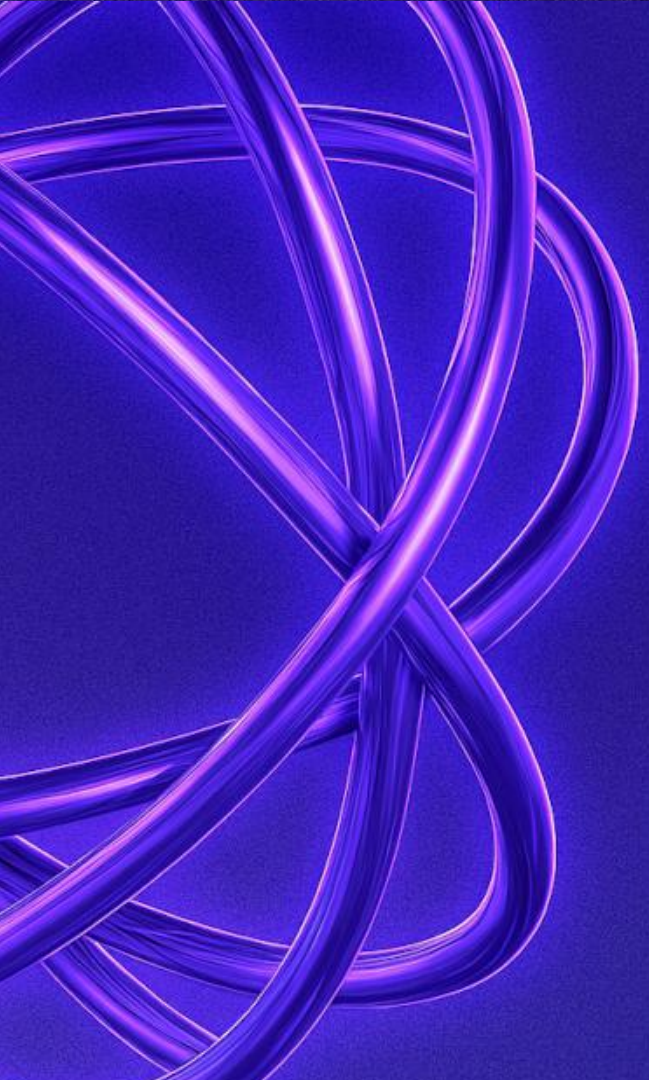


DataHackers | STATE OF DATA BRAZIL 2025

AULA 04 – ML

ANÁLISE COM MACHINE LEARNING

NA ÚLTIMA AULA VOCÊ APRENDEU

AGENDA

- ✓ Como fazer gráficos lindões e interativos com Plotly!
- ✓ Como criar um Data Product utilizando Streamlit!

AO FIM DESSA AULA VOCÊ VAI SABER

AGENDA

- ✓ Descrever o que é machine learning!
- ✓ Enumerar diferentes tipos de machine learning
- ✓ Criar um modelo de regrassão para ajuda na sua análise!
- ✓ Integrar um modelo de ML num data product com Streamlitnálise!

MACHINE LEARNING

CONCEITOS FUNDAMENTAIS

- ✓ Afinal, o que é Machine Learning?
- ✓ Como programas de machine learning são diferentes de programas "normais"?
- ✓ "Algoritmos de machine learning constroem modelos baseado em amostras de dados, conhecidos como 'dados de treinamento', a fim de fazer previsões ou decisões sem serem explicitamente programados para isso"*

MACHINE LEARNING

CONCEITOS FUNDAMENTAIS

MACHINE LEARNING

CONCEITOS FUNDAMENTAIS

- ✓ EXPLICITAMENTE programado com as características que EU identifiquei e apresentei para o algoritmo:

```
Se olho igual "biloca" preta:
   escreva("é cachorro!!")
Senão se olho igual "biloca" azul com um "rasgo":
   escreva("é gato!!")
Senão:
   escreva("sei que bicho é esse não!!")
```

