

EPI5717: Machine learning para previsões em saúde

Prof. Dr. Alexandre Chiavegatto Filho



TIME
SPECIAL EDITION

Artificial Intelligence

The Future of Humankind



THE DECLINE OF INTERNATIONAL S

FOREIGN AFFAIRS

Hi, Robot

Work and Life in the Age of Automation

The Perils of Special Counsels / This Is Your ISIS on Drugs

Newsweek

08.02.2017



THE DOCTOR WILL SEE YOU NOW

HOW AI IS GOING TO CURE OUR SICK HEALTH CARE SYSTEM

INSIDE: A 14-PAGE SPECIAL REPORT ON FINANCIAL TECHNOLOGY

The Economist

MAY 9TH - 15TH 2015

economist.com

- How to fix America's inner cities
- The self-service economy
- Time to open up Indonesia
- Inside the anti-bribery business
- Why humans cause heatwaves

Artificial Intelligence

The promise and the peril



OS DOIS RITMOS DA LAVA-JATO
Curitiba: 107 condenados
Brasília: nenhum

MADONNA FALA A VEJA
"Sou uma rebelde e serrei rebelde até o fim"

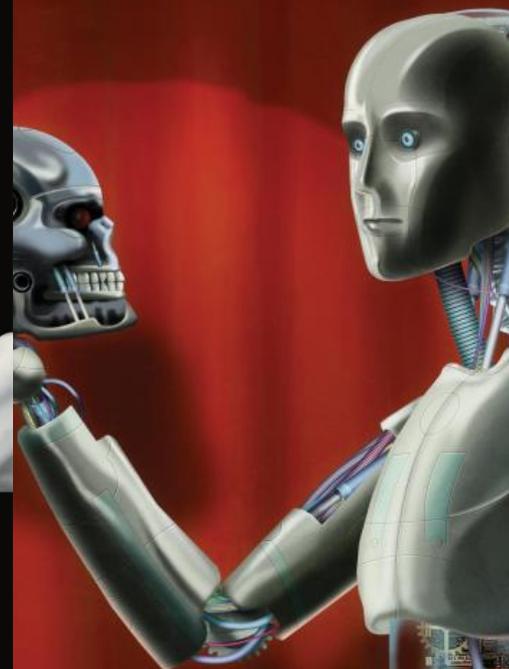
ASSINANTE
ABRIL

veja



DE MÃOS DADAS COM A INTELIGÊNCIA ARTIFICIAL

Longe dos cenários futurísticos da ficção científica, ela já faz parte do presente. Mas em que medida pode servir ao ser humano e, ao mesmo tempo, ameaçá-lo?



Science

510
17 JULY 2015
sciencemag.org

AAAS



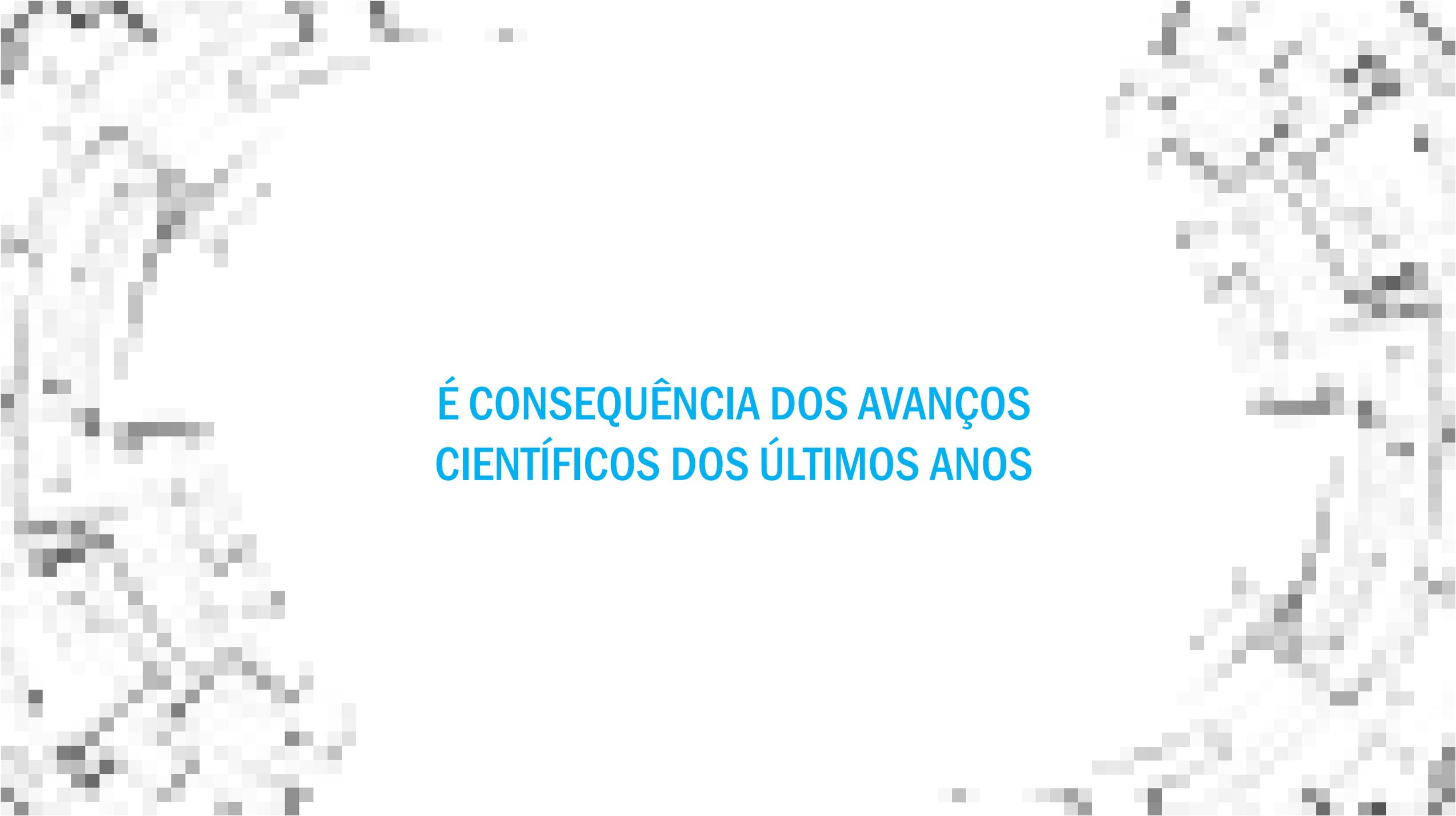
SPECIAL ISSUE

ARTIFICIAL INTELLIGENCE



Inteligência artificial

NÃO É HYPE CRIADO PELA MÍDIA



**É CONSEQUÊNCIA DOS AVANÇOS
CIENTÍFICOS DOS ÚLTIMOS ANOS**

**Por que têm ocorridos
avanços exponenciais
nos últimos 5 anos?**

**Por que têm ocorridos
avanços exponenciais
nos últimos 5 anos?**



Aumento da **quantidade de dados
(importante para melhorar performance)**

Por que têm ocorridos
avanços exponenciais
nos últimos 5 anos?



