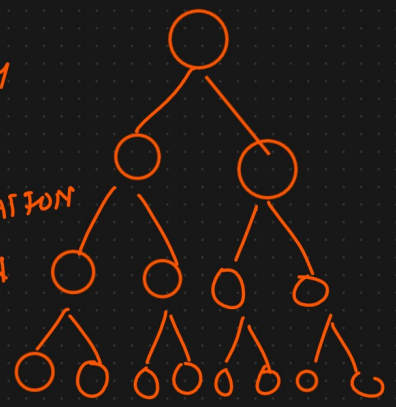
Decision Tree Regression

Dataset

Exp	Career Gap	Salary
2	Yes	40K
2.5	Yes	42K
3	No	52K
4	No	60K
4.5	Yes	56K

op

Entropy
GINI
INFORMATION GAIN



CART

[40K, 42K, 52K, 60K, 56K]

50K

60.8

≤ 2

Yes → 40K

No → 42, 52, 60, 50

V(C1) = 100

V(C2) = 51

[40K, 42K, 52K, 60K, 56K]

60.8

≤ 2.5

40, 42

52, 60, 56

Decision Tree Regressor

Var(C1) = $\frac{1}{2}\left[(40-50)^2 + (42-50)^2\right]$

$= \frac{1}{2}\left[100 + 64\right]$

$= \frac{164}{2} = 82$

Var(child2) = $\frac{1}{3}\left[4 + 100 + 36\right]$

$= \frac{140}{3} = 46.66$

Variance Reduction (Regression Problem)

① Variance $= \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$

Mean Square Error

→ Average

Variance (Root) $= \frac{1}{5}\left[(40-50)^2 + (42-50)^2 + (52-50)^2 + (60-50)^2 + (56-50)^2\right]$

$= \frac{1}{5}\left[100 + 64 + 4 + 100 + 36\right]$

$= 60.8$

Variance Reduction

$= 60.8 - \left[\frac{2}{5} \times 82 + \frac{3}{5} \times 46.66\right]$

$= 60.8 - \left[32.8 + 27.996\right]$

$= +0.004$

Variance (C1) $= \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$

$$= \frac{1}{1} (40-50)^2$$

$$= 100$$

$$\text{Variance (c2)} = \frac{1}{n} \sum_{i=1}^{n} (y-\bar{y})^2$$

$$= \frac{1}{4} \left[(42-50)^2 + (52-50)^2 + (60-50)^2 + (56-50)^2 \right]$$

$$= \frac{1}{4} \left[64 + 4 + 100 + 36 \right]$$

$$= 51$$

Variance Reduction ↓

$$= \text{Var(Root)} - \sum w_i \text{Var(child)}$$

$$= 60.8 - \left[\frac{1}{5} \times \overset{20}{\cancel{100}} + \frac{4}{5} \times 51 \right]$$

$$= 60.8 - 20 - 40.8$$

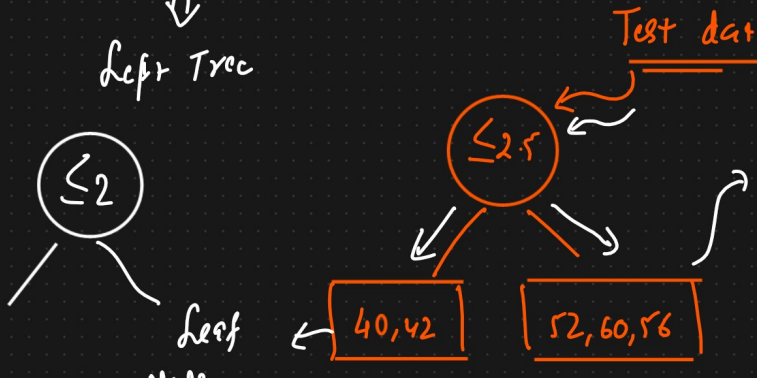
Variance Reduction = 0

0 0.004

Variance Reduction (Left Split) < VR (Right Split)

⇓

Left Tree



Test data

Decision Tree O/P Reg.

