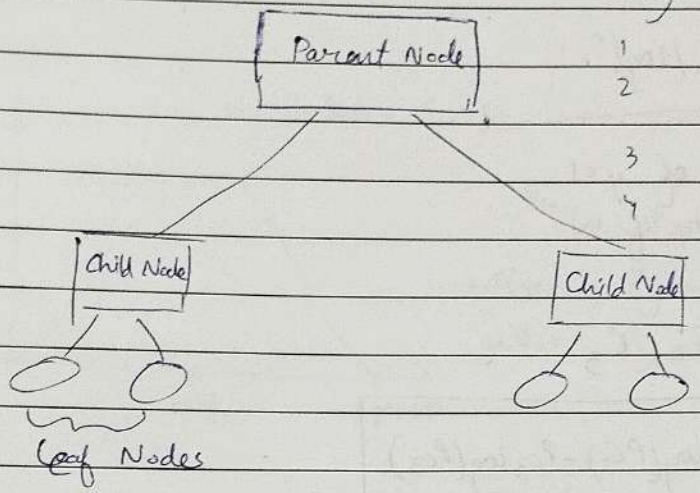
DECISION - TREE

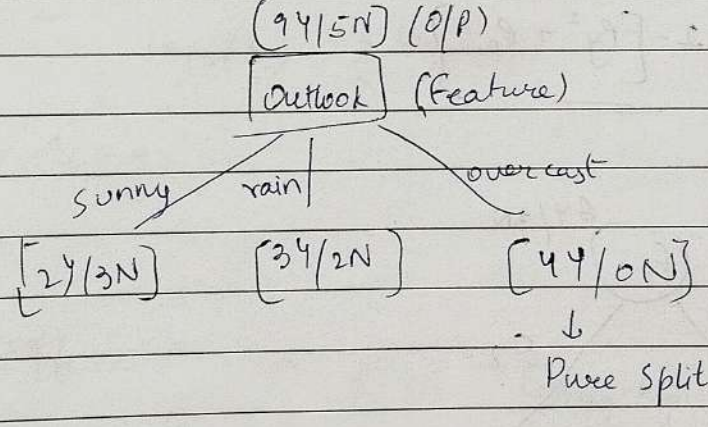
Dataset

Day	Outlook	Temp	Humidity	
1	Sunny	hot	High	Yes
2	Sunny	hot	low	No
3	Overcast	cold	High	No
4	rainfall	cold	low	Yes



① ID3
Iterative Decotomiser
(entropy)

② CART
Classification & Regression trees
Gini Impurity.



⇒ For checking the purity of the feature :-
① Entropy
② Gini coefficient / Gini impurity

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

Binary classification → $-P_y \log(P_y) - P_N(\log(P_N))$

P_y → Probability of yes
P_N → Probability of No.

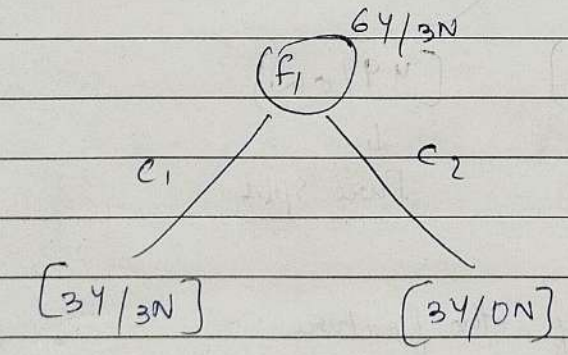
3 Class → C_1, C_2, C_3

Multi class Classification → $\boxed{-P_{C_1}\log(P_{C_1}) - P_{C_2}\log(P_{C_2}) - P_{C_3}\log(P_{C_3})}$

⇒ Gini coeff → $\boxed{1 - \sum_{i=1}^{n} P_i^2}$

→ $1 - [P_y^2 + P_N^2]$

Eg:



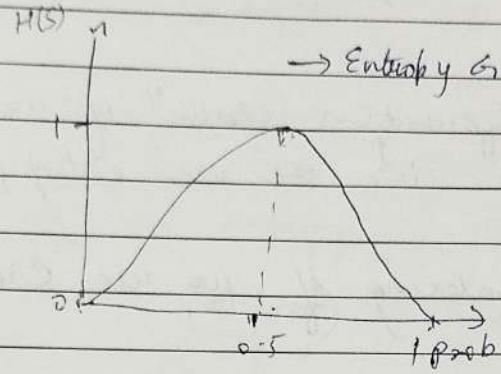
entropy = $H(S) = -\sum_{i=1}^{n} P_i \times \log_2(P_i)$

$= -0.5 \log_2(0.5) - 0.5 \log_2(0.5)$ (P_y, P_N)

$-1 \log_2(0.5) = -1 \times -1 = 1$

Entropy for other node = $-1\log_2(1) - 0\log_2(0)$

$= 0$



→ Entropy Graph wrt probability.

