DT-1

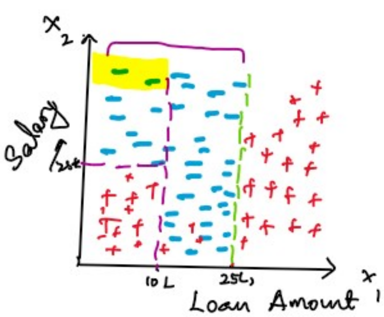
03 December 2022 20:50

Decision Trees

→ Classification (Binary)

$\boxed{ML} \longrightarrow$ $X \longrightarrow Y$ relation
Cost function
Optimization

Principle :- "Divide and Conquer"



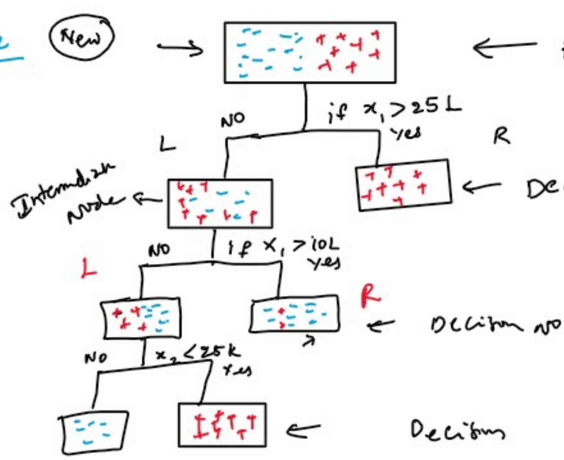
+ - Case (defaulted)
- paid up ✓

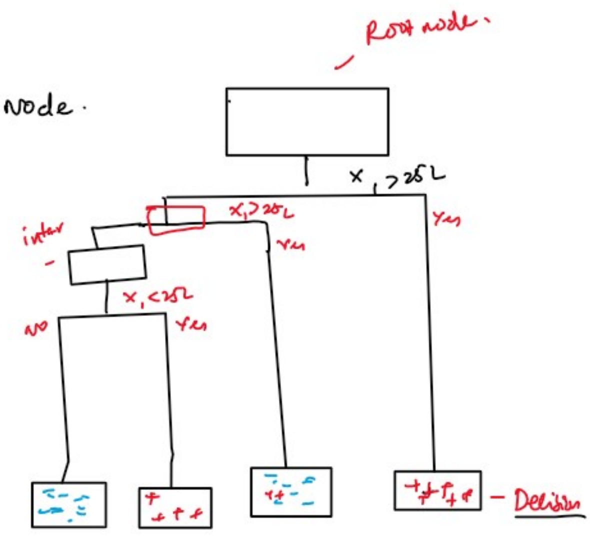
if $x_1 > 25L \rightarrow$ +ve ✓

elif if $x_1 > 10L \rightarrow$ -ve Cases

else if $x_2 < 25k \rightarrow$ +ve

Tree (New) →





(leaf nodes) ✓

Y - Price (100 - 5000)

Nor $[X]$ ← $[Y]$ [100 - 5000]
[0-1]

$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \cdots w_m x_m$

Inn [0-1] w_0 - high value, $w_1, w_2, \ldots w_m$
. 100, 300 . ...

Norm $[Y]$
[0-1]

$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 \cdots w_m x_m$

$Y \sim (\mu, \sigma)$ $w_0, w_1, w_2 \ldots \{0.1, 0.2\}$

$$z = \frac{x_i - \mu}{\sigma} \qquad z \cdot \sigma + \mu = x_i$$

$(\hat{y})$ $Y[0-1] \longrightarrow Y[100 - 5000]$

$$Z = \frac{y_i - \mu}{\sigma} =$$

$$\boxed{y_i = (\sigma \cdot z + \mu)}$$ ✓

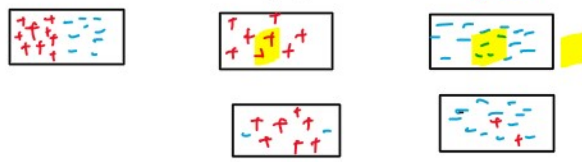
1. Where should we split the data? (for a given variable)

2. Which variable should split the data first?

Learning (Training)

