UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO DE JOINVILLE
CURSO DE ENGENHARIA MECATRÔNICA

ISAÍAS GUTTIERRES RIBEIRO FERNANDES DE BARROS

A EQUIDADE DO SISTEMA CRIMINAL: UMA ANÁLISE DO VIÉS RACIAL NO SISTEMA COMPAS COM MODELOS DE INTELIGÊNCIA ARTIFICIAL

Joinville
2024

ISAÍAS GUTTIERRES RIBEIRO FERNANDES DE BARROS

A EQUIDADE DO SISTEMA CRIMINAL: UMA ANÁLISE DO VIÉS RACIAL NO SISTEMA COMPAS COM MODELOS DE INTELIGÊNCIA ARTIFICIAL

Trabalho apresentado como requisito parcial para obtenção do título de bacharel em Engenharia Mecatrônica, no Curso de Engenharia Mecatrônica, do Centro Tecnológico de Joinville, da Universidade Federal de Santa Catarina.

Orientador: Dr. Ricardo José Pfitscher

Joinville
2024

Dedico este trabalho a todos aqueles que estão contra o sistema.

AGRADECIMENTOS

Primeiramente eu agradeço esse trabalho a mim mesmo por ter forças ao meio de adversidades por conseguir me manter focado na meta e conseguir concluir o curso com um orgulho tremendo.
Eu agradeço também a minha parceira Nathália por me dar a sustentação e ajuda necessária nos momentos de provação e desemparo.
Também a minha família por ter me ensinado a importância dos estudos desde criança.
Por último eu agradeço esse trabalho a todos aqueles que sempre lutaram contra o sistema e mostraram sempre que a nossa presença é um dos maiores atos de provocação existentes. Vidas pretas importam. Esse trabalho é por vocês e para vocês.

Lá na rua facada no mano, podiam ajudar, mas ninguém encostou, ninguém quer ter é culpa se é que cê me entende Neguin tá morrendo e os neguin só olhou. (Djonga, 2022)

RESUMO

O trabalho presente tem como objetivo investigar o sistema Compas que é utilizado para avaliar a periculosidade de indivíduos decidindo se são elegíveis para receber liberdade provisória. O estudo parte da hipótese de que o sistema possui viés racial entrelaçado entre suas decisões e utiliza dois modelos de inteligência artificial — *KNN - K-Nearest Neighbors* e *Decision Tree* — para avaliar os resultados do sistema Compas. Em seguida a pesquisa realiza uma classificação com a entrada de novos dados para comparar as decisões dos sistemas que foram treinados a partir da base de dados original com o objetivo de verificar se existem possíveis disparidade entre os resultados de pessoas afro-americanas e caucasianas, quando são colocadas com o mesmo histórico. Dessa forma, o trabalho foca na análise de três critérios presentes na base de dados que são os riscos de violência, de não comparecimento no tribunal e de reincidência. Assim, a pesquisa aborda a falta de transparência e opacidade dos modelos que mesmo possuindo grande potencial de utilização podem perpetuar vieses raciais ao classificar indivíduos de diferentes étnicas de maneiras distintas.

Palavras-chave: Sistema Compas; Viés Racial; KNN; Árvore de decisão.

ABSTRACT

The present study aims to investigate the COMPAS system, which is used to assess the dangerousness of individuals and decide whether they are eligible for provisional release. The research is based on the hypothesis that the system contains embedded racial bias in its decisions and uses two artificial intelligence models — KNN (K-Nearest Neighbors) and Decision Tree — to evaluate the results of the COMPAS system. The study then performs a classification with the input of new data to compare the decisions of the systems trained on the original database, aiming to verify if there are potential disparities between the results for African American and Caucasian individuals when they are presented with the same history. Thus, the work focuses on the analysis of three criteria present in the database: the risks of violence, failure to appear in court, and recidivism. Furthermore, the research addresses the lack of transparency and opacity in the models, which, despite their great potential for use, may perpetuate racial biases by classifying individuals of different ethnicities in distinct ways.

Keywords: Compas System; Racial bias; KNN; Decision tree.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

Compas Correctional Offender Management Profiling for Alternative Sanctions

IA Inteligência Artificial

KNN K-Nearest Neighbors

ML Machine Learning

SUMÁRIO

1 INTRODUÇÃO

A inteligência artificial é uma tecnologia que tem ganhado destaque em diversas áreas, desde desenvolvimento de produtos, melhoria de segurança até a identificação de pessoas em redes. No entanto, por trás dessa inovação aparentemente neutra, esconde-se um problema profundo e preocupante que são as microagressões geradas que perpetuam certos comportamentos sociais que dão origem ao que é conhecido como racismo algorítmico (Silva, 2020).

Dentre os problemas que mais críticos atualmente, destaca-se o viés existente em modelos de inteligência artificial que muitas vezes podem ser causados por dados que são enviesados ou até pela maneira em que seus algoritmos são treinados. Um exemplo grande desse caso é o sistema Compas que é um software utilizado nos Estados Unidos para avaliar a possibilidade de liberdade provisória para detentos. Alguns estudos como Angwin *et al*. (2016) revelam que o Compas apresenta uma taxa de erro significativa quando comparado os resultados de pessoas afro-americanas e pessoas caucasianas reforçando dessa forma as desigualdades sociais presentes nos sistemas de justiça.

Tendo isso em vista o problema do racismo algoritmico, o presente trabalho tem como objetivo explorar a influência dessa problemática em um setor crucial à sociedade, que é o sistema judicial. Modelos de avaliação como o Compas são sistemas utilizados em grandes escalas para detectar a periculosidade de indivíduos para com a população e auxiliar em decisões judiciais (Skeem; Louden, 2007).

Além disso, será discutido como a falta de transparência e a opacidade desses sistemas que dificultam a identificação e correção de preconceitos, além de auxiliar a má interpretação de dados como mostrado em (Ribeiro; Singh; Guestrin, 2016). A negação do racismo e a ideia de neutralidade tecnológica muitas vezes obscurecem a natureza desses modelos e mascaram todo o viés racista aplicado pelos mesmos através de características fenotípicas de usuários e dificultam a correta identificação e correção dos próprios vieses (Buolamwini; Gebru, 2018).

1.1 OBJETIVOS DO TRABALHO

Para resolver a problemática descrita propõe-se neste trabalho os seguintes objetivos.

1.1.1 Objetivo geral

Por meio de uma análise crítica, este trabalho tem como objetivo entender como o racismo algorítmico é disposto em modelos inteligentes, examinando especificamente a influência da categorização étnico racial nas saídas do sistema Compas. Além de tudo, este estudo busca explorar como modelos de inteligência artificial percorrem os resultados do Compas e caracterizam suas classificações em bases de dados em que ocorre uma diferenciação de pessoas afro-americanas e caucasianas, buscando entender se reforçam um julgamento diferenciado para pessoas de diferentes etnias.

1.1.2 Objetivos específicos

- Analisar dados do sistema Compas para verificar a influência de categorias étnico raciais dentro dos resultados presentes;
- Implementar e comparar diferentes modelos de inteligência artificial para avaliar a classificação dos detentos;
- Identificar potências enviesamentos raciais nas classificações realizadas pelo sistema;
- Propor direções para próximos trabalhos que ajudem a entender as necessidades dos sistemas.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentadas as referências bibliográficas e todo o embasamento teórico sobre inteligência artificial, modelos de explicabilidade, questões raciais oriundas de modelos de inteligência artificial, desafios ético-sociais além de estudos de caso e trabalhos relacionados.

2.1 INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é um campo da ciência da computação que se dedica ao estudo e ao desenvolvimento de máquinas e programas computacionais capazes de reproduzir o comportamento humano na tomada de decisões e na realização de tarefas, desde as mais simples até as mais complexas (Norvig; Intelligence, 2002).

A IA opera através da análise de dados e da identificação de padrões, utilizando uma variedade de métodos para replicar comportamentos humanos por meio do computador. Assim, existem vários tipos de IA, cada um com suas próprias características e métodos de funcionamento. Vamos destacar alguns desses tipos.

2.1.1 Aprendizado de máquina

É um campo da IA que se concentra no desenvolvimento de algoritmos e modelos que permitem que os sistemas computacionais aprendam e se desenvolvam automaticamente a partir de algumas experiências como explicado em (Alpaydin, 2020). A *ML* abrange várias aplicações como por exemplo reconhecimento de padrões, previsão de séries temporais, recomendação de produtos e diagnóstico médico, entre outras.

2.1.2 Técnicas de inteligência artificial

Existem diversos típicos de técnicas voltadas para as mais variadas aplicações dentro da IA, dessa forma, o foco será em algumas das aplicações mais utilizadas dentro das diversas finalidades até aqui apresentadas.

2.1.2.1 Aprendizado Supervisionado

Aprendizado supervisionado é uma técnica de *Machine Learning ML* em que o algoritmo é treinado em um conjunto de dados rotulados. Cada exemplo de entrada está associado a uma saída desejada conhecida. Durante o treinamento, o algoritmo ajusta seus parâmetros para minimizar a diferença entre as saídas previstas e os rótulos

reais dos dados de treinamento. Essa técnica é amplamente utilizada em problemas de classificação e regressão, onde o objetivo é fazer previsões com base em dados de entrada (Hastie; Tibshirani; Friedman, 2009).

2.1.2.2 Aprendizado Não Supervisionado

Em Bishop e Nasrabadi (2006) ve-se que no aprendizado não supervisionado, o algoritmo é treinado em um conjunto de dados não rotulados, e o objetivo é encontrar padrões intrínsecos nos dados, como grupos naturais ou estrutura subjacente. Neste tipo de aprendizado, o algoritmo não recebe feedback explícito sobre as previsões, o que significa que não há rótulos de saída para orientar o processo de aprendizado. O aprendizado não supervisionado é comumente utilizado em tarefas como clusterização, redução de dimensionalidade e detecção de anomalias.

2.1.2.3 Aprendizado por Reforço

O aprendizado por reforço é uma técnica que consiste em um agente aprende a tomar decisões sequenciais interagindo com um ambiente. O agente recebe feedback na forma de recompensas ou punições após cada ação, e seu objetivo é aprender uma política que maximize a recompensa cumulativa ao longo do tempo. Essa técnica é amplamente aplicada em problemas de controle e tomada de decisão, como jogos, robótica e otimização de sistemas (Sutton; Barto, 2018).

2.1.2.4 Aprendizado Semi-Supervisionado

Em Sanches (2003) ve-se que o aprendizado semi-supervisionado é uma abordagem que combina elementos do aprendizado supervisionado e não supervisionado. Neste tipo de aprendizado, o algoritmo é treinado em um conjunto de dados que contém tanto dados rotulados quanto não rotulados. O objetivo é usar a informação disponível nos dados rotulados para melhorar o desempenho do modelo em tarefas de predição, aproveitando também os dados não rotulados para aprender representações mais robustas dos dados.

2.1.2.5 Aprendizado Ativo

O aprendizado ativo é uma técnica em que o algoritmo seleciona ativamente exemplos de dados para rotular com o objetivo de melhorar o desempenho do modelo. Em vez de esperar por um conjunto de dados completamente rotulado, o algoritmo escolhe estrategicamente quais exemplos seriam mais informativos para rotular, economizando tempo e esforço humano. Essa técnica é útil em situações em que a rotulagem manual de dados é cara ou demorada (Dasgupta; Hsu, 2008).

2.1.2.6 Aprendizado Online

O aprendizado *online* é uma abordagem em que o modelo é atualizado continuamente à medida que novos dados chegam, sem a necessidade de retreinamento completo do modelo. Isso é particularmente útil em cenários em que os dados chegam em fluxo contínuo e a distribuição dos dados pode mudar ao longo do tempo. Essa técnica é comumente utilizada em sistemas de recomendação, detecção de fraudes e previsão de séries temporais (Bottou, 2012).

2.2 MODELOS DE INTELIGÊNCIA ARTIFICIAL

Como já visto, existem diversos modelos de inteligência artificial que são categorizados de acordo com o método e suas funcionabilidades nos campos de *Machine Learning* e *Deep Learning* (Shinde; Shah, 2018).
Da sua maneira, cada um dos modelos possui suas próprias características, sendo melhor encaixadas dependendo do tipo de problema que será resolvido através das mesmas, sendo esses problemas frutos de uma classificação, previsão, agrupamento de dados entre outros. Dessa forma, foram selecionados dois tipos de modelos de aprendizado supervisionado para estudo que são:

- *K-Nearst Neighbors*;
- *Decision Tree*.

2.2.1 KNN

O modelo tem como objetivo atribuir a uma nova amostra que não está classificada no sistema a uma classificação de um ponto mais próximo que ja tenha sido classificado de acordo com as suas características (Cover; Hart, 1967). Dessa forma, pode-se ver o *KNN* como um classificador que se baseia na similaridade entre os dados ja classificados para os não classificados. Assim, irá realizar uma comparação entre os pontos de amostra com aqueles que estão a sua volta e uma vez que ocorra similaridade o novo ponto irá ser classificado de acordo com os demais a sua volta.

