Universidade Estadual de Londrina

LUIZ FERNANDO PEREIRA NUNES

CAFIS — UM FRAMEWORK PARA ANONIMIZAÇÃO AUTOMATIZADA BASEADA EM CONTEXTO E SENSIBILIDADE

LONDRINA
2026

LUIZ FERNANDO PEREIRA NUNES

CAFIS — UM FRAMEWORK PARA ANONIMIZAÇÃO AUTOMATIZADA BASEADA EM CONTEXTO E SENSIBILIDADE

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. André Luís Andrade Menolli

LONDRINA

2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UEL

N972c Nunes, Luiz Fernando Pereira Nunes.
CAFIS — UM FRAMEWORK PARA ANONIMIZAÇÃO AUTOMATIZADA BASEADA EM CONTEXTO E SENSIBILIDADE / Luiz Fernando Pereira Nunes Nunes. - Londrina, 2026.
97 f.

Orientador: André Luís Andrade Menolli Menolli.
Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual de Londrina, Centro de Ciências Exatas, Programa de Pós-Graduação em Ciência da Computação, 2026.
Inclui bibliografia.

1. Computação - Tese. 2. LGPD - Tese. 3. Anonimização - Tese. 4. Proteção de Dados - Tese. I. Menolli, André Luís Andrade Menolli. II. Universidade Estadual de Londrina. Centro de Ciências Exatas. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDU 519

LUIZ FERNANDO PEREIRA NUNES

CAFIS — UM FRAMEWORK PARA ANONIMIZAÇÃO AUTOMATIZADA BASEADA EM CONTEXTO E SENSIBILIDADE

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da Universidade Estadual de Londrina para obtenção do título de Mestre em Ciência da Computação.

BANCA EXAMINADORA

Orientador: Prof. Dr. André Luís Andrade Menolli
Universidade Estadual de Londrina

Prof. Dr. Rodolfo Miranda de Barros
Universidade Estadual de Londrina – UEL

Prof. Dr. Marcelo Morandini
Universidade de São Paulo – USP

Londrina, 24 de novembro de 2026.

Este trabalho é dedicado ao meu pai José Elias Nunes (in-memorian), que tanta falta faz e sempre teve orgulho dos caminhos que eu escolhi. Obrigado meu velho e amigo!

AGRADECIMENTOS

A realização desta dissertação foi possível graças ao apoio, à colaboração e ao incentivo de diversas pessoas e instituições, às quais expresso minha sincera gratidão.

Agradeço, de forma especial, à minha esposa, Regiane, pelo apoio incondicional, pela paciência e pela compreensão ao longo de todo o percurso acadêmico, especialmente nos momentos de maior dedicação e renúncia. À minha mãe, Malvina, manifesto meu profundo agradecimento pelo constante incentivo, pelos valores transmitidos e pelo apoio irrestrito em todas as etapas da minha formação.

Registro minha gratidão ao meu orientador, Prof. Dr. André Menolli, pela orientação criteriosa, pela disponibilidade, pelo rigor acadêmico e pelas contribuições fundamentais que nortearam o desenvolvimento deste trabalho. Seu acompanhamento foi essencial para a consolidação da pesquisa e para o amadurecimento científico ao longo do mestrado.

Agradeço à Universidade Estadual de Londrina, em especial ao Programa de Pós-Graduação em Ciência da Computação, bem como a todo o corpo docente do Departamento de Computação, pela formação acadêmica de excelência, pelas disciplinas ministradas e pelo ambiente propício à pesquisa científica.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para o desenvolvimento deste trabalho, seja por meio de discussões, sugestões, apoio técnico ou incentivo, e que, de alguma forma, participaram desta trajetória acadêmica.

"Não vos amoldeis às estruturas deste mundo, mas transformai-vos pela renovação da mente, a fim de distinguir qual é a vontade de Deus: o que é bom, o que Lhe é agradável, o que é perfeito.
(Bíblia Sagrada, Romanos 12, 2))

NUNES, L. F.. **CAFIS — UM FRAMEWORK PARA ANONIMIZAÇÃO AUTOMATIZADA BASEADA EM CONTEXTO E SENSIBILIDADE**. 2026. 98f. Dissertação (Mestrado em Ciência da Computação) – Universidade Estadual de Londrina, Londrina, 2026.

RESUMO

A intensificação do tratamento de dados pessoais em ambientes digitais ampliou os riscos à privacidade e evidenciou a necessidade de soluções técnicas capazes de operacionalizar a conformidade com legislações de proteção de dados. Este trabalho tem como objetivo desenvolver e validar o CAFIS, um framework automatizado para anonimização de dados pessoais baseado na análise de contexto e na sensibilidade dos atributos, em conformidade com a Lei Geral de Proteção de Dados Pessoais. O método adotado fundamenta-se na Design Science Research, estruturada nas etapas de relevância, rigor, design, construção e avaliação do artefato. O framework integra módulos de análise contextual da base, classificação de atributos, avaliação de sensibilidade e risco de reidentificação, seleção automatizada de técnicas de anonimização e geração de indicadores de conformidade. A avaliação empírica foi realizada por meio da aplicação do CAFIS em diferentes bases de dados, considerando critérios técnicos, jurídicos e éticos, com foco na preservação da utilidade dos dados e na mitigação dos riscos à privacidade. Os resultados indicam que o CAFIS apoia o processo decisório de anonimização de forma sistemática e contextualizada, permitindo mensurar a aderência aos princípios da LGPD, especialmente finalidade, necessidade, adequação, prevenção e responsabilização. Conclui-se que o framework contribui para a redução da lacuna entre a normatização jurídica e a implementação técnica da proteção de dados, oferecendo uma abordagem integrada, auditável e adaptável a diferentes contextos organizacionais, sem comprometer a utilidade analítica das informações.

Palavras-chave: Anonimização de Dados. Proteção de Dados Pessoais. LGPD. Framework Automatizado. Privacidade da Informação.

NUNES, L. F.. **CAFIS — An Framework for Context- and Sensitivity-Based Automated Anonymization**. 2026. 98p. Master's Thesis (Master in Science in Computer Science) – State University of Londrina, Londrina, 2026.

ABSTRACT

The intensification of personal data processing in digital environments has increased privacy risks and highlighted the need for technical solutions capable of operationalizing compliance with data protection legislation. This dissertation aims to develop and validate CAFIS, an automated framework for personal data anonymization based on context analysis and attribute sensitivity, in compliance with the Brazilian General Data Protection Law. The adopted method is grounded in Design Science Research, structured into the stages of relevance, rigor, design, construction, and artifact evaluation. The proposed framework integrates modules for contextual dataset analysis, attribute classification, sensitivity assessment and reidentification risk evaluation, automated selection of anonymization techniques, and generation of compliance indicators. The empirical evaluation was conducted through the application of CAFIS to different datasets, considering technical, legal, and ethical criteria, with a focus on preserving data utility and mitigating privacy risks. The results indicate that CAFIS supports the anonymization decision-making process in a systematic and context-aware manner, enabling the measurement of adherence to the principles of the General Data Protection Law, particularly purpose limitation, necessity, adequacy, prevention, and accountability. It is concluded that the framework contributes to reducing the gap between legal regulation and the technical implementation of data protection, offering an integrated, auditable, and adaptable approach to different organizational contexts without compromising the analytical utility of the data.

Keywords: Data Anonymization. Personal Data Protection. LGPD. Automated Framework. Information Privacy.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
LGPD	Lei Geral de Proteção de Dados
GDPR	Regulamento Geral de Proteção de Dados
ANPD	Autoridade Nacional de Proteção de Dados
CCPA	California Consumer Privacy Act
DPO	Data Protection Officer
PIA	Privacy Impact Assessment
DSR	Design Science Research
NBR	Norma Brasileira

SUMÁRIO

1 INTRODUÇÃO

A sociedade contemporânea é marcada por um processo acelerado de digitalização da vida cotidiana, no qual os dados pessoais assumem papel central na operação de empresas, governos e plataformas tecnológicas. Essa transição para a denominada sociedade da informação ampliou as possibilidades de inovação, mas também trouxe desafios significativos à privacidade e à proteção de dados (MALDONADO; BLUM, 2020).

Nesse novo cenário, os dados deixaram de ser meros registros administrativos e passaram a constituir um ativo estratégico de alto valor econômico e social. Tal movimento, conforme Taal (2022), potencializa ganhos de eficiência e inovação, mas exige o redimensionamento dos marcos éticos e jurídicos da informação.

A intensificação do tratamento de dados pessoais, frequentemente realizada sem o devido conhecimento ou consentimento dos titulares, tem gerado uma crise de confiança e de soberania informacional. Vigliar (2021) destaca que o avanço das tecnologias de monitoramento e inferência comportamental introduziu riscos que transcendem a violação da intimidade, atingindo a própria liberdade e dignidade humana.

No Brasil, a promulgação da Lei nº 13.709/2018 — a Lei Geral de Proteção de Dados Pessoais (LGPD) — representou um marco regulatório fundamental ao estabelecer princípios e diretrizes unificadas para o tratamento de dados pessoais. Inspirada no Regulamento Geral sobre a Proteção de Dados da União Europeia (GDPR), a LGPD insere o país em um contexto global de convergência normativa e consolidação dos direitos digitais (MALDONADO; BLUM, 2020; KOHLS et al., 2021).

A legislação brasileira reconhece, em seu artigo 1º, que a proteção de dados pessoais visa assegurar os direitos fundamentais de liberdade, privacidade e livre desenvolvimento da personalidade. Essa formulação consolida a noção de autodeterminação informativa, originada no ordenamento jurídico alemão e posteriormente expandida ao contexto europeu (FABIANO, 2020). A ideia de que o sujeito deve manter controle sobre o fluxo de informações que o representa é central para o debate contemporâneo sobre governança e ética da informação (KOHLS et al., 2021).

A LGPD estabelece, no artigo 6º, um conjunto de princípios para o tratamento legítimo de dados, incluindo finalidade, adequação, necessidade, transparência, segurança e prevenção. Tais princípios não apenas orientam práticas técnicas, mas também reforçam valores democráticos e de *accountability* no tratamento da informação (BRASIL, 2018).

A emergência de sistemas baseados em inteligência artificial e aprendizado de máquina ampliou a complexidade das operações envolvendo dados pessoais. Alford (2020) observa que algoritmos são capazes de inferir características sensíveis a partir de informa-

ções aparentemente inofensivas, o que torna o processo de anonimização progressivamente mais desafiador. Nessa conjuntura, as escolhas de técnicas de anonimização devem considerar não apenas a sensibilidade dos atributos, mas também o contexto regulatório e a necessidade de preservar a utilidade analítica dos dados.

A anonimização de dados ocupa um papel fundamental tanto na proteção da privacidade quanto no cumprimento das normas de tratamento de dados. Para ser considerada eficaz, deve alcançar um equilíbrio entre a preservação da confidencialidade dos titulares e a manutenção da utilidade das informações. De acordo com a LGPD, a anonimização consiste na aplicação de meios técnicos razoáveis e disponíveis no momento do tratamento para impedir que um dado possa identificar seu titular (art. 5º, XI). Como destaca Vigliar (2021), essa é uma condição técnica e também temporal, pois sua efetividade depende das capacidades tecnológicas existentes em cada momento, exigindo avaliações contínuas sobre o risco de reidentificação.

A pseudonimização consiste em uma técnica de proteção de dados que substitui identificadores diretos por pseudônimos, reduzindo a exposição imediata do titular, mas sem eliminar a possibilidade de reidentificação. Dessa forma, embora útil como medida de segurança, a pseudonimização não descaracteriza juridicamente o dado como pessoal, uma vez que o vínculo com o indivíduo permanece reversível mediante o uso de informações adicionais mantidas sob controle do agente de tratamento. Fabiano (2021) reforça que a pseudonimização deve ser compreendida como um mecanismo de segurança e mitigação de riscos, e não como um processo de anonimização propriamente dita, o que implica a manutenção das obrigações legais previstas na legislação de proteção de dados.

Diante desse panorama, a implementação de soluções técnicas e jurídicas integradas é indispensável para garantir a proteção efetiva dos dados pessoais. A construção de frameworks automatizados que considerem o contexto, a sensibilidade dos atributos e a aplicabilidade das técnicas de anonimização surge como caminho promissor para assegurar equilíbrio entre inovação, utilidade e conformidade (MALDONADO; BLUM, 2020; BRASIL, 2022).

Ressalta-se que esta pesquisa não tem como objetivo comparar técnicas de anonimização ou bases de dados entre si, mas sim avaliar o framework CAFIS enquanto artefato computacional, verificando sua capacidade de apoiar o processo de anonimização orientado pelos princípios da LGPD, da proteção de dados e da segurança da informação.

1.1 Problema de Pesquisa

Apesar do avanço de marcos regulatórios como a LGPD e a GDPR, observa-se uma lacuna entre a normatização da proteção de dados e sua efetiva implementação técnica nas organizações. Persistem dificuldades em aplicar metodologias de anonimiza-

ção que conciliem a preservação da utilidade dos dados com a mitigação dos riscos de reidentificação. A literatura científica e as práticas industriais evidenciam a ausência de soluções integradas que aliem rigor técnico, respaldo jurídico e aplicabilidade social. Nesse contexto, formula-se o seguinte questão de pesquisa:

Como desenvolver e validar um framework automatizado de anonimização de dados capaz de equilibrar critérios técnicos de preservação da utilidade e mitigação de riscos, critérios jurídicos de conformidade regulatória e critérios éticos de responsabilidade social, viabilizando o uso legítimo e seguro dos dados em diferentes contextos organizacionais?

1.2 Justificativa Técnica, Jurídica e Social

A relevância desta pesquisa fundamenta-se em três dimensões complementares.

- **Dimensão técnica:** O desenvolvimento de um framework automatizado responde a um desafio central da ciência de dados: compatibilizar métricas avançadas de anonimização como *k-anonymity*, *l-diversity* e *t-closeness*, com escalabilidade, reprodutibilidade e integração a fluxos computacionais existentes. Ao propor uma solução sistematizada e aberta, busca-se fomentar a colaboração científica e tecnológica, promovendo a evolução contínua da ferramenta e o fortalecimento da engenharia da privacidade.
- **Dimensão jurídica:** A conformidade regulatória deixou de ser um diferencial competitivo e tornou-se uma exigência legal. Organizações que não demonstram adequação, correm riscos de sanções, prejuízos financeiros e danos reputacionais. O framework proposto busca traduzir, em termos operacionais, os princípios da LGPD e da GDPR, contribuindo para a efetivação do direito fundamental à proteção de dados e para a consolidação de práticas verificáveis de governança informacional.
- **Dimensão social:** O tratamento ético e seguro de dados transcende o âmbito técnico e jurídico, repercutindo diretamente na confiança pública em tecnologias digitais. Um framework que possibilite anonimização contextualizada reduz assimetrias de poder informacional, protege grupos vulneráveis contra discriminação algorítmica e amplia o uso responsável de dados em pesquisas, políticas públicas e inovação. Assim, a pesquisa propõe uma mediação entre inovação tecnológica e direitos fundamentais, contribuindo para o fortalecimento da cidadania digital.

1.3 Objetivos Gerais

Desenvolver e validar o CAFIS *(Context-Aware Framework for Information Secrecy*), um framework automatizado para anonimização de dados pessoais baseado em

critérios de contextualização do uso, sensibilidade dos atributos e emprego de técnicas avançadas, em conformidade com a LGPD e com foco na preservação da utilidade dos dados para análise, pesquisa e tomada de decisão.

1.3.1 Objetivos Específicos

- Classificar tipos de dados pessoais e sensíveis considerando o risco de reidentificação e o impacto sobre os direitos dos titulares;
- Mapear e comparar técnicas de anonimização, avaliando limitações e aplicabilidades contextuais;
- Projetar uma solução que integre análise de contexto, avaliação de sensibilidade e recomendação de técnicas adequadas;
- Implementar o CAFIS por meio de prototipagem computacional utilizando linguagens e bibliotecas apropriadas;
- Avaliar a eficácia do CAFIS quanto à privacidade, utilidade e aderência técnica e jurídica.

2 FUNDAMENTAÇÃO TÉORICA

A fundamentação teórica desta pesquisa apresenta os conceitos, princípios e abordagens que sustentam o desenvolvimento de métodos de anonimização de dados e sua conformidade jurídica no contexto da proteção de dados pessoais. Diante do avanço das tecnologias de coleta, armazenamento e inferência, compreender a natureza dos dados pessoais, suas classificações e os riscos associados ao tratamento inadequado torna-se essencial para estruturar soluções tecnicamente robustas e juridicamente alinhadas à Lei Geral de Proteção de Dados (LGPD) e a outros marcos regulatórios internacionais.

Além disso, esta seção discute as principais técnicas de anonimização, suas limitações e sua evolução ao longo do tempo, bem como modelos comparados de conformidade, como a GDPR europeu e o CCPA norte-americano. A revisão de literatura reúne contribuições de autores clássicos e contemporâneos, fornecendo uma base sólida para o desenvolvimento do CAFIS e permitindo integrar dimensões técnicas, jurídicas e éticas em uma abordagem coerente e fundamentada.

2.1 Conceitos Fundamentais: dados pessoais, sensíveis e pseudonimização

A compreensão dos conceitos centrais relativos à proteção de dados constitui condição indispensável para o desenvolvimento de soluções tecnológicas que estejam em conformidade com o ordenamento jurídico brasileiro. Entre esses conceitos, destacam-se os de dados pessoais, dados sensíveis e pseudonimização, previstos na Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018) e amplamente debatidos na literatura especializada.

De acordo com o artigo 5º da LGPD, dados pessoais são quaisquer informações relacionadas à pessoa natural identificada ou identificável. Essa definição segue a linha adotada pelo Regulamento Geral sobre a Proteção de Dados da União Europeia (GDPR), que também adota uma concepção ampla e contextual da identificação (EUROPEAN PARLIAMENT AND COUNCIL, 2016; MALDONADO; BLUM, 2020). A identificação pode ocorrer de forma direta — por meio de atributos como nome, CPF ou RG — ou indireta, quando a combinação de diferentes atributos permite inferir a identidade do indivíduo (VIGLIAR, 2021).

A LGPD diferencia três grandes categorias de informação: dados pessoais, dados sensíveis e dados anonimizados, conforme sintetizado no Quadro 1.

Quadro 1 – Classificação de dados segundo a LGPD

Categoria	**Exemplo**	**Nível de Sensibilidade**	**Tratamento Jurídico**
Pessoal	Nome, CPF	Médio	Sujeito às disposições da LGPD
Sensível	Religião, saúde	Alto	Requer base legal específica
Anonimizado	Dados estatísticos	Baixo	Fora do escopo da LGPD

Fonte: Adaptado da LGPD (Brasil, 2018).

Conforme o artigo 5º, inciso II, da LGPD, dados sensíveis são aqueles que envolvem origem racial ou étnica, convicção religiosa, opinião política, filiação sindical, dados referentes à saúde, à vida sexual, dados genéticos ou biométricos. O tratamento dessas informações exige cautela redobrada, pois seu uso indevido pode gerar discriminação ou violar direitos fundamentais (MALDONADO; BLUM, 2020; VIGLIAR, 2021). A legislação impõe, portanto, bases legais mais restritivas e medidas técnicas de segurança reforçadas. Essa diferenciação encontra paralelo na GDPR, que também classifica essas informações como special categories of personal data, exigindo proteção adicional (TAAL, 2022).

Já os dados anonimizados são aqueles que, mesmo mediante meios técnicos razoáveis e disponíveis, não permitem identificar o titular (art. 5º, XI). Por essa razão, deixam de estar sujeitos à incidência da LGPD, sendo tratados apenas como informações estatísticas ou científicas. Entretanto, a anonimização é considerada contextual e temporária, pois sua eficácia depende dos recursos tecnológicos disponíveis em determinado momento (VIGLIAR, 2021).

A LGPD adota, portanto, uma abordagem baseada em risco, reconhecendo que a possibilidade de um dado ser considerado pessoal depende do contexto tecnológico e social de uso. Cabe aos agentes de tratamento avaliar não apenas a informação isolada, mas também seu potencial de vinculação com outras bases de dados (KOHLS et al., 2021).

No interior da categoria de dados pessoais, a legislação prevê a subcategoria dos dados sensíveis, que demandam salvaguardas adicionais. Essa classificação decorre do princípio da proporcionalidade e busca mitigar riscos de dano moral, social ou econômico. Autores como DONEDA (2019) e RODOTÀ (2020) ressaltam que a proteção de dados sensíveis expressa o núcleo essencial da autodeterminação informativa, por se referir diretamente à dignidade e à integridade da pessoa humana.

Outro conceito central é o de pseudonimização, previsto no inciso XI do artigo 5º da LGPD e igualmente tratado na GDPR. A pseudonimização consiste no tratamento de dados de modo que eles não possam ser associados a um titular específico sem o uso de

informações adicionais mantidas em ambiente separado e seguro (EUROPEAN PARLIAMENT AND COUNCIL, 2016). Na prática, envolve técnicas como *hashing*, *tokenização*, substituição de identificadores e criptografia (CABRAL; SIRQUEIRA, 2023; FIESCHI et al., 2025).

Contudo, a pseudonimização não descaracteriza o dado como pessoal, pois mantém um vínculo reversível com o titular. Diferentemente da anonimização, que busca a irreversibilidade, a pseudonimização apenas introduz uma camada de proteção adicional. Assim, deve ser entendida como uma técnica de segurança da informação, e não como um processo de anonimização propriamente dito (FABIANO, 2021; MALDONADO; BLUM, 2020).

A distinção entre anonimização e pseudonimização é, portanto, crucial tanto no plano jurídico quanto no técnico. Enquanto a primeira, se eficaz e irreversível, exclui o dado do escopo da LGPD, a segunda apenas limita o risco de exposição, permanecendo o dado sob a tutela da lei. Essa diferenciação orienta as decisões de implementação, avaliação de risco e auditoria dos processos de proteção da informação.

Em síntese, a incorporação precisa desses conceitos — dados pessoais, sensíveis, anonimizados e pseudonimizados — constitui a base conceitual do CAFIS proposto nesta pesquisa. Ao adotar tais categorias como referência normativa e técnica, o trabalho assegura coerência entre proteção jurídica, eficiência tecnológica e preservação da utilidade analítica dos dados, em conformidade com os princípios da Lei Geral de Proteção de Dados Pessoais.

2.2 Técnicas de anonimização

A anonimização de dados constitui um conjunto de métodos e estratégias voltados à redução do risco de reidentificação de indivíduos em bases de dados. A literatura especializada reconhece que a eficácia desses métodos depende da capacidade de preservar a utilidade analítica das informações sem comprometer a privacidade dos titulares (SORIA-CHEVEDDEN et al., 2023; TAAL, 2022).

Historicamente, as técnicas de anonimização evoluíram em resposta ao aumento da disponibilidade de dados e à sofisticação dos mecanismos de inferência. As abordagens clássicas — como *k-anonymity*, *l-diversity* e t-*closeness* — estabelecem fundamentos matemáticos que permanecem relevantes, embora apresentem limitações diante de novos cenários de dados massivos e heterogêneos (SWEENEY, 2002; LI et al., 2007).

A técnica de *k-anonymity*, proposta por Sweeney (2002), estabelece que um conjunto de dados é considerado anônimo quando cada registro é indistinguível de, pelo menos, outros k – 1 registros quanto aos atributos quase-identificadores. Essa técnica busca reduzir a probabilidade de associação direta entre um conjunto de atributos e um

indivíduo. Contudo, *k-anonymity* é vulnerável a ataques de homogeneidade, nos quais todos os registros de um grupo compartilham o mesmo valor sensível (LI et al., 2007).

A técnica de *l-diversity* foi desenvolvida como uma extensão do modelo anterior, exigindo que cada grupo de equivalência contenha pelo menos l valores distintos para o atributo sensível. Essa variação aumenta a entropia e dificulta inferências diretas, mas ainda não impede ataques de proximidade, quando os valores distintos são semanticamente similares (MACHANAVAJJHALA et al., 2007).

Com o objetivo de superar tais limitações, foi introduzido o conceito de *t-closeness*, que propõe que a distribuição de valores sensíveis em cada grupo de equivalência não deve divergir da distribuição geral em mais de um limite t. Essa abordagem preserva melhor as propriedades estatísticas do conjunto de dados, embora sua aplicação prática seja mais complexa e custosa (LI et al., 2007). O Quadro 2 apresenta uma síntese comparativa dessas técnicas clássicas.

Quadro 2 – Comparativo entre técnicas clássicas de anonimização

Técnica	**Princípio Básico**	**Benefício Principal**	**Limitações**
k-anonymity	Indistinguibilidade entre k registros	Simplicidade e ampla aplicabilidade	Vulnerável a ataques de homogeneidade
l-diversity	Diversidade de valores sensíveis	Reduz inferências diretas	Frágil a ataques de proximidade
t-closeness	Similaridade da distribuição estatística	Preserva estatísticas globais	Complexidade computacional elevada

Fonte: Adaptado de Sweeney (2002) e Li et al. (2007).

Essas técnicas clássicas permanecem relevantes, sobretudo em contextos regulatórios e acadêmicos que demandam transparência e auditabilidade. No entanto, seu desempenho tende a decair em cenários de alta dimensionalidade, grandes volumes de dados ou quando há necessidade de análises preditivas baseadas em aprendizado de máquina.

O avanço das tecnologias de aprendizado de máquina e *big data* impulsionou o surgimento de novas técnicas de anonimização, mais robustas e adaptáveis. Entre elas, destacam-se a privacidade diferencial (*differential privacy*), a síntese de dados (*data synthesis*), a criptografia homomórfica e os métodos de aprendizado federado.

A privacidade diferencial, introduzida por Dwork e Roth (2014), fornece uma garantia formal de privacidade ao adicionar ruído estatisticamente calibrado às consultas ou resultados de análise. Esse ruído é proporcional a um parâmetro ε, que representa o limite de privacidade: quanto menor o valor de ε, maior a proteção, mas menor a precisão dos resultados. Essa técnica é amplamente utilizada em sistemas de grandes plataformas digitais, como Apple e Google, e é considerada o padrão-ouro para a publicação de dados sensíveis (DWORK; ROTH, 2014; GADOTTI et al., 2024).

Outra estratégia emergente é a síntese de dados, que consiste na geração de registros artificiais com base em distribuições estatísticas ou modelos de aprendizado profundo. Essa abordagem busca reproduzir padrões reais sem expor dados verdadeiros, sendo amplamente empregada em pesquisas médicas e estudos de políticas públicas (ZHAO et al., 2023). Modelos generativos, como Generative Adversarial Networks (GANs), são frequentemente utilizados para produzir bases sintéticas de alta fidelidade, preservando correlações e estruturas estatísticas.

