Decision Trees

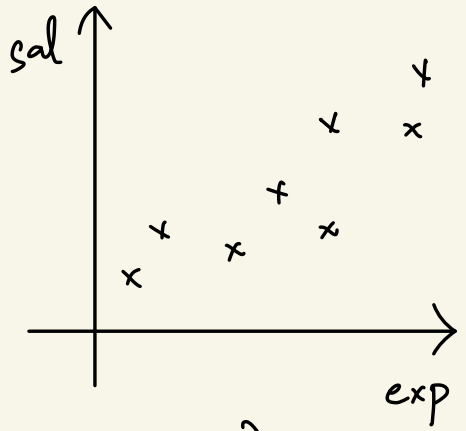
About Parametric and non-parametric algorithms

Parametric Algorithms :-

- ⇒ makes assumptions
- ⇒ fixed parameters
- ⇒ requires less data to train

eg: consider the following data:

experience	Salary
3	5
8	12
10	14
15	25



Linear data ⇒ $y = f(x)$

$Salary = f(experience)$

if done using linear regression:

$y = mx + b$ → assuming its a line

↓ ↓ fixed parameters

Non-parametric algorithms:-

→ does not assume
→ needs more data to train
→ chances of over fitting
    → gives importance to all choices
    → might consider noise also.

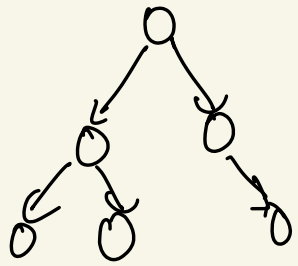
eg: Decision Tree

About Classification

→ dividing dataset into categories
→ in supervised → labels given.

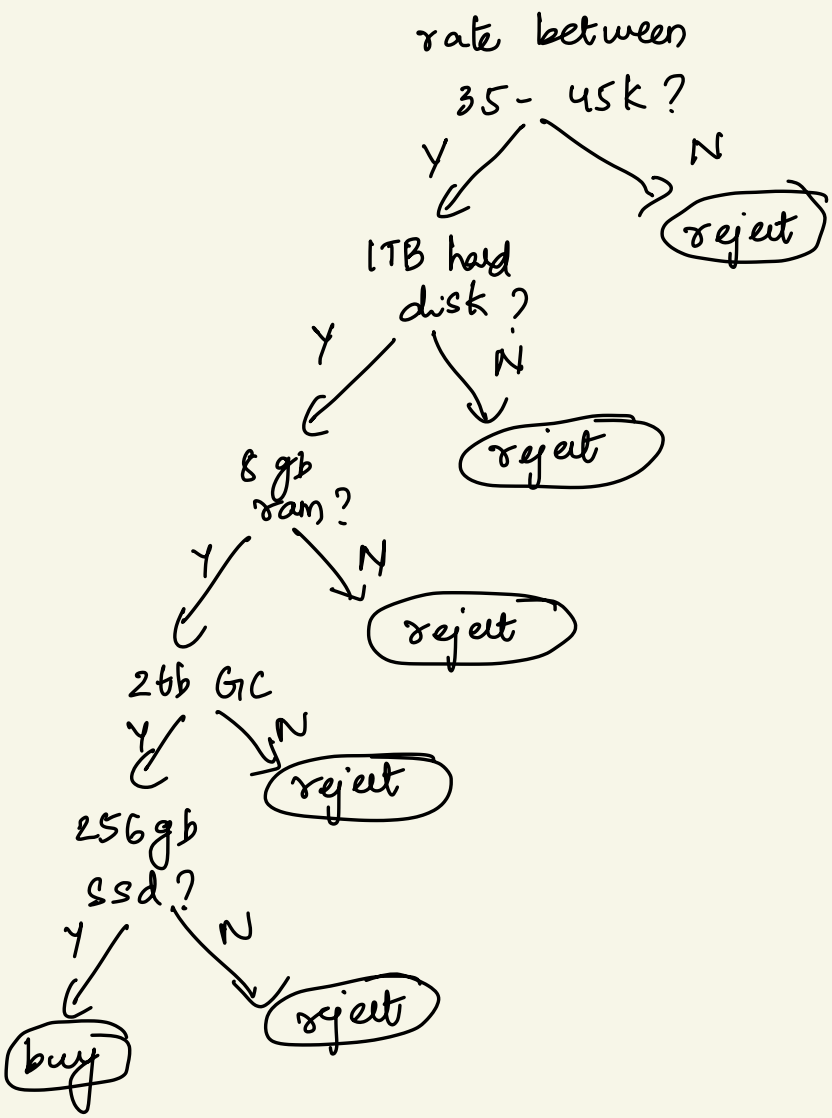
Decision Trees Algorithm

Transforming data into a tree structure.



→ root node
    → no incoming
    → outgoing
→ internal node
    → incoming & outgoing
→ leaf node - incoming, no outgoing, class label predicted here.

Give example of laptop buying :-



→ it is hierarchical

→ recursive splits used

→ each decision node m implements a function $f_m(x)$ with discrete outcomes labelling the branches

→ process starts at root

→ repeated recursively until leaf node is hit

→ value at leaf node is the output

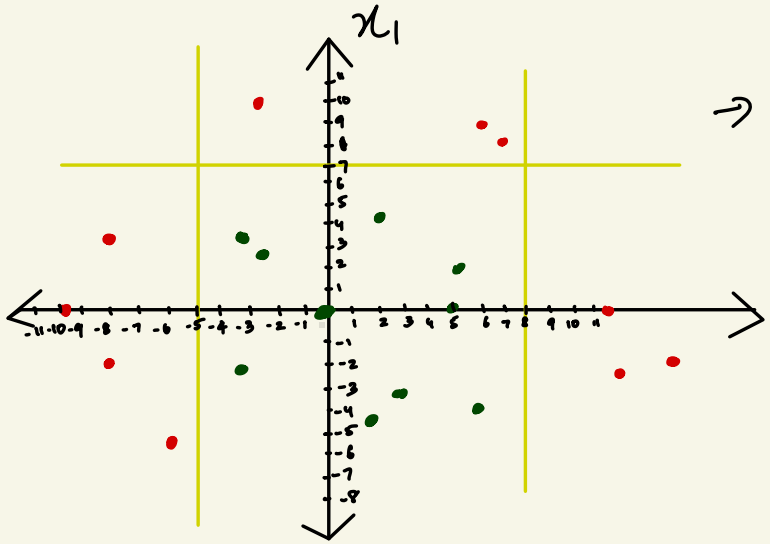
→ tree structure is not fixed

→ tree grows during learning depending on complexity

→

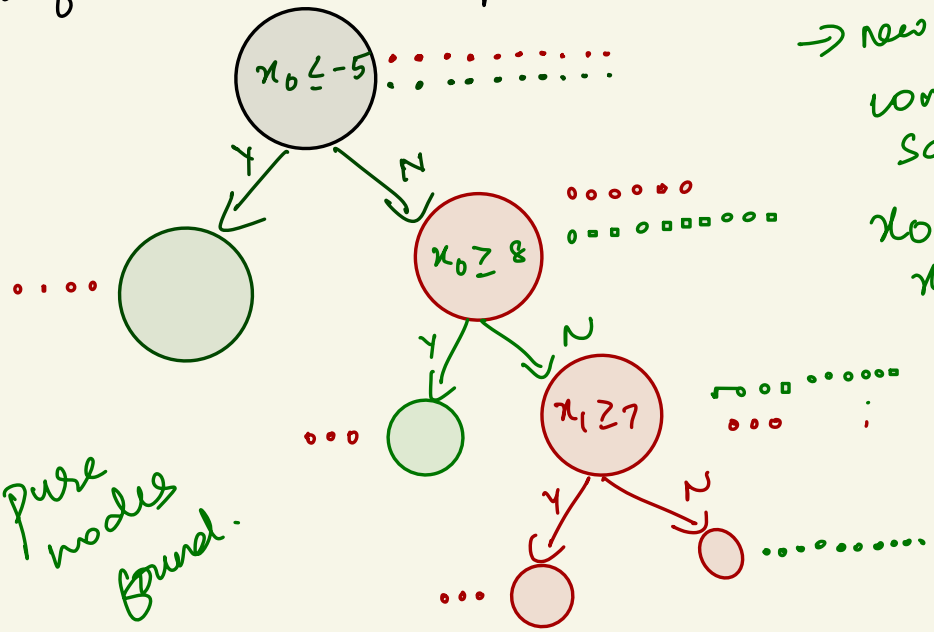
Lets consider a random dataset:

→ contains 2 features → x_0, x_1



→ Not linearly separable

→ recursively splits a dataset to obtain leaf nodes that's pure.