Aumento da **quantidade de dados**
(importante para melhorar performance)



Avanços em **capacidade computacional**
(modelos de *machine learning* exigem
muita memória).

Por que têm ocorridos avanços exponenciais nos últimos 5 anos?



Aumento da **quantidade de dados** (importante para melhorar performance)



Avanços em **capacidade computacional** (modelos de *machine learning* exigem muita memória).



Novos algoritmos para problemas mais complexos (*deep learning*).

A demanda por profissionais capazes de extrair informação relevante dos dados é muito grande no Brasil e no mundo



The most in-demand skills of 2020

Top 10 Hard Skills

- | | | |
|---|-------------------------|-----|
| 1 | Blockchain | new |
| 2 | Cloud computing | -1 |
| 3 | Analytical reasoning | - |
| 4 | Artificial intelligence | -2 |

A faint, light blue world map is visible in the background of the slide, centered behind the text.

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*Melhorar o
mundo*

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*melhorar o
mundo*



Desmatamento
da Amazônia

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*melhorar o
mundo*



Desmatamento
da Amazônia



Melhorar trânsito nas grandes
cidades.

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*melhorar o
mundo*



Desmatamento
da Amazônia



Melhorar trânsito nas grandes
cidades.



Desenvolvimento de
tecnologias verdes

MUITAS EMPRESAS TEM UTILIZADO O
MACHINE LEARNING PARA
*melhorar o
mundo*



Desmatamento
da Amazônia



Melhorar trânsito nas grandes
cidades.



Desenvolvimento de
tecnologias verdes



Melhoria da atenção à saúde e aumento
da qualidade de vida



INTELIGÊNCIA ARTIFICIAL

Capacidade de máquinas tomarem decisões inteligentes.

Várias definições para inteligência.

Possibilidade: “capacidade de tomar a melhor decisão possível dada a informação disponível. Com a capacidade de se adaptar a novas situações.”

Segundo essa definição, inteligência é um problema de análise de dados.

MACHINE LEARNING

Inteligência artificial *clássica*

Regras para a tomada de decisão ensinada por humanos

- Identificar spam via palavras-chave.
- Traduzir uma frase através de dicionário e regras de gramática.
- Identificar caras humanas por meio da forma de nariz, olho, boca etc.

Inteligência artificial com *machine learning*

Máquinas aprendendo sozinhas!

Tomada de decisão via identificação de padrões complexos nos dados.

É como uma criança aprende!

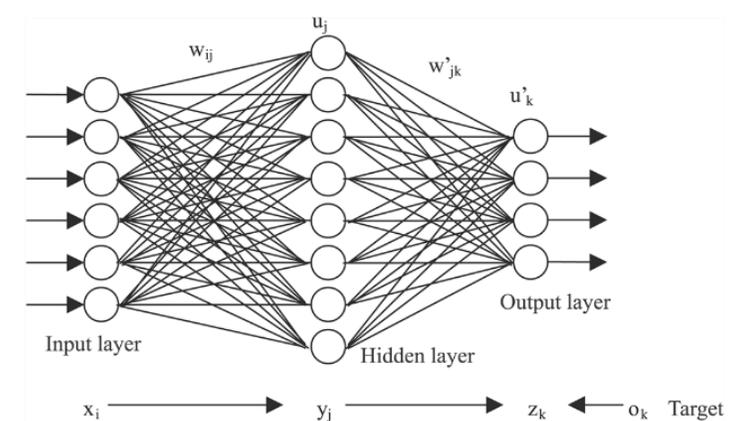
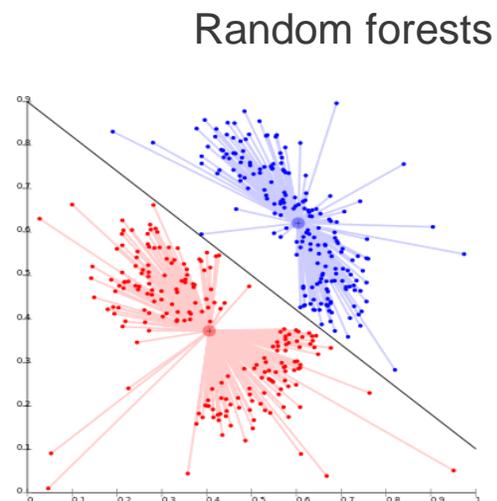
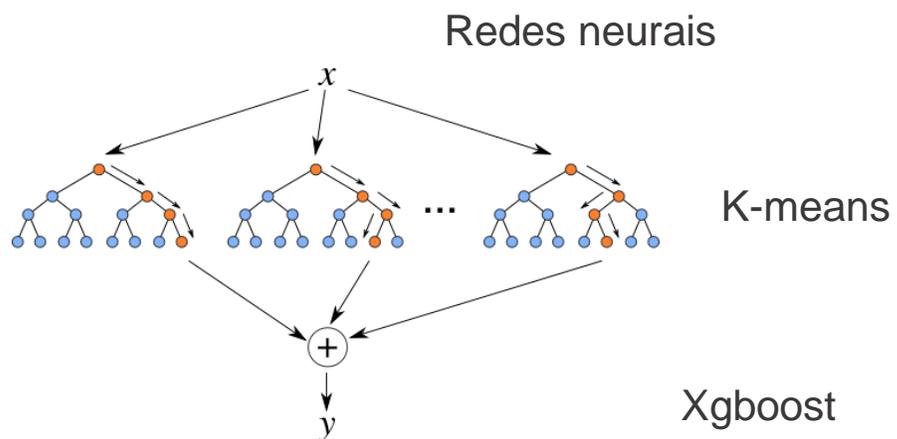


MACHINE LEARNING

Problemas **práticos** de predição (para a tomada de decisão)

Pouco interesse em *interpretar* os modelos.

