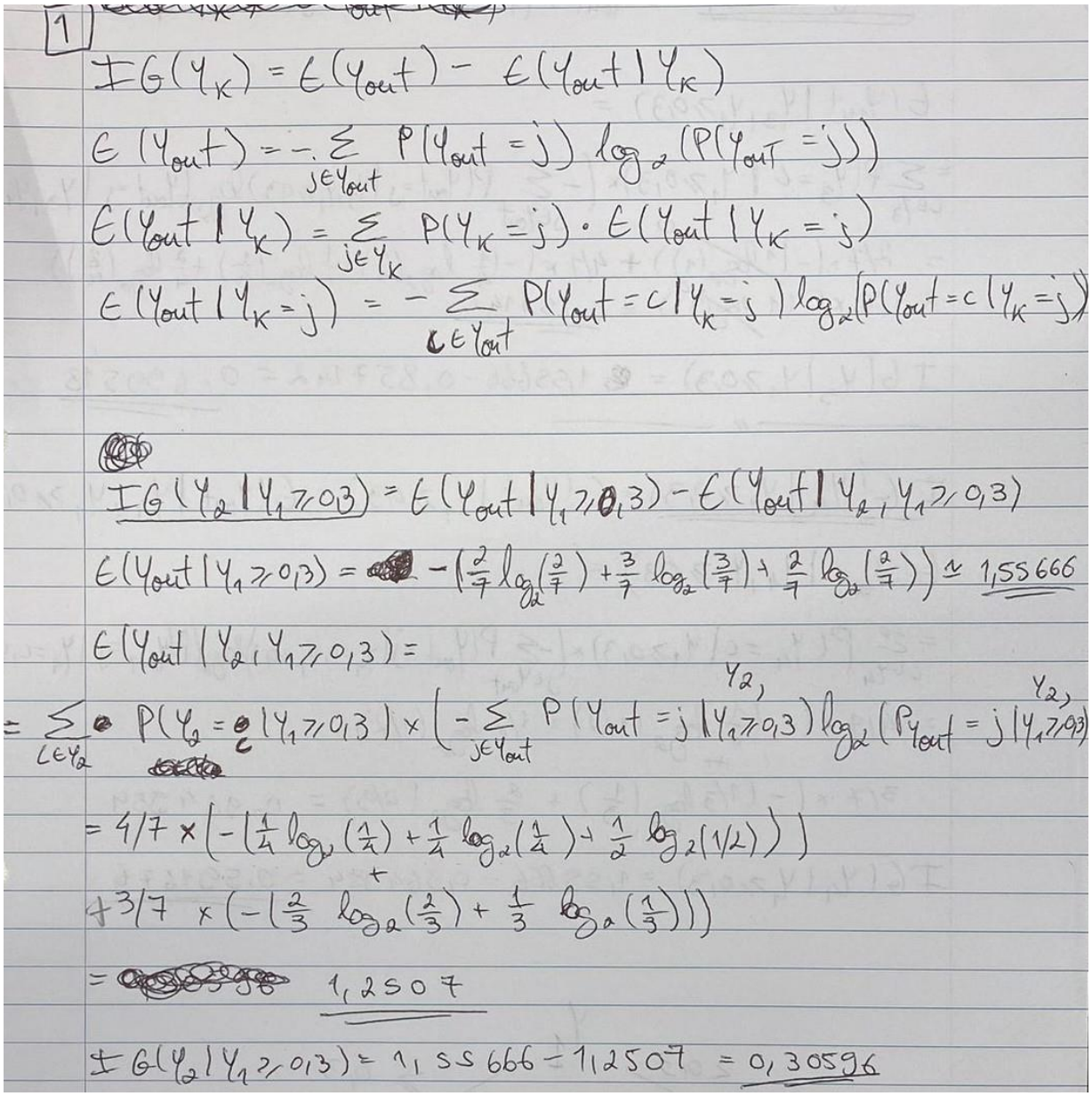
Aprendizagem 2024/25

Homework I – Group 036

(ist1106468, ist1106324)

I. Pen-and-paper

1)



