INSTITUTO POLITÉCNICO DE BRAGANÇA
Escola Superior de Tecnologia e Gestão

Exploração de indicadores para identificar víeis em diferentes conjuntos de dados

Rozeana Pereira - a42540

Trabalho realizado sob a orientação de

Professora Doutora Inês Sena

Professora Doutora Ana Isabel Pereira

Mestrado em Engenharia Eletrotécnica e de Computadores

2024-2025

Exploração de indicadores para identificar víeis em diferentes conjuntos de dados

Relatório da UC de Projeto

Mestrado em Engenharia Eletrotécnica e de Computadores

Escola Superior de Tecnologia e Gestão

Rozeana Pereira - a42540

2024-2025

A Escola Superior de Tecnologia e de Gestão não se responsabiliza pelas opiniões expressas neste relatório.

Declaro que o trabalho descrito neste relatório é da minha autoria e é da minha vontade que o mesmo seja submetido a avaliação.

Rozeana Pereira - a42540

Resumo

Os conjuntos de dados frequentemente contêm vieses que prejudicam determinados grupos. Geralmente, é muito difícil eliminar totalmente esses vieses. No entanto, deve-se sempre tentar minimizá-los. O presente estudo tem como principal objetivo identificar a presença de viés em diferentes conjuntos de dados. Para validar a abordagem, foram realizados testes práticos em Python, com diferentes conjuntos de teste. Foram utilizados dois conjuntos de dados de contextos distintos. Aplicaram-se os testes estatísticos Qui-Quadrado (χ^2) e ANOVA para identificar possíveis vieses, interpretando os resultados em função do nível de significância de 1% e dos testes de hipóteses. Além disso, aplicou-se a técnica de balanceamento de dados Random Oversampling, uma vez que os resultados continham vieses. A combinação dos testes estatísticos e da técnica de balanceamento permitiu analisar os resultados dos testes estatísticos antes e depois do balanceamento, demonstrando a eficácia das abordagens na deteção e mitigação de vieses, contribuindo para sistemas mais justos.

Palavras-chave: ANOVA, Balanceamento de dados, Inteligência Artificial, Qui-Quadrado, Viés algorítmico.

Abstract

Datasets often contain biases that disadvantage certain groups. It is generally very difficult to eliminate these biases. However, one should always try to minimize them. The main objective of this study is to identify the presence of bias in different datasets. To validate the approach, practical tests were performed in Python with different test sets. Two datasets from distinct contexts were used. Chi-square (χ^2) and ANOVA statistical tests were applied to identify possible biases, interpreting the results based on a significance level of 1% and hypothesis testing. In addition, the Random Oversampling data balancing technique was applied, as the results contained biases. The combination of statistical tests and the balancing technique allowed for the analysis of the statistical test results before and after balancing, demonstrating the effectiveness of the approaches in detecting and mitigating biases, contributing to fairer systems.

Keywords: Algorithmic Bias, ANOVA, Artificial Intelligence, Chi-Square, Data Balancing.

Isenção Responsabilidade

Neste relatório foi usado o ChatGPT para apoiar a redação de texto e revisão gramatical. Apesar do suporte, todo o conteúdo foi analisado, adaptado e validado pelo(a) autor(a), que assume integralmente a responsabilidade pela precisão, originalidade e veracidade das informações apresentadas.

Conteúdo

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A Inteligência Artificial (IA) vem-se tornando cada vez mais presente em diversos setores da economia mundial, abrangendo áreas como a saúde, a segurança, o recrutamento e o setor financeiro. As organizações procuram incessantemente qualidade, produtividade e redução de custos [16].

A crescente utilização de IA tem trazido avanços importantes para diversos setores. No entanto, esta modernização também levanta preocupações relevantes, como a presença de viés algorítmico, uma vez que os sistemas de IA dependem da qualidade e representatividade dos dados utilizados para o seu treino. Contudo, esses dados frequentemente carregam consigo preconceitos e desigualdades históricas, sociais e culturais, que acabam por ser reproduzidos ou até amplificados pelos algoritmos. Assim, o viés nos dados é uma das principais causas de discriminação em sistemas de IA, afetando desde a imparcialidade das decisões automatizadas até à confiabilidade das aplicações tecnológicas [61].

O viés algorítmico consiste numa tendência ou distorção nos dados que pode fazer com que as conclusões se afastem da realidade. Pode ocorrer de diversas formas: por normas organizacionais que favoreçam um grupo em detrimento de outro, viés inconsciente ou consciente dos programadores, uso de dados históricos preconceituosos, seleção de variáveis associadas a características discriminatórias e a falta de dados de grupos minoritários durante o treino do modelo. Sendo assim, torna-se fundamental compreender e assegurar que o processo de decisão não prejudique um grupo específico de pessoas ou, mesmo, um

indivíduo. Para tal, é necessário recorrer a técnicas que permitam detetar os vieses e, sempre que possível, corrigi-los.

Há desafios significativos na criação de sistemas de IA éticos e livres de vieses, mas os princípios éticos podem servir de base para que as organizações integrem a ética no desenvolvimento destes sistemas. A ética desempenha um papel fundamental, estabelecendo princípios e requisitos a seguir. Elementos como o respeito pela autonomia humana, privacidade e justiça devem ser incorporados, juntamente com transparência e responsabilidade, de modo a garantir a compreensibilidade e o controlo destes sistemas [16].

O propósito deste trabalho é explorar indicadores para identificar a presença de vieses em diferentes conjuntos de dados. O primeiro conjunto de dados refere-se a abordagens policiais na cidade de Minneapolis (EUA), utilizado para verificar se existiam diferenças significativas entre grupos raciais. O segundo conjunto aborda o desempenho académico no ensino a distância, onde foram analisadas variáveis como o tempo de ecrã, o sono e o nível de stress dos estudantes, com o objetivo de perceber se influenciam o desempenho académico. Para identificar a presença de viés, aplicaram-se os testes estatísticos Qui-Quadrado (χ^2) e Análise de Variância (ANOVA), que permitem verificar se há relações entre variáveis e se as diferenças observadas são estatisticamente relevantes. A análise dos resultados permitiu identificar a presença de vieses nos dados. Para validar o desempenho das técnicas aplicadas, aplicou-se a técnica *Random Oversampling* para equilibrar as classes e reduzir esses vieses, permitindo comparar os resultados antes e depois do balanceamento. A comparação mostrou que o *Random Oversampling* contribuiu para reduzir os vieses presentes nos dados, tornando os resultados mais equilibrados e consistentes. Assim, este trabalho demonstra que o uso combinado de testes estatísticos e técnicas de reamostragem pode ser útil para detetar e reduzir vieses em diferentes conjuntos de dados.

1.1 Objetivos

O objetivo principal deste trabalho é explorar indicadores para identificar viés em diferentes conjuntos de dados, para alcança-lo foram estabelecidos objetivos secundários, que incluem:

- Investigar os conceitos teóricos e os estudos já realizados sobre o viés, bem como os seus impactos e as diferentes estratégias utilizadas para a sua deteção.
- Selecionar e preparar conjuntos de dados adequados para a análise, garantindo que apresentem qualidade suficiente para o estudo.
- Aplicar testes estatísticos, como o Qui-quadrado e a ANOVA, com o intuito de identificar relações significativas entre variáveis e possíveis evidências de viés, bem como recorrer à técnica de *Random Oversampling* para efectuar o balanceamento dos dados e validar o desempenho dos testes estatísticos.
- Analisar e discutir os resultados obtidos, comparar os resultados antes e depois da aplicação da técnica de balanceamento dos dados.

1.2 Estrutura do Documento

Para os objetivos propostos neste trabalho, o documento foi estruturado em seis capítulos:

- **Capítulo 1: Introdução -** Apresenta o contexto geral do tema, destacando a relevância da IA e as preocupações associadas ao viés algorítmico. Além disso, explica o propósito do trabalho, os objetivos e a sua estrutura.
- **Capítulo 2: Revisão de Literatura -** Inclui estudos anteriores relacionados com o viés algorítmico, com exemplos práticos da aplicação dos testes Qui-Quadrado e ANOVA em contextos de análise de comportamento, bem como o uso da técnica *SMOTE* e *Random Oversampling* na classificação de dados desequilibrados.

- **Capítulo 3: Conceitos Teóricos -** Descreve os fundamentos teóricos, abordando temas como Inteligência Artificial, *Machine Learning*, tipos de viés, (Algoritmo, Humano) testes estatísticos (Qui-Quadrado e ANOVA) e técnicas de balanceamento de dados (*Oversampling, Undersampling* e métodos híbridos).

- **Capítulo 4: Metodologia -** Apresenta e explica detalhadamente os dois conjuntos de dados utilizados no estudo (operações policiais em Minneapolis e desempenho académico em ensino à distância) incluindo estatísticas descritivas e o procedimento seguido para identificar possíveis vieses.

- **Capítulo 5: Resultados e Discussão -** Demonstra e interpreta os resultados obtidos com os testes estatísticos, aplicados aos dois conjuntos de dados e gráficos, incluindo a comparação antes e depois da aplicação da técnica de balanceamento. Além disso, são analisados e comparados os métodos utilizados, avaliando a eficácia das técnicas estatísticas e de balanceamento.

- **Capítulo 6: Conclusões e Trabalhos Futuros -** Resume as principais conclusões alcançadas e propõe possíveis prolongamentos do trabalho.

Capítulo 2

Revisão da literatura

Este capítulo apresenta uma revisão detalhada da literatura que fundamenta a análise de indicadores para identificar vieses em diferentes conjuntos de dados, especialmente em contextos de abordagens policiais e ensino à distância. Além disso, descrevem-se as técnicas para detetar e equilibrar classes e reduzir esses vieses. Entre os estudos que abordam este tipo de análise, destaca-se um estudo conduzido por Onookome-Okome et al. que analisou mais de 170 mil abordagens policiais na cidade de Minneapolis a partir de dados abertos do Departamento de Polícia. O estudo identificou padrões significativos de abordagens policiais, focando em inspeções de pessoas e veículos segundo características raciais. Os resultados evidenciaram que indivíduos negros e indígenas têm maior probabilidade de serem submetidos a inspeções comparativamente a pessoas brancas. Mesmo controlando variáveis como horário da abordagem, motivo e tipo de infração, o viés racial manteve-se estatisticamente significativo, apontando para uma discriminação racial nos processos de abordagem policial [55].

No âmbito da educação à distância, um estudo investigou os fatores que influenciam o desempenho académico de estudantes em ambientes virtuais de ensino online. Entre as variáveis estudadas, apoio familiar, influência dos colegas, gestão financeira e ambiente de estudo/ensino, apenas o apoio familiar demonstrou influência estatisticamente significativa no desempenho académico. Esta constatação ressalta o papel fundamental do envolvimento da família para o sucesso dos estudantes em contextos remotos, onde

a ausência de acompanhamento presencial pode intensificar disparidades socioeducativas [46]. Por outro lado, El Refae et al indicaram que o desempenho académico no ensino à distância não sofreu impactos negativos, pelo contrário, verificou-se uma melhoria significativa nos resultados durante a implementação deste formato, evidenciando o potencial do ensino à distância para ampliar o acesso e melhorar o desempenho [26].

Tendo em conta que os vieses podem manifestar-se em diferentes domínios, incluindo segurança pública e educação, é relevante considerar o seu impacto em sistemas automatizados de decisão.

Nesse sentido Simonetta et al destacam que sistemas automatizados de decisão podem produzir impactos discriminatórios quando treinados com conjuntos de dados desequilibrados, reforçando vieses existentes. Assim, a análise prévia das características do conjunto de dados é essencial para identificar antecipadamente possíveis distorções e garantir decisões mais justas, especialmente em contextos sensíveis como abordagens policiais e educação [69].

Diversos estudos aplicaram técnicas estatísticas para identificar e corrigir esses vieses. O Reis et al utilizaram o teste Qui-quadrado para verificar associação entre variáveis, analisando a influência do sexo em decisões de risco, mostrando que o género pode afetar escolhas específicas mesmo diante de tendências gerais semelhantes entre grupos [58].

Já, o Moore et al aplicaram ANOVA para comparar médias entre grupos, medindo o efeito de estímulos monetários na predisposição psicológica de pedir ajuda [52].

No campo académico, estudos sobre o ensino de programação aplicaram a técnica *SMOTE* para corrigir desequilíbrios na distribuição de dificuldades das questões, permitindo o treino de classificadores mais equilibrados [24].

Além disso, Lobo et Martins avaliaram cinco algoritmos de aprendizagem automática para analisar sentimentos em manchetes de notícias sobre a Politec. Foram aplicadas técnicas de balanceamento de classes, como *Random Oversampling*, *SMOTE* e *SMOTE + Tomek Links*, para equilibrar os dados. Estas abordagens melhoraram a classificação de notícias positivas e negativas, sendo o *Random Oversampling* particularmente eficaz, ao permitir treinar modelos com distribuição mais uniforme e aumentar a precisão na

classificação [44].

Capítulo 3

Conceitos Teóricos

Neste capítulo, serão apresentados os principais conceitos, teorias e investigações relacionadas com os vieses algorítmicos na Inteligência Artificial (IA), com o objetivo de fundamentar a compreensão do tema abordado nesta dissertação.

3.1 Inteligência Artificial

A IA é uma área da computação dedicada ao desenvolvimento de algoritmos e sistemas que realizam tarefas que exigem inteligência humana, como comunicação em linguagem natural e reconhecimento de imagens. Entre os exemplos mais conhecidos do uso da IA, encontra-se a capacidade de comunicar na nossa linguagem, como os assistentes pessoais dos nossos telemóveis, ou de interpretar o mundo, como no reconhecimento de imagens realizado pelos carros autónomos [31].

Além disso, a IA permite que os sistemas técnicos identifiquem o contexto, resolvam problemas e ajam no sentido de alcançar um objetivo específico [29], podendo operar de forma autónoma, aprendendo e melhorando continuamente a partir de novas informações e experiências, oferecendo soluções avançadas para desafios do dia a dia, seja em negócios, saúde, transportes ou na vida quotidiana [21].

Existem métodos específicos dentro do amplo campo da IA que contribuem para a automatização de tarefas realizadas por humanos, entre os quais se destacam [42].

Machine learning (ML): É um método de análise de dados que ocorre de forma automatizada, utilizando um conjunto de algoritmos que extraem informação com base em dados [74].

Deep Learning (DL): Recorre a redes neuronais artificiais, inspiradas na estrutura do cérebro humano, para a identificação do contexto. Estas redes possuem várias camadas, responsáveis por processar dados em diferentes níveis de abstração. Desta forma, o DL permite executar tarefas como o reconhecimento de imagens, assistentes digitais, deteção de fraude de cartão de crédito [6].

Grande parte dos sistemas de IA são seguros e benéficos, proporcionando maior eficiência e oferecendo soluções para problemas complexos [17]. A IA trouxe uma série de benefícios para a sociedade, entre os quais podem-se destacar:

- **Automação e eficiência:** Reduzir a exposição das pessoas a situações e atividades de risco e diminuir a necessidade de realizar tarefas cansativas, repetitivas e monótonas, libertando os seres humanos para atividades mais agradáveis e desafiadoras [20].

- **Melhoria na saúde:** Avanços significativos na medicina, com a introdução de processos inovadores que ajudaram o trabalho de médicos e enfermeiros e, ao mesmo tempo, beneficiaram diretamente os pacientes, permitindo diagnósticos mais precisos e tratamentos personalizados [8, 20].

- **Avanços económicos:** A IA na economia tem o potencial de impulsionar um crescimento económico significativo. Ao simplificar processos, reduzir custos e promover a inovação, a IA pode atuar como um catalisador para ganhos de produtividade e eficiência económica em geral [60].

- **Segurança e transparência:** Tem o potencial de desempenhar um papel crucial na fiscalização de governos e empresas, promovendo maior transparência e reduzindo fraudes [20].

- **Serviços (públicos e privados):** É amplamente reconhecido que a IA tem um

grande potencial para melhorar os serviços públicos, permitindo que as empresas ofereçam soluções mais personalizadas, de acordo com as necessidades específicas dos consumidores, além de aumentar a qualidade e consistência dos serviços prestados e melhorar a conceção e a implementação de medidas políticas [20, 60].

Embora a maioria das aplicações da IA apresente poucos ou nenhuns riscos, algumas utilizações podem ser prejudiciais. Dentre esses riscos, destacam-se [17]:

- **Perda de empregos:** Muitas funções podem ser substituídas por máquinas, o que pode gerar desemprego em massa [20]. Para além disso, a substituição das interações humanas por sistemas automatizados, em setores como a saúde, a educação e o atendimento ao cliente, pode originar uma desumanização dos serviços, diminuindo a empatia, o acolhimento, o toque humano e a capacidade de compreender as necessidades individuais das pessoas, elementos fundamentais nestas áreas [50].

- **Desigualdade económica:** A IA tem o potencial de contribuir para a desigualdade económica, ao beneficiar de forma desproporcionada indivíduos e corporações ricas [48]. No entanto, empresas e países que investirem mais em IA terão vantagens competitivas [20].

- **Privacidade e segurança de dados:** À medida que os sistemas de IA se tornam mais avançados, os riscos relacionados com a segurança e a privacidade aumentam. Falhas no funcionamento e na comunicação entre diferentes partes do sistema, ataques adversarias e problemas técnicos podem comprometer o desempenho da IA e colocar em risco os utilizadores e a sociedade, violando a privacidade e provocando o vazamento de informações sensíveis [20, 72].

- **Falta de transparência:** A falta de transparência no funcionamento dos sistemas de IA gera processos pouco claros, limita a possibilidade de verificação, sobretudo quando desenvolvidos por fornecedores privados [72]. Além disso, algoritmos complexos podem tomar decisões que afetam vidas sem que as pessoas compreendam como essas escolhas são feitas [20].

- **Viés e discriminação:** Os sistemas de IA dependem das informações com que lhes são fornecidas. Muitos conjuntos de dados incorporam viés, o que pode resultar em decisões inconscientes. Casos de discriminação em processos de recrutamento, concessão de crédito e policiamento. [5].

- **Impacto ambiental:** O treino de modelos avançados de IA consome quantidades elevadas de energia, contribuindo para emissões de carbono e utilização excessiva de recursos naturais. À medida que cresce a procura por sistemas de IA, o impacto ambiental torna-se uma preocupação crítica [38].

- **Dependência tecnológica:** A dependência excessiva de sistemas inteligentes pode comprometer a capacidade humana de resolver problemas e tomar decisões, especialmente em situações críticas. Quando a IA falha ou se encontra indisponível, a ausência de competências humanas adequadas pode ter consequências graves [78].

O impacto e os desafios da IA na sociedade são questões que ganham ainda mais relevância à medida que a adoção dessa tecnologia cresce cada vez mais no nosso quotidiano. Vale a pena notar que a influência da IA já existe há anos [53]. A Figura 3.1 apresenta a evolução do interesse da sociedade em aplicações de IA, como o ChatGPT.

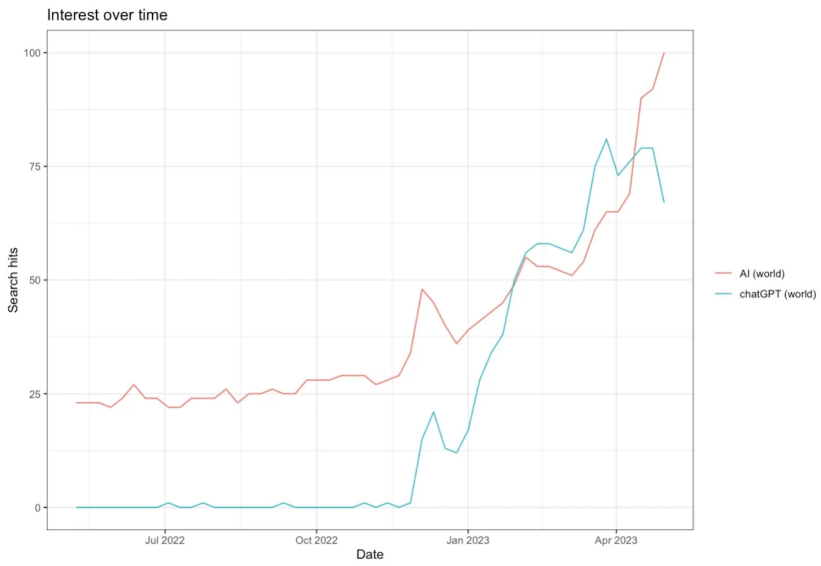

Figura 3.1: Dados de Pesquisa do Google para AI e ChatGPT ao longo do tempo [65].

O gráfico presente na Figura 3.1, demonstra o sucesso surpreendente de aplicações como o ChatGPT, que ultrapassou gigantes digitais como o TikTok ou o Instagram em termos de utilização no seu primeiro mês, é um reflexo claro de como a sociedade está cada vez mais preparada para adotar novas tecnologias [54]. Além disso, é possível notar uma grande explosão do interesse nos termos IA e ChatGPT no último mês de novembro, segundo os dados recolhidos de pesquisa da Google quando o GPT 3 foi lançado ao público. Desde então, as pesquisas no Google de ambos os termos têm crescido rapidamente, com outro pico em ChatGPT em março de 2023 [65].

3.1.1 *Machine Learning*

O Machine Learning (ML) é um subcampo da IA que usa algoritmos treinados em conjuntos de dados para criar modelos capazes de realizar tarefas que, de outra forma, só seriam possíveis para humanos, como categorizar imagens, analisar dados ou prever flutuações de preços [15].

O ML começa com a recolha de dados, números, fotos ou texto, como transações bancárias, fotos de pessoas ou até mesmo itens de padaria, registos de reparo, dados de séries temporais de sensores ou relatórios de vendas. Esses dados precisam ser preparados para que o modelo de ML possa aprender e fazer previsões precisas [10].

O ML depende da qualidade dos dados utilizados no treino. Se esses dados não representarem a realidade, o modelo irá fornecer vieses nos resultados. Dados fiáveis e variados são fundamentais para que o modelo seja eficaz. No entanto, a qualidade dos dados é crucial, pois a IA reproduzirá o conhecimento que está impregnado neles [47].

O ML auxilia as empresas em funções essenciais, como a deteção de fraudes, identificação de ameaças à segurança, personalização e recomendações, atendimento automatizado ao cliente através de chatbots, transcrição e tradução, análise de dados, entre outras. Esta tecnologia está também a impulsionar inovações promissoras para o futuro, como veículos autónomos, drones e aviões, realidade aumentada e virtual, bem como a robótica. Os algoritmos de ML podem ser classificados em diversas categorias, conforme o objetivo e

a natureza do aprendizado [57]. As principais categorias incluem :

- ***Supervised Learning*** - Essa técnica é baseada em dados rotulados. Neste processo um conjunto de dados rotulados é usado para elaborar modelos de forma a que estes reconheçam padrões permitindo que o sistema sugira decisões baseadas nesses exemplos. Assim, os modelos podem identificar respostas ou classificações futuros em dados, este processo pode ser aplicado em diversos contextos, como classificação de e-mails de spam [34], deteção de fraudes, previsão de vendas, reconhecimento e classificação de imagens [13].

- ***Unsupervised Learning*** - Esta técnica não requer que os dados sejam rotulados. Ou seja, o modelo pode identificar padrões e informações dos dados [34]. É amplamente utilizada em diversas áreas, entre as quais se destacam a segmentação de clientes, a recomendação de produtos e reconhecimento de imagens [13].

- ***Reinforcement Learning*** - Nesta técnica os modelos são melhorados durante o processo. Os modelos tentam diferentes soluções até encontrar a mais adequada para o problema que está a tentar resolver [34]. A sua aplicação estende-se a múltiplos domínios, entre os quais se incluem, a robótica e os Videojogos [13].

Embora muitas vezes sejam confundidos como sinónimos, ML e IA não são a mesma coisa [36]. A IA é o campo mais amplo, enquanto o ML é uma subcategoria específica dentro da IA, focada na capacidade das máquinas de aprender a partir de dados [22]. Tanto a IA como o ML tomam decisões com base em grandes volumes de dados, processando, interpretando e aprendendo com eles. No entanto, é importante lembrar que os algoritmos responsáveis por estas tarefas são programados por seres humanos. Por essa razão, preconceitos raciais, de género ou outros podem, mesmo que de forma não intencional, ser incorporados nos sistemas [19].

3.2 Viés em *Machine Learning*

O viés em ML ou viés de algoritmo refere-se a uma discriminação repetitiva e possivelmente indesejada nos resultados fornecidos por sistemas de IA sobre determinado tópico, decorrente de um julgamento ou favoritismo, levando a resultados distorcidos [35].

Além disso, o viés na IA pode ser incorporado nos modelos em consequência dos dados utilizados para o seu treino. Estes modelos identificam padrões e correlações nos dados, com o intuito de formular previsões e apoiar a tomada de decisões [62]. Quando os algoritmos detetam padrões associados a preconceitos históricos ou a disparidades sistémicas presentes nos dados de origem, as suas conclusões podem reproduzir e até amplificar esses mesmos preconceitos. Dado que os modelos de ML trabalham com elevadas quantidades de informação, mesmo pequenos enviesamentos nos dados de treino podem resultar em decisões discriminatórias em larga escala [62].

Os vieses podem variar em termos de consciencialização, mas todos têm o potencial de afetar negativamente a tomada de decisão. No entanto, eles manifestam-se de várias formas, dependendo do domínio, contexto e objetivo do sistema, sendo alguns dos mais comuns descritos de seguida.

3.2.1 Viés do algoritmo

O viés manifesta-se como uma inclinação irracional a atribuir um julgamento mais favorável ou desfavorável a alguma coisa, pessoa ou grupo. Ocorre quando o próprio algoritmo tem falhas, erros ou suposições que afetam seu desempenho ou lógica. Por exemplo, se o algoritmo usa um modelo simplificado ou impreciso, ou se baseia em uma determinada métrica ou critério que não é relevante ou apropriado [12]. O viés algorítmico na IA pode se manifestar de várias formas [65]:

- **Viés de Representação:** quando os dados usados para treinar modelos de IA não representam adequadamente a diversidade da população ou do problema que estão tentando resolver, resultando, por exemplo, em viés em determinado contexto.

- **Viés de Desempenho:** quando um modelo de IA funciona bem para um grupo de pessoas, mas não para outros devido a características específicas.

- **Viés de Confirmação:** quando os algoritmos de IA são projetados para favorecer informações que confirmam crenças ou estereótipos pré-existentes.

- **Viés de Dados:** quando os dados usados para treinar modelos de IA contêm preconceitos ou refletem desigualdades existentes na sociedade.

3.2.2 Viés humano

O viés humano ocorre quando os designers, desenvolvedores do algoritmo têm preferências, opiniões ou expectativas conscientes ou inconscientes que influenciam o sistema de IA.

Os sistemas de IA aprendem com dados existentes, identificando padrões e correlações para tomar decisões. Quando os dados usados no ML são tendenciosos, possuem preconceitos sociais ou estão desequilibrados, a IA pode replicá-los e até amplificá-los, o que pode comprometer a imparcialidade e a justiça dos resultados gerados [37, 45].

- **Dados de treino não representativos:** Se o conjunto de dados de treino não refletir a diversidade da população, o modelo poderá gerar resultados distorcidos.

- **Conjuntos de dados desequilibrados:** Quando uma classe se sobressai nos dados, o modelo tende a favorecer essa classe em detrimento das outras.

- **Dados não estruturados ou mal rotulados:** A falta de organização ou rotulação errada dos dados pode conduzir a modelos enviesados.

- **Baixa qualidade dos dados:** Dados imprecisos ou incompletos podem originar decisões enviesadas.

- **Algoritmos preconcebidos:** Algoritmos baseados em pressupostos preconcebidos podem gerar resultados tendenciosos.

Detetar viéses em sistemas de IA exige uma abordagem sistemática e rigorosa que envolve várias etapas e perspetivas. A análise de dados, a análise de algoritmos, a análise humana e a análise de contexto são métodos utilizados para detetar vieses em sistemas de IA [43].

- **A análise de dados:** Pode recorrer à estatística descritiva, visualização de dados, avaliação da qualidade dos dados e amostragem de dados.
- **A análise de algoritmos**: Utiliza revisão de código, teste de algoritmos, auditoria de algoritmos e transparência de algoritmos.
- **A análise humana:** Pode recorrer a investigação, entrevistas, discussões em grupo e testes com utilizadores.
- **A análise de contexto:** Pode recorrer a análise de cenários, avaliação de impacto, revisão ética e conformidade legal.

Todas essas técnicas podem ajudar a identificar quaisquer erros, lacunas, *outliers*, desequilíbrios, bugs, limitações, complexidades, opiniões, *feedbacks*, reclamações ou incompatibilidades que possam introduzir vieses.

Existem diversas ferramentas e *softwares* disponíveis que auxiliam na análise de vieses. *Softwares* como o R e o Python oferecem pacotes específicos para a deteção e correção de vieses. Além disso, plataformas de visualização de dados podem ajudar a identificar padrões que indicam a presença de vieses. A escolha da ferramenta adequada depende do tipo de análise e dos dados disponíveis [27].

Para usar essas técnicas e ferramentas de forma eficaz, deve-se definir o contexto e os objetivos do seu processo de deteção de viés, selecionar as técnicas e ferramentas apropriadas, aplicá-las ao seu sistema de IA e, posteriormente, analisar, relatar e tomar decisões com base nos resultados [43].

Os impactos do viés na IA podem ser vastos e profundos. Quando não é devidamente abordado, o viés pode acentuar desigualdades sociais, reforçar estereótipos e violar normas legais. Entre os impactos mais comuns, destacam-se [62]:

- **Desigualdades sociais:** O viés na IA pode agravar desigualdades sociais já existentes, afetando de forma desproporcionada comunidades marginalizadas e contribuindo para o aumento das disparidades económicas e sociais.

- **Reforço de estereótipos:** Sistemas de IA tendenciosos podem consolidar estereótipos negativos, perpetuando perceções e tratamentos discriminatórios com base na raça, género ou outras características.

- **Questões éticas e legais:** A presença de viés em sistemas de IA levanta sérias preocupações éticas e legais, colocando em causa a equidade e a justiça das decisões automatizadas. As organizações devem gerir cuidadosamente estas questões, de forma a garantir o cumprimento das normas legais e o respeito pelas responsabilidades éticas.

- **Impactos económicos:** Algoritmos enviesados podem prejudicar injustamente certos grupos, limitando oportunidades de emprego e perpetuando desigualdades no local de trabalho

- **Impactos nos negócios:** O viés nos sistemas de IA pode originar decisões erradas e reduzir a rentabilidade das empresas. Caso os enviesamentos presentes nas suas ferramentas de IA se tornem públicos, as empresas poderão sofrer danos reputacionais, perdendo a confiança dos consumidores e quota de mercado.

- **Impactos na saúde e segurança:** No sector da saúde, ferramentas de diagnóstico enviesadas podem conduzir a diagnósticos incorrectos ou a planos de tratamento subóptimos para determinados grupos populacionais, agravando desigualdades no acesso e qualidade dos cuidados de saúde.

- **Bem-estar psicológico e social:** A exposição repetida a decisões enviesadas por parte da IA pode provocar níveis elevados de stress e ansiedade nos indivíduos afectados, com consequências negativas para a sua saúde mental e bem-estar.

Mitigar o viés algorítmico é um desafio complexo, mas essencial para garantir que os modelos de IA sejam justos e inclusivos [73]. Embora seja quase impossível eliminar completamente o viés nos sistemas de IA, algumas ações podem ser tomadas para evitar que o viés ocorra [71].

- **Recolha e Validação de Dados Representativos:** A recolha de dados representativos é essencial para garantir que os modelos de IA sejam justos e precisos. É crucial que os dados utilizados representem a heterogeneidade da base de clientes, considerando aspetos como género, raça, faixa etária, localização geográfica e status socioeconómico, entre outros.

- **Revisão e Monitorização Contínua de Algoritmos:** A implementação de auditorias algorítmicas regulares permite a identificação de vieses ao longo do ciclo de vida do modelo. Isso inclui a realização de testes com diferentes subgrupos demográficos para avaliar o desempenho do modelo e identificar possíveis disparidades. Essas auditorias devem ser acompanhadas de ajustes contínuos nos algoritmos para corrigir quaisquer distorções identificadas.

- **Intervenção Humana nas Decisões:** A intervenção humana continua a ser crucial para garantir a equidade nas decisões. Isto pode significar a revisão manual de decisões críticas, como a aprovação de crédito para aquisição de serviços ou a definição de estratégias de retenção de clientes. O envolvimento humano permite a aplicação de um julgamento ético e a consideração de contextos específicos que os algoritmos podem não conseguir captar.

- **Diversidade na Identificação e Recolha de Amostras:** A redução do viés algorítmico não é apenas uma questão técnica, mas também cultural. As equipas de desenvolvimento e de gestão devem estar conscientes dos riscos associados ao viés e devidamente preparadas para os identificar e corrigir. Um processo de identificação e recolha de dados bem estruturado é essencial para garantir que as amostras recolhidas representem, de forma justa, a diversidade da população alvo.

