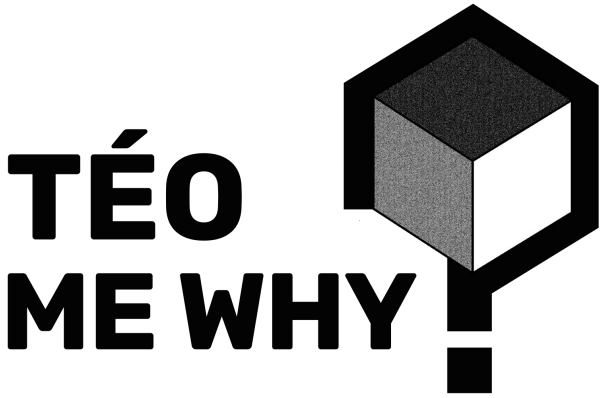
Introdução à Machine Learning

Utilizando a biblioteca scikit-learn



25 e 27 de Maio de 2021

Agenda

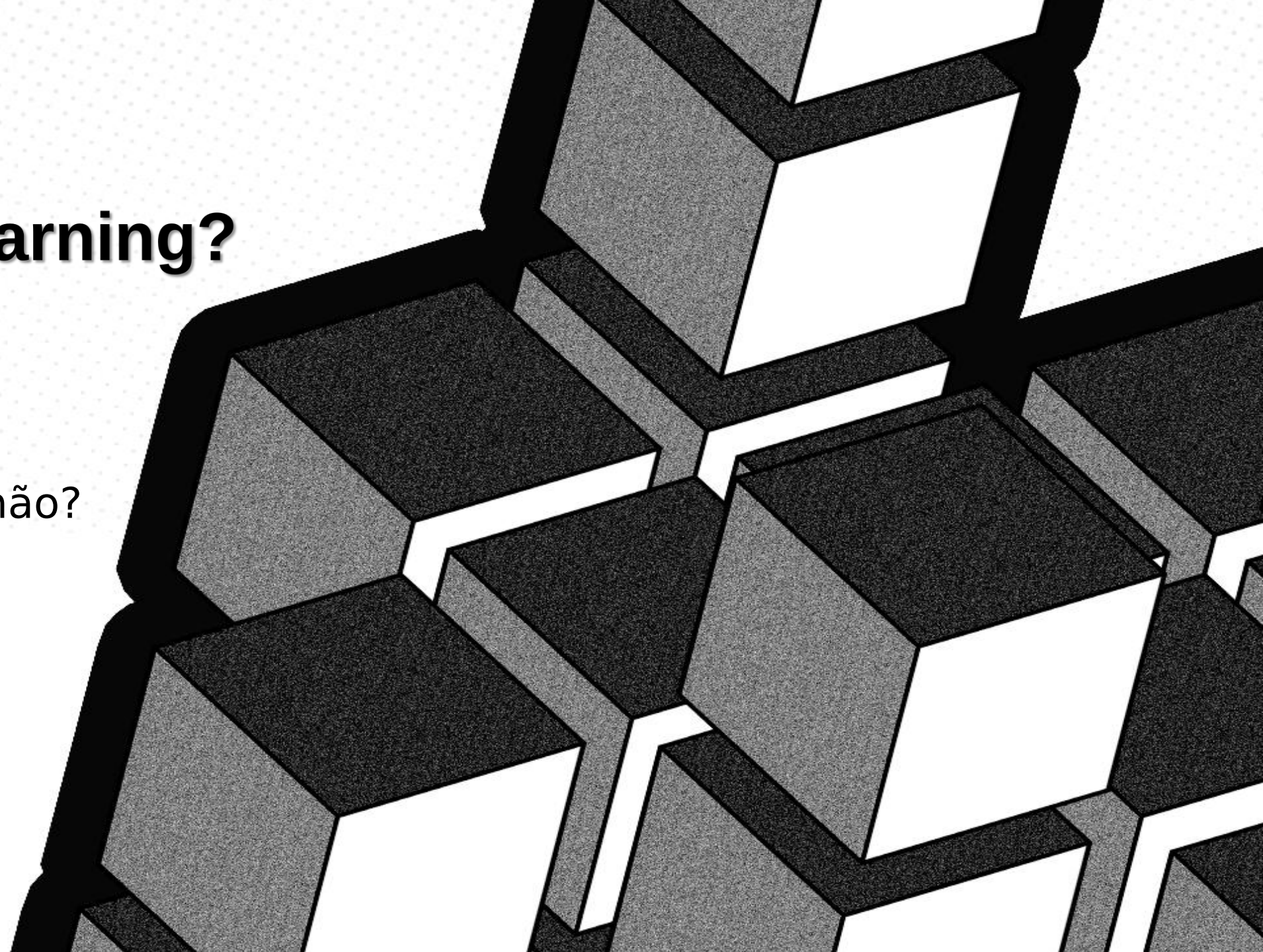
- O que é Machine Learning
- Tipos de Aprendizado
- Pré processamento
- Metricas de Ajuste

scikit learn

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]

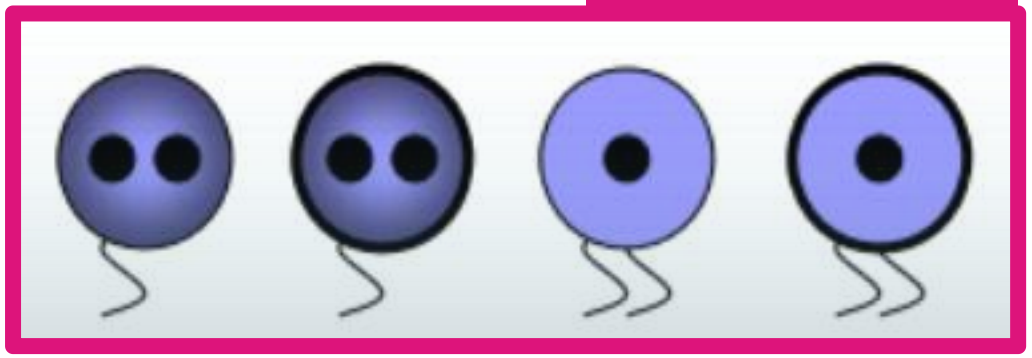
[1] https://www.britannica.com/technology/machine-learning

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

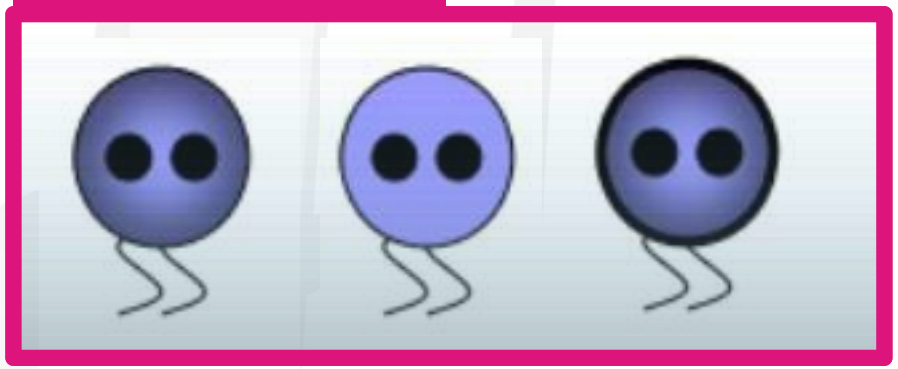
fonte: https://pt.wikipedia.org/wiki/Aprendizado_de_m%C3%A1quina

scikit learn

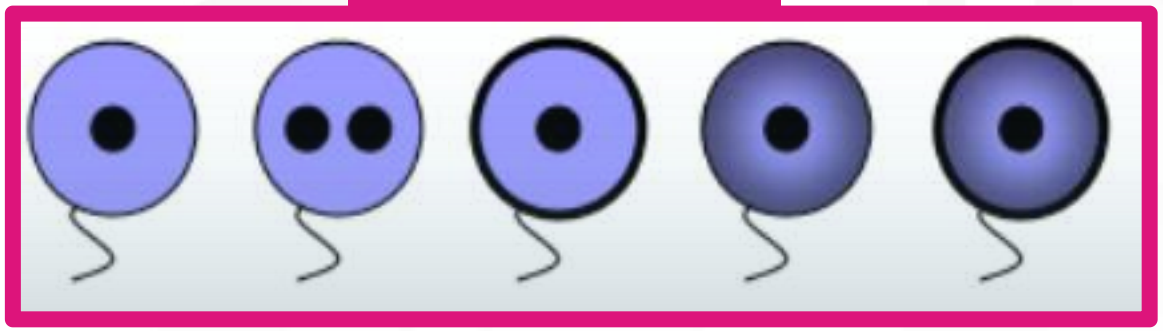
Saudável



Burpona



Lethargia

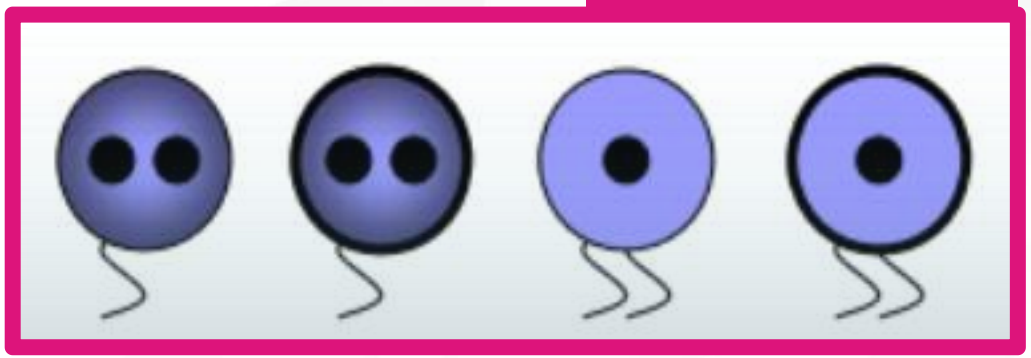


Quais atributos temos?

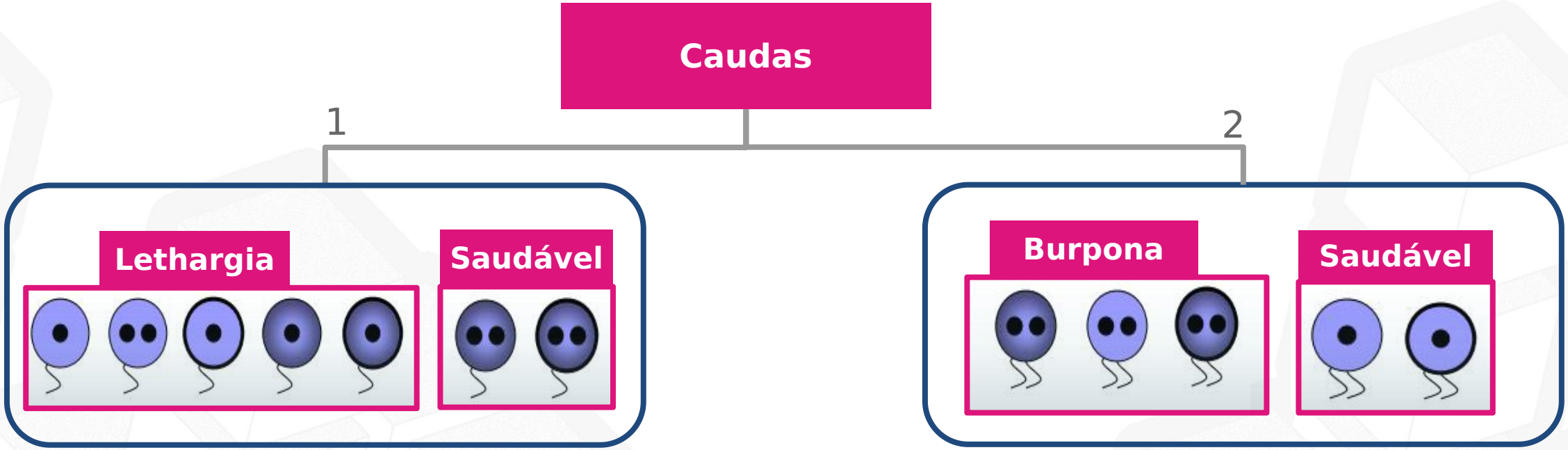
- Quantidade de caudas
- Quantidade de núcleos
- Tipo de membrana
- Cor

O que me difere das demais?

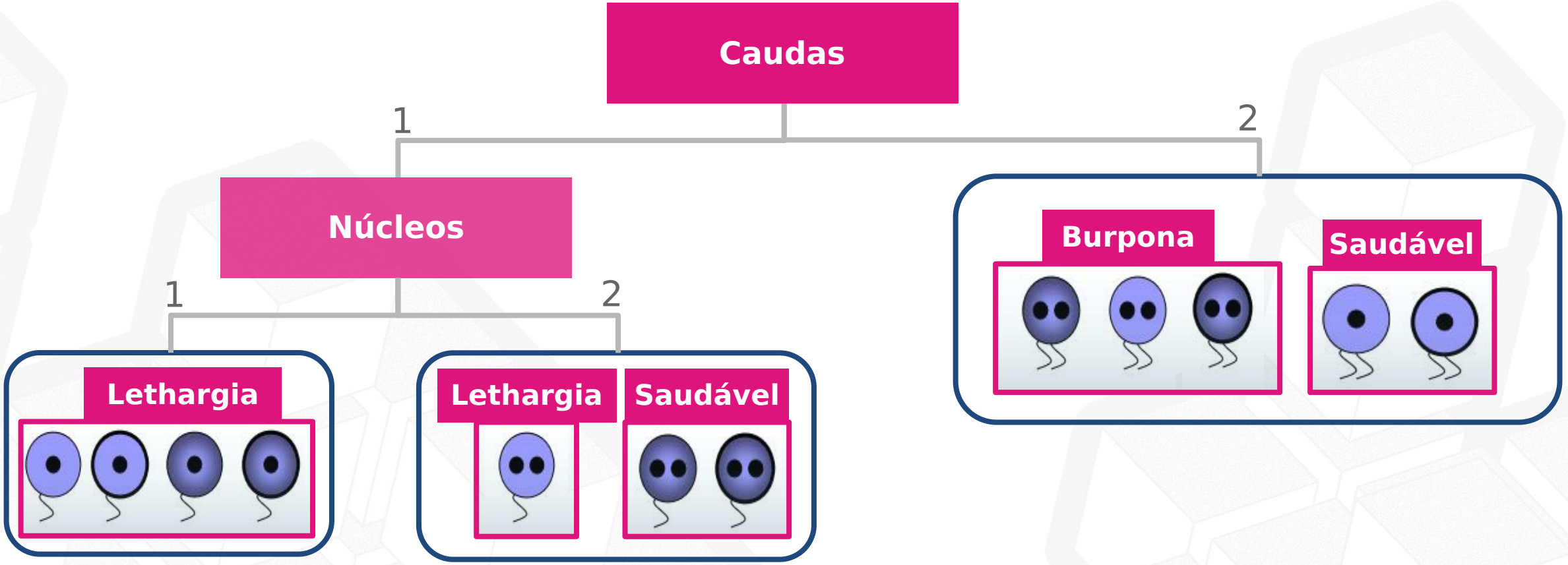
Saudável



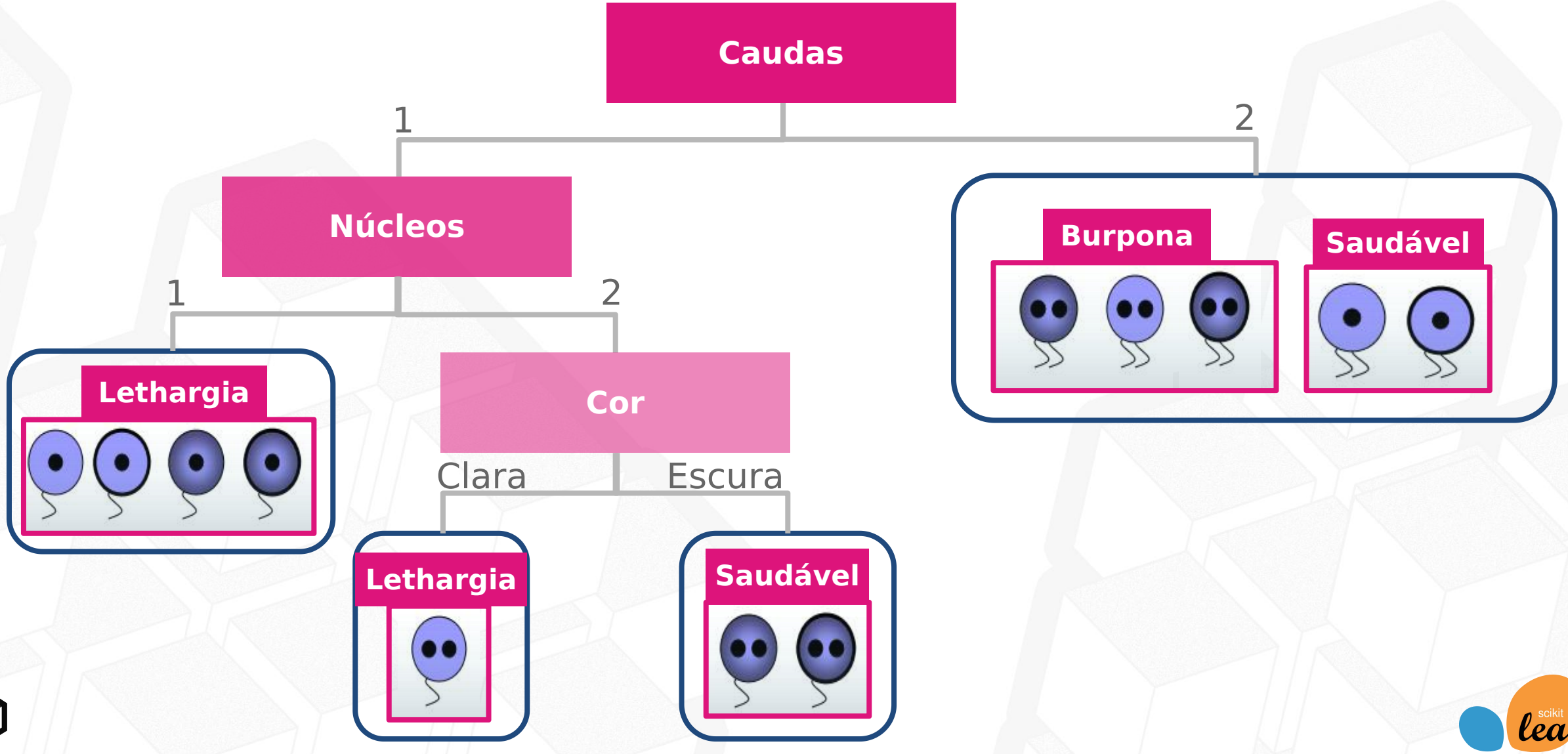
Criando regras



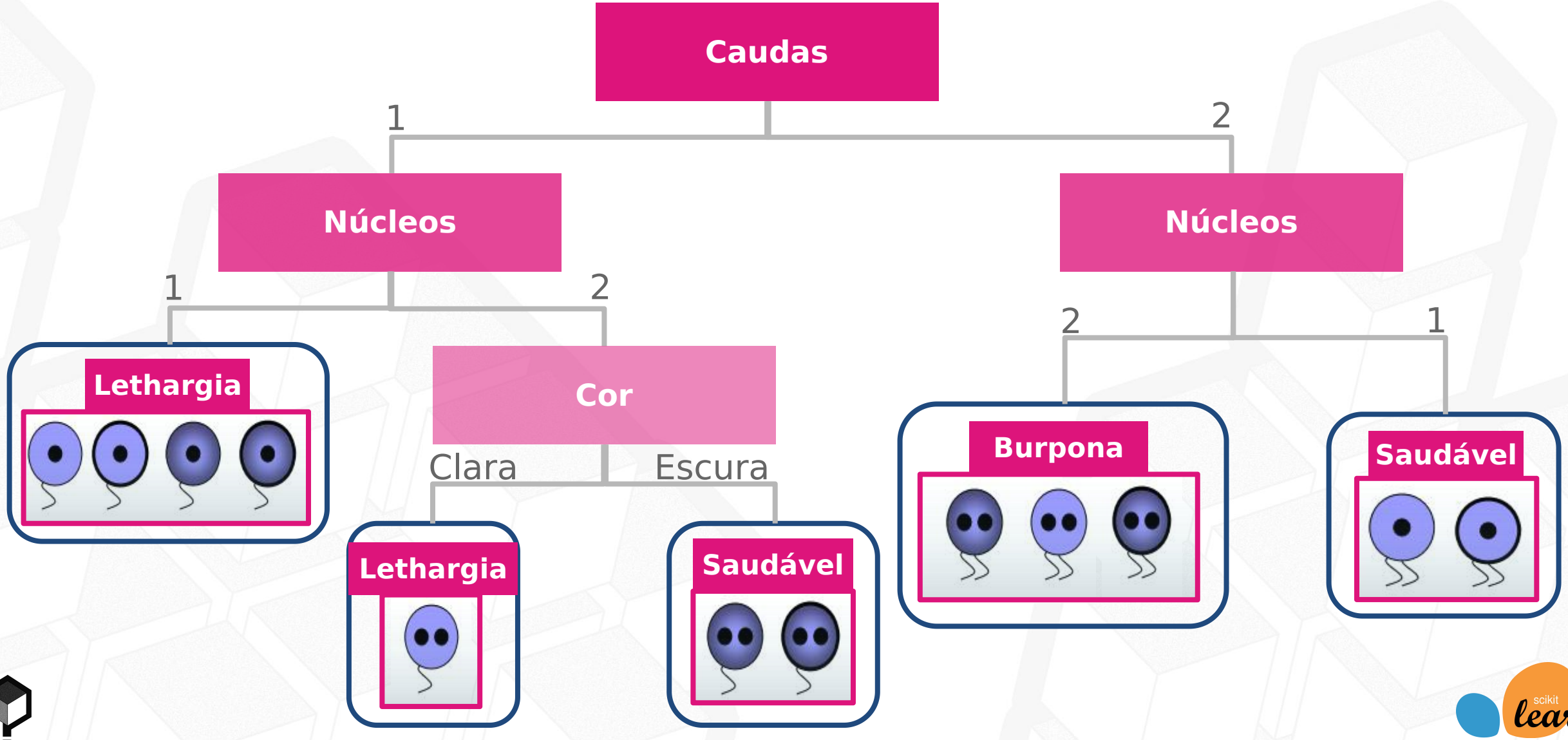
Criando regras



Criando regras



Criando regras



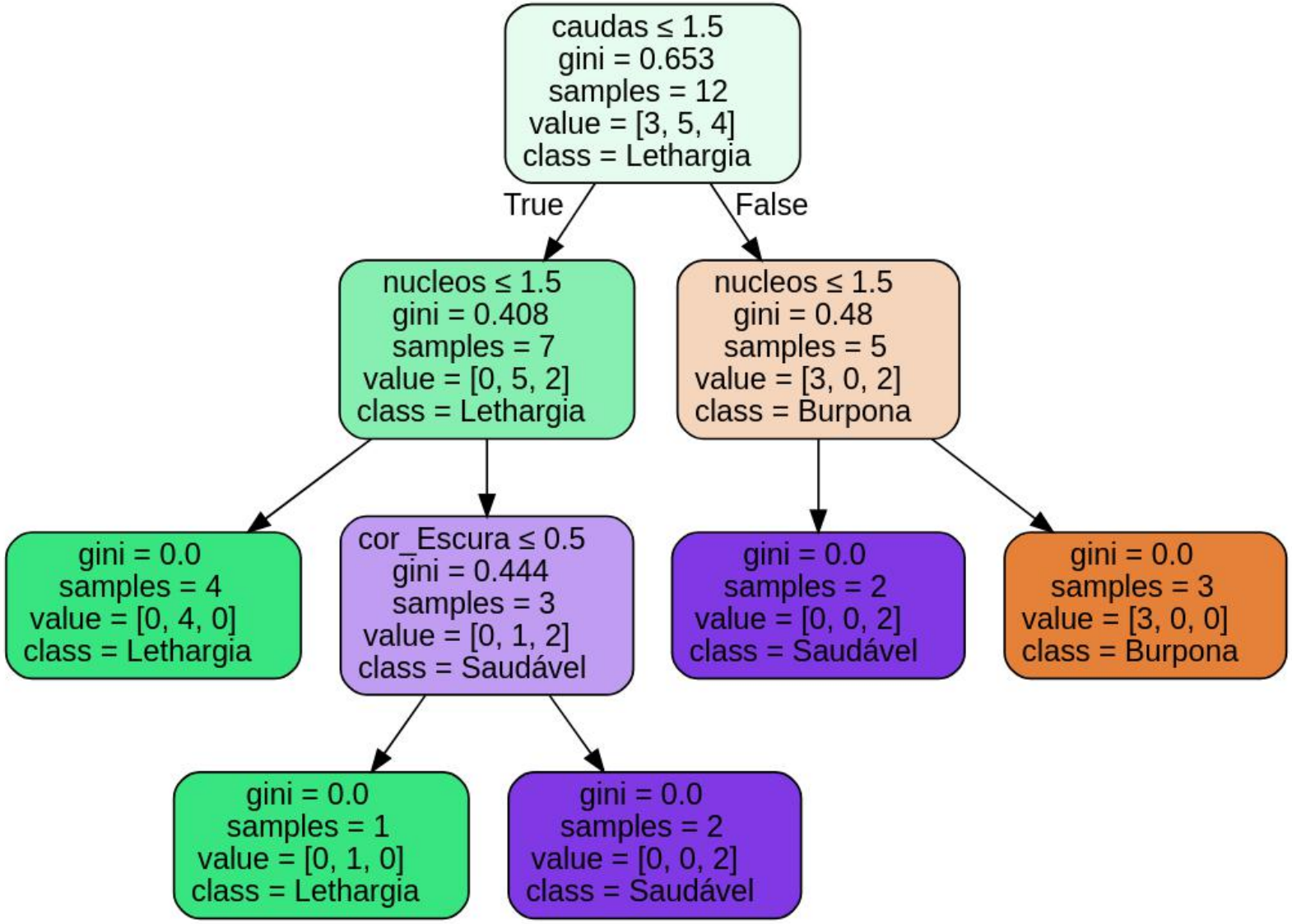
Tabela

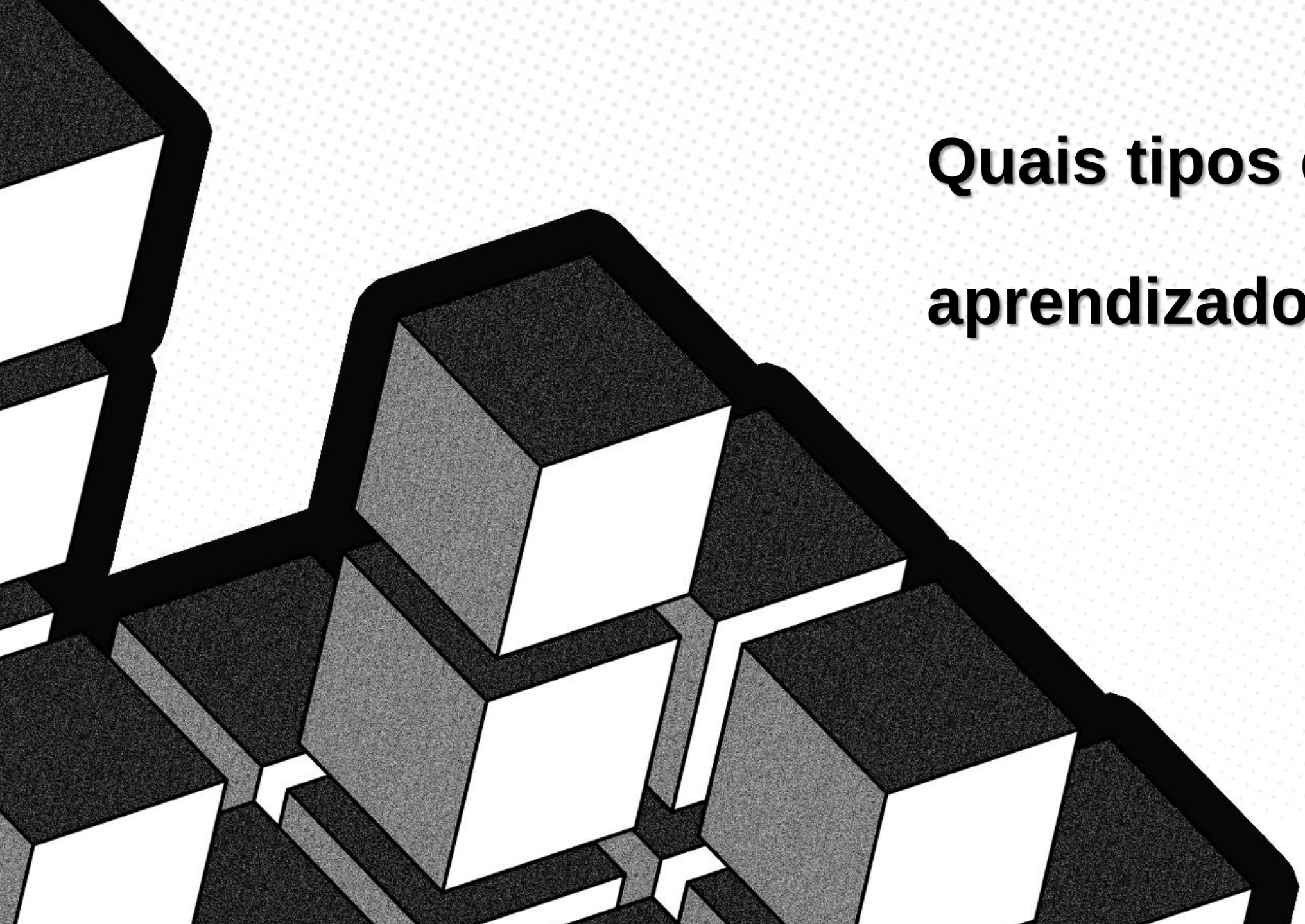
id	Núcleos	Caudas	Cor	Membrana	Classe
1	1	1	Clara	Fina	Lethargia
2	2	1	Clara	Fina	Lethargia
3	1	1	Clara	Grossa	Lethargia
4	1	1	Escura	Fina	Lethargia
5	1	1	Clara	Grossa	Burpona
6	2	2	Escura	Fina	Buporna
7	2	2	Escura	Fina	Burpona
8	2	2	Escura	Grossa	Burpona
9	2	1	Escura	Fina	Saudável
10	2	1	Escura	Grossa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

Atributos

Alvo

Árvore no Python





Quais tipos de aprendizado temos?

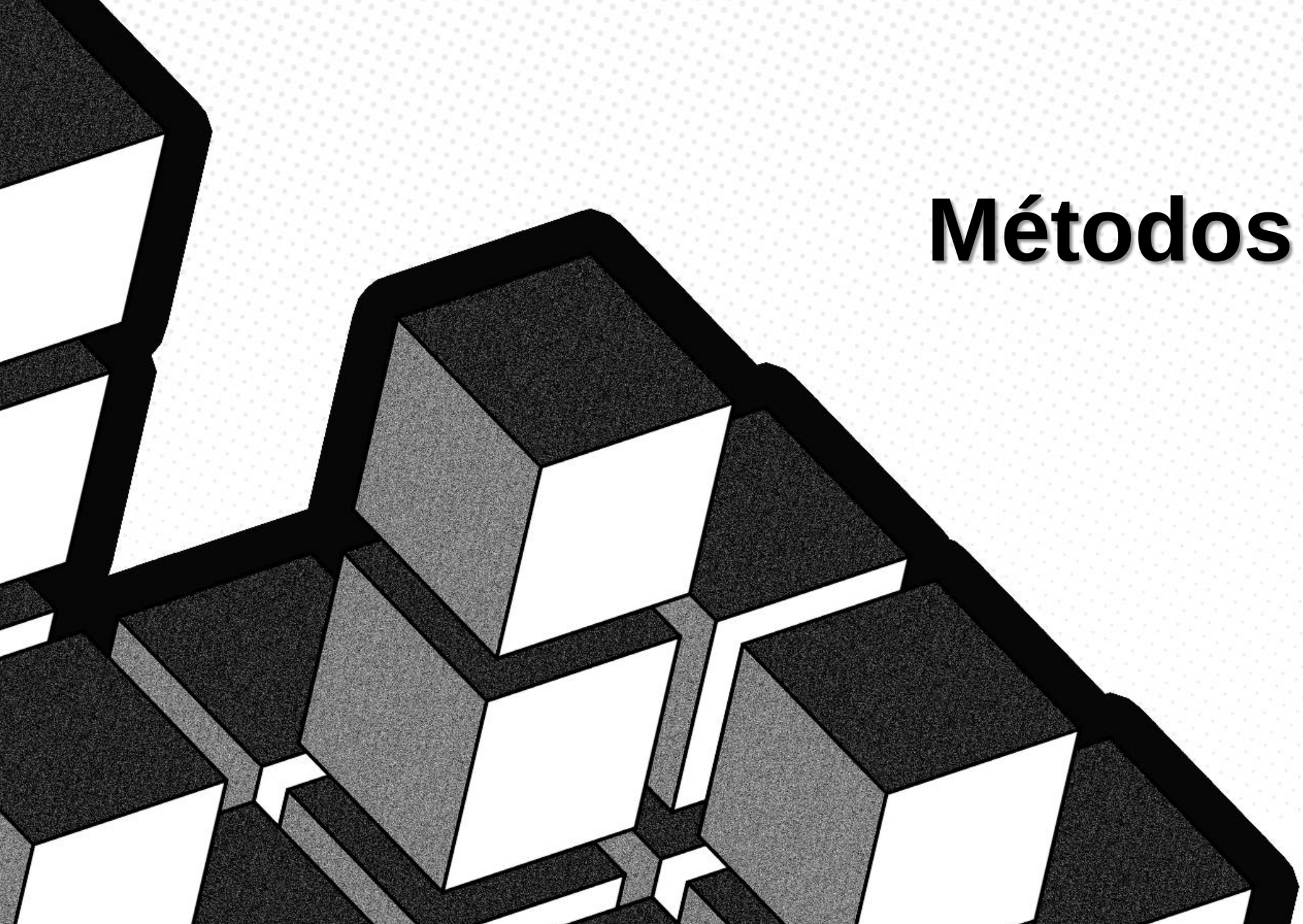

Aprendizado de Máquina
Não Supervisionado
Supervisionado
Clustering
Regressão
Classificação

