Vieses no Aprendizado de Máquina e suas Implicações Sociais: Um Estudo de Caso no Reconhecimento Facial

Lívia Ruback[1], Sandra Avila[2], Lucia Cantero[3]

[1]Departamento de Computação – Universidade Federal Rural do Rio de Janeiro (UFRRJ)
Seropédica, RJ – Brazil

[2]Instituto de Computação – Universidade Estadual de Campinas (Unicamp)
Campinas, SP – Brazil

[3]Department of International Studies – University of San Francisco
San Francisco, CA – EUA

liviaruback@gmail.com, sandra@ic.unicamp.br, luciacan@gmail.com

Abstract. *This work presents a study on biases generated in the machine learning process and its implications for society — moral, ethical, and social. We re-read a framework that positions the different types of biases in the machine learning process stages, from pre-processing, through data collection, to post-processing. We present a case study on facial recognition to illustrate the biases that can be potentially included during these machine learning stages, and their social implications.*

Resumo. *Este artigo apresenta um estudo sobre vieses gerados no aprendizado de máquina e as suas implicações na sociedade — morais, éticas e sociais. Fazemos uma releitura de um framework que posiciona os diferentes tipos de vieses nas etapas do processo de aprendizado de máquina, desde o pré-processamento, passando pela coleta dos dados, até o pós-processamento. Apresentamos um estudo de caso sobre reconhecimento facial para ilustrar os vieses que podem ser potencialmente incluídos durante estas etapas do aprendizado de máquina e as suas implicações sociais.*

1. Introdução

O documentário *The Coded Bias*[1], lançado em abril de 2021 na plataforma Netflix, começa com uma tentativa frustrada de Joy Boulawini, pesquisadora ganaense-americana do Instituto de Tecnologia de Massachusetts (*Massachusetts Institute of Technology* (MIT), do Inglês), de ter o seu rosto detectado por um software de visão computacional. Após inúmeras tentativas, tal sistema só foi capaz de detectar seu rosto após a pesquisadora utilizar uma máscara branca.

Sistemas para detecção de face, como o utilizado por Boulawini, utilizam aprendizado de máquina para ensinar a máquina a 'ver', fornecendo exemplos do que queremos que elas aprendam. Os algoritmos que eles utilizam extraem padrões a partir de grandes volumes de dados e são baseados em modelos estatísticos. São usados também para outras tarefas, além da detecção facial: identificar pessoas em sistemas de reconhecimento facial, determinar quem vai ser contratado ou demitido em uma empresa, conceder ou

[1]https://www.netflix.com/title/81328723

não empréstimos, diagnosticar doenças, determinar as chances de uma pessoa que cometeu um crime reincidir, entre outros. Nos sistemas de reconhecimento facial, os modelos aprendem a partir de bancos de dados com milhões de imagens de rostos, capturadas a partir de redes sociais, sites de compartilhamento de imagens e câmeras, que são armazenadas principalmente por gigantes de tecnologia como o Google e o Facebook[2].

Embora tenham sido criados para automatizar processos e aumentar a sua eficácia, pesquisas recentes [Vilarino e Vicente 2021; Buolamwini e Gebru 2018; Silva 2019] têm demonstrado que, assim como as pessoas e instituições, modelos de aprendizado de máquina podem apresentar vieses e privilegiar determinados grupos em relação a outros grupos. Estas pesquisas confrontam a ideia de sistemas de aprendizado de máquina como "tecnologias neutras", marcadas pela ausência de subjetividade da máquina [Rosa et al. 2020]. Nos sistemas de reconhecimento facial, por exemplo, que vêm sendo utilizados por vários países como tecnologia para a segurança pública, visando a identificação e prisão de suspeitos, os vieses podem ter consequências sociais graves, como acusar e prender pessoas injustamente[3].

Para mitigar os vieses no aprendizado de máquina, muitas soluções computacionais vêm sendo propostas nos últimos anos [Suresh e Guttag 2019; Mehrabi et al. 2019; Bellamy et al. 2018]. Porém, por se tratar de um problema complexo com muitas implicações sociais, soluções computacionais sozinhas não são suficientes para lidar com o problema. Tais sistemas, se utilizados sem restrições, podem amplificar disparidades e levar a graves consequências, como as que demonstraremos no decorrer do artigo. Estudos nas áreas das ciências sociais vem mostrando como as caixas-pretas, como são vistos alguns modelos de aprendizado de máquina, devem ser abertas para se avaliar como a ciência e o poder podem criar experiências desiguais [Pinch 1992; Latour 1999].

Neste trabalho, apresentamos uma releitura do *framework* proposto por [Suresh e Guttag 2019], que define uma terminologia para os principais tipos de vieses no aprendizado de máquina. Discutimos, através de um estudo de caso sobre os usos de sistemas de reconhecimento facial, alguns tipos de vieses que podem ser incluídos no aprendizado de máquina. Dessa forma, este trabalho traz as seguintes contribuições:

1. Fazemos uma releitura de um *framework* que define diferentes tipos de vieses nos sistemas de aprendizado de máquina que podem ser inseridos nas etapas do processo de aprendizado de máquina, desde o pré-processamento, passando pela coleta dos dados, até o pós-processamento.
2. Apresentamos um estudo de caso relacionado ao uso de sistemas de reconhecimento facial para ilustrar os diferentes vieses e as suas consequências — morais, éticas e sociais — na sociedade.

Na Seção 2, apresentamos alguns conceitos iniciais, relacionados ao aprendizado de máquina ao reconhecimento facial e aos vieses. Na Seção 3, apresentamos uma releitura do *framework* proposto por [Suresh e Guttag 2019], juntamente com um estudo de caso sobre reconhecimento facial. Finalmente, na Seção 4, concluímos e apresentamos os trabalhos futuros.

[2] https://cacm.acm.org/news/237592-who-owns-your-face/fulltext

[3] https://theintercept.com/2019/11/21/presos-monitoramento-facial-brasil-negros

2. Conceitos Iniciais

2.1. Aprendizado de Máquina

Aprendizagem de máquina, ou aprendizado de máquina, é um subcampo da ciência da computação que estuda a construção de algoritmos que extraem padrões a partir de grandes volumes de dados de exemplos de determinado fenômeno — também chamados de *dados de treinamento*. O termo também pode ser definido como o processo de resolução de problemas práticos por meio de 1) coleta de dados e 2) da construção algorítmica de um modelo estatístico baseado nos dados coletados [Burkov 2019]. Este modelo é usado então, para resolver, além do problema original, outros problemas similares.

Em alguns problemas, como a detecção de spam, detecção de doenças e a detecção facial, é preciso automaticamente atribuir um *rótulo* a um exemplo *não rotulado*. Um rótulo é um elemento de um conjunto de classes (geralmente duas), que podem ser, por exemplo, spam ou não spam, doente ou saudável. Em aprendizado de máquina, este problema é chamado de *classificação*. A classificação requer uma coleção de exemplos rotulados como entrada e produz um modelo que, a partir de uma nova entrada, infere o rótulo — ou um número que permita deduzir este rótulo. Nos classificadores usados no reconhecimento facial por, exemplo, os rótulos de saída dos programas não são originalmente rótulos, são uma probabilidade de a face fornecida como entrada corresponda a uma outra presente nos dados. A partir de um liminar desta probabilidade — por exemplo, 95% — o classificador produz uma saída binária: verdadeiro (para exemplos com chances menores do que 95%) ou falso (para chances a a partir de 95%).

Assim que o modelo é construído pelo algoritmo de aprendizagem a partir dos dados de treinamento, um outro conjunto é utilizado para avaliar o modelo, chamado de *dados de teste*. Os dados de testes são compostos por exemplos que o algoritmo de aprendizagem nunca viu antes. Um modelo que desempenha bem predizendo estas novas entradas é entendido como um modelo que generaliza bem [Burkov 2019]. Porém, para avaliar o quão bem o modelo usado para classificação funciona, uma tabela chamada de *matriz de confusão* é utilizada. A Figura 1 mostra um exemplo hipotético de matriz de confusão, que resume o quão bem sucedido é o modelo — ou classificador — na tarefa de predizer novos exemplos. Naturalmente, as métricas que avaliam o desempenho dos modelos consideram os seus acertos e, principalmente, seus erros. Os modelos são continuamente aperfeiçoados, através de *feedbacks*, principalmente sobre os erros, em um processo incremental.

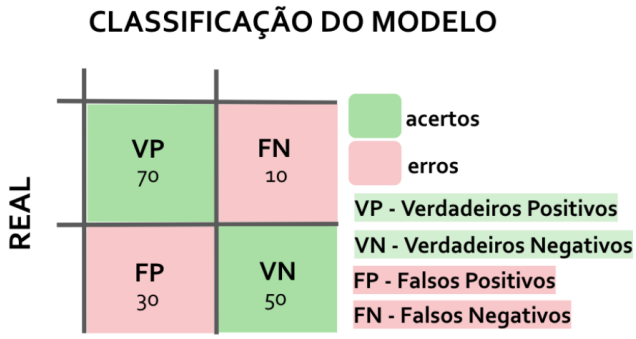


Figura 1. Exemplo de matriz de confusão para classificadores binários.

O eixo vertical se refere ao rótulo real do exemplo e o eixo horizontal se refere ao rótulo predito — ou à classificação. Os acertos do modelo são os "verdadeiros": VP (verdadeiros positivos) e VN (verdadeiros negativos). Os erros dos modelos são os "falsos": FP (falsos positivos) e FN (falsos positivos). No exemplo da Figura 1, temos um total de 160 predições, entre acertos e erros.

Os verdadeiros positivos indicam quantos foram preditos como positivos e são de fato positivos. Já os falsos positivos indicam quantos foram preditos como positivos, mas não eram positivos. No exemplo, de um total de 100 preditos como positivos, 70 eram realmente positivos e 30 foram incorretamente classificados. Se esta matriz de confusão representasse o desempenho de um modelo de reconhecimento facial, seriam 70 os casos onde o modelo reconheceu corretamente o rosto fornecido como entrada (VP) e 30 casos em que o modelo fez uma correspondência incorreta da face fornecida como entrada com uma outra face do conjunto de dados (FP), ou seja, identifica como indivíduo procurado aquele que não é o correto.

Os verdadeiros negativos indicam quantos foram preditos como negativos e são de fato negativos. Já os falsos negativos indicam quantos foram preditos como negativos, mas eram positivos. No exemplo, de um total de 60 preditos como negativos, 50 são realmente negativos e 10 foram incorretamente classificados. Em um sistema de reconhecimento facial, seriam 50 os casos onde o modelo não fez a correspondência, pois não havia o rosto fornecido como entrada nas imagens do conjunto de dados (VN) e 10 casos em que o modelo não fez a correspondência da face de entrada, mas que havia uma imagem correspondente com a face de entrada no conjunto de dados (FN).

Existem várias métricas para avaliar o desempenho (*performance*, do Inglês) dos modelos de aprendizado de máquina [Burkov 2019]. A métrica mais intuitiva é aquela que considera a proporção de acertos em relação ao total e é chamada de acurácia (*accuracy*, do Inglês). No exemplo da Figura 1, a acurácia do modelo seria de 120 (total de acertos) / 160 (total de predições), indicando que o modelo acertou 75% das vezes. Porém, algumas vezes, queremos evitar ao máximo os falsos positivos, como por exemplo, em sistemas de detecção de spam (para não perder e-mails importantes). Para estes casos, utilizamos a métrica precisão (*precision*, do Inglês). A precisão é a proporção entre os verdadeiros positivos e o total de classificados como positivos. Ou seja, quanto mais falsos positivos, menor será a precisão. No exemplo, a precisão seria de 70 (verdadeiros positivos) / 100 (total de positivos preditos), indicando que o modelo teve uma precisão de 70%.

Em outras vezes, queremos evitar ao máximo os falsos negativos, como por exemplo, em modelos de previsão de diagnósticos de doenças (um diagnóstico negativo de uma doença existente diminui as suas chances de tratamento). Para estes casos, utilizamos a métrica revocação (*recall*, do Inglês). A revocação é a proporção entre os verdadeiros positivos e o total de exemplos que são de fato verdadeiros. Ou seja, quanto mais falsos negativos, menor será a revocação. No exemplo da Figura 1, a revocação seria de 70 (verdadeiros positivos) / 80 (total de exemplos que são de fato verdadeiros), indicando que o modelo teve uma revocação de 87%.

2.2. Reconhecimento Facial

Os sistemas de aprendizado de máquina analisam a geometria das faces, considerando pontos que conectam, por exemplo, os olhos, o nariz, a boca e características como ta-

manho do queixo, distância entre olhos, entre outras, para criar uma "assinatura facial" (*faceprint*, do Inglês). Estas características são chamadas de pontos nodais — temos em média 80 pontos nodais nas faces — e são armazenadas em bancos de dados. Moraes chama este processo como "cadeia de processamento do reconhecimento facial", que começa com a captura da face por imagem/vídeo e lê a geometria da face através dos softwares [Moraes et al. 2020]. Esta assinatura facial, codificada, pode ser usada, em geral, para dois tipos de tarefas: (1) *análise facial* e (2) *reconhecimento facial*. A análise facial infere características a partir das marcas biométricas, como a idade, o gênero da pessoa, ou o estado emocional [Molina et al. 2020; Buolamwini e Gebru 2018].

Já o reconhecimento facial pode ser usado tanto para verificar a identidade da pessoa — como no desbloqueio automático de smartphones — quanto para identificar uma pessoa em meio a muitas outras, a partir de um escaneamento "um-para-muitos": é feita uma busca em um banco de dados até que haja uma correspondência com a face desejada, a partir das características dos pontos nodais de cada face [Castelvecchi 2020].

O uso de reconhecimento facial tem crescido no Brasil nos últimos anos, mas se tornou especialmente popular em 2019, quando uma comitiva do governo brasileiro foi à China, com o intuito de adquirir a tecnologia. O Instituto Igarapé levantou que, em 2019, 16 estados do Brasil usavam reconhecimento facial, contemplando 30 municípios diferentes. No total, trata-se de 48 iniciativas público-privadas em áreas como transporte, segurança pública, educação, controle de fronteiras, entre outros [Instituto Igarapé 2019].

2.3. Vieses no Aprendizado de Máquina

Vieses (*bias*, em Inglês) podem ser compreendidos sob vários pontos de vista, desde pela estatística, pela neurociência, até pela psicologia e pelo direito [ONU Mulheres 2016]. O dicionário Michaelis define viés como "tendência associada ou determinada por fatores externos". Podemos compreender vieses como mecanismos do nosso cérebro para tomar decisões, nem sempre racionais, feitas através de associações automáticas com base nas nossas experiências passadas e por heranças ancestrais. Tais associações muitas vezes se baseiam em olhares enviesados, suposições, julgamentos e preconceitos em relação a outras pessoas ou grupos, criando os chamados vieses inconscientes[4].

Um modelo de aprendizado de máquina que aprende a partir de dados faz predições injustas, privilegiando um grupo em relação a outros [Olteanu et al. 2019; Caton e Haas 2020]. Nos sistemas de reconhecimento facial, alguns levantamentos e estudos vêm comprovando que a maior parte dos vieses presentes em tais sistemas acontecem ao identificar pessoas negras[5] [Silva 2019; Buolamwini e Gebru 2018].

Muitos são os casos reportados de discriminação algorítmica. A Linha do Tempo do Racismo algorítmico[6], desenvolvida por Tarcízio Silva, apresenta casos, reportagens e reações ao racismo algorítmico, no Brasil e no mundo. Tarcízio também organizou o livro "Comunidades, Algoritmos e Ativismos Digitais – olhares afrodiaspóricos", que tem a sua versão digital gratuita[7] e reune reflexões diversas e multidisciplinares sobre as

[4] https://www.insper.edu.br/noticias/desconstruindo-preconceitos

[5] https://www1.folha.uol.com.br/cotidiano/2019/11/151-pessoas-sao-presas-por-reconhecimento-facial-no-pais-90-sao-negras.shtml

[6] https://tarciziosilva.com.br/blog/destaques/posts/racismo-algoritmico-linha-do-tempo

[7] https://literarua.commercesuite.com.br/livro/olhares-afrodiasporicos

interfaces dentre os fenômenos da comunicação digital, raça, negritude e branquitude nos últimos 20 anos, oferecendo material de referência para estudantes e pesquisadoras/es em diversos níveis" [Silva e Birhane 2020].

Vieses também estão sendo identificados em modelos gerados para prognóstico de doenças. [Bissoto et al. 2019] avaliaram vieses em modelos para prognóstico de câncer de pele. Os pesquisadores retiraram as informações das lesões nas imagens usadas para treinar alguns classificadores e então avaliaram novamente o desempenho dos modelos. Surpreendentemente, os modelos ainda mantiveram um alto desempenho, mesmo sem as informações que supostamente seriam essenciais para treiná-los. Experimentos como este demostram como os modelos de aprendizado de máquina podem se comportar como *caixas-pretas* — e os riscos de confiarmos cegamente nestes modelos.

3. Vieses Presentes no Reconhecimento Facial

Nesta seção, fazemos uma releitura do *framework* (Figuras 2a e 2b), proposto por [Suresh e Guttag 2019], para apresentar os principais vieses que podem ser inseridos no reconhecimento facial. Mais especificamente, extraímos 4 dos 6 vieses apresentados pelos autores no framework, que fazem sentido no estudo de caso de reconhecimento facial: *Historical Bias*, *Representation Bias*, *Evaluation Bias* e *Deployment Bias*. Não consideramos aqui o *Measurement Bias* e o *Aggregation Bias*, presentes no artigo original.

A Figura 2a mostra os vieses que podem ser inseridos nas etapas de coleta de dados e de pré-processamento e a Figura 2b exibe os que podem ser inseridos na criação do modelo, na sua avaliação e no pós-processamento. Estas etapas são detalhadas a seguir.

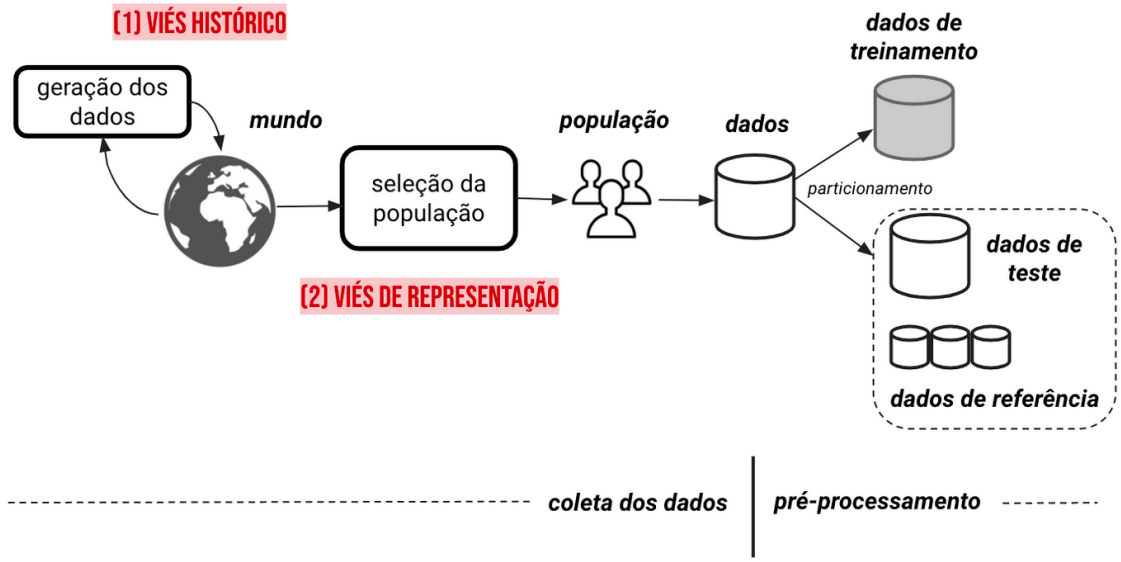
Coleta de dados. Após a geração de dados, eles são coletados: parte deles é selecionada, juntamente com algumas de suas características. Os programadores, muitas vezes, não realizam esta etapa, utilizando dados pré-existentes. Para sistemas de reconhecimento facial, por exemplo, alguns projetos disponibilizam dados com imagens de rostos para download[8] para serem utilizados na construção e treinamento dos modelos.

Pré-processamento. O pré-processamento envolve tarefas como limpeza dos dados — remoção ou substituição dos dados incompletos ou inválidos, seleção de atributos (*features*, em Inglês) usados pelo modelo, entre outras. Os dados são então particionados em dois: dados de treinamento (utilizados pelo modelo para aprender os padrões e fazer as previsões) e dados de teste (utilizados para avaliar o desempenho do modelo, i.e., o quão bem o modelo "aprendeu" – ver Seção 2.1).

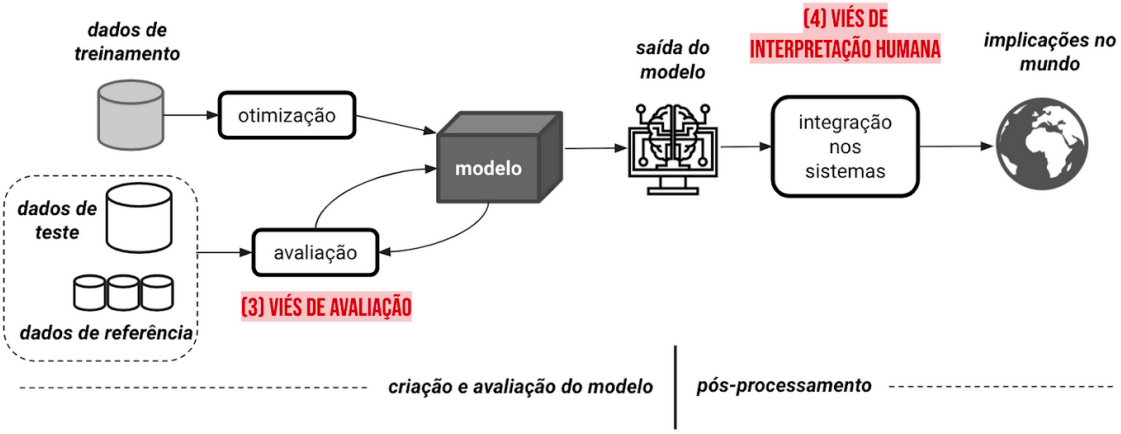
Criação do modelo. O modelo é construído utilizando os dados de treinamento, sem incluir os dados de teste (utilizados somente para avaliar o desempenho do modelo). Neste ponto, uma série de algoritmos podem ser usados para criar o modelo, como por exemplo algoritmos de aprendizado profundo, inspirados em redes neurais com várias camadas entre a entrada e a saída do modelo [Burkov 2019]. Para cada um dos algoritmos, são testados modelos com diferentes parâmetros e métodos de otimização e o com melhor desempenho é escolhido.

Avaliação do modelo. Nesta etapa, diferentes métricas de desempenho podem ser utilizadas (ver Seção 2.1). Após os testes com diferentes parâmetros, o modelo final, otimizado,

[8] https://www.face-rec.org/databases



(a) Viés Histórico e Viés de representação.



(b) Viés de avaliação e Viés de interpretação humana.

Figura 2. Ilustração de tipos de vieses, adaptado de [Suresh e Guttag 2019].

é escolhido. O desempenho final do modelo é calculado utilizando somente os dados de teste, que não são utilizados antes desta etapa. Isto garante que o desempenho do modelo representa de fato como ele se comporta com dados desconhecidos. Além dos dados de teste, outros conjuntos de dados, chamados de dados de referência (*benchmarks*, do Inglês), podem ser usados para se comprovar o desempenho do modelo e para comparar o seu desempenho com outros modelos já existentes. Como exemplos de dados de referência disponíveis, podemos citar os dados do UCI (*University of California, Irvine Machine Learning Repository*)[9], o ImageNet[10], entre outros.

Pós-processamento. Finalizado o modelo de aprendizado de máquina, algumas outras etapas podem ser necessárias para que o modelo seja de fato usado nas aplicações. Por exemplo, se o algoritmo escolhido para o modelo gera como saída uma porcentagem — probabilidade da face detectada na foto ser de determinada pessoa — e o sistema

[9] https://archive.ics.uci.edu/ml/datasets.php
[10] http://www.image-net.org

espera uma saída binária — sim ou não — é preciso escolher um limiar para converter tal probabilidade em uma classificação binária. É preciso, também, interpretar as saídas do modelo, de acordo com o propósito pelo qual o sistema foi construído.

O processo de desenvolvimento de um modelo de aprendizado de máquina geralmente é incremental, de forma que eles são retroalimentados com *feedbacks* — sobretudo indicando os erros nas predições. Dessa forma, os modelos aprendem com os próprios erros, o que permite a melhora contínua do seu desempenho. Os vieses podem surgir em várias destas etapas — e inclusive ao mesmo tempo. Neste artigo, nos limitamos a exemplificar, a partir de um estudo de caso no reconhecimento facial, quatro destes tipos de vieses — apresentados a seguir — que podem potencialmente ser inseridos nas etapas apresentadas.

(1) Viés histórico: Os vieses históricos acontecem na etapa anterior à coleta de dados (ver Figura 2a). Quando dados de entrada refletem na saída resultados passados, que podem ser discriminatórios, eles reforçam julgamentos e preconceitos dos indivíduos e instituições[11], como o racismo. Um levantamento feito pela Rede de Observatórios de Segurança, em 2019, mostrou que 90,5% dos presos por monitoramento facial no Brasil são negros[12]. Os erros na identificação dos suspeitos são decorrentes, principalmente, de falsos positivos nos modelos (ver Seção 2.1) e podem representar constrangimentos, prisões arbitrárias e violação dos direitos humanos [Nunes 2019]. Quando essas violações acontecem majoritariamente em grupos específicos — como o de pessoas negras, o problema é ainda maior. Um outro levantamento, feito pelo Colégio Nacional de Defensores Públicos Gerais (Condege)[13], mostrou que, entre 2012 e 2020, 81% das prisões injustas baseadas no reconhecimento facial no país foram de pessoas negras, perpetuando o racismo e agravando o encarceramento em massa de negros. Silvio de Almeida define racismo como "uma forma sistemática de discriminação que tem a raça como fundamento, e que se manifesta por meio de práticas conscientes ou inconscientes que culminam em desvantagens ou privilégios para os indivíduos, a depender do grupo racial ao qual pertençam" [Almeida 2019]. Os vieses históricos são, portanto, assim como o racismo estrutural, sistêmicos por natureza e, naturalmente, se refletem na construção de dados de treinamento desbalanceados.

(2) Viés de representação (ou de amostra): Os vieses de representação podem acontecer na etapa de coleta de dados (ver Figura 2a) e são incluídos na própria construção de dados de treinamento não representativos. Quando a amostra coletada não é representativa da população a ser modelada, de forma balanceada, o modelo irá errar muito mais em predizer rótulos para estes grupos sub-representados[14]. Joy Buolamwini, a protagonista do documentário *The Coded Bias*, pesquisadora do MIT, realizou uma das primeiras pesquisas que tratam de vieses em sistemas de reconhecimento facial. [Buolamwini e Gebru 2018] analisaram o desempenho de modelos de classificação de gênero por sistemas reconhecimento facial dos sistemas da Microsoft, da IBM e do Face++. A pesquisa, apresentada no documentário, concluiu que, no geral, homens e pessoas brancas foram melhor classificados pelos modelos do que os outros grupos. Uma visão interseccional da pesquisa revela que todos os classificadores avaliados tiveram um pior desempenho ao classificar

[11] https://medium.com/tecs-usp/inteligencias-artificiais-preconceitos-reais-f30c018cb2dd

[12] https://theintercept.com/2019/11/21/presos-monitoramento-facial-brasil-negros

[13] https://www.defensoria.rj.def.br/noticia/detalhes/11088-Relatorios-apontam-falhas-em-prisoes-apos-reconhecimento-fotografico

[14] https://tecs.ime.usp.br/etica/apresentacao-vies.pdf

mulheres negras. Estes vieses de representação estão presentes em muitos modelos de reconhecimento facial, que se baseiam em dados de treinamento desbalanceados, treinados majoritariamente em rostos de homens e pessoas de pele clara. A Figura 3 mostra um exemplo de dois conjuntos de dados com vieses de representação, apontados por [Buolamwini e Gebru 2018], utilizados em sistemas de reconhecimento facial. Pode-se observar a distribuição desbalanceada por gênero e tipo de pele nos dois conjuntos de dados, Adience e IJB-A. Nos dados da Adience, enquanto homens de pele clara representam 41,6% do total, as mulheres negras representam 7,4%. Já nos dados do IJB-A, a diferença é ainda maior: homens de pele clara representam 59,4% do total e mulheres negras representam somente 4,4% do total.

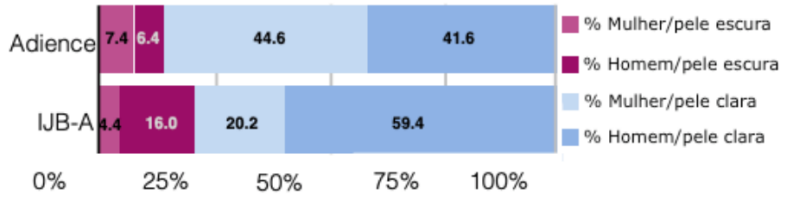


Figura 3. Exemplos de dados com vieses de representação [Buolamwini e Gebru 2018].

Sistemas de reconhecimento facial que utilizam dados de treinamento não balanceados, como os dados apresentados na Figura 3, geram sistemas com vieses de representação, gerando maiores taxas de erros — entre falsos positivos e falsos negativos — ao identificar os grupos sub-representados. A solução técnica para mitigar os vieses de representação é relativamente simples: basta construir bases de treinamento representativas. Porém, estas questões demandam um tipo diferente de sabedoria de cientistas de dados e criadores de algoritmos, em sua maioria homens brancos, que envolve questões sociais e de políticas públicas a qual estes profissionais têm pouca exposição [Silva 2019].

(3) Viés de avaliação. Os vieses de avaliação podem ser inseridos na etapa de avaliação do desempenho do modelo. O modelo aprende com os dados de treinamento, mas tem a sua qualidade avaliada a partir de dados de teste — ou de dados de referência (ver Figuras 2a e 2b). Dados usados como referência que não representam de forma balanceada os diferentes subgrupos da população — como os dados mostrados na Figura 3 — levam a modelos com vieses de avaliação. Para contornar este viés, os sistemas devem testar os seus modelos em dados de referência representativos. Joy Buolamwini e Timnit Gebru criaram, no projeto *Gender Shades*[15], um conjunto de dados balanceados, chamado PPB (*Pilot Partiament Benchmark*) (Figura 4) contendo 1270 imagens de rostos de pessoas, incluindo três países africanos e três países europeus, com uma boa distribuição de gênero e fenótipo quanto a cor da pele. Estes dados podem ser usados tanto como dados de treinamento quanto como dados de teste e de referência para modelos que implementam reconhecimento facial. As pesquisadoras, sozinhas, conseguiram gerar dados mais precisos do que os oferecidos por algumas gigantes de tecnologia. O impacto das pesquisas iniciadas pelas pesquisadoras foi tão grande que, em 2020, a IBM encerrou as suas pesquisas em reconhecimento facial, se posicionando contra o uso da tecnologia para monitoramento em massa e vigilância[16].

