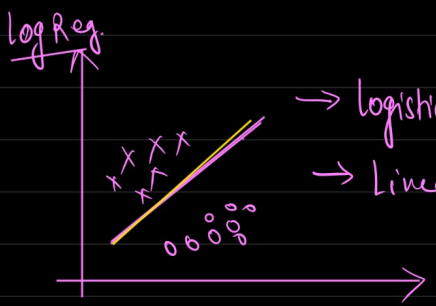
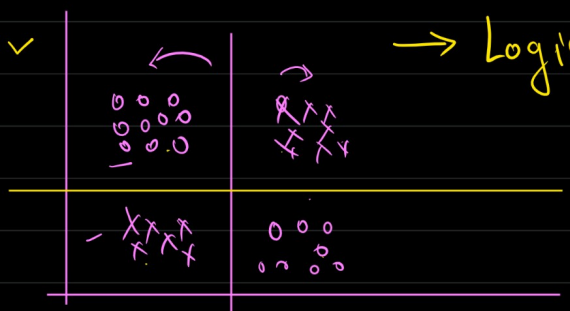
* Decision Tree



→ logistic regression works for linear decision boundary

→ Linear Regression captured linear relationship.

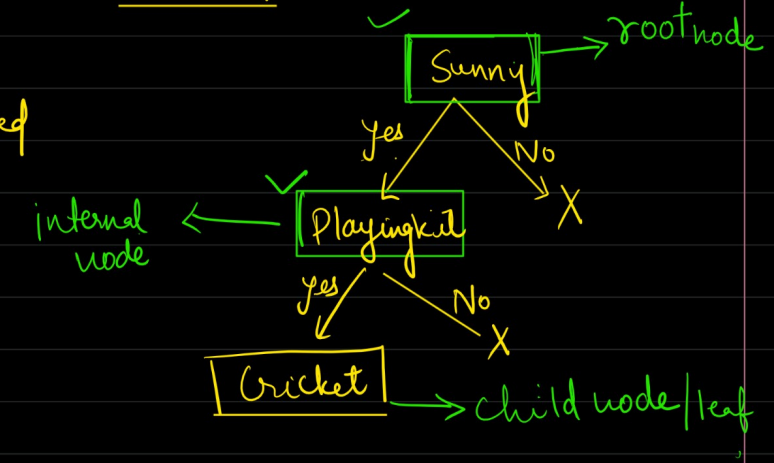


→ Logistic regression | Linear regression will not work with non linear data

* Decision tree → Classification problem (Decision tree classification)

→ Regression problem (Decision tree regressor)

Cricket



* Decision Tree works on nested if else condition.

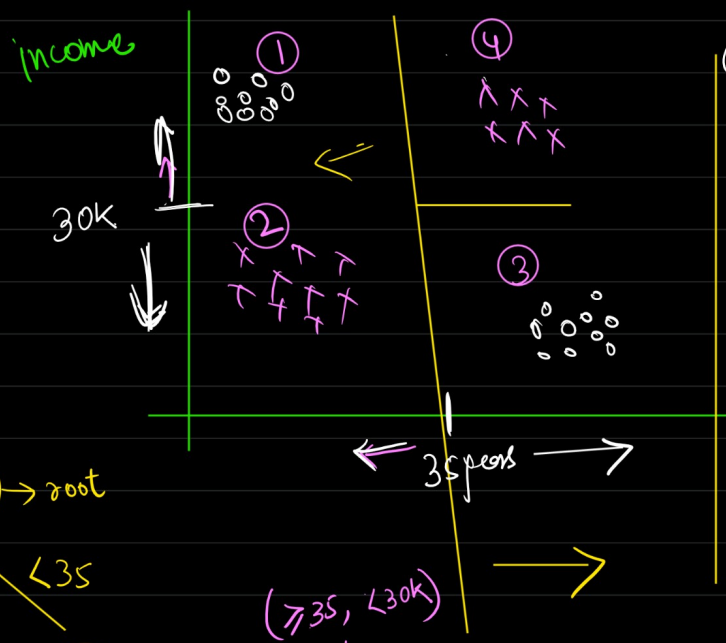
Decision Tree

↓

At every node a decision is taken

↓

It looks like a tree.

