MODULE 2

1. Give Decision trees for the following set of training examples

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
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

Solution :-

* $Entropy(S) \equiv -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

* Note,

(1) when all members of S belong to the same class then, $Entropy(S) = 0$

(2) If S contains an equal number of positive and negative examples then,

$Entropy(S) = 1$

* The first step is to find the topmost node of the decision tree. ID3 determines the information gain for each attribute, then selects the one with highest information gain.

⇒ Entropy of S : positive examples = 09
negative examples = 05

$$\text{Entropy}([9+, 5-]) = -\left(\frac{9}{14}\right)\log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right)\log_2\left(\frac{5}{14}\right)$$

$$= \underline{0.940}$$

⇒ Information gain of the attribute <u>Outlook</u> is calculated as.

Value (Outlook) = Sunny, Overcast, Rain

$S_{Sunny} \leftarrow [2+, 3-]$

$S_{Overcast} \leftarrow [4+, 0]$

$S_{Rain} \leftarrow [3+, 2-]$

$$\text{Gain}(S, \text{Outlook}) = \text{Entropy}(S) - \left[\left(\frac{5}{14}\right)\text{Entropy}(S_{Sunny}) + \left(\frac{4}{14}\right)\text{Entropy}(S_{Overcast}) + \left(\frac{5}{14}\right)\text{Entropy}(S_{Rain})\right]$$

* Entropy $(S_{sunny}) = -\left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right) - \left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right)$

$= -(0.4 * (-1.3219)) - (0.6 * (-0.7369))$

$= 0.52876 + 0.44214$

$= 0.970$

* Entropy $(S_{overcast}) = 0$ (because all members belong to same class)

* Entropy $(S_{rain}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)$

$= -(0.6 * (-0.7369) - (0.4 * (-1.3219))$

$= 0.44214 + 0.52876$

$= 0.970$

$= (0.940) - \left[\left(\frac{5}{14}\right) * 0.9709 + 0 + \left(\frac{5}{14}\right) 0.9709\right]$

$= 0.9409 - [0.3467 + 0.3467]$

$= \underline{0.246}$

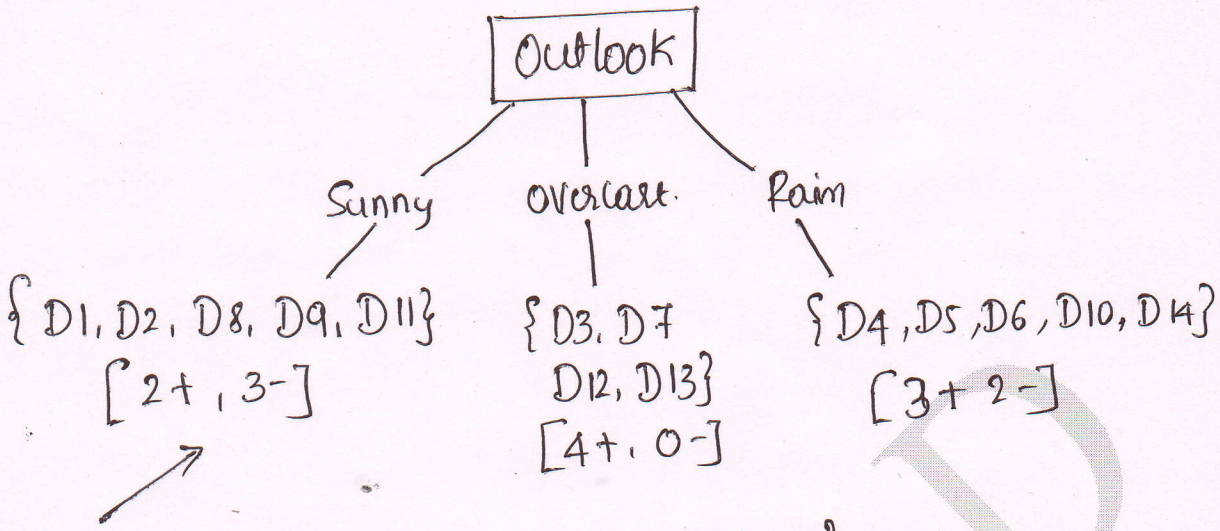
So Gain (S, outlook) = 0.246 ✓

Similarly Gain (S, Temperature) = 0.029

Gain (S, Humidity) = 0.151

Gain (S, wind) = 0.029

So root node will be <u>Outlook</u>



which attribute should be tested here?

$S_{sunny} = \{D1, D2, D8, D9, D11\}$

$Gain(Sunny, Humidity) = 0.970 - \left[\left(\frac{3}{5}\right) * 0.0 + \left(\frac{2}{5}\right) \cdot 0.0\right]$

$= \underline{0.970}$

$Gain(Sunny, Temperature) = 0.970 - \left[\left(\frac{2}{5}\right) * 0 + \left(\frac{2}{5}\right) * 1 + \left(\frac{1}{5}\right) \cdot 0\right]$

$= 0.970 - 0.4$

$= \underline{0.570}$

$Gain(Sunny, Wind) = 0.970 - \left[\left(\frac{3}{5}\right) * (0.918) + \left(\frac{2}{5}\right) \cdot 1\right]$

$= 0.970 - 0.9508$

$= \underline{0.0192}$

So, Attribute Humidity will be descendant node.