MACHINE LEARNING

CONCEITOS FUNDAMENTAIS

letsdata

IA E APLICAÇÕES

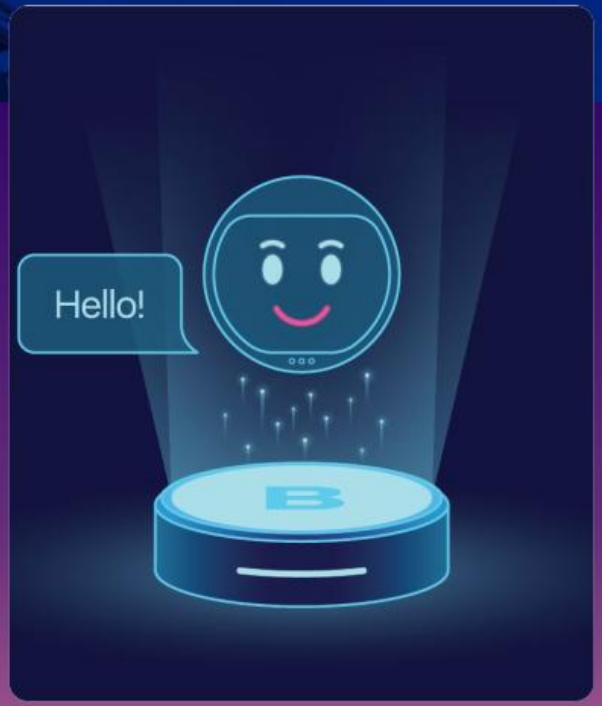
MACHINE LEARNING

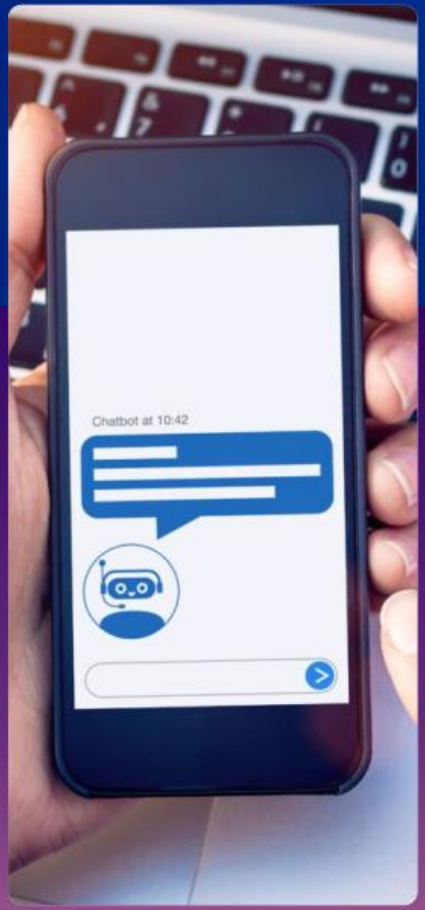
- ✓ Algoritmos de machine learning DESCOBREM os padrões por si mesmos!!
- ✓ Percebam que é uma quebra no paradigma e onde a revolução realmente acontece
- ✓ Portanto, todo algoritmo que não foi explicitamente programado para descobrir padrões, mas o faz por si só, pode ser considerado machine learning

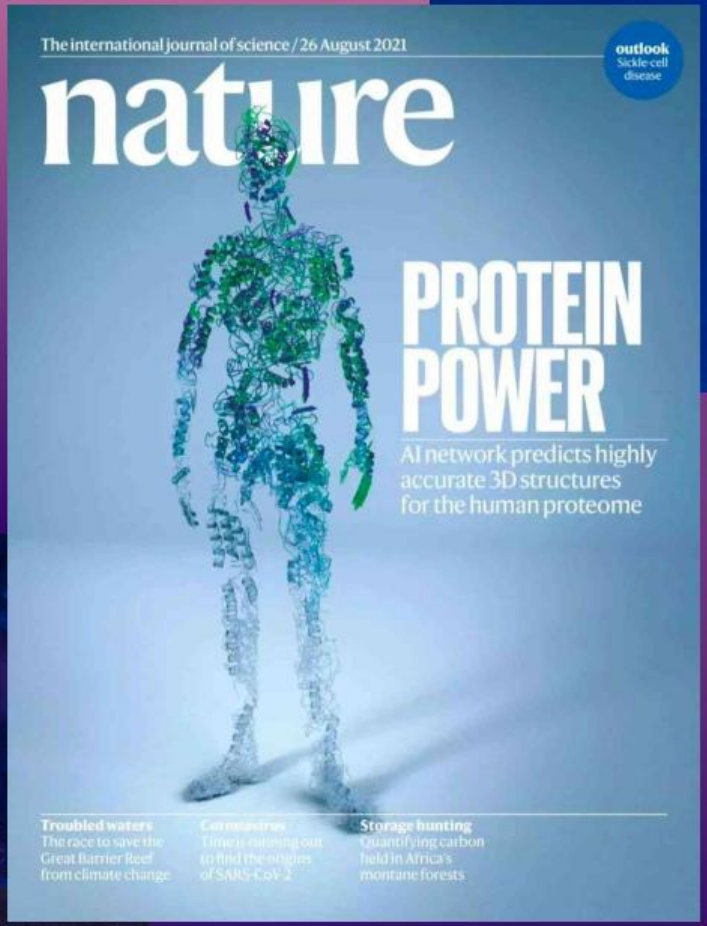
letsdata

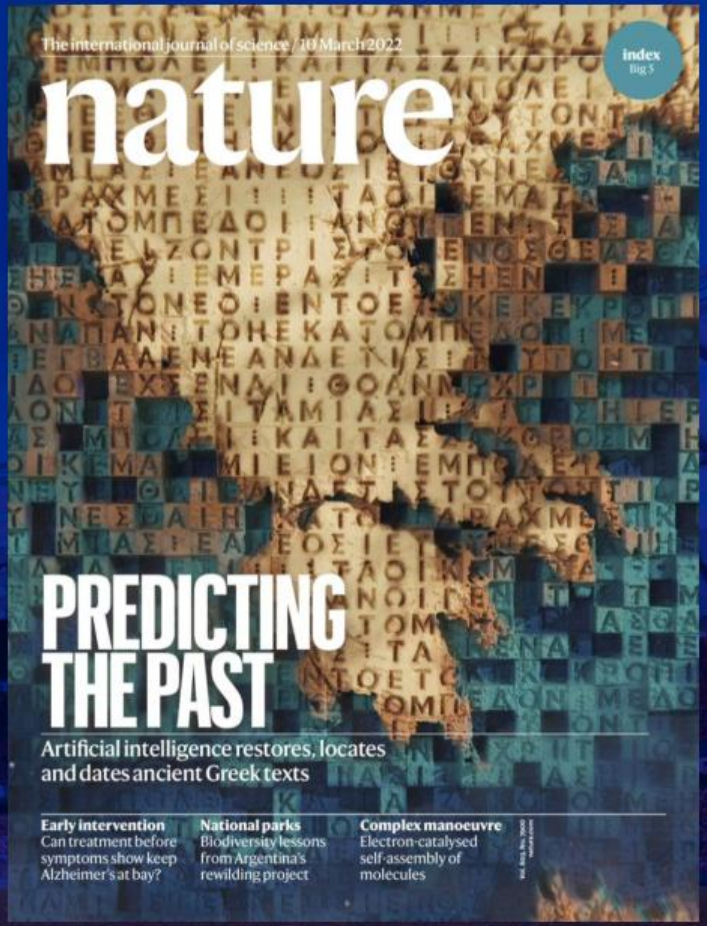
Veículos autônomos

Assistentes virtuais






The international journal of science / 26 August 2021
outlook
Sickle-cell disease
nature
PROTEIN POWER
AI network predicts highly accurate 3D structures for the human proteome
Troubled waters
The race to save the Great Barrier Reef from climate change
Coronavirus
Time is running out to find the origins of SARS-CoV-2
Storage hunting
Quantifying carbon held in Africa's montane forests


The international journal of science / 10 March 2022
index
Big 5
nature
PREDICTING THE PAST
Artificial intelligence restores, locates and dates ancient Greek texts
Early intervention
Can treatment before symptoms show keep Alzheimer's at bay?
National parks
Biodiversity lessons from Argentina's rewilding project
Complex manoeuvre
Electron-catalysed self-assembly of molecules

Sistemas de recomendação

Clientes

Análise de churn

Logística

Previsão de estoque

Saúde

Classificação de Raio-X

APRENDIZADO SUPERVISIONADO

CONCEITOS FUNDAMENTAIS

- ✓ Basicamente temos três tipos de aprendizado em machine learning
- ✓ Aprendizado supervisionado
- ✓ Aprendizado não supervisionado
- ✓ Aprendizado por reforço

APRENDIZADO SUPERVISIONADO

CONCEITOS FUNDAMENTAIS

- ✓ Utiliza dados rotulados para aprender!
- ✓ Imagine uma criança aprendendo com seus brinquedos
- ✓ Ela tem à frente carrinhos e caminhões e um adulto explicitamente ensina para ela alguns exemplos

APRENDIZADO SUPERVISIONADO

CONCEITOS FUNDAMENTAIS

- ✓ Nesse exemplo temos:
- ✓ Os brinquedos são os dados
- ✓ As características são: chassi, quantidade de rodas, tamanho de rodas, etc...
- ✓ Os rótulos são: carrinho e caminhão

APRENDIZADO SUPERVISIONADO

CONCEITOS FUNDAMENTAIS

CARRINHO

CAMINHÃO

letsdata

APRENDIZADO SUPERVISIONADO

CONCEITOS FUNDAMENTAIS

- ✓ A partir dos brinquedos à disposição (base de treinamento) a criança aprende as características que diferenciam os dois
- ✓ Com novas amostras (carrinhos na loja) a criança vai saber o que é cada carrinho
- ✓ Se for apresentado um barco: nada feito, ele nunca viu um barco

APRENDIZADO SUPERVISIONADO

CONCEITOS FUNDAMENTAIS

- ✓ Da mesma forma temos nos algoritmos de aprendizado supervisionado
- ✓ Classificar fraudes de transações bancárias conhecidas (com os rótulos de fraude x não fraude)
- ✓ Prever preços de carros a partir de dados de um site de vendas (com o "rótulo" sendo o preço final a partir das características do veículo: ano, estado de conservação, possui ar condicionado...)

APRENDIZADO SUPERVISIONADO

CONCEITOS FUNDAMENTAIS

- ✓ Principais tipos!
- ✓ Classificação: quanto o rótulo, ou seja, a variável que queremos aprender é discreta (ex: se paciente tem ou não diabetes)
- ✓ Regressão: quanto o rótulo é contínuo (ex: preço de carro)

APRENDIZADO SUPERVISIONADO

CONCEITOS FUNDAMENTAIS

- ✓ Algoritmos
- ✓ Baseados em árvore: árvores de decisão, random forests, XGBoost, Lightgbm, etc...
- ✓ Algoritmos lineares: regressão linear, regressão logística, etc...
- ✓ Redes neurais

letsdata

CONCEITOS FUNDAMENTAIS

APRENDIZADO SUPERVISIONADO

- ✓ Prever, prever, prever

https://martech.org/2018-ad-tech-predictions-didnt-expect-see/

letsdata

CONCEITOS FUNDAMENTAIS

APRENDIZADO SUPERVISIONADO

- ✓ Algoritmos de análise supervisionada são utilizados para prever!
- ✓ Como diz o subtítulo do ótimo livro "Análise preditiva" de Eric Siegel, machine learning tem "O Poder de predizer quem vai clicar, comprar, mentir ou morrer"
- ✓ Apesar do exagero, e da especialização em análise preditiva "tabular" (temos imagens, textos, sons), isso resume bastante bem o porquê do aprendizado supervisionado ser muito utilizado e solicitado pelas empresas

Siegel, Eric. *Análise Preditiva: o poder de prever quem vai clicar, comprar, mentir ou morrer.* Alta Books Editora, 2018.

letsdata

CONCEITOS FUNDAMENTAIS

DADOS TABULARES

- ✓ Apesar da aplicação cada vez mais premente de machine learning para imagens, textos e sons, a maior necessidade ainda são para dados tabulares
- ✓ Talvez pelo nível de maturidade das empresas, que não possuem equipes muito especializadas e tem problemas mais "concretos" e imediatos para resolver (nem todo mundo é Meta, Google ou NVIDIA)
- ✓ Previsão de fraudes, de preços, de churn, de turnover, de vendas, de estoque... o mundo real ainda é feito de dados tabulares*

* Não a toa Excel domina o mundo :D

letsdata

CONCEITOS FUNDAMENTAIS

DADOS TABULARES

- ✓ Numa análise supervisionada para caso de dados tabulares, basicamente precisamos de uma base de dados já rodulados, ou seja, com as respostas
- ✓ A partir dessa base de dados, os algoritmos “aprendem” os padrões a ponto de conseguirem prever os resultados para novos dados
- ✓ Temos, portanto, na base de dados, uma separação entre variáveis que são utilizadas para prever o nosso alvo
- ✓ Vamos aos exemplos para ficar mais claro

letsdata

CONCEITOS FUNDAMENTAIS

DADOS TABULARES

- ✓ Imagine uma concessionária que precise prever preços de carros usados para compor junto ao "feeling" dos funcionários mais experientes. Temos uma base rotulada (com os valores dos preços vendidos):

	marca	modelo	idade	quilometragem	combustivel	cambio	consumo	motor	potencia	assentos	preco_venda
0	Maruti	Alto	9	120000	Petrol	Manual	19.70	796	46.30	5	120000
1	Hyundai	Grand	5	20000	Petrol	Manual	18.90	1197	82.00	5	550000
2	Hyundai	i20	11	60000	Petrol	Manual	17.00	1197	80.00	5	215000
3	Maruti	Alto	9	37000	Petrol	Manual	20.92	998	67.10	5	226000
4	Ford	Ecosport	6	30000	Diesel	Manual	22.77	1498	98.59	5	570000

letsdata

CONCEITOS FUNDAMENTAIS

DADOS TABULARES

- ✓ A variável que queremos prever, portanto, é "preco_venda". A variável que queremos prever também pode ser chamada de **alvo, target, variável dependente ou label.**

	marca	modelo	idade	quilometragem	combustivel	cambio	consumo	motor	potencia	assentos	preco_venda
0	Maruti	Alto	9	120000	Petrol	Manual	19.70	796	46.30	5	120000
1	Hyundai	Grand	5	20000	Petrol	Manual	18.90	1197	82.00	5	550000
2	Hyundai	i20	11	60000	Petrol	Manual	17.00	1197	80.00	5	215000
3	Maruti	Alto	9	37000	Petrol	Manual	20.92	998	67.10	5	226000
4	Ford	Ecosport	6	30000	Diesel	Manual	22.77	1498	98.59	5	570000

CONCEITOS FUNDAMENTAIS

DADOS TABULARES

- ✓ As variáveis que vamos utilizar e buscar padrões para prever o preço do carro são todas as outras: marca, modelo, idade, quilometragem, combustível, cambio, consumo, motor, potencia, assentos

	marca	modelo	idade	quilometragem	combustivel	cambio	consumo	motor	potencia	assentos	preco_venda
0	Maruti	Alto	9	120000	Petrol	Manual	19.70	796	46.30	5	120000
1	Hyundai	Grand	5	20000	Petrol	Manual	18.90	1197	82.00	5	550000
2	Hyundai	i20	11	60000	Petrol	Manual	17.00	1197	80.00	5	215000
3	Maruti	Alto	9	37000	Petrol	Manual	20.92	998	67.10	5	226000
4	Ford	Ecosport	6	30000	Diesel	Manual	22.77	1498	98.59	5	570000

letsdata

CONCEITOS FUNDAMENTAIS

DADOS TABULARES

- ✓ As variáveis utilizadas para prever a variável target podem ser chamadas de independentes, preditoras ou features.