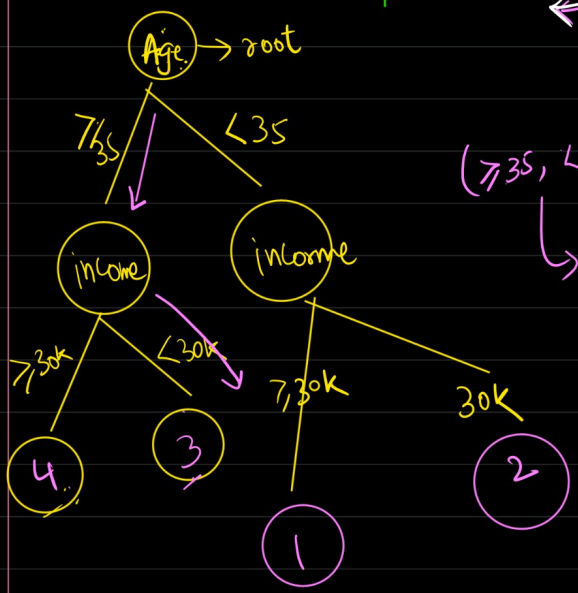


① If income ≥ 30K & Age ≤ 35 years
↳ group 1 → 0th class → Fraudster

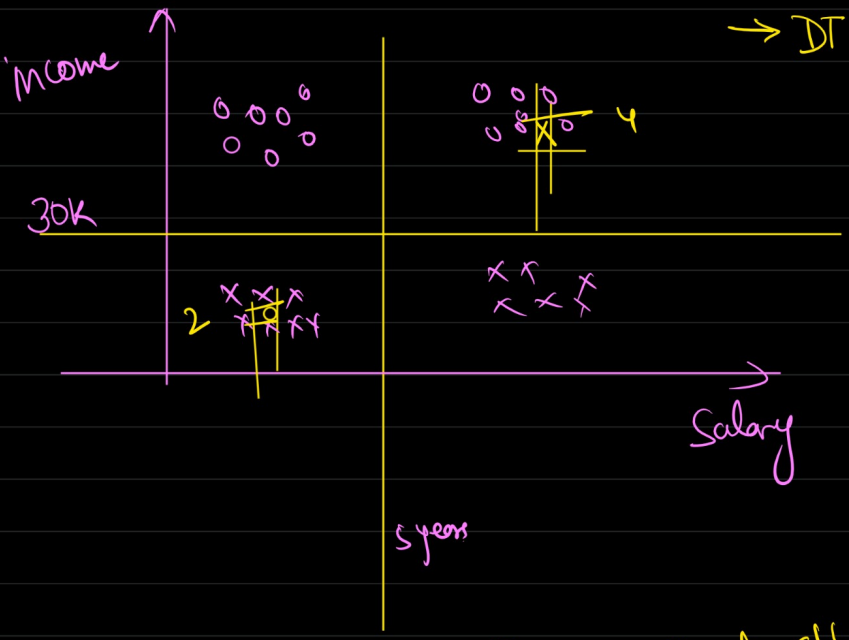
② If income < 30K & Age ≤ 35 years
↳ group-2
↳ xth class → Non fraudster

③ If income ≤ 30K & Age > 35 years
↳ group 3
↳ 0th class
↳ Fraudster

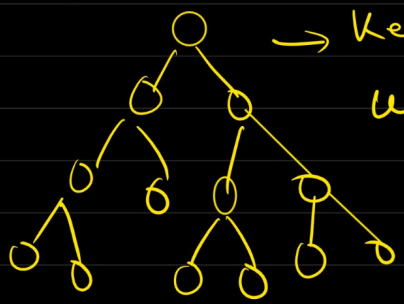
④ If income ≥ 30K & Age ≥ 35 years
↳ group 4
↳ xth class
↳ Non fraudster.