[15] http://gendershades.org

[16] https://www.theverge.com/2020/6/8/21284683/ibm-no-longer-general-purpose-facial-

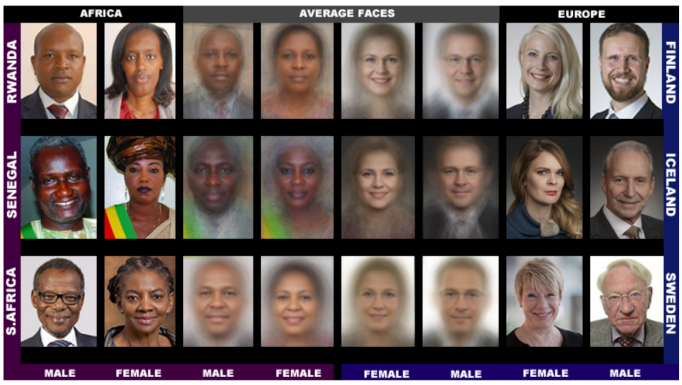


Figura 4. Dados do PPB (*Pilot Partiament Benchmark*) balanceados por gênero e tipo de pele, com parlamentares de 6 países [Buolamwini e Gebru 2018].

Uma outra forma de se inserir vieses de avaliação é através da escolha das métricas de avaliação de desempenho dos modelos (ver Seção 2.1). Em seu livro "Algoritmos de destruição em massa", Cathy O'Neil afirma que os modelos de aprendizado de máquina são "opiniões embutidas em matemática" e destaca que, se um modelo funciona ou não também é uma questão de opinião, afinal, um componente-chave de todo modelo é a sua definição de sucesso" [O'Neil 2020]. Ao escolher, por exemplo, uma métrica geral como a acurácia para avaliar o modelo, podemos esconder disparidades entre os diferentes subgrupos [Suresh e Guttag 2019]. Neste contexto, algumas pesquisas recentes tem proposto novas métricas para avaliar o desempenho dos modelos de forma que englobem noções de vieses, justiça e/ou discriminação [Caton e Haas 2020; Suresh e Guttag 2019; Mehrabi et al. 2019; Bellamy et al. 2018]. Uma forma relativamente simples de avaliar a presença de vieses de avaliação é aplicar a métrica — qualquer que seja — nos grupos separadamente. Por exemplo, um modelo de reconhecimento facial pode ter uma precisão geral de 80%, mas se formos considerar a precisão dentro do grupo que inclui mulheres negras, a precisão cai para 60%, enquanto que a precisão dentro do grupo que corresponde a homens de pele clara, a precisão sobe para 90%. Uma das métricas alternativas para lidar com vieses de avaliação é a métrica *impacto desproporcional* (*disparate impact*, do Inglês), que considera a razão entre grupos não privilegiados e privilegiados, ou seja, avalia o equilíbrio entre os valores preditivos positivos entre os grupos [Caton e Haas 2020].

(4) Viés de interpretação humana. Os vieses de interpretação humana podem ser inseridos na etapa de pós-processamento, durante a integração dos sistemas. A saída do modelo — por exemplo, a identificação de um suspeito, nos sistemas de reconhecimento facial — deve ser sempre interpretada por seres-humanos, de forma a evitar consequências injustas. Estes vieses ocorrem quando há uma incompatibilidade entre o problema que o modelo se propôs a resolver e a forma em que ele é usado na prática [Suresh e Guttag 2019]. Muitas vezes, quando acontece o reconhecimento facial de um suspeito por um sistema, as autoridades que buscam a punição criminal de alguém já consideram a resultado do modelo como prova da prática do crime, muitas vezes sem dar continuidade às investigações.

O estudo "Regulação do reconhecimento facial no setor público", lançado em 2020 pelo Instituto Igarapé e o Data Privacy Brasil [Francisco et al. 2020], mostra que a legislação

recognition-analysis-software

sobre o uso de reconhecimento facial na Inglaterra, França, Estados Unidos exigem a anuência expressa dos usuários sobre os possíveis usos das informações que eles fornecem, entre outras políticas de proteção de dados. No Brasil, em que pese a Lei Geral de Proteção de Dados Pessoais 13.709/2018, o governo editou a portaria n.793/2019, que estimula políticas de reconhecimento facial, sem, em contrapartida, desenvolver um marco de controle destes dispositivos, já utilizados na atividade policial [Francisco et al. 2020].

