UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Karla Felícia Carvalho da Silva

Impacto das técnicas de equilíbrio de classes na privacidade: uma análise dos ataques de inferência de associação

São Luís - MA

2025

Karla Felícia Carvalho da Silva

Impacto das técnicas de equilíbrio de classes na privacidade: uma análise dos ataques de inferência de associação

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Programa de Pós-Graduação em Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Luciano Reis Coutinho

Coorientador: Prof. Dr. Antonio de Abreu Batista-Jr

São Luís - MA

2025

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

da Silva, Karla Felicia Carvalho.
Impacto das técnicas de equilíbrio de classes na privacidade : uma análise dos ataques de inferência de associação / Karla Felicia Carvalho da Silva. - 2025.
53 f.

Coorientador(a) 1: Antonio de Abreu Batista Júnior.
Orientador(a): Luciano Reis Coutinho.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, São Luís, 2025.

1. Aprendizado de Máquina. 2. Ataque de Infereência. 3. Justiça. 4. Conjunto de Dados Desbalanceado. I. Batista Júnior, Antonio de Abreu. II. Reis Coutinho, Luciano. III. Título.

Karla Felícia Carvalho da Silva

Impacto das técnicas de equilíbrio de classes na privacidade: uma análise dos ataques de inferência de associação

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Trabalho defendido. São Luís - MA, 29 de agosto de 2025:

Prof. Dr. Luciano Reis Coutinho
Orientador
Universidade Federal do Maranhão

Prof. Dr. Antonio de Abreu Batista-Jr
Coorientador
Universidade Federal do Maranhão

Prof. Dr. Tiago Bonini Borchartt
Examinador Interno
Universidade Federal do Maranhão

Prof. Dr. Vinicius Ponte Machado
Examinador Externo
Universidade Federal do Piauí

São Luís - MA
2025

,

Dedico este trabalho ao meu esposo, José Emanuel, por me inspirar a iniciar esta jornada, e a Vicente, por me dar forças para persistir até o final.

Agradecimentos

Primeiramente, agradeço a Deus por me conceder saúde e coragem para não desistir durante a realização deste trabalho.

Agradeço ao meu esposo, José Emanuel, por toda ajuda, direta e indireta. Sua dedicação e apoio foram fundamentais para a concretização deste trabalho.

Aos meus orientadores, Luciano Reis Coutinho, pela paciência, e Antônio de Abreu Batista Júnior, pela valiosa ajuda, expresso minha sincera gratidão.

Agradeço também ao amigo de laboratório, Bruno Roberto Silva, pela motivação.

"As dificuldades preparam pessoas comuns para destinos extraordinários."

(C.S. Lewis)

Resumo

O uso de modelos de aprendizado de máquina treinados com conjuntos de dados desbalanceados que contêm informações confidenciais tem levantado preocupações em matéria de privacidade. Uma dessas preocupações é a vulnerabilidade ao ataque de inferência de associação (Membership Inference Attack - MIA), que revela se um ponto de dados específico foi ou não utilizado para treinar um modelo. Neste contexto, uma área que tem sido negligenciada é se o método para lidar com o desequilíbrio de classes para treinar classificadores com algoritmos de aprendizagem supervisionada padrão que esperam um cenário com classes equilibradas aumenta o vazamento de informações sensíveis. Para preencher esta lacuna, realizámos uma série de avaliações empíricas de MIA com conjuntos de dados desequilibrados do mundo real para avaliar esta questão. Os nossos resultados mostram que os métodos para garantir o equilíbrio do classificador podem aumentar significativamente o desempenho dos ataques MIA. Isto tem implicações para os cientistas que estudam as defesas contra ataques nestes cenários com conjuntos de dados desequilibrados.

Palavras-chave: Aprendizado de Máquina, Ataque de Infereência, Justiça, conjunto de dados desbalanceados.

Abstract

The use of machine learning models trained on unbalanced datasets with sensitive information has raised significant privacy concerns. One major issue is vulnerability to Membership Inference Attacks (MIAs), which aim to discover whether a particular data point was part of a model's training set. In this context, it has been largely overlooked whether class imbalance handling methods — a prerequisite for training standard supervised classifiers on such data — exacerbate the disclosure of sensitive information. To address this gap, we conducted a series of empirical MIA evaluations on real-world imbalanced datasets to investigate this question. Our results show that techniques employed to balance the dataset for classification can significantly improve the success rate of MIA attacks. This has clear implications for researchers developing protective measures against such attacks, especially in scenarios with imbalanced datasets.

Keywords: Machine Learning, Inference Attack, Fairness, imbalanced datasets.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ML *Machine Learning*

MIA *Membership Inference Attack*

DP *Differential Privacy*

A.I *Artificial Intelligence*

ROC *Receiver Operating Characteristic*

TP *True Positive*

FP *False Positive*

TN *True Negative*

FN *False Negative*

F-SCORE Métrica de Avaliação de Modelos de Classificação

DL Deep Learning

NN Neural Network

RN Redes Neurais

RNs Redes Neurais Artificais

APS American Physical Society

CNNs Redes Neurais Convolucionais

VC Visão Computacional

PLN Processamento de Linguagem Natural

LGPD Lei Geral de Proteção de Dados

GDPR Regulamento Geral de Proteção de Dados

SMOTE Synthetic Minority Over-sampling Technique

ROS Random OverSampling

RUS Random UnderSampling

LRT Ataque de Razão de verossimilhança.

MU Machine Unlearning

UCI University of California

XAI Inteligência Artificial Explicável

SVM Support Vector Machine

RF Random Forest

ANN Artificial Neural Networks

LR Logistic Regression

FDRTs *Facial Detection and Recognition Technologies*

Sumário

1 Introdução

A crescente integração dos sistemas de aprendizado de máquina em áreas críticas desencadeou um importante debate sobre a sua equidade (ARAGãO; SILVA; MACHADO, 2024). No entanto, o debate sobre a inteligência artificial (IA) confiável — que abrange os sistemas de IA desenvolvidos, implementados e utilizados de forma ética, robusta e socialmente benéfica, com atributos como justiça, transparência, segurança, privacidade e responsabilidade - estaria incompleto sem abordar a dimensão da privacidade.

Isto porque os modelos de aprendizado de máquina internalizam naturalmente os padrões dos dados de treino e podem, sem querer, memorizar informações sensíveis. Os ataques de inferência de associação (NIU et al., 2024), por exemplo, demonstram a capacidade de inferir se determinados dados estavam ou não presentes no conjunto original de treino do modelo. Esse vazamento não é apenas uma questão técnica, mas tem implicações profundas para a justiça algorítmica. A divulgação de atributos sensíveis, — como informações de saúde, demográficas ou financeiras, — pode levar à discriminação seletiva, prejudicando os esforços para construir sistemas equitativos.

A situação acima pode ocorrer em muitas aplicações do mundo real, incluindo a aplicação da previsão de reincidência criminal, em que diferentes grupos demográficos (por exemplo, acusados brancos e negros) devem receber tratamento igualitário, ou seja, a construção de modelos justos com precisão semelhante nas diferentes classes. Ao mesmo tempo, o fato de um indivíduo ter participado no conjunto de dados de treino já implica que cometeu uma infração, o que é altamente sensível e deve ser mantido em segredo. A revelação dessa participação por qualquer ataque pode resultar em discriminação seletiva, o que tornaria inútil todo o esforço para alcançar a justiça algorítmica. Portanto, tanto a justiça social (PINHEIRO; SILVA; MACHADO, 2023) quanto a proteção da privacidade (NGUYEN et al., 2024) são essenciais para o uso ético do aprendizado de máquina, e compreender a interação entre esses dois aspectos é crucial.

No entanto, muitos conjuntos de dados do mundo real, como os de reincidência, têm uma distribuição desequilibrada. Isso acontece porque eles tendem a conter significativamente mais exemplos de reincidentes negros do que brancos, um fato que reflete as desigualdades sociais e sistêmicas. Esse desequilíbrio compromete diretamente o desempenho dos algoritmos de aprendizado de máquina, pois os modelos inferidos a partir de dados com classes desequilibradas têm uma precisão comprovadamente inferior para a classe com menos exemplos. Como resultado, o modelo afeta de forma desproporcional os grupos que já são vulneráveis.

Para resolver este problema, foram desenvolvidas várias abordagens com o objetivo

de mitigar o desequilíbrio de classes e as suas consequências na equidade do modelo. Essas estratégias podem ser aplicadas principalmente de duas maneiras: ao nível dos dados (JOHNSON; KHOSHGOFTAAR, 2019), adicionando ou removendo instâncias aleatórias do conjunto de treino para restabelecer o equilíbrio; ao nível do algoritmo (THAI-NGHE; GANTNER; SCHMIDT-THIEME, 2010), modificando o classificador internamente através da manipulação da sua função objetivo para aumentar a importância da classe minoritária.

Neste trabalho, investigamos qual das abordagens discutidas anteriormente para reduzir os efeitos do desequilíbrio de classes leva a modelos que são mais suscetíveis a ataques de inferência de associação. A nossa hipótese é que os modelos treinados utilizando uma técnica ao nível de algoritmo são mais suscetíveis a esses ataques do que os modelos que utilizam amostragem de dados.

Para demonstrar isso, atacamos modelos de redes neurais usando técnicas que exploram a função de perda e a distribuição de erros do modelo. Comparamos o sucesso do ataque em um modelo treinado com subamostragem de dados com o de um modelo alternativo, no qual o desequilíbrio (quatro vezes mais exemplos para uma classe) é corrigido no nível do algoritmo. Nossos resultados confirmam nossa hipótese.

Este Trabalho defende que, embora as estratégias para lidar com desequilíbrios de classe sejam geralmente motivadas pela busca da equidade (e tenham como objetivo eliminar essas diferenças de desempenho), elas devem ser cuidadosamente examinadas do ponto de vista da privacidade nos casos em que o conjunto de dados contém informações pessoais sensíveis. Paradoxalmente, estas estratégias podem aumentar a suscetibilidade à MIA e, consequentemente, comprometer a justiça através da divulgação de dados sensíveis.

1.1 Motivação

O crescente uso de modelos de aprendizado de máquina em domínios sensíveis, como saúde efinanças cuasou preocupações importantes sobre privacidade e equidade. Em especial, algoritmos treinados com dados desbalanceados tendem a apresentar desempenho desigual entre classes, frequentemente prejudicando grupos vulneráveis e até mesmo resultando em proconceitos. Para mitigar esse problema, diversas estratégias de balanceamento de classes têm sido propostas, seja manipulando os dados diretamente ou ajustando o processo de aprendizado do modelo.

Entretanto, embora essas técnicas sejam projetadas com o objetivo de melhorar a justiça e a acurácia dos modelos, seus impactos sobre a privacidade dos dados ainda são pouco compreendidos. Um exemplo crítico dessa vulnerabilidade são os ataques de inferência de membros (Membership Inference Attacks – MIAs), que buscam identificar se um determinado dado foi utilizado durante o treinamento de um modelo. Tais ataques

representam uma ameaça direta à privacidade, especialmente em cenários envolvendo informações sensíveis.

Diante disso, torna-se essencial investigar não apenas a eficácia das técnicas de balanceamento, mas também o quanto elas podem tornar os modelos mais ou menos suscetíveis a vazamentos de informação. Compreender essa relação é fundamental para desenvolver práticas de treinamento mais seguras e que garantem a proteção da privacidade.

1.2 Hipótese

Este trabalho parte da hipótese de que técnicas de balanceamento de classes aplicadas em nível de algoritmo, como o aprendizado sensível ao custo (cost-sensitive learning), aumentam a vulnerabilidade dos modelos a ataques MIAs, quando comparadas a técnicas aplicadas em nível de dados. Acredita-se que ao ajustar internamente a função objetivo do modelo para dar maior peso à classe minoritária resulta em um algoritmo que tende a memorizar mais intensamente os exemplos dessa classe. Isso pode facilitar o trabalho de atacantes que exploram padrões de erro ou perdas do modelo para inferir a presença de dados no conjunto de treinamento. Dessa forma, embora o aprendizado sensível ao custo tenha como objetivo promover maior equidade entre as classes, ele pode expor informações sensíveis com maior facilidade e comprometer a privacidade e a justiça do sistema.

1.3 Objetivos

Analisar o impacto de diferentes abordagens de balanceamento de classes na privacidade de modelos de aprendizado de máquina sujeitos a ataques.

