Decision Tree

In Decision tree we measure Entropy and information Gain of a feature to decide which feature is better to split the data.

Entropy is similar to Gini Impurity. So, Lower the entropy, better the feature. Information Gain is opposite of these two, which helps the decision tree algorithm to identify the most informative feature to make splits that lead to clearer and more predictable classification. So, the more the Information Gain, the better the feature.

Formula —

$$\text{Entropy } H(X) = \sum_{i=1}^{c} -P_i \log_2 P_i$$

where c = number of classes.

$$\text{Information Gain}(S, A) = H(S) - \sum P(S|A) \times H(S|A)$$

Here, S = Dataset and A is a feature of the dataset.

Dataset →

	Loves popcorn	Loves Soda	Age	Loves Troll 2
①	Yes	Yes	7	NO
②	Yes	NO	12	NO
③	NO	Yes	18	Yes
④	NO	Yes	35	Yes
⑤	Yes	Yes	38	Yes
⑥	Yes	NO	50	NO
⑦	NO	NO	83	NO

features — target column

Now we have to find Information Gain for Loves popcorn, Loves soda and Age features to decide which should be the root node of the tree.

First let's find the Entropy of the dataset.

Entropy of Dataset — [If there was another class we would add that as well]

$$\text{info}(D) / H(D) / H(S) = -P(\text{Yes}) \log_2 P(\text{Yes}) - P(\text{NO}) \log_2 P(\text{NO})$$

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

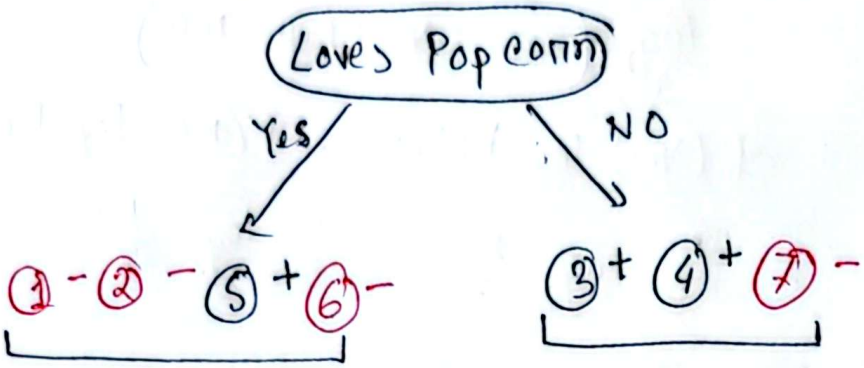
$$= 0.1985$$

$\frac{3}{7}$: there are a total 7 data points in the data set, among those there are 3 'Yes' in the target column

$\frac{4}{7}$: Among 7 data points there are 4 'NO' in the target column

'Loves popcorn feature' →

Let us find Entropies first and then we will find IG of Loves popcorn.



('Yes')

Entropy of ~~left~~ branch →

$$H(\text{Loves popcorn} = \text{Yes}) = -P(\text{Yes}) \log_2 P(\text{Yes}) - P(\text{NO}) \log_2 P(\text{NO})$$

$$= -P(\text{Loves Troll2} = \text{Yes} \mid \text{Loves popcorn} = \text{Yes}) \log_2 (\text{Loves Troll2} : \text{Yes} \mid \text{Loves popcorn} = \text{Yes}) - P(\text{Loves Troll2} = \text{NO} \mid \text{Loves popcorn} = \text{Yes}) \log_2 P(\text{Loves Troll2} = \text{NO} \mid \text{Loves popcorn} = \text{Yes})$$

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

$$= 0.811$$

Among 4 Loves popcorn = Yes, there is 1 Loves Troll2 = Yes

Among 4 Loves popcorn = Yes, there are 3 Loves Troll2 = NO

Entropy of No branch →

$$H(\text{Loves popcorn} = NO) = -P(\text{Troll} = Yes \mid pop = NO)\log_2 P(\text{Troll} = Yes \mid pop = NO) - P(\text{Troll} = NO \mid pop = NO)\log_2 P(\text{Troll} = NO \mid pop = NO)$$

$$= -P(Yes)\log_2 P(Yes) - P(NO)\log_2 P(NO)$$

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

$$= 0.918$$

Among Loves popcorn = NO, there are 2 Loves Troll2 = Yes

Among 3 Loves popcorn = No, there is 1 Loves Troll 2 = NO

∴ Weighted Entropy of Loves popcorn →

$$\text{Info}(\text{Loves popcorn}) = H(\text{Loves popcorn})$$

$$= P(\text{Loves popcorn} = Yes) \times H(\text{Loves popcorn} = Yes) + P(\text{Loves popcorn} = NO) \times H(\text{Loves popcorn} = NO)$$

$$= \frac{4}{7} \times 0.811 + \frac{3}{7} \times 0.918$$

$$= 0.857$$

[If there was another outcome of Lovespopcorn we would add that as well]

