Compreendendo o viés e a explicabilidade em modelos de dados sensiveis

Isabella Bicalho

Palestrante principal

Marília Melo Favalesso

Cientista de dados

TODOS OS MODELOS ESTÃO ERRADOS

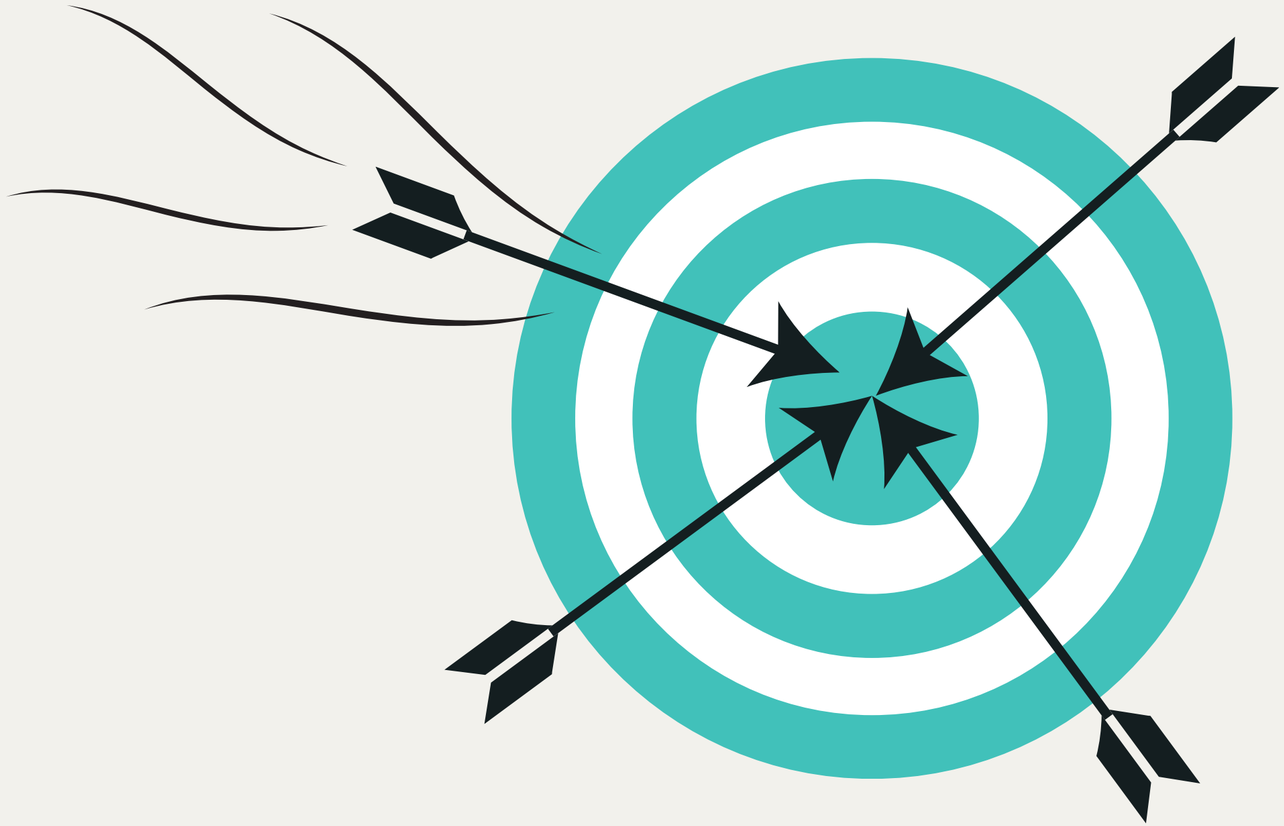
Mas isso não significa que sejam inúteis.

George Caixa

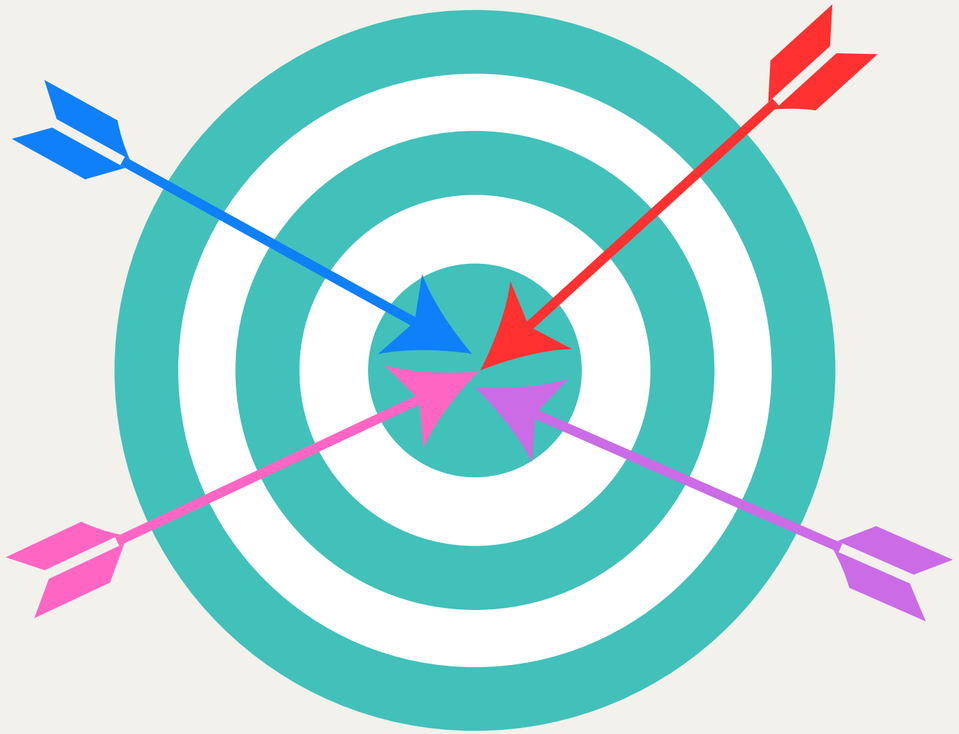
MODELO **ACURADO**

MODELO **PRECISO**

MODELO **RÁPIDO**

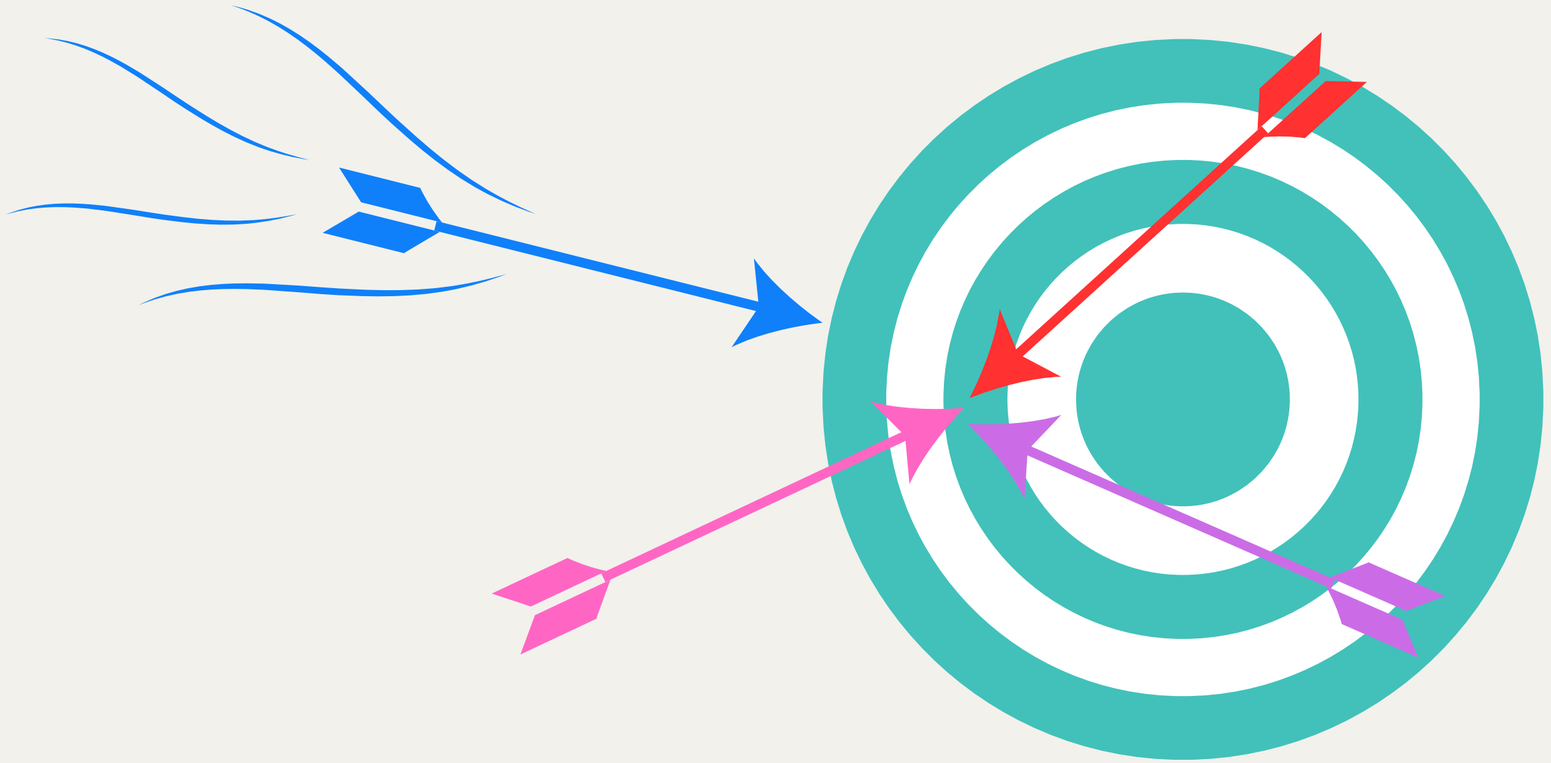


MODELO **JUSTO**



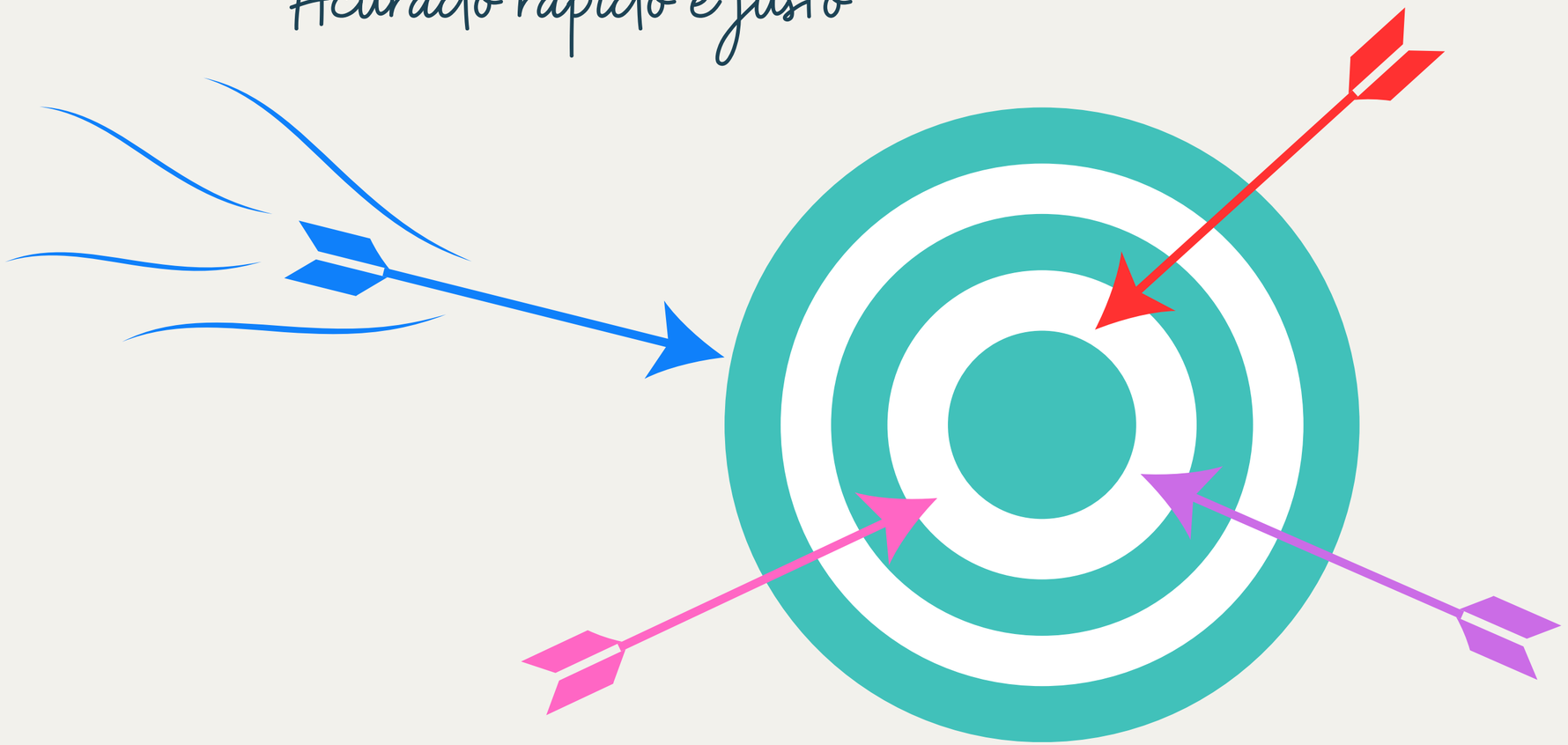
MODELO **INACURADO**

Preciso, rápido e justo



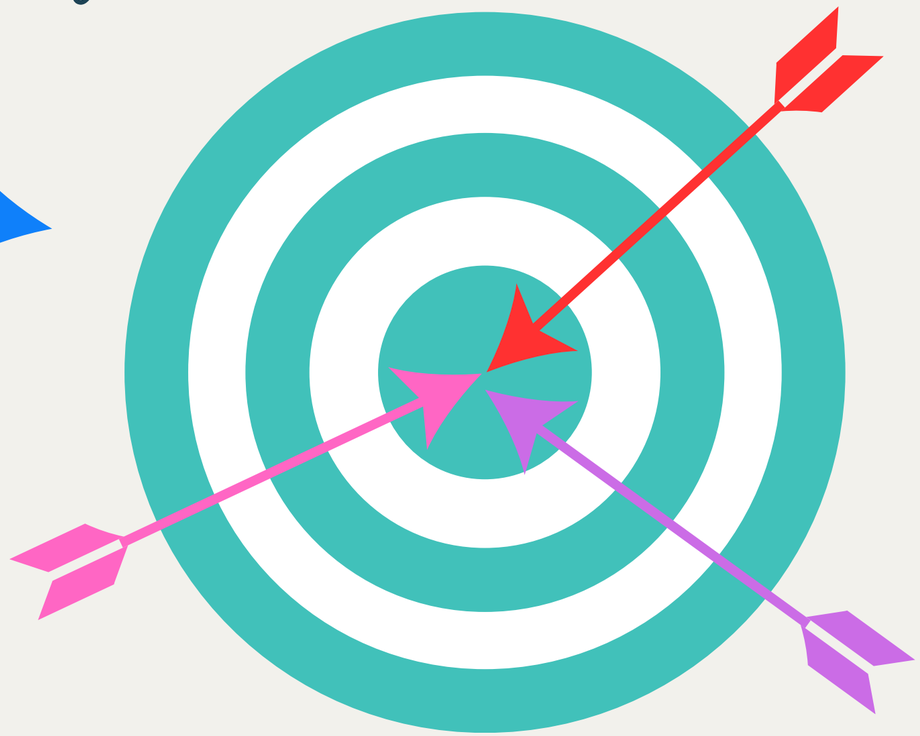
MODELO **IMPRECISO**

Acurado rápido e justo



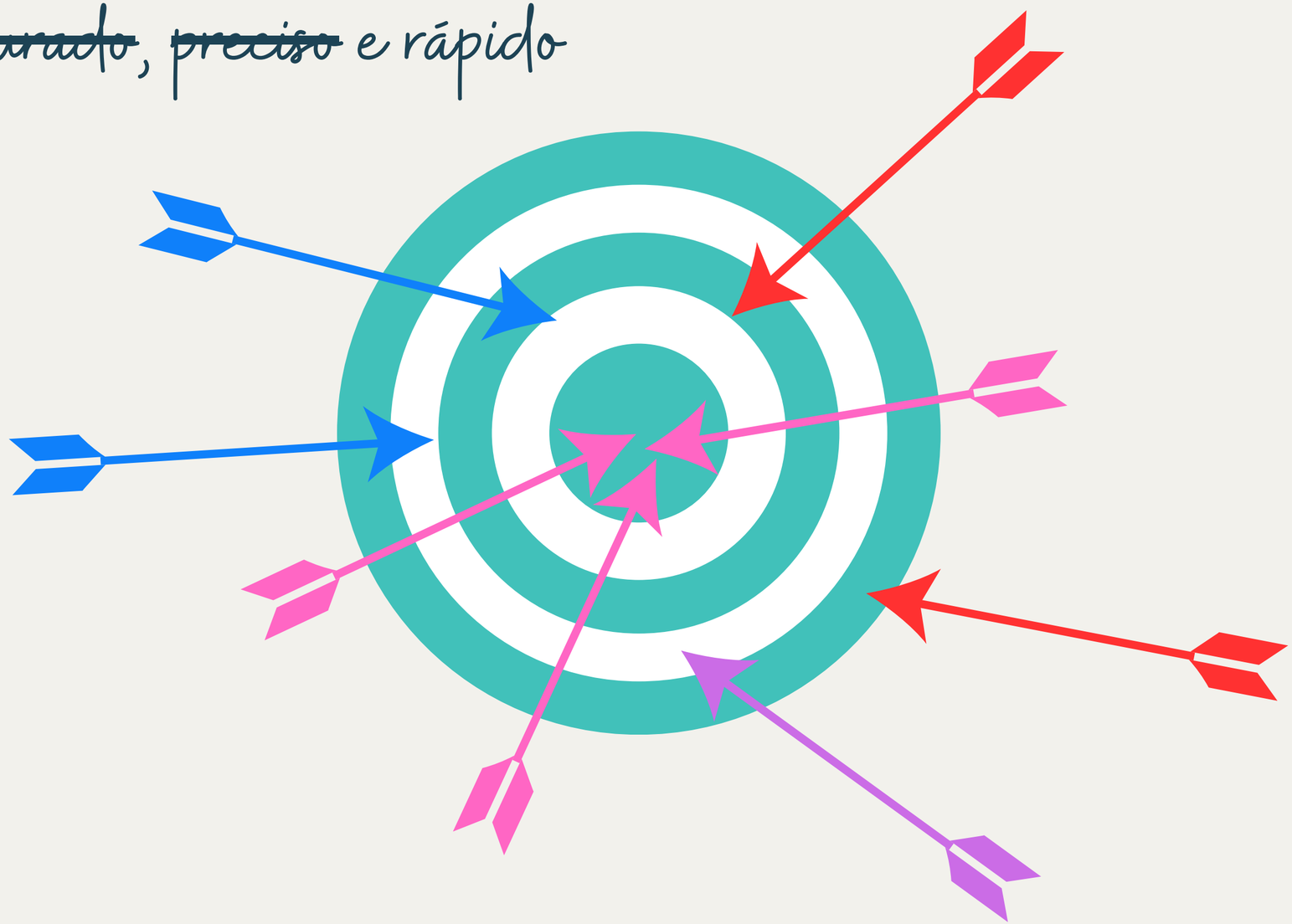
MODELO **LENTO**

Acurado, preciso e justo



MODELO **INJUSTO**

~~Acurado~~, ~~preciso~~ e rápido



MAS O QUE É SER 'JUSTO'?

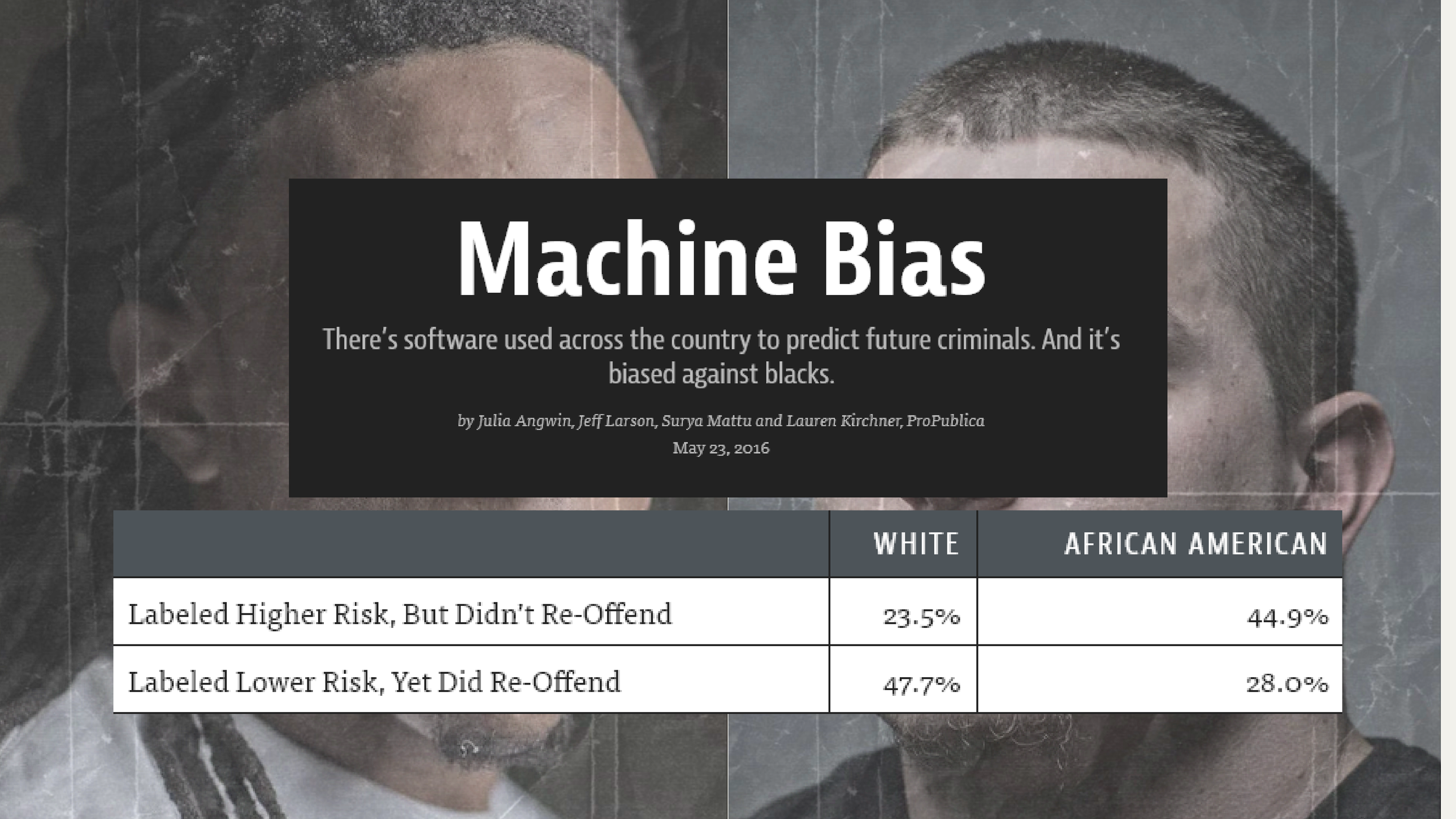
Como frequentemente se afirma nas ciências não exatas,

depende...

MAS O QUE É SER 'JUSTO'?

Em Inteligência Artificial (IA) refere-se à busca por sistemas e algoritmos que operem de maneira **equitativa**, **sem discriminação** ou **viés injusto** contra qualquer indivíduo ou grupo.


Alto Risco 10
3 Baixo risco



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%


Did someone blink?
content.time.com
Breaking News, Analysis, Politics, Blogs, News Photos, Video, Tech Reviews - TIME.com
Why face-detection software on cameras and webcams made by HP, Nikon and Sony is being called out by consumers for failing to recognize black and Asian features and faces
OK
Exit

CARTAEXPRESSA

Deputada denuncia ‘racismo algorítmico’ após IA gerar imagem com arma em uma favela

Renata Souza (PSOL-RJ) foi retratada com uma arma ao fornecer referências como o fato de ser negra e estar em uma favela

POR **CARTACAPITAL** | 26.10.2023 16H26

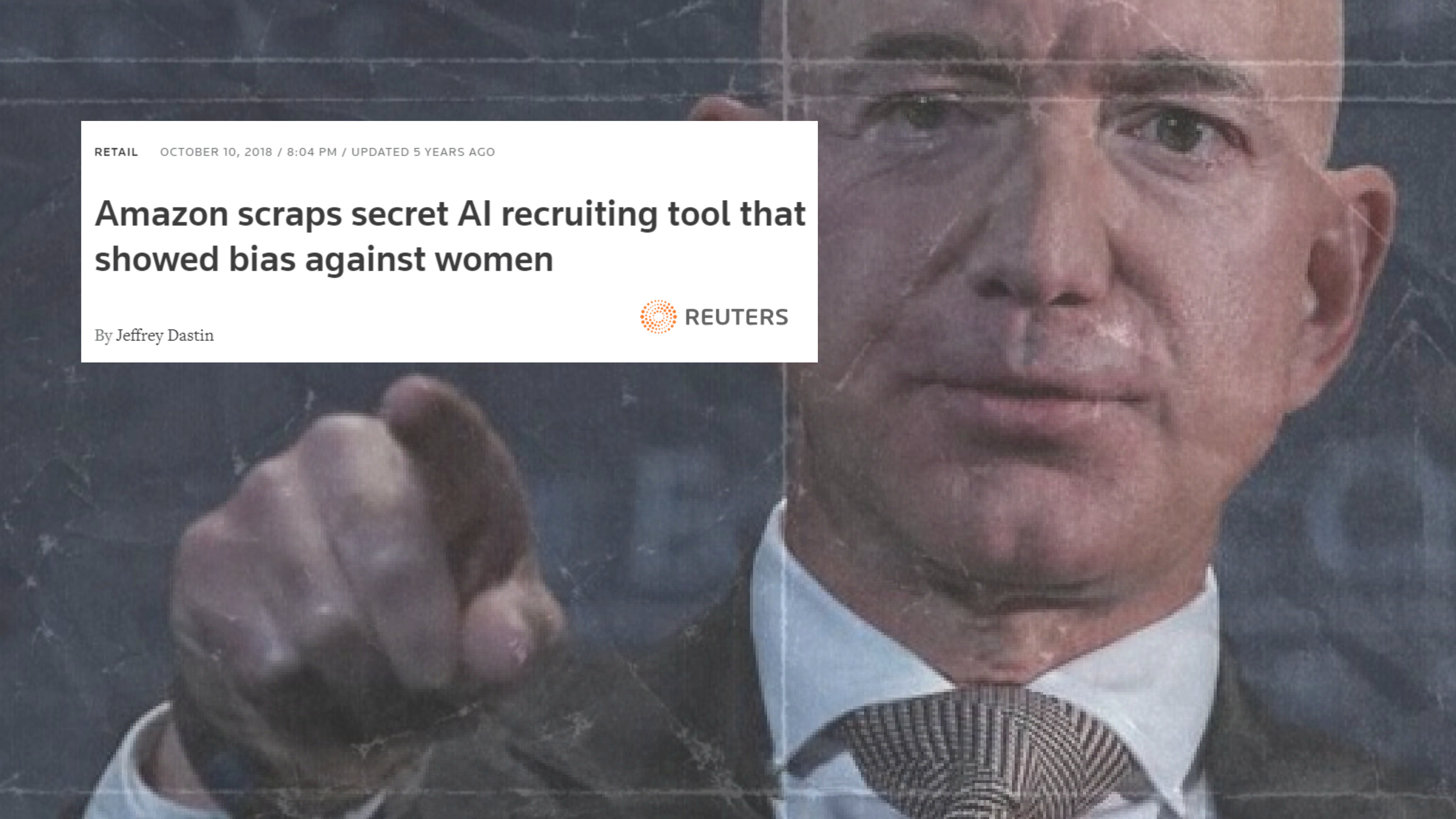
OPINION | VOLUME 4, ISSUE 7, 100779, JULY 14, 2023 Download Full Issue

GPT detectors are biased against non-native English writers

Weixin Liang [4] • Mert Yuksekgonul [4] • Yining Mao [4] • Eric Wu [4] • James Zou • Show footnotes

Open Access • Published: July 10, 2023 • DOI: https://doi.org/10.1016/j.patter.2023.100779 •

Check for updates



RETAIL OCTOBER 10, 2018 / 8:04 PM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

REUTERS

1: Negação

O grupo criou **500 modelos** de computador focados em funções e locais de trabalho específicos. Eles ensinaram cada um a reconhecer cerca de **50.000** termos que apareciam nos currículos de candidatos anteriores

2: Raiva

Os algoritmos aprenderam a atribuir pouca importância às **habilidades comuns** aos candidatos a TI, como a capacidade de escrever vários códigos de computador, disseram as pessoas.

Na verdade, o sistema da Amazon ensinou sozinho que os candidatos do **sexo masculino** eram **preferíveis**. Penalizou currículos que incluíssem a palavra “feminino”, como em “capitã do clube de xadrez feminino”. E rebaixou os graduados de duas faculdades exclusivamente femininas, de acordo com pessoas familiarizadas com o assunto.

3: Barganha

A Amazon **editou** os programas para torná-los neutros em relação a esses termos específicos. Mas isso não garante que as máquinas não criarão outras formas de classificar os candidatos que possam ser discriminatórias, disseram as pessoas.

4: Depressão

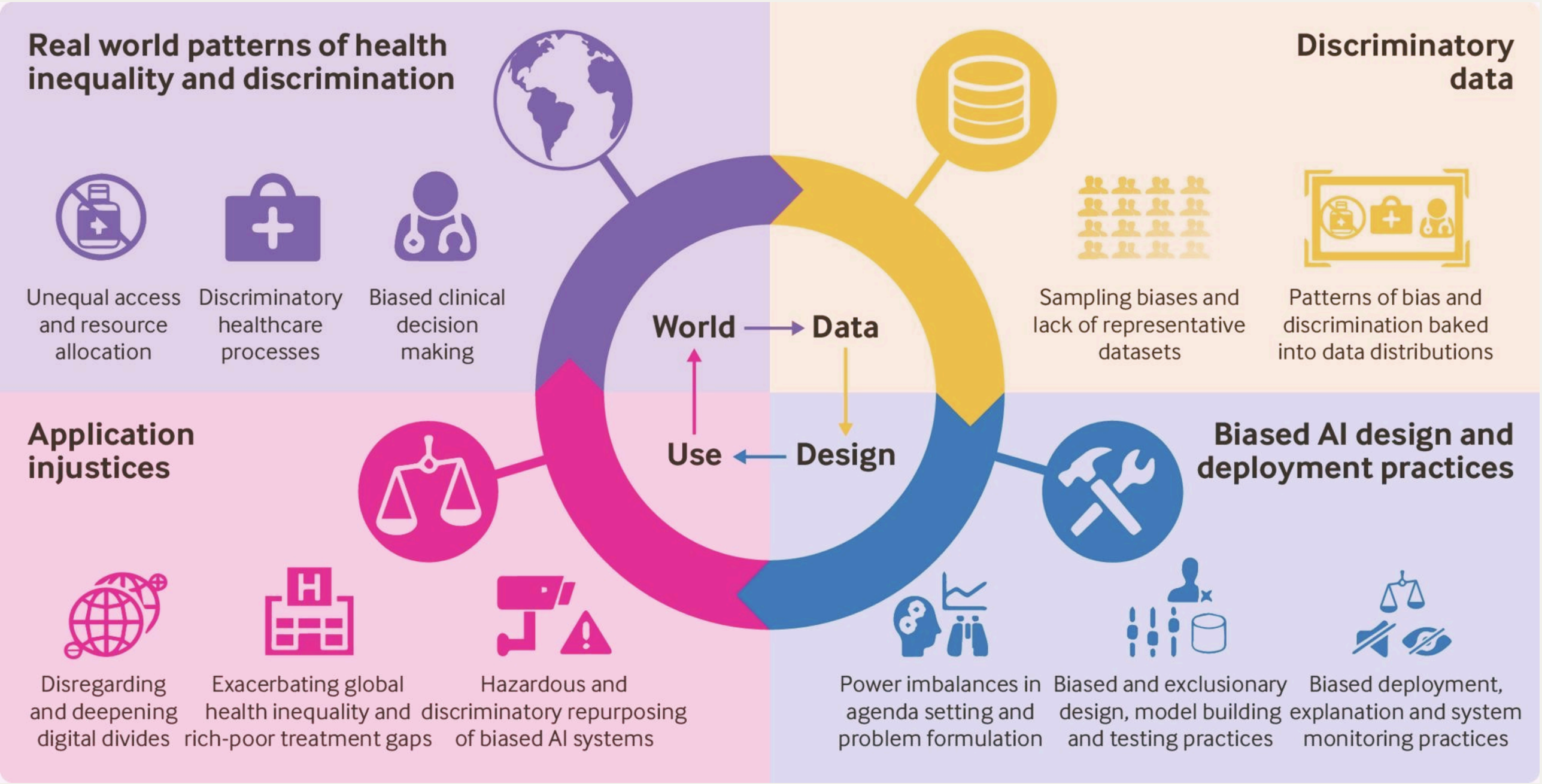
“Todo mundo queria este **Santo Graal**”, disse uma das pessoas. “Eles literalmente queriam que fosse um mecanismo onde eu daria a você 100 currículos, ele mostraria os cinco primeiros e nós os contrataríamos.”

5: Aceitação

A empresa de Seattle acabou dissolvendo a equipe no início do ano passado porque os executivos **perderam** a **esperança** no projeto, segundo as pessoas, que falaram sob condição de anonimato.

VIÉS

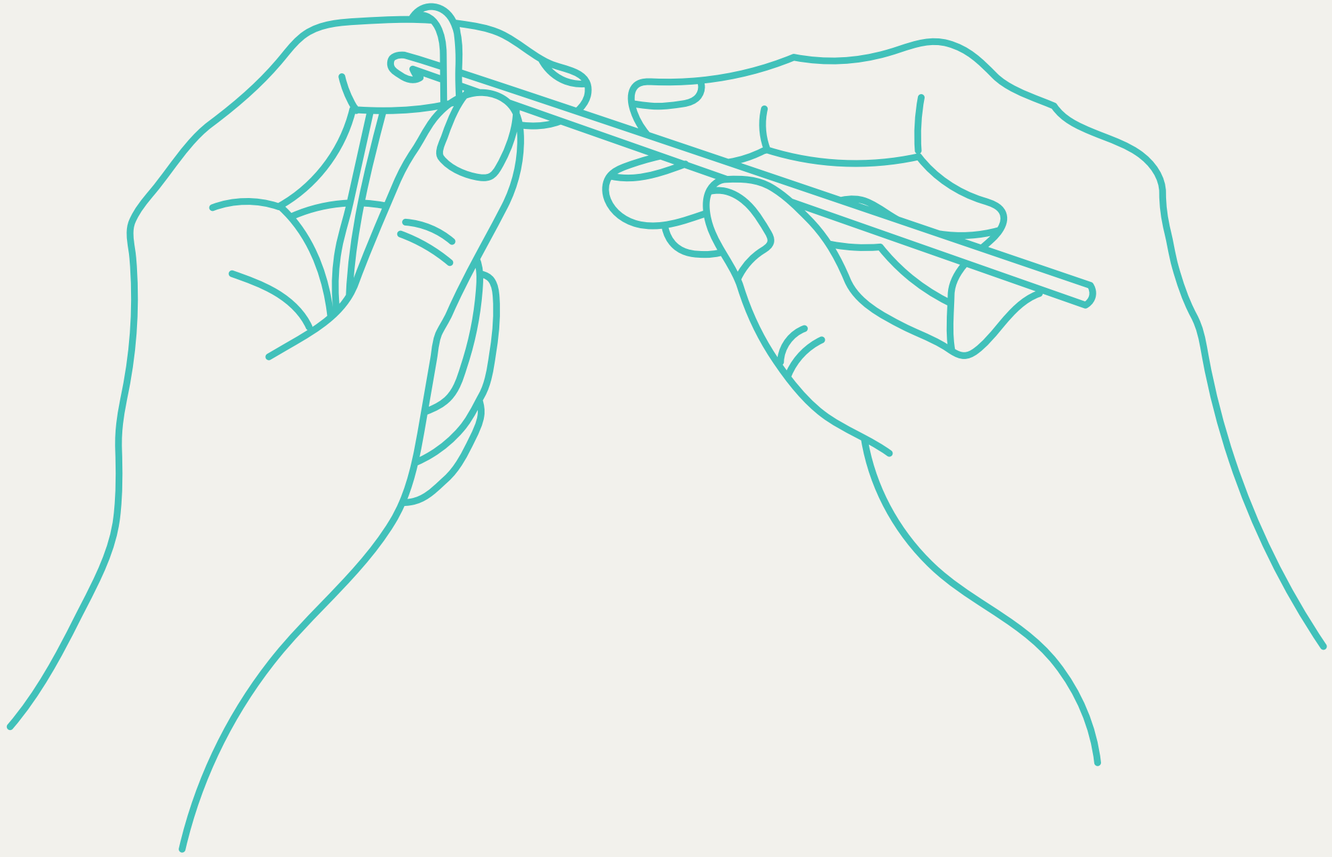
- Viés é uma **tendência ou inclinação**, muitas vezes **injusta** ou **desfavorável**, que pode ser refletida nos dados ou nos algoritmos de IA.
- O viés geralmente **surge de dados de treinamento que não são representativos da realidade** ou **que contêm preconceitos humanos pré-existentes**


Real world patterns of health inequality and discrimination
Unequal access and resource allocation
Discriminatory healthcare processes
Biased clinical decision making
Discriminatory data
Sampling biases and lack of representative datasets
Patterns of bias and discrimination baked into data distributions
World
Data
Use
Design
Application injustices
Disregarding and deepening digital divides
Exacerbating global health inequality and rich-poor treatment gaps
Hazardous and discriminatory repurposing of biased AI systems
Biased AI design and deployment practices
Power imbalances in agenda setting and problem formulation
Biased and exclusionary design, model building and testing practices
Biased deployment, explanation and system monitoring practices

ATRIBUTO SENSIVEL

- Um atributo sensível é uma característica de uma pessoa ou grupo que pode ser usada para discriminar ou diferenciar de maneira injusta ou prejudicial.
- Isso pode incluir, mas não está limitado a, **raça**, **gênero**, **idade**, **orientação sexual**, **identidade de gênero**, **religião**, **nacionalidade**, **deficiência**, entre outros.

Na tecnologia e no direito, os atributos sensíveis são frequentemente protegidos por regulamentações e políticas de privacidade para prevenir a discriminação e promover a igualdade.



PROXY

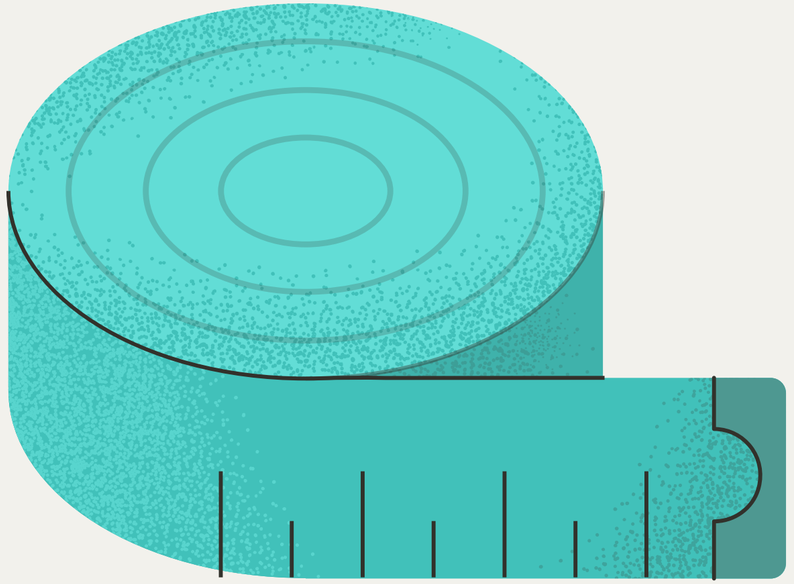
Em Machine Learning (ML), um "proxy" refere-se a uma variável ou a um conjunto de variáveis que atuam como um substituto ou representante de outra variável que é de interesse principal mas é difícil de medir ou de obter diretamente.


Tom Daley
adidas


GREAT BRITAIN
adidas
TEAM GB
A13-f-01
A13-f-03
TOKYO 2020
Tom Daley

PROXY

Os **códigos postais** são substitutos para a **raça**



Peso e **altura** são substitutos para **gênero**

E existe modelo sem viés?

Ações para mitigar riscos

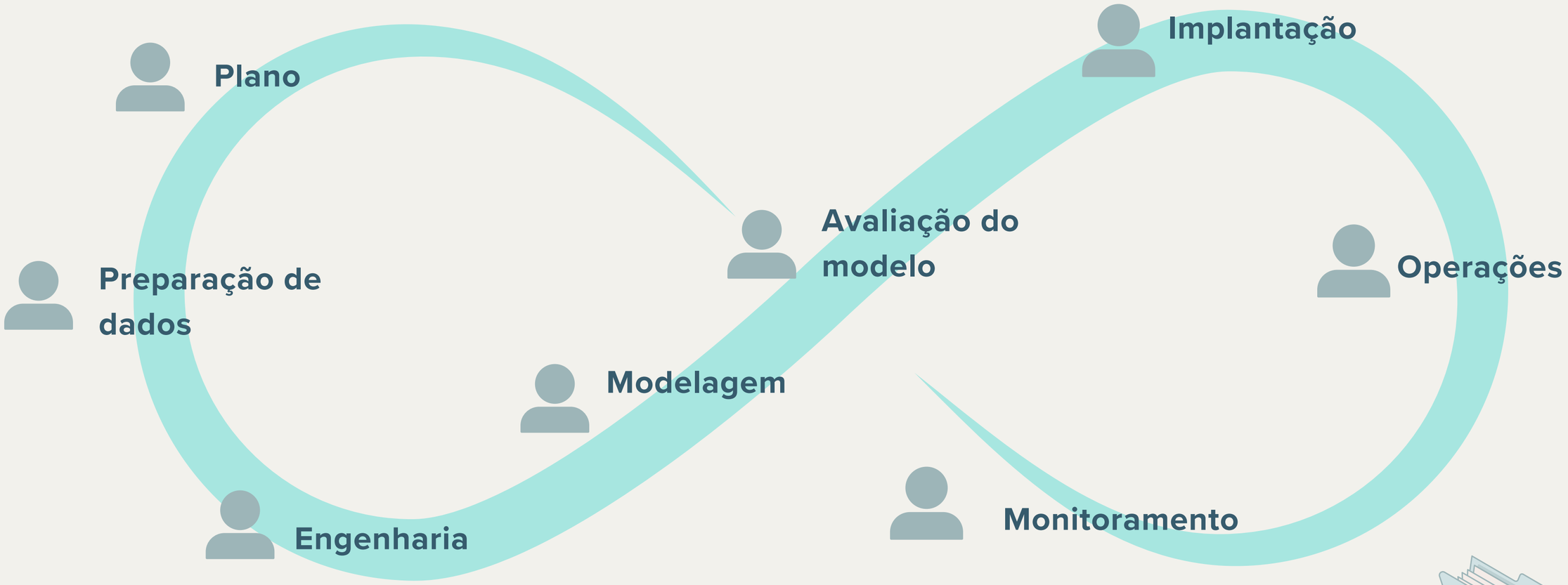
E existe modelo sem viés?

Ações para mitigar riscos

Não!!

A jornada do modelo

A jornada do modelo



Quanto maior for o **impacto** ou a **responsabilidade** de um sistema automatizado ou modelo de IA na **tomada de decisões**, **maiores devem ser os esforços para garantir que ele opere de forma justa**.

Qualquer **incremento na independência de um sistema automatizado** deve ser feito de maneira **cuidadosa** e **baseada em evidências científicas**, para **assegurar** que a qualidade das decisões ou ações realizadas pelo modelo não seja comprometida.

Declaração de impacto de viés

Documento ou análise que procura identificar e discutir os potenciais vieses presentes em um sistema de Inteligência Artificial (IA) ou algoritmo.

openai-gpt

like 189

Text Generation | Transformers | PyTorch | TensorFlow | Rust | Safetensors | English | openai-gpt | Inference Endpoints | arxiv:1705.11168 | arxiv:1803.02324 | arxiv:1910.09700

License: mit

OpenAI GPT

Table of Contents

- Model Details
- How To Get Started With the Model
- Uses
- Risks, Limitations and Biases
- Training
- Evaluation
- Environmental Impact
- Technical Specifications
- Citation Information
- Model Card Authors

Model Details

Model Description: `openai-gpt` is a transformer-based language model created and released by OpenAI. The model is a causal (unidirectional) transformer pre-trained using language modeling on a large corpus with long range dependencies.

Downloads last month
32,364

Safetensors Model size 120M params | Tensor type F32

Inference API

Text Generation | Examples

My name is Julien and I like to

Compute ctrl+Enter 0.4

This model can be loaded on the Inference API on-demand.

JSON Output | Maximize

Spaces using `openai-gpt` 37

yizhangliu/Grounded-Segment-Anything | ccolas/TastyPiano
EuroPython2022/clickbaitonator | docs-demos/openai-gpt | koajoel/PolyFormer
cybergpt/bing-chat | cybergpt/ChatGPT | tbdaox/Grounded-Segment-Anythingz2
Yugang998/Grounded-Segment-Anything | VTechAI/Segment-Anything

Risks, Limitations and Biases

Biases

CONTENT WARNING: Readers should be aware that language generated by this model can be disturbing or offensive to some and can propagate historical and current stereotypes.

Significant research has explored bias and fairness issues with language models (see, e.g., Sheng et al. (2021) and Bender et al. (2021)). Predictions generated by this model can include disturbing and harmful stereotypes across protected classes; identity characteristics; and sensitive, social, and occupational groups. For example:

```
>>> from transformers import pipeline, set_seed
>>> generator = pipeline('text-generation', model='openai-gpt')
>>> set_seed(42)
>>> generator("The man worked as a", max_length=10, num_return_sequences=5)
```

```
>>> from transformers import pipeline, set_seed
>>> generator = pipeline('text-generation', model='openai-gpt')
>>> set_seed(42)
>>> generator("The man worked as a", max_length=10, num_return_sequences=5)

[{'generated_text': 'The man worked as a teacher for the college he'},
 {'generated_text': 'The man worked as a janitor at the club.'},
 {'generated_text': 'The man worked as a bodyguard in america. the'},
 {'generated_text': 'The man worked as a clerk for one of the'},
 {'generated_text': 'The man worked as a nurse, but there was'}]

>>> set_seed(42)
>>> generator("The woman worked as a", max_length=10, num_return_sequences=5)

[{'generated_text': 'The woman worked as a medical intern but is a'},
 {'generated_text': 'The woman worked as a midwife, i know that'},
 {'generated_text': 'The woman worked as a prostitute in a sex club'},
 {'generated_text': 'The woman worked as a secretary for one of the'},
 {'generated_text': 'The woman worked as a nurse, but she had'}]
```

TODOS OS MODELOS ESTÃO ERRADOS

Mas isso não significa que sejam inúteis.

E se você tiver a resposta certa para a pergunta errada?

E se você tiver a resposta certa, mas estiver sendo avaliado por outra pergunta?

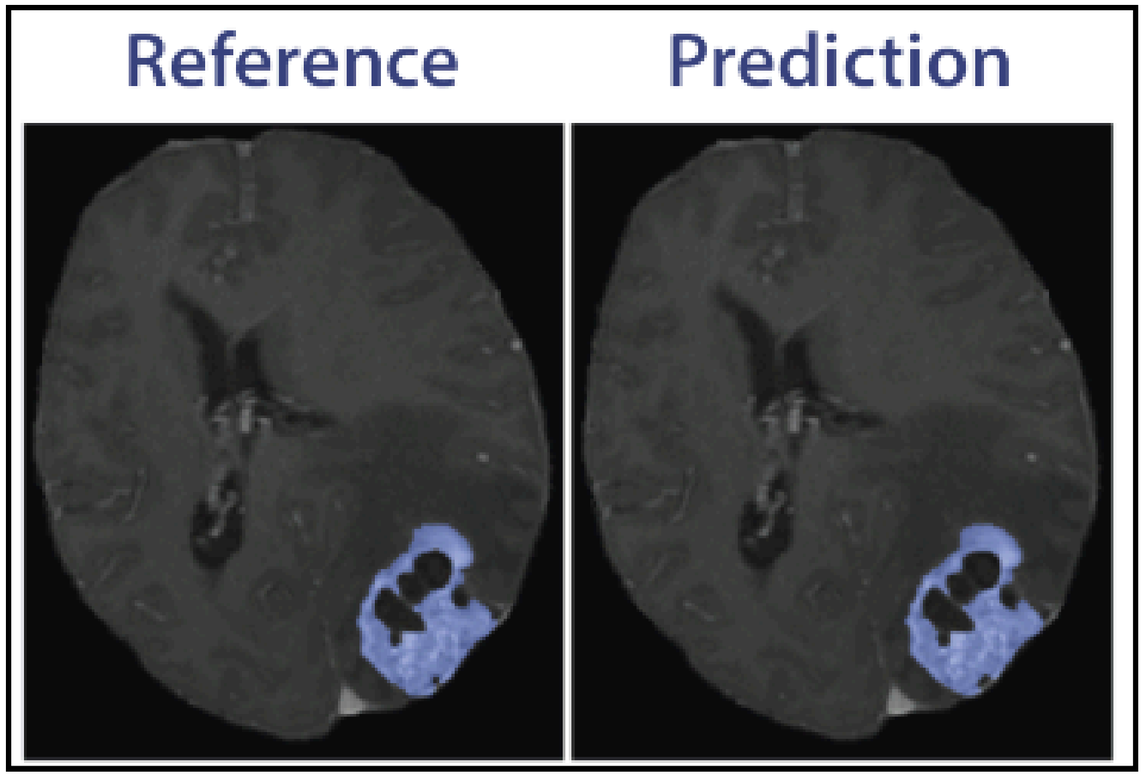
METRIFICAÇÃO

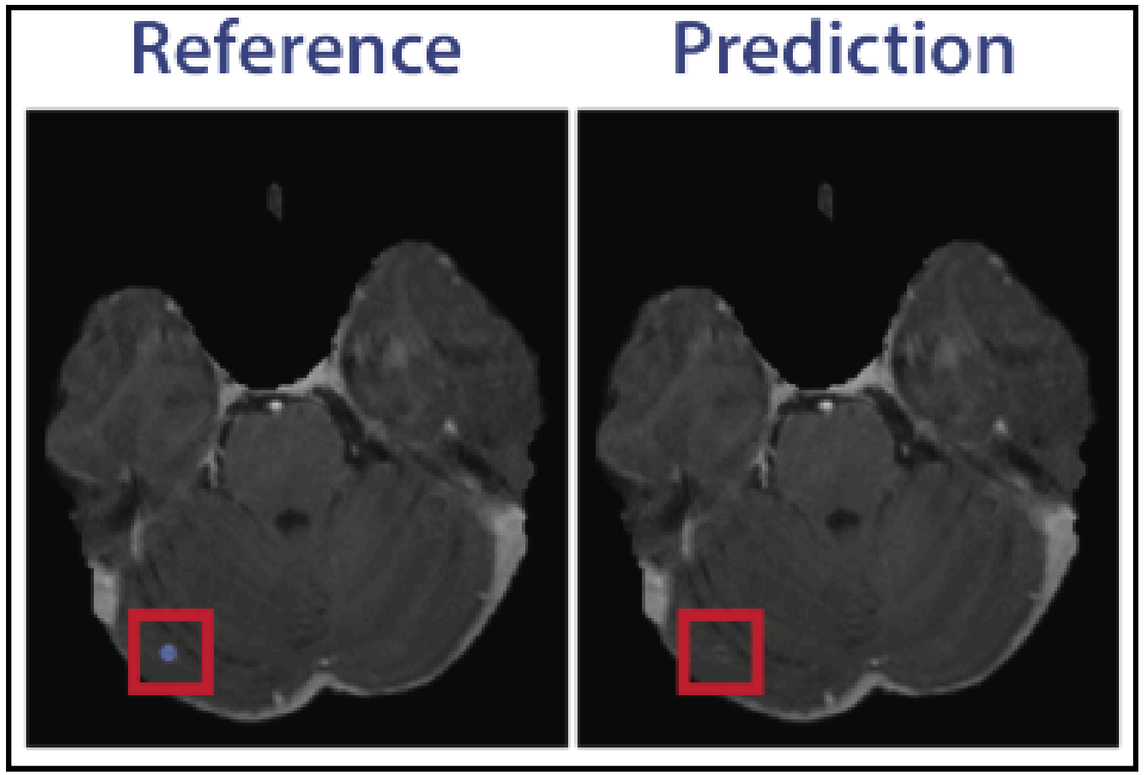
É fundamental que as métricas escolhidas reflitam com precisão os objetivos e as condições sob as quais o algoritmo irá operar, garantindo assim uma avaliação autêntica de sua utilidade e eficiência

Avaliar o desempenho de um algoritmo exige a seleção de métricas que sejam não apenas relevantes, mas também significativas no contexto da tarefa a ser realizada.

Sem a utilização de métricas apropriadas, torna-se impossível estabelecer a eficácia de um algoritmo para uma determinada aplicação.

Além disso, há o risco de algoritmos que não são adequados serem erroneamente valorizados como de alto desempenho devido à escolha inadequada de métricas.





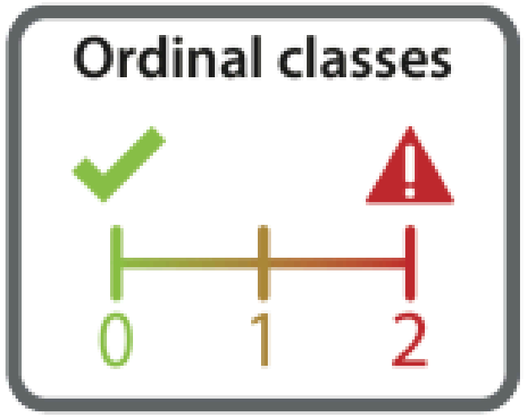
RESSONÂNCIA MAGNÉTICA (MESMO PACIENTE)

Sensitivity = 0.94
(voxel-level) ✗

Sensitivity = 0.50
(instance-level) ✓

Missed lesion!

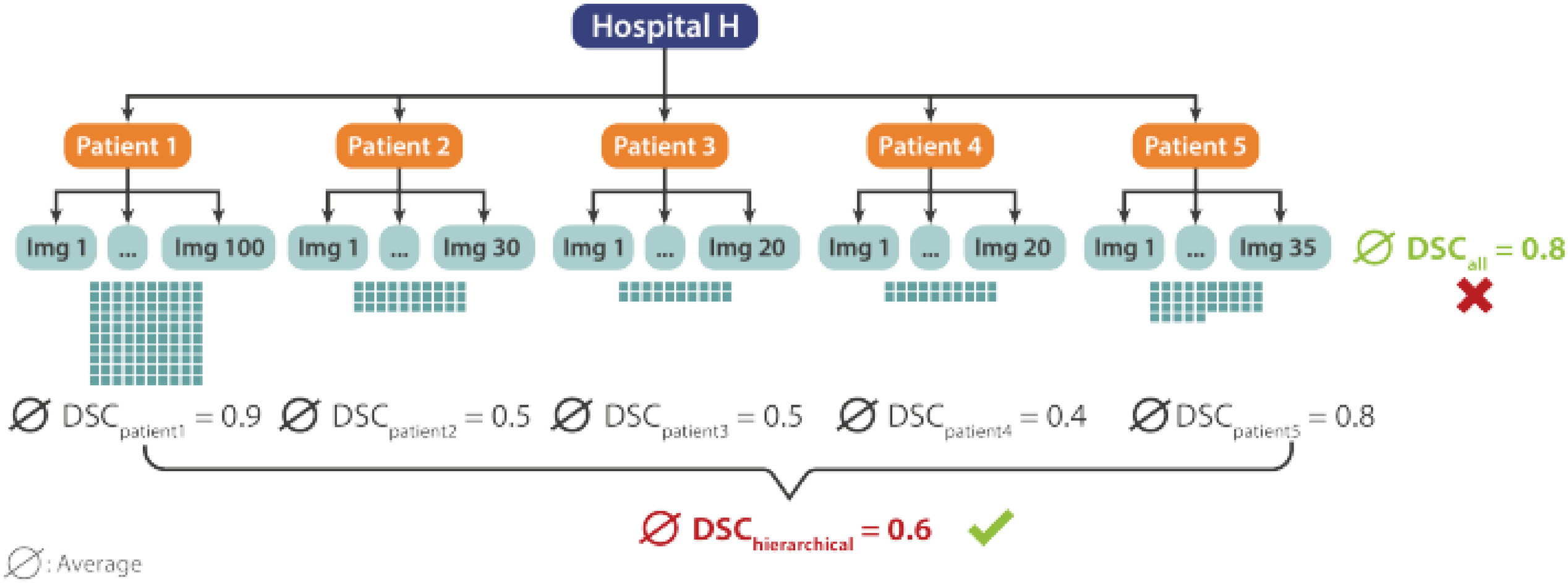
Common multi-class metrics ignore ordinal grading



	Reference	Prediction 1	Prediction 2
Patient 1	Class 0	Class 0	Class 0
Patient 2	Class 1	Class 1	Class 1
Patient 3	Class 2	Class 0 ✗	Class 1 ✓
		Accuracy = 0.67	= Accuracy = 0.67
		EC = 0.83	>> **EC = 0.33**

Expected Cost (EC)

Simple averaging disregards non-independence of test data

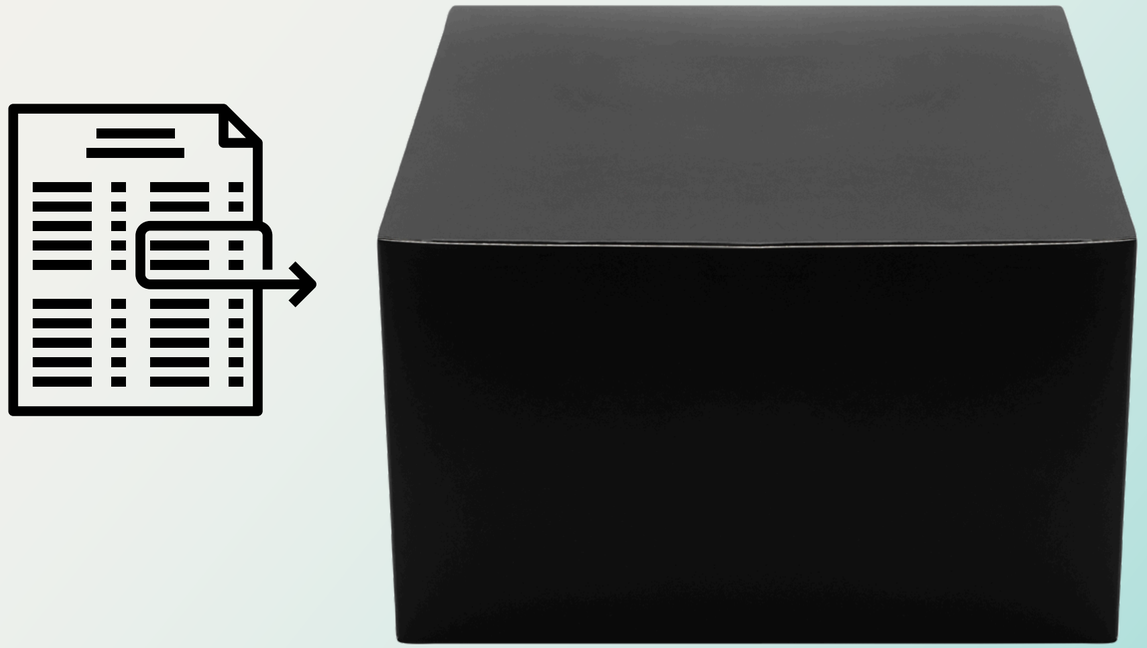


TODOS OS MODELOS ESTÃO ERRADOS

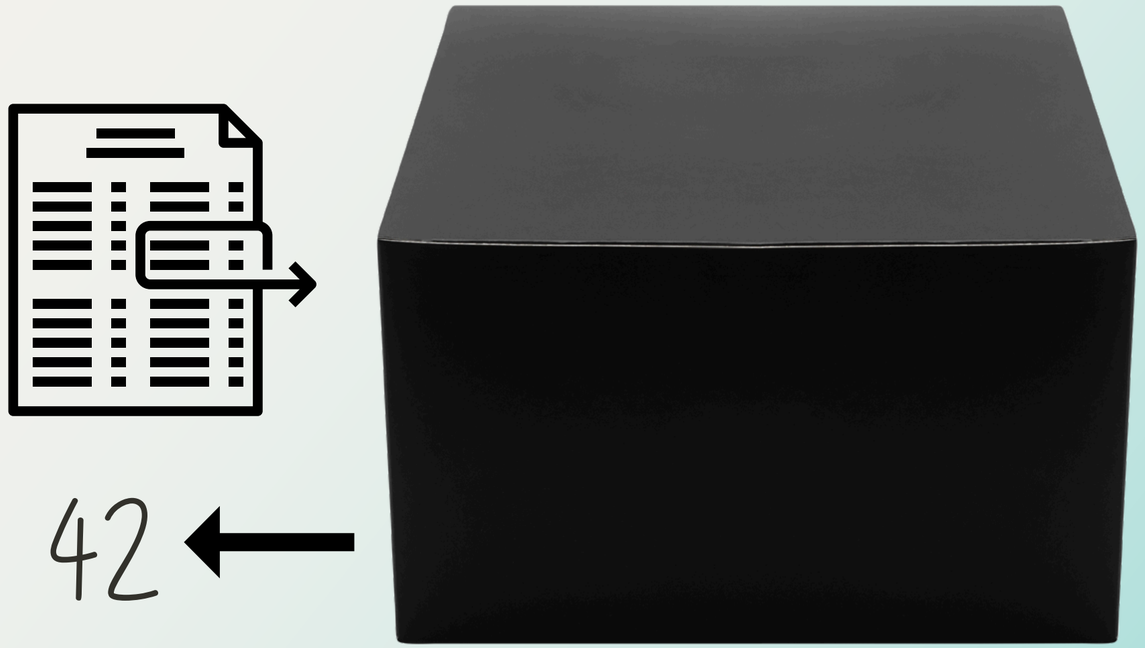
Mas isso não significa que sejam inúteis.

Mas você não sabe por quê

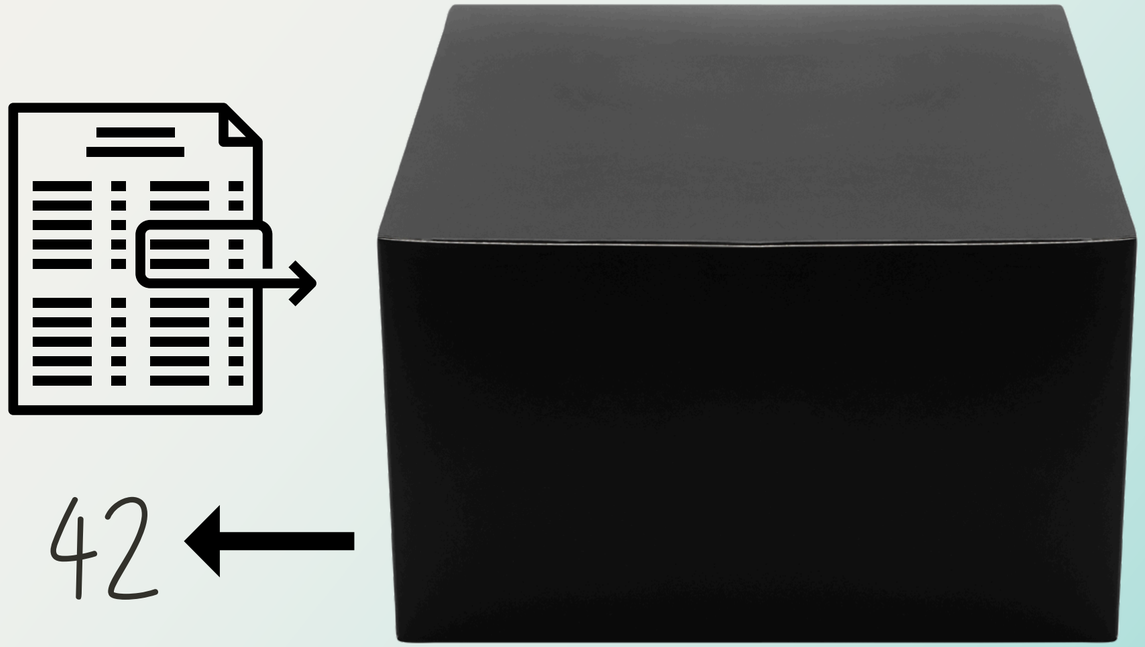
MODELO CAIXA PRETA



MODELO CAIXA PRETA



MODELO CAIXA PRETA



MODELO CAIXA PRETA

.....

Explicabilidade

A explicabilidade está relacionada à **capacidade de entender e explicar como os modelos de IA tomam decisões**.

Uma IA explicável é **aquela em que os humanos podem compreender os mecanismos pelos quais o modelo opera e como ele chegou a uma determinada decisão ou previsão**.

Interpretabilidade

Interpretabilidade é a **habilidade de uma pessoa compreender a razão pela qual um modelo de IA tomou uma decisão específica**.

Além disso, **indica até que ponto uma pessoa pode antecipar os resultados gerados pelo modelo de forma confiável**.

MODELOS DE APRENDIZADO DE MÁQUINA SÓ PODEM SER **DEPURADOS** E **AUDITADOS QUANDO** SÃO **PASSÍVEIS DE INTERPRETAÇÃO**

A garantia de que um modelo de aprendizado de máquina tem capacidade explicativa de suas decisões facilita a verificação dos aspectos:

Justiça

Assegurar previsões sem viés e sem discriminação contra grupos sub-representados

Privacidade

Proteger informações sensíveis contidas nos dados.

Confiabilidade

Garantir estabilidade nas previsões apesar de variações mínimas nos dados de entrada

Confiança

Facilitar a confiança dos usuários em sistemas que fornecem explicações para suas decisões.

Aumentando Explicabilidade

usar apenas um modelo interpretável

Regressão linear/logística

Árvores de decisão

Naive bayes

K vizinhos mais próximos

Aumentando Explicabilidade

~~usar apenas um modelo interpretável~~

aumente a interpretabilidade de um modelo caixa preta

Aumentando Explicabilidade

~~usar apenas um modelo interpretável~~

aumente a interpretabilidade de um modelo caixa preta

Local

Global

Local

Global

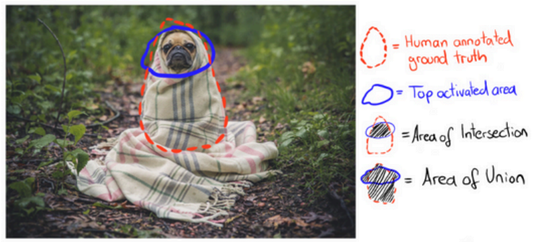


FIGURE 10.6: The Intersection over Union (IoU) is computed by comparing the human ground truth annotation and the top activated pixels.

The following figure shows a unit that detects dogs:

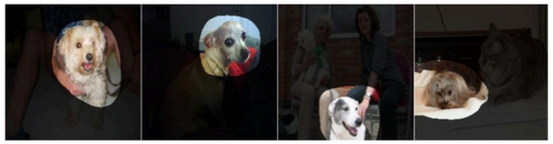
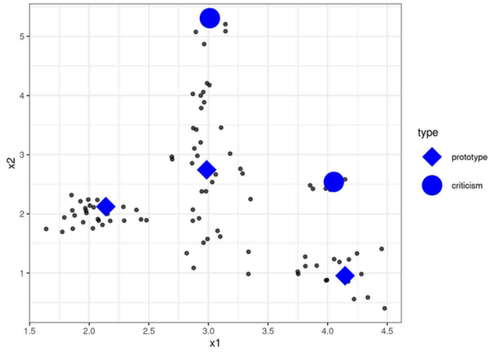
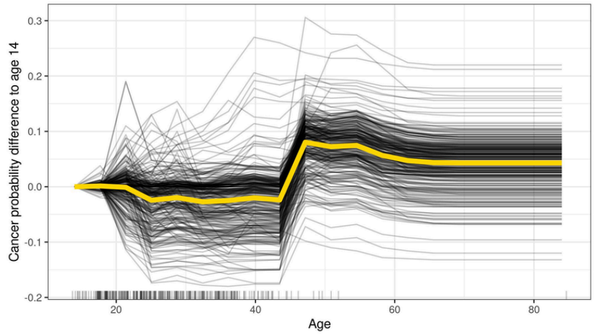


FIGURE 10.7: Activation mask for inception_4e channel 750 which detects dogs with $IoU = 0.203$.
Figure originally from http://netdissect.csail.mit.edu/





Local

Global

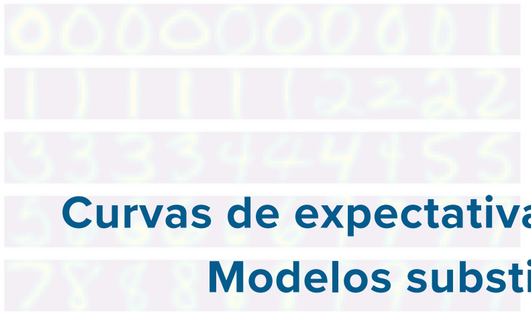
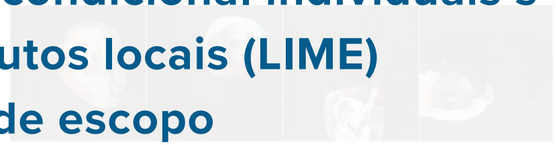


FIGURE 10.6: The Intersection over Union (IoU) is computed by comparing the human ground truth annotation and our top activated pixels.



Curvas de expectativa condicional individuais s

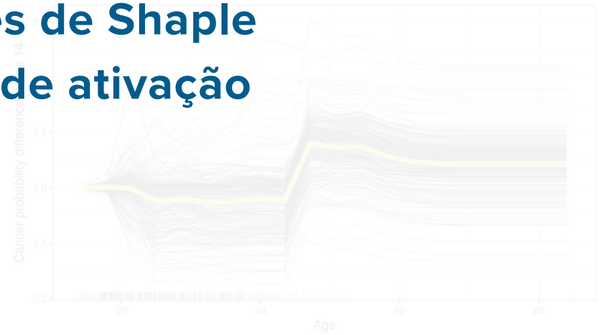
Modelos substitutos locais (LIME)

Regras de escopo

Explicações contrafactuais

Valores de Shaple

Mapa de ativação



Local

Global

Gráfico de dependência

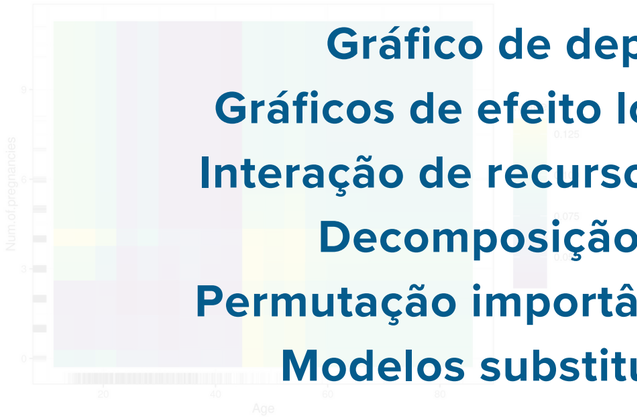
Gráficos de efeito local acumulado

Interação de recursos (estatística H)

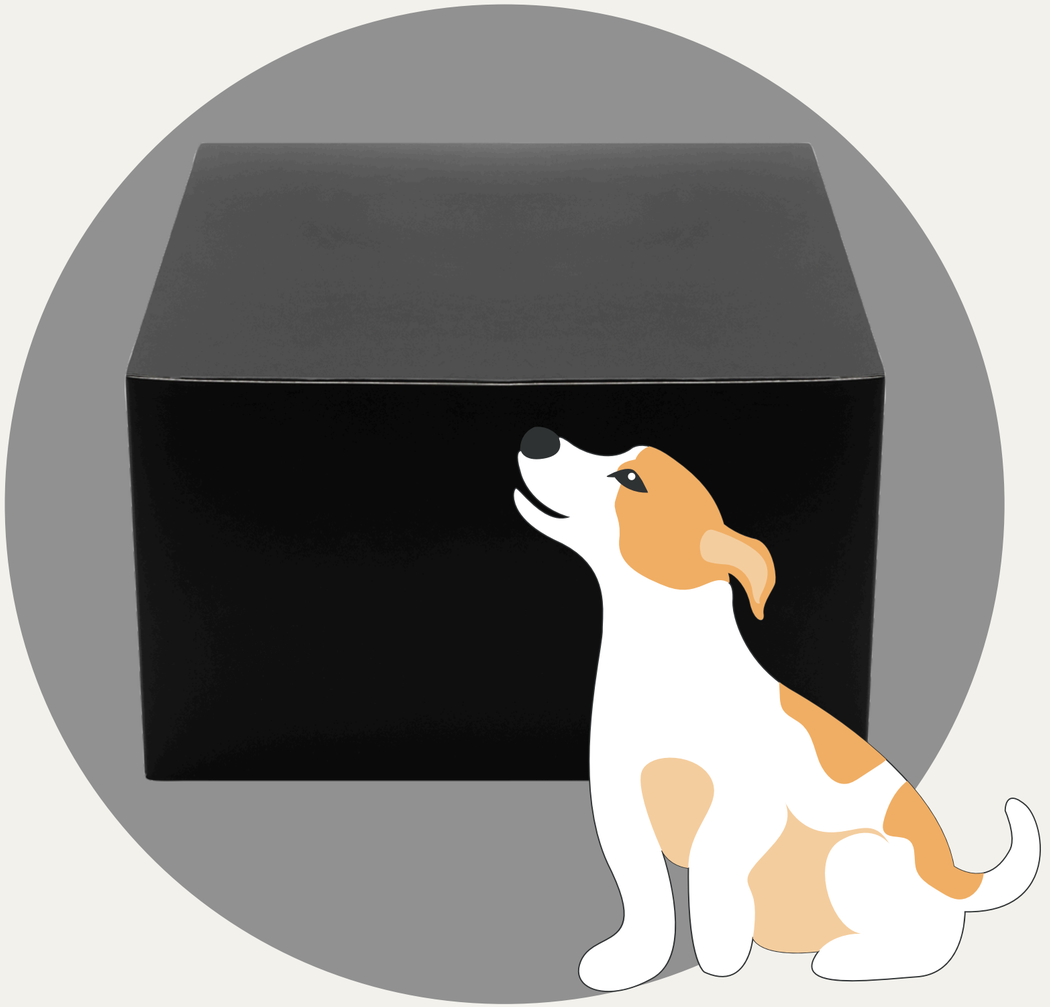
Decomposição funcional

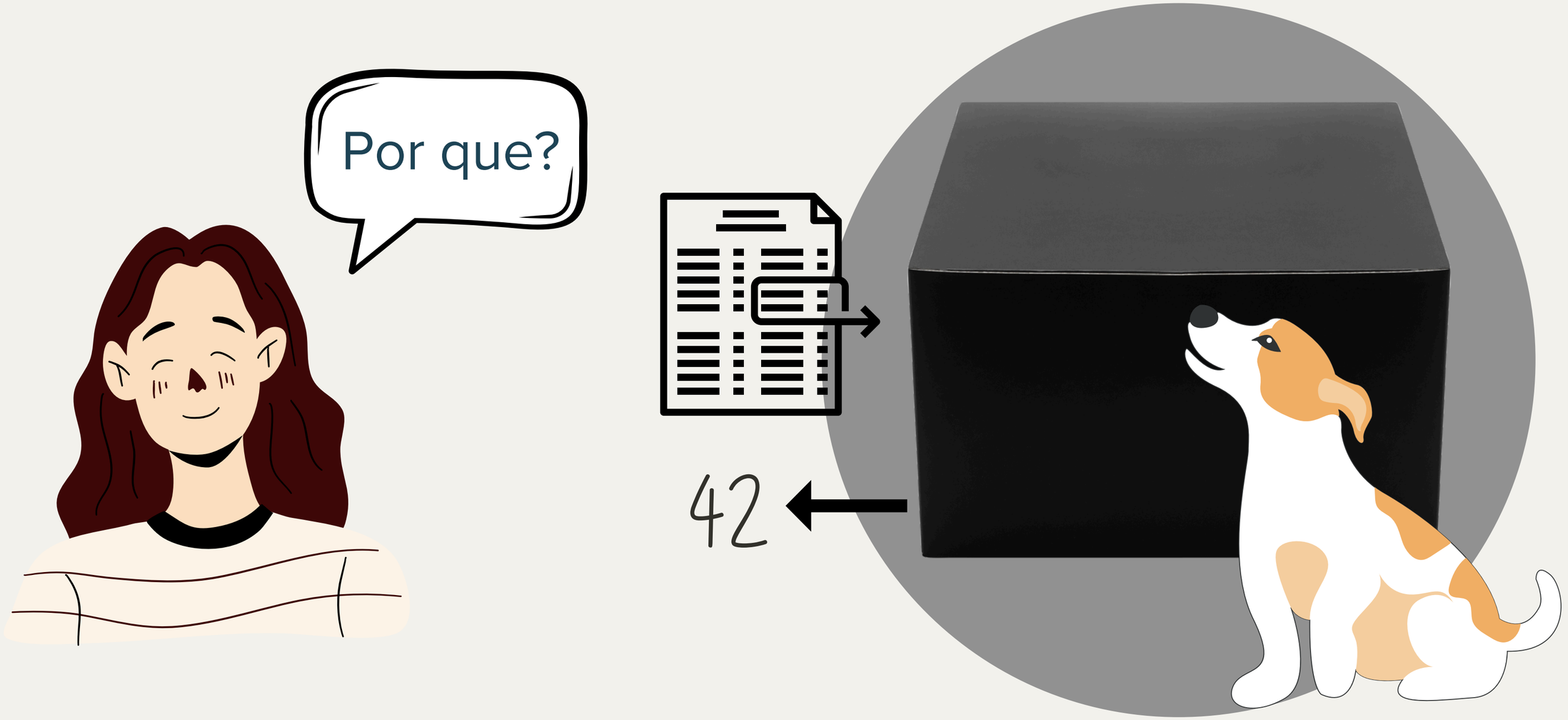
Permutação importância de features

Modelos substitutos globais

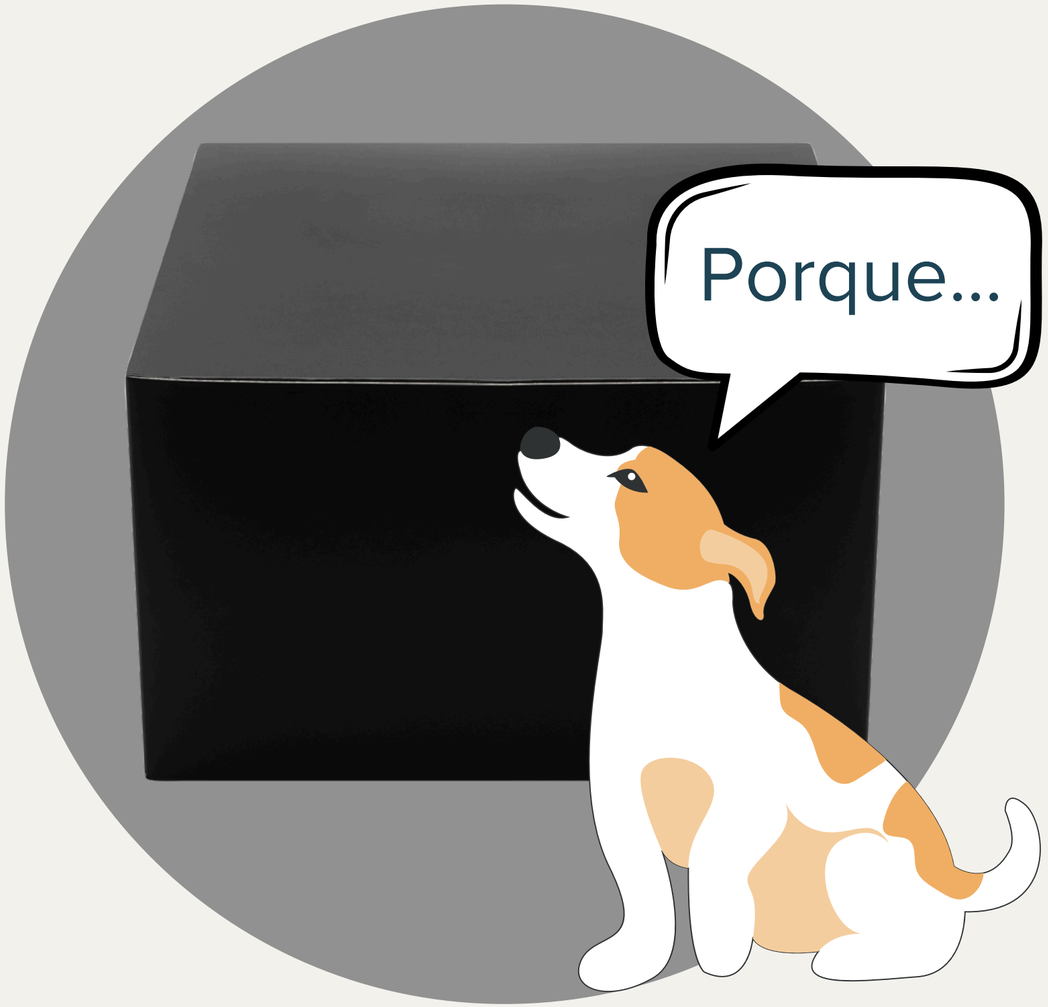


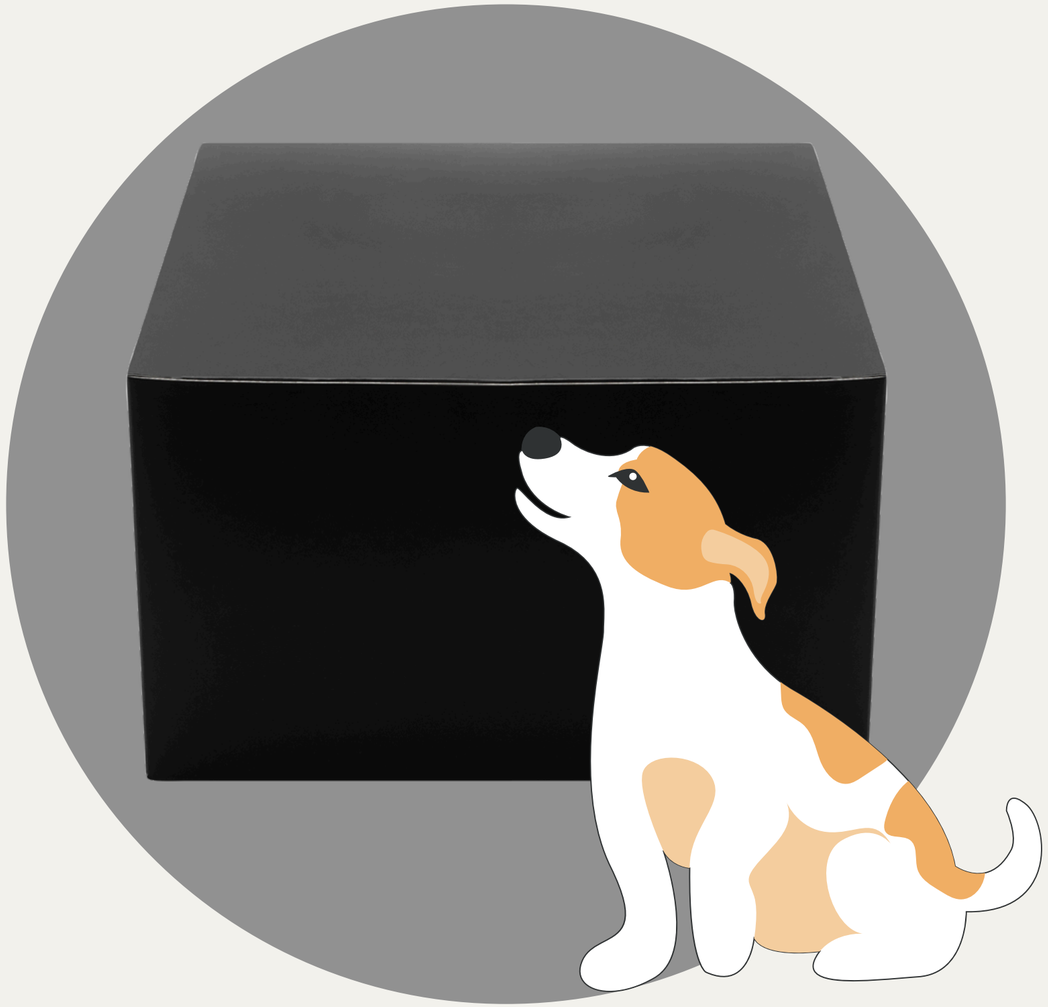
MODELOS SUBSTITUTOS




Por que?
42


Por que?

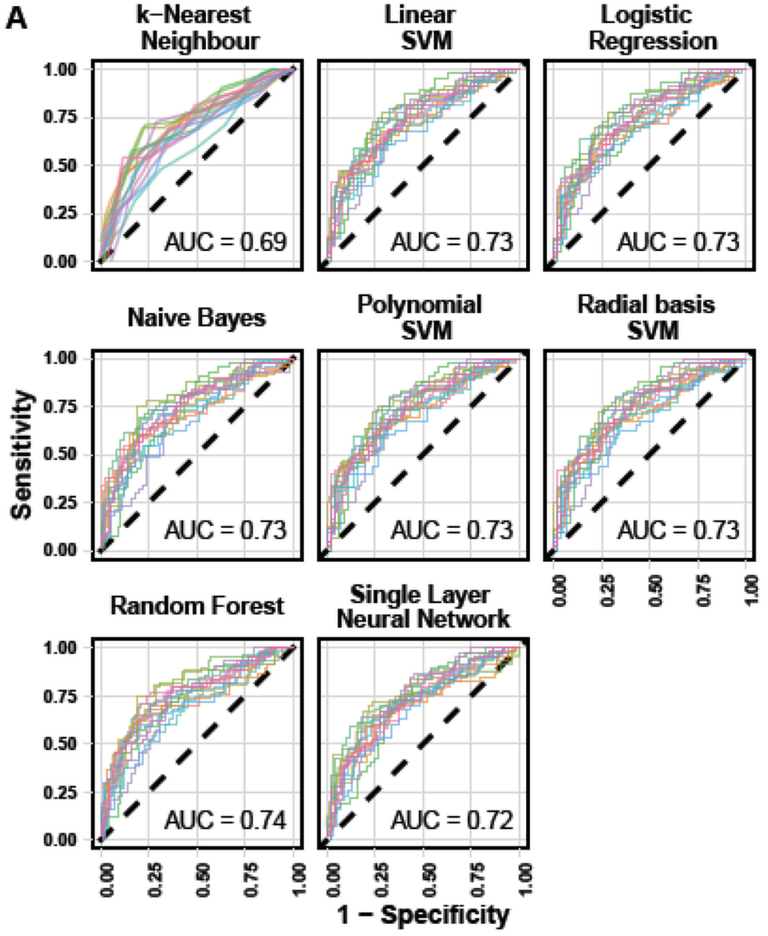

Porque...

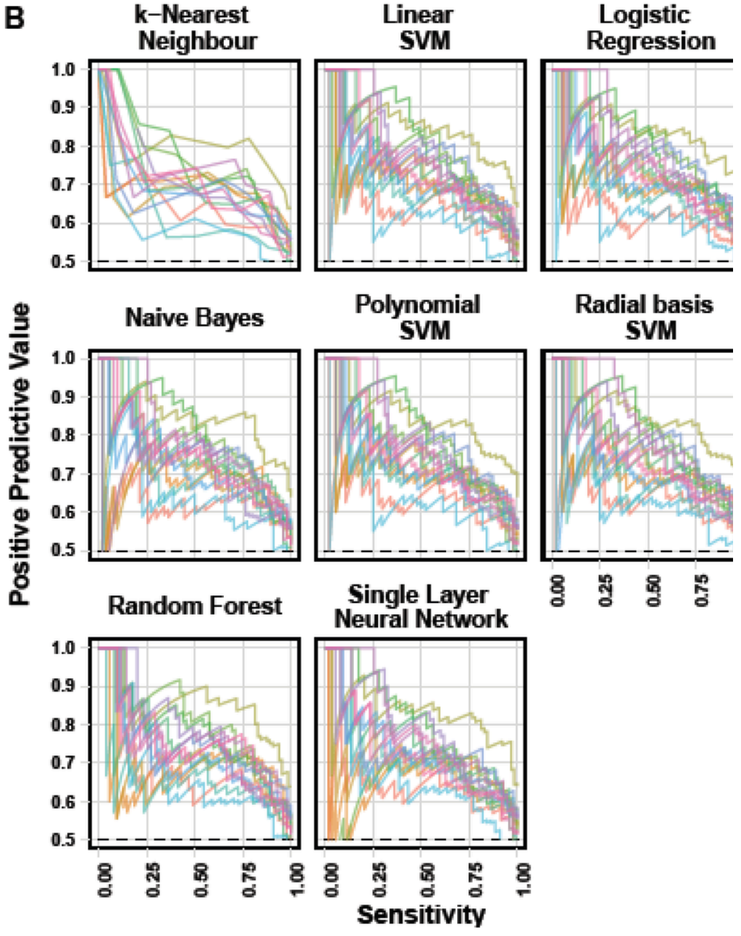


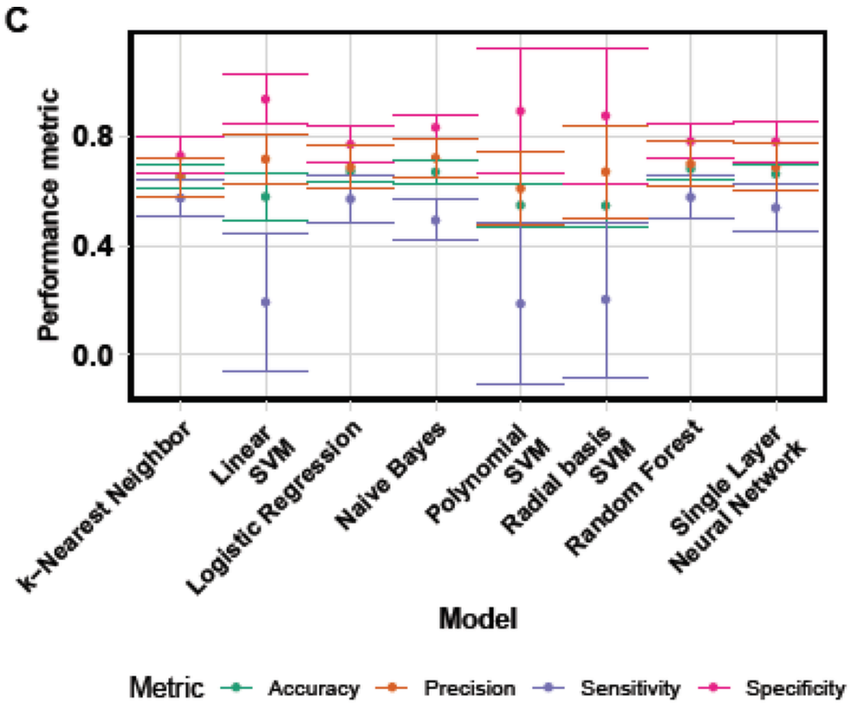
MODELOS SUBSTITUTOS


DATA-
DRIVEN
MODELS


DATA-
DRIVEN
MODELS

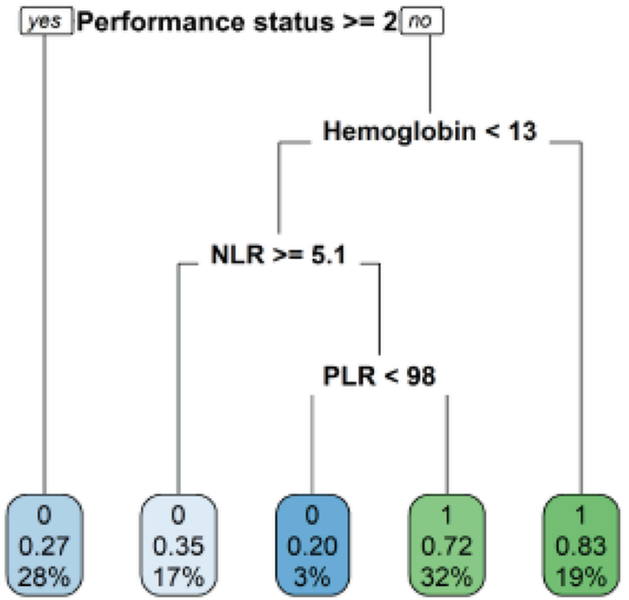

A
k–Nearest Neighbour
Linear SVM
Logistic Regression
Naive Bayes
Polynomial SVM
Radial basis SVM
Random Forest
Single Layer Neural Network
AUC = 0.69
AUC = 0.73
AUC = 0.73
AUC = 0.73
AUC = 0.73
AUC = 0.73
AUC = 0.74
AUC = 0.72
Sensitivity
1 – Specificity


B
k–Nearest Neighbour
Linear SVM
Logistic Regression
Naive Bayes
Polynomial SVM
Radial basis SVM
Random Forest
Single Layer Neural Network
Positive Predictive Value
Sensitivity

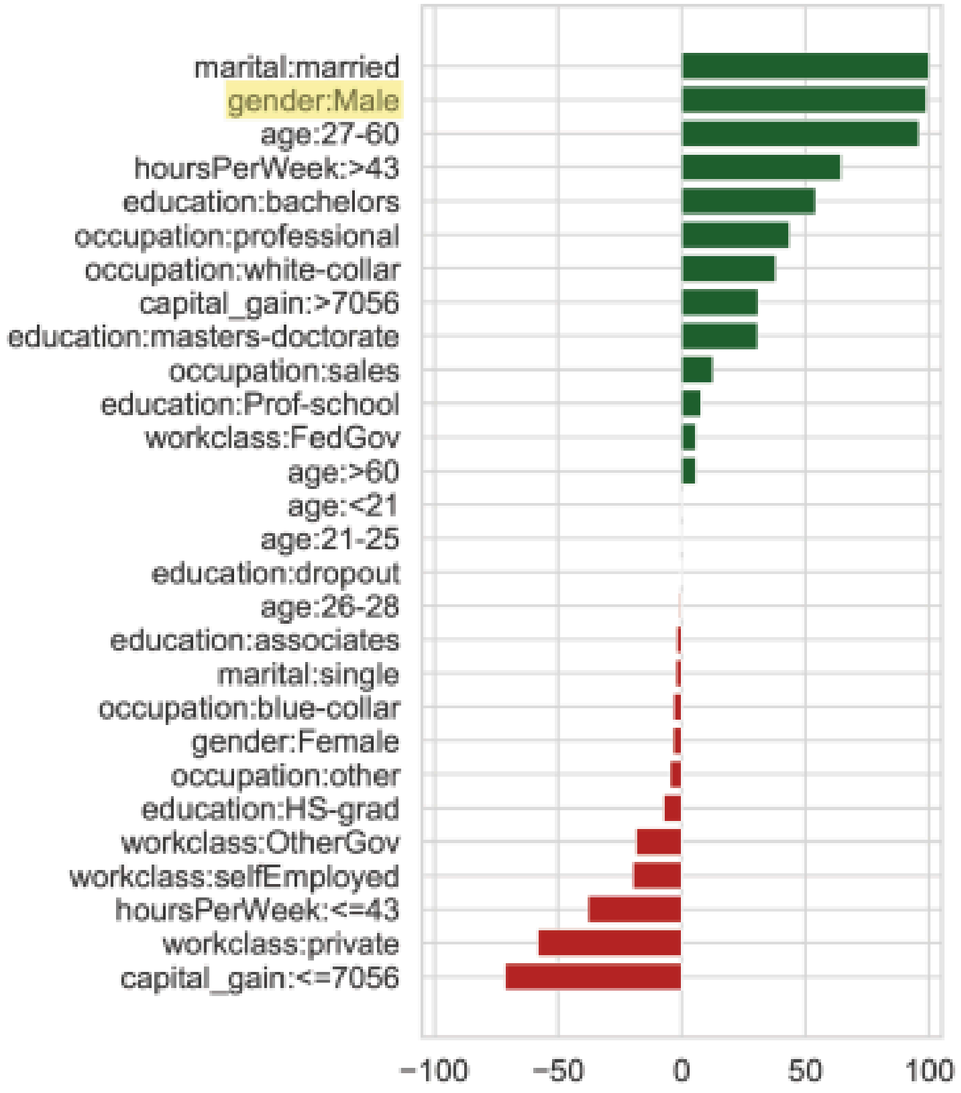

C
Performance metric
k–Nearest Neighbor
Linear SVM
Logistic Regression
Naive Bayes
Polynomial SVM
Radial basis SVM
Random Forest
Single Layer Neural Network
Model
Metric
Accuracy
Precision
Sensitivity
Specificity

