ASSIGNMENT 10

SRIHARI CHANDRAM OULI
1001529776

TASK 1

a) 80 people decide to wait
20 people decided not to wait

$K = K_1 + K_2$

$K = 80 + 20 = 100$

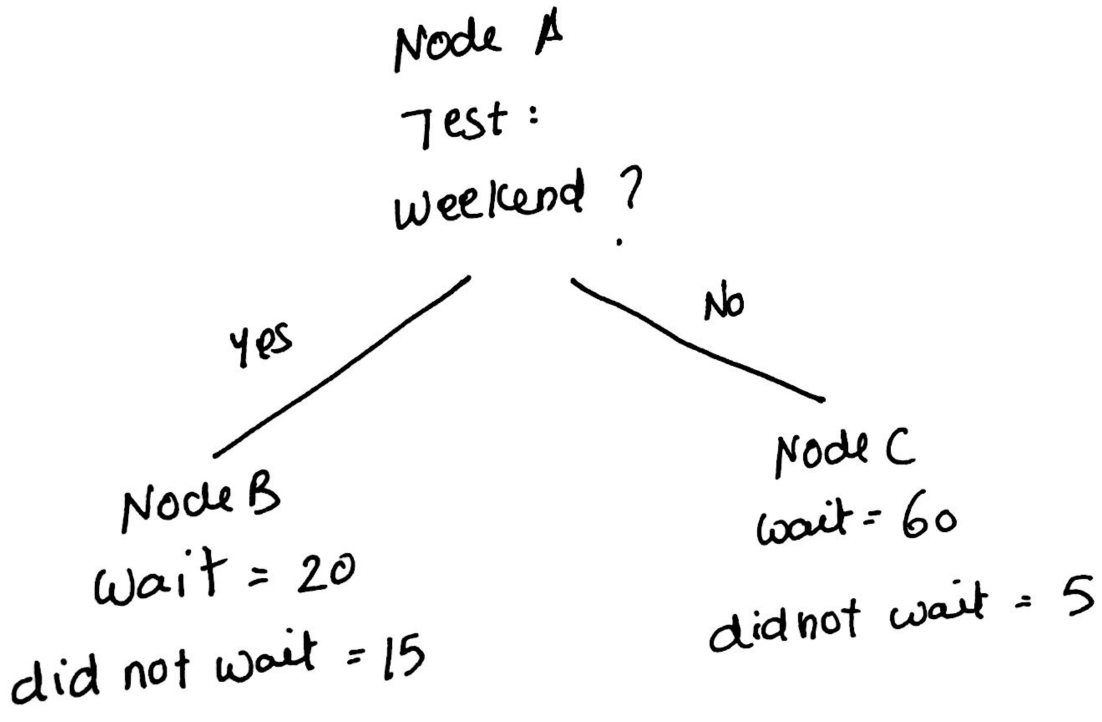
Initial Entropy at Node A is:

$$H\left(\frac{80}{100}, \frac{20}{100}\right) = \frac{-80}{100}\log_2\frac{80}{100} - \frac{20}{100}\log_2\frac{20}{100}$$

$$= 0.2575 + 0.46438$$

$$= \boxed{0.72188}$$

b)



* Let us find entropy at Node B :

$$H\left(\frac{20}{35}, \frac{15}{35}\right) \quad \text{where } K = 20+15$$
$$= 35$$

$$= \frac{-20}{35}\log_2\frac{20}{35} - \frac{15}{35}\log_2\frac{15}{35}$$

$$= \frac{-4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}$$

$$= 0.4613 + 0.5238 = \boxed{0.9852}$$

* Lets find entropy at C :

$K = 60 + 5 = 65$

$$H\left(\frac{60}{65}, \frac{5}{65}\right) = -\frac{60}{65}\log_2\frac{60}{65} - \frac{5}{65}\log_2\frac{5}{65}$$

$$= -\frac{12}{13}\log\frac{12}{13} - \frac{1}{13}\log_2\frac{1}{13}$$

$$= 0.10659 + 0.28460$$

$$= \boxed{0.3912}$$

* Information Gain

$$I(E, L) = H(E) - \sum_{i=1}^{L}\frac{K_i}{K}H(E_i)$$

$$0.72188 - \frac{35}{100}(0.9852) - \frac{65}{100}(0.3912)$$

P.T.O

$= \boxed{0.12275}$

c) As far as from Node A, it is a weekend At E, since it uses the exact test whether its a weekend or not, all attributes will be weekend and Node I will have no attributes.