Liberdade para modelar a complexidade do mundo real



Regressões penalizadas

QUATRO CATEGORIAS DE MACHINE LEARNING

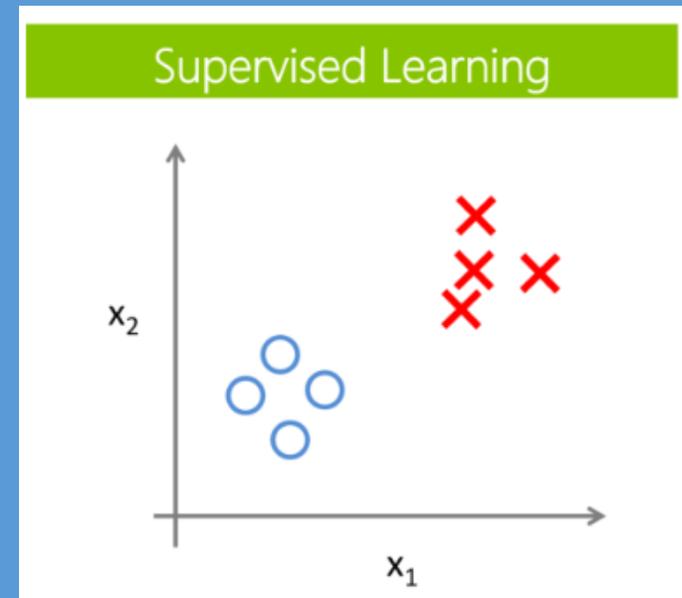
Objetivo: acertar
um resultado que
existe.

Treinar um modelo para obter a
melhor performance preditiva
possível para um *problema*

APRENDIZADO SUPERVISIONADO

Quando os dados incluídos
para treinar o algoritmo
incluem a solução desejada,
ou rótulo (“label”).

RESPOSTA
CERTA



Aprendizado supervisionado

Divididos em dois grandes grupos

CLASSIFICAÇÃO

Quando a variável a ser predita é categórica

Exemplo

Óbito em 5 anos, incidência de doença em 10 anos, etc.

REGRESSÃO

Quando a variável a ser predita é quantitativa

Exemplo

Quantos meses de vida a pessoa tem pela frente, qual será o seu IMC no próximo ano, etc.

Aprendizado não-supervisionado

Não existe rótulo (“label”).

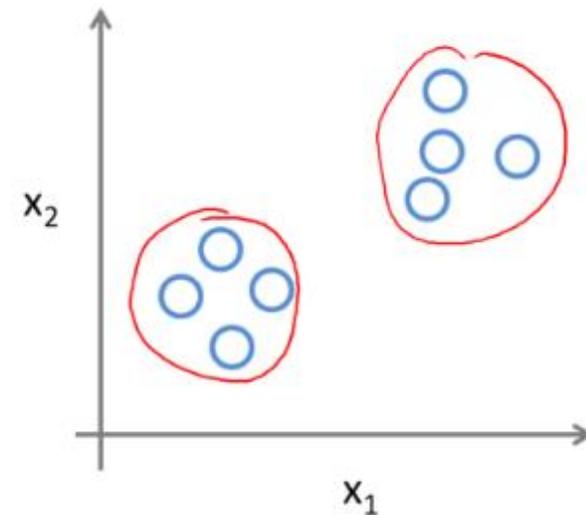
Algoritmo aprende sem uma resposta certa.

O objetivo é encontrar padrões nos dados.

Mais comuns: clustering (agrupamentos) e redução de dimensão (ACP).

QUATRO CATEGORIAS DE MACHINE LEARNING

APRENDIZADO NÃO-SUPERVISIONADO

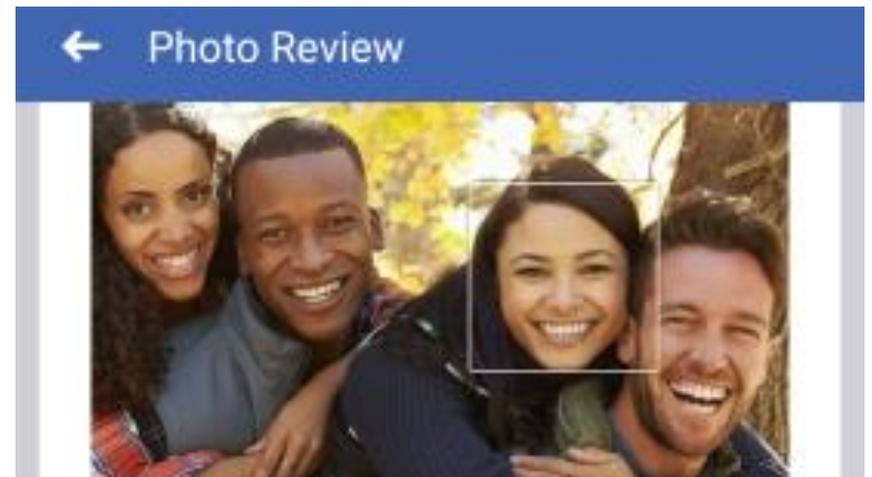


Aprendizado semi-supervisionado

Presença de alguns dados com rótulo e outros sem.

Identificação de fotos do Facebook: algoritmo identifica que a mesma pessoa está em várias fotos e só precisa de um rótulo.

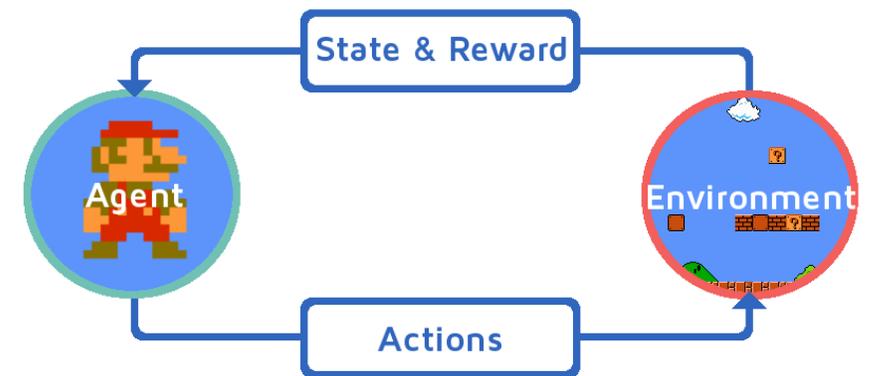
QUATRO CATEGORIAS DE MACHINE LEARNING



Aprendizado por reforço

- Agente interage com um ambiente dinâmico.
- Feedbacks constantes em termos de premiações e punições.
- Jogos.

QUATRO CATEGORIAS DE MACHINE LEARNING



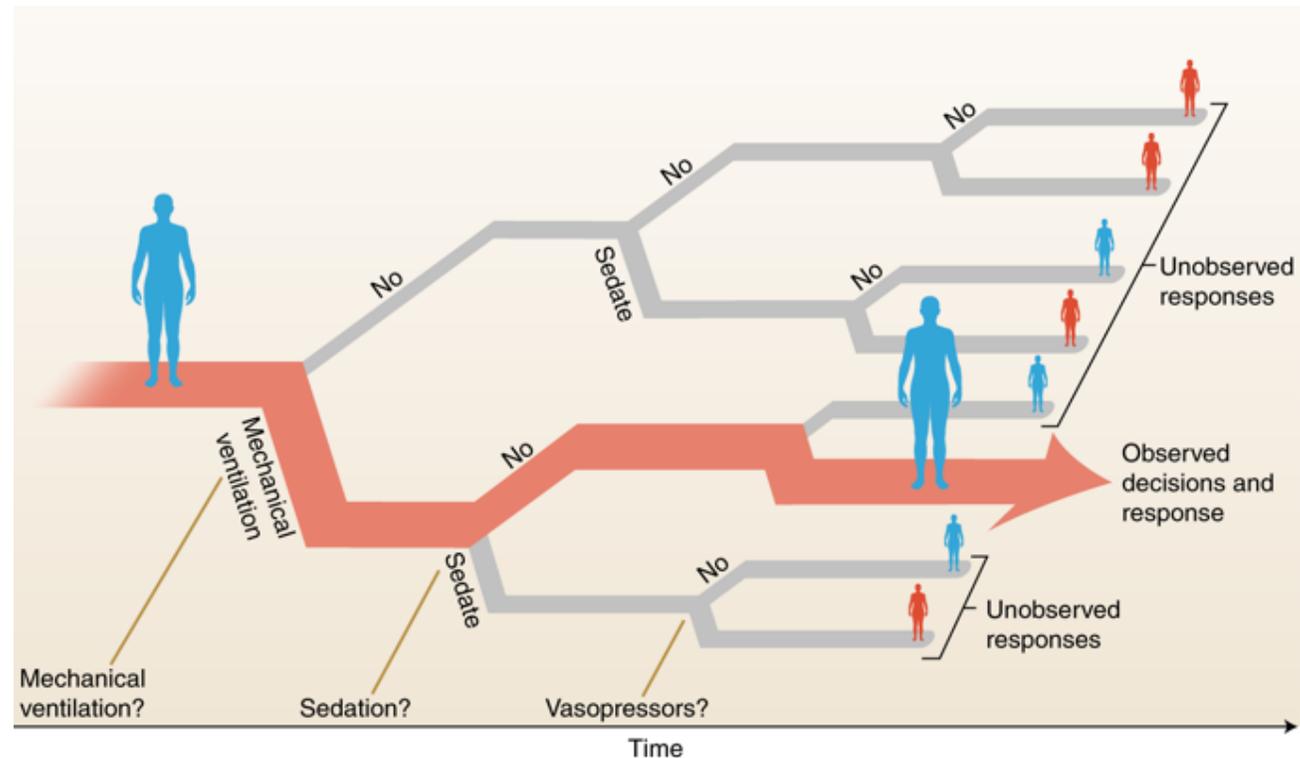
Aprendizado por reforço

QUATRO CATEGORIAS DE MACHINE LEARNING

Área promissora:

- Diferentes etapas do tratamento médico para identificar sequência ótima (ex: sepse).

Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, Celi LA. Guidelines for reinforcement learning in healthcare. Nat Med 2019. 25(1)16-18.

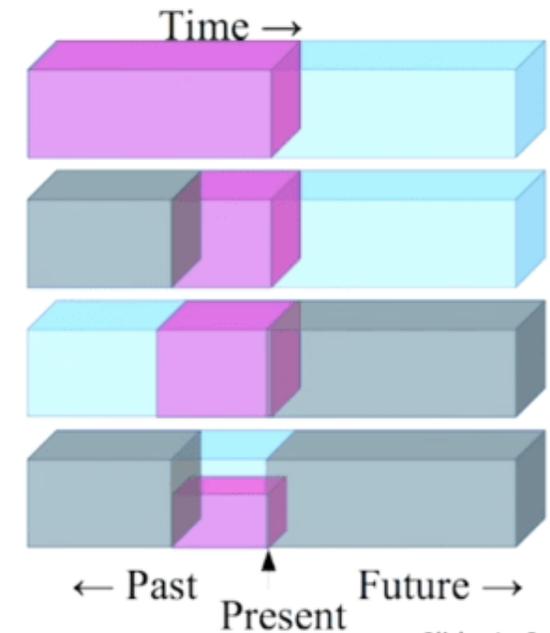


Nova: Autossupervisionado

- Um problema não supervisionado é transformado em um problema supervisionado pela geração automática dos rótulos.

- O processo de aprendizado autossupervisionado consiste em identificar partes ocultas dos dados.

- ▶ Predict any part of the input from any other part.
- ▶ Predict the **future** from the **past**.
- ▶ Predict the **future** from the **recent past**.
- ▶ Predict the **past** from the **present**.
- ▶ Predict the **top** from the **bottom**.
- ▶ Predict the **occluded** from the **visible**
- ▶ **Pretend there is a part of the input you don't know and predict that.**



PREDIÇÃO COM MACHINE LEARNING

- Trade-off entre predição e interpretação.
- Em predição estamos interessados em performance preditiva.
- Em inferência o interesse é entender a relação entre variáveis, normalmente como Y muda com uma alterações entre seus determinantes.
- Problema: modelos facilmente interpretáveis (regressões linear e logística, árvores de decisão) normalmente têm pior performance preditiva.
- Pensar bem sobre o objetivo da análise: é inferência ou predição?

PREDIÇÃO COM MACHINE LEARNING

Dados

Preferencialmente muitos e com boa qualidade (preenchidos corretamente e preditores fortes). Realizar o pré-processamento das variáveis.

Algoritmos

Inserir os dados no algoritmo de machine learning para aprender os parâmetros (regressão logística/linear) ou estruturas (árvores) que mapeiam os preditores aos resultados.

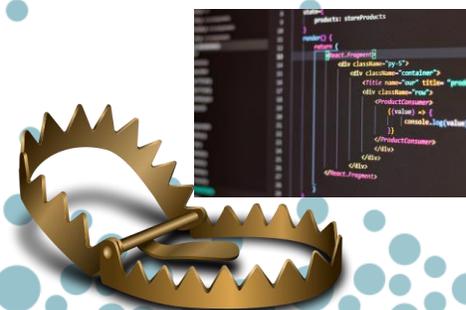
Testar no futuro

Inserir no algoritmo novos dados para testar a qualidade desse algoritmo para prever dados futuros.

Artigo: How to avoid machine learning pitfalls: a guide for academic researchers

Autor: Michael A. Lones*

*School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, Scotland, UK



Objetivo



- Elencar erros comuns ao aplicar técnicas de machine learning;
- Como evitar esses erros;
- Destina-se como um guia para estudantes;
- Foco em questões de pesquisa acadêmica;

Discussão



1. O que fazer antes de construir um modelo
2. Como construir modelos
3. Como avaliar modelos
4. Como comparar modelos
5. Como relatar resultados

1

Antes de começar a construir modelos



Entenda os dados



Não olhe para todos os seus dados



Certifique-se de que há dados suficientes



Converse com especialistas no assunto



Faça um levantamento de literatura



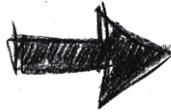
Pense em como o modelo será utilizado

1

Antes de começar a construir modelos



Entenda os dados



É importante saber:

- Fonte
- Forma de coleta
- Se há artigos que descreveram o banco de dados
- Se o artigo do banco de dados foi publicado em uma revista confiável
- Se os dados fazem sentido para o objetivo
- Comece sempre por uma análise exploratória

1

Antes de começar a construir modelos



Entenda os dados



Não olhe para todos os seus dados



Certifique-se de que há dados suficientes



Converse com especialistas no assunto



Faça um levantamento de literatura



Pense em como o modelo será utilizado

1

Antes de começar a construir modelos



Não olhe para
todos os seus
dados



- Evite analisar os dados de teste durante a análise exploratória inicial (suposições consciência ou inconscientes)
- Cuidado com vazamentos (acesso aos dados de teste)

1

Antes de começar a construir modelos



Entenda os dados



Não olhe para todos os seus dados



Certifique-se de que há dados suficientes



Converse com especialistas no assunto



Faça um levantamento de literatura



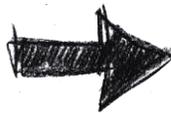
Pense em como o modelo será utilizado

1

Antes de começar a construir modelos



Certifique-se de que há dados suficientes



- Sem dados suficientes é impossível treinar um modelo que generalize
 - Descobrir isso pode não ser tão simples (sinal vs ruído)
 - Sinal forte = pode usar menos dados
 - Sinal fraco = precisa de mais dados
- ou usar validação cruzada
- Data Argumentation = aumento no número de observações



↓
Em classificação: importante em dados desbalanceados

↓
Não usar acurácia como medida de desempenho

1

Antes de começar a construir modelos



Entenda os dados



Não olhe para todos os seus dados



Certifique-se de que há dados suficientes



Converse com especialistas no assunto



Faça um levantamento de literatura



Pense em como o modelo será utilizado

1

Antes de começar a construir modelos



Converse com
especialistas no
assunto



- Ajuda para entender o problema (ex.: variáveis uteis)
- Modelo útil na prática real
- Checar de importância de variáveis fazem sentido e identificar vazamentos
- Escolha da revista adequada para publicação

1

Antes de começar a construir modelos



Entenda os dados



Não olhe para todos os seus dados



Certifique-se de que há dados suficientes



Converse com especialistas no assunto



Faça um levantamento de literatura



Pense em como o modelo será utilizado

1

Antes de começar a construir modelos



Faça um levantamento de literatura



- Entender o que foi feito anteriormente
- Progresso acadêmico = Processo iterativo
- Aproveitar o conhecimento existente

1

Antes de começar a construir modelos



Entenda os dados



Não olhe para todos os seus dados



Certifique-se de que há dados suficientes



Converse com especialistas no assunto



Faça um levantamento de literatura



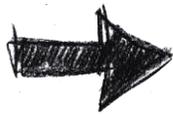
Pense em como o modelo será utilizado

1

Antes de começar a construir modelos



Pense em como
o modelo será
utilizado



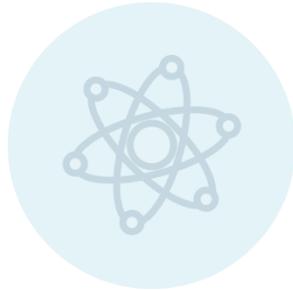
- "Por que você quer construir esse modelo?"
- Implementação em ambiente real
- Vinculação do modelo a um Sistema (prontuário eletrônico, aplicativo)

2

Como construir modelos confiáveis



Impeça vazamento de dados pro teste



Teste vários modelos



Não use modelos inapropriados



Otimize os hiperparâmetros



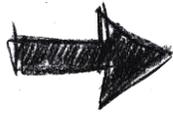
Cuidado com a otimização de hiperparâmetros e seleção de variáveis

2

Como construir modelos confiáveis



Impeça
vazamento de
dados pro teste



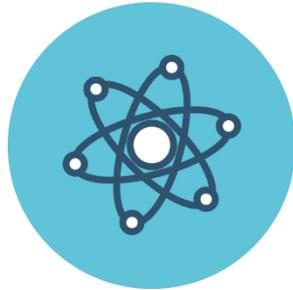
- Separe o conjunto de teste e não mexa nele!
- Vazamento torna modelos inúteis
- Falta de confiabilidade
- Exemplos:
 - Pré-processamento só no treino
 - Seleção de variáveis e otimização de hiperparâmetros só no treino

2

Como construir modelos confiáveis



Impeça vazamento de dados pro teste



Teste vários modelos



Não use modelos inapropriados



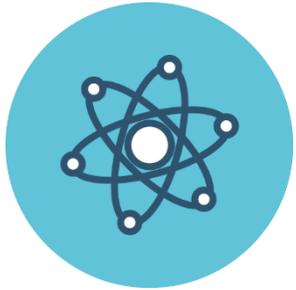
Otimize os hiperparâmetros



Cuidado com a otimização de hiperparâmetros e seleção de variáveis

2

Como construir modelos confiáveis



Teste vários
modelos



- Não existe um único algoritmo melhor
- Teorema do almoço grátis
- Encontrar algoritmo que gere a melhor predição
- Python: Scikit Learn

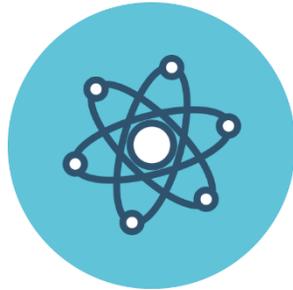


2

Como construir modelos confiáveis



Impeça vazamento de dados pro teste



Teste vários modelos



Não use modelos inapropriados



Otimize os hiperparâmetros



Cuidado com a otimização de hiperparâmetros e seleção de variáveis

2

Como construir modelos confiáveis



Não use
modelos
inapropriados



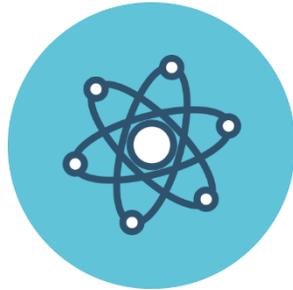
- Modelos inadequados aos dados: variáveis categóricas quando o modelo só aceita numéricas
- Modelos desnecessariamente complexos (deep learning para dados estruturados)
- Modelos mais recentes nem sempre são melhores (cuidado com artigos de benchmark em que autores querem que seu algoritmo vença)

2

Como construir modelos confiáveis



Impeça vazamento de dados pro teste



Teste vários modelos



Não use modelos inapropriados



Otimize os hiperparâmetros



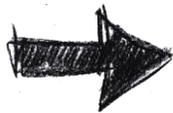
Cuidado com a otimização de hiperparâmetros e seleção de variáveis

