EFETIVIDADE DA K-ANONYMITY EM COMPARAÇÃO COM OUTRAS TÉCNICAS DE ANONIMIZAÇÃO CONFORME A LEI GERAL DE PROTEÇÃO DE DADOS

EFFECTIVENESS OF K-ANONYMITY COMPARED TO OTHER ANONYMIZATION TECHNIQUES UNDER THE LEI GERAL DE PROTEÇÃO DE DADOS

Gabriel Ferreira dos Santos [1]
Luiz Fernando Nunes [2]

RESUMO

Nos últimos anos, o aumento significativo de vazamentos de dados pessoais reforçou a urgência na proteção da privacidade dos indivíduos. Grandes incidentes, como o vazamento de dados de 223 milhões de brasileiros em 2021, demonstram a necessidade de práticas rigorosas de anonimização de dados para mitigar o risco de reidentificação. Este trabalho avalia a técnica de anonimização k-anonymity, destacada pela Lei Geral de Proteção de Dados (LGPD), em comparação com outras técnicas, como perturbação e generalização, para verificar sua eficácia na proteção de dados pessoais. Com a métrica de Risco de Reidentificação Mensurada (RRM), são analisadas as taxas de reidentificação, e os resultados mostram que a k-anonymity, especialmente quando combinada com supressão e generalização, reduz efetivamente o risco de reidentificação. Em comparação com as técnicas de perturbação e generalização isoladas, a k-anonymity apresenta-se como a mais promissora para atender às exigências da LGPD e manter um nível elevado de anonimização. A pesquisa conclui que o uso combinado de técnicas pode oferecer maior proteção aos dados pessoais, mas destaca que a k-anonymity, embora eficaz, não é completamente invulnerável. Assim, reforça-se a necessidade de avanços contínuos e abordagens híbridas para que a anonimização atenda plenamente às exigências legais e aos desafios tecnológicos futuros.

Palavras-chave: k-anonymity; anonimização; LGPD; privacidade; proteção de dados.

ABSTRACT

In recent years, the significant increase in personal data leaks has emphasized the urgency of protecting individuals' privacy. Major incidents, such as the data leak of 223 million Brazilians in 2021, underscore the need for rigorous data anonymization practices to mitigate the risk of reidentification. This study evaluates the k-anonymity anonymization technique, highlighted by Brazil's General Data Protection Law (LGPD), in comparison with other techniques, such as perturbation and generalization, to verify

[1] Gabriel Santos discente do curso de Ciência da Computação do Centro Centro Universitário Filadélfia de Londrina (UniFil)
[2] Luiz Nunes docente do curso de Ciência da Computação do Centro Universitário Filadélfia de Londrina (UniFil)

its effectiveness in protecting personal data. Using the Measured Reidentification Risk (RRM) metric, reidentification rates were analyzed, and the results show that k-anonymity, especially when combined with suppression and generalization, effectively reduces the reidentification risk. Compared to the standalone techniques of perturbation and generalization, k-anonymity emerges as the most promising approach to meet LGPD requirements and maintain a high level of anonymization. The study concludes that the combined use of techniques can provide greater personal data protection but highlights that k-anonymity, although effective, is not entirely invulnerable. This finding reinforces the need for continuous advancements and hybrid approaches to ensure that anonymization fully meets legal requirements and future technological challenges.

Keywords: k-anonymity; anonymization; LGPD; privacy; data protection.

AGRADECIMENTOS

Agradeço principalmente ao meu pai e à minha mãe, por todo apoio e carinho ao longo da formação. Em cada etapa vocês estiveram ao meu lado, oferecendo suporte, especialmente nos momentos mais desafiadores. Vocês foram e sempre serão meu alicerce, minha inspiração e a motivação constante que me impulsiona a seguir em frente.

Ao meu orientador e professor, Luiz Fernando Nunes, que com paciência e dedicação me ajudou e guiou durante todo o processo de pesquisa. Sou profundamente grato a ele por suas contribuições valiosas e seu apoio constante em todo o percurso da pesquisa.

Aos colegas e amigos que conheci durante a da faculdade e que compartilharam comigo momentos de aprendizado, dúvidas e conquistas, assim como conversas leves e descontraídas que tornaram os desafios menos desgastantes e o período mais leve.

Aos professores e coordenadores do curso de ciência da computação da UniFil, pelo conhecimento transmitido, pela paciência e pelo compromisso em proporcionar uma formação de excelência. Cada aula foi essencial para meu desenvolvimento acadêmico e pessoal.

Por fim, agradeço a todos aqueles que direta ou indiretamente, estiveram presentes nesta trajetória, ajudando a tornar este sonho uma realidade.

1 INTRODUÇÃO

Nos últimos anos, o volume de dados pessoais na internet aumentou exponencialmente, levando à necessidade urgente de proteger as informações dos indivíduos, principalmente com o aumento de vazamentos de dados recentes, como demonstrado pelo Centro de Prevenção, Tratamento e Resposta a Incidentes Cibernéticos de Governo (CTIR Gov) [3]as estatísticas sobre vazamentos de dados somados no período de 2020 a 2023 foi quase 3 vezes menor do que a quantidade de vazamentos total que ocorreu em 2024.

Vale destacar também um dos maiores casos de exposição de informações que ocorreu em 2021, onde mais de 223 milhões de brasileiros[4] tiveram o seu CPF, nome, endereço e mais alguns outros dados vazados, demonstrando assim a necessidade da aplicação de técnicas de anonimização, que de acordo com Marques e Bernardino (2020), consiste em um processo que visa remover e ofuscar a identidade das informações, visando dificultar a identificação do indivíduo que é dono daquela informação.

A Lei Geral de Proteção de Dados (LGPD)[5], define diversas regras rigorosas para a proteção dos dados, principalmente no tocante à anonimização de dados, são definidas normas e técnicas adequadas, que as empresas e organizações devem adotar para garantir a privacidade dos indivíduos. Vale ressaltar a diferença entre os tipos de dados sensíveis e pessoais, onde dados pessoais envolvem informações que podem identificar uma pessoa natural, como nome, endereço e telefone, enquanto dados sensíveis incluem informações que revelam origem racial, étnica, religião, orientação sexual entre outros.

A anonimização de dados possui diversas técnicas, como a *k-anonymity*, que conforme descrito por Vimercati *et al*. (2023), visa anonimizar a informação,

[3] Estatísticas resultantes do trabalho de detecção, triagem, análise e resposta a incidentes cibernéticos. Disponível em: https://www.gov.br/ctir/pt-br/assuntos/ctir-gov-em-numeros. Acesso em: 19 set. 2024.

[4] Megavazamento de dados de 223 milhões de brasileiros: o que se sabe e o que falta saber. Disponível em: https://g1.globo.com/economia/tecnologia/noticia/2021/01/28/ vazamento-de-dados-de-223-milhoes-de-brasileiros-o-que-se-sabe-e-o-que-falta-saber.ghtml. Acesso em: 19 set. 2024.

