

APRENDIZAGEM DE MÁQUINA

PROF. JOSENALDE OLIVEIRA

josenalde.oliveira@ufrn.br

<https://github.com/josenalde/machinelearning>

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - UFRN

OBJETIVOS

O(a) discente conseguirá selecionar tipo de modelo (supervisionado, não supervisionado, semi supervisionado) e algoritmo aplicável a determinado problema de negócio, com as etapas de processo de aprendizagem de máquina, avaliando o modelo e colocando em produção

CONTEÚDO – 60H (4 CR)

Preparação dos dados; Modelos preditivos; Modelos descritivos; Classificação; Regressão; Agrupamento Seleção/Extração de características (feature selection); Avaliação de modelos; Redução de dimensionalidade; Processo de deploy de modelos em produção

Embora [Fundamentos e Técnicas de Ciências de Dados](#) não seja pré-requisito oficial, este componente optativo segue o fluxo de Extração, Transformação (pré processamento) e Carga em modelos de ML. Usaremos Python. Conhecimento nas libs [NumPy](#) e [Pandas](#) é recomendado.

PLANO DE CURSO

- Referências

- [1] Referência da biblioteca scikit-learn: <https://scikit-learn.org/stable/index.html>
- [2] GÉRON, A. **Mãos à Obra**: aprendizado de máquina com Scikit-Learn, Keras & TensorFlow, 2. ed. O'Reilly (Alta Books), 2021.
- [3] HUYEN, C. **Projetando Sistemas de Machine Learning**: processo iterativo para aplicações prontas para produção. O'Reilly (Alta Books), 2023.
- [4] GRUS, J. **Data Science do Zero**: primeiras regras com o python, 2. ed. O'Reilly (Alta Books), 2021.
- [5] BROWNLEE, J. <https://machinelearningmastery.com/>
- [6] GitHub Profa. Laura Emmanuella (UFRN/UFRJ) de ML:
<https://github.com/lauraemmanuella/AprendizadoMaquina>
- [7] NUNES, M. <https://introbigdata.org/> (Material Prof. Marcus Nunes – Dep. Estatística UFRN)

PLANO DE CURSO

- Software usado nas aulas
 - Jupyter Lab + Python
 - Google Colab
 - Extensão Jupyter (Microsoft) para VS Code
- Fontes de datasets / imagesets
 - Open data
 - UC Irvine Machine Learning Repository (<http://archive.ics.uci.edu/ml/>)
 - Kaggle (<https://www.kaggle.com/datasets>)
 - AWS (<https://registry.opendata.aws/>)
 - Deeplake (<https://datasets.activeloop.ai/docs/ml/datasets/mnist/>)
 - Metaportais
 - Data Portals (<http://dataportals.org>)
 - OpenDataMonitor (<http://opendatamonitor.eu>)
 - Quandl (<http://quandl.com/>) NASDAK
 - Wikipedia (<https://homl.info/9>)
 - Quora (<https://homl.info/10>)
 - <https://dados.gov.br/home>



PLANO DE CURSO

- Avaliações (Semestre 2024.2: 16.set.24 – 01.fev.25)
 - *Feriados nas aulas de AM de 03.out (estadual) e 20.nov (nacional)*
 - Unidade I: **24.out.2024**
 - Unidade II: **05.dez.2024** (*Suspensão das atividades acadêmicas de 23/12-03/01*)
 - Retorno às aulas de AM em 2025 dia **08.jan**
 - Unidade III: **23.jan.2025** (projeto modelo/deploy/UI)
 - Prova reposição: **29.jan.2025** (FIM)

UM POUCO SOBRE APRENDIZAGEM DE MÁQUINA

Machine Learning - ML



Ser humano estabelece **conexões** para lidar com coisas novas

Similaridade pode ser óbvio para o humano, mas não para computadores

Máquinas operam sobre **tarefas frequentes**, com alto volume e velocidade

Desafio: máquinas serem ensinadas e depois aprenderem 'sozinhas'

UM POUCO SOBRE APRENDIZAGEM DE MÁQUINA

Machine Learning - ML

Como uma criança aprende que ambos são dinossauros?



E aqui?



APRENDIZAGEM DE MÁQUINA

Agentes, robôs, jogos,
raciocínio baseado em
memória, algoritmos
genéticos, ...

Classificação, agrupamento,
regressão, associação ...

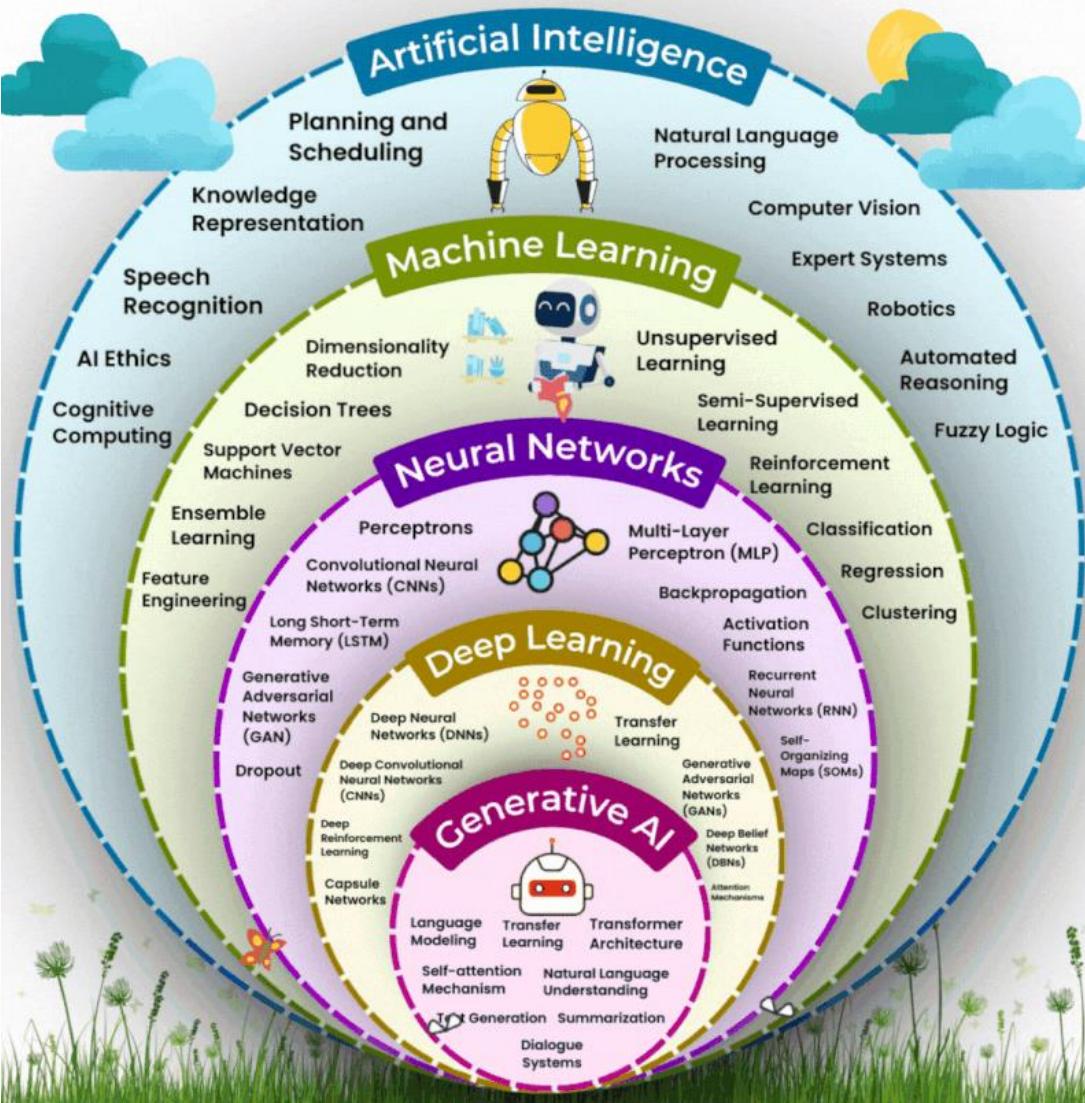
Redes neurais multilayer,
convolucionais, recorrentes,
long short-term memory
(LSTM), redes de hopfield ...



Aplicar métodos
científicos, processar
e armazenar grandes
volumes de dados, ...

<https://www.serpro.gov.br/menu/noticias/noticias-2019/democratizando-a-inteligencia-artificial>

The AI Universe



VÁRIAS VISÕES SOBRE “APRENDER”

aprender

a·pren·der

vtd e vint

1 Ficar sabendo, reter na memória, tomar conhecimento de: “*Então o Botelho [...] não perdoava a ninguém, amaldiçoando todo aquele [...] que gozava o que ele não desfrutara; que sabia o que ele não aprendera*” (AA1). “– Pede então à Rita que to ensine... [...] não terá muito que aprender!” (AA1).

vtd e vint

2 Adquirir habilidade prática (em): “*Eu, quando comecei a tocar violão, não queria aprender música...*” (LB2). Os chimpanzés aprendem com rapidez espantosa.

vtd, vti e vtdi

3 Passar a compreender (algo) melhor graças a um depuramento da capacidade de apreciação, empatia, percepção etc.: “– *Nunca vi uma criatura para aprender as coisas com tanta facilidade!*” (AA2). Aprendeu muito com os próprios erros. Aprenderam dos amigos a ausência nas horas difíceis.

ETIMOLOGIA

lat apprehendēre, como esp.

Fonte: Michaelis

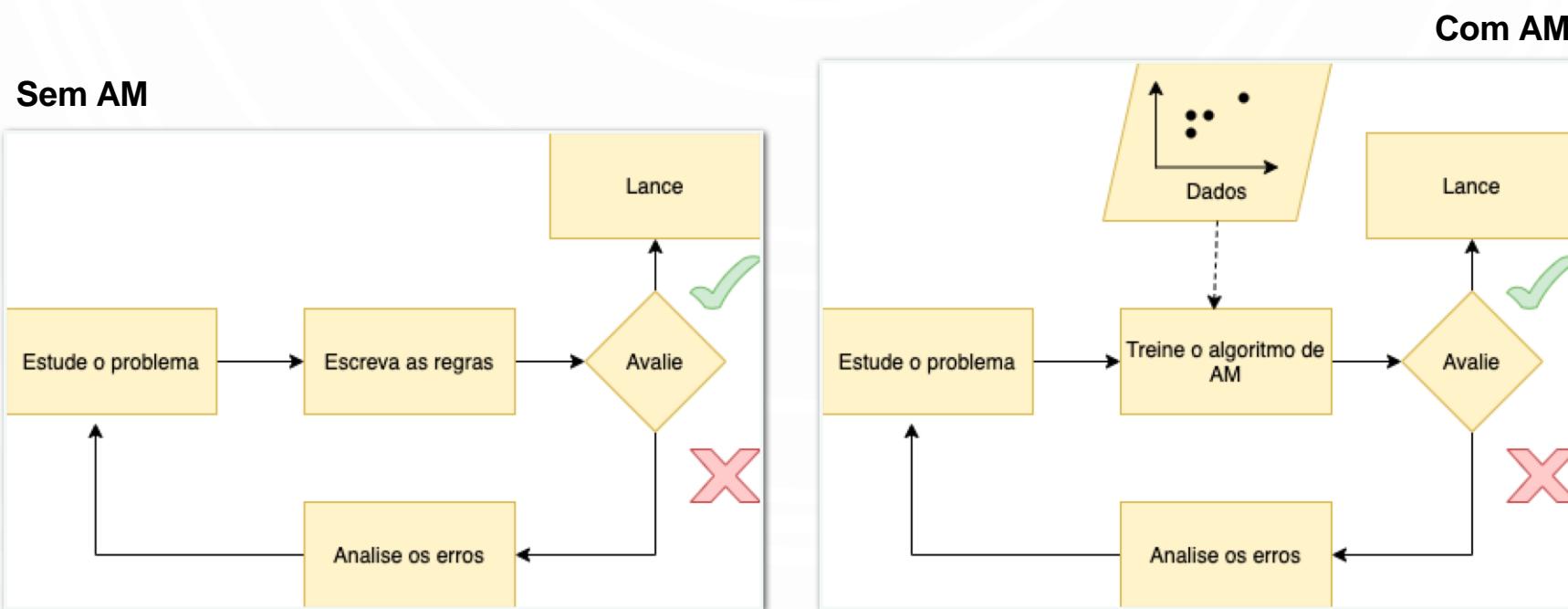
“AM é o campo de estudo que possibilita aos computadores a habilidade de aprender sem explicitamente programá-los”
[Arthur Samuel, 1959] – Ver história do OCR
*aplicação de AM disseminada na década 90: filtro spam

“O AM é uma abordagem para aprender padrões complexos (supondo que existem) a partir de dados existentes e usar esses padrões para fazer previsões sobre dados desconhecidos”
[Chip Huyen, 2023]

“Um programa de computador aprende pela experiência E em relação a alguma tarefa T e alguma medida de desempenho P SE o seu desempenho em T, conforme medido por P, melhora com a experiência E”
[Tom Mitchell, 1997]

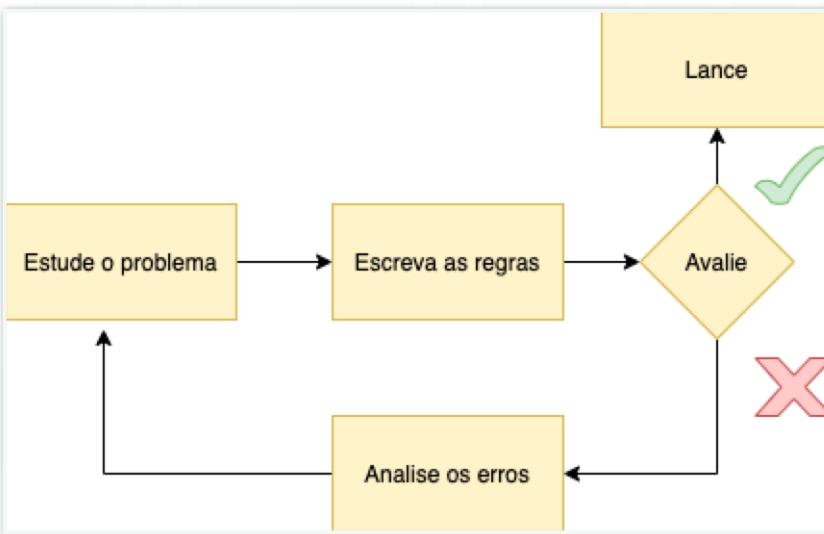
APRENDIZAGEM DE MÁQUINA

Área da Inteligência Artificial que investiga o **desenvolvimento de algoritmos que são capazes de aprender a partir dos dados**, adquirindo conhecimento de forma automática



APRENDIZAGEM DE MÁQUINA

Sem AM



Abordagem tradicional

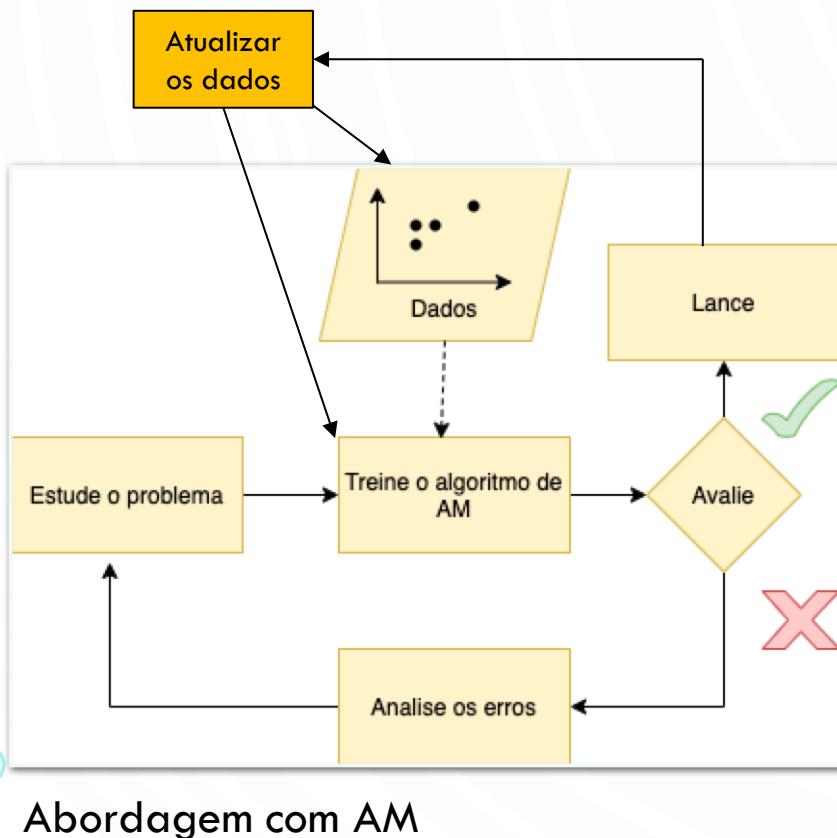
Seja a tarefa T de sinalizar e-mails novos como SPAM. Os usuários podem marcar e-mails como SPAM. Estes exemplos (amostras) serão o conjunto de treinamento. A experiência E é o dado de treinamento e a medida P poderia ser a proporção de e-mails sinalizados corretamente (acurácia)

Passos (abordagem tradicional):

1. Identificar palavras ou frases no campo ASSUNTO (“para você”, “cartão de crédito”, “de graça”, “promoção”, “imperdível”, etc.), nome do remetente no corpo do e-mail etc.
2. Escrever algoritmo para detectar estes padrões observados e sinalizar esses e-mails como SPAM
3. Testar e repetir passos 1 e 2 até julgar que está OK para produção

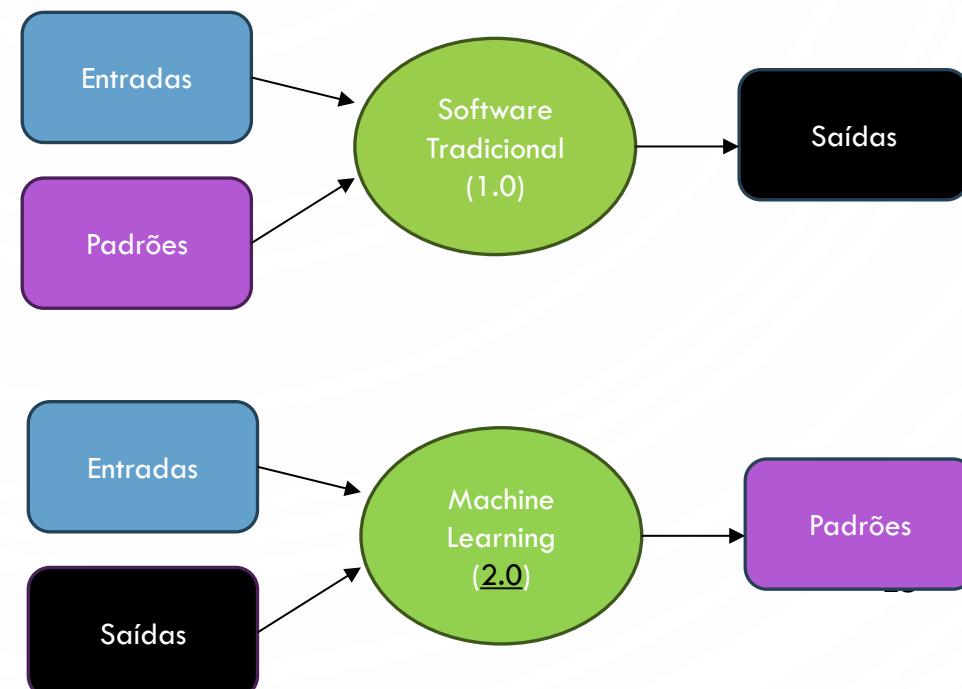
PROBLEMA: lista extensa; E se surgem novas palavras?

APRENDIZAGEM DE MÁQUINA



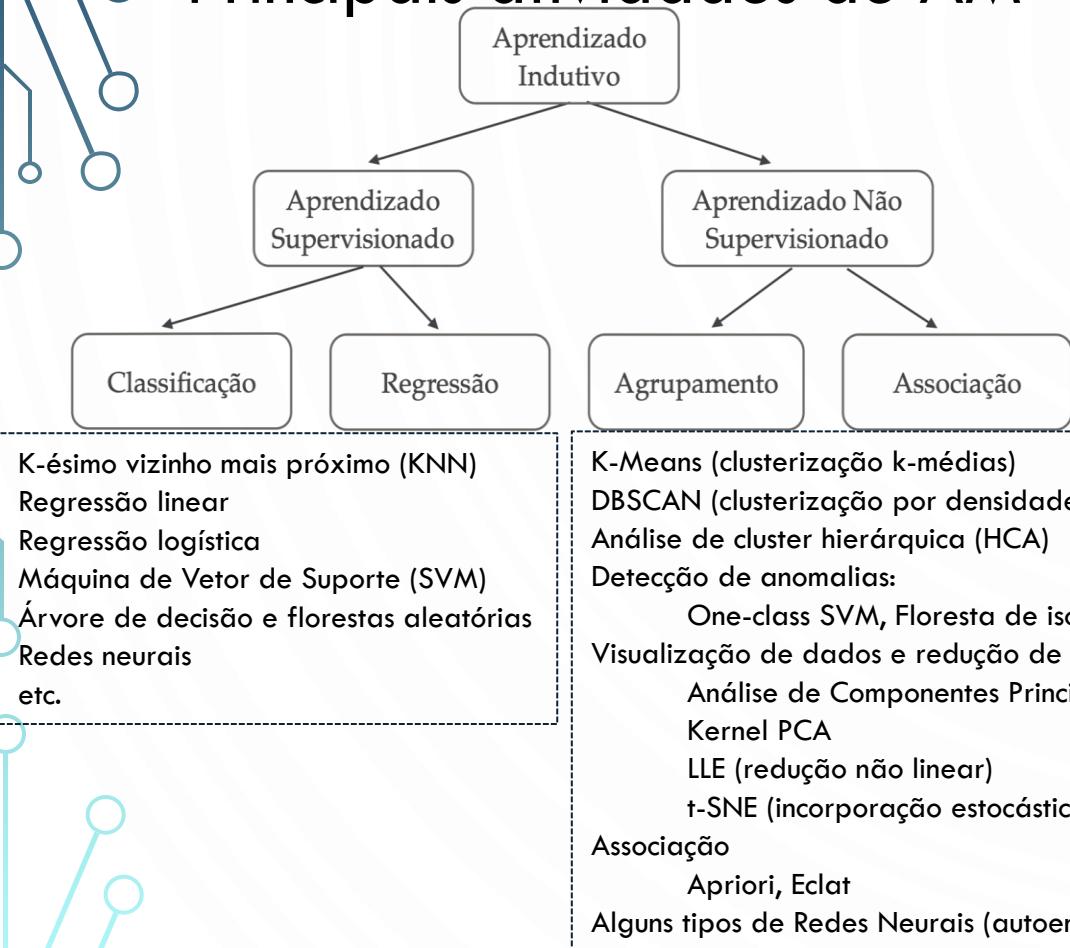
Abordagem com AM

Um filtro de spam baseado em AM aprende automaticamente quais palavras e frases são bons indicadores de spam, ao detectar os padrões de palavras inusitadamente frequentes em exemplos de SPAM quando comparado aos exemplos corriqueiros (HAM). Se surge novas palavras em e-mails marcados como SPAM pelos usuários, começa a marca-los sem a intervenção do desenvolvedor



APRENDIZAGEM DE MÁQUINA

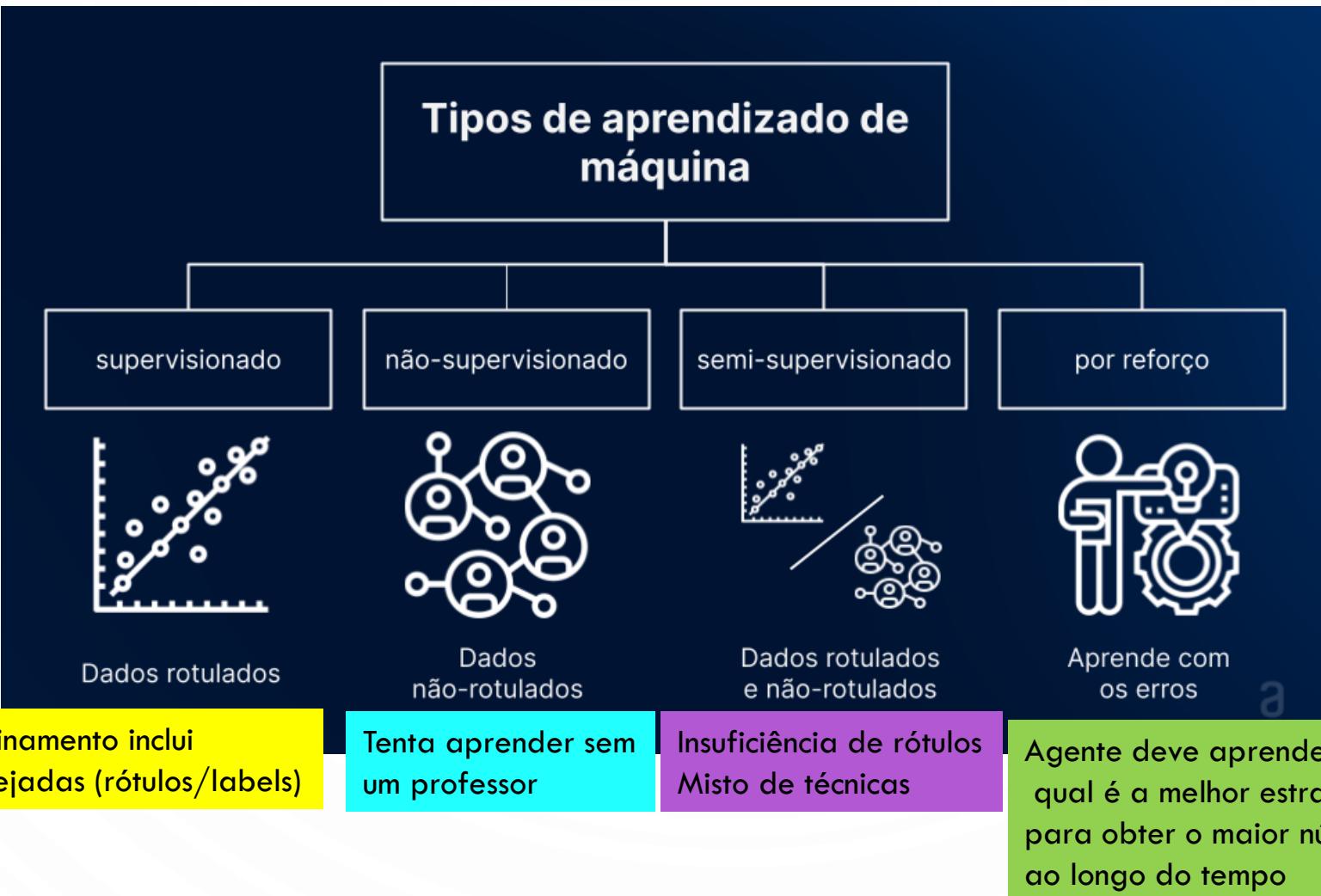
Principais atividades de AM



O aprendizado indutivo é aquele que faz a pessoa por si só descobrir as sementes do saber. Inicia-se com um desafio, uma pergunta inteligente, um problema real para resolver, um projeto para desenvolver, um estudo de caso para analisar, a observação de um fenômeno ou o resultado experimental de laboratório.

- Coleta de informações a partir da observação rigorosa da natureza;
- Reunião, organização sistemática e racional dos dados recolhidos;
- Formulação de hipóteses segundo a análise dos dados recolhidos;
- Comprovação das hipóteses a partir de experimentações.

AMPLIANDO OS TIPOS DE AM:



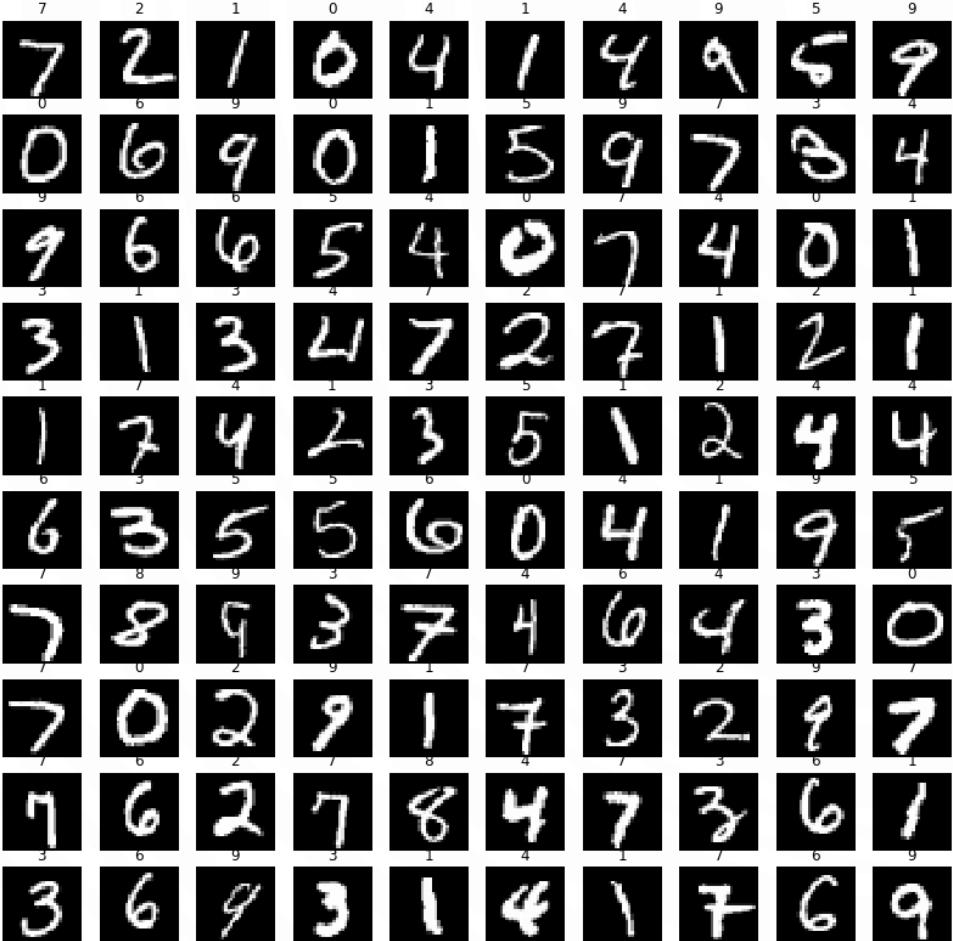
<https://deepmind.google/technologies/alphago/>

DATASETS

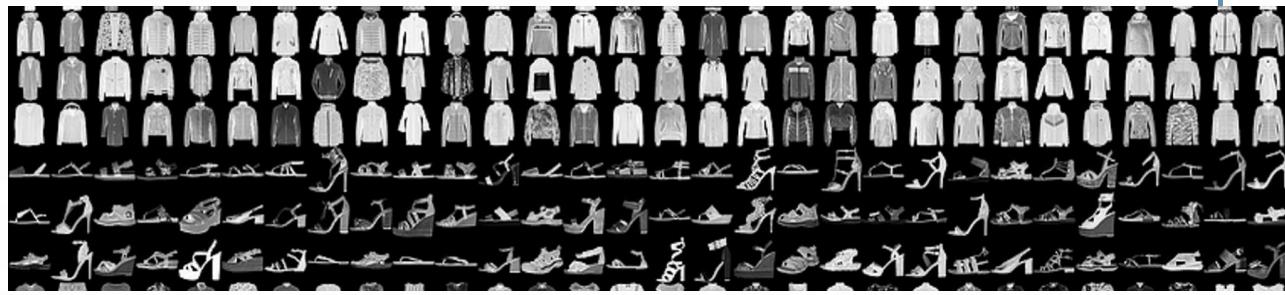
- Bases de dados para algoritmos de aprendizado de máquina são formadas por amostras do domínio/negócio que se deseja aprender
- Variáveis numéricas, categóricas (uni, bi, poli)
- Variáveis categóricas normalmente codificadas para AM (**encoder**)

Id.	Nome	Idade	Sexo	Peso	Manchas	Temp.	# Int.	Est.	Diagnóstico
4201	João	28	M	79	Concentradas	38,0	2	SP	Doente
3217	Maria	18	F	67	Inexistentes	39,5	4	MG	Doente
4039	Luiz	49	M	92	Espalhadas	38,0	2	RS	Saudável
1920	José	18	M	43	Inexistentes	38,5	8	MG	Doente
4340	Cláudia	21	F	52	Uniformes	37,6	1	PE	Saudável
2301	Ana	22	F	72	Inexistentes	38,0	3	RJ	Doente
1322	Marta	19	F	87	Espalhadas	39,0	6	AM	Doente
3027	Paulo	34	M	67	Uniformes	38,4	2	GO	Saudável

IMAGESET (EXEMPLOS CLÁSSICOS)



MNIST



FASHION MNIST



<https://www.kaggle.com/datasets/paramagarwal/fashion-product-images-dataset>

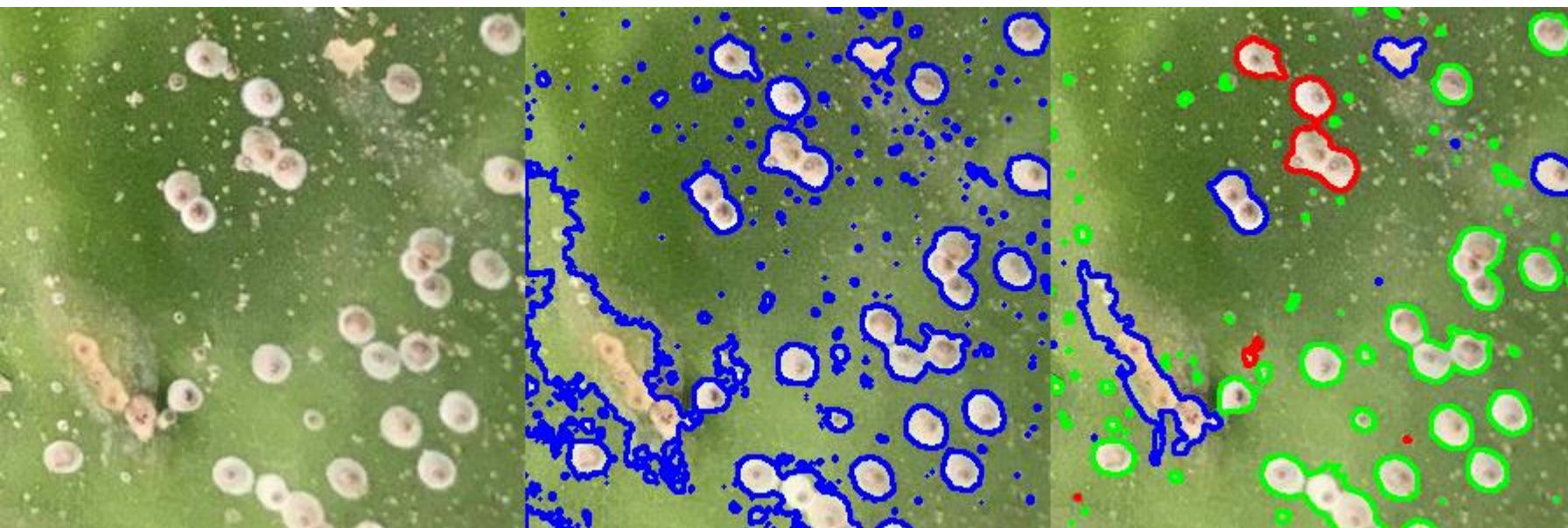
DATASETS

- As amostras são comumente chamadas de exemplos ou observações
- Em geral, eles são representados por um vetor de **características*** que os descreve, também denominados de **atributos, campos, features** (ou variáveis de decisão)
 - Cada amostra corresponde a uma ocorrência/observação/registro na base de dados
 - Cada atributo está associado a uma propriedade da amostra
 - **As features podem ser criadas a partir dos dados**
 - **Quais atributos são relevantes/significativos? (significância estatística, multicolinearidade, podem ser agrupados/reduzidos? (PCA, regularização)?**
 - **Feature selection X Feature extraction**

Em aprendizagem de máquina um atributo é um tipo de dado (por exemplo, quilometragem), ao passo que Característica assume vários significados; dependendo do contexto geralmente significa o atributo mais o seu valor (por exemplo: quilometragem=1500).

EXEMPLO DE CRIAÇÃO DE FEATURES

- Dada uma imagem com inseto cochonilha de escama de diferentes estágios de desenvolvimento, agrupar similares com base em características geométricas/morfológicas e de cor – aprendizado não supervisionado (agrupamento/clustering)



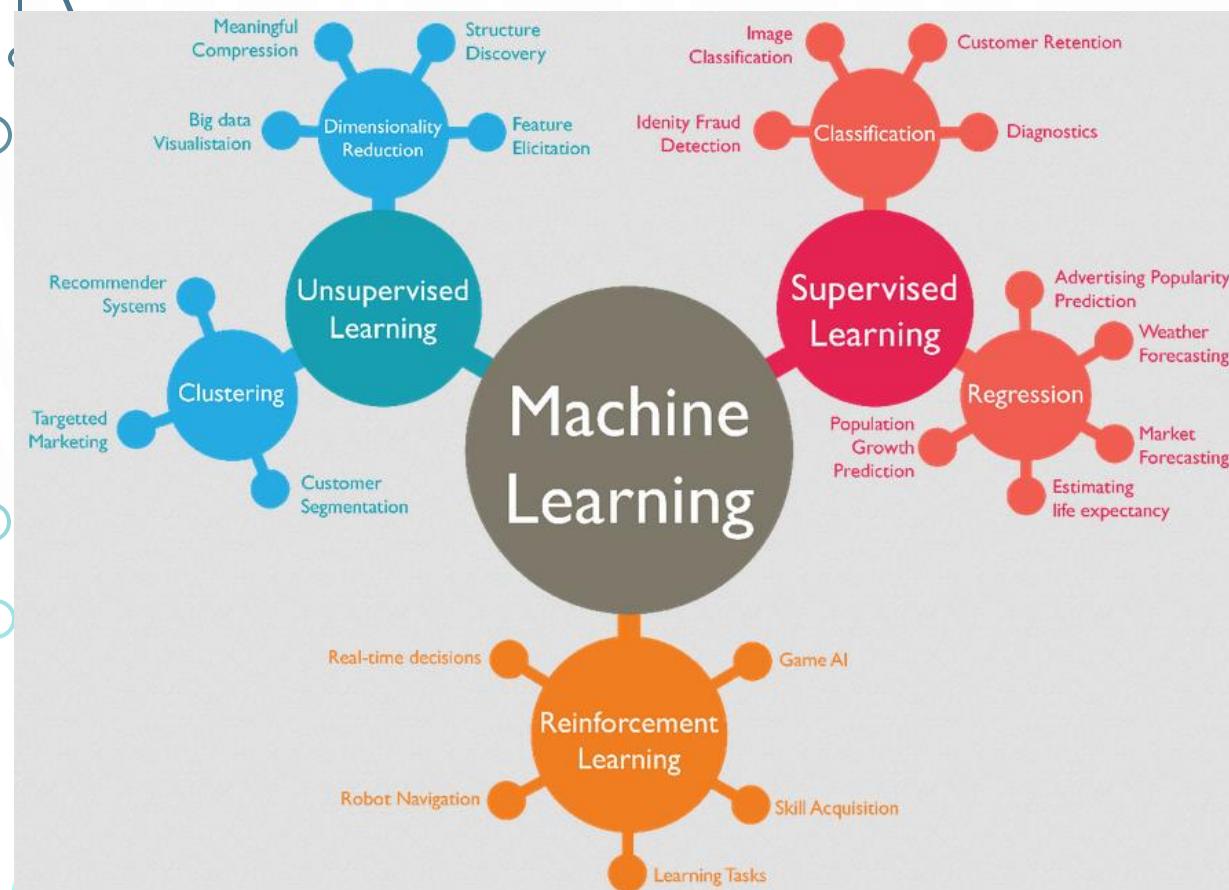
https://github.com/Gus-1003/Projeto_Pesquisa_PalmaS/blob/main/Projeto_Segmentado.ipynb

EXEMPLO DE CRIAÇÃO DE FEATURES A PARTIR DOS CONTORNOS DETECTADOS:

```
areas = [] # Armazena a área de cada objeto
perimeters = [] # Armazena o perímetro de cada objeto
centroids_x = [] # Armazena as coordenadas x do centróide de cada objeto
centroids_y = [] # Armazena as coordenadas y do centróide de cada objeto
aspect_ratio = [] # Armazena a relação entre Largura e altura do retângulo
# de contorno de cada objeto
extent = [] # Armazena a razão entre a área do contorno e a área do
# retângulo de contorno de cada objeto
solidity = [] # Armazena a razão entre a área do contorno e a área do
# casco convexo de cada objeto
equivalent_diameter = [] # Armazena o diâmetro equivalente (em pixels) do
# círculo com a mesma área de cada objeto
# r_mean, r_min, r_max e r_std: armazenarão a média, o mínimo, o máximo e
# o desvio padrão dos valores do canal vermelho (R) de cada objeto
# encontrado na imagem.
r_mean = [] r_min = [] r_max = [] r_std = []
# g_mean, g_min, g_max e g_std: armazenarão a média, o mínimo, o máximo e
# o desvio padrão dos valores do canal verde (G) de cada objeto encontrado
# na imagem. g_mean = [] g_min = [] g_max = [] g_std = []
# b_mean, b_min, b_max e b_std: armazenarão a média, o mínimo, o máximo e
# o desvio padrão dos valores do canal azul (B) de cada objeto encontrado na
# imagem.
b_mean = [] b_min = [] b_max = [] b_std = []
width = []
# Armazena a Largura do retângulo de contorno de cada objeto
height = []
# Armazena a altura do retângulo de contorno de cada objeto
angle = [] # Armazena o ângulo (em graus) do retângulo de contorno de cada
# objeto
radius = [] # Armazena o raio do círculo de contorno de cada objeto
```

APRENDIZAGEM DE MÁQUINA

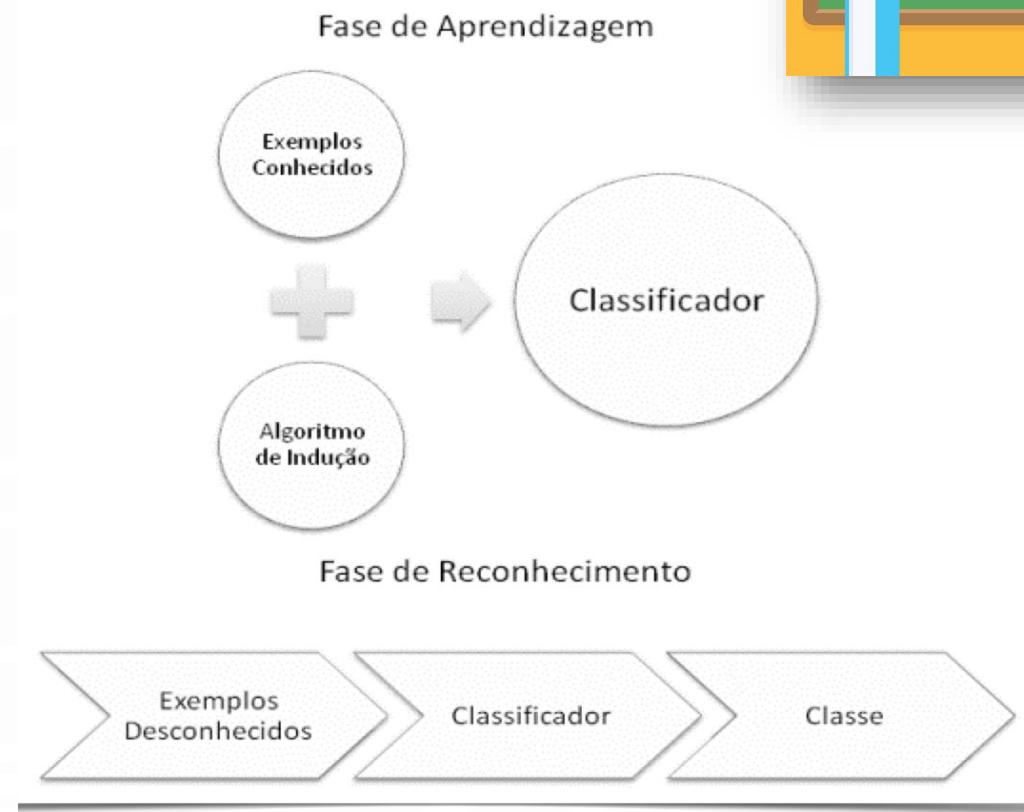
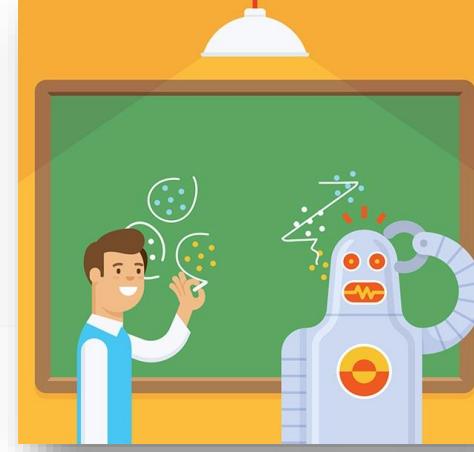
- Principais atividades de AM (alguns exemplos de aplicação e técnicas)



- Análise de imagens de produtos em linha de produção para classificação automática (redes neurais convolucionais – CNN)
- Detecção de tumores a partir de exames de imagens (segmentação semântica/CNN – localização exata e forma dos tumores)
- Classificação de artigos de notícias (PLN com redes neurais recorrentes, (RNNs), CNNs ou transformadores (transformers))
- Sinalização de comentários ofensivos em fóruns/social media
- Resumo automático de textos (PLN)
- Criação de chatbot
- Previsão do faturamento empresarial no próximo ano
- Responder a comandos de voz (RNN, CNN, transformer)
- Detecção de fraudes com cartão de crédito (detectar anomalia)
- Segmentação de clientes, público (clusterização)
- Representação de dados complexos de alta dimensão em diagrama claro e criterioso (PCA etc.)
- Recomendação de produtos, com base em compras anteriores (rede neural treinada com dados anteriores de todos os clientes)
- Criar bot inteligente para jogo (aprendizagem por reforço)

CLASSIFICAÇÃO DE PADRÕES

- Classificação de padrões é o **processo de atribuição de rótulos discretos, também chamados de classes, a amostras de um domínio**
- A classificação de padrões dá-se em duas fases distintas: fase de aprendizagem (**treino/fit**) e fase de reconhecimento (**teste/predict**)



CLASSIFICAÇÃO DE PADRÕES



Orquídea



Orquídea



Orquídea



Orquídea



Orquídea



Orquídea



Flor do Deserto



Flor do Deserto



Flor do Deserto



Flor do Deserto



Flor do Deserto



Flor do Deserto

CLASSIFICAÇÃO DE PADRÕES

Orquídea ou Flor do deserto?



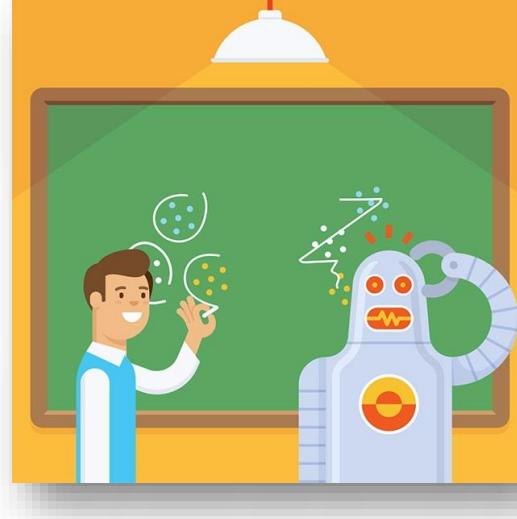
CLASSIFICAÇÃO DE PADRÕES

Orquídea ou Flor do deserto?

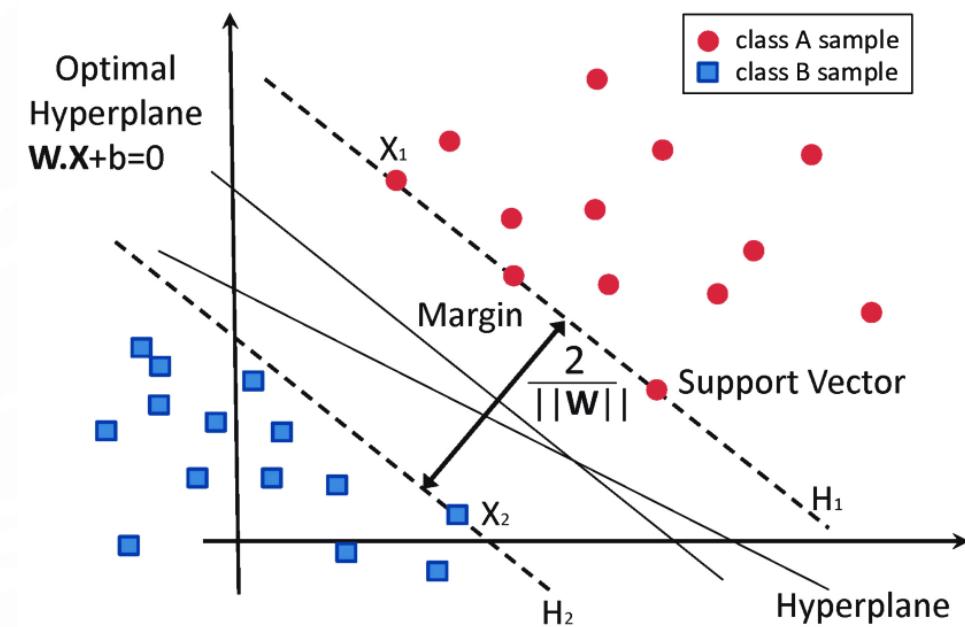
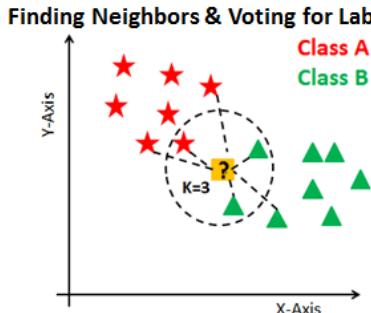
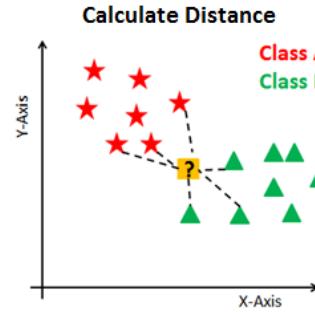
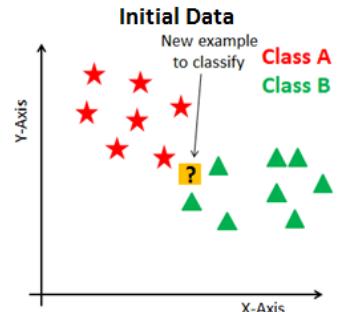
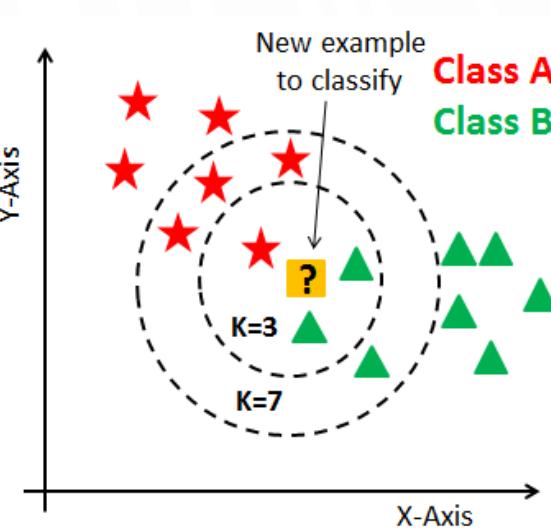


ASPECTOS DE MACHINE LEARNING

➤ Classificação: supervisionado – conjuntos de treino / teste, métricas de avaliação mais bem definidas, por comparar com o *ground truth*, matriz de confusão etc.



