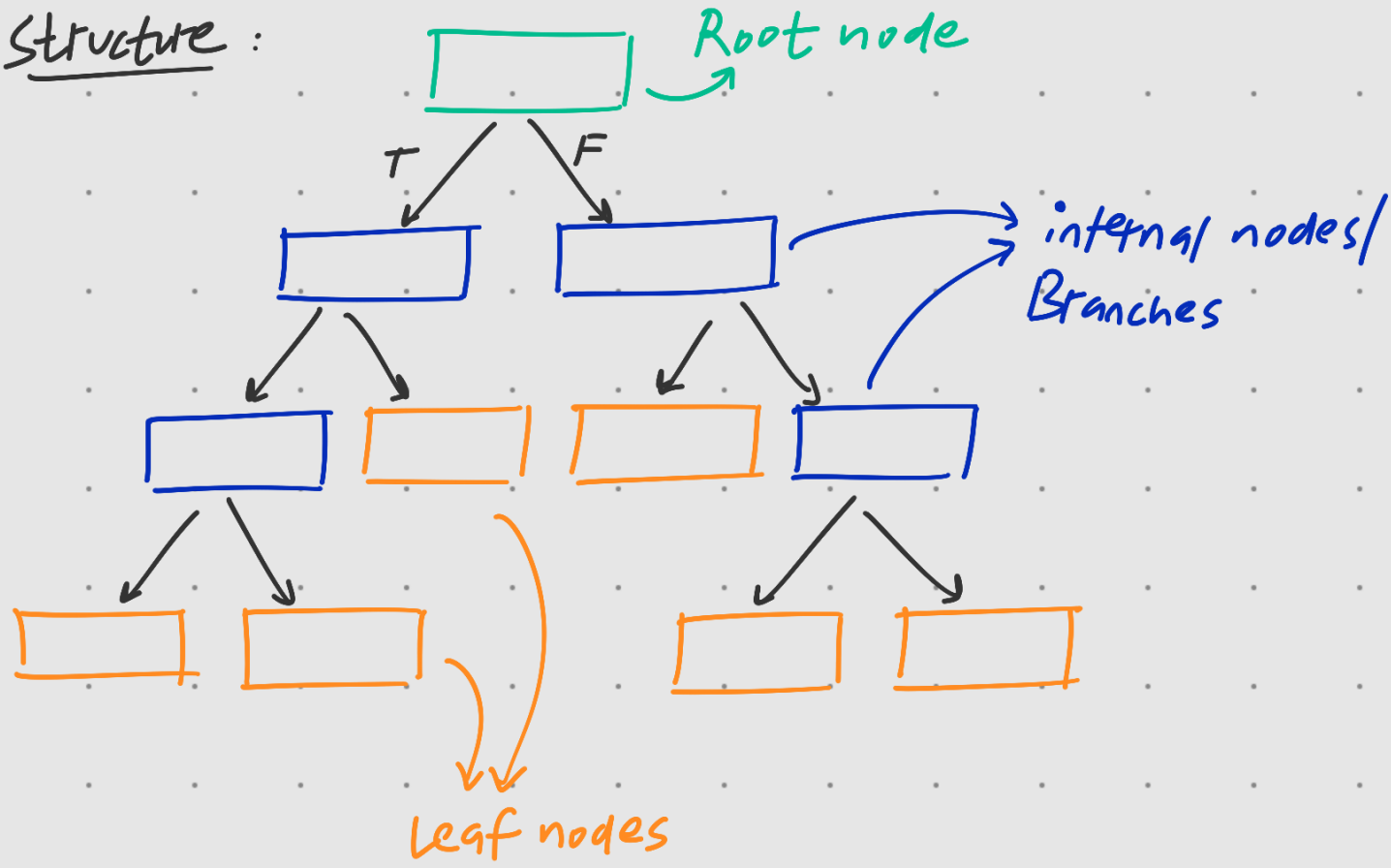
↳ Tree structure
Decision nodes
Leaf nodes
Splitting
Entropy
Information Gain
Pruning

- Decision Tree makes a statement & then makes decision based on whether or not statement is True or False.

- Assumption: If, statement is True → go Left.
Statement is False → go Right.

