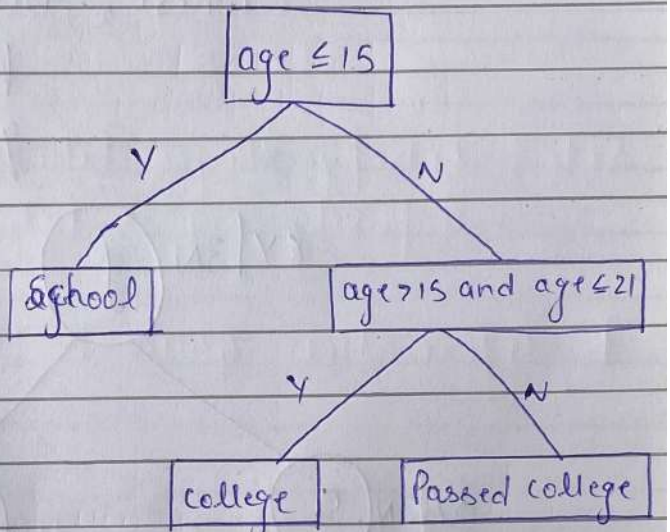
Decision Tree

→ It is a supervised machine learning algorithm which is used to solve both classification and regression problem by splitting the dataset repeatedly till leaf node achive.

```
age = 14
if age ≤ 15:
    print('school')
elif age >15 and age ≤21:
    print('college')
else:
    print('Passed college')
```



Q But how the splitting is happening and on which condition / feature ?

→

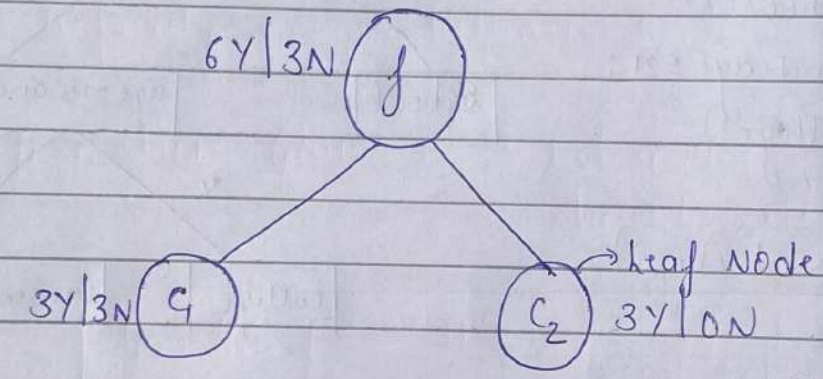
① Purity ⟶ Pure split ??
- ↳ Entropy
- ↳ Gini Impurity

② Gain / Information Gain ⟶ How the feature are selected ?

* Entropy : → Degree of disorganization or randomness

$$H(S) = -p \log_2 p - q \log_2 q$$

where, $H(S)$ → Entropy
p → probability of success
q → probability of failure.



• Entropy of C_1 :

$$H(C_1) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

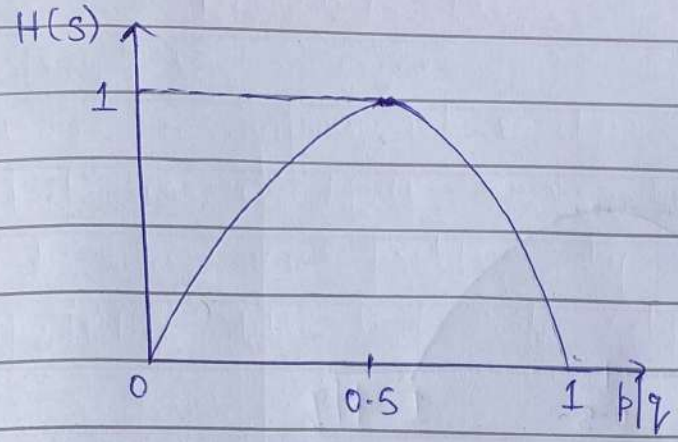
$$= 1 \rightarrow \text{Impure split}$$
↳ Need more splitting