[5] BRASIL. Lei n. 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 19 set. 2024.

garantindo que os indivíduos não sejam diferenciados dos dados de pelo menos outros k-1 indivíduos, dificultando assim a reidentificação dessa pessoa, utilizando características conjuntas de um grupo para dissolver a informação.

Entretanto, a k-anonymity enfrenta algumas críticas e desafios práticos, especialmente na utilização dos dados, conforme destacado por Marques e Bernardino (2020), que dependendo da aplicação da anonimização, pode ocorrer a perda da informação, o que acaba dificultando consideravelmente a utilização dos dados que passaram pela anonimização. Vale ressaltar também o que é descrito por Vimercati *et al*. (2023), onde que mesmo os dados estejam anonimizados pela k-anonymity, ao combinar as informações com outras colunas k-anônimas, pode ocorrer a facilidade de identificação do indivíduo através desse cruzamento de elementos distintos ou não no conjunto de dados.

O objetivo deste trabalho é avaliar a k-anonymity na anonimização de dados no quesito da LGPD em comparação com as técnicas de generalização e perturbação, principalmente na questão da proteção e privacidade do indivíduo presente no conjunto de dados. A avaliação das técnicas vai ser feita utilizando a métrica de risco de reidentificação mensurada(RRM), conforme descrito pela Autoridade Nacional de Proteção de Dados (ANPD)6.

2 ANONIMIZAÇÃO DE DADOS E O CONTEXTO LEGAL DA LGPD

Com o aumento dos casos de vazamentos de dados com diversas informações pessoais, a anonimização tem se mostrado uma técnica essencial para proteger a privacidade dos indivíduos donos da informação. Segundo Marques e Bernardino (2020), a anonimização envolve processos que visam transformar as informações dos indivíduos em dados não identificáveis, dificultando e impossibilitando que o elemento seja relacionado com o usuário real.

[6] Estudo técnico sobre anonimização de dados na LGPD: uma visão de processo baseado em risco e técnicas computacionais. Disponível em: https://www.gov.br/anpd/pt-br/ documentos-e-publicacoes/documentos-de-publicacoes/estudo_tecnico_sobre_anonimizacao_de_dados_na_lgpd_uma_visao_de_processo_baseado_em_risco_e_tecnicas_computacionais.pdf. Acesso em: 19 set. 2024.

No Brasil, a LGPD (2018) trouxe especificações para proteger as informações dos indivíduos, principalmente sobre a questão da anonimização de dados, estabelecendo diretrizes rigorosas sobre o tratamento de dados pessoais, que as organizações e empresas devem adotar. Embora a lei não trate diretamente cada tipo específico de anonimização, ela define algumas diretrizes que devem ser adotadas para garantir que os dados estejam de fato anonimizados, evitando a possibilidade de associação direta ou indireta a um indivíduo. Ela também estabelece, que para que um dado não seja considerado pessoal, a anonimização não pode ser revertida por meios próprios ou com algum tipo de esforço, demonstrando que as informações anonimizadas devem ter uma proteção robusta no conjunto de dados.

A anonimização de dados, conforme destacado por Ferreira (2024), é fundamental para cumprir as exigências que são impostas pela LGPD pois, segundo o autor, ela assegura que os dados quando estão anonimizados, não podem ser associados ao indivíduo dono daquela informação, mesmo se forem combinados com outras bases e informações. Isso é destacado no artigo 5º, inciso XI, da LGPD (2018), que define a anonimização como meios técnicos utilizados no tratamento dos dados, pelos quais se perde a possibilidade de associação direta ao sujeito ao passar pela anonimização.

No entanto, conforme é descrito por Carvalho (2024), a anonimização também apresenta alguns desafios técnicos e jurídicos, principalmente na questão da manutenção da eficácia da anonimização devido aos avanços tecnológicos que podem possibilitar a reidentificação da informação. Além disso, a autora ressalta que a LGPD, embora seja uma legislação robusta, não detalha todas as formas técnicas de como a anonimização deve ser implementada, deixando assim esse aspecto aberto para a interpretação da organização que é detentora do dado na aplicação da anonimização.

Vale lembrar o que é citado por Ferreira (2024), que destaca que a combinação de várias técnicas de anonimização pode ajudar a aumentar a segurança dos dados podendo ajudar com a diminuição da possibilidade de reidentificação do indivíduo nesse contexto. A autora Brasher (2018) exemplifica isso ao sugerir a adoção de múltiplos níveis de proteção que é utilizado pela *General Data Protection*

Regulation (GDPR)[7] na Europa com o intuito de reduzir reduzir a "linkabilidade "dos dados com os indivíduos criando assim uma camada extra de segurança nos dados anonimizados.

3 K-ANONYMITY

A k-anonymity consiste em uma técnica de anonimização que teve o seu conceito introduzido por Samarati e Sweeney (1998), onde foi estabelecido alguns fundamentos teóricos e práticos sobre a k-anonymity onde até o presente momento são referenciadas e citadas em pesquisas sobre essa técnica.

De acordo com Vimercati *et al.* (2023) a k-anonymity consiste na junção da supressão e da generalização de dados, com o objetivo de tornar os indivíduos de um conjunto de dados totalmente anônimos, sem que ocorra a identificação deles tornando os registros não relacionados a outros k-1 registros, onde informações como datas de nascimento, idades e Código de Endereçamento Postal (CEP) são alteradas, com o objetivo de garantir que os usuários não possam ser identificados diretamente.

Conforme descrito por Sweeney (2002),em um conjunto de dados que contenha o nome, CEP e a idade do usuário, ao aplicar a k-anonymity nesse conjunto, poderia ser realizada a supressão da coluna nome, removendo ou deixando um caractere padrão nessa coluna. No caso da idade e do CEP, seria aplicada a generalização, para deixar deixar os dados menos pessoais e mais relacionados a um grupo em comum, garantindo que os indivíduos não sejam identificados diretamente.

Segundo Vimercati *et al.* (2023), mesmo algumas técnicas de anonimização oferecendo uma camada robusta na proteção das informações, ela não é totalmente invulnerável, onde a possibilidade de reidentificação dos dados utilizando o cruzamento de informações de outras fontes é extremamente alta, principalmente se os valores de k utilizados forem baixos no momento da anonimização. Por outro lado, conforme declarado por Majeed e Lee (2020), dependendo do tipo da

[7] EUR-Lex. General Data Protection Regulation (GDPR). Disponível em: https://eur-lex.europa.eu/eli/reg/2016/679/oj. Acesso em: 26 out. 2024.

abordagem da anonimização, pode ocorrer a perda dos dados para utilizações futuras em eventuais consultas que possam ocorrer.

