Hare Krishna

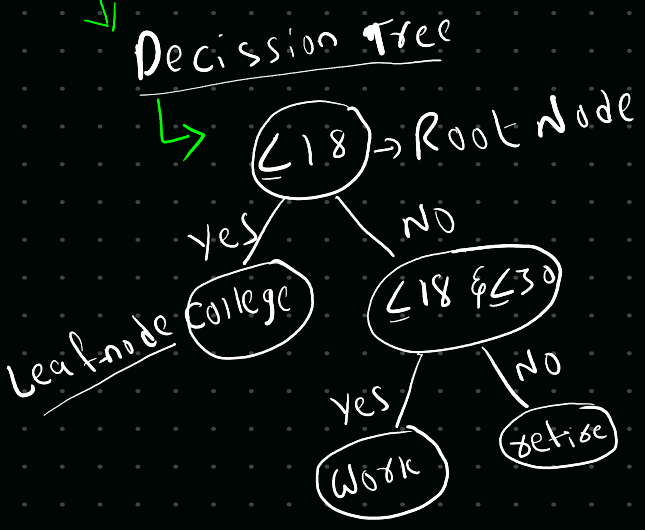
Agenda

1. Decission Tree Classification
2. Decission Tree Regression
3. Practical implementation
4. Ensemble Technique

} Agenda

Decission tree {Sloving many use case}
- Regression
- Classification }

```
if (age <18):
    Print("college")
Elif (age >18 and age <=35) :
    Print("Work")
Else :
    Print("retire")
```



Decission Trees

Nest if else ⇒ Decision tree

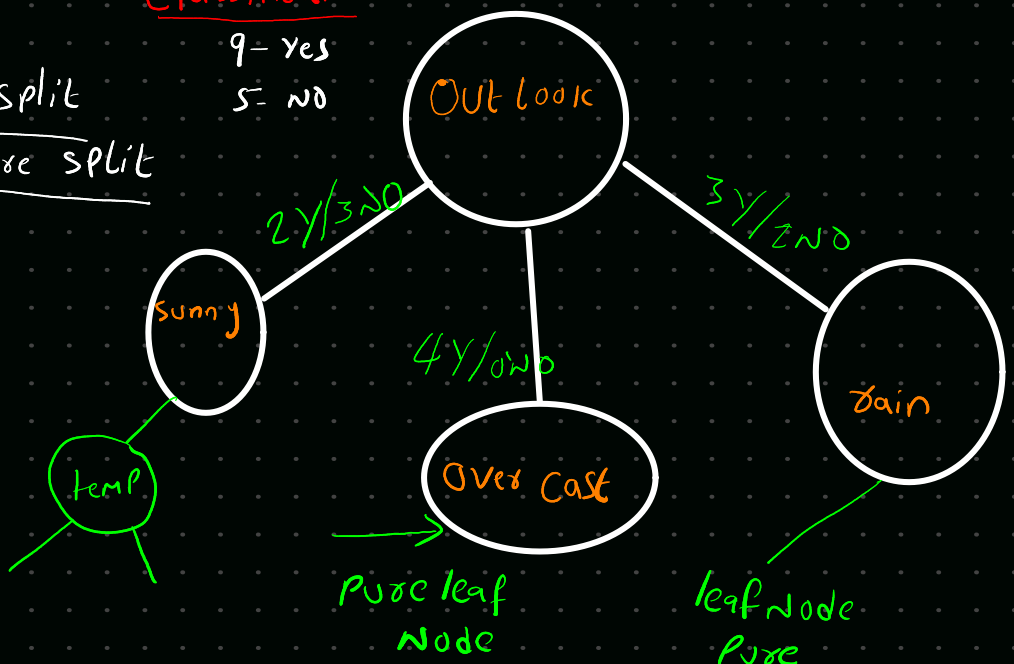
Classification
9- Yes
5- NO

Sunny — 2 Yes
3 NO

Over cast — 4 Yes
0 NO

Rain — 3 Yes
2 NO

Pure split
Impure split



① Purity → Pure Split ?!

