// Decision Tree

(i) DT classifier

(ii) DT Regressor

(i) DT classifier – Type of Techanique

(i) ID3

(ii) CART

✱ Entropy and Gini Index → purity split in dataset

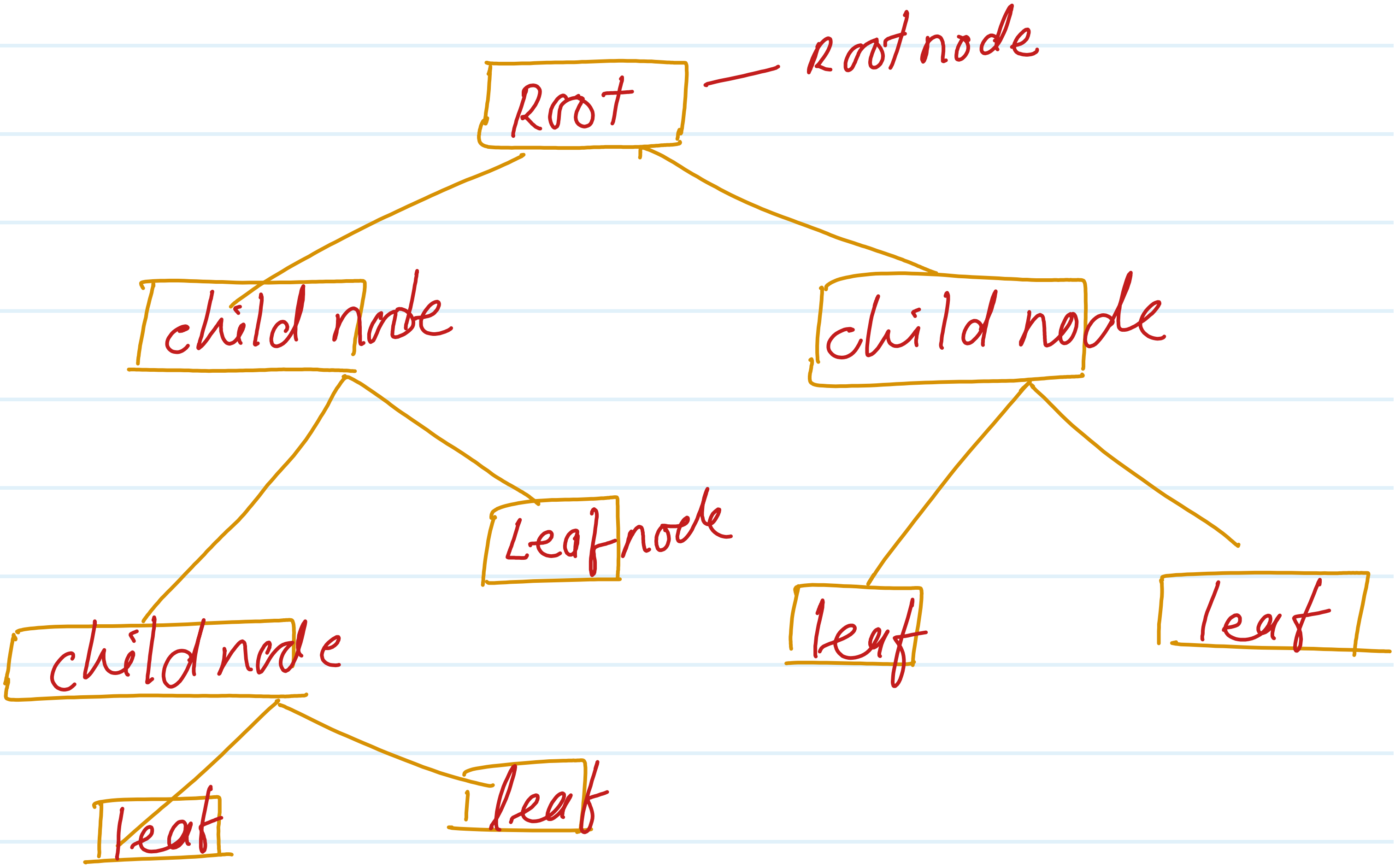
✱ Information Gain → DT feature split

weight	height	o/p obese/Noobese.
60	160	ob
70	170	No
80	180	ob
90	190	No
100	200	No

