Decision Tree Classification Algorithm.

i) Information Gain :-

$$I(P,n) = -\frac{P}{P+n}\log_2 \frac{P}{P+n} - \frac{n}{P+n}\log_2 \frac{n}{P+n}$$

where,

P = no. of elements with class Yes.

n = no. of elements with class No.

ii) Entropy :-

$$E(A) = \sum_{i=1}^{V} \frac{P_i + n_i}{P+n} I(P_i, n_i)$$

iii) Gain

$$Gain(A) = I(P,n) - E(A)$$

example 1 :-

Suppose, we want to decide whether the weather is amenable to play baseball. The target classification is "should we play baseball ?" which can be yes or no.

The weather attributes are outlook, humidity, temparature & wind speed. They have following values.

outlook = {Sunny, overcast, rain}
temparature = {hot, mild, cool}
humidity = {high, normal}
wind = {weak, strong}.

ay	outlook	Temp	humidity	wind	play basebal
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

Table :- Training Data set.

Solution :-

Step 1 :- CALCULATION OF INFORMATION GAIN.

class P=(play baseball="yes")

class N=(play baseball="No)

Total no. of records = 14

∴ no. of records with "yes" class = 9

& no. of records with "No" class = 5.

So,

Information Gain is,

$$I(P,N) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

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

$$= 0.940.$$

2:- COMPUTE ~~ENTROPY~~ Gain for ALL ATTRIBUTES

a) calculate information gain for outlook.

For outlook = sunny.

P_i = with "yes" class = 2

n_i = with "No" class = 3

So,

$$I(P_i, n_i) = I(2,3) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)$$

$$= 0.971.$$

similarly for all values of ~~ent~~ outlook

$I(P_i, n_i)$ is calculated as,

outlook	P_i	n_i	$I(P_i, n_i)$
sunny	2	3	0.971
overcast	4	0	0
Rain	3	2	0.971

So, Entropy is calculated as,

$$E(A) = \sum_{i=1}^{v} \frac{P_i + n_i}{P + n} I(P_i, n_i)$$

$E(outlook) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2$

$= 0.694$

$\therefore gain(S, outlook) = I(P,N) - E(outlook)$

$= 0.940 - 0.694$

$= \boxed{0.246}$

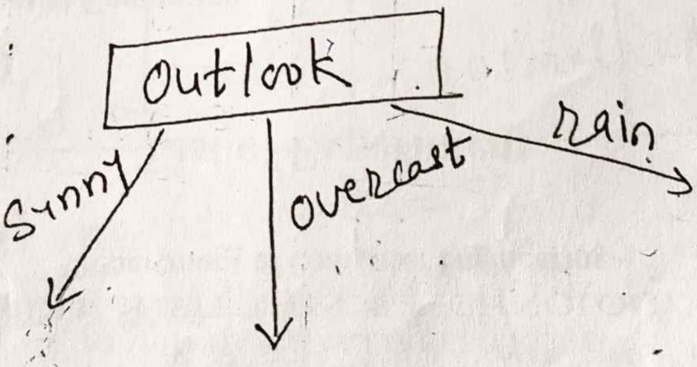
Similarly,

gain (S, Temparature) = 0.029.

gain (S, Humidity) = 0.151

gain (S, wind) = 0.048

As outlook attribute has highest gain, it is considered as root node with three branches (sunny, overcast, rain).



ep3 :-

As attribute outlook is considered as root node, we have to consider the remaining three attribute for Sunny branch node.

So, consider outlook = Sunny.

$S_{Sunny} = \{D1, D2, D8, D9, D11\}$
$= 5.$

Day	outlook	Temp	humidity	wind	play baseball
D1	Sunny	Hot	high	weak	No
D2	Sunny	Hot	high	strong	No
D8	Sunny	mild	high	week	No
D9	Sunny	cool	normal	weak	yes
D11	Sunny	mild	normal	strong	yes.

Total no. of tuples = 5

P = no. of tuples with 'yes' = 2

N = no. of tuples with 'No' = 3.