$\Rightarrow S_{Rain} = \{D_4, D_5, D_6, D_{10}, D_{14}\}$

$Entropy(S_{Rain}) = -\left(\frac{3}{5}\right)\log_2\left(\frac{3}{5}\right) - \left(\frac{2}{5}\right)\log_2\left(\frac{2}{5}\right)$

$= 0.970$

$Gain(Rain, Temperature) = 0.970 - \left[\left(\frac{0}{5}\right)*0 + \left(\frac{3}{5}\right)*0.918 + \left(\frac{2}{5}\right)*1\right]$

$= 0.0198$

$Gain(Rain, wind) = 0.970 - \left[\left(\frac{3}{5}\right)*0 + \left(\frac{2}{5}\right)*0\right]$

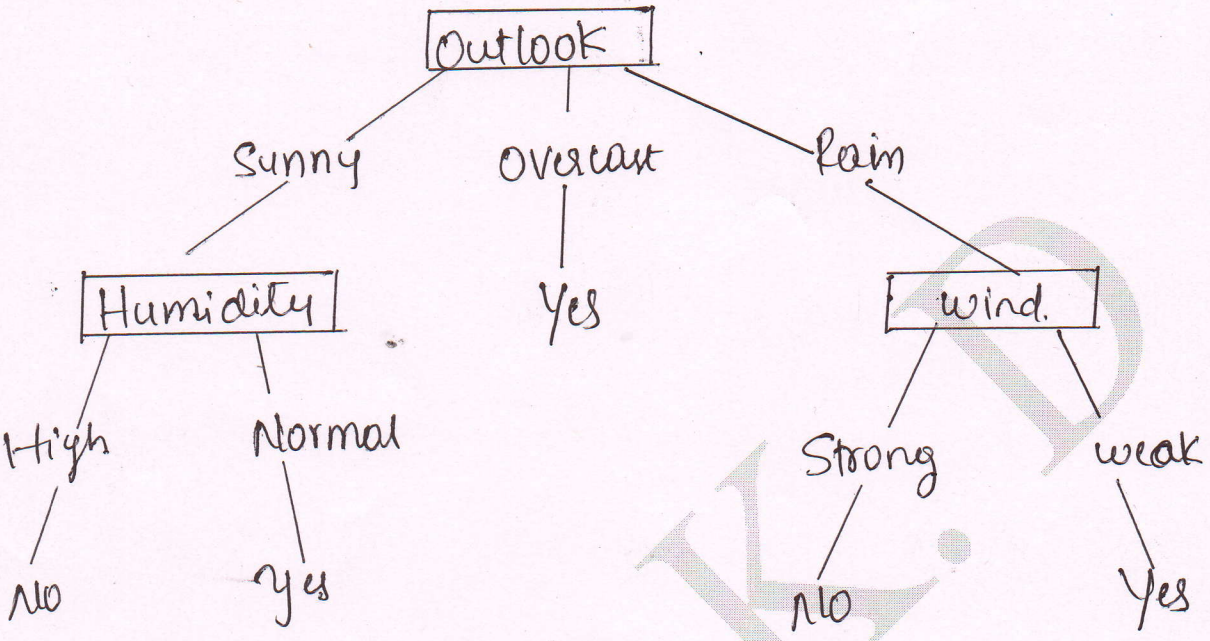
$= 0.970$

$Gain(Rain, Humidity) = 0.97 - \left[\left(\frac{2}{5}\right)*1 + \left(\frac{3}{5}\right)*0.917\right]$

$= 0.0198$

So, highest information gain is ~~code~~ attribute ~~Rain~~ Wind.

So, the final tree is.



2. Consider the following set of training examples.

a) What is the entropy of this collection of training example with respect to the target function classification?

b) What is the information gain of a_2 relative to these training examples?

Instance	Classification	a_1	a_2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

$\Rightarrow$ $Entropy(S) = -P_{\oplus} \log_2 P_{\oplus} - P_{\ominus} \log_2 P_{\ominus}$

Information gain

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

a) positive instances = 03

Negative instances = 03

$$Entropy(S) = -\left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right) - \left(\frac{3}{6}\right) \log_2 \left(\frac{3}{6}\right)$$

$Entropy(S) = 1$

when there are equal number of +ve & -ve instances, then $Entropy(S) = 1$.

b) $Gain(S, a_2) = Entropy(S) - \left[\frac{4}{6} Entropy(S_T) + \frac{2}{6} Entropy(S_F)\right]$

find

① $Entropy(S_T) = -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right)$

$= \underline{1}$

② $Entropy(S_F) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right)$

$= \underline{1}$

$\Rightarrow Gain(S, a_2) = 1 - \left[\frac{4}{6}(1) + \frac{2}{6}(1)\right]$

$= 1 - 1$

$Gain(S, a_2) = \underline{0}$

3. Give decision trees to represent the following Boolean functions.

i) A && ¬B

ii) A V [B && C]