When; $P_Y = 0.5$, $P_N = 0.5$, $H(S) = 1$,
when, $P_Y = 1$, $P_N = 0$, $H(S) = 0$
when, $P_Y = 0$, $P_N = 1$, $H(S) = 0$

Note :-

$H(S) = 1$ → very impure split

$H(S) = 0$ → Pure split

2Y/3N

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

$H(S) = 0.97$

(2) Gini coeff.

$$1 - \sum_{i=1}^{2} P_i^2$$

(1) 3Y/3N → Entropy - high → $H(S) = 1$ ⇒ Very impure split

(2) 3Y/0N → Entropy - 0 → Gini index = 0

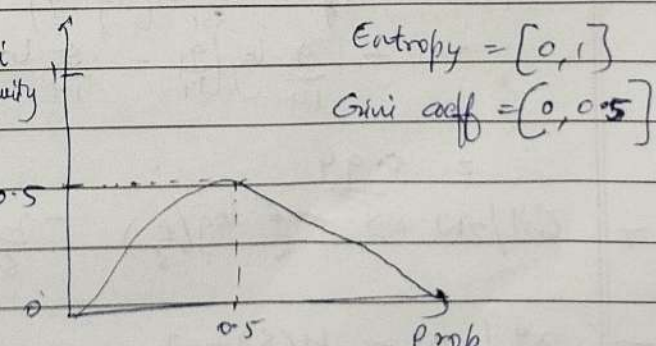
(3) 2Y/3N

3Y/3N $= 1 - \left[0.5^2 + 0.5^2\right]$

$= 0.5$

4Y/6N $= 1 - \left[\frac{1}{9} + \frac{4}{9}\right)$

$= 1 - \frac{5}{9} = \frac{4}{9} = 0.44$



Entropy = [0, 1]

Gini coeff = [0, 0.5]

→ $[8Y/2N] = 1 - \left[\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right]$

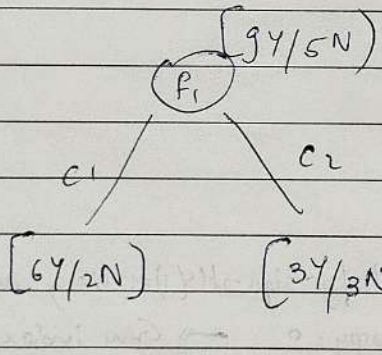
$= 1 - \left[\frac{16+1}{25}\right] = \frac{25-17}{25} = \frac{8}{25} = 0.32$

→ We have to calculate "Information Gain". If we consider ID3 algorithm to calculate IG (↑) we use entropy.

→ We have to use Gini indexing if we use CART algorithm to find "Information Gain".

ID3 { Entropy, I.G } Gini Impurity, I.G } CART → Nowadays this is used.

⇒ $\boxed{Gain(S, F_1) = H(S) - \sum \frac{|S_v|}{|S|} \left(H(S_v)\right)}$



H(S) ⇒ root feature entropy.

- $H(S) = -\sum_{i=1} P_i \log(P_i)$

$= -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right)$

$= 0.94$

- $6Y/2N \Rightarrow -\frac{6}{8}\log\left(\frac{6}{8}\right) - \frac{2}{8}\log\left(\frac{2}{8}\right) = 0.81$

- $3Y/3N = H(S) = 1$

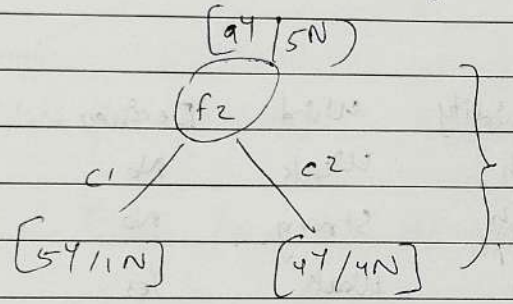
$Gain(S, F_1) = 0.94 - \left(\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right)$

$F_1 = 0.72$ — Information Gain wrt. to feature 1.

Σ_{Sv} represents total responses of child node [6Y/2N] → 8

Σ_S → Total responses

H_{Sv} → entropy of child Node.