A criptografia homomórfica representa um avanço relevante ao permitir o processamento de dados criptografados sem a necessidade de descriptografá-los. Essa técnica possibilita a realização de análises seguras em ambientes de terceiros, reforçando o princípio de *privacy by design* previsto no artigo 46 da LGPD (CABRAL; SIRQUEIRA, 2023).

Por sua vez, o aprendizado federado (*federated learning*) surge como alternativa para treinamento colaborativo de modelos de inteligência artificial sem a centralização de dados sensíveis. Em vez de transferir as bases de dados, transfere-se o modelo, que é treinado localmente e agregado por meio de parâmetros compartilhados (KONEČNÝ et al., 2016). Essa abordagem reduz significativamente o risco de exposição, além de promover conformidade com princípios de minimização e necessidade.

O Quadro 3 resume as novas principais técnicas, destacando suas características e contextos de aplicação.

Quadro 3 – Novas técnicas de anonimização e proteção de dados

Técnica	**Mecanismo**	**Vantagens**	**Limitações**	**Aplicações**
Differential Privacy	Adição de ruído estatístico calibrado	Garantia formal de privacidade	Perda de precisão	Publicação de estatísticas e IA
Data Synthesis	Geração de dados artificiais	Preserva padrões e anonimiza	Risco de sobreajuste	Pesquisa médica e testes de software
Criptografia homomórfica	Processamento de dados cifrados	Alta segurança	Alto custo computacional	Computação em nuvem e contratos inteligentes
Federated Learning	Treinamento descentralizado de modelos	Minimiza transferência de dados	Complexidade de orquestração	Saúde, finanças e IoT

Fonte: Adaptado de Dwork e Roth (2014); Konečný et al. (2016); Gadotti et al. (2024).

Essas técnicas refletem uma transição do paradigma puramente estatístico para abordagens computacionais e contextuais, nas quais a privacidade é tratada como uma propriedade mensurável e controlável dentro do ciclo de vida do dado. Tal evolução permite o desenvolvimento de frameworks híbridos, que combinam anonimização clássica, privacidade diferencial e aprendizado federado em fluxos automatizados e auditáveis.

A evolução das técnicas de anonimização revela a transformação da privacidade em um campo interdisciplinar, que articula estatística, ciência da computação, direito e

ética. No contexto da LGPD, a seleção de técnicas deve observar não apenas critérios técnicos, mas também princípios jurídicos, como necessidade, proporcionalidade e finalidade (BRASIL, 2018). Portanto, a escolha de um método de anonimização não pode ser dissociada do contexto de uso e do grau de risco associado aos dados tratados. Frameworks automatizados como o proposto nesta pesquisa buscam exatamente preencher essa lacuna, oferecendo mecanismos capazes de avaliar o contexto, recomendar técnicas adequadas e mensurar o equilíbrio entre privacidade e utilidade.

Neste trabalho, as técnicas de anonimização não são analisadas com o objetivo de aferir desempenho relativo ou superioridade técnica, mas sim como elementos de apoio à decisão no contexto do framework CAFIS, que atua na seleção e aplicação contextualizada dessas técnicas conforme requisitos legais, técnicos e éticos.

2.3 Princípios da LGPD Aplicados à Anonimização

A Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018) consolidou um marco regulatório que redefine as práticas de tratamento de dados no Brasil, introduzindo princípios orientadores que norteiam a governança da informação em ambientes digitais. Esses princípios, estabelecidos no artigo 6º da LGPD, visam assegurar que qualquer operação envolvendo dados pessoais seja realizada de forma ética, transparente e proporcional. No contexto da anonimização, tais princípios desempenham papel fundamental na definição de critérios técnicos, jurídicos e organizacionais para o uso legítimo e responsável dos dados.

Segundo Doneda (2019), os princípios da LGPD funcionam como balizas normativas que orientam não apenas a interpretação jurídica da lei, mas também o desenho de soluções tecnológicas compatíveis com a autodeterminação informativa. Assim, o processo de anonimização deve ser concebido como um mecanismo de conformidade dinâmica, capaz de traduzir valores jurídicos em procedimentos técnicos verificáveis.

2.3.1 Princípio da Finalidade

O princípio da finalidade determina que o tratamento de dados pessoais deve ocorrer para propósitos legítimos, específicos e informados ao titular (BRASIL, 2018). No contexto da anonimização, esse princípio implica que o dado deve ser tratado apenas até o ponto em que a sua identificação seja necessária para o objetivo declarado.

Conforme destaca Bygrave (2022), a anonimização deve ser orientada por finalidades claras e mensuráveis, evitando o tratamento genérico de dados sob justificativas amplas. Isso reforça a necessidade de design orientado ao propósito, em que a anonimização é planejada desde a concepção do sistema *(privacy by design)* e não apenas como uma etapa posterior.

2.3.2 Princípio da Necessidade

O princípio da necessidade complementa a finalidade ao exigir a limitação do tratamento ao mínimo necessário para a realização da atividade. No caso da anonimização, ele se manifesta na seleção adequada de técnicas e no controle do nível de granularidade dos dados. Maldonado e Blum (2020) afirmam que a minimização de dados é um componente essencial da conformidade, devendo-se reduzir tanto a quantidade quanto o grau de identificação das informações pessoais. Assim, a anonimização atua como instrumento prático de cumprimento desse princípio, eliminando atributos desnecessários e reduzindo riscos de exposição.

Em termos técnicos, a aplicação da necessidade envolve o uso de métricas de utility loss e privacy gain, avaliando o impacto da anonimização sobre a utilidade dos dados. Essa avaliação deve ser integrada ao processo de governança de dados, permitindo a mensuração contínua da proporcionalidade entre privacidade e desempenho informacional.

2.3.3 Princípio da Adequação

O princípio da adequação impõe que o tratamento de dados seja compatível com as finalidades informadas ao titular e com o contexto do tratamento. A adequação estabelece uma relação de coerência entre o uso pretendido e as expectativas legítimas do titular (RODOTÀ, 2020). No caso da anonimização, esse princípio exige que a técnica adotada seja coerente com o tipo e a sensibilidade dos dados e com o ambiente em que serão tratados. Por exemplo, técnicas de supressão ou generalização podem ser adequadas para conjuntos de dados administrativos, enquanto abordagens baseadas em differential privacy podem ser mais apropriadas para bases estatísticas ou de pesquisa científica.

A adequação também se manifesta na governança algorítmica: as decisões sobre anonimização devem ser transparentes, documentadas e justificáveis, assegurando rastreabilidade e possibilidade de auditoria, conforme defendem Floridi (2019) e Martin (2022).

2.3.4 Princípio da Prevenção

O princípio da prevenção determina que os agentes de tratamento adotem medidas proativas para evitar a ocorrência de danos. A anonimização, nesse contexto, atua como um mecanismo preventivo, reduzindo riscos antes que eles se materializem. Segundo Fabiano (2021), a prevenção não deve ser vista apenas como um requisito técnico, mas como um compromisso ético de antecipar vulnerabilidades e mitigar ameaças potenciais. Técnicas como *data masking*, *tokenization* e *k-anonymity* podem ser aplicadas como barreiras de primeira linha contra ataques de reidentificação, complementadas por auditorias periódicas de privacidade e reavaliação dos riscos associados à evolução tecnológica.

A cláusula presente na LGPD que menciona o uso de “meios técnicos razoá-

veis e disponíveis" evidencia o caráter dinâmico do princípio da prevenção, pois o que é tecnicamente suficiente em determinado momento pode tornar-se obsoleto com o avanço das técnicas de reidentificação (MALDONADO; BLUM, 2020).

2.3.5 Princípio da Transparência e da Responsabilização (*Accountability*)

A transparência e a responsabilização, princípios centrais da LGPD, reforçam a obrigação dos controladores de demonstrar conformidade e garantir a compreensão dos processos de tratamento. No contexto da anonimização, isso implica documentar as escolhas técnicas, justificar parâmetros adotados e manter registros que permitam auditoria e replicabilidade (BYGRAVE, 2022).

A transparência exige que o titular tenha acesso claro às políticas de proteção de dados e aos critérios de anonimização empregados, dentro dos limites da segurança da informação. Já o princípio da responsabilização — ou *accountability* — exige comprovação objetiva das medidas adotadas, incluindo logs, relatórios técnicos e registros de decisão.

Para Doneda (2019), a responsabilização é o ponto de convergência entre o direito e a engenharia de sistemas, pois transforma princípios abstratos em práticas verificáveis. A adoção de frameworks automatizados que gerem relatórios de conformidade, como o proposto nesta dissertação, representa uma forma efetiva de materializar esse princípio, permitindo monitoramento contínuo e padronizado.

2.3.6 A Interrelação dos Princípios e a Dinâmica de Conformidade

Os princípios da LGPD não devem ser compreendidos isoladamente, mas como um sistema normativo interdependente que orienta todo o ciclo de vida do dado. Na prática, isso significa que o processo de anonimização deve observar simultaneamente finalidade, necessidade, adequação, prevenção e accountability, em um fluxo contínuo de governança.

Essa inter-relaçãointerrelação reflete o caráter dinâmico e contextual da conformidade. Como observa Doneda (2019), o cumprimento da LGPD não se dá por mera aplicação formal da lei, mas pela integração de seus princípios à lógica de desenvolvimento tecnológico. A anonimização, portanto, deve ser concebida como uma ferramenta operacional de efetivação de direitos fundamentais, articulando técnica, ética e regulação.

Os princípios da LGPD aplicados à anonimização expressam a transposição de valores constitucionais como liberdade, privacidade e dignidade, para o domínio técnico da engenharia de dados. Eles asseguram que o processo de anonimização vá além da obliteração de identificadores, incorporando responsabilidade social, ética e transparência como fundamentos da prática científica e profissional.

Desse modo, a anonimização deixa de ser apenas um requisito de segurança e

passa a representar uma estratégia de governança, inserida na cultura organizacional e no ciclo de vida da informação. É nesse entrelaçamento entre norma e técnica que se estrutura a proposta deste trabalho: o desenvolvimento de um framework automatizado capaz de operacionalizar os princípios da LGPD em contextos computacionais reais, conciliando conformidade legal e eficiência analítica.

2.4 Modelos de Conformidade: GDPR x LGPD x CCPA

A proteção de dados pessoais consolidou-se como um eixo central da regulação digital em diversas jurisdições. Embora cada sistema jurídico adote especificidades culturais e normativas, observa-se uma tendência global de convergência em torno de princípios comuns, como transparência, finalidade, minimização e responsabilização. Entre os principais marcos regulatórios contemporâneos destacam-se o Regulamento Geral sobre a Proteção de Dados da União Europeia (GDPR), a Lei Geral de Proteção de Dados Pessoais (LGPD) do Brasil e a California Consumer Privacy Act (CCPA), nos Estados Unidos.

Essas legislações representam diferentes modelos de governança da privacidade: o europeu, baseado em direitos fundamentais; o brasileiro, de caráter híbrido e adaptativo; e o norte-americano, predominantemente orientado ao consumo e à transparência comercial (BYGRAVE, 2022; MALDONADO; BLUM, 2020). A comparação entre esses instrumentos permite compreender a posição da LGPD como um marco intermediário que dialoga tanto com a tradição europeia de proteção de direitos quanto com a flexibilidade regulatória norte-americana.

2.4.1 Regulamento Geral sobre a Proteção de Dados (GDPR)

A GDPR, em vigor desde 2018, é considerado o modelo paradigmático de proteção de dados, servindo de referência global para legislações posteriores (EUROPEAN PARLIAMENT AND COUNCIL, 2016). Fundamentado no artigo 8º da Carta de Direitos Fundamentais da União Europeia, o regulamento trata a proteção de dados como direito fundamental autônomo, associado à dignidade humana e à autodeterminação informativa (RODOTÀ, 2020).

A GDPR estrutura-se em princípios como licitude, lealdade e transparência, limitação da finalidade, minimização de dados, exatidão, limitação de armazenamento, integridade e confidencialidade e responsabilização (*accountability*). Além disso, introduz a figura do *Data Protection Officer* (DPO) e mecanismos de *Privacy Impact Assessment* (PIA), que institucionalizam a avaliação prévia de riscos.

No que tange à anonimização, a GDPR distingue claramente dados pessoais, dados pseudonimizados e dados anonimizados, conforme o artigo 4º, incisos 1 e 5. Dados

anonimizados são explicitamente excluídos do escopo da norma, desde que a reidentificação não seja possível "por meios razoáveis". Essa definição influenciou diretamente a formulação da LGPD, embora a legislação europeia enfatize a natureza dinâmica da anonimização e a necessidade de reavaliação contínua (BYGRAVE, 2022).

O regulamento também introduz o princípio do *privacy by design and by default*, obrigando que produtos e sistemas incorporem mecanismos de proteção de dados desde sua concepção. Essa exigência reforça a vinculação entre engenharia de software e responsabilidade jurídica, antecipando o que a literatura denomina "privacidade incorporada ao código" (*privacy embedded in code*) (DONEDA, 2019).

2.4.2 Lei Geral de Proteção de Dados Pessoais (LGPD)

A LGPD, sancionada em 2018 e inspirada fortemente na GDPR, representa um modelo híbrido de adequação regulatória, adaptado ao contexto jurídico e tecnológico brasileiro (BRASIL, 2018). Embora mantenha a estrutura principiológica europeia, a LGPD introduz nuances importantes, como a previsão de hipóteses legais específicas de tratamento e o papel da Autoridade Nacional de Proteção de Dados (ANPD) como órgão regulador e orientador.

A lei brasileira reconhece expressamente o direito fundamental à proteção de dados pessoais, incorporado à Constituição Federal pela Emenda nº 115/2022, reforçando a centralidade desse direito no sistema de garantias fundamentais.

No tocante à anonimização, a LGPD adota definição semelhante à da GDPR, mas com formulação própria: "utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo" (art. 5º, XI). Essa definição traz uma dimensão temporal e contextual, reconhecendo que a eficácia da anonimização depende do estado tecnológico e do risco de reidentificação (MALDONADO; BLUM, 2020).

Diferentemente da GDPR, a LGPD incorpora o conceito de pseudonimização de forma mais flexível, permitindo seu uso como medida de segurança, mas não como exclusão do escopo legal. A lei brasileira, portanto, combina rigor normativo e elasticidade interpretativa, adequando-se à diversidade de maturidade tecnológica das organizações.

Outro elemento distintivo da LGPD é a ênfase na educação e na cultura de proteção de dados, promovendo uma visão pedagógica de conformidade. Essa abordagem reflete o esforço de consolidar uma governança democrática e participativa, em que o cumprimento da lei depende da integração entre setores públicos, privados e acadêmicos (KOHLS et al., 2021).

2.4.3 *California Consumer Privacy Act* (CCPA)

A CCPA, aprovada em 2018 e em vigor desde 2020, constitui o principal marco regulatório de privacidade dos Estados Unidos. Ao contrário da GDPR e da LGPD, a CCPA tem orientação consumerista, sendo menos voltada a direitos fundamentais e mais à transparência de mercado. Seu objetivo principal é garantir aos consumidores o controle sobre o uso de suas informações pessoais por empresas, com base na lógica de consentimento e transparência (CALIFORNIA LEGISLATURE, 2018).

A lei confere aos titulares direitos específicos, como acesso, exclusão e *opt-out* da venda de dados, mas não impõe uma estrutura de princípios tão abrangente quanto o modelo europeu. Em relação à anonimização, a CCPA reconhece a categoria de dados desidentificados (deidentified data), definindo-os como aqueles que não podem ser razoavelmente associados a um consumidor específico, direta ou indiretamente. Contudo, a lei permite a reidentificação sob certas condições, especialmente quando necessária para auditorias ou segurança, o que reduz a robustez do modelo de proteção (GONZÁLEZ-BLAS et al., 2024).

A ausência de um órgão regulador nacional e a natureza fragmentada das legislações estaduais norte-americanas tornam o sistema menos uniforme, dependendo de autorregulação e ações judiciais para garantir conformidade. Ainda assim, a CCPA representa um avanço importante no contexto norte-americano, aproximando-se gradualmente dos padrões internacionais de privacidade.

2.4.4 Comparativo entre GDPR, LGPD e CCPA

A Quadro 4 apresenta uma síntese comparativa dos três marcos regulatórios, destacando aspectos estruturais e técnicos relevantes para o processo de anonimização.

Quadro 4 – Comparativo entre GDPR, LGPD e CCPA

Aspecto	**GDPR (UE)**	**LGPD (Brasil)**	**CCPA (EUA)**
Natureza jurídica	Direito fundamental	Direito fundamental (após EC 115/2022)	Direito do consumidor
Escopo	União Europeia e países associados	Território nacional e dados tratados por entidades brasileiras	Estado da Califórnia
Órgão regulador	Autoridades nacionais independentes	ANPD (Autoridade Nacional de Proteção de Dados)	Attorney General da Califórnia
Conceito de anonimização	Exclusão de dados irreversivelmente identificáveis	Meios técnicos razoáveis e disponíveis	*Deidentified data*, com possibilidade de reidentificação
Pseudoanonimização	Medida de segurança (art. 4º, 5º)	Técnica complementar de segurança	Raramente mencionada
Princípios fundamentais	Licitude, finalidade, minimização, accountability	Finalidade, necessidade, adequação, prevenção, transparência	Transparência e consentimento
Direitos dos titulares	Acesso, retificação, portabilidade, esquecimento	Acesso, correção, anonimização, portabilidade	Acesso, exclusão e *opt-out*
Penalidades	Multas até 4% do faturamento global	Multas até 2% do faturamento nacional	Multas fixas e ações coletivas

Fonte: Adaptado de GDPR (2016), LGPD (2018) e CCPA (2018).

A comparação entre os modelos evidencia que, embora compartilhem princípios estruturais, cada legislação reflete valores distintos de regulação da privacidade. A GDPR enfatiza a centralidade dos direitos fundamentais e o dever de responsabilidade técnica; a LGPD busca equilíbrio entre proteção e desenvolvimento econômico, promovendo um modelo adaptativo; e a CCPA foca na autonomia do consumidor e na transparência das relações comerciais.

Sob a perspectiva da anonimização, a GDPR apresenta o padrão mais rigoroso, tratando a irreversibilidade como condição essencial para exclusão do escopo normativo. A LGPD, por sua vez, adota um conceito contextual e tecnicamente orientado, alinhado à ideia de conformidade evolutiva. Já a CCPA privilegia mecanismos de opt-out e anonimização pragmática, sem exigências formais de irreversibilidade.

Essas diferenças reforçam a importância de frameworks flexíveis e contextuais, capazes de adaptar parâmetros técnicos conforme a jurisdição aplicável. O modelo proposto nesta dissertação insere-se nessa lógica, buscando harmonizar princípios normativos internacionais com práticas técnicas auditáveis e automatizadas.

2.5 Trabalhos Relacionados

O estudo da anonimização de dados e das técnicas de proteção da privacidade constitui um campo de investigação interdisciplinar, que integra fundamentos da ciência da computação, direito digital e ética da informação. A literatura nacional e internacional tem avançado na proposição de modelos teóricos, frameworks computacionais e métricas de avaliação de risco, buscando conciliar a utilidade analítica dos dados com o respeito aos direitos fundamentais (DWORK; ROTH, 2014; MALDONADO; BLUM, 2020; VIGLIAR, 2021).

A seguir são apresentados os principais trabalhos e iniciativas que dialogam com esta pesquisa, agrupados em três eixos: (i) modelos conceituais e métricos de anonimização; (ii) frameworks e ferramentas automatizadas; e (iii) abordagens jurídicas e normativas de conformidade.

2.5.1 Modelos Conceituais e Métricos de Anonimização

Os modelos conceituais de anonimização consolidaram-se a partir de abordagens estatísticas que buscaram enfrentar, de forma sistemática, o problema da reidentificação em bases de dados. Entre os marcos fundacionais mais relevantes desse campo destacam-se os modelos de *k-anonymity*, proposto por Sweeney (2002), *l-diversity*, desenvolvido por Machanavajjhala et al. (2007), e *t-closeness*, apresentado por Li, Li e Venkatasubramanian (2007), os quais introduziram estruturas formais para mensurar o grau de proteção conferido aos dados e os impactos decorrentes de sua transformação. Tais contribuições foram decisivas para o amadurecimento da anonimização como área de investigação, ao permitirem que o debate deixasse de ser exclusivamente intuitivo e passasse a apoiar-se em métricas quantitativas de risco e utilidade.

A técnica de *k-anonymity* representa um dos primeiros esforços robustos para formalizar a proteção contra reidentificação por meio da indistinguibilidade entre registros. Seu princípio central consiste em assegurar que cada combinação de quase-identificadores esteja associada a pelo menos k registros equivalentes, reduzindo a possibilidade de associação direta entre atributos e indivíduos específicos. Essa formulação exerceu forte influência sobre a literatura posterior por oferecer uma métrica clara, interpretável e operacionalizável. Ainda assim, sua relevância histórica não elimina o reconhecimento de que o modelo apresenta limitações importantes, especialmente em contextos em que grupos anonimizados mantêm forte homogeneidade nos atributos sensíveis.

Foi justamente em resposta a essas fragilidades que surgiram extensões como *l-diversity* e *t-closeness*. O modelo de *l-diversity* procurou ampliar a proteção exigindo diversidade de valores sensíveis dentro de cada grupo de equivalência, de modo a dificultar inferências diretas mesmo quando os registros compartilham quase-identificadores

semelhantes. Em seguida, *t-closeness* avançou ao incorporar uma noção de proximidade estatística entre a distribuição dos atributos sensíveis em cada grupo e a distribuição global da base, buscando reduzir riscos de inferência baseados em distorções distributivas. Em conjunto, esses modelos representam uma evolução importante do campo, pois evidenciam o esforço contínuo da literatura em aperfeiçoar mecanismos de proteção diante de cenários progressivamente mais complexos.

Entretanto, a literatura recente evidencia que a aplicação isolada dessas técnicas apresenta limitações significativas, sobretudo em ambientes caracterizados por alta dimensionalidade, grande volume de dados, forte correlação entre atributos e ampla disponibilidade de bases auxiliares para cruzamento de informações. Nesses contextos, mesmo mecanismos clássicos de anonimização podem tornar-se insuficientes para impedir ataques de homogeneidade, inferência semântica ou exploração de conhecimento prévio. Em outras palavras, a simples adoção de uma técnica isolada não garante, por si só, um nível satisfatório de proteção, especialmente quando o cenário envolve múltiplas fontes de dados e capacidades ampliadas de processamento e correlação.

Essa constatação é particularmente relevante porque desloca a discussão da anonimização de uma lógica centrada em técnicas individuais para uma perspectiva mais sistêmica. Modelos como *k-anonymity*, embora fundamentais, tornam-se vulneráveis quando empregados de forma independente, pois sua eficácia depende do comportamento estatístico da base, do perfil dos atributos envolvidos e do contexto em que os dados serão utilizados. O mesmo raciocínio se aplica a *l-diversity* e *t-closeness*, que, embora ampliem a robustez do tratamento, também não eliminam integralmente o risco de reidentificação nem asseguram, isoladamente, o equilíbrio ideal entre privacidade e utilidade.

Nesse sentido, a literatura contemporânea converge para o entendimento de que resultados mais consistentes em anonimização são obtidos a partir da combinação de múltiplas técnicas, aplicadas de forma complementar e orientada ao contexto. A articulação entre mecanismos como generalização, supressão, pseudonimização e adição de ruído permite mitigar vulnerabilidades específicas de cada abordagem individual, ao mesmo tempo em que preserva níveis adequados de utilidade analítica. Essa abordagem combinada representa um avanço em relação às estratégias isoladas, pois considera a natureza multifacetada do risco de reidentificação e a complexidade dos cenários reais de tratamento de dados.

A maturidade contemporânea da área decorre, precisamente, do reconhecimento de que a anonimização eficaz exige articulação entre técnicas, critérios contextuais e mecanismos adicionais de avaliação.

Com o avanço das aplicações de big data, inteligência artificial e aprendizado de máquina, surgiram modelos de privacidade formal que passaram a oferecer garantias matemáticas mais rigorosas. Nesse cenário, a privacidade diferencial (*differential privacy*)

tornou-se uma das abordagens mais influentes, conforme estabelecido por Dwork et al. (2014), ao definir limites formais para o vazamento informacional associado à presença ou ausência de um indivíduo em determinada base. Ao introduzir ruído calibrado e parâmetros explícitos de privacidade, essa abordagem redefiniu a forma de mensurar o equilíbrio entre precisão analítica e proteção de dados, deslocando a discussão para um terreno mais formalizado e mensurável.

A relevância da privacidade diferencial também decorre de sua capacidade de responder a desafios que ultrapassam as limitações dos modelos puramente estatísticos clássicos. Em vez de depender exclusivamente da generalização ou da supressão de atributos, essa abordagem formaliza a proteção em termos probabilísticos, permitindo controlar matematicamente o risco de exposição individual. Por essa razão, a privacidade diferencial passou a ser incorporada em ecossistemas tecnológicos de grande escala e em iniciativas de produção estatística oficial, consolidando-se como uma das referências mais importantes da literatura contemporânea sobre proteção de dados.

Paralelamente, estudos mais recentes têm explorado modelos híbridos que combinam anonimização estatística, criptografia, privacidade diferencial, síntese de dados e aprendizado federado, buscando construir mecanismos de proteção mais robustos e adaptáveis. Trabalhos como os de Taal (2022) e Soria-Chevedden et al. (2023) indicam que a integração entre múltiplas técnicas tende a ampliar a capacidade de resposta a cenários complexos de risco, nos quais uma única estratégia se mostra insuficiente. Esses resultados reforçam empiricamente a limitação das abordagens isoladas e evidenciam que a robustez da anonimização emerge da combinação estruturada de técnicas, e não da aplicação independente de um único modelo. Nessa perspectiva, a privacidade deixa de ser tratada como resultado direto da aplicação de um método isolado e passa a ser compreendida como uma propriedade emergente do sistema, construída pela combinação entre diferentes mecanismos de proteção e avaliação.

Essa mudança de perspectiva é particularmente importante para pesquisas que se situam na interseção entre ciência de dados, segurança da informação e conformidade regulatória. Em ambientes reais de tratamento de dados, os requisitos de proteção não são homogêneos, e o desempenho de uma técnica depende de fatores como a natureza dos atributos, a finalidade do uso, o grau de sensibilidade envolvido e as possibilidades de cruzamento com outras fontes. Nesse sentido, modelos híbridos oferecem maior flexibilidade analítica e operacional, pois permitem calibrar diferentes estratégias de anonimização de acordo com a complexidade do cenário. Mais do que um refinamento técnico, trata-se de uma evolução conceitual na forma de compreender a proteção da privacidade em sistemas informacionais contemporâneos.

