Anonimização de Dados e sua Influência na Performance de Modelos de Machine Learning: Uma Análise Comparativa

Maurício Alves da Silva Júnior[1], **Adiel T. de Almeida Filho**[1]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Av. Jornalista Aníbal Fernandes, s/n -– Cidade Universitária – 50.740-560
Recife-PE -– Brasil

(masj,adielfilho)@cin.ufpe.br

Abstract. *With the growing demand for data privacy and security, anonymization techniques have emerged as an alternative to reduce the likelihood of breaches during the training of machine learning models. Thus, this study aimed to present a comparative analysis between models trained with anonymized data and those trained with raw data, assessing the respective impacts on performance. The analysis involved the use of classification algorithms to evaluate the models' performance in both cases. The results indicate that the applied anonymization technique did not cause significant losses in model performance, making it possible to conclude that this technique is viable in scenarios where privacy is essential, without compromising model performance.*

Keywords: *machine learning; anonymization; privacy*

Resumo. *Com a crescente demanda por privacidade e segurança dos dados, as técnicas de anonimização emergem como uma alternativa para diminuir a possibilidade da violação dessas esferas no treinamento dos modelos de machine learning. Assim, o presente trabalho teve o objetivo de apresentar uma análise comparativa entre modelos treinados com dados anonimização e com dados puros, avaliando os respectivos impactos na performance. A análise consistiu no uso de algoritmos de classificação para avaliar o desempenho dos modelos em ambos os casos. Os resultados indicam que a técnica de anonimização aplicada não resultou em grandes prejuízos ao desempenho dos modelos, sendo possível concluir que essa técnica é viável em cenários onde a privacidade é essencial, e sem comprometer o desempenho dos modelos.*

Palavras-chave: *aprendizado de máquina; anonimização; privacidade.*

1. Introdução

Telecomunicações, saúde e negócios, áreas anteriormente entendidas como não relacionadas, hoje compartilham de algo em comum: dados. O desenvolvimento tecnológico atrelado a diferentes ramos tem impulsionado o aumento exponencial no que diz respeito à quantidade de informação diariamente gerada [Christen et al. 2020]. Estas podem ser analisadas e utilizadas das mais diversas formas.

O número de passos dados por alguém diariamente, o histórico de compras online e até mesmo os de batimentos cardíacos podem ser dados úteis dentro de algum contexto. Dados também têm sido importantes para analisar o impacto da desinformação

na democracia através do mapeamento do comportamento de usuários de mídias sociais [Lorenz-Spreen et al. 2023].

O avanço tecnológico, no entanto, não tem ocorrido sozinho. Junto a ele, o emergente debate sobre a privacidade e segurança de dados tem sido alvo de regulamentações específicas e discussões ao redor do mundo. A União Européia, uma das pioneiras no estabelecimento de legislação específica para a promoção da adoção da Inteligência Artificial através do chamado AI Act. A regulamentação tem por objetivo

> melhorar o funcionamento do mercado interno e promover a adoção de uma inteligência artificial (IA) centrada no ser humano e confiável, ao mesmo tempo em que garante um alto nível de proteção à saúde, segurança, direitos fundamentais consagrados na Carta, incluindo a democracia, o Estado de direito e a proteção ambiental, contra os efeitos prejudiciais dos sistemas de IA na União, além de apoiar a inovação. [União Europeia 2024]

Foi também pioneira no que diz respeito à proteção de dados com o início da vigência da *General Data Protection Regulation (GDPR*) em maio de 2018. O texto impõe obrigações às organizações de qualquer lugar do mundo que visem ou coletem dados relacionados aos cidadãos europeus.

Dois anos mais tarde, já no Brasil, a Lei Geral da Proteção de Dados (LGPD) entrou em vigor em setembro de 2020. A legislação brasileira exerce papel semelhante à da europeia. Ela "disciplina a proteção de dados pessoais" [Brasil 2020] fundamentando-se nos ideais de

> Respeito à privacidade; autodeterminação informativa; a liberdade de expressão, de informação, de comunicação e de opinião; a inviolabilidade da intimidade, da honra e da imagem; o desenvolvimento econômico e tecnológico e a inovação; a livre iniciativa, a livre concorrência e a defesa do consumidor; e os direitos humanos, o livre desenvolvimento da personalidade, a dignidade e o exercício da cidadania pelas pessoas naturais [Brasil 2020].

No Artigo 46° da LGPD, é dito que

> Os agentes de tratamento devem adotar medidas de segurança, técnicas e administrativas aptas a proteger os dados pessoais de acessos não autorizados e de situações acidentais ou ilícitas de destruição, perda, alteração, comunicação ou qualquer forma de tratamento inadequado ou ilícito [Brasil 2020].

Entende-se, então, que tanto os agentes de regulamentação quanto os responsáveis pela manipulação dos dados devem atentar-se à obrigação da implementação de medidas de segurança adequadas para a proteção de dados. Esta proteção pode ser alcançada através do auxílio de técnicas como criptografia, mascaramento e anonimização — esta última é o foco deste estudo.

Sendo assim, o desenvolvimento deste Trabalho de Conclusão de Curso deu-se sobre o seguinte questionamento: Como a anonimização de dados impacta a performance

de modelos de *machine learning*?

Quando informações pessoais são submetidas a técnicas de anonimização, elas deixam de ser consideradas dados sensíveis e, portanto, não precisam mais obedecer às regras das legislações. A GDPR prevê a possibilidade e descreve:

> Os princípios da proteção de dados não deverão, pois, aplicar-se às informações anónimas, ou seja, às informações que não digam respeito a uma pessoa singular identificada ou identificável nem a dados pessoais tornados de tal modo anónimos que o seu titular não seja ou já não possa ser identificado [União Europeia 2016].

Usar dados anonimizados durante o processo de treinamento de modelos de aprendizagem de máquina significa diminuir a possibilidade da violação de privacidade de indivíduos. Apesar disto, quando opta-se pelo uso dessas técnicas, um impacto significativo na precisão destes modelos poderá ser sentida, porque, ao generalizar ou suprimir informações, com a anonimização, a qualidade dos dados disponibilizados para o treinamento dos modelos pode ser afetada o que, consequentemente, pode diminuir sua precisão. ”A anonimização apresenta um dilema intrínseco: a proteção da privacidade versus a preservação da utilidade dos dados”[Aufschläger et al. 2023]

Pretendeu-se, com a pesquisa desenvolvida, explorar a relação entre a proteção de dados sensíveis e o treinamento de modelos de *machine learning*, ambos cada vez mais correlacionados nos debates e nas produções feitas no decorrer dos próximos anos.

2. Referencial Teórico

Para que fosse possível dar início aos estudos e experimentos sobre anonimização e machine learning, anteriormente, foi realizada uma pesquisa para levantamento da literatura relacionada acerca dos assuntos. A busca foi realizada em três bases: IEEE Xplore, Periódicos Capes e Papers With Code. A pesquisa justifica-se pois mapear o estado da arte é o primeiro passo para a execução de um projeto de pesquisa, porque como argumenta [Fonseca 2002], “permite ao pesquisador conhecer o que já se estudou sobre o assunto” — e neste conhecer descrito por ele, estão inclusas as descrições dos dois conceitos basilares para este trabalho.

2.1. Machine Learning

Definida por [Rebala et al. 2019] como o “campo da ciência da computação que estuda algoritmos e técnicas para automatizar soluções de problemas complexos que são difíceis de programar usando métodos convencionais”, machine learning (ou aprendizagem de máquina, em português), em resumo, é o processo através do qual o computador encontra padrões por ter sido submetido a um treinamento com dados tornando-se capaz de fazer previsões ou decisões. Há três tipos principais de aprendizado em machine learning: (1) supervisionado; (2) não supervisionado; e, (3) por reforço — será foco deste trabalho apenas o primeiro deles.

A aprendizagem supervisionada é entendida como

> uma tarefa que assume uma função a partir dos dados de treinamento rotulados. No aprendizado supervisionado, há uma variável de entrada (P) e uma variável de saída (Q). A partir da variável de entrada, a função do algoritmo é estudar a função de mapeamento para a variável de saída Q = f(P). O objetivo do aprendizado supervisionado é analisar os dados de treinamento para produzir uma função completa que possa ser utilizada para mapear novas instâncias. O algoritmo de aprendizado será capaz de analisar e generalizar corretamente os rótulos na classe a partir das instâncias não observadas. [Thomas and Gupta 2020]

Dessa forma, um modelo de aprendizagem supervisionada pode, por exemplo, ser capaz de prever se um e-mail é ou não spam. A partir de um dataset rotulado com entrada e saída, o modelo em questão será treinado e então estará apto para classificar a tal mensagem não previamente mapeada de forma precisa. Além do treinamento, o modelo será submetido a testes, que só poderão ser executados pois tal dataset foi dividido, desde o início do processo, em duas partes, uma que foi dedicada ao treinamento do sistema e a outra que será responsável por definir a sua acurácia.

Contudo, para isto, é necessário definir um algoritmo a depender da finalidade do modelo em questão. Estes podem ser categorizados como de classificação e de regressão. Conforme descreve [Nasteski 2017], enquanto no primeiro este valor é discreto, ou seja, pertence a um conjunto limitado de categorias, no segundo este valor é contínuo dentro de um intervalo.

Neste projeto, foram utilizados os seguintes algoritmos:

- **Support Vector Machine (SVM):** No aprendizado de máquina, as SVMs são modelos supervisionados com algoritmos associados que analisam dados para realizar tarefas de classificação e regressão. Além de executarem classificação linear, as SVMs podem realizar classificação não linear de maneira eficiente por meio do chamado ”truque do kernel”, que mapeia implicitamente as entradas em espaços de características de alta dimensionalidade. Basicamente, elas desenham margens entre as classes, garantindo que a distância entre a margem e as classes seja maximizada, o que minimiza o erro de classificação. [Mahesh 2019]
- **Decision Tree:** Uma árvore de decisão é representada graficamente como uma estrutura em forma de árvore, onde os nós correspondem a eventos ou escolhas, e as arestas indicam as regras de decisão ou condições. Cada árvore é composta por nós e ramos; os nós representam os atributos a serem classificados, enquanto os ramos indicam os valores que esses nós podem assumir. [Mahesh 2019]
- **Random Forest:** Uma árvore de decisão é representada graficamente como uma estrutura em forma de árvore, onde os nós correspondem a eventos ou escolhas, e as arestas indicam as regras de decisão ou condições. Cada árvore é composta por nós e ramos; os nós representam os atributos a serem classificados, enquanto os ramos indicam os valores que esses nós podem assumir. [Breiman 2001]
- **Naive Bayes:** A técnica de classificação baseada no Teorema de Bayes assume independência entre os preditores. Em termos simples, um classificador Naive Bayes presume que a presença de uma determinada característica em uma classe é independente da presença de qualquer outra característica. Naive Bayes é amplamente utilizado na indústria de classificação de textos, sendo aplicado princi-

palmente para fins de clustering e classificação, com base na probabilidade condicional de ocorrência. [Mahesh 2019]

- **k-Nearest Neighbors (kNN):** O algoritmo k-nearest neighbors (KNN) é um algoritmo simples e supervisionado de machine learning que pode ser utilizado tanto para resolver problemas de classificação quanto de regressão. Ele é fácil de implementar e entender, porém apresenta uma desvantagem significativa: torna-se consideravelmente mais lento à medida que o tamanho dos dados utilizados aumenta. [Mahesh 2019]

2.2. Anonimização

A anonimização é descrita por [Christen et al. 2020] como o processo de criptografar, modificar, mascarar ou remover informações identificáveis de modo que as entidades se tornem anônimas. Os autores continuam afirmando que os dados só podem ser considerados anonimizados quando não permitem a reidentificação dos indivíduos aos quais se referem por qualquer processamento adicional ou pela vinculação desses dados com outras informações públicas que estejam disponíveis ou que provavelmente estarão disponíveis no futuro.

Eles continuam argumentando que, com o passar dos anos, os estudos sobre anonimização ganharam força dada a necessidade de proteger as informações pessoais dos usuários, como nome, endereço e data de nascimento. Tal como [Ferreira 2023], que afirma que

A documentação e a implementação de modelos seguros de anonimização de dados pessoais são resultados importantes para garantir tanto a segurança e proteção dos dados armazenados, como a possibilidade de acesso eficiente a essas informações, quando se trata de acessos legítimos [Ferreira 2023].

Este trabalho utilizou especificamente duas técnicas de anonimização: generalização e supressão.

O Guia de Técnicas Básicas de Anonimização de Dados elaborado pela [Singapore Personal Data Protection Commission 2018] descreve supressão como remoção de uma parte inteira dos dados em um conjunto e generalização como uma "redução deliberada na precisão destes" — como por exemplo converter a idade de uma pessoa em um intervalo.

3. Metodologia

O início do processo experimental deu-se com a escolha da base de dados a ser utilizada para a comparação da acurácia entre o treinamento com dados anonimizados e dados puros. Estes conjuntos foram extraídos do *Kaggle* (2024), uma comunidade de compartilhamento de *datasets*, modelos, discussão e cursos de *machine learning* e inteligência artificial.

A base de dados escolhida foi a *AIDS Virus Infection Prediction* (1996). A descrição dada pela plataforma é de que o conjunto "contém estatísticas de saúde e informações categóricas sobre pacientes que foram diagnosticados com AIDS". O estudo foi feito com 2.139 participantes e descreveu 23 atributos diferentes para cada um deles, sendo destes 521 soropositivos.

Quanto às outras 1.618 pessoas que fazem parte do estudo, é informado, apenas, que estas não são pacientes com AIDS, e entendendo essa que é uma doença causada pela infecção do Vírus da Imunodeficiência Humana (HIV), não é possível concluir se estas são ou não portadoras do tal vírus.

A seguir, as características são apontadas e descritas:

- **time**: "Tempo até a falha ou censura" [Hammer et al. 1996].
- **trt**: "Indicador de tratamento, onde 0 representa apenas ZDV, 1 representa ZDV + ddI, 2 representa ZDV + Zal, e 3 representa apenas ddI" [Hammer et al. 1996].
- **age**: "Idade em anos no início do estudo" [Hammer et al. 1996].
- **wtkg**: "Peso em quilogramas no início do estudo" [Hammer et al. 1996].
- **hemo**: "Indicador de hemofilia (0 = não, 1 = sim)" [Hammer et al. 1996].
- **homo**: "Indicador de atividade homossexual (0 = não, 1 = sim)".
- **drugs**: "Histórico de uso de drogas intravenosas (0 = não, 1 = sim)" [Hammer et al. 1996].
- **karnof**: "Pontuação de Karnofsky, que varia de 0 a 100" [Hammer et al. 1996].
- **oprior**: "Terapia antirretroviral não ZDV antes de 175 dias (0 = não, 1 = sim)" [Hammer et al. 1996].
- **z30**: "Uso de ZDV nos 30 dias anteriores a 175 dias (0 = não, 1 = sim)" [Hammer et al. 1996].
- **preanti**: "Dias de terapia antirretroviral pré-175" [Hammer et al. 1996].
- **race**: "Raça (0 = Branco, 1 = não branco)" [Hammer et al. 1996].
- **gender**: "Gênero (0 = feminino, 1 = masculino)" [Hammer et al. 1996].
- **str2**: "Histórico de antirretroviral (0 = ingênuo, 1 = experiente)" [Hammer et al. 1996].
- **strat**: "Estratificação do histórico de antirretroviral, onde 1 representa 'Antirretroviral Naive', 2 representa '¿ 1 mas ¡= 52 semanas de terapia antirretroviral prévia', e 3 representa '¿ 52 semanas'" [Hammer et al. 1996].
- **symptom**: "Indicador sintomático (0 = assintomático, 1 = sintomático)" [Hammer et al. 1996].
- **treat**: "Indicador de tratamento, onde 0 representa apenas ZDV e 1 representa outros tratamentos" [Hammer et al. 1996].
- **offtrt**: "Indicador de interrupção do tratamento antes de 96 +/- 5 semanas (0 = não, 1 = sim)" [Hammer et al. 1996].
- **cd40**: "Contagem de CD4 no início do estudo" [Hammer et al. 1996].
- **cd420**: "Contagem de CD4 em 20 +/- 5 semanas" [Hammer et al. 1996].
- **cd80**: "Contagem de CD8 no início do estudo" [Hammer et al. 1996].
- **cd820**: "Contagem de CD8 em 20 +/- 5 semanas" [Hammer et al. 1996].
- **infected**: "Indicador de infecção por AIDS (0 = não, 1 = sim)" [Hammer et al. 1996].

Tais atributos foram divididos em quatro categorias, que foram:

- Informações pessoais (idade, peso, raça, gênero, atividade sexual).
- Histórico médico (hemofilia, histórico de uso de drogas intravenosas).
- Histórico de tratamento (histórico de tratamento com ZDV/não-ZDV).
- Resultados laboratoriais (contagens de CD4/CD8).

Após a escolha da base de dados foi realizada a escolha dos algoritmos para implementação dos modelos de *machine learning*. Foram utilizados os já detalhados SVM, *Decision Tree*, *Random Forest*, *Naive Bayes* e kNN.

Deu-se, então, início ao processo de anonimização dos dados. Foi utilizada a ferramenta ARX - *Data Anonymization Tool*. Descrita por seus distribuidores como um "software abrangente de código aberto para anonimizar dados pessoais confidenciais" [ARX 2024], a ferramenta permite a transformação dos dados através do uso de algoritmos e técnicas já inerentes ao sistema. Para o projeto, foram utilizadas as técnicas de generalização e de supressão, além do *privacy model population uniqueness* modelo Dankar com limite configurado a 0.01.

O algoritmo *population uniqueness* disponível na ARX Anonymization Tool é utilizado para estimar a probabilidade de que uma determinada combinação de atributos no dataset seja única em relação a uma população de referência. Esse conceito está relacionado à vulnerabilidade de reidentificação de indivíduos, especialmente quando dados anonimizados são cruzados com outras fontes de dados externas.

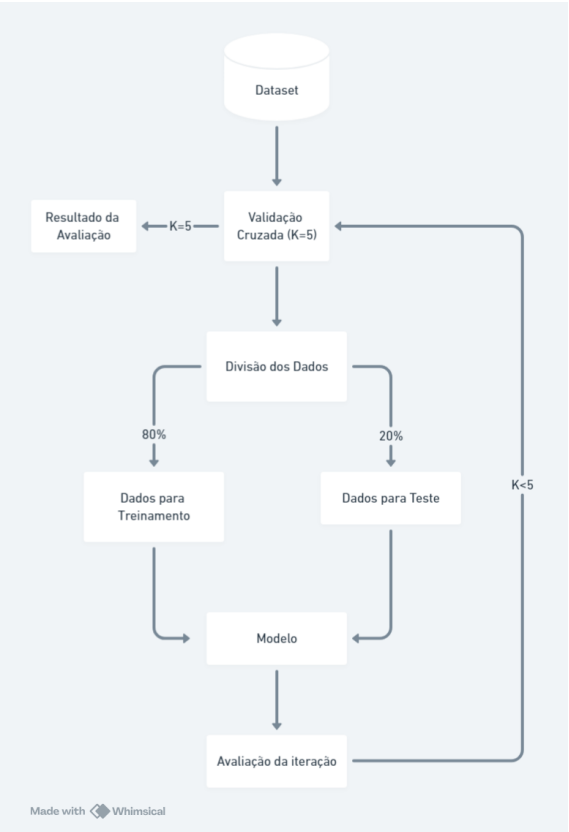
A ideia central do algoritmo é que, quanto mais rara ou única for uma combinação de atributos quasi-identificadores, maior é a probabilidade de que um indivíduo possa ser identificado dentro de uma população mais ampla. Para evitar isso, o algoritmo utiliza modelos estatísticos para calcular a probabilidade de reidentificação, levando em conta fatores como: distribuição de frequência, ajustes estatísticos e estimativa de risco.

Os atributos foram classificados como *insensitive*, *quasi-identifying* e *identifying*. Abaixo, são descritos um a um:

- **Insensitive**: Os classificados desta forma permanecem tal qual seus originais. Estes foram, *trt*, *hemo*, *karnof*, *oprior*, *z30*, *race*, *gender*, *str2*, *strat*, *symptom*, *treat* e *offtrt*.
- **Identifying**: Os dados agrupados neste parâmetro não constituem necessariamente dados que permitem a identificação direta de um indivíduo (que é a definição de dados identificadores). Contudo, foram assim distribuídos para que fosse possível suprimir as informações de tais, isto por serem dados de natureza sensível. Eles foram, *homo* e *drugs*.
- **Quasi-identifying**: São dados que, por si só, não identificam uma pessoa diretamente. No entanto, em conjunto a outras informações, indivíduos podem se tornar passíveis de identificação. Ao invés de serem postos como números exatos, foram transformados em um intervalo definido pelo modelo de privacidade previamente citado. Por exemplo, se um indivíduo tinha 33 anos, foi descrito como alguém com idade entre 30 e 35 anos. Estes foram, *time*, *age*, *wtkg*, *preanti*, *cd40*, *cd420*, *cd80* e *cd820*.

Feito isto, se iniciou o processo de implementação dos modelos de aprendizado de máquina com base nos algoritmos previamente escolhidos. A implementação foi feita a partir do uso da linguagem de programação Python na plataforma *Google Colab* — isto para ser possível determinar a acurácia dos dados puros e dos dados anonimizados em cada um dos algoritmos de classificação. Além disso, foi feito o uso de duas bibliotecas, a saber *Pandas* e *Sklearn*, e, deste último, os submódulos *classification_report*, *accuracy_score*, *train_test_split*, *StandardScaler*, específicos para os algoritmos supracitados.

Figura 1. Processo de implementação do modelo



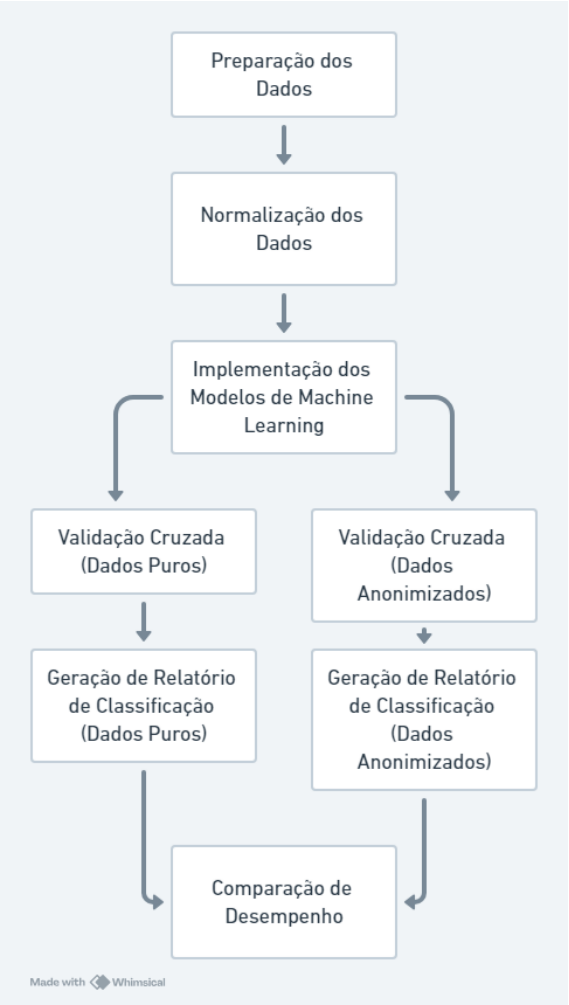
Fonte: Elaboração própria (2024).

Para o treinamento dos modelos, foram seguidos os seguintes passos:

1. **Preparação dos dados**: Inicialmente, o conjunto de dados puros e anonimizados foram carregados em duas estruturas diferentes, sendo feita a divisão entre as variáveis preditórias e as dependentes para a realização das demais etapas.
2. **Normalização dos dados**: Durante o processo de teste das implementações, foi observada a melhora de desempenho no treinamento dos modelos ao usar dados normalizados. A normalização envolve ajustar os valores para que fiquem em uma escala similar padronizada. Dessa forma, os dados utilizados para o treinamento e teste, este último contendo também dados anonimizados, foram submetidos, com o auxílio da função *StandardScaler* parte da biblioteca *sklearn*, ao processo de normalização.
3. **Implementação dos modelos**: Para a implementação dos modelos de *machine learning*, foi utilizada a biblioteca *Scikit-learn* (*Sklearn*). Foi criada uma instância do modelo para cada um dos algoritmos previamente citados.
4. **Validação cruzada**: Com a instância de cada modelo criada, o próximo passo foi a utilização da técnica de validação cruzada K-Fold, sendo o K definido como 5. Em cada iteração, um subconjunto é utilizado para teste, enquanto os outros quatro são usados para treinamento. Esse processo é repetido 5 vezes, e a média dos resultados é calculada. Ademais, foi utilizado o parâmetro *shuffle* para garantir que os dados sejam embaralhados antes de cada divisão, garantindo uma melhor generalização.

5. **Geração de relatório de classificação com o desempenho**: Por fim, foram gerados relatórios de classificação para os resultados das previsões em ambos os conjuntos de dados (puros e anonimizados). Eles incluíram métricas de desempenho como precisão, *recall* e *F1-score*. Comparando-as, foi possível avaliar o impacto das técnicas de anonimização na performance dos modelos de *machine learning*.

Figura 2. Fluxograma de Treinamento e Avaliação de Modelos de Machine Learning com Dados Puros e Anonimizados



Fonte: Elaboração própria (2024).

Após isto, os dados foram reunidos para posterior análise de tais resultados.

4. Resultados

Gerado o relatório de classificação com a acurácia de previsão de cada um dos modelos, chegou-se aos resultados descritos nas tabelas 1 e 2.

Para cada um dos algoritmos, o relatório gerou, ainda, informações específicas sobre a classificação. Tais informações serviram para fornecer uma visão detalhada sobre o comportamento do modelo em relação ao conjunto de dados. O que possibilitou,

Tabela 1. Comparação da Acurácia dos Dados Puros e Anonimizados

Algoritmo	Acurácia dos Dados Puros	Acurácia dos Dados Anonimizados
SVM	0,866	0,861
Decision Tree	0,849	0,847
Random Forest	0,890	0,890
Naive Bayes	0,822	0,821
kNN	0,821	0,822

Fonte: Elaboração própria (2024).

Tabela 2. Comparação do AUC dos Dados Puros e Anonimizados

Algoritmo	AUC dos Dados Puros	AUC dos Dados Anonimizados
SVM	0,888	0,886
Decision Tree	0,791	0,790
Random Forest	0,932	0,929
Naive Bayes	0,834	0,832
kNN	0,795	0,797

Fonte: Elaboração própria (2024).

inclusive, avaliar o desempenho do modelo em relação a cada classe em caso positivo e negativo de infecção.

Além da acurácia, as principais métricas presentes no relatório, e separadas por classe, foram precisão (proporção de verdadeiros positivos em relação ao total de previsões positivas feitas pelo modelo), recall (capacidade do modelo de identificar todos os exemplos positivos reais), f1-score (média harmônica entre precisão e recall) e support (número total de ocorrências de cada classe no conjunto de dados).

Por exemplo, para o algoritmo SVM foi gerado o seguinte relatório para os dados puros:

Tabela 3. Relatório de classificação do algoritmo SVM (dados puros)

Classe	Precisão	Recall	F1-Score	Support
0	0,89	0,94	0,91	1618
1	0,78	0,62	0,69	521

Fonte: Elaboração própria (2024).

Para os dados anonimizados, as informações foram as seguintes:

Logo, entende-se que a tabela mostra que os modelos em geral possuem uma

Tabela 4. Relatório de classificação do algoritmo SVM (dados anonimizados)

Classe	Precisão	Recall	F1-Score	Support
0	0,88	0,94	0,91	1618
1	0,77	0,61	0,68	521

Fonte: Elaboração própria (2024).

melhor performance na classificação de casos em que a infecção é negativa. O que ocorre, muito provavelmente, devido à distribuição dos dados, que se mostra maior em casos negativos do que positivos. Tabelas dos demais algoritmos são apresentadas na seção Anexos.

5. Análise dos dados e discussão

Embora a anonimização tenha alterado as características dos dados, o impacto desse processo na capacidade do modelo de prever corretamente as classes não foi negativo. É possível chegar a essa conclusão visto que as métricas de desempenho, como acurácia, AUC, precisão, recall e f1-score, dos modelos com dados anonimizados, apontam para uma performance próxima ou superior quando comparados com os de dados puros.

Outro ponto é a diferença de precisão nas classes. É possível afirmar que devido à distribuição não igualitária do dataset, com a discrepância considerável de mais de três quartos do total da quantidade de pessoas com AIDS (521) e sem AIDS (1.618), a precisão do modelo para identificar casos negativos é superior à sua capacidade para identificar casos positivos.

O AUC, por sua vez, sendo próximo de 1 em ambos os casos, sugere que o modelo possui uma boa capacidade de distinguir entre as classes, mesmo após a aplicação das técnicas de anonimização. A pequena diferença entre os valores da métrica em questão indica que a anonimização não comprometeu significativamente a habilidade do modelo de discriminar entre as classes positiva e negativa.

O resultado é significativo, pois demonstra que foi possível anonimizar dados sem sacrificar a qualidade das previsões — o que é crucial em contextos onde a privacidade dos dados é uma preocupação primária.

6. Conclusão

Apesar da crescente demanda por privacidade e segurança dos dados, especialmente em áreas sensíveis, o uso de dados pessoais é comum e necessário, sendo indispensável, assim, uma proteção robusta contra violações.

Os resultados obtidos a partir da análise comparativa entre entre dados normais e dados anonimizados, utilizando modelos de machine learning de cinco algoritmos diferentes para avaliar o impacto da anonimização na acurácia e em outras métricas relevantes de desempenho dos modelos de classificação, demonstraram que, embora a anonimização possa alterar as características dos dados, ela não necessariamente compromete a performance dos modelos de machine learning.

A partir das análises realizadas, foi possível concluir que a anonimização de dados é uma técnica viável para ser utilizada em cenários onde a cautela com a privacidade é

essencial, e isso sem que sejam causados prejuízos consideráveis ao desempenho dos modelos.

Entende-se, então, a importância do desenvolvimento e da aplicação de técnicas de anonimização como uma medida essencial na proteção de dados pessoais, especialmente em um contexto regulatório cada vez mais rigoroso, onde o respeito à privacidade e à segurança dos dados é agenda de governos e instituições privadas.

Este estudo, no entanto, encontra limitações no que diz respeito ao processo de anonimização em si. Por ter sido utilizada somente uma única técnica, a captura da totalidade dos impactos que os diferentes métodos de anonimização poderiam ter sobre o desempenho dos modelos pode não ter sido explicitada.

Métodos como anonimização baseada em *k-anonymity*, *l-diversity* ou *t-closeness*, por exemplo, podem ter efeitos variados nas métricas de desempenho, e a comparação entre essas técnicas poderia fornecer uma visão mais abrangente sobre quais métodos são mais adequados para diferentes tipos de dados e objetivos de proteção de privacidade.

Além disso, técnicas emergentes, como a *accuracy-guided anonymization* desenvolvida por [Goldsteen et al. 2020], oferecem avanços significativos em termos de utilidade alcançada, conforme seus resultados. [Gadotti et al. 2024] também mencionam que novas abordagens, como a *differential privacy* e a geração de dados sintéticos, têm se mostrado promissoras, oferecendo equilíbrio entre a preservação da privacidade e a minimização da perda de precisão dos modelos preditivos.

Investigações futuras a partir deste trabalho podem incluir uma análise mais detalhada do trade-off entre a perda de informação causada pela anonimização e o ganho em proteção de privacidade. Isso incluiria estudar o impacto da anonimização em modelos mais complexos, como redes neurais profundas, utilizar diferentes bases de dados e explorar o uso de técnicas de preservação da privacidade, como differential privacy, que oferecem garantias formais de privacidade.

Referências

ARX (2024). Arx - data anonymization tool. `https://arx.deidentifier.org/`.

Aufschläger, R., Folz, J., März, E., Guggumos, J., Heigl, M., Buchner, B., and Schramm, M. (2023). Anonymization procedures for tabular data: An explanatory technical and legal synthesis. *Information*.

Brasil (2020). *Lei Geral de Proteção de Dados*. Brasília.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Christen, P., Ranbaduge, T., and Schnell, R. (2020). *Linking Sensitive Data: Methods and Techniques for Practical Privacy-Preserving Information Sharing*. Springer, Austrália.

Ferreira, J. R. (2023). Aplicação da lei geral de proteção de dados com utilização de modelos de anonimização de dados em ambiente de nuvem pública. Master's thesis, Universidade de Brasília. Dissertação (Mestrado Profissional em Engenharia Elétrica), x, 44 f., il.

Fonseca, J. J. S. (2002). *Metodologia da Pesquisa Científica*. UEC, Fortaleza. Apostila.

Gadotti, A., Rocher, L., Houssiau, F., Creţu, A.-M., and de Montjoye, Y.-A. (2024). Anonymization: The imperfect science of using data while preserving privacy. *Science Advances*.

Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., and Farkash, A. (2020). Anonymizing machine learning models.

Hammer, S. et al. (1996). Aids virus infection prediction. `https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction`. Disponibilizada por Kaggle.

Lorenz-Spreen, P., Oswald, L., Lewandowsky, S., and et al. (2023). A systematic review of worldwide causal and correlational evidence on digital media and democracy. *Nature Human Behaviour*.

Mahesh, B. (2019). Machine learning algorithms - a review. *International Journal of Science and Research (IJSR)*, 9.

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons.B*, 4:51–62. University St. Kliment Ohridski - Bitola.

Rebala, G., Ravi, A., and Churiwala, S. (2019). Machine learning definition and basics. In *An Introduction to Machine Learning*, pages 1–17. Springer International Publishing.

Singapore Personal Data Protection Commission (2018). Guide to basic data anonymisation techniques. `https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf`. 39 p.

Thomas, R. N. and Gupta, R. (2020). A survey on machine learning approaches and its techniques. In *2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEEC)*, pages 1–6. IEEE.

União Europeia (2016). *General Data Protection Regulation*. Bruxelas.

União Europeia (2024). Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act). `https://artificialintelligenceact.eu/ai-act-explorer/`.

Anexos

Tabela 5. Relatório de classificação do algoritmo SVM (dados puros)

Classe	Precisão	Recall	F1-Score	Support
0	0,89	0,94	0,91	1618
1	0,78	0,62	0,69	521

Fonte: Elaboração própria (2024).

Tabela 6. Relatório de classificação do algoritmo SVM (dados anonimizados)

Classe	Precisão	Recall	F1-Score	Support
0	0,88	0,94	0,91	1618
1	0,77	0,61	0,68	521

Fonte: Elaboração própria (2024).

Tabela 7. Relatório de classificação do algoritmo Decision Tree (dados puros)

Classe	Precisão	Recall	F1-Score	Support
0	0,90	0,90	0,90	1618
1	0,70	0,68	0,69	521

Fonte: Elaboração própria (2024).

Tabela 8. Relatório de classificação do algoritmo Decision Tree (dados anonimizados)

Classe	Precisão	Recall	F1-Score	Support
0	0,90	0,90	0,90	1618
1	0,69	0,68	0,68	521

Fonte: Elaboração própria (2024).

Tabela 9. Relatório de classificação do algoritmo Random Forest (dados puros)

Classe	Precisão	Recall	F1-Score	Support
0	0,91	0,94	0,93	1618
1	0,81	0,72	0,76	521

Fonte: Elaboração própria (2024).

Tabela 10. Relatório de classificação do algoritmo Random Forest (dados anonimizados)

Classe	Precisão	Recall	F1-Score	Support
0	0,91	0,95	0,93	1618
1	0,81	0,71	0,76	521

Fonte: Elaboração própria (2024).

Tabela 11. Relatório de classificação do algoritmo Naive Bayes (dados puros)

Classe	Precisão	Recall	F1-Score	Support
0	0,88	0,89	0,88	1618
1	0,64	0,62	0,63	521

Fonte: Elaboração própria (2024).

Tabela 12. Relatório de classificação do algoritmo Naive Bayes (dados anonimizados)

Classe	Precisão	Recall	F1-Score	Support
0	0,88	0,89	0,88	1618
1	0,64	0,62	0,63	521

Fonte: Elaboração própria (2024).

Tabela 13. Relatório de classificação do algoritmo kNN (dados puros)

Classe	Precisão	Recall	F1-Score	Support
0	0,84	0,95	0,89	1618
1	0,72	0,43	0,54	521

Fonte: Elaboração própria (2024).

Tabela 14. Relatório de classificação do algoritmo kNN (dados anonimizados)

Classe	Precisão	Recall	F1-Score	Support
0	0,84	0,95	0,89	1618
1	0,72	0,44	0,54	521

Fonte: Elaboração própria (2024).