(≥35, <30K)
↳ 0th class.
↓
Fraud



→ DT divides the data into pockets (nodes)

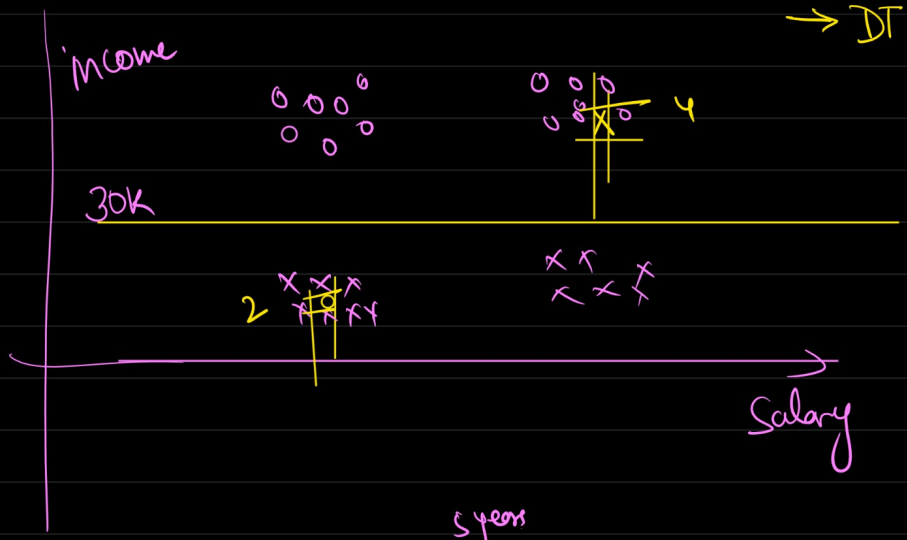


→ Keep splitting | dividing untill all the pockets where each data point are of same class

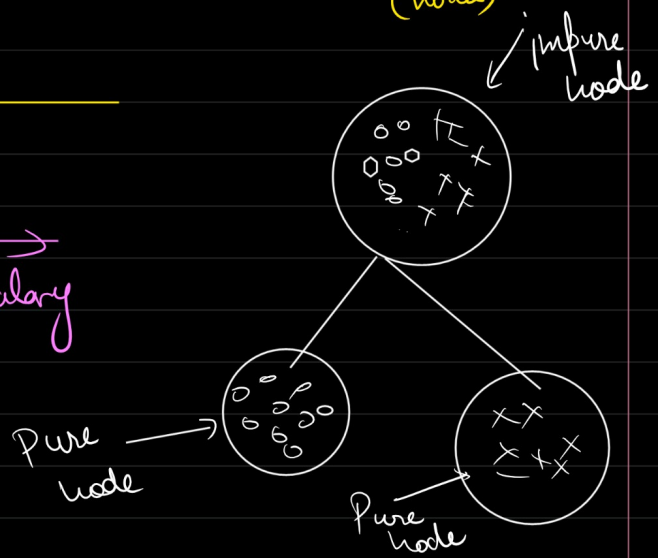
* Advantages

→ Simple to Understand

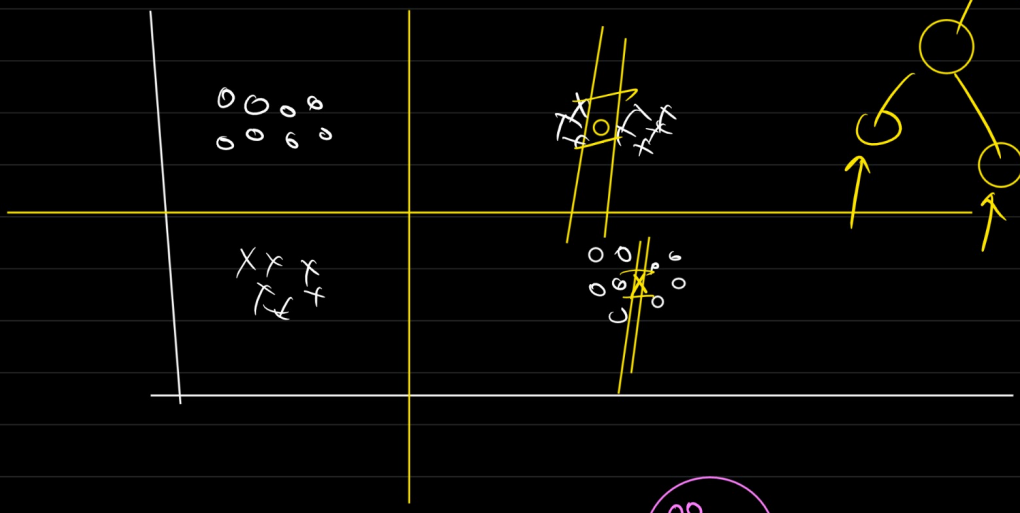
→ Captures non-linear relationship

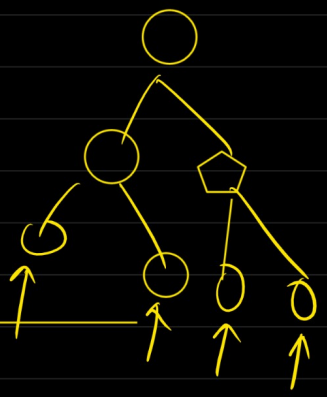


→ DT divides data into Pockets (nodes)

