Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
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

construct a Decision Tree using ID_3 Algorithm

→ we need to find the attribute giving maximum information among the available attributes.

→ 4 attributes are given, the IG of all 4 attributes needs to be calculated. The attribute with the max IG will be the root node & we start building the tree from here.

→ Consider the 1^{st} Attribute : outlook → sky condition.

The possible values for outlook

V(outlook) = <Sunny, Overcast → cloudy, Rain>

→ we need to calculate the Entropy of the whole data set (S) & the Entropy of individual values.

Entropy(S)

$$S = [9+, 5-] = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = \underline{0.94}$$

Similarly we need to find the Entropy for $\underline{V(\text{outlook})}$

→ Count the number of Examples with Sunny for outlook attribute; we need to count the number of Ex.

$$S_{Sunny} = [2+, 3-] \quad , \quad \text{So } Entropy(S_{Sunny}) = \underbrace{\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)}_{\text{proportion of positive examples}} - \underbrace{\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)}_{\text{proportion of negative example}} = \underline{0.97}$$

5 Examples with 2 positive & 3 Negative

$$S_{overcast} = [4+, 0-] \quad , \quad Entropy(S_{overcast}) = -\left(\frac{4}{4}\right)\log_2\left(\frac{4}{4}\right) - \left(\frac{0}{4}\right)\log_2\left(\frac{0}{4}\right)$$

$$= 0$$

4 Examples of overcast in outlook attribute

Note ① If there are only positive or negative examples for a particular value, we can write 0 directly. for this ie $\boxed{Entropy = 0}$

② If equal number of positive & negative examples, we can directly write $\boxed{Entropy = 1}$

9) $S_{Rain} = [3+, 2-]$ $Entropy(S_{Rain}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) = 0.971$

5 Examples

$$Gain(S, outlook) = Entropy(S) - \sum_{v \in (Sunny, Overcast, Rain)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Gain → Given the dataset S,

Entropy(S) → Entropy of whole data set

Note | | is Cardinality which is the number of distinct values for a column in a table.

S_v is the number of possible values for Sunny, Overcast & Rain

S_v is $\frac{S_{Sunny}}{S}$, $\frac{S_{Overcast}}{S}$ & $\frac{S_{Rain}}{S}$ & Entropy of that particular value.

$$Gain(S, outlook) = Entropy(S) - \frac{5}{14} \times Entropy(S_{Sunny}) - \frac{4}{14} \times Entropy(S_{overcast}) - \frac{5}{14} \times Entropy(S_{Rain})$$

$$= 0.94 - \left(\frac{5}{14}\right)0.971 - \left(\frac{4}{14}\right)0 - \left(\frac{5}{14}\right)0.971 = 0.2464$$

$$\boxed{Gain(S, outlook) = 0.2464}$$

Attribute : Temperature

Values(Temp) = (Hot, Mild, Cool)

$S = Entropy(S) = 0.94$

$S_{Hot} \leftarrow [2+, 2-]$ $Entropy(S_{Hot}) = 1.0$

$S_{Mild} \leftarrow [4+, 2-]$ $Entropy(S_{Mild}) = 0.9183$

$S_{Cool} = [3+, 1-]$ $Entropy(S_{Cool}) = 0.8113$

$$Gain(S, Temperature) = Entropy(S) - \sum_{v \in (Hot, Mild, Cool)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\text{Gain}(S, \text{Temperature}) = 0.94 - \left(\frac{4}{14}\right)\text{Entropy}(S_{Hot}) - \left(\frac{6}{14}\right)\text{Entropy}(S_{Mild}) - \left(\frac{4}{14}\right)\text{Entropy}(S_{Cool})$$ 3

$$\boxed{\text{Gain}(S, \text{Temperature}) = 0.0289}$$

Attribute : Humidity ; Calculate IG for Humidity Attribute

Values (Humidity) = (High, Normal)

$$\boxed{\text{Gain}(S, \text{Humidity}) = 0.1516}$$

Attribute : Wind.

Values (Wind) = (Strong, Weak)

