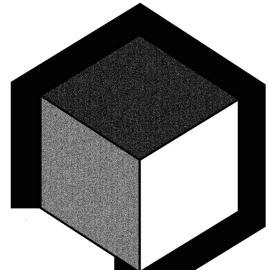


Introdução à Machine Learning

Utilizando a biblioteca scikit-learn

TÉO
ME WHY?



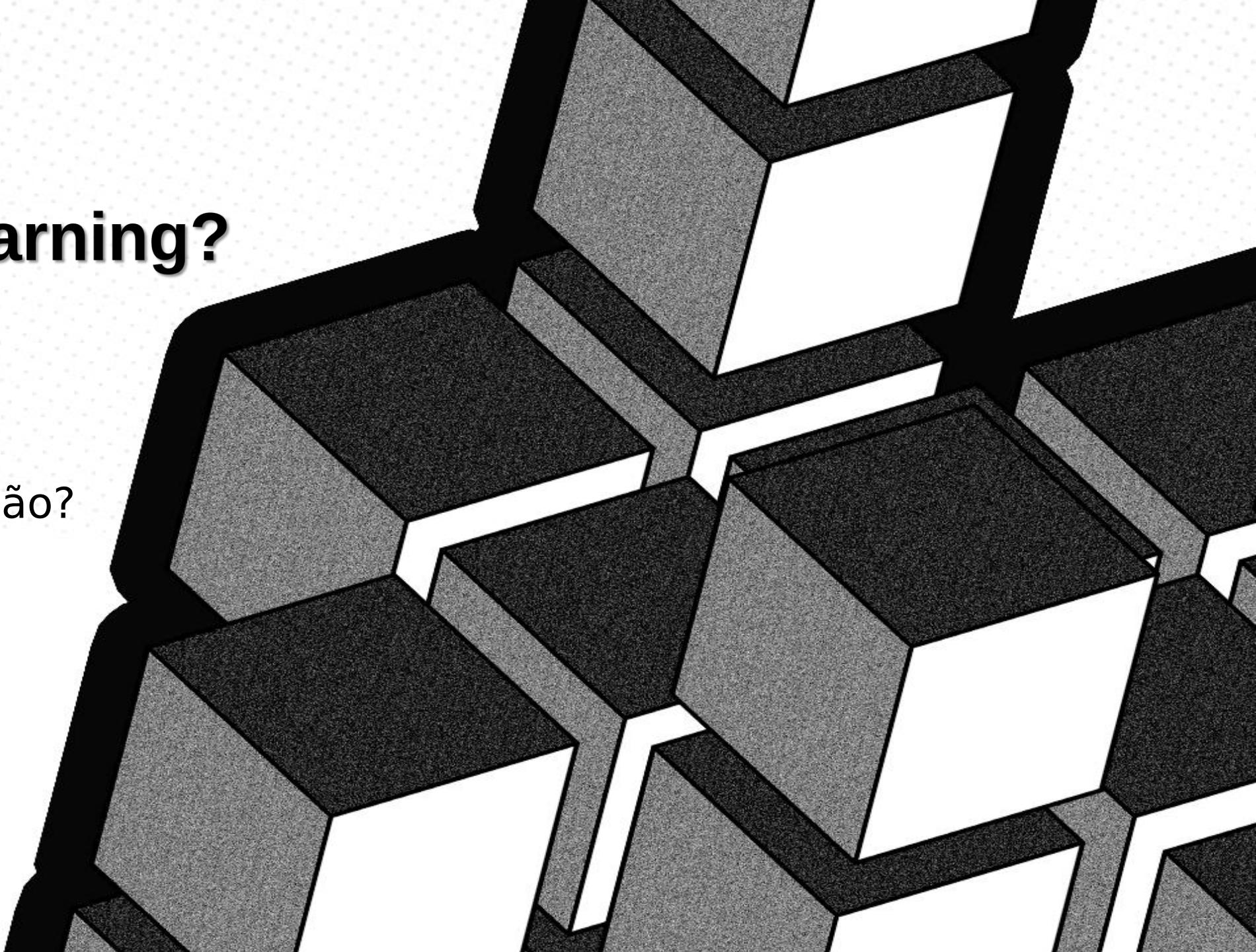
25 e 27 de Maio de 2021

Agenda

- O que é Machine Learning
- Tipos de Aprendizado
- Pré processamento
- Metricas de Ajuste

O que é Machine Learning?

Vamos fazer na mão?



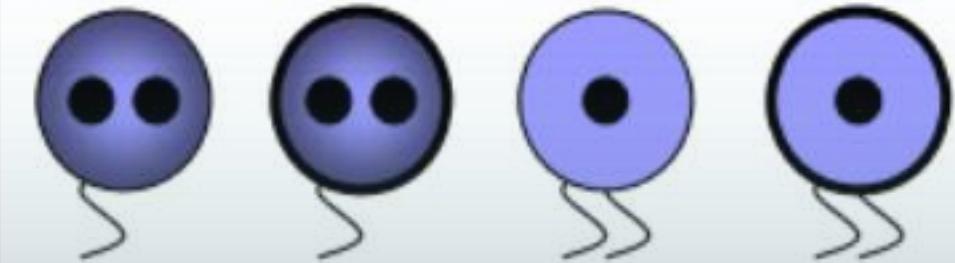
O que é Machine Learning?

O aprendizado automático (...) é um subcampo da Engenharia e da ciência da computação que evoluiu do estudo de **reconhecimento de padrões** e da teoria do aprendizado computacional em **inteligência artificial** [1]. Em 1959, Arthur Samuel definiu aprendizado de máquina como o "campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados"[2]

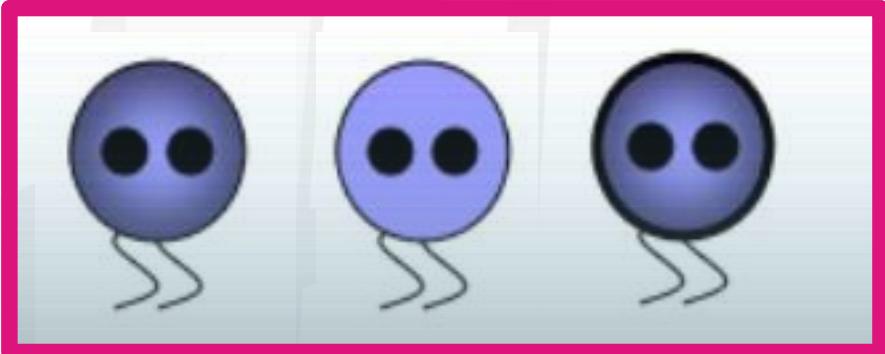
[1] <https://www.britannica.com/technology/machine-learning>

[2] https://books.google.com.br/books?id=Dn-Gdoh66sgC&pg=PA89&redir_esc=y#v=onepage&q&f=false

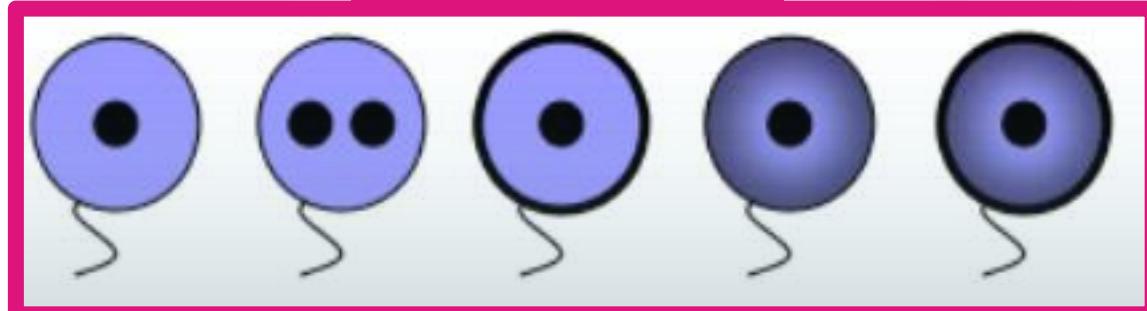
Saudável



Burpona



Lethargia



Quais atributos temos?

Quantidade de caudas

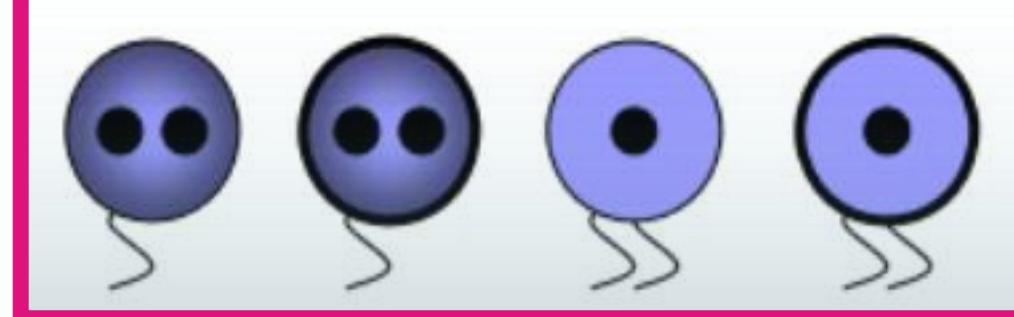
Quantidade de núcleos

Tipo de membrana

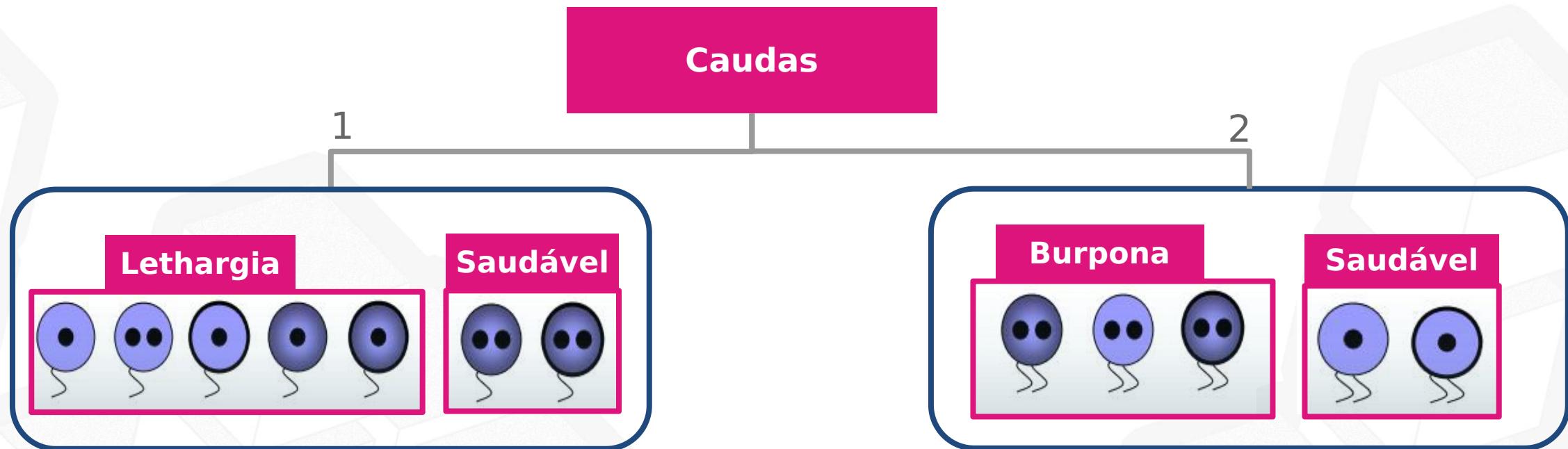
Cor

O que me difere
das demais?

