HW (1)

I

1) $IG(y_i, y_{out}) = H(y_{out}) - H(y_{out} \mid y)$

$$H_{y_{out}} = -\left(\frac{3}{12}\log\frac{3}{12} + \frac{4}{12}\log\frac{4}{12} + \frac{5}{12}\log\frac{5}{12}\right)$$

$$H_{y_{out}} = 1{,}55458$$

y_2:

$$H(y_{out} \mid y_2) = \frac{6}{12}S_0 + \frac{4}{12}S_1 + \frac{2}{12}S_2$$

$$= \frac{6}{12}\left(-\frac{1}{6}\log_2\frac{1}{6} - \frac{3}{6}\log_2\frac{3}{6} - \frac{2}{6}\log_2\frac{2}{6}\right)$$

$$+ \frac{4}{12}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)$$

$$+ \frac{2}{12}\underbrace{\left(\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)}_{\hookrightarrow\ = 0} = 1{,}229574$$

y_3:

$$H(y_{out} \mid y_3) = \frac{5}{12}S_0 + \frac{4}{12}S_1 + \frac{3}{12}S_2$$

$$= \frac{5}{12}\left(-\frac{2}{5}\log\frac{2}{5} - \frac{2}{5}\log\frac{2}{5} - \frac{1}{5}\log\frac{1}{5}\right)$$

$$+ \frac{3}{12}\left(-\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3}\right)$$

$$+ \frac{4}{12}\left(-\frac{1}{4}\log\frac{1}{4} - \frac{2}{2}\log\frac{2}{2} - \frac{2}{2}\log\frac{2}{2}\right)$$

$$= 1{,}0303$$

y_4:

$$H(y_{out} \mid y_4) = \frac{7}{12}S_0 + \frac{4}{12}S_1 + \frac{1}{12}S_2$$

$$= \frac{7}{12}\left(-\frac{2}{7}\log\frac{2}{7} - \frac{4}{7}\log\frac{4}{7} - \frac{1}{7}\log\frac{1}{7}\right)$$

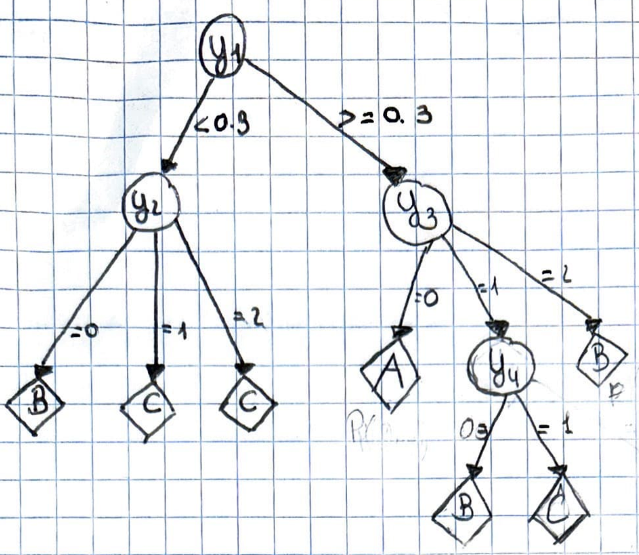
$$+ \frac{4}{12}\left(-\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4}\right)$$

$$+ \frac{1}{12}\underbrace{\left(-\frac{1}{1}\log\frac{1}{1}\right)}_{\hookrightarrow\ = 0} = 1{,}074716$$

$IG(y_3, y_{out}) = 1{,}55458 - 1{,}0303 = 0{,}52428$ ← maior

$IG(y_4, y_{out}) = 1{,}55458 - 1{,}074716 = 0{,}479864$

$IG(y_2, y_{out}) = 1{,}55458 - 1{,}229574 = 0{,}325006$



2-

Reality

		A	B	C
Expected	A	2	0	0
	B	0	4	0
	C	1	0	5

3- $\frac{1}{F_\alpha} = \alpha \left(\frac{1}{\beta} \right) + (1 . \alpha) \left(\frac{1}{r} \right)$

$F_1 : \frac{1}{F_1} = \frac{1}{2} \left(\frac{1}{P} + \frac{1}{r} \right) \Rightarrow F_1 = \frac{2 \times P \times R}{R + P}$

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

$Prec_A = \frac{TP_A}{TP_A + FP_A} = \frac{2}{2+0} = 1$ $\qquad$ $Rec_A = \frac{TP_A}{TP_A + FN_A} = \frac{2}{2+1} = \frac{2}{3}$