→ new data comes, say (15, 7)

$x_0 = 15$

$x_1 = 7$

what if we donot get pure nodes?

Impurity is calculated using entropy.

⇒ entropy measures impurity of collection of samples

⇒ depends on the proportion of classes

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

So if we have 2 classes ⇒ positive & negative:

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

P_+ ↓ proportion of +ve samples

P_- ↓ proportion of -ve samples

proportion means no of +ve/-ve samples out of total samples

Examples:

$$E(14+, 0-) = -\frac{14}{14}\log_2\left(\frac{14}{14}\right) - 0\log_2(0)$$

$$= -1\log_2(1) - 0\log_2(0)$$

$$= -1 \times 0 - 0 \Rightarrow 0$$

$$E(7+, 7-) = -\frac{7}{14}\log_2\left(\frac{7}{14}\right) -$$

$$\frac{7}{14}\log_2\left(\frac{7}{14}\right)$$

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

$$= -\frac{1}{2} \times -1 - \frac{1}{2} \times -1$$

$$= \frac{1}{2} + \frac{1}{2} = 1$$

Example dataset :-

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

If we want to find entropy of this dataset :

No. of +ve = 9

No. of -ve = 5

total rows = 14

So with eqn,

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

$$\approx 0.9 \Rightarrow$$

→ impurity measure is done between 0 & 1

→ 0.9 → highly impure.

Example - 1

Decision Tree algorithm - Entropy & Information Gain

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Step 1: Decide root node.

Method: -

→ find information gain of all attributes

→ attribute with highest IG will become root node.

→ Calculate impurity

Impurity measure → using Entropy.

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Here → $- p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$

Step 1: Entropy of entire dataset

$S[9+, 5-] =$

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

$\Rightarrow$ 0.94

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute : Outlook

Possible values - Sunny, overcast, rain

$S_{sunny} = [2+, 3-]$

$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \Rightarrow 0.97$

$S_{overcast} = [4+, 0-] \Rightarrow 0$

$S_{rain} = [3+, 2-] \Rightarrow -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \Rightarrow 0.97$

After entropy calculation → Calculate IG

$$Gain(S, Outlook) = Entropy(S) - \sum_{v \in (Sunny, rain, overcast)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\Rightarrow Entropy(S) - \frac{5}{14} Entropy(S_{sunny}) - \frac{4}{14} Entropy(S_{overcast}) - \frac{5}{14} Entropy(S_{rain})$$

$$\Rightarrow 0.94 - \frac{5}{14} \times 0.97 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.97$$

$$\Rightarrow 0.2464$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute :- Temp

Possible values - Hot, Mild, Cool

$S_{Hot} = [2+, 2-]$

$\Rightarrow 1$

$S_{Mild} = [4+, 2-]$

$\Rightarrow -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.91$

$S_{cool} = [3+, 1-] \Rightarrow -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.81$

$$Gain(S, Temp) = Entropy(S) - \sum_{v \in (hot, cool, mild)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= Entropy(S) - \frac{4}{14}E(S_{hot}) - \frac{6}{14}E(S_{mild}) - \frac{4}{14}E(S_{cool})$$

$$= 0.94 - \frac{4}{14}\cdot 1.0 - \frac{6}{14}\cdot 0.91 - \frac{4}{14}\cdot 0.81 = 0.028$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute: Humidity

Possible values = high, normal

$S_{high} = [3+, 4-]$

$E(S_{high}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}$

$= 0.98$

$S_{normal} = [6+, 1-]$

$E(S_{normal}) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.59$

$$Gain(S, humidity) = E(S) - \sum_{v \in (high, normal)} \frac{|S_v|}{|S|} E(S_v)$$

$= E(S) - \frac{7}{14} E(S_{high}) - \frac{7}{14} E(S_{normal})$

$= 0.94 - \frac{7}{14} \cdot 0.98 - \frac{7}{14} \cdot 0.59 = 0.15$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Attribute : wind

Possible values - weak, strong

$S_{strong} = [3+, 3-]$

$E(S_{strong}) = 1$

$S_{weak} = [6+, 2-]$

$$E(S_{weak}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.81$$

$$Gain(S, wind) = E(S) - \sum_{(v \in weak, strong)} \frac{|S_v|}{|S|} E(S_v)$$

$$= E(S) - \frac{6}{14}E(S_{strong}) - \frac{8}{14}E(S_{weak})$$

$$= 0.94 - \frac{6}{14} \times 1 - \frac{8}{14} \times 0.81 = 0.047$$

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

Comparing IGs:

$Gain(S, Outlook) = 0.2464$

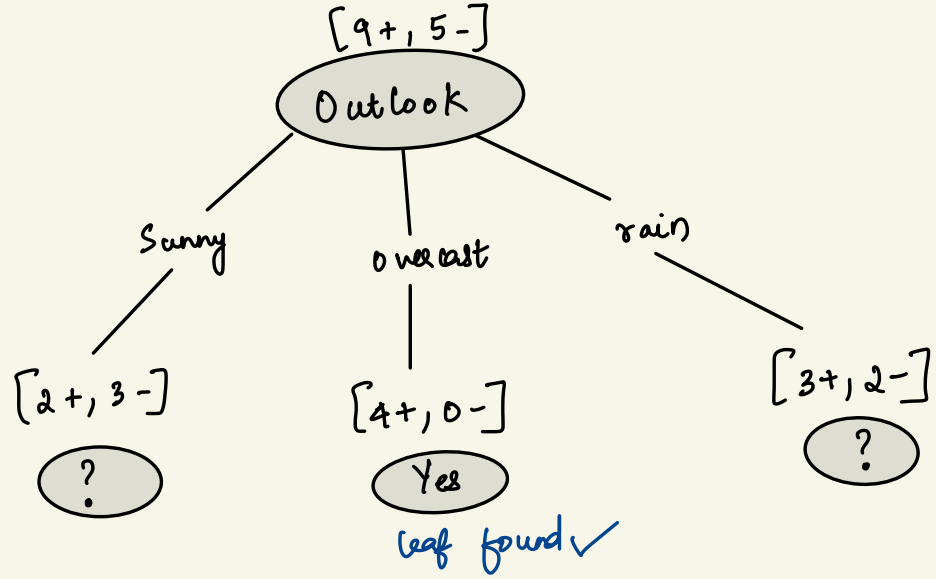
$Gain(S, temp) = 0.028$

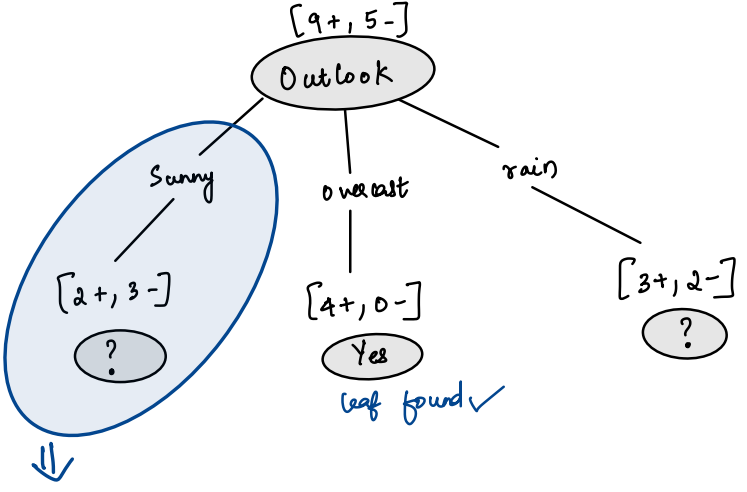
$Gain(S, humid) = 0.15$

$Gain(S, wind) = 0.047$

Humidity

high normal





take dataset with sunny values

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

→

Day	Temp	Humidity	wind	Play
D1	Hot	High	weak	No
D2	Hot	High	Strong	No
D8	Mild	High	weak	No
D9	Cool	normal	weak	Yes
D11	Mild	normal	Strong	Yes

Day	Temp	Humidity	wind	Play
D1	Hot	High	weak	No
D2	Hot	High	strong	No
D8	Mild	High	weak	No
D9	Cool	normal	weak	Yes
D11	Mild	normal	strong	Yes

Repeat steps here again:

Consider this as whole dataset.

$$E(S) = E(S_{sunny}) = [2+, 3-]$$

$$= -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$= 0.97$$

Consider attributes:

Attribute :- Temp → Possible values - hot, mild, cool

$S_{hot} \rightarrow [0+, 2-] \Rightarrow E(S_{hot}) = 0$

$S_{mild} \rightarrow [1+, 1-] \Rightarrow E(S_{mild}) = 1$

$S_{cool} \rightarrow [1+, 0-] \Rightarrow E(S_{cool}) = 0$

} Gain = 0.57

Day	Temp	Humidity	wind	Play
D1	Hot	High	weak	No
D2	Hot	High	Strong	No
D8	Mild	High	weak	No
D9	Cool	normal	weak	Yes
D11	Mild	normal	strong	Yes

Attributes – Humidity

Values – High, normal

$S_{high} = [0+, 3-] \Rightarrow E(S_{high}) = 0$

$S_{normal} = [2+, 0-] \Rightarrow E(S_{normal}) = 0$

$$Gain(S_{sunny}, humidity) = E(S) - \sum_{v \in (high, normal)} \frac{|S_v|}{|S|} E(S_v)$$

$$Gain = 0.97 - 0 \Rightarrow 0.97$$

Day	Temp	Humidity	wind	Play
D1	Hot	High	weak	No
D2	Hot	High	strong	No
D8	Mild	High	weak	No
D9	Cool	normal	weak	Yes
D11	Mild	normal	strong	Yes

Attribute = wind

values ⇒ strong, weak

$S_{strong} = [1+, 1-] = E(S_{strong}) = 1$

$S_{weak} = [1+, 2-] = E(S_{weak})$

$\Rightarrow -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.91$

$Gain = 0.97 - \frac{2}{5} \cdot 1 - \frac{3}{5} \cdot 0.91 = 0.019$

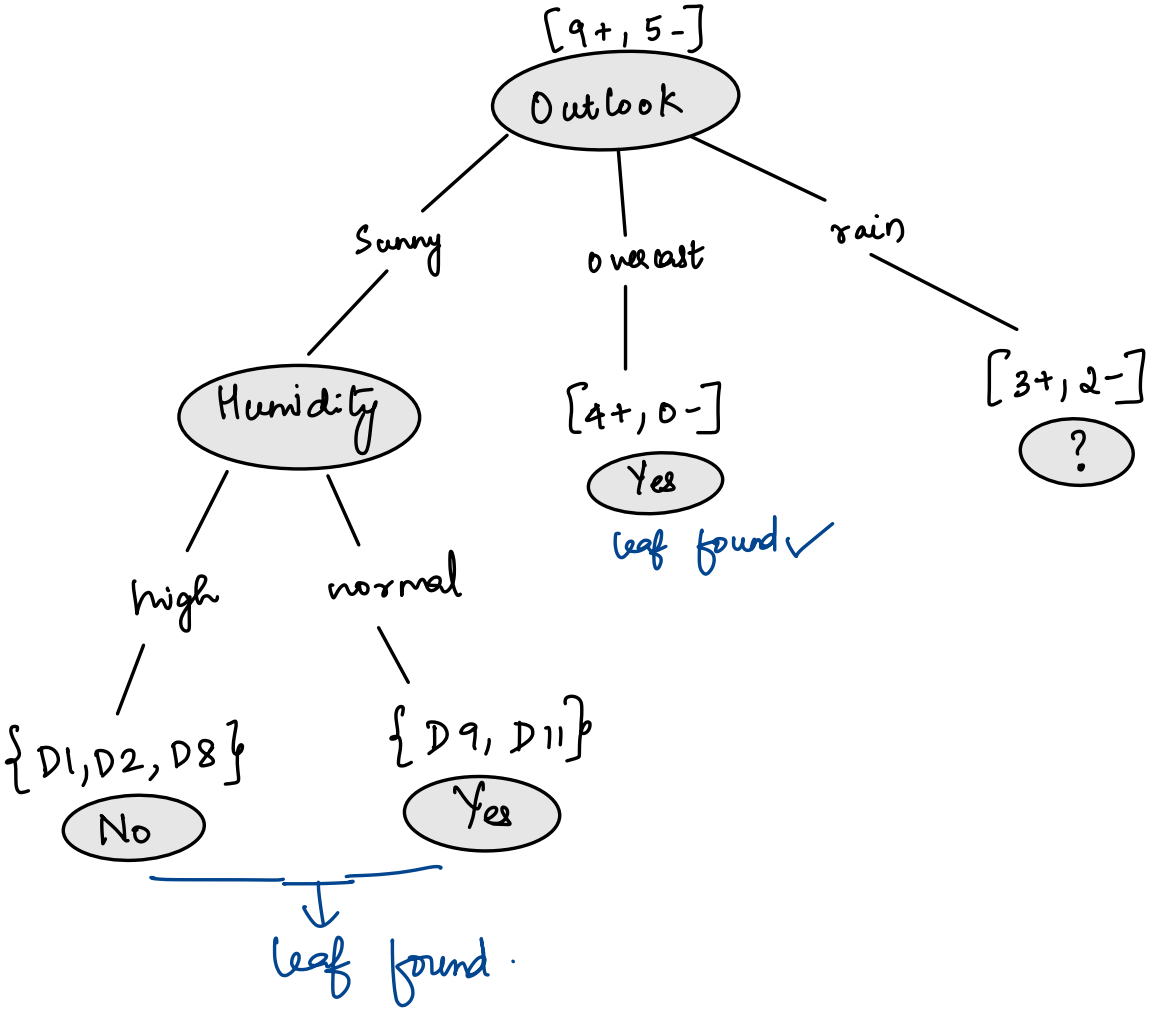
Day	Temp	Humidity	wind	Play
D1	Hot	High	weak	No
D2	Hot	High	strong	No
D8	Mild	High	weak	No
D9	Cool	normal	weak	Yes
D11	Mild	normal	strong	Yes

Compare gains:-

$Gain(S_{sunny}, temp) = 0.57$

$Gain(S_{sunny}, humid) = 0.97$

$Gain(S_{sunny}, wind) = 0.019$


[9+, 5-]
Outlook
Sunny
overcast
rain
Humidity
[4+, 0-]
Yes
leaf found ✓
[3+, 2-]
?
high
normal
{D1, D2, D8}
No
{D9, D11}
Yes
leaf found.

Day	Outlook	Temp	Humidity	Wind	Play Tennis
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

D4, D5, D6, D10, D14

rain rows

Day	Temp	Humidity	Wind	Play
D4	Mild	high	weak	Yes
D5	Cool	normal	weak	Yes
D6	Cool	normal	Stoong	No
D10	mild	normal	weak	Yes
D14	mild	high	Stoong	No

Day	Temp	Humidity	Wind	Play
D4	Mild	high	weak	Yes
D5	Cool	normal	weak	Yes
D6	Cool	normal	Stoong	No
D10	mild	normal	weak	Yes
D14	mild	high	Stoong	No

Repeat steps :-

Consider this as dataset :-

$S_{rain} \Rightarrow [3+, 2-]$

$E(S_{rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$

$= \underline{0.97}$

Attributes : Temp

$S_{hot} = [0+, 0-] \Rightarrow E(S_{hot}) = \underline{0}$

$S_{mild} = [2+, 1-] \Rightarrow E(S_{mild}) = \underline{0.91}$

$S_{cool} = [1+, 1-] = E(S_{cool}) = \underline{1}$

$gain = \underline{0.019}$

Day	Temp	Humidity	Wind	Play
D4	Mild	high	weak	Yes
D5	Cool	normal	weak	Yes
D6	Cool	normal	Strong	No
D10	mild	normal	weak	Yes
D14	mild	high	Strong	No

Attribute : Humidity

$S_{high} = [1+, 1-] = E(S_{high}) = 1$

$S_{normal} = [2+, 1-] = E(S_{normal}) = 0.91$

$Gain(S_{rain}, humid) = 0.019$

Attributes: Wind

$S_{strong} = [0+, 2-] \Rightarrow E(S_{strong}) = 0$

$S_{weak} = [3+, 0-] \Rightarrow E(S_{weak}) = 0$

$Gain[S_{rain}, wind) = 0.97$

Day	Temp	Humidity	Wind	Play
D4	Mild	high	weak	Yes
D5	Cool	normal	weak	Yes
D6	Cool	normal	Strong	No
D10	mild	normal	weak	Yes
D14	mild	high	Strong	No

$Gain(S_{rain}, temp) = 0.019$

$Gain(S_{rain}, humid) = 0.019$

$Gain(S_{rain}, wind) = 0.97$

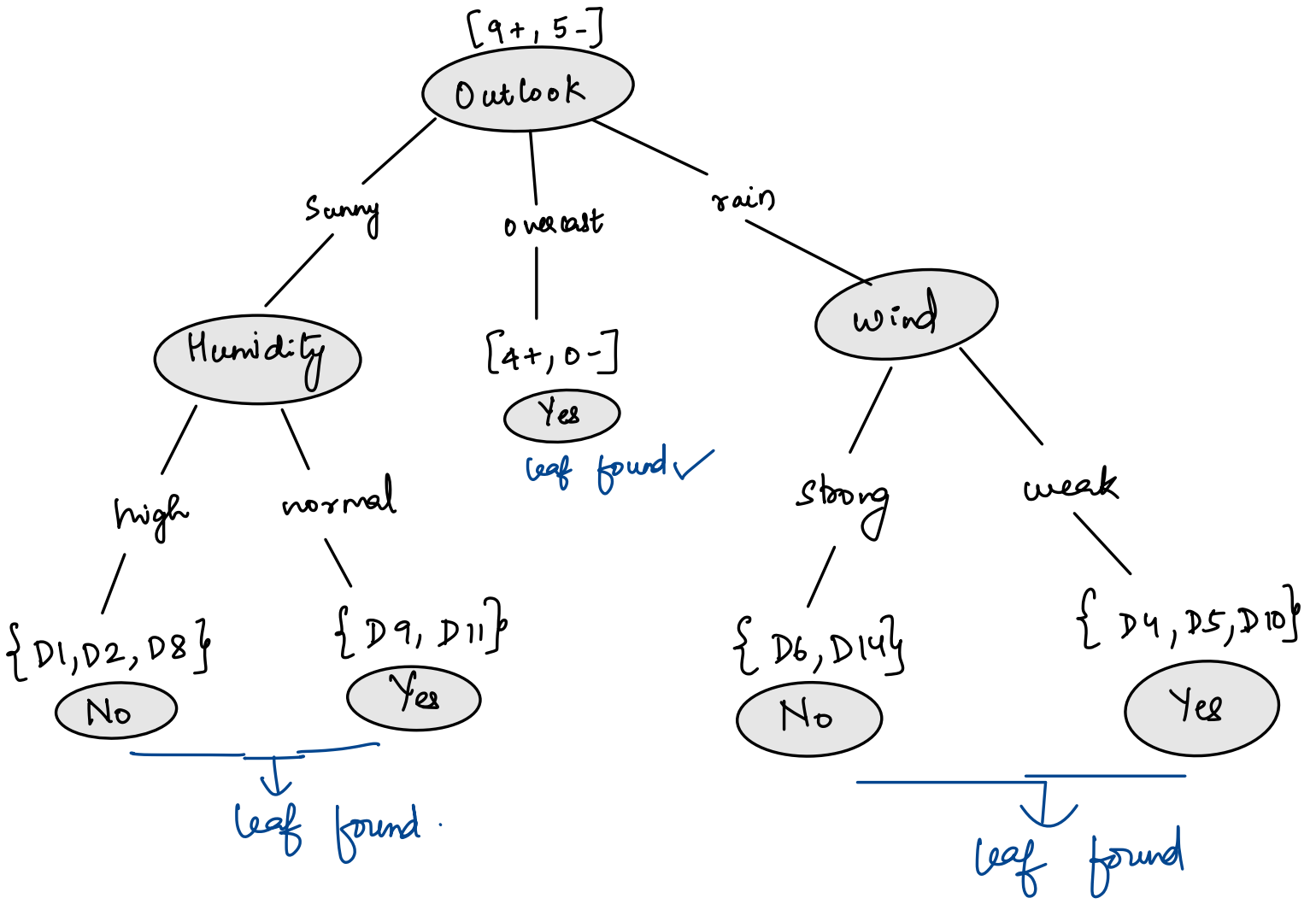
⇓

wind ⇒ weak , Strong

↓ ↓

3 Yes 0 Yes

0 No 2 No


[9+, 5-]
Outlook
Sunny
overcast
rain
Humidity
[4+, 0-]
Yes
leaf found ✓
wind
high
normal
strong
weak
{D1, D2, D8}
No
{D9, D11}
Yes
{D6, D14}
No
{D4, D5, D10}
Yes
leaf found.
leaf found

Gini Impurity Index

→ Measure of diversity in a dataset

Example.

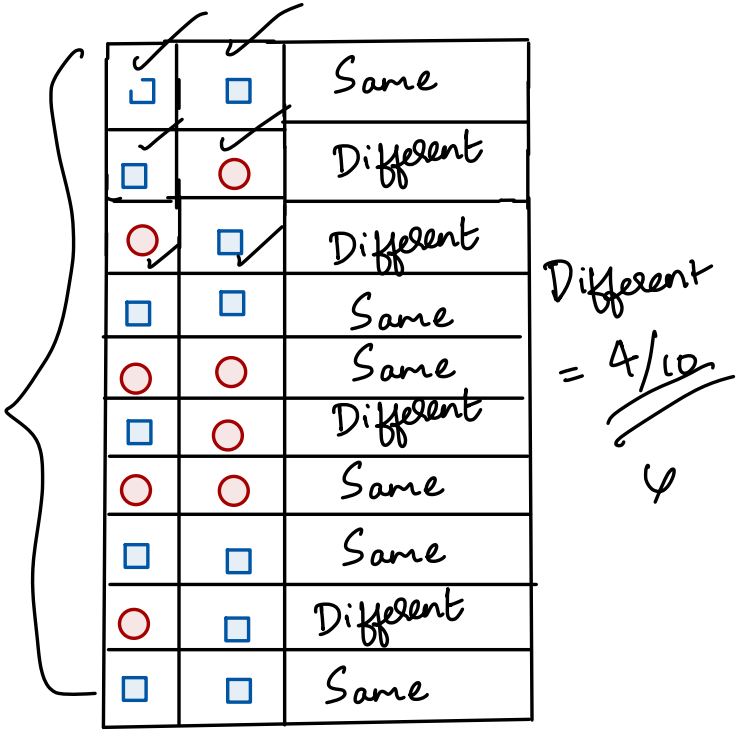
Dataset - 1

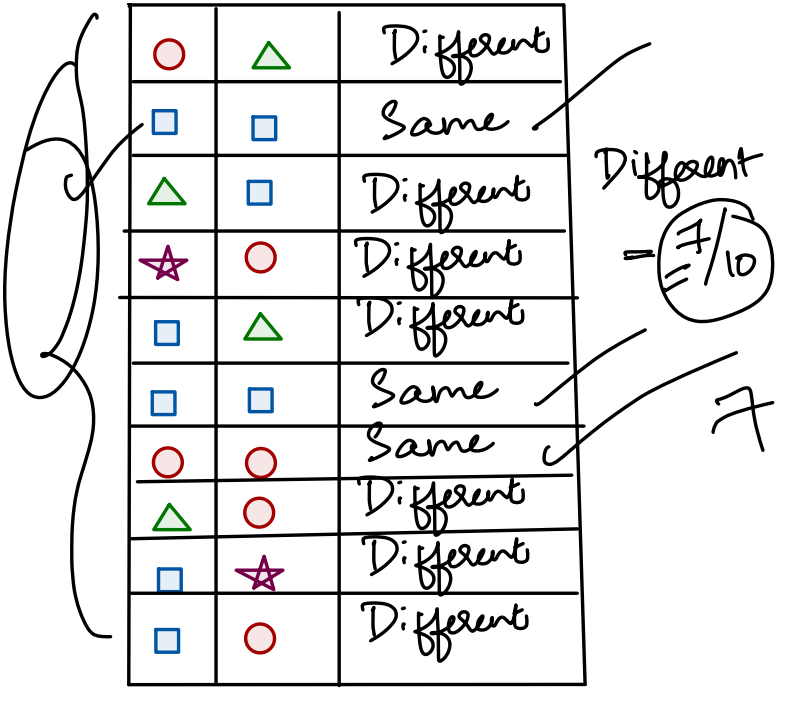
Dataset - 2

} which is more diverse?

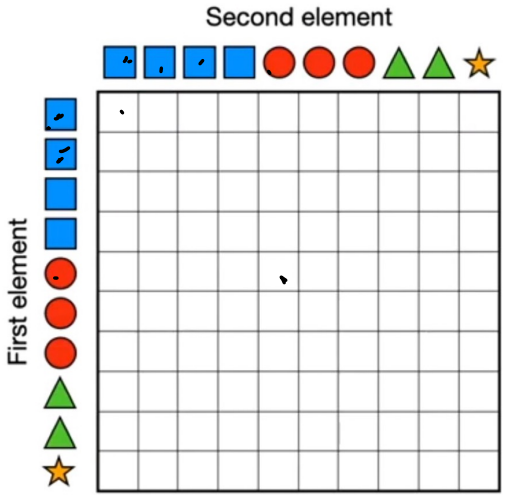
Finding impurity :-

→ picking 2 random elements & check if they are same



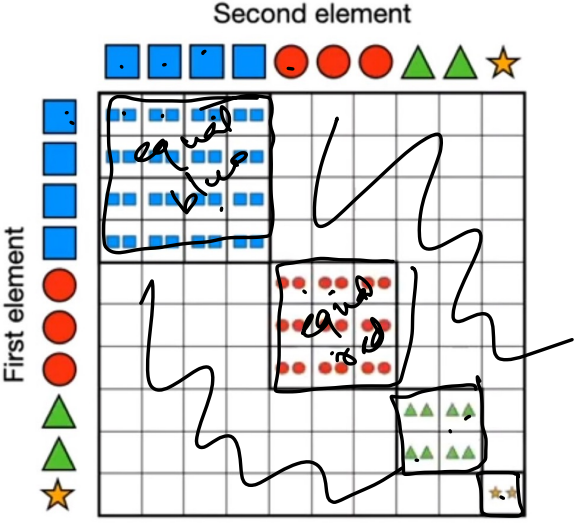


Gini Index = Probability of picking 2 distinct elements.

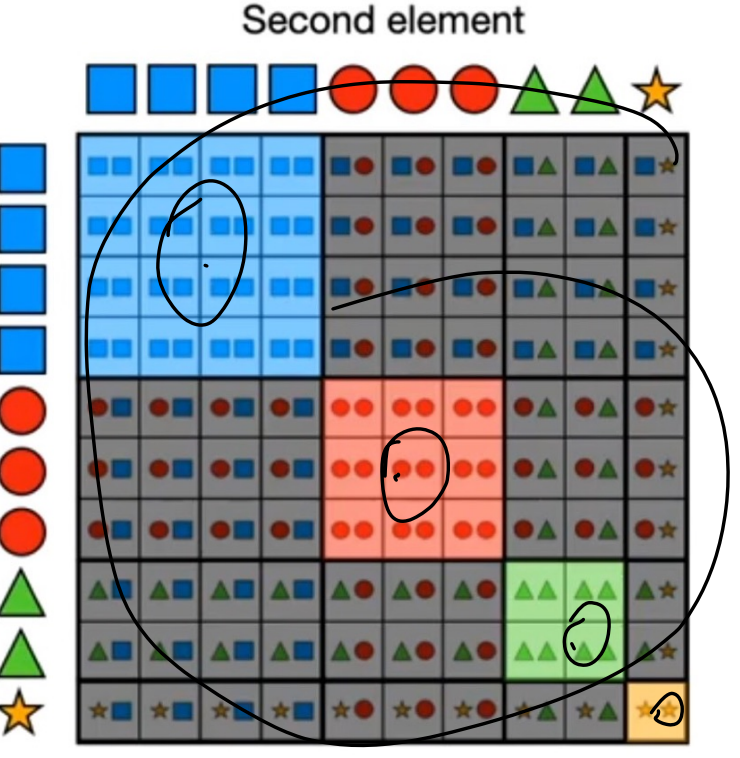


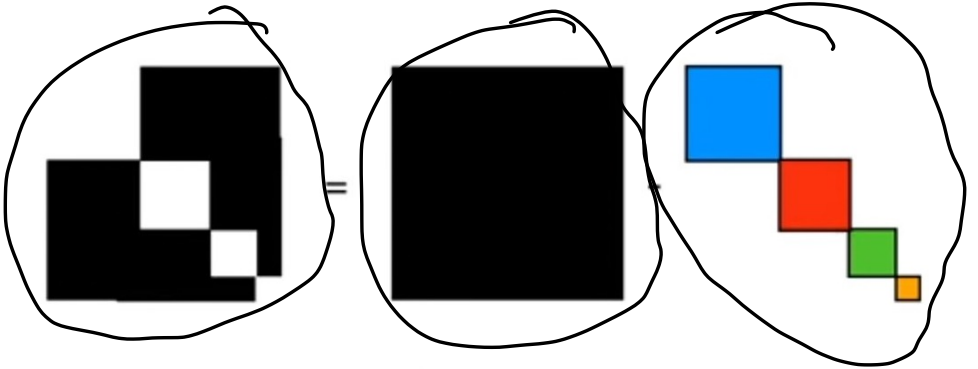
Considering entire data, all possibilities