Métodos mais comuns: KNN, SVM, Árvore de Decisão, Redes Neurais



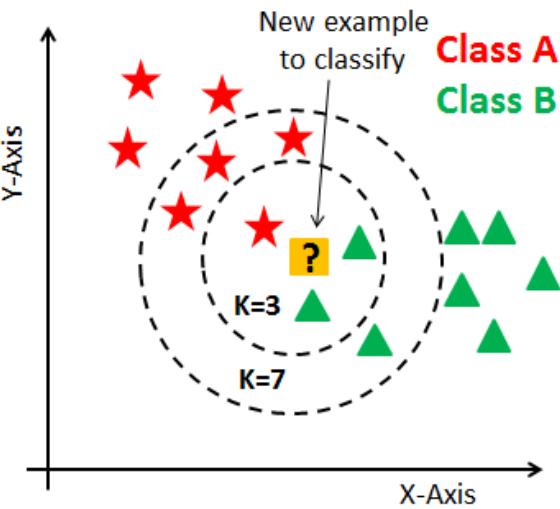
https://github.com/josenalde/machinelearning/blob/main/src/nb_knn1.ipynb

https://github.com/josenalde/machinelearning/blob/main/src/nb_svm1.ipynb

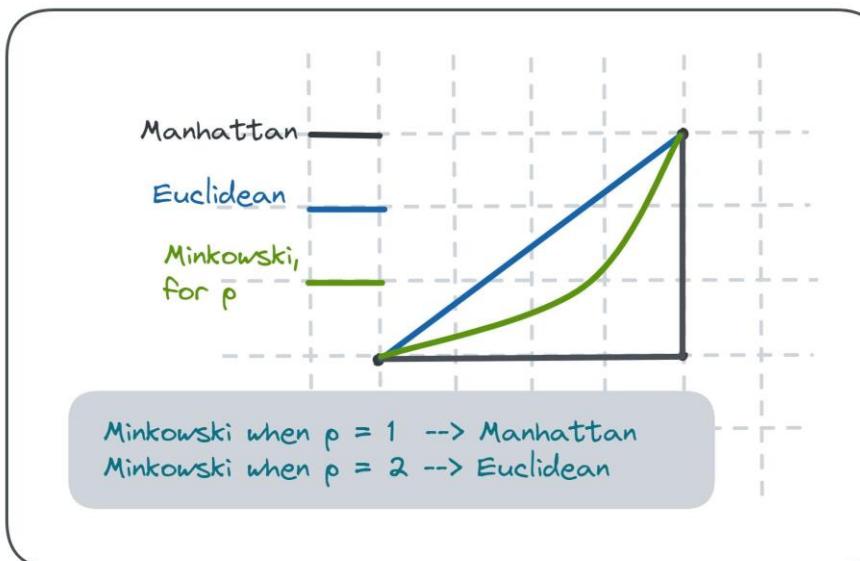
ASPECTOS DE MACHINE LEARNING

➤ Aprendizado baseado em instâncias

- O sistema aprende os exemplos por meio da memorização e depois generaliza em novos casos, ao empregar uma medida de similaridade a fim de compará-los a outros exemplos aprendidos (ou um conjunto deles)



- O K-vizinhos mais próximos (KNN) é um dos modelos preditivos mais simples que existe, pois não possui premissas matemáticas e não é custoso computacionalmente. Requer apenas uma noção de distância e uma premissa de que pontos que estão perto uns dos outros são similares!



$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p} \text{ for } p \geq 1$$

```
from scipy.spatial import distance
```

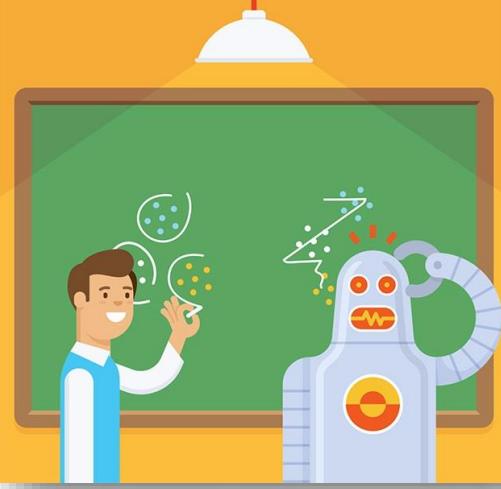
```
x = [3,6,9]
```

```
y = [1,0,1]
```

```
print(distance.euclidean(x,y))
```

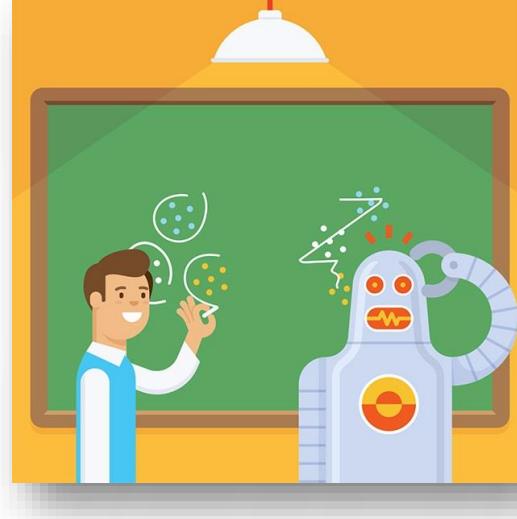
```
print(distance.cityblock(x,y))
```

```
print(distance.minkowski(x,y,p=3))
```

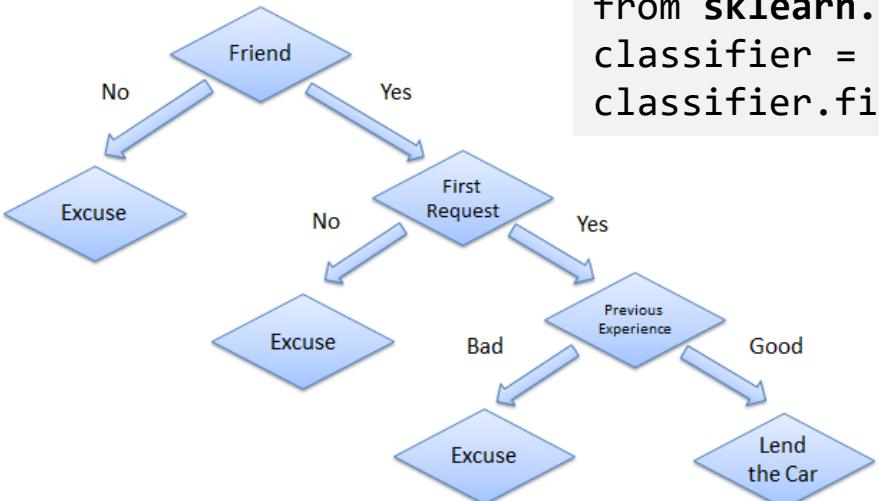


ASPECTOS DE MACHINE LEARNING

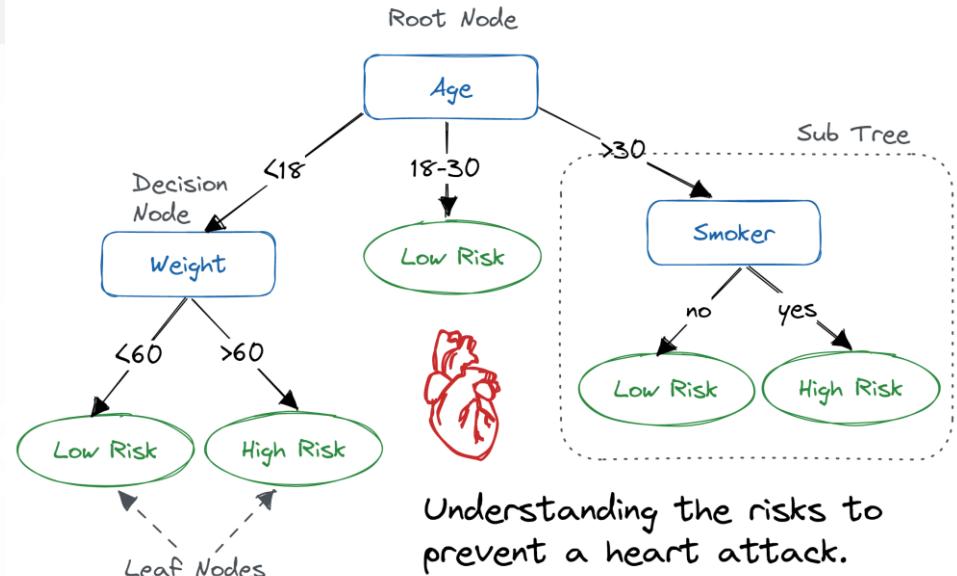
➤ **Classificação: supervisionado – conjuntos de treino / teste, métricas de avaliação mais bem definidas, por comparar com o ground truth, matriz de confusão etc.**



Métodos mais comuns: KNN, SVM, Árvore de Decisão, Redes Neurais



```
from sklearn.tree import DecisionTreeClassifier  
classifier = DecisionTreeClassifier()  
classifier.fit(X_train, y_train)
```



<https://www.datacamp.com/community/tutorials/decision-tree-classification-python>

REGRESSÃO

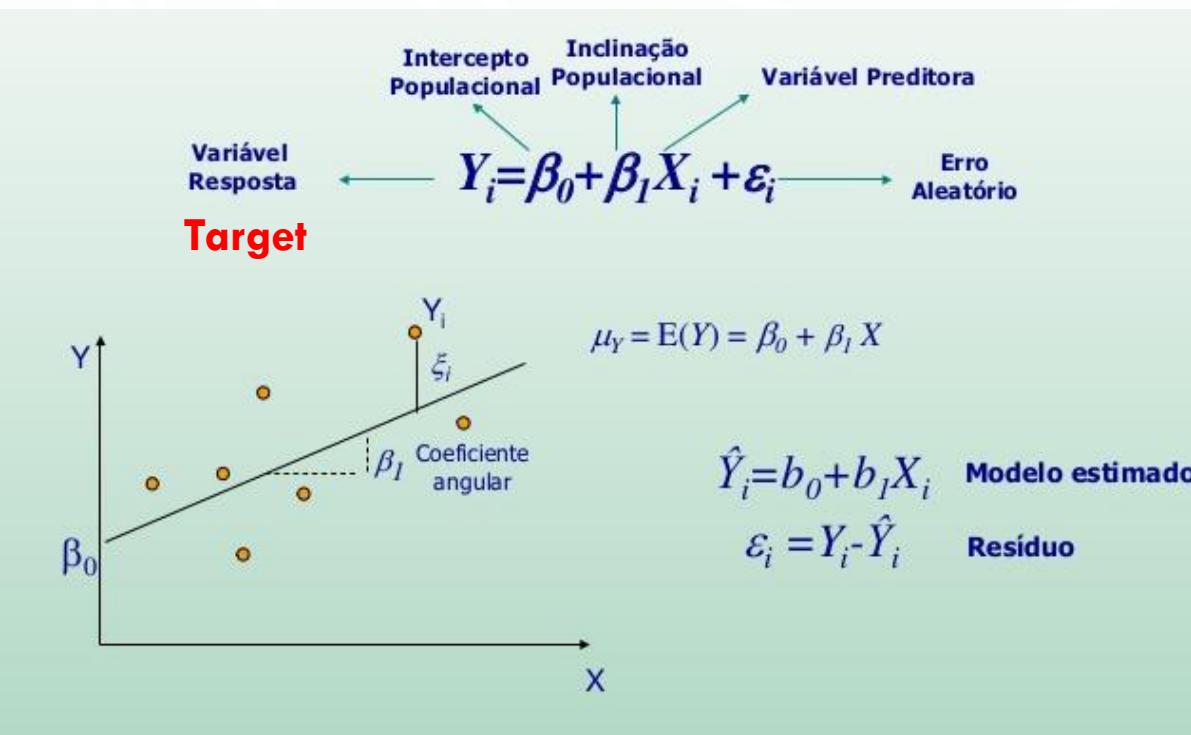
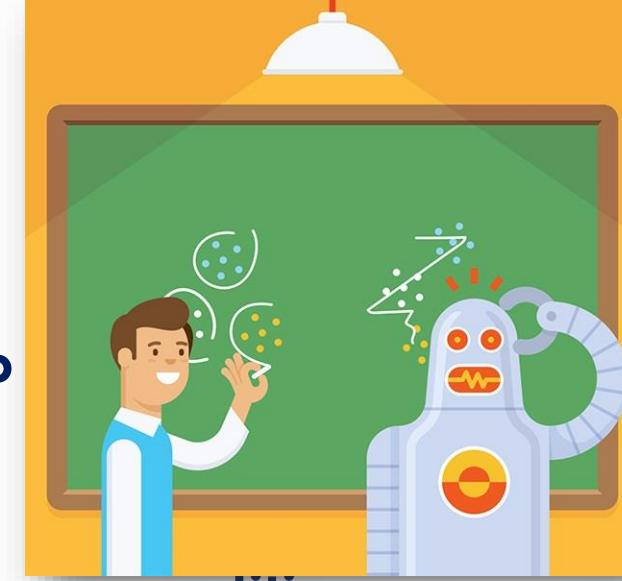
➤ Aprendizado baseado em modelo

➤ Modo de generalização de um conjunto de exemplos com um modelo matemático para fazer previsões

- A tarefa de regressão, também chamada de previsão ou estimativa, ocorre quando o atributo alvo da base de dados que se deseja aprender possui valor contínuo, como o preço de uma casa ou o lucro de uma empresa
- Nesse caso, o algoritmo de aprendizado de máquina deve encontrar um modelo matemático (equação) capaz de mapear as entradas numa saída esperada

ASPECTOS DE MACHINE LEARNING

- Um bom projeto de ML tem boa capacidade preditiva
- Acurácia nas decisões (acertos)
- Performance preditiva! Nem sempre a interpretação do processo é simples (pois decisões envolvem processos complexos)
- Em Inferência (regressão por exemplo), a relação entre as variáveis é melhor interpretável, mas usualmente pior performance preditiva

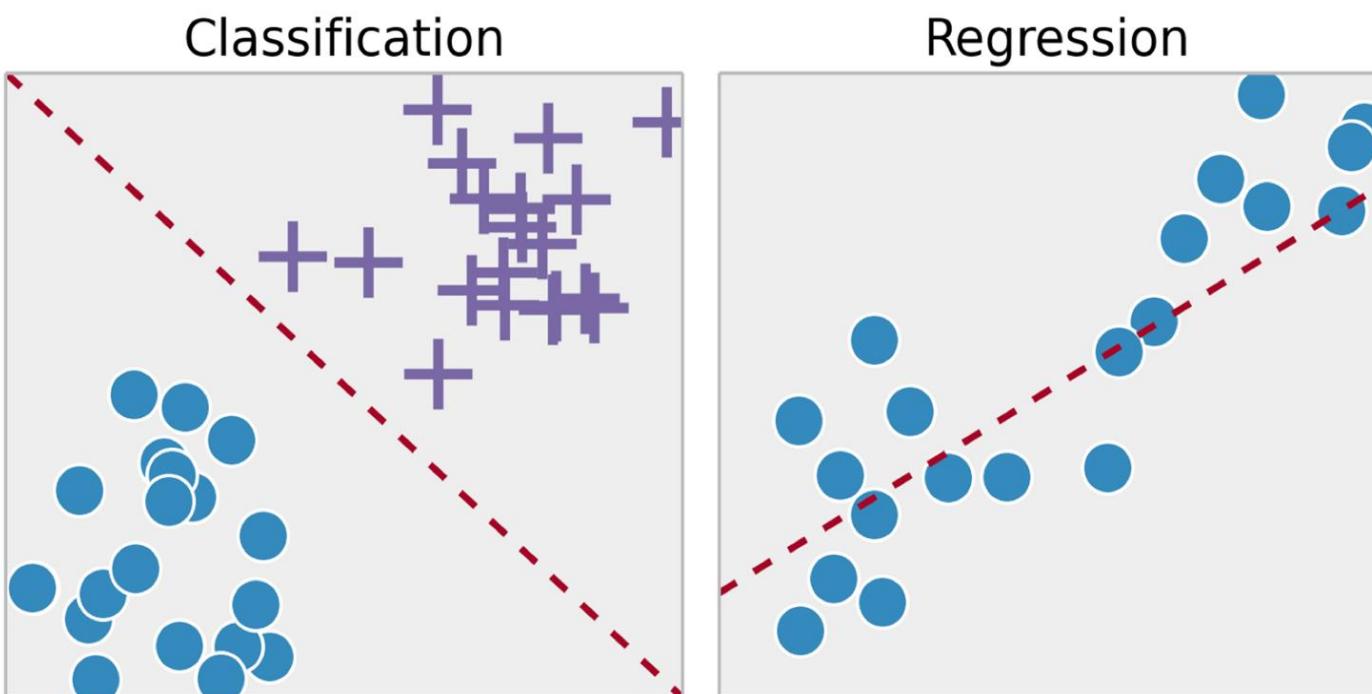


REGRESSÃO

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Normaltown	\$250,000
2	800	Hipsterton	\$300,000
2	850	Normaltown	\$150,000
1	550	Normaltown	\$78,000
4	2000	Skid Row	\$150,000

Bedrooms	Sq. feet	Neighborhood	Sale price
3	2000	Hipsterton	???

REGRESSÃO E CLASSIFICAÇÃO



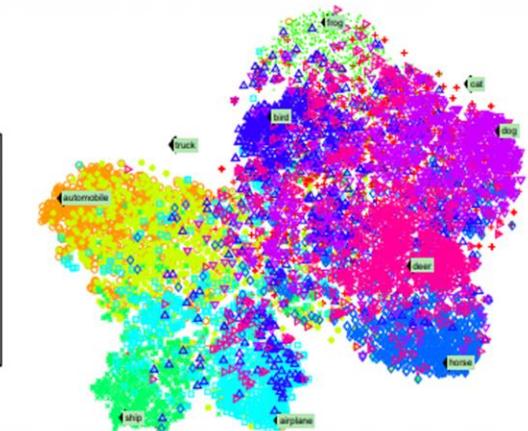
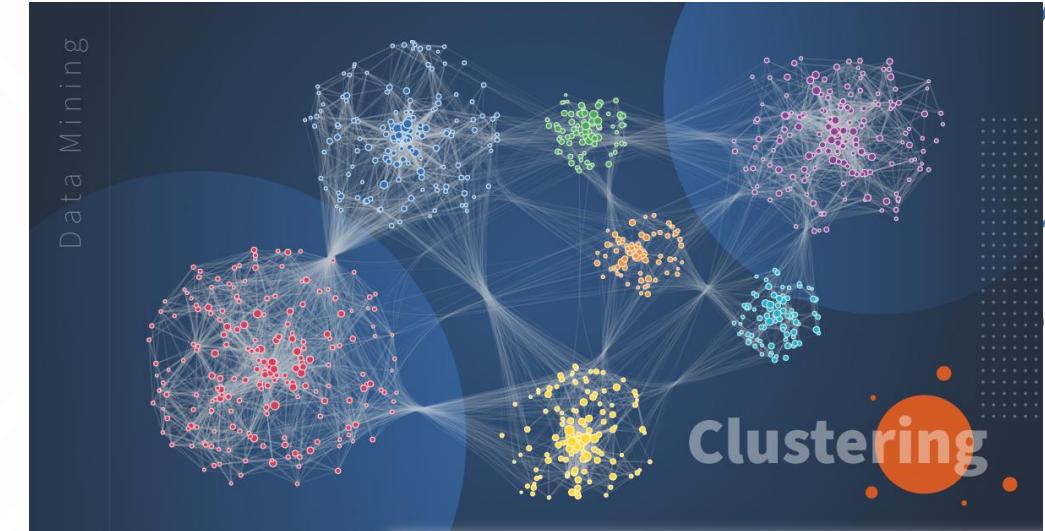
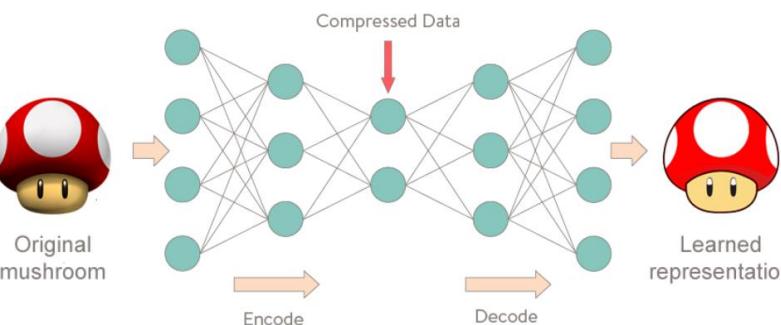
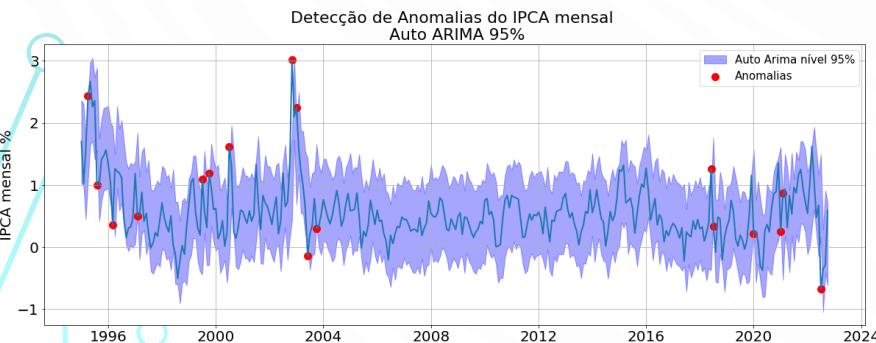
EXEMPLOS SUPERVISIONADOS