1

$IG(Y_k) = E(Y_{out}) - E(Y_{out} | Y_k)$

$E(Y_{out}) = -\sum_{j \in Y_{out}} P(Y_{out} = j) \log_2 (P(Y_{out} = j))$

$E(Y_{out} | Y_k) = \sum_{j \in Y_k} P(Y_k = j) \cdot E(Y_{out} | Y_k = j)$

$E(Y_{out} | Y_k = j) = -\sum_{c \in Y_{out}} P(Y_{out} = c | Y_k = j) \log_2 (P(Y_{out} = c | Y_k = j))$

$IG(Y_2 | Y_1 \geq 0{,}3) = E(Y_{out} | Y_1 \geq 0{,}3) - E(Y_{out} | Y_2, Y_1 \geq 0{,}3)$

$E(Y_{out} | Y_1 \geq 0{,}3) = -\left(\frac{2}{7} \log_2 \left(\frac{2}{7}\right) + \frac{3}{7} \log_2 \left(\frac{3}{7}\right) + \frac{2}{7} \log_2 \left(\frac{2}{7}\right)\right) \simeq 1{,}55666$

$E(Y_{out} | Y_2, Y_1 \geq 0{,}3) =$

$= \sum_{c \in Y_2} P(Y_2 = c | Y_1 \geq 0{,}3) \times \left(-\sum_{j \in Y_{out}} P(Y_{out} = j | Y_1 \geq 0{,}3)^{Y_2,} \log_2 (P_{Y_{out} = j | Y_1 \geq 0{,}3}^{Y_2,})\right)$

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

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

$= 1{,}2507$

$IG(Y_2 | Y_1 \geq 0{,}3) = 1{,}55666 - 1{,}2507 = 0{,}30596$

Homework I – Group 036

(ist1106468, ist1106324)

$IG(Y_3 | Y_1 \geq 0{,}3) = E(Y_{out} | Y_1 \geq 0{,}3) - E(Y_{out} | Y_3, Y_1 \geq 0{,}3)$

$E(Y_{out} | Y_3, Y_1 \geq 0{,}3) =$

$= \sum_{c \in Y_3} P(Y_3 = c | Y_1 \geq 0{,}3) \times \left(- \sum_{j \in Y_{out}} P(Y_{out} = j | Y_3, Y_1 \geq 0{,}3) \log_2 (Y_{out} = j | Y_3, Y_1 \geq 0{,}3) \right)$

$= 2/7 \times (-(1 \log_2 (1))) + 4/7 \times \left(-\left(\frac{1}{4} \log_2 \left(\frac{1}{4}\right) + \frac{1}{4} \log_2 \left(\frac{1}{4}\right) + \frac{2}{4} \log_2 \left(\frac{2}{4}\right)\right)\right)$
$+ 1/7 \times (-(1 \log_2 (1))) = 0{,}857142$

$IG(Y_3 | Y_1 \geq 0{,}3) = 1{,}55666 - 0{,}857142 = 0{,}699518$

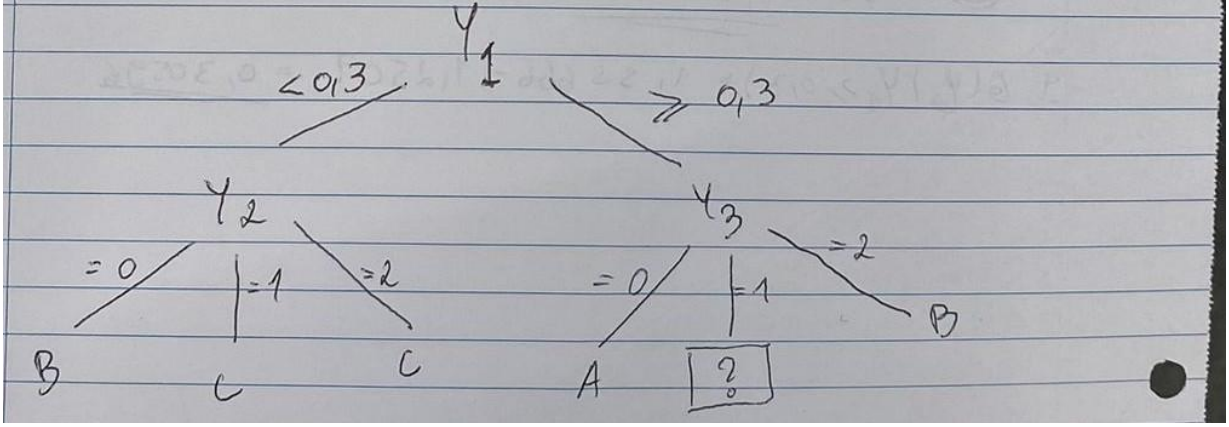
$IG(Y_4 | Y_1 \geq 0{,}3) = E(Y_{out} | Y_1 \geq 0{,}3) - E(Y_{out} | Y_4, Y_1 \geq 0{,}3)$

