Homework I – Group 108

(ist1103617, ist199991)

I. Pen-and-paper

1) Answer 1

$$IG(z|y) = E(z) - E(z|y) \qquad \boxed{z = y_{out}}$$

$$E(z|y_1 > 0{,}4) = -\left(\frac{3}{7}\log_2\left(\frac{3}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right) + \frac{2}{7}\log_2\left(\frac{2}{7}\right)\right)$$

$$= 1{,}557$$

$$E(z|y > 0{,}4, y_2) = -\frac{3}{7}\left(\frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3} + \frac{1}{3}\log\frac{1}{3}\right) - \frac{2}{7}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) - \frac{2}{7}\left(\cancelto{0}{1\log 1}\right) \approx 0{,}965$$

$$E(z|y > 0{,}4, y_3) = -\frac{1}{7}\left(\cancelto{0}{1\log 1}\right) - \frac{2}{7}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) - \frac{4}{7}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) = \frac{6}{7}$$

$$E(z|y > 0{,}4, y_4) = -\frac{2}{7}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) - \frac{3}{7}\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right) - \frac{2}{7}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)$$

Homework I – Group 108

(ist1103617, ist199991)

$IG(z \mid y_1 > 0{,}4, y_2) = 1{,}556 - 0{,}965 = 0{,}591$

$IG(z \mid y_1 > 0{,}4, y_3) = 1{,}556 - \frac{6}{7} = 0{,}7$

$IG(z \mid y_1 > 0{,}4, y_4) = 1{,}556 - 0{,}965 = 0{,}591$

Escolhemos o y_3, por ter maior gain.

$y_3 = 0 \rightarrow B$

$y_3 = 1 \rightarrow A, B$ 2 observações, empate escolhe-se A

$y_3 = 2 \rightarrow$ 4 observações, calcular IG

$E(z \mid y_1 > 0{,}4, y_3 = 2) = -\left(\frac{1}{2}\log\left(\frac{1}{2}\right) + \frac{1}{2}\log\left(\frac{1}{2}\right)\right) = 1$

$E(z \mid y_1 > 0{,}4, y_3 = 2, y_2) = -\left(\frac{1}{4}\cancel{\left(1\log 1\right)}^{\to 0} + \frac{1}{4}\cancel{\left(1\log 1\right)}^{\to 0}\right.$

$\left. + \frac{2}{4}\left(1\log 1\right)\right) = 0$

$E(z \mid y_1 > 0{,}4, y_3 = 2, y_4) = -\left(\frac{2}{4}\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) + \frac{1}{4}\cancel{\left(1\log 1\right)}^{\to 0}\right.$

$\left. + \frac{1}{4}\cancel{\left(1\log 1\right)}^{\to 0}\right) = 0{,}5$

$IG(z \mid y_1 > 0{,}4, y_3 = 2, y_2) = 1 - 0 = 1$

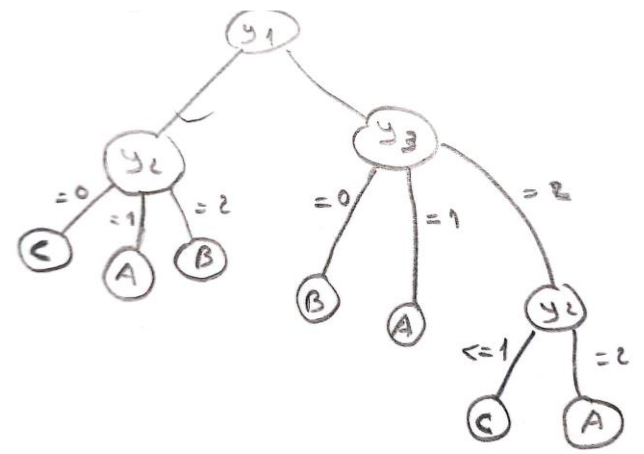
$IG(z \mid y_1 > 0{,}4, y_3 = 2, y_4) = 1 - 0{,}5 = 0{,}5$

Escolhe-se y_2, pois tem maior gain

$y_2 = 0 \rightarrow C$

$y_2 = 1 \rightarrow C$

$y_2 = 2 \rightarrow A$



2.

2) Answer 2

Aprendizagem 2022/23

Homework I – Group 108

(ist1103617, ist199991)

2.

predicado \ True	A	B	C	
A	4	1	0	5
B	0	2	0	2
C	0	1	4	5
	4	4	4	12

3) Answer 3

| | 4 | 4 | 4 | 12 |

3.

$prec_A = \frac{4}{5}$

$prec_B = 1$

$prec_C = \frac{4}{5}$

$\boxed{prec = precision}$

$recall_A = \frac{4}{4} = 1$

$recall_B = \frac{2}{4} = 0{,}5$

$recall_C = \frac{4}{4} = 1$

$$f_1 = \frac{2 \times prec \times recall}{prec + recall}$$

$$precision = \frac{TP_A}{TP_A + FP_A}$$

$f_{1A} = \frac{8}{9} \approx 0{,}8889$

$f_{1B} = \frac{2}{3} \approx 0{,}6667$

$f_{1C} = \frac{8}{9} \approx 0{,}8889$

O que tem menor f_1 é o B

4) Answer 4

Aprendizagem 2022/23

Homework I – Group 108

(ist1103617, ist199991)

4)

y_1	y_2		sorted y_1	rank y_1
0,24	1		0,04	1
0,06	2		0,06	2
0,04	0		0,24	3
0,36	0		0,32	4
0,32	0		0,36	5
0,68	2		0,44	6
0,9	0		0,46	7
0,76	2		0,52	8
0,46	1		0,62	9
0,62	0		0,68	10
0,44	1		0,76	11
0,52	0		0,90	12

$y_1 = [0,24; 0,06; 0,04; 0,36; 0,32; 0,68; 0,9; 0,76; 0,46; 0,62; 0,44; 0,52]$

$y_1' = [3; 2; 1; 5; 4; 10; 12; 11; 7; 9; 6; 8]$

y_2 sorted $\rightarrow y_2 = [0; 0; 0; 0; 0; 0, 1, 1, 1, 2, 2, 2]$

$y_{2\,rank} = [3,5; 3,5; 3,5; 3,5; 3,5; 3,5; 8; 8; 8; 11; 11; 11]$

$y_2' [8; 11; 3,5; 3,5; 3,5; 11; 3,5; 11; 8; 3,5; 8; 3,5]$

$\sum_{i=1}^{12} (y'_1)^2 = 650 \qquad \sigma(y'_1) = 3,6$

$\bar{y}'_1 = 6,5 \qquad \sum_{i=1}^{12} (y'_2)^2 = 628,5$

$\bar{y}'_2 = 6,5 \qquad \sigma(y'_2) = \sqrt{\frac{628,5 - 12 \times 6,5^2}{11}} = 3,32$

$pearson(y'_1, y'_2) = \frac{0,95455}{3,32 \times 3,6} = 0,0791$

$cov(y'_1, y'_2) = \qquad \sum_{i=1}^{12} (y'_1\, y'_2) = 517,5$

$= \frac{517,5 - 12 \times 6,5 \times 6,5}{11}$

$= 0,95454$

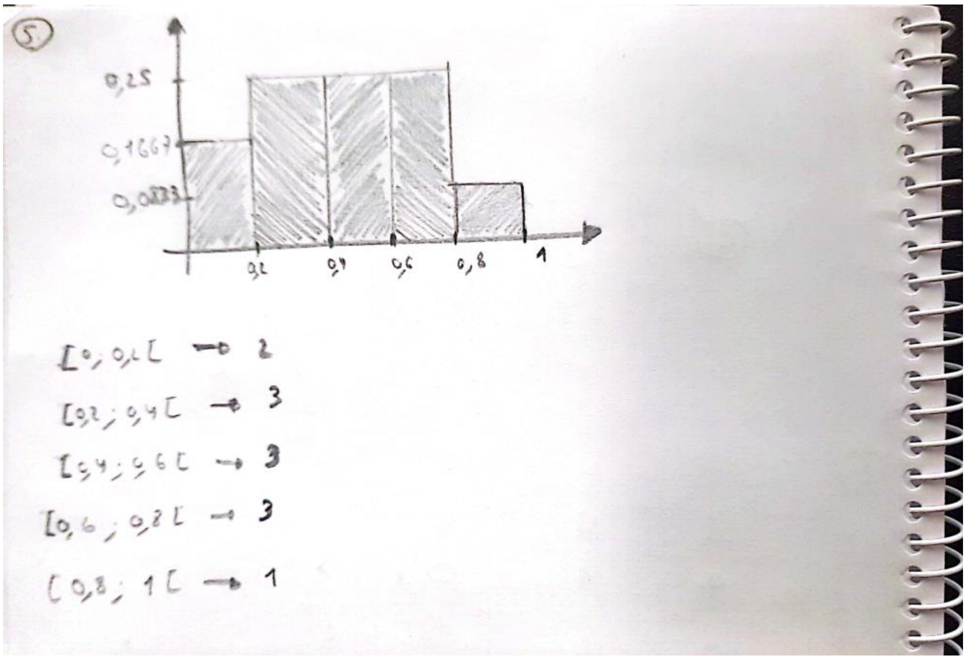
Como o valor do coeficiente é praticamente 0, a correlação entre y_1 e y_2 é fraca.

5) **Answer 5**

Aprendizagem 2022/23

Homework I – Group 108

(ist1103617, ist199991)

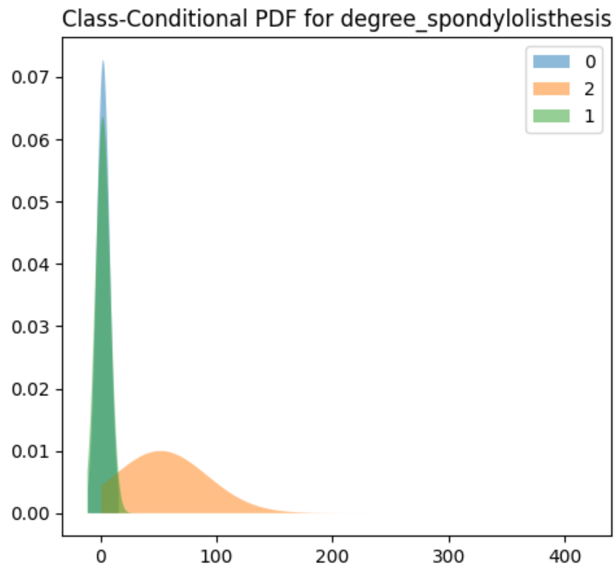


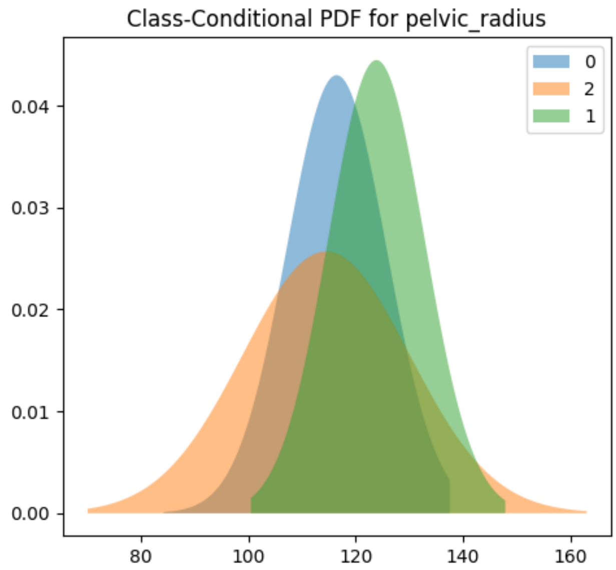
II. Programming and critical analysis

1.

Input variable with the highest discriminative power: degree_spondylolisthesis

Input variable with the lowest discriminative power: pelvic_radius





2.

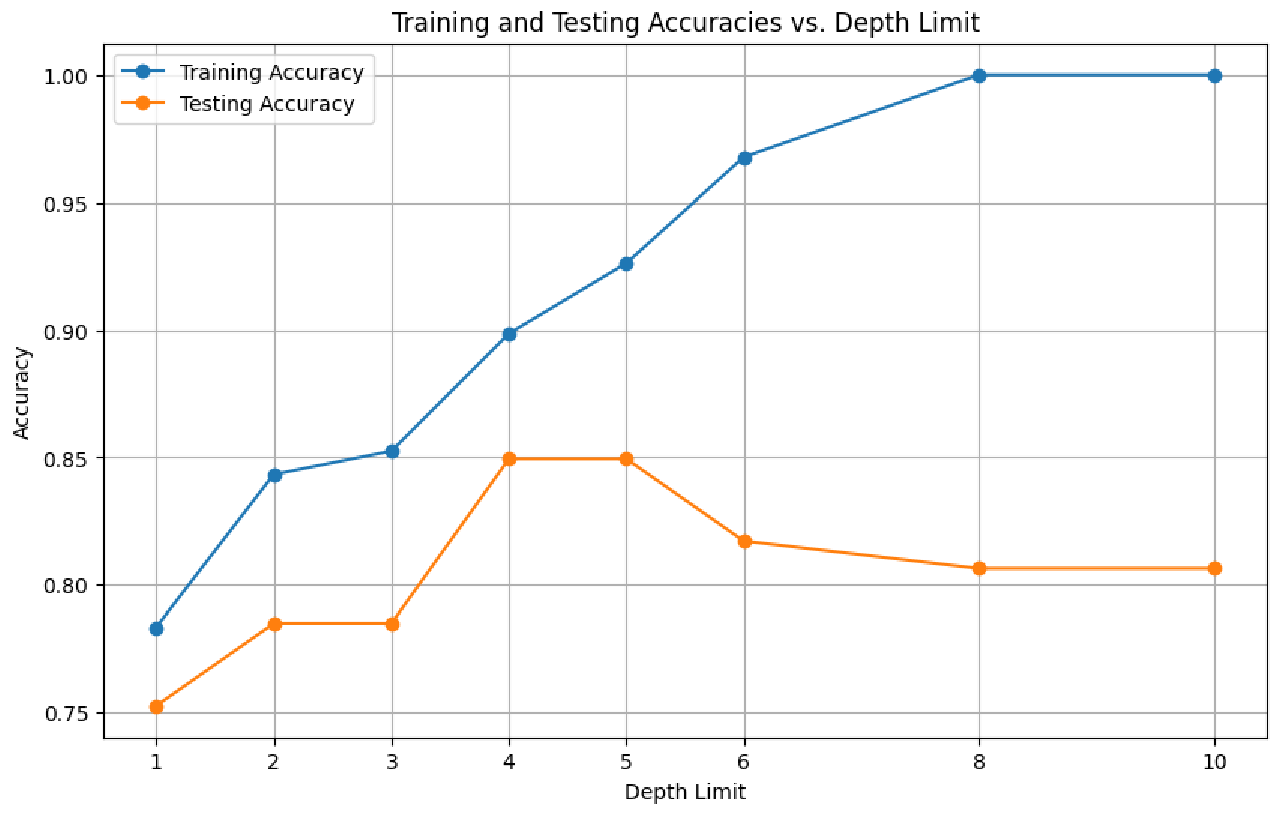
Depth Limit: 1, Training Accuracy: 0.7834, Testing Accuracy: 0.7527
Depth Limit: 2, Training Accuracy: 0.8433, Testing Accuracy: 0.7849
Depth Limit: 3, Training Accuracy: 0.8525, Testing Accuracy: 0.7849

Aprendizagem 2022/23

Homework I – Group 108

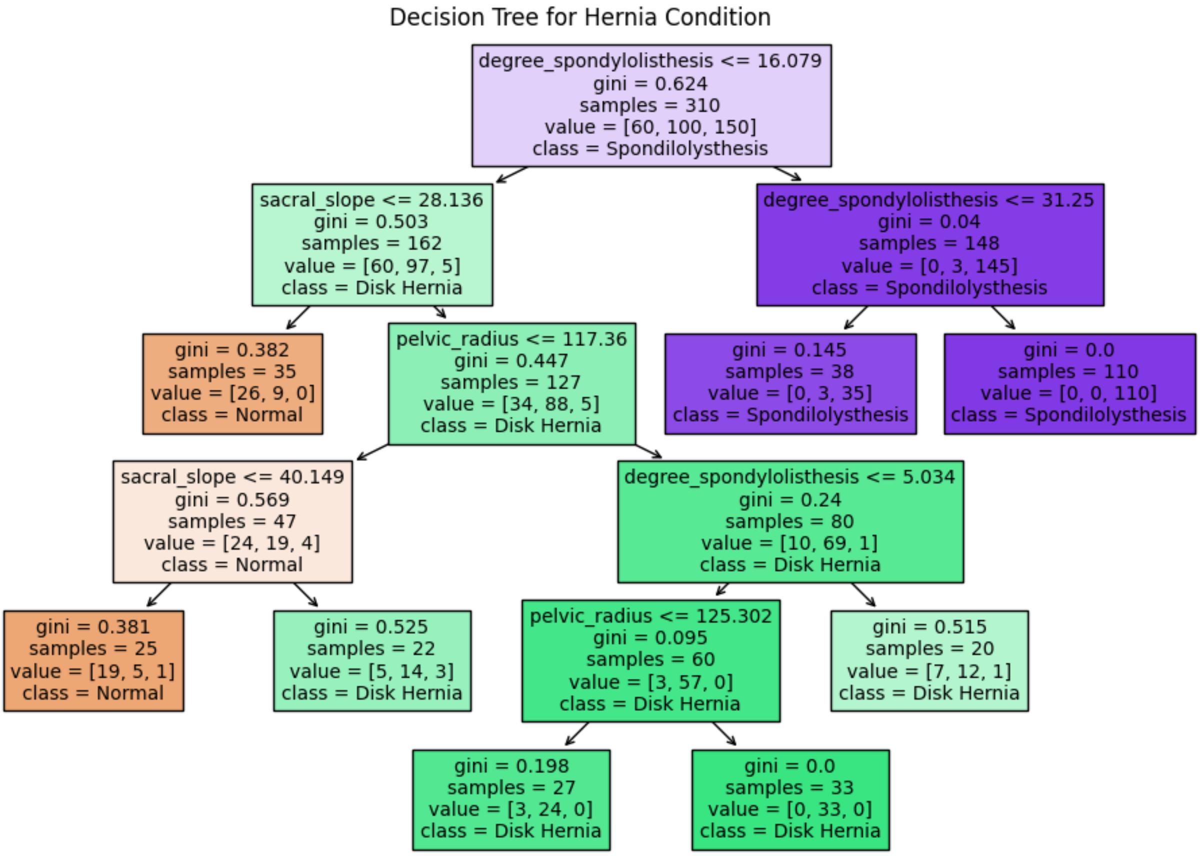
(ist1103617, ist199991)

Depth Limit: 4, Training Accuracy: 0.8986, Testing Accuracy: 0.8495
Depth Limit: 5, Training Accuracy: 0.9263, Testing Accuracy: 0.8495
Depth Limit: 6, Training Accuracy: 0.9677, Testing Accuracy: 0.8172
Depth Limit: 8, Training Accuracy: 1.0000, Testing Accuracy: 0.8065
Depth Limit: 10, Training Accuracy: 1.0000, Testing Accuracy: 0.8065



3. In summary, the results indicate that deeper decision trees (depth limit > 5) overfit the training data and do not generalize well to new data, leading to a decrease in testing accuracy. The optimal depth limit for this specific dataset is likely around 4 or 5, where the model achieves the highest testing accuracy while still maintaining a good generalization capacity.

4. i.



ii. Conditions with a degree_spondylolisthesis value <= 16.07 and a sacral_slope value <= 28.136 are tipically classed as having a Disk Hernia. In this group, having a pelvic_radius > 117.36 classifies you as "Normal", while other values classifiy you as having a Disk Hernia. In this new Disk Hernia group, having a degree_spondylolisthesis value <= 5.034 classifies you as having a Disk Hernia, while sacral_slope values <= 40.149 tipically classify you as "Normal", even though there are 19 patients who have a Disk Hernia (vs 24 "Normal" and 4 with Spondylolisthesis).

Homework1.ipynb