# Machine Learning (Day - 8) :-

## Decision Tree (Classifier) :-

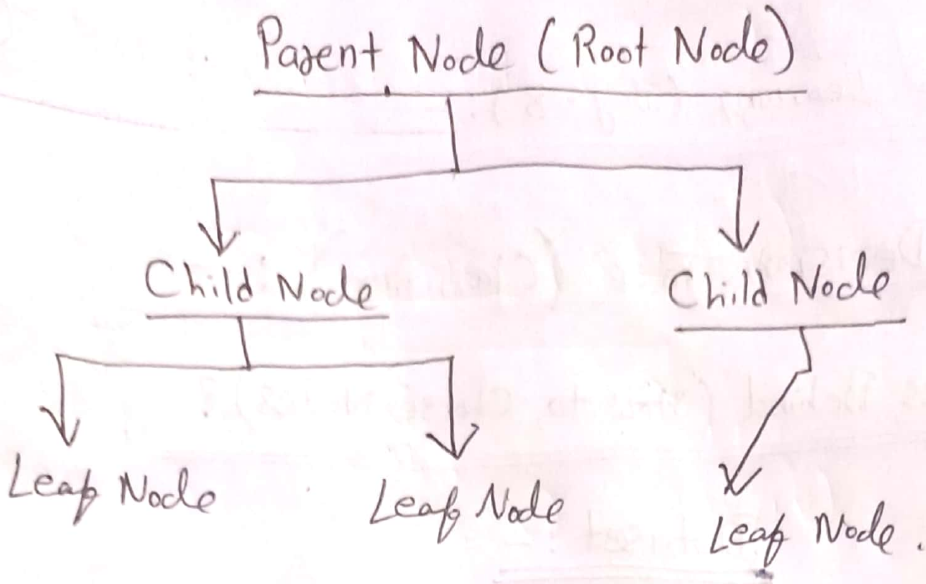
### *) Mathematics Behind (How to Choose Nodes)?

#### Dataset :-

| Day | Outlook | Temperature | Humidity | Wind | Decision |
|---|---|---|---|---|---|
| 1 | Sunny | Hot | High | Weak | No |
| 2 | S | H | H | Strong | N |
| 3 | Overcast | H | H | W | Yes |
| 4 | Rainfall | Mild | H | W | Y |
| 5 | R | Cool | Normal | 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 |
| | ($F_1$) | ($F_2$) | ($F_3$) | ($F_4$) | (Target) |

- ($F_1$) → Sunny, → Rainfall, → Overcast
- ($F_2$) → Hot, → Mild, → Cold
- ($F_3$) → High, → Normal
- ($F_4$) → Weak, → Strong
- (Target) → No, → Yes

P.T.O




Q) Question arises is How to Decide the Nodes?

⇒ Use the concept of Information Gain.

*) **Note :-**

Feature having highest Information Gain (IG), will be choosed as Node.

ID3 → Entropy is used to find IG

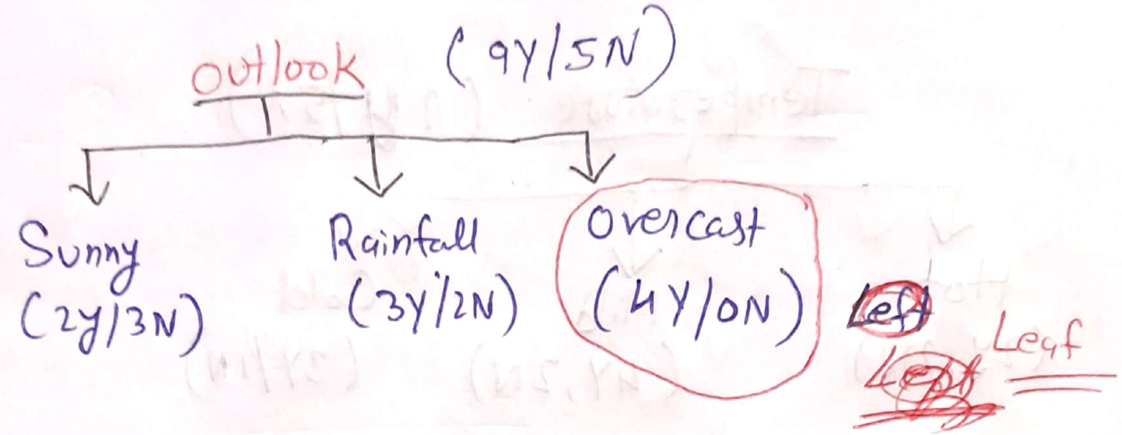
CART → Gini-Impurity is used to find IG.