Objetivos Específicos

Especificamente, este trabalho busca os seguintes objetivos aplicados ao problema de vazamento de dados por meio de ataques de inferência de associação (Membership Inference Attacks – MIA):

- Avaliar o impacto de métodos de balanceamento de classes no desempenho de ataques de inferência de associação.
- Comparar a eficácia de abordagens em nível de dados (como undersampling) com abordagens em nível de algoritmo (como aprendizado sensível ao custo) quanto à sua vulnerabilidade a ataques.

- Identificar práticas de treinamento que maximizem a privacidade dos dados sensíveis em cenários com classes desbalanceadas.
- Explorar a influência do balanceamento de classes no comportamento dos modelos sob diferentes conjuntos de dados reais (Adult, APS).

1.4 Contribuições

Destacam-se como principais contribuições:

- Avaliação empírica da relação entre técnicas de balanceamento de classes e a suscetibilidade a ataques de inferência de associação.
- Demonstração de que abordagens em nível de dados oferecem maior robustez frente a ataques de privacidade do que abordagens em nível de algoritmo.
- Proposição de recomendações práticas para cientistas de dados que trabalham com conjuntos de dados desbalanceados e desejam proteger informações sensíveis.

Este estudo contribui para o entendimento dos ataques de inferência em modelos de aprendizado de máquina, sobretudo em dados desbalanceqados, ressaltando a urgência de incorporar práticas de proteção à privacidade. Ao reunir, categorizar e comparar diferentes estratégias de ataque e defesa, evidenciamos como mesmo exposições limitadas de modelos podem colocar em risco informações sensíveis. Essa análise abre caminho para o desenvolvimento de modelos mais robustos, conscientes dos riscos de reidentificação e mais preparados para atuar em contextos reais.

1.5 Organização do Trabalho

A Dissertação está organizada da seguinte forma:

- O Capítulo 1 apresenta o contexto e a motivação para o desenvolvimento deste trabalho.
- O Capítulo 2 aborda a fundamentação teórica das técnicas utilizadas, apresentando conceitos sobre privacidade, ataques de inferência e justiça em algoritmos.
- O Capítulo 3 descreve os trabalhos relacionados ao tema, com ênfase em vulnerabilidades de algoritmos de aprendizado de máquina.
- O Capítulo 4 apresenta a metodologia proposta, detalhando os experimentos realizados em diferentes bases de dados e as diversas abordagens de ataques de privacidade empregadas.

- O Capítulo 5 apresenta e discute os resultados obtidos nos experimentos realizados.
- O Capítulo 6 traz as conclusões do trabalho, destacando as contribuições alcançadas, as limitações encontradas e as direções para pesquisas futuras na área de privacidade de dados.

2 Fundamentação Teórica

Este capítulo apresenta os principais conceitos e fundamentos teóricos que embasam o desenvolvimento do presente estudo. O foco está em três eixos centrais: justiça algorítmica, privacidade em sistemas e ataques de inferência de membros (Membership Inference Attacks – MIA). Esses temas são fundamentais para compreender os riscos associados ao uso de dados sensíveis em modelos de aprendizado de máquina e como diferentes abordagens de modelagem podem influenciar a exposição de informações confidenciais. A seguir, são discutidos os principais conceitos relacionados a esses eixos.

2.1 Inteligência Artificial e Aprendizado de Máquina

A Inteligência Artificial (IA) é uma área da ciência da computação que visa criar sistemas capazes de simular comportamentos inteligentes inspirados na inteligência humana, como raciocínio, aprendizado, percepção e tomada de decisão. Eles buscam replicar capacidades humanas utilizando algoritmos e modelos matemáticos. Atualmente já existem diferentes subáreas da IA, como a visão computacional (VC), o processamento de linguagem natural (PLN), e o aprendizado de máquina (ML), que tem ganhado destaque por sua capacidade de extrair padrões e realizar predições a partir de grandes volumes de dados.

Aprendizado de Máquina é um ramo da IA e da ciência da computação que se caracteriza por permitir que um algoritmo aprenda sozinho através de dados. Segundo Mitchell (1997), um programa de computador aprende com a experiência **E** no que diz respeito a alguma classe de tarefas **T** e alguma medida de desempenho **P**, se seu desempenho em **T**, medido por **P**, melhora com a experiência **E**. Essa definição ressalta que o aprendizado é avaliado com base na melhoria contínua do desempenho ao longo do tempo, com base nos dados recebidos por ele.

O avanço do aprendizado de máquina está intimamente ligado à disponibilidade de grandes volumes de dados e ao aumento da capacidade computacional. Técnicas modernas como *deep learning*, que utiliza redes neurais artificiais com múltiplas camadas, têm se mostrado particularmente eficazes em áreas como reconhecimento de fala, tradução automática, análise de imagens e detecção de padrões complexos.No entanto, o uso dessas tecnologias também traz desafios, incluindo questões éticas relacionadas à privacidade de dados, viés algorítmico e transparência nas decisões automatizadas. Por isso, conceitos como *explainable AI* (IA explicável) e aprendizado sensível a custos (*cost-sensitive learning*) vêm sendo explorados para tornar esses sistemas mais compreensíveis, justos e responsáveis.

2.2 Aprendizado Supervisionado

No aprendizado supervisionado, o modelo é treinado com um conjunto de dados que já contém entradas (features) conhecidas e saídas esperadas (rótulos). Portanto, o objetivo é o modelo aprender uma relação entre essas features e saídas para no processo de aprendizado prever o rótulo correspondente quando trabalhando em novos dados.O modelo funciona como um avaliador humano que corrige as respostas, pois ele já sabe previamente quais estão corretas ou erradas, tendo aprendido isso a partir do estudo e da análise dos exemplos anteriores. Um grande exemplo disso é a rede neural: para produzir determinado resultado, ele usa algumas entradas e executa uma ou mais camadas de transformação matemática com base no ajuste das ponderações dos dados.

.

2.3 Aprendizado Não Supervisionado

No aprendizado não supervisionado o modelo recebe dados de entrada fornecidos ao algoritmo sem nenhum dado de saída rotulado, ele recebe apenas as entradas (features) sem conter nenhum rótulo. Nesse formato, o objetivo é encontrar padrões e grupos dentro dos dados sem o conhecimento de como esses agrupamentos devem ser selecionados. Então, apenas após o trabalho do algoritmo de encontrar padrões que as estruturas são conhecidas e selecionadas da forma que mais fazem sentido e combinam entre si. Uma técnica comum dentro do aprendizado não supervisionado é o grupamento em clusters, ela agrupa determinadas entradas de dados, para que possam ser categorizadas como um todo. Existem vários tipos de algoritmos de agrupamento em clusters, dependendo dos dados de entrada. Um exemplo de clustering é a identificação de diferentes tipos de tráfego de rede para prever possíveis incidentes de segurança.

2.4 Considerações Éticas em Inteligência Artificial e Aprendizado de Máquina

Viés Algorítmico

O viés algorítmico ocorre quando os modelos de IA aprendem padrões que refletem desigualdades ou preconceitos existentes nos dados de treinamento. Isso pode levar a decisões discriminatórias , por exemplo, na área de reconhecimento facial, pode ocorrer uma maior taxa de erro para determinados grupos étnicos. No desenvolvimento das tecnologias de reconhecimento facial — *Facial Detection and Recognition Technologies* (FDRTs) —, os padrões históricos de discriminação exercem influência desde as fases iniciais, moldando tanto o design quanto a implementação desses sistemas. Conforme examinado

por Leslie (2020), a dinâmica do preconceito não se limita aos dados utilizados, mas também influencia decisões de escolhas de parâmetros e contextos de aplicação. A adoção de FDRTs enviesadas pode agravar injustiças ao negar oportunidades ou reconhecimento apropriado a determinados grupos; a identidade e a dignidade de grupos marginalizados podem ser inadequadamente representadas ou até negadas. O problema pode ocorrer em qualquer tipo de dado e surge porque os dados usados no treinamento não representam adequadamente toda a diversidade da população. Para corrigir tais problemas, é necessário realizar o controle da qualidade dos dados e aplicar técnicas de justiça algorítmica.

Privacidade de Dados

A privacidade de dados é um desafio central no desenvolvimento e implementação de modelos de aprendizado de máquina, pois esses sistemas geralmente dependem de grandes volumes de informações para atingir altos níveis de desempenho. Alé disso, esses modelos muitas vezes podem ser usados em dados privados e que lidam com informaões confidenciais. Frequentemente, esses conjuntos de dados incluem informações pessoais e sensíveis, o que introduz riscos significativos relacionados à proteção e ao uso ético desses dados. Um dos problemas mais críticos é o vazamento de informações, que pode ocorrer tanto por meio de ataques diretos a bases de dados quanto por vulnerabilidades nos próprios modelos. Shokri et al. (2017) demonstram que modelos de aprendizado profundo podem estar sujeitos a ataques de *membership inference*, nos quais um invasor consegue determinar se um determinado registro fez parte do conjunto de treinamento.

Outro risco importante é a reidentificação de indivíduos a partir de dados aparentemente anônimos. Pesquisas como as de Narayanan e Shmatikov (2008) mostram que, com dados auxiliares adequados, é possível reconstruir informações pessoais a partir de registros anonimizados, minando esforços de privacidade. Além disso, modelos complexos podem memorizar padrões específicos de dados individuais, criando a possibilidade de que informações sensíveis sejam indiretamente extraídas durante interações com o sistema Carlini et al. (2021)

Segurança e Uso Indevido

A crescente integração da inteligência artificial em sistemas digitais traz questões fundamentais de segurança. Para compreender os riscos associados ao uso indevido dessas tecnologias, é importante distinguir três conceitos-chave: vulnerabilidade, ataque e vazamento. Vulnerabilidade refere-se a uma falha ou fragilidade existente em um sistema que pode ser explorada. Um ataque, por sua vez, consiste na ação deliberada de um agente mal-intencionado que busca explorar essa vulnerabilidade para obter algum tipo de vantagem ou causar dano. Já o vazamento ocorre quando informações sensíveis são expostas ou acessadas sem autorização, seja como consequência direta de um ataque ou

por falhas na proteção dos dados. No contexto da IA, essas três dimensões se manifestam de formas inéditas, como a manipulação de modelos por meio de ataques adversariais, a extração de dados privados utilizados no treinamento ou a geração de conteúdos falsos com potencial de enganar usuários e comprometer a integridade de processos sociais e institucionais.

Dessa forma, percebe-se a possibilidade da I.A ser utilizada para fins maliciosos, como deepfakes para disseminar desinformação e automação de ataques cibernético. Por isso, o desenvolvimento desses sistemas seguros envolve estabelecer limites de uso e implementar mecanismos de controle e supervisão humana.

Os riscos associados aos *deepfakes* e a outras formas de conteúdo gerado por inteligência artificial são amplos e potencialmente perigosos, abrangendo desde danos individuais até ameaças à segurança nacional. Considerando o alto nível de confiança social depositado em registros audiovisuais e a infinidade de aplicações possíveis para esse tipo de material, torna-se simples imaginar cenários de manipulação e desinformação em larga escala. Conforme aponta Waldemarsson (2020), uma das utilizações mais preocupantes envolve a manipulação de processos eleitorais. Um exemplo seria a divulgação, às vésperas de uma eleição disputada, de um vídeo falso retratando um candidato em comportamentos comprometedores ou realizando declarações polêmicas, com potencial de alterar significativamente a percepção pública e influenciar o resultado final.

Diante desses riscos, torna-se essencial estabelecer limites claros para o uso da IA, definindo políticas públicas, diretrizes e técnicas que assegurem sua aplicação ética e segura. Isso envolve não apenas a restrição do acesso a modelos de alto risco, mas também a implementação de mecanismos de auditoria, monitoramento e supervisão humana em todo o ciclo de vida do sistema, desde seu desenvolvimento até a sua operação em campo. Ao mesmo tempo, é fundamental investir em pesquisas voltadas para a detecção e mitigação de usos indevidos e problemas de segurança que envolvem a I.A, fomentando a criação de ferramentas capazes de mitigar ou prevenir esses problemas Dessa forma, a segurança da IA deve ser compreendida como um compromisso contínuo que equilibre inovação e responsabilidade, prevenindo que seu potencial transformador seja comprometido por aplicações nocivas.

2.5 Justiça em Machine Learning

Neste trabalho, o conceito de justiça é definido como a capacidade do classificador de apresentar uma precisão equilibrada entre as diferentes classes, garantindo um desempenho equitativo para todos os grupos representados no conjunto de dados. Esta abordagem procura assegurar que o modelo não favoreça uma classe em detrimento de outra, mitigando os efeitos do desequilíbrio de classes e promovendo a equidade algorítmica.

