homework

Homework 1:

I. Pen-and-Paper:

1)

Shannon Entropy: $I(x) = \sum_{x \in X} p(x) \times I(x) = -\sum_{x \in X} p(x) \times \log_2(p(x))$

$$E(\text{class}) = E(y_{out}) = -\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)$$

$$= \log_2 3 = \boxed{1.58496}$$

$$IG(class|y_x) = E(class) - E(class|y_x)$$

$y_1 > 0.4$

$$E(class|y_2) = E(y_{out}|y_2=0) + E(y_{out}|y_2=1) + E(y_{out}|y_2=2)$$

$$= \frac{3}{7}\left(-\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)\right) +$$

$$+ \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right) + \frac{2}{7}\left(-\left(1\log_2(1)\right)\right)$$

$$= \frac{3}{7}\left(-\log_2\left(\frac{1}{3}\right)\right) + \frac{2}{7}\left(-\log_2\left(\frac{1}{2}\right)\right) + \frac{2}{7}\left(-\log_2(1)\right)$$

$$= \frac{3}{7}\log_2(3) + \frac{2}{7} \approx \boxed{0.964984}$$

$$E(class|y_3) = E(y_{out}|y_3)$$

$$= E(y_{out}|y_3=0) + E(y_{out}|y_3=1) + E(y_{out}|y_3=2)$$

$$= \frac{1}{7}\left(-1\log_2(1)\right) + \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right)$$

$$+ \frac{4}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right)$$

$$= \frac{1}{7}\log_2(1) + \frac{2}{7} + \frac{4}{7} = \frac{6}{7} \approx \boxed{0.857143}$$

$$E(Clan|y4) = E(yout|y4) = E(yout|y4=0) + E(yout|y4=1) + E(yout|y4=2)$$

$$= \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right) +$$

$$+ \frac{3}{7}\left(-\left(\frac{1}{3}\log_2\left(\frac{1}{3}\right) + \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)\right) + \frac{2}{7}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{1}{2}\right)\right)\right)$$

$$= \frac{2}{7}(1) + \frac{3}{7}\left(-\left(\frac{1}{3}(2 - 3\log_2(3))\right)\right) + \frac{2}{7}$$

$$= \frac{2}{7} + \frac{3}{7}\log_2(3) \simeq \boxed{0.964984}$$

$$IG(Clan|y2) = E(Clan) - E(Clan|y2)$$

$$= 1.58496 - 0.964984 = \boxed{0.619976}$$

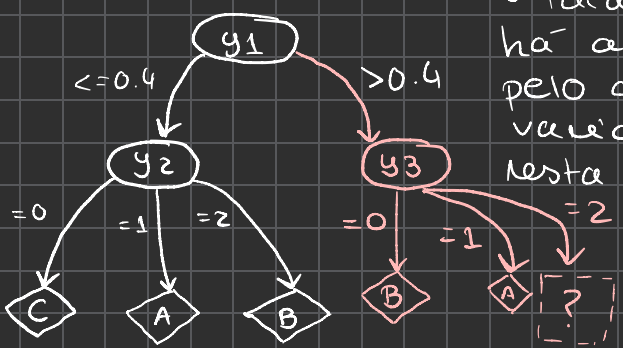
$$IG(Clan|y3) = E(Clan) - E(Clan|y3)$$

$$= 1.58496 - 0.857143 = \boxed{0.727817}$$

↳ maior IG, escolhe-se esta variável (y3)

$$IG(Clan|y4) = E(Clan) - E(Clan|y4)$$

$$= 1.58496 - 0.964984 = \boxed{0.619976}$$



- Como temos 4 observações para $y1 > 0.4 \wedge y3 = 2$, temos de recorrer novamente a IG.

