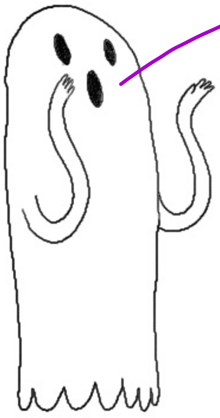
Decision tree 2



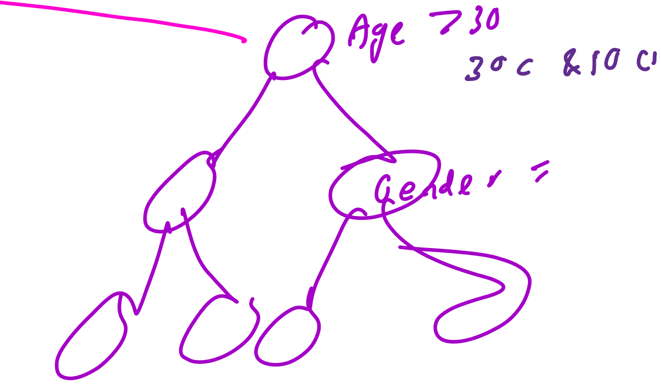
9:05 pao milte hai

DT

Entropy

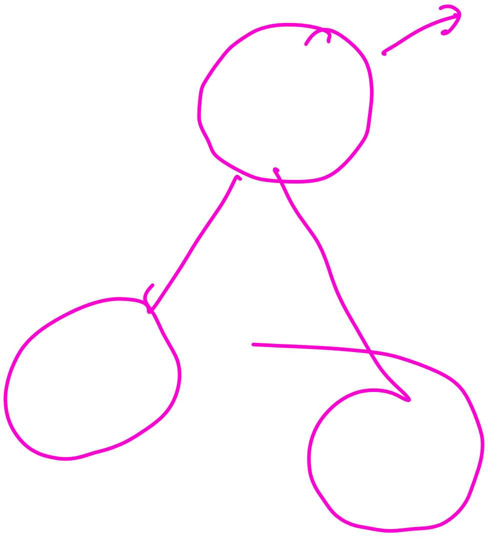
⇓

Impurity



$$H(y) = -p\log_2 p - (1-p)\log_2(1-p)$$

entropy(y):



y.value_counts()

⇓

# C \# NC

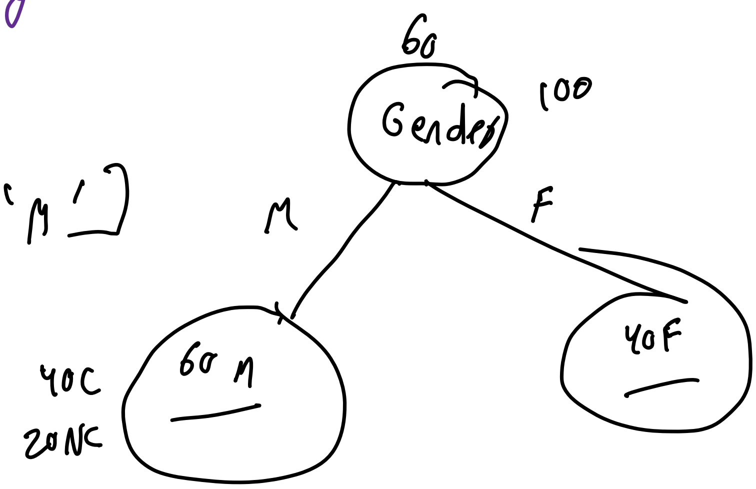
$$-\overset{\nearrow}{\frac{3}{5}}\log\left(\frac{3}{5}\right)-\frac{2}{5}\log\left(\frac{2}{5}\right)$$

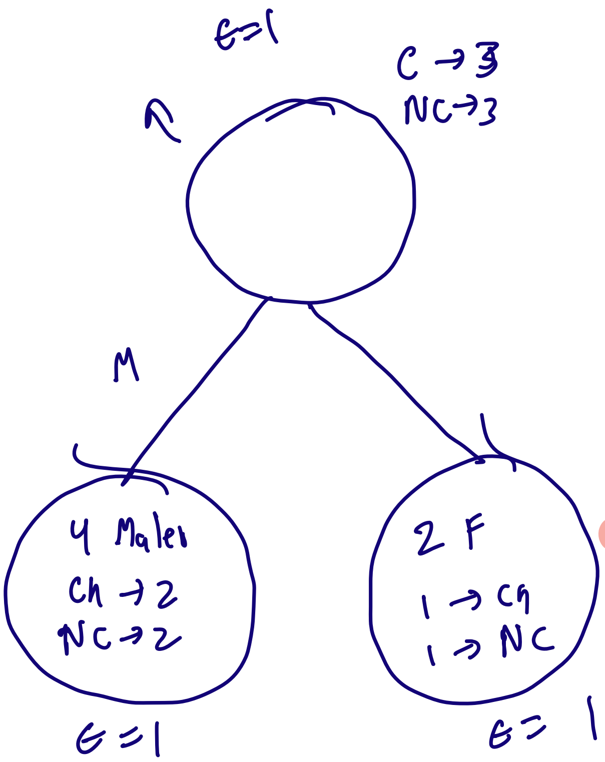
y.shape[0]