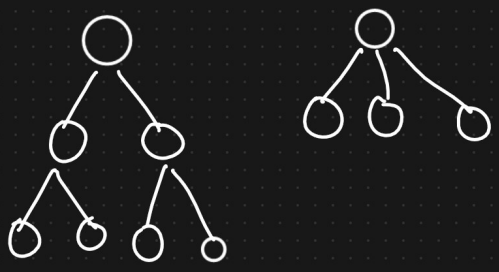
$\frac{52+60+56}{3} = 56$

O/P = $\left[\frac{40+42}{2} \right]$

$= 41$

Decision Tree Classifier

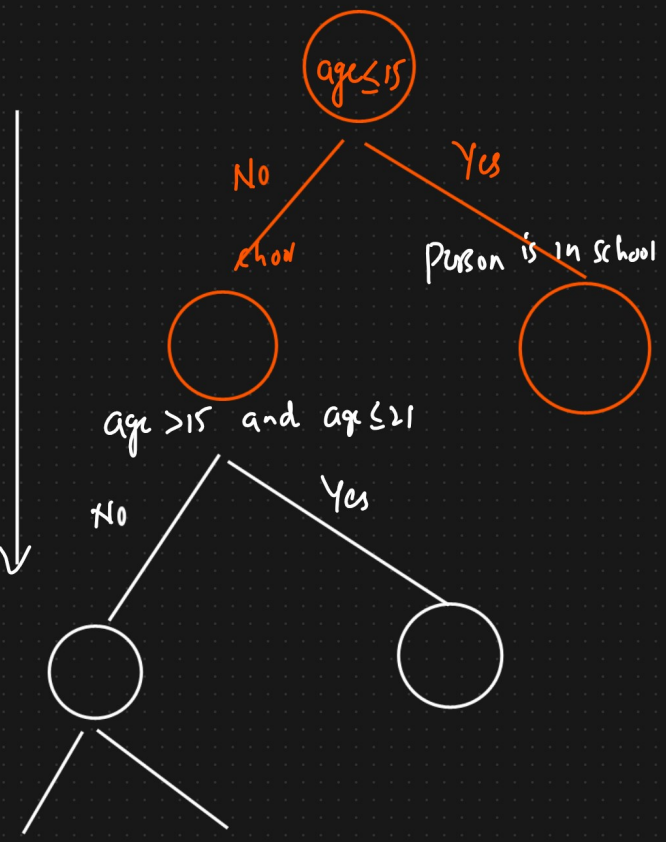
Decision Tree Classifier → ID3
→ CART ✓



a) Entropy and Gini Index → Purity Split

b) Information Gain → features to select for DT construction

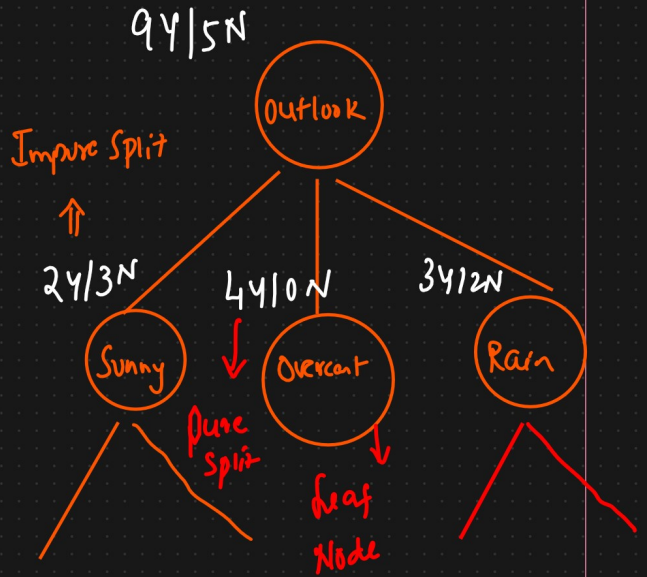
```
age = 14
if (age ≤ 15):
    Print ("The person is in School")
elif (age > 15 and age ≤ 21):
    Print ("The person may be college")
else:
    Print ("The person has passed")
```



