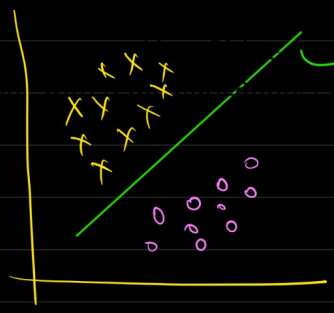
Agenda

→ Decision Tree

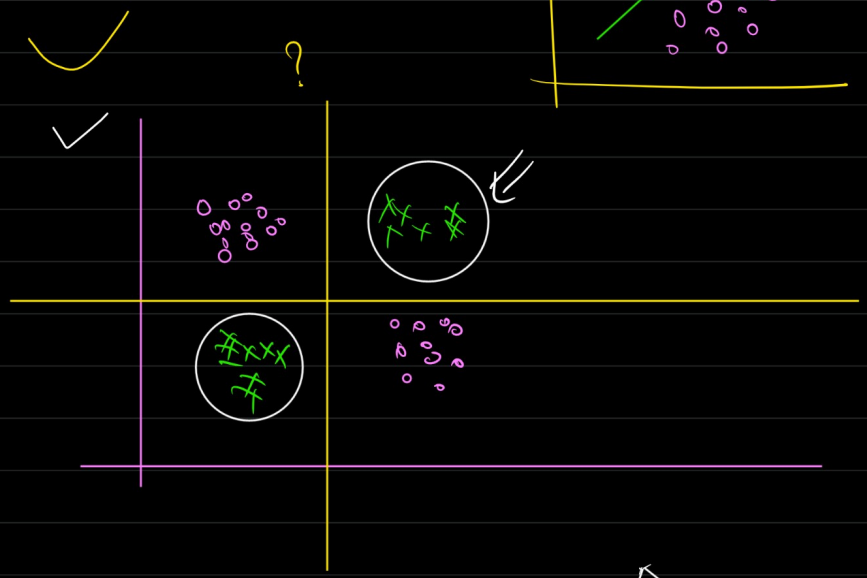
Logistic Regression → Linear Reg + Sigmoid function

⇓

→ Linear decision boundary



Problem wrt classification



?

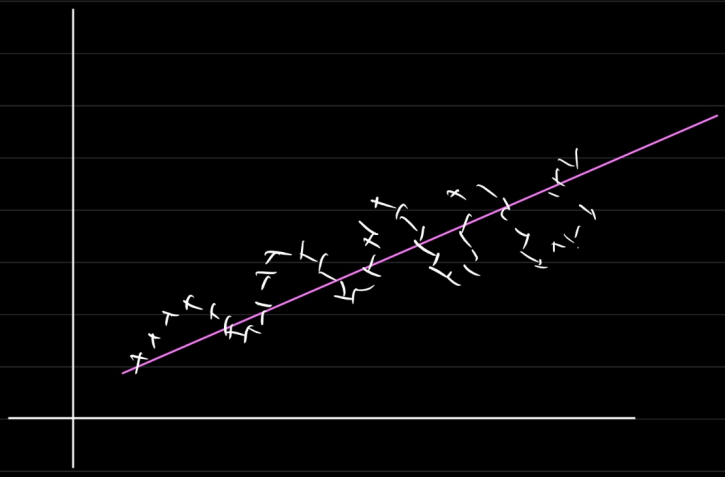
No

Why? Here data is non Linear

Decision Tree

SVM (support vector machine)

