

Árvores de Decisão

Eduardo Coppetti Radaelli



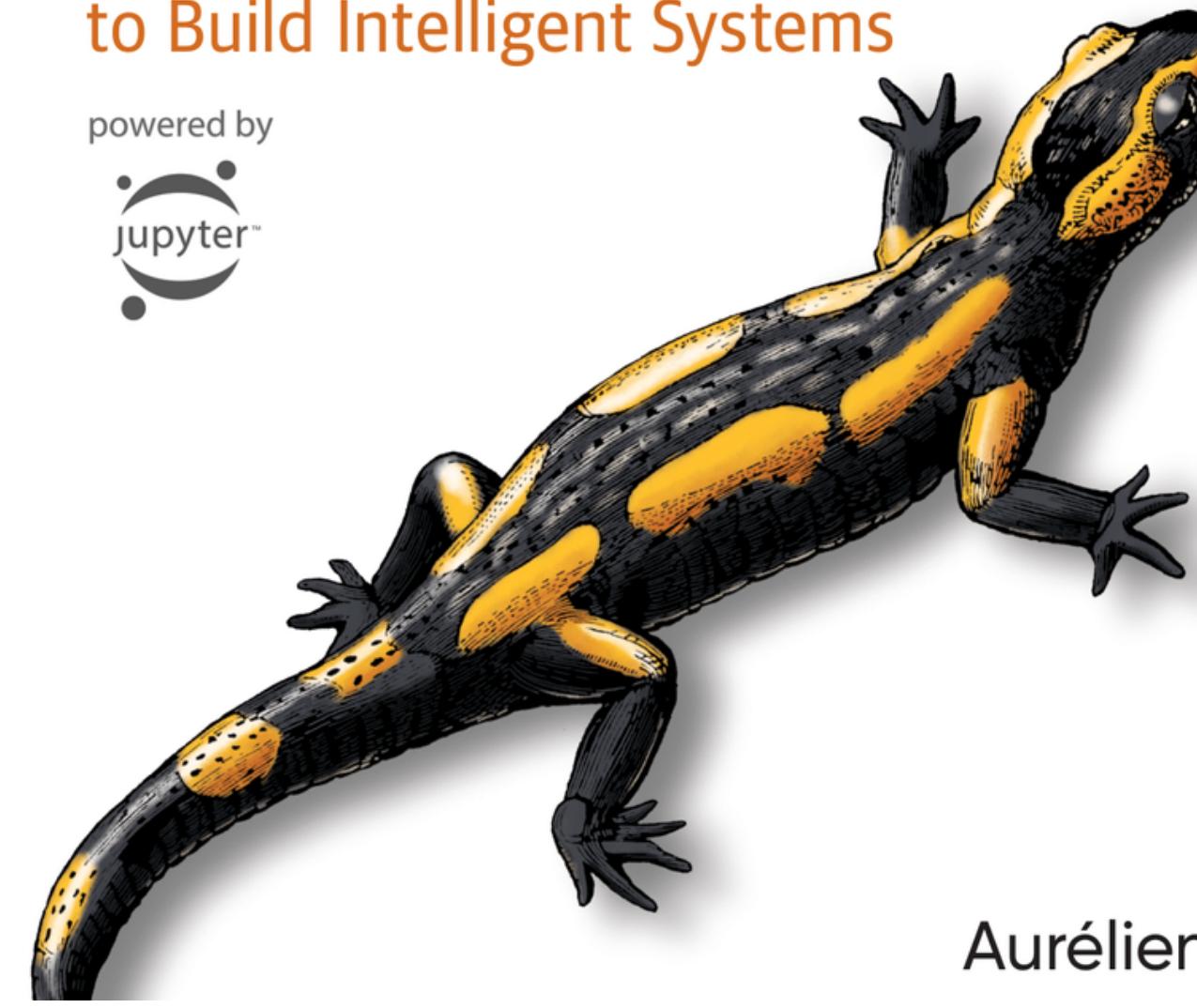
O'REILLY®

Third
Edition

Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow

Concepts, Tools, and Techniques
to Build Intelligent Systems

powered by



Aurélien Géron

Capítulo 6

Sumário

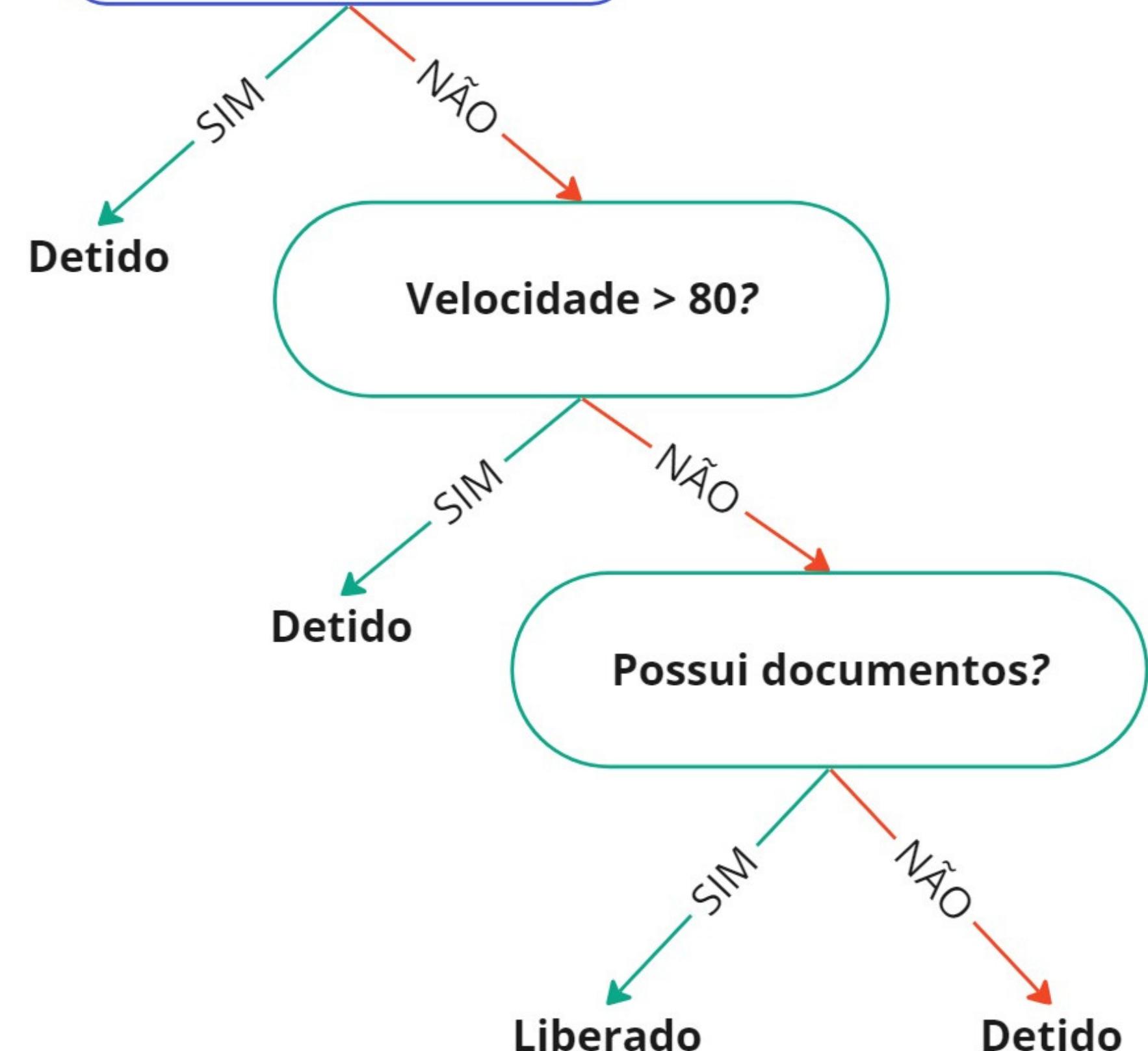
- Definição
- Objetivos
- Estrutura
 - Critérios de Divisão
- Exemplos
 - Classificação
 - Regressão
- Vantagens e Desvantagens
- Algoritmos Principais
- Técnicas de Ensemble

Definição

- Técnica de aprendizado de máquina utilizada para resolver problemas de classificação e regressão.
- Representa uma série de decisões em uma estrutura hierárquica de árvore.

O motorista será detido?

Está álcoolizado?



Objetivos

- Identificar padrões nos dados e depois utilizá-los para fazer previsões ou classificar novos dados em categorias.
- Dividir um problema complexo em perguntas simples para facilitar a tomada de decisões.
- Gerar modelos que possam ser facilmente compreendidos e interpretados por humanos.

