Welcome back everyone

2 Questions

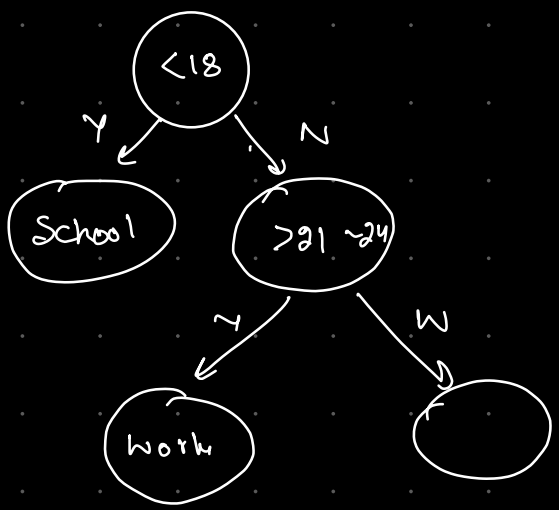
1. Classes in the morning from 10 - 1
2. If anyone has used a good Condenser mic.

Decision Tree

1. Classification
2. Regressor

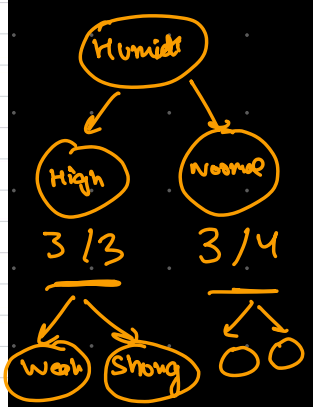
Nested if else Clause

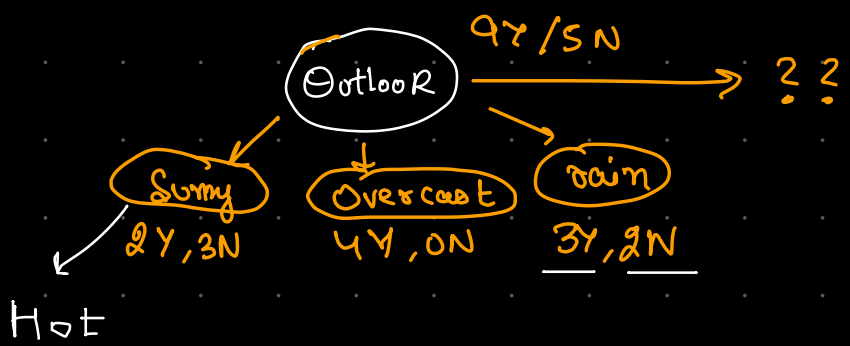
```
if (person < 18)
    ——— "college" school
elif person 18 - 22
    ——— 'College'
elif person 24 - 30
    ——— Job
else
    ——— Business.
```



Person ⇒ 14

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No





2 techniques in DT

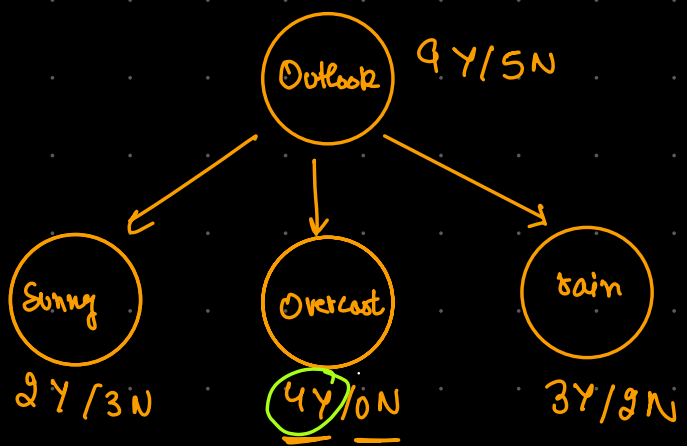
① ID3

② CART *

(binary)

f_1 f_2 f_3 f_4 Play

ID3 and CART have same functionality
but CART → binary classification
and ID3 → not lead to binary classification



split further | leaf node | split further

⇒ We will continue to split until we get a leaf node

1. How will the model know that it is a leaf node?
2. How to select column on which we should divide first?

Our trees can be a lot diff. depending on how we start.

