Group F6

Francisco Silva, 110409
Joana Cardoso, 109864

Homework 1

1. Pen-and-paper

1) $IG(class, Y_1) = H(class) - H(class|Y_i)$

Para $Y_1 > 0,4$

$$H(class) = -\sum_{x \in class} P(class = x) \cdot \log_2(P(class = x)) =$$

$$= -(0,25 \times \log_2(0,25) + (0,375 \times \log_2(0,375))|2| = 1,562 \text{ bits}$$

$$H(class|Y_2) = \frac{1}{8}(-(0 + 1\log 1 + 0)) + \frac{4}{8}(-(\tfrac{1}{2}\log(\tfrac{1}{2}) + (\tfrac{1}{4}\log(\tfrac{1}{4}) + 2)) +$$
$$+ \frac{3}{8}(-(0 + 2(\tfrac{1}{3}\log\tfrac{1}{3}) + (\tfrac{2}{3}\log\tfrac{2}{3}))) =$$

$$= 1,094$$

$$H(class|Y_3) = -\frac{3}{8} \times (3(\tfrac{1}{3}\log(\tfrac{1}{3}))) - \frac{2}{8} \times (2(\tfrac{1}{2}\log\tfrac{1}{2})) -$$

$$-\frac{3}{8}(\tfrac{3}{4}\log(3) + \tfrac{1}{4}\log(2)) = 1,439$$

$$H(class|Y_4) = -\frac{1}{8}(1\log 1) - \frac{3}{8}(\tfrac{2}{3}\log(\tfrac{2}{3}) + \tfrac{1}{3}\log(\tfrac{1}{3})) - \frac{4}{8}(0 + \tfrac{1}{4}\log\tfrac{1}{4} + \tfrac{3}{4}\log(\tfrac{3}{4}))$$

$$= 0,75$$

$IG(class, Y_2) = H(class) - H(class|Y_2) = 0,468$
$IG(class, Y_3) = H(class) - H(class|Y_3) = 0,123$
$IG(class, Y_4) = H(class) - H(class|Y_4) = 0,812$ ←

We choose Y_4 to be the next node because it has the highest Information Gain

Para: $Y_1 > 0,4 \wedge Y_4 == 0$:

mode (~~A~~B) = ~~A~~B (~~alphabetic order~~)

Para: $Y_1 > 0,4 \wedge Y_4 == 1$:

mode (A, A, B) = A

Para: $Y_1 > 0,4 \wedge Y_4 == 2$: Has ~~more~~ four (4) observations, therefore it will have to be expanded

Para $Y_1 > 0,4 \wedge Y_4 == 2$

$$H(class) = -\left(\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right) = 0.811$$

$$H(class|Y_2) = -\frac{1}{4}(1\log 1) - \frac{3}{4}\left(\frac{1}{3}\log\left(\frac{1}{3}\right) + \frac{2}{3}\log\left(\frac{2}{3}\right)\right) = 0,689$$

$$H(class|Y_3) = -\frac{1}{4}(1\times\log 1) - \frac{2}{4}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) = 0,5$$

