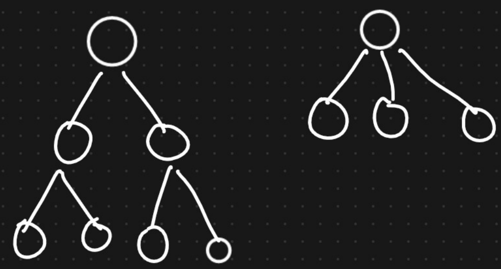
Decision Tree Classifier: Used for both classification and Regression problem.

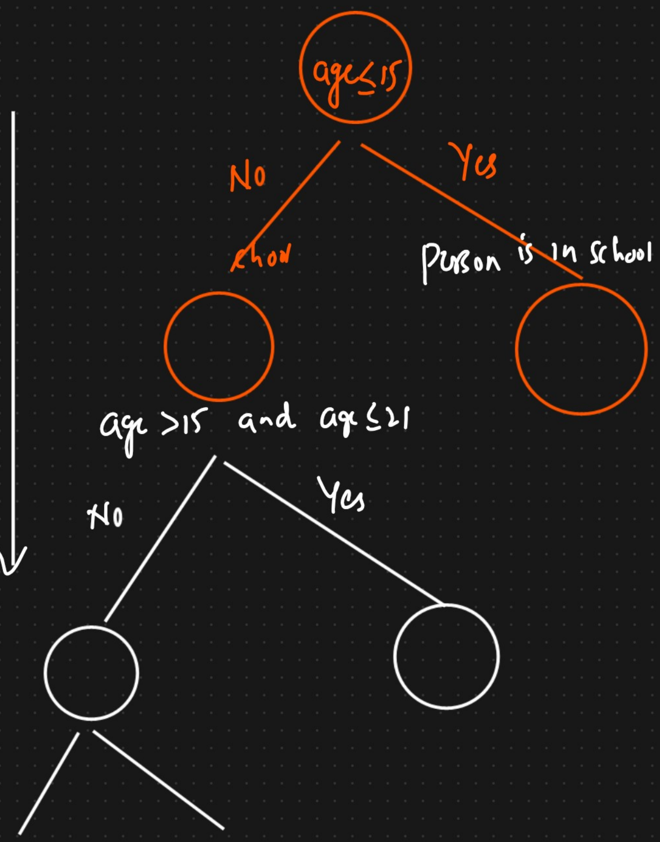


1. Decision Tree Classifier → ID3, CART ✓

a) Entropy and Gini Index → Purity Split

b) Information Gain → features to select for DT construction

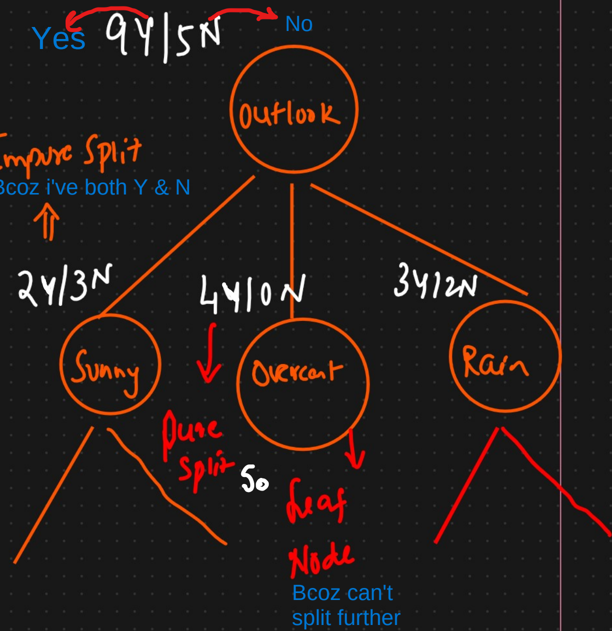
```
age = 14
if (age ≤ 15):
    Print ("The person is in School")
elif (age > 15 and age ≤ 21):
    Print ("The person may be college")
else:
    Print ("The person has passed")
```



Dataset

Binary Classification Problem

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

} To check pure or impure split Mathematically

② What feature you need select for splitting → Information Gain }

1
0

{Binary Classification}

1) Entropy $= -\sum p_i \log p_i$

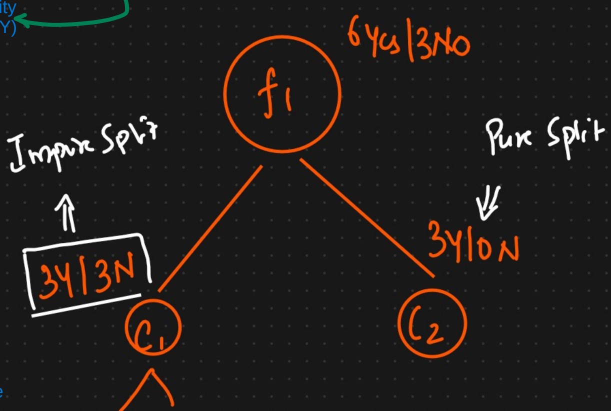
$$H(S) = -(P_+ \cdot \log_2 P_+) - (P_- \cdot \log_2 P_-) - \ldots\ldots$$

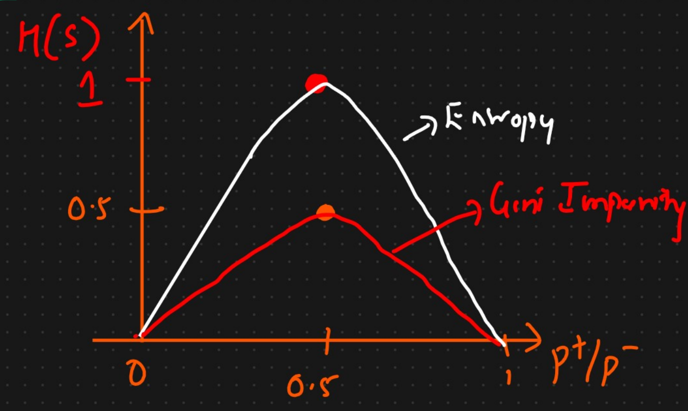
P+ means probability of being positive(1/Y)

p- means probability of being negative(0/N)

② Gini Impurity

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





Entropy of C1 node

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

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

$= 1 \Rightarrow$ Impure Split

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

$$= -1\log_2 1 \Rightarrow 0 \Rightarrow \text{Pure Split}$$

② Gini Impurity

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

$$= 1 - \left((p_+)^2 + (p_-)^2\right)$$

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

$$= 0.5 \Rightarrow \text{Impure Split}$$

3Y/0N

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

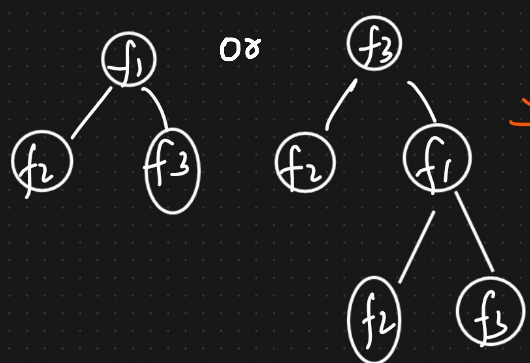
$$= 1 - 1$$

$$= 0 \Rightarrow \text{Pure Split}$$

How to decide which feature is select to make this decision tree split ?

f_1 f_2 f_3

Decision Tree Split



$\Downarrow$

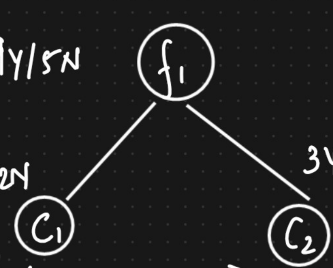
$\Rightarrow$ Information Gain

Information measure the effectiveness of a feature in reducing uncertainty (or impurity) in a dataset.

Information Gain

f_1	f_2	f_3	O/P

14 = 9Y/5N



8 = 6Y/2N

3Y/3N = 6 $\Downarrow$ Impure split

$\longleftrightarrow$

Entropy of the root node

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

Root Node

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

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

$$\approx 0.94$$

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

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

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

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

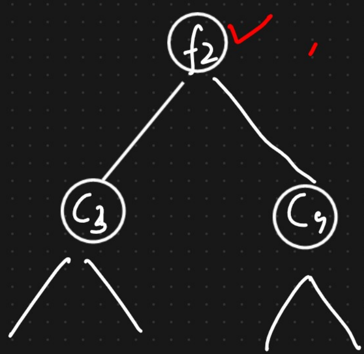
$|S_v|$ = Number of samples child node S_v.

$|S|$ = Total number of samples in the parent node

$H(S_v)$ = Entropy of S_v node.

k = Number of child nodes created by the split.

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



Note:-

Higher Information Gain

↓

Feature splits the data effectively and reduces impurity.

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

Information Gain is Basically calculated.

When should use Entropy vs Gini impurity ?

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 → Entropy

large → Gini Impurity