UNIVERSIDADE FEDERAL DE VIÇOSA

Anonimização Baseada em *k*-anonimato para Garantir a Privacidade de Dados em Internet das Coisas Aplicada na Saúde

Kristtopher Kayo Coelho
Doctor Scientiae

VIÇOSA - MINAS GERAIS
2025

KRISTTOPHER KAYO COELHO

Anonimização Baseada em *k*-anonimato para Garantir a Privacidade de Dados em Internet das Coisas Aplicada na Saúde

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Doctor Scientiae*.

Orientador: Jose Augusto Miranda Nacif

Coorientadora: Michele Nogueira Lima

VIÇOSA - MINAS GERAIS
2025

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
C672a
2025

Coelho, Kristtopher Kayo, 1989-
Anonimização baseada em k-anonimato para garantir a privacidade de dados em internet das coisas aplicada na saúde / Kristtopher Kayo Coelho. – Viçosa, MG, 2025.
1 tese eletrônica (89 f.): il. (algumas color.).

Inclui apêndice.
Orientador: José Augusto Miranda Nacif.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Informática, 2025.
Referências bibliográficas: f. 78-86.
DOI: https://doi.org/10.47328/ufvbbt.2025.572
Modo de acesso: World Wide Web.

1. Banco de dados - Medidas de segurança. 2. Proteção de dados. 3. Internet das coisas. 4. Saúde - Processamento de dados. I. Nacif, José Augusto Miranda, 1978-. II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22. ed. 005.8

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

KRISTTOPHER KAYO COELHO

Anonimização Baseada em *k*-anonimato para Garantir a Privacidade de Dados em Internet das Coisas Aplicada na Saúde

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Ciência da Computação, para obtenção do título de *Doctor Scientiae*.

APROVADA: 25 de julho de 2025.

Assentimento:

Kristtopher Kayo Coelho
Autor

Jose Augusto Miranda Nacif
Orientador

Dedico este trabalho a Deus e a todos que, de alguma forma, estiveram ao meu lado, apoiando e fortalecendo minha caminhada.

AGRADECIMENTOS

A Deus, meu mais sincero agradecimento. Cada passo dessa jornada só foi possível graças à Sua presença e à força que me concedeu para perseverar. Esta etapa marcante da minha vida, o encerramento de um ciclo tão significativo, só foi alcançada com o apoio inestimável de muitas pessoas. Em primeiro lugar, expresso minha profunda gratidão aos meus orientadores, professores José Augusto M. Nacif e Michele Nogueira, por toda a paciência, dedicação, orientação e constante motivação ao longo de todo o desenvolvimento deste trabalho. Não poderia deixar de reconhecer também a valiosa colaboração dos professores Alex Borges e Edelberto Silva, cuja contribuição foi essencial em diversos momentos. Agradeço igualmente à Universidade Federal de Viçosa (UFV) pela oportunidade de realizar a pós-graduação, a todos os colegas do programa, integrantes do laboratório e estudantes da UFV, que com apoio e amizade estiveram presentes durante essa caminhada. Aos funcionários da universidade, que sempre se mostraram solícitos e prontos para superar os desafios que surgiram, meu muito obrigado. Minha família merece um agradecimento especial, pelo amor e apoio incondicional que sempre me dedicaram, especialmente meus pais e irmãos, cuja presença foi uma fonte constante de força e encorajamento. Aos amigos, que mesmo diante das dificuldades e da distância, não deixaram de estar presentes, incentivando e motivando, meu eterno reconhecimento.

Este trabalho foi realizado com o apoio das seguintes agências de pesquisa brasileiras: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001, Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Agradeço ainda à Rede Nacional de Ensino e Pesquisa (RNP) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo número 2024/04923-0, pelo apoio financeiro e institucional.

"Vencer sem correr riscos é triunfar sem glórias!" (Ayrton Senna)

RESUMO

COELHO, Kristtopher Kayo, D.Sc., Universidade Federal de Viçosa, julho de 2025. **Anonimização Baseada em *k*-anonimato para Garantir a Privacidade de Dados em Internet das Coisas Aplicada na Saúde**. Orientador: Jose Augusto Miranda Nacif. Coorientadora: Michele Nogueira Lima.

O avanço acelerado da microeletrônica e a popularização de dispositivos da Internet das Coisas (IoT) aplicados à área da saúde (IoHT) têm impulsionado a coleta, o armazenamento e a transmissão de grandes volumes de dados sensíveis, como informações clínicas e pessoais. Embora tais dados sejam essenciais para viabilizar monitoramentos, diagnósticos e decisões médicas mais precisas, sua manipulação inadequada pode expor pacientes a riscos significativos de privacidade, resultando em violações legais e prejuízos éticos. Nesse cenário, torna-se crucial desenvolver soluções que conciliem a proteção da privacidade em conformidade com as legislações vigentes, preservando simultaneamente a utilidade dos dados, mesmo em ambientes distribuídos e heterogêneos característicos da IoHT. Para enfrentar esses desafios, esta tese propõe um *framework* abrangente para realizar *k*-anonimização de dados, fundamentado na abordagem de generalização antecipada e no agrupamento de *k*-membros. O *framework Generalization First k-Member Clustering* (GFKMC) aplica, paralelamente, a anonimização dinâmica por separatrizes a quase identificadores numéricos e a generalização hierárquica a quase identificadores categóricos. Posteriormente, para atender aos requisitos do *k*-anonimato, é aplicado um procedimento de reagrupamento eficiente. Os resultados demonstram a eficácia do *framework* GFKMC quanto à utilidade dos dados, apresentando uma perda constante de informação, em torno de 21%, para diversos valores de *k*, superando os métodos tradicionais em cenários que exigem níveis mais elevados de privacidade. Simultaneamente, o GFKMC garante forte proteção da privacidade ao reduzir o risco de reidentificação individual. Além disso, o GFKMC comprova sua robustez quando aplicado a modelos de aprendizado de máquina, apresentando degradação mínima nas métricas de desempenho. Essas evidências demonstram que o GFKMC equilibra de forma eficiente o compromisso entre privacidade e utilidade dos dados, tornando-o uma ferramenta valiosa para aplicação prática em ambientes IoHT.

Palavras-chave: privacidade de dados; anonimização de dados; k-anonimato; aprendizado federado; ferramenta de gerenciamento de segurança; aprendizado de máquina; IoHT; IoT

ABSTRACT

COELHO, Kristtopher Kayo, D.Sc., Universidade Federal de Viçosa, July, 2025. ***k*-anonymity-Based Anonymization to Ensure Data Privacy in the Internet of Healthcare Things**. Adviser: Jose Augusto Miranda Nacif. Co-adviser: Michele Nogueira Lima.

The rapid advancement of microelectronics and the popularization of Internet of Healthcare Things (IoHT) devices have driven the collection, storage, and transmission of large volumes of sensitive data, such as clinical and personal information. While such data is essential for enabling more accurate monitoring, diagnosis, and medical decision-making, its improper handling can expose patients to significant privacy risks, resulting in legal violations and ethical harm. In this scenario, it is crucial to develop solutions that balance privacy protection in compliance with current legislation, while preserving data utility even in distributed and heterogeneous environments typical of the IoHT. To address these challenges, this thesis proposes a comprehensive framework for performing k-anonymization based on generalization first and k-member clustering. The Generalization First k-Member Clustering (GFKMC) framework applies dynamic anonymization by separatrices to numerical quasi-identifiers and hierarchical generalization to categorical quasi-identifiers. Subsequently, to meet the k-anonymity requirements, an efficient reclustering procedure is applied. The results demonstrate the effectiveness of the GFKMC framework in terms of data utility, yielding a consistent information loss of around 21% for various k values, outperforming traditional methods in scenarios where higher levels of privacy are required. Simultaneously, GFKMC ensures strong privacy by reducing the risk of individual re-identification. Furthermore, GFKMC proves its robustness when applied to machine learning models, achieving minimal degradation in performance metrics. These findings show that GFKMC effectively balances the trade-off between privacy and data utility, making it a valuable tool for practical application in IoHT environments.

Keywords: data privacy; data anonymization; k-anonymity; federated Learning; security management tool; machine learning; IoHT; IoT

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AKA	*Adaptive k-anonymity*
DAS	*Dynamic Anonymization by Separatrices*
ECG	Eletrocardiograma
FL	*Federated Learning*
GDPR	*General Data Protection Regulation*
GFKMC	*Generalization First k-Member Clustering*
GH	*Generalization Hierarchies*
HIPAA	*Health Insurance Portability and Accountability*
HIT	*Healthcare Information Technologies*
ID	Identificador
IL	*Information Loss*
IoHT	*Internet of Healthcare Things*
IoT	Internet das Coisas
k-NN	*k-Nearest Neighbour*
LGPD	Lei Geral de Proteção de Dados
MD	*Minimal distortion*
ML	*Machine Learning*
NCP	*Normalized Certainty Penalty*
PD	Privacidade Diferencial
PPG	Fotopletismografia
QIs	Quase identificadores
RF	*Random Forests*
SVM	*Support Vector Machines*
WEF	*Wearable-exercise-frailty*
WIL	*Weight Information Los*
WKMCA	*Weighted K-member clustering algorithm*
XGBoost	*Extreme Gradient Boosting*

SUMÁRIO

Capítulo 1

Introdução

Nos últimos anos, o setor de saúde tem experimentado um crescimento significativo nos investimentos em Tecnologias da Informação em Saúde (*Healthcare Information Technologies* – HIT), impulsionado pelo avanço da miniaturização de dispositivos e pela disseminação da Internet das Coisas (IoT). Esse cenário favoreceu o surgimento da Internet das Coisas aplicada à Saúde (*Internet of Healthcare Things* – IoHT). IoHT é um ecossistema contemporâneo caracterizado pela integração de múltiplas fontes de dados, provenientes, sobretudo, de dispositivos vestíveis e/ou implantáveis Alexander and Wang (2025). Tal integração contribui para o aumento exponencial na geração e compartilhamento de dados, em especial dados potencialmente confidenciais Olatunji et al. (2022); Karagiannis et al. (2024).

A análise de grandes volumes de dados (*big data*) no domínio da saúde tem promovido transformações profundas, aprimorando a precisão diagnóstica, apoiando a tomada de decisões clínicas. Além disso, viabiliza o monitoramento remoto de pacientes, otimiza tratamentos e, ainda, contribui para a redução de custos operacionais e individuais Abouelmehdi et al. (2018); Yan et al. (2021); Batko and Slezak (2022); Alexander and Wang (2025). Contudo, a natureza sensível desses dados impõe desafios significativos em relação à preservação da privacidade. O compartilhamento de informações pessoais privadas entre distintas instituições de saúde agrava ainda mais tais desafios, exigindo conformidade com rigorosos marcos regulatórios Kumari and Prabha (2025). Dentre as principais regulamentações, destacam-se: a Lei Geral de Proteção de Dados (LGPD), no Brasil Garcia et al. (2020), o Regulamento Geral sobre a Proteção de Dados (*General Data Protection Regulation* – GDPR), na União Europeia Voigt and Von dem Bussche (2017), e a Lei de Portabilidade e Responsabilidade de Seguro de Saúde (*Health Insurance Portability and Accountability Act* – HIPAA), nos EUA Nosowsky and Giordano (2006).

Diante da necessidade crítica de assegurar a privacidade dos dados em saúde, métodos tradicionais como a privacidade diferencial (PD) e a criptografia leve têm sido empregados como mecanismos de segurança de dados Coelho et al. (2019); Coelho (2020); Coelho et al. (2022); Surbhi and Dahiya (2024). A criptografia, por sua vez, converte os dados originais em representações cifradas, de modo que, mesmo

diante de acessos não autorizados, o conteúdo somente possa ser revelado mediante a posse da chave de decodificação apropriada. No entanto, tais mecanismos podem comprometer a utilidade dos dados e, frequentemente, demandam elevados recursos computacionais, o que dificulta sua implementação em dispositivos de IoHT.

Ao utilizar a privacidade diferencial, a proteção é alcançada por meio da adição de ruído estatístico aos dados ou às consultas realizadas sobre eles. Essa abordagem, contudo, tende a reduzir a precisão das análises e pode afetar adversamente a preservação de correlações relevantes entre atributos sensíveis, tais como idade, condição médica e tratamento Farooqi et al. (2024). Ademais, a PD enfrenta limitações práticas, uma vez que ruídos adicionados podem ser atenuados ou revertidos por meio de inferências baseadas em triangulação estatística dos resultados, possibilitando a reidentificação de valores reais dos registros Eichner et al. (2024). A aplicação de mecanismos de PD, como os baseados nas distribuições de Laplace ou Exponencial, pode ainda demandar processamento computacional intensivo, sobretudo em cenários com grandes volumes de dados ou em ambientes dinâmicos. Além disso, a definição adequada do parâmetro ϵ — responsável pelo equilíbrio entre privacidade e utilidade — constitui um desafio adicional, pois depende do contexto de aplicação He and Zhang (2023); Franzen et al. (2022).

Como alternativa, técnicas de anonimização têm sido amplamente estudadas por sua viabilidade prática e por permitirem preservar a privacidade de dados sem comprometer a capacidade analítica Onesimu et al. (2022). Essas técnicas transformam registros sensíveis em dados genéricos e indistinguíveis por meio de operações computacionalmente eficientes. Dentre elas, destaca-se o k-anonimato, cuja premissa consiste em garantir que cada registro de um conjunto de dados seja indistinguível de, pelo menos, outros $k-1$ registros, com base em atributos denominados quase identificadores (QIs), como idade, gênero e código postal. A combinação desses QIs únicos com outras informações pode, potencialmente, permitir a reidentificação de indivíduos Karagiannis et al. (2024).

A aplicação do k-anonimato por meio de anonimização antecipada, utilizando-se de técnicas como generalização e microagregação, visa preservar o máximo de informação possível, ao mesmo tempo em que reduz a granularidade dos dados a níveis compatíveis com os requisitos de privacidade. A generalização consiste na substituição de valores específicos de QIs por categorias ou intervalos mais amplos (por exemplo, transformar uma idade exata no intervalo “30-34”), de modo a reduzir o risco de identificação sem comprometer a utilidade analítica dos dados Yan et al. (2021). Já a microagregação envolve o agrupamento de registros semelhantes e a substituição de seus valores por representações agregadas, tais como as modas, medianas ou médias dos grupos (*clusters*), com o objetivo de satisfazer os critérios de anonimato preservando os detalhes dos dados originais Ribeiro-Alves et al. (2022).

A anonimização de dados é um elemento central no campo da segurança e privacidade, tendo como objetivo proteger informações sensíveis enquanto mantém a aplicabilidade dos dados Hundepool et al. (2012); Kara et al. (2025). Ao garantir que os conjuntos de dados atendam aos requisitos mínimos de proteção, como o k-anonimato, torna-se possível mitigar significativamente os riscos de violações de privacidade, sem prejudicar a acurácia de modelos preditivos ou classificadores empregados em diagnósticos e decisões clínicas Domingo-Ferrer et al. (2022).

Apesar dos avanços, persistem desafios importantes, especialmente no tocante ao equilíbrio entre privacidade e utilidade. A definição do valor de k e a escolha adequada das estratégias de generalização influenciam diretamente tanto a eficácia da proteção quanto a qualidade dos dados anonimizados. Essa tensão inerente, comumente referida como *trade-off* entre privacidade e utilidade, evidencia a necessidade de aprofundamento teórico e metodológico no campo da anonimização de dados. Vale destacar, ainda, que alcançar a configuração ótima para o k-anonimato é um problema de complexidade NP-difícil, o que reforça os desafios computacionais envolvidos Yan et al. (2021).

Com o intuito de avançar o estado da arte na área, esta tese propõe um *framework* denominado *Generalization First k-member Clustering* (GFKMC), cujo objetivo é assegurar a privacidade de dados no contexto da IoHT. O núcleo da proposta reside na promoção de uma anonimização antecipada e eficiente, capaz de satisfazer os princípios fundamentais do k-anonimato. O GFKMC busca mitigar os desafios da proteção de dados em ambientes que exigem alta conformidade regulatória — como o setor de saúde — sem comprometer a qualidade diagnóstica e analítica de soluções baseadas em aprendizado de máquina, sejam elas centralizadas ou federadas.

1.1 Problema e sua Importância

A digitalização crescente do setor de saúde e a ampla adoção de dispositivos IoHT têm impulsionado a produção massiva de dados clínicos e pessoais altamente sensíveis. Nesse contexto, garantir a privacidade das informações tornou-se um requisito crítico, diante do risco de inferência ou reidentificação de indivíduos a partir de seus dados. A anonimização de dados, especialmente a baseada no conceito de k-anonimato, constitui uma abordagem promissora para mitigar tais riscos. Ao assegurar que cada registro seja indistinguível de pelo menos outros $k-1$ registros em relação a seus quase identificadores, o k-anonimato dificulta substancialmente a identificação de indivíduos, mesmo quando informações externas adicionais estão disponíveis. Assim, um *framework* fundamentado em k-anonimato reduz os riscos de vazamento de informações, o que é particularmente relevante para organizações de saúde que operam com grandes volumes de informações sensíveis.

A questão central desta pesquisa pode, portanto, ser expressa da seguinte forma: **como garantir a privacidade das informações, reduzindo a exposição de indivíduos, sem comprometer significativamente a utilidade e a qualidade dos dados?** Esse desafio decorre do fato de que, embora técnicas de anonimização protejam identidades, elas frequentemente introduzem perdas de informação que afetam a aplicabilidade dos dados em análises avançadas, como no treinamento de modelos de aprendizado de máquina. Torna-se, então, necessário investigar soluções que conciliem esses dois objetivos aparentemente conflitantes: reduzir os riscos de reidentificação e, ao mesmo tempo, preservar a consistência, a representatividade e o valor analítico dos dados em cenários práticos da IoHT.

Entre os principais desafios destaca-se o equilíbrio entre privacidade e utilidade. Técnicas de anonimização antecipada buscam minimizar a distorção decorrente da generalização ou microagregação, garantindo o k-anonimato sem comprometer excessivamente a qualidade das análises estatísticas e clínicas. A identificação desse ponto ótimo é essencial para assegurar tanto a proteção dos dados quanto sua utilidade. Esse equilíbrio torna-se ainda mais relevante em aplicações baseadas em aprendizado de máquina (*Machine Learning* – ML) e aprendizado federado (*Federated Learning* – FL), em que a qualidade e a integridade dos dados de treinamento impactam diretamente o desempenho dos modelos. Generalizações excessivas ou microagregações mal otimizadas podem degradar padrões estatísticos relevantes, reduzindo significativamente a acurácia. Por outro lado, a anonimização orientada por critérios de preservação da estrutura dos dados permite que modelos de ML e FL sejam treinados, mesmo em cenários de restrições rigorosas de privacidade.

Portanto, a anonimização baseada em k-anonimato configura-se como uma estratégia essencial para assegurar a privacidade de dados sensíveis na IoHT. Além de reforçar a segurança da informação, ela viabiliza análises preditivas e avançadas que contribuem para a inovação e o progresso da medicina moderna.

1.2 Contribuições

A anonimização antecipada busca minimizar perdas de informação e preservar a consistência estatística dos dados, mantendo sua relevância para aplicações clínicas e computacionais. Isso é particularmente importante para situações de análises históricas e treinamento de modelos de aprendizado de máquina. A preservação da qualidade dos dados possibilita que as análises mantenham níveis aceitáveis de precisão, evitando prejuízos significativos nos resultados e predições obtidas.

O objetivo central desta tese consiste em garantir a privacidade de dados sensíveis no contexto da IoHT, sem comprometer de forma significativa sua utilidade e qualidade para soluções baseadas em aprendizado de máquina. Para alcançar esse

objetivo, propõe-se o desenvolvimento do *framework Generalization First k-Member Clustering*. Esse framework foi concebido para lidar de forma integrada com diferentes tipos de quase identificadores (numéricos e categóricos), assegurando conformidade com os princípios fundamentais do *k*-anonimato. As principais contribuições as quais compõem esta pesquisa podem ser sintetizadas da seguinte forma:

- Adotar a Anonimização Dinâmica por Separatriz Coelho et al. (2024b), para generalizar informações dos quase identificadores numéricos, que possam vir a identificar um indivíduo (idade, CEP, renda, códigos regionais).
- Utilizar a generalização hierárquica baseada em árvores de taxonomia para generalizar QIs categóricos (cidade, sexo, hospital). Neste ponto, a informação de um QI categórico é generalizada para a representação contida em seu respectivo valor pai, reduzindo o custo computacional da geração de grupos.
- Proporcionar flexibilidade na generalização dos grupos com base na substituição pelo centróide do grupo, o registro mais comum ou o valor mais comum.
- Implementar um procedimento de reagrupamento para atender aos requisitos do *k*-anonimato. Inicialmente guloso, agrupa registros que compartilham valores semelhantes de quase identificadores. Se um grupo possuir, no mínimo *k* registros, uma classe de equivalência é formada. Caso contrário, as outras classes de equivalência são formadas aplicando operações de reagrupamento nos grupos restantes em vez de em registros individuais, reduzindo a sobrecarga computacional do procedimento de reagrupamento no *k*-anonimato.
- Minimização da perda de informações causada pela generalização de classes de equivalência com um algoritmo de agrupamento de *k*-membros eficiente.
- Integrar as soluções supracitadas em um *framework* completo e versátil que proporciona *k*-anonimização de atributos numéricos e categóricos baseada em generalização antecipada e agrupamento de *k*-membros para dados de saúde Coelho et al. (2024c) (*Submitted*).
- Proposição de uma metodologia inovadora para avaliar os efeitos da anonimização sobre o desempenho de modelos de aprendizado de máquina, considerando métricas relevantes de privacidade, como perda de informação, divulgação de atributos e desempenho Coelho et al. (2025a,b).
- Demonstração empírica de que o GFKMC alcança uma perda de informação constante e controlada, garantindo simultaneamente altos níveis de privacidade e manutenção da acurácia em aplicações ML.

A anonimização de dados antecipada com *k*-anonimato oferece uma abordagem eficaz para proteger a privacidade de dados, com o benefício adicional de causar menor interferência na qualidade dos dados. Essa abordagem equilibra as necessidades de segurança e utilidade, sendo uma solução relevante em ambientes que demandam alta proteção de dados e conformidade regulatória, como na área da saúde, sem sacrificar a capacidade de análise e predições. Embora a generalização antecipada obedecendo aos princípios do *k*-anonimato traga benefícios significativos, existem desafios a serem considerados. A escolha adequada dos métodos de generalização e do valor de *k* é crucial, pois um valor muito baixo pode não fornecer privacidade suficiente, enquanto um valor muito alto pode comprometer excessivamente a utilidade dos dados. Além disso, a estrutura proposta pode ser combinada com outros modelos de privacidade para aumentar a resistência contra diferentes tipos de ataques de reidentificação. A aplicação cuidadosa e combinada com outras técnicas de anonimização pode ajudar a mitigar limitações, garantindo a privacidade dos indivíduos sem comprometer a utilidade das soluções baseadas em aprendizado de máquina, sejam elas tradicionais ou federadas.

1.3 Estrutura da Tese

Esta tese está estruturada em conformidade com as Normas Gerais de Teses e Dissertações do Conselho Técnico de Pós-Graduação da Universidade Federal de Viçosa, UFV (2022). O formato desta tese é "Texto Corrido", composto por: Introdução Geral no Capítulo 1, descrevendo o problema e sua respectiva importância na área. O Capítulo 2 apresenta de maneira clara e sucinta os Fundamentos imprescindíveis para a compreensão sobre a anonimização de dados obtidos por dispositivos pertencentes à Internet das Coisas aplicada na área da saúde. Posteriormente, no Capítulo 3 são apresentados os Trabalhos Relacionados à anonimização de dados. No Capítulo 4 o *framework* GFKMC para *k*-anonimização baseada em generalização antecipada e agrupamento de *k*-membros para dados de saúde é detalhado por completo. O Capítulo 5 descreve a metodologia utilizada para avaliar os efeitos da anonimização sobre o desempenho de modelos de aprendizado de máquina, considerando métricas de privacidade, perda de informação e desempenho. O Capítulo 6 discute os resultados obtidos empiricamente. Finalmente, no Capítulo 7, são apresentadas as conclusões, além de relacionar as oportunidades de pesquisa em aberto. Ademais, o Apêndice A lista as demais publicações científicas produzidas durante o período de vigência do curso de pós-graduação.

