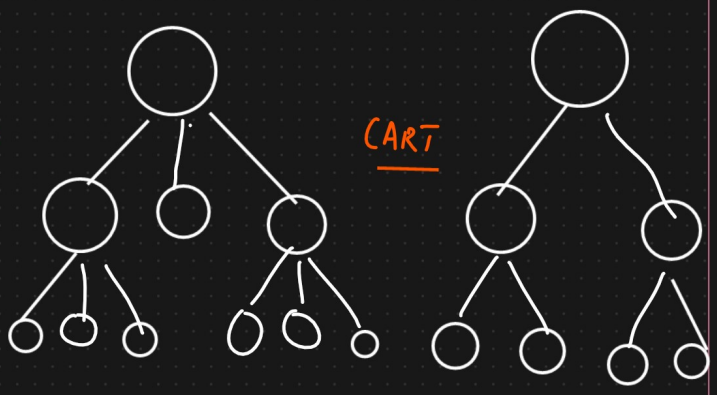
Decision Tree

① Decision Tree classifier [classification]

② Decision Tree Regressor [Regression]

ID 3

CART



Decision Tree Classifier =>

Two types

① ID3 [Iterative Dichotomiser 3]

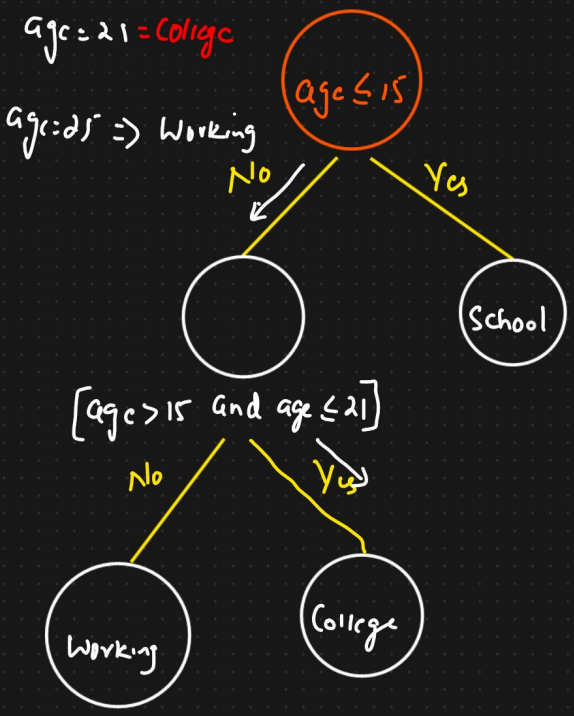
② CART [Classification And Regression Tree]

age = 14

```
if (age ≤ 15):
    Print ("School").
clif (age > 15 and age ≤ 21):
    Print ("College")
else:
    Print ("Working")
```

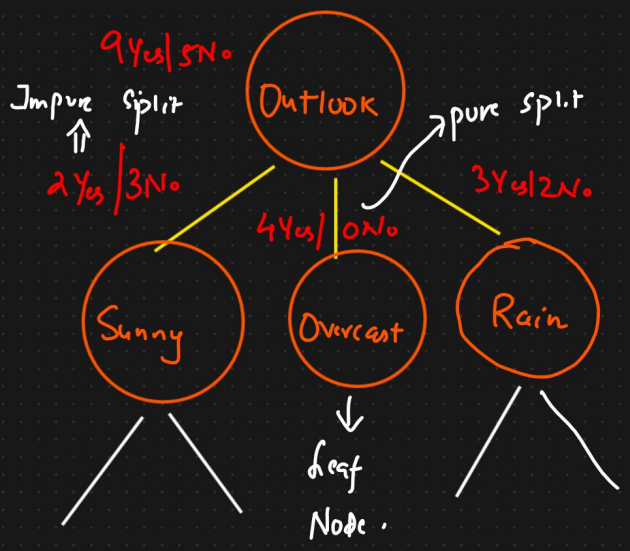
age = 21 = College

age = 25 => Working



Dataset → Predict Play Tennis OR Not

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 check : Pure Split or Impure Split

- Entropy
- Gini Impurity

} Measure of purity. ✓

② What feature you need to select to start the split ? → Information Gain

Binary classification

① Entropy

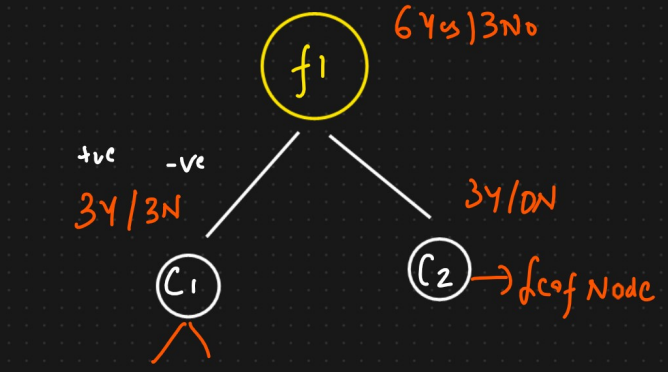
$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