	marca	modelo	idade	quilometragem	combustivel	cambio	consumo	motor	potencia	assentos	preco_venda
0	Maruti	Alto	9	120000	Petrol	Manual	19.70	796	46.30	5	120000
1	Hyundai	Grand	5	20000	Petrol	Manual	18.90	1197	82.00	5	550000
2	Hyundai	i20	11	60000	Petrol	Manual	17.00	1197	80.00	5	215000
3	Maruti	Alto	9	37000	Petrol	Manual	20.92	998	67.10	5	226000
4	Ford	Ecosport	6	30000	Diesel	Manual	22.77	1498	98.59	5	570000

letsdata

CONCEITOS FUNDAMENTAIS

DADOS TABULARES

- ✓ Finalmente, seguindo a convenção matemática de funções. As variáveis preditoras (independentes ou features) são chamadas de X enquanto a variável target (dependente ou label) é chamada de y

- ✓ No nosso exemplo portanto:

```
X.head()
```

	marca	modelo	idade	quilometragem	combustivel	cambio	consumo	motor	potencia	assentos
0	Maruti	Alto	9	120000	Petrol	Manual	19.70	796	46.30	5
1	Hyundai	Grand	5	20000	Petrol	Manual	18.90	1197	82.00	5
2	Hyundai	i20	11	60000	Petrol	Manual	17.00	1197	80.00	5
3	Maruti	Alto	9	37000	Petrol	Manual	20.92	998	67.10	5
4	Ford	Ecosport	6	30000	Diesel	Manual	22.77	1498	98.59	5

```
y.head()
```

	preco_venda
0	120000
1	550000
2	215000
3	226000
4	570000

letsdata

SEPARAÇÃO DE BASES

CONCEITOS FUNDAMENTAIS

- ✓ A forma mais utilizada de simular uma condição de dados que nunca foram vistos é a separação dos dados que temos em nossa posse
- ✓ Os dados já rotulados são separadas na base onde o algoritmo será treinado e na base onde o algoritmo será confrontado com a simulação da vida real
- ✓ Tornar o modelo generalizável (santo graal) é fazer com que o modelo consiga aprender na base de treinamento, sem overfit, ou seja, para que consiga ter uma boa performance em dados futuros (ou não vistos)

SEPARAÇÃO DE BASES

CONCEITOS FUNDAMENTAIS

- ✓ Mundo ideal: montão de dados
- ✓ Separação em:
  - ✓ Treino: onde será efetivamente treinado
  - ✓ Validação (ou desenvolvimento): onde sua base será validada, ou seja, os resultados DESSA BASE serão utilizados para saber qual o melhor modelo
  - ✓ Teste: só vai ser utilizado NO FINAL (simulação da realidade)

letsdata

SEPARAÇÃO DE BASES

CONCEITOS FUNDAMENTAIS

- ✓ Treino: CENTRO do modelo supervisionado (aprender com o histórico)
- ✓ Validação (ou desenvolvimento): selecionada uma métrica e vários algoritmos, como vou saber qual o melhor???
- ✓ Teste: um dia, meu modelo vai ver a luz do sol... mas como faço pra simular dados reais de produção para saber se meu modelo campeão vai se dar bem na vera?

SEPARAÇÃO DE BASES

CONCEITOS FUNDAMENTAIS

- ✓ Exemplo: classificação de fraude num sistema bancário
- ✓ Dados: transações rotuladas (labeled) em fraude x não fraude (812.039.709 de linhas, por exemplo)
- ✓ Separação em treino (80%), validação (10%) e teste (10%)

SEPARAÇÃO DE BASES

CONCEITOS FUNDAMENTAIS

- ✓ Treino: criar os modelos utilizando diferentes algoritmosusando esses dados (o modelo aprende com esses dados)
- ✓ Validação: com os modelos treinados lá na base de treino, realizar as predições aqui na validação e coletar as métricas
- ✓ Teste..... (larga o teste pra lá!!!)