Classification Tree

Structure :



Eg

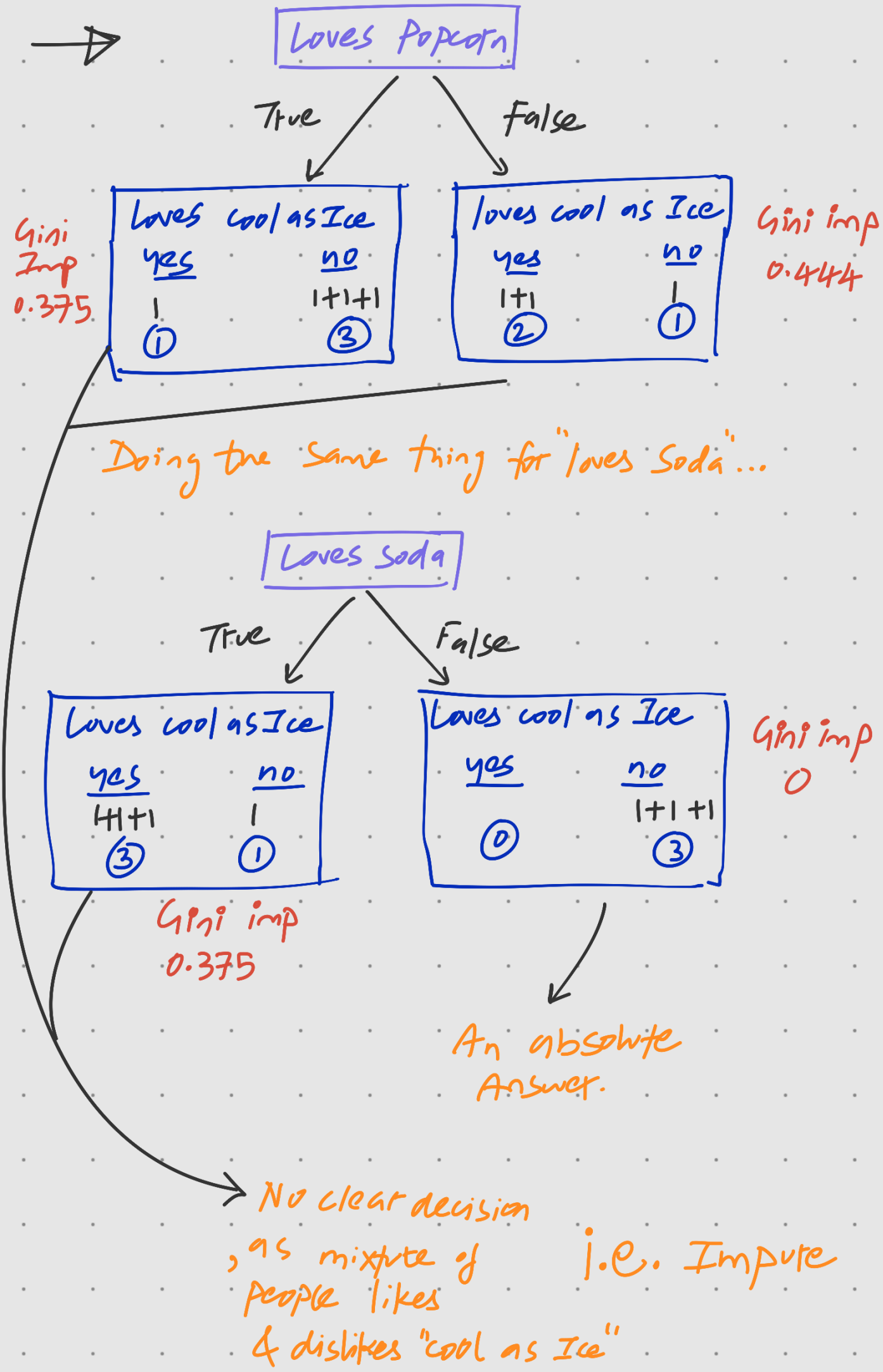
Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

→ To be predicted using, Previous 3 columns.

Soln

What should be the node?

Determining it... Hit & Trial (initially)


Loves Popcorn
True
False
Loves cool as Ice
yes
no
1
1+1+1
1
3
Gini Imp 0.375
loves cool as Ice
yes
no
1+1
1
2
1
Gini imp 0.444
Doing the same thing for "loves Soda"...
Loves Soda
True
False
Loves cool as Ice
yes
no
1+1+1
1
3
1
Gini imp 0.375
Loves cool as Ice
yes
no
0
1+1+1
3
Gini imp 0
An absolute Answer.
No clear decision, as mixture of people likes & dislikes "cool as Ice"
i.e. Impure

In "loves Popcorn" → Both leaves are impure (less information gain)

In "loves Soda" → only one leaf is impure (more information gain compared to previous.)

Does better Job at predicting who will "love cool as Ice"

A way to quantify the impurity of leaves...