P_+ = probability of positive category

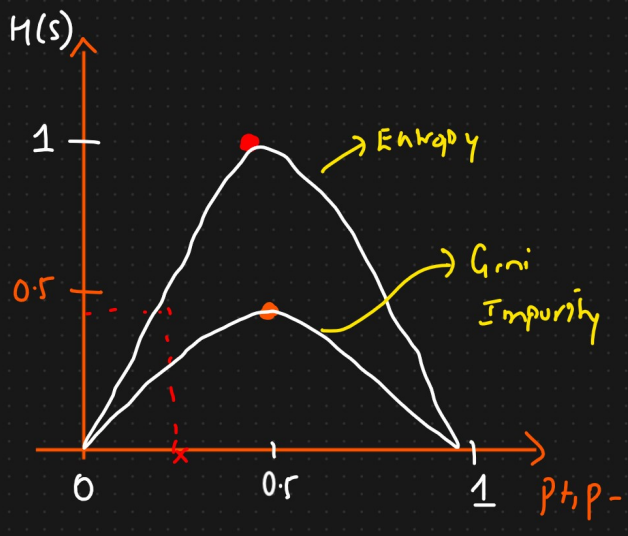
P_- = probability of negative category

② Gini Impurity

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



$$H(C_1) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$
$$= -\frac{3}{6} \log_2 \left(\frac{3}{6}\right) - \left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right)$$



$= -\frac{1}{2}\log_2(1/2) - (1/2)\log_2(1/2)$.

$= 1$ ⇒ Impure Split

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

$= 0$ ⇒ Pure Split

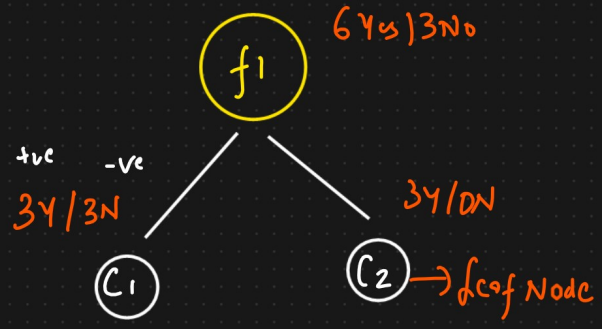
Multiclass

C_1 C_2 C_3
Yes | No | May Be

$H(S) = -P_{C_1}\log_2 P_{C_1} - P_{C_2}\log_2 P_{C_2} - P_{C_3}\log_2 C_3$

② Gini Impurity

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



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

$= 1 - \left[\left(\frac{3}{6}\right)^2 + (3/6)^2\right]$

$= 1 - \left[\frac{1}{4} + \frac{1}{4}\right]$

$= 1/2 = 0.5$ ⇒ Impure Split

$1 - [(3/3)^2 + (0/3)^2]$

$= 1 - 1$

$= 0/1$ ⇒ Pure Split

② What feature you need to select to start the split? → **Information Gain**

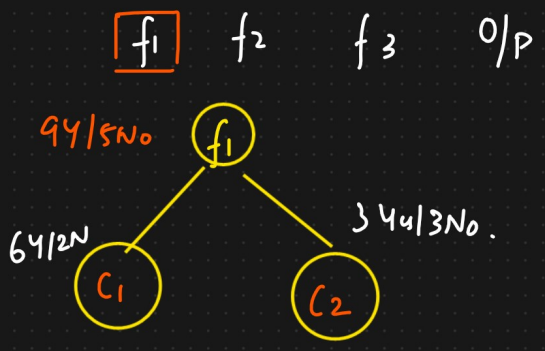
f1 f2 f3 O/P

f1 f2

Information Gain

Entropy of the root node

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



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

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

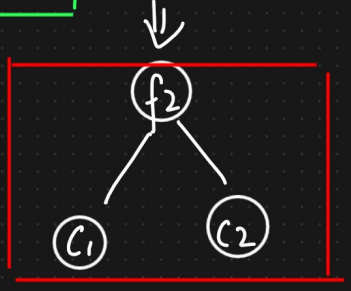
$$\approx 0.94 //$$

$$H(C_1) = -\frac{6}{8} \log_2 (6/8) - \frac{2}{8} \log_2 \frac{2}{8} \approx 0.81$$

$$H(C_2) = 1 //.$$

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

$$Gain(S, f_1) = 0.049$$



$$Gain(S, f_2) = 0.051 > Gain(S, f_1) = 0.049$$

Information gain is more when we split using f2.

Entropy Vs Gini Impurity

Whenever dataset is small → Entropy

dataset is huge → Gini Impurity