Capítulo 2

Fundamentos

2.1 Internet das Coisas Aplicada na Saúde

A Internet das Coisas representa um paradigma tecnológico baseado na interconexão de dispositivos heterogêneos, capazes de coletar, processar e transmitir dados de forma autônoma Din et al. (2018). Quando aplicada ao setor de saúde, dá origem à IoHT, um ecossistema caracterizado pela integração de sensores implantáveis, dispositivos vestíveis, sistemas de monitoramento e plataformas de análise Ketu and Mishra (2021). Todos esses têm o objetivo de aprimorar a prestação de serviços médicos, promover a saúde preventiva e otimizar diagnósticos e tratamentos.

O ambiente IoHT permite o monitoramento contínuo de parâmetros fisiológicos dos pacientes, tais como frequência cardíaca, pressão arterial, nível de glicose, sinais eletrocardiográficos (ECG) e fotopletismográficos (PPG), entre diversos outros Coelho et al. (2023b). Estas informações passam a compor o Prontuário Eletrônico do Paciente, uma versão digital do prontuário médico tradicional, que substitui o documento em papel por um registro eletrônico. Além de incluir informações sobre o histórico de saúde, essa ferramenta centraliza dados pessoais privados, como nome, endereço e documentos de identificação, entre outras informações relevantes. Esse conjunto de dados é fundamental para a personalização do atendimento médico, possibilitando a detecção precoce de condições clínicas e a intervenção em tempo hábil Paul (2025). Além disso, a adoção desse modelo contribui para a descentralização dos serviços de saúde, promovendo o acompanhamento remoto de pacientes, especialmente aqueles com doenças crônicas ou em situação de fragilidade Kolawole (2024).

No entanto, a geração massiva de dados sensíveis em ambientes IoHT impõe desafios significativos, principalmente no que tange à privacidade, à segurança da informação em conformidade com legislações de proteção de dados, como a LGPD, o GDPR e o HIPAA Shahid et al. (2022). Tais desafios exigem o desenvolvimento e a aplicação de mecanismos robustos de proteção, entre os quais se destaca a anonimização de dados.

2.2 Aprendizado de Máquina

O Aprendizado de Máquina é uma área da Inteligência Artificial que se concentra no desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e realizar previsões ou classificações sem a necessidade de programação explícita para cada tarefa Sarker (2021). O ML tem sido amplamente aplicado na área da saúde, permitindo desde a previsão de diagnósticos até a detecção de anomalias e o desenvolvimento de sistemas inteligentes de apoio à decisão clínica Asif et al. (2025).

Entre os paradigmas de ML, destaca-se o Aprendizado Supervisionado, onde os modelos são treinados a partir de dados rotulados para realizar previsões ou classificações. Já no Aprendizado Não Supervisionado, o modelo busca padrões ocultos em dados não rotulados Almuqati et al. (2024). Entre os modelos supervisionados, destacam-se algoritmos como *Support Vector Machines* (SVM), *Random Forests* (RF), *k-Nearest Neighbour* (k-NN), *logistic regression* (LR) e *Extreme Gradient Boosting* (XGBoost), amplamente utilizados em aplicações clínicas e biomédicas Binson et al. (2024).

Os algoritmos de aprendizado supervisionado supracitados serão utilizados como suporte para avaliar a respectiva degradação do desempenho. As florestas aleatórias constroem múltiplas árvores de decisão onde cada árvore é treinada com subconjuntos aleatórios de dados e atributos. A previsão final é feita por votação (classificação) ou média (regressão) entre as árvores. É robusto contra *overfitting* e muito aplicado na predição de risco clínico, diagnósticos e análises genômicas Binson et al. (2024). O XGBoost também é um método baseado em árvores de decisão, em que árvores são treinadas sequencialmente, e cada nova árvore corrige os erros da anterior. É conhecido por sua eficiência e precisão em competições de ciência de dados. Na saúde, é utilizado em tarefas como detecção precoce de doenças, análise preditiva e classificação de imagens médicas Routray and Choudhary (2025).

O SVM busca um hiperplano ótimo que separa as classes nos dados, maximizando a margem entre os pontos de dados de diferentes classes. É eficaz para tarefas de classificação e regressão, mesmo com conjuntos de dados complexos. É utilizado, por exemplo, para classificação de doenças a partir de exames laboratoriais ou sinais vitais. O k-NN é um método baseado em instâncias, que classifica um novo dado com base nos k vizinhos mais próximos em termos de distância (geralmente euclidiana). Simples e interpretável, é eficaz quando os dados estão bem distribuídos. É utilizado, por exemplo, em sistemas de apoio à decisão médica. A regressão logística é um modelo probabilístico que estima a probabilidade de uma classe binária com base em uma função logística. Ela é amplamente utilizada em epidemiologia, para avaliar fatores de risco e predição de desfechos clínicos tais como probabilidade de óbito,

infarto, entre outros Binson et al. (2024).

Recentemente, o Aprendizado Federado emergiu como uma abordagem inovadora, especialmente relevante para ambientes sensíveis à privacidade, como o IoHT Abbas et al. (2024). No FL, os modelos são treinados de forma descentralizada, permitindo que os dados permaneçam localmente nos dispositivos dos usuários. Apenas os parâmetros dos modelos, como gradientes ou pesos, são compartilhados com o servidor central, mitigando riscos de vazamento de dados sensíveis. Apesar das vantagens, o FL impõe novos desafios, incluindo a heterogeneidade dos dados distribuídos e o balanceamento entre desempenho e privacidade Beltrán et al. (2023).

2.3 Anonimização de dados

A anonimização de dados consiste em um conjunto de técnicas e procedimentos aplicados para remover ou transformar informações de forma que os indivíduos representados nos dados não possam ser identificados direta ou indiretamente Shamsinejad et al. (2025). Trata-se de uma abordagem fundamental para garantir a privacidade, especialmente em cenários que envolvem o compartilhamento e a análise de grandes volumes de dados sensíveis, como no contexto IoHT Zuo et al. (2021).

Diferentemente da criptografia, que protege os dados durante o armazenamento ou a transmissão, a anonimização visa garantir que, mesmo após o compartilhamento dos dados, não seja possível realizar a reidentificação dos indivíduos Shamsinejad et al. (2025). As técnicas principais de anonimização podem ser classificadas como generalização, remoção, supressão, perturbação através da adição de ruído e microagregação. A seção seguinte expande a descrição de cada uma delas.

As técnicas de anonimização são frequentemente avaliadas com base em dois critérios fundamentais, a privacidade e utilidade dos dados. O grau de privacidade garantido geralmente é mensurado por modelos formais como k-anonimato e l-diversidade, por exemplo. Já a preservação da utilidade dos dados é medida pela perda de informação e pela manutenção da acurácia em análises subsequentes.

2.4 Anonimização de Dados Baseada em k-anonimato

O modelo de k-anonimato é um dos pilares da anonimização de dados estruturados, amplamente utilizado para proteger a privacidade de indivíduos Zuo et al. (2021); Monteiro et al. (2022). A ideia central é garantir que, para qualquer registro de um conjunto de dados, existam pelo menos outros $k-1$ registros com a mesma

combinação de valores nos chamados quase identificadores (QIs). Assim, a probabilidade de reidentificação de um indivíduo com base nesses QIs é limitada a no máximo $1/k$ Domingo-Ferrer et al. (2016). Por exemplo, imagine uma base de dados com os atributos: Idade, Gênero, CEP, Condição Médica. Os três primeiros atributos podem ser considerados QIs. Caso uma combinação única como (34 anos, feminino, 36570-000) apareça apenas uma vez, ela representa um risco de reidentificação iminente. Aplicando o k-anonimato com $k = 3$, é necessário generalizar ou agrupar registros de modo que existam pelo menos três registros com a mesma combinação de QIs, tais como (30–40 anos, feminino, 3657*-*) Torra (2022).

A anonimização de dados objetiva proteger informações sensíveis enquanto mantém a utilidade para análises estatísticas Olatunji et al. (2024). Em seguida serão introduzidos os principais termos, conceitos gerais e aspectos técnicos necessários para a compreensão da anonimização de dados baseada em k-anonimato. Ademais, os conceitos básicos serão apresentados de maneira clara e sucinta, em formato de glossário. Deste modo, é fornecida uma visão holística da área, o que habilita discussões aprofundadas sobre a efetiva aplicação das técnicas de anonimização Domingo-Ferrer et al. (2016); Torra (2022); Fung et al. (2010); Personal Data Protection Commission (2018).

Framework

Um *framework*, em computação, é uma coleção de componentes reutilizáveis que fornecem uma base para o desenvolvimento de aplicações. São compostos por estruturas que fornecem uma variedade flexível de componentes de software que ajudam os desenvolvedores a acelerar o desenvolvimento de uma solução até a implantação em produção. No contexto de anonimização, o *framework* é um sistema composto por etapas e módulos reutilizáveis que operam em sequência para transformar os dados originais em um conjunto k-anonimizado. O GFKMC proposto nesta tese é um exemplo de *framework*, o qual atua com módulos de generalização antecipada, clusterização e reagrupamento.

Método

No contexto da Engenharia de Software, os métodos fornecem instruções práticas e técnicas para a construção de software. Eles compreendem várias tarefas, entre elas análise de requisitos, projeto e construção de programas. A definição do uso de generalização hierárquica via árvores de taxonomia para QIs categóricos ou a microagregação por separatrizes para dados numéricos são exemplos de métodos adotados no *framework* GFKMC.

Trade-off

Termo da língua inglesa que define uma situação em que há conflito de escolha. Ele é relacionado a uma ação que visa à resolução de um problema, mas provoca outro, obrigando a uma escolha. *Trade-off* pode ser traduzido como "compromisso". No contexto da anonimização, refere-se ao equilíbrio entre privacidade (quanto maior o k, mais anonimato) e utilidade (quanto menor a perda de informação). Um exemplo clássico: aumentar k de 3 para 10 aumenta a proteção, mas pode exigir generalizações tão amplas que o dado se torna pouco útil para análises.

Cluster

A tradução literal do inglês significa "grupo", "agrupamento" ou "aglomerado". Clusterização (*Clusterization*) é uma técnica utilizada em computação que agrupa vários recursos computacionais, no caso específico desta tese, registros e dados. São grupos de registros com características similares em seus QIs. Por exemplo, ao agrupar registros com idade entre 20 e 30 anos e CEPs da mesma região, pode-se formar um grupo que será generalizado como (20-30 anos, CEP 3657*-*).

Anonimização de Dados

A anonimização é uma técnica de processamento de dados que remove ou modifica informações que possam identificar uma pessoa. Essa técnica resulta em dados anonimizados, que não podem ser associados a nenhum indivíduo específico. Os algoritmos de anonimização são a implementação técnica que realiza operações sobre os dados (generalização, supressão, etc.) para garantir o k-anonimato. Um exemplo é o algoritmo Mondrian, que divide recursivamente os dados em partições até formar grupos com k ou mais elementos.

Modelos de privacidade

Um modelo de privacidade de dados é uma estrutura teórica ou um conjunto de regras que define como os dados devem ser protegidos para evitar a reidentificação de indivíduos. Tais modelos estabelecem critérios e propriedades que devem ser atendidos para que os dados possam ser considerados suficientemente anônimos. Dentre os modelos mais adotados na literatura, destaca-se o k-anonimato.

- **k-anonimato** O k-anonimato é um modelo de privacidade que visa impedir a reidentificação de registros com base em um subconjunto de atributos denominados quase identificadores, os quais, quando combinados, podem revelar a identidade de indivíduos. Um conjunto de dados está k-anonimizado

se, e somente se, cada combinação de valores dos QIs aparecer em pelo menos k registros distintos, ou seja, cada indivíduo é indistinguível de pelo menos outros $k-1$ indivíduos quanto a esses atributos.

O k-anonimato define uma condição a ser satisfeita, mas não estabelece os procedimentos específicos para alcançá-la. Portanto, é necessário aplicar operações de anonimização, tais como generalização, supressão, microagregação ou adição de ruído, a fim de transformar os dados de modo que o critério de k-anonimato seja atendido.

Exemplo: Considere a Tabela 2.1 com os atributos *Idade*, *CEP* e *Diagnóstico*. Os dois primeiros são considerados quase identificadores:

Tabela 2.1: Exemplo de dados brutos a serem anonimizados

Idade	**CEP**	**Diagnóstico**
29	36571-000	Diabetes
30	36579-000	Hipertensão
29	36570-000	Asma
40	36582-000	hepatite
42	36585-000	Câncer
43	36587-000	Câncer

Neste exemplo, a combinação (Idade, CEP) ocorre apenas uma vez para alguns registros, o que implica alto risco de reidentificação. Para garantir o k-anonimato com $k = 3$, é possível aplicar generalização, conforme ilustrado na Tabela 2.2

Tabela 2.2: Exemplo de dados anonimizados

Idade	**CEP**	**Diagnóstico**
20-30	3657*-*	Diabetes
20-30	3657*-*	Hipertensão
20-30	3657*-*	Asma
40-50	3658*-*	hepatite
40-50	3658*-*	Câncer
40-50	3658*-*	Câncer

Deste modo, os três primeiros registros formam uma classe de equivalência com os mesmos valores generalizados para os QIs (Idade e CEP), e o mesmo ocorre para os três últimos. O exemplo ilustra como a generalização pode ser usada para alcançar o k-anonimato. O exemplo também ilustra ainda o *trade-off*, demonstrando a potencial perda de granularidade e informação nos dados.

Algoritmos de privacidade

Os algoritmos para anonimização de dados implementam os modelos de privacidade, como k-anonimato, em um conjunto de dados. Esses algoritmos aplicam

operações como generalização, supressão ou particionamento para transformar os dados de modo que atendam ao critério de privacidade definido pelo modelo. O objetivo dos algoritmos é processar os dados brutos e aplicar transformações específicas para alcançar o modelo de privacidade desejado. Eles garantem que o conjunto de dados resultante cumpra com os requisitos de *k*-anonimato.

Classe de equivalência

Uma classe de equivalência é um subconjunto de registros com valores de quase identificadores idênticos. Ou seja, um subgrupo de indivíduos do conjunto original, os quais compartilham os mesmos valores para todos os atributos. Por exemplo, todos os registros generalizados para (30-40 anos, masculino, 3657*-*) pertencem à mesma classe de equivalência.

Caracterização dos atributos

A efetividade das técnicas de anonimização depende diretamente da correta classificação dos atributos. Os atributos também são chamados de campo de dados, coluna de dados ou variável. Uma informação que pode ser encontrada nos registros do conjunto de dados. Nome, gênero e endereço são exemplos de atributos.

- **Identificadores** Os atributos identificadores são aqueles que, mesmo sozinhos, podem ser usados para reidentificar um indivíduo diretamente. Para evitar tal reidentificação, esses atributos devem ser excluídos ou mascarados. Alguns exemplos de atributos identificadores são CPF e nome completo.
- **Quase identificadores (QIs)** Os atributos quase identificadores são aqueles que, sozinhos, não permitem a reidentificação direta de um indivíduo. No entanto, quando combinados (entre si ou com informações externas), podem revelar sua identidade e por isso devem ser anonimizados.
- **Sensíveis ou confidenciais** Os atributos sensíveis ou confidenciais são aqueles que possuem informações consideradas sensíveis sobre os indivíduos e não os reidentificam. O objetivo das técnicas de anonimização é evitar que seja possível relacionar o valor de um desses atributos a um indivíduo. Exemplos de atributos confidenciais são condições de saúde, salário, notas de avaliações e outros.

Tipos dos atributos

- **Numéricos** Atributos numéricos são aqueles que podem ser medidos em uma escala quantitativa. Ou seja, aplicar cálculos estatísticos numéricos sobre os valores faz sentido. Exemplos são peso, número de filhos e altura.

- **Categóricos** Atributos categóricos são aqueles que não possuem valores quantitativos, mas que classificam os indivíduos em diversas categorias. São exemplos: sexo, escolaridade e raça.

Operações de anonimização

Visando satisfazer os requisitos de privacidade especificados, faz-se necessário aplicar algumas operações de anonimização sobre o conjunto de dados inicial. A seguir, algumas das principais operações utilizadas em anonimização de dados são descritas.

- **Remoção e Supressão** A remoção é a operação mais simples e consiste em remover do conjunto de dados um atributo em sua totalidade ou apenas alguns registros específicos. Nota-se aqui que toda a informação referente aos registros ou atributos excluídos é perdida. Essa operação pode ser utilizada para, por exemplo, remover atributos identificadores. A remoção está intimamente relacionada com a supressão, pois a primeira é uma das formas de aplicação da segunda. No entanto, na supressão, ao invés de simplesmente remover os atributos, pode-se também, por exemplo, substituir os valores por um valor especial que indica que os valores substituídos não são divulgados.

- **Generalização** A generalização consiste em alterar os valores referentes a algum atributo de forma que os novos valores tenham menor probabilidade de serem únicos, ou seja, reduz-se deliberadamente a granularidade e precisão dos dados. Pode-se, por exemplo, transformar a idade de um indivíduo em uma faixa etária ou o peso em um intervalo.

- **Adição de Ruído** Esta operação consiste em adicionar um ruído aleatório aos dados originais. Assim, ela costuma ser aplicada apenas a dados numéricos contínuos. Existem diversas formas diferentes de realizar essa adição (PD é uma delas). No entanto, todas elas buscam preservar da melhor maneira possível as propriedades estatísticas dos dados originais.

- **Microagregação** A microagregação é comumente utilizada para satisfazer os requisitos do *k*-anonimato. Essa operação consiste em substituir os valores dos atributos de cada registro por métricas computadas em cima de pequenos grupos, como a média, por exemplo. Esses grupos são formados para maximizar a similaridade entre seus integrantes e, consequentemente, minimizar a perda de informação.

Modelo de dados relacionais

O modelo de *k*-anonimato é geralmente usado para garantir a preservação de privacidade de dados relacionais, onde as relações são salvas no formato de tabelas. Uma tabela geralmente usa linhas para registros e colunas para atributos. Um registro na tabela de dados relacionais corresponde a uma entidade específica (indivíduo na área da saúde) e descreve uma quantidade significativa de informações sobre a entidade por meio de diferentes atributos.

Métricas de informação

As métricas de informação avaliam quantitativamente a qualidade dos dados anonimizados em relação aos dados originais. Elas são responsáveis por guiar os algoritmos de anonimização, além de auxiliarem os desenvolvedores a avaliar o compromisso entre perda de informação e risco de reidentificação.

As métricas de informação podem ser separadas em três categorias de acordo com seu propósito: propósito geral, propósito específico e propósito de *trade-off*. Abaixo há a definição de cada uma e alguns exemplos.

- **Propósito geral** Essas métricas são utilizadas quando não se sabe ao certo como os dados disponibilizados serão analisados. Nesse caso, é interessante medir a semelhança entre os dados anonimizados e os dados originais.

  Uma das métricas de propósito geral mais simples é a *Minimal distortion* (MD). Nessa métrica, cobra-se uma penalidade toda vez que uma instância é generalizada ou suprimida. O cálculo ocorre em cima de cada atributo individualmente. Por exemplo, se a anonimização generalizou "engenheiro" para "profissional" em 10 instâncias, então tem-se 10 unidades de distorção.

  Além da MD, outras métricas de propósito geral são a IL (*Information Loss*), que também captura a perda de informação pela generalização, e a *Discernibility Metric*, que cobra uma penalidade para cada registro que for indistinguível de outros registros quando comparados os quase identificadores.

- **Propósito específico** As métricas de propósito específico são principalmente utilizadas para analisar dados cuja utilização possui um propósito claro. Esse é o caso de conjuntos de dados utilizados para treinamento de modelos de aprendizado de máquina supervisionados por exemplos. Para esses modelos, a generalização dos atributos deve ser feita com muita cautela. Enquanto ela pode ser extremamente prejudicial se feita em excesso, feita em pequena escala, ela até pode ajudar o modelo. Um exemplo é a *"Normalized Certainty Penalty"*, que mede a perda de informações penalizando valores de atributos que são transformados em valores mais generalizados após a anonimização.

- **Propósito de *trade-off*** Durante o processo de anonimização, é fundamental reduzir o risco de reidentificação, no entanto, essa operação está diretamente relacionada com a perda de informação que os dados carregam. As métricas com propósito de *trade-off* buscam uma forma de quantificar essa relação. Assim, um algoritmo de anonimização as usa como um guia.

2.5 Diretrizes para o desenvolvimento do GFKMC

A metodologia para desenvolver o *framework* GFKMC, que atenda aos requisitos de *k*-anonimato com base no agrupamento de *k*-membros, é inspirada no algoritmo de agrupamento ponderado de *k*-membros apresentado em Yan et al. (2021). Portanto, para entender e garantir a reprodutibilidade fiel do *framework* GFKMC proposto, as seguintes definições, baseadas na proposta de Yan et al. (2021), são fornecidas.

Definição 1 ***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:*

$$d_N(v_1, v_2) = \frac{|v_1 - v_2|}{max(N) - min(N)} \tag{2.1}$$

onde $max(N)$ e $min(N)$ se referem ao valor máximo e mínimo do atributo numérico N

Definição 2 ***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:*

$$d_C(v_1, v_2) = \begin{cases} 0, & v_1 = v_2 \\ \frac{|LCA(v_1,v_2)|}{|Tree\_C|}, & v_1 \neq v_2 \end{cases} \tag{2.2}$$

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

Definição 3 ***Distância entre dois registros*** *Para uma tabela de dados T com atributos numéricos $N_i (i = 1, \ldots, m)$ e atributos categóricos $C_j (j = 1, \ldots, n)$, a distância entre dois registros $r_1, r_2 \in T$ é definida como:*

$$d(r_1, r_2) = \sum_{i=1}^{m} d_N(r_1(N_i), r_2(N_i)) + \sum_{j=1}^{n} d_C(r_1(C_j), r_2(C_j)) \tag{2.3}$$

onde $d_N(\ldots)$ e $d_C(\ldots)$ são as funções de distância definidas anteriormente em 2.1 e 2.2.

Definição 4 ***Perda de informação (IL)*** *Seja e uma classe de equivalência na qual todos os registros têm quase identificadores numéricos $N_i(i = 1, \ldots, m)$ e quase identificadores categóricos $C_j(j = 1, \ldots, n)$. $Tree\_C_j$ é a árvore de taxonomia correspondente ao atributo categórico C_j. A quantidade de perda de informação causada pela generalização é definida como:*

$$IL(e) = |e| \left(\sum_{i=1}^{m} \frac{max(N_i) - min(N_i)}{|N_i|} + \sum_{j=1}^{n} \frac{H(Tree\_LCA(C_j))}{H(Tree\_C_j)} \right) \tag{2.4}$$

onde $|e|$ é o número de registros em e, $max(N)$ e $min(N)$ referem-se aos valores máximo e mínimo de N_i em e, $Tree\_LCA(C_j)$ é a árvore com raiz no antecessor comum mais baixo do conjunto de valores em e de C_j, e $H(\ldots)$ representa a altura da árvore.