$\boxed{\text{Cost}}$

Good Good



$-y_i \log(p_i) - (1-y_i)\log(1-p_i)$

$\boxed{\text{Metric}}$

50% 50%

Entropy :– $\underbrace{-\sum_{i=1}^{K} p_i \log_2(p_i)}_{-\text{ve sign}}$

(Impurity)

p_i is the probability of ith class

K is the number of classes

[K = 2]

+ve

$\Rightarrow -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right)$

$\Rightarrow -\frac{1}{2}\log_2 2^{-1} - \frac{1}{2}\log_2 2^{-1}$

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

$-\frac{10}{10}\log_2\left(\frac{10}{10}\right) - \underbrace{\frac{0}{10}\times(\quad)}_{0}$

$\Rightarrow 0$

Best worst

cost (0 – 1)

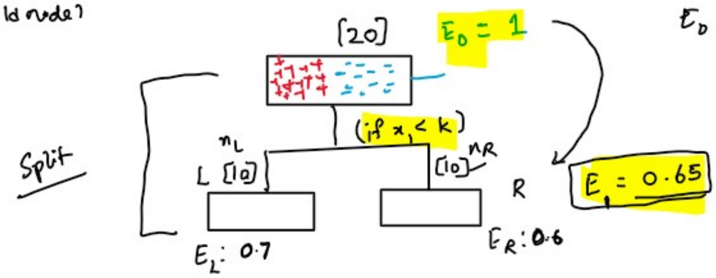
Parent (Child node)

Information gain : $E_0 - E_1$

$E_0 : 1 \quad E_1 = 0.65$

$IG = 1 - 0.65$

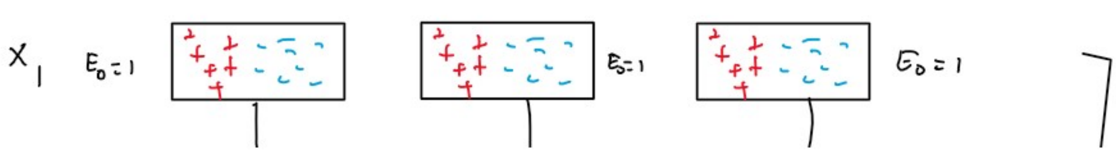
$I.G = 0.35$

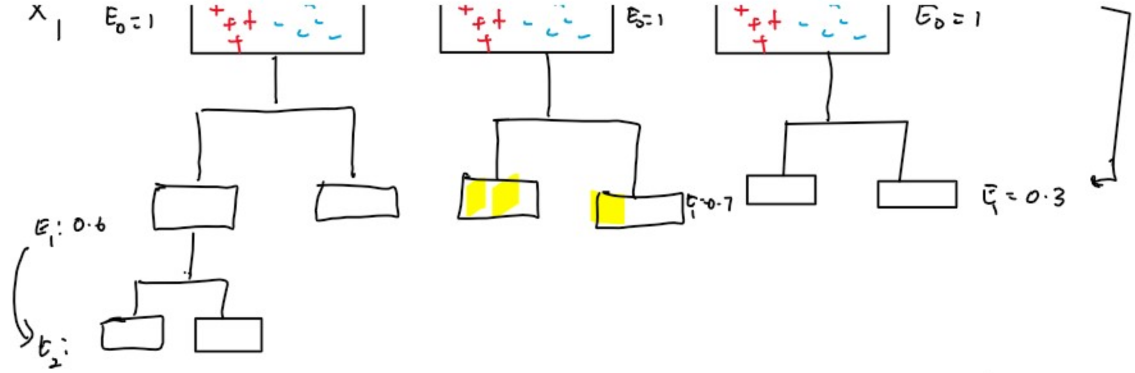


$\rightarrow \left(\frac{n_L}{n} \times E_L\right) + \left[\frac{n_R}{n} \times 0.6\right]$

$\left[\frac{10}{20} \times 0.7 + \frac{10}{20} \times 0.6\right] -$

$x_1 < k \qquad x_1 < Q \qquad \boxed{x_1 < R}$ Best split ✓





IG

0.7 ✓

$\textcircled{x_2}$ $x_2 < m$ $E_0 : 1$

$E_1 : 0.6$

$\boxed{x_2 < n}$ $E_0 : 1$

$E_1 : 0.2$ } 0.8

$x_2 < 0$ $E_0 : 1$

$E_1 : 0.5$

x_1 [ ]

x_2 [ ]

x_3 [ ]