10 elements
//
2 elements



Equal Possibilities

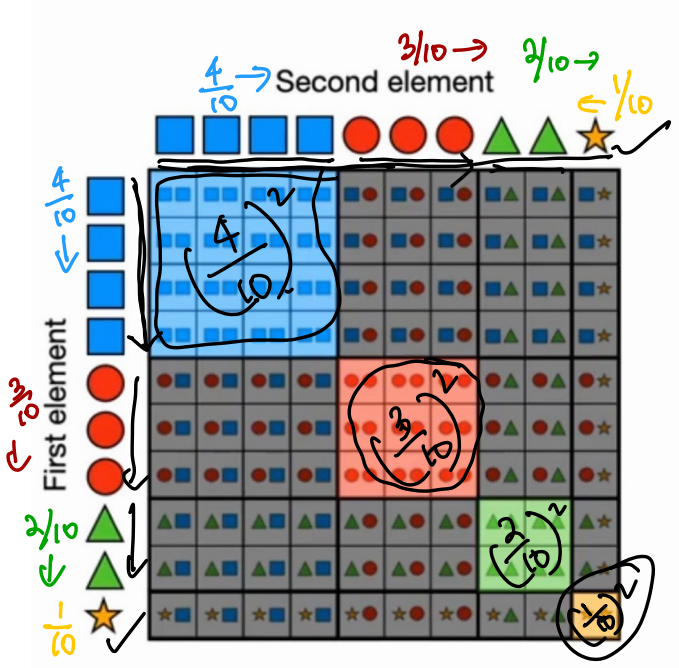


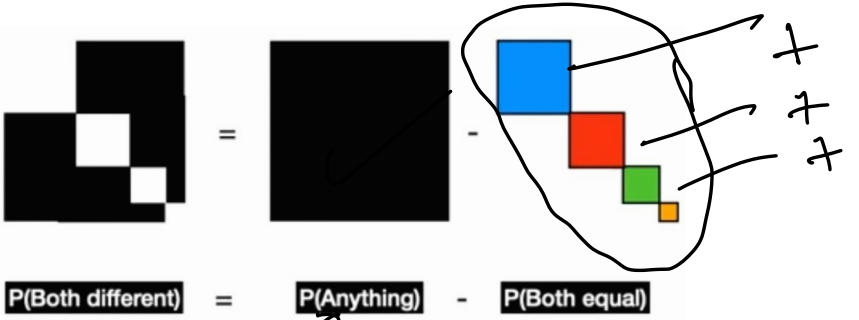


P(Both different) = P(Anything) - P(Both equal)

P(anything) = 1

⇒ 1 − (P(blue) + P(red) + P(green) + P(yellow))



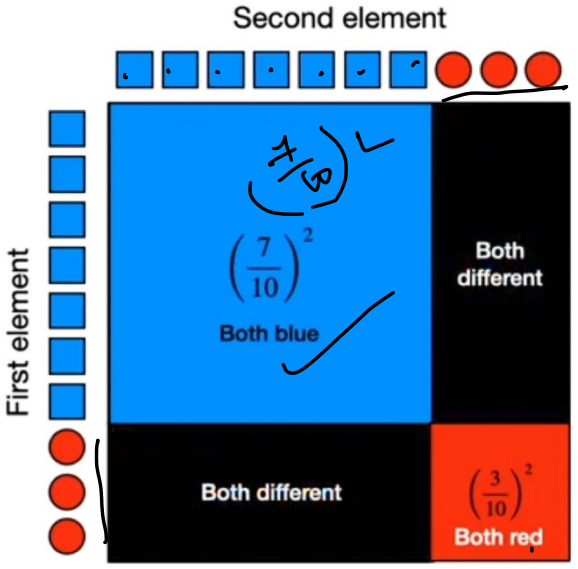


P(Both different) = P(Anything) - P(Both equal)

$= 1 - \left(\left(\frac{4}{10}\right)^2 + \left(\frac{3}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2\right)$

$= 1 - (0.16 + 0.09 + 0.04 + 0.01)$

$P(\text{diff}) = 1 - 0.3 = 0.7$



P(both different)

= P[Anything) − (P(both blue) + P(both red))

= $1 - \left(\left(\frac{7}{10}\right)^2 + \left(\frac{3}{10}\right)^2\right)$

= $1 - [0.49 + 0.09]$

= $1 - 0.58 = 0.42$

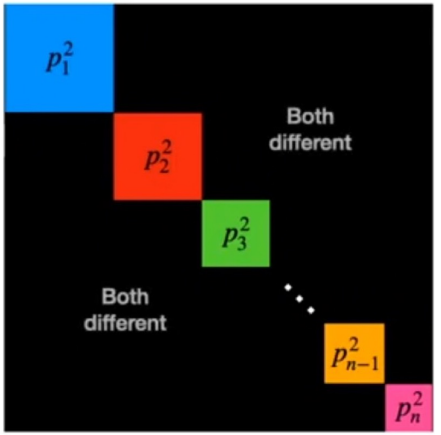
General formula

→ n classes

→ proportions → $p_1, p_2, p_3 \cdots p_n$

→ Gini impurity index = 1 − (sum of squares of proportions)

$$1 - (p_1^2 + p_2^2 + \cdots p_n^2)$$



Some more examples

$\rightarrow$ 0.42

$\rightarrow$ 0.7

$\rightarrow$ 0 (1 (anything) $-\left(\frac{10}{10}\right)^2$)

$= 1 - 1^2 = 0$

$\rightarrow$ all are different

$\Rightarrow$ every one will have $\left(\frac{1}{10}\right)^2$ area

$= 1 - \underbrace{0.1^2 + 0.1^2 + \dots 0.1^2}_{10 \text{ times}}$

$= 0.9$

Example Dataset

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

[10]

Step 1 ⇒ Compute Gini Index for entire dataset

possible values of target

⇒ cinema, tennis, stay in, shopping (4) (total 10)

→ Cinema ⇒ 6/10

→ tennis ⇒ 2/10

→ stay in ⇒ 1/10

→ shopping ⇒ 1/10

$$\text{Gini}(S) = 1 - \left(\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2 \right) = 0.58$$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

→ to get root node ⇒ compute Gini of all attributes

(1) Weather → Sunny, windy, rainy

Sunny (3) ⇒ 1 cinema, 2 tennis,

$$Gini(S_{sunny}) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right]$$

$$= 1 - [0.111 + 0.444] \cong 0.44$$

Gini (S_{windy}) :-

total = 4 ⇒ 3 cinema, 1 shopping

$$\Rightarrow 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right] = 1 - [0.5625 + 0.0625] = 0.375$$

Gini (S_{rainy}) :-

total = 3 ⇒ 2 cinema, 1 stay in

$$\Rightarrow 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = 0.44$$

Final Gini: weighted avg

$$0.44 \times \left(\frac{3}{10}\right) + 0.375 \times \left(\frac{4}{10}\right) + 0.44 \times \left(\frac{3}{10}\right) \Rightarrow 0.414$$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