(11) DT Regressor

Regression we use standard deviation / MSE / MAE

weight	height	BMI
60	160	21
70	170	22
80	180	20
85	190	23
90	195	24

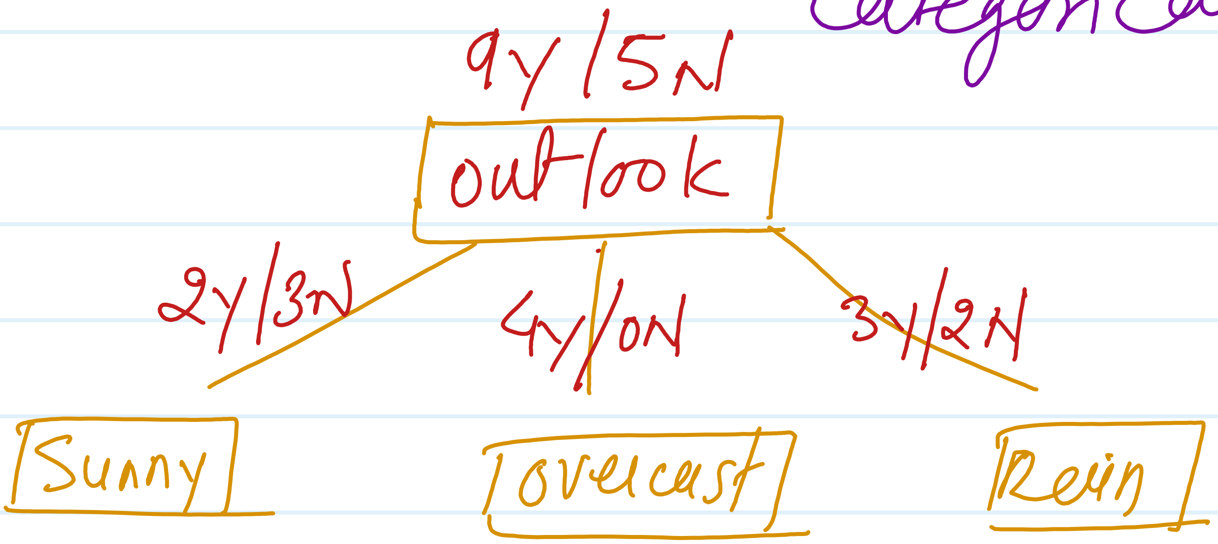


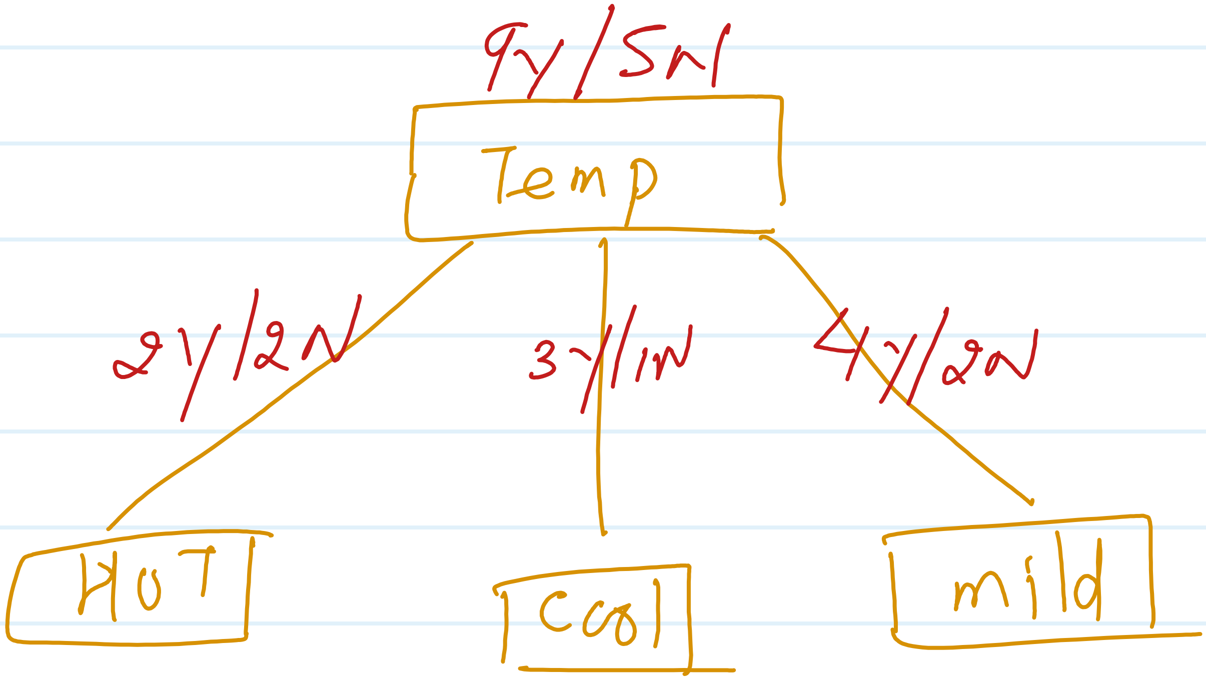
DT classifier

outlook	Temp	humidity	wind	play
sunny	H	High	weak	N
sunny	H	H	strong	N
overcast	H	H	W	Y
rain	M	H	W	Y
rain	C	Normal	W	Y
rain	C	N	S	N
overcast	C	N	W	Y
sunny	M	H	W	N
sunny	C	N	W	Y
rain	M	N	W	Y
sunny	M	N	S	Y
overcast	M	high	S	Y
overcast	H	N	W	Y
rain	M	H	S	N

