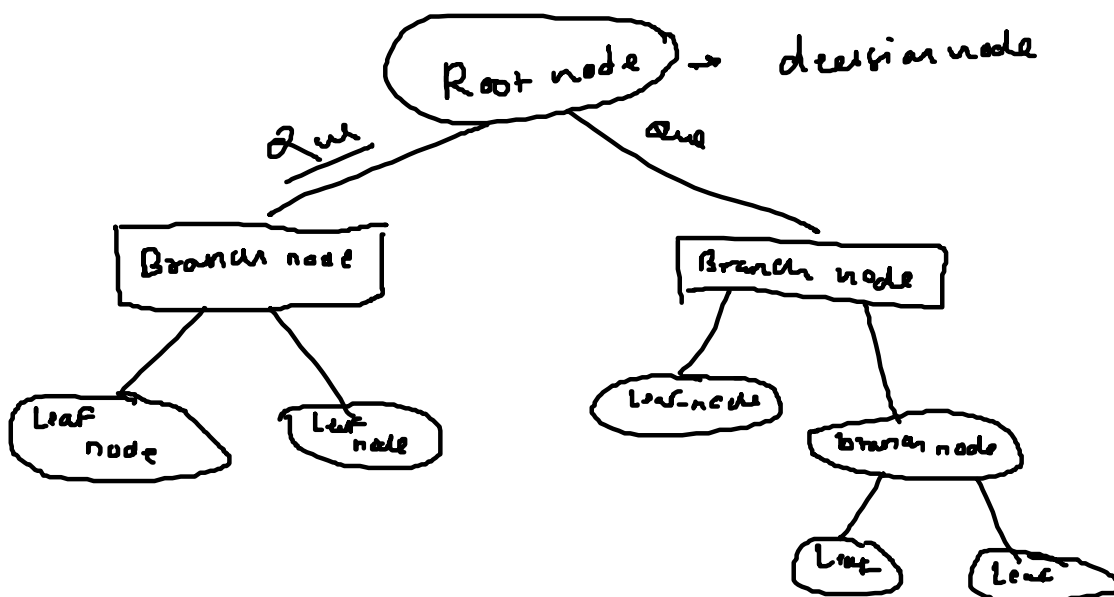
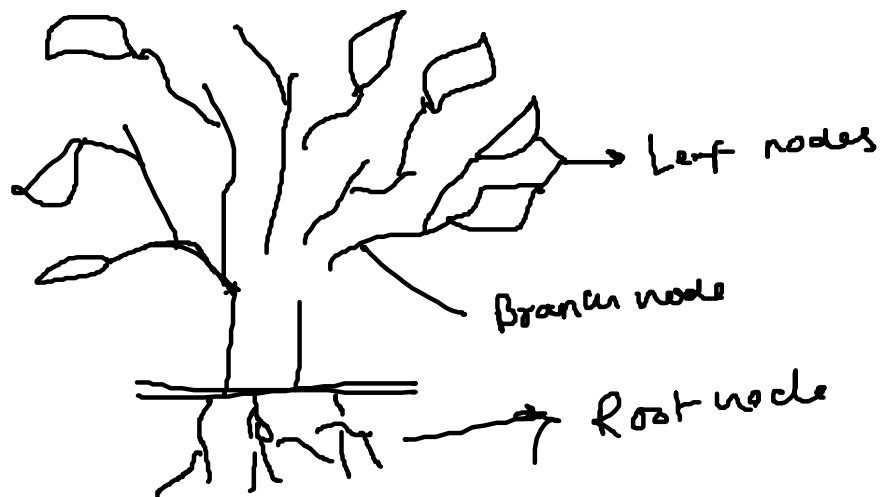
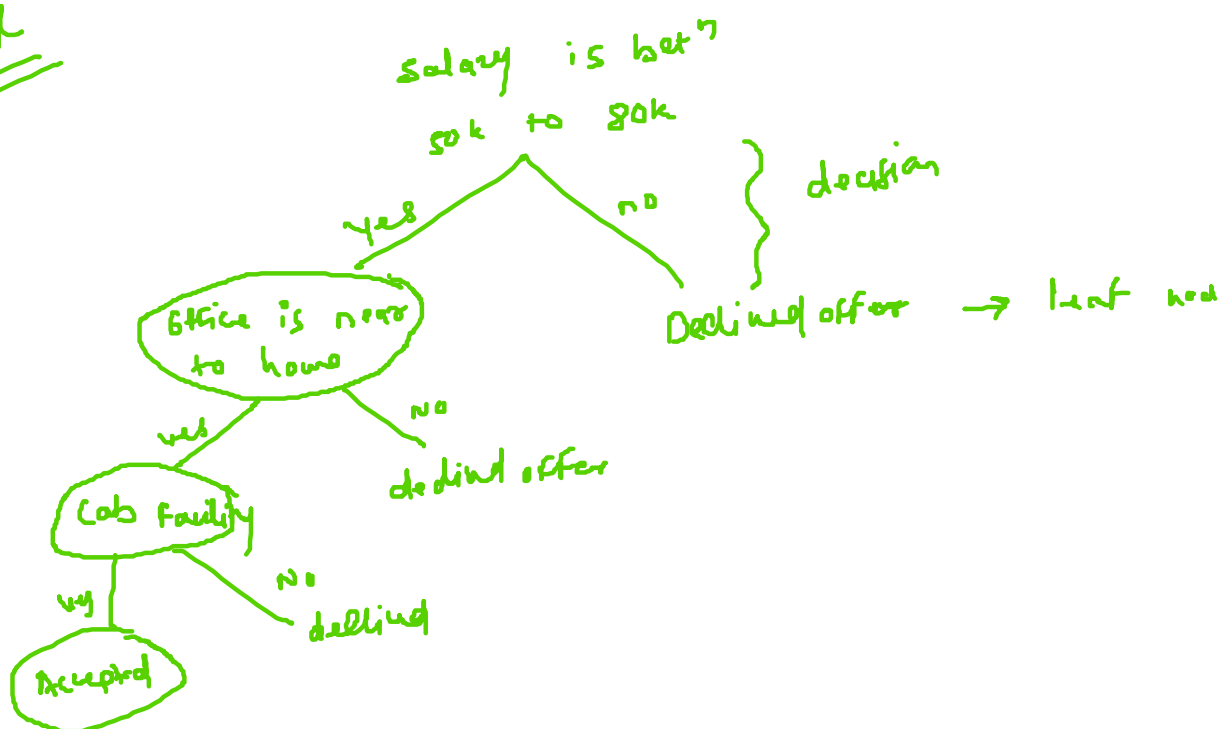


