ID3 :- It is one of the Most Common decision tree algo.

→ Dichomisation means dividing into two completely opposite things.

→ Algorithm iteratively divides attributes into two groups which are the most dominant attributes and others to construct a tree.

→ then, it calculates the Entropy and information gains of each attribute. In this way, the most dominant attribute can be found.

→ After then, the most dominant one is put on the tree as decision node.

→ Entropy and Gain scores would be calculated again among the other attributes.

→ Procedures continuous until reaching a decision for that branch.

→ Calculate the entropy of every attribute using the data set 'S'.

$$\text{Entropy}(S) = \sum -P(i) \cdot \log_2 P(I)$$

→ Split the set S into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or, equivalently, information gain is maximum)

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum [P(S/A) \cdot \text{Entropy}(S/A)]$$

→ Make a decision tree node containing that attribute

→ Recursive on subsets using remaining attributes.

Example

To go for outing or not

DAY	outlook	Temp	Humidity	wind	Decision
1	Sunny	HoT	High	weak	NO
2	Sunny	HoT	High	Strong	NO
3	overcast	HoT	High	weak	Yes
4	Rain	mild	High	weak	Y
5	Rain	cool	Normal	weak	Y
6	Rain	Cool	Normal	strong	N
7	overcast	cool	Normal	strong	Y
8	Sunny	Mild	High	weak	N
9	Sunny	Cool	Normal	weak	Y
10	Rain	mild	Normal	weak	Y
11	Sunny	Mild	Normal	Strong	Y
12	overcast	Mild	High	strong	Y
13	overcast	Hot	Normal	weak	Y
14	Rain	Mild	High	strong	NO

Data set 1

Calculate EntroPy :-

→ Decision colomn consists of 14 instances and includes two labels : "Yes" and "NO"

→ There are 9 decisions labelled "Yes" and
5 " " "NO".

$$Entropy(Decision) = -P(yes) * \log_2 P(yes) - P(NO) * \log_2 P(NO)$$

$$Entropy(Decision) = -(9/14) * \log_2(9/14) - (5/14) * \log_2(5/14)$$

$\therefore$ $Entropy(Decision) = 0.940$

Wind factor on decision

→ D

$$Gain(D, W) = Entropy(D) - \sum [P(D|W) * Entropy(D|W)]$$

→ wind attribute has two labels : weak & Strong.

$$Gain(D, W) = Entropy(D) - [P(D|W = weak) * Entropy(D|W = weak)] - [P(D|W = Strong) * Entropy(D|W = Strong)]$$

→ Now, we need to calculate (D|W = weak) and (D|W = Strong) respectively.

→

→ **Entropy(D|W = weak) =**

→ $-P(NO) * \log_2 P(NO) - P(yes) * \log_2 P(yes)$

$-(2/8) * \log_2(2/8) - (6/8) * \log_2(6/8)$

$\Rightarrow$ $0.811 = Entropy(D|W = weak)$

Day		Wind	Deci
1		weak	N
3		"	Y
4		"	Y
5		"	Y
8		"	N
9		"	Y
10		"	Y
13		"	Y

Entropy(D|W = Strong) $= -P(NO) * \log_2 P(NO) - P(yes) * \log_2 P(yes)$

$= -(3/6) * \log_2(3/6) - (3/6) * \log_2(3/6) = 1$

$\therefore$ $Entropy(D|W = Strong) = 1$

Day		Wind	Decis
2		Strong	No
6		"	No
7		"	Y
11		"	Y
12		"	Y
14		"	N

$\therefore$ Gain(D, W) = Entropy(D) − [P(D|W = weak) * entropy(D|W=weak)]

− [P(D|W = Strong) * Entropy(D|W = Strong)]

= 0.940 − [(8/14) * 0.811] − [(6/14) * 1]

= 0.048

<u>Other factors on Decision</u>

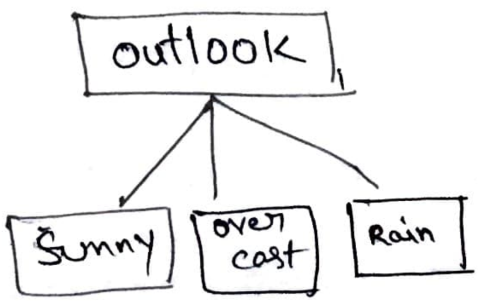
Applied Similar calculation on the other columns.