Feature can be numeric and categorical

Output can be numeric and categorical





ⓐ Entropy $H(s)$

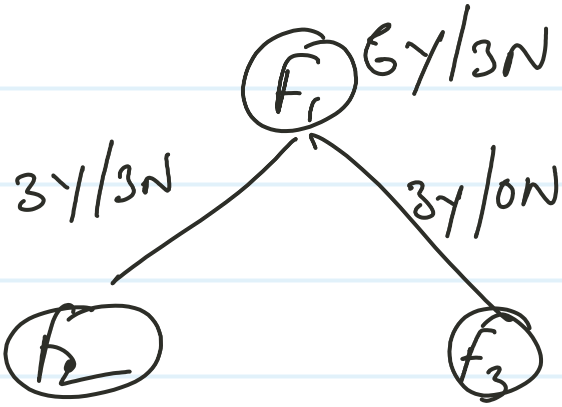
Formula (Binary)

$$H(s) = -P_{yes}\log_2(P_{yes}) - P_{No}\log_e(P_{No})$$

multiclass

$$H(s) = -P_{c_1}\log_2(P_{c_1}) - P_{c_2}\log_2(P_{c_2}) - P_{c_3}\log_2(P_{c_3})$$

Exm



$C_1 \Rightarrow H(s) = -\frac{3}{6}\log\left(\frac{3}{6}\right) - \frac{3}{6}\log\frac{3}{6}$

$\Rightarrow \underline{1}$

$C_2 \Rightarrow H(s) = -\frac{3}{3}\log\frac{3}{3} - \frac{0}{3}\log\frac{0}{3}$

$\Rightarrow 0$

For the pure split of feature

pure entropy should be zero (0)

for impure split = 1

③ Gini index (Impurities).

main formula –

$$G.I. = 1 - \sum_{i=1}^{n} (P)^2$$

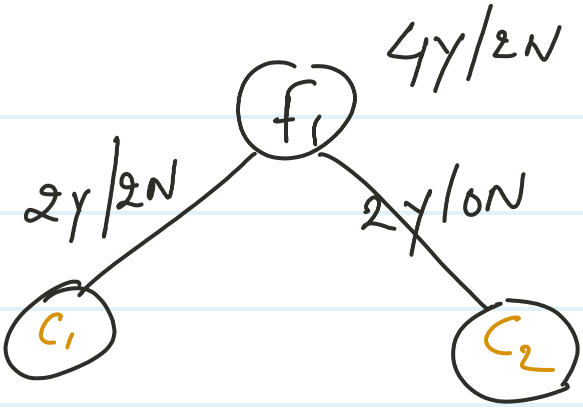
binary class

$$G.I. = 1 - \sum_{i=1}^{n} \left[(P_{c_1})^2 + (P_{c_2})^2 \right]$$

multiclass

$$G.I. = 1 - \sum_{i=1}^{n} \left[(P_{c_1})^2 + (P_{c_2})^2 + (P_{c_3})^2 + \cdots \right]$$

Exm



$$C_1 \Rightarrow G.I. = 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right]$$

$$\boxed{= 0.5}$$

$$C_2 \Rightarrow G.I. = 1 - \left[\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right]$$

