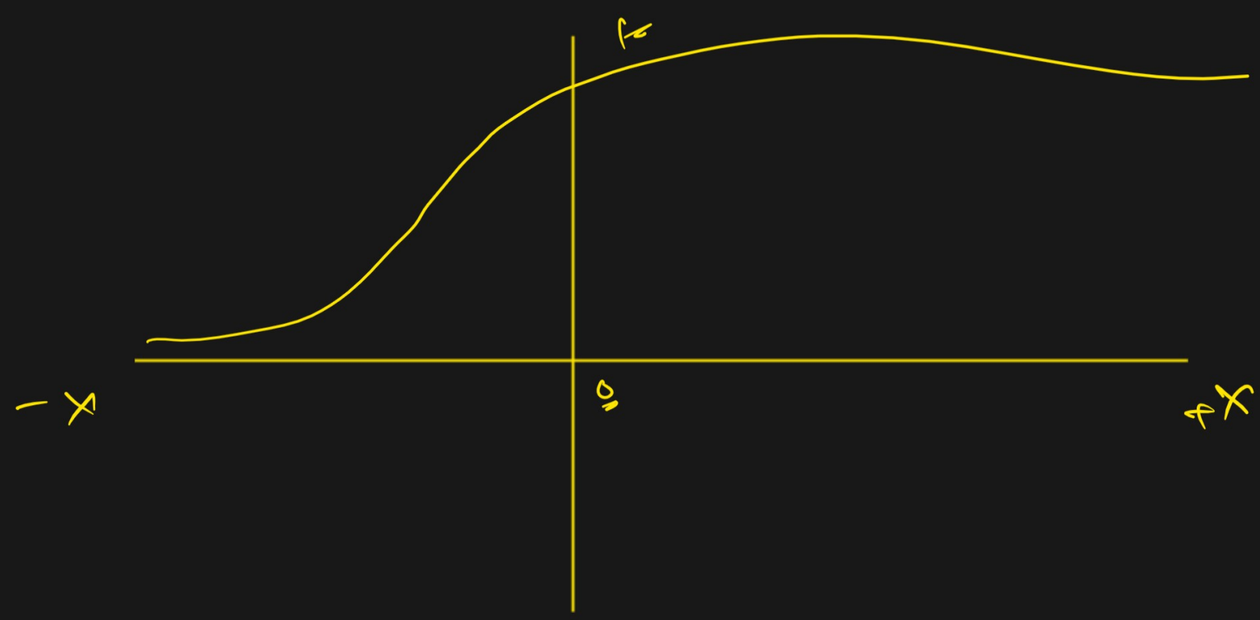
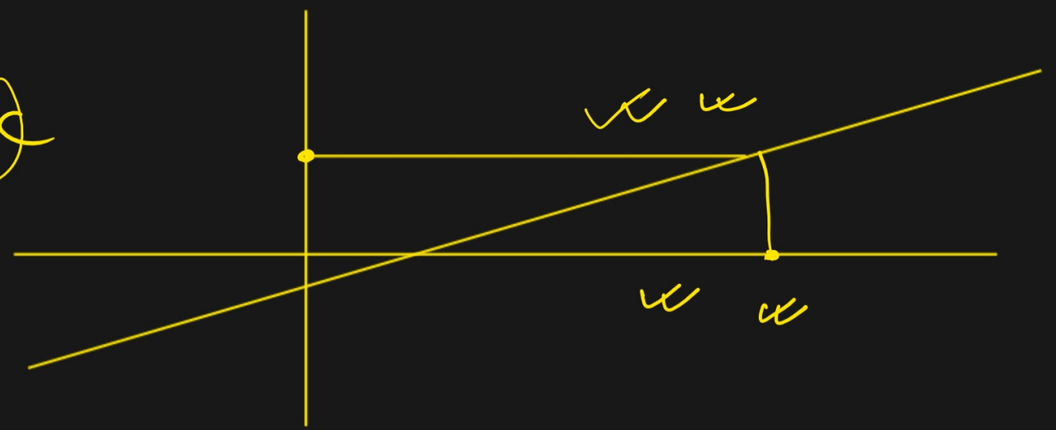
logistics Regression

$$\sigma(z) = \left(\frac{1}{1 + e^{(-z)}} \right) \approx \text{Sigmoid function}$$



$$z \simeq y = (mx + c$$



$$(\sigma) = \left(\frac{1}{1 + e^{-(mx+c)}}\right) \quad \begin{cases} 0 \\ 1 \end{cases}$$