Apesar desses avanços, a literatura ainda evidencia lacunas significativas no que se refere à aplicação contextualizada da anonimização. Grande parte dos modelos e

métricas disponíveis foi concebida em termos genéricos, sem incorporar de modo explícito as particularidades de domínios específicos, como saúde, educação, finanças ou administração pública. Essa limitação se torna ainda mais relevante quando se considera que o risco de reidentificação, a criticidade dos atributos e o impacto da perda de utilidade variam de acordo com o ambiente institucional e com a finalidade do tratamento. Assim, a simples existência de técnicas robustas não assegura, por si só, decisões adequadas em contextos concretos de uso.

Diante desse cenário, torna-se evidente a necessidade de frameworks capazes de articular modelos conceituais, métricas quantitativas e critérios contextuais de decisão. A lacuna não está apenas na ausência de técnicas, mas na dificuldade de selecionar, combinar e justificar sua aplicação de maneira coerente com a sensibilidade dos dados, com os riscos envolvidos e com os requisitos normativos do ambiente em que serão utilizados. É justamente nesse ponto que se insere a proposta desta pesquisa, ao defender uma abordagem que vá além da aplicação isolada de mecanismos de anonimização e incorpore sensibilidade contextual, avaliação automatizada de risco e apoio sistemático à decisão.

2.5.2 Frameworks e Ferramentas Automatizadas de Anonimização

Diversos frameworks e plataformas vêm sendo desenvolvidos para automatizar processos de anonimização, oferecendo suporte a empresas, órgãos públicos e pesquisadores na adequação à LGPD e a GDPR. Entre os mais relevantes, destacam-se:

- **ARX Data Anonymization Tool** (PRIVACY TOOLS PROJECT, 2021): ferramenta de código aberto que implementa técnicas de *k-anonymity*, l-diversity e *t-closeness*, permitindo análise de risco e geração de relatórios de conformidade. Embora seja amplamente utilizada, sua limitação reside na pouca adaptabilidade a contextos regulatórios não europeus e na ausência de integração com métricas de privacidade diferencial.
- **Amnesia Project** (EUROPEAN COMMISSION, 2020): desenvolvido pelo Athena Research Center, aplica anonimização hierárquica e oferece uma interface visual para generalização e supressão de dados. Seu foco, porém, permanece restrito a cenários de publicação de dados, sem contemplar fluxos dinâmicos de processamento contínuo.
- **OpenDP Initiative** (HARVARD; MICROSOFT, 2022): voltado à implementação de mecanismos de privacidade diferencial, fornece uma biblioteca modular para calibração de ruído e composição de consultas seguras. Destaca-se por sua precisão matemática e transparência, mas requer alto grau de conhecimento técnico para uso efetivo.

- **Google DP Library e TensorFlow Privacy** (GOOGLE, 2023): frameworks aplicados a modelos de aprendizado de máquina com ruído calibrado. Apesar da escalabilidade, são soluções proprietárias, com foco em contextos empresariais e pouca transparência sobre parâmetros internos de privacidade.
- **Fides e Gretel.ai** (FIDES, 2023; GRETEL, 2024): plataformas emergentes que aplicam geração de dados sintéticos e aprendizagem federada, oferecendo APIs que integram privacidade e análise preditiva. Embora inovadoras, ainda carecem de padronização normativa e rastreabilidade auditável, o que limita seu uso em ambientes regulados.

Essas iniciativas demonstram avanços significativos na operacionalização da privacidade, mas revelam uma ausência de soluções unificadas que combinem anonimização, auditoria, explicabilidade e conformidade legal. O framework proposto nesta dissertação busca preencher essa lacuna ao integrar dimensões técnicas, jurídicas e éticas em um modelo automatizado e auditável, aplicável a múltiplos contextos organizacionais.

O Quadro 5 apresenta um comparativo abrangente entre os principais frameworks e ferramentas automatizadas de anonimização identificados na literatura e em iniciativas tecnológicas recentes. Esses instrumentos representam diferentes abordagens de implementação de privacidade, variando desde modelos estatísticos clássicos, como *k-anonymity*, até mecanismos matematicamente formalizados, como a privacidade diferencial. A análise comparativa permite observar não apenas as características técnicas de cada solução, mas também suas limitações práticas em termos de adaptabilidade regulatória, escalabilidade e auditoria — aspectos fundamentais para conformidade com legislações como a LGPD e a GDPR.

Ferramentas consolidadas como o ARX *Data Anonymization Tool* (Privacy Tools Project, 2021) e o Amnesia Project (European Commission, 2020) apresentam maturidade na aplicação de técnicas clássicas de anonimização, oferecendo interfaces acessíveis e mecanismos de avaliação de risco. Entretanto, ambas ainda se mostram pouco adaptadas a cenários dinâmicos de tratamento contínuo de dados e carecem de integração com técnicas avançadas, como destacado por Sweeney (2002) e Li et al. (2007), que já apontavam limitações intrínsecas a modelos baseados apenas em generalização e supressão.

Por outro lado, iniciativas como OpenDP (Harvard & Microsoft, 2022) e as bibliotecas de privacidade diferencial do Google (Google, 2023) representam um avanço significativo ao incorporar mecanismos matemáticos robustos e calibrados, alinhados às recomendações de Dwork e Roth (2014). Essas ferramentas oferecem garantias formais de privacidade, mas apresentam elevada complexidade técnica, o que pode limitar sua adoção por organizações com baixa maturidade técnica ou infraestrutura restrita.

Plataformas mais recentes, como Fides (2023) e Gretel.ai (2024), exploram abordagens contemporâneas baseadas em dados sintéticos e aprendizado federado, dialogando com tendências descritas por Zhao et al. (2023) e Taal (2022). Apesar do potencial inovador, tais soluções ainda carecem de padronização normativa, mecanismos de *explainability* e rastreabilidade auditável — requisitos essenciais para a materialização do princípio da accountability presente na GDPR e na LGPD (BRASIL, 2018; BYGRAVE, 2022).

Assim, o comparativo evidencia um cenário em evolução, no qual cada ferramenta atende a demandas específicas, mas nenhuma oferece uma solução completa que integre análise contextual, técnicas híbridas, conformidade normativa e auditoria contínua. Essa lacuna reforça a relevância do desenvolvimento de frameworks como o CAFIS, que se propõe a unificar dimensões técnicas e jurídicas em uma abordagem sistematizada e aplicável a múltiplos contextos organizacionais.

Quadro 5 — Comparativo crítico entre frameworks e ferramentas automatizadas de anonimização

Ferramenta / Framework	**Tipo**	**Técnicas**	**Pontos fortes**	**Limitações (Análise Crítica)**
ARX Data Anonymization Tool (Privacy Tools Project, 2021)	Código aberto	k-anonymity, l-diversity, t-closeness; generalização e supressão	Interface madura; análise de risco; ampla adoção acadêmica	Não considera contexto de uso; ausência de suporte à LGPD; dependência de parametrização manual; não integra privacidade diferencial
Amnesia Project (European Commission, 2020)	Código aberto	Generalização; supressão	Interface simples; adequado para publicação de dados	Restrito a uso estático; sem análise de risco contextual; não contempla LGPD
OpenDP Initiative (Harvard & Microsoft, 2022)	Biblioteca	Privacidade diferencial	Alto rigor matemático; garantias formais	Alta complexidade; exige conhecimento técnico avançado; sem integração com governança
Google DP / TensorFlow Privacy (Google, 2023)	Framework proprietário	Privacidade diferencial em ML	Alta escalabilidade; integração com ML	Baixa transparência; dependência do ecossistema; não orientado à LGPD
Fides (2023)	Plataforma	Dados sintéticos; anonimização híbrida	Automação e integração com pipelines	Sem padronização normativa; baixa rastreabilidade; pouca explicabilidade
Gretel.ai (2024)	Plataforma comercial	Dados sintéticos; aprendizado federado	Alta qualidade; APIs modernas	Custos elevados; dependência do fornecedor; ausência de auditoria formal

Fonte: Adaptado de Privacy Tools Project (2021); European Commission (2020); Harvard e Microsoft (2022); Google (2023); Fides (2023); Gretel (2024).

3 METODOLOGIA

A presente seção tem por objetivo apresentar, de forma processual e estruturada, os procedimentos metodológicos adotados na condução da pesquisa, detalhando o percurso seguido para o desenvolvimento, implementação e avaliação do framework CAFIS. Para esse fim, adota-se o *Design Science Research* (DSR) como abordagem metodológica orientadora, não enquanto objeto de investigação em si, mas como um referencial científico capaz de estruturar a construção de artefatos voltados à solução de problemas reais em contextos computacionais complexos. Conforme Vaishnavi e Kuechler (2015), o DSR fornece um arcabouço metodológico adequado para pesquisas que buscam produzir conhecimento prescritivo por meio do design e da avaliação de artefatos.

A utilização do DSR permite articular, de maneira coerente, os principais elementos da investigação científica, incluindo a identificação do problema, a fundamentação teórica, o desenvolvimento técnico da solução e sua avaliação empírica. Segundo Hevner e Chatterjee (2010), essa abordagem é particularmente apropriada para pesquisas em Sistemas de Informação e áreas correlatas da Computação, nas quais o conhecimento emerge a partir de ciclos iterativos de construção e avaliação de artefatos, assegurando simultaneamente rigor científico e relevância prática. Nesse sentido, o DSR viabiliza o alinhamento entre as necessidades contextuais que motivam o estudo e as soluções propostas, evitando a dissociação entre teoria e aplicação.

Além disso, a natureza iterativa do DSR favorece o refinamento contínuo tanto da compreensão do problema quanto da solução desenvolvida, permitindo ajustes sucessivos ao longo do processo de pesquisa. Conforme discutido por vom Brocke, Hevner e Maedche (2020), essa característica contribui para maior maturidade do artefato final, bem como para a explicitação das decisões metodológicas adotadas durante o percurso investigativo. Assim, esta seção descreve como cada etapa metodológica foi planejada e aplicada, possibilitando compreender de forma clara e transparente o processo de concepção, aprimoramento e validação do framework de anonimização contextual apresentado neste trabalho.

3.1 Aplicação das Etapas do DSR na Pesquisa

A condução desta pesquisa foi estruturada em etapas alinhadas à lógica da DSR, com foco na identificação do problema, na construção do artefato, na fundamentação teórico-jurídica e na avaliação experimental, conforme apresentado na Figura 1. As atividades realizadas em cada etapa metodológica são descritas a seguir, organizadas conforme o fluxo adotado no desenvolvimento do framework CAFIS.

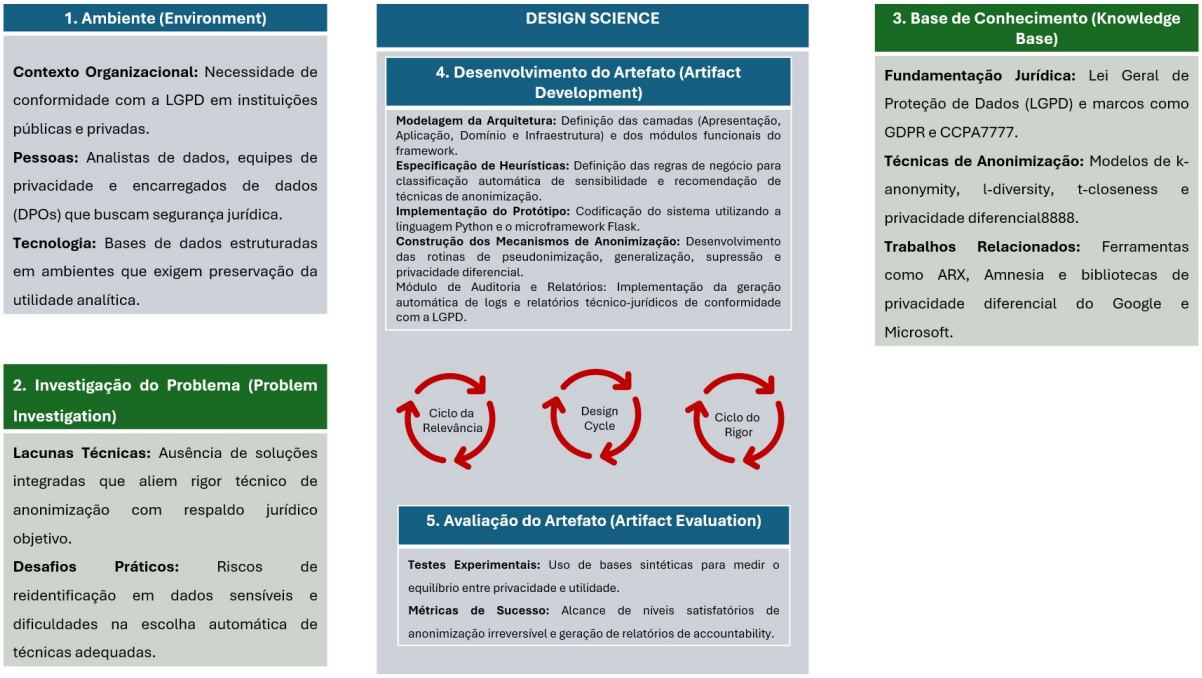


Figura 1 – Etapas da Design Science Research aplicadas ao desenvolvimento do framework CAFIS

3.1.1 Etapa de Relevância (o que foi feito)

Nesta etapa, foram analisados os desafios técnicos e jurídicos relacionados à anonimização de dados pessoais no contexto da LGPD. Realizou-se um levantamento sistemático das principais dificuldades observadas em organizações, incluindo riscos de reidentificação, ausência de critérios uniformes para classificação de sensibilidade, lacunas em processos de governança e limitações de soluções existentes. Essa análise permitiu consolidar os requisitos do framework, que incluem:

- capacidade de analisar sensibilidade de atributos;
- consideração do contexto de uso e finalidade do tratamento;
- recomendação automática de técnicas de anonimização;
- preservação da utilidade dos dados;
- geração de relatórios alinhados aos princípios da LGPD.

Esses requisitos serviram como base para o escopo funcional do CAFIS e orientaram as decisões de design e implementação.

3.1.2 Etapa de Rigor (o que fundamentou o trabalho)

A etapa de rigor consistiu na construção da base conceitual, científica e normativa que sustenta o framework. Foram integrados:

- conceitos jurídicos relativos a dados pessoais, sensíveis e anonimizados;
- princípios da LGPD relacionados a necessidade, adequação, finalidade e prevenção;
- estudos sobre risco de reidentificação, utilidade e equilíbrio entre privacidade e qualidade dos dados;
- técnicas clássicas e modernas de anonimização (k-anonymity, l-diversity, t-closeness, privacidade diferencial, criptografia homomórfica, dados sintéticos e aprendizado federado);
- comparação crítica de frameworks existentes (ARX, Amnesia, OpenDP, TensorFlow Privacy, Fides, Gretel.ai).

Essa fundamentação técnica e jurídica foi utilizada para definir métricas, parâmetros, heurísticas e critérios de decisão incorporados ao CAFIS, bem como para orientar seu processo de avaliação.

3.1.3 Etapa de Design e Construção (foco no desenvolvimento do artefato)

O desenvolvimento do CAFIS ocorreu de forma iterativa, envolvendo:

1. Modelagem da Arquitetura: definição dos módulos do framework (análise contextual, classificação de sensibilidade, motor de decisão, aplicação de técnicas, avaliação de risco/utilidade e geração de relatórios).
2. Especificação das Heurísticas: definição de regras de decisão baseadas na literatura, na LGPD e em técnicas de anonimização.
3. Implementação do Protótipo: construção do framework utilizando linguagem e bibliotecas adequadas, permitindo testes automatizados com diferentes bases de dados.
4. Integração das Métricas: inclusão de cálculos de utilidade, proximidade estatística e risco de reidentificação.
5. Geração de Relatórios: criação automática de documentos que descrevem técnicas aplicadas, parâmetros utilizados e justificativas alinhadas aos princípios jurídicos.
6. Essa etapa resultou em um protótipo funcional capaz de executar anonimização de forma contextualizada e auditável.

3.1.4 Etapa de Avaliação (como o artefato foi testado e validado)

A etapa de avaliação envolveu a aplicação do CAFIS em bases simuladas e cenários de teste representativos. Foram conduzidas as seguintes atividades:

- medição da perda de utilidade em diferentes técnicas e parâmetros;
- análise do risco residual de reidentificação;
- comparação dos resultados com métodos tradicionais;
- verificação da aderência aos princípios da LGPD;
- avaliação da clareza e completude dos relatórios gerados pelo framework.

Os resultados obtidos permitiram validar a eficácia técnica, a coerência jurídica e a aplicabilidade prática do CAFIS, além de identificar pontos de melhoria incorporados nos ciclos iterativos de refinamento.

3.2 Delimitações da Pesquisa

A presente pesquisa estabelece um conjunto rigoroso de delimitações metodológicas que orientam o escopo e garantem a consistência do trabalho. Definir esses limites é essencial para assegurar que o desenvolvimento do framework CAFIS ocorra dentro de um campo conceitual e técnico bem delimitado, permitindo que os resultados possam ser interpretados adequadamente e avaliados com precisão.

A primeira delimitação refere-se ao tipo de dado analisado. Esta investigação se restringe à anonimização de dados estruturados, isto é, informações organizadas em colunas e linhas, com atributos claros e relações formais entre variáveis. Esses dados representam a forma mais comum de armazenamento e processamento em ambientes organizacionais, abrangendo desde cadastros administrativos até sistemas corporativos amplamente utilizados.

A escolha por esse escopo específico não é arbitrária. Dados estruturados possuem maior maturidade técnica no campo da anonimização, oferecendo um conjunto consolidado de métricas, métodos e modelos teóricos que permitem análises mais aprofundadas e comparáveis. Além disso, constituem o tipo de dado que mais frequentemente fundamenta decisões organizacionais e operações reguladas, o que reforça sua relevância prática. Em contrapartida, dados não estruturados — como imagens, vídeos, áudios, textos livres e sinais brutos — foram deliberadamente excluídos deste estudo. Embora representem parcela crescente dos fluxos informacionais contemporâneos, sua anonimização apresenta desafios técnicos muito distintos, altamente dependentes de campos específicos, como visão computacional, processamento de linguagem natural e sistemas de reconhecimento de padrões.

Esses formatos carecem de padronização normativa e científica em relação a métodos de anonimização, além de exigirem abordagens tecnológicas que divergem significativamente das técnicas aplicáveis a dados tabulares. Portanto, abordá-los exigiria uma

estrutura metodológica própria, incompatível com o escopo central do CAFIS, que busca integrar privacidade, conformidade e decisão automatizada em dados estruturados.

Outra delimitação importante diz respeito às bases utilizadas nos experimentos. Para fins éticos, jurídicos e metodológicos, o trabalho utiliza apenas bases sintéticas e simuladas, construídas de forma controlada para representar características estatísticas de bases reais. Essa escolha é fundamental para evitar qualquer risco de reidentificação de indivíduos, respeitando os princípios da LGPD e assegurando integridade ética na condução da pesquisa.

O uso de bases sintéticas não compromete a validade dos resultados; ao contrário, permite o controle total sobre distribuições, tipos de atributos, comportamentos estatísticos e níveis de sensibilidade. Esse controle é essencial para testar de forma precisa o comportamento do framework sob diferentes cenários de risco e exigências de anonimização.

As bases simuladas foram construídas com diferentes composições de atributos, incluindo identificadores diretos, quase-identificadores, atributos sensíveis e atributos neutros. Essa diversidade estrutural garante que o framework possa ser avaliado de maneira abrangente, testando sua capacidade de detectar riscos, selecionar técnicas adequadas e preservar utilidade.

Além disso, a geração de dados sintéticos permite replicabilidade, uma característica essencial em trabalhos científicos. Qualquer pesquisador pode reproduzir o comportamento das bases e alcançar resultados semelhantes, o que fortalece a confiabilidade da abordagem adotada.

Há também uma delimitação regulatória explícita. As heurísticas, decisões e mecanismos implementados no CAFIS são fundamentados exclusivamente na LGPD, na GDPR e em modelos internacionais amplamente discutidos na literatura contemporânea. O framework não pretende abarcar legislações setoriais específicas, como a HIPAA (saúde nos EUA), FERPA (educação nos EUA) ou regulamentos financeiros próprios.

Essa decisão decorre da necessidade de manter coerência conceitual e jurídica com o ambiente brasileiro e europeu, que apresentam os marcos regulatórios mais amplos, estruturados e aplicáveis ao contexto geral do tratamento de dados pessoais. Incluir legislações setoriais exigiria adaptações significativas e comprometeria a universalidade do modelo proposto.

Ao se concentrar nesses marcos amplos, o framework mantém aplicabilidade geral e aderência às legislações predominantes em cenários organizacionais diversos, sem se comprometer com especificidades normativas que fogem ao propósito metodológico da pesquisa.

Também é importante explicitar que o CAFIS não busca substituir sistemas

completos de governança organizacional, auditoria jurídica ou políticas internas de conformidade. O framework atua como instrumento técnico de suporte à decisão, auxiliando analistas, equipes de privacidade e profissionais de tecnologia na escolha de técnicas de anonimização adequadas ao contexto.

Portanto, sua função é complementar, e não substitutiva. Ele oferece análises automatizadas, recomendações técnicas, avaliações de risco e relatórios estruturados, mas não se sobrepõe à necessidade de supervisão humana nem ao papel de estruturas formais de governança.

Esse posicionamento metodológico evita interpretações equivocadas sobre o propósito do artefato. O CAFIS não é uma solução regulatória completa, mas sim uma ferramenta que apoia práticas de governança já existentes, oferecendo uma camada adicional de análise técnica e contextual.

A delimitação funcional do framework também implica a exclusão de elementos não diretamente relacionados ao processo de anonimização, como gestão documental, controle de acesso em nível organizacional, classificação manual de informações ou processos gerais de compliance.

Da mesma forma, o CAFIS não realiza avaliação jurídica interpretativa nem substitui o papel de especialistas legais. Ele utiliza critérios jurídicos objetivos para embasar decisões técnicas, mas não executa análise normativa complexa ou interpretação de dispositivos legislativos em casos específicos.

Outra delimitação relevante diz respeito ao tratamento de dados sensíveis. Embora o framework seja capaz de identificar, classificar e aplicar técnicas adequadas a atributos sensíveis, ele não cobre situações extremas de hiperindividualização ou cenários que envolvem riscos éticos elevados, como dados genéticos completos ou informações biométricas altamente invasivas.

Esses casos exigem abordagens especializadas e avaliações de impacto complexas, que extrapolam o escopo deste estudo. O CAFIS é projetado para atuar em contextos amplamente encontrados no setor público e privado, como saúde administrativa, educação, cadastros, estatísticas e sistemas de gestão.

A delimitação metodológica também impacta o tipo de técnica avaliada. O framework utiliza técnicas clássicas e modernas, mas evita aprofundamento em modelos experimentais ainda pouco consolidados, como métodos totalmente probabilísticos de anonimização avançada ou técnicas extremas de homomorfismo pleno, cujo uso prático é limitado pela complexidade computacional.

Por fim, as decisões metodológicas refletem a busca por equilíbrio entre viabilidade técnica, aplicabilidade prática e rigor científico. As delimitações apresentadas não reduzem o alcance da pesquisa, mas contribuem para sua precisão e solidez. Ao estabelecer

claramente o que está dentro e fora do escopo, a metodologia assegura que o desenvolvimento e avaliação do CAFIS ocorram de maneira coerente, transparente e alinhada às exigências científicas e jurídicas contemporâneas.

Essas delimitações também favorecem a compreensão do leitor, que passa a saber exatamente quais terrenos conceituais e práticos o framework percorre, e quais permanecem como possíveis extensões para trabalhos futuros.

Elas garantem, ainda, que as conclusões alcançadas ao final da pesquisa sejam interpretadas com o devido rigor, evitando extrapolações inadequadas ou indevidas sobre contextos não analisados.

Assim, a definição cuidadosa do escopo não apenas organiza a metodologia, mas também protege a integridade científica do estudo, assegurando que o framework proposto seja avaliado de acordo com suas propostas, capacidades e limitações reais.

3.3 Bases Utilizadas

As bases de dados utilizadas nesta pesquisa foram integralmente compostas por dados sintéticos, gerados de forma controlada e orientada pelas características estatísticas observadas em bases reais. Assim, as bases sintéticas foram projetadas para manter elementos como distribuições, correlações, variabilidade e relações internas entre atributos, preservando a fidelidade estatística necessária para validação do framework.

A construção dessas bases considerou diferentes categorias de atributos, seguindo a taxonomia apresentada ao longo do referencial teórico. Foram incluídos identificadores diretos, como nome, CPF e e-mail, que representam atributos capazes de identificar imediatamente um indivíduo; quase-identificadores, como idade, CEP e sexo, que embora não identifiquem diretamente, podem levar à reidentificação quando combinados; atributos sensíveis, como condições de saúde, crenças religiosas ou opiniões políticas, cujo tratamento é estritamente regulado pela LGPD; e atributos neutros, compostos por informações administrativas ou operacionais sem potencial identificatório significativo. Essa composição diversificada permitiu simular cenários reais de risco, possibilitando a avaliação do CAFIS em diferentes níveis de sensibilidade e complexidade.

A opção por utilizar exclusivamente bases sintéticas também se justifica pela necessidade de observar rigorosamente os princípios de ética e proteção de dados, especialmente os de necessidade, prevenção e responsabilização previstos na LGPD. Ao adotar dados totalmente simulados, elimina-se qualquer possibilidade de violação de privacidade, garantindo total conformidade jurídica e permitindo a experimentação livre, transparente e reprodutível. Além disso, a literatura internacional em privacidade diferencial recomenda fortemente a utilização de dados sintéticos para testes, uma vez que esse tipo de dado possibilita avaliar técnicas de anonimização com segurança e robustez, preservando padrões

estatísticos relevantes.

Cada base foi desenvolvida para representar cenários típicos de aplicação, refletindo domínios onde técnicas de anonimização são amplamente necessárias. Foram contemplados quatro grandes contextos: saúde, reproduzindo características de registros clínicos e administrativos; educação, simulando bases estudantis e operacionais; pesquisas estatísticas, com atributos agregados e demográficos; e dados administrativos corporativos, com foco em processos internos, informações funcionais e cadastros organizacionais. Essa diversidade de cenários contribuiu para avaliar a capacidade do CAFIS de operar em condições heterogêneas, com diferentes tipos de atributos e demandas de conformidade.

Além de variados, os cenários foram construídos para refletir desafios reais enfrentados por organizações, incluindo presença de outliers, atributos altamente correlacionados, combinações de quase-identificadores e distribuições assimétricas. Essas características deram complexidade aos testes, permitindo avaliar o desempenho do framework em situações que exigem decisões automatizadas mais sofisticadas.

Por fim, o uso de dados sintéticos favoreceu a reprodutibilidade da pesquisa, possibilitando que qualquer pesquisador, organização ou avaliador externo replique os experimentos sem restrições legais. Essa característica reforça o compromisso da metodologia com transparência, auditoria e abertura científica, princípios coerentes com o propósito do CAFIS e com as orientações regulatórias sobre boas práticas de proteção de dados.

