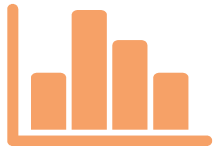
PROGRAMA de VERÃO
Faculdade de Saúde Pública USP

Discriminação Algorítmica

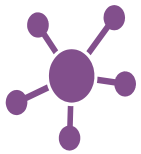
Prof. Dr. André Filipe M. Batista
Mariane Furtado Borba, Doutoranda no Programa de Pós-Graduação em Epidemiologia, FSP/USP
Pesquisadores do Laboratório de Big Data e Análise Preditiva em Saúde da Faculdade de Saúde Pública da USP

Avanços Científicos



- Grande quantidade de dados disponíveis online

- Computação de baixo custo



- Surgimento de novos algoritmos visando problemas cada vez mais complexos

- Necessidade de profissionais capacitados

Machine Learning

- O que é?
  - Aprendizado por meio do reconhecimento de padrões (aprendizado infantil ➡ análise de dados)

Categorias de Machine Learning

Como funciona?

Dados

- Preditores Fortes
- Qualidade dos dados

Algoritmos

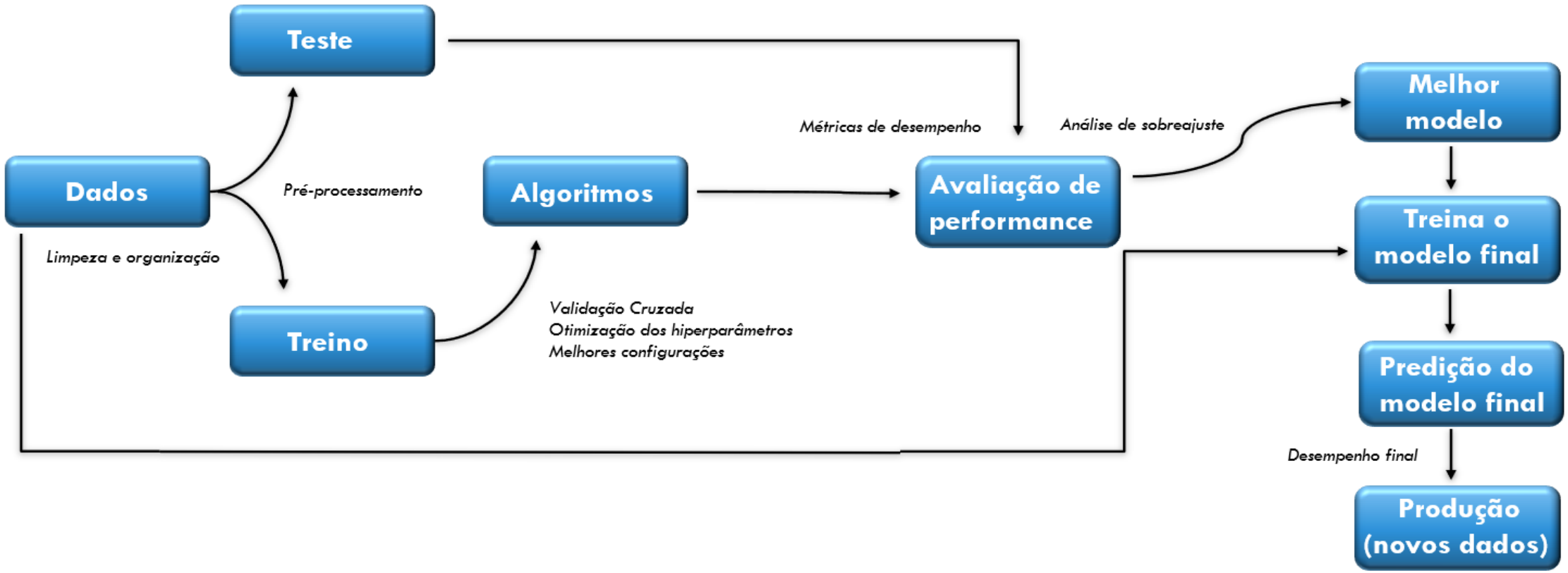
- Inserção dos dados
- Aprende com parâmetros
- Mapeiam os preditores aos resultados

Teste de Qualidade

- Novos dados
- Predição

Como funciona?