$H(S) = -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right)$

$= 0.94$

$Gain = 0.94 - \left(\frac{8}{14}\right) \times 0.65 - \left(\frac{8}{14}\right) \times 1$

$Gain(F_2) = 0.8$

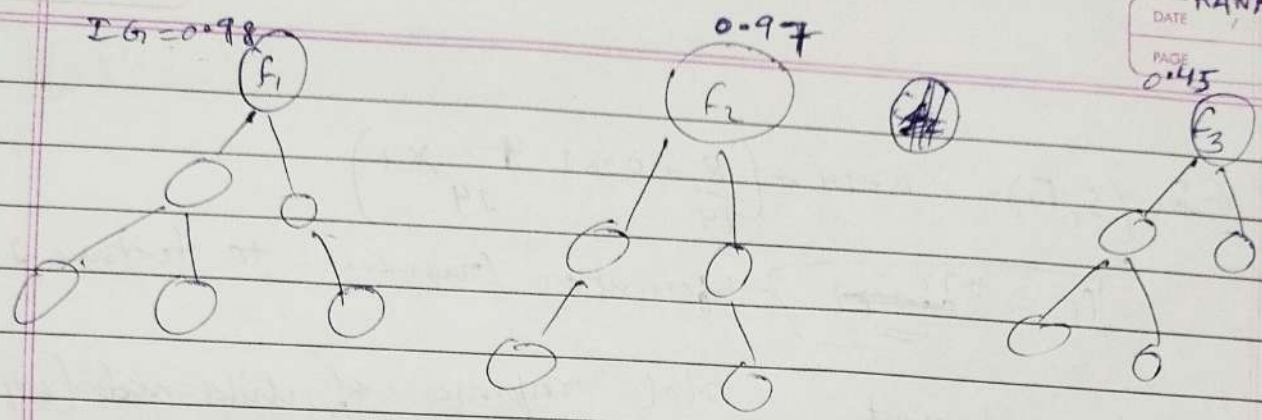
I.G of f_2 is more,

$IG(f_2) > IG(f_1)$

↓

$IG(f_2)$

This is greater & it is providing more information.



IG=0.98 is good

Day	Outlook	Temperature	Humidity	Wind	Decision
1	Sunny	hot	high	Weak	No
2	Sunny	hot	high	Strong	No
3	Overcast	hot	high	Weak	Yes
4	Rainfall	mild	high	Weak	Yes
5	Rainfall	cool	Normal	Weak	Yes
6	Rainfall	cool	Normal	Strong	No
7	Overcast	cool	Normal	Strong	Yes
8	Sunny	mild	High	weak	No
9	Sunny	cool	Normal	weak	Yes
10	Rainfall	mild	Normal	Weak	Yes
11	Sunny	mild	Normal	strong	Yes
12	Overcast	mild	High	strong	Yes
13	Overcast	hot	Normal	weak	Yes
14	Rainfall	mild	High	Strong	No

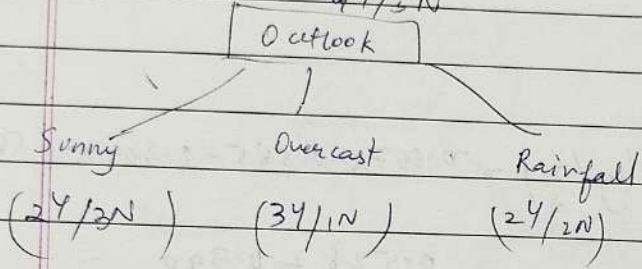
RANKA
DATE / /
PAGE

Assignment

1) Diff b/w Gini Impurity & Entropy.

2) Take the tennis play dataset & build Decision Tree from so

(2) Information Gain



$$Gain(S, F_i) = H_S - \sum_{|S|} \frac{|S_v|}{|S|} (H_{Sv})$$

$$H_S = -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = -0.64 \times -0.6374 - 0.35 \times -1.48$$
$$= 0.407 + 0.5194$$
$$= 0.9264$$

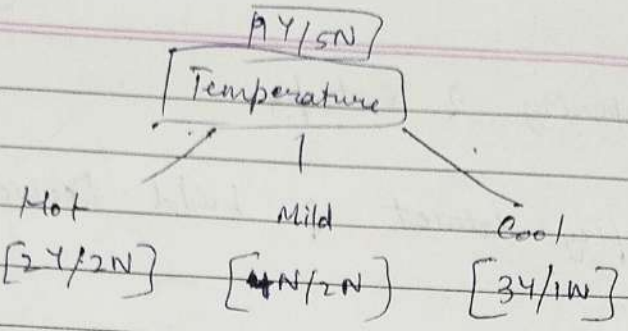
$$(H_S)_{Sunny} = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = -0.4 \times -1.322 - 0.6 \times -0.737$$
$$= 0.5288 + 0.4422$$
$$= 0.971$$

$$(H_{Sv})_{Overcast} = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) = -0.75 \times -0.415 - 0.25 \times -2$$
$$= 0.31125 + 0.5 = 0.81125$$

$$(H_{Sv})_{Rainfall} = -0.5\log_2(0.5) - 0.5\log_2(0.5)$$
$$= 2 \times -0.5 \times -1 = .2 \times +0.5 = 1$$

$$Gain = 0.9264 - \left[\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0.81125 + \frac{4}{14} \times 1\right]$$
$$= 0.9264 - [0.346 + 0.231 + 0.285] = 0.9264 - 0.8627$$
$$= 0.0637$$

⇒



$H_S = 0.9264$

⇒

$H_{(S,V)}_{(Hot)} = 1$

$H_{(S,V)_{Mid}} = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 0.667 \times 0.585 - 0.334 \times 1.585$

$= 0.528 + 0.390$

$= 0.918$

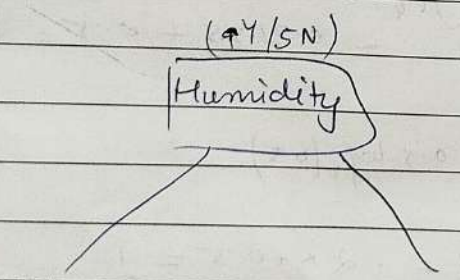
$H_{(S,V)\ Cool} = -\frac{3}{4}\log\left(\frac{3}{4}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) = 0.81125$

$Gain = 0.9264 - \left[\frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.81125\right]$

$= 0.9264 - \left[0.393 + 0.231 + 0.285\right]$

$= 0.9264 - 0.9097 = 0.0167$

⇒



$H_S = 0.9264$

$H_{SV(High)} = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right)$

$= 0.4285 \times 1.2224 - 0.5714$

$= 0.4613 + 0.5244 \times 0.8074$

$= 0.9857$

$H_{SV(Normal)} = -\frac{6}{7}\log\left(\frac{6}{7}\right) - \frac{1}{7}\log\left(\frac{1}{7}\right)$

$= 0.857 \times 0.2224 - 2.8074 \times 0.1428$

$= 0.4010 + 0.1905$

$= 0.59159$

⇒

RANKA
DATE / /
PAGE

Gain = $0.9264 - [0.5 \times 0.9857 + 0.5 \times 0.5917]$

$= 0.9264 - [0.49285 + 0.29585]$

$= 0.9264 - 0.78855 = 0.13785$

$\Rightarrow$

(9+/5-)

Wind

Weak [6+ / 2-]

Strong [3+ / 3-]

$H_S = 0.9264$

$H_{(S,U)}$ weak $= \frac{-6}{8}\log\left(\frac{6}{8}\right) - \frac{2}{8}\log\left(\frac{2}{8}\right)$

$= 0.81125$

H_{SU} strong $= \frac{1}{2}$

Gain $= 0.9264 - \left[\frac{8}{14} \times 0.81125 + \frac{6}{14} \times 1\right]$

