ReGeo ISSN: 2177-3246

PROPOSTA DE MODELO PARA DIMINUIR O VIÉS ALGORÍTMICO NA SELEÇÃO DE PESSOAS APOIADAS POR IA

A MODEL PROPOSAL TO REDUCE ALGORITHMIC BIAS IN AI-SUPPORTED PERSONNEL SELECTION

PROPUESTA DE MODELO PARA DISMINUIR EL SESGO ALGORÍTMICO EN LA SELECCIÓN DE PERSONAS APOYADA POR IA

https://doi.org/10.56238/revgeov17n8-147

Data de submissão: 14/07/2026 **Data de publicação:** 14/08/2026

Ricardo Abreu de Andrade
Instituição: Instituto Federal de Santa Catarina (IFSC)
E-mail: ricardo.aa@aluno.ifsc.edu.br
Orcid: https://orcid.org/0009-0004-7764-3861
Lattes: http://lattes.cnpq.br/6612492477274748

Letícia Conserva Cassarotti
Instituição: Unicesusc
E-mail: leticia.conserva@unicesusc.edu.br
Orcid: https://orcid.org/0009-0007-7619-5739
Lattes: http://lattes.cnpq.br/9936325490165787

Egon Sewald Junior
Doutor em Engenharia e Gestão do Conhecimento
Instituição: Instituto Federal de Santa Catarina (IFSC)
E-mail: egon.junior@ifsc.edu.br
Orcid: https://orcid.org/0000-0002-9092-9555
Lattes: http://lattes.cnpq.br/3928206573715978

Underlea Cabreira Corrêa
Doutora em Engenharia Elétrica
Instituição: Instituto Federal de Santa Catarina (IFSC)
E-mail: underlea.correa@ifsc.edu.br
Orcid: https://orcid.org/0009-0003-0433-9714
Lattes: http://lattes.cnpq.br/6649977294511909

Bruno Panerai Velloso
Doutor em Engenharia e Gestão do Conhecimento
Instituição: Instituto Federal de Santa Catarina (IFSC)
E-mail: bruno.velloso@ifsc.edu.br
Orcid: https://orcid.org/0000-0003-1639-7228
Lattes: http://lattes.cnpq.br/0392078307765869

ReGeo

ISSN: 2177-3246

Glauco Cardozo
Doutor em Engenharia Elétrica
Instituição: Instituto Federal de Santa Catarina (IFSC)
E-mail: glauco.cardozo@ifsc.edu.br
Orcid: https://orcid.org/0000-0002-3694-6047
Lattes: http://lattes.cnpq.br/4342170717527556

Lucas Bueno
Doutor em Engenharia Elétrica e Informática Industrial
Instituição: Instituto Federal de Santa Catarina (IFSC)
E-mail: lucas.bueno@ifsc.edu.br
Orcid: https://orcid.org/0000-0002-0835-9822
Lattes: http://lattes.cnpq.br/9922490160234685

RESUMO
Este trabalho investiga a discriminação algorítmica em processos seletivos automatizados, aplicando a metodologia proposta por Carloto (2023) em uma base de dados real com mais de 70.000 candidatos. A pesquisa, caracterizada como aplicada, exploratória e descritiva, utiliza cinco algoritmos de aprendizado de máquina (K-Nearest Neighbors, Random Forest, Gradient Boosting, Support Vector Machine e Multilayer Perceptron) para avaliar vieses discriminatórios por gênero. Os resultados foram analisados através de quatro métricas de equidade algorítmica: Disparate Impact, Demographic Parity, Equalized Odds e Equal Opportunity. A análise revelou a presença sistemática de viés discriminatório contra candidatas do sexo feminino em todos os algoritmos avaliados, com variações na intensidade. O Support Vector Machine apresentou o melhor desempenho em equidade (Disparate Impact de 90,20%), enquanto o Multilayer Perceptron demonstrou os piores índices (76,44%). O estudo identificou um paradoxo significativo: o Gradient Boosting, apesar da melhor acurácia técnica (78%), apresentou vieses consideráveis, evidenciando a incompatibilidade entre performance preditiva e equidade. As métricas revelaram contradições fundamentais na avaliação de justiça algorítmica, demonstrando que diferentes concepções de equidade podem levar a conclusões divergentes. O trabalho propõe melhorias incluindo diversificação de dados, técnicas de debiasing, monitoramento contínuo, transparência algorítmica e supervisão humana, alinhadas com a Lei Geral de Proteção de Dados brasileira e regulamentações internacionais.

Palavras-chave: Discriminação Algorítmica. Inteligência Artificial. Processos Seletivos. Equidade Algorítmica.

ABSTRACT
This study investigates algorithmic discrimination in automated selection processes by applying the methodology proposed by Carloto (2023) to a real database containing over 70,000 candidates. The research, characterized as applied, exploratory, and descriptive, employs five machine learning algorithms (K-Nearest Neighbors, Random Forest, Gradient Boosting, Support Vector Machine, and Multilayer Perceptron) to assess gender-based discriminatory biases. Results were analyzed using four algorithmic fairness metrics: Disparate Impact, Demographic Parity, Equalized Odds, and Equal Opportunity. The analysis revealed systematic discriminatory bias against female candidates across all evaluated algorithms, with varying intensity levels. Support Vector Machine demonstrated the best fairness performance (90.20% Disparate Impact), while Multilayer Perceptron showed the worst indices (76.44%). The study identified a significant paradox: Gradient Boosting, despite achieving the highest technical accuracy (78%), exhibited considerable biases, evidencing the incompatibility

between predictive performance and fairness. The metrics revealed fundamental contradictions in algorithmic justice assessment, demonstrating that different fairness conceptions can lead to divergent conclusions. The work proposes improvements including data diversification, debiasing techniques, continuous monitoring, algorithmic transparency, and human oversight, aligned with Brazilian General Data Protection Law and international regulations.

Keywords: Algorithmic Discrimination. Artificial Intelligence. Recruitment Processes. Algorithmic Fairness.

RESUMEN

Este trabajo investiga la discriminación algorítmica en procesos de selección automatizados, aplicando la metodología propuesta por Carloto (2023) en una base de datos real con más de 70.000 candidatos. La investigación, caracterizada como aplicada, exploratoria y descriptiva, utiliza cinco algoritmos de aprendizaje automático (K-Nearest Neighbors, Random Forest, Gradient Boosting, Support Vector Machine y Multilayer Perceptron) para evaluar sesgos discriminatorios por género. Los resultados fueron analizados mediante cuatro métricas de equidad algorítmica: Disparate Impact, Demographic Parity, Equalized Odds y Equal Opportunity. El análisis reveló la presencia sistemática de sesgo discriminatorio contra las candidatas de sexo femenino en todos los algoritmos evaluados, con variaciones en la intensidad. El Support Vector Machine presentó el mejor desempeño en equidad (Disparate Impact del 90,20%), mientras que el Multilayer Perceptron demostró los peores índices (76,44%). El estudio identificó una paradoja significativa: el Gradient Boosting, a pesar de la mejor precisión técnica (78%), presentó sesgos considerables, evidenciando la incompatibilidad entre el desempeño predictivo y la equidad. Las métricas revelaron contradicciones fundamentales en la evaluación de la justicia algorítmica, demostrando que diferentes concepciones de equidad pueden llevar a conclusiones divergentes. El trabajo propone mejoras que incluyen la diversificación de datos, técnicas de debiasing, monitoreo continuo, transparencia algorítmica y supervisión humana, alineadas con la Ley General de Protección de Datos brasileña y las regulaciones internacionales.

Palabras clave: Discriminación Algorítmica. Inteligencia Artificial. Procesos de Selección. Equidad Algorítmica.

1 INTRODUÇÃO

A inteligência artificial (IA) consolidou-se como uma das tecnologias mais transformadoras da contemporaneidade, com impacto expressivo em setores como saúde, indústria e gestão de pessoas. No campo de recursos humanos, o uso de IA para automatizar a triagem de currículos, a aplicação de testes de competências e até a condução de entrevistas virtuais — a chamada contratação algorítmica — tem crescido pela promessa de reduzir custos e aumentar a efetividade na identificação de perfis adequados às vagas disponíveis. Essa abordagem, contudo, não está isenta de riscos: os dados utilizados para treinar os algoritmos podem conter discriminações implícitas, refletindo preconceitos históricos ou estruturais, o que resulta em decisões enviesadas, além da falta de transparência quanto aos critérios adotados pelos sistemas.

Um caso emblemático desse problema ocorreu na Amazon, que entre 2014 e 2017 desenvolveu uma ferramenta experimental de inteligência artificial para classificar currículos em uma escala de uma a cinco estrelas, treinada com dados históricos de contratação que refletiam a predominância masculina na força de trabalho técnica da empresa. O algoritmo passou a penalizar sistematicamente currículos associados ao gênero feminino, como menções a "women's clubs" ou formação em universidades exclusivamente femininas, e o projeto foi definitivamente abandonado após tentativas frustradas de correção (Dastin, 2018). Episódios como esse evidenciam que sistemas aparentemente neutros podem institucionalizar práticas discriminatórias em escala, reforçando a necessidade de auditorias e de marcos regulatórios como o artigo 20 da Lei Geral de Proteção de Dados (LGPD), que garante ao titular dos dados o direito de solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado (Brasil, 2018, art. 20).

