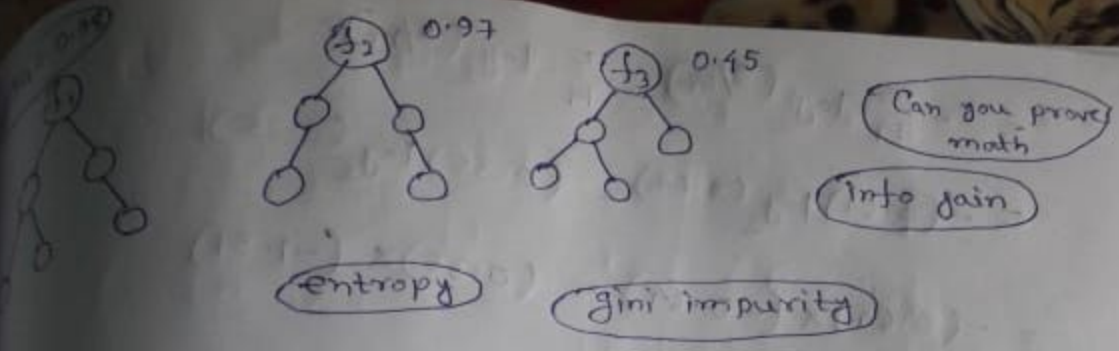
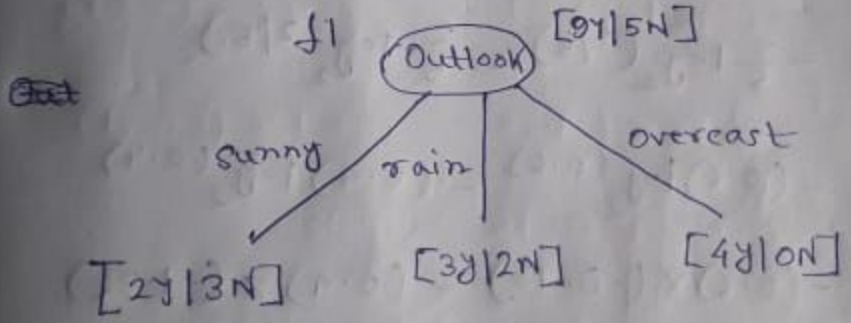


① entropy v/s gini_impurity [5 diff]

② take this data [tenis play] build DT from scratch (mathematical representation)

