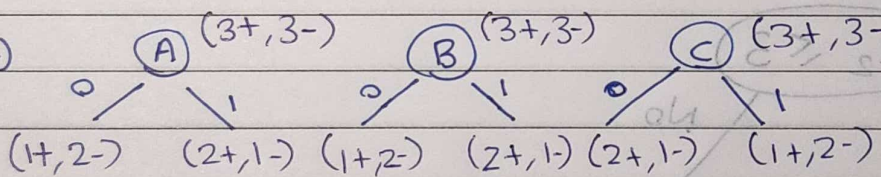
HOMEWORK-II

1.) The training dataset given is as follows

A	B	C	Y
0	1	0	Yes
1	0	1	Yes
0	0	0	No
1	0	1	No
0	1	1	No
1	1	0	Yes

(a.) The given dataset has instances with same values of A, B, C but different outcomes. Thus, a decision tree with a 100% accuracy on this training set cannot be formed, as no form of tree would satisfy both the instances

(b.)



$$\text{Entropy}(S) = -\left(P(Y=0)\log_2 P(Y=0) + P(Y=1)\log_2 P(Y=1)\right)$$

$$= -\frac{3}{6}\log_2 1/2 \; -\frac{1}{2}\log_2 1/2 = 1$$

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} P(v)\, Entropy(S_v)$$

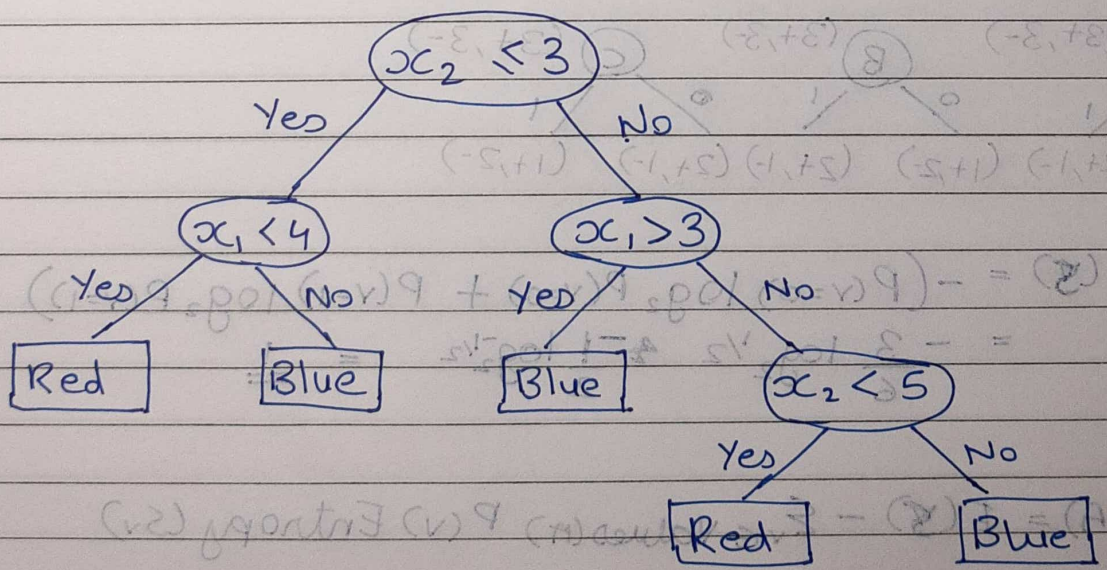
$$Gain(S,A) = 1 - \left[\frac{3}{6} E(1,2) + \frac{3}{6} E(2,1)\right] = 0.545$$

$$Gain(S,B) = 1 - \left[\frac{3}{6} E(1,2) + \frac{3}{6} E(2,1)\right] = 0.545$$

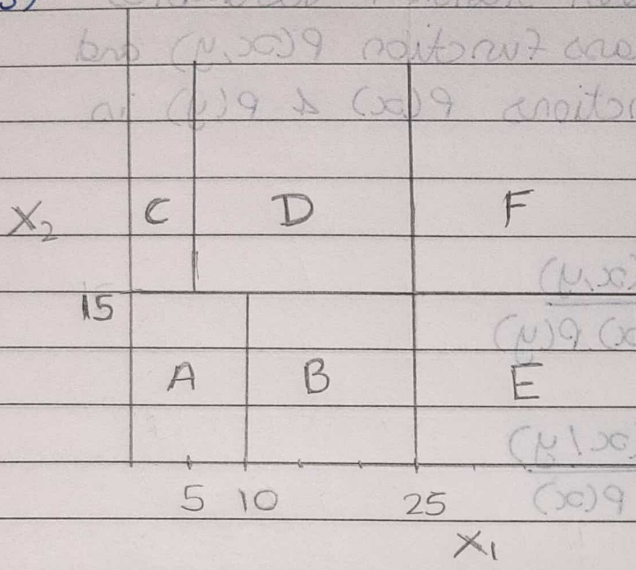
$$Gain(S,C) = 1 - \left[\frac{3}{6} E(2,1) + \frac{3}{6} E(1,2)\right] = 0.545$$

The attributes A, B and C all have equal information gain. Each attribute splits the training dataset in (3+,3-) into two: (1+,2-) & (2+,1-)

2) Decision tree for the given decision boundary is



3.)



4.) Let,

$T_P \rightarrow$ Test results are positive

$D_P \rightarrow$ Patient has disease

$T_n \rightarrow$ Test results are negative

$D_n \rightarrow$ Patient does not have disease

$P(T_P|D_P) = 0.99$

$P(T_n|D_P) = 1 - P(T_P|D_P)$

$= 0.01$

$P(D_P) = \frac{1}{10000} = 0.0001$

$P(T_n|D_n) = 0.99$

$P(T_P|D_n) = 1 - P(T_n|D_n)$

$= 0.01$

$P(D_n) = 1 - P(D_P) = 0.9999$

We need,

$$P(D_P|T_P) = \frac{P(T_P|D_P) \cdot P(D_P)}{P(T_P)}$$

$$= \frac{(0.99)(0.0001)}{0.010098}$$

$$P(T_P) = P(T_P|D_P) \cdot P(D_P) + P(T_P|D_n) \cdot P(D_n)$$

$$= (0.99)(0.0001) + (0.01)(0.9999)$$

$$= 0.000099 + 0.009999$$

$$= 0.010098$$

$$\boxed{P(D_P|T_P) = 0.00980}$$

5.) The mutual information between random variables X & Y with joint probability mass function $P(x,y)$ and marginal probability mass functions $P(x)$ & $P(y)$ is defined as

$$I(x,y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x,y)}{P(x).P(y)}$$

$$= \sum_{x \in X} \sum_{y \in Y} P(x,y) \log \frac{P(x/y)}{P(x)}$$

$$= -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(x) + \sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(x/y)$$

$$= H(X) - H(X|Y)$$

OR

$$H(Y) - H(Y|X)$$

We can infer that X says as much as about Y as Y says about X.

Thus,

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$