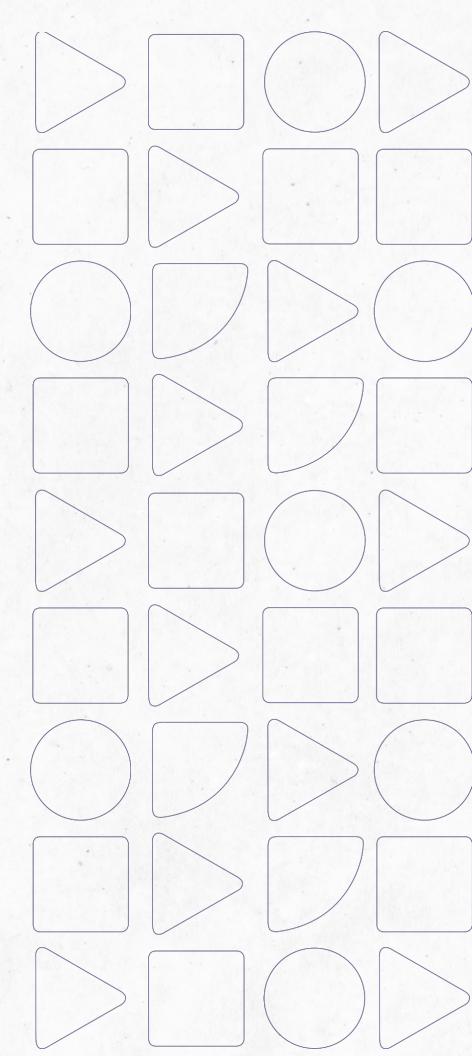


# *Machine learning* (aprendizado de máquina)

**Disciplina:** Sistemas Inteligentes



## Conteúdos:

*Machine learning* (aprendizado de máquina).

## Habilidade(s):

Construir algoritmos inteligentes.

# Bloco 1

---

Aprendizado de máquina e identificação de padrões.

# Como se fosse uma máquina, defina como você está se sentindo hoje!



Muito alegre



Sem paciência



Frustrado



Bem na paz



Estressado



Com raiva



Chateado



Rindo à toa



Com sono



Na choradeira

## Pane no sistema

O que você acha que é o aprendizado de máquina?

??



# Aprendizado de máquina

Aprendizado de máquina é um campo da inteligência artificial que se concentra na criação de programas de computadores capazes de aprender e melhorar o desempenho em uma tarefa específica, tomando como base a experiência, isto é, dados. O aprendizado de máquina é frequentemente definido em três componentes.

## Experiência (E)

Refere-se aos dados de entrada ou aos exemplos nos quais o algoritmo de aprendizado se baseia para melhorar o seu desempenho.

## Tarefa (T)

É a função ou o objetivo que o algoritmo de aprendizado está tentando realizar, como classificação, previsão, *clustering*, entre outros.

## Performance (P)

É a medida usada para avaliar quanto bem o algoritmo de aprendizado realiza a tarefa desejada. Essa medida pode variar dependendo do tipo de problema: precisão, F1-score, erro médio quadrático etc.

# E, T e P para diversos problemas

Tarefa (T)	Previsão meteorológica	Reconhecimento de fala
Experiência (E)	Valores de diversos sensores meteorológicos e estado meteorológico (por exemplo, chuva, sol, neblina etc).	Sons gravados e os respectivos textos transcritos.
Performance (P)	Quantidade de previsões corretamente identificadas.	Quantidade de transcrições corretas.

## Identificação de padrões

Um dos principais objetivos do aprendizado de máquina é identificar padrões nos dados. Isso envolve encontrar relações, correlações e tendências capazes de ajudar o algoritmo a tomar decisões ou fazer previsões.

Como exemplo, pode-se citar decidir jogar tênis com base nas condições meteorológicas. Os padrões identificados podem ser algumas regras, como “não jogar tênis em tempo ensolarado e quente” ou “jogar tênis em dias nublados”.

