Classification Tree

Loves Popcorn	Loves Soda	Age	Loves Troll 2
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 will predict if someone loves troll 2 or not.

→ We will solve the problem using classification tree.

⇒ When it comes to tree we need to have a root node.

— So what will be the root node?

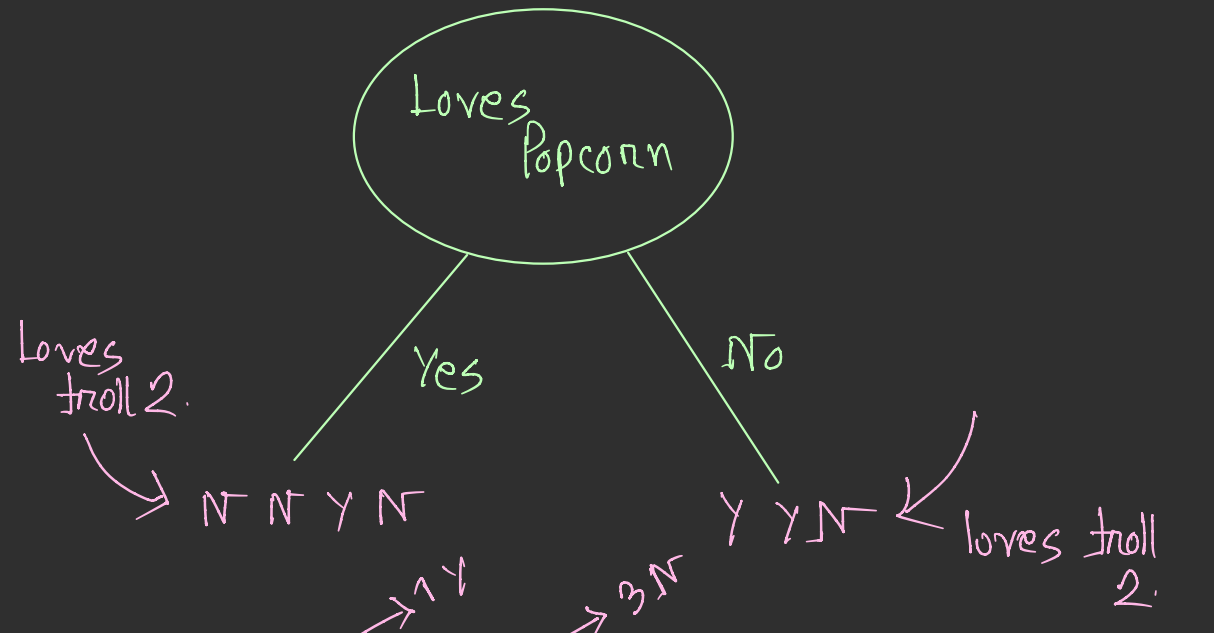
— How do we get this?

⇒ There are multiple ways but we will use "Gini Impurity"

↘ To calculate Gini impurity

— Pick an attribute from the dataset.

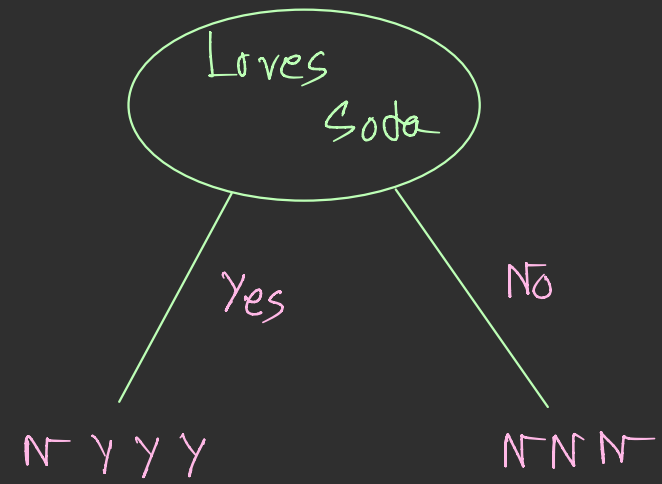
Let's say we pick "Loves Popcorn".



$$\Rightarrow \frac{4}{7}\left\{1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2\right\}+\frac{3}{7}\left\{1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right\}$$

