

## Aula 04

**Régressão Logística, Técnicas de Validação, Métricas de Validação e Balanceamento de dados.**

# Agenda

01

Regressão  
Logistica

02

Métricas de Avaliação

03

Técnicas de Validação

04

Técnicas de Balanceamento  
de dados

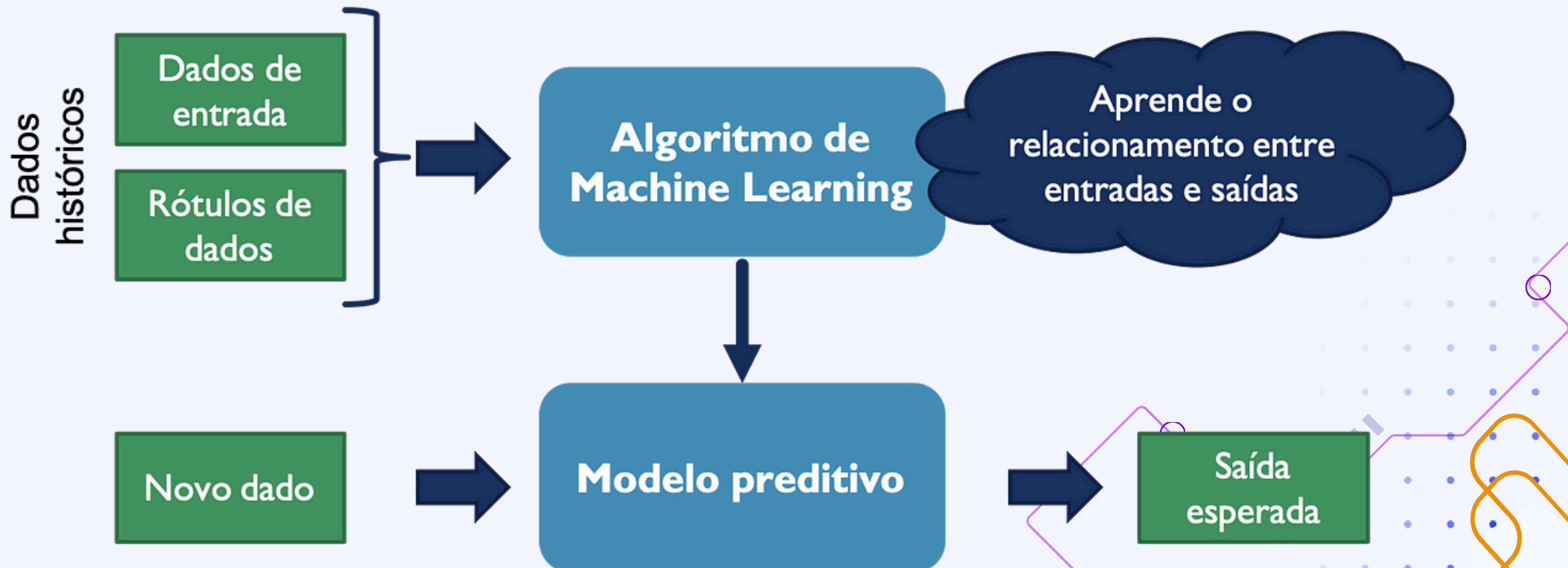
05

Implementando os algoritmos

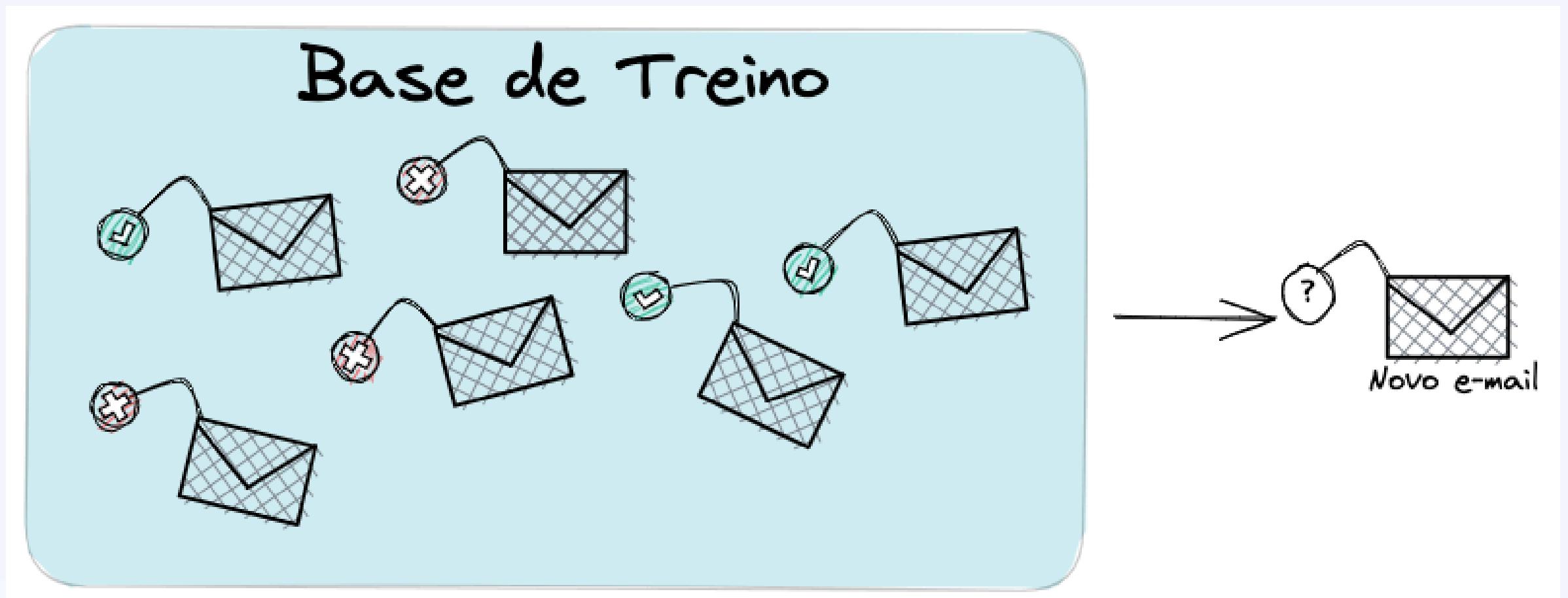
01

# Regressão Logística

# Modelando um fenômeno



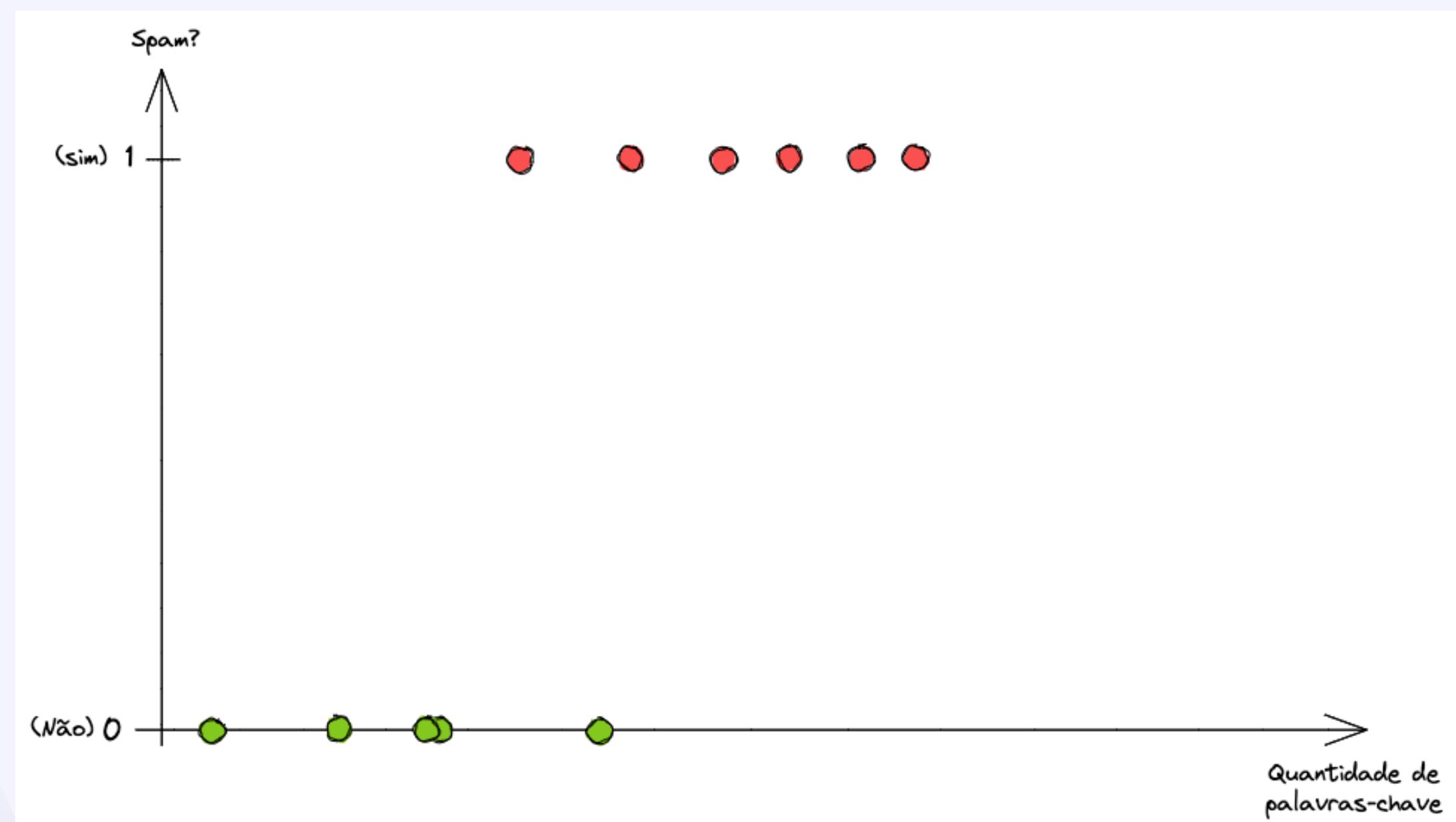
# Classificação Binária



Qtd Palavras	Spam?
7	Sim (1)
4	Não (0)
1	Não (0)
6	Sim (1)
8	Sim (1)
6	Não (0)
4	Não (0)
3	Não (0)
9	Sim (1)
5	Sim (1)
10	Sim (1)

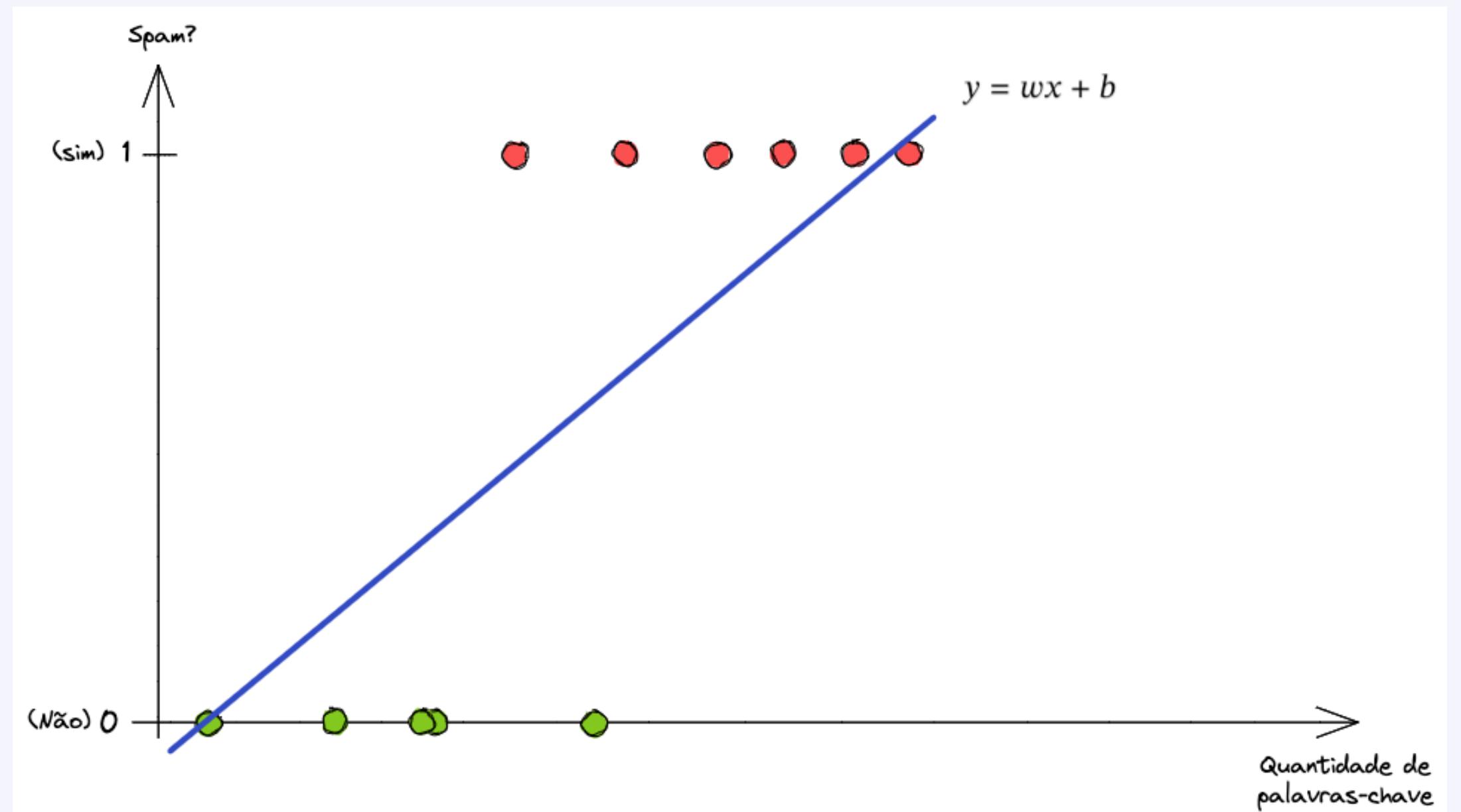
# Classificação Binária

Podemos tentar aplicar uma Regressão Linear?



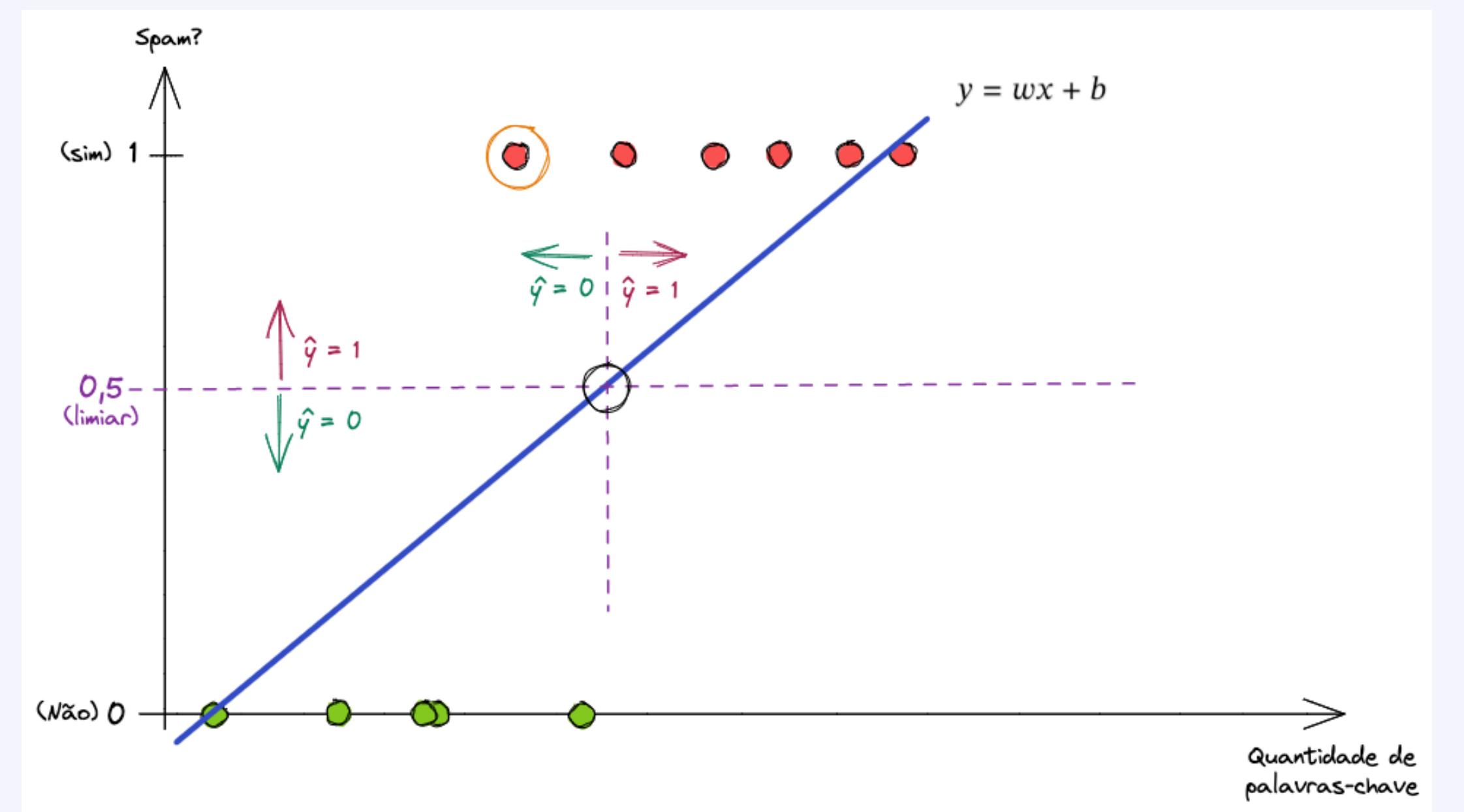
# Classificação Binária

Esse modelo satisfaz o que precisamos?



# Classificação Binária

## Analisando com com limiar



# Regressão Linear

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = \beta^T X$$

- onde  $X = [1, X_1, X_2, \dots, X_d]$
- Se consideramos a saída U como um valor inteiro, podemos usar o modelo de regressão para a realizarmos a classificação de dados

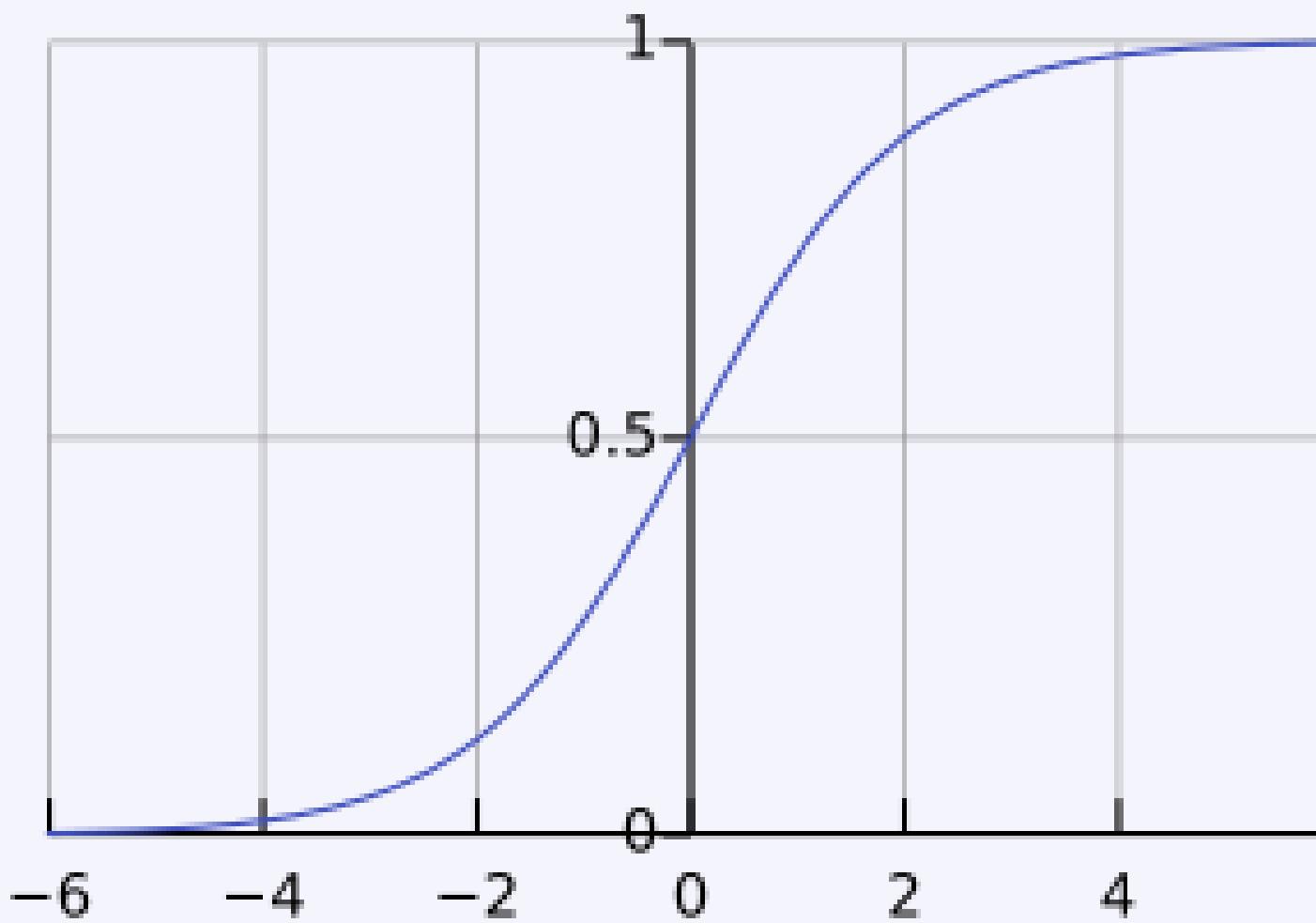
# Regressão Logística

- Vamos definir as probabilidades para o caso de duas classes

$$p(y = 0 | X) \quad \text{e} \quad p(y = 1 | X)$$

# Regressão Logística

Podemos usar a função logística em nosso problema



$$h(z) = \frac{e^z}{1 + e^z}$$

Essa função retorna valores no intervalo [0,1]

# Regressão Logística

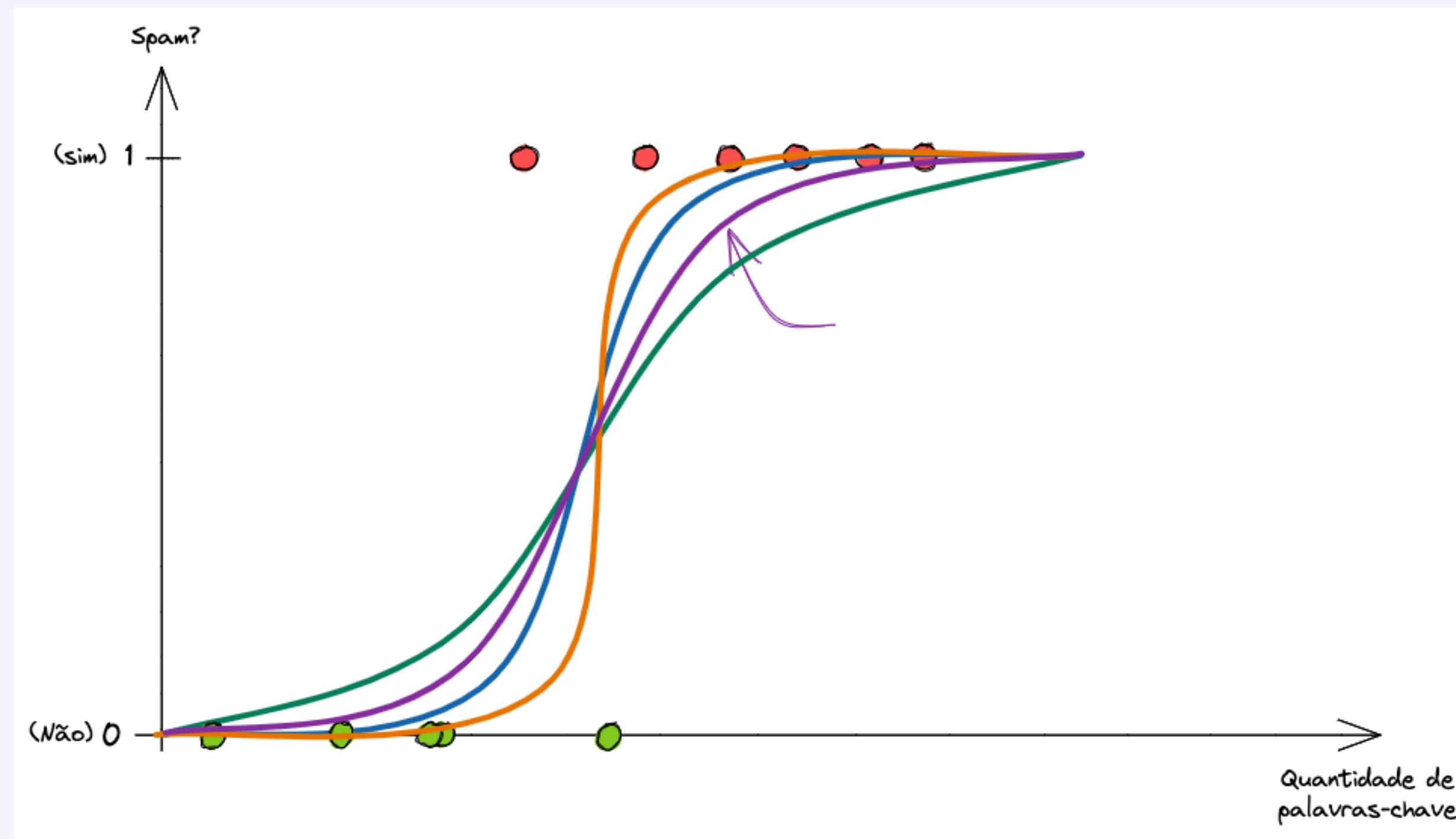
Usando a função logística

$$p(y = 1 | x) = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}}$$

$$p(y = 0 | x) = 1 - p(y = 1 | x) = \frac{1}{1 + e^{\beta^T X}}$$

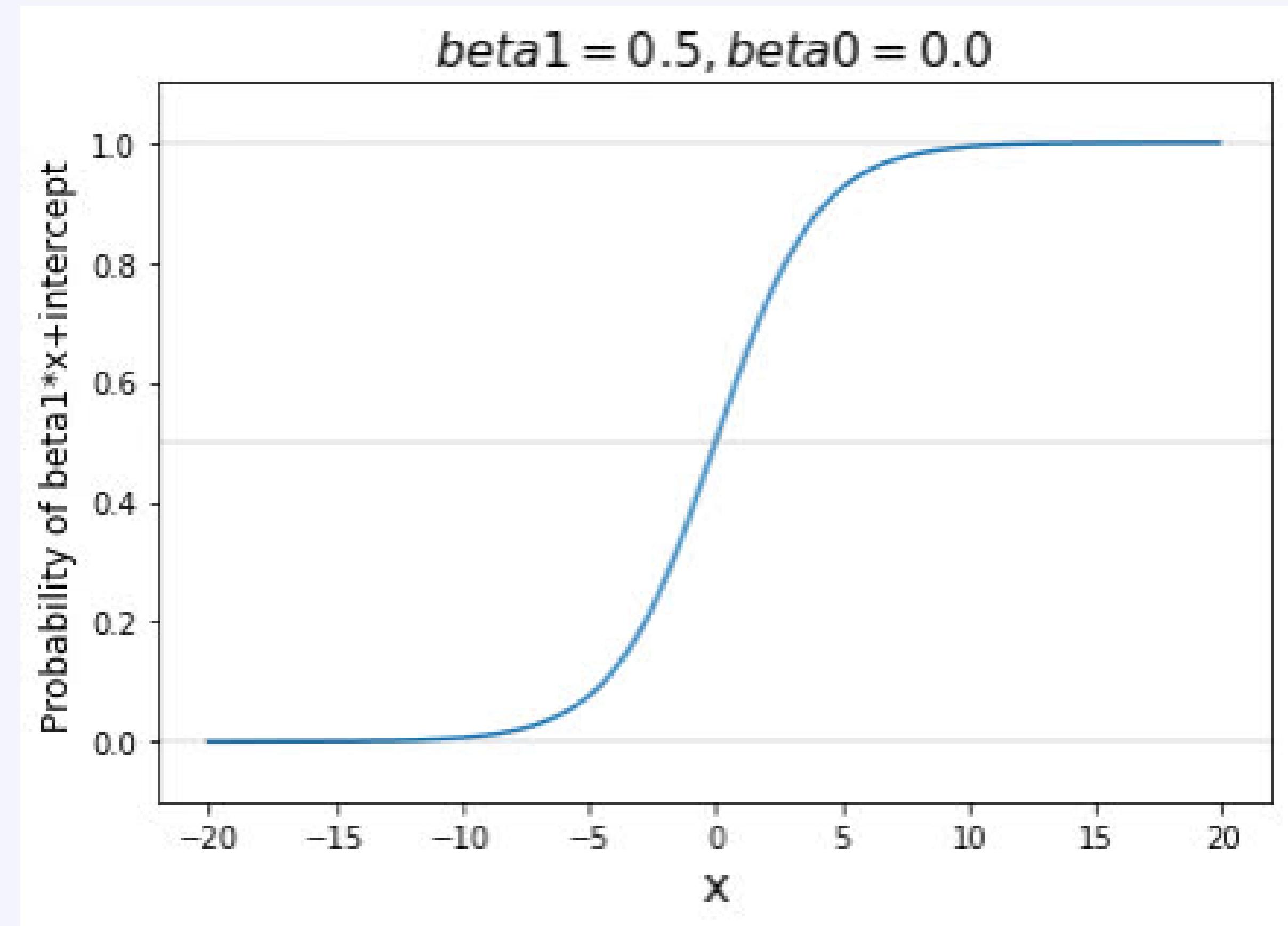
O aprendizado se resume a estimar o vetor de parâmetros  $\beta$   
Esse método é chamado regressão logística

# Regressão Logística



# Regressão Logística

$\text{beta1} = 0.5, \text{beta0} = 0.0$



# Regressão Logística

A superfície de decisão pode ser calculada usando:

$$p(y = 1 | x) = p(y = 0 | x)$$

$$\frac{e^{\beta^T X}}{1 + e^{\beta^T X}} = \frac{1}{1 + e^{\beta^T X}}$$

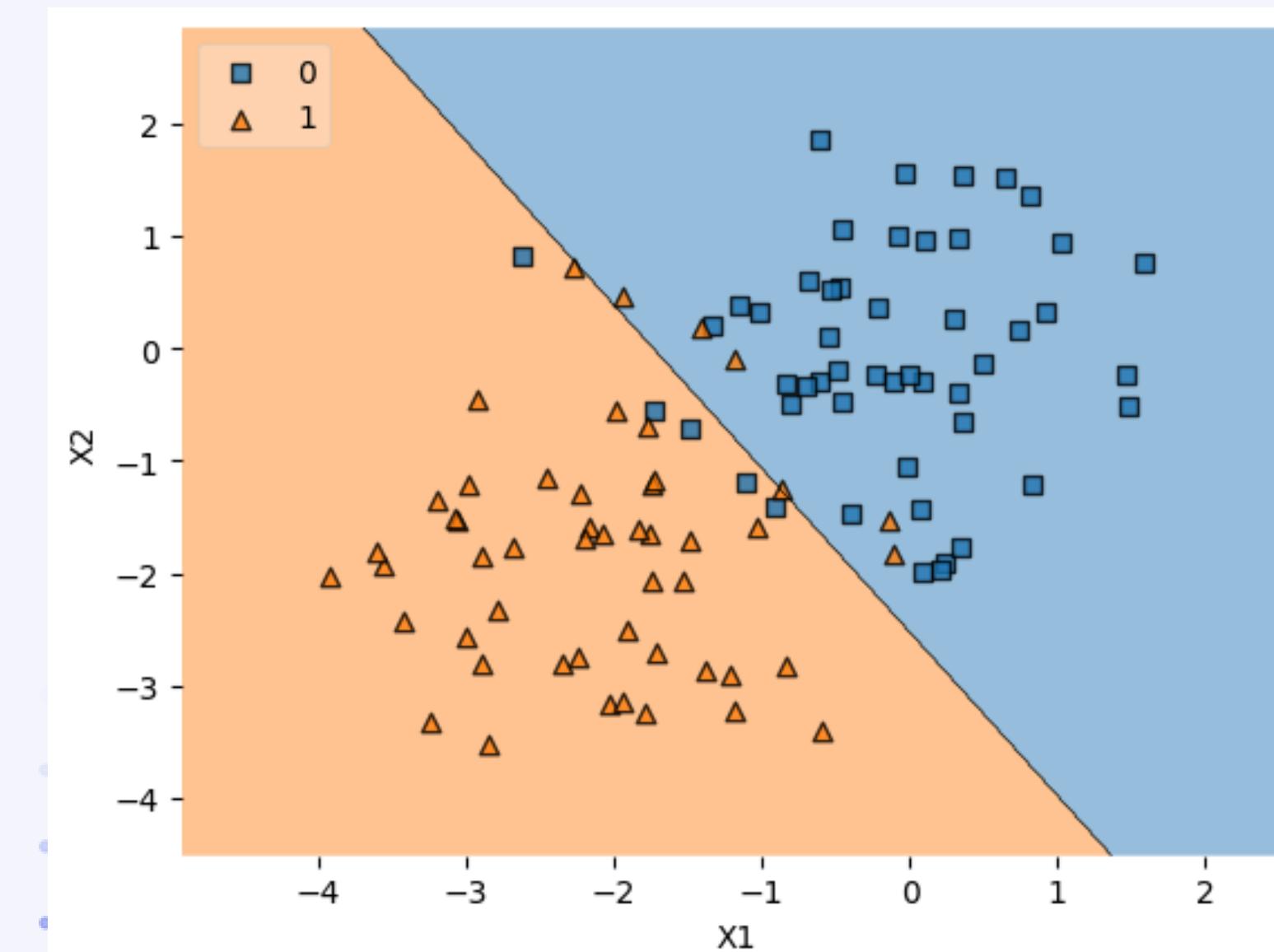
Ou seja, basta resolvemos:  $\beta^T X = 0$

# Regressão Logística

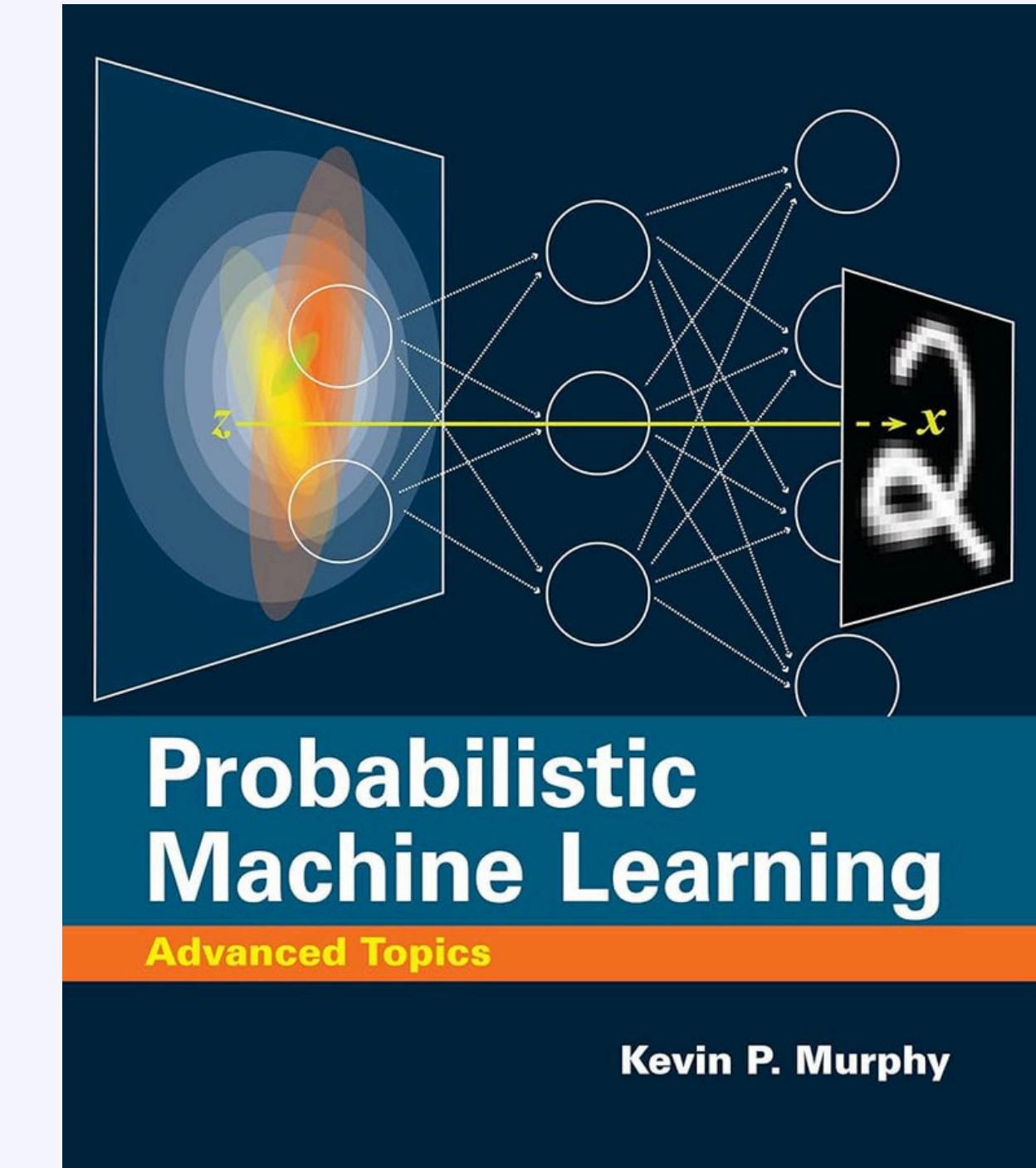
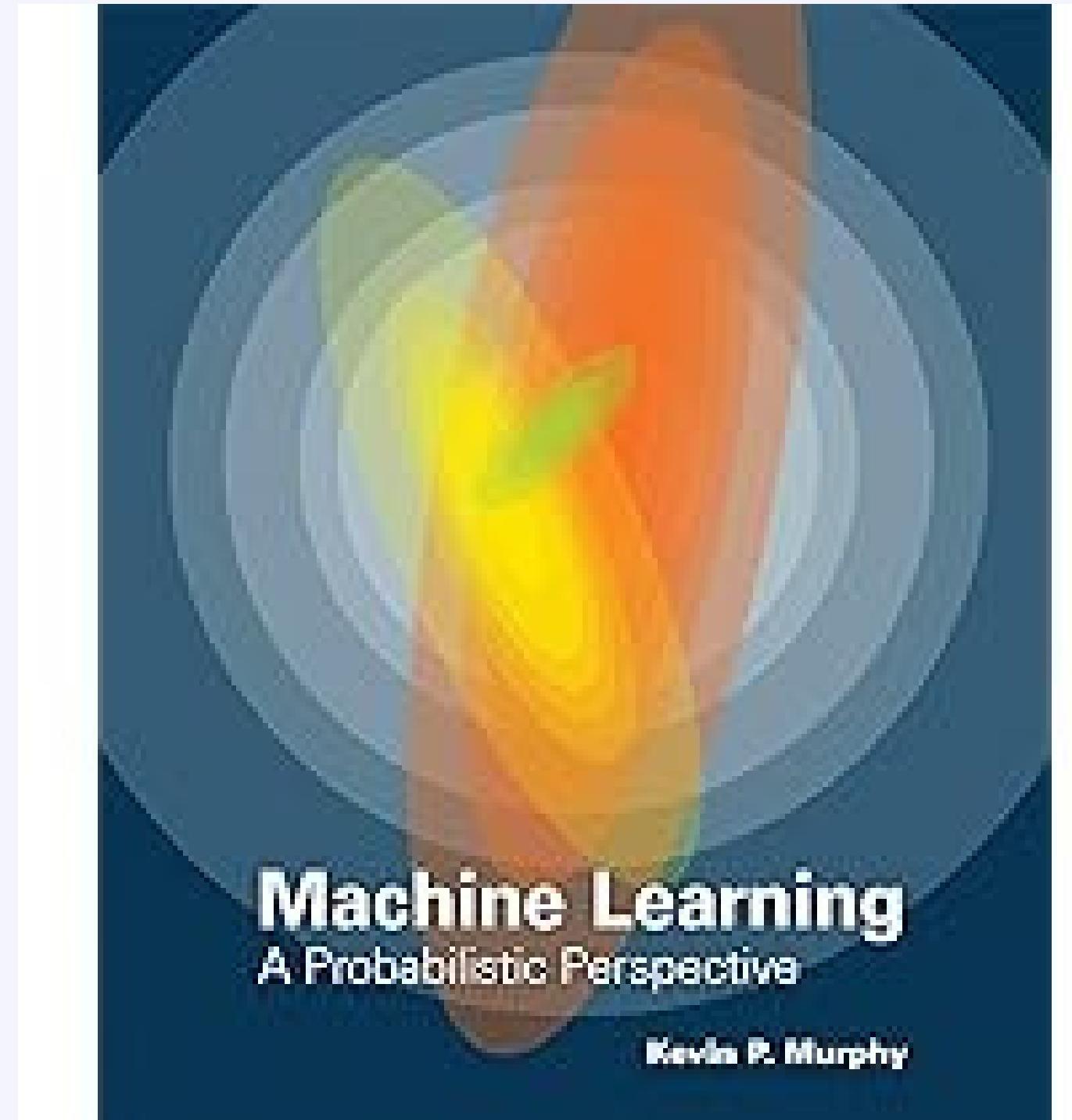
A solução são hiperplanos. Logo, a superfície de separação são hiperplanos (lineares):

$$Z = \beta^T X = 0$$

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$



# Regressão Logística

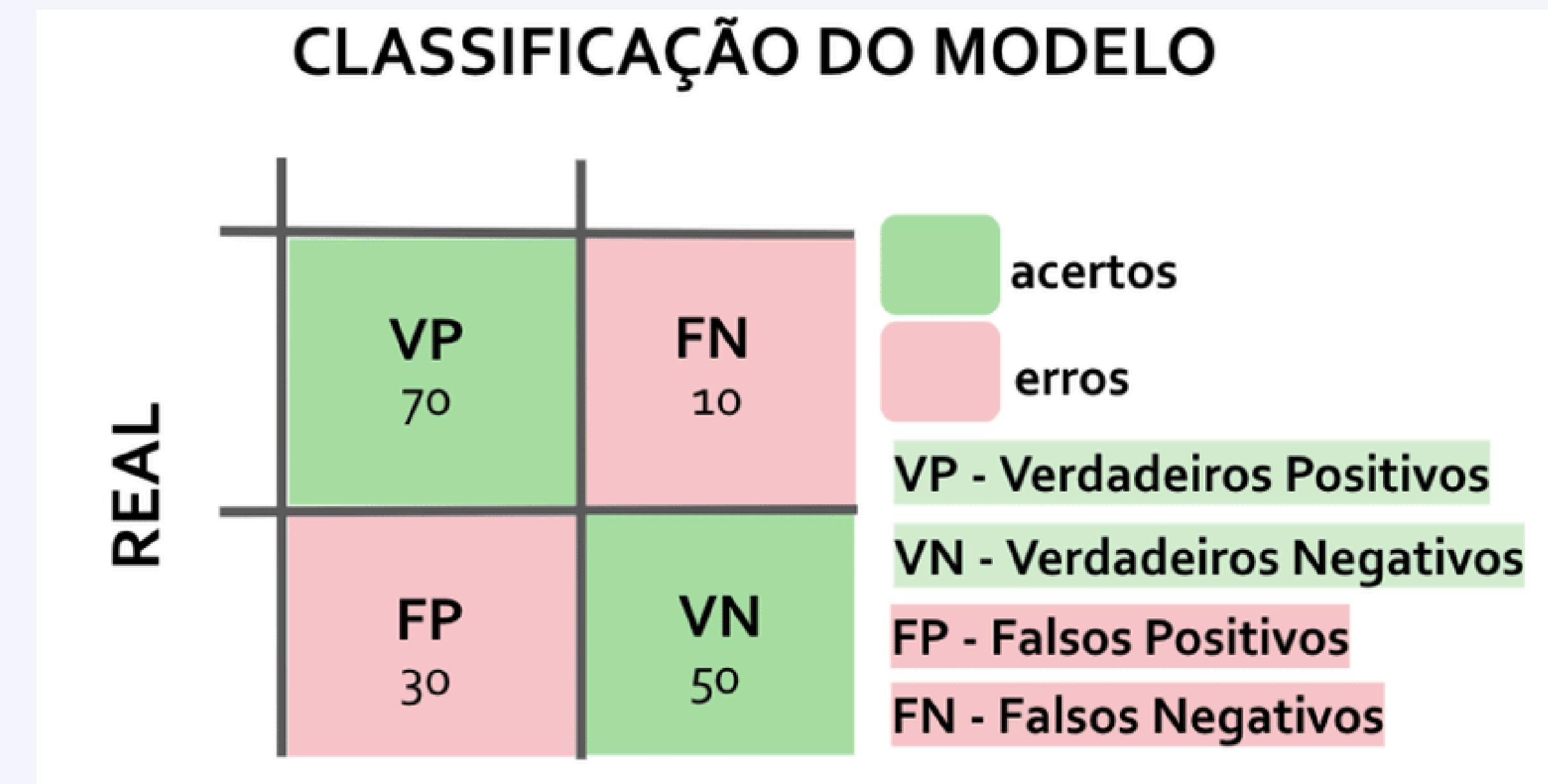


02

# Métricas de Avaliação

# Qual o método de avaliação do modelo?

A nossa métrica de avaliação será pela matriz de confusão



# MATRIZ DE CONFUSÃO

- **Verdadeiro positivo (true positive – TP): ocorre quando no conjunto real, a classe que estamos buscando foi prevista corretamente.**
- **Falso positivo (false positive – FP): ocorre quando no conjunto real, a classe que estamos buscando prever foi prevista incorretamente.**
- **Verdadeiro Negativo (true negative – TN): ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista corretamente.**
- **Falso negativo (false negative – FN): ocorre quando no conjunto real, a classe que não estamos buscando prever foi prevista incorretamente.**

# Métricas de Avaliação

A partir da matriz de confusão podemos obter outras métricas, um exemplo é Acurácia

$$acuracia = \frac{TP + TN}{TP + FP + TN + FN}$$

**Essa métrica diz quanto o modelo acertou as previsões**

# Métricas de Avaliação

Dentre todas as instâncias que o modelo previu como positivas,  
quantas eram realmente positivas?

$$precisao = \frac{VP}{VP + FP}$$

Então ela é usada para medir a exatidão das previsões  
positivas feitas por um modelo.

# Métricas de Avaliação

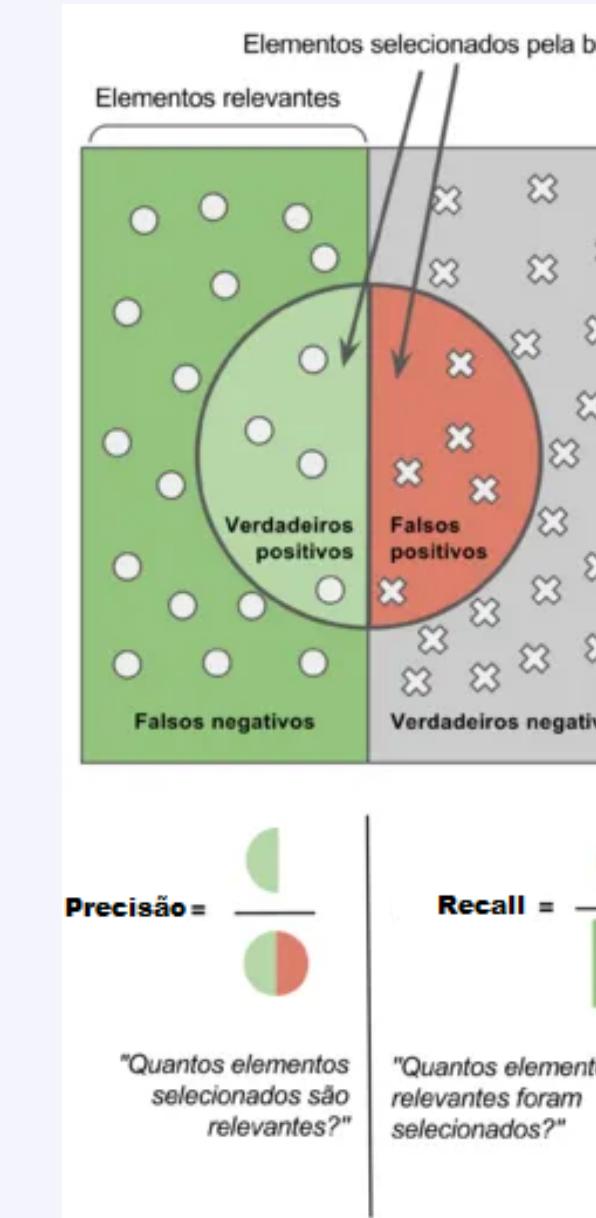
**Dentre todas as instâncias que são realmente positivas, quantas foram corretamente identificadas pelo modelo?**

$$\text{recall} = \frac{TP}{TP + FN}$$

**Essa métrica mede a capacidade do modelo de identificar corretamente todas as instâncias positivas.**

# Métricas de Avaliação

## Diferença entre Precisão e Recall



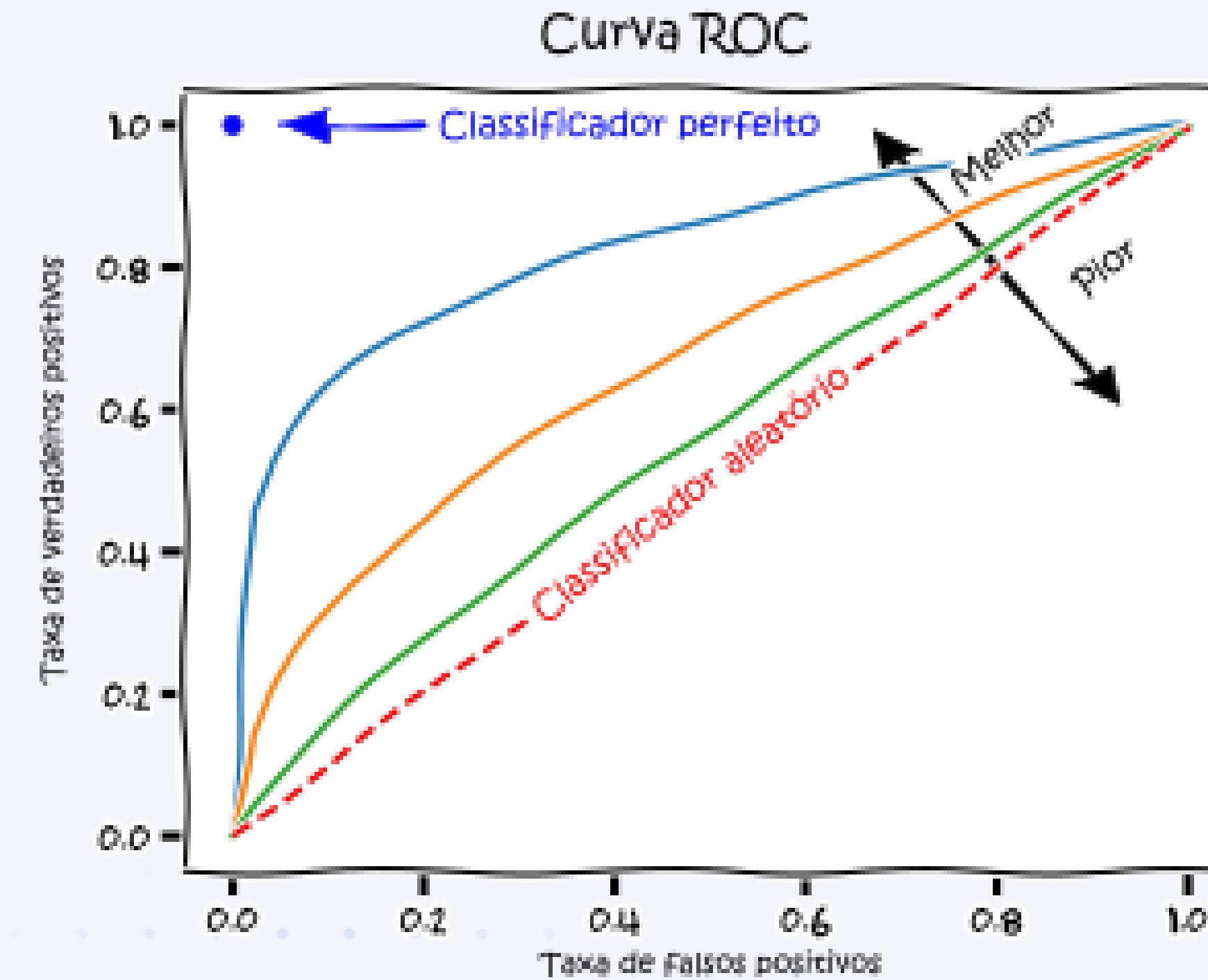
- **Precisão: Foco em minimizar falsos positivos (ser mais conservador).**
- **Recall: Foco em minimizar falsos negativos (capturar o máximo de verdadeiros positivos).**

# Métricas de Avaliação

Média harmônica entre a precisão e o recall.

$$f1 - score = 2 \times \frac{precisao \times recall}{precisao + recall}$$

# Métricas de Avaliação



# 03

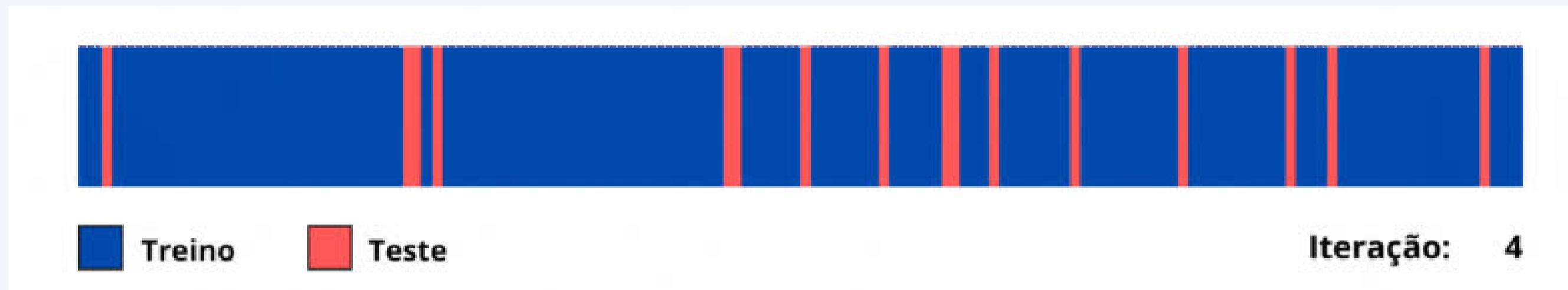
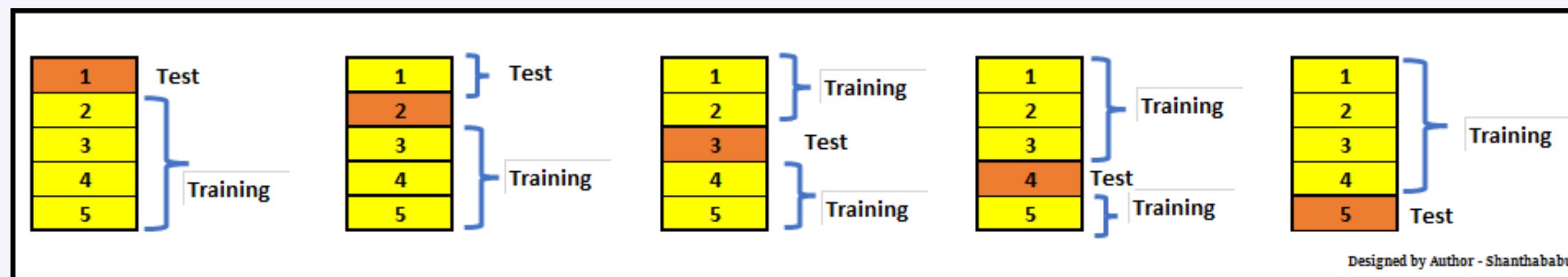
# Técnicas de Validação

# Validação Cruzada(k-fold cross-validation)

Na validação cruzada, o dataset é dividido aleatoriamente em “K” grupos. Quando definimos um número para “k”, usamos o número no lugar de “k” para fazer referência ao teste



# Validação Cruzada(k-fold cross-validation)



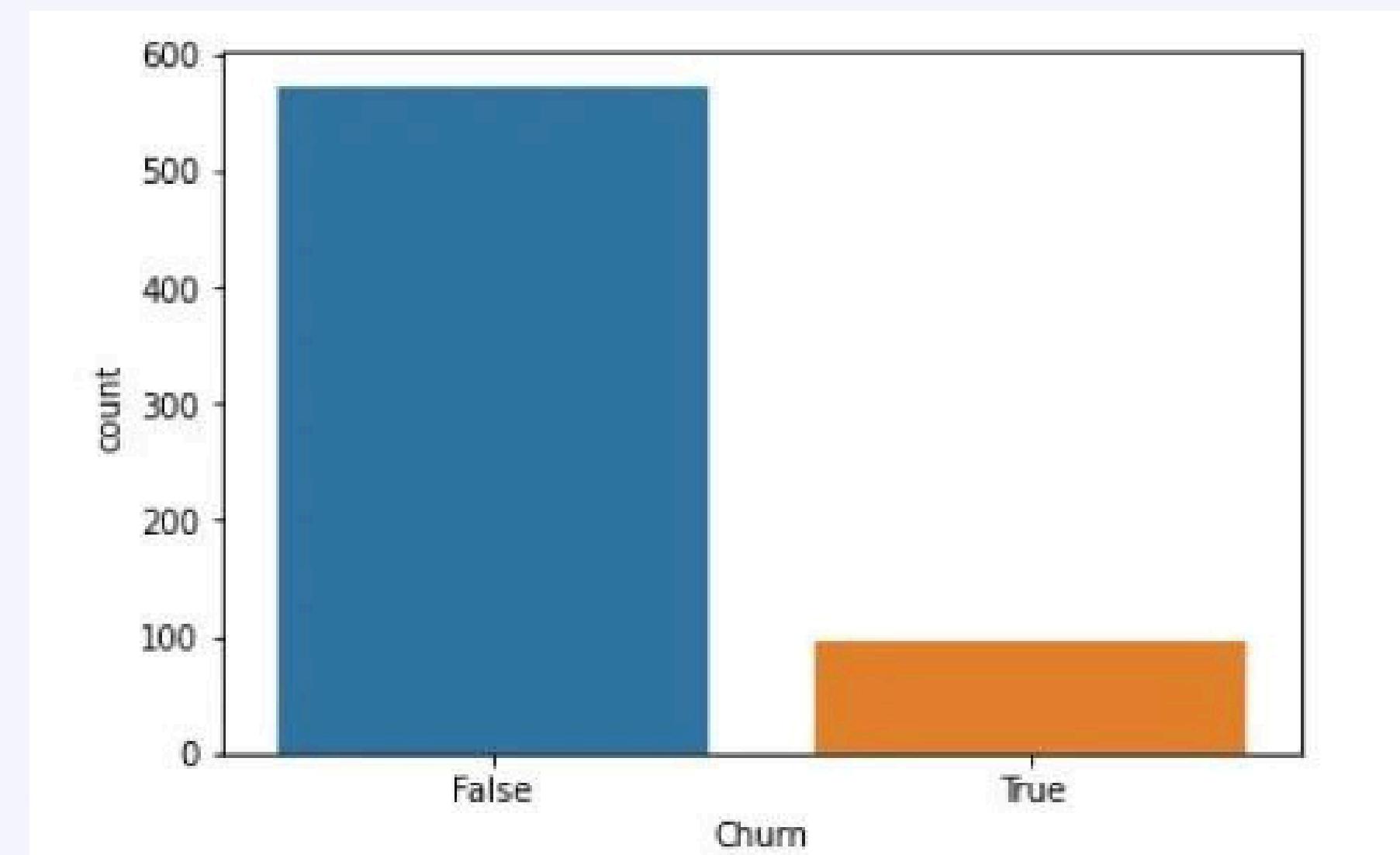
**04**

# Técnicas de Balancamento



# Dados Desbalanceados

O que fazer quando os dados de uma classe não são suficientes?.



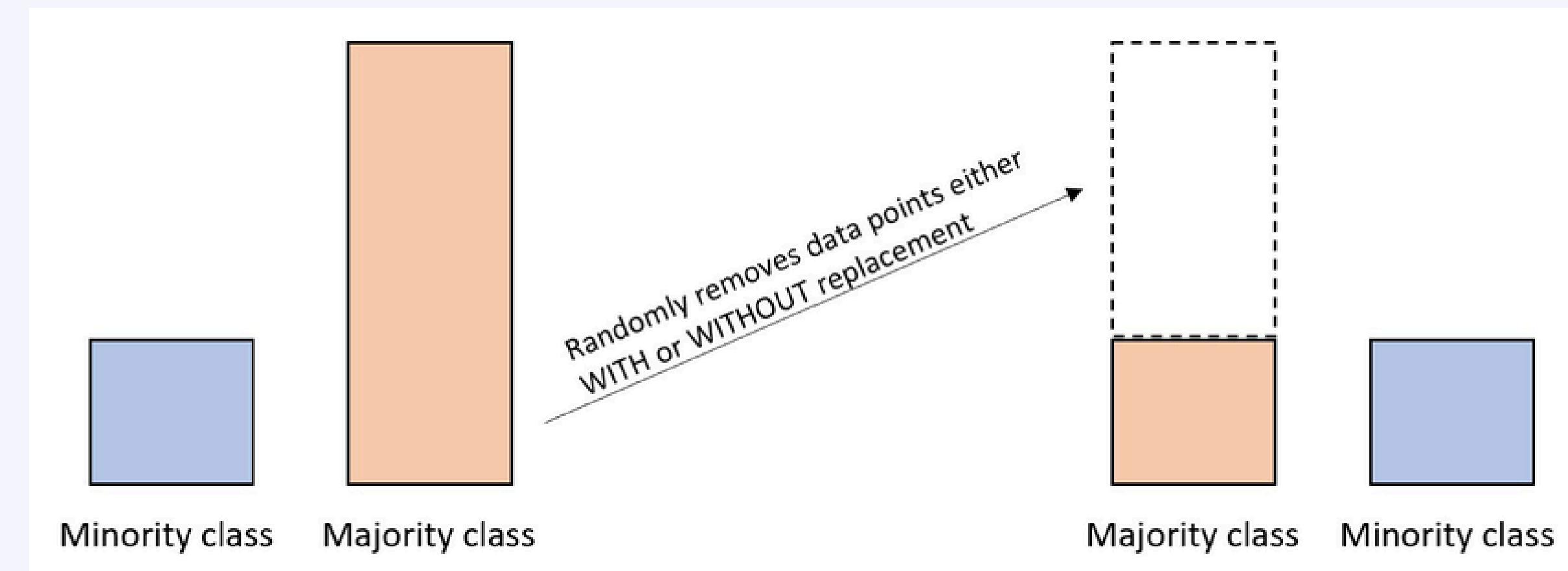
# Dados Desbalanceados

O que pode ocorrer se os dados forem desbalanceados?

- Overfitting: O algoritmo irá se ajustar demasiadamente aos resultados, classificando boa parte como não fraudes. Isso ocorre, uma vez que o modelo aprenderá de maneira excessiva os padrões da classe majoritária.
- Distorções: A análise de correlações é distorcida quando há uma classe dominante, gerando interpretações equivocadas e perdas de informações em relação a classe minoritária.

# Técnicas de Balanceamento

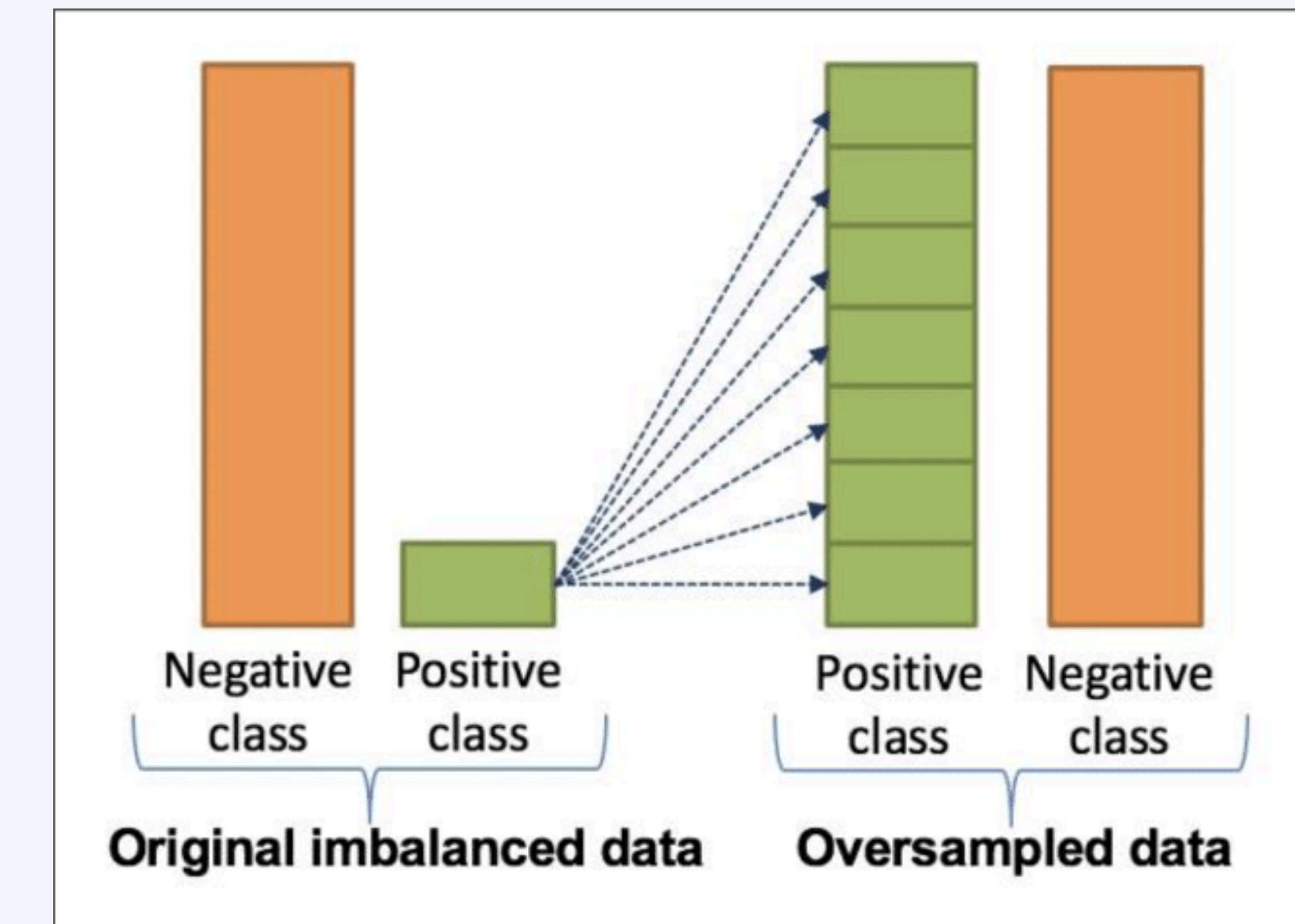
## UNDERSAMPLING



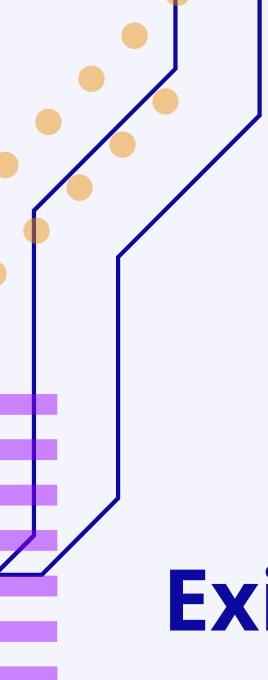
**Processo de reduzir aleatoriamente o número de amostras da classe majoritária em um conjunto de dados desbalanceado**

# Técnicas de Balanceamento

## OVERSAMPLING



**Processo de aumentar a quantidade de dados da classe minoritária até atingir um equilíbrio desejado entre as classes.**

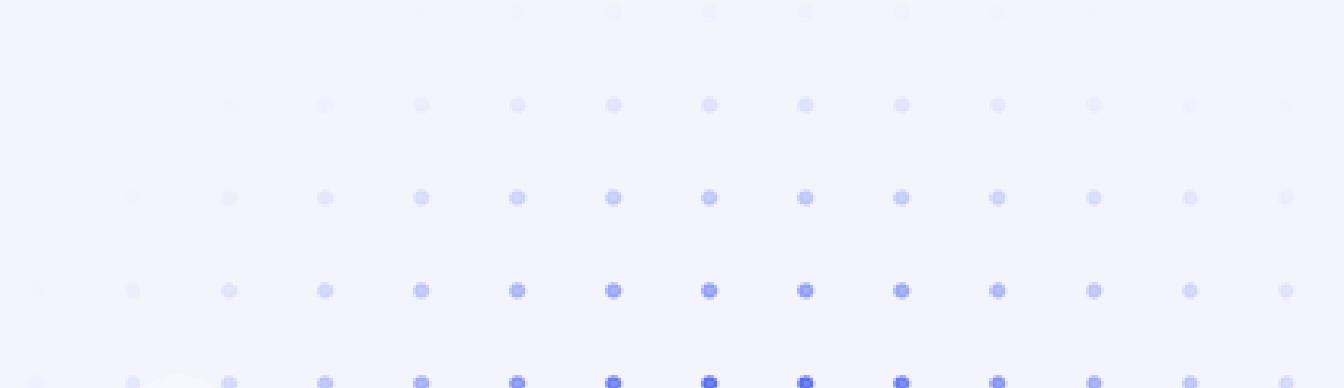


O fazer se a classe minoritária é muito pequena em comparação a classe majoritária?