$E(Y_{out} | Y_4, Y_1 \geq 0{,}3) =$

$= \sum_{c \in Y_4} P(Y_4 = c | Y_1 \geq 0{,}3) \times \left(- \sum_{j \in Y_{out}} P(Y_{out} = j | Y_4 = c, Y_1 \geq 0{,}3) \log_2 (P(Y_{out} = j | Y_4 = c, Y_1 \geq 0{,}3)) \right)$

$= 4/7 \times \left(- \left(\frac{1}{2} \log_2 (1/2) + 1/2 \log_2 (1/2)\right)\right)$
$+ 3/7 \times \left(- \left(1/3 \log_2 \left(\frac{1}{3}\right) + \frac{2}{3} \log_2 (2/3)\right)\right) = 0{,}964984$

$IG(Y_4 | Y_1 \geq 0{,}3) = 1{,}55666 - 0{,}964984 = 0{,}591676$



Aprendizagem 2024/25

Homework I – Group 036

(ist1106468, ist1106324)

$IG(Y_2 \mid Y_1 \geq 0{,}3, Y_3 = 1) =$

$= E(Y_{out} \mid Y_1 \geq 0{,}3, Y_3 = 1) - E(Y_{out} \mid Y_1 \geq 0{,}3, Y_3 = 1, Y_2)$

$E(Y_{out} \mid Y_1 \geq 0{,}3, Y_3 = 1) =$

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

$E(Y_{out} \mid Y_1 \geq 0{,}3, Y_3 = 1, Y_2) =$

$= 1 \times E(Y_{out} \mid Y_2 = 0, Y_1 \geq 0{,}3, Y_3 = 1) = 1{,}5$

$IG(Y_2 \mid Y_1 \geq 0{,}3, Y_3 = 1) = 1{,}5 - 1{,}5 = 0$

$IG(Y_4 \mid Y_1 \geq 0{,}3, Y_3 = 1) =$

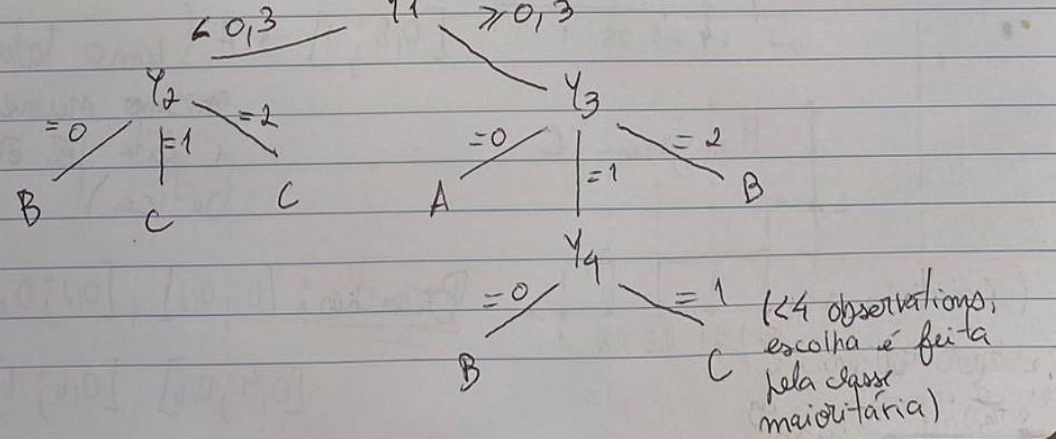
$= E(Y_{out} \mid Y_1 \geq 0{,}3, Y_3 = 1) - E(Y_{out} \mid Y_1 \geq 0{,}3, Y_3 = 1, Y_4)$

$E(Y_{out} \mid Y_1 \geq 0{,}3, Y_3 = 1, Y_4) =$

$= 1/4 \times E(Y_{out} \mid Y_4 = 0, Y_1 \geq 0{,}3, Y_3 = 1) + 3/4\, E(Y_{out} \mid Y_4 = 1, Y_1 \geq 0{,}3, Y_3 = 1)$

$= 3/4\left(-\left(1/3\log_2(1/3) + 2/3\log_2(2/3)\right)\right) = 0{,}688712$

$IG(Y_4 \mid Y_1 \geq 0{,}3, Y_3 = 1) = 1{,}5 - 0{,}918296 = 0{,}8113$

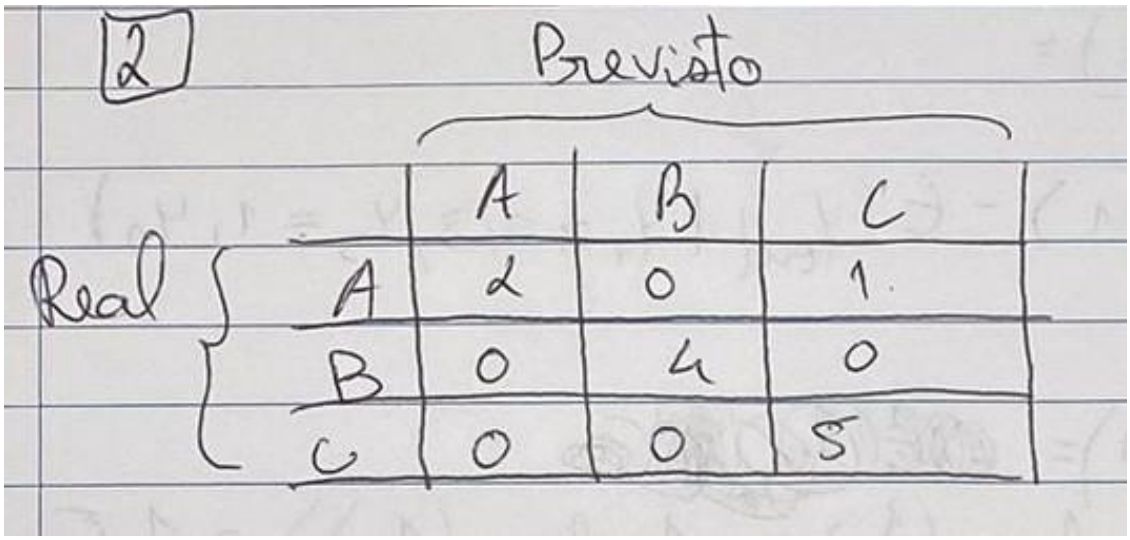


Aprendizagem 2024/25

Homework I – Group 036

(ist1106468, ist1106324)

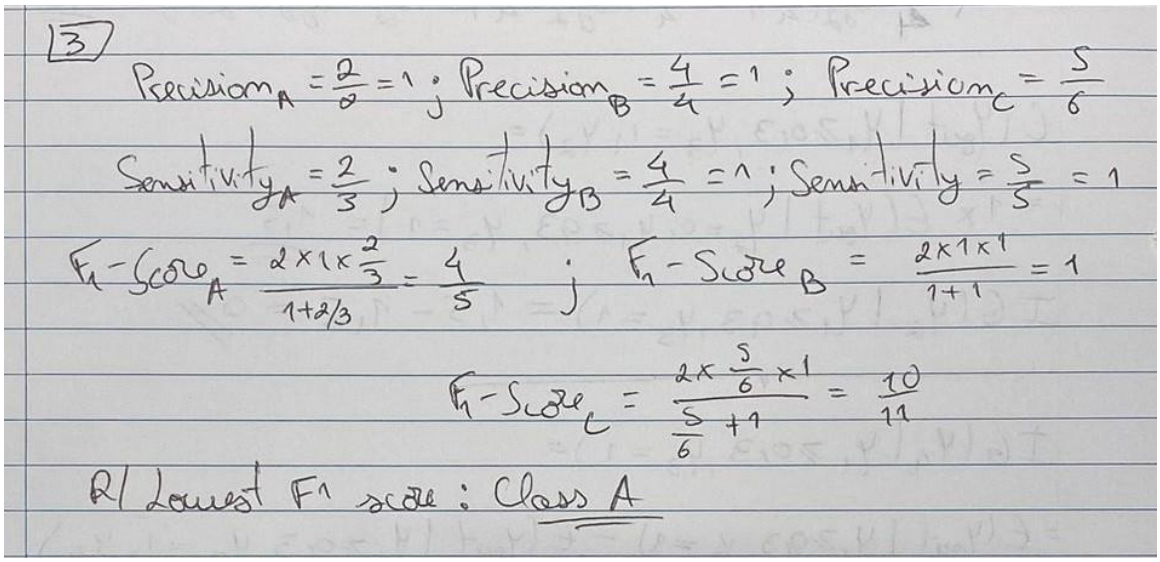
2)



2

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

3)



3

$Precision_A = \frac{2}{2} = 1$; $Precision_B = \frac{4}{4} = 1$; $Precision_C = \frac{5}{6}$

$Sensitivity_A = \frac{2}{3}$; $Sensitivity_B = \frac{4}{4} = 1$; $Sensitivity = \frac{5}{5} = 1$

$F_1\text{-}Score_A = \frac{2 \times 1 \times \frac{2}{3}}{1+2/3} = \frac{4}{5}$; $F_1\text{-}Score_B = \frac{2 \times 1 \times 1}{1+1} = 1$

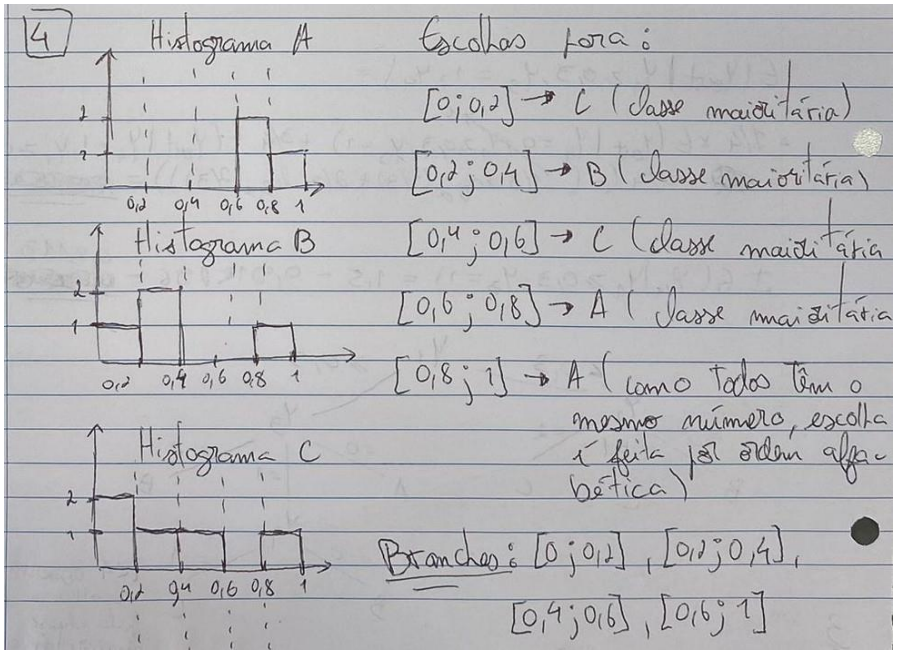
$F_1\text{-}Score_C = \frac{2 \times \frac{5}{6} \times 1}{\frac{5}{6}+1} = \frac{10}{11}$

