



# Entendendo Árvores de Decisão

Bem-vindos à nossa apresentação sobre árvores de decisão, um dos modelos mais intuitivos e poderosos em aprendizado de máquina. Vamos explorar como esse algoritmo funciona, suas aplicações práticas e como ele se compara com outros métodos populares.

# O que é uma Árvore de Decisão?

Uma árvore de decisão é um modelo de aprendizado de máquina que simula o processo de tomada de decisões humanas. Funciona como um fluxograma hierárquico, onde cada passo nos leva a uma conclusão com base em perguntas simples de sim/não.

- 1 Nós Internos
- 2 Ramificações
- 3 Nós Folha

Representam perguntas sobre características dos dados.  
Cada resposta direciona para um caminho diferente.

São as possíveis respostas para as perguntas dos nós,  
levando a novos nós ou folhas.

São os nós terminais que representam a decisão final ou a  
previsão do modelo.



A importância das árvores de decisão reside na sua interpretabilidade:  
você pode facilmente seguir o caminho das decisões para entender como  
o modelo chegou a uma determinada previsão.

# Aplicações em Machine Learning



## Classificação

Usada para decisões categóricas, como aprovar ou negar um empréstimo bancário.



## Regressão

Utilizada para prever valores numéricos, como o preço de uma casa.



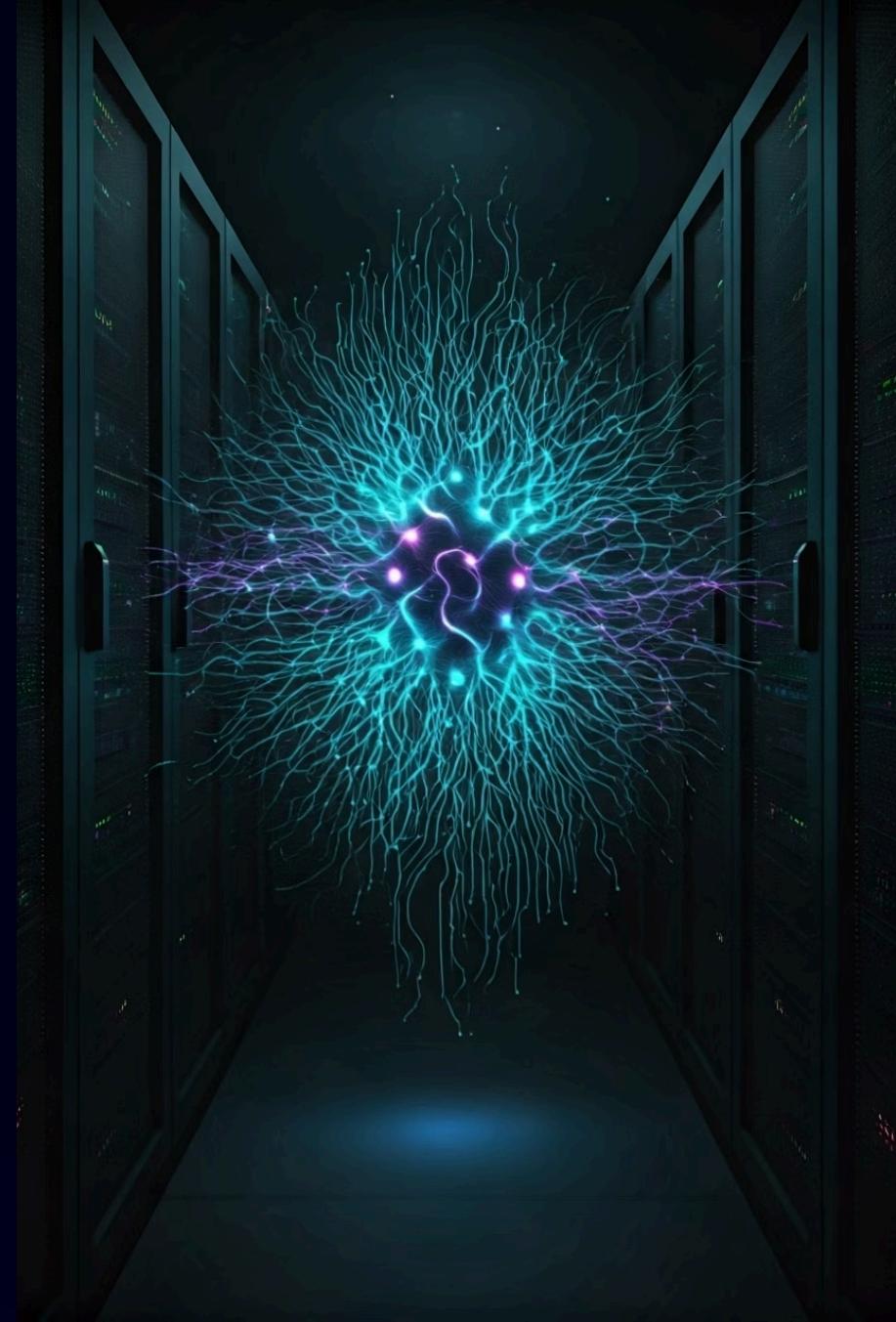
## Interpretabilidade

O algoritmo aprende automaticamente quais perguntas fazer (variáveis e pontos de corte) analisando os dados de treinamento.



## Visualização

A representação em forma de árvore permite visualizar quais atributos foram mais importantes para a previsão.



# Aplicações em Finanças e Negócios



Análise de risco de crédito



Segmentação de clientes

Devido à sua interpretabilidade, árvores de decisão são muito utilizadas em cenários onde é preciso explicar o porquê de uma decisão foi tomada



Detecção de fraude



Decisões de marketing

Em áreas como finanças e negócios, onde há regulação e necessidade de transparência, as árvores de decisão oferecem uma forma de IA explicável – um modelo em que cada decisão pode ser rastreada e justificada facilmente.

# Coeficiente de Gini e Impureza da Divisão

Para construir uma árvore de decisão, o algoritmo precisa escolher em cada nó qual variável (e em que ponto) usar para dividir os dados. Esse "melhor corte" é determinado medindo o quanto puras ficam as subdivisões em termos da variável alvo.

## O que é Impureza de Gini?

Mede o grau de mistura das classes em um conjunto de dados. Em um nó de decisão, se todos os exemplos pertencem à mesma classe, o nó é puro ( $Gini = 0$ ). Se os exemplos estão divididos meio a meio entre as classes, o nó é altamente impuro.



### Nó Puro

Quando todos os exemplos em um nó pertencem à **mesma classe**.

$Gini = 0$



### Nó Impuro

Quando os exemplos estão **misturados** entre diferentes classes.

Gini próximo de 0.5 (para 2 classes)

# Cálculo do Coeficiente de Gini

## Cálculo do Coeficiente de Gini

A impureza de Gini é calculada com base na proporção das classes dentro de um nó. Um valor mais baixo indica um nó mais puro.

### Fórmula do Gini

Para duas classes, o cálculo do Gini pode ser expresso como:

$$Gini = 1 - (p_1^2 + p_2^2)$$

onde  $p_1$  é a proporção de exemplos da classe 1 no nó, e  $p_2$  a proporção da classe 2.

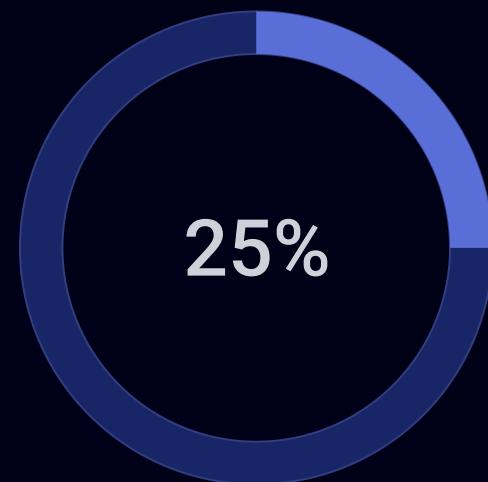
## Visualizando a Impureza



Gini (Nó Puro)



Gini (Nó Impuro)



25%

Gini (Impureza Média)

Todos os elementos são da mesma classe.

Classes divididas igualmente (máxima impureza para 2 classes).

Um nó com alguma mistura, mas não máxima.

# Analogia da Moeda

Podemos usar uma analogia da moeda para entender a impureza de Gini de forma intuitiva:



## Moeda Equilibrada (50% Cara, 50% Coroa)

Representa **máxima imprevisibilidade**, equivalente a  $\text{Gini} = 0,5$  (no caso binário). Assim como um nó da árvore que ainda tem uma mistura equilibrada de classes é "incerto" (impuro).



## Moeda Viciada (100% um Resultado)

Tem **imprevisibilidade nula**, equivalente a  $\text{Gini} = 0$ . Da mesma forma, um nó com casos todos iguais é totalmente certo (puro).



# Exemplo de Cálculo do Gini

Suponha um conjunto com 20 exemplos, dos quais 10 são da Classe A e 10 da Classe B. Antes de qualquer divisão, a impureza Gini seria 0,5 (situação mais impura). Após uma divisão que resulta em um nó com 18 exemplos da Classe A e 2 da Classe B, o Gini seria 0,18 - muito mais puro que o inicial.

**Gini Máximo**

**0,5**

Classes balanceadas (50% A, 50% B)

**Gini Médio**

**0,18**

Classes desbalanceadas (90% A, 10% B)

**Gini Mínimo**

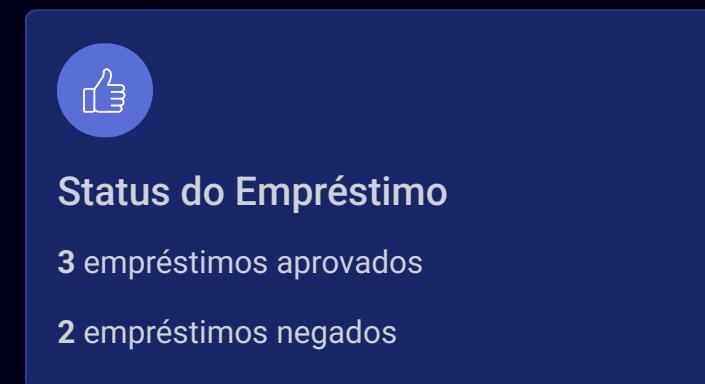
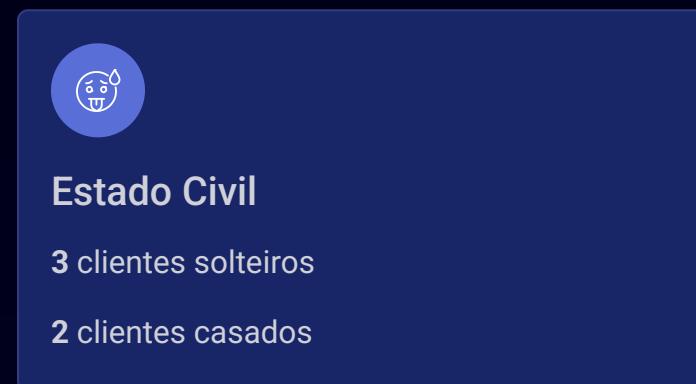
**0,0**

Classes puras (100% A, 0% B)

# Exemplo Prático: Base de Dados de Empréstimo

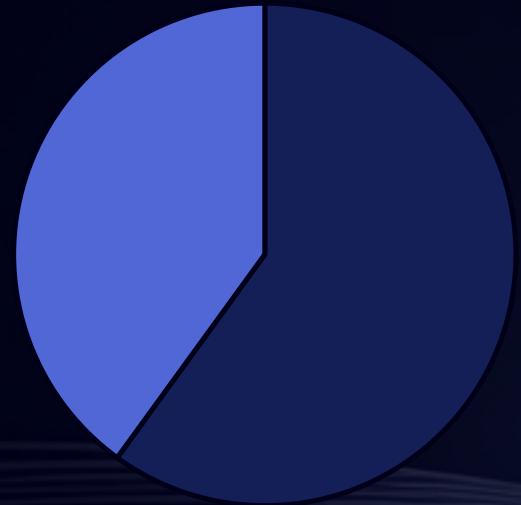
Vamos construir manualmente uma pequena árvore de decisão usando uma mini base de dados de solicitação de empréstimo:

Estado Civil	Renda Mensal	Empréstimo Aprovado?
Solteiro	1000	Não (0)
Casado	2000	Sim (1)
Solteiro	3500	Não (0)
Casado	4000	Sim (1)
Solteiro	5000	Sim (1)



# Construindo a Árvore: Passo 1

Distribuição dos empréstimos no nó raiz: 60% aprovados, 40% negados.



■ Aprovado (Sim) ■ Negado (Não)



## Índice Gini do Nó Raiz

A impureza Gini é de **0,48**.

Este valor é bastante alto, indicando uma mistura significativa das classes "Sim" e "Não" antes de qualquer divisão.

## Impureza do Nó Raiz

Antes de qualquer divisão, todos os 5 exemplos estão juntos no nó raiz. Vamos calcular a impureza inicial usando o índice Gini:

### Probabilidade de Aprovação

$$p(\text{Sim}) = 3 \text{ aprovados} / 5 \text{ total} = 0,6$$

### Probabilidade de Reprovação

$$p(\text{Não}) = 2 \text{ negados} / 5 \text{ total} = 0,4$$

Agora, aplicamos a fórmula do Gini para o nó raiz:

$$Gini_{raiz} = 1 - (p(\text{Sim})^2 + p(\text{Não})^2)$$

$$Gini_{raiz} = 1 - (0.6^2 + 0.4^2)$$

$$Gini_{raiz} = 1 - (0.36 + 0.16)$$

$$Gini_{raiz} = 1 - 0.52$$

$$Gini_{raiz} = 0.48$$

# Construindo a Árvore: Passo 2

## Primeiro Candidato de Divisão – Estado Civil

### Grupo Solteiro

Contém 3 exemplos: 2 negados, 1 aprovado

Proporção: 66,7% "Não", 33,3% "Sim"

Gsolteiros ≈ 0,44 (ainda impuro)

### Grupo Casado

Contém 2 exemplos: ambos aprovados

Proporção: 0% "Não", 100% "Sim"

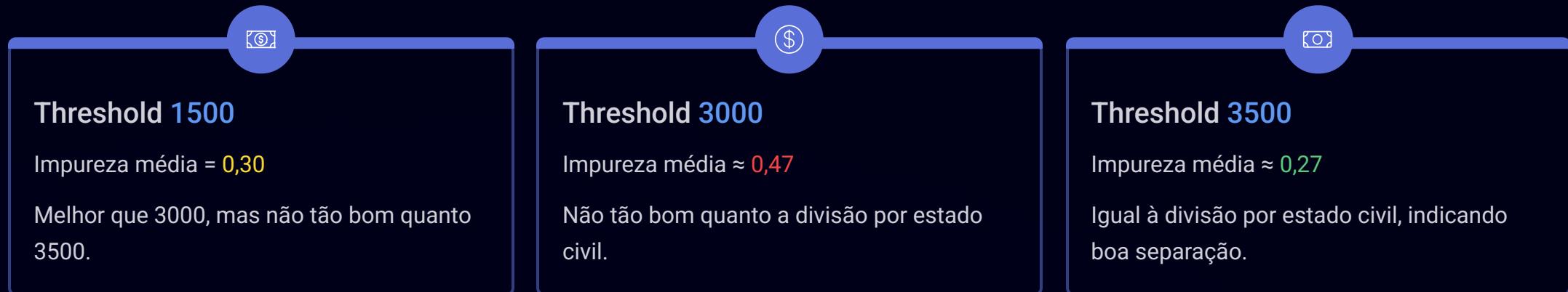
Gcasados = 0 (nó puro)

ⓘ Impureza média ponderada após esta divisão: Gpós-split =  $(3/5) \times 0,44 + (2/5) \times 0,0 \approx 0,27$

Comparando com Gini = 0,48 do nó raiz, houve uma redução significativa na impureza.

# Construindo a Árvore: Passo 3

## Segundo Candidato de Divisão – Renda Mensal



O melhor threshold encontrado foi em torno de 3500, que oferece uma boa separação entre quem tem renda mais alta (geralmente aprovados) e os demais. Isso resulta em uma impureza média mais baixa, comparável à divisão por estado civil.

# Árvore de Decisão Final



# Comparação: Árvores de Decisão vs. Outros Métodos



## Árvore de Decisão

Alta interpretabilidade, regras simples e transparentes. Lida bem com variáveis numéricas e categóricas. Pode sobreajustar se crescer demais.



## Regressão Logística

Interpretabilidade através de coeficientes. Supõe relação linear entre features e log-odds. Não captura bem interações entre variáveis.

# Comparação: Árvores de Decisão vs. Outros Métodos (cont.)



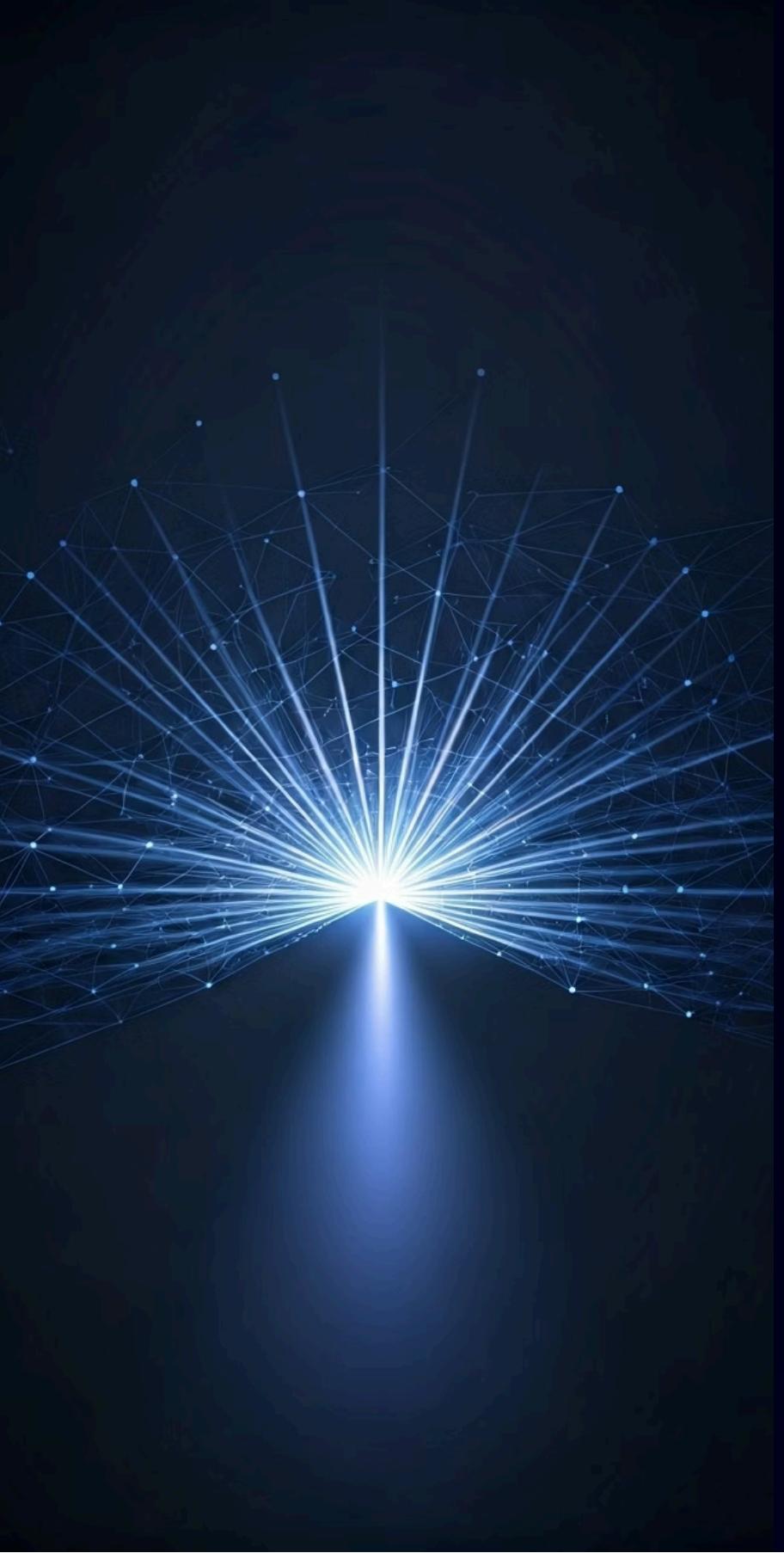
## K-Nearest Neighbors

Baixa interpretabilidade global. Não produz modelo explícito.  
Justificativa local baseada em vizinhos. Ineficiente com muitos dados  
e dimensões altas.



## Support Vector Machine

Modelo "caixa-preta" com baixa interpretabilidade. Pode modelar  
fronteiras complexas. Difícil explicar decisões em termos simples.



# Conclusão: Quando Usar Árvores de Decisão

Árvores de decisão fornecem um compromisso valioso entre interpretação e performance. São ideais quando:



A interpretabilidade é crucial (finanças, saúde)



Precisamos justificar decisões de forma clara



Queremos identificar os fatores mais importantes



Há interações complexas entre variáveis

Em muitos problemas de negócios, opta-se por um modelo ligeiramente menos preciso se ele puder fornecer justificativas claras para cada decisão – e é exatamente nessa situação que as árvores de decisão brilham.