DTR | RFR RFC

Making of Decision Tree (Tennis Play Dataset) and Mathematical Representation



H(S) => root featuring entropy

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

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

$$= -(0.4)\log_2(0.4) - (0.6)\log_2(0.6)$$

$$= -(0.4)\times(-1.3) - (0.6)\times(-0.7)$$

$$= 0.52 + 0.42$$

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

$= -(0.64)\log_2(0.64) - (0.35)\log_2(0.35)$

$= -(0.64)\times(-0.64) - (0.35)\times(-1.51)$

$= 0.40 + 0.52$

≈ 0.93

2Y|3N $\Rightarrow -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5)$

$= -(0.4)\log_2(0.4) - (0.6)\log_2(0.6)$

$= -(0.4)\times(-1.32) - (0.6)\times(-0.73)$

$= 0.52 + 0.43$

$= 0.95$

3Y|2N $\Rightarrow -\frac{3}{5}\log_2(3/5) - \frac{2}{5}\log_2(2/5)$

$= -(0.6)\log_2(0.6) - (0.4)\log_2(0.4)$

$= -(0.6)\times(-0.73) - (0.4)\times(-1.32)$

$= 0.43 + 0.52$

$= 0.95$

4Y|0N $\Rightarrow -\frac{4}{4}\log_2(4/4) - \frac{0}{4}\log_2(0/4)$

$= -1\log_2 1 - 0$

$= 0$

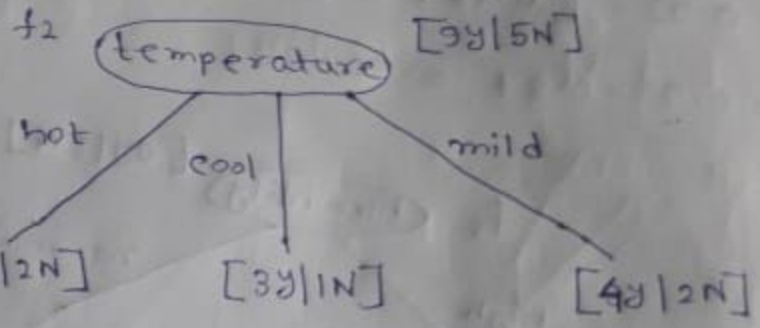
$\text{Gain}(S, f_1) = H_S - \sum_{\text{v\'e val}} \frac{|S_v|}{|S|} \times H(S_v)$

$= 0.93 - \left[\frac{5}{14}\times 0.95 + \frac{5}{14}\times 0.95 + \frac{4}{14}\times 0 \right.$

$= 0.93 - [0.33 + 0.33]$

$= 0.93 - 0.66$

$= 0.27$



9y|5N
= 0.93

$2y|2N \Rightarrow -\frac{2}{4}\log_2(2/4) - \frac{2}{4}\log_2(2/4)$

$= 1$

$3y|1N \Rightarrow -\frac{3}{4}\log_2(3/4) - \frac{1}{4}\log_2(1/4)$

$= -(0.75)\times\log_2(0.75) - (0.25)\log_2(0.25)$

$= -(0.75)\times(-0.41) - (0.25)\times(-2)$

$= 0.30 + 0.5 = 0.80$

$4y|2N \Rightarrow -\frac{4^2}{36}\log_2(4/6) - \frac{2^1}{3\,6}\log_2(2/6)$

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

$= -(0.66)\log_2(0.66) - (0.33)\log_2(0.33)$

$= -(0.66)\times(-0.59) - (0.33)\times(-1.59)$

$= 0.38 + 0.52$

$= 0.90$

$\text{Gain}(S, f_2) = H_S - \sum_{v \in val} \frac{|S_v|}{|S|} \times H(S_v)$

$= 0.93 - \left[\frac{4^2}{7\,14}\times 1 + \frac{4^2}{7\,14}\times 0.80 + \frac{6^3}{7\,14}\times 0.90\right]$

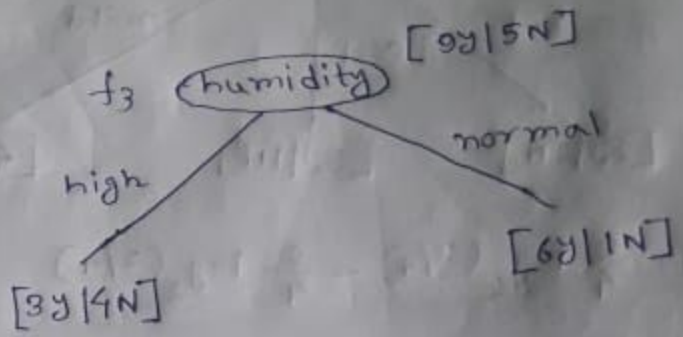
$\frac{4}{14}\times 0\Big]$

$= 0.93 - \left[\frac{2}{7} + \frac{2}{7} \times 0.80 + \frac{3}{7} \times 0.90\right]$

