Q.4 (a.) Total entropy (Emotion)

→ total instances of
S: 4
H: 5
Total: 9

$$\text{Entropy(Emotion)} = \frac{-5}{9}\log_2\frac{5}{9} - \frac{4}{9}\log_2\frac{4}{9}$$
$$= 0.4711 + 0.52$$
$$= 0.9911$$

(b.) Chi-square of Number of rings

Contingency Table: (observed)

		S (4/9)	H (5/9)	Total
(8/9)	no of rings 0	4	4	8
(1/9)	no of rings 3	0	1	1
	Total	4	5	9

Expected values:
no of rings 0 & S: $4/9 \times 8/9 \times 9 = 3.5556$
no of rings 0 & H: $5/9 \times 8/9 \times 9 = 4.4444$
no of rings 3 & S: $4/9 \times 1/9 \times 9 = 0.4444$
no of rings 3 & H: $5/9 \times 1/9 \times 9 = 0.5556$

$$\chi^2 = \frac{(4-3.5556)^2}{3.5556} + \frac{(4-4.4444)^2}{4.4444} + \frac{(0-0.4444)^2}{0.4444} + \frac{(1-0.5558)^2}{0.5558}$$
$$\Rightarrow 0.0575 + 0.0444 + 0.4444 + 0.3575$$
$$\Rightarrow 0.8998$$

(f) Information Gain

(i) Colour

$S_G = [0^{+ve}\ 3^{-ve}]$ $\quad -\frac{3}{3}\log_2\frac{3}{3} = 0$

$S_B = [1^{+ve}\ 1^{-ve}]$ $\quad 1$

$S_R = [4^{+ve}\ 0^{-ve}]$ $\quad -\frac{4}{4}\log_2\frac{4}{4} = 0$

$E(S, \text{colour}) = \text{Total Entropy} - \text{Entropy (colour)}$

$= 0.9911 - \frac{2}{9} \times 1$

$= 0.7689$

(ii) Contact lens

$S_Y = [1^{+ve}\ 1^{-ve}] = 1$

$S_N = [4^{+ve}\ 3^{-ve}] = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7}$

$= 0.4613 + 0.5239$

$= 0.9852$

$E(S, \text{lens}) = \text{Total Entropy} - \text{Entropy (lens)}$

$= 0.9911 - \left[\frac{2}{9} \times 1 + \frac{7}{9} \times 0.9852\right]$

$= 0.0026$

(iii) No of rings

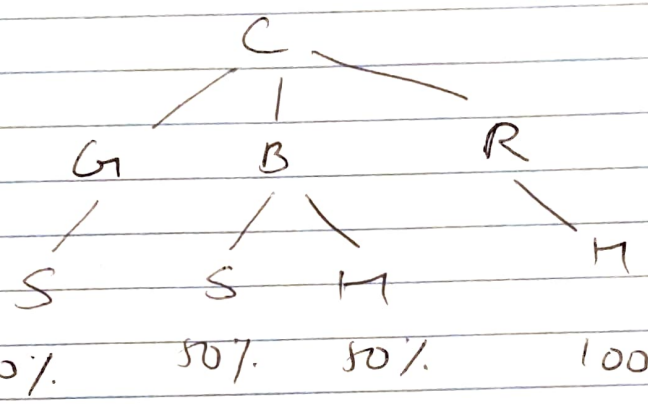
$S_0 = [4^{+ve}, 4^{-ve}] = 1$

$S_3 = [1^{+ve}, 0^{-ve}] = 0$

$$E(S, \text{avg}) = 0.9911 - \left[\frac{8}{9} \times 1\right]$$
$$= 0.1022$$

The maximum ~~entropy~~ attribute to contribute is Color using gain method.

(d.)



(e.) For accuracy,

$$\frac{\text{Total true cases}}{\text{Total cases}} \times 100$$

$$\Rightarrow \frac{8}{9} \times 100$$

$\Rightarrow$ 88.89% for the model.

(f) By looking at data, for last Row, the occurance of 3 rings for color R has only 1 occurence as "H" and cannot be used to generalize a general pattern. Alternatively this can be labelled as an outlier. To improve, dataset can be increased by increasing more sample and can use cross validation to evaluate model's performance even further.

Q.5 (a.) $Y_i = \beta_0 + \varepsilon_i$

$$y_i = \frac{1+(-1)+1}{3} = 1/3 = 0.33$$

(b) (i) $Y_i = \beta_1 x_i + \varepsilon_i$

$$\beta_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$x = [-1, 0, 2]$, $y = [1, -1, 1]$

$$\bar{x} = \frac{-1+0+2}{3} = 1/3 \Rightarrow \bar{y} = \frac{1}{3}$$

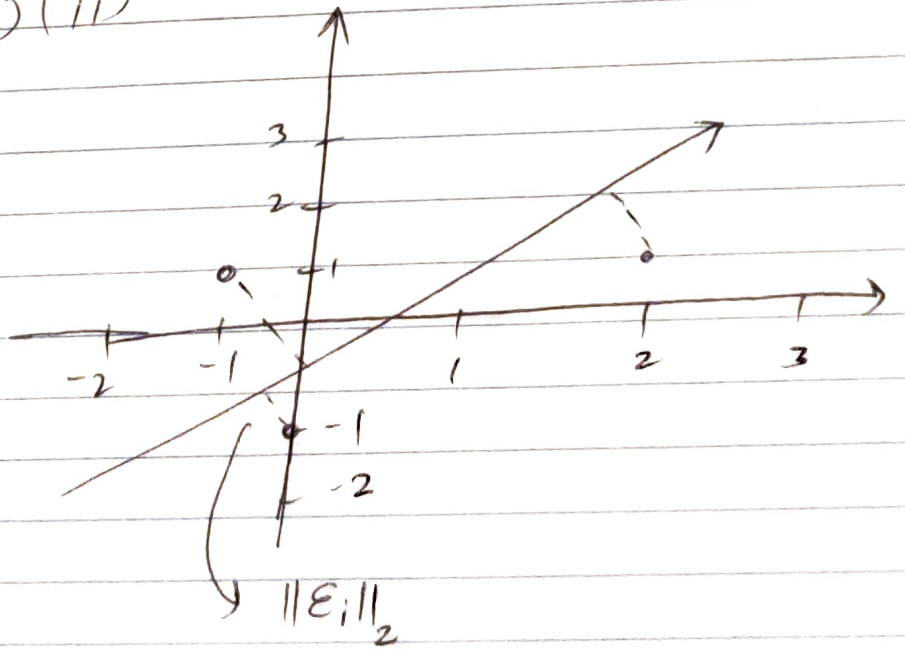
$$\Rightarrow \left(-1-\frac{1}{3}\right)\left(1-\frac{1}{3}\right) + \left(0-\frac{1}{3}\right)\left(-1-\frac{1}{3}\right) + \left(2-\frac{1}{3}\right)\left(1-\frac{1}{3}\right)$$

$$= \left(-\frac{4}{3}\right)\left(\frac{2}{3}\right) + \left(\frac{-1}{3}\right)\left(\frac{-4}{3}\right) + \left(\frac{5}{3}\right)\left(\frac{2}{3}\right) = \frac{6}{9} = \frac{2}{3} \quad (1)$$

$$\left(-1-\frac{1}{3}\right)^2 + \left(0-\frac{1}{3}\right)^2 + \left(2-\frac{1}{3}\right)^2 = \frac{42}{9} = \frac{14}{3} \quad (2)$$

$$\beta_1 = \frac{2/3}{14/3} = \frac{2}{3} \times \frac{3}{14} = 0.1428$$

Q.5 (b) (ii)



Q.6

True positive = 30
True negative = 30
False Positive = 10
False Negative = 30

(a) Sensitivity = $\frac{TP}{TP+FN} = \frac{30}{30+30} = \frac{30}{60} = 50\% \text{ or } 0.5$

(b) Specificity = $\frac{TN}{TN+FP} = \frac{30}{30+10} = \frac{30}{40} = 0.75 \text{ or } 75\%$

(c) FDR = $\frac{FP}{FP+TP} = \frac{10}{10+30} = \frac{10}{40} = 25\% \text{ or } 0.25$

(d) The test -ve for Cancer in the model can be reduced or Test +ve in No Cancer can be reduced. The methodology of taking test can be increased so that the probability of reducing false negative is high. We can recommend to take cancer test twice or thrice to get better result.

Q.7. (a.)

For a binary classification, a target variable can be either 0 or 1 ~~Yes~~, or Yes or No. The maximum training error for decision Tree that any dataset could have will then be 0.5 or 50%. Meaning when tree makes the prediction wrong for all instances in training set.

(b)

x_1	x_2	Y
0	0	1
0	1	1
1	0	0
1	1	0
0	0	1
0	1	1
1	0	0
1	1	0