WORKFLOW - APRENDIZADO SUPERVISIONADO



Como Interpretar?

SHAP (SHapley Additive exPlanations)

- Contribuição de cada variável para a predição (+ ou -)
- Base na teoria dos jogos
- Interpretar não é achar relação causal
- Fácil visualização

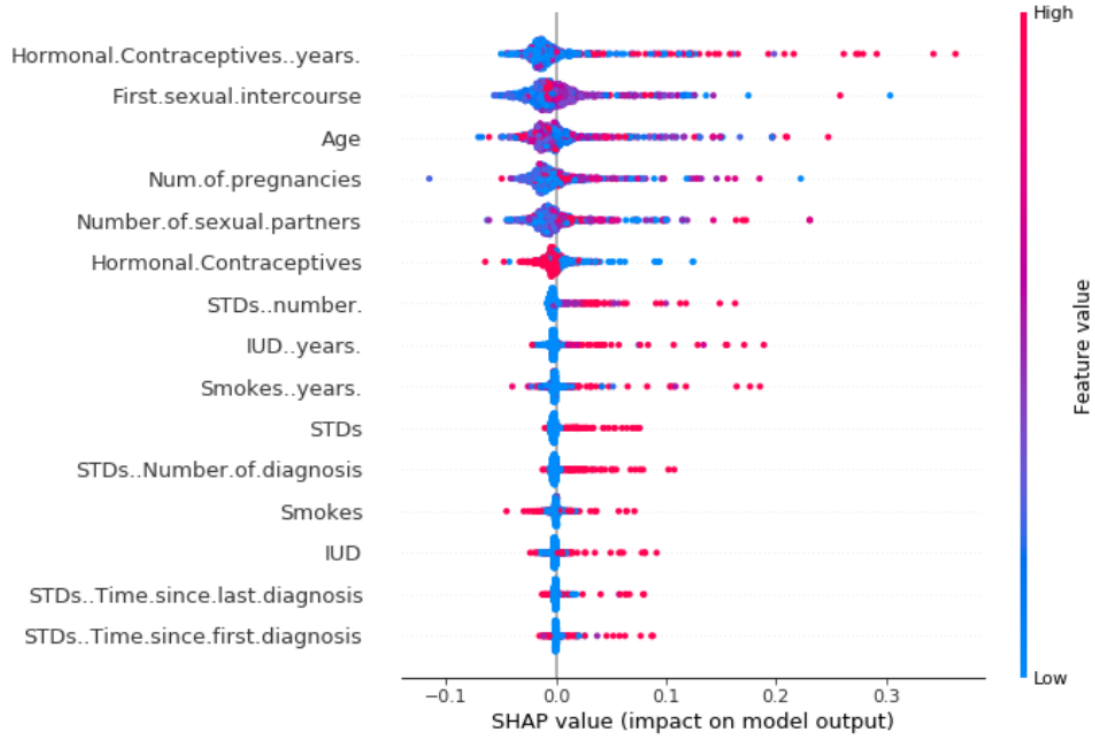


FIGURE 5.52: SHAP summary plot. Low number of years on hormonal contraceptives reduce the predicted cancer risk, a large number of years increases the risk. Your regular reminder: All effects describe the behavior of the model and are not necessarily causal in the real world.

https://christophm.github.io/interpretable-ml-book/shap.html

Benefícios do uso de machine learning

- Automatização de tarefas
- Otimização de tempo por meio de uma gama de análises simultâneas
- - tempo para o resultado
- - custos
- +eficiência