Dataset

Binary classification

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 → Pure or Impure Split

↳ Entropy

↳ Gini Impurity

② What feature you need select for Splitting → Information Gain

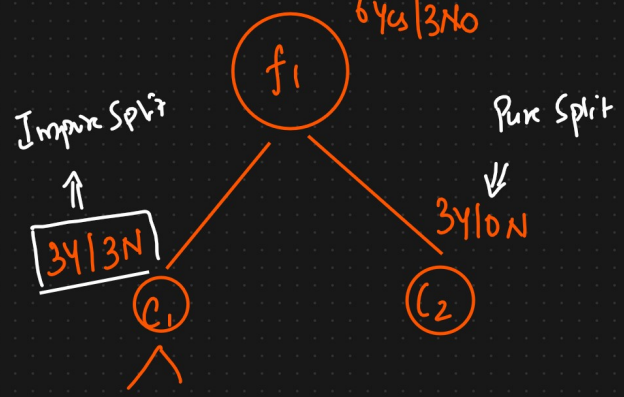
1

0

{Binary classification}

1) Entropy

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

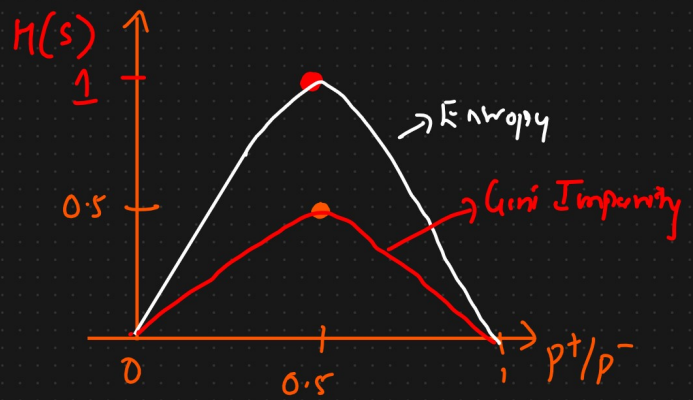


$$H(C_1) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

$$= -\frac{3}{6} \log_2 3/6 - \frac{3}{6} \log_2 3/6$$

2) Gini Impurity

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



$= 1 \Rightarrow$ Impure Split

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

$= -1\log_2 1 \Rightarrow 0 \Rightarrow$ Pure Split

② Gini Impurity

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

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

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

$= 0.5 \Rightarrow$ Impure Split

3Y/0N

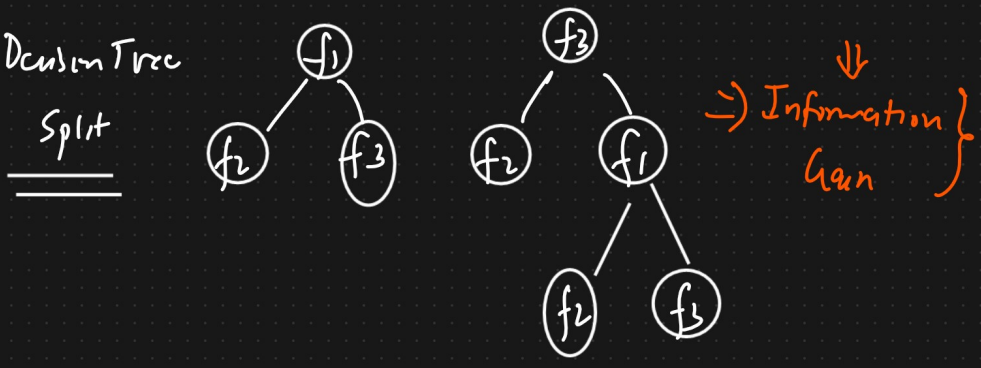
$= 1 - \left(\left(\frac{3}{3}\right)^2\right)$

