

Insper

Machine Learning

Aula 12 – Decision Trees

2021 – Engenharia
Fábio Ayres <fabioja@insper.edu.br>

Iris



Iris Versicolor

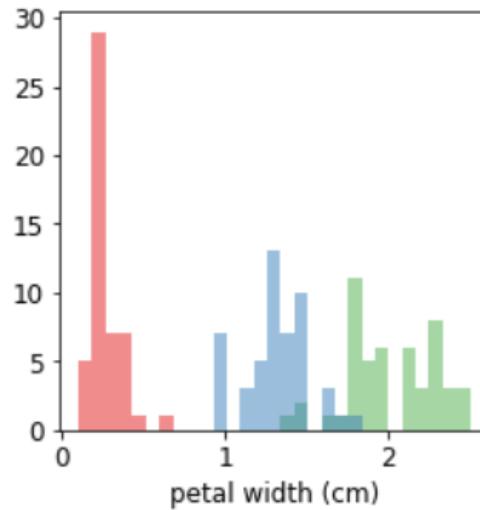
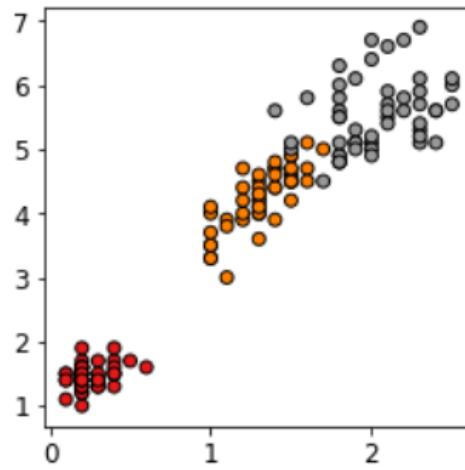
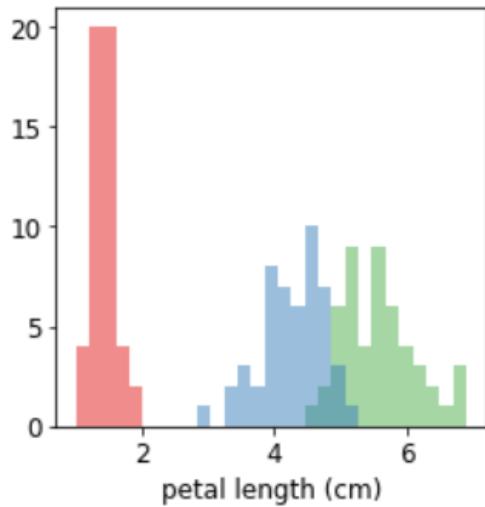


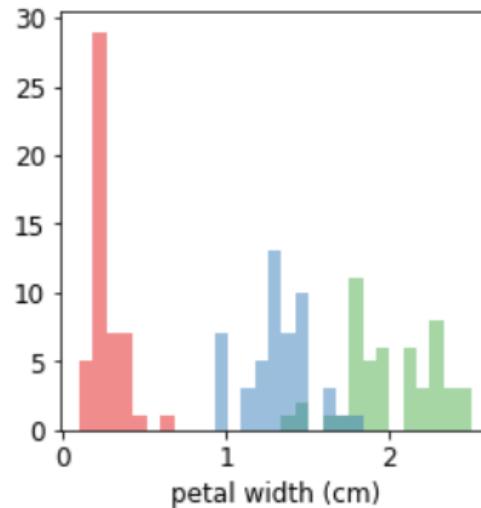
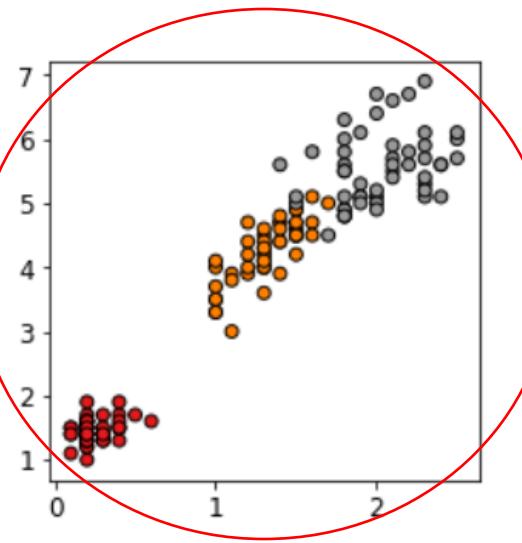
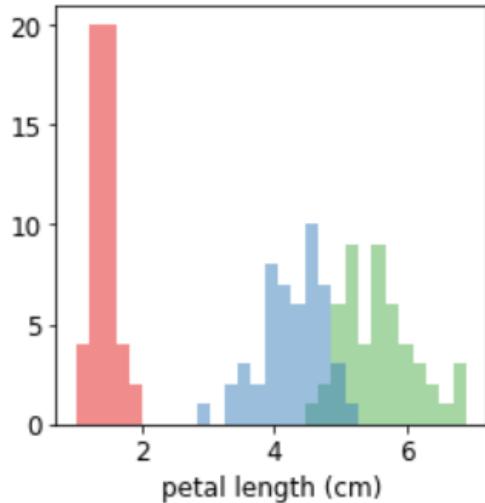
Iris Setosa

