"Decision Tree Using Entropy & Information Gain"

Day	Outlook	Temp	Humidity	Wind	PlayTennis
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

Dataset Comprising of '14 instances' & "four attributes" & Play Tennis is the target variable in this case.

If we want to draw decision tree, we need to find out the attribute which is giving maximum information gain out of the remaining attributes. So, here we are given four attributes, now information gain of every attribute is computed, & one with the maximum information gain will become the ["root node"]. & then we will start building the tree.

Let's start computing Information gain for each of the attributes :-

1) Attribute: Outlook

Values (Outlook) = [Sunny, Overcast, Rain]

if we want to compute the information gain of an attribute, we first need to compute the entropy of individual attribute values.

Let's first compute the entropy of entire dataset

In this dataset, we have "9 positive examples" & '5 negative examples".

P.T.O

$S = [9+, 5-]$ → It represents whole dataset

$\therefore$ Entropy$(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$

$= 0.94$

Proportion of +ive examples — Proportion of -ive examples

This is the entropy of entire dataset.

Now in a similar way we need to find the "entropy" of (Sunny, Overcast, Rain)

$S_{sunny} = [2+, 3-]$ · Entropy$(S_{sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = \boxed{0.971}$

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

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

So, we have computed entropy of entire dataset & entropy of individual values of the attribute ["Outlook"].

Now let us calculate the gain for "Outlook given the dataset S"

$$\text{Gain}(S, \text{outlook}) = \text{Entropy}(S) - \sum_{V \in (\text{Sunny, overcast, Rain})} \frac{|S_V|}{|S|}\,\text{Entropy}(S_V)$$

$$= \text{Entropy}(S) - \frac{5}{14}\,\text{Entropy}(S_{sunny}) - \frac{4}{14}\,\text{Entropy}(S_{overcast}) - \frac{5}{14}\,\text{Entropy}(S_{Rain})$$

$$\text{Gain}(S, \text{outlook}) = 0.94 - \frac{5}{14} * 0.971 - \frac{4}{14} * 0 - \frac{5}{14} * 0.971 = \boxed{0.2464}$$

Now let us compute the gain for second Attribute : Temp" (3)

Attribute : Temp

Values (Temp) = Hot, Mild, Cool

$S = [9+, 5-]$. Entropy $(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = \boxed{0.94}$

$S_{Hot} = [2+, 2-]$ Entropy $(S_{Hot}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = \boxed{1.0}$

$S_{Mild} = [4+, 2-]$ Entropy $(S_{Mild}) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = \boxed{0.9183}$

$S_{Cool} = [3+, 1-]$ Entropy $(S_{Cool}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = \boxed{0.8113}$

Now let's compute the Gain for Temp attribute given this data set.

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

$$\text{Gain}(S, \text{Temp}) = \text{Entropy}(S) - \frac{4}{14}\text{Entropy}(S_{Hot}) - \frac{6}{14}\text{Entropy}(S_{Mild}) - \frac{4}{14}\text{Entropy}(S_{Cool})$$

$$\text{Gain}(S, \text{Temp}) = 0.94 - \frac{4}{14}1.0 - \frac{6}{14}0.9183 - \frac{4}{14}0.8113 = \boxed{0.0289}$$

Now let us try to compute the gain for Third "attribute: Humidity"

Attribute : Humidity

Values (Humidity) = High, Normal

$S = [9+, 5-]$. Entropy $(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = \boxed{0.94}$

$S_{High} = [3+, 4-]$ Entropy $(S_{High}) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = \boxed{0.9852}$

$S_{Normal} = [6+, 1-]$ · $Entropy(S_{Normal}) = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.5916$

$$Gain(S, Humidity) = Entropy(S) - \sum_{v \in (High, Normal)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$Gain(S, Humidity)$

$$= Entropy(S) - \frac{7}{14} Entropy(S_{High}) - \frac{7}{14} Entropy(S_{Normal})$$

$$Gain(S, Humidity) = 0.94 - \frac{7}{14} 0.9852 - \frac{7}{14} 0.5916 = \boxed{0.1516}$$

Now let's compute the information gain for ['Attribute : Wind"]

Values (Wind) = Strong, Weak

$S = [9+, 5-]$ $\quad Entropy(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$