*) Entropy ⇒ $- \sum_{i=1}^{n} P_i \times \log_2 (P_i) \Big\} H(S)$

*) Here, we will use ID3 (Entropy) to find Nodes.

Info. Gain ⇒ $G(S,F) = H(S) - \sum \frac{|S_v|}{|S|} * H(S_v)$

Outlook : { Sunny (S) , Rainfall (R) , Overcast (O) } ;




## Entropy :-

$$H(\text{outlook} = \text{Sunny}) = -\sum P_i \times \log_2(P_i)$$

$$= -P_y \cdot \log_2(P_y) - P_N \cdot \log_2(P_N)$$

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

$$= 0.97$$

$$H(\text{outlook} = \text{Rainfall}) = -\frac{3}{5} \log_2(3/5) - \frac{2}{5} \log_2\left(\frac{2}{5}\right)$$

$$= 0.97$$

$$H(\text{outlook} = \text{Overcast}) = -\frac{4}{4} \log\left(\frac{4}{4}\right) - \frac{0}{4} \log\left(\frac{0}{4}\right)$$

$$= 0$$

## B) Information Gain :- $G(S, \text{outlook}) = H(S) - \sum \frac{|S_v|}{|S|} \times H(S_v)$

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

$$= 0.94$$

&

$|S| = 14$, $|S_{sunny}| = 5$, $|S_{Rainfall}| = 5$, $|S_{overcast}| = 4$

$\therefore$ So, $G(S, \text{outlook}) = 0.94 - \left[\frac{5}{14} \times 0.97 + \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0\right]$

$G(\text{outlook}) = 0.247$

Scanned with CamScanner

Temperature : { Hot , Mild , Cold }

Temperature (9Y/5N)

- Hot (2Y, 2N)
- Mild (4Y, 2N)
- Cold (3Y/1N)

## Entropy :-

$$H(\text{Temp}=\text{Hot}) \Rightarrow -\frac{2}{4}\log\left(\frac{2}{4}\right)-\frac{2}{4}\log\left(\frac{2}{4}\right) = 1$$

$$H(\text{Temp}=\text{Mild}) \Rightarrow \frac{4}{6}\log\left(\frac{4}{6}\right)-\frac{2}{6}\log\left(\frac{2}{6}\right) = 0.918$$

$$H(\text{Temp}=\text{Cold}) \Rightarrow \frac{3}{4}\log\left(\frac{3}{4}\right)-\frac{1}{4}\log\left(\frac{1}{4}\right) = 0.811$$

## Information Gain :-

$$G(S=\text{Temp}) = H(S) - \Sigma \frac{|S_v|}{|S|} \times H(S_v)$$

$$\therefore H(S) \Rightarrow -\frac{9}{14}\log\left(\frac{9}{14}\right)-\frac{5}{14}\log\left(\frac{5}{14}\right) = 0.94.$$

$|S| = 14$, $|S_{Hot}| = 4$, $|S_{Mild}| = 6$, $|S_{Cold}| = 4$

$$G(S=\text{Temp}) = 0.94 - \left[\frac{4}{14}\times 1 + \frac{6}{14}\times 0.918 + \frac{4}{14}\times 0.811\right]$$

$$= 0.94 - (0.28571 + 0.3934 + 0.2317)$$

$$= \boxed{0.029}$$ Temperature

## Humidity : { High, Normal }

Humidity (9y/5N)

High (3Y/4N) | Normal (6Y/1N)

## Entropy :—

$$H(\text{Hum} = \text{High}) = -\frac{3}{7}\log\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right) \Rightarrow 0.985$$

$$H(\text{Hum} = \text{Normal}) = -\frac{6}{7}\log\left(\frac{6}{7}\right) - \frac{1}{7}\log\left(\frac{1}{7}\right) \Rightarrow 0.59$$

