Decision tree

① ID3

Iterative Decotomiser.

(Entropy)

② CART

classification and Regression Tree.

(Gini impurity)

Day	outlook	Temp	Humid	wind	Decision.
1	S	H	H	W	N
2	S	H	H	S	N
3	O	H	H	W	Y
4	R	M	H	W	Y.
5	R	C	N	W	Y.
6	R	C	N	S	N
7	O	C	N	S	Y.
8	S	M	H	W	N
9	S	C	N	W	Y.
10	R	M	N	W	Y.
11	S	M	N	S	Y
12	O	M	H	S	Y
13	O	H	N	W	Y.
14	R	M	H	S	N.

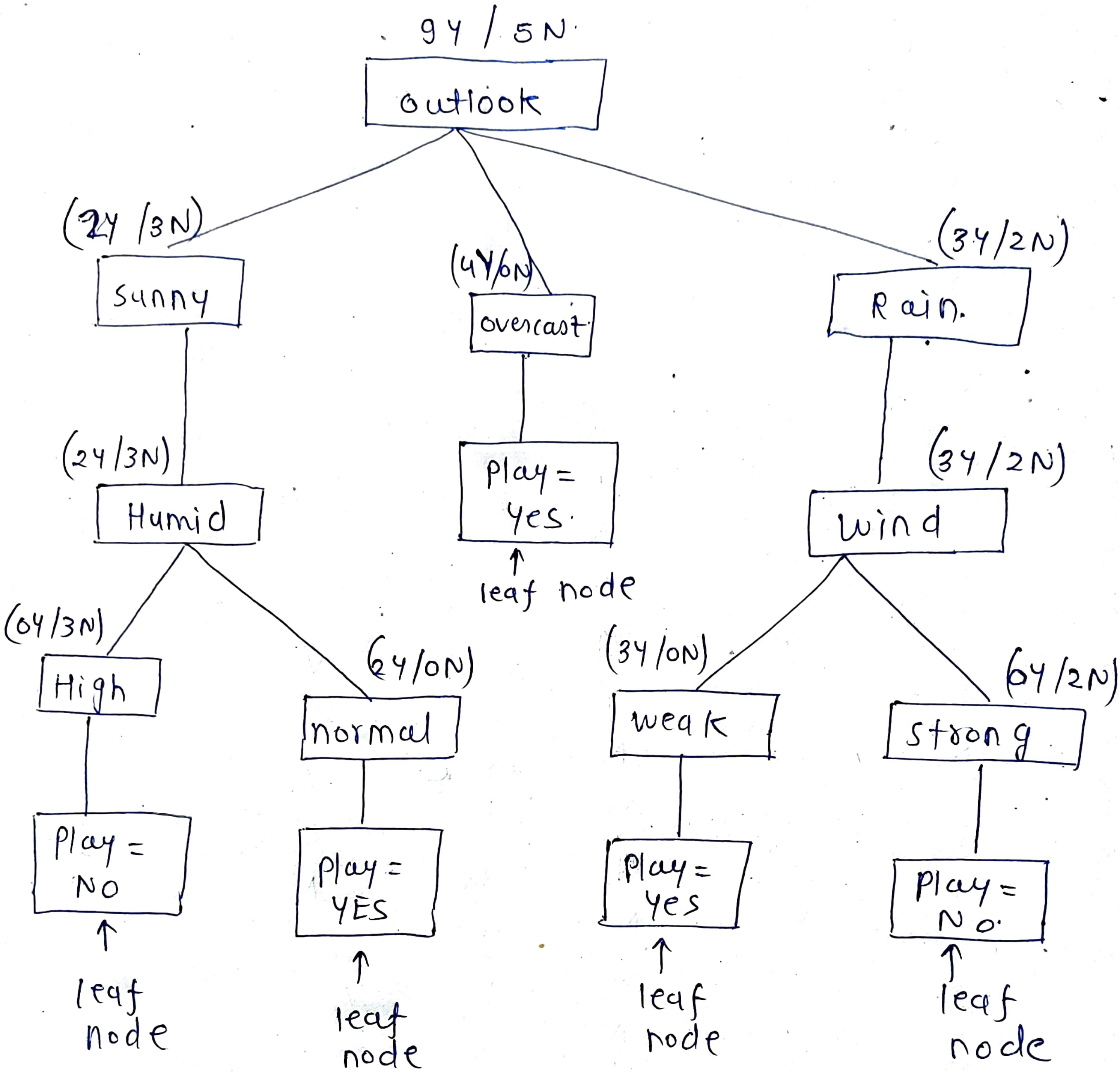
outlook { S → Sunny, O → overcast, R → Rain

