

Tera

Aula #17

Regressão Logística & Métricas de Classificação

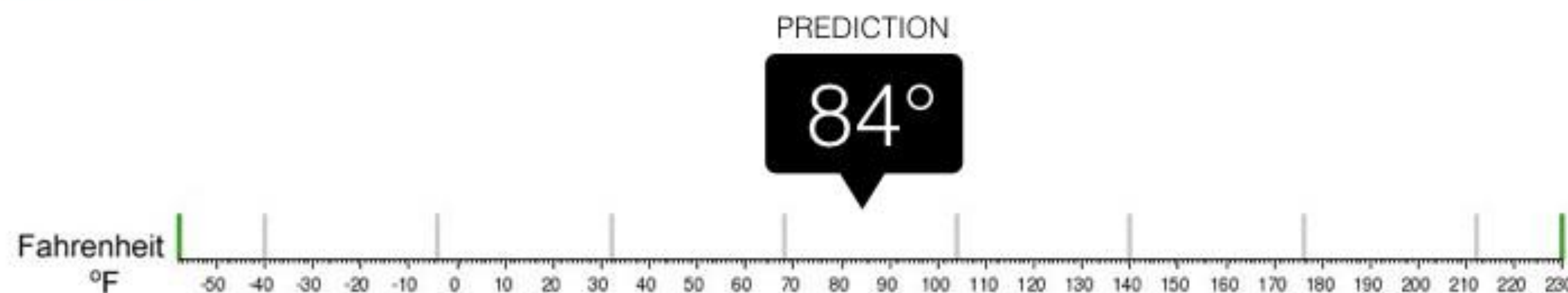
Gabriel Cypriano
07/nov/2018

Classificação vs Regressão



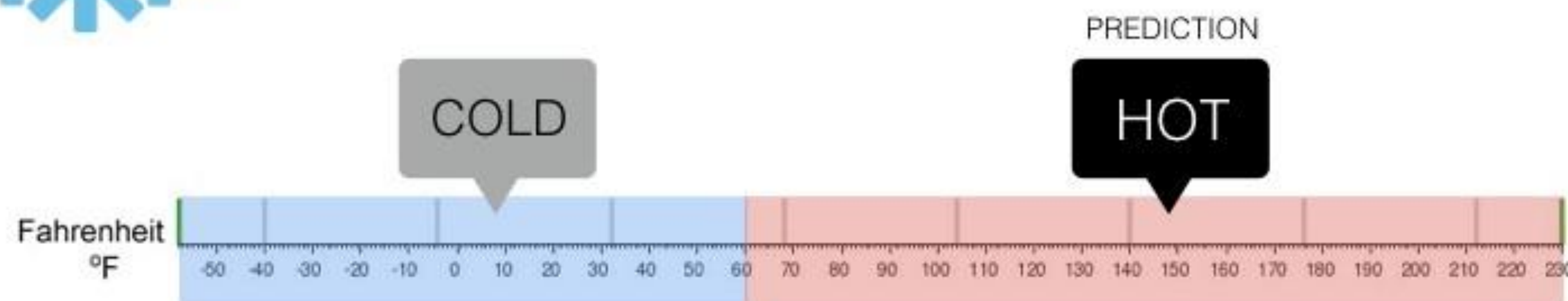
Regression

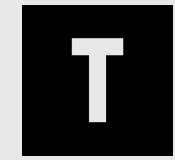
What is the temperature going to be tomorrow?



Classification

Will it be Cold or Hot tomorrow?





Classificação ou Regressão?

Entrada: características de um imóvel

Saída: valor/preço



Classificação ou Regressão?

Entrada: dados de uma transação bancária

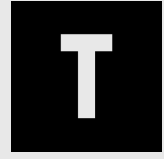
Saída: fraude ou genuína?



Classificação ou Regressão?

Entrada: altura dos pais

Saída: altura do filho



Classificação ou Regressão?

Entrada: imagem

Saída: pessoa, carro ou sinal de trânsito?

UCI

**Machine Learning Repository**Center for Machine Learning and Intelligent Systems[About](#) [Citation Policy](#)[Reposito](#)

Bank Marketing Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term dep

- Ligações de um banco português ofertando investimento financeiro
- Target: o cliente investiu ou não

Análise Exploratória



Discussão

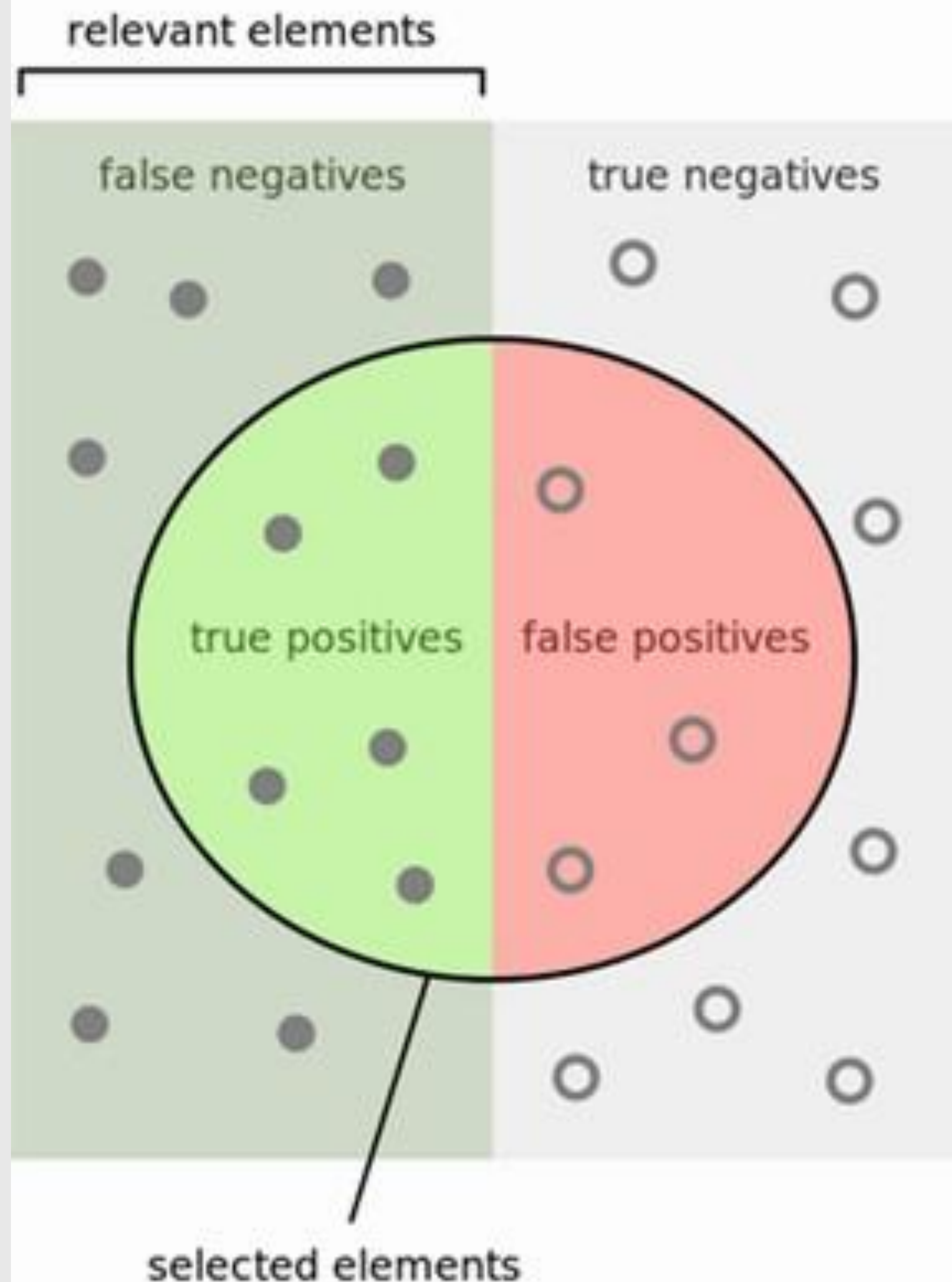
A Acurácia é sempre
a métrica ideal?