## Information Gain ⇒

$$G(S = \text{Humidity}) = H(S) - \left[\sum \frac{|S_v|}{|S|} \times H(S_v)\right]$$

$$\therefore H(S) = -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right) = 0.94$$

$$|S| = 14,\ |S_{High}| = 7,\ |S_{Normal}| = 7$$

$$\therefore G(S = \text{Humidity}) = 0.94 - \left[\frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.59\right]$$

$$= \boxed{0.153} \leftarrow \text{Humidity}$$

## ✱) Wind : { Weak, Strong }

Wind (9Y / 5N)

↓ Weak (6Y / 2N)

↓ Strong (3Y / 3N)

## ✱) Entropy :—

$$H(\text{Wind} = \text{Weak}) = -\frac{6}{8}\log\left(\frac{6}{8}\right) - \frac{2}{8}\log\left(\frac{2}{8}\right) \Rightarrow 0.811$$

$$H(\text{wind} = \text{Strong}) = \frac{-3}{6}\log\left(\frac{3}{6}\right) - \frac{3}{6}\log\left(\frac{3}{6}\right) \Rightarrow 1 \;//$$

## ✱) Information Gain :—

$$G(S = \text{Wind}) = 0.94 - \left[\frac{8}{14} \times 0.811 + \frac{6}{14} \times 1\right]$$

$$= \boxed{0.048} \leftarrow \text{Wind}$$

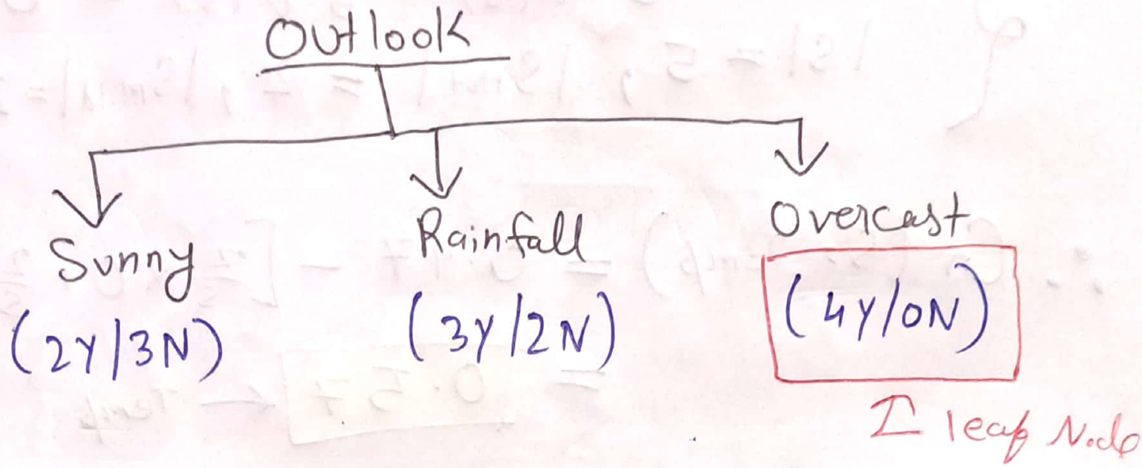
∴ Hence, out of 4 Features Outlook has highest Info. Gain (0.247).

So,

(Level-1 Node ⇒ Outlook)

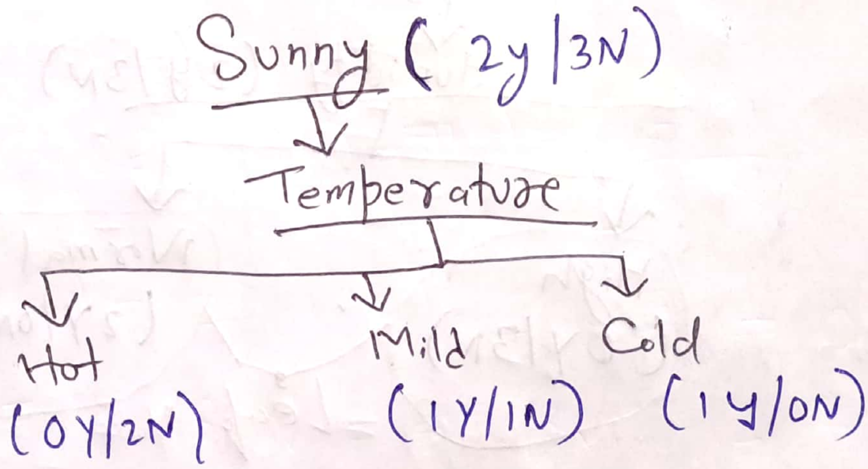
Now,

## Level-2 (Part-1)




Now,

**Sunny → Temperature :-**




**Entropy :-**

$$H(\text{Temp} = \text{Hot}) = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

$$H(\text{Temp} = \text{Mild}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$H(\text{Temp} = \text{Cold}) = -\frac{1}{1}\log\left(\frac{1}{1}\right) - \frac{0}{1}\log\left(\frac{0}{1}\right) = 0$$

