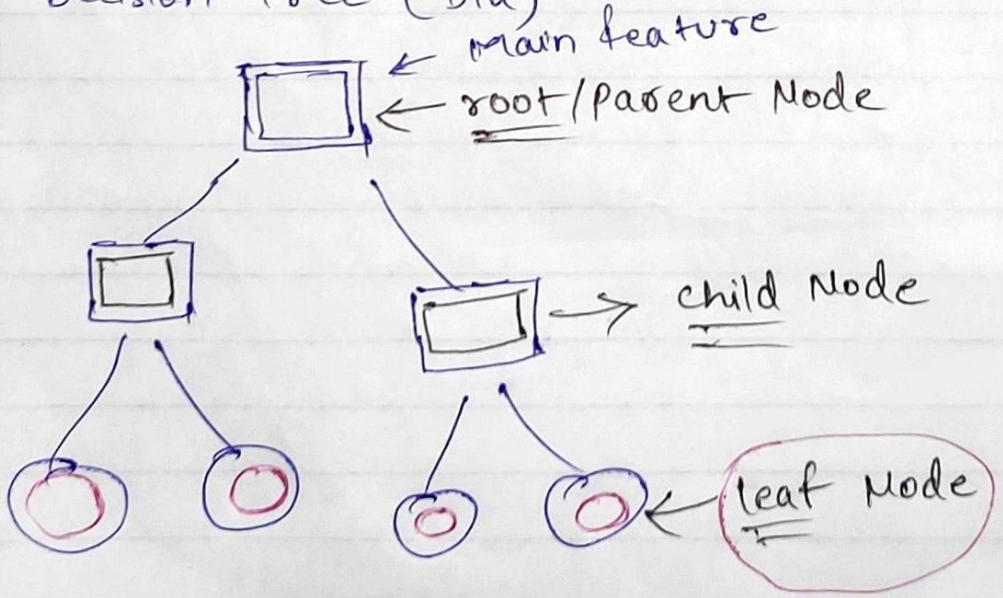
Notes

Machine learning (Decision Tree)

Decision Tree classifier (DTC) and Decision Tree Regressor (DTR)

	f1	f2	f3	f4	f5	O/P.
Day	outlook	temperature	Humidity	wind		yes/No 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	strong		yes
6	overcast	cool	normal	strong		NO
7	sunny	cool	normal	weak		yes
8	rainfall	mild	high	weak		yes

Decision Tree (Dia)



Date.....

Expt. No.....

Page. No.....

Decision Tree classifier (DTC) :

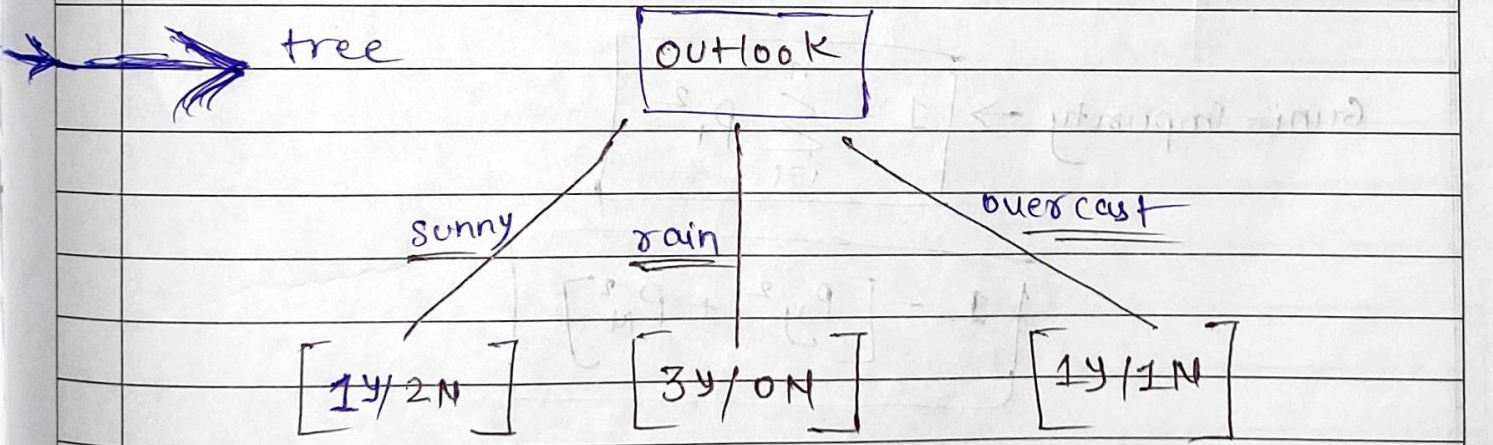
(Num/cat) feature → o/p (cate)

Decision Tree Regressor (DTR) :

features (Num/cate) → o/p (Num)

① ID_3 → Iterative Decotomiser

② CART → classification and Regression tree



Teacher's Signature.....

Purity → ① entropy

② Gini-coff, gini impurity

Entropy : $\sum_{i=1}^{n} P_i \times \log_2(P_i)$

Binary classification ⇒ 2 class → Y/N

Y → Yes N → No

$$-P_Y \log_2(P_Y) - P_N \log_2(P_N) \Rightarrow \text{2 class}$$

Gini coff → $1 - \sum_{i=1}^{n} P_i^2$

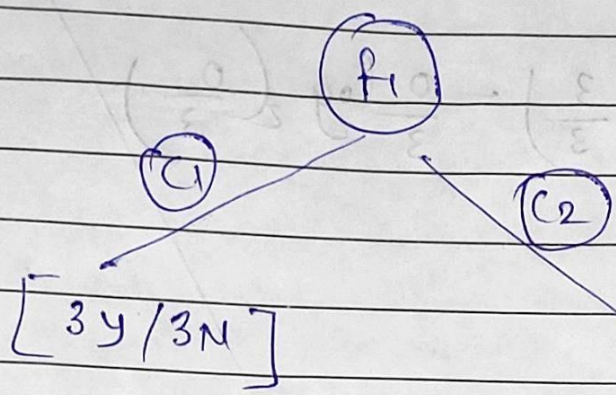
Multi-classification ⇒

3-class Entropy → C_1, C_2, C_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 → $1 - \sum_{i=1}^{n} P_i^2$

$$1 - [P_Y^2 + P_N^2]$$



↓ 50% yes
50% No
impurity
↓ ⇒ entropy = 1
Check
↓ using
① entropy
② Gini-coff / impurity

↓ only yes.
Pure split ∴ No
↓
entropy = 0

Class 1 :-

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

n = No
y = yes

$$= -P_y \log_2(P_y) - P_n \log_2(P_n)$$

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

$$= -\frac{1}{2}\left[\log_2(1) - \log_2(2)\right] - \frac{1}{2}\left[\log_2(1) - \log_2(2)\right]$$

$$= -\frac{1}{2}[0-1] - \frac{1}{2}[0-1]$$

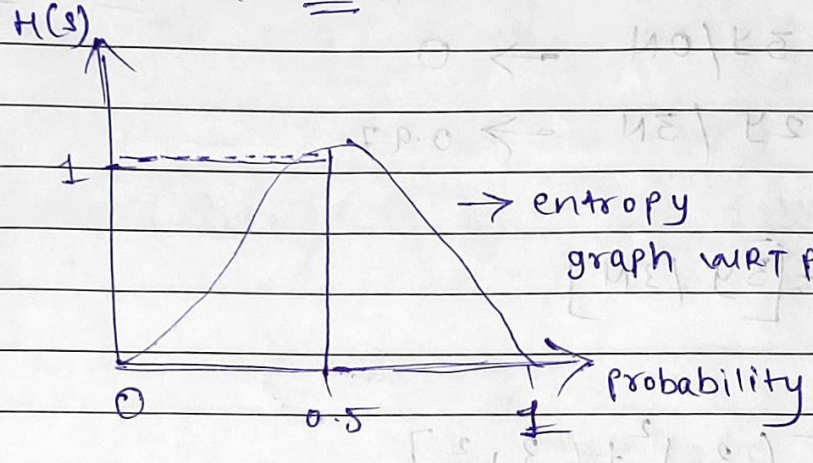
$$= \frac{1}{2} + \frac{1}{2} = 1$$

Teacher's Signature.....

class 2 :

$$-\frac{3}{3}\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$$

$$\Rightarrow 0$$



→ entropy graph WRT Probability.

(50%)

50-50%

$H(S) = 1$ → very impure split

$H(S) = 0$ → Pure split

2 yes / 3 No

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

$$H(S) = 0.97$$

Gini - impurity or Gini - coff :

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

Example : (1) 3Y / 3N → Entropy ⇒ H(S) = 1 (very impure)

(2) 3Y / 0N → 0

(3) 2Y / 3N → 0.97

(1) Gini impurity : [3Y / 3N]

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

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

$$\Rightarrow 1 - \left[\frac{1}{4} \right] + \left[\frac{1}{4} \right]$$

$$\Rightarrow 1 - \left[\frac{2}{4} \right] = 1 - \frac{1}{2} = \frac{1}{2} = \boxed{0.5}$$

(2) [4Y / 8N] = Gini ⇒ $1 - \left[\left(\frac{4}{12}\right)^2 + \left(\frac{8}{12}\right)^2 \right]$

Gini ⇒ 0.44

Date.....

Expt. No..... Page. No.....

(3) [8Y/2N]

$\Rightarrow 1 - \left[\left(\frac{8}{10}\right)^2 + \left(\frac{2}{10}\right)^2\right]$

Gini $\Rightarrow$ 0.32

1 split and 2nd split / or no of split

$$\text{Gain}(S, f_1) = H(S) - \sum \frac{|S_i|}{|S|} H(S_i)$$

To calculate information gain

f1 [9Y/5N]

C1 [6Y/2N]

C2 [3Y/3N]

step1:

(1) [9Y/5N] :

H(S) $\Rightarrow$ root feature entropy

$= -P_y \log(P_y) - P_n \log(P_n)$

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

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

≈ 0.94

Teacher's Signature.....

Date.....

Expt. No.....

Page. No.....

(2) 6Y / 2N

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

$$= 0.81$$

(3) 3Y / 5N $\Rightarrow$ H(S) = 1

$|S_v|$ $\Rightarrow$ total no. of sample after splitting.

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

$$= 0.94 \left[0.462 + 0.42\right]$$

$$\approx 0.049$$

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

H(S) $\Rightarrow$ entropy.

$|S_v|$ $\Rightarrow$ total no. of sample after splitting

$|S|$

$\Downarrow$

ID3 Method

Teacher's Signature.....

Grain(f_1) $\Rightarrow$ 0.049

step 2: (f_2) [9Y/5N]

[5Y/1N] [4Y/4N]

Grain = $0.94 - (6/14) \times 0.65 - (8/14) \times 1$

Grain(f_2) = 0.09 $\Rightarrow$ I.G(f_2)

* $\boxed{IG(f_2) > I.G(f_1)}$

this is greater and it is providing more info.