→ Gain (Decision, outlook) = 0.246

→ Gain (Decision, Temparature) = 0.029

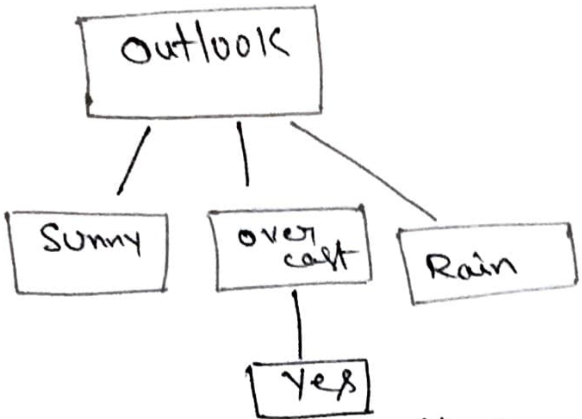
→ Gain (Decision, Humidity) = 0.151

⇒ outlook factor on decision produces the highest score. that's why, outlook decision will appear in the root node of the tree.



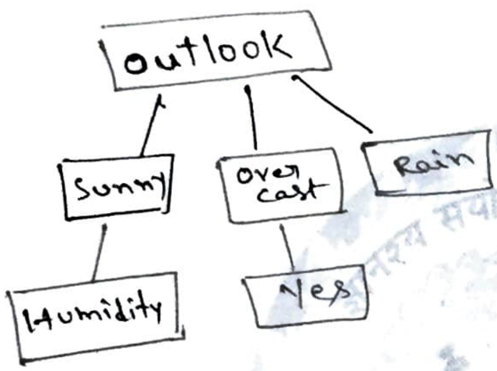
<u>overcast outlook on Decision</u>

Day	outlook	TEMP	Humidity	Wind	Decision
3	overcast	Hot	High	Weak	yes
7	"	cool	Normal	Strong	yes
12	"	Mild	High	Strong	yes
13	"	Hot	Normal	weak	yes



→ Decision will always be yes if outlook were overcast.

Sunny outlook decision :-



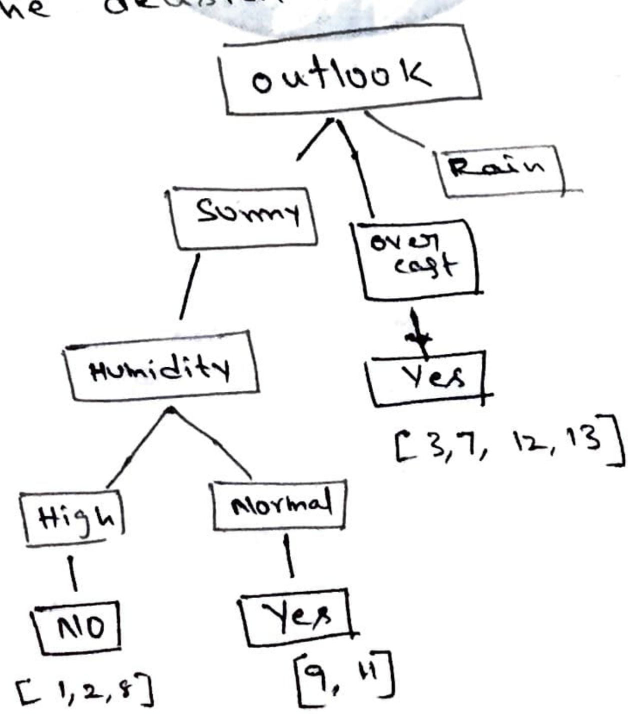
Day	outlook	Temp	Humi	Wind	De
1	Sunny	HOT	High	weak	NO
2	Sunny	HOT	High	Strong	NO
8	Sunny	Mild	High	weak	NO
9	Sunny	Cool	Normal	weak	Y
11	Sunny	Mild	Normal	Strong	Y

→ Gain (outlook = Sunny | Temparature) = 0.570

→ Gain (outlook = Sunny | Humidity) = 0.976

→ Gain (outlook = Sunny | Wind) = 0.019

Humidity is the decision.



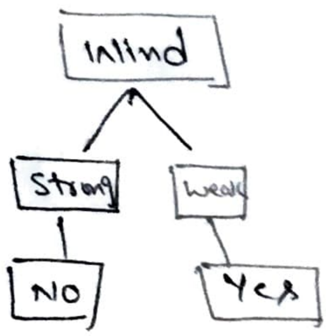
Rain outlook decision

Day	outlook	Temp	Humidity	wind	Decision
4	Rain	Mild	High	weak	Yes
5	"	cool	Normal	weak	Yes
6	"	cool	Normal	strong	NO
10	"	Mild	Normal	weak	Yes
14	"	mild	High	strong	NO

→ Gain (outlook = Rain | Temp) =

→ Gain (outlook = Rain | Humidity) =

→ Gain (outlook = Rain | Wind) = ✓



[6,14] [4,5,10]

∵ Wind produces the highest score if outlook were rain.

Complete Decision Tree