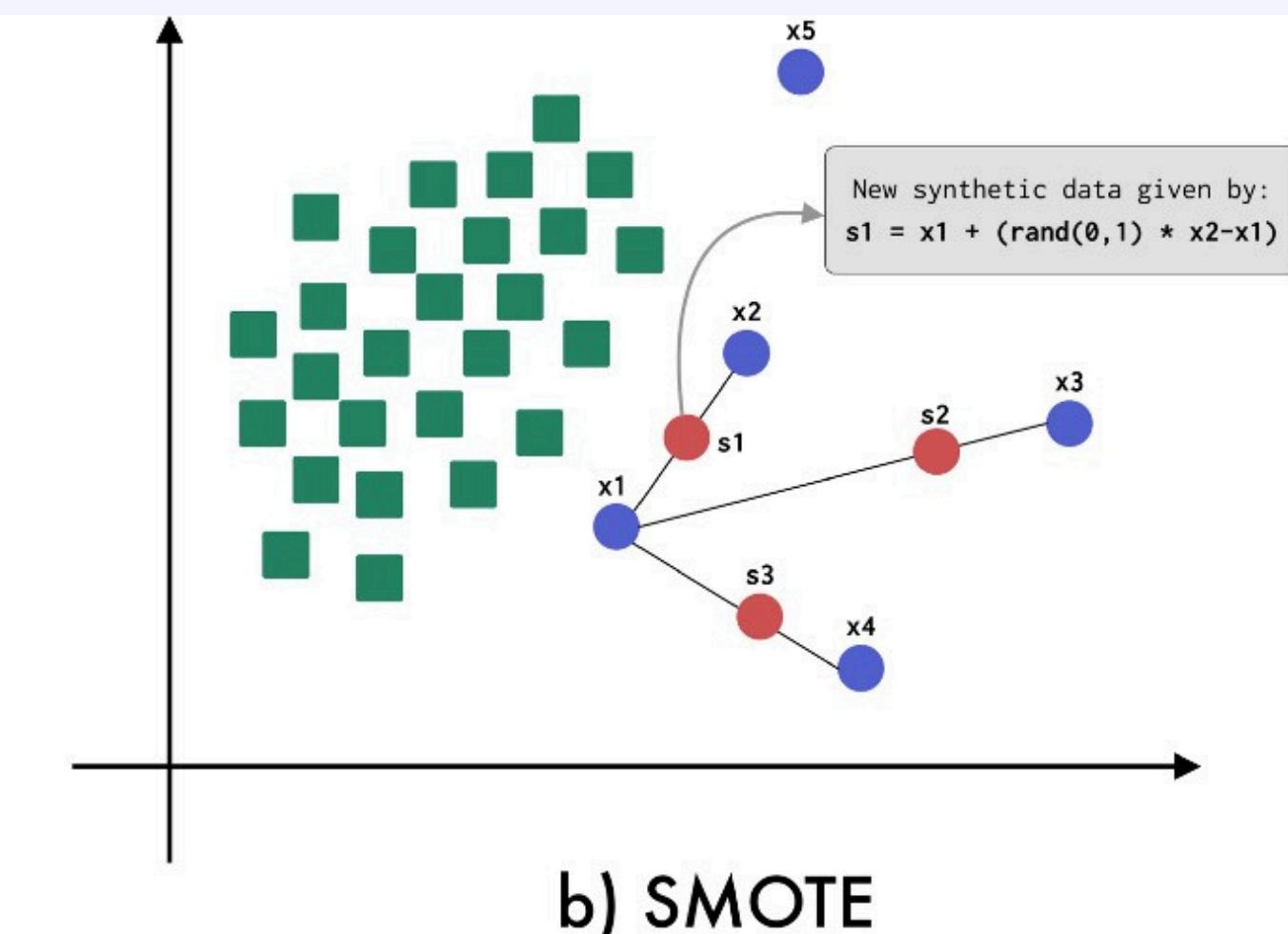
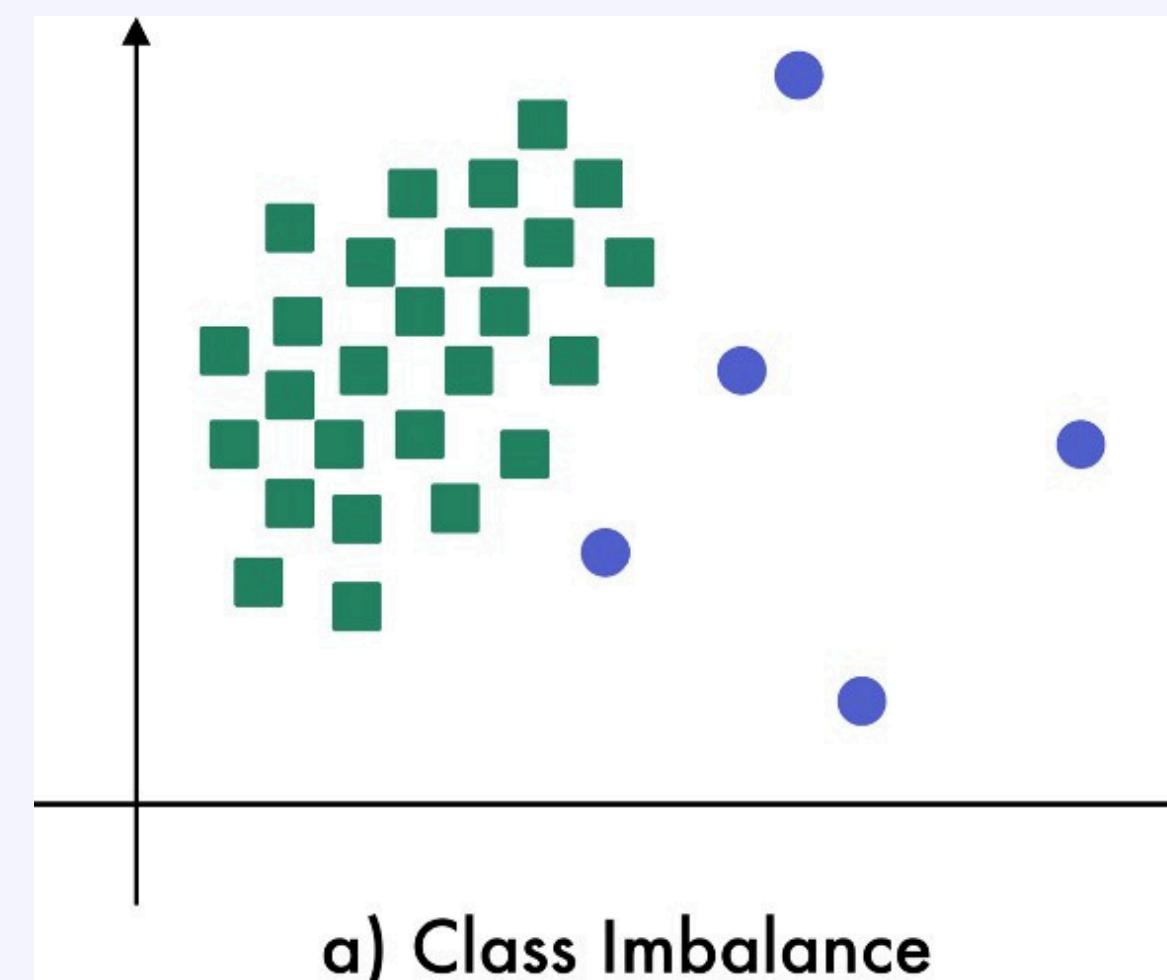
Existe uma técnica chamada **SMOTE (Synthetic Minority Oversampling Technique)**

Utilizamos o algoritmo KNN para identificar os k vizinhos mais próximos da observação selecionada.

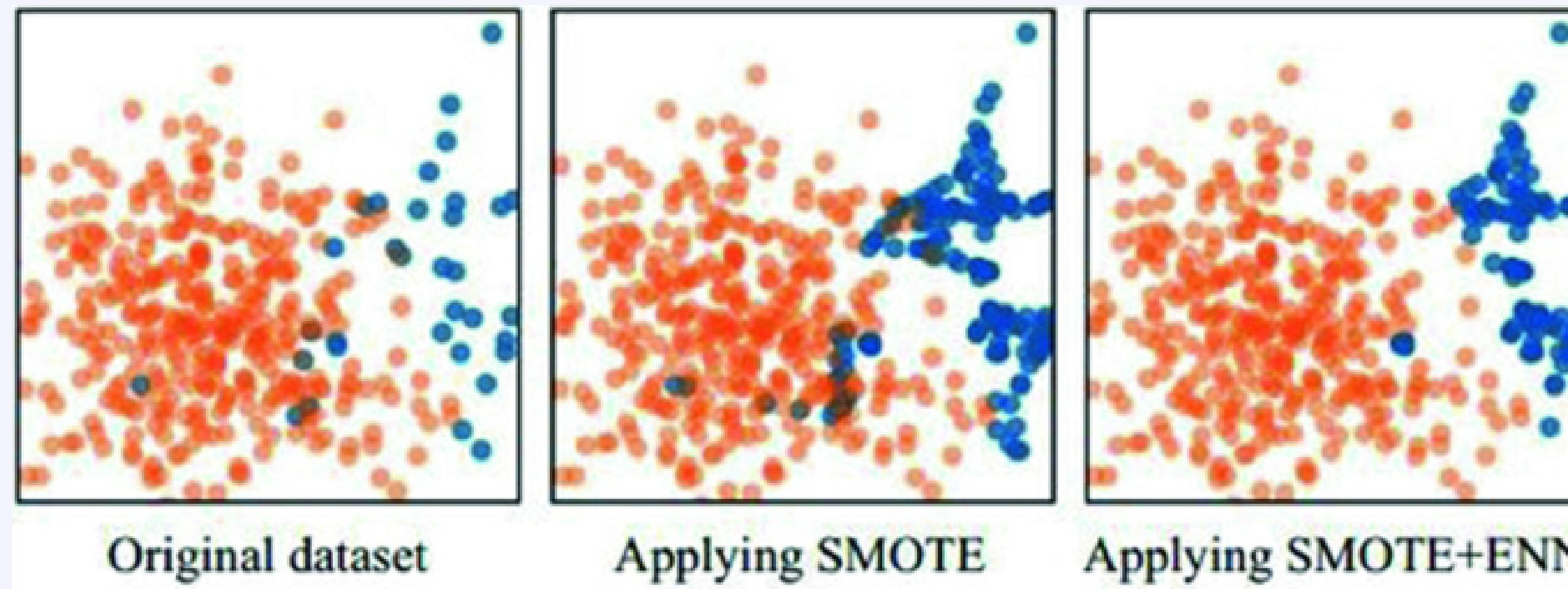
$$s_i = x_i + (x_{zi} - x_i) \times \lambda$$



# Técnicas de Balanceamento



# Técnicas de Balanceamento



# Obrigado pessoal!

Até próxima aula :)



**Iris Data Science UNICAMP**



**@irisdatascienceunicamp**

# Referências

Parte do material foi adaptado dos slides da Prof<sup>a</sup>. Sandra Avila apresentados na disciplina MC886.

## Livros utilizados:

- An Introduction to Statistical Learning  
(2023)
- Hands-On Machine Learning with  
Scikit-Learn, Keras, and Tensor Flow  
(2017)