$$\boxed{= 0}$$

Range of entropy = 0 to 1

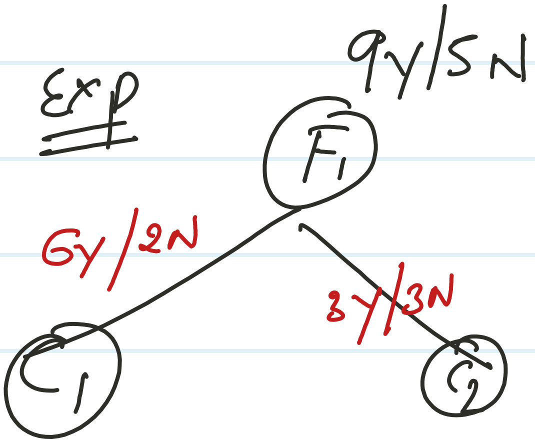
Gini impurity (index) = 0 to 0.5

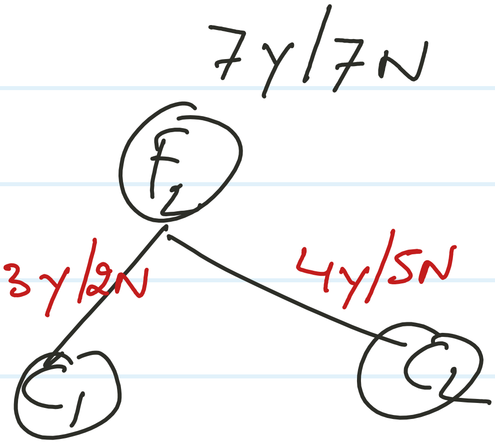
(c) Information Gain

formula -

$$gain(S, f_1) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$$

Exp





For F_1 ⇒

$$H(S) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14}$$

$$\boxed{H(S) = 0.94}$$

C_1 ⇒ $H(S) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}$

$$\boxed{H(S) = 0.81}$$

$C_2 \Rightarrow$

$$H(s) = -\frac{3}{3}\log\frac{3}{3} - \frac{0}{3}\log\frac{0}{3}$$

$$\boxed{H(s) = 1}$$

gain of $f_1 \Rightarrow$

$$\text{gain}(s, f_1) = 0.94 - \left[\frac{8}{14} \times 0.81 + \frac{6}{14} \times 1\right]$$

$$\text{gain}(S, f_1) = \underbrace{0.049}$$

$f_2 \rightarrow$ $$H(s) = -\frac{7}{7}\log\frac{7}{7} - \frac{7}{7}\log\frac{7}{7}$$

$$= 0$$

$C_1 \rightarrow$ $$H(s) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5}$$

$$= 0133 + 0.159$$

$$\boxed{= 0.29}$$

$$C_2 \Rightarrow H(S) = -\frac{4}{9}\log\frac{4}{9} - \frac{5}{9}\log\frac{5}{9}$$

$$\boxed{= 0.014}$$

$$f_2 \quad \text{gain}(S, f_2) = 0 - \left[\frac{5}{14}\times 0.29 + \frac{9}{14}\times 0.019\right]$$

$$= 0 - \left[0.10 + 0.009\right]$$

$$= -0.10$$

F_1	F_2	F_3	F_4
0.49	0.56	0.025	0.10

Since F_2 has highes value of information gain's among the all feature so that It will be our root node.

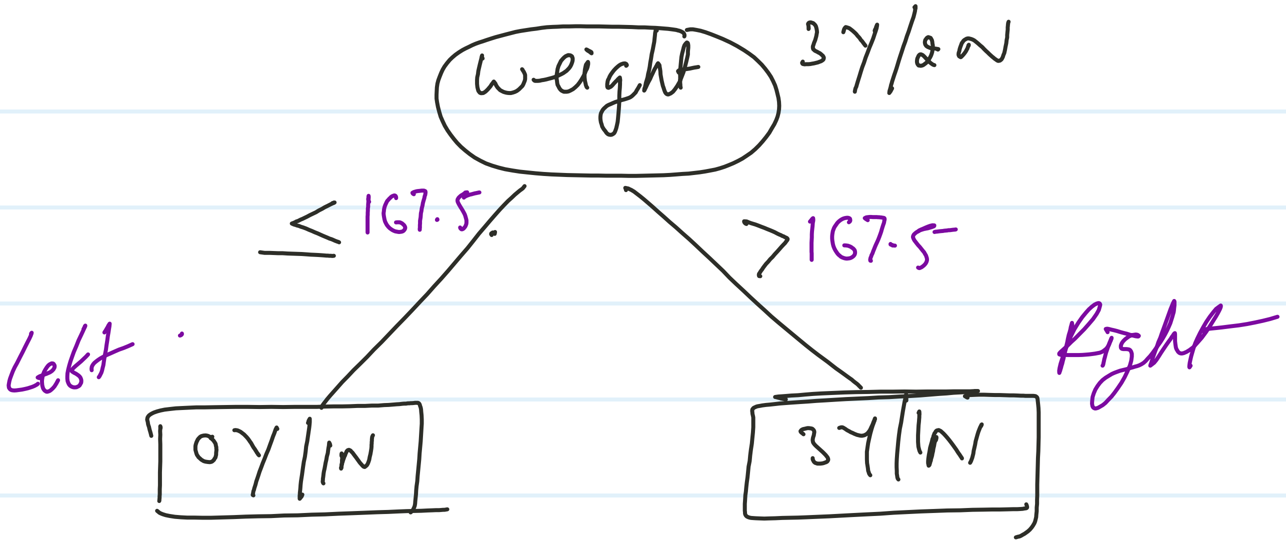
✫ Independent analysis befor making DT

build DT with numerical feature

weight	heart De.
220	Y
180	Y
225	Y
190	N
155	N

weight		Heart
155		N
	167.5	
180		Y
	185	
190		N
	204	
220		Y
	222.4	
225		Y

with respect to every point avg. value need to find out gini index / entropy



$$\text{Gini impurity} = 1 - \sum_{i=1}^{n} P_i^2$$

$$\text{gini (Left)} = 0$$

$$\text{gini (Right)} = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$$

$$= 0.375$$

$$\text{Information gain} = \text{G.I.}[\text{Root}] - \sum_{\text{Value}} \frac{|S_v|}{|S|}\,\text{G.I.}[\text{child}]$$

