Aprendizagem 2024/25

Homework I – Group 017

(ist1106992, ist1106154)

I. Pen-and-paper

1.

$E(y_{out}) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 1,5567$

$E(y_{out}|y_2) = \frac{4}{7}\left(-\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2}\right)\right) +$

$+ \frac{3}{7}\left(-\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)\right) = \frac{6}{7} + 0,3936 = 1,2507$

$E(y_{out}|y_3) = \frac{2}{7}\left(-\log_2 1\right) + \frac{1}{7}\left(-\log_2 1\right) +$

$+ \frac{4}{7}\left(-\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2}\right)\right) = \frac{6}{7} = 0,8571$

$E(y_{out}|y_4) = \frac{5}{7}\left(-\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right)\right) +$

$+ \frac{2}{7}\left(-\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{1}{3}\right)\right) = 0,9649$

lower entropy → y_3 → more advantageous to split

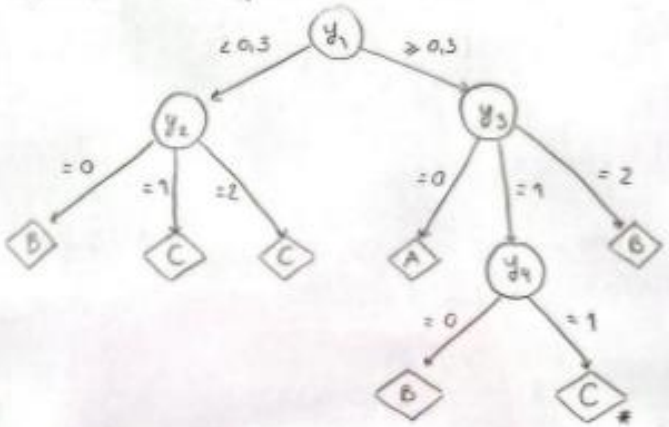
$E(y_{out}) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2}\right) = \frac{3}{2} = 1,5$

$E(y_{out}|y_2) = 1\left(-\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{2}\log_2\frac{1}{2}\right)\right) = \frac{3}{2} = 1,5$

$E(y_{out}|y_4) = \frac{1}{4}\left(-\left(1\log_2 1\right)\right) + \frac{3}{4}\left(-\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right)\right) = 0,68$

lower entropy → y_4 → more advantageous to split

→ at this point, there're only 3 observations (min. is 4) so stop the splitting



* out of the 3 observations (A, C, C) C is the most likely so we assume the branches' result as C.

Aprendizagem 2024/25

Homework I – Group 017

(ist1106992, ist1106154)

2.

Predicted \ Real	A	B	C
A	2	0	0
B	0	4	0
C	1	0	3

3.

$$F_1 = \frac{2PR}{P+R} \qquad R = \frac{TP}{P} = \frac{TP}{TP+FN} \qquad P = \frac{TP}{TP+FP}$$

A

$$P_A = \frac{2}{2} = 1$$

$$R_A = \frac{2}{2+1} = \frac{2}{3}$$

$$F_{1A} = 2\,\frac{1 \times \frac{2}{3}}{1+\frac{2}{3}} = \frac{4}{5}$$

B

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

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

$$F_{1B} = 2\,\frac{1\times 1}{1+1} = 1$$

C

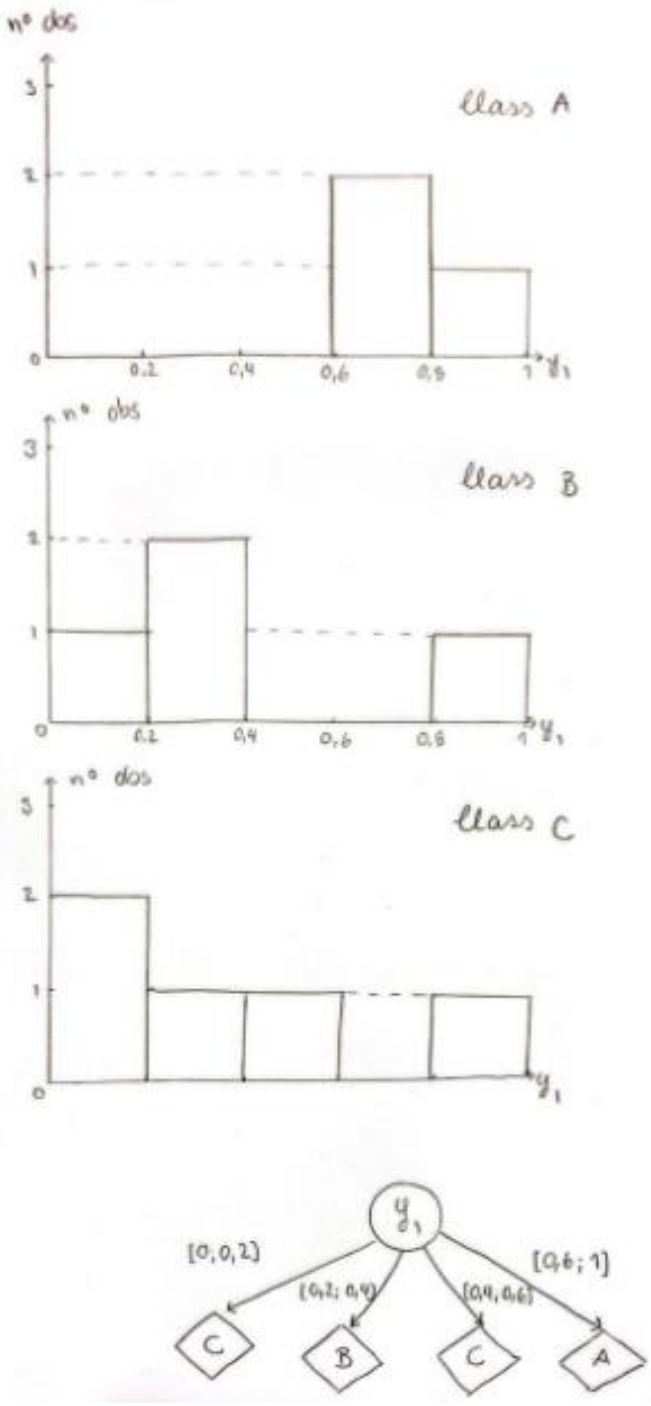
$$P_B = \frac{5}{5+1} = \frac{5}{6}$$

$$R_B = \frac{5}{5} = 1$$

$$F_{1C} = 2\,\frac{\frac{5}{6}\times 1}{\frac{5}{6}+1} = \frac{10}{11}$$

The A class has the lowest traing F1 score.

4.



For the [0;0,2], [0,2;0,4], [0,4;0,6] and [0.6;0,8] bins there is a class with majority (respectively, C, B, C and A). In the [0,8;1] all classes have the same number of observations so we chose (alphabetically) class A to represent this bin. Given this, both [0,6;0,8] and [0,8;1] have the main class as A so we fused them in the tree above and created the [0,6;1] branch as A.

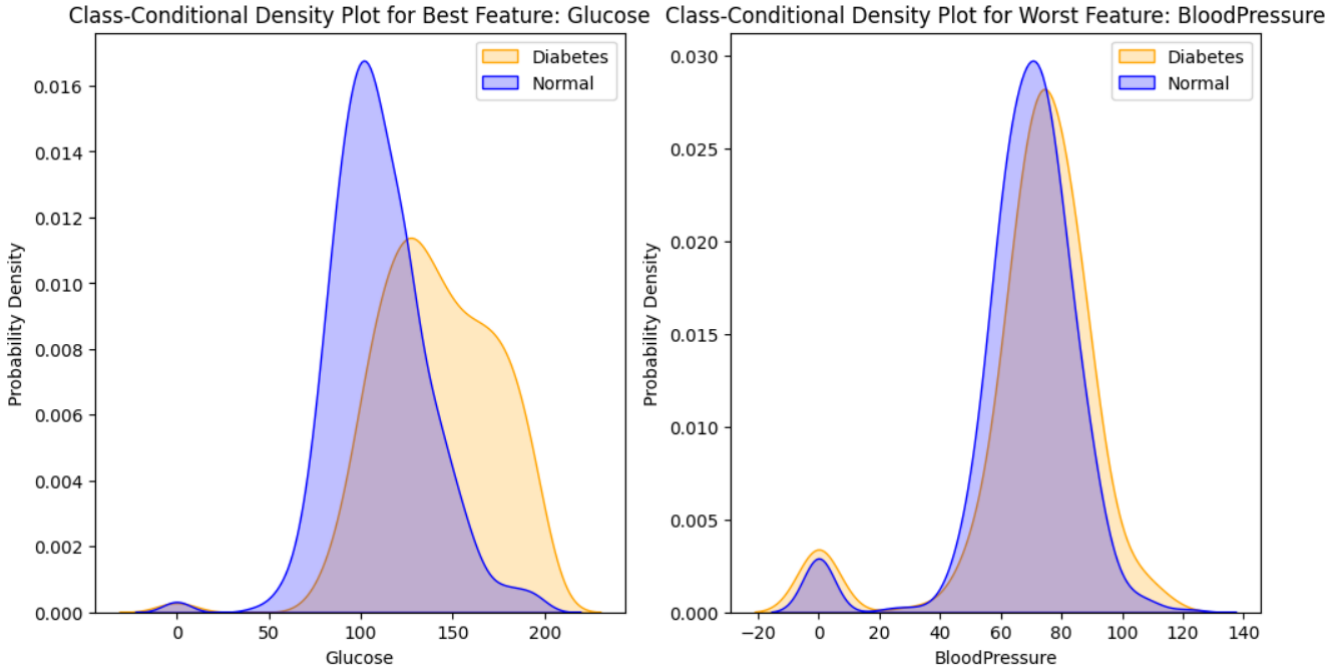
Aprendizagem 2024/25

Homework I – Group 017

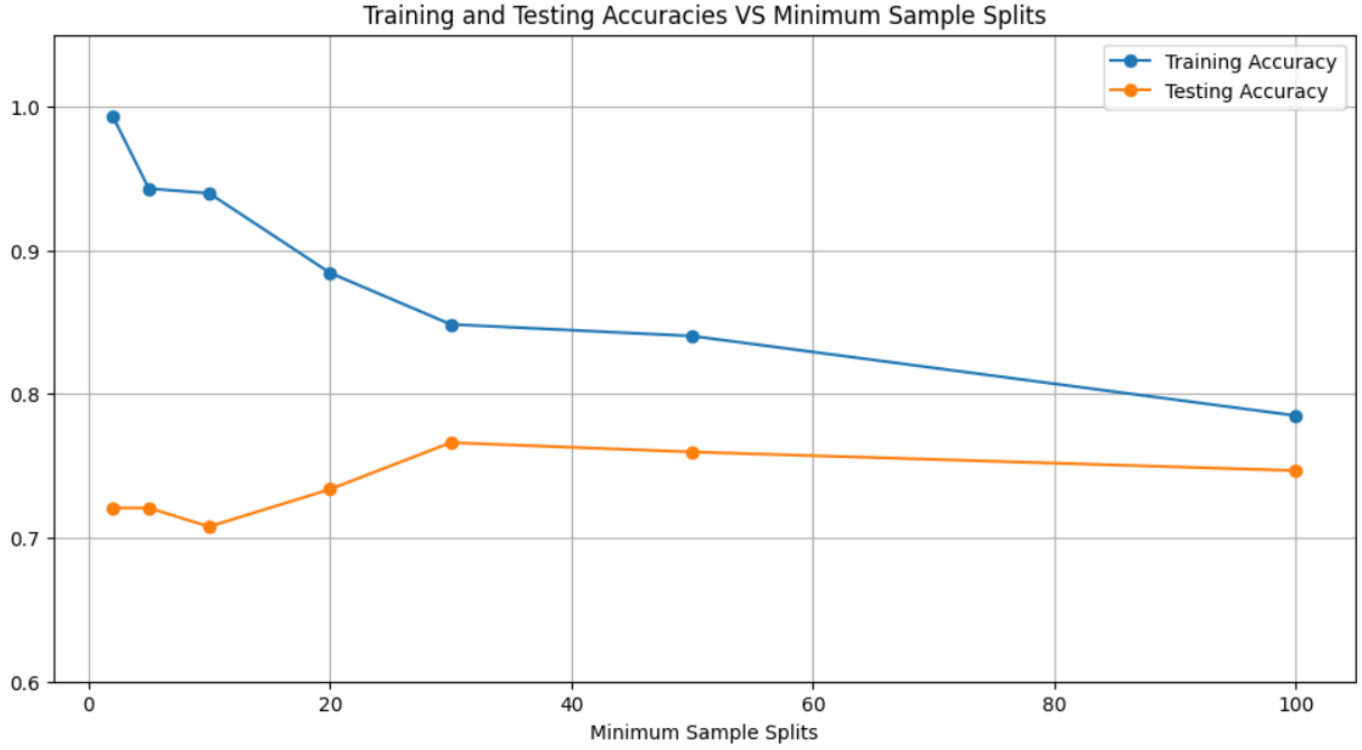
(ist1106992, ist1106154)