Gini Impurity

Start by calculating Gini Impurity of individual leaves...

$$\text{Gini Impurity} = 1 - (P(\text{yes}))^2 - (P(\text{no}))^2$$

→ Loves Popcorn (leaf 1)

$$\text{Gini impurity} = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2$$

$$= 1 - \frac{1}{16} - \frac{9}{16}$$

$$= \frac{6}{16} = 0.375$$

Loves Popcorn (leaf 2)

$$\text{Gini impurity} = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2$$

$$= 1 - \frac{4}{9} - \frac{1}{9}$$

$$= \frac{4}{9} = 0.444$$

→ ## Loves Soda (leaf 1)

$$\text{Gini Impurity} = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2$$

$$= 1 - \frac{9}{16} - \frac{1}{16} = \frac{6}{16}$$

$$= 0.375$$

Loves Soda (leaf 2)

$$\text{Gini impurity} = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2$$

$$= 1 - 0 - 1$$

$$= 0$$

Both leaves of "Loves Popcorn" doesn't have same no. of people...

Total Gini impurity = weighted Avg. of gini impurities of the leaves

calculating the weight...

weight = $\frac{\text{Total people in 1st leaf}}{\text{Total people in both leaves}}$ × Gini impurity (leaf 1)

Total Gini impurity (loves popcorn) = $\frac{4}{(4+3)} \cdot (0.375) + \frac{3}{(4+3)} \cdot (0.444)$

= 0.405 !

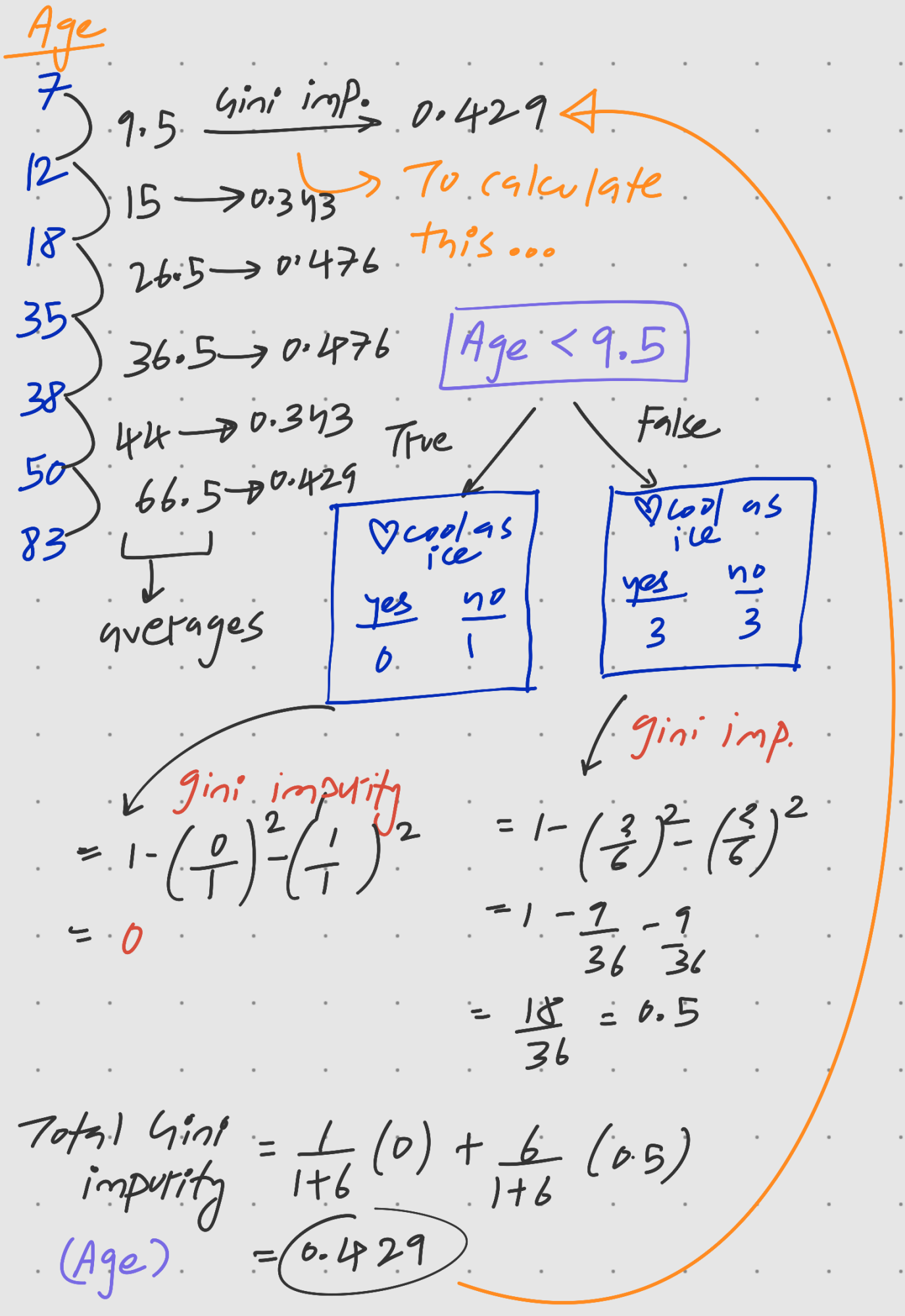
Total Gini impurity (loves soda) = $\frac{4}{(4+3)} \cdot (0.375) + \frac{3}{(4+3)} \cdot (0)$

