Question NO 1:-

As the training set given to us is.

C1	C2	C3	out
1	1	1	1
1	0	0	1
1	1	0	0
0	0	1	0

As we know entropy is given by

$$E(s) = \sum_{i=1}^{k} - pi \log_2 pi$$

so $$E(c1) = \left[\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3}\right]\frac{3}{4} + \frac{1}{4}\left[-1\log 1\right]$$

$$= (0.918)\,\frac{3}{4} + 0 = 0.688$$

$$E(c2) = \frac{2}{4}\left[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right] + \frac{2}{4}\left[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\right.$$

$$= \frac{2}{4}[1] + \frac{2}{4}[1] = \frac{4}{4}$$

$$= 1.$$

$$E(c_3) = \frac{2}{4}\left[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right] + \frac{2}{4}\left[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right]$$

$$= \frac{2}{4}[1] + \frac{2}{4}(1)$$

$$= 1.$$

For total Entropy

$$E(out) = \left[\frac{2}{4}\log\frac{2}{4}\right] + \left[\frac{-2}{4}\log\frac{2}{4}\right].$$

$$= 1$$

$$IG(out, c_1) = E(out) - E(c_1)$$

$$= 1 - 0.688 = 0.3112.$$

$$IG(out, c_2) = E(out) - E(c_2)$$

$$= 1 - 1 = 0$$

$$IG(out, c_3) = E(out) - E(c_3)$$

$$= 1 - 1 = 0$$

As gain of C1 is the best & thus the highest Information gain so we will select this.

$$E(C2) = [-1/2 \log 1/2 - 1/2 \log 1/2]\,2/3 +$$

$$[-1 \log 1]\,1/3$$

$$= 0.666\ldots$$

$$E(C3) = 2/3\,[-1/2 \log 1/2 - 1/2 \log 1/2] + 1/3[-1 \log$$

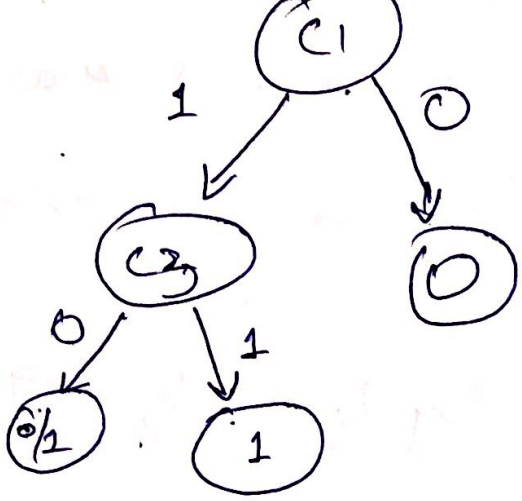
$$= 0.666\ldots$$

So our root node would be having the max info gain which is C1 & the next attribute we could choose any C2 or C3 as they have equa entropy. As if we have C1 = 0 then.

$$E(C2) = -1 \log 1 = 0$$

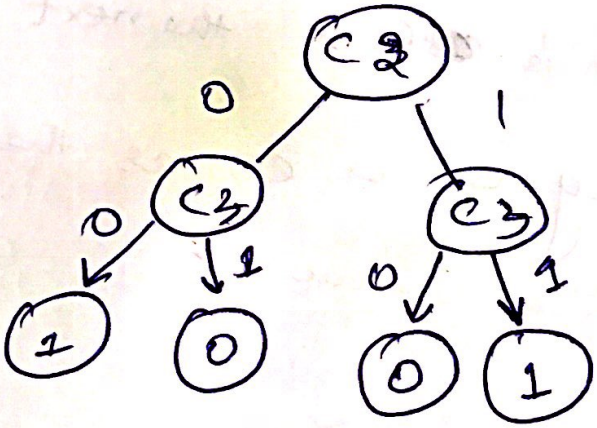
$$E(C3) = -1 \log 1 = 0.$$

So A-T-G-C



It will give us error in a case when, C1 is 1 & C3 is zero then is 1/4 probability that will occur in training data. So this is the training error that could occur.

6) Following is the decision tree of depth 2.



The above tree has training error zero as it satisfies all the given conditions in our data. & has depth 2 as well.

QUESTION NO 2:-

As we know we will use ID3(entropy) Algorithm.

$$E(S) = -p \log_2 p.$$

$$E(S) = -8/16 \log_2 8/16 - 8/16 \log_2 8/16$$

$$= 1$$

$$E(col) = 8/16 [-2/8 \log 2/8 - 6/8 \log 6/8]$$

$$+ 8/16 [-5/8 \log 5/8 - 3/8 \log 3/8]$$

$$= 8/6 (0.9) + 8/16 (0.8)$$

$$= 0.82.$$

$$E(size) = 8/16 [-3/8 \log 3/8 - 5/8 \log 5/8] +$$

$$8/16 [-6/8 \log 6/8 - 2/8 \log 2/8].$$

$$= 0.89.$$

$$E(Act.) = 8/16 [-6/8 \log 6/8 - 2/8 \log 2/8] +$$

$$[-3/8 \log 3/8 - 5/8 \log 5/8] 8/16. = 0.8$$

$$E(Age) = 8/16 [-3/8 \log 3/8 - 5/8 \log 5/8] + 8/16 [-2/8 \log 2/8 + 6/8 \log$$

$$= 0.8$$

$$IG(S, Col) = 1 - 0.8$$

$$= 0.12$$

$$IG(S, Size) = 1 - 0.8$$

$$= 0.12$$

$$IG(S, Act) = 1 - 0.8$$

$$= 0.12.$$

$$IG(S, Age) = 1 - 0.8$$

$$= 0.12.$$

As Information gain in all cases are equal so we can take any of it.

we are choosing size as our attribute - when size is small.

$$E(col) = 4/8\,[-1\log 1] + 4/8\,[-3/4 \log 3/4 - 1/4 \log 1/4]$$

$$= 0.405. \quad -(i)$$

$$E(Act) = 4/8\,[-2/4 \log 2/4 - 2/4 \log 2/4] + 4/8\,[-3/4 \log 3/4 - 1/4 \log 1/4]$$

$$= 0.965 \quad -(ii)$$

$$E(Age) = 4/8\,[-2/4 \log 2/4 - 2/4 \log 2/4] + 4/8\,[-3/4 \log 3/4 - 1/4 \log 1/4]$$

$$= 0.905. \quad -(iii)$$

As now we can see that eq i i.e. $E(col)$ has the smallest value so we will select this.

Now we will do the process for when we have size large.

$$E(Age) = 4/8(-1 \log 1) + 4/8[-2/4 \log 2/4 - 2/4 \log 2/4]$$

$$= 0.5$$

$$E(Col) = 4/8[-3/4 \log 3/4 - 1/4 \log 1/4] +$$

$$4/8[-1/4 \log 1/4 - 3/4 \log 3/4]$$

$$= 0.81$$

$$E(Act) = 4/8(-1 \log 1) + 4/8(-2/4 \log 2/4 - 2/4 \log 2/4$$

$$= 0.5$$

As we have 2 options here i.e. Age & Act as both have same value so we can select any but let us choose age as attribute.

col = purple

Age = Adult.

$E(Act) = -1 \log 1 = 0.$

$E(Size) = \frac{2}{4}[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}] + \frac{2}{4}[\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}].$

$= 1.$

size = Large

Col = yellow.

$E(Act) = \frac{2}{4}(-1\log 1) + \frac{2}{4}[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}].$

$= 0$ S.

$E(Age) = \frac{2}{4}(-1\log 1) + \frac{2}{4}[-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}].$

$= 0$ S.

color = purple

Age = child.

$E(Act) = -1\log 1 = 0.$

$E(Size) = -1\log 1 = 0.$

size = small

Col = yellow.

$E(Act) = 0$.

$E(Age) = 0$.

So from the above we can easily draw.

3 - level Decision tree for baloon classificatio