$Prec_B = \frac{TP_B}{TP_B + FP_B} = \frac{4}{4+0} = 1$ $\qquad$ $Rec_B = \frac{TP_B}{TP_B + FN_B} = \frac{4}{4+0} = 1$

$Prec_C = \frac{TP_C}{TP_C + FP_C} = \frac{5}{5+0} = 1$ $\qquad$ $Rec_C = \frac{TP_C}{TP_C + FN_C} = \frac{5}{5+1} = \frac{5}{6}$

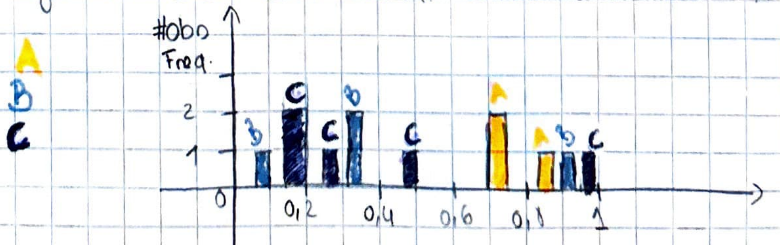
$F_{1A} = \frac{2 \times 1 \times \frac{2}{3}}{1 + \frac{2}{3}} = \underbrace{\frac{4}{5} = 0{,}8}_{\text{Lowest}}$ $\qquad$ $F_{1B} = \frac{2 \times 1 \times 1}{1 + 1} = 1$

$F_{1C} = \frac{2 \times 1 \times \frac{5}{6}}{1 + \frac{5}{6}} = 1{,}81818$

$F_{1A} < F_{1C} < F_{1B}$, logo a classe com menor valor é a A.