Em Peterson (2009) têm-se a definição de como modelo funciona seguindo a seguinte ordem:

- Definição dos parâmentros, escolhendo a quantidade de vizinhos próximos que será usado para realizar a classificação, além da seleção da métrica de distância que pode ser utilizada;
- Divisão do conjunto de dados em dados de treinamento, que será usado para determinar quais são os vizinhos e dados de teste que é onde as predições serão feitas;

- Realizar o cálculo das distâncias para cada um dos dados de teste em relação aos dados que pertencem ao conjunto de treinamento;
- Depois de realizar os cálculos de distância os valores dos k vizinhos mais próximos são escolhidos baseando-se nas menores distâncias calculadas;
- Por fim acontece a classificação da amostra de teste.

Dessa forma é possível perceber que o método é de simples implementação e possui boa adaptabilidade a problemas complexos. Entretanto, possui um custo operacional grande, uma vez que, computa as distâncias de todos os dados e possui necessidade de normalização dos dados previamente pois características diferentes podem vir a adulterar o cálculo.

2.2.2 Árvore de Decisão

As árvores de decisão são um modelo de aprendizado de máquina que são utilizados para realizar classificações e regressões, onde possuem como objetivo realizar a previsão de um conjunto de dados a partir de um aglomerado de dados de entrada. Para realizar essa função é contido em sua estrutura uma sequência lógica onde cada nó dentro da árvore possui a representação uma das características analisadas, elas são guiadas pelos ramos que são os resultados das condições apresentadas nos nós até as folhas que podem representar uma classe em casos de classificação ou um valor contínuo para casos de regressão (Loh, 2011). Sabendo-se disso, para montar a árvore de decisão segue-se os passos:

- Definição do problema, classificação ou regressão;
- Escolha do melhor atributo para dividir os dados;
- Depois da escolha da característica os dados devem ser divididos em subgrupos em relação aos valores da característica escolhida. Assim, cada nó filho da árvore será gerado por cada um desses subgrupos e o processo se repete para cada novo nó recursivamente;
- Repetição do procedimento de divisão até que seja atingido o critério de parada determinado, que podem ser como profundidade máxima, número mínimo de amostras;
- Atribuição de classes ou valores para as folhas da árvore;
- Por fim realizar a predição de dados novos baseados na árvore montada.

Desse modo, como mostrado em Lu (2010) o modelo de árvore de decisão possui fácil interpretabilidade, suporta dados categóricos, não possui necessidade de normalização dos valores e faz as seleções de características de maneira automática. Em contrapartida são propensos a overfitting em árvores que venham a ser profundas,

é um sistema instável que pode sofrer grande modificação a partir de alteração nos dados.

2.3 MÉTRICAS

As métricas de avaliação de um modelo de inteligência artificial são importantes pois fornecem de maneira objetiva como medir o desempenho de diversos modelos. Sendo assim, é possível avaliar a qualidade do sistema, realizar comparações entre modelos diferentes, identificar pontos fracos, realizar adaptabilidade do sistema, validar a confiabilidade e auxiliar na tomada de decisão de uso ou não de um modelo.

As métricas podem ser divididas em métricas de desempenho e de concordância, sendo as de desempenho utilizadas para avaliar a eficácia de sistemas e as de concordância medem a concordância entre os anotadores (Matos *et al.*, 2009). Dentro das métricas de desempenho se encontram:

- Precisão, mede a proporção de exemplos verdadeiros positivos em relação ao total de exemplos positivos;
- Sensibilidade, mede a proporção de verdadeiros positivos em relação a soma de verdadeiros positivos com falsos negativos;
- Acurácia, mede a relação de exemplos verdadeiros sobre todos os exemplos;
- F-1 *Score*, combina a precisão e a sensibilidade utilizando média harmônica entre elas;
- Especificidade, mede a relação de verdadeiros negativos sobre a soma de verdadeiros negativos com falsos positivos.

Para as métricas de concordância o artigo Matos *et al.* (2009) mostra a classificação Kapaa que é pertinente de ser usada para tarefas de classificação realizada por diversos anotadores. Seu resultado ajuda a avaliar o desempenho do sistema, se o conjunto de treinamento é valido e por último descartar exemplos controversos.

2.4 MODELOS DE EXPLICABILIDADE

Os modelos de explicabilidade em *AI* são técnicas e abordagens projetadas para tornar o processo de tomada de decisão mais transparente e compreensível para os usuários humanos. Em outras palavras, esses modelos buscam fornecer explicações ou justificativas para as decisões tomadas por algoritmos de modo que os usuários possam entender como e por que uma determinada decisão foi feita (Jacobs *et al.*, 2022).

A seguir, serão abordadas algumas das principais técnicas de explicabilidade, como a interpretabilidade do modelo, a geração de explicações pós-hoc, a sensibilidade a atributos e a auditoria do modelo, que visam tornar as decisões algorítmicas mais compreensíveis e transparentes.

2.4.1 Interpretabilidade de Modelo

A interpretabilidade de modelo visa tornar os modelos de inteligência artificial mais transparentes, permitindo aos usuários entenderem como as entradas são transformadas em saídas pelo modelo. Isso é especialmente importante em cenários onde a confiança e a compreensão do funcionamento interno do modelo são cruciais, como na área médica ou financeira. Técnicas de interpretabilidade, como visualizações de mapas de ativação em redes neurais convolucionais, permitem que os usuários observem quais características das entradas são mais relevantes para as decisões do modelo, facilitando assim a compreensão do raciocínio por trás das previsões ou classificações (Sundararajan; Taly; Yan, 2017).

2.4.2 Geração de Explicações Pós-Hoc

A geração de explicações pós-hoc é uma estratégia essencial para tornar os modelos de inteligência artificial mais compreensíveis e transparentes para os usuários humanos como explicado em Ribeiro, Singh e Guestrin (2016). Essa abordagem busca fornecer explicações adicionais após a tomada de decisão do modelo, permitindo que os usuários entendam o raciocínio por trás das previsões ou classificações.

Por exemplo, em sistemas de saúde, quando um modelo de IA recomenda um determinado tratamento, é crucial que os médicos possam entender as razões por trás dessa recomendação para tomar uma decisão informada. Métodos como a geração de explicações baseadas em linguagem natural e a identificação de características importantes nas entradas são aplicados para criar justificativas compreensíveis e intuitivas, isso não apenas aumenta a confiança dos usuários no modelo, mas também fornece *insights* valiosos sobre como o modelo está fazendo suas previsões, possibilitando melhorias e ajustes quando necessário.

2.4.3 Sensibilidade a Atributos

A sensibilidade a atributos é uma técnica fundamental para avaliar como as mudanças nas entradas afetam as saídas do modelo de inteligência artificial. Ao compreender quais características das entradas são mais influentes na decisão final do modelo, os usuários podem identificar potenciais vieses e entender melhor o funcionamento do sistema (Lundberg; Lee, 2017).

Essa compreensão é crucial para garantir a transparência e a justiça do sistema, permitindo que os desenvolvedores ajustem o modelo e mitiguem possíveis vieses antes de sua implantação. Ao fornecer insights sobre o processo de tomada de decisão do modelo, a sensibilidade a atributos desempenha um papel fundamental na construção de sistemas de inteligência artificial mais transparentes, éticos e equitativos.

2.4.4 Auditoria de Modelo

A auditoria de modelo é uma prática crucial para avaliar e monitorar continuamente o desempenho e comportamento dos modelos de inteligência artificial ao longo do tempo como demonstrado em Mittelstadt, Russell e Wachter (2019).

Essa abordagem envolve a análise sistemática dos dados de entrada, saída e do próprio modelo para identificar possíveis vieses, erros ou falhas que possam comprometer a qualidade das decisões tomadas pelo modelo.

Em cenários onde a transparência e a equidade são essenciais, como em sistemas de justiça criminal, a auditoria de modelo desempenha um papel fundamental na identificação e mitigação de possíveis preconceitos ou injustiças, sendo assim, a auditoria de modelo também pode ajudar a garantir que os modelos de IA estejam em conformidade com regulamentos e políticas éticas, promovendo assim a confiança e aceitação por parte dos usuários e da sociedade em geral.

2.5 COMPAS

O sistema Compas é uma ferramenta que foi desenvolvida para avaliar o risco de reincidência de infratores. Criado pela *Northpointe Institute for Public Management* é utilizado para informar decisões sobre liberação e gestão desses infratores com base nos dados que são coletados do histórico criminal e sua resposta a questionários. De acordo com o Skeem e Louden (2007) o sistema avalia uma série de fatores como comportamento criminal, necessidades sociais, personalidade entre outros. Além disso, utiliza dados normativos para comparar as notas de um detento com outros da mesma agência e gera estimativas de risco relacionadas a violência, reincidência e falhas de comparecimento a audiências.

O Compas é projetado para ser usado por profissionais de justiça criminal, como agentes de condicional, e é operado por meio de software que calcula notas para diferentes variáveis e gera um plano de gestão de caso para cada infrator. O sistema foi desenvolvido com base em teoria criminológica e pesquisa empírica, porém é flexível o que permite que as agências escolham quais escalas usar, dependendo de suas necessidades e realidades.

Dessa forma, de acordo com o Northpointe Inc. (2015) o Compas consiste em escalas preditivas utilizadas para a previsão de risco, possuindo escalas separadas

para identificar as necessidades de programas nas áreas de emprego, moradia, entre outras a partir de suas notas. As notas do Compas são transformadas em escalas decimais que variam entre 1 (mais baixo) a 10 (mais alto) que indicam a posição de avaliação do infrator em relação ao grupo normativo. Assim, a nota 1 corresponde aos 10% mais baixo, a nota 2 aos 10-20% e assim por diante. Porém de maneira geral o ranking possui a seguinte interpretação:

- 1 a 4: escala decimal é baixa em relação a outros infratores no grupo normativo;
- 5 a 7: escala decimal é média em relação a outros infratores no grupo normativo;
- 8 a 10: escala decimal é alta em relação a outros infratores no grupo normativo.

2.6 RACISMO E DISCRIMINAÇÃO EM SISTEMAS DE INTELIGÊNCIA ARTFICIAL

De acordo com Almeida (2018), o racismo é uma ideologia sistemática de discriminação que leva como parte central o conceito de raça onde pode se manifestar de maneiras conscientes ou inconscientes resultando assim em vantagens ou desvantagens para indivíduos diferenciando o seu grupo racial. Por sua vez, a descriminação racial é a atribuição de um diferente tratamento a pessoas de grupos identificados racialmente, assim tendo em seu cerne a questão do poder seja de maneira direta através de repúdios e proibições a esses grupos ou de maneira indireta onde as diversas situações vividas por diferentes pessoas de diferentes etnias são ignoradas pela sociedade.

O racismo e a descriminação em sistemas de inteligência artificial são fenômenos preocupantes que refletem e amplificam os preconceitos existentes na sociedade. Algoritmos de IA, quando treinados em conjuntos de dados desbalanceados ou enviesados, podem gerar resultados discriminatórios que afetam grupos minoritários de forma desproporcional (Angwin *et al.*, 2022).

Os algoritmos de IA podem reproduzir e amplificar preconceitos raciais presentes nos dados de treinamento. Isso ocorre porque os algoritmos de aprendizado de máquina aprendem com padrões nos dados fornecidos, e se esses dados contiverem viés racial, os algoritmos podem aprender a fazer previsões discriminatórias (Flores; Bechtel; Lowenkamp, 2016). Esses sistemas podem perpetuar desigualdades sociais e econômicas em diversas áreas, incluindo recrutamento automatizado e justiça criminal.

Em sistemas de recrutamento automatizado, algoritmos podem aprender a favorecer candidatos de determinadas origens étnicas com base em padrões históricos de contratação, perpetuando assim desigualdades raciais no mercado de trabalho. No contexto de justiça criminal, algoritmos de previsão de reincidência podem atribuir erroneamente taxas de risco mais altas a indivíduos pertencentes a grupos raciais minoritários, resultando em sentenças mais longas e injustas. Esses exemplos ilustram

como a IA pode amplificar preconceitos sociais, em vez de promover equidade.

A discriminação racial em inteligência artificial é um fenômeno complexo que levanta preocupações éticas e sociais significativas. Enquanto a IA é amplamente vista como uma ferramenta para aprimorar processos e tomar decisões de forma mais eficiente, sua aplicação indiscriminada pode resultar em consequências prejudiciais para grupos minoritários.

Como demonstrado em Buolamwini e Gebru (2018) esses sistemas discriminatórios podem causar danos significativos, ampliando as disparidades sociais e econômicas já existentes. Além disso, levantam questões éticas sobre a justiça e equidade na aplicação da IA. É crucial abordar o racismo em sistemas de IA implementando abordagens proativas para mitigar vieses e garantir a equidade e transparência na tomada de decisões algorítmicas. Isso inclui o desenvolvimento de conjuntos de dados mais representativos e diversificados, a implementação de técnicas de explicabilidade para entender e corrigir vieses algorítmicos e a promoção da diversidade e inclusão na equipe de desenvolvimento de IA.

2.6.1 Viés Algorítmico

O viés algorítmico é uma preocupação crescente no campo da inteligência artificial, onde algoritmos podem aprender e reproduzir preconceitos presentes nos dados de treinamento. Este fenômeno pode resultar em resultados discriminatórios que afetam grupos raciais minoritários de maneira desproporcional. Por exemplo, em sistemas de reconhecimento facial, algoritmos podem ser enviesados para identificar erroneamente pessoas de certas origens étnicas com maior frequência do que outras. O estudo de Caliskan, Bryson e Narayanan (2017) revelou que até mesmo dados linguísticos podem conter vieses que são reproduzidos por algoritmos de processamento de linguagem natural, destacando a importância de abordar essas questões.

Além disso, o viés algorítmico pode se manifestar em sistemas de recomendação, onde usuários de diferentes grupos demográficos podem receber recomendações de conteúdo diferentes com base em preconceitos incorporados nos algoritmos. Esses vieses podem perpetuar estereótipos e marginalizar grupos sociais, impactando negativamente sua experiência online e reforçando desigualdades existentes. Portanto, é crucial que os desenvolvedores de algoritmos considerem cuidadosamente o viés presente nos dados de treinamento e implementem técnicas para mitigar seu impacto.

2.6.2 Impacto Social

A discriminação racial em sistemas pode ter impactos sociais significativos, ampliando as disparidades e injustiças já existentes na sociedade, como explicitado em Obermeyer *et al*. (2019).

Por exemplo, em sistemas de justiça criminal, algoritmos de previsão de riscos podem resultar em sentenças mais longas e injustas para indivíduos pertencentes a grupos raciais minoritários, perpetuando assim o ciclo de desigualdade e injustiça como é demonstrado no sistema Compas nos Estados Unidos da América (Angwin *et al*., 2022).

Utilizando a mesma ideologia, além de sistemas prisionais podem ocorrer discriminações em sistemas de recrutamento e seleção, uma vez que grupos minoritários não são considerados hábeis a determinadas funções e cargos sociais, sendo assim, delimitando oportunidades de emprego e ascensão social para certos grupos raciais, contribuindo para disparidades econômicas e sociais que já são presentes na sociedade.

Logo, é crucial abordar a discriminação racial em *AI* para promover uma sociedade mais justa e equitativa para todos, evitando dessa forma as disparidades e pré-conceitos ja existentes na sociedade.

2.6.3 Técnicas de mitigação

Para mitigar o viés algorítmico em sistemas de inteligência artificial, são necessárias abordagens proativas que visam corrigir e compensar as disparidades presentes nos dados de treinamento. Uma abordagem comum é a coleta e análise cuidadosa dos dados para identificar e remover vieses existentes. Isso pode envolver o equilíbrio dos conjuntos de dados para garantir representatividade de todos os grupos raciais, bem como a aplicação de técnicas de pré-processamento para remover informações sensíveis que possam levar a discriminação (Mitchell *et al*., 2019).

Além disso, técnicas de aprendizado de máquina justas e equitativas estão sendo desenvolvidas para garantir que os modelos de IA não gerem resultados discriminatórios. Isso inclui o uso de métricas de equidade durante o treinamento do modelo e o ajuste dos algoritmos para garantir que as previsões sejam justas e imparciais para todos os grupos raciais. No entanto, é importante reconhecer que não existe uma solução única para mitigar o viés algorítmico, e uma abordagem multifacetada e contínua é necessária para garantir que sistemas de IA sejam éticos e equitativos em sua aplicação.

Essas técnicas de mitigação são fundamentais para promover a justiça e a equidade em sistemas de inteligência artificial, garantindo que esses sistemas não perpetuem ou ampliem desigualdades sociais existentes.

2.7 DESAFIOS ÉTICOS E SOCIAIS

O reconhecimento facial apresenta uma série de desafios éticos e sociais que são cada vez mais preocupantes à medida que essa tecnologia é mais amplamente adotada. Como demonstrado em Garvie (2016) um dos principais desafios é a questão da privacidade e da vigilância em massa. Sistemas de reconhecimento facial podem ser utilizados para monitorar e rastrear pessoas em espaços públicos sem seu consentimento, levantando preocupações sobre o direito à privacidade e o potencial para o uso indevido de dados pessoais. Além disso, a precisão desses sistemas pode ser prejudicada por vieses e imprecisões, especialmente quando se trata de identificar pessoas de diferentes grupos étnicos, levando a casos de discriminação e injustiça. Isso é especialmente preocupante em contextos como a aplicação da lei, onde decisões baseadas em sistemas de reconhecimento facial podem ter sérias consequências para as pessoas identificadas de forma incorreta.

Outro desafio ético importante está relacionado à falta de transparência e responsabilidade nos algoritmos de reconhecimento facial. Muitos desses algoritmos são caixas pretas, o que significa que é difícil entender como tomam decisões ou identificar e corrigir possíveis vieses. Isso levanta questões sobre a prestação de contas e a justiça dos sistemas de reconhecimento facial, especialmente quando são usados para tomar decisões importantes que afetam a vida das pessoas. Além disso, a falta de regulamentação e supervisão adequadas desses sistemas pode resultar em seu uso indiscriminado e potencialmente prejudicial, sem salvaguardas adequadas para proteger os direitos individuais e a dignidade humana.

Para lidar com esses desafios, é crucial adotar uma abordagem ética e responsável para o desenvolvimento e uso de tecnologia de reconhecimento facial. Isso inclui a implementação de políticas e regulamentos que protejam os direitos individuais à privacidade e à não discriminação, bem como a promoção da transparência e explicabilidade nos algoritmos de reconhecimento facial. Além disso, é importante envolver a sociedade civil, especialistas em ética, grupos de direitos civis e outros stakeholders na discussão e tomada de decisões sobre o uso dessa tecnologia, garantindo que seja usada de forma justa, equitativa e socialmente responsável.

2.8 TRABALHOS RELACIONADOS

Esta seção tem como objetivo trazer artigos e estudos que abordam o viés racial do modelo Compas que é uma ferramenta de predição utilizada no sistema penal nos Estados Unidos da América. Primeiramente, Angwin *et al*. (2016) mostraram que o sistema Compas tem uma tendência de realizar a classificação incorreta de pessoas negras em relação a pessoas brancas o que pode reforçar as desigualdades raciais presentes na sociedade.

Além disso, Chouldechova (2017) verificou qual o impacto do Compas nas comunidades consideradas minoritárias e como a utilização dessa ferramenta pode resultar em disparidades nos resultados judiciais. Dessa forma, mostrando a necessidade de métodos de avaliação que sejam mais justos e considerem os dados históricos e suas repercussões.

De forma semelhante, focando no efeito de impacto da raça sobre os resultados do Compas Khademi e Honavar (2020) conseguiu provar que a pontuação de reincidência do Compas exibe um viés racial contra pessoas classificadas como afro americanas e que esses réus tem uma chance de 1,87 vezes em relação a réus classificados como caucasianos a ser categorizados como reincidentes.

Sendo assim, este trabalho tem um posicionamento de reforçar as evidências existentes sobre o viés racial do Compas trazidas nos textos anteriores, porém avança sobre o tema ao aplicar técnicas de IA para fazer a verificação dessas disparidades ao possuir uma análise experimental que verifica se os padrões observados persistem em diferentes contextos, complementando dessa forma as análises estatísticas dos estudos relatados.

3 METODOLOGIA

O estudo começa com a análise do banco de dados da ferramenta Compas que foi desenvolvida pela Northpointe, Inc. Essa ferramenta computacional tem como objetivo fazer a classificação de um detento em três fatores, indicando assim suas possibilidades. Para fazer a classificação são separados três fatores:

- Risco de Violência;
- Risco de Reincidência;
- Risco de Não Comparecimento.

Para cada fator o réu recebe uma nota que varia de 1 a 10 e essas notas são classificadas da seguinte forma:

- *Low*: De 1 até 4;
- *Medium*: De 5 até 7;
- *High*: De 8 até 10.

Para realizar as análises em relação a existência de um viés racista na plataforma, foram usados dados do Condado de Broward na Flórida dos anos de 2013 e 2014 que são fornecidos pela (Angwin *et al*., 2016) através de um estudo realizado pela Propública sobre o Compas.

3.1 TRATAMENTO DE DADOS

O processo de tratamento de dados é fundamental para garantir a qualidade e a consistência das análises realizadas neste estudo. Os dados utilizados são extraídos do banco de dados público disponibilizado pela Angwin *et al*. (2016), contendo informações detalhadas sobre as avaliações realizadas pelo sistema Compas.

Primeiramente, os dados são importados para o software *Orange* por meio de um arquivo csv. Após a importação, é realizada uma amostragem de 70% dos dados originais, utilizando o módulo *Data Sampler*, para garantir uma análise eficiente e representativa.

Após isso, os dados passam por um processo de limpeza, onde são removidas inconsistências, como valores nulos diretamente pelo software *Orange*. As colunas irrelevantes para a análise são excluídas, enquanto as variáveis essenciais são identificadas e categorizadas como *features, targets* e metas.

3.2 CENÁRIOS DE IMPLEMENTAÇÃO

Nos cenários de implementação, são testados dois modelos de inteligência artificial, *Knn* e *Tree*, com o objetivo de avaliar o impacto da variável étnico-racial nas classificações realizadas. Para isso, são considerados dois cenários distintos:

No primeiro cenário os modelos são treinados utilizando a base de dados original, que inclui a variável de etnia como uma das *features*. Isso permite que os algoritmos considerem a informação étnico-racial ao realizar as predições, refletindo o comportamento do sistema original.

No segundo cenário, a variável de etnia é removida do conjunto de *features* antes do treinamento dos modelos. Dessa forma, os algoritmos realizam as predições sem considerar qualquer informação relacionada à classificação racial.

3.3 CENÁRIOS DE USO

A partir da validação dos resultados, utilizam-se dois cenários com os modelos de inteligência artificial salvos, que incluem os valores étnico-raciais dos detentos, para dar sequência à análise de dados e realizar a comparação entre diferentes etnias. As etnias selecionadas são afro-americana e caucasiana, por representarem a maior parte dos dados na base.

Com esses modelos salvos, são aplicados a dois novos arquivos, nos quais os modelos realizam a classificação das novas bases de dados. Os dois arquivos são idênticos, porém serão manipulados para poderem se diferenciar apenas nos valores da categoria etnia: em um, todos são classificados como caucasianos, e no outro, como afro-americanos. Esse processo permite verificar se há diferenças na classificação entre as duas etnias.

Os modelos salvos, *Knn* e *Tree*, são então utilizados para avaliar as bases de dados e classificá-las entre os valores *Low, Medium* e *High* nas categorias de risco de não comparecimento, risco de violência e risco de reincidência.

3.4 COMPARAÇÕES

A comparação entre os dois cenários de implementação permite analisar se a ausência da variável étnico-racial afeta significativamente os resultados das predições e, consequentemente, identificar possíveis vieses nos modelos originais. A análise foca em verificar se os modelos com e sem a variável de etnia apresentam discrepâncias nas classificações de risco *Low, Medium* e *High* nas categorias de risco de não comparecimento, risco de violência e risco de reincidência.

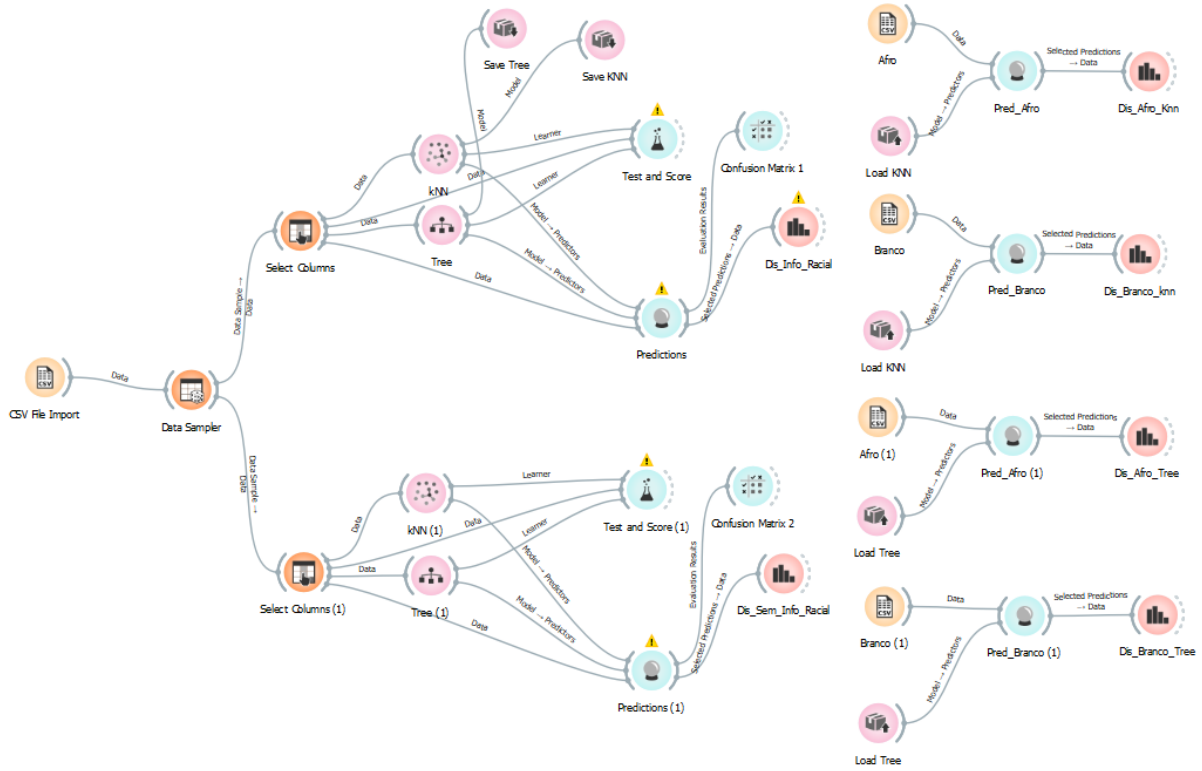
Por sua vez, a comparação entre os dois cenários de uso, permite analisar se a distinção entre a classificação étnica do detento afeta a maneira em que as predições

são realizadas. Assim, essa comparação foca na percepção da discrepância ou não de análises de acordo com a etnia e como essas analises podem prejudicar o julgamento de uma pessoa a classificando de maneira distinta de alguém que tenha o mesmo histórico porém outra etnia.

3.5 RELATÓRIO

Uma vez com o software instalado foi construido um fluxo inicial de aprendizado de dados onde é posta em evidência a informação étnico-racial de cada uma das pessoas. Esse fluxo foi pensado de acordo com a Figura 1

Figura 1 – Visualização Geral



Fonte: Elaborado pelo autor (2024).

O processo de análise começa com a importação do arquivo para o software por meio da funcionalidade *CSV File Import*. Em seguida, utiliza-se o *Data Sampler* para uma melhor visualização da amostragem de dados. Na etapa seguinte, os dados são organizados através do *Select Columns*, onde as variáveis são categorizadas como Features, Target, Metas e Ignorados.

Dois modelos de inteligência artificial são empregados: *Knn* e *Tree*. Após a configuração dos modelos, utiliza-se a célula *Test and Score* para avaliação de desempenho. Em seguida, cria-se uma célula de predição de dados e uma matriz de confusão, que permite a análise detalhada dos erros de classificação. Além disso, uma célula de distribuição de dados é utilizada para visualizar a dispersão das previsões.

Cada modelo é salvo individualmente para futuras análises. Posteriormente, os arquivos contendo dados de indivíduos identificados como afro-americanos e caucasianos são importados para o sistema. Os modelos de IA previamente treinados são carregados e utilizados em novas células de predição. Por fim, células de distribuição específicas são configuradas para cada novo carregamento, permitindo uma análise comparativa.

Segue uma descrição de cada uma das funções apresentadas na Figura 1

3.5.1 CSV File Import

A primeira ferramenta utilizada no fluxograma criada dentro do *Orange* é um importador de arquivo CSV. Possui como definição ler arquivos separados por vírgulas, ponto e vírgula, espaços, tabulações ou qualquer delimitador que venha a ser personalizado. Para a montagem do *prompt* foram utilizadas as seguintes definições:

- Codificação UTF-8;
- Delimitador por vírgula;
- Caractere de citação usado são aspas duplas.

O arquivo possui 28 Colunas que podem ser classificadas como auto, numérico, *categorical*, texto, data ou ignorado.

- A classificação automática é a configuração padrão do sistema, assim o software decidira qual é a melhor classificação para essa coluna de dados;
- A classificação numérica é utilizada para dados que possuam números contínuos;
- A classificação categórica é usada para dados que são classificados como categorias ou rótulos;
- A classificação texto é selecionada para dados que venham na forma de texto;
- A classificação data é para dados relacionados ao tempo;
- A classificação ignorar é para dados que não devem ser incluídos na análise.

Para definição da análise os seguintes dados foram ignorados pois não foram classificados como relevantes para a determinação dos graus de risco dos detentos:

- Id de pessoa;
- Id de avaliação;
- Id do caso;
- Agência;
- Último nome;
- Primeiro nome;

- Nome do meio;
- Id do conjunto de escalas;
- Conjunto de escalas;
- Razão da avaliação;
- Estado civil;
- Data de prisão;
- Tipo de Avaliação.

De mesma forma, os dados que serão utilizados são demonstrados a seguir junto com a sua categorização dentro do software.

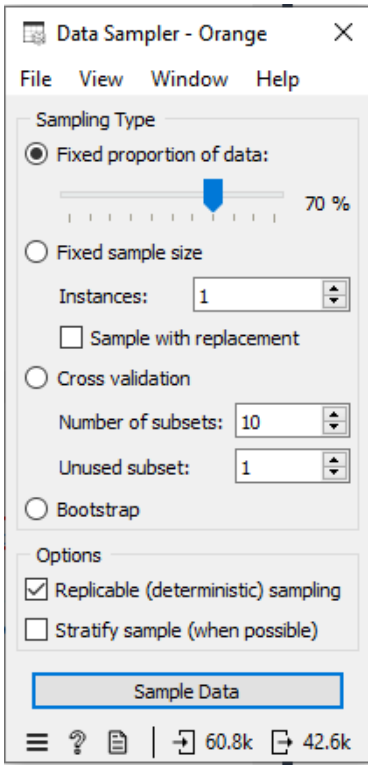
- Identificação - automático;
- Etnia - categórico;
- Data de nascimento - automático;
- Razão de avaliação - automático;
- Idioma - automático;
- Estado legal - automático;
- Estado de custódia - automático;
- Nota de supervisão - numérico;
- Texto de supervisão - categórico;
- Id de escala - numérico;
- Display de texto - categórico;
- Pontuação geral - numérico;
- Pontuação em sistema decimal - numérico;
- Texto de nota - categórico.

3.5.2 Data sampler

A ferramenta *Data Sampler* Data Sampler implementa diversos métodos de amostragem de dados, com isso um conjunto de dados amostrado e um conjunto complementar (com instâncias do conjunto de entrada que não estão incluídas no conjunto amostrado).

Essa ferramenta tem algumas configurações para realizar as amostragens de dados como demonstrado em sua página de ajuda, dessa forma foi selecionado de acordo com Gareth *et al*. (2013) a amostragem a partir de uma proporção fixa de dados em 70% para treinamento dos modelos como demonstrado na Figura 2.

Figura 2 – Amostragem de dados



Fonte: Elaborado pelo autor (2024).

3.5.3 Select Columns

Essa ferramenta é utilizada para compor de maneira manual o conjunto de dados, dessa forma possibilitando o usuário a decidir quais atributos vão ser utilizados para a análise e como serão utilizados. Dessa forma, como mostrado na Figura 3, os dados foram separados de forma a serem ignorados, utilizados como definição meta e como *features*.

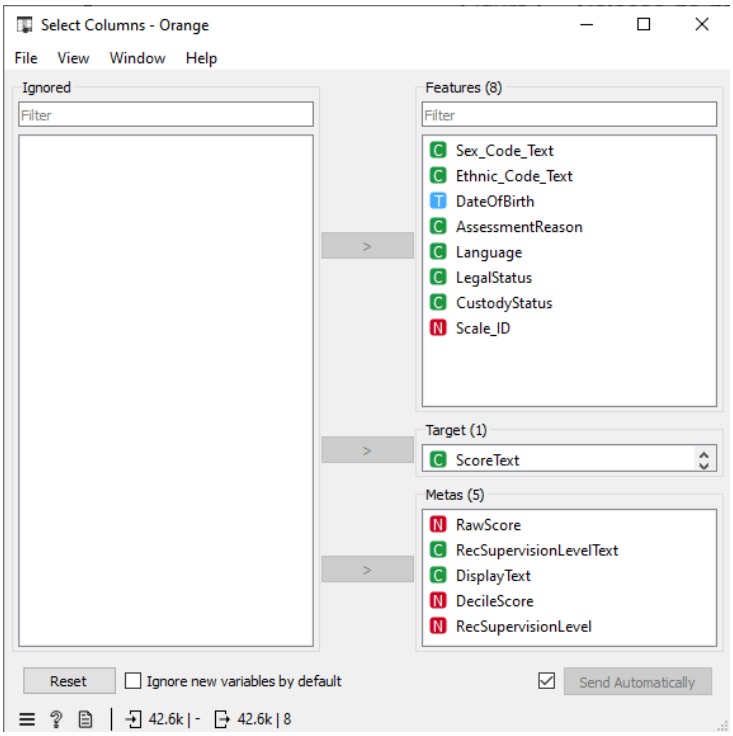
A partir dessa ferramenta serão exportados os dados para as comparações de teste e pontuação, predição e para os modelos de inteligência artificial que irão usar os dados para realização da modelagem.

De acordo com a Figura 3 é possível visualizar que como todos os dados ja foram ignorados previamente, não possui dados ignorados. As *Features* são os dados utilizados para poder fazer a classificação dos detentos. Por sua vez, o *Target* é o objetivo final da análise e por último as metas são os dados que são saídas do sistema Compas.

3.5.4 Modelos de Inteligência Artificial

Foram utilizados dois modelos de inteligência artificial para realização da modelagem dos dados. Assim, irá ser discutido textualmente como cada um foi configurado para realização dessa tarefa.

Figura 3 – Seleção de dados



Fonte: Elaborado pelo autor (2024).

3.5.4.1 Knn

O modelo *Knn* utiliza o algoritmo *Knn* que procura pelo k mais próximo nos exemplos de treinamento mais próximos dentro de um espaço de características e usa uma média como sua previsão.

Dentro do modelo pode-se escolher por 4 métricas diferentes que são:

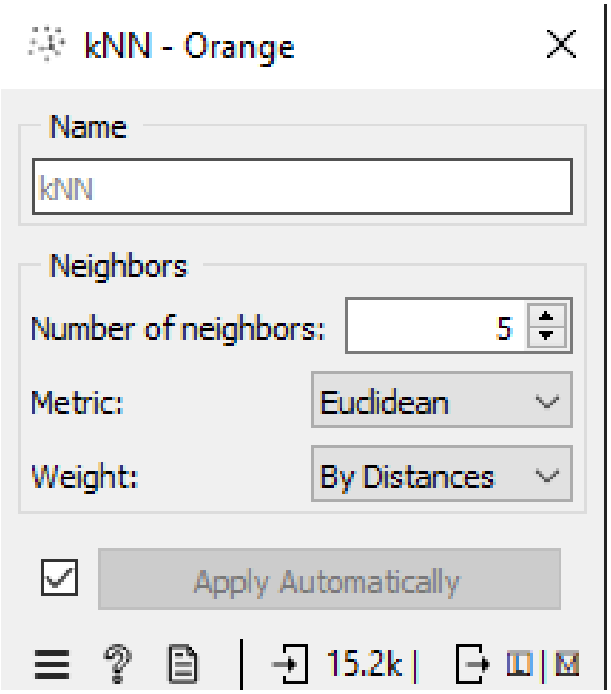
- Euclidiana, que calcula a linha reta entre 2 pontos.
- Manhattan, que calcula a soma das diferenças absolutas de todos os atributos;
- Maximal, que calcula a maior das diferenças absolutas entre os pontos;
- Mahalanobis, que calcula a distância entre o ponto e distribuição.

Além disso, deve-se selecionar o peso que os vizinhos exercem sobre uma célula k, e esse são categorizados da seguinte maneira:

- Uniforme, faz com que todos os pontos em cada vizinhança tenham o mesmo peso;
- Distância, esse peso faz com que os vizinhos mais próximos de um ponto de consulta tenham um maior peso de influência do que os vizinhos que estão mais distantes.

Depois de entender as configurações foi selecionado 5 vizinhos para funcionamento do modelo, pensando na resistência a ruídos nos dados, com métrica euclidiana, pois era a métrica mais simples e utilizada nos tutorias do *Orange*. Além disso foi utilizado peso de distância para influência dos dados, como mostrado na Figura 4

Figura 4 – Modelo Knn



Fonte: Elaborado pelo autor (2024).

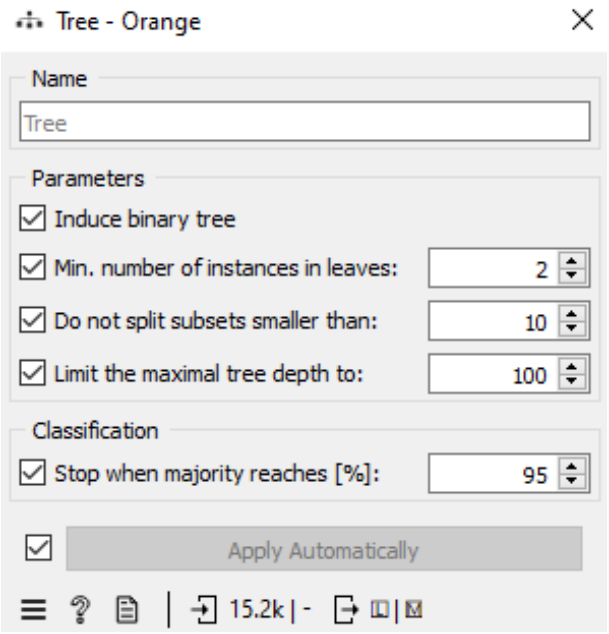
3.5.4.2 Árvore de decisão

Árvores de decisão são algoritmos simples que dividem os dados em nós e especificamente no Orange elas lidam com dados categóricos ou numéricos. Dentro da árvore são selecionadas as seguintes categorias:

- Arvore binária;
- Número mínimo de instâncias nas folhas;
- Tamanho mínimo para separação de dados;
- Tamanho máximo de profundidade da árvore;
- Parar a classificação quando a maioria chegar em uma porcentagem específica.

Através desses fatores, foram selecionados os dados demonstrados na Figura 5 para gerar a árvore de decisão.

Figura 5 – Modelo árvore de decisão



Fonte: Elaborado pelo autor (2024).

A imagem mostra a configuração de uma árvore de decisão no software *Orange*. A árvore é configurada para ser binária, com no mínimo 2 instâncias por folha e sem dividir subconjuntos menores que 10 instâncias. A profundidade máxima da árvore é limitada a 100 níveis, e a divisão de um nó é interrompida quando 95% das instâncias pertencem à mesma classe. Essas configurações visam equilibrar a complexidade do modelo, evitando overfitting e garantindo que a árvore seja mais generalizável e eficiente.

3.5.5 Test and Scores

Essa ferramenta tem como objetivo fornecer os valores de teste sobre os dados analisados. Para fornecer esses valores o modelo pode usar os seguintes métodos:

- Validação cruzada, realiza a validação cruzada, porém os *folds* são definidos pela classificação categórica no meta recursos;
- Amostragem aleatória, divide os dados de maneira aleatória em conjuntos de treinamento e teste na proporção especificada;
- *Leave-one-out*, é um método mais estável e confiável, porém bastante lento. O método retém uma instância por vez utilizando todas as outras para treinar o modelo em seguida classifica a instância que foi retida;
- Teste em dados de treinamento, utiliza todo o conjunto de dados para treinamento e em seguida ara teste;
- Teste em dados de teste, utiliza um conjunto separado de dados para avaliar o

modelo.

Para a classificação, é possível selecionar uma classe alvo do classificador, caso não seja selecionado uma classe alvo as pontuações vão ser calculadas em decorrência das médias ponderadas sobre todas as classes. Além disso, a ferramenta pode calcular uma série de estatísticas de desempenho que são:

- Área sob a Roc, área sob a curva do receptor;
- Acurácia de classificação, proporção de objetos classificados corretamente;
- F-1, medida harmônica ponderada;
- Precisão, é a proporção de verdadeiros positivos entre as instâncias classificadas como positivas;
- Recall, é a proporção de verdadeiros positivos entre todas as instâncias positivas;
- Especificidade, é a proporção de verdadeiros negativos entre todas as instâncias negativas;
- Logloss, é a perda de entropia que é considerada através da incerteza da previsão em relação ao quanto varia do rótulo real;
- Coeficiente de correlação de Matthews, conta os verdadeiros e falsos positivos e negativos;
- Tempo de treino, é o tempo gasto em segundos para treinar os modelos;
- Tempo de teste, é o tempo gasto em segundos para testar os modelos.

Dessa forma, os parâmetros selecionados foram a validação cruzada com 10 *folds* se baseando nos tutorias do *Orange*.

3.5.6 Predições

Essa ferramenta tem como entrada um conjunto de dados e um ou mais de um modelo preditivo, dessa forma, irá fornecer dados e previsões. A saída do *widget* consiste em um *data frame*, onde as previsões são concatenadas aos dados originais como novas colunas, assim o usuário tem a flexibilidade de selecionar quais colunas serão incluídas na saída final, permitindo a visualização de dados originais, previsões numéricas e, se aplicável, probabilidades de classe. Essa nova estrutura de dados pode ser facilmente explorada em uma tabela e, no caso de problemas de classificação, a acurácia do modelo pode ser avaliada através de uma matriz de confusão.

3.5.7 Matriz de Confusão

A ferramenta apresenta o número ou a proporão de instâncias entre as classes reais e as previstas. Dessa forma com essa aplicação pode-se visualizar os seguintes

segmentos:

- Número de instâncias, apresenta os valores classificados correta e incorretamente;
- Proporções previstas, mostra a porcentagem de instâncias classificadas em relação a classe verdadeira;
- Proporção real, exibe a porcentagem dos quais foram classificados corretamente dentro de cada classe.

3.5.8 Distribuição

A ferramenta de distribuição apresenta o arranjo de valores de atributos discretos ou contínuos. Assim, mostra quantas vezes cada valor de atributo aparece nos dados. Com essa funcionalidade é possível:

- Selecionar as variáveis para exibição;
- Classificar categorias por ordem de frequência;
- Definir a largura dos intervalos;
- Ajustar escala de precisão;
- Selecionar e ajustar distribuições ao gráfico.

3.5.9 Save and Load Model

A ferramenta *Save Model* tem como objetivo salvar modelos pré treinados dentro do sistema *Orange* para que possam ser utilizados em momentos posteriores.

Por sua vez a ferramenta *Load Model* tem como objetivo realizar o carregamento de modelos que foram salvos para que possam ser reutilizados sem a necessidade de um novo treinamento.

4 RESULTADOS E DISCUSSÃO

Neste capítulo, serão apresentados os resultados dos testes realizados de acordo com a metodologia descrita anteriormente. A partir disso, serão realizadas análises críticas e uma discussão em relação ao objetivo da pesquisa. A discussão dos resultados tem como objetivo responder às perguntas e verificar a validade das hipóteses que foram formuladas.

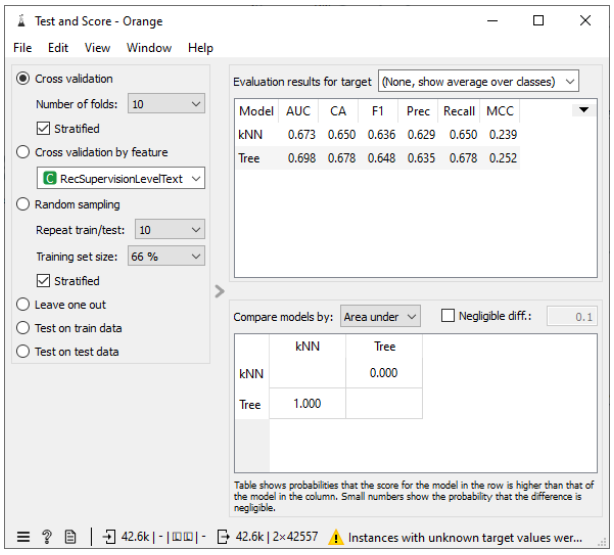
4.1 CENÁRIOS DE IMPLEMENTAÇÃO

Aqui serão mostrados os resultados e as diferenças dos dois cenários de implementação dos modelos de inteligência artificial.

4.1.1 Primeiro Cenário de Implementação

Para o primeiro cenário tem-se a utilização dos dados de maneira completa de acordo descrita na seção de metodologia. Ou seja, o objetivo é avaliar o quanto os modelos são capazes de replicar as anotações realizadas pelo modelo Compas para a métrica *ScoreText* utilizando toda a base de dados original. Com a utilização de dois modelos diferentes tem-se a aplicação de cada um em relação a base de dados. Primeiramente tem-se a análise da célula de *Test and Score* na Figura 6 que nos demonstra que a árvore de decisão tem um desempenho superior ao Knn em todas as suas métricas, sendo assim, nos leva ao direcionamento de que este é o modelo que melhor reproduz a classificação realizada pelo Compas.

Figura 6 – Test and Score



Fonte: Elaborado pelo autor (2024).

Através da Figura 6 é possível verificar que o conjunto de dados foi dividido em 10 partes, ou seja, sendo treinado em 9 partes e testado na parte restante respeitando a distribuição das classes. Além disso, os valores das métricas tem um valor razoável tanto para o *Knn* quanto para *Tree* variando entre 65% e 70%. Então o modelo *Tree* tem uma melhor métrica para no que diz respeito a capacidade de distinguir as classes e também em sensibilidade, porém, mostra que ambos os modelos têm valores baixos para a métrica de correlação entre as classes previstas e as classes reais.

A partir da análise dos valores das métricas, foi realizado a verificação dos valores das matrizes de confusão para poder-se visualizar quão fidedígno o modelo é em relação a classificação existente. Com isso, existem duas matrizes disponíveis uma para cada um dos modelos de inteligência artificial utilizados que podem ser verificados nas Figuras 7 e 8.

Figura 7 – Matriz de Confusão - KNN

		Predicted			
		High	**Low**	**Medium**	**Σ**
Actual	**High**	3181	1166	482	**4829**
	Low	1064	27099	826	**28989**
	Medium	1045	3919	3775	**8739**
	Σ	**5290**	**32184**	**5083**	**42557**

Fonte: Elaborado pelo autor (2024).

Figura 8 – Matriz de Confusão - Tree

		Predicted			
		High	**Low**	**Medium**	**Σ**
Actual	**High**	1962	2069	798	**4829**
	Low	766	26954	1269	**28989**
	Medium	868	4592	3279	**8739**
	Σ	**3596**	**33615**	**5346**	**42557**

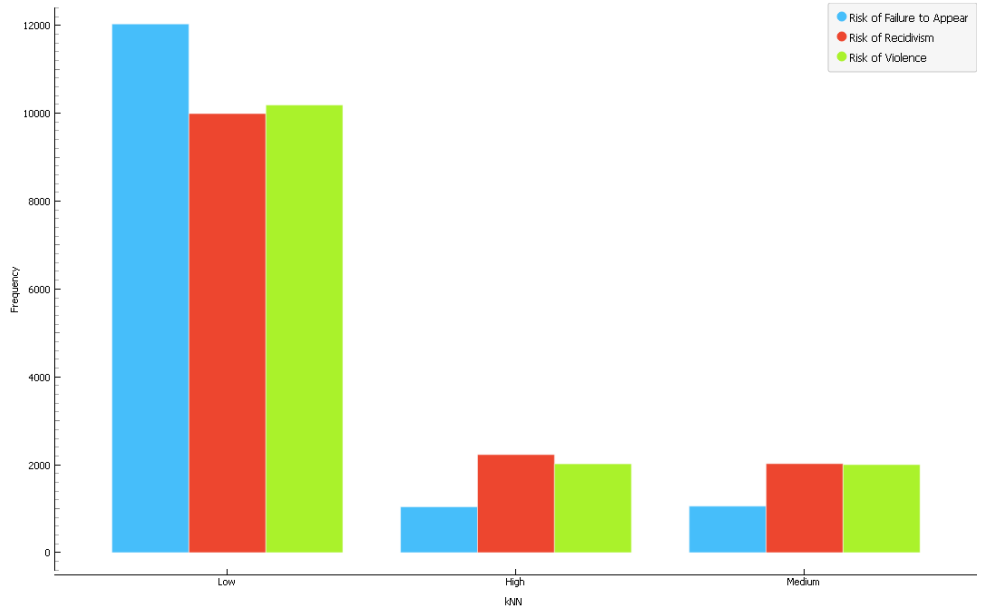
Fonte: Elaborado pelo autor (2024).

As matrizes de confusão mostram que o *Knn* e o *Tree* possuem uma grande taxa de acerto nas classificações *Low*, porém divergem mais nas outras categorias mostrando que são ótimos para classificações de baixo risco, porém possuem

dificuldades nas classificações de classes *Medium* e *High*.

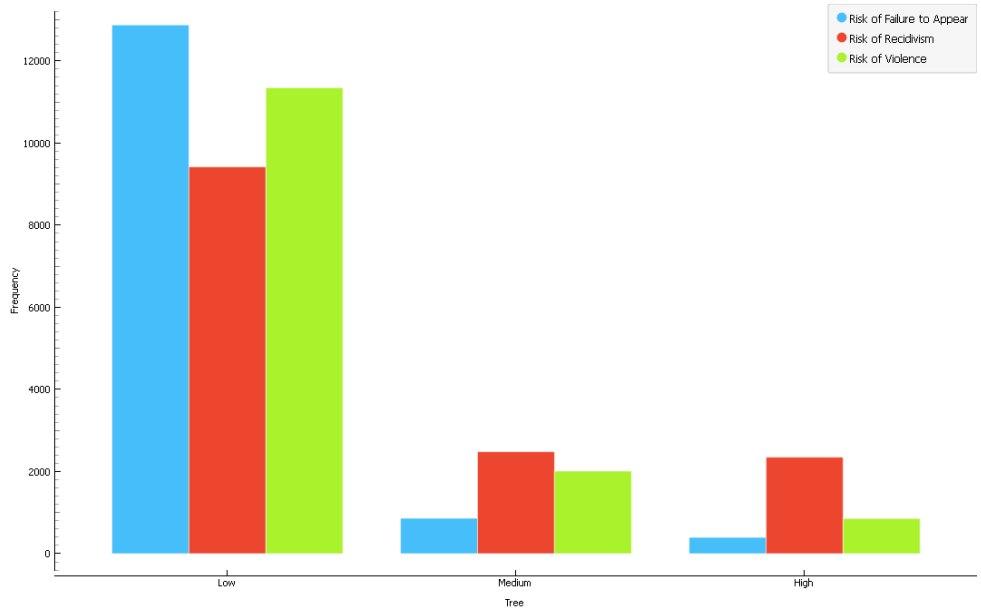
Além da verificação da matriz de confusão foi realizada também a distribuição de categorizada, sendo verificada pelo texto de display, *Low, Medium* e *High*, tanto em relação ao KNN e a árvore de decisão. Essas distribuições podem ser verificadas nas Figuras 9 e 10. Para melhorar a visualização dos valores os mesmos também podem ser visualizados na Figura 11 com o seu porcentual geral.

