→ Outlook এর gain বেশি, তাই এটা হবে decision tree এর root.

- Info gain dataset কে more appropriately ভাগ করে, Ay-
- If all the instances belongs to a Particular class then we would have a leaf node

C 4.5 Algorithm.

- C4.5 algorithm হল ID3 এর improvement.
- C4.5 এর জন্য প্রয়োজন হয় → Split $\text{Info}_A(D)$
  → Gain Ratio (A)
- ~~C4.5~~ →
- $ID_3 \xrightarrow{\text{Successor}} C4.5$

C4.5 এর ফর্মুলা:

→ $$\text{Split Info}_A(D) = -n\sum_{j=1}^{n} \frac{|Dj|}{|D|} \times \log_2\left(\frac{|Dj|}{|D|}\right)$$

→ $$\text{Gain Ratio}(A) = \frac{\text{Gain}(A)}{\text{Split Info}_A(D)}$$

- Gain Ratio তে Gain (A) দরকার হয়, Gain (A) এর ফর্মুলা

⇒ $$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D)$$

- অর্থাৎ ID_3 এর ফর্মুলাও কাজে লাগবে,

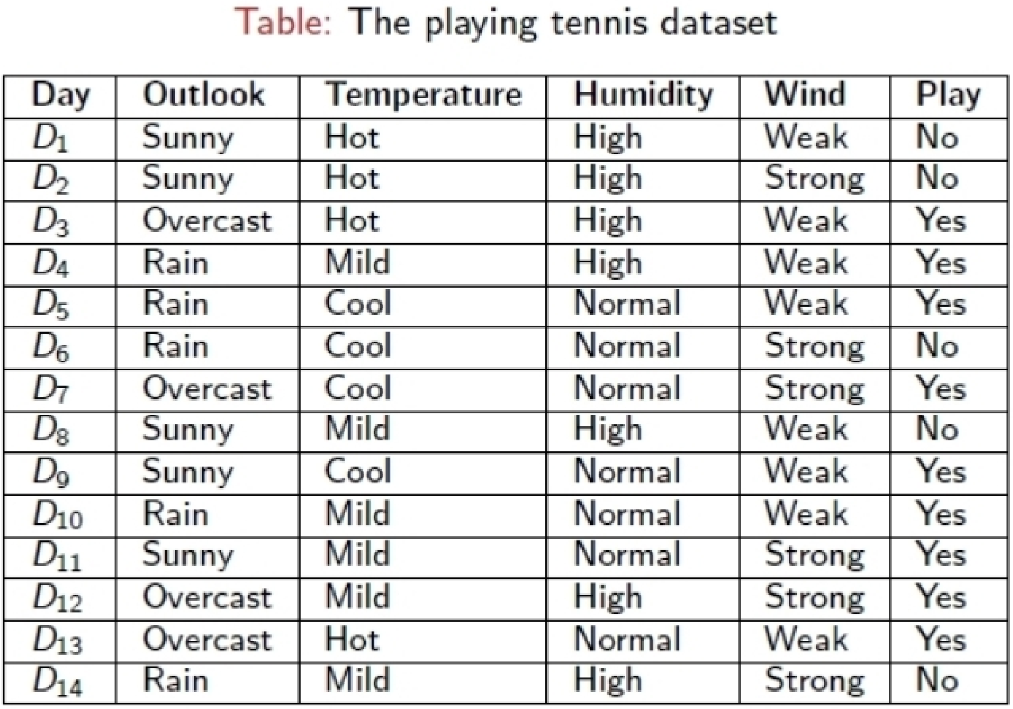

Table: The playing tennis dataset

Day	Outlook	Temperature	Humidity	Wind	Play
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