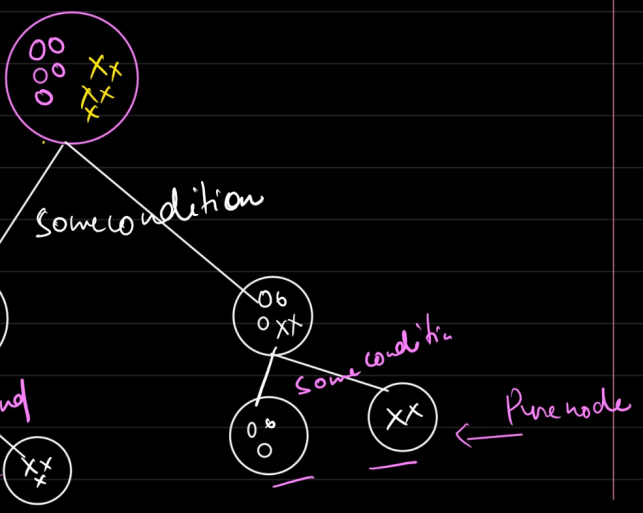


* The DT intend is to create pure leaf node/pockets where each data points belongs to one class





Depth of DT -2

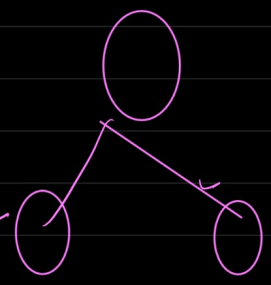


① How this split will happen? (purity / impurity)

② Which features to be used for splitting?

pure – homogenous node (same class)

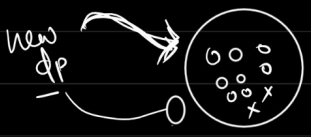
impure – heterogenous node

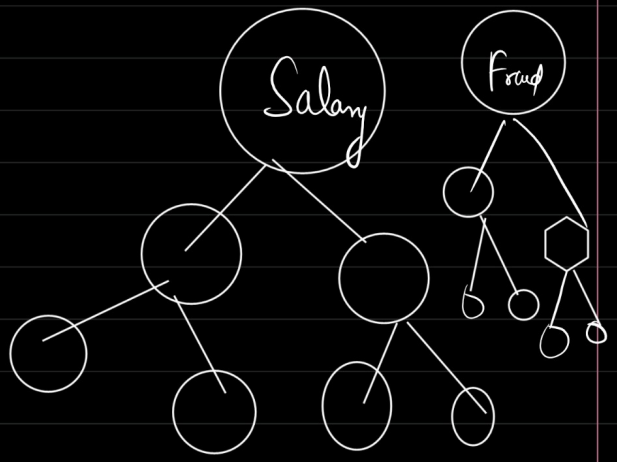


Impurity Measures

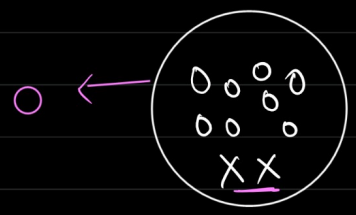
- Classification Error
- Gini
- Entropy.

Salary	Age	Fraud
–	–	0
–	–	1
–	–	0
–	–	1





① Classification Error



0 : 8

X : 2

Classification Error = $1 - \max p_i$

$P_0 = \frac{8}{8+2} = \frac{8}{10} = 0.8$

$P_X = \frac{2}{8+2} = \frac{2}{10} = 0.2$

$C.E = 1 - \max(0.8, 0.2)$

$= 1 - 0.8$

$= 0.2$

* If I assign everything to majority class then what is the error rate.