Decision Tree Day 1

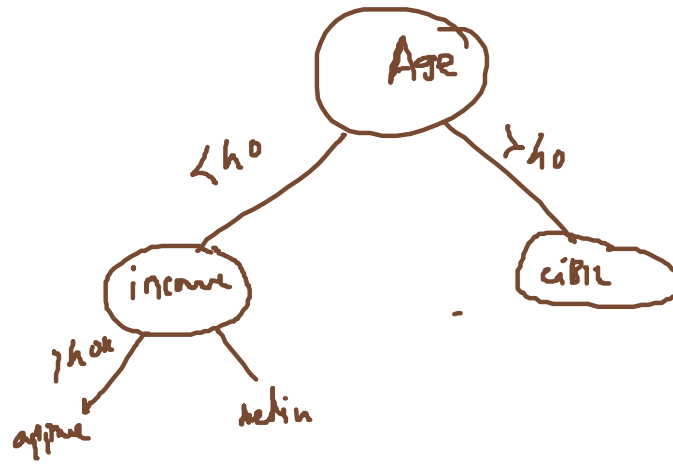
10 March 2022 07:11



ex



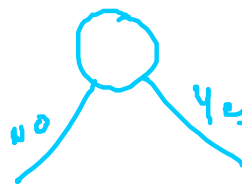
OK



OK

color	no	label
green	3	Mango
yellow	3	Mango
red	1	Grape
red	1	Grape
yellow	2	Lemon

R = 1 = Grape
R = 1 = Grape



no ≥ 2

3 = Mango
 1 = 3 = Mango
 4 = 2 = Lemon

Gini impurity

↓
Information Gain

for 1 IG > > entropy

case ①

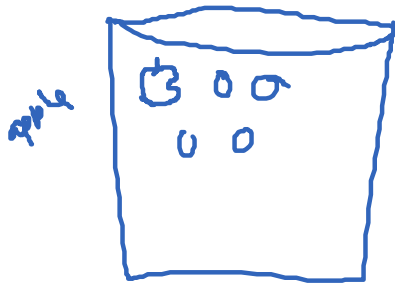
Apple



label

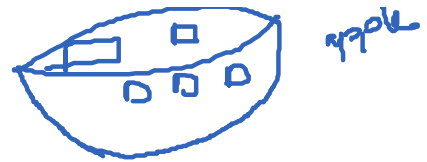


apple

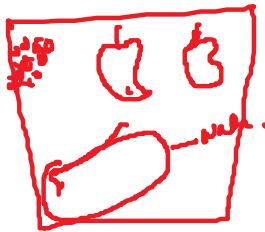


$$p = 1 \Rightarrow \text{highest}$$

$$\text{impurity} = \text{low} = 0$$



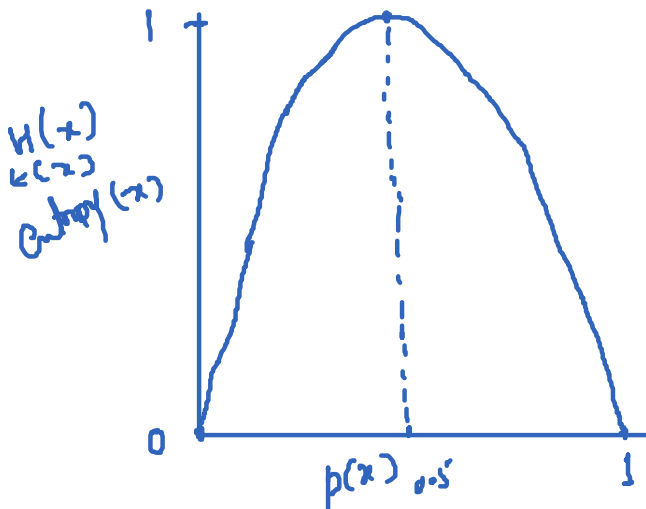
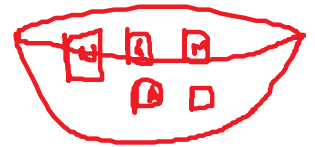
case 2



$$p = \text{low}$$

$$\neq 1$$

impurity $\neq 0$



$$\text{Entropy}(s) = -p(y) \log_2 p(y) - p(no) \log_2 p(N)$$

① IF no. of y = no. of no $\Rightarrow p(s) = \underline{\underline{0.5}}$

$\Rightarrow \text{entropy} = 1$

② all yes $\underline{\underline{or}}$ all no $\Rightarrow p(s) = \underline{\underline{1 or 0}}$

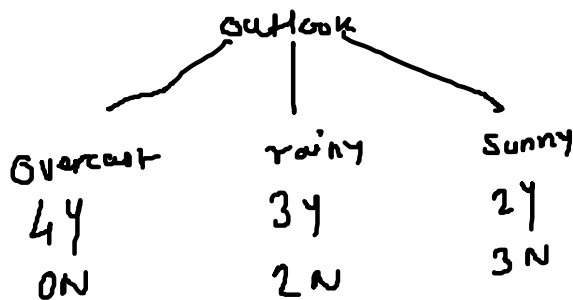
$$\rightarrow \text{entropy} = 0$$

$$H(S) : E(S) = - p(Y) \log_2 p(Y) \\ = - p(Y) \log_2 p(Y) - p(N) \log_2 p(N)$$

$$p(N) = p(Y) = 0.5$$

$E(S) =$

Information Gain



$$E(\text{outlook} = \text{overcast}) = - \frac{4}{4} \log_2 \frac{4}{4} - 0 \log_2 0 \\ = 0$$

$$E(\text{ov} = \text{rainy}) = - \frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} \\ = 0.971$$

$$E(\text{ov} = \text{sunny}) = - \frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \\ = 0.971$$



Information: \rightarrow

$$\text{wt ov} = \frac{4}{14} \times 0$$

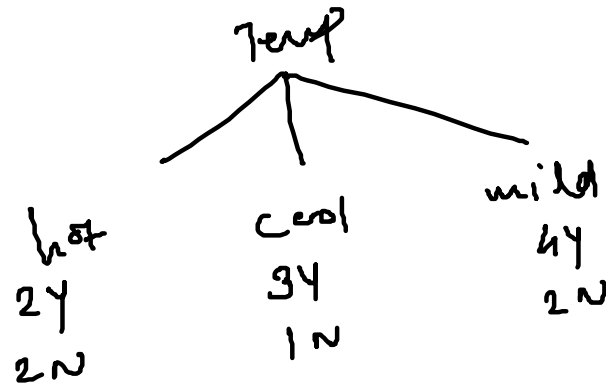
$$I(\text{outlook}) = \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 + \frac{5}{14} \times 0.971$$

$$= 0.693$$

$$I_{\text{outlook}} = 0.94 - 0.693$$

$$= 0.247$$

Temp



$$E(\text{hot}) = -\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4}$$

$$= 1$$

$$E(\text{cool}) = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4}$$

$$= 0.81$$

$$E(\text{mild}) = 0.91$$

$$I(\text{temp}) = 0.91$$

$$I_{\text{temp}} = 0.03$$

Humidity



hii

high
3Y
4N

normal
5Y
1N

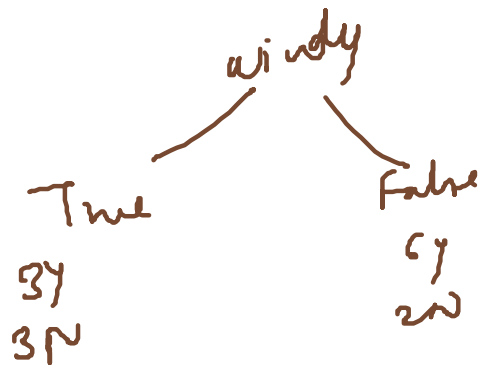
$$E(\text{high}) = 0.18$$

$$E(\text{norm}) = 0.59$$

$$I(\text{hum}) = 0.78$$

$$I_h(\text{hum}) = \underline{\underline{0.16}}$$

Windy

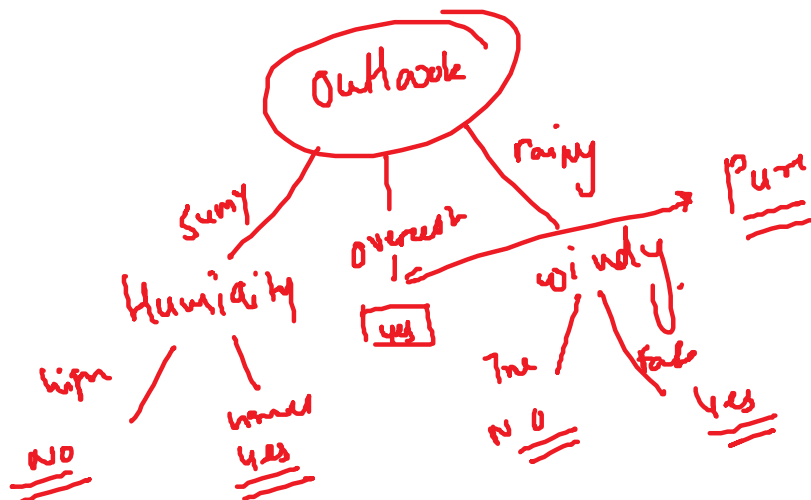


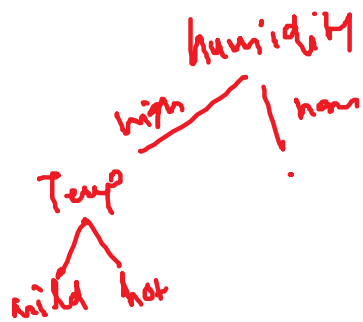
$$E(T) = 1$$

$$E(F) = 0.81$$

$$I(\text{windy}) = 0.89$$

$$I_h(\text{windy}) = 0.05$$





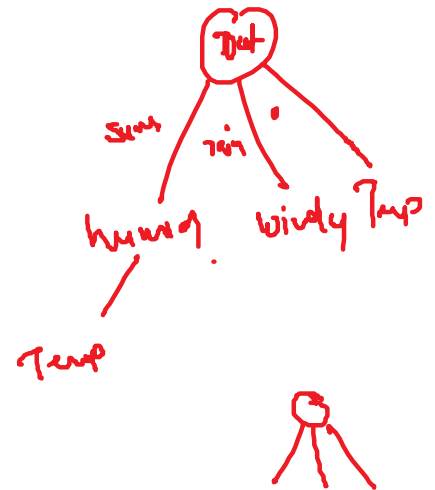
$IG_{out} : 0.24 \rightarrow$

$IG_{Hum} = 0.16 \rightarrow$

$24 - windy = 0.05$

$24 - Temp = 0.03$

Root node



Gini index

$$G.I = 1 - E(P)^2$$

$$= 1 - [P(Y)^2 + P(N)^2]$$



case ① All yes or All no $= P = 1$

$$G.I = 1 - (1)^2$$

$$= 0$$

case ② equal yes or equal no $= P = 0.5 \text{ and } 0.5$

$$G.I = 1 - [(0.5)^2 + (0.5)^2]$$

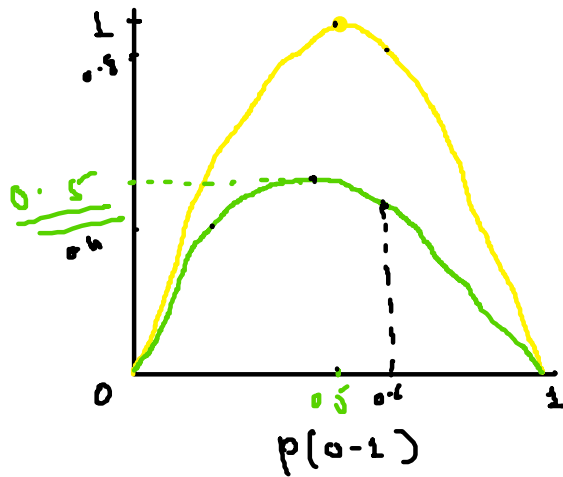
$$= 1 - [0.25 + 0.25]$$

$$= 1 - 0.5$$

$$\boxed{G.I = 0.5}$$

highest value of G.I

hi



Entropy = yellow
Gini = green

$$GI = 1 - [pY^2 + pN0^2]$$

$$= 1 - [1^2 + 0^2]$$

All Yes

$$= 1 - 1$$

$$= 0$$

$$GI = 1 - [0^2 + 1^2]$$

All No

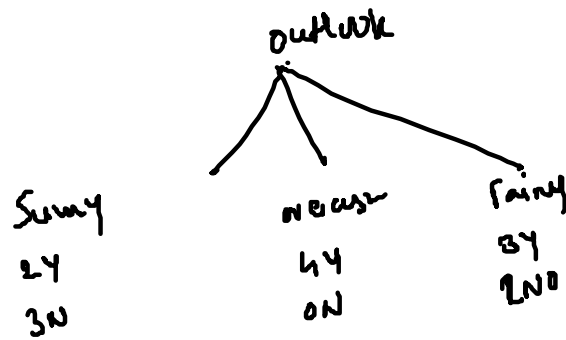
$$= 1 - 1$$

$$= 0$$

$$1) = Gini < Entropy$$

2) = Gini doesn't use log
it is faster in terms of entropy

A Q.L.



$$Q(s) = 1 - \left[\left(\frac{2}{5} \right)^2 + \left(\frac{3}{5} \right)^2 \right]$$

$$= 0.48$$

$$Q(0) = 1 - [1^2 + 0^2]$$

$$= 0$$

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

$$= 0.48$$

$$Q_{info} = \sum \frac{D_i}{D} \times Q_{info}(D_i)$$

$$= \frac{5}{14} \times 0.48 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.48$$

$$=$$

9 laes

75 k

All glar are fixed / no variable

Basics

= 4.5

50 %

HRA

= 2.5

20 % 25 %

—

TA
new.
Junt
Special -
PF → employee
P → employee

	700000	750000
ESLC	-	100
PF	=	1800
		1800
		<u>71200</u>

Variables: Joining =

Relocation =

18k

Special =

Quar → 3 4 5 6 = 18
annat =

regime

new → comp in her less

chat - → old tax → in her salary new

3 to 5 → 10 salary slip
30 29 7000

1

Emp. VAN

pm -

hi

Goal

impurity

purity

reduce

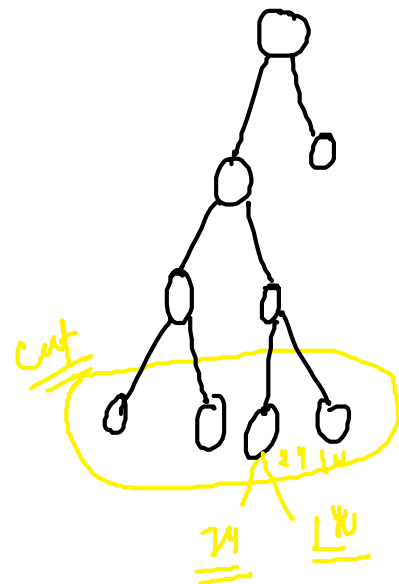
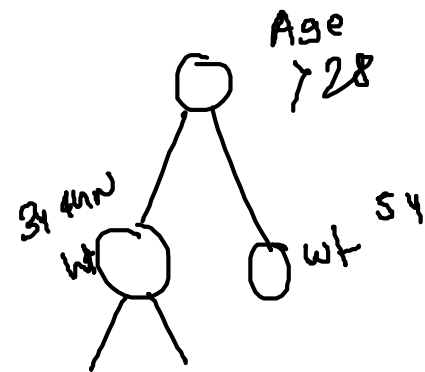
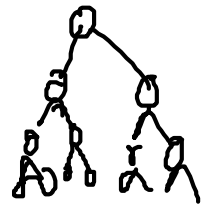
enhance

overfitD.T. is most overfitted
model# Pruning

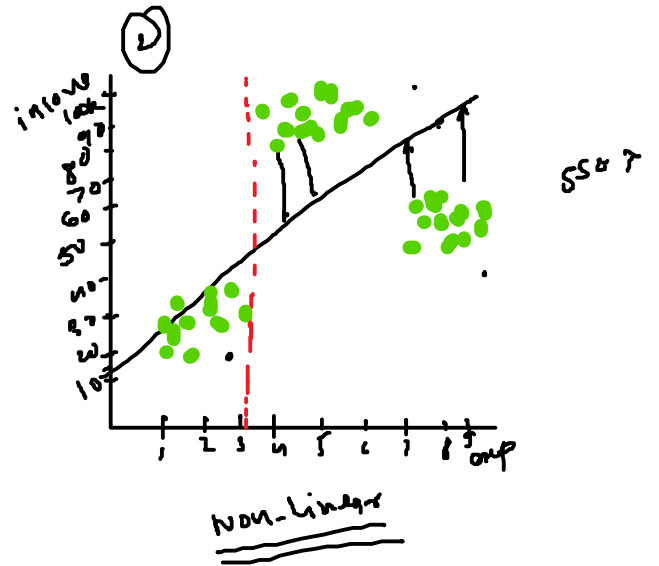
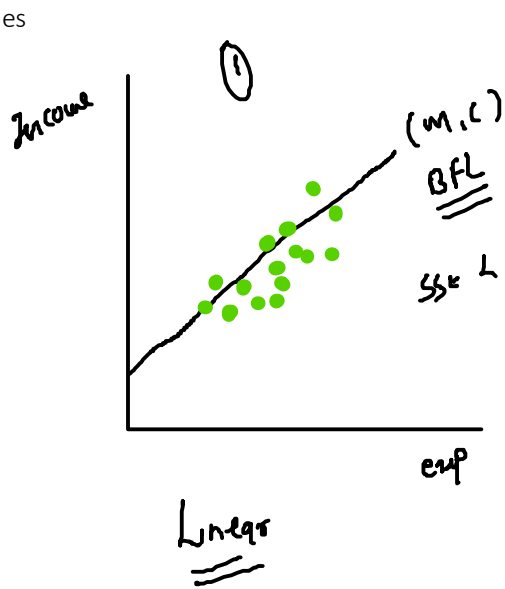
cut down the nodes