Purity of nodes ⇒ DT gonna check whether the node after split is pure or impure

(1) Entropy

(2) Gini Index.

For column to select first/prioritize, we use something called Information Gain.

① Entropy.

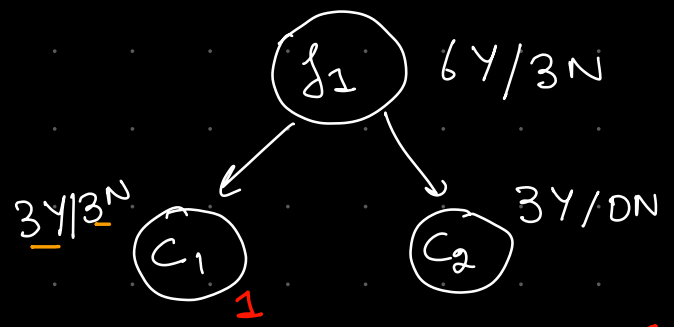
$$H(\text{of a node}) = -\sum p_i \log p_i$$

p_i = probability of getting the different categories of the node.

$$\text{Binary} = -p_+ \log_2 p_+ - p_- \log p_-$$

$$\Leftrightarrow -p_0 \log_2 p_0 - p_1 \log p_1$$

$$\text{Multiclass} = -p c_1 \log_2 c_1 - p c_2 \log_2 c_2 - p c_3 \log_2 c_3$$



$$H(S)_{C_1} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

(0.5 for each term; Y, N)

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

$$= 1$$

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

$$= 0$$

C_3 — 2Y/3N

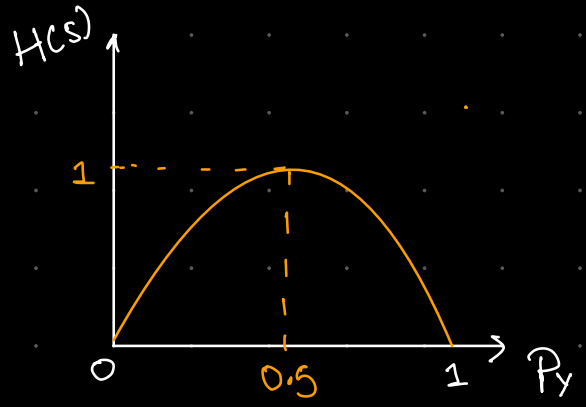
0 × + 1 ×

$$H(S)_{C_3} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.97$$

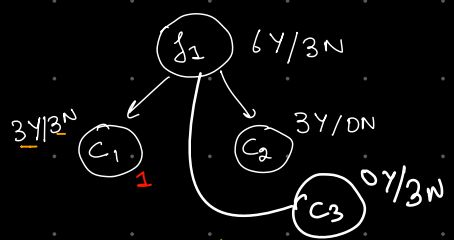
leaf node = Pure Split ⇒ 0 entropy

equal node = Highest entropy = 1



⇒ more entropy

⇒ equalish class distribution



3Y/0N

5Y/5N

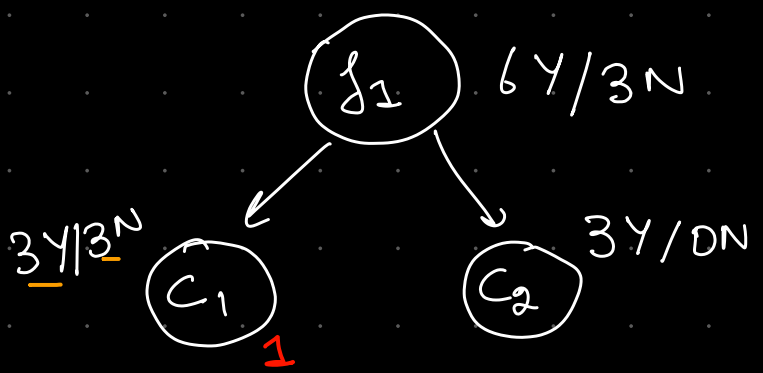
⇒ J1 impure, then we may split again.