- **Desenvolvimento de Modelos Justos e Inclusivos:** Para além de corrigir enviesamentos existentes, é fundamental adotar uma abordagem proativa no desenvolvimento de modelos que integrem, desde a sua conceção, princípios de justiça e inclusão. Isto pode incluir a definição de métricas específicas para avaliar a equidade dos modelos, bem como a aplicação de técnicas de *Fairness-Aware* ML, que ajustam os modelos de forma a minimizar disparidades entre diferentes grupos [73].

A redução do viés em sistemas de IA enfrenta diversos desafios, entre os quais se destacam [12]:

- **Equilíbrio entre viés e precisão:** Reduzir o viés sem comprometer a precisão dos modelos é uma tarefa complexa. Em muitos casos, remover dados sensíveis pode prejudicar o desempenho do sistema, sendo necessário recorrer a abordagens mais sofisticadas que conciliem justiça e eficácia.

- **Falta de transparência e documentação:** A escassez de documentação técnica clara sobre os algoritmos dificulta a compreensão por parte do público e dos reguladores, comprometendo a capacidade de identificar e corrigir enviesamentos. É fundamental encontrar um equilíbrio entre transparência e proteção de propriedade intelectual.

- **Falta de diversidade nas equipas de desenvolvimento:** A ausência de diversidade nas equipas limita a identificação de diferentes formas de viés. A inclusão de profissionais com diferentes origens, géneros, áreas de formação e experiências contribui para soluções mais justas e representativas.

- **Desafios regulatórios e de governação:** A legislação tem dificuldade em acompanhar a velocidade da inovação tecnológica, e muitas empresas, sobretudo pequenas e médias, ainda não dispõem de estruturas adequadas para assegurar a conformidade com padrões éticos e legais. Existe também a preocupação de que uma regulação excessiva possa travar a inovação.

- **Adaptação regional de tecnologias globais:** A importação de tecnologias desenvolvidas noutros contextos, sem a devida consideração pelas especificidades culturais, legais e sociais locais, pode acentuar desigualdades existentes ou criar novos enviesamentos. A regulação deverá assegurar que os sistemas de IA sejam devidamente adaptados às realidades regionais.

3.3 Métodos para Combater Viés

Esta secção descreve os métodos utilizados para identificar e analisar o viés presente num conjunto de dados, bem como os procedimentos adotados para validar a metodologia proposta.

De entre os diversos testes estatísticos existentes, existem várias formas de detetar vieses nos dados, incluindo **Teste T, Regressão Logística, Qui-quadrado, ANOVA**. Cada uma destas técnicas oferece ferramentas específicas para investigar se variáveis ou características de diferentes grupos podem influenciar de forma desigual os resultados de um estudo.

3.3.1 Teste T

É um tipo de teste estatístico usado para comparar as médias de dois grupos. São frequentemente aplicados em estudos que envolvem a comparação entre dois grupos independentes, com um grupo tratado com A e o outro com B [41]. A fórmula do Teste T é apresentada abaixo na expressão 1 [7].

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{1}$$

onde t é o valor do Teste T, X_1 e X_2 são as médias dos dois grupos que estão sendo comparados, s^2 é o erro padrão agrupado dos dois grupos e n_1 e n_2 são o número de observações em cada um dos grupos.

O teste T pode ser dividido em dois tipos [41]:

- **Teste T independente:** Que pode ser usado quando os dois grupos em comparação são independentes entre si.

- **Teste T pareado:** Que pode ser usado quando os dois grupos em comparação são dependentes entre si.

Para escolher o Teste T adequado, é necessário considerar dois fatores: se os grupos comparados pertencem a uma única população ou a populações diferentes, e se a diferença que se pretende testar é unidirecional [7].

3.3.2 Regressão Logística

A regressão logística é uma técnica que descobre a relação entre várias variáveis independentes (X_j) e uma variável dependente binária (Y). Este modelo descreve o valor esperado de Y por meio da expressão apresentada na Equação 2 [64].

$$E(Y) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{k} \beta_j X_j\right)\right]} \tag{2}$$

A expressão geral é dada pelas Equações 3 e 4.

$$f(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

$$z = \beta_0 + \sum_{j=1}^{k} \beta_j X_j \tag{4}$$

em que z é conhecido como *log-odds*, variando de $-\infty$ a $+\infty$, como se observa na Figura 2. Assim, a função logística $f(z)$ normaliza a saída do modelo para o intervalo $[0, 1]$, informando a probabilidade de ocorrência do evento de interesse.

A utilização da técnica de Regressão Logística é adequada em muitas situações, porque permite analisar o efeito de uma ou mais variáveis independentes (categóricas ou numéricas) [64]. Além disso, é possível calcular a probabilidade de ocorrência de um

evento através de um dado conjunto de características de uma observação. O modelo de Regressão Logística permite [18]:

- Modelar a probabilidade de um evento ocorrer, dependendo dos valores das variáveis independentes;
- Estimar a probabilidade de um evento ocorrer para uma observação específica.
- Prever o efeito do conjunto de variáveis sobre a variável dependente binária.
- Classificar observações, estimando a probabilidade de pertencerem a uma determinada categoria.

3.3.3 Qui-Quadrado

É uma técnica estatística utilizada para verificar se existe uma diferença significativa entre as frequências observadas e as esperadas em uma ou mais categorias. Este teste conhecido como teste do qui-quadrado (χ^2) é particularmente útil para avaliar a independência entre variáveis categóricas, assim como para avaliar a adequação de um modelo teórico aos dados observados. É especialmente eficaz em amostras grandes [23, 43]. Na Equação 5 apresenta-se a fórmula para calcular o teste:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{5}$$

onde χ^2 é o valor da estatística Qui-Quadrado, O_i representa as frequências observadas i, e E_i representa as frequências esperadas i.

O teste do qui-quadrado é particularmente útil quando se trabalha com dados categóricos, como género, preferências ou afiliações políticas, para testar relacionamentos e padrões [51]. Existem diferentes tipos de testes de qui-quadrado [51]:

- **Teste de adequação do Qui-Quadrado:** Uma variável categórica individual é testada para determinar se segue uma distribuição específica. Um modelo ou

dados históricos são frequentemente utilizados para verificar se os dados observados correspondem a uma distribuição esperada.

- **Teste Qui-quadrado de Independência:** Neste teste, duas variáveis categóricas são avaliadas quanto à sua independência. O teste analisa se a distribuição de uma variável varia consoante os níveis de uma segunda variável.

Para que os resultados do teste do qui-quadrado sejam válidos, é necessário cumprir três pressupostos principais [51]:

- **Amostragem aleatória:** Os dados devem ser recolhidos através de uma amostragem aleatória, garantindo que todos os elementos da população têm igual probabilidade de serem incluídos.

- **Variáveis categóricas:** O teste aplica-se apenas a variáveis categóricas, ou seja, dados que podem ser organizados em categorias distintas.

- **Frequência esperada adequada:** As categorias devem ter contagens esperadas suficientemente elevadas nas tabelas de contingência. Estas contagens baseiam-se na hipótese nula e garantem que o teste tem poder estatístico e resultados fiáveis.

O teste de Qui-Quadrado informa o quanto os dados observados diferem das frequências esperadas caso não exista relação entre as categorias. Um valor elevado de χ^2 indica que há uma grande diferença entre a frequência observada e a esperada. Por outro lado, um valor baixo de χ^2 indica que os dados observados estão muito próximos do que era esperado [51]. Na prática, a decisão é tomada através da comparação entre o valor-p obtido no teste e o nível de significância previamente definido.

3.3.4 ANOVA

A análise de variância (ANOVA) é uma técnica estatística utilizada para determinar se há diferenças significativas entre as médias de três ou mais grupos [30]. As equações seguintes apresentam as fórmulas utilizadas para o cálculo da Anova.

$$F = \frac{QM_{trat}}{QM_{erro}} \tag{6}$$

onde F é a estatística de teste da ANOVA, QM_{trat} representa a Média dos Quadrados entre grupos, QM_{erro} é a Média dos Quadrados do Erro dentro dos grupos [59].

$$QM_{trat} = \frac{n \sum_{j=1}^{g} \left(\bar{Y}_j - \bar{Y}\right)^2}{g - 1} \tag{7}$$

onde n é o número de observações por grupo, g é o número de grupos, $\bar{Y}_j$ é a média do grupo j, e $\bar{Y}$ é a média global[59].

$$QM_{erro} = \frac{\sum_{j=1}^{g} (n_j - 1) \cdot s_j^2}{N - g} \tag{8}$$

onde n_j é o número de observações no grupo j, s_j^2 é a variância amostral do grupo j, N é o número total de observações, g é o número de grupos [59].

Para a realização do teste de ANOVA, é necessário ter uma variável dependente numérica ou contínua, bem como pelo menos um fator independente categórico com dois ou mais níveis [30].

A ANOVA pode ser classificada de duas maneiras [59]:

- **ANOVA unilateral:** Neste tipo de análise de variância há apenas um fator em uma variável independente, determina se existem diferenças estatisticamente significativas entre as médias de três ou mais grupos.

- **ANOVA bidirecional:** Uma ANOVA bidirecional possui duas variáveis independentes afetando uma variável dependente, é usada para testar a interação entre os dois fatores e analisar o efeito de dois fatores ao mesmo tempo. Portanto, essa análise considera que variações pequenas podem ocorrer por acaso, enquanto diferenças maiores indicam a presença de causas reais [32].

Para determinar há diferença entre os valores podem-se utilizar diferentes métodos:

- **Teste de Hipóteses** envolve a formulação de duas hipóteses, a hipótese nula (H_0) e a hipótese alternativa (H_1) [28]. As hipóteses estatísticas são suposições ou afirmações feitas sobre um parâmetro, ou parâmetros, na população. Em qualquer teste de hipóteses existem duas hipóteses [63]:
  - **Hipótese nula (H_0)**: Assume que não existe qualquer diferença ou efeito estatisticamente significativo;
  - **Hipótese alternativa (H_1)**: Contradiz a hipótese nula, sugerindo a existência de uma diferença ou efeito real.
- **p-valor** indica a probabilidade de observar o valor da estatística de teste e determinar se a diferença encontrada é significativa ou se pode ser explicada por variação aleatória. É calculado a partir do teste estatístico (χ^2) [28, 63].
  - Se o valor-$p \leq$ nível de significância $\rightarrow$ rejeita-se H_0, considerando o resultado estatisticamente significativo;
  - Se o valor-$p >$ nível de significância, $\rightarrow$ não se rejeita H_0, por não existirem evidências estatísticas suficientes.
- **Níveis de Significância** constituem limites estatísticos definidos pelo investigador para apoiar a decisão de rejeitar ou não a hipótese nula num teste de hipóteses [75].

3.4 Técnicas de Balanceamento de Dados

O desequilíbrio de dados é um problema comum nos conjuntos de dados disponíveis, especialmente quando as classes dentro de um conjunto de dados apresentam distribuições desiguais. Esse desequilíbrio ocorre quando uma ou mais classes possuem significativamente mais instâncias do que outras, ou seja, quando a quantidade de dados numa classe específica (classe maioritária) é superior à de outra (classe minoritária). Isto pode levar a modelos preditivos tendenciosos e a um desempenho inferior na identificação das classes menos representadas [1, 14].

Para além das abordagens técnicas, é importante considerar os aspetos éticos e legais do viés algorítmico. A União Europeia aprovou o AI Act (2023), estabelecendo requisitos de transparência, responsabilidade e não discriminação para os sistemas de IA, promovendo uma IA centrada no ser humano e de confiança, e garantindo a proteção de direitos fundamentais, da saúde, da segurança e do ambiente. Organizações internacionais, como a OCDE, também emitem recomendações para o uso responsável da IA. Estes enquadramentos reforçam a necessidade de integrar métricas de justiça e mecanismos de auditoria nos sistemas algorítmicos, apoiando a inovação e a adoção segura das tecnologias de IA [25].

Existem dois tipos de problemas de desequilíbrio no conjunto de dados [67]:

- **Desequilíbrios entre classes :** Ocorre quando as amostras de duas classes apresentam um número significativamente desigual de dados, com uma classe a ter muitos mais dados do que a outra.

- **Desequilíbrios dentro da classe:** Acontecem quando, dentro das classes maioritária e minoritária, existem diferentes conceitos, sendo que alguns conceitos ocorrem com menor frequência do que outros.

Para lidar com estes problemas de desequilíbrio existem duas abordagens principais [68]:

- **Abordagens ao nível dos dados :** Estas envolvem a amostragem das classes maioritárias ou minoritárias de forma a equilibrar a quantidade de dados entre elas.

- **Abordagens ao nível do algoritmo:** Neste caso, consistem em modificar os métodos de classificação sem alterar os dados originais.

Ambas as metodologias apresentam vantagens e desvantagens. A abordagem ao nível dos dados é, geralmente, robusta e estável para a maioria dos classificadores. Contudo, pode levar a ajuste excessivo nos dados de treino (*overfitting*) ou perda de informação nos dados amostrados. Por outro lado, a abordagem ao nível do algoritmo preserva os dados

originais, mas o desempenho do modelo depende do classificador utilizado. Isto implica que um algoritmo desenvolvido pode não apresentar o mesmo nível de eficácia quando aplicado a conjuntos de dados com características distintas.

Entre essas duas abordagens, estudos indicam que as soluções ao nível dos dados são mais desenvolvidas para lidar com o desbalanceamento do que as baseadas em modificação de algoritmos. Em geral, as soluções ao nível dos dados utilizam técnicas de amostragem, que podem ser classificadas em três tipos: ***Oversampling, Undersampling, Híbrido***.

3.4.1 *Oversampling*

É uma técnica que tem como objetivo equilibrar a distribuição das classes, aumentando a quantidade de exemplos da classe minoritária no conjunto de treino, eliminando o viés da classe maioritária e melhorando o desempenho do modelo [11, 49]. A Figura 3.2 representa o processo de *oversampling*, mostrando a replicação de novas instâncias da classe menos representada, de forma a equilibrar o conjunto de dados.

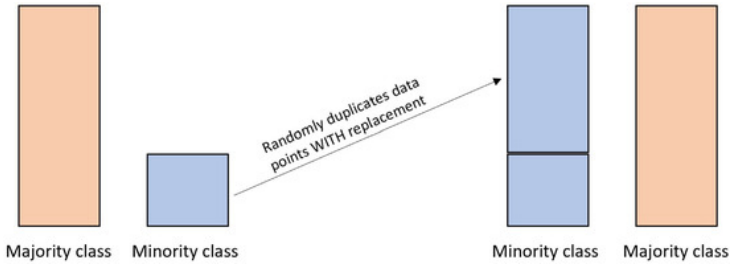

Figura 3.2: Processo de *oversampling* [76].

Está técnica apresenta as seguintes vantagens:

- **Aumento sem duplicação** - aumenta a classe minoritária criando novos exemplos, sem repetir os existentes.
- **Melhora desempenho desbalanceado** - melhora significativamente o desempenho de modelos de aprendizado de máquina [4].
- **Preserva informações originais** - não há perda de dados originais, pois tanto a classe minoritária quanto a maioritária são mantidas [11].

No entanto, tem como desvantagens:

- **Dados sintéticos distorcidos** - cria dados sintéticos para equilibrar o conjunto de dados, eles podem não refletir com precisão os dados reais, o que pode afetar o modelo.

- **Aumenta o risco de *overfitting*** - especialmente se a quantidade de novos exemplos sintéticos for muito grande [4].

- **Maior tempo de processamento** - aumentando a quantidade de exemplos da classe minoritária no conjunto de treino, aumenta tempo de processamento [11].

Apesar de ser uma técnica amplamente utilizada para lidar com conjuntos de dados desequilibrados, o *oversampling* apresenta alguns riscos que devem ser considerados, como a criação de amostras artificiais pode gerar padrões que não existem na realidade, distorcendo a representação dos dados. Além disso, os modelos podem tornar-se excessivamente confiantes com base nessas amostras sintéticas, o que pode conduzir a falhas graves quando aplicados a dados reais [9].

Entre as diversas estratégias para lidar com o desequilíbrio nos conjuntos de dados, destacam-se algumas técnicas específicas de *oversampling*, nomeadamente:

- ***Random Oversampling*:** é uma técnica simples, mas eficaz, para reamostragem [11], que duplica exemplos de classe minoritária aleatoriamente para equilibrar as distribuições de classe [3].

  - Ideal para conjuntos de dados pequenos que precisam de balanceamento rápido [9].

  - Não recomendado para conjuntos de dados complexos, isto é, conjuntos que apresentam sobreposição entre classes, falta de dados representativos e pequenas disjunções [9, 33].

- ***Synthetic Minority Over-sampling Technique (SMOTE)*:** É amplamente

utilizado para atenuar problemas de desequilíbrio de classes [3]. Em vez de simplesmente replicar exemplos da classe minoritária, como acontece nas técnicas tradicionais de oversampling, o *SMOTE* cria novos exemplos sintéticos ao calcular os k vizinhos mais próximos dos exemplos existentes dessa classe [39].

  - Indicado quando existe um número grande de amostras, permitindo capturar a variedade dos dados.
  - Não recomendado se há um número pequeno de amostras ou se os dados são muito dispersos [9].

- ***Adaptive Synthetic Sampling (ADASYN)*:** Semelhante ao *SMOTE*, mas concentra-se em gerar amostras sintéticas nas zonas mais próximas do limite de decisão entre as classes [3].

  - Ideal para conjuntos de dados complexos com regiões desafiadoras [9].
  - Não indicado se os dados forem simples e bem definidos [9].

3.4.2 *Undersampling*

O *Undersampling* visa equilibrar a distribuição das classes, reduzindo o número de exemplos da classe maioritária, tornando o conjunto de dados mais equilibrado e, consequentemente, melhorando o desempenho dos modelos de ML [4]. A Figura 3.3 ilustra o processo de *Undersampling*, mostrando a remoção aleatória de instâncias da classe mais representada.

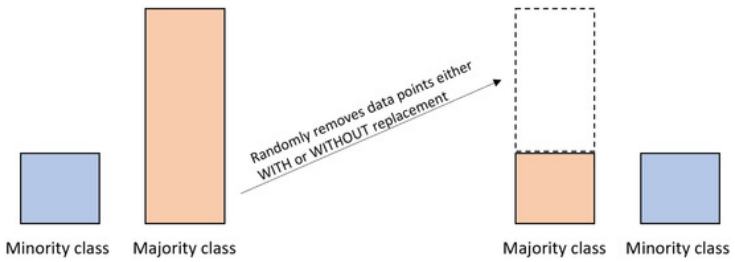

Figura 3.3: Processo de *Undersampling* [76].

Esta técnica tem como vantagens [4]:

- **Menor tempo de processamento** - reduzindo o número de exemplos da classe maioritária no conjunto de treino, diminui o tempo de processamento.
- **Não gera amostras sintéticas** - trabalha apenas com os dados existentes, evitando a introdução de possíveis ruídos.

No entanto, apresenta as desvantagens [4]:

- **Perda de informação** - pois a remoção de exemplos pode resultar na perda de informações importantes, especialmente se a classe maioritária tiver muitos exemplos relevantes.
- **Pode causar *overfitting*** - pois reduz o número de exemplos usados para treinar o modelo.

O *Undersampling*, embora utilizado para equilibrar conjuntos de dados, apresenta também alguns riscos importantes que devem ser tidos em conta. Esta técnica pode levar à perda permanente de informações, comprometendo a qualidade do conjunto de dados. Para além disso, existe o risco de eliminar, limites essenciais entre as classes, o que pode dificultar a correta compreensão do problema por parte do modelo. E também pode criar distribuições artificiais das classes que diferem significativamente das condições reais do mundo, afetando a capacidade de generalização dos resultados [9].

Para assegurar um melhor equilíbrio entre as classes nos conjuntos de dados, existem diversas técnicas específicas de *undersampling* que atuam na redução da classe maioritária:

- ***Tomek Links*:** Uma técnica de *undersampling* que remove exemplos da classe maior para tornar os limites entre as classes mais claros. É ideal quando há muita sobreposição entre as classes [9], e não é recomendado quando os grupos já estiverem bem separados [9].
- ***Near Miss:*** É uma técnica que reduz apenas a classe maioritária. Funciona mantendo apenas as amostras da classe maioritária que estão mais próximas das amostras da classe minoritária, criando, assim, um conjunto de dados mais equilibrado

[40]. Geralmente é utilizado quando se pretende um controle sobre quais exemplos manter [9], mas não indicado para quem precisa de uma solução rápida e simples [9].

- **ENN (Edited Nearest Neighbors):** É uma técnica utilizada para lidar com conjuntos de dados desequilibrados através da edição do próprio conjunto. Foca-se na remoção de amostras da classe maioritária que estão próximas das amostras da classe minoritária, garantindo uma distinção mais clara entre as classes [40]. Isto ajuda a eliminar ruído e a definir melhor os limites entre os grupos [9]. É ideal para limpar dados indesejados e eliminar *outliers*, no entanto, é desnecessário quando os dados já estiverem organizados e limpos [9].

3.4.3 Híbrido

É uma abordagem de balanceamento de dados que combina técnicas de *Undersampling* e *Oversampling*. Reduz-se o número de amostras da classe maioritária através da *Undersampling*, ao mesmo tempo que aumenta a quantidade de dados da classe minoritária com recurso à *Oversampling* [68]. A Figura 3.4 representa o processo híbrido.

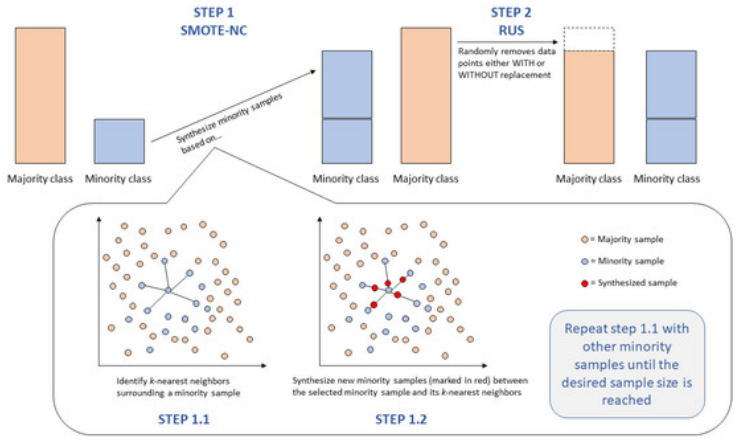

Figura 3.4: Processo híbrido [76].

Existem vários tipos de métodos híbridos que combinam técnicas de *undersampling* e *oversampling*. Estas abordagens procuram aproveitar as vantagens de cada método,

produzindo conjuntos de dados mais equilibrados e limpos. As técnicas híbridas, como:

- ***SMOTETomek*:** Com a combinação de sobreamostragem do tipo *SMOTE* e a subamostragem do tipo Tomek (*SMOTE + Tome*k). O *SMOTETomek* funciona primeiro criando novos exemplos para a classe minoritária utilizando o *SMOTE* e, em seguida, limpa os limites, removendo através do *Tomek Links*. Isto ajuda a criar um conjunto de dados mais equilibrado, com limites mais claros [9].

- ***SMOTEENN*:** É a combinação de *SMOTE + ENN*. O *SMOTEENN* funciona primeiro criando novos exemplos para a classe minoritária usando o *SMOTE* e, em seguida, limpa ambos os grupos, removendo exemplos que não se ajustam bem aos seus vizinhos, utilizando a técnica de *ENN*. Tal como o *SMOTETomek*, este método ajuda a criar um conjunto de dados mais limpo, com fronteiras mais definidas entre as classes [9].

Capítulo 4

Metodologia

Este capítulo visa a descrever os dois conjuntos de dados utilizados, bem como a metodologia adotada para aplicar os métodos estáticos na identificação de viés nos conjuntos de dados.

4.1 Caso de Estudo: Minneapolis

O conjunto de dados Minneapolis possui interações policiais na cidade de Minneapolis, localizada no estado de Minnesota, nos Estados Unidos, com uma população aproximada de 428.039 habitantes, com uma representatividade de 61.59% de pessoas brancas, 18.25% de pessoas negras, 5.21% pessoas asiáticas e 4.87% de outras raças e percentagens menores para populações nativas americanas, havaianas nativas ou das ilhas do Pacífico e multirraciais [77].

Este conjunto foi escolhido devido à ampla disponibilidade de dados públicos sobre policiamento, e tinha como objetivo analisar se a operação policial foi justa e se houve discriminação racial, e é composto por quatro variáveis categóricas:

- *CallDisposition* indica o resultado de cada ocorrência policial.
- *PersonSearch* informa se foi realizada uma busca pessoal no indivíduo abordado.
- *VehicleSearch* indica se houve busca ao veículo do indivíduo.

- *Race* representa a raça do indivíduo abordado e funciona como variável independente no estudo, permitindo investigar possíveis vieses nas interações policiais.

Neste conjunto de dados, a variável *Race* é a variável que se pretende analisar para explicar as relações entre as demais variáveis, servindo como referência para identificar possíveis diferenças de tratamento ou discriminação nas abordagens policiais.

Este conjunto de dados permite uma análise exploratória para identificar padrões e tendências nos dados. A Tabela 4.1 apresenta estatísticas descritivas das principais variáveis do conjunto de dados analisados. Para cada variável são indicados o número de observações, que representa o total de registos disponíveis, o número de valores únicos, que mostra a variedade de categorias presentes, o valor mais frequente ou moda, que indica a categoria mais frequente, a frequência, que mostra quantas vezes a moda aparece, e os valores ausentes, que mostram o número de valores em falta.

Tabela 4.1: Estatísticas descritivas da abordagem policial

Variável	**Observações**	**Únicos**	**Mais Frequente**	**Frequência**	**Valores Ausentes**
CallDisposition	159845	25	*ADV-Advised*	66908	2243
PersonSearch	141833	2	*NO*	123513	20255
VehicleSearch	141833	2	*NO*	130581	20255
Race	141833	8	*Black*	50029	20255

A base de dados analisada inclui um total de 162.088 interações policiais registadas, fornecendo informações detalhadas sobre o número de operações policiais realizadas, bem como dados sobre à população abordada. Os dados incluem informações sobre a raça dos indivíduos envolvidos, a ocorrência de buscas pessoais ou veiculares [70].

A Figura 4.1 ilustra a distribuição das abordagens policiais por raça, mostrando quantas vezes indivíduos de cada grupo racial foram abordados, permitindo visualizar a frequência dessas interações entre os diferentes grupos.

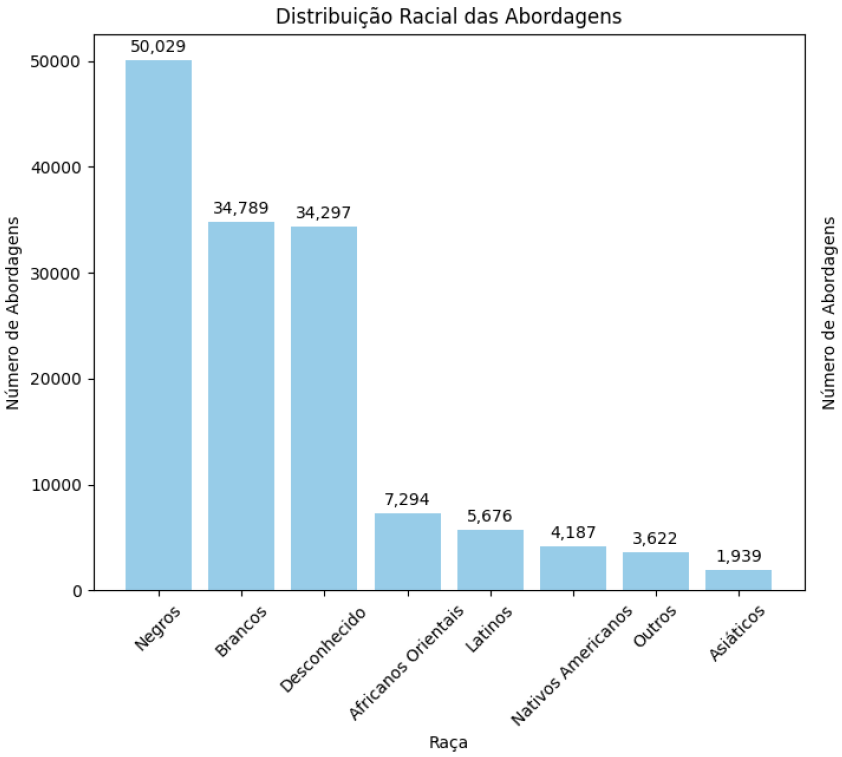

Figura 4.1: Distribuição das abordagens policiais por raça

Observando o gráfico, presente na Figura 4.1 percebe-se que a maior quantidade de abordagens ocorre entre população pela raça negra, seguida pela raça branca e pelo grupo cuja raça foi registada como desconhecida. Outros grupos raciais, como asiáticos, nativos americanos, latinos e africanos orientais , apresentam números significativamente menores de ocorrências. Essa distribuição evidencia como as abordagens estão concentradas entre os grupos raciais.

Já a Figura 4.2 apresenta a percentagem da população de Minneapolis por raça, dividida em três categorias, *White*, *Black* e *Other*. A categoria *Other* corresponde à soma das populações nativas americanas, nativas havaianas, asiáticas e multirraciais.

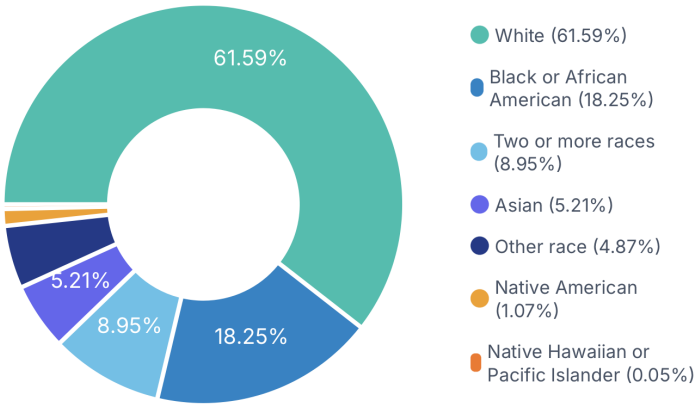

Figura 4.2: Percentagem da população em Minneapolis por raça [77]

Analisando o gráfico da Figura 4.2 observa-se que a população de Minneapolis é predominantemente composta por indivíduos de raça White, representando mais de metade do total. No entanto, também se observa a presença expressiva de outros grupos raciais, em particular indivíduos Black ou *African American*, que constituem a segunda maior parcela. As demais categorias, embora com percentuais menores, refletem a diversidade racial da cidade e servem de base para as análises seguintes, ajudando a entender como essa composição populacional pode influenciar os padrões de operações policiais explorados neste estudo.

4.2 Caso de Estudo: Desempenho Académico

Este estudo de caso baseia-se num conjunto de dados com respostas de 1.000 estudantes, recolhidas durante o período de ensino à distância [66]. Utilizaram-no para analisar de que forma o ensino online influenciou a mudança no desempenho académico dos alunos.

O conjunto de dados contém informação recolhida por inquéritos e abrange informações sobre o tempo de ecrã, a duração do sono, bem como sentimentos associados ao stress, divididas em quatro variáveis, duas numéricas (*Screen Time* e *Sleep Duration*) e duas

categóricas (*Stress Level* e *Academic Performance*).

- *Screen Time*, que representa o número de horas de exposição diária a ecrãs.
- *Sleep Duration*, que corresponde ao número de horas dormidas por dia.
- *Stress Level*, que indica o nível de stress reportado pelos estudantes, e *Academic Performance*, *Academic Performance* que descreve o seu desempenho académico, com três categorias piorou (*Declined*); melhorou (*Improved*); e manteve (*Same*).

A variável *Academic Performance* é considerada a variável independente, uma vez que constitui o resultado que se pretende explicar a partir das restantes variáveis. Assim, a análise permite compreender de que forma fatores como o tempo de ecrã, a duração do sono e o nível de stress influenciam o desempenho académico no contexto do ensino online.

A Tabela 4.2 apresenta uma descrição detalhada das principais variáveis incluídas no conjunto de dados. Para cada variável são apresentados o número total de observações, que corresponde ao total de registos recolhidos, o número de valores únicos, que revela a diversidade de valores existentes, a média e o desvio-padrão (*Std*) no caso das variáveis numéricas, bem como o valor mais frequente, ou moda, e a respetiva frequência, que indicam qual a categoria ou valor que ocorre com maior frequência nos dados. O conjunto analisado inclui quatro variáveis principais, duas numéricas e duas categóricas.

Tabela 4.2: Estatísticas descritivas do desempenho académico em ensino à distância

Variável	**Observações**	**Únicos**	**Mais Frequente**	**Frequência**	**Média**	**Std**
Education Level	1000	11	*MTech*	143	—	—
Screen Time	1000	—	—	—	6.91	2.90
Sleep Duration	1000	—	—	—	6.46	1.47
Stress Level	1000	3	*Medium*	492	—	—

A partir da Tabela 4.2 observam-se os principais valores associados a cada variável. Verifica-se que as variáveis numéricas (*Screen Time e Sleep Duration*) apresentam médias semelhantes, enquanto as variáveis categóricas são descritas apenas pelas suas frequências. Constata-se também que algumas células estão vazias, o que ocorre porque determinados indicadores estatísticos (como média, desvio-padrão) apenas se aplicam a variáveis

numéricas. Assim, nas variáveis categóricas esses valores não são calculados, permanecendo, por isso, em branco. Além disso, todas as variáveis apresentam o mesmo número de observações (1000).

4.3 Procedimento

Esta secção descreve os procedimentos adotados para a análise e identificação de viés algorítmico, com base em dois conjuntos de dados distintos, aplicando testes estatísticos e técnicas de balanceamento. Inicialmente, os dados foram organizados em Python, com a remoção de valores em falta. Para a identificação de viés, foram utilizados dois testes estatísticos: Qui-Quadrado e Análise de Variância (ANOVA). Para o balanceamento dos dados, aplicou-se a técnica de *Random Oversampling*.

4.3.1 Conjunto de dados: Minneapolis

Neste conjunto de dados, recorreu-se ao teste do Qui-Quadrado para analisar a existência de associação entre variáveis categóricas, como *CallDisposition*, *PersonSearch* e *VehicleSearch*, considerando *race* como variável independente. Adicionalmente, foi aplicada a ANOVA para verificar diferenças estatisticamente significativas nas médias dessas variáveis. Para validar o desempenho das técnicas aplicadas, dessa forma reduzindo o impacto de distribuições desiguais entre as categorias, aplicou-se a técnica de *Random Oversampling*, com o objetivo de equilibrar as classes, reduzir possíveis vieses, e validar se as técnicas deixavam de considerar enviesamento dos dados.

4.3.2 Conjunto de dados: Desempenho Académico

O mesmo procedimento foi aplicado para este conjunto de dados. O teste do Qui-quadrado foi utilizado para analisar a associação entre variáveis categóricas, como *Education Level* e *Stress Level*, e variáveis numéricas, como *Screen Time* e *Sleep Duration*, considerando

Education Level como variável independente. Foi também aplicada a ANOVA para verificar diferenças estatisticamente significativas nas médias destas variáveis. Da mesma forma, aplicou-se o *Random Oversampling* ao conjunto de dados para validar o desempenho das técnicas aplicadas.

Salienta-se que para ambas as análises, as hipóteses nula (H_0) e alternativa (H_1) foram definidas, adotando um nível de significância de 1%.

Capítulo 5

Resultados e Discussão

Neste capítulo são analisados e discutidos os resultados obtidos ao longo do estudo, interpretando-os com base nos objetivos e nas hipóteses inicialmente formuladas, considerando os dois conjuntos de dados distintos: um relacionado com operações policiais e outro sobre o desempenho académico no ensino online.

Para a análise estatística, foram utilizados os testes do Qui-Quadrado e da ANOVA, considerados os mais adequados às características das variáveis estudadas e aos objetivos da investigação. Estes testes permitem avaliar, de forma transparente, se as diferenças observadas entre os grupos são estatisticamente significativas, garantindo uma interpretação rigorosa e confiável da presença de possíveis vieses nos dados. Adicionalmente, em ambos os conjuntos de dados, aplicou-se a técnica de *Random Oversampling*, para validar o desempenho dos métodos aplicados, equilibrando os dados e reduzindo os vieses existentes.

5.1 Qui-Quadrado

A análise do teste do Qui-Quadrado é complementada pelo p-valor, que indica a probabilidade de observar uma diferença tão grande ou maior entre os grupos, assumindo que a hipótese nula é verdadeira. Para tal, adotou-se um nível de significância de 1% (0,01) como limite para determinar se o resultado é estatisticamente significativo. Dessa forma,

formularam-se as seguintes hipóteses: $\mathbf{H_0}$**:** Não existe associação entre as variáveis; $\mathbf{H_1}$**:** Existe associação entre as variáveis.

O p-valor obtido é comparado com o nível de significância definido. Se o p-valor for inferior ao nível de significância de $0,01$, rejeita-se a hipótese nula e conclui-se que existe associação entre as variáveis, mas caso o p-valor seja superior, não há evidência estatística suficiente para rejeitar a hipótese nula.

5.1.1 Conjunto de dados: Minneapolis

De seguida, apresentam-se os gráficos e a respetiva análise, elaborados com base nos dados recolhidos durante operações policiais em Minneapolis. Todas as variáveis analisadas — *VehicleSearch*, *PersonSearch*, *CallDisposition* e *Race* — são categóricas. A raça dos indivíduos abordados é considerada como variável independente, uma vez que se pretende verificar se existem diferenças estatisticamente significativas na forma como cada pessoa de cada raça foi abordada pela polícia.

A Figura 5.1 apresenta a distribuição racial para *VehicleSearch = YES*, indicando os casos em que os veículos foram revistados durante a operação policial. Neste caso, a variável *CallDisposition* representa a distribuição de raça em *callDisposition = LOP (Lights On Program)*, que oferece apoio sob a forma de vales para a reparação de viaturas em vez da aplicação de sanções.

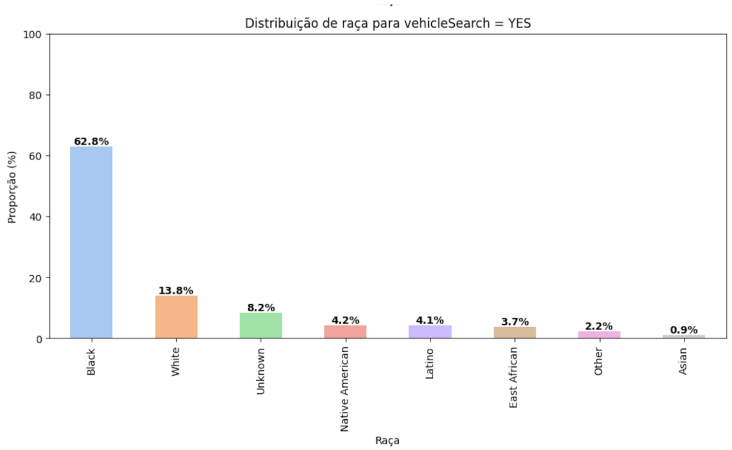

Figura 5.1: Distribuição das abordagens policiais por raça dos veículos revistados

O gráfico presente na Figura 5.1 apresenta a distribuição racial das abordagens policiais em que foi efetuada revista veicular. Do total de veículos fiscalizados, **(62.8%)** incidiram sobre indivíduos negros, enquanto apenas **(13.8%)** envolveram indivíduos brancos. Os restantes grupos raciais registaram percentagens significativamente inferiores, indivíduos com raça não identificada **(8.2%)**, nativos americanos **(4.2%)**, latinos **(4.1%)**, africanos orientais **(3.7%)**, outros **(2.2%)** e asiáticos **(0.9%)**. Analisando as distribuições dos dados no gráfico, conclui-se que os indivíduos negros são sujeitos a fiscalização de veículo de modo exagerado em relação às outras raças, revelando um padrão que sugere desigualdade racial nas práticas policiais, desfavorecendo principalmente essa população. Esta conclusão torna-se ainda mais evidente ao observar-se a composição da população de Minneapolis, que conta com 428.039 habitantes. A maioria é constituída por pessoas brancas **(61,59%)**, enquanto as pessoas negras representam **(18,25%)** da população, sendo as restantes percentagens distribuídas por outras raças, conforme referido no Capítulo 4, Secção 4.1, e ilustrado na Figura 4.2.

Esta tendência é comprovada pela aplicação do **teste do Qui-Quadrado**, cujo valor obtido é de $\chi^2 = 10280.14$ e o de $p - valor = 0.0000$. Dado que o p-valor é < 0.01, rejeita-se a hipótese nula de inexistência de associação entre as variáveis, concluindo-se a existência de associação e a presença de vieses raciais nas práticas policiais analisadas, o que demonstra que a distribuição das revistas por raça não ocorreu de forma aleatória.

Na Figura 5.2 são representados os casos quando *PersonSearch = YES*, correspondendo às situações em que os indivíduos foram revistados quando a polícia ordenou a paragem. Para esta situação, a variável *CallDisposition* apresenta a distribuição de raça em *callDisposition = RPT (Report)*, situações em que foi formalizado um registo oficial para possíveis investigações posteriores.

O gráfico presente na Figura 5.2 apresenta a distribuição das abordagens policiais por raça, considerando o total das pessoas submetidas à revista pessoal, que permite revelar que existe uma desigualdade racial nas abordagens que envolvem revistas pessoais. Os indivíduos negros representam **(61.9%)** das abordagens, um valor significativamente superior ao de qualquer outro grupo racial. Apenas **(17.2%)** das abordagens foram

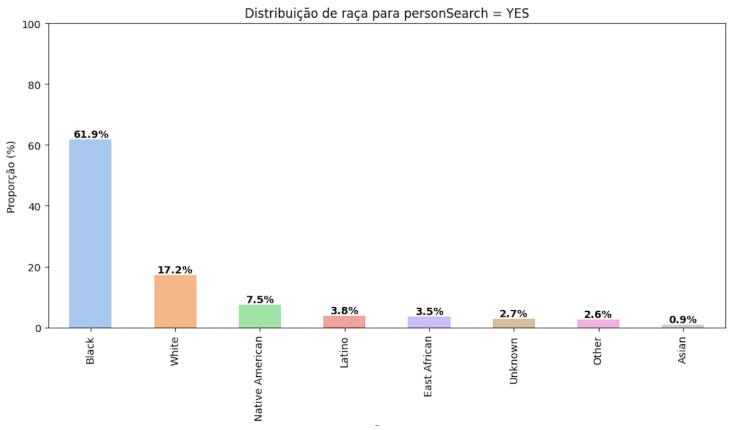

Figura 5.2: Distribuição das abordagens policiais por raça das pessoas revistadas.

direcionadas indivíduos brancas, enquanto os restantes grupos registaram percentagens ainda mais reduzidas.

Os resultados obtidos pela aplicação do **teste do qui-quadrado** foram $\chi^2 = 4618.14$ e $p - valor = 0.0000$, confirmando que esta distribuição não ocorre de forma aleatória. Como o p-valor é extremamente baixo (< 0.01), rejeita-se a hipótese nula, o que indica a existência de um padrão sistemático na realização destas abordagens. Este resultado confirma a presença de vieses raciais nas práticas policiais.

A Figura 5.3 evidencia a distribuição racial das abordagens policiais que resultaram na emissão de um Relatório para Futura Investigação **(RPT)** — isto é, situações em que foi formalizado um registo com vista a potenciais investigações futuras.

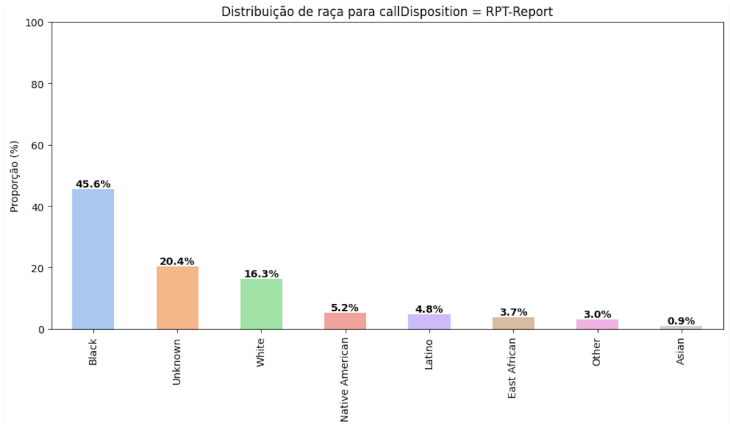

Figura 5.3: Distribuição racial das abordagens policiais que resultaram em RPT.

Analisando o gráfico da Figura 5.3, observa-se que do total dos registos feitos para investigações futuras, **(45.6%)** desses relatórios foram emitidos relativamente a indivíduos negros. Por outro lado, apenas **(16,3%)** dos relatórios foram emitidos relativamente a indivíduos brancos. A análise do gráfico permite constatar que a maioria dos relatórios foi direcionada a indivíduos negros, evidenciando um padrão de atuação policial que não se distribuiu igualmente entre os grupos raciais. O resultado do **teste do Qui-Quadrado** revelou $\chi^2 = 81080.81$ e $p - valor = 0.0000$, demonstrando que a distribuição dos relatórios por raça é estatisticamente significativa, rejeitando-se assim a hipótese nula de aleatoriedade. O p-valor extremamente reduzido reforça a existência de viés racial no registo destes relatórios, sugerindo que a tomada de decisão policial poderá estar influenciada por fatores discriminatórios.

Embora os gráficos anteriores e os resultados do teste do Qui-Quadrado tenham evidenciado a existência de viés racial nas abordagens policiais sobretudo no que se refere à população negra, a Figura 5.4 vem reforçar ainda mais essa desigualdade, mesmo num contexto que é considerado positivo.

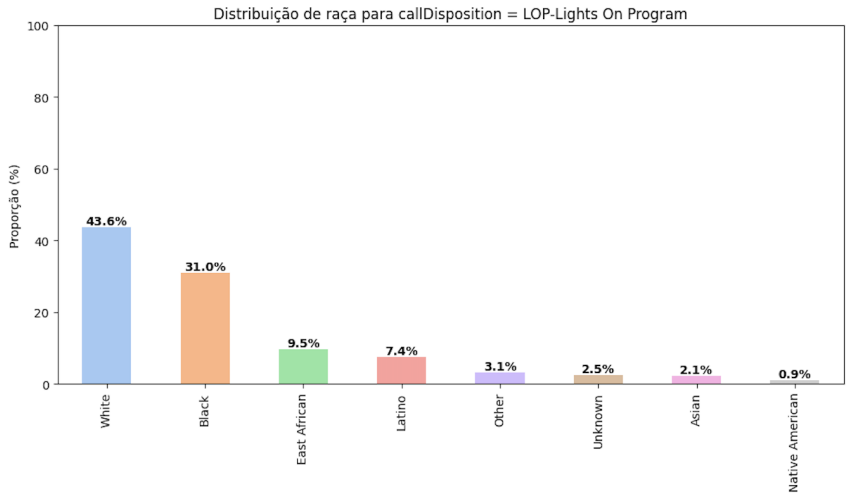

Figura 5.4: Distribuição racial das abordagens policiais que resultaram em LOP.

O LOP **(Lights On Program)**, que visa oferecer apoio sob a forma de vales para a reparação de viaturas em vez da aplicação de sanções, poderia ser interpretado como uma medida mais imparcial e orientada para a reabilitação. No entanto, a distribuição racial

dos beneficiários revela um padrão. Do total dos condutores de veículos abrangidos pelo programa, a maioria é constituída por indivíduos brancos **(43.6%)**, enquanto os condutores de veículos negros representam apenas **(31.0%)**, apesar de serem mais frequentemente alvo de abordagens policiais. Esses resultados indicam que, mesmo em iniciativas para apoiar em vez de penalizar, persistem desigualdades raciais. Os condutores de veículos, indivíduos brancos que foram abordados, continuam a ser, proporcionalmente, mais beneficiados. Assim, o viés racial mantém-se presente não apenas através da penalização desproporcional, mas também no acesso desigual aos próprios benefícios.

5.1.2 Conjunto de dados: Desempenho Académico

Para os dados recolhidos no contexto do ensino online, as variáveis analisadas incluem variáveis categóricas, como *Academic Performance* e *Stress Level*, bem como variáveis numéricas, como *Screen Time* e *Sleep Duration*, que foram convertidas em variáveis categóricas para permitir a sua análise adequada. Neste conjunto de dados, *Academic Performance* é considerada como variável independente, uma vez que se pretende compreender as diferenças significativas na forma como cada um destes fatores influencia o desempenho académico durante o ensino online.

Neste processo, os valores originais de *Sleep Duration* superiores a oito horas (>8h) foram convertidos para a categoria Excessivo, e os valores de *Screen Time* entre cinco e oito horas (5–8h) foram agrupados na categoria Moderado.

A Figura 5.5 apresenta a distribuição do desempenho académico entre os estudantes com um nível de *stress* classificado como baixo.