$x_2 < 20$, $x_1 < 24$

Split points

	Sweetness x_1	Y	Switch Case
	1	1	$x_1 < 2$
$x_1 < 2$ →	2	0	$x_1 < 3$
$x_1 < 3$ →	3	1	
→	4	1	
→	5	1	$x_1 < 6$
→	6	0	$x_1 < 7$
→	7	1	$x_1 < 8$
→	8	0	
→	9	0	
$x_1 < 10$ →	10	0	

$\boxed{x_1 < 6}$

0.6

x_2

20

5th → 36

10th → 45

15th → 56

$\boxed{x_2 < 30}$ 8th $= 0.8$

Break 10:28

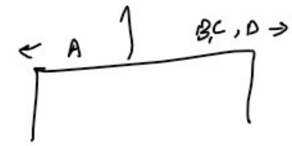
x_1 {A, B, C, D}

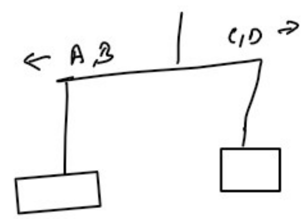
→ A | B, C, D

→ A, B | C D

→ A C | B D

→ A D | B D



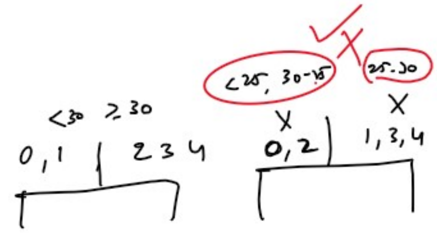


Ordinal variable → [0 1 2 3 4]

<25 25-30 30-35 35-40 40+

<25

0 | 1 2 3 4



5, 5   E - 1

+ + + +   - - - - - -

(6, 4)   (E - 0.8)

$$\text{Entropy} - \sum_{i=1}^{K} P_i \log_2(P_i)$$

$$= \boxed{-\frac{4}{10}\log_2\left(\frac{4}{10}\right) - \frac{6}{10}\log_2\left(\frac{6}{10}\right)}$$

$$= (0.8)$$

Gini Index : Measure of Impurity

$$\text{Gini} \rightarrow 1 - \sum_{i=1}^{K} (P_i)^2$$

P - Prob of a class

i - number of classes

+ + + + - - - -   + + + + +

$$= 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] \qquad = 1 - [1^2 + 0^2]$$

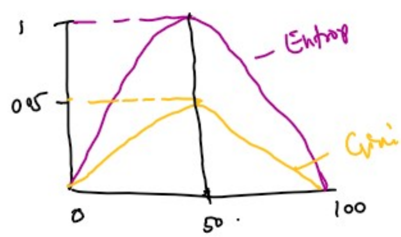
$$= 1 - \left[\frac{1}{4} + \frac{1}{4}\right] \qquad = 1 - 1$$

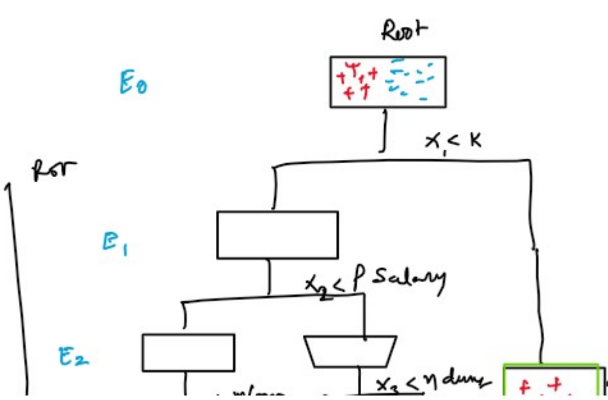
(1980)