Problem for Regression



⇓

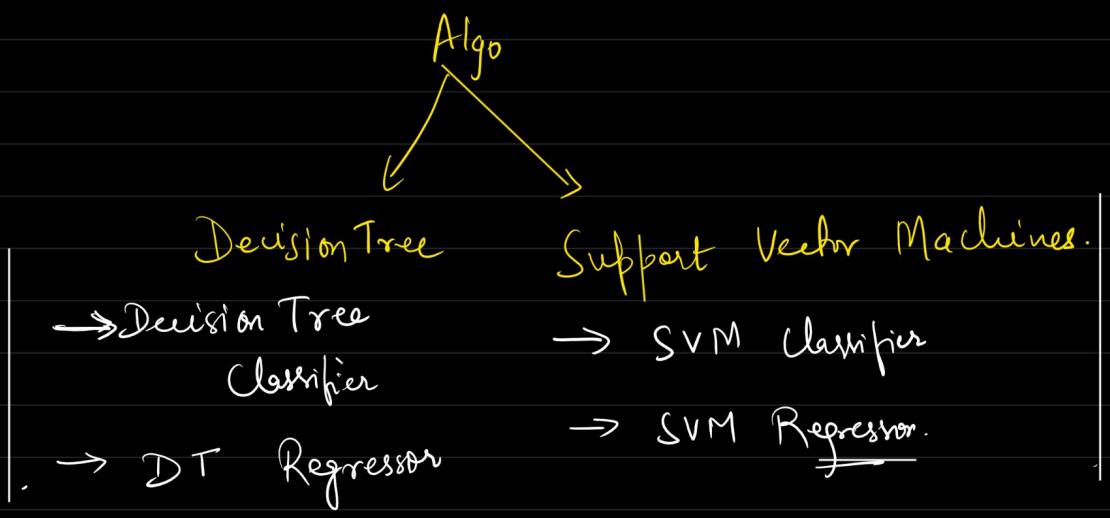
→ Polynomial regression

⇓

Preparing the data/ doing FE

⇓

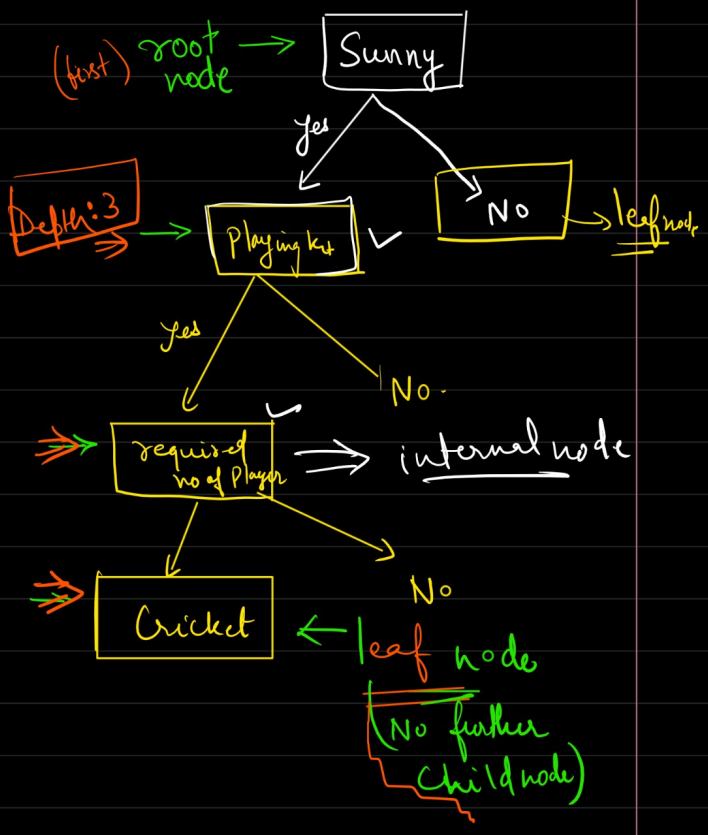
LR



* Decision Tree

→ DT cls
→ DT Reg.

Cricket

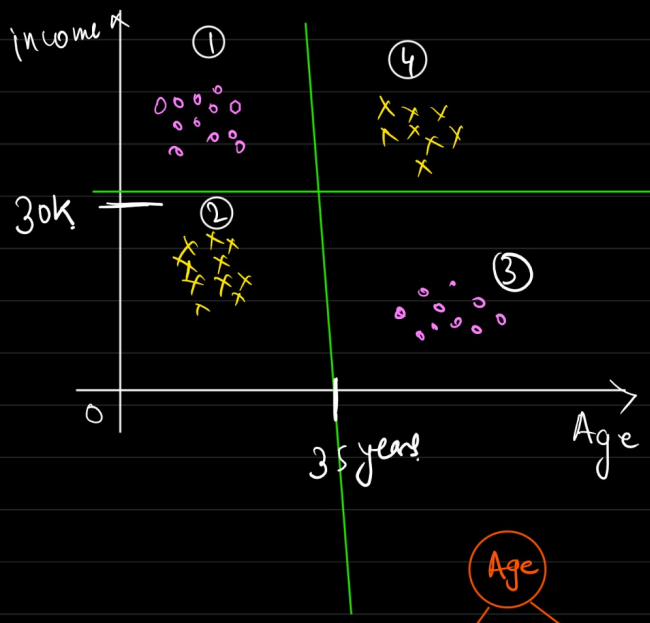


Tree

→ At every node, some decision is being made.

→ Decision Tree

Root node - first node of tree
leaf node - No child node
internal node - Node b/w internal & child
Depth → No of levels.