- Para $y1 > 0.4$ e $y3 = 0$, há apenas 1 observação, pelo que o valor da variável de destino será B nesta situação (única observação)

- Para $y1 > 0.4$ e $y3 = 1$, há apenas 2 observações, pelo que o nº não é dividido. Seguindo a ordem alfabética, o valor da variável de destino será A

$$E(class|y3=2 \wedge y2) = E(class|y3=2 \wedge y2=0) + E(class|y3=2 \wedge y2=1) + E(class|y3=2 \wedge y2=2)$$

$$= \frac{1}{4}\underbrace{\left(-1\log_2(1)\right)}_{\emptyset} + \frac{1}{4}\underbrace{\left(-1\log_2(1)\right)}_{\emptyset} + \frac{1}{2}\underbrace{\left(-1\log_2(1)\right)}_{\emptyset}$$

$$= \boxed{\emptyset}$$

$$E(class|y3=2 \wedge y4) = E(class|y3=2 \wedge y4=0) + E(class|y3=2 \wedge y2=1) + E(class|y3=2 \wedge y4=2)$$

$$= \frac{1}{2}\left(-\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right) +$$

$$+ \frac{1}{4}\underbrace{(-1\log(1))}_{\emptyset} + \frac{1}{4}\underbrace{(-1\log(1))}_{\emptyset}$$

$$= \frac{1}{2}(-(-1)) = \frac{1}{2} = \boxed{0.5}$$

$$IG(class|y3=2 \wedge y2) = E(class) - \underbrace{E(class|y3=2 \wedge y2)}_{\emptyset}$$

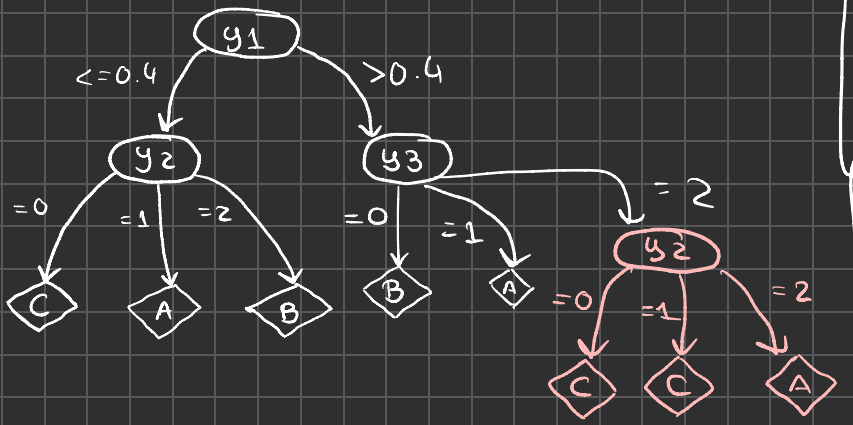
$$= E(class) = \boxed{1.58496}$$

↳ maior IG, escolhe-se esta variável

$$IG(class|y3=2 \wedge y4) = E(class) - E(class|y3=2 \wedge y4)$$

$$= 1.58496 - 0.5$$

$$= \boxed{1.08496}$$



Decision Tree

Para $y1 > 0.4 \wedge y3 = 2$ já só temos 4 observações: 1 observação para $y2 = 0$, 1 observação para $y2 = 1$ e 2 observações para $y2 = 2$. Desta forma os nós não serão expandidos. Para os ramos $y2 = 0$ e $y2 = 1$, como cada um tem apenas uma observação, o valor da variável de destino de cada ramo será o valor da observação (para ambos, C). Para o ramo $y2 = 2$, ambas as observações têm valor A, pelo que a variável de destino deste ramo tem valor A.

2.

previsto

real	A	B	C
A	4	0	0
B	1	2	1
C	0	0	4

TP

real

previsto	A	B	C
A	4	1	0
B	0	2	0
C	0	1	4

3.

Which class has the lower F1-score?

$$F1\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

TP - true positive
FP - false positive
TN - true negative
FN - false negative

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Sensitivity / Recall} = \frac{TP}{TP + FN}$$

Class A:

$$\text{Precision}_A = \frac{4}{4+0+1} = \frac{4}{5} = 0.80 \qquad \text{Recall}_A = \frac{4}{4} = 1$$

$$F1\text{-score}_A = 2 \times \frac{0.80 \times 1}{0.80 + 1} = 2 \times \frac{0.80}{1.80} = \frac{1.60}{1.80} = \frac{8}{9} = \boxed{0.8(8)}$$

Class B.

