①

Desion Tree - play Tennis
Gain calculation.

DAY	outlook	temperature	humidity	wind	Decision
1	Sunny	hot	high	weak	No
2	Sunny	hot	high	strong	NO
3	overcast	hot	high	weak	Yes
4	rainfall	mild	high	weak	Yes
5	raintall	cool	normal	weak	Yes
6	raintall	cool	normal	Stoorg	NO
7	overcast	cool	normal	strong	Yes
8	Sunny	mild	high	weak	NO
9	Sunny	cool	normal	weak	Yes
10	rainfall	mild	normal	weak	yes
11	Sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	Yes
13	overcast	hot	normal	weak	yes
14	rainfall	mild	high	strong	NO

$$Gain(S, f_i) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$$

~~CART~~ Purity – Entropy (from ID3)
– gini Impurity (from CART)

ID3 → Entropy = $-\sum_{i=1}^{n} P_i \log_2(P_i)$ Gini coeff = $1 - \sum_{i=1}^{n} P_i^2$

H(S) = Entropy → $0 \leq 1$ ⇒ 0 → Good Entropy (pure split)
Gini Coeff → $0 \leq 0.5$

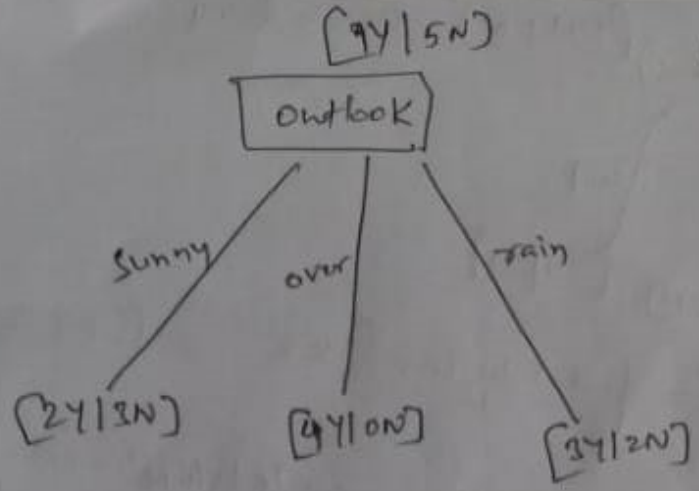
Conclusion:

8a.
Calculated information Gain individually for all 4 features.

Outlook has highest Gain ≃ 0.24. ✓

- Tempe Gain = 0.06.
- Windy = 0.04
- Humidity = 0.16

(2)



(Play Tennis)

$E_s = -\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{3}{5}\log\left(\frac{3}{5}\right) = 0.97$

$E_o = -\frac{4}{4}\log\left(\frac{4}{4}\right) - \frac{0}{4}\log\left(\frac{0}{4}\right) = 0$

$E_{rain} = -\frac{3}{5}\log\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = -0.6\log(0.6) - 0.4\log(0.4) = 0.97$

$H(S)$ ~~Gain (outlook)~~ $= E_{outlook} = -\frac{9}{14}\log\left(\frac{9}{14}\right) - \frac{5}{14}\log\left(\frac{5}{14}\right) = -0.64\log(0.64) - 0.357\log(0.357)$

$H(S) = 0.94$

$$Gain = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$$

$$= 0.94 - \left[\frac{5}{14}(0.97) + 0 + \frac{5}{14}(0.97)\right]$$

$$= 0.94 - \left[2 \times \frac{5}{14}(0.97)\right] = 0.94 - \left[2 \times 0.35 \times 0.97\right]$$

$$= 0.94 - 0.69$$

$$Gain_{[outlook]} = 0.24.$$

(Temperature) [9+|5-] $H(S) = 0.94$.

hot — [2+|2-]

mild — [4+|2-]

cool — [3+|1-]

$E_{hot} = \left[\frac{-2}{4}\log\left(\frac{2}{4}\right)\right] \times 2 = \left[-0.5\log(0.5)\right] \times 2 = [-0.5(-1)] \times 2 = 1$

$E_{mild} = \frac{4}{6} = 0.66 \Rightarrow \frac{2}{6} = 0.33 = -0.66\log(0.66) - 0.33\log(0.33)$

