Metodologia para Avaliação da Anonimização Baseada em k-Anonimato nos Modelos de Aprendizado de Máquina

Kristtopher K. Coelho[1], **Maurício M. Okuyama**[1], **Michele Nogueira**[2], **Alex Borges Vieira**[3], **Edelberto Franco Silva**[3], **José Augusto M. Nacif**[1]

[1]Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa (UFV - Campus Florestal)

[2]Departamento de Ciência da Computação (DCC)
Universidade Federal de Minas Gerais (UFMG)

[3]Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Universidade Federal de Juiz de Fora (UFJF)

{kristtopher.coelho,mauricio.okuyama,jnacif}@ufv.br

michele@dcc.ufmg.br,alex.borges@ufjf.edu.br,edelberto.franco@ufjf.br

Abstract. *The increasing volume of sensitive data generated by various domains demands robust approaches to privacy protection. Anonymization based on k-anonymity stands out for mitigating the risks of re-identification of personal data. However, the impact on the performance of machine learning models is commonly neglected. This work proposes a novel comparative method to evaluate the effects of anonymization on the performance of machine learning models, considering privacy, information loss, and performance metrics. The results provide insights for developing and improving k-anonymization-based solutions to reconcile privacy and efficiency in distributed environments.*

Resumo. *O crescente volume de dados sensíveis, gerados por diversos domínios, exige abordagens robustas para proteção da privacidade. A anonimização baseada em k-anonimato se destaca por mitigar os riscos de reidentificação de dados pessoais. Entretanto, o impacto sobre o desempenho de modelos de aprendizado de máquina é comumente negligenciado. Este trabalho propõe um método comparativo inovador para avaliar os efeitos da anonimização sobre o desempenho de modelos de aprendizado de máquina, considerando métricas de privacidade, perda de informação e desempenho. Os resultados fornecem insights para o desenvolvimento e aprimoramento de soluções baseadas em k-anonimato para conciliar privacidade e eficiência em ambientes distribuídos.*

1. Introdução

O volume de dados produzidos e compartilhados em redes de computadores e sistemas distribuídos entre diversas áreas, como saúde, financeira, redes sociais, entre outras, está aumentando exponencialmente, alavancado pela propagação de dispositivos da Internet das Coisas (IoT). Parte significativa desses dados é composta por informações pessoais sensíveis, como dados de localização, etnia, condição de saúde, opiniões políticas, identidade de gênero ou orientação sexual [Slijepčević et al. 2021]. Existem muitos desafios em proteger estes dados em repouso (armazenados em discos ou bases de dados) ou em

trânsito (durante uma transmissão). O principal é garantir a proteção da privacidade dos indivíduos cujos dados estão sendo armazenados e processados, mantendo a respectiva utilidade [Coelho et al. 2024a]. É preciso garantir que a privacidade dos dados esteja consoante a regulamentações rigorosas. Entre elas estão a Lei Geral de Proteção de Dados (LGPD) no Brasil, o Regulamento Geral de Proteção de Dados (*General Data Protection Regulation*–GDPR) na União Europeia e o Health Insurance Portability and Accountability Act (HIPAA) nos Estados Unidos.

Diante da necessidade crítica de proteger a privacidade de dados, a anonimização baseada em k-anonimato se posicionou como uma abordagem amplamente utilizada para mascarar dados pessoais. Dado um conjunto de informações sobre uma pessoa, a anonimização consiste em remover, alterar, generalizar ou agregar informações que possam identificar uma pessoa. O objetivo é tornar os dados indistinguíveis dentro de um conjunto, de modo que não possam ser associados a nenhum indivíduo [Torra 2013]. Garantir que os registros anonimizados sejam indistinguíveis de outros $k-1$ registros dificulta a identificação de indivíduos, mesmo quando dados adicionais estão disponíveis. A anonimização por k-anonimato promove proteção contra ataques de reidentificação, ou inferência, onde um invasor tenta associar informações anonimizadas a indivíduos específicos usando conhecimento prévio ou dados auxiliares. Entretanto, o benefício da proteção da privacidade fornecido pela anonimização está associado proporcionalmente à perda de informação derivada da intensidade das modificações sobre as informações originais [Ghinita et al. 2007]. Ou seja, quanto maior o compromisso com a privacidade, maior a generalização dos dados e a respectiva perda de informações. Portanto, um dos desafios centrais na anonimização é manter o compromisso entre privacidade e utilidade dos dados. Este equilíbrio é alcançado ao encontrar um ponto onde a generalização é suficiente para garantir o k-anonimato, mas não excessiva a ponto de prejudicar a qualidade e precisão de análises estatísticas, além da eficiência de modelos classificadores.

As informações pessoais, principalmente em grandes volumes, fornecem insumos para empresas, como instituições médicas, universidades e organizações a desenvolverem soluções baseadas em aprendizado de máquina (ML) para propósito específico ou uso geral. Portanto, a privacidade tornou-se uma importante preocupação também no âmbito do aprendizado de máquina. Nesse contexto, o aprendizado federado (FL) surge como uma solução inovadora para treinamento descentralizado de modelos, mantendo os dados em dispositivos locais e reduzindo riscos de quebra de privacidade durante a transmissão. No entanto, a combinação de anonimização e aprendizado federado apresenta desafios significativos. A generalização introduzida por técnicas de k-anonimato pode impactar negativamente o desempenho dos modelos, reduzindo sua capacidade de aprendizado, a precisão classificadora, especialmente em ambientes distribuídos. Portanto, é crucial avaliar o impacto de múltiplas abordagens de anonimização sobre o desempenho global dos modelos FL.

Ao longo do tempo uma variedade de algoritmos para garantir o k-anonimato foram propostos, entretanto a definição da melhor abordagem ainda é desafiadora, principalmente quando o objetivo é combiná-la com ML. Manter a perda de informações o menor possível é crucial, especialmente para análises automatizadas por meio de métodos de ML, que visam derivar padrões significativos dos dados subjacentes. Os autores de [Slijepčević et al. 2021] conduziram uma investigação sobre os efeitos de diferen-

tes algoritmos de anonimização baseados em k-anonimato sobre resultados dos modelos de aprendizado de máquina centralizado. Os resultados corroboram que quanto maior o tamanho do conjunto de dados que partilham os mesmos valores entre os atributos (valor atribuído a k), maior a degradação do desempenho da classificação. Outros estudos como [Coelho et al. 2024b], exploram a avaliação do desempenho de modelos ML sobre dados anonimizados para validar os métodos de anonimização. Avaliações equivalentes para soluções baseadas em FL, também objetivam analisar propostas específicas, como em [Kwatra and Torra 2021, Choudhury et al. 2020]. Portanto, o impacto combinado da anonimização sobre modelos descentralizados ainda carece de estudos aprofundados, especialmente no que diz respeito à proposição de uma metodologia robusta a qual permita avaliar o equilíbrio entre a preservação da privacidade, a utilidade dos dados e a eficiência dos modelos treinados.

Este artigo propõe uma metodologia para avaliação do impacto da anonimização de dados baseada em k-anonimato em modelos de aprendizado de máquina, com foco na preservação da privacidade em redes distribuídas. Uma metodologia robusta permite analisar os efeitos da anonimização de dados em relação à qualidade dos modelos ML. O método de avaliação é composto por cinco partes fundamentais, as quais contemplam a compreensão do universo de dados, os modelos de anonimização, modelos de aprendizado de máquina, o levantamento de métricas adequadas e as análises resultantes. O cenário de avaliação investiga como a generalização introduzida por diferentes métodos de anonimização baseados em k-anonimato afeta o desempenho de modelos de aprendizado federado. O objetivo é analisar comparativamente o desempenho dos modelos treinados com dados anonimizados em relação aos treinados com dados originais, avaliando métricas de desempenho como acurácia, AUC, Loss, além da perda de informação e risco de reidentificação. A metodologia proposta oferece *insights* valiosos sobre o equilíbrio entre privacidade, utilidade dos dados e eficiência dos modelos em sistemas distribuídos, contribuindo para o avanço de soluções que atendam às demandas de segurança e desempenho em redes modernas.

O restante deste artigo está organizado da seguinte forma. A Seção 2 explora os trabalhos relacionados. A Seção 3 detalha a metodologia proposta para a avaliação do impacto da anonimização baseada em k-anonimato nos modelos de aprendizado de máquina. A Seção 4 descreve o cenário para avaliação empírica sobre como a generalização por k-anonimato afeta as soluções baseadas em aprendizado federado, incluindo a descrição da base de dados, métricas, materiais e métodos. A Seção 5 apresenta e discute os resultados. Por fim, a Seção 6 conclui o artigo.

2. Trabalhos Relacionados

Os modelos de anonimização de dados e sua aplicação em cenários de aprendizado de máquina exigem constantes investigações devido à crescente necessidade de proteger dados sensíveis enquanto se preserva a utilidade para modelo, principalmente no contexto *big data* IoT. Atualmente, modelos de aprendizado de máquina federado têm sido amplamente explorados na literatura por serem promissores para o treinamento de modelos em ambientes descentralizados, como instituições médicas, por exemplo. Modelos FL garantem que os dados permaneçam nos dispositivos locais, sem a necessidade de aglomeração em uma entidade centralizada. No entanto, o impacto combinado da anonimização e da

descentralização ainda necessita de estudos aprofundados, principalmente sobre o equilíbrio entre a preservação da privacidade e a eficiência dos modelos treinados.

Estudos recentes, como em [Slijepčević et al. 2021] realizam uma comparação sistemática e investigam o efeito de diferentes técnicas de anonimização no desempenho de modelos de aprendizado de máquina centralizados. Os algoritmos baseados em k-anonimato, os quais compõem a avaliação, consistem em Optimal Lattice Anonymization (OLA), Mondrian, Top-Down Greedy Anonymisation (TDG), Clustering-Based Anonymisation (CB). Além desses algoritmos de k-anonimato, diferentes classificadores foram avaliados, entre eles Support Vector Machine (SVM), k-nearest neighbors (kNN), random forest (RFs) e eXtreme Gradient Boosting (XGBoost). Os experimentos indicam que quanto maior o valor de k, maior a degradação do desempenho da classificação. Este trabalho ainda destaca o melhor desempenho do algoritmo Mondrian em detrimento aos demais algoritmos de anonimização investigados.

O *framework* Generalization First k-Member Clustering (GFKMC) [Coelho et al. 2024b] propõe uma abordagem inovadora para anonimização de dados, principalmente ligados à Internet das Coisas (IoT). A ferramenta proporciona a anonimização dinâmica dos atributos numéricos [Coelho et al. 2024a] e generalização hierárquica a quase identificadores categóricos paralelamente. Esta generalização antecipada reduz o custo computacional durante a geração dos k grupos, de modo a atender aos requisitos de k-anonimato. Os autores avaliam o impacto do método GFKMC em relação ao desempenho de algoritmos de aprendizado de máquina K-NN e XGBoost. Os resultados demonstram que a perda constante de informação da ferramenta GFKMC tem implicações positivas para o desempenho dos modelos testados, especialmente em comparação com métodos como Mondrian, TDG e CB.

Em relação aos estudo do impacto de diferentes técnicas de anonimização, no desempenho de modelos distribuídos, o trabalho de [Kwatra and Torra 2021] propõe uma estrutura para preservação de privacidade usando o k-anonimato viabilizado pelo Mondrian para um classificador de árvore de decisão em uma configuração de Aprendizado Federado. Os resultados medem a degradação da acurácia para valores de $k \leq 50$ e constatam que, para o Mondrian a perda de desempenho é aceitável para conjuntos de dados específicos, mesmo grandes valores k, assim como na avaliação para modelos centralizados de [Slijepčević et al. 2021]. O trabalho de [Saleh 2022] também investiga os efeitos da anonimização no desempenho do modelo de aprendizado de máquina. O algoritmo Mondrian foi eleito para implementar o modelo de k-anonimato como consenso geral. O algoritmo Multi-layer Perceptron (MLP) com arquitetura de 5 camadas foi implementado em configuração de aprendizado federado. Os resultados para a acurácia do modelo, indicam que o k-anonimato tem o melhor desempenho para modelos ML em relação aos modelos de privacidade l-diversidade e t-proximidade.

A importância de modelos de aprendizado de máquina explicáveis está aumentando porque os usuários querem entender as razões por trás das decisões em modelos baseados em dados. Em [Choudhury et al. 2020] os autores estudam o efeito de técnicas de privacidade de dados para modelos elaborados para explicabilidade. O algoritmo de anonimização considerado é o MDAV (Maximum Distance to Average Vector), um dos algoritmos para modelos k-anônimos usando a distância euclidiana para medir a homogeneidade entre diferentes registros. O estudo avalia como a explicabilidade foi afetada pela

anonimização dos dados usando a técnica para explicabilidade TreeSHAP. Os experimentos mostram que é possível manter até certo grau tanto a precisão quanto a explicabilidade.

A análise da literatura reforça a necessidade de investigar sistematicamente os efeitos da anonimização de dados sobre soluções baseadas em aprendizado de máquina. Devido a demanda, este artigo propõe uma metodologia robusta de avaliação da seleção dos modelos de anonimização e ML, os quais devem garantir a eficiência, utilidade e a preservação da privacidade durante o trânsito dos dados. Foram considerando métricas relevantes relacionadas à perda de informação e ao risco de reidentificação dos dados, e desempenho.

3. Metodologia para Avaliação da Anonimização Baseada em k-Anonimato nos Modelos de Aprendizado de Máquina

Garantir a privacidade e utilidade de dados é uma necessidade fundamental em redes de computadores e sistemas distribuídos, especialmente diante do crescente uso ML sobre dados anonimizados. Esta seção apresenta uma metodologia estruturada para avaliar o impacto de diferentes técnicas de anonimização, sobre o desempenho de modelos ML. A abordagem visa analisar o compromisso entre a utilidade dos dados e a eficiência dos modelos enquanto assegura a privacidade.

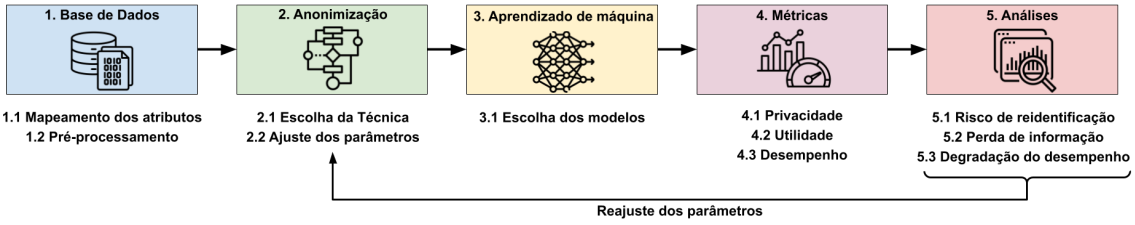


Figura 1. Estrutura da metodologia proposta

A Figura 1 ilustra o processo metodológico decomposto em cinco fases principais. A primeira fase do método consiste em compreender adequadamente o universo dos dados os quais pretende-se anonimizar. Estas informações sobre a caracterização da base de dados fornecem as diretivas essenciais para a escolha dos métodos de anonimização de dados e aprendizado de máquina. O mapeamento dos atributos consiste em categorizá-los conforme a caracterização consagrada na literatura [Domingo-Ferrer et al. 2022]. As categorias incluem atributos identificadores (PII - *Personally Identifiable Information*), quase identificadores e, sensíveis ou confidenciais. Os atributos identificadores são aqueles que, sozinhos, podem identificar um indivíduo diretamente, como CPF ou nome completo. Estes devem ser removidos da base ou mascarados para evitar reidentificação direta. Os atributos quase identificadores, sozinhos, não permitem a reidentificação direta de um indivíduo. No entanto, se combinados (endereço, sexo, idade), podem identificar um indivíduo e por isso devem ser anonimizados.

Atributos sensíveis ou confidenciais possuem informações consideradas sensíveis sobre os indivíduos, as quais são os objetos de interesse. O objetivo das técnicas de anonimização é evitar que seja possível relacionar qualquer dado desses atributos a um indivíduo. Exemplos de atributos sensíveis são condições de saúde, orientação sexual e salário. Os atributos ainda podem ser divididos entre os tipos dos dados, numéricos ou

categóricos. Os atributos numéricos são aqueles que podem ser medidos em uma escala quantitativa. Ou seja, aplicar cálculos estatísticos numéricos sobre os valores faz sentido. Os atributos categóricos são aqueles que não possuem valores quantitativos, mas que classificam os indivíduos em diversas categorias. Além de conhecer detalhadamente as características dos dados, é igualmente importante conhecer as características da base. Espera-se que os atributos sejam fornecidos em uma estrutura de dados tabular potencialmente heterogênea com eixos rotulados (linhas e colunas) (índices e números). Ademais, é necessário haver um pré-processamento adequado, tratando dados ausentes e eliminando atributos desnecessários.

Compreender em profundidade o universo dos dados é essencial para estabelecer diretrizes técnicas que orientem a escolha das técnicas e modelos de anonimização mais adequados. A seleção da técnica de anonimização mais apropriada depende do contexto e dos requisitos específicos do conjunto de dados, considerando fatores como o nível de proteção à privacidade desejado e a necessidade de preservar a utilidade dos dados. Os modelos de privacidade, como k-anonimato e outros, têm como objetivo assegurar a privacidade dos indivíduos ao impor que, após o processo de anonimização, o conjunto de dados atenda a condições específicas de privacidade definidas por cada técnica. Para alcançar esse objetivo, tais modelos frequentemente utilizam transformações nos registros, empregando técnicas como supressão, mascaramento, generalização, agregação, entre outras. Diferentes técnicas podem ser combinadas para transformar os dados de modo a alcançar um equilíbrio entre privacidade e usabilidade teorizado pelos modelos.

O modelo k-anonimato assegura que cada registro no conjunto de dados seja indistinguível de pelo menos outros $k - 1$ registros com base em atributos quase-identificadores. O parâmetro k deve ser ajustado para equilibrar o nível de anonimato e a perda de utilidade dos dados. É importante ressaltar que não existem abordagens específicas para determinar um valor ótimo para o parâmetro k, sendo este um problema NP-difícil [Domingo-Ferrer et al. 2022]. Dessa forma, cabe ao programador esta complexa tarefa de escolha empírica do valor. Esta metodologia contribui com análises eficientes, as quais produzem *insights* para que um valor adequado seja inferido, proporcionando um reajuste eficiente dos parâmetros.

A etapa 3 da metodologia refere-se ao levantamento de requisitos fundamentais, os quais conduzem a escolha adequada do modelo de aprendizado de máquina. A preferência por um modelo ML, seja ele centralizado ou federado, deve ser orientada pelo objetivo e contexto do problema, bem como pelas características específicas da base de dados identificadas na etapa inicial desta metodologia. É fundamental determinar a natureza do problema para selecionar entre modelos de classificação, regressão ou outras abordagens. Além disso, a escolha deve considerar os recursos computacionais disponíveis, assegurando que o modelo escolhido seja capaz de atender aos requisitos de desempenho, eficiência e aplicabilidade definidos para o cenário em questão.

A avaliação da solução inclui examinar o conjunto de dados anonimizados para verificar se o compromisso entre a privacidade e utilidade dos dados foi assegurado. Considerando a avaliação da preservação da privacidade, é fundamental que métricas apropriadas sejam adotadas. Não limitado a estas, destacam-se entre elas, risco de reidentificação, que mede a probabilidade de um invasor identificar indivíduos nos dados anonimizados. E o cálculo do risco residual quantifica o risco potencial de reidentificação após

a aplicação de técnicas de anonimização, considerando a presença de dados auxiliares. Para a avaliação da preservação da utilidade dos dados, devem ser consideradas métricas as quais avaliem a quantidade de perda de informações sobre a base resultante. A métrica *Normalized Certainty Penalty* (NCP) avalia a perda de precisão em atributos devido à generalização ou supressão. A *Discernibility Metric* (DM) mede o impacto da anonimização no agrupamento de registros, penalizando grupos excessivamente grandes ou altamente generalizados. Outras medidas estatísticas como desvio de distribuição, correlação de atributos e medidas de similaridade também são úteis para mensurar as variações e distância entre os registros originais e a base anonimizada. Ademais, avaliar métricas as quais avaliam o desempenho de modelos ML também é imprescindível. Acurácia, precisão, área sob a curva (AUC), pontuação F1, entre outras são fundamentais para avaliar o desempenho de modelos ML treinados com dados anonimizados em comparação aos treinados com dados originais.

A etapa 5 consiste em realizar profundas análises sobre o compromisso entre a utilidade dos dados e risco de identificação. Este é um dos principais desafios na anonimização de dados, especialmente em modelos baseados em k-anonimato, onde o parâmetro k determina o nível de anonimato. Avaliações precisas, guiadas pelas métricas consideradas, permitem realizar ajustes contínuos do parâmetro k de modo a alterar a granularidade da anonimização garantindo o equilíbrio entre utilidade e privacidade de acordo com o contexto específicos de cada base e modelo.

4. Avaliação

Esta seção apresenta um cenário de testes para investigar o impacto de diferentes abordagens de anonimização em modelos de aprendizado federado. O uso de aprendizado federado é destacado pela sua capacidade de preservar a privacidade de dados em trânsito em redes de computadores e sistemas distribuídos, garantindo que as informações sensíveis permaneçam localmente nos dispositivos enquanto permitem a construção de modelos globais eficientes.

A base ADULT [Lhoest et al. 2021], amplamente utilizada em tarefas de classificação [Liu et al. 2021, Salmeron and Arévalo 2024] e anonimização baseada em k-anonimato [Khan et al. 2020, Torra and Navarro-Arribas 2023], foi escolhida como estudo de caso. Seu principal objetivo é classificar se a renda anual de um indivíduo excede ou não 50 mil dólares, com base em atributos demográficos e ocupacionais. Para este trabalho, foram considerados os quase identificadores (sexo, idade, raça, estado civil, educação, país de origem, classe trabalhadora, ocupação) e o atributo sensível (classe salarial). Durante o pré-processamento, atributos não informativos foram removidos, garantindo maior consistência na análise.

O FL foi escolhido como abordagem central devido à sua capacidade de treinar modelos sem a necessidade de compartilhar dados entre os dispositivos participantes, mitigando os riscos de exposição de informações sensíveis em trânsito. No experimento, os registros da base foram divididos de forma independente e identicamente distribuída (IID) entre dez clientes, representando as unidades de federação. A propriedade IID dos dados desempenha um papel fundamental na determinação da precisão, confiabilidade e velocidade de convergência dos modelos de aprendizado de máquina [Qi et al. 2024]

Cada cliente aplicou localmente uma das quatro técnicas de anonimização em seu

subconjunto de dados: GFKMC [Coelho et al. 2024b], Mondrian [LeFevre et al. 2006], CB e TDG [Slijepčević et al. 2021]. Essas técnicas variam em abordagem, desde particionamento recursivo (como no Mondrian) até métodos baseados em agrupamento (como GFKMC e CB). A anonimização foi realizada para diferentes níveis de k-anonimato ($k = 3, 5, 10$ e 20), permitindo avaliar o impacto de graus variados de generalização. O valor atribuído ao k em k-anonimato indica a quantidade mínima de registros que cada registro de um conjunto de dados deve ser indistinguível. Os valores escolhidos para k, visam equilibrar proteção à privacidade e preservação da utilidade dos dados [Victor and Lopez 2020, Torra and Navarro-Arribas 2023].

O treinamento federado foi conduzido utilizando dois modelos: regressão logística (RL), que é um modelo linear amplamente utilizado por sua simplicidade e eficiência em tarefas de classificação binária, e XGBoost, um algoritmo baseado em árvores de decisão, escolhido por sua capacidade de capturar relações complexas nos dados. O processo de treinamento foi realizado ao longo de 100 rodadas, com agregação federada (*Federated Averaging* - FedAVG) para a RL e *FedXgbBagging* para o XGBoost. Em cada rodada, pelo menos oito dos dez clientes participaram, assegurando uma ampla contribuição dos dados distribuídos.

Os experimentos incluíram duas condições principais: (i) treinamento com os dados originais, sem anonimização, servindo como baseline, e (ii) treinamento com dados anonimizados por cada uma das técnicas consideradas. O desempenho dos modelos foi avaliado pelas métricas acurácia, AUC e *Loss* e pelo impacto na convergência do modelo ao longo das rodadas de treinamento. Além disso, foram investigadas a perda de informação causada pela anonimização e o risco residual de reidentificação. Essas análises fornecem uma visão abrangente sobre o compromisso entre a privacidade garantida pela anonimização e a utilidade dos dados em sistemas distribuídos, destacando a eficiência do FL para redes modernas.

4.1. Métricas

A perda de informação dos algoritmos é avaliada pela métrica *Normalized Certainty Penalty* (NCP) [Ghinita et al. 2007]. NCP é uma métrica amplamente aceita para avaliar a perda de informação em algoritmos de anonimização de dados. Para atributos categóricos, o NCP da classe de equivalência e é definido como:

$$NCP_C(e) = \sum_{A_C \in T_C} \begin{cases} 0, & |LCA(e_{A_C})| = 1 \\ \frac{|LCA(e_{A_C})|}{|T_C|}, & v_1 \neq v_2 \end{cases} \quad (1)$$

onde T_C representa os atributos categóricos de toda a tabela T, A_C é um atributo categórico de T_C, e_{A_C} representa os valores de atributos categóricos para A_C em e, $LCA(e_{A_C})$ é a árvore enraizada no menor ancestral comum de e_{A_C}. $|LCA(e_{A_C})|$ é o número de nós folha em $LCA(e_{A_C})$, e $|T_C|$ é o número de atributos categóricos distintos de toda a tabela T. Para atributos numéricos, o NCP da classe de equivalência e é definido como:

$$NCP_N(e) = \sum_{A_N \in T_N} \frac{max(e_{A_N}) - min(e_{A_N})}{max(T_N) - min(T_N)} \quad (2)$$

onde T_N representa os atributos numéricos de toda a tabela T, A_N é um atributo categórico de T_N, e_{A_N} representa os valores de atributos categóricos para A_N em e, e $max(\dots)$ e $min(\dots)$ representam o valor máximo e mínimo, respectivamente.

A divulgação de atributos ocorre quando um adversário tenta obter mais informações sobre um indivíduo. Se o adversário puder comparar os registros de QI de um indivíduo com algum conhecimento prévio, isso pode levar à divulgação de identidade e à exposição de atributos sensíveis. A vinculação de registros [Torra 2013, Domingo-Ferrer et al. 2022] é uma métrica essencial para avaliar a vulnerabilidade a tais ataques. Este trabalho emprega uma vinculação de registros baseada em distância para avaliar o risco de divulgação de atributos. A métrica reflete o número de correspondências de registros em relação ao número total de registros. Seja $d(r1, T)$ uma função de distância entre um registro $r1$ da tabela de dados anonimizada T^* e os registros da tabela de dados original T. Então, para cada registro $r1 \in T^*$, calcule $argmin(d(r1, T))$, onde $argmin(d(r1, T))$ retorna o índice de registro r_i com o menor valor de distância. Se r_i corresponde ao índice de $r1$, ele é contado como uma correspondência. Neste trabalho, a função de distância especificada na Definição 1 é utilizada para atributos categóricos. Para atributos numéricos, a função de distância da definição 2 é utilizada. Vale ressaltar que, para aplicar esta última definição em valores de intervalos, a média do intervalo foi considerada.

Definição 1 ***Distância para valores categóricos*** *Para qualquer atributo categórico C na tabela de dados T, a distância entre dois valores $v_1, v_2 \in N$ é definida como:*

$$d_C(v_1, v_2) = \begin{cases} 0, & v_1 = v_2 \\ \frac{|LCA(v_1,v_2)|}{|Tree\_C|}, & v_1 \neq v_2 \end{cases} \tag{3}$$

onde $Tree\_C$ é a árvore de taxonomia para um atributo categórico C, e $|Tree\_C|$ é o número de nós folha de $Tree\_C$. $LCA(v_1, v_2)$ é o menor ancestral comum de v_1 e v_2, e $|LCA(v_1, v_2)|$ é o número de nós folha da árvore enraizada em $LCA(v_1, v_2)$.

Definição 2 ***Distância para valores numéricos*** *Para qualquer atributo numérico N na tabela de dados T, a distância entre dois valores $v_1, v_2 \in N$ é definida como:*

$$d_N(v_1, v_2) = \frac{|v_1 - v_2|}{max(N) - min(N)} \tag{4}$$

onde $max(N)$ e $min(N)$ se referem aos valores máximo e mínimo do atributo numérico N, respectivamente.

Os resultados das métricas de desempenho acurácia indicam a proporção de classificações corretas em relação ao total de exemplos e podem ser com auxílio do Flower *framework* [Beutel et al. 2020]. A ampla maioria dos trabalhos relacionados consideram utilização da acurácia como métrica de avaliação da degradação da eficiência do modelo, principalmente [Kwatra and Torra 2021]. Por outro lado, a Área Sob a Curva (AUC) analisa a capacidade de discriminação do modelo em tarefas classificatórias com dados anonimizados. A AUC fornece uma medida agregada de desempenho em todos os limiares de classificação possíveis. Ela avalia como o modelo é capaz de distinguir entre as classes em diversos níveis de sensibilidade e especificidade, oferecendo uma visão global da eficácia do modelo. Para modelos de regressão linear, a perda (*Loss*) descreve o quão erradas são as previsões de um modelo. A perda mede a distância entre as previsões do modelo e os valores reais. O objetivo de treinar um modelo é minimizar a perda, reduzindo-a ao seu menor valor possível.

4.2. Materiais e métodos

Os modelos FL foram implementados na linguagem Python, com auxílio majoritário do *framework* Flower [Beutel et al. 2020], a qual é uma estrutura específica para implementação, análise e avaliação de aplicações FL amigável. O conjunto de dados ADULT pode ser utilizado de forma natural pelo Flower através da biblioteca *flwr_datasets*. Esta biblioteca interage diretamente da plataforma Hugging Face Datasets [Lhoest et al. 2021], onde a comunidade de aprendizado de máquina colabora em modelos, conjuntos de dados e aplicações. Outras bibliotecas python como numpy, pandas, matplotlib e scipy também foram importantes para a implementação. O código-fonte está disponível no GitHub[1]. Os experimentos foram conduzidos em uma máquina com 32 GB de RAM e processador AMD Ryzen 5 PRO 5675U.

5. Resultados e Discussões

A análise da perda de informação foi realizada utilizando a métrica NCP, que avalia empiricamente o impacto da anonimização em cada unidade federada. Os resultados apresentados, com valores médios e desvios padrão para os dez clientes do ambiente federado, consideram os níveis de $k = 3, 5, 10$ e 20 são ilustrados na Figura 2.

O algoritmo GFKMC demonstrou uma perda de informação essencialmente constante, em torno de 21%, independente do valor de k. Em contraste, os algoritmos Mondrian, CB e TDG exibiram perdas crescentes à medida que o valor de k aumentava. Especificamente, os algoritmos Mondrian e TDG já começam a ter uma perda de informação maior que o GFKMC para valores de k superiores a cinco, enquanto o algoritmo CB se aproxima substancialmente. A característica de estabilidade na perda de informação, proporcionada pela anonimização antecipada no *framework* GFKMC, é um diferencial importante no contexto de sistemas distribuídos que dependem de altos níveis de consistência entre privacidade e utilidade dos dados.

Em relação ao risco de reidentificação, a Figura 3 demonstra que o algoritmo TDG alcançou o menor risco médio, enquanto GFKMC e CB apresentaram desempenhos equivalentes, ainda melhores que o Mondrian. Esses resultados evidenciam os desafios em equilibrar as métricas de perda de informação e risco de reidentificação. Uma maior perda de informação, associada a generalizações mais amplas, tende a reduzir os riscos de reidentificação, mas pode impactar a utilidade dos dados.

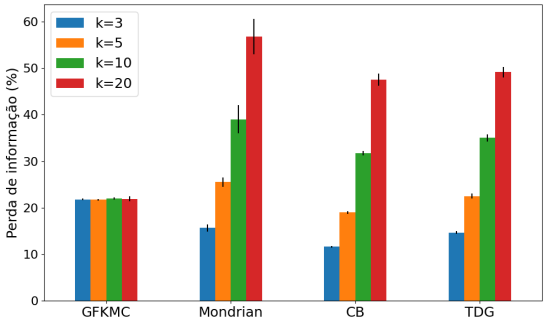


Figura 2. Perda de informação

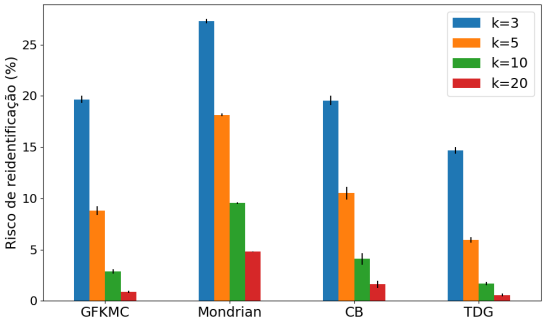


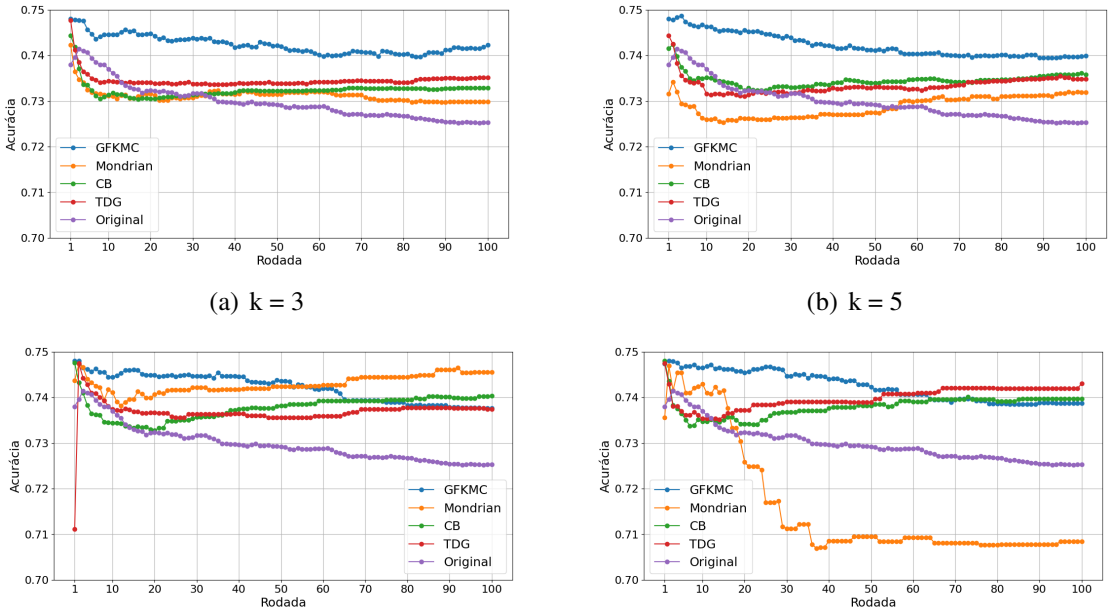
Figura 3. Risco de reidentificação

[1]https://github.com/mauriciokuyama/fl-anon

Os experimentos também avaliaram o impacto da anonimização sobre o desempenho de algoritmos de aprendizado federado, destacando a robustez do FL em sistemas distribuídos. O aumento da granularidade dos dados mostrou implicações mínimas para o desempenho dos modelos treinados. O *framework* GFKMC, por sua capacidade de equilibrar a perda de informações e a preservação de privacidade, apresentou um impacto significativamente menor no desempenho geral dos modelos, em comparação aos métodos Mondrian, CB e TDG.

A Figura 4 ilustra a convergência do modelo de RL a partir da rodada 20 de treinamento. O GFKMC obteve uma eficiência ligeiramente superior, com uma degradação mínima entre os valores de $k = 3, 5, 10$ e 20. A Figura 5 descreve o quão erradas são as previsões do modelo RL. As curvas obtidas mostram que há um rápido decaimento da perda. Portanto, os gráficos destacam a estabilidade na minimização da perda, mostrando que os modelos convergem de forma eficiente, mesmo com variações de k.

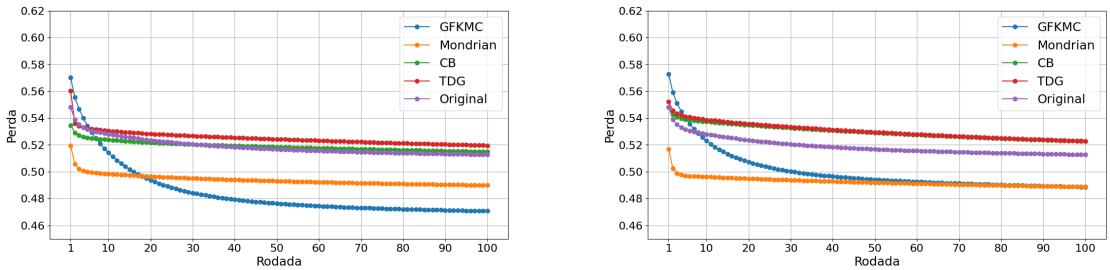
A Figura 6 permite observar a variação da acurácia durante as rodadas de treinamento do modelo XGBoost. Para este modelo, os resultados evidenciam uma leve degradação na eficiência em relação aos dados originais, especialmente conforme o valor de k aumenta. Apesar disso, o GFKMC manteve desempenho consistente, enquanto o Mondrian apresentou oscilações significativas para valores de $k \geq 10$. A Figura 7 complementa essa análise, mostrando que o XGBoost alcançou alta eficiência para $k = 3$ e 5, com forte capacidade de distinção entre classes. Entretanto, este desempenho começa a sofrer degradação para valores de $k \geq 10$, principalmente com o Mondrian apresentando grandes oscilações. Contudo, ressalta-se o desempenho do GFKMC, o qual mantém um desempenho médio eficiente para todos os valores de k.



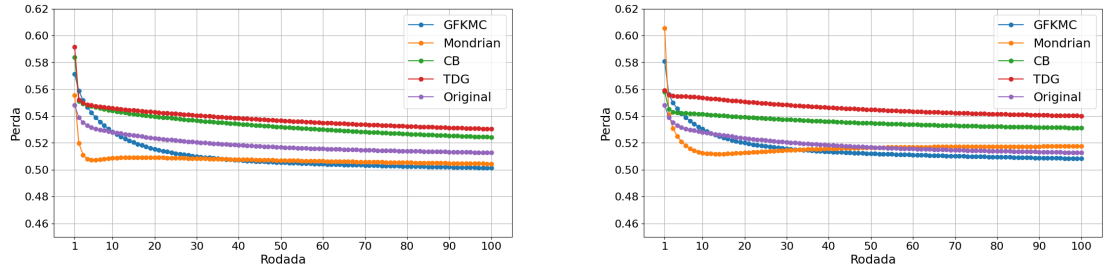
(a) k = 3 (b) k = 5

(c) k = 10 (d) k = 20

Figura 4. Acurácia do modelo Regressão logística para múltiplos métodos de anonimização

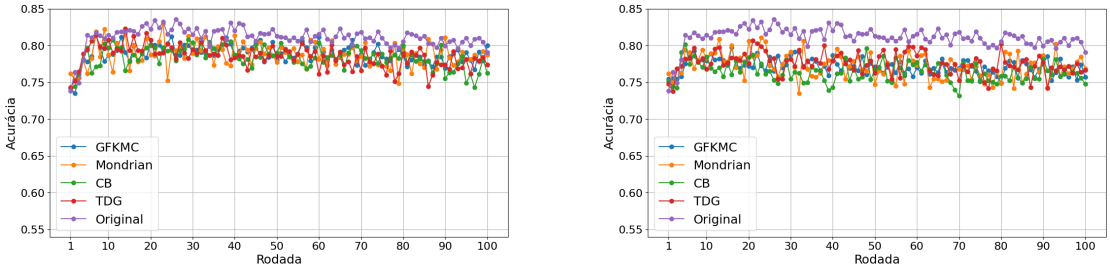


(a) k = 3 (b) k = 5

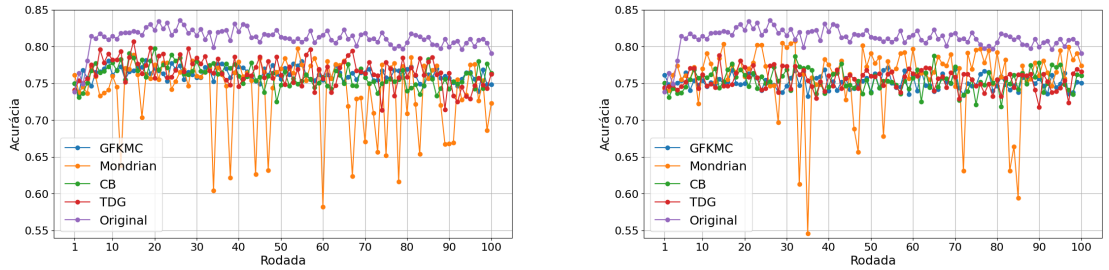


(c) k = 10 (d) k = 20

Figura 5. Perda do modelo Regressão logística para múltiplos métodos de anonimização



(a) k = 3 (b) k = 5

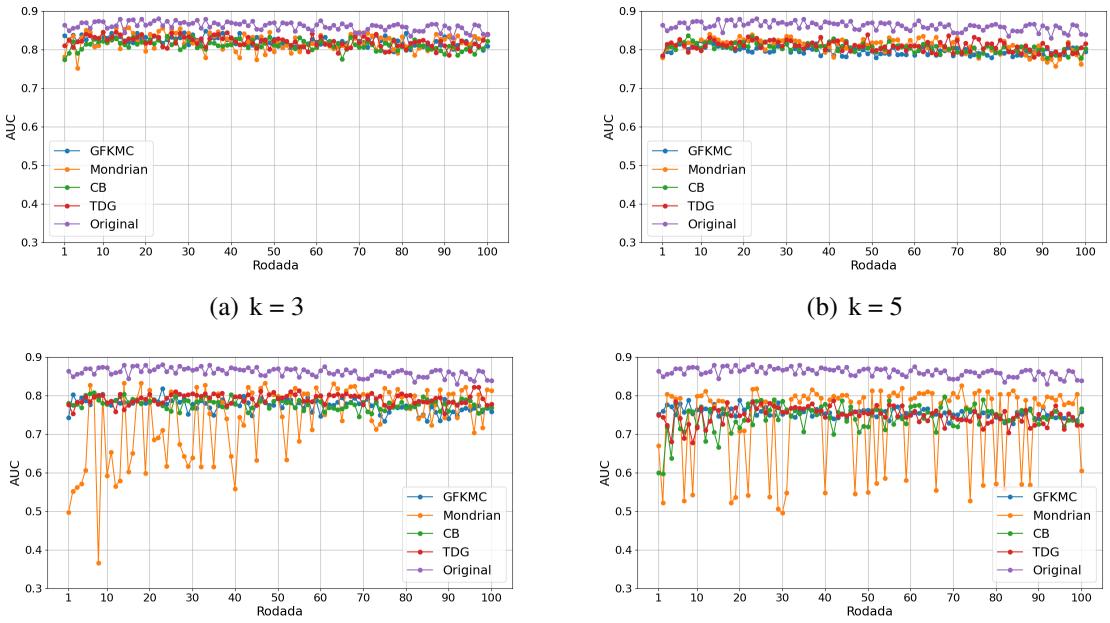


(c) k = 10 (d) k = 20

Figura 6. Acurácia do modelo XGBoost para múltiplos métodos de anonimização

6. Conclusões

A proteção de dados sensíveis em ambientes distribuídos, apresenta desafios críticos, principalmente na busca de um equilíbrio entre privacidade e utilidade dos dados. Este trabalho propôs uma metodologia robusta para avaliar o impacto da anonimização baseada em k-anonimato sobre a eficiência de modelos FL. O cenário de teste focou na análise



(a) k = 3 (b) k = 5

(c) k = 10 (d) k = 20

Figura 7. AUC do modelo XGBoost para múltiplos métodos de anonimização

de diferentes técnicas de anonimização, destacando a importância do FL em preservar a privacidade dos dados em trânsito entre sistemas distribuídos. Os resultados indicaram que níveis mais altos de anonimização aumentam a perda de informação usando técnicas tradicionais como Mondrian, CB e TDG. No entanto, o *framework* GFKMC demonstrou uma perda de informação constante, o que favoreceu seu desempenho, proporcionando maximizar a eficiência de modelos de classificação ML federados. Essa característica torna o GFKMC uma opção altamente relevante para anonimização de dados para sistemas distribuídos que exigem um compromisso entre privacidade, utilidade e estabilidade de dados. Como trabalhos futuros, propõe-se avaliar modelos ML mais complexos e modelos diversificados de anonimização, como l-diversidade em bases de dados heterogêneas, garantindo a utilidade e privacidade de dados em sistemas distribuídos.

Agradecimentos

Gostaríamos de agradecer o apoio financeiro da CAPES, Fapemig, Fapesp e CNPq.

Referências

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Li, K. H., Parcollet, T., de Gusmão, P. P. B., et al. (2020). Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*.

Choudhury, O., Gkoulalas-Divanis, A., Salonidis, T., Sylla, I., Park, Y., Hsu, G., and Das, A. (2020). Anonymizing data for privacy-preserving federated learning. *arXiv preprint arXiv:2002.09096*.

Coelho, K. K., Okuyama, M. M., Nogueira, M., Vieira, A. B., Silva, E. F., and Nacif, J. A. M. (2024a). A dynamic approach to health data anonymization by separatrices. In *2024 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE.

Coelho, K. K., Okuyama, M. M., Nogueira, M., Vieira, A. B., Silva, E. F., and Nacif, J. A. M. (2024b). A new k-anonymity method based on generalization first k-member clustering for healthcare data. In *Transactions on Dependable and Secure Computing*.

Domingo-Ferrer, J., Sánchez, D., and Soria-Comas, J. (2022). *Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections*. Springer Nature.

Ghinita, G., Karras, P., Kalnis, P., and Mamoulis, N. (2007). Fast data anonymization with low information loss. In *Proceedings of the 33rd international conference on Very large data bases*, pages 758–769.

Khan, R., Tao, X., Anjum, A., Kanwal, T., Malik, S. U. R., Khan, A., Rehman, W. U., and Maple, C. (2020). θ-sensitive k-anonymity: An anonymization model for iot based electronic health records. *Electronics*, 9(5):716.

Kwatra, S. and Torra, V. (2021). A k-anonymised federated learning framework with decision trees. In *International Workshop on Data Privacy Management*, pages 106–120. Springer.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *22nd International conference on data engineering (ICDE'06)*, pages 25–25. IEEE.

Lhoest, Q., Del Moral, A. V., Jernite, Y., Thakur, A., Von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., et al. (2021). Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*.

Liu, G., Ma, X., Yang, Y., Wang, C., and Liu, J. (2021). Federaser: Enabling efficient client-level data removal from federated learning models. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pages 1–10. IEEE.

Qi, P., Chiaro, D., Guzzo, A., Ianni, M., Fortino, G., and Piccialli, F. (2024). Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems*, 150:272–293.

Saleh, T. E. (2022). Comparison of the effects of data privacy preserving methods on machine learning algorithms in iot. Master's thesis, Marmara Universitesi (Turkey).

Salmeron, J. L. and Arévalo, I. (2024). Blind federated learning without initial model. *Journal of Big Data*, 11(1):56.

Slijepčević, D., Henzl, M., Klausner, L. D., Dam, T., Kieseberg, P., and Zeppelzauer, M. (2021). k-anonymity in practice: How generalisation and suppression affect machine learning classifiers. *Computers & Security*, 111:102488.

Torra, P. (2013). *Information Fusion in Data Mining*. Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg.

Torra, V. and Navarro-Arribas, G. (2023). Attribute disclosure risk for k-anonymity: the case of numerical data. *International Journal of Information Security*, 22(6):2015–2024.

Victor, N. and Lopez, D. (2020). Privacy preserving sensitive data publishing using (k, n, m) anonymity approach. *Journal of communications software and systems*, 16(1):46–56.