Age	exp	Churn
20	10	Y
30	20	N
40	30	Y
60	80	N
40	20	Y

entropy(y):

y[feature == 'M']





Gender	Age	Attrition
M	20	Yes
F	30	No
M	40	Yes
F	50	Yes
M	60	No
M	80	No

Gini Index

$$GI(y) = 1 - \sum_{i=1}^{k} (p(y_i))^2$$

$$-p \log p$$
$$\Downarrow$$
$$p^2$$

2 class :

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

Case 1 :

$p(+) = 0.5$
$p(-) = 0.5$ $\Rightarrow$ Entropy $= 1$

$$= 1 - \left[(0.5)^2 + (0.5)^2 \right)$$

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

$$= 0.5$$

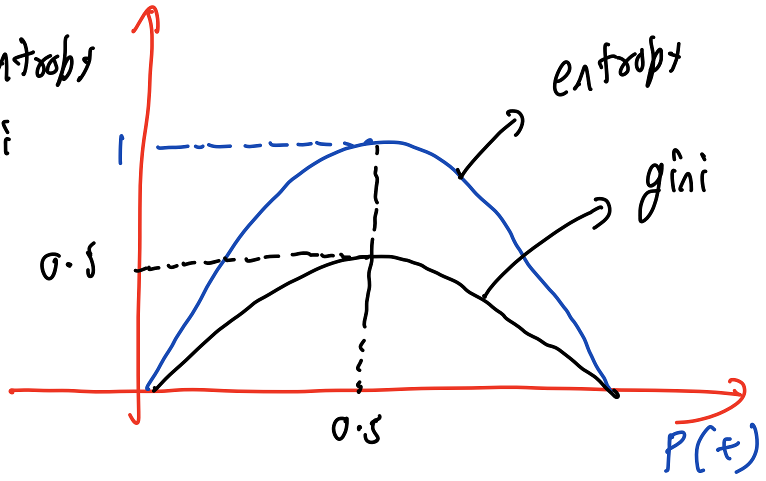
Case 2 :

