Day 4 - Machine Learning Algorithms

Agenda

1. Decision Tree CLASSIFICATION
2. DECISION TREE REGRESSION
3. PRACTICAL IMPLEMENTATION
4. Ensemble Techniques

Agenda

{DAY 1, DAY 2, DAY 3}

⇓

Experience

Decision Tree {Solving many usecase}

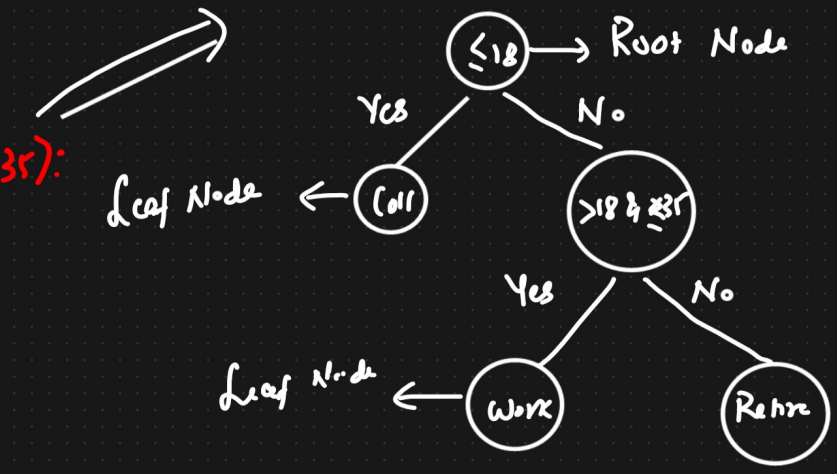
- Regression
- Classification

Regression & Classification

⇓

Decision Trees

```
if (age ≤ 18):
    Print ("College")
elif (age >18 and age ≤ 35):
    Print (Work)
else :
    Print ("Retire")
```



DECISION TREE

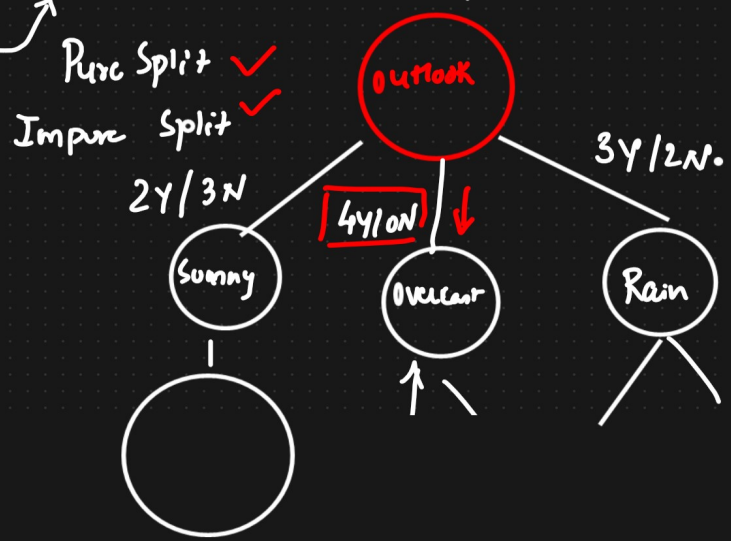
Nest if else ⇒ Decision Tree

CLASSIFICATION 9Y/5N

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Pure Split

Impure Split



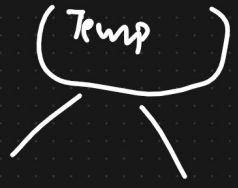
① Purity → Pure Split ??

↳ Entropy

↳ Gini Impurity

② How the features are selected

↳ Information Gain ?



Pure leaf Node

① Entropy

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

+ } Binary
− }

+ ⇒ Yes − ⇒ No

P_+ = Probability of Yes

→ Pure split

→ 0

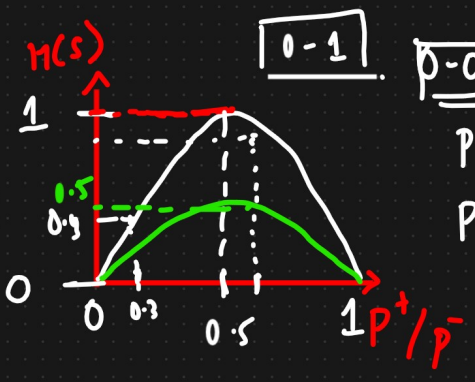
Entropy $H(S) = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$

