Data Science & AI

Machine Learning

Supervised Learning

Lecture No.- 04

By- Krish Naik Sir

Recap of Previous Lecture

Topic Ridge

Topic lasso

Topic Regression

Topic

Topic

Topics to be Covered

Topic — Decision Tree

Topic — Decision Tree Regressor

Topic — Decision Tree Classifier

Topic

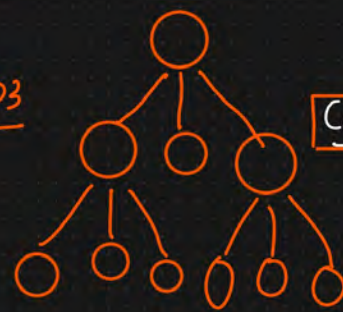
Topic

Decision Tree Classifier And Regressor

Agenda

① Decision Tree Classifier [classification]
② Decision Tree Regressor [Regression]

ID3



CART

① Decision Tree Classifier

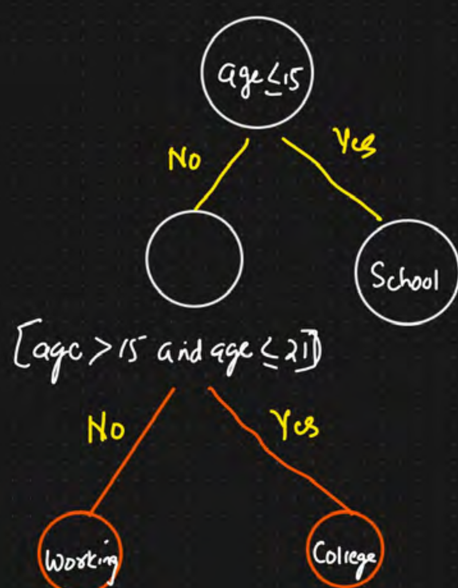
Two Types

① ID3 [Iterative Dichotomiser 3]

② CART [Classification And Regression Trees]

age = 14

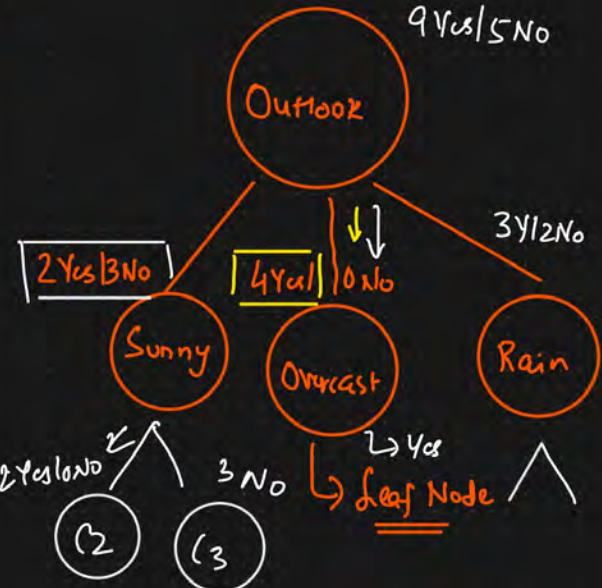
```
if (age ≤ 15):
    Print ("School")
elif (age > 15 and age ≤ 21):
    Print ("College")
else (age > 21):
    Print ("working")
```



Dataset : Predict Play Tennis or Not

O/P

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



① Purity check : Pure Split or Impure Split

→ Entropy
→ Gini Impurity

} ⇒ Measure of purity. }

② What feature you need to select to start the split? → Information Gain}.