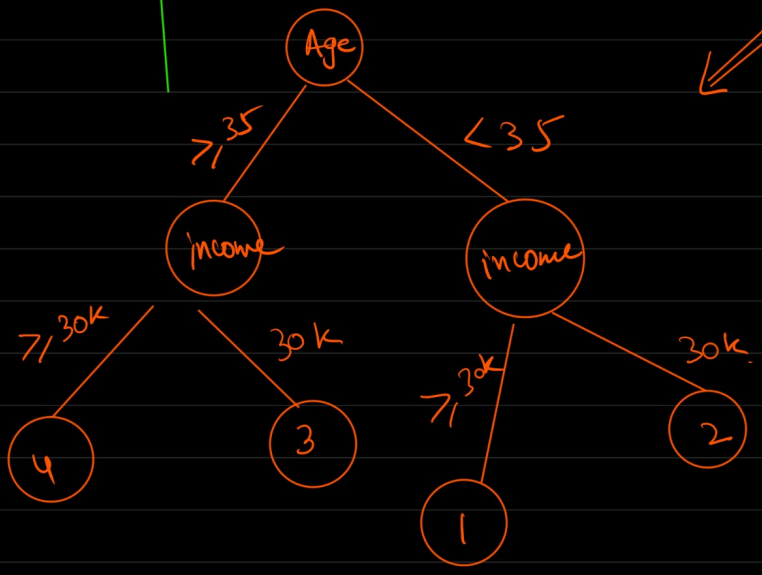


① income ≥ 30K & Age ≤ 35 years
↳ group-1 (0th class fraudster)

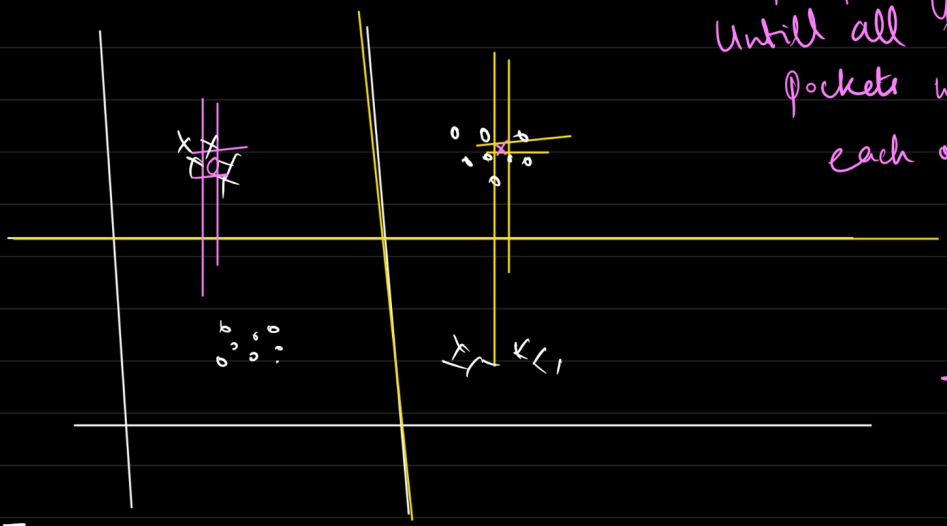
② income < 30K & Age ≤ 35 years
↳ group 2 → Xth class
↳ Non fraudster.

③ if income ≤ 30K and Age ≥ 35 years
↳ group 3
↳ 0th class fraudster

④ if income ≥ 30K & Age ≥ 35 yrs
↳ group-4
↳ Xth class
↳ fraud



⇒ DT divide the data into different Partition or into Pockets (nodes)



→ Keep splitting / dividing untill all the pockets where each of the dp is of same class.