$S_{Strong} = [3+, 3-]$ · $Entropy(S_{Strong}) = 1.0$

$S_{Weak} = [6+, 2-]$ $\quad Entropy(S_{Weak}) = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.8113$

$$Gain(S, Wind) = Entropy(S) - \sum_{v \in (Strong, Weak)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$Gain(S, Wind) = Entropy(S) - \frac{6}{14} Entropy(S_{Strong}) - \frac{8}{14} Entropy(S_{Weak})$$

$$= 0.94 - \frac{6}{14} 1.0 - \frac{8}{14} 0.8113 = \boxed{0.0478}$$

P.T.O

Gain (S, Outlook) = 0.2464
Gain (S, Temp) = 0.0289
Gain (S, Humidity) = 0.1516
Gain (S, Wind) = 0.0478

Since Gain of Outlook is maximum, so outlook will become our root node.

for outlook to be rootnode let's check the possibilities.
We have "Sunny", "Overcast" & "Rain" as the possibilities for this attribute "Outlook"

Root node is overcast & it has three possible values so we will draw three branches for this root (outlook)

When "Outlook" is "Sunny", it is appearing only 5 times in our dataset [D1, D2, D8, D9 & D11] Only these five examples we need to consider when ["Outlook"] is ["Sunny"].

When 'outlook' is 'Overcast', we need to consider [D3, D7, D12 & D13]

& when 'outlook' is 'Rain', we need to consider [D4, D5, D6, D10, D14]

So, these are the instances we need to consider.

Now when we consider sunny as one branch, let's look at the target value (variable) for D1, D2 & D8 target variable is 'No' for sunny
& for D9 & D11 target variable is "Yes"

So, out of these five instance '3' are -ive & 2 are +ive.
So, there is a dilemma. (we donot know whether to put a 'Yes' or a 'No'.)

Similarly we have to check for 'Overcast' & 'Rain'

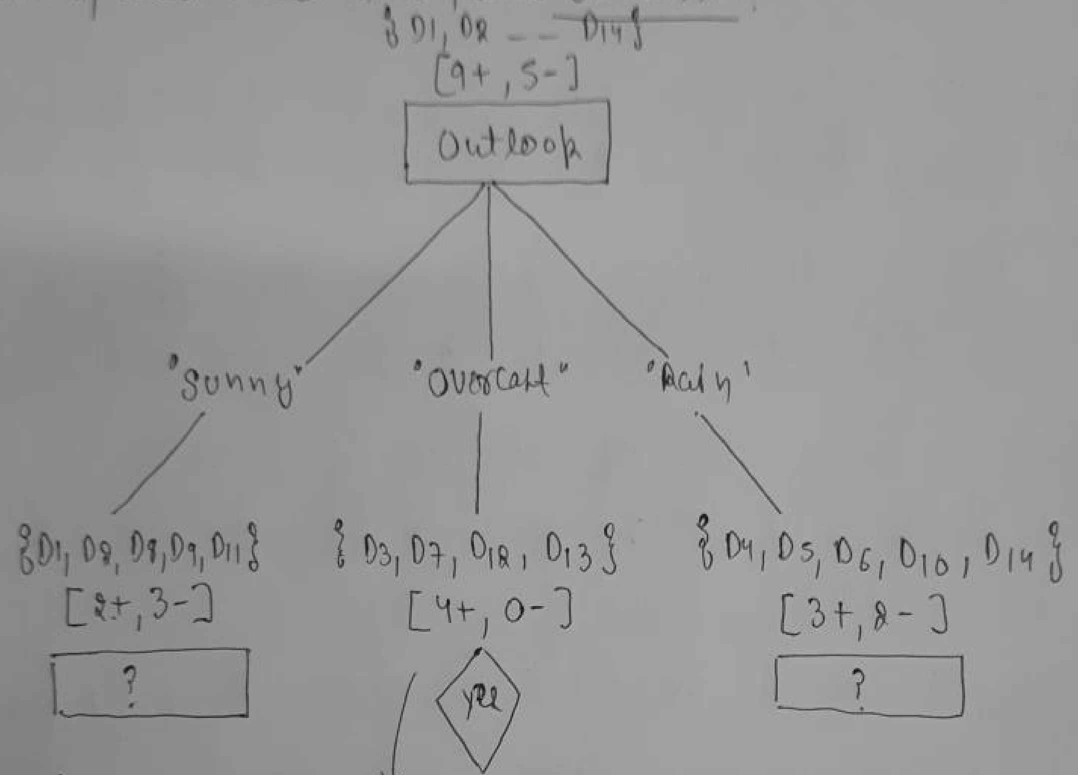
While considering 'Overcast', we need to look at [D3, D7, D12, D13] instances