Regressão

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

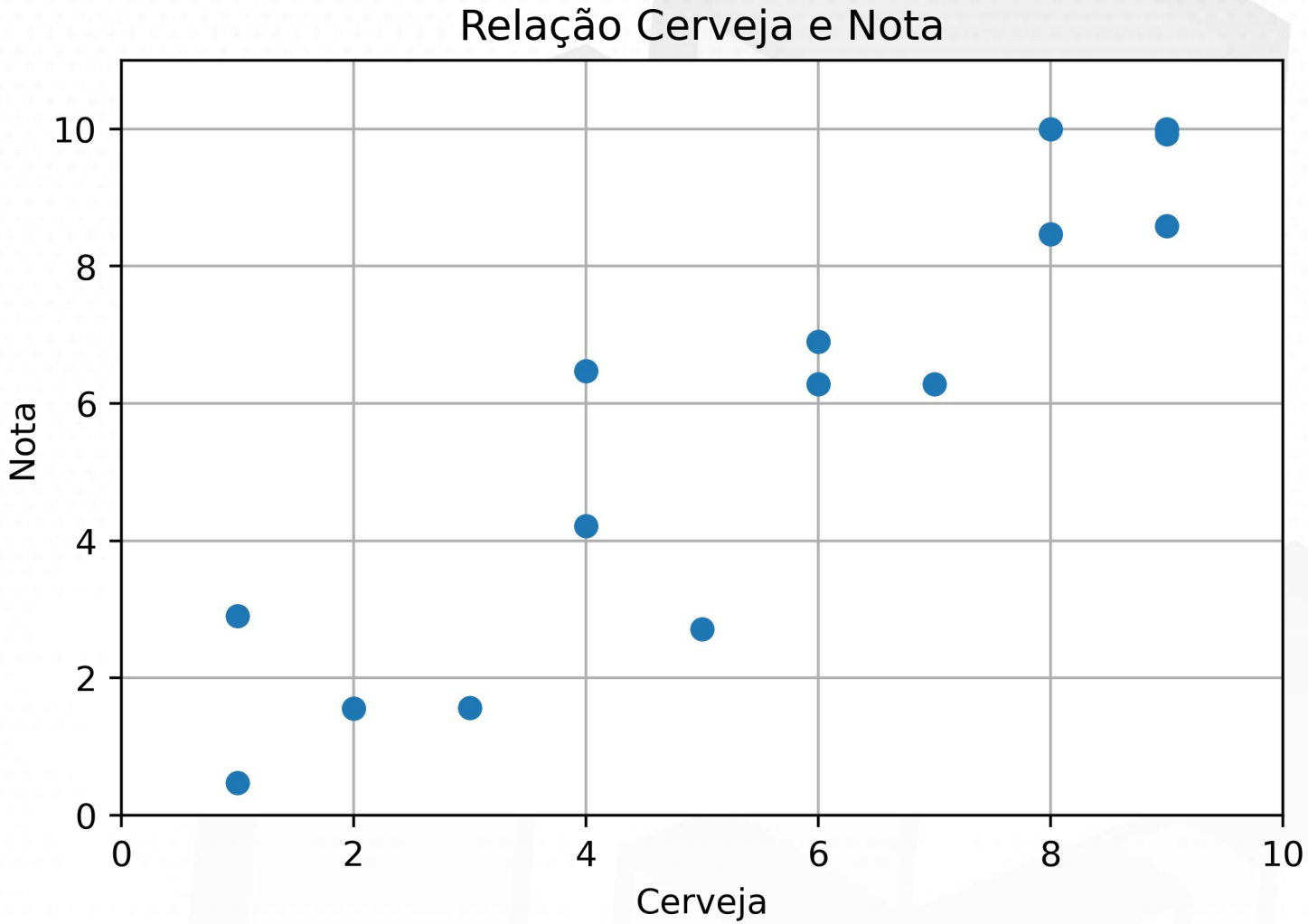
- Quantidade de vendas
- Receita presumida
- Valor de crédito
- Precificação de imóvel
- Volume de chuva

scikit learn

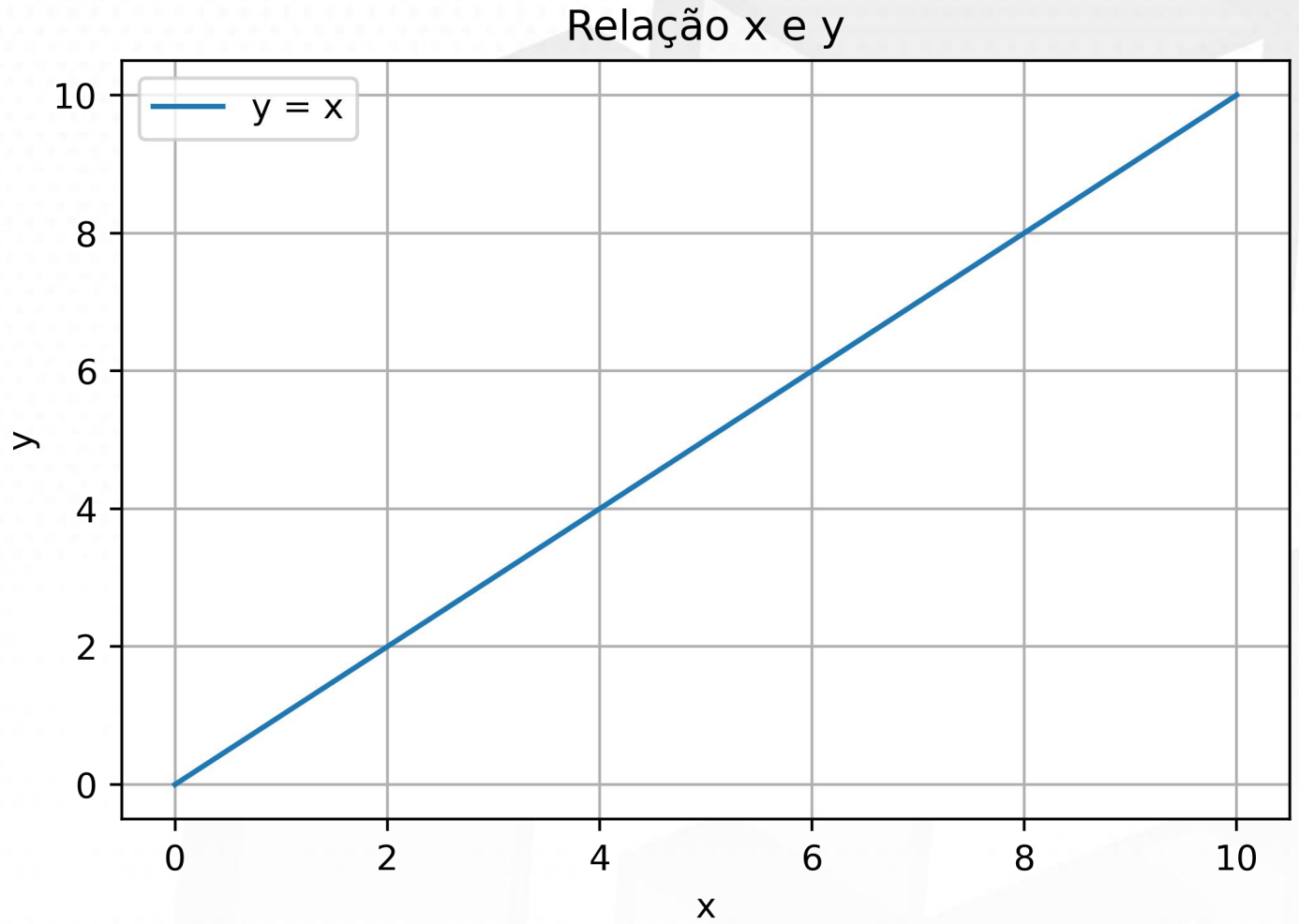


Métodos

Regressão Linear

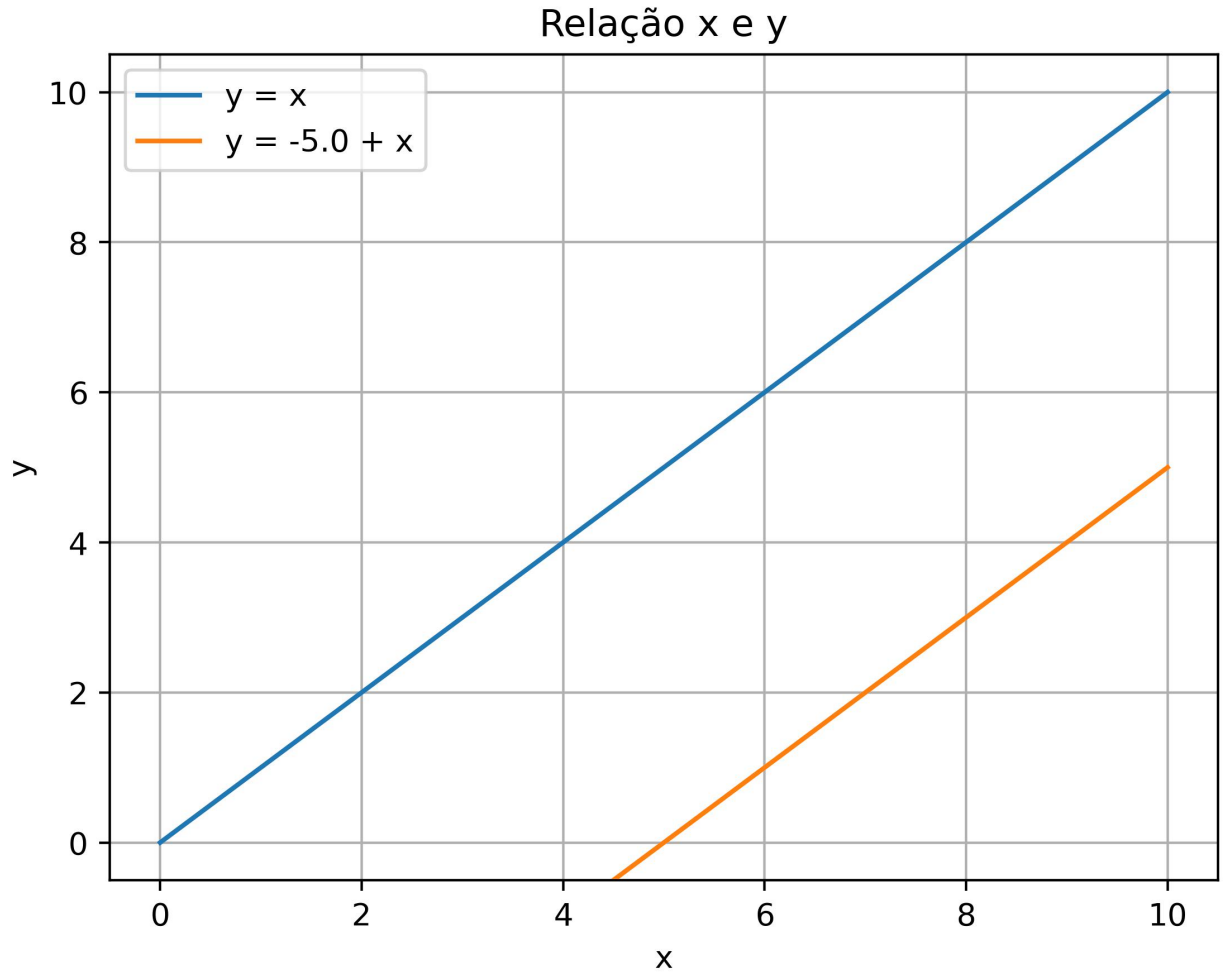


Regressão Linear



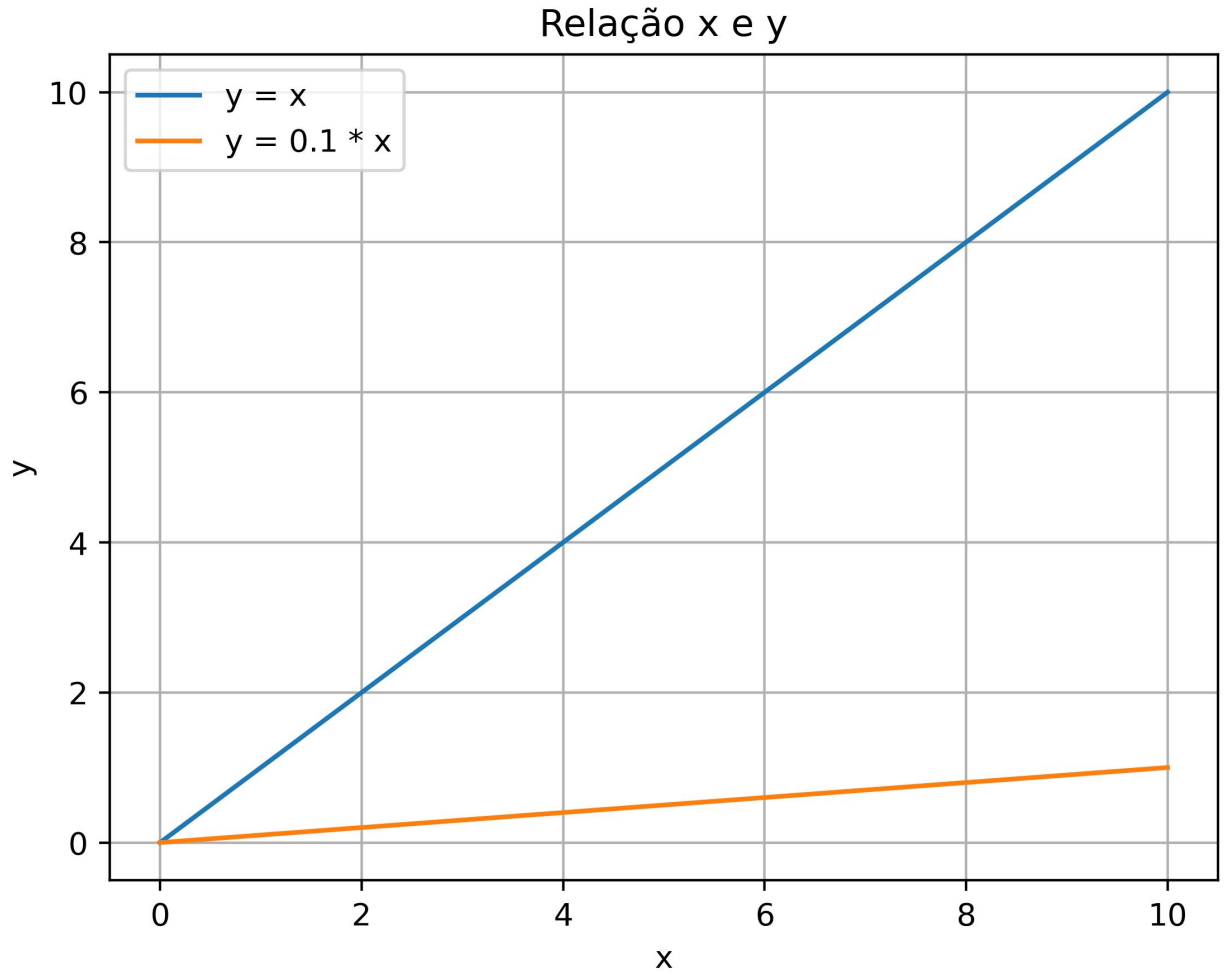
Regressão Linear

$$y = a + x$$



Regressão Linear

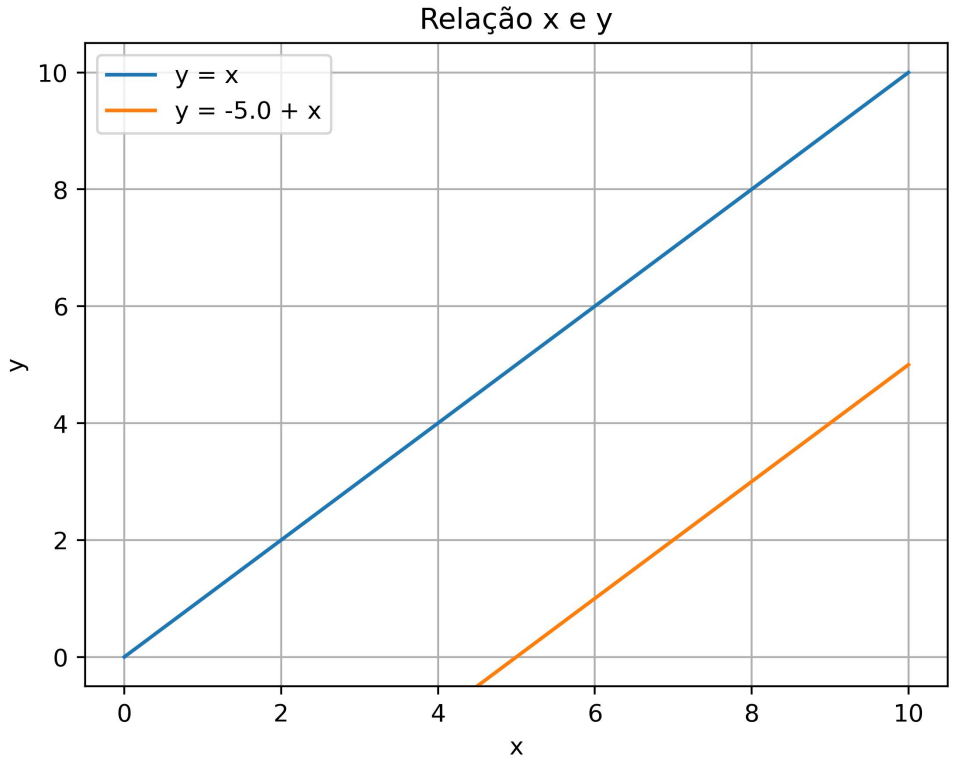
$$y = b.x$$

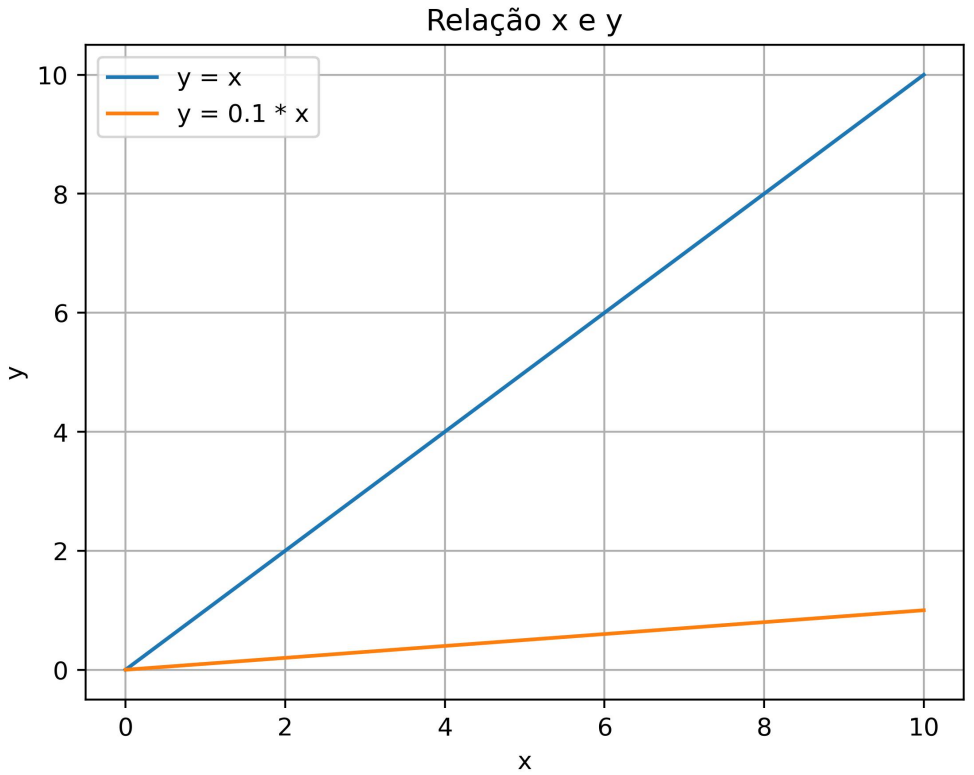


Regressão Linear

$$y = a + x$$

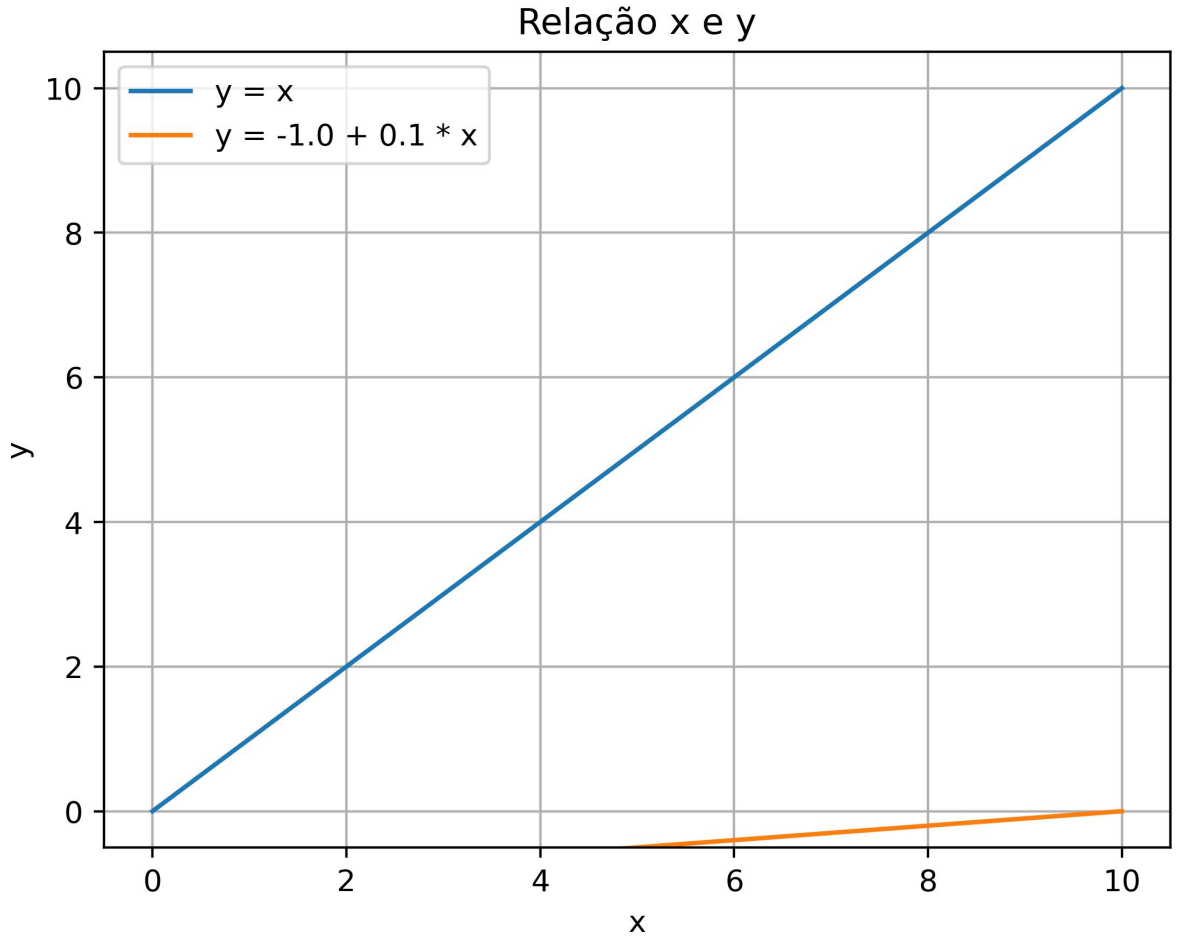
$$y = b.x$$



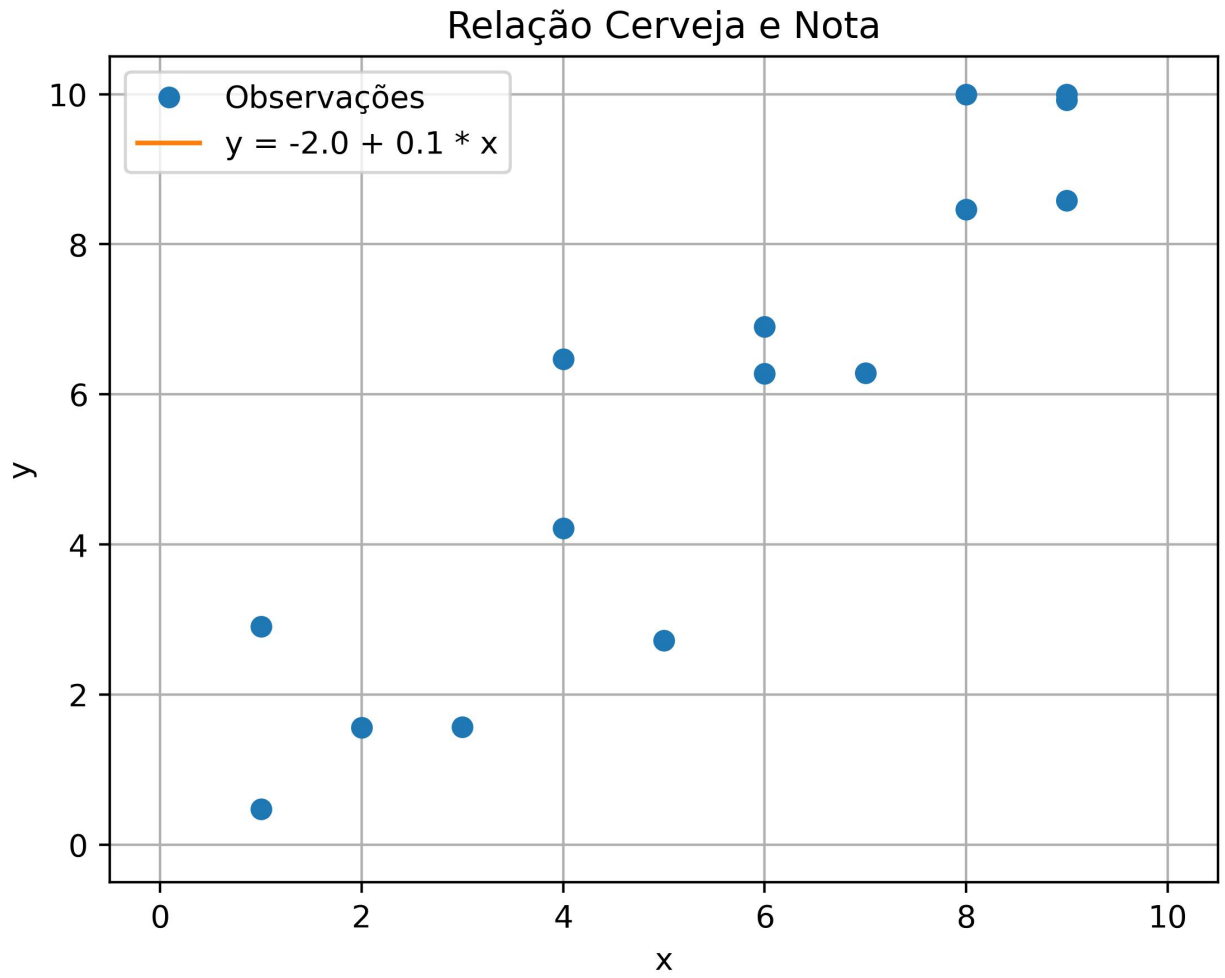


Regressão Linear

$$y = a + b.x$$



Regressão Linear



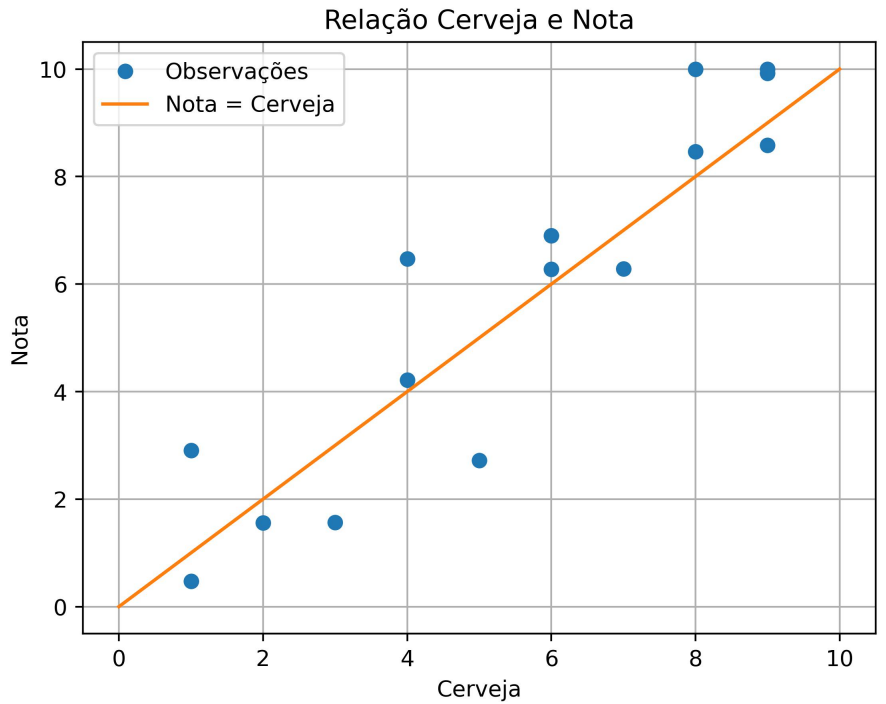
Regressão Linear

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

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

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

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



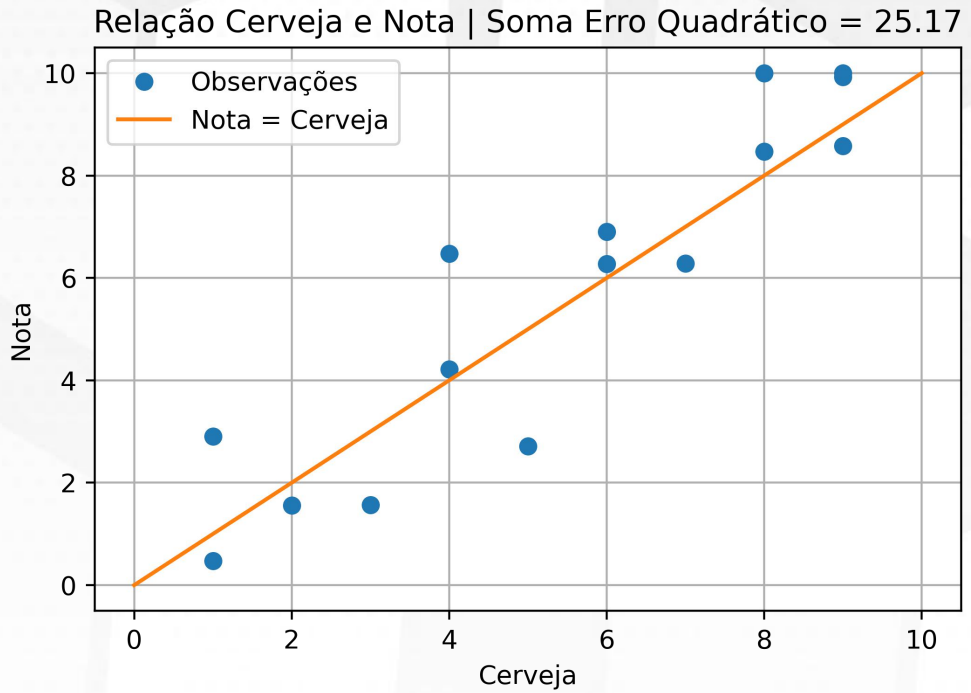
Regressão Linear

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

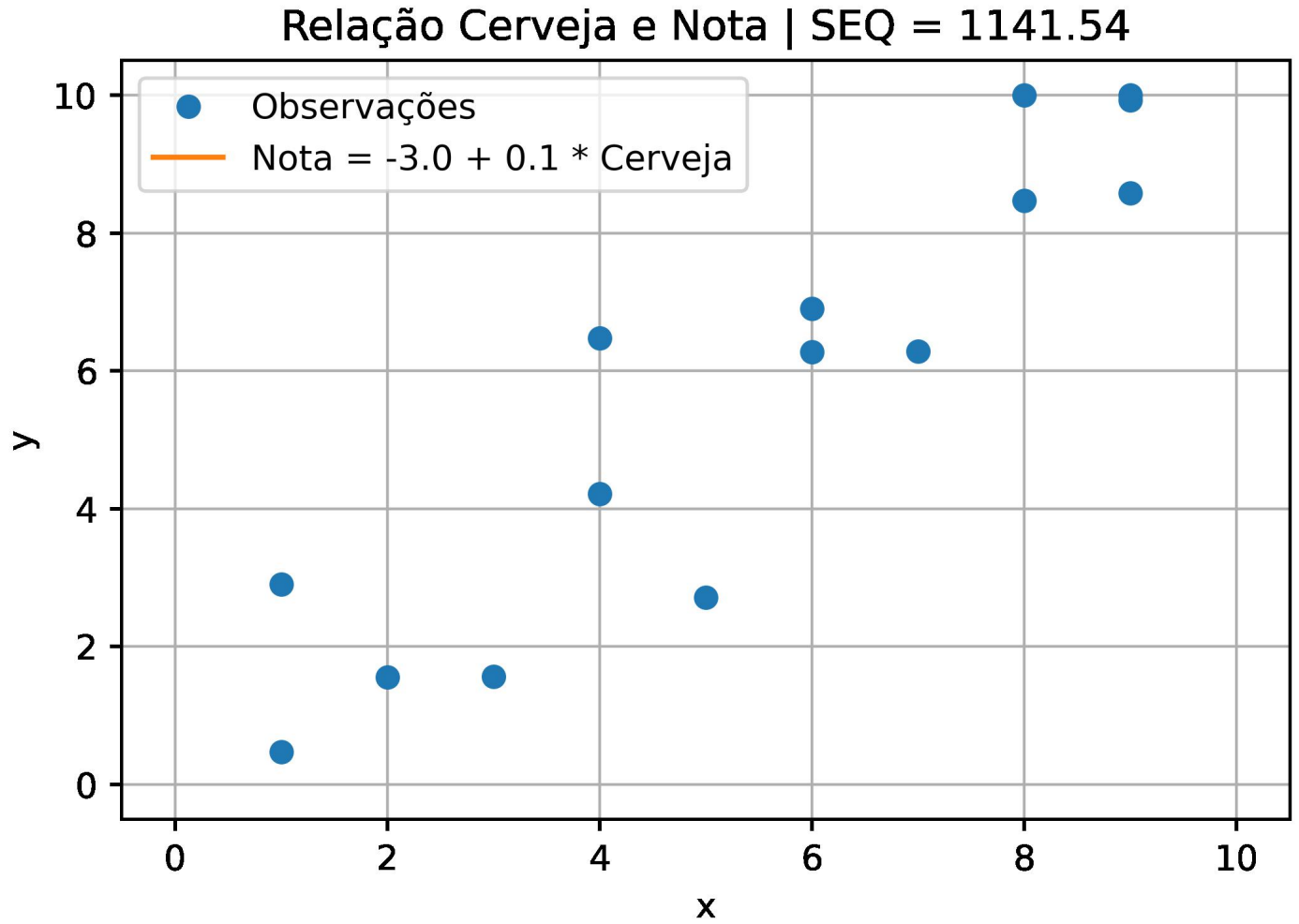
$$\text{Erro} = y - \hat{y}$$

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

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

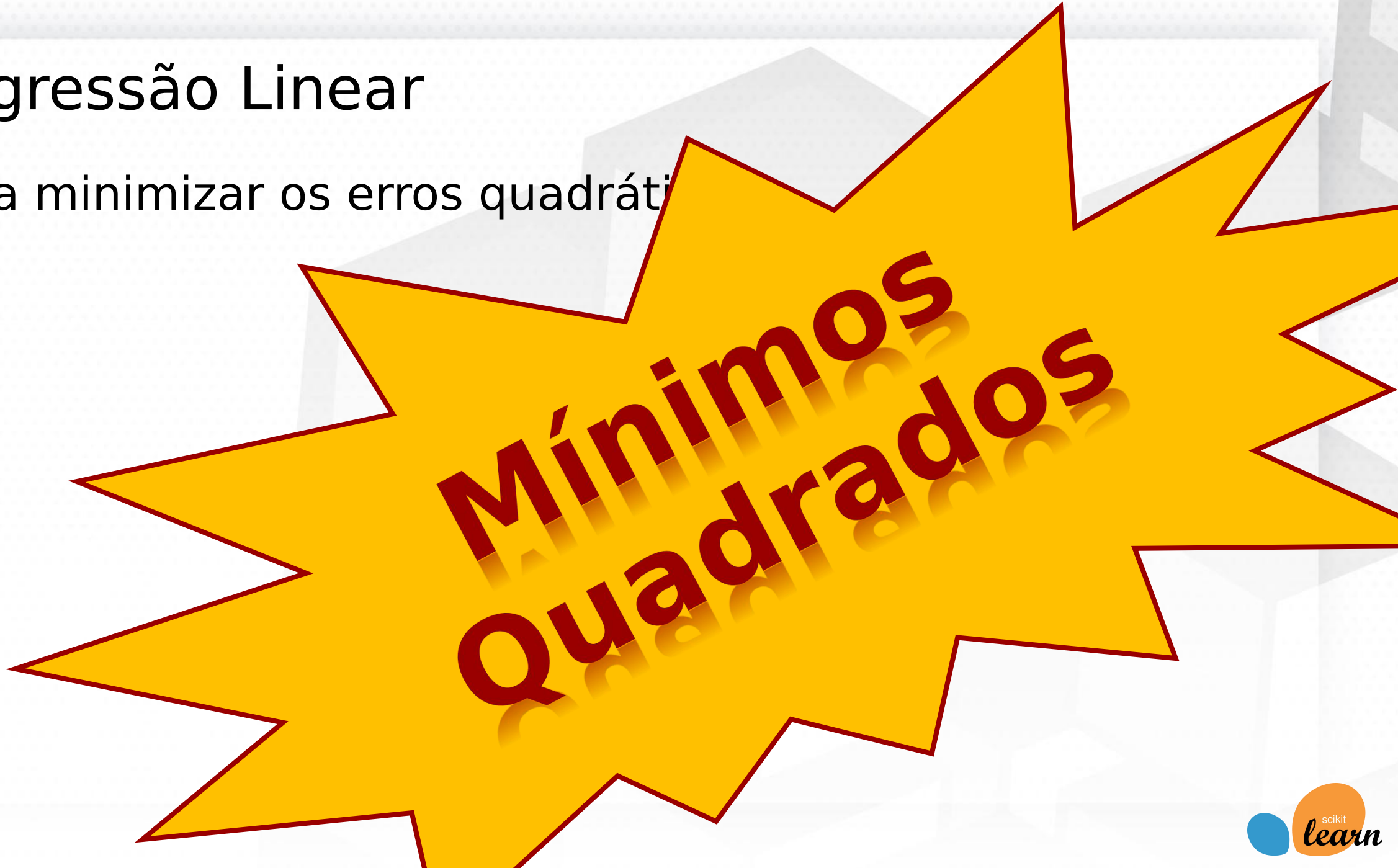


Regressão Linear



Regressão Linear

Bora minimizar os erros quadráti



Regressão Linear

Bora minimizar os erros quadráticos?

$$\hat{y} = a + b.x \Leftrightarrow N\hat{o}ta = a + b.\text{Cerveja}$$

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

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

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

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

scikit learn

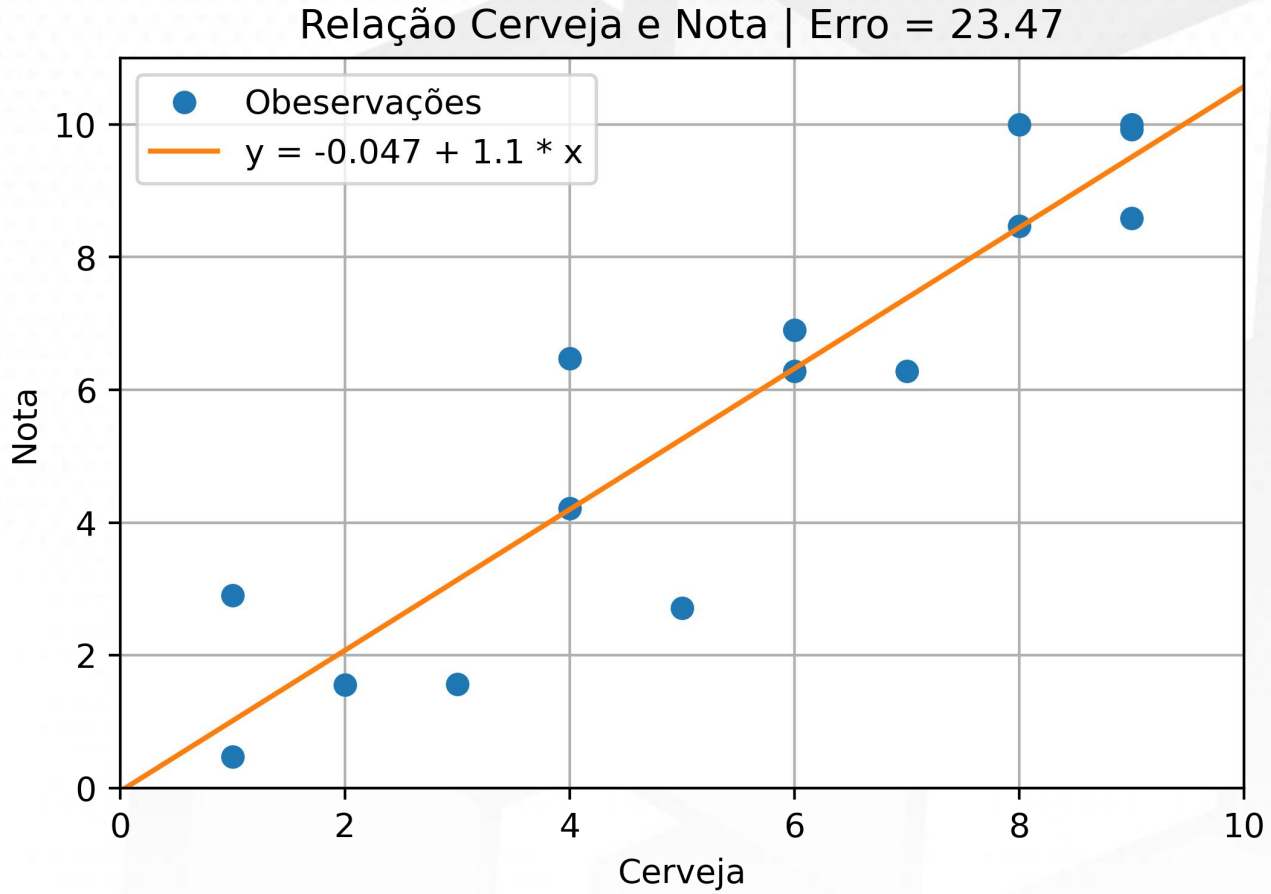
Regressão Linear

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 Múltipla

E se tivermos afim de utilizar mais variáveis?

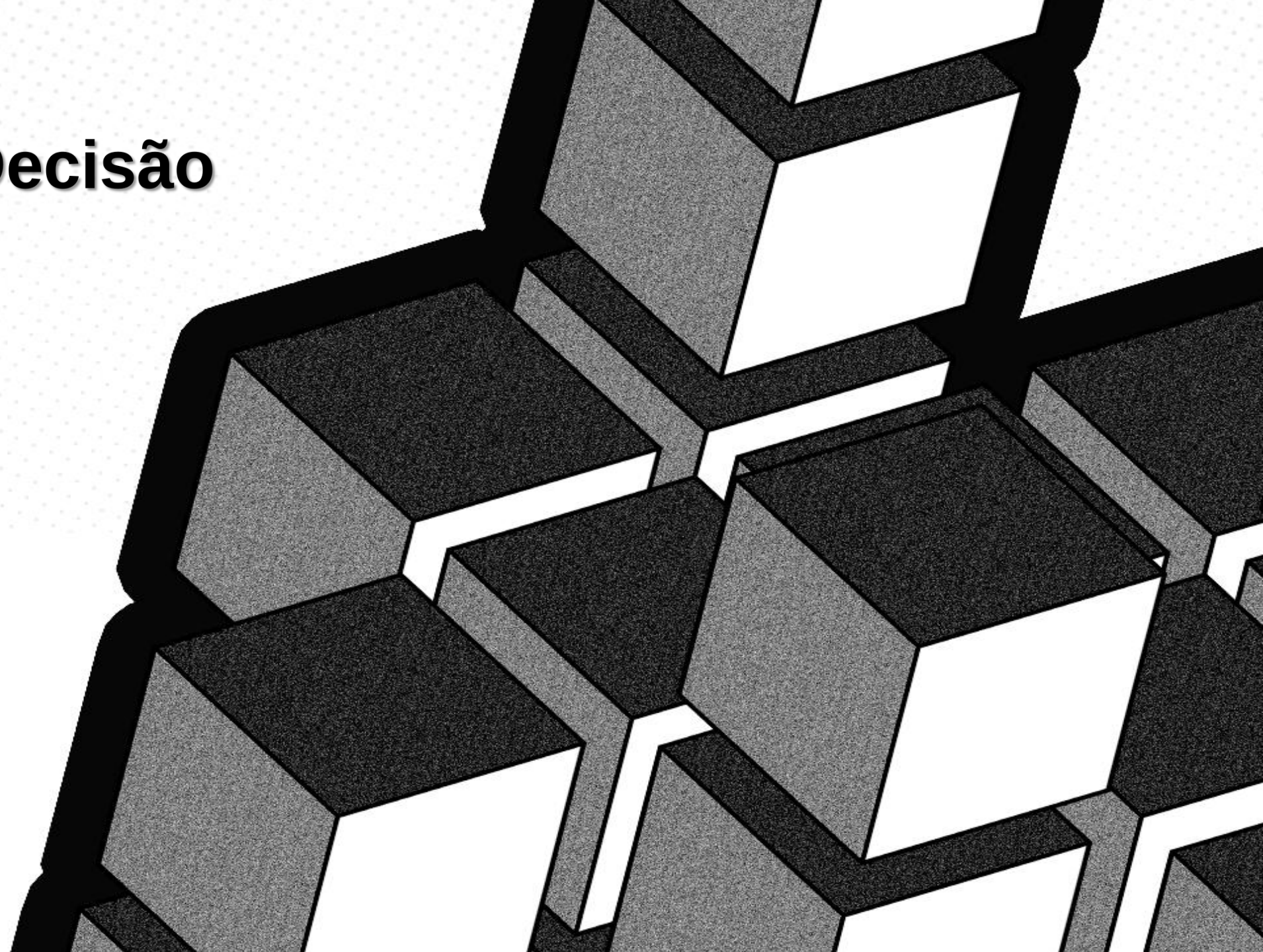
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

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

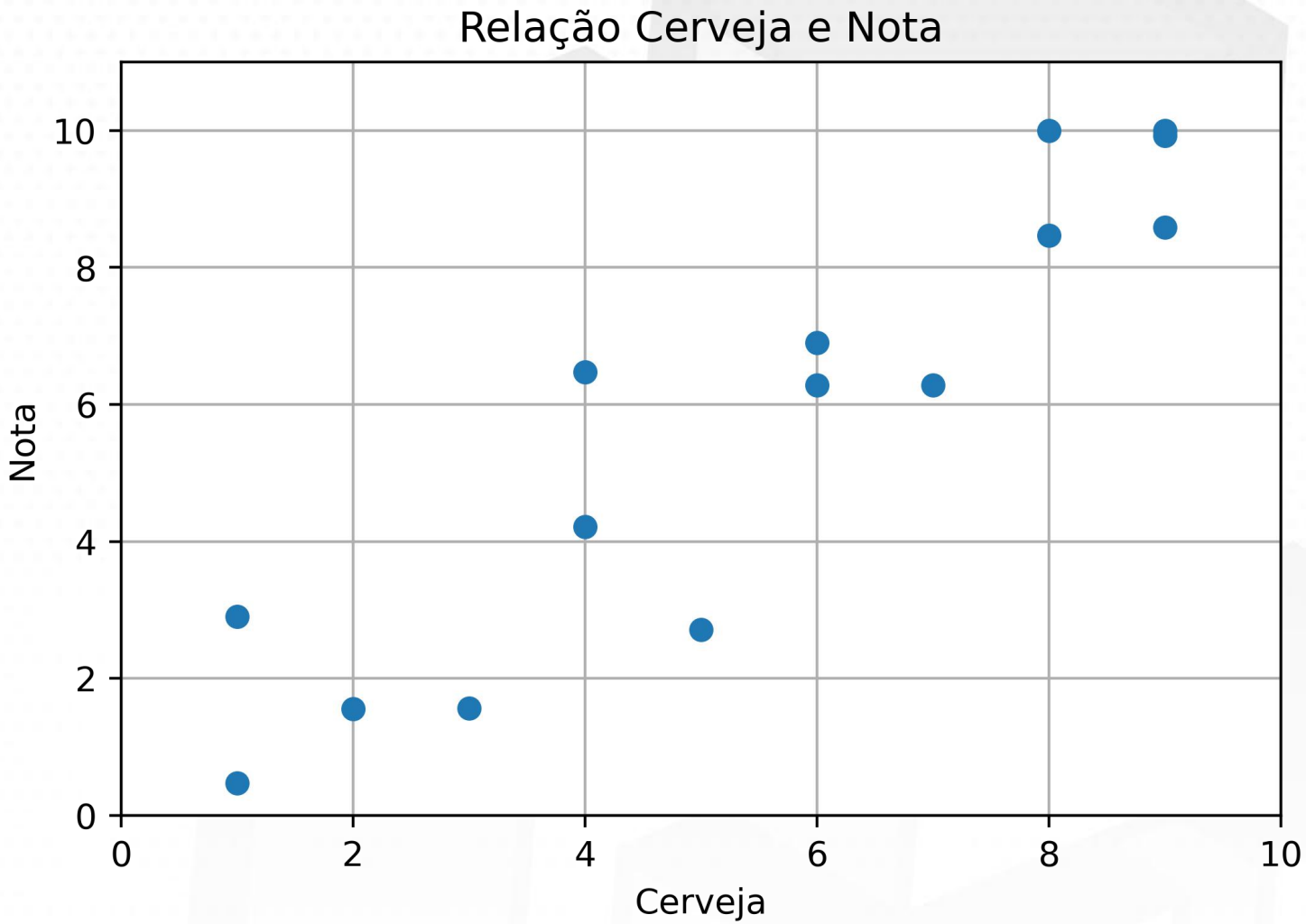
$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