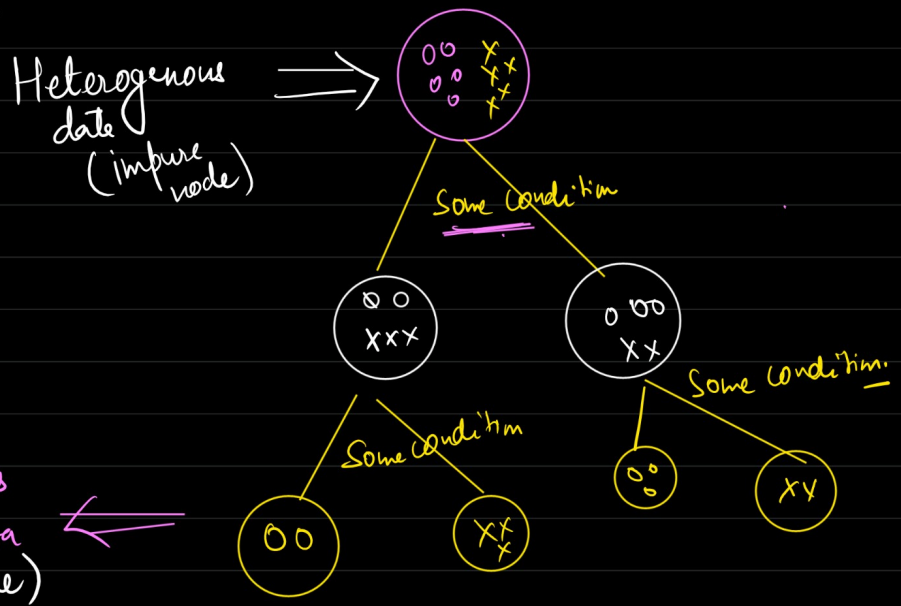
* Advantage of DT

→ Simple to Understand

→ Captures the non linear relationship

Intend :-

→ To create leaf node/pocket where each of the data points belong to one class.

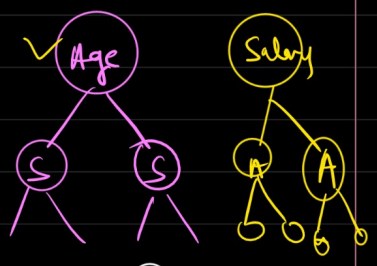


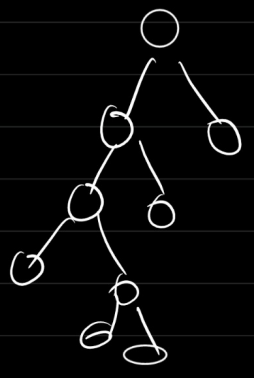
* Pure — homogenous (same class)
impure — heterogenous (mixture)

Q.1. How to split? → On what factor condition should be use?

Q2. Which feature to be used for splitting?

Q3. Till when to split?





* How to split

↳ Impurity measures

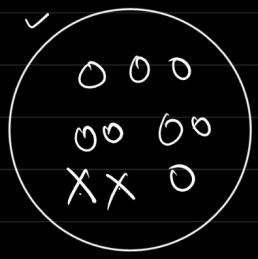
→ Classification Error

→ Gini

→ Entropy

Salary	Age	fraud
		O
		O
		O
		O
		X

① Classification error



O : 8
X : 2

$CE = 1 - \max p_i$

$p_x = \frac{2}{10} = 0.2$

$p_o = \frac{8}{10} = 0.8$

$CE = 1 - \max(0.8, 0.2)$
$= 1 - 0.8$
$= \underline{0.2}$

→ To make pure node, if I assign everything to majority class (o), what is the error rate?

↳ minority misclassified dP's.

* For multiclass

O − 4
X − 2
Δ − 4

$p_o = \frac{4}{10}$
$p_x = \frac{2}{10}$
$\Delta = \frac{4}{10}$

$1 - \max(0.4, 0.2, 0.4)$
$= \boxed{0.6}$

[1, 2, 3]
↑ ↑ r

② Gini

$G.I = 1 - \sum_{i=0}^{n-1} (p_i)^2$

i → from class 0 to n−1

$\Rightarrow 1 - \sum_{i=0}^{1} (p_i)^2 = 1 - [(p_o)^2 + (p_x)^2]$
$= 1 - [(0.8)^2 + (0.2)^2]$
$= 1 - (0.64 + 0.04)$
$= 1 - 0.68 = 0.32$

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

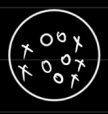
↳ for two class

$p_o(1 - p_o) + p_x(1 - p_x)$

$= 2p_o(1 - p_o)$

$p_o + p_x = 1$
$p_x = 1 - p_o$

$= 2 \times 0.8(1 - 0.8)$
$= 0.32$

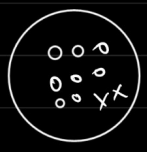


$\rightarrow$ Gini $\rightarrow$ $1 - \sum_{i=1}^{n} p_i$

$\rightarrow 1 - (p_o^2 + p_x^2)$

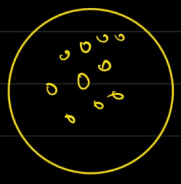
$= 1 - (0.5^2 + 0.5^2)$

$= 1 - 0.5 = 0.5$

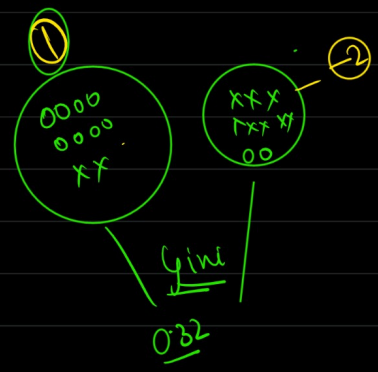


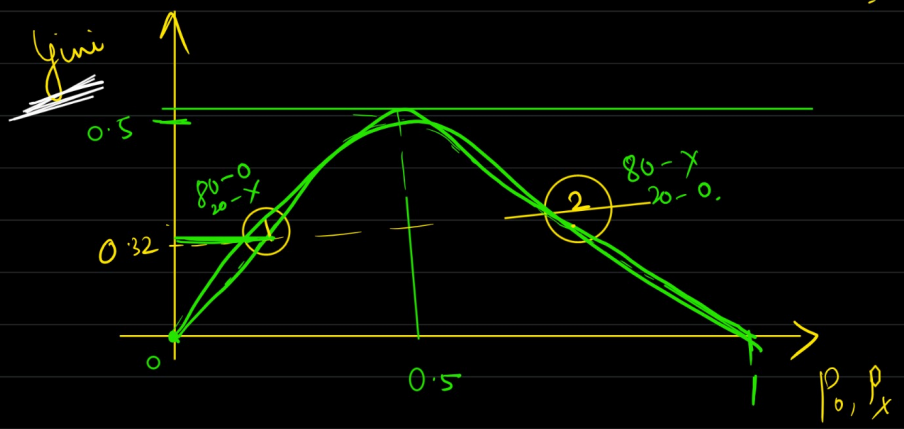
$\Rightarrow 0.32$

highest gini $\rightarrow$ 0.5

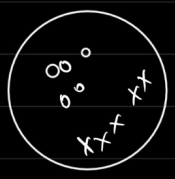


$\rightarrow$ Gini $\rightarrow$ 0 $\rightarrow$ all of are same $\rightarrow$ no impurity.





③ Entropy $= -\sum_{i=1}^{N} p_i \log_2 p_i$

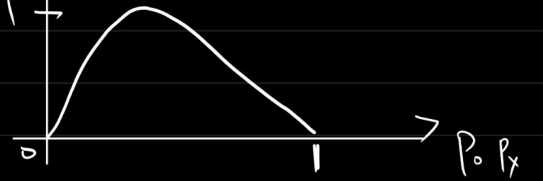


$= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$

a a

Highest entropy - 1

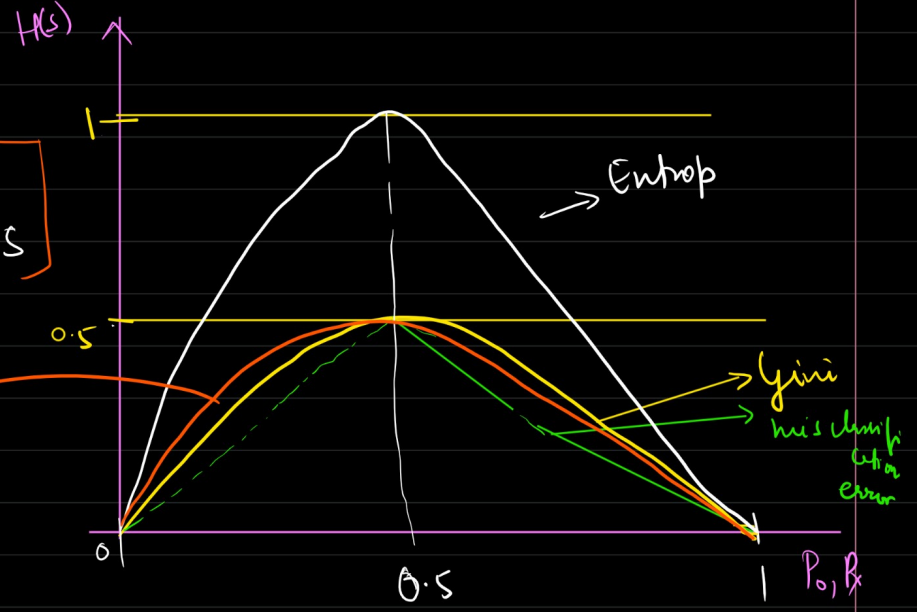
$= -2 \cdot \frac{1}{2}\log_2\frac{1}{2} = -\log_2\frac{1}{2} = -\log_2 2^{-1} = -1 \times -1 = 1$



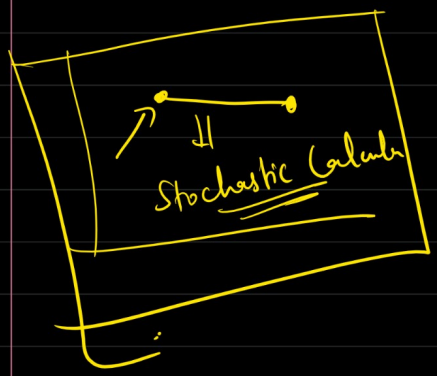
Entropy — from 0 to 1

Gini — 0 to 0.5

misclassification error = 0 - 0.5



Entropy Scaled = Entropy / 2



Gas-1

Brownian motion

always move → Entropy is more

Gas-2

always more → Entropy

Class 1	Class 2	Entropy
50%	50%	1
40%	60%	0.90
60%	40%	? 0.90
100%	0%	0
0%	100%	0

Multiclass classification

$$G.I = 1 - \sum_{i=1}^{n} p_i^2$$

$$= 1 - \left[p_1^2 + p_2^2 + p_3^2 \cdots p_n^2 \right]$$

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \log_2 p_i$$

$$- p_1 \log_2 p_1 - p_2 \log p_2 - p_3 \log p_3 \cdots - p_n \log p_n$$

* Which impurity measure to be used?

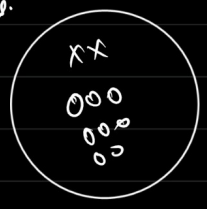
Gini Entropy

for small dataset → Entropy for large dataset → Gini (computation)

* On what features to split

↳ Information gain

disorganization/randomness →



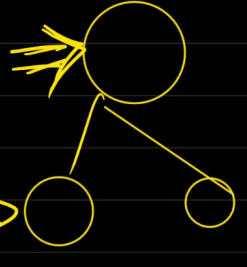
Aim of Splitting

we want pure nodes after splitting

↳ Such a split that maximise the purity.

* impurity (post split) < impurity (pre split)

↳ Split using that feature which does this best

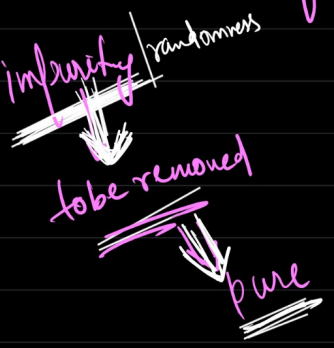


Post split →

impurity (post split) < impurity (pre split)

Δ impurity = impurity (pre split) − impurity (post split)

→ if randomness decreases, purity increase, then we say Information gain.



(randomness decreased) ⇒ Information Gain (Δ impurity)

* Information gain is a measure to define degree of disorganisation/randomness in a system ⇒ IG → GI, Entropy

IG ⇒ Δ impurity.

* Whichever feature will have maximum information, that feature will be used split.

* decrease in randomness ⇒ Information Gain.

Gender	class	Cricket (Yes/No)
M	IX	Yes
F	X	No
M	IX	Yes
...	...	...

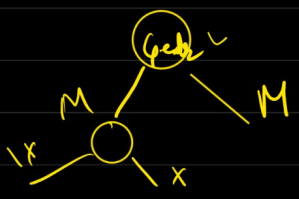
total Students : 30 → 50% play cricket Yes – 15, No – 15

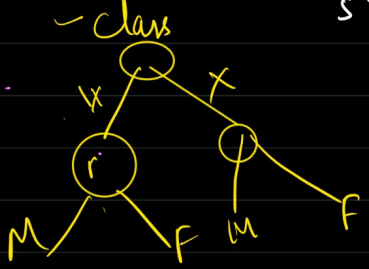
Gender — 10/2 — F – Play cricket (out of 10 female, 2 play cricket)

20/13 — M = Play cricket (out of 20 male, 13 play cricket)

Class — IX — 14/43% — out of 14 class IX students 43% play cricket

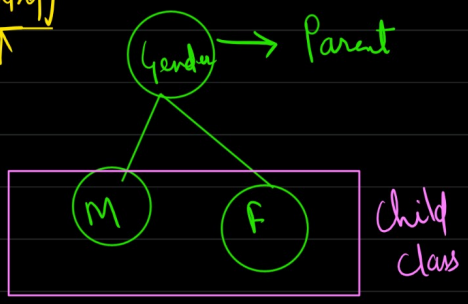
X — 16/57% — out of 16 class X student 57% play cricket





* Whichever feature has the highest information Gain, used for splitting.