Humidity { H → High, N → Normal

Play Y - Yes, N. - No

Temp { H → Hot, M → mild, C → cold

wind { W → weak, S → strong



* I have provided detailed mathematical solution of above problem using entropy at last of these notes.

* Entropy: measure of randomness (Disorder) or measure of purity or impurity.

$$\boxed{E = -\sum_{i=1}^{n} P_i \times \log_2(P_i)}$$

for 2 classes (Y/N)

$$\boxed{E_2 = -P_Y \log_2(P_Y) - P_N \log_2(P_N)}$$

for 3 classes (C_1, C_2, C_3)

$$\boxed{E_3 = -P_{C_1} \log_2(P_{C_1}) - P_{C_2} \log_2(P_{C_2}) - P_{C_3} \log_2(P_{C_3})}$$

Gini Impurity

$$\boxed{G = 1 - \sum_{i=1}^{n} p_i^2}$$

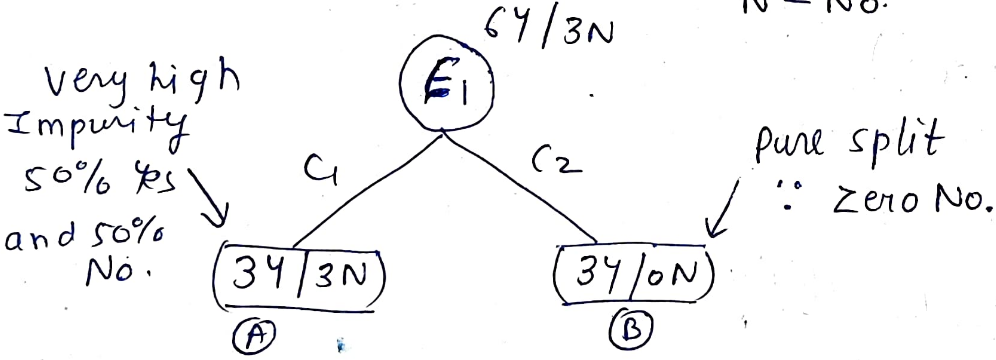
n ≡ no. of classes

for 2 classes (Y, N).

$$\boxed{G_2 = 1 - (P_Y^2 + P_N^2)}$$

Example

Y – Yes
N – No.



f1	Decision
C1	N
C2	Y
C1	Y
C1	N
C1	Y
C2	Y
C1	Y
C1	N
C2	Y

Checking Impurity of f1.

$$H(S) = E_A = -\sum_{i=1}^{n} p_i \times \log(p_i)$$

$$= -P_Y \log_2(P_Y) - P_N \log_2(P_N)$$

$$= -\frac{3}{6} \log_2\left(\frac{3}{6}\right) - \frac{3}{6} \log_2\left(\frac{3}{6}\right)$$

$$= -2 \times \frac{3}{6} \log_2(3/6) = -\log_2 \frac{1}{2} = 1$$

~~Gini impurity~~

$$E_B = -\cancel{\log}\; P_Y \log_2(P_Y) - P_N \log_2(P_N)$$

$$= -\frac{3}{3}\log_2(1) - 0 \times \log_2(0).$$

$$= -\log_2(1).$$

$$\boxed{E_B = 0}$$

Gini impurity

$$G_A = 1 - \sum_{i=1}^{n} P_i^2 = 1 - (P_Y^2 + P_N^2)$$

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

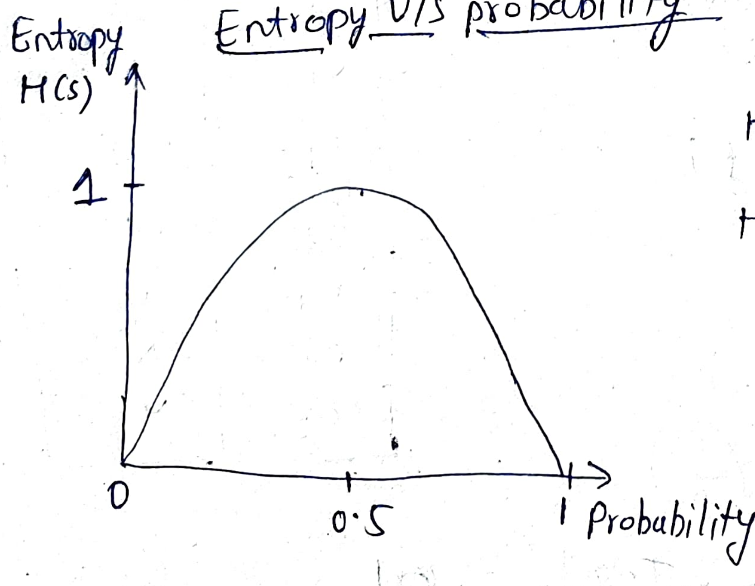
$$= 1 - \left(\frac{1}{4} + \frac{1}{4}\right) = 1 - \frac{1}{2} = \frac{1}{2}$$

$$G_B = 1 - \left[P_Y^2 + P_N^2\right]$$

$$= 1 - (1^2 + 0^2)$$

$$G_B = 0$$

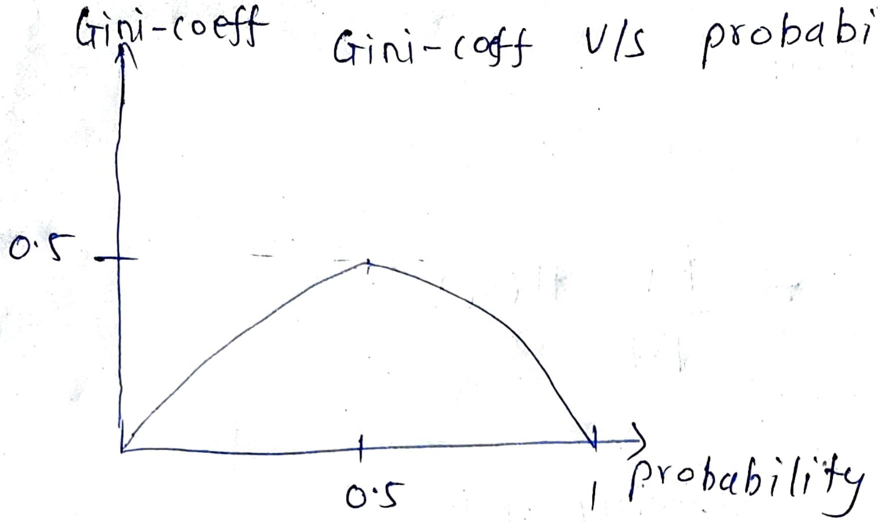
Entropy V/s probability



H(s) = 1 very impure split

H(s) = 0 pure split

Gini-coff v/s probability



Entropy $\Rightarrow$ 0 to 1

Gini-coeff $\Rightarrow$ 0 to 0.5

<u>4Y / 8N</u>

$$G = 1 - \left[\left(\frac{4}{12}\right)^2 + \left(\frac{8}{12}\right)^2\right] = 1 - \left(\frac{1}{9} + \frac{4}{9}\right)$$

$$= 1 - \frac{5}{9}$$

$$= 4/9$$

$$\boxed{G = 0.44}$$

(8Y, 2N)

$$G = 1 - \left[\left(\frac{8}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right] = 1 - \left(\frac{16}{25} + \frac{1}{25}\right)$$

$$= 1 - \frac{17}{25}$$

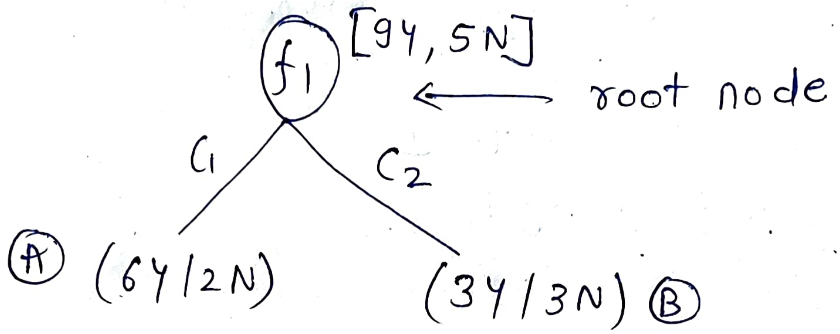
$$= 8/25$$

$$\boxed{G = 0.32}$$

Information Gain

1) Entropy

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



$H(S) \equiv$ root feature entropy.

$$H(S) = -P_Y \log_2(P_Y) - P_N(\log_2(P_N))$$

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

$$= 0.41 - (-0.53)$$

$$\boxed{H(S) = 0.94}$$

$$E_A = -6/8 \log(6/8) - \frac{2}{8}\log(2/8)$$

$$\boxed{E_A = 0.81}$$

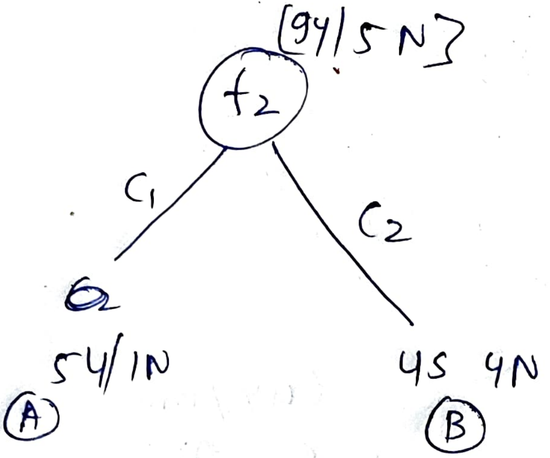
$$E_B = -3/6 \log(3/6) - 3/6 \log(3/6)$$

$$\boxed{E_B = 1}$$

Total Y/N in A Total Y/N in B.

$$Gain(S,f_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1.0\right]$$

$$= 0.048$$



$H(S) = 0.94$

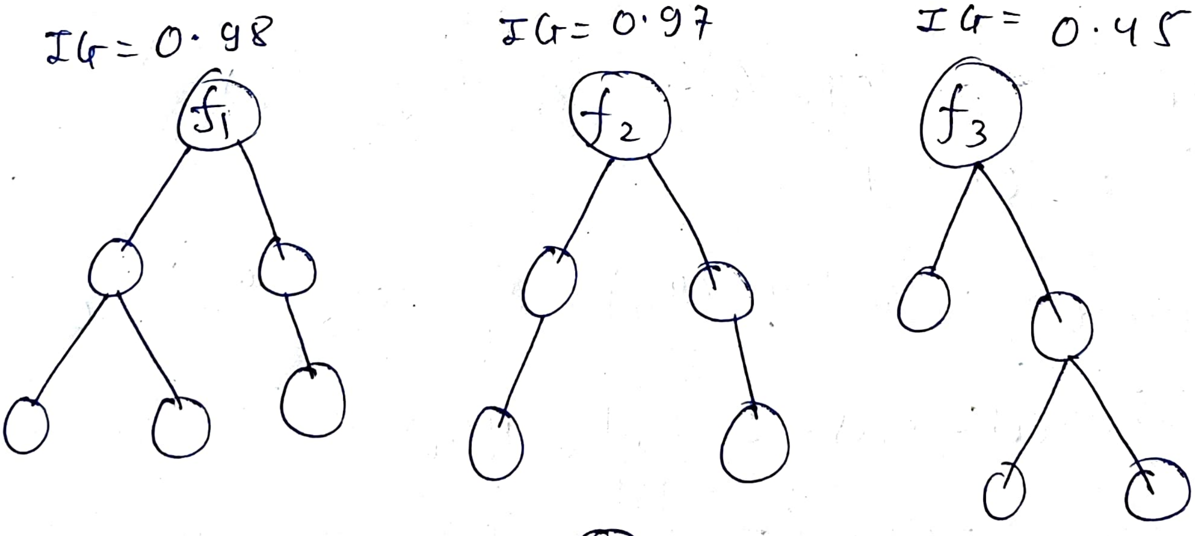
$$E_A = -\frac{5}{6}\log_2(5/6) - \frac{1}{6}\log_2(1/6)$$
$$= \cancel{0.82}\ 0.65$$

$$E_B = -4/8 \log_2(4/8) - 4/8 \log_2(4/8) = 1$$

$$\text{Gain}(S, f_2) = 0.94 - \left(\frac{6}{14} \times 0.65 + \frac{8}{14} \times 1\right)$$
$$= 0.09$$