1) reduce complexity

2) Pre-Pruning = max depth = 3

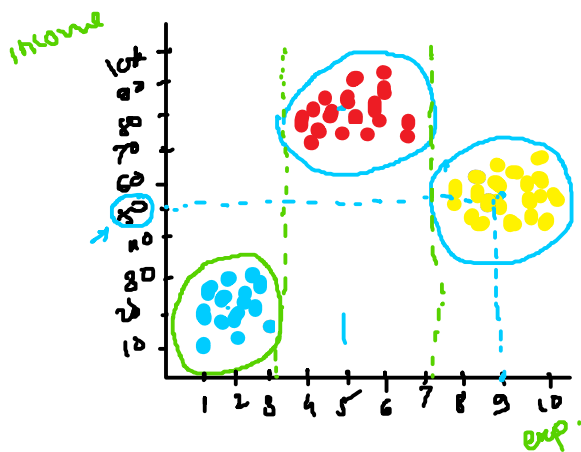
CCP \rightarrow cost complexity pruning
post pruning

hi

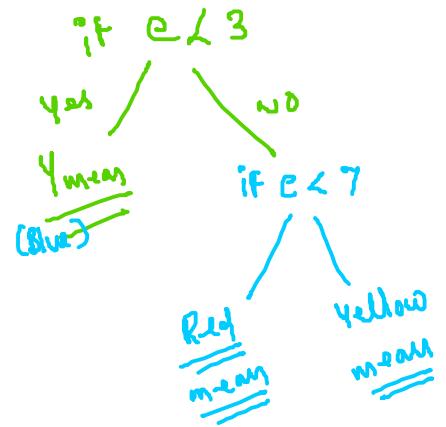


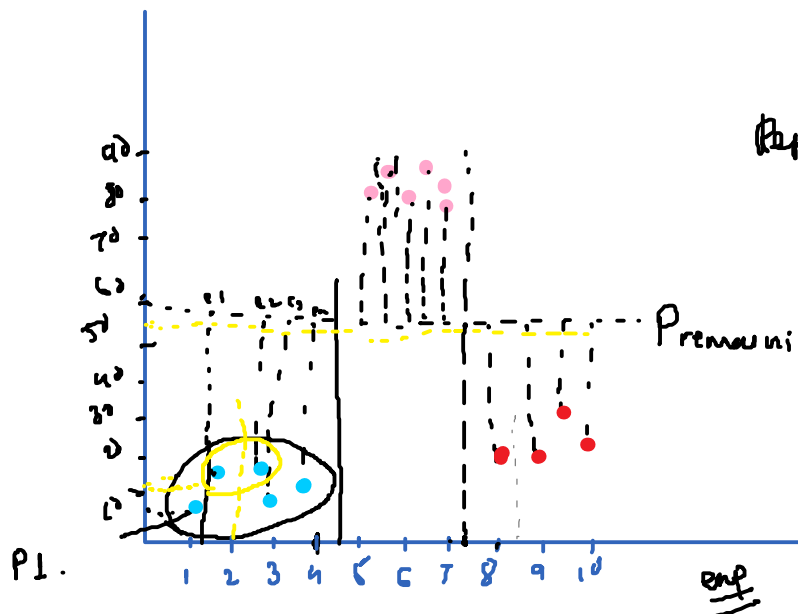
if $e < 3$

left comp. right comp.