Find the root using C4.5 algorithm from the given dataset

$$\text{Info}(D) = -\frac{9}{14}\log_2(9/14) - \frac{5}{14}\log_2(5/14)$$

$$= -0.64 \times (-0.64) - 0.35 \times (-1.51)$$

$$= 0.40 + 0.52$$

$$= 0.92$$

Attribute ⇒ Outlook

$$\text{Info}_{\text{outlook}}(D) = \underbrace{5/14 \times (-2/5 \log_2 2/5 - 3/5 \log_2 3/5)}_{\text{Sunny}} +$$

$$\underbrace{4/14 \times (-4/4 \log_2 4/4 - 0/4 \log_2 0/4)}_{\text{Overcast}} +$$

$$\underbrace{5/14 \times (-3/5 \operatorname{Log} 3/5 - 2/5 \log 2/5)}_{\text{Rain}}$$

$$= 0.35 \times \{-0.4 \times (-1.32) - 0.6 \times (-0.73)\} +$$

$$0.28 \times \{-1 \times 0 - \log_2(0/4)\} +$$

$$0.35 \times \{-0.6 \times (-0.73 - 0.4 \times (-1.32)\}$$

$$= 0.6762$$

$$\text{Gain (Outlook)} = \text{Info}(D) - \text{Info}_{\text{outlook}}(D)$$

$$= 0.92 - 0.6762$$

$$= 0.2438$$

$$\text{Split Info (Outlook)} = \underbrace{-5/14 \log_2 5/14}_{\text{Sunny}} - \underbrace{-4/14 \log_2 4/14}_{\text{Overcast}} - \underbrace{-5/14 \log_2 5/14}_{\text{Rain}}$$

$$= -0.35 \log_2 (0.35) - 0.28 \log_2 (0.28) - 0.35 \times \log_2 (0.35)$$

$$= -0.35 \times (-1.51) - 0.28 \times (-1.83) - 0.35 \times (-1.51)$$

$$= 0.52 + 0.51 + 0.52$$

$$= 1.55$$

$$\text{Gain Ratio (Outlook)} = \frac{0.2438}{1.55}$$

$$= 0.157.$$

Attribute => Temperature

$$\text{Info}_{\text{Temperature}}(D) = \underbrace{4/14 \times (-2/4 \log 2/4 - 2/4 \log_2 2/4)}_{\text{Hot}} + \underbrace{6/14 \times (-4/6 \log_2 4/6 - 2/6 \log_2 2/6)}_{\text{Mild}} + \underbrace{4/14 \times (-3/4 \log_2 3/4 - 1/4 \log_2 1/4)}_{\text{Cool}}$$

$$= 0.28 \times \{-0.5 \times (-1) - 0.5 \times (-1)\} + 0.42 \times \{-0.66 \times (-0.599) - 0.33 \times (-1.59)\} + 0.28 \times \{-0.75 \times (-0.41) - 0.25 \times (-2)\}$$

$= 0.892$

$\text{Gain(Temparature)} = \text{Info}(D) - \text{Info}_{\text{Temparature}}(D)$

$= 0.92 - 0.892$

$= 0.028$

$\text{Split Info (Temparature)} = -\underbrace{\tfrac{4}{14}\log_2 \tfrac{4}{14}}_{\text{Hot}} - \underbrace{\tfrac{6}{14}\log_2 \tfrac{6}{14}}_{\text{Mild}} - \underbrace{\tfrac{4}{14}\log_2 \tfrac{4}{14}}_{\text{Cool}}$

$= -0.28 \log_2(0.28) - 0.42\log_2(0.42) - 0.28\log_2(0.28)$

$= -0.28 \times (-1.83) - 0.42 \times (-1.25) - 0.28 \times (-1.83)$

$= 0.51 + 0.525 + 0.51$

$= 1.545$

$\text{Gain Ratio (Temparature)} = \dfrac{0.028}{1.545}$

$= 0.018$

Attribute ⇒ Humidity

$$\text{Info}_{\text{Humidity}}(D) = \underbrace{7/14 \times \{-3/7 \log_2 3/7 - 4/7 \log_2 4/7\}}_{\text{High}} + \underbrace{7/14 \times (-6/7 \log_2 6/7 - 1/7 \log_2 1/7)}_{\text{Normal}}$$

$$= 0.5 \times \{-0.42 \times (-1.25) - 0.57 \times (-0.81)\} + 0.5 \times \{-0.85 \times (-0.23) - 0.14 \times (-2.83)\}$$

$$= 0.78$$

$$\text{Gain(Humidity)} = \text{Info}(D) - \text{Info}_{\text{Humidity}}(D)$$

$$= 0.92 - 0.785$$

$$= 0.135$$

$$\text{Split Info (Humidity)} = \underbrace{-7/14 \log_2 7/14}_{\text{High}} - \underbrace{-7/14 \log_2 7/14}_{\text{Normal}}$$