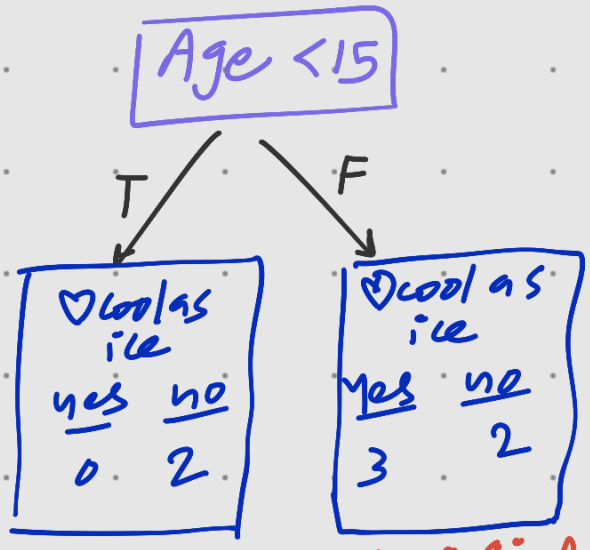
= 0.214 !

→ Gini Impurity by Age

Age
7
12
18
35
38
50
83

- First we sort the numeric data.
- Then, we calculate average for all adjacent people.
- Lastly, calculate Gini impurities for each average age.


Age
7
12
18
35
38
50
83
averages
9.5
15 → 0.343
26.5 → 0.476
36.5 → 0.476
44 → 0.343
66.5 → 0.429
Gini imp.
0.429
To calculate this...
Age < 9.5
True
False
♡cool as ice
yes
no
0
1
♡cool as ice
yes
no
3
3
gini impurity
gini imp.
= 1 - (0/1)² - (1/1)²
= 0
= 1 - (3/6)² - (3/6)²
= 1 - 1/36 - 9/36
= 18/36 = 0.5
Total Gini impurity (Age) = 1/1+6 (0) + 6/1+6 (0.5)
= 0.429