2

Como construir modelos confiáveis



Otimize os
hiperparâmetros



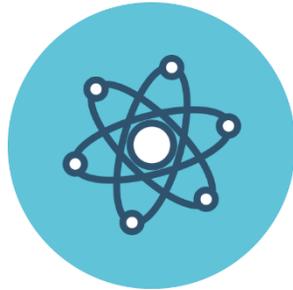
- Hiperparâmetros: parâmetros que afetam a configuração do modelo
- Hiperparâmetros afetam o desempenho
- Ajuste aos dados
- Estratégias de otimização (random search, grid search, hyperopt etc)

2

Como construir modelos confiáveis



Impeça vazamento de dados pro teste



Teste vários modelos



Não use modelos inapropriados



Otimize os hiperparâmetros



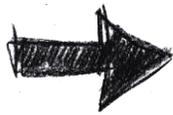
Cuidado com a otimização de hiperparâmetros e seleção de variáveis

2

Como construir modelos confiáveis



Cuidado com a
otimização de
hiperparâmetros e
seleção de
variáveis



- Erro comum: seleção de variáveis antes da divisão dos dados
- Seleção de variáveis e otimização de hiperparâmetros devem ser feitas com dados de treino
- Técnica: Validação cruzada aninhada (definição de hiperparâmetros e parâmetros dentro da mesma validação cruzada)

3

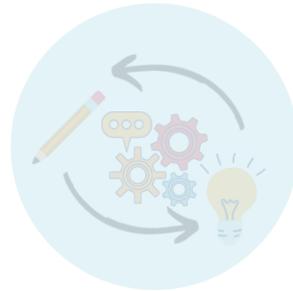
Como avaliar modelos



Use um conjunto de teste adequado



Use um conjunto de validação



Avalie o modelo várias vezes



Salve dados para uma avaliação final



Acurácia

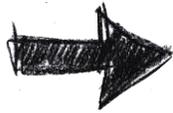
Não use a acurácia com dados desbalanceados

3

Como avaliar modelos



Use um conjunto
de teste
adequado



- Sempre use um conjunto de teste!
- Dados de teste apropriados (representativo)

3

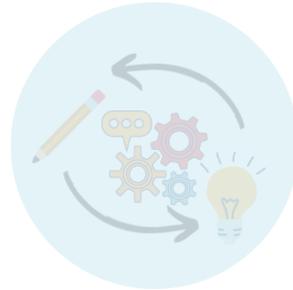
Como avaliar modelos



Use um conjunto de teste adequado



Use um conjunto de validação



Avalie o modelo várias vezes



Salve dados para uma avaliação final



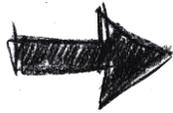
Não use a acurácia com dados desbalanceados

3

Como avaliar modelos



Use um conjunto de validação



- Conjunto de validação: conjunto de dados intermediário ao treino e teste, para medir desempenho
- Pode ser usado para orientar o treinamento

3

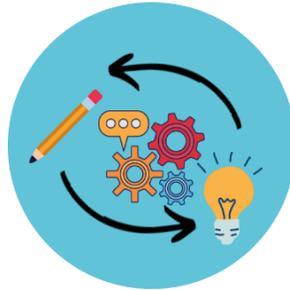
Como avaliar modelos



Use um conjunto de teste adequado



Use um conjunto de validação



Avalie o modelo várias vezes



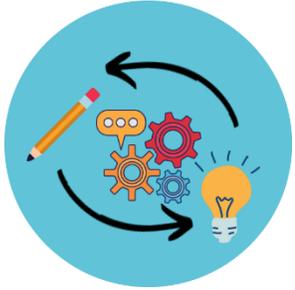
Salve dados para uma avaliação final



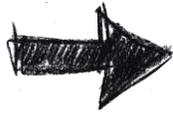
Não use a acurácia com dados desbalanceados

3

Como avaliar modelos



Avalie o modelo
várias vezes



- Alguns modelos são instáveis
- Pequenas mudanças no treino podem mudar o desempenho significativamente
- Validação Cruzada (com estratificação em problemas de classificação)

3

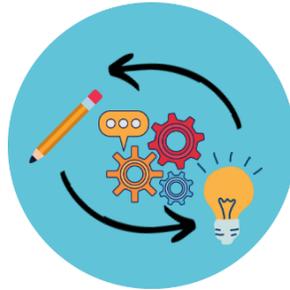
Como avaliar modelos



Use um conjunto de teste adequado



Use um conjunto de validação



Avalie o modelo várias vezes



Salve dados para uma avaliação final



Não use a acurácia com dados desbalanceados

3

Como avaliar modelos



Salve dados
para uma
avaliação final



- Validação cruzada possui um conjunto de teste pequeno – usar média da performance dos folds
- Usar teste se possível (deixar validação cruzada para seleção de hiperparâmetros)

3

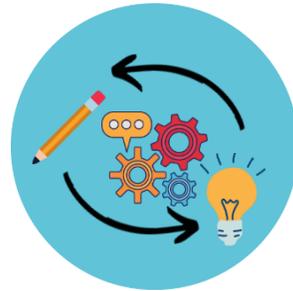
Como avaliar modelos



Use um conjunto de teste adequado



Use um conjunto de validação



Avalie o modelo várias vezes



Salve dados para uma avaliação final



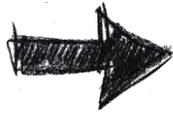
Não use a acurácia com dados desbalanceados

3

Como avaliar modelos



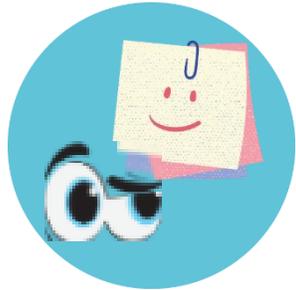
Não use a
acurácia com
dados
desbalanceados



- Cuidado com as métricas de avaliação!
- Acurácia: Em classificação só funciona com amostras balanceadas (mesmo número de observações de ambas classes)
- Os dados de saúde costumam ser desbalanceados

4

Como comparar modelos



Não presume
que valores
maiores
signifiquem
modelos
melhores



Use testes
estatísticos
para comparar
modelos



Faça correções
para múltiplas
comparações



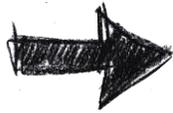
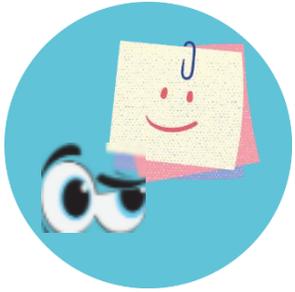
Nem sempre
acredite nos
resultados já
encontrados em
outros trabalhos



Considere
combinações de
modelos

4

Como comparar modelos

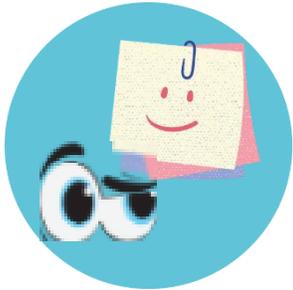


Não presume
que valores
maiores
signifiquem
modelos
melhores

- Cuidado ao compara seus resultados com o de outras pesquisas (diferentes bancos de dados)
- Comparações devem ser justas
- Desempenho depende de como o modelo foi desenvolvido (tipo de dados de treino, formas de avaliação, otimização dos hiperparâmetros, etc)

4

Como comparar modelos



Não presuma
que números
maiores
signifiquem
modelos
melhores



Use testes
estatísticos
para comparar
modelos



Faça correções
para múltiplas
comparações



Nem sempre
acredite nos
resultados já
encontrados em
outros trabalhos



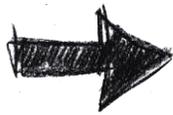
Considere
combinações de
modelos

4

Como comparar modelos



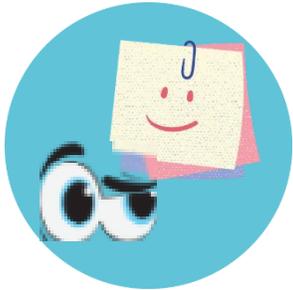
Use testes estatísticos para comparar modelos



- Use testes estatísticos para comparações entre algoritmos e com outros trabalhos (debatível)

4

Como comparar modelos



Não presuma
que valores
maiores
signifiquem
modelos
melhores



Use testes
estatísticos
para comparar
modelos



Faça correções
para múltiplas
comparações



Nem sempre
acredite nos
resultados já
encontrados em
outros trabalhos



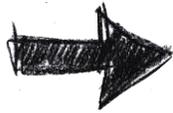
Considere
combinações de
modelos

4

Como comparar modelos



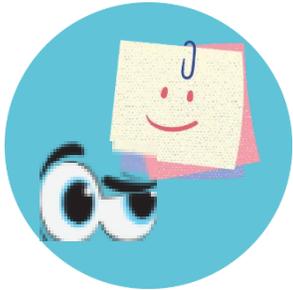
Faça correções
para múltiplas
comparações



- Necessidade de testes estatísticos para comparar mais de dois modelos
- Correção para múltiplos testes: Correção de Bonferroni

4

Como comparar modelos



Não presuma
que valores
maiores
signifiquem
modelos
melhores



Use testes
estatísticos
para comparar
modelos



Faça correções
para múltiplas
comparações



Nem sempre
acredite nos
resultados já
encontrados em
outros trabalhos



Considere
combinações de
modelos

4

Como comparar modelos



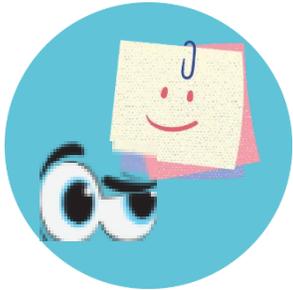
Nem sempre acredite nos resultados já encontrados em outros trabalhos



- Problema da utilização de dados de mesmos bancos
- Desconhecer se não olharam o teste antes
 - Coletivamente o teste está sendo usado muitas vezes (overfitting nesse teste)

4

Como comparar modelos



Não presume
que valores
maiores
signifiquem
modelos
melhores



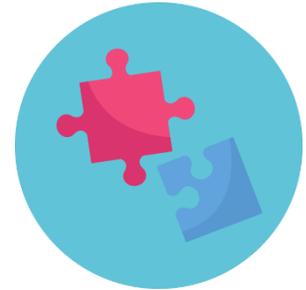
Use testes
estatísticos
para comparar
modelos



Faça correções
para múltiplas
comparações



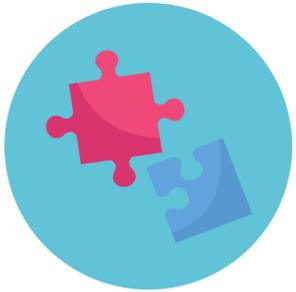
Nem sempre
acredite nos
resultados já
encontrados em
outros trabalhos



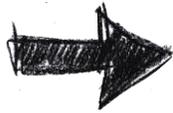
Considere
combinações de
modelos

4

Como comparar modelos



Considere
combinações de
modelos



- Compensar pontos fracos com pontos fortes
- Exemplo:
 - Bagging e Boosting
- Como combinar diferentes modelos de base?
(votação ou usar um algoritmo final para agregação de resultados)

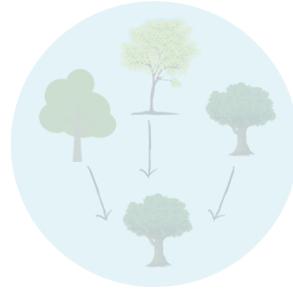
5 Como relatar resultados



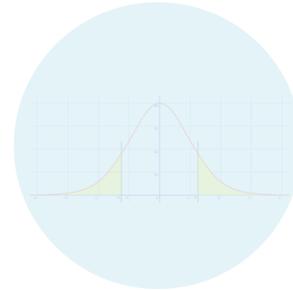
Seja transparente



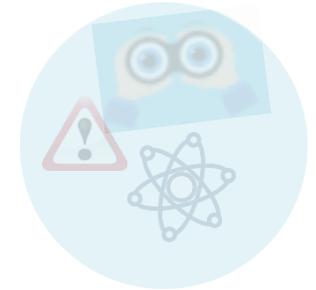
Relate o desempenho de várias formas



Não generalize para além do seu conjunto de dados



Cuidado ao relatar significância estatística



Olhe seus modelos

5 Como relatar resultados

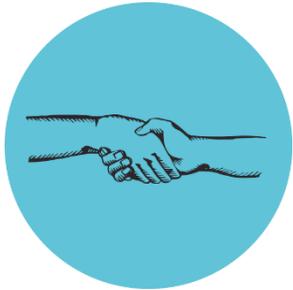


Seja
transparente



- Objetivo é contribuir com o conhecimento
- ML possui muitos trade-offs
- Compartilhe o modelo de forma acessível (ex.: github)
- Código limpo e organizado (documente as etapas)
- Facilitar a reprodutibilidade