Objetivos

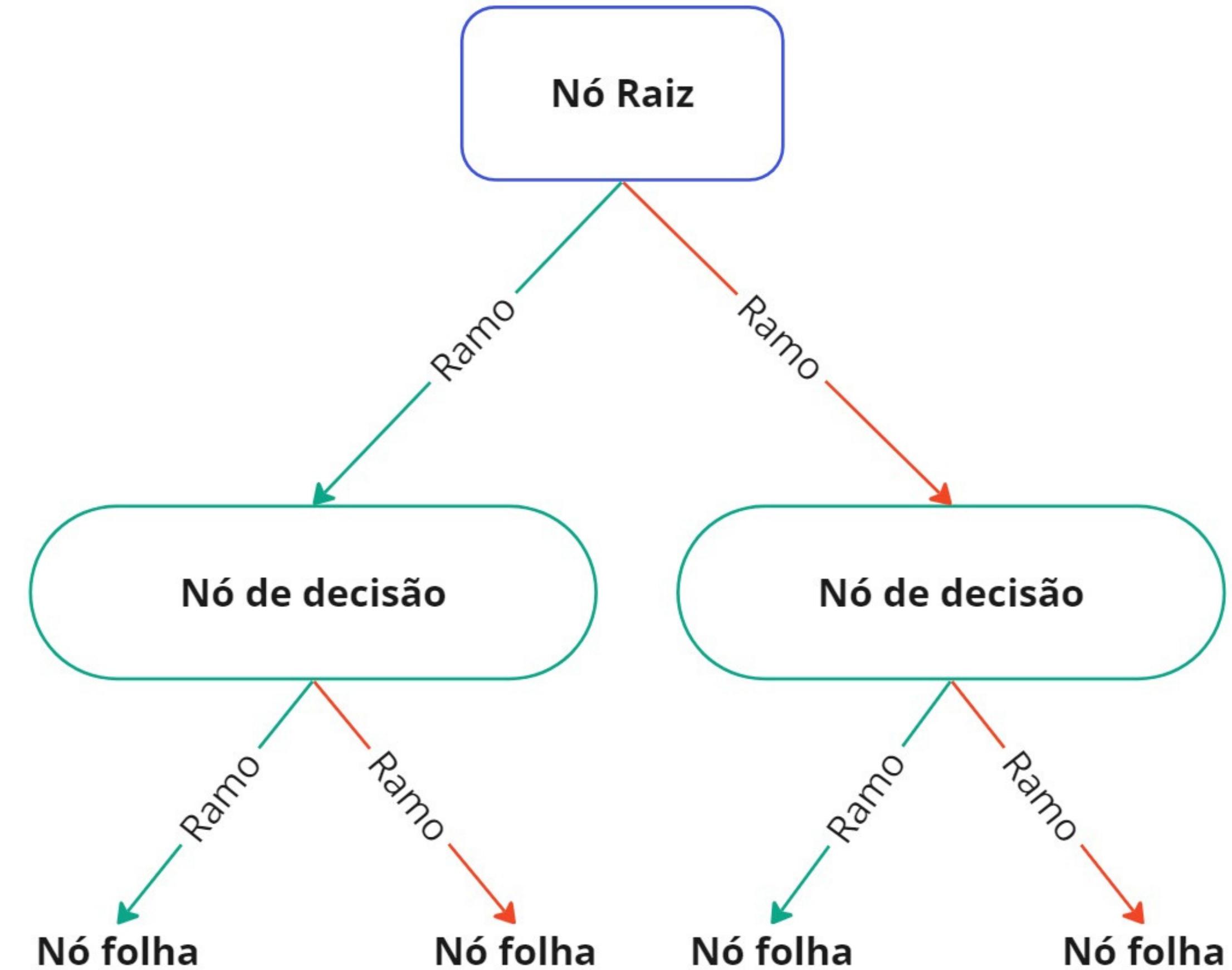
- Identificar quais características (features) são mais relevantes para o problema em questão.
- Tratar adequadamente classes desbalanceadas envolve atribuir pesos maiores às classes menos frequentes, para que sejam tratadas de forma igual à classe mais frequente.

Estrutura

- Nó raiz: Representa a primeira pergunta ou atributo que divide os dados.
- Nós de decisão: Locais onde ocorrem as divisões com base nos atributos dos dados.
- Nós folha: Representam as classes ou valores finais previstos para os dados.
- Ramos: Conectam os nós na árvore. Geralmente binários, dividindo-se em dois caminhos distintos em cada nó de decisão.

Estrutura

- Nó Raiz: Atributo principal (mais relevante em relação ao rótulo)
- Nós Internos: Atributos
- Folhas: Classes (rótulos)
- Ramos: Valores (atributos categóricos) ou intervalos (atributos numéricos)



Estrutura

- Profundidade da Árvore: Comprimento máximo do caminho da raiz até a folha mais distante.

Maior profundidade -> Relações mais complexas nos dados ->
Maior probabilidade de overfitting

- Critério de Divisão: Utilizado para dividir os dados em cada nó da árvore. Para **classificação**, comumente usa-se o **índice de Gini** ou a **entropia**. Para **regressão**, empregam-se critérios como **erro quadrático médio** e **erro absoluto médio**.

Estrutura

- Podas: Remoção de subárvores que não contribuem para a precisão da predição, prevenindo o overfitting e tornando a árvore mais generalizável.
- Ramificação Multiclasse: Capacidade de lidar com mais de duas classes, permitindo decisões com mais de dois resultados.
- Tratamento de Valores Ausentes ou Faltantes: Técnicas para lidar com dados faltantes, como ignorar exemplos, atribuir valores ou tratá-los como uma categoria separada.

Etapas para a Construção da Árvore de Decisão

1. Seleção da Raiz: Calcula-se a impureza para cada variável e escolhe-se aquela que resulta na menor impureza ou no maior ganho de informação como a raiz da árvore.
2. Divisão dos Nós: Os dados são divididos com base nos valores da variável raiz, criando subconjuntos representando diferentes ramificações.
3. Cálculo da Impureza dos Filhos: Para cada nó filho, calcula-se a impureza de Gini ou entropia.
4. Expansão dos Nós: Repetem-se os passos anteriores para cada nó filho.

Condição de Parada

- A divisão continua até que um critério de parada seja alcançado, como a **profundidade máxima da árvore**, o **número mínimo de amostras em um nó** ou um **limite de impureza**.
- Esses critérios são definidos previamente pelo desenvolvedor e são cruciais para controlar o crescimento da árvore e evitar o overfitting.
- Sem critérios de parada definidos, a árvore continuará crescendo até que todos os nós sejam puros ou até atingir sua profundidade máxima. Esses fatores podem causar overfitting e aumentar a complexidade da árvore.

Critérios de Divisão

O critério de divisão é uma medida usada para decidir como dividir os dados em cada nó da árvore de decisão durante o treinamento.

Uma escolha correta no critério de divisão pode impactar em:

- Melhora na capacidade de classificar corretamente novos exemplos.
- Redução do overfitting.
- Redução do tempo de treinamento de árvores em grandes conjuntos de dados.
- Melhor adaptação a conjuntos de dados desbalanceados.

| Métrica | Índice de Gini | Entropia |
|---------------|---|---|
| Definição | Métrica estatística que mede a desordem nos dados. | Medida estatística que mede a incerteza nos dados. |
| Origem | Desenvolvida para árvores de decisão e amplamente usada. | Teoria da informação de Shannon e termodinâmica. |
| Intervalo | Varia de 0 a 1. Um valor de 0 indica máxima pureza. | Varia de 0 a 1. Um valor de 0 indica máxima ordem. |
| Interpretação | Quanto maior, maior a impureza ou desordem nos dados. | Quanto maior, maior a incerteza ou desordem nos dados. |

Se todas as amostras forem da mesma classe,

Gini = 0 e Entropia = 0 (desordem mínima).

Se as amostras estiverem divididas igualmente entre as classes,

Gini = 1 e Entropia = 1 (desordem máxima).

Índice de Gini

$$G_i = 1 - \sum_{k=1}^n p_{i,k}^2$$

Entropia

$$H_i = - \sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^n p_{i,k} \log_2(p_{i,k})$$

Critérios para Classificação

| Métrica | Índice de Gini | Entropia |
|-----------------|---|--|
| Cálculo | Proporções de cada classe elevadas ao quadrado. | Distribuição de probabilidade direta das classes. |
| Impacto | Tende a isolar a classe mais frequente em seu próprio ramo. | Tende a produzir árvores mais balanceadas. |
| Velocidade | Mais rápido de calcular. | Um pouco mais lento devido ao cálculo. |
| Uso recomendado | Classificar corretamente as amostras (precisão). | Criar árvores mais平衡adas em termos de distribuição de classes. |

Critérios para Regressão

| Métrica | Erro Quadrático Médio (MSE) | Erro Absoluto Médio (MAE) |
|---------|---|---|
| Cálculo | Média dos quadrados dos erros entre as previsões e os valores reais. | Média dos valores absolutos dos erros entre as previsões e os valores reais. |
| Impacto | Sensível a grandes erros, útil quando deseja penalizar fortemente outliers. | Menos sensível a outliers, fornecendo uma visão mais equilibrada do desempenho. |

Exemplo de Classificação

| Dor Muscular | Grupo de Risco | Plaquetas | Temperatura | Dengue |
|--------------|----------------|-----------|-------------|--------|
| Sim | Sim | 150000 | 38.7 | Sim |
| Sim | Não | 90000 | 38.5 | Sim |
| Não | Sim | 180000 | 38.2 | Não |
| Não | Não | 115000 | 37.1 | Não |
| Não | Não | 175000 | 37.8 | Sim |
| Não | Não | 130000 | 38.3 | ??? |

Pergunta: Com base em todos esses atributos, qual será o diagnóstico do novo paciente?

Novo paciente

Cálculo para Dor Muscular

| Dor Muscular | Dengue |
|--------------|--------|
| Sim | Sim |
| Sim | Sim |
| Não | Não |
| Não | Não |
| Não | Sim |

| Variável | Dor Muscular = Sim (2/5) | Dor Muscular = Não (3/5) |
|--------------|--|---|
| Rótulo | $\text{Sim} = 2/2 = 1$ $\text{Não} = 0/2 = 0$ | $\text{Sim} = 1/3$ $\text{Não} = 2/3$ |
| Gini | $1 - ((1)^2 + (0))^2 = 0$ | $1 - ((1/3)^2 + (2/3))^2 =$ $4/9 = 0.44$ |
| Entropia (H) | $-(1 * \log_2(1) + 0 * \log_2(0)) = 0$ | $-(1/3 * \log_2(1/3) +$ $2/3 * \log_2(2/3)) = -$ $(-0.918) = 0.918$ |

Análise de resultados

A escolha do critério vai depender do objetivo principal:

- Gini: Se o foco está na simplicidade da árvore de decisão e na facilidade de interpretação. Além de separar os dados de forma que cada grupo tenha predominantemente uma única categoria.
- Entropia: Se a maior preocupação está com a sensibilidade aos dados desbalanceados e busca uma maior generalização do modelo, a entropia pode ser mais adequada.

Quanto menor o critério, menor a desordem dos dados.

Gini para variáveis contínuas

Para variáveis contínuas, como plaquetas (inteira) e temperatura (float), deve-se:

| Plaquetas | Dengue |
|-----------|--------|
| 150000 | Sim |
| 90000 | Sim |
| 180000 | Não |
| 115000 | Não |
| 175000 | Sim |

1. Ordenar os valores
2. Considerar cada valor como um ponto de divisão. Ex:
 - 1º Ponto: (Entre 90000 e 115000) = Plaquetas < 115000?)
3. Calcular o Gini para cada ponto de divisão. Ex:
 - (Plaquetas < 115000? 1º Gini)
 - (Plaquetas < 150000? 2º Gini)
 - (Plaquetas < 175000? 3º Gini)
 - (Plaquetas < 180000? 4º Gini)

Cálculo para Plaquetas

| Plaquetas | Dengue |
|-----------|--------|
| 90000 | Sim |
| 115000 | Não |
| 150000 | Sim |
| 175000 | Sim |
| 180000 | Não |

Plaquetas < 115000?

| Sim (1/5) | Não (4/5) |
|---------------|--|
| Sim = 1/1 | Sim = 2/4 = 1/2 |
| Não = 0/1 = 0 | Não = 2/4 = 1/2 |
| Gini = 0 | Gini = $1 - ((1/2)^2 + (1/2)^2) = 0.5$ |

$$\begin{aligned} \text{Gini (Plaquetas < 115000)} &= \\ &= (1/5 * 0) + (4/5 * 0.5) = 0.4 \end{aligned}$$

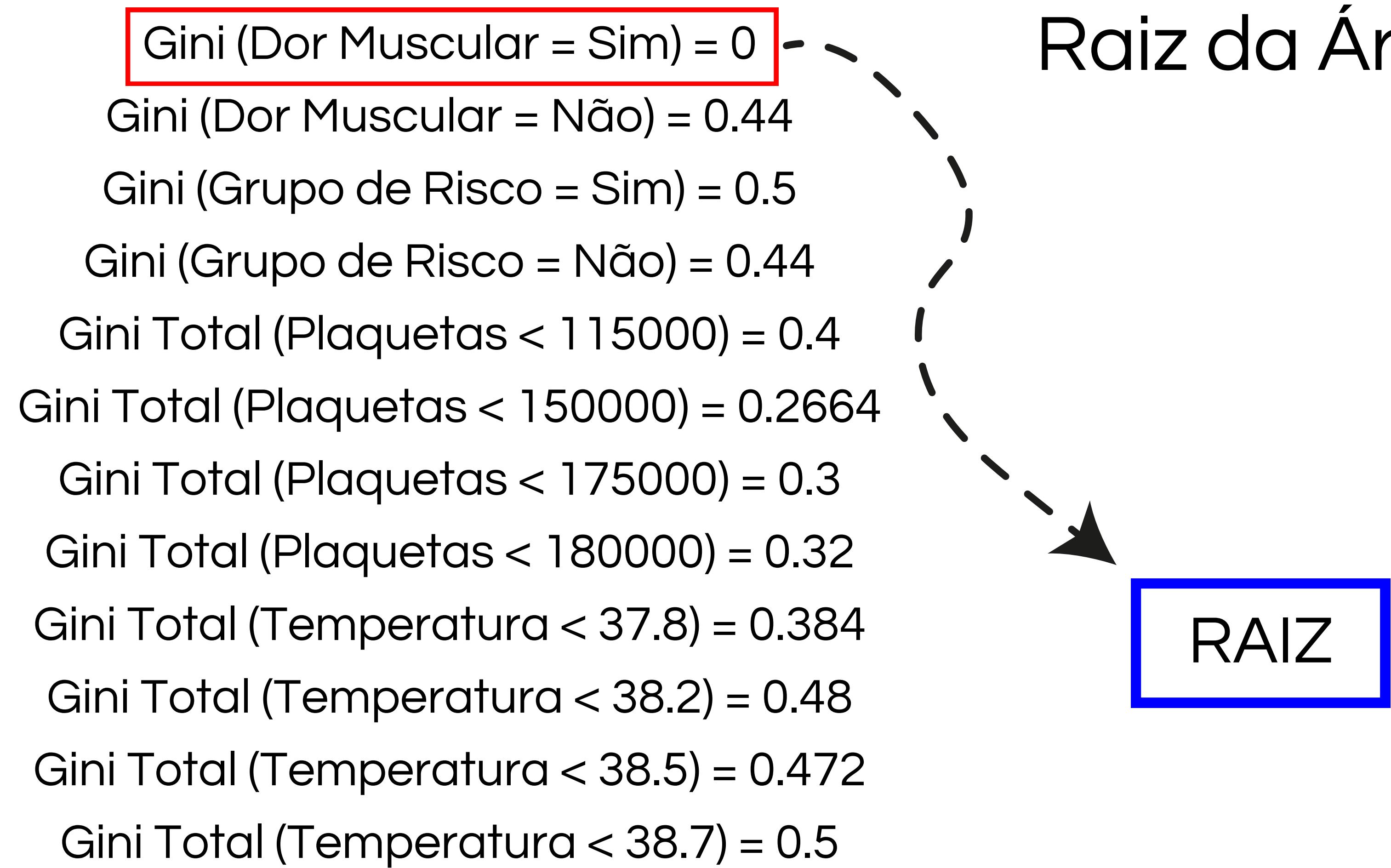
Cálculo para Plaquetas

Gini (Plaquetas < 115000) = 0.4

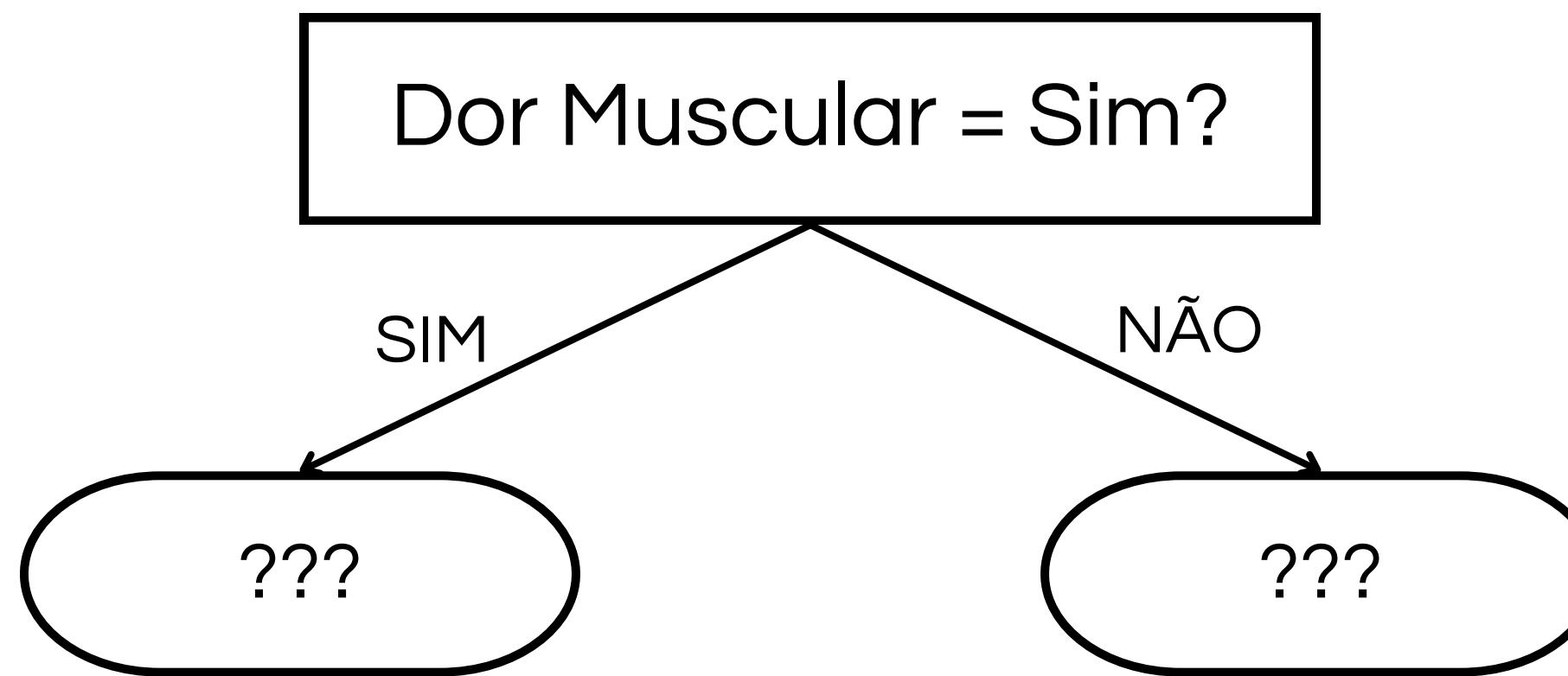
Gini (Plaquetas < 150000) = 0.2664

Gini (Plaquetas < 175000) = 0.3

Gini (Plaquetas < 180000) = 0.32



Árvore de Decisão



| Grupo de Risco | Plaquetas | Temp | Dengue |
|----------------|-----------|------|--------|
| Sim | 150000 | 38.7 | Sim |
| Não | 90000 | 38.5 | Sim |

| Grupo de Risco | Plaquetas | Temp | Dengue |
|----------------|-----------|------|--------|
| Sim | 180000 | 38.2 | Não |
| Não | 115000 | 37.1 | Não |
| Não | 175000 | 37.8 | Sim |

Árvore de Decisão

Gini (Grupo de Risco = Sim) = 0

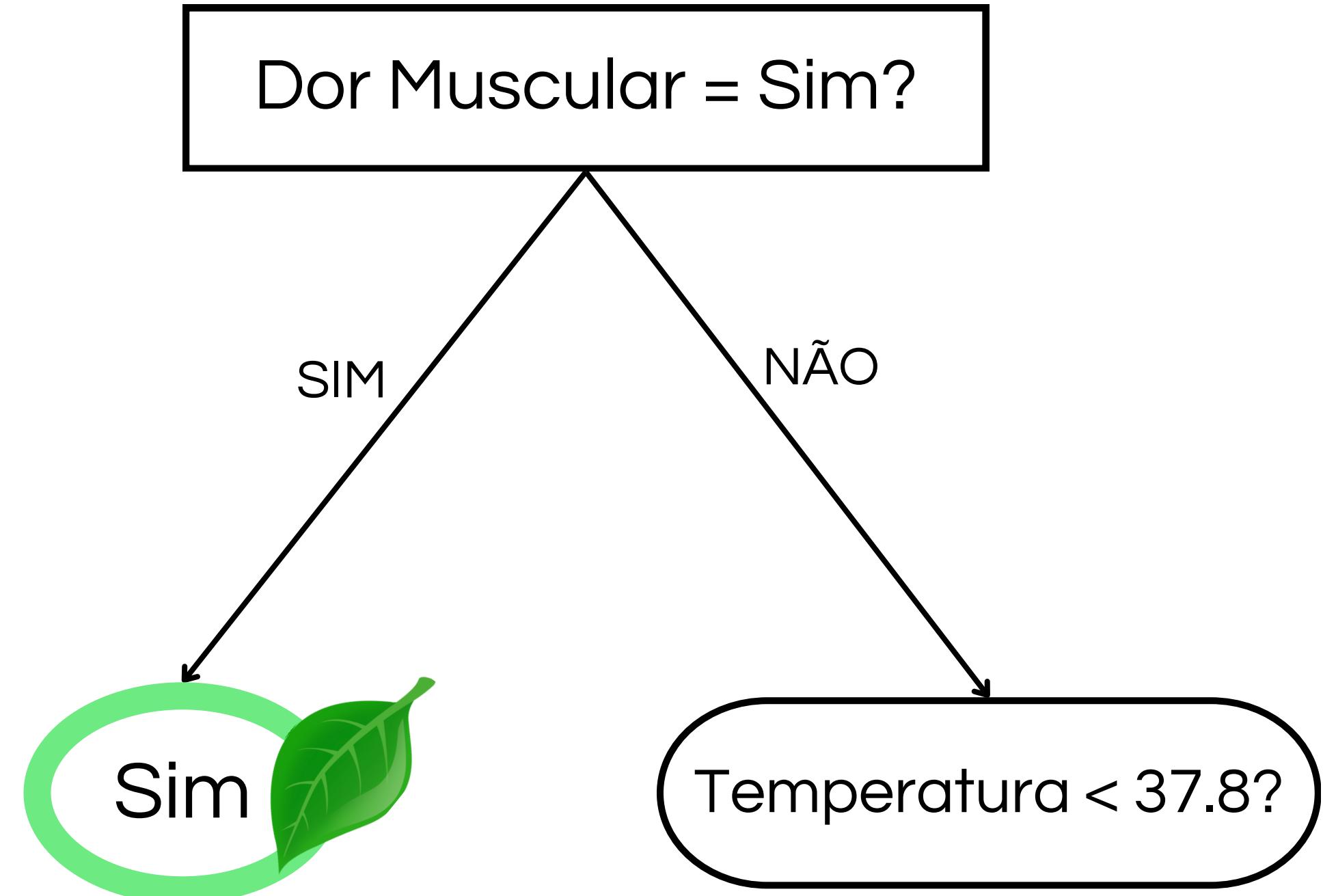
Gini (Grupo de Risco = Não) = 0.5

Gini (Plaquetas < 175000) = 0

Gini (Plaquetas < 180000) = 0.5

Gini (Temperatura < 37.8) = 0

Gini (Temperatura < 38.2) = 0.44

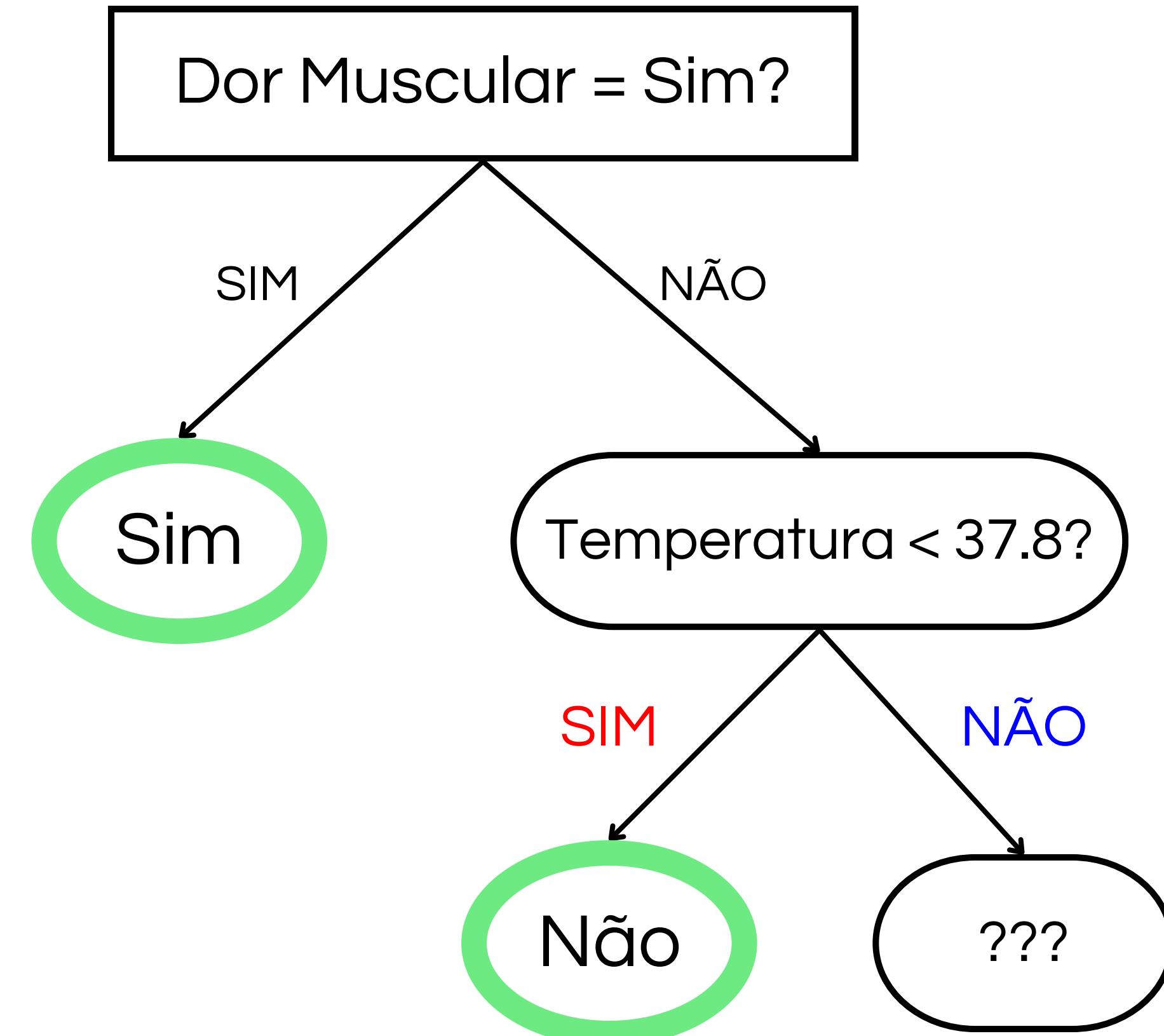


OBS: Quando tem mais de uma variável com o mesmo Gini, a escolha é do desenvolvedor.

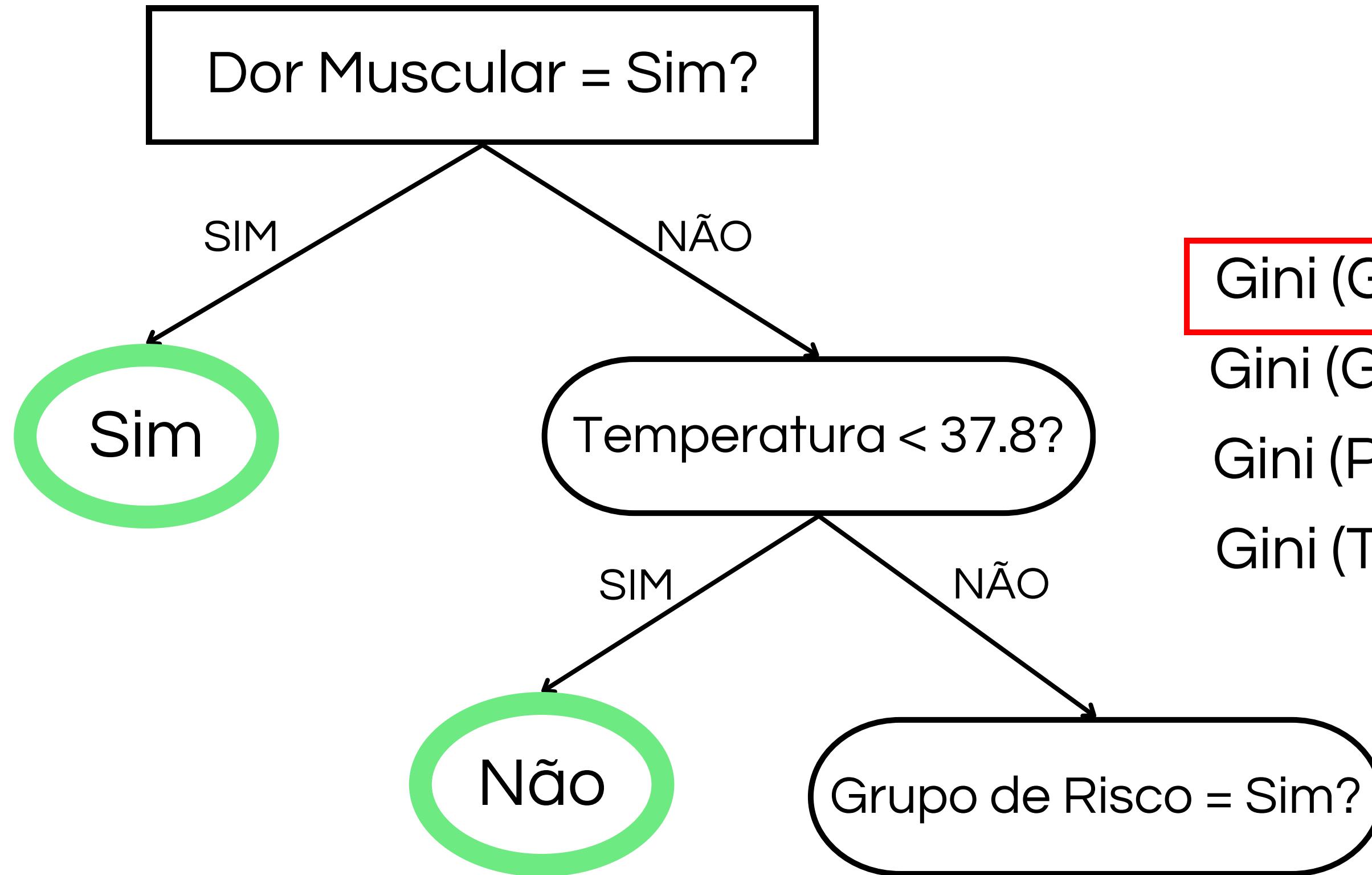
Árvore de Decisão

| Grupo de Risco | Plaquetas | Temp | Dengue |
|----------------|-----------|------|--------|
| Não | 115000 | 37.1 | Não |

| Grupo de Risco | Plaquetas | Temp | Dengue |
|----------------|-----------|------|--------|
| Sim | 180000 | 38.2 | Não |
| Não | 175000 | 37.8 | Sim |



Árvore de Decisão



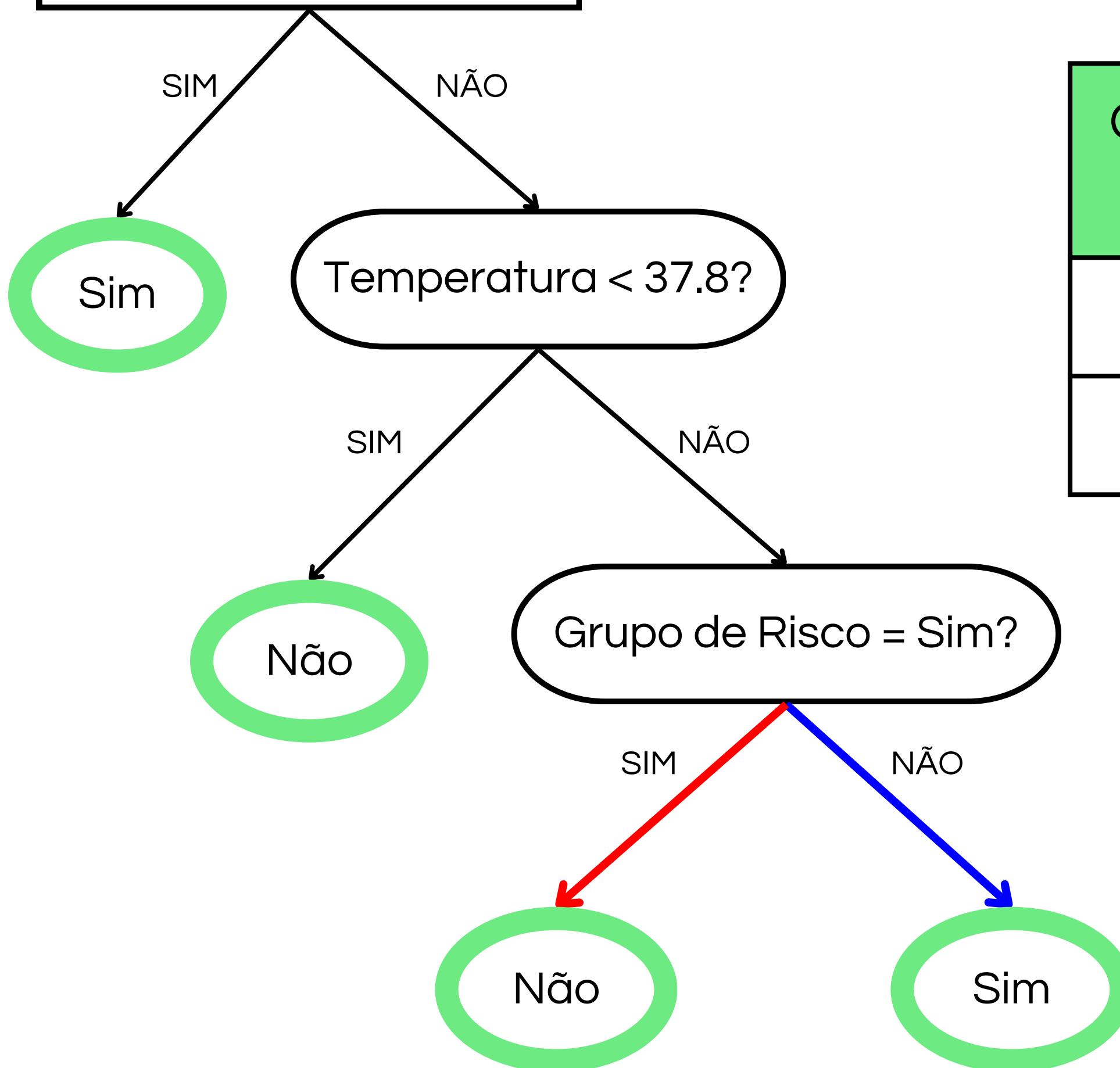
Gini (Grupo de Risco = Sim) = 0

Gini (Grupo de Risco = Não) = 0

Gini (Plaquetas < 180000) = 0.5

Gini (Temperatura < 38.2) = 0.5

Árvore de Decisão Final

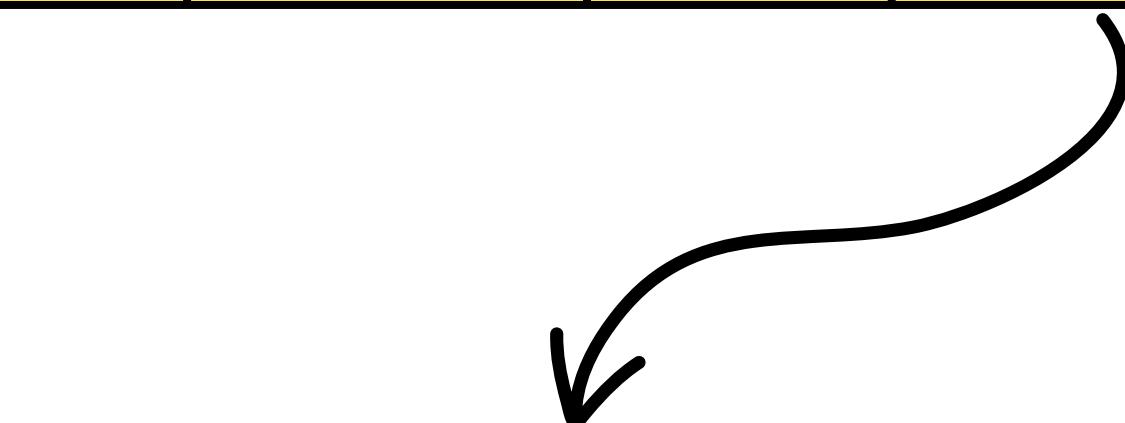
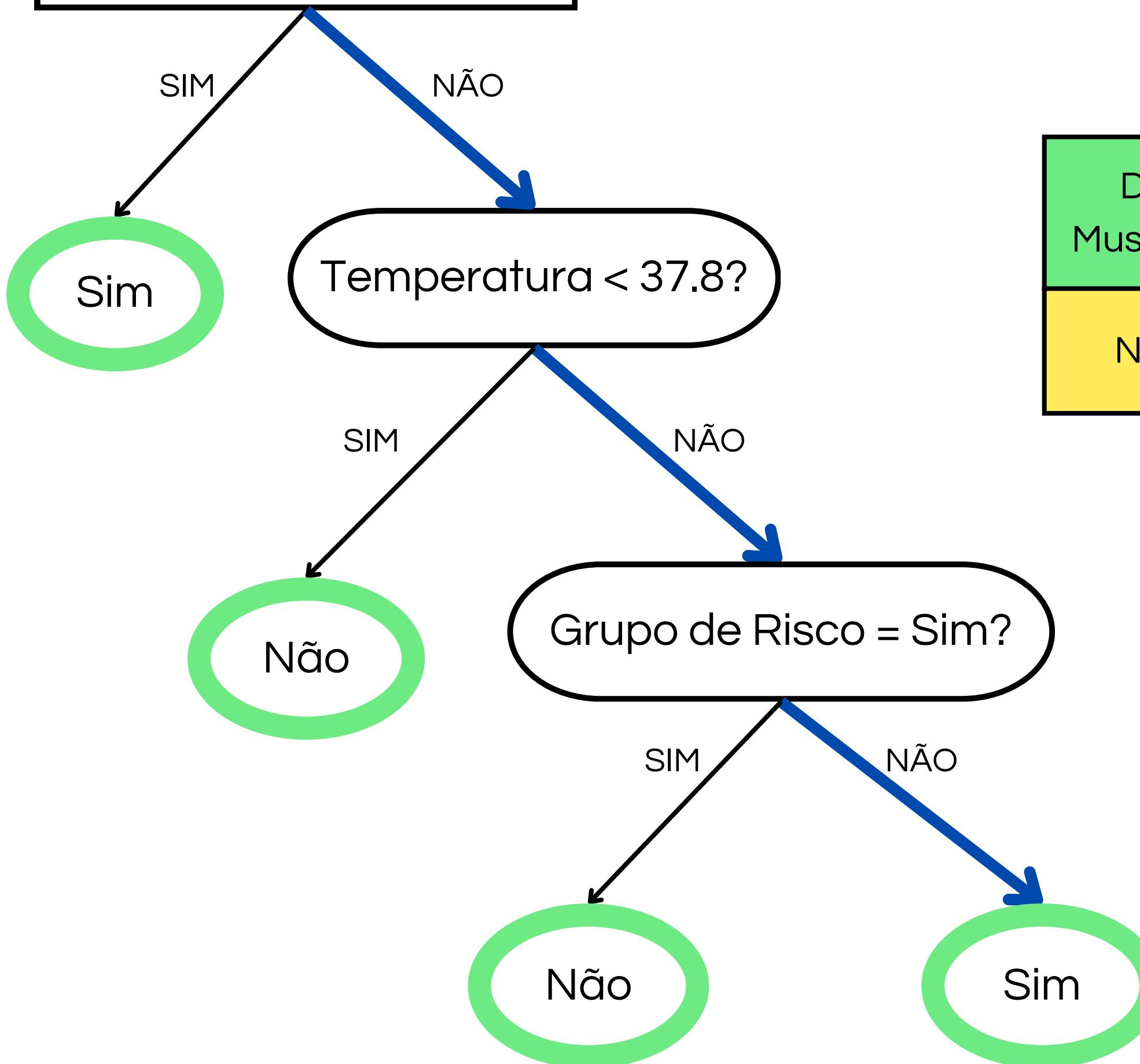


| Grupo de Risco | Plaquetas | Dengue |
|----------------|-----------|--------|
| Sim | 180000 | Não |
| Não | 175000 | Sim |

Árvore de Decisão Final

Novo Paciente:

| Dor Muscular | Grupo de Risco | Plaquetas | Temp | Dengue |
|--------------|----------------|-----------|------|--------|
| Não | Não | 130000 | 38.3 | ??? |



Dengue = Sim

Interpretação da Árvore de Decisão

- Dor Muscular é um indicador significativo de possível infecção por dengue.
- Temperatura baixa (< 37.8) pode indicar que o paciente não está com dengue.
- (GRUPO DE RISCO = SIM) -> (DENGUE = NÃO) ???
- A ausência da variável Plaquetas como critério na árvore indica que essa variável pode não ter sido tão relevante.
- OBS: Essas conclusões baseiam-se neste conjunto de dados específico e podem não ser universalmente aplicáveis.

Exemplo de Regressão

| Tamanho (m ²) | Número de Quartos | Área Urbana | Reformada | Preço (R\$) |
|---------------------------|-------------------|-------------|-----------|-------------|
| 150 | 3 | Sim | Não | 300000 |
| 200 | 4 | Não | Sim | 400000 |
| 120 | 2 | Sim | Não | 250000 |
| 180 | 3 | Não | Sim | 350000 |
| 220 | 4 | Sim | Sim | 450000 |
| 165 | 3 | Não | Sim | ??? |

Pergunta: Com base nas características fornecidas, qual é o preço estimado da casa?

Cálculo para as Variáveis

Para calcular o erro de previsão usando MSE e MAE em cada variável, seguem as abordagens:

- Para variáveis inteiras e decimais (float):
 - Ordene os valores da variável.
 - Considere cada valor como um ponto de divisão.
 - Calcule o erro de previsão para cada ponto de divisão usando MSE ou MAE.
- Para variáveis booleanas:
 - Calcule o erro de previsão para os dois valores booleanos possíveis (1 e 0).

Cálculo para Tamanho

| Tamanho (m ²) | Preço (R\$) |
|---------------------------|-------------|
| 120 | 250000 |
| 150 | 300000 |
| 180 | 350000 |
| 200 | 400000 |
| 220 | 450000 |

Escolher entre MSE e MAE e a calcular o erro para cada ponto de divisão especificado:

- Tamanho <= 150?
- Tamanho <= 180?
- Tamanho <= 200?
- Tamanho <= 220?

MSE

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Tamanho <= 150

MAE

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

$$\hat{y}_{\text{média}} = (250000 + 300000) / 2 \Rightarrow 275000$$

| Tamanho (m ²) | Preço (R\$) |
|---------------------------|-------------|
| 120 | 250000 |
| 150 | 300000 |

$$\hat{y}_{\text{média}} = (250000 + 300000) / 2 \Rightarrow 275000$$

$$\begin{aligned} MSE &= 1/2 [(250000 - 275000)^2 + \\ &(300000 - 275000)^2] = 1/2 [(25000)^2 + \\ &(25000)^2] = 1250000000 / 2 \Rightarrow \\ &\mathbf{625000000} \end{aligned}$$

$$\begin{aligned} MAE &= 1/2 [|250000 - 275000| + \\ &|300000 - 275000|] = 1/2 [25000 + \\ &25000] = 50000 / 2 \Rightarrow \mathbf{25000} \end{aligned}$$

MSE

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Tamanho <= 180

MAE

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|$$

\hat{y} (média) = $(250000 + 300000 + 350000) / 3 \Rightarrow 300000$

| Tamanho (m ²) | Preço (R\$) |
|---------------------------|-------------|
| 120 | 250000 |
| 150 | 300000 |
| 180 | 350000 |

\hat{y} (média) = $(250000 + 300000 + 350000) / 3 \Rightarrow 300000$

$$\begin{aligned} \text{MSE} &= 1/3 [(250000 - 300000)^2 + (300000 - 300000)^2 + (350000 - 300000)^2] = 1/3 [(25000)^2 + 0 + (25000)^2] = 1666666666.67 \end{aligned}$$

$$\begin{aligned} \text{MAE} &= 1/3 [|250000 - 300000| + |300000 - 300000| + |350000 - 300000|] = 1/3 [25000 + 0 + 25000] = 33333.33 \end{aligned}$$

Cálculo para Tamanho

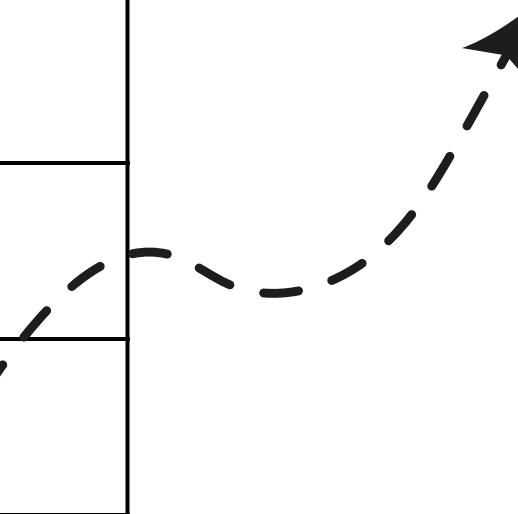
| Ponto de Divisão | MSE | MAE |
|------------------|----------------|----------|
| Tamanho <= 150 | 625000000 | 25000 |
| Tamanho <= 180 | 16666666666.67 | 33333.33 |
| Tamanho <= 200 | 6250000000 | 50000 |
| Tamanho <= 220 | 10000000000 | 60000 |

Quanto menor o erro, maior a precisão.

Raiz da Árvore

| Condição | MSE |
|-----------------------|----------------|
| Tamanho <= 150 | 625000000 |
| Tamanho <= 180 | 16666666666.67 |
| Tamanho <= 200 | 6250000000 |
| Tamanho <= 220 | 10000000000 |
| Número de Quartos < 3 | 3125000000 |
| Número de Quartos < 4 | 5000000000 |
| Área Urbana = Sim | 7962962.963 |
| Área Urbana = Não | 6250000 |
| Reformada = Sim | 1666666.67 |
| Reformada = Não | 2500000 |

RAIZ



Árvore de Decisão

| Tamanho (m ²) | Número de Quartos | Área Urbana | Preço (R\$) |
|---------------------------|-------------------|-------------|-------------|
| 200 | 4 | Não | 400000 |
| 180 | 3 | Não | 350000 |
| 220 | 4 | Sim | 450000 |

Reformada = Sim?

SIM

NÃO

| Tamanho (m ²) | Número de Quartos | Área Urbana | Preço (R\$) |
|---------------------------|-------------------|-------------|-------------|
| 150 | 3 | Sim | 300000 |
| 120 | 2 | Sim | 250000 |

???

???

Tamanho < 200: MSE = 12500000

Tamanho < 220: MSE = 33333333

Número de Quartos < 4: MSE = 12500000

Área Urbana = Sim: MSE = 0

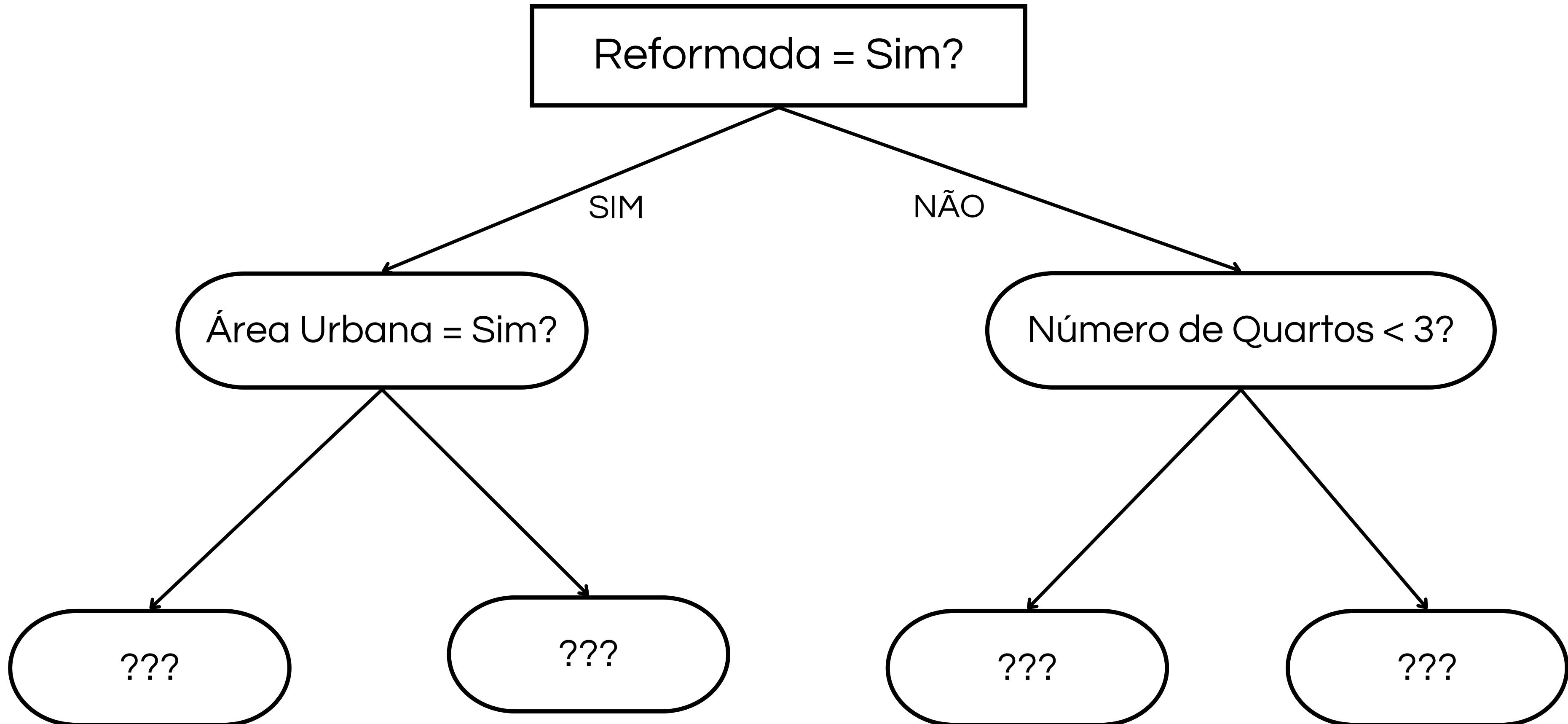
Área Urbana = Não: MSE = 12500000

Tamanho < 150: MSE = 31250000

Número de Quartos < 3: MSE = 0

Área Urbana = Sim: MSE = 31250000

Árvore de Decisão



Árvore de Decisão

| Tamanho (m ²) | Número de Quartos | Preço (R\$) |
|---------------------------|-------------------|-------------|
| 220 | 4 | 450000 |

Área Urbana = Sim?

SIM

NÃO

450000

| Tamanho (m ²) | Número de Quartos | Preço (R\$) |
|---------------------------|-------------------|-------------|
| 200 | 4 | 400000 |
| 180 | 3 | 350000 |

???

Tamanho \leq 200: MSE = 1250000

Número de Quartos < 4: MSE = 0

Árvore de Decisão

| Tamanho (m ²) | Número de Quartos | Área Urbana | Preço (R\$) |
|---------------------------|-------------------|-------------|-------------|
| 120 | 2 | Sim | 250000 |

Número de Quartos < 3?

| Tamanho (m ²) | Número de Quartos | Área Urbana | Preço (R\$) |
|---------------------------|-------------------|-------------|-------------|
| 150 | 3 | Sim | 300000 |

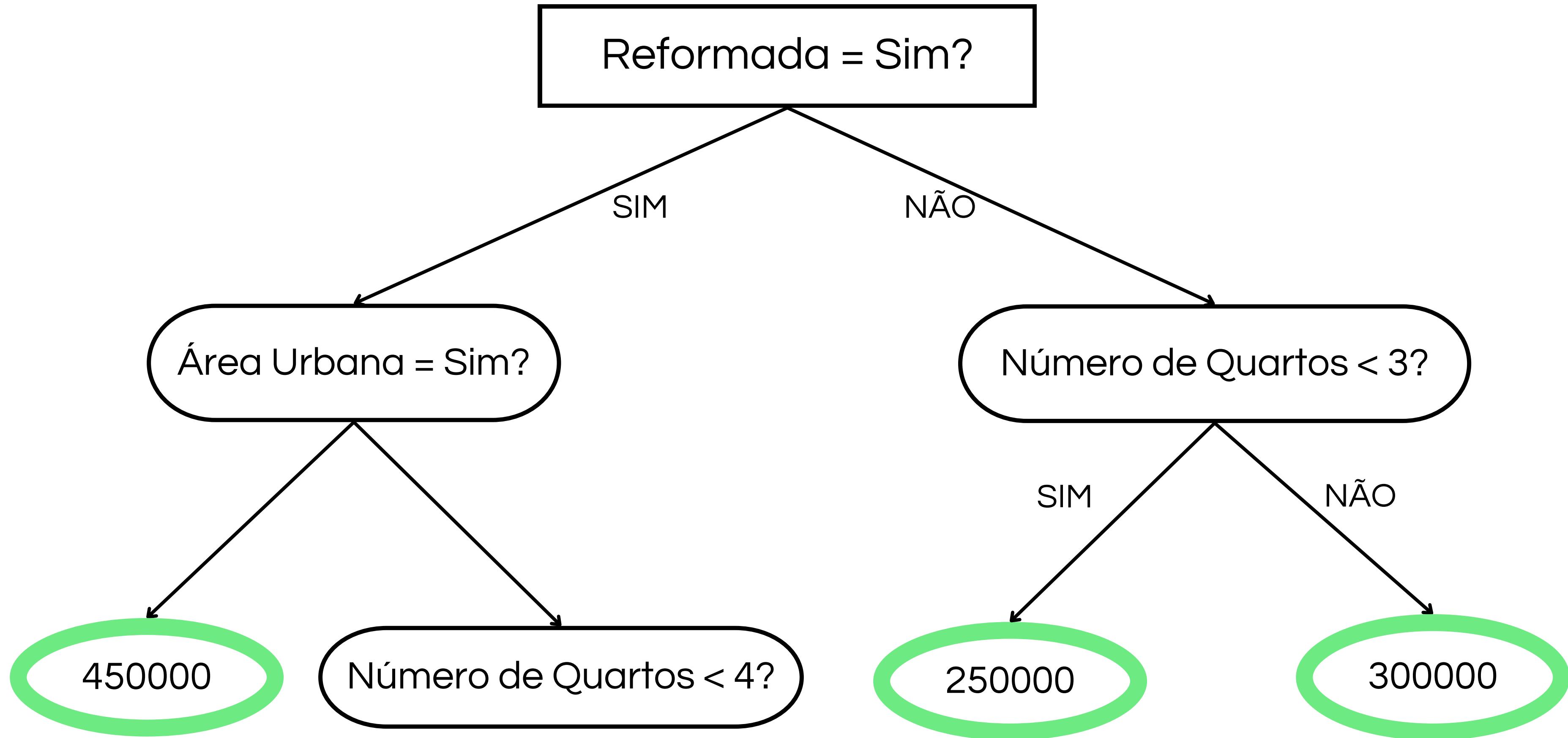
SIM

NÃO

250000

300000

Árvore de Decisão



Árvore de Decisão

| Tamanho (m ²) | Número de Quartos | Preço (R\$) |
|---------------------------|-------------------|-------------|
| 180 | 3 | 350000 |

Número de Quartos < 4?

| Tamanho (m ²) | Número de Quartos | Preço (R\$) |
|---------------------------|-------------------|-------------|
| 200 | 4 | 400000 |

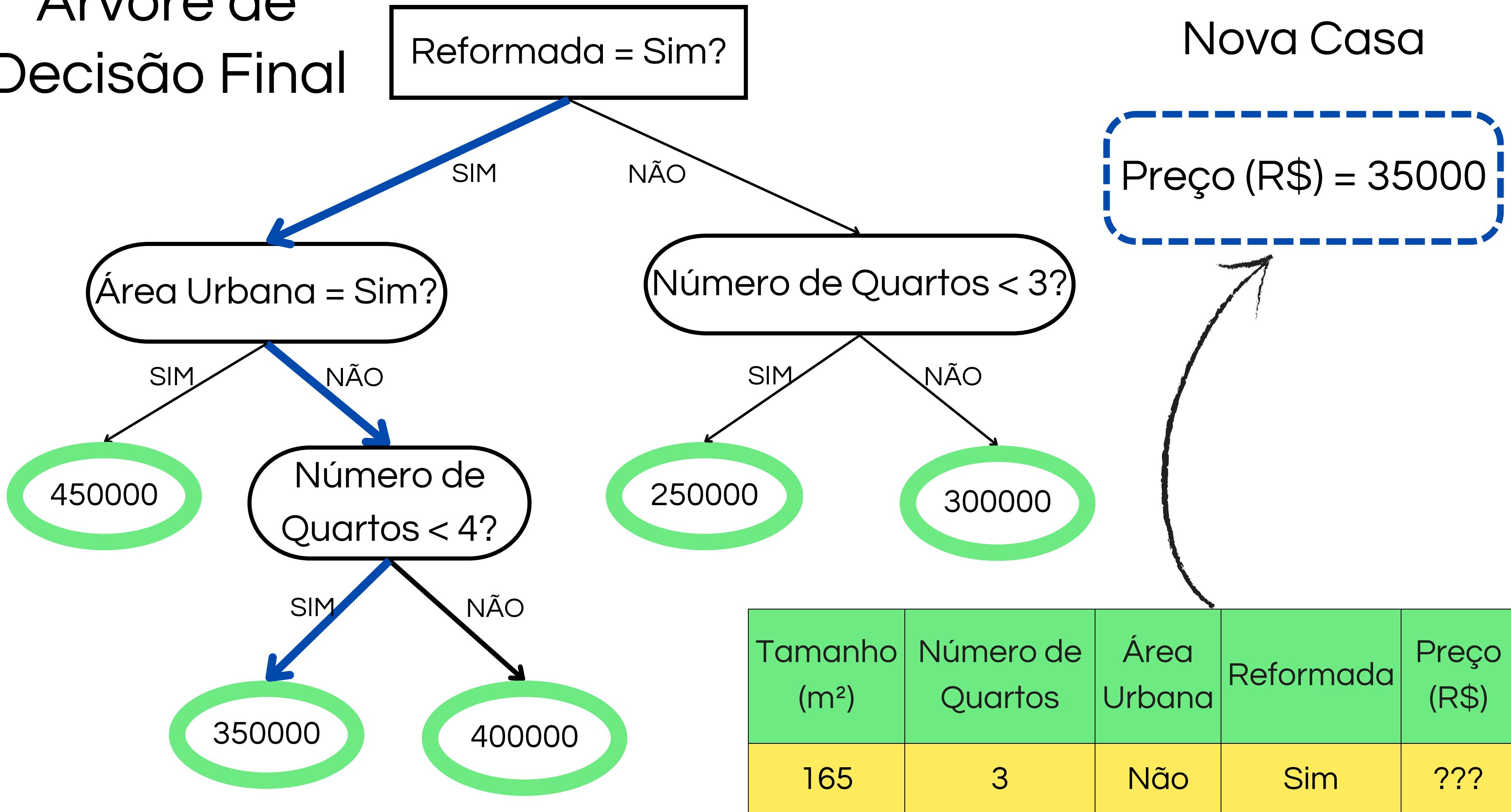
SIM

NÃO

350000

400000

Árvore de Decisão Final



Vantagens

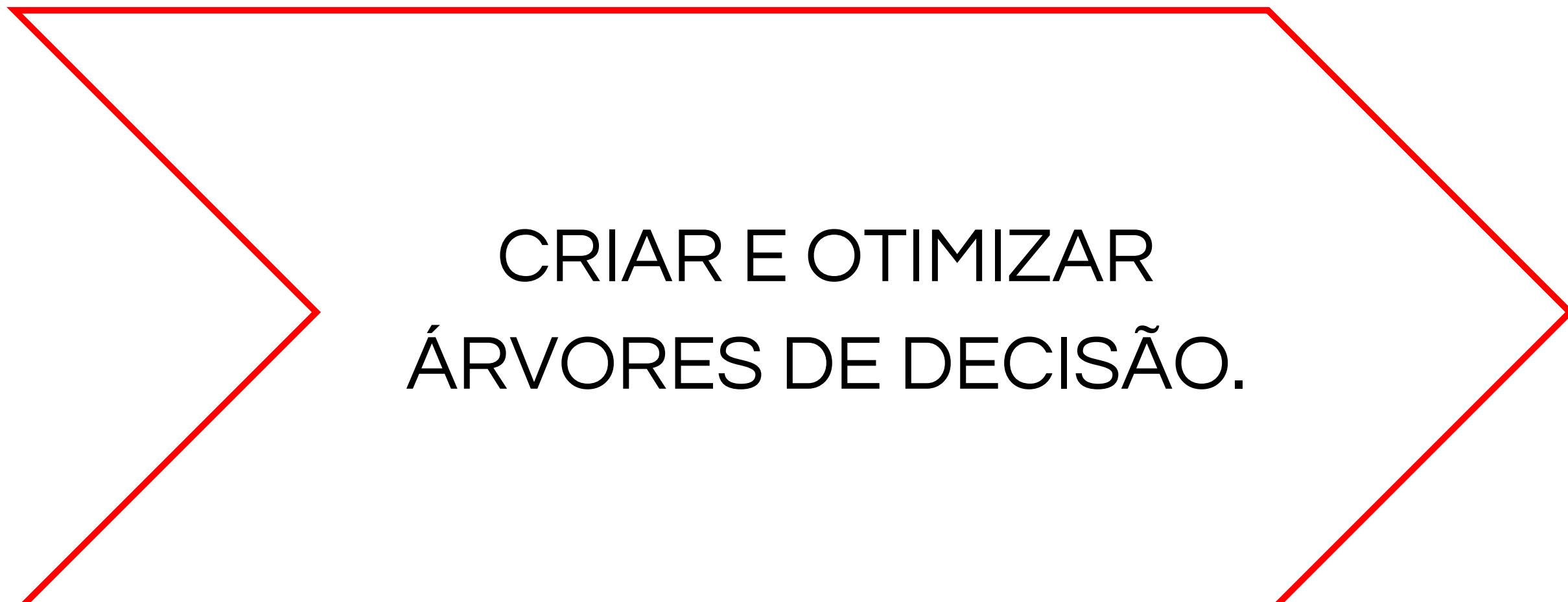
- Fáceis de entender e interpretar, mesmo para não especialistas.
- Lidam com diferentes tipos de dados sem necessidade de pré-processamento extenso.
- Podem capturar e modelar relações não lineares entre variáveis.
- Resistem a outliers e ruídos nos dados.
- Possuem grande versatilidade.

Desvantagens

- Tendência ao overfitting.
- Pequenas mudanças nos dados de treinamento podem resultar em árvores de decisão muito diferentes.
- Tendem a favorecer as classes majoritárias, especialmente em conjuntos de dados desbalanceados.
- Dificuldade em capturar relações muito complexas.
- Expansão exponencial.

Algoritmos Principais

- CART
- ID3
- C4.5



CRIAR E OTIMIZAR
ÁRVORES DE DECISÃO.

| Algoritmo | CART | ID3 | C4.5 |
|---------------------|--|--|---|
| Objetivo | Visa lidar eficazmente com atributos contínuos e valores ausentes, visando a construção de árvores generalizáveis. | Prioriza atributos que resultam em menor entropia, buscando simplificar as árvores e preservar sua interpretabilidade. | Procura equilibrar a seleção de atributos, resultando em árvores mais compactas e interpretáveis. |
| Critérios de parada | Profundidade máxima da árvore, número mínimo de amostras em um nó, impureza mínima. | Todas as amostras pertencem à mesma classe ou não há mais atributos para dividir. | Todas as amostras pertencem à mesma classe ou não há mais atributos para dividir. |

| Algoritmo | CART | ID3 | C4.5 |
|----------------------------|---------------------------|--|---------------------------|
| Atributos | Contínuos ou categóricos. | Categóricos. | Contínuos ou categóricos. |
| Quantidade de nós filhos | Binário. | Até o número de categorias do atributo dividido. | Ambos os anteriores. |
| Valores ausentes | Permitidos. | Não permitidos. | Permitidos. |
| Poda após a construção | Sim. | Não. | Sim. |
| Hiperparâmetros ajustáveis | Sim. | Não. | Sim. |

| Algoritmo | CART | ID3 | C4.5 |
|-------------|--|--|--|
| Codificação | Bibliotecas como scikit-learn em Python oferecem suporte direto ao algoritmo CART. | Implementação manual das principais funções, como cálculo de entropia, seleção de atributos e construção da árvore de decisão. | Implementação manual das principais funções, como cálculo de entropia, seleção de atributos e construção da árvore de decisão. |

Observação: Embora a implementação manual do ID3 e do C4.5 envolva a criação de funções principais, como cálculo de entropia e construção de árvores de decisão, bibliotecas prontas em Python, como scikit-learn, oferecem suporte a funcionalidades essenciais, simplificando o processo de implementação desses algoritmos.

Arvore Genérica X CART

- Na árvore de decisão genérica, a escolha da variável raiz é feita com base no menor índice de Gini entre os valores possíveis da variável, e esse processo é repetido para os nós subsequentes.
- No CART, a seleção da melhor divisão considera os índices de Gini dos subconjuntos resultantes, não apenas para os valores individuais das variáveis, utilizando uma média ponderada para levar em conta tanto a pureza quanto a quantidade de dados. Essa média é calculada pela função de custo do CART, descrita nas fórmulas abaixo, sendo a da esquerda para classificação e a da direita para regressão.

$$J(k, t_k) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}} \quad J(k, t_k) = \frac{m_{\text{left}}}{m} \text{MSE}_{\text{left}} + \frac{m_{\text{right}}}{m} \text{MSE}_{\text{right}}$$

ID3 X C4.5

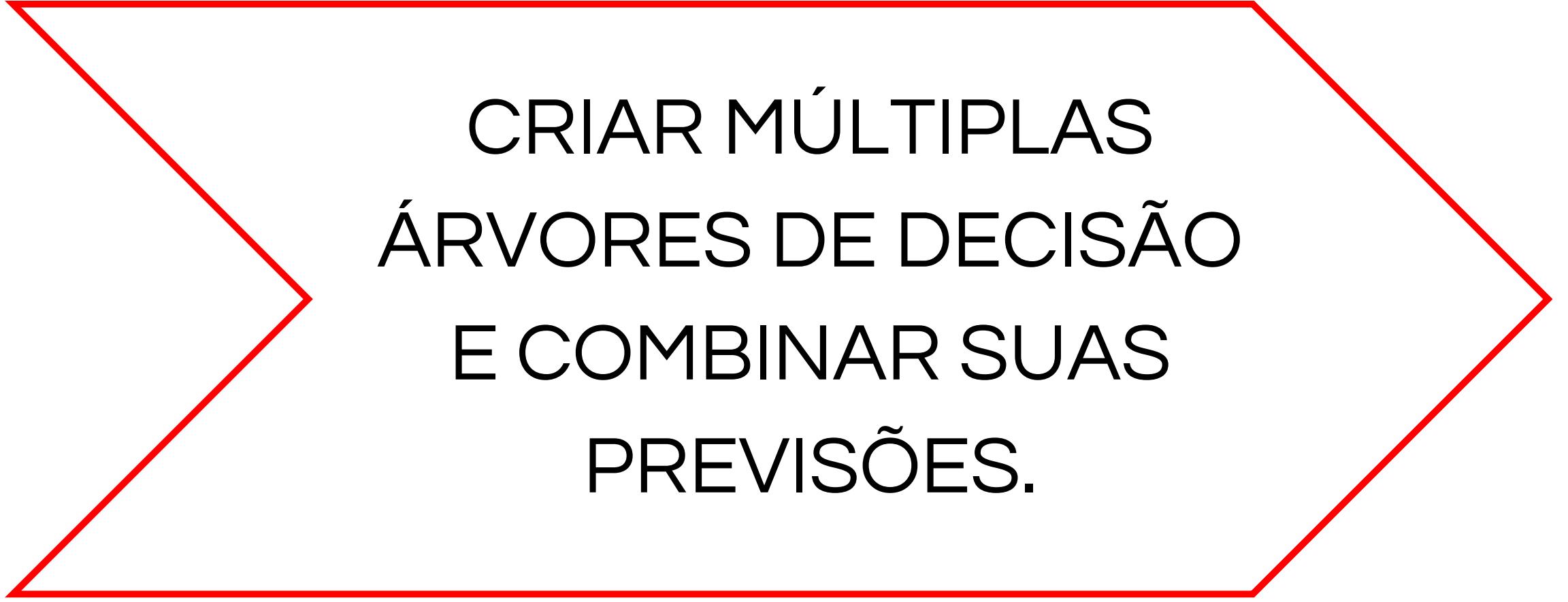
- Ambos os algoritmos ID3 e C4.5 utilizam o ganho de informação como critério principal para selecionar a variável raiz. No entanto, o C4.5 expande essa abordagem ao considerar também a proporção de ganho, o que o torna mais eficaz para atributos com diferentes números de valores possíveis.
- A proporção de ganho permite avaliar se o ganho de informação é alto devido à capacidade real do atributo de separar os dados em classes diferentes ou apenas porque o atributo tem muitas opções.
- Além disso, o C4.5 é capaz de lidar com atributos contínuos e valores ausentes, e realiza a poda da árvore para evitar o overfitting, tornando o C4.5 mais versátil e robusto em comparação com o ID3, permitindo uma construção mais eficiente e precisa da árvore de decisão.

C5.0

- Versão comercial (paga) e aprimorada do algoritmo C4.5, desenvolvida pela empresa RuleQuest Research.
- Oferece melhorias significativas em termos de desempenho, velocidade e precisão.
- Utiliza técnicas mais avançadas de otimização e ajuste de parâmetros, o que o torna mais eficiente na construção de árvores de decisão mais precisas e com menor probabilidade de overfitting.

Técnicas de Emsemble

- Random Forest
- Gradient Boosting Machines (GBM)
- XGBoost



CRIAR MÚLTIPAS
ÁRVORES DE DECISÃO
E COMBINAR SUAS
PREVISÕES.

| Algoritmo | Random Forest | Gradient Boosting | XGBoost |
|--------------------------|---|--|---|
| Objetivo | Tomada de decisão por meio da agregação de várias opiniões para reduzir overfitting e melhorar generalização. | Aprendizado com erros para melhorar progressivamente o desempenho do modelo. | Oferecer uma solução escalável e eficiente e lista de tarefas com regras (regularização) para evitar overfitting. |
| Metodologia | Combinação de árvores de decisão independentes. | Construção sequencial de árvores para corrigir erros. | Treinamento de árvores com regras para evitar overfitting. |
| Abordagem de Previsão | Votação ou média das previsões de árvores individuais. | Adição de árvores sequenciais para corrigir erros de previsão. | Controle de complexidade e overfitting por meio de regularização. |
| Sensibilidade a Outliers | Menos sensível devido à agregação de várias árvores. | Sensível, pois tenta corrigir erros de previsão. | Sensível, mas a regularização ajuda a mitigar impactos negativos. |

| Algoritmo | Random Forest | Gradient Boosting | XGBoost |
|---------------------------|---|--|---|
| Parâmetros | Número de árvores, profundidade máxima, número de features. | Taxa de aprendizado, número de iterações, profundidade máxima das árvores. | Taxa de aprendizado, número de iterações, profundidade máxima das árvores, parâmetros de regularização (L1, L2). |
| Velocidade de Treinamento | Rápido devido à independência das árvores. | Mais lento devido à construção sequencial das árvores. | Mais rápido que GBM devido à otimização e paralelismo. |
| Interpretabilidade | Menos interpretável devido à complexidade do modelo. | Menos interpretável comparado ao Random Forest, devido à adição sequencial de árvores. | Menos interpretável comparado ao Gradient Boosting, devido à adição sequencial de árvores e maior complexidade por meio da regularização. |

APLICAÇÕES

Random Forest:

- Identifica padrões anômalos nos dados que podem indicar atividade fraudulenta.
- Ajuda a identificar padrões em dados médicos para diagnosticar doenças com precisão.
- Usado para identificar as características mais relevantes.

Gradient Boosting:

- Utilizado para prever preços de propriedades com base em características.
- Classifica automaticamente sentimentos em postagens de mídia social, permitindo análises de opinião em larga escala.
- Recomenda produtos ou conteúdo personalizado para usuários com base em seu comportamento e preferências.

XGBoost:

- Amplamente utilizado em competições de machine learning e ciência de dados.
- Avalia o risco associado a empréstimos ou investimentos com base em dados financeiros e comportamentais.
- Processamento de linguagem natural (PLN): Classifica textos, extrai informações...

Muito Obrigado!