scikit learn

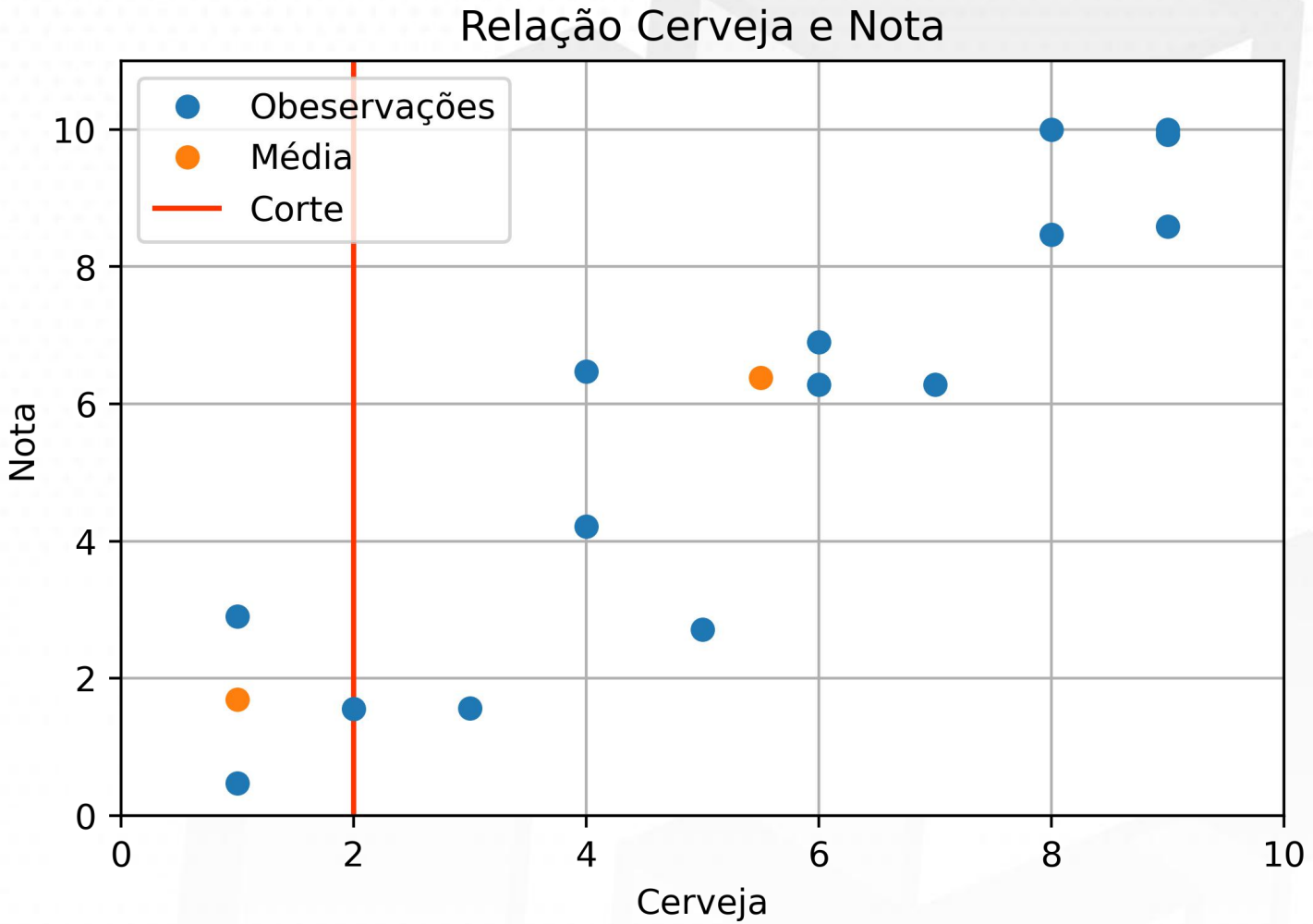
Árvore de Decisão



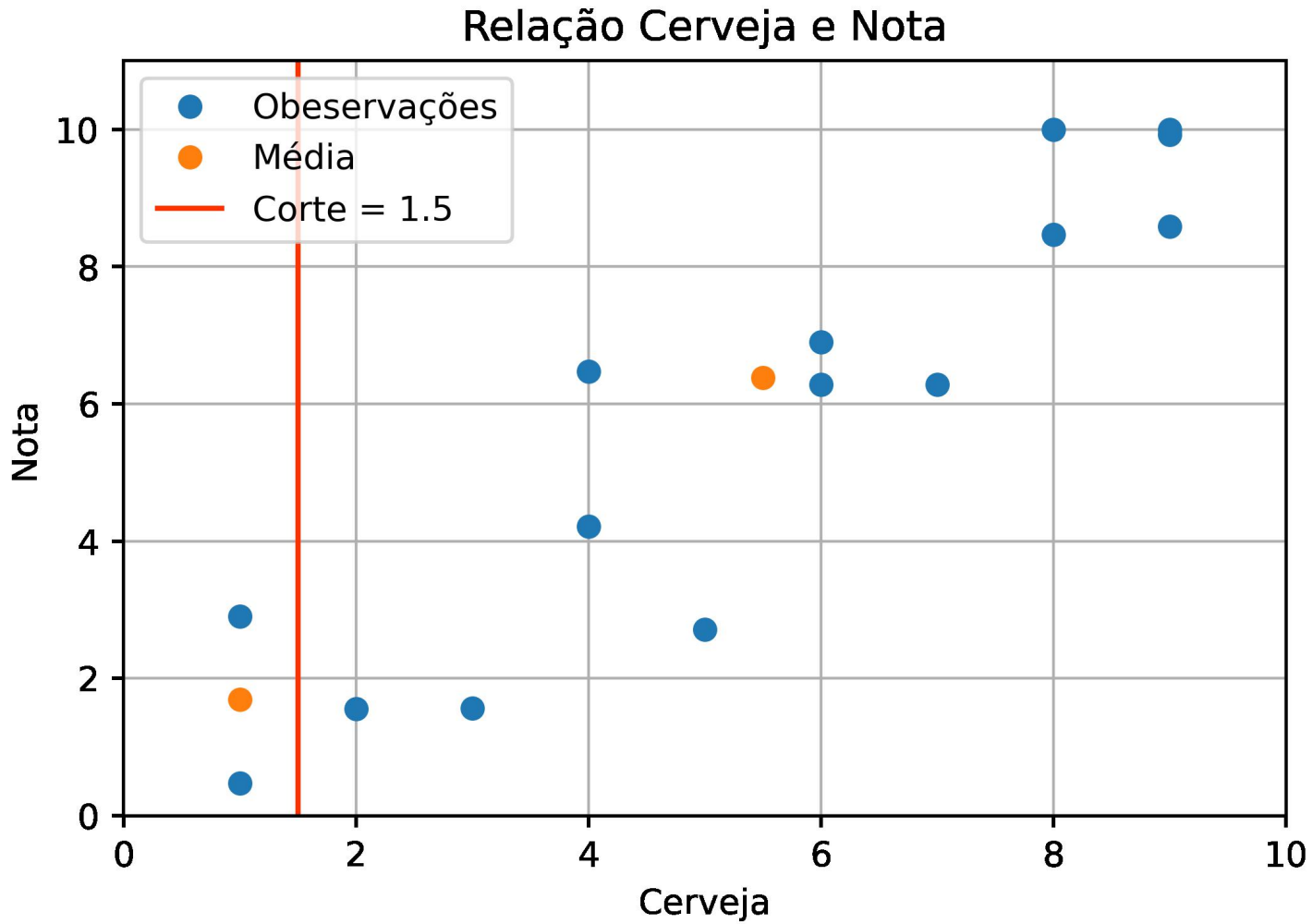
Voltando à Árvore de Decisão



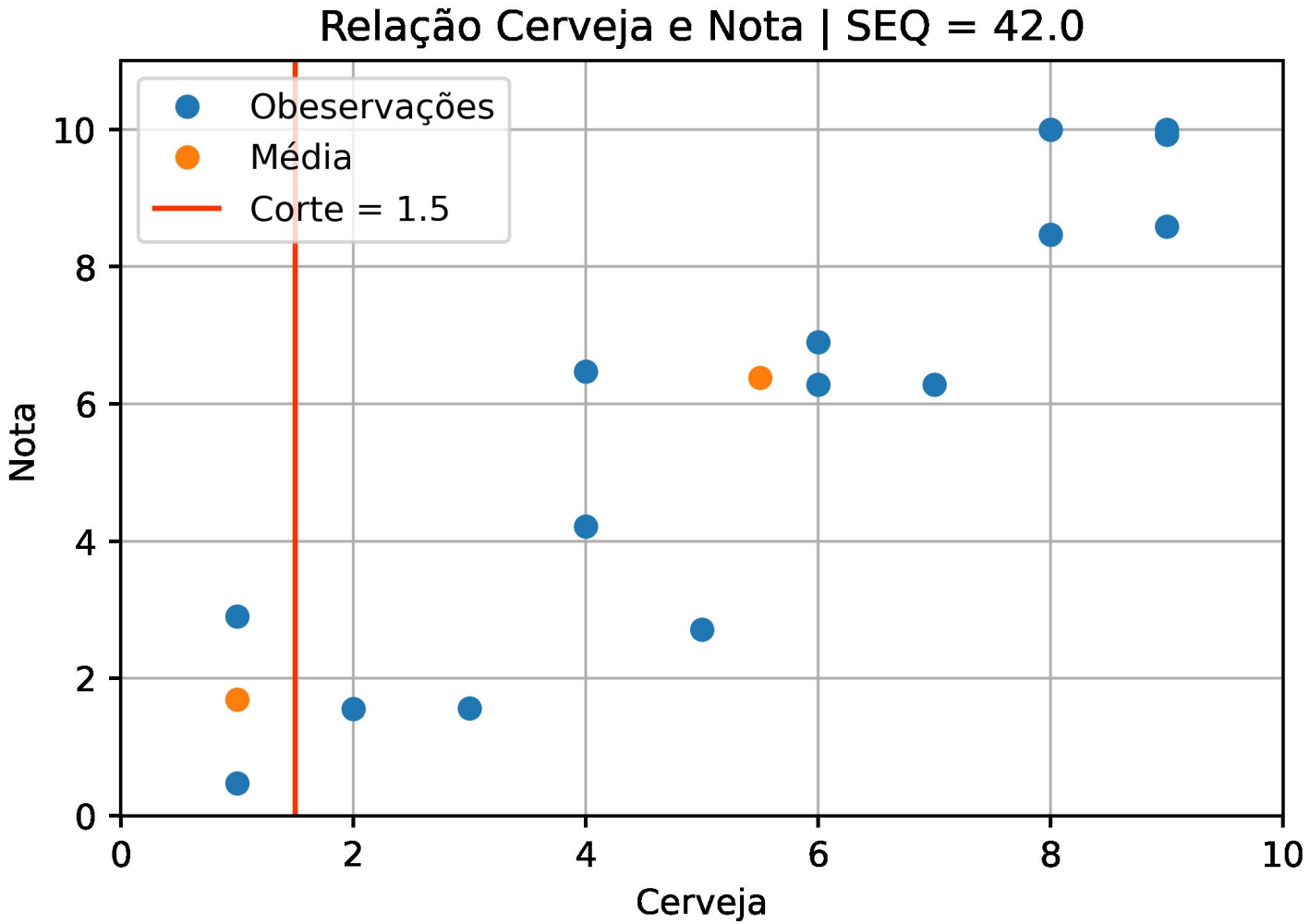
Voltando à Árvore de Decisão



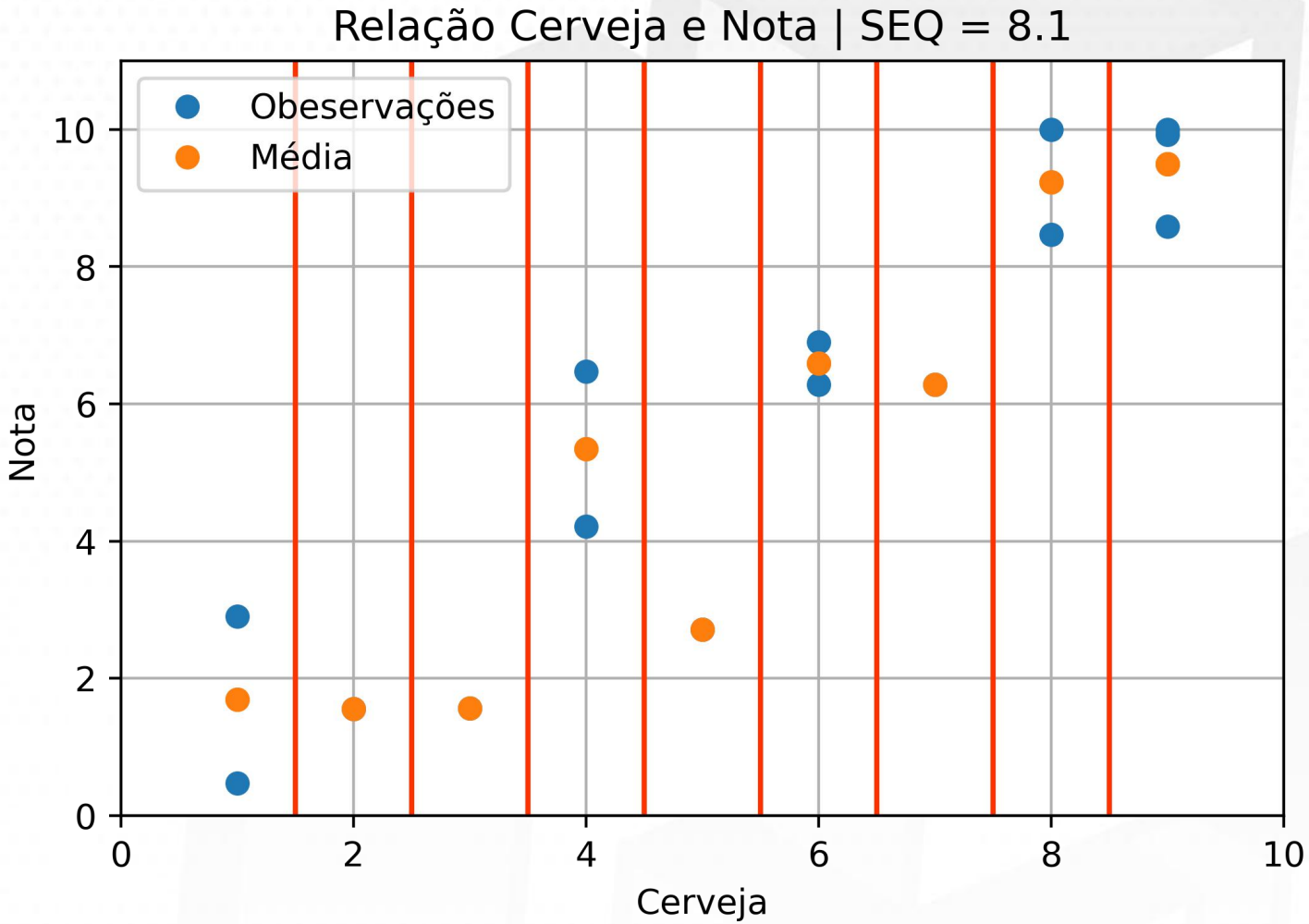
Voltando à Árvore de Decisão



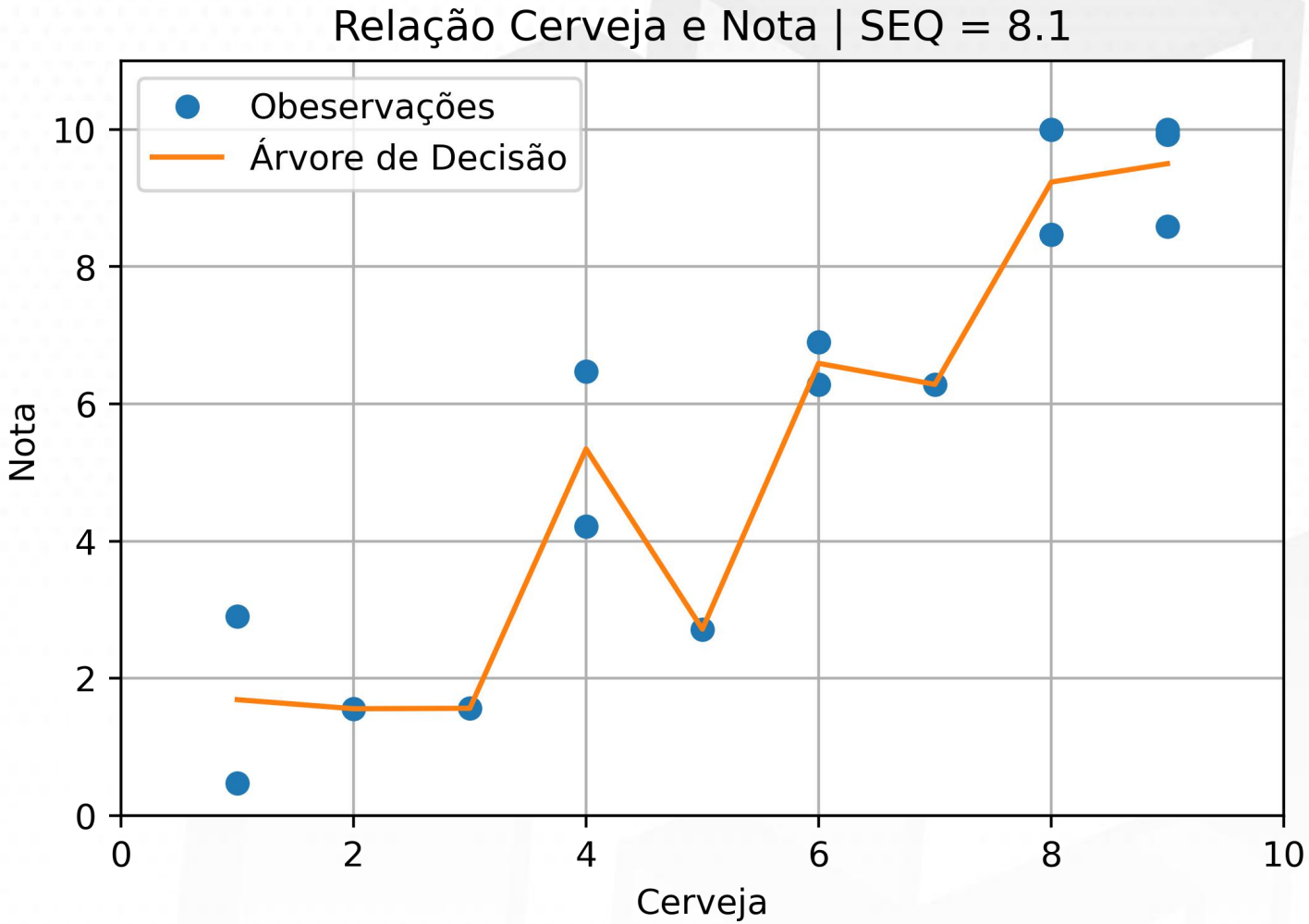
Árvore de Decisão



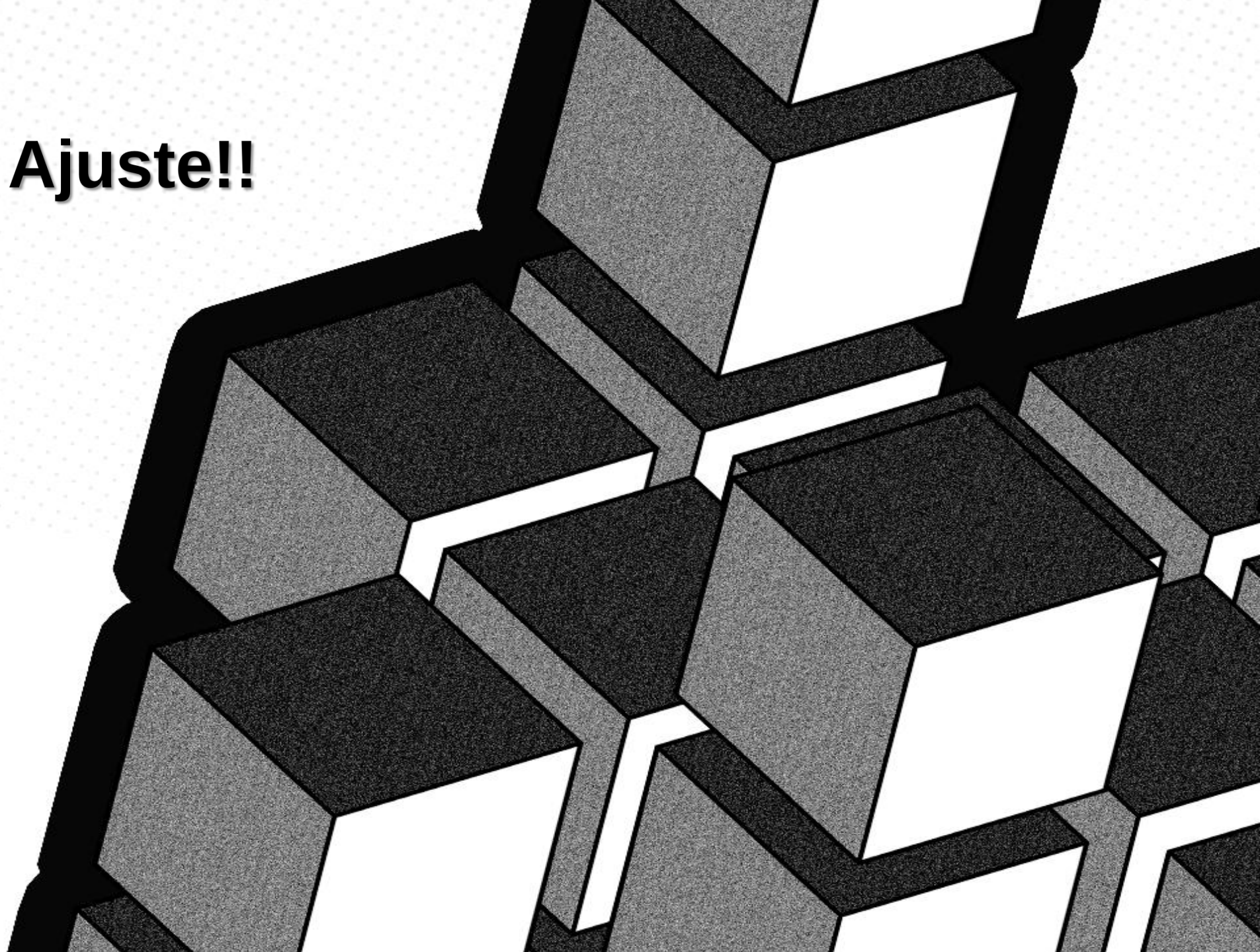
Árvore de Decisão



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

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

scikit learn

Agora é sentar o

dedo

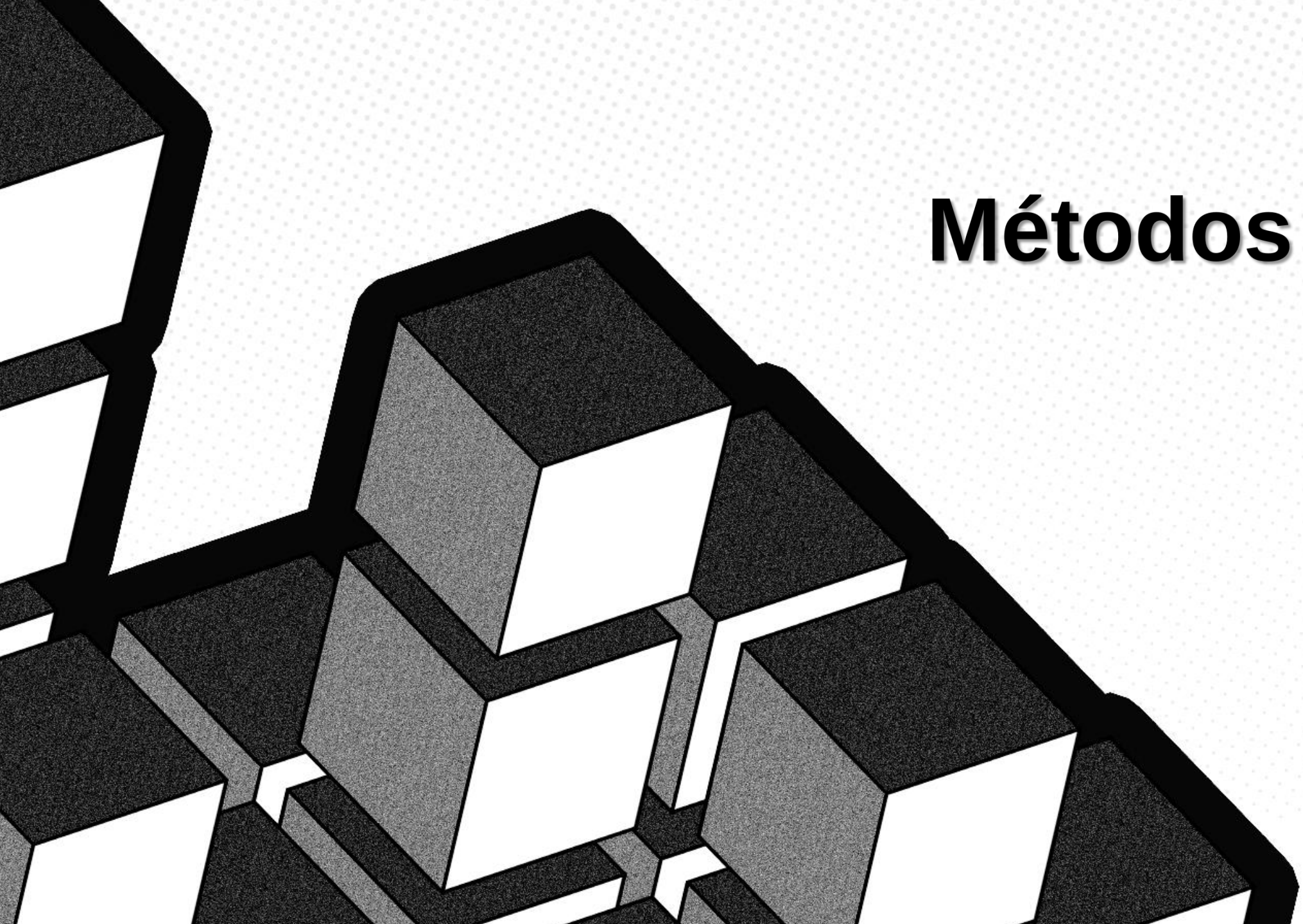
no código

Classificação

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

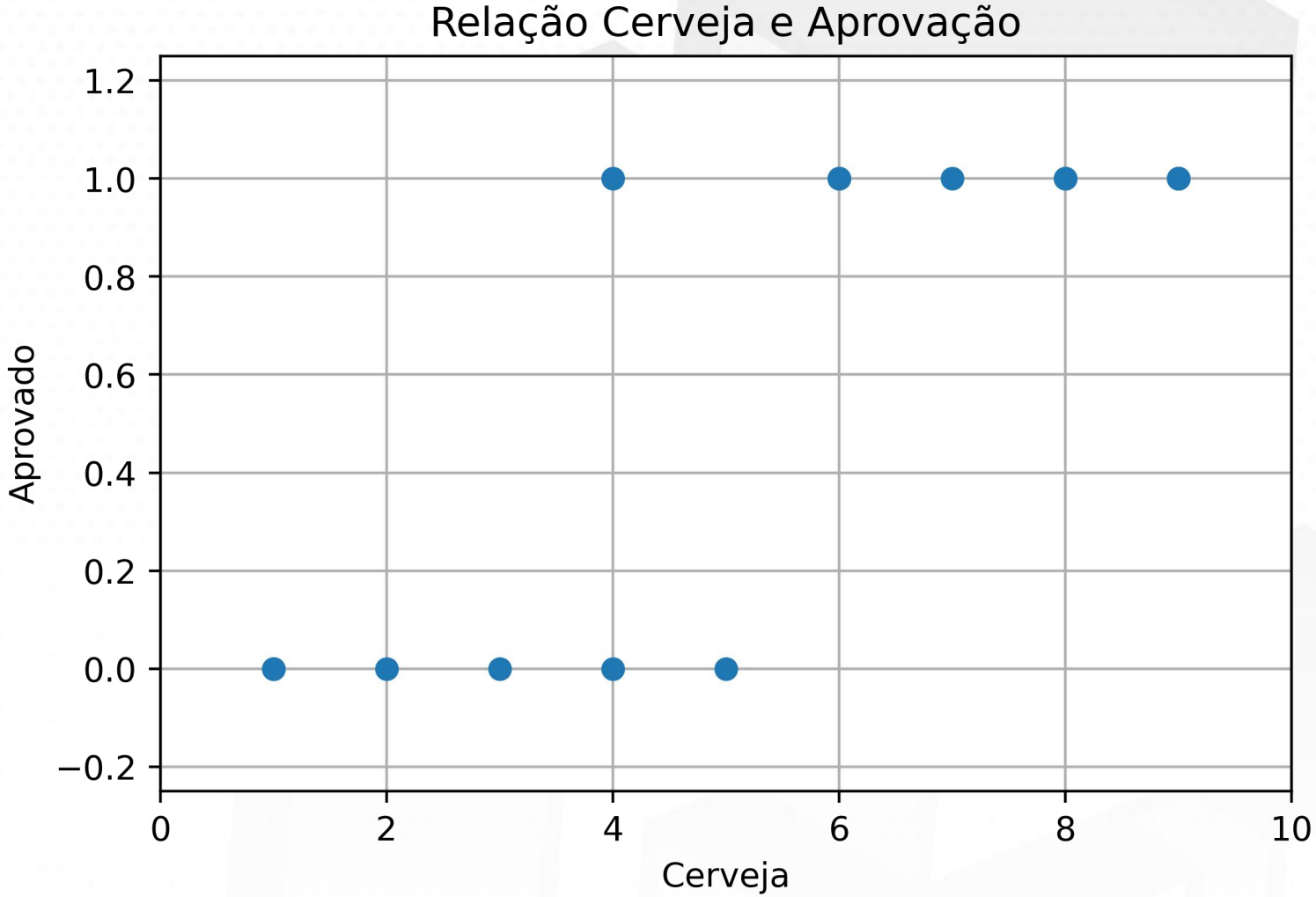
Por exemplo:

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

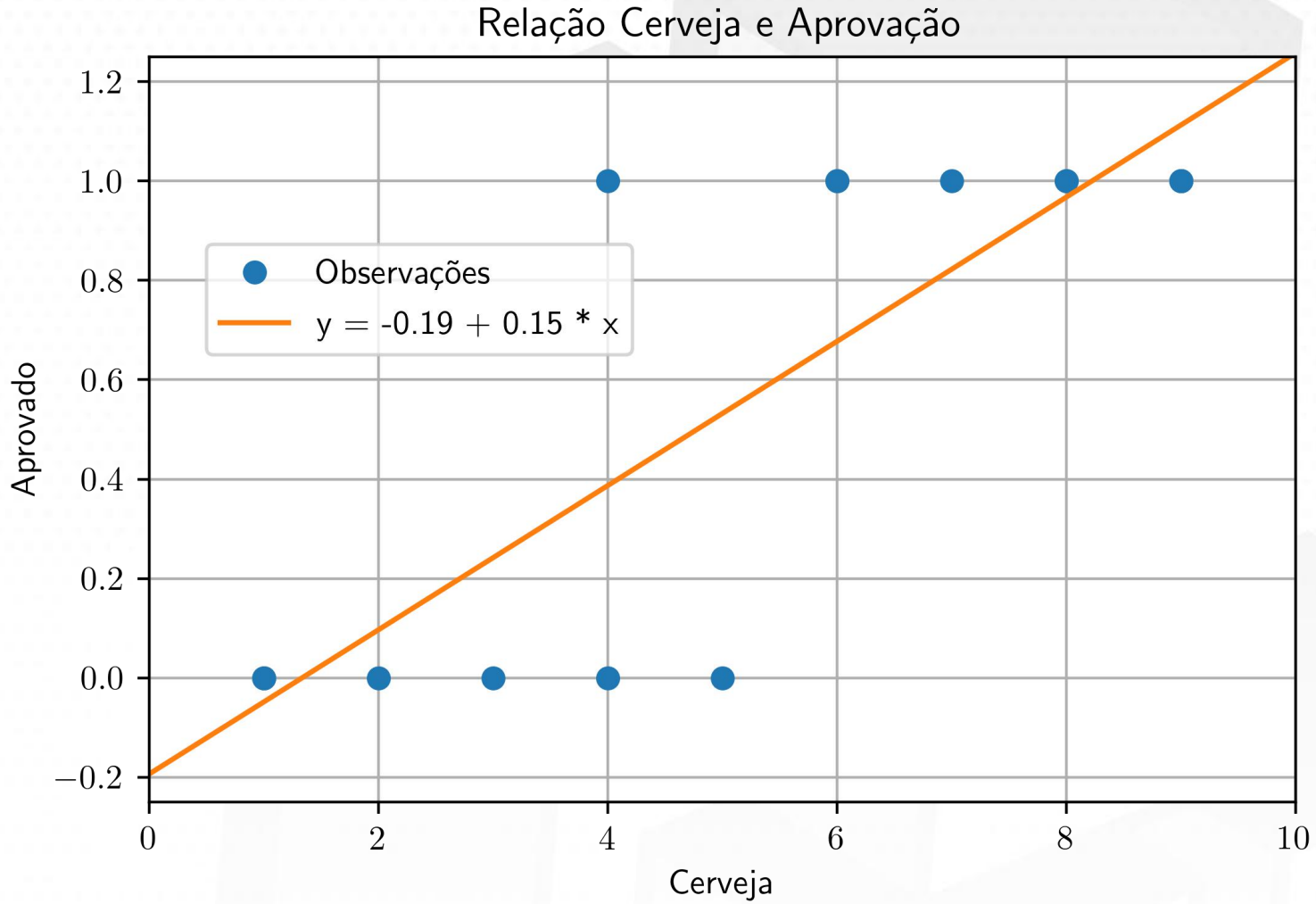


Métodos

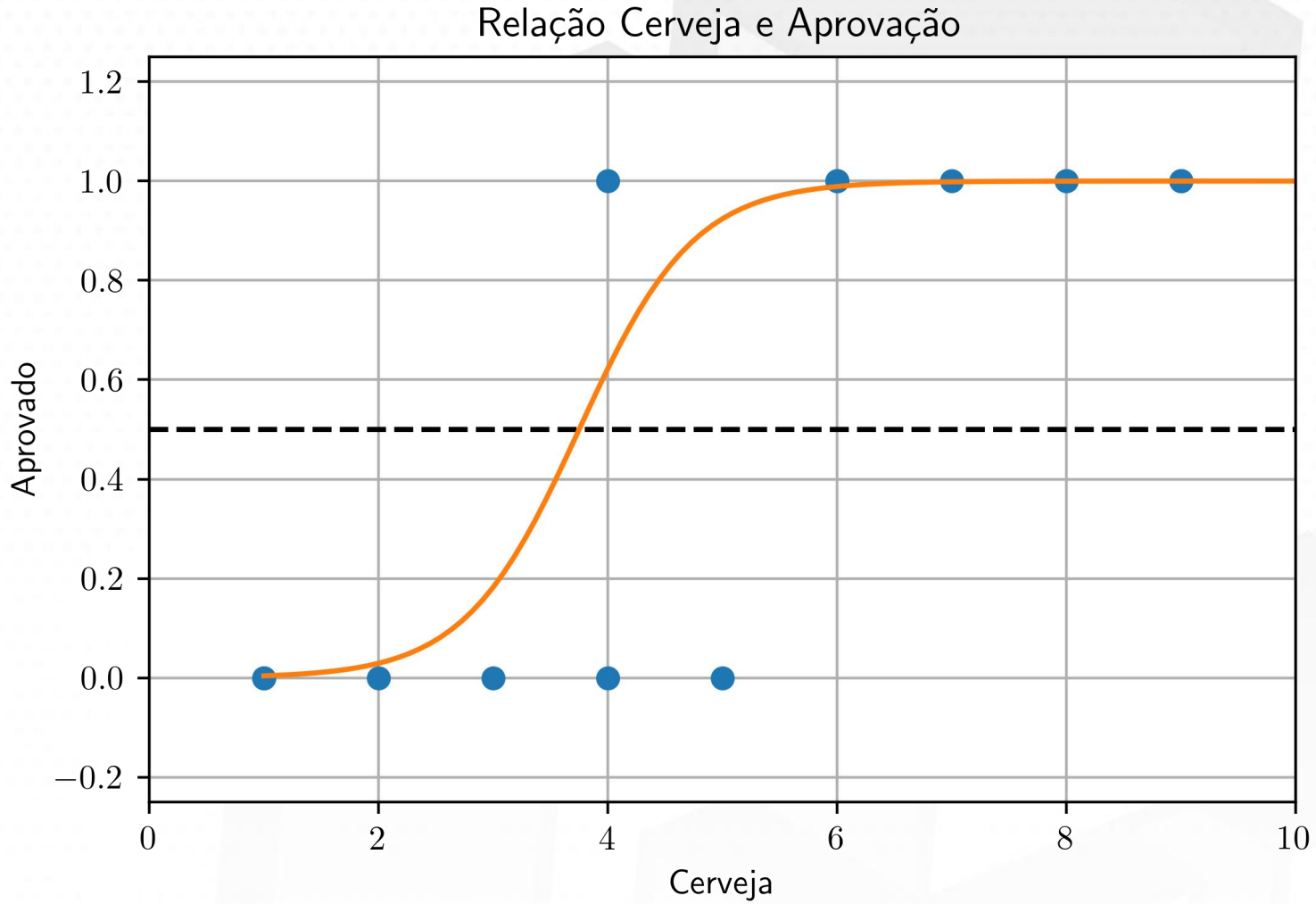
Regressão Logística



Regressão Logística



Regressão Logística

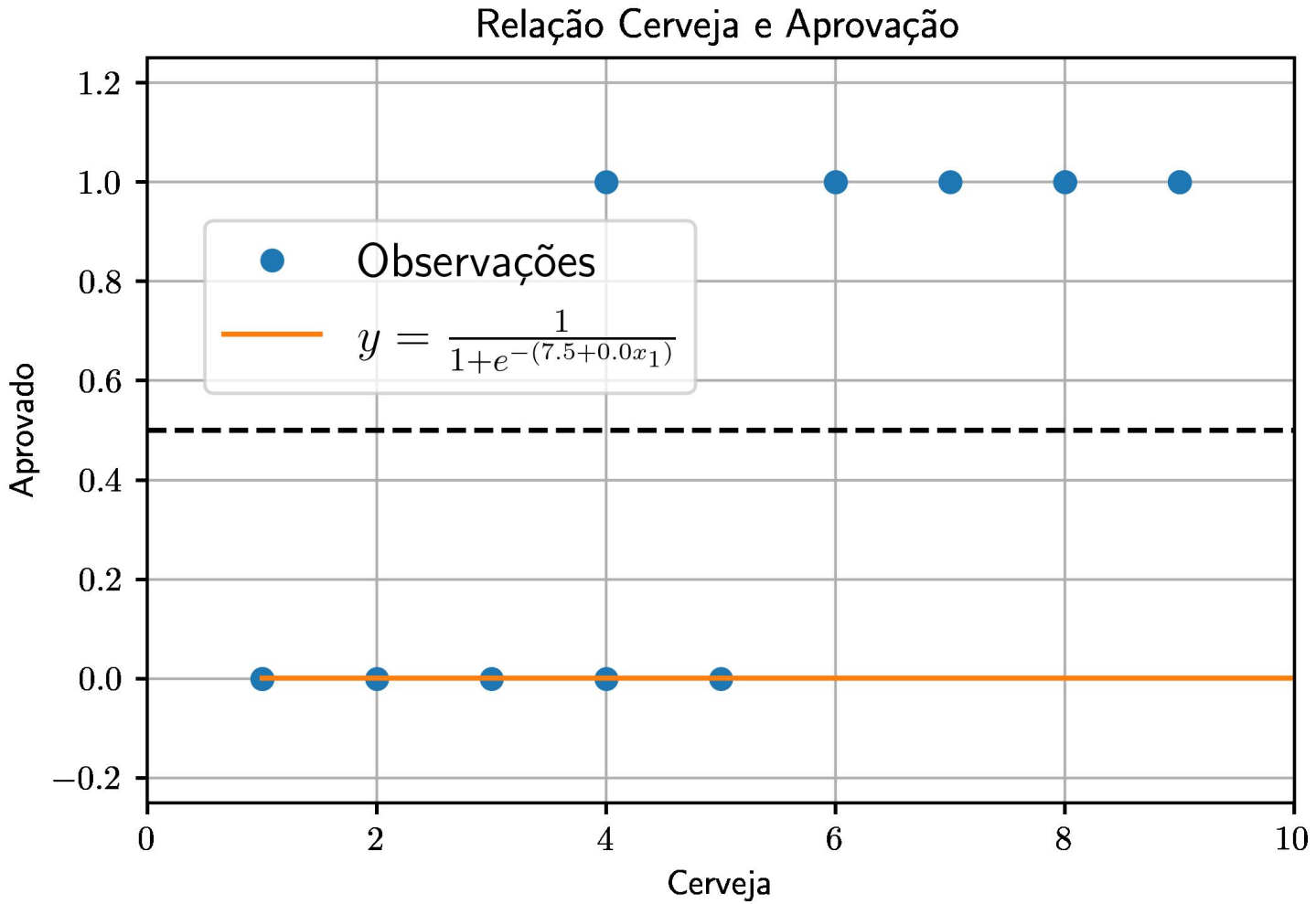


Regressão Logística

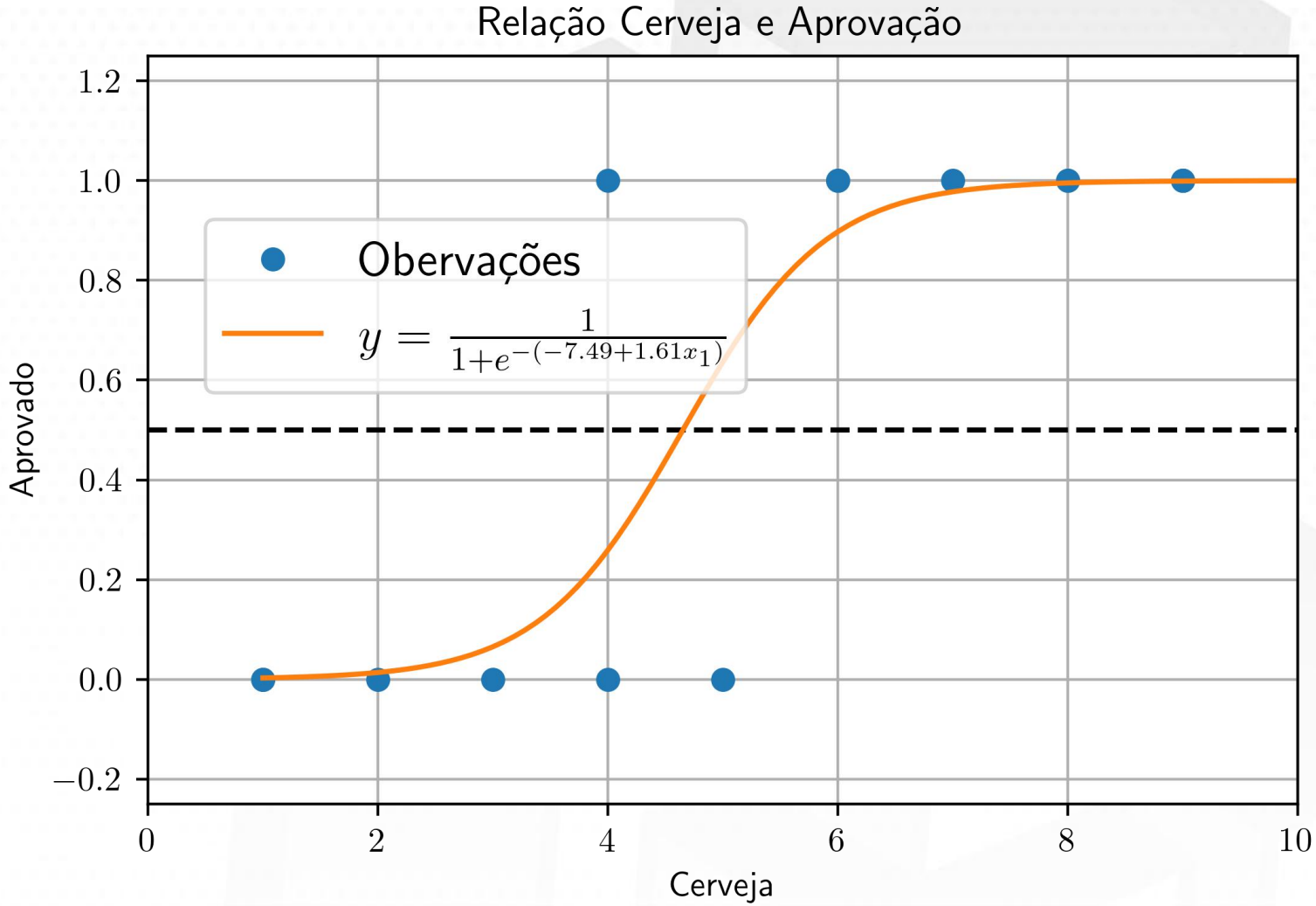
$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}}$$

scikit learn

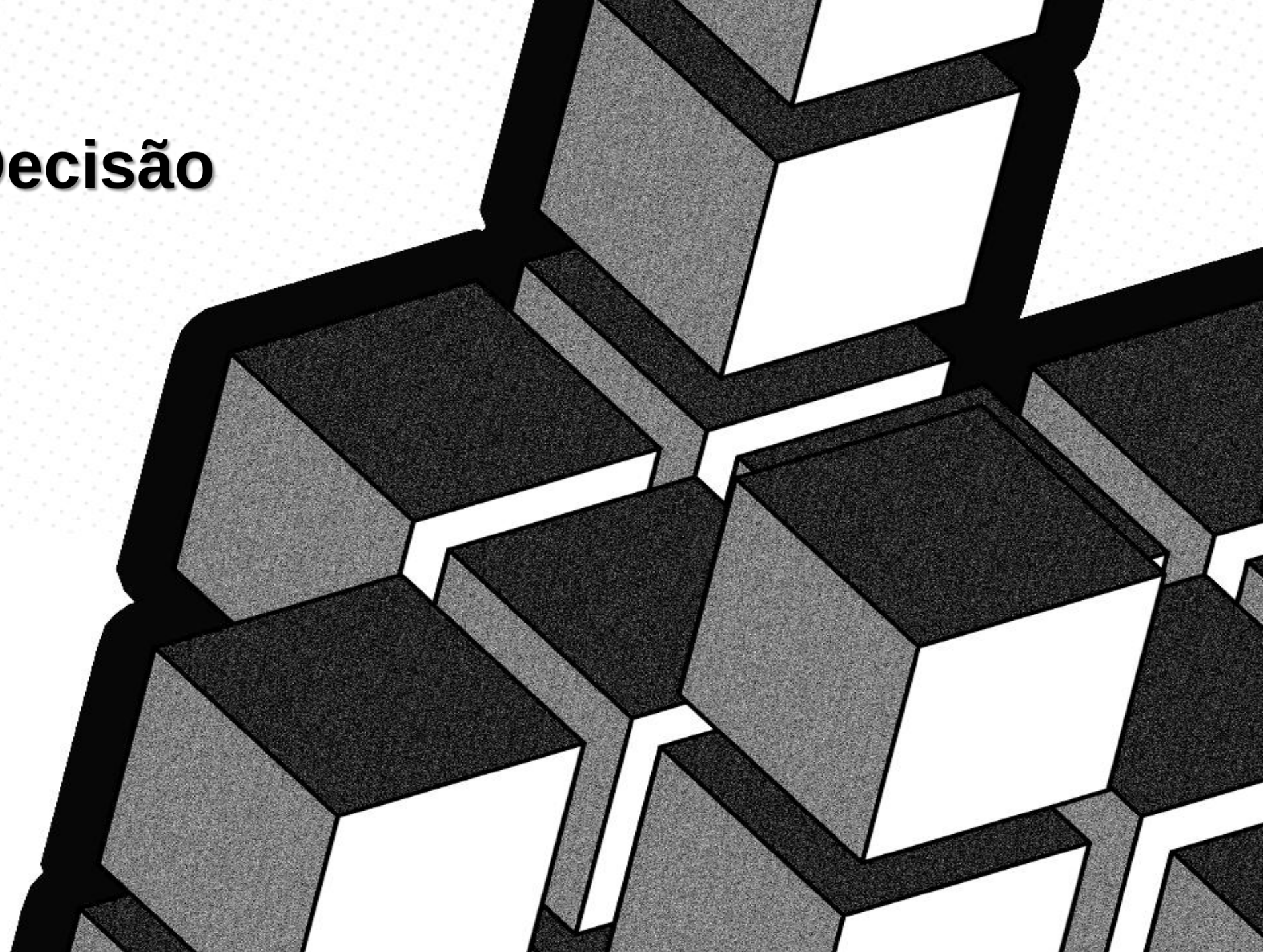
Regressão Logística



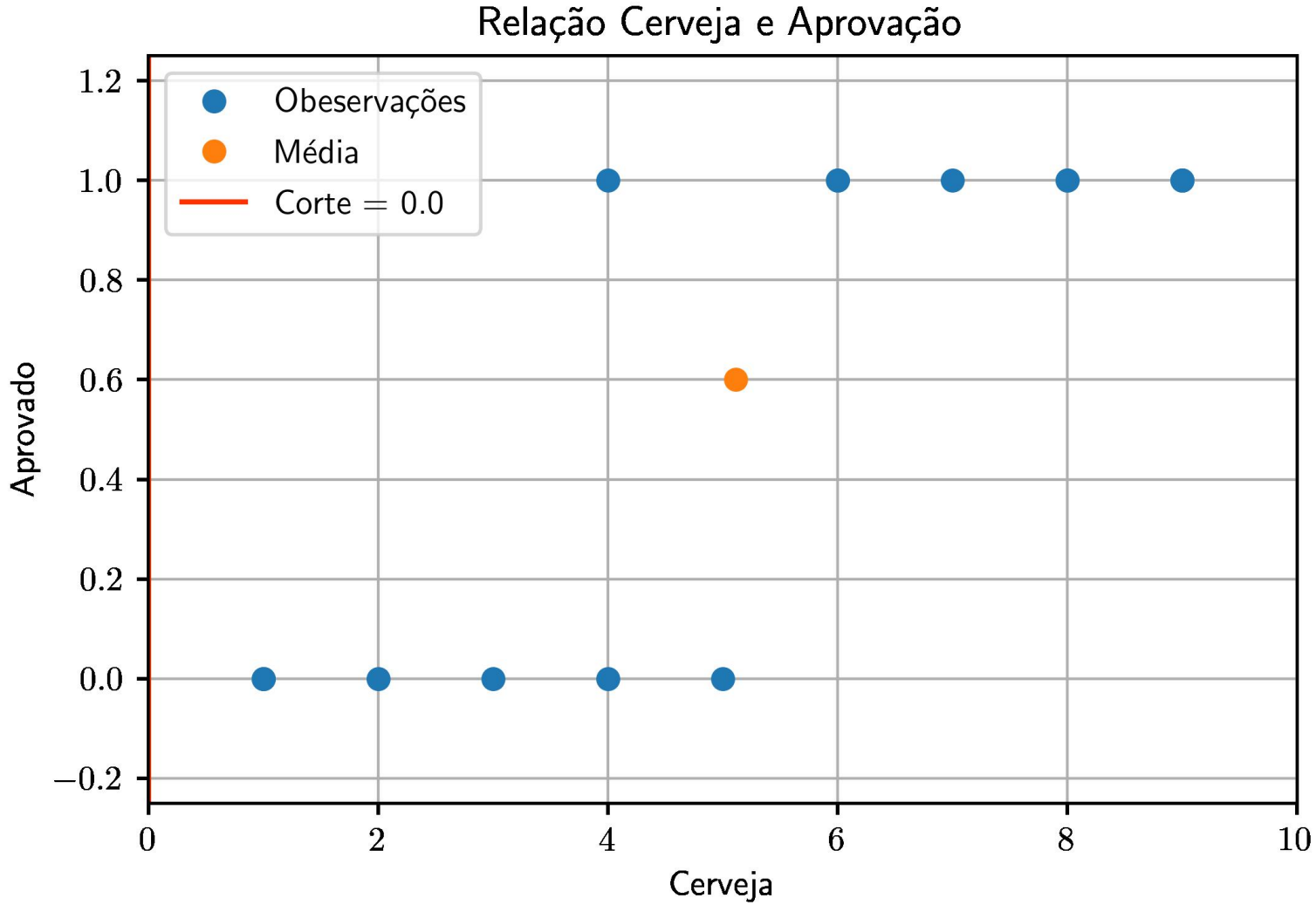
Regressão Logística



Árvore de Decisão



Árvore de Decisão



Árvore de Decisão

Anteriormente utilizamos a Soma dos Erros Quadráticos.

E Agora? Vamos usar qual medida?

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

scikit learn

Árvore de Decisão

Há outras medidas?

Sim!

Entropia!!

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

De uma maneira mais genérica

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

scikit learn

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

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

scikit learn

Agora é sentar o

dedo

no código

Clustering

Em problemas de Clustering, buscamos encontrar grupos de objetos (instâncias) com características similares, não havendo uma necessidade preditiva:

- Perfís de clientes
- Produtos com mesmas características
- Similaridade em eventos de Log
- Proximidade entre ocorrências

Obrigado

Téo Calvo

teo.bcalvo@gmail.com

/in/teocalvo

/teomewhy