DELTA Page No. Date / /

ASSIGNMENT - 4

Sol: 1 For x_1,

- Entropy at root node = $-\text{sum}(P(y)\log P(y))$

$= -(0.5 \log 0.5) + (0.5 \log 0.5)$

$= 1.$

- Weighted Entropy = $\frac{3}{4}\left(\left(-\frac{2}{3}\log\frac{2}{3}\right)+\left(-\frac{1}{3}\log\frac{1}{3}\right)\right)$

$+ \frac{1}{4}\left((-1\log 1)+(-0\log 0)\right)$.

$\Rightarrow 0.688$

- Gain = Entropy at root node - Weighted entropy

$= 1 - 0.688$

$= 0.312 \quad = (G)$

Split info (g) = $-P(T)\log P(T) - P(F)\log P(F)$

$= \left(-\frac{3}{4}\log\frac{3}{4}\right)+\left(-\frac{1}{4}\log\frac{1}{4}\right)$

$= 0.811$

Gain ratio = Gain / Split info $\{F, x_1\}$

$= 0.312 / 0.811$

$= 0.3847 \longrightarrow (1)$.

- For x_2,

- Entropy at root node = $-\text{sum}(P(y)\log P(y))$

$$= -\left(\frac{1}{2}\log\frac{1}{2}\right) + \left(-\frac{1}{2}\log\frac{1}{2}\right)$$

$$= \underline{1}$$

- Weighted entropy = $\frac{1}{2}\left(\left(-\frac{1}{2}\log\frac{1}{2}\right) + \left(-\frac{1}{2}\log\frac{1}{2}\right)\right)$

$$+ \frac{1}{2}\left(\left(-\frac{1}{2}\log\frac{1}{2}\right) + \left(-\frac{1}{2}\log\frac{1}{2}\right)\right)$$

$$\Rightarrow \underline{1}.$$

- Gain(G) = Entropy at root node − weighted entropy

$$= 1 - 1$$

$$= \underline{0}.$$

- Split info (g) = $-P(T)\log P(T) - P(F)\log P(F)$

$$= -\left(\frac{1}{2}\log\frac{1}{2}\right) - \left(\frac{1}{2}\log\frac{1}{2}\right)$$

$$= \underline{1}.$$

- Gain ratio = Gain / split info

$$= \underline{0}, \longrightarrow (2).$$

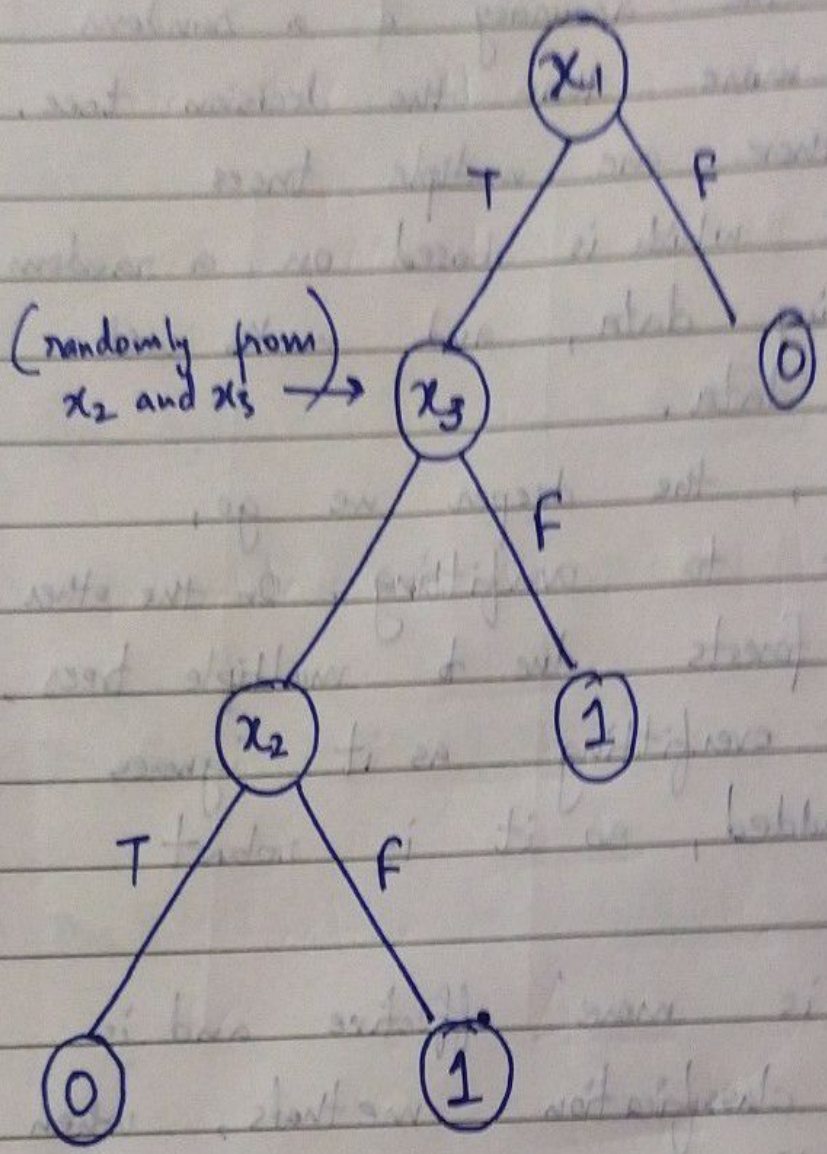
For x_2:

- Entropy at root node = $-\sum(P(s)\log P(s))$
  $= 1$
- Weighted entropy = $\frac{1}{2}\left(\left(-\frac{1}{2}\log\frac{1}{2}\right)+\left(-\frac{1}{2}\log\frac{1}{2}\right)\right)$
  $+\frac{1}{2}\left(\left(-\frac{1}{2}\log\frac{1}{2}\right)+\left(-\frac{1}{2}\log\frac{1}{2}\right)\right)$
  $= 1.$
- Gain = Entropy at root node − weighted entropy
  $= 1 - 1 = 0.$
- Split info (g) = $-P(T)\log P(T) - P(F)\log P(F)$
  $= -\left(\frac{1}{2}\log\frac{1}{2}\right) - \left(\frac{1}{2}\log\frac{1}{2}\right)$
  $= 1.$
- Gain ratio = Gain / split info (g)
  $= 0/1 = 0 \longrightarrow (3).$

So, from (1), (2) and (3), we take the maximum gain ratio for the starting node.

So, starting node is $\boxed{x_1}$

4.



Decision Tree

Sol:2 (i) Decision Tree accuracy : 92%.

(ii) Random forest accuracy : 98%.

(iii) We can see that the accuracy of a random forest classifier is more than the decision tree. This is because there are multiple trees in a random forest which is based on a random sample of the training data, and works well with non-linear data.

In Decision Trees, the deeper we go, it is more prone to overfitting. On the other hand, in random forests due to multiple trees, it is less prone to overfitting as it ignores errors or biases added, so it is robust to outliers.

So, random forest is more effective and is faster than other classification methods, when there are many features in the model, where the importance across is more balanced.

We can also see that the random forest is able to identify 100% : true positives (recall), while decision tree can find only 95%.