Aprendizagem 2024/25

Homework I – Group 028

(ist1106987, ist1107057)

I. Pen-and-paper

1) y_1 created the partitions: $\mathbb{Z}_1 = \{x_1, x_2, x_3, x_4, x_5\}, \mathbb{Z}_2 = \{x_6, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}\}$

$\mathbb{Z}_2$ takes as candidates y_2, y_3, y_4.

$\mathbb{U}_{y_{out}} = \{\text{A, B, C}\}$

$$H(y_{out}) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right) \approx 1{,}55666$$

$\mathbb{U}_{y_2} = \{0, 1\}$

$$IG(y_2) = 1{,}55666 - \left[\frac{4}{7} \times \left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right] - \left[\frac{3}{7} \times \left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right)\right] \approx 0{,}305962$$

$\mathbb{U}_{y_3} = \{0, 1, 2\}$

$$IG(y_3) = 1{,}55666 - \left[\frac{2}{7} \times (-1\ \log_2(1))\right] - \left[\frac{4}{7} \times \left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right] - \left[\frac{1}{7} \times (-1\ \log_2(1))\right] \approx 0{,}699517$$

$\mathbb{U}_{y_4} = \{0, 1\}$

$$IG(y_4) = 1{,}55666 - \left[\frac{4}{7} \times \left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right] - \left[\frac{3}{7} \times \left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)\right] \approx 0{,}591673$$

y_3 has the highest Information Gain. It creates the partitions $\mathbb{Z}_3 = \{x_8, x_{11}\}$, $\mathbb{Z}_4 = \{x_6, x_7, x_9, x_{10}\}$ and $\mathbb{Z}_5 = \{x_{12}\}$.

By i), $\mathbb{Z}_3$ and $\mathbb{Z}_5$ can't be split any further.

$\mathbb{Z}_4$ takes as candidates y_2, y_4.

$\mathbb{U}_{y_{out}} = \{\text{A, B, C}\}$

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

$\mathbb{U}_{y_2} = \{0\}$

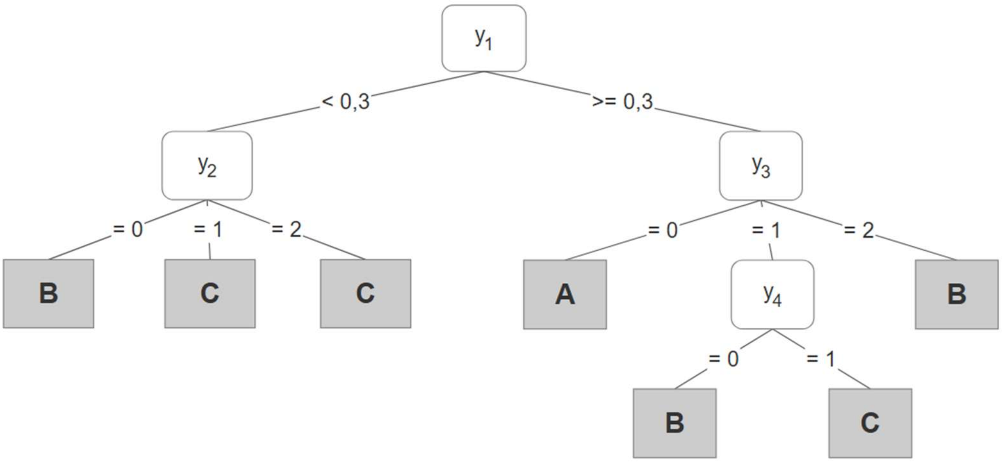
$$IG(y_2) = 1{,}5 - \left[1 \times \left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right] = 0$$

$\mathbb{U}_{y_4} = \{0, 1\}$

$$IG(y_4) = 1{,}5 - \left[\frac{1}{4} \times (-1\log_2(1))\right] - \left[\frac{3}{4} \times \left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right)\right] \approx 0{,}811278$$

y_4 has the highest Information Gain. It creates the partitions $\mathbb{Z}_6 = \{x_6\}$ and $\mathbb{Z}_7 = \{x_7, x_9, x_{10}\}$.

By i), $\mathbb{Z}_6$ and $\mathbb{Z}_7$ can't be split any further.



2) Confusion Matrix:

		Real		
		A	B	C
Previsto	A	2	0	0
	B	0	4	0
	C	1	0	5

3)

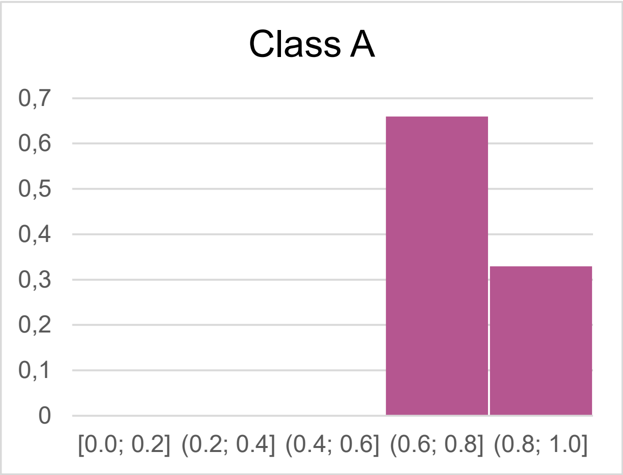
$$P_A = \frac{2}{2} = 1; \quad R_A = \frac{2}{2+1} = \frac{2}{3}; \quad F_A = \frac{2 \times 1 \times \frac{2}{3}}{1 + \frac{2}{3}} = \frac{4}{5} = 0.8$$

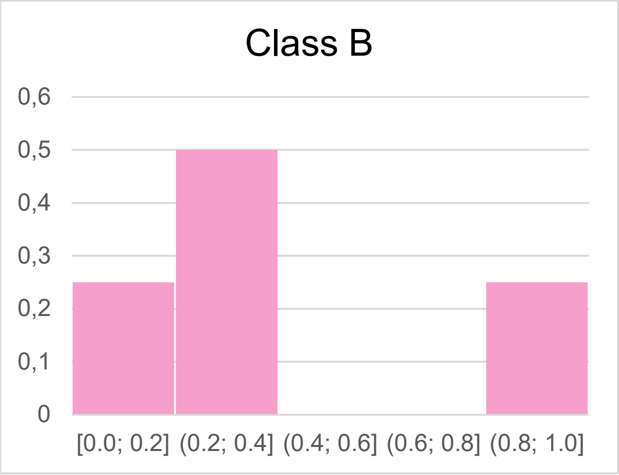
$$P_B = \frac{4}{4} = 1; \quad R_B = \frac{4}{4} = 1; \quad F_B = \frac{2 \times 1 \times 1}{1 + 1} = 1$$

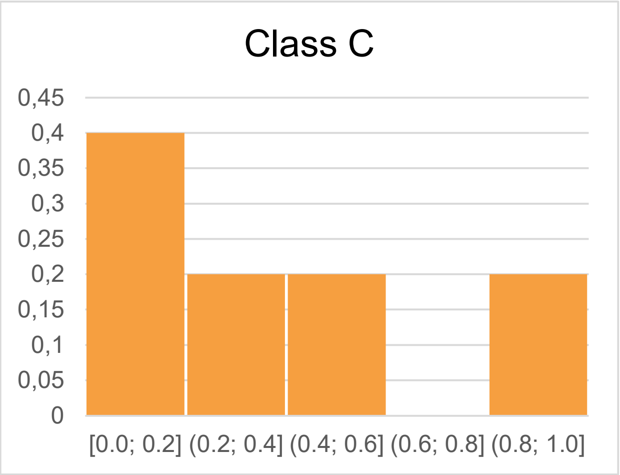
$$P_C = \frac{5}{5+1} = \frac{5}{6}; \quad R_C = \frac{5}{5} = 1; \quad F_C = \frac{2 \times \frac{5}{6} \times 1}{\frac{5}{6} + 1} = \frac{10}{11} \approx 0.9$$

Class A has the lowest F1.

4)







Class A dominates in (0.6; 1.0].

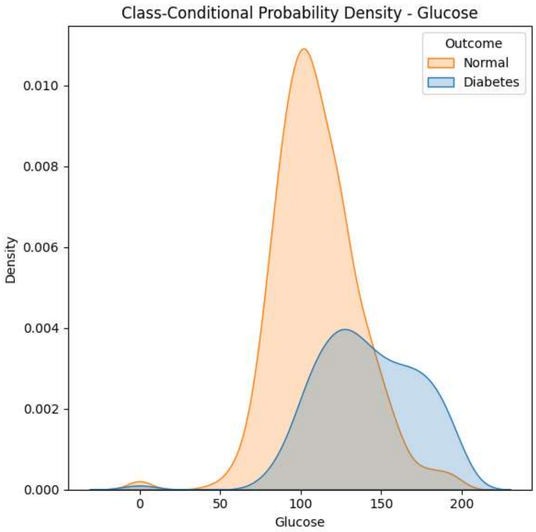
Class B dominates in (0.2; 0.4].

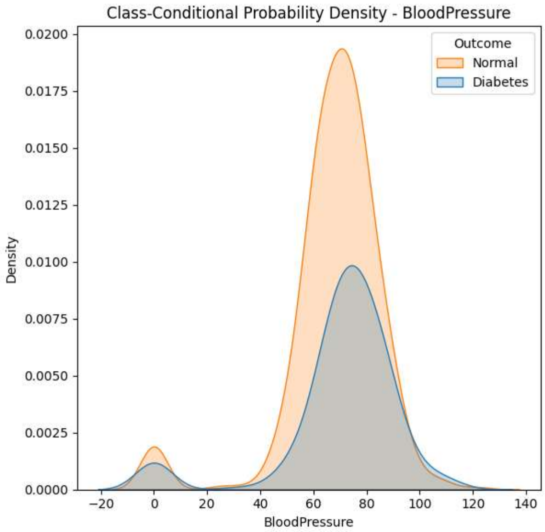
Class C dominates in [0.0; 0.2] and (0.4; 0.6].

Therefore, there will be a 4-ary root split:
[0.0; 0.2]: Class C
(0.2; 0.4]: Class B
(0.4; 0.6]: Class C
(0.6; 1.0]: Class A

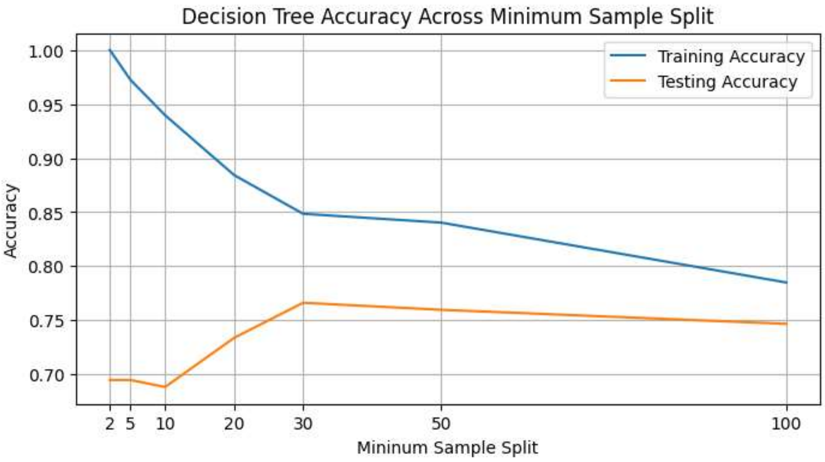
II. Programming and critical analysis

1)
Best feature: Glucose; Worst feature: Blood Pressure





2)



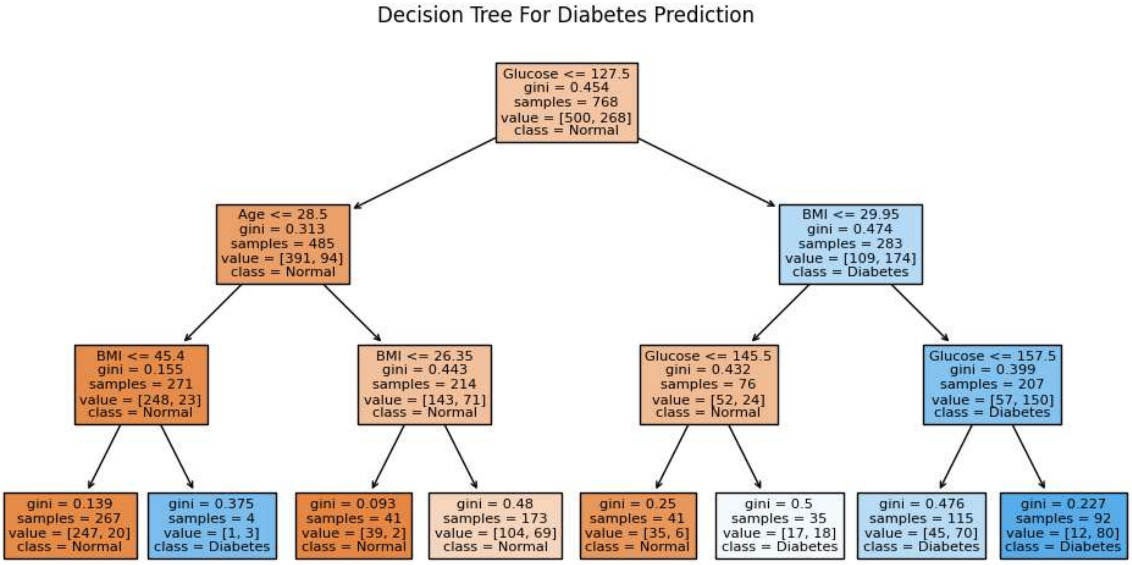
3) When the minimum sample split is low (2 and 5), the training accuracy is high, . On the other hand, the testing accuracy is much lower (around 0.70), which suggests the model is overfitting and does not have a good generalization capacity.

From the minimum sample split values of 5 to 30, the training accuracy drops steadily, which is expected considering the threshold makes the tree more constrained, preventing it from overfitting and increasing the model's generalization capacity. This results in a testing accuracy increase which peaks at 30.

From that point, the training accuracy continues to decrease, for the same reasons mentioned earlier, but, unlike before, the testing accuracy also decreases, indicating underfitting. This is due to the fact that the high minimum sample splits causes an oversimplification of the model which will lack the capacity to capture important relationships in the data.

In conclusion, the model suffers from overfitting and poor generalization capacity when the minimum sample split is low and from underfitting when it's too high. The optimal generalization performance is around the minimum sample split of 30, since the gap between training and testing accuracy is lower and both are relatively high.

4)
i.



ii.
Glucose higher than 127.5 are strongly associated with diabetes.
Individuals with Glucose below 127.5, but also Age below 28.5 and a BMI higher than 45.4 are also associated with diabetes.

In conclusion, diabetes is generally characterized by high Glucose levels or low Glucose with low Age and high BMI.

END