- Estimar o preço de uma casa
 - Atributos: tamanho, posição geográfica, material
 - Target: Preço (**regressão múltipla**)
- Estimar Índice de Vida Melhor da OCDE com base no PIB e da renda per capita do site do FMI
 - Atributo: PIB_per_capita [Code]
 - Target: satisfação_de_vida (**regressão simples**)
- Determinar se uma pessoa tem câncer benigno ou maligno
 - Atributos: Tamanho do tumor, formato do tumor, idade do paciente
 - Classe: Benigno ou maligno (**classificação**)



AGRUPAMENTO

- A tarefa de agrupamento busca reunir os exemplos por similaridade, criando grupos que serão posteriormente rotulados pelo cientista de dados
- Redução de dimensionalidade (simplificar dados sem perder muita informação) – merge de features correlacionadas (Feature Extraction)
- Detectar anomalias



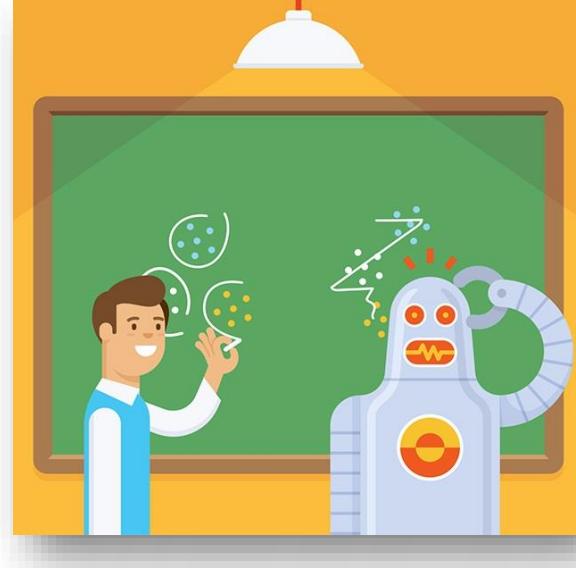
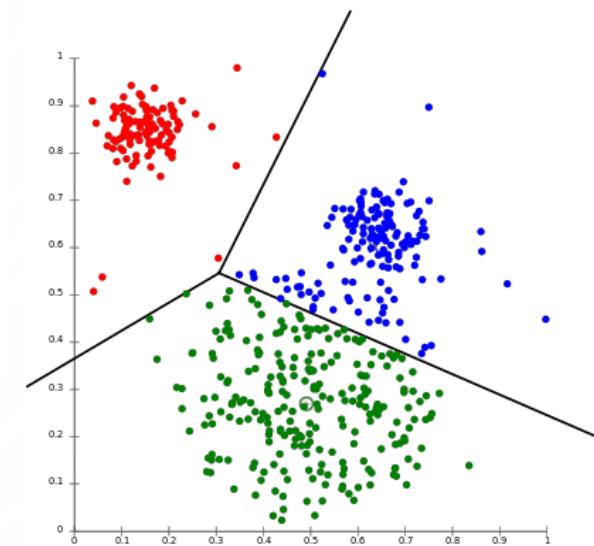
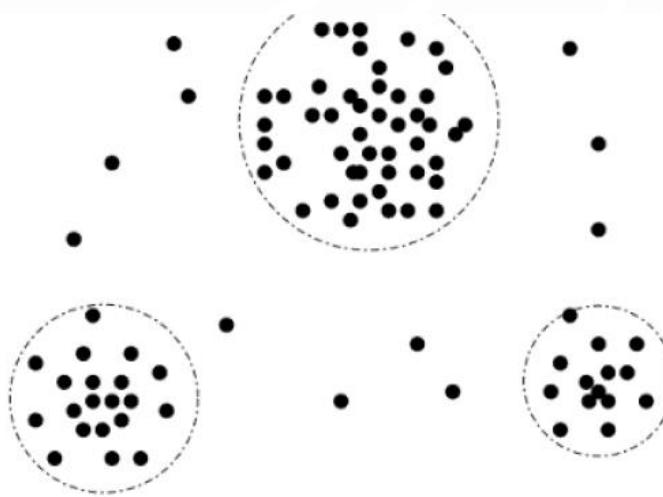
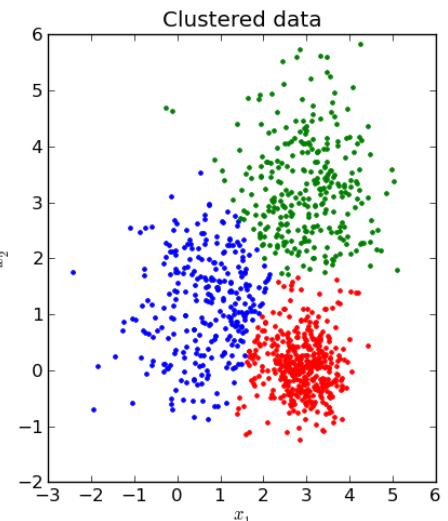
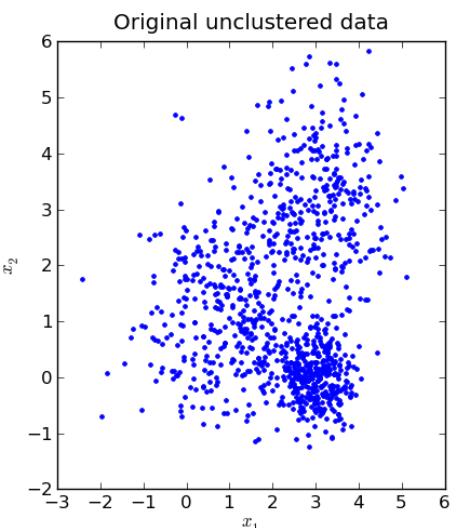
Fonte: <https://arxiv.org/pdf/1301.3666>

ASPECTOS DE MACHINE LEARNING

- Agrupamentos: não supervisionado – maximizar semelhanças (minimizar distâncias) dentro do cluster e maximizar diferenças (maximizar distâncias) entre clusters

Método K-means (incluindo Fuzzy c-means) é o mais usado

Mas existem vários outros métodos e variantes:
Hierárquicos, aglomerativos, incremental etc.



https://github.com/josenalde/machinelearning/blob/main/src/nb_kmeans1.ipynb

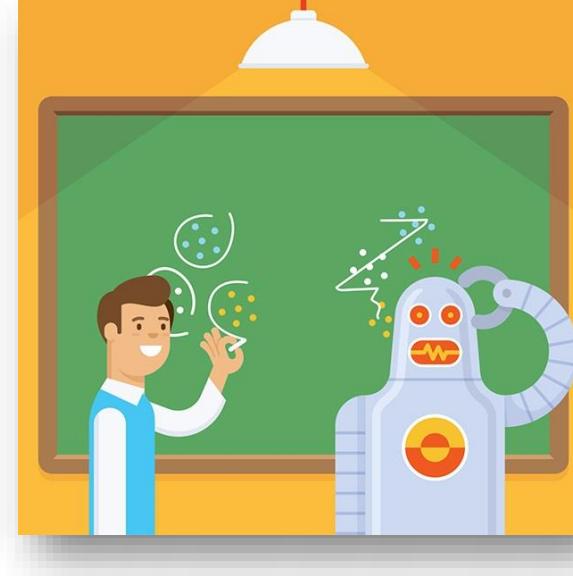
ASSOCIAÇÃO

transações

TID	Items
1	pão, leite
2	pão, fralda, cerveja, ovos
3	leite, fraldas, cerveja, coca
4	pão, leite, fraldas, cerveja
5	pão, leite, fraldas, coca

Exemplos de regras de associação

$\{\text{fraldas}\} \rightarrow \{\text{cerveja}\}$,
 $\{\text{leite, pão}\} \rightarrow \{\text{ovos,coca}\}$,
 $\{\text{cerveja, pão}\} \rightarrow \{\text{leite}\}$,



Implicação significa co-ocorrência, e não causa!!!

- Também chamado de mineração de regras de associação
- Essa tarefa busca identificar padrões nos dados analisados, como ocorrências de valores juntos em um mesmo exemplo da base
- No exemplo a seguir, espera-se que o algoritmo encontre os itens que normalmente são adquiridos juntos

EXEMPLOS NÃO SUPERVISIONADOS

- Identificar padrões de compras dos clientes de um supermercado (**Associação**)
- Identificar padrões de navegação em sites (**Associação**)
- Agrupar notícias semelhantes publicadas em várias fontes (**Agrupamento**)
- Numa rede social, identificar subgrupos de pessoas (**Agrupamento**)
- Identificar aves macho e fêmea em lotes mistos de frangos de corte em aviários, a partir de medidas de peso coletados por balança automática (**Agrupamento**)

TIPOS DE AM QUANTO À ATUALIZAÇÃO...

- Aprendizado em batch (lote, por ciclo)
 - Não aprende de forma incremental
 - Deve ser treinado a cada vez, com todos os dados disponíveis
 - Realizado em geral offline, pois demanda tempo e recursos

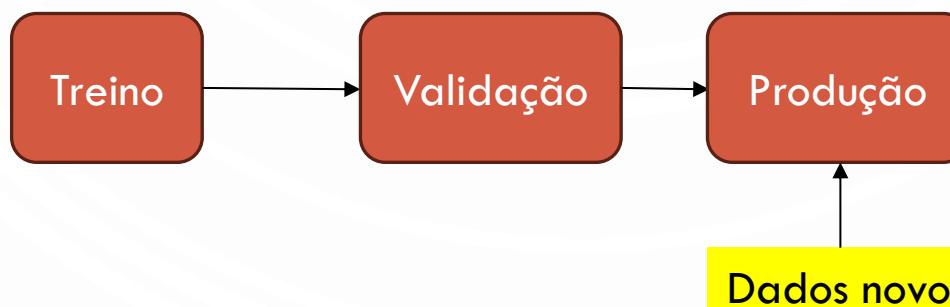


- Se surgem novos dados, treina nova versão a partir do zero e substitui o que está em produção – isto pode ser automatizado (24h, semanalmente (treino com muitos dados X custo \$)

TIPOS DE AM QUANTO À ATUALIZAÇÃO...

Aprendizado incremental (online)

- É possível treinar incrementalmente, fornecendo instâncias de forma sequencial, individual ou pequenos grupos (mini batches)
- Aprendizado rápido, custo baixo, aprende com dados novos em tempo real tão logo entrem
- Bom para dados em fluxo contínuo, adaptação às mudanças rápidas



- Grande conjunto de dados pode ser dividido (*out-of-core*) e treinado aos poucos
- Atenção à taxa de aprendizagem; **demandam monitoramento!**