↳ Entropy

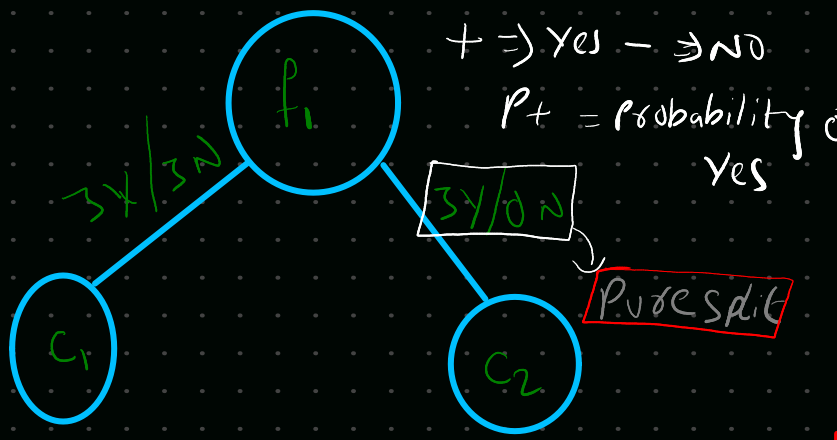
↳ Gini Impurity

② How the features are selected

↳ Information Gain!

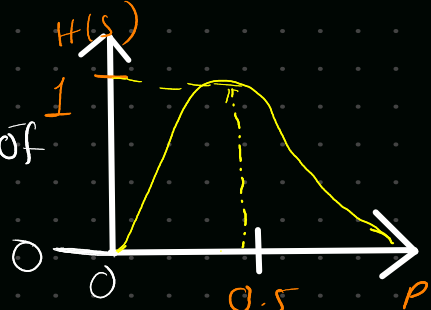
① Entropy

$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$



+ ⇒ Yes − ⇒ No

P_+ = probability of Yes



$P^+ = 0.5$

$P^- = 0.5$

$P = 1 - Q$

$Q = 1 - P$

Entropy $H(S) = \frac{-3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$

$= -1 \log_2 1$

$= 0$ → Pure split

↳ Entropy

$H(S) = \frac{3}{6}\log_2 3/6 - 3/6 \log_2 3/6$

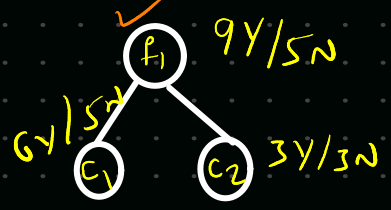
$= 1$

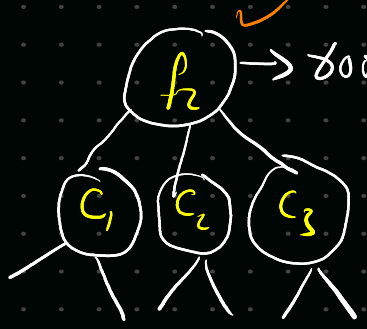
Impure split

① Gini impurity

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

② Which feature to take to split?