There are 4 Lovespopcorn = Yes

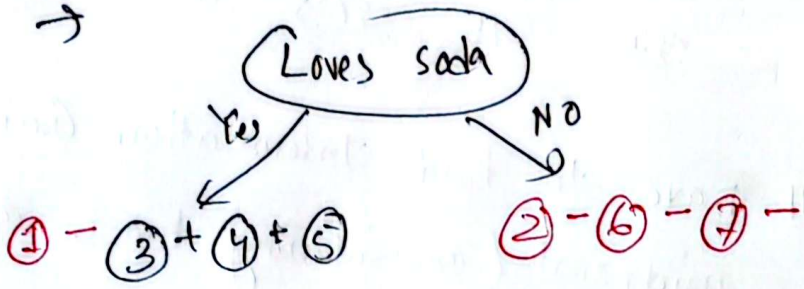
There are total 3 Loves popcorn = NO

Information Gain of Lovespopcorn →

$$\text{Info Gain}(\text{Loves popcorn}) = \text{Info}(D) - \text{Info}(\text{Loves popcorn})$$
$$= 0.985 - 0.857$$
$$= 0.128.$$

'Loves soda feature':

We can directly find the Info (Loves soda) or weighted entropy →



$$\text{Info}(\text{Loves Soda}) = \frac{4}{7}\left[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right]$$

$\frac{4}{7}$ → P(Loves soda = Yes)

$\left[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right]$ → H(Loves Soda = Yes)

$$+ \frac{3}{7}\left[-\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3}\right]$$

$\frac{3}{7}$ → P(Loves soda = NO)

$\left[-\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3}\right]$ → H(Loves soda = NO)

$$= 0.464$$

$$\therefore \text{Info Gain}(\text{Loves Soda}) = \text{Info}(D) - \text{Info}(\text{Loves soda})$$
$$= 0.985 - 0.464$$
$$= 0.521$$

Numerical Feature 'Age' →

As we did in Gini Impurity, we will sort the numerical features/column first in ascending order and the find the midpoints / Averages of adjacent data points.

Sorted order → 7 12 18 35 38 50 83

mid points 9.5 15 26.5 36.5 44 66.5

NOW we will have to find Information Gain for all of this midpoints (considering those as thresholds)

Let us find for Age < 36.5 only →

① - ② - ③ + ④ + ⑤ + ⑥ - ⑦ -

$$Info(Age < 36.5) = \frac{4}{7}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right] + \frac{3}{7}\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right]$$

$\frac{4}{7}$ = P(Age < 36.5); $\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right]$ = H(Age < 36.5)

$\frac{3}{7}$ = P(Age ≥ 36.5); $\left[-\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right]$ = H(Age ≥ 36.5)

$= 0.965$

features. So,

Info Gain(Age < 36.5) = 0.985 - 0.956 = 0.02

Similarly if we find information Gain for all of the mid points we get →

IG (Age < 12.5) IG (Age < 15) IG (Age < 26.5)

0.128 0.292 ✓ 0.02

IG (Age < 36.5) I (Age < 44) IG (Ag < 56.5)

0.02 0.292 ✓ 0.128

Among these the highest Information Gain is IG(Age < 15) = 0.292 and IG(Age < 44) = 0.292. we can pick any one. Let's pick Age < 15.

Info Gain of ~~Att~~ all of the features -

Info Gain (Loves popcorn) = 0.128

Info Gain (Loves Soda) = 0.521

Info Gain (Age < 15) = 0.292

Among these three, Loves Soda has the highest IG. So it will split the dataset better than the other features. So, it will be chosen as root of the tree.

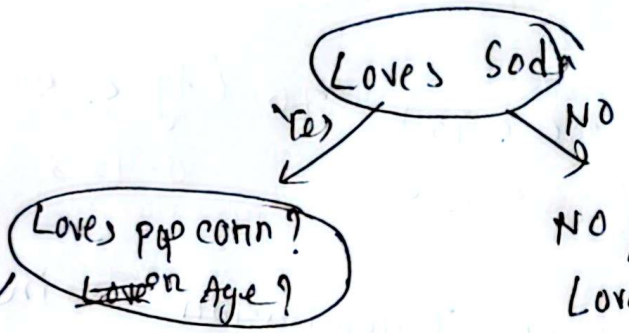
① - ③ + ④ + ⑤ +

impure leaf

② - ⑥ - ⑦ -

pure leaf → we can assign class/ Label here

We need to find another feature to split this branch.



To decide this we will now consider the data points where Loves soda = Yes only.

The new dataset →

	Loves popcorn	Loves Soda	Age	Loves Troll 2
①	Yes	Yes	7	NO
③	NO	Yes	18	Yes
④	No	Yes	35	Yes
⑤	Yes	Yes	38	Yes

Now we will find information gain for Loves popcorn and Age features for this dataset. Loves soda has been used as node already, we will skip it.

Let us find the Entropy of this small dataset Ansf

Entropy of dataset →

$$H(S) / Info(D) = -P(Yes) \log_2 P(Yes) - P(No) \log_2 P(No)$$

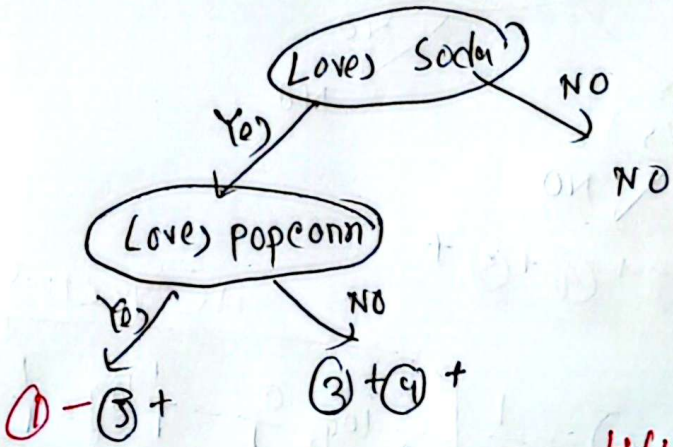
$$= -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4}$$

$$= 0.811$$

There are total 3 Yes in the loves Troll2 / target feature among 4 data points

There is only one No in the target feature.

Information Gain of Loves popcorn →



$$Info(\text{Loves popcorn}) = \frac{2}{4}\left[-\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] + \frac{2}{4}\left[-\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2}\right]$$

$H(\text{Loves popcorn} = Yes)$

$P(\text{Loves popcorn} = Yes)$

$P(\text{Loves popcorn} = No)$

$H(\text{Loves popcorn} = No)$

$$= 0.50$$

$\therefore$ Info Gain (Lovespopcorn) = Info (D) − Info (Lovespopcorn)

$= 0.811 - 0.50$

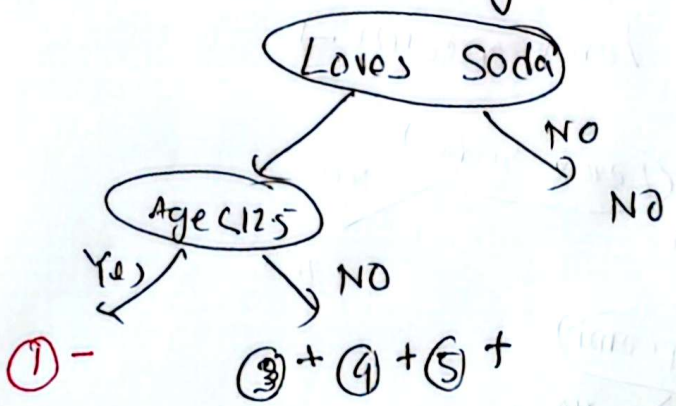
$= 0.311$

Information Gain of Age →

Age in Sorted order → 7 18 35 38

mid points → 12.5 26.5 36.5

Let us find info gain of Age < 12.5



$$\text{Info}(\text{Age}<12.5) = \frac{1}{4}\left[-\frac{0}{1}\log_2\frac{0}{1} - \frac{1}{1}\log_2\frac{1}{1}\right] + \frac{3}{4}\left[-\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{0}\log_2\frac{0}{0}\right]$$

H(Age<12.5) P(Age<12.5) H(Age>12.5) P(Age>=12.5)

$= 0.80$

$\therefore$ Info Gain (Age < 12.5) = Info (D) − Info (Age<12.5)

