PGCOMP - Programa de Pós-Graduação em Ciência da Computação
Universidade Federal da Bahia (UFBA)
Av. Milton Santos, s/n - Ondina
Salvador, BA, Brasil, 40170-110

https://pgcomp.ufba.br
pgcomp@ufba.br

Algoritmos de detecção e reconhecimento facial têm sido amplamente adotados para as mais diversas aplicações como, por exemplo, em redes sociais que automaticamente detectam e reconhecem todas as pessoas presentes em imagens publicadas. No entanto, com o crescimento do uso de algoritmos de Inteligência Artificial (IA) em geral, começaram a surgir questionamentos relacionados à existência de vieses. Em muitas situações foram encontrados vieses que afetam minorias historicamente oprimidas. Como exemplo, foi notado viés racial em muitos sistemas de reconhecimento facial utilizados pela polícia americana, o que levou à suspensão do uso dessa tecnologia em alguns estados, à descontinuação do desenvolvimento em algumas empresas, como a IBM, e pesquisadores a pedirem para seus colegas pararem de trabalhar nestes sistemas devido ao impacto sobre pessoas de diferentes raças e etnias. A problemática supracitada motiva o estudo e avaliação da existência de viés em um sistema, baseado em IA, para detectar fraudes no transporte público de Salvador (Brasil). Considerando que Salvador é a cidade brasileira com maior percentual de negros, qualquer erro pode afetar um número significativo de usuários, levando a um alto número de falsos positivos. Em estudos anteriores desenvolvidos pelo grupo de pesquisa em que o autor deste trabalho pertence, foram realizados testes estatísticos para verificar se há correlação entre a taxa de erro e a raça e gênero. Os resultados indicaram a existência dessa correlação, ou seja, há uma maior taxa de erro de detecção facial em usuários pretos ou pardos e mulheres. Com base em tais resultados, uma questão principal motivou o desenvolvimento deste trabalho: Há, de fato, uma relação causal entre a raça e a taxa de erros na detecção? Para avaliar essa questão, foi desenvolvido um modelo causal para estudar a influência da cor de pele no sistema de detecção facial utilizado no transporte público de Salvador.

Palavras-chave: Causalidade, Imparcialidade Causal, Detecção de Faces, Modelos Causais Estruturais, Aprendizado de Máquina, Grafos Direcionados Acíclicos.

MSC — 180 — 2024

Modelagem Causal para Estudo de Viés Racial em Sistemas de Detecção de Face

Ariel Menezes de Almeida Júnior

UFBA

Modelagem Causal para Estudo de Viés Racial em Sistemas de Detecção de Face

Ariel Menezes de Almeida Júnior

Dissertação de Mestrado

Universidade Federal da Bahia

Programa de Pós-Graduação em
Ciência da Computação

Maio — 2024

Universidade Federal da Bahia
Instituto de Computação

Programa de Pós-Graduação em Ciência da Computação

MODELAGEM CAUSAL PARA ESTUDO DE VIÉS RACIAL EM SISTEMAS DE DETECÇÃO DE FACE

Ariel Menezes de Almeida Júnior

DISSERTAÇÃO DE MESTRADO

Salvador
28 de maio de 2024

ARIEL MENEZES DE ALMEIDA JÚNIOR

MODELAGEM CAUSAL PARA ESTUDO DE VIÉS RACIAL EM SISTEMAS DE DETECÇÃO DE FACE

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Ricardo Araújo Rios
Co-orientador: Marcelo Magalhães Taddeo

Salvador
28 de maio de 2024

Ficha catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias Prof. Omar Catunda, SIBI – UFBA.

A447 Almeida Junior, Ariel Menezes de

Modelagem Causal para Estudo de Viés Racial em Sistemas de Detecção de Face / Ariel Menezes de Almeida Júnior. – Salvador, 2024.

44p.: il.

Orientador: Prof. Dr. Ricardo Araújo Rios
Coorientador: Prof. Dr. Marcelo Magalhães Taddeo.

Dissertação (Mestrado) – Universidade Federal da Bahia. Instituto de Computação, 2024.

1. Causalidade. 2. Imparcialidade Causal. 3. Detecção de Faces. I. Rios, Ricardo Araújo. II. Taddeo, Marcelo Magalhães III. Universidade Federal da Bahia. IV. Título.

CDU: 004.8:530.16

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DA BAHIA
INSTITUTO DE COMPUTAÇÃO
PGCOMP - Programa de Pós-Graduação em Ciência da Computação
http://pgcomp.ufba.br

"*Modelagem Causal para Estudo de Viés Racial em Sistemas de Detecção de Face*"

ARIEL MENEZES DE ALMEIDA JUNIOR

Dissertação apresentada ao Colegiado do Programa de Pós-Graduação em Ciência da Computação na Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Ciência da Computação.

Banca Examinadora

gov.br Documento assinado digitalmente
RICARDO ARAUJO RIOS
Data: 29/07/2024 12:18:32-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Ricardo Araújo Rios (Orientador - PGCOMP)

gov.br Documento assinado digitalmente
MARCELO MAGALHAES TADDEO
Data: 08/08/2024 12:31:32-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcelo Magalhaes Taddeo (Coorientador - UFBA)

gov.br Documento assinado digitalmente
RENATO PORFIRIO ISHII
Data: 29/07/2024 12:39:07-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Renato Porfirio Ishii (UFMS)

gov.br Documento assinado digitalmente
TATIANE NOGUEIRA RIOS
Data: 29/07/2024 12:22:05-0300
Verifique em https://validar.iti.gov.br

Profª. Drª. Tatiane Nogueira Rios (PGCOMP)

AGRADECIMENTOS

Antes de tudo, agradeço a Deus por tantas bênçãos e por me possibilitar chegar até aqui.

Agradeço à minha mãe por todo incentivo e apoio ao longo dessa vida. Você foi a primeira pessoa a acreditar em mim, sempre me encorajando e incentivando a estudar, persistir e lutar pelo meu sucesso.

Agradeço, especialmente, à minha amada Tauane Sales. Você é uma mulher excepcional, esteve ao meu lado em momentos difíceis, me apoiou durante o mestrado e nessa jornada que chamamos de vida. Você me incentivou e auxiliou na revisão deste trabalho e na produção de muitos outros. Sou eternamente grato pela pessoa que você é em minha vida, por todas as trocas, paciência, amor, cuidado e pela forma como mutuamente nos fazemos ir mais longe.

Agradeço ao meu orientador, Dr. Ricardo Rios, pela orientação, suporte e paciência ao longo de todo este trabalho, que fizeram toda a diferença para que eu conseguisse finalizar esse mestrado. Levo seu exemplo de profissionalismo, conhecimento e comunicação como uma meta pessoal.

A meu coorientador, Dr. Marcelo Taddeo, pelo efeito de suas aulas sobre causalidade, que causaram uma maior compreensão desse assunto através de suas explicações.

Aos professores que fizeram parte de minha banca de qualificação e defesa. Em especial, à Dr.ª Tatiane Nogueira, cujas sugestões elevaram a qualidade deste trabalho.

Aos meus colegas do grupo de pesquisa e demais colegas de curso, minha gratidão pela colaboração e por estarem disponíveis sempre que necessário.

À CAPES pelo suporte financeiro.

E a todos que participaram, direta ou indiretamente, deste trabalho e conquista.

RESUMO

Algoritmos de detecção e reconhecimento facial têm sido amplamente adotados para as mais diversas aplicações como, por exemplo, em redes sociais que automaticamente detectam e reconhecem todas as pessoas presentes em imagens publicadas. No entanto, com o crescimento do uso de algoritmos de Inteligência Artificial (IA) em geral, começaram a surgir questionamentos relacionados à existência de vieses. Em muitas situações foram encontrados vieses que afetam minorias historicamente oprimidas. Como exemplo, foi notado viés racial em muitos sistemas de reconhecimento facial utilizados pela polícia americana, o que levou à suspensão do uso dessa tecnologia em alguns estados, à descontinuação do desenvolvimento em algumas empresas, como a IBM, e pesquisadores a pedirem para seus colegas pararem de trabalhar nestes sistemas devido ao impacto sobre pessoas de diferentes raças e etnias. A problemática supracitada motiva o estudo e avaliação da existência de viés em um sistema, baseado em IA, para detectar fraudes no transporte público de Salvador (Brasil). Considerando que Salvador é a cidade brasileira com maior percentual de negros, qualquer erro pode afetar um número significativo de usuários, levando a um alto número de falsos positivos. Em estudos anteriores desenvolvidos pelo grupo de pesquisa em que o autor deste trabalho pertence, foram realizados testes estatísticos para verificar se há correlação entre a taxa de erro e a raça e gênero. Os resultados indicaram a existência dessa correlação, ou seja, há uma maior taxa de erro de detecção facial em usuários pretos ou pardos e mulheres. Com base em tais resultados, uma questão principal motivou o desenvolvimento deste trabalho: Há, de fato, uma relação causal entre a raça e a taxa de erros na detecção? Para avaliar essa questão, foi desenvolvido um modelo causal para estudar a influência da cor de pele no sistema de detecção facial utilizado no transporte público de Salvador.

Palavras-chave: Causalidade, Imparcialidade Causal, Detecção de Faces, Modelos Causais Estruturais, Aprendizado de Máquina, Grafos Direcionados Acíclicos.

ABSTRACT

Face detection and recognition algorithms have been widely adopted in a diversity of applications, such as social networks that automatically detect and recognize every person present in published images. However, with the growing adoption of Artificial Intelligence (AI) algorithms in general, questions related to the existence of biases began to arise. In many situations, it was found that there were biases affecting historically oppressed minorities. As an example, there was racial bias in many facial recognition systems used by American police, which led to the suspension of the use of this technology in some states, the discontinuation of development in some companies, such as IBM, and researchers to ask their colleagues to stop working on these systems because of the impact on people of different races and ethnicities. The aforementioned problem motivates the study and evaluation of the existence of bias in a fraud detection, AI-based, system used by the public transportation in Salvador (Brazil). Taking into account the fact that Salvador is the Brazilian city with the highest percentage of black people, any error can affect a significant number of users, leading to a high number of false positives. In previous studies developed by the research group, statistical tests were performed to verify if there is a correlation between the error rate and race and gender. The results indicated the existence of this correlation, that is, there is a higher error rate in face detection of black or brown users and women. From these results, a main question motivates the development of this project: Is there, in fact, a causal relationship between race and the error rate in detection? To evaluate this question, a causal model was developed to analyze the influence of skin color on the face detection system used in Salvador's public transportation.

Keywords: Causality, Causal Fairness, Face Detection, Strucutral Causal Models, Machine Learning, Directed Aciclic Graphs.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

Capítulo 1

INTRODUÇÃO

1.1 CONTEXTO E MOTIVAÇÃO

Algoritmos de detecção e reconhecimento facial têm sido amplamente adotados para as mais diversas aplicações como, por exemplo, em redes sociais que automaticamente detectam e reconhecem usuários após a publicação de imagens (SHARMA et al., 2017). Além das redes sociais, há diversas aplicações para entretenimento em jogos e realidade virtual, vigilância e investigações criminais, segurança em desbloqueio e autenticação de *smartphones*, e detectores de fraude (ZHAO et al., 2003; MISRA; GAJ, 2006).

De maneira geral, entende-se como detecção facial a capacidade de determinar se há ou não faces em imagens arbitrárias e, em caso positivo, suas respectivas localizações (SHARMA et al., 2017). O reconhecimento facial, que pode ser visto como uma extensão da detecção facial, é entendido como, dada imagens fixas ou de vídeo de uma cena, a capacidade de identificar ou verificar uma ou mais pessoas na cena usando um banco de dados previamente armazenado de faces (ZHAO et al., 2003).

Com os recentes avanços das técnicas de Inteligência Artificial (IA), principalmente os modelos de Aprendizado Profundo - Deep Learning (DL), nota-se uma melhoria expressiva no desempenho das tecnologias de reconhecimento facial, tornando-as populares em empresas de pequeno e médio porte. No entanto, o uso em grandes escalas tem levantado questionamentos relacionados à existência de possíveis vieses. Em muitas situações, foram encontrados erros que tendem a se propagar no reconhecimento de usuários que fazem parte de minorias historicamente oprimidas. Como exemplo, estudos mostraram que havia viés racial em muitos sistemas de reconhecimento facial utilizados pela polícia americana, levando à suspensão do uso dessa tecnologia em alguns estados, à descontinuação do desenvolvimento em algumas empresas, como a IBM, e a solicitações por parte de pesquisadores para que seus colegas interrompessem o desenvolvimento de novos modelos antes de compreender melhor o impacto dos vieses sobre pessoas de diferentes raças e etnias (RAJI et al., 2020; LUNTER, 2020; CASTELVECCHI, 2020).

A problemática supracitada motivou o estudo e a avaliação da existência de viés em um sistema de IA para detectar fraudes no transporte público de Salvador (Brasil). Considerando que Salvador é a cidade brasileira com maior percentual de negros (cerca de 80%

da população (CALVO-GONZALEZ; DUCCINI, 2010)), qualquer erro pode afetar um número significativo de usuários. Atualmente, mais de 1 milhão de passageiros utilizam diariamente ônibus do transporte público. Nesses ônibus, há um sistema de detecção de fraude que analisa as imagens de passageiros com algum benefício individual e intransferível, como descontos para estudantes, ao passarem pela catraca. Primeiramente, é feita a detecção de face nas imagens e, em seguida, é realizado o reconhecimento facial, comparando a face detectada com a imagem do cadastro do usuário no banco de dados da empresa de ônibus. Nessa situação, a fraude se caracteriza quando, por exemplo, estudantes compartilham seus cartões de desconto com outras pessoas. Em caso de fraude confirmada, o cartão é bloqueado.

Em um trabalho inicial, realizado pelo autor desta dissertação, foi conduzido um estudo (FERREIRA et al., 2021), no qual testes estatísticos foram considerados para verificar se há correlação entre a taxa de erro no algoritmos de detecção de face e a raça/gênero dos usuários. Com base nos resultados obtidos, foi observada a existência dessa correlação, i.e., há uma maior taxa de erro de detecção facial em usuários pretos ou pardos e mulheres. Entretanto, visando evitar conclusões com base em correlações espúrias, propôs-se um estudo que buscou estudar esse problema com base na investigação de duas questões principais: i) Há, de fato, uma relação causal entre a raça e a taxa de erros na detecção de face? ii) Técnicas de pré-processamento utilizadas em visão computacional são capazes de reduzir essa disparidade?

1.2 HIPÓTESE E OBJETIVO

A investigação das questões apresentadas anteriormente foi conduzida com o uso de conceitos de inferência causal para avaliar a imparcialidade no sistema de detecção de fraudes implantado no transporte público de Salvador.

Inferência causal, por definição, tem por objetivo avaliar o efeito de uma determinada exposição sobre um desfecho de interesse. Nos conceitos de imparcialidade causal, a ideia básica consiste em tratar uma determinada variável (atributo), que idealmente não deveria influenciar o desfecho do cenário observado, como exposição e avaliar seu efeito resultante (PEARL, 2009). Neste trabalho, avaliou-se a influência causal da cor da pele sobre a taxa de detecção de faces e verificou-se a possibilidade de quantificar e controlar, através de técnicas de pré-processamento, esse viés.

Considerando a contextualização e motivações apresentadas, esta dissertação de mestrado foi desenvolvida com base na seguinte hipótese:

O sistema de detecção facial desenvolvido pelo grupo de pesquisa em que o autor desta dissertação pertence é influenciado por técnicas de pré-processamento dependendo da cor de pele do usuário analisado.