# Algoritmos de aprendizado de máquina

Existem vários tipos de algoritmos de aprendizado de máquina que podem ser utilizados para identificar padrões nos dados e criar modelos. Esses algoritmos podem ser categorizados em diferentes classes, como:

## Algoritmos probabilísticos

Esses algoritmos usam a teoria das probabilidades para fazer inferências a partir dos dados. Exemplos incluem Naive Bayes e redes bayesianas.

## Algoritmos matemáticos

Algoritmos baseados em técnicas matemáticas, como regressão linear, regressão logística e árvores de decisão.

## Algoritmos simbólicos

Algoritmos que representam o conhecimento em forma de símbolos e regras, como sistemas de lógica fuzzy e redes neurais artificiais.

# Modelo de aprendizado

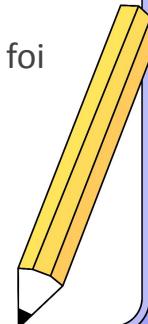
Os algoritmos de aprendizado de máquina, após serem treinados com dados de entrada e saída conhecidos, geram um modelo. Esse modelo é uma representação do conhecimento adquirido pelo algoritmo e pode ser usado para fazer previsões ou tomar decisões em novos dados não vistos durante o treinamento.

O aprendizado de máquina é uma abordagem poderosa que permite que os computadores identifiquem padrões nos dados e melhorem o seu desempenho em tarefas específicas, tomando a experiência como base.

# Pense como uma máquina!

Crie uma condição e faça a análise E, T,  
P.

Elabore uma tabela como a que foi  
mostrada durante a aula.

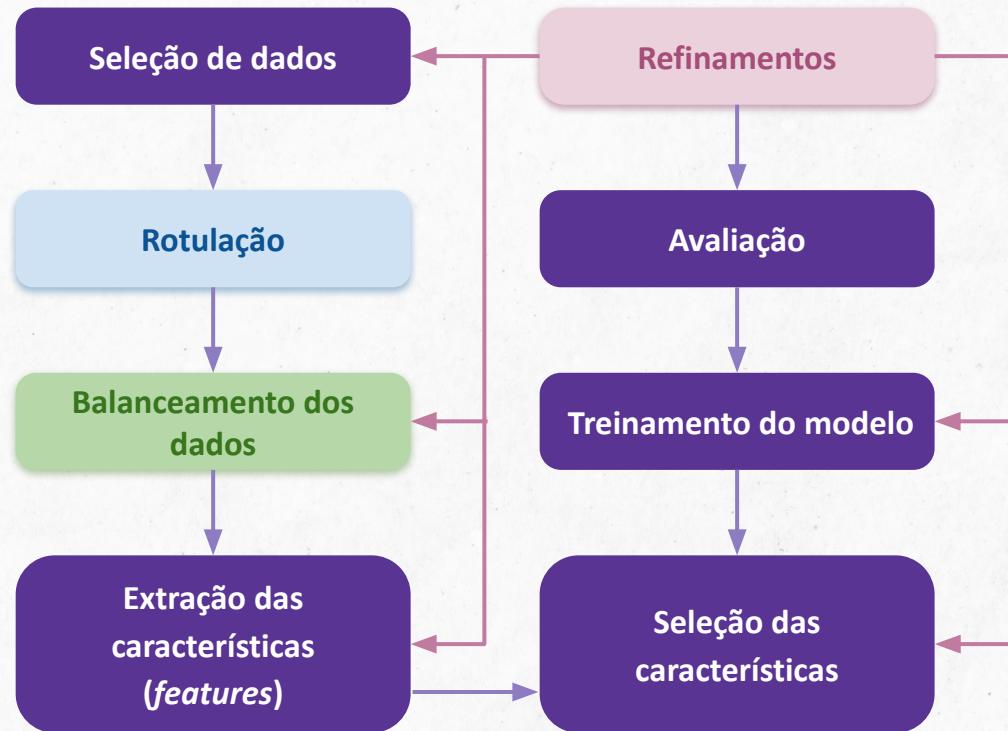


# Bloco 2

---

Etapas do aprendizado automático.

# Principais etapas do aprendizado automático



## Flashcard

Que tal criarmos *flashcards* para melhorar o entendimento?

Na frente deles, coloque a etapa. Atrás, o seu conceito ou um exemplo que ajude a entender o conteúdo.



# Desenvolvimento de uma solução

No desenvolvimento de uma solução com o uso de aprendizado automático, há etapas que são percorridas na maioria das vezes. Geralmente, refinamentos podem ser feitos entre as etapas, visando alcançar melhores resultados na avaliação.

Etapa	Conceito	Exemplo
Seleção dos dados	A seleção dos dados é crucial para o sucesso do aprendizado de máquina. Dados representativos aumentam a capacidade do modelo de generalização.	Para criar um atendente de <i>help desk</i> com aprendizado de máquina, colete dados de atendentes e usuários.
Rotulagem dos dados	A rotulagem dos dados é a atribuição das respostas esperadas (classes) para cada exemplo. A rotulagem é essencial para o treinamento supervisionado.	Em dados meteorológicos, rotule as condições do tempo (ensolarado, nublado, chuvoso).

# Desenvolvimento de uma solução

Etapa	Conceito	Exemplo
Balanceamento dos dados	Equilibrar os dados é crítico quando as classes não são distribuídas igualmente. Use <i>oversampling</i> (aumentar dados da classe minoritária) ou <i>undersampling</i> (reduzir dados da classe majoritária) para equilibrar.	Se 90% dos exemplos forem ensolarados e 0,1% chuvosos, o modelo pode ser enviesado.
Extração de <i>features</i>	A extração de atributos é essencial para representar os dados de forma significativa. Cada medida ou característica dos dados pode ser um atributo.	Em previsão meteorológica, as medições dos sensores são atributos.

# Desenvolvimento de uma solução

Etapa	Conceito	Exemplo
Seleção de <i>features</i>	Nem todos os atributos são igualmente importantes. A seleção de atributos ajuda a escolher os mais relevantes e alguns algoritmos fazem isso automaticamente.	Em detecção de <i>e-mails spam</i> , as palavras-chave podem ser selecionadas como atributos.



## Treinamento do modelo

Com base nos dados e tarefa, escolha o algoritmo de aprendizado apropriado e aplique-o aos dados de treinamento.

O modelo aprenderá os padrões associados às classes.

## Avaliação do modelo

Avaliar o desempenho do modelo é fundamental. Para isso, use métricas como precisão, cobertura, medida-F e acurácia.

A matriz de confusão é útil para avaliação.

Separe os dados em conjuntos de treinamento e teste para avaliar.

## Melhorias e ciclo de *feedback*

O aprendizado de máquina é um processo iterativo.

Identifique falhas e áreas de melhoria na avaliação e, conforme necessário, ajuste a seleção de atributos ou o algoritmo.

Continue aprimorando o modelo para obter melhores resultados.

# Bloco 3

---

Árvore de decisão.

# Aprendizado supervisionado

O aprendizado supervisionado utiliza dados para instruir algoritmos de treinamento, indicando a resposta esperada para cada exemplo.

# Árvore de decisão

## Paradigma simbólico

Árvores de decisão são um exemplo de algoritmo de aprendizado simbólico.

## Estrutura

Essas árvores têm nós internos (incluindo o nó raiz) e folhas.