R/ Lowest F1 score : Class A

Homework I – Group 036

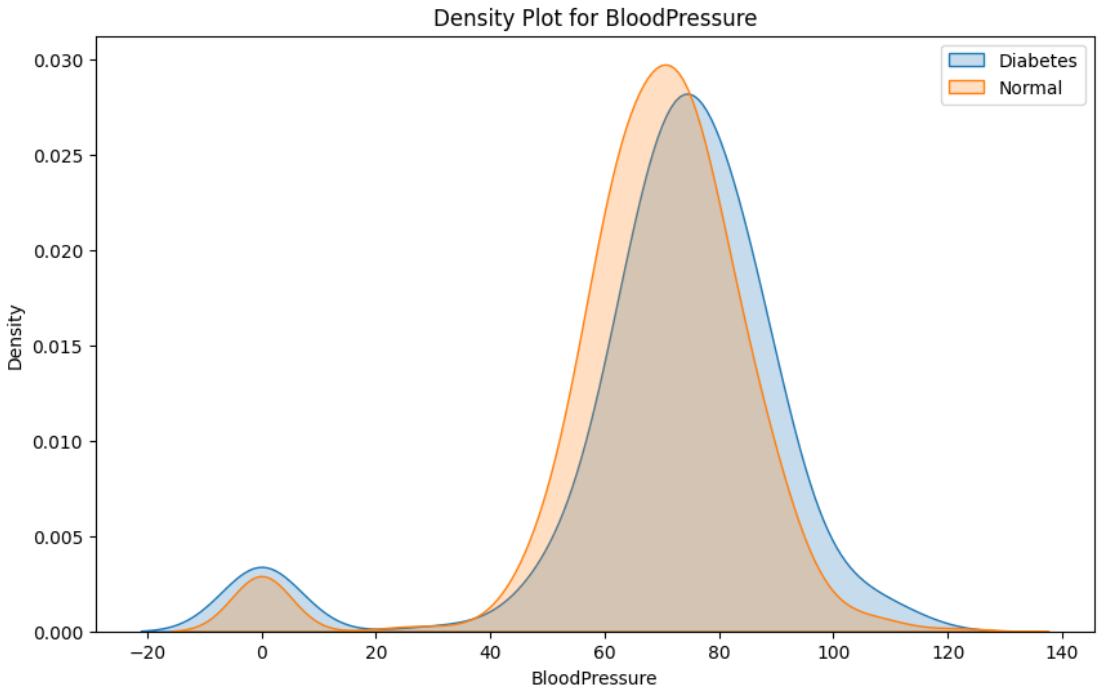
(ist1106468, ist1106324)

4)