(1Y, 3N)

$$\Rightarrow \text{Gini}(\text{Loves Popcorn}) = 0.405$$



$$\frac{4}{7}\left\{1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right\}+\frac{3}{7}\left\{1-\left(\frac{0}{3}\right)^2-\left(\frac{3}{3}\right)^2\right\}$$

$$\text{Gini}(\text{Loves Soda}) = 0.214$$

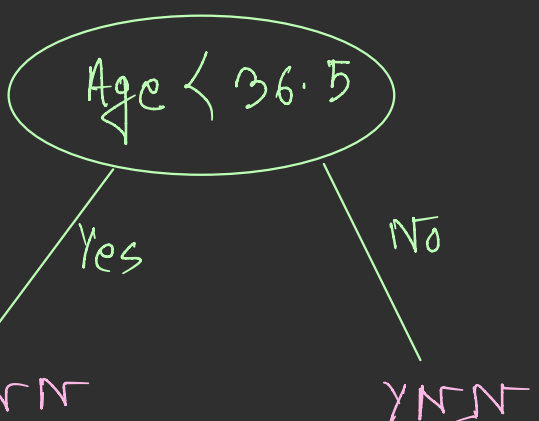
⇒ Now we will do the same for the age attribute.

- But this attribute contains numeric value So the process will be different.

→ first sort the number in ascending order, then compute average of two consecutive values.
also the label accordingly.

So,

7	12	18	35	38	50	83 (Age)
↓	↓	↓	↓	↓	↓	
9.5	15	26.5	36.5	44	66.5	(Avg)



$$G(\text{Age} < 36.5) = \frac{4}{7}\left\{1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right\} + \frac{3}{7}\left\{1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2\right\}$$

$$= 0.476$$

→ Similarly we will find the Gini impurity for all the average values of age.

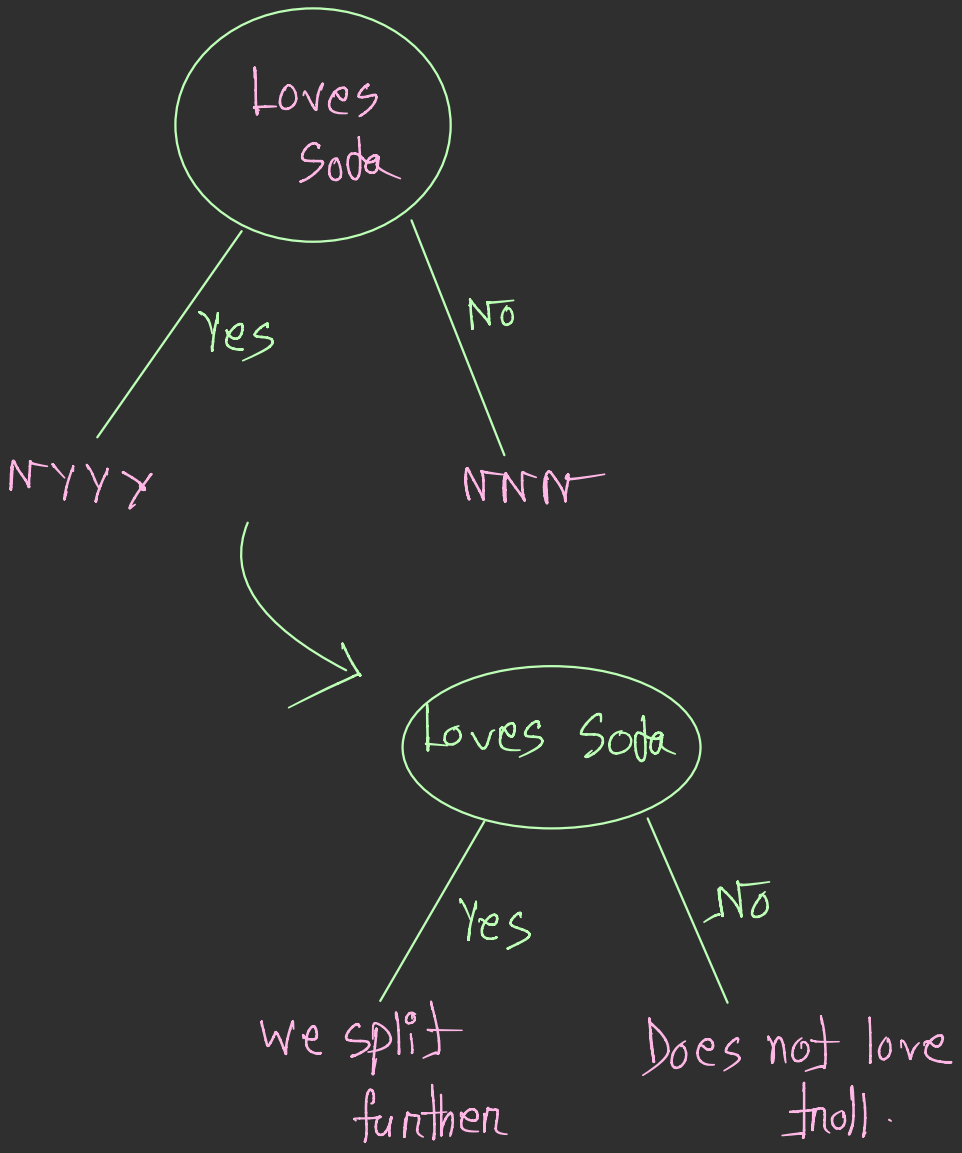
- Among all these A<15 or A<44 will have the lowest impurity. So we pick random one.
- Here we pick

$$G(\text{Age}<15) - 0.343$$

$$G(\text{Loves popcorn}) - 0.405$$

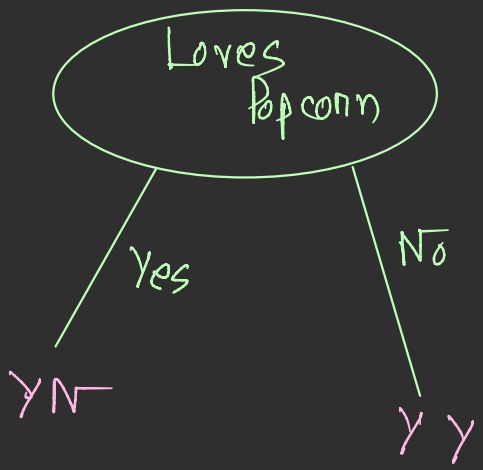
$$G(\text{Loves Soda}) - 0.214$$

⇒ So, for our dataset Loves Soda will be the root node.


Loves Soda
Yes
No
NYYY
NNNN
Loves Soda
Yes
No
We split further
Does not love troll.

⇒ We will continue for Love Soda = Yes

– Gini Impurity of Loves Popcorn



Gini (Loves popcorn) = 0.25

$$\frac{2}{4}\left\{1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right\}+\frac{2}{4}\left\{1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2\right\}$$