↓gini imp

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

$= 0$

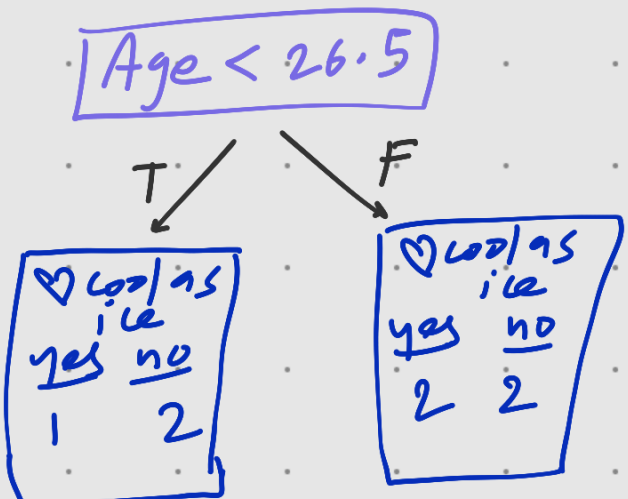
↓ gini imp

$1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2$

$= 0.48$

Total Gini impurity $= \frac{2}{7}(0)+\frac{5}{7}(0.48)$

$= 0.343$



↓G.I

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

$= 0.44$

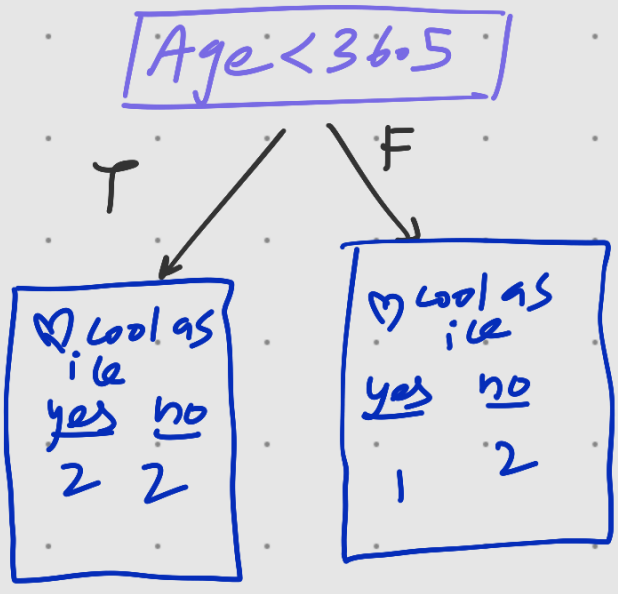
↓G.I

$1-\frac{4}{16}-\frac{4}{16}$

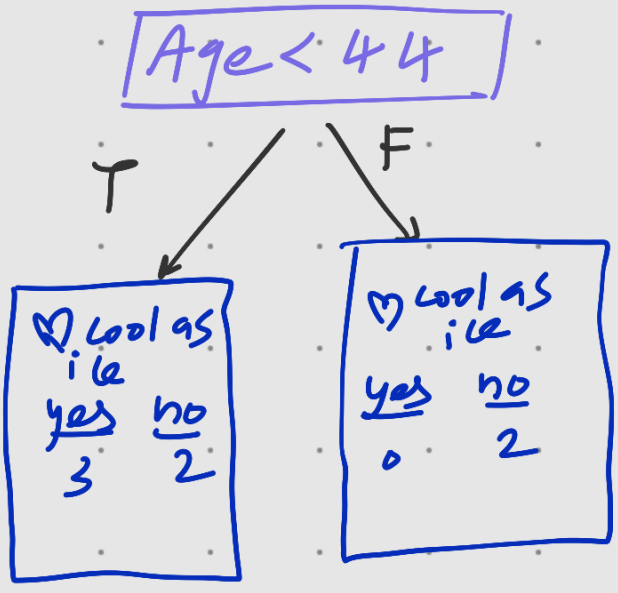
$= 0.5$

Total GI $= \frac{3}{7}(0.44)+\frac{4}{7}(0.5)$

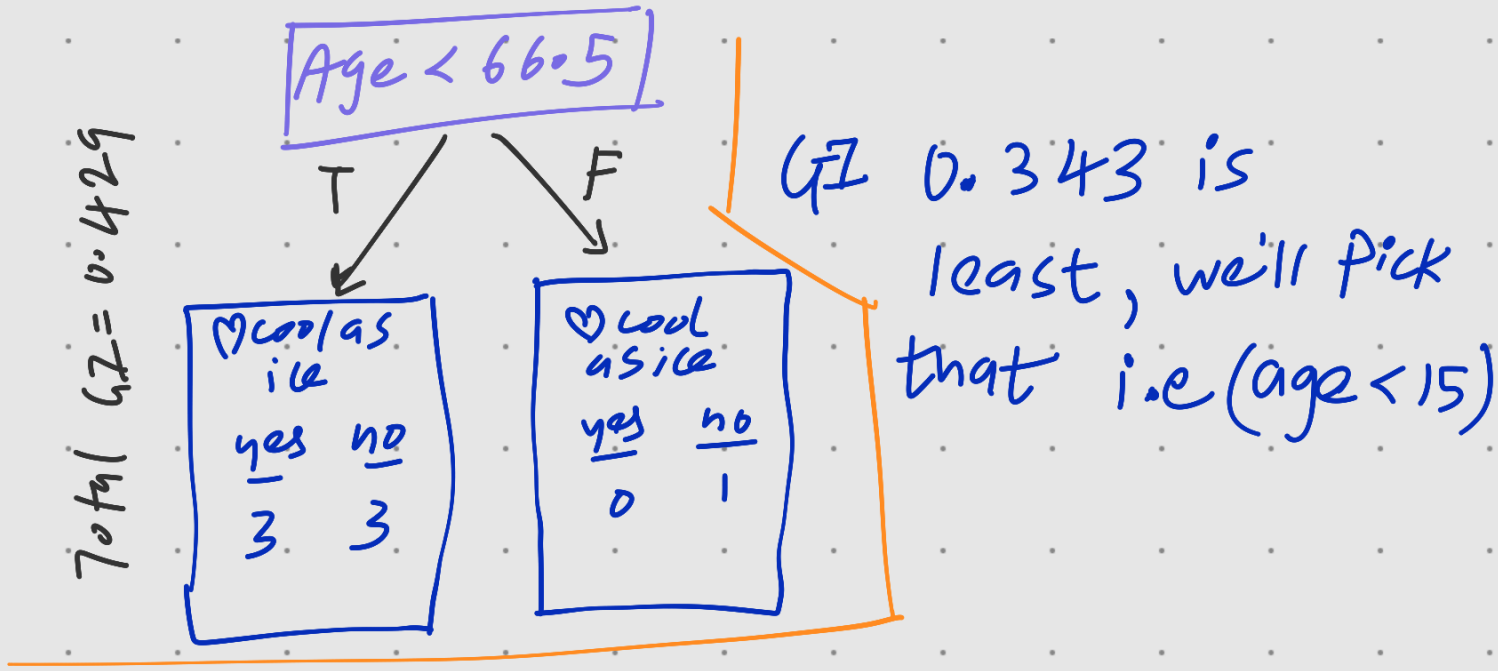
$= 0.476$



Total GI = 0.476
(direct calculation)