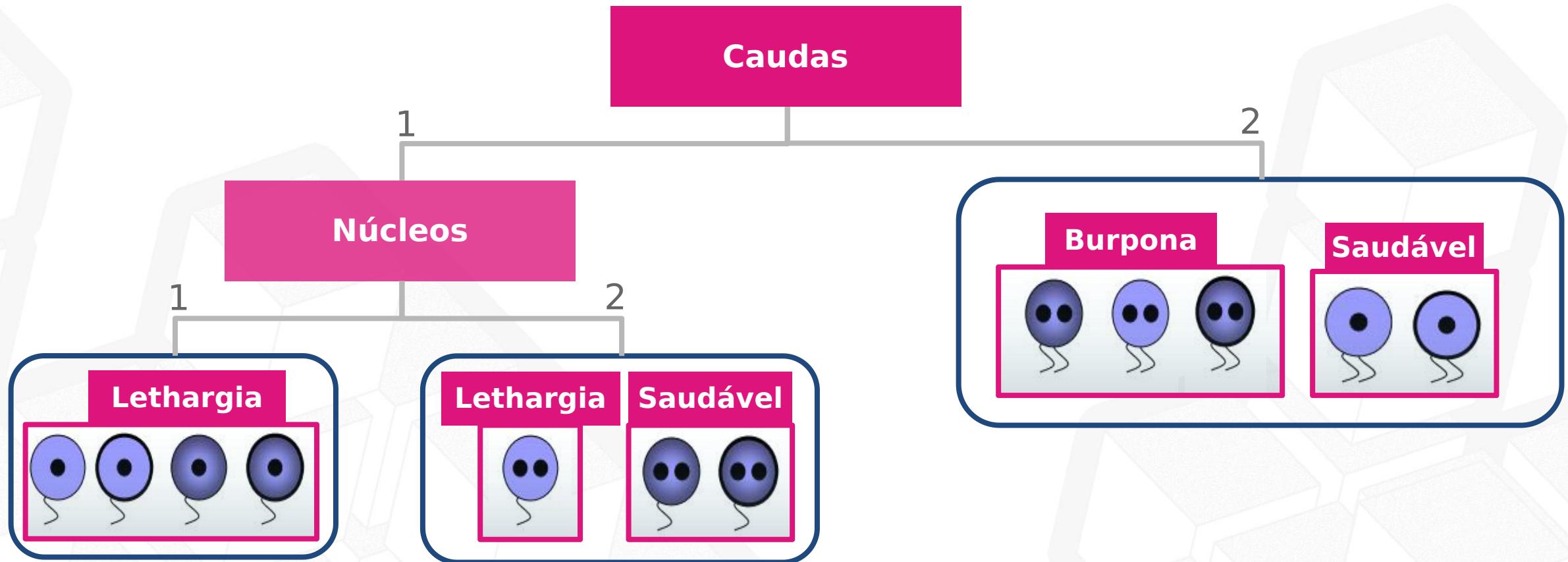
Saudável



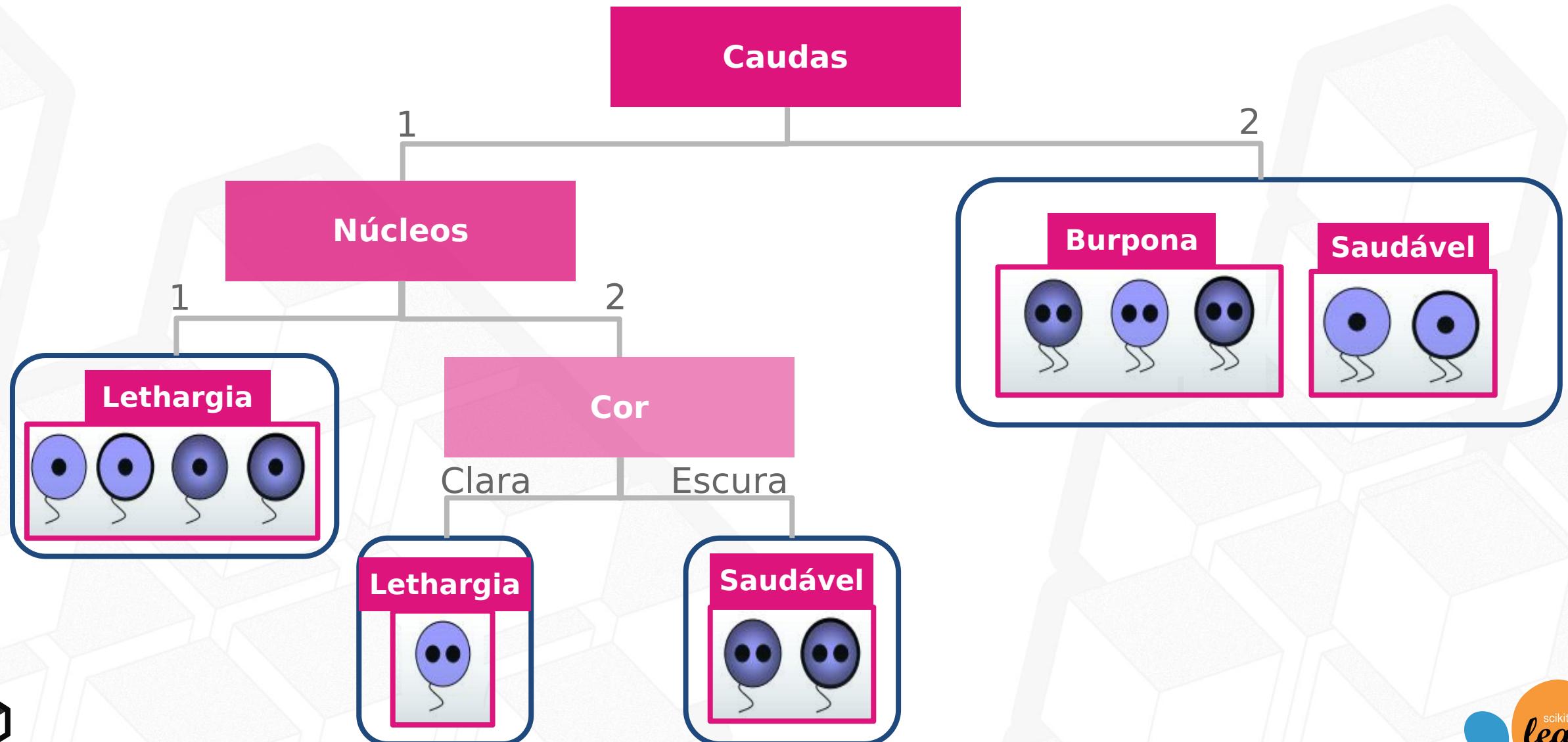
Criando regras



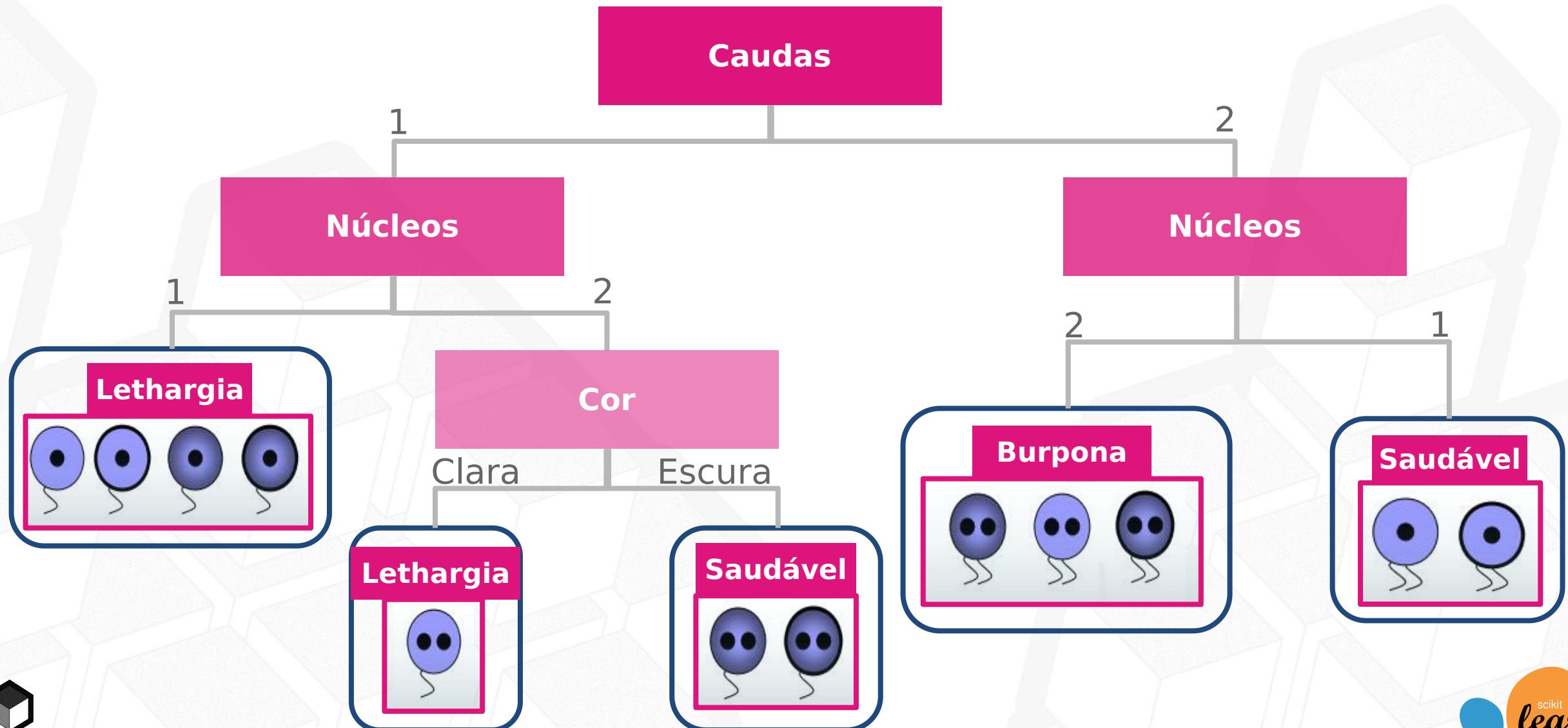
Criando regras



Criando regras



Criando regras



Tabela

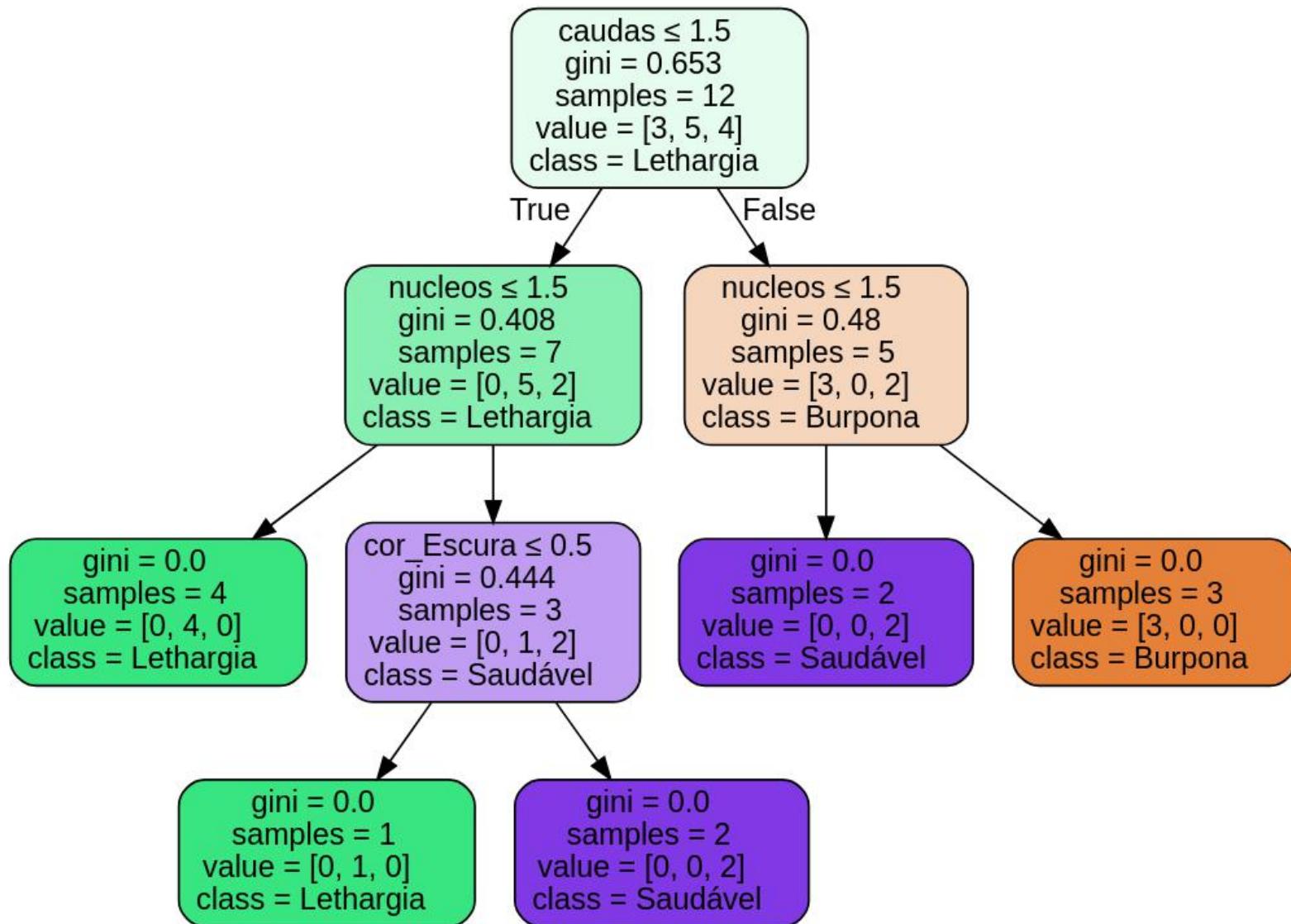
id	Núcleos	Caudas	Cor	Membrana	Classe
1	1	1	Clara	Fina	Lethargia
2	2	1	Clara	Fina	Lethargia
3	1	1	Clara	Grossa	Lethargia
4	1	1	Escura	Fina	Lethargia
5	1	1	Clara	Grossa	Burpona
6	2	2	Escura	Fina	Buporna
7	2	2	Escura	Fina	Burpona
8	2	2	Escura	Grossa	Burpona
9	2	1	Escura	Fina	Saudável
10	2	1	Escura	Grossa	Saudável
11	1	2	Clara	Fina	Saudável
12	1	2	Clara	Grossa	Saudável

Atributos

Alvo

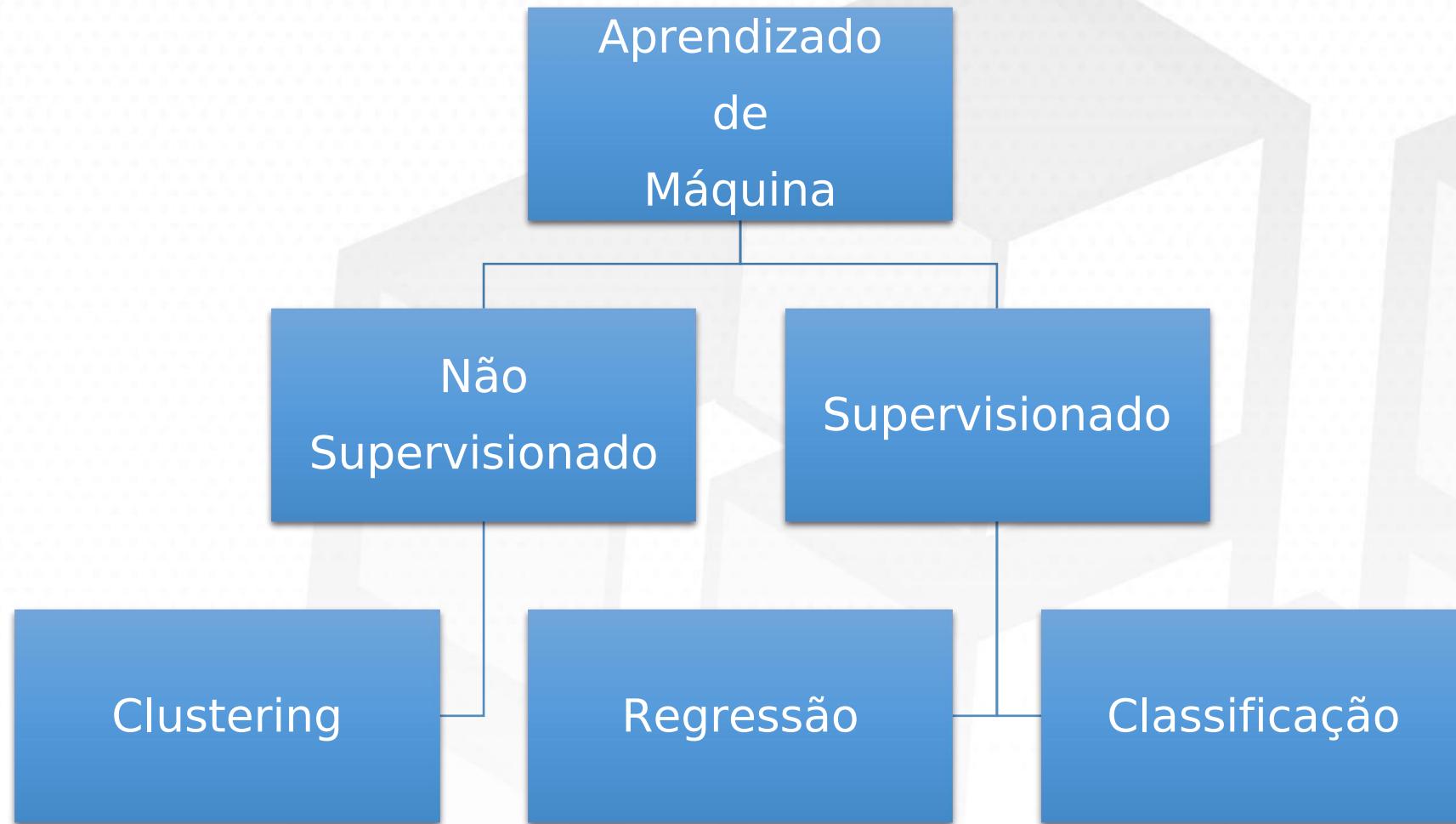


Árvore no Python



The background of the slide features a complex arrangement of black and white cubes. Some cubes are solid black, while others are white with black outlines. They are stacked and interconnected in a way that creates a sense of depth and perspective, resembling a 3D geometric puzzle.

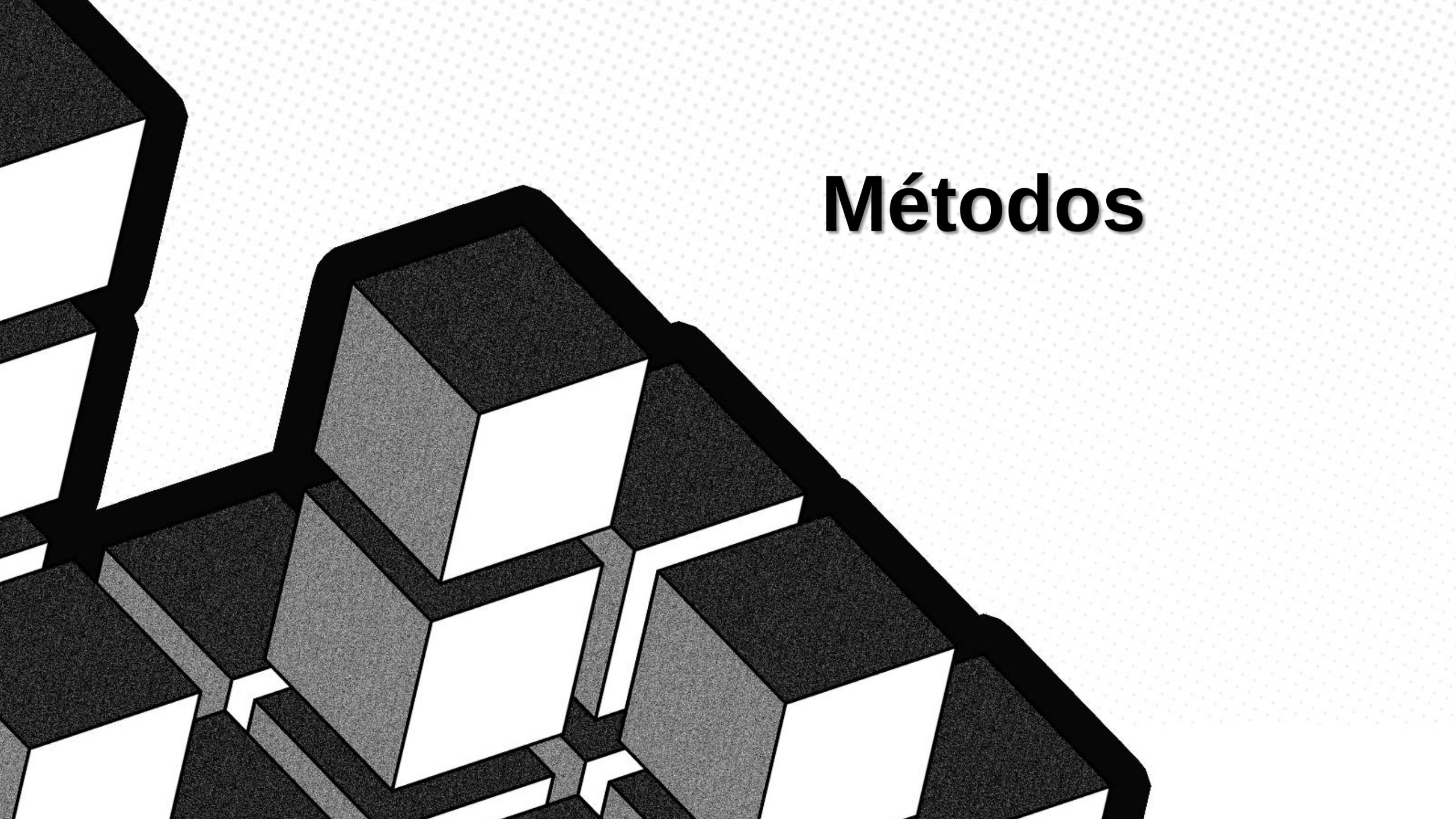
**Quais tipos de
aprendizado temos?**



Regressão

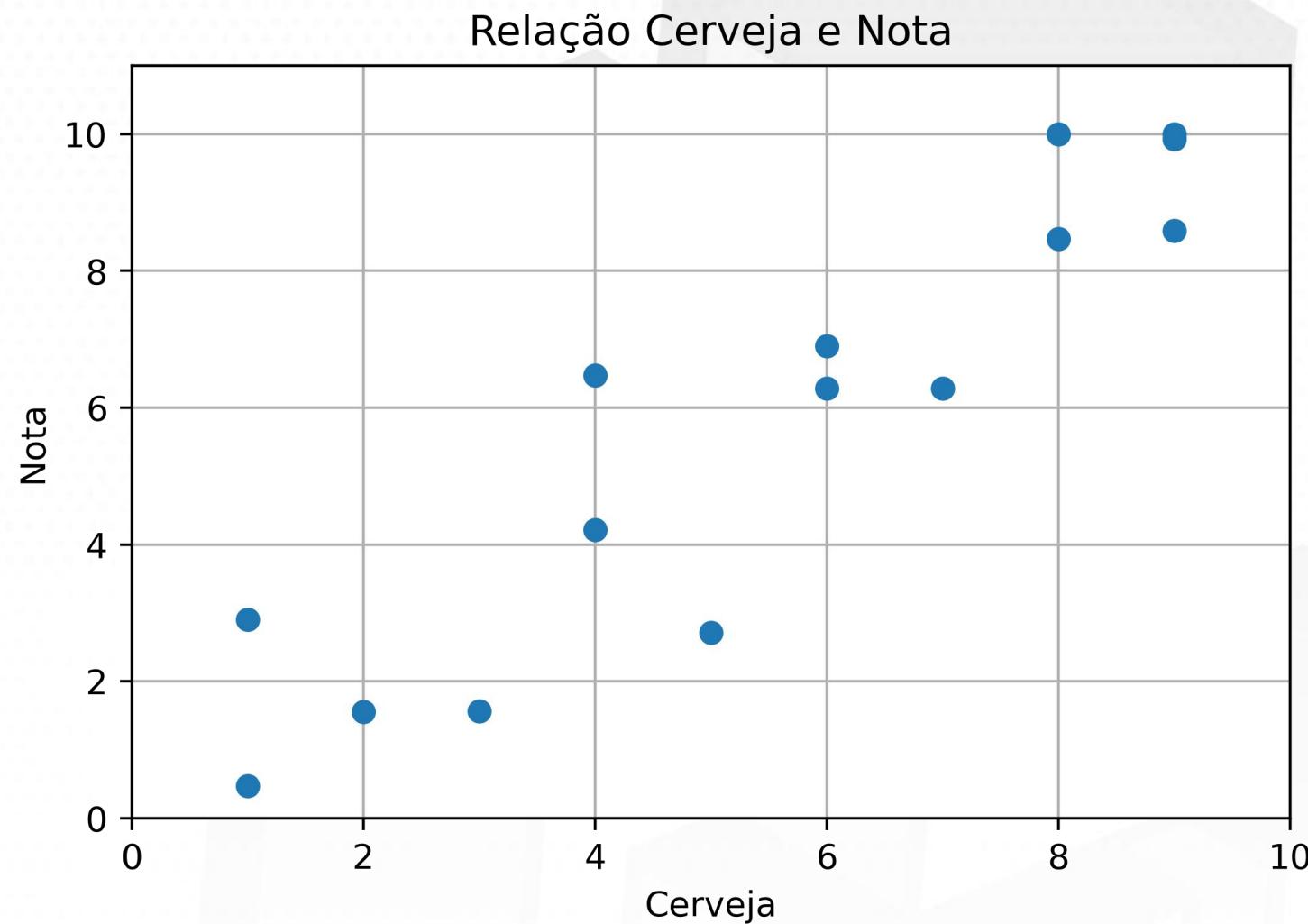
Problemas de regressão são voltados à estimativa alvo ('target'), sendo este um número, valor. Por exemplo:

- Quantidade de vendas
- Receita presumida
- Valor de crédito
- Precificação de imóvel
- Volume de chuva

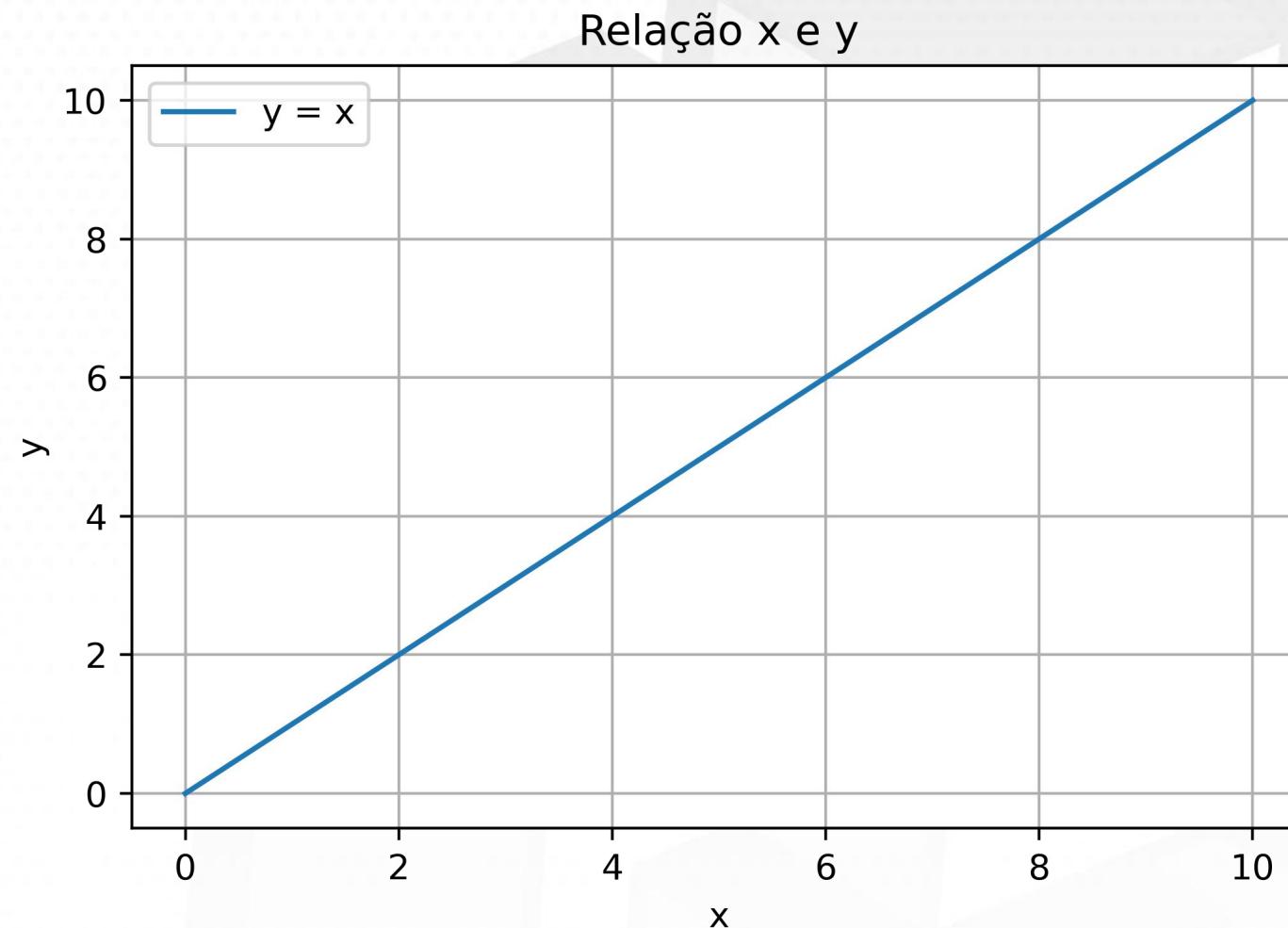
The background of the slide features a complex arrangement of black and white cubes. Some cubes are solid black, while others are white with black outlines. They are stacked and interconnected in a way that creates a sense of depth and perspective, resembling a 3D geometric puzzle. The overall aesthetic is minimalist and modern.