$= -1 \log_2 1$

$= 0$ → Pure split

① Gini Impurity

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



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

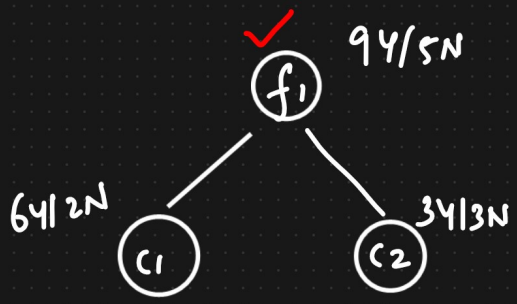
$= 1$ ✓

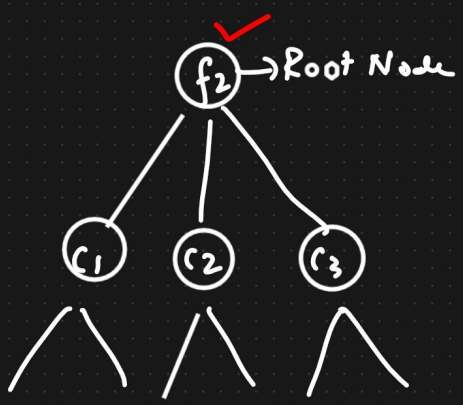
↳ Impure Split

Purity Test → Entropy

f1 f2 f3

② Which feature to take to split??

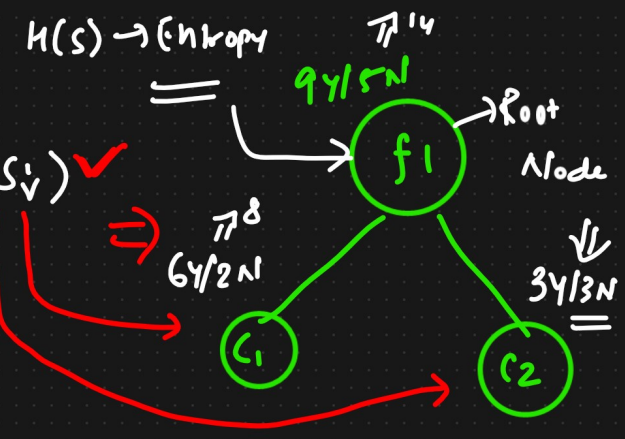




Information Gain

$$Gain(S, f_1) = H(S) - \sum_{v \in Val} \frac{|S_v|}{|S|} H(S_v)$$

(Root Node)