Total GI = 0.343

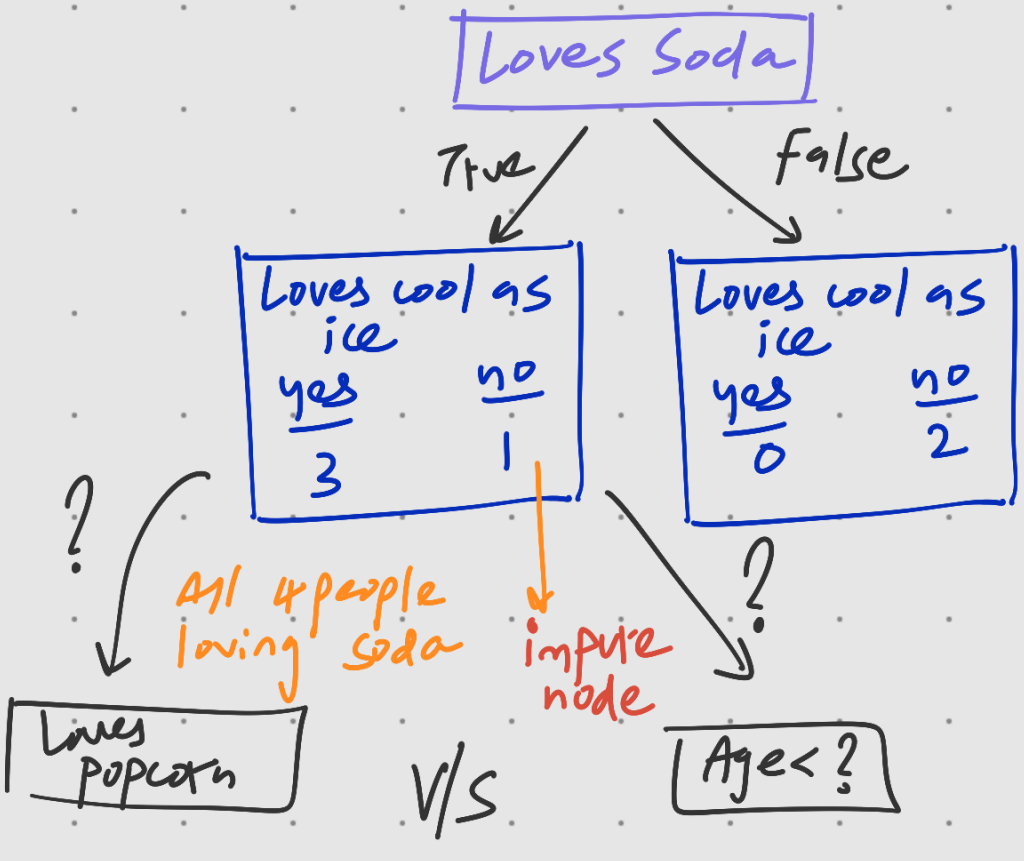


GI 0.343 is least, we'll pick that i.e (age < 15)

Total gini Impurity (loves popcorn) = 0.405

Total gini Impurity (loves soda) = 0.214 ↓ least impurity (i.e Root node)

Total gini Impurity (Age) = 0.343



→ loves popcorn

T F

♡ cool as ice: yes 1, no 1 → GI = 0.5

♡ cool as ice: yes 2, no 0 → GI = 0

Total GI (Popcorn) = $\frac{2}{4}(0.5)$ = 0.25

We'll start by asking 4 people that love Soda if they 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

12.5 · 26.5 · 36.5

→ Age <

Consider age of people loving Soda..