② Gini

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

Others formulae

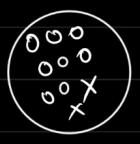
$$GI = \sum_{i=0}^{n-1} p_i (1 - p_i)$$

for two class $p(1-p_0) + p_x(1-p_x)$

$\sum_{i=0}^{n-1}$ ⇒ from class 0 to class n-1

$= 2p_0(1-p_0)$ $\quad (p_0 + p_x = 1,\ p_x = 1-p_0)$

$0.8(1-0.8) + 0.2(1-0.2)$
$= 0.8 \times 0.2 + 0.2 \times 0.8$
$= 0.32$



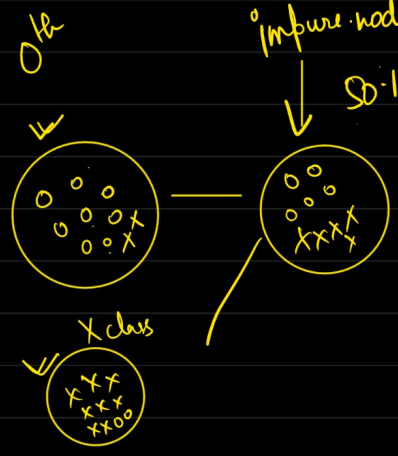
$$G.I = 1 - \sum_{i=0}^{2-1} (p_i)^2$$

$$= 1 - \sum_{i=0}^{1} (p_i)^2 = 1 - \left[(p_0)^2 + (p_x)^2\right]$$

$$= 1 - \left[(0.8)^2 + (0.2)^2\right]$$

$$= 1 - \left[0.64 + 0.04\right]$$

$$= 1 - 0.68 \Rightarrow 0.32$$

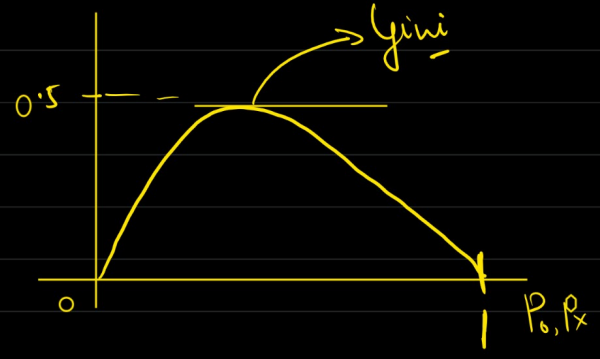


$$G.I = 1 - \left[(p_0)^2 + (p_x)^2\right]$$

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

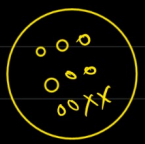
$$= 1 - \left(\frac{1}{4} + \frac{1}{4}\right) = 1 - \frac{2}{4_2} = \frac{1}{2} = 0.5$$

highest Gini impurity value = 0.5



③ Entropy

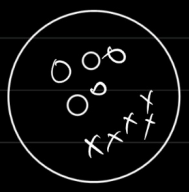
$$H(S) = \ominus \sum_{i=0}^{N-1} p_i \log_2 p_i$$



$$= -p_0 \log_2 p_0 - p_x \log_2 p_x$$

$$= -0.8 \log_2 0.8 - 0.2 \log_2 0.2 \Rightarrow 0.72$$

Highest H(S) → 50% - 50% of both Class



$$H(S) = -\sum_{i=0}^{N-1} p_i \log_2 p_i$$

$$= \underbrace{-\frac{1}{2}\log_2 \frac{1}{2}}_{a} - \underbrace{\frac{1}{2}\log_2 \frac{1}{2}}_{a}$$

