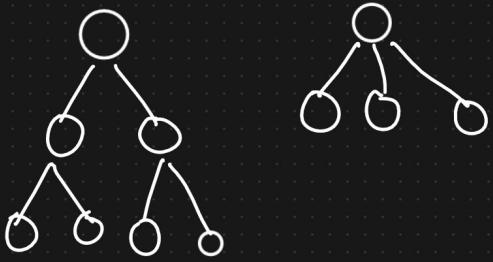


Decision Tree Classifier

Decision Tree Classifier

→ ID3
→ CART ✓



a) Entropy and Gini Index → Purity Split

b) Information Gain → features to select for

DT construction

$age = 14$

if ($age \leq 15$):

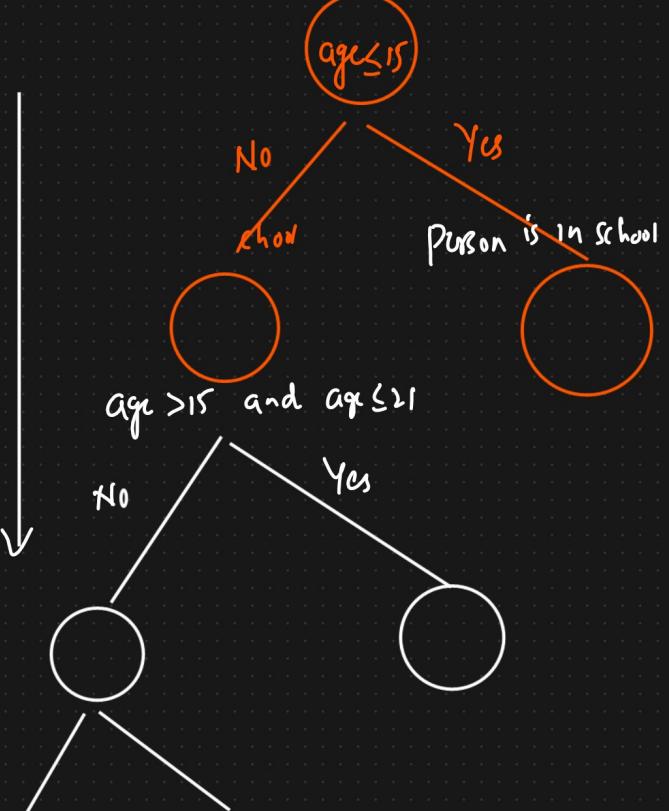
Print ("The person is in School")

elif ($age > 15$ and $age \leq 21$):

Print ("The person may be college")

else:

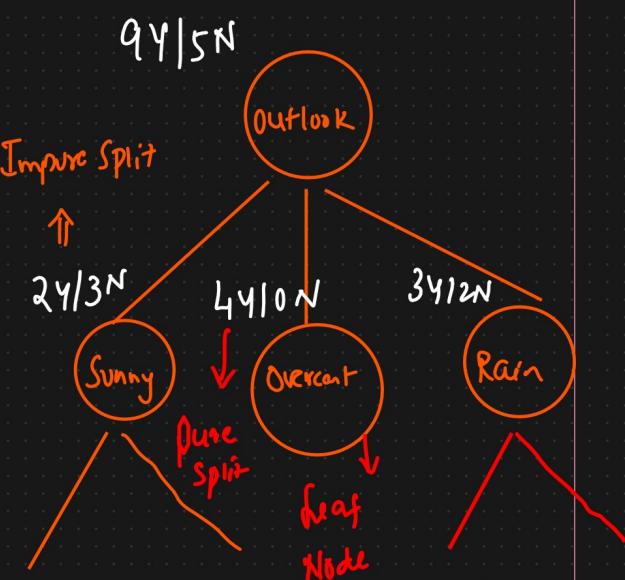
Print ("The person has passed")



Data set

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 Punty → Pure or Impure Split