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

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

$$\approx = \boxed{0.94}$$

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

$$\boxed{H(C_1) = 0.81} \qquad \boxed{H(C_2) = 1}$$

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

$$Gain(S, f_1) = 0.049$$

$$Gain(S, f_2) = 0.051$$

Using which feature should I start splitting first

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

② GINI Impurity ✓

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

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

{Entropy}

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

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

n = 2 output { Yes, No

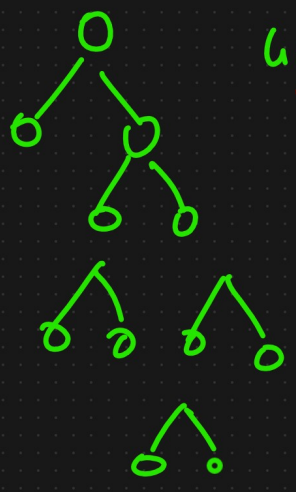
2Y/2N ⇒ Impure Split

Entropy = 1

Gini Impurity = 0.5

Entropy → fog

100



Gini Impurity → Simple Maths

Fast

Gini >> Entropy

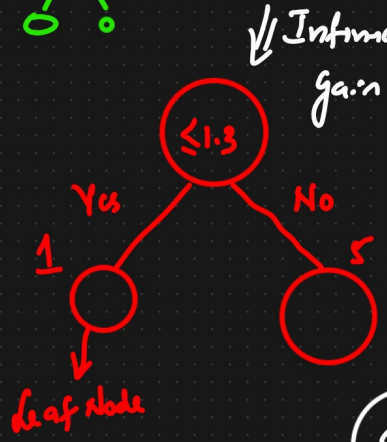
→ Continuous

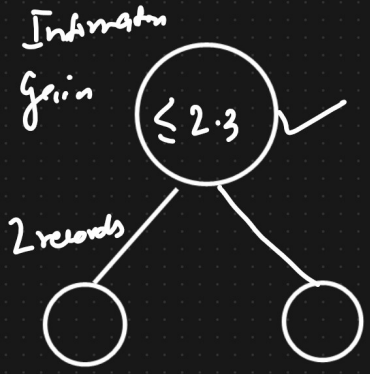
f1	O/P
2.3	
1.3	
4	
5	
7	
3	

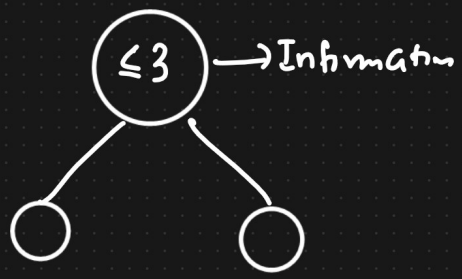
⇒

f1
1.3
2.3
3
4
5
7

O/P →



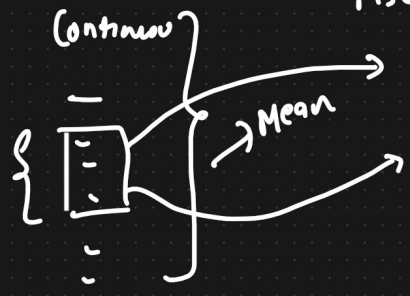


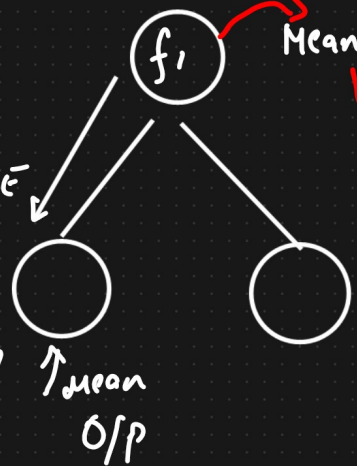


Best Info gain

Decision Tree Regressor

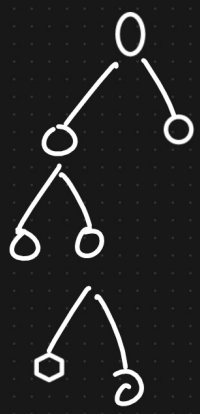
f1	f2	O/P





MSE or MAE

$$\frac{1}{2m}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$$



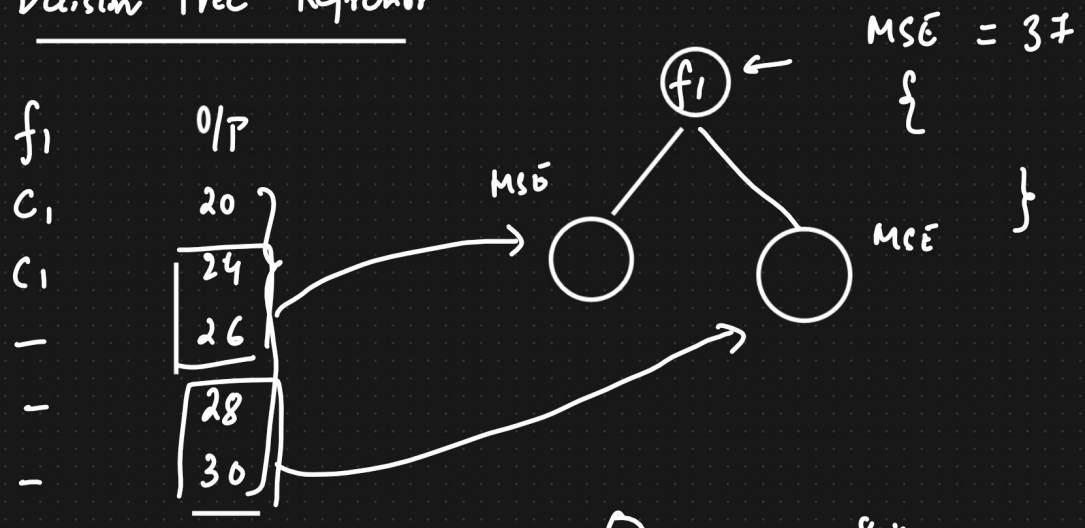
Overfitting

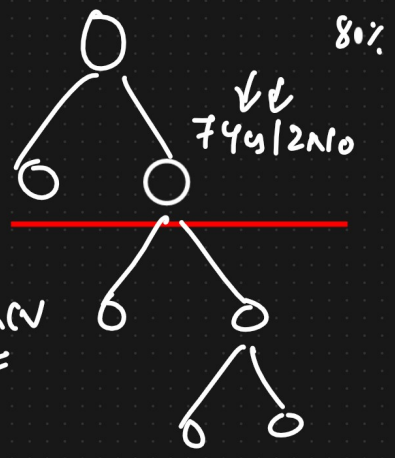
Hyperparameter

Decision → Overfitting

{ ① Post Pruning
② Pre Pruning }

Decision Tree Regressor





post pruning

pre pruning

Hyperparameter

GridSearchCV

max_depth, max_leaf