$P(+) = 1$
$P(-) = 0$ $\Rightarrow$ $E = 0$

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

$$= 0$$

$[0 - 1]$ → Entropy
$[0 - \underline{0.5}]$ → Gini



Break 10:11

Split Numerical Column

20

Age < 35

38

40

f_1	f_2	y
c_1	2.2	1
c_2	2.6	1
c_1	3.5	0
c_2	3.8	0
c_3	4.6	1
c_1	5.3	0

① $f_2 \le 2.2$

IG_1

② $f_2 \le 2.6$

IG_2

③ $f_2 \le 3.5$

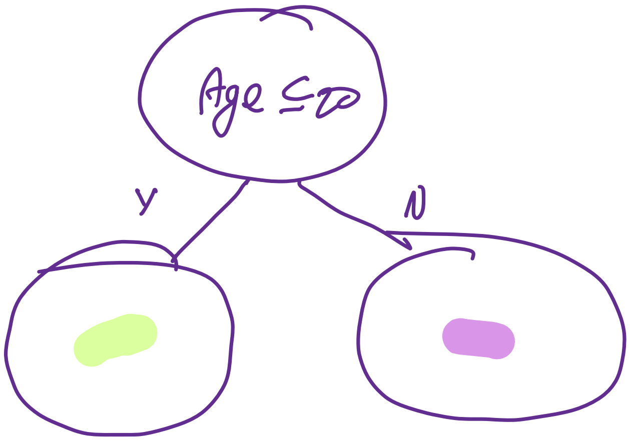
IG_3

$x \le Q_1$ $x \le Q_2$ $x \le Q_3$ $x \le Q_4$ ✓ ✓

IG_1

XGBoost

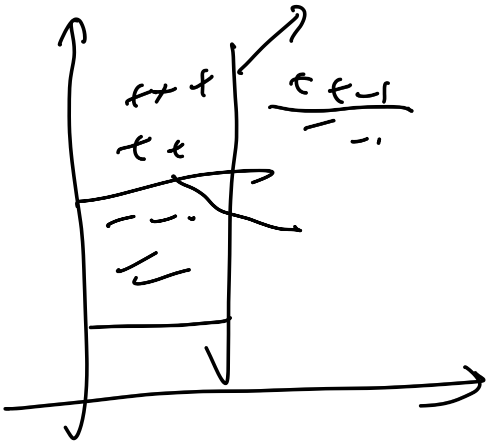
Age	Aff	
20	Y	🟩
18	N	🟩
20	Y	🟩
19	Y	🟩
21	N	🟪

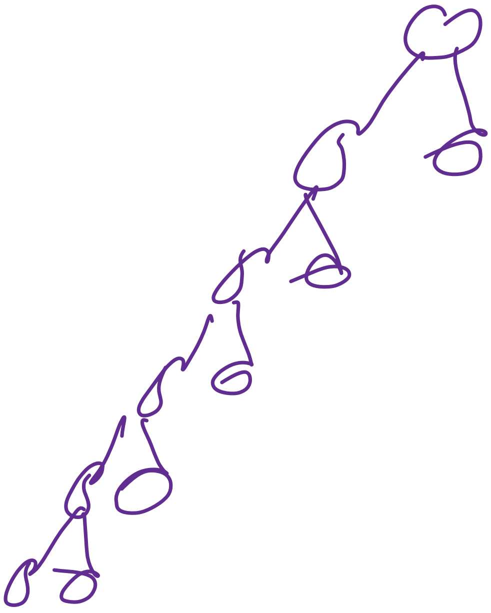


~~H~~ Huge depth

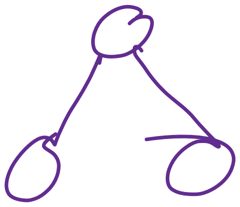
⇓

overfit

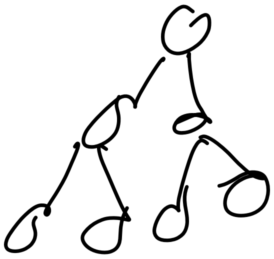




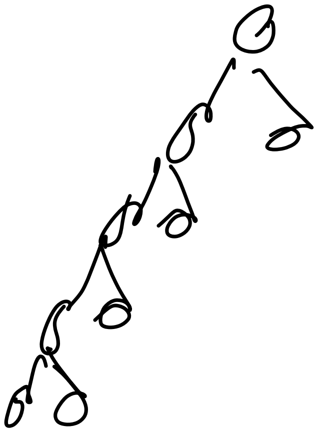
Decision stump →



shallow tree



Deep tree

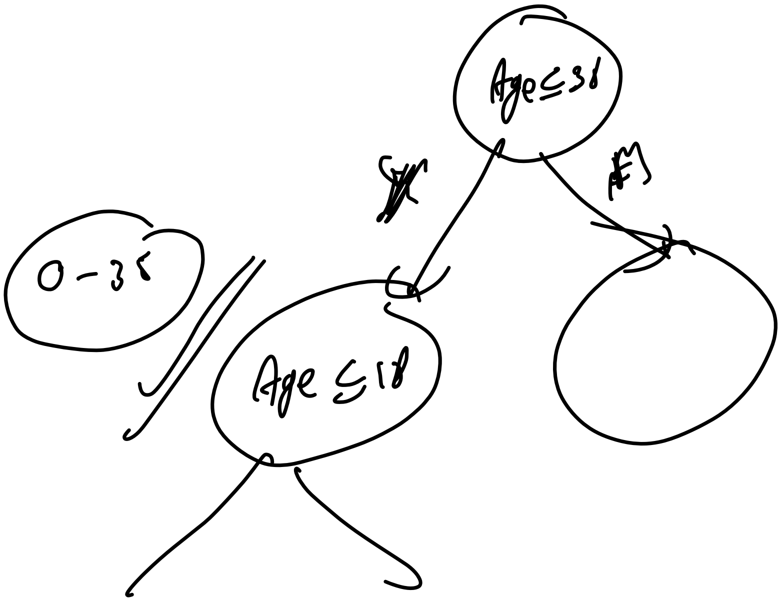


depth ⇒ 2 → cv _error

⇒ 3 →

⇒ 4 →

Train CV Test



2 minutes
remaining