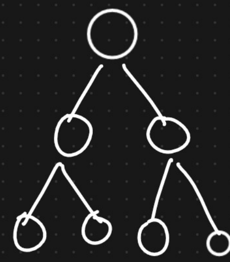
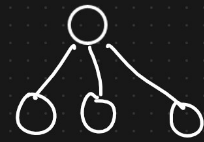
Decision Tree Classifier

Decision Tree Classifier → ID3
→ CART ✓

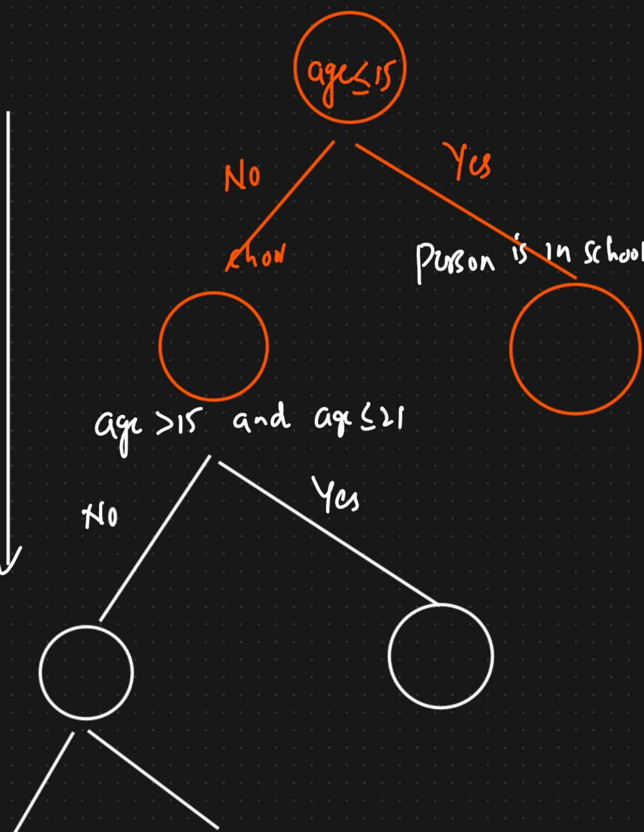




a) Entropy and Gini Index → Purity Split

b) Information Gain → features to select for DT construction

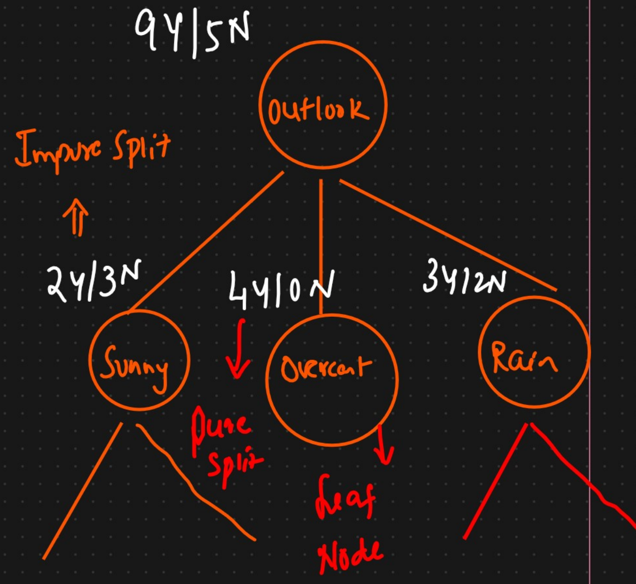
```
age = 14

if (age ≤ 15):
    Print ("The person is in School)
elif (age > 15 and age ≤ 21):
    Print ("The person may be college)
else:
    Print ("The person has passed")
```