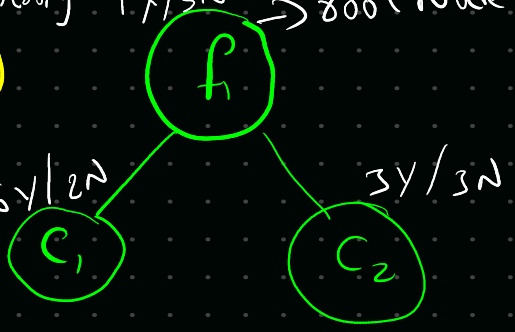




Information Gain :-

$$Gain(S, f_i) = H(S) - \sum_{v \in Val} \frac{|S_v|}{|S|} H(S_v)$$

H(S) → Entropy 9Y/5N → root Node



$$H(S) = -P_+ \log_2 P_+ - P_- \log_2 (P_-)$$

$$= -\frac{9}{14} \log_2 (9/14) - \frac{5}{14} \log_2 (5/14)$$

$$\approx 0.94$$

$$H(C_1) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8}$$

$H(C_1) = 0.81$ $H(C_2) = 1$

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

$$= 0.049$$

$Gain(S, f_1) = 0.049$

$Gain(S, f_2) = 0.051$

Using which feature should I start splitting first

$Gain(S, f_2) >> Gain(S, f_1)$

* Gini Impurity :-

$n = 2$ output { Yes / No

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