$= (-0.66 \times -0.59) - [0.33 \times (-1.5)]$

$= 0.38 + 0.49 = 0.87$

$E_{cool} = \frac{3}{4} = 0.75, \quad \frac{1}{4} = 0.25$

$= -0.75\log(0.75) - 0.25\log(0.25)$

$= -0.41 - -2$

$= +0.307 + 0.5 = 0.807.$

$Gain = 0.94 - \left[\frac{4}{14}(1) + \frac{6}{14}(0.87) + \frac{4}{14}(0.807)\right]$

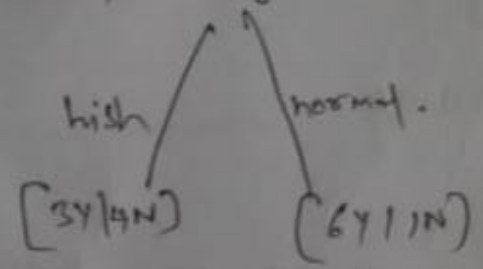
$= 0.94 - [0.28 + 0.372 + 0.22]$

$= 0.94 - [0.88] = 0.06.$

$Gain_{Temp} = 0.06$

(3)

[9Y|5N] humidity $H(S) = 0.94$.



$E_{high} \Rightarrow \frac{3}{7} = 0.42$, $\frac{4}{7} = 0.571$

$= -0.42 \log(0.42) - 0.571 \log(0.571)$

$= -0.42(-1.25) - 0.571(-0.801)$

$= 0.525 + 0.461 = 0.98$

$E_{normal} = -6 \log(6) - 0.16 \log(0.16)$

$= -6(2.58) - 0.16(-2.6)$

$= +5.4$

$E_{normal} = -\frac{6}{7} \log\left(\frac{6}{7}\right) - \frac{1}{7} \log\left(\frac{1}{7}\right)$

$= -0.85 \log(0.85) - 0.14 \log(0.14)$

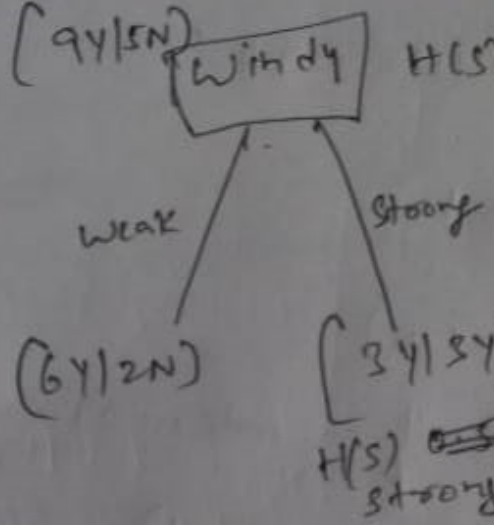
$= -0.85(-0.23) - 0.14(-2.83)$

$= 0.19 + 0.39 = 0.58$

Gain (humidity) $= 0.94 - \underbrace{\left[\left(0.98 \times \frac{7}{14}\right) + 0.58 \times \frac{7}{14}\right]}_{0.49 + 0.29}$

$= 0.94 - 0.78$

$= 0.16.$



$E_{weak} = \frac{6}{8} = 0.75 \quad \frac{2}{8} = 0.25$

$= -0.75 \log(0.75) - 0.25(\log 0.25)$

$= -0.75(-0.415) - 0.25(-2)$

$= 0.311 + 0.5 = 0.811$

$\text{Gain} = 0.94 - \left[0.811 \times \frac{8}{14} + \cdot 1 \times \frac{6}{14}\right]$

$= 0.94 - [0.463 + 0.42]$

Windy Gain $= 0.94 - 0.891 = 0.04.$

④

Entropy	Gini Impurity
① Algorithm used for calculation ID3	CART Algorithm.
② Value max purity = 0 max impurity = 1 $\Rightarrow [0, 1]$	$[0, 0.5] \rightarrow$ 0.5 is max purity.
③ Formula:- $-\sum_{i=1}^{n} p_i \log_2(p_i)$	$\Rightarrow 1 - \sum p_i^2$
④ measure of Information that Indicates disorderness of the features with target.	measures the frequency at which any element of the data set will be mislabeled when it is randomly labeld
⑤ use logarithms more complex.	calculation is faster