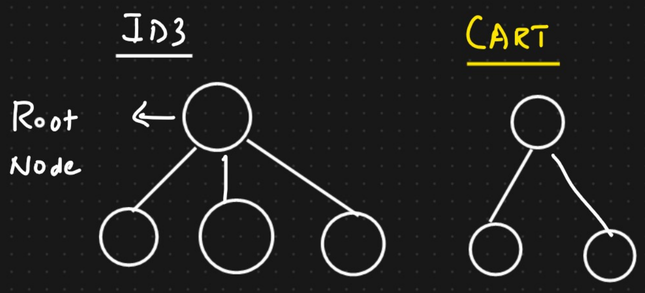
Decision Tree

① Decision Tree Classifier [classification]

② Decision Tree Regressor [Regression]



Decision Tree Classifier

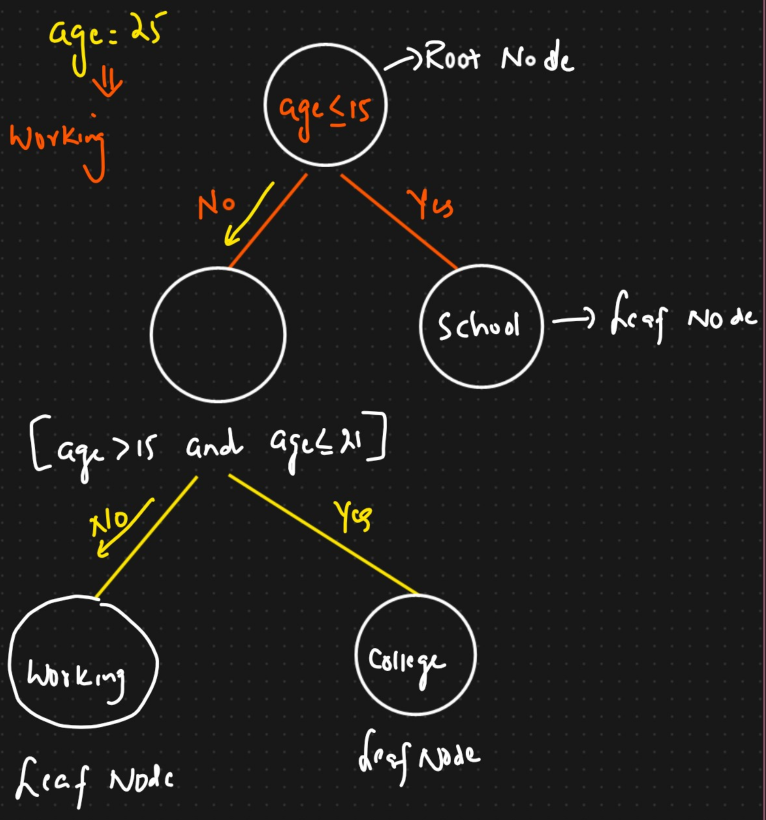
Two techniques

① ID3 [Iterative Dichotomisu 3]

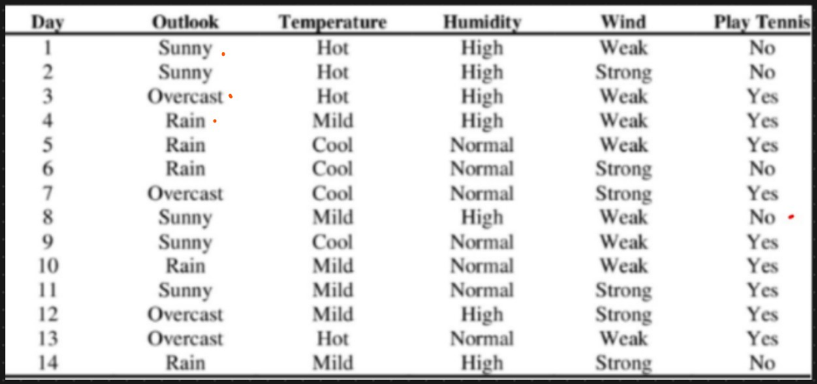
② CART [Classification And Regression Tree]