Na adaptação de exemplo da tabela 1 temos um conjunto de dados não anonimizados, onde contem algumas informações como o Nome, Idade, Cidade e Doença.

Tabela 1 – Tabela Não Anonimizada

Nome	Idade	Cidade	Doença
Renata Lima	46	Campinas - São Paulo	Asma
Felipe Mendes	57	Barretos - São Paulo	Asma
Júlia Martins	41	Santos - São Paulo	Asma
Vanessa Silva	42	Florianópolis - Santa Catarina	Hipertensão
Tiago Gomes	44	Londrina - Paraná	Hipertensão
Pedro Santos	32	Campo Grande - Mato Grosso do Sul	Epilepsia
Fernanda Almeida	37	Brasília - Distrito Federal	Epilepsia

Fonte: Adaptado de https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf. Acesso em: 26 out. 2024.

Ao aplicarmos a k-anonymity, conforme na tabela 2, em um conjunto de dados com o valor de k ≥ 2 a coluna nome seria suprimida, ocorrendo a remoção dela. A coluna idade seria generalizada e conteria a faixa etária, enquanto na coluna cidade também seria aplicada a generalização, apresentando apenas a região geográfica.

Tabela 2 – Tabela após a k-anonymity

Idade	Cidade	Doença
40-59	Sudeste	Asma
40-59	Sudeste	Asma
40-59	Sudeste	Asma
40-59	Sul	Hipertensão
40-59	Sul	Hipertensão
20-39	Centro-Oeste	Epilepsia
20-39	Centro-Oeste	Epilepsia

Fonte: Adaptado de: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf. Acesso em: 26 out. 2024.

4 PERTURBAÇÃO

A perturbação é uma técnica de anonimização que envolve a modificação das informações originais com a adição de ruídos aleatórios no conjunto de dados. Conforme destacado por Majeed e Lee (2020), a ideia central é substituir os valores originais por itens gerados sinteticamente de uma forma que o dado não seja relacionado com a pessoa real. Vale ressaltar que as informações estatísticas não podem se diferenciar muito do dado original, para manter a utilidade daquela informação em análises futuras.

No exemplo adaptado abaixo da tabela 3, temos um conjunto de dados contendo algumas informações fictícias que podem ser anonimizadas como na coluna Idade.

Tabela 3 – Tabela Não Anonimizada

Nome	Idade	Cidade	Doença
Carlos Oliveira	50	São Paulo - São Paulo	Diabetes
Renata Lima	46	Campinas - São Paulo	Asma
Felipe Mendes	57	Barretos - São Paulo	Asma
Júlia Martins	41	Santos - São Paulo	Asma
Ana Clara	28	Curitiba - Paraná	Ansiedade
Eduardo Costa	35	Porto Alegre - Rio Grande do Sul	Depressão
Vanessa Silva	42	Florianópolis - Santa Catarina	Hipertensão
Tiago Gomes	44	Londrina - Paraná	Hipertensão
Pedro Santos	32	Campo Grande - Mato Grosso do Sul	Esquizofrenia
Fernanda Almeida	37	Brasília - Distrito Federal	Epilepsia
Sofia Rocha	29	Barretos - São Paulo	Epilepsia

Fonte: Adaptado de: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/ Guide-to-Anonymisation_v1-(250118).pdf. Acesso em: 26 out. 2024.

Ao aplicar a perturbação na coluna Idade conforme na tabela 4, a informações da coluna idade teria a adição de um leve ruído, que alteraria o valor dos elementos contidos nesse conjunto de dados.

Tabela 4 – Tabela após a Perturbação

Nome	Idade	Endereço	Doença
Carlos Oliveira	53	São Paulo - São Paulo	Diabetes
Renata Lima	48	Campinas - São Paulo	Asma
Felipe Mendes	59	Barretos - São Paulo	Asma
Júlia Martins	44	Santos - São Paulo	Asma
Ana Clara	29	Curitiba - Paraná	Ansiedade
Eduardo Costa	38	Porto Alegre - Rio Grande do Sul	Depressão
Vanessa Silva	48	Florianópolis - Santa Catarina	Hipertensão
Tiago Gomes	46	Londrina - Paraná	Hipertensão
Pedro Santos	37	Campo Grande - Mato Grosso do Sul	Esquizofrenia
Fernanda Almeida	37	Brasília - Distrito Federal	Epilepsia
Sofia Rocha	32	Barretos - São Paulo	Epilepsia

Fonte: Adaptado de https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/ Guide-to-Anonymisation_v1-(250118).pdf. Acesso em: 26 out. 2024.

5 GENERALIZAÇÃO

A generalização é uma técnica que está presente dentro da k-anonymity e que nesse estudo foi analisada separadamente, ela envolve a substituição de valores que são específicos por categorias mais amplas. Conforme destacado por Vimercati *et al.* (2023), essa técnica transforma dados detalhados, como endereço completo e data de nascimento, em informações menos específicas, como uma região geográfica ou apenas o ano de nascimento. O objetivo dela é aumentar a dificuldade de identificação do indivíduo com informações mais amplas e menos detalhadas, que poderiam ajudar a achar o real dono daquele dado.

Na tabela 5, que contém um conjunto de dados, temos a coluna Idade e Cidade que poderiam ser generalizadas, com o intuito de anonimizar as informações desse conjunto de elementos.

Tabela 5 – Tabela Não Anonimizada

Nome	Idade	Cidade	Doença
Carlos Oliveira	50	São Paulo - São Paulo	Diabetes
Renata Lima	46	Campinas - São Paulo	Asma
Felipe Mendes	57	Barretos - São Paulo	Asma
Júlia Martins	41	Santos - São Paulo	Asma
Ana Clara	28	Curitiba - Paraná	Ansiedade
Eduardo Costa	35	Porto Alegre - Rio Grande do Sul	Depressão
Vanessa Silva	42	Florianópolis - Santa Catarina	Hipertensão
Tiago Gomes	44	Londrina - Paraná	Hipertensão
Pedro Santos	32	Campo Grande - Mato Grosso do Sul	Esquizofrenia
Fernanda Almeida	37	Brasília - Distrito Federal	Epilepsia
Sofia Rocha	29	Barretos - São Paulo	Epilepsia

Fonte: Adaptado de: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/ Guide-to-Anonymisation_v1-(250118).pdf. Acesso em: 26 out. 2024.

Ao aplicar a generalização na coluna a idade as informações seriam modificadas para algo mais abrangente como a faixa etária e a coluna cidade para regiões ou agrupamentos geográficos menos específicos.