Segundo Barocas, Hardt e Narayanan (2019), a discriminação algorítmica ocorre quando decisões automatizadas consideram de forma injusta características associadas a categorias sociais historicamente marginalizadas. O conceito de injustiça, nesse contexto, é específico de cada domínio, pois se relaciona diretamente a oportunidades que impactam a vida das pessoas. A introdução de algoritmos em processos decisórios amplia preocupações normativas relevantes, já que esses sistemas podem reforçar padrões sistemáticos de tratamento adverso com base em seus grupos.

Justiça algorítmica envolve a compreensão dos múltiplos tipos de discriminação que podem surgir em sistemas automatizados e o desenvolvimento de técnicas para mitigá-los, seja por meio do balanceamento dos dados, ajuste nos objetivos de aprendizado ou definição de métricas de equidade. Mehrabi et al. (2021) destacam que a justiça algorítmica pode assumir diferentes formas, como *demographic parity*, *equalized odds* ou *predictive parity*, cada uma com vantagens e limitações dependendo do contexto de aplicação. Por isso, a escolha de qual métrica utilizar deve considerar os objetivos do sistema e os potenciais impactos sociais do seu uso. Portanto, a Inteligência aritifical (I.A) deve garantir não só resultados computacionais preciso, mas garantir que ela não reproduzirá preconceitos de um pensamento humano.

2.6 O Trade-off entre Privacidade e Utilidade em Modelos de Machine Learning

A privacidade é um direito fundamental relacionado ao controle que indivíduos possuem sobre suas informações pessoais. Com as mudanças geradas pela tecnologia, a garantia de privacidade de forma digital precisou ser formalizada. Por isso, atualmente está sendo regulamentada por leis como a Lei Geral de Proteção de Dados (LGPD) no Brasil e o Regulamento Geral de Proteção de Dados (GDPR) na Europa. Essas normas estabelecem princípios que garantem a transparência dos dados, que devem ser seguidos por qualquer sistema que processe dados pessoais. Em contextos tecnológicos e computacionais, a privacidade refere-se à proteção desses dados contra acessos indevidos, vazamentos ou uso não autorizado. No contexto do Aprendizado de Máquina e da Inteligência Artificial, esse conceito ganha complexidade, pois os dados usados para treinar modelos muitas vezes contêm informações sensíveis. A pesquisa realizada por (KURTZ, 2020) ao IRIS - Instituto de Referência em Internet e Sociedade, demonstra que ao utilizar qualquer processo estatístico, os dados podem ser revertidos para informações pessoais, ressaltando a necessidade de precaução ao formar ou publicar qualquer banco de dados que envolva grupos de pessoas. Por esse motivo, torna-se fundamental compreender a importância da garantia de privacidade e dos estudos que visam assegurar essa proteção aos dados pessoais sensíveis.

Com o aumento do uso de aprendizado de máquina em aplicações reais nos últimos anos, surgiram preocupações significativas sobre a privacidade dos dados e a garantia de que não sejam expostos às pessoas, mesmo quando os dados parecem anônimos. Portanto, surge a necessidade de proteger esses dados, o que pode criar um problema denominado « trade-off» entre segurança e privacidade. Para compreender isso, basta observar que quanto mais os dados que a máquina irá utilizar são protegidos, mais isso pode afetar a qualidade desses dados. Portanto, apesar dos avanços em matéria de privacidade na aprendizagem automática, ainda existem desafios técnicos importantes, uma vez que muitos mecanismos de proteção implicam compromissos com a precisão e a utilidade do modelo. O modelo de trade-off demonstra que, à medida que a privacidade aumenta, a garantia de utilidade diminui, enquanto a utilidade do modelo tende a melhorar.

Por outro lado, a busca por um modelo de aprendizado de máquina justo, que minimize o impacto do desequilíbrio de classes, pode paradoxalmente aumentar os riscos de privacidade. A medida que estratégias de balanceamento de classes são aplicadas, surge um trade-off evidente: a tentativa de alcançar a equidade no desempenho do algoritmo pode, ao mesmo tempo, elevar a vulnerabilidade do modelo a ataques de inferência de associação, aumentando o risco de vazamento de dados confidenciais.

2.7 Ataques de inferência

O conceito de Ataque de Inferência de associação foi introduzido pela primeira vez por Homer et al. (2008), que demonstrou como um invasor poderia usar estatísticas públicas de um conjunto de dados genômicos para inferir a presença de um genoma específico nesse conjunto. Posteriormente, em um contexto de ataques a modelos de aprendizado de máquina, Shokri et al. (2017) propôs formalmente o MIA. Nesse cenário, o objetivo do invasor é construir um modelo de ataque que seja capaz de reconhecer as diferenças no comportamento do modelo-alvo e, a partir da sua saída, distinguir se um registro de dados específico foi ou não utilizado no treinamento. Assim, um invasor pode identificar se um dado foi usado para treinar um classificador de rede neural apenas com base no error de previsão desse registro.

Metodologia do Ataque

A metodologia do ataque, conforme proposta por (SHOKRI et al., 2017), baseia-se na criação de modelos sombra para simular o comportamento do modelo-alvo. Esses modelos sombra são treinados em conjuntos de dados disjuntos, ou seja, que não contêm nenhum registro do conjunto de dados privado do modelo-alvo.

Para entender a dinâmica, formalizamos o processo:

Seja f_{target} o modelo-alvo, treinado em um conjunto de dados privado D_{target}. Esse conjunto é composto por registros rotulados (x_i, y_i), onde x_i representa as características do ponto de dado e y_i é o rótulo verdadeiro. A saída do modelo-alvo é um vetor de probabilidade de tamanho c_{target}, onde c_{target} é o número de classes.

O invasor, por sua vez, constrói diversos modelos sombra que imitam o comportamento do modelo-alvo. A saída desses modelos sombra é então utilizada para treinar um modelo de ataque, cuja função é distinguir entre saídas de dados que foram usados no treinamento (membros) e saídas de dados que não foram (não-membros).

A principal premissa deste ataque é que os modelos de aprendizado de máquina tendem a se comportar de maneira diferente ao processar dados de treinamento versus dados que nunca viram, e essa diferença pode ser explorada.

A seguir, descrevem-se outras categorias de ataques de inferência de associação.

Ataques de inferência baseados na confiança da predição

Salem et al. (2019) propuseram um ataque com limiar baseado na maior probabilidade de predição do modelo (top-1). Nesse ataque, o adversário obtém a confiança de predição do modelo-alvo f_T para um dado-alvo x_{target}, representada por $f_T(x_{\text{target}})$. Em seguida, ele extrai a maior probabilidade posterior desse vetor de confiança e verifica se esse valor máximo é superior a um limiar pré-definido (τ). Se a condição for satisfeita, o ponto de dado é classificado como pertencente ao conjunto de treinamento do modelo-alvo; caso contrário, é considerado um não-membro. O ataque $M_{\text{conf}}(f_T(x))$ é formalmente definido da seguinte maneira:

$$M_{\text{conf}}(f_T(x)) = 1\left\{\max f_T(x) \geq \tau\right\}, \tag{2.1}$$

onde τ é o limiar pré-definido e $1(\cdot)$ é uma função indicadora, definida como:

$$1(A) = \begin{cases} 1, & \text{se o evento } A \text{ ocorrer,} \\ 0, & \text{caso contrário.} \end{cases} \tag{2.2}$$

Ataques de inferência baseados na perda predição

(YEOM et al., 2018) também propuseram um ataque baseado em limiar, utilizando o vetor de probabilidade $f_T(x)$ para uma instância x com rótulo verdadeiro y. A perda de predição, $\mathcal{L}(x, y) = -\log f_T(x)_y$, onde $f_T(x)_y$ é a probabilidade atribuída ao rótulo verdadeiro y. O atacante prediz que x é uma instância do conjunto de treinamento (ou seja, é membro) se $\mathcal{L}(x, y)$ for menor que a perda média de todas as instâncias de treinamento. A intuição era que o modelo treinado minimiza a perda de predição, de modo que a perda

de uma instância de treinamento tende a ser menor do que a de uma instância fora do treinamento. Assim, o ataque $M_{\text{loss}}(x)$ é definido da seguinte forma:

$$M_{\text{loss}}(f_T(x), y) = 1\left\{\mathcal{L}(x, y) \leq \tau\right\}, \tag{2.3}$$

onde $\mathcal{L}(\cdot)$ é a função de perda de entropia cruzada, e τ é o limiar, normalmente definido como a perda média sobre os dados de treinamento. Esse método também requer a estimativa da perda média de todas as instâncias de treinamento no modelo alvo.

Ataque de inferência baseado em rótulo de previsão

Yeom et al. (2018) propuseram um ataque de inferência baseado no rótulo de predição, que não requer um modelo sombra. Esse ataque se baseia na discrepância entre a classe prevista pelo modelo e a classe real, classificando uma amostra como membro se a predição do modelo corresponder ao rótulo verdadeiro. Caso contrário, a amostra é inferida como não-membro. A intuição é que o modelo foi treinado para prever corretamente seus dados de treinamento, mas esse desempenho pode não se generalizar bem para dados não vistos.

O ataque $\mathcal{M}_{\text{label}}(\cdot)$ é formalizado da seguinte maneira:

$$\mathcal{M}_{\text{label}}(F_T(x), y) = 1\left[\arg\max F_T(x) = y\right],$$

onde $F_T(x)$ é a saída do modelo para a entrada x, y é o rótulo verdadeiro, e 1 é a função indicadora. .

Ataques de inferência baseados na entropia da predição

Salem et al. (2018) propuseram um ataque do tipo top1 com limiar, cujo limiar é baseado na característica top-1, e esse limiar pode ser calculado utilizando a perda por entropia cruzada. Uma amostra é considerada um membro quando sua entropia cruzada é menor do que um limiar pré-definido τ_{entr}. O ataque $\mathcal{M}_{\text{entr}}(\cdot)$ é definido da seguinte forma:

$$\mathcal{M}_{\text{entr}}(F_T(x), y) = 1\left\{-\log(F_T(x)_y) \leq \tau_{\text{entr}}\right\}, \tag{2.4}$$

onde $\log(F_T(x)_y)$ representa o valor da probabilidade para o rótulo verdadeiro y, e τ_{entr} é o limiar de entropia cruzada predefinido.

Ataques de inferência baseados em perturbações adversariais das amostras

Choquette-Choo et al. (2021) introduziram um ataque label-only que, em vez disso, avaliou a robustez dos rótulos previstos (rótulos duros) do modelo sob perturbações da

entrada, para inferir a pertença. A estratégia deles consistiu em computar "proxies"somente com rótulos para a confiança do modelo avaliando sua robustez frente a perturbações estratégicas da entrada x, sejam elas sintéticas (isto é, por aumento de dados) ou adversariais (exemplos).

Seguindo uma perspectiva de margem máxima, eles previram que os pontos de dados que exibiam alta robustez eram pontos de dados de treino.

Ataques de inferência baseados em teste de hipótese

Long et al. (2020) consideraram um adversário pragmático que selecionava cuidadosamente os alvos com base na (percebida) vulnerabilidade à inferência de pertencimento, e tentavam minimizar falsos positivos ao mesmo tempo que equilibravam a cobertura com precisão.

No ataque de inferência direta, eles estimaram a confiança de que os dados-alvo r estavam presentes no conjunto de treino realizando um teste de hipótese unilateral: sob a hipótese nula H_0, r não estava presente no conjunto de treino (isto é, a função de perda logarítmica $\mathcal{L}(F_T, r)$ foi extraída de uma distribuição D_L); enquanto sob a hipótese alternativa, r foi usado para treinar o modelo alvo F_T (isto é, $\mathcal{L}(F_T, r)$ foi menor devido à influência de r no treino). Portanto, o valor-p foi calculado da seguinte forma:

$$p = F(\mathcal{L}(F_T, r)). \tag{2.5}$$

Isso deu a confiança de que r foi usado para treinar F_T apenas se p fosse menor que um limiar (por exemplo, 0,01), de modo que a hipótese nula fosse rejeitada.

Principais Defesas

As defesas contra ataques de inferência de associação podem ser analisadas não apenas pelo seu funcionamento individual, mas também pelo contexto em que se aplicam, pelos compromissos entre desempenho e segurança e pelas tendências de evolução dessas abordagens.