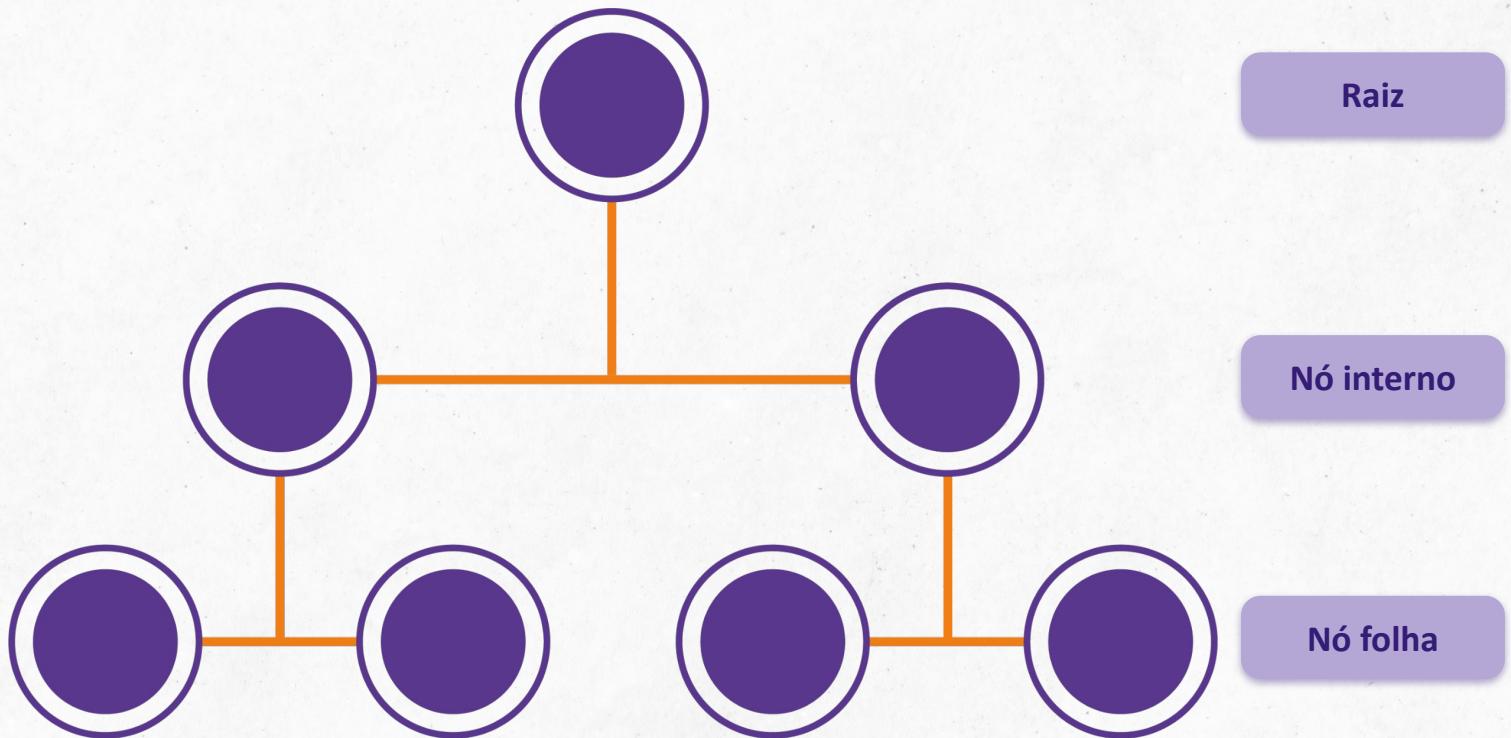
## Tomada de decisões

A cada nó interno, uma decisão é definida com base nos valores de entrada (atributos ou *features*).

## Flexibilidade

Árvores de decisão podem lidar tanto com valores numéricos quanto com valores nominais.

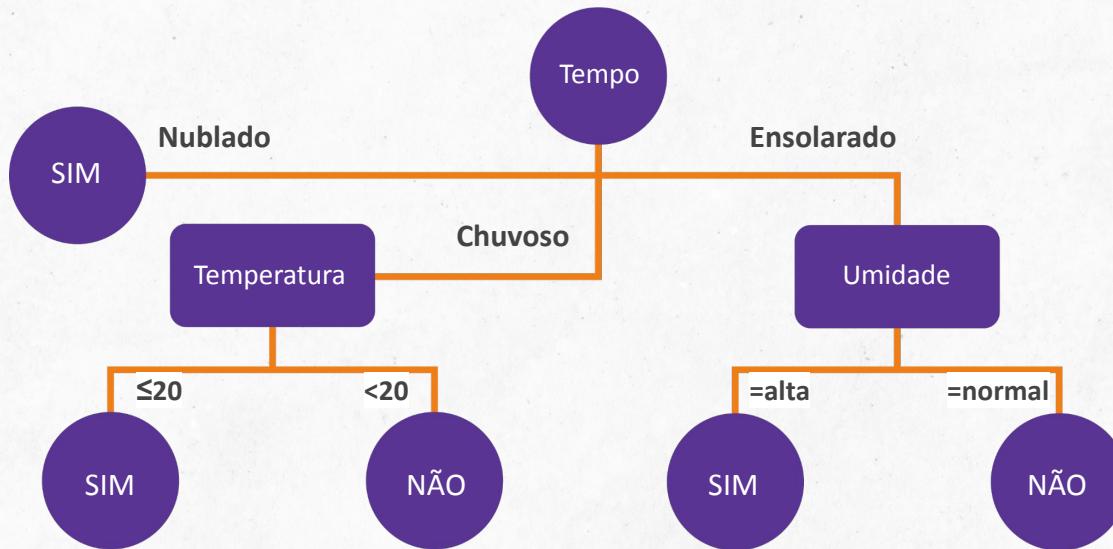
# Árvore de decisão



# Árvore de decisão

Ao considerar atributos como aparência do tempo, temperatura, umidade e vento na hora de decidir se deve-se sair ou não para jogar tênis (uma decisão de **sim** ou **não**), é possível a árvore de decisão ilustrada.

Há outras possíveis árvores de decisão? Elas são mais ou menos eficientes?



# Processo de criação de árvore de decisão



## Indução da árvore

O processo de criação de uma árvore de decisão é chamado de indução.



## Objetivo

Algoritmos de indução buscam criar árvores com a menor profundidade possível, mantendo a precisão de classificação.



## Eficiência

Isso garante que as decisões sejam tomadas de maneira eficiente, mesmo com grandes conjuntos de atributos.

## Estratégia gulosa

Uma abordagem eficiente para encontrar árvores de decisão é a estratégia gulosa. Ela prioriza o uso dos atributos mais importantes, dividindo o problema em subproblemas.

- **Aperfeiçoamento gradual:** a árvore é construída até que todos os exemplos de treinamento sejam corretamente classificados;
- **Eficiência na tomada de decisões:** isso garante que a árvore tome decisões de forma eficiente, mesmo em conjuntos de dados complexos.

# Vamos praticar?

Utilize uma folha para realizar a atividade.

## Primeiro momento

10 minutos

Pense em uma situação e crie uma árvore de decisão.

## Segundo momento

10 minutos

Compartilhe o que você elaborou com a turma!



# Bloco 4

---

*Support Vector Machines (SVM), regressão linear e  
classe contínua.*

# Support Vector Machines (SVM)

*Support Vector Machines* (SVM) é um paradigma matemático amplamente usado para criar classificadores em aprendizado supervisionado. Exploraremos três razões principais pelas quais o SVM é amplamente adotado: generalização eficaz, truque inteligente com *kernels* e sua natureza não paramétrica.

## Generalização para dados não vistos

O SVM é conhecido por sua habilidade em generalizar bem para dados não vistos, graças à definição de um hiperplano de margem máxima em um espaço multidimensional.

## Transformação com *kernels*

Embora crie uma separação linear inicial, o SVM pode usar *kernels* para transformar o problema em um espaço de dimensão superior, permitindo a separação de problemas não linearmente separáveis.

## Armazenamento de dados indicativos

O SVM não é paramétrico e armazena os dados de treinamento mais indicativos para o separador de margem máxima, conhecidos como *Support Vector Machines*.

# Lidando com valores nominais no SVM

1

## Desafio

O SVM lida, principalmente, com valores numéricos. Então, como representar valores nominais em um espaço numérico?

## Solução *one-hot-encoding*

Na prática, considere o atributo “tempo”, com valores chuvoso, nublado e ensolarado. A técnica cria três atributos binários para representar esses valores.

Chuvoso

1

Nublado

0

Ensolarado

0

Representa “Chuvoso”

0

1

0

Representa “Nublado”

0

0

1

Representa  
“Ensolarado”

# Problemas de classificação multiclasse no SVM

2

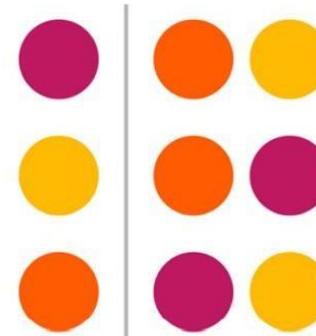
## Desafio

O SVM é, essencialmente, uma técnica de classificação binária, definindo um hiperplano de separação entre duas classes. **Como lidar quando o problema possui mais de duas classes?**

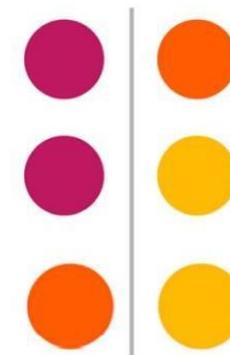
## Solução *one-hot-encoding*

Um contra todos (*one vs. all*) e um contra um (*one vs. one*).

Um contra todos



Um contra um



# Vamos praticar?

Use uma folha para realizar a atividade.

## Primeiro momento

15 minutos

Pense em uma situação e crie uma representação com *Support Vector Machines*, utilizando a solução *one-hot-encoding* e a solução “um contra todos” ou “um contra um”.

## Segundo momento

10 minutos

Mostre o que criou para a turma.



## Lidando com classes contínuas: regressão linear

Em problemas nos quais a classe é um valor contínuo, como preços de produtos, a regressão linear é uma técnica matemática eficaz. Ela busca minimizar o erro de uma equação cuja representação geométrica se ajusta a um conjunto de pontos.

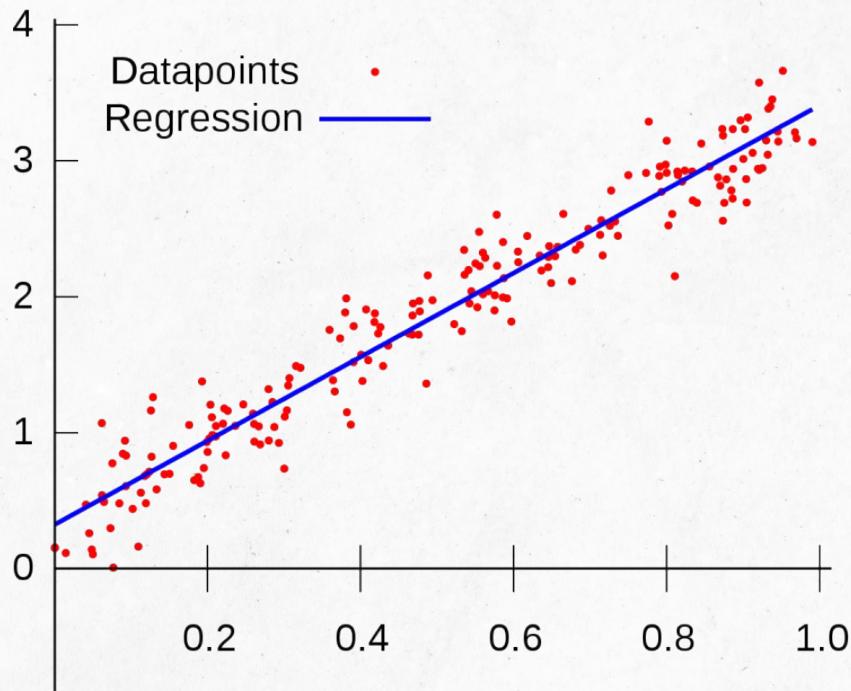
A regressão linear define uma equação que relaciona atributos de entrada X (dados de treinamento) com classes Y (valores contínuos).

$$Y = \alpha + \beta X_i$$

Quando atributos de entrada são valores nominais, eles devem ser convertidos em valores numéricos para serem usados na regressão linear.

## Lidando com classes contínuas: regressão linear

Essa equação define uma reta na qual os pontos vermelhos são as instâncias de treinamento da regressão. A reta azul (regressão) é utilizada para, dado o valor de X, definir o valor de Y (classe).



# Bloco 5

---

Aprendizagem não supervisionada.

??

## E aí?

O que você acha que deve ser feito quando não sabemos quais respostas esperar de uma situação?



## Aprendizado não supervisionado

Imagine conjuntos de dados como listas de compras em supermercados, *logs* em sistemas bancários ou informações de sensores meteorológicos. O aprendizado não supervisionado ajuda nessas situações.

No entanto, aprender algo útil sem a definição das classes do problema é o desafio do aprendizado não supervisionado.

Nesse tipo de aprendizado, destacamos duas tarefas principais: associação e agrupamento.

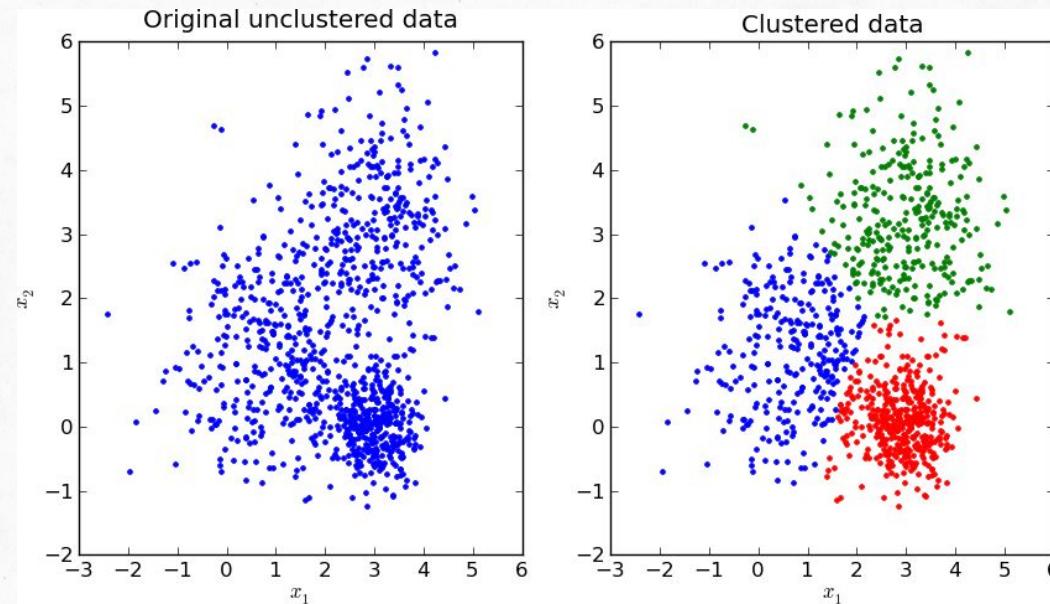
# K-means: uma abordagem de agrupamento

O **K-means** é um dos métodos mais populares no aprendizado não supervisionado, especificamente na tarefa de agrupamento. Ele busca identificar K grupos (*clusters*) nas instâncias de treinamento.

- **Processo de agrupamento:** o algoritmo escolhe centróides para cada grupo e aloca novas instâncias ao grupo mais próximo, usando medidas de similaridade;
- **Aplicações:** o K-means é usado em diversas aplicações, desde análise de mercado até segmentação de clientes e reconhecimento de padrões em dados não rotulados.

# K-means: uma abordagem de agrupamento

A figura abaixo mostra como o algoritmo K-means funciona, com K = 3 (*clusters* azul, verde e vermelho).



# Vamos praticar?

Formem grupos para realizar a atividade.

## Primeiro momento

20 minutos

Busquem uma situação na qual o K-means é utilizado. Se quiserem, podem usar o celular. A partir disso, criem um *folder* que mostre essa aplicação.

## Segundo momento

10 minutos

Compartilhem o que criaram com o resto da turma.



# Bloco 6

---

Aprendizagem semi-supervisionada.

## Aprendizado semi-supervisionado

Entre o aprendizado supervisionado e o não supervisionado, existe uma abordagem poderosa: o aprendizado semi-supervisionado.

Essa abordagem tira proveito de dados rotulados e não rotulados, geralmente disponíveis em grandes quantidades. Esse processo é conhecido como *self-training* e é um método de treinamento semi-supervisionado.