msc
mpe

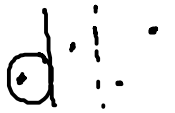
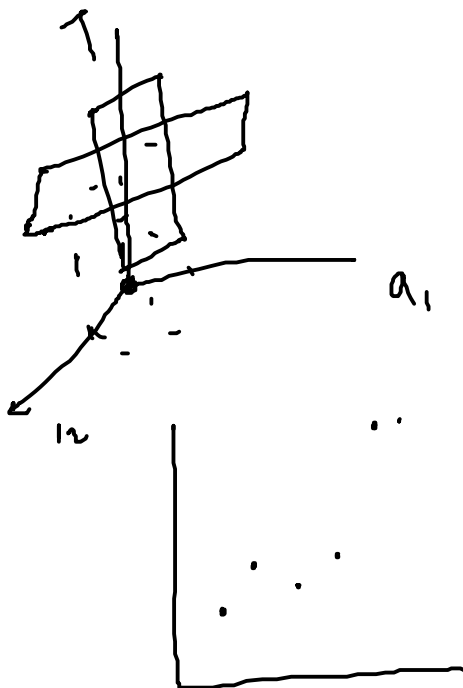
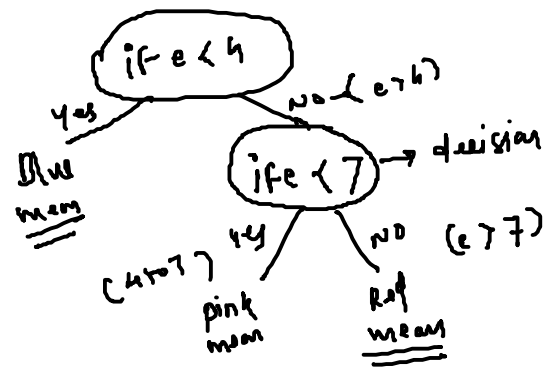
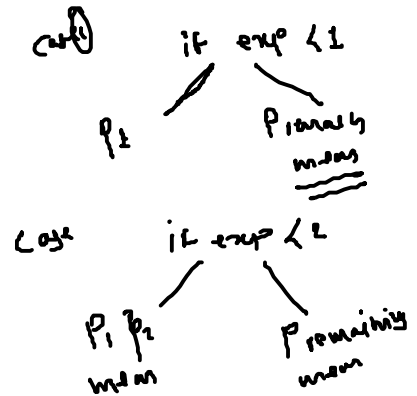
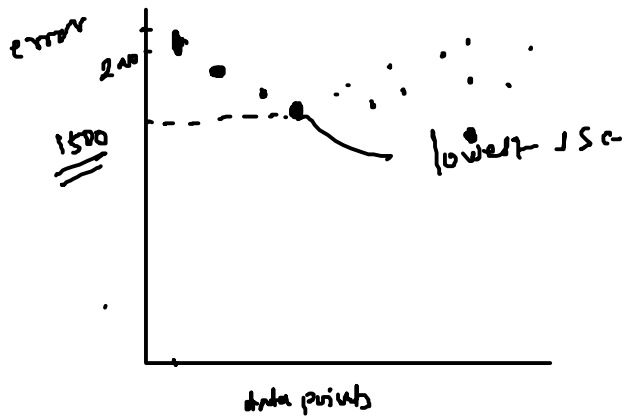


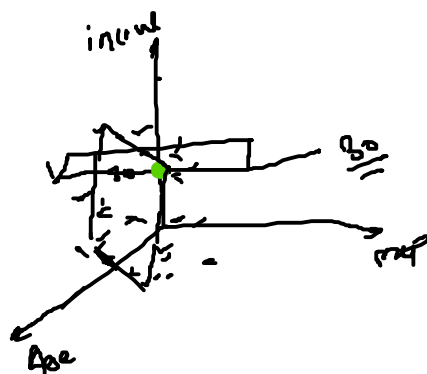
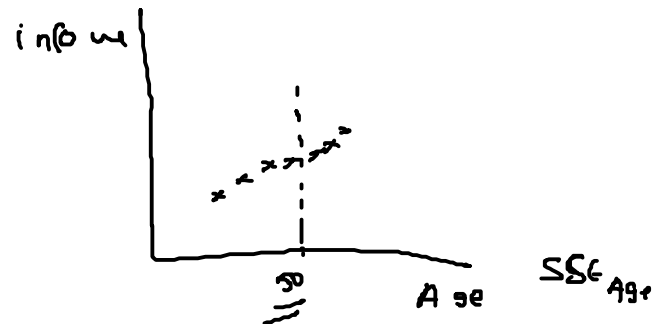
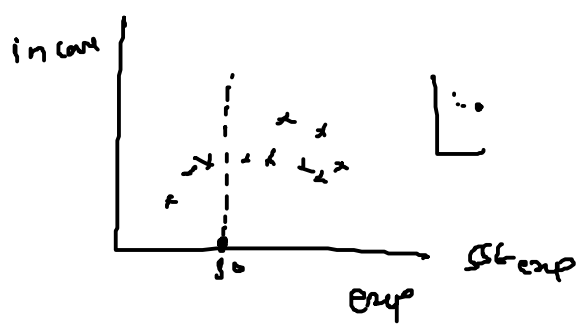
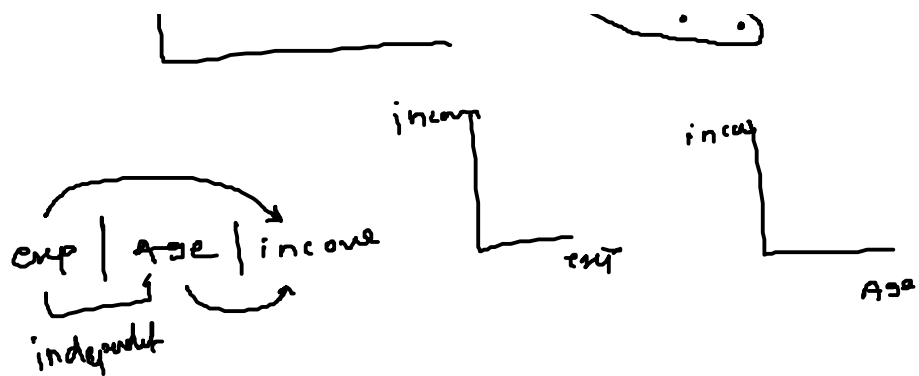