$$= 1 - \frac{1}{2} = \boxed{0.5} \qquad = \underline{0} \quad \text{Best}$$

$$\text{Gini} - [0 - 0.5]$$

Gini - Computationally faster



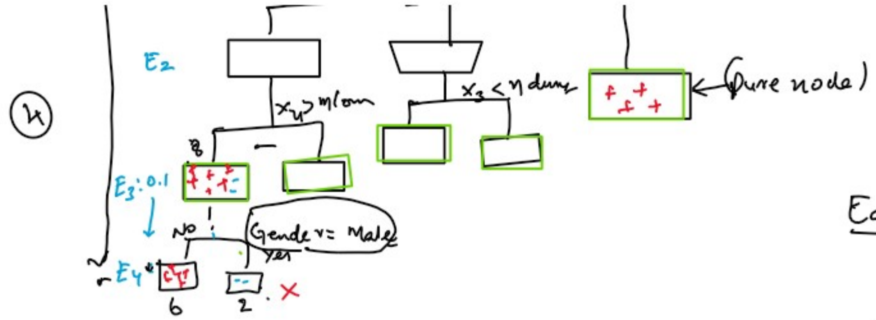


(Pure node)

→ **Stopping** :-

① When we have a pure node.

② When there is no further I.G.



(2) When there is no further I.G

Early Stopping Criteria (Pruning) →

1 → Highly prone to overfitting ✓

→ Greedy

✓ 1 Depth of the tree ✓

$\boxed{\text{Max-depth} = 3}$

→ ✓ 2 Min no. of observation to split ✓

10 ×

→ ✓ 3 Min no. of observations in leaf node = (5) ✓

Entropy

(0 - 1)

$P_+ = P$

$P_- = 1 - P$

$-y_i \log(P_i) - (1-y_i)\log(1-P_i)$

P_+ P_-

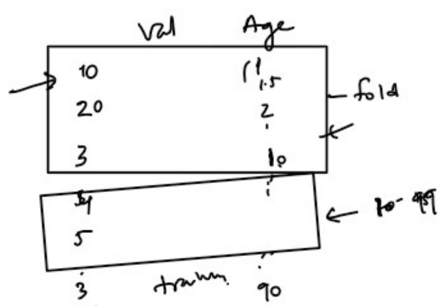
$-P\log_2(P) - (1-P)\log_2(1-P)$ ←

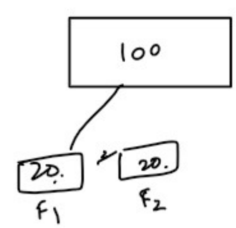
$$\text{Entropy} = -\sum_{i=1}^{K} P_i \log_2(P_i)$$

$$\text{Gini} = 1 - \sum_{i=1}^{k} (P_i)^2$$

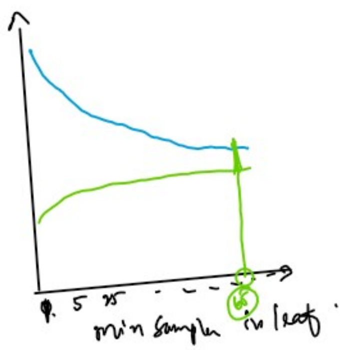
Start : 9:05

1 folds = KFold($\boxed{\text{n\_splits} = 5}$, $\boxed{\text{Shuffle} = \text{True}}$, random_state = 1)