Multinested if else clause

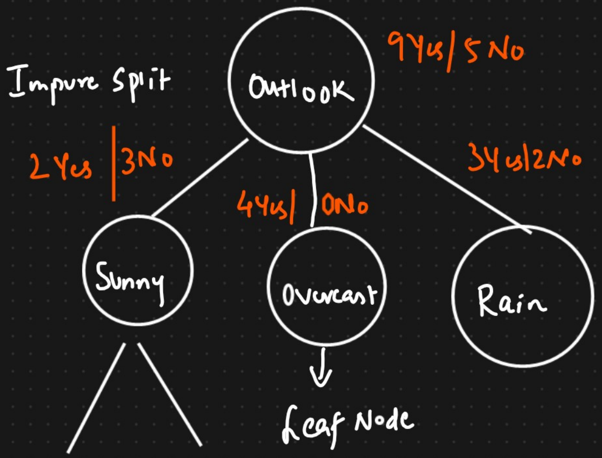
```
age = 14

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



Dataser → Problem statement



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 Split 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.

① Purity Check

Binary classifiation

① Entropy

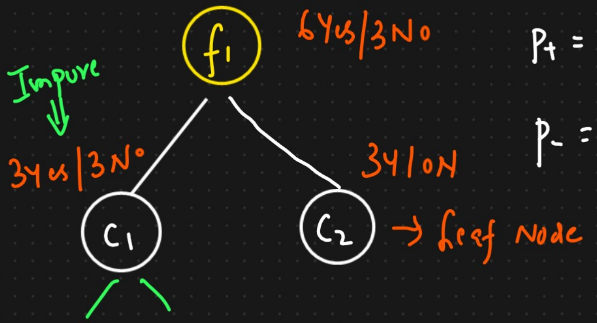
② Gini Impurity

$$GI = 1 - \sum_{i=1}^{n} (p)^2$$

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

P_+ = probability of positive category

P_- = " " negative category



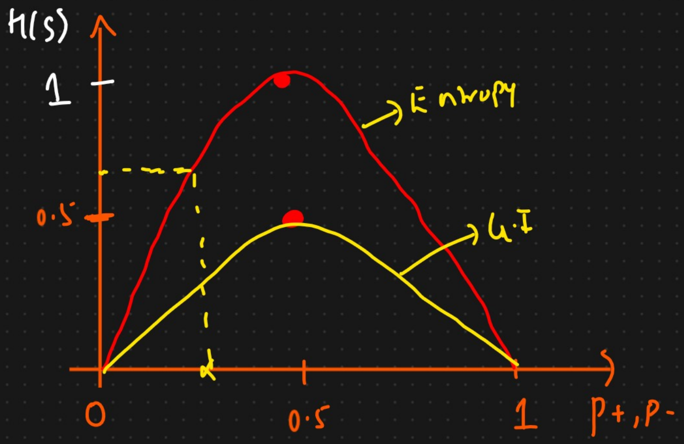
$$P_+ = \frac{3}{6} = \frac{1}{2}$$

$$P_- = \frac{3}{6} = \frac{1}{2}$$

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

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

$$\boxed{H(C_1) = 1} \Rightarrow \text{Impure Split}$$



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

$$H(C_2) = 0 \Rightarrow \text{Pure Split}$$

② Gini Impurity

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

$$G.I(C_1) = 1 - \left[(P_+)^2 + (P_-)^2\right]$$

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

$$= 1 - \frac{1}{2} = \underline{\underline{0.5}} \Rightarrow \text{Impure Split}$$

3Yes/0No

$$G.I(C_2) = 1 - \left[\left(3/3\right)^2 + \left(0/3\right)^2\right]$$

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

