

FUNDAMENTOS DE CIÊNCIA DE DADOS



PPGI

PROGRAMA
DE PÓS-GRADUAÇÃO
EM INFORMÁTICA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Machine Learning



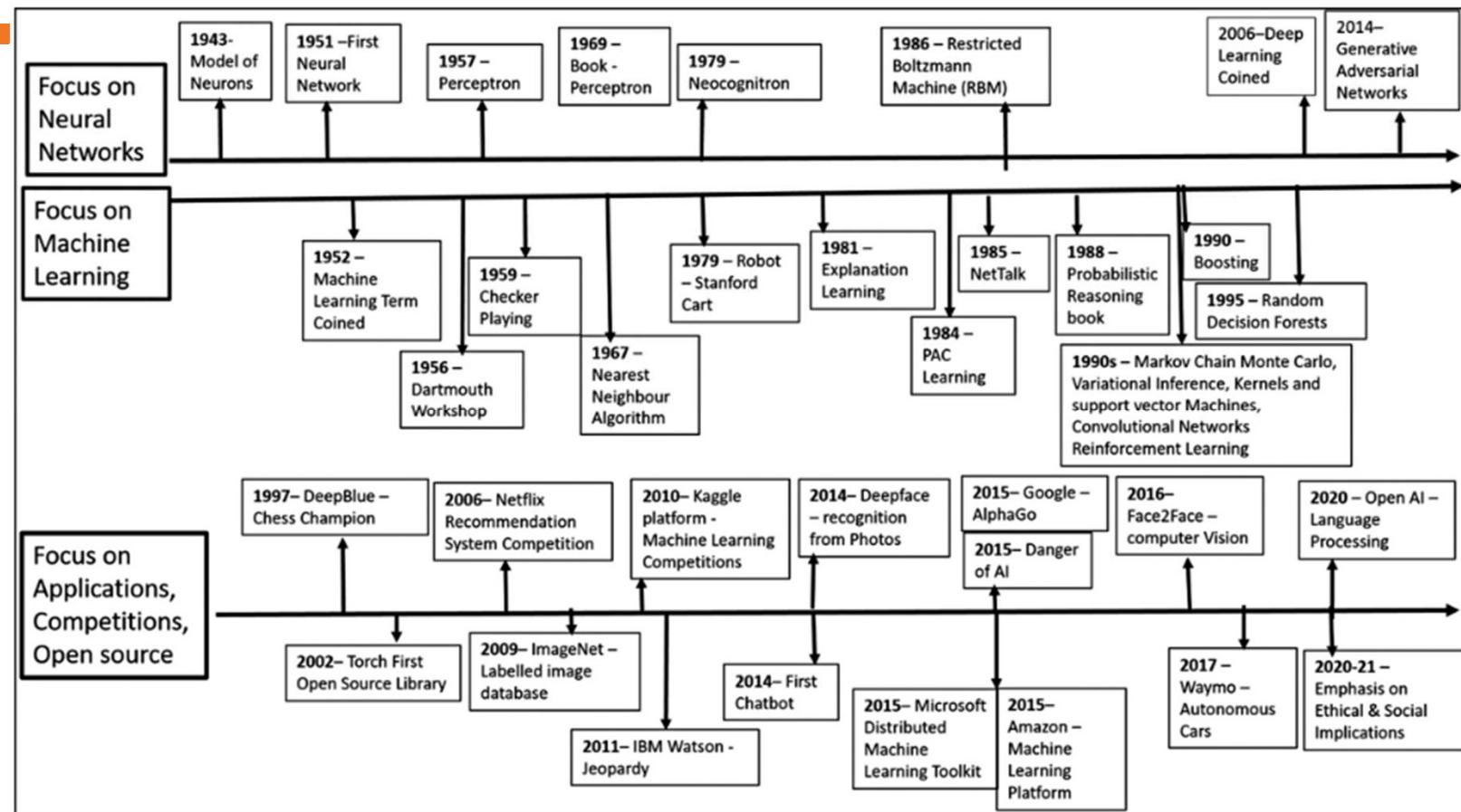
Prof. Jorge Zavaleta
Prof. Sergio Serra

RJ, Outubro de 2024

zavaleta@pet-si.ufrj.br

FUNDAMENTOS DE CIÊNCIA DE DADOS

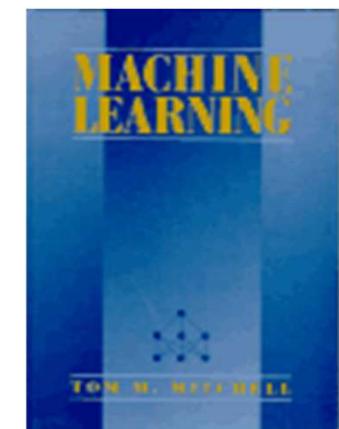
Machine Learning (ML) – História antiga



@Geetha2023

Machine Learning – O que é?

- “Aprendizado de máquina é o campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados” - Arthur Samuels, 1959.
- “Um programa de computador aprende com a experiência E no que diz respeito a alguma tarefa T e a alguma medida de desempenho P, se o seu desempenho em T, conforme medido por P, melhorar com a experiência E” - Tom Mitchell, 1997



Machine Learning

O que resolve ML?



Problemas com grande número de regras ou processos manuais repetitivos.



Problemas complexos sem boas soluções.



Obter soluções de grandes volumes de dados .



...



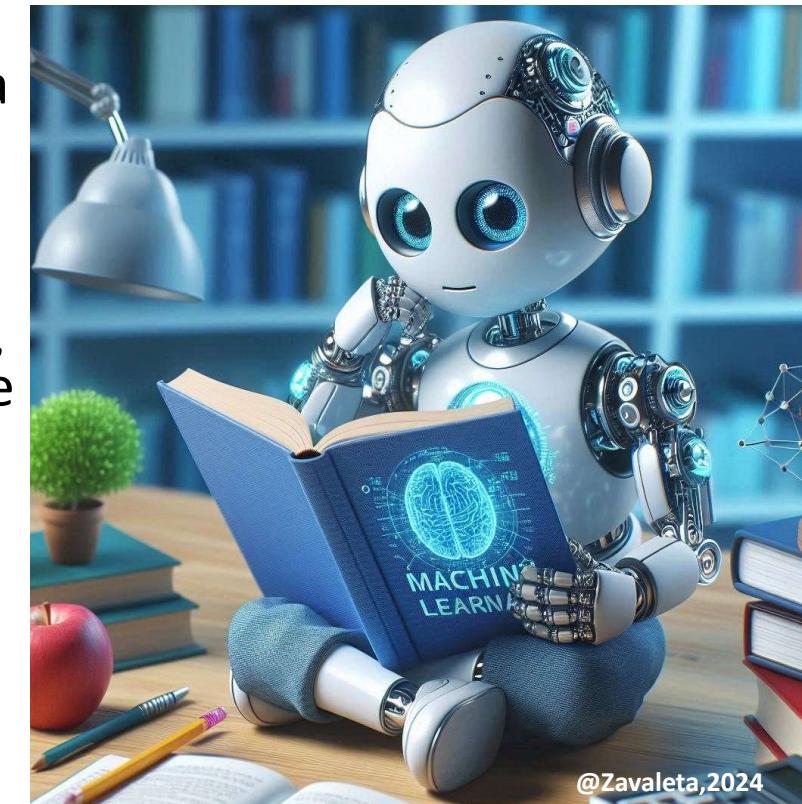
@Zavaleta.2024

O que é inteligência?

- “Inteligência é a capacidade de aprender, compreender ou lidar com situações novas ou difíceis (não é aprendizagem mecânica) e a capacidade de aplicar o conhecimento para manipular o ambiente ou pensar abstratamente conforme medido por alguns critérios objetivos.” (Dicionário Online Merryam Webster 2024).
- A inteligência envolve adaptação ou a capacidade de aplicar o que aprendemos ao ambiente em mudança de amanhã, ou por outras palavras, a inteligência exige que a aprendizagem seja dinâmica.
- Algumas das tarefas que requerem inteligência incluem raciocínio para resolver quebra-cabeças e fazer julgamentos, planejar sequências de ação, aprendizagem, processamento de linguagem natural, integração de habilidades e capacidade de sentir e agir.

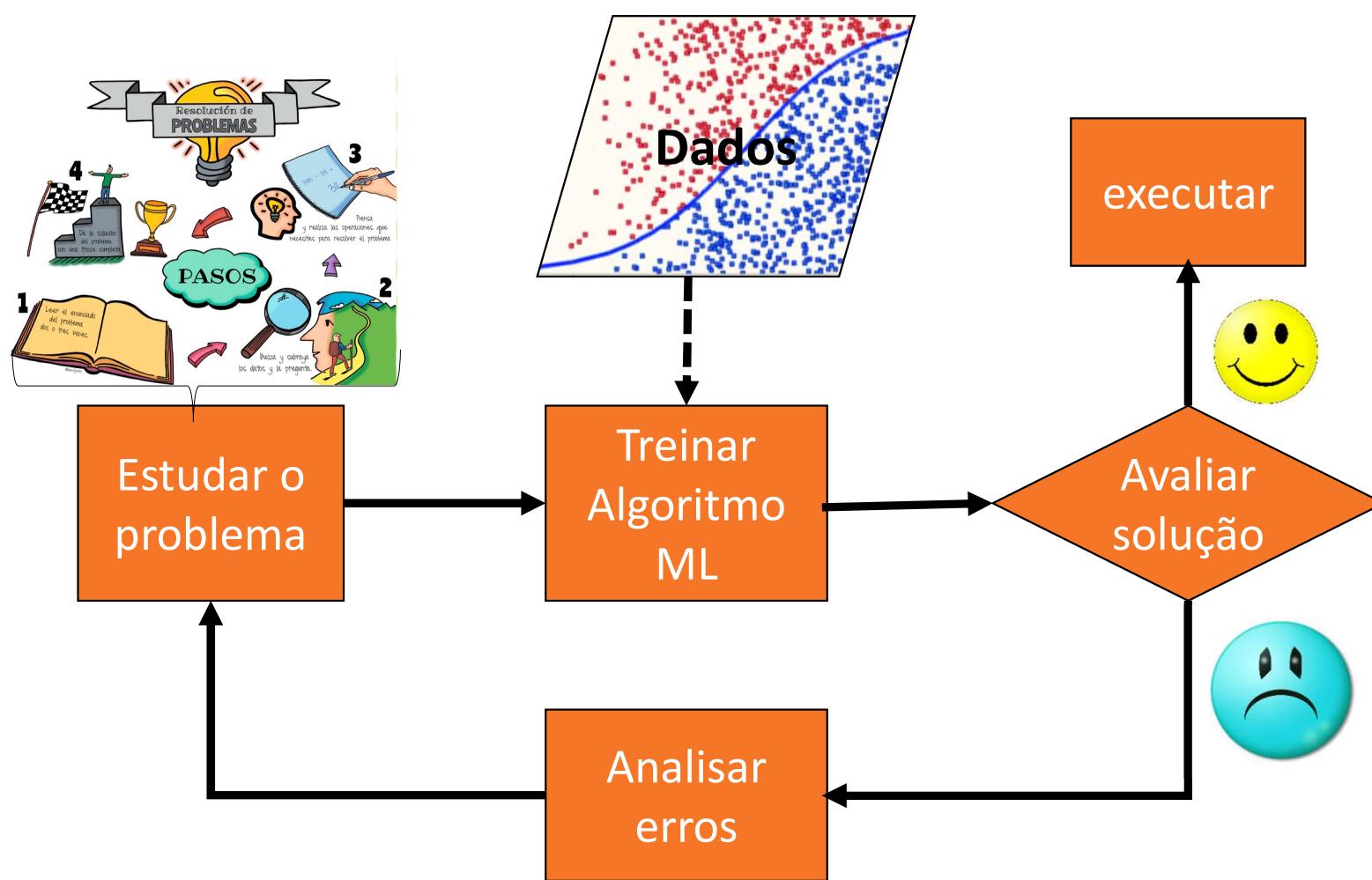
O que é aprendizagem?

- “Aprendizagem é qualquer processo pelo qual um sistema melhora seu desempenho a partir da experiência” (Herbert Simon, 1983).
- Para uma máquina, a experiência vem na forma de dados.
- A aprendizagem é o núcleo de muitas atividades, incluindo a cognição de alto nível, a realização de inferências com uso intensivo de conhecimento, a construção de sistemas inteligentes adaptativos, o tratamento de dados confusos do mundo real e a análise de dados.
- Inteligência + aprendizagem = Aprendizagem de máquina

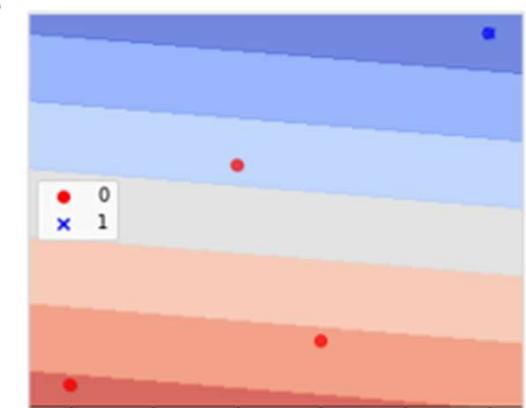
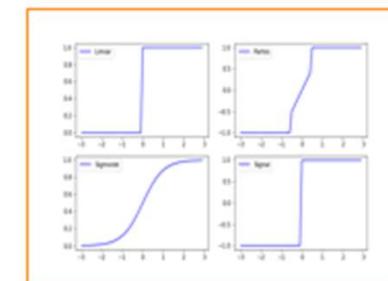
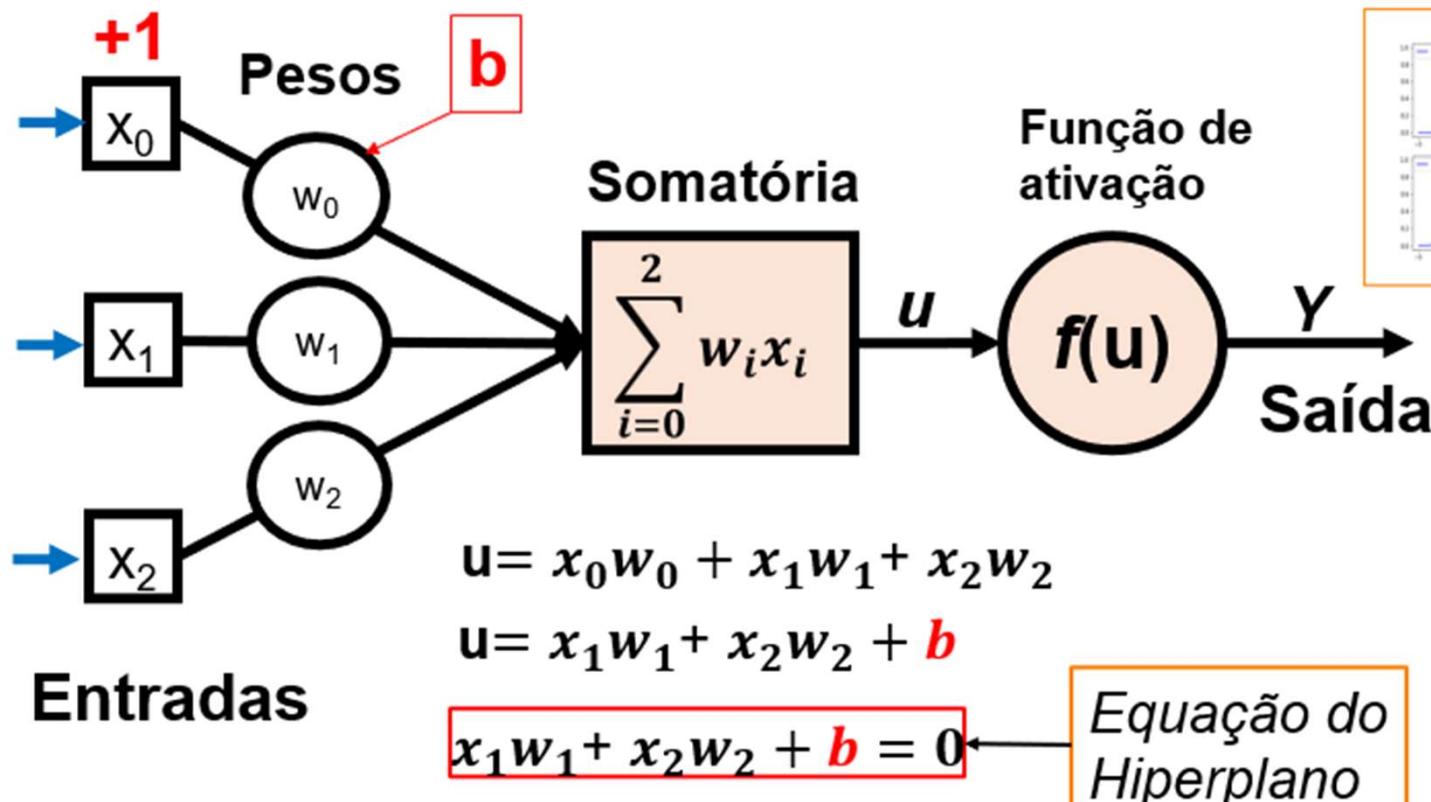


@Zavaleta, 2024

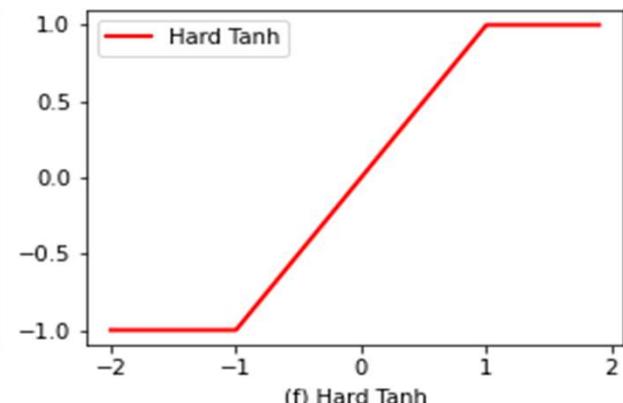
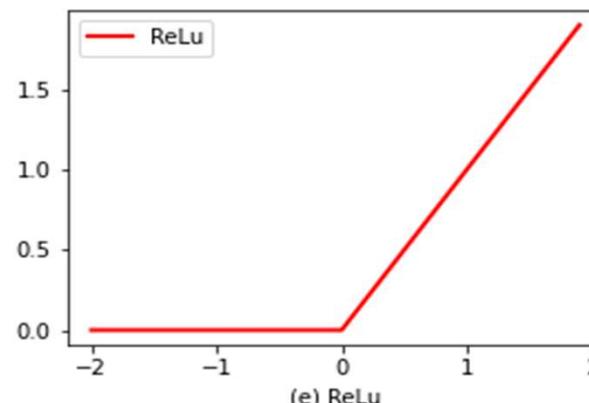
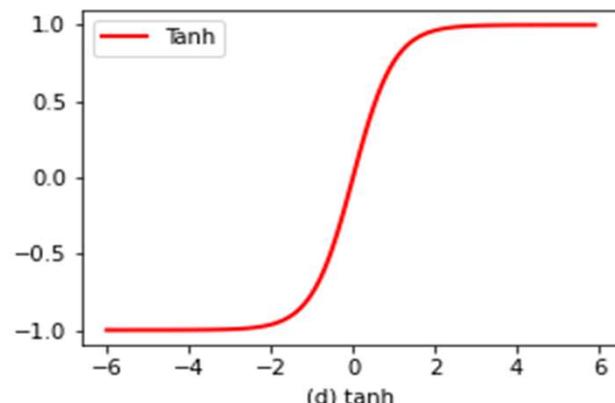
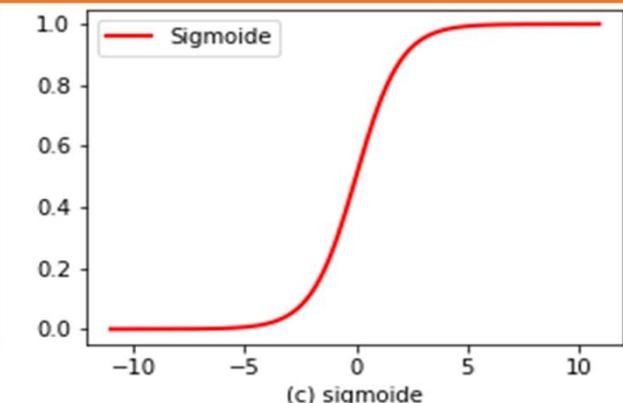
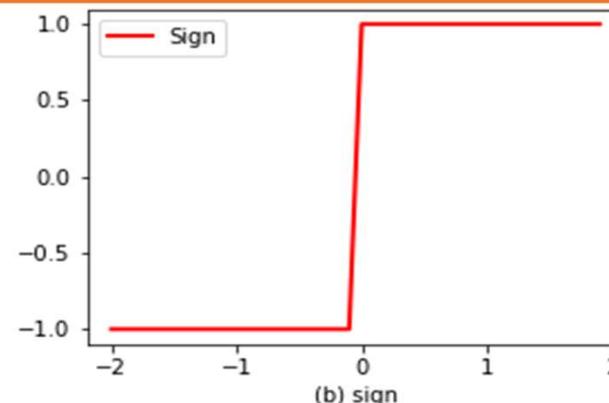
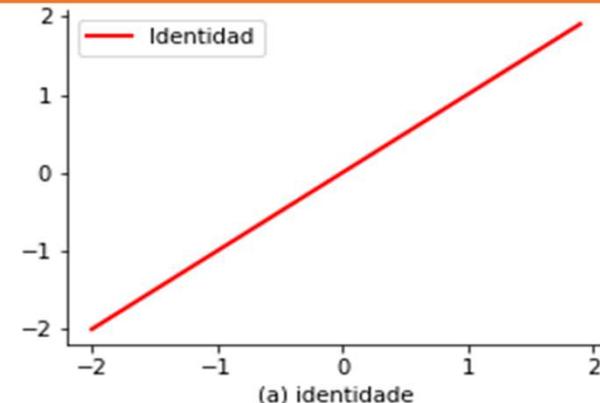
FUNDAMENTOS DE CIÊNCIA DE DADOS



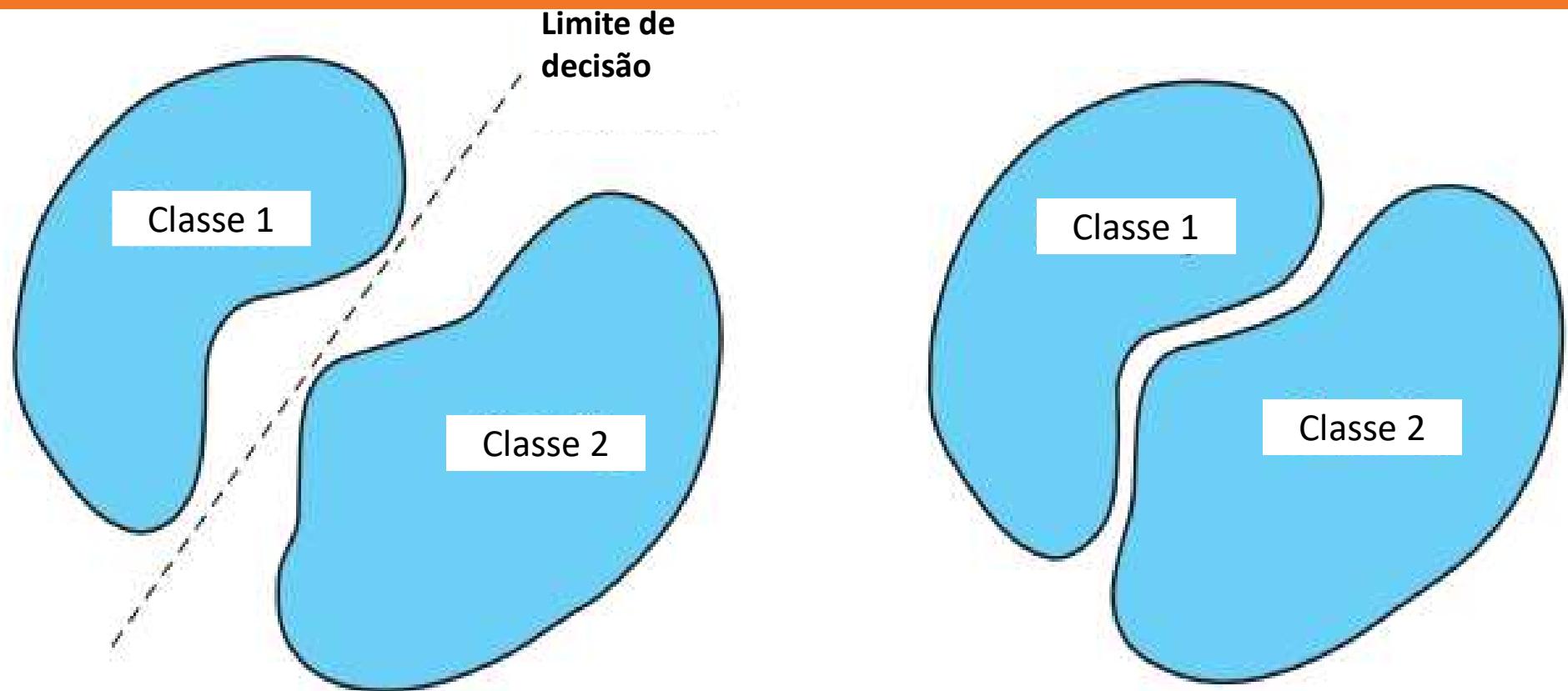
Aprendizagem – Modelo matemático



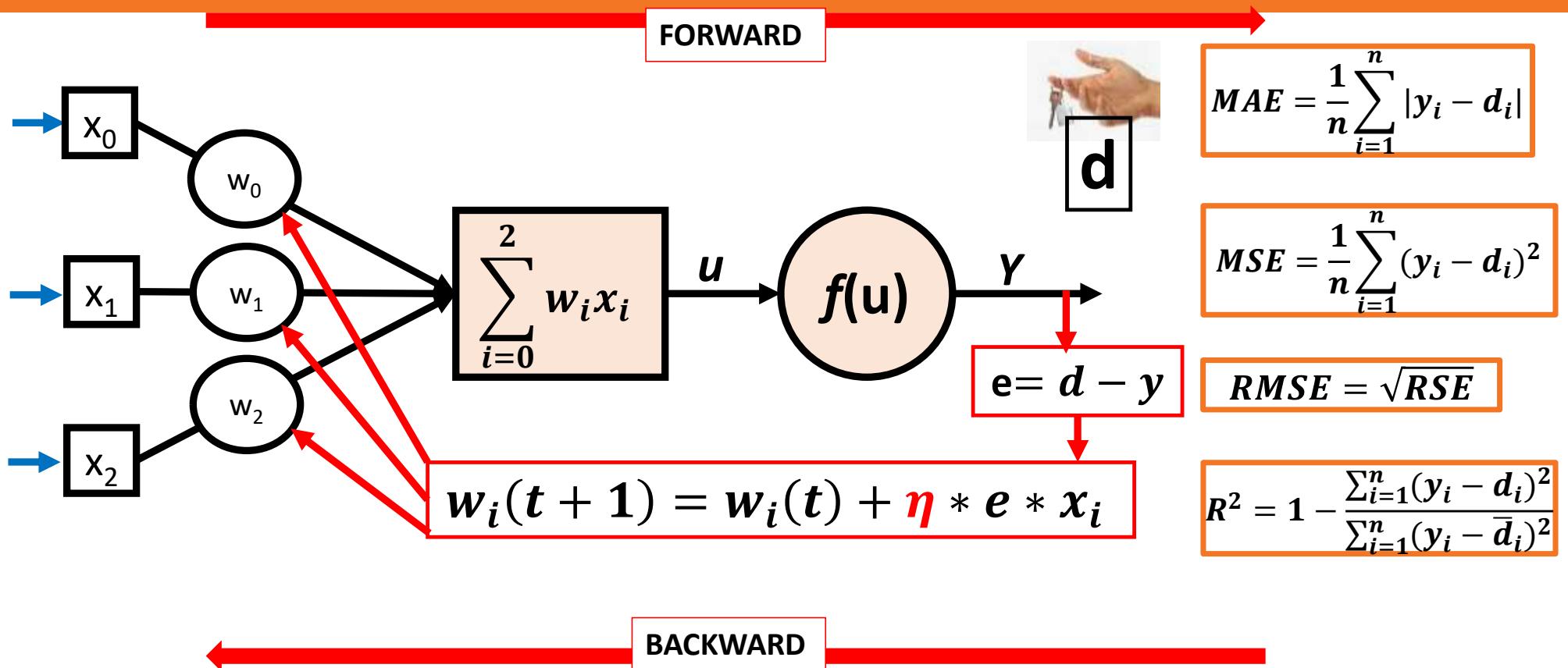
Funções de Ativação



Hiperplano



Aprendizagem (perceptron)



ML – Tipos de Aprendizado



Aprendizagem



Supervisionado

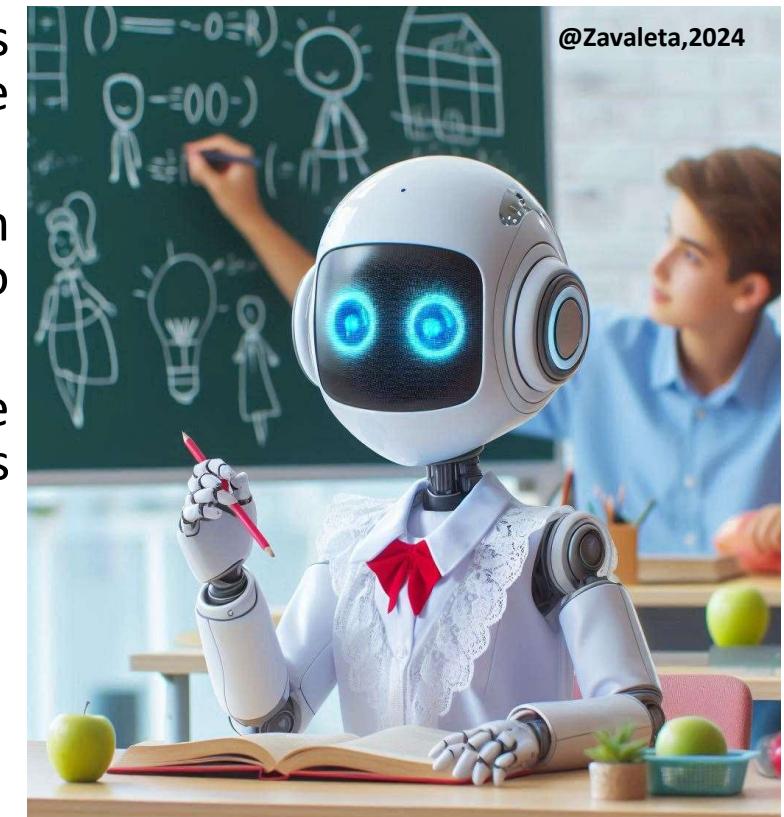


Não Supervisionado

Por Reforço

Aprendizado Supervisionado

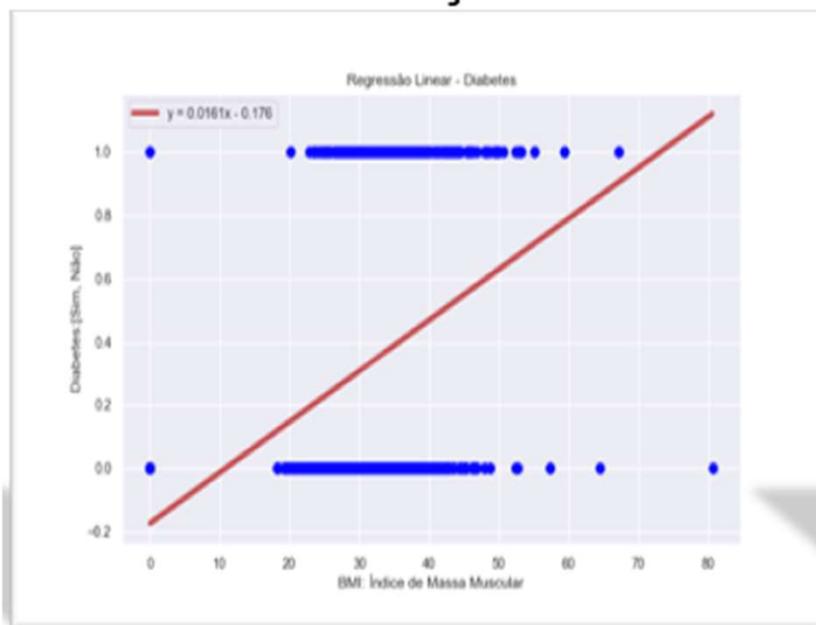
- Os algoritmos de aprendizado supervisionados fazem previsões com base em um conjunto de exemplos de entrada.
- Neste tipo de aprendizagem existe um “professor/tutor” que avalia a resposta da rede ao padrão atual de entradas.
- No aprendizado supervisionado, o conjunto de treinamento que se alimenta o algoritmo inclui as soluções desejadas, chamadas rótulos.
- O aprendizado supervisionado realiza tarefas de:
 - **Régressão**
 - **Classificação**



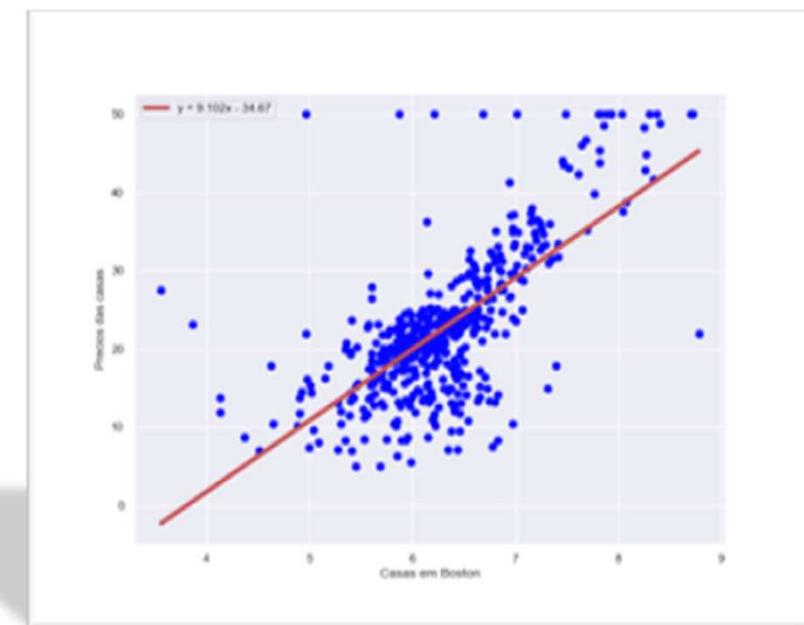
FUNDAMENTOS DE CIÊNCIA DE DADOS

Aprendizado Supervisionado

- **Regressão:** métodos de regressão buscam encontrar como uma variável evolui em relação a outras.

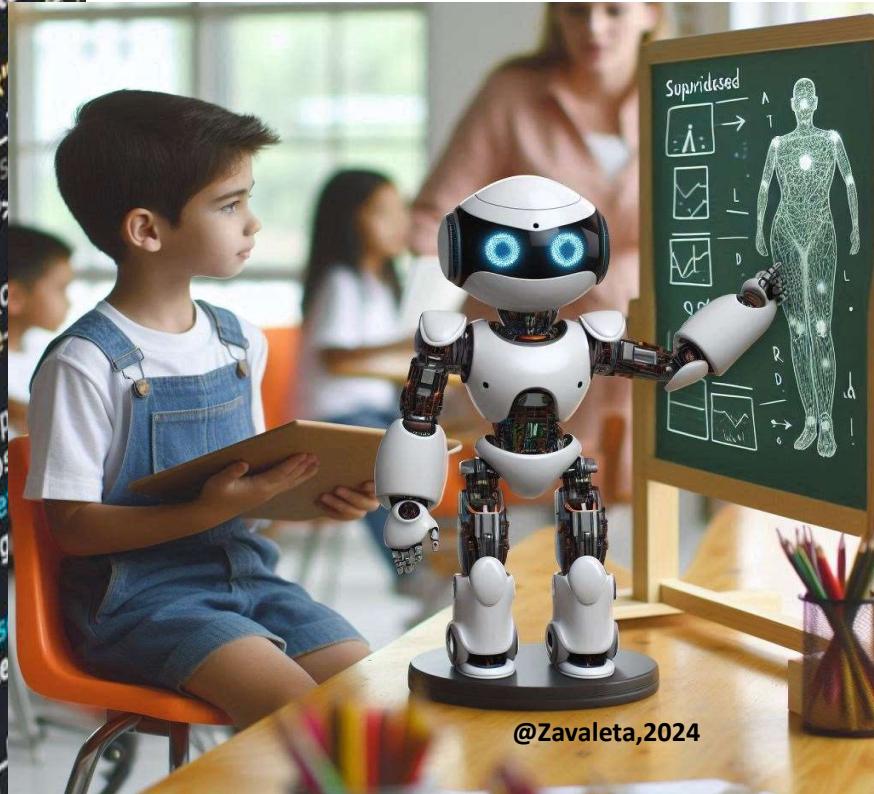


Diabetes



Preços de casas em Boston

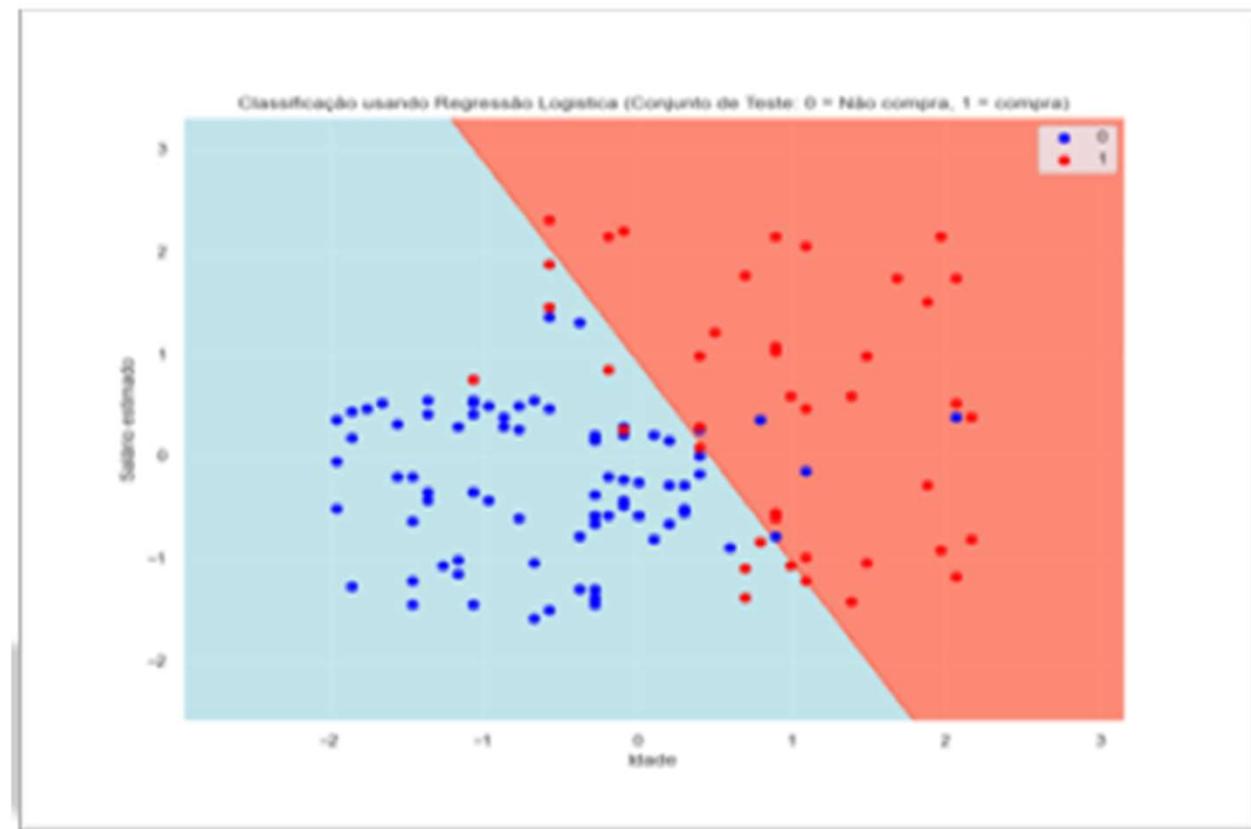
Regressão - Aplicações



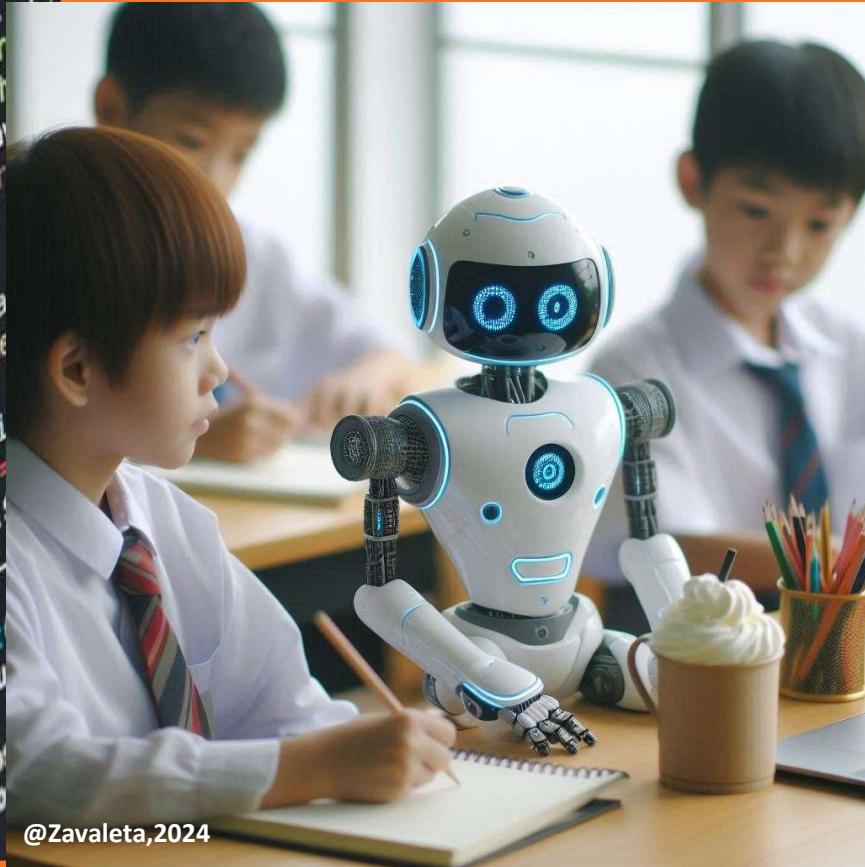
- Previsão do mercado de ações,
- Previsão de demanda,
- Estimativa de preços,
- Otimização de lances de anúncios,
- Gerenciamento de riscos,
- Gerenciamento de ativos,
- Previsão do tempo (clima),
- Previsão de esportes

Aprendizado Supervisionado

- **Classificação:** são métodos que buscam explicar uma variável categórica, com duas categorias (variável binária) ou mais.



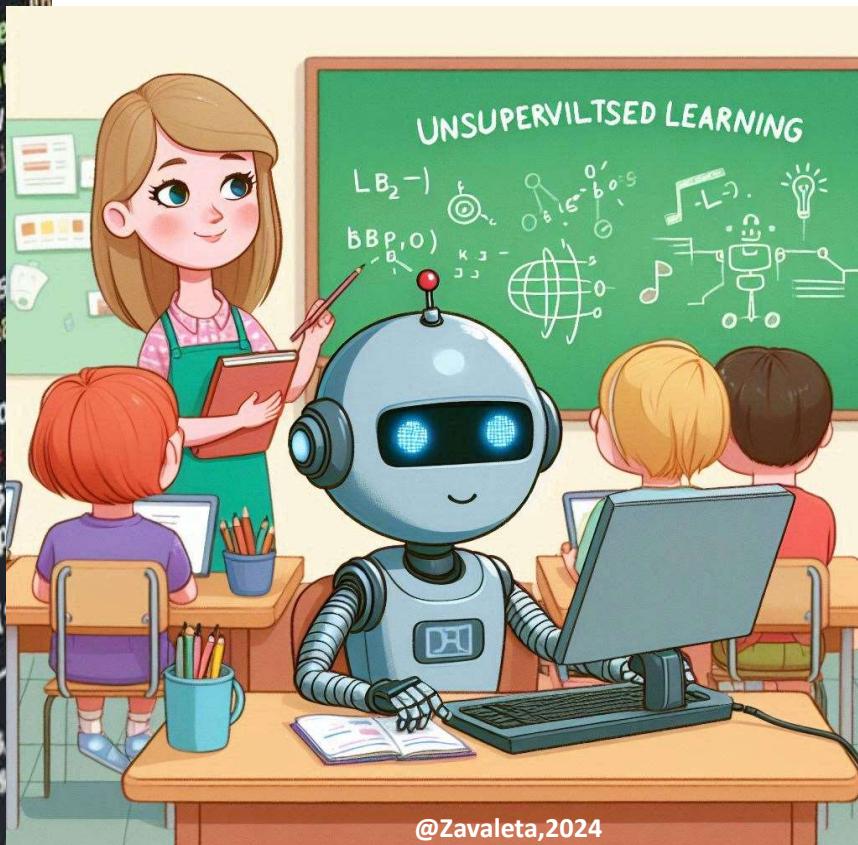
Classificação - Aplicações



@Zavaleta,2024

- Filtros de spam,
- análise de sentimentos,
- detecção de fraude,
- segmentação de anúncios de clientes,
- previsões de rotatividade,
- sinalização de casos de suporte,
- personalização de conteúdo,
- segmentação de clientes,
- descoberta de eventos,
- eficácia de medicamentos,
- detecção de defeitos de fabrica etc.

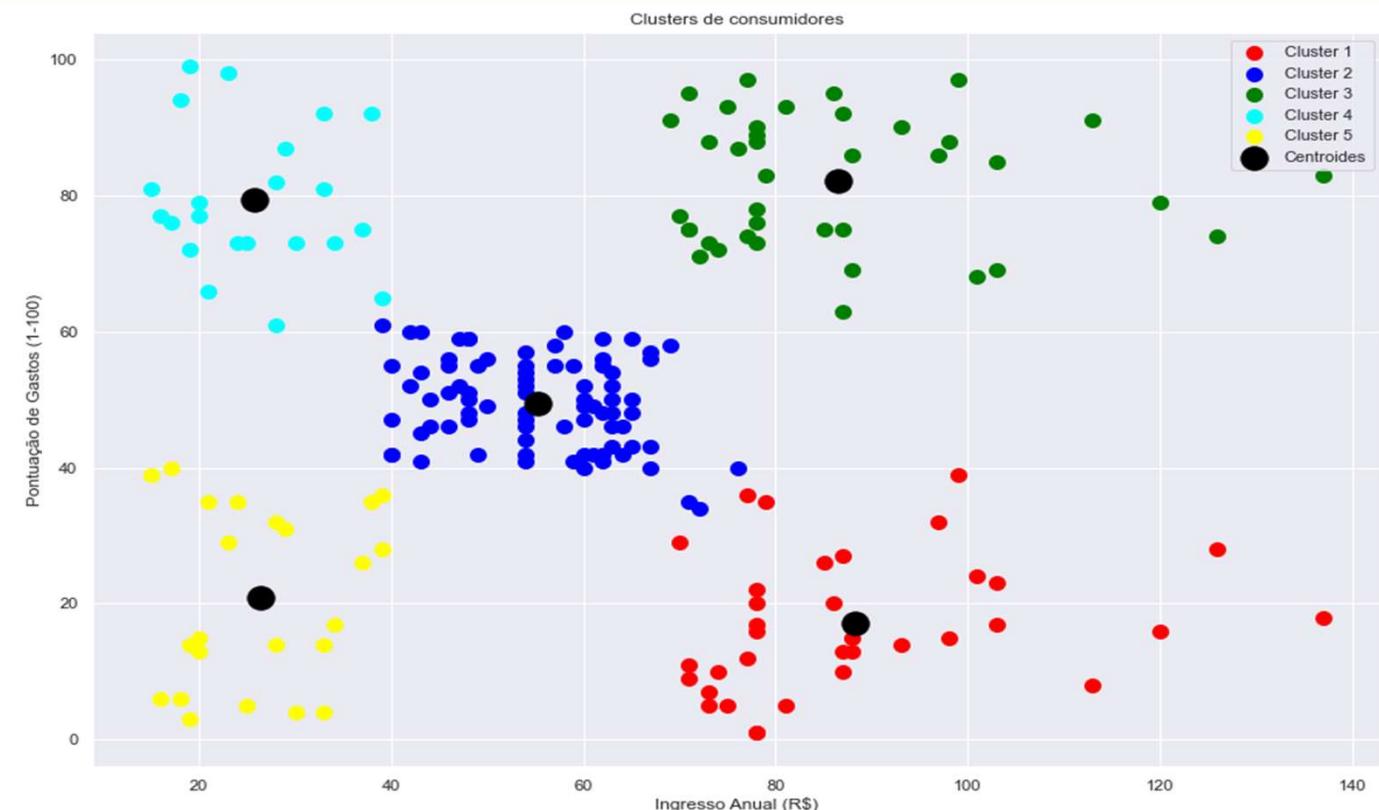
Aprendizado Não Supervisionado



- Não existe “**professor/tutor**”.
- A rede tem que descobrir sem ajuda as relações, padrões, regularidades ou categorias nos dados que lhe vão sendo apresentados e codificá-las nas saídas.
- Os dados de treinamento são não rotulados.
- As tarefas típicas da aprendizagem não supervisionada são:
 - **Clusterização**
 - **Análises de Componentes Principais (PCA)**
 - **Redução de dimensões**, etc.

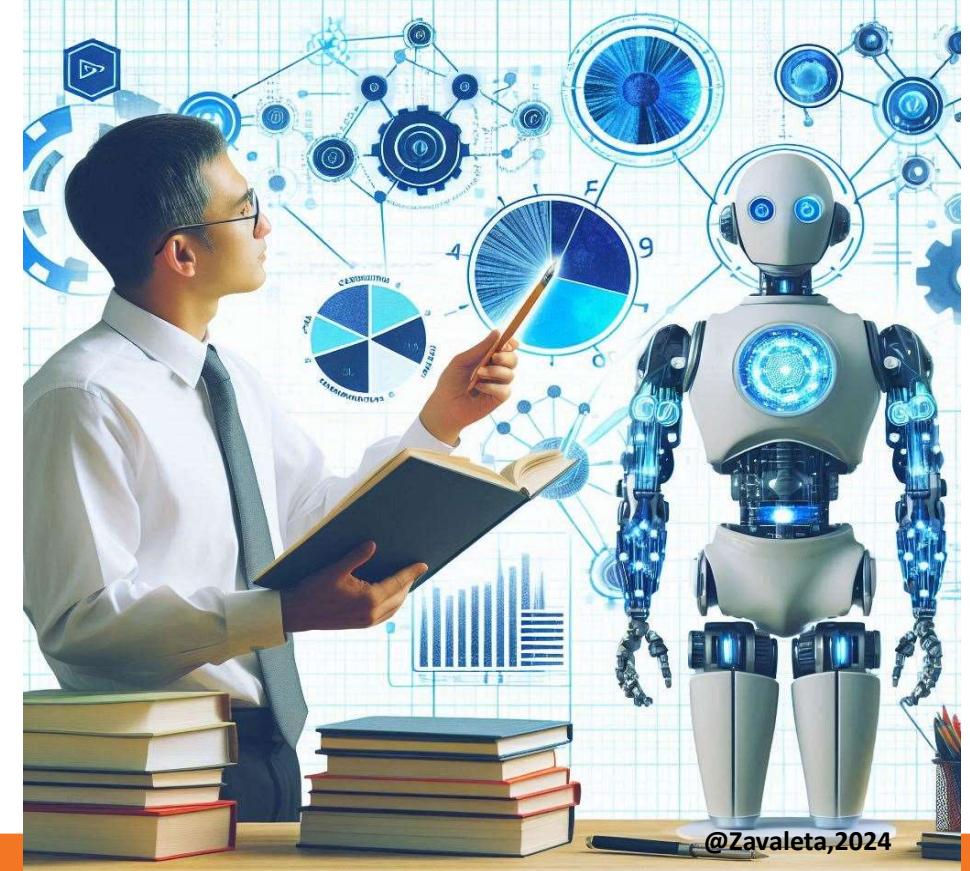
Aprendizado Não Supervisionado

- **Clusterização** é a tarefa de dividir os pontos de dados em vários grupos com características semelhantes.
- Cada grupo possui um ponto central, denominado de centroide.



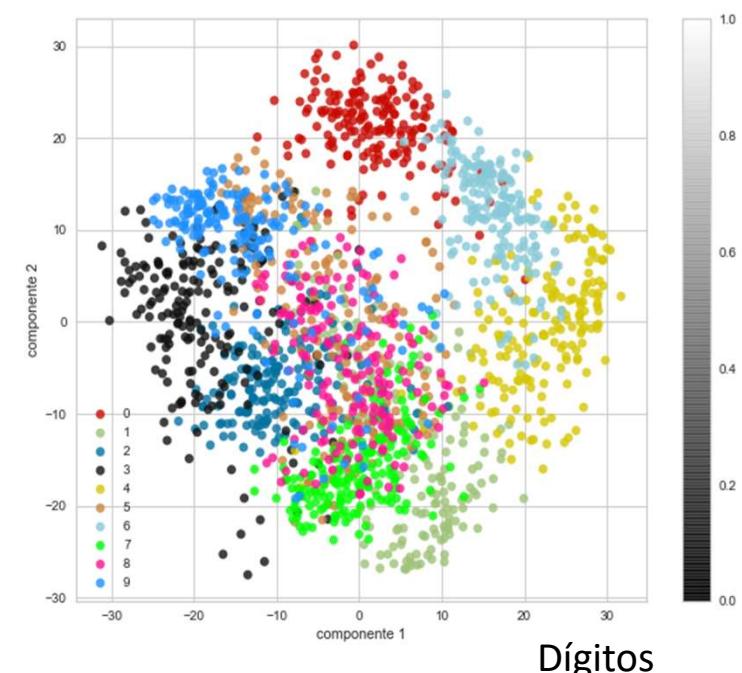
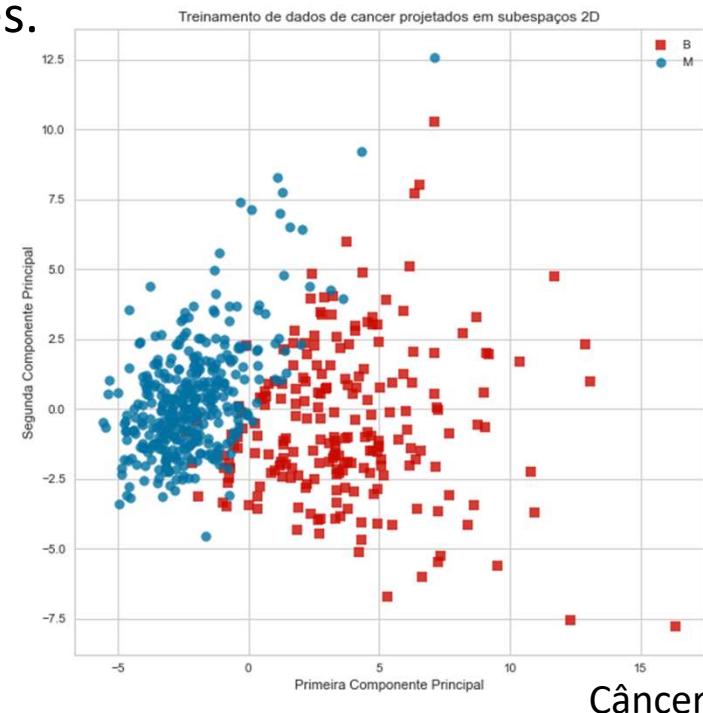
Clusterização - Aplicações

- Segmentação de Mercado
- Análise de Imagens
- Biologia e Genômica
- Detecção de Anomalias
- Agrupamento de Documentos
- Sistemas de Recomendação
- Análise de Dados de Clientes
- Exploração de Dados
- Modelagem de Associações:
- Planejamento Urbano



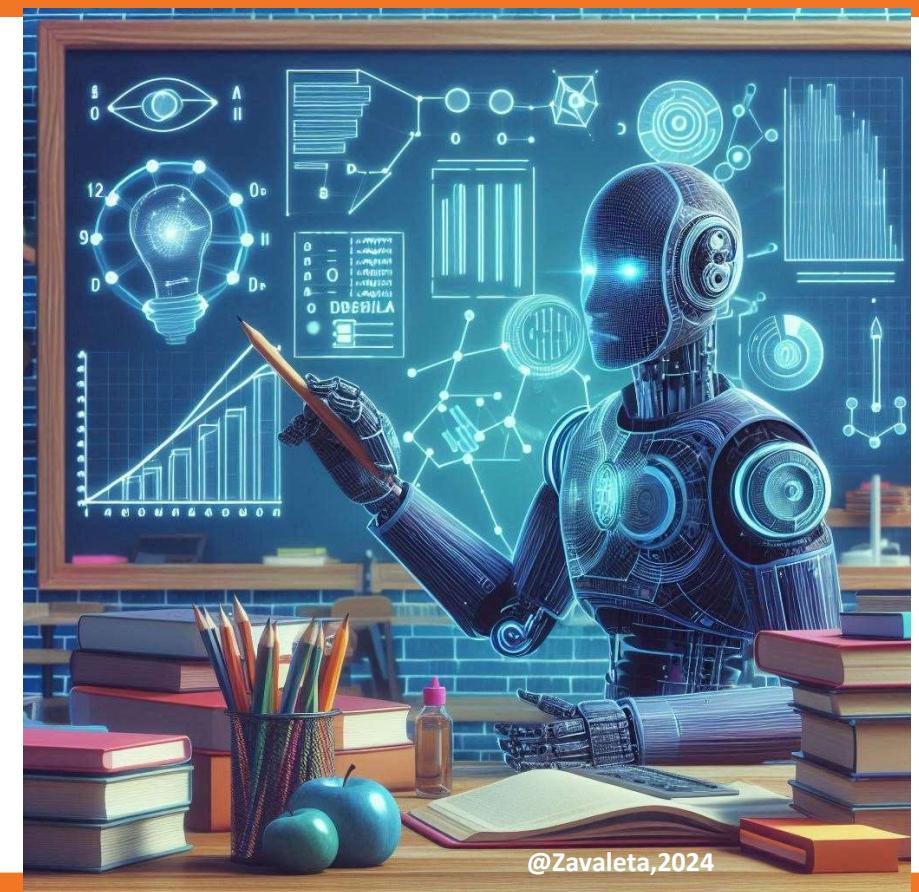
Aprendizado Não Supervisionado

- **Principal Component Analysis (PCA -Análise de Componentes Principais)** é um procedimento matemático que permite reduzir variáveis n-dimensionais em variáveis lineares.



PCA - Aplicações

- Redução de Dimensionalidade
- Compressão de Dados
- Pré-processamento para Modelagem Preditiva
- Visualização de Dados
- Detecção de Anomalias
- Reconhecimento de Padrões
- Genômica e Bioinformática
- Análise Financeira
- Processamento de Imagens
- Análise de Dados Meteorológicos



Resumo dos tipos de aprendizagem

Supervised Learning

Labeled Data
Direct Feedback
Classification and Regression

Unsupervised Learning

Unlabeled Data
No Feedback
Clustering and Dimensionality Reduction

Semi-supervised Learning

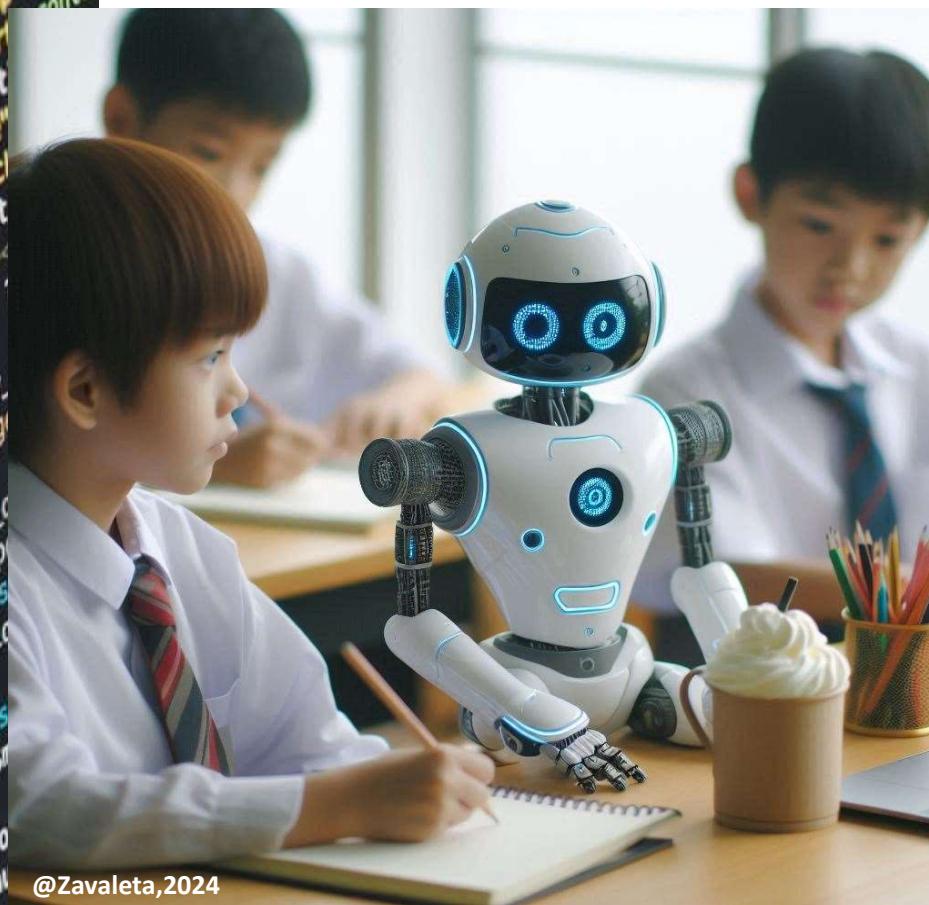
Labeled and Unlabeled Data
Some Feedback
Classification and Regression

Reinforcement Learning

Reward Based Learning
Direct Feedback
Learn series of actions

@Yalcin2021

Métricas

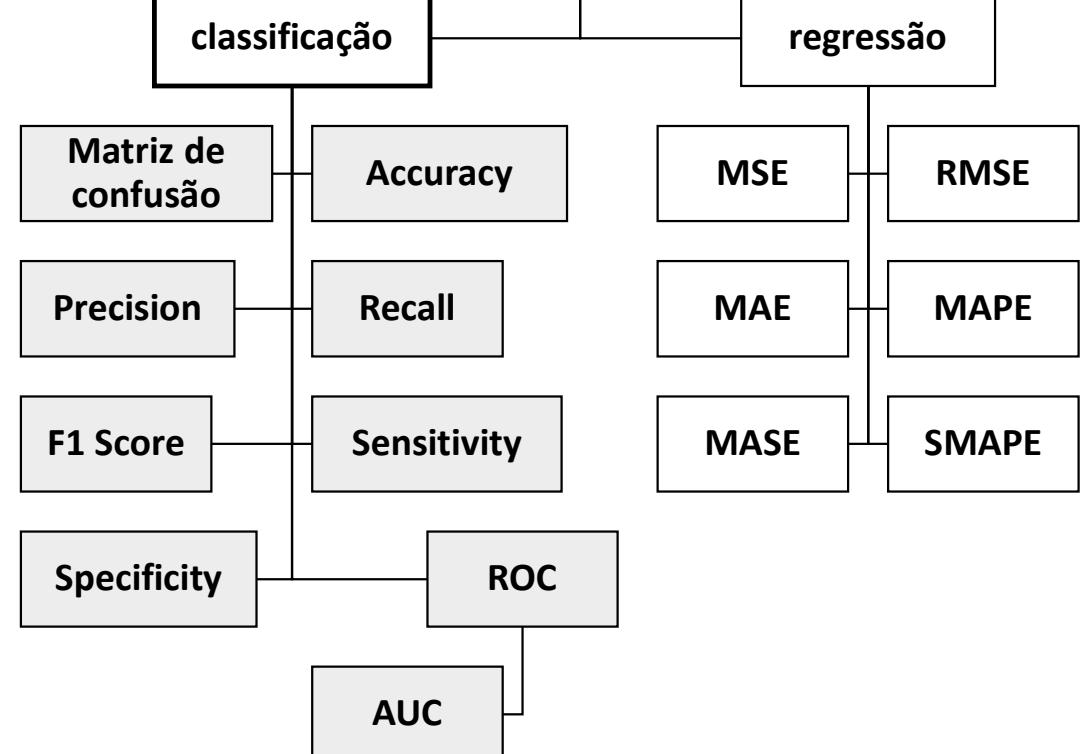


@Zavaleta,2024

Prof. Dr. Jorge Zavaleta

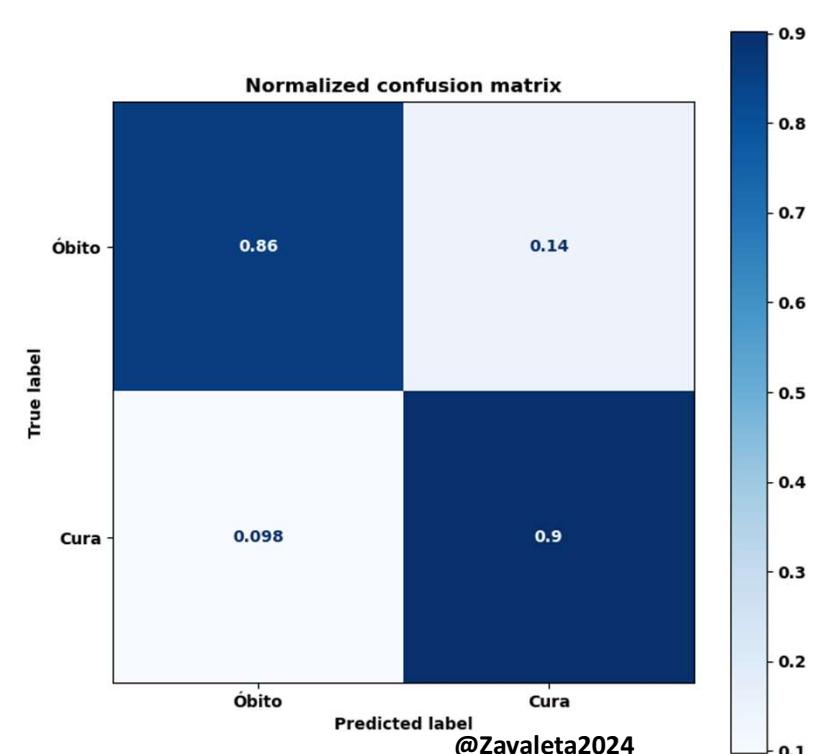
zavaleta@pet-si.ufrj.br

Desempenho



Métricas - Matriz de confusão

- A acurácia de um algoritmo é implementada usando a matriz de confusão.
- Uma matriz de confusão ilustra a precisão da solução para um problema de classificação.
- Uma matriz de confusão contém informações sobre classificações reais e previstas feitas por um sistema de classificação.



FUNDAMENTOS DE CIÊNCIA DE DADOS

Métricas – Matriz de confusão

		VERDADEIROS		PREVISTOS	
		POSITIVO (1)	NEGATIVO (0)	POSITIVO (1)	NEGATIVO (0)
PREVISOS	POSITIVO (1)	TP Verdadeiro Positivo	FN Falso Negativo	TP Verdadeiro Positivo	FN Falso Negativo
	NEGATIVO (0)	FP Falso Positivo	TN Verdadeiro Negativo	FP Falso Positivo	TN Verdadeiro Negativo
VERDADEIROS	POSITIVO (1)			POSITIVO (1)	
	NEGATIVO (0)			NEGATIVO (0)	

Métricas - Accuracy

		PREVISTOS	
		POSITIVO (1)	NEGATIVO (0)
VERDADEIROS	POSITIVO (1)	TP Verdadeiro Positivo	FN Falso Negativo
	NEGATIVO (0)	FP Falso Positivo	TN Verdadeiro Negativo

- Número de previsões corretas dividido pelo número total de previsões.
- A acurácia é uma métrica que fornece a fração de previsões corretas

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

Métricas - Precision

		PREVISTOS	
		POSITIVO (1)	NEGATIVO (0)
VERDADEIROS	POSITIVO (1)	TP Verdadeiro Positivo	FN Falso Negativo
	NEGATIVO (0)	FP Falso Positivo	TN Verdadeiro Negativo

- É a capacidade do modelo de não classificar um evento negativo como positivo.
- Fornece a fração que identifica corretamente como positiva dentre todos os positivos que foram previstos.

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

Métricas – Recall/Sensitivity

		PREVISTOS	
		POSITIVO (1)	NEGATIVO (0)
VERDADEIROS	POSITIVO (1)	TP Verdadeiro Positivo	FN Falso Negativo
	NEGATIVO (0)	FP Falso Positivo	TN Verdadeiro Negativo

- A recall/sensibilidade fornece a fração que identifica corretamente como positiva dentre todos os positivos ou indica o quanto preciso é o modelo durante a previsão.
- Esta medida mostra a fração de casos positivos corretamente identificados entre todos os casos positivos reais

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

Métricas - Specificity

		PREVISTOS	
		POSITIVO (1)	NEGATIVO (0)
VERDADEIROS	POSITIVO (1)	TP Verdadeiro Positivo	FN Falso Negativo
	NEGATIVO (0)	FP Falso Positivo	TN Verdadeiro Negativo

- Mede a capacidade do modelo em identificar corretamente os casos negativos.
- É a proporção de verdadeiros negativos em relação ao total de indivíduos que não têm a condição.

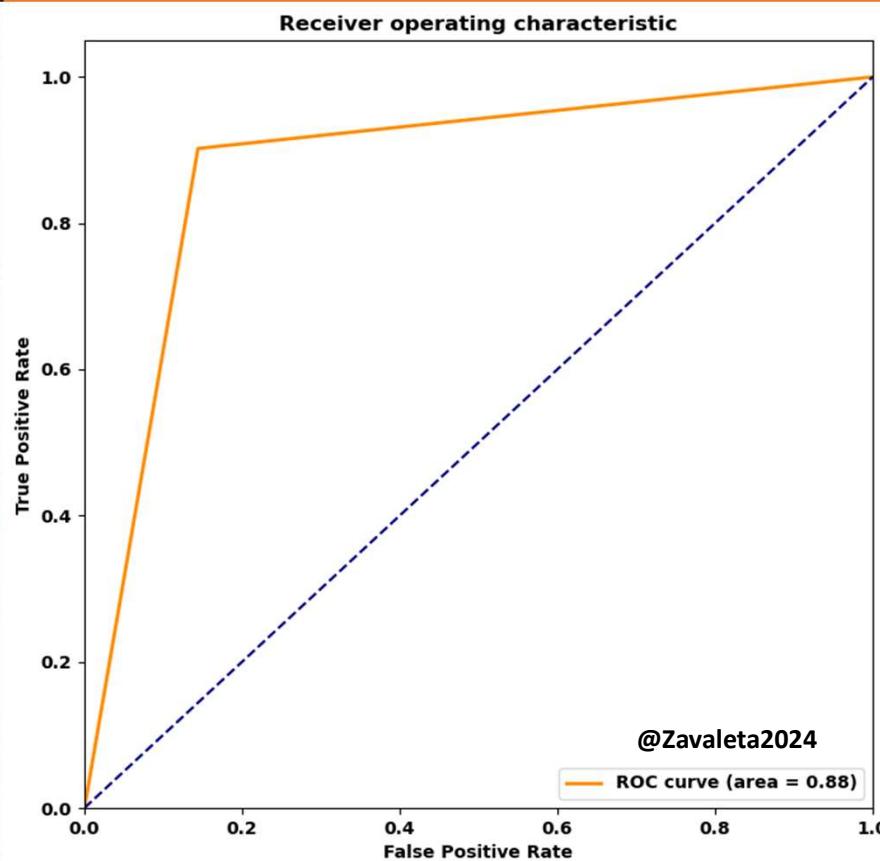
$$\text{specificity} = \frac{TN}{TN + FP}$$

Métricas – F1 Score

- F1- score é a média harmônica entre a Precisão e o Recall e fornece um equilíbrio entre essas duas métricas.
- F1- score é útil quando se deseja levar em consideração tanto os falsos positivos quanto os falsos negativos.

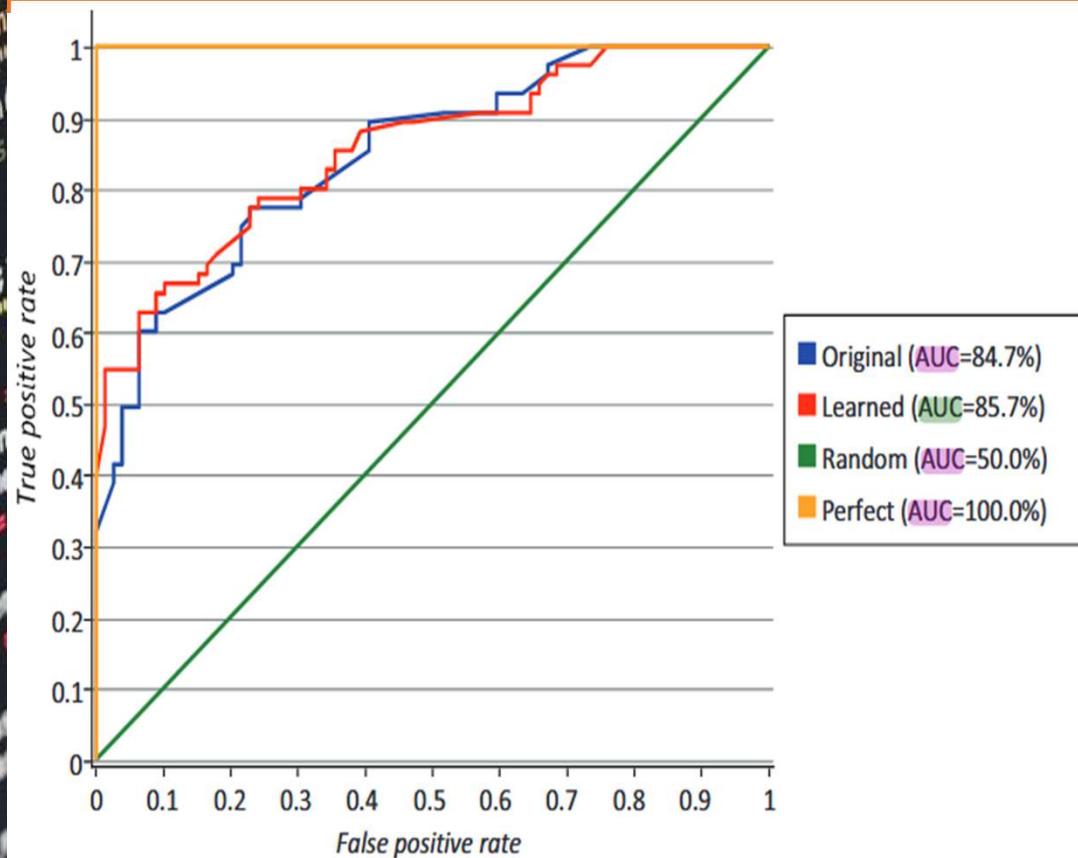
$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Métricas - Curva ROC



- A curva ROC (*Receiver Operating Characteristic*) é outra ferramenta comum usada com classificadores binários.
- A curva ROC mostra o quanto bom o modelo criado pode distinguir entre duas classes.
- A curva ROC representa graficamente a taxa de verdadeiros positivos em relação à taxa de falsos positivos (FPR).

Métricas - AUC



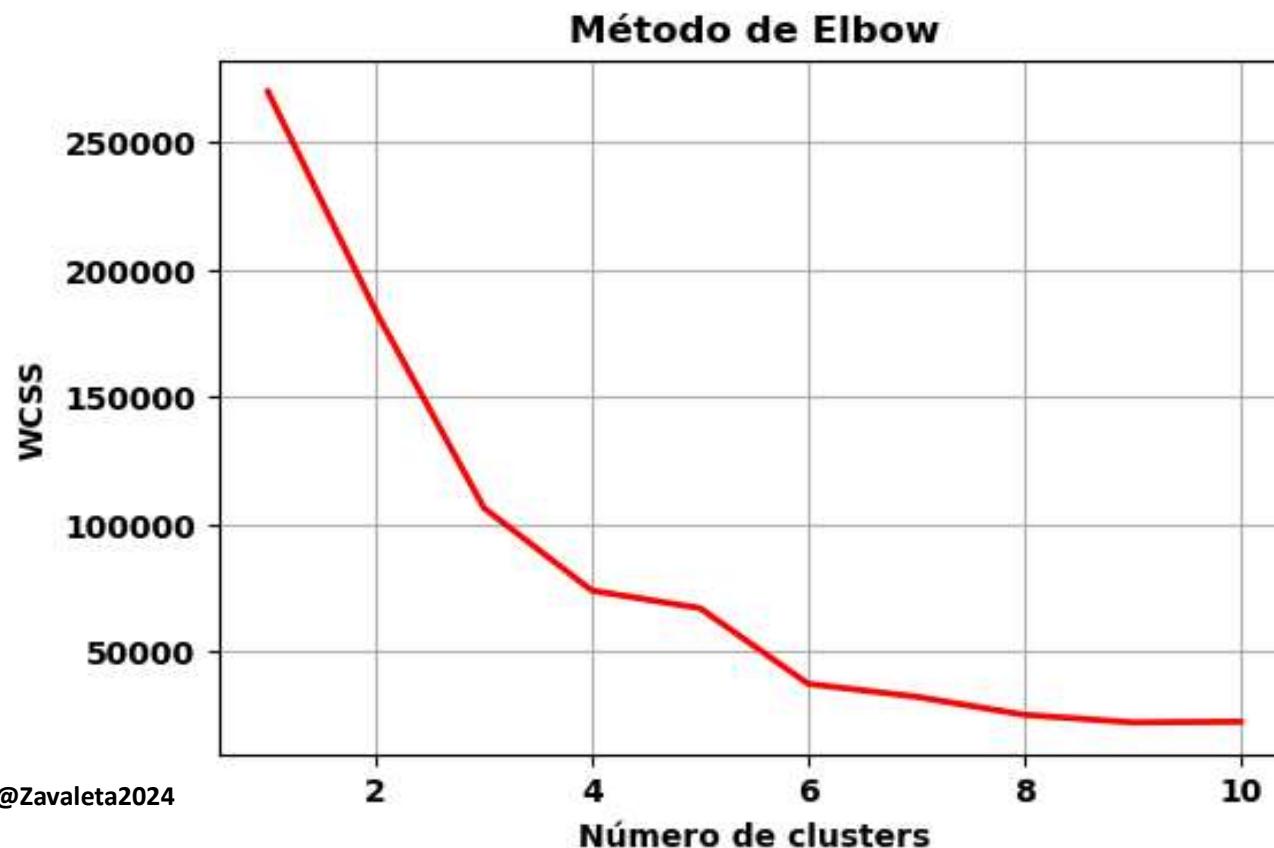
- A AUC (*area under the curve*) é uma medida útil para comparar o desempenho de dois modelos diferentes, desde que os conjuntos de dados estejam mais equilibrados.
- A AUC avalia como o modelo é capaz de distinguir entre as classes em diversos níveis de sensibilidade e especificidade, oferecendo uma visão global da eficácia do modelo.

Clusterização - Escolher o K ideal

- Método **Elbow** ou método **cotovelo** é uma das abordagens comuns para encontrar o valor ideal de k.
 1. Para cada valor k, inicializar k-means e identificar a soma das distâncias quadradas das amostras até o centro do cluster mais próximo.
 2. Traçar um gráfico entre vários valores de k e a soma das distâncias quadradas.
 3. Identificar um ponto no gráfico correspondente a k, além do qual a soma das distâncias quadradas começa a diminuir.
- Este ponto é conhecido como ponto de **cotovelo**, e o valor k é escolhido como o valor ideal de k.

FUNDAMENTOS DE CIÊNCIA DE DADOS

Escolher o K ideal - Elbow



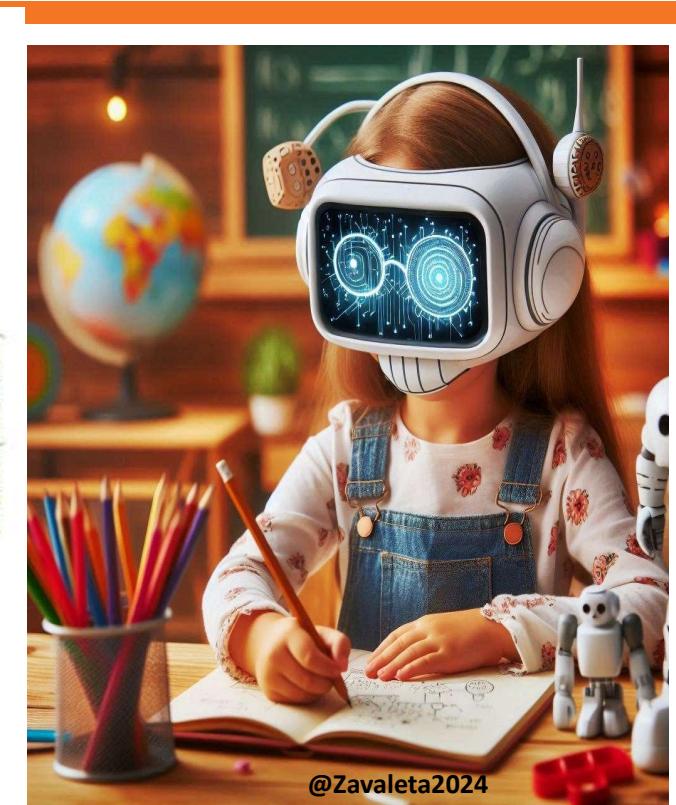
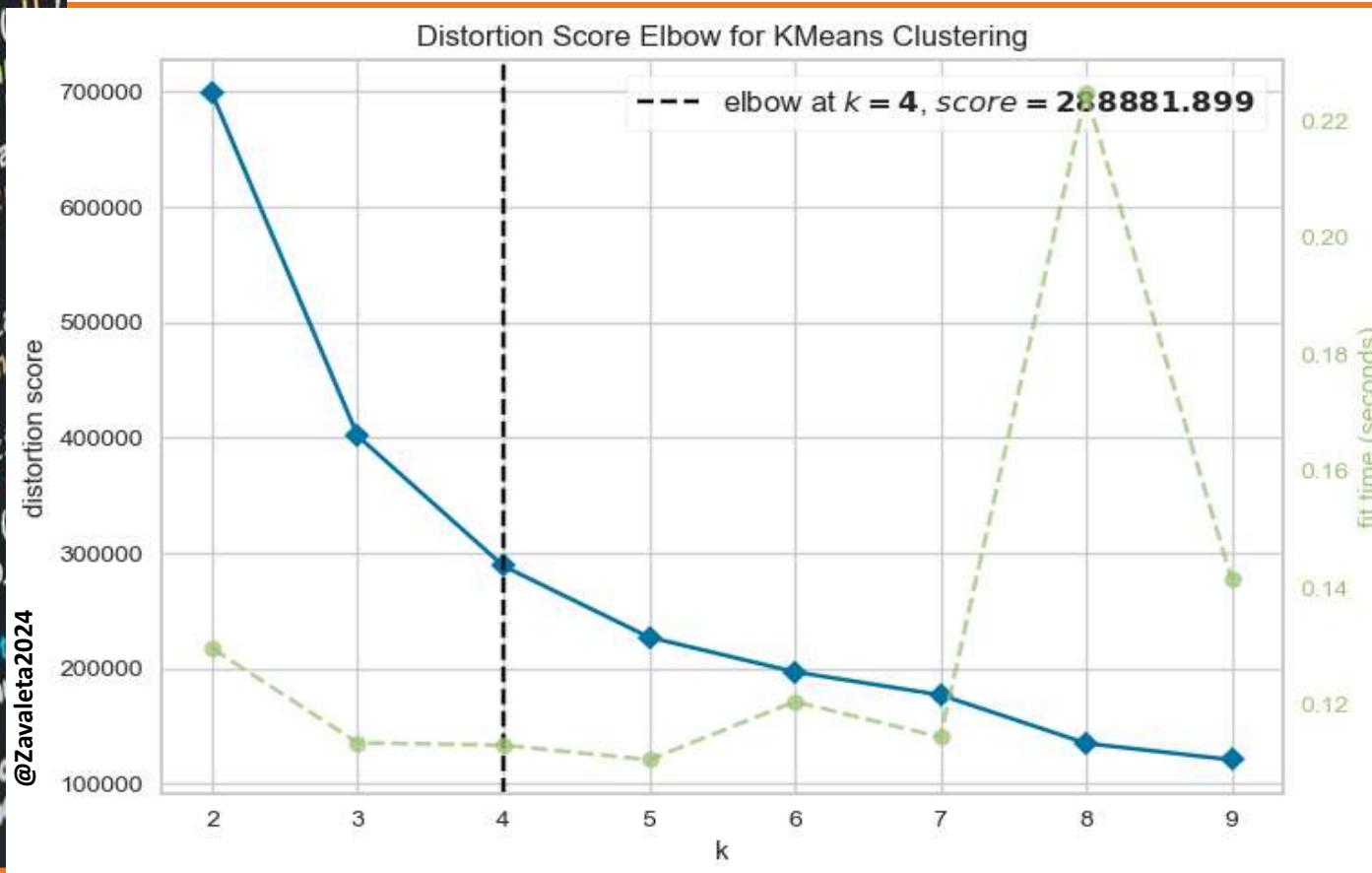
Clusterização - Escolher o K ideal

- Método de **Silhouette** é a melhor abordagem para determinar o número de clusters a serem formulados a partir do dataset.
- Assumir que os dados já foram agrupados em **k** clusters por k-means
- Com as informações disponíveis sobre os clusters, o coeficiente de silhueta $s(i)$ é dado conforme a equação

$$s(i) = \frac{x(i) - y(i)}{\max(x(i), y(i))}$$

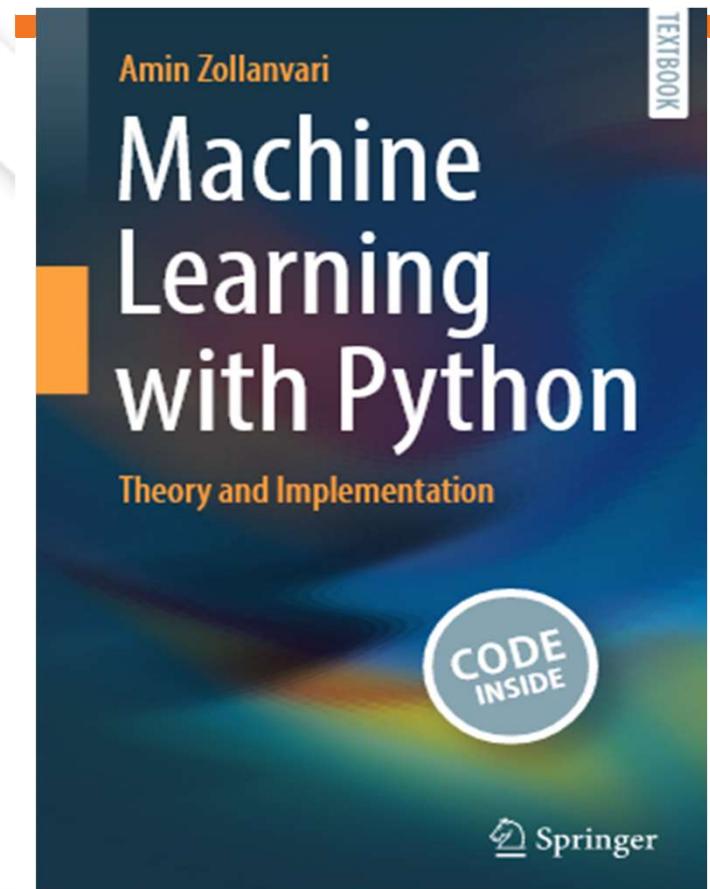
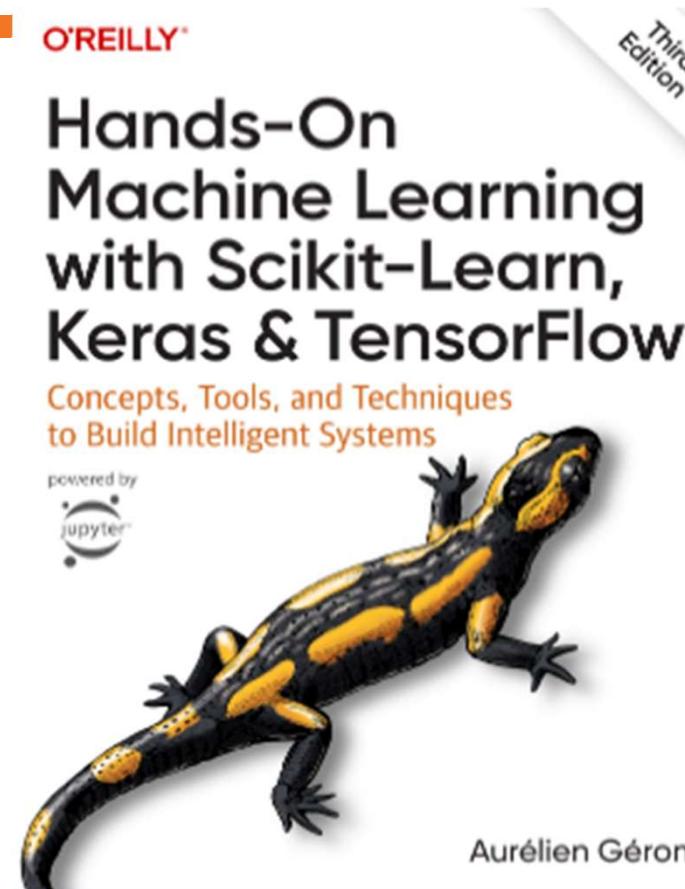
FUNDAMENTOS DE CIÊNCIA DE DADOS

Escolher o K ideal - Silhouette



FUNDAMENTOS DE CIÊNCIA DE DADOS

Referências



Prof. Dr. Jorge Zavaleta

zavaleta@pet-si.ufrj.br

FUNDAMENTOS DE CIÊNCIA DE DADOS



Hands on...

NOTEBOOKS:

- MACHINE_LEARNING