Para avaliar essa hipótese, este trabalho tem como objetivo principal criar um modelo causal que permita avaliar a influência da cor de pele e do pré-processamento no sistema de detecção facial utilizado no transporte público de Salvador.

1.3 ORGANIZAÇÃO DO TRABALHO

O estudo apresentado nesta dissertação está organizado da seguinte maneira: no Capítulo 2, é estruturada uma breve introdução sobre inferência causal e suas principais métricas, trabalhos relacionados e, ainda, algumas possibilidades de integração entre inferência causal e o problema da imparcialidade algorítmica; No Capítulo 3 os experimentos e resultados deste trabalho são discutidos em detalhes. Em seguida, as conclusões são apresentadas, juntamente com as vantagens e limitações da investigação e trabalhos futuros.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

Em muitas aplicações, saber se (ou quais) eventos são correlacionados pode não ser o suficiente para compreender o comportamento de sistemas em geral. Correlações podem aparecer em diferentes contextos por múltiplas razões. Por exemplo, pode-se supor que o pertencimento a um determinado grupo de crianças e o aproveitamento escolar ao final do primeiro ano do ensino fundamental sejam positivamente correlacionados. Contudo, não se pode, a partir dessa correlação, inferir qualquer conclusão favorável àquele grupo de crianças. Há razões pelas quais deve-se suspeitar de tal correlação como indicativo causal, i.e., o aproveitamento final pode ser consequência de outros elementos que afetam conjuntamente o desenvolvimento na qual a criança tipicamente se encontrava no início dos seus estudos e sua performance como estudante. Inferência causal aborda questões dessa natureza, ou seja, busca-se avaliar se um determinado evento tem *de fato* algum efeito sobre outro. É importante notar que não se trata puramente de realizar previsões no sentido usual. Trata-se de saber se a alteração nos níveis de exposição relativa a uma característica qualquer implica em mudança nos níveis do desfecho de interesse. Trata-se, portanto, de mensurar essa mudança através do que denominamos *efeito causal* (SPIRTES, 2010; HERNÁN, 2004).

A análise de causalidade em um contexto empírico é complexo, pois a mera observação de uma sequência de eventos, por maior que seja, não garante a existência de uma relação de causa e efeito. No entanto, sob certas condições, o efeito de uma exposição sobre um determinado desfecho pode, sim, ser avaliado. Nesse sentido, as próximas seções serão abordados conceitos fundamentais de inferência causal necessários para a nossa discussão.

2.1 MODELOS CAUSAIS

De maneira geral, pode-se afirmar que há duas grandes linhas de desenvolvimento da inferência causal baseadas em (i) equações estruturais e teoria de grafos (PEARL, 2009), e (ii) em respostas potenciais (IMBENS; RUBIN, 2015). Neste trabalho, o foco é a utilização de grafos para a análise causal da hipótese apresentada. Sendo assim, as próximas subseções são dedicadas à descrição de suas principais características.

2.1.1 Grafos Causais

Na abordagem via modelos estruturais e grafos causais, as relações causais entre as variáveis de interesse são representadas por meio de um conjunto de equações (modelo estrutural) e graficamente via grafos. A Figura 2.1 ilustra uma configuração com 5 variáveis

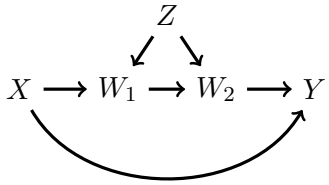

Figura 2.1: Exemplo de grafo causal.

causalmente relacionadas entre si. Nela, a variável X "causa" diretamente W_1 e Y e indiretamente W_2, mas não causa Z. Ou seja, *intervenções* em X impactam diretamente W_1 e Y, e indiretamente W_2, mas não devem impactar o comportamento de Z. O mesmo raciocínio vale em relação a todas outras variáveis. Embora intuitivamente clara, essa representação necessita ser conectada a um aparato matemático para ser interpretada de modo bem definido e para que se torne operacional. Embora os conhecimentos necessário sobre a teoria de grafos para estudar causalidade não estejam muito além do básico (veja, por exemplo, Gould (2012) e Pearl (2009)), é importante conceituar formalmente a noção de grafos para compreender melhor sua conexão com a inferência causal.

Um grafo $\mathcal{G}$ é um objeto matemático composto de *vértices* (ou *nós*) e *arestas* (que conectam os vértices). Formalmente, portanto, pode-se escrever $\mathcal{G} = (\mathcal{X}, \mathcal{A})$ em que

$$\mathcal{X} = \{X_1, ..., X_k\}$$

denota o conjunto correspondente de vértices com k elementos e

$$\mathcal{A} \subseteq \{(X_i, X_j) : i, j \in \{1, \ldots, k\} \text{ e } i \neq j\}$$

representa o conjunto de todas as arestas de $\mathcal{G}$. Nesse contexto, $(X_i, X_j) \in \mathcal{A}$ significa que há uma aresta conectando X_i a X_j. Essas arestas podem ser direcionadas de modo a apresentar vértices de origem e destino. Ou seja, existe a relação direcionada de X_i para X_j se $(X_i, X_j) \in \mathcal{A}$, mas $(X_j, X_i) \notin \mathcal{A}$. Em grafos, arestas direcionadas são representadas por setas "partindo" da primeira variável e "chegando" na segunda variável (e.g. $X_i \to X_j$). Se $(X_i, X_j) \in \mathcal{A}$ e $(X_j, X_i) \in \mathcal{A}$, a aresta é chamada *não-direcionada* e é representada em grafos por um traço ligando as duas variáveis ou por uma seta para ambas variáveis ($X_i \leftrightarrow X_j$). Um grafo no qual *todas arestas* são direcionadas é denominado *grafo direcionado*. Todo grafo causal $\mathcal{G}$ é direcionado de modo que

1. Os vértices representam *variáveis aleatórias* de interesse;

2. As arestas representam a "direção causal" entre as variáveis conectadas.

Além da condição de direcionalidade, um grafo causal $\mathcal{G}$ também deve ser acíclico. Mais precisamente, um grafo com k elementos é considerado cíclico caso possua uma coleção de ℓ vértices $X_1, X_2, \ldots, X_\ell$, tal que $\ell \leq k$, onde $X_i \to X_{i+1}$ para todo $i = 1, \ldots, \ell - 1$ e $X_\ell \to X_1$, ou seja, o último elemento, ℓ, completa o ciclo com uma aresta para o primeiro elemento, como ilustrado na Figura 2.2.

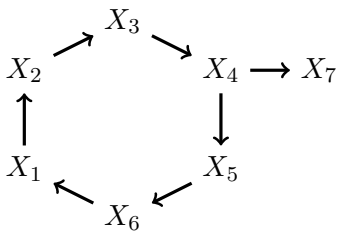

Figura 2.2: Ciclos.

Um grafo que não possua ciclo algum é chamado *acíclico*. Em particular, grafos acíclicos cujas arestas são direcionadas chamam-se *grafos direcionados acíclicos* ou Directed Acyclic Graphs (DAGs). Vale observar que os DAGs, além de acessíveis e de uso simples, tornam as relações de causa ao efeito visíveis e facilmente identificáveis (LüBKE et al., 2020).

Dois vértices X_i e X_j são considerados adjacentes se $X_i \to X_j$ ou $X_j \to X_i$. Um vértice X_i é "pai" de outro vértice X_j (que é "filho" do primeiro) se $X_i \to X_j$. Nesse caso, escrevemos $X_i \in PA_{X_j}$ (X_i pertence ao conjunto dos *pais* de X_j) e $X_j \in CH_{X_i}$ (X_j pertence ao conjunto dos *filhos* de X_i). De maneira semelhante, usamos os termos "ancestral" ($X_i \in AN_{X_j}$) e "descendente" ($X_j \in DE_{X_i}$).

2.1.2 Modelos Estruturais

Como vimos na seção anterior, as relações de causa e efeito em termos de equações podem ser representada via DAGs. Alternativamente, no entanto, ele também pode ser representado através de um sistema de equações estruturais (PETERSEN, 2011). Ou seja, as relações em um DAG podem ser descritas matematicamente por

$$X_i := \sum_{j \neq i} \alpha_{X_i X_j} X_j + U_{X_i}, \ i = 1, ..., k \tag{2.1}$$

Nessa equação, X_i denota a i-ésima variável, $\alpha_{X_i X_j}$ representa o coeficiente que quantifica a influência da variável X_j sobre X_i, sendo zerado na ausência de influência (quando não há aresta direcionada), e U_{X_i} é o termo de erro associado a X_i, que capta influências não modeladas no valor de X_i: ele captura todas as outras influências sobre X_i não explicadas pelas variáveis presentes na análise. Por fim, o índice i varia de 1 a k, onde k é o número total de variáveis no modelo. Ela pode ser utilizada para modelar qualquer número de variáveis e suas inter-relações conforme representado em um DAG, permitindo uma representação matemática da estrutura de dependência entre as variáveis.

Como exemplo, o grafo da Figura 2.3 pode ser representado pelas equações estruturais

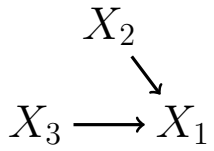

Figura 2.3: DAG exemplo.

$$X_1 := \alpha_{X_1X_2}X_2 + \alpha_{X_1X_3}X_3 + U_{X_1} \tag{2.2}$$

$$X_2 := U_{X_2} \tag{2.3}$$

$$X_3 := U_{X_3} \tag{2.4}$$

Dessa forma, o Directed Acyclic Graph (DAG) é uma representação visual simplificada das dependências entre as variáveis. Nesse exemplo, o grafo mostra três nós (vértices), que representam as variáveis X_1, X_2, e X_3, com setas indicando as relações de dependência entre elas. Especificamente, X_2 e X_3 apontam para X_1, implicando que X_1 é influenciado por X_2 e X_3. Em adição, a ausência de arestas direcionadas tem um significado importante, X_2 e X_3 são independentes: na equação estrutural, seu comportamento é definido por um fator externo não medido que assume-se independente dos demais.

A definição de modelos estruturais causais, conforme apresentada em Peters, Janzing e Schlkopf (2017, Definição 6.2, p. 83-84), segue abaixo.

Definição 1. *(Modelo Causal Estrutural) Um modelo causal estrutural* $\mathfrak{C} := (\mathbf{S}, P_{\mathbf{U}})$ *consiste em uma coleção* $\mathbf{S}$ *de* k *atribuições (estruturais)* $X_i := f_i(\mathbf{PA}_i, U_i)$*;* $i = 1, ..., k$*. Onde* PA_i *são os pais de* X_i*,* U_i *é um ruído aleatório com distribuição* $P_{\mathbf{U}}$ *conjuntamente independente e* f_i *é a função que relaciona o ruído e pais de* X_i *com* X_i*.*

Dadas as atribuições estruturais, pode-se criar um grafo ao atribuir um vértice para cada variável X_i *e desenhando setas direcionadas de seus pais para ela. Nesse caso, os pais de* X_i *podem ser chamados também de causas diretas de* X_i *e* X_i *é um efeito direto de seus pais. Modelo(s) Causal(is) Estrutural(is) (MCE) também são conhecidos como Modelos de Equações Estruturais (não-lineares).*

Note que a Definição 1 se trata apenas de outra forma de representar a Equação 2.1. Como exemplo, a Figura 2.1 possui o modelo causal estrutural representado nas Equações 2.5 a 2.9.

$$X := U_X \tag{2.5}$$

$$W_1 := \alpha_{W_1X}X + \alpha_{W_1Z}Z + U_{W_1} \tag{2.6}$$

$$W_2 := \alpha_{W_2W_1}W_1 + \alpha_{W_2Z}Z + U_{W_2} \tag{2.7}$$

$$Z := U_Z \tag{2.8}$$

$$Y := \alpha_{YX}X + \alpha_{YW_2}W_2 + U_Y \tag{2.9}$$

As variáveis U_{X_i} são os ruídos que afetam cada variável X_i. Apesar de, nesse caso, essa variável estar incluída em todas as equações, por padrão, a literatura sempre assume a existência dessas variáveis e costuma-se omiti-las nos grafos em que assume-se independência entre elas. Dessa forma, possibilita-se uma representação mais reduzida e direta das principais variáveis em análise. Além disso, como uma forma de simplificar a análise, geralmente considera-se que a relação entre as variáveis é linear: assume-se $\alpha_{X_iX_j}$, valor que define a relação direta (causal) de uma variável X_i com uma outra X_j, como constantes. No entanto, não necessariamente são lineares para um caso qualquer.

Em resumo, um conjunto de equações na forma da Equação 2.1, em que cada equação representa um mecanismo autônomo, é o chamado de modelo estrutural; se cada variável tem uma equação distinta na qual aparece no lado esquerdo (chamada de variável dependente), então esse modelo é chamado de modelo causal estrutural ou, simplesmente, modelo causal (PEARL, 2009).

2.2 PROPRIEDADES DOS GRAFOS CAUSAIS

2.2.1 Propriedades Markovianas, pais Markovianos, compatibilidade e cobertura de Markov

Pearl (2009) e Peters, Janzing e Schlkopf (2017) trazem diversas propriedades e teoremas sobre como DAGs se comportam ao utilizá-los como ferramentas para descrever a relação entre variáveis de MCE. De acordo Peters, Janzing e Schlkopf (2017), quando uma distribuição é Markoviana em relação a um grafo, este grafo codifica certas independências na distribuição que podem ser exploradas.

Uma propriedade extremamente importante e prevalente no estudo de causalidade é a propriedade de Markov:

Definição 2. *(Propriedade de Markov) Condicionada a seus pais (causas diretas), cada variável é independente de seus não descendentes. (PEARL, 2009)*

Essa propriedade permite a seguinte fatoração (PETERS; JANZING; SCHLKOPF, 2017):

$$p(\mathcal{X}) = p(X_1, ..., X_n) = \prod_i p(X_i | PA_i^{\mathcal{G}}) \tag{2.10}$$

onde $\mathcal{X}$ é um vetor contendo todos os vértices X_i, e $PA_i^{\mathcal{G}}$ são os vértices que são pais de X_i no grafo $\mathcal{G}$. $PA_i^{\mathcal{G}}$ também são chamados de pais Markovianos (PEARL, 2009, Definição 1.2.1) de X_i. Outra definição importante trazida por Pearl (2009, Definição 1.2.2) é a compatibilidade de Markov:

Definição 3. *(Compatibilidade de Markov) Se uma função de probabilidade P admite a fatoração 2.10 em relação ao DAG $\mathcal{G}$, diz-se que $\mathcal{G}$ representa P, que $\mathcal{G}$ e P são compatíveis, ou que P é Markov (ou Markoviano) em relação a $\mathcal{G}$.*

Peters, Janzing e Schlkopf (2017) também tratam a propriedade chamada cobertura de Markov.

Definição 4. *(Cobertura de Markov) Considere um DAG $\mathcal{G} = (\mathcal{X}, \mathcal{A})$ e um dado nó X_1. A cobertura Markov de X_1 é o menor conjunto M tal que $X_1 \perp\!\!\!\perp_{\mathcal{G}} \mathcal{X} \setminus \{X_1 \cup M\}$. Se $P_{\mathcal{X}}$ for Markoviano em relação a $\mathcal{G}$, então $X_1 \perp\!\!\!\perp \mathcal{X} \setminus \{X_1 \cup M\}$.*

Dessa forma, outras variáveis em $\mathcal{X}$ não fornecem nenhuma informação adicional sobre X_1 ao conhecer $M = PA_{X_1} \cup CH_{X_1} \cup PA_{CH_{X_1}}$. Em um cenário de regressão idealizado, seria necessário apenas incluir as variáveis em M para realizar predições em X_1 (PETERS; JANZING; SCHLKOPF, 2017).

Em outras palavras, entende-se que M, as variáveis que compõem a Cobertura de Markov de X_1, se tratam do conjunto dos pais, filhos e pais dos filhos de X_1. Ao conhecer essas variáveis, outras variáveis no grupo de vértices $\mathcal{X}$ não fornecem nenhuma informação adicional sobre X_1. Por exemplo, na Figura 2.1, a Cobertura de Markov de Y é o conjunto $\{X, W_2\}$. Com essas duas variáveis já é possível estudar o comportamento de Y sem a necessidade de adicionar as demais variáveis à análise.

2.2.2 Estruturas

Grafos com uma grande quantidade de vértices e arestas, apesar de representarem relações complexas, podem ser analisados de forma qualitativa e intuitiva (RAITA et al., 2021). No entanto, para isso, é necessário conhecer e entender as estruturas mínimas e mais comuns de como um grupo de variáveis podem se relacionar. Além disso, torna-se mais visualmente intuitivo sair da representação comum de vértices utilizando apenas uma letra (ex.: $X_1, X_2, ...$) e passar a representar vértices com algum significado específico com letras que indiquem esse significado.

Neste trabalho, as variáveis analisadas como causas são representadas por X, enquanto as que são consideradas efeitos são denotadas por Y. Variáveis que atuam como intermediárias no caminho entre X e Y são representadas por W. Já aquelas que podem introduzir vieses na relação entre X e Y são indicadas por Z. Por fim, para explicitar a presença de variáveis ou ruídos não considerados na análise — seja por impossibilidade de coleta ou por estarem fora do escopo — utiliza-se U.

2.2.2.1 Caminho causal, correntes causais e mediação Dablander (2020) define um *caminho* de X para Y como "uma sequência de nós e arestas que começam e terminam nos nós X e Y, respectivamente". É importante observar que essa definição independe de direcionamento de qualquer aresta. Um DAG conectado em que cada nó tem no máximo um pai é chamado de árvore e uma árvore em que cada nó tem no máximo um filho é chamada de *corrente* (PEARL, 2009). Nesse caso, as notações "pai" e "filho" já implicam direcionalidade, ou seja, uma corrente é um caminho direcionado, portanto, um caminho causal também.

Na Figura 2.4 há dois caminhos causais de X para Y: $X \to Y$ e $X \to W \to Y$. No primeiro, X é uma causa direta de Y. No segundo, que é uma corrente, o efeito de X sobre Y é mediado pela variável W. A soma dos efeitos sobre todos os caminhos causais resulta no efeito total de X sobre Y, considerando que não há variáveis externas não modeladas no grafo.

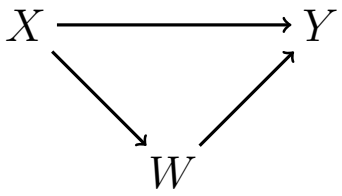

Figura 2.4: DAG de caminhos causais com mediação.

2.2.2.2 Bifurcações, colisores e confundimento A definição de confundimento e variáveis confundidoras pode variar na literatura (VANDERWEELE; SHPITSER, 2013). De forma simplificada, pode-se entender como variável confundidora qualquer variável que possa causar um enviesamento da relação entre duas outras. Como exemplo, a Figura 2.5 em comparação com a Figura 2.4, X deixa de ser uma causa direta de Y, há a inversão da direção da aresta (X, W) para (W, X) e substitui-se a representação da variável W pela letra Z, para manter a consistência de seu significado. Dessa forma, não há mais relação causal entre X e Y, o grafo deixa de ter um mediador do efeito de X, e a variável Z se torna uma causa em comum de X e Y, dado que existem as arestas (Z, X) e (Z, Y).

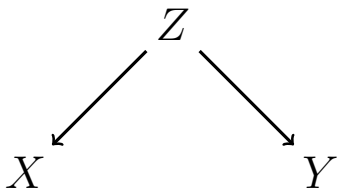

Figura 2.5: Exemplo de DAG de confundimento com bifurcação.

O caminho $X \leftarrow Z \rightarrow Y$, chamado de bifurcação, gera uma correlação espúria entre X e Y que é independente de se há uma relação causal entre as variáveis em questão. Mesmo sem haver a aresta entre X e Y nesse exemplo, ambas ainda são correlacionadas através de Z. Essa correlação espúria, que originou o Paradoxo de Simpson, pode ser desfeita ao condicionar ou estratificar em Z (DABLANDER, 2020; FORNEY; MUELLER, 2021; PEARL, 2014; PETERS; JANZING; SCHLKOPF, 2017). Ela também pode ser representada como um arco tracejado bi-direcionado entre X e Y em situações onde Z é uma variável desconhecida (PEARL, 2009).

Outra estrutura que gera confundimento ou viés em análises é a chamada de colisor. Ao inverter a direção das arestas da Figura 2.5, obtém-se a Figura 2.6, que é o exemplo mínimo de colisor. Nesse caso, ao contrário do exemplo anterior, condicionar ou estratificar em Z levaria a uma correlação espúria entre X e Y. Por padrão, não há correlação entre eles. Esse é o caso conhecido como o Paradoxo de Berkson ou Viés de Seleção (CUMMISKEY et al., 2020; DABLANDER, 2020; PETERS; JANZING; SCHLKOPF, 2017)

Compreender a natureza e o impacto de bifurcações e colisores é importante para a definição de estratégias adequadas de análise. A capacidade de identificar e ajustar corretamente por variáveis confundidoras e remover os colisores da análise é necessária

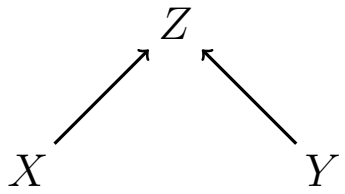

Figura 2.6: Exemplo de DAG de confundimento com colisores.

para ser possível determinar relações de causa e efeito a partir de dados observacionais.

2.2.3 *d*-separação

Pearl (2009) define *d*-separação como:

Definição 5. *Um caminho p é dito ser d-separado (ou bloqueado) por um conjunto de nós Z se e somente se:*

1. p contém uma corrente $X_i \to X_m \to X_j$ ou uma bifurcação $X_i \leftarrow X_m \to X_j$ de modo que o nó do meio X_m está no conjunto Z, ou

2. p contém um colisor $X_i \to X_m \leftarrow X_j$ tal que o nó do meio X_m não está no conjunto Z e tal que nenhum descendente de X_m está em Z.

Diz-se que um conjunto Z d-separa X de Y se e somente se Z bloqueia todos os caminhos de um nó em X para um nó em Y.

Ao compreender as estruturas mais básicas, pode-se afirmar que na Figura 2.5 X e Y são *d*-conectados pelo conjunto vazio e *d*-separados pelo conjunto $\{Z\}$. Em outras palavras, condicionar em Z *d*-separa X e Y: todos os caminhos entre X e Y são bloqueados e a correlação espúria entre eles deixa de existir. Uma análise similar pode ser feita sobre a Figura 2.6: X e Y são *d*-separados pelo conjunto vazio e *d*-conectado pelo conjunto $\{Z\}$. Em outras palavras, condicionar em Z conecta X e Y e gera confundimento.

A Figura 2.7 traz um exemplo um pouco mais complexo. Nesse caso, tanto X quanto Y possuem uma associação espúria com Z através de U_1 e U_2. Observa-se também que o caminho entre X e Y é bloqueado pelo colisor em Z e, portanto, que X e Y são *d*-separados pelo conjunto vazio e *d*-conectados ao condicionar/estratificar em Z, ou seja, X e Y não são causalmente relacionados.

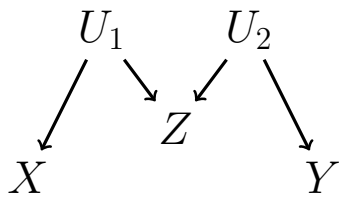

Figura 2.7: Estrutura M espúria.

2.3 OBSERVAÇÕES, INTERVENÇÕES, CONTRAFACTUAIS E O OPERADOR "do"

Seguindo a Hierarquia Causal proposta por Pearl (2019), o estudo começa definindo observações, depois intervenções e, por fim, contrafactuais. Observações são o tipo de informações comumente utilizadas e presentes na estatística usual: informações facilmente visíveis nos dados. As intervenções são os dados que se obtém, por exemplo, em Ensaio(s) Clínico(s) Randomizado(s) (ECR) bem planejados e desenvolvidos. Já contrafactuais são os dados que, por definição, não se tem acesso no mundo real porque, simplificadamente, contrafactual é "o fato que não ocorreu".

Além da impossibilidade do acesso a dados contrafactuais de forma direta, dados intervencionais podem ser de difícil obtenção. Por exemplo, ECRs em pesquisa clínica e epidemiologia psiquiátrica são, muitas vezes, impossíveis de conduzir por uma série de razões logísticas, práticas e éticas. Para ilustrar, durante um estudo do efeito do cigarro sobre o desenvolvimento de câncer, não é possível forçar pessoas a fumar para produzir dados. Por razões como essa, a resposta de muitas questões críticas nessas áreas é inviabilizada (RAITA et al., 2021; OHLSSON; KENDLER, 2020; DABLANDER, 2020). Em adição, mesmo ao considerar as ECRs que são possíveis de serem realizadas, há fatores que podem enviesar os resultados obtidos (OHLSSON; KENDLER, 2020).

Como consequência, a estatística tradicional não consegue responder a perguntas de nível intervencional fora de ECRs, muito menos a nível contrafactual. No entanto, com métodos como os apresentados em Pearl (2009), é possível superar essa limitação considerando os diferentes níveis da Hierarquia Causal. Por exemplo. critérios gráficos permitem decidir, sem olhar os dados, se é possível remover os vieses que levam às correlações espúrias. De maneira complementar, foi desenvolvido o conceito de *do-calculus* para simular intervenções e realizar outros tipos de inferências.

2.3.1 *do*-Calculus

Conforme citado por Peters, Janzing e Schlkopf (2017, p.119), Pearl (2009) desenvolveu o chamado *do-calculus*, que consiste em três regras. Dado um grafo $\mathcal{G}$, com grupo de variáveis X, Y, Z, W cuja representação segue o definido na subseção 2.2.2, as regras são:

1. Inserção/exclusão de observações:

$$p^{\mathfrak{C};do(X:=X_1)}(Y_1|Z_1, W_1) = p^{\mathfrak{C};do(X:=X_1)}(Y_1|W_1)$$

se $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}$, ou seja: Y e Z são d-separados por $\{X, W\}$ em um grafo onde as arestas de entrada em X foram removidas, exemplificado na Figura 2.8.

2. Troca de ação/observação:

$$p^{\mathfrak{C};do(X:=X_1,Z=Z_1)}(Y_1|W_1) = p^{\mathfrak{C};do(X:=X_1)}(Y_1|Z_1, W_1)$$

se $(Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}$, ou seja: Y e Z são d-separados por $\{X, W\}$ em um grafo onde as arestas de entrada em X e as arestas de saída de Z foram removidas, exemplificado na Figura 2.9.

3. Inserção/exclusão de ações:

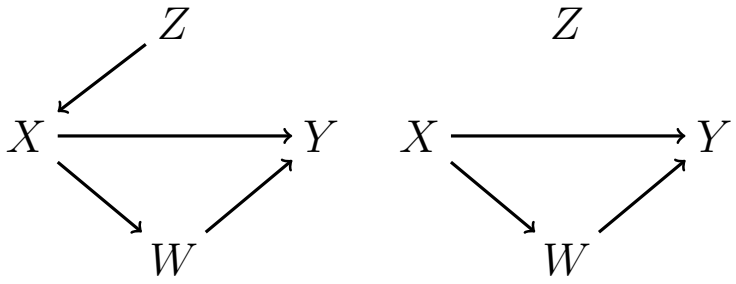

Figura 2.8: Y e Z d-separados com arestas de entrada em X removidas.

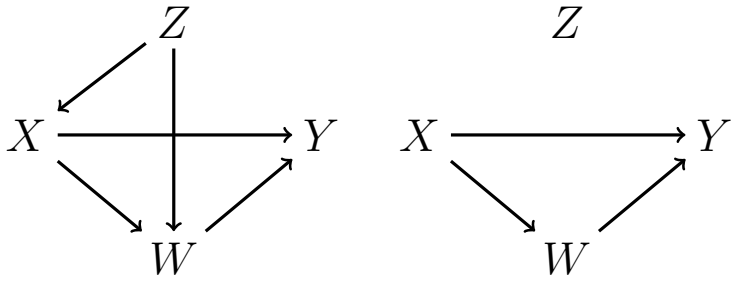

Figura 2.9: Y e Z d-separados com arestas de entrada em X e saída de Z removidas.

$$p^{\mathfrak{C};do(X:=X_1,Z=Z_1)}(Y_1|W_1) = p^{\mathfrak{C};do(X:=X_1)}(Y_1|W_1)$$

se $(Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X},\overline{Z(W)}}}$, ou seja: Y e Z são d-separados por $\{X,W\}$ em um grafo onde as arestas de entrada em X e $Z(W)$ foram removidas. Aqui, $Z(W)$ é o subconjunto de nós em Z que não são ancestrais de nenhum nó em W em um grafo obtido de $\mathcal{G}$ após a remoção de todas as arestas que entram em X. A Figura 2.8 representa essa situação também.

O operador "*do*" simula uma intervenção mínima: $P^{do(X=X_1)}(Y)$ é equivalente ao obtido através de um ECR. Com as regras supracitadas é possível calcular o efeito de uma intervenção, sem efetivamente realizá-la, a partir de qualquer distribuição intervencional identificável. No grafo $X \to Y$, por exemplo, $P^{do(X=x)}(Y)$ é equivalente e exatamente igual à $P(Y|X=x)$.

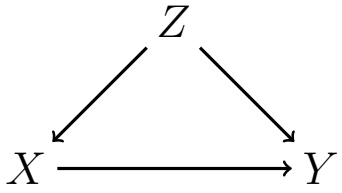

Figura 2.10: Confundimento com bifurcação e relação causal entre X e Y.

De acordo o calculado por Pearl (2009), uma intervenção em X na Figura 2.10 resulta

em

$$P^{do(X=X_1)}(Y_1) = \sum_Z P(Y = Y_1|X = X_1, Z = Z_1)P(Z = Z_1) \tag{2.11}$$

Em grafos maiores, $P^{do(X=X_1)}(Y)$ se torna rapidamente mais complexo de se obter em termos observacionais.

2.3.1.1 Identificabilidade Com a inclusão desse novo operador, considerando seu significado e a forma como ele é utilizado, torna-se necessário abordar a ideia de identificabilidade.

A identificação de um efeito causal consiste em determinar se ele é computável a partir de uma combinação de suposições qualitativas sobre o sistema subjacente (ex.: um gráfico causal) e distribuições coletadas desse sistema (LEE; BAREINBOIM, 2021). Em outras palavras, um problema ser identificável significa que ele é estimável de dados observacionais ao fazer uma sequência de transformações que removem os termos com "*do*" (HUANG; VALTORTA, 2012).

Huang e Valtorta (2012) provam que as regras supracitadas são "completas", no sentido de que, se um efeito causal é identificável, existe uma sequência de aplicações dessas regras que transforma a fórmula do efeito causal em uma que inclui apenas quantidades observacionais.