Impurity → Gini, Entropy



IG = Imp. Presplit − Impurity Post split

= En (Parent) − En (child combined)

En Parent ⇒ $-\frac{15}{30}\log_2\frac{15}{30} - \frac{15}{30}\log_2\frac{15}{30} = 1$

En(child)

En(Femlchild) $= -\frac{2}{10}\log_2\frac{2}{10} - \frac{8}{10}\log\frac{8}{10}$

$= 0.72$

En(Male child) $= -\frac{13}{20}\log_2\frac{13}{20} - \frac{7}{20}\log_2\frac{7}{20}$

$= 0.92$

wt = Some total

En(Childcombined) = $\frac{\text{No of Females}}{\text{Total Students}} \times E_{\text{Femle child}} + \frac{\text{No of males}}{\text{Total Studes}} \times E_{\text{male child}}$

Weighted Entropy

$= \frac{10}{30} \times 0.72 + \frac{20}{30} \times 0.92 \Rightarrow 0.85.$

Class

En(Parent) $= -\frac{15}{30}\log_2\frac{15}{30} - \frac{15}{30}\log_2\frac{15}{30} = 1$

En child class

$En(IX) = -\frac{43}{100}\log_2\frac{43}{100} - \frac{57}{100}\log_2\frac{57}{100} = x_1$

$En(X) = -\frac{57}{100}\log_2\frac{57}{100} - \frac{43}{100}\log_2\frac{43}{100} = x_2$

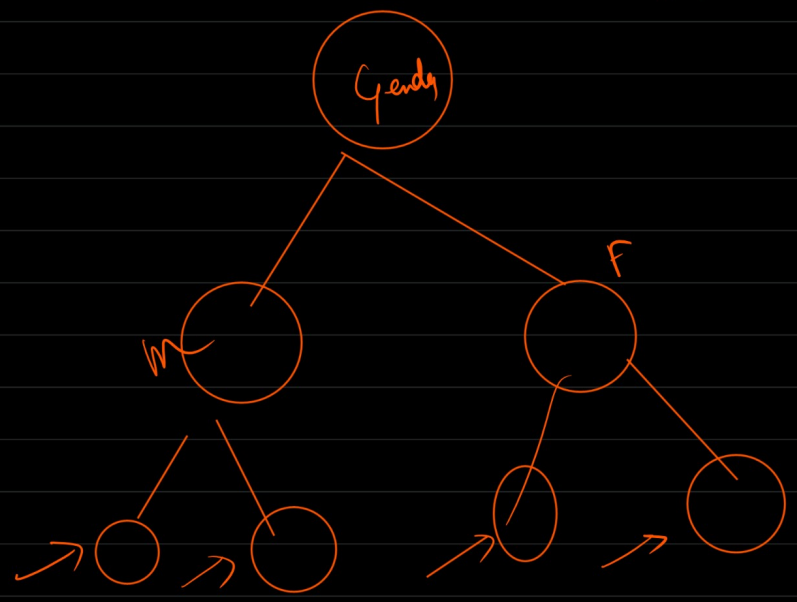
En Combined $= \frac{14}{30}x_1 + \frac{16}{30}x_2$

$= 0.99$

$IG = E_p - E_c = 1 - 0.99 = 0.01$

$$IG = E_p - E_c$$
$$= 1 - 0.85$$
$$= 0.15$$

Since $IG_{Gender} > IG_{Class}$
∴ We will split based on Gender



Algorithm of DT

→ Step-1 → Recursive binary splitting | Partitioning the data into smaller subset

→ Step-2 - Select the feature to split (Information gain)

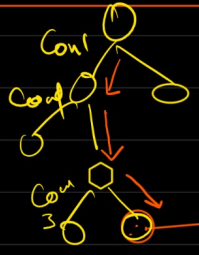
Step-3 Apply the split

Step-4 - Repeat the process for the subset obtained

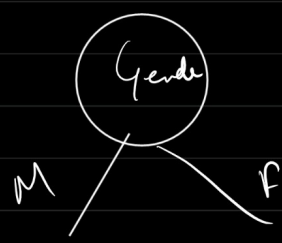
Step-5 → Continue the process untill every node is a pure node / Stopping criteria is reached.