Tabela 6 – Tabela após a Generalização

Nome	Idade	Cidade	Doença
Carlos Oliveira	50-59	Sudeste	Diabetes
Renata Lima	40-49	Sudeste	Asma
Felipe Mendes	50-59	Sudeste	Asma
Júlia Martins	40-49	Sudeste	Asma
Ana Clara	20-29	Sul	Ansiedade
Eduardo Costa	30-39	Sul	Depressão
Vanessa Silva	40-49	Sul	Hipertensão
Tiago Gomes	40-49	Sul	Hipertensão
Pedro Santos	30-39	Centro-Oeste	Esquizofrenia
Fernanda Almeida	30-39	Centro-Oeste	Epilepsia
Sofia Rocha	20-29	Sudeste	Epilepsia

Fonte: Adaptado de https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/ Guide-to-Anonymisation_v1-(250118).pdf. Acesso em 26 de outubro de 2024.

6 METODOLOGIA

A metodologia seguiu o fluxograma de funcionamento apresentado na Figura 1. Inicialmente, foi utilizada uma tabela com dados sintéticos para a aplicação da

técnica de anonimização que seria analisada por meio de um algoritmo. A anonimização foi aplicada nas colunas de Nome, Telefone, Endereço, E-mail e Idade, sendo as informações contidas nessas colunas classificadas como dados pessoais. Após a execução da técnica nos dados, foi realizado o Cálculo do Risco de Reidentificação Mensurada (RRM), que apresenta o risco de reidentificação de cada coluna anonimizada.

Figura 1 – Fluxograma de funcionamento.

Fonte: Os autores

7 REPRESENTAÇÕES DA EXPERIMENTAÇÃO

Na condução da análise da eficácia da anonimização foram utilizados conjuntos de dados sintéticos gerados utilizando as bibliotecas Mimesis e Faker do Python, que permitiram a criação de tabelas com informações simuladas para refletir dados realistas. A utilização desse conjunto de dado visa evitar quaisquer implicações éticas e legais que poderiam ocorrer com a utilização de informações reais de usuários. Foi criado um conjunto de dados contendo informações como Nome, Gênero, Telefone, E-mail, Endereço, Religião, Idade, Tipo Sanguíneo, Condição Médica, Data de Admissão, Médico, Plano de Saúde, Valor da Fatura, Número do Quarto, Tipo de Admissão, Resultado do Teste, Data de Alta, Hospital e Medicação Específica.

A k-anonymity foi aplicada no conjunto de dados não anonimizados através

de um algoritmo[8] desenvolvido em Python, onde conforme descrito por Sweeney (2002) a k-anonymity deve garantir que cada registro que está presente no conjunto de dados não pode ser distinguido de pelo menos outros k-1 registros. Na implementação da k-anonymity foi utilizado a generalização e a supressão conforme descrito sobre a técnica, onde ela engloba essas duas maneiras de anonimização, a coluna suprimida foi o Nome e as generalizadas foram a coluna Telefone, E-mail, Endereço e Idade, onde na coluna Telefone foi deixado apenas os 3 últimos dígitos, em Endereço foi deixado apenas o estado (UF), no E-mail apenas o provedor(@provedor.com), na coluna Idade foi definido um intervalo.

Em algumas colunas do conjunto de dados foi utilizado o valor de k igual a 20 onde foi estabelecido algumas regras em que caso a quantidade de informações ou valor de k for menor que 20 na coluna telefone ele generaliza para apenas os 2 últimos dígitos do telefone, na coluna endereço ele generaliza para as regiões geográficas e a idade para uma faixa etária mais abrangente que engloba mais informações.

A aplicação da generalização foi realizada nas mesmas colunas e conjuntos de dados que foi realizada a k-anonymity, através também de um algoritmo que foi feito em Python, respeitando as regras e descrições dessa técnica de anonimização. Na implementação do algoritmo, a coluna Nome teve suas informações generalizadas para o termo “Paciente” seguido por um valor numérico, Telefone foi deixado apenas os 3 últimos dígitos, o Endereço ficou apenas com o UF, no E-mail foi deixado apenas o provedor e na Idade foi definido um intervalo.

Na implementação do algoritmo em python da técnica de perturbação, foram adicionados pequenos ruídos nas informações, onde a coluna Nome teve a adição de letras aleatórias nas informações, o Telefone teve os seus 3 últimos dígitos alterados para valores aleatórios,o E-mail teve apenas a sua primeira parte que contém o usuário alterada para um valor aleatório, no Endereço foi substituído de 1 a 2 palavras por ruídos aleatórios e na coluna Idade foi subtraído ou adicionado aleatoriamente 5 anos em todas as informações contidas nas colunas.

Na implementação do algoritmo em python da técnica de perturbação, foram

[8] Implementação do algoritmo de anonimização utilizado no estudo. O código-fonte e detalhes sobre a aplicação da k-anonymity e outras técnicas de anonimização podem ser encontrados no reposi- tório do GitHub juntamente com com o conjunto de dados: https://github.com/Gabriel-Santos11/Algoritmo_Anonimizacao_LGPD_TCC. Acesso em: 27 out. 2024.

adicionados pequenos ruídos nas informações, onde a coluna Nome teve a adição de letras aleatórias nas informações, o Telefone teve os seus 3 últimos dígitos alterados para valores aleatórios, o E-mail teve apenas a sua primeira parte que contém o usuário alterada para um valor aleatório, no Endereço foi substituído de 1 a 2 palavras por ruídos aleatórios e na coluna Idade foi subtraído ou adicionado aleatoriamente 5 anos em todas as informações contidas nas colunas.

Além da aplicação da técnicas de anonimização também foi realizado a medição do RRM conforme descrito no estudo técnico da ANPD que é definida pela formula abaixo:

$$\text{Risco de Reidentificação Mensurado} = Vc \times \theta$$

Fonte: ANPD (2023). Estudo técnico sobre anonimização de dados na LGPD: uma visão de processo baseado em risco e técnicas computacionais.