② Parents ⇒ Yes, no

gini (S_{yes}) ⇒ total 5

⇒ 5 cinema targets

⇒ $1 - \left[\frac{5}{5}\right]^2 = 1 - 1^2 = 0$

gini (S_{no}) ⇒ 5 total

⇒ 2 tennis, 1 stay in, 1 cinema, 1 shopping

$$\Rightarrow 1 - \left[\left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right]$$

$$= 1 - [0.16 + 0.04 + 0.04 + 0.04] = 0.72$$

weighted average:-

$$0 * \frac{5}{10} + 0.72 * \frac{5}{10} \Rightarrow 0.36$$

Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W2	Sunny	No	Rich	Tennis
W3	Windy	Yes	Rich	Cinema
W4	Rainy	Yes	Poor	Cinema
W5	Rainy	No	Rich	Stay In
W6	Rainy	Yes	Poor	Cinema
W7	Windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping
W9	Windy	Yes	Rich	Cinema
W10	Sunny	No	Rich	Tennis

Repeat for money attribute.

weighted average of money = 0.486

Now compare all gini indexes:-

gini (weather) = 0.414

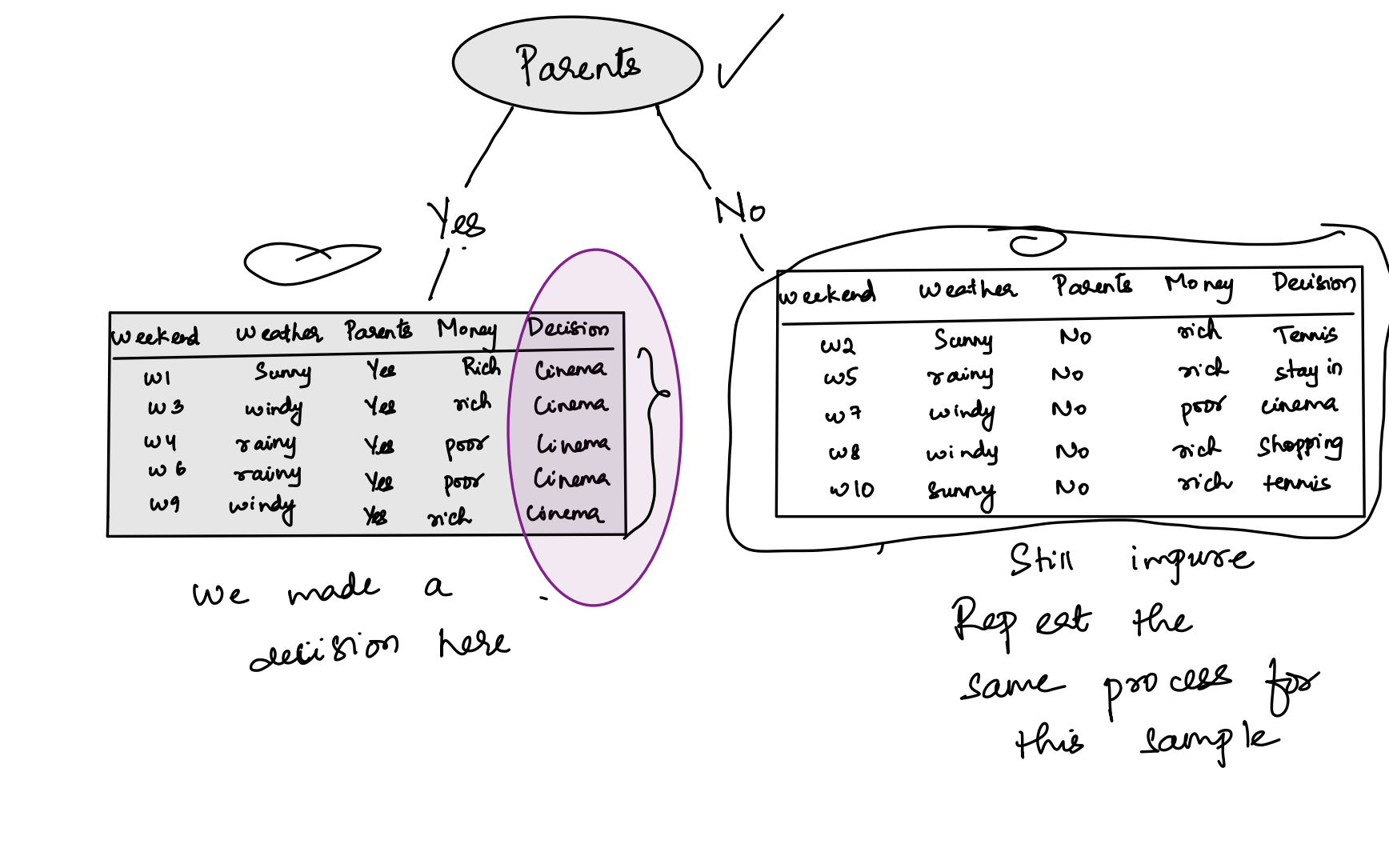
gini (parents) = 0.36

gini (money) = 0.486

Select the attribute with least gini impurity

⇒ parents selected as root node.

⇒ from root, split the dataset with possible values ⇒ (here, yes or no)


Parents
Yes
No
Weekend Weather Parents Money Decision
W1 Sunny Yes Rich Cinema
W3 windy Yes rich Cinema
W4 rainy Yes poor Cinema
W6 rainy Yes poor Cinema
W9 windy Yes rich Cinema
Weekend Weather Parents Money Decision
W2 Sunny No rich Tennis
W5 rainy No rich stay in
W7 windy No poor cinema
W8 windy No rich Shopping
W10 Sunny No rich tennis
We made a decision here
Still impure
Repeat the same process for this sample

weekend	weather	Parents	Money	Decision
w2	Sunny	No	rich	Tennis
w5	rainy	No	rich	stay in
w7	windy	No	poor	cinema
w8	windy	No	rich	shopping
w10	Sunny	No	rich	tennis

gini (S weather) ⇒ sunny, rainy, windy

gini (S sunny) ⇒ total 2 ; both tennis ⇒ 0 ✓

gini (S rainy) = 1 value ⇒ stay in ⇒ 0 ✓

gini (S windy) = 2 values ⇒ cinema, shopping

$$= 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 \right] = 0.5 \checkmark$$

$$\text{weighted average} = 0 \times \frac{2}{5} + 0 \times \frac{1}{5} + 0.5 \times \frac{2}{5} = 0.2$$

weekend	weather	Parents	Money	Decision
w2	Sunny	No	rich	Tennis
w5	rainy	No	rich	stay in
w7	windy	No	poor	cinema
w8	windy	No	rich	shopping
w10	Sunny	No	rich	tennis

gini (S_{money}) $\Rightarrow$ rich, poor

gini (S_{rich}) $\Rightarrow$ total 4 values

2 tennis, 1 stay in, 1 shopping

$$\Rightarrow \quad 1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{1}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right] = \underline{0.625}$$

gini (S_{poor}) $\Rightarrow$ 1 value $\Rightarrow$ cinema $\Rightarrow \underline{0}$

$$\text{Weighted average} = 0.625 \times \frac{4}{5} + 0 \times \frac{1}{5} = \underline{0.5}$$