Definição 5 ***Pontuação Numérica (Numerical Score)*** *A pontuação numérica é uma medida para detecção de outliers dos atributos numéricos de um registro. Sejam $N_i(i = 1, \ldots, m)$ os atributos numéricos da tabela de dados T. Então, a pontuação numérica de um registro $r_j \in T$ é definida como:*

$$N_{score}(r_j) = \sum_{i=1}^{m} ||r_j(N_i) - ave(N_i)||_1 \tag{2.5}$$

onde $||r_j(N_i) - ave(N_i)||_1$ é a forma normal L_1 de $r_j(N_i) - ave(N_i)$, e $ave(N_i)$ é o valor médio do atributo numérico N_i.

Definição 6 ***Pontuação categórica (Categorical Score)*** *A pontuação categórica é uma medida da detecção de outliers dos atributos categóricos de um registro. Sejam C_i $(i = 1, \ldots, n)$ os atributos categóricos da tabela de dados T. Então, a pontuação categórica de um registro $r_j \in T$ é definida como:*

$$C_{score}(r_j) = \frac{1}{n} \sum_{i=1}^{n} f(r_j(C_i)) \tag{2.6}$$

onde $f(r_j(C_i))$ denota a frequência do valor de C_i em r_j.

Definição 7 ***Pontuação média de classificação (Average ranking score – AR)*** *A pontuação média de classificação é definida para uma tabela de dados com atributos numéricos e categóricos. Para um registro $r_j \in T$, a pontuação AR é definida como:*

$$AR_{score}(r_j) = \frac{rank(N_{score}(r_j)) + rank(C_{score}(r_j))}{2} \tag{2.7}$$

onde $rank(\ldots)$ denota a função de classificação Li (2011). Nesta tese, a função de classificação usa uma ordem decrescente para uma pontuação numérica e uma ordem crescente para uma pontuação categórica.

Definição 8 ***Pontuação de Peso (Weight Score)*** *Para a tabela de dados T, a pontuação de peso de um registro $r_j \in T$ é definida como:*

$$W(r_j) = |T| - AR_{score}(r_j) \tag{2.8}$$

onde $|T|$ é o número total de registros na tabela T.

Definição 9 ***Distância de Peso (Weight Distance)*** *Com a pontuação de peso, a distância ponderada pode ser calculada. A distância de peso de dois registros $r_1, r_2 \in T$ é definida como:*

$$W_d(r_1, r_2) = \sqrt{W(r_1)^2 + W(r_2)^2} \cdot d(r_1, r_2) \tag{2.9}$$

onde $W(r_1)$ e $W(r_2)$ são a pontuação de peso dos registros r_1 e r_2, e $d(r_1, r_2)$ é a distância entre os dois registros definidos em 2.3.

Definição 10 ***Perda de informação de peso (Weight Information Loss – WIL)*** *Seja $E = \{e1, \ldots, e_S\}$ que representa o conjunto de classes de equivalência geradas pelo algoritmo de agrupamento. Então, a quantidade de perda de informação de peso (WIL) para a tabela anonimizada T^* é definida como:*

$$WIL(T^*) = \sum_{e \in E} |W(e)|_2 \cdot IL(e) \tag{2.10}$$

onde $||W(e)||_2$ é a forma normal L_2 da pontuação de peso para a classe de equivalência e, e $IL(e)$ se refere à perda de informação de e definida em 2.4.

2.6 Resumo

Este capítulo apresentou os fundamentos teóricos essenciais para a compreensão dos desafios e soluções ligados à preservação da privacidade de dados na IoHT. Foram discutidos os principais conceitos da IoHT, destacando-se seu papel no monitoramento remoto de pacientes e na geração massiva de dados sensíveis. Em seguida, abordaram-se os princípios do aprendizado de máquina, aplicado à saúde, tanto em cenários centralizados quanto descentralizados, como o aprendizado federado. Também foram detalhados os conceitos de anonimização de dados, com ênfase no modelo de privacidade baseado no *k*-anonimato. Por fim, foi apresentada a fundamentação teórica que sustenta o desenvolvimento do *framework* GFKMC proposto. No próximo capítulo, são explorados os principais trabalhos relacionados ao tema, com foco tanto nas abordagens de anonimização baseadas em *k*-anonimato quanto nas metodologias utilizadas para avaliar seus impactos sobre modelos de aprendizado de máquina, fornecendo o embasamento para a compreensão do estado da arte na área.

Capítulo 3

Trabalhos Relacionados

A proteção da privacidade de dados privados, especialmente no contexto da IoHT, impõe desafios significativos à proteção da privacidade. A anonimização, particularmente baseada no modelo de *k*-anonimato, é uma estratégia amplamente adotada para mitigar riscos de reidentificação. Contudo, este modelo enfrenta limitações quanto ao equilíbrio entre privacidade, utilidade dos dados e escalabilidade. A literatura tem proposto diversos algoritmos para aprimorar a eficácia da anonimização, incorporando principalmente técnicas de agrupamento e generalização. Além disso, observa-se uma crescente preocupação com os impactos da anonimização no desempenho de modelos de aprendizado de máquina, tanto centralizados quanto federados. Este capítulo organiza-se em duas partes, a primeira aborda as principais técnicas de anonimização baseadas em *k*-anonimato e a segunda discute metodologias para avaliar seus efeitos sobre modelos de aprendizado, considerando privacidade, utilidade e desempenho.

Anonimização de Dados Baseada em k-Anonimato

Dentre as abordagens clássicas para anonimização de dados, o modelo baseado em *k*-anonimato tem sido amplamente adotado. Diversas propostas visam aprimorar sua eficiência, escalabilidade e capacidade de preservação da utilidade dos dados. Em El Ouazzani and El Bakkali (2018), os autores propõem um algoritmo de *k*-anonimato que não requer o conhecimento prévio do valor de *k*. O método agrupa iterativamente registros com combinações idênticas de quase identificadores até que todos os elementos do grupo apresentem características homogêneas. Embora o trabalho almeje contemplar grandes volumes de dados, as demonstrações conceituais são desenvolvidas sobre bases reduzidas, compostas de atributos quase identificadores fictícios. Isto resultou em um $k = 3$. Apesar deste trabalho não aprofundar a discussão sobre os resultados, os autores acompanham a literatura inferindo que um valor de *k* baixo implica diretamente em menor privacidade.

O algoritmo *Adaptive k-Anonymity* (AKA) Arava and Lingamgunta (2020) é outro exemplo de solução voltada para ambientes computacionalmente restritos, como

serviços em nuvem na área da saúde. Como este tipo de serviço é cobrado conforme o uso, refazer a anonimização para obter agrupamentos ideais e aliados à perda mínima de informações é custoso. A proposta combina técnicas tradicionais de agrupamento, tais como *k-member*, *C-means*, *One-Pass k-mean* e *Efficient Systematic Clustering* para mitigar o custo computacional com agrupamentos. No entanto, encontrar uma boa definição de k inicial, escolhendo de forma aleatória, é um desafio. Portanto, os autores se baseiam no *Enhanced Clustering Method* para definir o valor k no k-anonimato.

Khan et al. (2020) propuseram o modelo θ–Sensitive, projetado para mitigar ataques de variância sensível e de similaridade categórica. A proposta calcula θ como o produto da variância (σ^2) de uma classe de equivalência diversa por um valor observado (μ). Além disso, se a privacidade desejada não for alcançada, acrescentam-se pequenas quantidades de ruído para aumentar a variabilidade em uma classe de equivalência. Embora eficaz, o método depende da calibração criteriosa dos parâmetros μ e do nível de ruído, que impactam diretamente a utilidade dos dados.

O trabalho de Onesimu et al. (2022) apresenta um esquema de publicação de dados com preservação de privacidade com foco em atributos numéricos e categóricos. Para atributos numéricos, a abordagem de intervalo fixo é adotada. Nessa abordagem, os valores originais dos dados de saúde são substituídos por um valor generalizado. Os atributos categóricos são protegidos por fatiamento l-diverso dos dados, horizontal e verticalmente, generalizando-os para evitar vazamentos de privacidade. Embora eficaz na mitigação de riscos, essa técnica apresenta aumento significativo na perda de informação, especialmente devido ao fatiamento. Ademais, esta abordagem exige que o valor de k seja fornecido previamente, o que pode interferir diretamente na garantia da privacidade.

No contexto específico de dados provenientes de dispositivos vestíveis, Liu and Li (2018) propõem uma técnica baseada em agrupamento, utilizando generalização e supressão para proteger quase identificadores. A abordagem garante que todos os registros de um grupo sejam suficientemente semelhantes, mas, como outras técnicas baseadas em k-anonimato, permanece vulnerável a ataques de divulgação de atributos.

Torra and Navarro-Arribas (2023) demonstraram que algoritmos tradicionais como MDAV Domingo-Ferrer and Mateo-Sanz (2002) e Mondrian LeFevre et al. (2006, 2005a) não são imunes à divulgação de atributos, reforçando as limitações do estado da arte. O MDAV Templ (2008) é um método heurístico de microagregação multivariada para espaços n-dimensionais (>1). Já o Mondrian LeFevre et al. (2006, 2005b) consiste em uma anonimização multidimensional gananciosa com um algoritmo de aproximação simples. É baseado em divisão recursiva de baixo para

cima (*bottom-up*) da base de dados em duas até que cada agrupamento contenha entre k e $2k - 1$ registros. Ele exige a pré-configuração do valor de k pelo usuário, o que os torna suscetíveis a ataques caso sejam mal configurados.

A proposta apresentada por Xu et al. (2006) introduz o conceito de recodificação local. Ela diferencia-se das abordagens de anonimizações globais ao permitir que a generalização ocorra de forma localizada em microrregiões do espaço de dados. A maioria das abordagens de anonimização é global. Desse modo, um valor de quase identificador específico é generalizado para um valor global. Na recodificação local (microagregação), o espaço de dados é particionado em muitas microrregiões, e o mapeamento do quase identificador para o valor generalizado é local para essa região. Deste modo melhora a utilidade preservada, ainda que com maior complexidade de implementação.

No mesmo sentido de garantir a privacidade dos indivíduos, especificamente em relação aos atributos categóricos, Khatir et al. (2023) propôs um algoritmo baseado em agrupamento ganancioso. Esse algoritmo começa agrupando os dados considerando tanto a similaridade dos quase identificadores quanto a diversidade dos atributos sensíveis. O algoritmo é projetado para operar sobre uma tabela de dados binários gerada a partir da tabela de dados original. Para atingir k-anonimato, a proposta substitui os valores quase identificadores pelo valor do centróide do grupo. Embora a abordagem seja eficaz para quase identificadores com poucos valores de domínio, como gênero (masculino, feminino), ela se torna cada vez mais complexa com um número maior de valores de domínio. Isso se deve à complexidade do cálculo da métrica de distância variacional, a qual mede a similaridade dos quase identificadores, que aumenta significativamente à medida que o número de valores de domínio quase identificadores se expande. Portanto, embora eficiente em cenários com atributos categóricos de domínio reduzido, a escalabilidade do método é limitada à medida que cresce o número de categorias.

Por fim, Yan et al. (2021) desenvolveram o algoritmo de agrupamento ponderado de k-membros *Weighted k-Member Clustering Anonymization* (WKMCA). Este algoritmo usa uma série de indicadores de peso para avaliar a discrepância dos registros, a distância entre os registros e a perda de informações dos dados publicados. Esses recursos ajudam a reduzir a influência de dados com valores fora da curva (*outliers*) produzindo grupos otimizados. O WKMCA inspira diretamente abordagens modernas, como o *framework* GFKMC Coelho et al. (2024c). O GFKMC se distingue das outras abordagens ao aplicar generalização a atributos quase identificadores no início do procedimento e ao operar em grupos de registros em vez de registros individuais. Esta estratégia reduz o custo computacional de geração de agrupamentos e minimiza a perda de informações para valores mais altos de k. Para satisfazer os requisitos de k-anonimato, Yan et al. usam microagregação após

agrupar os atributos, substituindo os dados do grupo pelo respectivo valor do centróide do grupo. Em contraste, a metodologia proposta permite o uso de maneiras distintas para generalizar o agrupamento com base na substituição pelos valores do centróide do grupo, o registro mais comum ou os valores mais comuns. Ao agrupar quase identificadores antecipadamente, o GFKMC obtém classes de equivalência que satisfazem o *k*-anonimato com uma operação de filtro simples, mesmo antes das etapas de agrupamento.

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

As técnicas de anonimização, embora fundamentais para a preservação da privacidade, podem impactar negativamente a acurácia de modelos de aprendizado de máquina, especialmente quando aplicadas a dados sensíveis. Este impacto tem motivado uma linha crescente de pesquisa dedicada à avaliação dos efeitos da anonimização na performance dos modelos.

Slijepčević et al. Slijepčević et al. (2021) conduziram uma análise sistemática dos impactos de diferentes técnicas de *k*-anonimato, incluindo Mondrian, CB, TDG e OLA, sobre modelos supervisionados como SVM, k-NN, Random Forest e XGBoost. Os resultados evidenciam que a degradação da acurácia é proporcional ao aumento do parâmetro *k*, com destaque para o desempenho superior do Mondrian entre os algoritmos avaliados.

No contexto federado, Kwatra et al. Kwatra and Torra (2021) analisaram os efeitos do *k*-anonimato, aplicado via Mondrian, sobre classificadores baseados em árvores de decisão. Os resultados indicam que, mesmo para valores elevados de *k*, a perda de desempenho é aceitável em alguns cenários, corroborando as conclusões obtidas em ambientes centralizados. Saleh et al. Saleh (2022) aprofundaram essa análise ao comparar o impacto de diferentes modelos de privacidade — *k*-anonimato, *l*-diversidade e *t*-proximidade — sobre redes neurais do tipo *Multi-Layer Perceptron* (MLP) em ambientes de aprendizado federado. Os resultados apontam o *k*-anonimato como a estratégia que melhor equilibra privacidade e desempenho.

O estudo conduzido por Choudhury et al. Choudhury et al. (2020) insere uma perspectiva adicional ao avaliar o impacto da anonimização em modelos explicáveis. Utilizando o algoritmo MDAV em conjunto com a técnica de interpretabilidade TreeSHAP, os autores demonstram que é possível preservar, até certo ponto, tanto a acurácia quanto a interpretabilidade dos modelos. O *framework* GFKMC Coelho et al. (2024c), desenvolvido no âmbito desta tese, também foi submetido à avaliação

empírica quanto ao seu impacto sobre modelos de ML. Resultados comparativos com Mondrian, CB e TDG indicam que a estratégia de generalização antecipada, aliada ao agrupamento eficiente, permite reduzir significativamente a perda de informação, mantendo a estabilidade dos modelos, tanto em contextos centralizados quanto federados Coelho et al. (2025b). De forma geral, a literatura reforça que o impacto da anonimização sobre modelos de aprendizado de máquina é uma questão crítica, que demanda metodologias de avaliação robustas, capazes de mensurar não apenas o risco de reidentificação, mas também a degradação da performance dos modelos.

3.1 Resumo

A análise dos trabalhos relacionados evidencia que, embora existam avanços significativos na construção de algoritmos de anonimização, desafios importantes continuam em aberto. Estes desafios estão principalmente associados à necessidade de balancear de forma eficiente privacidade e utilidade dos dados. Além disso, são raros os modelos metodológicos robustos que avaliam de forma conjunta os impactos sobre privacidade, utilidade e desempenho de modelos preditivos. Neste contexto, o desenvolvimento do *framework* GFKMC busca endereçar parte dessas lacunas, propondo uma abordagem que alia baixo custo computacional, estabilidade na perda de informação e preservação da performance de modelos de aprendizado de máquina, contribuindo, assim, de forma significativa para o avanço do estado da arte em anonimização de dados sensíveis.

Capítulo 4

Um Framework de Anonimização Baseado em k-Anonimato para Garantir a Privacidade de Dados para Internet das Coisas Aplicada na Saúde

Considerando a importância fundamental da preservação da privacidade dos dados de saúde e o impacto dos métodos tradicionais na utilidade desses dados, torna-se imprescindível o desenvolvimento de novas abordagens para a anonimização. É necessário que estas abordagens sejam capazes de equilibrar a proteção da informação com a manutenção de sua relevância para análises e aplicações científicas. Portanto, este capítulo detalha por completo o *framework* projetado para realizar *k*-anonimização baseada em generalização antecipada e agrupamento de *k*-membros para dados de saúde GFKMC (*Generalization First k-Member Clustering*). A representação visual da solução é ilustrada pela Figura 4.1. Esta ferramenta versátil aplica anonimização dinâmica por separatrizes a atributos quase identificadores numéricos e generalização hierárquica a quase identificadores categóricos. Simultaneamente, as árvores de taxonomia fornecem uma hierarquia clara de generalização para dados categóricos. Os valores de quase identificadores categóricos são generalizados para seus respectivos valores pais, reduzindo o custo computacional da geração dos *k* grupos. Para atender aos requisitos de *k*-anonimato, um procedimento de reagrupamento é implementado. Ele começa agrupando registros que compartilham valores de quase identificadores semelhantes. Se um grupo tiver *k* ou mais registros, uma classe de equivalência é formada por estes registros. Caso contrário, as outras classes de equivalência são formadas aplicando operações de reagrupamento nos grupos restantes, em vez dos registros individuais, reduzindo efetivamente a sobrecarga computacional do procedimento de agrupamento.

Em resumo, este capítulo detalha por completo o *framework* para garantir a privacidade de dados da saúde proposto. A Seção 4.1 apresenta o método de

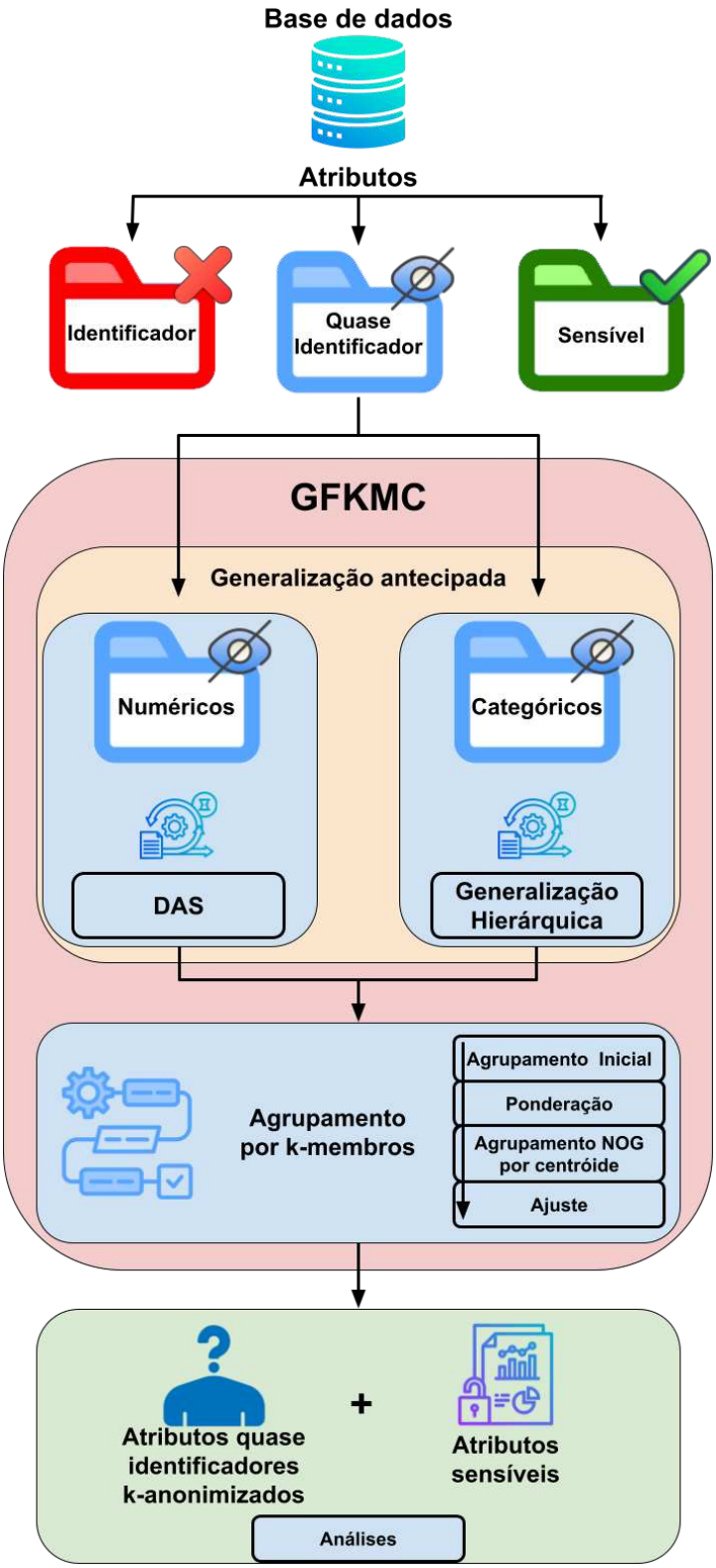


Figura 4.1: Representação visual do *framework* GFKMC.

anonimização dinâmica por separatriz para anonimizar atributos numéricos. A Seção 4.2 descreve os métodos para anonimizar os atributos categóricos, enfatizando a fase inicial de generalização que precede o agrupamento. A Subseção 4.3 detalha o

algoritmo de anonimato completo para garantir que cada agrupamento atenda aos critérios de *k*-anonimato, aumentando assim a privacidade dos dados e mantendo a utilidade dos dados.

4.1 Anonimização Dinâmica de Atributos Quase Identificadores Numéricos

Esta seção apresenta o método para Anonimização Dinâmica por Separatriz (*Dynamic Anonymization by Separatrices* – DAS), para anonimizar atributos numéricos. Este método objetiva definir o valor ideal k e para o agrupamento dinâmico dos atributos numéricos a serem anonimizados usando medidas de separatrizes. Esta anonimização de dados baseada em grupos ordenados em partes definidas por *k*-percentil é capaz de modificar informações que identifiquem uma pessoa (atributos quase identificadores numéricos) para garantir a preservação de privacidade e maior fidelidade em relação aos dados puros. O método é dinâmico, pois define o valor ideal para a quantidade de grupos (k) utilizando o método estatístico conhecido como "método cotovelo" ou método de Elbow. Em seguida, os grupos são delimitados pelos respectivos percentis. Todos os valores de um agrupamento pertencente a um *k*-percentil são generalizados, substituindo-os pela respectiva média de valores deste intervalo. Em resumo, o método DAS possui duas contribuições principais. Primeiro, ela define o valor ideal para a quantidade de grupos (k). Segundo, ela realiza a anonimização de atributos numéricos com baixo custo computacional e agrupamentos definidos por separatrizes. A definição dinâmica do valor ideal de k e a aplicação da anonimização por separatrizes em cada tipo de atributo numérico proporcionam maior diversidade/heterogeneidade entre as classes de equivalências dos atributos.

O método DAS Coelho et al. (2024b) toma como base a construção dos agrupamentos por separatrizes. As separatrizes são valores que ocupam determinados lugares de uma distribuição de frequência Correa (2003). Podemos classificá-las de acordo com o número de partes iguais em que os dados são particionados. Dessa forma tem-se, por exemplo, quartis (4 partes) e decis (10 partes) Correa (2003). Nesse sentido, as separatrizes permitem dividir uma distribuição em n partes iguais. A Figura 4.2 ilustra a distribuição em quartis.

O DAS trata o problema de escolha do valor de k para atributos numéricos, uma vez que este é escolhido dinamicamente. Testar diversos valores de k é uma solução cara e escolher um valor arbitrário de k representa um risco significativo quanto à divulgação de atributos. Os exemplos na literatura indicam um valor de $k = 3$. Entretanto, alguns autores Victor and Lopez (2020); Torra and Navarro-Arribas

Figura 4.2: Representação da divisão dos dados em medidas de separatrizes (quartil).

(2023) sugerem o valor de $k = 5$. Porém, quanto maior o k, menor o risco, uma consequência natural da dimensionalidade no agrupamento. Portanto, aliado à escolha dinâmica, o método DAS contribui para garantir a eficiência quanto à privacidade dos dados, aumentando o valor associado a k. Além da eficiência inerente à privacidade, o método DAS tem uma construção de agrupamento anônimo de dados com baixo custo computacional, o que lhe permite operar em *hardware* com recursos limitados.

O DAS aplica o método de Elbow que é tradicionalmente utilizado para encontrar a quantidade ideal de grupos em técnicas de agrupamento, e este pode ser estendido à anonimização. Portanto, o método de Elbow é responsável por definir o valor ideal de k dinamicamente, o qual define o respectivo k-percentil conforme a distribuição dos atributos numéricos. O método de Elbow Bholowalia and Kumar (2014) é um método gráfico utilizado para determinar um valor adequado referente ao número de grupos k, de forma que a adição de outro agrupamento não reduza significativamente a função de custo a ser minimizada. A ideia é começar com $k = 2$ e incrementá-lo em cada etapa, calculando o custo para cada valor. O custo é baseado no cálculo da distorção, ou seja, a média das distâncias quadradas dos centros dos respectivos grupos até cada ponto de dados. Normalmente, a métrica de distância euclidiana é usada. Em algum ponto, o custo diminui drasticamente e depois se estabiliza para valores maiores Kodinariya et al. (2013). Este ponto representa o valor ótimo de k, que pode ser detectado automaticamente utilizando algoritmos como o *Kneedle* Satopaa et al. (2011). Embora existam outras abordagens na literatura para determinar o número ideal de grupos, o método de Elbow se destaca por seu tempo de execução menor, comparado com os outros métodos da literatura (Estatística de Gap, Método da silhueta e *Canopy* Yuan and Yang (2019). O valor ótimo de k define as medidas de separatrizes, as quais ocupam determinados lugares de uma distribuição de frequência, delimitando assim os k grupos de atributos com características a serem generalizadas.

Ao DAS, espera-se que os atributos QIs numéricos sejam fornecidos em uma estrutura de dados tabular potencialmente heterogênea com eixos (linhas e colunas) rotulados (identificadores – IDs e números). O Algoritmo 4.1 descreve o conjunto de instruções necessárias para realizar a anonimização dinâmica dos atributos numéricos usando separatrizes. O valor de k, o qual será utilizado como limiar para

definir os percentis, é definido individualmente para cada QI. Em seguida, o algoritmo processa cada QI de forma ordenada. Para obter cada valor separatriz, calcula-se o q-ésimo percentil dos dados ao longo do eixo (QI) especificado. O método DAS utiliza o parâmetro "*closest_observation*", o qual estima a observação mais próxima de um valor ideal para quantidade de percentis Hyndman and Fan (1996); Developers (2024).

O algoritmo processa cada quase identificador individualmente. Na linha 3, o método de Elbow define o valor ideal de k (k-percentil) para o QI. Na Linha 4, o algoritmo ordena o QI de forma crescente. Para todos os registros pertencentes ao QI, na Linha 7, calcula-se o valor do q-ésimo percentil. Em seguida, na Linha 8, encontra-se o maior índice que corresponde ao respectivo q-ésimo percentil, delimitando assim a faixa de amostras as quais serão anonimizadas por generalização. Na Linha 9, calcula-se o valor médio das amostras pertencentes ao respectivo q-ésimo percentil. Posteriormente, na Linha 10, todas as amostras pertencentes ao intervalo são substituídas pelo valor médio calculado no passo anterior. Na Linha 12, os valores anonimizados dinamicamente por agrupamento baseado em q-ésimo percentil são atribuídos à tabela de *QIs_anonimos*. Este fluxo se repete até processar todos os QIs. Por fim, o algoritmo retorna a tabela anonimizada *QIs_anonimos*, na linha 13.

Algoritmo 4.1: Anonimização Dinâmica por Separatriz para atributos numéricos.

```
   Entrada: QIs
   Saída: QIs_anonimos
1  início
2      para cada QI em QIs faça
3          k = Elbow (Unicos(QI));
4          QI = Ordena(QI);
5          id_inicio = 0;
6          para i ← 1 até k + 1 faça
7              separatriz = Percentil(QI, ((100 / k) * i),
               metodo="closest_observation");
8              id_fim = Maior(onde(QI == separatriz));
9              QI_media = Media(QI, id_inicio, id_fim);
10             QI = Substitui_media(QI, QI_media, id_inicio, id_fim);
11             id_inicio = id_fim + 1;
12         Inserir(QIs_anonimos, QI)
13     retorna QIs_anonimos
```

A Tabela 4.1 ilustra como é representada a estrutura tabular para dados brutos, os quais serão tratados pelo método DAS. Já a Tabela 4.2 apresenta os respectivos dados anonimizados pelo método proposto. Neste exemplo, considerando $k = 3$, os QIs são

anonimizados na ordem em que aparecem (idade, altura em centímetros e peso em quilogramas). Observa-se pelos IDs 1 e 4 da Tabela 4.2 que o algoritmo não satisfaz as condições do k-anonimato para o valor $k = 3$, definido por Elbow. Neste caso, para cada um dos registros da tabela, não existem pelo menos $k - 1$ outros registros com valores idênticos de QIs. Entretanto, por não exigir a condição do k-anonimato, o DAS gera maior heterogeneidade entre os dados em respectivos agrupamentos por percentil e, portanto, menor perda de informação útil.

Tabela 4.1: Dados brutos.

ID	Idade	Altura (cm)	Peso (Kg)
0	21	160	50,55
1	24	154	60,60
2	25	158	48,80
3	30	170	76,80
4	34	169	54,70
5	33	176	67,90
6	38	183	79,00
7	41	190	80,60
8	39	180	83,10

Tabela 4.2: Dados anonimizados.

ID	Idade	Altura (cm)	Peso (Kg)
2	23	157	51,35
0	23	157	51,35
4	32	171	51,35
1	23	157	68,43
5	32	171	68,43
3	32	171	68,43
6	39	184	80,90
7	39	184	80,90
8	39	184	80,90

4.2 Generalização Antecipada de Quase Identificadores Categóricos

Esta subseção detalha os métodos usados no GFKMC para anonimizar quase identificadores categóricos. Estendemos a anonimização para atributos categóricos usando Hierarquias de Generalização (*Generalization Hierarchies* – GH) para garantir anonimização robusta. Ambas as abordagens (DAS e GH) são projetadas para garantir o anonimato de quase identificadores com perda mínima de informações. Além disso, antecipar a anonimização melhora o desempenho do método de agrupamento para k-anonimato, pois aumenta o número de classes de equivalência, melhorando assim sua eficiência geral.

Utilizar árvores de taxonomia é uma maneira eficiente para atingir a anonimização de dados categóricos. Essas árvores descrevem uma hierarquia de generalização a qual é fundamental no processo de anonimização. As representações dos valores dos atributos são localizadas nos nós da árvore. Os nós folha representam os valores originais, enquanto os nós pais significam os valores menos específicos. O nó raiz, por outro lado, representa a generalização completa, geralmente culminando na supressão completa da informação, comumente denotada pelo valor "*".

Antes da etapa de agrupamento, outra tarefa preparatória essencial envolve o uso

de árvores de taxonomia predefinidas. Essas árvores desempenham um papel significativo no processo, pois generalizam todos os valores categóricos de quase identificadores para seu valor pai. Por exemplo, na Figura 4.3a, Japão e Índia seriam generalizados para a Ásia. No entanto, para qualquer atributo categórico C e sua árvore de taxonomia $Tree_C$, se a altura da árvore de taxonomia for um, como na Figura 4.3b, então C não é generalizado para reduzir a perda de informações. Isso ocorre porque C seria generalizado para "*", removendo qualquer informação útil.

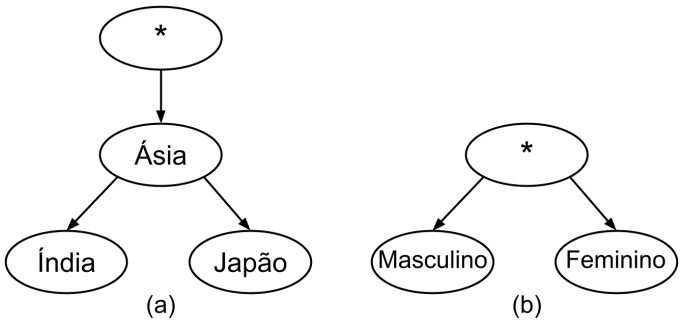


Figura 4.3: Representação de árvores de taxonomia para (a) países e (b) gêneros.

4.3 Agrupamento por k-membros

O algoritmo agrupamento por k-membros é construído em quatro fases cruciais, com a Fase Inicial de Agrupamento servindo como a base fundamental da proposta. Conforme ilustrado no Algoritmo 4.2, esta fase agrupa registros com os mesmos valores de quase identificador previamente anonimizados para o mesmo grupo e somente forma uma classe de equivalência com um grupo se o tamanho do grupo for maior ou igual ao valor k. A pré-generalização dos quase identificadores é uma etapa fundamental nesta fase, pois aumenta significativamente a quantidade de grupos k-anônimos formados de antemão. Este aumento dos grupos k-anônimos desempenha um papel fundamental na redução da sobrecarga computacional das fases subsequentes, tornando o *framework* GFKMC uma solução eficiente.

O *framework* GFKMC, em sua fase inicial, toma como entrada a estrutura de dados tabular T_G que foi previamente generalizada com o algoritmo DAS para QIs numéricos e árvores de taxonomia para os QIs categóricos. Os registros contidos em T_G com os mesmos valores de QI são agrupados na linha 4. Entre as linhas 5 e 9, os grupos são atribuídos às listas CLS e RG com base em seus respectivos tamanhos. CLS é a lista de grupos que satisfazem k-anonimato, enquanto RG representa a lista de grupos restantes que não têm k ou mais registros em cada classe de equivalência. Na linha 9, **Cluster**(G) denota a criação de um grupo que consiste nos registros do

grupo G, com um registro arbitrário de G como o centróide do grupo. Finalmente, na linha 10, o algoritmo retorna as listas CLS e RG.

Algoritmo 4.2: Fase Inicial de Agrupamento

```
   Entrada: T_G, k
   Saída: CLS, RG
1  início
2      CLS = Lista();
3      RG = Lista();
4      groups = Registros do grupo T_G com valores QI equivalentes;
5      para cada G ∈ groups faça
6          se Tamanho(G) < k então
7              Acrescentar(RG, G);
8          senão
9              Acrescentar(CLS, Cluster(G))
10     retorna CLS, RG
```

Em seguida, o algoritmo 4.3 detalha a Fase de Ponderação. O objetivo desta fase é filtrar os grupos *outliers* avaliando-os de acordo com sua pontuação de peso e parâmetro β. Ele recebe como entrada a lista de grupos restantes RG, que não satisfazem k-anonimato, e o parâmetro β. Nesta tese, o valor β é definido como $0,05|T_G|$, onde $|T_G|$ é o número de registros na tabela T_G.

Na linha 2, para cada grupo $G \in RG$, a pontuação de peso de todos os registros em G é computada de acordo com a Definição 8. Como a pontuação de peso considera apenas valores de QI, ela é constante para todos os registros dentro de um grupo. Portanto, o algoritmo calcula apenas a pontuação de peso para um único registro de cada grupo, reduzindo cálculos redundantes. Nas linhas 3 a 7, o valor β é utilizado para determinar grupos *outliers* e não *outliers*. Especificamente, nas linhas 3 e 4, se β for zero, todos os grupos em RG serão considerados não *outliers*. Caso contrário, na linha 6, os registros serão ordenados em ordem decrescente de pontuação de peso. Na linha 7, o valor da pontuação de peso do β_{th} registro na lista *sorted_scores* classificado é selecionado como o valor limite. Esse limite é usado para classificar um grupo como *outlier* ou não *outlier*. Então, nas linhas 8 a 13, os grupos são atribuídos a OG ou NOG, de acordo com o limite $outlier\_value$. Finalmente, ao final do processamento, na linha 14, as listas NOG e OG são retornadas.

A Fase de Agrupamento é descrita no Algoritmo 4.4. Ele recebe como entrada a lista NOG e o valor k, e retorna a lista de novos agrupamentos new_CLS. Nesta fase, os grupos são formados usando a métrica WIL, proveniente da Definição 10. O objetivo é minimizar a perda de informação devido à generalização dos grupos recém-formados new_CLS. Isso ocorre porque os QIs de cada registro dentro de um grupo serão generalizados para satisfazer o k-anonimato na fase subsequente.

Algoritmo 4.3: Fase de Ponderação

```
   Entrada: T_G, RG, β
   Saída: NOG, OG
1  início
2      weight_scores = Weight_score(RG);
3      se β == 0 então
4          NOG = RG;
5      senão
6          sorted_scores = Ordenar(weight_scores, reverse=True);
7          outlier_value = sorted_scores[β];
8          para cada G ∈ RG faça
9              r = Selecione qualquer registro de G;
10             se weight_scores[r] >= outlier_value então
11                 Acrescentar(OG, G)
12             senão
13                 Acrescentar(NOG, G);
14     retorna NOG, OG
```

O algoritmo começa na linha 2, onde a quantidade de registros não *outliers* é calculada computando a soma do tamanho de todos os grupos em *NOG*. Na linha 3, N_G é o número de grupos em *RG*, e na linha 4, N_C é a quantidade máxima de novos grupos a serem formados. Então, na linha 5, uma matriz de distância $N_G \times N_C$ vazia *DM* é criada para ser usada no Algoritmo 4.5. Na linha 6, uma lista vazia para novos grupos *new_CLS* é criada. A linha 7 inicia um *loop while* para formar grupos até que o número de registros em *NOG* seja insuficiente para formar um grupo de tamanho *k*. Entre as linhas 8 e 11, se *new_CLS* estiver vazio, o algoritmo seleciona um grupo aleatório de *NOG* como o grupo *CG*. Caso contrário, o algoritmo Encontrar o Próximo Centróide 4.5 determina *CG*. Na linha 12, um novo grupo é criado com *CG*. Então, na linha 13, *CG* é removido da lista *NOG*, e o tamanho do grupo removido é subtraído de *NO_count* na linha 14.

Na linha 15, para cada grupo $G \in NOG$, o algoritmo seleciona qualquer registro *r* em *G* e calcula a alteração no valor da Perda de Informação de Peso (WIL) ao adicionar *r* ao *cluster*. Essa mudança é computada como **WIL**($cluster \cup r$) - **WIL**(*cluster*), onde **WIL** é calculado de acordo com a Definição 10, e *cluster* é o *cluster* da linha 12. Posteriormente, os valores computados em *groups_wil* são classificados em ordem crescente. A ideia é mesclá-los em grupos *cluster*, minimizando a mudança de WIL. Esta etapa de mesclagem ocorre nas linhas 17 a 28, onde para cada $G \in sorted\_groups$, o número de registros *n* necessários para *cluster* atingir o tamanho de *k* é computado na linha 18. Na linha 19, se o grupo já possuir *k* registros, o algoritmo pula para a linha 29, onde o grupo é adicionado a *new_CLS*. Caso contrário, entre as linhas 21 a

28, se o tamanho de G for menor ou igual a n, ele é mesclado em *cluster* e removido de *NOG*. Caso contrário, n registros são movidos de G para *cluster*.

Algoritmo 4.4: Fase de Agrupamento

```
   Entrada: NOG, k
   Saída: new_CLS
1  início
2      NO_count = Soma do tamanho de todos os grupos em NOG;
3      N_G = Tamanho(RG);
4      N_C = Floor(NO_count / k);
5      DM = Crie uma matriz de distância vazia N_G × N_C;
6      new_CLS = Lista();
7      enquanto NO_count >= k faça
8          se new_CLS == Vazia então
9              CG = Selecione um grupo aleatório do NOG;
10         senão
11             CG = Encontrar_Próximo_Centróide(NOG, new_CLS, DM);
12         cluster = Cluster(CG);
13         Remove(NOG, CG);
14         NO_count −= Tamanho(CG);
15         groups_wil = Para cada G ∈ NOG, selecione qualquer registro r em G e
           calcule WIL(cluster ∪ r) - WIL(cluster);
16         sorted_groups = Ordenar(groups_wil);
17         para cada G ∈ sorted_groups faça
18             n = k − Tamanho(cluster);
19             se n == 0 então
20                 break;
21             se Tamanho(G) <= n então
22                 cluster = cluster ∪ G;
23                 Remove(NOG, G); NO_count −= Tamanho(G);
24             senão
25                 subgroup = Selecione n registros de G;
26                 cluster = cluster ∪ subgroup;
27                 NO_count −= Tamanho(subgroup);
28                 G = G \ subgroup;
29         Acrescentar(new_CLS, cluster);
30     retorna new_CLS
```

O algoritmo 4.5 objetiva Encontrar o Próximo Centróide. Ele usa a métrica *Weight Distance* da Definição 9 para selecionar o centróide do grupo *CG* com distância máxima entre os grupos formados anteriormente *new_CLS*. O algoritmo depende essencialmente da matriz de distância *DM* para reter valores de distância previamente computados, reduzindo efetivamente computações redundantes.

Algoritmo 4.5: Encontrar Próximo Centróide

```
   Entrada: NOG, new_CLS, DM
   Saída: CG
1  início
2      dists = Lista();
3      para cada G ∈ NOG faça
4          r = Seleciona qualquer registro de G;
5          para cada cluster ∈ new_CLS faça
6              centroid = Centroid(cluster);
7              dist = DM[G][cluster];
8              se dist == Vazia então
9                  dist = Weight_distance(r, centroid);
10                 DM[G][cluster] = dist;
11             Acrescentar(dists, [G, dist]);
12     retorna CG
```

A fase de ajuste é o último ponto do *framework* GFKMC, que é descrito no Algoritmo 4.6. Esta fase é responsável por alocar os grupos restantes de *NOG* e os grupos *outliers OG* para um novo grupo em *new_CLS* que minimiza a mudança em WIL. Posteriormente, para garantir a conformidade com as restrições de *k*-anonimato, todos os grupos que não atendem a esse requisito são generalizados. Para fornecer maior flexibilidade na aplicação do GFKMC, a generalização dos agrupamentos pode ser obtida por meio de uma das seguintes técnicas: centróide do grupo, registro mais comum ou valor mais comum. A técnica do centróide do grupo envolve a substituição dos valores dos quase identificadores dos registros dentro de cada grupo por seus respectivos centróides. O método de registro mais comum substitui os QIs de cada grupo pelos valores do registro mais comum, ou seja, pela linha QI mais recorrente. Por fim, a abordagem de valor mais comum substitui cada QIs do grupo pelo valor mais comum de cada coluna QI.

Neste ponto, o algoritmo recebe as listas *CLS*, *new_CLS*, *NOG* e *OG* como entrada e gera a tabela final anonimizada T_{anon}. Nas linhas 2 a 6, enquanto *NOG* não estiver vazio, um grupo *G* é selecionado aleatoriamente. Com tal grupo, na linha 4, o algoritmo encontra o agrupamento que minimiza a mudança no WIL como **WIL**($cluster \cup G$) - **WIL**($cluster$). Vale ressaltar que todo o grupo *G* é passado para calcular o WIL, em contraste com o Algoritmo 4.4. Isso ocorre porque, nesse caso, o grupo inteiro deve ser mesclado (linha 5), enquanto no Algoritmo 4.4, o número de registros do grupo que serão mesclados é desconhecido naquele estágio do algoritmo. Na linha 7, as mesmas operações das linhas 2 a 6 são repetidas, mas dessa vez para a lista *OG* em vez da lista *NOG*. Finalmente, nas linhas 8 e 9, para cada grupo $c \in new\_CLS$, todos os valores quase identificadores de registros dentro de *c*

são substituídos por c valores quase identificadores referentes ao centróide. Então, *CLS* e *new_CLS* são mesclados, e o resultado é a tabela final completamente anonimizada que é retornada na linha 10.

Algoritmo 4.6: Fase de Ajuste

```
   Entrada: CLS, new_CLS, NOG, OG
   Saída: T_anon
1  início
2      enquanto Não_Vazia(NOG) faça
3          G = Selecione um grupo do NOG aleatoriamente;
4          min_cluster = Para cada cluster ∈ new_CLS, encontre o grupo que
           minimiza WIL(cluster ∪ G) - WIL(cluster);
5          cluster = cluster ∪ G;
6          Remove(NOG, G);
7      Repita as mesmas operações das linhas 2 a 6 para OG;
8      new_CLS = Generalizar registros de cluster;
9      T_anon = CLS ∪ new_CLS;
10     retorna T_anon;
```

4.4 Resumo

Este capítulo apresenta, de forma detalhada, a concepção e o desenvolvimento do *framework Generalization First k-Member Clustering*, elaborado com o propósito de assegurar a privacidade de dados sensíveis no contexto da Internet das Coisas aplicada à saúde. O *framework* combina a técnica de anonimização dinâmica por separatrizes e generalização hierárquica. Para atributos numéricos, é utilizado o método DAS, que agrupa dados com base em separatrizes estatísticas, ajustando dinamicamente o valor de k para cada atributo. Para atributos categóricos, emprega-se uma abordagem de generalização antecipada baseada em hierarquias taxonômicas. Além disso, o processo de agrupamento por k-membros tem como objetivo constituir classes de equivalência que satisfazem rigorosamente os princípios do k-anonimato, permitindo, assim, um balanceamento adequado entre os níveis de privacidade e a preservação da utilidade dos dados. O capítulo seguinte descreve a metodologia adotada para a avaliação experimental do *framework* proposto, considerando métricas de privacidade, perda de informação e impacto no desempenho de modelos de aprendizado de máquina.

Capítulo 5

Metodologia de Avaliação do GFKMC

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. Entretanto, quanto maior o compromisso com a privacidade, maior a generalização dos dados e a respectiva perda de informações Ghinita et al. (2007). 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 e comparativas, além da eficiência de modelos classificadores.

As informações pessoais, principalmente em grandes volumes, fornecem insumos para empresas, como instituições médicas, universidades e organizações a desenvolver soluções baseadas em aprendizado de máquina para propósito específico ou uso geral. Portanto, a privacidade tornou-se uma importante preocupação também no âmbito do aprendizado de máquina. Deste modo, o aprendizado federado (FL) tem ganhado destaque como uma solução promissora para treinar modelos de aprendizado de máquina de forma descentralizada, preservando a privacidade dos dados ao mantê-los nos dispositivos locais. Nesse contexto, o FL surge como uma solução inovadora para treinamento descentralizado de modelos, mantendo os dados em dispositivos locais e reduzindo riscos 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 foi proposta Ciriani et al. (2008), 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 diferentes 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 partilha os mesmos valores entre os atributos (valor atribuído a *k*), maior a degradação do desempenho da classificação. Outros trabalhos ainda realizam avaliações de desempenho para soluções baseadas em FL e objetivam analisar propostas específicas, como em Kwatra and Torra (2021); Choudhury et al. (2020). Entretanto, o impacto combinado da anonimização sobre modelos ML, centralizados e descentralizados, ainda carece de estudos aprofundados, especialmente no que diz respeito à proposição de uma metodologia robusta que permita avaliar o equilíbrio entre a preservação da privacidade, a utilidade dos dados e a eficiência dos modelos treinados.

Este capítulo 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 Coelho et al. (2025a,b). 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, os modelos de aprendizado de máquina, o levantamento de métricas adequadas e as análises resultantes. 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. A Figura 5.1 ilustra o processo metodológico decomposto em cinco fases principais.

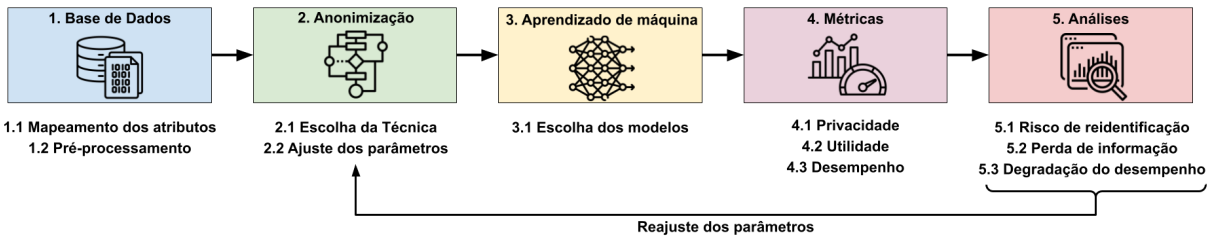


Figura 5.1: Estrutura da metodologia proposta

5.1 Definição da Base de Dados

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 (QI - *quasi-identifiers*) 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.

5.2 Escolha da Técnica de Anonimização e Ajustes dos Parâmetros

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, 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. É importante ressaltar que não existem abordagens específicas para determinar um valor ótimo para o parâmetro k, e consequentemente obter o equilíbrio ideal entre privacidade e utilidade Domingo-Ferrer et al. (2022). Dessa forma, cabe ao programador esta complexa tarefa de escolha empírica do valor. Portanto, esta metodologia contribui com análises eficientes, as quais produzem *insights* para que um valor adequado seja inferido, proporcionando um reajuste eficiente dos parâmetros.

5.3 Escolha dos Modelos de Aprendizado de Máquina

A etapa 3 da metodologia refere-se ao levantamento de requisitos fundamentais, os quais conduzem à 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.

5.4 Levantamento das Métricas de Avaliação

A avaliação da solução inclui examinar o conjunto de dados anonimizados para verificar se o compromisso entre a privacidade e a 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 o 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 que avaliem a quantidade de perda de informações sobre a base resultante. Como exemplos detas métricas cita-se a *Normalized Certainty Penalty* e a *Discernibility Metric*. 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.

5.5 Análises

A etapa 5 consiste em realizar profundas análises sobre o compromisso entre a utilidade dos dados e o 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 específico de cada base e modelo. Essa metodologia fornece uma base sistemática para compreender o impacto da generalização dos dados introduzida por técnicas de anonimização em aplicações ML, promovendo o desenvolvimento de soluções que conciliem privacidade, utilidade dos dados e eficiência em sistemas distribuídos.

5.6 Resumo

Este capítulo apresentou uma metodologia sistemática destinada à avaliação da eficiência do *framework* GFKMC, especificamente sob a ótica da preservação da privacidade e da manutenção da utilidade dos dados. A metodologia é estruturada em cinco etapas principais, (i) definição das bases de dados empregadas nos experimentos, (ii) seleção dos métodos de anonimização, incluindo o ajuste criterioso de seus parâmetros, (iii) escolha dos modelos de aprendizado de máquina utilizados nas análises, (iv) estabelecimento das métricas de avaliação, englobando indicadores de privacidade, perda de informação e desempenho preditivo e finalmente, (v) com a apresentação das análises experimentais. Tal abordagem metodológica busca proporcionar uma avaliação abrangente, permitindo quantificar de forma objetiva os efeitos da anonimização sobre diferentes cenários de aplicação, tanto em ambientes centralizados quanto federados. Na sequência, o Capítulo 6 apresenta e discute os resultados obtidos, evidenciando o desempenho do *framework* GFKMC frente aos métodos tradicionais de anonimização.

Capítulo 6

Resultados e Discussões

Este capítulo apresenta o cenário para realizar a avaliação do *framework* GFKMC proposto. A avaliação é composta por duas partes principais. Primeiro será avaliado de forma independente o método de anonimização dinâmica de atributos quase identificadores numéricos. Esta avaliação compreende a descrição das bases de dados utilizadas, as métricas de avaliação empregadas e as configurações detalhadas do ambiente computacional utilizado para a execução dos experimentos.

Em seguida, guidado pela metodologia apresentada no Capítulo 5, serão apresentados cenários de avaliação os quais permitem analisar o impacto da anonimização na preservação da privacidade e na utilidade dos dados para o *framework* GFKMC como um todo. Além disso, o capítulo introduz cenários de testes voltados para investigar e comparar os efeitos de diferentes abordagens de anonimização, incluindo o GFKMC, em modelos de aprendizado de máquina tradicionais e federados. Para estes cenários, também são apresentadas as bases de dados e suas características, as métricas utilizadas para a avaliação do desempenho dos modelos e os detalhes da infraestrutura computacional empregada na realização dos experimentos.

6.1 DAS

Esta seção descreve a metodologia de avaliação do método DAS. Serão apresentadas as especificações das bases de dados. Também são apresentadas as métricas de avaliação e detalhadas as configurações da máquina utilizada para execução dos experimentos. Por fim, são divulgados e discutidos os resultados obtidos.

6.1.1 Base de Dados

Adult Becker and Kohavi (1996) é uma base de dados construída para prever se o salário de um indivíduo ultrapassa $50 mil por ano, baseada em dados de senso, incluindo quase identificadores como idade, sexo, educação, ocupação e raça, além do atributo sensível salário. Ela é utilizada para avaliação na maioria das propostas

para métodos de *k*-anonimato Bache and Lichman (2013); Khan et al. (2020); Onesimu et al. (2022); Torra and Navarro-Arribas (2023); Byun et al. (2007), inclusive em cenário de aplicação IoHT Arava and Lingamgunta (2020). A base de dados Adult é composta por 48.842 registros com 14 atributos, onde 7 são categóricos e 7 numéricos. Como o DAS destina-se a anonimizar dados numéricos, apenas estes foram considerados neste artigo. Os dados foram pré-processados para remover as entradas com atributos ausentes. Assim, o arquivo final contém 30.162 registros.

A técnica de anonimização de atributos numéricos proposta também foi avaliada considerando dados reais de saúde, evidenciando ainda mais sua utilidade. Portanto, além da base Adult, a base de dados de saúde WEF – *wearable-exercise-frailty* Sokas et al. (2022) também é utilizada neste trabalho. A base WEF contém dados reais de saúde, incluindo atributos quase identificadores, tais como idade, sexo, altura e peso, além de atributos sensíveis, como eletrocardiograma e aceleração triaxial. A base de dados é constituída por 80 registros e 45 atributos, sendo estes numéricos, categóricos e sensíveis. Para esta base de dados, foram considerados apenas os três atributos numéricos disponíveis, os quais são idade, altura e peso.

6.1.2 Anonimização

A proposta de Anonimização Dinâmica por Separatriz para atributos numéricos foi implementada na linguagem de programação Python, com suporte das bibliotecas scikit-learn, yellowbrick, numpy, pandas e matplotlib. Os experimentos foram conduzidos utilizando uma máquina com 20 GB de memória RAM DDR4 3200 MHz, processador AMD Ryzen 5 5500u. A identificação do valor ideal de k ocorreu através da média de 10 iterações do método de Elbow para cada atributo QI. Em seguida, os quase identificadores foram anonimizados pelas abordagens comparadas (intervalo fixo Onesimu et al. (2022), MDAV, Mondrian Torra and Navarro-Arribas (2023) e DAS), empregando valores de k sugeridos pela literatura e o valor ideal obtido por Elbow. O desempenho do método DAS foi confrontado com o desempenho das demais técnicas, por meio da aferição de métricas de risco de divulgação de atributos e de perda de informação, as quais são descritas na sequência.

6.1.3 Métricas

Privacidade

A divulgação de atributos ocorre quando um atacante tenta adquirir mais conhecimento sobre um indivíduo (por exemplo, diagnóstico). Quando o atacante consegue identificar registros QI de um indivíduo e correlacionar com algum conhecimento prévio, pode ocorrer a vinculação e consequentemente divulgação de

identidade e exposição de atributos sensíveis Onesimu et al. (2022). Uma métrica para avaliar o risco desse tipo de ataque é a vinculação de registros baseada em distância (*distance-based record linkage*) Christen et al. (2020). Este trabalho utiliza a implementação e descrição apresentada em Jiang and Torra (2023). O algoritmo realiza o ataque com o objetivo de calcular a quantidade de registros que possam ser divulgados. Para cada registro $r1$ da base de dados anonimizada $DB1$, calcula-se a distância de $r1$ em relação a cada registro $r2$ da base de dados original $DB2$ (a função de distância utilizada foi a distância euclidiana). Em seguida, seleciona-se o registro $r'(r1)$ mais similar (próximo) de $r1$. Se o registro selecionado $r'(r1)$ corresponder ao registro anonimizado $r1$, então os dois registros foram vinculados corretamente. Assim, a métrica descrita reflete o número de vínculos corretos obtidos pelo algoritmo, em relação ao número total de registros. A correspondência entre os registros é conhecida pelo fato de as bases de dados $DB1$ anonimizada e $DB2$ original serem objetos de estudo deste trabalho.

Utilidade

A avaliação da perda de informação a atributos numéricos ocasionada pela anonimização praticada pelas técnicas é obtida por meio da métrica *Normalized Certainty Penalty* (NCP) Ghinita et al. (2007); Ayala-Rivera et al. (2014), descrita pela equação 6.1. Sejam min_i e max_i o valor mínimo e máximo de um atributo numérico i, respectivamente. Um registro j pertence a uma classe de equivalência com valor máximo max_{ij} e mínimo min_{ij}, para o atributo i. O NCP de uma tabela anonimizada T^* é definido como:

$$NCP(T^*) = \frac{1}{|T| \times n} \times \sum_{i=1}^{n} \sum_{j=1}^{|T|} \frac{max_{ij} - min_{ij}}{max_i - min_i} \tag{6.1}$$

onde $|T|$ é o número de registros e n o número de atributos. A métrica é baseada no conceito de que valores que representam um intervalo maior são menos precisos que valores que representam intervalos menores Ayala-Rivera et al. (2014). O valor do NCP varia entre 1 e 0, em que 0 representa nenhuma perda de informação (dados originais) e 1 representa perda total de informação. Assim, valores baixos, próximos a 0, são desejáveis Ayala-Rivera et al. (2014).

6.1.4 Análises

O estado da arte busca soluções para ataques externos à privacidade de dados sensíveis, especialmente em IoHT. A divulgação de atributos é um dos principais responsáveis pelo vazamento de informações. A reidentificação das informações é

originada a partir de atributos numéricos e categóricos. Exclusivamente, o método DAS se destina a coibir os riscos à privacidade inerentes aos atributos numéricos. Ao considerar soluções baseadas em k-anonimato, um fator imprescindível passa pela escolha do valor de k. As soluções estáticas (Onesimu et al. (2022), MDAV ou Mondrian) atribuem a responsabilidade da segurança ao usuário, o qual define os intervalos fixos de k. A literatura recomenda valores de $k = 3$ ou $k = 5$, porém, esta estimativa depende das características dos dados a serem anonimizados, o que implica diretamente no nível de proteção à privacidade e na utilidade dos dados.

Para o conjunto de dados WEF, ao ser aplicada a definição dinâmica para o valor de k, baseada no método de Elbow, encontra-se o valor ideal de $k = 6$ para peso e altura, como identificado pela interseção da linha tracejada com a linha azul na Figura 6.1. A figura também exibe a quantidade de tempo de ajuste do modelo de agrupamento para cada valor de k como uma linha verde tracejada. Em geral, o valor maior de k é comprovadamente mais seguro. Entretanto, inferir um valor arbitrariamente alto afeta o compromisso entre anonimização e utilidade dos dados, uma vez que a generalização total não representa características de todos os indivíduos. A Tabela 6.1 exibe informações sobre a divulgação de registros para os valores de k sugeridos na literatura e o valor obtido dinamicamente. Para a base de dados WEF, o método de anonimização dinâmica baseada em separatriz obteve desempenho semelhante ao intervalo fixo com $k = 6$, porém, eliminando a responsabilidade de escolha do usuário e o respectivo custo. Em relação à perda de informação das técnicas, a Tabela 6.2 exibe informações que respaldam a aplicação do DAS mesmo em conjuntos menores de dados, com um índice de perda de apenas 0,101, considerando os valores dinâmicos de k.

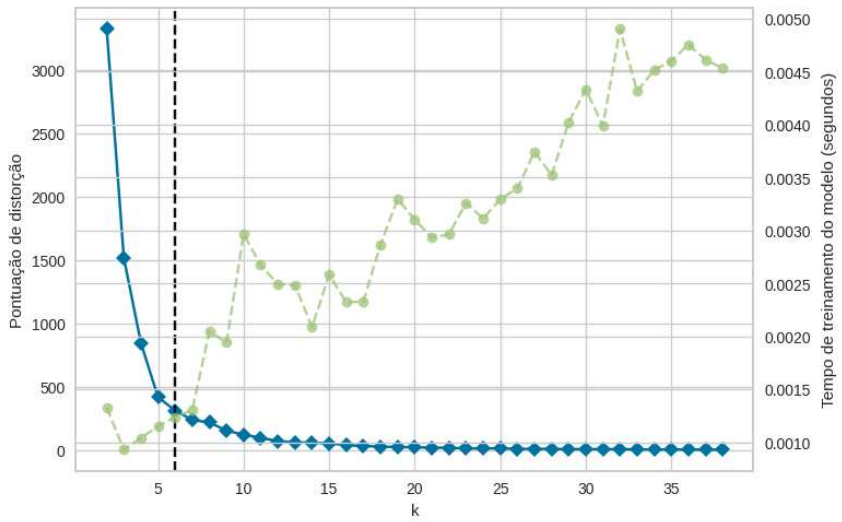


Figura 6.1: Definição do valor ideal de k usando Elbow para a base de dados WEF, considerando o QI peso.

Tabela 6.1: Número de registros anonimizados vinculados corretamente aos 80 registros da base de dados original WEF.

WEF	Intervalo fixo	MDAV	Mondrian	DAS
$k = 3$	17	21	13	-
$k = 5$	38	15	13	-
$k = 6$	48	12	8	-
$idade : k = 4$ $altura : k = 6$ $peso : k = 6$	-	-	-	51

Tabela 6.2: Perda de informação de dados originais para a base WEF.

WEF	Intervalo fixo	MDAV	Mondrian	DAS
$k = 3$	0,184	0,094	0,160	-
$k = 5$	0,107	0,158	0,160	-
$k = 6$	0,086	0,185	0,240	
$idade : k = 4$ $altura : k = 6$ $peso : k = 6$	-	-	-	0,101

O conjunto de dados Adult é consideravelmente maior e com grande aderência junto às avaliações de técnicas de anonimização de dados. Para esta base de dados, o valor ideal atribuído pelo método de Elbow considerando o atributo horas por semana é $k = 9$, como ilustrado pela Figura 6.2. É importante ressaltar que, quanto maior a base de dados, maior a possibilidade de divulgação de atributos Torra and Navarro-Arribas (2023). Na Tabela 6.3 é possível observar a descoberta dos atributos para MDAV e Mondrian para $k = 3$ e $k = 5$, a qual é significativamente reduzida ao aplicar o valor de $k = 9$. Além disso, o DAS obteve desempenho superior aos outros métodos considerando o valor de $k = 9$. Quanto à perda de informação das técnicas, a Tabela 6.4 ilustra o eficiente desempenho do método DAS sob um vasto conjunto de dados, com um índice de perda de apenas 0,033, considerando os valores dinâmicos de k.

Tabela 6.3: Número de registros anonimizados vinculados corretamente aos 30.162 registros da base de dados original ADULT.

Adult	Intervalo fixo	MDAV	Mondrian	DAS
$k = 3$	26	3367	2200	-
$k = 5$	85	2219	1186	-
$k = 9$	447	1471	683	-
$idade : k = 8$ $educacao\_num : k = 5$ $horas\_por\_semana : k = 9$	-	-	-	111

O desempenho computacional do DAS é intrinsecamente ligado à complexidade

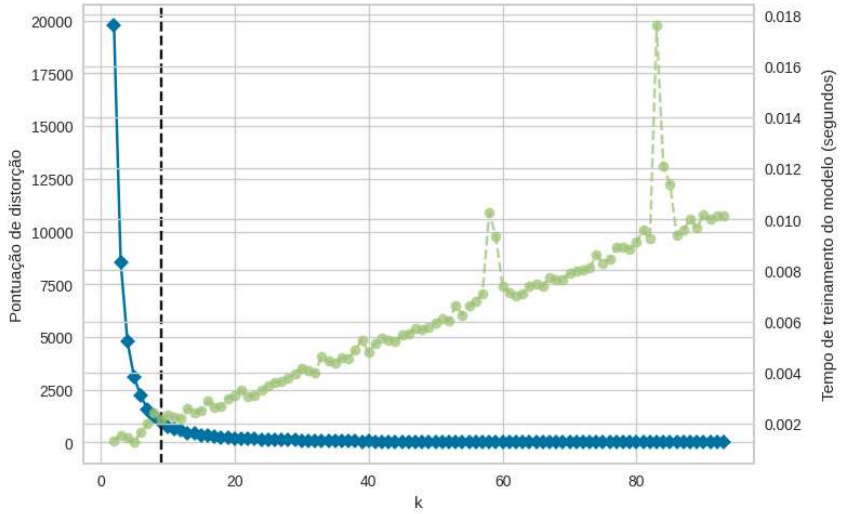


Figura 6.2: Definição do valor ideal de k usando Elbow para a base de dados ADULT, considerando o QI horas por semana.

Tabela 6.4: Perda de informação de dados originais para a base ADULT.

Adult	**Intervalo fixo**	**MDAV**	**Mondrian**	**DAS**
$k = 3$	0,318	0,008	0,021	-
$k = 5$	0,195	0,016	0,037	-
$k = 9$	0,092	0,028	0,056	-
$idade : k = 8$ $educacao\_num : k = 5$ $horas\_por\_semana : k = 9$	-	-	-	0,033

computacional do método de Elbow, etapa do algoritmo que possui maior custo. Entretanto, Elbow se destaca positivamente em relação aos demais métodos (*Gap Statistic*, *Silhouette Coefficient* e *Canopy*). O DAS utiliza a implementação do método de Elbow *KElbowVisualizer* da biblioteca yellowbrick, que aplica o algoritmo Kneedle para encontrar o valor ideal de k, com complexidade conhecida de $O(n^2)$ em função da quantidade de registros Satopaa et al. (2011). Como os atributos possuem características unidimensionais, a entrada para o método de Elbow é limitada aos valores únicos dos atributos, reduzindo significativamente a quantidade de entradas e o tempo de processamento.

O desempenho do método DAS na etapa de anonimização dos dados é obtido através da baixa complexidade computacional, combinando o custo do algoritmo Kneedle com o custo da função de ordenação, a qual possui limite superior $NlogN$, multiplicado pela quantidade de atributos. Em contrapartida, as soluções baseadas em k-anonimato pertencem à classe de problemas NP-difícil Onesimu et al. (2022). Um fator determinante para alcançar o exímio desempenho quanto ao ataque de divulgação de atributos é o fato de a proposta de anonimização dinâmica baseada

em separatrizes gerar agrupamentos, ou classes de equivalência desbalanceadas. Deste modo, é possível coibir efetivamente ataques de divulgação de atributos. Além da segurança proporcionada, a anonimização dos dados permite que estes sejam divulgados publicamente para abastecer a comunidade científica com dados equivalentes aos dados puros. Consequentemente, possibilita assim extrair estatísticas e processamentos precisos e equivalentes aos obtidos com dados puros, ideal quando consideramos um cenário IoHT.

6.2 GFKMC

Esta seção é responsável por descrever a avaliação geral do *framework* GFKMC. Ademais, são apresentados os cenários gerais de testes para investigar o impacto da anonimização em relação à privacidade, à utilidade e ao desempenho em modelos de aprendizado de máquina, tradicionais e federados. A seguir, expõem-se as especificações das bases, as métricas de avaliação e o conjunto de *hardwares* usados para executar os experimentos. Por fim, são discutidos todos os resultados.

6.2.1 Base de Dados

O *framework* GFKMC espera que os atributos QI numéricos e categóricos sejam fornecidos em uma estrutura de dados tabular potencialmente heterogênea com eixos rotulados (linhas e colunas) (índices e números). De modo a atender este requisito, assim como para avaliação anterior do DAS, opta-se por utilizar a base de dados ADULT Lhoest et al. (2021) para fornecer os insumos adequados para avaliar o GFKMC. Nesta nova análise, o banco de dados Adult também foi pré-processado para remover os atributos que continham valores ausentes. Deste modo, a base final considerada nesta análise é composta por 30.162 instâncias compostas por 14 atributos, sendo 7 deles numéricos e 7 categóricos.

A base de dados Adult possui destaque de excelência não só em cenários de anonimização baseada em k-anonimato Khan et al. (2020); Torra and Navarro-Arribas (2023), onde permite investigar o impacto da anonimização em relação à privacidade, à utilidade dos dados. Mas também pela vasta aplicação em tarefas de classificação Liu et al. (2021); Salmeron and Arévalo (2024), o que proporciona apurar o impacto da anonimização sobre o desempenho dos modelos. Em relação às tarefas para treinar modelos classificadores, o 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 esta avaliação, foram considerados os quase identificadores (sexo, idade, raça, estado civil, educação, país de origem, classe trabalhadora, ocupação) e o atributo sensível (classe salarial).

Para avaliar o desempenho dos métodos baseados em aprendizado federado, a base de dados ADULT utilizada é disponibilizada pela biblioteca *flwr_datasets*. Esta biblioteca interage diretamente com a plataforma Hugging Face Datasets Lhoest et al. (2021), onde a comunidade de aprendizado de máquina colabora em modelos, conjuntos de dados e aplicações.

6.2.2 Anonimização

O *framework* GFKMC proposto, um *framework* completo e prático para realizar a anonimização de atributos numéricos e categóricos, foi implementado em linguagem de programação Python, com suporte de bibliotecas como numpy, pandas, matplotlib e scipy. A análise do desempenho do GFKMC em relação à privacidade, utilidade dos dado foi aferido e comparado com os algoritmos tradicionais como, Mondrian, K-NN Clustering-Based Anonymization (CB) e Top-Down Greedy Anonymization (TDG) do repositório de código aberto[1] por Slijepčević et al. (2021), fornecendo uma compreensão abrangente do estado da arte. Os experimentos para avaliação do GFKMC foram conduzidos em uma máquina com 32 GB de RAM e um processador AMD Ryzen 5 PRO 5675U.

6.2.3 Modelos de Aprendizado de Máquina

A anonimização de dados é uma prática que garante a privacidade e proteção de dados ao modificar os dados originais para impedir a reidentificação de indivíduos. Entretanto, a anonimização pode impactar significativamente a qualidade e a utilidade desses dados, especialmente em aplicações que dependem de aprendizado de máquina. Dessa forma, torna-se essencial avaliar o impacto da anonimização na performance dos modelos classificadores, garantindo que as técnicas aplicadas mantenham um equilíbrio adequado entre privacidade e utilidade.

Aprendizado de Máquina Centralizado

Modelos de aprendizado de máquina centralizados dependem fortemente da qualidade dos dados de entrada para garantir um bom desempenho em tarefas como classificação, regressão e agrupamento. Quando os dados são anonimizados, a granularidade e a representatividade dos padrões presentes nos dados originais podem ser reduzidas. Essa perda de informação pode levar à degradação na precisão dos modelos, impactando diretamente as principais métricas como acurácia, precisão, etc. Portanto, avaliar empiricamente esse impacto é crucial para

[1] https://github.com/fhstp/k-AnonML

compreender até que ponto a anonimização pode ser aplicada sem comprometer a eficácia dos modelos de aprendizado de máquina.

A avaliação do impacto da anonimização em modelos de aprendizado de máquina centralizados é essencial para garantir que a privacidade dos indivíduos seja protegida sem comprometer a utilidade dos dados. Deste modo, para avaliar o impacto de dados anonimizados em modelos de aprendizado de máquina centralizados, foram considerados métodos populares. Entre esses métodos, incluem *k*-NN, XGBoost, SVM e RF. Essa análise permite aprimorar técnicas de anonimização, adaptar algoritmos de aprendizado e desenvolver metodologias robustas para lidar com desafios na interseção entre privacidade e ML.

Aprendizado Federado

Aplicações baseadas em aprendizado federado se destacam por sua capacidade de preservar a privacidade de dados em redes de computadores e sistemas distribuídos, garantindo que as informações sensíveis permaneçam localmente nos dispositivos enquanto permitem a construção de modelos globais eficientes. 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. No cenário de avaliação que compreende FL, 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)

Os modelos FL para Regressão Logística e XGBoost foram implementados na linguagem Python, com auxílio majoritário do *framework* Flower Beutel et al. (2020), o qual é uma estrutura específica para implementação, análise e avaliação de aplicações FL amigável. Outras bibliotecas Python como numpy, pandas, matplotlib e scipy também foram importantes para a implementação. O código-fonte está disponível no GitHub[2].

Cada cliente aplicou localmente uma das quatro técnicas de anonimização em seu subconjunto de dados: o *framework* GFKMC proposto no Capítulo 4, Mondrian LeFevre et al. (2006), CB e TDG Slijepčević et al. (2021). Essas técnicas variam em abordagem, desde particionamento recursivo (como no Mondrian) a 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

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

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.

6.2.4 Métricas

Privacidade

Analogamente à avaliação do DAS, a vinculação de registros Torra (2013); Domingo-Ferrer et al. (2022); Jiang and Torra (2023) é a métrica adotada para avaliar a vulnerabilidade em relação à divulgação de atributos. Esta tese 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 corresponder ao índice $r1$, ele será contado como uma correspondência. Para a avaliação do GFKMC é empregada a função de distância da Definição 1 para atributos numéricos. É importante enfatizar que, para aplicar esta definição para conjuntos de dados anonimizados onde atributos numéricos foram generalizados em intervalos, a média do intervalo é considerada. Em relação aos atributos categóricos, a função de

distância considerada é especificada na Definição 2.

Utilidade

A perda de informações dos algoritmos é avaliada pela métrica NCP Ghinita et al. (2007). A NCP é uma métrica amplamente aceita para avaliar a perda de informações em algoritmos de anonimização de dados. Para atributos numéricos, o cálculo NCP de uma classe de equivalência e é uma evolução do cálculo NCP para uma tabela anonimizada T^* 6.1, definido pela equação:

$$NCP_N(e) = \sum_{A_N \in T_N} \frac{max(e_{A_N}) - min(e_{A_N})}{max(T_N) - min(T_N)} \tag{6.2}$$

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

Para atributos categóricos, o NCP da classe de equivalência e é definido como:

$$NCP_C(e) = \sum_{A_C \in T_C} \begin{cases} 0, & |LCA(e_{A_C})| = 1 \\ \frac{|LCA(e_{A_C})|}{|T_C|}, & v_1 \neq v_2 \end{cases} \tag{6.3}$$

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 antecessor comum mais baixo 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. Com as equações 6.2 e 6.3, o NCP da tabela anonimizada T^* é definido por:

$$NCP(T^*) = \frac{\sum_{e \in T^*} |e| \cdot (NCP_N(e) + NCP_C(e))}{|T_A| \cdot |T|} \tag{6.4}$$

Desempenho

A avaliação de modelos de aprendizado de máquina depende diretamente da escolha de métricas adequadas para medir seu desempenho em diferentes cenários. Em modelos tradicionais de aprendizado de máquina, métricas como acurácia, precisão, recall e F1-Score são amplamente utilizadas para avaliar o equilíbrio entre a capacidade de previsão correta e os erros cometidos pelo modelo. Já no contexto do aprendizado federado, métricas como a Área Sob a Curva (AUC) e a função de perda (loss) são fundamentais para monitorar a estabilidade e a convergência do treinamento descentralizado. A análise combinada dessas métricas é essencial para

compreender o impacto das técnicas de anonimização sobre a qualidade dos modelos preditivos e sua capacidade de generalização.

Para os métodos tradicionais de aprendizado de máquina, as métricas de desempenho de acurácia, precisão, recall e F_1-Score foram computadas. A acurácia mede a proporção total de previsões corretas em relação ao total de amostras. A precisão avalia a proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo. É especialmente relevante quando os custos de falsos positivos são altos. Já o recall mede a capacidade do modelo de identificar corretamente todas as instâncias positivas dentro do conjunto de dados. É fundamental em cenários onde minimizar falsos negativos é crucial, como no diagnóstico médico. O F1-Score representa a média harmônica entre precisão e recall, sendo útil para balancear ambos os fatores e proporcionar uma métrica mais robusta em conjuntos de dados desbalanceados. Essas métricas são essenciais para entender como diferentes métodos de anonimização afetam o desempenho dos modelos tradicionais, garantindo que a privacidade dos dados não comprometa excessivamente a eficácia da análise preditiva.

No aprendizado federado, a avaliação do desempenho dos modelos ocorre em um cenário descentralizado, onde os dados permanecem distribuídos em diferentes dispositivos ou servidores. A avaliação de modelos de FL é um processo que objetiva determinar a precisão e a eficiência de um modelo classificador descentralizado. Ao avaliar um modelo FL, é importante considerar os efeitos da anonimização de dados empregada pelo modelo de privacidade *k*-anonimato de modo empírico. Neste cenário, os resultados das métricas de desempenho acurácia, Área Sob a Curva (AUC) e a perda (*Loss*) podem ser obtidos com auxílio do Flower *framework* Beutel et al. (2020). A ampla maioria dos trabalhos relacionados considera a utilização da acurácia como métrica de avaliação da degradação da eficiência do modelo, principalmente Kwatra and Torra (2021). Por outro lado, a Área AUC mede 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. Essa métrica é importante no aprendizado federado, pois permite avaliar a qualidade do modelo mesmo quando os dados locais apresentam distribuições diferentes (não-iid). Para modelos de regressão linear, a perda (*Loss*) representa a diferença entre as previsões do modelo e os valores reais, sendo um indicador fundamental da qualidade da otimização do modelo. No aprendizado federado, a perda pode ser afetada por fatores como heterogeneidade dos dados, ruído introduzido por anonimização e convergência do modelo global ao agregar atualizações de múltiplos clientes. O objetivo de treinar um modelo é

minimizar a perda, reduzindo-a ao seu menor valor possível.

6.2.5 Análises

A avaliação criteriosa das métricas de utilidade, privacidade e desempenho sobre o *framework* GFKMC proposto visa entender os efeitos da técnica de anonimização adotada em relação aos ataques externos à privacidade de dados sensíveis e à qualidade de classificação dos modelos ML especialmente em ambientes IoHT. Em relação à privacidade, a divulgação de atributos numéricos e categóricos é uma das principais causas de vazamento de informações. No entanto, é de grande importância manter o compromisso entre a privacidade dos dados e sua utilidade prática. Portanto, a análise da perda de informações é realizada usando a métrica NCP e a avaliação empírica do impacto nos métodos de classificação de aprendizado de máquina.

Utilidade

Com base nas métricas e experimentos conduzidos em cenário centralizado, os resultados obtidos na análise da perda de informações, para variações entre $k = 2, 3, \ldots, 100$, são ilustrados na Figura 6.3. Os resultados indicam que o *framework* GFKMC, com suas variações na generalização dos grupos, apresenta uma perda de informações essencialmente constante, em torno de 25%, para todos os valores de k. Em contraste, os algoritmos Mondrian, CB e TDG Slijepčević et al. (2021) mostram um aumento na perda de informações à medida que k aumenta. Especificamente, o algoritmo Mondrian tem uma perda de informação maior que o GFKMC para valores de k maiores que 15. Em contraste, os outros algoritmos (CB e TDG) mostram perdas maiores que o GFKMC quando k excede o valor de 40. A característica de perda de informação constante é proveniente da anonimização antecipada. Uma vez que a metodologia proposta generaliza os dados, a perda de informação do reagrupamento posterior para atender ao k-anonimato é mínima. Um fator essencial para a perda de informação relativamente baixa do GFKMC é a não generalização de atributos categóricos em casos em que a árvore de taxonomia possui altura igual a um. Ou seja, não suprimir informações é uma prioridade da proposta.

Ao repetir a análise da perda de informação para avaliar empiricamente o impacto da anonimização sobre o cenário federado, os resultados corroboram as informações obtidas no ambiente centralizado. A Figura 6.4 ilustra os valores médios e desvios padrão para os dez clientes do ambiente federado, considerando os principais níveis de k, sendo eles $3, 5, 10$ e 20 ilustrados na Figura 6.4. Considerando o cenário de FL, *framework* GFKMC demonstrou uma perda de informação

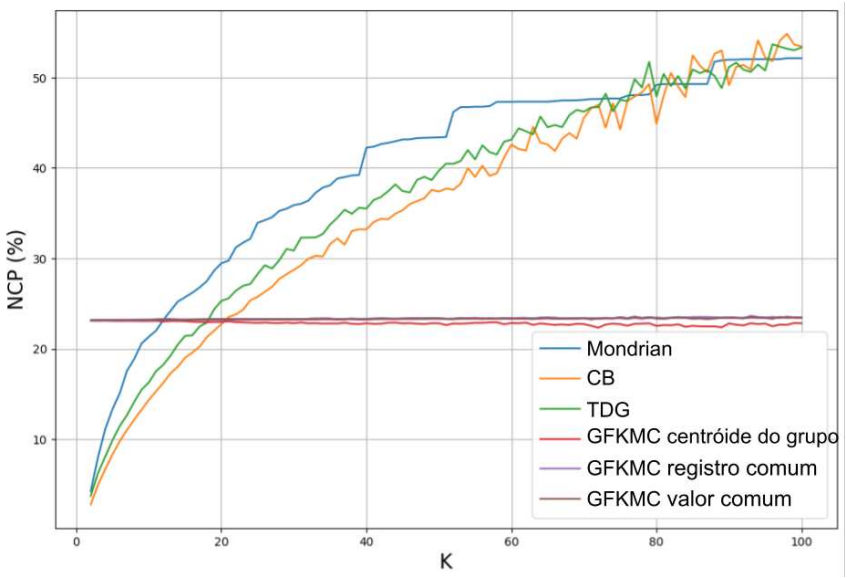


Figura 6.3: Análise de perda de informação em cenário centralizado.

essencialmente constante, em torno de 21%, independentemente 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.

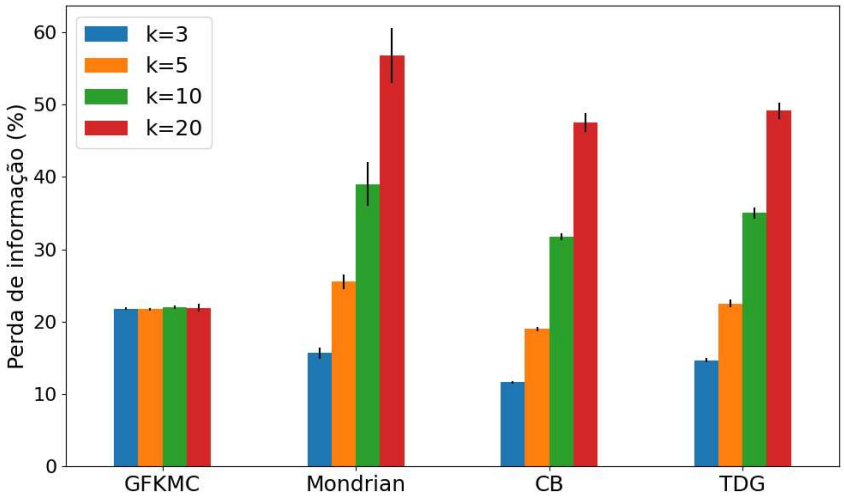


Figura 6.4: Análise de perda de informação em cenário de aprendizado federado.

Privacidade

Em relação à privacidade e ao risco de vinculação de registros no cenário centralizado, a Figura 6.5 ilustra o desempenho dos métodos de anonimização avaliados também para variações entre $k = 2, 20, \ldots, 100$. As variações do GFKMC se destacam em comparação aos outros algoritmos, como Mondrian, CB e TDG, especialmente para valores de k menores que 20. Isso se deve ao maior nível de generalização do GFKMC nesta faixa de valores. Para valores de k maiores que 20, o desempenho do GFKMC é semelhante aos outros algoritmos, mantendo assim um equilíbrio entre a proteção da privacidade e a preservação da utilidade dos dados.

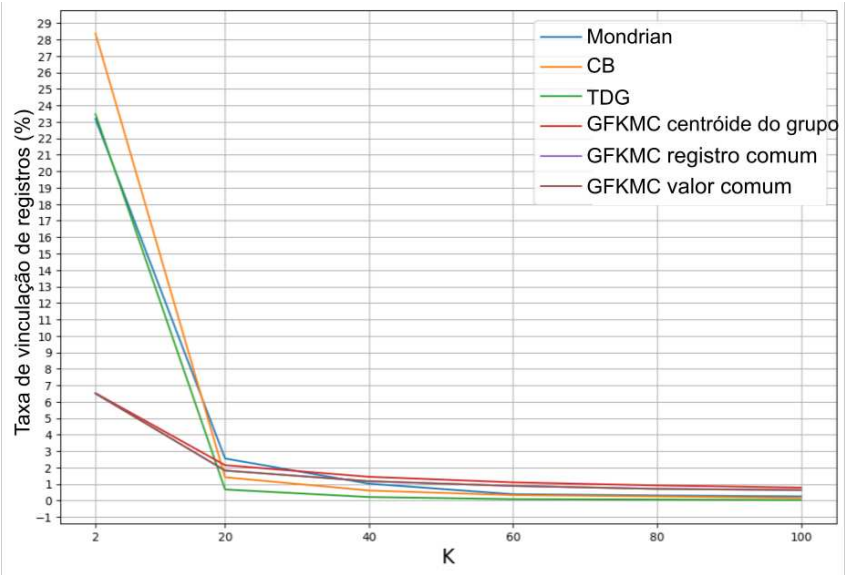


Figura 6.5: Análise para vinculação de registros (*Record linkage*) para cenário centralizado.

Ao avaliar o risco de reidentificação em um cenário federado, a Figura 6.6 reitera 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.

De modo geral, estes resultados destacam a dificuldade em se obter um bom compromisso entre perda de informação e risco de identificação, especialmente em cenários onde o tamanho de k é pequeno. No entanto, ao apresentar uma perda de informação constante, mantendo um nível aceitável de granularidade nos dados anonimizados, o *framework* de generalização antecipada do GFKMC contribui significativamente para garantir que a tomada de decisão sobre o valor ideal de k

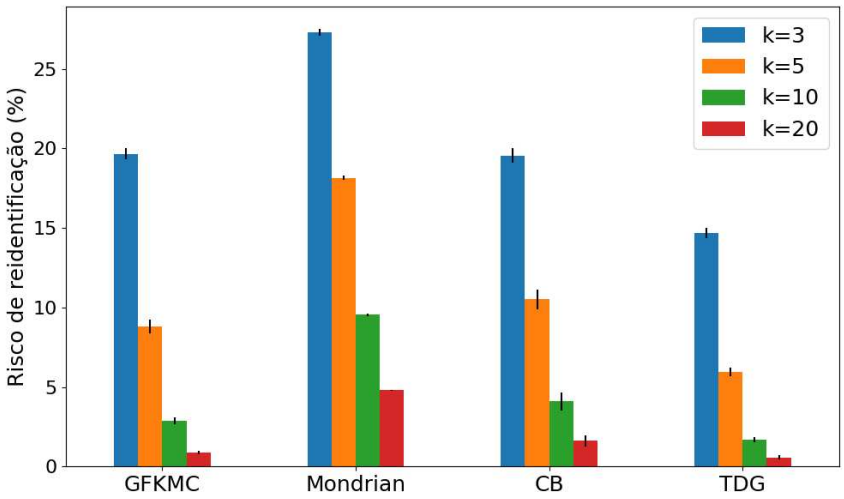


Figura 6.6: Risco de reidentificação

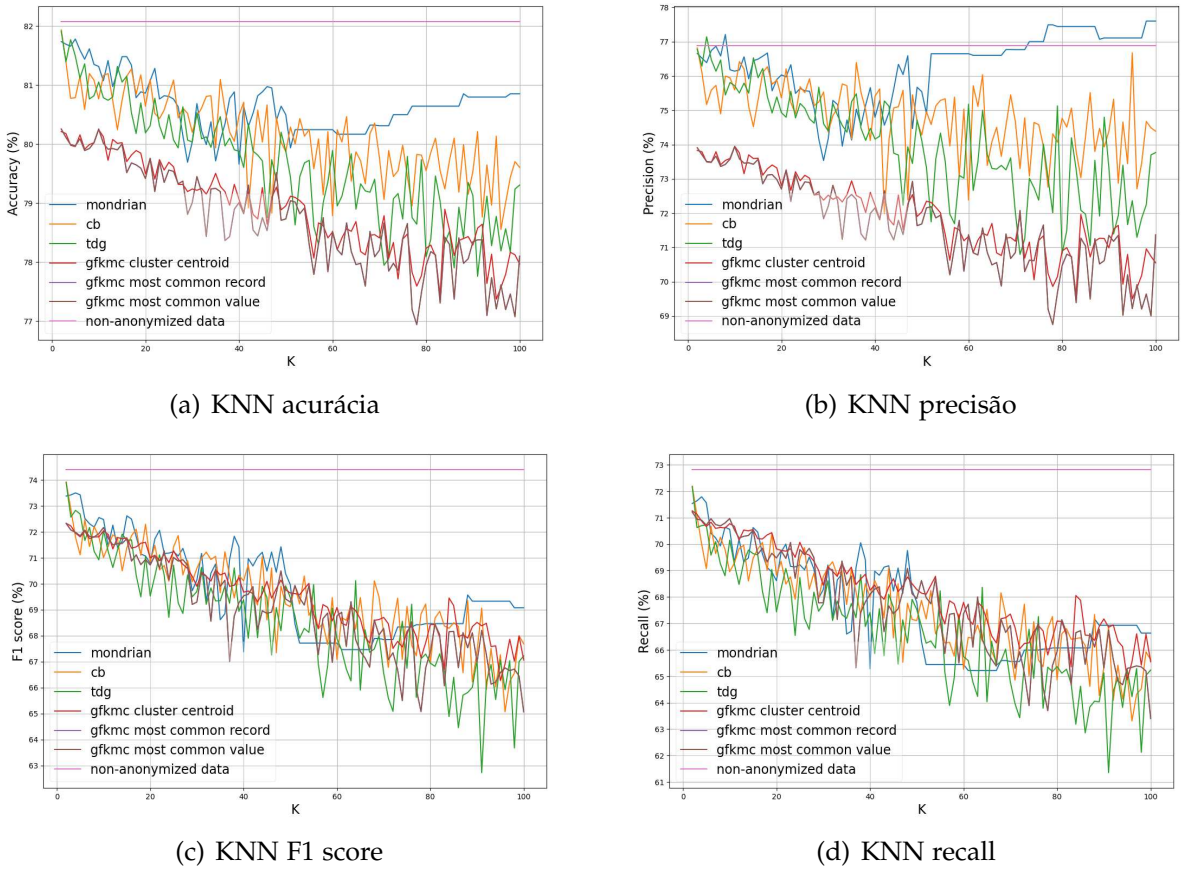
seja baseada apenas no risco de reidentificação, o que é crucial em aplicações sensíveis como assistência médica.

Desempenho

Os experimentos para avaliar o impacto do *framework* GFKMC e variações no desempenho de algoritmos de aprendizado de máquina centralizados demonstram que a perda constante de informação desse método tem implicações positivas para o desempenho dos modelos testados, especialmente em comparação com métodos como Mondrian, TDG e CB. Em relação ao modelo K-NN, a Figura 6.7 (a) mostra uma redução na precisão de menos de dois pontos percentuais em comparação com os dados originais quando o valor de k é baixo. Essa redução é comparável à observada nos outros métodos de anonimização. No entanto, à medida que os valores de k aumentam, o desempenho dos outros métodos de anonimização começa a cair significativamente. Por outro lado, as variações do GFKMC, exceto o valor mais comum, mantêm um comportamento de desempenho mais estável. Esse padrão se repete na métrica de precisão (Figura 6.7 (b)). No entanto, para F1-score (Figura 6.7 (c)) e recall (Figura 6.7 (d)), o GFKMC obtém desempenho equivalente aos outros métodos, indicando que o GFKMC preserva melhor a qualidade dos dados anonimizados em cenários de k mais altos.

A acurácia do GFKMC para o método RF é equivalente aos outros métodos para todos os valores de k. Entretanto, existe um ligeiro destaque para o Mondrian quando valores de k são altos, como ilustrado na Figura 6.8. Em relação à F1-score e Recall, as variações do GFKMC são mais eficientes perante o TDG e o CB. Entretanto, para a RF, o GFKMC desempenhou uma precisão ligeiramente inferior.

Já considerando o método SVM, o desempenho do GFKMC foi equivalente ao



(a) KNN acurácia

(b) KNN precisão

(c) KNN F1 score

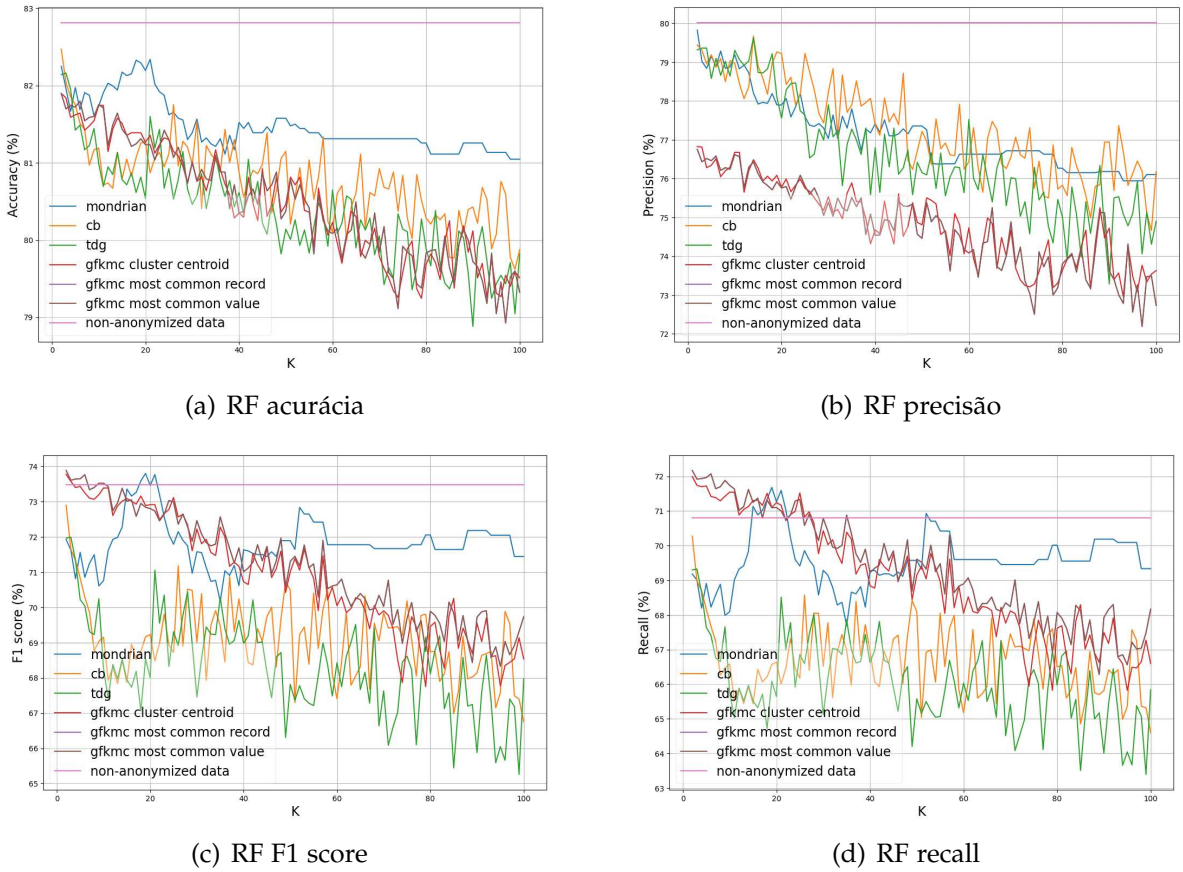
(d) KNN recall

Figura 6.7: Análise de desempenho para o método KNN.

apresentado pelos demais métodos em todas as métricas para valores de k até 50, como ilustrado na Figura 6.9. O desempenho ainda se equivale para todas as métricas para valores de k maiores que 50 em relação ao TDG e CB, com leve vantagem para Mondrian.

Por fim, para o XGBoost, considerando os resultados ilustrados pela Figura 6.10, o GFKMC apresentou um desempenho médio equivalente aos outros métodos para todos os valores de k. Portanto, isso reforça a robustez do GFKMC em manter a integridade de dados anonimizados em vários cenários. Todos esses resultados indicam que o GFKMC oferece uma vantagem significativa na manutenção do desempenho em modelos de aprendizado de máquina, especialmente em contextos onde é necessário anonimizar dados sem comprometer a utilidade analítica.

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) RF acurácia

(b) RF precisão

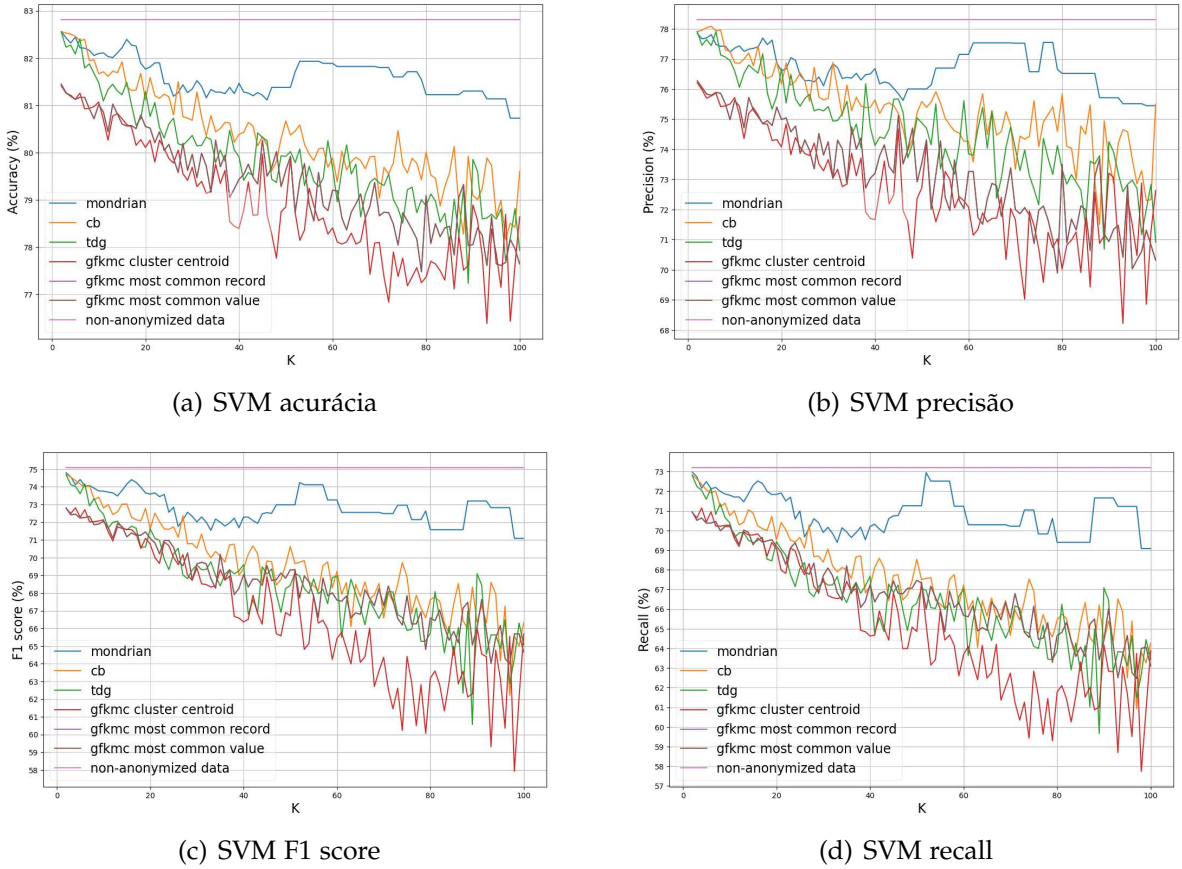
(c) RF F1 score

(d) RF recall

Figura 6.8: Análise de desempenho para o método RF.

A Figura 6.11 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 6.12 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 6.13 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 6.14 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) SVM acurácia

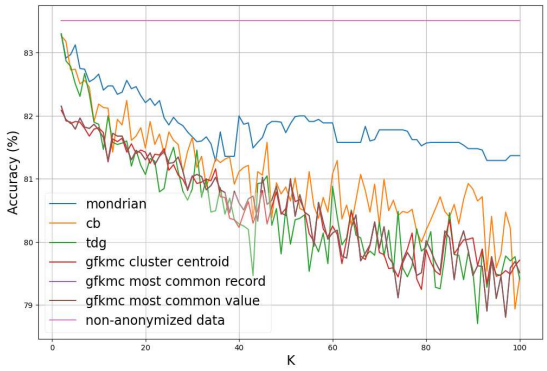
(b) SVM precisão

(c) SVM F1 score

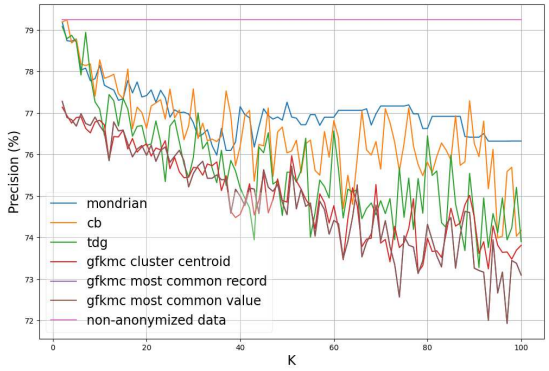
(d) SVM recall

Figura 6.9: Análise de desempenho para o método SVM.

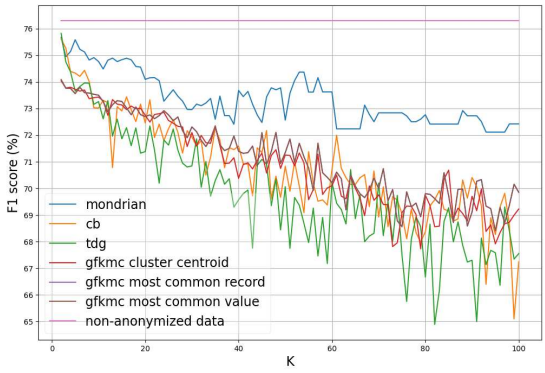
A avaliação criteriosa das métricas de desempenho em modelos de aprendizado de máquina tradicionais e federados é essencial para entender os efeitos das técnicas de anonimização sobre a qualidade preditiva dos modelos. No aprendizado tradicional, a análise de acurácia, precisão, recall e F1-Score permite verificar a eficácia dos modelos em diferentes cenários, especialmente em problemas de classificação desbalanceada. Já no aprendizado federado, métricas como AUC e função de perda são fundamentais para avaliar a robustez do treinamento descentralizado e a estabilidade da convergência. Dessa forma, a escolha adequada das métricas permite inferir de forma empírica o melhor equilíbrio entre privacidade e utilidade dos dados, assegurando que os modelos preservem sua capacidade de classificação mesmo em cenários com dados anonimizados.



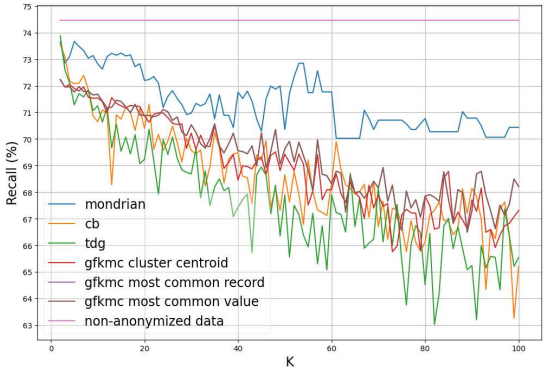
(a) XGB acurácia



(b) XGB precisão

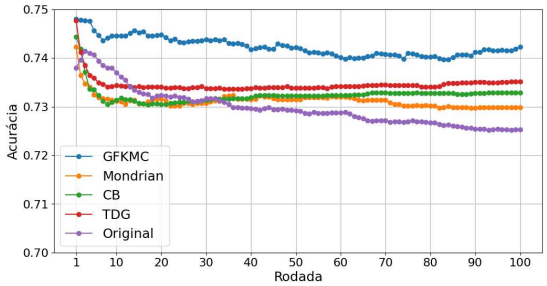


(c) XGB F1 score

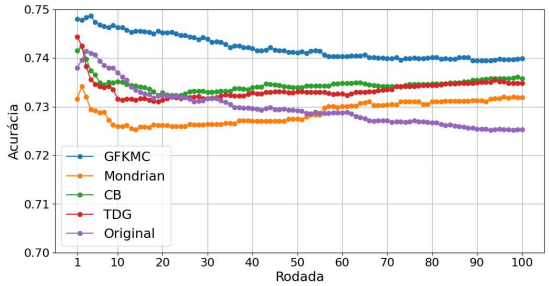


(d) XGB recall

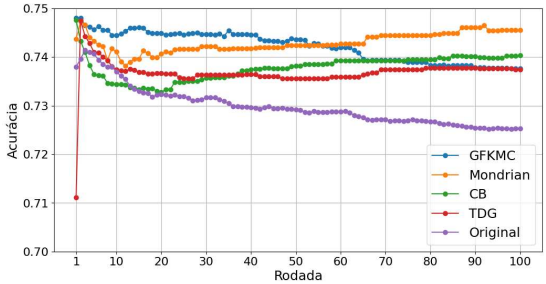
Figura 6.10: Análise de desempenho para o método XGB.



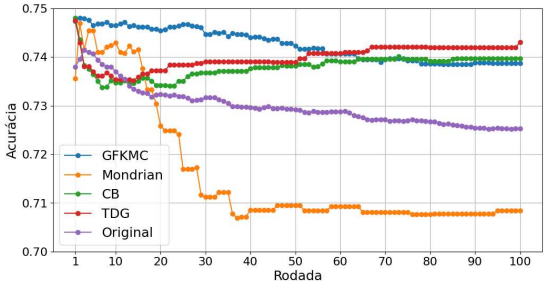
(a) k = 3



(b) k = 5

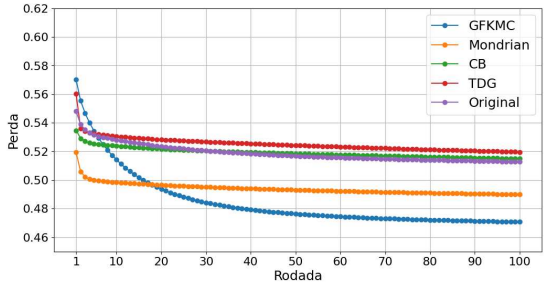


(c) k = 10

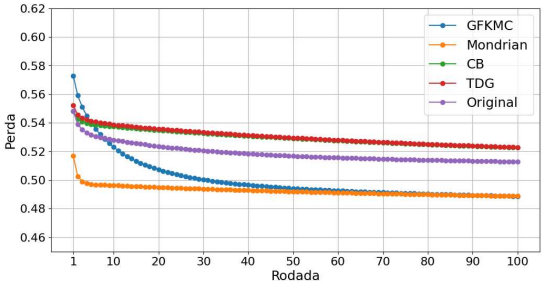


(d) k = 20

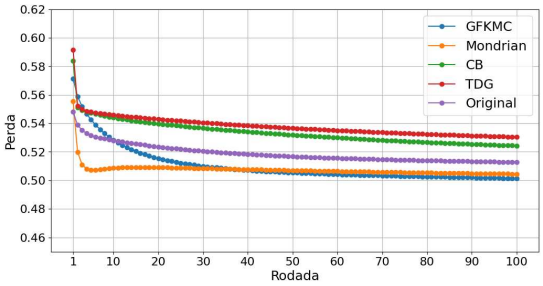
Figura 6.11: Acurácia do modelo Regressão logística para múltiplos métodos de anonimização



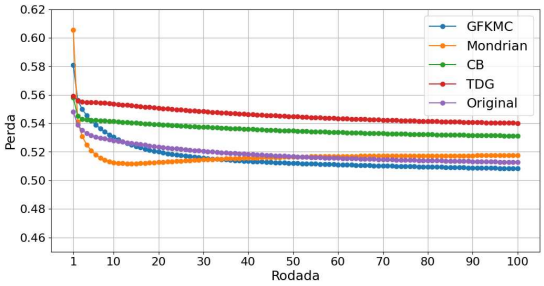
(a) k = 3



(b) k = 5

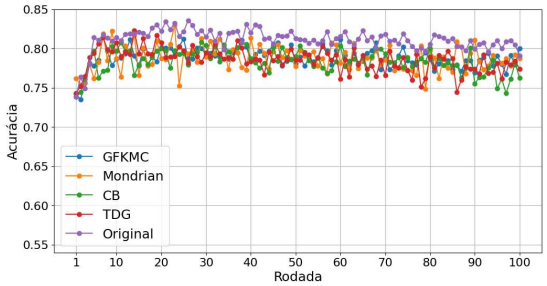


(c) k = 10

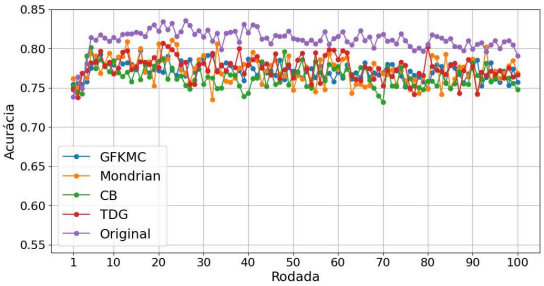


(d) k = 20

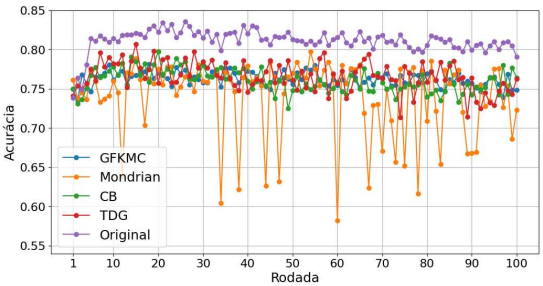
Figura 6.12: Perda do modelo regressão logística para múltiplos métodos de anonimização



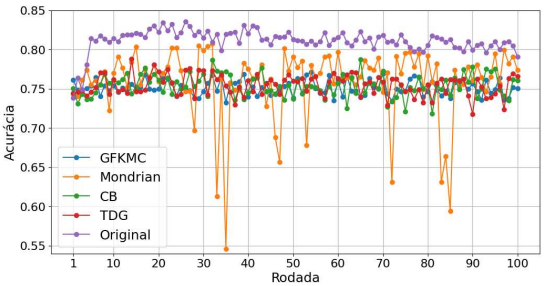
(a) k = 3



(b) k = 5

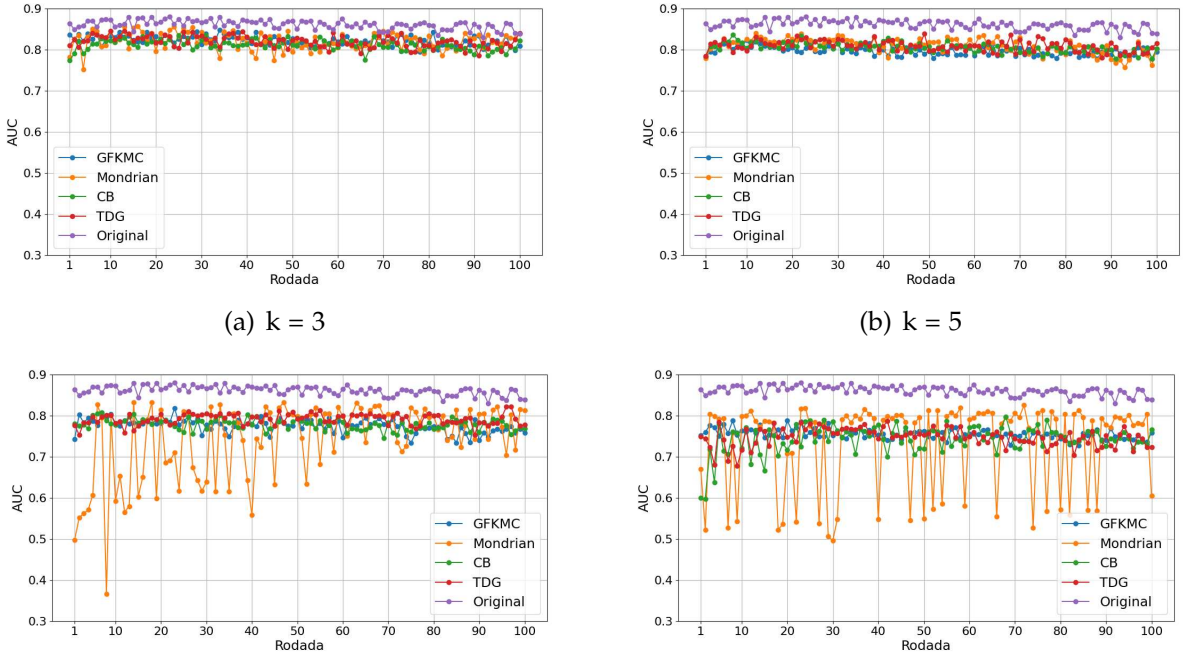


(c) k = 10



(d) k = 20

Figura 6.13: Acurácia do modelo XGBoost para múltiplos métodos de anonimização



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

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

Figura 6.14: AUC do modelo XGBoost para múltiplos métodos de anonimização

6.3 Resumo

Este capítulo contemplou a apresentação e discussão dos resultados experimentais obtidos a partir da aplicação do *framework* GFKMC sobre dados privados. Inicialmente, avalia-se isoladamente o método DAS, destacando seu desempenho na anonimização de atributos numéricos. Além disso, é considerado o impacto do DAS na preservação da privacidade e na minimização da perda de informação. Posteriormente, avalia-se o desempenho do *framework* GFKMC em sua totalidade, considerando sua aplicação em bases de dados de saúde e seu efeito sobre o desempenho de modelos de aprendizado de máquina, tanto em ambientes tradicionais quanto em aprendizado federado. Os resultados demonstram que a abordagem proposta apresenta desempenho superior em relação aos métodos de referência, como Mondrian, CB e TDG, destacando-se pela capacidade de proporcionar menor perda de informação e maior robustez na mitigação do risco de reidentificação. Além disso, constata-se que a aplicação do GFKMC exerce impacto mínimo sobre as métricas de desempenho dos modelos de aprendizado de máquina, evidenciando sua aplicabilidade prática em ambientes que demandam elevado grau de privacidade. Por fim, o Capítulo 7 consolida as conclusões deste trabalho, sintetiza as principais contribuições e limitações e sugere perspectivas para pesquisas futuras no domínio da anonimização de dados privados.

Capítulo 7

Conclusões

O compartilhamento de dados desempenha um papel essencial na promoção do avanço científico em diversas áreas, especialmente na área da saúde, onde a análise de grandes volumes de dados pode contribuir significativamente para a melhoria da qualidade dos serviços, da acurácia diagnóstica e da eficácia terapêutica. No entanto, preocupações relacionadas à privacidade e à exposição indevida de informações privadas dificultam a colaboração entre instituições. Deste modo, a adoção de mecanismos robustos de anonimização torna-se um requisito indispensável para viabilizar o compartilhamento seguro desses dados. Diante desse cenário, diversas técnicas de anonimização têm sido propostas, com o objetivo de mitigar os riscos associados à reidentificação, ao mesmo tempo em que buscam preservar a utilidade dos dados para fins analíticos e científicos.

Este trabalho teve como objetivo principal o desenvolvimento de um *framework* para anonimização de dados, denominado *Generalization First k-Member Clustering*. O GFKMC é capaz de assegurar elevados níveis de privacidade para dados privados, especialmente no domínio da Internet das Coisas aplicada à Saúde. A proposta integra mecanismos inovadores, como a Anonimização Dinâmica por Separatrizes para atributos numéricos e generalização hierárquica antecipada para atributos categóricos, permitindo balancear adequadamente o compromisso entre privacidade e utilidade dos dados.

O módulo DAS, responsável pela anonimização dos atributos quase identificadores numéricos, demonstrou elevada capacidade em garantir privacidade, mantendo, simultaneamente, a fidelidade estatística dos dados originais. Isso foi possível por meio da identificação dinâmica da configuração ótima de agrupamento, aplicada de maneira individualizada a cada atributo, o que resultou em ganhos significativos em termos de segurança e utilidade dos dados anonimizados. Complementarmente, o *framework* GFKMC, em sua totalidade, destacou-se por proporcionar uma perda de informação constante para os tamanhos de k considerados, característica determinante para assegurar a viabilidade dos dados anonimizados em análises estatísticas e em modelos de aprendizado de máquina.

Os resultados experimentais demonstraram que o GFKMC supera metodologias

tradicionais, como Mondrian, CB e TDG, tanto na preservação da utilidade dos dados quanto na mitigação dos riscos de reidentificação. Além disso, constatou-se que o *framework* apresenta um comportamento robusto em ambientes de aprendizado de máquina centralizado e federado. Em ambos os cenários, os modelos ML mantêm a estabilidade frente às transformações dos dados impostas pela anonimização.

Particularmente, no contexto do aprendizado federado, o GFKMC demonstrou-se capaz de reduzir os impactos negativos que elevados níveis de anonimização normalmente exercem sobre o desempenho dos modelos distribuídos. Tal característica adquire especial relevância diante da crescente adoção de arquiteturas descentralizadas em cenários sensíveis à privacidade, como na saúde digital, onde dados permanecem localmente nos dispositivos.

A continuidade desta pesquisa poderá seguir múltiplas direções, entre elas, a mais trivial é integrar extensão para garantir os princípios da l-Diversidade e t-proximidade. Embora o *framework* GFKMC atenda satisfatoriamente aos princípios do k-anonimato, sabe-se que tal abordagem não é imune a ataques de homogeneidade ou de ligação, especialmente quando os grupos formados apresentam pouca diversidade nos atributos sensíveis. Assim, uma evolução natural consiste na incorporação dos princípios de l-diversidade e t-proximidade. Estes modelos de privacidade têm o objetivo de assegurar que os atributos sensíveis dentro de cada grupo sejam suficientemente diversos ou estatisticamente próximos da distribuição global dos dados. O desenvolvimento de algoritmos que satisfaçam simultaneamente essas propriedades, de forma eficiente do ponto de vista computacional e sem comprometer a utilidade dos dados, constitui um desafio relevante e uma contribuição significativa para o estado da arte.

Atualmente, o GFKMC foi projetado especificamente para dados estruturados em formato tabular. Contudo, em cenários típicos da IoHT, é comum a presença de dados não tabulares, como imagens médicas, sinais biomédicos (ECG, PPG, entre outros) e textos clínicos. A expansão do *framework* para esses domínios representa uma linha de pesquisa promissora, a qual demandará a definição de métricas de similaridade adequadas para espaços vetoriais complexos, além da adaptação dos conceitos de generalização e agrupamento. Trata-se de um desafio metodológico e teórico de elevada complexidade, cuja superação ampliará substancialmente o escopo de aplicabilidade da proposta.

Ainda considerando a natureza dinâmica dos sistemas IoT e IoHT, uma direção relevante para trabalhos futuros consiste na adaptação do *framework* para operar em fluxos contínuos de dados (*streaming*). Essa evolução permitiria a anonimização em tempo real, à medida que os dados são gerados, mantendo os requisitos de privacidade previamente estabelecidos. Tal adaptação impõe desafios associados à baixa latência, à escalabilidade e à necessidade de atualização incremental dos

modelos de agrupamento e generalização, para garantir a conformidade com os princípios do *k*-anonimato, e potencialmente, da *l*-diversidade e *t*-proximidade.

Embora o *k*-anonimato proporcione garantias de anonimização na indistinguibilidade dentro de grupos, ele não oferece proteção contra ataques baseados em conhecimento externo arbitrário. Assim, a integração do GFKMC com os princípios da privacidade diferencial surge como uma abordagem complementar, capaz de oferecer garantias probabilísticas robustas. Essa integração requer a definição de mecanismos de perturbação ou adição de ruído que sejam compatíveis com os dados anonimizados por generalização, além da modelagem do impacto desses mecanismos na utilidade dos dados e na viabilidade computacional.

Apesar de o aprendizado federado mitigar os riscos associados ao compartilhamento direto de dados, ele permanece vulnerável a ataques realizados por participantes maliciosos, tais como inferência de propriedades, reconstrução de dados e envenenamento de modelos. Nesse contexto, uma linha de pesquisa altamente promissora consiste em aprimorar o GFKMC para que ele ofereça resistência aprimorada contra esses vetores de ataque, por meio de técnicas de detecção de anomalias em redes federadas. Tal aprimoramento é permitiria assegurar a robustez dos sistemas em ambientes descentralizados de alta criticidade, como redes IoHT.

Referências Bibliográficas

Abbas, S. R., Abbas, Z., Zahir, A., and Lee, S. W. (2024). Federated learning in smart healthcare: a comprehensive review on privacy, security, and predictive analytics with iot integration. In *Healthcare*, volume 12, page 2587. MDPI.

Abouelmehdi, K., Beni-Hessane, A., and Khaloufi, H. (2018). Big healthcare data: preserving security and privacy. *Journal of big data*, 5(1):1–18.

Alexander, C. and Wang, L. (2025). Cybersecurity benefits and challenges of advanced healthcare technologies. *J Inform Techn Int*, 2(1):104.

Almuqati, M. T., Sidi, F., Mohd Rum, S. N., Zolkepli, M., and Ishak, I. (2024). Challenges in supervised and unsupervised learning: A comprehensive overview. *International Journal on Advanced Science, Engineering & Information Technology*, 14(4).

Arava, K. and Lingamgunta, S. (2020). Adaptive k-anonymity approach for privacy preserving in cloud. *Arabian Journal for Science and Engineering*, 45(4):2425–2432.

Asif, S., Wenhui, Y., ur Rehman, S., ul ain, Q., Amjad, K., Yueyang, Y., Jinhai, S., and Awais, M. (2025). Advancements and prospects of machine learning in medical diagnostics: unveiling the future of diagnostic precision. *Archives of Computational Methods in Engineering*, 32(2):853–883.

Ayala-Rivera, V., McDonagh, P., Cerqueus, T., Murphy, L., et al. (2014). A systematic comparison and evaluation of k-anonymization algorithms for practitioners. *Trans. Data Priv.*, 7(3):337–370.

Bache, K. and Lichman, M. (2013). UCI machine learning repository.

Batko, K. and Slezak, A. (2022). The use of big data analytics in healthcare. *Journal of big Data*, 9(1):3.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Beltrán, E. T. M., Pérez, M. Q., Sánchez, P. M. S., Bernal, S. L., Bovet, G., Pérez, M. G., Pérez, G. M., and Celdrán, A. H. (2023). Decentralized federated learning: Fundamentals, state of the art, frameworks, trends, and challenges. *IEEE Communications Surveys & Tutorials*, 25(4):2983–3013.

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*.

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

Binson, V., Thomas, S., Subramoniam, M., Arun, J., Naveen, S., and Madhu, S. (2024). A review of machine learning algorithms for biomedical applications. *Annals of Biomedical Engineering*, 52(5):1159–1183.

Byun, J.-W., Kamra, A., Bertino, E., and Li, N. (2007). Efficient k-anonymization using clustering techniques. In *International Conference on Database Systems for Advanced Applications*, pages 188–200. Springer.

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*.

Christen, P., Ranbaduge, T., and Schnell, R. (2020). Linking sensitive data. *Methods and techniques for practical privacy-preserving information sharing. Cham: Springer*.

Ciriani, V., Di Vimercati, S. D. C., Foresti, S., and Samarati, P. (2008). k-anonymous data mining: A survey. *Privacy-Preserving Data Mining: Models and Algorithms*, pages 105–136.

Coelho, K., Damião, D., Noubir, G., Borges, A., Nogueira, M., and Nacif, J. (2019). Cryptographic algorithms in wearable communications: An empirical analysis. *IEEE Communications Letters*, 23(11):1931–1934.

Coelho, K., Okuyama, M., Nogueira, M., Vieira, A., Silva, E., and Nacif, J. (2024a). Uma abordagem dinâmica para anonimização de dados de saúde por separatrizes. In *Anais do XLII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 826–839, Porto Alegre, RS, Brasil. SBC.

Coelho, K., Okuyama, M., Nogueira, M., Vieira, A., Silva, E., and Nacif, J. (2025a). Metodologia para avaliação da anonimização baseada em k-anonimato nos modelos de aprendizado de máquina. In *Anais do XLIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 742–755, Porto Alegre, RS, Brasil. SBC.

Coelho, K. K. (2020). *Um sistema para garantir a segurança de informações médicas em redes corporais*. PhD thesis, Universidade Federal de Viçosa.

Coelho, K. K., Nogueira, M., Marim, M. C., Silva, E. F., Vieira, A. B., and Nacif, J. A. M. (2022). Lorena: Low memory symmetric-key generation method for based on group cryptography protocol applied to the internet of healthcare things. *IEEE Access*, 10:12564–12579.

Coelho, K. K., Nogueira, M., Vieira, A. B., Silva, E. F., and Nacif, J. A. M. (2023a). A survey on federated learning for security and privacy in healthcare applications. *Computer Communications*, 207:113–127.

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

Coelho, K. K., Okuyama, M. M., Nogueira, M., Vieira, A. B., Silva, E. F., and Nacif, J. A. M. (2025b). Methodology for evaluating k-anonymity-based anonymization in machine learning models. In *2025 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7. IEEE.

Coelho, K. K., Tristão, E. T., Nogueira, M., Vieira, A. B., and Nacif, J. A. (2023b). Multimodal biometric authentication method by federated learning. *Biomedical Signal Processing and Control*, 85:105022.

Correa, S. (2003). Probabilidade e estatística.

Developers, N. (2024). numpy.percentile.

Din, I. U., Guizani, M., Hassan, S., Kim, B.-S., Khan, M. K., Atiquzzaman, M., and Ahmed, S. H. (2018). The internet of things: A review of enabled technologies and future challenges. *IEEE access*, 7:7606–7640.

Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002). Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and data Engineering*, 14(1):189–201.

Domingo-Ferrer, J., Sánchez, D., and Soria-Comas, J. (2016). *Database anonymization: privacy models, data utility, and microaggregation-based inter-model connections*. Morgan & Claypool Publishers.

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

Eichner, H., Ramage, D., Bonawitz, K., Huba, D., Santoro, T., McLarnon, B., Van Overveldt, T., Fallen, N., Kairouz, P., Cheu, A., et al. (2024). Confidential federated computations. *arXiv preprint arXiv:2404.10764*.

El Ouazzani, Z. and El Bakkali, H. (2018). A new technique ensuring privacy in big data: K-anonymity without prior value of the threshold k. *Procedia Computer Science*, 127:52–59. Proceedings Of The First International Conference On Intelligent Computing In Data Sciences, ICDS2017.

Farooqi, S. A., Abd Rahman, A., and Saad, A. (2024). Differential privacy based federated learning techniques in iomt: A review. In *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, pages 1–7. IEEE.

Fernandes, C., Sena, S., Coelho, K., Nogueira, M., Silva, E., Vieira, A., and Nacif, J. A. (2022). Avaliação de protocolos de acordo de chave baseados em sinais fisiológicos para redes corporais sem fio. In *SBESC 2022 ()*.

Franzen, D., Nuñez von Voigt, S., Sörries, P., Tschorsch, F., and Müller-Birn, C. (2022). Am i private and if so, how many? communicating privacy guarantees of differential privacy with risk communication formats. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pages 1125–1139.

Freitas, L. L. C., Coelho, K. K., Nogueira, M., Vieira, A. B., Nacif, J. A. M., and Silva, E. F. (2024a). Context-sensitive access control and zero trust for security in e-health. In *2024 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6.

Freitas, L. L. d. C., Coelho, K. K., Nogueira, M., Vieira, A. B., Nacif, J. A. M., and Silva, E. F. (2024b). Controle de acesso sensıvel ao contexto e zero trust para a segurança em e-health. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos SBRC*.

Fung, B. C., Wang, K., Fu, A. W.-C., and Yu, P. S. (2010). *Introduction to privacy-preserving data publishing: Concepts and techniques*. Chapman and Hall/CRC.

Garcia, L. R., Aguilera-Fernandes, E., Gonçalves, R. A. M., and Pereira-Barretto, M. R. (2020). *Lei Geral de Proteção de Dados (LGPD): guia de implantação*. Editora Blucher.

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.

He, X. and Zhang, S. (2023). Differential privacy with fine-grained provenance: Opportunities and challenges. *Data Engineering*, page 21.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., and De Wolf, P.-P. (2012). *Statistical disclosure control*. John Wiley & Sons.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365.

Jiang, L. and Torra, V. (2023). Data protection and multi-database data-driven models. *Future Internet*, 15(3).

Kara, B. C., Eyupoglu, C., and Karakus, O. (2025). (r, k, ε)-anonymization: Privacy-preserving data publishing algorithm based on multi-dimensional outlier detection, k-anonymity and ε-differential privacy. *IEEE Access*.

Karagiannis, S., Ntantogian, C., Magkos, E., Tsohou, A., and Ribeiro, L. L. (2024). Mastering data privacy: leveraging k-anonymity for robust health data sharing. *International Journal of Information Security*, 23(3):2189–2201.

Ketu, S. and Mishra, P. K. (2021). Internet of healthcare things: A contemporary survey. *Journal of Network and Computer Applications*, 192:103179.

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.

Khatir, R. A., Izadkhah, H., and Razmara, J. (2023). Designing a novel approach using a greedy and information-theoretic clustering-based algorithm for anonymizing microdata sets. *Entropy*, 25(12):1613.

Kodinariya, T. M., Makwana, P. R., et al. (2013). Review on determining number of cluster in k-means clustering. *International Journal*, 1(6):90–95.

Kolawole, O. O. (2024). Iot and ai-based remote patient monitoring for chronic disease management.

Kumari, S. and Prabha, C. (2025). Internet of medical things ecosystem: regulations, challenges of standards, security mechanisms and future perspectives. In *Blockchain and Digital Twin for Smart Healthcare*, pages 519–535. Elsevier.

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., and Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. In *22nd International Conference on Data Engineering (ICDE'06)*, pages 25–25.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2005a). Incognito: Efficient full-domain k-anonymity. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 49–60.

LeFevre, K., DeWitt, D. J., and Ramakrishnan, R. (2005b). Multidimensional k-anonymity. Technical report, University of Wisconsin-Madison Department of Computer Sciences.

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*.

Li, H. (2011). *Learning to Rank for Information Retrieval and Natural Language Processing, Second Edition*, volume 4. Springer Cham.

Liu, F. and Li, T. (2018). A clustering k-anonymity privacy-preserving method for wearable iot devices. *Security and Communication Networks*, 2018:1–8.

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.

Marim, M. C., Coelho, K. K., Vieira, A. B., Nacif, J. A. M., Nogueira, M., and Silva, E. F. (2023). Análise de desempenho de um esquema de acordo de chaves de conferência para iot. In *Anais Estendidos do XIII Simpósio Brasileiro de Engenharia de Sistemas Computacionais*, pages 67–70. SBC.

Monteiro, S., Oliveira, D., António, J., Sá, F., Wanzeller, C., Martins, P., and Abbasi, M. (2022). Data anonymization: techniques and models. In *International Conference on Marketing and Technologies*, pages 73–84. Springer.

Nogueira, M., Borges, L. F., de Neira, A. B., Albano, L., and Coelho, K. K. (2024). Ciência de dados aplicada à cibersegurança: Teoria e prática. *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais SBSEG*.

Nosowsky, R. and Giordano, T. J. (2006). The health insurance portability and accountability act of 1996 (hipaa) privacy rule: implications for clinical research. *Annu. Rev. Med.*, 57(1):575–590.

Olatunji, I. E., Rauch, J., Katzensteiner, M., and Khosla, M. (2022). A review of anonymization for healthcare data. *Big data*.

Olatunji, I. E., Rauch, J., Katzensteiner, M., and Khosla, M. (2024). A review of anonymization for healthcare data. *Big data*, 12(6):538–555.

Onesimu, J. A., Karthikeyan, J., Eunice, J., Pomplun, M., and Dang, H. (2022). Privacy preserving attribute-focused anonymization scheme for healthcare data publishing. *IEEE Access*, 10:86979–86997.

Paul, J. (2025). Real-time predictive health monitoring using ai-driven wearable sensors: Enhancing early detection and personalized interventions in chronic disease management. .

Personal Data Protection Commission, o. S. (2018). Guide to basic data anonymisation techniques.

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.

Ribeiro-Alves, M., Franco, C. M., et al. (2022). Manual prático de anonimização de dados de pesquisa com o r.

Routray, S. and Choudhary, P. K. (2025). Extreme gradient booster model (xgb). In *Leveraging Emerging Technologies and Analytics for Empowering Humanity, Vol. 1: International Conference Proceedings of LEAD-2024, IIM Shillong, December 10–12*, page 103. Springer Nature.

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.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a"kneedle"in a haystack: Detecting knee points in system behavior. In *2011 31st international conference on distributed computing systems workshops*, pages 166–171. IEEE.

Shahid, J., Ahmad, R., Kiani, A. K., Ahmad, T., Saeed, S., and Almuhaideb, A. M. (2022). Data protection and privacy of the internet of healthcare things (iohts). *Applied Sciences*, 12(4):1927.

Shamsinejad, E., Banirostam, T., Pedram, M. M., and Rahmani, A. M. (2025). A review of anonymization algorithms and methods in big data. *Annals of Data Science*, 12(1):253–279.

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.

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.

Sokas, D., Butkuvienė, M., Tamulevičiūtė-Prascienė, E., Beigienė, A., Kubilius, R., Petrėnas, A., and Paliakaitė, B. (2022). Wearable-based signals during physical exercises from patients with frailty after open-heart surgery. *PhysioNet*.

Surbhi, N. R. C. and Dahiya, N. (2024). of healthcare things (ioht): Critical analysis. *Trends in Mechatronics Systems: Industry 4. 0 Perspectives*, page 59.

Templ, M. (2008). Statistical disclosure control for microdata using the r-package sdcmicro. *Transactions on Data Privacy*, 1(2):67–85.

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

Torra, V. (2022). *A Guide to Data Privacy*. Springer.

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.

Tristão, E., Coelho, K., Silva, E., Vieira, A. B., Nogueira, M., and Nacif, J. A. (2022). Autenticação biométrica baseada em ppg e ecg utilizando aprendizado profundo. In *SBESC 2022*.

Tristão, E. T., Coelho, K. K., Menezes, C., Freitas, L., Nogueira, M., Vieira, A. B., Silva, E. F., and Nacif, J. A. M. (2025). Enhancing biometric security with multimodal eeg and ppg identification. In *2025 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–6. IEEE.

UFV, P. P.-G. S. S. (2022). Normas gerais de teses e dissertacoes. https://www.ppg.ufv.br/wp-content/uploads/2012/08/Normas-gerais-de-Teses-e-Dissertac%CC%A7o%CC%83es-1.pdf.

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.

Voigt, P. and Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676):10–5555.

Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., and Fu, A. W.-C. (2006). Utility-based anonymization using local recoding. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 785–790.

Yan, Y., Herman, E. A., Mahmood, A., Feng, T., and Xie, P. (2021). A weighted k-member clustering algorithm for k-anonymization. *Computing*, pages 1–23.

Yuan, C. and Yang, H. (2019). Research on k-value selection method of k-means clustering algorithm. *J*, 2(2):226–235.

Zuo, Z., Watson, M., Budgen, D., Hall, R., Kennelly, C., and Al Moubayed, N. (2021). Data anonymization for pervasive health care: systematic literature mapping study. *JMIR medical informatics*, 9(10):e29871.

Apêndice A

Resumo das Contribuições Científicas

Ainda durante o período vigente do curso de doutorado, além dos trabalhos que compõem esta tese Coelho et al. (2024a,b,c, 2025a,b), foram desenvolvidas diversas outras publicações em colaboração com docentes e discentes de graduação e pós-graduação. As publicações serão brevemente apresentadas em ordem cronológica. Entre estas, destaca-se o artigo ***LORENA: Low memORy symmEtric-key geNerAtion method for based on group cryptography protocol applied to the Internet of Healthcare Things*** Coelho et al. (2022), no qual é proposto um método de geração de chave simétrica de baixa exigência de memória (LORENA), fundamentado em protocolo de acordo de chave secreta de grupo, voltado para ambientes IoHT. Este trabalho foi publicado no periódico *IEEE Access*.

Adicionalmente, o trabalho intitulado **Avaliação de Protocolos de Acordo de Chave Baseados em Sinais Fisiológicos para Redes Corporais sem Fio** Fernandes et al. (2022) apresenta uma análise empírica sobre o desempenho e o consumo de recursos computacionais por protocolos de autenticação em redes corporais, tendo sido publicado nos anais do *XII Simpósio Brasileiro de Engenharia de Sistemas Computacionais* (SBESC). De forma complementar, o artigo **Autenticação Biométrica Baseada em PPG e ECG utilizando Aprendizado Profundo** Tristão et al. (2022) propõe um método de autenticação biométrica multimodal, baseado na combinação de duas redes neurais convolucionais em cascata, também publicado no *SBESC*.

Dentre as contribuições internacionais, destaca-se a publicação da revisão sistemática ***A Survey on Federated Learning for Security and Privacy in Healthcare Applications*** Coelho et al. (2023a), no periódico *Computer Communications*, a qual oferece uma análise abrangente sobre o estado da arte do aprendizado federado aplicado à segurança e privacidade em aplicações na área da saúde.

Outro trabalho de relevância internacional intitula-se ***Multimodal Biometric Authentication Method by Federated Learning*** Coelho et al. (2023b), publicado no periódico *Biomedical Signal Processing and Control*. Este estudo apresenta um método de autenticação biométrica multimodal que combina sinais de fotopletismografia e eletrocardiograma por meio de redes neurais convolucionais aplicadas em contexto de aprendizado federado, resultando em melhorias significativas na precisão da

identificação.

No âmbito nacional, destaca-se o artigo **Uma Análise de Desempenho de um Esquema de Acordo de Chaves de Conferência para IoT** Marim et al. (2023), que realiza uma análise de viabilidade computacional e propõe um protocolo de acordo de chaves unidirecional, baseado em equações quadráticas, aplicado a ambientes IoT. Este trabalho foi apresentado no *XIII Simpósio Brasileiro de Engenharia de Sistemas Computacionais* (SBESC).

Adicionalmente, os trabalhos intitulados **Controle de Acesso Sensível ao Contexto e Zero Trust para a Segurança em E-Health** Freitas et al. (2024b) e ***Context-Sensitive Access Control and Zero Trust for Security in E-Health*** Freitas et al. (2024a) propõem e avaliam um modelo de controle de acesso contextualizado, fundamentado no paradigma *Zero Trust*, aplicado ao aumento da segurança em ambientes de e-health. O primeiro trabalho foi publicado nos anais do *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos* (SBRC), enquanto o segundo foi publicado no *IEEE Symposium on Computers and Communications* (ISCC).

Também se destaca o capítulo de livro **Ciência de Dados Aplicada à Cibersegurança: Teoria e Prática** Nogueira et al. (2024), publicado nos anais do *Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais* (SBSEG). Este capítulo, estruturado como minicurso, apresenta de forma abrangente o pipeline de Ciência de Dados aplicado à Cibersegurança, abordando conceitos, metodologias e técnicas, com os seguintes objetivos: disseminar a cultura de uso de Ciência de Dados no contexto da segurança cibernética; demonstrar o potencial de técnicas de inteligência artificial e aprendizado de máquina na área; e fomentar colaborações interinstitucionais.

No ano de 2025, foi publicado no *IEEE Symposium on Computers and Communications* (ISCC) o artigo ***Enhancing Biometric Security with Multimodal EEG and PPG Identification*** Tristão et al. (2025), o qual investiga uma abordagem de autenticação biométrica multimodal baseada na combinação dos sinais de fotopletismografia e eletroencefalografia (EEG). O estudo adapta um modelo previamente validado, originalmente fundamentado em PPG e ECG, substituindo o componente de ECG por EEG, cuja performance é otimizada por meio de ajuste de hiperparâmetros, resultando em ganhos em robustez e precisão.

Expandindo o escopo de atuação, o trabalho intitulado **Comparação de Desempenho de Rede Mesh em Campo com os Padrões Wi-SUN FAN 1.0 (FSK) e FAN 1.1 (OFDM)** foi apresentado no evento *SENDI 2025*, considerado o maior evento da América Latina voltado para Distribuição de Energia. Este estudo tem enfoque industrial e avalia, por meio de experimentos realizados em campo, as melhorias proporcionadas pelo padrão Wi-SUN FAN 1.1 em relação ao FAN 1.0. Paralelamente, a análise também foi elaborada para impulsionar a automatização de

usinas eólicas. Deste modo, o artigo **Análise de Desempenho de Rede Mesh com Padrões Wi-SUN FAN 1.0 (FSK) e FAN 1.1 (OFDM) para Parques Eólicos** foi submetido ao evento *Brazil Windpower 2025*. Ainda explorando o padrão WI-SUN, o trabalho ***Field Performance Analysis of Wi-SUN FAN 1.0 and FAN 1.1 Mesh Networks Using the Okumura-Hata Propagation Model*** foi submetido ao *Symposium on Computing Systems Engineering (SBESC) 2025*.

Por fim, destaca-se que se encontra em desenvolvimento o trabalho intitulado ***A Brain-Heart Multimodal Biometric Dataset with Physical and Mental Stimulation***, cuja proposta consiste na construção de um conjunto de dados biométricos composto por sinais de PPG e EEG de 86 indivíduos. Durante a coleta, os participantes foram submetidos a uma série de estímulos físicos e mentais, objetivando proporcionar uma base robusta para estudos futuros em autenticação biométrica multimodal e análise de sinais fisiológicos.