$$\boxed{\text{Gain}(S, \text{Wind}) = 0.0478}$$

④ Comparing Gain of all 4 attributes

Gain (S, outlook) = 0.2464 → Highest Gain & will be Considered as Root Node

Gain (S, Temp) = 0.0289

Gain (S, Humidity) = 0.1516

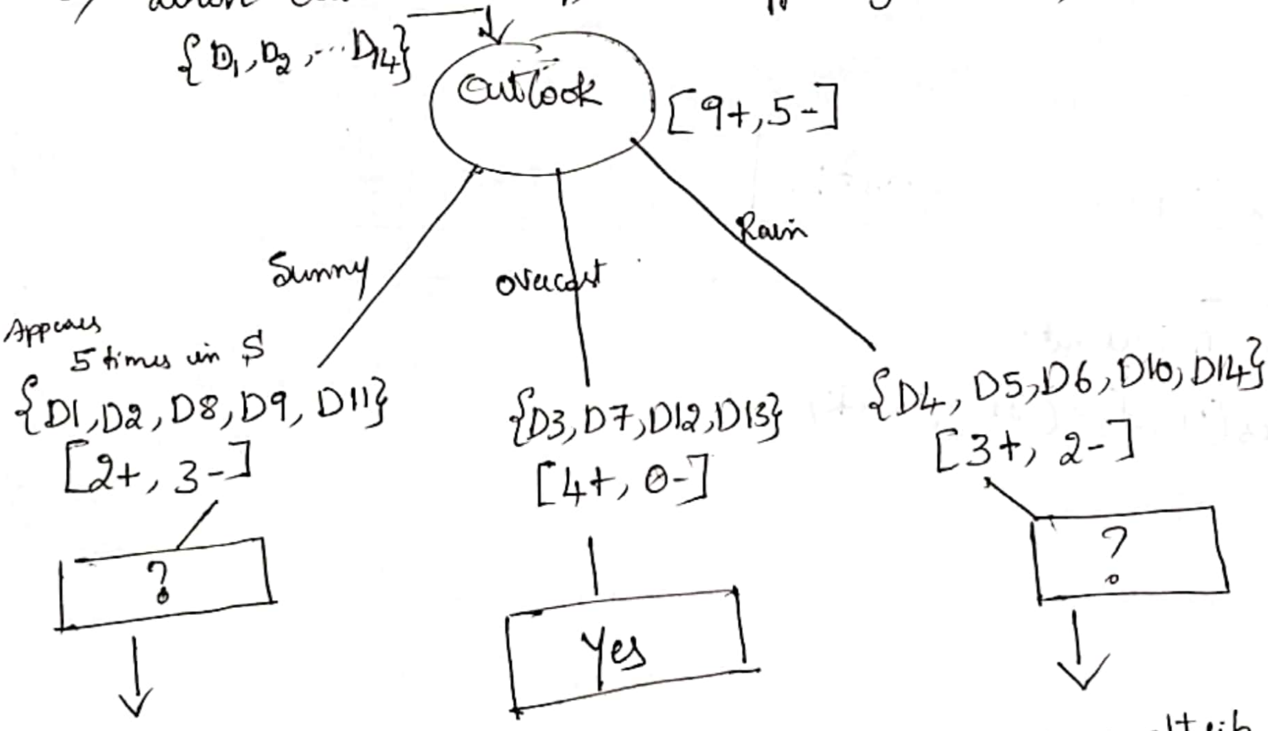
Gain (S, wind) = 0.0478

→ outlook has 3 values, So we draw 3 branches for out look attribute

For overcast, all labels are yes, so we can directly write yes; if combination of yes & No, we cannot write directly

→ when outlook = Sunny, it is appearing 5 times,

{D1, D2, ... D14}



Here we need to introduce a new attribute with max IG

new attrib with max IG.

→ we need to continue the calculation for Sunny & Rain

Scanned with CamScanner

(b) → We shall start with left side

→ From the 14 examples $D_1, D_2 \ldots D_{14}$, only 5 examples should be considered for Sunny i.e $\{D_1, D_2, D_8, D_9, D_{11}\}$.

→ we need not consider outlook since it is already considered we shall consider only Temp, Humidity, wind.

→ Sunny

Day	Temp	Humidity	Wind	playTennis
D_1	Hot	High	Weak	NO
D_2	Hot	High	Strong	No
D_8	Mild	High	Weak	No
D_9	Cool	Normal	Weak	Yes
D_{11}	Mild	Normal	Strong	Yes

Consider Attribute Temperature

$$\text{Values}(\text{Temp}) = (\text{Hot}, \text{Mild}, \text{Cool})$$

$$S_{Sunny} = [2+, 3-] = \text{Entropy}(S_{Sunny}) = -\left(\frac{2}{5}\right)\log_2 \frac{2}{5} - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)$$

$$= 0.97$$

$$S_{Hot} = [0+, 2-] = \text{Entropy}(S_{Hot}) = 0.0$$

(Since one is Zero)

$$S_{Mild} = [1+, 1-] = \text{Entropy}(S_{Mild}) = 1.0$$

(Since equal no. of positive & Negative values)

$$S_{Cool} = [1+, 0-] = \text{Entropy}(S_{Cool}) = 0.0$$

$$\text{Gain}(S_{Sunny}, \text{Temp}) = \text{Entropy}(S_{Sunny}) - \sum_{v \in (\text{Hot}, \text{Mild}, \text{Cool})} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

(6) $Gain(S_{sunny}, Temp) = Entropy(S_{sunny}) - \frac{2}{5} Entropy(S_{Hot}) - \frac{2}{5} Entropy(S_{Mild}) - \frac{1}{5} Entropy(S_{Cool})$

Gain (S_{sunny}, Temp)

$\boxed{\text{Gain of Temp wrt } S_{sunny} = 0.570}$

Consider the attribute Humidity, calculate the Gain of Humidity wrt S_{sunny} i.e Gain(S_{sunny}, Humidity)

Entropy(S_{sunny}) will be same as previous

$S_{sunny} = [2+, 3-]$ Entropy(S_{sunny}) = 0.97.