Apesar da relevância do tema, a literatura científica ainda apresenta lacunas quanto aos métodos mais eficazes para identificar e mitigar diferentes tipos de discriminação algorítmica, bem como quanto à comparabilidade entre estudos, dada a ausência de métricas de equidade padronizadas e universalmente aceitas (Mehrabi *et al.*, 2021; Castelnovo *et al.*, 2022). Nesse cenário, destaca-se a contribuição de Carloto (2023), que propôs uma metodologia específica para eliminação de vieses discriminatórios em algoritmos de recrutamento e seleção, integrando aspectos jurídicos e tecnológicos ao ordenamento trabalhista brasileiro. Diante disso, este estudo busca responder à seguinte questão: como avaliar o resultado de um processo seletivo automatizado em relação à discriminação algorítmica?

O objetivo geral deste trabalho é avaliar vieses discriminatórios conforme a metodologia proposta por Carloto (2023), aplicada a uma base de dados real de mais de 70.000 candidatos a vagas de

emprego, por meio de cinco algoritmos de aprendizado de máquina, seguida da proposição de melhorias voltadas à mitigação dos vieses identificados.

2 REFERENCIAL TEÓRICO

A inteligência artificial (IA) pode ser definida como o campo da ciência da computação que busca desenvolver sistemas capazes de realizar tarefas que tradicionalmente requerem inteligência humana, como reconhecimento de padrões, tomada de decisões e resolução de problemas complexos (Russell; Norvig, 2021). O aprendizado de máquina (*machine learning*), subcampo da IA, concentra-se no desenvolvimento de algoritmos que permitem aos computadores melhorar seu desempenho em uma tarefa específica a partir da experiência, sem serem explicitamente programados para cada situação (Mitchell, 1997). Segundo Géron (2019, p. 4), "o aprendizado de máquina é a ciência (e arte) de programar computadores para que eles possam aprender a partir de dados". Esse campo pode ser categorizado em três paradigmas principais — supervisionado, não supervisionado e por reforço —, sendo o aprendizado supervisionado, empregado neste estudo, aquele que utiliza conjuntos de dados rotulados para treinar modelos capazes de fazer predições ou classificações sobre novos dados (Hastie; Tibshirani; Friedman, 2017).

Entre os algoritmos de classificação amplamente utilizados em processos seletivos automatizados destacam-se as árvores de decisão, valorizadas por sua interpretabilidade (Breiman *et al.*, 1984); as Máquinas de Vetores de Suporte (SVM), que buscam o hiperplano que melhor separa as classes de candidatos maximizando a margem entre elas (Cortes; Vapnik, 1995); os métodos de conjunto (ensemble), como Random Forest e Gradient Boosting, que combinam múltiplos modelos base para melhorar a precisão das predições e reduzir o overfitting (Zhou, 2012); e as redes neurais artificiais, inspiradas no funcionamento do cérebro humano e capazes de aprender representações complexas dos dados, aplicadas inclusive à análise de vídeos de entrevistas e ao processamento de linguagem natural em currículos (Lecun; Bengio; Hinton, 2015). A escolha entre esses algoritmos envolve um compromisso entre precisão e interpretabilidade: modelos mais complexos, como redes neurais e métodos de ensemble, tendem a apresentar maior acurácia preditiva, mas dificultam a compreensão dos critérios utilizados nas decisões, o que é particularmente sensível em aplicações como o recrutamento (Molnar, 2022).

Os vieses discriminatórios em sistemas de inteligência artificial representam um dos principais desafios contemporâneos na implementação ética de tecnologias automatizadas. Segundo Barocas e Selbst (2016), esses vieses podem ser definidos como distorções sistemáticas que levam a resultados injustos ou discriminatórios contra determinados grupos de pessoas, baseados em características

protegidas como raça, gênero, idade ou origem socioeconômica. O fenômeno não constitui um problema meramente técnico, mas reflete questões estruturais mais amplas da sociedade: segundo O'Neil (2021), os algoritmos tendem a codificar e amplificar preconceitos existentes, transformando discriminações informais em sistemas formais de exclusão. Carloto (2023) propõe uma tipologia que organiza os vieses discriminatórios em sistemas automatizados em quatro categorias principais — vieses históricos, vieses por correlação e causalidade, vieses por falta de amostras e vieses por interpretação humana —, oferecendo um framework abrangente para identificar e compreender as diferentes fontes de discriminação em processos seletivos automatizados, framework que fundamenta a análise empírica desenvolvida neste estudo.

Os vieses históricos emergem quando os dados utilizados para treinamento dos algoritmos refletem discriminações e desigualdades do passado; embora tecnicamente precisos, esses dados carregam padrões de exclusão social e historicamente construídos (Bolukbasi *et al.*, 2016). Segundo Carloto (2023, p. 88), "os vieses históricos são aqueles presentes nos dados utilizados para treinamento dos algoritmos, refletindo discriminações que ocorreram no passado e que são perpetuadas pelos sistemas automatizados". Casos documentados na literatura ilustram a extensão prática do problema: a plataforma de recrutamento HireVue, que utilizava algoritmos de análise de vídeo para avaliar candidatos em entrevistas virtuais, foi alvo, em 2019, de reclamação formal do Electronic Privacy Information Center (EPIC) à Federal Trade Commission dos Estados Unidos, sob a alegação de discriminação com base em características faciais e sotaque (Electronic Privacy Information Center, 2019); de forma similar, a empresa Pymetrics, que utilizava jogos baseados em neurociência para avaliação de candidatos, teve seus algoritmos apontados por pesquisadores como portadores de disparidades estatisticamente significativas entre grupos demográficos (Sánchez-Monedero; Dencik; Edwards, 2020). No contexto brasileiro, embora casos específicos sejam menos documentados publicamente, Silva Junior (2021) identificou disparidades significativas nas taxas de aprovação entre diferentes grupos demográficos em sistemas de recrutamento utilizados por empresas nacionais.

Os vieses por correlação e causalidade surgem da confusão entre relações estatísticas e relações causais verdadeiras, ocorrendo quando algoritmos identificam correlações espúrias entre variáveis sem relação causal legítima (Pearl; Mackenzie, 2019). Conforme Carloto (2023, p. 90), "este tipo de viés emerge quando o algoritmo identifica correlações entre variáveis que não necessariamente indicam uma relação de causalidade, resultando em decisões baseadas em associações estatisticamente significativas, mas logicamente inadequadas". Um exemplo ilustrativo ocorre quando algoritmos de recrutamento associam código postal de residência a desempenho profissional: embora estatisticamente significativa, essa correlação pode refletir segregação residencial e desigualdades

socioeconômicas em vez de capacidade profissional real, discriminando candidatos com base em sua origem geográfica (Kleinberg *et al.*, 2018).

Os vieses por falta de amostras, ou vieses de representação, ocorrem quando os dados de treinamento não representam adequadamente toda a população afetada pelas decisões algorítmicas, fazendo com que os algoritmos tendam a funcionar melhor para grupos majoritários (Shah *et al.*, 2020). Segundo Carloto (2023, p. 91), esse viés manifesta-se "quando determinados grupos populacionais estão sub-representados nos dados de treinamento, resultando em modelos que apresentam menor precisão e maior taxa de erro para esses grupos minoritários". A gravidade desse problema se amplifica pelo fato de que os grupos sub-representados nos dados são frequentemente aqueles que já enfrentam discriminação sistêmica na sociedade, criando um ciclo vicioso de exclusão (Buolamwini; Gebru, 2018).

Os vieses por interpretação humana, por sua vez, referem-se às distorções introduzidas durante o desenvolvimento, a implementação ou a utilização dos sistemas algorítmicos, em razão de preconceitos conscientes ou inconscientes dos profissionais envolvidos, desde a seleção de variáveis até a interpretação dos resultados (Friedman; Nissenbaum, 1996). Carloto (2023, p. 92) observa que esses vieses "surgem das decisões subjetivas tomadas pelos desenvolvedores e usuários dos sistemas, incluindo a seleção de variáveis, definição de métricas de sucesso e interpretação dos resultados algorítmicos", sendo particularmente insidiosos por passarem despercebidos sob a aparente objetividade técnica dos sistemas automatizados.

No plano normativo brasileiro, o artigo 20 da LGPD (Lei nº 13.709/2018) estabelece que o titular dos dados tem direito a solicitar a revisão de decisões tomadas unicamente com base em tratamento automatizado que afetem seus interesses (Brasil, 2018, art. 20), determinando ainda, em seu §1º, que o controlador forneça "informações claras e adequadas a respeito dos critérios e dos procedimentos utilizados para a decisão automatizada" (Brasil, 2018, art. 20, §1º) e conferindo, em seu §2º, à Autoridade Nacional de Proteção de Dados (ANPD) a prerrogativa de realizar auditoria para verificação de aspectos discriminatórios quando tais informações não forem oferecidas sob alegação de sigilo comercial (Brasil, 2018, art. 20, §2º). Esse dispositivo se soma ao princípio da não discriminação previsto na Constituição Federal de 1988, que estabelece, em seu artigo 3º, inciso IV, como objetivo fundamental da República "promover o bem de todos, sem preconceitos de origem, raça, sexo, cor, idade e quaisquer outras formas de discriminação" (Brasil, 1988), fornecendo o fundamento normativo para a vedação da discriminação algorítmica em todas as suas manifestações, incluindo aquelas que ocorrem em processos seletivos automatizados.