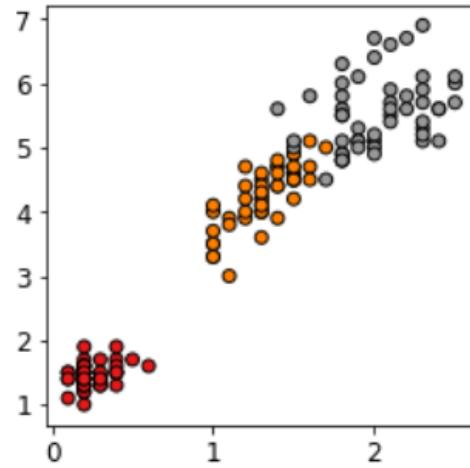
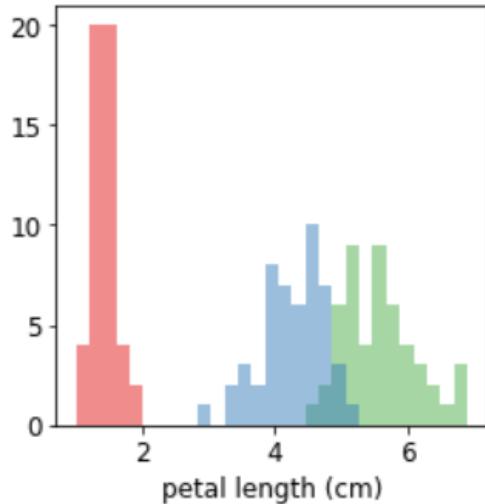


Iris Virginica

Fonte: <https://www.datacamp.com/community/tutorials/machine-learning-in-r>

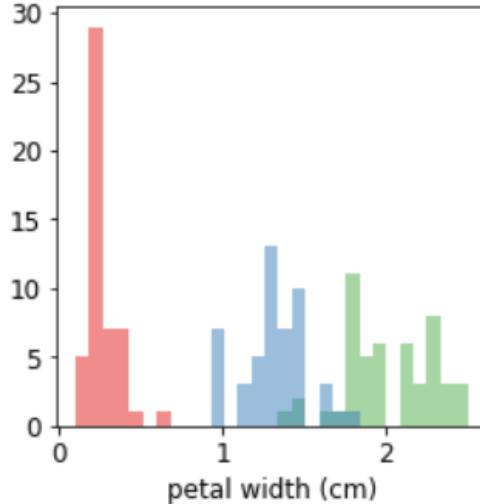


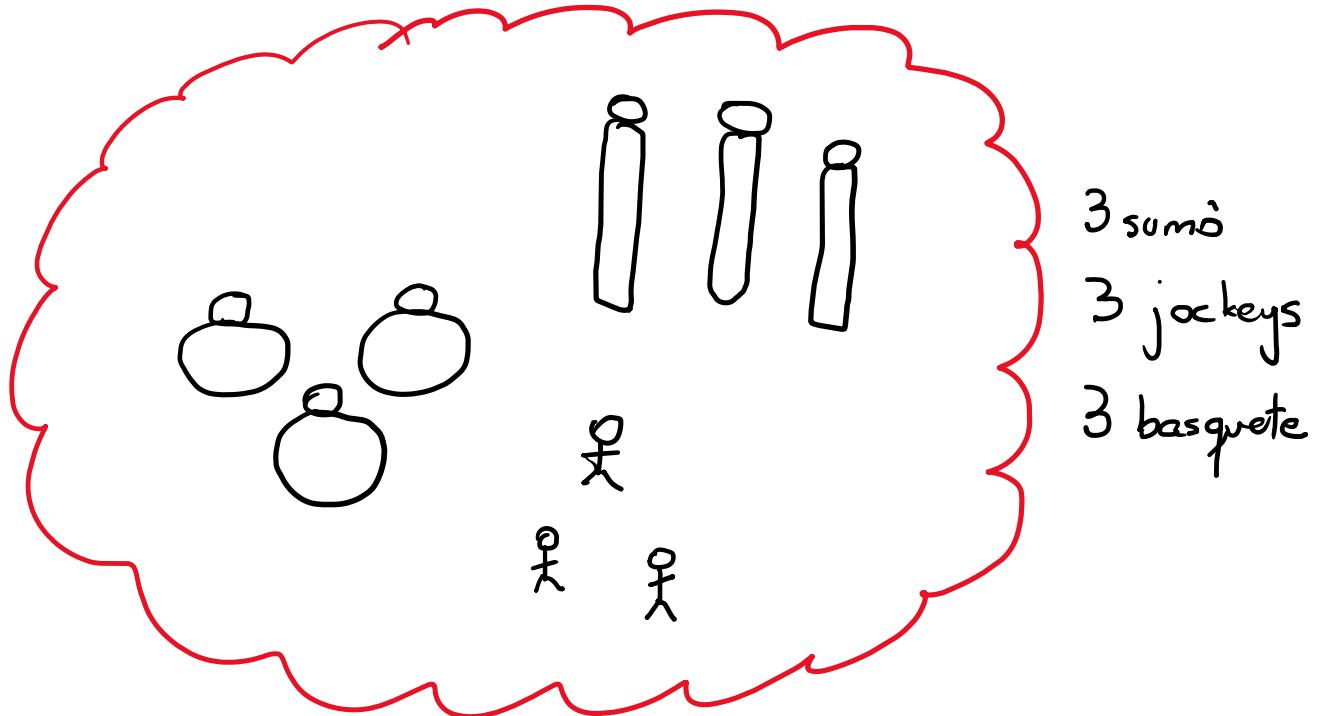




Ideia:

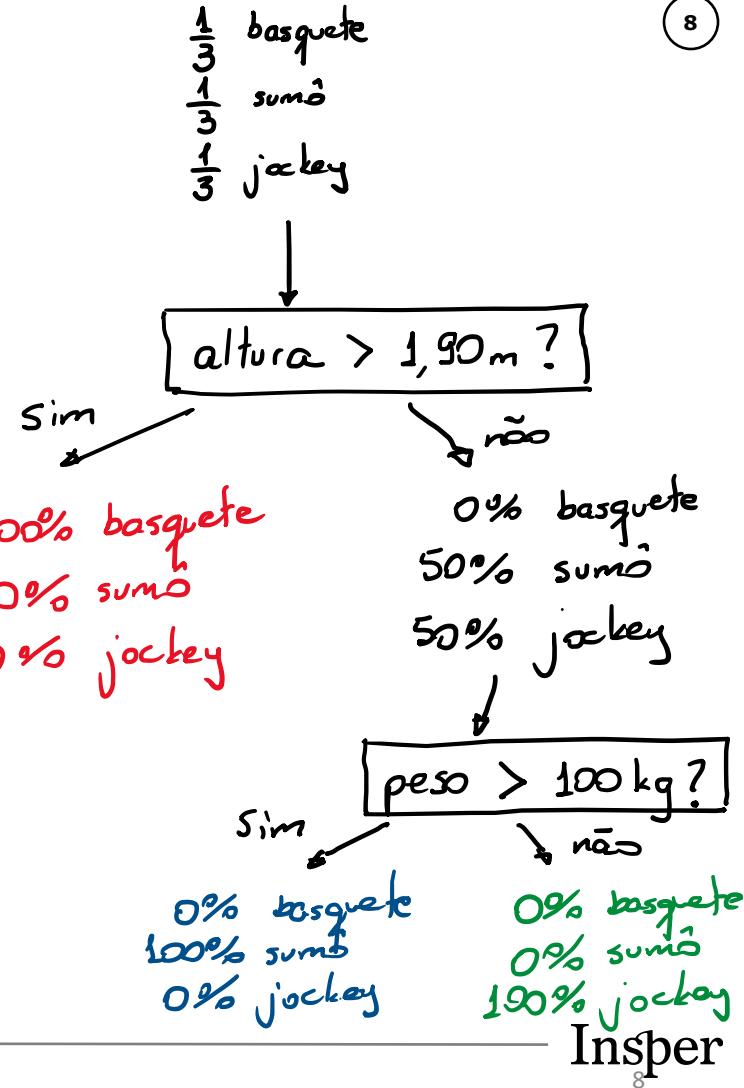
- escolhe uma feature
- escolhe um limiar
- separa em conjuntos
mais homogêneos





X_{train}		y_{train}	
peso	altura	esporte	
140	2,15	basquete	$\frac{1}{3}$ basquete
130	1,70	sumô	$\frac{1}{3}$ sumô
79	1,70	jockey	$\frac{1}{3}$ jockey
65	1,72	jockey	$\frac{1}{3}$ jockey
120	1,70	sumô	
125	1,75	sumô	
120	2,05	basquete	
75	1,79	jockey	
115	1,95	basquete	

x_{train}	y_{train}	
peso	altura	esporte
140	2,15	basquete
130	1,70	sumô
79	1,70	jockey
65	1,72	jockey
120	1,70	sumô
125	1,75	sumô
120	2,05	basquete
75	1,79	jockey
115	1,95	basquete



CART

Classification and Regression
Trees

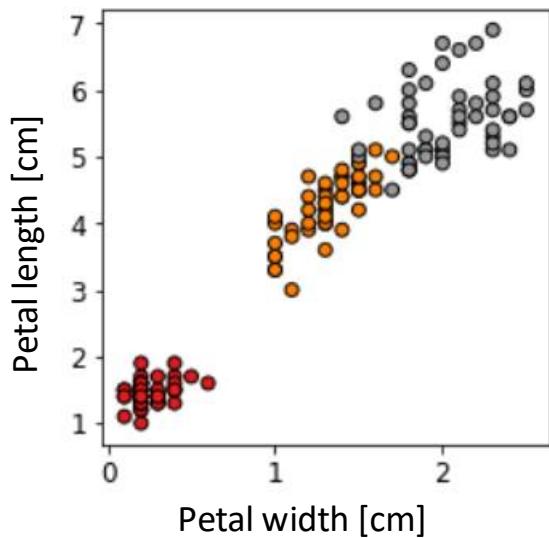
Medidas de “impureza”

- Queremos uma medida de impureza que seja
 - Zero para um conjunto completamente homogêneo
 - Se eu doubro (triplico, etc) o número de elementos em cada classe, a impureza é a mesma
 - Só a proporção de elementos por classe importa

Medidas de “impureza”

- Queremos uma medida de impureza que seja
 - Aumente para conjuntos mais misturados
 - Se o número de elementos por classe for o mesmo para todas as classes, a impureza é máxima para aquele número de classes
 - Quanto maior o número de classes existentes, maior a impureza máxima

Exemplo



Frequencias
por classe:

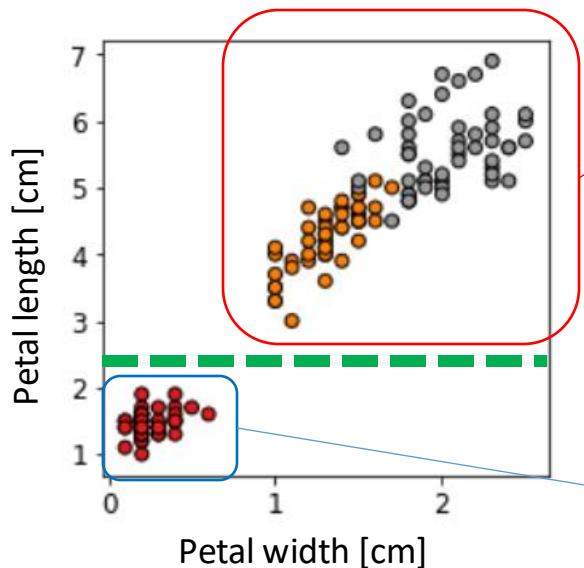
- 50 setosa
- 50 versicolor
- 50 virginica

Proporções
por classe:

$$p = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right]$$

Dividindo por limiar para uma feature escolhida

Mais puro!



Frequencias
por classe:

- 0 setosa
- 50 versicolor
- 50 virginica

Frequencias
por classe:

- 50 setosa
- 0 versicolor
- 0 virginica

Proporções
por classe:

$$p = \left[0, \frac{1}{2}, \frac{1}{2}\right]$$

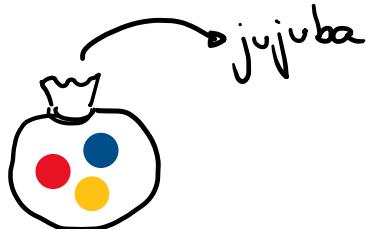
Proporções
por classe:

$$p = [1, 0, 0]$$

Medidas de impureza comuns

Medida	Coeficiente Gini	Entropia
Definição	$G = 1 - \sum p_i^2$	$E = -\sum p_i \log_2 p_i$
Conjunto homogeneo $p_1 = 1$, no resto $p_i = 0$	$\begin{aligned} G &= 1 - (1^2 + 0^2 + \dots + 0^2) \\ &= 1 - 1 \\ &= 0 \end{aligned}$	$E = - \left(\begin{aligned} 1 \times \log_2 1 \\ + 0 \times \log_2 0 \\ + \dots \\ + 0 \times \log_2 0 \end{aligned} \right) = 0$
Conjunto heterogeneo $p_i = 1/C$	$\begin{aligned} G &= 1 - \left(\sum \left(\frac{1}{C} \right)^2 \right) \\ &= 1 - C \times \frac{1}{C^2} \\ &= 1 - \frac{1}{C} \end{aligned}$	$\begin{aligned} E &= - \left(\sum \frac{1}{C} \times \log_2 \frac{1}{C} \right) \\ &= -C \times \frac{1}{C} \times \log_2 \frac{1}{C} \\ &= \log_2 C \end{aligned}$

Entropia



i) $P(\bullet) = 1$
 $P(\bullet) = 0$
 $P(\bullet) = 0$

→ não tem emoções!

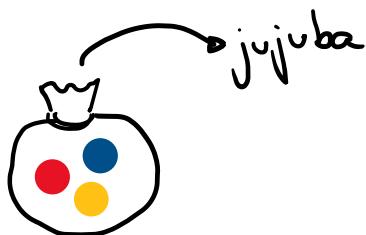
Tirei uma jujuba e deu \bullet : você está surpreso?
 - Claro que não!

ii) $P(\bullet) = \frac{1}{3}$
 $P(\bullet) = \frac{1}{3}$
 $P(\bullet) = \frac{1}{3}$

→ tem emoções!

Tirei uma jujuba e deu \bullet : você está surpreso?
 - Sim, podia ser outra cor!

Entropia



$$\text{iii) } P(\bullet) = 0,95$$

$$P(\bullet) = 0,04$$

$$P(\bullet) = 0,01$$

Tirei uma jujuba e
deu : você está surpreso? → não muito

Tirei uma jujuba e
deu : você está surpreso? → sim! baixa

Tirei uma jujuba e → sim, muito! baixissima
deu ☺ : você está surpreso?

Entropia

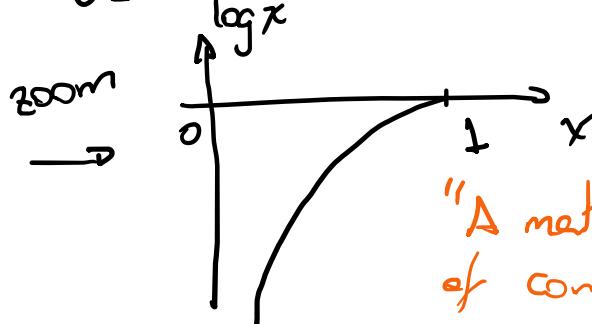
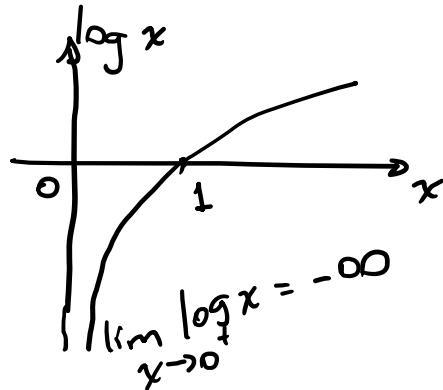
Medida da informação média de uma fonte de dados

↳ O quanto eu não sabia e agora sei.

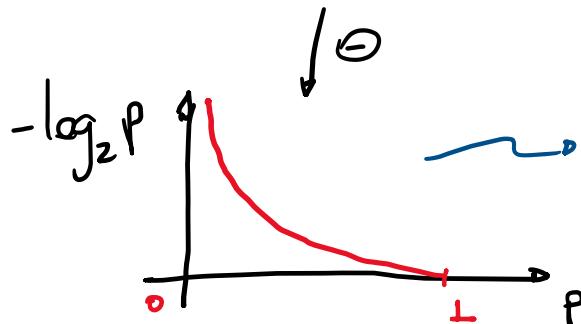
- Algo comum não é surpreendente!
→ não traz muita informação!
- Algo raro é surpreendente!
→ traz muita informação!

Medida de informação

Ideia: usar $-\log_2 p$ como medida de informações



*"A mathematical Theory
of communication"
Claude Shannon*



tem algumas
propriedades
desejáveis.

Unidade de medida da informação

$-\log_b \text{prob}$

b = 2 : bits
b = e : nats
b = 10 : hartleys

“Surpresa” média

jujuba surpresa

$$\begin{aligned} & -\log_2(p_{\text{red}}) \\ & -\log_2(p_{\text{red}}) \\ & -\log_2(p_{\text{red}}) \\ & -\log_2(p_{\text{blue}}) \\ & -\log_2(p_{\text{yellow}}) \end{aligned}$$

:

Na média:

$$\mathbb{E}[-\log_2(p_{\text{jujuba}})]$$

$$= \sum p_{\text{jujuba}} \cdot (-\log_2 p_{\text{jujuba}})$$

$$= - \sum_{i \in \{\text{red}, \text{blue}, \text{yellow}\}} p_i \log p_i$$

Entropia



$$64 \text{ casas} = 2^6$$

$\Rightarrow 6$ perguntas

$$8 \text{ linhas} = 2^3$$

$\Rightarrow 3$ perguntas

$$8 \text{ colunas} = 2^3$$

$\Rightarrow 3$ perguntas

X_{train}		y_{train}
peso	altura	esporte
140	2,15	basquete
130	1,70	sumô
79	1,70	jockey
65	1,72	jockey
120	1,70	sumô
125	1,75	sumô
120	2,05	basquete
75	1,79	jockey
115	1,95	basquete

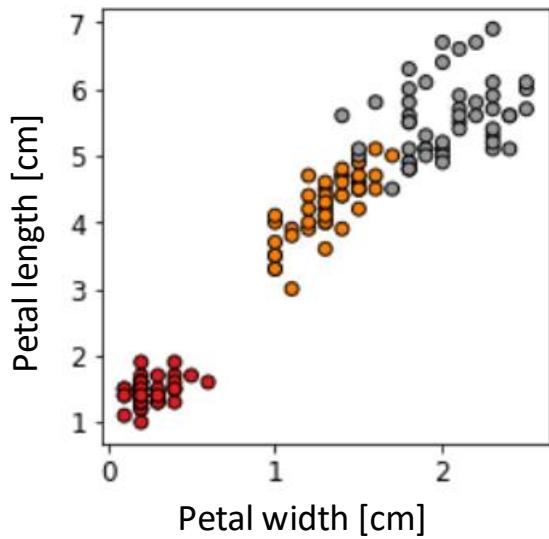
Jogo de perguntas
(árvore de decisões)

- Cada pergunta te traz mais informações



diminui a incerteza!

Exemplo



Frequencias
por classe:

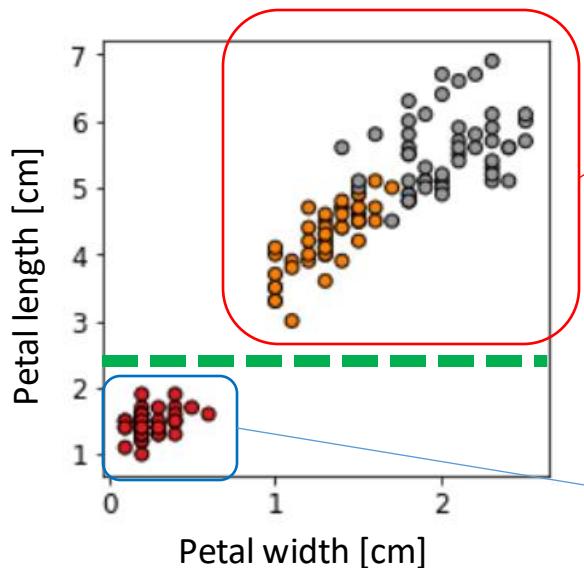
- 50 setosa
- 50 versicolor
- 50 virginica

Proporções
por classe:

$$p = \left[\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right]$$

Dividindo por limiar para uma feature escolhida

Mais puro!



Frequencias
por classe:

- 0 setosa
- 50 versicolor
- 50 virginica

Frequencias
por classe:

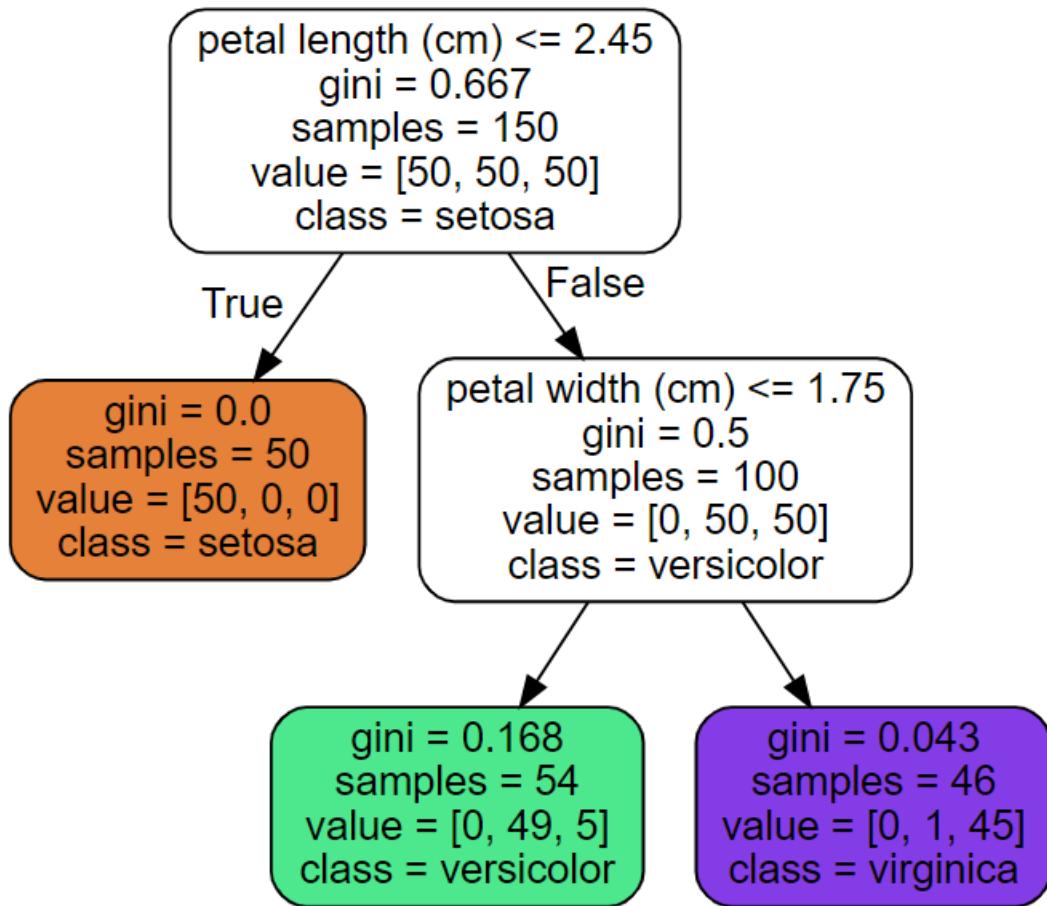
- 50 setosa
- 0 versicolor
- 0 virginica