$S_{high} = [0+, 3-]$ Entropy(S_{High}) = 0.0

$S_{Normal} = [2+, 0-]$ Entropy(S_{Normal}) = 0.0

$Gain(S_{sunny}, Humidity) = Entropy(S_{sunny}) - \left(\frac{3}{5}\right) Entropy\ S_{High} - \left(\frac{2}{5}\right) Entropy\ S_{Normal}$

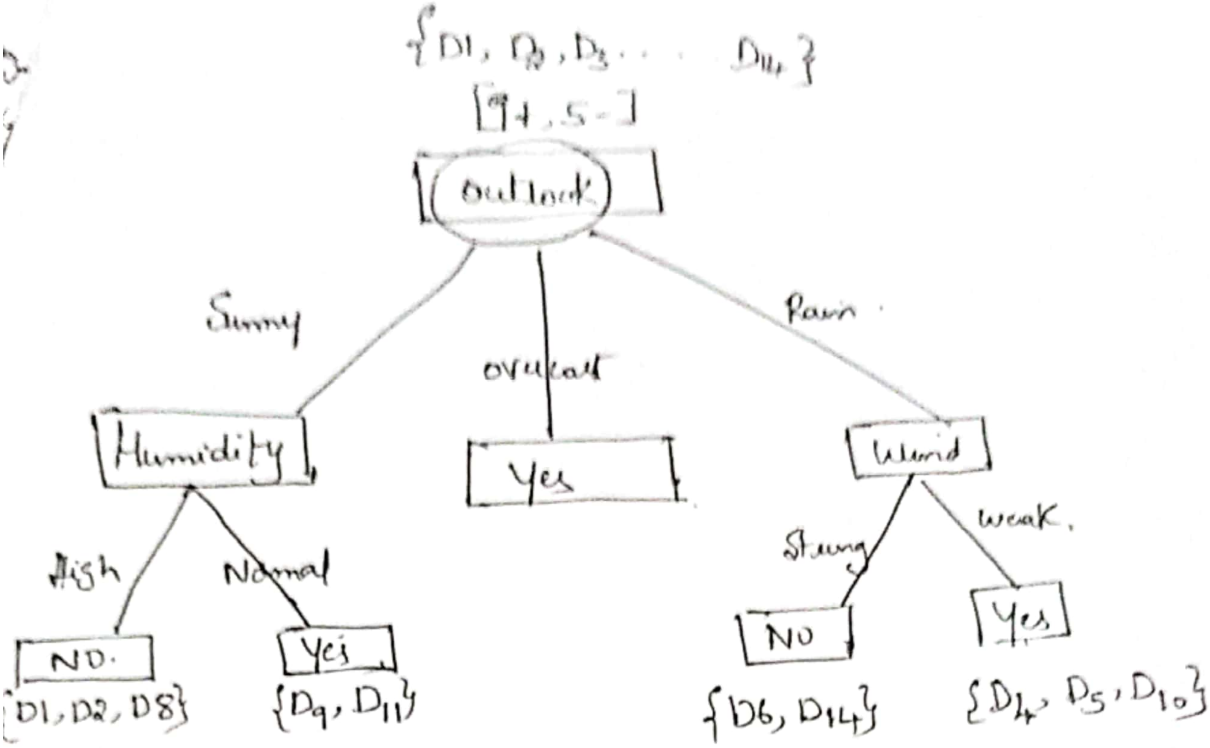
Gain(S_{sunny}, Humidity) = 0.97.

Compute

Gain (S_{sunny}, wind) = 0.0192

Gain (S_{sunny}, Temp) = 0.570

Gain for Humidity is Highest, so the humidity attribute is considered as next node



Day Temp Humidity wind play

For Rain value, we shall consider Temperature wind since outlook & Humidity has been consid

Rain

Day	Temp	wind	playTennis
D4	Mild	Weak	Yes
D5	Cool	Weak	yes
D6	Cool	Strong	No
D10	Mild	Weak.	Yes
D14	Mild	Strong.	NO

$Gain(S_{Rain}, Temp) = 0.0202$.

$Gain(S_{Rain}, Wind) = 0.971$