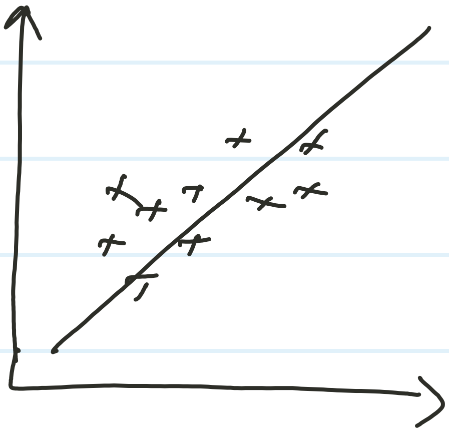
$$\text{G.I. [Root]} = 0\ 48$$

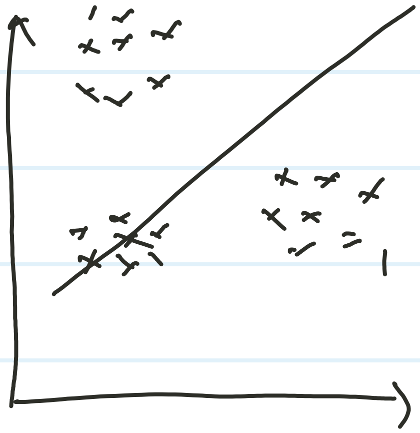
$$\text{I.G.}[167.5] = 0.48 \left[\frac{1}{5} \times 0 + \frac{4}{5} \times 0.375\right]$$

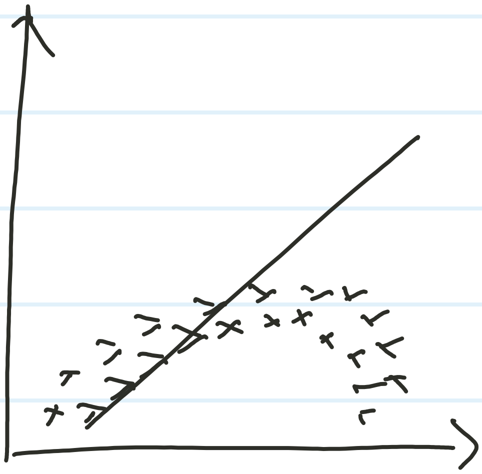
$$\text{I.G.}[167.5] = 0.18$$

Information Gain should be high and Gini Index should be low.

✵ DT Regression :-







Ex

	height	weight
	165	65
162.5		
	160	50
167.5		
	180	90
172.5		
	170	85
177.5		
	175	70

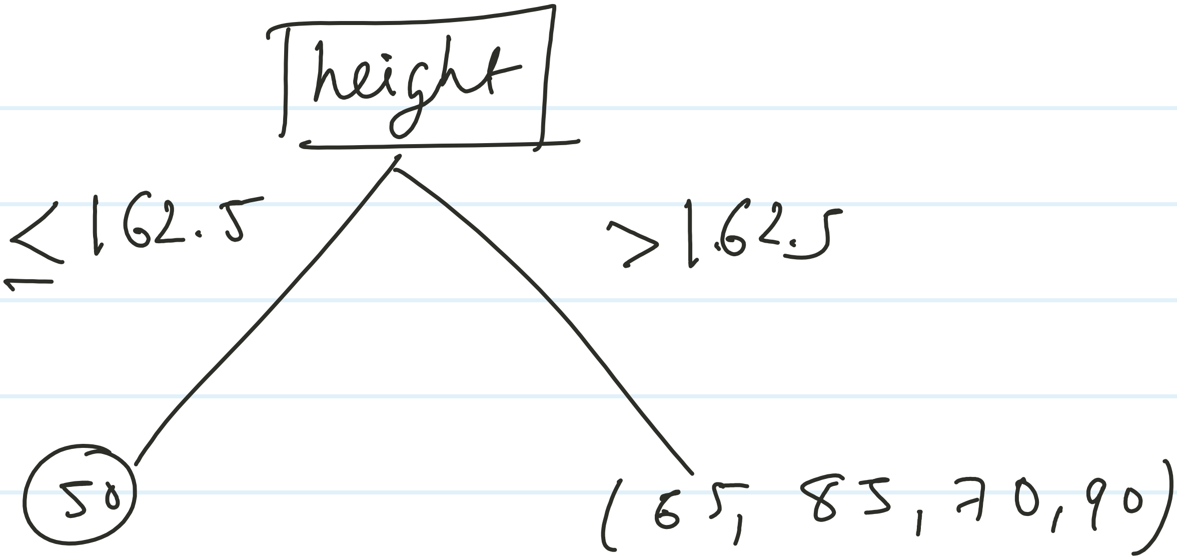
160
165
170
175
180

Regression problem weight calculated with respect to height

① Step: - Short the value of height column (x Featu)