Proporções
por classe:

$$p = \left[0, \frac{1}{2}, \frac{1}{2}\right]$$

Proporções
por classe:

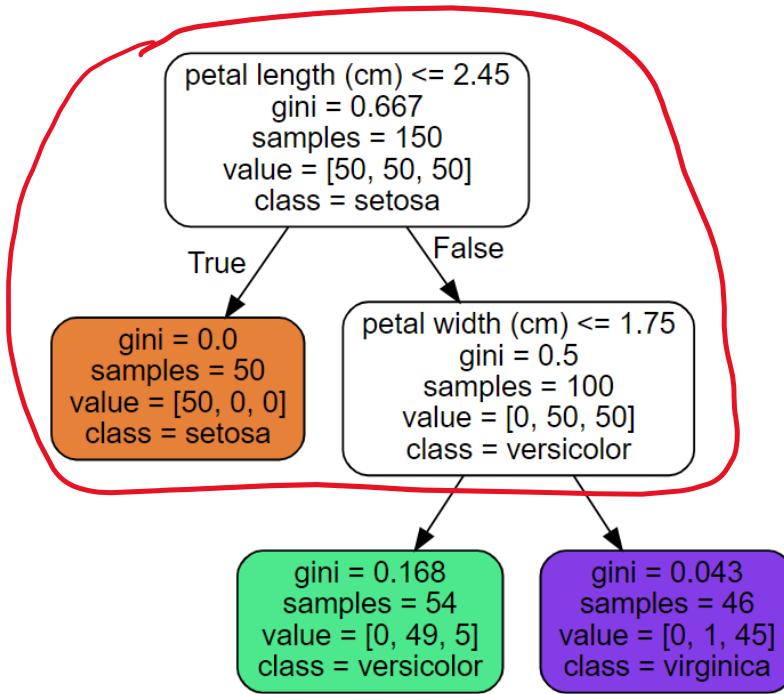
$$p = [1, 0, 0]$$



Comparando pais e filhos

- $G_{\text{medio}} = \frac{m_{\text{left}}}{m} \cdot G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}}$
- Escolher a partição que maximiza

$$G_{\text{pai}} - G_{\text{médio}}$$



$$G_{\text{par}} = 0,667 \quad m = 150$$

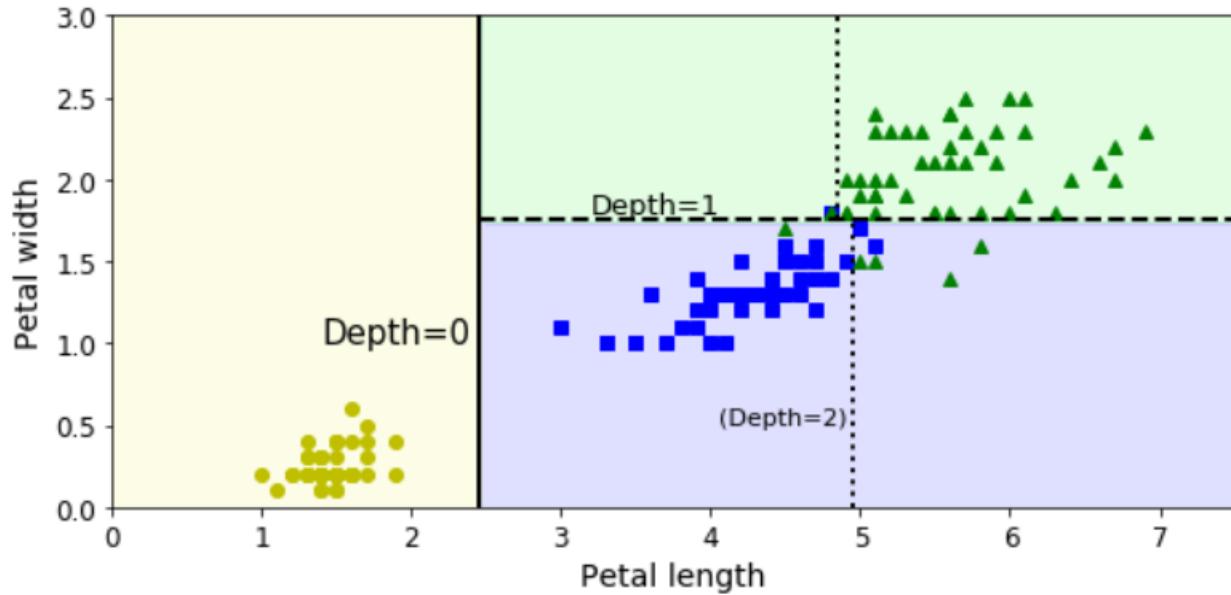
$$G_{\text{left}} = 0 \quad m_{\text{left}} = 50$$