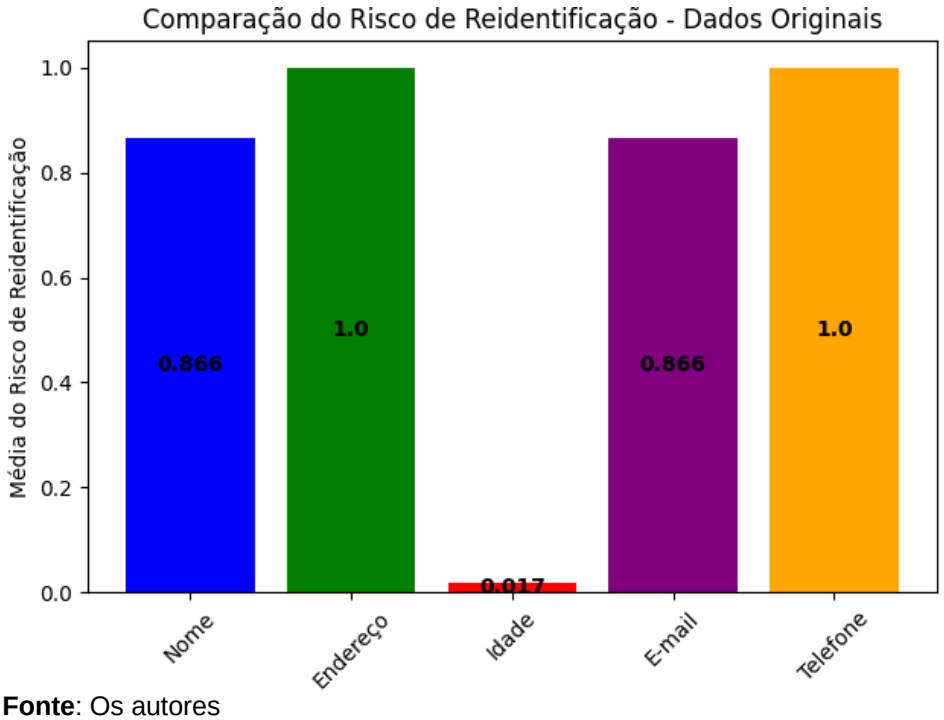
- Vc consiste em um valor que regula a medição do risco de acordo com variáveis contextuais, podendo ser elas a sensibilidade dos dados, a forma como ele é armazenado além de outras características que podem influenciar o risco de reidentificação. Quando não há variáveis contextuais essa valor é definido como 1.
- θ é o valor geral da métrica que é calculado com base na equivalência de classe. A equivalência de classe consiste na representação dos grupos de registros, que compartilham o mesmo valor, onde cada grupo é tratado como uma unidade. Um grupo menor de individuos em uma classe de equivalencia indica um risco maior de reidentificação, pois a distinção entre os registro aumenta.

Esse cálculo foi utilizado para avaliar o risco de reidentificação nas colunas do conjunto de dados, principalmente em cenários de cruzamentos de informações com outras bases ou casos que o atacante tenha um conhecimento prévio das informações do seu alvo.

8 ANÁLISE DOS RESULTADOS OBTIDOS DAS TÉCNICAS DE ANONIMIZAÇÃO

Conforme a figura 2, a analise dos dados sem possuírem nenhum tipo de técnica de anonimização contem varias colunas criticas com o risco de reidentificação extremamente alto como no caso da coluna Endereço e Telefone que apresentam uma taxa de 1,0 ou 100%. A coluna Idade apresentou uma taxa de 0,017 ou 1,7%, porém caso essa coluna fosse cruzadas por exemplo com as informações com a taxa alta facilmente seria possível identificar o individuo.

Figura 2 – Resultado dos dados Originais



Fonte: Os autores

Na figura 3 temos os 17 primeiros dados originais do conjunto de dados antes da ser aplicada qualquer técnica de anonimização, nesses dados é possível perceber que facilmente conseguiríamos identificar o individuo através dessas informações.

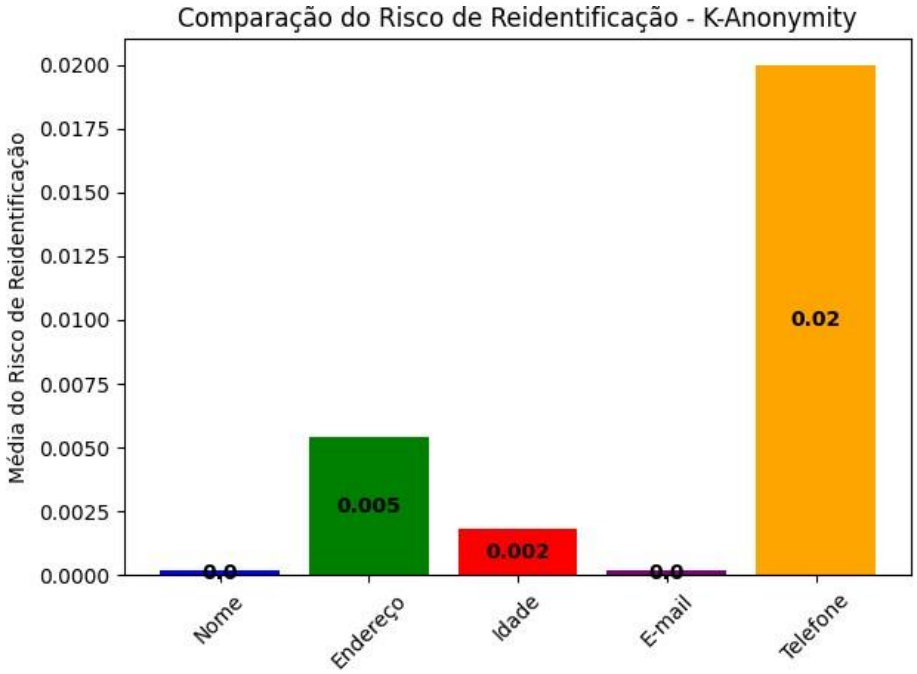
Figura 3 – Dados originais não anonimizados

Nome	Endereço	Idade	E-mail	Telefone
Leandro Tavares	Setor Vitória Jesus, 465 Flavio Marques Lisboa 48337-672 Nogueira / PI	27	leandro.tavares@example.com	+55-21 6727-1602
Joana Serra	Estação Martins, 44 Jardim Guanabara 79561262 Jesus / RR	38	joana.serra@example.com	+55 (87) 3614-8615
Lucas Tavares	Loteamento Camargo, 80 Vila Santo Antônio Barroquinha 04902-818 Cunha dos Dourados / MG	68	lucas.tavares@example.com	+55-91 6192-5136
Nicole Galvão	Setor Josué Farias, 39 Virginia 46752134 Martins / GO	97	nicole.galvão@example.com	+55 (24) 2319-0105
Murilo Barbosa	Vale Bruna Castro, 2 Aarão Reis 72451724 Pacheco / PR	93	murilo.barbosa@example.com	+55-91 9992-2115
Rebeca Oliveira	Recanto de da Cruz, 87 Conjunto Califórnia I 35356-860 Dias / TO	15	rebeca.oliveira@example.com	(11) 13778-2794
Diego Chaves	Jardim Rios, 5 Conjunto Califórnia I 23628982 da Conceição / TO	68	diego.chaves@example.com	(19) 65095-6102
Sophia Nascimento	Lagoa Emanuel Borges Rio Branco 42653303 Sampaio do Norte / AM	62	sophia.nascimento@example.com	+55-21 8050-1966
Fábio Aguiar	Condomínio de Leão, 69 Ademar Maldonado 69770-611 Brito de Moraes / AL	95	fábio.aguiar@example.com	+55-21 8531-3253
Eloá Neves	Campo Nicolas Duarte, 21 Vila Independência 1ª Seção 02788395 Rios / AC	23	eloá.neves@example.com	+55 (24) 2178-3397
Théo Cavalcanti	Residencial Costa, 568 Vila Suzana Primeira Seção 92484-796 Gomes / RO	25	théo.cavalcanti@example.com	+55-91 6128-4964
Washington Freitas	Favela Jade Moura, 90 Vila Puc 44825-717 Gonçalves de Aparecida / PI	43	washington.freitas@example.com	(48) 88131-2243
Marcela Sales	Praça Pacheco, 19 Laranjeiras 23415-509 Costela Verde / MT	21	marcela.sales@example.com	+55-91 2453-3409
Nicolas Couto	Feira de Cunha, 343 Sion 40817-717 Rodrigues do Norte / SE	31	nicolas.couto@example.com	(90) 4165-9184
Gabriela Pimentel	Avenida de Castro, 7 Mariano De Abreu 69777469 das Neves / PE	32	gabriela.pimentel@example.com	+55 (11) 4821-3222
Adriana Galvão	Lagoa Danilo Martins Baleia 36193077 Macedo da Praia / RJ	37	adriana.galvão@example.com	(11) 9760-6736
Mariana Moreira	Aeroporto de Alves, 56 Marieta 3ª Seção 26391219 Moreira de Goiás / PI	85	mariana.moreira@example.com	+55-21 2557-4364

