Name- Yunika Upadhayaya
ID - 1001631183

Assignment-5

Task-1

Part-a :

Let

K_1 be the number of people decided to wait = 65

K_2 be the number of people decided not to wait = 35

$K = K_1 + K_2 = 65+35 = 100$

We know that,

entropy is given by:

$$H\left(\frac{K_1}{K}, \frac{K_2}{K}\right) = \sum_{i=1}^{N} -\frac{K_i}{K}\left(\log_2 \frac{K_i}{K}\right)$$

Computing initial entropy,

$$H(A) = \left(-\frac{65}{100}\log_2\frac{65}{100}\right) + \left(-\frac{35}{100}\log_2\frac{35}{100}\right)$$

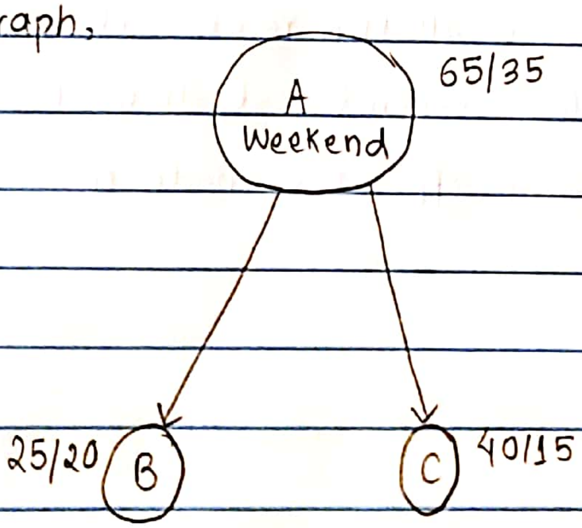
$$= 0.934 \ //$$

Part-b:

Information Gain is given by,

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

From the graph,



We know,

$H(A) = 0.934$

$$H(B) = -\frac{25}{45}\log_2\frac{25}{45} - \frac{20}{45}\log_2\frac{20}{45}$$

$$= 0.991$$

$$H(C) = -\frac{40}{55}\log_2\frac{40}{55} - \frac{15}{55}\log_2\left(\frac{15}{55}\right)$$

$$= 0.845$$

Here,

$$\text{Information Gain} = \text{Entropy}(A) - \frac{45}{100} \times \text{Entropy}(B) - \frac{55}{100} \times \text{Entropy}(C)$$

$$= 0.934 - \frac{45}{100} \times 0.99 - \frac{55}{100} \times 0.845$$

$$= 0.0233$$

Part-c:

The information gain at Node E of using the weekend test will be 0 because all the records at node E will be classified as "Yes" for the weekend test. It is because weekend test has already been applied at node A.

Part-d :

→ This test case will end up at Node D.

→ The decision tree output is class will wait.

Task-2

Class $X = 5$

Class $Y = 5$

Total $= 5+5 = 10$

Entropy at root $= -\frac{5}{10}\log_2\left(\frac{5}{10}\right) - \frac{5}{10}\log_2\left(\frac{5}{10}\right)$

$= 1$

Calculation for attribute A :

Attribute A

Class	1	2	3
X	3	1	1
y	0	3	2
Total:	3	4	3

Entropy (1) $= -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3}$

$= 0$

Entropy (2) $= -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4}$

$= 0.81124$

Entropy (3) $= -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) = 0.9182$

$$\text{Information Gain for } A = 1 - \frac{3}{10} \times 0 - \frac{4}{10} \times 0.81127 - \frac{3}{10} \times 0.9182$$

$$= 0.4.$$

Calculation for Attribute B:

	B		
Class	1	2	3
X	1	3	1
Y	3	1	1
Total	4	4	2

$$\text{Entropy}(1) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.8113$$

$$\text{Entropy}(2) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113$$

$$\text{Entropy}(3) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$\text{Information gain at } B = 1 - \frac{4}{10} \times 0.8113 - \frac{4}{10} \times 0.8113 - \frac{2}{10} \times 1$$

$$= 0.151$$

Calculation for attribute C:

	C		
Class	1	2	3
X	1	3	1
Y	4	1	0
Total :	5	4	1

$$\text{Entropy}(1) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) = 0.7219$$

$$\text{Entropy}(2) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.8113$$

$$\text{Entropy}(3) = -\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right) = 0$$

$$\text{Information gain at } C = 1 - \frac{5}{10}\times 0.7219 - \frac{4}{10}\times 0.8113 - \frac{1}{10}\times 0$$

$$= 0.3145$$

$\therefore$ Attribute A achieves the highest information gain at root which is 0.400 //

Task-3

Part-a:

↳ Highest possible entropy at node N is 2 when 1000 training examples are equally distributed among class A, B, C, D.

↳ Lowest possible entropy at node N is 0 when all the training examples belongs to one particular class

Part-b:

(2)

↳ Highest possible information gain can be same as the entropy of parent node N when we choose an attribute K. It happens when the entropy of each children node is 0. i.e. all the records in each children node belongs to one class.

↳ Lowest possible information gain can be 0 when the entropy of each children node is 1. It is possible when records in each children node is equally distributed among all the classes (A, B, C, D).

Task-4

↳ The accuracy of current classifier is 28%. We can achieve (100-28) = 72% accurate binary classifier by just swapping the class output of the classifier.

Yes, I can guarantee achieving better than 60% accuracy.