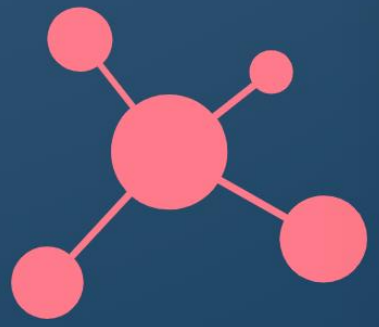
Tech&Dados
BY @ITALLEIRAS

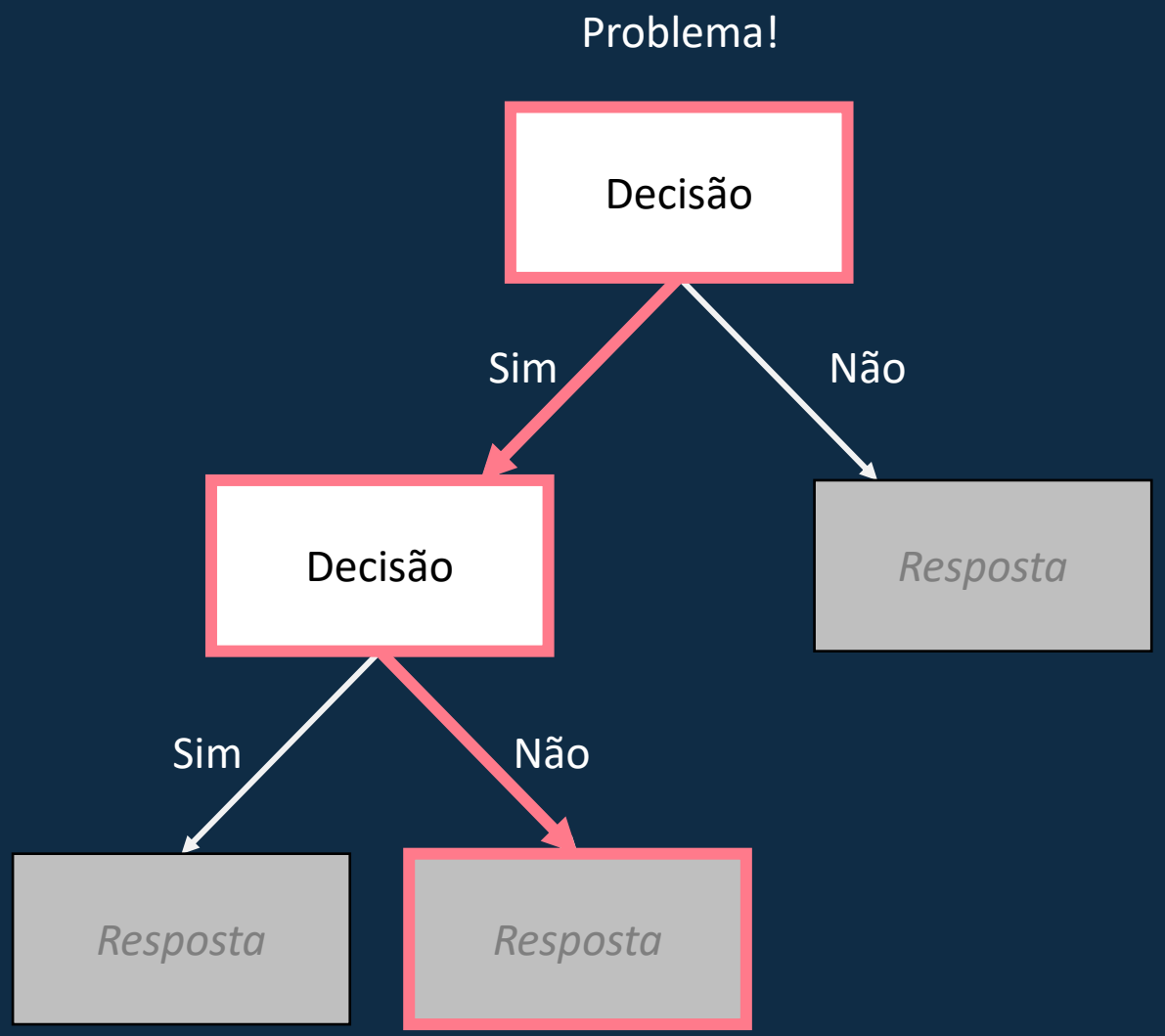


# Classificação: Árvores de Decisão

## (Decision Trees)

Funcionamento, exemplos, código e mais.

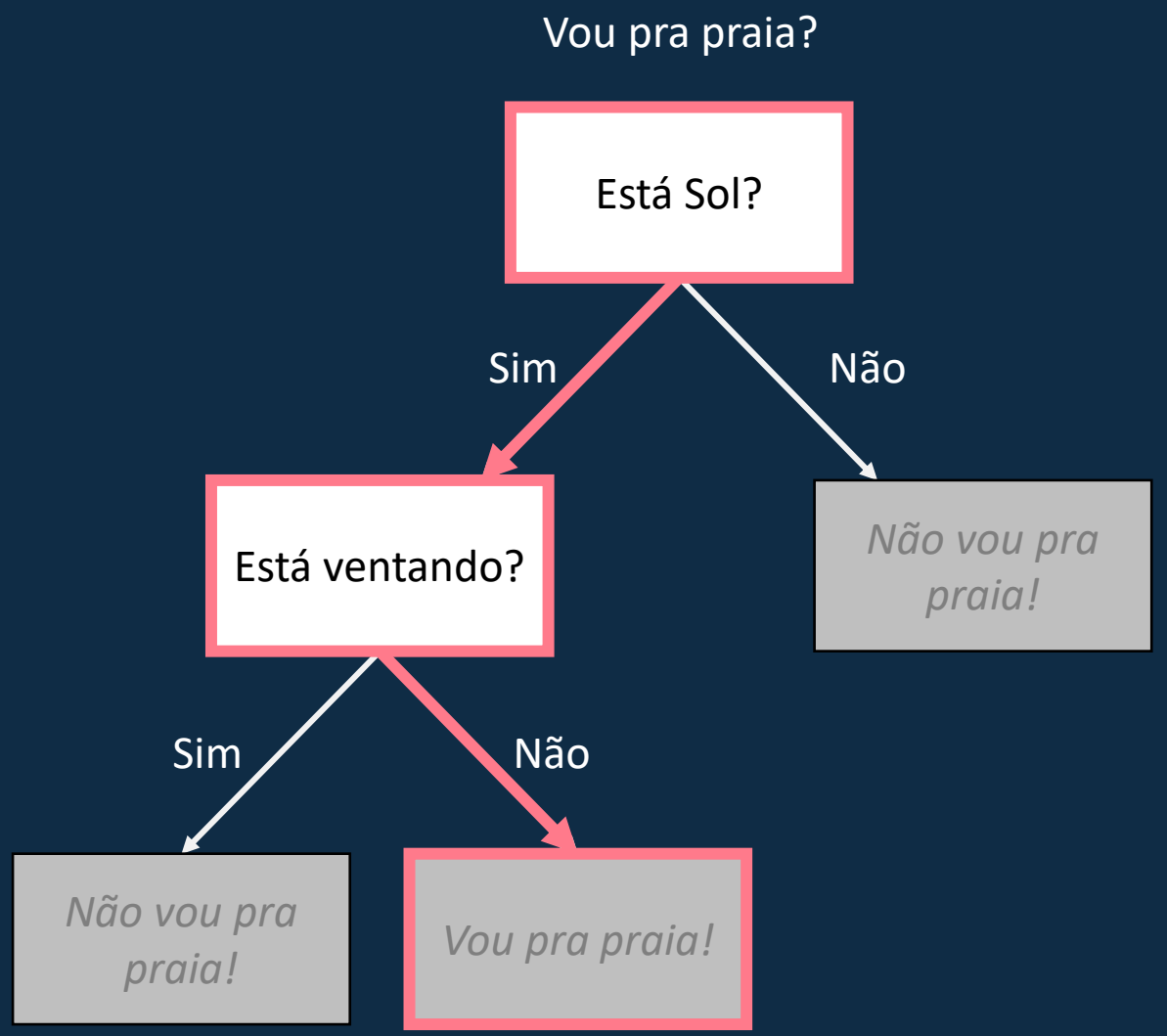
Setembro de 2023




Tipos de nó

Decisão

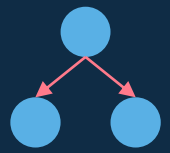
*Folhas*




Tipos de nó

Decisão

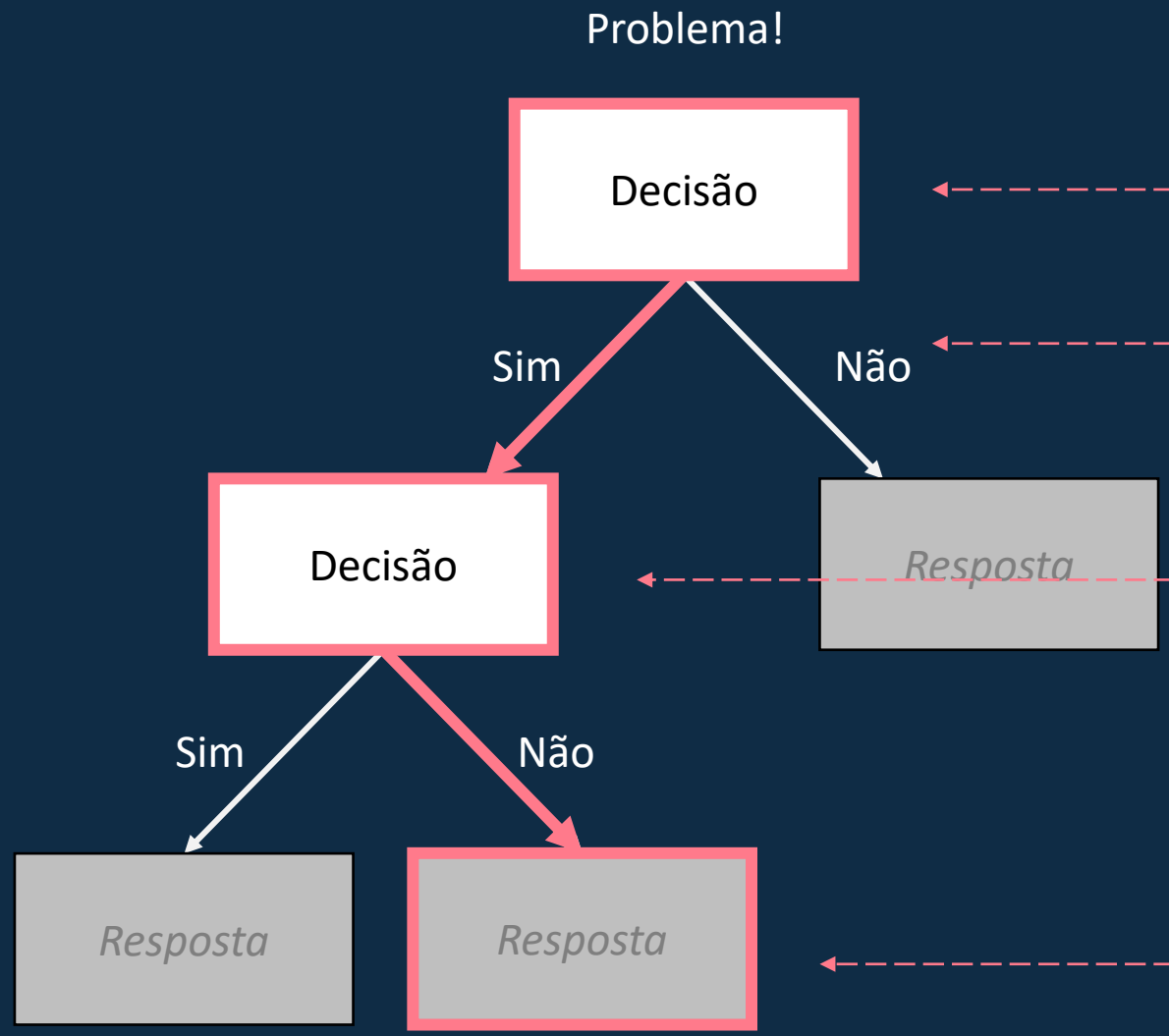
*Folhas*



# Árvore de Decisão

- Um algoritmo supervisionado
- Usado para classificação e regressão

# O que compõem uma Árvore de Decisão




Nó de decisão principal (todos os dados estão aqui!)

Os dados serão divididos em grupos menores

Nós de decisão continuarão a dividir o conjunto de dados

Até atingirmos os nós folhas
(aqui os registros são classificados)

Tipos de nó

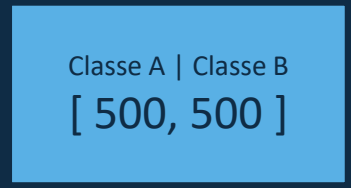
Decisão

*Folhas*

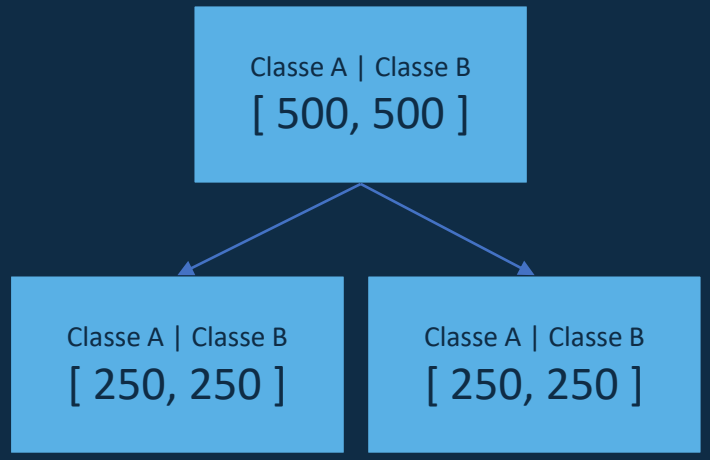
# Como as divisões são feitas?

Suponha uma base de dados com 1000 registros
- 500 pertencente a classe A
- 500 pertencente a classe B

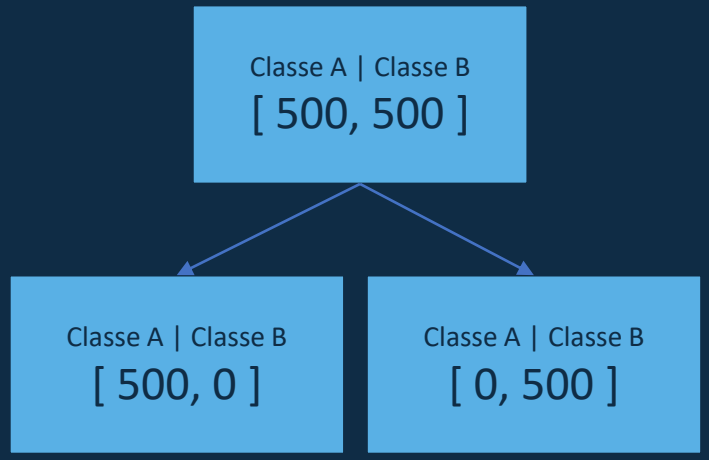
Precisamos separar os dados de forma a
diminuir a entropia / impureza
nos próximos nós e folhas.




Entropia alta



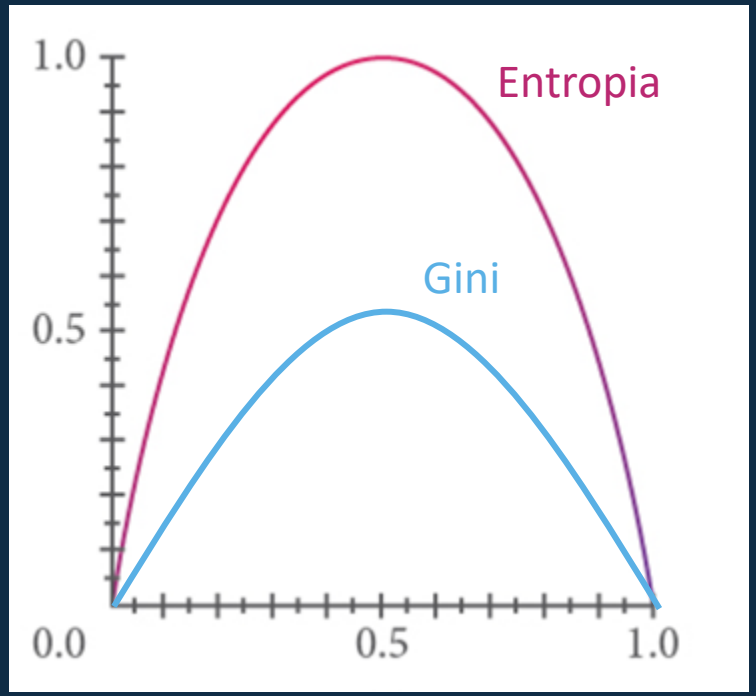

Entropia baixa



# Como as divisões são feitas?

Podemos realizar a divisão dos dados a partir de duas formas:

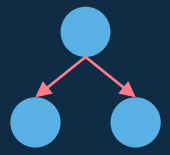
- Impureza de Gini
- Ganho de Informação (redução de entropia)




$$G.I. = Entropy\ (S) - \Sigma\ (\frac{|Sv|}{|S|} \times Entropy(Sv))$$

$$Entropy\ (S) = -\ \Sigma\ pi\ \log_2 pi$$

$$Gini = 1 - \Sigma\ pi^2$$



# Como as divisões são feitas?

Jogar ou não tênis?

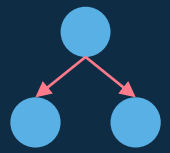
| Vento | Umidade | Temperatura | Jogar |
|---|---|---|---|
| Fraco | Alta | Quente | Sim |
| Forte | Baixa | Quente | Não |
| Fraco | Baixa | Quente | Não |
| Fraco | Baixa | Fria | Sim |
| Forte | Baixa | Fria | Sim |
| Forte | Alta | Quente | Não |
| Fraco | Alta | Fria | Sim |

A escolha dos atributos deve ser feita pensando em reduzir a entropia dos próximos nós e folhas.

7 Amostras
[ 4 , 3 ]
Sim , Não

$$Gini = 1 - \Sigma\, pi^2$$

Onde ***pi*** corresponde a proporção de dados em cada classe



# Como as divisões são feitas?

Jogar ou não tênis?

| Vento | Umidade | Temperatura | Jogar |
|---|---|---|---|
| Fraco | Alta | Quente | Sim |
| Forte | Baixa | Quente | Não |
| Fraco | Baixa | Quente | Não |
| Fraco | Baixa | Fria | Sim |
| Forte | Baixa | Fria | Sim |
| Forte | Alta | Quente | Não |
| Fraco | Alta | Fria | Sim |

3 registros “Sim” (3/4)
1 registros “Não” (1/4)

4 registros “Fraco” (4/7)
3 registros “Forte” (3/7)

$$Gini(Vento) = 1 - \Sigma\, pi^2$$

$$Gini(Vento|Fraco) = 1 - \left(\frac{3}{4}^2 + \frac{1}{4}^2\right) = 0{,}375$$

# Como as divisões são feitas?

Jogar ou não tênis?

| Vento | Umidade | Temperatura | Jogar |
|---|---|---|---|
| Fraco | Alta | Quente | Sim |
| Forte | Baixa | Quente | Não |
| Fraco | Baixa | Quente | Não |
| Fraco | Baixa | Fria | Sim |
| Forte | Baixa | Fria | Sim |
| Forte | Alta | Quente | Não |
| Fraco | Alta | Fria | Sim |

1 registros “Sim” (1/3)
2 registros “Não” (2/3)

4 registros “Fraco” (4/7)
3 registros “Forte” (3/7)

$$Gini(Vento) = 1 - \Sigma\, pi^2$$

$$Gini(Vento|Fraco) = 1 - \left(\frac{3}{4}^2 + \frac{1}{4}^2\right) = 0{,}375$$

$$Gini(Vento|Forte) = 1 - \left(\frac{1}{3}^2 + \frac{2}{3}^2\right) = 0{,}444$$

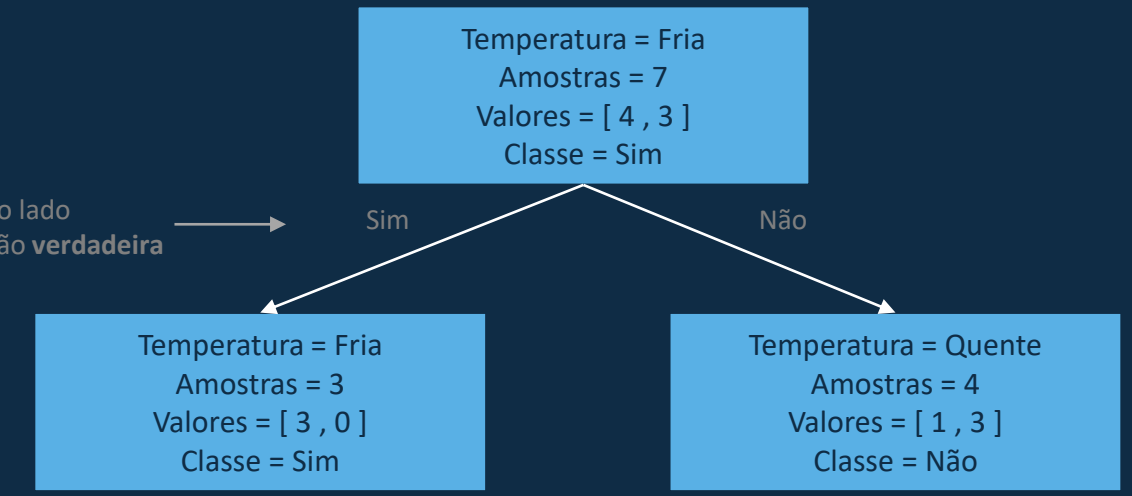
$$Gini\ Ponderada\ (VENTO) = \frac{4}{7} \times 0{,}375 + \frac{3}{7} \times 0{,}444 = \mathbf{0,404762}$$

# Como as divisões são feitas?

Jogar ou não tênis?

| Vento | Umidade | Temperatura | Jogar |
|---|---|---|---|
| Fraco | Alta | Quente | Sim |
| Forte | Baixa | Quente | Não |
| Fraco | Baixa | Quente | Não |
| Fraco | Baixa | Fria | Sim |
| Forte | Baixa | Fria | Sim |
| Forte | Alta | Quente | Não |
| Fraco | Alta | Fria | Sim |

Por convenção, a ramificação do lado esquerdo sempre leva a condição **verdadeira**




$$Gini\ Ponderada\ (Vento) = \mathbf{0,404762}$$

$$Gini\ Ponderada\ (Umidade) = \mathbf{0,476}$$

$$Gini\ Ponderada\ (Temperatura) = \mathbf{0,2143}$$

Menor impureza de Gini!

| Vento | Umidade | Temperatura | Jogar |
|---|---|---|---|
| Fraco | Baixa | Fria | Sim |
| Forte | Baixa | Fria | Sim |
| Fraco | Alta | Fria | Sim |

| Vento | Umidade | Temperatura | Jogar |
|---|---|---|---|
| Fraco | Alta | Quente | Sim |
| Forte | Baixa | Quente | Não |
| Fraco | Baixa | Quente | Não |
| Forte | Alta | Quente | Não |

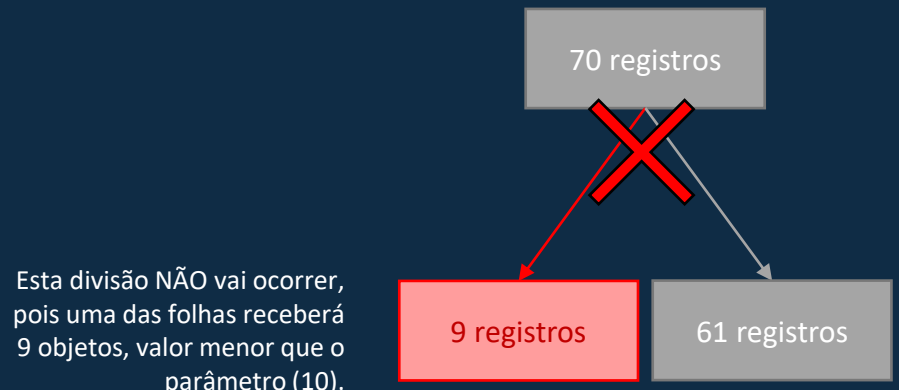
# Hiperparâmetros de uma árvore

**Principais hiperparâmetros para pré-poda:**

Min_samples_split: Número mínimo de elementos que um nó precisa ter para permitir que haja uma divisão.

Ex: Min_Samples_split = 20


Min_samples_leaf = Número mínimo de elementos necessários dentro de uma folha. Caso a divisão de um nó anterior acarrete em um número menor que este parâmetro em qualquer folha, a divisão não é realizada
Ex: Min_Samples_leaf = 10



# Hiperparâmetros de uma árvore

**Principais hiperparâmetros para pré-poda:**

Max_depth = defini a profundidade máxima da árvore (níveis).

Ex: Max depth= 2


Mesmo que haja condições para que a árvore continue crescendo, esta árvore chegará apenas até o nível 2 de profundidade.

# Vantagens

- Não é sensível a outliers. A árvore consegue lidar bem e “desprezar” dados com características  muito diferentes dos demais dados.
- Facilmente interpretável e pode ser implantada com simples regras if/else.

# Desvantagens

- Podem apresentar baixo viés e alta variância, ou seja, tem alto risco de overfitting na etapa de treinamento.
- Pode ser instável, isto é, pequenas variações no conjunto de treino podem produzir grandes variações na árvore final.

Técnicas de podagem da árvore e evolução do modelo para Random Forest.

# Para se aprofundar mais

- Como funciona o corte de variáveis contínuas.
- Conhecer os tipos e evoluções de árvores (ID3, C4. 5, CART, CHAID, MARS)

Modelo

# Árvore de Decisão

| | |
|---|---|
| Tipo | Classificação |
| Categoria | Supervisionado |

**Objetivo Resumido**

Cria uma árvore de decisão sobre os dados. O objetivo é criar divisões nos dados da melhor maneira possível, de modo que no topo da árvore (raiz) estejam todos os dados e na base da árvore (folhas) estejam as classes dos dados. O objetivo é criar regras nas quais os dados serão submetidos (regras If/Else) até que seja classificado com uma determinada classe.

| | | | |
|---|---|---|---|
| Premissas para o funcionamento | - Dentro do Scikit Learn, os dados precisam ser numéricos | Hiperparâmetros | - Profundidade máxima<br>- Min_sample_split<br>- Min_sample_leaf |

**Funcionamento detalhado (Steps, informações detalhadas etc.)**

**STEPS:**

1. Com base nos dados de entrada, é realizado um cálculo sobre todos os atributos para avaliar qual deles possui menor índice de impureza. Esse atributo é posicionado como raiz da árvore de decisão com uma regra bem definida (maior que, menor que, igual a, diferente de etc.)
2. São geradas então duas sub-árvores, uma quando a condição for verdadeira, outra quando a condição for falsa.
3. Com todos os dados que foram separados no nó, as etapas 1 e 2 são repetidas inúmeras vezes até que sejam atingidos quaisquer um dos parâmetros de parada da árvore.

As árvores possuem um nó raiz (topo da árvore), nós internos e folhas, onde os objetos são classificados.

| | | | |
|---|---|---|---|
| Vantagens | - Robusta a outliers. A árvore consegue lidar bem e "desprezar" dados com características muito diferentes dos demais dados.<br>- Facilmente interpretável e pode ser implantada com simples regras if/else. | Desvantagens | - Podem apresentar baixo viés e alta variância, ou seja, tem alto risco de overfitting na etapa de treinamento.<br>- Pode ser instável, isto é, pequenas variações no conjunto de treino podem produzir grandes variações na árvore final. |
| Como avaliar o desempenho | - O desempenho durante o teste pode ser medido com cálculos de acurácia, precisão, recall, curva ROC etc. | Como corrigir ou compensar desvantagens | Técnicas de pré-poda ou pós-poda podem ser aplicados para aumentar o poder de generalização do algoritmo, encontrando um melhor equilíbrio entre viés e variância. |

# Obrigado!

youtube.com/@Tech_dados

linkedin.com/in/itallo-dias/