UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FLÁVIO DA SILVA NEVES

SMART ANONYMITY: Um Mecanismo para Recomendação de Algoritmos de Anonimização de Dados Baseado no Perfil dos Dados para Ambientes IoT

Recife

2023

FLÁVIO DA SILVA NEVES

SMART ANONYMITY: Um Mecanismo para Recomendação de Algoritmos de Anonimização de Dados Baseado no Perfil dos Dados para Ambientes IoT

Tese de Doutorado apresentada ao Centro de Informática da Universidade Federal de Pernambuco em cumprimento parcial dos requisitos para o título de Doutor em Ciência da Computação. Área de Concentração: Engenharia de Software e Linguagens de Programação.

Orientador Dr. Vinicius Cardoso Garcia (Universidade Federal de Pernambuco, Brasil)

Co-orientador Dr. Michel Sales Bonfim (Universidade Federal do Ceará, Brasil)

Recife

2023

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

N518s Neves, Flávio da Silva
Smart anonymity: um mecanismo para recomendação de algoritmos de anonimização de dados baseado no perfil dos dados para ambientes IoT / Flávio da Silva Neves. – 2023.
147 f.: il., fig., tab.

Orientador: Vinicius Cardoso Garcia.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2023.
Inclui referências.

1. Engenharia de software. 2. Internet das coisas. I. Garcia, Vinicius Cardoso (orientador). II. Título.

005.1 CDD (23. ed.) UFPE - CCEN 2023-170

Flávio da Silva Neves

“Smart Anonymity: Um Mecanismo para Recomendação de Algoritmos de Anonimização de Dados Baseado no Perfil dos Dados para Ambientes IoT”

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de Concentração: Engenharia de Software e Linguagens de Programação

Aprovado em: 22/06/2023.

_____

Orientador: Prof. Dr. Vinicius Cardoso Garcia

BANCA EXAMINADORA

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Centro de Informática / UFPE

_____

Profa. Dra. Jessyka Flavyanne Ferreira Vilela
Centro de Informática / UFPE

_____

Prof. Dr. Eduardo Luzeiro Feitosa
Instituto de Computação / UFAM

_____

Prof. Dr. Fernando Antônio Aires Lins
Departamento de Computação / UFRPE

_____

Prof. Dr. Paulo Antônio Leal Rego
Departamento de Computação / UFC

Dedico esse trabalho a minha família e amigos, em especial aos meus pais, Irani Maria da Silva Neves e José Maria Neves da Silva.

AGRADECIMENTOS

Primeiramente agradeço a Deus por ter me dado a motivação necessária ao longo do processo de doutoramento.

Ao meu orientador, professor Dr. Vinicius Cardoso Garcia, que me apoiou, sempre mostrando as melhores soluções, com sua experiência e conhecimento. Ao meu coorientador, Michel Sales Bonfim, por todo apoio no desenvolvimento desta Tese.

Aos meus pais Irani e José Maria, que apesar de muitas dificuldades me deram as condições e apoio necessários para que eu pudesse estudar, e também a toda a minha família, que sempre me apoiaram e incentivaram a continuar estudando.

Um agradecimento especial à minha esposa, Juliana Valença de S. Neves, por ter me incentivado e apoiado durante todos os momentos. Muito obrigado por estar ao meu lado me apoiando e incentivando em todos os momentos.

Aos demais docentes do programa de Pós-Graduação do Centro de Informática da UFPE, pela competência com que transmitiram os conteúdos e ensinamentos.

Agradeço a todos os amigos que tive a oportunidade de conhecer no Centro de Informática, que por muitas vezes me ajudaram durante o processo de doutoramento, pois sem as trocas de ideias esta pesquisa não seria possível.

A todos os meus amigos que de alguma forma contribuíram e participaram na realização deste trabalho.

À Capes e ao CIn – UFPE, pelo apoio financeiro para realização desta pesquisa.

A todos, os meus mais sinceros agradecimentos.

RESUMO

A Internet das Coisas (IoT) prevê um mundo onde os dispositivos do dia a dia estão conectados à internet, interagindo entre si e com o ambiente ao seu redor. Os dados coletados pelos dispositivos IoT são processados para fornecer serviços aos seus usuários. Existem inúmeros dispositivos espalhados por vários locais, tais como casas inteligentes, carros, locais públicos, bem como dispositivos que as pessoas usam em seu corpo, sem saber das suas reais capacidades, como, por exemplo, *smartwatchs*. Esses dispositivos coletam os mais variados tipos de dados dos seus usuários e a exposição desses dados pode colocar a privacidade de seus usuários em risco. Diante disso, o objetivo desta pesquisa é desenvolver o *Smart Anonymity*, que é uma solução que recomenda o algoritmo de anonimização de dados mais adequado para um conjunto de dados de acordo com suas características. As principais contribuições desta pesquisa são: (i) desenvolvimento do *Smart Anonymity*; (ii) criação dos critérios para escolha de algoritmos de anonimização baseado nas características dos dados; (iii) duas ontologias para dar suporte à classificação dos dados; (iv) o uso de *Machine Learning* para melhorar os resultados da classificação realizada pelas ontologias. Com base nos resultados das avaliações realizadas no decorrer desta tese, é possível concluir que o *Smart Anonymity* alcançou resultados promissores para a classificação e recomendação dos algoritmos de anonimização para dados gerados por dispositivos IoT. Também é possível concluir que o uso de *Machine Learning* traz melhorias nos resultados do processo de classificação dos dados gerados por dispositivos da IoT.

Palavras-chave: internet das coisas; privacidade; anonimização de dados; k-anonymity; segurança; recomendação.

ABSTRACT

The Internet of Things (IoT) envisions a world where everyday devices are connected to the Internet, interacting with each other and the environment around them. The data collected by IoT devices is processed to provide services to their users. There are countless devices scattered around various locations, such as smart homes, cars, and public places, as well as devices that people wear on their bodies, unaware of their true capabilities, such as smartwatches. These devices collect all kinds of data from their users, and exposing this data can put their users' privacy at risk. Because of this, the aim of this research is to develop Smart Anonymity, which is a solution that recommends the most suitable data anonymization algorithm for a data set according to its characteristics. The main research's contributions are (i) the development of Smart Anonymity, (ii) the creation of criteria for choosing anonymization algorithms based on data characteristics, (iii) two ontologies to support data classification; (iv) the use of Machine Learning to improve the ontologies' results. Based on the evaluations conducted in this thesis, it is possible to conclude that Smart Anonymity has achieved promising results in the classification and recommendation of anonymization algorithms for data generated by devices. It is also possible to conclude that using Machine Learning improves classifying data generated by IoT devices.

Keywords: internet of things; privacy; data anonymization; k-anonymity; security; recommendation.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AWS	*Amazon Web Services*
cPPDC	*Cooperative Privacy-Preserving Data Collection protocol*
CSV	*Comma-Separated Values*
DL	*Description Logic*
DSR	*Design Science Research*
ESOT	*Enhanced Semantic Obfuscation Technique*
FPGA	*Field-Programmable Gate Array*
GDPR	*General Data Protection Regulation*
IA	Inteligência Artificia
IoT	*Internet of Thing*
K-VARP	*K-anonymity for VARied data stream via Partitioning*
KNN	*K-nearest neighbor*
KR	*Knowledge Representation*
LBS	*Location Based Service*
LGPD	Lei Geral de Proteção de Dados Pessoais
OWL	*Web Ontology Language*
PPM	*privacy preservation module*
RDF	*Resource Description Frameworks*
RMSE	*Root Mean Squared Error*
SIoT	*Social Internet of Things*
SLR	*Systematic Literature Review*
SOT	*Semantic Obfuscation Technique*
SVM	*Support vector machine*
TSM	*Trajectory Selection Mechanism*
UDCMi	*Urban Design Center Misono*

XML *Markup Language*

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta a contextualização, motivação e justificativa para realização desta pesquisa, bem como os problemas de pesquisa e a definição das hipóteses de pesquisa. São apresentados os objetivos geral e específicos desta pesquisa. Também são descritas as contribuições e a metodologia usada para a condução desta pesquisa. A organização de todo o trabalho é apresentada no final deste capítulo.

1.1 CONTEXTUALIZAÇÃO

A *Internet of Thing* (IoT) consiste essencialmente em sensores, responsáveis por captar os mais variados tipos de dados sobre o meio ambiente e atuadores que executam tarefas de acordo com o que foi captado pelos sensores. Portanto, a IoT é capaz de conectar dispositivos e incorporá-los ao sistema de comunicação de modo a processar inteligentemente suas informações específicas e tomar decisões autônomas (ULLAH; SHAH, 2016).

Para Borgia (2014), a aplicação de IoT vem sendo utilizada em três grandes domínios: (i) domínio industrial, (ii) domínio de cidades inteligentes, e (iii) domínio da saúde e bem-estar. A IoT possibilita o desenvolvimento de novas aplicações inteligentes. Isso acontece principalmente devido à sua dupla capacidade de executar e detectar situações (por exemplo, coletar informações sobre fenômenos naturais, parâmetros médicos, ou os hábitos do usuário), e, a partir dessas informações, oferecer serviços personalizados. Independente da aplicação, a IoT destina-se a oferecer uma melhor qualidade de vida às pessoas, e tem profundo impacto sobre a economia e a sociedade.

A partir dos três domínios citados anteriormente, a IoT ainda pode ser subdividida em mais 9 subdomínios (BORGIA, 2014), são eles: (i) Gerenciamento de logística e vida útil do produto; (ii) Agricultura e criação de animais; (iii) Processamento industrial; (iv) Mobilidade inteligente e Turismo inteligente; (v) *Smart Grid*; (vi) Casa / edifício inteligente; (vii) Monitor de segurança pública e meio ambiente; (viii) Vida independente; e (ix) Medicina e saúde. Na literatura, ainda é possível encontrar mais um subdomínio, que é o (x) Campus inteligente (HOSSAIN; DAS; RASHED, 2019; SASTRA; WIHARTA, 2016; DU; MENG; GAO, 2016; ALGHAMDI; SHETTY, 2016). Em sua pesquisa, Pekar et al. (2020) divide a IoT em oito domínios. Essa divisão é bem semelhante à proposta de Borgia (2014).

Vale salientar que dentre a subdivisão dos domínios citados por Borgia (2014), nem todas as aplicações da IoT têm atualmente o mesmo nível de maturidade. Alguns aplicativos, normalmente os mais simples e mais intuitivos para o usuário, já fazem parte da vida cotidiana das pessoas como, por exemplo, os s*mart watch*. Muitos outros ainda estão em fase experimental, pois eles exigem uma maior cooperação entre os vários participantes do ecossistema como, por exemplo, carros autônomos. Finalmente, outros são mais futuristas, e estão em um estágio inicial de seu desenvolvimento como, por exemplo, carros voadores.

1.2 MOTIVAÇÃO

A IoT está sendo usada nos mais variados campos e já faz parte da vida cotidiana de milhões de pessoas no mundo. Mesmo que as pessoas não consigam perceber diretamente a IoT no seu cotidiano, os dispositivos da IoT já fazem parte de suas vidas como, por exemplo, relógios inteligentes. Esse paradigma prevê um mundo onde tudo está conectado, desde os mais diversos sensores (ex. sensor de temperatura, de luminosidade ou até mesmo de presença) até os objetos do dia a dia (cafeteira, geladeira, lâmpadas), possibilitando o monitoramento e o controle desses objetos remotamente (COLLINA et al., 2014).

A IoT apresenta uma vasta gama de potencialidades, desde a ajuda na melhoria na qualidade de vida das pessoas idosas, até o acompanhamento da saúde das pessoas, bem como o monitoramento da segurança pública, ou controle de processos industriais, ou ainda o monitoramento de plantações em fazendas. A IoT pode ser usada nas principais áreas de atuação do homem para ajudar a melhorá-las, assim os benefícios contribuem para a produtividade, redução de custos de produção, melhoria da qualidade de vida, entre outras possibilidades (BORGIA, 2014; ALDEEN; SALLEH, 2019).

As perspectivas para a IoT no futuro são otimistas. Presume-se que cada vez mais os dispositivos IoT irão auxiliar as pessoas no sentido de tornar a vida humana mais cômoda e produtiva. Cada vez mais os dispositivos irão se espalhar e tornar-se parte do cotidiano das pessoas, direta ou indiretamente (ALDEEN; SALLEH, 2019; BHATTACHARYA; PANDEY, 2020). A quantidade de dispositivos IoT vem crescendo muito nos últimos anos. Por exemplo, a empresa de consultoria *Gartner* fez um estudo que estima que o número de conexões IoT irá crescer de 6 milhões em 2015 para 27 bilhões em 2025 (GARTNER, 2020).

Atrelada a um grande número de dispositivos e o enorme volume de dados gerados por esses dispositivos, surge um impasse que pode ser considerado preocupante: a privacidade

dos dados dos usuários desses dispositivos IoT. A maioria dos usuários não está disposta a compartilhar seus dados pessoais diretamente com terceiros, seja para pesquisa acadêmica ou análise comercial, porque os dados pessoais contêm informações privadas ou confidenciais, como situação econômica ou hábitos de vida (LIU et al., 2019). Diante disso, a coleta autônoma dos dados pessoais torna a privacidade uma das principais preocupações legais, éticas e/ou tecnológicas com relação à IoT na atualidade (BERREHILI; BELMEKKI, 2016; HARADAT et al., 2018).

Diante de problemas relacionados a privacidade, surgiram várias iniciativas não só no contexto de IoT. Atualmente existem algumas iniciativas legais para tratar problemas relacionados a privacidade dos dados dos usuários. Nos EUA, existem leis para tratar problemas relacionados a privacidade de dados; na União Europeia, tem a *General Data Protection Regulation* (GDPR)[1]; e, no Brasil, em 14 de agosto de 2018, foi sancionada a primeira lei para privacidade de dados: a Lei Geral de Proteção de Dados Pessoais (LGPD)[2]. Essas iniciativas mostram a importância e a preocupação de se manter a privacidade dos dados dos usuários.

Há sensores e controladores em diversos locais que prometem economia de energia, automação e outros tipos de controles de gastos e conveniências referente aos custos (SCHURGOT; SHINBERG; GREENWALD, 2015; BORGIA, 2014). Estes tipos de dispositivos acabam por capturar inúmeras informações do cotidiano dos seus usuários, informações estas que podem ser usadas para os mais diversos fins. Estas informações podem ser usadas para melhorar as funcionalidades destes dispositivos, bem como, podem ajudar a oferecer produtos mais adequados ao gosto dos usuários de acordo com seu perfil (ZIEGELDORF; MORCHON; WEHRLE, 2014). Contudo, há a possibilidade de essas informações serem vendidas para terceiros, ou até mesmo numa eventualidade, utilizadas por uma pessoa mal-intencionada a fim de prejudicar os usuários, seja financeiramente ou por outra finalidade escusa.

Existe uma gama de possibilidades para o uso de informações pessoais, possibilidades de uso que nem sempre têm por objetivo oferecer algo benéfico para as pessoas. Inclusive, em algumas ocasiões, tem-se por objetivo prejudicar como, por exemplo, identificar se há pessoas na casa ou quais os horários essas pessoas costumam sair. Este é o tipo de informação que pode ser bastante útil para pessoas mal-intencionadas, que poderiam usar essas informações para invadirem uma casa quando não houver ninguém.

[1] https://op.europa.eu/en/publication-detail/-/publication/3e485e15-11bd-11e6-ba9a-01aa75ed71a1/language-en

[2] https://www.planalto.gov.br/ccivil$_03/_{ato}2015-2018/2018/lei/l13709.htm$

1.3 PROBLEMA E QUESTÕES DE PESQUISA

Os dispositivos IoT podem ter acesso aos dados pessoais, particularmente em aplicativos de IoT relacionados à vida humana, como dispositivo pessoal, casa inteligente, saúde inteligente, cidade inteligente, entre outros. Em tais aplicativos, esses dispositivos são usados para coletar uma enorme quantidade de dados pessoais usados pelos provedores desses aplicativos para inúmeros objetivos, como criação de perfil para publicidade, *marketing* digital, estatísticas, dentre outros. O uso dos dados pessoais, sem uma autorização do proprietário dos dados, pode ser considerada uma violação do direito humano universal, que é a privacidade dos seus dados (BERREHILI; BELMEKKI, 2016).

Os desafios apresentados na literatura sobre a IoT estão normalmente relacionados à falta de padronização na comunicação de diferentes tipos de dispositivos e a heterogeneidade dos dispositivos. Dentre os desafios, um dos fatores considerados preocupantes é a privacidade dos dados gerados por esses dispositivos (ABOMHARA; KØIEN, 2014; HARADAT et al., 2018), que têm características únicas, dada a limitação de recurso de *hardware* destes dispositivos (capacidade de processamento, armazenamento e capacidade de energia). As limitações dos *hardwares* usados na IoT limitam o uso de criptografia convencional usada em aplicações locais ou *Web* em função da quantidade crescente de dados. Estes dados são provenientes de localização, históricos de pesquisa no site e dados de uso de energia, entre outros, e são usados em vários tipos de serviços. No entanto, esses dados não podem ser usados facilmente para outros fins, devido aos problemas de privacidade. Portanto, a proteção da privacidade é necessária no processo de aplicação desses dados em usos secundários, onde o anonimato é a solução usual (HARADAT et al., 2018).

A anonimização dos dados permite que dados úteis sejam publicados e até mesmo armazenados com segurança, sem que se revelem informações pessoais. Esse é o recurso mais característico e marcante do anonimato e o difere de outros métodos de fornecimento de serviço criptografados. Dados anonimizados podem ser entregues a terceiros para fornecer serviços úteis com o uso secundário destes. Contudo, mesmo se os dados anonimizados forem expostos por meio de vazamentos, o impacto social será significativamente menor do que quando os dados originais são divulgados (OTGONBAYAR; PERVEZ; DAHAL, 2016; OTGONBAYAR et al., 2018; SHOHATA; NAKAMURA; NISHI, 2018; LI; PALANISAMY, 2019).

Dentre as várias soluções existentes para tratar privacidade em IoT, a anonimização de dados se mostra uma solução bem aceita na comunidade científica (ELKHODR; SHAHRESTANI;

CHEUNG, 2012; SAMANI; GHENNIWA; WAHAISHI, 2015; BERREHILI; BELMEKKI, 2016; DAVOLI; PROTSKAYA; VELTRI, 2017; TAKBIRI et al., 2018). Existem várias técnicas usadas para anonimizar dados em IoT como, por exemplo, perturbação e ofuscação. Essas técnicas são usadas em vários contextos de aplicação da IoT, contudo, ainda não se tem uma técnica considerada ampla o suficiente para atender a várias áreas do IoT.

Neste contexto, baseado nos resultados da *Systematic Literature Review* (SLR) descrita no Capítulo 3, uma solução baseada em anonimização de dados que seja aplicável nos diferentes contextos de uso da IoT, até o presente momento, é considerada uma questão em aberto. Além disso, conforme a SLR, nenhum outro trabalho na literatura forneceu uma solução que possa ser aplicada nos vários campos de aplicação da IoT.

Diante do exposto, o problema que permeia esta pesquisa é: **Como desenvolver uma solução para privacidade de dados, baseada em anonimização, que seja autoadaptável para vários ambientes de uso da IoT.**

Em face deste cenário, fica evidente que uma solução automatizada, que possa recomendar o algoritmo de anonimização mais adequado para um conjunto de dados gerado por dispositivos IoT tem um impacto significativo para melhorar o tratamento da privacidade dos dados coletados pelos dispositivos IoT. Para isso, esta Tese se baseia nas seguintes premissas:

1. A IoT pode ser dividida em onze subdomínios, descritos na Seção 4.1. Cada subdomínio tem uma vasta gama de tipos de dispositivos que coletam uma grande quantidade de dados dos seus usuários. Em contrapartida, existem diversas formas de tratar a privacidade dos usuários. Uma abordagem muito utilizada é a anonimização (BERREHILI; BELMEKKI, 2016; DAVOLI; PROTSKAYA; VELTRI, 2017; TAKBIRI et al., 2018). Porém, existem várias técnicas de anonimização de dados implementadas em diversos algoritmos que podem ser usadas para tratar a privacidade dos usuários de dispositivos IoT, conforme as apresentadas por Neves et al. (2023) em sua Revisão Sistemática da Literatura.

2. Para selecionar o melhor algoritmo de anonimização para os diferentes tipos de dados, é necessário que as características dos dados sejam consideradas. Então, para a recomendação dos algoritmos de anonimização, uma abordagem baseada em sistemas de recomendação é promissora. Os sistemas de recomendação conseguem recomendar o algoritmo de anonimização mais adequado para um determinado conjunto de dados considerando suas características. Por sua vez, as ontologias baseadas em lógica de descrição conseguem classificar dados de acordo com suas características e inferir novos

conhecimentos (COSTA, 2020). Dessa forma, podem ser usadas para classificar os dados em categorias, para que um sistema de recomendação possa recomendar o algoritmo de anonimização mais adequado para determinado conjunto de dados. Por sua vez, a *Machine Learning* pode ser utilizada para tornar a classificação dos dados mais assertiva e inteligente conforme as características dos dados.

Assim, com base no problema citado acima e as tecnologias apresentadas, serão investigadas as seguintes hipóteses:

- **Hipótese 1:** o uso de recomendação de algoritmos de anonimização de dados vai possibilitar que seja aplicado o algoritmo de anonimização mais adequado para um determinado conjunto de dados. Desta forma, haverá maior privacidade em relação aos dados pessoais para os usuários de dispositivos IoT nos vários subdomínios da IoT, e maiores possibilidades para o uso dos dados por parte das empresas; e
- **Hipótese 2:** o uso de *Machine Learning* melhora os resultados da classificação e recomendação para identificar as melhores técnicas de anonimização baseado nas características dos dados no contexto de IoT.

Neste sentindo, esta Tese propõe o *Smart Anonymity*, uma solução inteligente para a recomendação de algoritmos de anonimização, sensível ao contexto dos dados. Para tanto, ele faz uso de ontologias *Web Ontology Language* (OWL) e *Description Logic* (DL) para fazer a classificação dos dados. Para melhorar os resultados da classificação feita por meio das ontologias desenvolvidas, a abordagem desenvolvida usa *Machine Learning*, a mesma também é utilizada para fazer a recomendação dos algoritmos de anonimização, para essa recomendação é usado o algoritmo *Random Forest*, porque ele poderia ser usado em situações mais complexas no futuro com a expansão da quantidade de subdomínios e dos algoritmos de anonimização. Diante do exposto, esta pesquisa procurou responder a cinco questões de pesquisa:

- **RQ1:** Como anonimização de dados pode fornecer privacidade de dados aos usuários considerando o contexto e os vários ambientes de uso da IoT?
- **RQ2:** quais critérios devem ser considerados para recomendar algoritmos de anonimização para diferentes tipos de dados?

- **RQ3:** Como Lógica de Descrição (DL) pode contribuir para que o mecanismo proposto possa definir qual algoritmo de anonimização recomendar, de acordo com as características dos dados no contexto de ambientes IoT?
- **RQ4:** Como a *Machine Learning* pode contribuir para que o mecanismo proposto seja capaz de aprender conforme analisa os dados?
- **RQ5:** Como medir e analisar a qualidade da classificação e recomendação baseado nas características dos dados?

1.4 OBJETIVO GERAL

Esta Tese tem por objetivo geral **desenvolver uma solução inteligente que recomende qual o algoritmo de anonimização de dados é o mais adequado para o conjunto de dados de acordo com suas características em ambientes IoT**. Para alcançar o objetivo geral desta pesquisa, a classificação dos dados é feita por meio de duas ontologias OWL baseadas em Lógica de Descrição. Já para melhorar a classificação feita pelas ontologias, é utilizada *Machine Learning*. Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos (OE):

1.5 OBJETIVOS ESPECÍFICOS

- **OE1:** definir os critérios para fazer a recomendação dos algoritmos de anonimização para os diferentes subdomínios da IoT;
- **OE2:** desenvolver uma solução para recomendação de algoritmos de anonimização que usa formalismo baseado em lógica de descrição (*Description Logic* - DL);
- **OE3:** usar *Machine Learning* para melhorar a precisão do processo de classificação feito pelas ontologias; e
- **OE4:** avaliar o desempenho do *Smart Anonymity* na classificação dos dados de IoT.

1.6 CONTRIBUIÇÕES

Nesta Tese são abordados problemas relacionados a privacidade de dados gerados por dispositivos IoT. A principal contribuição desta Tese foi o desenvolvimento do ***Smart Anonymity*: Um Mecanismo para Recomendação de Algoritmos de Anonimização de Dados Baseado no Perfil dos Dados para Ambientes IoT**. A solução desenvolvida usa o conceito de anonimização de dados para tratar o problema abordado. Para implementar *Smart Anonymity*, foram desenvolvidas duas ontologias para fazer a classificação dos dados e, para melhorar os resultados da classificação, foi usado um modelo de *Machine Learning*.

Para atingir o objetivo desta Tese, o *Smart Anonymity* foi dividido em dois subsistemas independentes, considerados como contribuições secundárias:

1. **Módulo de Classificação:** responsável por fazer a classificação dos dados recebidos pela aplicação. A classificação é feita pela base de conhecimento de duas ontologias. A primeira ontologia, a *SensorCategory*, é responsável por classificar os sensores em categorias. A segunda ontologia, que foi nomeada como *IoTSubdomains*, é responsável por usar a classificação das categorias para identificar a qual subdomínio da IoT a base de dados analisada pertence. Para melhorar a classificação feita pelas ontologias, também é usado *Machine Learning*. Para tanto foi feita uma avaliação com o uso de três algoritmos *K-nearest neighbor* (KNN), *Support vector machine* (SVM) e *Random Forest*. A escolha desses três algoritmos foi devido à sua ampla utilização na literatura, e ao seu reconhecido uso para classificação. Outro fator que levou à escolha destes algoritmos foi por serem de diferentes classes; e

2. **Módulo de Recomendação Personalizada:** recomenda o algoritmo de anonimização mais adequando para determinado conjunto de dados, de acordo com a classificação feita pelo Módulo de Classificação. Basicamente, o Módulo de Recomendação Personalizada recebe um conjunto de dados já classificado, de acordo com o subdomínio em que ele foi classificado e suas características e, através de tabela de decisão, recomenda o algoritmo de anonimização mais adequado para esse conjunto de dados em questão.

Além do desenvolvimento do *Smart Anonymity* e todos os seus componentes, outra contribuição desta Tese foi a definição dos critérios para que os dados sejam classificados. A definição dos critérios foi o primeiro passo que possibilitou o desenvolvimento das ontologias e

suas respectivas bases de conhecimento. Outra importante contribuição desta foi a Revisão Sistemática da Literatura sobre as técnica de anonimozação de dados utilizadas em IoT, descrita no capítulo 3.

1.7 METODOLOGIA

Para condução e execução desta pesquisa, foi escolhido como método principal o *Design Science Research* (DSR) (DRESCH; LACERDA; JÚNIOR, 2015), uma abordagem com duplo objetivo: (1) desenvolver um artefato para resolver um problema prático num contexto específico e (2) gerar novos conhecimentos técnicos e científicos. O conhecimento técnico necessário para a construção de um artefato é diferente do conhecimento científico. O conhecimento técnico não é mais importante que o conhecimento científico, mas é preciso reconhecer que são conhecimentos distintos, ainda que frequentemente confundidos (PIMENTEL; FILIPPO; SANTOS, 2020).

A DSR é o método que fundamenta e operacionaliza a condução da pesquisa quando o objetivo a ser alcançado é um artefato ou uma prescrição. Como método de pesquisa orientado à solução de problemas, a DSR busca, a partir do entendimento do problema, construir e avaliar artefatos que permitam transformar situações, alterando suas condições para estados melhores ou desejáveis. A DSR é utilizada nas pesquisas como forma de diminuir o distanciamento entre teoria e prática (DRESCH; LACERDA; JÚNIOR, 2015).

Hevner et al. (2004) ressaltam que a DSR procura identificar e tratar problemas reais, propondo soluções práticas e apropriadas para solucionar tais problemas por meio da construção e aplicação de um artefato. Esses artefatos podem ser constructos, modelos, métodos ou instâncias de um sistema, construídos e avaliados sob o prisma do problema que se propõe a tratar.

Para condução de uma pesquisa utilizando o método DSR, Hevner et al. (2004) definem sete critérios a serem considerados pelos pesquisadores, que são listados a seguir:

- **Design como Artefato:** as pesquisas desenvolvidas pelo método da DSR devem produzir artefatos viáveis, na forma de um constructo, modelo, método, ou de uma instanciação;
- **Relevância do Problema:** o objetivo da DSR é desenvolver soluções para resolver problemas importantes e relevantes para as organizações;

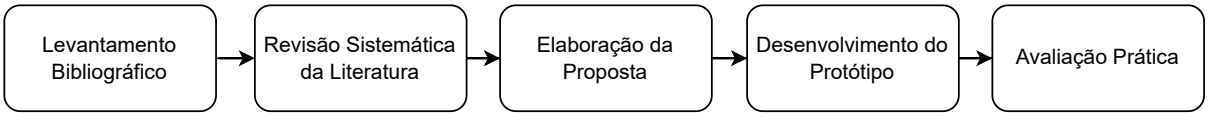
- **Avaliação do Design:** a utilidade, a qualidade e a eficácia do artefato devem ser rigorosamente demonstradas por meio de métodos de avaliação bem executados;
- **Contribuições da Pesquisa:** uma pesquisa conduzida pelo método da DSR deve prover contribuições claras e verificáveis nas áreas específicas dos artefatos desenvolvidos e apresentar fundamentação clara em fundamentos de design e/ou metodologias de design;
- **Rigor da Pesquisa:** a pesquisa deve ser baseada em uma aplicação de métodos rigorosos, tanto na construção como na avaliação dos artefatos;
- ***Design* como Processo de Pesquisa:** a busca por um artefato efetivo exige a utilização de meios que estejam disponíveis para alcançar os fins desejados, ao mesmo tempo que satisfaçam as leis que regem o ambiente em que o problema está sendo estudado; e
- **Comunicação da Pesquisa:** as pesquisas conduzidas pelo método da DSR devem ser apresentadas tanto para o público mais orientado à tecnologia quanto para aquele mais orientado à gestão.

Com base nos argumentos e definições apresentados por Dresch, Lacerda e Júnior (2015) e Hevner et al. (2004), é possível resumir que a DSR permite sanar um problema de pesquisa, sem desconsiderar o rigor científico necessário quando se produz novos conhecimentos, tampouco se desconsidera a aplicabilidade desta solução em uma situação real (JAHN, 2017). Diante disto, esses conceitos vão ao encontro desta Tese.

Além do DSR, mais uma metodologia de apoio foi utilizada para condução dessa pesquisa: uma Revisão Sistemática da Literatura (SLR). Finalmente, a avaliação experimental é conduzida por meio de um *Benchmark* do mecanismo de recomendação proposto. Para medir sua eficácia no processo de classificação e recomendação dos algoritmos de anonimização, as métricas avaliadas foram: A acurácia, precisão *Recall* e *F1-Score*, descritas da seção 5.1.1.

A Figura 1 ilustra as etapas de pesquisa conduzidas nesta pesquisa. A seguir, são descritas cada uma delas.

Figura 1 – Fluxo de Execução da Pesquisa.



Fonte: Autor.

Etapa 1: a primeira etapa da pesquisa teve como objetivo realizar um estudo exploratório *ad hoc* da literatura para compreender os principais problemas relacionados à segurança e a privacidade de dados no contexto de IoT, e as tecnologias mais promissoras para solucionar os problemas de privacidade. Os resultados deste estudo ajudaram a compreender as melhores técnicas usadas para fornecer privacidade de dados em ambientes IoT. Além disso, ajudou na compreensão das limitações das abordagens existentes e na identificação das áreas de melhoria. Os resultados mostraram que não existe uma solução ampla e metódica o suficiente para resolver problemas de privacidade em IoT que possa ser aplicada em vários cenários de aplicação da IoT. Foi identificado também que a anonimização surgiu como uma técnica promissora para tratar os problemas de privacidade dos dados em IoT.

Etapa 2: foi conduzida uma SLR que possibilitou mapear as principais técnicas de anonimização para resolver problemas de privacidade de dados em IoT. Para tal, foi feita uma classificação das técnicas de anonimização listando os pontos fortes e fracos das técnicas selecionadas como sendo as mais promissoras. Para um melhor entendimento desta etapa, no Capítulo 3 foram descritos os procedimentos adotados, bem como seus resultados.

Etapa 3: com as informações coletadas na Revisão Sistemática da Literatura, foi feita uma classificação dos tipos de dados presentes nos subdomínios da IoT. A partir desta classificação, foi possível identificar a qual subdomínio um conjunto de dados pertence para que então seja indicado o algoritmo de anonimização mais adequado. Após a classificação, foi projetado o mecanismo de recomendação usado para analisar as características de um conjunto de dados e, a partir desta análise, identificar a que subdomínio de IoT ele pertence. Após a classificação, o mecanismo de recomendação vai indicar qual o algoritmo de anonimização é o mais adequado para este conjunto de dados de acordo com seu perfil. A arquitetura da proposta é apresentada no Capítulo 4.

Etapa 4: esta etapa foi dedicada ao desenvolvimento do Mecanismo de Recomendação proposto, seguindo os fundamentos propostos na Etapa 3.

Etapa 5: o mecanismo proposto nesta Tese foi colocado em funcionamento, usando diferentes bases de dados, e com características distintas. Foi conduzida uma avaliação experimental, em que foi feita a comparação dos resultados após a execução do mecanismo com os resultados esperados, conforme as características dos dados. Para esta avaliação, foram usadas como métricas: A acurácia, precisão *Recall* e *F1-Score*, o protocolo de avaliação está melhor descrito no Capítulo 5.

1.8 ESTRUTURA DA TESE

O restante da Tese está organizado da seguinte forma:

- **No Capítulo** 2 são descritos os principais conceitos e definições abordados nesta Tese, tais como as definições para IoT, sua composição e áreas de aplicação. São apresentadas também as definições de privacidade de dados e anonimização, bem como são descritas a técnicas de anonimização usadas na Tese;
- **O Capítulo** 3 descreve o estado da arte sobre IoT e privacidade de dados e os trabalhos relacionados a Sistemas de Recomendação, também são apresentadas as principais técnicas de anonimização usadas nos trabalhos analisados.
- **No Capítulo** 4 é apresentada uma descrição da principal contribuição desta Tese, que é o *Smart Anonimity*, juntamente com seus principais componentes. Também é apresentada a estratégia usada para o desenvolvimento da aplicação;
- **O Capítulo** 5 apresenta a descrição da avaliação do mecanismo de recomendação proposto. Neste capítulo também são apresentados os resultados da avaliação; e
- Por fim, no **Capítulo** 6 são apresentadas as contribuições, as ameaças à validade da Tese, direções para trabalhos futuros e as conclusões desta Tese.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são descritos os principais conceitos e definições abordados nesta Tese, a começar pelas definições sugeridas para IoT, as camadas que constituem sua arquitetura, a estrutura da sua composição, as suas áreas de aplicação, a importância da privacidade dos dados para os usuários, o valor e a relevância da anonimização dos dados disseminados pela literatura e, por fim, é apresentado o conceito de ontologia OWL (*Ontology Web Language*) e Lógica de Descrição (DL).