4-
Divisão da Raiz

]0; 0,2[-> C
]0,2; 0,4[-> B
]0,4; 0,6[-> C
]0,6; 0,8[-> A
]0,8; 1 [-> A

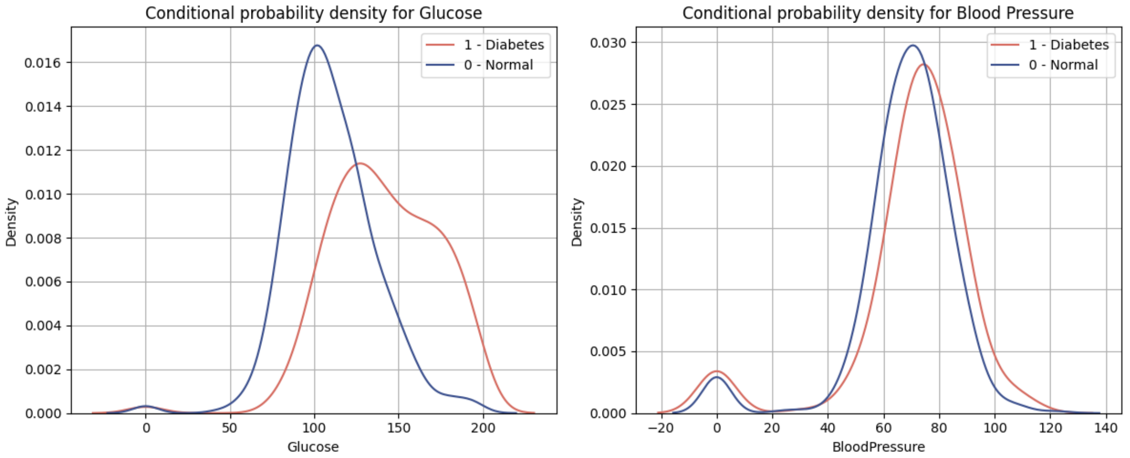


Parte II

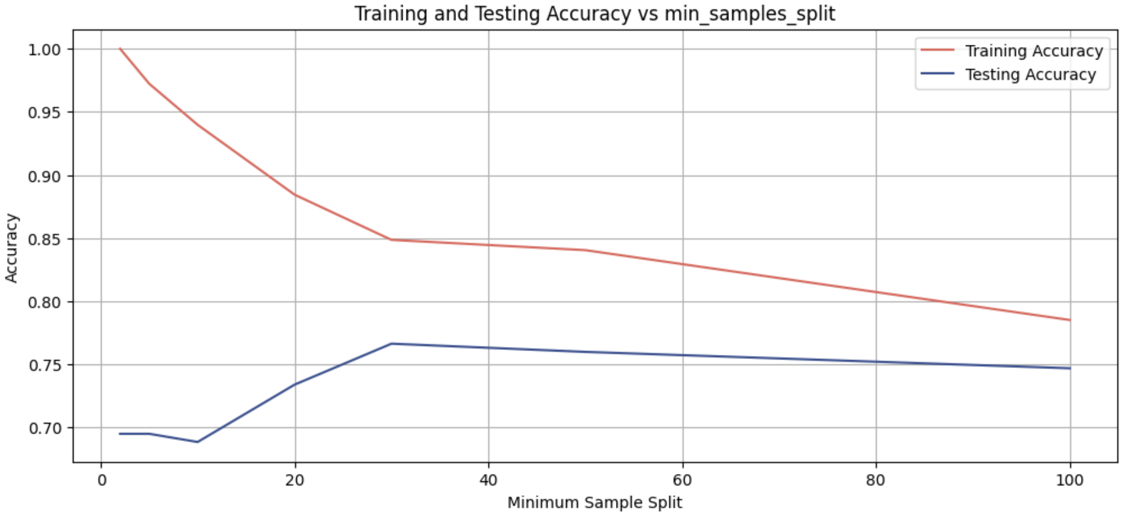
1.

Input variables with the best discriminative power. Glucose

Input variable with the worst discriminative power: BloodPressure



2.



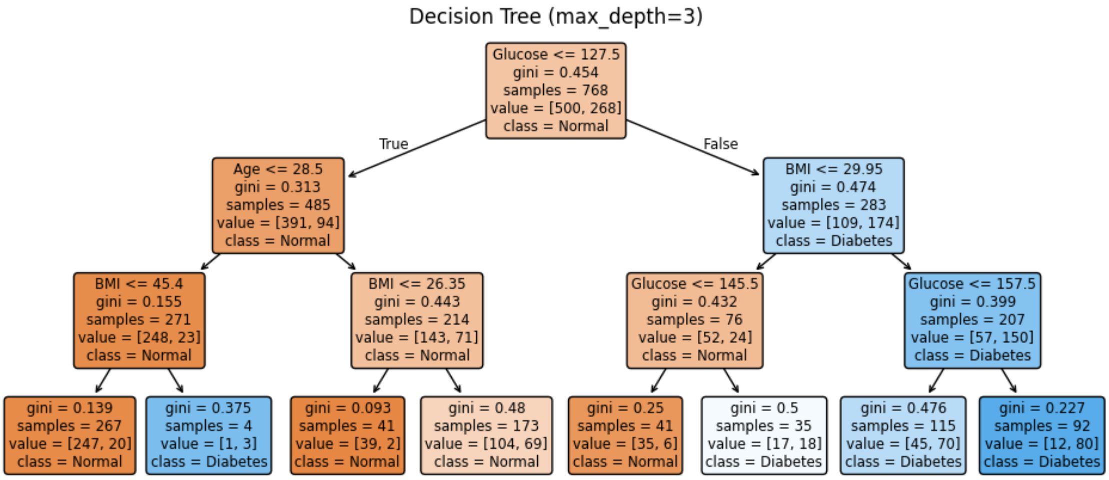
3.

As the minimum sample split increases, training accuracy declines, with a sharp drop initially until the split reaches around 30, after which the decrease becomes more gradual. In contrast, testing accuracy improves as the minimum sample split increases, peaking around the same value (approximately 30), before gradually declining. This pattern indicates that a lower minimum sample split leads to overfitting, where the model closely matches the training data but performs poorly on unseen data.

Increasing the minimum sample split helps reduce overfitting, resulting in better generalization and improved testing accuracy. However, beyond a certain point, the model starts to underfit, leading to a decline in both training and testing performance. Therefore, an optimal balance between overfitting and underfitting is achieved around a minimum sample split of 30, where testing accuracy is highest.

4.

I.



II.

In this decision tree model, the primary factor characterizing diabetes is glucose levels above 127.5. When glucose exceeds this threshold, the model further splits based on BMI and glucose levels again to refine the prediction. This indicates that individuals with high glucose levels, especially those with elevated BMI, are more likely to have diabetes. These conditional associations suggest that high glucose and BMI are the most significant predictors of diabetes in the dataset.

On the other hand, glucose levels below 127.5 are the key factor characterizing non-diabetic individuals. In this case, the model also considers Age and glucose in subsequent splits to predict a normal outcome. This implies that lower glucose levels, coupled with Age, reduce the likelihood of diabetes. Therefore, the model highlights glucose levels as the most critical feature for determining both diabetic and non-diabetic outcomes, with Age serving as an additional important factor.