$= 0.93 - [0.28 + 0.28 \times 0.80 + (0.42) \times 0.90]$

$= 0.93 - [0.28 + 0.22 + 0.37]$

$= 0.93 - 0.87 = 0.06$



$9y|5N = 0.93$

$3y|4N \Rightarrow -\frac{3}{7}\log_2(3/7) - \frac{4}{7}\log_2(4/7)$

$= -0.42\log_2(0.42) - 0.57\log_2(0.57)$

$= -0.42 \times (-1.25) - 0.57 \times (-0.81)$

$= 0.52 + 0.46 = 0.98$

$6y|1N \Rightarrow -\frac{6}{7}\log_2(6/7) - \frac{1}{7}\log_2(1/7)$

$= -(0.85)\log_2(0.85) - (0.14)\log_2(0.14)$

$= -(0.85) \times (-0.23) - (0.14) \times (-2.83)$

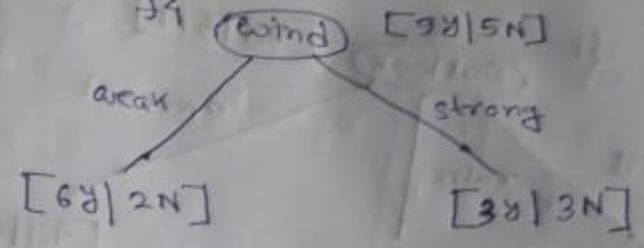
$= 0.19 + 0.39$

$= 0.58$

$Gain(S, f_3) = H_S - \sum_{v \in val} \frac{|S_v|}{|S|} \times H(S_v)$

$= 0.93 - \left[\frac{7}{14} \times 0.98 + \frac{7}{14} \times 0.58\right]$ (0.49, 0.29)

$= 0.93 - 0.78 = 0.15$



$9Y|5N = 0.93$

$6Y|2N \Rightarrow -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right)$

$= -\frac{3}{4}\log_2(3/4) - \frac{1}{4}\log_2(1/4)$

$= -0.75 \times \log_2(0.75) - 0.25 \times \log_2(0.25)$

$= -0.75 \times (-0.41) - 0.25 \times (-2)$

$= 0.30 + 0.50$

$= 0.80$

$3Y|3N \Rightarrow 1$

$$Gain(S, f_4) = H_S - \sum_{v \in val} \frac{|S_v|}{|S|} \times H(S_v)$$

$$= 0.93 - \left[\frac{8}{14} \times 0.80 + \frac{6}{14} \times 1\right]$$

$$= 0.93 - [0.45 + 0.42]$$

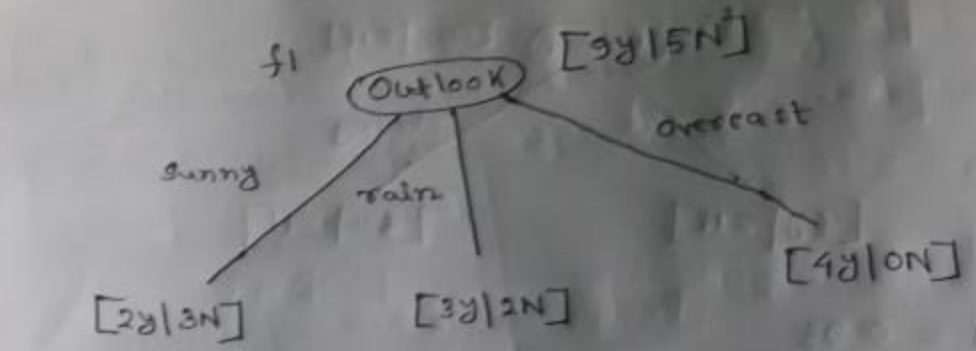
$$= 0.93 - 0.87$$

$$= 0.06$$

$Gain(S, f_1) = 0.27$, $Gain(S, f_2) = 0.06$

$Gain(S, f_3) = 0.15$, $Gain(S, f_4) = 0.06$

$\therefore$ Root Node = $Gain(S, f_1)$ = Outlook



Outlook	temperature	humidity	wind	Decision
Sunny	hot	high	weak	N
sunny	hot	high	strong	N
sunny	mild	high	weak	N
sunny	cool	normal	weak	y
sunny	mild	normal	strong	y

2y|3N ⇒ 0.95

0y|2N ⇒ 0

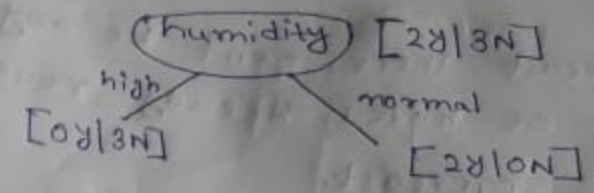
1y|1N ⇒ 1

1y|0N ⇒ 0

Gain (S temp) = $0.95 - \left(\frac{2}{5} \times 0 + \frac{2}{5} \times 1 + \frac{1}{5} \times 0\right)$

$= 0.95 - 0.40$

$= 0.55$

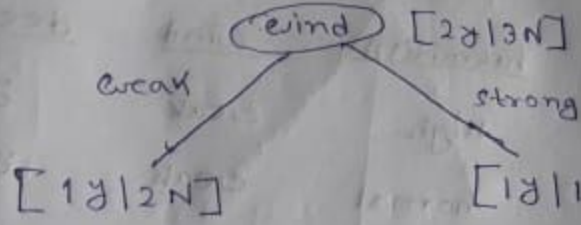


0Y|3N ⇒ 0

2Y|0N ⇒ 0

Gain(S, humidity) = 0.95 − (0+0)

= 0.95



1Y|2N ⇒ $-\frac{1}{3}\log_2(1/3) - \frac{2}{3}\log_2(2/3)$

$= -(0.33)\log_2(0.33) - (0.66)\log_2(0.66)$

$= -(0.33) \times (-1.59) - (0.66) \times (-0.59)$

$= 0.52 + 0.38$

≈ 0.91

1Y|1N ⇒ 1

Gain(S, wind) = $0.95 - \left(\frac{3}{5} \times 0.91 + \frac{1}{5} \times 1\right)$

$= 0.95 - (0.54 + 0.4)$

$= 0.95 - 0.94$

$= 0.01$

Gain(S, humidity) > Gain(S, temperature) > Gain(S, wind)

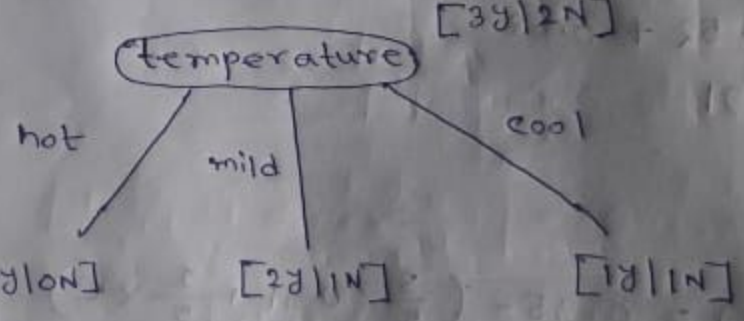
∴ Gain(S, humidity) is highest i.e (0.95)

∴ Humidity is decision node for sunny.

* for overcast there can't be any decision node because we have already reached the leaf node for overcast.

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

Outlook	temp	humidity	wind	decision
rainfall	mild	high	weak	yes
rainfall	cool	normal	weak	yes
rainfall	cool	normal	strong	no
rainfall	mild	normal	weak	yes
rainfall	mild	high	strong	no



$0y|0N \Rightarrow 0$

$2y|1N \Rightarrow -\frac{2}{3}\log_2(2/3) - \frac{1}{3}\log_2(1/3)$

≈ 0.91

$1y|1N \Rightarrow 1$

$Gain(S, temp) = 0.95 - \left[0 + \frac{3}{5} \times 0.91 + \frac{2}{5} \times 1\right]$