$= 1 - 1$

$= 0 \Rightarrow$ Pure Split

f1 f2 f3

Decision Tree Split

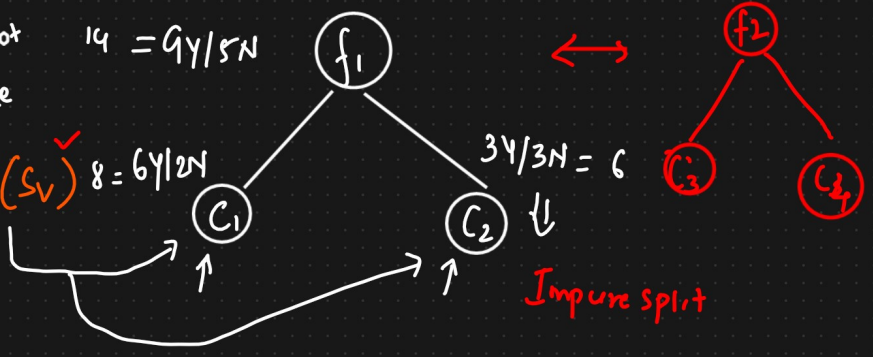


⇓ ⇒ Information Gain

Information Gain

f1 f2 f3 O/P

14 = 9Y/5N



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

↑ Entropy of the root node

↑ Root Node

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

$= -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$

≈ 0.94

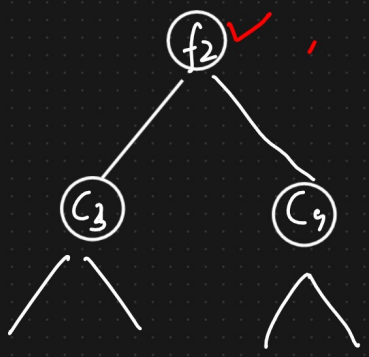
$H(C_1) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8}$

$H(C_1) = 0.81$

$H(C_2) = 1$

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

$Gain(S, f_1) = 0.049$



$Gain(S, f_2) = 0.051 > Gain(S, f_1) = 0.049$

Gain

Information is Basically calculated.

Entropy Vs Gini Impurity

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

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

O/P = 3 categories

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

Whenever dataset is small → Entropy

large → Gini Impurity

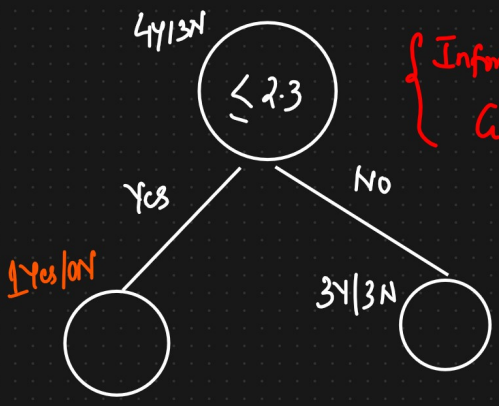
Decision Tree Split for Numerical Feature.

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

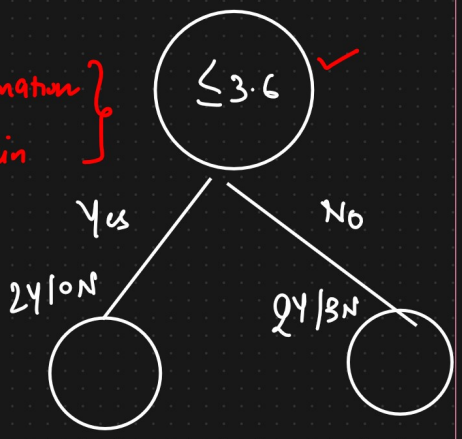
① Sort the feature value

f1	O/P
2.3	Yes
3.6	Yes
4	No
5.2	No
6.7	Yes
8.9	No
10.5	Yes

① Threshold = 2.3



② Threshold = 3.6



{ Information Gain }

Millions of records

Time Complexity ↑↑