Date ___/___/___

Assigment 10

Task 1

1001571136

Part (a) Entropy:

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

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

$$= -0.8\log(0.8) - .2\log(.2)$$

$$= -0.8 \times (-0.322) - (0.2)(-2.322)$$

$$= 0.2575 + 0.4643$$

$$\boxed{H(A) = 0.7218}$$

Part. (b) Info Gain

$$\text{Info Gain} = H(A) - \frac{35}{100} \times H\left(\frac{20}{35}, \frac{15}{35}\right) - \frac{65}{100} \times H\left(\frac{5}{65}, \frac{60}{65}\right)$$

$$= 0.7218 - \frac{35}{100} \times \left(-\frac{20}{35}\log_2\left(\frac{20}{35}\right) - \frac{15}{35}\log_2\left(\frac{15}{35}\right)\right) - \frac{65}{100} \times \left(-\frac{60}{65}\log\left(\frac{60}{65}\right) - \frac{5}{65}\log\left(\frac{5}{65}\right)\right)$$

$$= 0.7218 - \left(0.35 \times \left(-0.571 \times -0.80735\right) - (0.4285 \times -1.2239\right) - 0.65(-0.923 \times (-0.1154) - 0.0769 \times (-3.700)))$$

$$= 0.7218 - 0.344 - 0.2543$$

$$= 0.7218 - 0.5990$$

$$\boxed{\text{Info Gain} = 0.1229}$$

Part(C) The Info Gain would be 0, it's repeated, hence no change would be observed.

$A \to C \to F$

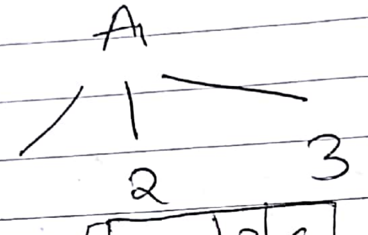
Part(D). leaf node : F
Output : will wait

$A \to B \to E \to H$

Part (E) leaf node : H
Output : will not wait

Task 2 Considering 1st case with $\boxed{\text{Rootnode} = A}$

①



Class	A=1	B	C
X		2	1
X		3	3
X		2	2

Total=3
② X = 3
y = 0

Cla	A=2	B	C
X		1	2
Y		1	1
Y		1	2
Y		1	1

Total=4
③ X=1
y=3

A = 3	B	C	Class
	2	2	X
	1	1	Y
	3	1	Y

Total=3
④ X=1
y=2

① $H(E)=1$

② $H(E_1)=0$

③ $H(E_2) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)$

$= 0.5 + 0.31127$

$= 0.81127$

④ $H(E_3) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(2/3\right)$

$= 0.5278 - 0.3956$

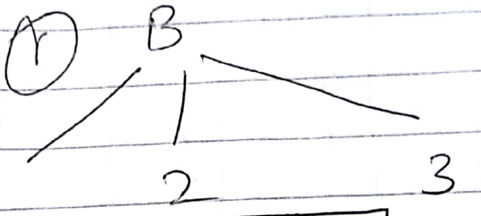
$= 0.9234$

(3) ~~E~~ Information Gain(A) $= H(E) - \frac{4}{10}(H(E_2)) - \frac{3}{10}H(E_3)$

$= 1 - 0.4 \times 0.811 - 0.3 \times 0.9234$

$= 0.39858$

Considering 2nd Case w.r.t : Rootnode = B



A

~~B~~=1	~~B~~	C	Class
	3	2	x
	2	1	y
	3	1	y
	2	1	y

(2) Total=4
x=1
y=3

~~B~~=2	B	C	Class
	1	1	x
	3	2	x
	1	2	x
	2	2	y

(3) Total=4
x=3
y=1

A

~~B~~=3	~~B~~	C	Class
	1	3	x
	3	1	y

Total=2
x=1
y=1

(4)

(1) $H(E)=1$

(2) $H(E_1) = -\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right) = 0.81127$

(3) $H(E_2) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.81127$

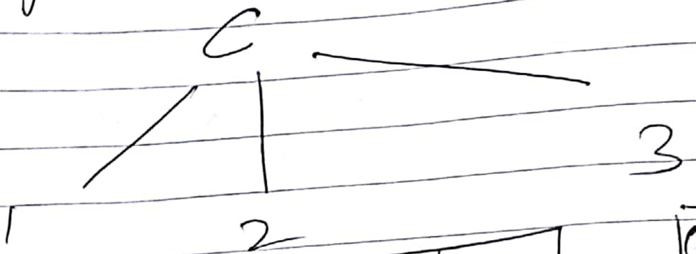
(4) $H(E_3) = 1$

Information Gain(B) $= 1 - \frac{4}{10}H(E_1) - \frac{4}{10}H(E_2) - \frac{2}{10}H(E_3)$

$= \cancel{E_1}\,1 - 0.4 \times 0.811 - 0.4 \times 0.811 - 0.2 \times 1$

$= 0.1512$

Considering 3rd case with Rootnode = C



C = 1	A	B	Class
	1	2	X
	2	1	Y
	3	1	Y
	2	3	Y
	2	1	Y

Total = 5
X = 1
y = 4

C = 2	A	B	Class
	2	1	X
	3	2	X
	1	2	X
	2	2	Y

Total = 4
X = 3
y = 1

C=3	A	B	Class
	1	0	X

Total = 1
X = 1

$H(E) = 1$

$H(E_1) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)$

$= -0.2\log_2(0.2) - 0.8\log_2(0.8)$

$= 0.4643 + 0.2575$

$=$ ~~0.7128~~ $= 0.7218$

$H(E_2) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\Big)\log_2\left(\frac{1}{4}\right)$

= 0.8[illegible]1127

$H(E_3) = 0$

Information Gain(C) $= H(E) - \frac{5}{10} H(E_1) - \frac{4}{10} H(E_2) - \frac{1}{10} H(E_3)$

$= 1 - \left(\frac{1}{2} \times 0.7218\right) - \left(\frac{2}{5} \times 0.81127\right)$

$= 1 - 0.3609 - 0.3245$

$= 0.3146$

Since, Info Gain (A) > Info Gain(C) > Info Gain(B)

0.39858 > 0.3146 > 0.1512

Hence, A receives the highest Info Gain

Task3 (a) Highest entropy is incase of even distribution.

A B C D

i.e 1000 → 250, 250, 250, 250

Hence Entropy would be $= H(E_A) + H(E_B) + H(E_C) + H(E_D)$

$= 4 \times \left(-\frac{250}{1000} \times \log_2\left(\frac{250}{1000}\right)\right)$

$= -4 \times 0.25 \times \log_2 0.25$

$= 2$

lowest entropy would mean, putting all the examples in a single class. i.e 1000 → 1000

Hence, Entropy would be $= H(E_A)$

$= -\frac{1000}{1000} \times \log_2\left(\frac{1000}{1000}\right)$

$= 0$

Task3

(b) Highest possible Entropy is 2

Lowest possible Entropy is 0

Task 4

We can improve the classifier depending on true cases in the data set. i.e we would need a more vibrant and diverse data set.

Completely depends on the Data Set. We cannot guarantee 60%. It would depend on the training Data.

Task 5

The total no. of distinct decision trees with n boolean attributes is equal to the distinct truth table with 2^n rows will be 2^{2^n}

Hence, with 5 boolean variables $= 2^{2^5}$

$= 2^{32}$

$= 4294967296$