A seguir, as defesas contra ataques de inferência de associação são comparadas de forma estruturada. A Tabela 1 resume as principais técnicas, avaliando cada uma em relação à sua metodologia, ao impacto no desempenho do modelo e à sua robustez frente a adversários.

Tabela 1 – Comparação entre diferentes defesas contra ataques de inferência de pertencimento (adaptado de Niu et al., 2024)

Defesa	Vantagens	Desvantagens
Redução do overfitting baseada em defesa	Reduz o overfitting do modelo e melhora a generalização.	Falta em fornecer compensações aceitáveis entre privacidade e utilidade.
Privacidade diferencial	Oferece garantias formais de privacidade. Pode ser eficaz quando o orçamento de privacidade é pequeno.	Introduz muito ruído que prejudica o objetivo do aprendizado ou resulta em compensações inaceitáveis.
Mascaramento do score de confiança	Preserva a utilidade com resiliência razoável ao ataque.	Vulnerável a ataques baseados em limiar. Não protege completamente.
Métodos baseados em amostragem	Diminui a exposição dos dados confidenciais.	Amostragem inadequada pode levar a modelos menos precisos.
Perturbação dos dados de treino	Pode reduzir o risco de memorização excessiva dos dados.	Afeta a acurácia e pode distorcer o aprendizado.
Adversarial training	Pode aumentar a robustez do modelo contra diferentes tipos de ataque.	Custo computacional elevado e dificuldade de generalização.
Defesas específicas contra MIA	Desenvolvidas especificamente para mitigar ataques de inferência de pertencimento.	Podem não ser eficazes contra ataques fora do escopo.

A redução do overfitting baseada em defesa busca melhorar a generalização do modelo, diminuindo a memorização de exemplos de treino. Embora eficaz nesse aspecto, essa estratégia tende a não oferecer garantias explícitas de privacidade, exigindo um

equilíbrio delicado entre desempenho e proteção de dados.

A privacidade diferencial se destaca por fornecer garantias formais de privacidade ao adicionar ruído controlado ao processo de aprendizado. No entanto, o excesso de ruído pode comprometer a utilidade do modelo, especialmente quando o orçamento de privacidade é reduzido.

O mascaramento do score de confiança preserva a utilidade dos modelos e oferece certa resistência aos ataques, mas continua vulnerável a abordagens baseadas em limiar, que exploram pequenas variações nas respostas do modelo.

Já os métodos baseados em amostragem reduzem a exposição de dados confidenciais, porém, uma amostragem inadequada pode resultar em modelos menos precisos ou enviesados.

A perturbação dos dados de treino, por sua vez, visa limitar a memorização de exemplos sensíveis, mas pode distorcer o processo de aprendizado e afetar negativamente a acurácia.

O treinamento adversarial é uma técnica mais robusta que aumenta a resiliência do modelo frente a diferentes tipos de ataque. No entanto, o custo computacional é elevado e há maior dificuldade em manter boa capacidade de generalização.

Por fim, as defesas específicas contra ataques MIA são desenvolvidas com foco direto nesse tipo de ameaça, mas tendem a ter eficácia limitada quando aplicadas a outros tipos de ataque ou cenários distintos, indicando a necessidade de abordagens mais abrangentes.

Com o crescimento do uso de modelos de ML em áreas sensíveis, aumentaram também as preocupações com sua segurança. Diversos tipos de ataques têm sido propostos para explorar vulnerabilidades desses sistemas, gerando a necessidade de diferentes estratégias de defesa e contramedidas que têm sido desenvolvidas com o objetivo de mitigar ou impedir tais ataques, preservando o desempenho do modelo. Embora diversas estratégias de defesa tenham sido propostas, nenhuma delas garante proteção absoluta, sendo necessário avaliar o equilíbrio entre privacidade e desempenho em cada cenário de uso. Cabe à comunidade científica e aos profissionais da área considerar essas ameaças desde o projeto dos modelos até sua aplicação final, promovendo uma cultura de segurança e responsabilidade com os dados sensíveis dos usuários.

Considerações Finais sobre os Ataques

Os ataques de inferência representam uma ameaça crescente à privacidade em sistemas baseados em aprendizado automático. A partir da análise das diversas abordagens desenvolvidas nos últimos anos, incluindo os ataques mencionados, pode-se observar que a exposição dos modelos preditivos, mesmo que de forma limitada, pode comprometer as informações sensíveis das pessoas presentes nos dados de treinamento. Esses ataques

demonstram que, mesmo quando os modelos operam com alta precisão e generalização, ainda podem ser explorados por adversários. Portanto, o estudo dos MIA não apenas alerta sobre a vulnerabilidade dos sistemas atuais, mas também reforça a importância da privacidade como requisito fundamental no desenvolvimento de soluções de aprendizado automático.

2.8 O desbalanço de Classes e suas implicações

No aprendizado de máquina o desequilíbrio de classes pode surgir num conjunto de dados que tem uma classe desproporcional em relação a outra, por exemplo, numa base que tem acusados negros e brancos, pode haver um número maior de um grupo do que do outro. O problema acontece quando o classificador tende a prever que o grupo maioritário tem maior propensão a reincidir, o que pode ser problemático, uma vez que o classificador preditivo deve ter uma precisão preditiva semelhante.

Uma base de dados está desequilibrada quando a quantidade de dados de uma classe é muito superior à quantidade de outra classe. É importante ter em conta esta característica da base, pois pode dar origem a problemas de enviesamento em possíveis análises. Um exemplo conhecido deste problema é a base de dados chamada COMPAS (sigla para Correctional Offender Management Profiling for Alternative Sanctions), na qual os réus negros reincidentes em crimes são mais numerosos do que os réus brancos, o que pode levar a que sejam julgados incorretamente e corram um maior risco de reincidência de acordo com o algoritmo. O erro também pode ocorrer na classe menos numerosa. Neste exemplo, os acusados brancos tinham mais probabilidades do que os negros de serem incorretamente identificados como de baixo risco.

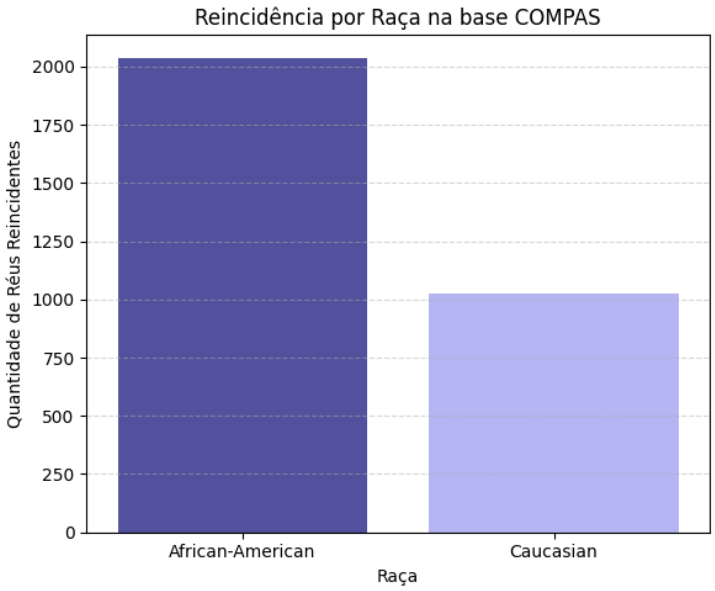


Figura 1 – Desbalanceamento de reincidência por raça na base COMPAS.

A Figura 1 ilustra o desbalanceamento na base de dados mencionada COMPAS. Observa-se que há um número significativamente maior de réus reincidentes classificados como African-American (afro-americanos) em comparação aos classificados como Caucasian (brancos). Esse desbalanceamento é relevante, pois pode introduzir viés nos modelos de aprendizado de máquina que utilizam essa base. Modelos treinados com dados desbalanceados tendem a aprender padrões distorcidos resultando em decisões injustas. No caso da COMPAS, estudos indicam que réus negros têm mais chance de serem classificados como de alto risco, mesmo quando não reincidem, o que na prática além de problemas em resultado de treinamento errôneo também pode gerar implicações éticas.

2.9 Técnicas de Tratamento de Dados Desbalanceados

Dados desbalanceados geralmente requerem algum tipo de tratamento antes de serem utilizados para treinar um modelo de machine learning. A seguir, são apresentadas algumas das técnicas mais comuns.

Técnicas a Nível de Dados

As técnicas a nível de dados atuam diretamente no conjunto de dados, modificando a distribuição das classes.

Oversampling

Essa técnica visa aumentar o número de exemplos da classe minoritária. O objetivo é equilibrar a proporção entre as classes para que o algoritmo de aprendizado de máquina não favoreça a classe majoritária durante o treinamento. Existem diferentes formas de aplicar o oversampling, sendo o **SMOTE** (*Synthetic Minority Over-sampling Technique*) uma das mais utilizadas. O SMOTE gera novas amostras sintéticas da classe minoritária com base nas características das amostras existentes, o que ajuda a evitar o problema de *overfitting*. O balanceamento obtido contribui para modelos mais justos e com melhor desempenho.

Para exemplificar, a Figura 2 mostra a distribuição de dados antes e depois da aplicação do SMOTE. Observa-se uma concentração predominante de indivíduos da raça *African-American*, enquanto as demais raças (*Hispanic*, *Other*, *Native American* e *Asian*) apresentam uma frequência menor. Para mitigar esse problema, foi aplicada a técnica SMOTE, que gera exemplos sintéticos para as classes minoritárias com base na interpolação entre os exemplos existentes. Após a aplicação do SMOTE, observa-se uma distribuição equilibrada entre todas as raças, como mostra o lado direito da Figura 2. Com essa nova distribuição, é possível treinar modelos mais robustos e reduzir o viés algorítmico.

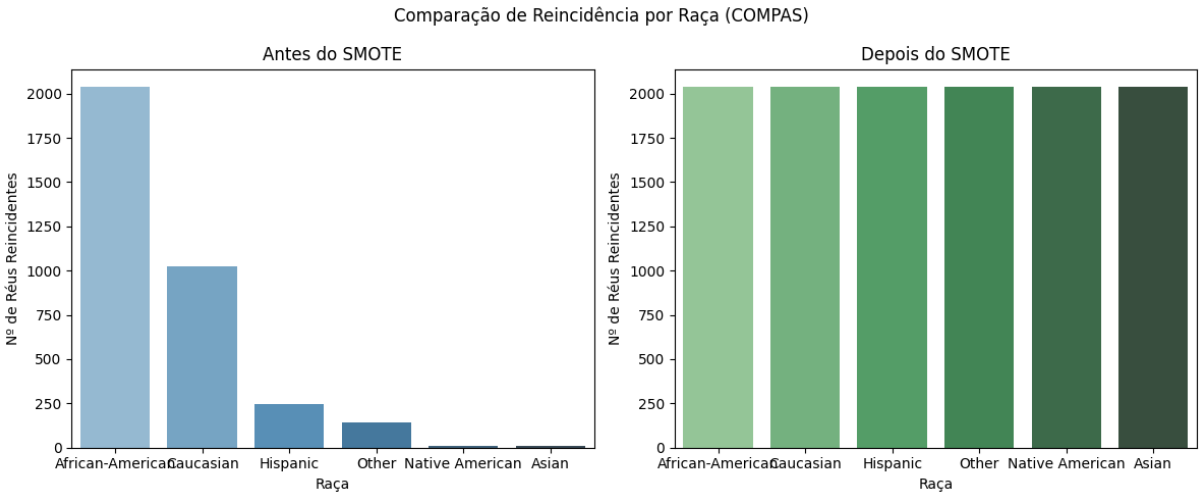


Figura 2 – Comparação da distribuição de reincidência por raça após a aplicação do SMOTE no conjunto de dados COMPAS.

Todos os exemplos e visualizações com a base COMPAS foram desenvolvidos para ilustrar os conceitos, gerados utilizando a linguagem Python e bibliotecas como Scikit-learn, Matplotlib e Seaborn.

De forma geral, oversampling sempre lida com conjuntos de dados desbalanceados da forma mostrada na Figura 3. No lado esquerdo, observa-se uma desproporção entre os rótulos 0 e 1, onde a classe 0 (barra laranja) possui muito menos amostras do que a classe 1 (barra azul). Para corrigir esse desequilíbrio, a técnica de oversampling replica aleatoriamente exemplos da classe minoritária 0 até que ela tenha a mesma quantidade de exemplos da classe majoritária. Isso resulta em um conjunto de dados balanceado, o que ajuda o modelo de aprendizado de máquina a não ser tendencioso em relação à classe com mais exemplos.

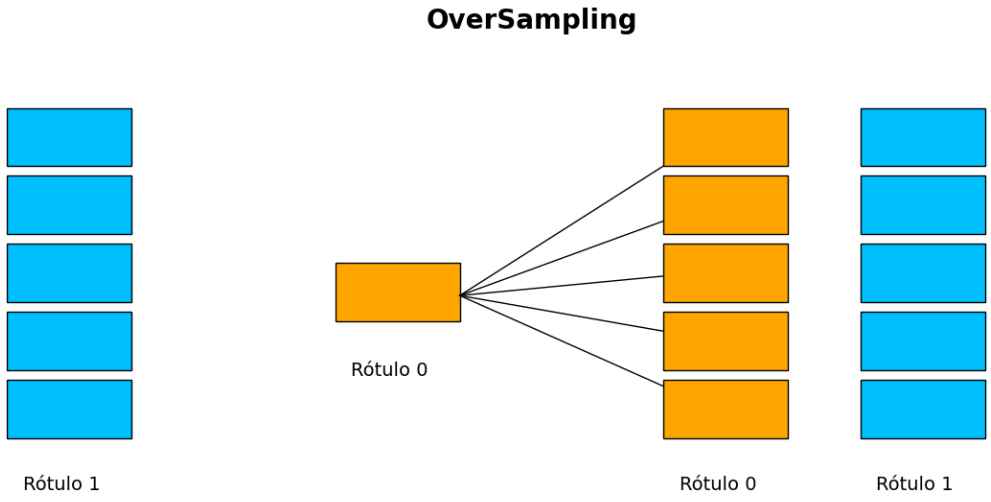


Figura 3 – Representação visual do Oversampling.

Undersampling

O Undersampling é outra técnica para equilibrar dados desproporcionais. No entanto, ela atua pela remoção de instâncias da classe majoritária. (JOHNSON; KHOSHGOFTAAR, 2019) destacam que, em cenários de desequilíbrio severo, técnicas de oversampling aleatório (*Random OverSampling* – ROS) tendem a gerar melhor desempenho em modelos de *deep learning* do que técnicas de undersampling aleatório (*Random UnderSampling* – RUS), que podem comprometer a representatividade da classe majoritária ao remover amostras importantes.

Os benefícios do undersampling incluem a redução do tempo de treinamento do modelo, pois o excesso de dados é removido, e a colaboração para amenizar o problema de desequilíbrio da base.

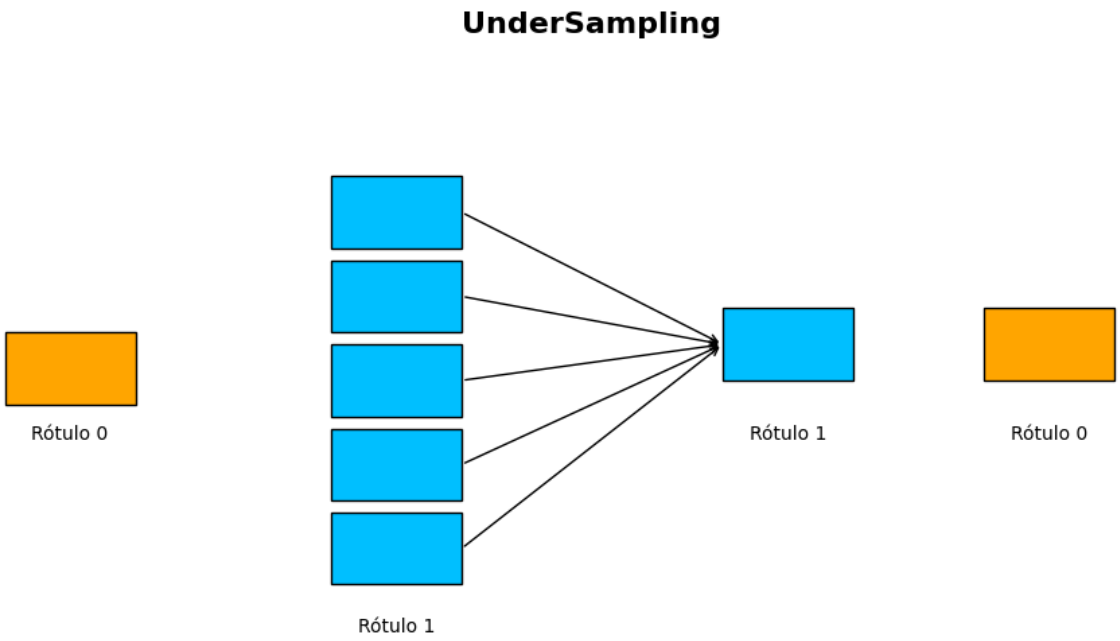


Figura 4 – Representação visual do Undersampling.

Na Figura 4, a subamostragem reduz o número de amostras da classe majoritária (representada pela barra azul). Isso é feito selecionando aleatoriamente um subconjunto das amostras dessa classe, de forma que o número de exemplos fique mais equilibrado em relação à classe minoritária (laranja).

Estratégias Híbridas (Oversampling + Undersampling)

As estratégias híbridas combinam Oversampling e Undersampling. (JOHNSON; KHOSHGOFTAAR, 2019) apontam que essas técnicas, que combinam ROS e RUS, conseguem equilibrar o conjunto de dados de forma adequada, aliando eficácia preditiva e eficiência computacional. Por essa razão, são altamente recomendadas para bases com desbalanceamento extremo.

Técnicas a Nível de Algoritmo

Cost-sensitive learning

No aprendizado de máquina, o *Cost-sensitive learning* é uma técnica que reconhece os custos variáveis associados a erros de classificação incorreta em conjuntos de dados desbalanceados. Sua principal vantagem é a capacidade de manter os dados originais, sem a necessidade de over ou undersampling. Em vez de alterar o conjunto de dados, essa abordagem ajusta os pesos das classes diretamente na função de perda. Dessa forma, o modelo aprende a dar mais importância aos erros da classe minoritária, sem duplicar ou excluir dados. Essa técnica pode ser aplicada em diversos algoritmos de classificação supervisionada, como:

- **SVM (Support Vector Machine)**: um modelo que busca encontrar o hiperplano ótimo que separa as classes com a maior margem possível. No caso de *Cost-sensitive learning*, o parâmetro `class_weight` pode ser utilizado para penalizar mais os erros da classe minoritária.
- **RF (Random Forest)**: um conjunto de árvores de decisão que votam para decidir a classe final. Também permite ponderação de classes.
- **ANN (Artificial Neural Networks)**: redes neurais artificiais que aprendem padrões complexos nos dados; a penalização por classe pode ser ajustada diretamente na função de perda.
- **LR (Logistic Regression)**: modelo estatístico simples e interpretável, que também suporta ponderação por classe para lidar com desbalanceamento.

Segundo (HE; GARCIA, 2010), algoritmos que tratam todos os erros de classificação da mesma forma não conseguem lidar adequadamente com dados desbalanceados. Isso ocorre porque a classe minoritária, além de ser sub-representada, exige uma maior relevância na sua classificação correta. Esses fatores tornam a classificação desbalanceada um dos desafios mais difíceis no aprendizado de máquina.

Os custos de classificação incorreta podem ser categorizados em quatro tipos, conforme a Tabela 2.

Custos de Classificação incorreta

Esta abordagem pressupõe que os custos de classificação incorreta (custos de falsos negativos e falsos positivos) não são iguais. O custo de classificar incorretamente um não terrorista como terrorista é muito menor do que o custo de classificar incorretamente um terrorista real carregando uma bomba em um voo.

Tabela 2 – Classificações possíveis em um contexto de duas classes.

Tipo	**Descrição**
Verdadeiro Positivo (VP)	A instância positiva foi corretamente classificada como positiva.
Falso Positivo (FP)	A instância negativa foi incorretamente classificada como positiva.
Verdadeiro Negativo (VN)	A instância negativa foi corretamente classificada como negativa.
Falso Negativo (FN)	A instância positiva foi incorretamente classificada como negativa.

Este estudo se concentra em problemas de classificação binária; denotamos a classe positiva (+ ou +1) como a minoria e a classe negativa (- ou -1) como a maioria. Seja $C(i, j)$ o custo de prever um exemplo que pertence à classe i, embora, na verdade, pertença à classe j; a matriz de custos é mostrada na Tabela 3.

Tabela 3 – Matriz de custos para classificação binária.

	Classe Real	
	-1	+1
Classe Prevista -1	$C(-1,-1)$	$C(-1,+1)$
Classe Prevista +1	$C(+1,-1)$	$C(+1,+1)$

O objetivo do aprendizado sensível ao custo é construir um modelo com custos mínimos de classificação incorreta (custo total):

$$\text{CustoTotal} = C(-,+) \times FN + C(+,-) \times FP$$

onde FN e FP são o número de exemplos de falso negativo e falso positivo, respectivamente.

3 Trabalhos Relacionados

A proteção de dados tornou-se uma questão central no aprendizado de máquina. Isso se deve ao fato de que a proteção da privacidade e a transparência são dois pilares fundamentais para o desenvolvimento de sistemas de aprendizado confiáveis. Nesse contexto, os *Membership Inference Attacks* (MIA) representam um problema sério.

Shokri et al. (2017) demonstraram que modelos podem ser vulneráveis a esse tipo de ataque. Entre os ataques mais populares estão o *loss threshold* e o *likelihood ratio attack* (LRT). O *loss threshold* detecta se um ponto de dado foi incluído no conjunto de treinamento comparando a taxa de erro do modelo com um valor de limiar, o que requer acesso aos rótulos e aos detalhes do modelo (SABLAYROLLES et al., 2019). O LRT, por sua vez, utiliza *shadow models* para comparar os níveis de confiança de dados presentes ou não no conjunto de treinamento e calcula uma razão de verossimilhança para prever a associação, sem exigir acesso direto ao modelo (CARLINI et al., 2022).

Shokri, Strobel e Zick (2021) investigaram os riscos à privacidade associados a técnicas de explicação de modelos. Eles utilizaram um ataque baseado em limiar, que infere a associação com base na confiança do modelo ou em sua saída explicativa.Pawelczyk, Lakkaraju e Neel (2023) desenvolveram um ataque baseado em distância, no qual o ponto é considerado membro do conjunto de treinamento se a *loss* estiver abaixo de um determinado limiar. (ALVES et al., 2024) desenvolveram uma abordagem de *Differential Privacy* aplicada a *Gradient Boosting Decision Trees*, utilizando técnicas de particionamento para dados categóricos. Essa abordagem atende a altos requisitos de privacidade, ao mesmo tempo em que preserva a acurácia do modelo.

Com base no tema da justiça em aprendizado de máquina, outros pesquisadores também conduziram análises semelhantes. Por exemplo, Sena e Machado (2024) conduziram um estudo usando o Conjunto de Dados Adultos da UCI que analisou especificamente vieses relacionados a características sensíveis como etnia e gênero. Sua abordagem foi reduzir a dimensionalidade do conjunto de dados para mitigar esses vieses. Eles avaliaram o desempenho e a imparcialidade de três modelos comuns de aprendizado de máquina — regressão logística, floresta aleatória e aumento de gradiente — tanto na inclusão quanto na exclusão de atributos sensíveis. Os principais resultados mostram que, embora as métricas de desempenho tenham permanecido estáveis, as métricas de imparcialidade forneceram insights importantes, destacando a necessidade de considerar a imparcialidade juntamente com o desempenho em aplicações reais de aprendizado de máquina.

Chang e Shokri (2021) investigou a complexa interseção entre o aprendizado de máquina justo e a privacidade dos dados. O estudo analisa como a aplicação de técnicas

para minimizar a discriminação contra grupos específicos pode influenciar o vazamento de informações sobre os dados de treino do modelo. Para quantificar este vazamento, os autores utilizam ataques de inferência de associação. A principal conclusão da análise é que existe um trade-off significativo: quanto mais tendenciosos são os dados de treino, maior é o custo de privacidade - medido pela vulnerabilidade a esses ataques - necessário para alcançar a equidade para os subgrupos não privilegiados.

Neste trabalho, investigamos quantitativamente de que forma modelos de aprendizado de máquina podem revelar informações sobre os *datasets* individuais nos quais foram treinados. Nosso foco está em abordagens voltadas ao tratamento de dados desbalanceados — uma área ainda pouco explorada.