② step - Find Adjcent Avg. value b/w data point

③ step - Find Information gain with help of entropy and Gini Index.



mean = 77.5

Regression -

① mean = 77.5

② MSE, RMSE, MAE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2$$

$$\frac{50 + 65 + 85 + 70 + 90}{5} = 72.$$

$$\text{height (variance)} = \frac{(72-50)^2 + (72-65)^2 + (72-85)^2 + (72-70)^2 + (72-90)^2}{5}$$

hight (variance) = 206

$$\text{var(Right)} = \frac{(77.5-65)^2 + (77.5-85)^2 + (77.5-70)^2 + (77.5-90)^2}{4}$$

$\text{var(Right)} = 106.25$

$\text{var(left)} = 50$

Reduction in variance

$$= \text{var(root)} - \sum_{i=1}^{n} w_i \times \text{var}[\text{child}]$$

$$= 206 - \left[\frac{1}{5} \times 0 + \frac{4}{5} \times 106.25\right]$$

Reduction variance $= 121$

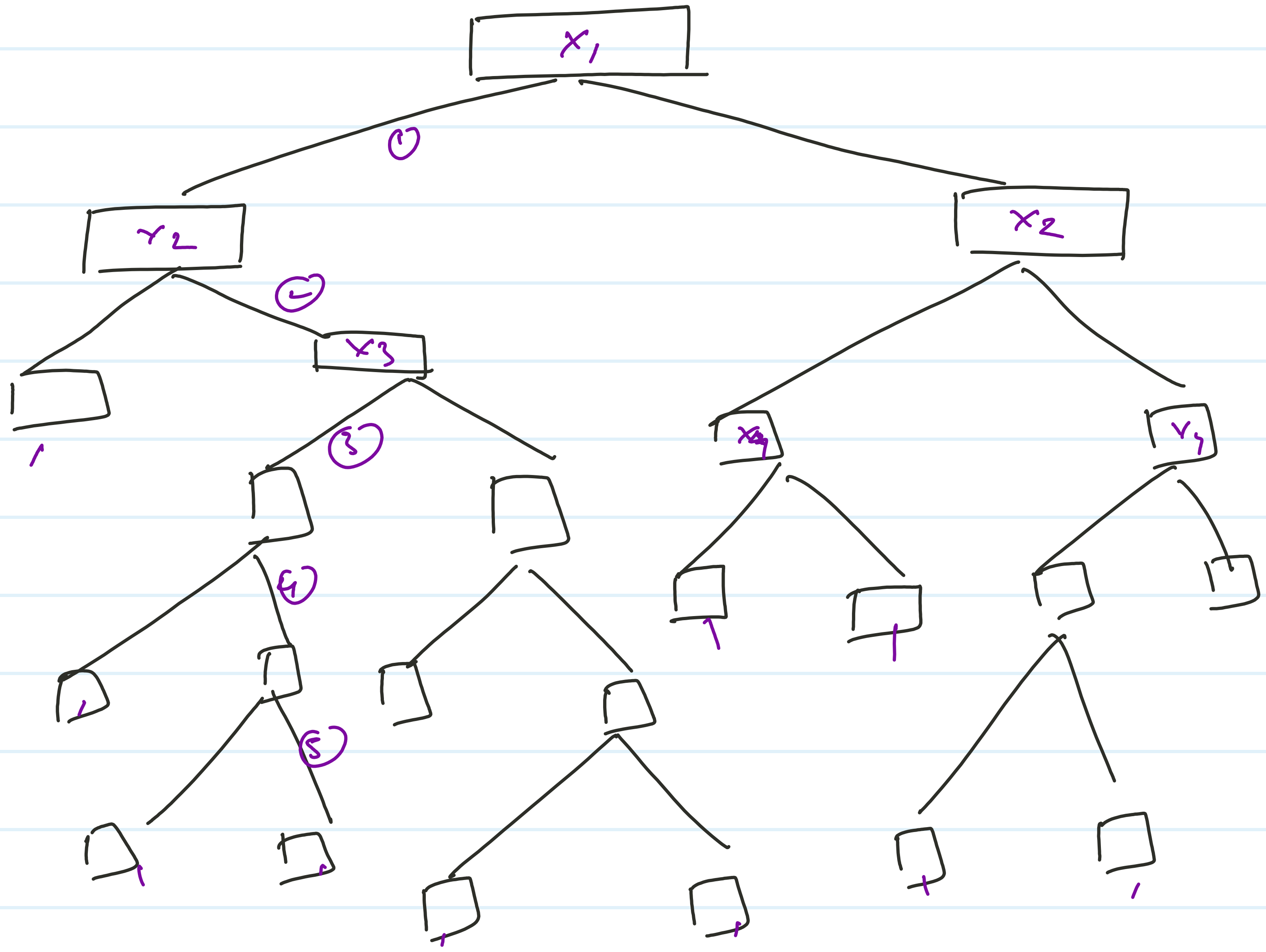
we calculate MSE for all the datapoint whichever is less will be thresold.