Métodos

Regressão Linear

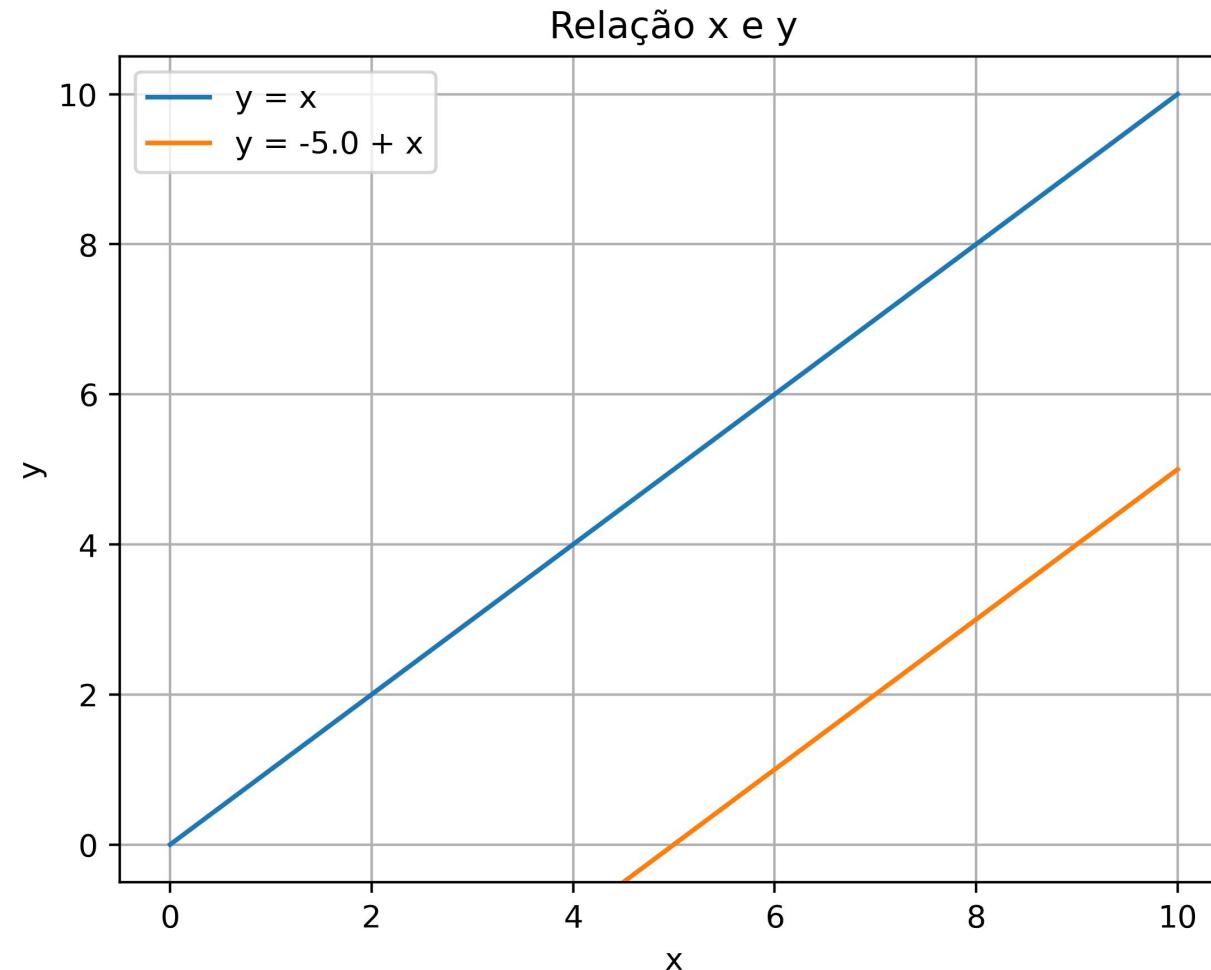


Regressão Linear



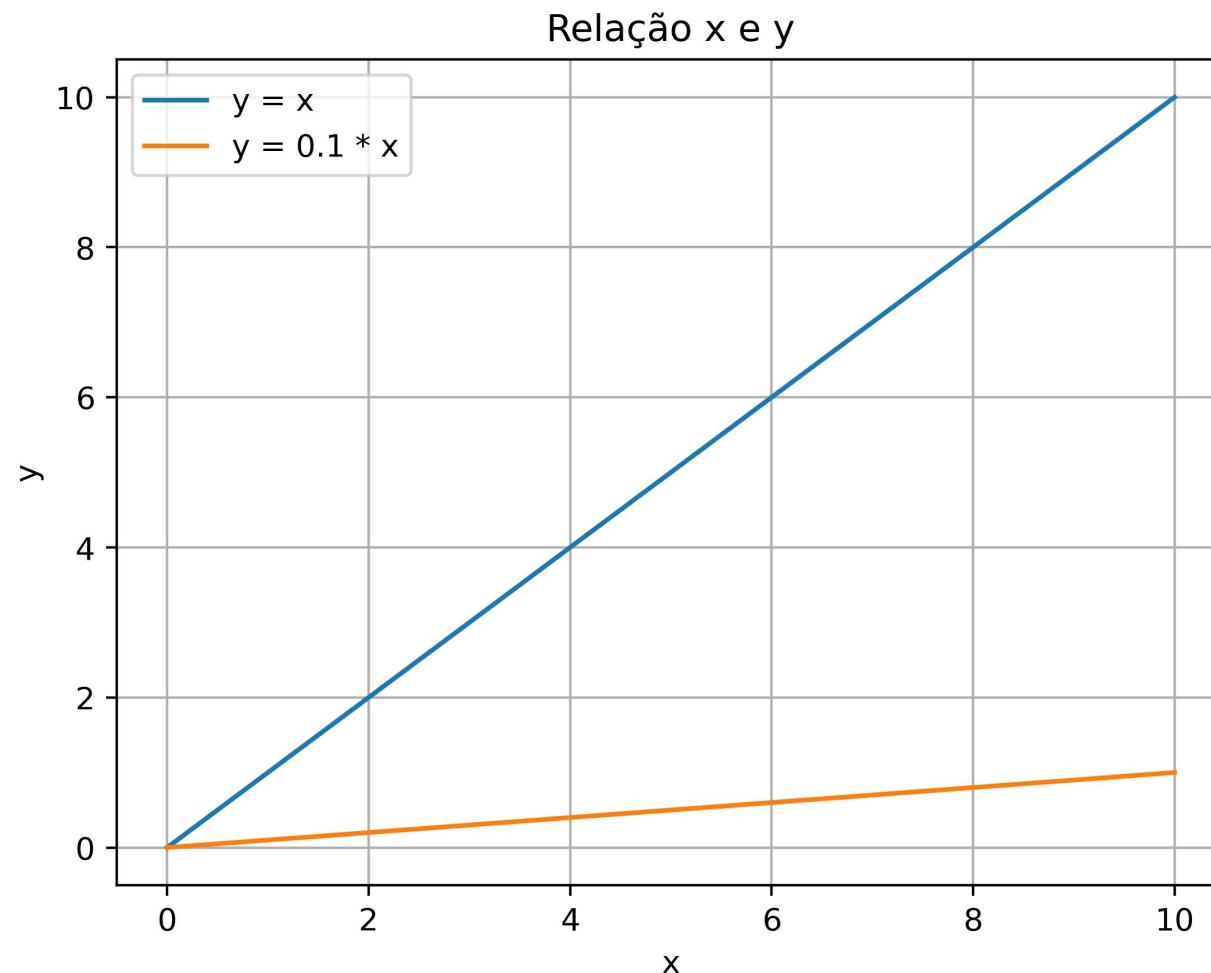
Regressão Linear

$$y = a + x$$



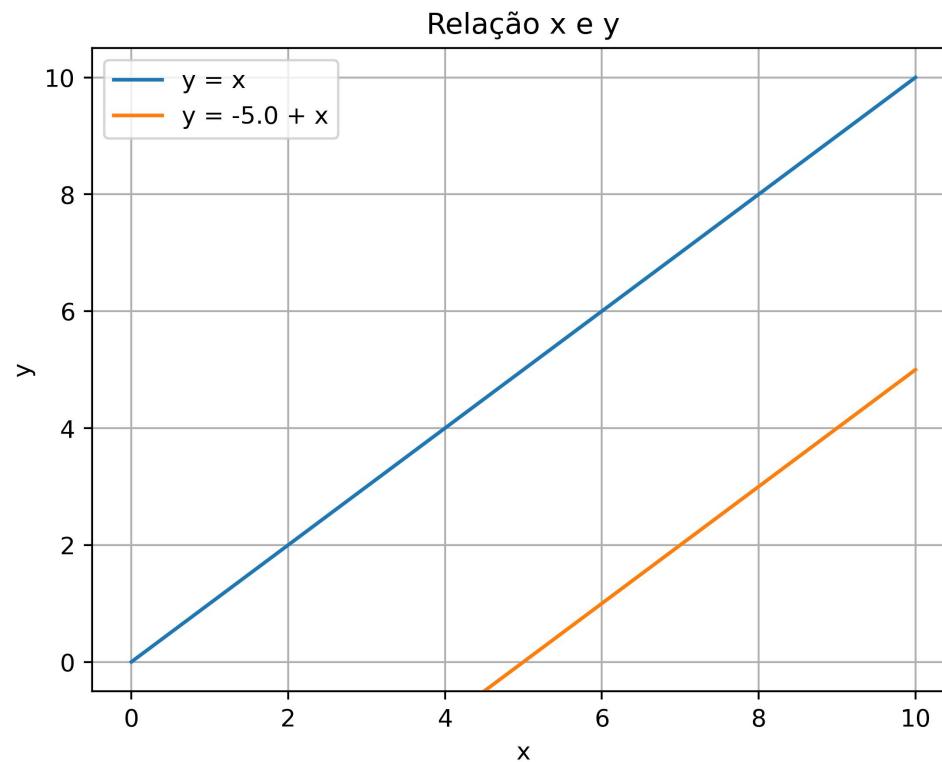
Regressão Linear

$$y = b \cdot x$$

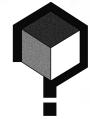
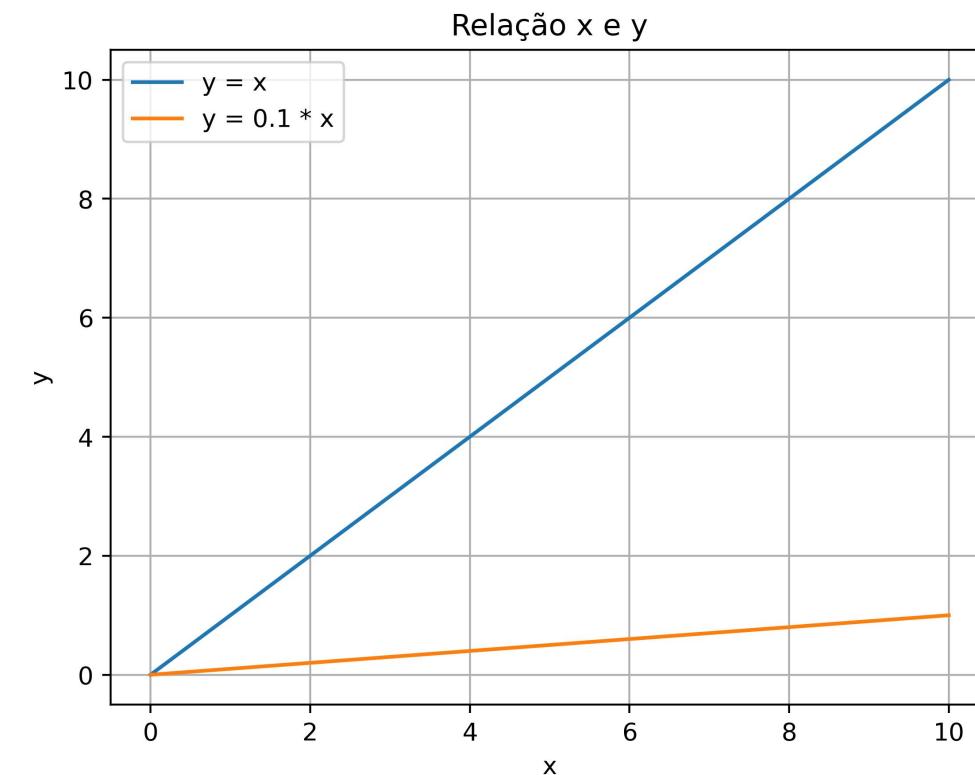