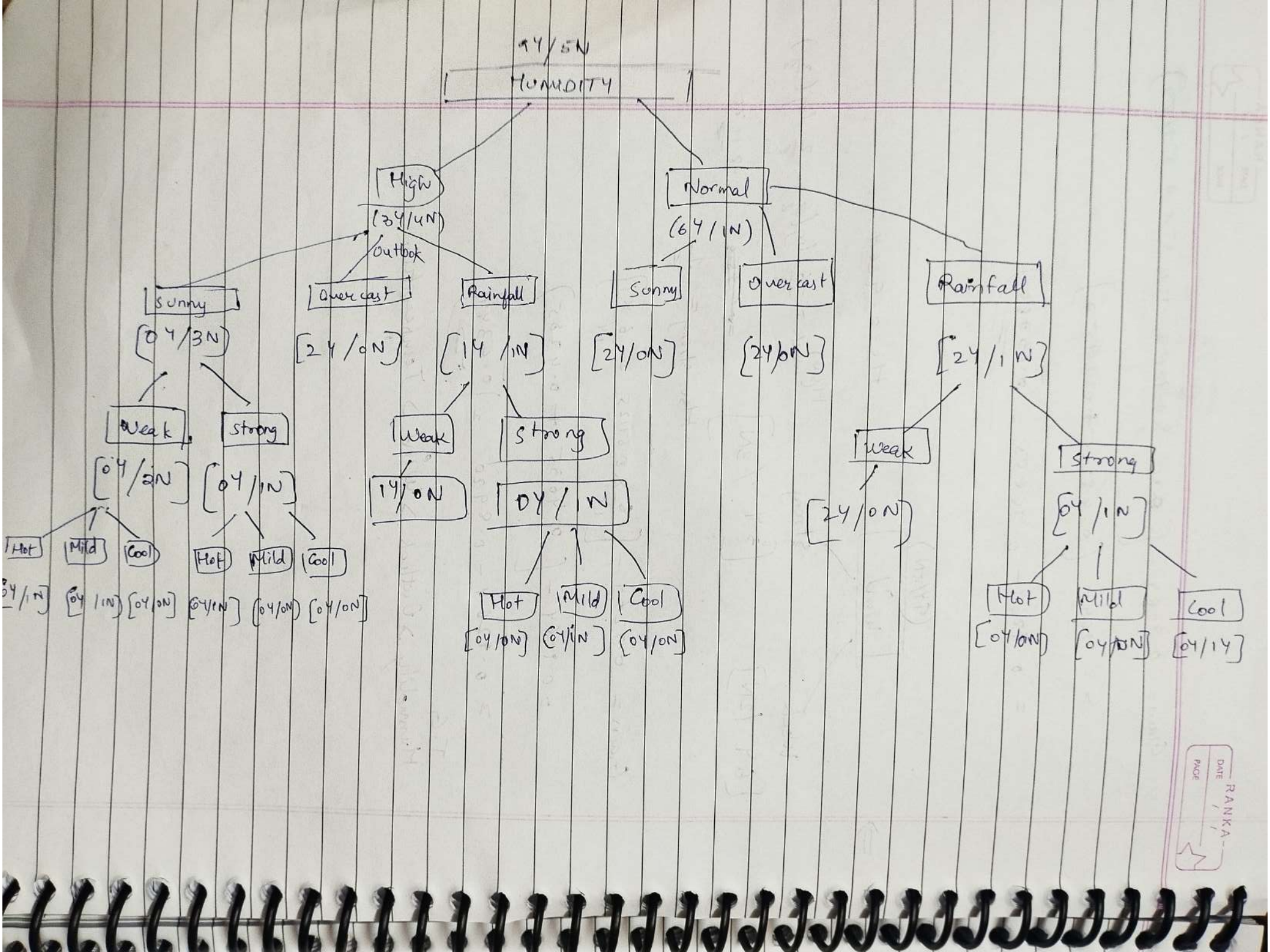
$= 0.9264 - [0.4635 + 0.4285]$

$= 0.9264 - 0.8920 = 0.0344$

$\Rightarrow$

I.G

Humidity > Outlook > Wind > Temperature


9+/5N
HUMIDITY
High
(3+/4N)
Normal
(6+/1N)
Outlook
Sunny
Overcast
Rainfall
Sunny
Overcast
Rainfall
[0+/3N]
[2+/0N]
[1+/1N]
[2+/0N]
[2+/0N]
[2+/1N]
Weak
Strong
Weak
Strong
Weak
Strong
[0+/2N]
[0+/1N]
1+/0N
0+/1N
2+/0N
[0+/1N]
Hot
Mild
Cool
Hot
Mild
Cool
Hot
Mild
Cool
Hot
Mild
Cool
[0+/1N]
[0+/1N]
[0+/0N]
[0+/1N]
[0+/0N]
[0+/0N]
[0+/0N]
[0+/1N]
[0+/0N]
[0+/0N]
[0+/0N]
[0+/1+]

	Entropy	Gini Index
1)	Range → (0-1) 0 → Perfectly pure, all data pts belong to one class 1 → Max. impurity → equal distribution of classes.	Range - 0 - 0.5 0 → Perfectly pure (Data pts belong to 1 class) 0.5 → Max Impurity (equal distribution)
2)	More Robust than Gini Index.	less Robust than Entropy.
3)	Entropy is logarithmic measure	Gini Index is linear measure.
4)	used in ID3 & C4.5 algorithms	Gini Index is used in CART
5)	Entropy measures the amount of uncertainty or randomness in a set.	Probability of misclassifying a randomly chosen element in a set