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

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

$$= -1 \times -1 = 1$$

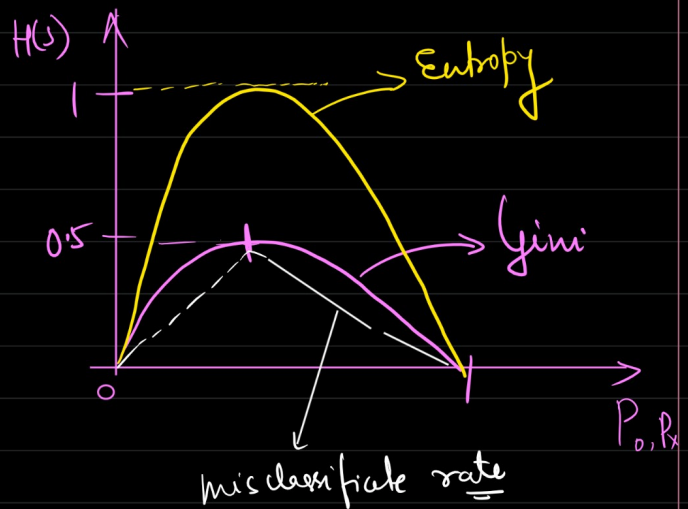
→ Highest Entropy value is 1

→ 1

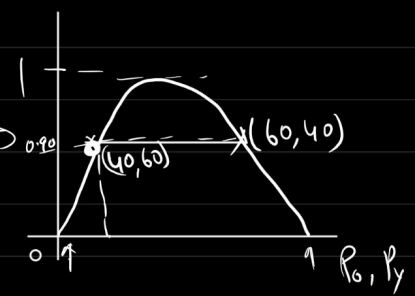
Entropy → from 0 to 1

Gini → 0 to 0.5

misclassification Error → 0 to 0.5



	Class 1	Class 2	Entropy
impure →	50%	50%	1
	40%	60%	0.90
	60%	40%	0.90
Pure →	100%	0%	0
	0%	100%	0



Multiclass classification

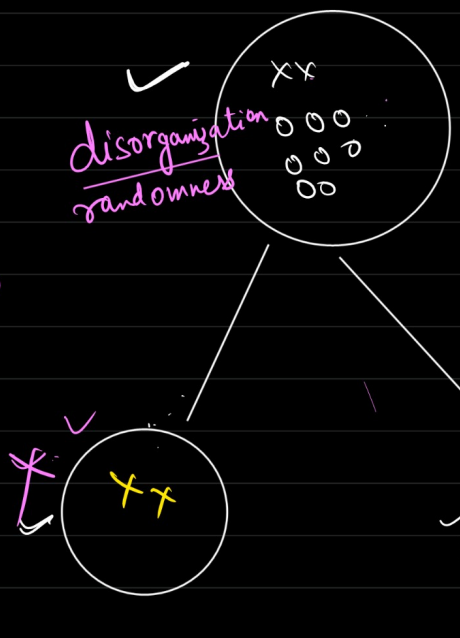
$$\text{Gini Impurity} = 1 - \sum_{i=1}^{n} (p_i)^2$$

$$= 1 - [(p_1)^2 + (p_2)^2 + (p_3)^2]$$

$$\text{Entropy}(H(s)) = -\sum_{i=1}^{N} p_i \log_2 p_i$$

$$= -p_1 \log_2 p_1 - p_2 \log p_2 - p_3 \log p_3$$

Q On what feature to split ? ⇒ Information gain



* Less impure node requires less information to describe it

and

More impure nodes requires more information.

Aim of splitting

↳ We want pure nodes after splitting.

↳ Such a split condition that maximises the purity of nodes.

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

→ Split the does this best is chosen as best split.

* This change in impurity is given by the difference of post split impurity and pre split impurity

$$\Delta \text{impurity} = \text{impurity (pre split)} - \text{impurity (post split)}$$

‖

Information gain

* Whichever feature has the information gain, that will be used for splitting.

* If disorganization / randomness decreases, then we say Information gain — Δ impurity

* Information theory is a measure to define the degree of disorganization in a system ⇒ IG ← GI / Entropy

Gender	Class	Cricket (Y/N)
M	IX	Yes
F	X	No
M	X	Yes
M	IX	
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—
—	—	—

30 students

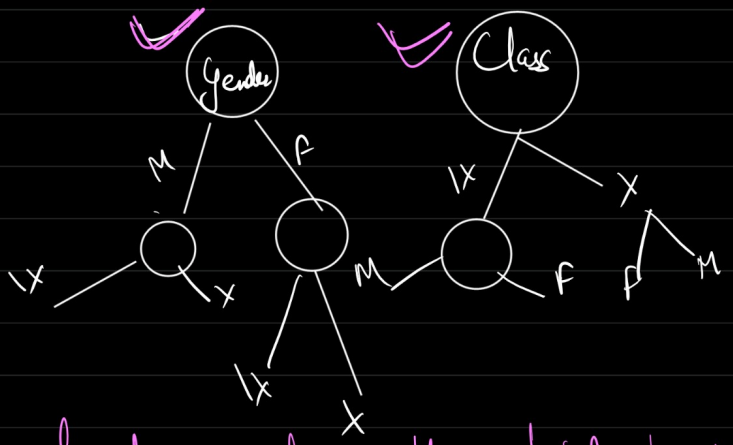
30 → 50% play cricket

Gender → 10/2 – F ⇒ Play cricket (out of 10 female 2 play cricket)

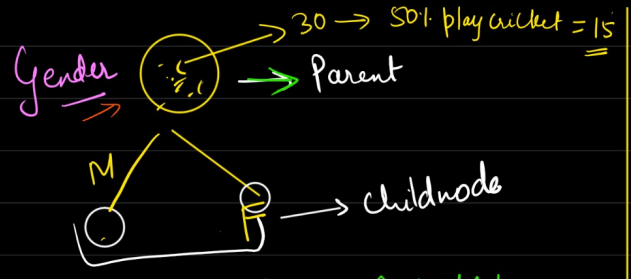
Gender → 20/13 – M ⇒ Play cricket (out of 20 male, 13 play cricket)

Class → IX – 14/43% – out of 14 class IX students, 43% play cricket

Class → X – 16/57% → out of 16 class X students 57% play cricket.



* Whichever feature has the highest information gain, that will be used for splitting

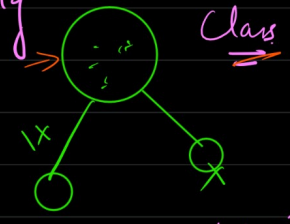


ENTROPY / GI

IG = Imp. Presplit – Impurity Post split

= En (Parent) – En (child combined)

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



IG = En (Parent) – En (Child)

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

En child Class

$$En(IX) \rightarrow \frac{-43}{100}\log_2\frac{43}{100} - \frac{57}{100}\log_2\frac{57}{100} - X_1$$

En child Gender

$E_{female\ child} = -\frac{2}{10}\log_2\frac{2}{10} - \frac{8}{10}\log\frac{8}{10}$

$= 0.72$

$E_{male\ child} = -\frac{13}{20}\log_2\frac{13}{20} - \frac{7}{20}\log_2\frac{7}{20} = 0.92$

$En(child\ combined) = \left(\frac{No\ of\ female}{Total\ student} \times E_{female\ child}\right) + \left(\frac{No\ of\ males}{Total\ stud} \times E_{male\ child}\right)$

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

$= 0.85$

$IG_{Gender} = E_p - E_{cc}$

$= 1 - 0.85 = 0.15$

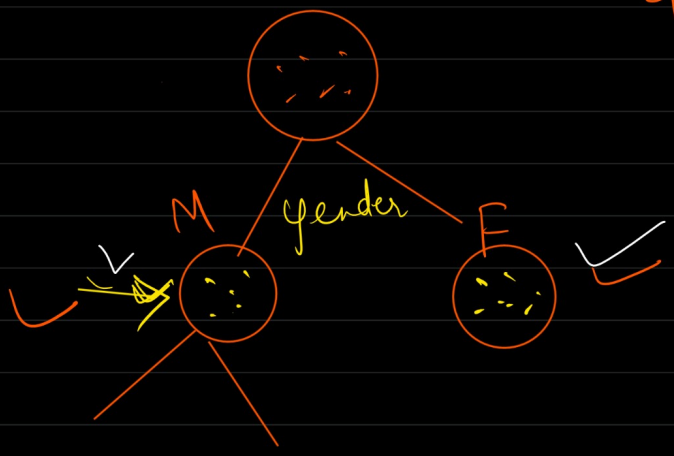
$En(x) = -\frac{57}{100}\log_2\frac{57}{100} - \frac{43}{100}\log\frac{43}{100} - X_2$

$En_{combined} = \frac{14}{30}x_1 + \frac{16}{30}x_2 = 0.99$

$IG_{class} = 1 - 0.99 = 0.01$

Since $IG_{gender} > IG_{class}$

∴ We will Gender to split.



Algorithm for DT

Step-1 Recursive binary splitting | partitioning the data into smaller subset.

