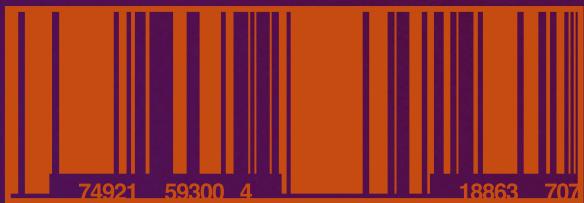
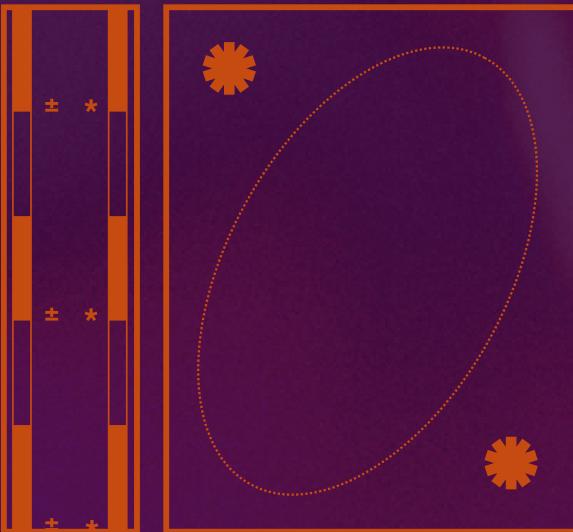


# APRENDIZADO DE MÁQUINA



DIA 1



# **ANTES DE COMEÇARMOS. . .**

## **QUEM SOMOS NÓS? QUEM É O HYPE?**



# **GRUPO DE ESTUDOS E EXTENSÃO DE ALUNOS DA USP FOCADO EM DADOS E INTELIGÊNCIA ARTIFICIAL**

- **Criação de Projetos e pesquisas**
- **Produção de Eventos internos e externos**



**NOS SIGA NO YOUTUBE, LINKEDIN E INSTAGRAM  
PARA SABER MAIS DE EVENTOS FUTUROS!**



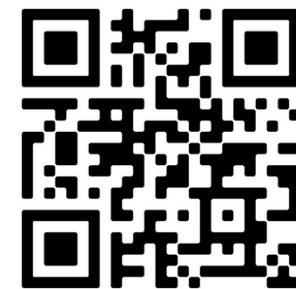
**Hype - Data & AI**



**@hype.usp**



**Hype - Data & AI**



# INSTRUÇÕES:

- Ficará fixado o formulário de presença no chat do Youtube
- Preencha apenas uma vez!



# INSTRUÇÕES:

- **Fizemos exercícios no google colab para prática e revisão!**
- **Link na descrição do Youtube**

The screenshot shows a Google Colab notebook titled "Dia\_1\_Curso\_ML\_Hype.ipynb". The interface includes a toolbar with "Arquivo", "Editar", "Ver", "Inserir", "Ambiente de execução", "Ferramentas", and "Ajuda". Below the toolbar are tabs for "Comandos", "+ Código", "+ Texto", and "Executar tudo". A sidebar on the left lists sections: "3. Treinar o modelo" (selected), "Selecionamos", "Removemos", "Convertemos", "Separamos", and "df". The main content area contains Python code for data preprocessing and model training:

```
[ ] from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression

# Selecionamos características (colunas) de interesse
df = df[["age", "education-num", "hours-per-week", "income-per-year"]]

# Removemos amostras em que alguma coluna tem valor inválido
df = df.dropna()

# Convertemos nossa coluna de previsão (feature target) para 0 ou 1, pois esse modelo não trabalha com strings (palavras), apenas inteiros
df["income-per-year"] = df["income-per-year"].map({">50K": 1, "<=50K": 0})

# Separamos o dataset em X (características) e y (target)
X = df.drop("income-per-year", axis=1)
y = df["income-per-year"]
```

# APRESENTADOR



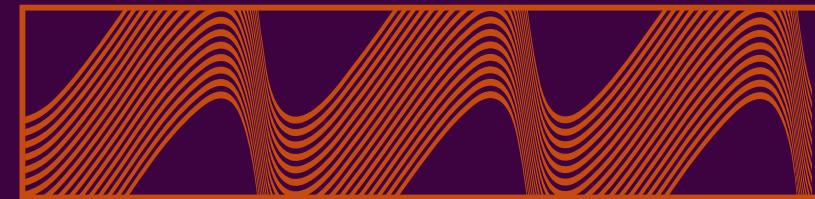
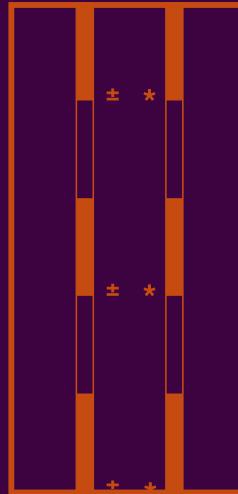
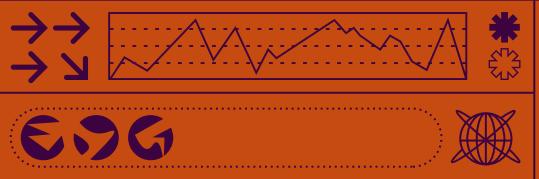
Heitor Gama Ribeiro

Convidado Hype

# APRENDIZADO DE MÁQUINA



## Tópicos explorados na Aula:



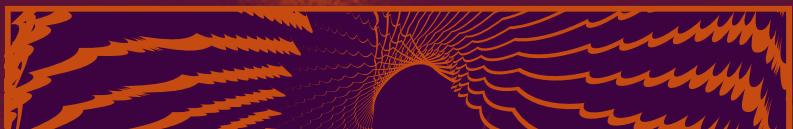
- » Fundamentos de Machine Learning
- » Problemas Clássicos
- » Etapas de um Projeto
- » Modelos Supervisionados e Não Supervisionados
- » Métricas e Avaliação

AGENDA AGENDA AGENDA AGENDA AGENDA AGENDA AGENDA



# O que é Aprendizado de Máquina?

Métodos computacionais baseados em algum tipo de experiência que melhorem a performance de uma tarefa ou façam previsões acuradas



# APLICAÇÕES

## Jogos

### Movimentação de NPCs

Aprendizado por Reforço para um agente (NPC) evitar obstáculos e aprender trajetórias sem supervisão humana

## Finanças

## Saúde

## Redes Sociais

## Vendas

# APLICAÇÕES

## Jogos

### Movimentação de NPCs

Aprendizado por Reforço para um agente (NPC) evitar obstáculos e aprender trajetórias sem supervisão humana

## Redes Sociais

## Finanças

### Previsão do mercado

Uso dos dados históricos de ações e sentimentos de twitts para "tentar" prever o preço de uma ação de maneira legal

## Saúde

## Vendas

# APLICAÇÕES

## Jogos

### Movimentação de NPCs

Aprendizado por Reforço para um agente (NPC) evitar obstáculos e aprender trajetórias sem supervisão humana

## Redes Sociais

## Finanças

### Previsão do mercado

Uso dos dados históricos de ações e sentimentos de twitts para "tentar" prever o preço de uma ação de maneira legal

## Vendas

## Saúde

### Diagnóstico precoce

Usar o histórico do paciente e imagens de exames para averiguar a possibilidade de desenvolvimento de doenças com alta precisão

# APLICAÇÕES

## Jogos

### Movimentação de NPCs

Aprendizado por Reforço para um agente (NPC) evitar obstáculos e aprender trajetórias sem supervisão humana

## Finanças

### Previsão do mercado

Uso dos dados históricos de ações e sentimentos de twitts para "tentar" prever o preço de uma ação de maneira legal

## Saúde

### Diagnóstico precoce

Usar o histórico do paciente e imagens de exames para averiguar a possibilidade de desenvolvimento de doenças com alta precisão

## Redes Sociais

### Recomendação de conteúdos

Utilização das diferentes variáveis como texto, tempo de tela e número de interações com outros usuários para recomendar conteúdos

## Vendas

# APLICAÇÕES

## Jogos

### Movimentação de NPCs

Aprendizado por Reforço para um agente (NPC) evitar obstáculos e aprender trajetórias sem supervisão humana

## Finanças

### Previsão do mercado

Uso dos dados históricos de ações e sentimentos de twitts para "tentar" prever o preço de uma ação de maneira legal

## Saúde

### Diagnóstico precoce

Usar o histórico do paciente e imagens de exames para averiguar a possibilidade de desenvolvimento de doenças com alta precisão

## Redes Sociais

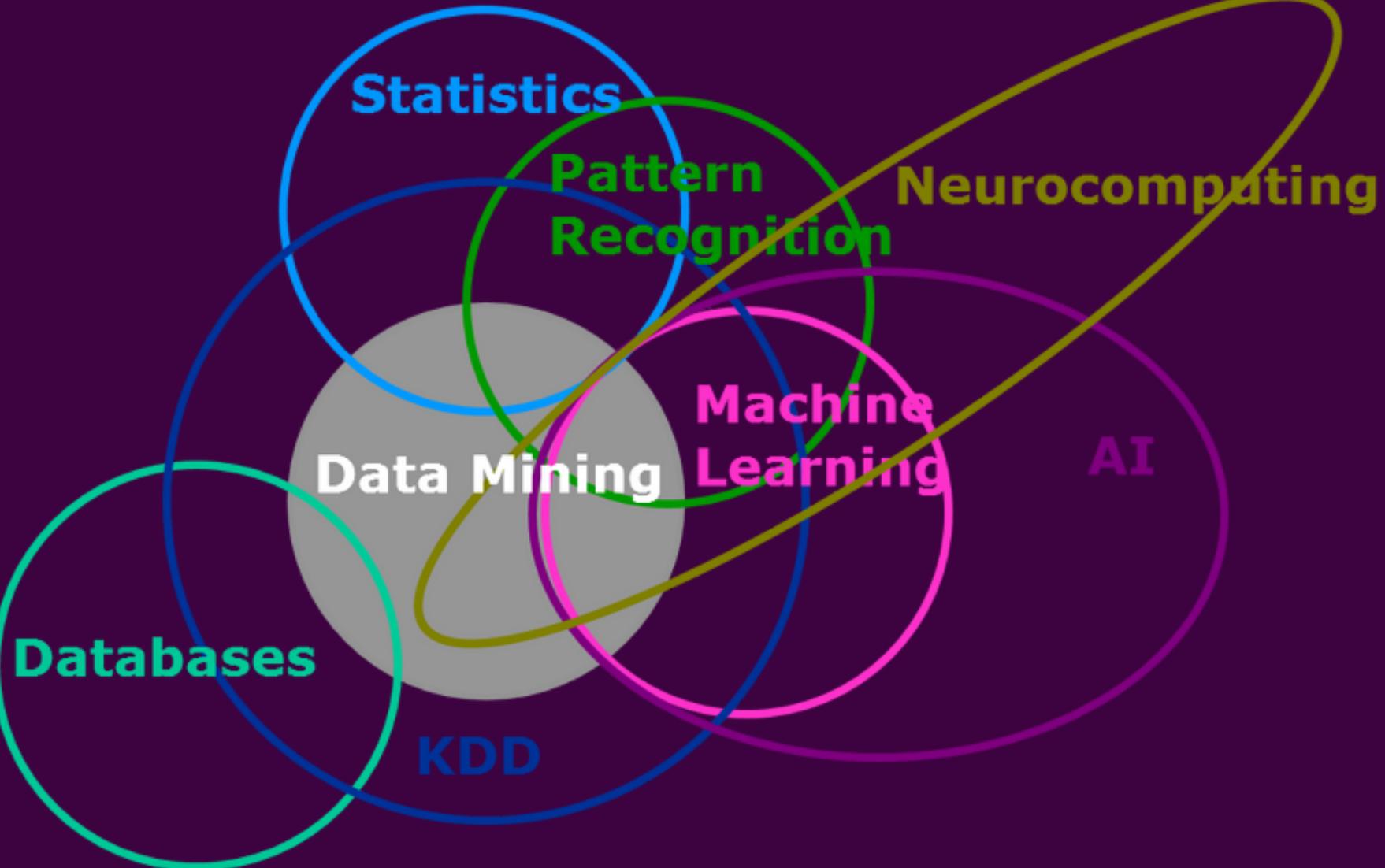
### Recomendação de conteúdos

Utilização das diferentes variáveis como texto, tempo de tela e número de interações com outros usuários para recomendar conteúdos

## Vendas

### Amazon e os produtos que você pode gostar

Reconhecimento dos "gostos" de um cliente e a utilização de informações demográficas e de outros usuários para oferecer produtos de seu interesse



# OS PASSOS DE UM PROJETO DE MACHINE LEARNING

## PASSO 1

### Coletar os dados

Pesquise sobre seu problema e procure dados para resolvê-lo. Sempre se preocupe com a qualidade deles

## PASSO 2

## PASSO 3

## PASSO 4

## PASSO 5

# OS PASSOS DE UM PROJETO DE MACHINE LEARNING

## PASSO 1

### Coletar os dados

Pesquise sobre seu problema e procure dados para resolvê-lo. Sempre se preocupe com a qualidade deles

## PASSO 2

### Preparar seus dados

Modelos de Machine Learning precisam receber seus dados em um certo formato

## PASSO 3

## PASSO 4

## PASSO 5

# OS PASSOS DE UM PROJETO DE MACHINE LEARNING

## PASSO 1

### Coletar os dados

Pesquise sobre seu problema e procure dados para resolvê-lo. Sempre se preocupe com a qualidade deles

## PASSO 2

### Preparar seus dados

Modelos de Machine Learning precisam receber seus dados em um certo formato

## PASSO 3

### Escolher e treinar um modelo

Treinar um modelo no conjunto de dados evitando enviesar a maneira em que são feitas as decisões do modelo

## PASSO 4

## PASSO 5

# OS PASSOS DE UM PROJETO DE MACHINE LEARNING

## PASSO 1

### Coletar os dados

Pesquise sobre seu problema e procure dados para resolvê-lo. Sempre se preocupe com a qualidade deles

## PASSO 2

### Preparar seus dados

Modelos de Machine Learning precisam receber seus dados em um certo formato

## PASSO 3

### Escolher e treinar um modelo

Treinar um modelo no conjunto de dados evitando enviesar a maneira em que são feitas as decisões do modelo

## PASSO 4

### Avaliar os resultados

Gerar previsões do seu modelo treinado e verificar se elas parecem condizentes ao esperado

## PASSO 5

# OS PASSOS DE UM PROJETO DE MACHINE LEARNING

## PASSO 1

### Coletar os dados

Pesquise sobre seu problema e procure dados para resolvê-lo. Sempre se preocupe com a qualidade deles

## PASSO 2

### Preparar seus dados

Modelos de Machine Learning precisam receber seus dados em um certo formato

## PASSO 3

### Escolher e treinar um modelo

Treinar um modelo no conjunto de dados evitando enviesar a maneira em que são feitas as decisões do modelo

## PASSO 4

### Avaliar os resultados

Gerar previsões do seu modelo treinado e verificar se elas parecem condizentes ao esperado

## PASSO 5

### Aplicar o modelo e/ou melhorá-lo

Aplicar seu projeto em uma ambiente real, fazendo os ajustes necessários para melhoria continua

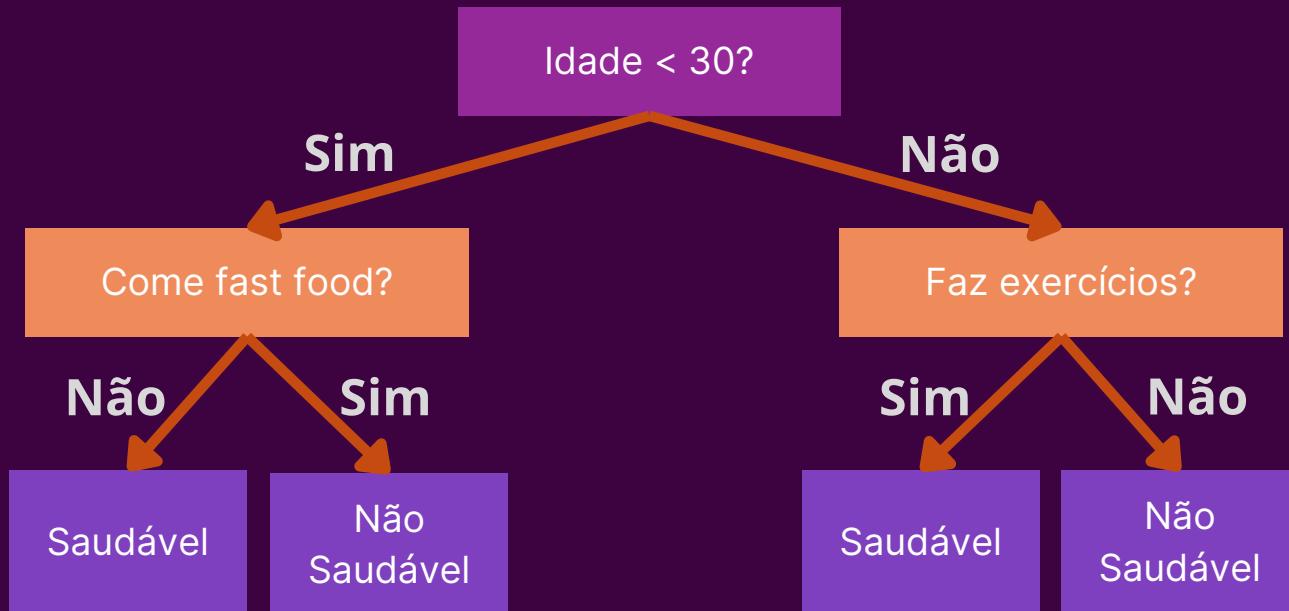
# EXEMPLO SIMPLES DE MÓDULO

Uma pessoa é saudável?