$$= -0.5 \log_2(0.5) - 0.5 \log_2(0.5)$$

$$= -0.5 \times (-1) - 0.5 \times (-1)$$

$$= 0.5 + 0.5$$

$$= 1$$

$$\text{Gain Ratio (Humidity)} = \frac{0.135}{1}$$

$$= 0.135$$

Attribute ⇒ Wind

$$\text{Info}_{wind}(D) = \underbrace{8/14\,(-6/8\,\log_2 6/8 - 2/8\,\log_2 2/8)}_{\text{Weak}} +$$

$$\underbrace{6/14\,(-3/6\,\log_2 3/6 - 3/6\,\log_2 3/6)}_{\text{Strong}}$$

$$= 0.57 \times \{-0.75 \times (-0.41) - 0.25 \times (-2)\} + 0.42 \times \{-0.5 \times (-1) - 0.5 \times (-1)\}$$

$$= 0.88$$

$$\text{Gain (Wind)} = \text{Info}_D - \text{Info}_{wind}(D)$$

$$= 0.92 - 0.88$$

$$= 0.04$$

$$\text{Split Info (Wind)} = \underbrace{-8/14\,\log_2 8/14}_{\text{Weak}} - \underbrace{6/14\,\log_2 6/14}_{\text{Strong}}$$

$$= -0.57 \times \log_2(0.57) - 0.42 \times \log_2(0.42)$$

$$= -0.57 \times (-0.81) - 0.42 \times (-1.25)$$

$$= 0.46 + 0.525$$

$$= 0.985$$

$$\text{Gain Ratio}(\text{Wind}) = \frac{0.04}{0.985}$$
$$= 0.04$$

Here

Gain Ratio (Outlook) = 0.2438
Gain Ratio (temperature) = 0.0181
Gain Ratio (Humidity) = 0.135
Gain Ratio (Wind) = 0.04

Gain Ratio of outlook is higher, so it would be the spliting feature.

প্র Write the usage of Gini Index

Ans. The Gini index is used in classification & Regression Trees (CART) algorithm.

Gini (CART) Formula

a. $Gini(D) = 1 - \sum_{i=1}^{n} P_i^2$

b. $Gini_A(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$ [এটা Slide এ আছে]

c. $\Delta Gini(A) = Gini(D) - Gini_A(D)$

d. $Gini_A(D) = \sum_{j=1}^{n} \left(\frac{|D_j|}{|D|} \times Gini|D_j| \right.$ [এটা দিয়ে $Gini_A(D)$ easily বের হয়]

Attribute Total Instance.