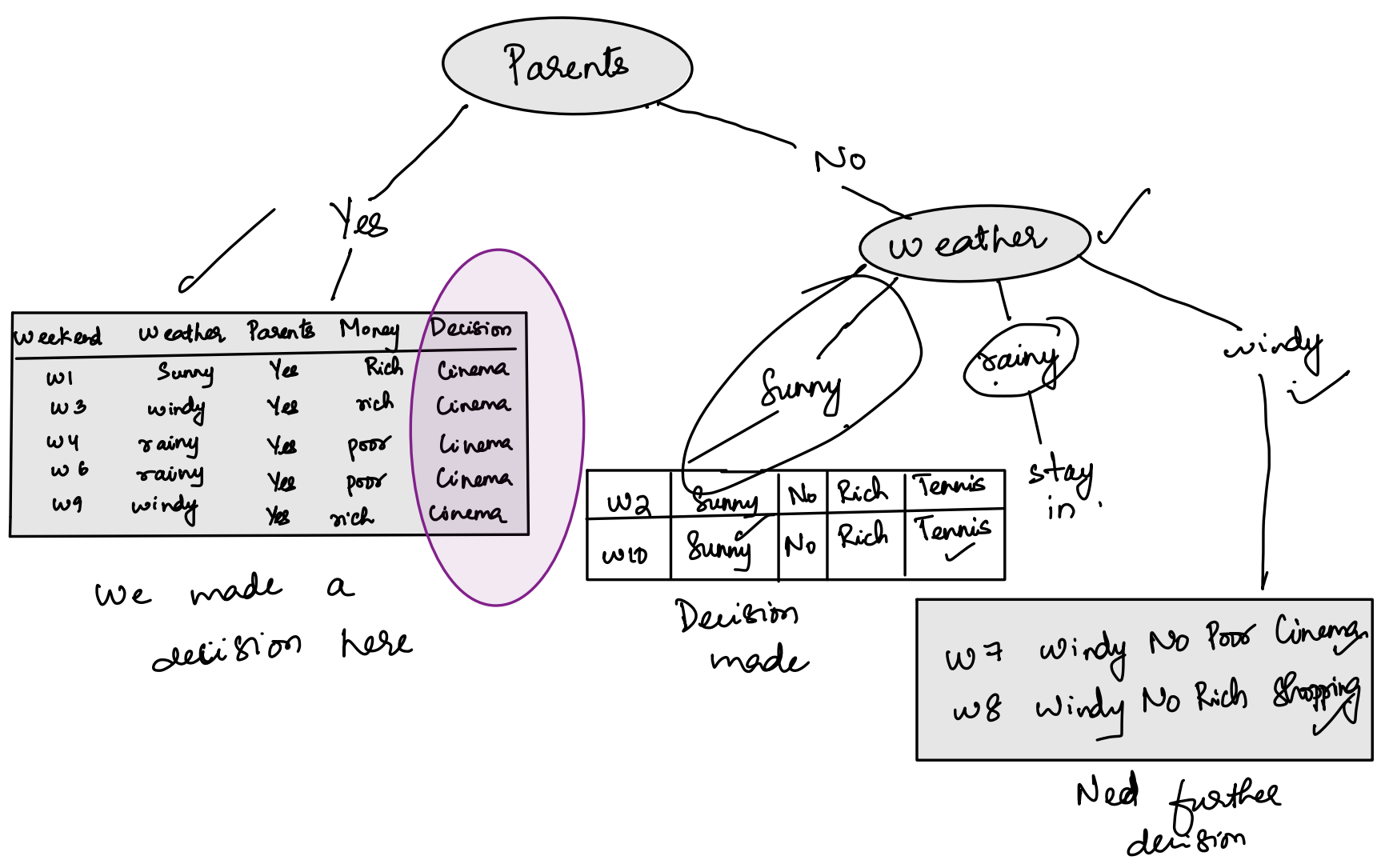
weekend	weather	Parents	Money	Decision
w2	Sunny	No	rich	Tennis
w5	rainy	No	rich	stay in
w7	windy	No	poor	cinema
w8	windy	No	rich	Shopping
w10	Sunny	No	rich	tennis

Comparing the impurities :-

Weather = 0.2

Money = 0.5

⇒ selected



Weekend	Weather	Parents	Money	Decision
W1	Sunny	Yes	Rich	Cinema
W3	windy	Yes	rich	Cinema
W4	rainy	Yes	poor	Cinema
W6	rainy	Yes	poor	Cinema
W9	windy	Yes	rich	Cinema

W2	Sunny	No	Rich	Tennis
W10	Sunny	No	Rich	Tennis

W7	windy	No	Poor	Cinema
W8	Windy	No	Rich	Shopping

Weekend	Weather	Parents	Money	Decision
W7	windy	NO	Poor	Cinema
W8	windy	NO	rich	Shopping

→ only 2 values for money

→ rich & poor

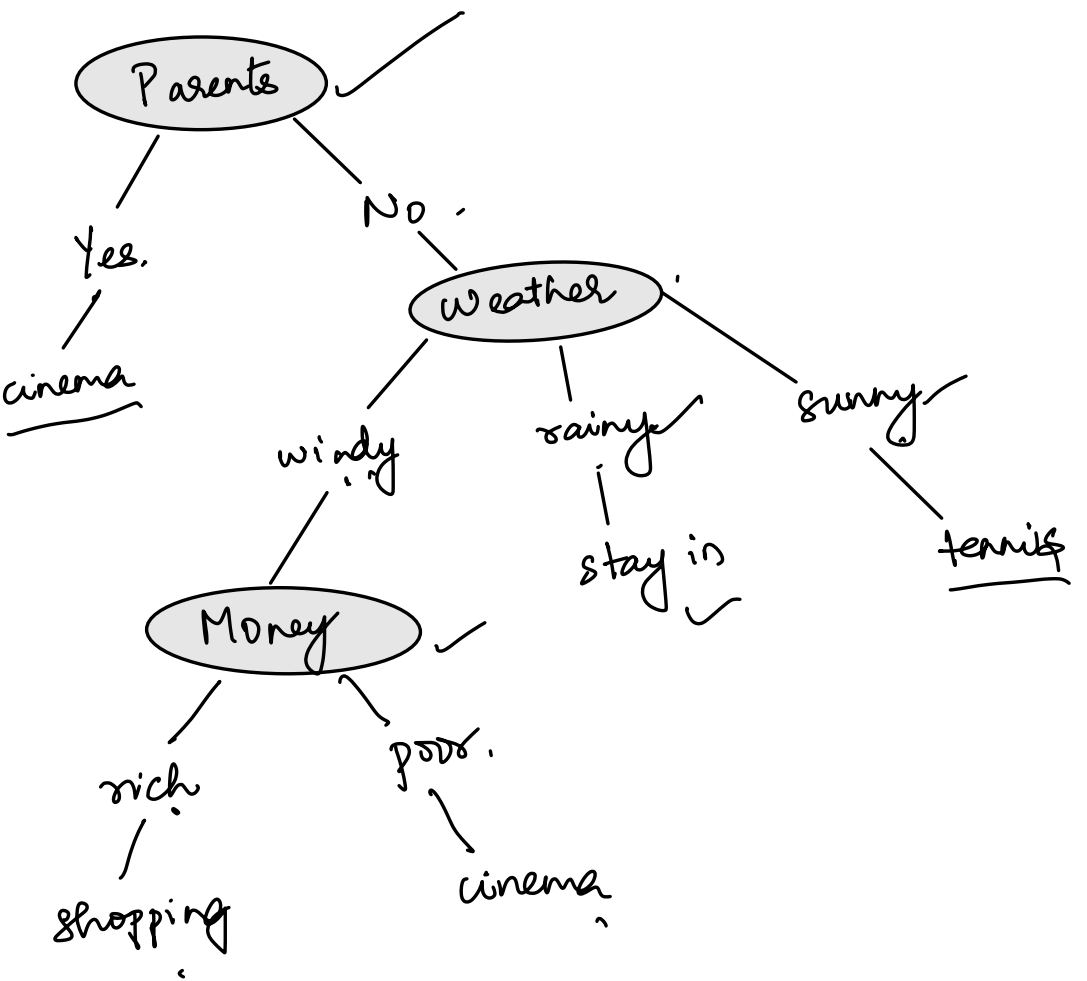
→ can directly

build a tree

⇒ poor ⇒ cinema

⇒ rich ⇒ shopping

Final tree


Parents
Yes.
No.
cinema
Weather
windy
rainy
sunny
Money
stay in
tennis
rich
poor.
shopping
cinema