Analisando o gráfico da Figura 5.5, observa-se uma distribuição em que do total dos estudantes, a maioria **(40.1%)** manteve o seu rendimento durante o período de ensino à distância. Este resultado sugere que o nível de stress baixo no ensino online, por si só, não teve um impacto significativo no desempenho académico deste grupo de estudantes. Embora um nível de stress equilibrado possa contribuir para um ambiente mental mais saudável e propício à aprendizagem, continuam a existir vários fatores que podem gerar

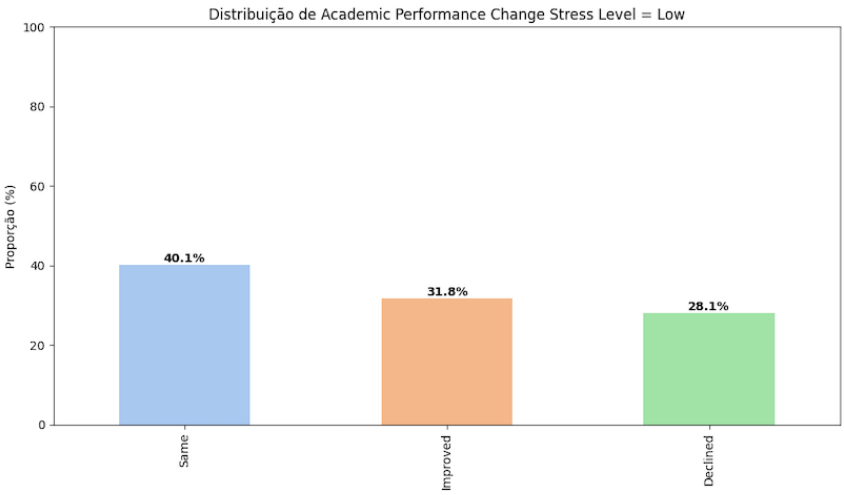

Figura 5.5: Desempenho académico com stress baixo.

stress e influenciar o desempenho académico [2]. Contudo, ao alargar a análise ao grupo, o **teste do Qui-Quadrado** apresentou um valor de $\chi^2 = 1.61$ e de $p - valor = 0.8075$, o que significa que $p > 0.01$ e, consequentemente, não se rejeita a hipótese nula. Estes resultados indicam que não existem vieses e que não há evidência de que o stress tenha influenciado o desempenho académico.

A Figura 5.6 representa a distribuição do desempenho académico entre os estudantes que indicaram dormir, em média, sono excessivo por noite.

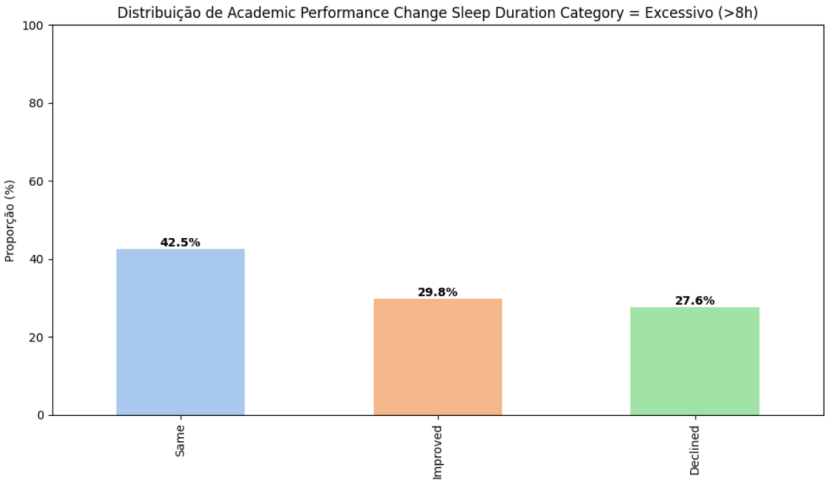

Figura 5.6: Desempenho académico com hora do sono excessivo(>8).

Observando o gráfico da Figura 5.6, verifica-se uma tendência positiva, com uma

maior proporção de alunos que mantiveram ou melhoraram o seu rendimento durante o período de ensino à distância. Embora o sono seja um fator reconhecidamente importante, não constitui, por si só, o único determinante do rendimento escolar. Ainda assim, os resultados estão em consonância com o que é amplamente descrito na literatura científica e validado pela prática pedagógica, que referem que os estudantes com um padrão de sono regular e adequado tendem a apresentar melhor desempenho, beneficiando de maior capacidade de concentração, memorização e estabilidade emocional [56]. O **teste do Qui-Quadrado** apresentou um valor de $\chi^2 = 1.28$ e de $p-valor = 0.8656$, o que indica que $p > 0.01$ e, consequentemente, não se rejeita a hipótese nula. Assim, os dados não evidenciam a existência de enviesamentos que indiquem uma relação estatisticamente significativa entre o número médio de horas de sono e o desempenho académico dos estudantes.

A Figura 5.7 demonstra a distribuição dos estudantes que reportaram passar entre 5 e 8 horas diárias em frente a ecrãs.

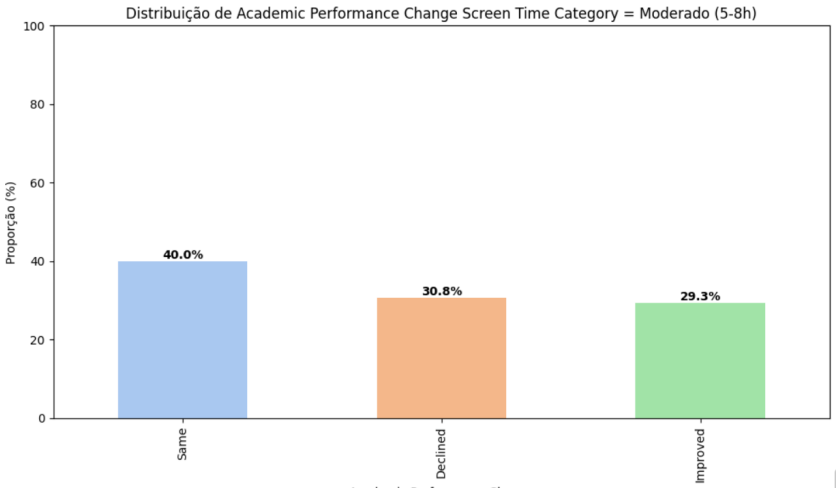

Figura 5.7: Desempenho académico para tempo de ecrã 5–8h/dia.

Analisando o gráfico da Figura 5.7, observa-se que do total dos estudantes, **(40,0%)** dos alunos manteve o mesmo nível de desempenho académico, enquanto **(30,8%)** registaram uma diminuição e **(29,3%)** verificaram uma melhoria no rendimento escolar. Estes resultados indicam que o efeito do tempo de ecrã varia, dado que proporções semelhantes de estudantes melhoraram ou pioraram o desempenho. Muitos alunos mantiveram

o mesmo nível de desempenho, sugerindo que um tempo de ecrã moderado (5–8 h/dia) não teve um impacto negativo relevante. A análise estatística, através do **teste do Qui-Quadrado** obteve-se $\chi^2 = 1.61$ e o $p-valor = 0.8072$, não revelou diferenças significativas entre os grupos, o que significa que $p > 0.01$ e, consequentemente, não se rejeita a hipótese nula. Assim, conclui-se que do ponto de vista estatístico, o tempo de ecrã declarado pelos alunos no ensino à distância não evidenciou um impacto negativo comprovado no desempenho académico.

5.2 ANOVA

A análise de variância (ANOVA) também é complementada pelo p-valor, que indica a probabilidade de se observar uma diferença tão grande entre as médias dos grupos, assumindo que a hipótese nula é verdadeira. Adotou-se um nível de significância de 1% (0.01). Para o teste, foram formuladas as seguintes hipóteses:

- $\mathbf{H_0}$**:** Não existem diferenças significativas entre as médias dos grupos.
- $\mathbf{H_1}$**:** Existem diferenças significativas entre as médias dos grupos.

O p-valor obtido é comparado com o nível de significância definido. Quando é < 0.01, rejeita-se a hipótese nula, indicando diferenças estatisticamente significativas entre as médias dos grupos. Caso contrário, não há evidência suficiente para rejeitar a hipótese nula.

5.2.1 Conjunto de dados: Minneapolis

De seguida, apresentam-se os gráficos e a respetiva análise, elaborados com base nos dados recolhidos durante operações policiais em Minneapolis, onde foram avaliadas diferenças estatisticamente significativas entre as médias dos diferentes grupos raciais. As variáveis analisadas foram *PersonSearch*, *VehicleSearch*, *CallDisposition*, sendo a raça a variável independente, cujo significado é explicado no Capítulo 4, Secção 4.1.

Como as variáveis *PersonSearch*, *VehicleSearch* e *CallDisposition* são originalmente categóricas, foi necessário convertê-las em variáveis numéricas para permitir a aplicação adequada da ANOVA. Esta transformação possibilitou o cálculo da média de ocorrência por grupo racial. A Figura 5.8 evidencia uma variação clara nas médias da variável *PersonSearch* entre os diferentes grupos raciais.

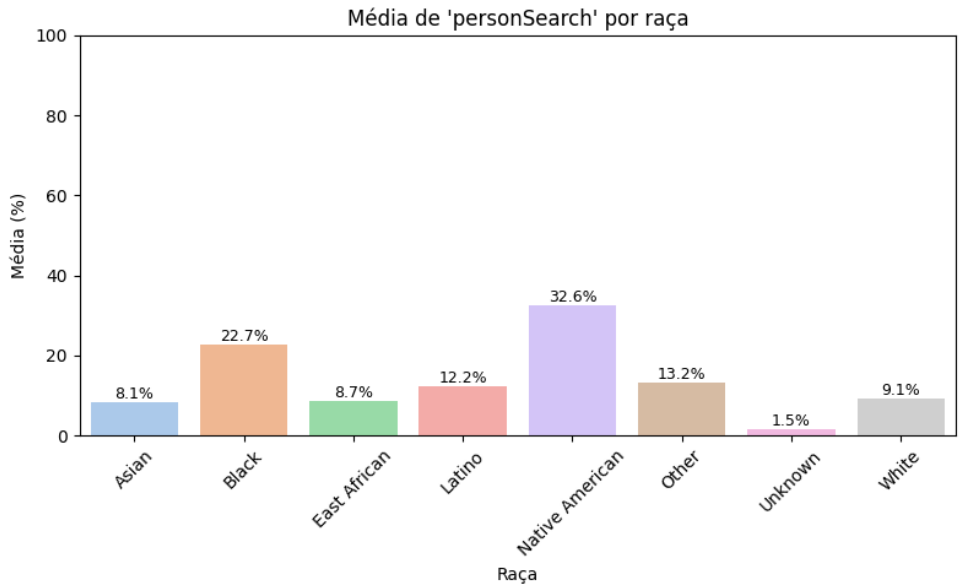

Figura 5.8: Distribuição racial da variável *PersonSearch*.

A análise das percentagens apresentadas no gráfico da Figura 5.8, mostra que determinados grupos são sujeitos a buscas pessoais com uma frequência superior à de outros. O grupo *Native American* apresenta a média mais elevada, com **(32,6%)**, seguido do grupo *Black*, com **(22,7%)**, valores superiores aos registados pelos restantes grupos. Além disso, o teste estatístico **ANOVA** confirma que estas diferenças entre as médias são estatisticamente significativas.

A análise revelou um valor de $F = 1583.26$ e um $p - valor = 0.0000$. Como $p < 0.01$, rejeita-se a hipótese nula de igualdade das médias, concluindo-se que os grupos raciais apresentam diferenças estatisticamente significativas na frequência de buscas pessoais.

A Figura 5.9 apresenta diferenças claras na média da variável *VehicleSearch* entre os diferentes grupos raciais.

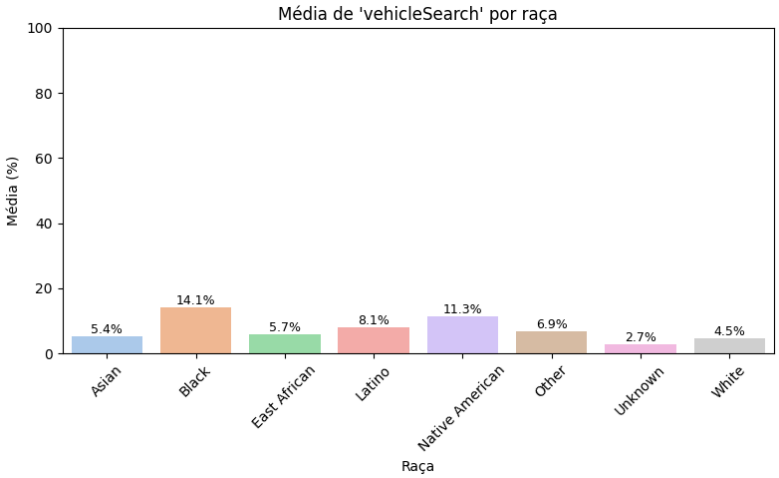

Figura 5.9: Distribuição racial da variável *vehicleSearch*.

O gráfico da Figura 5.9, demonstra que o grupo dos negros apresenta a média mais elevada com **(14.1%)**, seguido do grupo *Native American* com **(11.3%)**, com base nesta análise é possível indicar a existência de práticas diferenciadas nas abordagens policiais. A análise estatística **ANOVA** revelou um valor de $F = 681.90$ e de $p - valor = 0.0000$, dado que $p < 0.01$, rejeita-se a hipótese nula de igualdade das médias, concluindo-se que as diferenças observadas entre os grupos são estatisticamente significativas.

A Figura 5.10 apresenta diferenças nas médias da variável *callDisposition* entre os diversos grupos raciais.

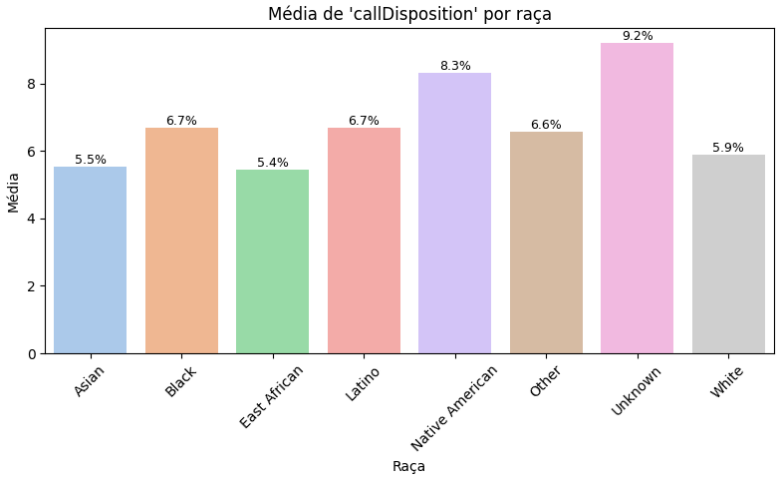

Figura 5.10: Distribuição racial da variável *callDisposition*.

Observando o gráfico da Figura 5.10 verifica-se que os grupos *Native American* e *Unknown* registam as médias mais elevadas, enquanto o grupo *East African* apresenta a média mais baixa. Esta variação apresenta diferenças nos desfechos das chamadas policiais relacionadas aos grupos raciais envolvidos. A análise estatística **ANOVA** revelou um valor de $F = 492.48$ e de $p - valor = 0.0000$. Sendo $p < 0.01$, rejeita-se a hipótese nula, confirmando que as diferenças observadas entre os grupos são estatisticamente significativas.

5.2.2 Conjunto de dados: Desempenho Académico

De seguida, apresentam-se os gráficos e a respetiva análise, realizados com base nos dados recolhidos no contexto do ensino online. As variáveis analisadas incluem tanto variáveis categóricas, *Academic Performance* e *Stress Level*, e variáveis numéricas, Screen Time e *Sleep Duration*. Como descrito no Capítulo 4, Secção 4.2, onde se explica o significado de cada uma das variáveis. Para permitir a realização da **ANOVA**, a variável *Stress Level* foi convertida em numérica. No presente conjunto de dados, *Academic Performance* é considerada a variável independente.

A Figura 5.11 mostra a média do nível de stress dos estudantes de acordo com a forma como o seu desempenho académico mudou.

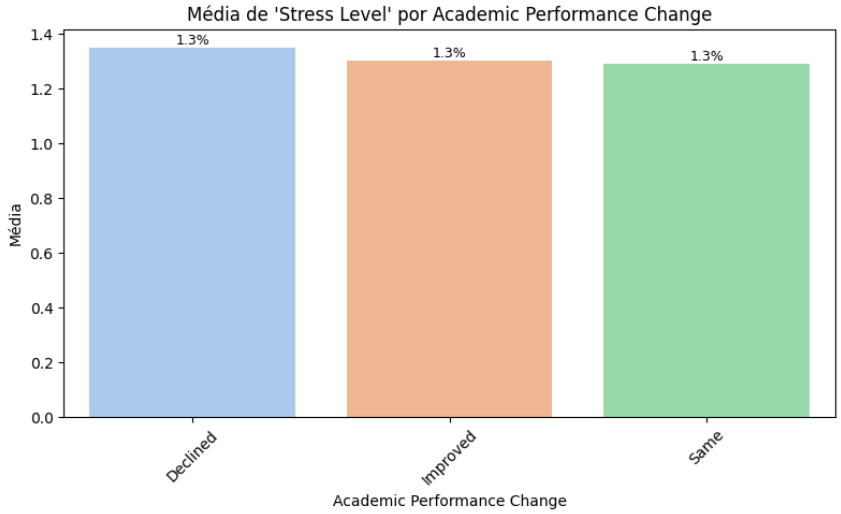

Figura 5.11: Média do nível de stress em função da mudança no desempenho académico.

