Name : Riddha Majumder

ID : 23341023

Sec : 13

Course : CSE 422

Assignment : 04

Ans to the question no: 1

(a)

Dog_id is not significant for determining whether the dog has a disease, as it is not a feature describing the dog's characteristics. It's just a label

(b)

Here, lets calculate

$E(\text{Disease}) = -P(\text{yes}) \log_2 P(\text{yes}) - (\text{No}) \log_2 P(\text{No})$

$= -4/8 \log_2 4/8 - 4/8 \log_2 4/8 \Rightarrow 1$

$E(\text{Black}) = -P(Y|B) \log_2 P(Y|B) - P(N|B) \log_2 P(N|B)$

$\Rightarrow -\frac{3}{5} \log_2 3/5 - 3/5 \log_2 3/5 \Rightarrow 0.970$

$E(\text{White}) = -P(Y|W) \log_2 P(Y|W) - P(N|W) \log_2 P(N|W)$

$\Rightarrow -\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \Rightarrow 0.918$

$\therefore IG(\text{Fur colour}) = \text{Entropy}(D) - P(B).E(B) - P(W).E(W)$

$= 1 - 5/8 \times 0.970 - 3/8 \times 0.918 \Rightarrow 0.0495$

$E(\text{large}) = -P(Y|L) \log_2 P(Y|L) - P(N|L) \log_2 P(N|L)$

$= -1/4 \log_2 1/4 - 3/4 \log_2 3/4 \Rightarrow 0.811$

$E(\text{small}) = -P(Y|S) \log_2 P(Y|S) - P(N|S) \log_2 P(N|S)$

$= -3/4 \log_2 3/4 - 1/4 \log_2 1/4 \Rightarrow 0.811$

$IG(\text{size}) \Rightarrow E(D) - P(L) E(L) - P(S) E(S)$

$\Rightarrow 1 - 4/8 \times 0.811 - 4/8 \times 0.811 \Rightarrow 0.189$

∴ The tail length in sorted order will be

1.2, 1.4, 2.2, 2.3, 3.5, 3.8, 4.2, 5.6 [n=8]

Now, we calculate the median $\Rightarrow \frac{(8/2)th + (8/2+1)th}{2}$

$\Rightarrow 2.3 + 3.5/2$

$\Rightarrow 2.9$

As we can see, the value is less than 2.9. So, it will be short, or else long

Dog_ID	Fur color	size	Tail_length	disease
1	Black	Large	Long	NO
28	white	Large	short	Yes
3	Black	small	Long	yes
34	Black	small	Long	yes
26	Black	Large	short	NO
11	White	small	short	NO
32	Black	small	Long	yes
13	white	Large	short	NO

Now, $E(long) \Rightarrow -P(y|l)\log_2 P(y|l) - P(N|l)\log_2 P(N|l)$

$\Rightarrow -\frac{3}{4}\log_2 3/4 - 1/4 \log_2 1/4 \Rightarrow 0.811$

$E(short) \Rightarrow -1/4 \log_2(1/4) - 3/4 \log_2(3/4) \Rightarrow 0.811$

∴ $IG(Tail\ length) \Rightarrow E(D) - P(L)E(L) - P(S)E(S)$

$\Rightarrow 1 - 0.811 \times 4/8 - 0.811 \times 4/8$

$\Rightarrow 0.189$

So, we got IG's. Now, we can say by looking them, two features among three has high information gain and equal. So, I will say, size is the best suited for the root node.

(c)

Given, $n = 8$

∴ Maximum possible value for entropy will be $H(X) = 3$ and this only happens when all the possible outcomes have equal probability that is 1/8.

$$\therefore \max H(X) = \sum_{1}^{8} -1/8 \log_2 1/8$$

$$= 3$$

(Ans:)