height	gender	weight
160	M	65
165	F	70
170	M	80
175	M	90
180	F	100

from height / Gender, . choose root node

for height mse = 55.5
for gender mse = 53

so value of gender mse is less
It will be our root node.



Pre-pruning and post-pruning

max-Depth = 5

mini-sample-leaf = 10

min-sample-split = 8

max-feature = 6

These 4 hyperparameter selected for pre-pruning before build DT. algorithms.

Post pruning =>

① make DT till end

② cut DT. using ccp value.

③ ccp value is nothing but thersold for gini / Entropy.

ccp value is responsible for depth of Tree.. If ccp is less, the depth will be less.
High ccp value the depth will be more.

ccp = [0.4, 0.5, 0.6, 0.01]

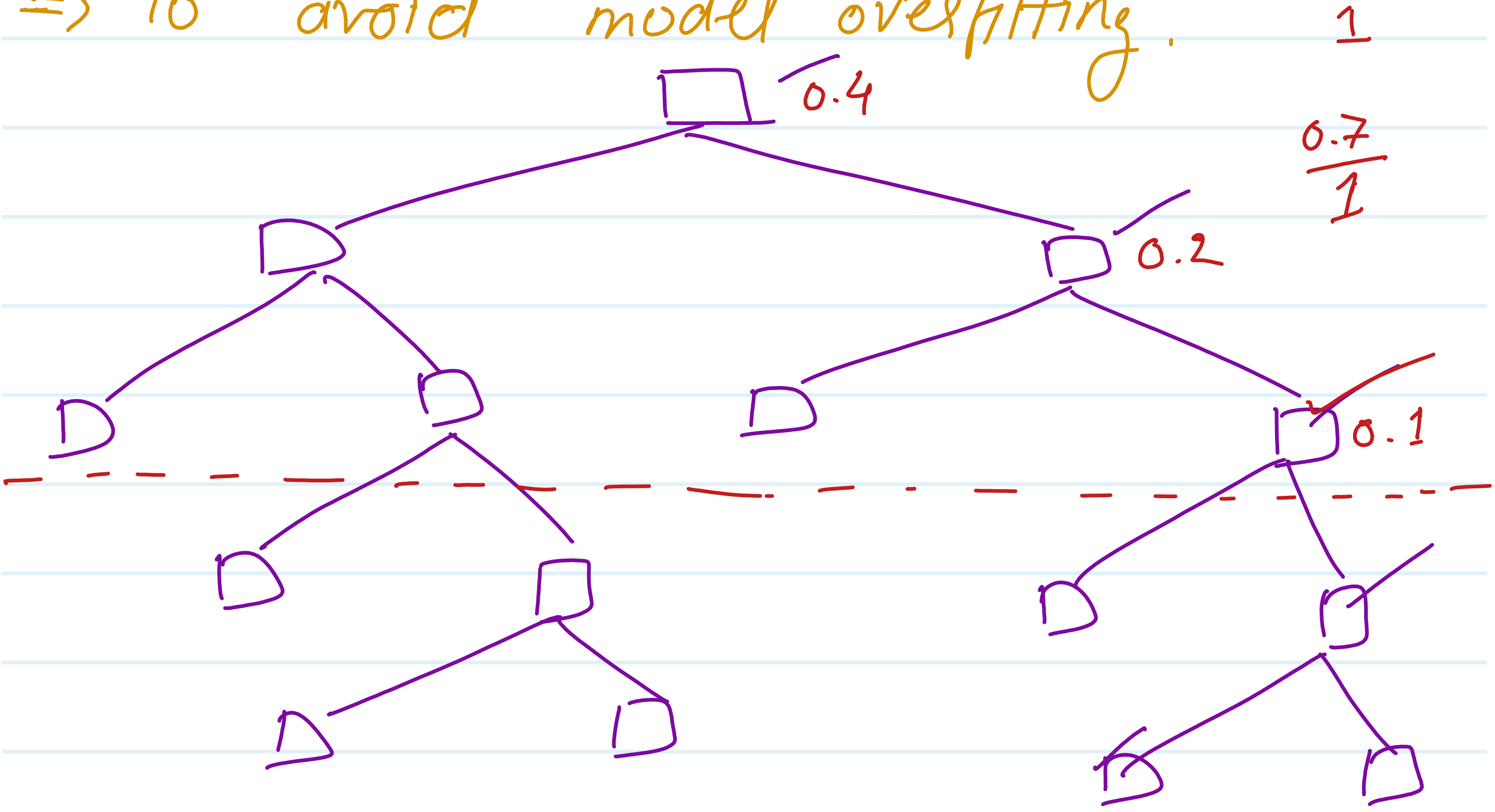
For model training either we can use pre-prunning or post-prunning.