P.T.O

## Info. Gain ⇒

$$G(S = Temp) = H(S) - \sum \frac{|S_v|}{|S|} \times H(S_v)$$

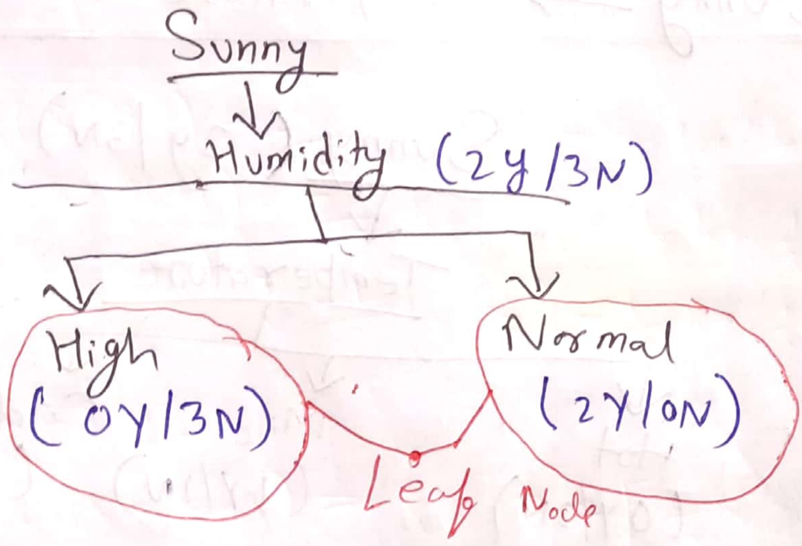
$$H(S) = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.97$$

$$|S| = 5,\ |S_{Hot}| = 2,\ |S_{Mild}| = 2,\ |S_{Cold}| = 1$$

$$\therefore G(S = Temp) = 0.97 - \left[\frac{2}{5}\times 0 + \frac{2}{5}\times 1 + \frac{1}{5}\times 0\right]$$

$$= \boxed{0.57} \leftarrow \text{Temp}$$

## Sunny → Humidity :




## Entropy :-

$$H(Humi = High) = -\frac{0}{3}\log\left(\frac{0}{3}\right) - \frac{3}{3}\log\left(\frac{3}{3}\right) = 0$$

$$H(Humi = Normal) = -\frac{2}{2}\log\left(\frac{2}{2}\right) - \frac{0}{2}\log\left(\frac{0}{2}\right) =$$

## Info. Gain ⇒

$$G(S = Humidity) = H(S) - \sum \frac{|S_v|}{|S|} \times H(S_v)$$

$$= 0.97 - \left[\frac{3}{5}\times 0 + \frac{2}{5}\times 0\right]$$

$$= 0.97 \leftarrow \text{Humidity}$$

## Sunny → Wind :—

Sunny
↓
Wind (2Y/3N)

Weak (1Y/2N) | Strong (1Y/1N)

### ∴ Entropy :—

$$H(\text{Wind} = \text{weak}) = -\frac{1}{3}\log\left(\frac{1}{3}\right) - \frac{2}{3}\log\left(\frac{2}{3}\right) = 0.918$$

$$H(\text{wind} = \text{Strong}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

### Info. Gain :—

$$G(S = \text{wind}) = 0.97 - \left[\frac{3}{5} \times 0.918 + \frac{2}{5} \times 1\right]$$

$$= \boxed{0.0192} \leftarrow \text{Wind}$$

Hence,

Out of 3 Feature, Humidity has Highest Info. Gain (0.97).

So, we have (Sunny > Humidity)

# Level 2 → [Part - 2]

Rainfall

↓

Temperature (3Y/2N)

Hot (0y/0N) — Mild (2Y/1N) — Cold (1Y/1N)

Here,

$$H(S) = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.97$$

## ✶) Entropy :-

$$H(S=\text{Hot}) = -\frac{0}{0}\log\left(\frac{0}{0}\right) - \frac{0}{0}\log\left(\frac{0}{0}\right) = 0$$

$$H(S=\text{Mild}) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.918$$

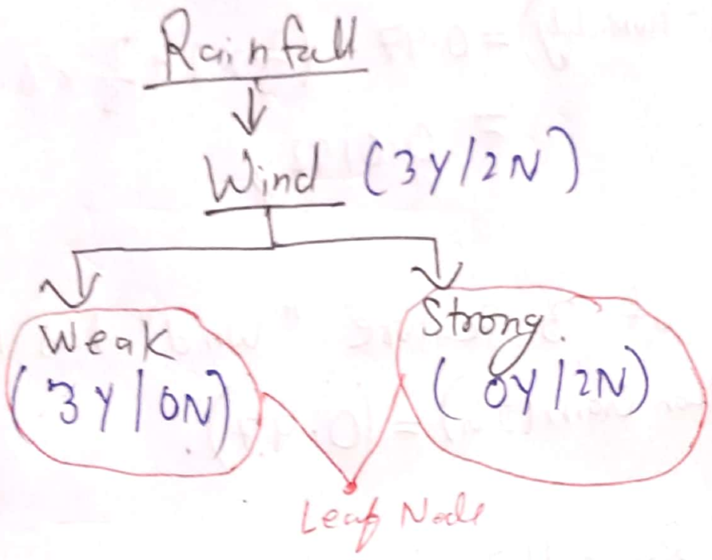
$$H(S=\text{Cold}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

## ✶) Info. Gain :-

$$G(S=\text{Temp}) = 0.97 - \left[\frac{0}{5}\times 0 + \frac{3}{5}\times 0.918 + \frac{2}{5}\times 1\right]$$

$$= \boxed{0.0192}$$

## Rainfall → Wind :–




**Entropy :–**

$$H(\text{Wind} = \text{Weak}) = -\frac{3}{3}\log\left(\frac{3}{3}\right) - \frac{0}{3}\log\left(\frac{0}{3}\right) = 0$$

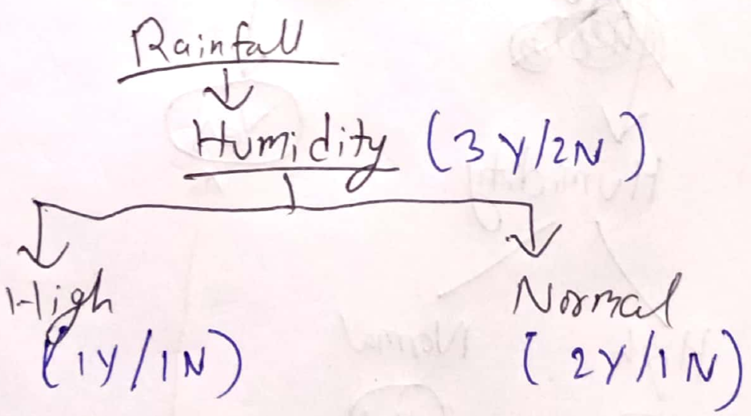
$$H(\text{Wind} = \text{Strong}) = -\frac{0}{2}\log\left(\frac{0}{2}\right) - \frac{2}{2}\log\left(\frac{2}{2}\right) = 0$$

∴ **Info. Gain :–**

$$G(S = \text{Wind}) = 0.97 - \left[\frac{3}{5}\times 0 + \frac{2}{5}\times 0\right]$$

$$= \boxed{0.97}$$

## Rainfall → Humidity




**Entropy :–**

$$H(\text{Hum} = \text{High}) = -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1$$

$$H(\text{Hum} = \text{Normal}) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = \boxed{0.918}$$

d) Info. Gain ⇒

$$G(S = \text{Humidity}) = 0.97 - \left(\frac{2}{5} \times 1 + \frac{3}{5} \times 0.918\right)$$
$$= 0.0192$$

∴ Hence, out of 3 Feature "Wind" has Highest Information Gain (IG) = (0.97).

So, Rainfall > Wind

Hence

Final Decision Tree