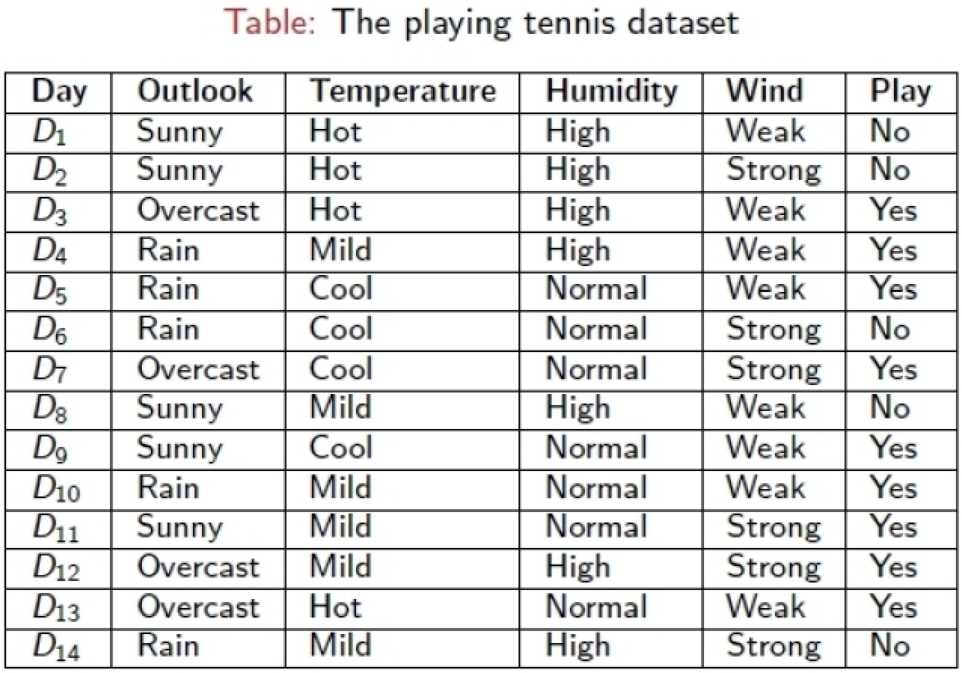

Table: The playing tennis dataset

Day	Outlook	Temperature	Humidity	Wind	Play
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

Find the root using CART(Gini Index) algorithm from the given dataset

Solve:

$$Gini\,(D) = 1 - \sum_{i=1}^{n} p_i^2$$

$$= 1 - \{ (9/14)^2 + (5/14)^2 \}$$

$$= 1 - \{ (0.64)^2 + (0.35)^2 \}$$

$$= 1 - (0.40 + 0.12)$$

$$= 1 - 0.52$$

$$= 0.48$$

Attribute ⇒ Outlook

~~$Gini_{Sunny}(D) =$~~

$$Gini_{Outlook}(D) = \underbrace{5/14 \{ 1 - ((2/5)^2 + (3/5)^2) \}}_{Sunny} +$$

$$\underbrace{4/14 \{ 1 - ((4/4)^2 + (0/4)^2) \}}_{Overcast} +$$

$$\underbrace{5/14 \{ 1 - ((3/5)^2 + (2/5)^2) \}}_{Rainy}$$

$$= 0.35 \times (1 - 0.16 - 0.36) + 0.28 \times (1 - 1) + 0.35 \times (1 - 0.36 - 0.16)$$

$$= 0.35 \times 0.48 + 0.35 \times 0.48$$

$$= 0.998$$

$$\Delta \text{Gini (Outlook)} = 0.48 - 0.998$$
$$= -0.518$$

Attribute ⇒ Temperature

$$\text{Gini}_{\text{Temperature}}(D) = \frac{4/14\{1 - ((2/4)^2 + (2/4)^2)\}}{\text{Hot}} +$$

$$\frac{6/14\{1 - ((4/6)^2 + (2/6)^2)\}}{\text{Mild}} +$$

$$\frac{4/14\{1 - ((3/4)^2 + (1/4)^2)\}}{\text{Cool}}$$

$$= 0.28 \times 0.5 + 0.42 \times 0.44 + 0.28 \times 0.375$$

$$= 0.4298$$

$$\Delta \text{Gini (Temperature)} = 0.48 - 0.4298$$
$$= 0.050$$

Attribute → Humidity

$$\text{Gini}^{(D)}_{\text{Humidity}} = \underbrace{7/14\,\{1-((3/7)^2+(4/7)^2)\}}_{\text{High}} +$$

$$\underbrace{7/14\,\{1-((6/7)^2+(1/7)^2)\}}_{\text{Normal}}$$

$$= 0.5 \times 0.48 + 0.5 \times 0.244$$

$$= 0.362$$

$$\Delta\text{Gini}(\text{Humidity}) = 0.48 - 0.362$$

$$= 0.118$$

Attribute ⇒ Wind

$$\text{Gini}^{(D)}_{\text{Wind}} = \underbrace{8/14\,\{1-((6/8)^2+(2/8^2)\}}_{\text{Weak}} +$$

$$\underbrace{6/14\,\{1-((3/6)^2+(3/6)^2)\}}_{\text{Strong}}$$

$$= 0.57 \times 0.375 + 0.42 \times 0.5$$

$$= 0.42$$

$$\Delta\text{Gini}(\text{Wind}) = 0.48 - 0.42$$

$$= 0.06$$

Here

ΔGini (Outlook) = -0.518

ΔGini (Temperature) = 0.050

ΔGini (Humidity) = 0.118

ΔGini (Wind) = 0.06

Here highest Delta gini comes from Humidity. So, it would be the root.