↳ Entropy
↳ Gini Impurity }

② What feature you need Select for
Splitting → Information Gain }

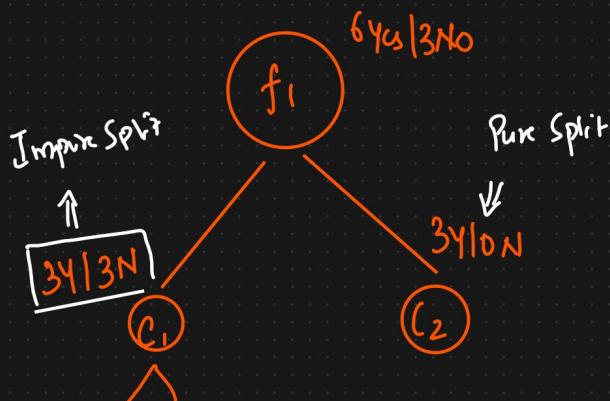
1

0

{Binary Classification}

1) Entropy

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

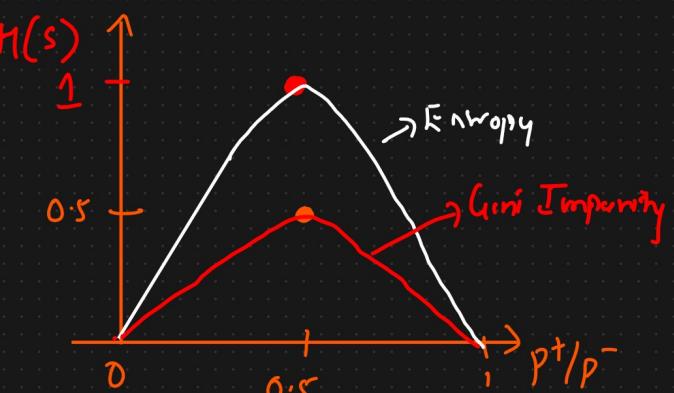


$$H(C_1) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$

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

2) Gini Impurity

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



\Rightarrow Impure Split

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

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

(2) Min Impurity

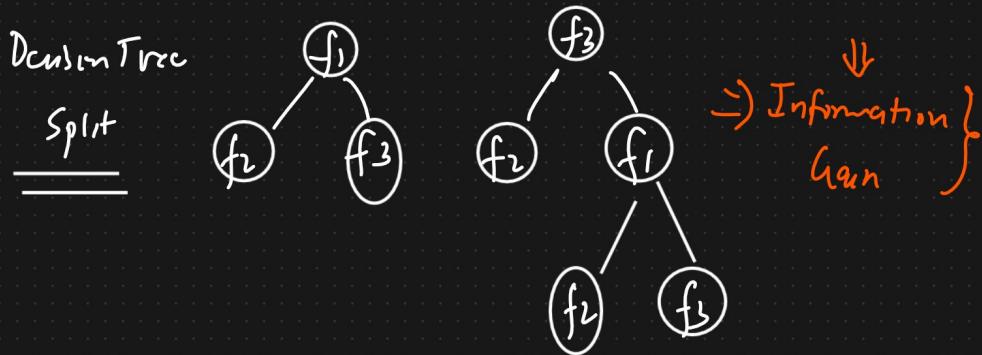
$$\begin{aligned} G \cdot I &= 1 - \sum_{i=1}^n (P_i)^2 \\ &= 1 - \left((P_1)^2 + (P_2)^2 \right) \\ &= 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right) \\ &= 0 \Rightarrow \text{Pure Split} \\ &\equiv 0.5 \Rightarrow \text{Impure Split} \end{aligned}$$

BY ION

$$\begin{aligned} &= 1 - \left(\left(\frac{3}{5}\right)^2 \right) \\ &= 1 - 1 \end{aligned}$$

\Rightarrow Pure Split

$f_1 \quad f_2 \quad f_3$



Information Gain

$\text{Gain}(S, f_1) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$

Root Node

Entropy of the root node

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

$H(S) = 1 - \left(\frac{3}{5} \log_2 \frac{3}{5} + \frac{2}{5} \log_2 \frac{2}{5} \right) = 0.971$

$H(S_1) = 1 - \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1 - 1 = 0$

$H(S_2) = 1 - \left(\frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3} \right) = 0.918$

$H(S_3) = 1 - \left(\frac{1}{2} \log_2 \frac{1}{2} + \frac{1}{2} \log_2 \frac{1}{2} \right) = 1 - 1 = 0$

$\text{Gain}(S, f_1) = 0.971 - \frac{3}{5} \cdot 0 - \frac{2}{5} \cdot 0.918 = 0.971 - 0.367 = 0.604$

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

$f_1 \rightarrow C_1$

$f_2 \rightarrow C_2$

$f_3 \rightarrow C_3$

$C_1 \rightarrow C_4$

$C_2 \rightarrow C_3$

$C_3 \rightarrow C_4$

$14 = 9y/5N$

$8 = 6y/2N$

$3y/3N = 6$

\leftrightarrow

\downarrow

Impure split

$$= -\frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} \quad H(C_1) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8}$$

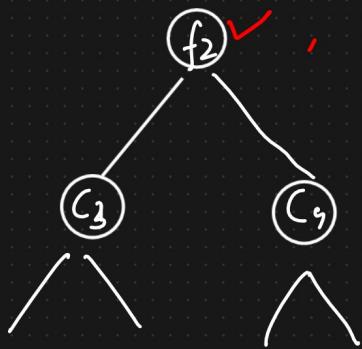
≈ 0.94

$$H(C_1) \approx 0.81$$

$$H(C_2) = 1$$

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

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



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

Information $\frac{\text{Gain}}{\text{Gain}}$ is Basically calculated.

Entropy \checkmark Vs Gini Impurity \checkmark Text Text

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_- \quad G.I. = 1 - \sum_{i=1}^n (P_i)^2 \Rightarrow$$

O/P = 3 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



Text