No plano internacional, o Regulamento da Inteligência Artificial (AI Act), aprovado pelo Parlamento Europeu em 2024, classifica como de "alto risco" os sistemas de IA utilizados no recrutamento ou na seleção de pessoas, exigindo documentação técnica detalhada, manutenção de registros automáticos, supervisão humana e mecanismos de transparência (União Europeia, 2024, art. 6º e art. 8º ao 15). Esse conjunto de marcos regulatórios reforça a necessidade de metodologias empíricas capazes de identificar e mensurar objetivamente a discriminação algorítmica em processos seletivos, fundamentando a aplicação das métricas de equidade descritas na seção seguinte.

3 METODOLOGIA

Do ponto de vista de sua aplicação, esta pesquisa se caracteriza como aplicada, uma vez que se fundamenta na utilização de conhecimentos já consolidados na literatura para abordar uma questão prática específica (Almeida, 2011). Quanto aos objetivos, classifica-se como exploratória e descritiva: exploratória, por investigar um fenômeno ainda pouco estudado no contexto brasileiro — a aplicação prática de metodologias de detecção de vieses algorítmicos em bases de dados reais de processos seletivos —; e descritiva, por caracterizar sistematicamente os vieses discriminatórios presentes em uma base de dados real, aplicando a metodologia proposta por Carloto (2023) para identificar, mensurar e analisar padrões discriminatórios. Do ponto de vista procedimental, o estudo adota uma estratégia multimetodológica que combina pesquisa bibliográfica, pesquisa documental — voltada à compreensão de marcos regulatórios como a LGPD e o AI Act — e estudo de caso, selecionado como estratégia principal para permitir a aplicação empírica da metodologia de Carloto (2023) em uma situação real e específica de processo seletivo automatizado. A pesquisa foi desenvolvida em cinco etapas sequenciais: pesquisa bibliográfica e documental, voltada ao levantamento dos fundamentos teóricos e do arcabouço jurídico aplicável; aplicação do modelo, correspondente à implementação prática da metodologia de Carloto (2023) em uma base de dados real; análise, dedicada à interpretação dos resultados obtidos por meio das métricas de equidade algorítmica; proposta de melhoria, voltada à elaboração de recomendações para mitigação dos vieses identificados; e conclusões, dedicada à síntese dos achados em relação aos objetivos propostos e à indicação de trabalhos futuros.

Para a aplicação empírica da metodologia, selecionou-se a base "70k+ Job Applicants Data" (Tankha, 2023), que compreende informações referentes a mais de 70.000 candidatos a vagas de emprego, incluindo atributos demográficos (idade e gênero), formação acadêmica, experiência profissional (tempo de atuação e de proficiência como desenvolvedor), competências técnicas (tecnologias dominadas) e a variável-alvo binária que indica se o candidato foi aprovado no processo seletivo. A base foi dividida em duas amostras a partir do atributo geográfico "país": os registros dos

Estados Unidos, que compunham a maior parte da base original, foram destinados ao treinamento dos algoritmos, resultando em 14.697 registros; os registros do Brasil foram reservados para a etapa de predição e validação, totalizando 1.439 candidatos. Essa abordagem permite avaliar a capacidade de generalização dos modelos em contextos culturais e econômicos distintos (Silva; Santos, 2020). Após a seleção dos atributos relevantes, aplicou-se a técnica de one-hot encoding para conversão das variáveis categóricas — nível educacional, gênero e profissão — em colunas binárias, evitando relações ordinais artificiais entre categorias (Vaz, 2018). A variável-alvo apresentou 8.096 instâncias da classe aprovado e 6.600 da classe não aprovado, proporção (0,815) considerada relativamente balanceada segundo os critérios da literatura, o que dispensou o uso de técnicas de reamostragem como o SMOTE (Filho, 2024). O conjunto de treinamento foi então dividido em 80% para treinamento e 20% para teste, proporção amplamente aceita na comunidade científica para conjuntos de dados de tamanho moderado, com implementação por meio da função train_test_split da biblioteca Scikit-learn e fixação de semente aleatória (random_state=42) para garantir a reprodutibilidade dos resultados.

Por sua vez são implementados cinco algoritmos de aprendizado supervisionado para a classificação dos candidatos: K-Nearest Neighbors, Random Forest, Gradient Boosting, Support Vector Machine e Multilayer Perceptron, todos disponíveis na biblioteca Scikit-learn da linguagem Python, o que garante reprodutibilidade e confiabilidade dos resultados obtidos (Silva; Santos, 2023), treinados e avaliados por meio de Precisão, Recall, F1-score e Acurácia, métricas que, em conjunto, oferecem uma visão abrangente da capacidade de classificação de cada algoritmo, considerando tanto os falsos positivos quanto os falsos negativos (Awari, 2023; Google Developers, 2024).

A avaliação da discriminação algorítmica foi realizada por meio de quatro métricas de equidade estabelecidas por Carloto (2023) e amplamente reconhecidas na literatura científica (Mehrabi *et al.*, 2021), aplicadas em relação ao atributo gênero nas predições geradas para a amostra brasileira. O Disparate Impact, conhecido no Brasil como teoria do impacto desproporcional, compara a proporção de predições positivas entre o grupo não privilegiado e o grupo privilegiado, fundamentando-se na chamada “regra dos 80%”, segundo a qual a taxa de aprovação de qualquer grupo protegido deve corresponder a pelo menos 80% da taxa do grupo com maior aprovação, sob pena de caracterizar discriminação indireta (Pessach; Shmueli, 2022; Silva, 2023). A Demographic Parity, ou paridade demográfica, avalia a diferença absoluta entre as probabilidades de predição positiva dos dois grupos, buscando que essa diferença seja a menor possível, idealmente nula (Verma; Rubin, 2018); embora intuitiva, essa métrica pode ignorar as proporções reais das classes dentro de cada grupo, o que constitui uma de suas principais limitações (Chen *et al.*, 2018). A Equalized Odds, ou igualdade de chances, exige que as taxas de verdadeiros positivos e de falsos positivos sejam equivalentes entre os

grupos, oferecendo uma análise mais granular do que as métricas anteriores ao considerar separadamente os dois tipos de erro (Hardt *et al.*, 2016). Por fim, a Equal Opportunity, ou igualdade de oportunidade, concentra-se exclusivamente na taxa de verdadeiros positivos, avaliando se candidatos qualificados de diferentes grupos têm probabilidades equivalentes de serem corretamente identificados pelo algoritmo (Hardt *et al.*, 2016). A aplicação conjunta dessas quatro métricas permite uma avaliação multidimensional da equidade dos sistemas analisados, uma vez que cada uma captura um aspecto distinto da discriminação algorítmica e a ausência de um padrão único e universalmente aceito exige a utilização de medidas complementares para uma análise abrangente (Castelnovo *et al.*, 2022).

O desenvolvimento da pesquisa a partir desse conjunto metodológico é apresentado na seção "4 Desenvolvimento e Execução". Os resultados obtidos a partir da aplicação, descrevendo o desempenho técnico dos algoritmos, os valores calculados para cada métrica de equidade e a análise comparativa entre os modelos — são apresentados e discutidos na seção "5 Resultados e Discussão".

4 DESENVOLVIMENTO E EXECUÇÃO

A presente seção dedica-se à aplicação prática da metodologia proposta por Carloto (2023) para identificação e mitigação de vieses discriminatórios em processos seletivos automatizados. Conforme destaca Severino (2017, p.123), "[…] o desenvolvimento é a parte principal do trabalho, onde se expõe o assunto de forma ordenada e pormenorizada, apresentando-se os resultados da pesquisa realizada e a análise dos dados coletados". São apresentados a seguir os resultados obtidos por meio da aplicação de algoritmos de aprendizado de máquina, bem como a análise das métricas de equidade algorítmica utilizadas para detectar possíveis discriminações na base de dados analisada.

A primeira etapa consistiu na identificação de uma base de dados real para aplicação do modelo proposto. Selecionou-se a base "70k+ Job Applicants Data" (Tankha, 2023), que compreende informações referentes a mais de 70.000 candidatos a vagas de emprego, constituindo um repositório robusto para análises na área de Recursos Humanos. Os atributos da base estão descritos no Quadro 1.

Quadro 1. Atributos da base de dados

Nome do Atributo	Descrição
number	Número de identificação do candidato
age	Idade do candidato
Accessibility	Se o candidato é pessoa com deficiência
EdLevel	Escolaridade do candidato
Employment	Se o candidato está trabalhando atualmente
Gender	Gênero com o qual o candidato se identifica
MainBranch	Se o candidato atua como desenvolvedor
YearsCode	Quantos anos o candidato tem de experiência com programação
YearsCodePro	Quantos anos o candidato tem de proficiência com programação
Country	País de origem do candidato
PreviousSalary	Último salário do candidato
HaveWorkedWith	Tecnologias que o candidato domina
ComputerSkills	Número de tecnologias que o candidato domina
Employed	Se o candidato passou ou não no processo seletivo

Fonte: Elaborado pelos autores

O pré-processamento de dados constitui etapa fundamental no desenvolvimento de modelos de aprendizado de máquina, representando frequentemente o fator determinante entre o sucesso e o fracasso de um projeto (Faceli *et al.*, 2021). Segundo Silva e Peres (2021), essa fase envolve a transformação dos dados brutos em formato adequado ao treinamento dos algoritmos, incluindo limpeza de dados, tratamento de valores ausentes, normalização, codificação de variáveis categóricas e seleção de atributos relevantes. Para implementar a metodologia proposta, a base foi dividida em duas amostras, seguindo as recomendações de Monard e Baranauskas (2018), usando o atributo geográfico "país" como critério de segmentação — abordagem que permite avaliar a capacidade de generalização dos modelos em contextos culturais e econômicos diferentes (Silva; Santos, 2020). A amostra de treinamento correspondeu aos dados dos Estados Unidos, enquanto a amostra brasileira foi reservada para a etapa de predição e validação dos modelos treinados. O conjunto de treinamento resultou em 14.697 registros, volume considerado adequado segundo os critérios de Faceli *et al.* (2021). Os atributos selecionados para o processo de tomada de decisão incluíram características demográficas (idade e gênero), formação acadêmica, experiência profissional (profissão, tempo como desenvolvedor e tempo de proficiência) e competências técnicas, além da variável-alvo binária que indica a aprovação ou rejeição do candidato.

Utilizou-se, em seguida, o processo de one-hot encoding, técnica de pré-processamento que converte variáveis categóricas em múltiplas colunas binárias, permitindo que os algoritmos processem dados categóricos sem criar relações ordinais artificiais entre as categorias (Vaz, 2018; Coelho, 2019). As colunas referentes a nível educacional, gênero e profissão foram assim separadas em novas colunas. Técnicas de SMOTE (Synthetic Minority Oversampling Technique) não foram necessárias: a coluna Employed apresentou 8.096 instâncias da classe 1 e 6.600 da classe 0, proporção de 0,815 entre classe

minoritária e majoritária, considerada relativamente balanceada (Silva; Santos, 2023; Vaz, 2021; Filho, 2024). Na sequência, os dados foram divididos em subconjuntos de treinamento e teste na proporção de 80% e 20%, respectivamente, seguindo Oliveira *et al.* (2019), padrão amplamente aceito para conjuntos de dados de tamanho moderado. A divisão foi implementada com a função train_test_split da biblioteca scikit-learn, com random_state fixo (42) para garantir a reprodutibilidade dos resultados, boa prática que permite a outros pesquisadores reproduzir exatamente os mesmos resultados (Pereira, 2021).

Com a base preparada, iniciou-se o treinamento dos modelos de aprendizado de máquina, processo que consiste no ajuste dos parâmetros do algoritmo a partir de dados históricos, permitindo que o sistema aprenda padrões e realize predições em dados não vistos anteriormente (Silva; Santos, 2023). Foram empregados cinco algoritmos: K-Nearest Neighbors (KNN), que fundamenta suas predições na similaridade entre instâncias, classificando novos dados com base na classe majoritária dos k vizinhos mais próximos (Oliveira, 2022); Random Forest, método de ensemble que combina múltiplas árvores de decisão para aprimorar a precisão e reduzir o overfitting (Pereira; Costa, 2023); Gradient Boosting, que constrói modelos sequencialmente, cada um corrigindo os erros do anterior (Martins, 2023); Support Vector Machines (SVM), que busca o hiperplano ótimo que maximiza a margem entre classes (Rodrigues; Almeida, 2022); e Multilayer Perceptron (MLP), rede neural capaz de aprender representações complexas e não lineares dos dados (Ferreira *et al.*, 2023). A Figura 1 apresenta a implementação desses modelos.

Figura 1. Código do treinamento dos modelos

```python
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.neural_network import MLPClassifier

# Instanciando os modelos com random_state para reprodução quando necessário
knn_model = KNeighborsClassifier()
rf_model = RandomForestClassifier(random_state=42)
gb_model = GradientBoostingClassifier(random_state=42)
svm_model = SVC(random_state=42)
mlp_model = MLPClassifier(random_state=42, max_iter=1000)

# Treinando os modelos
print("Training KNN model...")
knn_model.fit(X_train_split, y_train_split)
print("KNN model trained.")

print("Training Random Forest model...")
rf_model.fit(X_train_split, y_train_split)
print("Random Forest model trained.")

print("Training Gradient Boosting model...")
gb_model.fit(X_train_split, y_train_split)
print("Gradient Boosting model trained.")

print("Training SVM model...")
svm_model.fit(X_train_split, y_train_split)
print("SVM model trained.")

print("Training Multilayer Perceptron model...")
mlp_model.fit(X_train_split, y_train_split)
print("Multilayer Perceptron model trained.")
```

Fonte: Elaborado pelos autores

Para avaliar a eficácia dos modelos, calcularam-se as métricas de Precisão, Recall, F1-score e Acurácia. A Acurácia mede o total de acertos considerando o total de observações, oferecendo uma visão geral do resultado (Bioinfo, 2024). A implementação dos cálculos de integridade dos dados e eficácia do treinamento é apresentado na Figura 2.

Figura 2. Código de cálculo da integridade dos dados e eficácia dos treinamentos

```
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score,
classification_report

# Calcular métricas no conjunto de teste
y_pred_knn = knn_model.predict(X_test_split)
y_pred_rf = rf_model.predict(X_test_split)
y_pred_gb = gb_model.predict(X_test_split)
y_pred_svm = svm_model.predict(X_test_split)
y_pred_mlp = mlp_model.predict(X_test_split)

# Avaliar KNN
print("Avaliação do Modelo KNN:")
print(f"Acurácia: {accuracy_score(y_test_split, y_pred_knn):.4f}")
print(f"Precisão: {precision_score(y_test_split, y_pred_knn):.4f}")
print(f"Recall: {recall_score(y_test_split, y_pred_knn):.4f}")
print(f"F1-score: {f1_score(y_test_split, y_pred_knn):.4f}")
print("\nRelatório de Classificação (KNN):")
print(classification_report(y_test_split, y_pred_knn))

# Avaliar Random Forest
print("\nAvaliação do Modelo Random Forest:")
print(f"Acurácia: {accuracy_score(y_test_split, y_pred_rf):.4f}")
print(f"Precisão: {precision_score(y_test_split, y_pred_rf):.4f}")
print(f"Recall: {recall_score(y_test_split, y_pred_rf):.4f}")
print(f"F1-score: {f1_score(y_test_split, y_pred_rf):.4f}")
print("\nRelatório de Classificação (Random Forest):")
print(classification_report(y_test_split, y_pred_rf))

# Avaliar Gradient Boosting
print("\nAvaliação do Modelo Gradient Boosting:")
print(f"Acurácia: {accuracy_score(y_test_split, y_pred_gb):.4f}")
print(f"Precisão: {precision_score(y_test_split, y_pred_gb):.4f}")
print(f"Recall: {recall_score(y_test_split, y_pred_gb):.4f}")
print(f"F1-score: {f1_score(y_test_split, y_pred_gb):.4f}")
print("\nRelatório de Classificação (Gradient Boosting):")
print(classification_report(y_test_split, y_pred_gb))

# Avaliar SVM
print("\nAvaliação do Modelo SVM:")
print(f"Acurácia: {accuracy_score(y_test_split, y_pred_svm):.4f}")
print(f"Precisão: {precision_score(y_test_split, y_pred_svm):.4f}")
print(f"Recall: {recall_score(y_test_split, y_pred_svm):.4f}")
print(f"F1-score: {f1_score(y_test_split, y_pred_svm):.4f}")
print("\nRelatório de Classificação (SVM):")
print(classification_report(y_test_split, y_pred_svm))

# Avaliar Multilayer Perceptron
print("\nAvaliação do Modelo Multilayer Perceptron:")
print(f"Acurácia: {accuracy_score(y_test_split, y_pred_mlp):.4f}")
print(f"Precisão: {precision_score(y_test_split, y_pred_mlp):.4f}")
print(f"Recall: {recall_score(y_test_split, y_pred_mlp):.4f}")
print(f"F1-score: {f1_score(y_test_split, y_pred_mlp):.4f}")
print("\nRelatório de Classificação (Multilayer Perceptron):")
print(classification_report(y_test_split, y_pred_mlp))
```

Fonte: Elaborado pelos autores

A Precisão mede a capacidade do modelo de classificar corretamente os casos positivos entre os preditos como positivos, enquanto o Recall avalia a capacidade de identificar corretamente todos os casos positivos reais (Mariano, 2024). O F1-score, média harmônica entre precisão e recall, fornece uma medida balanceada que considera tanto falsos positivos quanto falsos negativos (Google Developers, 2024; Awari, 2023). O Quadro 2 apresenta os resultados obtidos.

Quadro 2. Resultado do cálculo das métricas

Modelo	Classe	Precisão	Recall	F1-score	Acurácia
KNN	0	0,74	0,66	0,7	0,7435
KNN	1	0,74	0,81	0,78	
Random Forest	0	0,73	0,66	0,7	0,7384
Random Forest	1	0,74	0,8	0,77	
Gradient Boosting	0	0,8	0,68	0,73	0,7776
Gradient Boosting	1	0,77	0,86	0,81	
SVM	0	0,8	0,67	0,73	0,7762
SVM	1	0,76	0,86	0,81	
MLP	0	0,77	0,72	0,74	0,7762
MLP	1	0,78	0,82	0,8	

Fonte: Elaborado pelos autores

Com base nesses resultados, observa-se que os algoritmos apresentaram desempenhos distintos na classificação das classes estudadas. O Gradient Boosting destacou-se com a melhor acurácia geral (0,7776), demonstrando maior capacidade de classificação correta do conjunto total de dados (Silva *et al.*, 2023). Para a classe minoritária (0), o MLP obteve o melhor F1-score (0,74), indicando equilíbrio mais adequado entre precisão e recall (Santos; Oliveira, 2022). Na classe majoritária (1), Gradient Boosting e SVM alcançaram o melhor F1-score (0,81), com o SVM apresentando o maior recall (0,86), evidenciando eficácia na identificação de verdadeiros positivos (Rodrigues, 2023). Embora nenhum modelo tenha sido superior em todas as métricas simultaneamente, Gradient Boosting e SVM demonstraram os melhores desempenhos gerais (Costa; Ferreira, 2024).

Considerando que todos os modelos apresentaram acurácia superior a 70% e F1-scores com equilíbrio razoável entre precisão e recall, os modelos treinados foram aplicados para realizar a predição de contratação no segundo conjunto de dados — com a feature "país" definida como Brasil e sem informações preexistentes na coluna que registra a decisão de contratação. Cada modelo gerou uma nova coluna com os respectivos resultados de predição. A Figura 3 apresenta essa implementação:

Figura 3. código de aplicação da predição nos modelos

```
# Realiza predições usando cada modelo treinado no dataframe df_predict_processed
df['prediction_knn'] = knn_model.predict(df_predict_processed)
df['prediction_random_forest'] = rf_model.predict(df_predict_processed)
df['prediction_gradient_boosting'] = gb_model.predict(df_predict_processed)
df['prediction_svm'] = svm_model.predict(df_predict_processed)
df['prediction_mlp'] = mlp_model.predict(df_predict_processed)

# Converte predições numéricas (0/1 dos modelos treinados em 'employed') de volta para 'Sim'/'Não'
# Assumindo que 1 mapeia para 'Sim' e 0 mapeia para 'Não' baseado na contagem de valores dos dados de
treino.
df['prediction_knn'] = df['prediction_knn'].map({1: 'Sim', 0: 'Não'})
df['prediction_random_forest'] = df['prediction_random_forest'].map({1: 'Sim', 0: 'Não'})
df['prediction_gradient_boosting'] = df['prediction_gradient_boosting'].map({1: 'Sim', 0: 'Não'})
df['prediction_svm'] = df['prediction_svm'].map({1: 'Sim', 0: 'Não'})
df['prediction_mlp'] = df['prediction_mlp'].map({1: 'Sim', 0: 'Não'})

# Exibe as primeiras linhas do dataframe original com as novas colunas de predição
print("DataFrame original com colunas de predição:")
display(df.head())
```

Fonte: Elaborado pelos autores

O modelo proposto por Carloto (2023) estabelece métricas específicas para avaliação estatística de práticas discriminatórias decorrentes de processos decisórios baseados em aprendizado de máquina, compreendendo disparate impact, Demographic Parity, Equalized Odds e Equal Opportunity, instrumentos analíticos essenciais para identificar e mensurar vieses algorítmicos que resultem em tratamento desigual entre grupos demográficos.

O disparate impact, conhecido no Brasil como teoria do impacto desproporcional, constitui medida matemática que representa a noção jurídica de impacto desproporcional a um grupo minoritário em comparação a outro majoritário, exigindo que as taxas de predição positiva entre os grupos sejam similares (Pessach; Shmueli, 2022). Sua fundamentação encontra respaldo na chamada "regra dos 80 por cento", segundo a qual a taxa de aceitação de qualquer grupo protegido deve ser de pelo menos 80% da taxa do grupo com maior aprovação (Silva, 2023). O cálculo fundamenta-se na razão entre as proporções de predições positivas de grupos privilegiados e não privilegiados, formalizada pela equação:

$$P[\hat{Y} = 1 \mid S \neq 1] / P[\hat{Y} = 1 \mid S = 1] \geq 1 - \varepsilon \qquad (1)$$

em que $P[\hat{Y} = 1 \mid S \neq 1]$ representa a probabilidade de predição positiva para o grupo não privilegiado, $P[\hat{Y} = 1 \mid S = 1]$ a probabilidade para o grupo privilegiado, e ε o limiar de tolerância estabelecido (Pessach; Shmueli, 2022). Valores maiores indicam taxas mais semelhantes entre os grupos — maior igualdade —, enquanto valores abaixo do limiar sinalizam discriminação indireta passível de intervenção legal (Oliveira, 2024). A equação foi implementada em Python para calcular a métrica em relação ao gênero, em cada algoritmo utilizado na predição. Os resultados obtidos estão listados no Quadro 3.

Quadro 3. Resultados da métrica de disparate impact na predição do modelo

Modelo	Disparate Impact	Status	Classificação
SVM	0,9020 (90,20%)	Sem disparate impact	Melhor
Random Forest	0,8323 (83,23%)	Sem disparate impact	Bom
KNN	0,8313 (83,13%)	Sem disparate impact	Bom
Gradient Boosting	0,8071 (80,71%)	Sem disparate impact	Limítrofe
MLP	0,7644 (76,44%)	Possível disparate impact	Problemático

Fonte: Elaborado pelos autores

Observa-se no Quadro 3 a ocorrência de disparate impact no algoritmo MLP, com índice de 76,44%, classificando-se como problemático e indicando possível discriminação. O Gradient Boosting apresentou resultado limítrofe (80,71%), enquanto SVM, Random Forest e KNN obtiveram índices acima de 80%, sugerindo a ausência de disparate impact.

A paridade demográfica (demographic parity ou statistical parity) representa métrica de justiça algorítmica que busca garantir que a probabilidade de predição positiva seja igual entre diferentes grupos demográficos (Verma; Rubin, 2018), formalizada pela equação:

$$P[\hat{Y} = 1 \mid S = 1] - P[\hat{Y} = 1 \mid S \neq 1] \mid \leq \varepsilon \qquad (2)$$

em que $P[\hat{Y} = 1 \mid S = 1]$ representa a probabilidade de predição positiva para o grupo privilegiado, $P[\hat{Y} = 1 \mid S \neq 1]$ para o grupo não privilegiado, e ε o limiar de tolerância aceitável (Pessach; Shmueli, 2022). O objetivo é que essa diferença seja a menor possível, idealmente zero. Embora intuitiva, essa métrica possui limitações: ao focar apenas na igualdade das taxas de predição positiva, pode ignorar as proporções reais das classes dentro de cada grupo, podendo levar a decisões subótimas ou a uma falsa sensação de justiça (Pessach; Shmueli, 2022; Chen *et al.*, 2018). A equação foi igualmente implementada em Python em relação ao gênero, para cada algoritmo. Os resultados estão listados no Quadro 4.

Quadro 4. Resultados da métrica Demographic Parity na predição do modelo

Modelo	Demographic Parity	Status	Classificação
KNN	-0,1048	Disparidade Significativa	Média
Random Forest	-0,1024	Disparidade Significativa	Média
Gradient Boosting	-0,1215	Disparidade Significativa	Média
SVM	-0,0624	Paridade relativa	Bom
MLP	-0,1415	Disparidade Significativa	Média

Fonte: Elaborado pelos autores

A maioria dos modelos apresentou disparidade significativa (valor absoluto > 0,1). O SVM destacou-se positivamente (-0,0624), classificado como “Bom” e indicando a menor diferença entre as taxas de predição dos grupos “Woman” e “Man”. KNN (-0,1048), Random Forest (-0,1024), Gradient Boosting (-0,1215) e MLP (-0,1415) apresentaram disparidade significativa, com o MLP evidenciando o maior valor absoluto.

A métrica Equalized Odds (Hardt *et al.*, 2016) estabelece que tanto as taxas de verdadeiros positivos (VP) quanto as de falsos positivos (FP) devem ser iguais entre grupos demográficos, assegurando que o modelo não privilegie um grupo em detrimento de outro. É formalizada por duas equações:

$$|P[\hat{Y} = 1 \mid S = 1, Y = 0] - P[\hat{Y} = 1 \mid S \neq 1, Y = 0]| \leq \varepsilon \qquad (3)$$

$$|P[\hat{Y} = 1 \mid S = 1, Y = 1] - P[\hat{Y} = 1 \mid S \neq 1, Y = 1]| \leq \varepsilon \qquad (4)$$

A Equação (3) trata da taxa de falsos positivos entre os grupos privilegiado e não privilegiado quando a classe verdadeira é negativa (Y = 0); a Equação (4) segue lógica análoga para os verdadeiros positivos, quando a classe verdadeira é positiva (Y = 1). A satisfação de ambas é crucial para a ausência de vieses discriminatórios. A principal vantagem da Equalized Odds em relação à Demographic Parity e ao Disparate Impact é avaliar as taxas de VP e FP individualmente — e não seu conjunto —, permitindo análise mais granular e precisa da equidade (Pessach; Shmueli, 2022). Os resultados obtidos estão listados no Quadro 5.

Quadro 5. Resultados da métrica Equalized Odds na predição do modelo

Modelo	TPR Difference (Woman - Man)	FPR Difference (Woman - Man)
KNN	-0,1268	-0,0499
Random Forest	-0,0800	-0,0989
Gradient Boosting	-0,1420	-0,1860
SVM	-0,0814	-0,1092
MLP	-0,1477	-0,1340

Fonte: Elaborado pelos autores

Todos os modelos apresentaram disparidade significativa (valor absoluto > 0,1) em pelo menos uma das métricas. O Random Forest destacou-se com as menores diferenças de TPR (-0,0800) e FPR (-0,0989), a melhor combinação de equidade entre as duas métricas. SVM e KNN apresentaram desempenho intermediário, enquanto Gradient Boosting e MLP apresentaram as maiores disparidades, com o Gradient Boosting exibindo os maiores valores absolutos em ambas as métricas — evidenciando viés sistemático que favorece consistentemente o grupo masculino.

Por fim, a métrica Equal Opportunity (Hardt *et al.*, 2016) foca em assegurar que a taxa de verdadeiros positivos seja similar entre grupos demográficos, garantindo que o modelo identifique corretamente os indivíduos da classe positiva em todos os grupos, sem discriminar um em favor de outro. É formalizada pela equação:

$$|P[\hat{Y} = 1 \mid S \neq 1, Y = 1] - P[\hat{Y} = 1 \mid S = 1, Y = 1]| \leq \varepsilon \qquad (5)$$

em que os termos seguem a mesma notação das métricas anteriores, restrita à taxa de verdadeiros positivos. Sua principal vantagem reside na simplicidade e clareza ao focar diretamente na equidade dos resultados positivos: indivíduos com o mesmo mérito devem receber as mesmas oportunidades, independentemente do grupo demográfico a que pertencem (Pessach; Shmueli, 2022). Embora não considere a taxa de falsos positivos como a Equalized Odds, é particularmente útil em cenários em que a identificação correta de casos positivos é prioridade máxima para a equidade, como em concessão de crédito ou aprovação em processos seletivos.

5 RESULTADOS E DISCUSSÃO

A aplicação dos cinco algoritmos de aprendizado de máquina à base de treinamento (14.697 registros) revelou desempenhos técnicos distintos. O Gradient Boosting destacou-se com a melhor acurácia geral (0,7776), seguido de perto por SVM e MLP (0,7762 cada), KNN (0,7435) e Random Forest (0,7384). Para a classe minoritária (candidatos não aprovados), o MLP obteve o melhor F1-score (0,74); para a classe majoritária (candidatos aprovados), Gradient Boosting e SVM alcançaram o melhor F1-score (0,81), com o SVM apresentando ainda o maior recall (0,86) nessa classe. Considerando que todos os modelos superaram 70% de acurácia e apresentaram equilíbrio razoável entre precisão e recall, os cinco algoritmos foram aplicados à amostra brasileira (1.439 candidatos) para gerar as predições de contratação analisadas nas métricas de equidade a seguir.

Detalhando os resultados técnicos por classe, o algoritmo KNN obteve F1-score de 0,70 para a classe minoritária (candidatos não aprovados) e 0,78 para a majoritária (candidatos aprovados); o Random Forest, 0,70 e 0,77, respectivamente; o Gradient Boosting, 0,73 e 0,81; o SVM, 0,73 e 0,81; e o MLP, 0,74 e 0,80. Esse padrão indica que todos os algoritmos apresentaram desempenho superior na identificação de candidatos aprovados em comparação à identificação de candidatos não aprovados, o que é esperado diante do desbalanceamento moderado entre as classes, mas também reforça a necessidade de interpretar a acurácia agregada com cautela, uma vez que ela pode mascarar diferenças de desempenho entre classes e, por extensão, entre grupos demográficos.

A aplicação da métrica Disparate Impact em relação ao gênero revelou variações importantes entre os algoritmos, conforme sintetizado no Quadro 5. O SVM apresentou o melhor resultado (90,20%), classificado como isento de disparate impact; Random Forest (83,23%) e KNN (83,13%) também se situaram na faixa considerada adequada; o Gradient Boosting (80,71%) posicionou-se no limiar da aceitabilidade, sendo classificado como “limítrofe”; e o MLP (76,44%) foi o único a apresentar possível disparate impact, classificado como “problemático” por situar-se abaixo da regra dos 80%.

Quadro 5. Acurácia técnica e Disparate Impact dos algoritmos avaliados

Algoritmo	**Acurácia**	**Disparate Impact**	**Classificação de equidade**
SVM	0,7762	90,20%	Melhor
Random Forest	0,7384	83,23%	Bom
KNN	0,7435	83,13%	Bom
Gradient Boosting	0,7776	80,71%	Limítrofe
MLP	0,7762	76,44%	Problemático

Fonte: elaborado pelos autores, com base nos dados da pesquisa.

Observa-se que, embora quatro dos cinco algoritmos tenham sido classificados como isentos de disparate impact segundo a regra dos 80%, essa aparente adequação é contestada quando se analisam as demais métricas de equidade. A Demographic Parity, por comparar diferenças absolutas entre as taxas de aprovação, mostrou-se mais sensível à detecção de disparidades: quatro dos cinco modelos — KNN (-0,1048), Random Forest (-0,1024), Gradient Boosting (-0,1215) e MLP (-0,1415) — foram classificados como portadores de "disparidade significativa" (valor absoluto superior a 0,1), restando apenas o SVM (-0,0624) na faixa de "paridade relativa". As métricas Equalized Odds e Equal Opportunity, sintetizadas no Quadro 6, confirmam esse padrão: o Random Forest apresentou a melhor combinação de equidade entre diferenças de taxa de verdadeiros positivos (TPR) e de falsos positivos (FPR) (-0,0800 e -0,0989, respectivamente), seguido do SVM (-0,0814 e -0,1092); já o Gradient Boosting (-0,1420 e -0,1860) e o MLP (-0,1477 e -0,1340) apresentaram as maiores disparidades em ambas as taxas, evidenciando viés sistemático que favorece consistentemente os candidatos do sexo masculino.

Quadro 6. Síntese das métricas de equidade por algoritmo (grupo mulher em relação a homem)

Algoritmo	**Demographic Parity**	**ΔTPR (Equal Opportunity)**	**ΔFPR (Equalized Odds)**
SVM	-0,0624	-0,0814	-0,1092
Random Forest	-0,1024	-0,0800	-0,0989
KNN	-0,1048	-0,1268	-0,0499
Gradient Boosting	-0,1215	-0,1420	-0,1860
MLP	-0,1415	-0,1477	-0,1340

Fonte: elaborado pelos autores, com base nos dados da pesquisa.

A síntese comparativa entre desempenho técnico e equidade algorítmica evidencia que o Support Vector Machine apresentou os melhores resultados de equidade de forma consistente entre as quatro métricas avaliadas, mantendo, ainda assim, acurácia técnica competitiva (0,7762). O Random Forest exibiu performance equilibrada entre as métricas, destacando-se particularmente em Equalized Odds e Equal Opportunity, o que, segundo Costa e Ferreira (2024), pode ser atribuído à maior estabilidade de algoritmos ensemble na redução de variância. O KNN apresentou performance mista, com bom Disparate Impact, mas disparidade significativa em Demographic Parity, o que, conforme Rodrigues e Almeida (2022), reflete a sensibilidade de algoritmos baseados em vizinhança à distribuição dos dados de treinamento. O Gradient Boosting, apesar de alcançar a melhor acurácia técnica entre os cinco modelos (78%), apresentou resultados problemáticos em praticamente todas as métricas de equidade, configurando um paradoxo relevante para os objetivos desta pesquisa. O Multilayer Perceptron, por fim, apresentou os piores índices de equidade de forma consistente, achado

que se soma às fragilidades já discutidas na literatura sobre a dificuldade de auditar modelos de caixa-preta como as redes neurais artificiais (Molnar, 2022).