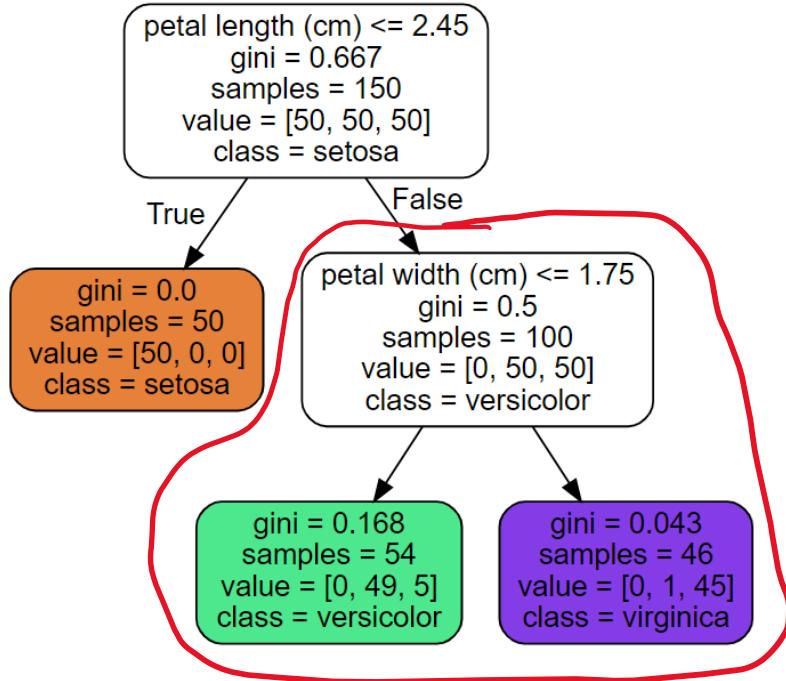
$$G_{\text{right}} = 0,5 \quad m_{\text{right}} = 100$$

$$G_{\text{medio}} = \frac{50}{150} \cdot 0 + \frac{100}{150} \cdot 0,5$$

$$= \frac{1}{3}$$

$G_{\text{medio}} < G_{\text{par}}$





$$G_{\text{pai}} = 0,5 \quad m = 100$$

$$G_{\text{left}} = 0,168 \quad m_{\text{left}} = 54$$

$$G_{\text{right}} = 0,043 \quad m_{\text{right}} = 46$$

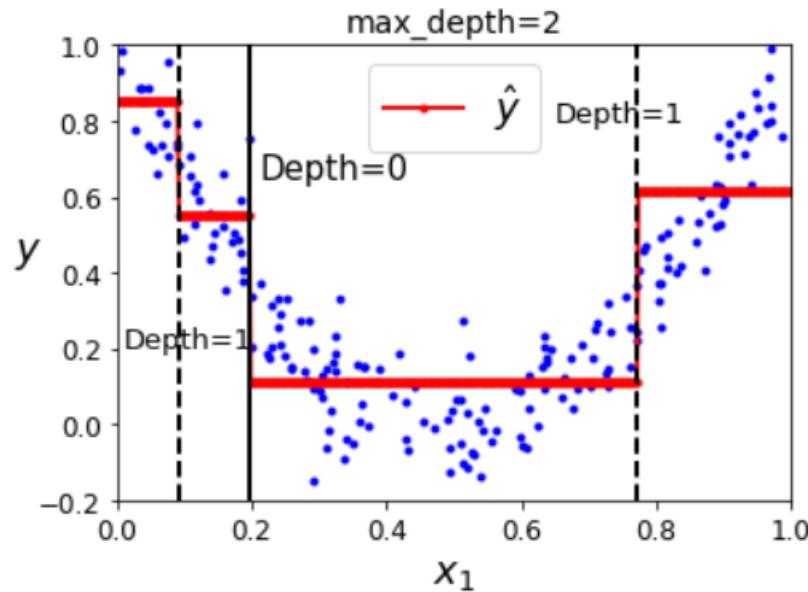
$$G_{\text{medio}} = \frac{54}{100} \cdot 0,168 + \frac{46}{100} \cdot 0,043 \\ = 0,11$$