4 Procedimento Metodológico

Neste trabalho, propomos uma investigação quantitativa sobre como o desequilíbrio entre classes nos conjuntos de dados pode influenciar a vulnerabilidade de modelos de aprendizado supervisionado a ataques de inferência de pertencimento. Mais especificamente, comparamos o desempenho de MIA sob duas estratégias distintas para lidar com o desequilíbrio de classes: a abordagem tradicional de balanceamento em nível de dados, que busca equilibrar as classes antes do treinamento, e a abordagem alternativa de aprendizado sensível ao custo, que atua diretamente na função de perda do modelo. A seguir, detalhamos as etapas adotadas em cada abordagem, bem como os critérios de avaliação utilizados para comparar seus impactos na privacidade e no desempenho dos modelos.

Comparamos o sucesso de um ataque de inferência em um conjunto de dados desbalanceado. Queremos saber qual das duas técnicas para lidar com o desequilíbrio o ataque apresenta melhor desempenho. A abordagem padrão de aprendizado de máquina, na qual aplicamos primeiro técnicas em nível de dados para balancear o conjunto de dados, ou a alternativa, na qual esse problema é abordado em nível de algoritmo, ajustando a função de perda.

Nossos passos para a abordagem de amostragem de dados são os seguintes: Selecionamos aleatoriamente amostras da classe majoritária em número igual ao da classe minoritária, de modo que as classes minoritária e majoritária contenham (aproximadamente) o mesmo número de amostras. Seguindo esse procedimento, aplicamos a abordagem padrão de aprendizado supervisionado. Na abordagem de aprendizado sensível ao custo, ajustamos as perdas para alcançar a imparcialidade, ponderando as amostras da classe minoritária com mais intensidade, modificando assim a função objetivo da abordagem padrão de aprendizado supervisionado.

4.1 Experimento de associação

Utilizamos experimentos de associação para avaliar o MIA. Um experimento de associação é um jogo teórico entre um proprietário de dados e um oponente, no qual o oponente tenta adivinhar a associação de uma amostra de dados. O jogo determina como o proprietário dos dados gera os dados de treinamento e quais informações estão disponíveis para o invasor. Utilizamos uma variação do procedimento de Yeom et al. (2018). Em nosso método, este experimento não recebe o conjunto de treinamento S como entrada, mas o conjunto de teste D. O proprietário dos dados treina a com S e, em seguida, seleciona um bit $b \in \{0, 1\}$ uniformemente e aleatoriamente. Este bit determina se o oponente recebe um elemento (z selecionado aleatoriamente de S) ou um não-elemento (uma nova amostra

$z \sim \mathcal{D}$). O oponente recebe o elemento z e o modelo treinado a, o algoritmo de treinamento $\mathcal{A}$ e o conjunto de dados de teste $\mathcal{D}$. O invasor usa essas informações para realizar um ataque $\text{Att}(z, a, \mathcal{A}, \mathcal{D})$, que gera um bit que indica a associação prevista de z. O invasor é bem-sucedido se o ataque inferir corretamente o status de associação de z.

4.1.1 Ataque de inferência de Associação considerado neste trabalho

O ataque funciona da seguinte forma: Primeiramente, o atacante observa o erro do modelo ao prever a saída para uma determinada entrada, por exemplo, um erro de 1. Em seguida, a probabilidade condicional da base correspondente ao treinamento (B_{train}) é calculada se o classificador cometeu um erro de 1 (E_1).

$$\Pr(B_{train} \mid E_1) = \frac{A * \Pr(B_{train})}{\Pr(E_1)}$$

onde $A = \Pr(E_1 \mid B_{train})$ é a probabilidade condicional de um erro de 1 do classificador E_1 se a base for a base de treinamento, ou seja, 1 menos a precisão do classificador na base de treinamento. As probabilidades $\Pr(B_{train} \mid E_0)$, $\Pr(B_{test} \mid E_1)$ e $\Pr(B_{test} \mid E_0)$ são calculadas usando a mesma lógica de antes. Se o invasor observar um erro de 1, ele escolhe aleatoriamente um valor para b (zero ou um) com probabilidade $\Pr(B_{train} \mid E_1)$ para o treinamento (b=0) e $\Pr(B_{test} \mid E_1)$ para o teste (b=1).

A vantagem de associação no experimento é definida como:

$$\text{Vantagem} = \Pr(\text{Att} = 0 \mid b = 0) - \Pr(\text{Att} = 0 \mid b = 1).$$

Nesta equação:

- O primeiro termo, $\Pr(\text{Att} = 0 \mid b = 0)$, representa a Taxa de Verdadeiros Positivos (TPR).
- O segundo termo, $\Pr(\text{Att} = 0 \mid b = 1)$, representa a Taxa de Falsos Positivos (FPR).

Assumimos que $b = 0$ indica um exemplo positivo, ou seja, o ataque descrito prevê corretamente que a rede foi treinada na entrada específica.

4.2 Conjuntos de Dados

Realizamos nossos experimentos utilizando dois *datasets* de acesso público, cada um representando um cenário de aplicação distinto e um conjunto específico de atributos. Abaixo, apresentamos uma breve descrição de ambos os *datasets*:

UCI Adult dataset (Census Income) [1] O UCI Adult Dataset, amplamente utilizado em pesquisas de aprendizado de máquina, foi obtido a partir do UCI Machine Learning Repository, mantido pela University of California, Irvine. Ele tem sua origem no Censo dos Estados Unidos de 1994, conduzido pelo U.S. Census Bureau, e foi posteriormente disponibilizado ao público. Criado pela Universidade da Califórnia, em Irvine, o repositório foi desenvolvido para reunir e compartilhar bases de dados destinadas à pesquisa e ensino em aprendizado de máquina. O conjunto Adult foi incluído no UCI em 1996 por pesquisadores da Silicon Graphics com o intuito de servir como um exemplo clássico para tarefas de classificação supervisionada. Desde então, o dataset tornou-se um dos mais utilizados no campo da ciência de dados.

Esse conjunto de dados contém 48.842 registros e 14 atributos, derivados de informações do censo norte-americano de 1994. Os atributos incluem idade, gênero, escolaridade, estado civil, ocupação, horas de trabalho semanais, capital ganho/perdido e país de origem.

A tarefa de classificação é binária, consistindo em prever se um indivíduo possui renda superior a 50.000 dólares anuais com base em seus atributos socioeconômicos. Essa base é amplamente utilizada em estudos que investigam viés, equidade e privacidade em modelos de aprendizado supervisionado, por conter atributos sensíveis, como raça e sexo, que podem introduzir desigualdades preditivas.

Em nosso estudo, o Adult Dataset foi utilizado para o treinamento dos modelos alvo (target models), devido à sua natureza sensível e ao desbalanceamento moderado entre as classes (“50K” e “>50K”), que torna o conjunto adequado para investigar como diferentes abordagens de balanceamento em nível de dados e de algoritmo impactam a vulnerabilidade a ataques de inferência de membros (MIA).

APS dataset [2] Utilizamos o *dataset* gerenciado pela APS como fonte de dados para nossas análises, identificando os autores por meio dos nomes presentes nos metadados dos artigos. Definimos a idade acadêmica dos autores com base no intervalo entre a primeira e a última publicação. Selecionamos autores que começaram a publicar após 1985 e que possuíam ao menos 8 anos de idade acadêmica.

Para calcular o valor dos preditores, consideramos as publicações realizadas até três anos antes do ano de pico de produtividade, enquanto as publicações subsequentes foram consideradas como publicações futuras. Utilizamos todas as citações de um determinado artigo dentro do *dataset* e, por isso, não definimos uma janela temporal para o acúmulo de citações.

A tarefa de classificação (binária) consiste em prever se um(a) cientista publicará

[1] https://archive.ics.uci.edu/dataset/2/adult
[2] https://journals.aps.org/datasets

mais de 3 artigos, cada um com o mesmo número de citações, nos anos seguintes ao ano de previsão.

A tarefa de classificação (binária) consiste em prever se um cientista publicará mais de 3 publicações, cada uma com o mesmo número de citações, nos anos seguintes ao ano da previsão. Usamos os seguintes índices bibliométricos do cientista como características preditoras: H, o índice H do cientista u do ano específico y de carreira até o ano da previsão; co, o número de diferentes coautores do cientista u do ano y de carreira até o ano da previsão; c, o número total de citações do cientista u do ano y de carreira até o ano da previsão; I_{iu}, o número total de artigos do cientista u, cada um com i citações, do ano y de carreira até o ano da previsão; e v, o número total de publicações em que o cientista u publicou do ano y de carreira até o ano da previsão.

Para descrever com precisão o desequilíbrio em nosso projeto experimental, fornecemos as distribuições de amostragem exatas para os conjuntos de dados UCI Adult e APS usados em nosso estudo.

- **UCI Adult Dataset:** Este *dataset* possui um total de 47.621 instâncias. A classe majoritária contém 39.780 amostras, enquanto a classe minoritária possui 7.841 amostras, resultando em uma razão de desbalanceamento aproximada de 1:5,07.
- **APS Dataset:** O *dataset* APS utilizado em nossos experimentos contém 28.481 instâncias. Especificamente, a classe majoritária inclui 23.168 amostras, enquanto a classe minoritária é composta por 5.313 amostras. Isso resulta em uma razão de desbalanceamento de aproximadamente 1:4,36, indicando uma disparidade significativa entre as classes.

A Tabela 4 apresenta a descrição completa de cada uma das colunas do conjunto de dados APS, que é orientado para a previsão do sucesso científico. De forma análoga, a Tabela 5 detalha os atributos do conjunto de dados Adult, amplamente utilizado em tarefas de classificação de rendimentos e análise de justiça em algoritmos.

Tabela 4 – Descrição das colunas do APS dataset

Coluna	**Descrição**
`name`	Nome do autor
`y1`, `y2`	Anos de início (`y1`) e fim (`y2`) da janela de análise (ex: 2000–2005)
`co`	Número total de coautores
`co_5`	Número de coautores nos últimos 5 anos
`h`	Índice h acumulado até o ano final da janela
`h_5`	Índice h nos últimos 5 anos
`h_3`	Índice h nos últimos 3 anos
`c`	Total de citações
`c_5`	Citações nos últimos 5 anos
`c_3`	Citações nos últimos 3 anos
`i_10_5`	Número de artigos com 10 ou mais citações nos últimos 5 anos
`i_5_5`	Número de artigos com 5 ou mais citações nos últimos 5 anos
`i_3_5`	Número de artigos com 3 ou mais citações nos últimos 5 anos
`v_5`	Variância no número de citações nos últimos 5 anos
`p_5`	Número de publicações nos últimos 5 anos
`p_3`	Número de publicações nos últimos 3 anos
`h_future_3`	Índice h projetado para os próximos 3 anos
`i3_future_3`	Índice i3 projetado para os próximos 3 anos
`i4_future_3`	Índice i4 projetado para os próximos 3 anos
`i5_future_3`	Índice i5 projetado para os próximos 3 anos
`i3_future3_gt1`	Indicador binário: se o índice i3 futuro será maior que 1
`i3_future3_gt1_complement`	Complemento da variável anterior (ex: 1 se i3_future3_gt1 for 0)

Tabela 5 – Descrição das colunas do Adult dataset

Coluna	**Descrição**
`age`	Idade do indivíduo
`workclass`	Categoria de emprego (ex: Private, Self-emp-not-inc, Federal-gov, etc.)
`education`	Nível educacional do indivíduo (ex: Bachelors, HS-grad, etc.)
`education_num`	Representação numérica do nível educacional
`marital_status`	Estado civil (ex: Married-civ-spouse, Never-married, etc.)
`occupation`	Tipo de ocupação profissional (ex: Sales, Tech-support, Craft-repair, etc.)
`relationship`	Relação familiar dentro do domicílio (ex: Husband, Not-in-family, etc.)
`race`	Raça/etnia do indivíduo (ex: White, Black, Asian-Pac-Islander, etc.)
`sex`	Sexo do indivíduo (Male ou Female)
`capital_gain`	Quantia de ganho de capital obtido (valores numéricos positivos)
`capital_loss`	Quantia de perda de capital sofrida (valores numéricos positivos)
`hours_per_week`	Número de horas trabalhadas por semana
`native_country`	País de origem do indivíduo
`income`	Classe de renda (<=50K ou >50K); variável-alvo (target)

A escolha do conjunto de dados para este estudo justifica-se pela presença de atributos sensíveis, tais como raça, sexo e renda, que viabilizam a análise de vulnerabilidades a ataques de inferência de pertencimento (MIAs). Adicionalmente, o notório desbalanceamento da variável-alvo, em que a maioria das instâncias representa indivíduos com renda inferior a 50 mil dólares, constitui um fator que, além de impactar o desempenho dos modelos, pode intensificar sua suscetibilidade ao vazamento de informações. Esta última relação é a hipótese central a ser explorada na presente pesquisa.

5 Resultados e Discussão

Para testar a hipótese de que a abordagem de aprendizado sensível ao custo no nível do algoritmo revela mais do que a abordagem de amostragem de dados e, portanto, fornece um valor maior para a métrica de vantagem de associação, conduzimos um experimento controlado. Comparamos o desempenho do ataque em um cenário com uma abordagem tradicional de aprendizado supervisionado, em que o conjunto de dados é primeiramente balanceado no nível dos dados (a abordagem de amostragem de dados), com o desempenho observado com uma abordagem de aprendizado supervisionado que aborda esse problema no nível do algoritmo e atua diretamente na função de perda (a abordagem de aprendizado sensível ao custo no nível do algoritmo). Figure 5 mostra o desempenho das NETs (Table 6) ataque ao conjunto de dados adultos e Figure 6 mostra o desempenho das NETs (Table 7) ataque ao conjunto de dados APS. Usamos perda BCEWithLogitsLoss($weight$) com $weight = 4.5$ para a abordagem sensível ao custo e $weight = 1$ para a abordagem alternativa. Tamanho de entrada da rede NET para adult 100 e saída 1 e tamanho de entrada da rede NET para APS 13 e saída 1.

Nossos resultados revelam que a técnica em nível de algoritmo leva a uma maior divulgação de dados privados do que a abordagem em nível de dados. Conforme mostrado na Figura 5, que descreve a métrica de Vantagem (Adv) do ataque, e na Figura 6, que mostra a Taxa de Verdadeiros Positivos (TPR) versus a Taxa de Falsos Positivos (FPR), o desempenho do ataque é consistentemente superior ao da técnica em nível de algoritmo.

Em ambas as figuras, a linha diagonal (linha tracejada cinza) representa o desempenho de um ataque aleatório. Na Figura 5 (vantagem), valores acima dessa linha indicam que o invasor obteve alguma informação sobre se uma entrada fazia parte dos dados de treinamento. Na Figura 6 (FPR vs. TPR), pontos acima dessa linha e que tendem a ser posicionados no canto superior esquerdo (indicando menos falsos positivos em uma determinada TPR) também indicam um ataque mais eficaz.

Vale ressaltar que, na técnica em nível de algoritmo, há mais redes (NETs) cujos resultados estão acima da linha diagonal em ambos os gráficos e tendem para o canto superior esquerdo, especialmente no gráfico TPR vs. FPR. Isso demonstra claramente que essa abordagem favorece a divulgação de dados privados e supera tanto um ataque aleatório quanto a abordagem alternativa, independentemente da métrica utilizada para visualizar o desempenho da MIA.

Nossos resultados revelam que o Ataque de Inferência de Membros (MIA) contra redes treinadas com a técnica em nível de algoritmo atinge valores médios de Adv mais altos, até **0,3**, do que a abordagem alternativa (técnica em nível de dados). Conforme

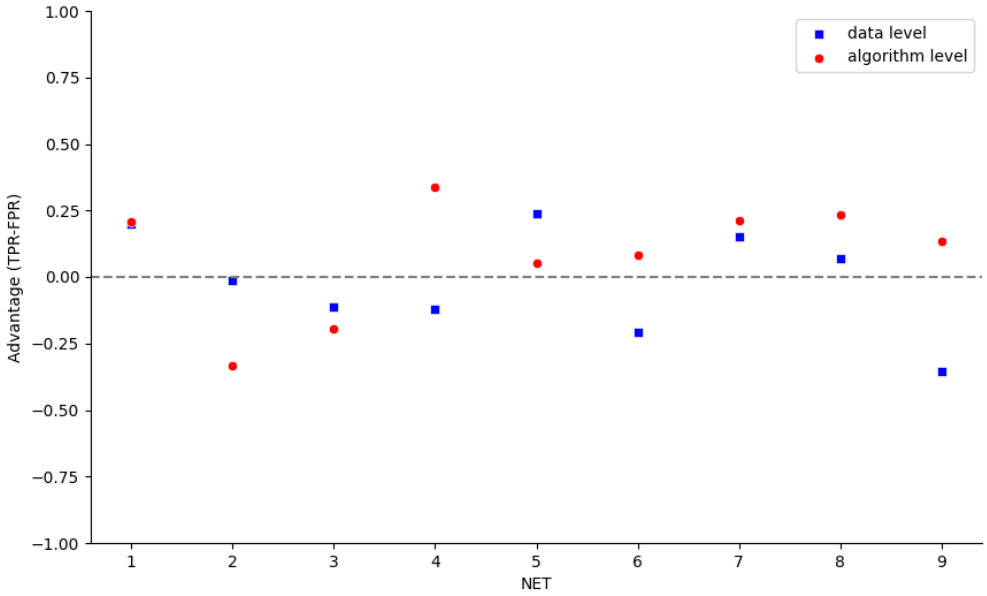
definido, um valor de Adv igual a 0 significa que o invasor determina aleatoriamente se uma entrada faz parte do conjunto de dados de treinamento, enquanto um valor igual a 1 representa uma inferência perfeita. O desempenho observado de 0,3, que está bem acima do acaso, demonstra claramente que o invasor obteve uma vantagem não trivial ao determinar se um ponto de dados pertence ao conjunto de treinamento.

Este resultado confirma claramente a existência de vazamento de informações dos modelos em relação aos seus dados de treinamento. Embora o invasor não tenha obtido uma inferência perfeita, um Adv de 0,3 já é um indicador preocupante de vulnerabilidade de privacidade. Isso sugere que os modelos revelam padrões reconhecíveis sobre se uma entrada faz parte dos dados de treinamento, o que poderia ser explorado por um invasor. Mesmo essa divulgação parcial representa um risco significativo à privacidade, pois permite uma identificação mais precisa do que aleatória de indivíduos cujos dados sensíveis podem ter sido usados no treinamento, abrindo caminho para discriminação ou outras violações de privacidade.

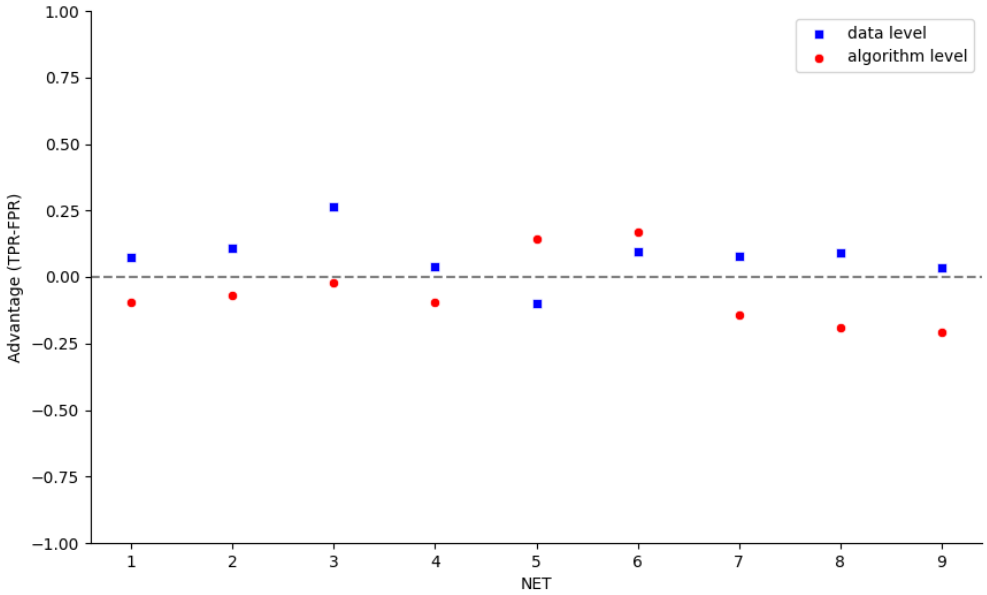
É importante observar que o ataque ao conjunto de dados APS teve um desempenho ruim e mostrou menos variação entre as técnicas. No entanto, houve casos em que o ataque a algumas redes treinadas com a abordagem em nível de algoritmo revelou mais dados privados do que a abordagem em nível de dados.

Nossos experimentos confirmam descobertas anteriores de que é possível obter alguma informação sobre se uma rede foi treinada com uma entrada específica. No entanto, demonstramos que isso se torna mais claro quando a abordagem de aprendizado sensível ao custo em nível de algoritmo é utilizada para lidar com um conjunto de dados desbalanceado.

Por outro lado, é necessária cautela, pois apenas modelos de rede neural feedforward foram testados em dois conjuntos de dados. Um ataque mais preciso poderia separar melhor o desempenho de cada técnica. Também é importante experimentar outros métodos para lidar com dados desbalanceados. Fornecemos evidências que apontam para a confirmação de nossa hipótese. No entanto, mais testes são necessários.

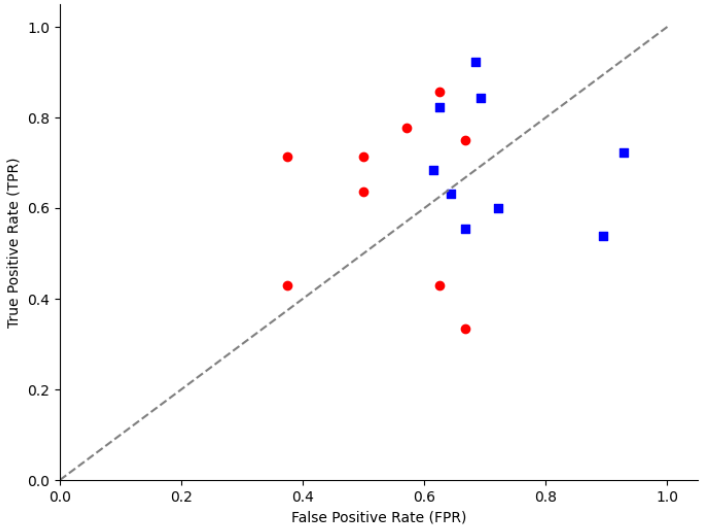


(a) Adult dataset

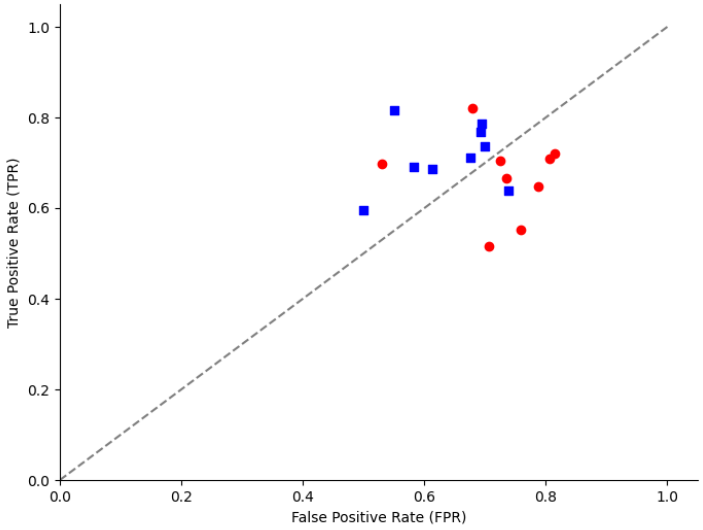


(b) APS dataset

Figura 5 – Avaliação do desempenho do MIA em diferentes redes, dependendo da técnica usada para a métrica de vantagem.



(a) Adult dataset



(b) APS dataset

Figura 6 – Avaliação do desempenho do Ataque de Inferência de Membro (MIA) (TPR vs. FPR) em diferentes redes, agrupadas por técnica.

Mostramos que a abordagem em nível de algoritmo para lidar com *datasets* desbalanceados em tarefas de classificação aumenta o risco de divulgação de dados privados. Ainda são necessários mais estudos que considerem a privacidade individual e abordagens para lidar com *datasets* desbalanceados. Esperamos que testes futuros confirmem nossas descobertas.

Trabalhos futuros devem se concentrar em melhorar a qualidade das defesas contra ataques de *membership inference* no treinamento de modelos de aprendizado supervisionado, por meio da adaptação da *loss* para lidar com dados desbalanceados.

