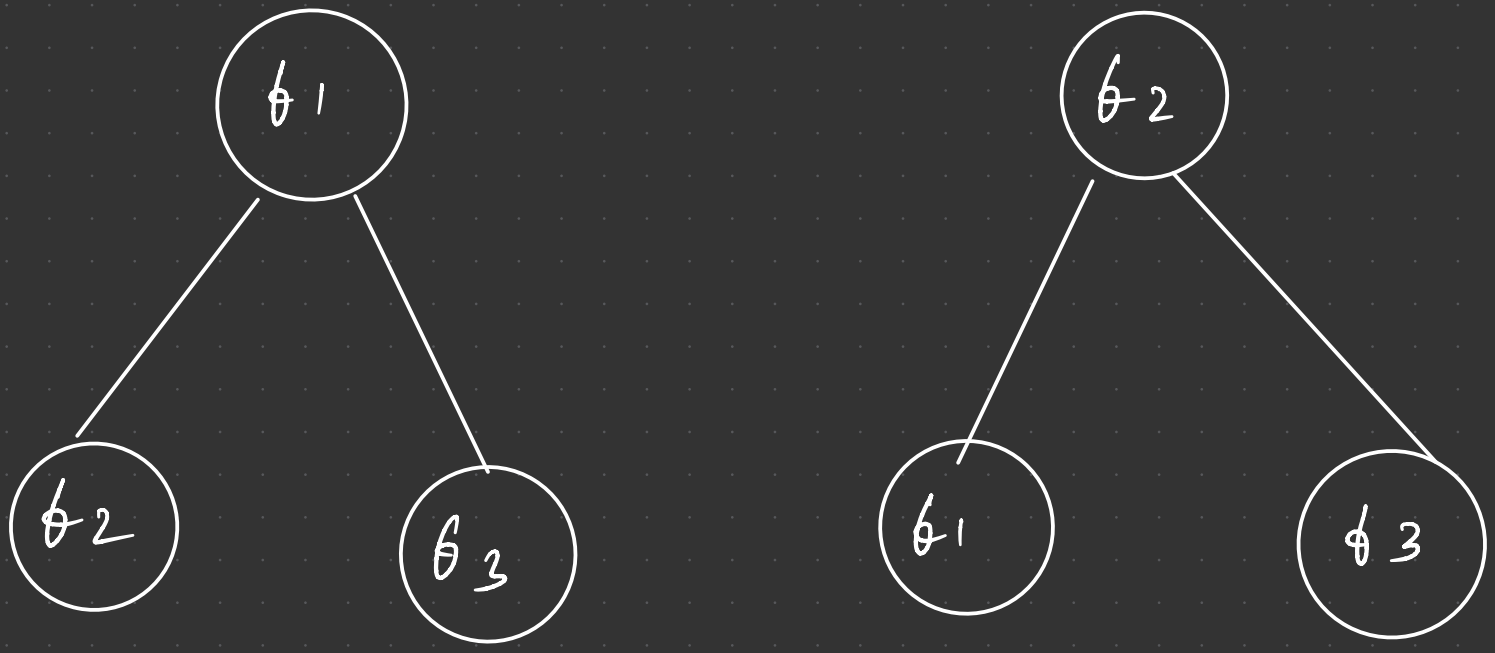
Information Gain

Say for example I have below dataset

θ_1	θ_2	θ_3	O/P

now, I can do the split as below



how do we determine, from which feature we need to start the split, for this we can use Information Gain.

Formula :