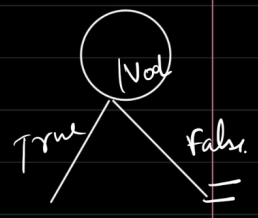
→ Step-6 Majority value in the leaf node will be prediction

Age-100, Salary 50k, Gender M



* For categorical feature

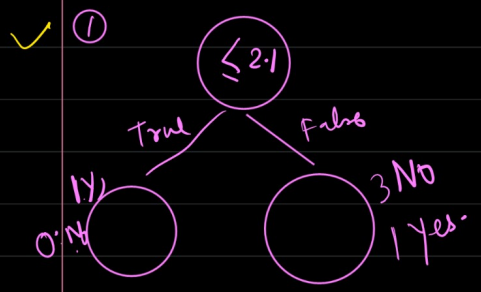


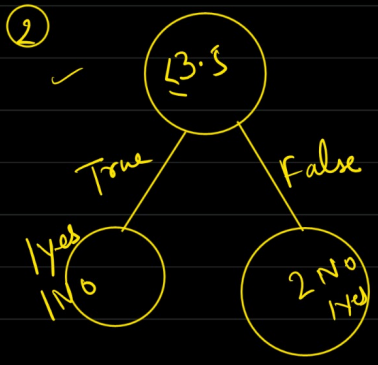


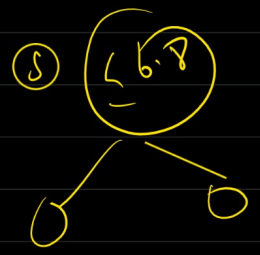
* What if feature is continuous

f_1	output
2.1	Yes
3.5	No
4.1	No
5.3	No
6.8	Yes

① Sort the feature (f_1)
② treat each value/row as a cutoff/threshold and try to build decision tree and whichever threshold gives you the best information gain that will be used for splitting.



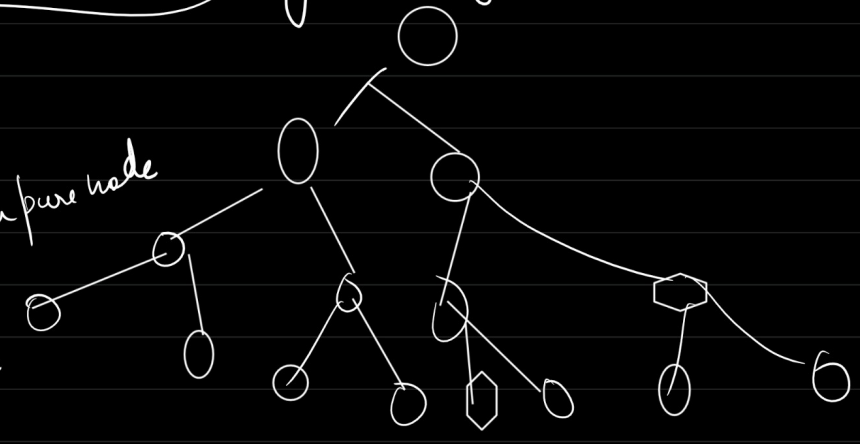


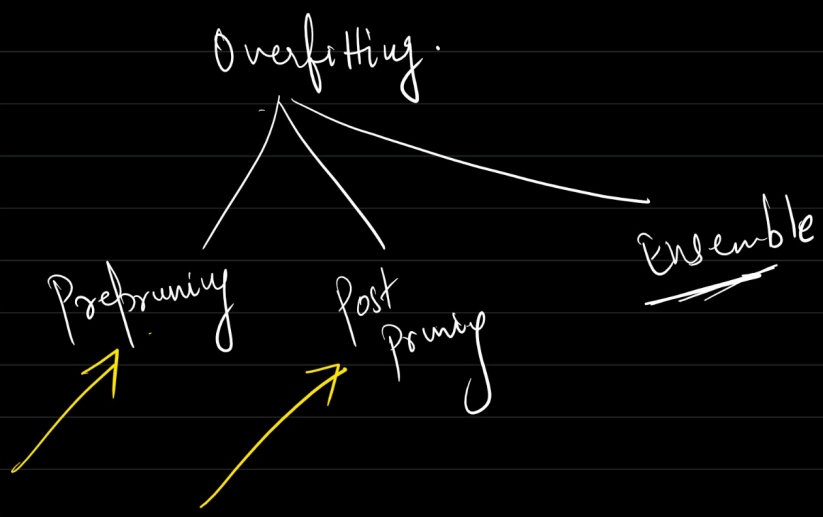


→ whichever split gives the highest IG that will be used for splitting. (feature/value)

* Decision tree is a greedy algorithm.

↳ It keeps on splitting until every leaf node is a pure node
↓
memorise the data
↓





Overfitting → training acc ↑

testing acc ↓