Fonte: Os autores

Ao aplicar a técnica k-anonymity nas colunas é possível perceber que houve uma redução alta no risco de reidentificação principalmente em colunas que anteriormente na tabela original estavam com 100% e agora apresentam respectivamente 0,5% e 2%. Com a aplicação da supressão na coluna Nome a taxa de reidentificação caiu para 0% reduzindo assim a chance que essa informação seja diretamente ligada a um usuário.

Figura 4 – Resultado dos dados anonimizados com a k-anonymity



Fonte: Os autores

Na figura 5 temos o conjunto de dados após realizar a k-anonymity, onde algumas informações foram suprimidas e outras foram generalizadas.

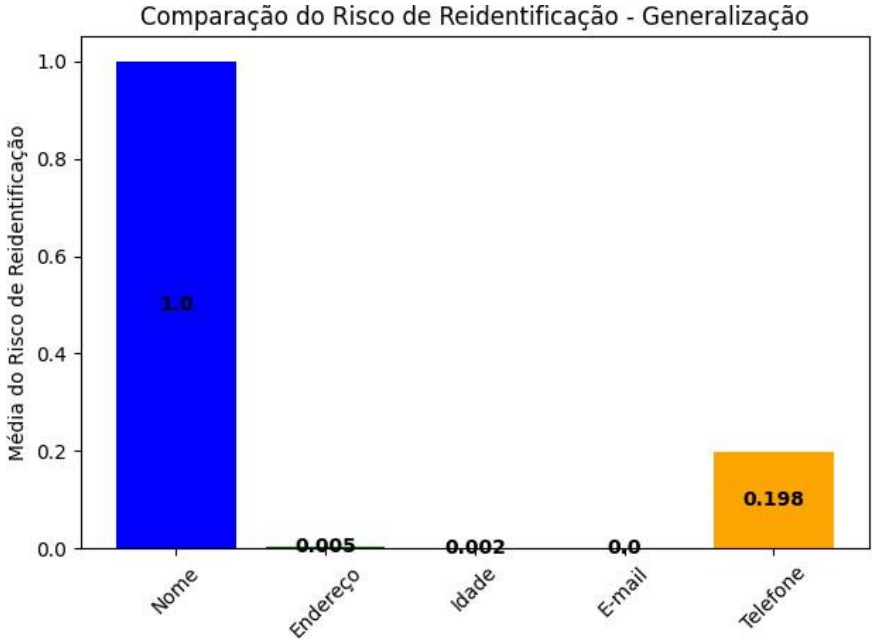
Figura 5 – Dados anonimizados com a k-anonymity

Nome	Endereço	Idade	E-mail	Telefone
*	PI	21-30	***@example.com	*****02
*	RR	31-40	***@example.com	*****15
*	MG	61-70	***@example.com	*****36
*	GO	91-100	***@example.com	*****05
*	PR	91-100	***@example.com	*****15
*	TO	10-20	***@example.com	*****94
*	TO	61-70	***@example.com	*****02
*	AM	61-70	***@example.com	*****66
*	AL	91-100	***@example.com	*****53
*	AC	21-30	***@example.com	*****97
*	RO	21-30	***@example.com	*****64
*	PI	41-50	***@example.com	*****43
*	MT	21-30	***@example.com	*****09
*	SE	31-40	***@example.com	*****84
*	PE	31-40	***@example.com	*****22
*	RJ	31-40	***@example.com	*****36
*	PI	81-90	***@example.com	*****64

Fonte: Os autores

Na aplicação da generalização conforme na figura 6, o risco de reidentificação da coluna nome ficou extremamente alto e as colunas seguintes tiveram uma baixa taxa de reidentificação. O único diferencial alto foi a coluna telefone que teve 0,198 ou 19,8%, onde mesmo que ocorra o cruzamento das informações com as outras colunas a chance de identificar o indivíduo é extremamente baixa.

Figura 6 – Resultado dos dados anonimizados com a Generalização



Fonte: Os autores

No entanto, conforme a figura 7 os nomes não estão identificáveis, onde é possível perceber que elas estão apenas com o termo paciente e um número tendo assim até que um grau de anonimização já que não é possível identificar diretamente o nome real que estava presente naquele local.

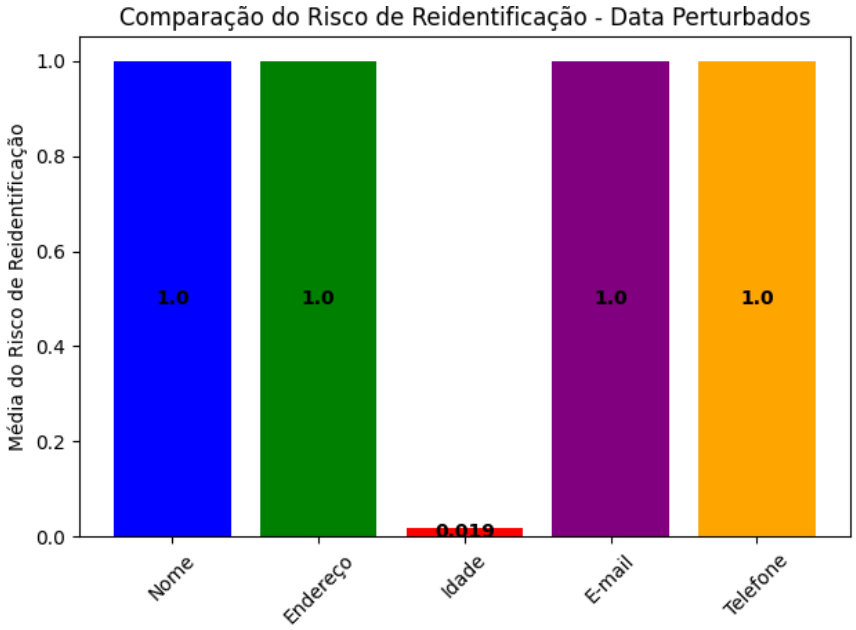
Figura 7 – Dados anonimizados com a generalização

Nome	Endereço	Idade	E-mail	Telefone
Paciente 1	PI	21-30	****@example.com	****602
Paciente 2	RR	31-40	****@example.com	****615
Paciente 3	MG	61-70	****@example.com	****136
Paciente 4	GO	91-100	****@example.com	****105
Paciente 5	PR	91-100	****@example.com	****115
Paciente 6	TO	11-20	****@example.com	****794
Paciente 7	TO	61-70	****@example.com	****102
Paciente 8	AM	61-70	****@example.com	****966
Paciente 9	AL	91-100	****@example.com	****253
Paciente 10	AC	21-30	****@example.com	****397
Paciente 11	RO	21-30	****@example.com	****964
Paciente 12	PI	41-50	****@example.com	****243
Paciente 13	MT	21-30	****@example.com	****409
Paciente 14	SE	31-40	****@example.com	****184
Paciente 15	PE	31-40	****@example.com	****222
Paciente 16	RJ	31-40	****@example.com	****736
Paciente 17	PI	81-90	****@example.com	****364

Fonte: Os autores

Com a aplicação da perturbação no conjunto de dados conforme na figura 8, 4 das 5 colunas tiveram a sua taxa de reidentificação com mais de 1,0 ou 100% demonstrando uma baixa proteção da informação onde apenas a coluna Idade teve um certo grau de proteção que é anulada pelas outras colunas que tiveram números mais altos.

Figura 8 – Resultado dos dados anonimizados com a Perturbação



Fonte: Os autores

Conforme apresentado na figura 9, a perturbação foi aplicada ao conjunto de dados, adicionando ruídos às informações. Com essa adição, algumas informações se tornaram únicas e diferentes da original, resultando em altos índices de identificação em determinadas colunas. Isso demonstra que, embora os dados possam estar anonimizados, as chances de reidentificação são factíveis.

Figura 9 – Dados anonimizados com a perturbação

Nome	Endereço	Idade	E-mail	Telefone
LeanDro Tavares	Setor Vitória bRFGKj 465 Flavio Marques Lisboa 48337-672 Nogueira / PI	31	1P4x1He1H1pyKBc@example.com	+55-21 6727-1712
Joana oerra	Estação Martins, 44 Jardim Guanabara 79561262 oqJRp / RR	42	bZIkS0B1bR1@example.com	+55 (87) 3614-8002
Lucas TaAaBAs	Loteamento Camargo, 80 Vila Santo Antônio Barroquinha NyLcCCXuS PskJB dos Dourados / MG	69	kjkrCB1dPwrW4@example.com	+55-91 6192-5705
licofe Galvão	Setor hENeA Farias, 39 Virgínia 46752134 Martins / GO	99	BwIUJhx7wEc5o@example.com	+55 (24) 2319-0686
Murilo Barxosa	Vale Bruna Castro, 2 Aarão Reis 72451724 BPOIwRy / PR	92	IrkfFvxmNv5Ltj@example.com	+55-91 9992-2081
Rebeca Oliveirn	Recanto de da Cruz, uM QzOeKGyV Califórnia I 35356-860 Dias / TO	15	stW1HHMtkfLIL63@example.com	(11) 13778-2975
DiPgo ChJves	Jardim Rios, 5 Conjunto Califórnia I 23628982 da Conceição / oJ	64	HLkjgQB0Rhpf@example.com	(19) 65095-6205
Sophia NasRimento	Lagoa Emanuel Borges Rio Branco 42653303 gHMIQgP do ndCqP / AM	67	1Ze2M0UTsOLu5mAuq@example.com	+55-21 8050-1830
Fásco AguiMr	Condomínio de Leão, 69 gAXHTX Maldonado 69770-611 Brito de Moraes / ON	99	HD6ySG0Uo5CJ@example.com	+55-21 8531-3922
Eqoá Nwmes	Campo Nicolas PfEGnJC 21 Vila Independencia 1ª Seção 02788395 Rios / AC	28	bS09gKhY08@example.com	+55 (24) 2178-3683
Théo Cavnlcants	Residencial Costa, 568 Vila Suzana Primeira Seção 92484-796 Gomes / Ib	23	Kf6dzCCSuuuOjzA@example.com	+55-91 6128-4478
WashingtoS Freitjs	VGCSzG Jade Moura, 90 Vila Puc 44825-717 Gonçalves de WDbmPsHFj / PI	39	b4CCXXlvlK68taJwVZ@example.com	(48) 88131-2878
carcelH Sales	Praça Pacheco, 19 AKuYBAGWTSU 23415-509 Costela Verde U MT	17	U7aVex8VO5Hxh@example.com	+55-91 2453-3513
NicYlas Couto	Feira de Cunha, 343 Sion 40817-717 Rodrigues do Norte N SE	35	VCSU8loq2Eww6@example.com	(90) 4165-9453
Gabriela PimePtel	Avenida de Castro, 7 Mariano De Abreu 69777469 das dlToU R PE	31	bF8hhRXVIC4GO1LUI@example.com	+55 (11) 4821-3481
bGriana Gatvão	Lagoa Danilo Martins Baleia 36193077 Macedo da ZPcXF / RJ	41	Jr2gULU7SmKoWy@example.com	(11) 9760-6914
MarCana Moreira	Aeroporto fA Alves, 56 Marieta 3ª Seção 26391219 Moreira de Goiás / PI	81	xwO3WDX3mWFrWmY@example.com	+55-21 2557-4045

Fonte: Os autores

9 COMPARAÇÃO DAS TÉCNICAS

A k-anonymity com a união da supressão e generalização aparenta ser a mais eficaz no tangente a anonimização das informações, já que a maior taxa de reidentificação ficou na coluna telefone com 0,02 ou 2%, se esse campo for comparado com o mesmo campo das outras tabelas que foi aplicada a perturbação e a generalização é possível perceber um valor consideravelmente baixo por parte da k-anonymity. Vale destacar também que mesmo que ocorra o cruzamento das informações, buscando identificar algum indivíduo utilizando as outras colunas ou informações prévias, a baixa taxa de reidentificação de 0,027 ou 2,7% somando todas as colunas ainda continua baixa, todavia ela não é totalmente nula, demonstrando assim que uma possível iden- tificação pode ocorrer, mesmo que seja dificultosa ou trabalhosa.