2.3.2 Critério da porta dos fundos

O critério gráfico da porta dos fundos é caracterizado como uma das soluções para os problemas de confundimento identificados em grafos e caracterizado como uma "solução geral e formal do problema de ajuste usando a linguagem amigável de gráficos causais" (PEARL, 2009). Ele é utilizado para lidar com situações de confundimento como a da Figura 2.10, onde encontra-se uma bifurcação gerando viés na associação entre as duas variáveis de interesse.

Definição 6. *(Portas dos fundos) Um conjunto de variáveis Z satisfaz o critério da porta dos fundos em relação a um par ordenado de variáveis (X, Y) em um DAG $\mathcal{G}$ se:*

1. *nenhum nó em Z é descendente de X; e*
2. *Z bloqueia todo caminho entre X e Y que contém uma aresta entrando em X*

Quando Z bloqueia todos os caminhos das porta dos fundos de X a Y, definir $do(X = X_1)$ tem o mesmo efeito em Y que estratificar os dados em Z e diretamente condicionar em $X = X_1$ (PEARL, 2009), como exemplificado na Equação 2.11.

2.4 ANÁLISE DE MEDIAÇÃO

Conforme visto anteriormente, modelos causais dependem fortemente de premissas estabelecidas *a priori*. Tais premissas podem ser dadas em termos de relações de independência condicional ou através da descrição do grafo causal e do modelo estrutural associado a ele. A compreensão dos mecanismos que compõem o grafo causal vai além das estimativas dos efeitos de uma variável sobre a outra. Pode-se, por exemplo, tentar

entender como esse fluxo causal percorre os diversos caminhos entre uma variável e outra. Ou seja, tomando X como exposição e Y como desfecho, é possível decompor a relação de variação entre X e Y em termos dos (possivelmente) diversos caminhos causais e espúrios que partem da exposição em direção ao desfecho.

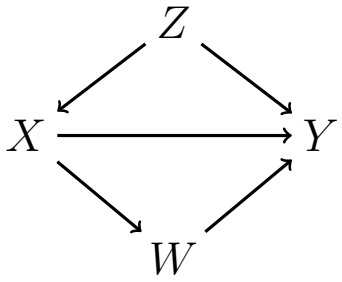

Figura 2.11: Mediação causal.

A Figura 2.11 ilustra uma possível configuração em que a decomposição supracitada pode ser útil. Nela, o caminho $X \to Y$ representa o efeito direto de X sobre Y, o caminho $X \to W \to Y$ representa o efeito indireto de X sobre Y, pois ele é mediado pela variável W, enquanto o caminho de portas dos fundos $X \leftarrow Z \to Y$ representa o efeito/caminho espúrio entre X e Y, pois é confundido por Z. Embora intuitivamente claro, a formalização e avaliação desses efeitos não é trivial. Na sequência, além dos principais conceitos relacionados a esses efeitos, serão apresentados resumidamente as premissas necessárias para a sua identificação e as fórmulas que resultam da sua aplicação. Para mais detalhes e aplicações, recomenda-se a leitura de VanderWeele (2015), Plecko e Bareinboim (2022).

O efeito indireto, que opera através de um fator (ou grupo de fatores) intermediário de interesse (mediador, W), é algumas vezes referido como efeito mediador. O fenômeno pelo qual uma causa (X) afeta um intermediário (W) e a mudança nesse mediador passa a afetar o resultado (Y) é geralmente chamado de fenômeno de mediação. O conjunto de técnicas que um pesquisador utiliza parar avaliar a magnitude relativa desses efeitos diretos e indiretos às vezes é chamado de análise de mediação.

Na análise de mediação é comum tentar entender e calcular o "efeito direto controlado", "efeito natural direto" e "efeito natural indireto". O primeiro expressa o quanto o resultado mudaria em média se o mediador fosse fixado no nível m uniformemente na população, mas o tratamento fosse alterado do nível X_0 para o nível X_1 ($Y_{X_1 m} - Y_{X_0 m}$). O segundo expressa o quanto o resultado mudaria se a exposição fosse definida no nível X_1 versus nível X_0, mas para cada indivíduo o mediador fosse mantido no nível que teria na ausência da exposição ($Y_{X_1 W_{X_0}} - Y_{X_0 W_{X_0}}$). Assim, captando qual seria o efeito da exposição sobre o resultado se o caminho da exposição ao mediador fosse desabilitado. O terceiro, em contraste, captura o efeito da exposição no resultado por via exclusiva do mediador: ele expressa o quanto o resultado mudaria, em média, ao fixar a exposição em X_1 e variar apenas o valor assumido pelo mediador em um universo com a exposição contra um universo sem a exposição ($Y_{X_1 W_{X_1}} - Y_{X_1 W_{X_0}}$).

Entende-se como variação total a variação sobre Y da mudança de $X = X_0$ para $X =$

X_1 ($\mathbb{E}[Y|X_1] - \mathbb{E}[Y|X_0]$). A variação total pode ser subdividida em efeito espúrio e efeito total (estritamente causal). Esse último pode ser subdividido mais uma vez em efeito direto e efeito indireto. Basicamente, o efeito total engloba todo o efeito de X sobre Y através de seus caminhos causais, enquanto a contribuição de cada caminho no efeito total pode ser especificada através do efeito direto e efeitos indiretos. Uma forma de demonstrar que o efeito total é uma soma do efeito direto e dos indiretos é através da seguinte formulação: $Y_{X_1} - Y_{X_0} = Y_{X_1 W_{X_1}} - Y_{X_0 W_{X_0}} = (Y_{X_1 W_{X_1}} - Y_{X_1 W_{X_0}}) + (Y_{X_1 W_{X_0}} - Y_{X_0 W_{X_0}})$.

Os efeitos totais são identificados se, condicional a algum conjunto de co-variáveis medidas Z, o efeito de X sobre Y não é confundido dado Z ($Y_{X_k} \perp\!\!\!\perp X|Z$). Se $Y_{X_1 W_1} \perp\!\!\!\perp X|Z$ e $Y_{X_1 W_1} \perp\!\!\!\perp W|X, Z$, o efeito controlado direto (do inglês *Controlled Direct Effect* (CDE)) médio é identificado e dado por

$$CDE(Z_1) = \mathbb{E}[Y_{X_1 W_1} - Y_{X_0 W_1}|Z_1] = \mathbb{E}[Y|X_1, W_1, Z_1] - \mathbb{E}[Y|X_0, W_1, Z_1] \tag{2.12}$$

Em adição, se $W_{X_1} \perp\!\!\!\perp X|Z$ e $Y_{X_1 W_1} \perp\!\!\!\perp W_{X_0}|Z$ os efeitos naturais direto (*Natural Direct Effect* (NDE)) e indiretos (*Natural Indirect Effect* (NIE)) médios são identificados e dados por:

$$NDE = \mathbb{E}[Y_{X_1 W_{X_0}} - Y_{X_0 W_{X_0}}|Z_1] = \sum_{W_1} \{\mathbb{E}[Y|X_1, W_1, Z_1] - \mathbb{E}[Y|X_0, W_1, Z_1]\} P(W_1|X_0, Z_1) \tag{2.13}$$

$$NIE = \mathbb{E}[Y_{X_1 W_{X_1}} - Y_{X_1 W_{X_0}}|Z_1] = \sum_{W_1} \mathbb{E}[Y|X_1, W_1, Z_1]\{P(W_1|X_1, Z_1) - P(W_1|X_0, Z_1)\} \tag{2.14}$$

2.5 IMPARCIALIDADE CAUSAL

Com o crescimento da utilização de algoritmos de Aprendizado de Máquina (AM) para tomada de decisões que afetam diretamente a vida de pessoas, também houve o crescimento da preocupação da imparcialidade (*fairness*) dos mesmos (KAMISHIMA et al., 2012; PLEISS et al., 2017). Essa preocupação surge justamente por esperar que esses algoritmos não sejam discriminatórios e que as decisões provenientes dos mesmos sejam imparciais (*fair*, ou "justas") (KAMISHIMA et al., 2012; BRUN; MELIOU, 2018). No entanto, há estudos que têm não apenas levantado hipóteses, mas até mesmo demonstrado a existência de certos vieses relacionados a raça, gênero, entre outros.

Kamishima et al. (2012) discorre sobre haver muitos pesquisadores tentando desenvolver técnicas de análise que sejam imparciais/não-discriminatórias. Ele afirma que já foi demonstrado que simplesmente evitar o uso de variáveis protegidas é insuficiente para eliminar vieses em decisões, devido à influência indireta dessas variáveis através de outras que podem ser afetadas pelas mesmas. Ele cita que a imparcialidade de decisões em AM está relacionada à inferência causal porque a decisão final se torna injusta se a decisão depende de uma variável protegida. No entanto, apesar da proposição de técnicas para regularização que consideram essas variáveis, não é feito uma análise causal que descreva

claramente como essas variáveis afetam as decisões tomadas, muito menos as interações delas com outras variáveis.

Brun e Meliou (2018) trazem que há sistemas modernos repletos de exemplos de comportamentos enviesados como por exemplo: tradução injetando estereótipos de gênero, sistemas de visão computacional que não conseguem detectar faces de certas raças, entre outros. Eles falam que vieses nesses sistemas podem emergir de diversas situações diferentes, como: utilização de dados tendenciosos para treinar o algoritmo, especificações de requisitos ambíguos ou incompletos, *bugs* (erros) de implementação e interações não intencionais de componentes, entre outros. Eles argumentam que sistemas de qualidade deveriam ser imparciais e trazem questionamentos e sugestões sobre o que é imparcialidade e como obtê-la.

2.5.1 Noções de Imparcialidade

Na literatura, é possível encontrar diversos tipos de noções/definições de imparcialidade. Com isso, pode ser muito difícil, ou mesmo impossível, utilizar muitas ao mesmo tempo (BRUN; MELIOU, 2018). Por vezes, também, há problemas em que pode ser difícil escolher uma definição de imparcialidade para trabalhar: dependendo da relação entre o atributo em questão e os dados, algumas definições podem na verdade aumentar a discriminação (KUSNER et al., 2017). Alguns exemplos de noções são, em tradução livre, "probabilidades equalizadas", "calibração", "paridade demográfica/impacto díspar", "imparcialidade individual" e "imparcialidade causal" (LOFTUS et al., 2018).

2.5.2 Modelo Padrão de Imparcialidade

O grafo da Figura 2.12 foi nomeado como "Modelo Padrão de Imparcialidade (MPI)" e é proposto como base para desenvolver e estudar a relação entre variáveis em um problema qualquer de forma "reduzida" (ao separar as variáveis em grupos).

A utilização do mesmo é defendida através da ideia que não é necessário conhecer completamente as relações de todos os fatores do problema, apenas as principais. Em outras palavras, um modelo minimalista já é informativo o suficiente para estudar e aproximar o valor real das relações entre as variáveis de um problema (prova no apêndice de Plecko e Bareinboim (2022)).

No MPI, X representa um grupo de variáveis "protegidas" (ou atributos protegidos). Essas são variáveis em que é desejável que não haja influência sobre um (ou um grupo de) desfecho(s) de interesse Y. Como exemplo, é comum desejar imparcialidade na decisão de contratar alguém com relação a raça, religião, gênero, etc, dessa pessoa. W representa os caminhos indiretos (variáveis mediadoras) entre X e Y. É necessário analisar se é aceitável haver influência através desses caminhos para cada problema/atributo. Por fim, Z representa possíveis variáveis que confundem a relação entre X e Y. Nesse contexto X são variáveis em que idealmente não deveriam ter efeito causal sobre Y.

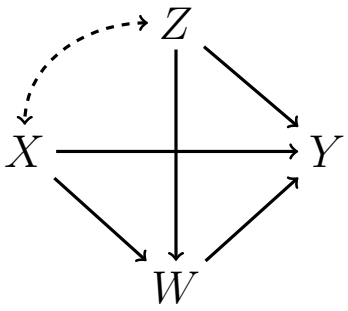

Figura 2.12: Grafo do Modelo Padrão de Imparcialidade.

2.5.3 Medidas de Imparcialidade

Uma vez que o pesquisador tenha definido o MPI, para calcular o impacto do atributo protegido no desfecho é necessário a utilização de métricas adequadas que levem em consideração as relações entre as variáveis. Apesar de que há muitas formas de se calcular esse impacto, muitas podem apresentar um resultado nulo e ainda assim haver um viés não capturado pela métrica (PLECKO; BAREINBOIM, 2022).

Por exemplo, ao se calcular a variação total de Y com relação a uma mudança em X, é possível que o efeito causal se cancele com o espúrio. Ao calcular o efeito total, o efeito direto e indireto podem se cancelar e também resultar em um viés não capturado. Mesmo o NDE e NIE, se utilizados sem uma análise causal minuciosa, nem sempre capturam se há algum viés relacionado ao atributo protegido. Eles também podem ser igual a zero e ainda existir efeito direto/indireto estrutural quando, por exemplo o NDE for uma medida agregada de duas subpopulações distintas em que o efeito nas duas subpopulações se cancela.

O indicado para cálculo de métricas de imparcialidade é, inicialmente, a subdivisão da variação total em três: efeito estrutural direto (*Structural Direct Effect (Str-DE)*), efeito estrutural indireto (*Structural Indirect Effect (Str-IE)*) e efeito estrutural espúrio (*Structural Spurious Effect (Str-SE)*). Essas três métricas devem ser avaliadas separadamente e calculadas preferivelmente dentro de estratos (ex.: cálculo do NIE condicionado a um valor específico de alguma variável). Se houver, por exemplo, um efeito indireto que não for aceitável, o efeito por aquele caminho deve ser zero em todos os estratos. Alguns exemplos de estratos poderiam ser os níveis de X, que seria equivalente a calcular o efeito do tratamento sobre os tratados, os níveis de Z, ambos, etc. O cálculo das métricas de imparcialidade em subpopulações resulta em métricas, em tradução livre, "mais fortes" e que capturam melhor o viés, se houver algum.

2.6 DISCUSSÃO E EXEMPLOS

O objetivo da inferência causal é estimar o efeito de um tratamento em um resultado de interesse. Um efeito causal é definido como a diferença no resultado observado em uma população quando um tratamento é aplicado, em comparação com o resultado na mesma população na ausência desse tratamento (CUMMISKEY et al., 2020). Com a utilização

de grafos para modelar um dado problema, e com as regras do "*do-calculus*", é possível calcular o efeito causal de uma intervenção apenas com dados observacionais.

No entanto, nem todas as situações são solucionáveis com esses métodos. Há casos em que algumas das variáveis, nas quais os métodos gráficos identificam que devem ser controladas (ex.: estratificadas), podem ser impossíveis de medir. Modelar grafos não significa que todos os vieses possíveis serão removidos do problema, apenas significa que a maioria deles será conhecido.

Nas próximas subseções serão apresentados alguns exemplos de análise utilizando dados sintéticos criados com a linguagem de programação "R". X, Y e Z são variáveis aleatórias normalmente distribuídas, onde cada variável é criada com $rnorm(100000)$ somado às variáveis que causam a variável atual e é utilizada a semente (*seed*) "42" para geração dos dados. Para cada exemplo, será demonstrado como diferentes análises de regressão em modelos lineares ("$lm(y \sim x)$" em R, por exemplo) podem se aproximar ou afastar da verdadeira relação entre essas variáveis.

2.6.1 Viés de Seleção/Berkson

O viés de seleção pode ser representado graficamente com o DAG $X \rightarrow Z \leftarrow Y$, onde $X \perp\!\!\!\perp Y$ e Z é um colisor. Assim como inicialmente abordado na Subseção 2.2.2.2, mesmo que X e Y sejam independentes, encontra-se um viés (associação espúria) ao condicionar/estratificar um colisor. A literatura demonstra que o correto é não condicionar e a analise a seguir demonstrará como a estimativa de associação entre X e Y realmente é a correta ao seguir essa recomendação.