$IG(class, Y_2) = 0,811 - 0,689 = 0,121$

$IG(class, Y_3) = 0,811 - 0,5 = 0,311$ ←

we choose Y_3 to be the next node because it has the highest information Gain

Para $Y_1 > 0,4 \wedge Y_4 == 2 \wedge Y_3 == 0$

mode(C) = C

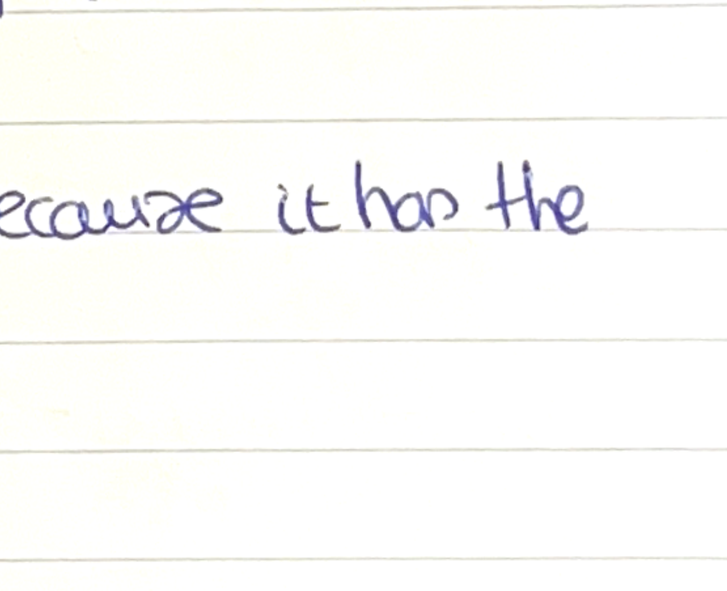
Para $Y_1 > 0,4 \wedge Y_4 == 2 \wedge Y_3 == 1$

mode(C, B) = B (alphabetic order)

Para $Y_1 > 0,4 \wedge Y_4 == 2 \wedge Y_3 == 2$

mode(C) = C

Assim, obtemos a nossa tree:



2) Confusion Matrix

		True A	True B	True C
Predictions	A	4	1	0
	B	0	3	1
	C	0	0	3

$\rightarrow \begin{bmatrix} 4 & 1 & 0 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}$

3) TP: True Positive (Previsões corretas) real = classe; Previsto = classe

FP: False Positive (Previsões erradas) real ≠ classe; Previsto = classe

TN: True Negative (Falha em prever) real = classe; Previsto ≠ classe

FN: False Negative (Previsões corretas de não-classe) real ≠ class; previsto ≠ classe

$$F_1 = \frac{2 \times P \times R}{P+R} \quad ; \underbrace{P = \frac{TP}{TP+FP}}_{\text{Precision}} \qquad \underbrace{R = \frac{TP}{TP+FN}}_{\text{Recall / sensitivity}}$$

Classe A: $TP_A = 4 \quad FP_A = 1 \quad FN_A = 0$

$P_A = \frac{4}{4+1} = \frac{4}{5} \quad R_A = \frac{4}{4+0} = 1 \quad F_1 = \frac{2 \times 0.8 \times 1}{1+0.8} = 0.89$

classe B: $TP_B = 3 \quad FP_B = 1 \quad FN_B = 1$

$P_B = \frac{3}{3+1} = \frac{3}{4} \quad R_B = \frac{3}{3+1} = \frac{3}{4} \quad F_1 = \frac{2 \times 0.75 \times 0.75}{1,5} = \frac{3}{4} = \underline{0.75}$

classe C:

$P_C = \frac{3}{3} = 1 \quad R_C = \frac{3}{3+1} = \frac{3}{4} \quad F_1 = \frac{2 \times \frac{3}{4}}{1,75} = 0.86$

Class B is the one with the lowest training score

4)

A: 0,12 0,33 0,62 0,71

B: 0,18 0,52 0,83 0,90

C: 0,25 0,45 0,58 0,95

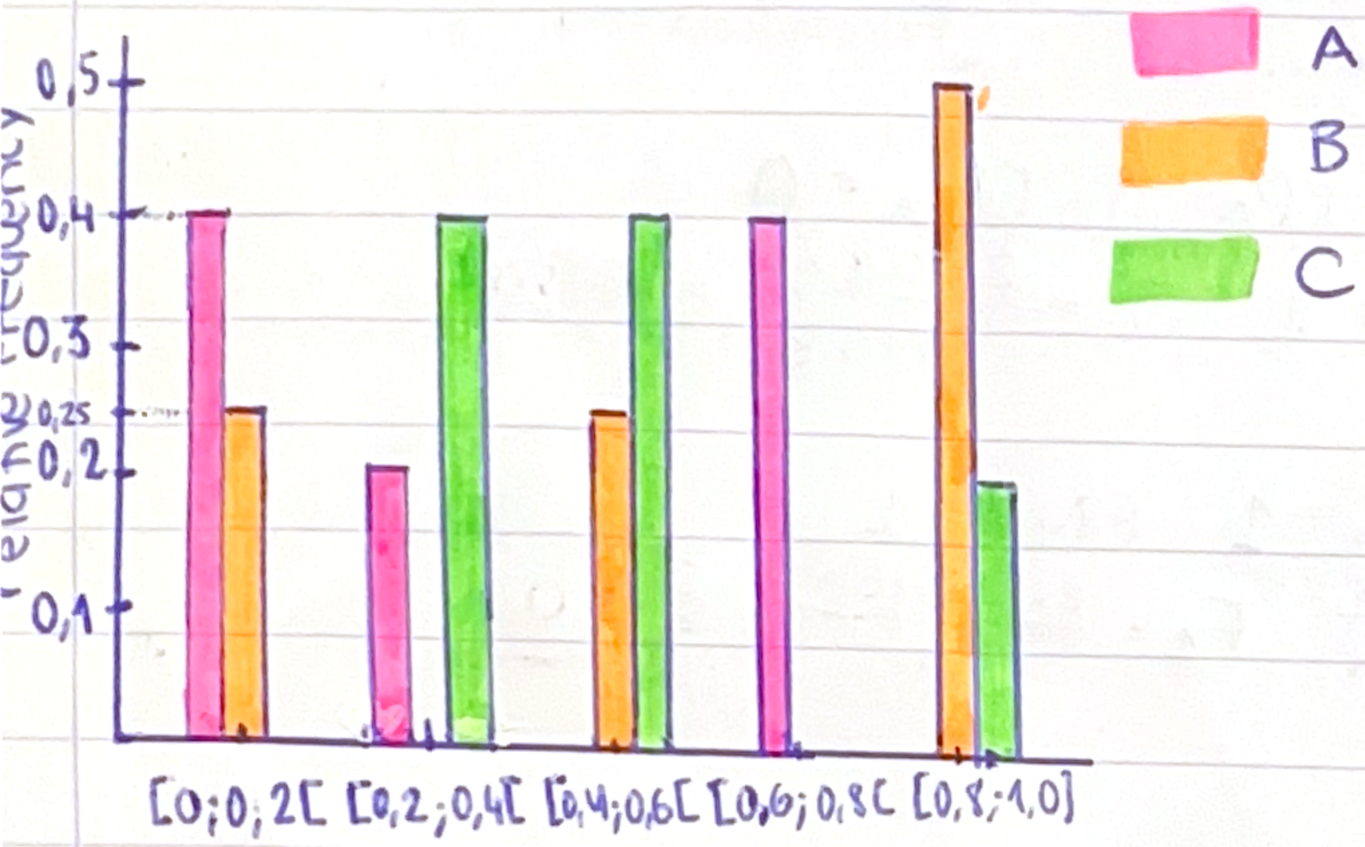
[0,00;0,20[	[0,20;0,40[	[0,40;0,60[	[0,60;0.80[	[0,80;1]
AAB	ACC	BCC	AA	BBC

A: [0,40 ; 0,20; 0; 0,40; 0]

B: [0,25; 0; 0,25; 0; 0,5]

C: [0; 0,40; 0,40; 0; 0,20]

Histograma A B e C



domina: { A em [0,0;0,20[∧ [0,60,0,80[; C em [0,20;0,60[; B em [0,80;1] }

Assim, n é 4 (dividimos o intervalo [0,1] em 4]

5. Neste dataset D, para encontrar outliers temos de utilizar o método Interquartile Range (IQR), aplicado apenas a variáveis contínuas/numéricas, logo V_1.

1° sort dos dados

0.12 0.18 0.25 0.33 0.45 0.52 0.58 0.62 0.71 0.83 0.90 0.95

Q_1 Q_2 Q_3 Q_4

IQR = ~~max(Q1) -~~ $max(Q_3) - min(Q_2)$ = 0.71 − 0,33

= 0.38

intervalo dos valores que não são outliers: $[Q_1 - 1,5 \times IQR, Q_3 + 1,5 \times IQR]$ =

$= \left[\frac{0,25+0,33}{2} - 1,5 \times 0.38, \frac{(0.71+0.83)}{2} + 1,5 \times 0.38\right] = [-0,28; 1,34]$

Como todos os dados pertencem ao intervalo $[-0,28; 1,34]$ podemos concluir que no dataset D não há outlier, pelo IQR method.