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

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

$$= 1 - \left[\frac{1}{2}\right] = 0.5$$

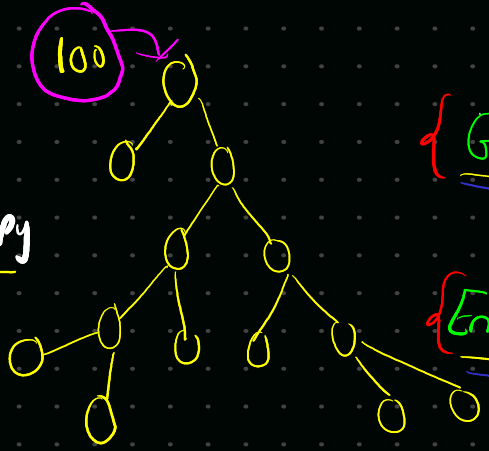
2Y/2N ⇒ Impure split

Entropy = 1

Gini Impure = 0.5

Entropy ⇒ log

Gini impurity ⇒ Simple Math

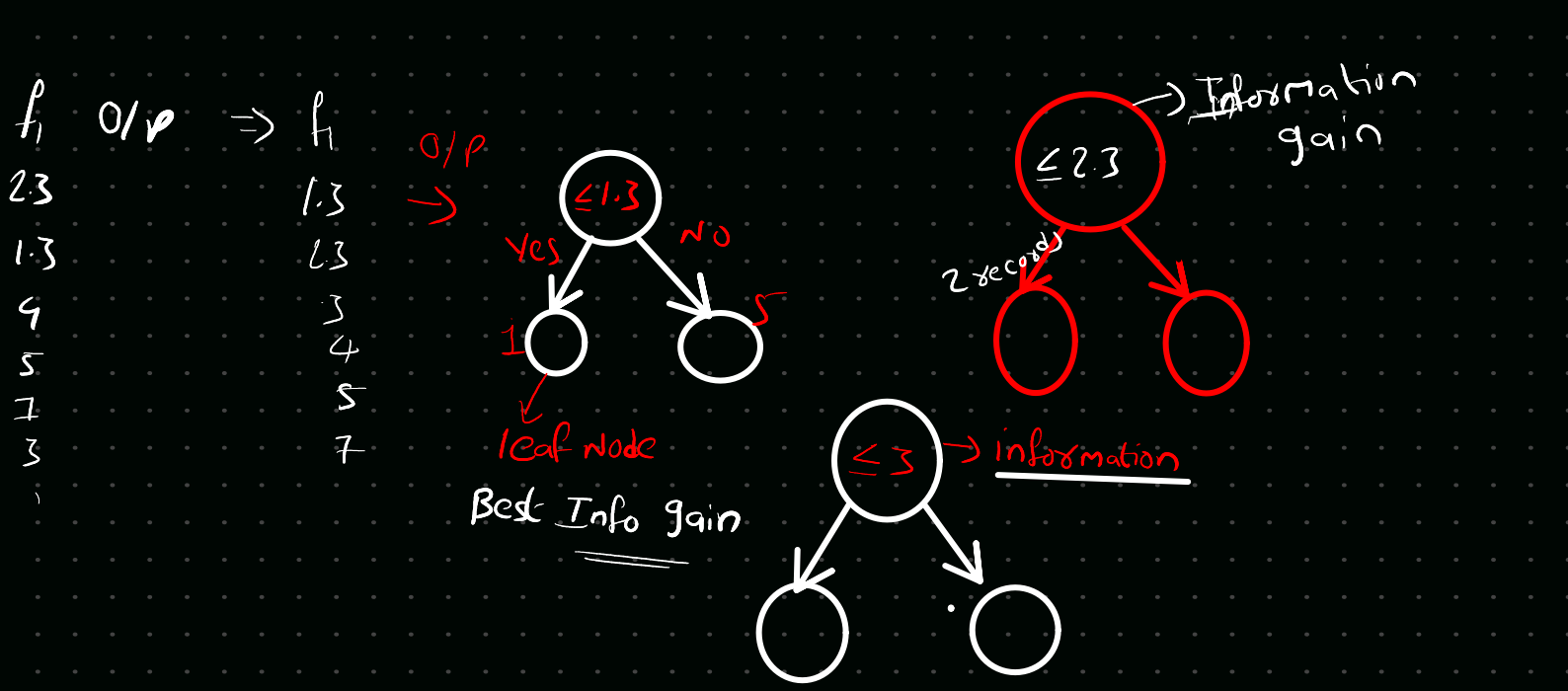


Fast

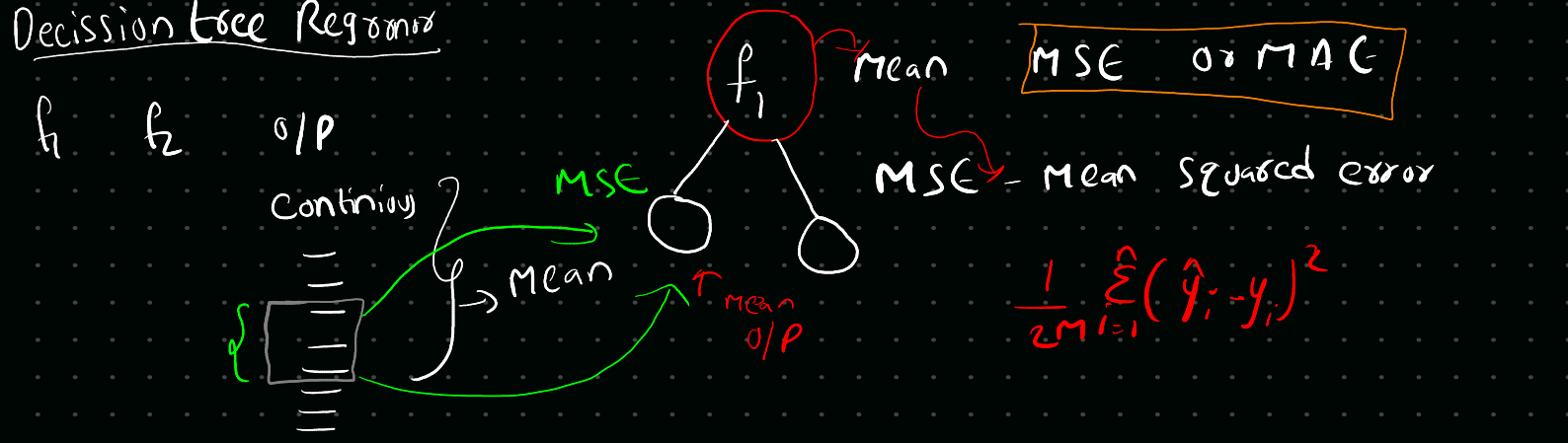
Gini >> Entropy

{Gini} → should be used when huge numbers are there

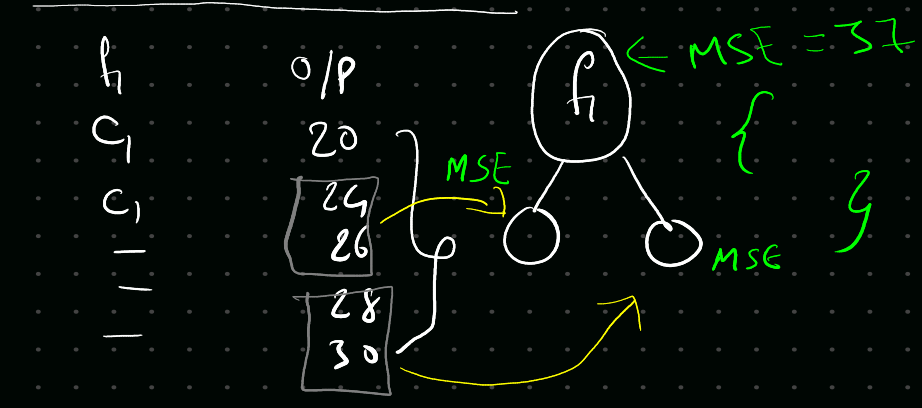
{Entropy} → should be used less no are there

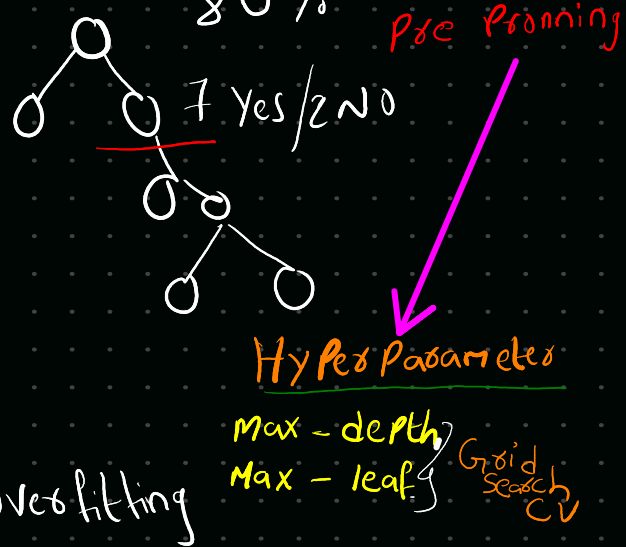


Decission tree Regressor



Decision tree regressor





HyPer Parameters :-

Decision → over fitting

① Post Pruning

② Pre Pruning