Nesse exemplo, foi utilizado dados sintéticos em que X e Y são independentes, mas são d-conectados pelo conjunto $\{Z\}$ conforme a Figura 2.6 e a discussão já realizada sobre a mesma. Ao realizar uma regressão de X em Y obtém-se os resultados da Tabela 2.1

Tabela 2.1: Regressão de X em Y para viés de seleção, com $X \perp\!\!\!\perp Y$.

$lm(Y \sim X)$	Coef. Estimado	D. Padrão
(Intercepto)	-0.001502031	0.003169564
X	-0.002855772	0.003159365

que é a real relação utilizada na geração dos dados: X e Y são independentes. O valor do coeficiente estimado se aproxima de zero: $Y = 0 \cdot X$. Apenas não sendo zero devido a pequenas variações que são comuns nos dados.

Por outro lado, se a variável Z for adicionada à análise, X e Y passam a estar d-conectados e um viés é introduzido. Dessa forma, ao realizar uma regressão de X e Z em Y, observa-se um enviesamento na Tabela 2.2

onde a regressão indica haver uma correlação entre X e Y, nesse caso $Y = -0.5 \cdot X + 0.5 \cdot Z$, mas sabe-se que X e Y são gerados de maneira completamente independente: na verdade Z é causado por X e Y enquanto X e Y são independentes (Figura 2.6). Ou seja, adicionar o colisor à análise gera um enviesamento no resultado da regressão.

Tabela 2.2: Regressão de X e Z em Y para viés de seleção, com $X \perp\!\!\!\perp Y$.

$lm(Y \sim X + Z)$	Coef. Estimado	D. Padrão
(Intercepto)	-0.000348398	0.002242974
X	-0.502858063	0.002739794
Z	0.501089239	0.001587071

2.6.2 Reversão de Simpson/Confundimento

A Reversão de Simpson pode ser representada graficamente com o DAG $X \leftarrow Z \rightarrow Y$, onde $X \perp\!\!\!\perp Y$ e Z é uma bifurcação. Nesse caso, apesar de $X \perp\!\!\!\perp Y$ e isso poder ser constatado visualmente, encontra-se um viés quando Z **não** é condicionado. A literatura demonstra que o correto nessa situação **é condicionar** e a analise a seguir demonstrará como a estimativa de associação entre X e Y realmente é a correta ao seguir essa recomendação.

No exemplo a seguir, foi utilizado dados sintéticos em que X e Y são independentes, mas são d-conectados pela conjunto vazio, conforme a Figura 2.5. Ao realizar uma regressão de X em Y obtém-se os resultados da Tabela 2.3

Tabela 2.3: Regressão de X em Y para confundimento, com $X \perp\!\!\!\perp Y$.

$lm(Y \sim X)$	Coef. Estimado	D. Padrão
(Intercepto)	-0.002126956	0.003884465
X	0.498943167	0.002743088

a qual demonstra um enviesamento: $Y = 0.5X$. No entanto, sabe-se que os dados foram gerados com $X \perp\!\!\!\perp Y$. Esse enviesamento é gerado pela variável Z, que não foi levada em conta nessa análise. Por outro lado, ao adicioná-la na análise, ocorre a d-separação de X e Y a análise retorna à real relação entre X e Y, presente na Tabela 2.4.

Tabela 2.4: Regressão de X e Z em Y para confundimento, com $X \perp\!\!\!\perp Y$.

$lm(Y \sim X + Z)$	Coef. Estimado	D. Padrão
(Intercepto)	-0.000805813	0.003162689
X	-0.003723759	0.003155447
Z	1.004399725	0.004454029

Nesse caso, a regressão resultou em $Y = 0 \cdot X + 1 \cdot Z$, que é a relação real entre as variáveis. Demonstrando mais uma vez como a escolha das variáveis da regressão e o

conhecimento do grafo causal pode afetar o resultado da análise.

Como exemplo final, foi utilizado a Figura 2.10 como base para gerar as equações estruturais e testar o critério da porta dos fundos. Dessa vez, os dados sintéticos são construídos de forma que X causa Y e com o efeito de X sobre Y enviesado por Z: X e Y são d-conectados pelo conjunto vazio e d-separados por $\{Z\}$.

Assim, ao regredir X em Y, obtém-se os dados da Tabela 2.5

Tabela 2.5: Regressão de X em Y para confundimento, com $X \not\perp\!\!\!\perp Y$.

$lm(Y \sim X)$	Coef. Estimado	D. Padrão
(Intercepto)	-0.002126956	0.003884465
X	1.498943167	0.002743088

onde $Y = 1.5 \cdot X$. Este é um efeito enviesado por Z que, nesse caso, a literatura indica que deve ser incluído na análise de regressão. Ao adicionar Z na análise, obtém-se os dados da Tabela 2.6.

Tabela 2.6: Regressão de X e Z em Y para confundimento, com $X \not\perp\!\!\!\perp Y$.

$lm(Y \sim X + Z)$	Coef. Estimado	D. Padrão
(Intercepto)	-0.000805813	0.003162689
X	0.996276241	0.003155447
Z	1.004399725	0.004454029

Nesse caso, obtém-se $Y = 1 \cdot X + 1 \cdot Z$, que é a modelagem correta do efeito simulado com os dados sintéticos.

Dessa forma, fica evidente a importância de conhecer as relações entre as variáveis: a modelagem do grafo e escolha da variáveis da análise afetam diretamente os resultados obtidos e as conclusões que podem ser feitas com os dados. Ao escolher corretamente as variáveis, ou seja, não utilizar colisores e incluir as bifurcações que confundem X e Y, a mesma análise estatística, que inicialmente apenas indica uma correlação entre duas variáveis, passa a informar o verdadeiro efeito causal de X sobre Y.

2.7 TRABALHOS RELACIONADOS

Esta seção apresenta os trabalhos relacionados a este trabalho, abordando a investigação de questões éticas causadas por vieses de gênero e raça, que influenciam o desempenho de métodos de IA, causalidade e imparcialidade no Aprendizado de Máquina (AM). Nesse sentido, Farinella e Dugelay (2012) publicaram um manuscrito focado em entender se gênero e etnia se afetam mutuamente durante o processo de classificação. Segundo os autores, essas características não são afetadas uma pela outra.

O estudo escrito por Mao et al. (2022) investiga as representações visuais em tarefas de reconhecimento de objetos e seu impacto na generalização fora de distribuição. Ele observa que os classificadores de imagens geralmente lutam com novos ambientes devido a correlações espúrias entre recursos e rótulos não robustos. Essa falta de transportabilidade entre as configurações leva a um desempenho ruim em amostras fora de distribuição. Para resolver isso, os pesquisadores propõem um algoritmo que estima o efeito causal da classificação de imagens, que permanece invariável em diferentes ambientes. O algoritmo usa representações de modelos profundos como proxies para derivar o efeito causal sem a necessidade de variáveis adicionais observadas. A abordagem é suportada por análises teóricas, resultados empíricos e visualizações, demonstrando sua capacidade de capturar invariâncias causais e melhorar o desempenho geral de generalização para classificadores de imagens.

No que diz respeito à justiça, Plecko e Bareinboim (2023) exploram o campo da justiça, como um exemplo, sobre a quantificação e medição da discriminação. Eles discorrem sobre a situação na qual várias medidas de justiça foram propostas, mas algumas se mostraram incompatíveis, levando a uma falta de consenso sobre a medida adequada. Isso dificulta as aplicações práticas do aprendizado de máquina justo. O artigo se concentra em um resultado de impossibilidade relacionado à paridade estatística e preditiva, derivando uma nova fórmula de decomposição causal para medidas de justiça e fornecendo percepções sobre suas conexões com doutrinas jurídicas como tratamento díspar, impacto díspar e necessidade comercial. Os resultados revelam que a paridade estatística e preditiva não são mutuamente exclusivas, mas complementares e formam um espectro de noções de justiça através do conceito de necessidade do negócio. O artigo também demonstra o significado prático dessas descobertas com um exemplo do mundo real.

O trabalho de Kumar, Kaur e Kumar (2019) fala sobre o aumento dos bancos de dados de vídeo e imagem que criou uma necessidade premente de sistemas inteligentes para entender e analisar informações automaticamente, pois o processamento manual se torna impraticável. Eles trazem à tona o fato de que os rostos são essenciais para transmitir identidade e emoções nas interações sociais e argumenta que as máquinas podem identificar rostos melhor do que os humanos, tornando os sistemas de detecção automática de rosto cruciais para aplicações como reconhecimento facial, reconhecimento de expressão facial, estimativa de pose da cabeça, e interação humano-computador. Como tem sido um tópico proeminente na literatura de visão computacional, os autores apresentam uma pesquisa abrangente de várias técnicas de detecção facial e explora os desafios e aplicações da detecção facial. Ele também lista diferentes bancos de dados padrão para detecção de rosto e seus recursos. O artigo conclui com discussões sobre aspectos práticos para o desenvolvimento de sistemas robustos de detecção de face e sugere direções promissoras para pesquisas futuras.

Furl, Phillips e O'Toole (2002) conduziram um estudo para verificar a precisão de modelos de reconhecimento facial, lidando com diferentes etnias. Os resultados foram focados em caucasianos e asiáticos. Phillips et al. (2011) discutiram os riscos de usar rostos com características diferentes durante as fases de treinamento e teste.

Trabalhos semelhantes foram publicados por Klare et al. (2012), cuja principal contribuição foi estender a análise para diferentes características demográficas, como gênero,

raça e idade. Os autores identificaram que mulheres negras e jovens apresentaram as maiores taxas de erro. Para superar essa situação, eles recomendam a utilização de conjuntos de dados balanceados, com amostras representativas da população onde o sistema será implantado. Além disso, os autores sugerem a criação de conjuntos de dados públicos para apoiar sistemas desenvolvidos para cenários semelhantes, garantindo uma melhor representatividade e reduzindo potenciais vieses no desempenho do sistema.

Manuscritos mais recentes mostram contribuições na organização de conjuntos de dados, bem como na investigação da existência de um viés racista. Nesse sentido, Karkkainen e Joo (2021) focaram em desenvolver um conjunto de dados balanceado, chamado FairFace, que contém informações sobre sete grupos raciais: Branco, Negro, Indiano, Leste Asiático, Sudeste Asiático, Oriente Médio e Latino. Esse conjunto de dados foi comparado com outros na literatura e usado em APIs comerciais, trazendo melhorias na tarefa de classificação racial.

Buolamwini e Gebru (2018) exploraram três modelos criados para a detecção de rostos, utilizados em soluções comerciais (Microsoft, IBM, Face++), para avaliar a presença de preconceito de gênero e raça. Sua contribuição mais significativa para a literatura foi a criação de um conjunto de dados balanceado por gênero e etnia, integrando outras imagens públicas. Os resultados apresentados pelos autores mostram que maiores taxas de erro foram obtidas no grupo composto por mulheres negras.

Wang et al. (2019) avaliaram o viés racista em raças/etnias, desenvolvendo um conjunto de dados e um modelo capaz de reduzir o viés racista no reconhecimento facial. O conjunto de dados criado, chamado Racial Faces in-the-Wild (RFW), possui etnias caucasianas, indianas, asiáticas e africanas. Os experimentos mostraram que o reconhecimento facial de africanos apresentou taxas de erro duas vezes maiores do que o reconhecimento facial de caucasianos.

Raji et al. (2020) investigaram a ética nas tecnologias de processamento facial desenvolvidas por Amazon, Microsoft e Clarifai, verificando a precisão de seus sistemas em classificar pessoas de acordo com raça, gênero, idade, expressão e detecção facial. Todos os experimentos foram realizados utilizando um conjunto de dados criado pelos autores, composto apenas por celebridades. O principal resultado dessa pesquisa é o alerta sobre a importância de uma análise adequada das questões éticas nos modelos considerados.

2.8 CONSIDERAÇÕES FINAIS

O capítulo aborda a importância da inferência causal na compreensão das relações entre eventos e como distinguir correlação de causalidade. Através de modelos causais e *Directed Acyclic Graphs (DAGs)*, discute-se como identificar e mensurar efeitos causais, introduzindo conceitos como intervenções, contrafactuais e o operador "do". Exemplos práticos demonstram o impacto de variáveis confundidoras e a importância de escolher corretamente as variáveis para análise. A seção sobre imparcialidade causal destaca a relevância de considerar atributos protegidos em algoritmos de aprendizado de máquina para evitar vieses discriminatórios. A seção sobre exemplos sintéticos ilustra a aplicação de análises causais em situações de viés de seleção e Reversão de Simpson, ressaltando como a análise causal pode revelar as verdadeiras relações entre variáveis e evitar in-

terpretações equivocadas. E o capítulo conclui com a apresentação de alguns trabalhos relacionados. O próximo capítulo se trata dos experimentos e resultados obtidos a partir dos mesmos, utilizando o que foi abordado nesse capítulo como base para decisões e análise dos resultados.

Capítulo 3

EXPERIMENTOS E RESULTADOS

3.1 CONSIDERAÇÕES INICIAIS

O grupo de pesquisa, no qual o autor dessa dissertação está inserido, desenvolveu um sistema completo de identificação e reconhecimento facial para detecção de fraude no uso de cartões com benefícios no transporte público de Salvador (Bahia, Brasil).

Esse sistema analisa imagens capturadas em todos os 1.680 ônibus que atuam nas 345 linhas distribuídas na cidade. De maneira resumida, o objetivo desse sistema é realizar a detecção de fraudes quando passageiros passam pela catraca usando irregularmente cartões que dão direito a algum benefício individual e intransferível, como descontos para estudantes.

O comportamento geral do sistema está ilustrado na Figura 3.1. Inicialmente, o sistema aplica um conjunto de técnicas de pré-processamento comumente utilizadas em Visão Computacional como, por exemplo, a redução da imagem e a transformação de seus canais de cores. Em sistemas computacionais, uma imagem é representada como um tensor na forma $[C \times M \times N]$, sendo que C representa o número de canais de cores e $M \times N$ refere-se ao tamanho da imagem ou à quantidade de *pixels* usada para representá-la. Neste exemplo, foram aplicadas as transformações nos canais de cores de $C = 3$ para $C = 1$, i.e., saindo de uma imagem colorida para escala de cinza, e reduzindo a imagem para $m \times n$, tal que $M > m$ e $N > n$.

Após o pré-processamento, uma Rede Neural Profunda - Deep Neural Network (DNN), especificamente uma arquitetura de Rede Neural Convolucional - Convolutional Neural Network (CNN), é utilizada para detectar faces nas imagens de uso e cadastro. Por fim, uma nova CNN é utilizada para extrair características da face detectada. A principal propriedade dessa rede neural é a ausência da camada final de classificação, ou seja, utiliza-se apenas os vetores de características produzidos das camadas intermediárias. Na fase final, os vetores de característica dessa CNN são comparados entre si usando uma métrica de distância. Se a distância resultante d for maior quem um limiar λ, então um alerta de fraude é disparado.

Ao considerar que a cidade em que esse sistema está implantado é composta por uma população majoritariamente negra, qualquer erro na detecção da face devido a um

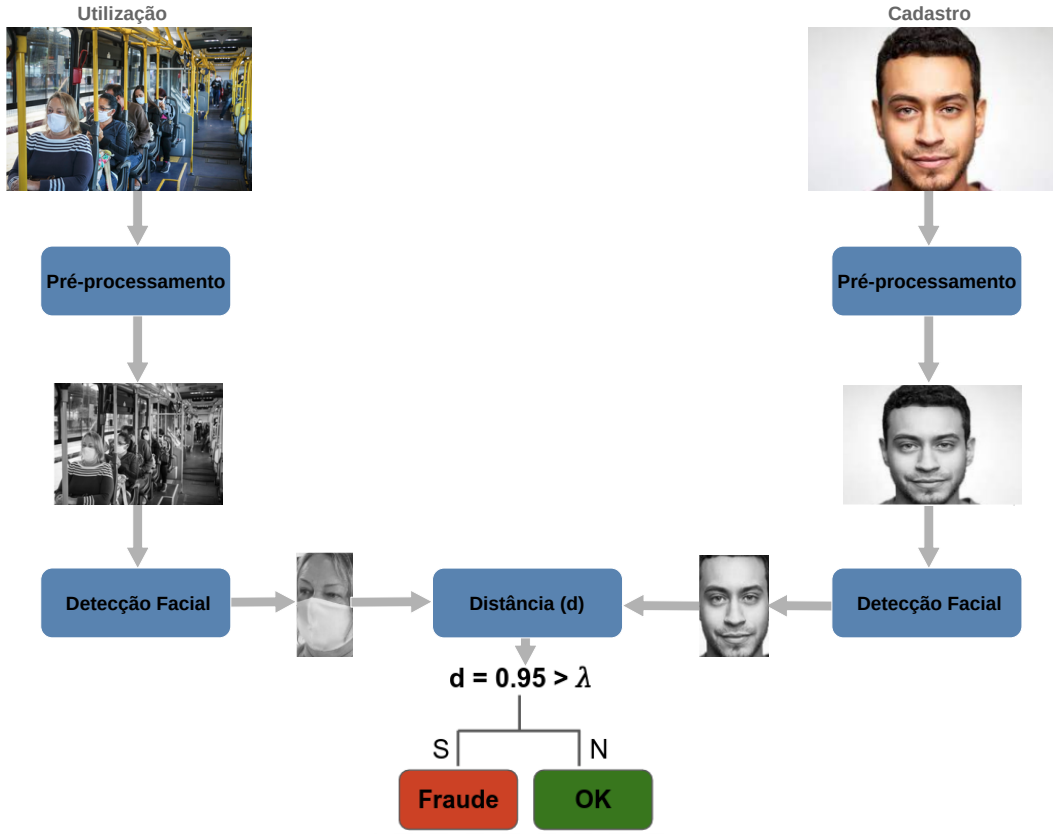

Figura 3.1: Etapas do processo de reconhecimento facial.

viés racial pode impactar na produção de alertas de fraude. Falsos positivos de fraude e errôneos bloqueios de cartões devido a um viés são situações fortemente indesejadas. Sendo assim, as recentes preocupações com a imparcialidade de algoritmos e o viés racial notado em sistemas de detecção e reconhecimento facial motivou a problemática desta pesquisa e a proposta de um modelo causal para estudo da imparcialidade na situação supracitada. Esse modelo busca avaliar se a cor da pele influencia causalmente a taxa de detecção de faces, assim como verificar a possibilidade de quantificar e controlar esse viés usando técnicas de pré-processamento.

A hipótese deste trabalho, definida no Capítulo 1, diz que *"o sistema de detecção facial desenvolvido pelo grupo de pesquisa é influenciado por técnicas de pré-processamento dependendo da cor de pele do usuário analisado"*. Visando comprovar essa hipótese, criou-se neste trabalho um modelo causal que permite avaliar a influência da cor de pele e do pré-processamento no sistema de detecção facial utilizado no transporte público de Salvador.

Para alcançar esse objetivo principal, esse trabalho foi dividido em 4 etapas. A primeira etapa (Seção 3.2) consistiu em realizar a modelagem do problema apresentado, descrevendo a relação das possíveis variáveis que afetam a taxa de detecção de faces por intermédio de um Directed Acyclic Graph (DAG). Essa modelagem seguiu as indicações

de Plecko e Bareinboim (2022), que visam em desenvolver um Modelo Padrão de Imparcialidade (MPI) capaz de descrever o problema de forma "reduzida", mas permitindo, ainda, avaliar as influências das variáveis de interesse sobre o desfecho.

Na segunda etapa (Seção 3.3), realizou-se a coleta de dados (imagens), a execução o algoritmo de identificação de faces e a anotação manual dos dados relacionados a cor de pele e correta detecção das faces.

Na terceira etapa (Seção 3.4), foram selecionadas e aplicadas diferentes técnicas de processamento de imagens para medir seu efeito no sistema. Por fim, na quarta etapa, foram aplicadas técnicas para estudo da taxa de erro de detecção, avaliando o efeito total, i.e., foi quantificado o efeito do atributo protegido e diferentes métodos de pré-processamento, sobre o desfecho. Nessa etapa, utilizou-se do modelo causal proposto e a teoria apresentada na fundamentação teórica como base para considerar o efeito encontrado, que a priori seria uma correlação, como efeito causal. Dessa forma, foi avaliado o efeito da cor de pele sobre a taxa de detecção facial para cada método.

Com o resultado das etapas acima, foi conduzido um processo de avaliação (Seção 3.5) para quantificar o possível viés de detecção facial, juntamente com uma sugestão de uma forma de reduzi-lo. A contribuição mais importante deste trabalho é apresentar um modelo causal, detalhado na seção seguinte, que pode ser usado, e ampliado, para apoiar a implantação de detectores faciais focados na imparcialidade dos sistemas de decisão.

3.2 REPRESENTAÇÃO DO PROBLEMA COMO UM MODELO CAUSAL

Utilizando como base o Modelo Padrão de Imparcialidade (MPI), o qual foi apresentado na Figura 2.12, desenvolveu-se o grafo da Figura 3.2 para descrever as relações dos atributos relevantes neste estudo de forma objetiva. Os atributos são representados da seguinte maneira:

1. "Cor da pele" é a variável protegida X;
2. "Horário" e "Luminosidade" são as variáveis que armazenam o instante da coleta da imagem e as condições de iluminação, consideradas como variáveis confundidoras Z_1 e Z_2, respectivamente;
3. "Pré-processamento" e "Imagem" são os mediadores W_1 e W_2, respectivamente;
4. "Rosto identificado" é o desfecho Y.

De acordo com o conhecimento do especialista que desenvolveu e implantou o sistema em questão, nenhum tipo de informação externa é utilizada na identificação de faces além dos *pixels* da imagem. Isso implica que, considerando o sistema final, todos os caminhos no grafo até o desfecho devem passar pela imagem (W_2), ou seja, não há efeito direto de X sobre Y (*Str-DE*=0) e não há confundimento entre X e Y (*Str-SE*=0).

Dentre os demais atributos, acredita-se que um fator importante na composição da imagem é a iluminação do ambiente, representada pela variável "Luminosidade". A captura dessa informação é realizada por meio de um modelo de IA que automaticamente

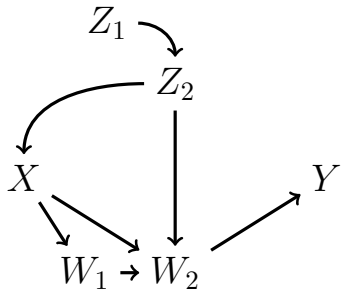

Figura 3.2: Modelo Padrão de Imparcialidade do problema avaliado.

classifica a imagem de acordo com diferentes condições de iluminação. Devido às características do ambiente em que as fotos são tiradas, há grandes variações de iluminação, sendo o horário um dos grandes fatores para sua determinação. A iluminação afeta a percepção de cor de pele (mais luz implica maior intensidade dos *pixels* e percepção de pele mais clara). A cor de pele afeta a composição final da imagem e pode afetar o pré-processamento, também influenciando a imagem final. Como um exemplo, a equalização de histograma depende da frequência dos *pixels*: diferentes cores de pele implica diferentes frequências de *pixels* que implica em diferentes equalizações.

Um exemplo da influência da luminosidade na imagem pode ser observada na Figura 3.1, na qual a foto frontal do rosto de cadastro possui uma fonte de luz diretamente à sua frente. Nesse caso, a parte mais central da face está muito mais clara que as bordas. Dessa forma, dependendo da direção da fonte de luz e sua intensidade, o rosto possuirá pontos mais claros ou escuros e sombreamentos diferentes. Uma equalização de histograma ajusta a frequência dos pixels de forma diferente para cada tipo de iluminação, mesmo que a foto seja retirada de maneira exatamente igual.

3.3 COLETA DE DADOS

As imagens utilizadas nos experimentos foram coletadas do transporte público de Salvador. Quando um passageiro passa pela catraca dentro dos ônibus, quatro fotos são tiradas e armazenadas em um banco de dados. Desta base de dados, selecionou-se aleatoriamente 10.000 imagens capturadas em 10 de dezembro de 2019. Esta data era um dia normal de trabalho e foi considerada sem qualquer motivo relevante, mas o fato de ser anterior ao surto da pandemia de COVID-19 para evitar a análise de imagens de usuários usando máscaras, o que pode afetar o desempenho dos detectores faciais selecionados. Inicialmente foram rotuladas, manualmente, mais de 1000 imagens, informando se havia rosto, a raça e o gênero, ignorando imagens com a qualidade extremamente baixa, repetidas ou com oclusão da face. Para esse estudo, foram descartadas as imagens sem rosto e imagens de qualidade baixa (extremamente escuras/claras/desfocadas).

A rotulação da imagem foi feita através de uma ferramenta criada, em um trabalho anterior, especificamente para auxiliar as tarefas de definição dos rótulos de gênero e raça e validação dos resultados do detector facial, conforme apresentado na Figura 3.3. Ressalta-se que, embora tenha-se conhecimento das diferentes classes de raça adotadas

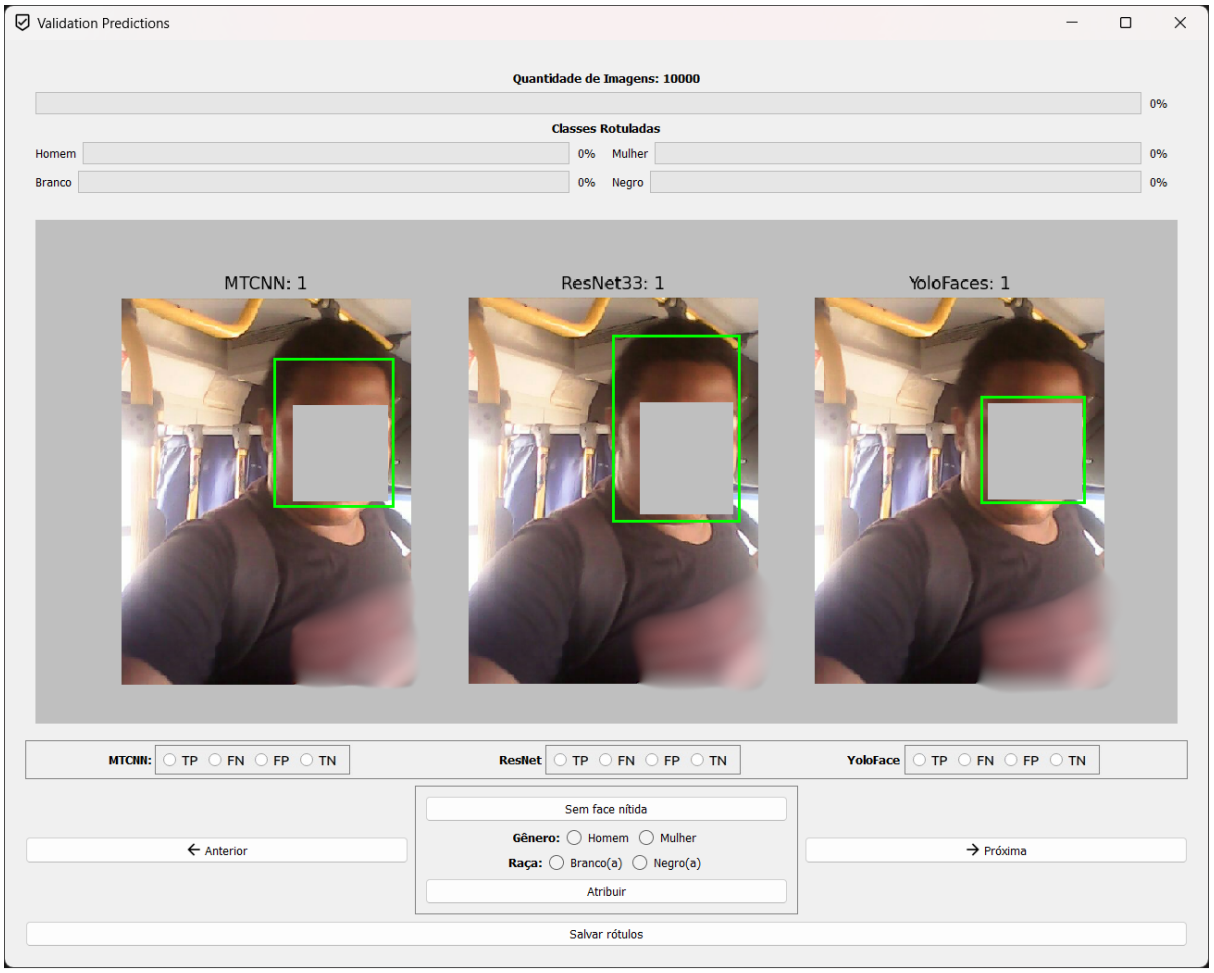

Figura 3.3: Interface GUI para imagens e previsões de modelos.

pelo IBGE (Instituto Brasileiro de Geografia e Estatística), o escopo deste estudo limitou-se apenas a negro e branco.

3.4 MODELOS DE DETECÇÃO FACIAL

Os modelos de detecção facial YoloFace, MTCNN e ResNet foram selecionados para estudar seus respectivos comportamentos e possíveis vieses de detecção.

YoloFace é um modelo de detecção de objetos de estágio único com ênfase na velocidade de inferência, baseado no modelo YOLOv3 (CHEN et al., 2021). Ao contrário de outras arquiteturas de modelo que operam em dois estágios: i) usando um subconjunto de rede para processar o primeiro estágio, a localização de objetos, e ii) usando outro subconjunto para processar o segundo estágio, a classificação de objetos. A análise de estágio único do Yolo é possível porque divide a imagem em uma grade, gerando previsões para cada sub-região e limitando a localização e o nível de confiança dos objetos encontrados. A seguinte arquitetura Yolo foi usada: vinte e seis camadas convolucionais, quatro pooling e duas totalmente conectadas. Além disso, o modelo pré-treinado usado

neste trabalho foi treinado com o conjunto de dados Wider Face (YANG et al., 2016).

MTCNN (Multi-task Cascaded Convolutional Networks) é um modelo de detecção e alinhamento de faces (ZHANG et al., 2016) que opera em cascata com três redes neurais convolucionais, cada uma refinando as detecções para melhorar a precisão. Primeiramente, a CNN P-Net (Proposal Network) visa localizar regiões candidatas e possui três camadas convolucionais, com dez, dezesseis e trinta e dois filtros, seguidas por uma camada de pooling. Em seguida, a CNN R-Net (Rede de Refinamento) refina as detecções de P-Net para melhorar a precisão e é composta por três camadas convolucionais, com vinte e oito, quarenta e oito e sessenta e quatro filtros, duas camadas de agrupamento e uma totalmente camada conectada. Por fim, a cascata O-Net (Output Network) realiza a detecção final de faces e marcos faciais (olhos, nariz e boca). O O-Net tem quatro camadas convolucionais, com trinta e dois, sessenta e quatro, sessenta e quatro e cento e vinte e oito filtros, duas camadas de agrupamento e uma camada totalmente conectada. O MTCNN utilizado foi treinado com o conjunto de dados Face Detection e Benchmark (JAIN; LEARNED-MILLER, 2010) e Wider Face dataset (YANG et al., 2016).