2Yes/2No

$P_+ = \frac{2}{4} = 0.5$

① Entropy

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

P_+ = probability of +ve category

P_- = probability of negative category

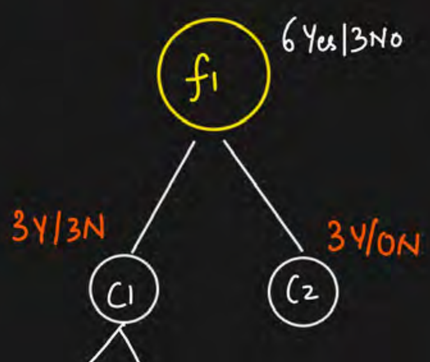
ii) Gini Impurity

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

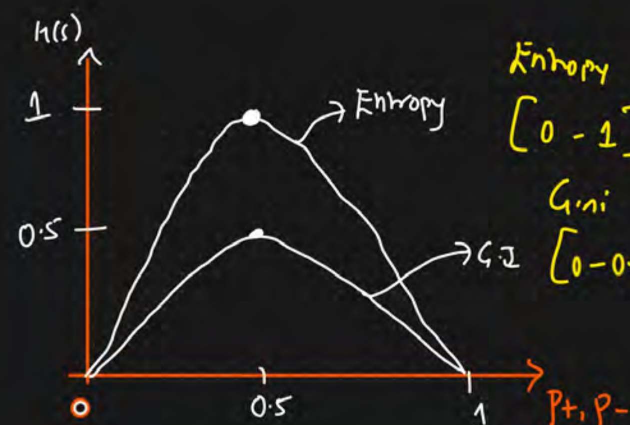
$\boxed{2Y/3N}$

⇓

0 to 1



$$H(C_1) = -P_+ \log_2 P_+ - P_- \log_2 P_-$$



Entropy [0 - 1]

Gini [0 - 0.5].

$$= -\frac{1}{2}\log_2(1/2) - \frac{1}{2}\log_2(1/2)$$

$$= 1 \Rightarrow \text{Impure Split}$$

Multiclass Classification $[C_1, C_2, C_3]$

$$H(S) = -P_{C_1}\log_2 P_{C_1} - P_{C_2}\log_2 P_{C_2} - P_{C_3}\log_2 P_{C_3}.$$

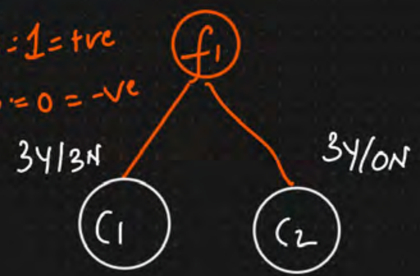
$$H(C_2) = -3/3\log_2(3/3) - \frac{0}{3}\log_2(0/3)$$

$$= 0 \Rightarrow \text{Pure Split}$$

② Gini Impurity

Yes = 1 = +ve

No = 0 = -ve



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

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

$$= 1 - [(3/6)^2 + (3/6)^2]$$

$$= 1 - [1/4 + 1/4] = 0.5$$

⇓

Impure Split.

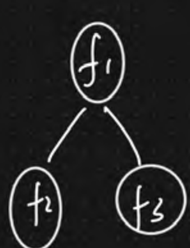
$$G.I(C_2) = 1 - [(1)^2 + (0)^2]$$

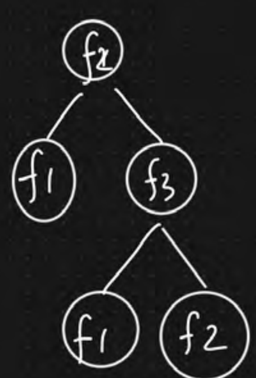
$$= 1 - 1$$

$$= 0 \Rightarrow \text{Pure Split.}$$

② What feature you need to select to start the split? → Information Gain

f1 f2 f3 O/P



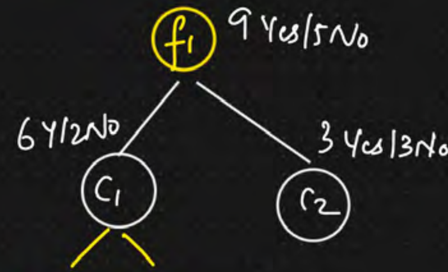


Information Gain

f_1	f_2	f_3	O/P

Entropy of the root Node

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



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

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

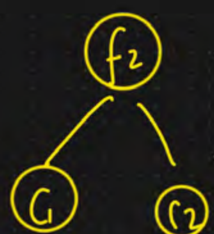
$$Gain(S, f_1) = 0.94 - \left[\frac{8}{14} * 0.81 + \frac{6}{14} * 1\right].$$

$$\boxed{H(S) \approx 0.94}$$

$$\boxed{Gain(S, f_1) = 0.049}$$

$$H(C_1) = -\frac{6}{8} \log_2 \frac{6}{8} - \frac{2}{8} \log_2 \frac{2}{8} \approx \boxed{0.81}$$