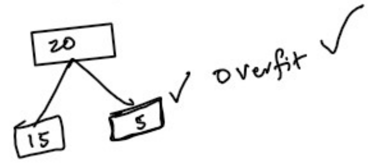




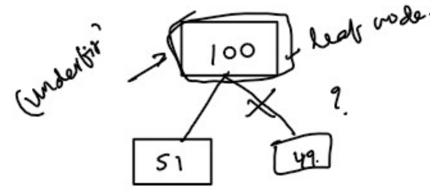
Plot



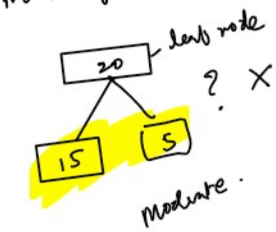
min Samples in leaf = 5 ✓



min Sample in leaf: 50



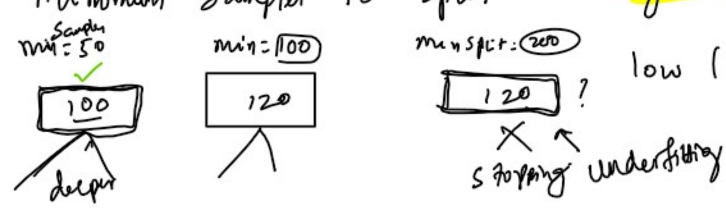
min Sample in leaf = 10



→ Minimum Samples to Split

high (?) — underfit

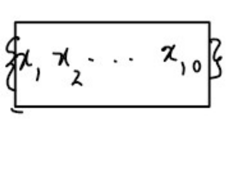
low (?) — overfit

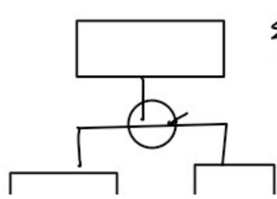


max features = 4

Split points < 'best' approach / 'Random' ←

age{10 — 99}





{ x_1, x_3, x_8, x_9 }

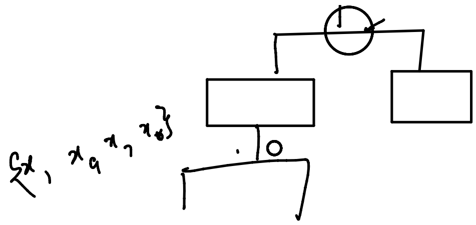
age < 21

age < 36

age < 61

age < 20

age < 25



age < 60

max_features = 'auto'

Classification $\sqrt{m}$ m no. of features

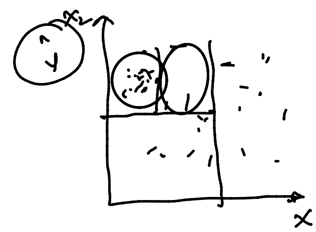
Regressn $\frac{m}{3}$ m no. of features

Classification

Regression ?

Entropy } for regression ?
Gini

$-\sum p_i \log_2(p_i)$ ✗



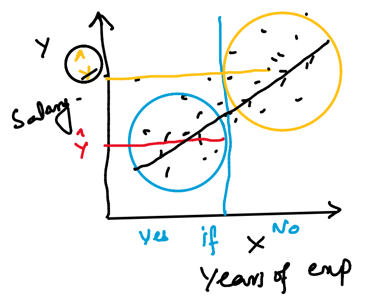
if $x < 5$ $\hat{y}$ = mean(y_L)

$x > 5$ $\hat{y}$ = mean(y_R)

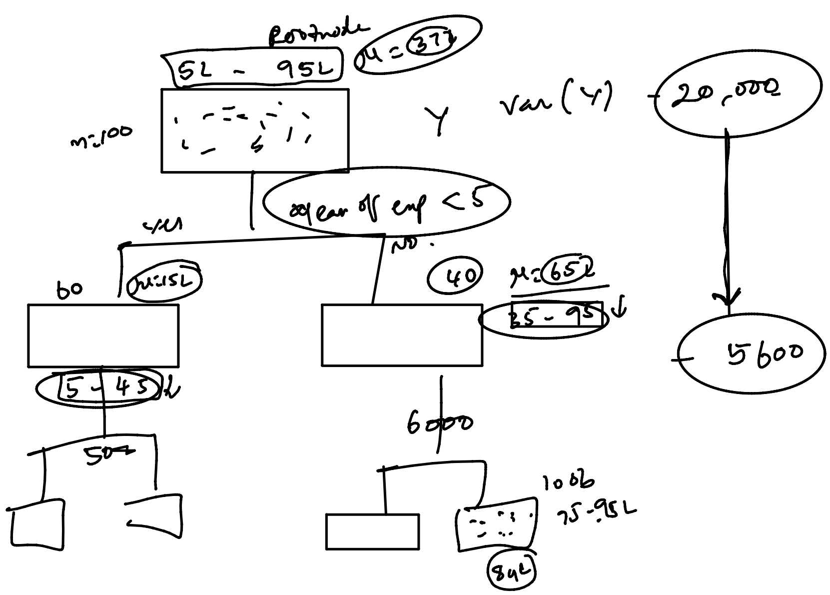
MSE

if years < 5 avg salary 15L

> 5 avg sal 30L



if $x < 5$



Multiclass Classification - Entropy, Gini work?

$$- \sum_{i=1}^{K} P_i \log_2(P_i)$$

K-classes

$\boxed{10 \quad 20 \quad 30}$

$$- \frac{10}{60} \log_2\left(\frac{10}{60}\right) - \frac{20}{60} \log_2\left(\frac{20}{60}\right) - \frac{30}{60} \log_2\left(\frac{30}{60}\right)$$

(Binary)

Classification and Regression Trees (CART) → Ensembles