

I. Pen-and-paper

- 1) Para completar a árvore de decisão que nos foi dada, apenas precisamos de considerar as observações nas quais $y_1 \geq 0,3$. Assim, analisamos as seguintes observações:

D	y_2	y_3	y_4	y_{out}
x_6	0	1	0	B
x_7	0	1	1	A
x_8	1	0	0	A
x_9	0	1	1	C
x_{10}	0	1	1	C
x_{11}	1	0	0	A
x_{12}	1	2	0	B

Calculamos a Entropia de Shannon de y_{out} para estas observações:

$$H(y_{out}) = -\frac{3}{7} \cdot \log_2\left(\frac{3}{7}\right) - \frac{2}{7} \cdot \log_2\left(\frac{2}{7}\right) - \frac{2}{7} \cdot \log_2\left(\frac{2}{7}\right) \approx 1.5567$$

E o Ganho de Informação para cada uma das variáveis:

$$\begin{aligned} IG(y_2) &= H(y_{out}) - \left(\frac{4}{7} \cdot H(y_{out}|0) + \frac{3}{7} \cdot H(y_{out}|1) \right) = \\ &= 1.5567 - \frac{4}{7} \left(-\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) - \frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) - \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right) \right) - \\ &\quad - \frac{3}{7} \left(-\frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) - \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) \right) \approx \\ &\approx 0.31 \end{aligned}$$

$$\begin{aligned} IG(y_3) &= H(y_{out}) - \left(\frac{2}{7} \cdot H(y_{out}|0) + \frac{4}{7} \cdot H(y_{out}|1) + \frac{1}{7} \cdot H(y_{out}|2) \right) = \\ &= 1.5567 - \frac{2}{7} \cdot (0) - \frac{4}{7} \cdot \left(-\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) - \frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) - \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right) \right) - \frac{1}{7} \cdot (0) \approx \\ &\approx 0.70 \end{aligned}$$

$$\begin{aligned} IG(y_4) &= H(y_{out}) - \left(\frac{4}{7} \cdot H(y_{out}|0) + \frac{3}{7} \cdot H(y_{out}|1) \right) = \\ &= 1.5567 - \frac{4}{7} \cdot (1) - \frac{3}{7} \cdot \left(-\frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right) - \frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) \right) \approx \\ &\approx 0.59 \end{aligned}$$

Como y_3 tem o maior Ganho de Informação entre os calculados, é essa variável que vamos escolher para continuar a nossa árvore de decisão.

Atentando nos valores de y_{out} consoante os valores de y_3 , notamos que $y_{out} = A$ quando $y_3 = 0$ e que $y_{out} = B$ quando $y_3 = 2$. No entanto, não temos nenhuma certeza acerca do valor de y_{out} quando $y_3 = 1$, então vamos dividir esse nó, visto que temos 4 observações.

Basta-nos, a partir de agora, considerar as observações em que $y_1 \geq 0,3$ e $y_3 = 1$:

D	y_2	y_4	y_{out}
x_6	0	0	B
x_7	0	1	A
x_9	0	1	C
x_{10}	0	1	C

Calculamos novamente a Entropia de Shannon de y_{out} para estas observações:

$$H(y_{out}) = -\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) \cdot 2 - \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right) = 1.5$$

E agora o Ganho de Informação para as variáveis:

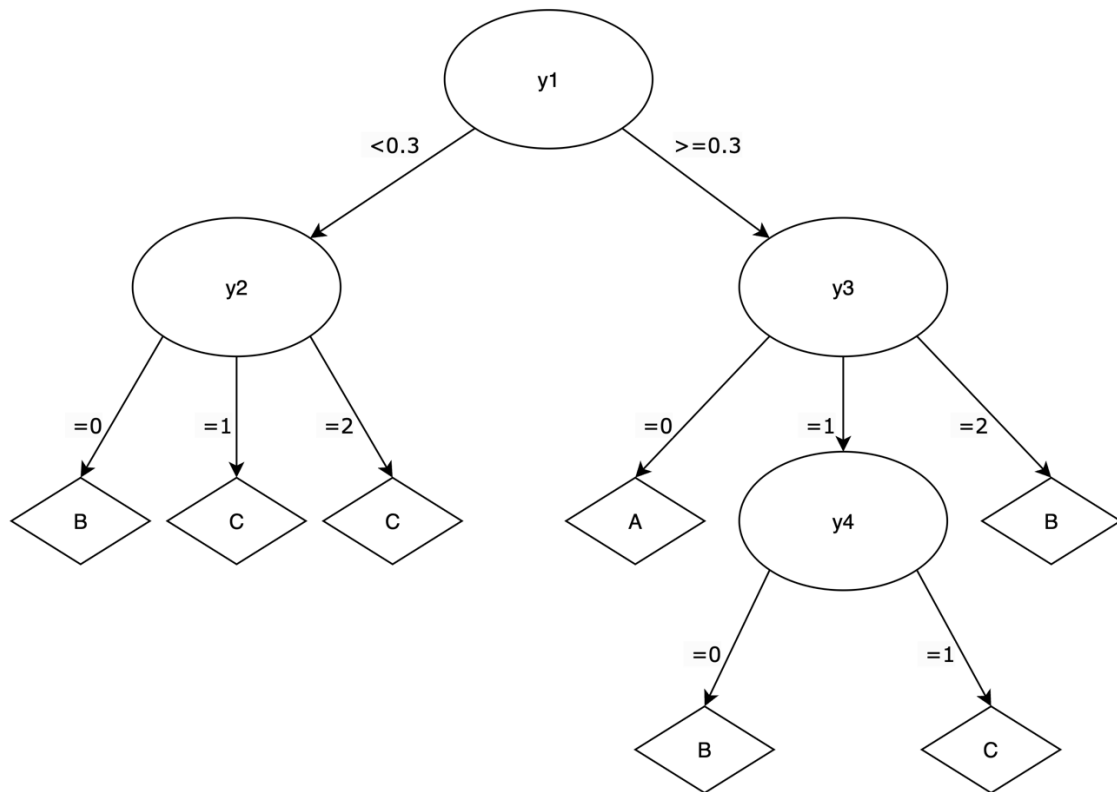
$$\begin{aligned} IG(y_2) &= H(y_{out}) - H(y_{out}|0) = \\ &= H(y_{out}) - \left(\frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) + \frac{1}{4} \cdot \log_2\left(\frac{1}{4}\right) + \frac{2}{4} \cdot \log_2\left(\frac{2}{4}\right)\right) = \\ &= 1.5 - 1.5 = 0 \end{aligned}$$

$$\begin{aligned} IG(y_4) &= H(y_{out}) - \left(\frac{1}{4} \cdot H(y_{out}|0) + \frac{3}{4} \cdot H(y_{out}|1)\right) = \\ &= 1.5 - \frac{1}{4} \cdot (0) - \frac{3}{4} \cdot \left(-\frac{1}{3} \cdot \log_2\left(\frac{1}{3}\right) - \frac{2}{3} \cdot \log_2\left(\frac{2}{3}\right)\right) \approx \\ &\approx 0.81 \end{aligned}$$

Como y_4 tem o maior Ganho de Informação, é essa variável que vamos escolher para continuar a nossa árvore de decisão.

Ao analisar os valores, notamos facilmente que $y_{out} = B$ quando $y_4 = 0$, mas não temos certezas sobre y_{out} quando $y_4 = 1$. Poderíamos pensar então em dividir o nó, mas desta vez só temos 3 observações que verificam esse caso, então não o podemos fazer.

Temos assim de escolher $y_{out} = C$, uma vez que para as 3 observações em que $y_4 = 1$, obtemos os seguintes valores para y_{out} : {A, C, C}. Como há uma classe que predomina, é essa que escolhemos. Assim, $y_{out} = C$ quando $y_4 = 1$, pelo que a nossa árvore de decisão fica:



- 2) Para desenhar a Matriz de Confusão, utilizamos a árvore de decisão construída na alínea anterior, através da qual podemos prever y_{out} para cada uma das observações e comparar assim os resultados obtidos com os valores reais:

Real	C	B	C	B	C	B	A	A	C	C	A	B
Prev	C	B	C	B	C	B	C	A	C	C	A	B

Com estes dados podemos então construir a matriz de confusão que tem em conta as múltiplas classes:

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

- 3) Para identificar a classe que tem o menor valor de treino de F1 score, precisamos antes de calcular a *Precision* e o *Recall* de cada uma das classes tendo em conta as observações que temos.

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

Este cálculo torna-se bastante simples ao utilizarmos a Matriz de Confusão para identificarmos os valores que necessitamos:

$$\text{Precision}_A = \frac{2}{2+0} = 1 \quad \text{Precision}_B = \frac{4}{4+0} = 1 \quad \text{Precision}_C = \frac{5}{5+1} \approx 0.83$$

$$\text{Recall}_A = \frac{2}{2+1} \approx 0.67 \quad \text{Recall}_B = \frac{4}{4+0} = 1 \quad \text{Recall}_C = \frac{5}{5+0} = 1$$

Calculamos assim a F1 score de cada classe:

$$\text{F-measure} = \frac{1}{\alpha \cdot \left(\frac{1}{\text{Precision}}\right) + (1 - \alpha) \cdot \left(\frac{1}{\text{Recall}}\right)} \quad \text{com } \beta^2 = \frac{1 - \alpha}{\alpha}$$

$$\text{F-measure} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

$$\text{Como } \beta = 1 : \quad \text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1}_A = 2 \cdot \frac{1 \cdot 0.67}{1 + 0.67} \approx 0.80 \quad \text{F1}_B = 2 \cdot \frac{1 \cdot 1}{1 + 1} = 1 \quad \text{F1}_C = 2 \cdot \frac{0.83 \cdot 1}{0.83 + 1} \approx 0.91$$

Concluimos que a classe que tem o menor valor de treino de F1 score é a **classe A**.

- 4) O primeiro passo para resolver este exercício é dividir as observações que temos pelos 5 intervalos igualmente espaçados em $[0, 1]$:

$$[0; 0.2[: x_2, x_3, x_5$$

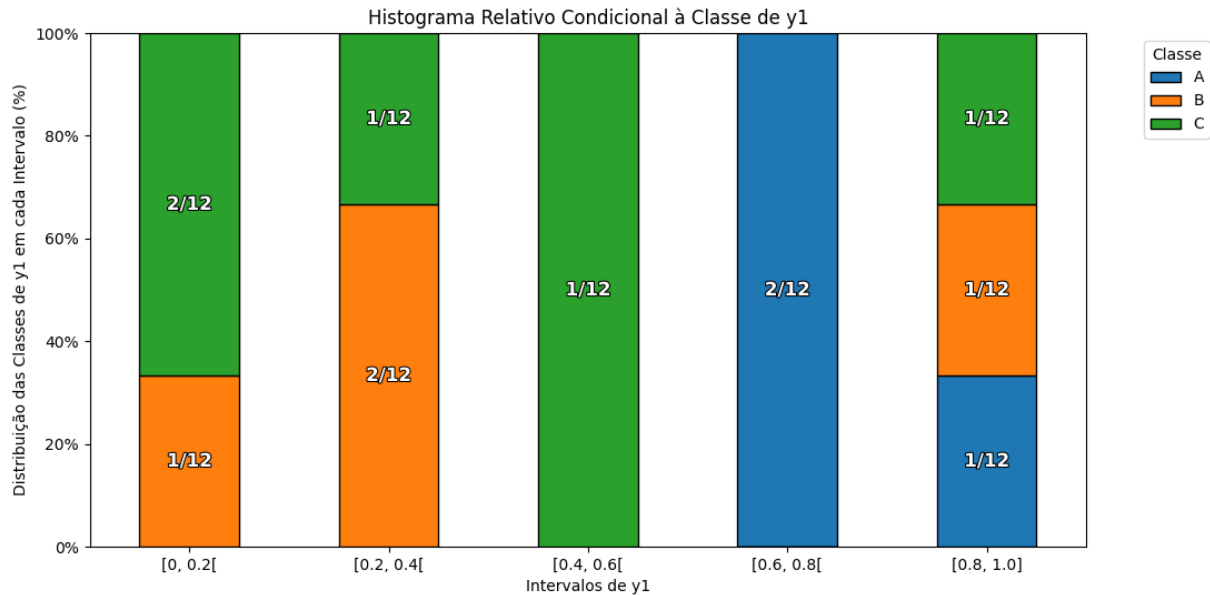
$$[0.2; 0.4[: x_1, x_4, x_6$$

$$[0.4; 0.6[: x_{10}$$

$$[0.6; 0.8[: x_7, x_{11}$$

$$[0.8; 1]: x_8, x_9, x_{12}$$

Agora contruímos o histograma tendo em conta a classe (y_{out}) e o intervalo a que cada pertence cada observação pelo valor de y_1 :



Nota sobre o gráfico: a frequência relativa é dada pelo número de observações de cada classe no respetivo intervalo em relação ao total de observações ($n=12$).

Por fim, para fazer o *n-ary root split*, temos de categorizar os dados das observações dada a frequência de cada classe nos intervalos. Por outras palavras, o objetivo é dividir o intervalo $[0, 1]$ em sub-intervalos menores, onde o modelo retorna a classe predominante para cada um desses sub-intervalos.

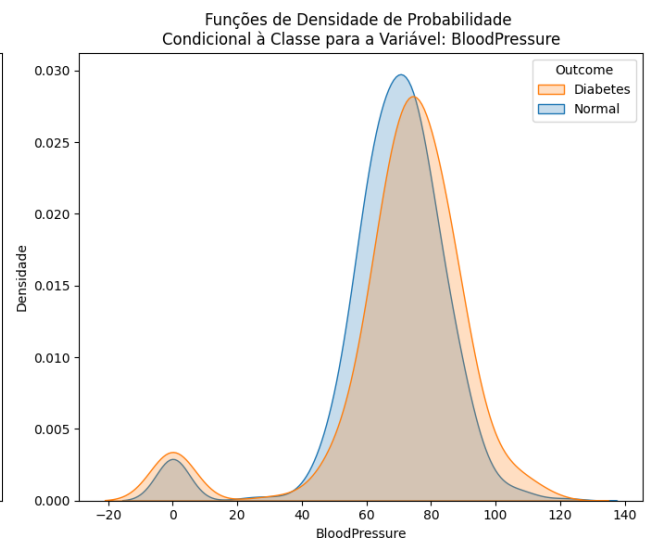
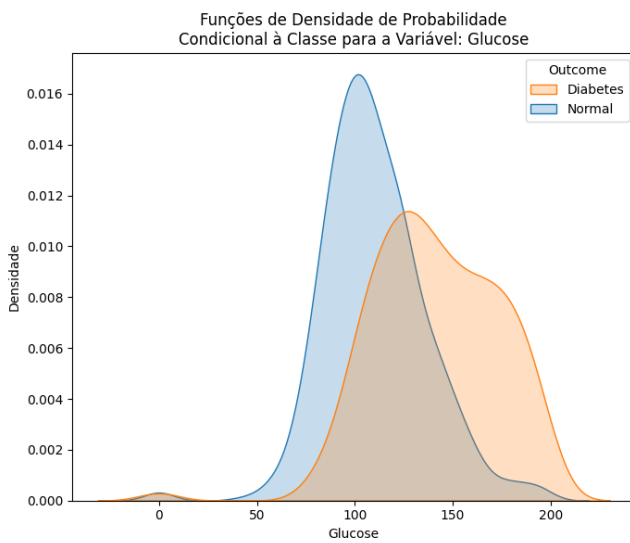
$$\hat{y}_{out}(y_1) = \begin{cases} C, & \text{se } 0 \leq y_1 < 0.2 \\ B, & \text{se } 0.2 \leq y_1 < 0.4 \\ C, & \text{se } 0.4 \leq y_1 < 0.6 \\ A, & \text{se } 0.6 \leq y_1 \leq 1 \\ n.d., & \text{caso contrário} \end{cases}$$

Note-se que, como no intervalo $[0.8; 1]$ todas as classes se encontram na mesma proporção, juntámos os intervalos $[0.6; 0.8]$ e $[0.8; 1]$ num só para já ser possível identificar uma classe predominante. Neste caso, foi a classe A.

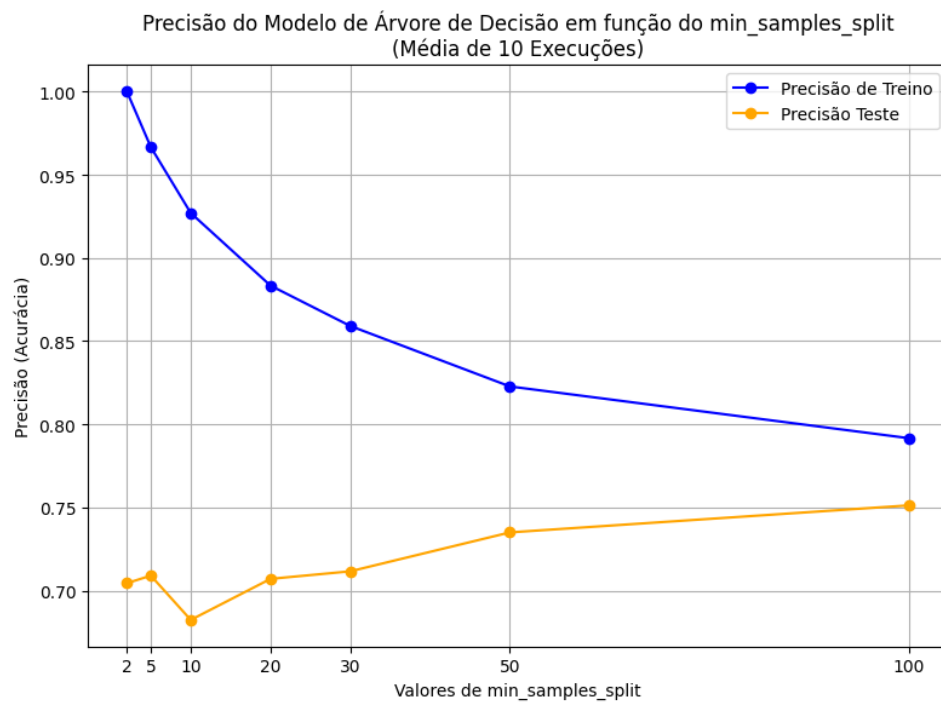
II. Programming and critical analysis

1) Análise do poder discriminativo das variáveis:

- Melhor poder discriminativo: **Glucose** → F-score ≈ 213.162
- Pior poder discriminativo: **BloodPressure** → F-score ≈ 3.257



2) Gráfico obtido:



- 3) O gráfico acima mostra que, para valores baixos de *min_samples_split* (como 2 ou 5), o modelo de árvore de decisão apresenta uma precisão quase perfeita no conjunto de treino, mas esta precisão elevada não se reflete no conjunto de teste. Na verdade, isto sugere um claro **overfitting**, ou seja, o que acontece realmente é que o modelo se ajusta excessivamente aos dados de treino, acabando por capturar padrões irrelevantes. Assim, o desempenho do teste fica comprometido devido à sua fraca capacidade de generalização.

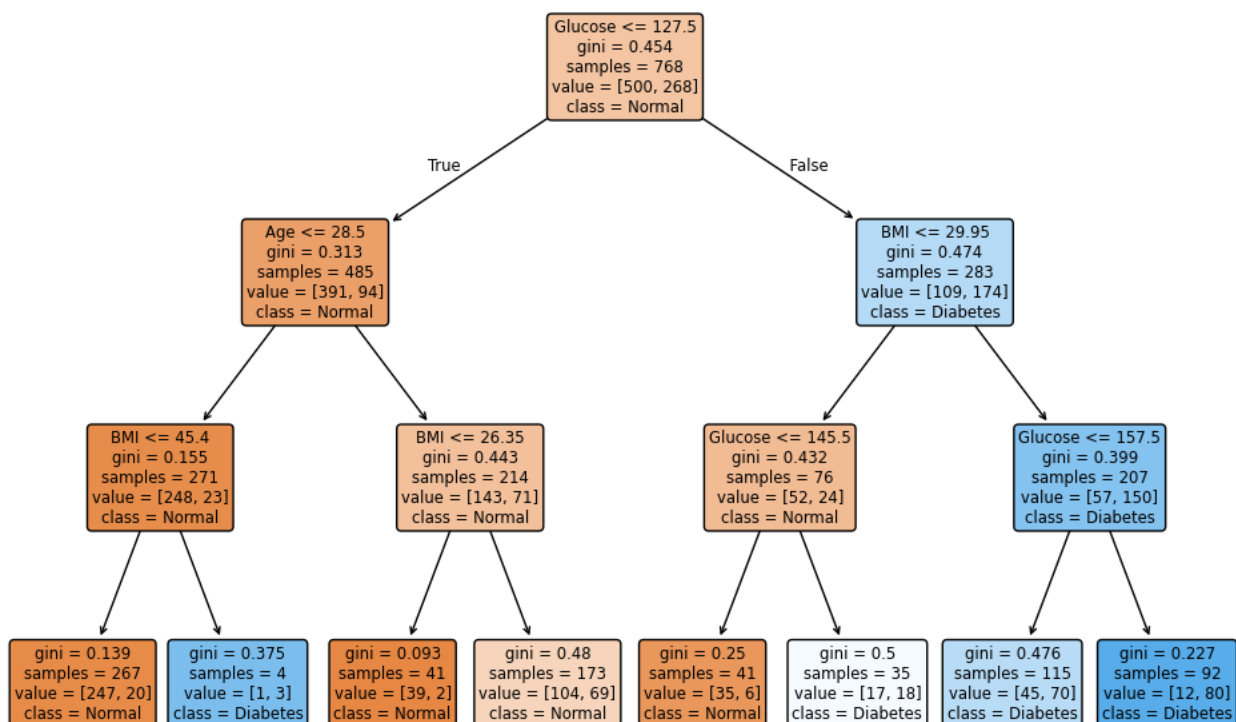
À medida que o valor de *min_samples_split* aumenta, a precisão no treino diminui gradualmente, enquanto que a precisão no teste melhora. Isto indica que o modelo se torna mais equilibrado e generaliza melhor, levando a uma árvore menos complexa. Para valores moderados de *min_samples_split* (como 20, 30 ou 50), verifica-se um equilíbrio entre as precisões de treino e teste, o que indica que o modelo está a capturar as características mais importantes dos dados.

No entanto, com valores mais altos de *min_samples_split* (como 100), observa-se uma convergência das precisões no treino e teste, mas ambas começam a diminuir. Nestas situações diz-se que estamos perante um caso de **underfitting**, isto é, o modelo torna-se excessivamente simples, perdendo a capacidade preditiva tanto no treino quanto no teste e, conseqüentemente, o modelo deixa de conseguir capturar a complexidade necessária para um bom desempenho.

Concluindo, o gráfico demonstra a importância de ajustar corretamente o parâmetro *min_samples_split* para evitar tanto overfitting como underfitting. Valores intermédios, por exemplo entre 30 e 50, mostram-se mais adequados para maximizar a capacidade de generalização e precisão do modelo.

4)

Árvore de Decisão - Diagnóstico de Diabetes



Regras da Árvore de Decisão:

```
|--- Glucose <= 127.50
| |--- Age <= 28.50
| | |--- BMI <= 45.40
| | | |--- class: 0
| | |--- BMI > 45.40
| | | |--- class: 1
| |--- Age > 28.50
| | |--- BMI <= 26.35
| | | |--- class: 0
| | |--- BMI > 26.35
| | | |--- class: 0
|--- Glucose > 127.50
| |--- BMI <= 29.95
| | |--- Glucose <= 145.50
| | | |--- class: 0
| | |--- Glucose > 145.50
| | | |--- class: 1
| |--- BMI > 29.95
| | |--- Glucose <= 157.50
| | | |--- class: 1
| | |--- Glucose > 157.50
| | | |--- class: 1
```

Nota sobre regras:

- **Classe 0** corresponde a **Normal**;
- **Classe 1** corresponde a **Diabetes**.

Probabilidades Posteriores para cada Folha:

- [Classe 0 – Normal; Classe 1 – Diabetes]
- Folha 3: [0.925; 0.075]
- Folha 4: [0.250; 0.750]
- Folha 6: [0.951; 0.049]
- Folha 7: [0.601; 0.399]
- Folha 10: [0.854; 0.146]
- Folha 11: [0.486; 0.514]
- Folha 13: [0.391; 0.609]
- Folha 14: [0.130; 0.870]

Conclusão Final:

A árvore de decisão divide as observações inicialmente pela variável *Glucose*, que é um dos fatores mais relevantes para o diagnóstico de diabetes, já que níveis mais altos de glicose estão associados a uma maior probabilidade da condição.

As probabilidades posteriores foram calculadas para cada folha da árvore, indicando a proporção de amostras diagnosticadas como "Normal" ou "Diabetes" em cada uma delas.

Folhas com níveis elevados de *Glucose* e *BMI* apresentam maior probabilidade para "Diabetes", enquanto folhas com valores mais baixos destas variáveis estão associadas a um diagnóstico de "Normal".

END