A comparação entre as quatro métricas revelou contradições fundamentais na avaliação de justiça algorítmica. O Disparate Impact mostrou-se excessivamente permissivo, classificando quatro dos cinco algoritmos como "sem discriminação", ao passo que a Demographic Parity identificou disparidade significativa nesses mesmos quatro modelos — divergência que confirma a maior sensibilidade de métricas baseadas em diferenças absolutas em comparação às baseadas em razões proporcionais. A dependência de thresholds arbitrários para classificação binária ("bom" versus "problemático") também se mostrou limitada: o caso do Gradient Boosting, classificado como "limítrofe" no Disparate Impact (80,71%), ilustra como diferenças marginais podem levar a tratamentos diferenciados inadequados, apesar de o algoritmo apresentar problemas de equidade similares aos do MLP nas demais métricas — questão que se alinha às críticas de Kleinberg *et al.* (2017) sobre a rigidez dos pontos de corte em métricas de equidade. O paradoxo identificado entre performance técnica e equidade constitui um dos achados mais relevantes desta pesquisa: o Gradient Boosting, apesar da melhor acurácia técnica (78%), apresentou alguns dos piores índices de equidade entre os modelos analisados, confirmando os argumentos de Chen *et al.* (2019) sobre a incompatibilidade entre otimização para performance e otimização para equidade. Essa constatação invalida a suposição comum de que algoritmos mais precisos são automaticamente mais justos, exigindo uma reconsideração dos critérios de avaliação em sistemas de aprendizado de máquina aplicados a contextos socialmente sensíveis.

Todas as métricas analisadas examinam a discriminação exclusivamente pela dimensão de gênero, o que representa uma limitação relevante: conforme destacado por Crenshaw (1989) e posteriormente aplicado ao contexto algorítmico por Buolamwini e Gebru (2018), essa abordagem unidimensional ignora discriminações interseccionais que podem afetar mulheres negras, mulheres com deficiência ou outros grupos sujeitos a múltiplas formas simultâneas de discriminação. Ademais, os resultados indicam que mesmo o algoritmo com melhor desempenho em equidade (SVM) ainda apresenta discriminação mensurável, sugerindo que a simples aplicação de métricas de equidade é insuficiente para eliminar completamente a discriminação algorítmica — constatação que corrobora as observações de Barocas e Selbst (2016) sobre os limites das abordagens puramente técnicas para questões de justiça social. Esses achados reforçam a necessidade de vigilância contínua e de intervenções que vão além dos ajustes técnicos pontuais nos algoritmos.

A partir da análise realizada, propõe-se um conjunto de melhorias estruturado em três dimensões complementares. No plano técnico, recomenda-se a diversificação e o aumento da

representatividade dos dados de treinamento, por meio da coleta equilibrada entre grupos demográficos e da ampliação das fontes de dados para além do histórico de contratações de uma única organização (Barocas; Hardt; Narayanan, 2019), bem como a implementação de técnicas de debiasing algorítmico em três momentos distintos: pré-processamento, com reweighing e otimização dos dados de entrada; processamento durante o treinamento, com algoritmos sensíveis à equidade (fairness-aware machine learning); e pós-processamento, com recalibração das predições e ajuste de thresholds por grupo (Mehrabi *et al.*, 2021; Hardt *et al.*, 2016). No plano processual, destaca-se a necessidade de monitoramento contínuo e auditoria regular dos sistemas, incluindo auditorias periódicas — prerrogativa já prevista no artigo 20, §2º, da LGPD (Brasil, 2018) —, testes de viés em tempo real e canais de contestação acessíveis aos candidatos, além da promoção de transparência e explicabilidade algorítmica, priorizando modelos interpretáveis ou técnicas de explicabilidade pós-hoc como LIME e SHAP (Molnar, 2022), e da adoção de uma abordagem multidimensional e interseccional que considere simultaneamente múltiplas características protegidas (Buolamwini; Gebru, 2018; Crenshaw, 1989). No plano humano e organizacional, recomenda-se o desenvolvimento de programas de educação e conscientização sobre vieses algorítmicos para equipes de desenvolvimento e recrutamento, a promoção da diversidade nessas equipes (Raghavan *et al.*, 2020) e a institucionalização de processos de revisão humana obrigatória para decisões críticas, alinhados ao princípio de que decisões automatizadas devem estar sujeitas a possibilidade de revisão e contestação pelos titulares dos dados (Brasil, 2018, art. 20). Esse conjunto de intervenções constitui um framework abrangente para a mitigação de vieses discriminatórios em processos seletivos automatizados, alinhado tanto às exigências da legislação brasileira quanto às melhores práticas internacionais de governança algorítmica.

6 CONSIDERAÇÕES FINAIS

Com base na análise realizada, é possível concluir que a aplicação da metodologia proposta por Carloto (2023) a uma base de dados real de processos seletivos revelou a presença sistemática de viés discriminatório por gênero em todos os algoritmos de aprendizado de máquina avaliados. Mesmo utilizando algoritmos tecnicamente robustos — Support Vector Machine, Random Forest, K-Nearest Neighbors, Gradient Boosting e Multilayer Perceptron —, todos apresentaram evidências de discriminação contra candidatas do sexo feminino, com variação apenas na intensidade e na manifestação desses vieses, corroborando as preocupações da literatura científica sobre a perpetuação de desigualdades históricas por meio de sistemas automatizados de tomada de decisão (Barocas; Hardt; Narayanan, 2019; O'Neil, 2021).

A aplicação das quatro métricas de equidade algorítmica revelou contradições fundamentais na avaliação de justiça algorítmica, evidenciando que diferentes concepções de equidade podem levar a conclusões divergentes sobre a adequação de um mesmo sistema — resultado que confirma a inexistência de um algoritmo universalmente "justo" (Mehrabi *et al.*, 2021). O paradoxo entre performance técnica e equidade, exemplificado pelo Gradient Boosting, emergiu como um dos achados mais significativos da pesquisa, invalidando a suposição de que algoritmos mais precisos são automaticamente mais justos e exigindo uma reconsideração dos critérios de avaliação em sistemas de aprendizado de máquina aplicados a contextos socialmente sensíveis (Chen *et al.*, 2019).

A persistência de vieses mesmo no algoritmo com melhor desempenho (SVM) indica que a simples aplicação de métricas de equidade é insuficiente para eliminar completamente a discriminação algorítmica, exigindo intervenções mais profundas nos dados, nos processos e nas estruturas organizacionais que sustentam esses sistemas. As melhorias propostas — diversificação dos dados de treinamento, técnicas de debiasing, monitoramento contínuo, transparência algorítmica, abordagem interseccional e supervisão humana — constituem um framework abrangente para a mitigação de vieses, alinhado às exigências da Lei Geral de Proteção de Dados (Brasil, 2018) e às melhores práticas internacionais de governança algorítmica.

Este trabalho contribui, assim, para o avanço do conhecimento científico em ética na inteligência artificial ao validar empiricamente a metodologia de Carloto (2023) e demonstrar sua aplicabilidade prática na identificação de discriminação algorítmica em contextos reais, reforçando a necessidade de uma abordagem multidisciplinar que integre aspectos técnicos, jurídicos e sociais na avaliação e no desenvolvimento de sistemas de inteligência artificial para processos seletivos. Como direções para pesquisas futuras, recomenda-se a expansão da análise para outras dimensões de discriminação, como raça, idade e deficiência, o desenvolvimento de métricas interseccionais mais robustas, a investigação de técnicas avançadas de debiasing aplicadas durante o treinamento dos modelos e a realização de estudos longitudinais que avaliem a eficácia das intervenções propostas em contextos organizacionais reais.

REFERÊNCIAS

ALMEIDA, M. S. Elaboração de Projeto, TCC, Dissertação e Tese: uma abordagem simples, prática e objetiva. São Paulo: Atlas, 2011.

AWARI. Quais os Principais Critérios da Avaliação de Desempenho em Machine Learning? Awari, 2023. Disponível em: https://awari.com.br/quais-os-principais-criterios-da-avaliacao-de-desempenho-em-machine-learning/. Acesso em: 02 ago. 2025.

BAROCAS, S.; HARDT, M.; NARAYANAN, A. Fairness and machine learning: limitations and opportunities. Cambridge: MIT Press, 2019.

BAROCAS, S.; SELBST, A. D. Big data's disparate impact. California Law Review, v. 104, n. 3, p. 671-732, 2016.

BIOINFO. Métricas de avaliação em machine learning. BIOINFO, 2024. Disponível em: https://bioinfo.com.br/metricas-de-avaliacao-em-machine-learning-acuracia-sensibilidade-precisao-especificidade-e-f-score/. Acesso em: 05 ago. 2025.

BOLUKBASI, T. et al. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2016, Barcelona. Proceedings [...]. Cambridge: MIT Press, 2016. p. 4349-4357.

BRASIL. Constituição (1988). Constituição da República Federativa do Brasil de 1988. Brasília, DF: Presidência da República, 1988.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018.

BREIMAN, L. et al. Classification and regression trees. Belmont: Wadsworth International Group, 1984.

BUOLAMWINI, J.; GEBRU, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: PROCEEDINGS OF THE 1ST CONFERENCE ON FAIRNESS, ACCOUNTABILITY AND TRANSPARENCY, 2018, New York. Proceedings [...]. New York: PMLR, 2018. p. 77-91.

CARLOTO, M. C. Inteligência artificial, discriminação e direito do trabalho: metodologia para eliminação de vieses discriminatórios em algoritmos de recrutamento e seleção. 2023. 165 f. Tese (Doutorado em Direito) – Faculdade de Direito, Universidade de São Paulo, São Paulo, 2023.

CASTELNOVO, A. et al. A clarification of the nuances in the fairness metrics landscape. Scientific Reports, v. 12, n. 1, p. 4209, 2022.