SEPARAÇÃO DE BASES

CONCEITOS FUNDAMENTAIS

- ✓ Validação: com as métricas coletadas, escolher o melhor modelo!
- ✓ Escolha do modelo: e o campeão foi.... XGBoost (os kagglers adoram)!!! Hooray, XG, XG, XG...
- ✓ "Calma lá!!! Com o modelo escolhido, como vou saber se ele vai funfar direito em produção???"
- ✓ A resposta: você NUNCA saberá por antecipação

letsdata

SEPARAÇÃO DE BASES

CONCEITOS FUNDAMENTAIS

- ✓ Mesmo sem saber com será em produção, temos uma forma de ao menos ter uma NOÇÃO de como o modelo performa em dados nunca vistos...
- ✓ Finalmente: utilizamos o modelo treinado no teste!
- ✓ O critério para saber se o modelo entra ou não em produção deve ser acordado com os donos do processo de trabalho cujo problema você está resolvendo

SEPARAÇÃO DE BASES

CONCEITOS FUNDAMENTAIS

- ✓ Mundo não ideal: não tem montão de dados
- ✓ Separação em
  - ✓ Treino: onde será efetivamente treinado
  - ✓ Teste: só vai ser utilizado NO FINAL
- ✓ “Uai... e onde valida?”

letsdata

VALIDAÇÃO CRUZADA

CONCEITOS FUNDAMENTAIS

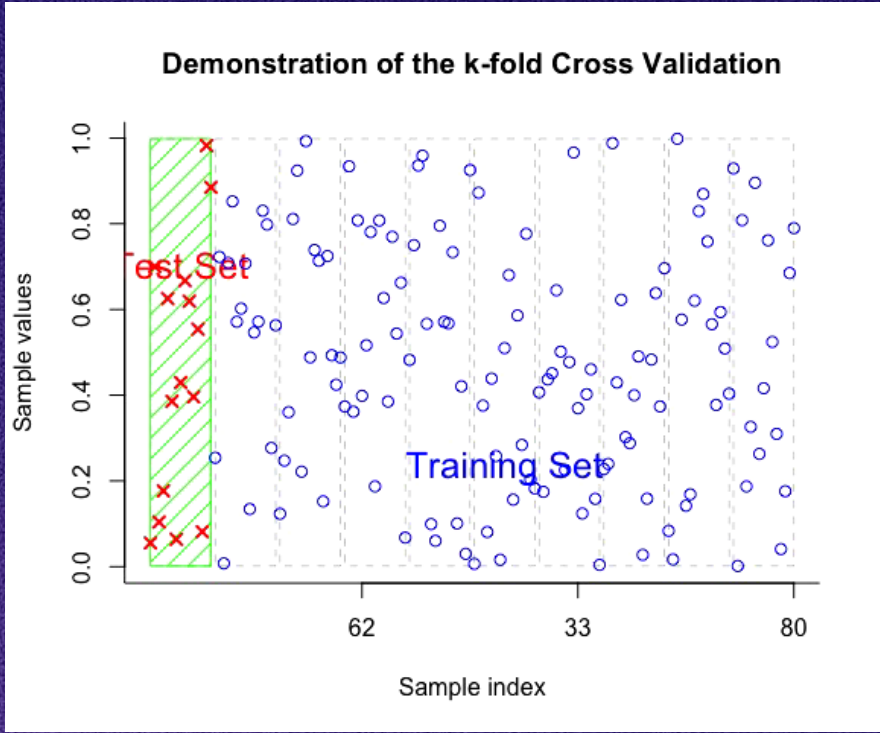
- ✓ Bom e famoso: validação cruzada (cross validation)!
- ✓ Ajuda a garantir que sua amostra de validação não foi enviesada (ou seja, é uma gambiarra boa que deve ser usada até quando se tem muitos dados)

VALIDAÇÃO CRUZADA

CONCEITOS FUNDAMENTAIS

- ✓ Esse vídeo do mestre Yihui Xie explica muito bem como funciona uma validação cruzada (10-fold cross validation):