Regressão Linear

$$y = a + x$$

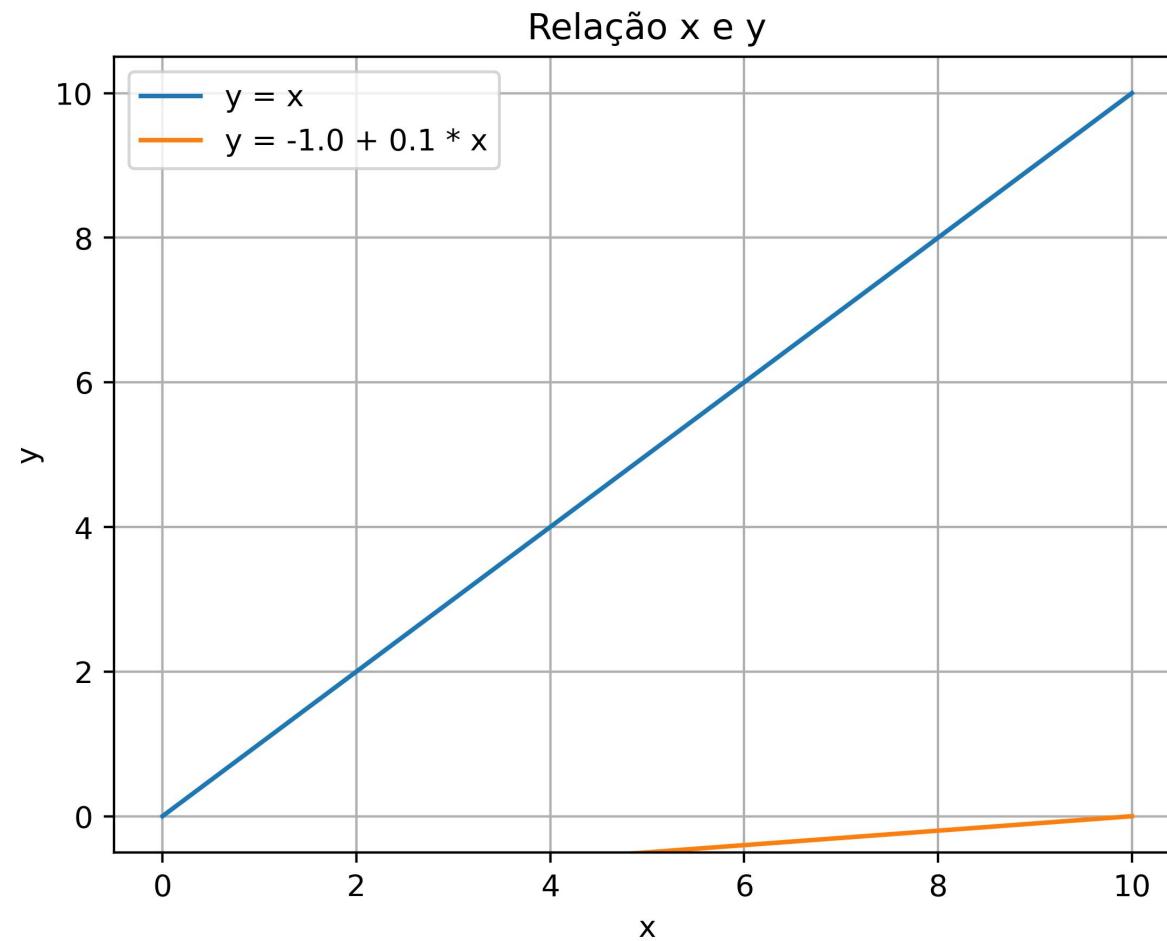


$$y = b \cdot x$$

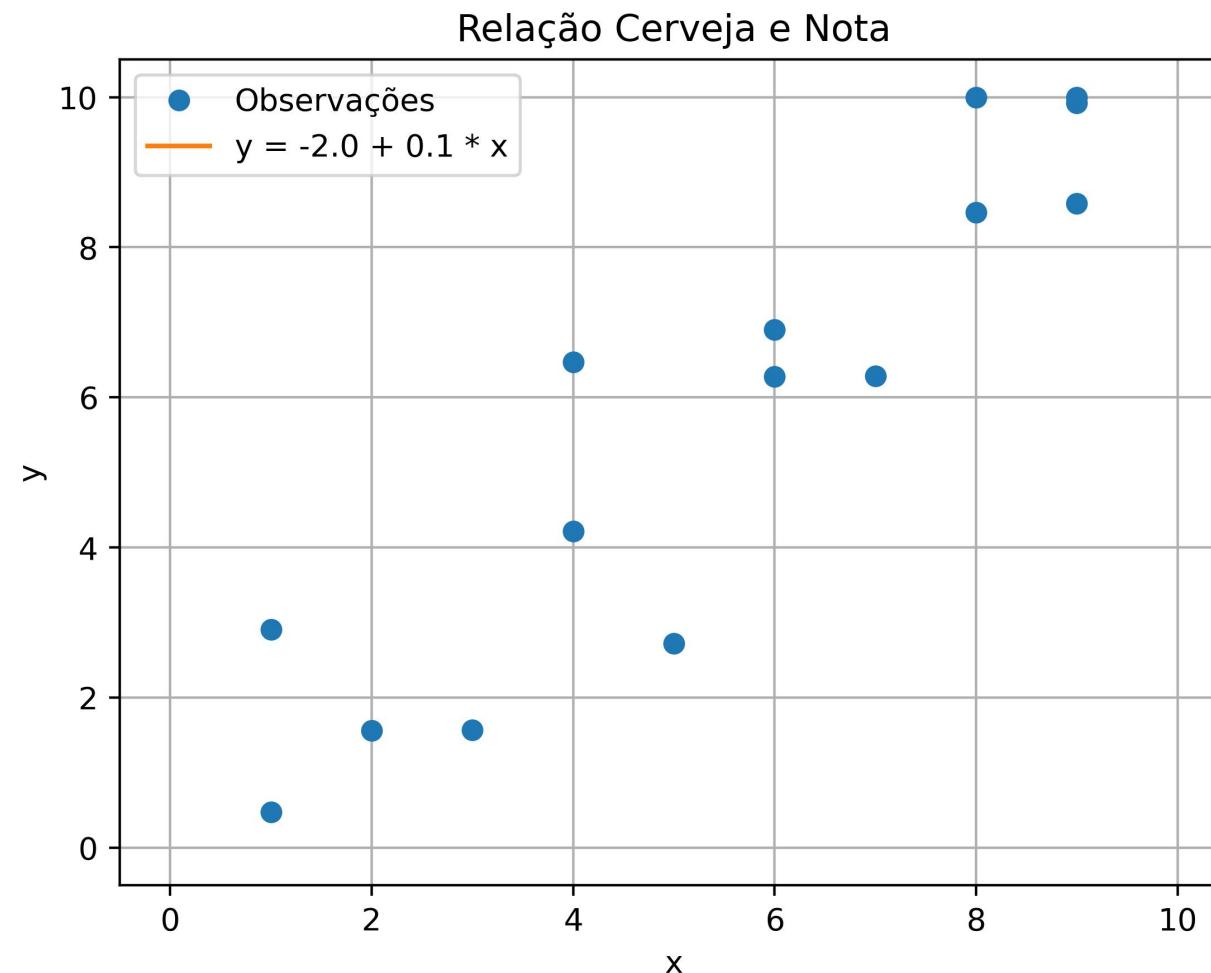


Regressão Linear

$$y = a + b \cdot x$$



Regressão Linear



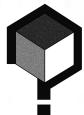
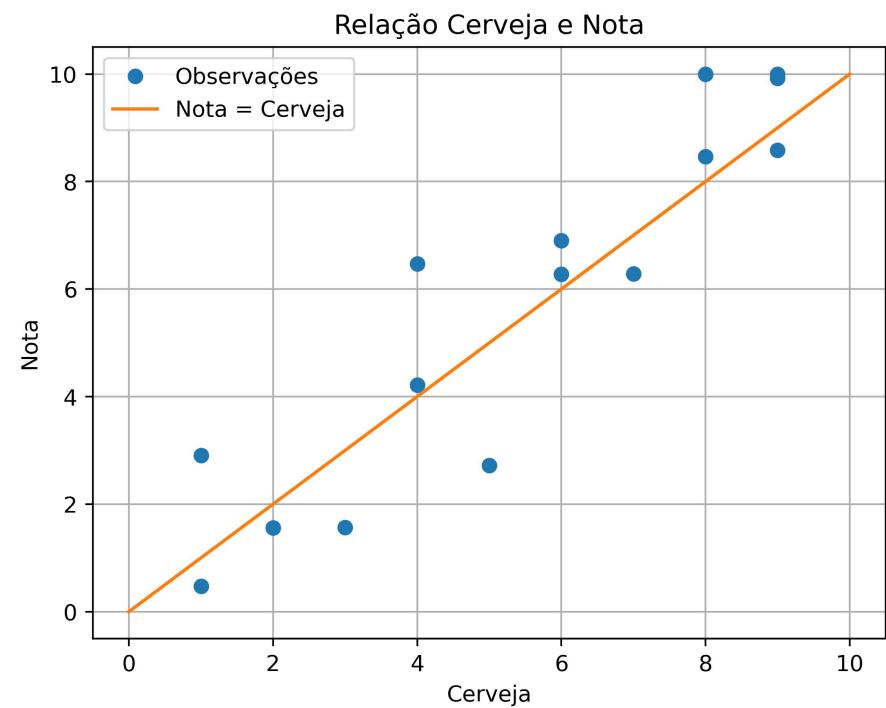
Regressão Linear

$$\hat{y} = a + b \cdot x \Leftrightarrow \text{Nota} = a + b \cdot \text{Cerveja}$$

$$\text{Erro} = y - \hat{y}$$

$$\text{Erro Quadrático} = (y - \hat{y})^2$$

$$\text{Soma do Erro Quadrático} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



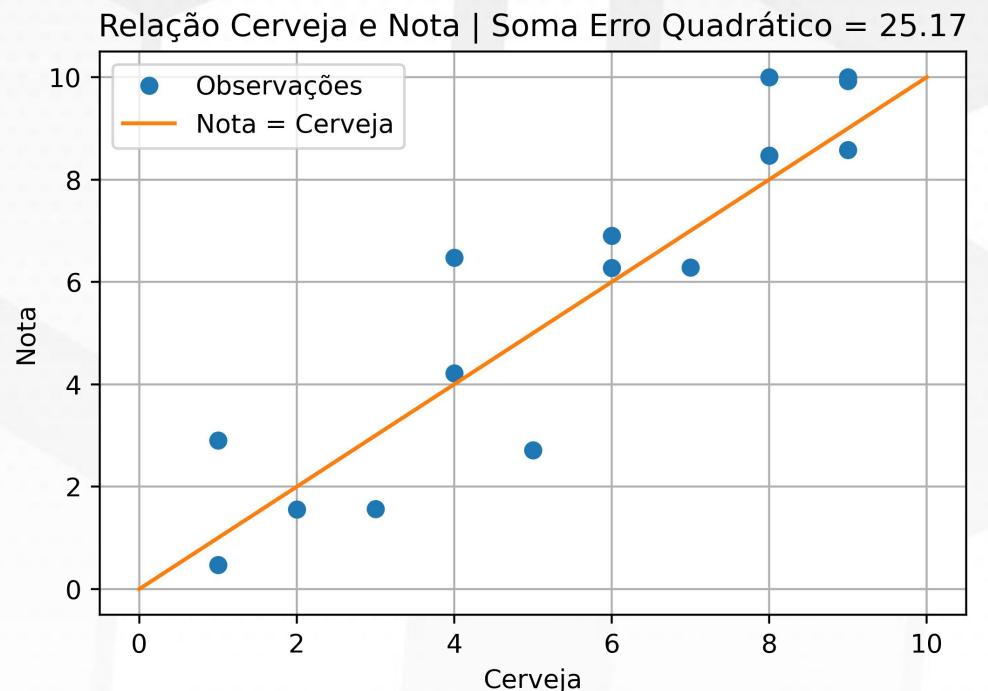
Regressão Linear

$$\hat{y} = a + b \cdot x \Leftrightarrow \text{Nota} = a + b \cdot \text{Cerveja}$$

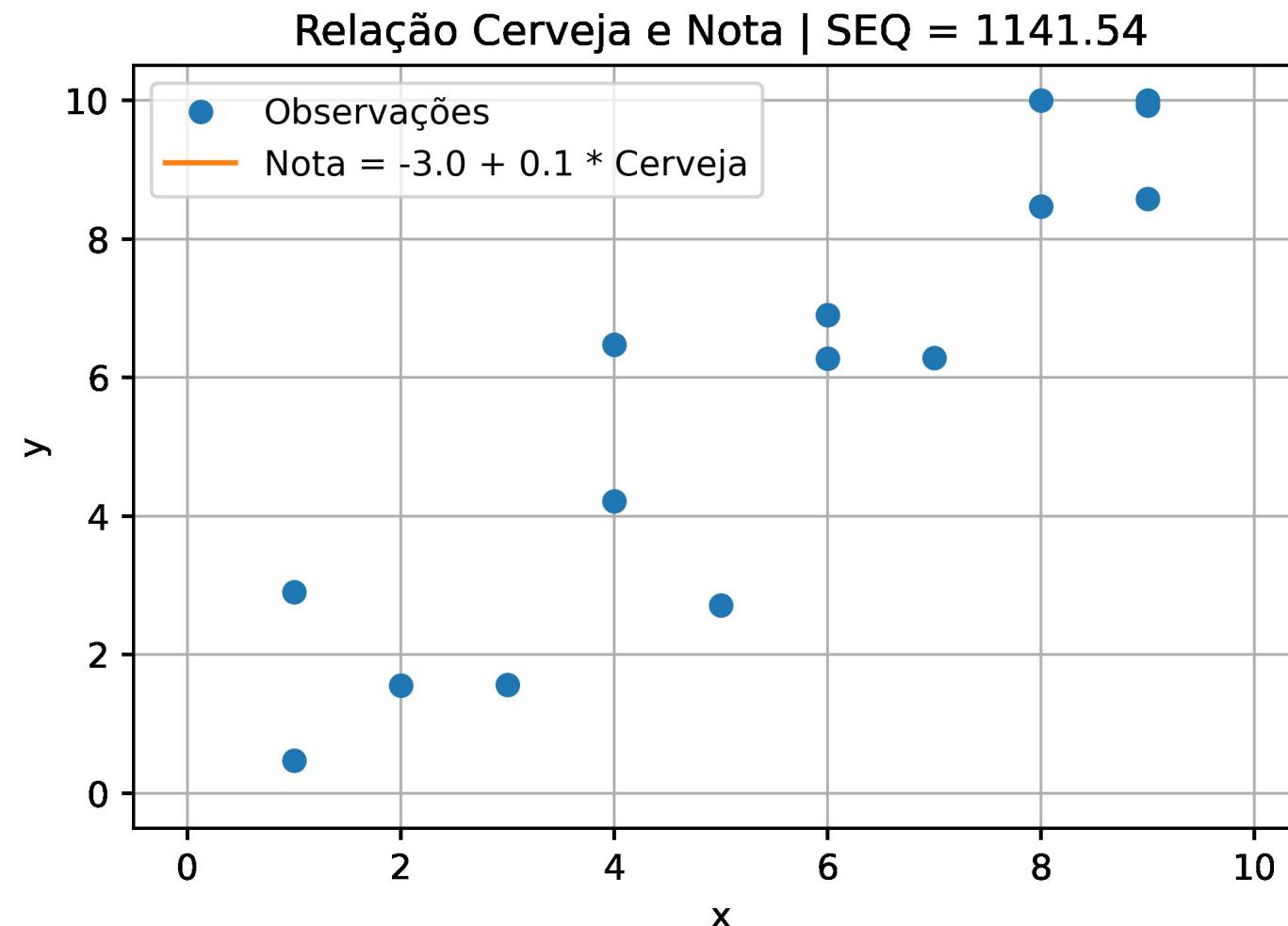
$$\text{Erro} = y - \hat{y}$$

$$\text{Erro Quadrático} = (y - \hat{y})^2$$

$$\text{Soma do Erro Quadrático} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



Regressão Linear



Regressão Linear

Bora minimizar os erros quadráticos

Mínimos
Quadrados



Regressão Linear

Bora minimizar os erros quadráticos?

$$\hat{y} = a + b \cdot x \Leftrightarrow \text{Nota} = a + b \cdot \text{Cerveja}$$

$$\text{Erro} = y - \hat{y}$$

$$\text{Erro Quadrático} = (y - \hat{y})^2$$

$$\text{Soma do Erro Quadrático} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{Soma do Erro Quadrático} = \sum_{i=1}^n (y_i - (a + b x_i))^2$$



Regressão Linear

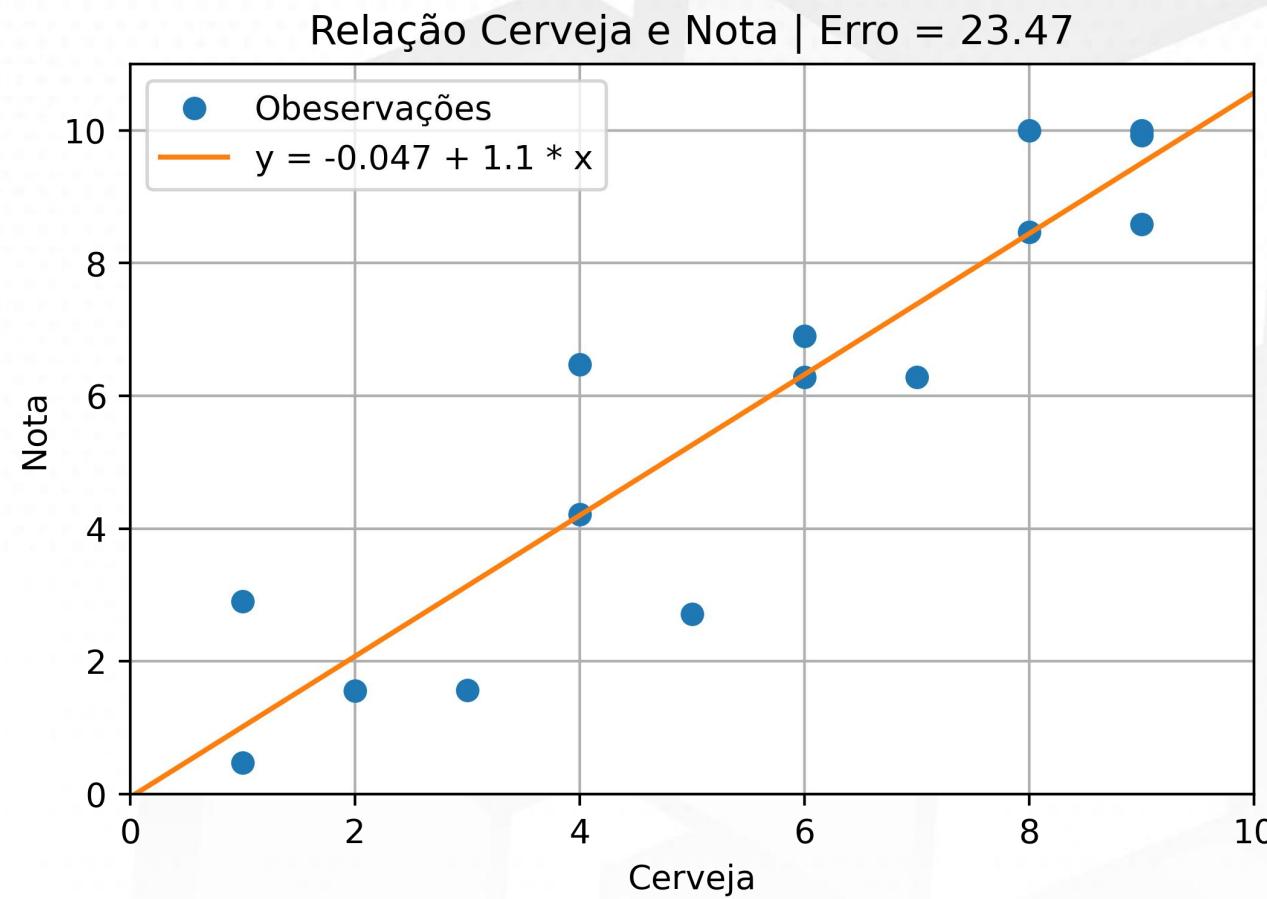
Bora minimizar os erros quadráticos?

$$\frac{\partial}{\partial a} \sum_{i=1}^n (y_i - (a + bx_i))^2 = 0$$

$$\frac{\partial}{\partial b} \sum_{i=1}^n (y_i - (a + bx_i))^2 = 0$$



Regressão Linear



Regressão Linear

Mínimos Quadrados é o único método de ajustada reta?

Método de Máxima Verossimilhança (Inferência Clássica)

Método Bayesiano (Inferência Bayesiana)



Regressão Linear

$$y = a + bx$$

$$y = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x_1$$

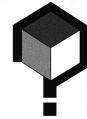
Regressão Linear Múltipla

E se tivermos afim de utilizar mais variáveis?

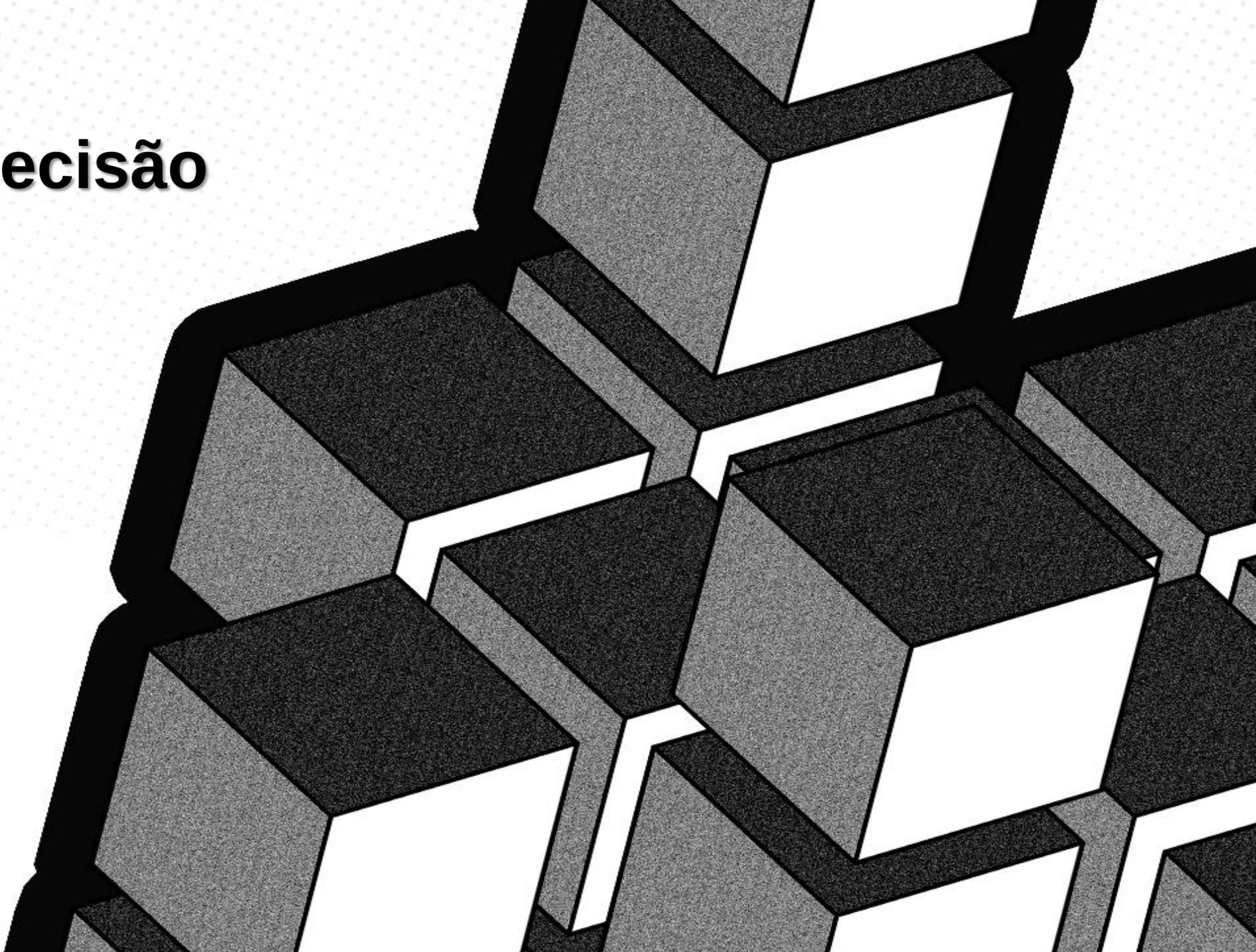
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p$$