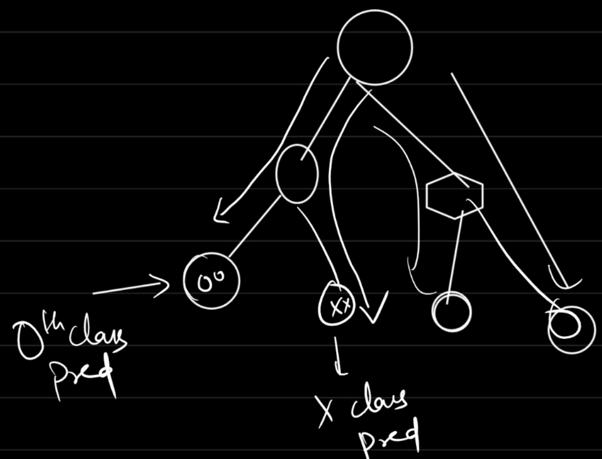
Step-2 Select the best variable for split (Information Gain)

Step-3 Apply the split

Step-4 - Repeat the process for the subset obtained.

Step-5 - Continue the process until every node is a pure node | stopping criteria is reached.

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

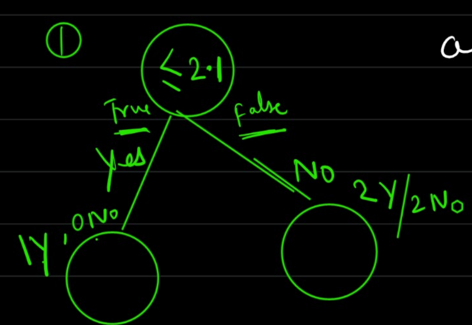


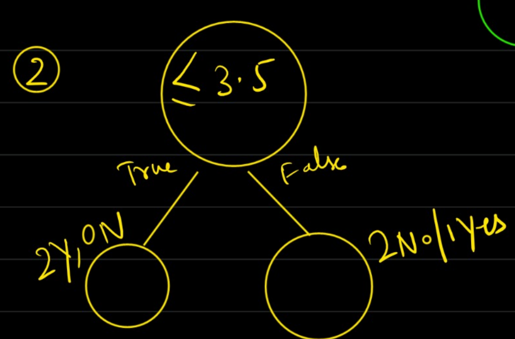
* What if the feature is continuous

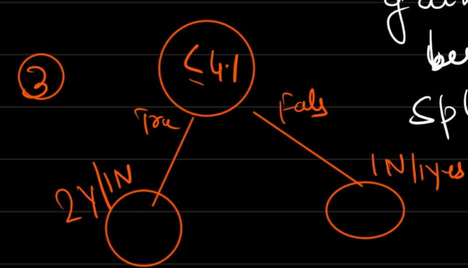
f1 (No of hours studied)	Output (cricket or not)
2.1	yes
3.5	yes
4.1	No
5.3	No
6.8	yes

① Sort the feature (f1)

② treat each value/row as a cut off | threshold and try to build decision tree and whichever threshold gives you the best information gain that will be used for splitting.







* Disadvantage

Time-complexity ↑↑

* Decision tree is a Greedy Algorithm

⇓

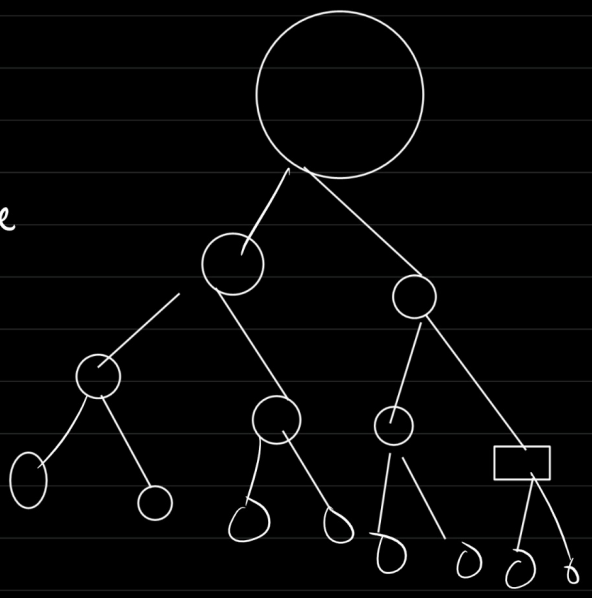
It keeps on splitting until every leaf node is a pure node

⇓

memorise the data

⇓

Overfitting



- Prepruning
- Post pruning
- Ensembles