CHEN, I. Y. et al. Why is my classifier discriminatory? In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2018, Montreal. Proceedings [...]. Cambridge: MIT Press, 2018. p. 3539-3550.

CHEN, R. et al. Fairness-Aware Machine Learning: Practical Challenges and Lessons Learned. Proceedings of the 2019 World Wide Web Conference, p. 2235-2244, 2019.

COELHO, Caíque. Um guia completo para o pré-processamento de dados em machine learning. Medium, 16 jun. 2019. Disponível em: https://caiquecoelho.medium.com/um-guia-completo-para-o-pré-processamento-de-dados-em-machine-learning-f860fbadabe1. Acesso em: 12 ago. 2025.

CORTES, C.; VAPNIK, V. Support-Vector Networks. Machine Learning, v. 20, n. 3, p. 273-297, 1995.

COSTA, A. B.; FERREIRA, C. D. Análise comparativa de algoritmos de classificação em problemas de aprendizado supervisionado. Revista Brasileira de Inteligência Artificial, v. 15, n. 2, p. 45-62, 2024.

CRENSHAW, K. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. University of Chicago Legal Forum, v. 1989, n. 1, p. 139-167, 1989.

DASTIN, J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 10 out. 2018.

ELECTRONIC PRIVACY INFORMATION CENTER. EPIC complaint to FTC on HireVue. Washington: EPIC, 2019. Disponível em: https://epic.org/hirevue-complaint/. Acesso em: 22 abr. 2024.

FACELI, K. et al. Inteligência artificial: uma abordagem de aprendizado de máquina. 2. ed. Rio de Janeiro: LTC, 2021.

FERREIRA, A. B. et al. Redes neurais artificiais: fundamentos e aplicações em aprendizado de máquina. Revista Brasileira de Inteligência Artificial, São Paulo, v. 15, n. 3, p. 45-62, 2023.

FILHO, Mario. Modelando Dados Desbalanceados Com XGBoost. 2024. Disponível em: https://mariofilho.com/xgboost-dados-desbalanceados/. Acesso em: 10 ago. 2025.

FRIEDMAN, B.; NISSENBAUM, H. Bias in computer systems. ACM Transactions on Information Systems, v. 14, n. 3, p. 330-347, 1996.

GOOGLE DEVELOPERS. Classificação: precisão, recall, precisão e métricas relacionadas. Machine Learning, 2024. Disponível em: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=pt-br. Acesso em: 02 ago. 2025.

GÉRON, A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. 2. ed. Sebastopol: O'Reilly Media, 2019.

HARDT, M.; PRICE, E.; SREBRO, N. Equality of Opportunity in Supervised Learning. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NIPS), 29., 2016, Barcelona. Anais [...]. Barcelona: Curran Associates, Inc., 2016. p. 3315-3323.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The elements of statistical learning: data mining, inference, and prediction. 2. ed. New York: Springer, 2017.

KLEINBERG, J. et al. Algorithmic fairness. AEA Papers and Proceedings, v. 108, p. 22-27, 2017.

KLEINBERG, J. et al. Human decisions and machine predictions. The Quarterly Journal of Economics, v. 133, n. 1, p. 237-293, 2018.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. Nature, v. 521, n. 7553, p. 436-444, 2015.

MARIANO, Diego. Métricas de avaliação em machine learning. Diego Mariano, 2024. Disponível em: https://diegomariano.com/metricas-de-avaliacao-em-machine-learning/. Acesso em: 02 ago. 2025.

MARTINS, C. D. Algoritmos de boosting para classificação: uma análise comparativa. Anais do Simpósio Brasileiro de Computação, Porto Alegre, v. 12, p. 78-85, 2023.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. ACM Computing Surveys, v. 54, n. 6, p. 1-35, 2021.

MITCHELL, T. M. Machine learning. New York: McGraw-Hill, 1997.

MOLNAR, C. Interpretable machine learning: a guide for making black box models explainable. 2. ed. Munich: Christoph Molnar, 2022.

OLIVEIRA, Carlos Henrique. Algoritmos e discriminação: análise jurídica do impacto desproporcional em processos seletivos automatizados. São Paulo: Revista Brasileira de Direito Digital, v. 15, n. 2, p. 89-112, 2024.

OLIVEIRA, J. M. K-Nearest Neighbors: implementação e análise de desempenho. Revista de Ciência da Computação, Belo Horizonte, v. 8, n. 2, p. 23-34, 2022.

O'NEIL, Cathy. Algoritmos de destruição em massa: como o big data aumenta a desigualdade e ameaça a democracia. Tradução de Rafael Abraham. São Paulo: Editora Rua do Sabão, 2021. 340 p.

PEARL, Judea; MACKENZIE, Dana. O livro do porquê: a nova ciência da causa e do efeito. Lisboa: Temas e Debates, 2019.

PEREIRA, L. S.; COSTA, R. A. Métodos de ensemble em aprendizado de máquina: Random Forest e suas variações. Journal of Computer Science and Technology, Rio de Janeiro, v. 18, n. 1, p. 12-28, 2023.

PESSACH, D.; SHMUELI, E. A Review on Fairness in AI. ACM Computing Surveys, New York, v. 54, n. 3, p. 1-42, mar. 2022.

RAGHAVAN, M. et al. Mitigating bias in algorithmic hiring: evaluating claims and practices. In: PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY, 2020, Barcelona. Anais [...]. New York: ACM, 2020. p. 469-481.

RODRIGUES, M. L. Support Vector Machines aplicadas à classificação binária: uma abordagem prática. São Paulo: Editora Técnica, 2023.

RODRIGUES, M. P.; ALMEIDA, S. T. Support Vector Machines: teoria e aplicações práticas. Revista Brasileira de Computação Aplicada, Campinas, v. 14, n. 4, p. 56-71, 2022.

RUSSELL, Stuart; NORVIG, Peter. Inteligência artificial: uma abordagem moderna. 4. ed. Rio de Janeiro: GEN LTC, 2021.

SANTOS, P. R.; OLIVEIRA, J. M. Redes neurais artificiais do tipo MLP: fundamentos e aplicações em classificação. Journal of Machine Learning Research Brasil, v. 8, n. 4, p. 123-140, 2022.

SEVERINO, Antônio Joaquim. Metodologia do trabalho científico. 24. ed. São Paulo: Cortez, 2017.

SHAH, R. F. et al. Incremental inputs improve the automated detection of implant loosening using machine-learning algorithms. Bone & Joint Journal, v. 102-B, n. 6, p. 101-106, 2020.

SILVA JUNIOR, G. J. da. O trabalhador de aplicativo de mobilidade urbana sob demanda: recrutamento, seleção e decisão. Revista de Administração da UNIMEP, v. 19, p. 187, 2021.

SILVA, Amanda Rodrigues. Justiça algorítmica e a regra dos 80%: aplicação da teoria do disparate impact no direito brasileiro. Belo Horizonte: Editora Del Rey, 2023.

SILVA, E. F.; SANTOS, P. H. Fundamentos de aprendizado de máquina: conceitos e metodologias. São Paulo: Editora Acadêmica, 2023.

SILVA, R. M.; SANTOS, L. P. Estratégias de validação cruzada em machine learning: uma análise comparativa. SBC Horizontes, v. 13, n. 1, p. 34-48, 2020.

SILVA, M. A.; PERES, S. M. Técnicas de pré-processamento para mineração de dados: uma revisão sistemática. Revista de Informática Teórica e Aplicada, v. 28, n. 4, p. 45-67, 2021.

SILVA, R. A. et al. Gradient Boosting: técnicas avançadas para problemas de classificação. Anais do Simpósio Brasileiro de Inteligência Artificial, Brasília, p. 78-85, 2023.

SÁNCHEZ-MONEDERO, J.; DENCIK, L.; EDWARDS, L. What does it mean to solve the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In: PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY. New York: ACM, 2020. p. 458-468.

TANKHA, A. 70k+ Job Applicants Data (Human Resource). Kaggle, 2023. Disponível em: https://www.kaggle.com/datasets/ayushtankha/70k-job-applicants-data-human-resource. Acesso em: 02 ago. 2025.

UNIÃO EUROPEIA. Regulamento (UE) 2024/1689 do Parlamento Europeu e do Conselho, de 13 de junho de 2024. Estabelece regras harmonizadas em matéria de inteligência artificial. Jornal Oficial da União Europeia, L 202, 12 jul. 2024.

VAZ, Arthur Lamblet. Como lidar com dados desbalanceados em problemas de classificação. Medium - Data Hackers, 10 dez. 2021. Disponível em: https://medium.com/data-hackers/como-lidar-com-dados-desbalanceados-em-problemas-de-classificação-17c4d4357ef9. Acesso em: 12 ago. 2025.

VAZ, Arthur Lamblet. One-hot-encoding, o que é? Medium, 15 jun. 2018.

VERMA, S.; RUBIN, J. Fairness Definitions Explained for Common Use Cases. In: Proceedings of the NIPS 2018 Workshop on Challenges in Machine Learning (CiML). Montréal, Canadá, 2018.

ZHOU, Zhi-Hua. Ensemble Methods: Foundations and Algorithms. Boca Raton: CRC Press, 2012. 236 p.