2.1 INTERNET DAS COISAS

É possível encontrar várias definições na literatura para o termo Internet das Coisas IoT, contudo, ainda não existe um consenso a respeito desse conceito. No decorrer deste trabalho, serão apresentadas algumas das definições encontradas na literatura.

A Internet das Coisas (IoT) consiste em dispositivos inteligentes e objetos com sensores/atuadores embutidos. Segundo Ashton et al. (2009), o termo "Internet das Coisas" (IoT) foi mencionado pela primeira vez em 1999, como título de uma apresentação que Ashton fez na empresa *Procter & Gamble* (P & G).

O paradigma IoT prevê um mundo onde tudo está conectado e pode ser monitorado e controlado remotamente. É possível melhorar a eficiência e reduzir os custos, se os dados relevantes forem coletados e analisados. Além disso, controlar remotamente casas pode nos conduzir a uma nova onda de ganhos de eficiência energética, proporcionando um ambiente inteligente, conectando coisas e pessoas (COLLINA et al., 2014; ZHOU; ZHANG, 2014). Para Santos et al. (2016), a Internet das Coisas "nada mais é que uma extensão da internet atual", para proporcionar que objetos do dia a dia (quaisquer que sejam) se conectem à Internet e permite que os próprios objetos sejam acessados como provedores de serviços.

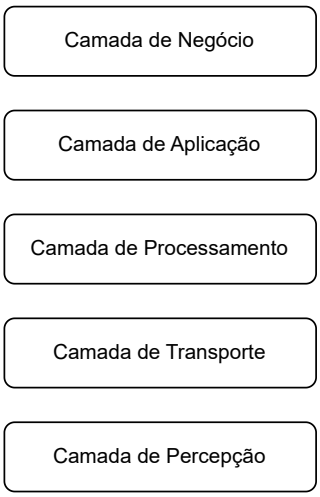
Umas das grandes características e desafios da IoT é a comunicação entre dispositivos heterogêneos (dispositivos com capacidade de processamento diferente), com diferentes formas de conexão (*Wi-Fi, Bluetooth, 4G*), ou seja, a falta de padronização é uma das características mais marcantes da IoT atualmente, bem como a privacidade dos dados gerados pelos dispositivos IoT (AMARAN et al., 2015; FRIGIERI; MAZZER; PARREIRA, 2015; JUNG et al., 2015; LIM et al., 2018; MALEKZADEH et al., 2019).

2.1.1 Arquitetura IoT

A implementação da IoT requer uma arquitetura aberta, baseada em várias camadas para maximizar a interoperabilidade entre sistemas heterogêneos e recursos distribuídos. Na literatura pesquisada, encontra-se uma quantidade variada de trabalhos sobre estudos no tocante às diferentes instâncias da arquitetura da IoT (WU et al., 2010; KHAN et al., 2012; ABOMHARA; KØIEN, 2014; WANG et al., 2016; IEEE, 2019; JAMALI et al., 2020).

Para esta pesquisa, será adotada arquitetura proposta por Jamali et al. (2020). A arquitetura proposta por eles explica melhor os recursos e a conotação da IoT, e essa definição tem sido amplamente aceita dentro da literatura pertinente da área, até o presente momento. Eles dividiram a IoT em cinco camadas, que são: a camada de negócios, a camada de aplicação, a camada de processamento, a camada de transporte e a camada de percepção. Na Figura 2, é possível visualizar essa divisão proposta pelos autores e, a seguir, explicitadas da camada mais inferior à mais superior.

Figura 2 – Divisão da IoT em camadas.



Fonte: Jamali et al. (2020).

- **Camada de Percepção:** a principal tarefa da Camada de Percepção é perceber as propriedades físicas dos objetos (como temperatura, localização, etc.) por meio de vários sensores (como sensores infravermelhos, RFID, código de barras) e converter essas informações em sinais digitais, o que é mais conveniente para transmissão em rede. Os vários sensores e equipamentos na camada de percepção são como o elemento de rede na rede de gerenciamento de telecomunicações;

- **Camada de Transporte:** a Camada de Transporte, ou também chamada de Camada de Rede, é responsável pela transmissão de dados recebidos da Camada de Percepção para o centro de processamento por meio de várias redes, como redes sem fio ou a cabo, ou até mesmo pela rede local da empresa. As principais técnicas nessa camada incluem *FTTx*, *3G*, *Wi-Fi*, *bluetooth*, *Zigbee*, *UMB*, tecnologia infravermelha e assim por diante. Portanto, a principal função da camada de transporte é fazer o transporte dos dados gerados pelos dispositivos;

- **Camada de Processamento:** a Camada de Processamento armazena, analisa e processa principalmente as informações dos objetos recebidos da camada de transporte. As principais técnicas incluem banco de dados, processamento inteligente, computação em nuvem, computação onipresente, etc. A computação em nuvem e a computação onipresente são as principais tecnologias dessa camada;

- **Camada de Aplicação:** a tarefa da Camada de Aplicação é baseada nos dados processados e desenvolve diversas aplicações das Coisas, como transporte inteligente, gerenciamento de logística, autenticação de identidade, *Location Based Service* (LBS), segurança, entre outros. Essa camada fornece todos os tipos de aplicativos para cada setor; e

- **Camada de Negócios:** esta camada é responsável pelo gerenciamento de aplicativos, pelo modelo de negócios relevantes, entre outros negócios. A Camada de Negócios não apenas gerência a liberação e cobrança de vários aplicativos, mas também a pesquisa sobre modelo de negócios e modelo de lucro. Essa camada também permite gerenciar a privacidade dos usuários, que é igualmente importante na IoT.

2.1.2 Composição da IoT

Al-Fuqaha et al. (2015) descrevem a IoT como sendo composta por seis componentes: Identificação; Detecção; Comunicação; Computação; Serviços; e Semântica. A seguir, serão apresentadas a definição e o objetivo de cada um desses componentes:

- **Identificação:** é crucial para a IoT nomear e combinar serviços com sua demanda. Muitos métodos de identificação estão disponíveis para IoT, tais como códigos de produtos eletrônicos (EPC) e códigos *ubíquos (uCode)*. Além disso, o endereçamento dos objetos

IoT é crítico para diferenciar entre o ID do objeto e seu endereço. Object ID refere-se ao seu nome como "T1"para um sensor de temperatura específico e o endereço do objeto se refere ao seu endereço dentro de uma rede de comunicações;

- **Detecção:** significa reunir dados de objetos relacionados dentro da rede e enviá-los para um *data warehouse*, banco de dados ou nuvem. Os dados coletados são analisados para tomar ações específicas com base nos serviços necessários;
- **Comunicação:** as tecnologias de comunicação IoT conectam objetos heterogêneos para fornecer serviços inteligentes específicos. Normalmente, os nós IoT devem operar com baixa potência na presença de links de comunicação com perdas e ruídos;
- **Computação:** as unidades de processamento (por exemplo, microcontroladores, microprocessadores, etc.) e aplicações de software representam o "cérebro"e a capacidade computacional da IoT. Várias plataformas de *hardware* foram desenvolvidas para executar aplicações IoT como, por exemplo, o *Arduino*;
- **Serviços IoT:** de modo geral, os serviços IoT podem ser classificados em quatro classes: serviços relacionados à identidade, serviços de agregação de informações, serviços de colaboração e serviços ubíquos; e
- **Semântica:** na IoT se refere à capacidade de extrair conhecimento inteligentemente por diferentes máquinas para fornecer os serviços necessários. A extração de conhecimento inclui a descoberta e o uso de recursos e informações de modelagem. Além disso, ele inclui o reconhecimento e a análise de dados.

2.1.3 Áreas de Aplicação da IoT

Como dito anteriormente, Borgia (2014) classifica a aplicação da IoT em três domínios, que serão detalhadas a seguir:

- **Domínio Industrial:** a IoT pode ser explorada em todas as atividades industriais que envolvem desde transações comerciais a financeiras. Alguns exemplos são: logística, manufatura, acompanhamento de processos, setor de serviços, bancos, autoridades governamentais financeiras, intermediários, etc. Segundo Borgia (2014), o Domínio Industrial, por sua vez, é dividido nos seguintes subdomínios:

  - Gestão logística e tempo de vida do produto;
  - Agricultura e criação de animais; e
  - Processos industriais.

- **Domínio de Cidades Inteligentes:** a IoT é capaz de auxiliar no aumento da sustentabilidade ambiental das cidades e a qualidade de vida das pessoas, com ênfase na energia e como controlá-la de forma eficiente, além de buscar soluções inteligentes para oferecer serviços que podem melhorar a qualidade de vida das pessoas. Segundo Borgia (2014) o Domínio de Cidades Inteligentes, por sua vez, é dividido nos seguintes subdomínios:
  - Mobilidade inteligente e turismo inteligente;
  - *Smart grid*;
  - Casas/edifícios inteligentes; e
  - Segurança pública e monitoramento ambiental.

- **Domínio de Saúde e bem-estar:** a IoT vai desempenhar um papel essencial no desenvolvimento de serviços inteligentes para apoiar e melhorar a saúde das pessoas e as atividades da sociedade. Estes serviços são direcionados desde os cidadãos até as comunidades, permitindo que as pessoas vivam de forma independente ou para manter suas relações sociais, para melhorar a saúde e assistência social. Segundo (BORGIA, 2014), o Domínio de Saúde e bem-estar, por sua vez, é dividido nos seguintes subdomínios:
  - Medicina e cuidados de saúde; e
  - Vida independente.

Al-Fuqaha et al. (2015), afirmam que "a IoT oferece uma grande oportunidade de mercado para fabricantes de equipamentos, provedores de serviços de Internet e desenvolvedores de aplicativos".

2.2 PRIVACIDADE DE DADOS

A ideia por trás da privacidade é garantir a confidencialidade dos dados relacionados à vida privada de uma pessoa. Dessa maneira, á privacidade garante que os dados pessoais não sejam legíveis, exceto pelo proprietário ou entidades com autorizações explícitas. O processo e o uso

dos dados coletados pelo usuário devem ser regulamentados e, absolutamente, não devem ser vendidos às partes interessadas para fins de *marketing*, publicidade direcionada ou usado para pressionar e/ou chantagear os usuários que tiveram seus dados coletados (BERREHILI; BELMEKKI, 2016). Atualmente existem algumas iniciativas legais para tratar problemas relacionados a privacidade dos dados dos usuários como, por exemplo, a lei para privacidade de dados da União Europeia (*General Data Protection Regulation* – GDPR), ou a Lei Geral de Proteção de Dados Pessoais (LGPD) no Brasil.

A preservação da privacidade é definida como tolerância à prevenção de vazamento de informações causada pela combinação de dados publicados com outros dados. Portanto, os métodos de anonimização devem ser tolerantes na mesclagem com outros dados (NAKAMURA; NISHI, 2018).

Conforme a análise de risco apresentada por Berrehili e Belmekki (2016) em sua pesquisa, as soluções para preservar privacidade devem garantir a confidencialidade dos dados relacionados aos usuários e sua vida privada. Caso o destinatário da informação resultante da análise dos dados tenha acesso aos dados que podem identificar seu proprietário, a privacidade do usuário não pode ser ferida.

2.3 ANONIMIZAÇÃO DE DADOS

Uma das principais técnicas para tratar a privacidade de dados é a anonimização. Ela remove ou substitui as informações que podem ser exploradas por um invasor para comprometer a privacidade de um usuário. Portanto, a anonimização permite que os indivíduos permaneçam ocultos de ameaças em potencial quando seus dados são publicados para fins analíticos ou comerciais. Informações confidenciais ou identificadoras de indivíduos, que não devem ser publicadas ao domínio público, são chamadas de informações confidenciais (LI; PALANISAMY, 2019).

As abordagens de anonimização são classificadas em duas classes principais: a anonimização de dados estáticos e a anonimização de fluxo de dados. A anonimização de dados estáticos trabalha com um conjunto de dados pré-gravados. Já a anonimização de fluxo de dados processa os dados à medida que chegam (OTGONBAYAR et al., 2018). O anonimato apropriado satisfaz um determinado nível de anonimização, causando a modificação mínima dos dados, e isso é desejável para alcançar a preservação da privacidade e o uso secundário eficaz dos dados (NAKAMURA; NISHI, 2018).

Os métodos convencionais de anonimização acrescentam grande ruído aos dados devido à dificuldade de definir o anonimato, o que dificulta o uso prático. Além disso, é verdade que diferentes graus de anonimização são necessários, uma vez que os níveis exigidos de proteção a privacidade variam de acordo com indivíduos ou propósitos. Por exemplo, algumas pessoas não hesitam em revelar sua idade, enquanto outras não gostariam de expor. No entanto, os métodos convencionais não consideram essas diferenças de percepção ou gosto pessoal dos indivíduos (NAKAMURA; NISHI, 2018).

Portanto, o objetivo dos algoritmos de anonimização é encontrar a melhor maneira de ofuscar/esconder os dados, de uma forma que garanta a privacidade dos dados pessoais dos usuários, maximizando a utilidade dos dados, mas ocultado a identidade do seu proprietário (LIM et al., 2018).

2.3.1 Métodos para Anonimização de Dados

Na literatura é possível encontrar vários métodos de anonimização usados para omitir ou até mesmo ocultar dados pessoais. Existem métodos usados de maneira mais ampla, que podem ser aplicados em vários contextos, tanto no cenário IoT como em outras áreas. Também existem métodos que foram desenvolvidos exclusivamente para IoT, e que podem ser usados em vários contextos de aplicação da IoT. Existem ainda outros métodos de anonimização que foram criados para fins mais específicos como, por exemplo, o método proposto por Ullah e Shah (2016), que foi criado para anonimizar a localização dos usuários.

2.3.1.1 Ofuscação de Dados

A ofuscação de dados é um formato de mascaramento que altera dados confidenciais de forma que sejam de pouca ou nenhuma utilidade para intrusos não autorizados, mas podem ser usados por um indivíduo legítimo. Essas informações podem ser confidenciais em termos de comércio, saúde, entre outros. O mascaramento de dados substitui os dados reais por dados plausíveis, mas falsos, para proteger a privacidade (THIRUMALAISAMY et al., 2022).

Para Thirumalaisamy et al. (2022), a substituição é uma das melhores técnicas de Ofuscação que podem ser usadas para manter a aparência real dos registros de dados. A substituição mascara o valor original substituindo os dados por um valor diferente. Vários tipos diferentes de dados podem ser ofuscados usando este método.

A ofuscação é uma técnica usada para proteger a privacidade como, por exemplo, da localização do usuário. Nessa técnica, o local original do usuário é ligeiramente alterado para outro local. Dessa maneira, a localização original da pessoa é ocultada do adversário. Várias técnicas foram propostas para a proteção da privacidade do local como, por exemplo, as técnicas de dispersão, técnicas de randomização, técnica de perturbação, técnica de ofuscação semântica, etc. Mas, nenhuma delas deu atenção ao equilíbrio entre privacidade e utilidade. Algumas delas forneceram bons resultados do ponto de vista da privacidade. O usuário pode obter um bom resultado da perspectiva de privacidade, mas pode não obter bons resultados com a análise dos dados resultantes do processo de ofuscação (ULLAH; SHAH, 2016).

2.3.1.2 Generalização

A Generalização é uma técnica para substituir valores de atributos por uma categoria mais ampla. O método inclui amostragem, recodificação global e recodificação local (SHOHATA; NAKAMURA; NISHI, 2018). Para Murthy et al. (2019), a generalização é um processo de substituição do valor por um valor menos específico, mas semanticamente consistente. Essa técnica aplica-se ao nível da célula, onde alguns valores originais são mantidos com confusão adicional. Isso irá dificultar o trabalho do atacante para inferir dados sensíveis.

Segundo Murthy et al. (2019), a técnica de generalização não pode ser codificada de forma dinâmica. Deve ser codificada com base em atributos apropriados. Os diferentes atributos podem utilizar uma técnica de generalização diferente. Por exemplo, a utilização do consumo pode ser generalizada através da definição de um valor de gama. Entretanto, para o endereço, o número da casa é removido.

2.3.1.3 Perturbação

Para Shohata, Nakamura e Nishi (2018), a perturbação é uma técnica para perturbar os dados usando ruído multiplicativo ou aditivo. Os processos randomizados também são um método popular de perturbação. O método de perturbação pode ser concluído apenas executando uma operação simples nos dados, e possui vantagens em relação à quantidade necessária de recursos.

Essa técnica consiste em substituir os valores originais dos atributos por valores fictícios, dessa forma, as informações estatísticas calculadas a partir dos dados originais não se diferen-

ciem significativamente de informações estatísticas calculadas sobre os dados perturbados. A perturbação é feita através da adição de ruído, que consiste na adição ou multiplicação do valor original por um *offset*, é normalmente usada para valores numéricos.

Na perturbação, diferentes registros são trocados entre si. Nessa abordagem, os valores de um mesmo atributo de dois registros diferentes são permutados. Isso mantém algumas características estatísticas dos dados, como frequência dos atributos e contagem. Essa abordagem, semelhante à adição de ruído, preserva algumas propriedades estatísticas (como contagem e frequência).

2.3.1.4 k-anonymity

O *k-anonymity* é um método típico de preservação da privacidade, considerando o nível de anonimato. No conceito de *k-anonymity* proposto por Sweeney (2002), é dito que "uma liberação de dados possui a propriedade *k-anonymity* se a informação de cada pessoa contida na liberação não puder ser distinguida de pelo menos *k-1* indivíduos, cujas informações também aparecem nesse conjunto de dados".

Para Nayahi e Kavitha (2017), *k-anonymity* é o processo de anonimizar os registros de modo que *k* indivíduos se tornam indistinguíveis um do outro. Esse mecanismo protege o registro da divulgação de identidade ou ataque de vinculação. Assim, Ataque de Vinculação ou Divulgação de Identidade é a possibilidade de um invasor divulgar o valor do atributo sensível de uma pessoa com os valores conhecidos.

O *k-anonymity* é propenso a alguns dos ataques comuns, como ataque de homogeneidade, ataque de similaridade, ataque de conhecimento em segundo plano e ataque de inferência probabilística (NAYAHI; KAVITHA, 2017). O objetivo do algoritmo *k-anonymity* original é garantir que nenhum indivíduo possa ser identificado exclusivamente dentro de um grupo de k pessoas (LIAO et al., 2017).

2.4 ONTOLOGIAS

O termo Ontologia é uma palavra derivada do grego *ontos* (ser) e *logos* (palavra). Os estudos seminais sobre ontologias foram iniciados por Aristóteles, em um contexto filosófico em que ele atribuía esse termo a um ramo da Metafísica para classificar as coisas, ou seja, como uma ciência para descrever "o ser", isto é, o estudo dos atributos que pertencem às coisas devido à sua própria natureza (GUARINO; OBERLE; STAAB, 2009).

Para Borst (2006): “Uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada”. Já Almeida e Bax (2003) afirmam que essa definição, “formal” significa legível para computadores; “especificação explícita” diz respeito a conceitos, propriedades, relações, funções, restrições, axiomas explicitamente definidos; “compartilhado” quer dizer conhecimento consensual; e, “conceitualização” diz respeito a um modelo abstrato de algum fenômeno do mundo real.

Na ciência da computação, o interesse pelo uso de ontologias surgiu a partir do momento em que os engenheiros do conhecimento perceberam que não era suficiente a preocupação apenas com mecanismos de representação do conhecimento (regras, *frames*, redes neurais, lógica *fuzzy*, etc.) se não existisse um bom conteúdo e organização sobre o conhecimento do domínio em que se deseja trabalhar (CHANDRASEKARAN; JOSEPHSON; BENJAMINS, 1999; OLIVEIRA; WERNECK, 2003).

De acordo com Costa (2020), as ontologias são tipicamente especificadas em linguagens que permitem a abstração de estruturas de dados e estratégias de implementação. As linguagens das ontologias estão mais próximas à lógica de primeira ordem do que as linguagens usadas para modelar bancos de dados. Costa (2020) complementa afirmando que as ontologias são consideradas no nível "semântico", enquanto os esquemas de banco de dados são modelos de dados no nível "lógico"ou "físico".

Na área da Inteligência Artificial, uma ontologia O pode ser definida como um relacionamento de quatro elementos, representado por $O = (C, R, I, A)$, onde (KIRYAKOV; OGNYANOV; MANOV, 2005):

- C - é o conjunto de classes que representam os conceitos em um dado domínio de interesse;
- R - é o conjunto de relações ou associações entre os conceitos do domínio;
- I - é o conjunto de instâncias derivadas das classes, ou ainda, os exemplos de conceitos representados em uma ontologia; e
- A - é o conjunto de axiomas do domínio, que servem para modelar restrições e regras inerentes às instâncias e conceitos.

Ontologias são usadas para capturar conhecimento sobre algum domínio de interesse. Uma ontologia descreve os conceitos no domínio e também os relacionamentos que existem entre esses conceitos. Diferentes linguagens de ontologia fornecem diferentes facilidades. O

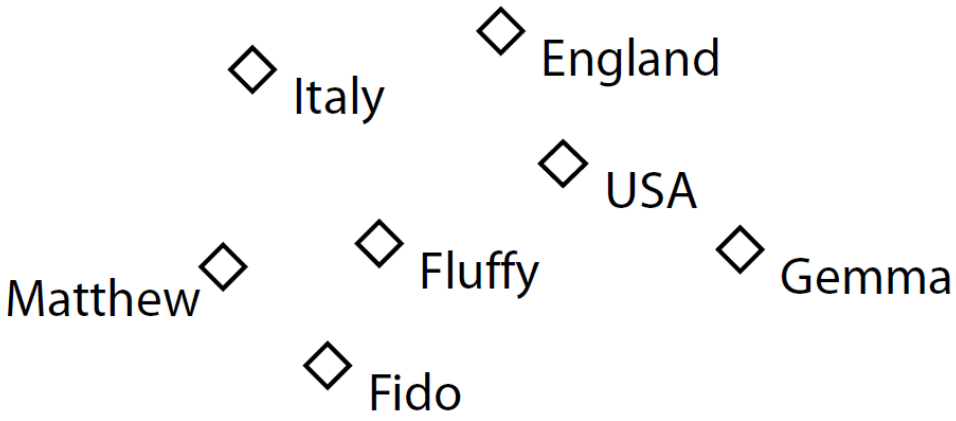
desenvolvimento mais recente em linguagens de ontologia padrão é o OWL do *World Wide Web Consortium* (W3C).

O OWL possibilita a descrição de conceitos, mas também oferece novas facilidades. Ele tem um conjunto mais rico de operadores - por exemplo, intersecção, união e negação. Baseia-se num modelo lógico diferente, que permite definir e descrever conceitos. Portanto, conceitos complexos podem ser construídos em definições a partir de conceitos mais simples. Além disso, o modelo lógico permite o uso de um raciocinador que pode verificar se todas as declarações e definições na ontologia são mutuamente consistentes e também pode reconhecer quais conceitos sob quais definições. O raciocinador pode, portanto, ajudar a manter a hierarquia corretamente. Isso é particularmente útil ao lidar com casos em que as classes podem ter mais de um pai (HORRIDGE et al., 2009).

Com OWL, além de se poder expressar a estrutura dos conceitos e relacionamentos, é possível descrever características especiais sobre os conceitos e os relacionamentos por meio de axiomas lógicos, essa estrutura, assim definida, constitui uma ontologia (VIEIRA et al., 2005).

Em ontologias OWL, indivíduos representam objetos no domínio de interesse. Dois nomes diferentes podem realmente referir-se à mesma pessoa. Por exemplo, "Rainha Elizabeth", "A Rainha" e "Elizabeth Windsor" podem se referir ao mesmo indivíduo. Em OWL, deve ser explicitamente declarado que os indivíduos são iguais entre si ou diferentes entre si. A Figura 3 mostra uma representação de alguns indivíduos em algum domínio (HORRIDGE et al., 2009).

Figura 3 – Representação de indivíduos.

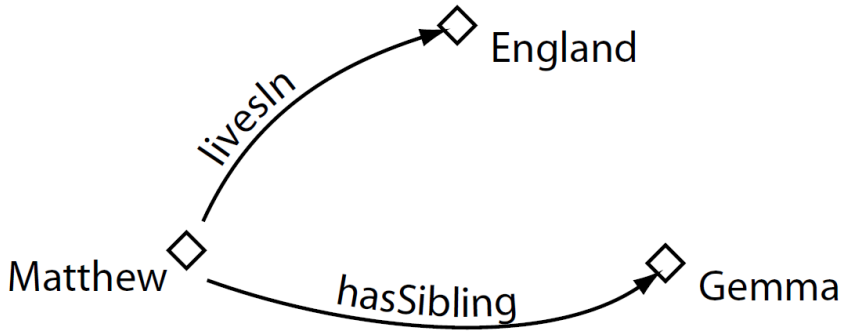


Fonte: Horridge et al. (2009).

As Propriedades são relações binárias em indivíduos, ou seja, propriedades ligam dois

indivíduos. Por exemplo, a propriedade *hasSibling* pode vincular o indivíduo Matthew a cada Gemma, ou a propriedade *hasChild* pode vincular o indivíduo Peter ao indivíduo Matthew. As propriedades podem ter inversos. Por exemplo, o inverso de *hasOwner* é *isOwnedBy*. As propriedades podem ser limitadas a ter um único valor, ou seja, ser funcionais. A Figura 4 mostra uma representação de algumas propriedades ligando alguns indivíduos (HORRIDGE et al., 2009).

Figura 4 – Representação de propriedades.



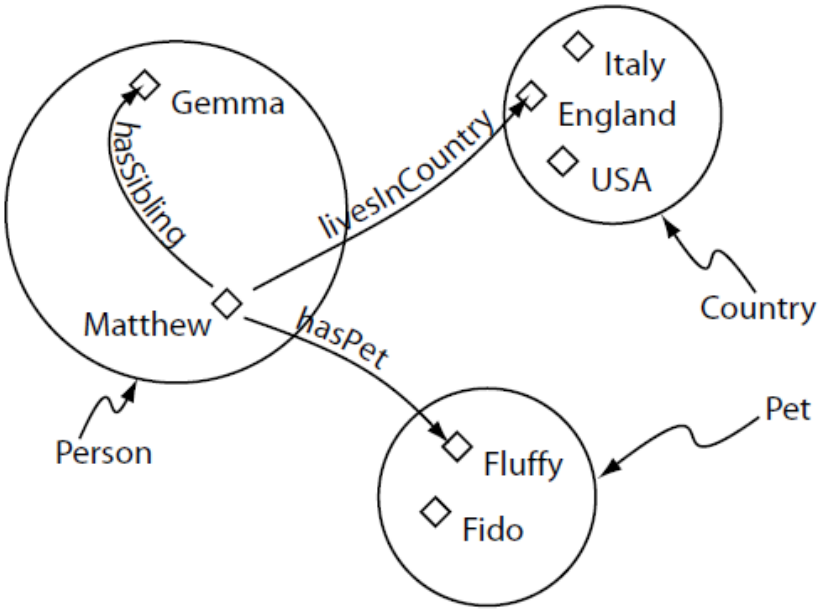
Fonte: Horridge et al. (2009).

As classes em ontologias OWL são interpretadas como conjuntos que contêm indivíduos. Eles são descritos usando descrições formais (matemáticas) que definem precisamente os requisitos para a associação à classe. Por exemplo, uma classe Gato conteria todos os indivíduos que são gatos no domínio de interesse. As classes podem ser organizadas em uma hierarquia superclasse-subclasse, que também é conhecida como taxonomia. As subclasses se especializam, são incluídas por suas superclasses. Por exemplo, considere as classes Animal e Gato (Gato pode ser uma subclasse de Animal, então Animal é a superclasse de Gato). Isso diz que todos os gatos são animais, todos os membros da classe Gato são membros da classe Animal. Ser um Gato implica que você é um Animal, e Gato é classificado como Animal. A Figura 5 mostra uma representação de algumas classes contendo indivíduos (HORRIDGE et al., 2009).

2.4.1 Lógica de Descrição

Lógica de Descrições (DL) refere-se a uma família de linguagens formais de representação do conhecimento. A DL é usada na Inteligência Artificial para descrever e raciocinar sobre

Figura 5 – Representação de classes.



Fonte: Horridge et al. (2009).

conceitos relevantes de um domínio de aplicação (COSTA, 2020).

Em *Web* Semântica, DL provê o formalismo lógico para linguagem OWL, permitindo que sejam criados axiomas lógicos para ontologias. O uso de DL em ontologias OWL permite que se tenham alguns perfis em OWL conforme a sua expressividade em DL (COSTA, 2020).

Costa (2020) afirma ainda que uma das principais características da DL é que ela suporta padrões de inferência que ocorrem em muitas aplicações de sistemas de processamento de informação inteligentes, e que também podem ser usados pelos seres humanos para classificação de conceitos e indivíduos.

Vieira et al. (2005) definem as DLs como uma evolução de representação do conhecimento baseado em objetos (redes semânticas e *frames*), ao qual corresponde um subconjunto estruturado da lógica de primeira ordem. Em termos gerais, as lógicas de descrições são formalismos para representar conhecimento e raciocinar sobre ele. Vieira et al. (2005) complementam ainda que a sintaxe deste formalismo foi definida para facilitar o raciocínio, tornando-o computacionalmente menos custoso. A OWL é baseada na lógica de descrição, havendo uma correspondência entre a linguagem *Markup Language* (XML) para expressar ontologias e uma lógica de descrição.

A sintaxe das DLs é formada por símbolos representando conceitos e papéis, construtores, e quantificadores. Os conceitos são representações de classes, conjuntos de indivíduos que

apresentam as mesmas características gerais. Podem ser conceitos-base (conceitos primitivos), que não dependem de outros conceitos ou relacionamento para serem definidos ou conceitos complexos, formados a partir da utilização de outros conceitos previamente declarados. Os construtores são operadores que permitem a criação de conceitos complexos, dando significado especial à interpretação de conceito. Os papéis são propriedades dos conceitos. Eles representam relacionamentos entre os elementos da base do conhecimento (conceitos e instâncias). Os quantificadores são operadores que quantificam os papéis (VIEIRA et al., 2005). Na Tabla 1 é apresentada a sintaxe de DL.

Tabela 1 – Sintaxe e Famílias de DL.

Descrição	Sintaxe	Linguagem
conceito	A	$\mathcal{FL}$
papel	R	
conjunção ou intersecção de conceitos	$C \sqcap U$	
quantificador universal	$\forall R.C$	
quantificador existencial	$\exists R$	
conceito superior	$\top$	$\mathcal{AL}$*
conceito inferior	$\bot$	
negação	$\neg A \neg C$	
disjunção ou união de conceitos	$A \sqcup D$	
restrição existencial	$\exists R.C$	
restrição numérica	$(\geq n\ R)(\leq n\ R)$	
coleção de indivíduos	$\{a_1 ... a_n\}$	
hierarquia de papéis	$R \sqsubseteq S$	$\mathcal{H}$
inversão de papéis	R^-	$\mathcal{I}$
restrição numérica qualificada	$(\geq n\ R.C)(\leq n\ R.C)$	$\mathcal{Q}$

Fonte: Adaptado de Costa (2020).

Baader et al. (2003) definem a DL como o nome mais recente para uma família de formalismos de representação do conhecimento, *Knowledge Representation* (KR) que representam o conhecimento de um domínio de aplicação (o "mundo"). Primeiro, os conceitos relevantes do domínio são definidos e, em seguida, esses conceitos são usados para especificar propriedades de objetos e indivíduos que ocorrem no domínio. Uma das características dessas linguagens é que, ao contrário de algumas de suas predecessoras, elas são equipadas com uma semântica formal baseada em lógica. Outra característica distinta é a ênfase no raciocínio como um serviço central: o raciocínio permite inferir o conhecimento representado implicitamente a partir do conhecimento explicitamente contido na base de conhecimento.

Para Baader et al. (2003), as principais tarefas de Raciocínio em DL são satisfabilidade, subsunção, equivalência e disjunção. Neste contexto, DL fornece inferência decidível, sólida e completa para algumas tarefas de raciocínio, que podem ser definidas formalmente como:

- **Satisfabilidade**: um conceito C é *satisfatível* em relação a $\mathcal{T}$ se existir um modelo de $\mathcal{I}$ para $\mathcal{T}$ tal que $C^{\mathcal{I}}$ não seja vazia. Nesse caso, é dito que $\mathcal{I}$ é um modelo de C;
- **Subsunção**: um conceito C é *subconceito* de D em relação a $\mathcal{T}$ se $C^{\mathcal{I}} \sqsubseteq D^{\mathcal{I}}$ para cada modelo $\mathcal{I}$ de $\mathcal{T}$. Neste caso, é escrito $C \sqsubseteq_{\mathcal{T}} D$ ou $\mathcal{T} \models C \sqsubseteq D$;
- **Equivalência**: dois conceitos C e D são *equivalentes* em relação a $\mathcal{T}$ se $C^{\mathcal{I}} = D^{\mathcal{I}}$ para cada modelo $\mathcal{I}$ de $\mathcal{T}$. Neste caso, escrevemos $C \equiv_{\mathcal{T}} D$ ou $\mathcal{T} \models C \equiv D$; e
- **Disjunção**: dois conceitos C e D são *disjuntos* em relação a $\mathcal{T}$ se $C^{\mathcal{I}} \sqcap D^{\mathcal{I}} = \emptyset$ para cada modelo $\mathcal{I}$ de $\mathcal{T}$.

Para Harmelen, Lifschitz e Porter (2008), DLs são uma família de linguagens de representação de conhecimento que podem ser usadas para representar o conhecimento de um domínio de aplicação de uma forma estruturada e formalmente bem compreendida. A lógica de descrição de nomes é motivada pelo fato de que as noções importantes do domínio são descritas por descrições de conceitos. Essas expressões são construídas a partir de conceitos atômicos (predicados unários) e papéis atômicos (predicados binários), usando os construtores de conceito e função fornecidos pela DL particular. Por outro lado, DLs diferem de seus predecessores, como redes semânticas e *frames*, por serem equipados com uma semântica formal baseada em lógica.

2.4.2 Software Protégé

O Protégé é um ambiente para criação e edição de ontologias e bases de conhecimento desenvolvido pela University of Stanford, que permite ao usuário iniciar o projeto de uma ontologia rápida e intuitivamente (Protégé , 2023). De acordo com (COSTA, 2020), o Protegé é uma das ferramentas mais utilizadas para modelagem de ontologias.

O Protégé Desktop suporta a criação e edição de uma ou mais ontologias em um único espaço de trabalho por meio de uma interface de usuário totalmente personalizável. As ferramentas de visualização permitem a navegação interativa dos relacionamentos da ontologia.

O suporte a explicações avançadas auxilia no rastreamento de inconsistências. O Protégé permite operações de refatoração, incluindo fusão de ontologias, movimentação de axiomas entre ontologias, renomeação de várias entidades e muito mais (Protégé , 2023).

2.5 *MACHINE LEARNING*

O princípio fundamental do *Machine Learning* é que uma máquina (computador) pode aprender automaticamente com base em treinamento. Embora os aplicativos de *Machine Learning* sejam diferentes, sua função geral é semelhante a de um aplicativo para outro (OZDEMIR et al., 2021). Já para (CHIBANI; COUDERT, 2020), o *Machine Learning* visa desenvolver algoritmos que podem aprender e criar modelos estatísticos para análise e previsão de dados. Os algoritmos de ML devem ser capazes de aprender sozinhos com base nos dados fornecidos e fazer previsões precisas, sem terem sido programados especificamente para uma determinada tarefa. Nesta Tese, foram implementados três algoritmos. São eles: o KNN, o SVM e o *Random Forest*.

2.5.1 K-nearest Neighbor (KNN)

O K-nearest neighbor (KNN) é um algoritmo que mantém todas as instâncias e classifica novas instâncias dependendo da medida de semelhança do algoritmo KNN. É um dos algoritmos baseados na distância mais frequentemente utilizados, que se baseia na distância como único critério de classificação. Outro benefício é que o KNN não precisa de conhecimentos e de preservar um determinado modelo. Portanto, o modelo pode ajustar-se a mudanças rápidas (ATALLAH et al., 2019).

No algoritmo KNN, as instâncias de formação do conjunto de dados são extraídas como pontos de dados no espaço de características e divididas em várias classes separadas. Para prever a classe de um novo ponto de instância Pt, primeiro, é estimado no espaço de características proposto. Depois. são calculados os espaços entre Pt e o Kth exemplos mais próximos. Finalmente, um voto maioritário dos seus vizinhos classifica Pt (ATALLAH et al., 2019).

KNN é um algoritmo de aprendizagem supervisionada por máquinas e utiliza um método não paramétrico. Na análise de regressão, a produção é calculada utilizando a média dos valores dos k vizinhos mais próximos. É um dos exemplos mais simples de técnica de aprendizagem mecânica, também chamada aprendizagem preguiçosa, em que a função é estimada localmente.

Assim, o KNN é sensível à estrutura local de dados. A distância Euclidiana, métrica comumente utilizada, é utilizada para variáveis contínuas. O melhor valor de k depende dos dados, mas o bom valor de k pode ser selecionado por várias técnicas heurísticas. O caso especial é chamado algoritmo do vizinho mais próximo, em que a etiqueta ou classe prevista é a mesma que a etiqueta ou classe da amostra de treino mais próxima (isto é, quando k = 1). A precisão do algoritmo k-NN pode ser gravemente degradada na presença de características ruidosas ou irrelevantes (GOEL et al., 2020).

2.5.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) é um modelo matemático e um aproximador universal eficaz, que pode ser utilizado para resolver problemas de classificação e regressão. SVM baseia-se principalmente no conceito de minimização de risco estrutural, em oposição ao conceito de minimização de erro empírico em Redes Neurais. A SVM executa eficientemente a classificação linear e a não linear. Isto é feito usando projeção (mapeamento não linear usando o truque do núcleo) do conjunto de dados de treino para um espaço dimensional superior no qual um determinado hiperplano (vector de suporte) separa as categorias dos dados de treino (BAJAJ et al., 2023).

O SVM é um dos cálculos de Aprendizagem Supervisionada mais utilizados, utilizado para classificação apenas como questões de regressão. No entanto, é utilizado para questões de Classificação na Aprendizagem Automática. O objetivo do cálculo SVM é fazer a melhor linha ou limite de escolha que possa isolar o espaço n-dimensional em classes para que pesquisadores possam, sem grande esforço, colocar mais tarde o novo ponto de informação na classificação correta. Ese melhor limite de escolha é conhecido como um hiperplano. SVM escolhe os focos/vetores ultrajantes que ajudam a fazer o hiperplano. Esses casos ultrajantes são chamados Vetores de Suporte e, consequentemente, o cálculo é denominado Máquina de Vetores de Suporte (KURANI et al., 2023).

2.5.3 Random Forest Classifier (RF)

Random forest classifier é um classificador baseado em árvores de decisão com gerações de árvores de decisão aleatória (ONG; ZULVIA; PRASETYO, 2023). *Random forest* é utilizado para classificação bem como para problema de regressão. É um algoritmo de aprendizagem em

conjunto. O *Random foresta* é adequado para problemas em que a formação é feita na árvore de decisão como dados, e a produção é obtida sob a forma de previsão média. (GOEL et al., 2020).

Random forest utiliza a técnica de ensacamento (*Bootstrap Aggregation*) que utiliza árvores de decisão múltipla. Cada árvore de decisão é treinada numa amostra de dados diferente, onde a amostragem dos dados é feita com substituição. Assim, em vez de dependendo de árvores de decisão individuais, a produção final é determinada pela combinação de múltiplas árvores de decisão (GOEL et al., 2020).

A principal vantagem de utilizar *Random Forest* é que, em vez de procurar a característica mais vital enquanto divide um nó, ele procura a melhor característica entre um subconjunto aleatório de características. Devido a essa característica, dá bons resultados em caso de extração de dados (GOEL et al., 2020).

2.6 SOFTWARES USADOS PARA ANONIMIZAÇÃO DE DADOS

Existem várias ferramentas que podem ser usadas para anonimizar dados. Em sua pesquisa Prasser et al. (2020) listam algumas ferramentas de código aberto:

- **Amnesia:** é um *software* disponibilizado gratuitamente para a comunidade. Suporta supressão e pseudonimização (sendo os campos simplesmente substituídos por caracteres aleatórios ou por símbolos, mascarando-os). Funciona com o modelo de privacidade k-anonimidade. O objetivo da ferramenta é deixar os dados k-anonimizados, para um determinado k especificado pelo utilizador. Não tem suporte para correr diretamente sobre bases de dados relacionais, apenas permitindo o carregamento de ficheiros com os dados (RAMOS, 2022);

- **Argus:** o nome ARGUS é um acrônimo para *AntiRe-Identification General Utility System*. Tem o objetivo de tornar os microdados seguros. Para a anonimização usa o modelo de privacidade k-anonimato na maioria das etapas, e também é possível aplicar transformações adicionais como, por exemplo, supressão local, agrupamento de categorias, adição de ruído e dados sintéticos (STENERSEN, 2020);

- **ARX:** é um *software* open source para tornar anônimos dados pessoais sensíveis. Suporta uma grande variedade de (1) modelos de privacidade e risco, (2) métodos para transformar

dados e (3) métodos para analisar a utilidade dos dados de saída. O ARX é capaz de lidar com grandes conjuntos de dados em hardware de base e possui uma interface gráfica de utilizador intuitiva e multiplataforma (ARX, 2023);

- **Cornell Anonymization Toolkit:** é uma ferramenta gratuita que permite anonimização de dados. A ferramenta suporta o algoritmo *Incognito* e os modelos de privacidade l-diversidade e t-proximidade (MAIER, 2013);

- **sdcMicro:** o *sdcMicro* foi criado para auxiliar pesquisas na geração de microdados para uso público. No *sdcMicro*, são utilizados dois modelos de privacidade, k-anonimato e l-diversidade, também são utilizados métodos para transformação de dados, tais como, randomização, *top and bottom coding*, supressão e *recoding* (ZUO et al., 2021);

- **SECRETA:** o *System for Evaluating and Comparing RElational and Transaction Anonymization algorithms* (SECRETA) tem como objetivo a análise da eficácia e da eficiência de algoritmos de anonimização para dados tabulares e com valor definido (POULIS et al., 2014);

- **TIAMAT:** a *Tool for Interactive Analysis of Microdata Anonymization Techniques* (TIAMAT) possibilita o uso de algoritmos de anonimização como, por exemplo, *Mondrian e k-Member*, bem como modelos de privacidade k-anonimato, l-diversidade e t-proximidade (DAI et al., 2009); e

- **UTD Anonymization Toolbox:** implementa seis métodos de anonimização para uso público por pesquisadores. Os algoritmos podem ser aplicados diretamente a um conjunto de dados ou podem ser usados como funções de biblioteca dentro de outros aplicativos. Os métodos implementados são: (i) *Datafly*, (ii) *Mondrian Multidimensional k-Anonymity*, (iii) *Incognito*, (vi) *Incognito* com l-diversidade, (v) *Incognito* com t-proximidade e (vi) *Anatomy* (ToolBox, 2023).

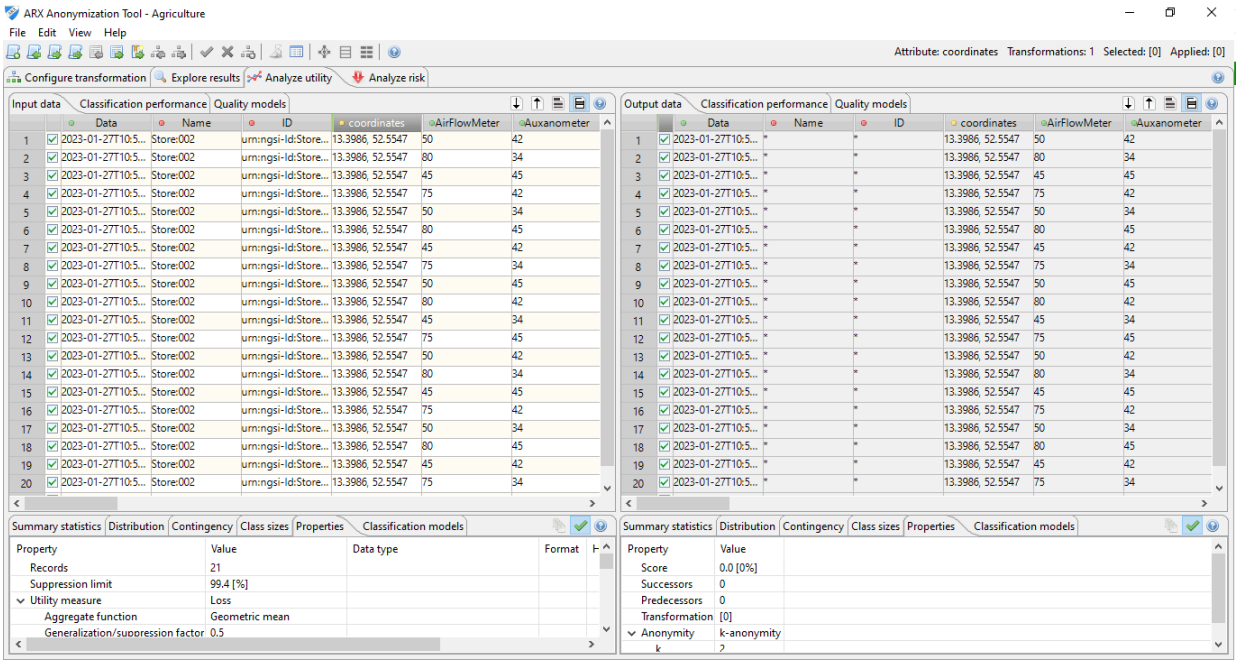
Para Prasser et al. (2020), a ferramenta de anonimização de dados ARX fornece um software aberto que alcance um alto grau de automação e, ao mesmo tempo, fornecer suporte a uma ampla gama de técnicas.

2.6.1 Software ARX

O ARX está dividido em quatro perspectivas, que modelam diferentes aspectos do processo de anonimização. Como se mostra abaixo na Figura 6, essas perspectivas suportam 1) a configuração de modelos de privacidade; medidas de utilidade e métodos de transformação; 2) a exploração do espaço de soluções; 3) a análise da utilidade dos dados; e 4) a análise dos riscos de privacidade (ARX, 2023).

O ARX exibe os dados de entrada e saída em tabelas com cabeçalhos que indicam os tipos de atributos por cores diferentes. A seguir, a Figura 6 ilustra a tela com dados sendo anonimizados. Do lado esquerdo, é mostrado uma tabela com os dados recebidos como entrada no ARX. Nesse momento, os dados podem ser classificados por cores conforme as características dos dados. Já do lado direito, são mostrados os dados após o processo de anonimização. Nesse caso, duas colunas serão removidas porque existem dados sensíveis e podem colocar a privacidade dos proprietários dos dados em risco. Esses dados podem ser exportados em um arquivo 'CSV'.

Figura 6 – Visualização da tela inicial da ferramenta ARX versão 3.9.1.



Fonte: ARX (2023).

Na ferramenta ARX, os dados analisados podem ser classificados como:

- Os atributos de identificação serão removidos do conjunto de dados;
- Atributos quase identificadores serão transformados;

- Atributos confidenciais serão mantidos como estão, mas podem ser protegidos usando modelos de privacidade, como t-proximidade ou l-diversidade; e
- Atributos insensíveis serão mantidos inalterados.

Onde as cores representam a classificação dos tipos de dados por cor:

- **Vermelho:** indica um atributo identificador;
- **Amarelo:** indica um atributo quase identificador;
- **Roxo:** indica um atributo sensível; e
- **Verde:** indica um atributo insensível.

A ferramenta ARX suporta vários tipos de dados. A seguir são listados os tipos suportados:

- ***String*:** uma sequência genérica de caracteres. Este é o tipo de dados padrão;
- **Inteiro:** um tipo de dados para números sem um componente fracionário;
- **Decimal:** um tipo de dados para números com componente fracionário;
- **Data/hora:** um tipo de dados para datas (com ou sem hora); e
- **Ordinal:** variáveis de string com uma escala ordinal.

2.7 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentados os principais conceitos utilizados nesta Tese. Inicialmente foram apresentados as definições de IoT, bem como sua divisão e aplicabilidade. Posteriormente foram apresentados os conceitos sobre privacidade de dados e anonimização, também foram descritos os principais algoritmos de anonimização. Foram apresentados os conceitos e definições sobre ontologias, bem como sua importância e funcionamento. São descritas ainda as principais definições sobre lógica de descrição (DL). Também foram descritos conceitos sobre *Machine Learning* e o *software* ARX e, outras ferramentas semelhantes, usadas para anonimização de dados.

3 REVISÃO SISTEMÁTICA DA LITERATURA E TRABALHOS RELACIONADOS

Este capítulo apresenta os trabalhos relacionados com a descrição do problema desta Tese. Nesse contexto, são apresentados os procedimentos adotados para a seleção de fontes de pesquisa para os estudos. Vale ressaltar que, para realizar este estudo, foram seguidas as Diretrizes propostas por Kitchenham e Charters (2007).

3.1 QUESTÕES DE PESQUISA

Especificar as questões de pesquisa é parte crucial do planejamento de uma Revisão Sistemática da Literatura (*Systematic Literature Review* - SLR), uma vez que as questões direcionam o processo de revisão e fornecem base para decidir: (i) quais estudos primários serão incluídos na revisão (direcionando a estratégia de busca); (ii) quais dados serão extraídos; e (iii) como os dados serão sintetizados para responder às questões de pesquisa (KITCHENHAM; CHARTERS, 2007; KITCHENHAM; BUDGEN; BRERETON, 2015).

Para alcançar o objetivo elencado neste estudo, as questões descritas a seguir devem ser respondidas em modo de desenvolvimento de soluções usando anonimização de dados para privacidade de dados em IoT. Diante disso, para condução desta pesquisa foi definida a seguinte questão de pesquisa: Quais são as principais abordagens que propõem o fornecimento de privacidade usando anonimização existentes na área de IoT? A partir deste questionamento, derivaram-se as seguintes subquestões:

- **RQ1:** Quantos estudos com foco da pesquisa sobre anonimização de dados para IoT foram publicados entre 2009 e 2021?
- **RQ2:** Quais indivíduos e organizações são mais ativos em pesquisa na área de anonimização de dados para IoT?
- **RQ3:** Quais dados não podem ser divulgados, pois o anonimato do indivíduo pode ser ferido?
- **RQ4:** Quais dados podem ser divulgados sem prejuízo para o anonimato do indivíduo?
- **RQ5:** Quais são as categorias relacionadas às métricas, técnicas, métodos, ferramentas, tecnologias aplicadas à anonimização de dados para IoT?

- **RQ6:** Quais são os desafios em aberto sobre anonimização de dados para IoT?

Estas perguntas buscam esclarecimentos sobre como as técnicas de anonimização de dados são usadas para tratar a privacidade de dados no contexto de IoT. Para respondê-las, é necessário fazer uma classificação desses trabalhos. Para isso, será usado um subconjunto de características para classificação dos estudos. Essas características incluem: (i) *design* da técnica/abordagem; (ii) área de aplicação; (iii) métodos usados; e (iv) métricas usadas na avaliação da anonimização.

Ao final desta SLR, foi feita uma classificação das diferenças entre esses estudos com base no tipo de ambiente alvo, isto é, onde essas técnicas aplicadas a anonimização de dados para IoT são usadas ou para que foram criadas e apontar as principais vantagens e desvantagens no cenário abordado.

3.2 BUSCA DOS ESTUDOS

A estratégia para condução desta revisão é usar uma combinação de busca manual com busca automática. A busca e seleção dos estudos foi definida com base nas questões de pesquisa. A organização da *String* usada para a busca automática se refere aos termos e sinônimos respectivamente usados para encontrar os trabalhos disponíveis nas bases de trabalhos científicos.

Inicialmente, a *String* de busca foi elaborada com base nos resultados de uma busca manual. Para obter uma *String* de busca final que atendesse aos objetivos desta pesquisa, foram executados diversos testes de maneira interativa e incremental. Após várias tentativas, foi definida a versão final da *String*:

(“Data Anonymization”) AND (“IoT"OR “Internet Of Things”)

Para as buscas automáticas, foram utilizadas bases reconhecidamente utilizadas na literatura. Ao todo foram usadas 7 bases, todas disponíveis via internet. A seguir, na Tabela 2 é apresentada a lista dos engenhos de busca utilizados e seus respectivos endereços eletrônicos.

A Tabela 3 mostra as sequências de pesquisa por mecanismo de busca da *Web*, suas respectivas *Strings* de busca usadas e a quantidade de trabalhos retornados por cada uma delas. Uma informação a ser destacada e que se fez necessária é uma adaptação da *String* para a sintaxe de alguns engenhos de busca. É importante salientar que, inicialmente, foi aplicado

Tabela 2 – Bases usadas para pesquisa automática.

Mecanismo de Pesquisa na *Web*	***Link* de Acesso**
ACM Digital Library	https://dl.acm.org/
Engineering Village	https://www.engineeringvillage.com
IEEEXplore	http://ieeexplore.ieee.org/
Science Direct	https://www.sciencedirect.com
Scopus	http://www.scopus.com
SpringerLink	https://link.springer.com/
Web of Science	https://apps.webofknowledge.com

Fonte: O Autor.

um filtro para artigos escritos entre 2009 a 2019, porém, posteriormente, a busca foi expandida até 2021. O engenho de busca que retornou o maior número de trabalhos foi a *SpringerLink* com 295 trabalhos. Ao final das buscas, foram retornados um total de 523 trabalhos.

Tabela 3 – Lista de sequências de pesquisa por mecanismo de busca.

Mecanismo de busca	*String*	Qtd
ACM Digital Library	("*Data Anonymization*") AND ("IoT"OR "*Internet Of Things*")	14
Engineering Village	("*Data Anonymization*") AND ("IoT"OR "*Internet Of Things*")	23
IEEEXplore	("*Data Anonymization*") AND ("IoT"OR "*Internet Of Things*")	36
Science Direct	("*Data Anonymization*") *AND* ("IoT"OR "*Internet Of Things*")	119
Scopus	("*Data Anonymization*") AND ("IoT"OR "*Internet Of Things*")	22
SpringerLink	("*Data Anonymization*") AND ("IoT"OR "*Internet Of Things*")	295
Web of Science	("*Data Anonymization*") AND ("IoT"OR "*Internet Of Things*")	14
Total de trabalhos retornados		523

Fonte: O Autor.

3.3 SELEÇÃO DOS ESTUDOS

Para auxiliar no processo de seleção de estudos considerados relevantes para esta pesquisa, foram estabelecidos alguns critérios de inclusão e exclusão. Para que um estudo seja incluído, ele

deve satisfazer pelo menos um dos critérios de inclusão estabelecidos a seguir, por outro lado, para que um determinado estudo seja eliminado, este mesmo estudo deverá estar relacionado a pelo menos um dos critérios de exclusão.

Critério de Inclusão:

- Estudos que apresentam conceitos, teorias, métodos, ferramentas, modelos de referência, guidelines, lições aprendidas e relatos de experiência sobre a aplicação de anonimização de dados em IoT.

Critérios de Exclusão:

- Editoriais, resumos, tutoriais, *keynote*, relatórios de *workshop*, opiniões, sumários de conferências, teses, dissertações, relatórios técnicos, livros;
- Estudos secundários ou terciários;
- Estudos primários que não apresentam de forma explícita a aplicação de anonimização de dados;
- Artigos que não estejam acessíveis, via *Web*;
- Duplicidade de publicação (Indexados);
- Artigos não escritos em língua inglesa; e
- Artigos que não estejam associados às áreas de Ciência da Computação e Engenharia.

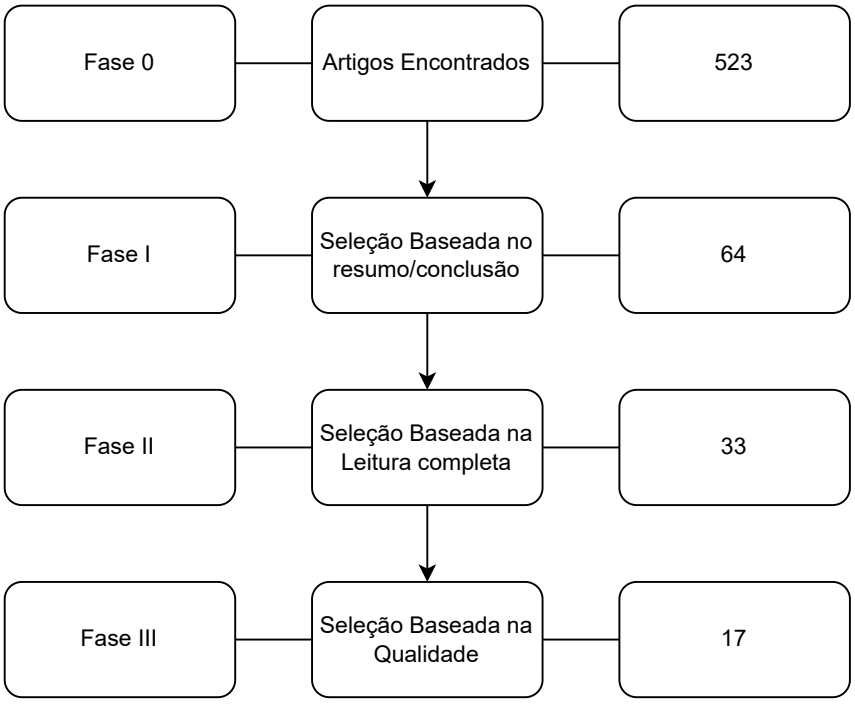
A condução da SLR foi dividida em quatro fases, descritas em detalhes em seguida. Estas fases foram executadas para a seleção dos trabalhos considerados relevantes para esta pesquisa.

- **Fase 0:** conforme explicado anteriormente, o quantitativo de artigos encontrados obteve o montante inicial de 523 artigos, que foram obtidos usando os critérios estabelecidos na seção 3.2;
- **Fase I:** uma busca manual foi executada. Nela foram aplicados os critérios de inclusão e exclusão alinhados com a pergunta de pesquisa e com o conhecimento do tópico de pesquisa. Em seguida, uma busca automática foi realizada nas bases definidas na seção anterior e os revisores aplicaram os critérios de inclusão e exclusão usando o título e o resumo do artigo. Nesse momento, a ênfase é na inclusão do artigo, a menos que eles não estejam alinhados aos objetivos da pesquisa;

- **Fase II:** artigos incluídos na busca automática foram reunidos em um conjunto de artigos candidatos a serem inseridos. As decisões finais de inclusão ou exclusão foram tomadas quando os artigos foram lidos completamente durante a extração de dados e avaliação da qualidade; e
- **Fase III:** nesta Fase foi realizada uma validação da busca e seleção. A validação da qualidade dos trabalhos foi baseada em Dybå e Dingsøyr (2008), e a avaliação da qualidade dos trabalhos segue um conjunto de critérios descritos na seção 3.4.

Na Figura 7, é possível observar os resultados de cada uma das quatros fases executadas para seleção dos trabalhos na SLR desenvolvida nesta pesquisa. Após a Fase 0, previamente explicada, segue-se, para a segunda análise (Fase I) e nela foram pré-selecionados 64 estudos. Nesta análise, foram lidos o título e o resumos/conclusão de todos os estudos retornados nas buscas. Depois, na Fase II foram selecionados 33 artigos. Na terceira e última fase (Fase III), foi feita a Análise da Qualidade (análise está detalhada e pormenorizada na seção 3.4). Nesta fase os artigos foram lidos na íntegra e classificados de acordo com sua relevância, resultando em um montante final de 17 artigos selecionados.

Figura 7 – Processo de seleção dos trabalhos.



Fonte: Autor.

3.4 ANÁLISE DE QUALIDADE

Na fase III, foi realizada a priorização dos artigos selecionados anteriormente para a execução da Análise de Qualidade. Para isso, foi realizada a leitura na íntegra dos artigos. O processo de Análise de Qualidade foi realizado com base em Dybå e Dingsøyr (2008). Para essa análise foram usados 9 critérios, divididos em 4 categorias: (i) quanto à qualidade do relato; (ii) rigor; (iii) credibilidade; e (iv) relevância do estudo, os quais são apresentados na Tabela 4. Em seguida, para avaliar a qualidade dos estudos de tipos diferentes, foi utilizado um mesmo *checklist*. Para tanto, foi usada uma escala baseada em *Likert* com os seguintes intervalos: -2 (totalmente em desacordo); -1 (em desacordo parcial); 1 (parcialmente em acordo); e 2 (totalmente em acordo). É essencial avaliar a qualidade dos estudos primários incluídos na revisão, porque se o resultado de um determinado estudo for inválido ou se possuir algum viés, então este resultado deve ser considerado no processo de síntese (KITCHENHAM; BUDGEN; BRERETON, 2015).

Os estudos primários que não atendem aos critérios de qualidade listados na Tabela 4 foram excluídos do processo de síntese. Entretanto, antes de efetuar a exclusão, é necessário verificar o impacto da exclusão para o mapeamento. Os artigos remanescentes, que não foram excluídos na Análise de Qualidade, foram analisados nas próximas fases.

Tabela 4 – Critérios para Análise de Qualidade.

Categoria	Critério de Qualidade
Qualidade do Relato	1 . Existe uma declaração explícita dos objetivos da pesquisa?
	2. Existe uma descrição adequada do contexto em que a pesquisa foi realizada?
Rigor	3. O design de pesquisa foi claramente apresentado?
	4. Os grupos de controle foram claramente estabelecidos para a comparação do uso dos tratamentos?
	5. Os dados foram coletados de uma forma clara abordando a questão ou objetivo de pesquisa?
	6. A análise dos dados realizada foi suficientemente rigorosa?
Credibilidade	7. Existe uma declaração clara dos resultados, com justificativa das conclusões?
	8. O estudo apresenta valor para a pesquisa ou prática?
Relevância	9. O estudo aponta novas possibilidades de pesquisa e evoluções do estudo?

Fonte: Adaptado de Dybå e Dingsøyr (2008).

O nível de concordância atingido na Análise de Qualidade usando valores para o número de questões consideradas apropriadas e a média da pontuação para cada artigo será mensurada utilizando o coeficiente de correlação de Pearson (PEARSON, 1914). Esses critérios de Inclusão foram aplicados na fase III, após a Análise de Qualidade, foram selecionados 17 trabalhos para então executar a próxima etapa, que é a extração dos dados. Na Tabela 5, são listados os trabalhos selecionados.

Tabela 5 – Lista de trabalhos selecionados.

Autores	Legenda
(ULLAH; SHAH, 2016)	Work01
(HARADAT et al., 2018)	Work02
(SHOHATA; NAKAMURA; NISHI, 2018)	Work03
(OTGONBAYAR; PERVEZ; DAHAL, 2016)	Work04
(OTGONBAYAR et al., 2018)	Work05
(BERREHILI; BELMEKKI, 2016)	Work06
(DAVOLI; PROTSKAYA; VELTRI, 2017)	Work07
(LIM et al., 2018)	Work08
(MALEKZADEH et al., 2019)	Work09
(NAKAMURA; NISHI, 2018)	Work10
(NAYAHI; KAVITHA, 2017)	Work11
(LIAO et al., 2017)	Work12
(MAHANAN; CHAOVALITWONGSE; NATWICHAI, 2020)	Work13
(RODRIGUEZ-GARCIA; CIFREDO-CHACÓN; QUIRÓS-OLOZÁBAL, 2020)	Work14
(PURI; KAUR; SACHDEVA, 2020)	Work15
(KANWAL et al., 2021)	Work16
(JEON et al., 2021)	Work17

Fonte: O Autor.

3.5 EXTRAÇÃO DOS DADOS E RESULTADOS

O formulário utilizado para a extração de dados foi construído com base no trabalho de (Dybå Dingsøyr, 2008) e está exposto na Tabela 6. É importante que seja realizado um mapeamento entre os itens a serem extraídos e as questões de pesquisa, como meio para guiar o processo de extração, e facilitar o processo de análise e síntese dos dados.

No processo de extração dos dados, todos os trabalhos foram lidos e relidos, executando uma análise criteriosa, já descrita anteriormente, para que pudessem ser extraídas as informações

Tabela 6 – Formulário para extração de dados.

Dados	Descrição
Referências Bibliográficas	Título, Autores, Ano de Publicação.
Instituição de Desenvolvimento da Pesquisa	Universidade, empresas, etc.
Local da publicação	Nome do Journal, conferência, *workshop*, etc.
Objetivos do Estudo	Objetivo principal do estudo.
Método para Avaliação de Sistemas	Métricas avaliadas: Confiabilidade; Disponibilidade; Desempenho; Eficiência energética; Consumo de Recursos do *Hardware*.
Método para Avaliação dos Resultados	Avaliação Quantitativa ou Qualitativa.
Experimento	Foi desenvolvido algum protótipo para avaliação.
Aplicações da Proposta	Identificação das aplicações do método; algoritmo; arquitetura e se foi colocado em prática.

Fonte: Adaptado de Dybå e Dingsøyr (2008).

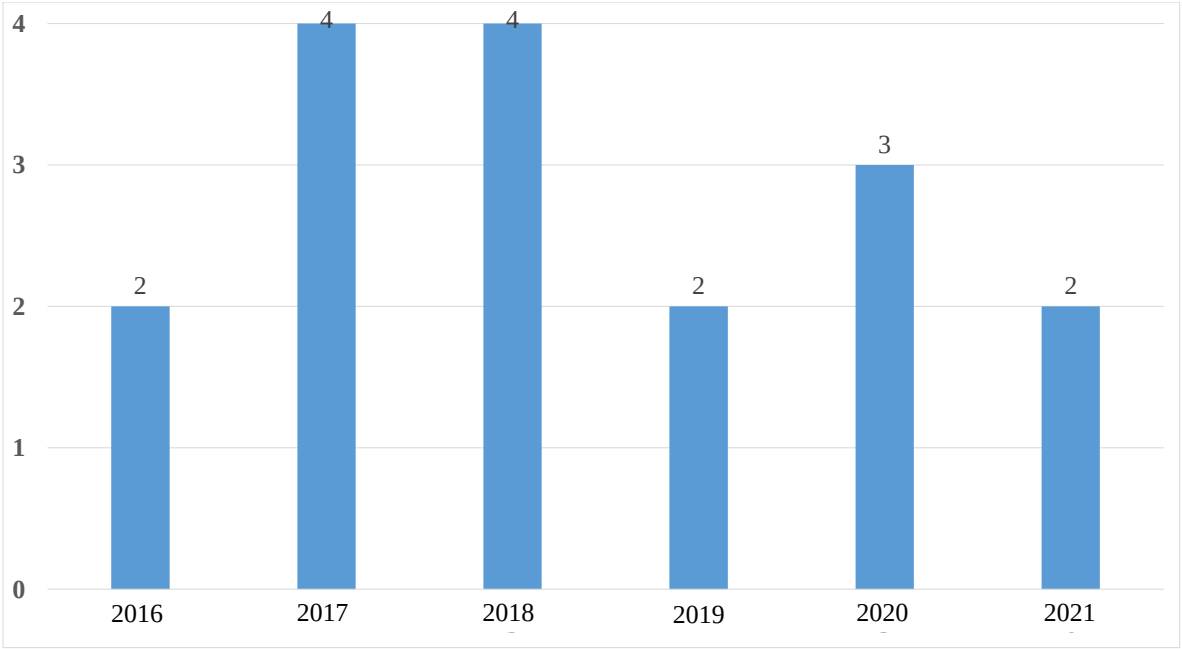
que dessem suporte para as respostas das questões de pesquisa, descritas na seção 3.1. A Tabela 7 apresenta os dados extraídos dos artigos selecionados na revisão. É igualmente importante destacar que alguns dados (objetivos) foram omitidos a fim de respeitar os limites de espaçamento delimitados do documento, contudo é possível encontrar esses dados/objetivos na parte de síntese.

Observando a Tabela 7, é possível identificar que os pesquisadores da *Keio University*, no Japão, se destacam em relação às outras universidades, pois, dos doze trabalhos analisados, três são deles: (HARADAT et al., 2018; SHOHATA; NAKAMURA; NISHI, 2018; NAKAMURA; NISHI, 2018). O pesquisador Nakamura participa desses três trabalhos, em um deles, ele é o autor principal e, em outros dois, ele participa como coautor. Um dos trabalhos foi desenvolvido por pesquisadores da *"IBM T. J. Watson Research Center / Cardiff University"* (LIM et al., 2018); e quase todos os demais trabalhos avaliaram de alguma forma suas propostas. Contudo, apenas um dos trabalhos, o de Berrehili e Belmekki (2016) da *STRS Lab, National Institute of Posts and Telecommunications* não fez avaliação.

A Figura 8 ilustra a distribuição gráfica dos trabalhos em relação ao ano de publicação. Dos dezessete trabalhos analisados, os anos 2016, 2019, e 2021 tiveram dois cada; 2017 e 2018 somam oito das obras analisadas. Para 2020, três das obras foram analisadas. Logo,

dos dezessete trabalhos analisados, quinze deles fizeram avaliações quantitativas. Contudo, Berrehili e Belmekki (2016) não fizeram avaliação. No caso deles, foi apenas proposto o método, mas não foi realizado nenhum experimento. Já Davoli, Protskaya e Veltri (2017) fizeram uma avaliação qualitativa. E todos os trabalhos implementaram pelo menos um algoritmo para testar sua proposta.

Figura 8 – Quantitativo dos trabalhos analisados e os seus respectivos anos de publicação (2016-2021).



Fonte: Autor.

Tabela 7 – Dados extraídos dos artigos selecionados durante o processo de revisão.