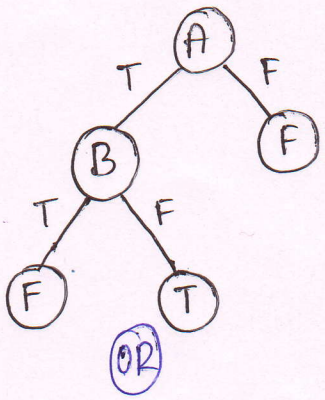
iii) A XOR B

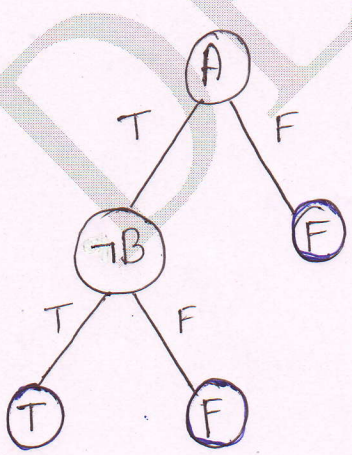
iv) [A && B] V [C && D]

Solution

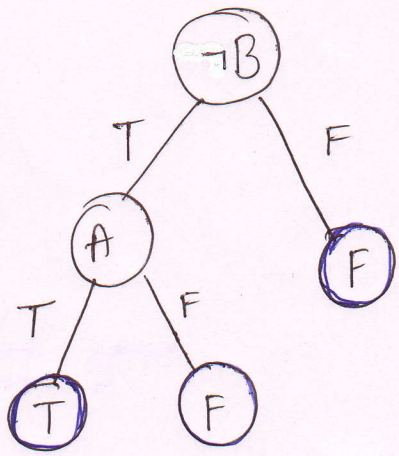
i) A && ¬B

A	B	¬B	A && ¬B
T	T	F	F (-)
T	F	T	T (+)
F	T	F	F (-)
F	F	T	F (-)



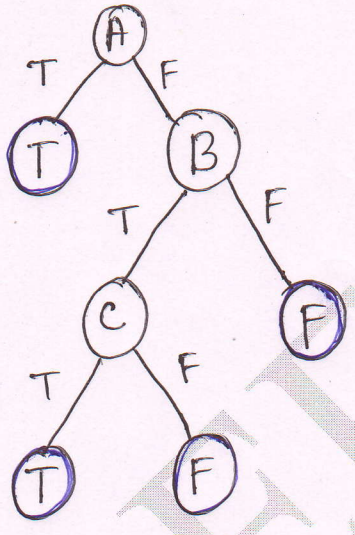


OR

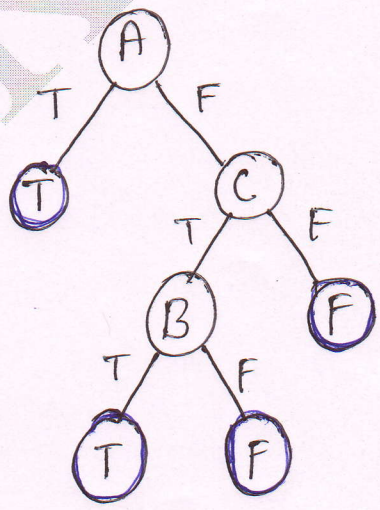


ii) A V [B && C]

A	B	C	B && C	A V [B && C]
T	T	T	T	T
T	T	F	F	T
T	F	T	F	T
T	F	F	F	T
F	T	T	T	T
F	T	F	F	F
F	F	T	F	F
F	F	F	F	F

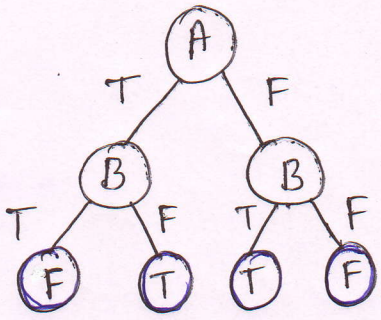


OR.



iii) A XOR B

A	B	A XOR B
T	T	F
T	F	T
F	T	T
F	F	F



iv) [A && B] v [C && D]

A	B	C	D	A && B	C && D	(A && B) v (C && D)
T	T	T	T	T	T	T
T	T	T	F	T	F	T
T	T	F	T	T	F	T
T	T	F	F	T	F	T
T	F	T	T	F	T	T
T	F	T	F	F	F	F
T	F	F	T	F	F	F
T	F	F	F	F	F	F
F	T	T	T	F	T	T
F	T	T	F	F	F	F
F	T	F	T	F	F	F
F	T	F	F	F	F	F
F	F	T	T	F	T	T
F	F	T	F	F	F	F
F	F	F	T	F	F	F
F	F	F	F	F	F	F