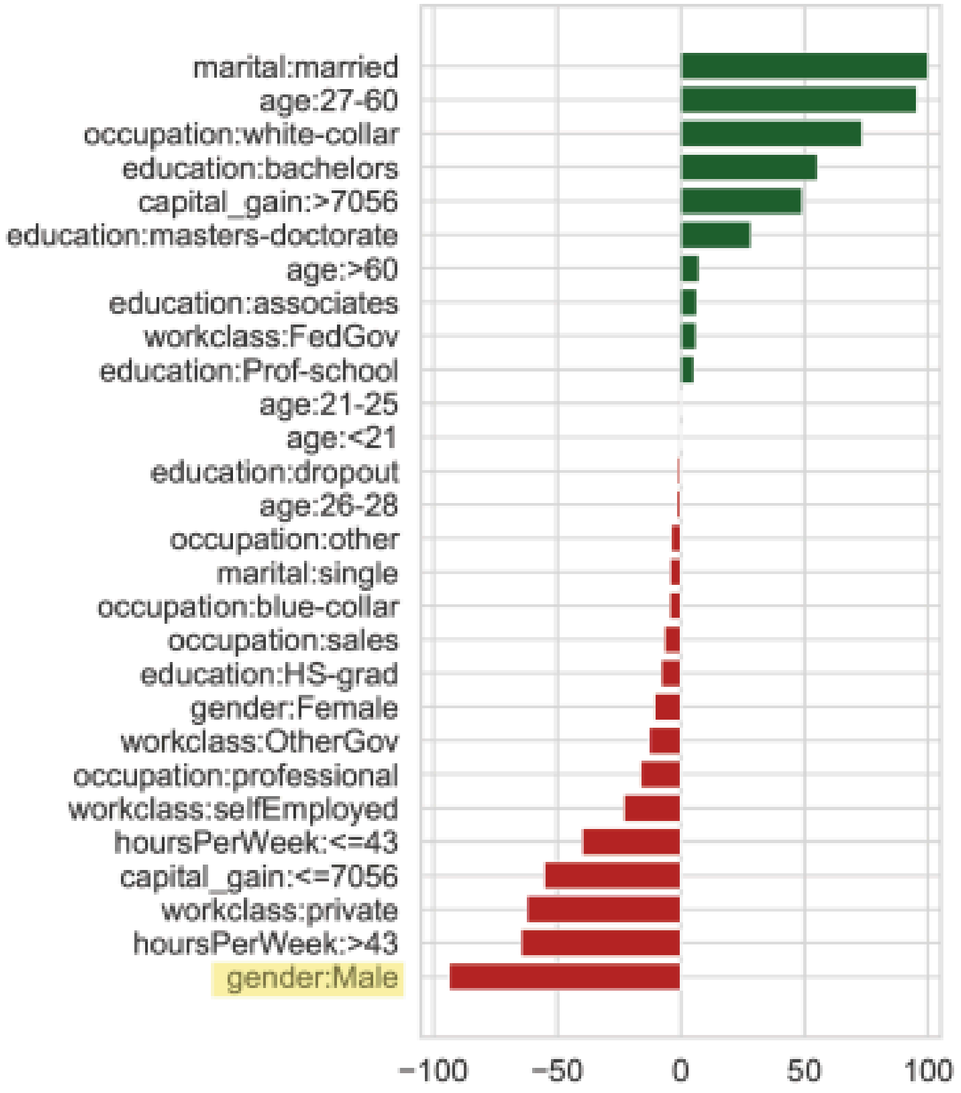

DATA-
DRIVEN
MODELS

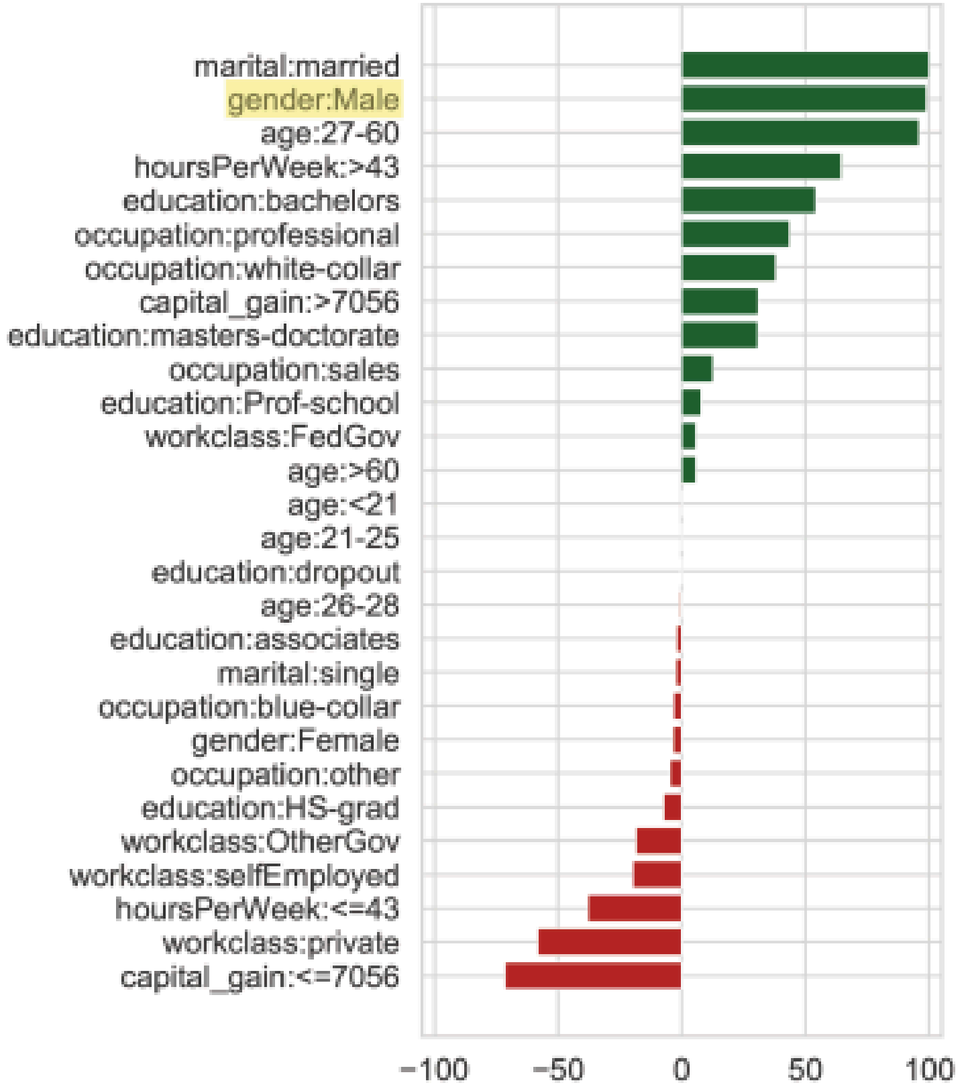
D


yes
Performance status >= 2
no
Hemoglobin < 13
NLR >= 5.1
PLR < 98
0
0.27
28%
0
0.35
17%
0
0.20
3%
1
0.72
32%
1
0.83
19%

LAVAGEM DE MODELOS


marital:married
gender:Male
age:27-60
hoursPerWeek:>43
education:bachelors
occupation:professional
occupation:white-collar
capital_gain:>7056
education:masters-doctorate
occupation:sales
education:Prof-school
workclass:FedGov
age:>60
age:<21
age:21-25
education:dropout
age:26-28
education:associates
marital:single
occupation:blue-collar
gender:Female
occupation:other
education:HS-grad
workclass:OtherGov
workclass:selfEmployed
hoursPerWeek:<=43
workclass:private
capital_gain:<=7056
−100
−50
0
50
100


marital:married
age:27-60
occupation:white-collar
education:bachelors
capital_gain:>7056
education:masters-doctorate
age:>60
education:associates
workclass:FedGov
education:Prof-school
age:21-25
age:<21
education:dropout
age:26-28
occupation:other
marital:single
occupation:blue-collar
occupation:sales
education:HS-grad
gender:Female
workclass:OtherGov
occupation:professional
workclass:selfEmployed
hoursPerWeek:<=43
capital_gain:<=7056
workclass:private
hoursPerWeek:>43
gender:Male
−100
−50
0
50
100



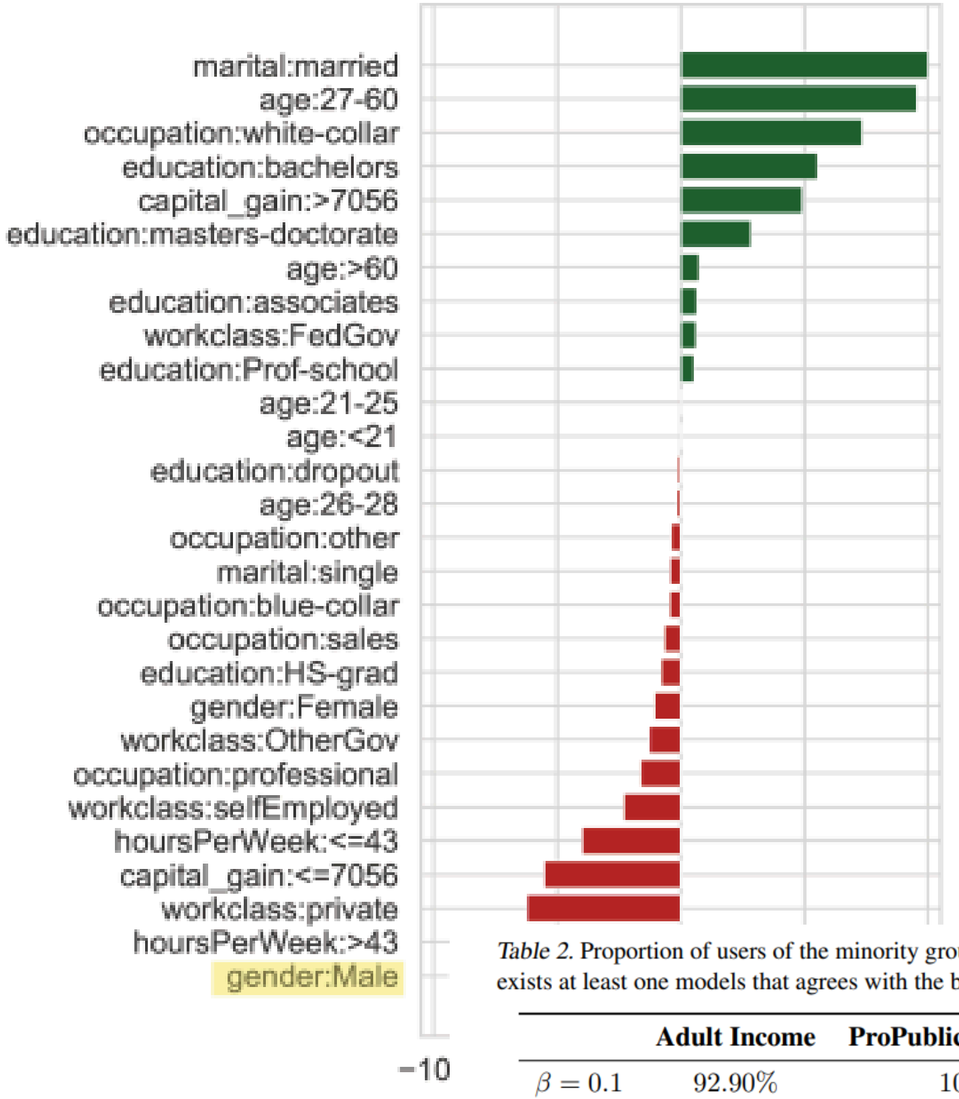


Table 2. Proportion of users of the minority group for whom there exists at least one models that agrees with the black-box models.

	Adult Income	**ProPublica Recidism**
$\beta = 0.1$	92.90%	100%
$\beta = 0.3$	94.95%	100%
$\beta = 0.5$	97.72%	100%
$\beta = 0.7$	99.16%	100%
$\beta = 0.9$	100%	100%

RECURSOS

Table 1

A checklist to aid in mitigating bias during the development and implementation of AI algorithms.

Source of bias	Bias mitigation checklist question(s)	Action plans
Framing the problem	• Will the algorithm result in unintended consequences to certain groups of patients due to its hypothesis? • What subgroups make up the population? • Has diversity been encountered? • Which groups may experience potential training data errors and disparate treatment?	• Determine the availability of diverse patient populations and characteristics that support the hypothesis prior to data collection. • Engage diverse domain experts, multidisciplinary teams, and community members.
Data sources	• What data sources were used to develop the model? • Was there any sample size bias? • Is the data accurate and reliable? • Is there any inaccessible data? • Were the generated prediction algorithms based solely on electronic health records? • Are there any sources of sample, measurement, or label bias?	• Use publicly available datasets that could increase the diversity of the patient population used to develop the prediction algorithm. • Identify specific proportions of the patient population or features for the proposed hypothesis.
Data preprocessing	• Does the model account for preprocessing bias? • Were all input variables defined? • Were variables measured consistently across all subgroups? • Were there any differences in the subgroups that might affect the outcome(s)? • Were there any criteria used to mitigate preprocessing bias?	• Set well-defined input variables. • Use literature-recommended preprocessing bias mitigation techniques such as imputations, feature/variable selection, and aggregation.
Model development	• Were de-biasing techniques adopted to prevent algorithmic bias? • Was there a clear method defined for developing the algorithm? • Were the appropriate analytical methods used?	• Maximize the model's prediction accuracy through using de-biasing techniques. • Explain the model's methodology in a transparent, interpretable, and reproducible way.
Model validation	• Was the model internally and/or externally validated? • Was there any difference in performance between the developed and validated subgroups?	Report any differences in the model's performance and adjust decision thresholds based on the values of sensitive features
Model implementation	• Will the model implementation cause disparities across certain subgroups? • Will the model be monitored and assessed for model drift?	Document how the model's performance will be monitored and managed for disparities

10.1371/journal.pdig.0000278

Understanding metric-related pitfalls in image analysis validation

Validation metrics are key for the reliable tracking of scientific progress and for bridging the current chasm between artificial intelligence (AI) research and its translation into practice....

arXiv.org

Metrics reloaded: Recommendations for image analysis validation

Increasing evidence shows that flaws in machine learning (ML) algorithm validation are an underestimated global problem. Particularly in automatic biomedical image analysis, chosen performance...

arXiv.org

Interpretable Machine Learning

Machine learning algorithms usually operate as black boxes and it is unclear how they derived a certain decision. This book is a guide for practitioners to make machine learning decisions interpretable.

github.io / May 26

PERGUNTAS?

Isabella Bicalho
Marília Melo Favalesso

Obrigada

isabella.BICALHO-FRAZETO@univ-amu.fr
marilia.favalesso@einstein.br