Gini Index (G.I.)

Another purity check

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

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



$$G.I.C_1 = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right]$$

(Y, N)

$$= 0.5$$

$$G.I.C_2 = 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right]$$

(Y, N)

$$= 0$$

C_4

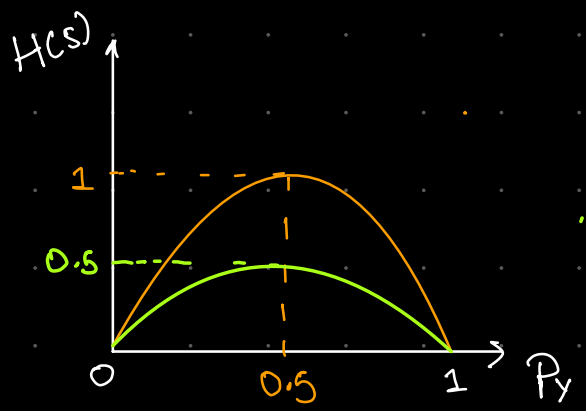
4Y/8N

8Y/2N

$$G.I_{C_4} = 4/9 = 0.44$$

$$= 0.32$$

87/4N = 0.44



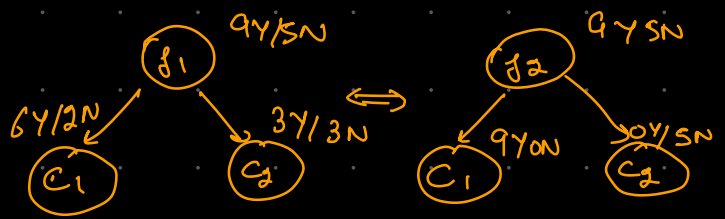
G.I. ⇒ larger dataset

Entropy ⇒ Smaller dataset [log is expensive]

Information Gain

Why outlook and not wind, humidity etc?

f_1 f_2



I.G. helps in finalizing which node to split on.

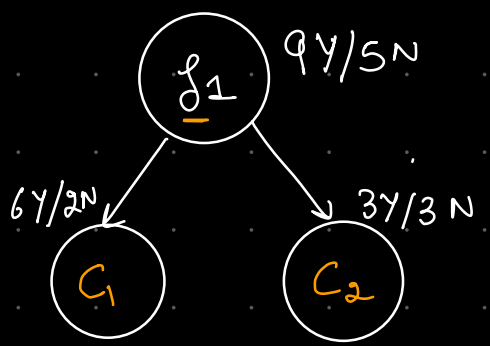
It calculates the information which f will gain from f_1 & f_2 and divide on variable which give me more information.

I. G. (feature):

$$H(\text{root}) - \sum_{v \in \text{val}} \frac{|S_v|}{|S|} H(S_v)$$

($H(S_v)$: entropy of child nodes)

entropy but you can use G.I.



$$H(f_1) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$$

$$= 0.94$$

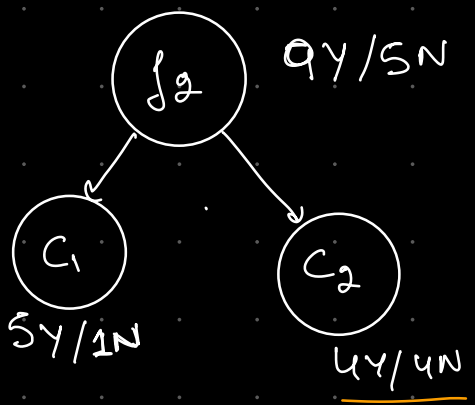
$$\frac{|S_v|}{|S|} = \frac{\text{total datapoint in Node}}{\text{total in Parent Node}}$$