Dataset

Binary classification

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



① Purity → Pure or Impure Split
- → Entropy
- → Gini Impurity

② What feature you need select for splitting → Information Gain

1

0

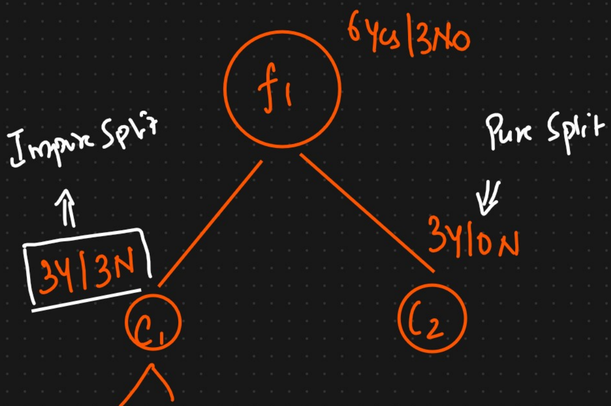
{Binary classification}

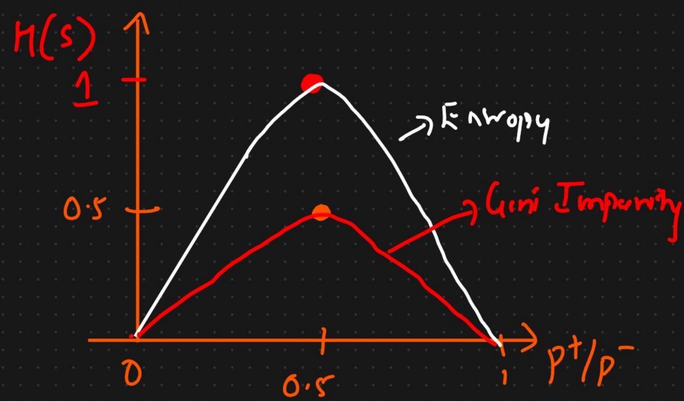
1) Entropy

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

② Gini Impurity

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





$$H(C_1) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

$$= -\frac{3}{6} \log_2 3/6 - \frac{3}{6} \log_2 3/6$$

$= 1 \Rightarrow$ Impure Split

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

$= -1 \log_2 1 \Rightarrow 0 \Rightarrow$ Pure Split

② Gini Impurity

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

$= 1 - ((p_+)^2 + (p_-)^2)$

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

$= 0.5 \Rightarrow$ Impure Split

3Y/0N

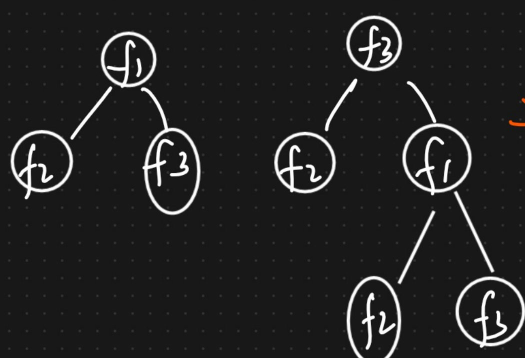
$= 1 - \left(\left(\frac{3}{3}\right)^2\right)$

$= 1 - 1$

$= 0 \Rightarrow$ Pure Split

$f_1 \quad f_2 \quad f_3$

Decision Tree Split



$\Downarrow$

$\Rightarrow$ Information Gain }

Information Gain

$f_1 \quad f_2 \quad f_3 \quad O/P$

14 = 9Y/5N

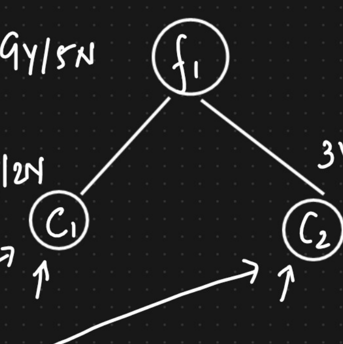
Entropy of the root node

$$\text{Gain}(S, f_1) = \boxed{H(S)} - \sum_{v \in Val} \frac{|S_v|}{|S|} H(S_v)$$

Root Node

8 = 6Y/2N

3Y/3N = 6



$\Downarrow$

Impure split

$\longleftrightarrow$

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

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

≈ 0.94

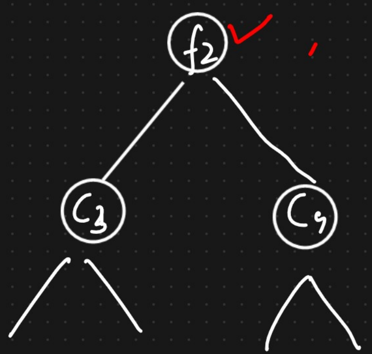
$H(C_1) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$

$\boxed{H(C_1) = 0.81}$

$\boxed{H(C_2) = 1}$

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

$\boxed{\text{Gain}(S,f_1) = 0.049}$



$\boxed{\text{Gain}(S,f_2) = 0.051} > \boxed{\text{Gain}(S,f_1) = 0.049}$

Information Gain is Basically calculated.

Entropy Vs Gini Impurity

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

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

$\boxed{O/P = 3 \text{ categories}}$

$H(S) = -P_{c_1}\log_2 P_{c_1} - P_{c_2}\log_2 P_{c_2} - P_{c_3}\log_2 P_{c_3}$

Whenever dataset is small $\rightarrow$ Entropy

large $\rightarrow$ Gini Impurity