II. Programming and critical analysis

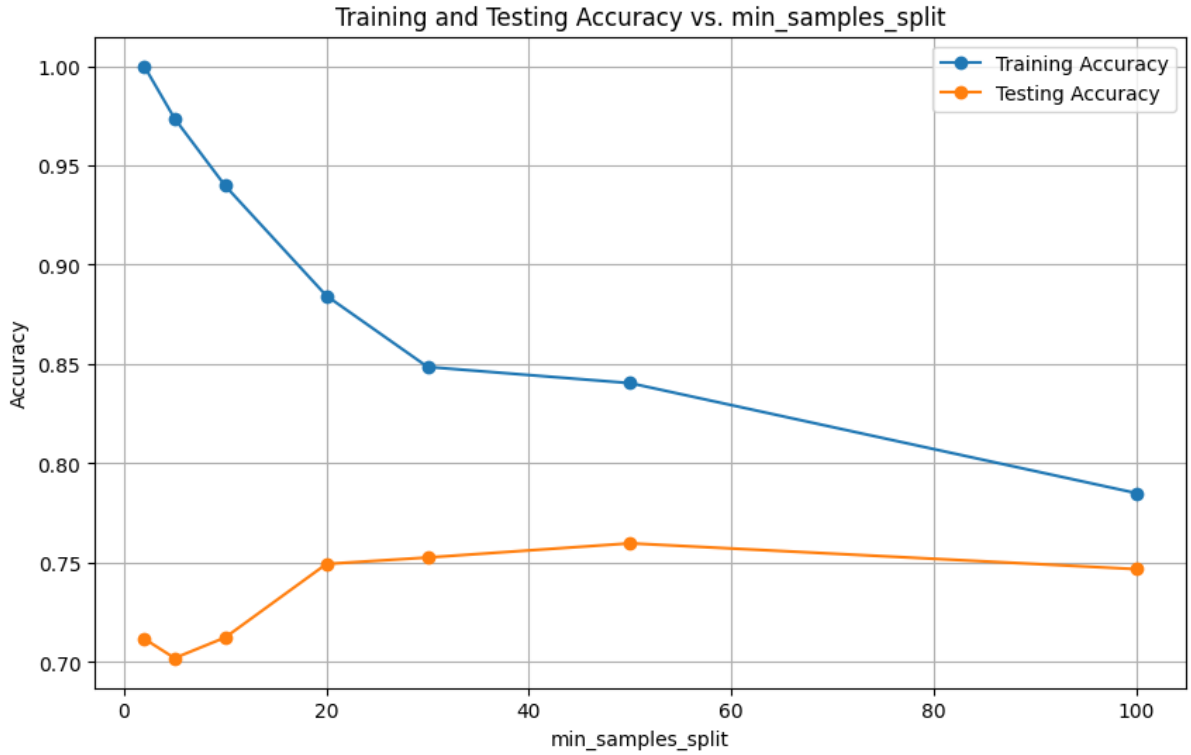
5) Answer 5



Homework I – Group 036

(ist1106468, ist1106324)

6) Answer 6



7) Answer 7

Os resultados obtidos ao variar o parâmetro min_samples_split na árvore de decisão revelam informações importantes sobre a capacidade de generalização do modelo.

Quando min_samples_split toma **valores baixos**, a árvore de decisão torna-se excessivamente complexa, pois ocorrem muitos splits. Isto permite que a árvore se ajuste de forma muito precisa aos dados de treino, o que leva a uma alta precisão nesse conjunto.
No entanto, essa complexidade resulta em overfitting, o que significa que o modelo não generaliza bem para dados novos. Como consequência, a precisão nos dados de teste tende a ser significativamente inferior, tal como se verifica no gráfico acima, uma vez que o modelo está excessivamente adaptado aos dados de treino e não consegue captar as variações presentes em dados exteriores.

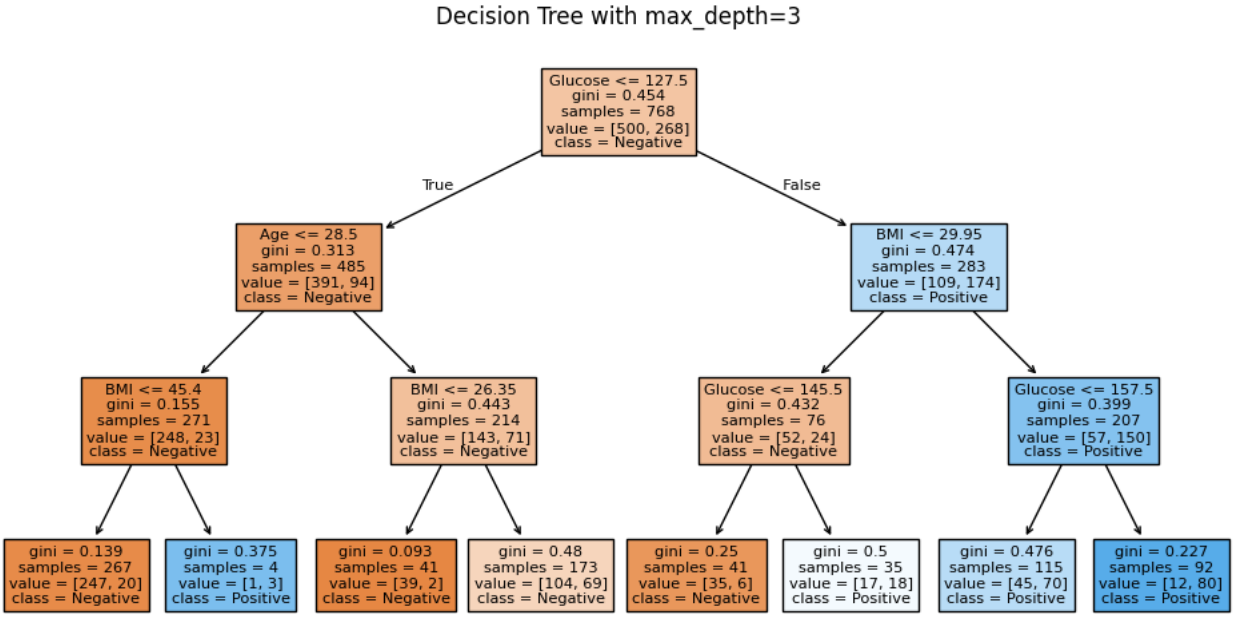
Quando min_samples_split é definido para **valores mais altos**, a árvore de decisão torna-se mais simples e menos profunda, pelo facto de se darem menos splits. Isso resulta num modelo mais generalizado, que consegue captar melhor os dados externos, evitando possíveis variações em comparação com os dados de teste.
Como consequência, a precisão nos dados de treino diminui, uma vez que a árvore não está tão overfitted e não capta tanto os "ruídos" e variações presentes nesses dados. Em contrapartida, a precisão nos dados externos tende a aumentar, pois o modelo é mais abrangente e generalizável.