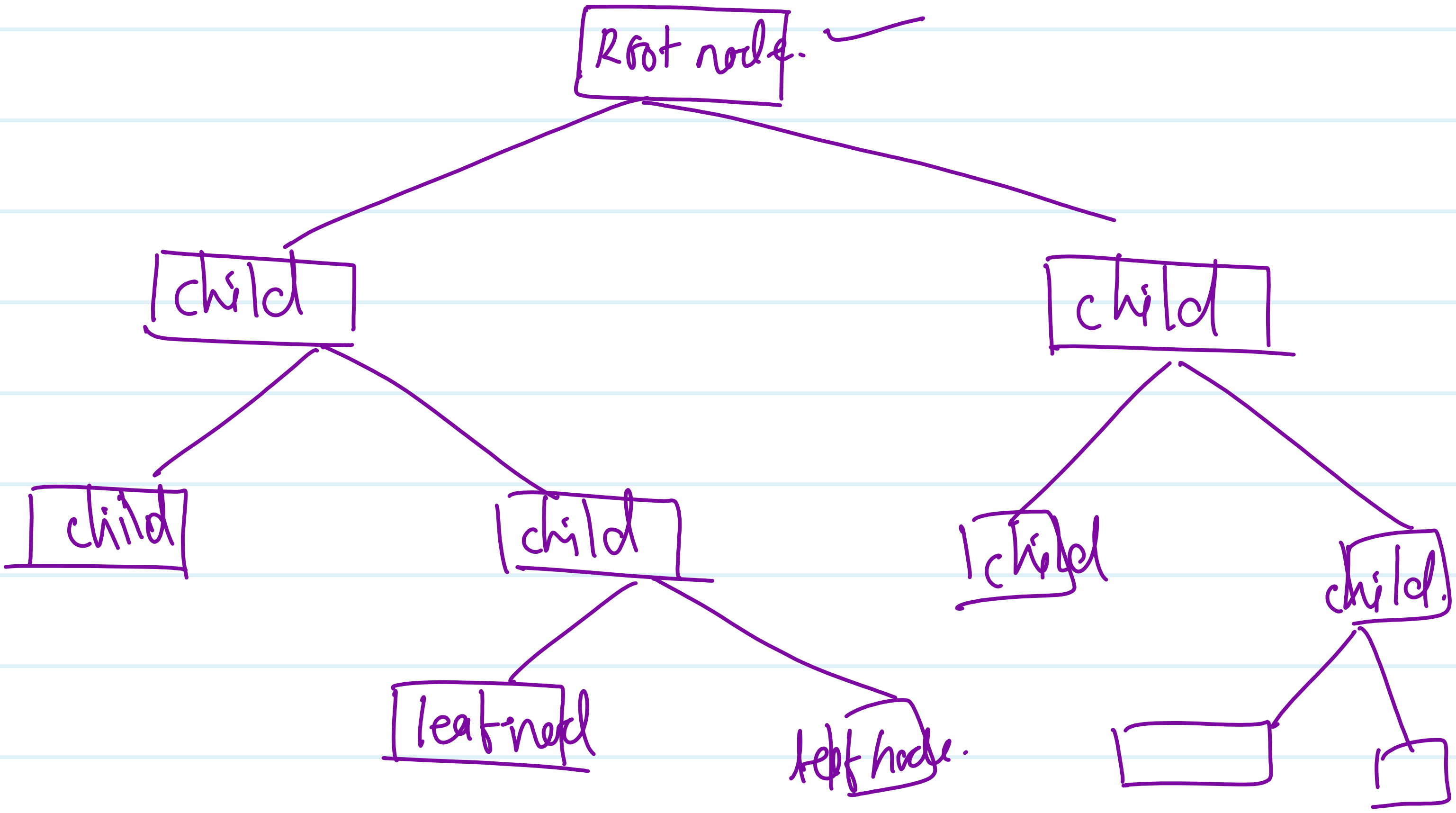
① When we have large dataset at this time we use pre-prunning.

② When we have small dataset at this time we use postprunning.

Why we use post or pre-pruning?

⇒ To avoid model from overfitting.




Root node.
child
child
child
child
child
child
leaf node
leaf node.

x_1	x_2	x_3	Y

$X_4 =$