Information gain is calculated as,

$$I(P,n) = I(2,3)$$
$$= -(2/5)\log_2(2/5) - (3/5)\log_2(3/5)$$
$$= 0.971$$

calculate gain for all values of Temparature

Temparature	Pi	ni	I(Pi,ni)
hot	0	2	0
mild	1	1	1
cold	1	0	0

entropy for temparature,

$$E(temp) = \sum_{i=1}^{N} \frac{Pi+ni}{P+n} I(Pi,ni)$$

$$E(temparature) = 2/5 * I(0,2) + (2/5) * (1,1) + (1/5) * I(1,0)$$

$$= \boxed{0.4}$$

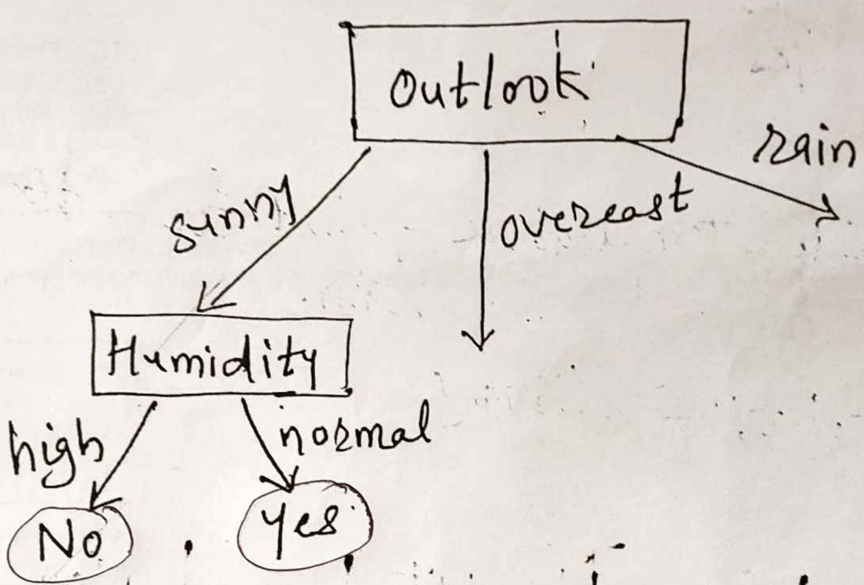
$$\therefore Gain(S_{sunny}, temparature) = I(P,n) - E(temp)$$
$$= 0.971 - 0.4$$
$$= 0.571$$

nilarly ;

gain (S_{sunny}, Humidity) = 0.971

gain (S_{sunny}, wind) = 0.02

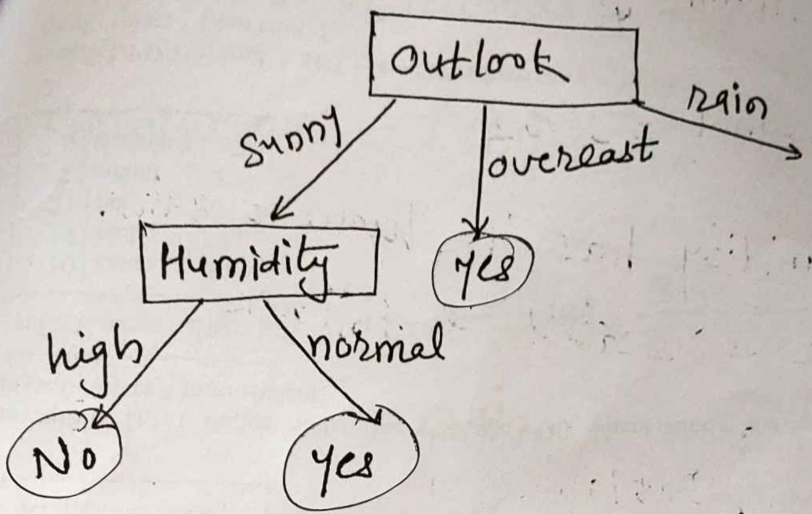
As Humidity has the highest gain, decision tree is created as follows,



step 4 :- Now consider temparature & wind for outlook = overcast.

Day	outlook	Temp	humidity	wind.	Play baseball
D3	overcast	hot	high	weak	yes
D7	overcast.	Cool	normal	strong	yes
D12	overcast	mild	high	strong	yes

Since for attribute temparature & wind assign class "yes" to overcast.



step 5 :-

Now, Consider temparature & wind for outlook= rain.