$G_{\text{pai}} > G_{\text{medio}}$

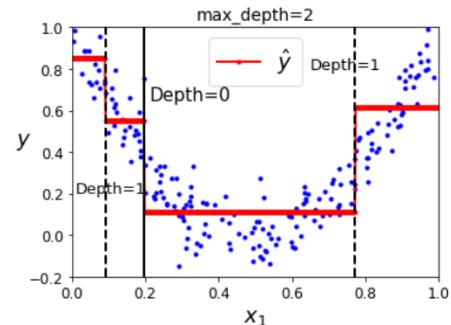
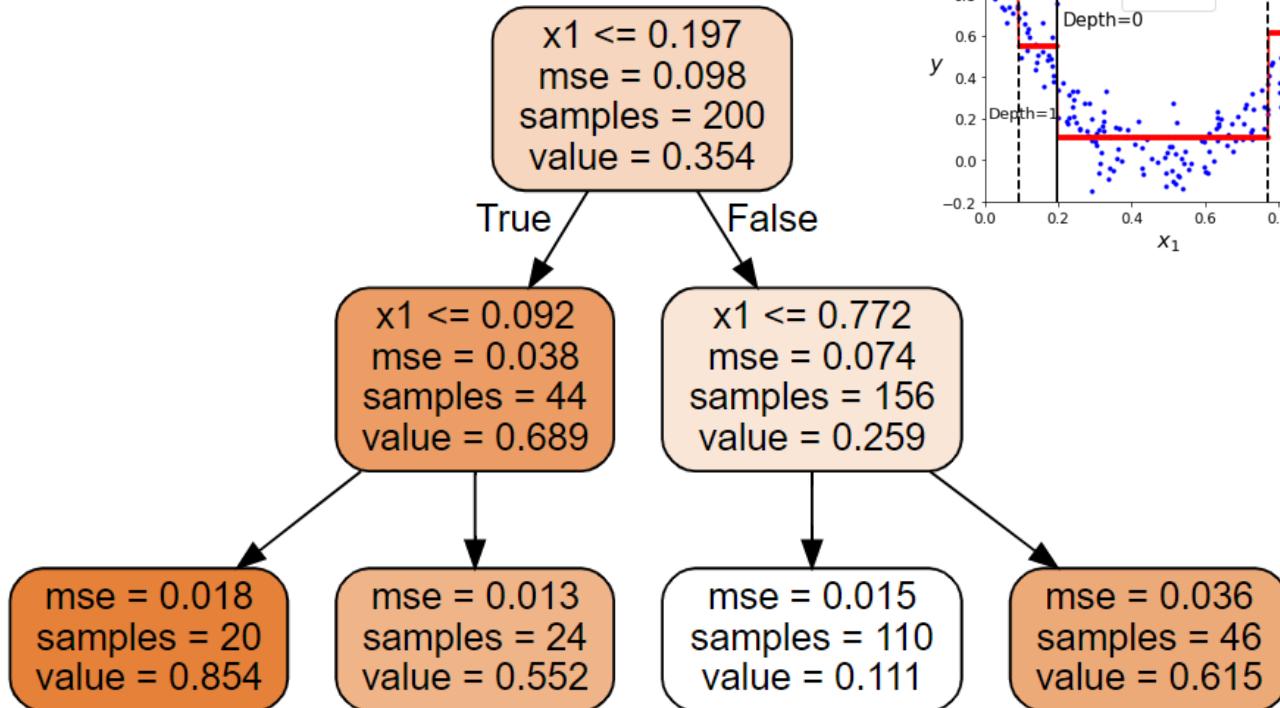
Algoritmo CART

CART: Classification and **Regression** Trees

- Sim, regressão também! Basta trocar a medida de impureza!



CART para regressão



Algoritmo CART: treinamento

- Testa todas as features e todos os thresholds
 - Basta testar os thresholds correspondentes aos valores das amostras
- Para cada combinação feature e threshold, avaliar a função de custo do CART:

$$J = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

O que acontece com J se o conjunto já for puro?

Algoritmo CART: treinamento

- Se a melhor combinação (feature, threshold) efetivamente melhora a função de custo, dividir o conjunto de pontos de treinamento.
- Repetir recursivamente o algoritmo para cada partição

Algoritmo CART: predição

Para uma nova amostra:

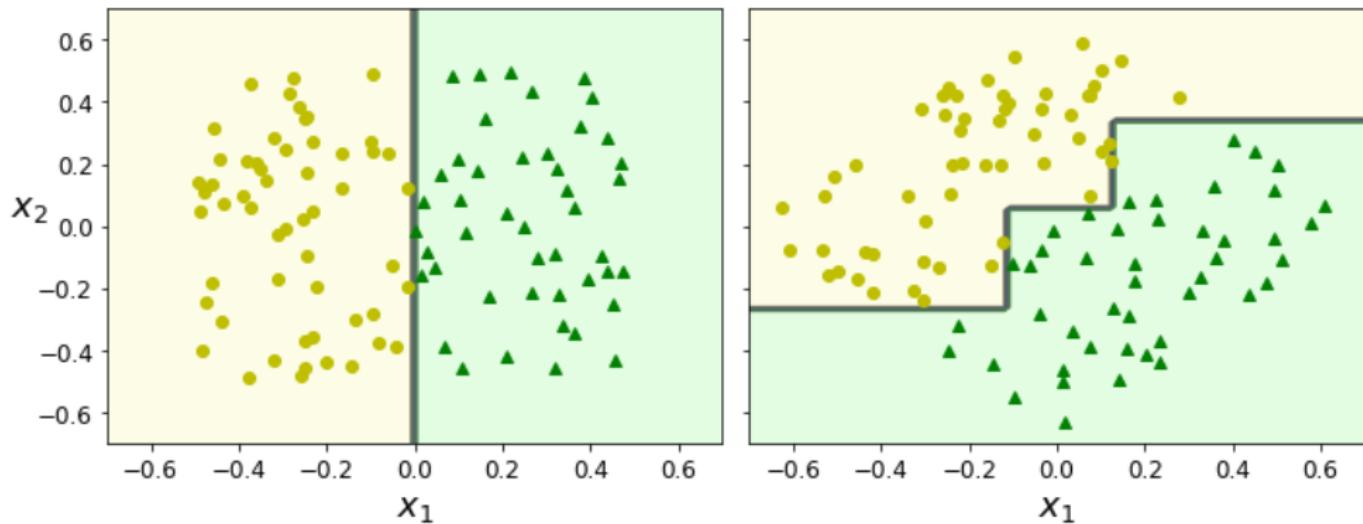
- Percorre a árvore até chegar na folha
- Retorna o valor da decisão na folha
 - Classificação: retorna a classe mais proeminente
 - Regressão: retorna o valor médio das amostras da folha

Vantagens da árvore de decisão

- Não precisa de *scaling* como a SVM
- Fácil de implementar
- Paralelizável
- **INTERPRETÁVEL**
 - Features mais importantes aparecem mais cedo na árvore!
 - Podemos saber a incerteza da predição olhando a impureza do nó de decisão

Desvantagens

- Preferência por fronteiras de decisão ortogonais e alinhadas com os eixos cartesianos
- Não é invariante à rotação





Insper