Figura 9 – Distribuição - KNN



Fonte: Elaborado pelo autor (2024).

Figura 10 – Distribuição - Tree



Fonte: Elaborado pelo autor (2024).

Figura 11 – Valores Distribuição

	KNN							
	Não comparecimento		Reincidêndia		Violência		Total	
Low	12.031	28,25%	9.992	23,46%	10.187	23,92%	**32.210**	**75,63%**
Medium	1.058	2,48%	2.025	4,75%	2.003	4,70%	**5.086**	**11,94%**
High	1.043	2,45%	2.231	5,24%	2.021	4,75%	**5.295**	**12,43%**
Total	**14.132**	**33,18%**	**14.248**	**33,45%**	**14.211**	**33,37%**	**42.591**	**100,00%**

	Tree							
	Não comparecimento		Reincidêndia		Violência		Total	
Low	12.876	30,23%	9.419	22,12%	11.346	26,64%	**33.641**	**78,99%**
Medium	861	2,02%	2.481	5,83%	2.010	4,72%	**5.352**	**12,57%**
High	395	0,93%	2.348	5,51%	855	2,01%	**3.598**	**8,45%**
Total	**14.132**	**33,18%**	**14.248**	**33,45%**	**14.211**	**33,37%**	**42.591**	**100,00%**

Fonte: Elaborado pelo autor (2024).

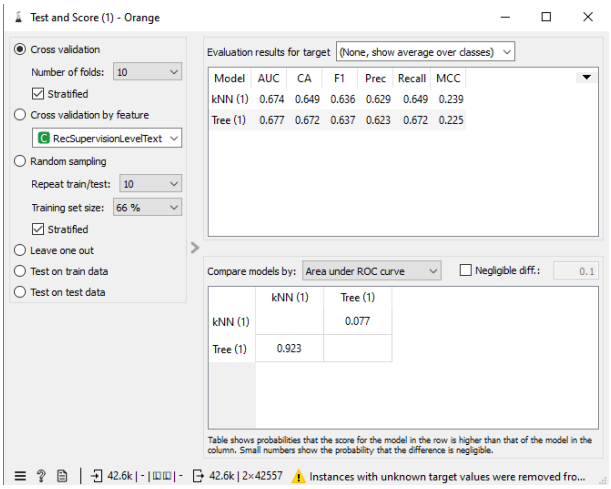
As figuras apresentam as distribuições de valores para os dois algoritmos, *Knn* e *Tree*, em três categorias de risco: *Low*, *Medium*"e *High*. Cada categoria é analisada para três critérios que são os de não comparecimento, reincidência e violência. No *Knn*, a maioria dos casos está na categoria *Low*, correspondendo a 75,63% do total. Já no *Tree*, a distribuição similarmente concentra 78,99% dos casos no nível *Low*. As categorias *Medium* e *High* possuem percentuais menores em ambos os modelos com um destaque para a categoria *High* onde possuem uma diferença de aproximadamente 8% na distribuição total.

4.1.2 Segundo Cenário Implementação

Para o segundo cenário serão mostrados os mesmos resultados do primeiro cenário, a única diferença dentro de seu fluxograma é que foi ignorada propositalmente a coluna que representa a classificação étnica racial dos detentos dentro da célula *Select Collumns*, dessa forma, os modelos serão treinados uma segunda vez não possuindo a acesso a essa classificação e com isso será observado se existe alguma variação dentro de sua própria classificação.

Primeiramente, será visualizado o resultado da célula *Test and Scores* para avaliar a pontuação das métricas presentes, essa visualização está presente na Figura 12 que vem a seguir.

Figura 12 – Test and Score - Sem Classificação Étnico Racial



Fonte: Elaborado pelo autor (2024).

Através da Figura 12 é possível verificar que o conjunto de dados continua dividido em 10 partes respeitando a distribuição de classes. O modelo *Knn* possui métricas de caracterização razoáveis igualmente ao modelo *Tree*, sendo assim as métricas estão valendo entre 62% até 67%. Entretanto o modelo *Tree* tem melhores notas em todos os quesitos menos em precisão e no coeficiente de correlação.

A partir disso também é possível verificar o resultado da matriz de confusão para os modelos propostos que são demonstrados nas Figuras 13 e 14.

Figura 13 – Matriz de Confusão - KNN - Sem Classificação Étnico Racial

		Predicted			
		High	**Low**	**Medium**	**∑**
Actual	**High**	3166	1178	485	**4829**
	Low	1055	27099	835	**28989**
	Medium	1042	3908	3789	**8739**
	∑	**5263**	**32185**	**5109**	**42557**

Fonte: Elaborado pelo autor (2024).

Figura 14 – Matriz de Confusão - Tree - Sem Classificação Étnico Racial

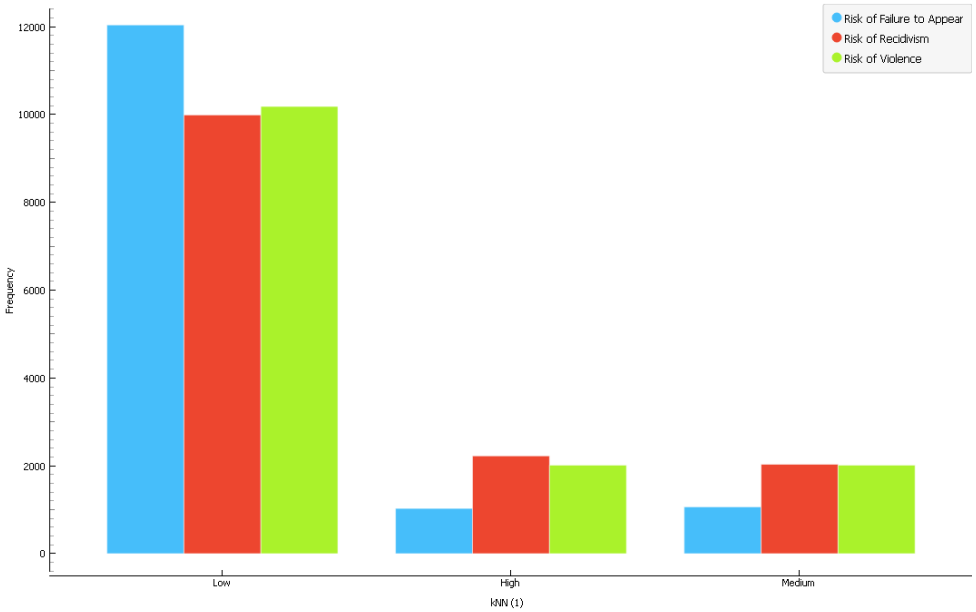
Actual \ Predicted	High	Low	Medium	Σ
High	1815	2248	766	**4829**
Low	837	27025	1127	**28989**
Medium	789	4933	3017	**8739**
Σ	**3441**	**34206**	**4910**	**42557**

Fonte: Elaborado pelo autor (2024).

As matrizes de confusão mostram que o *Knn* e o *Tree* continuam tendo um grande acerto na classificação *Low* porém baixo acerto nas classificações *Medium* e *High*. As matrizes demonstram como os dois modelos quase não sofrem variações por terem ou não terem lidado com o parâmetro étnico racial dos detentos.

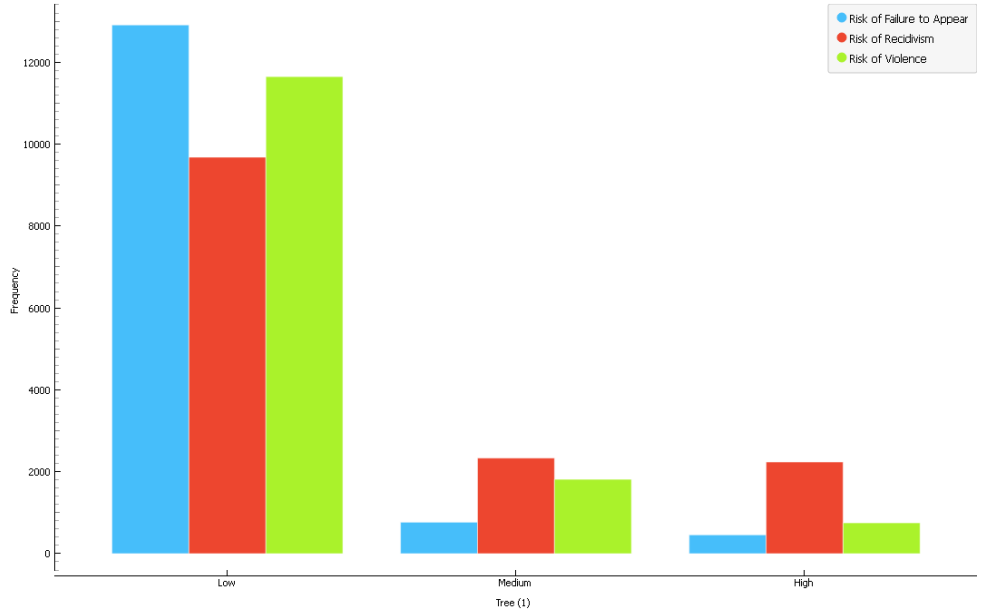
A partir desses valores é possível realizar uma distribuição de valores para que dessa forma seja possível verificar a quantidade pertencente a cada grupo dentro das classificações ja especificadas anteriormente. Com isso, de forma igual para o primeiro cenário, são mostrados nas Figuras 15, 16 e 17 a distribuição para cada um dos modelos além do seu percentual geral.

Figura 15 – Distribuição - KNN - Sem Classificação Étnico Racial



Fonte: Elaborado pelo autor (2024).

Figura 16 – Distribuição - Tree - Sem Classificação Étnico Racial



Fonte: Elaborado pelo autor (2024).

Figura 17 – Valores Distribuição - Sem Classificação Étnico Racial

	KNN							
	Não comparecimento		Reincidência		Violência		Total	
Low	12.039	28,27%	9.989	23,45%	10.183	23,91%	**32.211**	**75,63%**
Medium	1.064	2,50%	2.034	4,78%	2.014	4,73%	**5.112**	**12,00%**
High	1.029	2,42%	2.225	5,22%	2.014	4,73%	**5.268**	**12,37%**
Total	**14.132**	**33,18%**	**14.248**	**33,45%**	**14.211**	**33,37%**	42.591	**100,00%**

	Tree							
	Não comparecimento		Reincidência		Violência		Total	
Low	12.907	30,30%	9.678	22,72%	11.646	27,34%	**34.231**	**80,37%**
Medium	767	1,80%	2.333	5,48%	1.814	4,26%	**4.914**	**11,54%**
High	458	1,08%	2.237	5,25%	751	1,76%	**3.446**	**8,09%**
Total	**14.132**	**33,18%**	**14.248**	**33,45%**	**14.211**	**33,37%**	42.591	**100,00%**

Fonte: Elaborado pelo autor (2024).

Através dessas figuras é possível observar que para o modelo *Knn* ocorreu uma estabilidade dos valores totais referentes as categorizações com 75% para risco *Low* e 12% para riscos *Medium* e *High*. Entretanto, o modelo *Tree* possium um aumento de aproximadamente 1.5% na categoria *Low*, um aumento de 1% na categoria *Medium* e um declínio de 0.36% na categoria *High*.

4.1.3 Comparação dos Cenários de Implementação

Dado os dois cenários de implementação, é possível verificar por meio das Figuras 11 e 17 que existem poucas diferenças entre as classificações finais propostas. Entretanto, segundo os valores das métricas de *Test and Score* apresentados é visto que na implementação dos dados nos quais existes a classificação étnica as métricas possuem valores praticamente similares no uso do *Knn* possuindo variações de 0,01% para mais em acurácia e sensibilidade e para menos no modelo de classficiação

binária AUC. Isso sugere que para esse modelo o uso da categorização étnica não é uma variável de impacto para que exista uma alteração na avaliação das categorias propostas.

Por sua vez, o modelo *Tree* demonstra melhores resultados utilizando os dados étnicos do que quando não utiliza esses valores, tendo dessa forma a percepção que essa categoria nos dados têm um impacto nos resultados apresentados que podem ser vistos na diminuição de 0,21% na classificação binária AUC, 0,06% em acurárica 0,09% no F-1*Score*, 0,12% em precisão e 0,06% em sensibilidade uma vez que não possuem mais os dados étnicos. Além disso, também apresenta valores melhores do que os *KNN* sugerindo que consegue ter uma leitura e interpretação dos dados de maneira mais acurada.

Por sua vez nas matrizes de confusão representadas pelas Figuras 7, 13, 8 e 14, é possível verificar que as diferenças são ligerias assim como no *Test and Score* com os dois modelos tendo uma grande porcentagem de acerto na classificação da categoria *Low* e baixos acertos nas categorias *Medium* e High.

Por ultimo nas distribuições representadas pelas Figuras 11 e 17 perbece-se uma ligeira redução nos valores totais, sendo dessa forma correspondente ao que é mostrado nas matrizes de confusão.

Em resumo, os modelos apresentam desempenhos na casa de 60 a 70%, o que demonstra que possuem um desempenho moderado, mas que ainda necessitam de melhorias a serem realizadas para que consigam estar de acordo com a classificação do sistema Compas. Tendo em vista a necessidade dessas melhorias algumas das sugestões seriam a necessidade de ajuste dos parâmetros dos modelos, a utilização de outros modelos mais complexos como por exemplo Redes Neurais Artificias, *Random Forest* ou superiores e principalmente ter base de dados que seja balanceada em relação as classificações ajudando assim com que não ocorra sobreamostragem ou subamostragem.

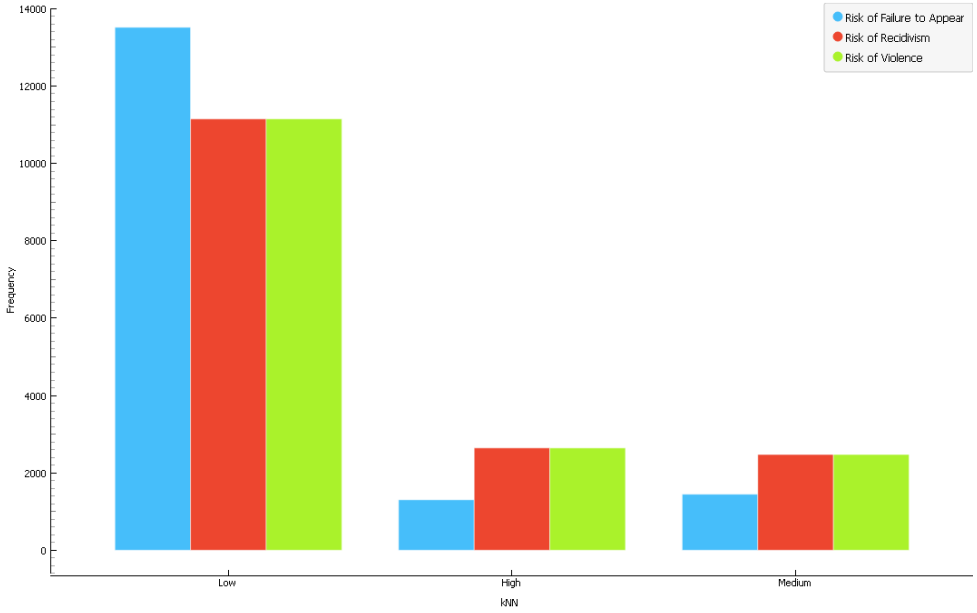
4.2 CENÁRIOS DE USO

Aqui serão mostrados os resultados e as diferenças dos dois cenários de uso dos modelos de inteligência artificial salvos.

4.2.1 Primeiro cenário de Uso - Modelo KNN

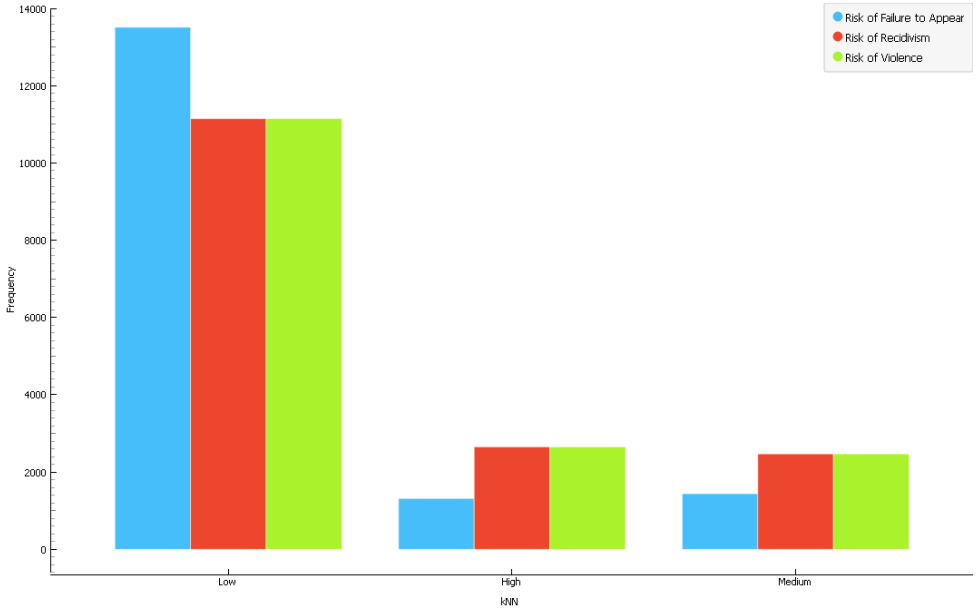
Para o modelo *Knn* foi obtido a distribuição contida nas Figuras 18 e 19, além disso, é possível verificar os valores e suas porcentagens na Figura 20 que demonstra a comparação entre os valores obtidos entre as duas classificações.

Figura 18 – Distribuição - KNN - Caucasiano



Fonte: Elaborado pelo autor (2024).

Figura 19 – Distribuição - KNN - Afro-Americano



Fonte: Elaborado pelo autor (2024).

Figura 20 – Valores Distribuição - KNN

	KNN - Caucasiano							
	Não comparecimento		Reincidência		Violência		Total	
Low	13.517	27,70%	11.149	22,85%	11.150	22,85%	**35.816**	**73,39%**
Medium	1.447	2,97%	2.474	5,07%	2.474	5,07%	**6.395**	**13,10%**
High	1.303	2,67%	2.644	5,42%	2.643	5,42%	**6.590**	**13,50%**
Total	**16.267**	**33,33%**	**16.267**	**33,33%**	**16.267**	**33,33%**	**48.801**	**100,00%**

	KNN - Afro Americano							
	Não comparecimento		Reincidência		Violência		Total	
Low	13.513	27,69%	11.147	22,84%	11.148	22,84%	**35.808**	**73,38%**
Medium	1.439	2,95%	2.469	5,06%	2.469	5,06%	**6.377**	**13,07%**
High	1.315	2,69%	2.651	5,43%	2.650	5,43%	**6.616**	**13,56%**
Total	**16.267**	**33,33%**	**16.267**	**33,33%**	**16.267**	**33,33%**	**48.801**	**100,00%**

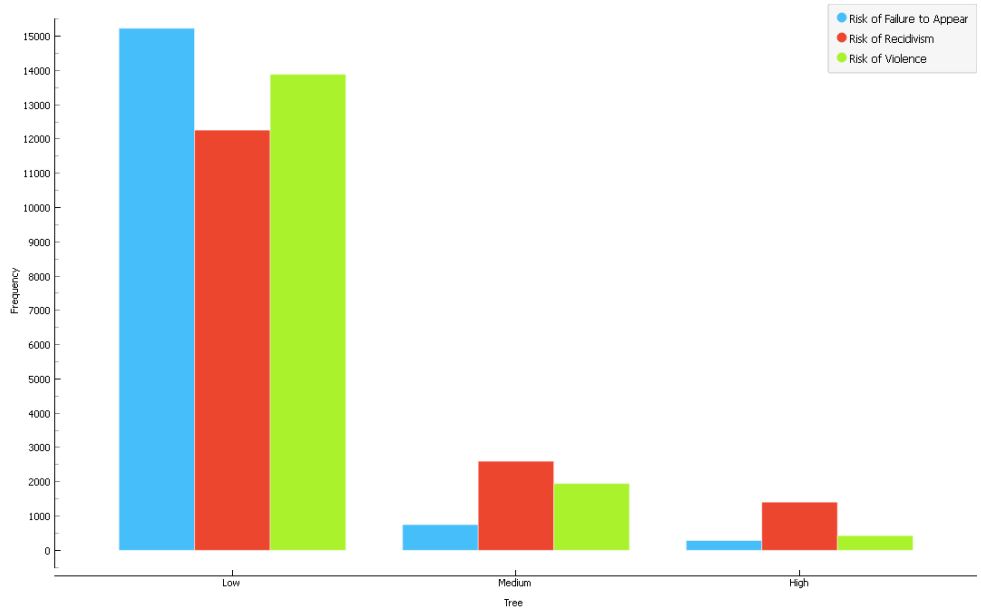
Fonte: Elaborado pelo autor (2024).

Dessa forma é possível verificar que para o modelo *Knn* a distribuição de resultados possui uma mudança que é na casa de 0,01% mostrando dessa forma que para esse modelo não existe diferença entre as categorizações étnico raciais o que pode ser interpretado como o não uso dessa categoria para realizar as classificações de risco dos detentos presentes na base de dados.

4.2.2 Segundo cenário de Uso - Modelo Tree

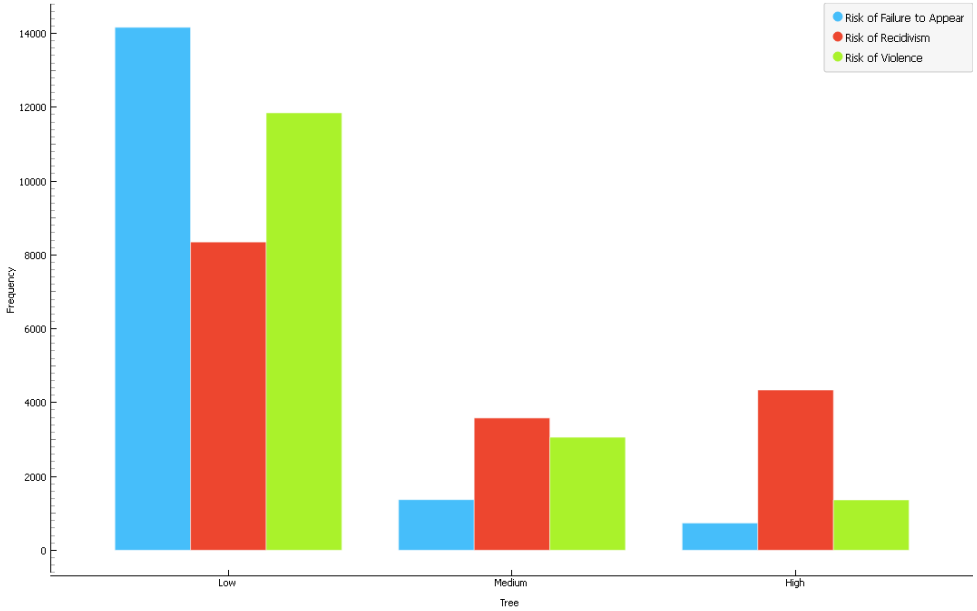
O segundo cenário de uso é feito de maneira similar ao primeiro, entretanto é utilizado o modelo *Tree* ao invés do modelo *Knn* para realizar as classificações dos detentos que são mostradas nas Figuras 21 e 22 com o resultado de avaliação do modelo em relação as duas bases de dados.

Figura 21 – Distribuição - Tree - Caucasiano



Fonte: Elaborado pelo autor (2024).

Figura 22 – Distribuição - Tree - Afro-Americano



Fonte: Elaborado pelo autor (2024).

Com isso, é verificado as porcentagens das duas avaliações com esse modelo na Figura 23 demonstrada abaixo.

Figura 23 – Valores Distribuição - Tree

	Tree - Caucasiano							
	Não comparecimento		Reincidência		Violência		Total	
Low	15.232	31,21%	12.261	25,12%	13.889	28,46%	**41.382**	**84,80%**
Medium	749	1,53%	2.600	5,33%	1.949	3,99%	**5.298**	**10,86%**
High	286	0,59%	1.406	2,88%	429	0,88%	**2.121**	**4,35%**
Total	**16.267**	**33,33%**	**16.267**	**33,33%**	**16.267**	**33,33%**	**48.801**	**100,00%**

	Tree - Afro Americano							
	Não comparecimento		Reincidência		Violência		Total	
Low	14.163	29,02%	8.347	17,10%	11.846	24,27%	**34.356**	**70,40%**
Medium	1.368	2,80%	3.581	7,34%	3.060	6,27%	**8.009**	**16,41%**
High	736	1,51%	4.339	8,89%	1.361	2,79%	**6.436**	**13,19%**
Total	**16.267**	**33,33%**	**16.267**	**33,33%**	**16.267**	**33,33%**	**48.801**	**100,00%**

Fonte: Elaborado pelo autor (2024).

O modelo *Tree* de acordo com as figuras acima demonstra uma grande variação em sua classificação perante as duas etnias propostas. Primeiramente na classificação de baixo risco, *Low*, ocorre uma diminuição de 14% quando se altera a etnia de caucasiano para afro-americano o que demonstra que o modelo possui um viés de considerar o primeiro grupo de pessoas com menos riscos para a sociedade do que o segundo. Levando em consideração que o modelo desde o início do trabalho demonstra um grande percentual de acerto para essa categoria fica claro que possui uma distinção que não deveria existir para o mesmo ser considerado imparcial.

Já nas categorias *Medium* e *High* o modelo demonstra um aumento de 5,55% e 8,84% respectivamente, demonstrando dessa maneira que considera as pessoas

afro-americanas com um maior risco de periculosidade para com a sociedade.