$\because$ $\text{Gain}(S, f_2) > \text{Gain}(S, f_1)$

$\therefore$ f_2 will be better as root feature.

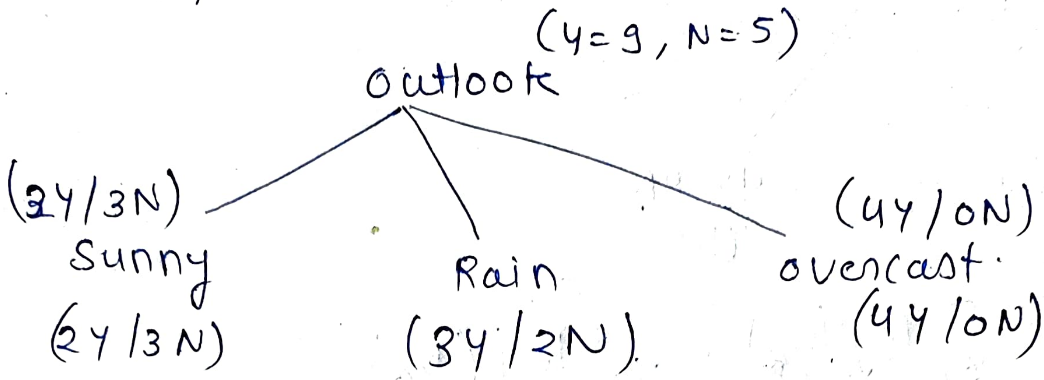


$\therefore$ Root node $\Rightarrow$ f_1

=> Solution of Problem on First page.

Total Y and N => Y = 9 N = 5

① outlook feature.



$$H(S) = -\Sigma P_i \log_2 (P_i)$$

$$= -P_Y \log_2 (P_Y) - P_N \log_2 (P_N)$$

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

$$H(S) = 0.94$$

$$E_{sunny} = -P_Y \log_2 (P_Y) - P_N \log_2 (P_N)$$

$$= -\frac{2}{5} \log_2 (2/5) - \frac{3}{5} \log_2 (3/5)$$

$$= 0.971$$

$$E_{Rain} = -3/5 \log_2 (3/5) - 2/5 \log_2 (2/5)$$

$$= 0.971$$

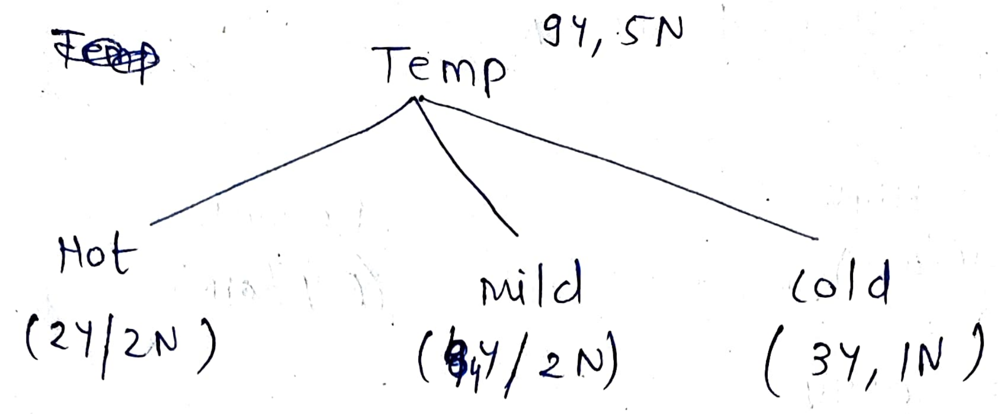
$$E_{overcast} = -4/4 \log_2 (4/4) - \frac{0}{4} \log_2 (0/4) = 0$$

$$Gain(s, outlook) = H(S) - \Sigma \frac{|S_v|}{|S|} \times H(S_v)$$

$$= 0.94 - \left(\frac{5}{14} \times 0.971 + \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0\right)$$

$$Gain(s, outlook) = 0.246$$

(2)