Aprendizagem 2024/25

Homework I – Group 036

(ist1106468, ist1106324)

8) Answer 8



A análise da árvore de decisão revela que os principais fatores de classificação para pessoas diabéticas são **a glicose**, **o IMC (Índice de Massa Corporal)** e **a idade**. Para verificarmos isso com mais detalhe em cada folha, vamos analisar qual a probabilidade de o paciente ser diabético se tiver certas características determinadas pelos nós das árvores:

Quando:

- **Glicose ≤ 125.5**:
  - **Idade ≤ 28.5**:
    - **BMI ≤ 45.4**: 7.50%
    - **BMI ≥ 45.4**: 75.00%
  - **Idade ≥ 28.5**:
    - **BMI ≤ 26.35**: 4.88%
    - **BMI ≥ 26.35**: 39.88%
- **Glicose ≥ 125.5**:
  - **BMI ≤ 29.95**:
    - **Glicose ≤ 145.5**: 14.63%
    - **Glicose ≥ 145.5**: 51.43%
  - **BMI ≥ 29.95**:
    - **Glicose ≤ 157.5**: 60.87%
    - **Glicose ≥ 157.5**: 86.96%

Aprendizagem 2024/25

Homework I – Group 036

(ist1106468, ist1106324)

Análise Detalhada dos Resultados:
- No **primeiro nó**, a divisão é feita com base nos níveis de **glicose**, indicando que este é um fator forte na previsão de diabetes.
- Quando os níveis de glicose são menores ou iguais a **125.5**, a próxima divisão é realizada com base na **idade**. Caso a glicose seja maior, o próximo critério utilizado é o **IMC**. Isso mostra que **glicose**, **IMC** e **idade** são fatores significativos na determinação do risco de diabetes.

Nas subfolhas, **glicose** e **IMC** são novamente utilizados para as divisões, reforçando o seu poder de previsão no diagnóstico de diabetes. Observamos que:
- Uma pessoa com **glicose elevada** e **IMC elevado** tem uma probabilidade muito maior de ser diabética, em comparação com aquelas que apresentam baixos valores de ambos os fatores.
- No entanto, mesmo indivíduos com **glicose baixa** ($\leq$ 125.5), se forem **mais velhos** (idade $>$ 28.5) e tiverem um **IMC elevado** também apresentam maior probabilidade de desenvolver diabetes.

Percebemos, portanto, que **altos níveis de glicose** e **IMC elevado** são fatores decisivos no diagnóstico de diabetes. Para além disso, a **idade** também exerce um impacto significativo, indicando que pessoas mais velhas tendem a ter maior predisposição para a doença, embora o seu efeito não seja tão marcante quanto os níveis de glicose e IMC.

END