$\sigma \geq 0.5 \approx 1$ ✓

$\sigma < 0.5 \approx 0$

log-loss ((or) Entropy loss)

$$= -\frac{1}{m} \sum_{i=1}^{m} \left[y \log(h_\theta(x)) + (1-y) \log(1 - h_\theta(x)) \right]$$

y = the actual outcome

$(\hat{y}) \approx h_\theta(x)$ = Predicted Probability.

m = no of training data.

$$\sigma(z) = \left(\frac{1}{1 + e^{-(mx+c)}} \right)$$

x_1	x_2	~~x_3~~	~~x_4~~	y
				A B C D

(OUR)

(OUR)

$$m_n = m_0 - \eta \left(\frac{\partial L}{\partial m} \right) \quad \text{(LR)} \quad \text{(LR)} \quad \rightarrow (10)$$

$$c_n = c_0 - \eta \left(\frac{\partial L}{\partial c} \right) \quad \text{(LR)} \quad \text{(LR)} \quad \rightarrow (11)$$

$$\text{Softmax}(x) = \left(\frac{(e^x)}{\sum_{i=1}^{n} e^x} \right)$$

(OUT)

$$\begin{bmatrix} 1.3 \\ 5.1 \\ 2.2 \\ 0.7 \\ 1.1 \end{bmatrix} \rightarrow \begin{pmatrix} \frac{e^{1.3}}{\sum e^{x}} \\ \frac{e^{x}}{\sum e^{x_i}} \end{pmatrix} = \begin{bmatrix} 0.02 \\ 0.90 \\ 0.70 \\ 0.01 \\ 0.01 \end{bmatrix}$$

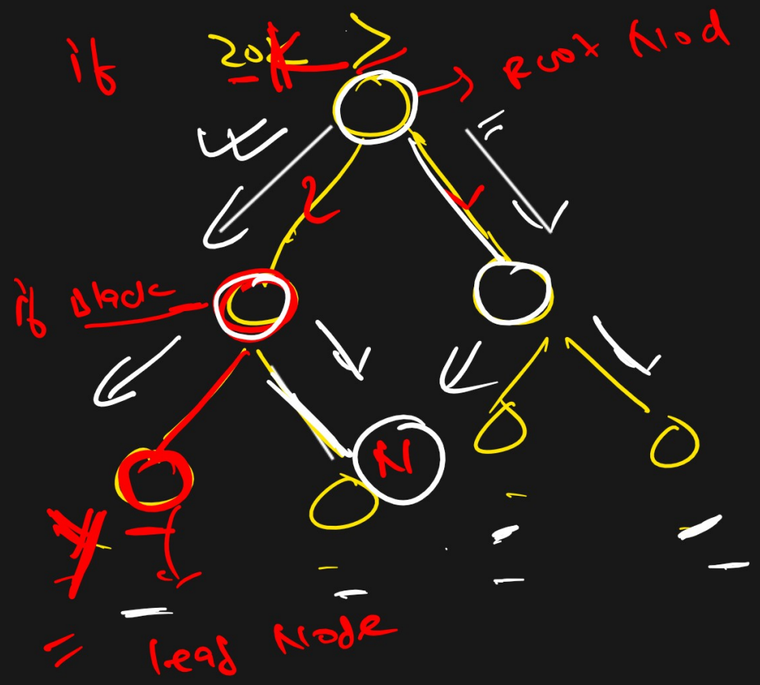
$$e^{1.3} + e^{5.1} + e^{2.2} + e^{0.7} + e^{1.1}$$

$$\left(\frac{1}{1+e^{-(mx+c)}}\right) \quad (\text{OVO})$$

$$\left(\frac{e^{(mx+c)}}{\sum e^{(mx+c)}}\right) \quad (\text{OVR})$$

$$= -\frac{1}{m}\sum_{i=1}^{n}\left(y\log(\hat{y}_p) + (1-y)\log(1-\hat{y}_p)\right)$$