3.4 Avaliação Empírica

Os procedimentos de avaliação adotados nesta pesquisa foram estruturados com o objetivo de avaliar, de maneira organizada e controlada, a capacidade do framework CAFIS de identificar riscos, aplicar técnicas adequadas de anonimização e preservar a utilidade dos dados.

O ponto inicial do processo experimental consistiu na classificação dos atributos presentes nas bases sintéticas utilizadas. Essa etapa buscou identificar corretamente identificadores diretos, quase-identificadores, atributos sensíveis e atributos neutros, conforme definições detalhadas na fundamentação teórica. A classificação foi realizada de forma automatizada pelo CAFIS, utilizando heurísticas derivadas tanto de elementos jurídicos — especialmente da definição de dados pessoais e sensíveis prevista na LGPD — quanto de padrões técnicos amplamente aceitos, como os utilizados em modelos de risco baseados em *k-anonymity* e métricas de privacidade diferencial.

Após a etapa de classificação, iniciou-se o processo de análise de risco preliminar, no qual o framework identificou possíveis pontos de vulnerabilidade que poderiam facilitar a reidentificação. Esse processo incluiu a análise de combinações entre quase-identificadores, a avaliação de presença de atributos únicos ou raros e a inspeção de

correlações internas que pudessem contribuir para inferências não desejadas. Como indicado no texto-base, a literatura destaca que riscos de reidentificação não decorrem apenas da presença de identificadores explícitos, mas também de padrões estatísticos capazes de revelar perfis individuais.

Com a avaliação inicial de riscos realizada, o CAFIS passou à aplicação das técnicas de anonimização mais adequadas ao contexto de cada base e ao perfil de risco de cada atributo. Esse processo seguiu os critérios definidos previamente na modelagem do framework, contemplando tanto técnicas clássicas — como generalização, supressão, *k-anonymity*, *l-diversity* e *t-closeness* — quanto técnicas mais avançadas, como mecanismos de ruído calibrado e métodos inspirados em privacidade diferencial.

A seleção da técnica não ocorreu de forma arbitrária; ao contrário, foi orientada por uma matriz decisória interna que leva em conta sensibilidade, finalidade do uso, impacto esperado na utilidade dos dados e nível de risco residual permitido. Essa matriz decisória é parte fundamental do CAFIS, concebida para fornecer recomendações contextualizadas, indo além das abordagens puramente automatizadas encontradas em ferramentas tradicionais de anonimização.

Após a aplicação das técnicas, iniciou-se a fase de avaliação do impacto produzido sobre a utilidade dos dados. Essa etapa foi essencial para verificar se a anonimização preservou, de maneira satisfatória, o valor informacional e o potencial analítico das bases. Foram aplicadas métricas estatísticas de proximidade, variação de distribuição, mudanças estruturais e perda de granularidade, todas reconhecidas pela literatura como métodos adequados para mensurar o equilíbrio entre privacidade e utilidade.

Paralelamente à mensuração da utilidade, foi calculado o risco residual de reidentificação, etapa necessária para verificar se o processo de anonimização atingiu níveis adequados de segurança. Esse cálculo incorporou tanto métricas tradicionais, como risco de registro único e probabilidade de ligação, quanto métricas derivadas de modelos probabilísticos, sempre respeitando os limites da base sintética e as características do cenário testado.

A partir dos resultados obtidos, iniciou-se o processo de refinamento do framework. Cada ciclo gerou relatórios contendo informações sobre as técnicas aplicadas, parâmetros utilizados, desempenho obtido e eventuais limitações observadas. Esses relatórios foram essenciais para aperfeiçoar heurísticas, ajustar limiares e revisar regras de decisão, permitindo que o CAFIS evoluísse progressivamente até alcançar um equilíbrio adequado entre precisão técnica, segurança jurídica e viabilidade prática.

Todas as anáises foram conduzidos de forma repetível e transparente, com registro dos parâmetros utilizados e documentação dos resultados obtidos, assegurando a reprodutibilidade científica necessária. Esse compromisso com a transparência metodoló-

gica é coerente não apenas com boas práticas de pesquisa, mas também com os princípios de privacidade e responsabilização previstos na LGPD, reforçando o caráter ético da investigação.

Assim, a avaliação empírica adotada permitiram avaliar o comportamento do CAFIS em múltiplos cenários, fornecendo uma análise robusta sobre sua eficácia no tratamento de dados estruturados e validando sua capacidade de servir como um instrumento confiável de suporte à decisão em processos de anonimização.

3.5 Critérios de Avaliação do Framework

A avaliação do framework CAFIS foi estruturada a partir de uma abordagem multidimensional, contemplando aspectos técnicos, jurídicos e éticos, em conformidade com a Lei Geral de Proteção de Dados (Lei nº 13.709/2018). Essa abordagem busca garantir que o processo de anonimização não apenas reduza o risco de identificação dos titulares, mas também preserve a utilidade analítica dos dados e mantenha aderência aos princípios regulatórios. Conforme discutido por Fung et al. (2010), a anonimização de dados envolve um trade-off inerente entre privacidade e utilidade, sendo necessária a utilização de métricas específicas para avaliar esse equilíbrio.

3.5.1 Critérios Técnicos

Os critérios técnicos têm como objetivo avaliar a eficiência computacional do CAFIS, bem como sua capacidade de preservar a qualidade dos dados e reduzir o risco de reidentificação. Esses critérios são fundamentais para validar a aplicabilidade do framework em cenários reais de tratamento de dados.

3.5.1.1 Tempo de Execução

O tempo de execução corresponde ao tempo médio necessário para a realização do processo de anonimização. Essa métrica permite avaliar a eficiência computacional e a escalabilidade do framework, especialmente em bases de dados de grande volume.

Formalmente, o tempo médio de execução pode ser representado por:

$$T_{exec} = \frac{1}{n} \sum_{i=1}^{n} t_i$$

em que t_i representa o tempo de execução em cada experimento e n o número total de execuções realizadas.

A análise desse indicador permite verificar se o CAFIS apresenta desempenho compatível com aplicações práticas, evitando que a complexidade das técnicas de anonimização inviabilize sua adoção em ambientes reais.

3.5.1.2 Preservação da Utilidade dos Dados

A preservação da utilidade dos dados refere-se à capacidade do processo de anonimização manter propriedades estatísticas relevantes do conjunto de dados original. Esse aspecto é essencial para garantir que os dados anonimizados permaneçam úteis para análise, em conformidade com o princípio da adequação previsto na LGPD.

Para avaliar essa dimensão, são utilizadas métricas consolidadas na literatura.

Uma das principais métricas é o erro quadrático médio (RMSE), amplamente empregado para medir a diferença entre valores originais e anonimizados:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}$$

onde x_i representa o valor original, $\hat{x}_i$ o valor anonimizado e N o número de registros. Valores menores de RMSE indicam maior preservação da utilidade dos dados.

Além disso, a preservação das relações entre atributos pode ser avaliada por meio do coeficiente de correlação de Pearson:

$$\rho = \frac{Cov(X, \hat{X})}{\sigma_X \cdot \sigma_{\hat{X}}}$$

Essa métrica permite verificar se a estrutura relacional entre variáveis é mantida após a anonimização.

Para atributos categóricos, a perda de informação pode ser estimada com base em medidas de entropia, conforme proposto por Shannon (1948):

$$I(D) = -\sum p(x) \log p(x)$$

A partir dessa definição, a perda de informação pode ser expressa como:

$$IL = 1 - \frac{I(D_{anon})}{I(D_{orig})}$$

Valores próximos de zero indicam menor perda informacional, preservando a diversidade dos dados.

3.5.1.3 Risco de Reidentificação

A avaliação do risco de reidentificação é central no contexto da anonimização de dados, estando diretamente relacionada ao conceito de anonimização definido na LGPD (Art. 5º, inciso XI). Esse critério busca verificar se os dados tratados não permitem a identificação de indivíduos, considerando meios técnicos razoáveis disponíveis.

Para essa análise, o CAFIS utiliza modelos clássicos da literatura.

O modelo de *k-anonymity*, proposto por Sweeney (2002), estabelece que cada registro deve ser indistinguível de pelo menos $k-1$ outros registros:

$$R_k = \min_{r \in D} |EC(r)|$$

onde $EC(r)$ representa a classe de equivalência do registro r. Esse modelo reduz o risco de identificação direta, garantindo anonimato por agrupamento.

Complementarmente, o modelo de *l-diversity*, apresentado por Machanavajjhala et al. (2007), exige diversidade nos atributos sensíveis dentro de cada grupo:

$$l = \min_{EC \in D} |\{valores\ sensíveis(EC)\}|$$

Esse critério evita ataques por inferência, nos quais informações sensíveis podem ser deduzidas mesmo em dados aparentemente anonimizados.

Por fim, o modelo de *t-closeness*, proposto por Li, Li e Venkatasubramanian (2007), avalia a proximidade entre a distribuição dos atributos sensíveis em cada grupo e a distribuição global do conjunto de dados:

$$t = \max_{EC \in D} d(P(EC), P(D))$$

em que d representa uma medida de distância entre distribuições, frequentemente baseada na Earth Mover's Distance (EMD). Valores menores indicam maior preservação da distribuição original e menor risco de inferência.

Esses modelos permitem avaliar diferentes dimensões do risco de reidentificação, sendo utilizados no CAFIS como critérios para validação da efetividade das técnicas de anonimização.

3.5.2 Critérios Jurídicos

Os critérios jurídicos avaliam o alinhamento do framework com os princípios e obrigações da LGPD. Entre os principais aspectos considerados estão os princípios da finalidade, adequação, necessidade, segurança e responsabilização, previstos no Art. 6º da referida legislação.

O CAFIS incorpora esses princípios ao estruturar o processo de anonimização de forma controlada e auditável, garantindo que o tratamento de dados ocorra dentro de limites legais bem definidos. Conforme argumenta Doneda (2020), a proteção de dados pessoais envolve não apenas a preservação da privacidade, mas também o controle sobre o fluxo informacional e sua utilização.

Além disso, o framework prevê a geração de evidências de conformidade, como relatórios técnicos e registros de processamento, alinhando-se ao princípio da accountability. A utilização do Relatório de Impacto à Proteção de Dados (RIPD), previsto no Art. 38 da LGPD, reforça a capacidade do sistema de demonstrar conformidade regulatória.

3.5.3 Critérios Éticos

Os critérios éticos complementam a avaliação técnica e jurídica, considerando os impactos sociais do tratamento de dados e a necessidade de proteção dos direitos fundamentais dos titulares. Esses critérios estão relacionados à prevenção de danos, à transparência e à equidade no uso dos dados.

No contexto do CAFIS, a dimensão ética manifesta-se na preocupação com a redução de riscos de discriminação e uso indevido dos dados, bem como na garantia de que o processo de anonimização não introduza vieses que comprometam análises futuras.

Adicionalmente, o framework promove rastreabilidade e auditabilidade das operações realizadas, permitindo revisão e justificativa das decisões tomadas. Essa característica fortalece a confiança no uso de dados e contribui para uma governança responsável e transparente, alinhada às demandas contemporâneas de proteção de dados.

3.6 Tecnologias Utilizadas

O desenvolvimento do framework CAFIS exigiu a adoção de um conjunto de tecnologias que garantissem robustez computacional, capacidade analítica e reprodutibilidade científica. A escolha dessas tecnologias foi orientada tanto pelas necessidades técnicas da pesquisa quanto pelas recomendações internacionais para implementação de sistemas de anonimização e mecanismos de privacidade. Nesse sentido, as ferramentas selecionadas foram utilizadas de forma integrada, permitindo a construção de um ambiente de experimentação seguro, flexível e alinhado às melhores práticas em ciência de dados.

A linguagem adotada para implementação do CAFIS foi Python, devido à sua ampla aceitação acadêmica, diversidade de bibliotecas especializadas e forte ecossistema voltado à análise de dados, estatística e experimentação computacional. Python oferece não apenas flexibilidade e facilidade de prototipação, mas também módulos consolidados que facilitam a manipulação de bases, a aplicação de algoritmos de anonimização e o cálculo de métricas.

Para operações estatísticas, análise de distribuições e avaliação de risco, foram utilizadas bibliotecas como NumPy, Pandas e SciPy. Essas ferramentas forneceram os fundamentos matemáticos necessários para manipular grandes volumes de dados sintéticos, calcular padrões de correlação, estimar variabilidade e medir a proximidade entre bases

originais e anonimizadas. A reprodutibilidade dos experimentos depende diretamente dessas estruturas, que permitem que cada etapa seja registrada e repetida com consistência.

Técnicas de anonimização clássicas, como generalização, supressão e implementação de k-anonymity, foram apoiadas por bibliotecas inspiradas em ferramentas como ARX e Amnesia — discutidas nas páginas referentes à comparação entre frameworks no documento. Embora o CAFIS não dependa diretamente dessas ferramentas, a experiência acumulada por esses projetos fundamentou a implementação das rotinas internas do framework, especialmente no que diz respeito a estratégias de agregação hierárquica e definição de níveis de granularidade.

Para mecanismos inspirados em privacidade diferencial, foram utilizados módulos que implementam funções de ruído calibrado, com destaque para bibliotecas que seguem os princípios do OpenDP e do TensorFlow Privacy. Embora o framework não implemente integralmente modelos de aprendizado profundo ou privacidade diferencial avançada, utiliza funções analíticas baseadas em Laplace e Gauss para simular comportamentos típicos dessas abordagens. Essa escolha garante coerência científica sem comprometer a simplicidade e interpretabilidade necessárias ao protótipo.

A geração de dados sintéticos é fundamental para a condução de testes éticos e fez uso de algoritmos probabilísticos e modelos inspirados em técnicas de simulação estatística. Foram aplicadas combinações de distribuições normais, exponenciais, binomiais e mistas, com o objetivo de criar bases realistas que reproduzissem correlações internas observáveis em cenários reais. A literatura presente nas páginas 18 a 20 do documento reforça a importância de dados sintéticos como recurso para evitar riscos e, simultaneamente, permitir experimentação.

Para garantir rastreabilidade e auditabilidade, o CAFIS incorporou ferramentas de registro automático, permitindo armazenar logs de decisões, parâmetros utilizados e técnicas aplicadas. Esses registros não apenas reforçam o princípio de accountability previsto na LGPD, mas também proporcionam ao pesquisador e às organizações um instrumento útil de auditoria técnica. Em termos práticos, foram utilizadas estruturas próprias da linguagem Python para logging, complementadas por formatos padronizados de exportação como JSON e CSV.

A avaliação das técnicas aplicadas exigiu mecanismos adicionais de comparação e visualização. Bibliotecas como Matplotlib e Seaborn foram utilizadas para representar graficamente distribuições, perdas de utilidade e padrões de correlação antes e depois da anonimização. Visualizações coerentes são fundamentais para interpretar resultados de maneira intuitiva, reforçando a importância da integração entre análise quantitativa e interpretação qualitativa.

O ambiente de execução do framework foi configurado de modo a assegurar

previsibilidade computacional. As experiências foram conduzidas em ambiente local controlado, utilizando máquinas com capacidade suficiente para manipulação de dados sintéticos de médio porte. A escolha por um ambiente local, e não distribuído, está alinhada ao caráter prototípico do CAFIS e à necessidade de garantir reprodutibilidade total dos experimentos.

A integração entre todas essas tecnologias foi projetada para manter o caráter modular do framework. Cada bloco funcional — classificação, decisão, anonimização, avaliação e relatório — foi implementado de forma independente, possibilitando que pesquisadores futuros possam substituir ou aprimorar componentes específicos sem comprometer a estrutura geral. Essa modularidade está em sintonia com tendências identificadas na literatura contemporânea, que enfatizam a importância de arquiteturas flexíveis para sistemas de privacidade.

Assim, o conjunto de tecnologias adotado não apenas viabiliza o funcionamento do CAFIS, mas também assegura que ele permaneça fiel aos princípios científicos, jurídicos e metodológicos que fundamentam o trabalho. A combinação de linguagem acessível, bibliotecas robustas e ambiente transparente constitui uma base sólida para continuidade da pesquisa, permitindo que o framework evolua e seja aplicado em novos cenários organizacionais.

3.7 Bases de Dados Utilizadas no Estudo

A avaliação empírica do framework CAFIS foi conduzida por meio da aplicação controlada do artefato em bases de dados sintéticas, especialmente elaboradas para fins acadêmicos, experimentais e metodológicos. A adoção desse tipo de base constitui uma decisão deliberada, fundamentada tanto em critérios técnicos quanto jurídicos, assegurando a conformidade com os princípios éticos da pesquisa científica e com o arcabouço normativo estabelecido pela Lei Geral de Proteção de Dados Pessoais (LGPD).

Sob a perspectiva jurídica, a utilização de dados reais, especialmente em domínios sensíveis como saúde, educação ou relações de consumo, implicaria riscos significativos de violação aos direitos dos titulares, além de demandar bases legais específicas, consentimento informado, mecanismos adicionais de governança e autorizações institucionais. A LGPD estabelece que dados pessoais sensíveis exigem salvaguardas reforçadas e tratamento estritamente necessário, o que poderia deslocar o foco da pesquisa do objetivo central — a avaliação do framework — para aspectos procedimentais e administrativos. Dessa forma, o uso de bases sintéticas elimina qualquer possibilidade de identificação direta ou indireta de indivíduos, afastando completamente o risco de reidentificação e garantindo que os experimentos realizados estejam fora do escopo de incidência material da legislação de proteção de dados.

Do ponto de vista técnico, as bases sintéticas possibilitam maior controle experimental sobre variáveis críticas para a análise de anonimização, tais como distribuição de atributos, presença de quase-identificadores, níveis de sensibilidade e complexidade contextual. Diferentemente de bases reais, que refletem contingências específicas e, muitas vezes, não documentadas, os dados simulados permitem a construção de cenários parametrizáveis, nos quais é possível observar de forma padronizada o comportamento do CAFIS frente a diferentes configurações de risco, finalidade e contexto de uso. Essa característica é particularmente relevante em pesquisas fundamentadas na Design Science Research, nas quais a avaliação do artefato deve ocorrer de maneira controlada, reproduzível e orientada a objetivos claramente definidos.

A base principal utilizada no estudo corresponde a um cenário sintético de pacientes hospitalares, concebido para representar um ambiente de alta sensibilidade informacional. Essa base foi escolhida como eixo central da avaliação por reunir características típicas de sistemas de saúde, tais como dados pessoais, dados sensíveis e atributos quase-identificadores, exigindo decisões rigorosas quanto à classificação dos dados, à avaliação de sensibilidade e à seleção das técnicas de anonimização. Embora sua estrutura reflita padrões realistas de sistemas hospitalares, nenhum registro corresponde a pacientes reais, sendo todos os dados integralmente gerados de forma artificial.

Complementarmente, foram utilizadas bases sintéticas adicionais que representam outros contextos organizacionais, como ambientes comerciais, educacionais e demográficos. Essas bases não foram empregadas com o objetivo de comparação estatística entre conjuntos de dados, mas sim para avaliar a capacidade adaptativa do CAFIS frente a diferentes cenários de tratamento de dados. Tal abordagem permite observar como o framework ajusta seus mecanismos de análise contextual, classificação de atributos, avaliação de sensibilidade e geração de indicadores de conformidade conforme o domínio de aplicação e o perfil informacional da base analisada.

O uso de múltiplas bases com níveis distintos de sensibilidade possibilita demonstrar que o CAFIS não opera com regras fixas ou genéricas, mas adota uma lógica contextualizada, orientada pelos princípios da LGPD, tais como finalidade, necessidade, adequação, prevenção e responsabilização. Essa característica reforça o caráter dinâmico do framework e sua aderência à concepção contemporânea de conformidade baseada em risco.

O Quadro 5 apresenta uma síntese das bases de dados utilizadas na avaliação do CAFIS, destacando o contexto de aplicação, o tipo de dados predominante e o nível de sensibilidade informacional associado a cada cenário analisado.

Quadro 5 – Síntese das bases de dados utilizadas na avaliação do CAFIS

Base de Dados	**Contexto de Aplicação**	**Tipo de Dados Predominante**	**Nível de Sensibilidade**
Pacientes Hospitalares (sintética)	Saúde	Dados pessoais e dados sensíveis simulados	Alto
Clientes de Loja (sintética)	Comercial	Dados pessoais não sensíveis e dados transacionais	Médio
Alunos Universitários (sintética)	Educacional	Dados pessoais e dados acadêmicos simulados	Médio
Base Demográfica (sintética)	Estatístico / Público	Dados agregados e demográficos	Baixo
Cenários Parametrizados de Proteção	Experimental	Combinação controlada de atributos e níveis de risco	Variável

Fonte: Elaboração própria.

A utilização combinada dessas bases permitiu demonstrar que o CAFIS é capaz de operar de forma consistente em diferentes domínios, respeitando os limites jurídicos da anonimização e, simultaneamente, preservando a utilidade analítica dos dados. Ao empregar exclusivamente bases sintéticas, o estudo reforça seu compromisso com a ética da pesquisa, com a segurança da informação e com a conformidade regulatória, ao mesmo tempo em que assegura rigor metodológico e reprodutibilidade científica.

Essa estratégia metodológica evidencia que os resultados obtidos não dependem de particularidades de um conjunto específico de dados, mas refletem o comportamento estrutural do framework enquanto artefato computacional, validando sua aplicabilidade em contextos organizacionais reais, desde que observadas as salvaguardas legais e técnicas pertinentes.

4 FRAMEWORK CAFIS

O CAFIS — *Context-Aware Framework for Information Sensitivity* — constitui o principal produto técnico desenvolvido nesta pesquisa. Trata-se de um framework concebido para auxiliar organizações, equipes de privacidade e analistas de dados na identificação contextualizada da sensibilidade de atributos, na avaliação de riscos de reidentificação e na recomendação automatizada de técnicas de anonimização alinhadas às exigências legais da Lei Geral de Proteção de Dados Pessoais (LGPD) e de marcos internacionais como a GDPR. O framework organiza, de forma integrada, elementos técnicos, jurídicos e estatísticos que historicamente têm sido analisados de maneira fragmentada, fornecendo uma interface de apoio à decisão fundamentada em modelos formais de risco e critérios legais.

O propósito central do CAFIS é disponibilizar um instrumento técnico capaz de orientar a anonimização com base no contexto de uso da informação, levando em conta não apenas a natureza intrínseca dos dados, mas também finalidades, relações entre atributos, níveis de sensibilidade, estruturas estatísticas e requisitos regulatórios. Com isso, o framework busca preencher uma lacuna identificada na literatura e nas práticas organizacionais: a carência de soluções que integrem análise jurídica, classificação contextual, métricas de privacidade e mecanismos automatizados de recomendação em um único ambiente interpretável e auditável.

4.1 Definição Geral do Artefato

O CAFIS é definido como um framework computacional de apoio à decisão, capaz de:

- classificar dados conforme tipologias jurídicas e técnicas;
- avaliar sensibilidade contextual;
- estimar riscos de reidentificação;
- recomendar técnicas adequadas de anonimização;
- aplicar e testar essas técnicas em bases sintéticas;
- gerar relatórios interpretáveis que documentem todo o processo.

Ele não substitui modelos de governança, políticas organizacionais ou análise jurídica especializada, mas complementa esses mecanismos ao oferecer uma camada objetiva, reproduzível e tecnicamente fundamentada para orientar decisões sobre tratamento de dados.

4.2 Visão de Uso do Sistema

O uso de sistemas voltados à anonimização de dados não se resume à execução de procedimentos técnicos sobre uma base informacional. Em contextos organizacionais e acadêmicos, esse uso envolve decisões relacionadas à finalidade do tratamento, ao nível de sensibilidade dos atributos, ao risco de reidentificação e à necessidade de preservar a utilidade analítica dos dados. Desse modo, a interação entre usuário e sistema torna-se um elemento relevante da própria proposta metodológica, uma vez que a qualidade do processo de anonimização depende não apenas dos mecanismos implementados, mas também da forma como o sujeito conduz, valida e interpreta cada etapa. É nesse sentido que a Figura 2 apresenta a visão de uso do CAFIS, evidenciando o percurso de interação estabelecido entre o usuário e o sistema ao longo do processo.

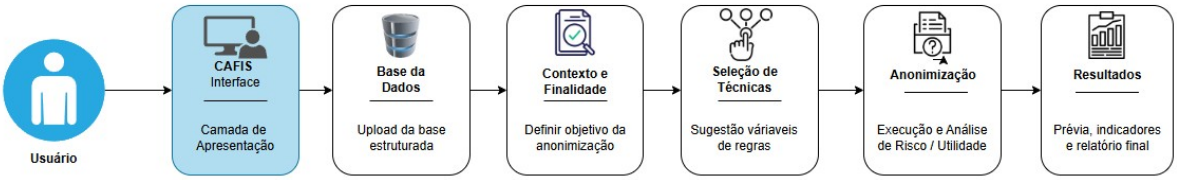


Figura 2 – Visão de uso do sistema CAFIS. Representação do fluxo de interação do usuário com o CAFIS, desde a inserção da base de dados e definição dos objetivos de anonimização até a execução do processo e a apresentação dos resultados, indicadores e relatório final.

Os usuários que interagem com o CAFIS correspondem, em geral, a profissionais e pesquisadores que atuam em atividades relacionadas ao tratamento, análise, compartilhamento ou governança de dados. Entre esses atores, destacam-se analistas de dados, pesquisadores, profissionais de tecnologia da informação, encarregados de proteção de dados pessoais, equipes de privacidade, gestores institucionais e, em certos contextos, servidores públicos ou profissionais administrativos responsáveis pela manipulação de bases estruturadas. Ainda que apresentem perfis distintos, esses usuários compartilham uma mesma necessidade: utilizar dados de forma tecnicamente segura, juridicamente adequada e metodologicamente justificável, sobretudo em cenários marcados por exigências de conformidade regulatória.

A jornada representada na figura inicia-se no próprio usuário, o que revela uma escolha conceitual importante. Antes de qualquer processamento automatizado, existe uma intenção humana que motiva a utilização do sistema. Há sempre uma demanda concreta que antecede a execução técnica: preparar dados para pesquisa, viabilizar compartilhamento institucional, reduzir riscos de exposição, atender critérios de conformidade ou organizar informações para fins analíticos. Assim, o fluxo não parte da base de dados nem das técnicas de anonimização, mas do sujeito que reconhece a necessidade de tratar

a informação de maneira responsável e controlada.

Na sequência, o usuário acessa a interface do CAFIS, correspondente à camada de apresentação. Essa etapa representa o primeiro contato operacional com a solução e possui função mediadora entre a complexidade interna do framework e a experiência prática de uso. A interface organiza as ações do usuário, apresenta as possibilidades disponíveis e estrutura o encadeamento das etapas posteriores. Sob essa perspectiva, ela não deve ser compreendida apenas como componente visual, mas como ambiente em que a lógica técnica do sistema é convertida em interação compreensível, permitindo que diferentes perfis profissionais utilizem o CAFIS de modo orientado.

Após esse ingresso no sistema, o usuário realiza o carregamento da base de dados estruturada. Esse momento marca a passagem de uma necessidade abstrata de anonimização para uma operação concreta sobre um conjunto específico de informações. A base inserida pode conter identificadores diretos, quase-identificadores, atributos sensíveis e variáveis neutras, exigindo que o sistema passe a operar sobre um cenário real de exposição informacional. Para o usuário, essa etapa é decisiva porque define o material empírico que será submetido às análises e recomendações do framework, vinculando a jornada de uso a um problema efetivo de tratamento de dados.

Em seguida, o fluxo se desloca para a definição do contexto e da finalidade da anonimização. Essa etapa adquire especial importância porque impede que o processo seja conduzido de forma genérica ou descontextualizada. O usuário precisa explicitar o objetivo do tratamento, indicando se a base será utilizada para pesquisa, análise estatística, testes computacionais, compartilhamento institucional ou outras finalidades legítimas. Essa informação orienta o comportamento do sistema e reforça o entendimento de que a anonimização não constitui um procedimento uniforme, mas uma prática dependente do contexto de uso e das exigências associadas à proteção dos titulares.

A etapa de seleção de técnicas evidencia uma das principais contribuições do CAFIS para a experiência do usuário: a oferta de suporte estruturado à decisão. Em vez de exigir que o operador defina sozinho, e de forma inteiramente manual, quais mecanismos devem ser aplicados, o sistema apresenta sugestões baseadas em variáveis, regras e parâmetros relacionados ao cenário informado. Para o usuário, isso significa atuar em um ambiente mais seguro e racionalizado, no qual as escolhas deixam de ser intuitivas ou dispersas e passam a ser apoiadas por critérios sistematizados. Tal característica é particularmente relevante para contextos em que nem todos os usuários possuem especialização aprofundada em técnicas de anonimização.

Superada a etapa decisória, o usuário acompanha a execução da anonimização e da análise de risco e utilidade. Esse momento representa a materialização do processamento técnico propriamente dito, no qual a base original é transformada segundo estratégias selecionadas ou recomendadas pelo sistema. Sob a perspectiva da jornada de uso,

trata-se da fase em que o usuário percebe o valor operacional do CAFIS, uma vez que o framework converte parâmetros e objetivos previamente informados em ações concretas de proteção e avaliação. Mais do que executar alterações sobre os dados, o sistema produz uma resposta estruturada ao problema inicial apresentado pelo usuário.

O estágio final da jornada é representado pela apresentação dos resultados. Nesse ponto, o usuário recebe uma devolutiva composta por prévia da base tratada, indicadores e relatório final. Essa entrega é relevante porque amplia a compreensão do processo realizado e favorece a interpretação dos efeitos produzidos pela anonimização. O usuário não recebe apenas um arquivo modificado, mas também elementos que permitem avaliar o impacto das técnicas adotadas, refletir sobre o risco residual e considerar o grau de preservação da utilidade dos dados. A experiência de uso, assim, não se encerra com a execução automática do processamento, mas se estende à análise crítica dos resultados disponibilizados.

Ao representar esse percurso de forma encadeada, a Figura 2 evidencia que o CAFIS foi concebido para acompanhar o usuário em uma sequência progressiva de ações, decisões e validações. Cada etapa do fluxo responde a uma necessidade prática do tratamento de dados e busca reduzir a distância entre a complexidade técnica da anonimização e a realidade operacional dos profissionais que lidam com informação estruturada. O sistema, portanto, não se limita a automatizar tarefas, mas organiza uma experiência de uso orientada por contexto, finalidade e responsabilidade.

4.3 Arquitetura do Framework

A arquitetura do framework foi concebida com base no paradigma orientado a camadas (layered architecture), amplamente adotado em sistemas corporativos e científicos por favorecer separação de responsabilidades, modularidade e escalabilidade. Essa escolha metodológica está alinhada às boas práticas descritas por Pressman e Maxim (2020) e Sommerville (2019), que destacam a importância da estrutura em camadas para a manutenção de sistemas complexos e a mitigação de riscos relacionados ao acoplamento entre componentes.

O modelo proposto, ilustrado na Figura 3, organiza-se em quatro níveis principais — Camada de Apresentação, Camada de Aplicação, Camada de Domínio e Camada de Infraestrutura — cada um com funções, fluxos e artefatos específicos. Essa estrutura reflete a filosofia de Clean Architecture de Robert C. Martin (2019), na qual o fluxo de dependências é unidirecional, das camadas externas para as internas, garantindo que o núcleo de negócio permaneça independente de frameworks, bibliotecas e detalhes de implementação.

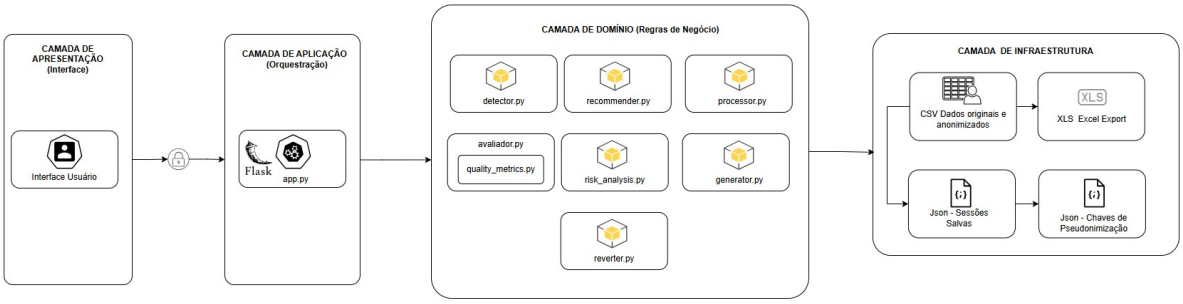


Figura 3 – Arquitetura CAFIS

4.3.1 Camada de Apresentação (Interface)

A Camada de Apresentação é responsável pela interação entre o usuário e o sistema, servindo como ponto de entrada para todas as requisições do framework. Ela foi desenvolvida com ênfase em usabilidade, clareza e acessibilidade, seguindo princípios de user-centered design e recomendações da norma ISO 9241-210 sobre ergonomia da interação humano-computador.

A interface web foi implementada utilizando o microframework Flask, que atua como intermediário entre o front-end e a lógica de aplicação. Nessa camada, são tratadas as entradas do usuário, validações básicas e chamadas de rotas HTTP definidas no módulo principal app.py. O usuário pode carregar arquivos CSV contendo dados sensíveis, configurar parâmetros de anonimização e solicitar a geração de relatórios técnicos e jurídicos.

Do ponto de vista da segurança, essa camada adota práticas de sanitização de entrada e controle de sessão para evitar injeções maliciosas e garantir que as interações permaneçam dentro do contexto autorizado. Todas as comunicações são encapsuladas por meio de um canal seguro, estabelecendo o primeiro nível de defesa do framework.

4.3.2 Camada de Aplicação (Orquestração)

A Camada de Aplicação desempenha o papel de núcleo de orquestração do sistema, sendo responsável por coordenar o fluxo de execução entre a interface e o domínio. Ela implementa o padrão arquitetural que separa claramente as responsabilidades de controle, visualização e manipulação de dados.

O módulo app.py concentra as rotas e o controle de dependências, mapeando as requisições do usuário para os componentes adequados da camada de domínio. Essa camada também realiza a gerência de estado da aplicação, controle de erros e logging de eventos, permitindo rastrear todas as ações executadas. Além da função de mediação, essa camada é responsável por garantir a consistência transacional entre os módulos, coordenando a execução sequencial das operações: carregamento de dados → aplicação

das técnicas → cálculo de métricas → geração de relatório. Esse fluxo é sustentado por um pipeline orquestrado que mantém a integridade dos dados anonimizados e os resultados de cada operação registrados em log para futura auditoria.

4.3.3 Camada de Domínio (Regras de Negócio)

A Camada de Domínio constitui o núcleo lógico e conceitual do framework, concentrando as regras de negócio, políticas de anonimização e módulos de análise. Essa camada é completamente independente das tecnologias externas, conforme recomendações de Martin (2019), permitindo que as decisões de negócio e algoritmos permaneçam isolados das camadas de interface e persistência.

4.3.4 Descrição dos Módulos

O sistema é composto por um conjunto de módulos que se organizam em três grandes camadas: (i) interface e orquestração, (ii) serviços de anonimização e (iii) avaliação e conformidade. A seguir, descrevem-se em detalhe os principais componentes.

O arquivo app.py representa a camada de apresentação e orquestração. Desenvolvido em Flask, ele implementa todas as rotas do sistema, estruturando o fluxo interativo em etapas sequenciais: upload do arquivo de dados, detecção de atributos, recomendação de técnicas, aplicação de anonimização, visualização de métricas, geração de relatórios, exportação de resultados e, em casos autorizados, reversão de pseudonimização. Este módulo é responsável também pelo controle de perfis de acesso, diferenciando administradores, que possuem maior autoridade, de analistas, que têm permissões restritas, pela persistência de dados temporários e pelo gerenciamento das sessões do usuário. Assim, app.py atua como núcleo orquestrador, conectando os módulos técnicos às necessidades de interação do usuário.

O arquivo detector.py constitui o primeiro ponto de contato com os dados submetidos. Ele realiza a classificação automática das colunas com base em regras léxicas e dicionários cuidadosamente elaborados, categorizando atributos em identificadores diretos, quase-identificadores, dados sensíveis e não sensíveis. Essa etapa é fundamental, pois fornece o insumo inicial para o processo de anonimização, estabelecendo quais atributos devem ser tratados com técnicas mais rigorosas, de acordo com sua relevância para a proteção da identidade do titular.

Na sequência, o arquivo recommender.py implementa o mecanismo de recomendação automática de técnicas de anonimização. Baseado em heurísticas explicáveis, este módulo sugere, por exemplo, o uso de pseudonimização com chave para identificadores diretos, supressão ou privacidade diferencial para dados sensíveis, e microagregação ou generalização para quase-identificadores. Essa camada assegura que as recomendações

sejam não apenas técnicas, mas também passíveis de justificativa em relatórios, atendendo ao princípio da explicabilidade.

O arquivo `processor.py` é responsável por operacionalizar as recomendações, aplicando efetivamente as técnicas sobre o conjunto de dados. As funcionalidades implementadas incluem pseudonimização simples e com chave, generalização por faixas ou categorias, hashing e supressão de valores, privacidade diferencial (com injeção de ruído laplaciano em atributos numéricos) e microagregação.

No caso da pseudonimização com chave, o sistema gera e armazena arquivos JSON contendo os mapeamentos entre valores originais e tokens, permitindo eventual reversão sob controle. Um segundo módulo, `pseudonimizador_com_chave.py`, oferece uma abordagem alternativa baseada em tokens aleatórios alfanuméricos, reforçando a separação entre dados tratados e chaves de reidentificação, ampliando a robustez do sistema.

O arquivo `reverter.py` implementa a reversão controlada da pseudonimização, restrita a usuários com perfil de administrador. Ele identifica o formato das chaves existentes e aplica a substituição inversa, recuperando valores originais quando há justificativa legal para tanto. Essa funcionalidade garante alinhamento com a LGPD, que permite a pseudonimização reversível em determinados contextos regulatórios e contratuais.

No âmbito da avaliação, o arquivo `quality_metrics.py` contém a implementação das métricas técnicas de anonimização, como *k-anonymity*, *l-diversity* e *t-closeness*, além de índices de perda de informação e cobertura de atributos sensíveis. O arquivo `avaliador.py`, por sua vez, amplia a análise, reunindo essas métricas em um quadro comparativo mais abrangente, incluindo tempo de execução e grau de preservação da utilidade dos dados.

É neste ponto metodológico que as métricas são efetivamente calculadas e capturadas, logo após a aplicação das técnicas de anonimização pelo `processor.py`, servindo tanto para retroalimentar o sistema quanto para compor os relatórios de conformidade.

O arquivo `risk_analysis.py` introduz uma dimensão complementar: a análise de risco residual de reidentificação. Este módulo atribui níveis de risco (baixo, médio ou alto) com base na diversidade dos atributos e na robustez das técnicas aplicadas, traduzindo o impacto técnico em linguagem acessível para tomadores de decisão.

Finalmente, o arquivo generator.py é responsável pela produção do relatório técnico-jurídico.É apresentado: descrição das técnicas aplicadas, justificativas técnicas e legais (com menções expressas à LGPD), métricas de anonimização, análise de risco e recomendações. Esse relatório é o principal produto de evidência de conformidade, articulando os resultados técnicos com o discurso normativo.

4.3.5 Fluxo de Uso do Sistema

O CAFIS (*Context-Aware Framework for Information Sensitivity*) foi concebido para estruturar e automatizar todo o processo de anonimização, análise de risco e posterior reversão autorizada de dados sensíveis. Seu fluxo de operação é organizado de forma sequencial, permitindo ao usuário percorrer cada etapa com clareza, transparência e rastreabilidade. A seguir, apresenta-se o fluxo completo, correlacionando cada fase às telas do sistema (que deverão ser inseridas como figuras conforme referência indicada).

1ª Etapa - Acesso ao Sistema

O fluxo inicia-se pelo login seguro, onde o usuário informa suas credenciais institucionais. Conforme a Figura de Login, Figura 4, a interface apresenta campos de usuário e senha, reforçando a política de acesso restrito e autenticado, alinhada às exigências da LGPD para controle de acesso e segurança da informação.

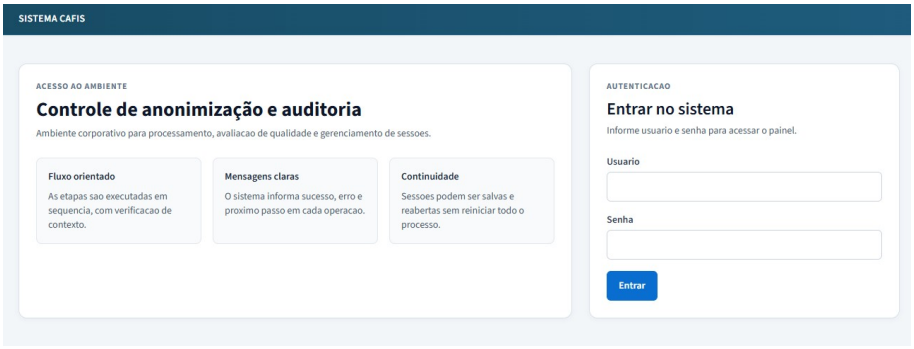


Figura 4 – Tela de Acesso

2ª Etapa - Upload do Arquivo CSV

Após a autenticação, o usuário é direcionado à etapa inicial do fluxo: o upload da base de dados que será processada. Na Figura 5 correspondente ao Upload, o sistema orienta sobre o formato aceito (.csv) e permite arrastar o arquivo ou selecioná-lo. Ao carregar o arquivo, o CAFIS exibe a confirmação visual de que o dataset foi reconhecido, garantindo transparência e conferência prévia dos dados.

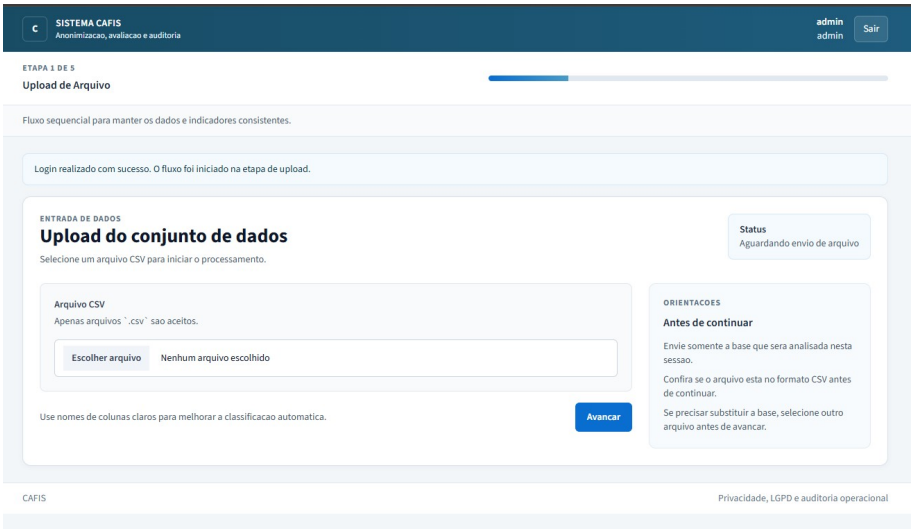


Figura 5 – Upload de Base de Dados

3ª Etapa - Detecção Automática de Atributos

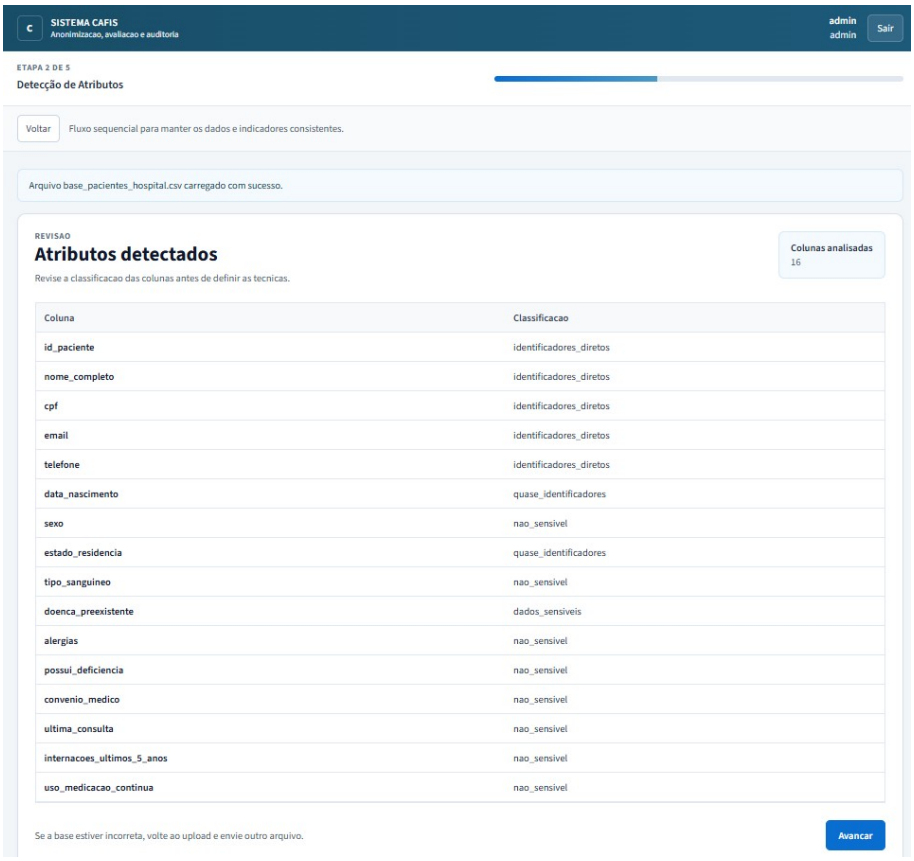


Figura 6 – Detecção Automática de Atributos

Na etapa seguinte, o CAFIS realiza uma varredura semântica e estrutural sobre as colunas do arquivo enviado. A tela de atributos detectados, aprensentada na Figura 6 lista cada

campo identificado e sua classificação preliminar:

- identificadores diretos;
- quase-identificadores;
- atributos sensíveis;
- atributos neutros.

Essa identificação é fundamental para aplicação correta das técnicas de anonimização, conforme princípios de minimização e necessidade estabelecidos pela LGPD.

4ª Etapa - Escolha das Técnicas de Anonimização

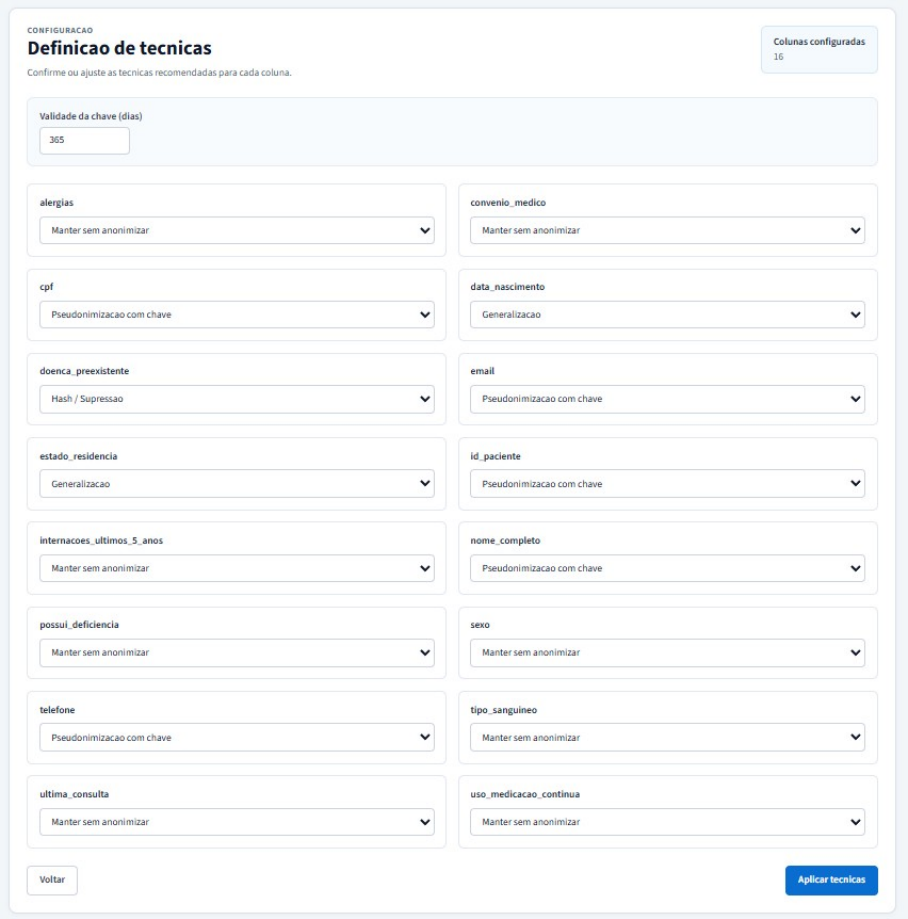


Figura 7 – Escolha das Técnicas de Anonimização

Com base nos atributos detectados, o CAFIS permite ao usuário definir a técnica de anonimização mais adequada para cada coluna. A interface de escolha das técnicas, mostrada na Figura 7, apresenta opções como:

- pseudonimização simples ou com chave,
- generalização,
- supressão,
- hashing irreversível,
- manutenção (para atributos neutros).

Nesta mesma tela, o usuário também pode ajustar o grau de variabilidade ou granularidade (por exemplo, generalização de idade para faixas etárias).

5ª Etapa Prévia dos Dados Anonimizados

	id_paciente	nome_completo	cpf	email	telefone	data_nascimento	sexo	estado_residencia	tipo_sanguineo
0	gAAAAA670221	gAAAAA894298	gAAAAA761284	gAAAAA875001	gAAAAA764230	categoria	Feminino	categoria	B+
1	gAAAAA688416	gAAAAA289839	gAAAAA491225	gAAAAA725485	gAAAAA615738	categoria	Masculino	categoria	A+
2	gAAAAA776418	gAAAAA956673	gAAAAA615919	gAAAAA211001	gAAAAA618263	categoria	Não-binário	categoria	O+
3	gAAAAA536961	gAAAAA111210	gAAAAA818843	gAAAAA614780	gAAAAA284066	categoria	Masculino	categoria	O+
4	gAAAAA912321	gAAAAA674414	gAAAAA280453	gAAAAA673355	gAAAAA165955	categoria	Não-binário	categoria	O-
5	gAAAAA408985	gAAAAA75162	gAAAAA249574	gAAAAA805423	gAAAAA852493	categoria	Não-binário	categoria	A-
6	gAAAAA495471	gAAAAA268782	gAAAAA276253	gAAAAA423126	gAAAAA957414	categoria	Masculino	categoria	O+
7	gAAAAA771940	gAAAAA909590	gAAAAA674463	gAAAAA587637	gAAAAA273210	categoria	Não-binário	categoria	A-
8	gAAAAA403632	gAAAAA437082	gAAAAA363714	gAAAAA891069	gAAAAA417219	categoria	Feminino	categoria	B+
9	gAAAAA745310	gAAAAA879206	gAAAAA548570	gAAAAA444869	gAAAAA340154	categoria	Não-binário	categoria	B-

Figura 8 – Prévia dos Dados Anonimizados

Após as escolhas técnicas, o CAFIS gera uma pré-visualização da base anonimizada, permitindo auditoria e conferência antes da aplicação definitiva. A Figura 8, que apresenta uma Prévia de Dados exibe a tabela resultante com:

- identificadores substituídos por pseudônimos,
- campos generalizados,
- atributos sensíveis tratados adequadamente,
- remoção ou supressão quando necessária.

Essa etapa garante transparência e evita erros de configuração de técnicas.

6ª Etapa - Painel CAFIS — Indicadores de Anonimização

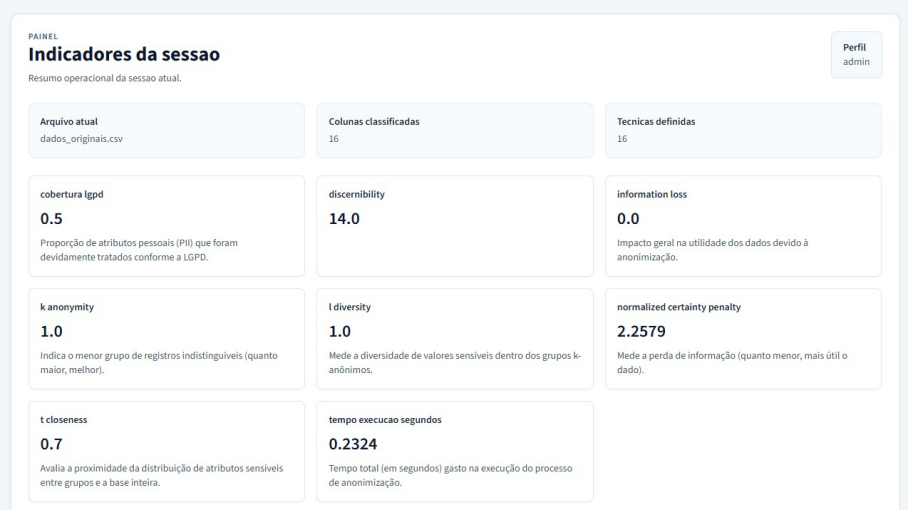


Figura 9 – Painel de Indicadores

Após a aplicação final das técnicas, o CAFIS apresenta um painel de métricas quantitativas que avaliam a robustez do processo. Conforme a Figura 9 de Indicadores, são exibidos parâmetros como:

- K-anonymity,
- L-diversity,
- T-closeness,
- Information Loss,
- Discernibility,
- Normalized Certainty Penalty (NCP),
- Tempo de Execução.