Information gain:

$$H(E) - \sum_{i=1}^{k} H(E_i)$$

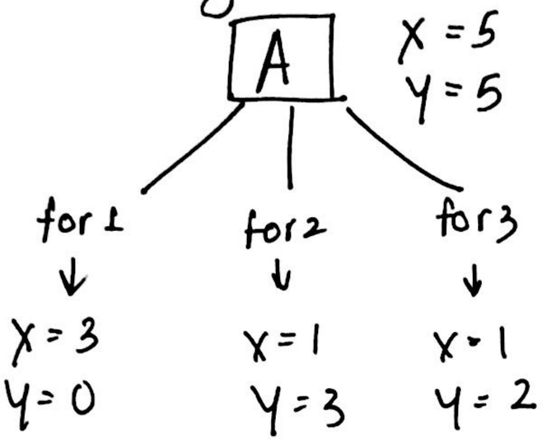
$$= 7H(E) - H(E) = 0$$

d) Since Tuesday is not a weekend and it is rainy, by looking at the graph, the leaf node will end up in $\boxed{\text{Node F}}$. Output is that he will wait.

e) According to the question, the leaf node will end up in $\boxed{\text{Node H}}$. The output for that case is that he will not wait.

TASK 2

* Checking for root as A



$$Root = \frac{-5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$$

$$= 0.5 + 0.5 = 1$$

for 1

$$H\left(\frac{3}{3}, \frac{0}{3}\right) = \frac{-3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log\frac{0}{3} = \boxed{0}$$

for 2

$$H\left(\frac{1}{4}, \frac{3}{4}\right) = \frac{-1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

$$= (0.25 \times 2) + (0.75 \times 0.415) = \boxed{0.81125}$$

for 3

$$H\left(\frac{1}{3}, \frac{2}{3}\right) = \frac{-1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.533 + 0.3851$$

$$\Rightarrow \boxed{0.9183}$$

* Checking for root as B

$$Root = \frac{-5}{10}\log\frac{5}{10} - \frac{5}{10}\log\frac{5}{10} = 1$$

for 1

$$H\left(\frac{1}{4},\frac{3}{4}\right) = \frac{-1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$$

$$= \boxed{0.81125}$$

for 2

$$H\left(\frac{3}{4},\frac{1}{4}\right) = \frac{-3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= \boxed{0.81125}$$

for 3

$$H\left(\frac{1}{2},\frac{1}{2}\right) = \frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}$$

$$= \boxed{1}$$

* Checking for root as C

Root = 1

for 1

$$H\left(\frac{1}{5},\frac{4}{5}\right) = \frac{-1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$$

$$= \boxed{0.7218}$$

for 2

$$H\left(\frac{3}{4},\frac{1}{4}\right) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$$

$$= \boxed{0.81125}$$

for 3

$$H\left(\frac{1}{1},\frac{0}{1}\right) = -1\log_2 1 = \boxed{0}$$

* Information gain

1. A as root

$$1 - 0 - \frac{4}{10}(0.81125) - \frac{3}{10}(0.9183)$$

$$= 1 - 0.3245 - 0.27549 = \boxed{0.4001}$$

2. B as root

$$1 - \frac{4}{10}(0.81125) - \frac{4}{10}(0.81125) - \frac{2}{10}(1)$$

$$= 1 - 0.3245 - 0.3245 - 0.2 = \boxed{0.151}$$

3. C as root

$$1 - \frac{5}{10}(0.7218) - \frac{4}{10}(0.81125) - 0$$

$$= 1 - 0.3609 - 0.3245$$

$$= \boxed{0.3146}$$

Therefore, A has the most information gain i.e A is the root.

TASK 3

Part a

$A = 1000$, $B = 0$, $C = 0$, $D = 0$

Lowest Entropy $= -\frac{1000}{1000} \log_2 \frac{1000}{1000} = \boxed{0}$

Highest Entropy

A = 250 , B = 250 , C = 250 , D = 250

$$\Rightarrow -\frac{250}{1000}\log_2\frac{250}{1000} - \frac{250}{1000}\log_2\frac{250}{1000} - \frac{250}{1000}\log_2\frac{250}{1000} - \frac{250}{1000}\log_2\frac{250}{1000}$$

$$= 4 \times 0.5 = \boxed{2}$$

Part b

* Highest Info. gain

A = 250 , B = 250 , C = 250 , D = 250

$$I.G = \left[H\left(\frac{250}{1000}, \frac{250}{1000}, \frac{250}{1000}, \frac{250}{1000}\right) - \left[\sum_{i=1}^{L} H\left(\frac{250}{250}, \frac{0}{250}, \frac{0}{250}, \frac{0}{250}\right)\right]\frac{250}{1000}\right]$$

$$\Rightarrow 2 - 0 - 0 - 0 - 0$$

$$= \boxed{2}$$

* Lowest Information Gain

$$IG = \left[H\left(\frac{1000}{1000}, \frac{0}{1000}, \frac{0}{1000}, \frac{0}{1000}\right) - \sum_{i=1}^{L} H(E_i)\right]$$

$$= \boxed{0}$$

TASK 4

We will be able to achieve the task by reversing all the outputs. If we reverse, we get an accuracy of 72%.

Therefore, we can guarantee achieving better than 60% accuracy.

TASK 5

In Decision Tree concept, largest possible number of elements for X is 2^{2^n}.

In our case, we have 5 boolean variables, thus, the largest possible number of elements for X will be:

$$\boxed{2^{2^5} = 2^{32}}$$