$H(S) = 0.94$. for 9Y and 5N.

$$E_{Hot} = -P_Y \log_2(P_Y) - P_N \log_2(P_N)$$
$$= -\frac{2}{4}\log_2(2/4) - \frac{2}{4}\log_2(2/4)$$
$$= 1$$

$$E_{mild} = -\frac{4}{6}\log(4/6) - \frac{2}{6}\log(2/6)$$
$$= 0.918$$

$$E_{cold} = -\frac{3}{4}\log(3/4) - \frac{1}{4}\log(1/4)$$
$$= 0.811$$

$$Gain(S, Temp) = H(S) - \sum \frac{|S_V|}{|S|} \times H(S_V)$$
$$= 0.94 - \left(\frac{4}{14}\times 1 + \frac{6}{14}\times 0.918 + \frac{4}{14}\times 0.811\right)$$
$$= 0.029$$

③ Humid 9Y/5N

High (3Y/4N)

Normal. (6Y/1N)

$H(S) = 0.94$

$$E_{High} = -P_Y \log_2(P_Y) - P_N \log_2(P_N)$$

$$= -\frac{3}{7}\log_2(3/7) - \frac{4}{7}\log_2(4/7)$$

$$= 0.985$$

$$E_{Normal} = -\frac{6}{7}\log_2(6/7) - \frac{1}{7}\log_2(1/7)$$

$$= 0.592$$

$$Gain(S, Humid) = 0.94 - \left(\frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.592\right)$$

$$= 0.151$$

④ wind (9Y/5N)

weak (6Y/2N)