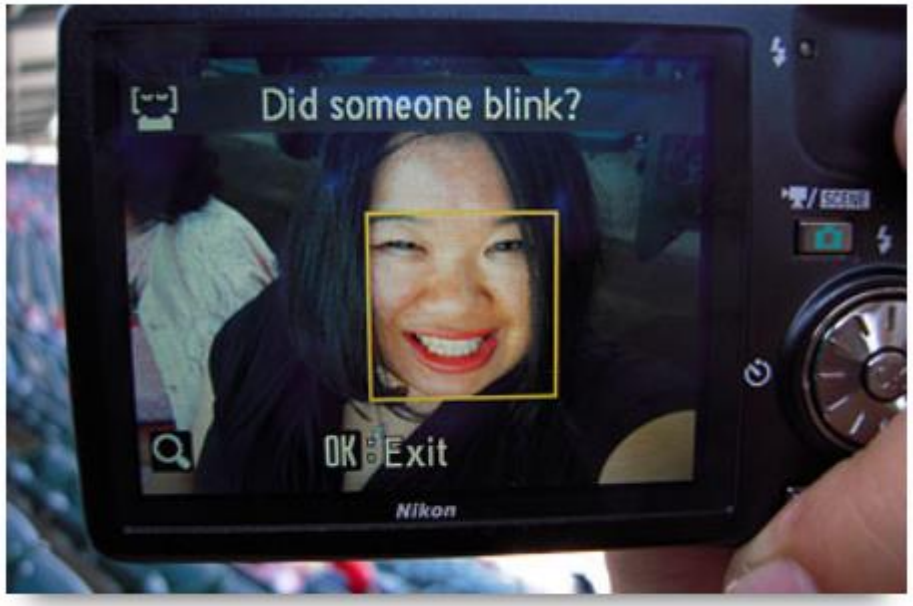
Para pensar...

- Resultados ruins existem por alguma razão
- Seu problema é mesmo preditivo?
- Resultados muito bons exigem cautela

Machine Learning também exige atenção!

- A máquina aprende padrões que nem sempre queremos reproduzir
- É necessário bom senso!
- Conhecer bem os dados é fundamental
- Modelos não devem ser implementados sem testes para a análise de possíveis resultados com vieses humanos

O que queremos evitar



JANEIRO 2010

CÂMERAS DA NIKON NÃO ENTENDEM ROSTOS ASIÁTICOS

Recurso para evitar selfies com olhos fechados se confunde com olhos de asiáticos

- SILVA, Tarcízio. Linha do Tempo do Racismo Algorítmico. **Blog do Tarcízio Silva**, 2020. Disponível em: <http://https://tarciziosilva.com.br/blog/posts/racismo-algoritmico-linha-do-tempo>. Acesso em: 03 de fevereiro de 2021.

O que queremos evitar

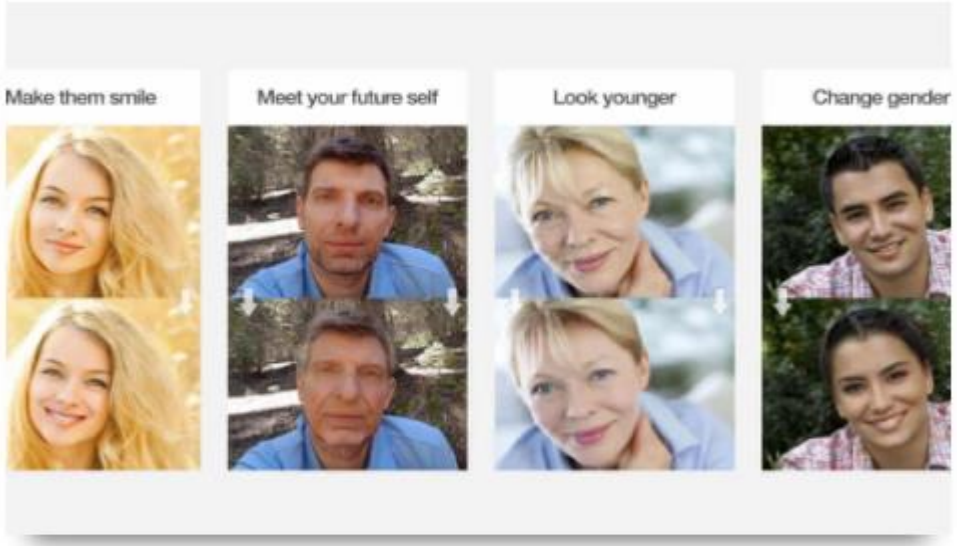
MARÇO 2016

CHATBOT DA MICROSOFT TORNA-SE RACISTA EM MENOS DE UM DIA

A chatbot Tay, que constrói discurso a partir de aprendizado de máquina, virou racista e xenófoba em menos de um dia, mostrando falta de compreensão da sociedade pelos engenheiros da empresa.

- SILVA, Tarcízio. Linha do Tempo do Racismo Algorítmico. **Blog do Tarcízio Silva**, 2020. Disponível em: <http://https://tarciziosilva.com.br/blog/posts/racismo-algoritmico-linha-do-tempo>. Acesso em: 03 de fevereiro de 2021.

O que queremos evitar



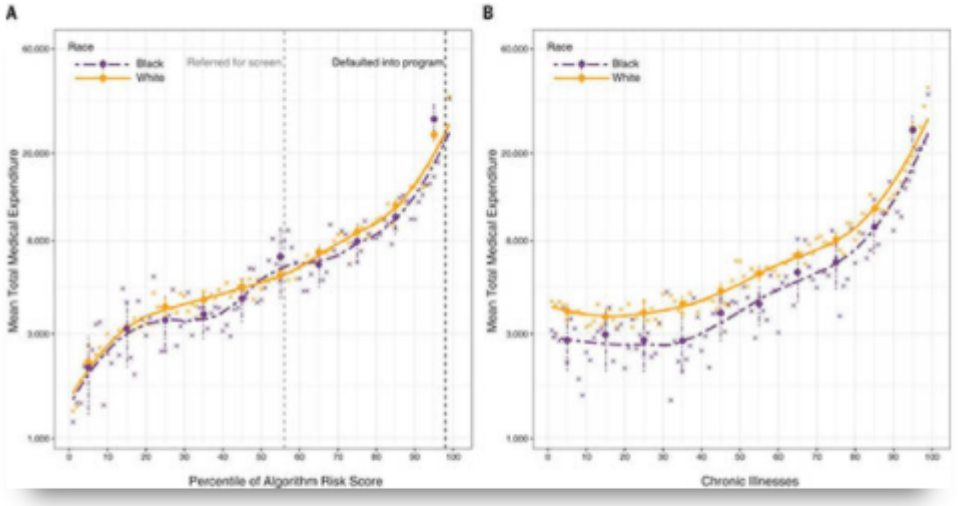
ABRIL 2017

APP QUE TRANSFORMA SELFIES EQUIPARA BELEZA À BRANCURA

Aplicativo "FaceApp" viralizou com filtros de vários tipos. O que torna as selies "mais bonitas" também torna os rostos brancos ou mais brancos.

- SILVA, Tarcízio. Linha do Tempo do Racismo Algorítmico. **Blog do Tarcízio Silva**, 2020. Disponível em: <http://https://tarciziosilva.com.br/blog/posts/racismo-algoritmico-linha-do-tempo>. Acesso em: 03 de fevereiro de 2021.

O que queremos evitar



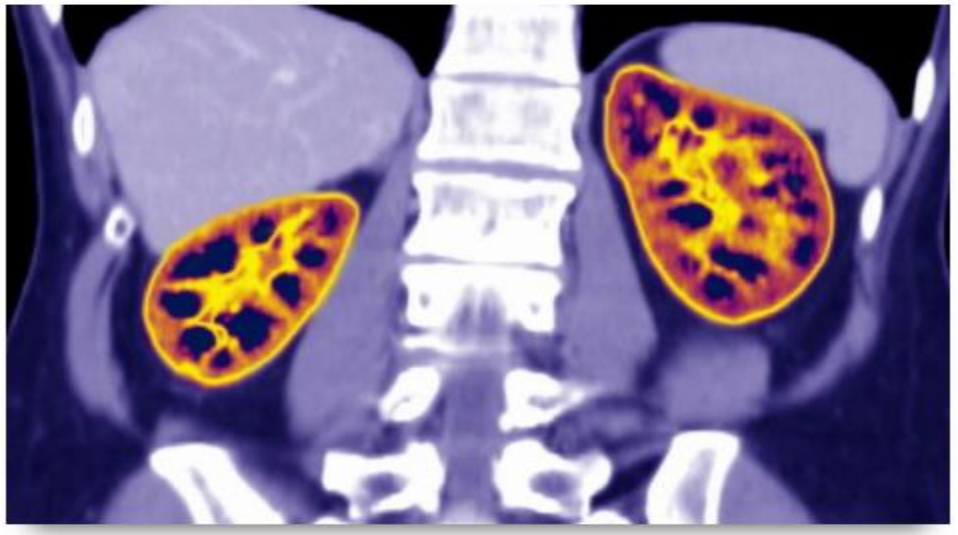
OUTUBRO 2019

ESCORES DE RISCO EM SAÚDE PENALIZAM PACIENTES NEGROS

Algoritmos de cálculo de necessidade de saúde penalizam pacientes negros pois se baseiam em indicadores de gastos já racistas

- SILVA, Tarcízio. Linha do Tempo do Racismo Algorítmico. **Blog do Tarcízio Silva**, 2020. Disponível em: <http://https://tarciziosilva.com.br/blog/posts/racismo-algoritmico-linha-do-tempo>. Acesso em: 03 de fevereiro de 2021.

O que queremos evitar



OUTUBRO 2020

ALGORITMO IMPEDE PACIENTES NEGROS DE RECEBER TRANSPLANTE DE RIM

Estudo identificou como presunções de diferença racial em algoritmo para estimativa de funcionamento dos rins impediu < a href="https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/">ao menos 64 pacientes negros de receber o procedimento

- SILVA, Tarcízio. Linha do Tempo do Racismo Algorítmico. **Blog do Tarcízio Silva**, 2020. Disponível em: <http://https://tarciziosilva.com.br/blog/posts/racismo-algoritmico-linha-do-tempo>. Acesso em: 03 de fevereiro de 2021.

Mais exemplos de vieses

- Ferramentas de recrutamento online

  O caso da Amazon: Preconceito de gênero
  (Dados utilizados: Currículos,10 anos, homens brancos)

- Associações de palavras

  Software de conexão de palavras: Preconceitos raciais e de gênero

- Anúncios online

  Preconceitos raciais

- Reconhecimento facial

  Preconceitos raciais

- Justiça criminal

  Preconceitos raciais

Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Causas e Fontes de Vieses

- Falhas nos dados
- Preconceitos humanos históricos
- Dados incompletos
- Dados não representativos

Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Algoritmos não criam preconceitos, humanos criam

Analisar com rigor

- O que pode causar vieses na sua base de dados
- Testar > Avaliar > Executar
- O caso do excesso de representação
- Cuidado após a correção do viés
- Cuidado com as variáveis proxy (o algoritmo não é cego!)

Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Higiene Algorítmica

- Reconhecer a possível existência de vieses
- Governança ética para IA
- Interpretação humana de justiça
- Alfabetização algorítmica

Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Checklist sobre ocorrência de vieses

- O que a decisão automatizada irá fazer?
  - Quem é o público do algoritmo e quem será mais afetado por ele?
  - Temos dados de treinamento para fazer as previsões corretas sobre a decisão?
  - Os dados de treinamento são suficientemente diversificados e confiáveis? Qual é o ciclo de vida dos dados do algoritmo?
  - Com quais grupos estamos preocupados no que tange erros de dados de treinamento, tratamento desigual e impacto?