Srain = {D4, D5, D6, D10, D14}
= 5

Day	outlook	Temp	humidity	wind	play baseball
D4	Rain	mild	high	weak	Yes
D5	Rain	Cool	normal	weak	Yes
D6	Rain	Cool	normal	Strong	No
D10	Rain	mild	normal	weak	Yes
D14	Rain	~~cool~~ mild	high	strong	No

class P : play baseball = "yes" = 3

class N : play baseball = "No" = 2

Total No. of records = 5.

$$\text{Information gain} = I(P,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

$$\therefore I(3,2) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}.$$

$$= 0.970$$

Info. gain $I(P_i, n_i)$ for wind

wind	P_i	n_i	$I(P_i, n_i)$
weak	3	0	0
strong	0	2	0

Entropy for wind

$$E(\text{wind}) = \frac{3}{5} I(3,0) + \overset{2/5}{\cancel{2(5)}} I(0,2)$$

$$= 0$$

$$\text{Gain}(S_{rain}, \text{wind}) = I(P,n) - E(\text{wind})$$

$$= 0.970 - 0$$

$$= 0.970$$

Info. gain I(Pi, ni) for temparature

Temparature	Pi	ni	I(Pi, ni)
Hot	0	0	0
mild	2	1	0.918
cool	1	1	1

Calculate Entropy,

$$E(\text{temparature}) = \left(\tfrac{0}{5}\right) I(0,0) + \left(\tfrac{3}{5}\right) I(2,1) + \left(\tfrac{2}{5}\right) I(1,1)$$

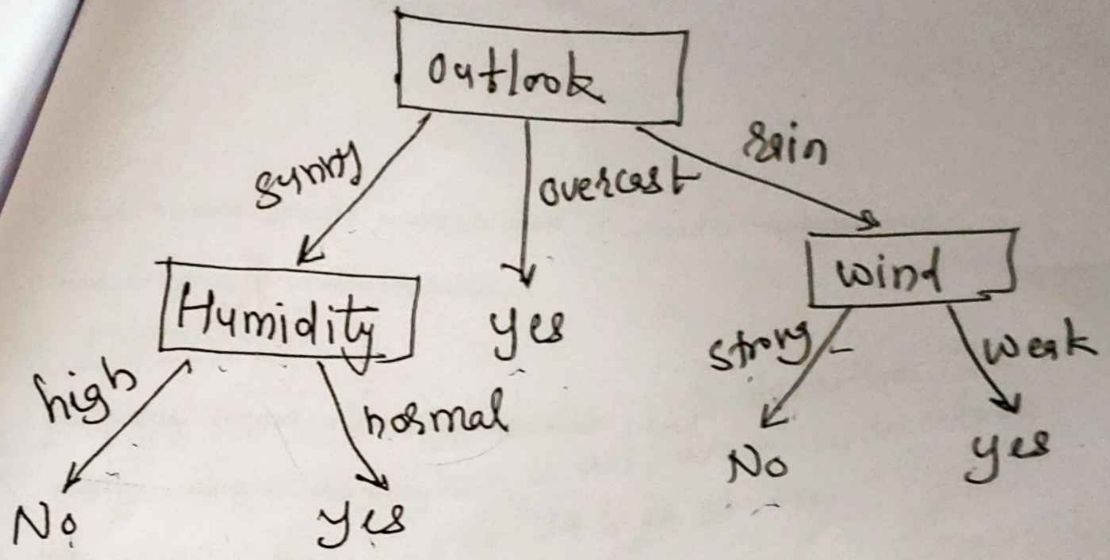
$$= 0.951$$

$\therefore$ Gain (Srain, temparature) = I (P.n) − E (temp

$$= 0.970 - 0.951$$

$$= 0.019$$

As wind has highest gain, it will be placed below outlook = rain



The decision tree can be represented in rule format as,

If outlook = sunny and humidity = high
then play baseball = yes.

If outlook = overcast then play baseball = yes

If outlook = rain and wind = strong
then play baseball = no

if outlook = rain and wind = weak
then play baseball = yes.

Accuracy By Class :-

TP rate	FP rate	Precision	Recall	F-measure	Roc Area	Class
$\frac{TN}{N}$	$\frac{FP}{P}$	$\frac{TN}{TN+FN}$	$\frac{TN}{N}$	$\frac{(2*Precision*recall)}{(Precision+recall)}$	$\frac{(TP/P)}{(FP/FP+TN)}$	No
$\frac{TP}{P}$	$\frac{FP}{N}$	$\frac{TP}{TP+FP}$	$\frac{TP}{P}$	$\frac{(2*Precision*recall)}{(Precision+recall)}$	$\frac{(TP/P)}{(FP/FP+TN)}$	Yes

Note: -

Calcutale everything in calculation and also solve the eg. using ID3 & match tree with your o/p: