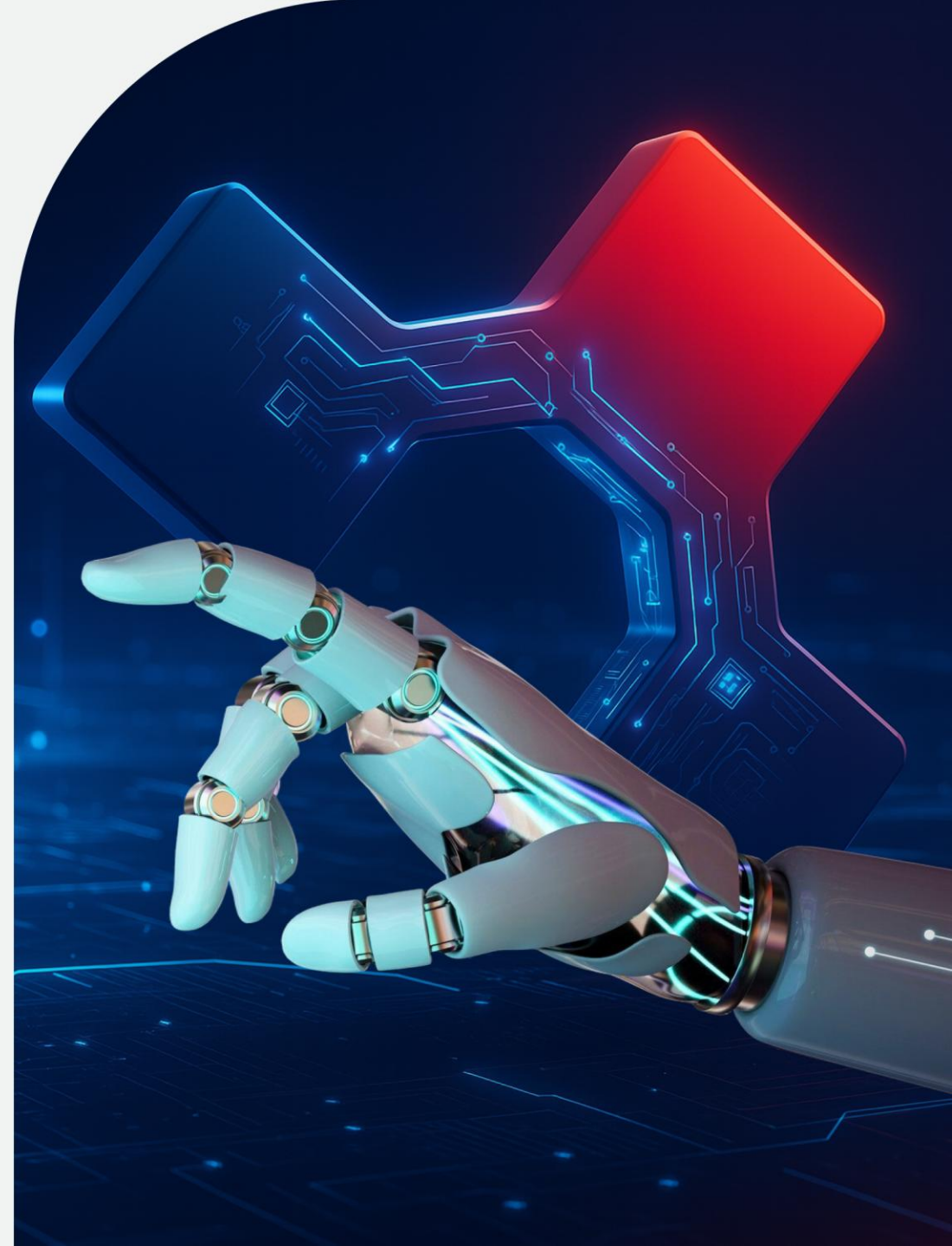
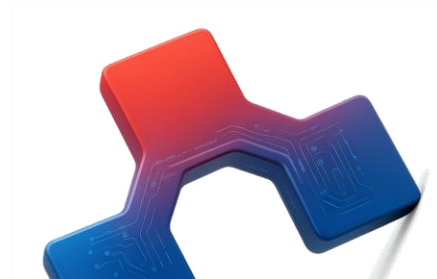




Núcleo de Capacitação em Inteligência Artificial







Projeto de Machine Learning de ponta a ponta

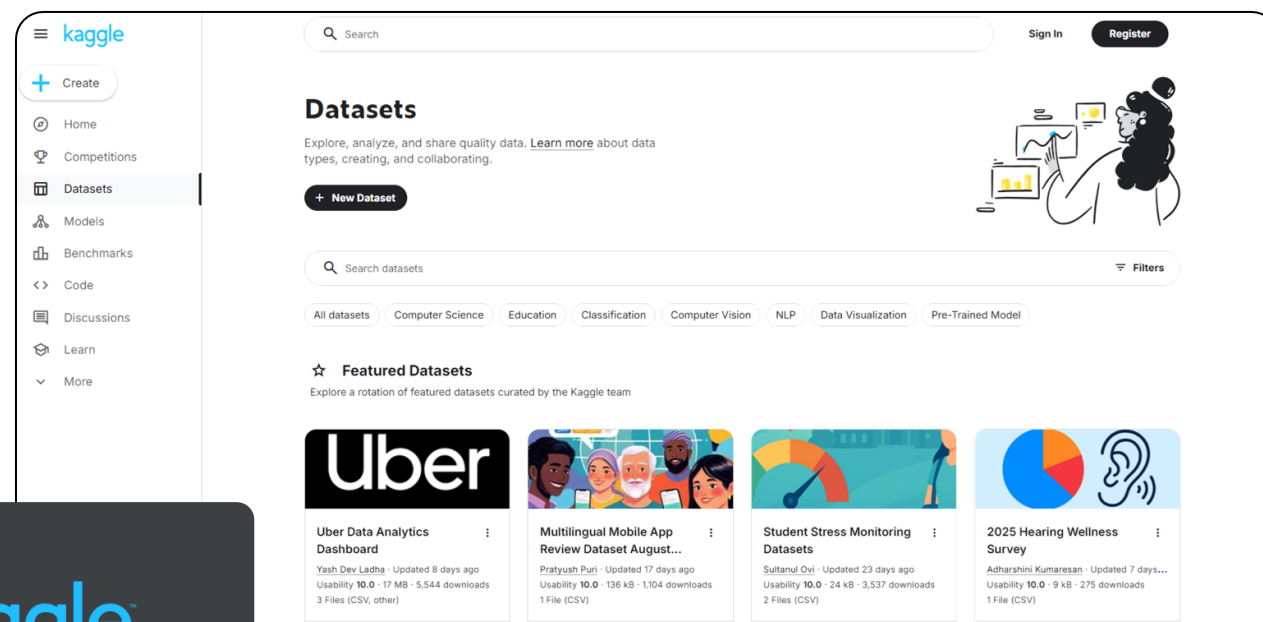
Look at the Big Picture, Get the Data, Explore and Visualize the Data to Gain Insights, Prepare the Data for Machine Learning Algorithms, Select and Train a Model, Fine-Tune Your Model, Launch, Monitor, and Maintain Your System



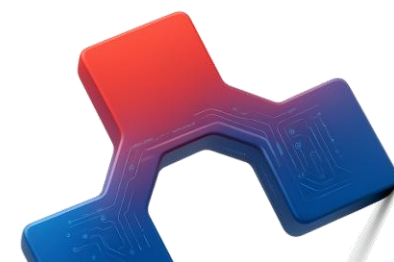
Principais plataformas de dados abertos

1. Kaggle

- Um hub central de datasets para ML e Data Science.
- Forte pela comunidade, competições e notebooks compartilhados.
- Abrange praticamente todas as áreas (texto, imagem, áudio, tabular, séries temporais).



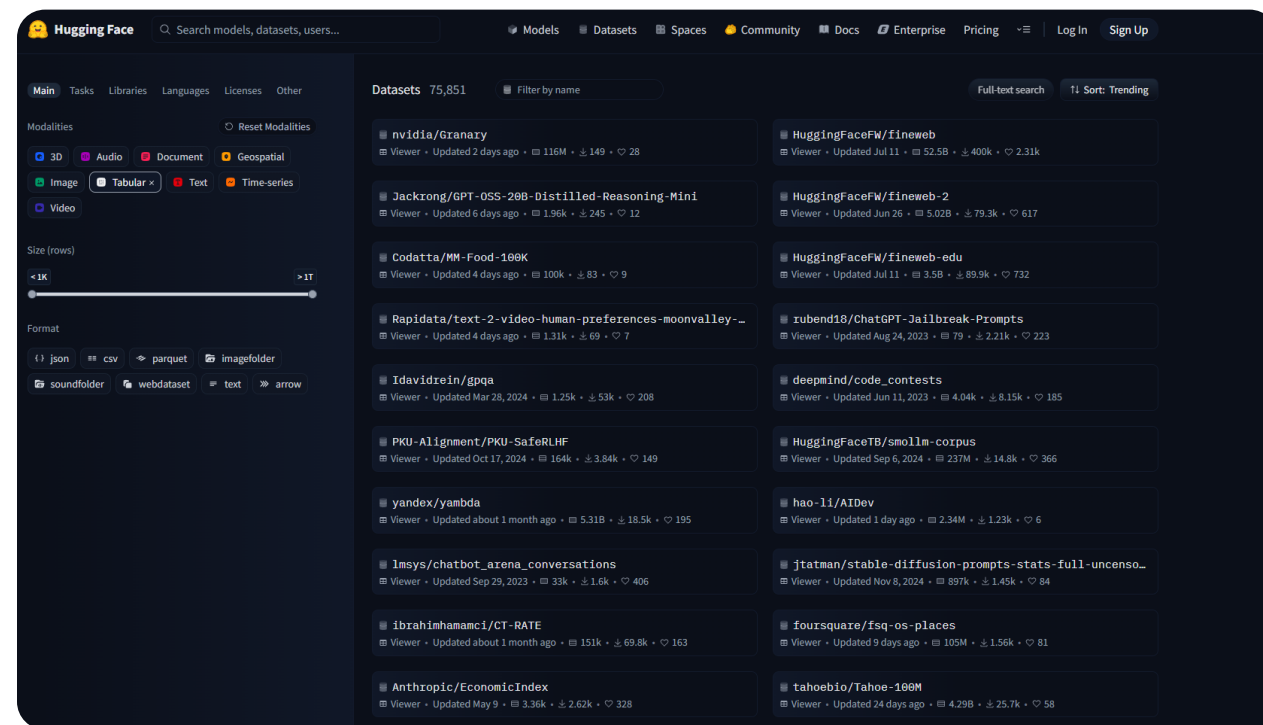
kaggle



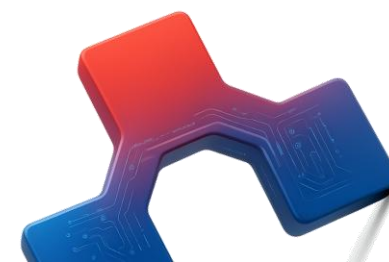
Principais plataformas de dados abertos

2. Hugging Face Datasets

- Hoje é o padrão de fato para quem trabalha com IA moderna.
- Enorme biblioteca de datasets (texto, imagens, áudio, multimodal).
- Integração direta via código (datasets library).



Hugging Face

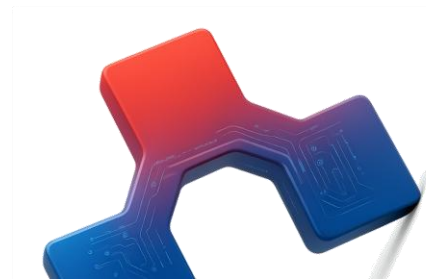
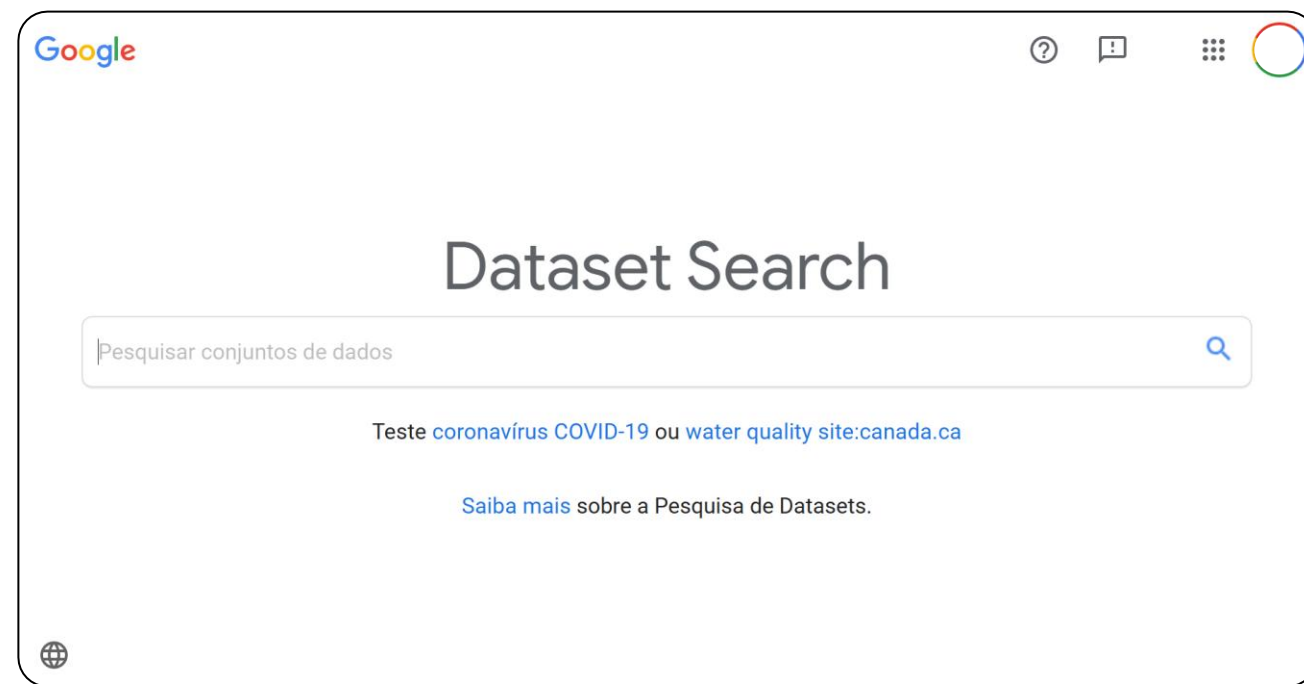




Principais plataformas de dados abertos

3. Google Dataset Search

- O meta-buscador mais eficiente para encontrar datasets em múltiplas fontes (acadêmicas, governamentais, científicas).
- Essencial quando você precisa de algo muito específico ou raro.



Principais plataformas de dados abertos

4. Registry of Open Data on AWS (RODA)

- Fonte para grandes datasets (ex.: Common Crawl, NASA, genômica, satélites).
- Excelente para Big Data e IA em nuvem.

Registry of Open Data on AWS



About

This registry exists to help people discover and share datasets that are available via AWS resources. See [recent additions](#) and [learn more about sharing data on AWS](#).

Get started using data quickly by viewing [all tutorials with associated SageMaker Studio Lab notebooks](#).

See [all usage examples for datasets listed in this registry](#).

See datasets from [EPA](#), [Allen Institute for Artificial Intelligence \(AI2\)](#), [Digital Earth Africa](#), [Data for Good at Meta](#), [NASA Space Act Agreement](#), [NIH STRIDES](#), [NOAA Open Data Dissemination Program](#), [Space Telescope Science Institute](#), and [Amazon Sustainability Data Initiative](#).

Search datasets (currently 788 matching datasets)

Add to this registry

If you want to add a dataset or example of how to use a dataset to this registry, please follow the instructions on the [Registry of Open Data on AWS GitHub repository](#).

Unless specifically stated in the applicable dataset documentation, datasets available through the Registry of Open Data on AWS are not provided and maintained by AWS. Datasets are provided and maintained by a variety of third parties under a variety of licenses. Please check dataset licenses and related documentation to determine if a dataset may be used for your application.

please [tell us about it](#). We may [log post](#).

The Human Sleep Project

[bioinformatics](#) [deep learning](#) [life sciences](#) [machine learning](#) [medicine](#) [neurophysiology](#) [neuroscience](#)

The Human Sleep Project (HSP) sleep physiology dataset is a growing collection of clinical polysomnography (PSG) recordings. Beginning with PSG recordings from from ~15K patients evaluated at the Massachusetts General Hospital, the HSP will grow over the coming years to include data from >200K patients, as well as people evaluated outside of the clinical setting. This data is being used to develop CAISR (Complete AI Sleep Report), a collection of deep neural networks, rule-based algorithms, and signal processing approaches designed to provide better-than-human detection of conventional PSG...

[Details →](#)

Usage examples

- [Classification algorithms for predicting sleepiness and sleep apnea severity. Journal of Sleep Research. 2012 Feb;21\(1\):101-12. PMID: PMC3698244.](#) by Eiseman NA, Westover MB, Mietus JE, Thomas RJ, Bianchi MT
- [Automated Scoring of Respiratory Events in Sleep with a Single Effort Belt and Deep Neural Networks. IEEE Transactions on Biomedical Engineering. 2021 Dec 20;PP. doi: 10.1109/TBME.2021.3136753. Epub ahead of print. PMID: PMC9119908.](#) by Nassi TE, Ganglberger W, Sun H, Bucklin AA, Biswal S, van Putten MJAM, et al.
- [The Challenge of Undiagnosed Sleep Apnea in Low-Risk Populations: A Decision Analysis. Military Medicine 2014 Aug;179\(8S\):47-54. PMID: PMC6788752.](#) by Bianchi MT, Hershman S, Bahadoran M, Ferguson M, Westover MB.
- [The sleep and wake electroencephalogram over the lifespan. Neurobiol Aging. 2023 Jan 19;124:60-70. doi: 10.1016/j.neurobiolaging.2023.01.006. Epub ahead of print. PMID: 36739622.](#) by Sun H, Ye E, Paixao L, Ganglberger W, Chu CJ, Zhang C, et al.
- [Algorithm for automatic detection of self-similarity and prediction of residual central respiratory events during continuous positive airway pressure. Sleep. 2021 Apr 9;44\(4\):zsaa215. doi: 10.1093/sleep/zsaa215. PMID: PMC8631077.](#) by Oppersma E, Ganglberger W, Sun H, Thomas RJ*, Westover MB*

[See 37 usage examples →](#)

Common Crawl

[encyclopedic](#) [internet](#) [natural language processing](#) [web archive](#)

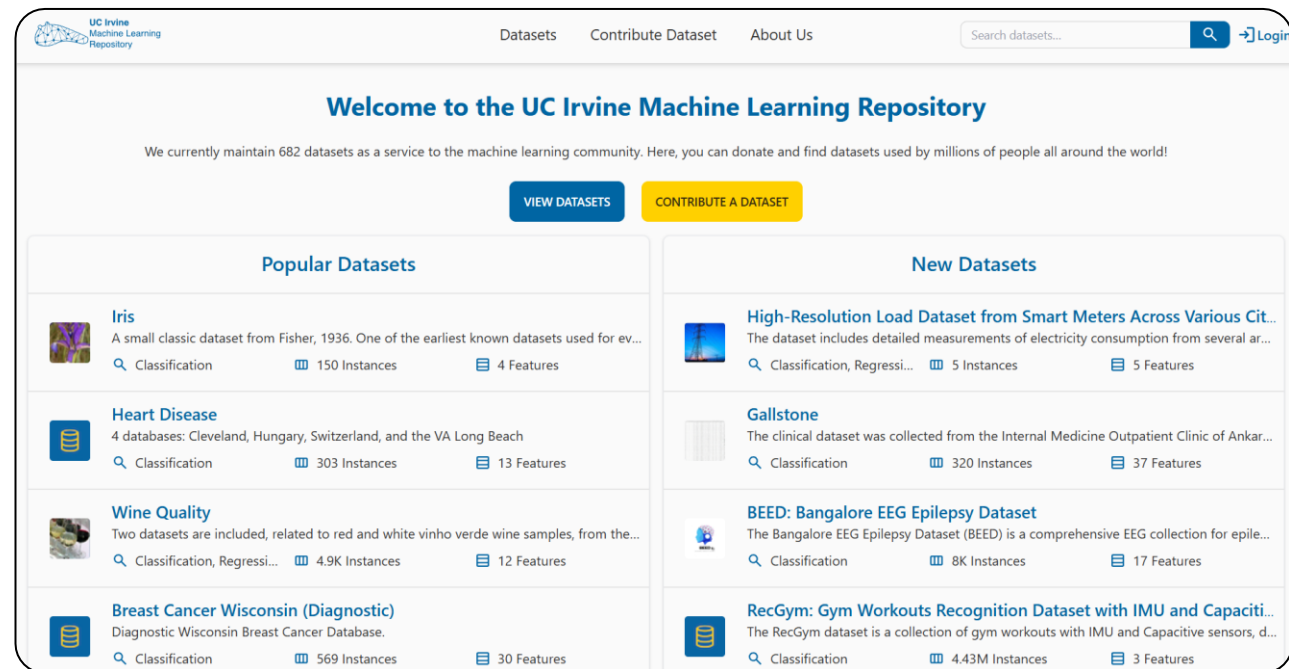
A corpus of web crawl data composed of over 50 billion web pages.



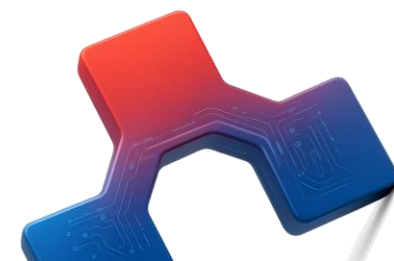
Principais plataformas de dados abertos

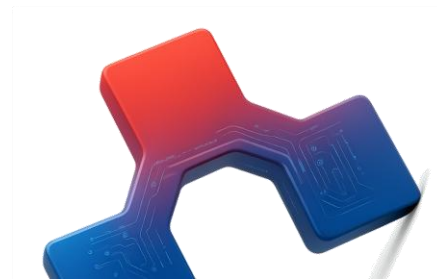
5. UCI Machine Learning Repository

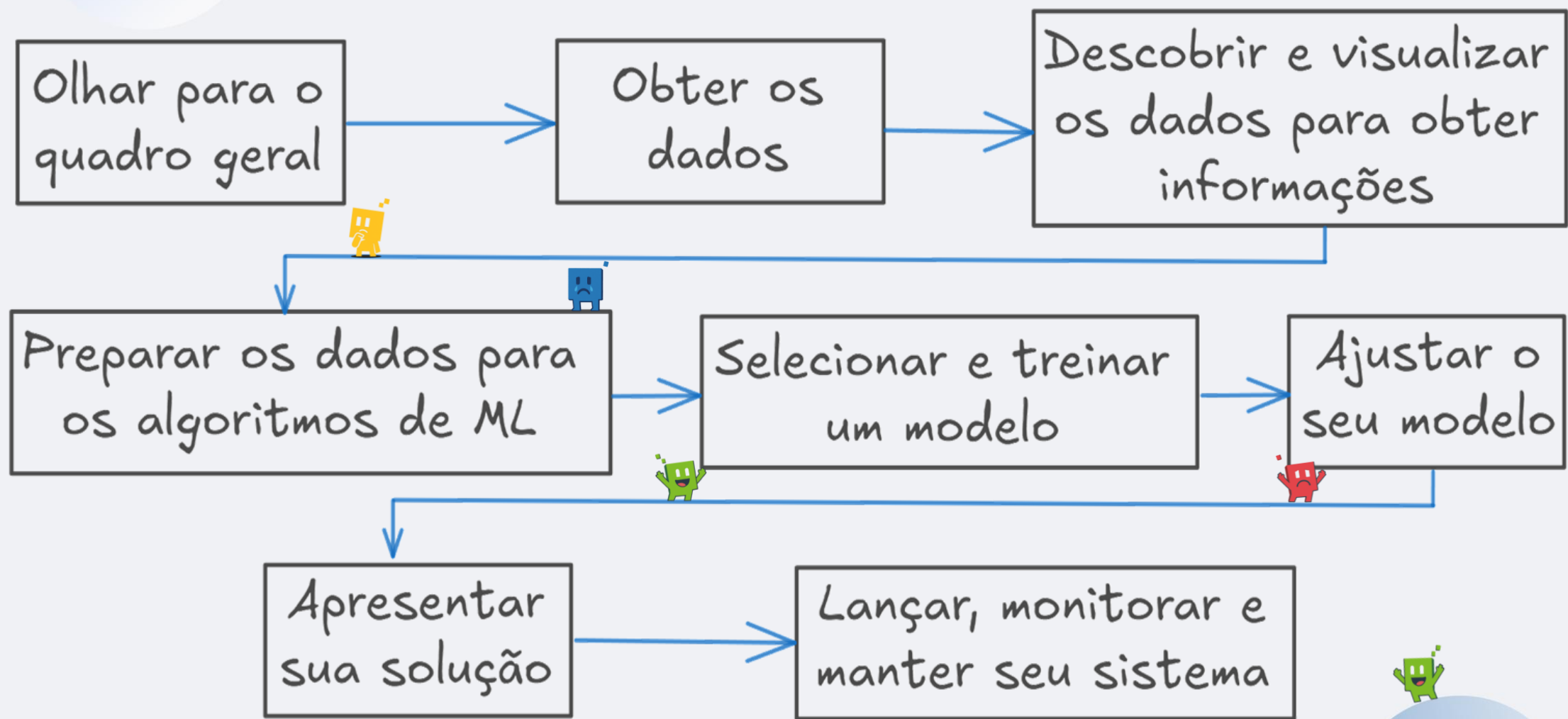
- Clássico, mas ainda útil por datasets curados, limpos e didáticos.
- Ótimo para aprendizado, prototipagem e benchmarks.



UC Irvine
Machine Learning
Repository







Qual o objetivo Comercial?

Qual o objetivo Comercial?

Objetivo do Projeto: Previsão de Preços de Imóveis na Califórnia

Missão: Construir um modelo de Machine Learning para prever o valor mediano de imóveis.

Fonte de Dados: Censo da Califórnia, com dados demográficos e econômicos por "distrito" (quarteirão).

Entrada do Modelo (Features): Métricas como população, renda mediana, etc.

Saída do Modelo (Alvo): A previsão do preço mediano de um imóvel em qualquer distrito

Qual o objetivo Comercial?

Em que Paradigma de treinamento o sistema se enquadra?

Laboratório de Paradigmas

Problemas de negócio

Qual é o problema de negócio a ser resolvido?

Reduzir a perda de clientes (churn)

Otimizar o estoque

Detectar fraudes

Personalizar recomendações

Como o sucesso do projeto será medido em termos de negócio?

Aumento de X% na receita

redução de Y% nos custos operacionais

melhoria de Z% na satisfação do cliente (NPS)

Monitoramento de performance

Quem são os stakeholders e os usuários finais do sistema?

Analistas de marketing

Gerente de logística

Cliente final

Qual é a solução atual (se houver) e quais são suas limitações?

Processo Manual

Baseado em regras simples

Já utiliza algum modelo mas não performa bem

Sobre os dados

Quais são as fontes de dados disponíveis?

Bancos de dados
internos (SQL, NoSQL)

Planilhas

APIs externas

logs de sistema

Qual é o volume de dados disponível?

Pequeno (cabe na
memória de uma máquina)

Médio (requer
processamento
distribuído, como Spark)

Grande (Big Data)

Como os dados serão coletados e atualizados?

Conjunto de dados estático
(exportação única)

Há um fluxo contínuo de
novos dados (streaming)

Qual é a qualidade e o estado dos dados?

Dados faltantes

Ruído

Inconsistências

Outliers

Modelagem e Paradigmas de Aprendizado

Qual é o tipo de supervisão de treinamento

Supervisionado: Temos dados com rótulos (respostas) corretos

Não Supervisionado: Não temos rótulos; o objetivo é encontrar estrutura nos dados

Semissupervisionado: Temos uma grande quantidade de dados sem rótulos e poucos dados rotulados.

Autossupervisionado (Self-Supervised): Os rótulos são gerados a partir dos próprios dados (ex: prever a próxima palavra em uma frase)

Por Reforço (Reinforcement Learning): O modelo aprende por tentativa e erro, recebendo recompensas ou punições.

Qual é a tarefa principal do modelo?

Regressão: Prever um valor numérico contínuo (ex: preço, idade, temperatura)

Classificação: Prever uma categoria ou classe

Clusterização (Agrupamento): Agrupar dados semelhantes sem conhecimento prévio dos grupos (ex: segmentação de clientes)

Deteção de Anomalias: Identificar pontos de dados que fogem do padrão normal (ex: deteção de falhas em equipamentos)

Redução de Dimensionalidade: Reduzir o número de variáveis (features) mantendo a informação mais importante

Sistema de Recomendação: Sugerir itens relevantes para os usuários

Qual técnica de treinamento será utilizada

Em Lote (Batch Learning): O modelo é treinado de uma só vez com todos os dados disponíveis.

Online (Online Learning): O modelo é treinado de forma incremental, com mini-lotes de dados.

Aprendizado por Transferência (Transfer Learning): Usar um modelo pré-treinado em uma tarefa semelhante como ponto de partida

Avaliação de Desempenho

Quais métricas de performance serão usadas para avaliar o modelo?

Regressão

- Erro Quadrático Médio (MSE)
- Raiz do Erro Quadrático Médio (RMSE)
- Erro Absoluto Médio (MAE), R^2

Classificação

- Acurácia
- Precisão
- Recall (Sensibilidade)
- F1-Score
- Curva ROC (AUC)
- Matriz de Confusão

Clusterização

- Coeficiente de Silhueta
- Índice de Davies-Bouldin

Qual será a estratégia de validação do modelo

Divisão simples em
treino/teste

Validação Cruzada
(Cross-Validation)

Validação Cruzada
Estratificada
(para dados esbalanceados)

Qual é a linha de base (baseline) de performance a ser superada?

Um modelo muito simples

Desempenho da solução atual

Produção e Implantação (MLOps)

Como o modelo fará as previsões?

Em Lote (Batch Prediction):
O modelo processa um grande volume de dados de forma agendada

Em Tempo Real (Real-time Prediction):
O modelo está disponível via uma API para fazer previsões sob demanda

Qual é a infraestrutura necessária para a implantação?

Servidores locais
(on-premise)

Nuvem (AWS, GCP, Azure)

Containers
(Docker, Kubernetes)

Como o modelo será monitorado em produção?

Monitoramento de performance
(degradação do modelo)

Desvio de dados
(data drift)

Consumo de recursos
computacionais

Com que frequência o modelo precisará ser retreinado?

Diariamente

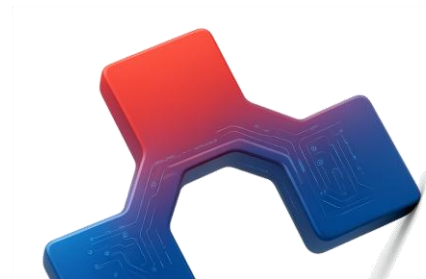
Semanalmente

Quando a performance cair
abaixo de um limiar



As Limitações do Processo Atual

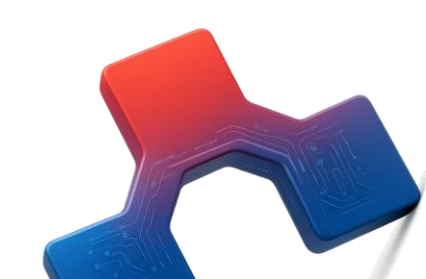
O método de estimativa manual feito por especialistas, embora bem-intencionado, apresenta falhas críticas. Ele é extremamente **caro e demorado**, consumindo muitos recursos. Mais preocupante ainda é sua **baixa precisão**: em muitos casos, foi verificado que as estimativas manuais tinham um **erro superior a 30%** quando comparadas aos valores reais. Um erro dessa magnitude representa um risco financeiro significativo para a empresa.





Oportunidades com Machine Learning

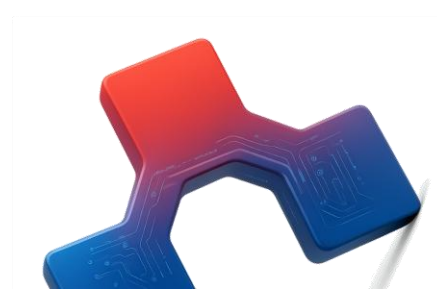
Diante dessas limitações, a empresa vê uma grande oportunidade no uso de Machine Learning. A proposta é treinar um **modelo** para **automatizar e otimizar essa tarefa**. Acreditamos que um modelo bem treinado pode fornecer **previsões muito mais precisas, rápidas e consistentes**, permitindo que a empresa tome **decisões de investimento** mais seguras e em maior escala.





Por que usar os dados do Censo?

O conjunto de dados do censo da Califórnia é a ferramenta perfeita para este desafio. Ele é robusto e contém exatamente o que precisamos: o preço mediano dos imóveis (nossa variável alvo) para milhares de distritos, além de dezenas de outras métricas (nossas *features*), como população e renda, que podem ser usadas para ensinar o modelo a encontrar os padrões.

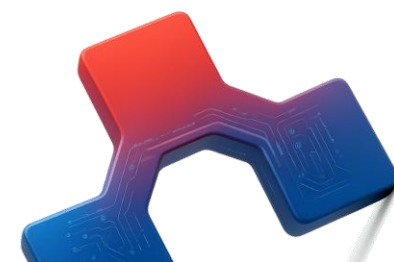




Select a Performance Measure

Para avaliar nosso modelo, precisamos de uma métrica de **performance**. Em problemas de regressão, uma escolha possível é o **Erro Médio Absoluto (MAE)**. Essa métrica calcula o "**desvio padrão**" dos erros de previsão. Sua principal é penalizar o modelo calculando o erro gerado entre a previsão e o rótulo.

$$\text{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^m |h(\mathbf{x}^{(i)}) - y^{(i)}|$$

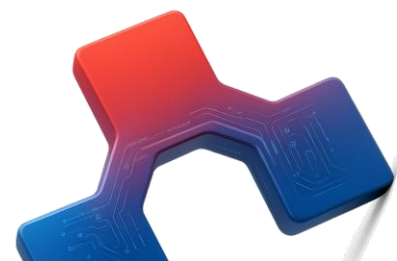




Select a Performance Measure

Para avaliar nosso modelo, precisamos de uma métrica de **performance**. Em problemas de regressão, a escolha mais comum é o **Raiz do Erro Quadrático Médio (RMSE)**. Essa métrica calcula o "**desvio padrão**" dos erros de previsão. Sua principal característica é dar um **peso muito maior aos erros grandes**, o que é geralmente desejável, pois queremos evitar previsões muito distantes da realidade.

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m \left(h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}$$



Na regressão linear , a função de custo é o erro quadrático médio definido pela seguinte igualdade :



$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m \left(h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}$$

Na regressão linear , a função de custo é o erro quadrático médio definido pela seguinte igualdade :

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m \left(h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}$$

Função do erro
quadrático médio

Na regressão linear , a função de custo é o **erro quadrático médio** definido pela seguinte igualdade :

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m \left(h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}$$

Função do erro
quadrático médio

Número de
amostras

Na regressão linear , a função de custo é o **erro quadrático médio** definido pela seguinte igualdade :

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m \left(h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}$$

Função do erro
quadrático médio

Número de
amostras

Na regressão linear , a função de custo é o **erro quadrático médio** definido pela seguinte igualdade :

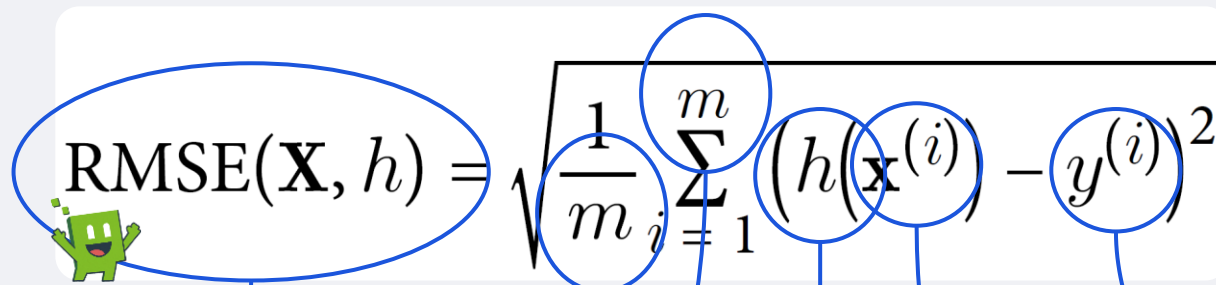
$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(\mathbf{x}^{(i)}) - y^{(i)})^2}$$

Função do erro
quadrático médio

Número de
amostras

Previsão do modelo

Na regressão linear , a função de custo é o **erro quadrático médio** definido pela seguinte igualdade :



The diagram shows the formula for the Root Mean Square Error (RMSE) with several parts circled in blue and arrows pointing to labels. The formula is:
$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(\mathbf{x}^{(i)}) - y^{(i)})^2}$$
 Annotations:

- A green robot icon points to the $\text{RMSE}(\mathbf{X}, h)$ term.
- The $\frac{1}{m}$ term is circled, with an arrow pointing to "Número de amostras".
- The summation symbol \sum is circled, with an arrow pointing to "Número de amostras".
- The upper limit m of the summation is circled, with an arrow pointing to "Número de amostras".
- The term $h(\mathbf{x}^{(i)})$ is circled, with an arrow pointing to "Previsão do modelo".
- The term $\mathbf{x}^{(i)}$ is circled, with an arrow pointing to "Features do exemplo de número i".
- The term $y^{(i)}$ is circled, with an arrow pointing to "Ground Truth".

Função do erro quadrático médio

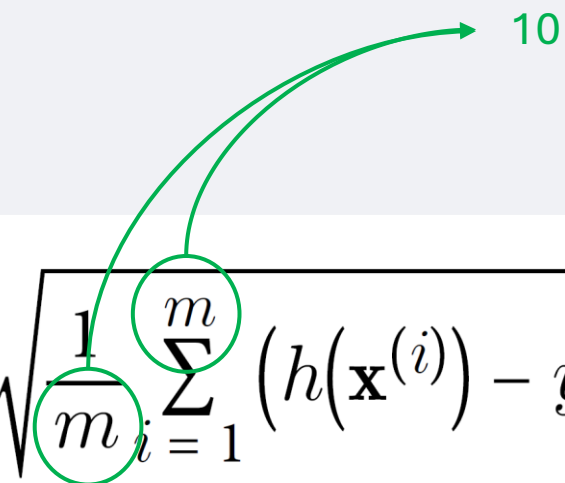
Número de amostras

Previsão do modelo

Features do exemplo de número i

Ground Truth

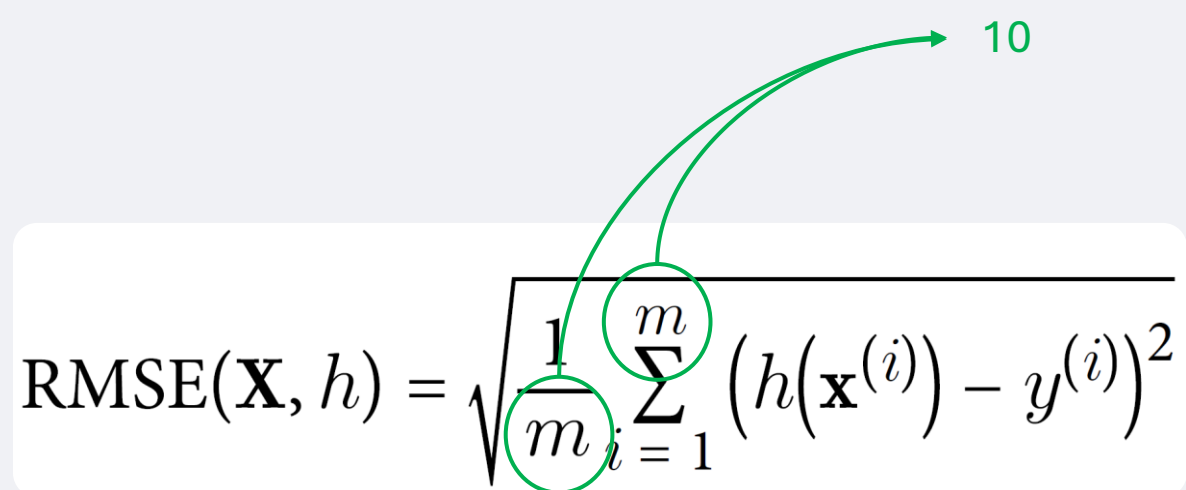
Se estivermos trabalhando com um universo de 10 bairros:



The diagram shows a green arrow pointing from the number 10 to the variable m in the RMSE formula. The variable m is circled in green, and the number 10 is also circled in green. The formula is:

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m \left(h(\mathbf{x}^{(i)}) - y^{(i)} \right)^2}$$

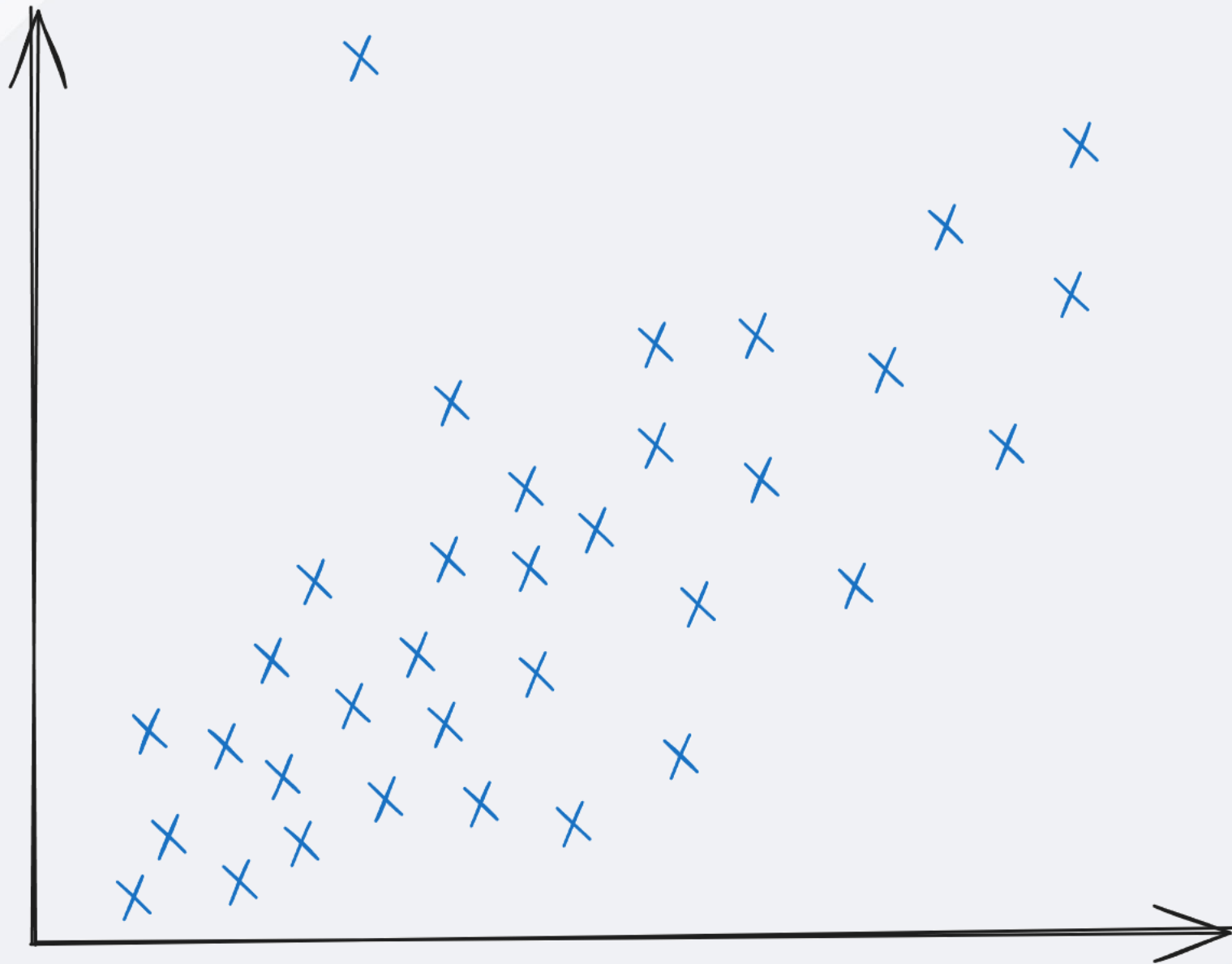
h é a função de previsão do sistema, também conhecida como função hipótese. Quando o sistema recebe o vetor de características $x^{(i)}$ tal que $y'^{(i)} = h(x^{(i)})$

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(\mathbf{x}^{(i)}) - y^{(i)})^2}$$


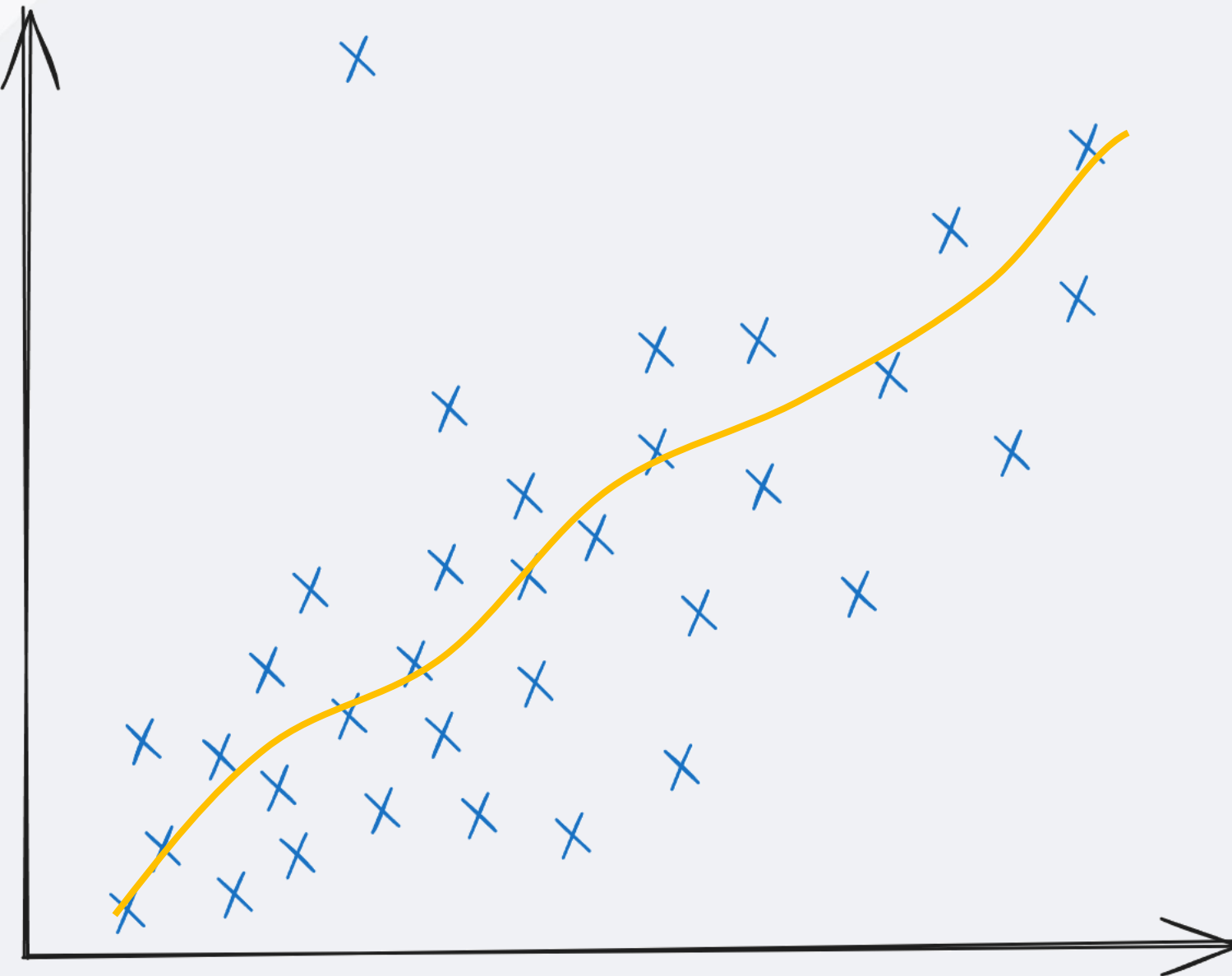
$x^{(i)}$: Vetor com todos os valores de características da i -ésima instância do conjunto de dados e $y^{(i)}$ é o rótulo, valor desejado na saída para aquela i -ésima instância

i	$y^{(i)}$	$h(\mathbf{x}^{(i)})$	$h(\mathbf{x}^{(i)}) - y^{(i)}$	$(h(\mathbf{x}^{(i)}) - y^{(i)})^2$	$\text{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^m h(\mathbf{x}^{(i)}) - y^{(i)}$	$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(\mathbf{x}^{(i)}) - y^{(i)})^2}$

i	$y^{(i)}$	$h(\mathbf{x}^{(i)})$	$h(\mathbf{x}^{(i)}) - y^{(i)}$	$(h(\mathbf{x}^{(i)}) - y^{(i)})^2$	$\text{MAE}(\mathbf{X}, h) = \frac{1}{m} \sum_{i=1}^m h(\mathbf{x}^{(i)}) - y^{(i)}$	$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m} \sum_{i=1}^m (h(\mathbf{x}^{(i)}) - y^{(i)})^2}$
1	6	7				
2	16	14				
3	8	8				
4	2	5				
5	20	17				
6	12	15				
7	12	11				
8	10	10				
9	10	10				



MAE



RMSE