Trabalhos	**Instituição**	**Aval**	**Experi**	**Aplica**
Work01	COMSATS Institute of Information Technology	Quant	Sim	Alg
Work02	Keio University	Quant	Sim	Alg
Work03	Keio University	Quant	Sim	hard/Alg
Work04	University of the West of Scotland	Quant	Sim	Alg
Work05	University of the West of Scotland Paisley	Quant	Sim	Alg
Work06	STRS Lab, National Institute of Posts and Telecommunications	-	Não	Alg
Work07	University of Parma	Qual	Sim	prot/Alg
Work08	IBM T. J. Watson Research Center / Cardiff University	Quant	Sim	Arq
Work09	Queen Mary University of London / Imperial College London	Quant	Sim	Alg
Work10	Keio University	Quant	Sim	Alg
Work11	Anna University Regional Campus - Tirunelveli, Tirunelveli / University College of Engineering, Kancheepuram Campus, Kancheepuram	Quant	Sim	Alg
Work12	University of Electronic Science and Technology of China / Guangdong Institute of Electronic and Information Engineering, UESTC / Xi'an Jiaotong Liverpool University	Quant	Sim	Alg
Work13	Chiang Mai University/University of Arkansas	Quant	Sim	Alg
Work14	Universidad de Cádiz	Quant	Sim	prot
Work15	Jaypee Institute of Information Technology Noida/ National Insititute of Technology Delhi	Quant	Sim	Arq
Work16	Comsats University Islamabad/ Cybernetica AS, Estonia/ Aberystwyth University/ University of Hull/ Southern University of Science and Technology	Quant	Sim	Alg
Work17	Hoseo University/ Jeju National University/ Kwangwoon University	Quant	Sim	Alg

Abreviações - **Aval:** Avaliação; **Experi:** Experimento; **Aplica:** Aplicação da Proposta; **Quant:** Quantitativa; **Qual:** Qualitativa; **Alg:** Algoritmo; **prot:** Protocolo; **hard:** *hardware*; **Arq:** Arquitetura

Fonte: O Autor.

3.6 ABORDAGENS PARA ANONIMATO DE DADOS EM IOT

A seguir, é apresentada uma síntese dos dezessete trabalhos previamente selecionados para análise desta revisão, conforme anteriormente apresentados e detalhados na Tabela 7. Nesta seção, são apreciados os pontos fortes e fracos de cada uma das abordagens selecionadas.

3.6.1 *A novel model for preserving Location Privacy in Internet of Things (Work01)*

Em sua pesquisa, Ullah e Shah (2016) propõem a *Enhanced Semantic Obfuscation Technique* (ESOT), que foi projetada para preservar a privacidade da localização de dispositivos gerais da IoT. Essa técnica é baseada na semântica da localização do usuário ou dispositivo. Sua principal preocupação é ocultar a localização do dispositivo IoT do adversário, que pode querer encontrar uma localização real para violar a privacidade.

Nesse cenário, os autores supracitados propõem proteger a localização do usuário do Serviço Baseado em Localização (*Location Based Service* - LBS), pois o LBS não é uma parte confiável. Recomenda-se, portanto, o uso da Ocultação, que é um tipo de técnica utilizada para proteger a privacidade da localização do usuário. Nessa técnica, o local original do usuário é ligeiramente alterado para outro local. Dessa maneira, a localização original da pessoa é ocultada do adversário. A ESOT alcança melhor o desempenho em termos de privacidade do local e utilitário de serviço comparado com a *Semantic Obfuscation Technique* (SOT).

3.6.2 *Anonymization method based on sparse coding for power usage data (Work02)*

Em Haradat et al. (2018), é proposto um método para anonimizar dados de demanda de energia, em que a codificação esparsa é usada para resolver os três problemas que afetam o método convencional. O método proposto pode anonimizar dados de séries temporais e permite que os dados sejam analisados no momento escolhido. O método proposto foi utilizado para anonimizar os dados de uso de energia do *Urban Design Center Misono* (UDCMi) e a taxa de erro experimental diminuiu em comparação com o método convencional.

Prosseguindo com Haradat et al. (2018), existem três padrões de privacidade diferentes chamados *k-anonymity, l-diversity, e t-closeness.* O método proposto pelos autores foca no *k-anonymity*. Esse método atinge os três objetivos principais a seguir: primeiramente, método proposto permite que os dados de séries temporais sejam anonimizados; o objetivo seguinte diz

que a taxa de erro é reduzida em comparação com os métodos convencionais; e, por fim, as informações de uso adquiridas em eletrodomésticos podem ser armazenadas em um ambiente anônimo.

3.6.3 *Hardware for Accelerating Anonymization Transparent to Network (Work03)*

Em sua pesquisa, Shohata, Nakamura e Nishi (2018) propuseram um *hardware* de anonimização de dados que obtém anonimização transparente de dados de fluxos de pacotes de dados de dispositivos IoT. A arquitetura é implementada em um arranjo de portas programável em campo *Field-Programmable Gate Array* (FPGA). O *hardware* proposto é alocado no local intermediário de uma rede para capturar o pacote dos dispositivos IoT. O pacote capturado é anonimizado no *hardware* e encaminhado para o próximo nó enquanto o cabeçalho é modificado sem influenciar seu roteamento.

Como o processo de anonimização não afeta o protocolo de comunicação ou o roteamento de pacotes, o *hardware* proposto obtém anonimização transparente à rede, semelhante a um cabo de comunicação. Essa transparência permite que uma função de anonimização seja instalada em todos os tipos de dispositivos, incluindo dispositivos IoT, sem modificar os dispositivos. O mecanismo proposto alcançou menor consumo de energia e maior rendimento que o processamento de *software*. Além disso, foi confirmado que o FPGA realizou o anonimato de pacotes transparentes para a rede.

Conforme descrito pelos autores Shohata, Nakamura e Nishi (2018), o *k-anonymity* pode ser adotado, mas, neste estudo em especial, eles implementaram a perturbação por adição de ruído, que é um método simples de anonimização que envolve a adição de números pseudoaleatórios a valores numéricos. Os valores-alvo a serem anonimizados foram os dados de uso de energia registrados por um medidor inteligente. O uso não autorizado de dados de uso de energia viola a privacidade das pessoas que vivem em casas. Portanto, a preservação da privacidade dos dados de uso de energia é necessária. O formato dos dados de uso de energia era simplesmente uma matriz de valores numéricos e a faixa de ruído que era previamente fornecida. Os autores focaram na transparência do *hardware* proposto, e em um método simples de anonimização e formato de dados foram empregados.

Apesar de os autores afirmarem que a técnica pode ser usada em outros dispositivos IoT, os mesmos só a implementaram em dois tipos de dispositivos e em um cenário controlado. A avaliação poderia ser mais ampla, com o uso de mais dados e usando conexões diferentes para

ver o comportamento.

3.6.4 *Toward Anonymizing IoT Data Streams via Partitioning (Work04)*

Neste trabalho, Otgonbayar, Pervez e Dahal (2016) apresentam um novo algoritmo de anonimato que publica fluxos de dados da IoT gerados a partir de vários dispositivos sob o modelo de privacidade do *k-anonymity*. Para isso, foi usada a técnica de Janela Deslizante baseada no tempo para manipular fluxos de IoT particionando tuplas com base em sua descrição. Essa operação preliminar ajudou a formar o *cluster* mais rapidamente, localizando tuplas e suportando a fusão das partições quando necessário. O algoritmo proposto superou a abordagem convencional de anonimização do fluxo de dados modificada para anonimizar os fluxos de dados da IoT.

No experimento, foi demonstrado que o anonimato de fluxo convencional não pode ser aplicado diretamente aos dados de streaming da IoT e exigia pesquisa e desenvolvimento significativos. Segundo os autores Otgonbayar, Pervez e Dahal (2016), seria interessante investigar novos modelos de privacidade para agrupar anonimamente com dados ausentes.

3.6.5 *K-VARP: K-anonymity for varied data streams via partitioning (Work05)*

Em seu trabalho, Otgonbayar et al. (2018) propõem o *K-anonymity for VARied data stream via Partitioning* (K-VARP) (*k-anonymity* para fluxo de dados variados via particionamento) para anonimizar variados fluxos de dados. O objetivo é anonimizar e publicar fluxos de dados variados com atraso mínimo e menos perda de informações. O algoritmo *K-VARP* usa métodos de particionamento e marginalização para anonimizar fluxos de dados variados em uma Janela Deslizante baseada no tempo.

Os resultados demonstraram a eficácia do *K-VARP*, pois usa R-similaridade para identificar partições semelhantes na mesclagem. Além disso, uma combinação de marginalização e reutilização flexível tem um impacto significativo no anonimato de fluxos de dados variados. O K-VARP anonimiza fluxos de dados variados com 3% a 9% menos perda de informações e 10% a 20% menos perda de informações em comparação com outros dois algoritmos comparados, gastando tempo semelhante para o cálculo. A usabilidade de dados do *K-VARP* é melhor que os outros dois algoritmos porque o algoritmo proposto não atribui valores ausentes e anonimato tardio impraticável.

3.6.6 *Privacy Preservation in the Internet of Things (Work06)*

Em seu trabalho, Berrehili e Belmekki (2016) analisam a privacidade no contexto da IoT com base em um estudo de caso que propõe mecanismos para melhorar a segurança e preservar a privacidade. Esta análise considera também as vantagens econômicas relacionadas ao uso de IoT como nova oportunidade de negócios sem divulgação de dados pessoais.

O artigo contribui com o campo de privacidade do usuário na IoT, apresentando uma análise de risco em profundidade da ameaça a privacidade e propõe duas abordagens para preservar a privacidade. A primeira são recomendações de melhores práticas que podem reduzir o risco de violação da privacidade na Internet das coisas em aplicações relacionadas à vida humana, e boas práticas para desenvolvedores de aplicativos para IoT. A segunda é o uso de anonimização para ocultar dados gravados que podem ser usados como ameaça para violar a privacidade.

3.6.7 *An anonymization protocol for the Internet of Things (Work07)*

Os pesquisadores Davoli, Protskaya e Veltri (2017) criaram um protocolo de anonimato. O *Social Internet of Things* (SIoT) foi projetado especificamente para as comunicações máquina a máquina IoT. O sistema de anonimato proposto é baseado no conceito de roteamento do *Browser Tor*[1], mas, diferentemente do Tor, o SIoT é completamente orientado a datagrama, com menos protocolos e sobrecarga criptográfica. Além disso, a seleção do caminho é feita por pacote, a fim de aumentar o nível geral de anonimato. Dois modos de caminho de anonimato diferentes foram projetados para suportar facilmente os modelos de anonimização de nó de ponta a ponta e de origem para saída.

O protocolo de anonimato foi implementado como uma camada de datagrama de usuário de anonimato que estende a camada UDP padrão (*Java DatagramSocket*). Isso permitiu uma integração simples e fácil em qualquer aplicativo baseado em UDP, por exemplo, como os clientes e servidores CoAP usados, simplesmente substituindo a camada UDP padrão (*Java DatagramSocket*) pela nova camada de datagrama de anonimato.

Ainda segundo os autores Davoli, Protskaya e Veltri (2017), o escopo principal desta implementação foi validar os princípios de design e a correção do protocolo. Porém, no processo de avaliação, foram adicionados 100 objetos IoT (virtuais) executados em dois dispositivos

[1] https://www.torproject.org/

Raspberry Pi (50 nós virtuais por dispositivo único), ou seja, a avaliação não usou objetos reais, os mesmos foram simulados.

3.6.8 *Learning Light-Weight Edge-Deployable Privacy Models (Work08)*

Segundo os autores Lim et al. (2018), a estrutura desenvolvida por eles utiliza o *Apache Spark* para criar rapidamente um modelo de anonimização implementável na borda para uma enorme quantidade de dados. Depois que um modelo é criado, a estrutura permite que até usuários ou dispositivos de ponta da IoT ofusquem seus próprios registros sem grande sobrecarga de computação e conhecimento de dados inteiros.

Para refletir as naturezas variáveis dos dados no tempo, a estrutura também fornece uma função de verificação para validar se um modelo de anonimização satisfaz as restrições de privacidade de dados desejadas. Portanto, os administradores de dados não precisam calcular continuamente o modelo de anonimização dos registros recebidos, e eles só treinam um modelo de anonimização novamente apenas quando um modelo atual não passa no procedimento de validação.

Neste artigo, os autores implementam e avaliam a estrutura de anonimização de dados escalável e leve para a implantação flexível da função de anonimização que eles propõem. Segundo Lim et al. (2018), a estrutura agiliza o processo de aprendizado de regras de anonimização, aplicando computação paralela e gera modelos viáveis para serem implementados nos dispositivos de última geração. Também foram investigados vários fatores que afetam o desempenho da anonimização, bem como a paralelização. Os resultados experimentais mostram que a estrutura é capaz de reduzir o tempo para construir modelos de anonimização. As avaliações mostram que a estrutura aprende modelos de anonimização até 16 vezes mais rápido que uma abordagem de anonimização sequencial e preserva informações suficientes em dados anonimizados para aplicativos orientados a dados.

3.6.9 *Mobile sensor data anonymization (Work09)*

Em suas pesquisa Malekzadeh et al. (2019), propõem uma transformação no dispositivo de dados do sensor a ser compartilhada para aplicativos específicos, como o monitoramento de atividades diárias selecionadas, sem revelar informações que permitam a identificação do usuário. Os autores formularam o problema de anonimização usando uma abordagem teórica

da informação e propuseram uma nova função de perda multi-objetiva para o treinamento de autocodificadores profundos. Essa função de perda ajuda a minimizar as informações de identidade do usuário e a distorção dos dados para preservar o utilitário específico do aplicativo.

De acordo com Malekzadeh et al. (2019), para remover os recursos identificáveis pelo usuário inclusos nos dados, foi considerado não apenas o extrator de recursos do modelo de rede neural (codificador), como também o reconstrutor (decodificador), que foi forçado a moldar a saída final independentemente de cada usuário no conjunto de treinamento, portanto, o modelo final treinado é um modelo generalizado que pode ser usado por um novo usuário invisível. A proposta dos autores garante que os dados transformados sejam minimamente perturbados, para que um aplicativo ainda possa produzir resultados precisos, por exemplo, para reconhecimento de atividades. A solução proposta por Malekzadeh et al. (2019) é importante para garantir o anonimato da detecção participativa, quando os indivíduos contribuem com dados registrados por seus dispositivos pessoais para análise de dados de saúde e bem-estar.

3.6.10 *TMk-Anonymity: Perturbation-Based Data Anonymization Method for Improving Effectiveness of Secondary Use (Work10)*

No trabalho de Nakamura e Nishi (2018), foi desenvolvida uma solução para os problemas de privacidade dos dados divulgados. Os autores definiram o *TMk-anonymity* e propuseram três métodos diferentes de anonimato: (i) o método de condensação; (ii) o método de adição de ruído; e (iii) o método de adição de condensação e ruído (CoNoA), que combina os dois métodos.

O *TMk-anonymity* pode ser mais prático da perspectiva de que pode dar diferentes graus de anonimato a diferentes indivíduos em comparação com o *k-anonymity* convencional. Além disso, o *Root Mean Squared Error* (RMSE) dos métodos de anonimização propostos foi de aproximadamente 0,7% em comparação com o do *Pk-anonymization* convencional. Os métodos propostos, especialmente o método CoNoA, podem reter os recursos dos dados originais para os dados normais de distribuição e os dados de localização GPS. Isso ocorre porque o método CoNoA usa o método de adição de ruído que adiciona ruídos conforme os dados originais e o método de condensação que permite o anonimato dos dados com RMSE menor em comparação com demais métodos.

3.6.11 *Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop (Work11)*

A pesquisa de Nayahi e Kavitha (2017) propõe um algoritmo de anonimização baseado em *cluster* e resiliente a ataques de similaridade e ataques de inferência probabilística. Os dados anonimizados são distribuídos no *Hadoop Distributed File System*. Os autores defendem que o algoritmo alcança um melhor compromisso entre privacidade e utilidade. O desempenho é medido em termos de precisão e *FMeasure* em relação a diferentes classificadores.

Neste trabalho é empregado o *k-anonymity* para preservar a privacidade dos dados em um ambiente distribuído. Os autores afirmam que propuseram um algoritmo de agrupamento para alcançar *k-anonymization* e *l-diversidade* resiliente ao ataque de similaridade e ataque de inferência probabilística. Mais tarde, os conjuntos de dados anonimizados preservados pela privacidade são distribuídos no *Hadoop*.

Segundo Nayahi e Kavitha (2017), os proprietários dos dados podem distribuir os dados de maneira segura, aplicando os algoritmos de *cluster* propostos em qualquer estrutura distribuída. As informações sigilosas nos dados são protegidas contra ataques de ligação, ataque de homogeneidade, ataque de similaridade e ataque de inferência probabilística usando os algoritmos de *cluster* propostos.

3.6.12 *The framework and algorithm for preserving user trajectory while using location-based services in IoT-cloud systems (Work12)*

O trabalho de Liao et al. (2017) tem o foco principal na solução de dois problemas: (i) preservação da privacidade da localização para uma única consulta; (ii) preservação da privacidade da trajetória para consultas contínuas. Eles propõem um algoritmo eficiente baseado na técnica de *k-anonymity* para proteger a privacidade da trajetória do usuário em serviços baseados em localização. E, para preservar melhor a privacidade do local e reduzir a complexidade do tempo, o mecanismo de *k-anonymity* proposto é baseado em Janela Deslizante.

Segundo os autores Liao et al. (2017), o *k* Janela Deslizante seleciona locais fictícios; já o *Trajectory Selection Mechanism* (TSM) seleciona as trajetórias fictícias. Os resultados da simulação mostram que o algoritmo reduz a complexidade do tempo em comparação com as soluções existentes para uma única consulta e preserva efetivamente a privacidade da trajetória dos usuários para consultas contínuas.

3.6.13 *Data anonymization: a novel optimal k-anonymity algorithm for identical generalization hierarchy data in IoT (Work13)*

Na sua investigação, Mahanan, Chaovalitwongse e Natwichai (2020) apresentam um novo algoritmo de k-anonimato para dados idênticos da hierarquia de generalização (IGH), que é o principal tipo de dados no ambiente IoT. Os autores propuseram um novo método para fornecer uma solução globalmente otimizada de k-anonimato para os conjuntos de dados IGH. Os algoritmos propostos determinam uma solução ótima baseada em características idênticas de dados da hierarquia de generalização (IGH), visitando e avaliando apenas os nós essenciais da malha de generalização que satisfazem o k-anonimato. Uma vez que o problema de k-anonimização é do tipo NP-difícil, foi demonstrado que o algoritmo poderia encontrar eficazmente soluções ótimas de k-anonimato, explorando características especiais dos dados IGH, ou seja, a otimização entre nós a diferentes níveis da rede de generalização. A partir dos resultados experimentais, segundo os autores, é evidente que o algoritmo é muito mais eficiente do que os algoritmos comparativos, exigindo menos pesquisa na árvore dada.

De acordo com Mahanan, Chaovalitwongse e Natwichai (2020), a ideia chave do algoritmo é analisar apenas os nós necessários, que são os que se encontram no nível mais baixo de generalização, encontrados como nós k-anônimos, em contraste com outros algoritmos da literatura onde se tem de examinar todos os nós. O algoritmo encontra primeiro as rotas do nó raiz da malha de generalização, por exemplo, (000), até ao nó de nível mais alto, utilizando o método de travessia de pré-encomenda. *K-anonymity* determina todos os nós nas rotas a partir do nó ao nível mais baixo. Os nós k-anônimos devem ser marcados, e o nível mais baixo encontrado k-anônimo, chamado nível k-anônimo, é definido. O algoritmo atravessa as outras rotas e visita apenas os nós no nível mais baixo que o anônimo até que todos os nós no nível mais baixo que o anônimo tenham sido encontrados e marcados.

3.6.14 *Cooperative Privacy-Preserving Data Collection Protocol Based on Delocalized-Record Chains (Work14)*

Os autores Rodriguez-Garcia, Cifredo-Chacón e Quirós-Olozábal (2020) introduzem um novo mecanismo de comunicação anônima colaborativa destinado a ambientes multiusuários. A proposta caracteriza-se por ser uma solução autônoma adaptada à natureza distribuída de um ambiente IoT, em que os utilizadores interessados em obter anonimato trabalham em sinergia

para anonimizar as suas transmissões de dados. Como a solução carece de intermediários de terceiros, é particularmente apropriada para redes privadas, tais como as redes privadas de IoT.

A nova coleção de dados gera um protocolo denominado *Cooperative Privacy-Preserving Data Collection protocol* (cPPDC) que oferece condições de preservação de privacidade tanto na recolha de dados como na publicação sem limitar a recolha de dados. Esse método pode ser utilizado para k-anonimizar o conjunto de dados. Este protocolo é resistente a ataques de análise de tráfego de rede, utilizando a cadeia de registos deslocalizados como meio de transmissão de dados na fase de recolha. Esse protocolo pode gerar conjuntos de dados k-anónimos em ambientes IoT, protegendo, na sua fonte, os dados pessoais que um conjunto de dispositivos envia a um colector central. Para alargar o requisito de privacidade à fase de recolha de dados, o protocolo utiliza um novo mecanismo de comunicação anônima colaborativa denominada cadeia de registos deslocalizada. Uma vez que o protocolo não requer canais de comunicação anônima de terceiros, a sua aplicação é especialmente relevante em ambientes de IoT implantados em redes privadas. (RODRIGUEZ-GARCIA; CIFREDO-CHACÓN; QUIRÓS-OLOZÁBAL, 2020).

3.6.15 *Data Anonymization for Privacy Protection in Fog Enhanced Smart Homes (Work15)*

Em sua pesquisa Puri, Kaur e Sachdeva (2020) apresentam a arquitetura de um ambiente doméstico inteligente e melhorado com *fog-enhanced* para preservar a privacidade dos indivíduos quando os dados das suas casas inteligentes são partilhados com terceiros. É utilizada uma técnica de anonimização de dados para prevenir violações da privacidade por indivíduos. A arquitetura e a técnica propostas são avaliadas num conjunto de dados do mundo real, e os resultados indicam a sua eficácia na preservação da privacidade.

Na arquitetura proposta por Puri, Kaur e Sachdeva (2020), o módulo de preservação da privacidade *privacy preservation module* (PPM) é colocado sobre a camada de nevoeiro. Os dados sensíveis recolhidos de casas inteligentes numa comunidade ou complexo de apartamentos inteligentes são alimentados pelo PPM. O PPM aplica uma técnica de corte (Li et al., 2012) aos dados agregados. Ele remove a ligação entre os campos constituintes dos registos dos utilizadores. Esta remoção anonimiza os dados e, consequentemente, não é possível associar um registo de dados a um determinado utilizador ou casa. Essa impossibilidade protege os dados do utilizador contra violações da privacidade.

3.6.16 *A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes (Work16)*

Em sua pesquisa, Kanwal et al. (2021) formalizaram o comportamento de um adversário, realizando revelações de identidade e atributos num modelo equilibrado de k-anonimato p+-sensível com a ajuda de cenários adversos, uma vez que o modelo equilibrado de k-animato p+-sensível não é suficiente para 1 a M com múltiplos atributos sensíveis (*multiple sensitive attributes* - MSAs) em termos de preservação da privacidade. Os autores cunharam um modelo de privacidade alargado denominado "1: MSA-(p, l)-diversidade"para 1: conjunto de dados M com MSAs.

Em seguida, realizaram a modelação formal e verificação do modelo proposto, utilizando redes de Petri de alto nível para confirmar a invalidação do ataque a privacidade. Os resultados experimentais mostram que o modelo proposto "1: MMSA-(p, l)-diversidade"é eficiente e fornece uma maior utilidade dos dados publicados (KANWAL et al., 2021).

Os autores Kanwal et al. (2021) propõem uma abordagem baseada em conjuntos de dados contendo MSAs e registos múltiplos de um único doente (1: M) em registo sanitário eletrônico. É feita para classificar os cenários de privacidade para a divulgação de identidade e atributos sensíveis num modelo equilibrado de k-animato p+ sensível e propõem uma versão melhorada chamada "1: MSA- (p, l)-diversidade"para anonimizar os dados dos registos sanitários eletrônicos.

3.6.17 *Distributed L-diversity using Spark-based algorithm for large resource description frameworks data (Work17)*

Nesta pesquisa, os autores Jeon et al. (2021) propõem um método de desidentificação da anatomia da l-diversidade que pode ultrapassar as limitações do k-anonimato e garantir uma proteção de privacidade mais forte do que o k-anonimato. Além disso, como este processo de anonimização de dados é computacionalmente intensivo em termos de tempo, a computação distribuída *Spark* foi usada para proporcionar uma reidentificação rápida para aumentar a sua utilidade. A preservação da l-diversidade também foi proposta para conjuntos de dados *Resource Description Frameworks* (RDF) em evolução dinâmica. Os resultados experimentais mostram que a distribuição proposta do algoritmo de l-diversidade processa os dados de forma mais eficiente do que as abordagens convencionais.

Segundo os autores Jeon et al. (2021), as principais contribuições desta pesquisa são as seguintes: (1) aplicação do algoritmo de anatomia existente do modelo l-diversidade ao modelo RDF; (2) fornecimento de um algoritmo de anatomia baseado em *Spark* para um processamento mais eficiente de dados RDF em grande escala (*Spark* é amplamente utilizado para o grande processamento de dados e estende o modelo *MapReduce*); (3) desenvolvimento de um algoritmo de anonimização de dados em l-diversidade para conjuntos de dados RDF em evolução dinâmica; (4) apresentação de uma abordagem mais eficiente em comparação com os métodos convencionais, de acordo com resultados experimentais

3.7 PRINCIPAIS TÉCNICAS DE ANONIMIZAÇÃO USADAS

Os trabalhos selecionados foram classificados levando-se em consideração qual técnica usam para fornecer privacidade aos dados. Dessa forma, todas as técnicas usadas nos trabalhos selecionados e analisados executam de alguma maneira a anonimização dos dados. É importante destacar que o tópico anonimização é destinado a trabalhos que não categorizaram especificamente qual técnica foi adotada na execução de suas respectivas propostas (Tabela 8). Os trabalhos marcados como *k-anonymity* significa que, ou eles usaram a técnica sem modificação, ou a usaram como base para fazer uma versão melhorada, ou seja, eles a usaram indiretamente.

Após a análise dos dezessete trabalhos selecionados na revisão, foi identificado que onze entre eles Haradat et al. (2018), Shohata, Nakamura e Nishi (2018), Otgonbayar, Pervez e Dahal (2016), Otgonbayar et al. (2018), Lim et al. (2018), Nakamura e Nishi (2018), Nayahi e Kavitha (2017), Liao et al. (2017), Mahanan, Chaovalitwongse e Natwichai (2020), Rodriguez-Garcia, Cifredo-Chacón e Quirós-Olozábal (2020), Kanwal et al. (2021) usam a técnica "*k-anonymity*". Desses trabalhos que usam o "*k-anonymity*" quatro deles Shohata, Nakamura e Nishi (2018), Otgonbayar et al. (2018), Lim et al. (2018), Liao et al. (2017) usam em conjunto com o "*k-anonymity* mais uma técnica. Outros dois trabalhos, Davoli, Protskaya e Veltri (2017), Berrehili e Belmekki (2016) usam anonimização para fornecer privacidade aos dados. Os mesmos autores só especificam que usam anonimização, e não especificam se usam algumas das técnicas contidas na Tabela 6, ou outra técnica. A técnica de Janela Deslizante foi usada em dois trabalhos Liao et al. (2017), Otgonbayar et al. (2018). Dos doze trabalhos analisados, quatro, Liao et al. (2017), Shohata, Nakamura e Nishi (2018), Otgonbayar et al. (2018), Lim et al. (2018) usam duas técnicas ao mesmo tempo.

Tabela 8 – Principais técnicas de anonimização.

Trabalhos	Ofusca	Pertuba	k-anonymi	Anonimiz	Janela Desliz	Generaliza
Work01	Sim	Não	Não	Não	Não	Não
Work02	Não	Não	Sim	Não	Não	Não
Work03	Não	Sim	Sim	Não	Não	Não
Work04	Não	Não	Sim	Não	Não	Não
Work05	Não	Não	Sim	Não	Sim	Não
Work06	Não	Não	Não	Sim	Não	Não
Work07	Não	Não	Não	Sim	Não	Não
Work08	Sim	Não	Sim	Não	Não	Não
Work09	Não	Sim	Não	Não	Não	Não
Work10	Não	Não	Sim	Não	Não	Não
Work11	Não	Não	Sim	Não	Não	Não
Work12	Não	Não	Sim	Não	Sim	Não
Work13	Não	Não	Sim	Não	Não	Não
Work14	Não	Não	Sim	Não	Não	Não
Work15	Não	Não	Não	Não	Não	Sim
Work16	Não	Não	Sim	Não	Não	Não
Work17	Não	Não	Não	Sim	Não	Não

abbreviation - Ofusca: Ofuscação; Pertuba: Perturbação; k-anonymi: *k-anonymity*; Anonimiz: Anonimização; Janela Desliz: Janela Deslizante; Generaliza: Generalização

Fonte: O Autor.

Com base nos trabalhos analisados até o momento da elaboração/realização desta pesquisa, não foi identificada uma técnica de anonimização que seja aplicada em vários cenários de uso da IoT. A maioria dos trabalhos analisados usam técnicas distintas para anonimizar os dados ou fazem a combinação de mais de uma técnica. Entretanto, é importante destacar que a técnica *k-anonymity* é usada na maioria dos trabalhos analisados, inclusive em alguns casos é usada em combinação com outra técnica, ou são feitas modificações no *k-anonymity*.

Diante dos dados extraídos nesta revisão, foi possível demonstrar que a anonimização de dados é uma solução promissora no fornecimento de privacidade aos dados gerados pelos dispositivos IoT. Existem várias técnicas que estão sendo usadas atualmente para prover anonimização no contexto de IoT, dependendo do contexto de uso uma técnica ou outra pode ser mais usada.

3.8 TRABALHOS POR DOMÍNIO DE APLICAÇÃO

Na Tabela 9, é feita uma análise comparativa entre os trabalhos selecionados, considerando a área de aplicação de cada estudo. O objetivo desta análise comparativa entre os trabalhos é identificar que áreas são mais estudadas e usadas para tratar a privacidade de dados em IoT.

Tabela 9 – Classificação dos trabalhos de acordo com os domínios da IoT.

Trabalhos	**Industrial**	**Smart city**	**HealthCare**	**Indefinido**
Work01	Não	Sim	Não	Não
Work02	Não	Sim	Não	Não
Work03	Não	Não	Não	Sim
Work04	Não	Não	Não	Sim
Work05	Não	Não	Não	Sim
Work06	Não	Não	Não	Sim
Work07	Não	Não	Não	Sim
Work08	Não	Não	Não	Sim
Work09	Não	Não	Sim	Não
Work10	Não	Não	Não	Sim
Work11	Não	Não	Não	Sim
Work12	Não	Não	Sim	Não
Work13	Não	Não	Não	Sim
Work14	Não	Não	Não	Sim
Work15	Não	Sim	Não	Não
Work16	Não	Não	Sim	Não
Work17	Não	Não	Não	Sim

Fonte: O Autor.

Os domínios são analisados conforme a divisão proposta por Borgia (2014). Uma vez que alguns trabalhos não especificam onde se encontra a sua aplicação ou onde testaram as suas técnicas, foi acrescentada uma quarta categoria: "Indefinido". Essa última categoria destina-se a obras que não foram atribuídas ou dirigidas a algum domínio devido a uma falta de dados de informação utilizados nos testes.

Conforme a análise dos trabalhos, foi possível identificar que o Domínio Cidade Inteligente e o Domínio Bem-estar Saudável tinham ambos três trabalhos, que desenvolveram algum mecanismo para tratar a privacidade utilizando a anonimização. Com as informações presentes nos trabalhos pesquisados, não foi possível identificar nenhum que tivesse desenvolvido algo para o Domínio Industrial. Onze trabalhos foram classificadas como Domínio indefinido porque,

nesses trabalhos, não havia informações que os classificassem em qualquer domínio. Esta análise permite observar que a maioria das obras são genéricas: em primeiro lugar, foram desenvolvidas para a IoT como um todo, e os autores não se concentram em nenhuma área da IoT; outra hipótese é que não colocaram informação suficiente devido ao espaço regulamentado nas suas respectivas obras.

3.9 TRABALHOS RELACIONADOS

Nesta Seção, são apresentados os trabalhos relacionados com esta Tese. Mais especificamente, são apresentados trabalhos relacionados ao uso de sistemas de recomendação para a área de IoT. Os trabalhos descritos são comparados com a abordagem adotada nesta Tese.

Em Cepeda-Pacheco e Domingo (2022), é proposto um sistema de recomendação de atrações turísticas para cidades inteligentes baseado em aprendizagem profunda. São integradas informações secundárias sobre os usuários em uma rede neural profunda. O sistema coleta dados de IoT sobre a visita do turista na cidade inteligente para fazer recomendações de atrações em tempo real. São incluídos os atrativos turísticos já visitados para melhorar o rigor das recomendações com as escolhas do próprio turista. Segundo os autores, o classificador de aprendizagem profunda multirrótulo proposto pode recomendar com sucesso atrações turísticas para o primeiro caso [(a) pesquisar e planejar atividades antes de viajar] com a perda, exatidão, precisão, *recall* e *F1-score* de 0,5%, 99,7%, 99,9%, 99,9% e 99,8%, respectivamente. Ele também pode recomendar com sucesso atrações turísticas para o segundo caso [(b) procurar atividades dentro da cidade inteligente] com perda, exatidão, precisão, *recall* e *F1-score* de 3,7%, 99,5%, 99,8%, 99,7% e 99,8%, respectivamente.

A pesquisa de Qi et al. (2022) oferece um modelo de recomendação de categoria de ponto de interesse (*point-of-interest - POI*) baseado em preferências de grupo PPCM (*preference-based POI category recommendation model*). O PPCM que pode construir grupos de usuários semelhantes de forma rápida e automática, protegendo a privacidade dos usuários e, em seguida, usar diferentes codificadores de aprendizagem profunda para recomendar categorias de POI para diferentes grupos. O trabalho pode ser resumido em três aspectos: (i) são usados LSH (*locality-sensitive hashing*) para encontrar usuários com tempo semelhante em várias plataformas IoT e, em seguida, são consideradas as semelhanças entre grupos de usuários para recomendar POIs; (ii) é apresentado um modelo de recomendação de categoria de POI baseado em preferência de grupo (PPCM) para recomendação de categoria de POI. Considerando a

importância da influência do grupo, é feita a combinação da influência dos grupo dos usuários, e também é usado o LSTM *(long short-term memory*) para recomendar categorias de POI; e (iii) dois conjuntos de dados reais de cidades coletados por sensores que são usados para construir experimentos comparativos. Segundo os autores, o PPCM apresenta melhor desempenho de recomendação do que os outros quatro métodos comparados em sua pesquisa.

Em sua pesquisa, Forestiero (2022) projetou uma metodologia capaz de realizar recomendações/sugestões de serviços/dispositivos úteis em um ambiente inteligente, caracterizado pela alta dinamicidade e falta de confiabilidade das entidades envolvidas. Objetos inteligentes são representados usando vetores de valor real obtidos através do Doc2Vecmodel[2]: uma técnica de incorporação de palavras capaz de capturar o contexto semântico que representa documentos e frases com vetores densos. Os vetores estão associados a agentes móveis que se movem num espaço virtual 2D, seguindo um modelo bioinspirado, denominado de modelo flocagem, no qual os agentes realizam operações simples e locais de forma autônoma, obtendo uma organização global inteligente. Uma regra de similaridade baseada nos vetores atribuídos, foi projetada para permitir que os agentes discriminem entre eles. Um posicionamento mais próximo (*clustering*) apenas de agentes semelhantes é alcançado. O posicionamento inteligente permite a identificação de objetos inteligentes semelhantes, possibilitando, assim, operações de seleção rápidas e eficazes. Avaliações experimentais permitiram demonstrar a validade da abordagem, e como a metodologia proposta permite obter um aumento de desempenho de cerca de 50%, em termos de qualidade e relevância do agrupamento, em comparação com outras abordagens existentes.

Na pesquisa de Huang et al. (2019), é proposto um novo modelo baseado em aprendizagem de representação multimodal (*multimodal representation learning-based model* - MRLM) para realizar recomendações precisas em IoT. No MRLM, dois módulos estreitamente relacionados foram treinados simultaneamente. Eles são o aprendizado de representação de recursos globais (*global feature representation learning* - GFRL) e o aprendizado de representação de recursos multimodais (*multimodal feature representation learning* - MFRL), e a função de perda conjunta desses dois módulos foi minimizada. Além disso, a eficácia do modelo MRLM foi estudada através de extensos experimentos em dois conjuntos de dados do mundo real. Os resultados mostraram que o MRLM superou significativamente os modelos de última geração em termos de diversas métricas de avaliação em IoT.

Em sua pesquisa, Cui et al. (2020) propõem um novo algoritmo de filtragem colaborativa

[2] https://radimrehurek.com/gensim/auto$_{e}xamples/tutorials/run_{d}oc2vec_{l}ee.html$

baseado no coeficiente de correlação de tempo (TCC) e nas médias CSK (TCCF). TCCF usa algoritmo *CSK-means* para separar problemas de big data em vários problemas menores gerenciáveis. O método de clustering é um pré-processamento que agrupa usuários semelhantes para recomendações mais rápidas e precisas. Primeiro, é empregado um novo algoritmo de otimização inteligente, Cuckoo search, para otimizar o algoritmo *K-means* a fim de melhorar o efeito de agrupamento. Em seguida, foi projetado um fator tempo para resolver o desvio de juros ao longo do tempo. Finalmente, foi projetado um modelo de recomendação eficaz e personalizado baseado no padrão de preferência (PTCCF) para melhorar a qualidade do TCCF. Ele pode fornecer uma recomendação de maior qualidade, analisando o comportamento do usuário. Para avaliar o desempenho, foram conduzidos experimentos em dois conjuntos de dados reais de MovieLens e Douban, e a precisão do modelo melhorou cerca de 5,2% em comparação com o modelo MCoC. Resultados experimentais sistemáticos demonstraram que os modelos TCCF e PTCCF são eficazes para cenários de IoT.

Existem várias técnicas que podem ser utilizadas para o desenvolvimento de um sistemas de recomendação. Os vários tipos de sistemas de recomendação existentes estão sendo utilizados em diversas áreas, como entretenimento para recomendação de filmes, ou, por exemplo, na área de vendas para a recomendação de produtos conforme o perfil do cliente, também são amplamente utilizados em redes sociais para recomendação de conteúdo. É possível encontrar na literatura várias iniciativas para desenvolver sistemas de recomendação para a área de IoT, como os descritos anteriormente.

O Smart Anonymity foi desenvolvido para recomendar algoritmos de anonimização para dados de IoT. As principais contribuições e características que diferenciam o *Smart Anonymity* dos demais citados nesta Tese, é que ele usa ontologias OWL para classificar os dados analisados para fazer a recomendação. Além disso, também usa *Machine Learning* para melhorar os resultados da classificação feita pelas ontologias. Outro ponto que o diferencia dos demais é a sua finalidade, pois o *Smart Anonymity* foi desenvolvido para recomendar algoritmos de anonimização para dados de ambientes IoT. Dessa forma, é possível aplicar o algoritmo de anonimização mais adequado para um determinado conjunto de dados, e isso possibilita que as empresas responsáveis por armazenar dados de dispositivos IoT possam fornecer uma maior privacidade aos dados dos usuários.

3.10 CONSIDERAÇÕES FINAIS

Por meio de uma análise pormenorizada e interpretação abrangentes dos dados coletados, o panorama apresentado indica uma falta de consenso quanto ao uso das técnicas no que diz respeito às metodologias e padrões de uso. No entanto, mediante os resultados alcançados nesta Revisão Sistemática da Literatura, a técnica *k-anonymity* mostrou-se bastante útil, sendo utilizada tanto isolada como combinada com outras técnicas mencionadas nos trabalhos pesquisados. Sendo seu uso, portanto, um desafio extremamente viável.

Por meio da análise das pesquisas encontradas, o objetivo dessa SLR foi alcançado. Após sua finalização foi possível identificar e ordenar as principais categorias, técnicas aplicadas, e métodos de anonimização de dados usadas no contexto de IoT. As principais técnicas mapeadas são: (i) Ofuscação de Dados; (ii) Generalização; (iii) Perturbação; (iv) *k-anonymity*; e (v) Janela Deslizante. Foi possível responder aos questionamentos feitos inicialmente sobre a quantidade de estudos com foco da pesquisa sobre anonimização de dados para IoT. Foram identificados os indivíduos e as organizações que são mais ativas nas pesquisas baseadas em anonimização de dados para IoT. Por fim, foi possível compreender os desafios encontrados, dessa forma a compreensão resultou em direções e oportunidades e desafio no tocante às pesquisas futuras na área de anonimização de dados em IoT.

Entre os trabalhos analisados, não foi identificada a existência de uma solução para anonimização de dados no cenário da IoT que possa ser aplicada aos vários ambientes de uso da IoT. Como consequência disso, este é um problema em aberto, visto que existem diversos tipos de dados, com características diversas para serem tratados, porém, as soluções existentes foram criadas para anonimizar os dados em cenários bem específicos. Diante disso, foi possível identificar a necessidade de uma solução autoadaptável que possa recomendar o algoritmo de anonimização mais adequado para um conjunto de dados considerando suas características. A solução proposta nesta Tese tem por objetivo recomendar um algoritmo de anonimização para um determinado conjunto, porém, para tornar a anonimização mais eficiente, as características dos dados analisados precisam ser consideradas.

4 SMART ANONYMITY: UM MECANISMO PARA RECOMENDAÇÃO DE ALGORITMOS DE ANONIMIZAÇÃO DE DADOS BASEADO NO PERFIL DOS DADOS PARA AMBIENTES IOT

Como visto no Capítulo 3, até o presente momento, não foi identificada na literatura uma solução para privacidade de dados no contexto de IoT que seja ampla o suficiente para ser aplicada em vários subdomínios de uso da IoT. Nesse contexto, este capítulo apresenta a principal contribuição desta Tese, que é uma solução de recomendação que possa fornecer privacidade aos dados gerados por dispositivos IoT, indicando qual algoritmo de anonimização de dados é o mais adequado para um conjunto de dados de acordo com suas características.

4.1 TAXONOMIA PARA CLASSIFICAÇÃO DOS DADOS EM IOT

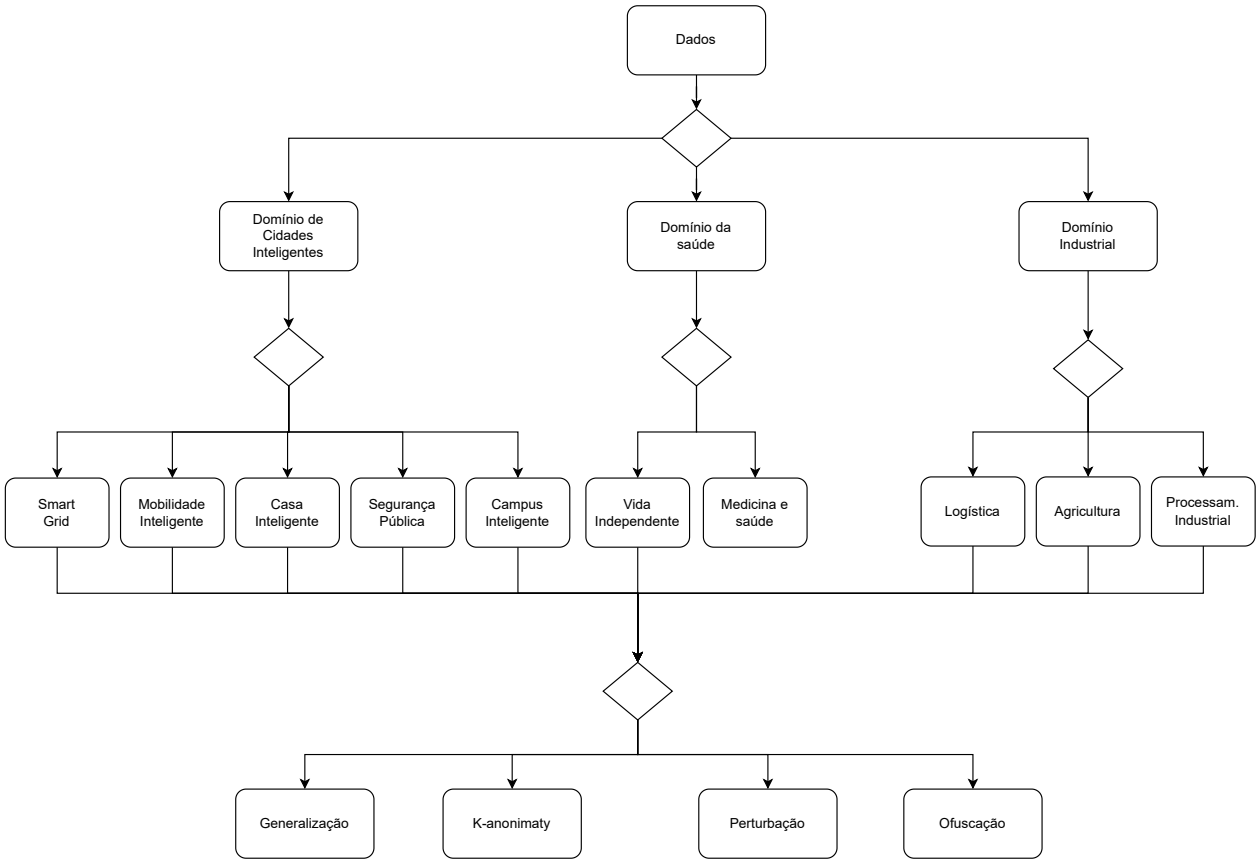
Para o desenvolvimento do *Smart Anonymity*, é feita a divisão da IoT em 10 subdomínios. Essa divisão é baseada no trabalho de Borgia (2014), que fez a classificação das áreas da IoT em 9 subdomínios. Outro autor usado como referência para essa divisão foi Pekar et al. (2020), que fez uma divisão semelhante, dividindo a IoT em 8 áreas de aplicação. Vale ressaltar que, nesta Tese, foram adicionados mais dois subdomínios: o Smart Campus e o Indefinido. Estes 11 subdomínios apresentados, conforme a Tabela 10, ainda são classificados como público ou privado conforme as características dos dados. A Figura 9 ilustra a taxonomia para classificação dos dados e os algoritmos de anonimização utilizados nesta Tese.

Tabela 10 – Classificação dos Subdomínios.

Subdomínio	Privado	Público
Logística	Sim	Não
Agricultura	Sim	Não
Processamento industrial	Sim	Não
Mobilidade inteligente	Não	Sim
Energia Inteligente	Sim	Não
Casa/edifício inteligente	Sim	Não
Segurança pública	Não	Sim
Campus Inteligente	Não	Sim
Vida Independente	Sim	Não
Medicina e Saúde	Sim	Não

Fonte: O Autor.

Figura 9 – Taxonomia de Classificação dos subdomínios.



Fonte: Autor.

Os subdomínios são utilizados para fazer a classificação dos Datasets de dados gerados por dispositivos IoT. Os dados de cada subdomínios foram listados usando o trabalho da autora Borgia (2014) como referência, posteriormente, foi feita uma busca *ad hoc* para aperfeiçoar essa classificação. Diante disto, as categorias e seus respectivos dados são listados a seguir:

1. **Logística e gerenciamento de vida do produto** - os dados a serem considerados, foram definidos baseados no trabalho de Borgia (2014); posteriormente foi feita uma busca *ad hoc* para aperfeiçoar essa classificação; e os trabalhos de Hsu et al. (2015), Saputra e Suryanegara (2019) foram usados para esse aperfeiçoamento. A seguir, são listados os dados referentes a este subdomínio da IoT:

    - Dados de identificação de produtos – como data de validade, origem, destino;
    - Dados de qualidade do produto;
    - Dados com localização de onde os produtos estão armazenados;
    - Dados de localização de veículos;

- Dados de sensores de porta;
- Dados de consumo de combustível do veículo;
- Dados de controle de velocidade do veículo; e
- Dados de nível de estoque.

2. **Agricultura e criação** - os dados a serem levados em consideração foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi realizada uma busca *ad hoc* para aperfeiçoar essa classificação; e os trabalhos de Zhao et al. (2010), Dholu e Ghodinde (2018), Dagar, Som e Khatri (2018), Kjellby et al. (2019), Nagaraja et al. (2019) foram usados. A seguir, são listados os dados referentes a este subdomínio da IoT:

   - Dados de umidade no solo;
   - Dados de qualidade do solo (pH);
   - Dados de saúde de animais;
   - Dados de monitoramento de plantações;
   - Dados de pressão do ar;
   - Dados de luminosidade;
   - Dados de temperatura;
   - Dados de umidade relativa do ar; e
   - Dados de irrigação.

3. **Processamento industrial** - os dados a serem levados em consideração foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi feita uma busca *ad hoc* para aperfeiçoar essa classificação; e os trabalhos de Kadiyala et al. (2017), Nguyen-Hoang e Vo-Tan (2019), Forkan et al. (2019), Wang, Nixon e Boudreaux (2019) também foram usados. A seguir, são listados os dados referentes a este subdomínio da IoT:

   - Dados de pressão dos pneus;
   - Dados do motor;
   - Dados de consumo de combustível;
   - Dados de localização;

- Dados de velocidade;
- Dados de distância de outros veículos;
- Dados de tempo de condução;
- Dados de paradas;
- Dados de presença do motorista;
- Dados de produção;
- Dados do funcionamento dos equipamentos industriais;
- Dados de gerenciamento industrial;
- Dados de produção industrial; e
- Dados de sensores conectados a contêineres que transportam mercadorias perigosas.

4. **Mobilidade inteligente e turismo inteligente** - os dados a serem levados em consideração, foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi feita uma busca *ad hoc* para aperfeiçoamento da classificação; e os trabalhos de Sundharam, Fejoz e Navet (2016), Desima et al. (2017), Faria et al. (2017), Zenkert et al. (2018) também foram usados. A seguir, são listados os dados referentes a este subdomínio da IoT:

   - Dados de câmeras de monitoramento;
   - Dados de telas para consultas;
   - Dados de alto-falantes;
   - Dados de comunicações veículo a veículo (V2V) e veículo a infraestrutura (V2I);
   - Dados de localização de veículos;
   - Dados de velocidade de veículos;
   - Dados de trânsito (congestionamento);
   - Dados de estacionamento (vagas disponíveis);
   - Dados do sistema de transporte público;
   - Dados de semáforos inteligentes; e
   - Dados de multidões de cidadãos.

5. ***Smart Grid*** - dados de locais públicos (governo) são públicos, porém, por exemplo, se forem dados de uma casa inteligente ou uma empresa podem ser considerados dados privados. E os dados a serem levados em consideração foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi feita uma busca *ad hoc* para aperfeiçoar essa classificação, e os trabalhos de Shu-wen (2011), Alam, St-Hilaire e Kunz (2014), Babadi, Nouri e Khalaj (2017), Schreiber, Nadal e Valente (2019), Khan et al. (2020) foram usados na classificação. A seguir, são listados os dados referentes a este subdomínio da IoT:

    - Dados de geração de energia;
    - Dados consumo de energia (medidores inteligentes);
    - Dispositivos conectados a energia (consumo de cada dispositivo);
    - Horário em que cada dispositivo é ligado e desligado;
    - Dados de transmissão de energia;
    - Dados de armazenamento de energia; e
    - Dados da potência da energia.

6. **Casa/edifício inteligente** - os dados a serem levados em consideração foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi feita uma busca *ad hoc* para aperfeiçoar essa classificação; e os trabalhos de Salman et al. (2016), Yadav et al. (2017), Malche e Maheshwary (2017), Chang e Nam (2019), Wang et al. (2019), Ray e Bagwari (2020) foram usados como referência. A seguir, são listados os dados referentes a este subdomínio da IoT:

    - Dados de *gateways* de banda larga;
    - Dados de telefones celulares;
    - Dados de laptops;
    - Dados de PCs;
    - Dados de TVs;
    - Dados de alto-falantes;
    - Dados de eletrodomésticos;

- Dados de tomadas;
- Dados de luzes;
- Dados de persianas automatizadas;
- Dados de termostatos.
- Dados de vigilância por vídeo;
- Dados de gerenciamento de acesso,
- Dados de automação de serviços (por exemplo, iluminação, irrigação);
- Dados de hábitos dos usuários;
- Dados de sensores de som;
- Dados de temperatura da casa;
- Dados de trancas eletrônicas;
- Dados de sensores de movimento;
- Dados de localização da casa; e
- Dados de detecção de intrusão.

7. **Monitor de segurança pública e meio ambiente** - os dados a serem levados em consideração foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi feita uma busca *ad hoc* para aperfeiçoar essa classificação; e os trabalhos de Butun et al. (2016), Merege e Ueda (2018), Blasch et al. (2019), Xu et al. (2019) também são usados. A seguir, são listados os dados referentes a este subdomínio da IoT:

   - Dados coletados por câmeras fixas localizadas na cidade;
   - Dados da segurança de edifícios públicos e privados pode ser reforçada usando a tecnologia de sensores de presença, que acionará alarmes;
   - Dados de sensores dedicados e câmeras inteligentes, bem como tecnologias GPS e sem fio, fornecendo localização em tempo real;
   - Dados de rastreamento pode ser usado para formar um mapa completo do evento, prever suas tendências (por exemplo, direção e/ou velocidade de propagação do fogo, principais áreas de risco);
   - Dados de dispositivos que detectam sons de tiros;

- Dados de alertas de desastres;
- Dados de reconhecimento facial por câmeras inteligentes; e
- Dados de sensoriamento inteligente.

8. **Campus Inteligente** - os dados a serem levados em consideração foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi feita uma busca *ad hoc* para aperfeiçoar essa classificação; e os trabalhos de Alghamdi e Shetty (2016), Du, Meng e Gao (2016), Sastra e Wiharta (2016), Hossain, Das e Rashed (2019) foram usados. A seguir, são listados os dados referentes a este subdomínio da IoT:

   - Dados de estacionamento inteligente;
   - Dados de gerenciamento inteligente de cantinas;
   - Dados de sistema de monitoramento ambiental;
   - Dados de iluminação inteligente;
   - Dados de reconhecimento facial;
   - Dados de sistema de detecção de incêndio;
   - Dados de sistema de biblioteca inteligente;
   - Dados de sistema de gerenciamento de resíduos inteligente;
   - Dados de sistema de rastreamento de ônibus inteligente;
   - Dados de sistema de monitoramento de visitantes inteligentes;
   - Dados de sensores biométricos;
   - Dados de sistema de segurança inteligente;
   - Dados de sistema de gerenciamento de banheiros; e
   - Dados de controle de energia do Campus.

9. **Vida Independente** - os dados a serem levados em consideração foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi feita uma busca *ad hoc* para aperfeiçoar essa classificação; e os trabalhos de Islam et al. (2015), Laplante e Laplante (2016), Bhattacharya e Pandey (2020) também foram usados. A seguir, são listados os dados referentes a este subdomínio da IoT:

   - Dados de condição/status do idoso;

- Dados de monitoramento de funções vitais (por exemplo, temperatura, pressão arterial, frequência cardíaca, nível de colesterol);
- Dados de redes de área corporal, formadas por dispositivos vestíveis conectados entre si, permitem que os médicos continuem o monitoramento remoto do paciente fora do hospital;
- Dados de identificação de materiais e instrumentação médica; e
- Dados de camas equipadas com terminais inteligentes com tela sensível ao toque, permitindo que os pacientes acessem serviços de entretenimento.

10. **Medicina e Saúde** - os dados a serem levados em consideração foram definidos baseados no trabalho de Borgia (2014); posteriormente, foi feita uma busca *ad hoc* para aperfeiçoar essa classificação; e os trabalhos de Islam et al. (2015), Laplante e Laplante (2016), Quist-Aphetsi e Xenya (2019), Bhattacharya e Pandey (2020) foram usados como referência. A seguir, são listados os dados referentes a este subdomínio da IoT:
    - Dados de sinais fisiológicos;
    - Dados de assistentes pessoais disponíveis em telas de PC ou TV usadas para estimular as pessoas a fazerem exercícios;
    - Dados de monitoramento de sono; e
    - Dados de exercícios físicos.
11. **Indefinido** - os dados que não foram classificados em nenhum subdomínio são classificados como Indefinidos.
    - Dados indefinidos.

4.2 SMART ANONYMITY

Para mitigar problemas relacionados a privacidade de dados em IoT, esta pesquisa propõe o *Smart Anonymity*. O *Smart Anonymity* busca manter a utilidade dos dados para que análises futuras sejam realizadas e, simultaneamente, garantir a privacidade dos usuários dos dispositivos. Este é um problema que tem recebido bastante atenção nos últimos anos. Diante disso, o grande desafio da anonimização é transformar/anonimizar os dados de tal forma que a privacidade dos indivíduos seja protegida, e a utilidade dos dados também seja mantida.

O *Smart Anonymity* tem por objetivo recomendar qual o algoritmo de anonimização de dados é o mais adequado para o conjunto de dados, de acordo com suas características, usando como critérios para fazer a recomendação as principais características dos dados. Para a recomendação, é necessária ser feita uma análise dos dados, considerando a origem deles, o tipo e a heterogeneidade, para que, de acordo com essas características, seja recomendado o algoritmo de anonimização mais eficaz para proteger os dados privados dos usuários e, ao mesmo tempo, possibilitar que os dados sejam úteis para análises futuras.

A primeira classificação oriunda do mecanismo proposto é feita conforme a proveniência/origem dos dados (ambiente IoT em que foi gerado, e se são públicos ou privados). Os dados públicos não necessariamente significam que são gerados por algum órgão do governo. Eles podem ser gerados por dispositivos que pertencem a alguma empresa como, por exemplo, dados provenientes dos dispositivos instalados em um shopping ou um prédio comercial. Porém, eles também podem ser gerados por algum órgão público, como uma universidade que tenha dispositivos inteligentes, ou até mesmo por meio de sensores instalados em uma grande cidade para as mais variadas finalidades.

Já os dados privados são gerados por dispositivos que uma pessoa instalou em sua residência. Esses dispositivos podem estar em suas casas ou podem ser dispositivos que eles andam com eles no dia a dia (relógio inteligente, celular, etc.). Também podem ser de uma empresa que tenha dispositivos que monitoram seu funcionamento. Essas informações são confidenciais e sua divulgação pode colocar a empresa em risco.

Dependendo do conjunto de dados que o mecanismo irá receber como entrada, os campos que precisarem ser anonimizados podem mudar de acordo com as características do conjunto de dados, no decorrer do texto foi usado o termo *Dataset* para se referir ao termo conjunto de dados. Logo, o algoritmo de anonimização irá decidir quais dados precisam ser anonimizados para preservar a privacidade dos usuários.

O *Smart Anonymity* pode, por exemplo, receber um conjunto de dados para serem anonimizados. Esse conjunto de dados pode ter vários tipos de dados, como nome, IP, CPF, ID, idade, e localização. Porém, a localização pode ser uma informação muito útil para gerar conhecimento a partir desses dados, então pode ser usada a técnica de ofuscação sobre a localização, por exemplo, realizando a alteração de uma ou duas ruas da real localização do usuário.

Nesse caso, a empresa que tiver autorização dos usuários para vender os dados gerados pelos dispositivos pode garantir aos usuários que seus dados só serão vendidos quando forem anonimizados. Consequentemente, esses usuários não terão que se preocupar com essa anoni-

mização, pois o sistema fará isso para eles. Neste cenário, a empresa poderá lucrar de duas maneiras: oferecendo um serviço de melhor qualidade, pois a análise dos dados conduz ao melhor entendimento do gosto e das preferências do usuário em relação ao produto/serviço sem colocar a privacidade de seus clientes em risco; e ainda pode ganhar com a venda dos dados para terceiros, se esse for um mercado de interesse da empresa.

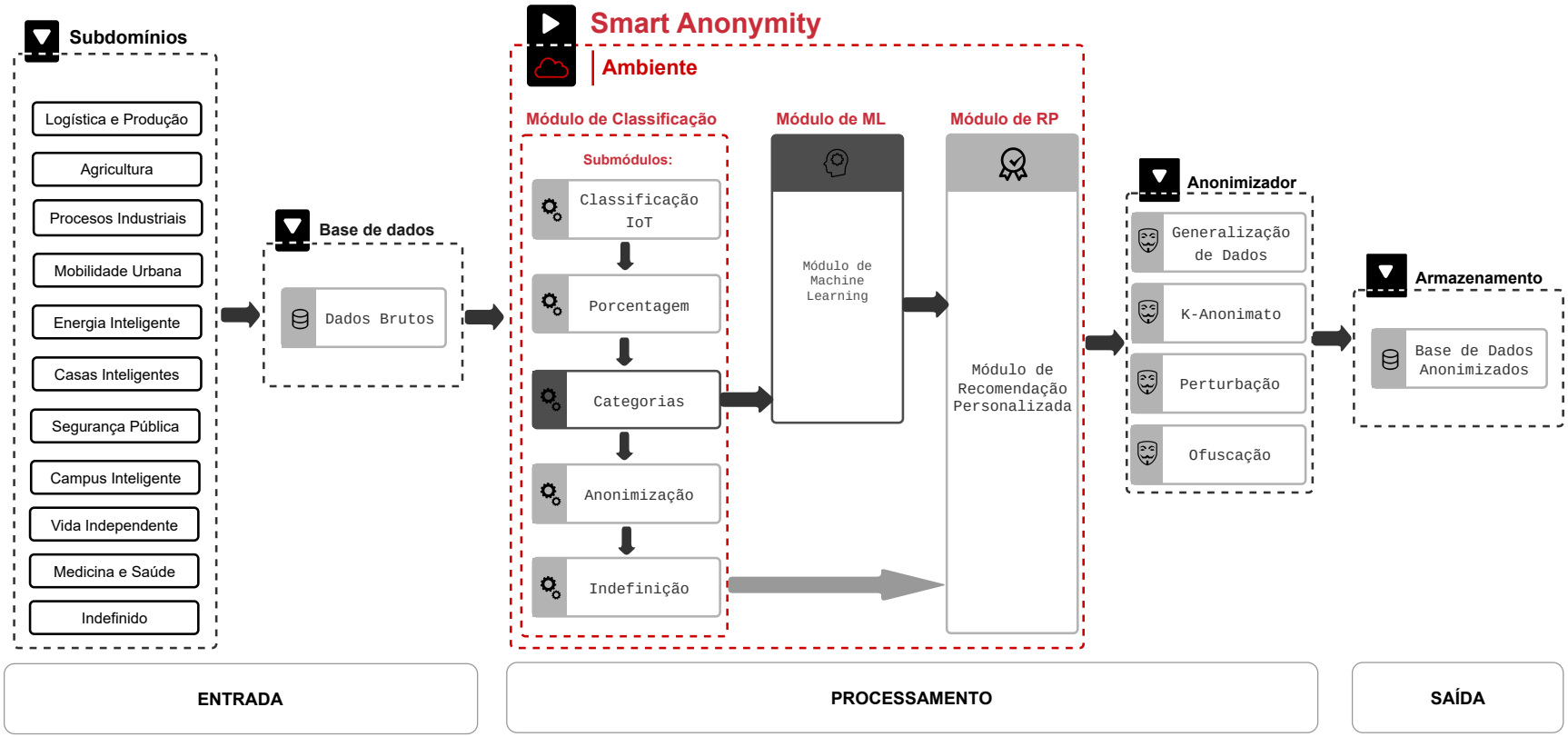
O *Smart Anonymity* irá funcionar no âmbito de aplicação, não será necessário mexer com a complexidade dos algoritmos de anonimização porque eles serão usados e o *Smart Anonymity* irá decidir qual algoritmo será usado, de acordo com as características dos dados.

Os dados que precisarem ser anonimizados são classificados como dados identificadores, e estes são os dados diretamente relacionados à identidade de um indivíduo. Contudo, é importante destacar que, após a anonimização do conjunto de dados, este conjunto de dados ainda tem que ser útil para gerar informação em análises futuras. Ainda é necessário salientar que, normalmente, os dados que precisam ser anonimizados são dados pessoais, como localização, nome, telefone, ID, etc.

Isso posto, a Figura 10 ilustra o funcionamento do mecanismo de recomendação proposto. O *Smart Anonymity* é dividido em **Módulo de Classificação**, Módulo de ***Machine Learning*** e **Módulo de Recomendação Personalizada**. Após a recomendação, os dados são enviados para o Anonimizador que aplica o algoritmo de anonimização recomendado, e então envia os dados já anonimizados para o Armazenamento, que é o local onde os dados anonimizados ficam armazenados.

O Módulo de Classificação provê vários Submódulos de classificação de subdomínios, dentre eles: O Submódulo de Classificação IoT, Submódulo de Porcentagem, Submódulo Categorias, Submódulo Anonimização e o Submódulo Indefinição. Esses submódulos são executados na ordem apresentada e apenas é utilizado o módulo seguinte, caso o anterior não tenha sido capaz de classificar uma amostra em um subdomínio específico. É importante destacar que a Machine Learning recebe como entrada a saída após o processamento do Submódulo Categorias.

Figura 10 – Arquitetura do *Smart Anonymity*.



Fonte: Autor.

Algorithm 1 Pseudocódigo da classificação feita pelo Módulo de Classificação

Input: SensorCategory, IoTSubdomains: Duas ontologias
Output: Um algoritmo de anonimização recomendado

$X_s \leftarrow$ Tamanho de X, onde X é uma variável
$X_n \leftarrow$ Nome do domínio, onde X é uma variável
$X^c \leftarrow$ Categorias de um domínio, onde X é uma variável
$X_s^c \leftarrow$ Tamanho das categorias de um domínio, onde X é uma variável
$X_n^1 \leftarrow$ Nome do primeiro elemento do conjunto X, onde X é uma variável
$D \leftarrow Database$
$Sc \leftarrow OntologiaSensorCategory$
$Is \leftarrow OntologiaIoTSubdomains$

$Domonios, Categorias \leftarrow$ submoduloClassificação(D, Sc, Is)
if $Domais_s = 1$ **then**
    DefinaAlgoritmoDeAnonimização(Dominios)
else
    $Porcentagens \leftarrow$ SubmoduloPorcentagem(Dominios, Categorias)
    if $Max(Porcentagens) > 50\%$ e $Max(Porcentagens)_s = 1$ **then**
        DefinaAlgoritmoDeAnonimização($Max(Porcentagens)_n$)
    else
        $P \leftarrow$ SubmoduloPorcentagem(Dominios, Categorias)
        if $Max(P) > 50\%$ e $Max(P)_s = 1$ **then**
            DefinaAlgoritmoDeAnonimização($Max(P)_n$)
        else
            $NC \leftarrow$ SubmoduloCategorias(Dominios, Categorias)
            if $Max(NC)_s = 1$ **then**
                DefinaAlgoritmoDeAnonimização($Max(NC)_n$)
            else
                $SA \leftarrow$ SubmoduloAnonimizacao(Dominios, Categorias)
                if SA **then**
                    DefinaAlgoritmoDeAnonimização($Categorias_n^1$)
                else
                    DefinaAlgoritmoDeAnonimização($Indefinido$)
                end if
            end if
        end if
    end if
end if

O Pseudocódigo 1 ilustra o funcionamento do Módulo de Classificação, detalhando quando cada submódulo pode ser usado para fazer a classificação. A seguir, são descritos os submódulos do Módulo de Classificação:

- **Submódulo de Classificação IoT:** este submódulo determina o subdomínio de uma amostra usando apenas as ontologias *SensorCategory* e *IoTSubdomains*, que são descritas

Algorithm 1 Pseudocódigo da classificação feita pelo Módulo de Classificação (continuação)

procedure SUBMODULOPORCENTAGEM(D, C)
$P \leftarrow$ Determine o quanto as categorias (C) raciocinadas representam um determinado domínio ($d \in D$) IoT. Ou seja, $P \leftarrow ((C \cap d_c)/d_s^c)$.
Retorne P
end procedure

procedure SUBMODULOCLASSIFICAÇÃO(D, Sc, Is)
$Categorias \leftarrow$ raciocine em D usando Sc
$Domonios \leftarrow$ raciocine em categories usando Is
Retorna Domonios, Categorias
end procedure

procedure DEFINAALGORITMODEANONIMIZAÇÃO($Dominios$)
if $Domais_n =$ Logística Processamento ou $Domais_n =$ Processamento Industrial ou $Domais_n =$ Casa/edifício Inteligente **then**
Retorne Ofuscação
else if $Domais_n =$ Agricultura e Criação ou $Domais_n =$ Medicina e Saúde ou $Domais_n =$ Vida Independente **then**
Retorne Pertubação
else if $Domais_n =$ Mobilidade Inteligente ou $Domais_n =$ Campus Inteligente **then**
Generalização de Dados
else if $Domais_n =$ Smart Grid ou $Domais_n =$ Segurança Pública ou $Domais_n =$ Indefinido **then**
K-anonimato
else
K-anonimato
end if
end procedure

procedure SUBMODULOCATEGORIAS(D, C)
$NC \leftarrow$ Determine quanto as categorias (C) (em valor absoluto) raciocinadas representam um determinado domínio ($d \in D$) IoT. Ou seja, $NC \leftarrow C \cap d_c$.
Retorne NC
end procedure

procedure SUBMODULOANONIMIZACAO(D, C)
$SA \leftarrow$ Determine se todas as categorias (C) são de um único tipo de anonimização.
Retorne SA
end procedure

em detalhes na seção 4.3.1. O *output* da *SensorCategory* é utilizado para classificação do subdomínio na Ontologia *IoTSubdomains*. Caso uma amostra seja classificada em dois subdomínios, este Submódulo não é capaz de finalizar a classificação. Dessa forma a classificação passa para o Submódulo seguinte;

- **Submódulo Porcentagem:** este submódulo determina a porcentagem que uma dada subamostra (entrada) tem em cada subdomínio IoT. Nesse caso, é considerada a quantidade de sensores que pertencem a cada subdomínio, por exemplo, determinado *Dataset* tem dados de 50 tipos de sensores, e 26 desses tipos de sensores são pertencentes ao subdomínio de agricultura, então esse *Dataset* é classificado como sendo do subdomínio de agricultura. O *software* classifica a amostra no subdomínio com a maior porcentagem. Se dois ou mais subdomínios têm a mesma porcentagem, ou se a amostra não tem pelo menos 50% dos sensores do subdomínio, o Submódulo Porcentagem não consegue finalizar a classificação. Nesse caso, a classificação para o próximo submódulo;

- **Submódulo Categorias:** este submódulo classifica uma dada subamostra baseado na quantidade de sensores por subdomínio. Para cada subdomínio, a aplicação calcula quantas categorias da amostra estão presentes nos subdomínios mapeados. O subdomínio com mais sensores é o escolhido. Se dois ou mais subdomínios têm a mesma quantidade de sensores, o Submódulo Categorias não consegue fazer a classificação. Neste caso, a classificação passa para o submódulo seguinte;

- **Submódulo anonimização:** este submódulo classifica uma dada subamostra baseado no conjunto de subdomínios classificados nos submódulo anteriores. A aplicação analisa os subdomínios e avalia se os tipos de categorias que eles têm pertencem a uma mesma classe de anonimização. Caso positivo, o algoritmo não é classificado em um subdomínio específico, mas recebe um algoritmo de anonimização eficaz. Caso os subdomínios não façam parte do mesmo algoritmo de anonimização, a classificação passa para o próximo submódulo; e

- **Submódulo Indefinição:** caso nenhuma submódulo consiga classificar a amostra em um subdomínio/algoritmo de anonimização, foi optado por indicar o algoritmo de anonimização K-Anonimato porque ele é genérico e, segundo uma revisão feita nesta pesquisa, ele é o algoritmo que se aplica em mais campos da IoT.

Conforme a Arquitetura do *Smart Anonymity* apresentada na Figura 10, a classificação pode ser feita exclusivamente usando as bases de conhecimento das ontologias, ou pode ser feitas com o auxílio do Módulo de *Machine Learning*.

A Figura 11 representa o cenário de execução do *Smart Anonymity* usando apenas as bases de conhecimento das ontologias. Nesse cenário, a classificação é feita usando os submódulos

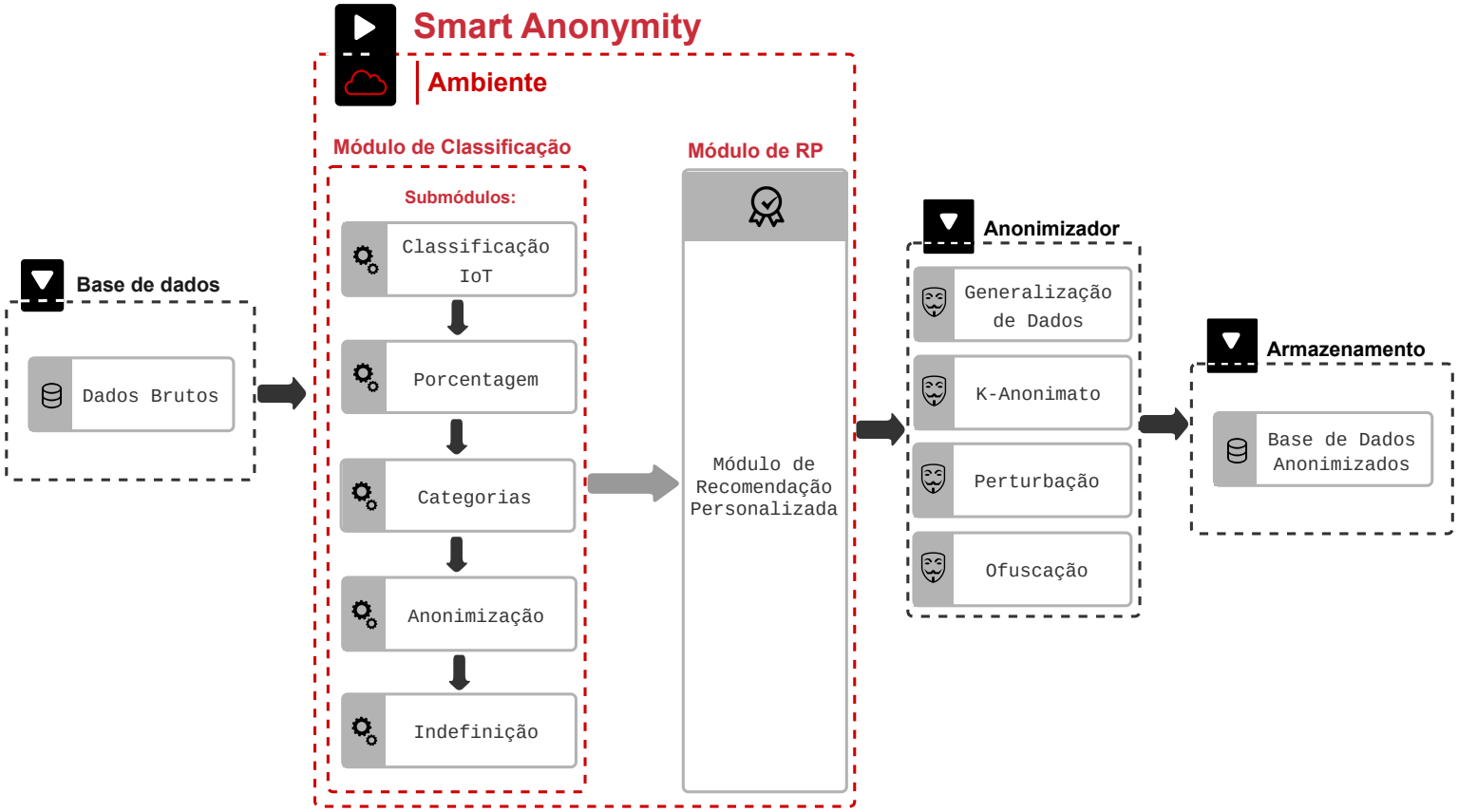
contidos dentro do **Módulo de Classificação**, descritos acima. Inicialmente, o **Submódulo de Classificação IoT** recebe o *Dataset* e tenta fazer a classificação. Caso a classificação não seja definida pelo **Submódulo de Classificação IoT**, o **Submódulo Porcentagem** analisa o *Dataset* para fazer a classificação. Caso o **Submódulo Porcentagem** não consiga finalizar a classificação, o **Submódulo Categorias** recebe o *Dataset* para fazer a classificação. Se a classificação ainda não for finalizada no **Submódulo Categorias**, o **Submódulo anonimização** entra em ação para fazer a classificação. Por fim, se nenhum dos submódulos forem capazes de fazer a classificação, o *Dataset* é repassado para o **Submódulo Indefinição** que irá classificar o *Dataset* como indefinido. Após a classificação, o *Dataset* é analisado pelo **Módulo de Recomendação Personalizada** para que seja recomendado o algoritmo de anonimização para o *Dataset* em questão.

Já a Figura 12 representa a execução do *Smart Anonymity* usando *Machine Learning* para melhorar a classificação feita pelas ontologias, ver Seção 4.3.2. Nesse cenário, os dados contidos no *Dataset* são classificados em categorias pelo **Submódulo Categorias** que faz parte do **Módulo de Classificação**. Após a classificação inicial feita pelo **Submódulo Categorias**, o *Dataset* que está sendo analisado é repassado para que o **Módulo de Machine Learning** juntamente com as informações das categorias que ele contém. Com o *Dataset* e as informações das categorias, o **Módulo de Machine Learning** consegue finalizar a classificação e repassar os dados para o **Módulo de Recomendação Personalizada**, para que então seja feita a recomendação do algoritmo de anonimização para o *Dataset* que está sendo analisado.

4.3 FLUXO DE EXECUÇÃO DO *SMART ANONYMITY*

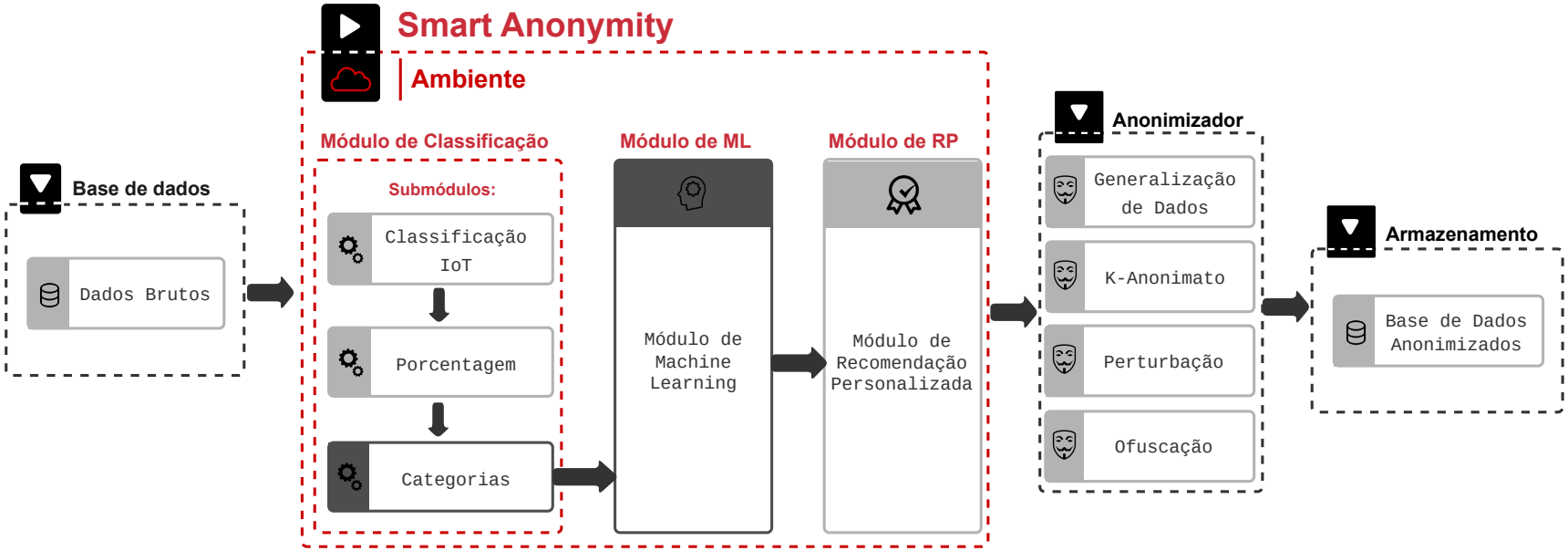
Os dados são gerados pelos dispositivos IoT que podem ser de 11 subdomínios distintos, dispostos na Figura 10. Esses dados normalmente são enviados para serem armazenados em alguma plataforma na nuvem (por exemplo: *Amazon Web Services* (AWS), *Microsoft Azure* e *FIWARE*), ou pode ser armazenado localmente no servidor da empresa proprietária dos dados, representada na Figura 10 como "Dados Brutos", local onde os dados são armazenados depois que são gerados. Essa base de dados pode ser atualizada constantemente pelos dispositivos ou podem ser dados já armazenados por um determinado tempo.

Figura 11 – Fluxo de execução do *Smart Anonymity* usando apenas as bases de conhecimento das ontologias.



Fonte: Autor.

Figura 12 – Fluxo de execução do *Smart Anonymity* usando ML.



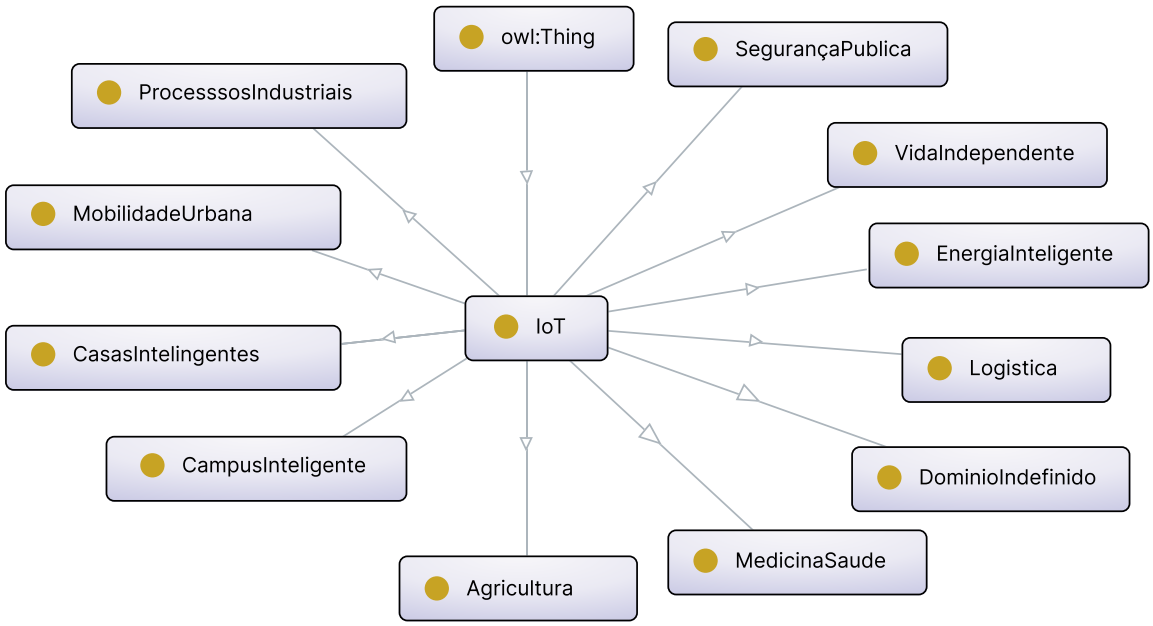
Fonte: Autor.

O *Smart Anonymity* está dividido em três módulos. São eles: (i) **Módulo de Classificação**, responsável por fazer a classificação dos dados de acordo com suas características; (ii) **Módulo de Machine Learning**, responsável por melhorar a classificação feita pelo Módulo de Classificação; e (iii) **Módulo de Recomendação Personalizada**, responsável por recomenda o algoritmo de anonimização mais adequado.

4.3.1 Módulo de Classificação

O Módulo de Classificação recebe um conjunto de dados, e faz a classificação do conjunto de dados levando em consideração os 11 subdomínios e as características dos dados que cada subdomínio tem, ver Figura 13. Essa classificação é feita por meio de duas ontologias OWL que usam formalismo baseado em lógica de descrição (DL) para analisar os dados e classificá-los de acordo com suas características.

Figura 13 – Representação ontologica para classificação de dados.



Fonte: Autor.

A primeira Ontologia nomeada de *sensorCategory* classifica os dados contidos em uma base de dados considerando 366 tipos de sensores [1] que posteriormente podem ser classificados como pertencentes a uma das 101 categorias mapeadas nesta pesquisa. O mapeamento das

[1] https://en.wikipedia.org/wiki/List_of_sensors

categorias foi realizado baseado nas classificações feitas no trabalho de Borgia (2014). A segunda Ontologia nomeada de *IoTSubdomains* classifica as categorias mapeadas na base de dados em um dos 11 subdomínios da IoT mostrados na Figura 13.

Ontologia 01: *sensorCategory* é responsável por classificar os sensores em categorias. As bases de dados são consumidas por essa Ontologia e, de acordo com as características dos sensores, eles são classificados em categorias. É importante destacar que nessa Ontologia, as categorias não são disjuntas porque têm elementos (Sensores) que podem pertencer a mais de uma categoria. A seguir, são listadas as restrições implementadas nessa Ontologia para que a classificação seja feita. Todas as restrições descritas nesta pesquisa são implementadas na base de conhecimento da Ontologia e estão consequentemente implementadas na aplicação desenvolvida para a avaliação desta pesquisa.

Restrições para os sensores:

- Todos os sensores são disjuntos (distintos);
- Cada sensor deve pertencer a pelo menos uma Categoria; e
- Sensores que não são classificados em nenhuma categoria, são pertencentes a categoria "DataUndefined".

Restrições para as Categorias:

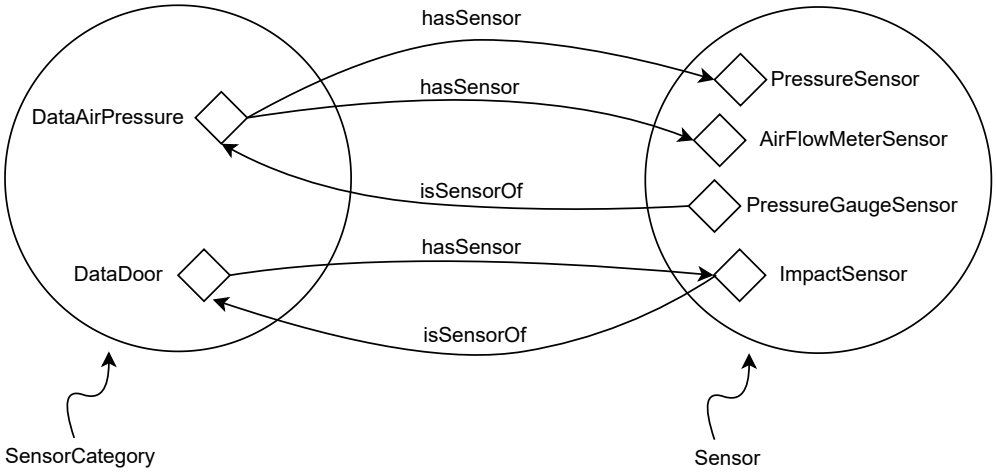
- Cada Categoria deve pertencer a pelo menos um subdomínio;
- Cada Categoria deve ter pelo menos um sensor; e
- Categorias que não são classificadas em nenhum subdomínio são pertencentes ao domínio *"Undefined"*.

A ontologia sensorCategory tem duas propriedades:

- **hasSensor:** onde a expressão *'SensorCategory hasSensor some Sensor'* significa que um indivíduo de SensorCategory tem algum indivíduo Sensor; e
- **isSensorOf:** já a expressão *'Sensor isSensorOf some SensorCategory'* significa que um indivíduo Sensor é de algum indivíduo SensorCategory.

A Figura 14 ilustra o funcionamento das propriedades da Ontologia *SensorCategory*. Em OWL, as classes são conjuntos que contêm os indivíduos. No exemplo ilustrado na Figura 14, a Ontologia *SensorCategory* tem a classe *SensorCategory* e a classe *Sensor*, onde a expressão *'DataAirPressure hasSensor some PressureSensor'* significa que um indivíduo *DataAirPressure* tem algum indivíduo *PressureSensor*. Outro tipo de expressão é *'PressureGaugeSensor isSensorOf some DataAirPressure'*. Nesse caso, essa expressão significa que o indivíduo *PressureGaugeSensor* é de algum indivíduo *DataAirPressure*.

Figura 14 – Representação do uso das propriedades na ontologia SensorCategory.



Fonte: Autor.

Ontologia 02: *ioTSubdomains* responsável por usar a classificação das categorias para identificar a qual subdomínio a base de dados analisada pertence.

Restrições para os subdomínios:

- Todos os subdomínios são disjuntos;
- Cada subdomínio deve ter pelo menos uma categoria; e
- subdomínio *"Undefined"* tem apenas categoria *"DataUndefined"*.

Restrições para as Categorias:

- Todas as Categorias são disjuntas;
- Cada Categoria deve pertencer a pelo menos um subdomínio;
- Cada Categoria deve ter pelo menos um sensor; e

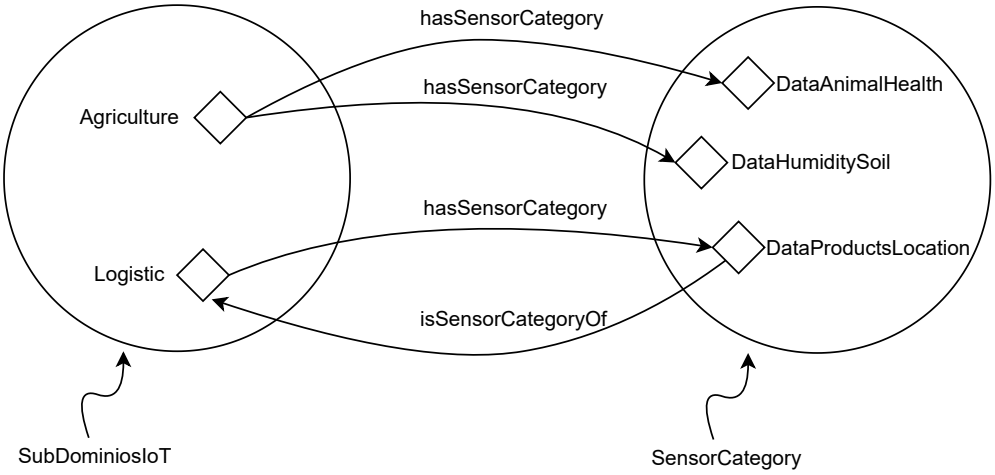
- Categorias que não são classificadas em nenhum subdomínio são pertencentes ao subdomínio *"Undefined"*.

A ontologia SubDomainIoT tem duas propriedades:

- ***hasSensorCategory***: onde a expressão *'SubDomainIoT hasSensorCategory some SensorCategory'* significa que um indivíduo de SubDomainIoT tem algum indivíduo SensorCategory; e
- ***isSensorCategoryOf***: já a expressão *'SensorCategory isSensorCategoryOf some SubDomainIoT'* significa que um indivíduo de *SensorCategory* é de algum indivíduo *SubDomainIoT*.

A Figura 15 ilustra o funcionamento das propriedades da Ontologia *ioTSubdomains*. No exemplo ilustrado na Figura 15, a Ontologia *ioTSubdomains* tem a classe *SubDominiosIoT* e a classe *SensorCategory*, onde a expressão *'Agriculture hasSensorCategory some DataAnimalHealth'* significa que um indivíduo *Agriculture* tem algum indivíduo *DataAnimalHealth*. Outro tipo de expressão é *'DataProductsLocation isSensorCategoryOf some Logistic'*. Nesse caso essa expressão significa que o indivíduo *DataProductsLocation* é de algum indivíduo *Logistic*.

Figura 15 – Representação do uso das propriedades na ontologia ioTSubdomains.



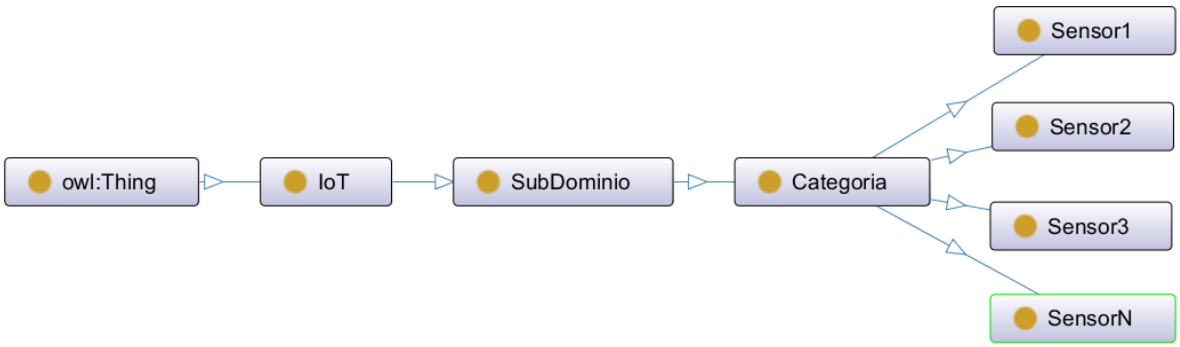
Fonte: Autor.

Ontologias baseadas em DL fornecem um vocabulário para representação do conhecimento. Esse vocabulário tem por trás uma conceitualização que o sustenta, evitando interpretações ambíguas, ainda auxilia a elucidação de ambiguidades de compreensão existentes no texto. Portanto, usando ontologias, pode-se anotar informações semânticas em artefatos de informação

não estruturadas visando, assim, à obtenção de resultados mais precisos em pesquisas de informação (GRUBER, 1993; GUARINO; OBERLE; STAAB, 2009). Um dos principais motivos para se usar a ontologia é a verificação de inconsistência e descoberta de novos fatos para validar dados.

A Figura 13 apresenta uma ilustração do Módulo de Classificação da ontologia, onde a classe *Thing* é a superclasse de todas as outras classes, e a classe IoT contém todas as classes que representam os subdomínios da IoT. Os subdomínios são divididos em 101 categorias listadas na seção 4.1 e as categorias são divididas em 366 tipos de sensores. A Figura 16 ilustra o cenário do fluxo de execução descrito anteriormente.

Figura 16 – Representação do Fluxo de classificação dos dados.



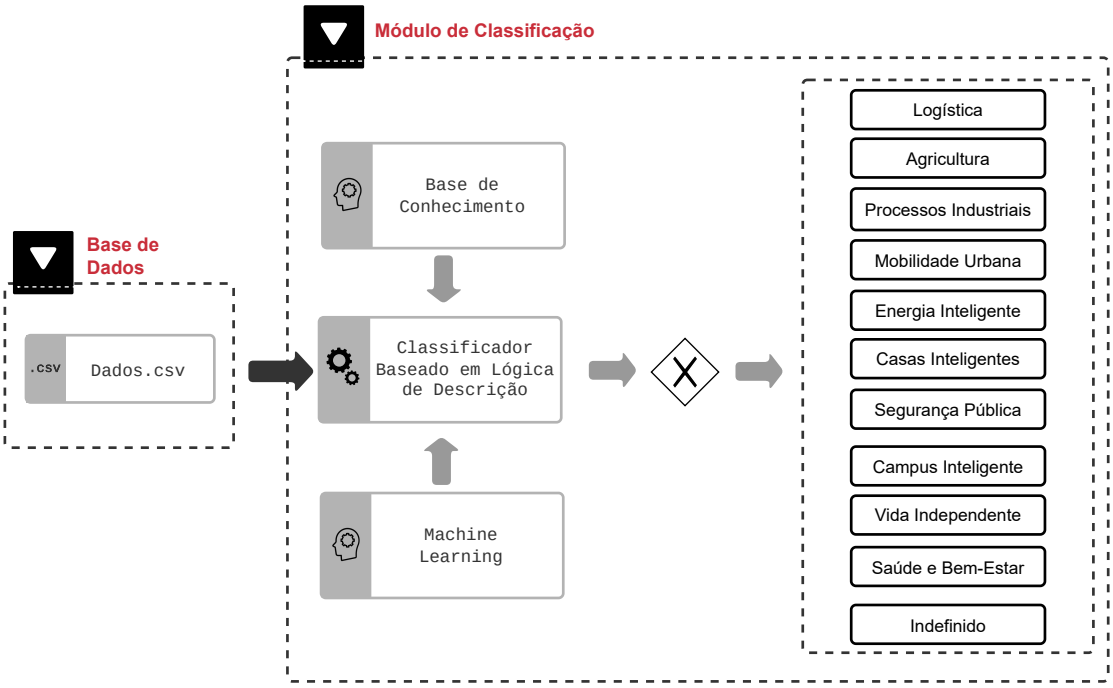
Fonte: Autor.

Segundo Costa (2020), o raciocinador é responsável por deduzir novos fatos em uma base de conhecimento, que é umas das mais importantes tarefas de Inteligência Artificia (IA). O raciocinador é responsável por realizar inferências lógicas no código OWL. Os serviços de raciocínio em DL são fornecidos por raciocinadores, como *HermiT* (GLIMM et al., 2014), *FaCT ++* (TSARKOV; HORROCKS, 2006), *Pellet* (SIRIN et al., 2007), *RacerPro* (HAARSLEV; MÖLLER, 2001), *CEL* (BAADER; LUTZ; SUNTISRIVARAPORN, 2006), e *RACCOOM* (FILHO; FREITAS; OTTEN, 2017). Nesta pesquisa é utizado o *Pellet*, que é um raciocinador OWL compatível para verificação de inconsistência. Além disso, inclui otimizações para melhorar o desempenho do raciocínio e que, de acordo com testes realizados no decorrer da pesquisa, obteve melhores resultados.

Além da classificação por meio das ontologias, é usada *Machine Learning* para ajudar a melhorar a classificação das bases de dados. A Figura 17 ilustra o funcionamento da classificação, a base de conhecimento das ontologias é usada para classificar um determinado subdomínio e, para melhorar os resultados da classificação, é utilizado *Machine Learning* em conjunto com as

ontologias. Os sensores presentes em um determinado *Dataset* podem ser classificados em uma ou mais categorias (101). A partir da classificação em categorias, o *Machine Learning* ajuda a melhorar os resultados da classificação, posteriormente, as categorias podem ser classificadas e um determinado subdomínio (11).

Figura 17 – Arquitetura do Módulo de Classificação.



Fonte: Autor.

A atualização da base de conhecimento da ontologia do Modelo de Classificação dar-se-á da seguinte maneira: podem ser atualizados os axiomas existentes, ou adicionados novos axiomas, para que sejam adicionadas novas características a um determinado subdomínio da IoT, para que, então, a classificação seja realizada. Também podem ser atualizadas as propriedades existentes, as classes e os indivíduos. Porém, com a atualização da base de conhecimento de uma ontologia, podem acontecer inconsistências como, por exemplo, a adição de um novo axioma pode causar alguma inconsistência na ontologia, dessa maneira, esse novo axioma pode entrar em conflito com outros já existentes. Para que a consistência da ontologia seja mantida, podem ser utilizadas técnicas de revisão de crenças (GÄRDENFORS, 1992).

4.3.2 Módulo de Machine Learning

Para melhorar os resultados da classificação feita pelas ontologias, foi desenvolvido o Módulo de *Machine Learning*. Este módulo recebe o *output* do Submódulo Categorias para fazer a

classificação utilizando *Machine Learning*. Para que este módulo possa funcionar, é necessário que as ontologias façam a classificação inicial dos sensores em categorias. Essa classificação inicial é feita utilizando a base de conhecimento das ontologias. Este módulo utiliza *Machine Learning* para finalizar a classificação iniciada pelo Módulo de classificação. Os *Datasets* já pré-classificados em categorias são processados pelos algoritmos de *Machine Learning* que, ao final do processo, identificam a qual subdomínio cada *Dataset* analisado pertence. Esse processo é ilustrado pela Figura 12.

Neste módulo foram avaliados três algoritmos descritos na seção 2.5. Para analisar se os algoritmos ajudam a melhorar os resultados da classificação em relação às ontologias, foi feita uma avaliação comparando os resultados da classificação feita exclusivamente usando as antologias e a classificação feita por três algoritmos de *Machine Learning*. Também é feita uma comparação entre os três algoritmos de *Machine Learning*. Esta avaliação é melhor descrita na seção 5.1.3.2.

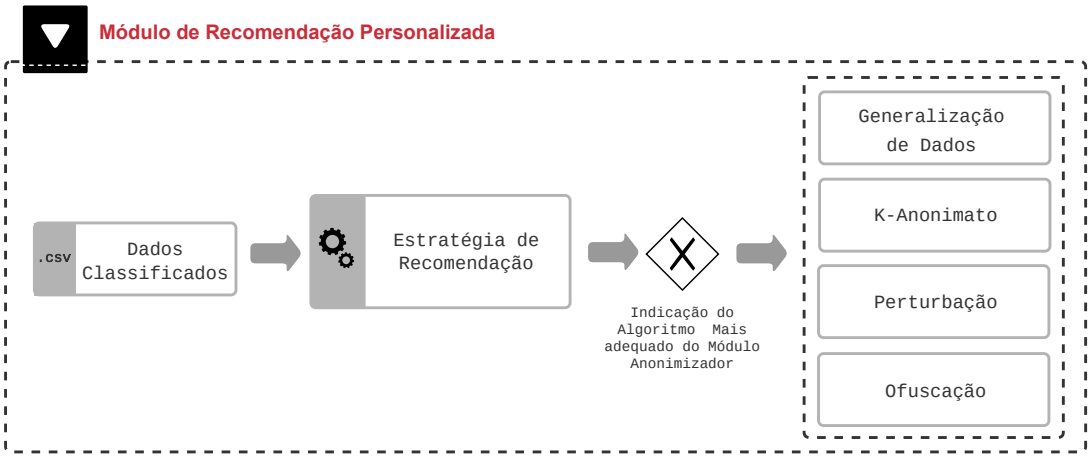
4.3.3 Módulo de Recomendação Personalizada

O Módulo de Recomendação Personalizada recomenda o algoritmo de anonimização mais adequado para determinado conjunto de dados, de acordo com a classificação feita pelo Módulo de Classificação. Basicamente o Módulo de Recomendação Personalizada recebe como entrada um conjunto de dados (Dataset) já classificado como pertencente a um dos subdomínios da IoT. De acordo com o subdomínio em que ele foi classificado e suas características, é recomendado o algoritmo de anonimização mais adequado para este conjunto de dados em questão. A Figura 18 ilustra o processo de recomendação, feito por meio do uso do algoritmo de IA *Random Forest*.

A seguir, na Tabela 11, são listados os algoritmos de anonimização que podem ser usados, juntamente com a definição de quais algoritmos usar em cada subdomínio. Essa classificação inicial foi elaborada baseada na Revisão Sistemática da Literatura descrita no capítulo 3. Nesta revisão foi feito um levantamento dos principais algoritmos de anonimização, suas principais características e seus pontos fortes e fracos.

Conforme os resultados obtidos por meio da Revisão Sistemática da Literatura conduzida por Neves et al. (2023), que também está descrita no capítulo 3, a Generalização de Dados se aplica melhor aos dados pertencentes aos subdomínios Mobilidade inteligente e Campus inteligente. O K-anonimato, por sua vez, é aplicado em dados pertencentes ao subdomínio

Figura 18 – Arquitetura do módulo de recomendação.



Fonte: Autor.

Smart Grid e Segurança pública. A Perturbação é utilizada em três subdomínios, Agricultura e Criação, Vida Independente e Medicina e Saúde e, por fim, a técnica de Ofuscação é aplicada em dados pertencentes aos subdomínios de Logística, Processamento industrial e Casa/edifício inteligente. O Módulo de Recomendação Personalizada foi desenvolvido seguindo essa estrutura, para que a recomendação do algoritmo de anonimização seja feita.

O algoritmo *Random Forest* foi escolhido para fazer a recomendação, pois ele permite tornar a aplicação mais robusta, adicionando mais subdomínios e mais algoritmos de anonimização. Isso permite que o *Smart Anonymity* seja expandido e possar ser usado em situações mais complexas, com mais tipos de dados.

Tabela 11 – Aplicação dos algoritmos de anonimização.

Subdomínio	**Generalização de Dados**	**k-anonimato**	**Perturbação**	**Ofuscação**
Logística	Não	Não	Não	Sim
Agricultura e Criação	Não	Não	Sim	Não
Processamento Industrial	Não	Não	Não	Sim
Mobilidade Inteligente	Sim	Não	Não	Não
Smart Grid	Não	Sim	Não	Não
Casa/edifício Inteligente	Não	Não	Não	Sim
Segurança Pública	Não	Sim	Não	Não
Campus Inteligente	Sim	Não	Não	Não
Vida Independente	Não	Não	Sim	Não
Medicina e Saúde	Não	Não	Sim	Não

Fonte: O Autor.

4.3.4 Anonimizador de Dados

O Anonimizador de Dados é composto pelos algoritmos de anonimização, para que o Módulo de Recomendação Personalizada possa indicar qual algoritmo é o mais adequado para o conjunto de dados em questão. Então o algoritmo de anonimização escolhido é aplicado ao conjunto de dados em questão, para que ele seja anonimizado e enviado para ser armazenado. Um ponto importante a ser destacado é que cada conjunto de dados é submetido a apenas um algoritmo de anonimização.

Para a anonimização, é utilizada a ferramenta ARX[2], que é um software de código aberto para a anonimização de dados pessoais sensíveis. O ARX suporta uma grande variedade de (1) modelos de privacidade e risco, (2) métodos de transformação de dados e (3) métodos de análise da utilidade dos dados de saída.

A ferramenta ARX é capaz de lidar com grandes conjuntos de dados. O ARX utiliza um algoritmo de pesquisa globalmente otimizado e bastante eficiente para que os dados sejam transformados com a generalização e supressão (PRASSER et al., 2020). A ferramenta ARX é capaz de suportar o uso de milhões de registros, oferecendo uma interface gráfica para o usuário, tutoriais de ajuda e visualizações que orientam os usuários em diferentes aspectos durante o processo de anonimização (SILVA et al., 2019). Além disso, a ferramenta ARX suporta os principais algoritmos de anonimização, incluindo os presentes na Figura 18.

4.3.5 Armazenamento de Dados

O Armazenamento de dados é responsável por manter os dados após anonimizados. Esses dados anonimizados ficam armazenados para que possam ser consumidos por alguma aplicação ou usados em análises futuras. Fica a critério da organização que usar o *Smart Anonymity* escolher o local mais apropriado para salvar os dados após o procedimento de anonimização. Esses dados podem ser armazenados em bancos de dados em diversas plataformas, como AWS, Microsoft Azure e a plataforma FIWARE, dentre outras. A escolha do tipo de Banco de dados que será usado para armazenamento dos dados, fica a critério da empresa proprietária dos dados.

[2] https://arx.deidentifier.org/

4.4 IMPLEMENTAÇÃO DA ABORDAGEM PROPOSTA

Na primeira etapa de desenvolvimento do *Smart Anonymity*, foram desenvolvidas duas ontologias OWL (sensorCategory e ioTSubdomains), disponíveis no repositório do projeto no *GitHub* em: <ontologies>. Elas são responsáveis por fazer a classificação das bases de dados analisadas utilizando as propriedades e axiomas descritos na seção 4.3.1. Para criação das ontologias, foi utilizada a ferramenta *Protegé* 5.2[3]. Nas ontologias, foram especificadas as definições de todos os 366 sensores e as 101 categorias que cada sensor faz parte, bem como os 11 subdomínios de que cada categoria faz parte.

Em um segundo momento, foi desenvolvido o *Smart Anonymity*. Para essa parte do desenvolvimento da aplicação, foi utilizada a linguagem de programação Java, versão 17. A aplicação lê os dados e os processa usando as bases de conhecimento das ontologias para fazer a classificação. Para o desenvolvimento do *Smart Anonymity*, foram usadas várias outras tecnologias, dentre elas o *Openllet Jena* versão 4.5.0[4], que é um raciocinador *OWL* DL de código aberto para Java. Ele foi usado no carregamento, processamento e execução das ontologias. Para a leitura dos arquivos em formato 'CSV' foi utilizado Opencsv versão 5.6, o código-fonte da aplicação em Java está disponível em: <SmartAnonymity>.

Devido a características únicas pertencentes ao domínio abordado nesta pesquisa, algumas regras foram implementadas diretamente na aplicação desenvolvida para testar a solução proposta. A seguir, são listadas as regras implementadas na aplicação.

- Maior percentual de Sensores pertencentes a uma categoria define o *Dataset* como pertencente a esta categoria;
- O *DataSet* que tem maior percentual de categorias que pertencem a um determinado subdomínio, é classificado como sendo desse mesmo subdomínio; e
- Quando um *Dataset* tem todos os sensores pertencentes a uma determinada categoria e mais alguns pertencentes a outra, se essa outra tiver todos os sensores do *Dataset*, esse *Dataset* vai ser classificado sendo dessa categoria (a categoria que tiver o maior número de sensores dentro de um *DataSet*, é usada para classificá-lo). a. EX: *DataSet* "A" tem sensores (A,B,C,D,E,F,G), a Categoria "X" tem sensores (A,B,C,F) e a Categoria "Y" tem sensores (A,B,C,E,F,G), o *Dataset* "A" vai ser classificado sendo da categoria "Y".

[3] https://protege.stanford.edu/
[4] https://github.com/Galigator/openllet

Para melhorar a classificação da aplicação, foram implementados 3 algoritmos de *Machine Learning*, KNN (*K-NEAREST NEIGHBORS ALGORITHM*), SVM (*SUPPORT VECTOR MACHINE*) e o RF (*RANDOM FOREST CLASSIFIER*). Esses algoritmos recebem os *Datasets* já rotulados em categorias para finalizar a classificação. Essa parte da aplicação foi desenvolvida usando a linguagem de programação Python versão 3, foi usada a biblioteca Sklearn[5]. Esta parte da aplicação está disponível em: <Machine Learning>.

Para o desenvolvimento do Módulo de Recomendação Personalizada foi usada a linguagem de programação Python versão 3. Também foi utilizada a biblioteca Sklearn[6] para treinamento do *Random Forest Classifier* (RF), responsável por fazer a recomendação dos algoritmos de anonimização. Também foi utilizado o *Flask*[7] como servidor Web para o Módulo de Recomendação Personalizada. Sua escolha foi devido a ele ser uma das bibliotecas mais usadas no Python, e ele também fornece configuração e convenções padrão.

O Fluxo de execução da aplicação funciona da seguinte maneira: inicialmente, o *software* recebe como *input* um conjunto de dados (Dataset) referente aos sensores IoT. Esses sensores são classificados em N categorias usando a ontologia *SensorCategory*. Em seguida, essas categorias são classificadas em um subdomínio.

Quando são adicionados sensores ainda não mapeados na ontologia, a classificação deixa de ser feita por percentagem e passa a ser feita por categorias. Isso acontece porque existem vários sensores nessa lista. Esses vários sensores classificam todas as categorias como 100%, ou seja, existem diversas classes com a mesma probabilidade. Logo, não há como porcentagem resolver e isso fica em encargo da maior quantidade de categorias.

4.5 CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentada a principal contribuição desta Tese, que é o *Smart Anonymity*, um mecanismo de recomendação de algoritmos de anonimização.

O *Smart Anonymity* propõe uma solução para recomendação de algoritmos de anonimização, de acordo com as características dos dados. Essa solução visa possibilitar que os dados gerados por dispositivos IoT possam ser tratados/anonimizados para que a privacidade das pessoas cujos dados foram capturados seja preservada. Para a preservação da privacidade dos usuários dos dispositivos IoT, são utilizadas duas ontologias para classificar esses dados de acordo com

[5] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[7] https://flask.palletsprojects.com/en/2.3.x/

suas características. Foi utilizada *Smart Anonymity* para melhorar os resultados da classificação retornados pela Ontologia. Após a classificação, é feita a recomendação do algoritmo de anonimização mais adequado para os dados.

Para validação do *Smart Anonymity*, foi implementada uma aplicação usando a linguagem de programação *Java*. A parte de *Machine Learning* foi implementada usando a linguagem de programação *Python*.

Com a anonimização promovida pelo *Smart Anonymity*, as empresas podem usar os dados gerados pelos dispositivos IoT sem colocar em risco a privacidade dos seus usuários. As empresas podem usar os dados anonimizados para diversos fins, os dados podem ser analisados para extração de alguma informação relevante para a empresa, os dados podem ser armazenados para análises futuras e, se alguma pessoa ou entidade não autorizada tiver acesso aos dados, a privacidade dos usuários não será colocada em risco, pois os dados estão anonimizados. Outra possibilidade para uso dos dados está ligada a mais uma fonte de receita para as empresas, que seria a venda desses dados para outras empresas, isso se for do interesse da empresa ingressar nesse mercado. Nesse caso seria necessário a autorização do usuário para a venda dos dados.

5 AVALIAÇÃO E RESULTADOS

Neste capítulo, é apresentada a avaliação do mecanismo de recomendação proposto e desenvolvido nesta Tese (*Smart Anonymity*), descrito em detalhes no Capítulo 4. São descritos os cenários a serem avaliados, bem como as métricas usadas na avaliação. Para estruturação da avaliação, foram adotadas as diretrizes baseadas nas recomendações de Jain (1990). As diretrizes adotadas foram: i) definir o objetivo da avaliação; ii) estabelecer o cenário de avaliação (Domínios da IoT); iii) selecionar métricas; iv) projetar a avaliação; v) analisar os resultados; e vi) apresentar os resultados.

Para avaliar a eficácia do *Smart Anonymity*, foi executado um estudo para avaliar o desempenho da classificação e da recomendação. Posteriormente, foi conduzida uma prova de conceito para validar o uso do *Smart Anonymity*. Para essa prova de conceito, foram criados três ambientes de uso da IoT na plataforma FIWARE.

A avaliação de desempenho foi feita usando *Datasets* sintéticos para avaliar a classificação e recomendação feitas pelo *Smart Anonymity.* Na prova de conceito, são criados três cenários de uso da IoT usando a plataforma *FIWARE*, os dados gerados são usados como *input* para que o *Smart Anonymity* possa classificar os dados e recomendar o algoritmo de anonimização mais adequado de acordo com as características dos dados.

5.1 OBJETIVO E TÉCNICA DE AVALIAÇÃO

Esta avaliação tem por objetivo avaliar a eficácia do *Smart Anonymity* no que diz respeito à classificação dos dados recebidos e à recomendação personalizada, considerando as características dos dados. Conforme descrito no Capítulo 4, o *Smart Anonymity* é dividido em módulos, e cada módulo tem suas funcionalidades específicas.

Os cenários de experimentação foram definidos da seguinte forma: i) para a avaliação de desempenho, inicialmente foi aplicada toda a carga de trabalho (121 *Datasets*) para avaliar a classificação da aplicação, utilizando apenas a base de conhecimento das ontologias; ii) Em seguida, procurou-se comparar a eficiência da classificação da aplicação usando algoritmos de *Machine Learning*. Para tanto, foram utilizados 61 *Datasets* para o treinamento dos algoritmos de *Machine Learning*. Os outros 60 *Datasets* foram aplicados para coletar as métricas nos seguintes cenários:

- **Cenário 1:** classificação base (usando somente a base de conhecimento das ontologias);
- **Cenário 2:** classificação base complementada por um modelo do algoritmo KNN;
- **Cenário 3:** classificação base complementada por um modelo do algoritmo SVM; e
- **Cenário 4:** classificação base complementada por um modelo do algoritmo *Random Forest*.

Para essa avaliação, foi usada uma máquina com sistema operacional Windows 10 (64 bits), processador Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz, 8 Gigabytes de memória RAM. Os algoritmos de *Machine Learning* foram implementados usando a IDE *Visual Studio Code*. Para executar os algoritmos, foram usadas as seguintes bibliotecas:

- Pandas;
- *Scikit Learn;*
- numpy;
- seaborn; e
- matplotlib

5.1.1 Métricas de Avaliação

Nesta avaliação, as métricas usadas foram a Acurácia (formula 5.1), a Precisão (formula 5.2), o *Recall* (formula5.3) e o *F1-Score* (formula 5.4). Essas métricas foram escolhidas por serem métricas comumente adotadas na avaliação de Sistemas de recomendação (ESPÍNDOLA; EBECKEN, 2005) e sistemas de inteligência artificial (MARTINS, 2016). Essas métricas serão descritas a seguir. Contudo, primeiramente é preciso introduzir os seguintes conceitos:

- TP são valores verdadeiros positivos (*True Positives*): ocorre quando valores são classificados como positivos, e que realmente são positivos;
- FP são valores falso positivo (*False Positive*): ocorre quando valores são classificados como positivos, mas que realmente são negativos;
- TN são valores verdadeiro negativo (*True Negative*): ocorre quando valores são classificados como negativos, e que realmente são negativos; e

- FN são valores falso negativo (*False Negative*): ocorre quando valores são classificados como negativos, mas que realmente são positivos.

Acurácia: é a proporção entre o número de previsões corretas e o número total de amostras de entrada:

$$Acuracia = \frac{TP + TN}{TP + FP + FN + TN} \tag{5.1}$$

Precisão: é a proporção de exemplos positivos previstos que realmente são positivos:

$$Precisão = \frac{TP}{TP + FP} \tag{5.2}$$

***Recall*:** também chamada de taxa de acerto ou *Recall*, mede o quanto um classificador pode reconhecer exemplos positivos:

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

***F1-Score*:** O F1-score combina a precisão e *Recall* de um classificador em uma única métrica, tomando sua média harmônica. É usado principalmente para comparar o desempenho de dois classificadores. Suponha que o classificador A tenha um *Recall* maior e o classificador B tenha uma precisão maior. Nesse caso, as pontuações F1 para ambos os classificadores podem ser usadas para determinar qual deles produz melhores resultados:

$$F1 - Score = \frac{2(Precisão * Recall)}{Precisão + Recall} \tag{5.4}$$

5.1.2 Testes A/B

Para comparar o desempenho de cada uma das avaliações, e identificar qual delas tem o melhor desempenho, será usado o conceito de teste A/B (ANDERSON, 2015). Para Deng e Shi (2016), testes A/B são experimentos em que são criadas duas versões de uma mesma peça/sistema (anúncio, *e-mail*, *Landing Page*, etc.) para descobrir qual delas possui maior taxa de conversão, ajudando a otimizar as estratégias de *Marketing Digital*. O autor Deng e Shi (2016) descreve teste A/B como: "teste A/B é o termo usado para experimentar aleatoriamente

uma variável de controle (A) e uma variável de experimento (B) com o propósito de testar estatisticamente uma hipótese".

Já Kompella (2015) define teste A/B como: "É um processo no qual você escolhe a versão de melhor desempenho de uma página da *Web*, exibindo aleatoriamente diferentes versões de seu site aos visitantes e avaliando o desempenho de cada variante em relação a uma métrica desejada".

Em seu trabalho, Kompella (2015) afirma que o teste A/B não é novo, porém, até pouco tempo, apenas grandes organizações com grandes equipes de *marketing* tinham condições de utilizar este tipo de avaliação, devido à complexidade estatística do processo. Porém, segundo o autor, atualmente já existem várias ferramentas disponíveis que podem ajudar a executar esse teste, ferramentas que podem ser consideradas simples ou até mais sofisticadas.

Segundo Anderson (2015), em um teste A/B, o avaliador configura um controle, como o estado atual do seu site (o Avariante). É enviado metade do tráfego do site para essa versão. Esses visitantes são o grupo (A). A outra metade é enviada para outra versão do site com algumas alterações, por exemplo, o botão de *checkout* indique 'compre' em vez de 'compre agora' (a variante B). Esses visitantes são o grupo (B). O avaliador determina o que está testando, a métrica de sucesso pode ser: o texto do botão afeta a receita média por visitante? O avaliador executa o experimento por um número predeterminado de dias ou semanas e, em seguida, executa uma análise estatística. Então o avaliador analisa se houve uma diferença estatisticamente significativa no comportamento focal. Nesse exemplo, receita por visitante entre o grupo (A) e o grupo (B). Se houver uma diferença, qual é a causa? Se tudo foi controlado (ou seja, foi o mesmo, exceto por esta pequena mudança), existem duas possibilidades. Pode ser devido ao acaso, o que pode acontecer se o tamanho da amostra for muito pequeno (ou seja, o experimento está abaixo do esperado). Alternativamente, a mudança entre as variantes (A) e (B) pode ser um fator causal. Os dados mostram que esse recurso causou essa mudança de comportamento.

Nos casos de uso de teste A/B descritos por Anderson (2015), o teste é feito em várias configurações de visualizações de sistemas *Web* para identificar qual retorna os melhores resultados em relação a consumo por parte dos usuários, ou que chama mais atenção para uma campanha de *marketing*, e consegue trazer melhores resultados.

Uma das vantagens do teste A/B, é que não é necessário ter uma explicação causal a priori porque algo deve funcionar, só é necessário testar, explorar e encontrar as melhorias com um impacto positivo (ANDERSON, 2015).

Para avaliar a classificação feita pelas ontologias e a classificação feita pelos algoritmos de *machine learning*, foi utilizado o conceito de teste A/B, onde foi comparado o desempenho das ontologias e dos três algoritmos usados (KNN, SVM e *Random Forest*) para avaliar qual deles teve o melhor desempenho na classificação dos *Datasets*.

5.1.3 Avaliação com Datasets sintéticos (Carga de trabalho)

Para a avaliação, foram gerados 121 *Datasets* que foram estruturados conforme a mesma estrutura que a plataforma *FIWARE* usa para estruturar seus ambientes IoT. O *FIWARE* armazena informações de contexto para qualquer tipo de ambiente IoT. Nele, é possível armazenar o histórico dos dados coletados pelos sensores, bem como as informações do contexto em que tais sensores estão inseridos (tipo de sensor, localização, características de uso, etc). Neste sentido, os *Datasets* usados buscam simular diferentes subdomínios da IoT com dados que poderiam ser extraídos de diferentes implantações de *middlewares*, como o *FIWARE*. A Figura 19 ilustra a estrutura de um *Dataset* usado nas avaliações.

Figura 19 – Ilustração da estrutura do *Dataset* usado na avaliação.

	A	B	C	D	E	F	G	H
1	ID	address	postalCode	coordinates	Sensor01	Sensor02	Sensor03	Sensor04
2	01	Berlin	10439	13.3986, 52.5547	143	93	143	1155
3	01	Berlin	10439	14.3986, 52.5547	108	58	108	1120
4	01	Berlin	10439	15.3986, 52.5547	141	91	141	1153
5	01	Berlin	10439	16.3986, 52.5547	126	76	126	1138
6	01	Berlin	10439	18.3986, 52.5547	132	82	132	1144
7	01	Berlin	10439	20.3986, 52.5547	133	83	133	1145
8	01	Berlin	10439	23.3986, 52.5547	107	57	107	1119
9	01	Berlin	10439	25.3986, 52.5547	116	66	116	1128

Fonte: Autor.

A primeira coluna é responsável por armazenar o 'ID' do *Dataset*, as demais para representar os sensores para o ambiente IoT e informações de localização, datas em que os dados foram gerados entre outras informações sobre o *Dataset*. Essas informações são inseridas na primeira linha do *Dataset*. As demais linhas são usadas para armazenar o histórico dos dados gerados pelos sensores. Também podem ser adicionadas colunas para armazenar informações como, localização, nome, entre outras informações.

Os *Datasets* usados na avaliação foram criados apenas com o título das colunas. Esses títulos representam os tipos de sensores usados na IoT. Não é necessária a inserção dos dados para a classificação, pois a classificação é feita baseada nos tipos de sensores, que são representados

pelos títulos das colunas nos *Datasets* em formato *Comma-Separated Values* (CSV). Para a avaliação, foi usado um único arquivo 'CSV' contendo todos os *Datasets* gerados.

Foram criados 11 *Datasets* completos, um para cada subdomínio. Posteriormente, foram criadas 12 variações para cada um desses *Datasets*, com exceção do *dataset* indefinido, que não permite a criação de variações, por este motivo, ele só tem 1 *Dataset*. As variações foram criadas com a retirada de forma aleatória da quantidade de sensores e, ao mesmo tempo, modificando a ordem dos sensores nos *Datasets*. Dessa forma, ao todo foram criados 121 *Datasets* para serem usados na avaliação.

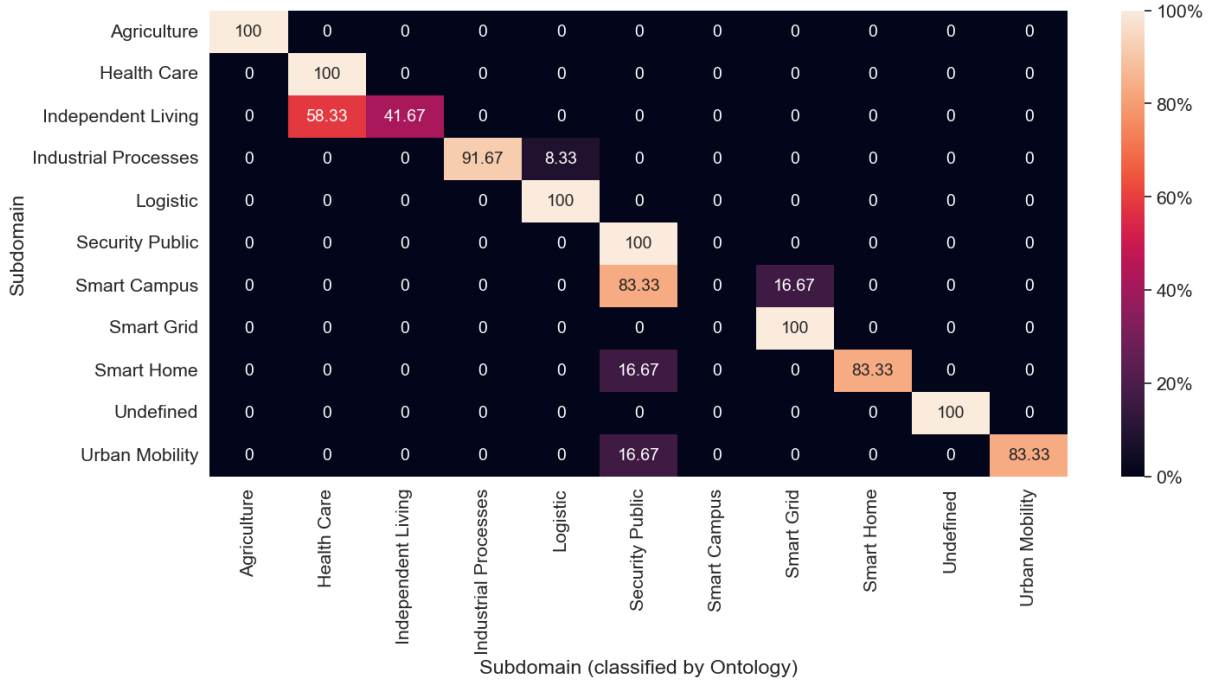
5.1.3.1 Classificação Usando a Base de Conhecimento das Ontologias com 121 Datasets

Um *Dataset* pode ter dados de sensores que podem pertencer a várias categorias. Por sua vez, uma categoria pode ser classificada em mais de um subdomínio. A aplicação usa as bases de conhecimento das ontologias para fazer a classificação. Portanto, a primeira etapa da avaliação teve como objetivo avaliar a classificação e a recomendação feita exclusivamente pelas ontologias, usando apenas as suas respectivas bases de conhecimento.

Para essa etapa da avaliação, todos os Datasets criados (121) foram inseridos como input para o *Smart Anonymity*. Neste cenário, a aplicação processa os dados de um *Dataset* criando instâncias dos sensores na ontologia *sensorCategory*. Em seguida, ao executar o raciocinador DL, os sensores são classificados em categorias de sensores, considerando a quantidade de sensores que pertencem a cada categoria. Por sua vez, as categorias retornadas são criadas como instâncias na ontologia *iotSubDomains*. Novamente, ao executar o raciocinador DL, as categorias são classificadas como parte de um ou mais subdomínios. Por fim, no *output* da aplicação, é impresso a lista dos *Datasets* e o algoritmo de anonimização recomendado.

Na Figura 20, é apresentada uma matriz de confusão e, nela, são apresentados os resultados da classificação usando apenas as bases de conhecimento das ontologias. Nessa imagem, é possível observar que o subdomínio *Smart Campus* está sendo classificado como *Public Security* e *Smart Grid*. É importante destacar que, de acordo com a análise dos *Datasets*, todos os sensores de *Smart Campus* também pertencem a outros subdomínios. Devido a isso, a aplicação não conseguiu classificar nenhum *Dataset* como sendo de *Smart Campus*. *IndependentLiving* também apresenta muitos sensores similares ao subdomínio *HealthCare*, devido a essa similaridade entre os dois subdomínio. De acordo com a classificação, 58,33% dos *Datasets*, que deveriam ser classificados como *IndependentLiving*, foram classificados como

Figura 20 – Matriz de confusão da classificação usando apenas a Ontologia.



Fonte: Autor.

Tabela 12 – Métricas avaliada na matriz de confusão da classificação usando apenas a Ontologia.

	Precision	**Recall**	**F1-score**	**Support**
Agriculture	1.00	1.00	1.00	12
HealthCare	0.63	1.00	0.77	12
IndependentLiving	1.00	0.42	0.59	12
IndustrialProcesses	1.00	0.92	0.96	12
Logistic	0.92	1.00	0.96	12
SecurityPublic	0.46	1.00	0.63	12
Smart Campus	0.00	0.00	0.00	12
Smart Grid	0.86	1.00	0.92	12
SmartHome	1.00	0.83	0.91	12
Undefined	1.00	1.00	1.00	1
UrbanMobility	1.00	0.83	0.91	12
Accuracy			0.80	121
Macro avg	0.81	0.82	0.79	121
Weighted avg	0.79	0.80	0.77	121

Fonte: O Autor.

HealthCare.

Na Tabela 12, são apresentados os resultados da avaliação para as métricas listadas na

seção 5.1.1 usando a aplicação com apenas as bases de conhecimento das ontologias. Nesta avaliação foi obtida uma acurácia geral de 80%. O subdomínio *Agriculture* atingiu 1.00 em todas as métricas avaliadas. Devido aos resultados obtidos nesta avaliação, existe a necessidade de se ter outro classificador para resolver falhas na classificação da ontologia ou possíveis disputas.

5.1.3.2 Classificação por Ontologia e Machine Learning com 61 Datasets

Para esta avaliação, os *Datasets* descritos na seção 5.1.3 foram divididos em dois subconjuntos com 61 e 60 *Datasets*, utilizados para treino e teste. Após a divisão, 61 *Datasets* foram usados para o treinamento dos três algoritmos de *Machine Learning*. Cada *Dataset* corresponde a um elemento que se pretende classificar nesta etapa. Diante disso, o (N) amostral corresponde a 61 elementos. O conjunto de dados foi normalizado, removendo a média e dividindo pelo desvio padrão. Foi empregado o método de validação cruzada do tipo Kfold, com o (K) definido em 5. Para os três algoritmos, foi utilizado um *grid* de pesquisa com um *range* de possíveis valores que poderiam ser utilizados no treino dos modelos. Nos testes, cada algoritmo foi utilizado no processo de classificação dos 60 *Datasets* utilizados na avaliação.

Para cada um dos algoritmos utilizados no *Smart Anonymity*, os parâmetros e valores usados para o treinamento foram definidos por meio de vários testes e baseado na literatura e documentação disponíveis. Na Tabela 13, todos os parâmetros e valores utilizados são destacados.

Tabela 13 – Parâmetros e valores utilizados para o treinamento dos algoritmos de *Machine Learning*.

KNN		**SVM**		**RF**	
Parâmetro	**Valor**	**Parâmetro**	**Valor**	**Parâmetro**	**Valor**
N_neighbors	3	*Kernel RBF*	Polinômio grau 3	*Criterion*	*Gini*
Weights	*Uniform*			*Max_depth*	10
		Gamma	*1/(n_features* X.var())*	*Min_samples_leaf*	1
Algorithm	*Auto*			*Min_samples_split*	5
				N_estimators	50

Fonte: O Autor.

No processo de treinamento do algoritmo KNN, foram utilizados os parâmetros, *'n_neighbors = 3'*, o *weights* foi definido como *'uniform'* e o parâmetro *algorithm* foi definido como *'auto'*, neste caso, o próprio modelo seleciona o melhor algoritmo com base no treinamento.

No processo de treinamento do algoritmo SVM, foi utilizado o *kernel RBF* com polinômio de grau 3 e o parâmetro *gamma* foi definido como *'1 / (n_features * X.var())'*. Já no processo de treinamento do algoritmo *Random Forest* foram utilizados os seguintes parâmetros:

- ***criterion: gini***: mede a qualidade do subconjunto;
- ***max_depth: 10***: define a profundidade máxima da árvore;
- ***min_samples_leaf: 1***: define o número mínimo de amostras para criar um nó folha;
- ***min_samples_split: 5***: define a quantidade mínima de amostras necessárias para dividir um nó interno; e
- ***n_estimators: 50***: define o número de árvores na floresta.

Para o desenvolvimento e treinamento de todos os algoritmos de *Machine Learning*, foi utilizada a linguagem de programação Python, e a biblioteca *scikit learn*. O ambiente de desenvolvimento usado foi o *Visual Studio Code*, que foi instalado em uma máquina com sistema operacional Windows 10 (64 bits), processador Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz, 8 Gigabytes de memória RAM.

Após o processo de treinamento dos algoritmos de *Machine Learning*, os outros 60 *Datasets* foram executados na aplicação com apenas as ontologias. Os mesmos 60 *Datasets* também foram usados na versão da aplicação com as ontologias e os algoritmos de *Machine Learning*. A qualidade do treinamento dos algoritmos de *Machine Learning* foi avaliada no conjunto com 60 *Datasets* separados para o teste. Dessa forma, foi possível realizar a comparação dos resultados entre a classificação apenas com a ontologia e a classificação da Ontologia em conjunto dos algoritmos de *Machine Learning*.

Com esta avaliação, também é possível avaliar o desempenho dos algoritmos de *Machine Learning* para identificar qual obtém melhores resultados na classificação em conjunto com as ontologias. Nesse sentido, é importante destacar que o principal objetivo do uso dos algoritmos de *Machine Learning* é para melhorar os resultados da classificação feita pelas ontologias.

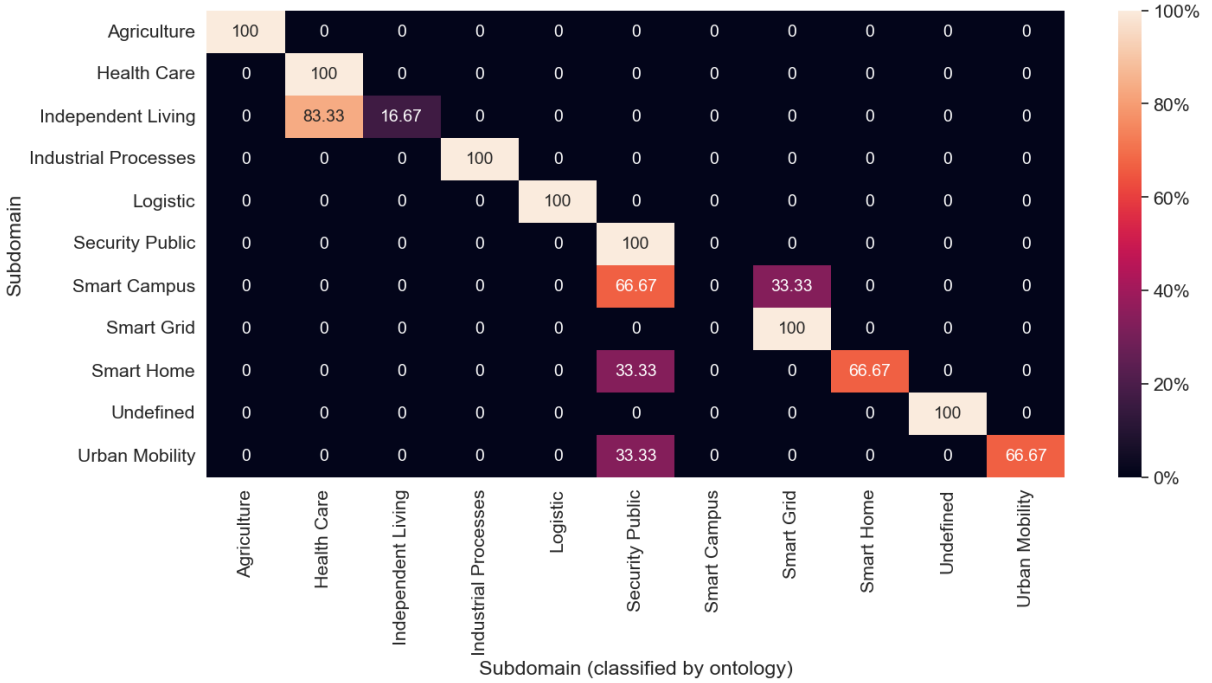
Para que a classificação seja executada pelos algoritmos de *Machine Learning*, eles recebem como entrada os dados pré-classificados pelas ontologias em categorias. A partir dessa classificação inicial, os algoritmos de *machine learning* conseguem finalizar a classificação. Além disso, eles conseguem melhores resultados comparados à classificação feita usando apenas a base de conhecimento das ontologias. É importante destacar que, nesse ponto, as ontologias

sempre conseguem classificar o *Dataset* Indefinido. O motivo para isso, é que esse *Dataset* não tem dados para serem classificados em categorias.

5.1.3.2.1 *Classificação Usando a Base de Conhecimento das Ontologias com 61 Datasets*

Nesta avaliação, a aplicação está usando apenas as bases de conhecimento das ontologias. Para essa avaliaçã,o a aplicação recebeu como entrada 61 *Datasets* para serem classificados e recomendado o algoritmo de anonimização mais adequado, considerando as características dos dados. A Figura 21 apresenta os resultados da avaliação em uma matriz de confusão usando a aplicação com apenas as ontologias. Dos onze subdomínios, nove foram classificados de forma correta, dentre eles, sete tiveram 100% de acerto na classificação com as ontologias. O resultados dessa avaliação são similares aos resultados obtidos na seção 5.1.3.1. Nesse cenário, 66.67% que deveriam se classificados como *Smart Campus* foram classificados como *Security Public*.

Figura 21 – Matriz de confusão da classificação usando apenas a Ontologia com 61 *Datasets*.

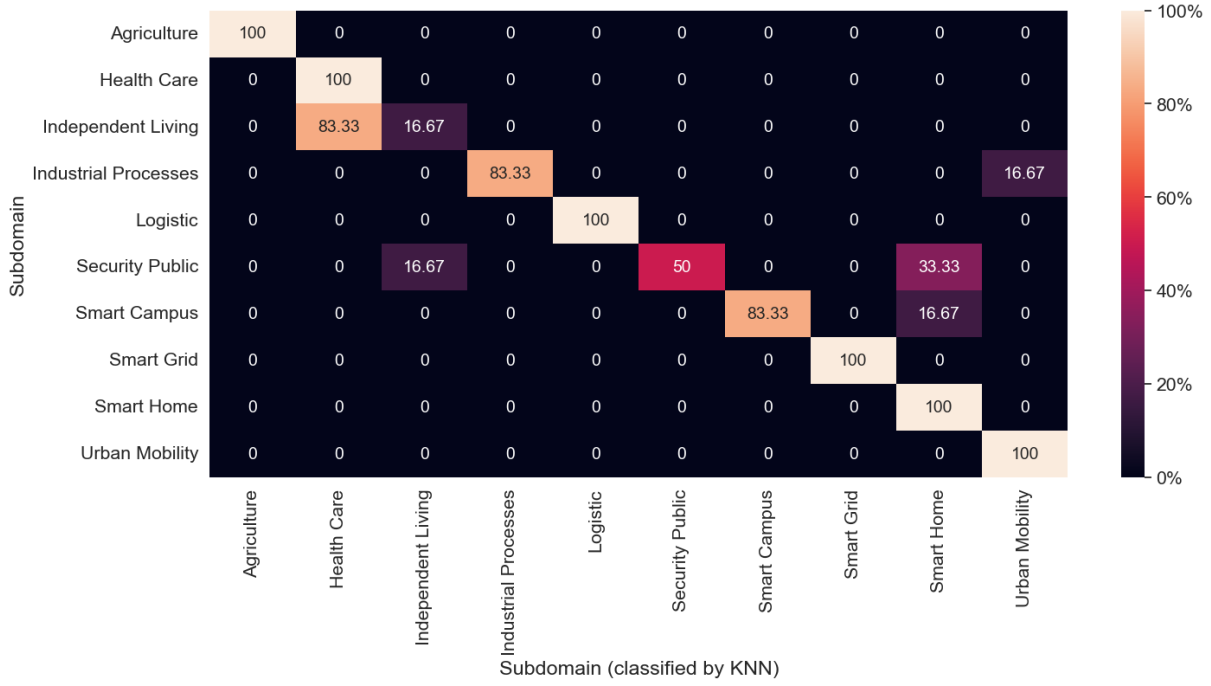


Fonte: Autor.