Repeating the process of GI of averages

Age < 12.5

T F

♡ cool as ice: yes 0, no 1 → GI = 0

♡ cool as ice: yes 3, no 0 → GI = 0

Total GI = 0

The least among ages & less than Popcorn too. i.e the branch

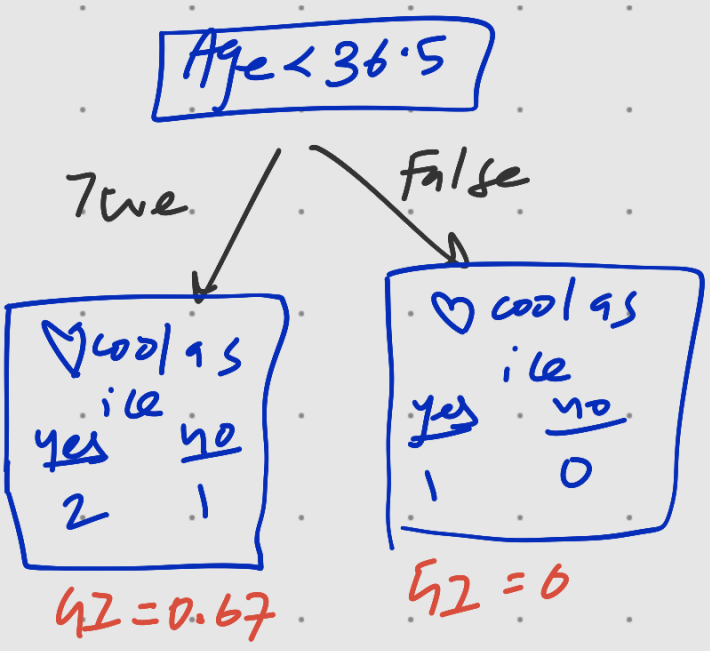
Age < 26.5

T F

♡ cool as ice: yes 1, no 1 → GI = 0.5

♡ cool as ice: yes 2, no 0 → GI = 0