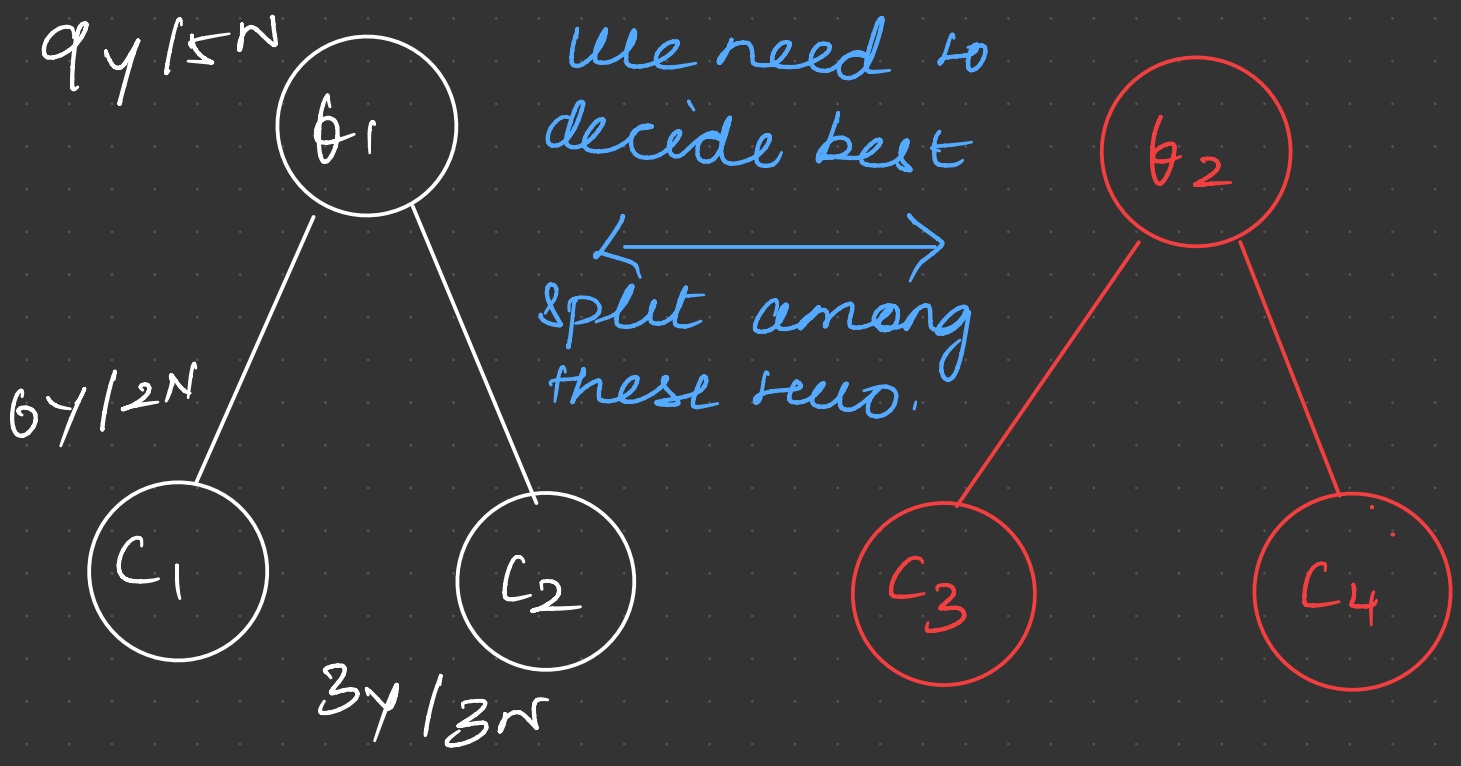
Entropy of the root node.

$$Gain(S, \theta_1) = H(S) - \sum_{v \in val} \frac{|S_v|}{|S|} H(S_v)$$

Entropy of the categories

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

Lets consider, we have below splits



Lets take the first split

$$-\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$\approx 0.94 \Rightarrow$ Entropy of the root node.

$$H(C_1) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$$

$= 0.81 \Rightarrow$ Entropy of category 1

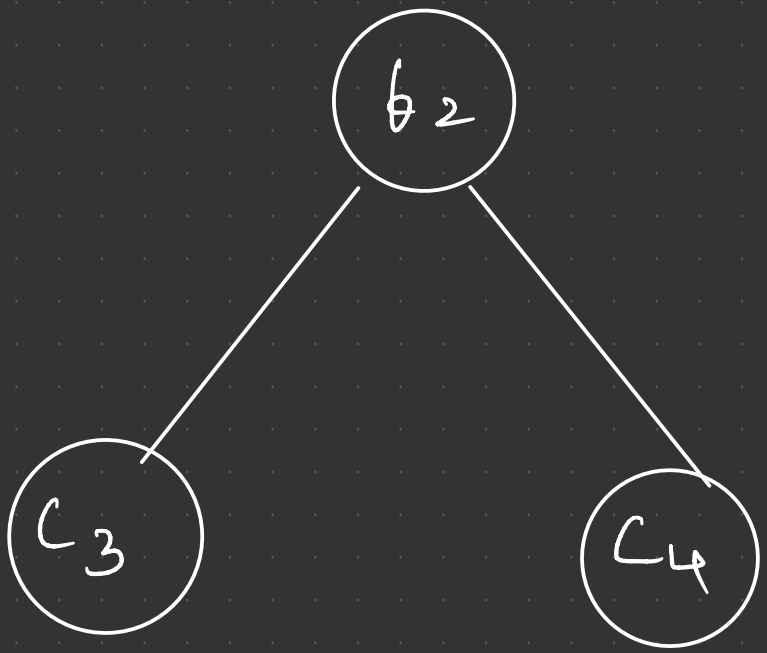
$$H(C_2) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{3}{3}\log_2\frac{3}{3}$$

$= 1 \Rightarrow$ Perfect impure.

↳ Entropy of category 1

$$Gain(S, \phi_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$\boxed{Gain(S, \phi_1) = 0.049}$$



Lets consider gain of this split

$$Gain(S, \phi_2) = 0.051$$

this gain is slightly greater than ϕ_1 split.

to summarize.

$Gain(S, \phi_1) = 0.049$

$Gain(S, \phi_2) = 0.051$

In this case we need to start the split from ϕ_2 as it has higher information gain.

This is how the Information gain is calculated.