4.2.3 Comparação entre os Cenários de Uso

A comparação dos casos de uso é realizada considerando as diferenças nos valores gerados pelos modelos nas duas bases de dados analisadas. Na Figura 20, observa-se que os percentuais das categorias *Low, Medium* e *High* para os riscos de não comparecimento, reincidência e violência são praticamente iguais entre as etnias caucasiana e afro-americana. Isso demonstra que o modelo *Knn* não utiliza a categorização étnica como um fator determinante na avaliação dos dados, resultando em distribuições semelhantes para os dois grupos. Essa uniformidade reforça que o modelo trata as informações de maneira equitativa, sem evidências de viés relacionado à etnia.

Entretanto, o modelo *Tree* possui uma grande diferença nas classificações mesmo estando na mesma margem de 60 a 70 % que o modelo *KNN*. Na Figura 23 é possível verificar que quando o modelo analisa uma base de pessoas caucasianas classifica como 84,80% na categoria *Low*, 10,86% na categoria *Medium* e 4.35% na categoria *High* e o mesmo modelo por sua vez quando esta analisando uma base de pessoas afro-americanas possui como resultado 70,40% na categoria *Low*, 16,41% na categoria *Medium* e 13,19% na categoria *High*.

Usando esses valores para realizar uma comparação em relação ao percentual de predição de classificação do modelo demonstrado na Figura 11, onde por sua vez ocorre a análise de toda base de dados com a categorização racial presente resultando dessa forma na seguinte comparação demonstrada na Figura 24.

Figura 24 – Diferença Percentual do Caso de Uso em Relação ao Modelo Tree

	Não comparecimento		Reincidência		Violência		Total	
	Caucasiano	Afro Americano	Caucasiano	Afro Americano	Caucasiano	Afro Americano	Caucasiano	Afro Americano
Low	0,98%	-1,21%	3,01%	-5,01%	1,82%	-2,37%	5,81%	-8,59%
Medium	-0,49%	0,78%	-0,50%	1,51%	-0,73%	1,55%	-1,71%	3,85%
High	-0,34%	0,58%	-2,63%	3,38%	-1,13%	0,78%	-4,10%	4,74%

Fonte: Elaborado pelo autor (2024).

A partir da Figura 24 é possível perceber as seguintes tendências:

- Na classificação de pessoas caucasianas sempre existe uma maior quantidade de detentos classificados como *Low* em todos os seguimentos apresentados;
- Também aplicado a detentos caucasianos, estão sempre abaixo em relação ao percentual de predição do modelo nas classificações *Medium* ou *High* demonstrados na Figura 11 para o modelo *Tree*;
- Detentos categorizados como afro-americanos possuem uma menor representação na categoria *Low* para todos as classificações no modelo

Tree;

- Também aplicados a detentos afro-americanos, são sempre classificados em um porcentual maior do que o de predição da Figura 11 nas categorias *Medium* ou *High*.

Com essas observações é possível avistar que o modelo *Tree* que foi salvo à partir dos dados do sistema Compas tem um viés racial que aumenta o risco de pessoas afro-americanas em relação as pessoas caucasianas, oferecendo dessa forma um modelo que não existe equidade racial e julgamento justo para pessoas que venham a cometer o mesmo crime, mas sejam de etnias diferentes.

5 CONCLUSÕES

Este estudo teve como objetivo investigar se existe equidade racial no sistema Compas através da análise dos dados do sistema utilizando modelos de inteligência artificial para fazer a classificação e averiguação dos resultados do próprio sistema. Uma vez com a classificação dos dados foi realizado a separação dos dados utilizados para teste do modelo com 70% dos dados disponíveis onde foram verificados 2 modelos de implementação para verificar a primeira hipótese da diferenciação de tratamento caso os dados utilizassem ou não a categorização étnico racial para realizar a classificação.

A partir desses cenários foi possível perceber que os valores das métricas nas duas ocasiões são parecidas para os dois modelos de inteligência artificial utilizados. O modelo *KNN* quase não apresenta variações quando existe a classificação por etnia diferentemente do modelo *Tree* que mesmo mantendo uma métrica com pontuação moderada apresenta variações nas distribuições indicando que a categoria étnica é uma categoria de importância para o modelo.

Ao decorrer da pesquisa foi possível observar que quando se utiliza esses modelos para realizar uma nova classificação o modelo *Knn* continua apresentando uma não diferenciação entre os valores com ou sem a informação étnica, diferentemente do modelo *Tree*, que deixa claro que possui uma enviesamento racial ao demonstrar em suas análises que possui uma probabilidade maior em classificar pessoas caucasianas com risco menores nas três categorias de estudo diferentemente de pessoas que sejam categorizadas como afro americanas, que por sua vez tem a probabilidade de serem consideradas nas categorias de alto risco a sociedade.

Embora os modelos *Knn* e *Tree* utilizados no estudo tenham taxas de acerto de aproximadamente 70% é importante destacar que nenhum dos modelos conseguiu de fato refletir perfeitamente a classificação do Compas. Além disso, também observou-se que o modelo *Tree* tomou como critério preditivo a variável étnica que é um comprtamento que reforça vieses indesejados e compromete a imparcilidade necessária para as classificações.

Partindo do princípio de que esses sistemas têm como objetivo julgar a periculosidade de uma pessoa para a sociedade e decidir se está apta a esperar seu julgamento em liberdade ou em prisão, é de suma importância levar em consideração de que é possível existir um viés racial na base de dados utilizada e modelo de funcionamento do sistema. que pode ser demonstrado nos seguintes pontos:

- Dados históricos de desigualdade social e discriminação estrutural;
- Forma que o sistema foi projetado, utilizando uma base de dados com um

desbalanceamento nas classificações e julgamentos.

Apesar das limitações apresentadas nas métricas pela simplicidade dos modelos propostos, a pesquisa contribui significativamente para o estudo dos impactos da inteligência artificial na reprodução de classificações humanas. Uma vez que a mesma é aplicada para realizar avaliações de grupos minoritários na sociedade que são historicamente marginalizados, o sugerido seria que para trabalhos futuros o sistema seja analisado com modelos de inteligência artificial que possuam uma complexidade maior em relação aos que foram utilizados nessa pesquisa, como por exemplo Redes Neurais Artificias e *Random Forest*.

Dessa forma, os modelos teriam uma maior capacidade de verificar e analisar as nuances entre as classes, realizando dessa maneira uma classificação mais acurada para determinar exatamente se existe ou não a presença de um viés racial nos dados analisados.

REFERÊNCIAS

ALMEIDA, S. L. de. **O que é racismo estrutural?** Belo Horizonte: Letramento, 2018.

ALPAYDIN, E. **Introduction to machine learning**. Cambridge, Massachusetts: MIT Press, 2020.

ANGWIN, J. *et al*. How we analyzed the compas recidivism algorithm. **ProPublica**, 2016. Disponível em: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-. Acesso em: 2024 oct.10.

ANGWIN, J. *et al*. Machine bias. In: MARTIN, K. (Ed.). **Ethics of data and analytics**. New York: Auerbach Publications, 2022. p. 254–264.

BISHOP, C. M.; NASRABADI, N. M. **Pattern recognition and machine learning**. New York: Springer, 2006. v. 4.

BOTTOU, L. Stochastic gradient descent tricks. In: MONTAVON, G.; ORR, G.; MÜLLER, K.-R. (Ed.). **Neural Networks**: Tricks of the trade. 2nd. ed. New York: Springer, 2012. p. 421–436.

BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: FRIEDLER, S. A.; WILSON, C. (Ed.). **Proceedings of the 1st Conference on Fairness, Accountability and Transparency**. New York: PMLR, 2018. (Proceedings of Machine Learning Research, v. 81), p. 77–91. Disponível em: https://proceedings.mlr.press/v81/buolamwini18a.html.

CALISKAN, A.; BRYSON, J. J.; NARAYANAN, A. Semantics derived automatically from language corpora contain human-like biases. **Science**, American Association for the Advancement of Science, v. 356, n. 6334, p. 183–186, 2017.

CHOULDECHOVA, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. **Big Data**, v. 5, n. 2, p. 153–163, 2017. PMID: 28632438. Disponível em: https://doi.org/10.1089/big.2016.0047.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967.

DASGUPTA, S.; HSU, D. Hierarchical sampling for active learning. In: **Proceedings of the 25th International Conference on Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2008. (ICML '08), p. 208–215. ISBN 9781605582054. Disponível em: https://doi.org/10.1145/1390156.1390183.

FLORES, A. W.; BECHTEL, K.; LOWENKAMP, C. T. False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. **Fed. Probation**, HeinOnline, v. 80, p. 38, 2016.

GARETH, J. *et al*. **An introduction to statistical learning: with applications in R**. 1. ed. New York: Spinger, 2013. Disponível em: https://doi.org/10.1007/978-1-4614-7138-7.

GARVIE, C. **The perpetual line-up: Unregulated police face recognition in America**. Georgetown, DC: Georgetown Law, Center on Privacy & Technology, 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: Data mining, inference and prediction. New York: [s.n.], 2009. v. 2. Disponível em: https://doi.org/10.1007/978-0-387-84858-7.

JACOBS, A. S. *et al*. Ai/ml and network security: The emperor has no clothes. In: **Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security**. New York, NY, USA: Association for Computing Machinery, 2022. (CCS '22).

KHADEMI, A.; HONAVAR, V. Algorithmic bias in recidivism prediction: A causal perspective (student abstract). **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 34, n. 10, p. 13839–13840, Apr. 2020. Disponível em: https://ojs.aaai.org/index.php/AAAI/article/view/7192.

LOH, W.-Y. Classification and regression trees. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 1, p. 14 – 23, 01 2011.

JOHN LU, Z. Q. The elements of statistical learning: Data mining, inference, and prediction. **Journal of the Royal Statistical Society Series A: Statistics in Society**, v. 173, n. 3, p. 693–694, 06 2010. ISSN 0964-1998. Disponível em: https://doi.org/10.1111/j.1467-985X.2010.00646_6.x.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. **Advances in neural information processing systems**, v. 30, 2017.

MATOS, P. F. *et al*. Relatório técnico "métricas de avaliaçao". **Universidade Federal de Sao Carlos**, São Carlos, 2009.

MITCHELL, M. *et al*. Model cards for model reporting. In: **Proceedings of the Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 2019. p. 220–229. ISBN 9781450361255. Disponível em: https://doi.org/10.1145/3287560.3287596.

MITTELSTADT, B.; RUSSELL, C.; WACHTER, S. Explaining explanations in ai. In: **Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency**. New York, NY, USA: Association for Computing Machinery, 2019. p. 279–288. ISBN 9781450361255. Disponível em: https://doi.org/10.1145/3287560.3287574.

Northpointe Inc. **COMPAS Risk Scales: Practitioner's Guide**. [S.l.], 2015. Disponível em: https://archive.epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI-FOIA-201600805-COMPASPractionerGuide.pdf.

NORVIG, P. R.; INTELLIGENCE, S. A. A modern approach. **Prentice Hall Upper Saddle River, NJ, USA: Rani, M., Nayak, R., & Vyas, OP (2015). An ontology-based adaptive personalized e-learning system, assisted by software agents on cloud storage. Knowledge-Based Systems**, v. 90, p. 33–48, 2002.

OBERMEYER, Z. *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. **Science**, American Association for the Advancement of Science, v. 366, n. 6464, p. 447–453, 2019.

PETERSON, L. K-nearest neighbor. **Scholarpedia**, v. 4, p. 1883, 01 2009.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?": Explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining**. New York, NY, USA: Association for Computing Machinery, 2016. (KDD '16), p. 1135–1144. ISBN 9781450342322. Disponível em: https://doi.org/10.1145/2939672.2939778.

SANCHES, M. K. **Aprendizado de máquina semi-supervisionado: proposta de um algoritmo para rotular exemplos a partir de poucos exemplos rotulados**. Tese (Doutorado) — Universidade de São Paulo, 2003.

SHINDE, P. P.; SHAH, S. A review of machine learning and deep learning applications. In: **2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)**. Pune, India: Institute of Electrical and Electronics Engineers, 2018. p. 1–6.

SILVA, T. Racismo algorítmico em plataformas digitais: microagressões e discriminação em código. **Comunidades, algoritmos e ativismos digitais: olhares afrodiaspóricos**, Editora Literarua São Paulo, p. 121–135, 2020.

SKEEM, J.; ENO LOUDEN, J. Assessment of evidence on the quality of the correctional offender management profiling for alternative sanctions (compas). **Unpublished report prepared for the California Department of Corrections and Rehabilitation.**, 2007. Disponível em: https://bpb-us-e2.wpmucdn.com/sites.uci.edu/dist/0/1149/files/2013/06/CDCR-Skeem-EnoLouden-COMPASeval-SECONDREVISION-final-Dec-28-07.pdf.

SUNDARARAJAN, M.; TALY, A.; YAN, Q. Axiomatic attribution for deep networks. In: PRECUP, D.; TEH, Y. W. (Ed.). **Proceedings of the 34th International Conference on Machine Learning**. PMLR, 2017. (Proceedings of Machine Learning Research, v. 70), p. 3319–3328. Disponível em: https://proceedings.mlr.press/v70/sundararajan17a.html.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning**: An introduction. Cambridge, Massachusetts: MIT press, 2018.