• Entropy of C_2 :

$$H(C_2) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$$

$$= -1 \log 1$$

$$= 0 \rightarrow \text{Pure split}$$
↳ leaf Node
↳ No more splitting

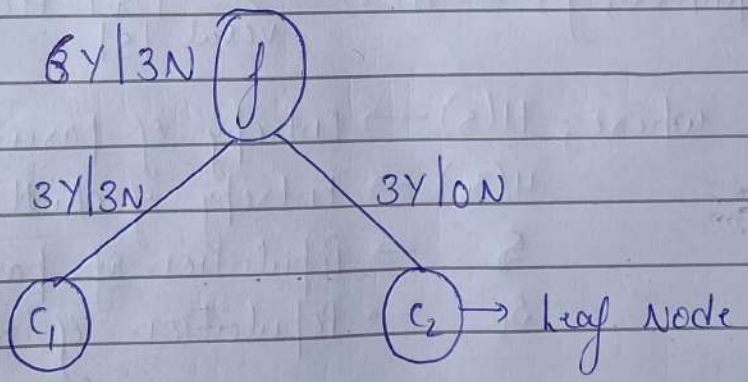


Entropy Graph

* Gini Impurity : → purity of split

$$G.I = 1 - \sum_{i=1}^{n} (p)^2$$

$$\Rightarrow \boxed{G.I = 1 - [p^2 + q^2]}$$



• Gini Impurity of C_1 :

$$GI = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right]$$

$$= 1 - \frac{1}{2} = 0.5 \rightarrow \text{Impure split}$$

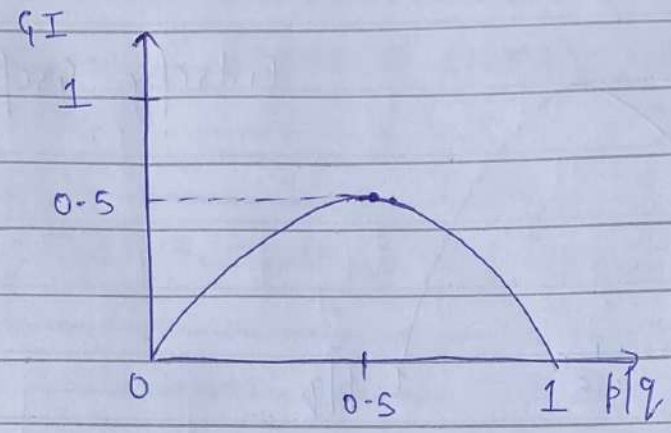
↳ Need more splitting

• Gini Impurity of C_2 :

$$GI = 1 - \left[\left(\frac{3}{3}\right)^2 + 0^2\right] = 1 - 1$$

$$= 0 \longrightarrow \text{Pure split} \rightarrow \text{leaf Node}$$

↳ No more splitting



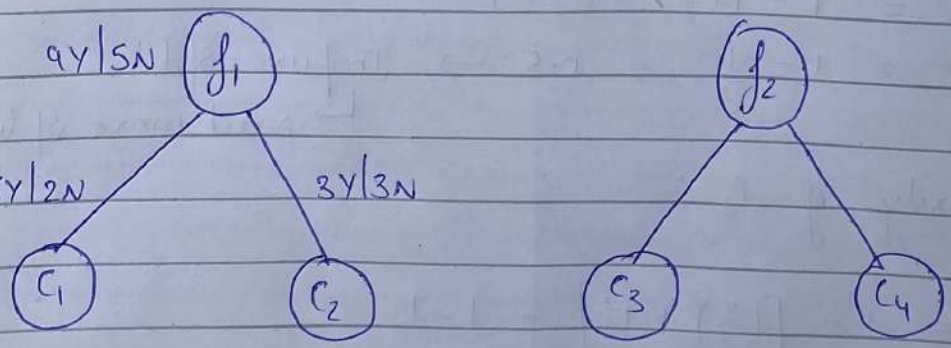
Note : Entropy ranges between 0 to 1 whereas gini impurity ranges between 0 to 0.5.

* Information Gain : → Determine which feature should select for splitting.

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

where, $H(S)$ → Entropy of parent node
$H(S_v)$ → Entropy of child node
S → Population of parent node
S_v → Population of child node

Eg : we have two features (f_1 and f_2). which feature should we select for splitting.



• Entropy :

$$H(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} \;\Big\}\rightarrow \text{Parent Node}$$

$$= 0.94$$

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

$$= 0.81$$

$$H(C_2) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} \;\Big\}\rightarrow C_2$$

$$= 1$$

$\Big\}\rightarrow$ ~~Root~~ ~~N~~ Child Nodes

• Information Gain :

$$Gain(S, f_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$= 0.049$$

Similarly, let's suppose information gain for f_2

$$Gain(S, f_2) = 0.052.$$

$\therefore Gain(S, f_2) > Gain(S, f_1)$

So, feature f_2 will be selected for splitting