Traduzido e adaptado de:
Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Checklist sobre ocorrência de vieses

- Como a tendência potencial será detectada?
  - Como e quando o algoritmo será testado? Quem serão os alvos dos testes?
  - Qual será o limite para medir e corrigir o viés no algoritmo, especialmente no que se refere a grupos protegidos?

Traduzido e adaptado de:
Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Checklist sobre ocorrência de vieses

- Quais são os incentivos do operador?
  - O que ganharemos no desenvolvimento do algoritmo?
  - Quais são os resultados potencialmente ruins e como saberemos?
  - Quão aberto (por exemplo, em código ou intenção) faremos o processo de design do algoritmo para parceiros internos e clientes?
  - Que intervenção será tomada se prevermos que pode haver resultados ruins associados ao desenvolvimento ou implantação do algoritmo?

Traduzido e adaptado de:
Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Checklist sobre ocorrência de vieses

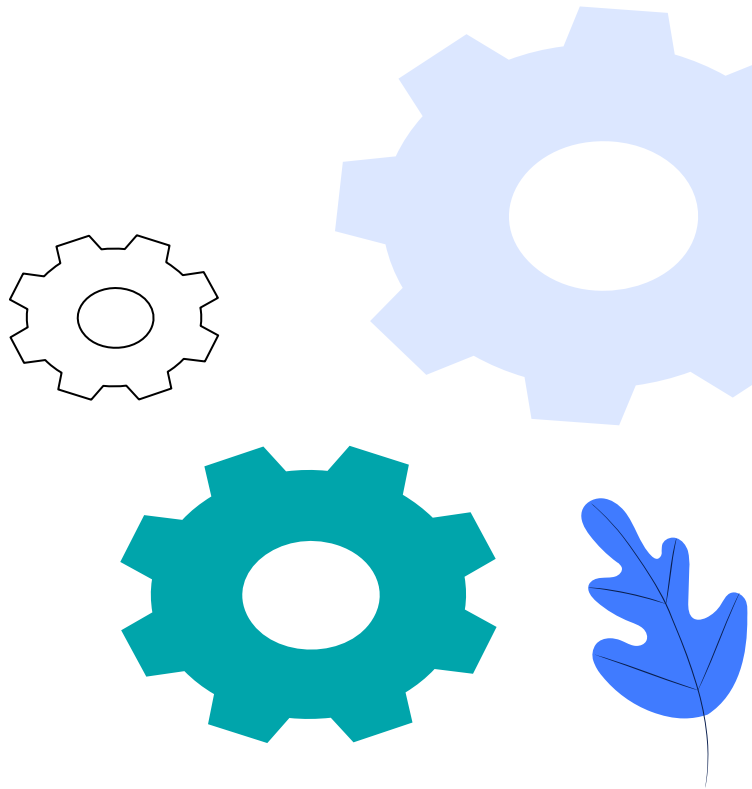
- Como outras partes interessadas estão sendo engajadas?
  - Qual é o ciclo de feedback do algoritmo para desenvolvedores, parceiros internos e clientes?
  - Existe um papel para as organizações da sociedade civil no projeto do algoritmo?

Traduzido e adaptado de:
Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Checklist sobre ocorrência de vieses

- A diversidade foi considerada no projeto e execução?
  - O algoritmo terá implicações para grupos culturais e funcionará de maneira diferente em diferentes contextos culturais?
  - A equipe de design é representativa o suficiente para capturar essas nuances e prever a aplicação do algoritmo em diferentes contextos culturais? Se não, que medidas estão sendo tomadas para tornar esses cenários mais salientes e compreensíveis para os designers?
  - Dado o propósito do algoritmo, os dados de treinamento são suficientemente diversificados?
  - Existem proteções legais que as empresas devem revisar para garantir que o algoritmo seja legal e ético?

Traduzido e adaptado de:
Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms
https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/



Contato:
marianefurtado@usp.br
andrefmb@usp.br