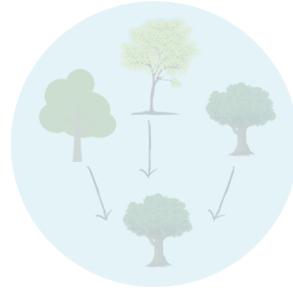
5 Como relatar resultados



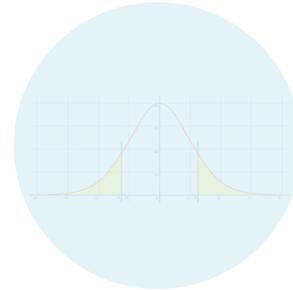
Seja transparente



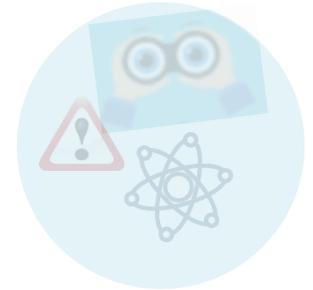
Relate o desempenho de várias formas



Não generalize para além do seu conjunto de dados



Cuidado ao relatar significância estatística



Olhe seus modelos

5 Como relatar resultados



Relate o
desempenho
de várias
formas



- Múltiplos conjuntos de dados para teste
- Relate várias métricas

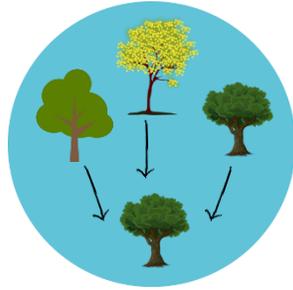
5 Como relatar resultados



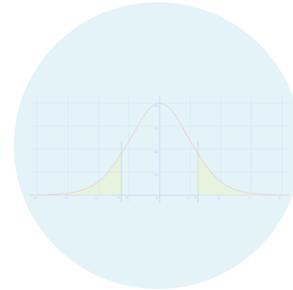
Seja transparente



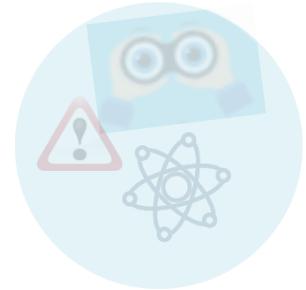
Relate o desempenho de várias formas



Não generalize para além do seu conjunto de dados

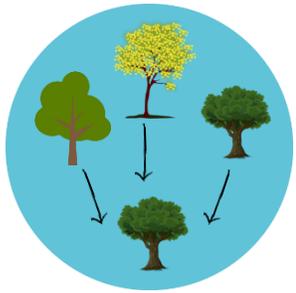


Cuidado ao relatar significância estatística

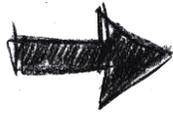


Olhe seus modelos

5 Como relatar resultados

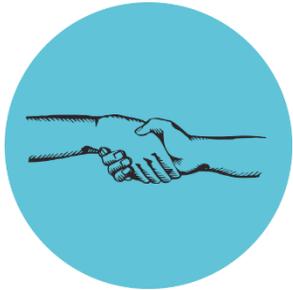


Não generalize
para além do seu
conjunto de
dados



- Não faça afirmações exageradas
- Esteja ciente das limitações do estudo
- Variações com a qualidade dos dados e diferenças de distribuição

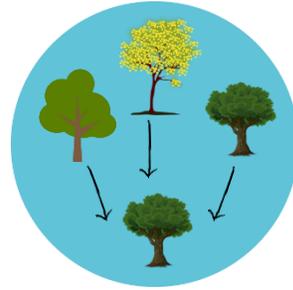
5 Como relatar resultados



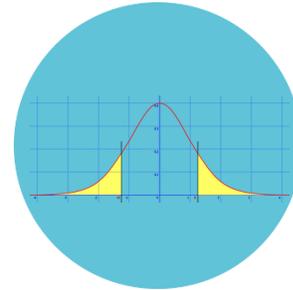
Seja transparente



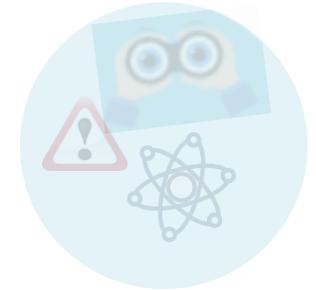
Relate o desempenho de várias formas



Não generalize para além do seu conjunto de dados

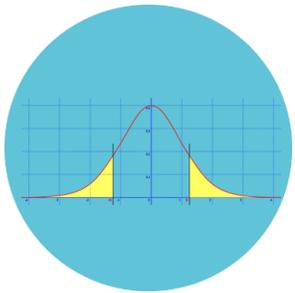


Cuidado ao relatar significância estatística



Olhe seus modelos

5 Como relatar resultados



Cuidado ao
relatar
significância
estatística



- Testes estatísticos não são perfeitos!
- Significância estatística: relate o valor p e deixe que o leitor os interprete

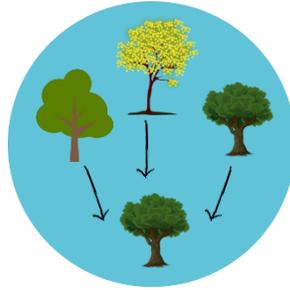
5 Como relatar resultados



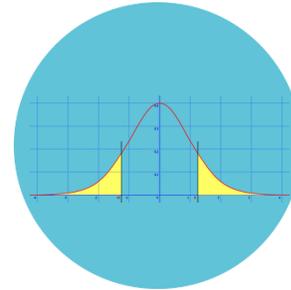
Seja transparente



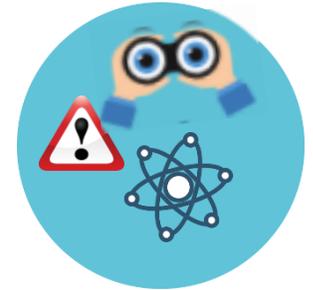
Relate o desempenho de várias formas



Não generalize para além do seu conjunto de dados

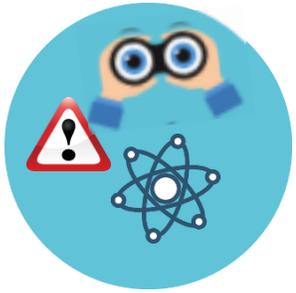


Cuidado ao relatar significância estatística

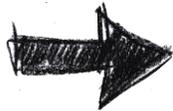


Olhe seus modelos

5 Como relatar resultados



Olhe seus
modelos



- Mostre como o modelo aprendeu e tomou as decisões
- Objetivo: gerar conhecimento
- Forneça visualizações
- Use técnicas de explicabilidade (Shapley)

Em ML assim como em qualquer tipo de pesquisa:

- mantenha a mente aberta
- acompanhe o desenvolvimento da rea
- aceite que você não sabe tudo

Referência

LONES, Michael A. How to avoid machine learning pitfalls: a guide for academic researchers. arXiv preprint arXiv:2108.02497, 2021.



LABDAPS

LABORATÓRIO DE BIG DATA E
ANÁLISE PREDITIVA EM SAÚDE



Obrigado!

Alexandre Chiavegatto Filho



<http://labdaps.fsp.usp.br>



@SaudenoBR



@labdaps



alexdiasporto@usp.br