Ao observar o gráfico da Figura 5.11, percebe-se que os três grupos apresentam valores muito semelhantes, todos próximos de **(1.3%)** variando apenas em algumas casas decimais que não estão representadas no gráfico. Esta semelhança indica que, em média, o stress não tem influência de forma clara no desempenho académico dos estudantes. A análise estatística **(ANOVA)** confirmou essa observação inicial, revelando um valor de $F = 0.54$ e um $p - valor = 0.5804$. Como o p-valor é superior ao nível de significância adotado (p $>$ 0,01), não se rejeita a hipótese nula de igualdade das médias. Assim, conclui-se que não existem diferenças estatisticamente significativas nos níveis médios de *stress* entre estudantes com diferentes trajetórias de desempenho académico.

Entretanto, essa semelhança nas médias não exclui a possibilidade de que o *stress* afete o desempenho académico de cada estudante de forma diferente. A média é apenas um valor geral e não mostra como o *stress* está distribuído dentro de cada grupo. Ou seja, em cada grupo existem estudantes com níveis altos, médios e baixos de *stress*, e essa variabilidade individual não é captada quando analisamos apenas as médias.

A Figura 5.12 apresenta a média da duração diária do sono dos estudantes em função da mudança no seu desempenho académico.

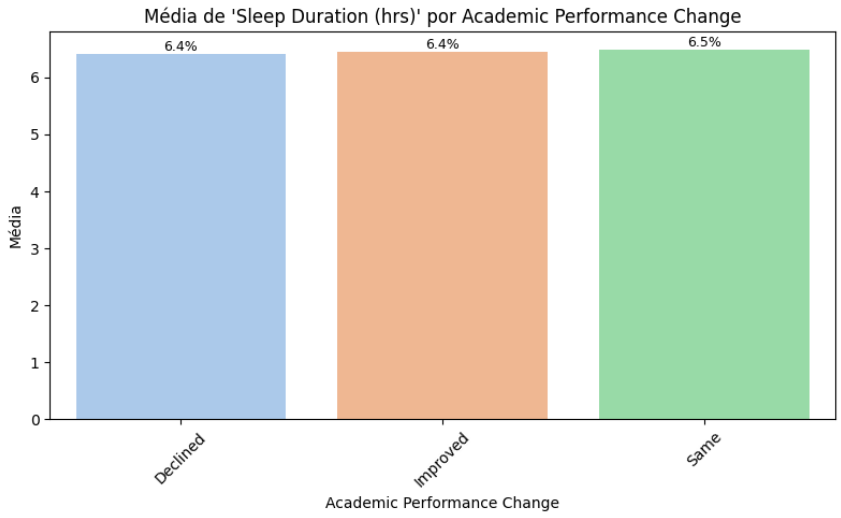

Figura 5.12: Média da duração do sono em função da mudança no desempenho académico.

Com o gráfico da Figura 5.12, observa-se que os três grupos exibem médias muito

próximas, variando ligeiramente entre **(6.4% e 6.5%)**. Essa proximidade sugere que, de forma geral, a quantidade média de sono dos estudantes pode não estar associada a alterações no desempenho académico. A análise estatística **ANOVA** confirma essa interpretação, revelando $F = 0.28$ e de p-valor de 0.7589. Como p-valor é superior ao nível de significância adotado ($p > 0.01$), não se rejeita a hipótese nula de igualdade das médias. Assim, conclui-se que não existem diferenças estatisticamente significativas na duração média do sono entre os estudantes. Apesar disso, é importante entender que a média reflete apenas um valor global e não mostra como a duração do sono está distribuída dentro de cada grupo. Em todos os grupos encontram-se estudantes que dormem mais, outros que dormem menos e outros que mantêm um padrão.

A Figura 5.13 apresenta a média do tempo de ecrã (horas/dia) em função da variação no desempenho académico.

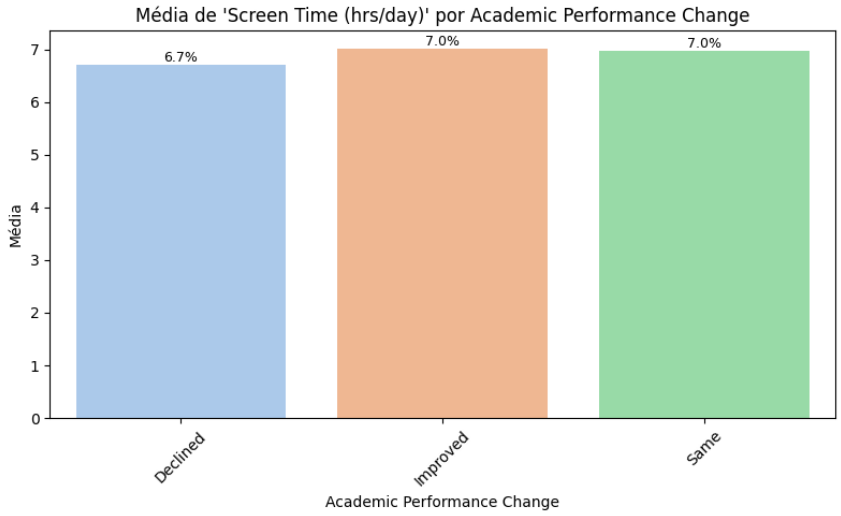

Figura 5.13: Média do tempo de ecrã (horas/dia) em função da mudança no desempenho académico.

Através da análise do gráfico presente na Figura 5.13, observa-se que os três grupos apresentam médias muito próximas, oscilando entre aproximadamente (6.7% e 7%). Essa semelhança indica que a alteração no desempenho académico não está associada a diferenças relevantes no tempo médio de exposição ao ecrã. A análise ANOVA revelou um valor de $F = 1.01$ e $p-valor = 0.3649$. Como o p-valor é superior ao nível de significância

adotado $p > 0.01$, não se rejeita a hipótese nula de igualdade das médias. Conclui-se, portanto, que não existem diferenças estatisticamente significativas no tempo médio de ecrã entre os grupos.

Entretanto, a semelhança das médias não exclui a possibilidade de que o tempo de exposição ao ecrã possa afetar o desempenho académico de cada estudante de maneira distinta. Além disso, a média do tempo total de exposição não diferencia o tipo ou a qualidade do uso, que pode variar entre atividades académicas produtivas e momentos de lazer ou distração. Mesmo com médias semelhantes, a forma como cada estudante utiliza o tempo em frente ao ecrã pode influenciar o desempenho académico de maneiras diferentes.

5.3 Balanceamento dos Dados

Nesta secção, serão analisados e discutidos os resultados obtidos através do balanceamento dos dados. Para avaliar o desempenho das técnicas exploradas, foi aplicado o método *Random Oversampling*, que replica exemplos das classes menos representadas até que todas tenham a mesma contagem. Esse procedimento de balanceamento foi aplicado tanto para o teste estatístico Qui-Quadrado quanto para o teste ANOVA. Os resultados serão apresentados de forma detalhada nas tabelas a seguir, contendo os valores dos testes ANOVA e Qui-Quadrado antes e depois do balanceamento.

5.3.1 Teste do Qui-Quadrado

Para o **conjunto de dados Minneapolis**, considerando as variáveis *PersonSearch*, *VehicleSearch* e *Call-Disposition = RPT-Report*, foi aplicado o *Random Oversampling* para equilibrar a variável *Race*. A Tabela 5.1 apresenta os resultados antes e depois do balanceamento dos dados.

Analisando a Tabela 5.1, observa-se que o teste do Qui-Quadrado depois de balanceamento para todas as variáveis apresentou um resultado de $\chi^2 = 0.00$ e $p-valor = 1.0000$). Sendo o p-valor superior ao nível de significância 0.01, não se rejeita a hipótese nula, o

Tabela 5.1: Distribuição do χ^2 antes e depois de balanceamento da operação policial.

	Antes		**Depois**	
	χ^2	p	χ^2	p
PersonSearch=Yes	4618.14	0.00	0.00	1.0000
VehicleSearch=Yes	10280.14	0.00	0.00	1.0000
CallDisposition = RPT-Report	81080.81	0.00	0.00	1.0000

que confirma a ausência de associação estatisticamente significativa entre as variáveis no conjunto balanceado e valida a eficácia do método aplicado. Antes do balanceamento, o p-valor era inferior a 0.01, indicando associação significativa, contudo, após o processo, o valor de χ^2 tornou-se zero.

Este resultado é esperado, uma vez que o teste do **Qui-Quadrado** analisa a diferença entre as frequências observadas e as esperadas. Quando os dados são balanceados através de técnicas como o *Random Oversampling*, são criados valores sintéticos que fazem com que as frequências observadas coincidam exatamente com as esperadas, desta forma, todas as classes passam a ter a mesma quantidade de observações, o que elimina diferenças entre os grupos e resulta num valor de $\chi^2 = 0$, confirmando a ausência de vieses no conjunto de dados balanceado. Dessa forma, valida-se o bom desempenho do teste do **Qui-Quadrado** para identificar viés neste caso.

Este procedimento também foi aplicado para o **conjunto de dados do Desempenho Académico**, considerando as variáveis, *Stress Level Low*, *Sleep* (>8h) e *Screen Time* (5–8h). Na Tabela 5.2 são apresentados os resultados do teste do Qui-Quadrado aplicado antes e depois do balanceamento.

Tabela 5.2: Distribuição do χ^2 antes e depois de balanceamento do desempenho académico.

	Antes		**Depois**	
	χ^2	p	χ^2	p
Stress Level, Low	1.61	0.8075	0.00	1.0000
Sleep, (>8h)	1.28	0.8656	0.00	1.0000
Screen Time,(5-8h)	1.61	0.8072	0.00	1.0000

Antes do balanceamento, o valor do p era superior ao nível de significância de 0.01, indicando que não existiam vieses estatisticamente significativos nas variáveis analisadas. No entanto, pequenas diferenças entre as frequências observadas estavam presentes. Após o balanceamento, o valor do **Qui-Quadrado** passou a ser $\chi^2 = 0.00$, com um p-valor de 1.0000, eliminando totalmente qualquer diferença entre as categorias. Este resultado demonstra que o processo de balanceamento igualou o número de observações em cada grupo, removendo qualquer associação entre as variáveis e garantindo que os dados fiquem equilibrados para análises futuras, confirmando o bom desempenho do Qui-Quadrado a identificar o viés.

5.3.2 ANOVA

Esta análise de desempenho também foi realizada para a ANOVA no **conjunto de dados Minneapolis** considerando as variáveis *PersonSearch*, *VehicleSearch* e *Call-Disposition*. Dessa forma, a Tabela 5.3 compara os valores obtidos da aplicação da ANOVA antes e após o processo de balanceamento do conjunto de dados.

Tabela 5.3: Distribuição da ANOVA antes e depois de balanceamento da operação policial

	Antes		**Depois**	
	F	p	F	p
PersonSearch	1583,26	0.00	1.39	0.2049
VehicleSearch	681.90	0.00	1.86	0.0712
CallDisposition	492.48	0.00	1.63	0.1221

Analisando a Tabela 5.3, observa-se que após o balanceamento, os valores do p foram diferentes para cada categoria, mas mantiveram-se superiores ao nível de significância de 0,01, demonstrando que não existem diferenças estatisticamente significativas entre os grupos. A **ANOVA**, que compara as médias dos grupos, evidencia que, mesmo após o balanceamento, as médias não são exatamente iguais, devido a pequenas variações naturais nos dados amostrais. Este resultado confirma que o balanceamento igualou o número de observações em cada grupo, mantendo apenas pequenas diferenças, e garante que os

dados se encontram equilibrados para análises futuras, sem introduzir vieses significativos.

Para o **conjunto de dados do Desempenho Académico**,considerando as variáveis, *Stress Level*, *Sleep* e *Screen Time*, também analisou-se a comparação da aplicação da ANOVA no conjunto de dados antes e após balancemaneto dos dados, como pode ser observável na Tabela 5.4.

Tabela 5.4: Distribuição da ANOVA antes e depois de balanceamento do desempenho académico

	Antes		**Depois**	
	F	p	F	p
Stress Level	0.54	0.5804	0.37	0.6910
Sleep Duration	0.28	0.7589	0.32	0.7275
Screen Time	1,01	0,3649	2.33	0.0979

Com a Tabela 5.4, é possível retirar que antes do balanceamento, os resultados já indicavam que não existiam diferenças estatisticamente significativas entre os grupos, o que mostra que os dados estavam praticamente equilibrados. Depois do balanceamento, observaram-se pequenas alterações nos valores de p em alguns casos ficaram ligeiramente menores e, noutros, um pouco maiores. Este resultado é esperado, pois o balanceamento cria novas observações sintéticas nas categorias com menos dados, o que pode provocar pequenas diferenças nos resultados obtidos. Embora o número de observações tenha ficado proporcional entre os grupos, as médias não se tornam exatamente iguais, já que é natural existirem pequenas diferenças entre os dados. Estas alterações nos valores de p não significam que tenham surgido novos vieses, mas apenas que ocorreram ajustamentos normais durante o processo de equilíbrio dos dados. De forma geral, os resultados confirmam que os dados permanecem equilibrados, sem diferenças estatisticamente significativas entre os grupos analisados.

5.4 Discussão de Resultados

A análise dos dois conjuntos de dados permitiu compreender como os testes estatísticos, Qu i-Quadrado e ANOVA, contribuem para identificar e minimizar viés.

No conjunto de dados de operação policial, os testes estatísticos revelaram viés racial evidente. O Qui-Quadrado apresentou valores de χ^2 elevados e p-valor igual a zero, inferior ao nível de significância 0.01, para todas as variáveis (*PersonSearch, VehicleSearch* e *CallDisposition*), enquanto a ANOVA indicou p-valor < 0.01, mostrando diferenças significativas entre as médias dos grupos raciais. Estes resultados confirmam que certos grupos foram desproporcionalmente abordados, evidenciando desigualdades significativas.

No conjunto de dados desempenho académico, por outro lado, as variáveis *Stress Level, Sleep Duration* e *Screen Time* não apresentaram diferenças significativas entre os grupos, com p-valores superiores ao nível de significância 0.01 em ambos os testes, Qui-Quadrado e ANOVA, sugerindo que estes fatores não influenciam de forma relevante o desempenho académico dos estudantes.

Após a aplicação do *Random Oversampling*, verificou-se que os testes refletiram o efeito do balanceamento. No Qui-Quadrado, os valores passaram a $\chi^2 = 0$ e $p-valor = 1$, validando o desempenho do método aplicado. Na ANOVA, os resultados diferiram ligeiramente dos valores originais devido a pequenas variações nas médias, mas a interpretação geral manteve-se. A ANOVA revelou-se mais adequada para comparar médias de variáveis numéricas, enquanto o Qui-Quadrado mostrou-se mais simples de interpretar, rápido de executar e útil para identificar padrões em variáveis categóricas.

Com os resultados alcançados, os dois testes estatísticos, ANOVA e Qui-Quadrado, cumpriram o objetivo de identificar vieses, e a técnica de balanceamento *Random Oversampling* cumpriu o objetivo de minimizar o viés nos dois conjuntos de dados, validando a eficácia dos dois testes estatísticos.

Capítulo 6

Conclusões e Trabalhos Futuros

O presente trabalho teve como principal objetivo identificar e analisar a presença de viés em conjuntos de dados frequentemente utilizados, em contextos sociais relevantes, como as abordagens policiais e o desempenho académico de estudantes.

Este estudo demonstrou que, através da aplicação de testes estatísticos ANOVA e Qui-Quadrado, foi possível detetar diferenças significativas associadas a variáveis, especialmente nos dados relacionados com operações policiais, revelando vieses raciais. Já no desempenho académico, fatores como stress, sono e tempo de tela não influenciaram de forma relevante.

A aplicação do balanceamento dos dados por meio do *Random Oversampling* eliminou esses vieses, permitindo análises estatísticas mais fiáveis. Os resultados dos testes Qui-Quadrado e ANOVA revelaram que, após o balanceamento, as diferenças significativas entre grupos desaparecem, confirmando a eficácia do procedimento.

É importante destacar que o viés em conjuntos de dados não é apenas um problema técnico, mas também ético e social. Modelos treinados com dados enviesados podem reproduzir ou até amplificar desigualdades estruturais existentes na sociedade. Por isso, a construção e utilização de sistemas inteligentes devem ser acompanhadas de responsabilidade, transparência e consciência crítica.

Este estudo pretende, assim, contribuir para uma utilização mais justa dos dados e para aprofundar o conhecimento sobre métodos de identificação e mitigação do viés.