Accuracy Paradox

Predictive models with a given level of accuracy may have greater predictive power than models with higher accuracy.

Métricas relevantes


- Precisão
- Recall

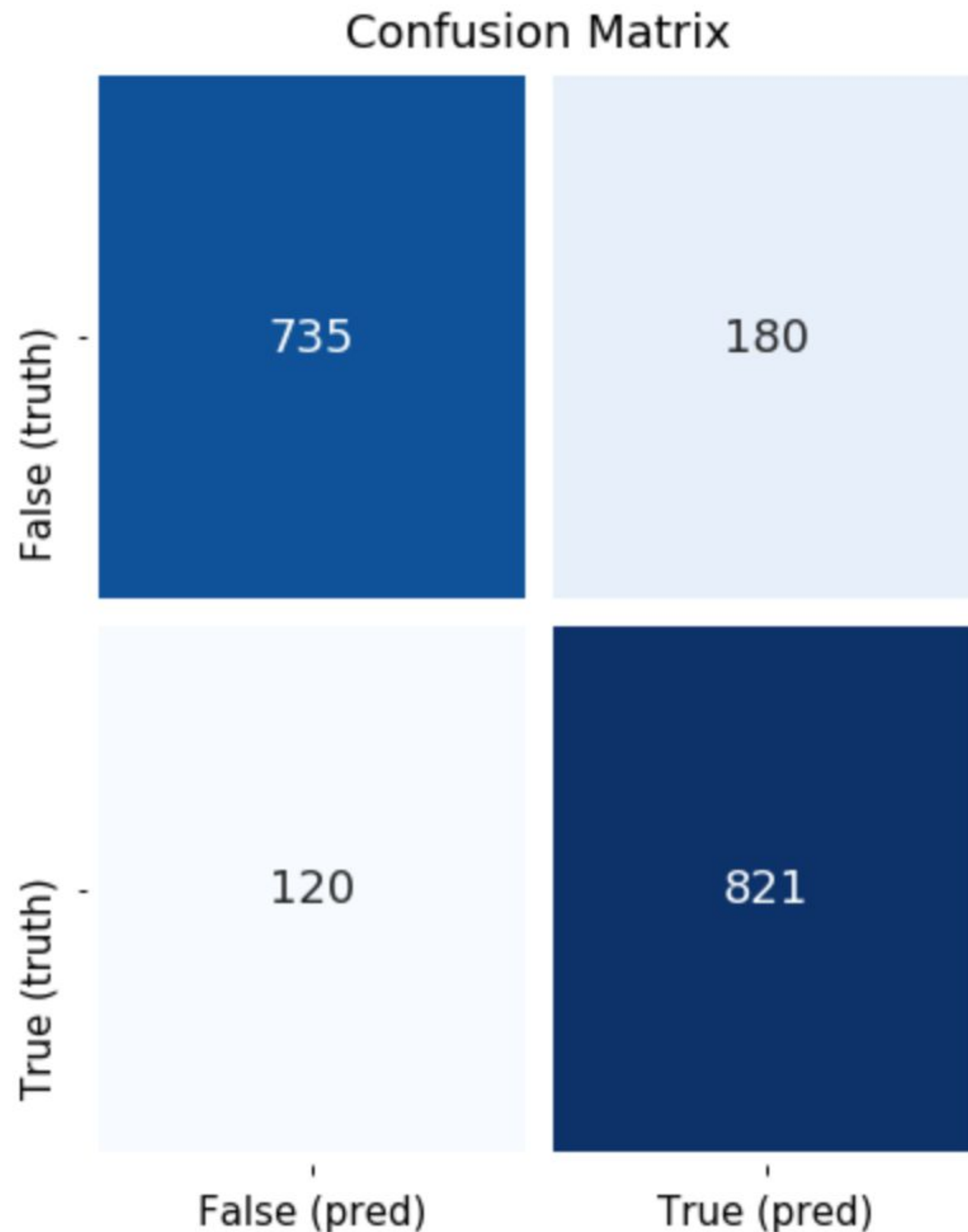


How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$


How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$


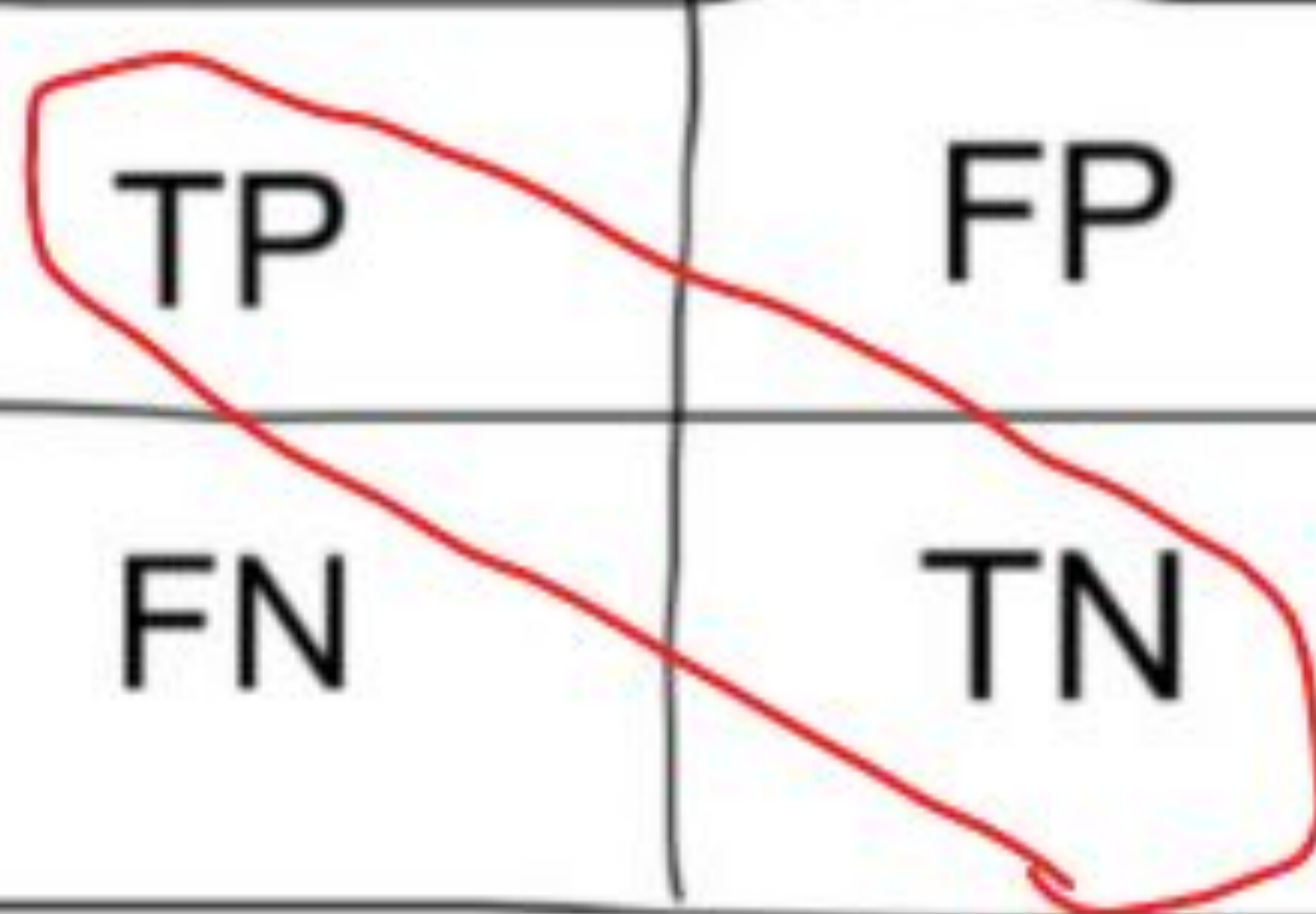


Matriz de Confusão

- Azuis escuros
 - Verdadeiros negativos (esq.)
 - Verdadeiros positivos (dir.)
- Azuis claros
 - Falsos negativos (esq.)
 - Falsos positivos (dir.)

■ Matriz de Confusão: Acurácia

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN



$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

■ Matriz de Confusão: Precisão

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

T Matriz de Confusão: Recall

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

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

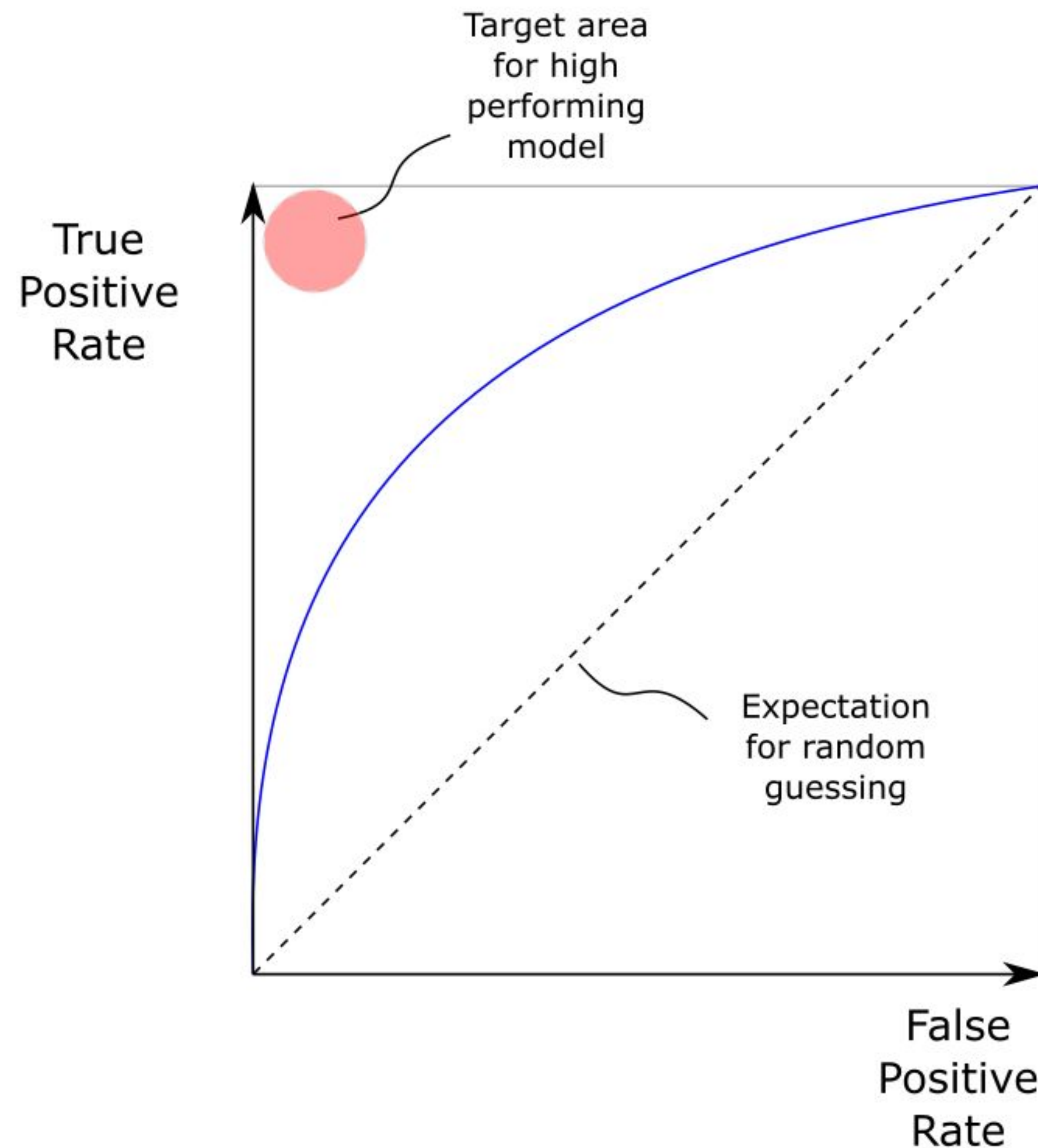
Balanceando entre Precisão e Recall: **F1-score**

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$



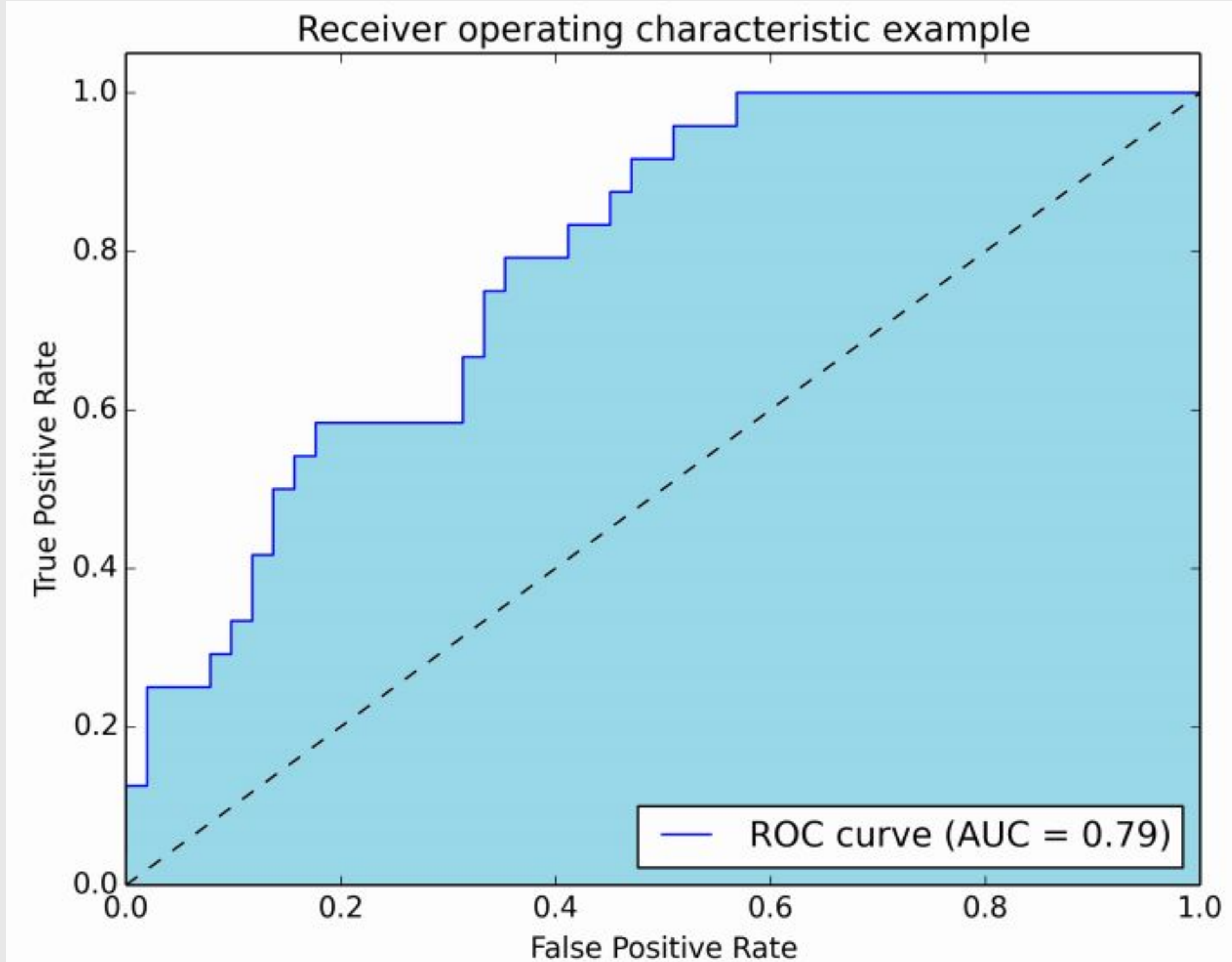
Curva ROC

- Eixo y: verdadeiros positivos
- Eixo x: falsos positivos



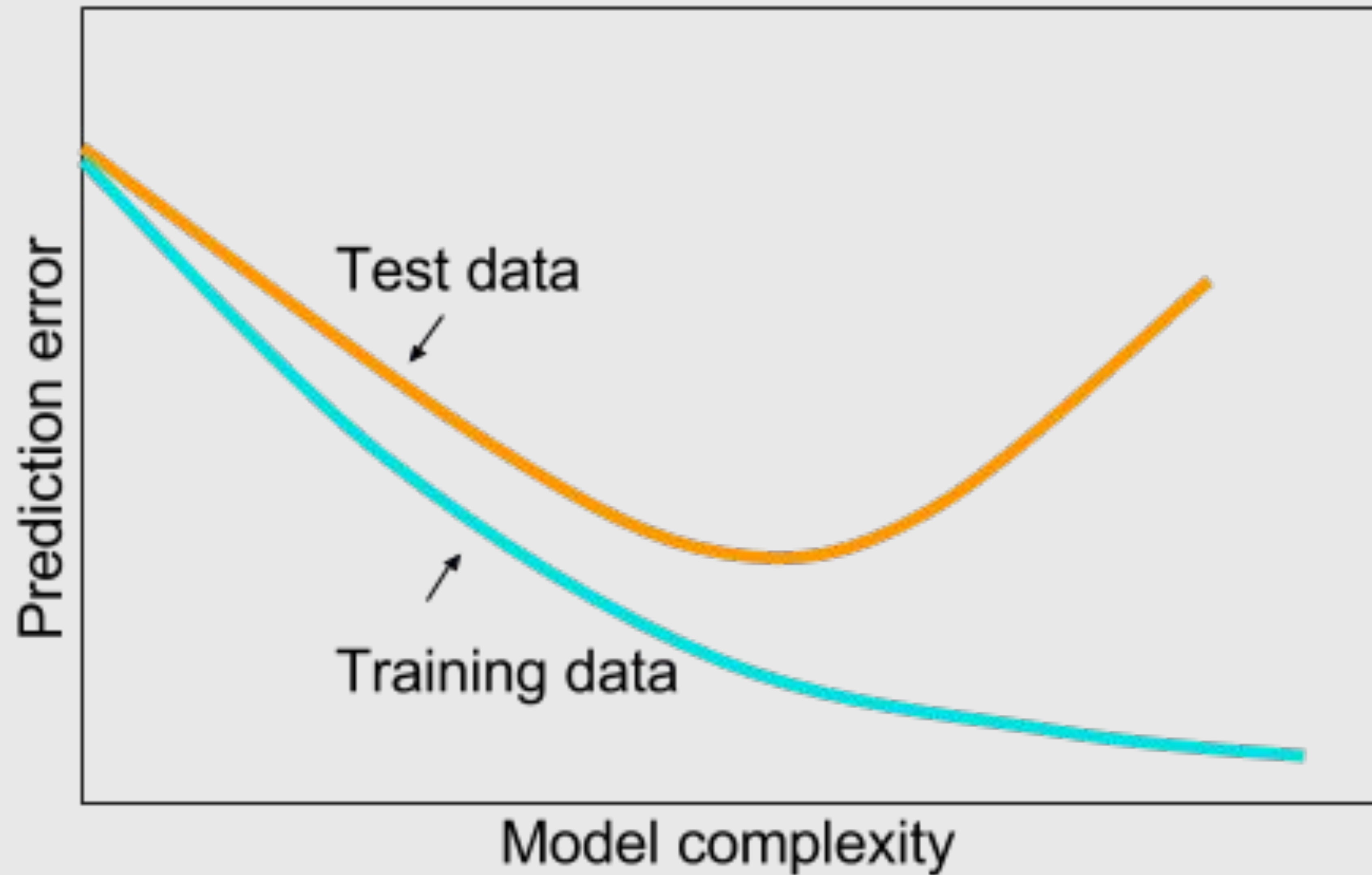
AUC (Area Under Curve)

- 0.50 equivale a um classificador aleatório
- Quando mais próximo de 1, melhor a ordenação



Detecção de Overfitting

- Estratégia simples: comparar qualidade das predições nos datasets de treino vs validação/teste



Relembrando: regressão linear

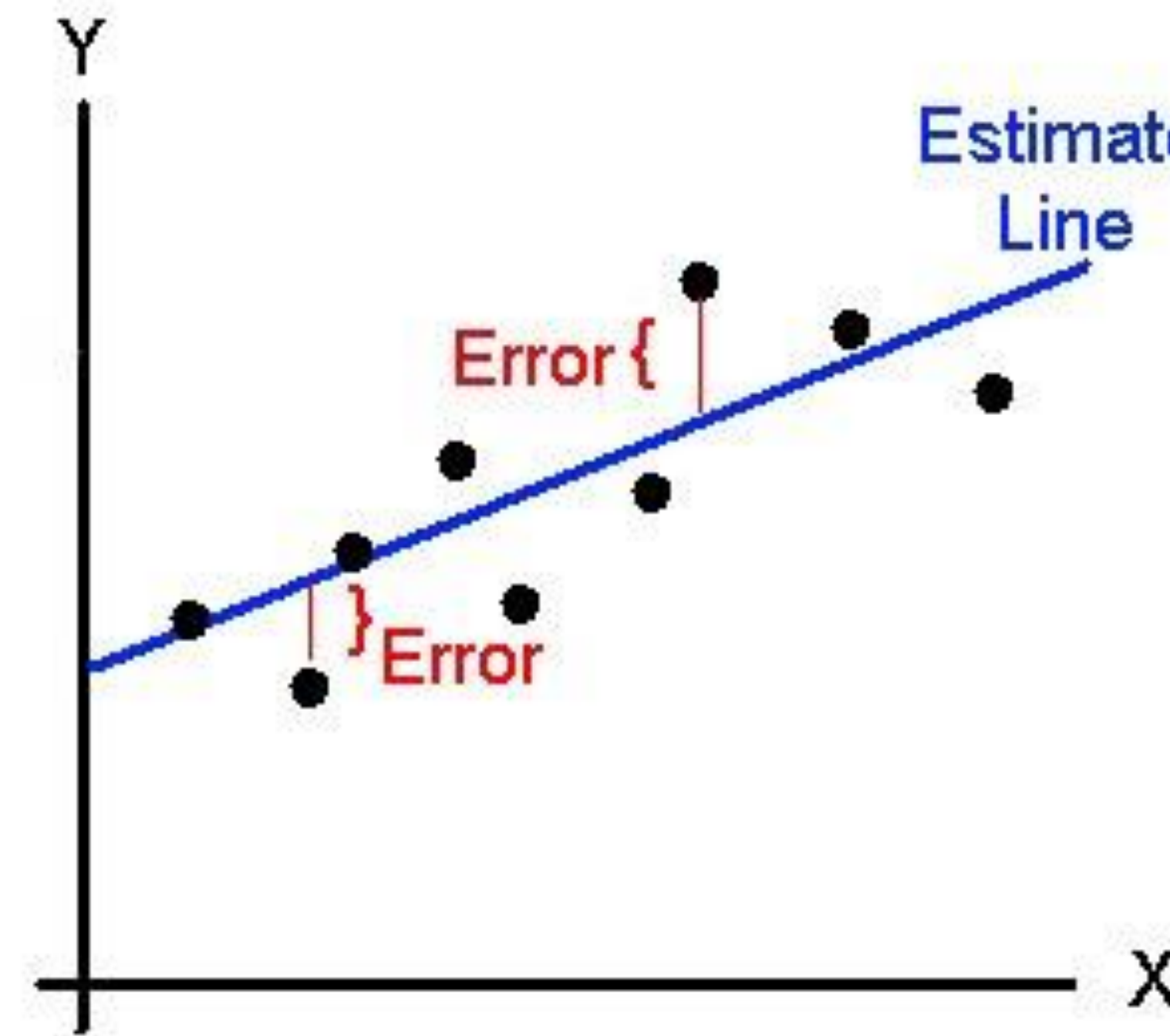
Estimated
(or predicted)
Y value for
observation i

Estimate of
the regression
intercept

Estimate of the
regression slope

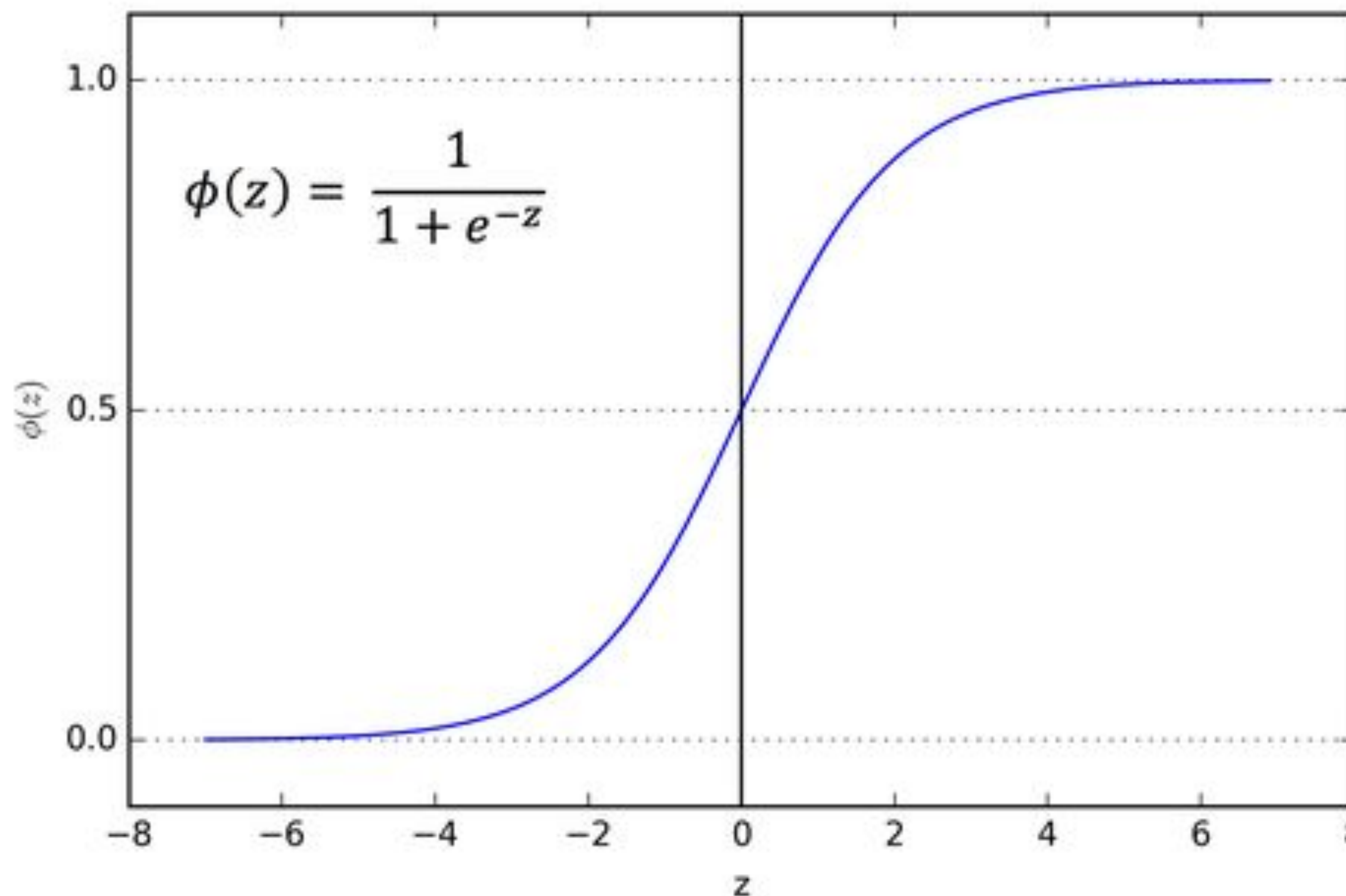
Value of X for
observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$



Função sigmoid (ou logística)

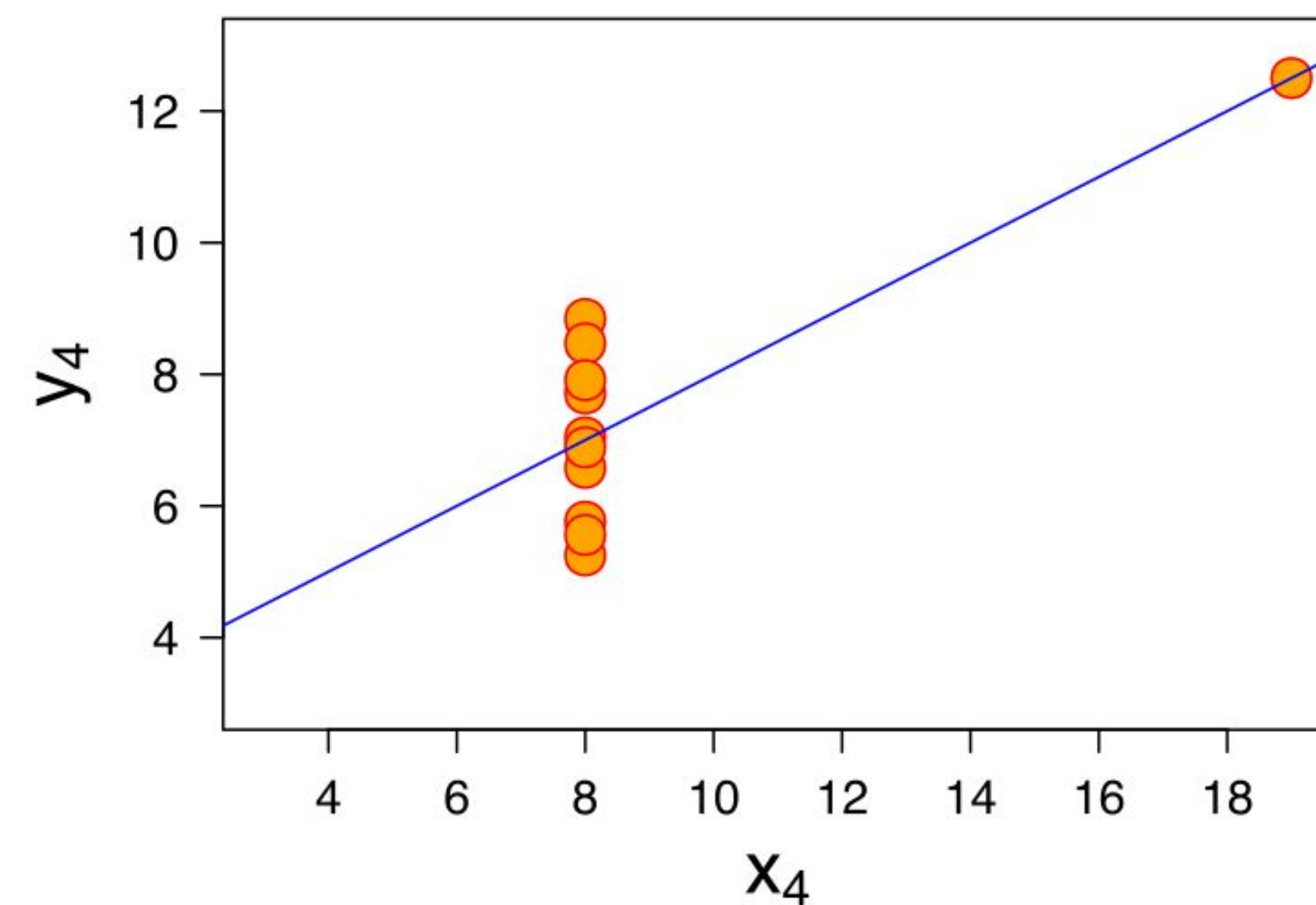
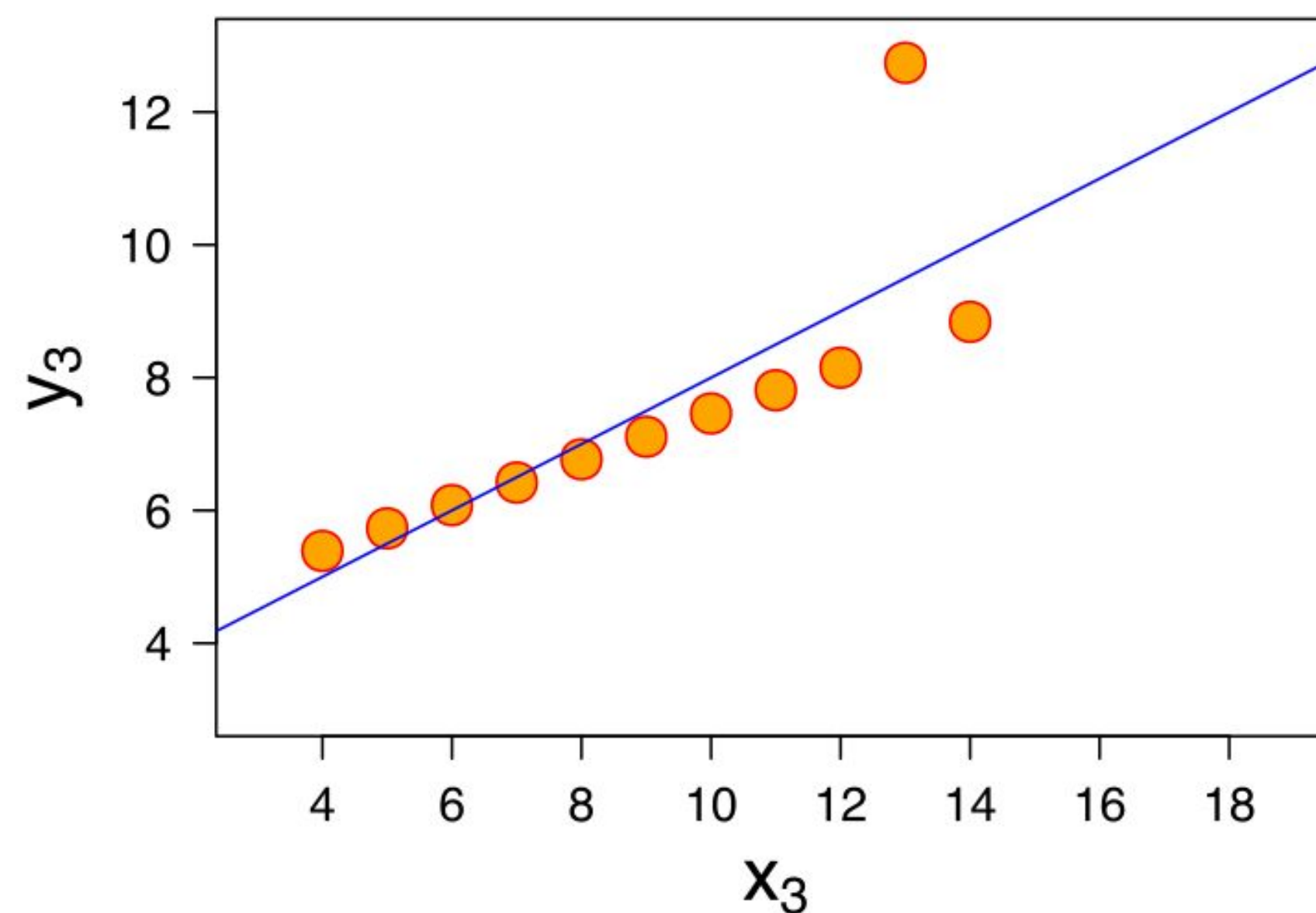
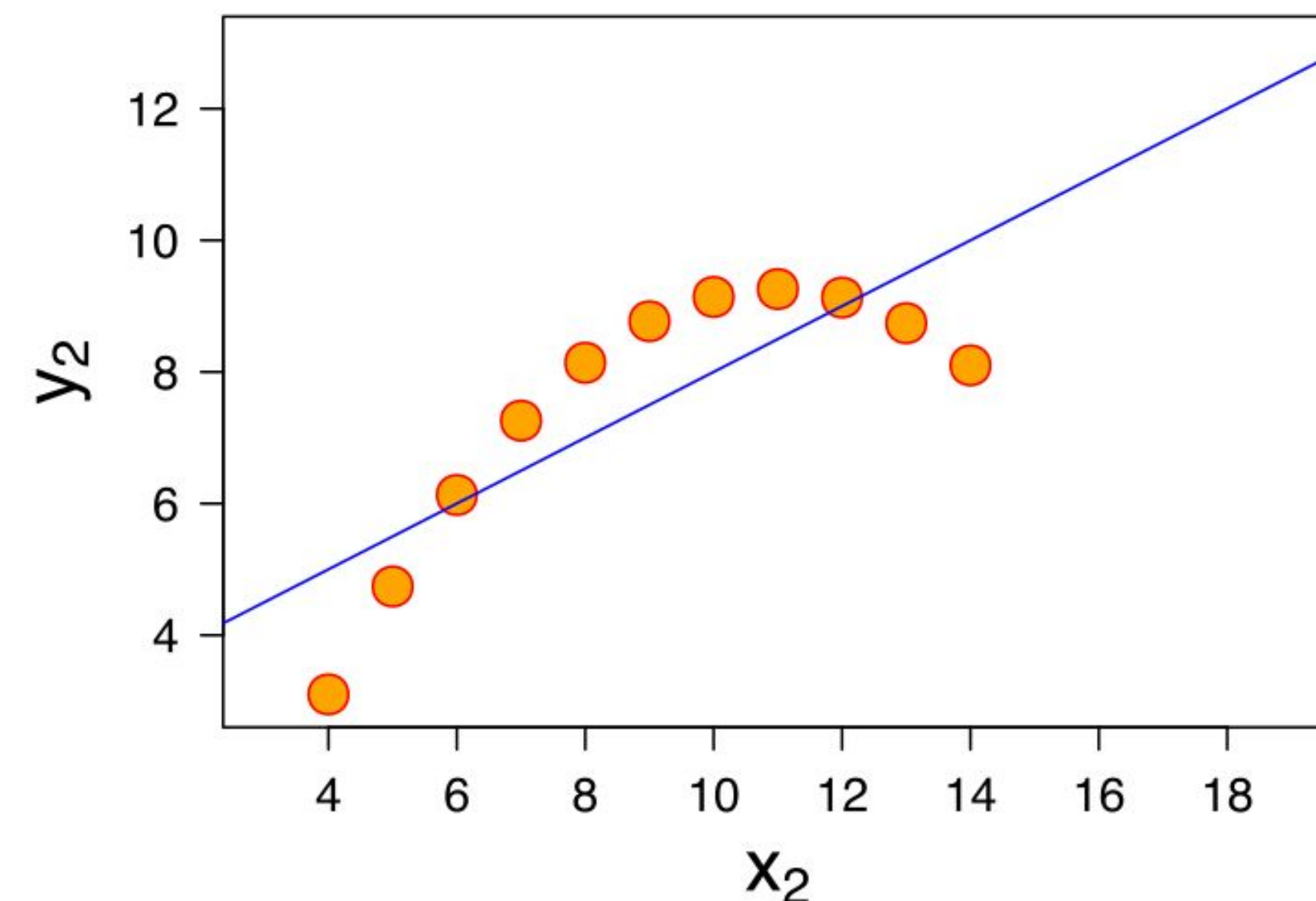
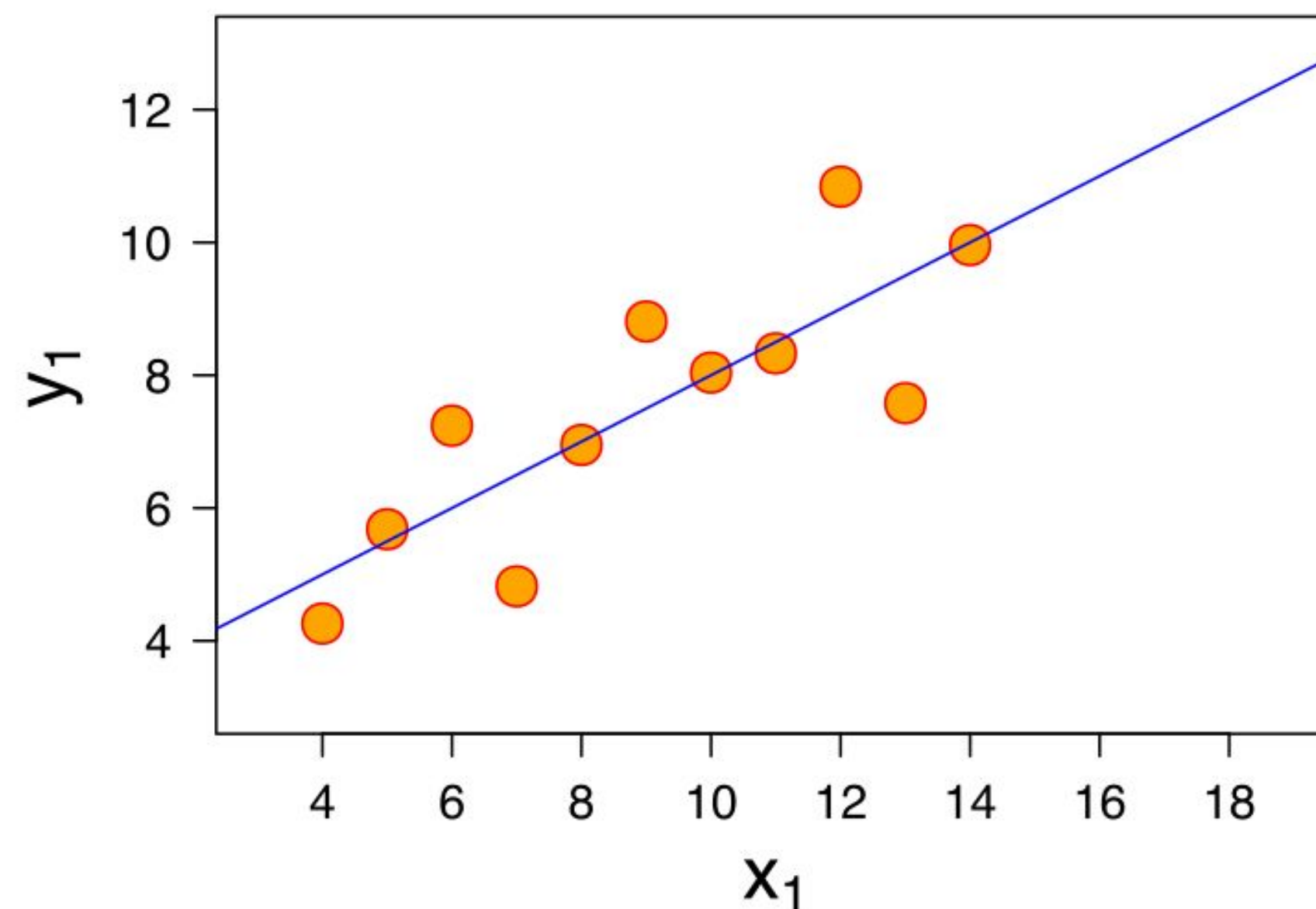
- Adapta a Regressão Linear para problemas de Classificação



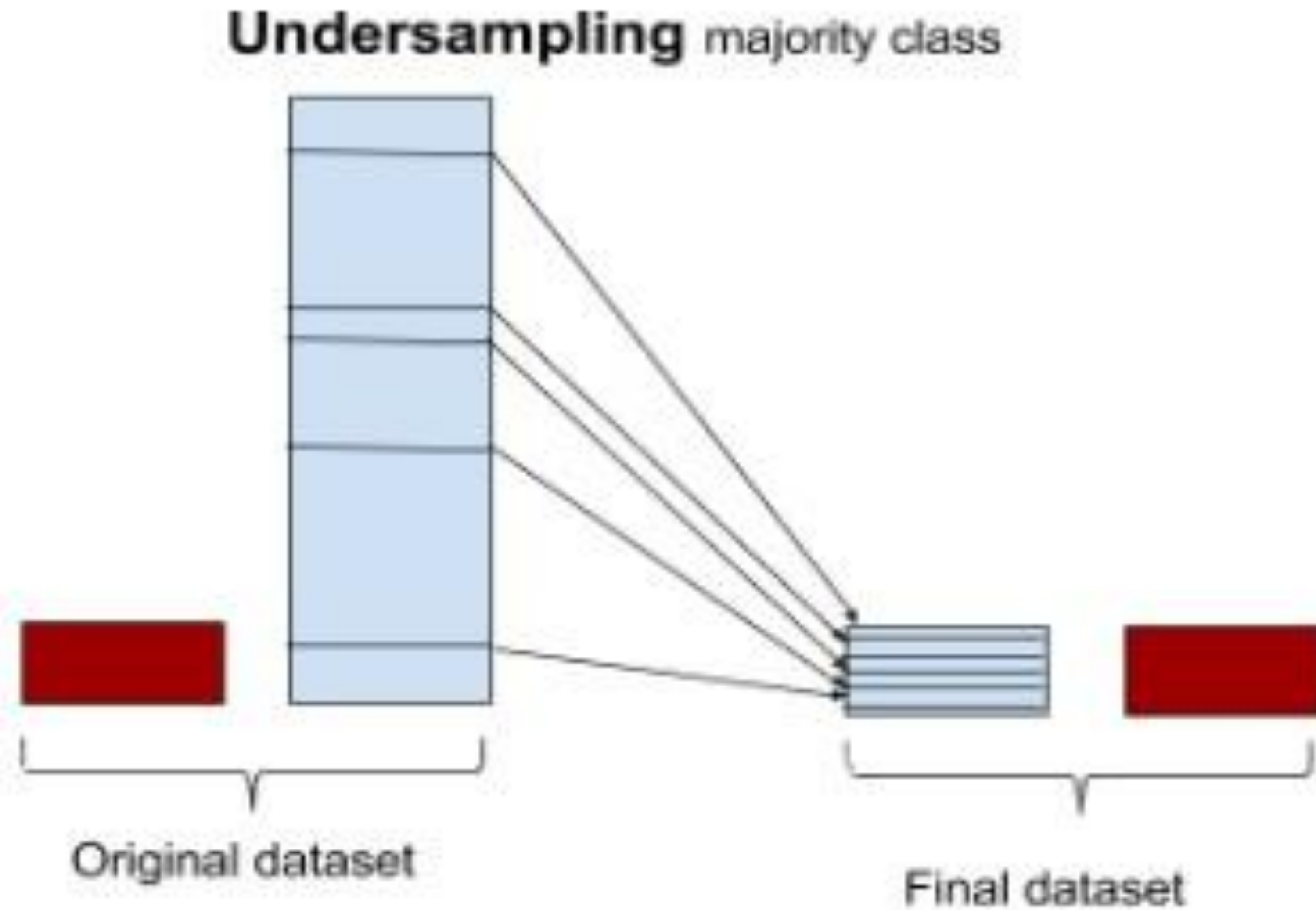
T

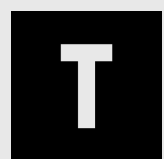
Outliers

- Modelos Lineares não são robustos contra outliers
- É necessário tratá-los



■ Balanceamento: undersampling





DÚVIDAS?