- ÁRVORE DE DECISÃO  
PARA UM PROBLEMA  
INTRODUTÓRIO

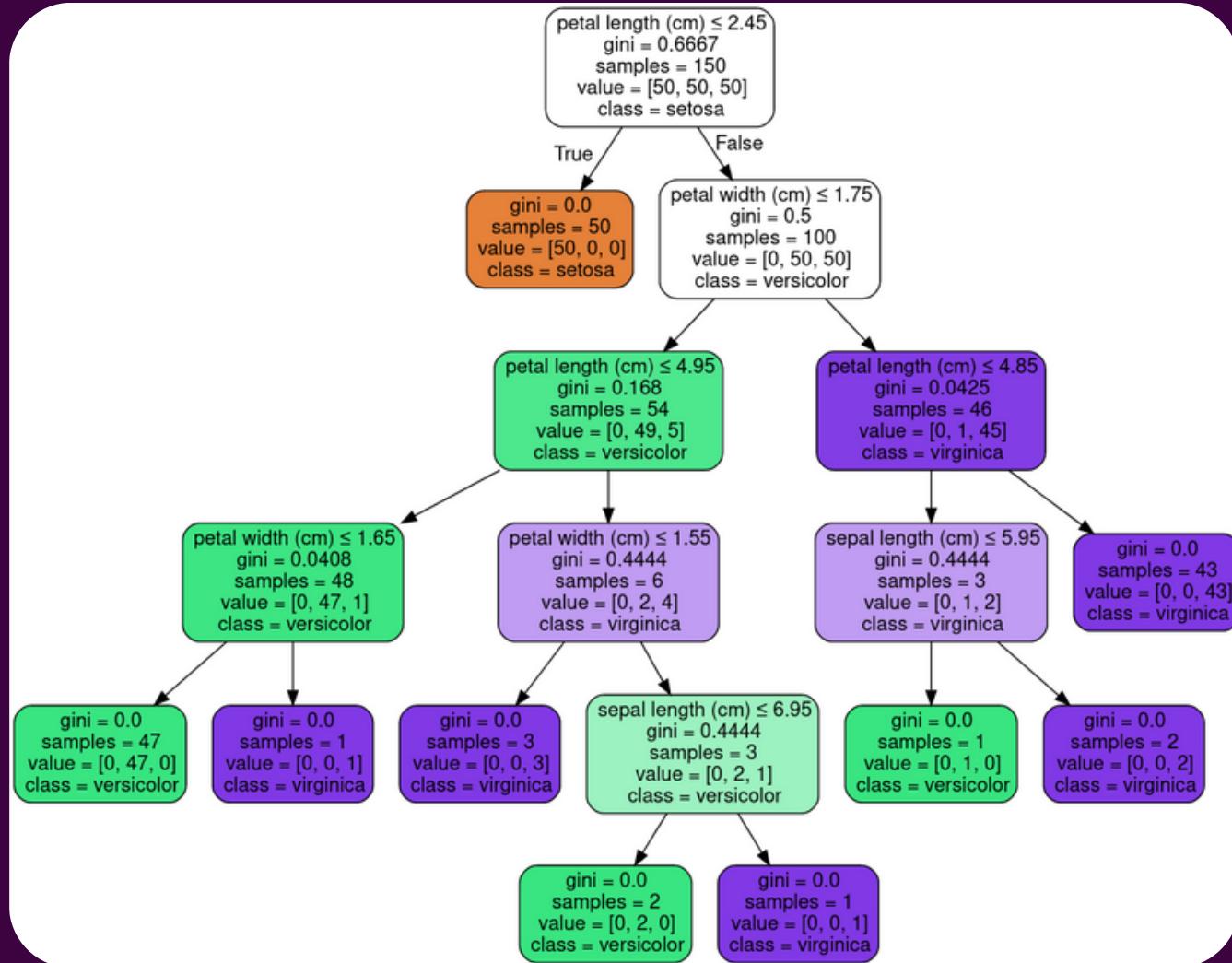


O modelo parece  
pertinente?



# OUTRO MODELO

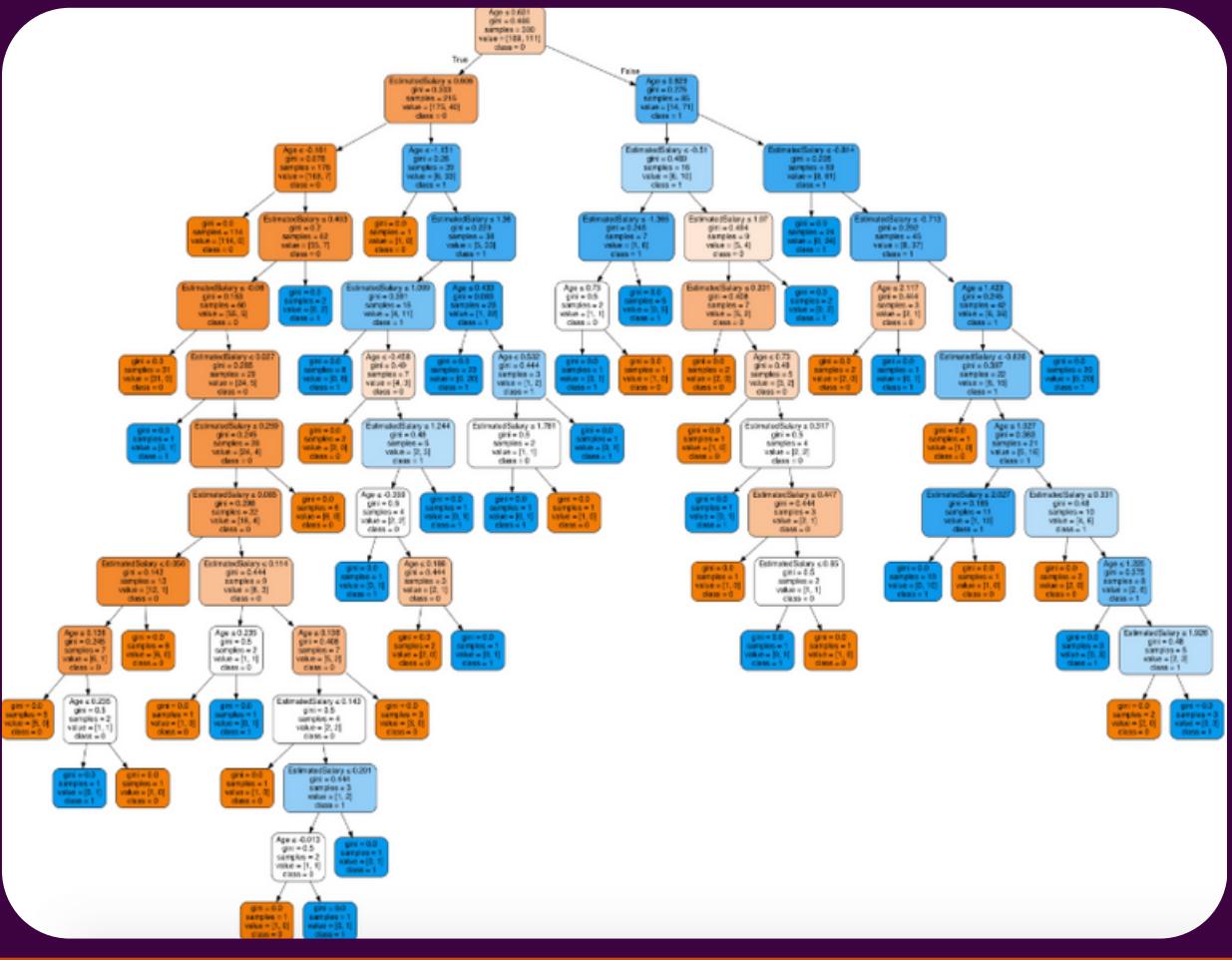
**Uma árvore de decisão  
para um problema com  
mais possibilidades**



# COMPLEXIDADE

## Complexidade X Interpretabilidade

- Muitos caminhos
- Resultados mais precisos?





# GRUPOS DE MODELOS

- Como devem ser os dados usados em um algoritmo de ML?
- Como os modelos operam sobre os dados?



# TRÊS GRANDES GRUPOS

**De acordo com a estrutura dos dados e a maneira como o algoritmo opera sobre eles, os modelos de Machine Learning podem ser divididos em:**

**Modelos  
Supervisionados**

**Modelos Não  
Supervisionados**

**Modelos por  
Reforço**

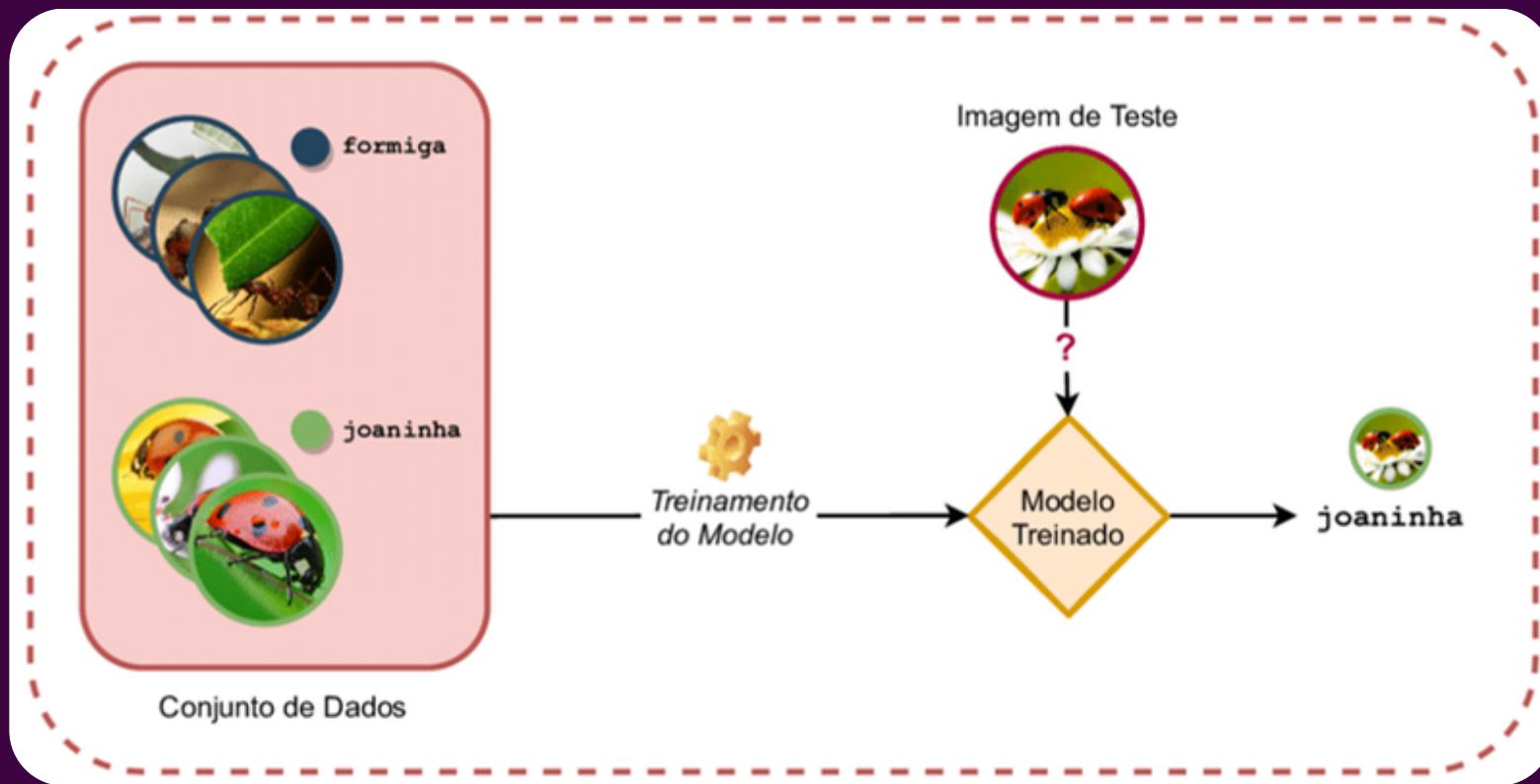
**Seus objetivos e casos de uso são distintos. Portanto, é de extrema importância conhecê-los, para saber quando cada um deve ser aplicado.**

# MODELOS SUPERVISIONADOS

**Palavra-chave:**  
Rótulos

- Os dados utilizados para treinar modelos supervisionados devem necessariamente ser rotulados.
  - Ou seja, um de seus atributos deve ser a resposta esperada para aquele dado.

# MODELOS SUPERVISIONADOS



**Fonte: Research Gate**

# MODELOS SUPERVISIONADOS

**Esse grupo de modelos atende a dois principais tipos de problemas:**

## Problemas de Classificação

- **Problemas de Classificação:**
  - **Rótulos categóricos**
  - **Exemplo de caso → indicar se uma célula é cancerígena / não**
  - **Algoritmos → Árvores de Decisão, KNN, etc.**

## Problemas de Regressão

- **Problemas de Regressão:**
  - **Rótulos são números contínuos**
  - **Exemplo de caso → projetar o valor de uma ação na bolsa**
  - **Algoritmos → Regressão Linear, Regressão Lasso, etc.**

# MODELOS SUPERVISIONADOS

- Problemas de Classificação:



**Fonte: Stoodi**

# MODELOS SUPERVISIONADOS

- Problemas de Regressão:

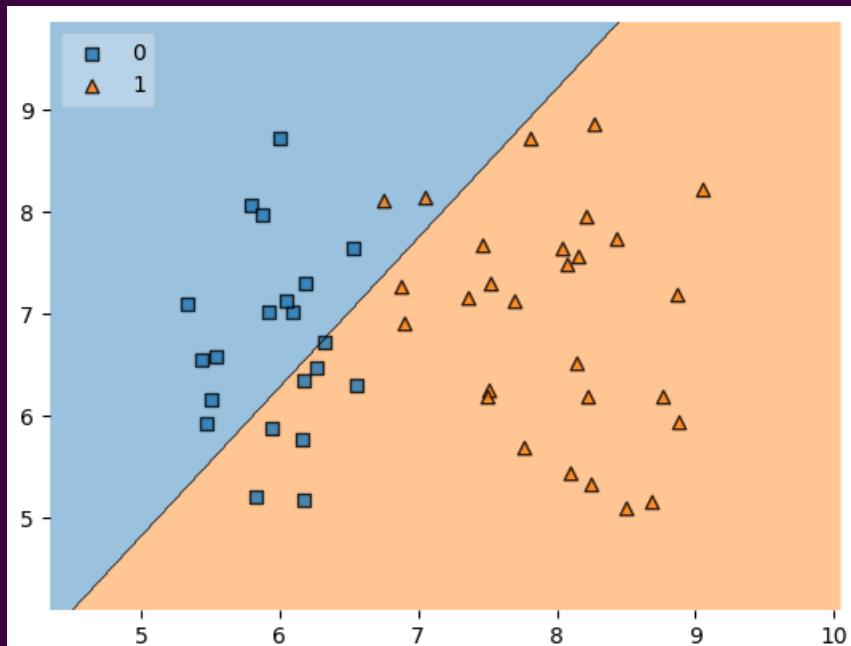


Fonte: Riconnect

# Classificação

O rótulo é uma categoria

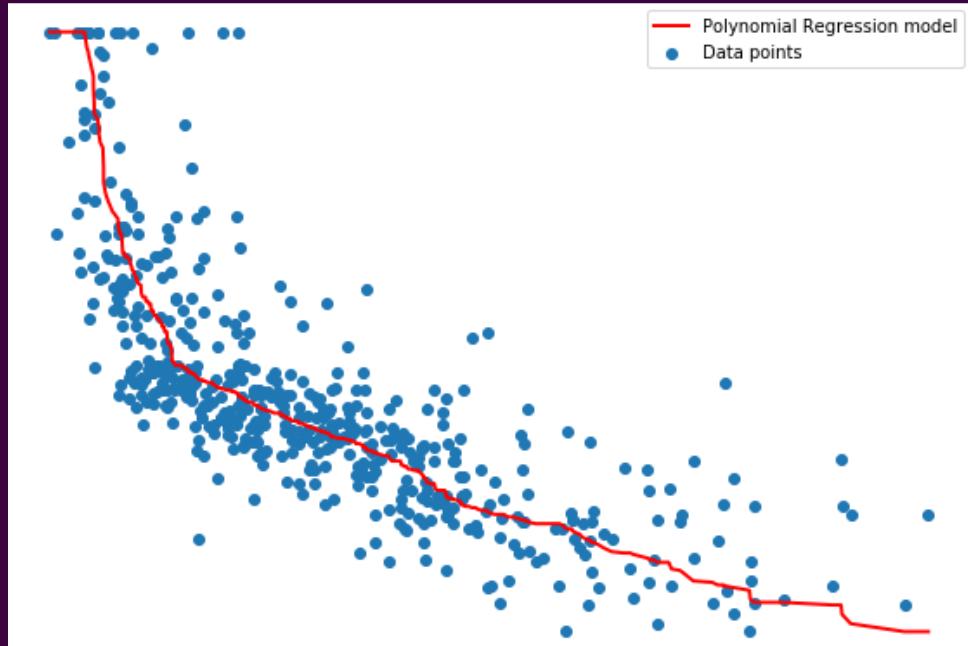
Exemplo: Email spam ou não



# Regressão

O rótulo é um valor numérico

Exemplo: Preço de casas

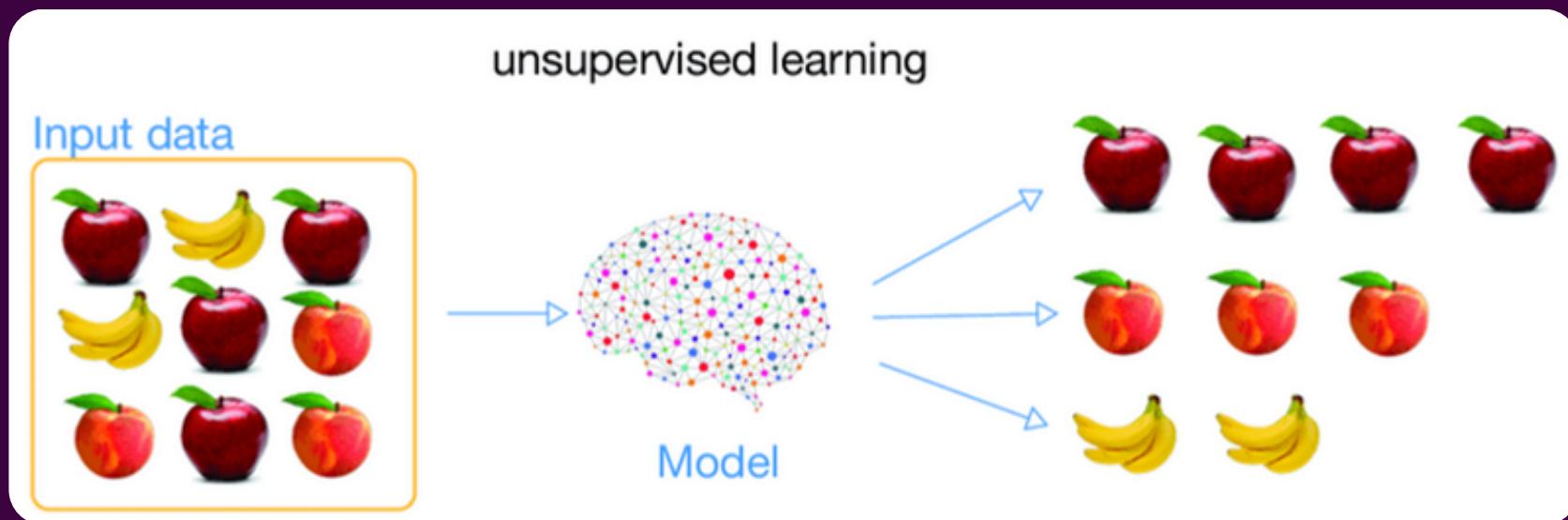


# **MODELOS NÃO SUPERVISIONADOS**

**Palavra-chave:**  
**Descoberta de Padrões**

- Os dados utilizados para treinar modelos não supervisionados não podem ter rótulos.**
- O algoritmo se encarrega de identificar padrões nos dados e, com isso, propor segmentações.**

# MODELOS NÃO SUPERVISIONADOS



**Fonte: Research Gate**

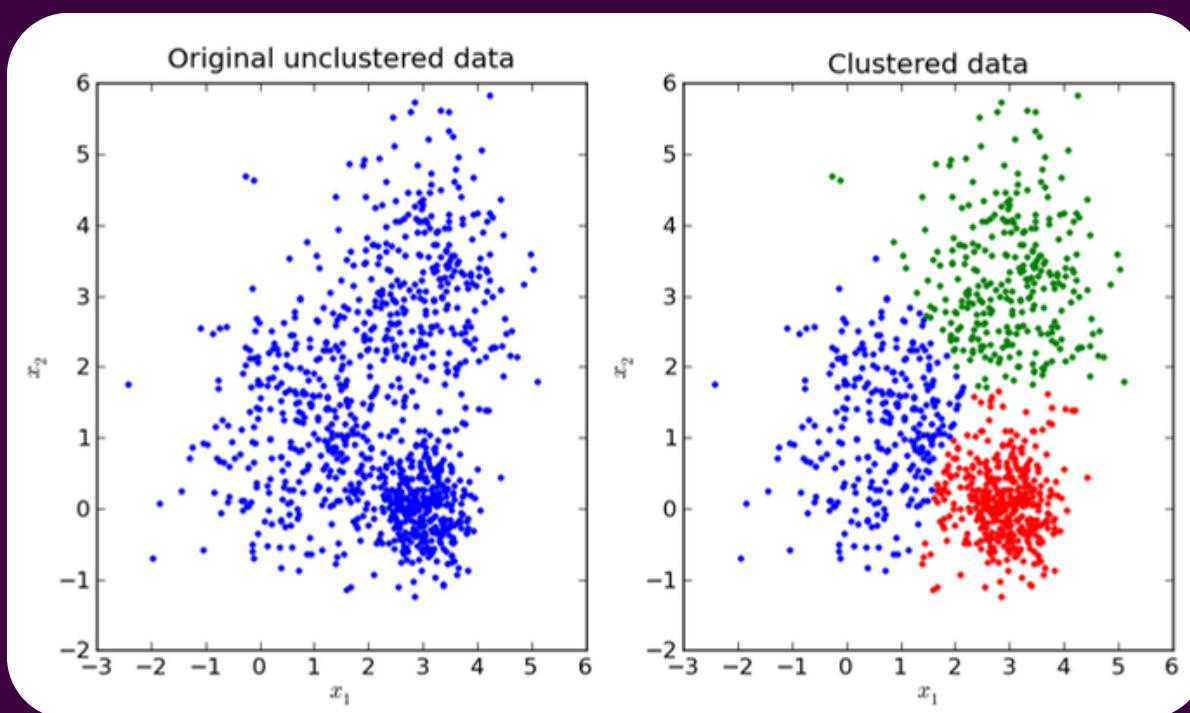
# **MODELOS NÃO SUPERVISIONADOS**

**Esse grupo de modelos atende principalmente a problemas de:**

## **Clusterização**

- Separar os dados em grupos**
- Maximizando a semelhança entre dados do mesmo grupo**
- Minimizando a semelhança entre dados de grupos diferentes**
- Exemplo de caso → segmentar clientes em um projeto de marketing**
- Algoritmos → K-Means, ROM, etc.**

# MODELOS NÃO SUPERVISIONADOS



Fonte: Linedata

# APRENDIZADO POR REFORÇO

**Esse grupo de modelos é arquitetado com os seguintes elementos:**

Estado

Ação

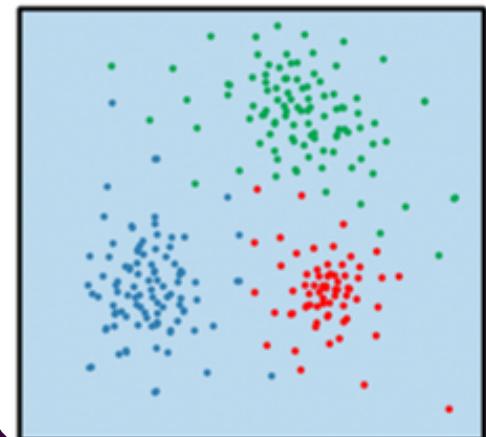
Recompensa

- **Interação agente-ambiente:** estado → ação → recompensa.
- **Objetivo:** maximizar a recompensa acumulada.
- **Exploração × exploração:** achar novas estratégias sem abandonar as que já funcionam.
- **Exemplo:** agente jogando Atari ou controlando um robô.
- **Algoritmos:** Q-Learning, DQN, PPO, Policy Gradient.

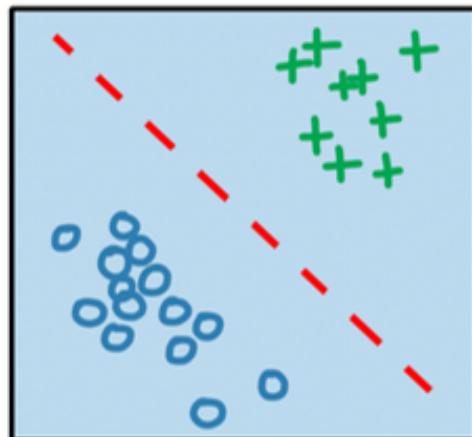
# VISÃO GERAL

## machine learning

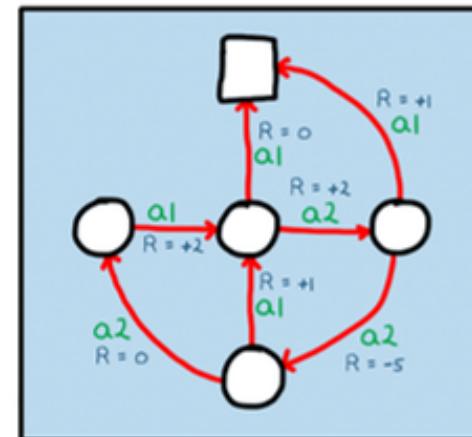
unsupervised  
learning



supervised  
learning

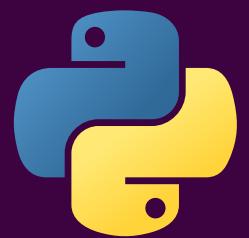


reinforcement  
learning



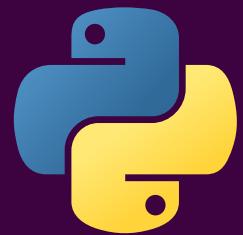
Fonte: DevOpsSchool

**E COMO APLICAR ISSO  
TUDO?**



python





# python

- **linguagem acessível, simples e fácil de entender;**
- **Comunidade ativa para auxílio;**
- **Bibliotecas diversas e especializadas em Machine Learning:**
- **Pandas, Matplotlib, Seaborn, Tensorflow, Scikit-learn, etc.**



- Classificação
- Regressão
- Clusterização
- Seleção e Validação de modelo
- Pré-processamento

# scikit-learn

Machine Learning in Python

[Getting Started](#)[Release Highlights for 1.1](#)[GitHub](#)

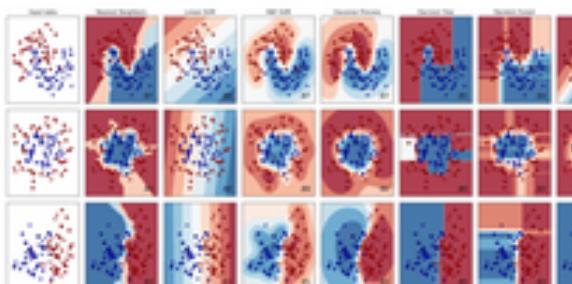
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

## Classification

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, and more...

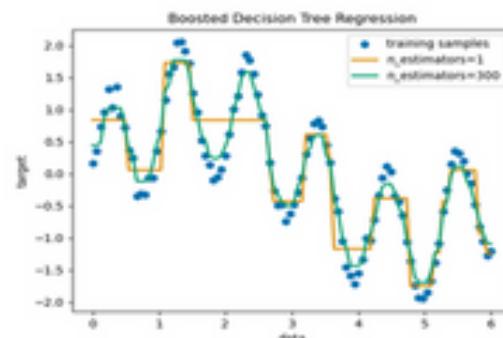
[Examples](#)[Dimensionality reduction](#)

## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, nearest neighbors, random forest, and more...

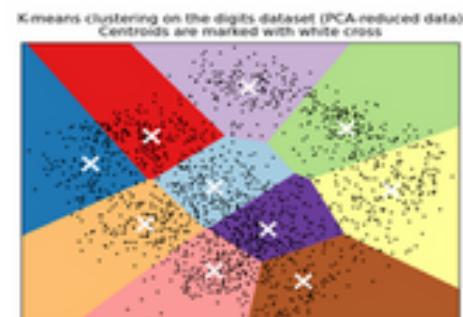
[Examples](#)[Model selection](#)

## Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, and more...

[Examples](#)[Preprocessing](#)



- Organização
- Executar o código localmente
- Facilidade de uso

# **PROBLEMAS NOS TREINOS**

**Temos um Problema!**

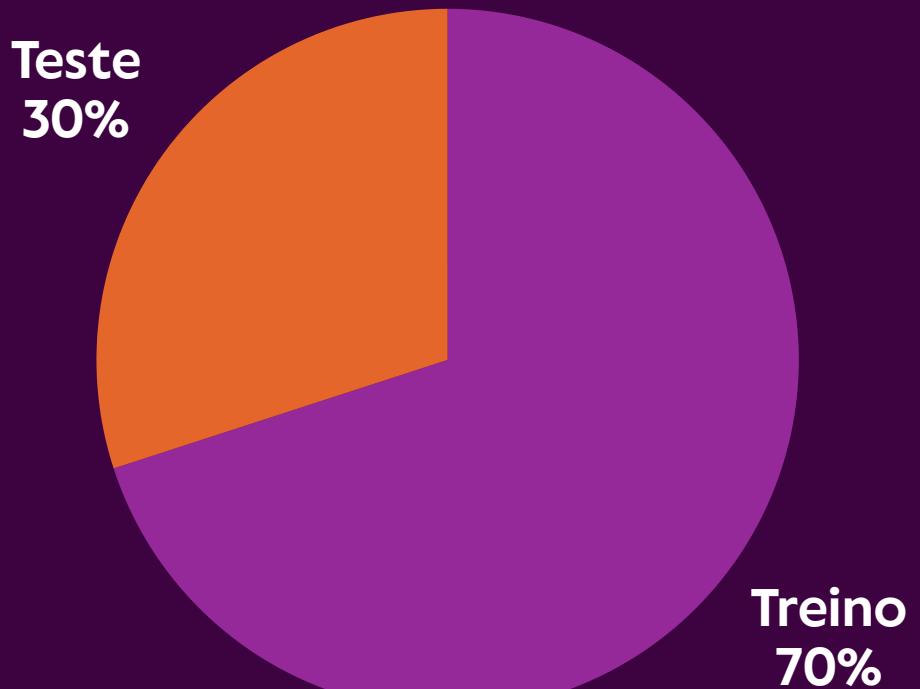
**O modelo está sendo treinado e  
avaliado com os mesmos dados**

**Por que isso não faz sentido?**

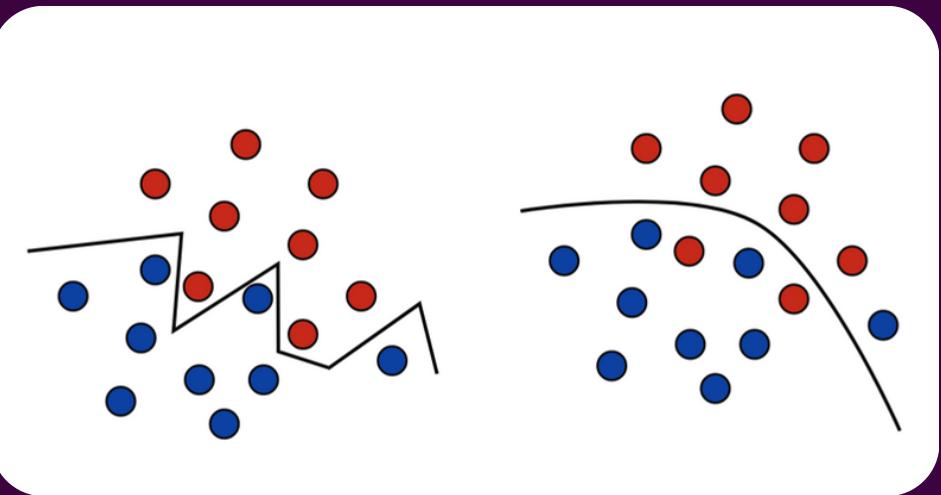
# Separação em treino e teste

**NUNCA SE MEXE EM TESTE, APÓS SEPARADO!**

- Amostragem aleatória
- Eventuais pré-processamentos precisam acontecer em treino e teste
- 70% treino e 30% teste - Questionável (quantidade de dados)



# Generalização



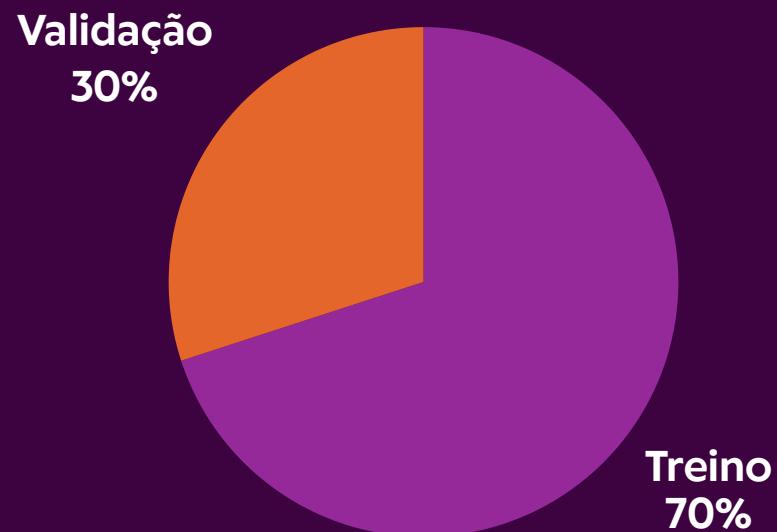
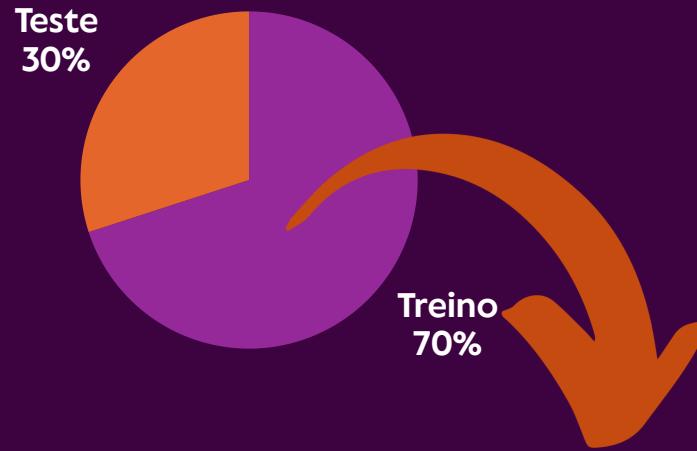
MODELOS PRECISAM SER BONS PARA  
DADOS NÃO VISTOS!

- Como saber se nosso modelo realmente é bom para dados não vistos?
  - Treino e Teste
- Como chegar à generalização?
  - Viés com os dados de treinamento
  - Variância para poder acertar os dados de teste
- Também há influência do modelo e seus hiperparâmetros

# Validação

**SE NÃO VEMOS TESTE, USAMOS TREINO PARA VALIDAR**

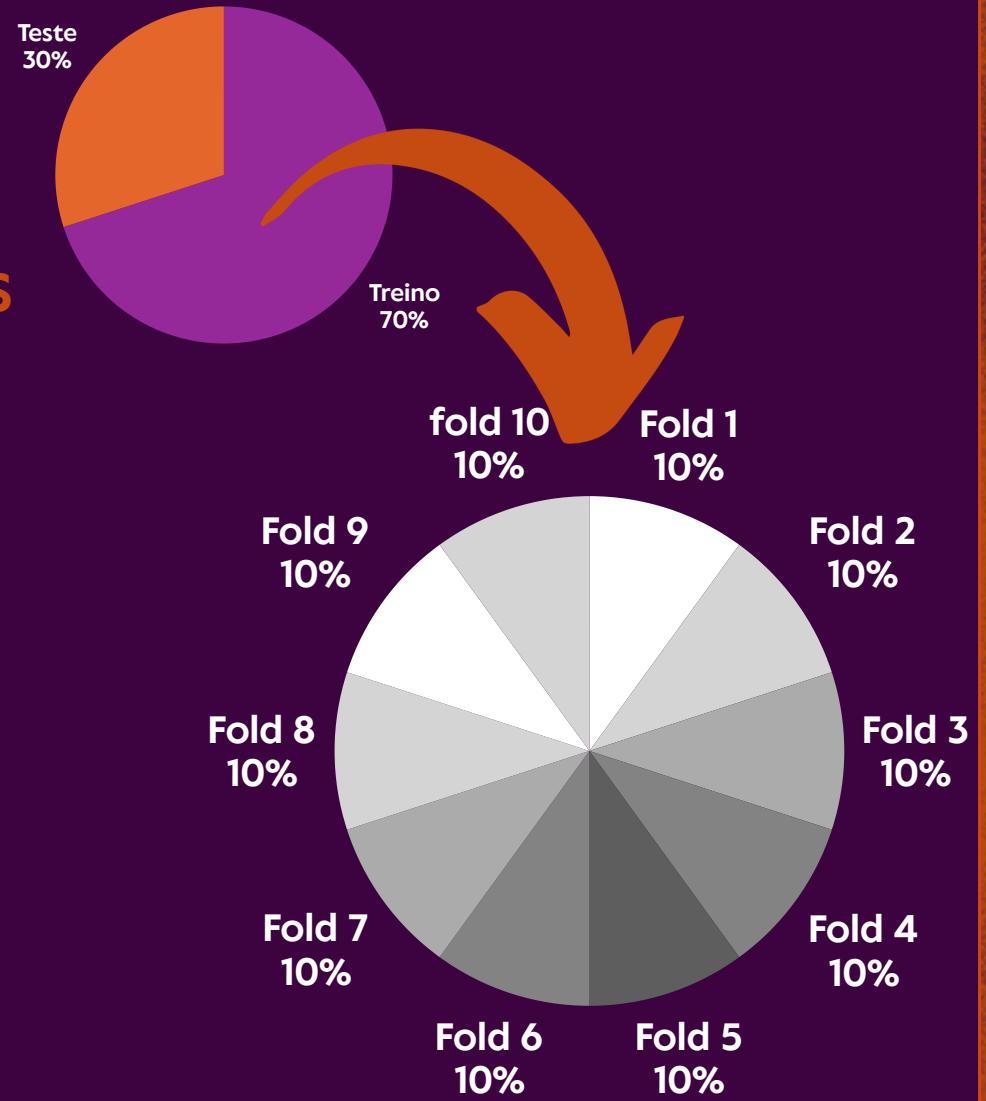
- O que validar?
  - Hiperparâmetros
  - Modelos diferentes
  - Generalização - Resultado das métricas em treino e validação
- Podemos separar novamente em 70% treino e 30% teste
- Ou, para evitar viéses, usar validação cruzada



# Validação cruzada

## SEPARAR OS DADOS DE TREINO EM K-FOLDS

- Normalmente K=10
- Treina-se com K-1, e valida-se com 1
  - Repete-se o processo K vezes
- Verificar a diferença no desempenho de diferentes modelos



# **PRÉ-PROCESSAMENTO**

# Pré-Processamento

- **Variáveis categóricas**
  - `OrdinalEncoding`
  - `OneHotEncoding`
- **Variáveis numéricas**
  - `MinMaxScaler`
  - `StandardScaler`
- **Valores Perdidos**

# Ordinal

- Uma coluna - Se mapeiam os dados para valores numéricos que seguem uma ordem



# OneHot

- Para cada categoria é criada uma coluna

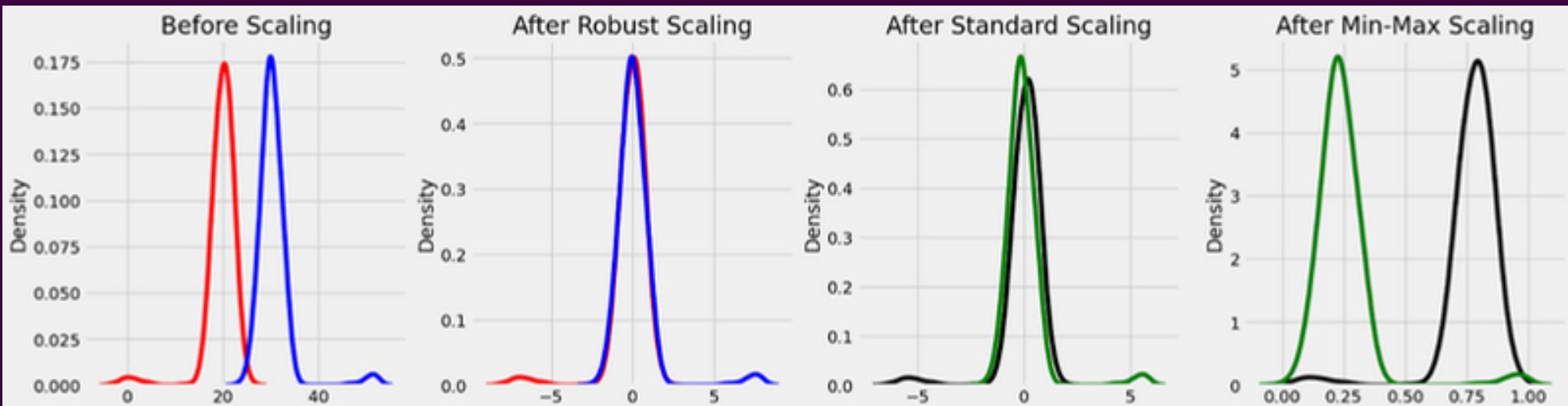
Original
Livros
Eletrônicos
Alimentos
Livros

Livros	Eletrônicos	Alimentos
1	0	0
0	1	0
0	0	1
1	0	0



# MinMaxScaler

- Valores numéricos ficam no intervalo 0-1
- Fraco contra outliers



# StandardScaler

- Assume que os valores seguem uma distribuição normal e os centraliza com média 0 e desvio padrão 1

# Valores faltantes



- Remoção das linhas
- Remoção das colunas
- Inferência por algum algoritmo
- Substituição por moda, média ou mediana
- Interpolação

# **PROBLEMAS E CUIDADOS NECESSÁRIOS**

# Sites para procurar conjuntos de dados

[Kaggle](#)

[UCI](#)

[Google Datasets](#)

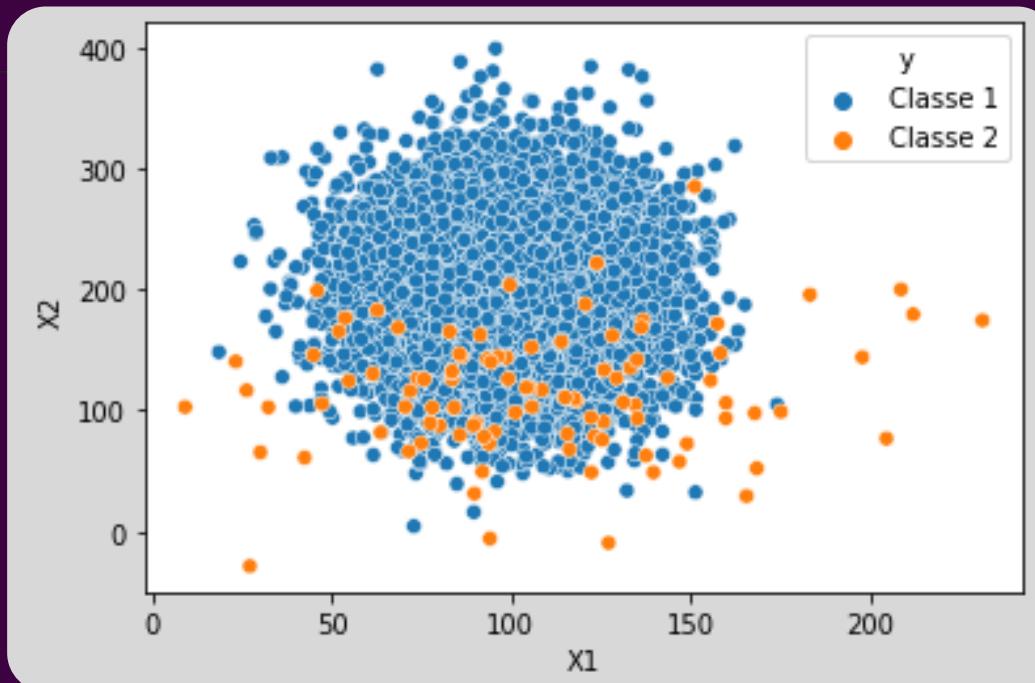
[dados.gov](#)