II. Programming and critical analysis

1.



2.



Homework I – Group 017

(ist1106992, ist1106154)

3.

With the analysis of the resulting plot from exercise 2, it's possible to assess the relation between the number of min_sample_splits and the testing and training accuracies and, therefore, the model's generalization capacity.

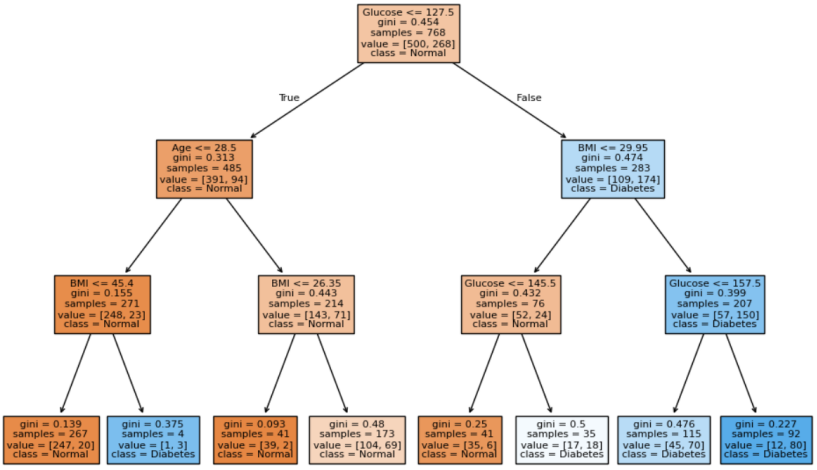
For min_sample_splits between 0 and 25, the training accuracy is quite high as opposed to the testing accuracy, as seen in the plot. This happens because smaller min_sample_split values allow the decision tree to get deeper and more complex, causing the model to memorize the data rather than generalize patterns, leading to overfitting.

On the other hand, for min_sample_split between 40 and 100, as the values increase, testing accuracy decreases slightly, and training accuracy drops more significantly. This happens because increasing the min_sample_split results in a simpler and shallower decision tree, making the model too simplistic to effectively capture patterns in the training data, leaning towards underfitting.

In conclusion, lowering the min_sample_split value increases the likelihood of overfitting, while raising it increases the risk of underfitting. In this case, the optimal min_sample_split value is around 30, as that is where testing accuracy reaches its peak, indicating improved generalization capacity.

4.

i.



ii.

Upon analysing the decision tree, the main features that characterize diabetes are, in order of importance: Glucose, BMI and Age. In the following table, we present the interpretation of the decision tree for the class “Diabetes” and the conditions for each features along with the probability of the class in the tree’s leaves.

P(class = Diabetes)	Conditions
3/4	Glucose <= 127.5, Age <= 28.5 and BMI > 45.4
18/35	Glucose > 145.5 and BMI <= 29.95
70/115	127.5 < Glucose <= 157.5 and BMI > 29.95
80/92	Glucose > 157.5 and BMI > 29.95