$$\text{Precision}_B = \frac{2}{2+0} = 1 \qquad \text{Recall}_B = \frac{2}{4} = \frac{1}{2} = 0.5$$

$$F1\text{-score}_B = 2 \times \frac{1 \times 0.5}{1 + 0.5} = 2 \times \frac{0.5}{1.5} = 2 \times \frac{1}{3} = \frac{2}{3} = \boxed{0,6(6)}$$

class C:

$$\text{Precision}_C = \frac{4}{4+1+0} = \frac{4}{5} = 0.8 \qquad \text{Recall}_C = \frac{4}{4} = 1$$

$$\text{F1-score}_C = 2 \times \frac{0.8 \times 1}{0.8+1} = \frac{1.60}{1.80} = \frac{8}{9} = \boxed{0.8(8)}$$

R: Class B is the one with the lowest training F1 score, with a value of approximately 0.67.

4.

	rank y_1	rank y_2	$(\text{rank } y_1)^2$	$(\text{rank } y_2)^2$	$\text{rank}y_1 \cdot \text{rank}y_2$
x1	3	8	9	64	24
x2	2	11	4	121	22
x3	1	3,5	1	12.25	3.5
x4	5	3,5	25	12.25	17.5
x5	4	3,5	16	12.25	14
x6	10	11	100	121	110
x7	12	3,5	144	12.25	42
x8	11	11	121	121	121
x9	7	8	49	64	56
x10	9	3,5	81	12.25	31.5
x11	6	8	36	64	48
x12	8	3,5	64	12.25	28
	$\sum_{x \in X}(\text{rank}y_1) = 78$	$\sum_{x \in X}(\text{rank}y_2) = 78$	$\sum_{x \in X}(\text{rank}y_1)^2 = 650$	$\sum_{x \in X}(\text{rank}y_2)^2 = 628.5$	$\sum_{x \in X}(\text{rank}y_1)(\text{rank}y_2) = 517.5$

Há rank ties, pelo que o clássico Spearman é substituído pelo PCC dos ranks:

$$\text{Spearman}(y_1, y_2) = \text{PCC}((3\ 2\ 1\ 5\ 4\ 10\ 12\ 11\ 7\ 9\ 6\ 8), (8\ 11\ 3,5\ 3,5\ 3,5\ 11\ 3,5\ 11\ 8\ 3,5\ 8\ 3,5))$$

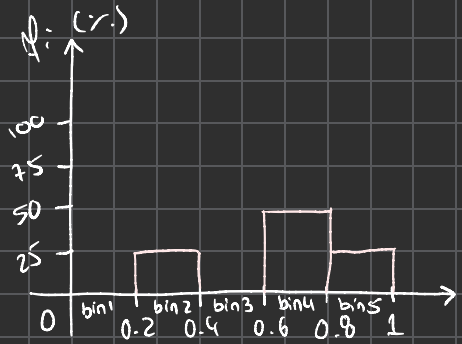
$$r = \frac{\sum(\text{rank}y_1)(\text{rank}y_2) - \frac{\sum(\text{rank}y_1)\sum(\text{rank}y_2)}{n}}{\sqrt{\left(\sum(\text{rank}y_1)^2 - \frac{(\sum \text{rank}y_1)^2}{n}\right) \cdot \left(\sum(\text{rank}y_2)^2 - \frac{(\sum(\text{rank}y_2))^2}{n}\right)}}$$

$$r = \frac{517.5 - \frac{78 \times 78}{12}}{\sqrt{\left(650 - \frac{78^2}{12}\right)\left(628.5 - \frac{78^2}{12}\right)}} \approx \boxed{0.0796587}$$

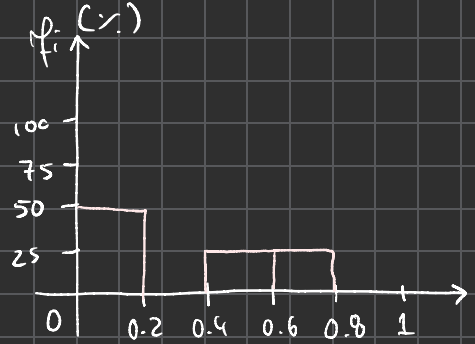
R: Como $r = 0.0796587$, y_1 e y_2 são loosely related.

5.

Classe A:



Classe B:



Classe C: