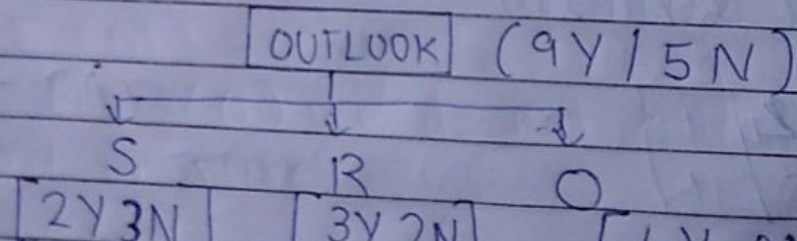
ROC → Receiver Operating Characteristics
AUC → Area Under Curve

Page No.
Date :

* Decision Tree

DAY	OUTLOOK	TEMPERATURE	HUMIDITY	WIND	DECISION
1	S	H	Hi	W	N
2	S	H	Hi	S	N
3	O	H	Hi	W	Y
4	R	M	Hi	W	Y
5	R	C	No	W	Y
6	R	C	No	S	N
7	O	C	No	S	Y
8	S	M	Hi	W	N
9	S	C	No	W	Y
10	R	M	No	W	Y
11	S	M	No	S	Y
12	O	M	Hi	S	Y
13	O	H	No	W	Y
14	R	M	Hi	S	N



↳ Pure Split
↳ Leaf Node

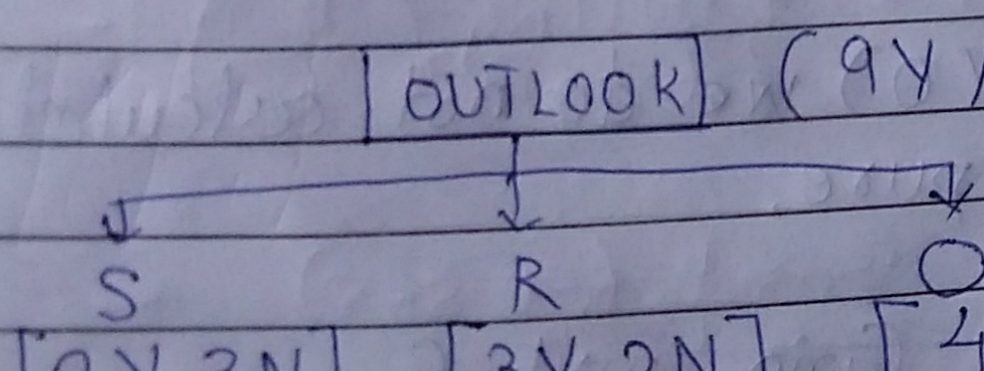
For checking purity we use (a) Entropy, (b) Gini Impurity

(a) Entropy → $\sum_{i=1}^{n} P_i \times \log(P_i)$

(b) Gini Impurity → $1 - \sum_{i=1}^{n} P_i^2$

Task (For Tennis Dataset)

(1) FOR OUTLOOK AS ROOT NODE



Entropy of Root Node

~~$H(y) = P_y \log(P_y)$~~ $H(S) = -P(Y) \log(P_Y) - P(N) \log(P_N$

$$= -\frac{9}{14} \times \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right)$$

$$= \frac{-9}{14}(-0.64) - \frac{5}{14}(-1.48)$$

Page No.
Date :

$= 0.411 + 0.52 \approx 0.94$

FOR (S)

Entropy $= -P_Y \log(P_Y) - P_N \log(P_N)$

$= \frac{-2}{5} \log\left(\frac{2}{5}\right) - \frac{3}{5} \log\left(\frac{3}{5}\right)$

$= \frac{-2}{5}(-1.32) - \frac{3}{5}(-0.737)$

$= 0.528 + 0.4422$

$= 0.9702$

~~Gain~~ → For (R)

Entropy $= -P_Y \log(P_Y) - P_N \log(P_N)$

$= -\frac{3}{5}\left(\log\left(\frac{3}{5}\right)\right) - \frac{2}{5} \log\left(\frac{2}{5}\right)$

$= 0.9702$

For (O) we do not calculate as the split is pure.