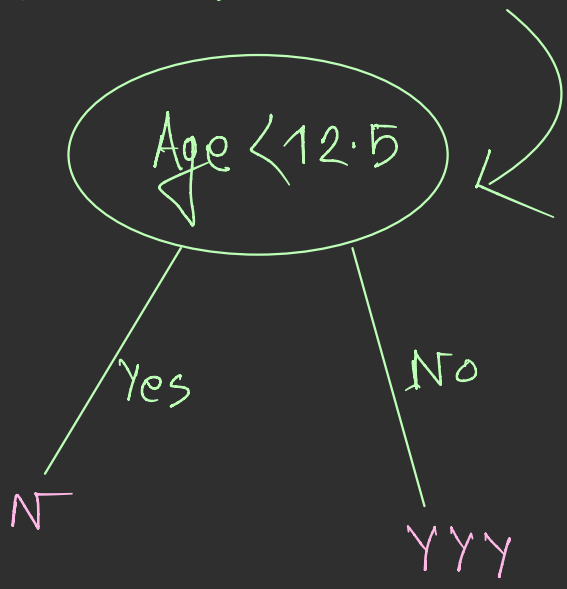
$$= 0.5$$

Age

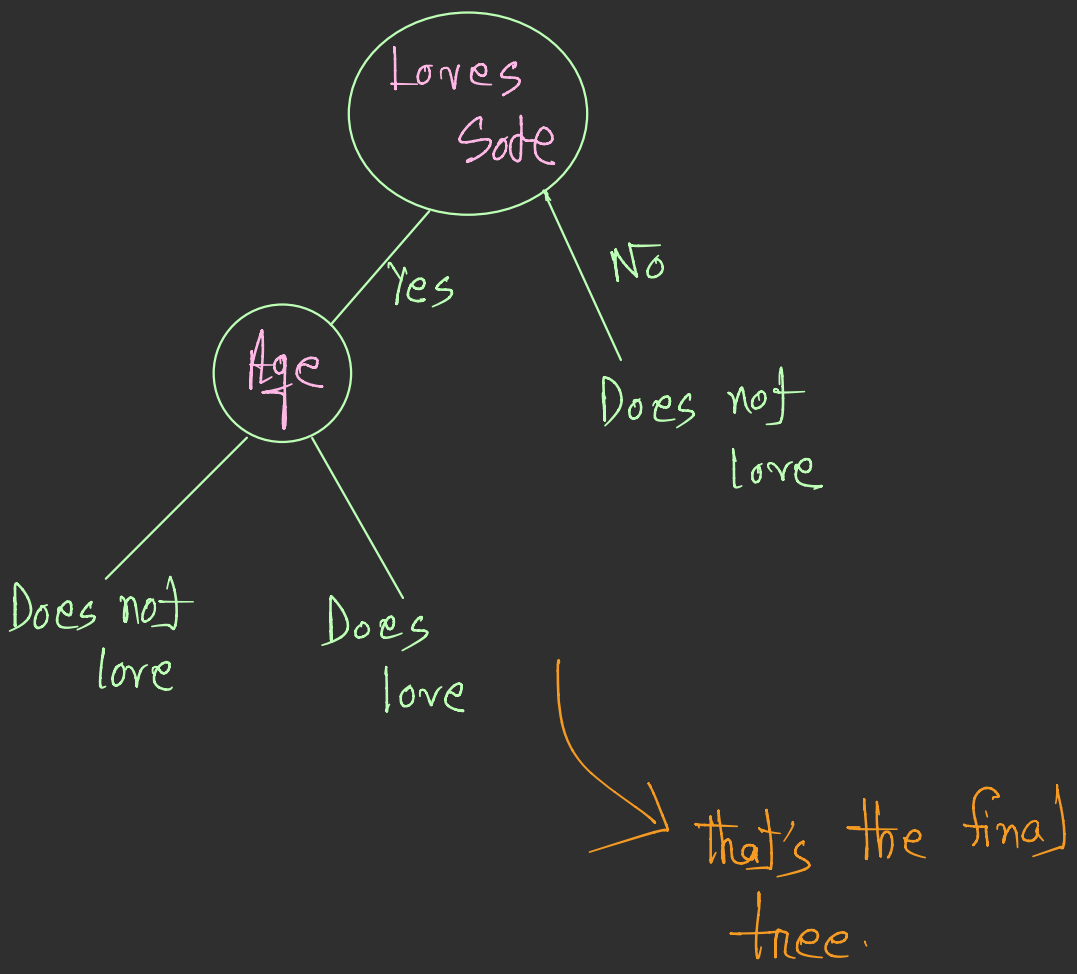
7 18 35 38

12.5 26.5 36.5

⇒ For Age < 12.5 = 0 is the minimum Gini impurity among all the age values.



→ Between Age<12.5 and Loves Popcorn Age will be picked.



that's the final tree.

⇒ Solve the same problem using 'Information Gain'.

$$\text{Entropy} = \sum_{i=1}^{n} -P \log_2 P$$

→ i is the number of lables.

→ Computes impurity

⇒ Entropy (Troll 2)

$$= -P(\text{Yes}) \log_2 P(\text{Y}) - P(\text{N}) \log_2 P(\text{N})$$

$$= -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7}$$

$$= 0.985$$

⇒ Entropy (Loves popcorn = Yes)

$$= -P(\text{Yes}) \log_2 P(\text{Y}) - P(\text{N}) \log_2 P(\text{N})$$

$$= -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.811$$

→ Entropy (Loves popcorn = No)

$$= -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.918$$

→ Information Gain (Loves popcorn)

= Entropy (troll 2) − [P(Loves popcorn = yes) * Entropy (Loves popcorn = yes)] − [P(Loves popcorn = No) * Entropy (Loves popcorn = No)]

$$= 0.985 - \left(\frac{4}{7} * 0.811\right) - \left(\frac{3}{7} * 0.918\right)$$

$$= 0.128$$

→ Information Gain (Loves Soda)

$$= 0.985 - \left[\frac{4}{7} * \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)\right] - \left[\frac{3}{7} * \left(-\frac{0}{3}\underbrace{\log_2\frac{0}{3}}_{0} - \frac{3}{3}\underbrace{\log_2\frac{3}{3}}_{0}\right)\right]$$

$$= 0.985 - 0.463 - 0$$

$$= 0.522$$

- Next calculate information gain for Age like before.
- Among these three attribute Loves Soda has the max information gain.

So, root will be Loves Soda.

→ then continue the spliting untill you meet the conditions.