# Oversampling

Aumento da quantidade de dados da classe minoritária

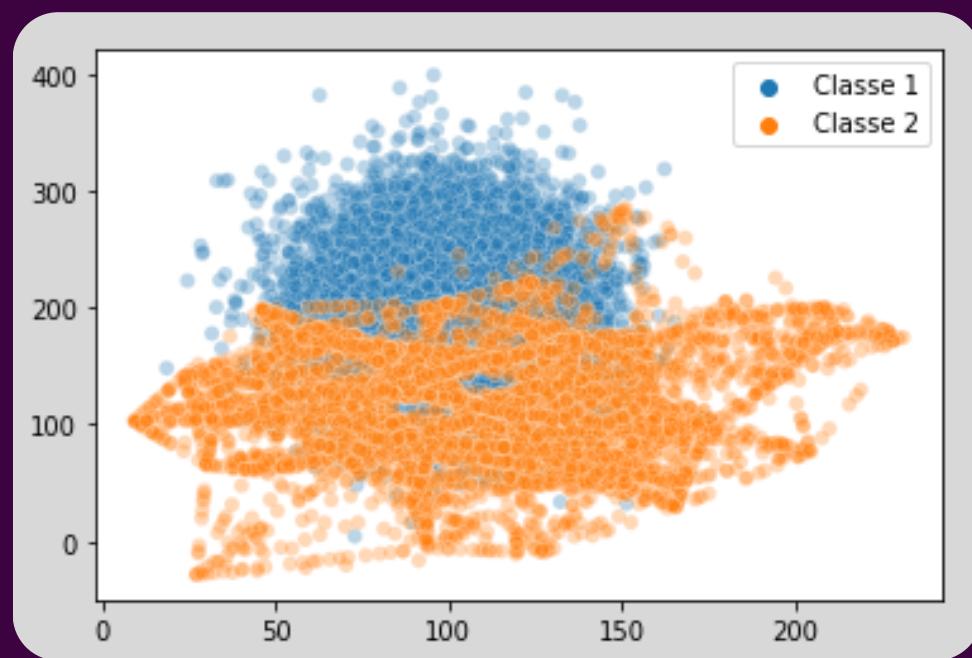
# Undersampling

Diminuição da quantidade de dados da classe majoritária



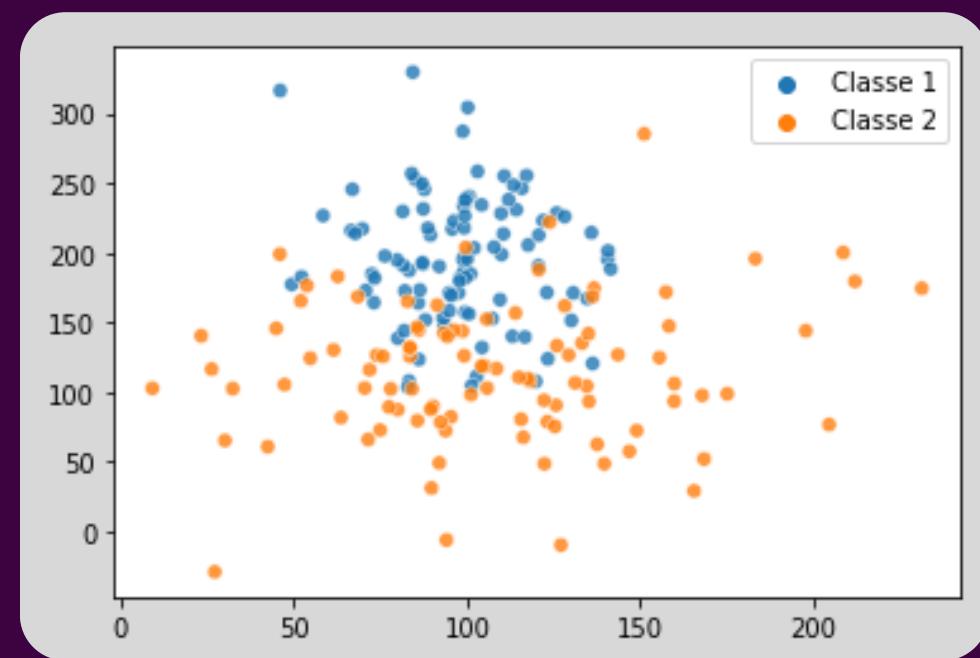
# Oversampling

Aumento da quantidade de dados da classe minoritária



# Undersampling

Diminuição da quantidade de dados da classe majoritária



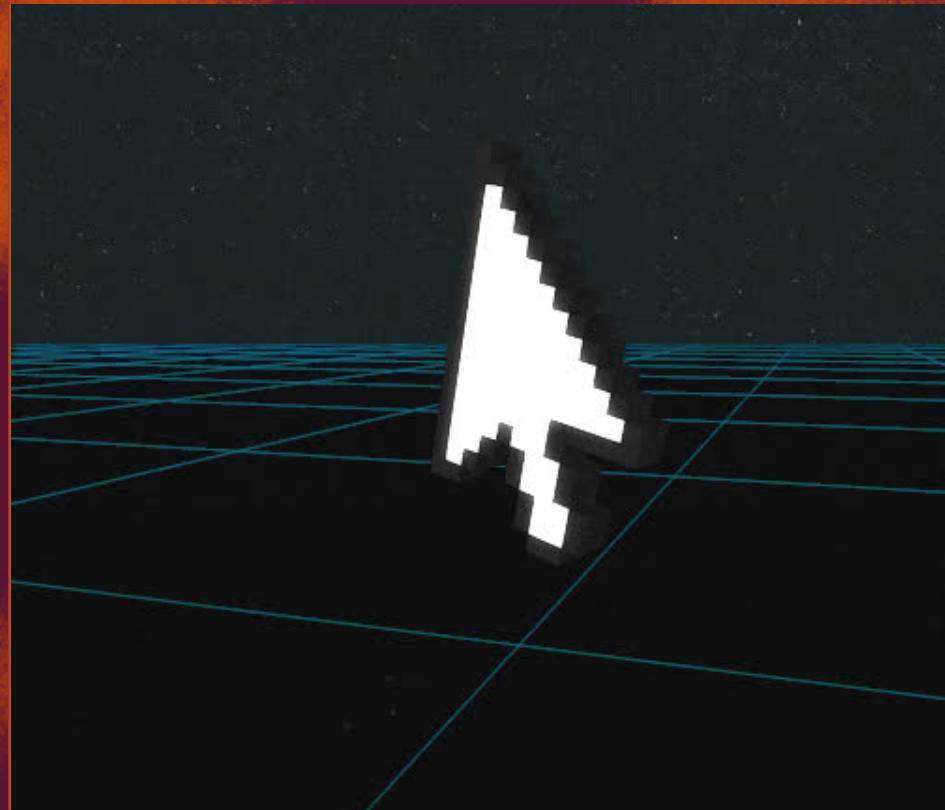


ESG

SG

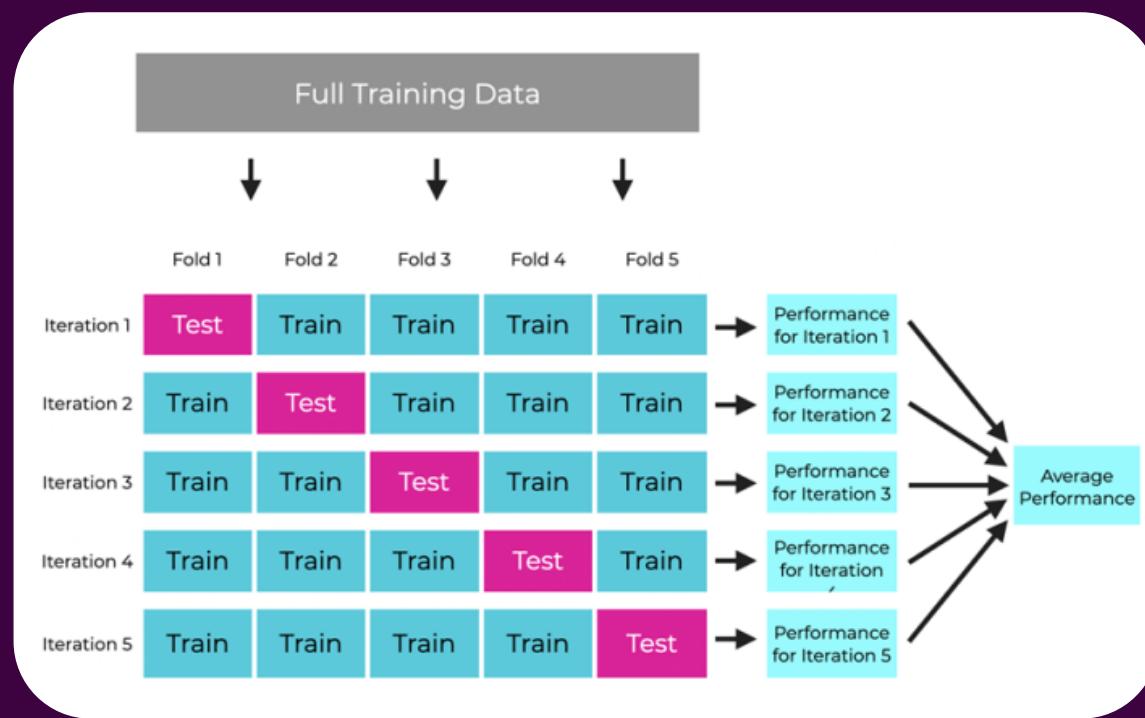
# MÉTRICAS E AVALIAÇÃO

- Como podemos saber se um modelo está adequado?
- Há métricas mais indicadas conforme o contexto?



# AVALIAÇÃO

## CROSS-VALIDATION



# AVALIAÇÃO

## CROSS-VALIDATION

- Visa dar maior credibilidade para a métrica de qualidade
- Dividir o conjunto de dados em K subconjuntos (“folds”);
- Treinar o modelo K vezes:
  - Em cada treino, utilizar o k-ésimo fold para treino e os demais para teste
- Teremos K modelos → K acurácia
- Se a média das K acurácia for alta e seu desvio padrão for baixo → modelo robusto e com muitos acertos

# AVALIAÇÃO

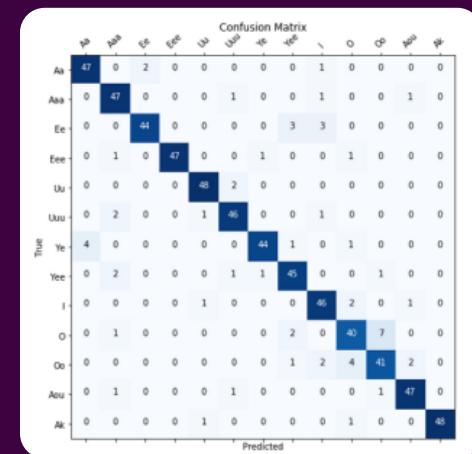
## MATRIZ DE CONFUSÃO

- Matriz onde:
  - Cada linha é o rótulo real de um dado;
  - Cada coluna é o rótulo predito pelo modelo para o dado.
- Ao realizar cada teste, é adicionado 1 na célula correspondente da matriz.

		CLASSIFICAÇÃO DO MODELO	
		REAL	
REAL	VP	70	FN
	FP	30	VN
		50	

acertos  
erros

VP - Verdadeiros Positivos  
VN - Verdadeiros Negativos  
FP - Falsos Positivos  
FN - Falsos Negativos



# AVALIAÇÃO

## MATRIZ DE CONFUSÃO - PROBLEMA BINÁRIO

- **True positive:**
  - Rótulo real é positivo
  - Modelo prediz positivo
- **True negative:**
  - Rótulo real é negativo
  - Modelo prediz negativo
- **False positive:**
  - Rótulo real é negativo
  - Modelo prediz positivo
- **False negative:**
  - Rótulo real é positivo
  - Modelo prediz negativo
- **Para pensar: lembram-se destes termos na Pandemia de COVID-19?**

# MÉTRICAS E AVALIAÇÃO

**É importante medir a confiabilidade do modelo construídos:**

- Qual é sua confiabilidade?
- É seguro usá-lo (confiabilidade é suficiente)?

**E é aí que entram os dados separados para teste!**

# MÉTRICAS

## Acurácia

- Ideal para conjuntos de dados balanceados;
- Mede a taxa de acertos dentre todos os testes realizados.

**Acerto: classe predita pelo modelo = classe real do dado (rótulo)**

## Taxa de Erro

- Medida complementar à acurácia;
- Taxa de erro =  $1 - \text{acurácia}$

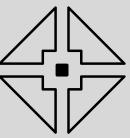
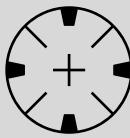
# MÉTRICAS

## Sensibilidade

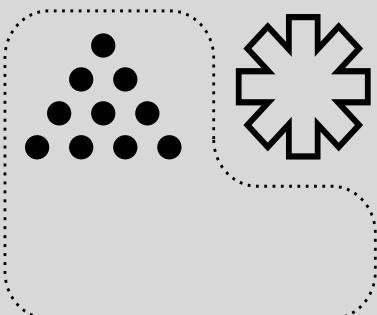
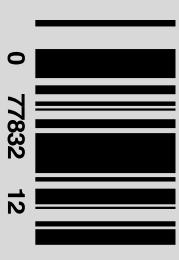
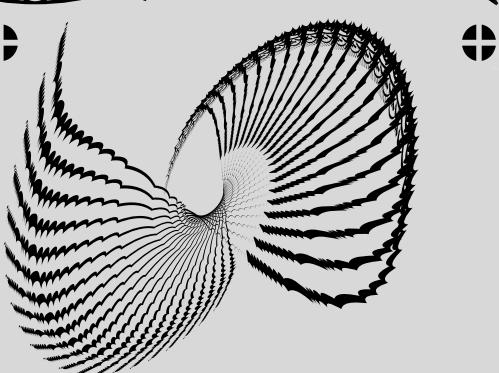
- Especialmente útil para conjuntos de dados desbalanceados
- Dentre todos os casos realmente positivos, quantas vezes o modelo aponta que é positivo?
  - Exemplo: a pessoa está doente e o modelo aponta isso

## Precisão

- Especialmente útil para conjuntos de dados desbalanceados
- Dentre tudo que o modelo aponta como positivo, quantos casos de fato são positivos?
  - Exemplo: o modelo apontou que a pessoa está doente e ela de fato está



00004512 // 99374128 // 1005934



**OBRIGADO PELA  
PARTICIPAÇÃO!**