Residual / error = $\frac{MAE}{\sqrt{SSC}}$

$$e_1 + e_2 + e_3 + \dots + e_{10} = 200$$

$SSC = 1800$





if $exp \neq 30$

41 (3)

13 (13)

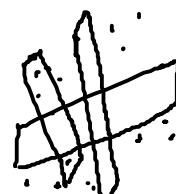
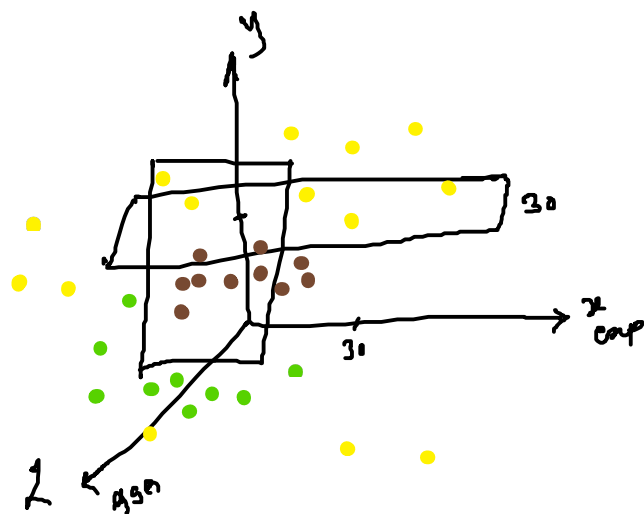
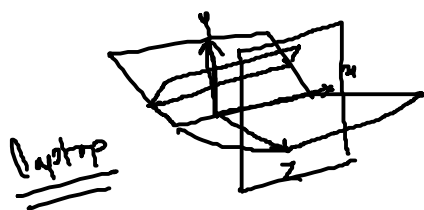
$SS_{exp} = 1500$

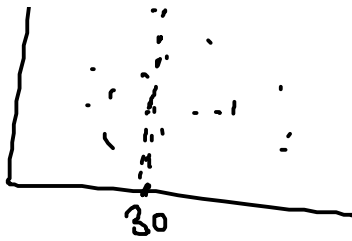
$SS_{Age} = 1800$

if $exp < 30$

2 data point

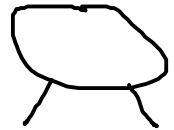
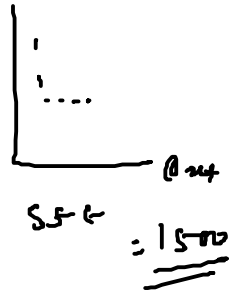
if > 30





T e l e n

ind. analysis



act pred = err

Ensemble technique.
 1) Bagging → RF
 2) Boosting → Ada.B.

Resume :-

Cognizant
infos
Calsoft

Corp. Idg →
Bitwise

Amol D.
Data Scientist
Capital Invest. Pvt. Ltd.

3.2 yrs in Capital as DS.

capita → what type
→ address
→ Building
→ Floor → 2nd floor
→ Cabinet
→ Commute (transport)
→ own vehicle

Cognizant → visit portal

Where ever → Capita, Pune.
desire → DS, ML, P.D, DE
current CTC → cost to company
→ 6 L LPA

Money list :- for may 0

profession :- every possible way

Varities :- No / any find

In line : 61 → 50983
47533
68333
47033

warred, anand@danishook . cu. c.
a. i. . a. i.

Chet. bngal@vite me. co. in
hr @ vite me. co. in

Arased 4th

2nd

3rd

→ HR :-

capita but head and 042 client
fixed - Any shubis.

D.S Project :-

1) Manager →

2) Project manager →

3) Team lead / consultant / Associate

swati bhar

shena bhar

swati Jadh

27 praj - ...
 2) Teru kod / consul / Associat swast Jadh
 4) BA .
 3) Jr. ml gr } 32
 6) D.S. for }
 7) St. denier }

Sole. print :- use decision.