O último modelo utilizado foi a arquitetura ResNet, baseada na implementada no framework Dlib (KING, 2009), projetada para detectar e reconhecer faces. A grande vantagem do ResNet é o aprendizado residual (otimização de gradiente e ajuste de peso), permitindo treinar uma rede com várias camadas sem degradar a taxa de aprendizado. A arquitetura baseada em ResNet usada tem trinta e três camadas convolucionais e totalmente conectadas. No entanto, o ResNet utilizado neste trabalho teve o número de camadas convolucionais ajustado para vinte e seis para otimizar a tarefa de detecção de faces. Por fim, o modelo pré-treinado foi treinado com o conjunto de dados scrub (NG; WINKLER, 2014), conjunto de dados VGG (PARKHI; VEDALDI; ZISSERMAN, 2015) e imagens coletadas por pesquisadores do framework Dlib.

3.5 AVALIAÇÃO

A metodologia de avaliação empregada neste estudo adere às práticas convencionais do campo de Visão Computacional (CV) para tarefas de detecção de objetos. Em resumo, a detecção é avaliada comparando-se as áreas entre as caixas delimitadoras (bounding boxes) esperadas (anotações pré-definidas – G) e as previstas (predições realizadas pelo modelo – P). Ao considerar essas áreas, utilizou-se a razão de Interseção sobre União (IOU) conforme definido na Equação 3.1 para estimar os números de verdadeiros positivos (VP), falsos positivos (FP) e falsos negativos (FN). Essa estimativa é baseada em um limiar, que funciona como um intervalo de confiança para aceitar ou rejeitar a previsão. Em suma, se $IOU \geq \tau$, há sobreposição suficiente entre a área prevista e a anotação pré-definida para classificar o objeto detectado como verdadeiro positivo (TP). Se a área prevista não tem sobreposição suficiente com a anotação pré-definida, um falso positivo (FP) é detectado. De forma similar, quando não é gerada uma caixa de previsão quando existe uma anotação pré-definida, é considerada um falso negativo (FN). Essa investigação avaliou a taxa de erro na classe positiva (detecção de rostos).

$$IOU = \frac{G \cap P}{G \cup P} \tag{3.1}$$

Da mesma forma que (FERREIRA et al., 2021), os três detectores foram executados, mas este estudo se concentra em um subconjunto dos dados utilizados nesse estudo inicial. O subconjunto utilizado contém apenas as imagens mais nítidas do conjunto de dados usado naquele estudo anterior. Considerando os objetivos deste estudo, foram utilizadas apenas imagens que continham pelo menos um rosto. Um conjunto de algoritmos de pré-processamento foi selecionado e executado nas imagens e, com isso, os detectores também foram executados novamente nessas imagens pré-processadas. Como consequência, cada detector produziu novas imagens com caixas delimitadoras desenhadas nas regiões onde as faces foram estimadas. As caixas delimitadoras nas imagens pré-processadas foram classificadas como desenhadas corretamente ou não por sua Interseção sobre União (IoU) sendo maior que 0,5 com as detecções anotadas do conjunto de dados original.

3.6 ERROS PERCENTUAIS

Em termos de valores absolutos, o YOLO se destaca como o de melhor desempenho, conforme pode ser observado na Tabela 3.1. Notavelmente, quando nenhum filtro é aplicado, o erro absoluto é mais pronunciado para tons de pele mais escuros. Observou-se uma redução na taxa de erro com filtros e que esses filtros podem afetar de forma diferente por luminosidade e hora do dia. Entre os diferentes filtros testados, o filtro 4x4 apresenta o pior desempenho, enquanto o filtro 8x8 surge claramente apresentando o melhor desempenho geral. Há uma redução na taxa de detecção de pessoas com pele mais clara em ambientes claros e pessoas com pele mais escura em ambientes mais escuros.

A Tabela 3.1 contém os resultados de erro percentual obtidos ao executar os modelos (YoloFace, MTCNN, ResNet) sobre as imagens pré-processadas com um determinado filtro. Os filtros são Raw (a imagem sem pré-processamento), clipLim1-tile4x4, clipLim2-tile8x8 e clipLim3-tile16x16 (diferentes parâmetros selecionados) respectivamente. A coluna Raça separa o conjunto de dados em Negro e Branco. As colunas a seguir são referentes aos estratos dos subconjuntos, definidos pela análise causal do problema em questão, que podem possivelmente confundir a relação entre a variável protegida, raça, e a variável de interesse, erro percentual. Utilizando a primeira linha da tabela como exemplo, a coluna Erro Base é o erro percentual do modelo YoloFace nas imagens, sem pré-processamento (Raw), de rostos negros. A coluna Baixa L. (Baixa Luminosidade) é o erro percentual no subconjunto de imagens com luminosidade abaixo da média de luminosidade das imagens dos conjunto. A coluna Alta L. (Alta Luminosidade) se trata do erro percentual no subconjunto de imagens com alta luminosidade. As colunas Manhã, Tarde e Noite seguem a mesma ideia para os diferentes turnos do dia, 05–12 horas, 12–18 horas e 18–23 horas, respectivamente.

Os valores apresentados nessas colunas foram definidos considerando o modelo causal definido, no qual foi utilizada a variável protegida X como "Raça", as variáveis de confusão Z_1 e Z_2 como dados de tempo e luminosidade, e o mediador W_1 como filtros. As imagens rotuladas usadas nos experimentos são representadas pela variável mediadora W_2 e as variáveis de resultado Y são representadas pelas taxas de erro de detecção. Os resultados obtidos com o MTCNN e ResNet foram incluidos, no entanto seus erros em todos os cenários foram significativamente elevados e nenhuma conclusão pode ser tirada

deles. Dessa forma, são apresentados apenas para ilustrar outros modelos e a discussão principal será focada nos resultados do YoloFace.

Observando os dois primeiros conjuntos de experimentos (Linhas 1 e 2), é possível notar o desempenho do YoloFace sem filtro após analisar usuários negros e brancos, respectivamente. Os erros dos usuários negros foram maiores que os brancos em todas as situações, exceto quando a imagem se caracteriza por apresentar alta luminosidade. Após utilizar o primeiro filtro (ClipLim1-tile4x4), nota-se que o YoloFace reduziu significativamente os erros em todos os cenários. Além disso, é possível notar comportamento similar mas com erros mais significativos em usuários negros. Ao utilizar o segundo filtro, os valores de erro são ainda menores, sendo semelhantes ao comparar as duas raças. O último filtro apresentou comportamento muito semelhante ao primeiro em termos de taxas de erro e desempenho nos cenários analisados. Tanto no primeiro quanto no terceiro filtro, os usuários brancos apresentaram maiores erros em alta luminosidade e no turno da manhã.

Tabela 3.1: Erros percentuais por filtro, modelo, raça e subconjuntos.

Filtro	Modelo	Raça	Erro Base	Baixa L.	Alta L.	Manhã	Tarde	Noite
Raw	YoloFace	Negro	0.2682	0.3878	0.2082	0.2279	0.3005	0.3636
		Branco	0.2482	0.301	0.2179	0.2098	0.2821	0.3182
Filtro 1	YoloFace	Negro	0.1682	0.2449	0.1297	0.1302	0.202	0.2273
		Branco	0.1383	0.1456	0.1341	0.1678	0.1197	0.0455
Filtro 2	YoloFace	Negro	0.1205	0.1633	0.099	0.1023	0.1478	0.0455
		Branco	0.1241	0.1165	0.1285	0.1469	0.1111	0.0455
Filtro 3	YoloFace	Negro	0.1523	0.2041	0.1263	0.1302	0.1823	0.0909
		Branco	0.1489	0.1456	0.1508	0.1678	0.1453	0.0455
Raw	MTCNN	Negro	0.6455	0.6871	0.6246	0.6372	0.6453	0.7273
		Branco	0.5993	0.6408	0.5754	0.5944	0.5812	0.7273
Filtro 1	MTCNN	Negro	0.6295	0.6803	0.6041	0.6186	0.6256	0.7727
		Branco	0.5638	0.5728	0.5587	0.5664	0.547	0.6364
Filtro 2	MTCNN	Negro	0.6295	0.6803	0.6041	0.6279	0.6158	0.7727
		Branco	0.5567	0.5728	0.5475	0.5664	0.547	0.5455
Filtro 3	MTCNN	Negro	0.6477	0.6871	0.628	0.6512	0.6404	0.6818
		Branco	0.6418	0.6893	0.6145	0.6364	0.641	0.6818
Raw	ResNet	Negro	0.6886	0.5646	0.7509	0.7163	0.6847	0.4545
		Branco	0.7624	0.6311	0.838	0.7762	0.7863	0.5455
Filtro 1	ResNet	Negro	0.7636	0.6599	0.8157	0.8047	0.7438	0.5455
		Branco	0.805	0.6893	0.8715	0.8042	0.8462	0.5909
Filtro 2	ResNet	Negro	0.8477	0.7823	0.8805	0.8465	0.8522	0.8182
		Branco	0.8582	0.7767	0.905	0.8462	0.8889	0.7727
Filtro 3	ResNet	Negro	0.8318	0.7891	0.8532	0.814	0.8473	0.8636
		Branco	0.8688	0.8058	0.905	0.8462	0.906	0.8182

Existem situações em que o pré-processamento pode ter efeitos diferentes em cada cor de pele. Por exemplo, ao analisar a coluna do estrato “Manhã” no modelo YOLO,

observa-se que o uso de imagens sem pré-processamento (Raw) levam a uma taxa de erro mais elevada em tons de pele mais escuros. Em contrapartida, obtêm-se uma taxa de erro maior em tons de pele mais claros com qualquer uma das configurações de filtro apresentadas. Já no turno da noite, os parâmetros podem afetar significativamente a taxa de detecção de indivíduos com pele mais escura.

Além disso, apesar da taxa de erro mais elevada nos demais modelos apresentados, é possível notar algumas equivalências. Como, por exemplo, há uma tendência do YoloFace e MTCNN possuírem uma taxa de erro maior em imagens de baixa luminosidade e à noite, especialmente em tons de peles mais escuros.

3.7 TESTES DE HIPÓTESE

Além das diferenças nos erros absolutos, também é possível analisar em qual cenário as variáveis de confusão e os filtros (mediadores) eram estatisticamente mais relevantes para influenciar os resultados. Nesse sentido, foi utilizado o teste z para analisar, cuja hipótese nula é que os erros são maiores quando nenhum filtro é considerado. Os resultados estão resumidos na Tabela 3.2, em que z está relacionado ao teste z utilizado para análise estatística dos resultados, apresentando nenhuma relação direta com o modelo causal além da escolha de variáveis: inclusão das variáveis confundidoras na análise, já que nenhuma se trata de um colisor.

Apesar de em muitos casos poder ser observado uma grande disparidade no erro (maior erro em negros), muitos dos testes de hipótese não são indicativos de bias no YOLO. Um dos parâmetros de pré-processamento evidencia um bias durante a noite (maior erro em negros). Os filtros reduziram significativamente os erros na maioria das situações.

A primeira tabela dos testes de hipótese tem como variável de análise principal a X (raça). Dado o grafo que compõe o comportamento do problema, ela testa a hipótese que o erro de detecção nos negros é maior que o erro nos brancos dentro de subconjuntos que possam remover possíveis confundimentos entre as variáveis no caminho de raça para detecção.

Já essa segunda tabela tem como variável de análise principal a W_1: filtro. Da mesma maneira, tenta-se remover possíveis confundimentos entre essa variável e a detecção. Ela analisa se o erro nas imagens sem filtro é maior que o erro nas imagens com filtro. Com isso, a primeira coluna define qual filtro está sendo analisado. As colunas "Subconjunto" e "Raça" definem o conjunto de dados em questão. Na maioria dos casos, a aplicação do filtro é indicativo de melhoria da taxa de detecção. É possível notar que no período da manhã, nenhuma das três parametrizações do filtro demonstram indicação de melhoria na detecção de brancos. Já no período da noite, apenas a primeira parametrização não possui indicação de melhoria de detecção de negros. Essa redução é observável por meio de testes de hipótese dos horários da manhã e da noite da Tabela 3.2.

3.8 CONSIDERAÇÕES FINAIS

Neste capítulo foi abordada a metodologia e resultados desse estudo que visa entender como o pré-processamento e outros fatores (como a luminosidade e o horário do dia)

Tabela 3.2: Hipótese (YOLO): Err(Raw) > Err(Filtro)

Filtro	Subconjunto	Raça	Z	p-value	IC 95%
Filtro 1	Dados Base	Negro	3.591	0.000	(0.055, 0.145)
Filtro 1	Dados Base	Branco	3.305	0.001	(0.056, 0.164)
Filtro 2	Dados Base	Negro	5.536	0.000	(0.105, 0.191)
Filtro 2	Dados Base	Branco	3.786	0.000	(0.071, 0.177)
Filtro 3	Dados Base	Negro	4.219	0.000	(0.071, 0.161)
Filtro 3	Dados Base	Branco	2.956	0.002	(0.044, 0.154)
Filtro 1	Baixa L.	Negro	2.536	0.006	(0.028, 0.129)
Filtro 1	Baixa L.	Branco	2.082	0.000	(0.018, 0.15)
Filtro 2	Baixa L.	Negro	3.666	0.000	(0.061, 0.158)
Filtro 2	Baixa L.	Branco	2.235	0.013	(0.024, 0.155)
Filtro 3	Baixa L.	Negro	2.656	0.004	(0.031, 0.132)
Filtro 3	Baixa L.	Branco	1.637	0.051	(-0.0, 0.134)
Filtro 1	Alta L.	Negro	2.634	0.42	(0.055, 0.231)
Filtro 1	Alta L.	Branco	2.678	0.37	(0.062, 0.249)
Filtro 2	Alta L.	Negro	4.308	0.0	(0.142, 0.307)
Filtro 2	Alta L.	Branco	3.258	0.06	(0.094, 0.275)
Filtro 3	Alta L.	Negro	3.45	0.03	(0.098, 0.269)
Filtro 3	Alta L.	Branco	2.678	0.37	(0.062, 0.249)
Filtro 1	Manhã	Negro	2.642	0.41	(0.037, 0.158)
Filtro 1	Manhã	Branco	0.907	18.21	(-0.034, 0.118)
Filtro 2	Manhã	Negro	3.508	0.02	(0.068, 0.184)
Filtro 2	Manhã	Branco	1.389	8.24	(-0.011, 0.137)
Filtro 3	Manhã	Negro	2.642	0.41	(0.037, 0.158)
Filtro 3	Manhã	Branco	0.907	18.21	(-0.034, 0.118)
Filtro 1	Tarde	Negro	2.288	1.11	(0.028, 0.169)
Filtro 1	Tarde	Branco	3.1	0.1	(0.078, 0.247)
Filtro 2	Tarde	Negro	3.689	0.01	(0.086, 0.22)
Filtro 2	Tarde	Branco	3.291	0.05	(0.088, 0.254)
Filtro 3	Tarde	Negro	2.783	0.27	(0.049, 0.187)
Filtro 3	Tarde	Branco	2.553	0.53	(0.05, 0.224)
Filtro 1	Noite	Negro	0.991	16.09	(-0.087, 0.36)
Filtro 1	Noite	Branco	2.345	0.95	(0.094, 0.452)
Filtro 2	Noite	Negro	2.615	0.45	(0.134, 0.502)
Filtro 2	Noite	Branco	2.345	0.95	(0.094, 0.452)
Filtro 3	Noite	Negro	2.158	1.55	(0.076, 0.469)
Filtro 3	Noite	Branco	2.345	0.95	(0.094, 0.452)

influenciam a detecção facial, com o objetivo de mitigar possíveis vieses raciais. Foram realizados testes com diferentes configurações de pré-processamento, utilizando um con-

Tabela 3.3: Hipótese (YOLO): Err(Black) > Err(White)

Filtro	Subconjunto	Z	p-value	IC 95%
Raw	Dados Base	0.597	27.51	(-0.035, 0.075)
Filtro 1	Dados Base	1.079	14.03	(-0.015, 0.075)
Filtro 2	Dados Base	-0.144	55.73	(-0.045, 0.038)
Filtro 3	Dados Base	0.124	45.05	(-0.041, 0.048)
Raw	Baixa L.	1.414	7.86	(-0.013, 0.186)
Filtro 1	Baixa L.	1.918	2.76	(0.018, 0.181)
Filtro 2	Baixa L.	1.037	14.98	(-0.025, 0.119)
Filtro 3	Baixa L.	1.185	11.8	(-0.021, 0.138)
Raw	Alta L.	-0.25	59.88	(-0.074, 0.054)
Filtro 1	Alta L.	-0.137	55.46	(-0.057, 0.048)
Filtro 2	Alta L.	-0.993	83.97	(-0.08, 0.021)
Filtro 3	Alta L.	-0.754	77.47	(-0.079, 0.03)
Raw	Manhã	0.404	34.29	(-0.055, 0.091)
Filtro 1	Manhã	-0.989	83.87	(-0.101, 0.026)
Filtro 2	Manhã	-1.271	89.82	(-0.104, 0.015)
Filtro 3	Manhã	-0.989	83.87	(-0.101, 0.026)
Raw	Tarde	0.348	36.39	(-0.068, 0.105)
Filtro 1	Tarde	1.879	3.01	(0.015, 0.15)
Filtro 2	Tarde	0.927	17.7	(-0.026, 0.1)
Filtro 3	Tarde	0.851	19.74	(-0.033, 0.107)
Raw	Noite	0.318	37.54	(-0.189, 0.28)
Filtro 1	Noite	1.757	3.95	(0.018, 0.346)
Filtro 2	Noite	0.0	50.0	(-0.103, 0.103)
Filtro 3	Noite	0.597	27.52	(-0.079, 0.17)

junto de dados de imagens rotuladas manualmente, para analisar o impacto na taxa de erro de detecção facial. Os resultados sugerem que certos filtros de pré-processamento podem melhorar a precisão do sistema, especialmente para negros, destacando a importância de considerar as variáveis de confusão e mediadores na análise. No próximo capítulo, serão discutidas as implicações desses resultados para o projeto de sistemas de reconhecimento facial imparciais e como eles podem ser aplicados para melhorar a equidade em tecnologias de identificação facial.

Capítulo 4

CONCLUSÕES

Este trabalho apresenta uma investigação sobre vieses raciais e o impacto de variáveis confundidoras e mediadoras na detecção facial, através de um estudo da relação causal entre as mesmas. Com o objetivo de realizar este estudo, foi criado um cenário metodológico utilizando uma aplicação do mundo real focada na detecção de rostos em imagens coletadas no transporte público de Salvador, Brasil. Embora três detectores faciais clássicos tenham sido utilizados nos experimentos, o objetivo principal não é dedicado à análise de vieses presentes em modelos específicos de Inteligência Artificial. Nesta dissertação, discute-se uma investigação que utiliza uma abordagem metodológica com inferência causal para analisar preconceitos raciais em diferentes cenários. A principal expectativa é mostrar como os sistemas baseados em IA podem se beneficiar da causalidade para implantar modelos justos.

Dessa forma, a motivação para esta pesquisa surge da necessidade de compreender e mitigar possíveis vieses raciais em sistemas de detecção facial, uma preocupação crescente dada a diversidade da população brasileira, particularmente a predominância de indivíduos negros em Salvador. O estudo emprega um quadro metodológico fundamentado em um modelo causal que explora a relação entre a cor da pele (variável protegida X), a presença de variáveis confundidoras como horário e luminosidade (Z), e o papel dos mediadores representados pelas imagens das faces e diferentes filtros aplicados (W). Este enquadramento permite uma análise detalhada de como essas variáveis interagem e influenciam a taxa de erro de detecção (Y), oferecendo uma compreensão sobre a presença e magnitude de vieses raciais no sistema.

Os resultados obtidos destacam a complexidade da detecção facial em ambientes reais e variados, evidenciando a influência significativa da luminosidade e do horário nas taxas de erro, especialmente quando não são utilizados filtros. A aplicação de filtros, como uma variável mediadora, mostrou-se uma estratégia eficaz para reduzir esses erros, embora seus efeitos variem de acordo com a luminosidade, o horário do dia entre outros fatores. Dessa forma, evidência-se a importância de considerar a interação entre as características da imagem, ambiente e variáveis protegidas no desenvolvimento de sistemas de detecção facial imparciais.

A importância deste estudo transcende a análise técnica, destacando a necessidade crítica de incorporar considerações éticas e de justiça na implementação de tecnologias de inteligência artificial em contextos sociais. Ao demonstrar a aplicabilidade de modelos causais para identificar e quantificar vieses, este trabalho fornece uma base metodológica valiosa para futuras pesquisas focadas na promoção da equidade em sistemas baseados em IA. Em termos de contribuições práticas, este estudo oferece evidências que podem informar o desenvolvimento e a implementação de diretrizes e práticas para a criação de sistemas de detecção facial mais justos e imparciais. Ao identificar os fatores que contribuem para os vieses de detecção e explorar estratégias de mitigação eficazes, como a aplicação de filtros específicos, os desenvolvedores de sistemas podem aprimorar a precisão e a equidade de seus algoritmos.

Além disso, as derivações deste estudo têm implicações significativas para a governança e regulamentação de tecnologias de IA. Ao iluminar as complexidades e desafios associados à garantia de sistemas imparciais, este trabalho reforça a necessidade de uma abordagem colaborativa entre pesquisadores, desenvolvedores, reguladores e a sociedade em geral para estabelecer padrões éticos e práticas recomendadas que orientem o uso responsável da IA.

Finalmente, este estudo abre caminho para pesquisas futuras que podem expandir a análise para um conjunto de dados mais amplo, incorporar modelos de detecção facial mais recentes e explorar uma separação mais granular por níveis de luminosidade e horários específicos. Através de uma investigação mais detalhada, os pesquisadores podem continuar a aprofundar a compreensão dos vieses em sistemas de detecção facial e desenvolver soluções inovadoras para abordá-los, garantindo que a tecnologia sirva equitativamente a toda a sociedade.

REFERÊNCIAS BIBLIOGRÁFICAS

BRUN, Y.; MELIOU, A. Software fairness. In: *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. New York, NY, USA: Association for Computing Machinery, 2018. (ESEC/FSE 2018), p. 754–759. ISBN 9781450355735. Disponível em: ⟨https://doi.org/10.1145/3236024.3264838⟩.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: PMLR. *Conference on fairness, accountability and transparency*. [S.l.], 2018. p. 77–91.

CALVO-GONZALEZ, E.; DUCCINI, L. On 'black culture'and 'black bodies': State discourses, tourism and public policies in salvador da bahia, brazil. *Tourism, power and culture: Anthropological insights*, Channel View Publications Toronto, p. 134–152, 2010.

CASTELVECCHI, D. Mathematicians urge colleagues to boycott police work in wake of killings. *Nature*, v. 582, p. 465, 2020.

CHEN, W. et al. Yolo-face: a real-time face detector. *The Visual Computer*, Springer, v. 37, p. 805–813, 2021.

CUMMISKEY, K. et al. Causal inference in introductory statistics courses. *Journal of Statistics Education*, Taylor & Francis, v. 28, n. 1, p. 2–8, 2020. Disponível em: ⟨https://doi.org/10.1080/10691898.2020.1713936⟩.

DABLANDER, F. *An Introduction to Causal Inference*. PsyArXiv, 2020. Disponível em: ⟨psyarxiv.com/b3fkw⟩.

FARINELLA, G.; DUGELAY, J.-L. Demographic classification: Do gender and ethnicity affect each other? In: IEEE. *2012 International Conference on Informatics, Electronics & Vision (ICIEV)*. [S.l.], 2012. p. 383–390.

FERREIRA, M. V. et al. Ethics of ai: Do the face detection models act with prejudice? In: BRITTO, A.; DELGADO, K. V. (Ed.). *Intelligent Systems*. Cham: Springer International Publishing, 2021. p. 89–103. ISBN 978-3-030-91699-2.

FORNEY, A.; MUELLER, S. Causal inference in ai education: A primer. In: . [S.l.: s.n.], 2021.

FURL, N.; PHILLIPS, P.; O'TOOLE, A. J. Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis. *Cognitive*

Science, v. 26, n. 6, p. 797–815, 2002. ISSN 0364-0213. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0364021302000848⟩.

GOULD, R. J. *Graph Theory*. Reprint. [S.l.]: Dover Publications, 2012. (Dover Books on Mathematics). ISBN 0486498069,9780486498065.

HERNÁN, M. A. A definition of causal effect for epidemiological research. *Journal of Epidemiology & Community Health*, BMJ Publishing Group Ltd, v. 58, n. 4, p. 265–271, 2004. ISSN 0143-005X. Disponível em: ⟨https://jech.bmj.com/content/58/4/265⟩.

HUANG, Y.; VALTORTA, M. *Pearl's Calculus of Intervention Is Complete*. 2012.

IMBENS, G. W.; RUBIN, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. USA: Cambridge University Press, 2015. ISBN 0521885884.

JAIN, V.; LEARNED-MILLER, E. *Fddb: A benchmark for face detection in unconstrained settings*. [S.l.], 2010.

KAMISHIMA, T. et al. Fairness-aware classifier with prejudice remover regularizer. In: FLACH, P. A.; BIE, T. D.; CRISTIANINI, N. (Ed.). *Machine Learning and Knowledge Discovery in Databases*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 35–50. ISBN 978-3-642-33486-3.

KARKKAINEN, K.; JOO, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. [S.l.: s.n.], 2021. p. 1548–1558.

KING, D. E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, v. 10, n. 60, p. 1755–1758, 2009. Disponível em: ⟨http://jmlr.org/papers/v10/king09a.html⟩.

KLARE, B. F. et al. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, v. 7, n. 6, p. 1789–1801, 2012.

KUMAR, A.; KAUR, A.; KUMAR, M. Face detection techniques: a review. *Artificial Intelligence Review*, Springer, v. 52, p. 927–948, 2019.

KUSNER, M. J. et al. Counterfactual fairness. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. v. 30. Disponível em: ⟨https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf⟩.

LEE, S.; BAREINBOIM, E. Causal identification with matrix equations. *Advances in Neural Information Processing Systems*, v. 34, p. 9468–9479, 2021.

LOFTUS, J. R. et al. *Causal Reasoning for Algorithmic Fairness*. arXiv, 2018. Disponível em: ⟨https://arxiv.org/abs/1805.05859⟩.

LUNTER, J. Beating the bias in facial recognition technology. *Biometric Technology Today*, v. 2020, n. 9, p. 5–7, 2020. Disponível em: ⟨https://doi.org/10.1016/S0969-4765(20)30122-3⟩.

LüBKE, K. et al. Why we should teach causal inference: Examples in linear regression with simulated data. *Journal of Statistics Education*, Taylor & Francis, v. 28, n. 2, p. 133–139, 2020. Disponível em: ⟨https://doi.org/10.1080/10691898.2020.1752859⟩.

MAO, C. et al. *Causal Transportability for Visual Recognition*. 2022.

MISRA, D.; GAJ, K. Face recognition captchas. In: *Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services (AICT-ICIW'06)*. [S.l.: s.n.], 2006. p. 122–122.

NG, H.-W.; WINKLER, S. A data-driven approach to cleaning large face datasets. In: IEEE. *2014 IEEE international conference on image processing (ICIP)*. [S.l.], 2014. p. 343–347.

OHLSSON, H.; KENDLER, K. S. Applying Causal Inference Methods in Psychiatric Epidemiology: A Review. *JAMA Psychiatry*, v. 77, n. 6, p. 637–644, 06 2020. ISSN 2168-622X. Disponível em: ⟨https://doi.org/10.1001/jamapsychiatry.2019.3758⟩.

Deep face recognition. [S.l.]: British Machine Vision Association, 2015.

PEARL, J. *Causality*: Models, reasoning, and inference. 2. ed. Cambridge, UK: Cambridge University Press, 2009. ISBN 978-0-521-89560-6.

PEARL, J. Comment: Understanding simpson's paradox. *The American Statistician*, Taylor & Francis, v. 68, n. 1, p. 8–13, 2014. Disponível em: ⟨https://doi.org/10.1080/00031305.2014.876829⟩.

PEARL, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 62, n. 3, p. 54–60, fev. 2019. ISSN 0001-0782. Disponível em: ⟨https://doi.org/10.1145/3241036⟩.

PETERS, J.; JANZING, D.; SCHLKOPF, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. [S.l.]: The MIT Press, 2017. ISBN 0262037319.

PETERSEN, M. L. Compound treatments, transportability, and the structural causal model: the power and simplicity of causal graphs. *Epidemiology*, LWW, v. 22, n. 3, p. 378–381, 2011.

PHILLIPS, P. J. et al. An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception (TAP)*, ACM New York, NY, USA, v. 8, n. 2, p. 1–11, 2011.

PLECKO, D.; BAREINBOIM, E. *Causal Fairness Analysis*. arXiv, 2022. Disponível em: ⟨https://arxiv.org/abs/2207.11385⟩.

PLECKO, D.; BAREINBOIM, E. *Reconciling Predictive and Statistical Parity: A Causal Approach*. 2023.

PLEISS, G. et al. On fairness and calibration. *Advances in neural information processing systems*, v. 30, 2017.

RAITA, Y. et al. Big data, data science, and causal inference: A primer for clinicians. *Frontiers in Medicine*, v. 8, p. 998, 2021. ISSN 2296-858X. Disponível em: ⟨https://www.frontiersin.org/article/10.3389/fmed.2021.678047⟩.

RAJI, I. D. et al. Saving face: Investigating the ethical concerns of facial recognition auditing. *URL https://doi.org/10.1145/3375627.3375820*, 2020.

SHARMA, M. et al. Facial detection using deep learning. In: IOP PUBLISHING. *IOP Conference Series: Materials Science and Engineering*. [S.l.], 2017. v. 263, n. 4, p. 042092.

SPIRTES, P. Introduction to causal inference. *Journal of Machine Learning Research*, v. 11, n. 5, 2010.

VANDERWEELE, T. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, 2015. ISBN 9780199325870. Disponível em: ⟨https://books.google.com.br/books?id=Ob_coQEACAAJ⟩.

VANDERWEELE, T. J.; SHPITSER, I. On the definition of a confounder. *Annals of statistics*, NIH Public Access, v. 41, n. 1, p. 196, 2013.

WANG, M. et al. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. [S.l.: s.n.], 2019. p. 692–702.

YANG, S. et al. Wider face: A face detection benchmark. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016.

ZHANG, K. et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, IEEE, v. 23, n. 10, p. 1499–1503, 2016.

ZHAO, W. et al. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 35, n. 4, p. 399–458, 2003.