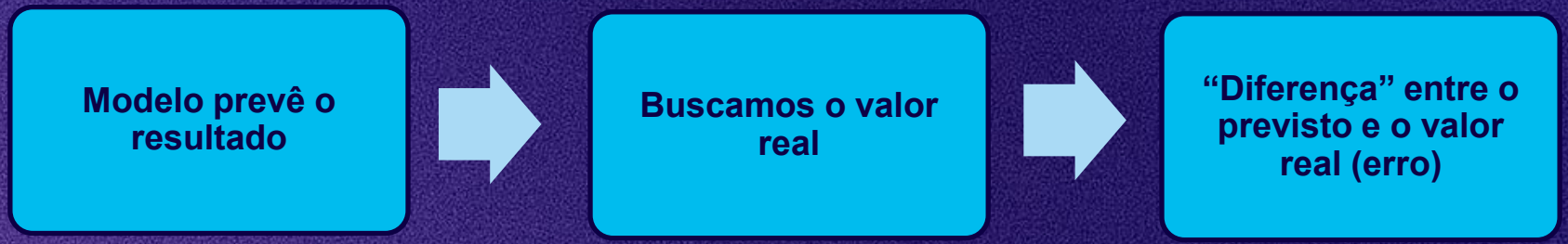
VALIDAÇÃO CRUZADA



letsdata

MÉTRICAS

CONCEITOS FUNDAMENTAIS



MÉTRICAS

CONCEITOS FUNDAMENTAIS

- ✓ Temos métricas para avaliação diferentes de acordo com o tipo de problema
- ✓ Métricas de Regressão (variável dependente contínua): calculamos a diferença numérica do valor previsto e do valor real
- ✓ Métricas de Classificação (variável dependente discreta: binária ou classes): "contamos" os erros de predição de cada classe

MÉTRICAS

CONCEITOS FUNDAMENTAIS

- ✓ Métricas de Regressão: Mean Squared Error, Mean Absolute Error, Root Mean Square Value
- ✓ Métricas de Classificação: acurácia, sensibilidade, sensitividade, área sob a curva ROC, f1-score
- ✓ Nomes bonitos e chiques para um mesmo padrão: identificar o quanto as previsões do nosso modelo está distante dos valores reais


scikit
learn

SCIKIT LEARN

CONCEITOS FUNDAMENTAIS

- ✓ Biblioteca Python de Machine Learning de código aberto
- ✓ Interface consistente: mesmo padrão para treino (fit), predição (predict) e validação
- ✓ Um montão de algoritmos: regressão, classificação, clustering, pré-processamento, métricas
- ✓ “Entende” pandas como fonte de dados

SCIKIT LEARN

PASSO 01

- ✓ Ler CSV com pandas.read_csv()
- ✓ Separar features (X) e target (y)
- ✓ Verificar faltantes, tipos de dados e distribuições
- ✓ Usar print() ou df.describe() para ter visão rápida do dataset

SCIKIT LEARN

PASSO 02

- ✓ Limpeza: remover ou preencher valores nulos essenciais
- ✓ Conversão de faixas salariais → valor numérico (apply + mapeamento)
- ✓ Codificação de categóricas: LabelEncoder ou OneHotEncoder
- ✓ train_test_split() para criar conjuntos treino (80%) e teste (20%)

SCIKIT LEARN

PASSO 03

- ✓ Escolher algoritmo: RandomForestRegressor para regressão não-linear robusta
- ✓ Definir hiperparâmetros principais (n_estimators, max_depth etc.)
- ✓ Ajustar: modelo.fit(X_train, y_train)
- ✓ Guardar encoders e lista de features para uso posterior

SCIKIT LEARN

PASSO 04

- ✓ Métricas-chave: r2_score, RMSE e MAE para regressão
- ✓ feature_importances_ mostra variáveis mais influentes
- ✓ Salvar o modelo TREINADO com pickle.dump(modelo_completo, 'modelo_salarios.pkl')
- ✓ Esse arquivo será usado lá no streamlit pra prever novos salários!

letsdata

BORA PRO MÃO NA MASSA!

AGORA VOCÊ JÁ SABE

AGENDA

- ✓ Descrever o que é machine learning!
- ✓ Enumerar diferentes tipos de machine learning
- ✓ Criar um modelo de regrassão para ajuda na sua análise!
- ✓ Integrar um modelo de ML num data product com Streamlitnálise!