Multiclass Classification Problem : 3 categories In O/P

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

$$G.I = 1 - \left[(P_{C_1})^2 + (P_{C_2})^2 + (P_{C_3})^2\right]$$

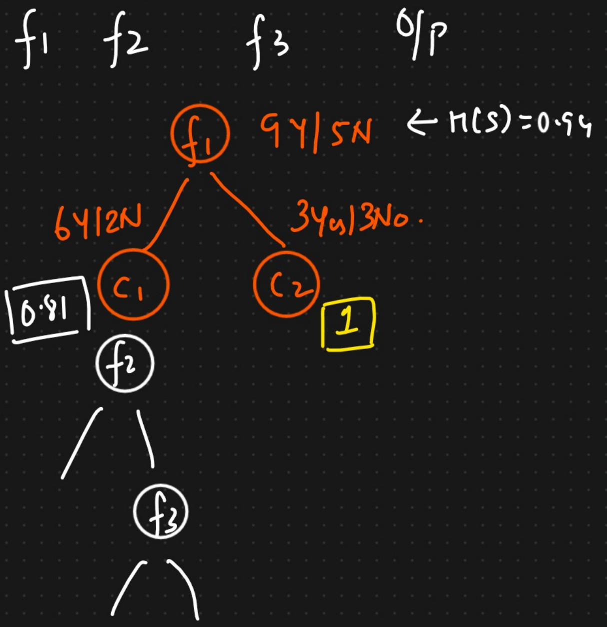
② Information Gain → Which feature to Select to start the split?

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

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

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

$$\approx 0.94$$



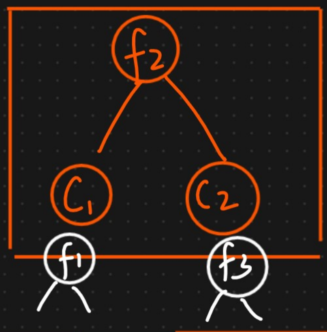
$$H(C_1) = -\frac{6}{8} \log_2\left(\frac{6}{8}\right) - \frac{2}{8} \log_2 \frac{2}{8} \approx \boxed{0.81}$$

$$H(C_2) = 1$$

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

$$= 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

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



=) Information Gain = 0.051

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

We need to Splitting by using f2 features

Entropy vs Gini Impurity

When dataset is small → Entropy [log formula]

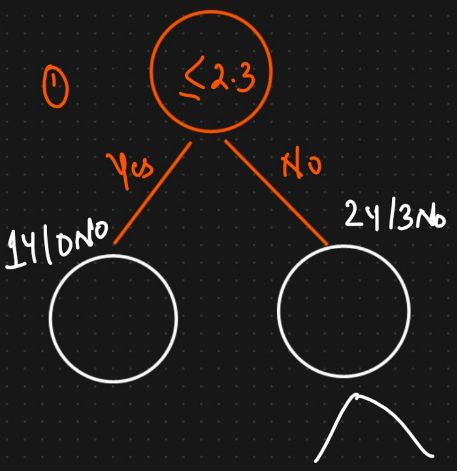
When dataset is huge → Gini Impurity [Simple Maths]

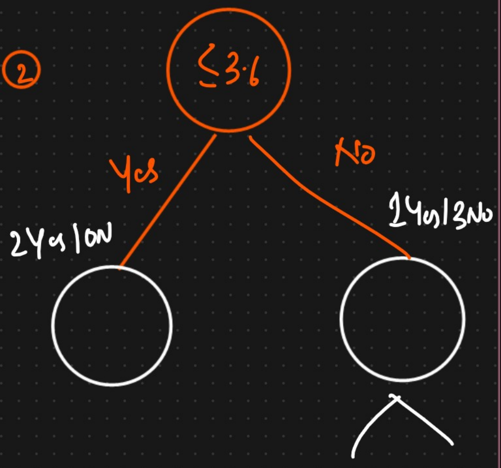
ⓓ What if my feature is continuous.

f1	O/p
→ 2.3	Yes
→ 3.6	Yes
4	No
5.2	No
6.7	Yes
7.8	No

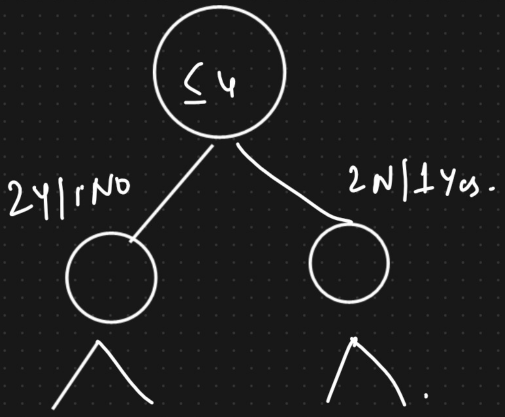
① Sort the feature f1

① Threshold = 2.3





③ Threshold = 4



Time Complexity ↑↑

DATASET ↑↑.