$= 0.95 - [0 + 0.54 + 0.40]$

$= 0.95 - 0.94 = 0.01$

1y|1N ⇒ 1

2y|1N ⇒ 0.91

Gain (S, humidity) = 0.01

3y|0N = 0

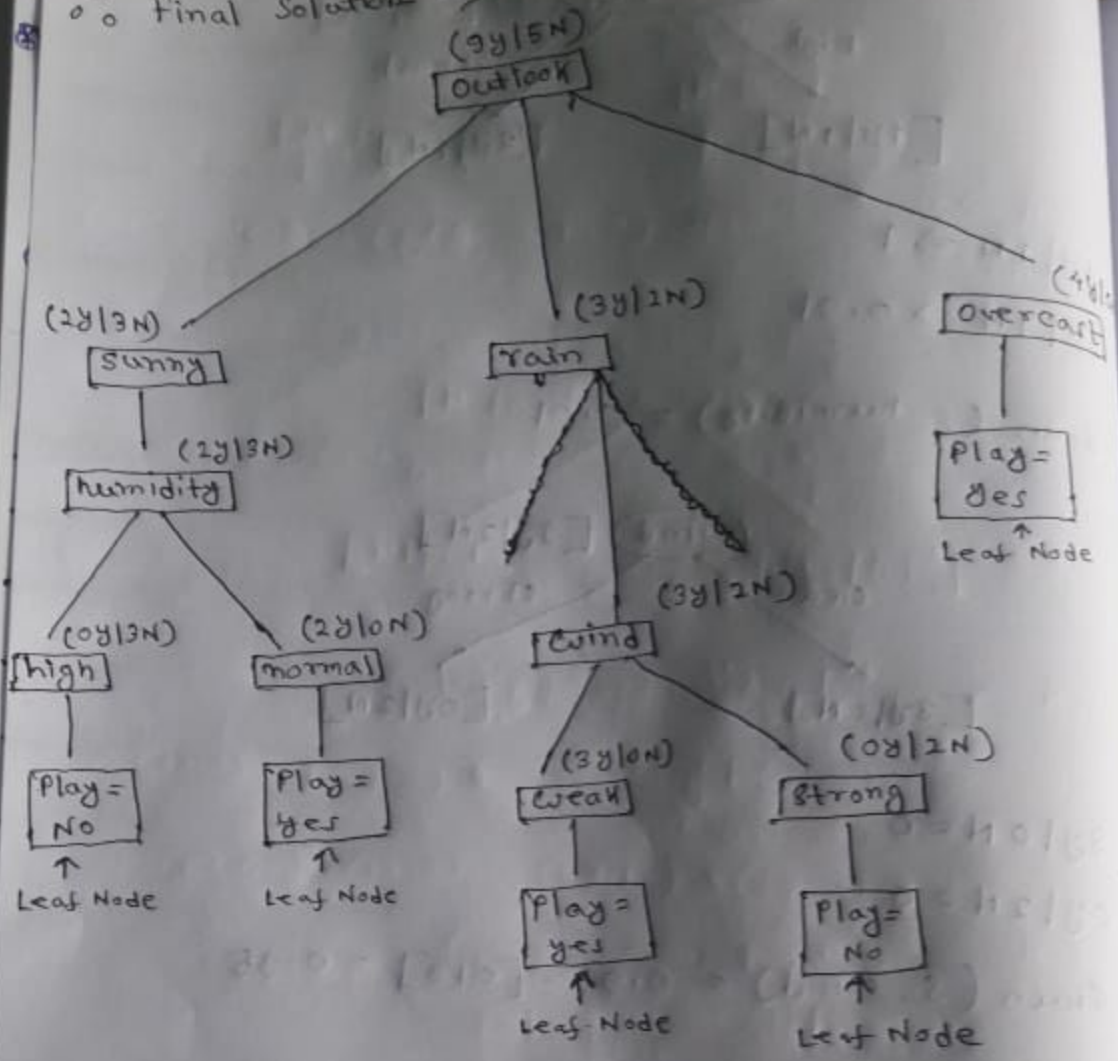
0y|2N = 0

Gain (S, wind) = 0.95 - [0+0] = 0.95

∵ Gain (S, wind) is highest

∴ Wind is our decision node for Rain.

∴ Final Solution ⇒



✱ Entropy Vs Gini Impurity

① If we use ID3 approach for calculating information gain we will ~~use~~ calculate entropy.

If we use CART approach for calculating information gain then we will calculate gini impurity

② Gini impurity is faster than Entropy becau in Entropy we use log function to calculate

entropy and in Gini impurity we don't use log function to calculate Gini impurity

② Entropy should be used when less number of features are present in the data set and Gini Impurity should be used when more number of ~~tree~~ features are present in the data set.

③ The range of Entropy lies in between 0 to 1 and the range of Gini impurity lies in between 0 to 0.5.

⑤ formula of ~~Gini~~ Entropy is

$$\sum_{i=1}^{n} p_i \times \log(p_i)$$

formula of Gini Impurity is $1 - \sum_{i=1}^{n} p_i^2$