# Decision Tree Classifier with Example

| outlook | temp | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |

## ID3 Approach

Entropy is defined as the measure of randomness in the information being processed. It is given by the formula

$$H(S) = \text{Entropy} = -p_y \log_2 p_y - p_N \log_2 p_N$$

Information Gain is defined as

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

We consider that node as the root node of a Decision Tree which has the highest Information Gain.


| outlook | temp | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |


Calculations:-

Target variable play has 9 yes & 5 Nos.

$$\text{Entropy(play)} = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= -\frac{9}{14}\left[\frac{\log_{10}\left(\frac{9}{14}\right)}{\log_{10} 2}\right] - \frac{5}{14}\left[\frac{\log_{10}\left(\frac{5}{14}\right)}{\log_{10} 2}\right]$$

$$= -0.64\left[\frac{-0.19}{0.3010}\right] - 0.35\left[\frac{-0.44}{0.3010}\right]$$

$$= 0.4016 + 0.511$$

$$\simeq 0.94$$


| outlook | temp | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |


Values(outlook) = sunny, overcast, rainy

play = [9Y, 5No]

$\text{play}_{\text{sunny}} = [2Y, 3No]$

$\text{play}_{\text{overcast}} = [4Y, 0No]$

$\text{play}_{\text{rainy}} = [3Y, 2No]$

$\therefore IG(\text{play}, \text{outlook}) = \text{Entropy(play)} - \sum_{v \in [\text{sunny, overcast, rainy}]} \frac{|S_v|}{S} \text{Entropy}(S_v)$

$$= 0.94 - \frac{5}{14}\text{Entropy}(\text{play}_{\text{sunny}}) - \frac{4}{14}\text{Entropy}(\text{play}_{\text{overcast}}) - \frac{5}{14}\text{Entropy}(\text{play}_{\text{rainy}})$$

$$\text{Entropy}(\text{play}_{\text{sunny}}) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = +0.52 + 0.44 = +0.96$$

$$E(\text{play}_{\text{overcast}}) = 0$$

| outlook | temp | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |

$$E(Play_{rainy}) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log\left(\frac{2}{5}\right) = 0.44 + 0.52 = +0.96$$

$\therefore$ Information Gain(Play, Outlook)

$$= 0.94 - \frac{5}{14}(2 \times 0.96)$$

$$= 0.254$$

| outlook | temp | humidity | windy | play |
|---|---|---|---|---|
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |

Similarly we compute information gain for wind feature w.r.t target play

Values (wind) : FALSE, TRUE

play = [9Y, 5No]

$play_{FALSE}$ = [6Y, 2No]

$play_{TRUE}$ = [3Y, 3No]

$\therefore$ Information Gain = Entropy (Play) -

$$\sum_{v \in [True, false]} \frac{|S_v|}{S} \text{Entropy } S_v$$

$$= 0.94 - \frac{8}{14}\text{Entropy}(Play_{FALSE}) - \frac{6}{14}\text{Entropy}(Play_{TRUE})$$

$$= 0.94 - \frac{8}{14}(0.811) - \left(\frac{6}{14}\right)\cdot$$

$$= 0.048$$

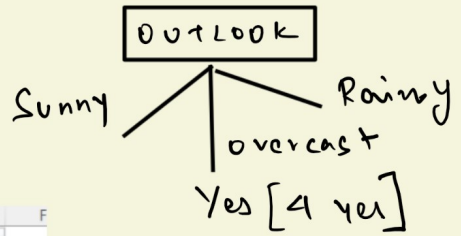
Similarly

I.G (Play, Temp) = 0.029

IG (Play, Humidity) = 0.152

Since outlook feature has the highest I.G, we choose it as rootnode.

$\therefore$ our decision Tree currently looks as below




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | outlook | temp | humidit | windy | play | |
| 2 | sunny | hot | high | FALSE | no | |
| 3 | sunny | hot | high | TRUE | no | |
| 9 | sunny | mild | high | FALSE | no | |
| 10 | sunny | cool | normal | FALSE | yes | |
| 12 | sunny | mild | normal | TRUE | yes | |
| 16 | | | | | | |
| 17 | | | | | | |

Next we consider this table on the left

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | outlook | temp | humidit | windy | play | |
| 2 | sunny | hot | high | FALSE | no | |
| 3 | sunny | hot | high | TRUE | no | |
| 9 | sunny | mild | high | FALSE | no | |
| 10 | sunny | cool | normal | FALSE | yes | |
| 12 | sunny | mild | normal | TRUE | yes | |
| 16 | | | | | | |
| 17 | | | | | | |

$$\text{Entropy}(\text{Sunny}) = -P_Y \log_2 P_Y - P_N \log_2 P_N$$
$$= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2(3/5)$$
$$= 0.971$$

Now we calculate the new I.G.

$$IG(\text{Sunny}, \text{Temp}) = E(\text{Sunny}) - \sum_{v \in [\text{hot}, \text{mild}, \text{cold}]} \frac{|S_v|}{S} \text{Entropy}(S_v)$$

Values (Temp) = hot, mild, cold

$\text{play} = [2Y, 3No]$

$\text{play}_{hot} = [0Y, 2No]$

$\text{play}_{mild} = [0Y, 1No]$

$\text{play}_{cold} = [2Y, 0No]$

$$\therefore IG(\text{Sunny}, \text{Temp}) = 0.971 - \frac{2}{5}\text{Entropy}(\text{Play}_{hot}) - \frac{1}{5}\text{Entropy}(\text{Play}_{mild}) - \frac{2}{5}\text{Entropy}(\text{Play}_{cold})$$
$$= 0.971 - 0.4 = 0.571$$

Similarly $IG(\text{Sunny}, \text{Humidity}) = 0.971$

$IG(\text{Sunny}, \text{Windy}) = 0.02$

Hence we choose Humidity as a node under Sunny.

Similarly we can find nodes under rainy

Hence currently our D.T looks like one below.