Com a aplicação da generalização sozinha as colunas endereço e idade tiverem o mesmo valor da aplicação da k-anonymity porém sem o valor de k a coluna telefone tem uma taxa bem alta com cerca de 0,178 ou 17,8% de diferença para a k-anonymity. Mesmo a coluna nome tendo uma taxa extremamente alta o fato dos nomes dos pacientes serem apenas o nome com um número acaba por dificultar o reconhecimento direto do indivíduo, conforme destacado anteriormente a alta taxa da coluna telefone pode ser algo prejudicial no conjunto de dados demonstrando uma possível falha, já que caso ocorra o cruzamento das informações com outros itens a chance de identificar o usuário mesmo com a aplicação da generalização se torna bem alta.

Na aplicação da perturbação as taxas vieram extremamente altas se comparadas com a k-anonymity onde 4 das 5 colunas tiveram uma taxa de 1,0 ou 100%, demonstrando assim que dependendo da adição dos ruídos a segurança fica extremamente baixa, possibilitando assim o reconhecimento da informação, vale lembrar também que com a adição de muitos ruídos no conjunto de dados a informação acaba se diferenciando totalmente do real, acabando por dificultar possíveis análises futuras que possam ocorrer principalmente em informações que envolvem números como no caso da idade.

10 CONCLUSÃO

A análise dos resultados obtidos a partir da utilização da métrica de Risco de Reidentificação Mensurada(RRM), indicou que entre as 3 técnicas de anonimizações utilizadas no conjunto de dados a k-anonymity foi a que mais se destacou tendo uma baixa taxa de reidentificação em comparação com as outras técnicas, principalmente em possíveis cenários em que as informações anonimizadas podem ser combinadas ou cruzadas com dados não anonimizados.

Observando as outras técnicas que tiveram uma alta taxa de reidentificação a Generalização foi a que se saiu melhor, principalmente pelo fato do Nome não ter a informação direta do usuário, dificultando possíveis reconhecimentos que poderiam ocorrer caso essa informação estivesse no campo. Na pertubação, mesmo os dado tendo pequenos ruídos na informação a possibilidade de cruzamento com informa ções não anonimizadas é extremamente alto e factível, já que se for adicionado muitos ruídos nas informações, o dado seria totalmente inutilizado para analises futuras e dependendo do tipo do conjunto de dados essa adicção de ruido pode danificar informações que podem ser extremamente uteis.

Olhando pelo lado da LGPD a k-anonymity quando utilizada corretamente em um conjunto de dados atende às exigências da legislação, principalmente quando a LGPD (2018) define que a anonimização consiste em meios e técnicas utilizadas para evitar com que o dado seja associado direta ou indiretamente a um individuo. Por outro lado, conforme também é destacado na LGPD, a k-anonymity pode não atender completamente todos os requisitos da lei em questão da proteção da informação, já que em algumas colunas o risco de reidentificação não chegou a 0,0 ou 0%, demonstrando assim um possível caso de reversão dos dados para o estado originais deles, tornando a informação novamente um dado pessoal, mesmo não tendo a informação do nome no conjunto anonimizado o reconhecimento de outras informações poderia levar ao individuo alvo.

Conforme destacado anteriormente sobre a LGPD a Generalização e a Pertubação não atenderiam diretamente os requisitos da lei, já que nessas técnicas principalmente na pertubação a chance de reversão dos dados anonimizados é extremamente alta, principalmente em informações com um baixo nível de ruido. Na

generalização mesmo a informação estando mais generalizada para grupos maiores de indivíduos o processo de reversão poderia ocorrer principalmente na questão dos dígitos finais do telefone, onde pegando essa informação juntamente com uma região especifica, tendo o conhecimento da idade do alvo facilmente acharia a pessoa real, entretanto não consiguiria pegar informações do usuário, já que o nome é o endereço não estaria presentes nos dados anonimizados.

Analisando o estudo técnico da ANPD, onde é falado sobre o Risco de Reidentificação Mensurada (RRM), a generalização e a pertubação não obtiveram um resultado satisfatório utilizando essa métrica, principalmente por apresentaram altas taxas de 1,0 ou 100% em algumas colunas anonimizadas mesmo ela estando anonimizadas demonstrando assim uma passibilidade de reversão em alguns casos.

Recapitulando, a k-anonymity mostrou-se a técnica mais promissora entre os outros métodos de anonimização, especialmente quando considerada a métrica de Risco de Reidentificação Mensurada (RRM). Com isso, este estudo contribui significativamente para a comunidade científica, fornecendo uma base sobre a eficácia e os limites das técnicas analisadas, e abre um caminho para futuras investigações que possam desenvolver abordagens híbridas ou novos métodos que combinem essas técnicas de outras maneiras.

Portanto é recomendado que em pesquisas futuras seja explorada a união de mais de uma técnica de anonimização em um conjunto de dados, com o objetivo de garantir um equilíbrio entre a utilização e a anonimização da informação principalmente com a finalidade de atender à LGPD no quesito da anonimização de dados.

REFERÊNCIAS

BRASHER, E. Addressing the failure of anonymization: Guidance from the European Union's General Data Protection Regulation. **Columbia Business Law Review**, v. 2018, n. 1, p. 209–253, 2018. Disponível em: https://journals.library.columbia.edu/index.php/CBLR/article/view/1217. Acesso em: 26 out. 2024.

CARVALHO, F. P. **O ser atrás do dado**: limites e desafios da anonimização e seus reflexos nos requisitos estabelecidos pela LGPD. 2024. Disponível em: http://repositorio2.unb.br/jspui/handle/10482/48043.

FERREIRA, J. R. **Aplicação da Lei Geral de Proteção de Dados com utilização de modelos de anonimização de dados em ambiente de nuvem pública**. 2024. Disponível em: http://repositorio2.unb.br/jspui/handle/10482/47940.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: https://www.planalto.gov.br/ccivil_03/_ato20152018/2018/lei/l13709.htm.

MAJEED, A.; LEE, S. Anonymization techniques for privacy preserving data publishing: A comprehensive survey. **IEEE Access**, v. 9, p. 8512–8545, 2020. Disponível em: https://ieeexplore.ieee.org/abstract/document/9298747.

MARQUES, J. F.; BERNARDINO, J. Analysis of data anonymization techniques. **KEOD**. [S.l.: s.n.], 2020. p. 235–241. Disponível em: https://www.scitepress.org/PublishedPapers/2020/101423/101423.pdf.

SAMARATI, P.; SWEENEY, L. **Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression**. Technical report, SRI International, 1998. Disponível em: https://dataprivacylab.org/dataprivacy/projects/kanonymity/paper3.pdf.

SWEENEY, L. k-anonymity: A model for protecting privacy. **International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems**, World Scientific, v. 10, n. 05, p. 557–570, 2002. Disponível em: https://doi.org/10.1142/S0218488502001648.

VIMERCATI, S. D. C. di et al. k-anonymity: From theory to applications. **Trans. Data Priv.**, v. 16, n. 1, p. 25–49, 2023. Disponível em: https://spdp.di.unimi.it/papers/tdp2023.pdf.