Decision Tree (C, R)



P	C	D	
20	B	Y	✓
40	R	N	✓
10	Blue	Y	✓
18	B	N	✓

Classification

① Gini Impurity $= \left(1 - \sum p_i^2\right)$

p_i = Prob of class i in the given node.

② Entropy (Information Gain)

$$E = -\sum p_i \log(p_i)$$

$$IG = E_{parent} - \sum \left(\frac{|s_i|}{|s|} E_i \right)$$

Regression -> (DT)

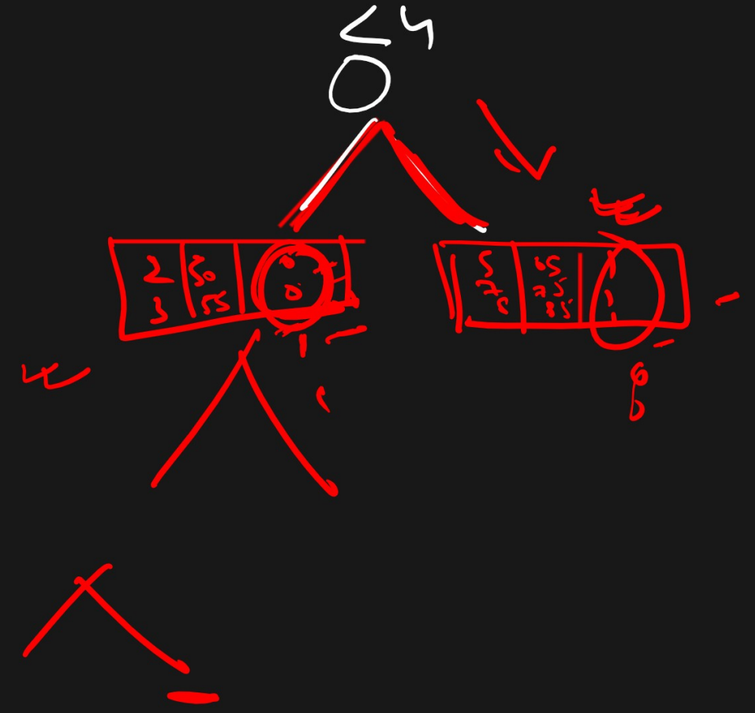
$$MSE = \frac{1}{N} \sum (y - \hat{y})^2$$

$$Variance = \frac{1}{N} \sum (y_i - \bar{y})^2$$

$$G_L = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2 \right) = 0.44$$

$$G_R = 1 - \left(\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2 \right) = 0 \qquad G_{split} = \frac{2}{5} \times 0.44 + \frac{3}{5} \cdot 0 = 0.18$$

Student	Study Hrs	Previous Score	Pass	Fail
A	2	50	0	
B	3	55	0	
C	5	65	1	
D	7	75	1	
E	8	80	1	
F	6	68	1	



$$G = \left(1 - \sum p_i^2\right)$$

$P_P = 3/5$

$P_F = 2/5$

$$G = 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right)$$

$$= 1 - \left((0.6)^2 + (0.4)^2\right)$$

$$= 1 - \left(0.36 + 0.16\right) = 0.48$$

Gini Index = 0.48

Split at study Hours = (4) ≤

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

$$G_R = 1 - \left(\left(\frac{0}{3}\right)^2 + \left(\frac{3}{3}\right)^2\right) = 0$$

$$G_{split(4)} = \frac{2}{5} \cdot 0 + \frac{3}{5} \cdot 0$$

$$= 0$$

DT → Regression

House	Size (Sq ft)	Price
H1	1000	50000
H2	1200	60000
H3	1500	70000
H4	1800	85000
H5	2000	95000
H6	1700	?

(1700, 2000)

$\bar{y}_L = 55000$

$MSE_L = 25000000$

$\bar{y}_R = 83333.33$

$MSE_R = 105555556$

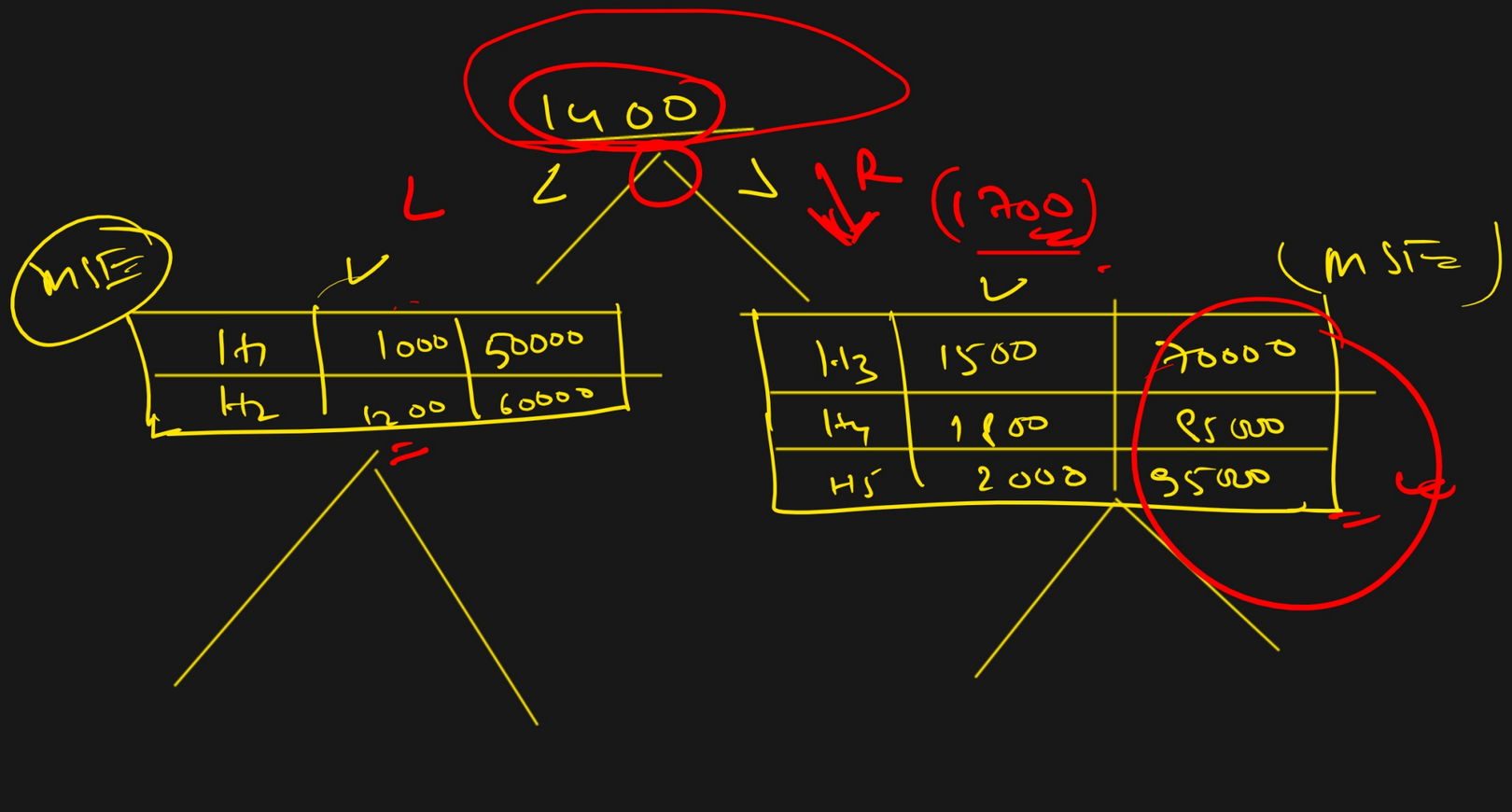
$MSE_{split} = \frac{2}{5} \times 25000000 + \frac{3}{5} \times 105555556 = 73333333$