Tendo em conta os resultados obtidos, identificam-se várias linhas de investigação promissoras que poderão ser exploradas em projetos futuros. Destaca-se a deteção de viés em sistemas baseados em imagem, aplicando os mesmos princípios de análise de viés a dados visuais (imagens ou vídeo). Em vez de analisar apenas tabelas com variáveis numéricas ou categóricas, seria possível estudar algoritmos de reconhecimento facial, detetar emoções ou analisar expressões faciais, avaliando se funcionam de forma desigual consoante a raça, o género ou a idade das pessoas analisadas.

Bibliografia

[1] R. Agra, F. N. B. De Souza, P. C. Soares, and G. L. Da Silva. A comparative analysis between undersampling and oversampling approaches to data balancing. *IEEE Xplore*, 2023. Acesso em 2025.

[2] Ana Paula Amaral and Carlos Fernandes da Silva. Estado de saúde, stress e desempenho académico numa amostra de estudantes do ensino superior. *Revista Portuguesa de Pedagogia*, pages 111–133, 2008. Acedido em: 2025.

[3] A. Ashraf. Oversampling for better machine learning with imbalanced data. `https://medium.com/@abdallahashraf90x/oversampling-for-better-machine-learning-with-imbalanced-data-\68f9b5ac2696`, 2025. Acesso em 2025.

[4] J. Barreto. Oversampling ou undersampling: Qual método de balanceamento usar? `https://pt.linkedin.com/pulse/oversampling-ou-undersampling-qual-método-de-usar-joseferson-\barreto-dgwvf`, 2024. Acesso em 2025.

[5] R. Barroso, L.˙ Inteligência artificial: promessas, riscos e regulação. algo de novo debaixo do sol. *Revista de Desenvolvimento e Políticas Públicas*, 2024. Acesso em: 2025.

[6] Dave Bergmann. What is deeplearning?, 2025. Acesso em 2025.

[7] Rebecca Bevans. An introduction to t tests | definitions, formula and examples. https://www.scribbr.com/statistics/t-test/?utm_source=chatgpt.com, 2023. Revised June 22, 2023.

[8] Brasil Escola. Inteligência artificial. https://brasilescola.uol.com.br/informatica/inteligencia-artificial.htm, 2025. Acesso em 2025.

[9] C. Brown. Oversampling and undersampling explained: A visual guide with mini 2d dataset. https://medium.com/data-science/oversampling-and-undersampling-explained-a-visual-guide-with-mini-2d\-dataset-1155577d3091, 2025. Acesso em 2025.

[10] Sara Brown. Machine learning, explained. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained, Abril 2021. Acesso em 2025.

[11] A. Y. c. Liu. The effect of oversampling and undersampling on classifying imbalanced text datasets. Master's thesis, Universidade do Texas em Austin, 2004. Acesso em 2025.

[12] CEPED. Os muitos desafios da mitigação de vieses no desenvolvimento de algoritmos. https://medium.com/o-centro-de-ensino-e-pesquisa-em-inovaç~ao-está/os-muitos-desafios-da-mitigaç~ao-de-vieses-no-desenvolvimento-de-\algoritmos-83a48888e734, 2023. Acesso em 2025.

[13] Chrystal R. China. Tipos de machine learning (ml). https://www.ibm.com/br-pt/think/topics/machine-learning-types, 2025. Acesso em: 2025.

[14] L. S. Coradeli and D. R. Pereira. Comparação entre métodos de aumento de dados desbalanceados para inteligência artificial. Trabalho de Conclusão de Curso, Universidade Estadual Paulista "Júlio de Mesquita Filho", 2023. Acesso em 2025.

[15] Coursera. What is machine learning? definition, types, and examples. https://www.coursera.org/articles/what-is-machine-learning, Fevereiro 2025. Acesso em 2025.

[16] Aline Alves da Silva. *Vieses em Modelos de Redes Neurais do Setor Financeiro: uma análise comparativa entre sistemas de IA no Brasil e de Portugal*. PhD thesis, Pontifícia Universidade Católica de São Paulo (PUC-SP) e Universidade de Coimbra, 2025.

[17] Conselho da União Europeia. Benefícios e riscos da inteligência artificial. https://www.consilium.europa.eu/pt/policies/benefits-and-risks-of-ai/, Fevereiro 2025. Acesso em 2025.

[18] Leandro de Azevedo Gonzalez. Regressão logística e suas aplicações. *Universidade Federal do Maranhão (UFMA) – Monografia de Graduação*, 2018.

[19] Bruno Fediuk de Castro and Gilberto Bomfim. A inteligência artificial, o direito e os vieses. *Revista Ilustração*, 1(3):31–45, 2020. set.-dez.

[20] André Carlos Ponce de Leon Ferreira de Carvalho. Inteligência artificial: riscos, benefícios e uso responsável. *Estudos Avançados*, 35(101):21–36, 2021.

[21] Distrito. Inteligência artificial (ia): o que é? entenda tudo sobre o assunto. https://distrito.me/blog/inteligencia-artificial-ia-o-que-e-entenda-tudo-sobre-o-assunto/, 2025. Acesso em 2025.

[22] Autor(es) do artigo. Título do artigo. *Journal of Personalized Medicine*, 11(32):..., 2021. Acedido em: 2025.

[23] Universidade Federal do Paraná. Teste do qui-quadrado. http://www.leg.ufpr.br/lib/exe/fetch.php/disciplinas:ce001:teste_do_qui-quadrado.pdf, 2025. Acesso em 2025.

[24] Pedro H. C. dos Santos, Leandro S. G. Carvalho, Elaine H. T. Oliveira, and David B. F. de Oliveira. Classificação de dificuldade de questões de programação com base na inteligibilidade do enunciado. *Anais do XXX Simpósio Brasileiro de Informática na Educação (SBIE 2019)*, pages 1886–1895, 2019. Acesso em 2025.

[25] Parlamento Europeu e Conselho da União Europeia. Regulamento (ue) 2024/1689 do parlamento europeu e do conselho de 13 de junho de 2024 que estabelece regras harmonizadas sobre inteligência artificial e altera os regulamentos, 2024.

[26] Ghada A. El Refae, Abdallah Kaba, and Shadi Eletter. The impact of demographic characteristics on academic performance: Face-to-face learning versus distance learning implemented to prevent the spread of covid-19. *International Review of Research in Open and Distributed Learning*, 22(1):91–110, 2021.

[27] Estatística Fácil. O que é análise de vieses estatísticos. `https://estatisticafacil.org/glossario/o-que-e-analise-de-vieses-estatisticos/`, 2025. Acesso em 2025.

[28] EstatísticaFácil. O que são testes estatísticos – guia completo, 2025. Acesso em 2025.

[29] Parlamento Europeu. O que é a inteligência artificial e como funciona? `https://www.europarl.europa.eu/topics/pt/article/20200827STO85804/o-que-e-a-inteligencia-artificial-e-como-funciona`, Setembro 2020. Atualizado em junho de 2023.

[30] FM2S. Anova: o que é e como utilizar. `https://www.fm2s.com.br/blog/anova`, July 2024.

[31] Ana Cristina Bicharra Garcia. Ética e inteligência artificial. *Computação Brasil*, Novembro 2020.

[32] Amanda Munari Guimarães. Análise de variância (anova) one-way e tukey usando r. `https://medium.com/omixdata/`

análise-de-variância-anova-one-way-e-tukey-usando-r-f91b6f79240e, 8 2019.

[33] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, September 2009.

[34] Kamal Kant Hiran, Ritesh Kumar Jain, Kamlesh Lakhwani, and Ruchi Doshi. *Machine Learning: Master Supervised and Unsupervised Learning Algorithms ....* BPB Publications, 2023.

[35] James Holdsworth. O que é viés de ia? https://www.ibm.com/br-pt/think/topics/ai-bias, Dezembro 2023. Acesso em 2025.

[36] IBM. Ai vs. machine learning vs. deep learning vs. neural networks, 2023. Acedido em: 2025.

[37] IBM. Viés algorítmico: Compreendendo e mitigando o viés na inteligência artificial. https://www.ibm.com/br-pt/think/topics/algorithmic-bias, 2024. Acesso em 2025.

[38] P. Iglecias. Inteligência artificial (ia) e dano ambiental, 2024. Acesso em: 10 Nov 2025.

[39] Kaggle. Discussão sobre problemas com desbalanceamento de classes. https://www.kaggle.com/discussions/general/520989, 2025. Acesso em 2025.

[40] Kaggle. Discussão sobre técnicas de balanceamento de dados. https://www.kaggle.com/discussions/general/424515, 2025. Acesso em 2025.

[41] Tae Kyun Kim. T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6):540–546, 2015.

[42] Jakub Kufel, Katarzyna Bargieł-Łączek, Szymon Kocot, Maciej Koźlik, Wiktoria Bartnikowska, Michał Janik, Łukasz Czogalik, Piotr Dudek, Mikołaj Magiera, Anna

Lis, et al. What is machine learning, artificial neural networks and deep learning?—examples of practical applications in medicine. *Diagnostics*, 13(15):2582, 2023.

[43] LinkedIn. Como detetar vieses em sistemas de ia. `https://pt.linkedin.com/advice/0/how-do-you-detect-bias-ai-systems-skills-artificial-\intelligence?lang=pt`, 2025. Acesso em 2025.

[44] Thiago Ruiz Lobo and Cláudia Aparecida Martins. Comparativo de algoritmos de aprendizado de máquina para a classificação de notícias sobre a politec em mato grosso. In *Anais do SIBGRAPI – Simpósio Brasileiro de Computação Gráfica e Processamento de Imagens*, 2024.

[45] Mailchimp. Viés no aprendizado de máquina. `https://mailchimp.com/pt-br/resources/bias-in-machine-learning/?utm_source=chatgpt.com`, 2024. Acesso em 2025.

[46] Isnewati Ab Malek, Amira Atiqah Mohd Nazri, Nur Iffa Aisha Jamshah, Nur Syarah Md Hashim, and Haslinda Ab Malek. Factors affecting students' academic performance through online distance learning. *Journal of Exploratory Mathematical Undergraduate Research (JEMUR)*, 2:35–42, 2024.

[47] ManageEngine. Cuidado com estes 4 riscos do aprendizado de máquina. `https://blogs.manageengine.com/corporate/general/2024/04/26/cuidado-com-estes-4-riscos-do-aprendizado-de-maquina.html`, Abril 2024. Acesso em 2025.

[48] Bernard Marr. Os 15 maiores riscos da inteligência artificial. *Forbes Brasil*, 2023. Acesso em 2025.

[49] MDPI. Análise de algoritmos e implicações Éticas em inteligência artificial. *Information*, 14(1):54, 2024. Acesso em 2025.

[50] Gabriel Damazio Nobre Mendes. Apoio tecnológico no atendimento supermercadista: mitigando temores e entendendo a satisfação do cliente, 2023. Orientação: Prof. Dr. Rodrigo Parron Santos. Disponível em: https://repositorio.ufrn.br/server/api/core/bitstreams/726f7430-87fb-49eb-8ec7-4e3fd3cbdcfa/content.

[51] Mind the Graph. Teste de qui-quadrado: Como e quando usá-lo. https://mindthegraph.com/blog/pt/chi-square-test/, 2025. Acesso em 2025.

[52] David S. Moore, William I. Notz, and Michael A. Fligner. *The Basic Practice of Statistics*. W. H. Freeman and Company, New York, NY, 7 edition, 2015. Acesso em 2025.

[53] Chloe Ng. How has interest and coverage around ai changed since chatgpt?, May 2023.

[54] Observador. Que impacto tem a inteligência artificial no nosso dia a dia? https://observador.pt/opiniao/que-impacto-tem-a-inteligencia-artificial-no-nosso-dia-a-dia/, Setembro 2020. Acesso em 2025.

[55] T. Onookome-Okome. Characterizing patterns in police stops by race in minneapolis. *PLOS ONE*, 17(5):e0267564, 2022.

[56] Ana Perdigão, Andreia Cristina, and Filipe Sousa. Qualidade do sono e sonolência em estudantes de enfermagem. *RevSALUS*, 2023. Acedido em 2025.

[57] Predize. Tipos de aprendizado de máquina. https://predize.com/blog/tipos-de-aprendizado-de-maquina/, 2024. Acesso em 2025.

[58] Aleff Fonseca Reis, Handerson Leonidas Sales, André Luiz Mendes Athayde, and Lucinéia Lopes Bahia Ribeiro. Vieses do comportamento financeiro dos estudantes de administração da universidade federal de minas gerais - campus montes claros. *Revista Ciências Administrativas*, 21(1):195–223, 2024. Acesso em 2025.

[59] Leonardo Rodrigues. O que é o teste anova? aprenda calcular e quando utilizar! https://voitto.com.br/blog/artigo/anova, 8 2019.

[60] Jose Alexandre Sa. O impacto da inteligencia artificial na economia. https://www.ordemeconomistas.pt/xportalv3/file/XEOCM_Documento/74260924/file/Jose%20Alexandre%20Sa.pdf, 2025. Acesso em 2025.

[61] Aline Bessa Sampaio. Inteligência artificial no poder judiciário: Análise do risco de vieses algorítmicos em decisões judiciais. Master's thesis, Universidade Federal do Ceará, 2025. 2025.

[62] SAP. What is ai bias? https://www.sap.com/resources/what-is-ai-bias, 2023. Acesso em 2025.

[63] SBOC. Leitura crítica - c4. https://www.sboc.org.br/app/webroot/leitura-critica/LEITURA-CRITICA_C4.pdf, 2024. Acesso em 2025.

[64] Lisiane Priscila Roldão Selau and José Luis Duarte Ribeiro. Uma sistemática para construção e escolha de modelos de previsão de risco de crédito. *Gestão & Produção*, 16(3):398–413, 2009.

[65] Erick Morais de Sena. Viés na ia: como o viés algorítmico influencia na perpetuação de estereótipos e desigualdades existentes. Artigo (Bacharelado em Ciência da Computação) – Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande, 2023. 12f.

[66] Utkarsh Sharma. Student mental health analysis. https://www.kaggle.com/datasets/utkarshsharma11r/student-mental-health-analysis, 2023. Acesso em 2025.

[67] M. Shelke. A review on imbalanced data handling using undersampling and oversampling technique. https://www.academia.edu/71149653/A_Review_on_Imbalanced_Data_Handling_Using_Undersampling_and_Oversampling_Technique, 2017. Acesso em 2025.

[68] J. Silva, M. Pereira, and L. Santos. Análise comparativa de técnicas de reamostragem para dados desequilibrados. *Revista de Inteligência Computacional*, 15(3):123–130, 2019. Acesso em 2025.

[69] Alessandro Simonetta, Andrea Trenta, Maria Cristina Paoletti, and Antonio Vetrò. Metrics for identifying bias in datasets. In *Proceedings of the AI Ethics and Society Workshop*, volume 3118 of *CEUR Workshop Proceedings*, pages 1–12.

[70] Shubham K Singh. Racial biased police & violence: Minneapolis. `https://www.kaggle.com/code/shubhamksingh/racial-biased-police-violence-minneapolis/input`, 2021. Acesso em: 15 nov. 2025.

[71] Robert Smith. Ai bias: Why it happens and how to address it. `https://robertsmith.com/blog/ai-bias/`, 2024. Acesso em 2025.

[72] M. Syukrina and R. Nugraha. Artificial intelligence risk identification: Challenges, impacts, and mitigation strategies. *International Journal of Economics, Commerce and Business Enterprise*, 7(2):109–122, 2023. Acesso em: 2025.

[73] TECHLIVEN. Viés algorítmico na inteligência artificial. `https://pt.linkedin.com/pulse/viés-algorítmico-na-inteligência-artificial-techliven-ixirf`, 2025. Acesso em 2025.

[74] Pedro Gabriel Fernandes Vieira. Deep learning para identificação de mutações genéticas patogénicas. Master's thesis, Universidade de Aveiro, Aveiro, Portugal, 2017. Acesso em 2025.

[75] Jack Virag. Understanding significance levels: A key to accurate data analysis. Blog post on Statsig, jul 2024.

[76] Tarid Wongvorachan, Surina He, and Okan Bulut. A comparison of undersampling, oversampling, and SMOTE-NC methods for dealing with imbalanced classification in educational data mining. *Information*, 14(1):54, 2023.

[77] World Population Review. Minneapolis, minnesota population 2025. `https://worldpopulationreview.com/us-cities/minnesota/minneapolis`, 2025. Acesso em 2025.

[78] Shunan Zhang, Xiangying Zhao, Tong Zhou, and Jang Hyun Kim. Do you have ai dependency? *International Journal of Educational Technology in Higher Education*, 2024. Acesso em: 2025.