Tabela 6 – Configuração das redes treinadas no conjunto de dados adult.

NET	Hidden 1	Hidden 2
1	50	100
2	50	50
3	100	75
4	200	50
5	150	50
6	100	150
7	200	100
8	100	100
9	200	200

Tabela 7 – Configuração das redes treinadas no conjunto de dados APS.

NET	Hidden 1	Hidden 2
1	25	25
2	50	25
3	100	25
4	25	50
5	50	50
6	100	50
7	25	100
8	50	100
9	100	100

6 Considerações Finais

6.1 Contribuições do Estudo

Este estudo explorou como diferentes estratégias para lidar com o desequilíbrio de classes afetam o risco de exposição de dados privados por meio de ataques de inferência de associação (MIAs). Foram comparadas duas abordagens principais: a amostragem em nível de dados e o aprendizado sensível a custos em nível de algoritmo, aplicadas em redes neurais treinadas sobre os conjuntos de dados APS e UCI Adult.

Os resultados demonstraram que a abordagem sensível a custos apresenta maior vulnerabilidade a vazamentos de privacidade, especialmente no conjunto UCI Adult, onde o sucesso dos ataques foi significativamente superior. Em contrapartida, a abordagem em nível de dados se mostrou mais robusta e segura frente aos ataques, ainda que apresente pequenas variações de desempenho em relação à acurácia.

As principais contribuições deste trabalho incluem:

- Avaliação empírica da relação entre técnicas de balanceamento e vulnerabilidade a ataques de inferência de associação.
- Demonstração de que abordagens em nível de dados reduzem o risco de vazamento de informações sensíveis.
- Identificação de um potencial trade-off entre justiça e privacidade, em que melhorias de desempenho em classes minoritárias podem aumentar a exposição de dados privados.
- Proposição de recomendações práticas para cientistas de dados, destacando a importância de: (i) priorizar o balanceamento em nível de dados; (ii) avaliar a privacidade como métrica adicional de desempenho; e (iii) documentar o processo de pré-processamento de forma transparente registrando como o balanceamento foi realizado e seus impactos sobre o comportamento do modelo.

6.2 Limitações

Apesar dos resultados promissores, este estudo apresenta algumas limitações. Primeiramente, os experimentos foram conduzidos apenas com redes neurais feedforward, o que restringe a generalização dos achados para outras arquiteturas, como redes convolucionais ou modelos baseados em atenção. Além disso, foram utilizados apenas dois conjuntos de dados públicos, o que pode limitar a variabilidade dos cenários analisados.

Outra limitação está na escala dos ataques de inferência considerados, que se concentraram em uma configuração clássica do MIA. Diferentes variações de ataque, com ajustes de complexidade ou uso de modelos adversários mais sofisticados, poderiam revelar novas nuances sobre a relação entre balanceamento e privacidade.

6.3 Direções para Pesquisas Futuras

Trabalhos futuros devem avaliar se esses resultados se generalizam para outras arquiteturas de modelos e conjuntos de dados e investigar defesas que, em conjunto, abordem o desequilíbrio de classes e os riscos de privacidade. Também vale a pena explorar se certas classes ou grupos demográficos são desproporcionalmente afetados por MIAs, um aspecto que permanece amplamente não examinado na literatura.

Então, percebe-se a necessidade de trabalhos que se concentrem na integração de estratégias que conciliem medidas como o de justiça algorítmica com técnicas de proteção à privacidade, investigando, por exemplo, como a aplicação combinada de métodos de balanceamento e aprendizado diferencialmente privado podem influenciar a eficácia dos ataques. A intenção é desenvolver um método que não apenas otimize o desempenho preditivo, mas que também reduza os riscos de vazamento de informações sensíveis, atendendo aos requisitos éticos e legais contemporâneos. Em última análise, a continuidade desta pesquisa visa aprofundar a compreensão das relações entre a robustez dos modelos de aprendizado de máquina e sua vulnerabilidade a ataques de inferência, oferecendo subsídios para a construção de sistemas que sejam simultaneamente precisos e seguros.

Referências

ALVES, A. G. M.; PEREIRA, F.; CHAVES, I.; MACHADO, J. Privacidade diferencial em gradient boosting decision trees com técnicas de particionamento para dados categóricos. In: *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*. Porto Alegre, RS, Brasil: SBC, 2024. p. 444–456. ISSN 2763-8979. Disponível em: <https://sol.sbc.org.br/index.php/sbbd/article/view/30712>. Citado na página 35.

ARAGãO, L. R.; SILVA, M.; MACHADO, J. The inefficiency of achieving fairness with protected attribute suppression. In: *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*. Porto Alegre, RS, Brasil: SBC, 2024. p. 813–819. ISSN 2763-8979. Citado na página 13.

BAROCAS, S.; HARDT, M.; NARAYANAN, A. *Fairness and machine learning: Limitations and opportunities*. [S.l.]: fairmlbook.org, 2019. <https://fairmlbook.org/>. Citado na página 22.

CARLINI, N.; CHIEN, S.; NASR, M.; SONG, S.; TERZIS, A.; TRAMÈR, F. Membership inference attacks from first principles. In: *2022 IEEE Symposium on Security and Privacy (SP)*. [S.l.]: IEEE, 2022. Citado na página 35.

CARLINI, N.; TRAMER, F.; WALLACE, E.; JAGIELSKI, M.; HERBERT-VOSS, A.; LEE, K.; ROBERTS, A.; BROWN, T. B.; SONG, D.; ERLINGSSON, U.; OPREA, A.; RAFFEL, C. Extracting training data from large language models. In: *30th USENIX Security Symposium (USENIX Security 21)*. [s.n.], 2021. p. 2633–2650. Disponível em: <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>. Citado na página 20.

CHANG, H.; SHOKRI, R. On the Privacy Risks of Algorithmic Fairness . In: *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. Los Alamitos, CA, USA: IEEE Computer Society, 2021. p. 292–303. Citado na página 35.

CHOQUETTE-CHOO, R. B.; CARLINI, N.; NASR, M.; PAPERNOT, N.; TRAMèR, F. Label-only membership inference attacks. In: *International Conference on Machine Learning (ICML)*. [S.l.: s.n.], 2021. Citado na página 25.

HE, H.; GARCIA, E. A. Cost-sensitive learning methods for imbalanced data. In: *IEEE Transactions on Knowledge and Data Engineering*. [S.l.: s.n.], 2010. v. 24, n. 9, p. 1379–1395. Citado na página 33.

HOMER, N.; SZELINGER, S.; REDMAN, M.; DUGGAN, D.; TEMBE, W.; MUEHLING, J.; PEARSON, J.; STEPHAN, D.; NELSON, S.; CRAIG, D. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS Genet*, v. 4, n. 8, p. e1000167, 2008. Citado na página 23.

JOHNSON, J. M.; KHOSHGOFTAAR, T. M. Survey on deep learning with class imbalance. *Journal of Big Data*, v. 6, n. 1, p. 27, Mar 2019. ISSN 2196-1115. Disponível em: <https://doi.org/10.1186/s40537-019-0192-5>. Citado 2 vezes nas páginas 14 e 32.

KURTZ, L. *Privacidade Diferencial: O Ruído Anonimizador*. 2020. <http://irisbh.com.br/privacidadediferencial-o-ruido-anonimizador/>. Acesso em: 30 nov. 2023. Citado na página 22.

LESLIE, D. *Understanding bias in facial recognition technologies: an explainer*. 2020. The Alan Turing Institute. Available via SSRN. Citado na página 20.

LONG, Y.; WANG, L.; BU, D.; BINDSCHAEDLER, V.; WANG, X.; TANG, H.; GUNTER, C. A.; CHEN, K. A pragmatic approach to membership inferences on machine learning models. In: *2020 IEEE European Symposium on Security and Privacy (EuroSP)*. [S.l.: s.n.], 2020. p. 521–534. Citado na página 26.

MEHRABI, N.; MORSTATTER, F.; SAXENA, N.; LERMAN, K.; GALSTYAN, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, ACM, v. 54, n. 6, p. 1–35, 2021. Citado na página 22.

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997. Citado na página 18.

NARAYANAN, A.; SHMATIKOV, V. Robust de-anonymization of large sparse datasets. In: IEEE. *2008 IEEE Symposium on Security and Privacy (SP)*. [S.l.], 2008. p. 111–125. Citado na página 20.

NGUYEN, T. T.; HUYNH, T. T.; REN, Z.; NGUYEN, T. T.; NGUYEN, P. L.; YIN, H.; NGUYEN, Q. V. H. Privacy-preserving explainable ai: a survey. *Science China Information Sciences*, v. 68, n. 1, p. 111101, Nov 2024. ISSN 1869-1919. Disponível em: <https://doi.org/10.1007/s11432-024-4123-4>. Citado na página 13.

NIU, J.; LIU, P.; ZHU, X.; SHEN, K.; WANG, Y.; CHI, H.; SHEN, Y.; JIANG, X.; MA, J.; ZHANG, Y. A survey on membership inference attacks and defenses in machine learning. *Journal of Information and Intelligence*, v. 2, n. 5, p. 404–454, 2024. ISSN 2949-7159. Citado na página 13.

PAWELCZYK, M.; LAKKARAJU, H.; NEEL, S. On the privacy risks of algorithmic recourse. In: RUIZ, F.; DY, J.; MEENT, J.-W. van de (Ed.). *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. PMLR, 2023. (Proceedings of Machine Learning Research, v. 206), p. 9680–9696. Disponível em: <https://proceedings.mlr.press/v206/pawelczyk23a.html>. Citado na página 35.

PINHEIRO, M. V.; SILVA, M. M.; MACHADO, J. Strategies selection for a fair classification in logistic regression: A comparative analysis. In: *Anais Estendidos do XXXVIII Simpósio Brasileiro de Bancos de Dados*. Porto Alegre, RS, Brasil: SBC, 2023. p. 15–21. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbbd_estendido/article/view/25608>. Citado na página 13.

SABLAYROLLES, A.; DOUZE, M.; SCHMID, C.; OLLIVIER, Y.; JEGOU, H. White-box vs black-box: Bayes optimal strategies for membership inference. In: CHAUDHURI, K.; SALAKHUTDINOV, R. (Ed.). *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019. (Proceedings of Machine Learning Research, v. 97), p. 5558–5567. Disponível em: <https://proceedings.mlr.press/v97/sablayrolles19a.html>. Citado na página 35.

SALEM, A.; ZHANG, Y.; HUMBERT, M.; BERRANG, P.; FRITZ, M.; BACKES, M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: *Network and Distributed Systems Security Symposium (NDSS)*. [S.l.: s.n.], 2018. Citado na página 25.

SALEM, A.; ZHANG, Y.; HUMBERT, M.; BERRANG, P.; FRITZ, M.; BACKES, M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: *NDSS*. [S.l.: s.n.], 2019. Citado na página 24.

SENA, L.; MACHADO, J. Evaluation of fairness in machine learning models using the uci adult dataset. In: *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*. Porto Alegre, RS, Brasil: SBC, 2024. p. 743–749. ISSN 2763-8979. Citado na página 35.

SHOKRI, R.; STROBEL, M.; ZICK, Y. On the privacy risks of model explanations. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY, USA: Association for Computing Machinery, 2021. (AIES '21), p. 231–241. ISBN 9781450384735. Disponível em: <https://doi.org/10.1145/3461702.3462533>. Citado na página 35.

SHOKRI, R.; STRONATI, M.; SONG, C.; SHMATIKOV, V. Membership inference attacks against machine learning models. In: IEEE. *2017 IEEE Symposium on Security and Privacy (SP)*. [S.l.], 2017. p. 3–18. Citado 3 vezes nas páginas 20, 23 e 35.

THAI-NGHE, N.; GANTNER, Z.; SCHMIDT-THIEME, L. Cost-sensitive learning methods for imbalanced data. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2010. p. 1–8. Citado na página 14.

WALDEMARSSON, C. *Disinformation, Deepfakes and Democracy: The European Response to Election Interference in the Digital Age*. 2020. Copenhagen: Alliance of Democracies. April 27, 2020. Citado na página 21.

YEOM, S.; GIACOMELLI, I.; FREDRIKSON, M.; JHA, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In: IEEE. *IEEE Computer Security Foundations Symposium (CSF)*. [S.l.], 2018. p. 268–282. Citado 3 vezes nas páginas 24, 25 e 37.