$MSE = \frac{1}{N} (y - \bar{y})^2$

$\bar{y} = 72000$

$MSE_{initial} = (265\ 200000)$

$MSE_{split} < MSE_{initial}$

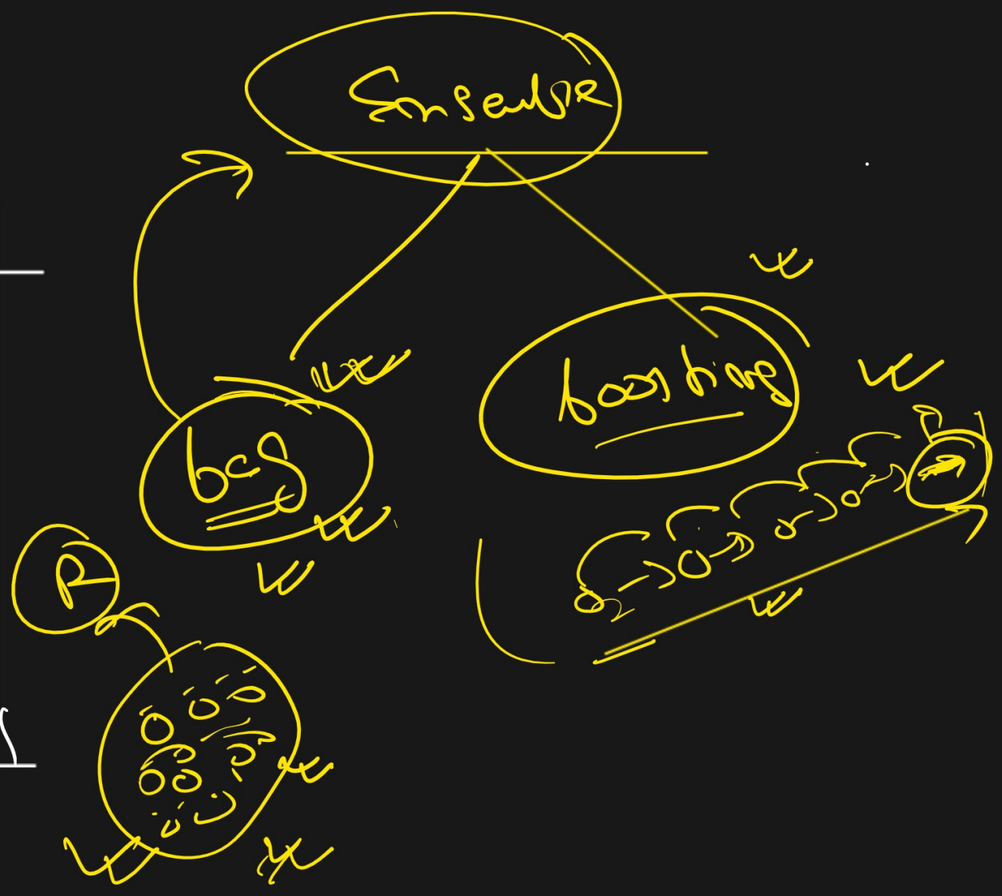


(Random Forest) (C/R)

Ensemble learning tech :-

Data

DS	C1	C2	C3	C4	Y
A					
B					
C					
D					
E					
F					
G					
H					



Bootstrap Sampling :-