Strong 3Y/3N

$H(S) = 0.94$

$$E_{weak} = -\frac{6}{8}\log_2(6/8) - \frac{2}{8}\log_2(2/8)$$

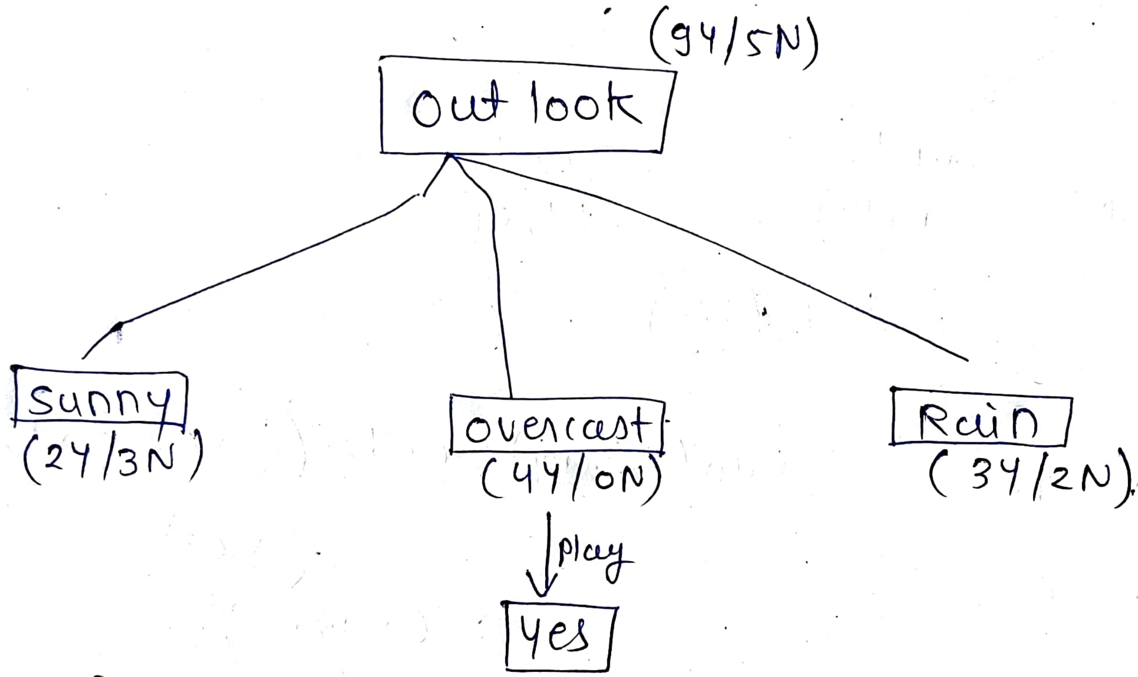
$$= 0.811$$

$$E_{strong} = -\frac{3}{6}\log_2(3/6) - 3/6 \log_2(3/6)$$
$$= 1$$

$$Gain(S, wind) = 0.94 - \left(\frac{8}{14} \times 0.811 + \frac{6}{14} \times 1\right)$$
$$= 0.048$$

$\because$ Gain (S, outlook) is Highest ie 0.246

$\therefore$ our root node is outlook



* Now Doing calculations for Decision Node (Sunny)
* for 2Y and 3N.

$$H(S) = -P_Y \log_2(P_Y) - P_N \log_2(P_N)$$
$$= -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5)$$
$$H(S) = 0.97$$

outlook	Temp	Humid	wind	Decision
S	H	H	W	N
S	H	H	S	N
S	M	H	W	N
S	C	N	W	Y
S	M	N	S	Y

① Temp (2Y, 3N)

Hot (0Y, 2N)

mild (1N, 1Y)

cool. (1Y, 0N)

$E_{Hot} = 0 - \frac{2}{2}\log_2(2/2) = 0$

$E_{mild} = 1$

$E_{cool} = 0$

$Gain(S, Temp) = 0.97 - (\frac{2}{5}\times 0 + \frac{2}{5}\times 1 + \frac{1}{5}\times 0)$

$= 0.97 - 2/5$

$= 0.57$

② Humid (2Y, 3N)

High (0Y, 3N)

Normal (2Y, 0N)

$E_{High} = 0$

$E_{Normal} = 0$

$Gain(S, Humid) = 0.97 - (0 + 0)$

$= 0.97$

(3) wind (2Y, 3N)

weak (1Y, 2N) — strong (1Y, 1N)

$$E_{weak} = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.918$$

$$E_{strong} = 1$$

$$Gain(S, wind) = 0.97 - \left(\frac{3}{5} \times 0.918 + \frac{2}{5} \times 1\right)$$

$$= 0.0192$$

$\because$ Gain(S, Humid) is Highest ie (0.97)

$\therefore$ ~~Gain(s~~ Humid is our Decision Node for sunny.

(2) for overcast there won't be any Decision node because we have already reached leaf node for overcast.

(3) Now Doing calculations for Decision Node (Rain)
(3Y and 2N)

Outlook	Temp	Humid	wind	Decision
R	M	H	W	Y
R	C	N	W	Y
R	C	N	S	N
R	M	N	W	Y
R	M	H	S	N

* for 3Y and 2N

$$H(s) = -\frac{3}{5}\log_2(3/5) - \frac{2}{5}\log_2(2/5) = 0.97.$$

1) Temp (3Y, 2N)

Hot (0Y, 0N) | mild (2Y, 1N) | cold (1Y, 1N)

$E_{Hot} = 0$

$$E_{mild} = -\frac{2}{3}\log_2(2/3) - \frac{1}{3}\log_2(1/3) = 0.918$$

$E_{cold} = 0$

$$Gain(s, Temp) = 0.97 - \left(0 + \frac{3}{5} \times 0.918 + 0\right)$$
$$= 0.4192$$

(2) Humid (3Y, 2N)

High (1Y, 1N) | Normal (2Y, 1N)

$E_{High} = 0$

$E_{Normal} = 0.918$

$$Gain(s, Humid) = 0.97 - \left(0 + \frac{3}{5} \times 0.918\right) = 0.4192$$

③ wind (3Y, 2N)

weak (3Y, 0N)   strong (0Y, 2N)

$E_{weak} = 0$

$E_{strong} = 0$

$$Gain(S, wind) = 0.97 - (0 + 0)$$

$$= \underline{0.97}$$

∵ Gain(S, wind) is Highest ∴ wind is our decision node for Rain.

Final Solution ⇒