For nodes $= -\left[\frac{8}{14} * \left[-\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}\right] + \right.$ (6Y/2N → C_1)

$$\left. + \frac{6}{14}(1) \right]$$ (3Y/3N → C_2)

$$I.G.(S, f_1) = 0.045$$

★ Above is for feature f_1



$I.G.(S, f_2)$

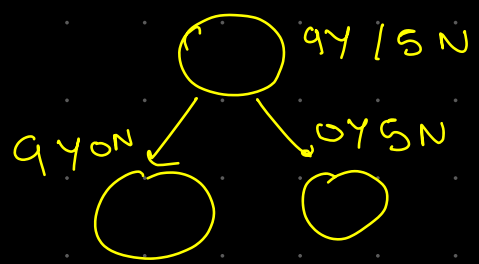
$H(\text{root} \mid f_2) = 0.94$

$$\Rightarrow \underset{f_2}{0.94} - \left[\frac{6}{14} * \underset{C_1}{\left[\frac{5}{6}\log\frac{5}{6} + \frac{1}{6}\log\frac{1}{6}\right]} + \frac{8}{14}\underset{C_2}{[1]}\right]$$

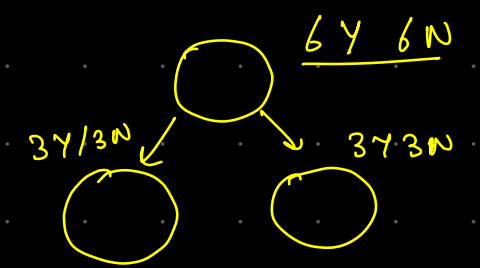
$$\Rightarrow 0.59$$

I.G. $(S, f_1) = 0.45$ I.G. $(S, f_2) = 0.59$

$\Rightarrow$ f_2 split is better than f_1 split so we are gonna first split using f_2



$$\Rightarrow 0.94 - \left[\frac{9}{14} \cdot 0 + \frac{5}{14} \cdot 0\right]$$



$$\Rightarrow 1 - \left[\frac{6}{12} \cdot 1 - \frac{6}{12} \cdot 1\right]$$

$\Rightarrow 0$

f_1 f_2

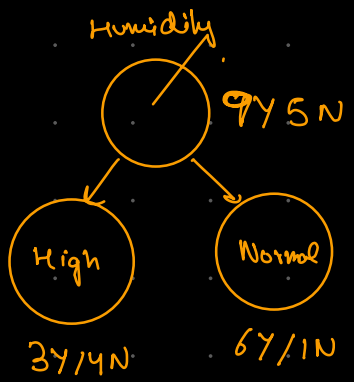
f_2 f_1

f_3 f_3

f_4 f_4

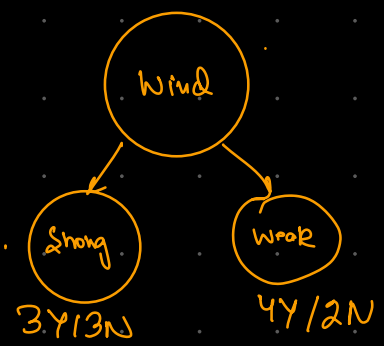
$(4!)$

Outlook	Temperature	Humidity	Wind	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No



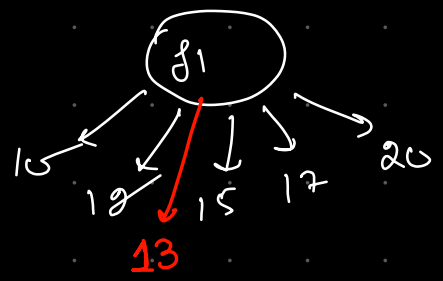
I.G. (Humidity)

I.G. (Wind)



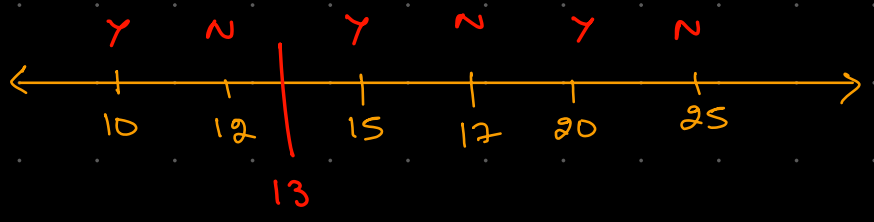
How to deal with numeric features?

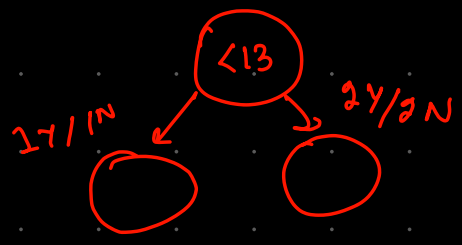
f1	O/P
10	1
12	0
15	1
17	0
20	0
25	1

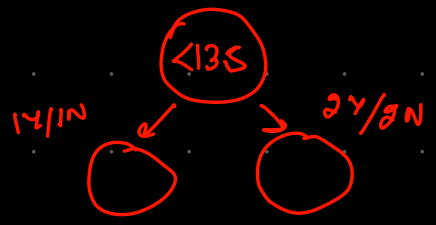


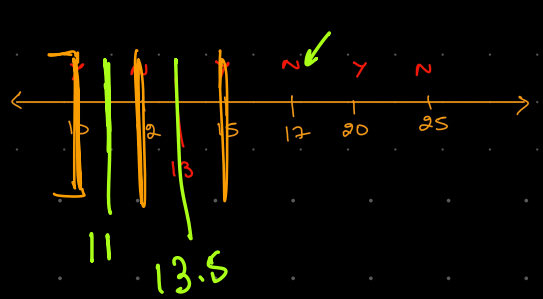
test data 13

13.5



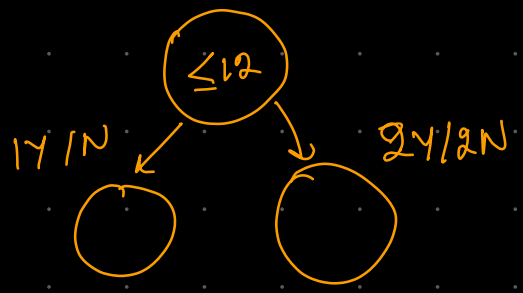


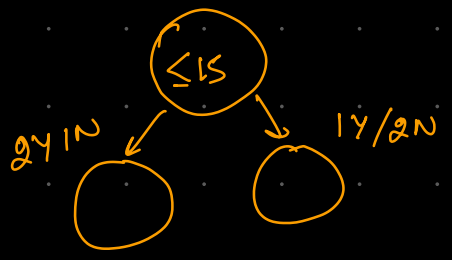



Y
N
N
Y
N
10
12
15
17
20
25
13
11
13.5


≤10
<11
1Y/0N
12 – 25
2Y/3N

I.G.


≤12
1Y/1N
2Y/2N


≤15
2Y/1N
1Y/2N

Decision Tree Regressor

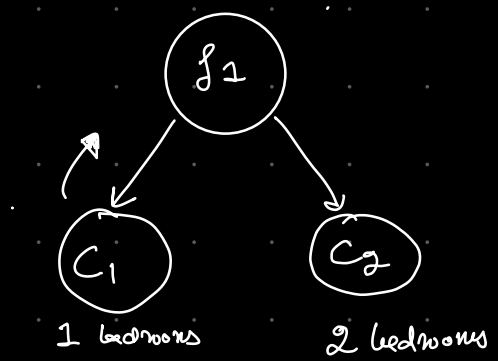
↓

Continous O/P feature

Price of house

(Bedrooms)

f_1	f_2	Price
1		20
1		24
1		28
2		14
2		18
2		20



MSE / MAE