$$H(C_2) = 1$$



$$\boxed{Gain(S, f_2) = 0.051}$$

$$\boxed{Gain(S, f_2) = 0.051} > \boxed{Gain(S, f_1) = 0.049}$$

Higher the IG ↑

Entropy Vs Gini Impurity

Whenever dataset is small → Entropy

dataset is huge → Gini Impurity

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

→ Outlook Temperature $\boxed{\text{Play}}$

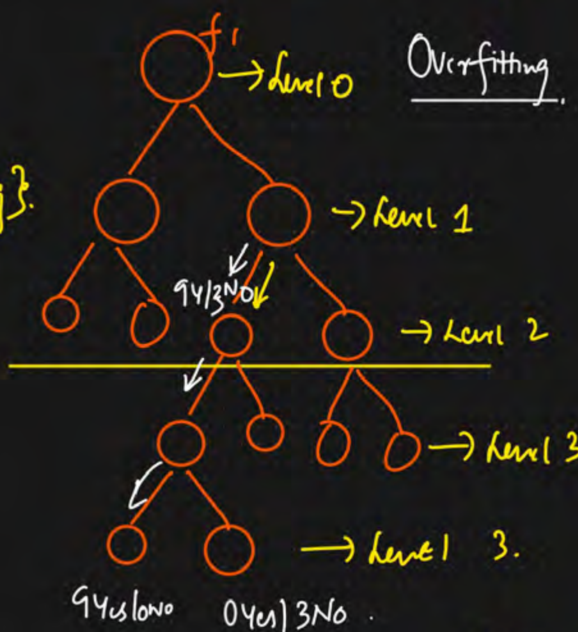
Sunny Cold $\boxed{\text{Yes}}$

Decision Tree Post Prunning And Pre prunning [Post Prunning And Pre prunning].

Post prunning {Reduce Overfitting}.

⇓ Small datant

⇓ Yes



Select the best parameters

Pre Prunning → DATASET → Suitable parameter ⇒ Hyperparameter Tuning.

⇓

Max depth, Split,

Criterion

Gini

Entropy

max-depth = 4

→ max depth = 3

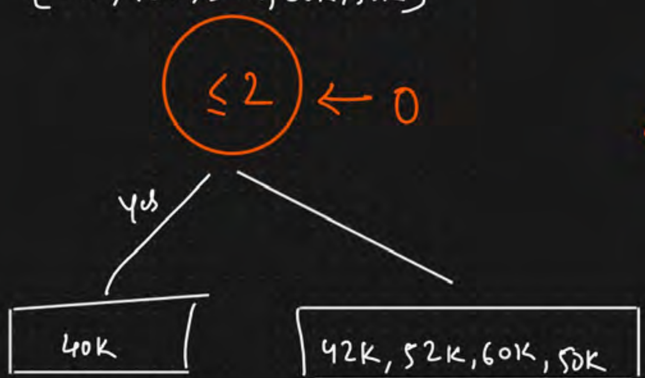
⇒ Accuracy ⇒ Hyperparameter Tuning.

④ Decision Tree Regressor

DATASET

Exp	Career Gap	Salary (↓ o/p)
→ 2	Yes	40K
2.5	Yes	42K
→ 3	No	52K
4	No	60K
4.5	Yes	56K
		$\bar{y}$ = 50K

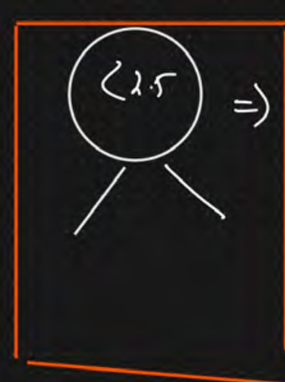
[40K, 42K, 52K, 60K, 50K]



<

↓ Split {Variance Reduction}.

⇓



⇒ Var (0.04).

Variance Reduction

$$\text{Variance (Root)} = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2 \quad [\text{MSE}].$$

↑ mean of y.

$$= \frac{1}{5}\left[(10)^2 + (8)^2 + (2)^2 + 10^2 + 6^2\right]$$

Variance Reduction :

$$\text{Var(Root)} - \sum W_i \,\text{Var(child)}$$

$$60.8 - \left[\frac{1}{5} * 100 + \frac{4}{5} * 51\right]$$

$Variance(Root) = 60.8$

$Var(Right) = \underline{\underline{51}} = \underline{\underline{0}}$.

$Var(Left) = \frac{1}{1}\left[(40-50)^2\right]$

$Var(Left) = 100$

THANK - YOU