Page No.
Date :

IG wrt R = $0.94 - \left[\frac{5}{14} \times (0.9702)\right]$

≈ 0.5935

Entropy for all 3 branches

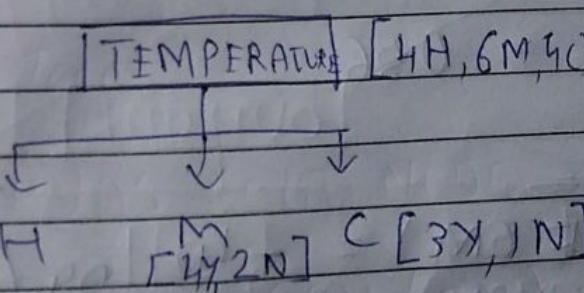
$= \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.97$

$= 0.69$

Information Gain = $0.94 - 0.69$

$= 0.25$

IF TEMPERATURE IS SELECTED AS ROOT NODE



	Y	N
H	2	2
M	4	2
C	3	1

Entropy of Root Node = 0.94

For H it is an impure split

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

$= -\frac{2}{4} \log_2\left(\frac{2}{4}\right) - \frac{2}{4} \log_2\left(\frac{2}{4}\right)$

$= \frac{1}{2} + \frac{1}{2} = 1$

~~E(A) =~~

$$E(M) = -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$$

$$= -\frac{4}{6}(-0.585) - \frac{2}{6}(-1.585)$$

$$= 0.39 + 0.52$$

$$= 0.918$$

$$E(I) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$$

$$= -\frac{3}{4}(-0.415) - \frac{1}{4}(-2)$$

$$= 0.811$$

Total Entropy For three branches =

$\frac{4}{14} \times 1 + \frac{6}{14} \times 0.91 + \frac{4}{14} \times 0.81$

Page No.
Date :

$\frac{4}{14} + 0.39 + 0.23 = 0.90$

$IG = 0.94 - 0.9057 \approx 0.03$

IF ~~TEMP~~ HUMIDITY is selected as root node

HUMIDITY (9Y, 5N)

Hi (3Y, 4N) No (6Y, 1N)

$E(Hi) = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right)$

$= 0.52 + 0.45$

$= 0.977$

$E(No) = -\frac{6}{7}\log\left(\frac{6}{7}\right) - \frac{1}{7}\log\left(\frac{1}{7}\right)$

$= 0.18 + 0.4 = 0.58$

Total $E = 0.97 \times \frac{7}{14} + 0.58 \times \frac{7}{14}$

$= 0.48 + 0.29$

$= 0.77$

$IG = 0.94 - 0.77 \approx 0.17$

IF WIND is selected as root node

WIND

S [3Y, 3N] W [6Y, 2N]

$E(S) = 1$

$$E(W) = -\frac{6}{8}\log\left(\frac{6}{8}\right) - \frac{2}{8}\log\left(\frac{2}{8}\right)$$

$= 0.3071 \quad 0.5$ ~~0.10~~

$= 0.807$

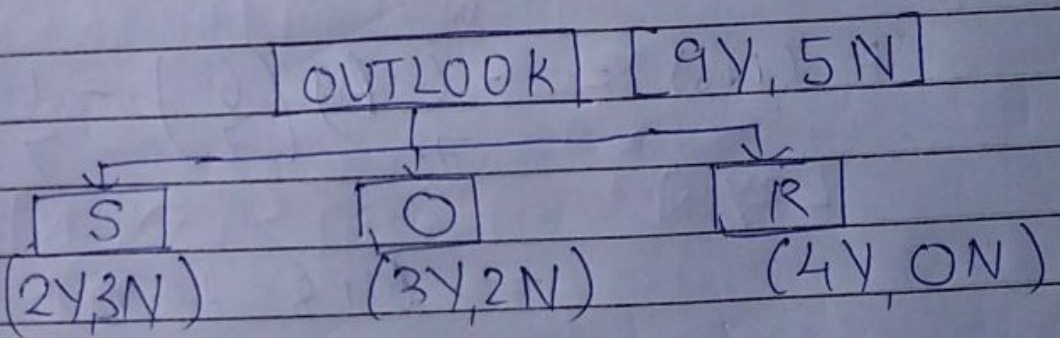
$$\text{Total entropy} = 0.807 \times \frac{8}{14} + \frac{6}{14} \times 1$$

$= 0.46 + 0.42$

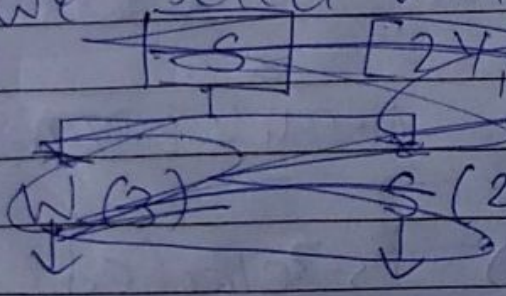
$= 0.88$

$IG = 0.94 - 0.88 = 0.051$

As IG is maximum for OUTLOOK Feat
we make it as root node.



Now we consider 'S' as root node
& decide the next feature on wh
the split is to be made
Suppose we select "wind"



Wind De
S W
S S
S W
S W
S S

S (2Y, 3N)

W(3) S(2)

(1Y, 2N) (1Y, 1N)

Page No.
Date :

Entropy of root node

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

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

$= 0.528 + 0.442$

$= 0.970$

Entropy (W) $= -\frac{1}{3} \log\left(\frac{1}{3}\right) - \frac{2}{3} \log\left(\frac{2}{3}\right)$

$= 0.528 + 0.39$

$= 0.918$

Entropy (S) $= 1$

Total entropy for 2 branches $= \frac{3}{5} \times 0.97 + \frac{2}{5} \times 1$

$= 0.982$

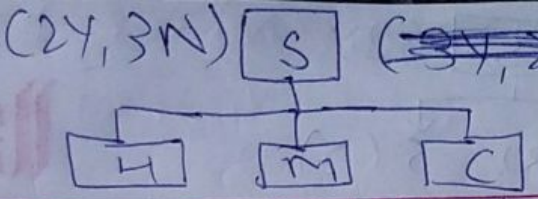
IG = Total Entropy of 2 branches − Entropy of root node

$= 0.982 - 0.918$

$= 0.064$

If we select Temperature as child Node

S	H	N
S	H	N
S	M	N
S	C	Y
S	M	Y



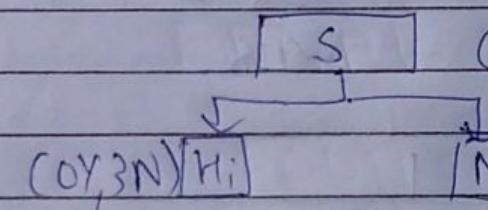
Page No.
Date :

Entropy H = 0 , Entropy C = 0
Entropy M = ~~$-\frac{1}{2}\log\frac{1}{2}$~~ 1

Total entropy of 3 branches = $0 + 0 + \frac{2}{5}$
$= \frac{2}{5}$

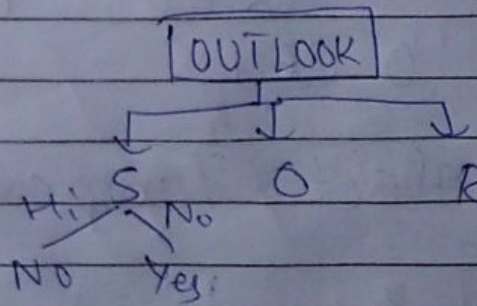
$IG = \frac{2}{5} - 0.970 = -0.4 + 0.970$
$= +0.57$

* If Humidity is selected for next split

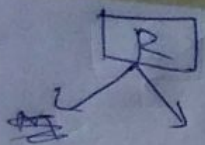


Entropy $H_i = 0$ Entropy $N_o = 0$
$\therefore IG = 0.97 - 0$
$= 0.97$

Thus For 'S' as root node its child node will be "Humidity" as it provides highest IG

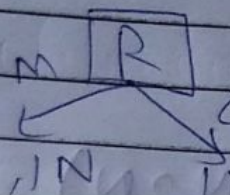


* Performing similar calculation for R branch



Page No.
Date :

If we select temperature for making further splits



R	M	Y
R	C	Y
R	C	N
R	M	Y
R	M	N

Entropy (R) = 0.970

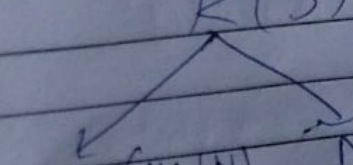
$$\text{Entropy}(M) = P_Y(\log P_Y) - P_N \log(P_N)$$
$$= \frac{2}{3}(\log_2(\frac{2}{3})) - \frac{1}{3}\log_2^2(\frac{1}{3})$$
$$= 0.39 + 0.528$$
$$= 0.918$$

Entropy (C) = 1

$$\text{Total entropy of 2 Branches} = \frac{1 \times 2}{5} + \frac{0.918 \times 3}{5}$$
$$= 0.950$$

$$\therefore IG = 0.97 - 0.95 = 0.019$$

If Humidity is selected as child node



R	Hi	Y
R	No	Y
R	No	N
R	No	Y
R	Hi	N

Entropy (Hi) = 1

$$\text{Entropy (No)} = -\frac{1}{3}\log(\frac{1}{3}) - \frac{2}{3}\log(\frac{2}{3})$$
$$= 0.528 + 0.39 = 0.918$$

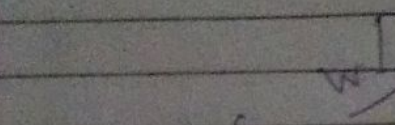
Total entropy = $1 \times \frac{2}{5} + 0.91 \times \frac{3}{5}$

$= 0.946$

$IG = 0.97 - 0.946 = 0.024$

Page No.
Date :

* If wind is selected as child node



Pure splits

R	W	Y
R	W	Y
R	S	N
R	W	Y
R	S	N

Entropy(W) = 0, Entropy(S) = 0

$IG = 0.97 - 0$

$= 0.97$

Thus we select ~~outlook~~ wind as final split

FINAL DT