Decision Tree (ID3 Algorithm)

Ex

S.No	Outlook	Temp	Humidity	Windy	Play Tennis
1	Sunny	Hot	High	weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	weak	Yes
4	Rainy	Mild	High	weak	Yes
5	Rainy	Cool	Normal	weak	Yes
6	Rainy	cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	weak	No
9	Sunny	Cool	Normal	weak	Yes
10	Rainy	mild	Normal	weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	weak	Yes
14	Rainy	mild	High	Strong	No

→ A decision Tree is a Tree where each node represents a feature (attributes), each link (branch) represents a decision (rule) and each leaf represent an outcome.

→ ID3 algorithm uses * Entropy function
* Information Gain

→ How to choose attribute to be a root of Tree?
Ans The attribute that best classifies the training data.

* calculate Entropy

$$\text{Entropy} \Rightarrow \frac{-p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

* Average Information

$$I(\text{Attribute}) = \sum \frac{p_i + n_i}{p+n}\,\text{Entropy}(A)$$

* calculate Information Gain (Difference in Entropy before and after splitting dataset on A)

$$\text{Gain} = \text{Entropy}(S) - I(\text{Attribute})$$

p stands for positive output
n stands for negative output.

ID3 Algorithm

1. Compute the entropy for the dataset Entropy(S)
2. For every attribute / feature :-
   1. calculate entropy for all other values Entropy(
   2. Take average information Entropy
   3. calculat Gain
3. Pick the Highest Gain attribute
4. Repeat until we get desired Tree.

So In the given example, total outcome = 14

$P = 9$

$N = 5$

$$\text{Entropy} = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

Entropy (S) = 0.940

⇒ Now calculate entropy for outlook → Sunny, Rainy, Overcast

①

Outlook	Play Tennis
Sunny	No
Sunny	No
Sunny	No
Sunny	Yes
Sunny	Yes

②

Outlook	Play Tennis
Rainy	Yes
Rainy	Yes
Rainy	No
Rainy	Yes
Rainy	No

③

Outlook	Play Tennis
Overcast	Yes
Overcast	Yes
Overcast	Yes
Overcast	Yes

outlook	P	N	Entropy
Sunny	2	3	0.971
Rainy	3	2	0.971
overcast	4	0	0

$$I(\text{outlook}) = \frac{P_{sunny} + N_{sunny}}{P+n} * \text{Entropy (Sunny)} + \frac{P_{rainy} + N_{rainy}}{P+N} * \text{Entropy (Rainy)} + \frac{P_{overcast} + N_{overcast}}{P+N} * \text{Entropy (overcast)}$$

$$I(\text{outlook}) = \frac{3+2}{14} * 0.971 + \frac{2+3}{9+5} * 0.971 + \frac{4+0}{9+5} * 0$$

$$\Rightarrow 0.693$$

Calculate Gain = Entropy (S) − I(outlook)

Gain (outlook) ⇒ 0.940 − 0.693 = 0.247

⇒ Now calculate Entropy for Temperature

Temp	P	N	Entropy
Hot	2	2	1
mild	4	2	0.918
cool	3	1	0.811

I(Temp) = 0.911

Gain(Temp) = 0.940 − 0.911 = 0.029

⇒ Now calculate Entropy for Humidity

Humidity	P	N	Entropy
High	3	4	0.985
Normal	6	1	0.591

I(Humidity) = 0.788

Gain(Humidity) = 0.940 − 0.788 = 0.152

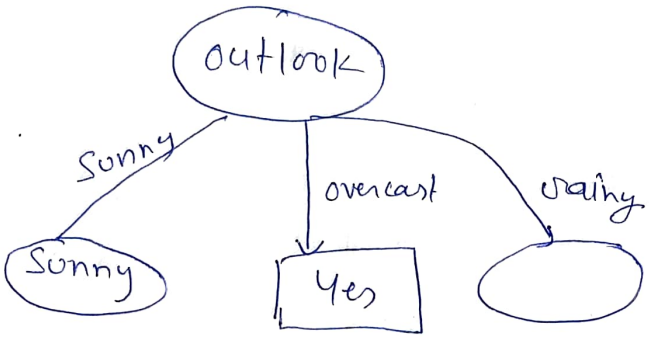
⇒ Now calculate Entropy for Windy

Windy	P	N	Entropy
Strong	3	3	1
Weak	6	2	0.811

I(Windy) = 0.892

Gain(Windy) = 0.940 − 0.892 = 0.048

Attribute	Gain
outlook	0.247
Temperature	0.029
Humidity	0.152
Windy	0.048



Now we have Two data sets.

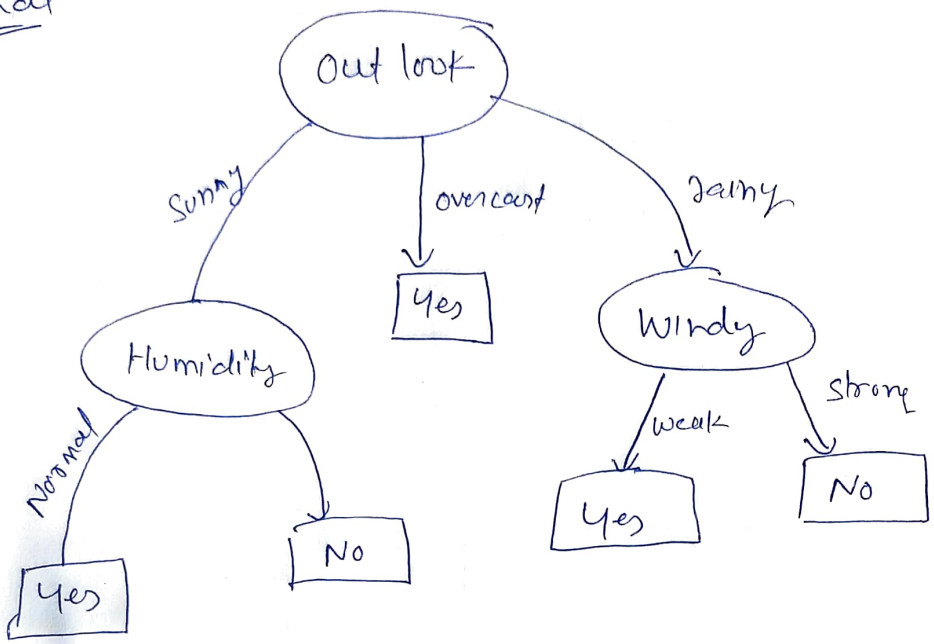
①

outlook	Temp	Humidity	Windy	Play Tennis
Sunny	Hot	High	Weak	No
Sunny	Hot	High	strong	No
Sunny	mild	High	Weak	No
Sunny	Cool	Normal	weak	Yes
Sunny	mild	Normal	Strong	Yes

②

Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Rainy	mild	Normal	weak	Yes
Rainy	mild	High	Strong	No

final


Out look
Sunny
overcast
Rainy
Humidity
Yes
Windy
Normal
High
weak
strong
Yes
No
Yes
No