4. Conclusão

Neste trabalho, apresentamos um estudo sobre vieses gerados nas etapas do aprendizado de máquina, que vem sendo utilizados para automatizar várias tarefas anteriormente realizadas por seres humanos. Nos sistemas de reconhecimento facial, estes modelos estão sendo utilizados, no Brasil e no mundo, para identificar suspeitos, e estão apresentando maiores taxas de erro para identificar pessoas de grupos que já sofrem preconceitos. Fizemos uma releitura do *framework* proposto por [Suresh e Guttag 2019] para ilustrar quatro tipos de vieses — e suas implicações sociais — que podem ser incluídos durante a construção de um modelo para reconhecimento facial: viés histórico, viés de representação, viés de avaliação e viés de interpretação humana.

Através dos exemplos de discriminação algorítmica apresentados neste trabalho, mostramos como o reconhecimento facial tem reforçado preconceitos já existentes na sociedade e tem "se mostrado uma atualização *high-tech* para o velho e conhecido racismo que está na base do sistema de justiça criminal brasileiro" [Nunes 2019]. Como trabalhos futuros, estamos investigando as métricas alternativas para avaliar o desempenho dos modelos para torná-los mais justos e inclusivos. Além disso, estamos investigando como a interdisciplinaridade pode ser utilizada para tornar os modelos de aprendizado de máquina mais "inteligentes", de acordo com uma visão mais ampla de inteligência, que abarca vários campos da ciência cognitiva, como a filosofia e a antropologia.

Referências

[Almeida 2019] Almeida, S. (2019). *Racismo estrutural*. Pólen Produção Editorial LTDA.

[Bellamy et al. 2018] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.

[Bissoto et al. 2019] Bissoto, A., Fornaciali, M., Valle, E., and Avila, S. (2019). (De)Constructing bias on skin lesion datasets. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[Buolamwini e Gebru 2018] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

[Burkov 2019] Burkov, A. (2019). The hundred-page machine learning book (em português).

[Castelvecchi 2020] Castelvecchi, D. (2020). Is facial recognition too biased to be let loose? *Nature*, 587(7834):347–349.

[Caton e Haas 2020] Caton, S. and Haas, C. (2020). Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*.

[Francisco et al. 2020] Francisco, P. A. P., Hurel, L. M., and Rielli, M. M. (2020). Regulação do reconhecimento facial no setor público. *Data Privacy Brasil*. https://igarape.org.br/wp-content/uploads/2020/06/2020-06-09-Regulao-do-reconhecimento-facial-no-setor-pblico.pdf.

[Instituto Igarapé 2019] Instituto Igarapé (2019). Reconhecimento facial no brasil. https://igarape.org.br/infografico-reconhecimento-facial-no-brasil.

[Latour 1999] Latour, B. (1999). *Pandora's hope: essays on the reality of science studies*. Harvard University Press.

[Mehrabi et al. 2019] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

[Molina et al. 2020] Molina, D., Causa, L., and Tapia, J. (2020). Toward to reduction of bias for gender and ethnicity from face images using automated skin tone classification. In *International Conference of the Biometrics Special Interest Group*, pages 281–289.

[Moraes et al. 2020] Moraes, T. G., Almeida, E. C., and de Pereira, J. R. L. (2020). Smile, you are being identified! risks and measures for the use of facial recognition in (semi-) public spaces. *AI and Ethics*, pages 1–14.

[Nunes 2019] Nunes, P. (2019). Novas ferramentas, velhas práticas: reconhecimento facial e policiamento no Brasil. Retratos da violência: cinco meses de monitoramento, análise e descobertas (Rede de Observatório de Segurança). http://observatorioseguranca.com.br/wp-content/uploads/2019/11/1relatoriorede.pdf.

[Olteanu et al. 2019] Olteanu, A., Castillo, C., Diaz, F., and Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13.

[O'Neil 2020] O'Neil, C. (2020). *Algoritmos de Destruição em Massa*. Editora Rua do Sabão, 1ª edition.

[ONU Mulheres 2016] ONU Mulheres, Insper, M. M. . P. B. (2016). Vieses incosscientes, equidade de gênero e o mundo corporativo: lições da oficina vieses inconscientes. https://www.onumulheres.org.br/wp-content/uploads/2016/04/Vieses_inconscientes_16_digital.pdf.

[Pinch 1992] Pinch, T. J. (1992). Opening black boxes: Science, technology and society. *Social Studies of Science*, 22(3):487–510.

[Rosa et al. 2020] Rosa, A., Pessoa, S. A., and Lima, F. S. (2020). Neutralidade tecnológica: reconhecimento facial e racismo. *REVISTA V! RUS*, 21. http://www.nomads.usp.br/virus/virus21/?sec=4&item=9&lang=pt.

[Silva 2019] Silva, T. (2019). Visão computacional e vieses racializados: branquitude como padrão no aprendizado de máquina. *II COPENE Nordeste: Epistemologias Negras e Lutas Antirracistas*, pages 29–31.

[Silva e Birhane 2020] Silva, T. and Birhane, A. (2020). *Comunidades, algoritmos e ativismos digitais: olhares afrodiasporicos*. LiteraRua.

[Suresh e Guttag 2019] Suresh, H. and Guttag, J. V. (2019). A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*.

[Vilarino e Vicente 2021] Vilarino, R. and Vicente, R. (2021). Dissecting racial bias in a credit scoring system experimentally developed for the brazilian population. *arXiv preprint arXiv:2011.09865*.