O algoritmo SVM, por exemplo, fornece uma distância ao hiperplano de separação entre as classes. Quanto maior essa distância, maior a confiabilidade da classificação.



# Self-training



## Geração de novos rótulos

Esse modelo inicial é aplicado aos dados não rotulados para gerar novos rótulos de forma automática.



## Cuidado com erros

É importante notar que esses novos rótulos podem conter erros, pois são gerados automaticamente.



## Seleção de rótulos confiáveis

A escolha de um método que indique a confiabilidade das classificações é fundamental para selecionar apenas rótulos confiáveis.

**Imagine que estamos construindo um sistema de detecção de *spam* em e-mails.**

# *Self-training*

1

**Dados rotulados:** inicialmente, temos um conjunto de *e-mails* já rotulados como "spam" ou "não spam". Esses são nossos dados rotulados.

2

**Modelo inicial:** usamos esses dados rotulados para treinar um modelo inicial de detecção de *spam*, como um classificador de Naive Bayes.

3

**Classificação inicial:** agora, com o nosso modelo treinado, aplicamos o classificador a um grande conjunto de *e-mails* não rotulados.

4

**Geração de novos rótulos:** com base nas classificações do modelo, os *e-mails* são automaticamente rotulados como "spam" ou "não spam".

5

**Avaliação de confiabilidade:** para garantir a qualidade dos novos rótulos, usamos métricas, avaliando a confiabilidade das classificações do modelo.

6

**Seleção de rótulos confiáveis:** selecionamos apenas os rótulos gerados com alta confiabilidade para serem adicionados aos nossos dados rotulados originais.

# **Self-training**

7

**Treinamento aprimorado:** agora, treinamos um novo modelo de detecção de *spam* com mais dados rotulados, incluindo os novos rótulos gerados automaticamente.

8

**Melhoria contínua:** repetimos esse processo iterativamente, gerando novos rótulos confiáveis e melhorando o nosso modelo com cada iteração.

9

**Resultado:** ao final, nosso sistema de detecção de *spam* se torna mais preciso e eficaz, mesmo sem a necessidade de rotular manualmente todos os *e-mails*.

# Vamos praticar?

Use uma folha para realizar a atividade.

## Atividade

20 minutos

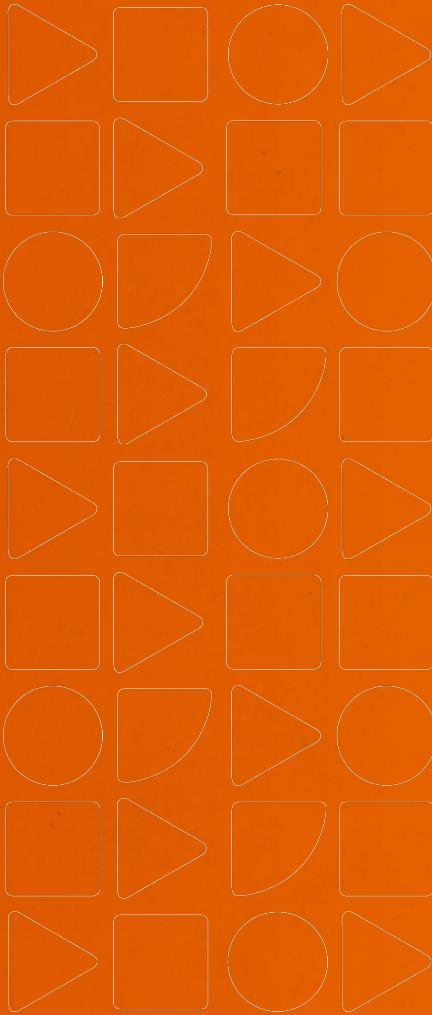
Pense em uma situação e crie um passo a passo, como o que foi visto anteriormente, aplicando o *self-training*.



## Fechamento

Diga uma palavra que resuma o que você achou da aula.





# Referências Bibliográficas

PROZ EDUCAÇÃO. *Apostila de Sistemas Inteligentes*. 2023.