Essas métricas servem tanto para validação técnica quanto para documentação de conformidade com a LGPD e normas internacionais.

7ª Etapa - Visualizações e Gráficos de Execução

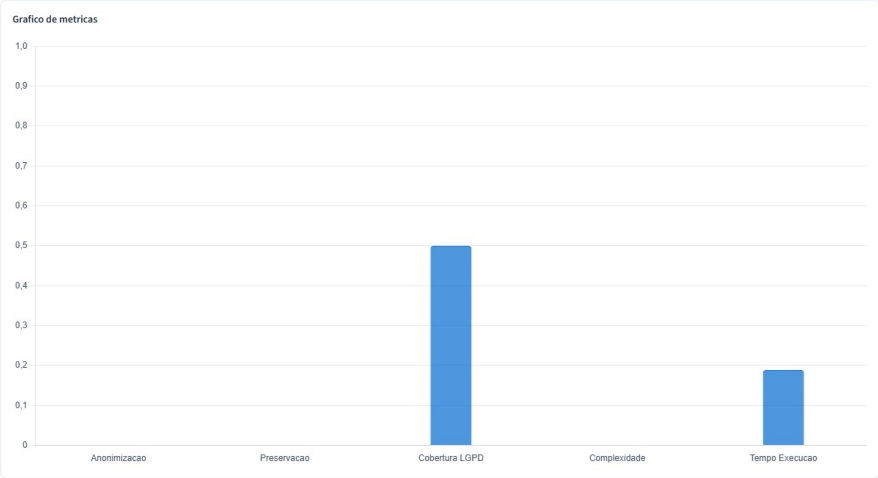


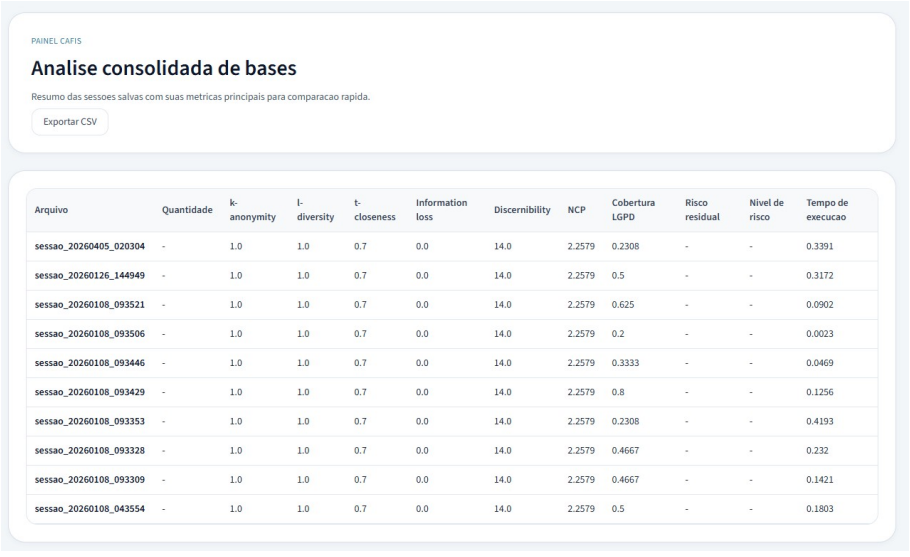
Figura 10 – Visualização de Resultados

Em seguida, o CAFIS apresenta gráficos que auxiliam na interpretação dos níveis de anonimização alcançados, conforme apresentado na Figura 10. Esses gráficos demonstram:

- qualidade global da anonimização,
- perda de informação,
- comparativo de métricas por sessão,
- tempos de execução.

A interface permite salvar sessões, gerar relatórios e consultar o histórico.

8ª Etapa - Análise Consolidada de Bases



Arquivo	Quantidade	k-anonymity	l-diversity	t-closeness	Information loss	Discernibility	NCP	Cobertura LGPD	Risco residual	Nivel de risco	Tempo de execucao
sessao_20260405_020304	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.2308	-	-	0.3391
sessao_20260126_144949	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.5	-	-	0.3172
sessao_20260108_093521	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.625	-	-	0.0902
sessao_20260108_093506	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.2	-	-	0.0023
sessao_20260108_093446	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.3333	-	-	0.0469
sessao_20260108_093429	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.8	-	-	0.1256
sessao_20260108_093353	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.2308	-	-	0.4193
sessao_20260108_093328	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.4667	-	-	0.232
sessao_20260108_093309	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.4667	-	-	0.1421
sessao_20260108_043554	-	1.0	1.0	0.7	0.0	14.0	2.2579	0.5	-	-	0.1803

Figura 11 – Análise de Bases Anonimizadas

O CAFIS também possui uma visão comparativa entre múltiplas sessões ou múltiplas bases anonimizada em diferentes rodadas. A Figura 11 reúne, em formato tabular, indicadores de todas as sessões registradas, permitindo auditoria longitudinal e avaliação da consistência dos processos realizados.

9ª Etapa - Comparação: Anonimização x Reversão Autorizada

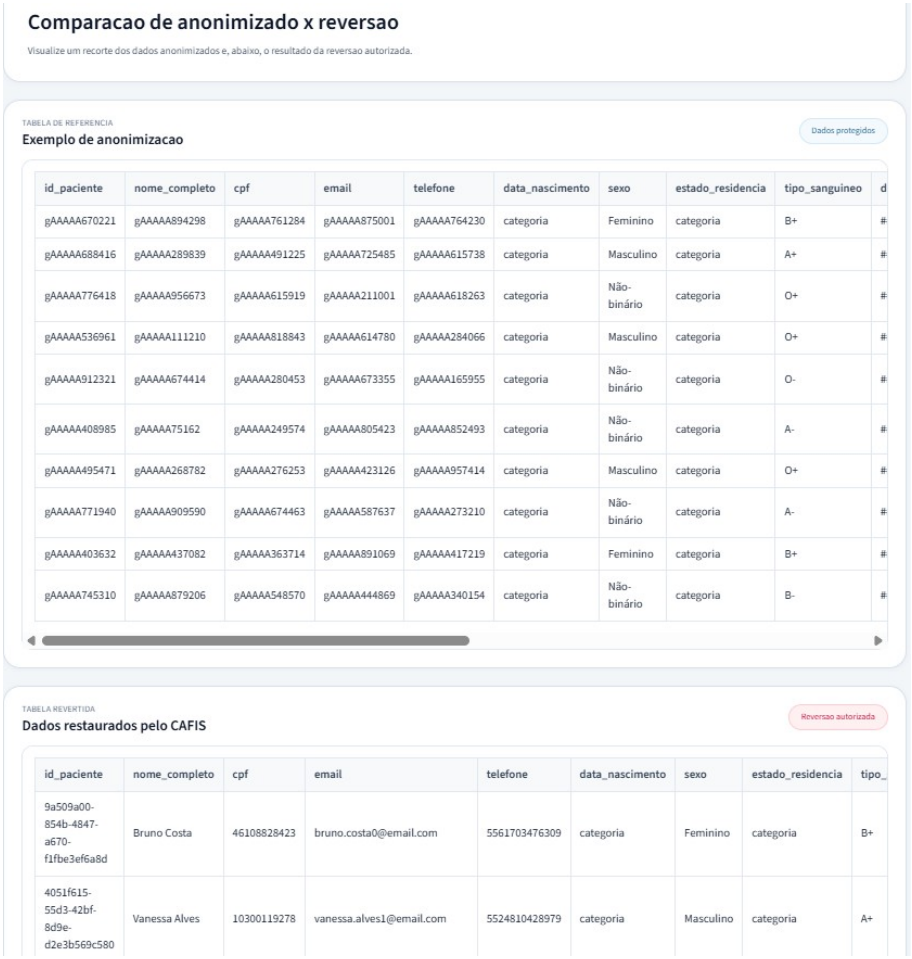


Figura 12 – Reversão e Comparação

Um dos diferenciais do CAFIS é a capacidade de simular e, quando autorizado, executar a reversão segura dos dados. Na Figura 12, o sistema exibe lado a lado:

- versão anonimizada,
- versão restaurada (quando há chave autorizada e solicitação justificada).

Por fim, na Figura de Restauração, visualiza-se a tabela com dados reidentificados sob autorização controlada — funcionalidade alinhada aos requisitos de accountability e finalidade legítima previstos na LGPD.

4.3.6 Processo de Governança LGPD e Execução do CAFIS

A Figura 13 apresenta o processo integrado de governança e tratamento de dados proposto pelo framework CAFIS, modelado por meio da notação BPMN. O fluxo evidencia a articulação entre duas camadas complementares: a **Camada de Governança (LGPD/DPO)**, responsável pela definição dos requisitos legais, formalização dos instrumentos de governança e validação da conformidade; e a **Camada Técnica de Tratamento de Dados (CAFIS)**, na qual são executadas as atividades de classificação, análise de risco, anonimização e avaliação dos resultados obtidos. Essa separação permite estabelecer pontos explícitos de interação entre decisões jurídicas e procedimentos técnicos, evitando que a anonimização seja tratada como uma atividade exclusivamente computacional e desvinculada das condições que legitimam o tratamento dos dados pessoais.

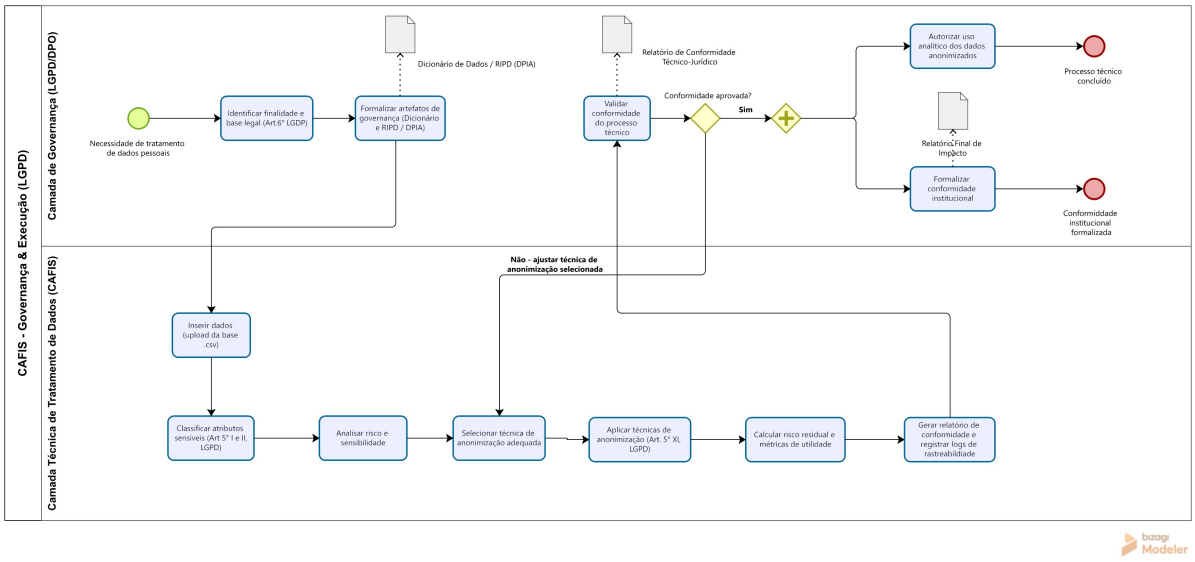


Figura 13 – Processo de Governança LGPD e Execução do CAFIS

O processo inicia-se na Camada de Governança a partir da identificação de uma **necessidade de tratamento de dados pessoais**. Diante dessa necessidade, a primeira atividade consiste em identificar a finalidade do tratamento e a respectiva base legal, conforme os requisitos estabelecidos pela LGPD. Essa etapa constitui um ponto fundamental do processo, uma vez que a legitimidade das operações subsequentes depende da existência de propósito determinado e de fundamento jurídico adequado. A atividade encontra respaldo especialmente nos princípios da finalidade, adequação e necessidade previstos no artigo 6º da Lei nº 13.709/2018. Conforme argumenta Doneda (2020), o estabelecimento da finalidade constitui importante mecanismo de limitação do poder informacional, pois restringe a utilização dos dados aos propósitos previamente determinados e juridicamente legitimados.

Após a definição da finalidade e da base legal, são **formalizados os artefatos de governança**, representados no processo pelo Dicionário de Dados e pelo Relatório de Impacto à Proteção de Dados (RIPD/DPIA). Esses instrumentos documentam as características do tratamento, os dados envolvidos, seus respectivos significados e os riscos relacionados às operações realizadas. O RIPD assume especial relevância nesse contexto, considerando sua previsão no artigo 38 da LGPD como instrumento destinado à descrição dos processos de tratamento que possam gerar riscos às liberdades civis e aos direitos fundamentais, bem como das medidas adotadas para mitigá-los. Dessa forma, a documentação produzida estabelece uma referência de governança para a execução das etapas técnicas subsequentes e contribui para a demonstração da responsabilidade e da prestação de contas.

A partir da formalização desses artefatos, o fluxo é direcionado à **Camada Técnica de Tratamento de Dados do CAFIS**, iniciando-se pela inserção dos dados, representada no BPMN pelo *upload* da base em formato CSV. Na sequência, os atributos presentes no conjunto de dados são classificados quanto à sua sensibilidade, considerando as categorias estabelecidas pelo artigo 5º, incisos I e II, da LGPD. Essa classificação permite reconhecer atributos que demandam maior proteção e fornece os elementos necessários para orientar as decisões técnicas de anonimização. A identificação e classificação dos dados também contribuem para a observância do princípio da necessidade, uma vez que permitem compreender quais informações estão efetivamente envolvidas no tratamento e quais riscos estão associados à sua utilização.

Após a classificação dos atributos, o CAFIS realiza a **análise de risco e sensibilidade** dos dados. Nessa etapa, são avaliadas as características do conjunto de dados e os potenciais riscos relacionados à exposição, associação ou reidentificação dos titulares. Essa análise estabelece os parâmetros necessários para a escolha das medidas técnicas subsequentes e materializa uma abordagem orientada a risco. Em vez de assumir que todos os conjuntos de dados apresentam o mesmo nível de exposição, o processo considera suas particularidades para determinar o nível de proteção necessário, aproximando a decisão técnica dos riscos efetivamente identificados.

Com base nos resultados dessa análise, o processo avança para a **seleção da técnica de anonimização adequada**. A escolha é realizada de maneira contextual, considerando a natureza e a sensibilidade dos atributos, os riscos previamente identificados e a necessidade de preservação da utilidade dos dados. Uma vez selecionada a técnica, são aplicados os mecanismos de anonimização previstos pelo CAFIS, em consonância com o conceito estabelecido no artigo 5º, inciso XI, da LGPD, segundo o qual a anonimização corresponde à utilização de meios técnicos razoáveis e disponíveis no momento do tratamento para impedir a associação direta ou indireta dos dados a um indivíduo.

Concluída a aplicação das técnicas de anonimização, o framework realiza o **cál-**

culo do risco residual e das métricas de utilidade. Essa etapa é necessária porque a aplicação de uma técnica de anonimização não implica, por si só, a eliminação automática de qualquer possibilidade de associação ou reidentificação. O risco remanescente deve ser mensurado e confrontado com os critérios estabelecidos para o tratamento. Paralelamente, a avaliação da utilidade permite verificar em que medida as transformações realizadas preservaram as características necessárias para a finalidade analítica pretendida. Dessa forma, o CAFIS busca estabelecer equilíbrio entre redução do risco de reidentificação e manutenção da utilidade informacional do conjunto de dados.

Os resultados dessas operações são então consolidados na atividade de **geração do relatório de conformidade e registro dos logs de rastreabilidade**. Esse relatório reúne evidências sobre as operações realizadas durante o tratamento, incluindo as técnicas selecionadas, os resultados da avaliação de risco residual e as informações necessárias para reconstruir o percurso técnico adotado. Os registros de rastreabilidade complementam esse mecanismo ao permitir o acompanhamento das ações executadas, fortalecendo a capacidade de auditoria e a demonstração de conformidade. Dessa maneira, o CAFIS incorpora mecanismos de prestação de contas não apenas como documentação posterior, mas como componentes integrantes do próprio processo de tratamento.

Após a geração das evidências técnicas, o fluxo retorna à **Camada de Governança**, na qual é realizada a atividade de validação da conformidade do processo técnico. Nesse momento, os resultados produzidos pelo CAFIS são confrontados com os requisitos legais, os artefatos de governança previamente estabelecidos e os critérios definidos para o tratamento. A validação resulta na elaboração de um **Relatório de Conformidade Técnico-Jurídico**, que funciona como elemento de ligação entre as evidências produzidas tecnicamente e a decisão de governança sobre a aceitabilidade do tratamento.

O BPMN estabelece, nesse ponto, um **gateway de decisão**, representado pela questão “Conformidade aprovada?”. Caso a avaliação resulte em uma decisão negativa, o processo não é encerrado nem o uso dos dados é autorizado. O fluxo retorna à Camada Técnica, especificamente à etapa de **seleção da técnica de anonimização**, indicando a necessidade de ajustar a estratégia anteriormente adotada. Uma nova técnica ou configuração pode então ser selecionada, aplicada e novamente submetida ao cálculo de risco residual, à avaliação de utilidade, à geração das evidências de conformidade e à validação técnico-jurídica. Esse mecanismo caracteriza um **ciclo iterativo de melhoria e adequação**, no qual o tratamento somente avança quando os resultados técnicos atendem aos critérios de conformidade estabelecidos pela governança.

Quando a conformidade é aprovada, o processo alcança um **gateway paralelo**, a partir do qual são executados dois desdobramentos complementares. O primeiro corresponde à **autorização do uso analítico dos dados anonimizados**, permitindo que o conjunto de dados resultante seja utilizado para a finalidade previamente definida.

Esse caminho encerra-se com a indicação de que o processo técnico foi concluído. Dessa forma, a autorização não decorre simplesmente da execução de uma técnica de anonimização, mas da aprovação de todo o ciclo de análise, transformação, mensuração de risco e validação de conformidade.

O segundo desdobramento ocorre paralelamente e corresponde à **formalização da conformidade institucional**. Essa atividade consolida os resultados e evidências produzidos durante o processo, estando associada à elaboração do **Relatório Final de Impacto**. Seu objetivo é registrar institucionalmente as condições sob as quais o tratamento foi realizado e aprovado, estabelecendo documentação capaz de subsidiar processos de governança, auditoria e prestação de contas. Esse fluxo encerra-se com a conformidade institucional formalizada, complementando a conclusão técnica representada pelo primeiro ramo do processo.

A estrutura apresentada na Figura 13 evidencia, portanto, que o CAFIS não considera a anonimização como uma operação isolada ou meramente técnica. O processo estabelece uma relação contínua entre **governança, avaliação de risco, transformação dos dados, mensuração dos resultados, validação técnico-jurídica e autorização de uso**. Particularmente relevante é o mecanismo de retroalimentação previsto quando a conformidade não é aprovada, pois impede que um conjunto de dados avance para utilização analítica enquanto os níveis de risco ou os requisitos de conformidade permanecerem inadequados. Essa lógica aproxima o processo da concepção defendida por Doneda (2020), segundo a qual a proteção de dados deve abranger o controle sobre todo o ciclo de circulação e utilização das informações pessoais, e não apenas sua confidencialidade.

Assim, o BPMN apresentado na Figura 13 materializa operacionalmente a integração entre a **Camada de Governança (LGPD/DPO)** e a **Camada Técnica de Tratamento de Dados (CAFIS)**. A governança estabelece as condições de legitimidade, documenta o tratamento e decide sobre sua conformidade, enquanto o CAFIS fornece os mecanismos técnicos para classificar os dados, avaliar riscos, selecionar e aplicar técnicas de anonimização, mensurar o risco residual e produzir evidências rastreáveis. A combinação dessas duas dimensões estabelece um fluxo iterativo e verificável no qual a utilização analítica dos dados anonimizados somente é autorizada após a validação da conformidade, enquanto, paralelamente, os resultados são formalizados institucionalmente. Dessa forma, o framework busca traduzir requisitos jurídicos e princípios de proteção de dados em atividades técnicas concretas, documentadas e passíveis de auditoria.

4.3.7 Heurísticas de Decisão do CAFIS

A operacionalização das decisões no framework CAFIS não ocorre de forma arbitrária, sendo guiada por um conjunto estruturado de heurísticas que orientam tanto a classificação dos dados quanto a seleção das técnicas de anonimização. Essas heurísticas

constituem o núcleo decisório do sistema, permitindo transformar regras abstratas de proteção de dados em critérios operacionais aplicáveis ao tratamento automatizado de informações.

No contexto do CAFIS, as heurísticas são organizadas em duas camadas complementares: (i) heurísticas de classificação de atributos, baseadas em análise léxica, e (ii) heurísticas de recomendação de técnicas de anonimização, baseadas na categoria atribuída aos dados. Essa separação permite modularidade e clareza no processo decisório, distinguindo a etapa de identificação do tipo de dado da etapa de definição da técnica de proteção mais adequada.

A primeira camada consiste na classificação automática de atributos a partir da análise dos nomes das colunas do conjunto de dados. Para isso, o framework utiliza um dicionário de palavras-chave sensíveis, estruturado em três categorias principais: identificadores diretos, quase-identificadores e dados sensíveis. A partir da normalização dos nomes das colunas, o sistema verifica a presença de termos associados a cada categoria, classificando os atributos de forma automática. Essa abordagem léxica permite identificar, de maneira eficiente, possíveis dados pessoais mesmo na ausência de metadados estruturados, sendo particularmente útil em cenários exploratórios ou com dados heterogêneos.

Os identificadores diretos incluem atributos que permitem a identificação imediata do titular, como nome, CPF, e-mail ou telefone. Os quase-identificadores correspondem a atributos que, isoladamente, não identificam diretamente um indivíduo, mas que podem levar à sua reidentificação quando combinados com outras informações, como idade, localização ou código postal. Já os dados sensíveis compreendem informações relacionadas a aspectos íntimos da pessoa, como saúde, religião ou origem étnica, cuja proteção é reforçada pela Lei Geral de Proteção de Dados (Lei nº 13.709/2018), conforme definido em seu Art. 5º.

A segunda camada de heurísticas corresponde à recomendação das técnicas de anonimização com base na categoria atribuída a cada atributo. Essa camada implementa um mapeamento direto entre categorias de dados e técnicas de proteção, permitindo automatizar a escolha de estratégias de anonimização. De forma geral, atributos classificados como identificadores diretos são submetidos à pseudonimização com chave, dados sensíveis são tratados por técnicas mais restritivas, como hash ou supressão, e quase-identificadores são transformados por meio de generalização. Atributos não classificados como sensíveis são mantidos sem alteração.

Essa lógica de decisão reflete uma abordagem baseada em risco, na qual a intensidade da técnica de anonimização é proporcional à sensibilidade do dado. Tal princípio está alinhado às boas práticas de proteção de dados e à lógica regulatória da LGPD, que exige a adoção de medidas adequadas e proporcionais para garantir a proteção dos titulares.

A estrutura geral das heurísticas adotadas no CAFIS pode ser sintetizada no Quadro6.

Quadro 6 – Estrutura das heurísticas de decisão do CAFIS

Camada	**Entrada**	**Regra**	**Saída**
Classificação léxica	Nome da coluna	Identificação de termos sensíveis por meio de correspondência com palavras-chave previamente definidas no dicionário de atributos	Categoria do atributo
Decisão de técnica	Categoria do atributo	Aplicação de mapeamento heurístico entre categoria de dado e técnica de anonimização mais adequada, considerando o nível de sensibilidade	Técnica recomendada

No fluxo operacional do sistema, essas heurísticas são aplicadas após o carregamento dos dados. Inicialmente, o conjunto de dados é analisado para identificação automática das colunas sensíveis. Em seguida, o sistema recomenda técnicas de anonimização para cada atributo, com base nas regras definidas. O usuário pode revisar e ajustar essas recomendações antes da aplicação final, o que caracteriza o framework como um sistema híbrido, combinando automação e intervenção humana.

Embora, na implementação atual, essas heurísticas estejam codificadas diretamente no sistema, sua estrutura pode ser representada de forma explícita como um artefato configurável, denominado “arquivo de heurísticas”. Essa representação permite tornar visíveis os critérios utilizados no processo decisório, favorecendo a auditabilidade e a reprodutibilidade do framework.

Um exemplo de representação lógica dessas heurísticas é apresentado a seguir:

```
{
  "sensitive_keywords": {
    "identificadores_diretos": ["nome", "cpf", "email", "telefone", "id"],
    "quase_identificadores": ["cep", "cidade", "estado", "idade"],
    "dados_sensiveis": ["doenca", "diagnostico", "religiao", "etnia"]
  },
  "categoria_para_tecnica": {
    "identificadores_diretos": "pseudonimizacao_com_chave",
    "quase_identificadores": "generalizacao",
    "dados_sensiveis": "hash/supressao",
    "nao_sensivel": "manter"
  }
}
```

A explicitação dessas regras como um artefato independente reforça a transparência do processo de anonimização, permitindo que decisões automatizadas sejam compreendidas, auditadas e, quando necessário, ajustadas. Essa característica está alinhada ao princípio da responsabilização (*accountability*) previsto na LGPD, ao tornar explícitos os critérios adotados no tratamento de dados pessoais.

Por fim, a formalização das heurísticas no CAFIS contribui para transformar o processo de anonimização em um procedimento sistemático, replicável e governável, integrando eficiência técnica, controle regulatório e suporte à tomada de decisão.

5 RESULTADOS E DISCUSSÕES

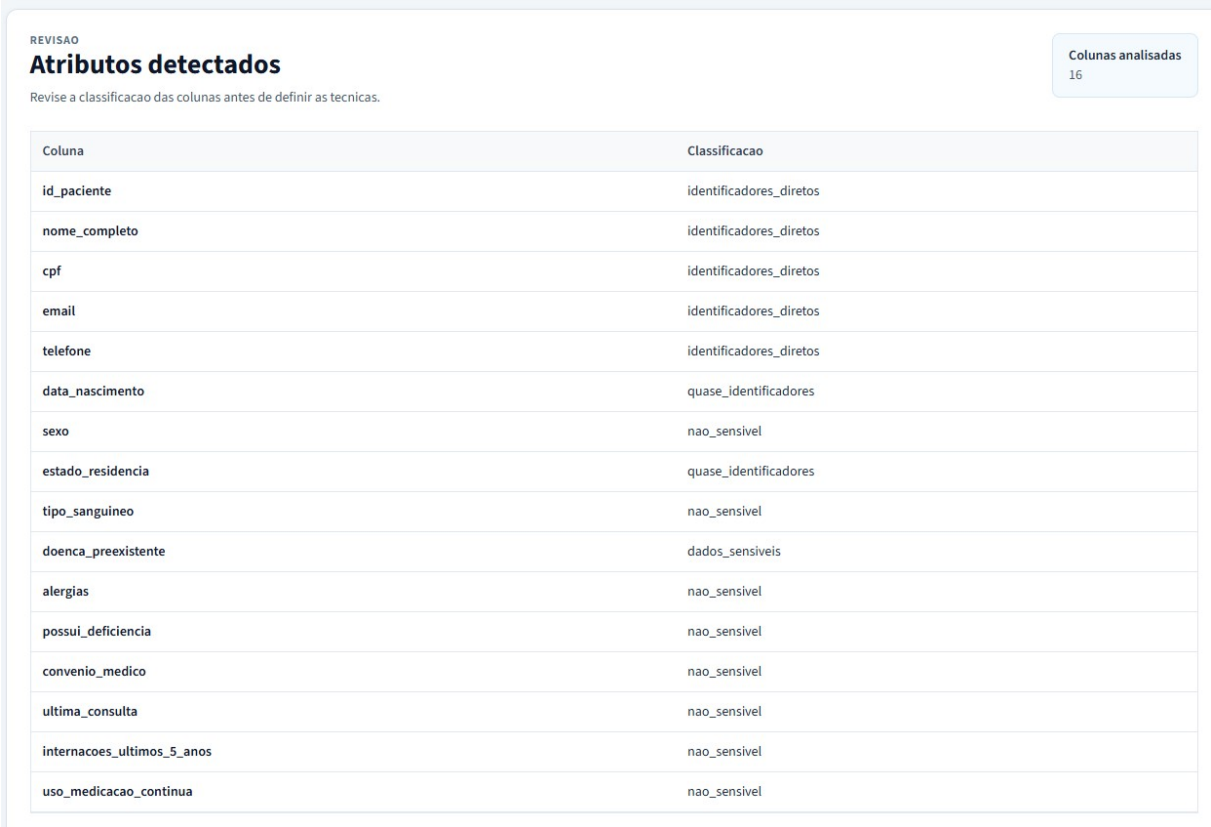
Esta seção apresenta os resultados obtidos a partir da aplicação do framework CAFIS sobre o dataset experimental inserido no sistema. O objetivo da análise foi avaliar o comportamento do framework frente a uma base realista, contendo dados pessoais e sensíveis, verificando sua capacidade de identificar atributos, classificar riscos, aplicar técnicas adequadas de anonimização e preservar a utilidade informacional dos dados, em conformidade com a LGPD.

5.1 Resultado da Análise Contextual da Base

Após a inserção do dataset, o CAFIS identificou que a base analisada pertence ao domínio da saúde, reunindo informações cadastrais, clínicas e administrativas de pacientes. Esse enquadramento inicial caracterizou a base como de alto risco regulatório, uma vez que envolve dados sensíveis nos termos do artigo 5º, inciso II, da LGPD. Como resultado dessa análise, o framework ativou automaticamente critérios mais restritivos de avaliação de risco e recomendação de técnicas, priorizando mecanismos de proteção mais robustos para atributos críticos. Observou-se, portanto, que a análise contextual influenciou diretamente todas as etapas subsequentes do processo de anonimização, confirmando o papel estruturante dessa fase no CAFIS.

5.2 Resultado da Detecção e Classificação dos Atributos

Na etapa de detecção automática, o framework CAFIS realizou a inspeção sistemática das colunas do *dataset*, promovendo a classificação dos atributos em categorias distintas de acordo com seu potencial de identificação individual e grau de sensibilidade jurídica, conforme ilustrado na Figura 12. Nesse processo, foram identificados diversos identificadores diretos, tais como *id_paciente*, *nome_completo*, *cpf*, *email* e *telefone*, bem como quase-identificadores, a exemplo de *data_nascimento* e *estado_residencia*. Adicionalmente, atributos de natureza clínica, como *doenca_preexistente*, foram corretamente classificados como dados sensíveis, enquanto variáveis de caráter administrativo e estatístico foram enquadradas como dados de baixo risco contextual.



REVISAO

Atributos detectados

Revise a classificacao das colunas antes de definir as tecnicas.

Colunas analisadas
16

Coluna	Classificacao
id_paciente	identificadores_diretos
nome_completo	identificadores_diretos
cpf	identificadores_diretos
email	identificadores_diretos
telefone	identificadores_diretos
data_nascimento	quase_identificadores
sexo	nao_sensivel
estado_residencia	quase_identificadores
tipo_sanguineo	nao_sensivel
doenca_preexistente	dados_sensiveis
alergias	nao_sensivel
possui_deficiencia	nao_sensivel
convenio_medico	nao_sensivel
ultima_consulta	nao_sensivel
internacoes_ultimos_5_anos	nao_sensivel
uso_medicacao_continua	nao_sensivel

Figura 14 – Etapa de Classificação dos Atributos

O resultado dessa etapa evidencia a capacidade do CAFIS de diferenciar automaticamente atributos com base em seu potencial de identificação e na sensibilidade jurídica associada. Tal classificação possibilitou que o framework evitasse a aplicação homogênea de técnicas de anonimização, permitindo a adoção de um tratamento diferenciado, proporcional e alinhado aos princípios de minimização e adequação, conforme preconizado pela legislação de proteção de dados pessoais.

5.3 Resultado da Avaliação de Sensibilidade e Risco de Reidentificação

A partir da classificação dos atributos, o CAFIS realizou a avaliação do risco de reidentificação associado a cada campo da base. Os identificadores diretos foram considerados de risco elevado, pois possibilitam associação imediata a indivíduos específicos. Os quase-identificadores foram avaliados quanto ao risco de reidentificação indireta por correlação, enquanto os dados sensíveis foram classificados como de alto impacto potencial sobre os direitos fundamentais dos titulares.

Como resultado dessa análise, o framework demonstrou capacidade de reconhecer que determinados atributos, embora não identificadores isoladamente, podem aumentar significativamente o risco quando combinados. Esse resultado reforça a abordagem

contextual e baseada em risco adotada pelo CAFIS, alinhada à definição de anonimização prevista na LGPD, que considera os meios técnicos razoáveis disponíveis no momento do tratamento.

5.4 Resultado da Seleção e Aplicação das Técnicas de Anonimização

Com base na avaliação de risco contextual e na classificação previamente atribuída aos atributos, o framework CAFIS procedeu à aplicação diferenciada de técnicas de anonimização, conforme evidenciado na Figura 13. Nessa etapa, os identificadores diretos, como *cpf*, *nome_completo*, *email* e *id_paciente*, foram submetidos à técnica de pseudonimização com chave, a qual substitui os valores originais por representações artificiais consistentes e reversíveis sob controle criptográfico, permitindo rastreabilidade controlada quando estritamente necessária.

Os quase-identificadores, a exemplo de *data_nascimento* e *estado_residencia*, foram tratados por meio de generalização, reduzindo deliberadamente o nível de granularidade das informações com o objetivo de mitigar riscos de reidentificação por inferência ou correlação externa. Para atributos classificados como dados sensíveis de maior criticidade jurídica e ética, como *doenca_preexistente*, o CAFIS aplicou técnicas mais restritivas, incluindo *hash* criptográfico e supressão, eliminando a possibilidade de associação direta ao titular dos dados.

Por outro lado, atributos avaliados como de baixo risco contextual ou de relevância administrativa e estatística, tais como *sexo*, *internacoes_ultimos_5_anos*, *alergias* e *possui_deficiencia*, foram mantidos sem anonimização, preservando sua integridade informacional e utilidade analítica. O resultado dessa etapa demonstra que o CAFIS foi capaz de aplicar anonimização de forma seletiva, proporcional e orientada ao risco, equilibrando a mitigação de ameaças de reidentificação com a preservação do valor informacional do *dataset*, evitando sua descaracterização completa e assegurando aderência aos princípios de minimização e necessidade previstos na legislação de proteção de dados.

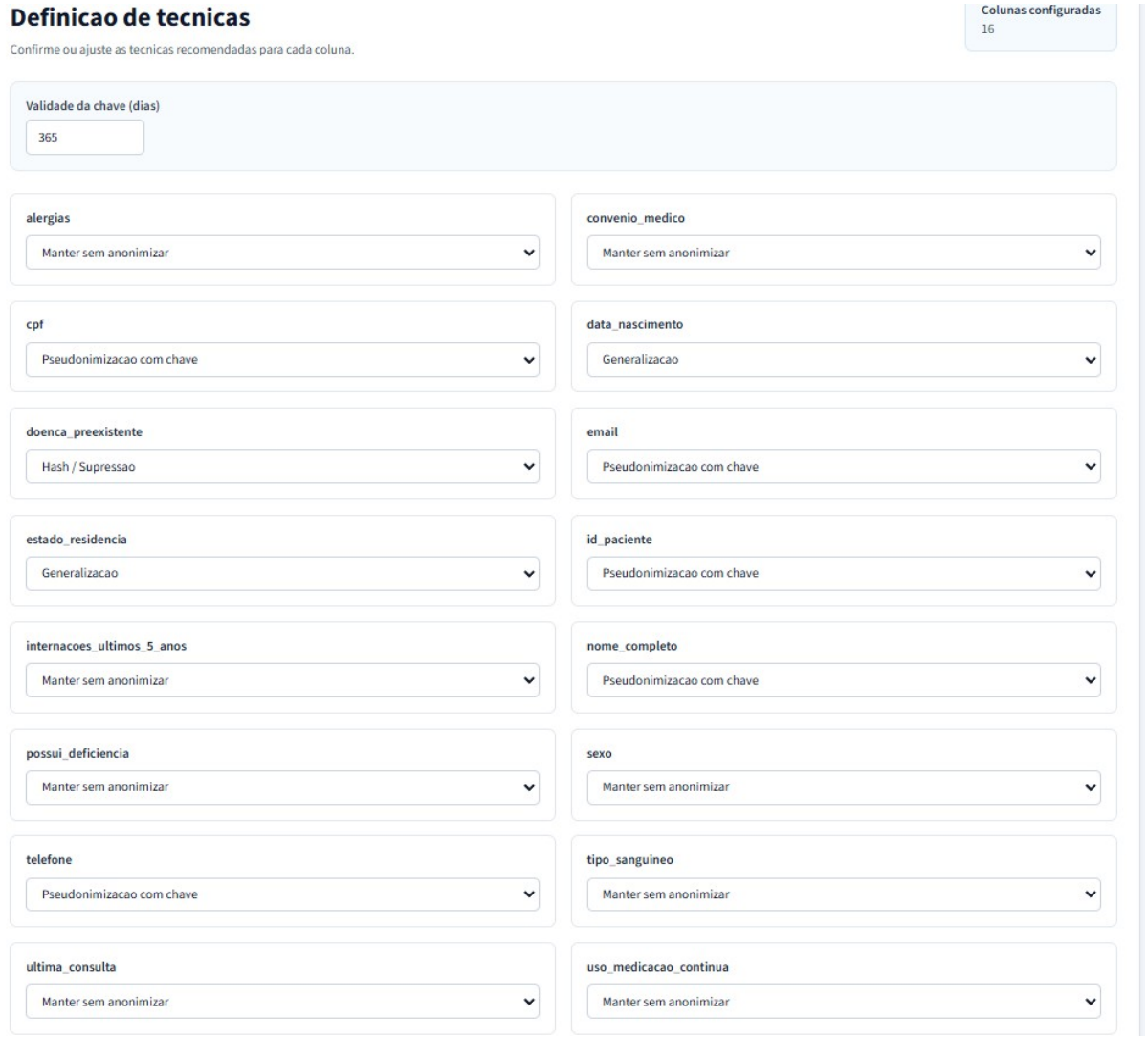


Figura 15 – Escolha das Técnicas de Anonimização

5.5 Resultado da Prévia e Validação dos Dados Anonimizados

A aplicação das técnicas de anonimização selecionadas na etapa anterior pode ser observada de forma concreta nas Figuras 13 e 14, que ilustram, respectivamente, o tratamento de dados sensíveis e de dados pessoais no *dataset* analisado. De acordo com a Lei Geral de Proteção de Dados Pessoais (LGPD – Lei nº 13.709/2018), considera-se dado pessoal toda informação relacionada a pessoa natural identificada ou identificável, enquanto dado pessoal sensível corresponde a informações sobre origem racial ou étnica, convicções religiosas, opiniões políticas, dados referentes à saúde ou à vida sexual, dados genéticos ou biométricos, entre outros.

	id_paciente	nome_completo	cpf	email	telefone
0	gAAAAA670221	gAAAAA894298	gAAAAA761284	gAAAAA875001	gAAAAA764230
1	gAAAAA688416	gAAAAA289839	gAAAAA491225	gAAAAA725485	gAAAAA615738
2	gAAAAA776418	gAAAAA956673	gAAAAA615919	gAAAAA211001	gAAAAA618263
3	gAAAAA536961	gAAAAA111210	gAAAAA818843	gAAAAA614780	gAAAAA284066

Figura 16 – Resultado da Anonimização de Dados Pessoais

tipo_sanguineo	doenca_preexistente	alergias	possui_deficiencia	convenio_medico	ultima_consulta
B+	###	Lactose	Não	Particular	2023-03-01
A+	###	Glúten	Não	Bradesco Saúde	2025-01-25
O+	###	Nenhuma	Não	Amil	2025-06-17
O+	###	Penicilina	Sim	Amil	2022-11-12
O-	###	Dipirona	Não	SulAmérica	2025-01-01

Figura 17 – Resultado da Anonimição de Dados Sensíveis

No contexto apresentado, atributos como *id_paciente*, *nome_completo*, *cpf*, *email* e *telefone* enquadram-se como dados pessoais identificadores diretos, uma vez que permitem a identificação imediata do titular. Conforme demonstrado na Figura 14, esses atributos foram submetidos à técnica de pseudonimização com chave, resultando na substituição dos valores originais por identificadores artificiais consistentes, preservando a estrutura relacional do conjunto de dados e viabilizando análises internas sem exposição direta da identidade dos indivíduos.

Por sua vez, dados relacionados à saúde, como *doenca_preexistente*, enquadram-se como dados pessoais sensíveis nos termos do art. 5º, inciso II, da LGPD, demandando um nível de proteção mais rigoroso. Conforme ilustrado na Figura 13, o CAFIS aplicou técnicas de *hash* criptográfico e supressão para esses atributos, tornando impossível a reidentificação direta ou indireta do titular, ao mesmo tempo em que sinaliza a existência do dado sem revelar seu conteúdo específico.

Observa-se ainda que outros atributos clínicos e administrativos, como *tipo_sanguineo*,

alergias, *possui_deficiencia*, *convenio_medico* e *ultima_consulta*, foram mantidos ou parcialmente preservados conforme sua classificação de risco contextual, garantindo utilidade analítica sem violar os princípios de necessidade e minimização de dados. Dessa forma, o CAFIS demonstrou a capacidade de aplicar anonimização de maneira seletiva, proporcional e juridicamente orientada, mitigando riscos de reidentificação e assegurando conformidade com os requisitos legais, sem comprometer integralmente o valor informacional do *dataset*.

5.6 Resultado Global da Análise do Dataset pelo CAFIS

De forma consolidada, os resultados da aplicação do CAFIS ao *dataset* evidenciam a efetividade do framework tanto sob a perspectiva de proteção de dados quanto sob o critério de utilidade informacional. Conforme apresentado na Figura 16, a tela de análise de indicadores de anonimização consolida métricas quantitativas que permitem avaliar, de maneira objetiva, o impacto das técnicas aplicadas sobre a segurança, a privacidade e a utilidade analítica da base de dados, funcionando como um instrumento de validação operacional do CAFIS.

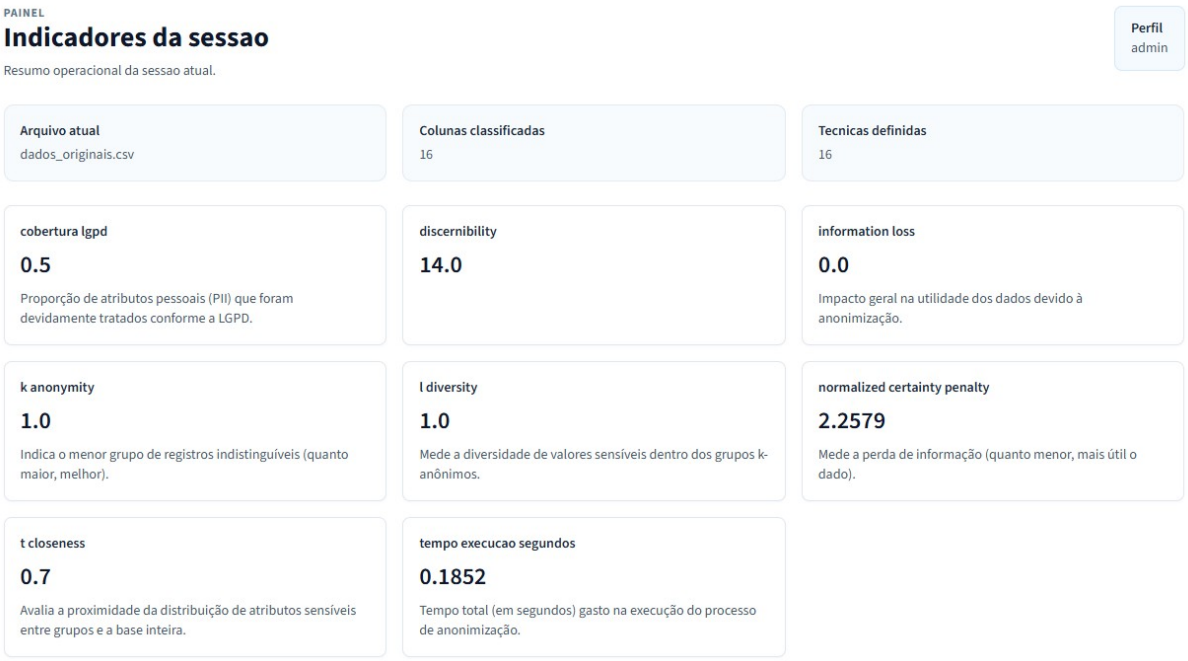


Figura 18 – Resultados e Calibração do CAFIS

A partir dessa análise integrada, observa-se que o framework:

- identificou corretamente atributos classificados como dados pessoais, dados pessoais sensíveis e dados de baixo risco contextual;

- avaliou os riscos de reidentificação de forma contextualizada, considerando tanto a natureza dos atributos quanto sua combinação no conjunto de dados;
- aplicou técnicas de anonimização proporcionais ao grau de sensibilidade e criticidade jurídica das informações;
- preservou a utilidade analítica do *dataset*, conforme evidenciado pelos baixos índices de perda de informação apresentados nos indicadores;
- gerou métricas técnicas e evidências quantitativas compatíveis com os princípios da Lei Geral de Proteção de Dados Pessoais (LGPD), em especial os princípios da necessidade, adequação, minimização e responsabilização.

Os indicadores apresentados na Figura 16 — como *k-anonymity*, *l-diversity*, *t-closeness*, perda de informação normalizada e cobertura LGPD — permitem não apenas mensurar o nível de proteção alcançado, mas também avaliar a eficiência do processo de anonimização em termos de preservação de valor informacional. Dessa forma, a tela de indicadores configura-se como um mecanismo essencial de transparência, auditabilidade e apoio à tomada de decisão, possibilitando que gestores, analistas e responsáveis pelo tratamento de dados compreendam os trade-offs entre privacidade e utilidade.

Em síntese, esses resultados indicam que o CAFIS é capaz de operacionalizar, de forma automatizada, mensurável e auditável, os requisitos legais de proteção de dados pessoais em um cenário realista, validando sua aplicabilidade prática, sua aderência normativa e sua utilidade como artefato tecnológico alinhado aos pressupostos da Design Science Research.

5.7 Discussão Crítica dos Resultados sob a Perspectiva da Lei Geral de Proteção de Dados Pessoais

A aplicação do framework CAFIS ao dataset experimental possibilitou avaliar, de forma empírica, o grau de aderência entre as decisões técnicas automatizadas e os comandos normativos estabelecidos pela Lei Geral de Proteção de Dados Pessoais (Lei nº 13.709/2018). Diferentemente de abordagens meramente declarativas de conformidade, os resultados obtidos permitem analisar como os princípios e artigos da LGPD foram operacionalizados ao longo do processo de anonimização.

5.7.1 Aderência ao Princípio da Finalidade (Art. 6º, I)

Os resultados demonstram que o CAFIS incorporou o princípio da finalidade de maneira estrutural, uma vez que a análise contextual inicial do dataset orientou todas as decisões subsequentes. Ao identificar que a base pertence ao domínio da saúde, o

framework passou a tratar os dados sob um regime de maior rigor, compatível com a natureza sensível das informações. Essa conduta está em conformidade com o artigo 6º, inciso I, da LGPD, que exige que o tratamento de dados seja realizado para propósitos legítimos, específicos e explícitos. O CAFIS evita, assim, a aplicação genérica de técnicas de anonimização, prática comum em soluções tradicionais, que desconsideram o contexto e podem violar expectativas legítimas dos titulares.

5.7.2 Conformidade com o Princípio da Adequação (Art. 6º, II)

A adequação entre finalidade, contexto e técnica aplicada é evidenciada pelos resultados da classificação automática de atributos e da seleção diferenciada de técnicas. O CAFIS demonstrou capacidade de ajustar o nível de anonimização conforme o tipo de dado, aplicando pseudonimização, generalização ou supressão apenas quando tecnicamente justificável. Essa abordagem atende ao artigo 6º, inciso II, da LGPD, ao assegurar que o tratamento seja compatível com as finalidades informadas e com o contexto do tratamento. Em comparação com frameworks que adotam anonimização uniforme, o CAFIS apresenta maior aderência normativa, pois evita tanto a subproteção quanto a descaracterização excessiva da base.

5.7.3 Aplicação Efetiva do Princípio da Necessidade (Art. 6º, III)

Os resultados indicam que o CAFIS restringiu o tratamento de dados ao mínimo necessário, preservando atributos de baixo risco contextual e anonimização apenas aqueles que ampliariam significativamente o risco de reidentificação. Esse comportamento reflete diretamente o princípio da necessidade, previsto no artigo 6º, inciso III, da LGPD. Do ponto de vista crítico, observa-se que o framework supera abordagens tradicionais que eliminam indiscriminadamente informações relevantes, comprometendo a utilidade dos dados. O CAFIS, ao contrário, demonstra que é possível reduzir riscos sem inviabilizar análises legítimas, reforçando a proporcionalidade exigida pela legislação.

5.7.4 Mitigação de Riscos e Princípio da Prevenção (Art. 6º, VIII)

A etapa de avaliação de sensibilidade e risco de reidentificação evidenciou a aderência do CAFIS ao princípio da prevenção. Ao antecipar cenários de reidentificação direta e indireta, o framework adotou medidas técnicas antes da efetiva disponibilização da base anonimizada. Essa conduta está alinhada ao artigo 6º, inciso VIII, da LGPD, que impõe aos agentes de tratamento o dever de adotar medidas para prevenir a ocorrência de danos. A análise baseada em risco realizada pelo CAFIS reforça a visão da LGPD de que a proteção de dados não deve ser reativa, mas preventiva e contínua.

5.7.5 Distinção Jurídica entre Anonimização e Pseudonimização (Art. 5º, XI)

Os resultados evidenciam que o CAFIS respeita a distinção normativa entre anonimização e pseudonimização estabelecida no artigo 5º, inciso XI, da LGPD. Identificadores diretos foram pseudonimizados com chave, mantendo o dado sob o escopo da lei, enquanto técnicas mais restritivas foram aplicadas a atributos sensíveis conforme o risco. Essa diferenciação é juridicamente relevante, pois evita a falsa suposição de que pseudonimização equivale à anonimização plena. O CAFIS demonstra maturidade normativa ao tratar a pseudonimização como medida de segurança, e não como exclusão automática da incidência legal.

5.7.6 Transparência e Prestação de Contas (Art. 6º, VI e X)

A disponibilização da prévia dos dados anonimizados e a rastreabilidade das decisões técnicas evidenciam aderência aos princípios da transparência e da responsabilização, previstos nos incisos VI e X do artigo 6º da LGPD. O framework permite compreender quais técnicas foram aplicadas, a quais atributos e com base em quais critérios.

Em termos críticos, essa funcionalidade diferencia o CAFIS de soluções opacas, que não oferecem meios de auditoria ou explicação das decisões automatizadas. A geração de evidências técnicas fortalece a governança de dados e facilita a demonstração de conformidade perante autoridades reguladoras e auditorias internas. De forma consolidada, os resultados indicam que o CAFIS apresenta elevado grau de aderência à LGPD, não apenas em nível declarativo, mas operacional. O framework traduz princípios legais em decisões técnicas verificáveis, promovendo:

- proporcionalidade no tratamento dos dados;
- mitigação efetiva de riscos de reidentificação;
- preservação da utilidade informacional;
- fortalecimento da accountability.

Essa análise crítica sugere que o CAFIS contribui para reduzir a lacuna entre norma jurídica e implementação técnica, oferecendo um modelo de anonimização alinhado à lógica de conformidade dinâmica prevista na LGPD.

5.8 Justificativa da Cobertura Parcial (50%) de Aderência à LGPD na Base Analisada

A cobertura aproximada de 50% de aderência à Lei Geral de Proteção de Dados Pessoais (LGPD), identificada na análise da base de dados por meio do framework CAFIS, não configura uma falha do modelo proposto. Trata-se, ao contrário, de um resultado esperado e tecnicamente justificável, decorrente de limitações inerentes ao *dataset* analisado, ao escopo metodológico adotado e à própria estrutura normativa da legislação brasileira de proteção de dados.

Essa aderência parcial decorre, sobretudo, da distinção conceitual entre conformidade técnico-operacional e conformidade jurídica plena, distinção esta que orienta a interpretação dos resultados apresentados a seguir.

5.8.1 A LGPD não se resume à anonimização técnica

A LGPD estabelece um conjunto abrangente de requisitos que extrapolam o tratamento técnico dos dados, abrangendo dimensões organizacionais, jurídicas e procedimentais. Embora o CAFIS atue de forma eficaz na identificação de atributos sensíveis, na aplicação de técnicas de anonimização e na mitigação de riscos de reidentificação, diversos dispositivos legais não podem ser integralmente avaliados apenas a partir da análise de uma base de dados isolada.