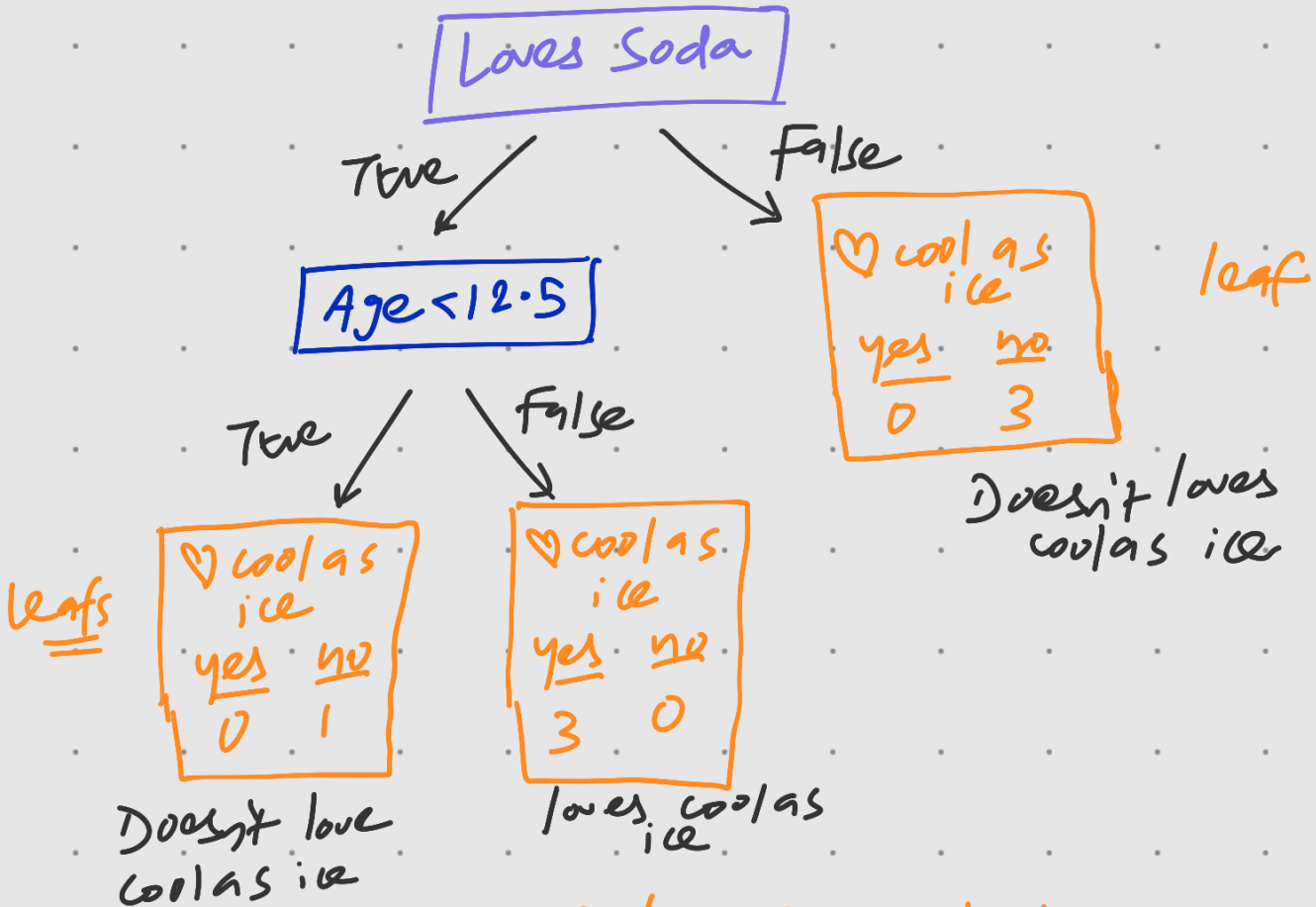
Total GI = $\frac{1}{2}(0.5) = 0.25$



GI = 0.67

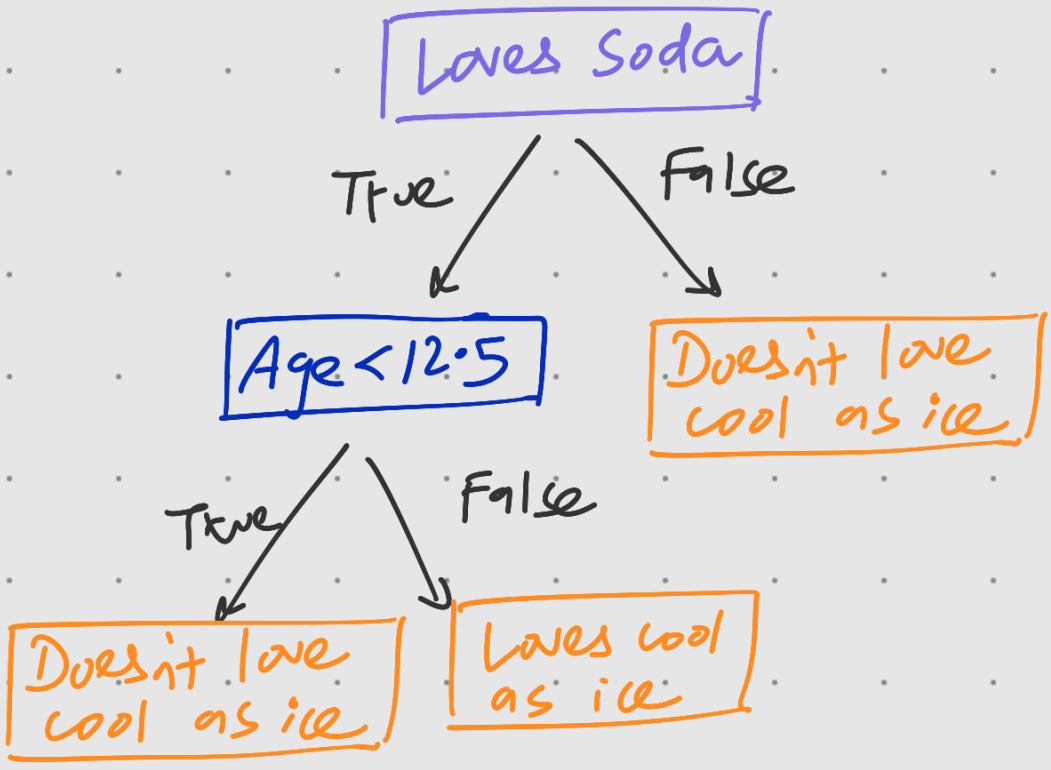
GI = 0

Total GI = $\frac{3}{4}$ (0.67) = 0.5



we need to assign output values for each leaf. i.e category

Final Tree



→ overfitting, because only 1 person reaches this node | Prune

Check for no. of values per node by cross validation.