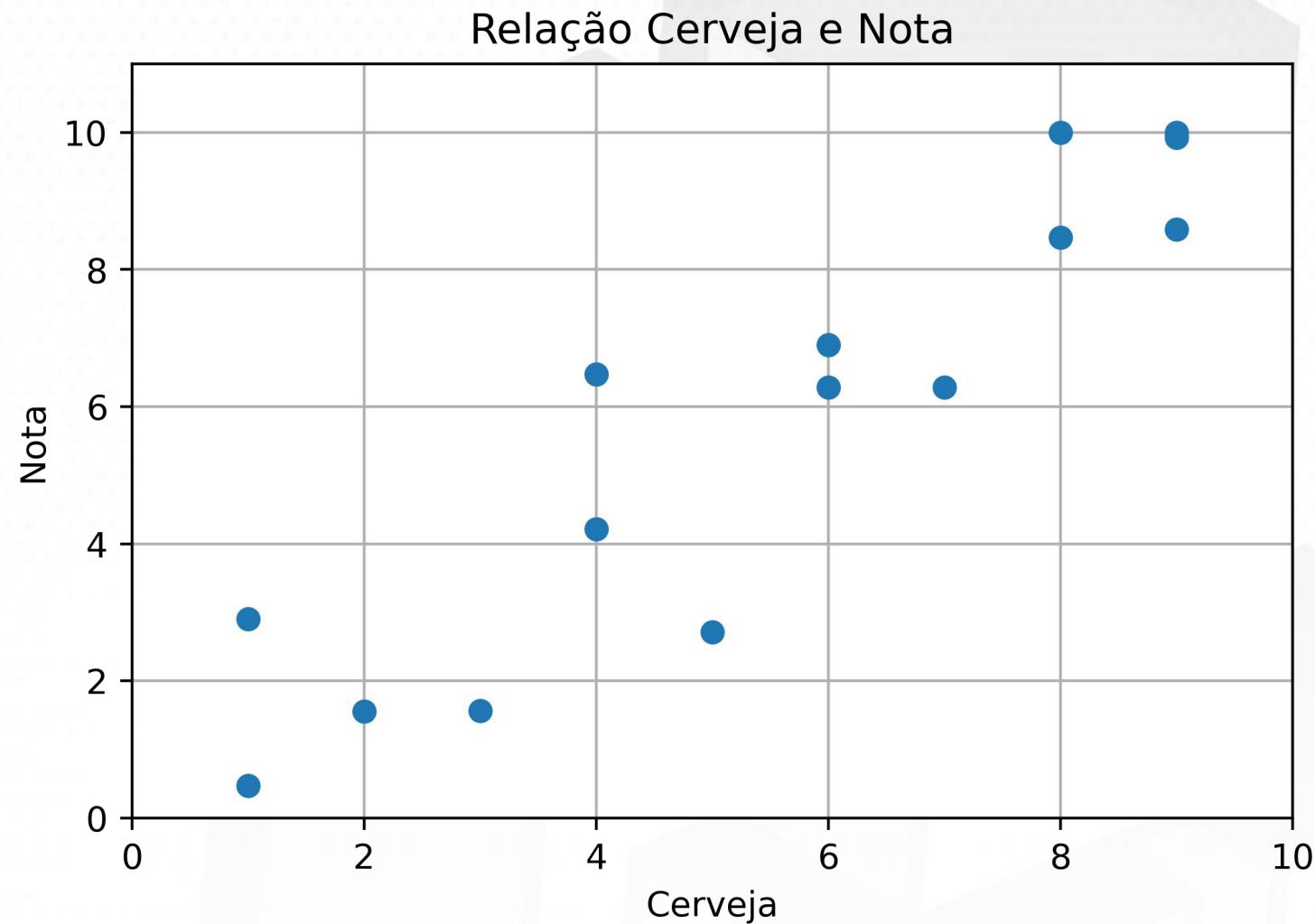
$$y = \beta_0 + \sum_{i=1}^p \beta_i x_i$$



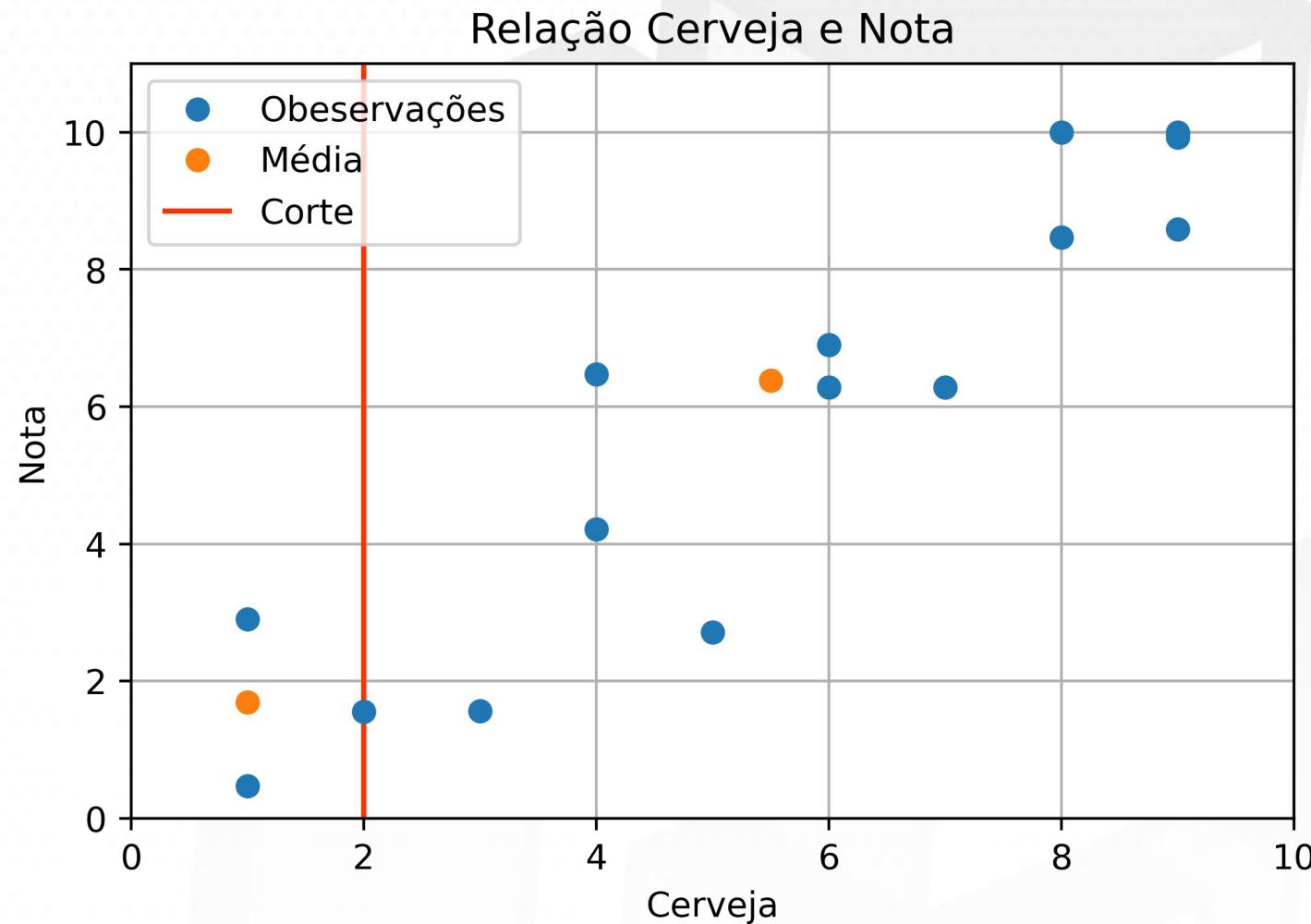
Árvore de Decisão



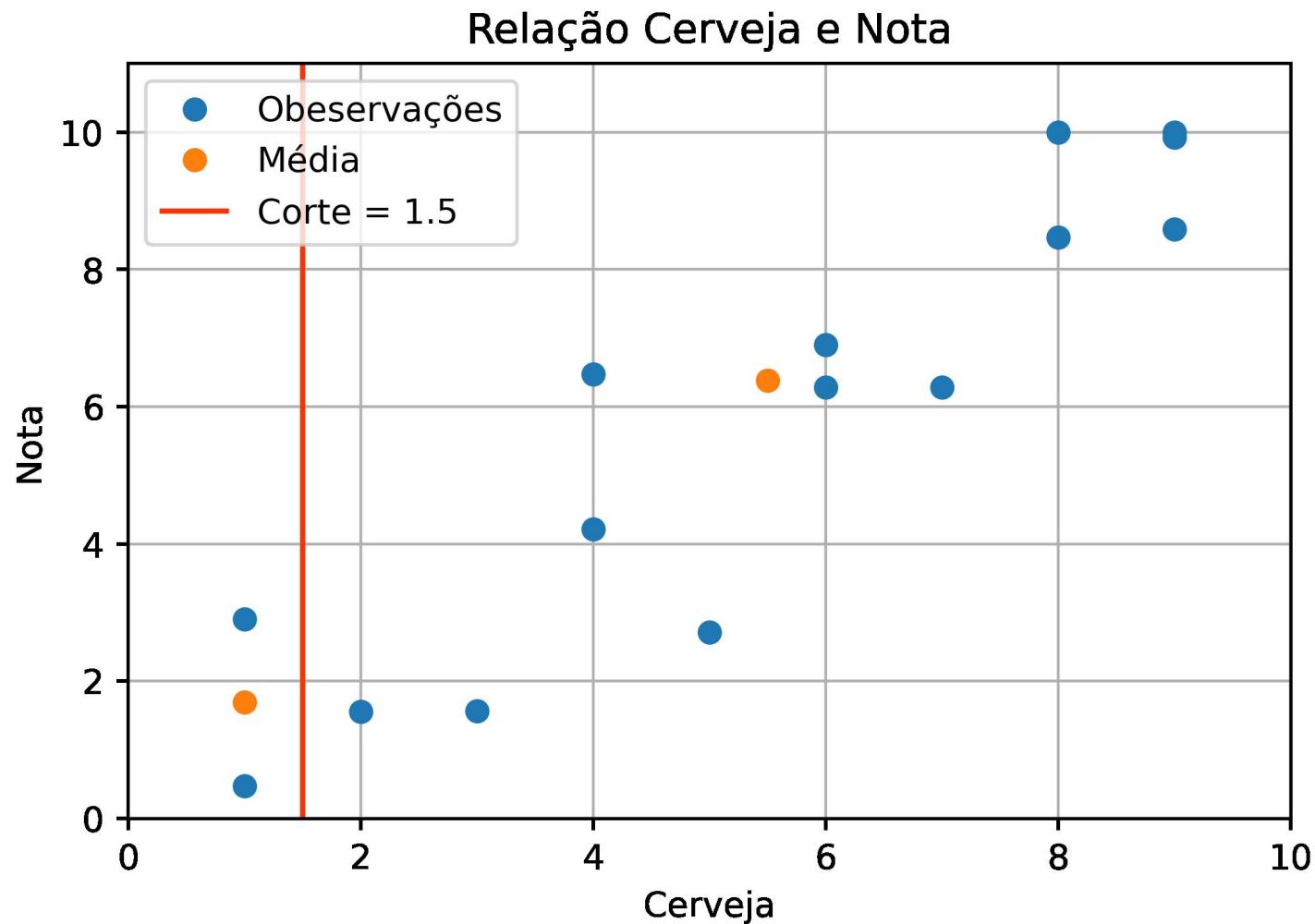
Voltando à Árvore de Decisão



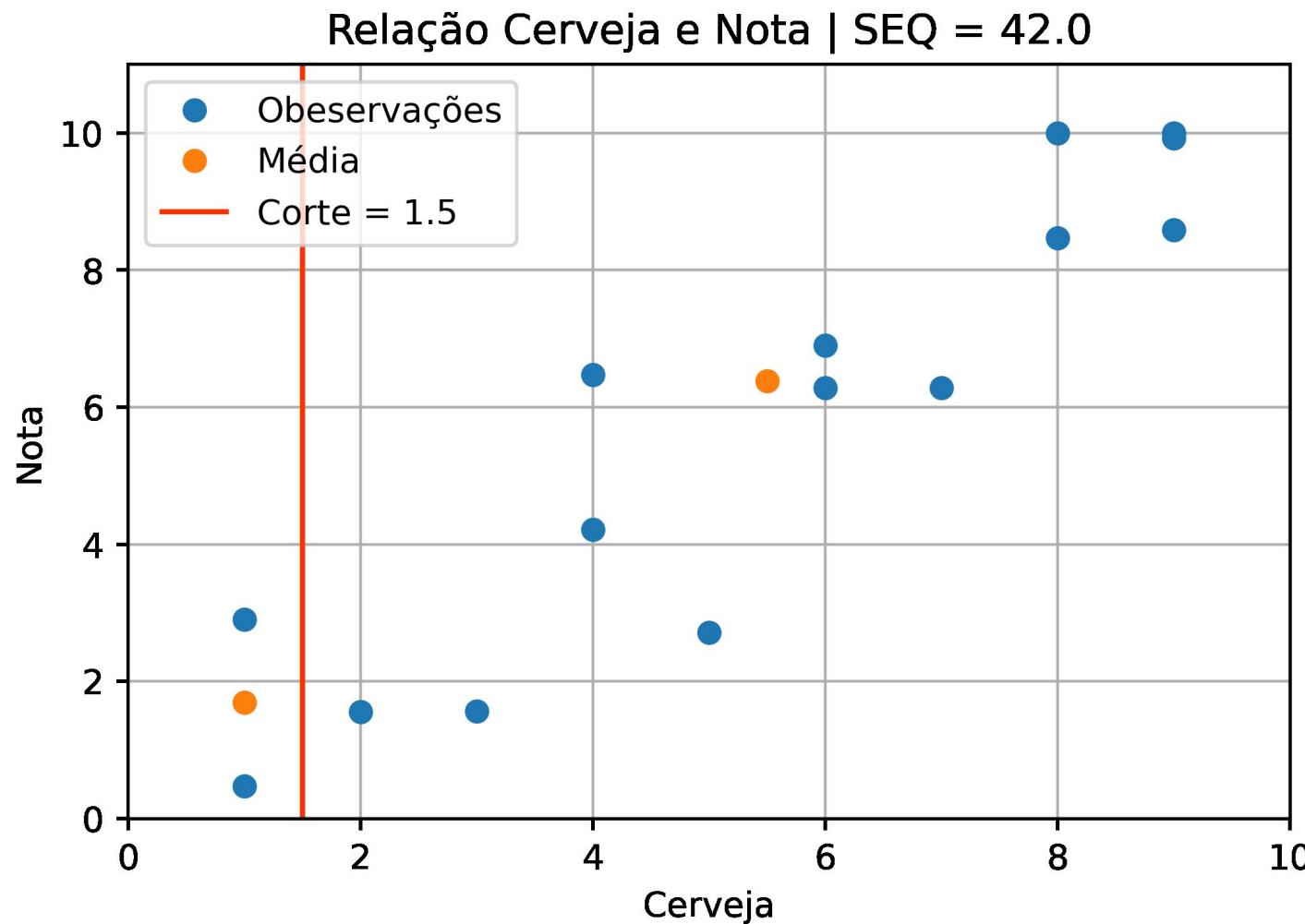
Voltando à Árvore de Decisão



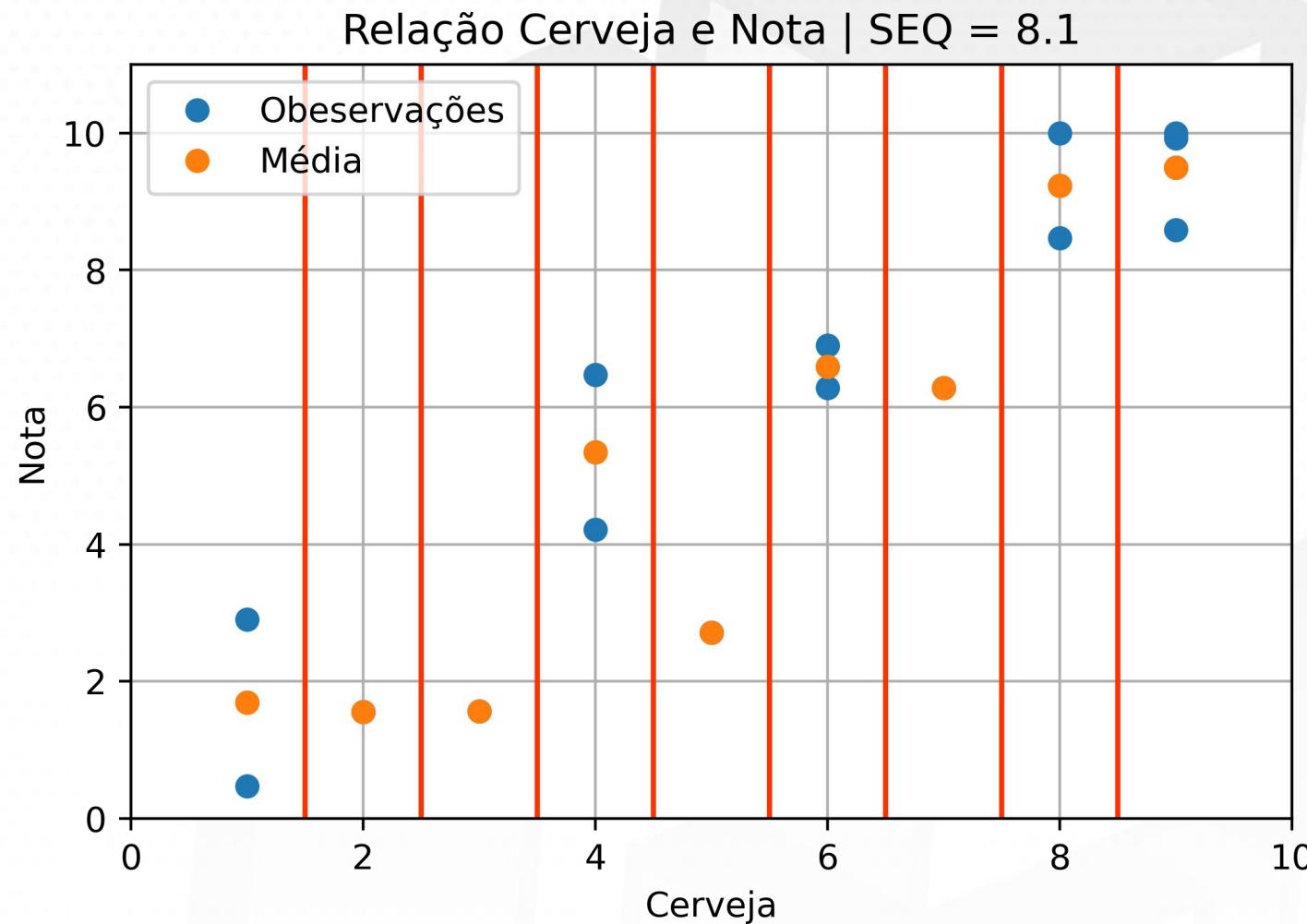
Voltando à Árvore de Decisão



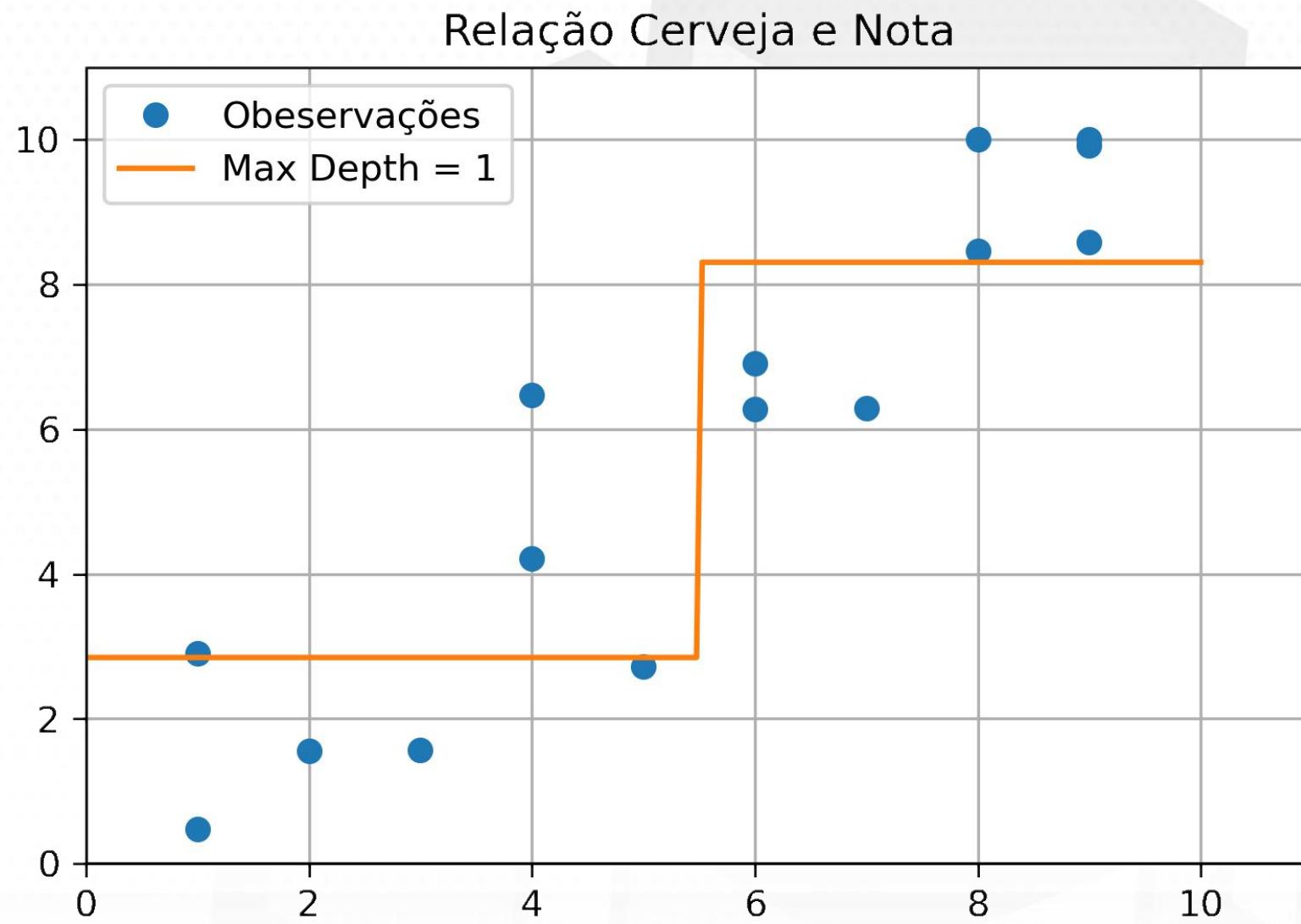
Árvore de Decisão



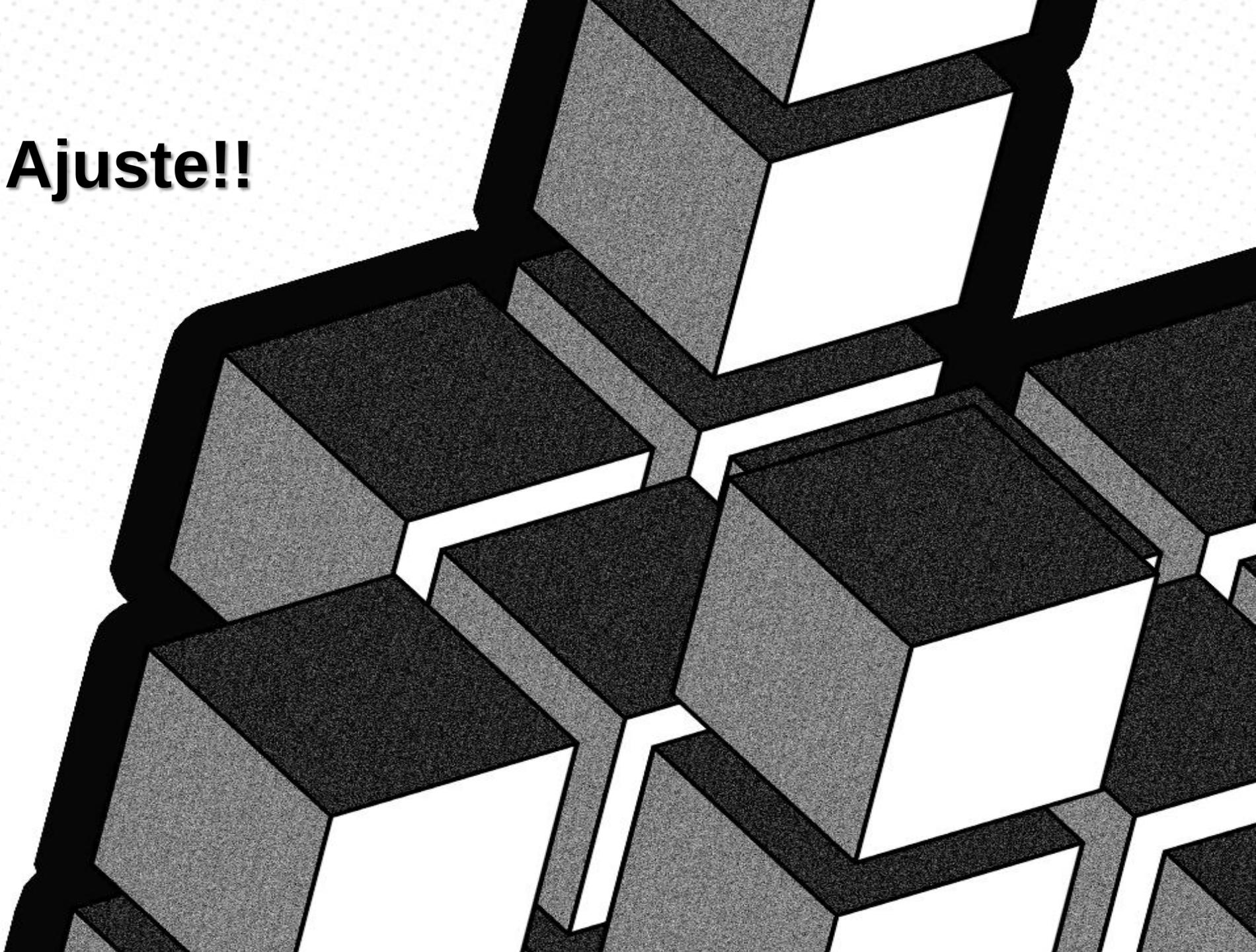
Árvore de Decisão



Árvore de Decisão



Métricas de Ajuste!!



Métricas de Ajuste

https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

Erro médio absoluto

$$MAE = \frac{\sum_{i=1}^n \|y - \hat{y}\|}{n}$$

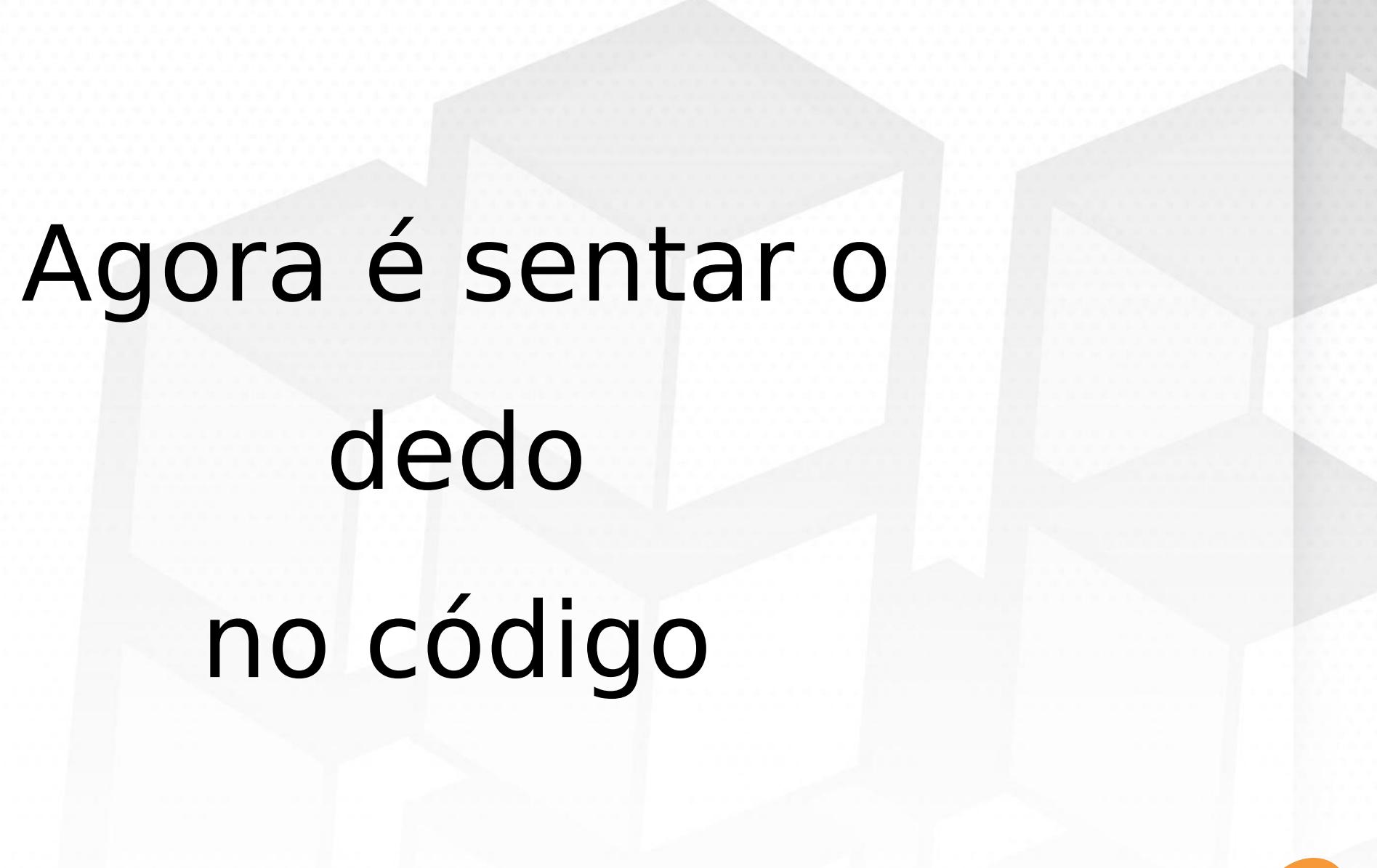
Erro Quadrático Médio

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

R2

$$R^2 = 1 - \frac{\sum_{i=1}^n (y - \hat{y})^2}{\sum_{i=1}^n (y - \bar{y})^2}$$





Agora é sentar o
dedo
no código

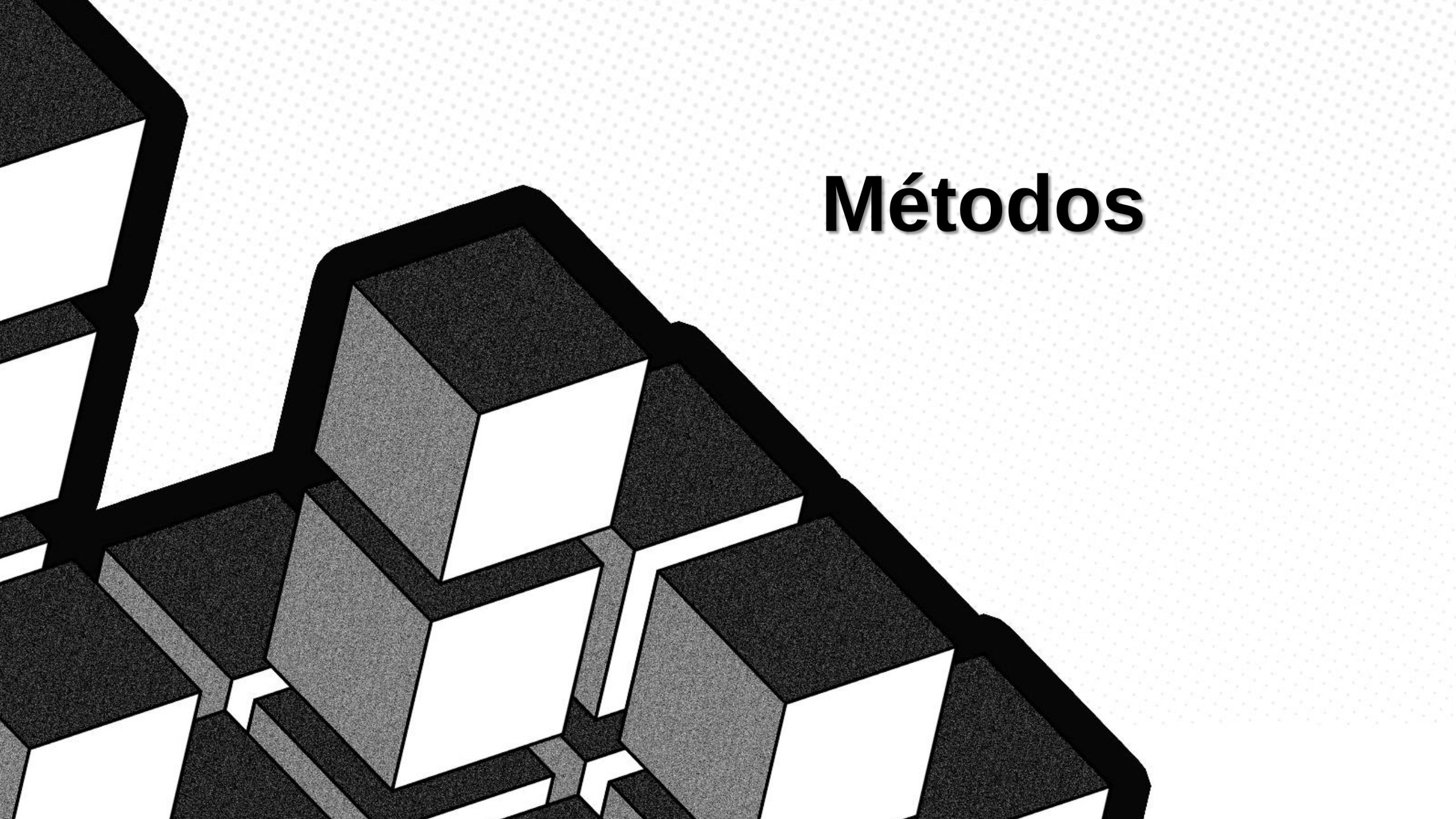


Classificação

Problemas de classificação são voltamos à estimativa alvo, sendo este um rótulo ou classe.

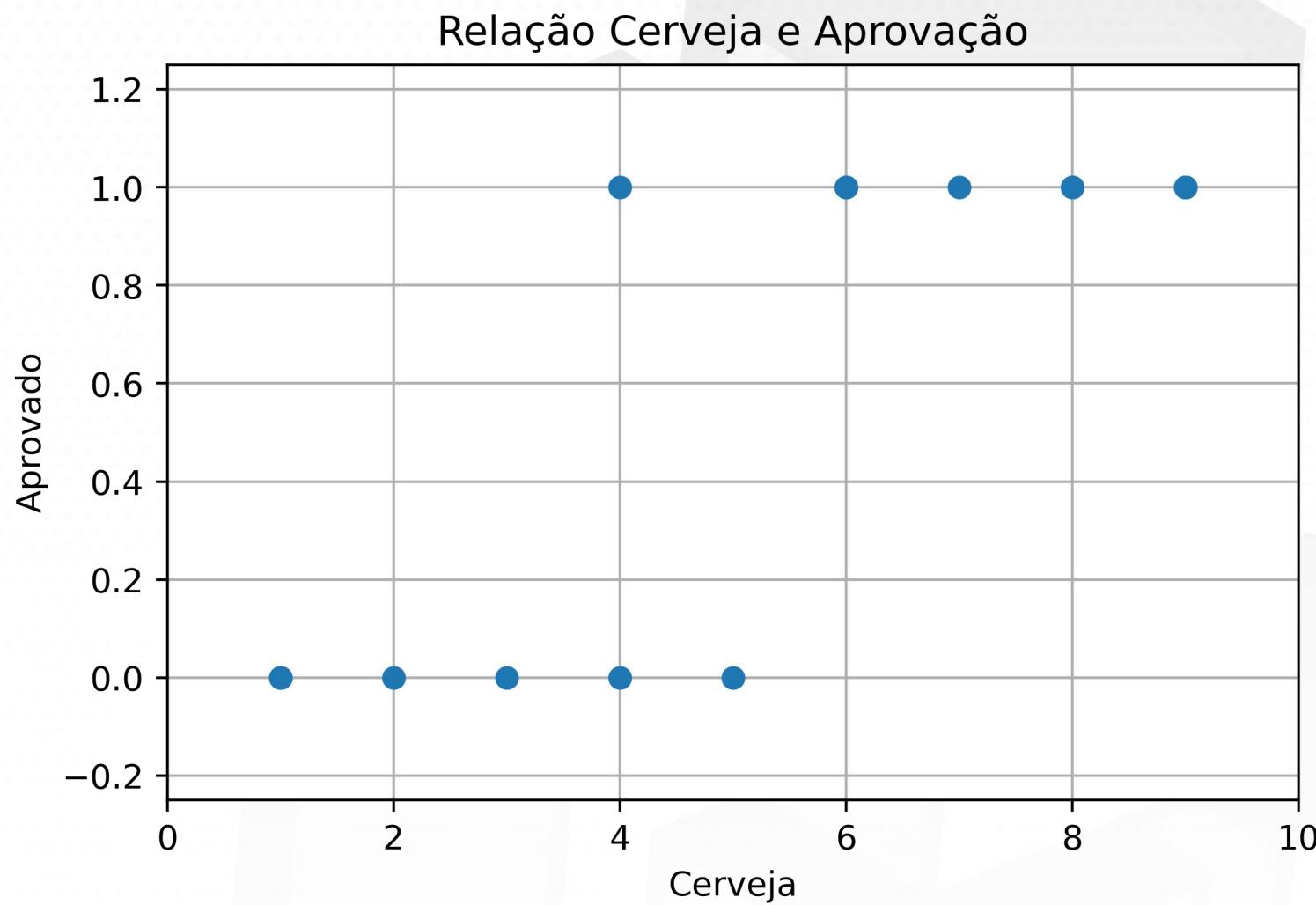
Por exemplo:

- Compradores vs Não compradores
- Churn vs Não Churn
- Objeto em uma imagem
- Inadimplente vs Adimplente (default)
- Propenção à câncer

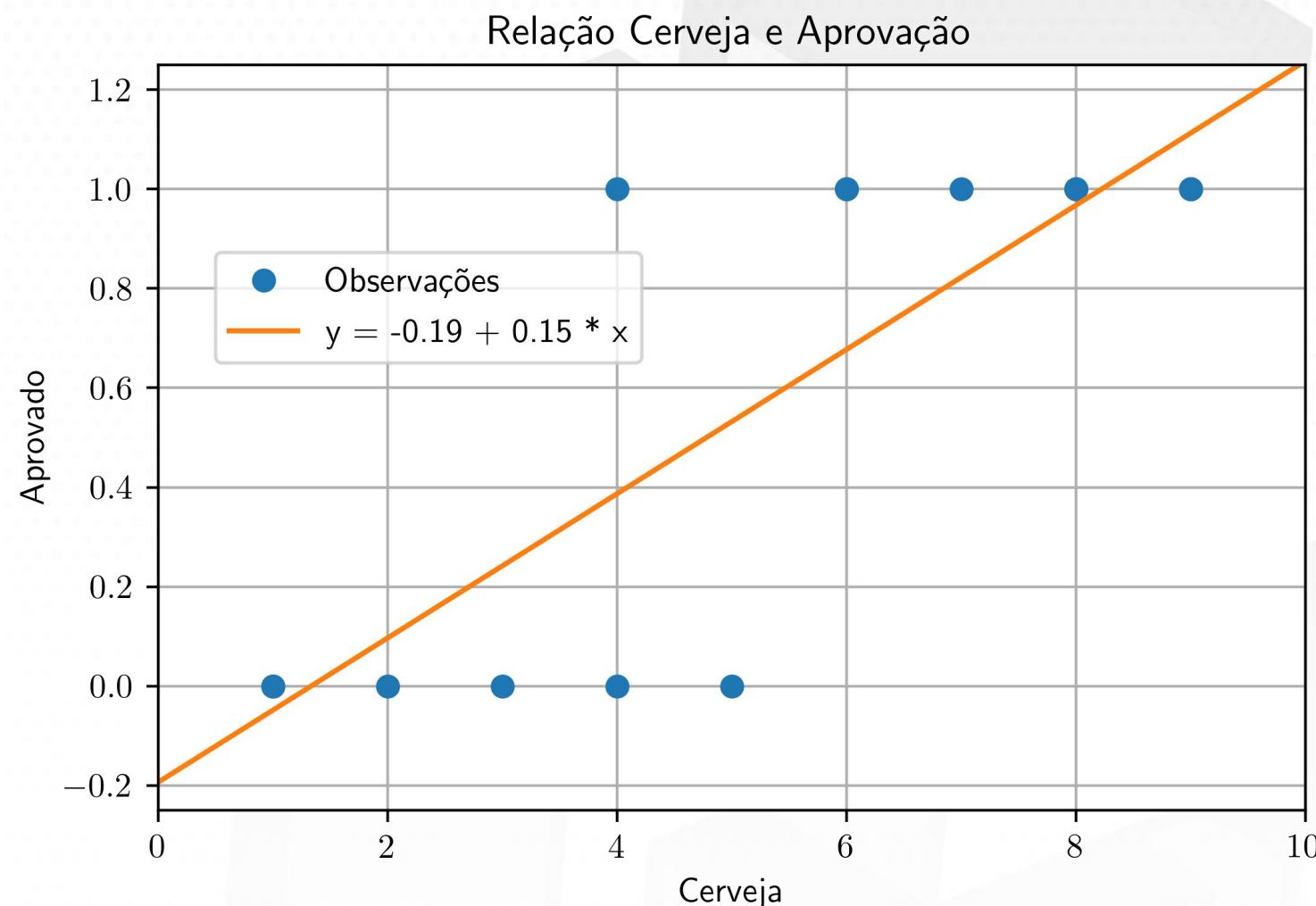
The background of the slide features a complex arrangement of black and white cubes. Some cubes are solid black, while others are white with black outlines. They are stacked and interconnected in a way that creates a sense of depth and perspective, resembling a 3D geometric puzzle. The overall aesthetic is minimalist and modern.

Métodos

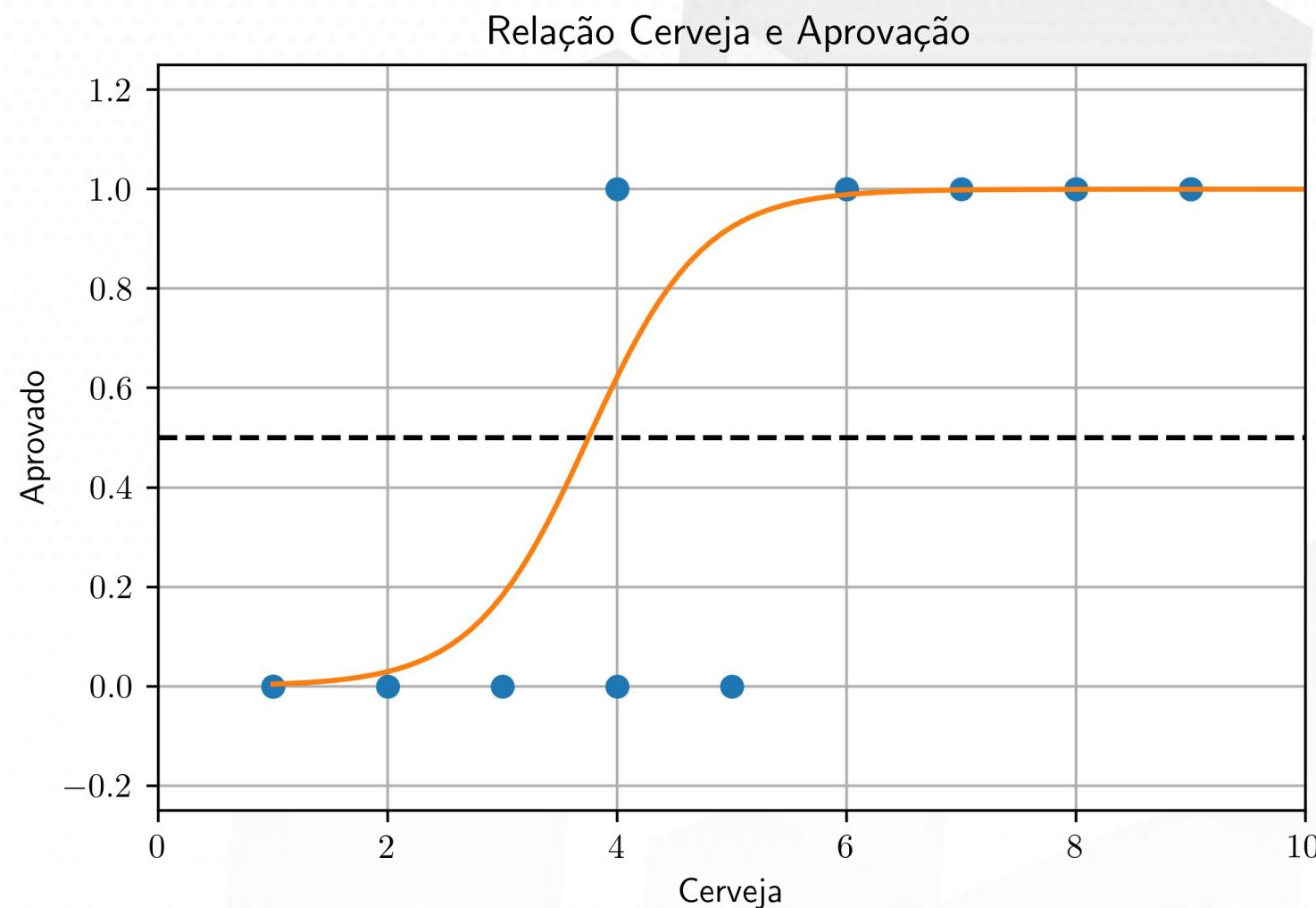
Regressão Logística



Regressão Logística



Regressão Logística

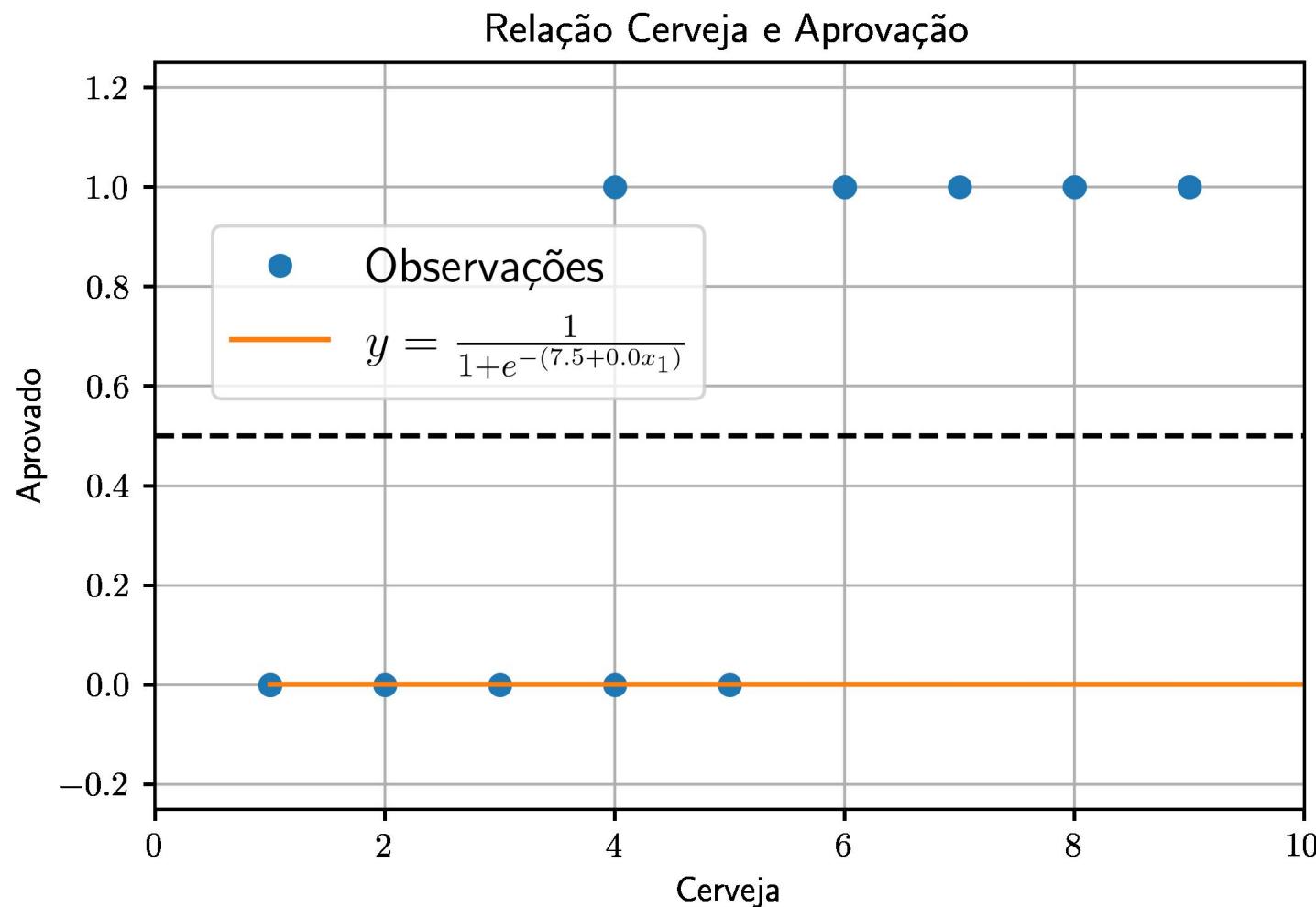


Regressão Logística

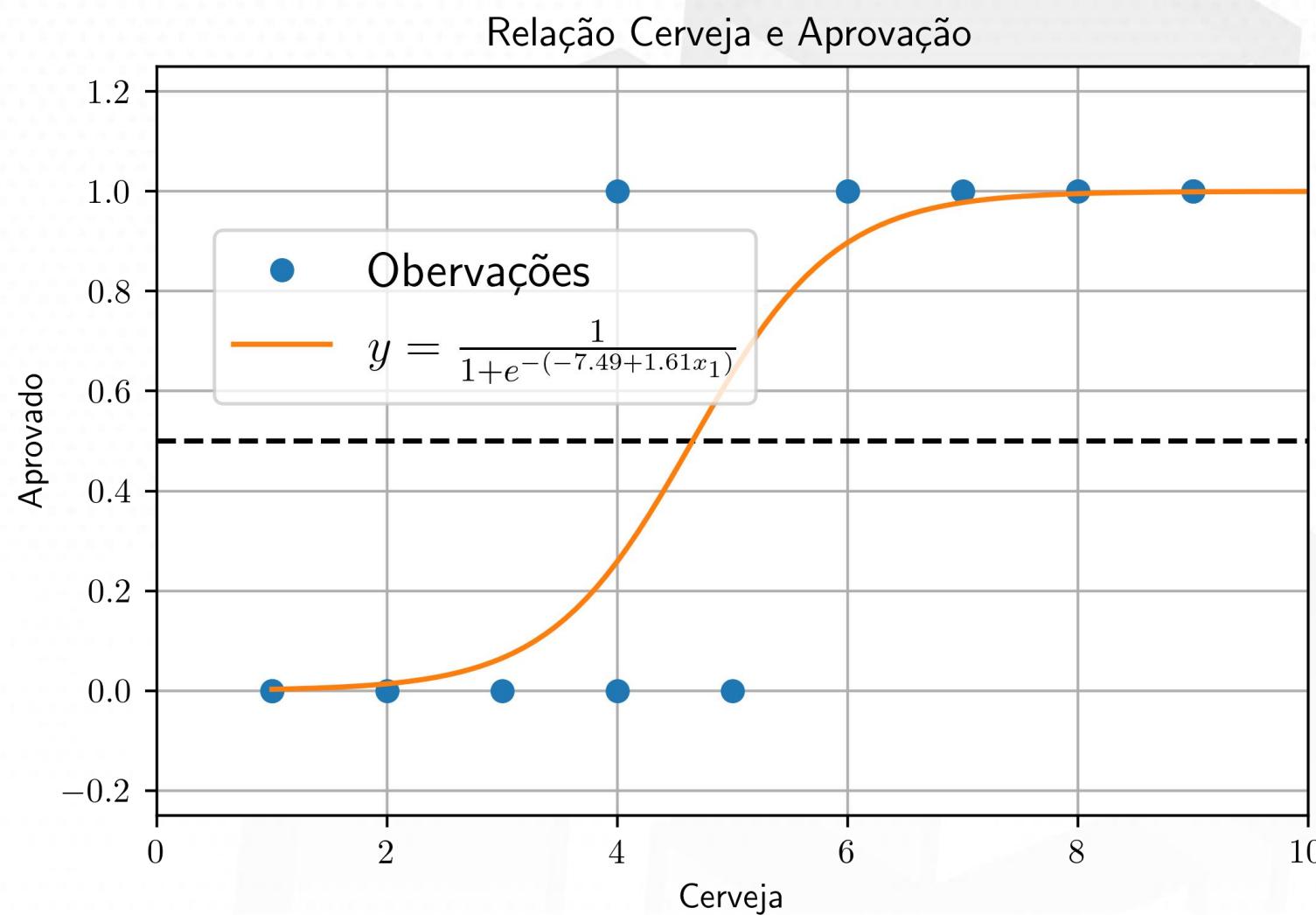
$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}}$$



Regressão Logística



Regressão Logística



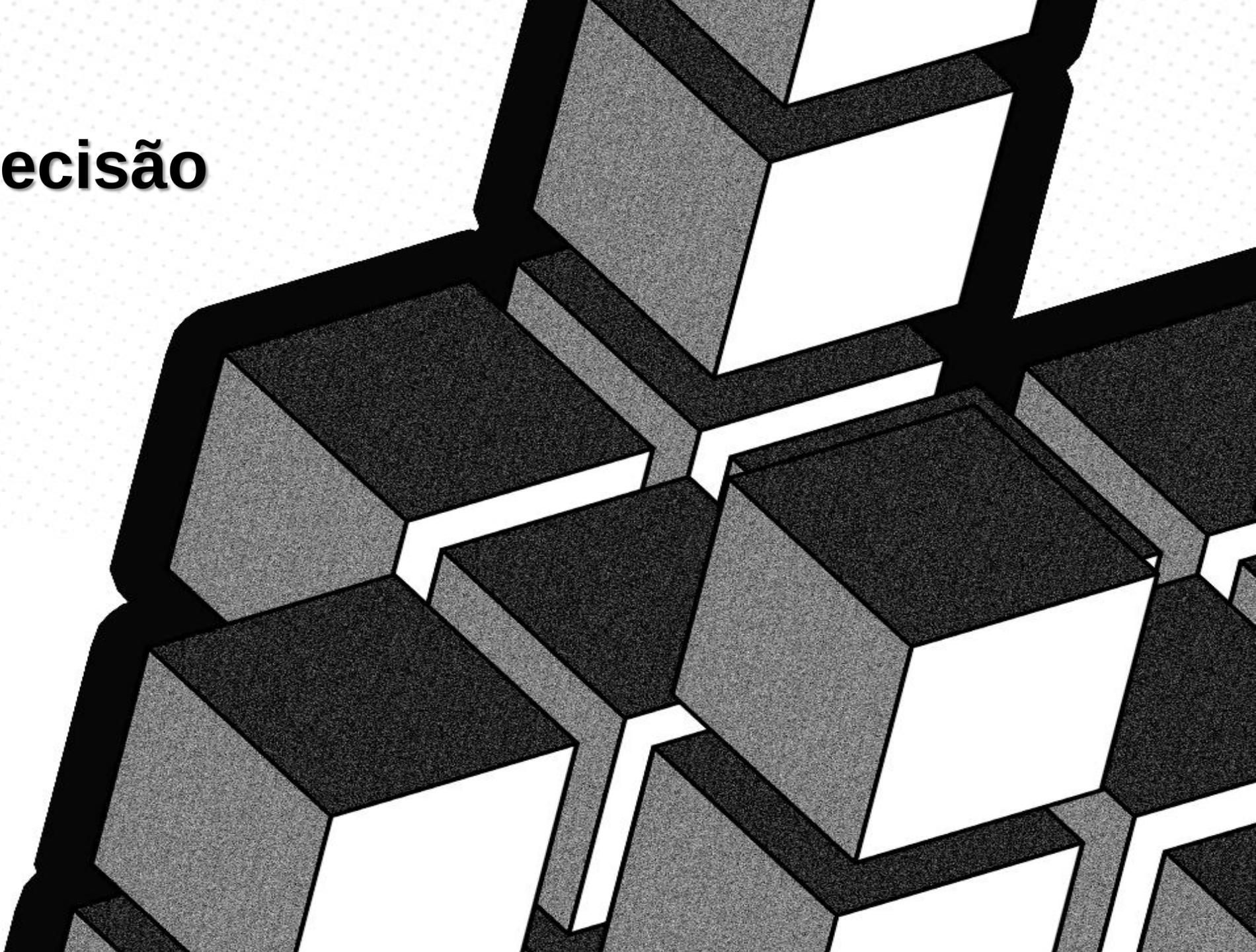
Regressão Logística

Log Loss

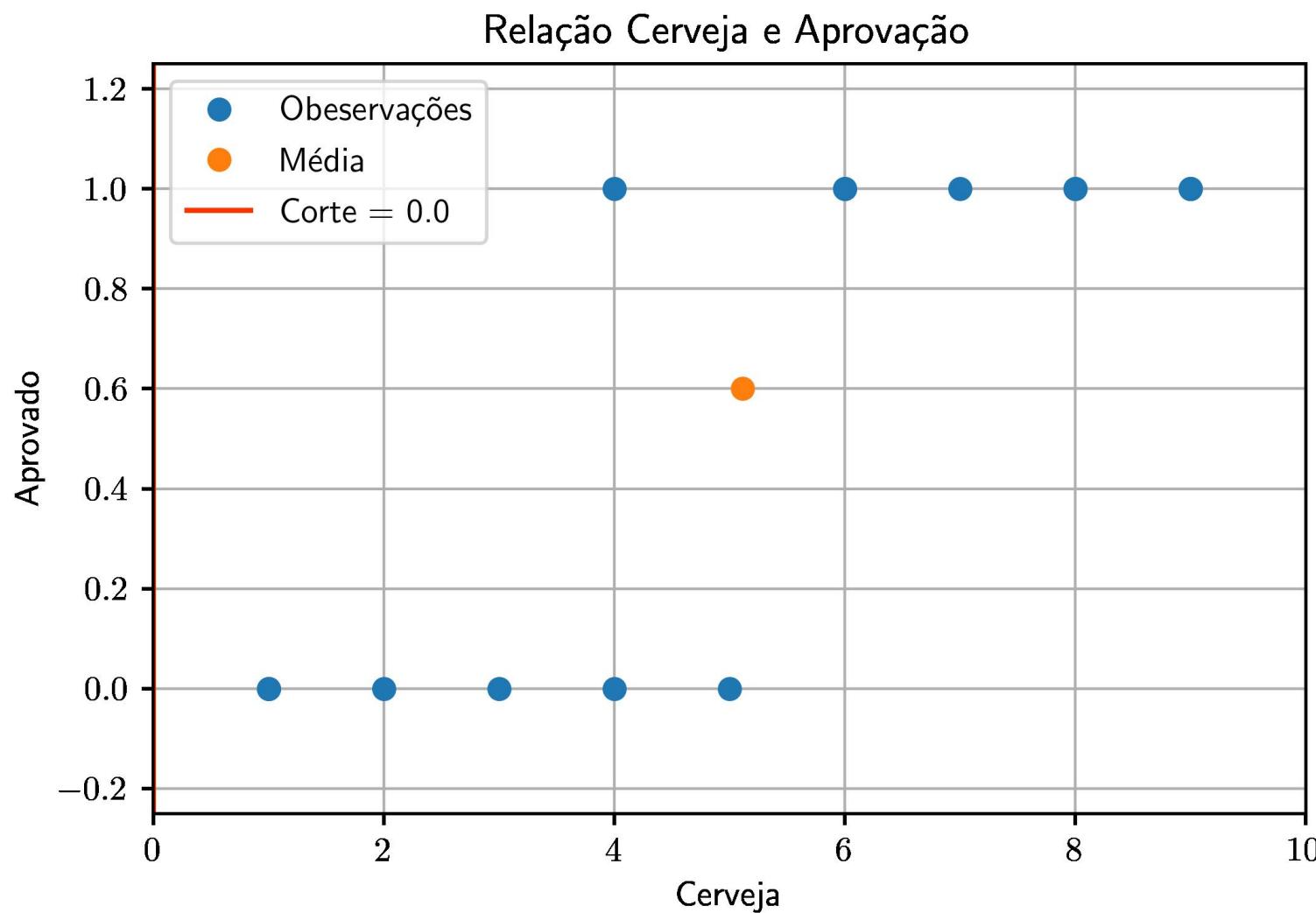
$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p))$$

$$\sum_{i=1}^n L_{\log}(y_i, p_i) = - \sum_{i=1}^n (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

Árvore de Decisão



Árvore de Decisão



Árvore de Decisão

Anteriormente utilizamos a Soma dos Erros Quadráticos.

E Agora? Vamos usar qual medida?

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n \|x_i - x_j\|}{2n^2 \bar{x}}$$



Árvore de Decisão

Há outras medidas?

Sim!

Entropia!!

$$H = -[p \log_2(p) + (1 - p) \log_2(1 - p)]$$

De uma maneira mais genérica

$$H = - \sum_{i=1}^c p_i \log_2(p_i)$$

Árvore de Decisão

Gini

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n \|x_i - x_j\|}{2n^2 \bar{x}}$$

Entropia

$$H = - \sum_{i=1}^c p_i \log_2(p_i)$$



Naive-Bayes

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Naive-Bayes

y	x		
Diabetes	Histórico Familiar	Acima do Peso	Atividade Física
1	1	1	1
1	1	1	0
1	0	1	0
1	1	0	0
0	1	0	1
0	0	1	1
0	0	0	0

$y = \text{Diabetes}$

$x_1 = \text{Histórico Familiar}$

$x_2 = \text{Acima do Peso}$

$x_3 = \text{Atividade Física}$



Naive-Bayes

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

$$P(y|X) = P(y|x_1, x_2, x_3) = \frac{P(y)P(x_1, x_2, x_3|y)}{P(x_1, x_2, x_3)}$$

$$P(y|X) = \frac{P(y)P(x_1|y)P(x_2|y)P(x_3|y)}{P(x_1, x_2, x_3)} = \frac{P(y) \prod_{i=1}^p P(x_i|y)}{P(X)}$$



Naive-Bayes

y	x		
Diabetes	Histórico Familiar	Acima do Peso	Atividade Física
1	1	1	1
1	1	1	0
1	0	1	0
1	1	0	0
0	1	0	1
0	0	1	1
0	0	0	0

$$P(y = 1) =$$

$$P(x_1|y = 1) =$$

$$P(x_2|y = 1) =$$

$$P(x_3|y = 1) =$$

Naive-Bayes

y	x		
Diabetes	Histórico Familiar	Acima do Peso	Atividade Física
1	1	1	1
1	1	1	0
1	0	1	0
1	1	0	0
0	1	0	1
0	0	1	1
0	0	0	0

$$P(y = 0) =$$

$$P(x_1|y = 0) =$$

$$P(x_2|y = 0) =$$

$$P(x_3|y = 0) =$$

Naive-Bayes

$$P(x_1, x_2, x_3)$$

$$\begin{aligned} P(x_1, x_2, x_3) &= P(y = 1)P(x_1, x_2, x_3|y = 1) \\ &\quad + P(y = 0)P(x_1, x_2, x_3|y = 0) \end{aligned}$$

Naive-Bayes

Juntando tudo



Naive-Bayes

Complicando um pouco, mas nem tanto

Distribuição Bernoulli

$$f(x_i; \theta) = \theta^{x_i} (1 - \theta)^{1 - x_i}$$

$$f(x_i; \theta|y) = \theta_y^{x_i} (1 - \theta_y)^{1 - x_i}$$

Naive-Bayes

y	x		
Diabetes	Histórico Familiar	Acima do Peso	Atividade Física
1	1	1	1
1	1	1	0
1	0	1	0
1	1	0	0
0	1	0	1
0	0	1	1
0	0	0	0

$$f(x_1; \theta) =$$

$$f(x_1; \theta | y = 1) =$$

$$f(x_1; \theta | y = 0) =$$



Naive-Bayes

Para “Acima do Peso”

y	x		
Diabetes	Histórico Familiar	Acima do Peso	Atividade Física
1	1	1	1
1	1	1	0
1	0	1	0
1	1	0	0
0	1	0	1
0	0	1	1
0	0	0	0



Naive-Bayes

Para “Atividade Física”

y	x		
Diabetes	Histórico Familiar	Acima do Peso	Atividade Física
1	1	1	1
1	1	1	0
1	0	1	0
1	1	0	0
0	1	0	1
0	0	1	1
0	0	0	0



Naive-Bayes

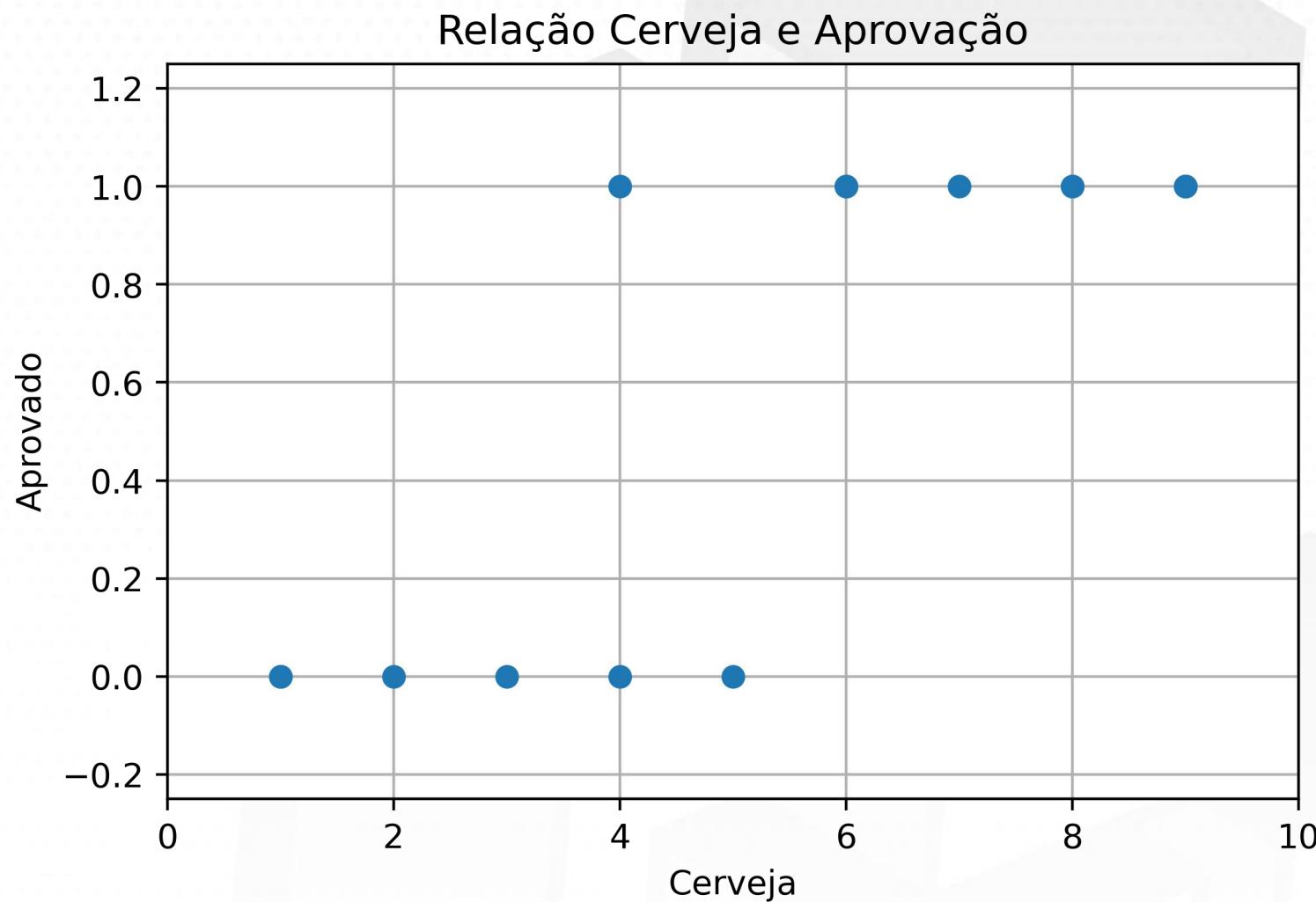
A fórmula mais genérica

$$P(y_c|X) = \frac{P(y_c) \prod_{i=1}^p f(x_i; \theta|y_c)}{P(X)}$$

$$P(y_c|X) = \frac{P(y_c) \prod_{i=1}^p f(x_i; \theta|y_c)}{\sum_{c=0}^C P(y_c) \prod_{i=1}^p f(x_i; \theta|y_c)}$$



Voltando à Cerveja e Aprovação



Voltando à Cerveja e Aprovação

Nossas variáveis são numéricas e não mais booleanas ou categóricas!!

Podemos usar outra distribuição para x?

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

[Distribuição Normal](#)

Voltando à Cerveja e Aprovação

Estatísticas de **Cerveja**

Média: 5,46

Variância: 8.26

Desvio Padrão: 2.87

Estatísticas de **Cerveja | Nota = 0**

Média: 2,66

Variância: 2.66

Desvio Padrão: 1.63

Estatísticas de **Cerveja | Nota = 1**

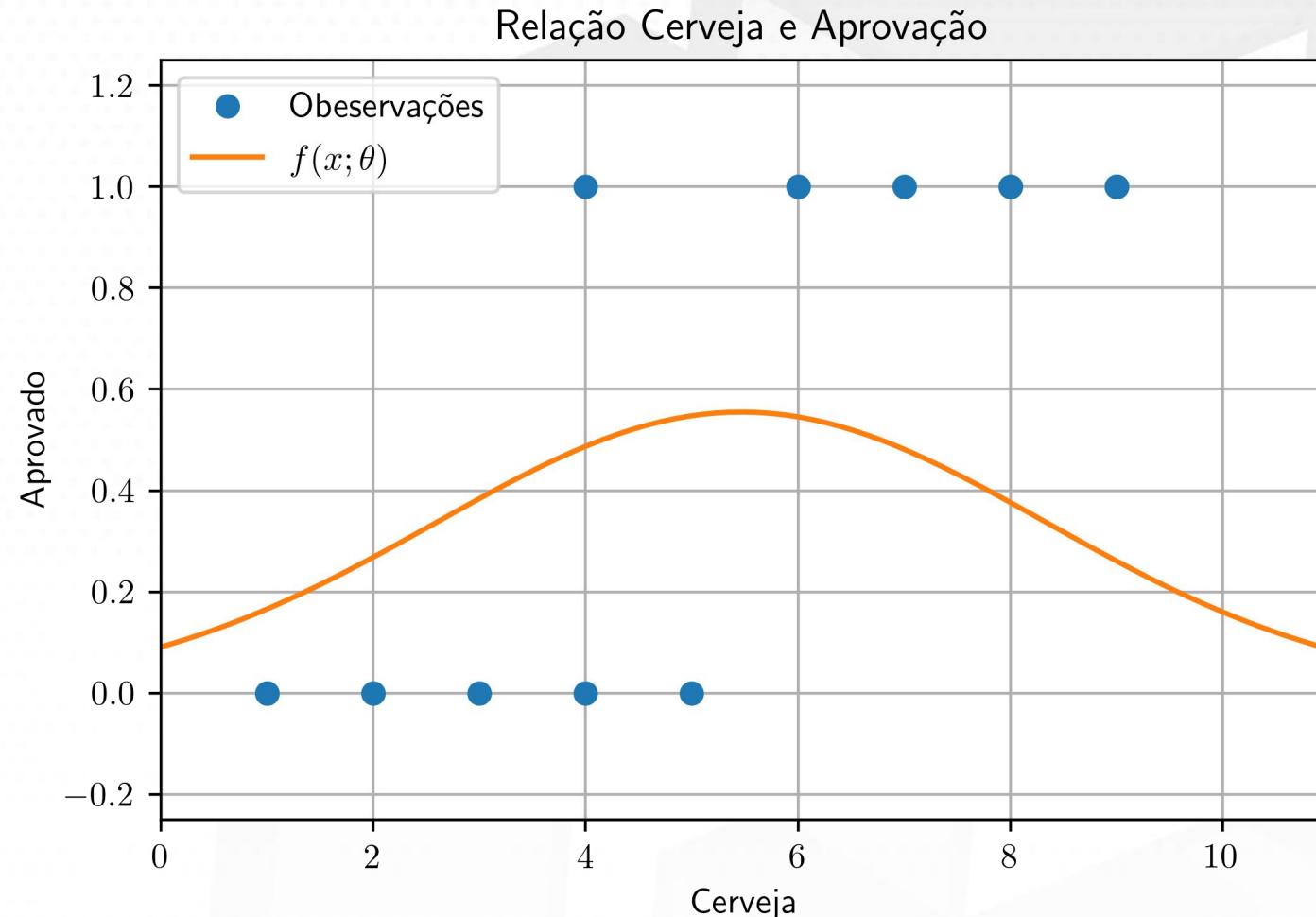
Média: 7,33

Variância: 3.00

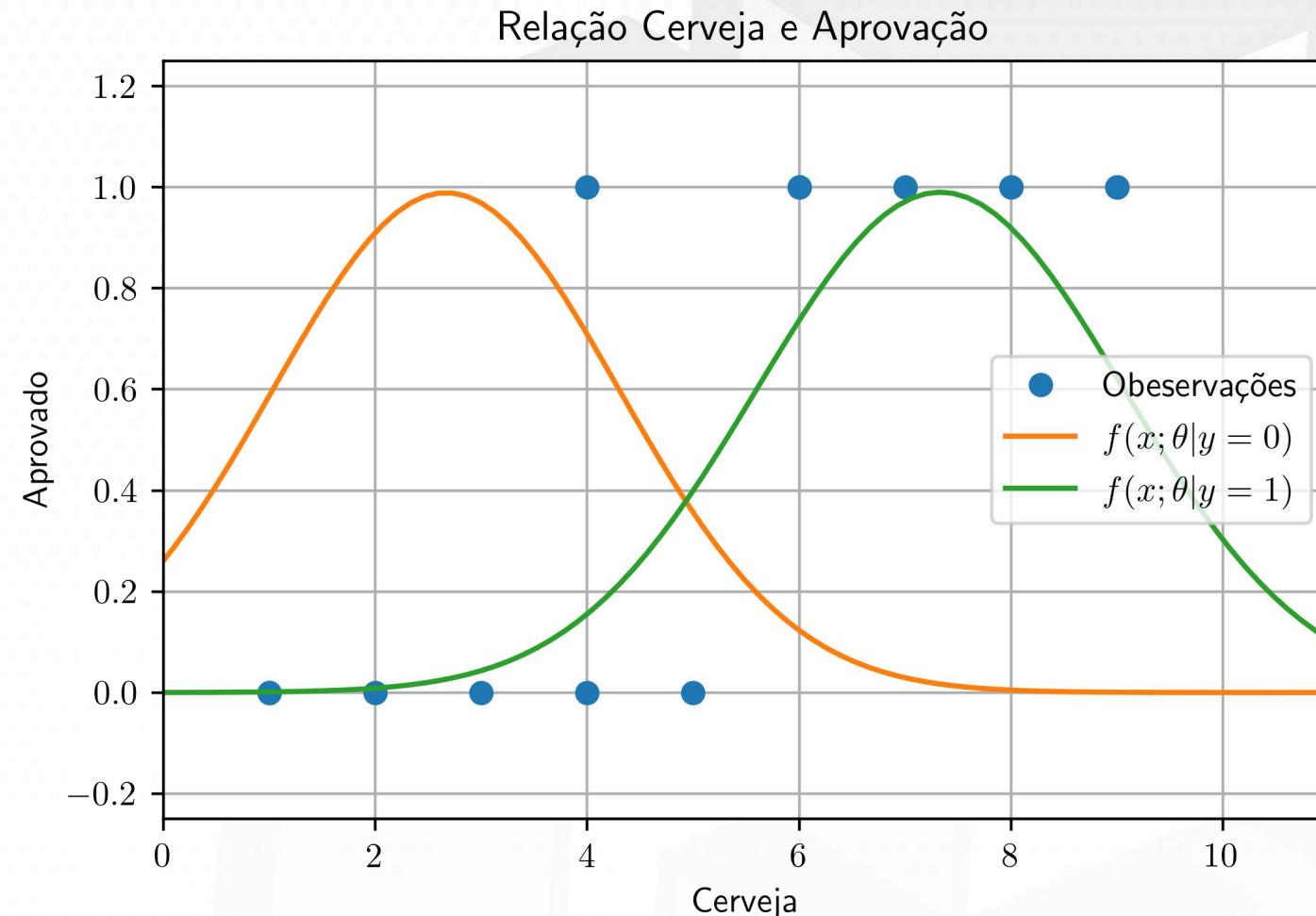
Desvio Padrão: 1.73



Voltando à Cerveja e Aprovação



Voltando à Cerveja e Aprovação



Naive-Bayes

$$\hat{y} = \arg \max_y P(y)P(X|y)$$

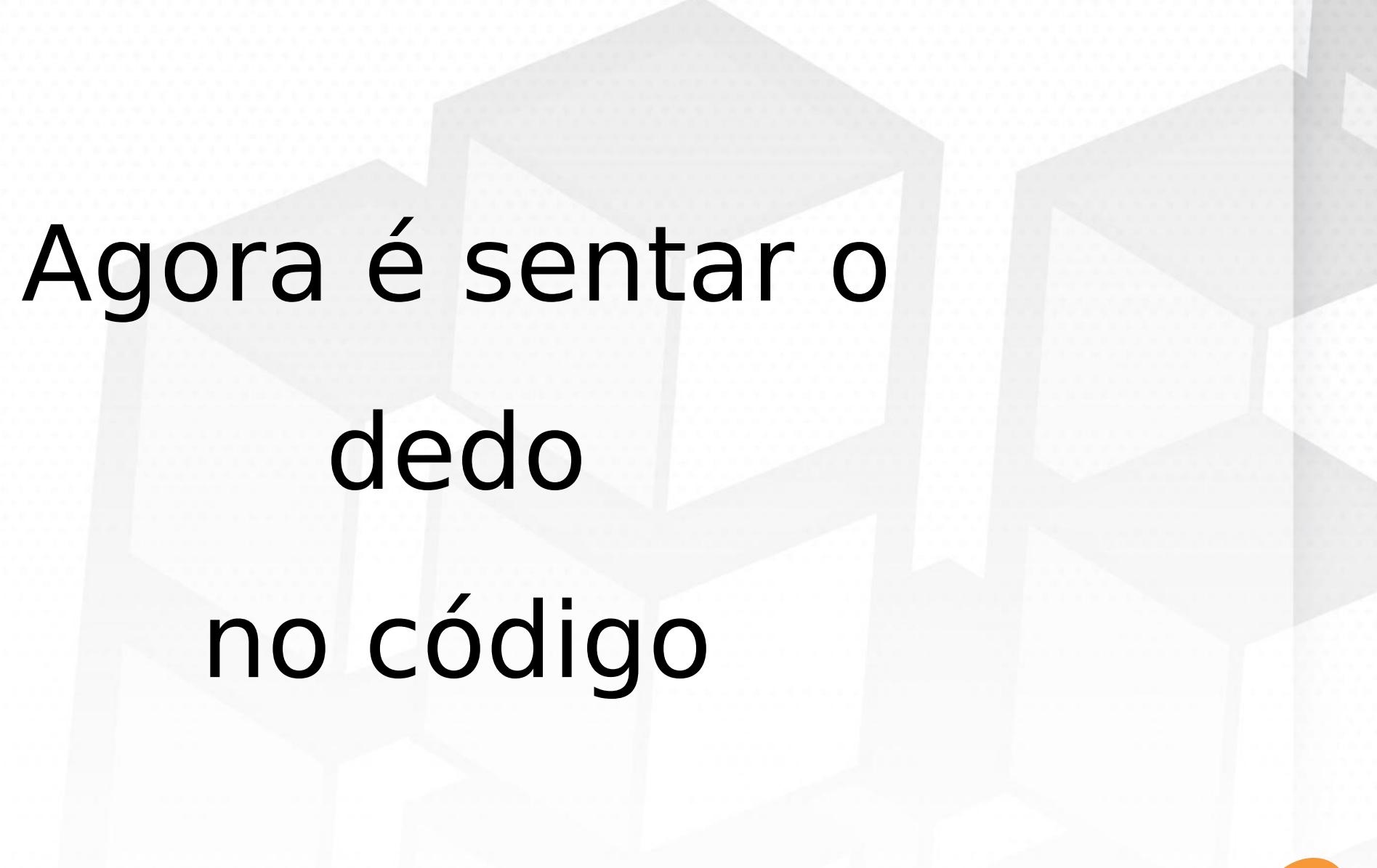


Métricas de Ajuste

[Canal Téo Me Why](#)

- Matriz de Confusão
- Acurácia
- Precisão
- Recall
- Curva ROC





Agora é sentar o
dedo
no código



Obrigado

Téo Calvo

teo.bcalvo@gmail.com

 /in/teocalvo

 /teomewhy