In all these instance for overcast, target variable value is 'yes'

So in that case we can directly write a 'yes'

Similarly while considering 'Rain', we need to look at [D4, D5, D6, D10, D14]

This is having a combination of 'Yes & 'No'. Again there is a dilemma, we are not sure, what to write.



This is our leaf node, & we will continue growing the tree for Sunny & Rain attribute

Day	Temp	Humidity	Wind	Play Tennis
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

Outlook is already considered. Now we dont need to look at it

Out of these "5" 3 are -ive examples & 2 are +ive examples

Now again we need to compute the information Gain for these attributes.

Attribute : Temp

Values (Temp) = Hot, Mild, Cool

$S_{Sunny} = [2+, 3-]$. Entropy $(S_{Sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

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

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

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

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

$$= \text{Entropy}(S) - \frac{2}{5}\text{Entropy}(S_{Hot}) - \frac{2}{5}\text{Entropy}(S_{Mild}) - \frac{1}{5}\text{Entropy}(S_{Cool})$$

$$= 0.97 - \frac{2}{5}0.0 - \frac{2}{5}1 - \frac{1}{5}0.0 = \boxed{0.570}$$

Attribute : Humidity

Values (Humidity) = High, Normal

$S_{Sunny} = [2+, 3-]$. Entropy $(S) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = \boxed{0.97}$

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

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

$$\text{Gain}(S_{Sunny}, \text{Humidity}) = \text{Entropy}(S) - \sum_{V \in (\text{High, Normal})} \frac{|S_v|}{|S|}\text{Entropy}(S_v)$$

$$= \text{Entropy}(S) - \frac{3}{5}\text{Entropy}(S_{High}) - \frac{2}{5}\text{Entropy}(S_{Normal})$$

$$= 0.97 - \frac{3}{5}0.0 - \frac{2}{5}0.0 = \boxed{0.97}$$

P.T.O

Similarly we need to find out the gain with respect to "outlook = sunny"

Attribute : Wind

Values (Wind) = Strong, Weak

$S_{sunny} = [2+, 3-]$ $\quad$ $Entropy(S) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

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

$S_{weak} = [1+, 2-]$ $\quad$ $Entropy(S_{weak}) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9183$

$$Gain(S_{sunny}, Wind) = Entropy(S) - \sum_{v \in (strong, weak)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= Entropy(S) - \frac{2}{5} Entropy(S_{strong}) - \frac{3}{5} entropy(S_{weak})$$

$$= 0.97 - \frac{2}{5}1.0 - \frac{3}{5}0.918 = \boxed{0.0192}$$

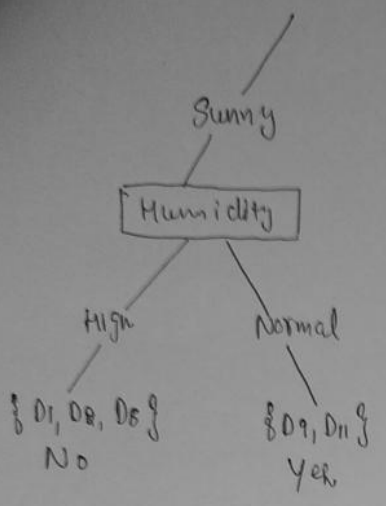
$\therefore$ $Gain(S_{sunny}, Temp) = \boxed{0.570}$

$Gain(S_{sunny}, Humidity) = \boxed{0.97}$

$Gain(S_{sunny}, Wind) = \boxed{0.0192}$

$\therefore$ at this particular level we will consider 'Humidity' as a node as its Gain is maximum

P.T.O



Now let's move to the outloop branch of Rain.

Proceed in similar way