Entre os elementos que se encontram fora do escopo direto de verificação técnica do *dataset*, destacam-se:

- a definição da base legal do tratamento (art. 7º);
- a existência de consentimento válido ou outra hipótese legal aplicável;
- o registro das operações de tratamento (art. 37);
- mecanismos de governança organizacional e políticas internas;
- o exercício dos direitos do titular (arts. 18 a 20);
- práticas de compartilhamento de dados com terceiros.

Dessa forma, a cobertura de 50% reflete exclusivamente os princípios e dispositivos passíveis de verificação técnica automatizada, especialmente aqueles relacionados à minimização de dados, prevenção de riscos e segurança da informação.

5.8.2 Presença de dados pessoais mantidos por necessidade funcional

Outro fator determinante para a redução do índice global de aderência foi a manutenção deliberada de determinados atributos sem anonimização, tais como sexo, tipo sanguíneo, uso contínuo de medicação, histórico de internações e datas de eventos clínicos.

Esses atributos foram preservados por apresentarem, de forma isolada, baixo risco de reidentificação e elevada relevância analítica, em conformidade com o princípio da necessidade previsto no art. 6º, inciso III, da LGPD.

Todavia, do ponto de vista jurídico, tais informações permanecem caracterizadas como dados pessoais, o que implica que a base, em sua totalidade, não pode ser considerada completamente excluída do escopo de incidência da LGPD. Esse aspecto impacta diretamente o cálculo do índice global de aderência, sem que isso represente uma omissão ou deficiência técnica do framework.

5.8.3 Uso de pseudonimização em detrimento da anonimização irreversível

O CAFIS aplicou técnicas de pseudonimização com uso de chave a diversos identificadores diretos, como CPF, nome completo, endereço de e-mail e telefone. Embora essa abordagem reduza significativamente os riscos de exposição indevida e reidentificação, ela não descaracteriza juridicamente o dado como pessoal, conforme estabelecido no art. 5º, inciso XI, da LGPD.

A legislação é explícita ao afirmar que apenas dados anonimizados por meios técnicos razoáveis e disponíveis, de forma irreversível, estão excluídos de seu escopo de aplicação. Assim, como parte relevante da base foi submetida apenas à pseudonimização, tais dados permanecem juridicamente protegidos pela LGPD, influenciando negativamente o índice final de aderência.

Esse resultado reflete uma escolha técnica consciente, orientada pela necessidade de preservar rastreabilidade, auditabilidade e possibilidade de validação dos dados, em detrimento de uma anonimização absoluta que inviabilizaria tais processos.

5.8.4 A aderência mensurada refere-se a princípios, e não a artigos isolados

O cálculo de aderência realizado pelo CAFIS fundamenta-se predominantemente nos princípios estabelecidos no art. 6º da LGPD, tais como finalidade, adequação, necessidade, prevenção, segurança e responsabilização.

Entretanto, princípios como transparência plena, livre acesso do titular e *accountability* organizacional exigem a existência de metadados, processos documentais e registros externos ao *dataset*, os quais não estavam disponíveis no contexto da análise realizada.

Assim, a ausência desses elementos não deve ser interpretada como descumprimento normativo, mas como indisponibilidade de evidência técnica mensurável, o que limita a aferição de uma aderência total à legislação.

5.8.5 A lógica de conformidade dinâmica e contextual da LGPD

A LGPD não adota um modelo binário de conformidade, baseado em classificações estritas de conformidade ou não conformidade. Ao contrário, a legislação fundamenta-se em uma abordagem orientada ao risco, à proporcionalidade e ao contexto, conforme evidenciado nos arts. 5º, inciso XI, e 6º.

Sob essa perspectiva, a cobertura de 50% indica que os riscos mais críticos foram mitigados, os dados sensíveis receberam proteção reforçada, a utilidade analítica da base foi preservada e as decisões técnicas adotadas são justificáveis e passíveis de auditoria.

5.8.6 Interpretação crítica dos resultados

Do ponto de vista científico, a cobertura parcial observada evidencia que o CAFIS é eficaz na operacionalização dos princípios técnicos da LGPD, ao mesmo tempo em que demonstra que a conformidade plena depende da integração com camadas organizacionais, jurídicas e procedimentais adicionais.

O resultado reforça a tese central deste trabalho, segundo a qual não existe anonimização universalmente conforme, mas sim processos de anonimização proporcionais ao contexto, à finalidade do tratamento e ao risco associado, sendo a aderência à LGPD um fenômeno progressivo, mensurável e essencialmente contextual.

5.9 Análise Comparativa dos Resultados do CAFIS em Diferentes Bases de Dados

Com o objetivo de avaliar a robustez e a consistência do framework CAFIS em cenários distintos, realizou-se uma análise comparativa dos resultados obtidos a partir da aplicação do framework em múltiplas bases de dados. A Tabela X apresenta os percentuais de cobertura da LGPD, bem como a distribuição de dados pessoais, sensíveis e gerais em cada sessão analisada.

Essa análise comparativa permite compreender como a composição informacional de cada base influencia diretamente o nível de aderência à LGPD alcançado pelo CAFIS, evidenciando o caráter contextual e não determinístico do processo de anonimização.

5.10 Análise Comparativa dos Resultados do CAFIS em Diferentes Bases de Dados

Com o objetivo de avaliar a robustez e a consistência do framework CAFIS em cenários distintos, realizou-se uma análise comparativa dos resultados obtidos a partir da aplicação do framework em múltiplas bases de dados. A Tabela 1 apresenta os percentuais de cobertura da Lei Geral de Proteção de Dados Pessoais (LGPD), bem como a distribuição de dados pessoais, dados sensíveis e dados gerais em cada sessão analisada.

Tabela 1 – Cobertura da LGPD e classificação dos dados por base analisada

nome_arquivo	Cobertura LGPD (%)	Dados Pessoais	Dados Sensíveis	Dados Gerais
sessao_20260108_093521	62,50	3	2	5
sessao_20260108_093506	20,00	0	2	4
sessao_20260108_093446	33,33	0	2	3
sessao_20260108_093429	80,00	1	2	5
sessao_20260108_093353	23,08	0	1	12
sessao_20260108_093328	46,67	5	0	10
sessao_20260108_093309	46,67	6	1	9
sessao_20260108_043554	50,00	5	1	9

Essa análise comparativa permite compreender como a composição informacional de cada base influencia diretamente o nível de aderência à LGPD alcançado pelo CAFIS, evidenciando o caráter contextual e não determinístico do processo de anonimização.

5.10.1 Variação da Cobertura da LGPD entre as Bases

Os resultados demonstram uma variação significativa na cobertura da LGPD, com valores que oscilam entre 20,00% e 80,00%. Essa dispersão indica que a aderência normativa não é uma propriedade fixa do framework, mas sim um resultado emergente da interação entre:

- a composição da base de dados;
- a proporção de dados pessoais e dados sensíveis;
- as decisões técnicas de anonimização aplicáveis a cada contexto.

Bases com maior cobertura, como `sessao_20260108_093429` (80,00%) e `sessao_20260108` (62,50%), apresentam uma combinação mais equilibrada entre dados pessoais, sensíveis e gerais, permitindo ao CAFIS aplicar técnicas proporcionais sem comprometer excessivamente a utilidade dos dados.

Por outro lado, bases com baixa cobertura, como `sessao_20260108_093506` (20,00%) e `sessao_20260108_093353` (23,08%), evidenciam limitações estruturais que impactam diretamente a mensuração da aderência à LGPD.

5.10.2 Influência da Presença de Dados Sensíveis

A análise comparativa revela que a presença de dados sensíveis constitui um dos fatores mais críticos na redução da cobertura da LGPD. Observa-se que bases contendo dois atributos classificados como dados sensíveis, mas poucos ou nenhum dado pessoal adicional, tendem a apresentar baixa cobertura percentual.

Esse resultado decorre do fato de que dados sensíveis, mesmo quando submetidos a técnicas de proteção, permanecem fortemente regulados pela LGPD, exigindo salvaguardas adicionais e restringindo a possibilidade de exclusão do escopo legal. Assim, a simples presença desses dados limita a aderência plena, independentemente do volume total da base.

5.10.3 Relação entre Dados Pessoais e Cobertura Normativa

Outro aspecto relevante refere-se à relação entre o número de dados pessoais e a cobertura da LGPD. Os resultados indicam que bases com maior quantidade de dados pessoais não apresentam, necessariamente, menor aderência normativa. Observa-se que bases como `sessao_20260108_093328` e `sessao_20260108_093309`, apesar de conterem cinco ou mais dados pessoais, alcançam cobertura intermediária (46,67%).

Esse resultado evidencia que a quantidade absoluta de dados pessoais não é o fator determinante para a redução da aderência, mas sim a possibilidade de aplicação eficaz de técnicas de mitigação de risco, como pseudonimização, generalização e supressão seletiva. O CAFIS demonstra, portanto, capacidade de lidar com bases densas em dados pessoais, desde que o contexto permita a adoção de técnicas proporcionais.

5.10.4 Impacto da Predominância de Dados Gerais

Bases com elevada quantidade de dados gerais, como `sessao_20260108_093353`, que apresenta doze atributos dessa natureza, obtiveram baixa cobertura normativa. Esse comportamento indica que, embora dados gerais não estejam diretamente sob o escopo da LGPD, sua predominância não contribui automaticamente para o aumento da aderência, especialmente quando coexistem com dados sensíveis ou quando inexistem evidências explícitas de governança e finalidade do tratamento.

Esse resultado reforça que a cobertura da LGPD não deve ser interpretada como uma métrica puramente quantitativa, mas sim como um indicador qualitativo de conformidade contextual.

5.10.5 Discussão Comparativa dos Resultados

A análise comparativa entre as diferentes bases confirma que o CAFIS responde de forma coerente às variações estruturais dos datasets, ajustando automaticamente suas

decisões técnicas conforme o contexto informacional identificado. A heterogeneidade dos resultados observados não indica inconsistência do framework, mas sim sensibilidade contextual, característica desejável em sistemas alinhados à lógica principiológica da LGPD.

Do ponto de vista científico, esses resultados reforçam três conclusões centrais:

- a aderência à LGPD é gradual e contextual, não binária;
- bases com maior complexidade informacional exigem decisões técnicas mais restritivas, o que tende a reduzir a cobertura percentual;
- o CAFIS é capaz de evidenciar limites reais de conformidade, evitando falsas percepções de adequação plena.

6 CONCLUSÃO

A presente dissertação teve como propósito investigar, estruturar e operacionalizar o processo de anonimização de dados pessoais em ambientes regulados, caracterizados pela interseção entre exigências técnicas, jurídicas e éticas. Conforme discutido ao longo do trabalho, a crescente intensificação do uso de dados, aliada à complexidade dos sistemas baseados em inteligência artificial, evidenciou a existência de uma lacuna significativa entre os princípios normativos estabelecidos pela Lei Geral de Proteção de Dados Pessoais (LGPD) e sua efetiva implementação em sistemas computacionais reais.

Diante desse cenário, foi proposto o CAFIS (Context-Aware Framework for Information Sensitivity), concebido sob a perspectiva da Design Science Research (DSR), como um artefato capaz de integrar análise contextual, classificação de sensibilidade, avaliação de risco de reidentificação e recomendação automatizada de técnicas de anonimização. Diferentemente de abordagens tradicionais, centradas exclusivamente em métricas isoladas ou técnicas específicas, o CAFIS foi projetado para operar de forma sistêmica, incorporando critérios técnicos, jurídicos e éticos em um fluxo contínuo, auditável e orientado à tomada de decisão.

Os resultados apresentados demonstraram que o framework foi capaz de identificar de forma consistente atributos pessoais, sensíveis e quase-identificadores, aplicando heurísticas de classificação e mecanismos de análise contextual que refletem a lógica normativa da LGPD. A etapa de avaliação de risco evidenciou a capacidade do CAFIS de detectar vulnerabilidades associadas à reidentificação, considerando não apenas atributos isolados, mas também suas combinações e correlações internas, o que está alinhado com a literatura contemporânea sobre privacidade baseada em risco.

No que se refere à aplicação das técnicas de anonimização, observou-se que o CAFIS foi capaz de selecionar e aplicar estratégias adequadas de forma proporcional à sensibilidade dos dados, equilibrando proteção e utilidade. A análise dos resultados indicou que técnicas como generalização, supressão e pseudonimização foram empregadas de maneira contextualizada, evitando tanto a exposição indevida quanto a perda excessiva de informação. Esse comportamento reforça a aderência do framework aos princípios da necessidade e da adequação previstos no artigo 6º da LGPD.

A avaliação da utilidade dos dados demonstrou que, mesmo após a anonimização, os conjuntos tratados mantiveram propriedades estatísticas relevantes, permitindo sua utilização em análises, pesquisas e processos decisórios. Esse resultado é particularmente relevante, pois evidencia que a anonimização, quando orientada por critérios contextuais e métricas adequadas, não implica necessariamente perda significativa de valor

informacional, contrariando uma percepção comum na literatura e na prática organizacional.

Sob a perspectiva jurídica, os resultados evidenciaram que a conformidade com a LGPD não deve ser interpretada como um estado binário, mas como um processo dinâmico e gradual, dependente do contexto de uso, da finalidade do tratamento e das características da base de dados. A cobertura parcial de aderência observada em alguns cenários, conforme discutido na Seção 5.8, não representa uma limitação do framework, mas sim um reflexo da própria lógica normativa da legislação, que exige proporcionalidade, prevenção e responsabilização, e não a eliminação absoluta de dados pessoais.

Além disso, o CAFIS demonstrou capacidade de materializar o princípio da accountability por meio da geração de relatórios técnicos e jurídicos que documentam as decisões adotadas ao longo do processo de anonimização. Essa funcionalidade aproxima o domínio técnico do domínio jurídico, permitindo que decisões computacionais sejam compreendidas, auditadas e justificadas, o que representa um avanço relevante no campo da governança de dados.

Outro aspecto relevante observado nos resultados refere-se à variabilidade do desempenho do framework em diferentes bases de dados. A análise comparativa evidenciou que fatores como presença de dados sensíveis, quantidade de quase-identificadores e estrutura da base influenciam diretamente o nível de anonimização alcançado e a cobertura normativa obtida. Esse comportamento reforça a premissa central do trabalho: a anonimização deve ser compreendida como um processo contextual, e não como uma aplicação uniforme de técnicas.

Do ponto de vista científico, esta pesquisa contribui ao propor um modelo que integra, de forma estruturada, conceitos da engenharia de privacidade, da ciência de dados e do direito digital. Ao operacionalizar princípios abstratos da LGPD em mecanismos computacionais verificáveis, o CAFIS reduz a distância entre teoria e prática, oferecendo um instrumento capaz de apoiar decisões reais em ambientes organizacionais complexos.

Por fim, conclui-se que o CAFIS constitui uma contribuição relevante tanto para a academia quanto para a prática profissional, ao oferecer uma abordagem automatizada, auditável e adaptável para anonimização de dados pessoais. O framework demonstra que é possível conciliar proteção de dados, conformidade regulatória e preservação da utilidade analítica, desde que o processo seja orientado por critérios contextuais, métricas adequadas e princípios jurídicos bem definidos.

6.1 Trabalhos Futuros

Apesar dos avanços alcançados com o desenvolvimento e avaliação do CAFIS, diversas oportunidades de aprimoramento e expansão do framework podem ser exploradas

em trabalhos futuros, tanto do ponto de vista técnico quanto jurídico e metodológico.

Uma primeira linha de evolução consiste na incorporação de técnicas baseadas em aprendizado de máquina para a calibragem dinâmica das estratégias de anonimização. Modelos supervisionados ou por reforço poderiam ser utilizados para ajustar automaticamente parâmetros como níveis de generalização, intensidade de supressão ou aplicação de ruído, considerando métricas contínuas de risco de reidentificação e perda de utilidade. Essa abordagem permitiria ao framework evoluir de um sistema heurístico para um sistema adaptativo, capaz de aprender com diferentes cenários de dados.

Outra direção promissora envolve a integração de mecanismos de Explainable Artificial Intelligence (XAI), com o objetivo de tornar as decisões de anonimização mais transparentes e interpretáveis. A explicabilidade das recomendações técnicas pode fortalecer a auditoria, facilitar a compreensão por parte de profissionais não técnicos e contribuir para a conformidade com os princípios de transparência e accountability previstos na LGPD.

Adicionalmente, a expansão do framework para suportar múltiplas jurisdições representa um avanço relevante. A adaptação do CAFIS para contemplar regulamentações como a GDPR europeia e a CCPA norte-americana permitiria sua aplicação em ambientes internacionais, exigindo a incorporação de regras normativas específicas e mecanismos de parametrização regulatória.

Do ponto de vista técnico, destaca-se a possibilidade de extensão do framework para dados não estruturados, como textos, imagens e dados biométricos. Esse desafio implica a integração com técnicas de processamento de linguagem natural, visão computacional e anonimização multimodal, ampliando significativamente o escopo de aplicação do CAFIS.

Outra linha relevante refere-se à utilização de dados sintéticos e aprendizado federado como estratégias complementares à anonimização tradicional. A geração de bases sintéticas pode reduzir riscos de exposição, enquanto o aprendizado federado permite a análise distribuída sem centralização de dados sensíveis, alinhando-se aos princípios de minimização e necessidade.

Também se destaca a necessidade de aprofundar a integração do CAFIS com processos organizacionais de governança de dados, incluindo gestão de consentimento, controle de acesso, políticas internas e fluxos de auditoria contínua. Essa evolução permitiria transformar o framework em um componente mais abrangente dentro de ecossistemas corporativos de conformidade.

Por fim, recomenda-se a realização de avaliações empíricas em ambientes reais, com bases de dados operacionais e participação de organizações públicas e privadas. Essa validação prática poderá evidenciar limitações não observadas em cenários simulados e

contribuir para o refinamento do framework em contextos de uso efetivo.

Assim, os trabalhos futuros apontam para a evolução do CAFIS em direção a um sistema mais inteligente, explicável, adaptativo e integrado, capaz de responder às demandas crescentes de proteção de dados em um cenário tecnológico e regulatório em constante transformação.

REFERÊNCIAS

ABNT. Associação Brasileira de Normas Técnicas. NBR ISO/IEC 27002:2013 – Tecnologia da informação – Técnicas de segurança – Código de prática para controles de segurança da informação. Rio de Janeiro, 2013.

ALFORD, Kenneth. Machine learning and privacy: risks and opportunities. New York: Springer, 2020.

ALMEIDA, M. B.; SILVA, F. A.; SOUZA, R. P. Privacidade e segurança em serviços digitais: desafios contemporâneos. Revista Brasileira de Computação, v. 12, n. 3, p. 45–60, 2020.

BRASIL. Autoridade Nacional de Proteção de Dados (ANPD). Regulamentações complementares à LGPD. Brasília: ANPD, 2022.

BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, 2018.

CASTELLS, M. A sociedade em rede. São Paulo: Paz e Terra, 1999.

CAVOUKIAN, A. Privacy by design: the 7 foundational principles. Toronto: Information and Privacy Commissioner of Ontario, 2011.

DE STEFANI, Federica. Le Regole della Privacy: Guida Pratica al Nuovo GDPR. Milano: Franco Angeli, 2018.

DOMINGO-FERRER, Josep; SÁNCHEZ, David; SORIA-COMAS, Jordi. Database Anonymization: Privacy Models, Data Utility, and Microaggregation Methods. Cham: Springer, 2016.

DONEDA, Danilo. Da privacidade à proteção de dados pessoais. Rio de Janeiro: Renovar, 2006.

ERYUREK, Evren; GILAD, Uri; LAKSHMANAN, Valliappa; et al. Data Governance: The Definitive Guide. Sebastopol: O'Reilly Media, 2021.

EUROPEAN PARLIAMENT. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016. General Data Protection Regulation (GDPR). Official Journal of the European Union, 2016.

EUROPEAN UNION. General Data Protection Regulation (GDPR). Official Journal of the European Union, 2016. Disponível em: https://eur-lex.europa.eu/eli/reg/2016/679/oj. Acesso em: 26 out. 2025.

FABIANO, André. Autodeterminação informativa e proteção de dados pessoais. São Paulo: Revista dos Tribunais, 2020.

FABIANO, André. Privacidade, anonimização e pseudonimização na LGPD. São Paulo: Revista dos Tribunais, 2021.

FABIANO, Nicola. GDPR e Privacy: consapevolezza e opportunità. Roma: Key Editore, 2019.

FABIANO, Nicola. GDPR e Privacy: L'approccio con il Data Protection and Privacy Relationships Model (DAPPREMO). Roma: Key Editore, 2020.

FERETZAKIS, Georgios; et al. Responsible Use of Synthetic and Anonymized Data in the Age of AI. Future Internet, v. 17, n. 151, 2025. DOI: 10.3390/fi17010151.

FIESCHI, Andrea; HIRMER, Pascal; STACH, Christoph. Characterising and categorising anonymization techniques: A literature-based approach. ICISSP 2025. Lisboa: SCITEPRESS, 2025. DOI: 10.5220/0013379100003899.

FIESCHI, Andrea; HIRMER, Pascal; STACH, Christoph. Discovering suitable anonymization techniques: A privacy toolbox for data experts. BTW 2025 – Lecture Notes in Informatics. Bonn: Gesellschaft für Informatik, 2025. DOI: 10.18420/BTW2025-48.

FUNG, Benjamin C. M. et al.Privacy-preserving data publishing: a survey of recent developments. ACM Computing Surveys, v. 42, n. 4, p. 1–53, 2010.

GADOTTI, Andrea; ROCHER, Luc; HOUSSIAU, Florimond; CREȚU, Ana-Maria; DE MONTJOYE, Yves-Alexandre. Anonymization: The imperfect science of using data while preserving privacy. Science Advances, v. 10, eadn7053, 2024. DOI: 10.1126/sciadv.adn7053.

HANISCH, Simon; et al. A False Sense of Privacy: Towards a Reliable Evaluation Methodology for the Anonymization of Biometric Data. PETS, Dresden, 2024. Disponível em: https://arxiv.org/abs/2304.01635v2. Acesso em: 26 out. 2025.

HEVNER, Alan R.; et al. Design Science in Information Systems Research. MIS Quarterly, v. 28, n. 1, p. 75–105, 2004.

HOLZENBERGER, Nils; MAXWELL, Winston. A Quantitative Approach to the GDPR's Anonymisation and Pseudonymisation Tests. Télécom Paris Working Paper, 2024. Disponível em: https://arxiv.org/abs/2412.06001v1. Acesso em: 26 out. 2025.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. ISO/IEC 29100:2024 — Privacy Framework. Genebra: ISO, 2024.

ISO. ISO 9241-210: Ergonomics of human-system interaction. Geneva: ISO, 2019.

KARAGIANNIS, Stylianos; et al. Mastering Data Privacy: Leveraging K-Anonymity for Robust Health Data Sharing. IJIS, v. 23, p. 2189–2201, 2024. DOI: 10.1007/s10207-024-00838-8.

KOHLS, César; OLIVEIRA, Thiago; FERREIRA, Bruno. Autodeterminação informativa e governança de dados na LGPD. Revista Brasileira de Direito Digital, v. 3, n. 2, p. 55–74, 2021.

LESTYÁN, Szilvia; et al. Anonymity-washing. arXiv preprint, 2025. Disponível em: https://arxiv.org/abs/2505.18627v1. Acesso em: 26 out. 2025.

LI, Ninghui; LI, Tiancheng; VENKATASUBRAMANIAN, Suresh. t-closeness: privacy beyond k-anonymity and l-diversity. In: IEEE 23rd International Conference on Data Engineering. 2007. p. 106–115.

MACHANAVAJJHALA, Ashwin et al.l-diversity: privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data, v. 1, n. 1, p. 1–52, 2007.

MALANIN, Vladyslav. Anonymization Techniques for Large-Scale Health Databases: A Critical Review. IJIRSET, v. 14, n. 5, p. 11719–11735, maio 2025. DOI: 10.15680/IJIRSET.2025.1405015.

MALDONADO, Viviane; BLUM, Renato. Lei Geral de Proteção de Dados comentada. São Paulo: Thomson Reuters, 2020.

MARTIN, James. Advanced privacy-preserving techniques: from k-anonymity to differential privacy. Oxford: Oxford University Press, 2022.

MARTIN, Robert C. Clean Architecture. Boston: Prentice Hall, 2019.

MENDES, L. C.; OLIVEIRA, R. S. Proteção de dados e governança digital no Brasil. Revista de Direito e Tecnologia, v. 5, n. 2, p. 115–140, 2021.

MONTEIRO, Mariana; et al. Patterns for anonymization, pseudonymization and perturbation. EuroPLoP 2024. New York: ACM, 2024. DOI: 10.1145/3698322.3698360.

MONTEIRO, Mariana; et al. Patterns of data anonymization. EuroPLoP 2024. New York: ACM, 2024. DOI: 10.1145/3698322.3698337.

PAU, David; et al. Comparison of anonymization techniques regarding statistical reproducibility. PLOS Digital Health, v. 4, n. 2, e0000735, 2025. DOI: 10.1371/journal.pdig.0000735.

PEFFERS, Ken; et al. A Design Science Research Methodology for Information Systems Research. JMIS, v. 24, n. 3, p. 45–77, 2007.

PISSARRA, David; et al. Unlocking the Potential of Large Language Models for Clinical Text Anonymization. arXiv preprint, 2024. Disponível em: https://arxiv.org/abs/2406.000 Acesso em: 26 out. 2025.

PRESSMAN, Roger S.; MAXIM, Bruce. Engenharia de Software. Porto Alegre: AMGH, 2020.

RAGHUNATHAN, Balaji. The Complete Book of Data Anonymization. Boca Raton: CRC Press, 2013.

REMPE, Moritz; et al. De-Identification of Medical Imaging Data. arXiv preprint, 2024. Disponível em: https://arxiv.org/abs/2410.12402v1. Acesso em: 26 out. 2025.

SOLOVE, Daniel J. Understanding privacy. Harvard University Press, 2008.

SOMMERVILLE, Ian. Engenharia de Software. São Paulo: Pearson, 2019.

SHANNON, Claude E. A mathematical theory of communication. Bell System Technical Journal, v. 27, p. 379–423, 623–656, 1948.

SWEENEY, Latanya. k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, v. 10, n. 5, p. 557–570, 2002.

TAAL, Walter. Information law and ethics in the digital society. Amsterdam: Elsevier, 2022.

TAVANI, H. T. Ethics and Technology. New Jersey: Wiley, 2016.

THE MITRE CORPORATION. Privacy Maturity Model. Bedford: MITRE, 2019.

VIGLIAR, José. Privacidade e dignidade na sociedade da informação. Rio de Janeiro: Forense, 2021.

WARREN, S. D.; BRANDEIS, L. D. The right to privacy. Harvard Law Review, v. 4, n. 5, p. 193–220, 1890.

WEBER, Rolf H.; HEINRICH, Ulrike I. Anonymization: The Global Legal Perspective. Berlin: Springer, 2012.