$= 0.811 - 0.00 = 0.811$

Similarly IG of other thresholds,

Age < 12.5	Age < 26.5	Age < 36.5
0.811	0.311	0.123

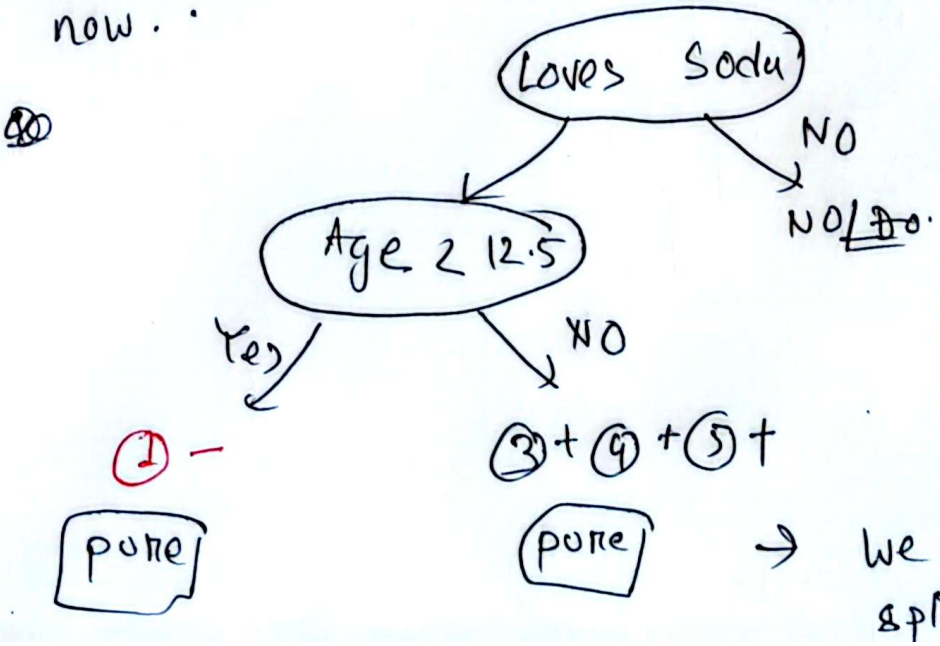
So Info Gain (Age < 12.5) is Greater than other two. we can also decide it by seing the weighted Entropy, Info (Age < 12.5) = 0.00. This is the lowest Entropy possible. So, Age < 12.5 will be chosen

So, Now IG of all features →

Info Gain (Loves popcorn) = 0.311

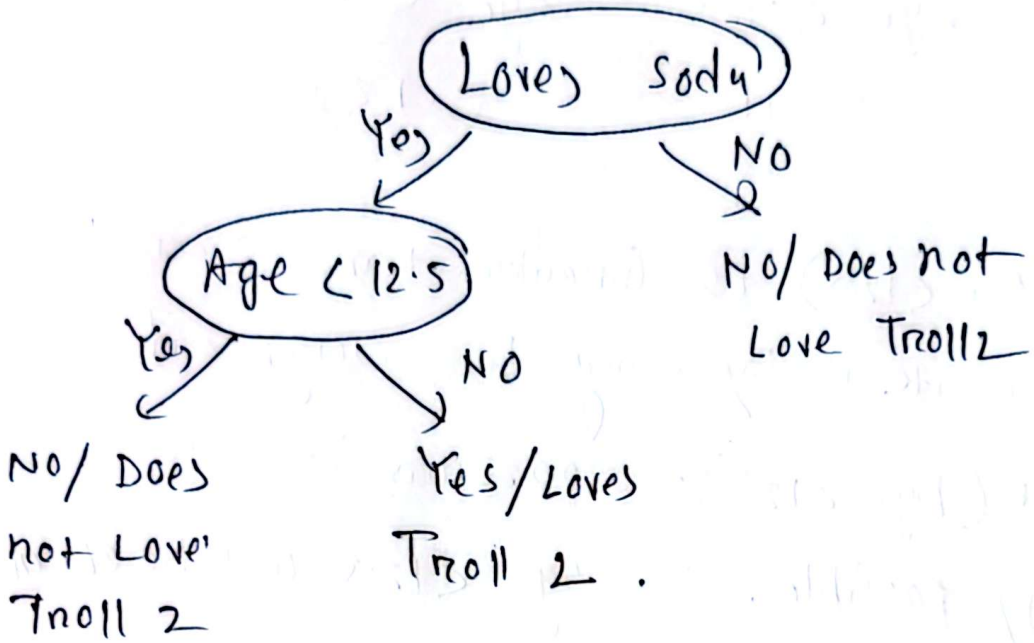
Info Gain (Age < 12.5) = 0.811 winner

Here Age < 12.5 has the highest IG. So it will be chosen as the node of the tree now..



→ we do not need to split any further.

∴ The Final Decision Tree →



This is the same tree we found in classification tree (Gini impurity) Process.