5.1.3.2.2 Classificação Usando KNN com 61 Datasets

Nesta avaliação, foram submetidos 60 *Datasets* para serem processados pela aplicação. Nessa classificação, foram usadas as bases de conhecimento das ontologias, porém, para melhorar os resultados da classificação, também foi usado o algoritmo de *Machine Learning* KNN. Como dito anteriormente na seção 5.1.3, nesse momento, os *Datasets* já foram pré-classificados em categorias, para que o KNN possa finalizar a classificação. A seguir, a Figura 22 apresenta os resultados da avaliação em uma matriz de confusão. Dos dez subdomínios, nove foram classificados de forma correta, e seis tiveram 100% de acerto na classificação com as ontologias e o KNN. Com o uso desse algoritmo, a classificação foi capaz de classificar o *Smart Campus* de forma correta. Porém, o subdomínio *IndependentLiving* teve 83.33% dos *Datasets* classificados como *Health Care*.

Figura 22 – Matriz de confusão da classificação usando apenas o KNN com 61 *Datasets*.



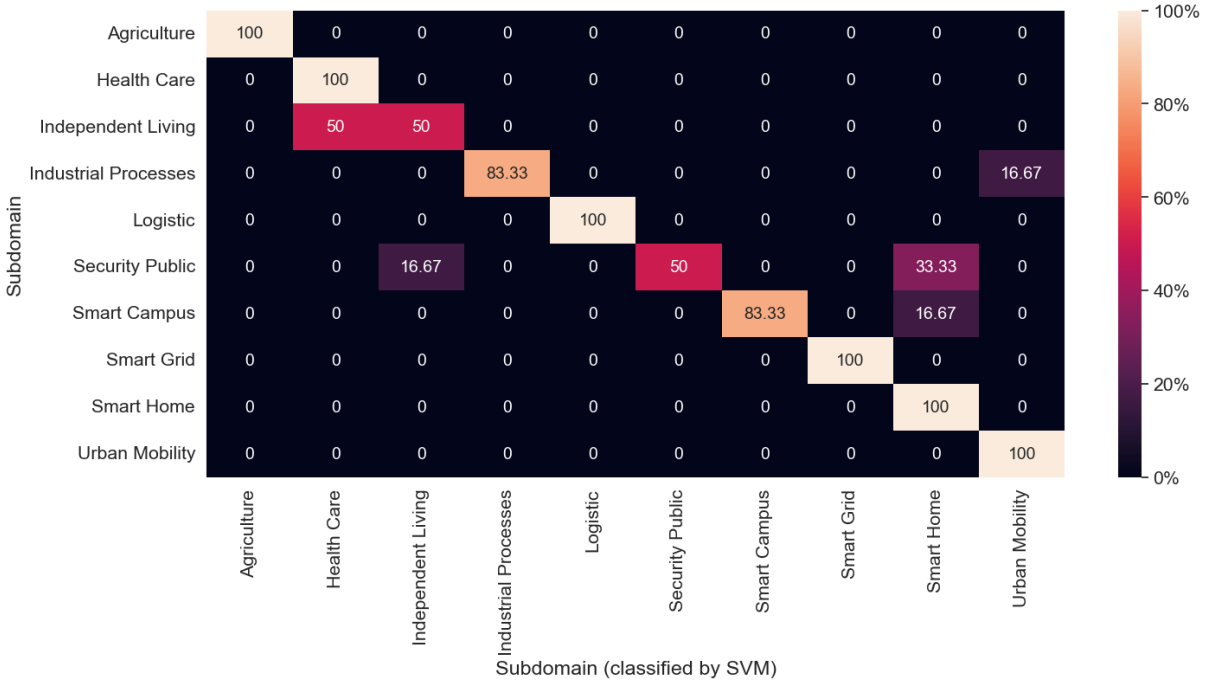
Fonte: Autor.

5.1.3.2.3 Classificação Usando SVM com 61 Datasets

Para esta avaliação, foram usados 60 *Datasets* para análise na aplicação. Nessa classificação, foram usadas as bases de conhecimento das ontologias, contudo, para melhorar os resultados da classificação, também foi usado o algoritmo de *Machine Learning* SVM. A seguir, na

Figura 23, são apresentados os resultados da avaliação em uma matriz de confusão. Os dez subdomínios foram classificados de forma correta, e seis tiveram 100% de acerto na classificação com as ontologias e o SVM. Com o uso do algoritmo SVM os subdomínios *Smart Campus* e *IndependentLiving* foram classificados de forma correta.

Figura 23 – Matriz de confusão da classificação usando apenas o SVM com 61 *Datasets*.

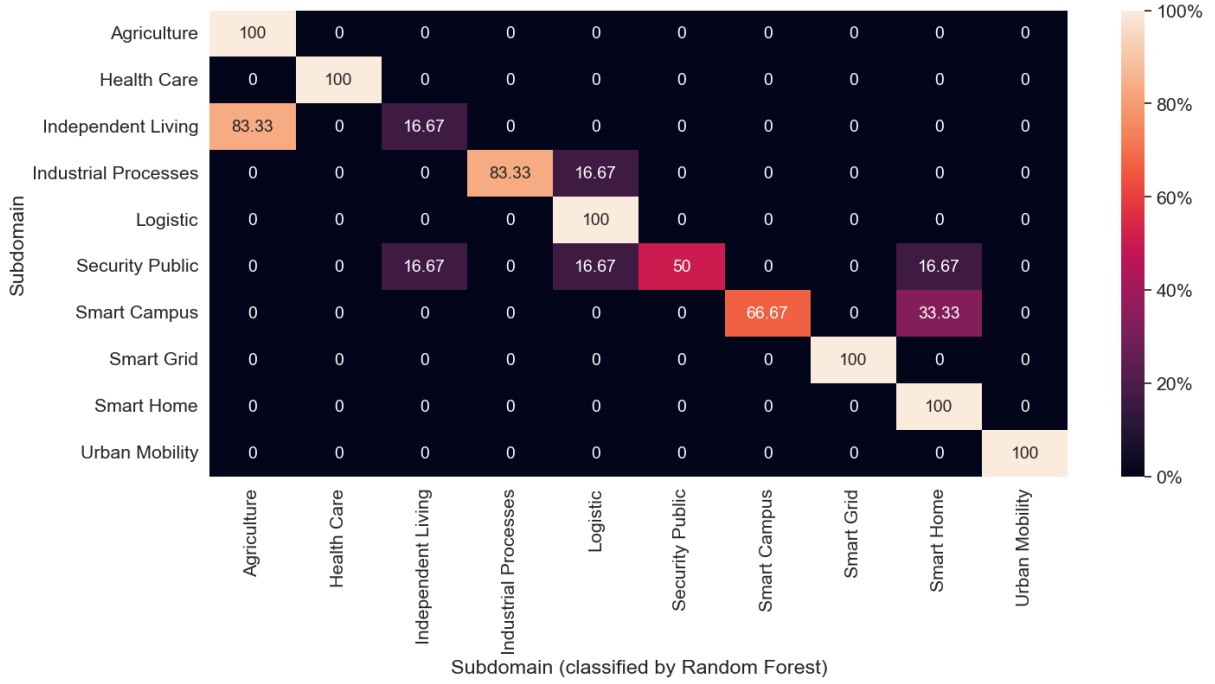


Fonte: Autor.

5.1.3.2.4 Classificação Usando Random Forest com 61 Datasets

Nesta avaliação, também foram submetidos 60 *Datasets* na aplicação para classificação e recomendação de algoritmos de anonimização. São usadas as bases de conhecimento das ontologias mais o algoritmo de *Machine Learning Random Forest*. A seguir, a Figura 24 apresenta os resultados da avaliação em uma matriz de confusão usando a aplicação com as ontologias e o *Random Forest* para melhorar os resultados da classificação. Dos dez subdomínios, nove foram classificados de forma correta, e seis tiveram 100% de acerto na classificação com a Ontologia e o *Random Forest*. Com o uso do *Random Forest* e a base de conhecimento das ontologias, a aplicação foi capaz de classificar o *Smart Campus* de forma correta. *Security Public* teve *Datasets* classificados como *Independent Living*, *Logistic* e *Smart Home*.

Figura 24 – Matriz de confusão da classificação usando apenas o RF com 61 *Datasets*.



Fonte: Autor.

5.1.3.2.5 Análise Comparativa Entre Ontologia e Algoritmos de Machine Learning

Com o uso dos algoritmos de *Machine Learning*, foram obtidos melhores resultados na classificação em relação ao uso apenas das bases de conhecimento das ontologias. Na Tabela 14, é possível visualizar os dados obtidos na avaliação e comparar os resultados usando apenas a Ontologia para classificação e os resultados usando os algoritmos de *Machine Learning*. Como o objetivo da aplicação é classificar um *Dataset* usado como *input* na aplicação em um dos 11 subdomínios, o *target* da avaliação descrita na Tabela 14 é justamente a lista dos 11 subdomínios em que um determinado *Dataset* pode ser classificado.

Conforme os resultados obtidos, foi possível verificar que a *Machine Learning* junto com as ontologias alcançam melhores resultados. Já entre os algoritmos de *Machine Learning* aplicados nessa avaliação, o SVM obtive melhores resultados com 87% de acurácia geral. Ele também apresentou melhores resultados em todas as outras métricas analisadas. Desta forma, é possível concluir que o uso de *Machine Learning* ajuda a melhorar os resultados da classificação e, dentre os algoritmos usados, o que obteve melhores resultados foi o SVM.

Nesta etapa da avaliação, foi usado o conceito de teste A/B para comparar os resultados da classificação usando apenas as ontologias e a versão usando os algoritmos de *Machine Learning* com as ontologias. Também são comparados os resultados para cada algoritmo de *Machine*

Learning. Diante disso, foi possível concluir que o teste A/B foi muito importante para identificar as melhores configurações no mecanismo proposto para que os melhores resultados fossem alcançados nas avaliações. A comparação feita nessa avaliação utiliza as métricas descritas na seção 5.1.1. Com essa avaliação, foi identificado que a melhor versão do mecanismo de recomendação proposto baseado na comparação do desempenho é o uso do algoritmo SVM juntamente com as bases de conhecimento das ontologias.

Tabela 14 – Métricas avaliadas na matriz de confusão da classificação usando os 4 métodos.

SubDomain	Ontologia				KNN				SVM				Random Forest			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
Agriculture	1.00	1.00	1.00	6	1.00	1.00	1.00	6	**1.00**	**1.00**	**1.00**	**6**	0.55	1.00	0.71	6
Health Care	0.55	1.00	0.71	6	0.55	1.00	0.71	6	**0.67**	**1.00**	**0.80**	**6**	1.00	1.00	1.00	6
Independent Living	1.00	0.17	0.29	6	0.50	0.17	0.25	6	**0.75**	**0.50**	**0.60**	**6**	0.50	0.17	0.25	6
Industrial Processes	1.00	1.00	1.00	6	1.00	0.83	0.91	6	**1.00**	**0.83**	**0.91**	**6**	1.00	0.83	0.91	6
Logistic	1.00	1.00	1.00	6	1.00	1.00	1.00	6	**1.00**	**1.00**	**1.00**	**6**	0.75	1.00	0.86	6
Security Public	0.43	1.00	0.60	6	1.00	0.50	0.67	6	**1.00**	**0.50**	**0.67**	**6**	1.00	0.50	0.67	6
Smart Campus	0.00	0.00	0.00	6	1.00	0.83	0.91	6	**1.00**	**0.83**	**0.91**	**6**	1.00	0.67	0.80	6
Smart Grid	0.75	1.00	0.86	6	1.00	1.00	1.00	6	**1.00**	**1.00**	**1.00**	**6**	1.00	1.00	1.00	6
Smart Home	1.00	0.67	0.80	6	0.67	1.00	0.80	6	**0.67**	**1.00**	**0.80**	**6**	0.67	1.00	0.80	6
Undefined	1.00	1.00	1.00	1	-	-	-	-	-	-	-	-	-	-	-	-
Urban Mobility	1.00	0.67	0.80	6	0.86	1.00	0.92	6	**0.86**	**1.00**	**0.92**	**6**	1.00	1.00	1.00	6
Accuracy			0.75	61			0.83	60			**0.87**	**60**			0.82	60
Macro avg	0.79	0.77	0.73	61	0.86	0.83	0.82	60	**0.89**	**0.87**	**0.86**	**60**	0.85	0.82	0.80	60
Weighted avg	0.78	0.75	0.71	61	0.86	0.83	0.82	60	**0.89**	**0.87**	**0.86**	**60**	0.85	0.82	0.80	60

abbreviation - M1: Precision; M2: Recall; M3: F1-score; M4: Support;

Fonte: O Autor.

Foi realizado o teste de normalidade shapiro-wilk para todas as métricas listadas na seção 5.1.1. Para todos os casos, não foi verificada a existência de normalidade dos dados (p-Valor menor que 0.05). Diante disso, foi aplicado o teste não paramétrico de Wilcoxon para amostras pareadas.

Com base nos resultados apresentados na Tabela 15, foi possível determinar diferenças estatísticas apenas para a comparação entre ontologias e o algoritmo SVM para a métrica *F1-Score* (p-Valor = 0.0425). Isso confirma que a utilização do algoritmo de Machine Learning SVM proporcionou uma melhora nos resultados da classificação dos *Datasets*, principalmente pelo fato da métrica *F1-Score* ser mais robusta que a Precisão e o *Recall*.

Tabela 15 – Resultado da comparação da classificação pelas ontologias e da classificação por ontologias combinadas com os algoritmos de *Machine Learning*.

Comparação	**Precisão**		**Recall**		**F1-Score**	
	Teste t	**p-Valor**	**Teste t**	**p-Valor**	**Teste t**	**p-Valor**
Ontologia x KNN	8.0	0.6001	5.0	0.4982	4.0	0.1729
Ontologia x SVM	6.5	0.4003	6.0	0.3401	2.0	0.0425*
Ontologia x RF	15.0	0.6736	5.0	0.4982	16.0	0.4405
SVM x KNN	0	0.1088	0	0.3173	0	0.1797
SVM x RF	3.0	0.4614	0	0.1797	5.0	0.2488
RF x KNN	4.5	0.8539	0	0.3173	5.5	0.5879

*, estatisticamente diferente ao nível de 5% de significância.

Fonte: O Autor.

O resultado desse teste deve ser interpretado de forma cautelosa, pois as métricas Precisão e *Recall* estão infladas com o valor '1', o que causa muitos empates na hora do ranqueamento do teste de Wilcoxon, afetando a precisão do mesmo.

5.2 AVALIAÇÃO COM PROVA DE CONCEITO

Para validar a aplicabilidade do *Smart Anonymity*, foi desenvolvida uma prova de conceito com três cenários de uso da IoT (*Agriculture*, *HealthCare* e *Smart Grid*). Esses três cenários foram escolhidos por serem de áreas totalmente distintas e também por serem compostas por diferentes tipos de sensores. A prova de conceito tem o objetivo de produzir conhecimento a respeito do uso do *Smart Anonymity* testando seu funcionamento em cenários de uso da IoT.

Para tanto, foi usada a plataforma *FIWARE* como *middleware* para criação dos cenários. Essa é uma plataforma de código aberto para IoT, e a escolha dessa plataforma se deu porque

ela é muito usada pela comunidade que pesquisa sobre IoT para criação de ambientes IoT. A plataforma *FIWARE* disponibiliza vários tutoriais para treinamentos[1]. Para o desenvolvimento da prova de conceito, foi utilizada como base a estrutura disponível no tutorial *Querying Time Series Data (Crate-DB)*[2]. Nesta prova de conceito, a infraestrutura necessária para usar o *FIWARE* foi instalada em uma máquina com sistema operacional Linux (Ubuntu 22.04 LTS), processador Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz, 8 Gigabytes de memória RAM.

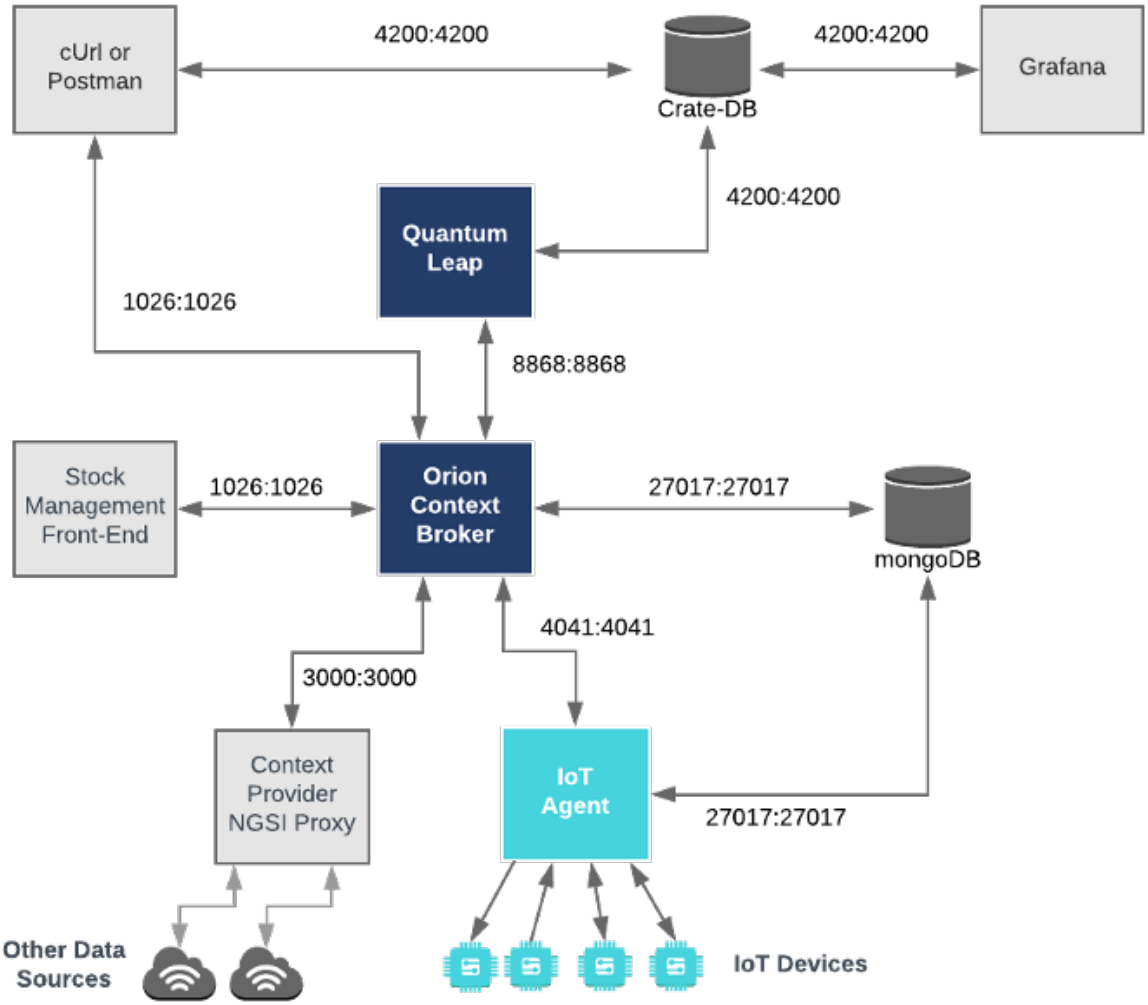
Para desenvolver a prova de conceito, foi levantada uma infraestrutura com o *Docker Compose*. Por padrão, ele já cria um cenário com sensores de uma loja (exemplo: sensor de porta, sensor de campainha, sensor de movimento e sensor de iluminação). Contudo, para viabilizar a prova de conceito, foram alterados somente os tipos de sensores, pois é o que importa para a aplicação. Foram criados sensores usando requisições *REST*. Os dados dos sensores também foram inseridos usando requisições *rest*. A Figura 25 ilustra a arquitetura da infraestrutura usada na prova de conceito.

Na prova de conceito são usados três componentes *FIWARE*: o *Orion Context Broker*, o IoT *Agent for Ultralight 2.0* e o *QuantumLeap* (FIWARE, 2023). A arquitetura global foi constituída, usando os seguintes elementos:

- O *FIWARE Generic Enablers*:
  - O *FIWARE Orion Context Broker*, que receberá pedidos utilizando o NGSI-v2.
  - O *FIWARE IoT Agent for Ultralight 2.0*, que receberá medições dos dispositivos IoT fictícios no formato *Ultralight 2.0* e as converterá em pedidos NGSI-v2 para que o corretor de contexto altere o estado das entidades de contexto.
  - O *FIWARE QuantumLeap* subscreve as alterações de contexto e persiste em um *CrateDB database*.
- Uma *MongoDB database*:
  - Utilizado pelo *Orion Context Broker* para guardar informações sobre dados de contexto, tais como entidades de dados, subscrições e registros.
  - Utilizado pelo agente IoT para guardar informações sobre o dispositivo, tais como URLs e chaves do dispositivo.

[1] (https://fiware-tutorials.readthedocs.io/en/latest/index.html)
[2] https://fiware-tutorials.readthedocs.io/en/latest/time-series-data.html

Figura 25 – Infraestrutura usada na prova de conceito.



Fonte: FIWARE (2023).

- Uma *CrateDB database*:
  - Utilizado para guardar dados de contexto histórico baseados no tempo.
  - Oferece um ponto *endpoint* HTTP para interpretar consultas de dados baseadas no tempo.
- Um *Context Provider*: - Um servidor *Web* que atua como um conjunto de dispositivos IoT fictícios, utilizando o protocolo *Ultralight 2.0* em execução sobre HTTP.

Os dados gerados pelos cenários criados para a prova de conceito são capturados por meio de uma consulta do QuantumLeap. Nessa consulta, são retornados os dados em formato Json, e

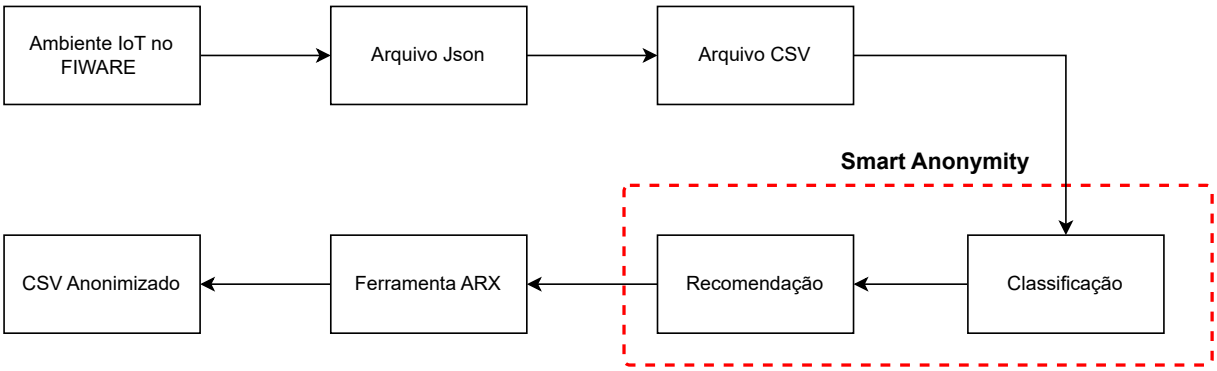
posteriormente, esses dados são transformados em um arquivo 'CSV' para serem anonimizados. A anonimização é executada pelo ARX versão 3.9.1.

Consulta utilizada para capturar todos os dispositivos provisionados:

curl -X GET 'http://localhost:4041/iot/devices' -H 'fiware-service: openiot' -H 'fiware-servicepath: /'

A partir dessa infraestrutura, foram montados cenários seguindo as características de alguns subdomínios da IoT. Nesses cenários, foram criados todos os sensores para cada subdomínio. A Figura 26 apresenta o fluxo de execução da prova de conceito.

Figura 26 – Fluxo de execução da prova de conceito.



Fonte: Autor.

Após a criação dos cenários na plataforma *FIWARE*, é retornado um arquivo Json com as informações dos cenários como, por exemplo, todos os sensores que fazem parte do cenário, bem como as informações sobre o cenário (ID, localização, endereço). A partir desse arquivo Json, foram extraídos os dados e organizados em um arquivo 'CSV'. O arquivo 'CSV' é submetido à aplicação desenvolvida para que sejam analisadas suas características e seja classificada em um dos subdomínios da IoT e, posteriormente, seja recomendado o algoritmo de anonimização mais adequado. Após a classificação e recomendação do algoritmo de anonimização, o arquivo 'CSV' é utilizado para executar a anonimização usando o ARX.

É importante destacar que a anonimização também pode ser feita por outras ferramentas, ou a aplicação pode ser estendida com a implementação dos algoritmos de anonimização. Conforme descrito anteriormente, a Figura 26 ilustra o fluxo de execução das provas de conceito.

5.2.1 Execução da Prova de Conceito

Nesta prova de conceito, o ARX foi instalado em uma máquina com sistema operacional Windows 10 (64 bits), processador Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz, 8 Gigabytes de memória RAM. Segundo a documentação disponível no site de ferramenta, também funciona no sistema operacional como Linux. O ARX disponibiliza uma API para que aplicações externas possam acessar e usar seus recursos.

5.2.2 Cenário 01 (*Agriculture*)

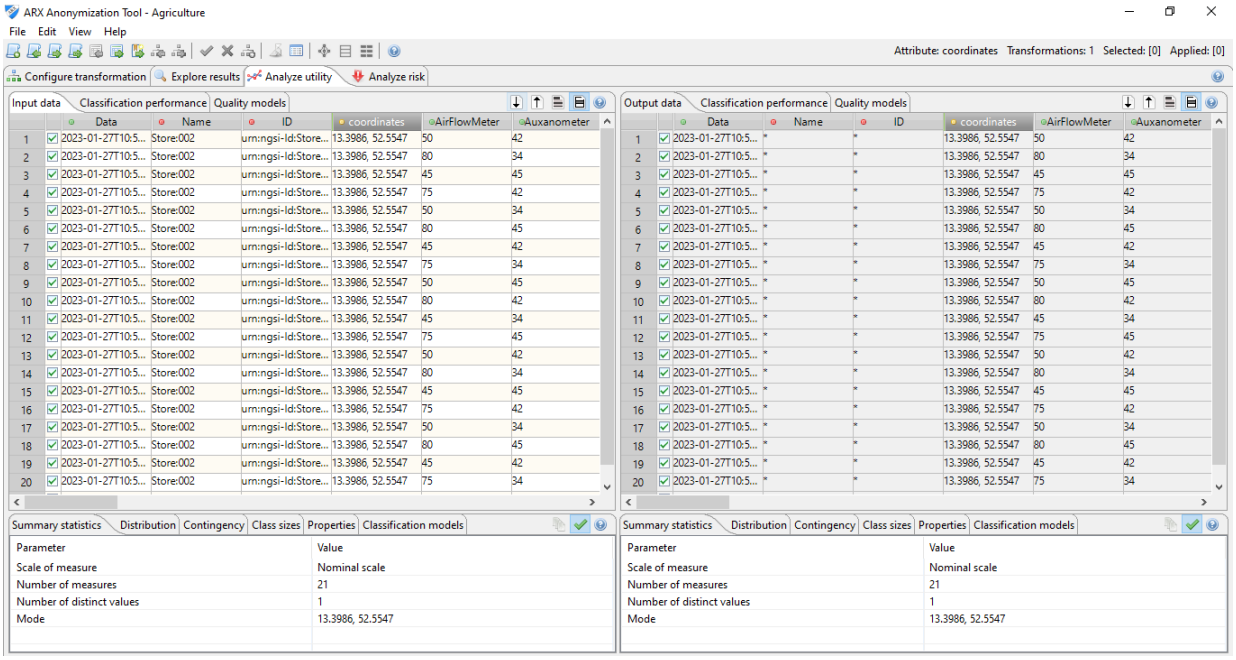
Foi criado um ambiente de *Agriculture*. Esse cenário contém todos os sensores que pertencem ao referido subdomínio. A nomenclatura e estrutura dos sensores seguiu a mesma estrutura adotada pelas ontologias desenvolvidas e usadas na aplicação. Essa mesma estrutura para criação dos sensores também é adotada pela plataforma FIWARE. Foram utilizados os seguintes sensores:

- AirFlowMeter;
- Auxanometer;
- BackIlluminated;
- Barograph;
- Barometer;
- Biochip;
- Bio;
- BoostGauge;
- Calorimeter;
- CapacitanceProbe;
- FrequencyDomain;
- Humistor;

- Hydrometer;
- Hyperspectral;
- InfraredThermometer;
- Leaf;
- LedAsLight;
- Lysimeter;
- Nano;
- Pressure;
- PressureGauge;
- Psychrometer;
- Pyrometer;
- QuartzThermometer;
- ResistanceThermometer;
- SiliconBandgapTemperature;
- SoilMoisture;
- TemperatureGauge;
- Thermocouple;
- Thermometer;
- TimePressureGauge;
- UnattendedGround; e
- WaterMeter.

Na Figura 27, é possível visualizar, do lado esquerdo, os dados antes da anonimização e, do lado direito, os dados após a anonimização. Após o processo de anonimização, o ARX permite que os dados anonimizados sejam exportados em um arquivo 'CSV'. Como os dados agora estão anonimizados, eles agora podem ser distribuídos para terceiros sem preocupações relacionadas à exposição de dados privados dos usuários. A Figura 27 mostra os dados do cenário de *Agriculture* sendo analisados e anonimizados.

Figura 27 – Visualização do ARX anonimizando dados de *Agriculture*.



Fonte: Autor.

5.2.3 Cenário 02 (*HealthCare*)

Nesse cenário, foi criado um ambiente de *HealthCare*. Esse cenário contém todos os sensores correspondentes ao referido subdomínio. Para nomenclatura e estrutura dos sensores, foi seguida a mesma estrutura adotada pelas ontologias na aplicação. Segue a lista dos sensores usados:

- Accelerometer;
- Actigraphy;
- Bio;
- Biochip;

- DiffusionTensorImaging;
- FunctionalMagneticResonanceImaging;
- GroundSpeedRadar;
- Heartbeat;
- MagneticResonanceImaging;
- Nano;
- NanoTetherball;
- PiezoelectricAccelerometer; e
- Location.

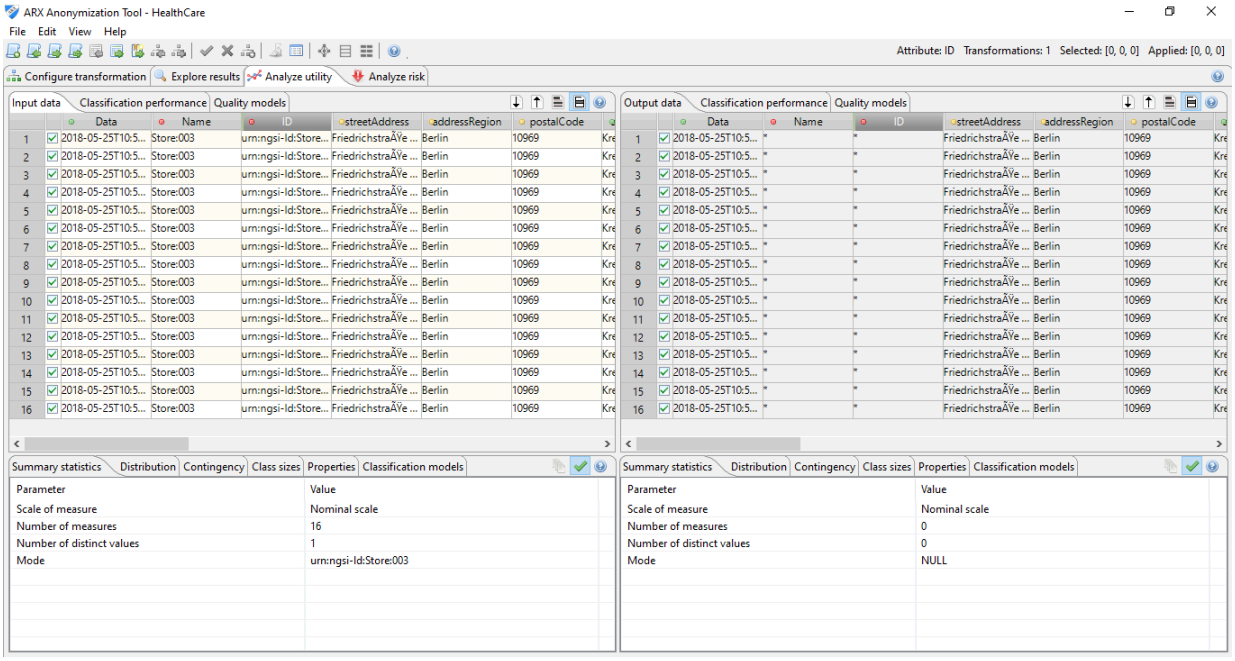
Na Figura 28, é possível visualizar que, do lado esquerdo, são apresentados os dados antes da anonimização e, do lado direito, ficam os dados após a anonimização. Após o processo de anonimização, o ARX possibilita que os dados já anonimizados sejam exportados em um arquivo 'CSV'. Como os dados já estão anonimizados, eles podem ser distribuídos para terceiros sem preocupações relacionadas à exposição de dados privados dos usuários. A Figura 28 mostra os dados do cenário de *HealthCare* sendo analisados e anonimizados.

5.2.4 Cenário 03 (*Smart Grid*)

Por fim, foi criado um cenário que representa o ambiente de *Smar Grid*. Para este cenário, foram usados todos os sensores correspondentes ao referido subdomínio. A nomenclatura e estrutura dos sensores seguiu a mesma estrutura adotada pelas ontologias na aplicação. A lista de sensores usados foi a seguinte:

- Current;
- Galvanometer;
- Photodiode;
- ProportionalCounter;
- StrainGauge;

Figura 28 – Visualização do ARX anonimizando dados de *Health Care*.

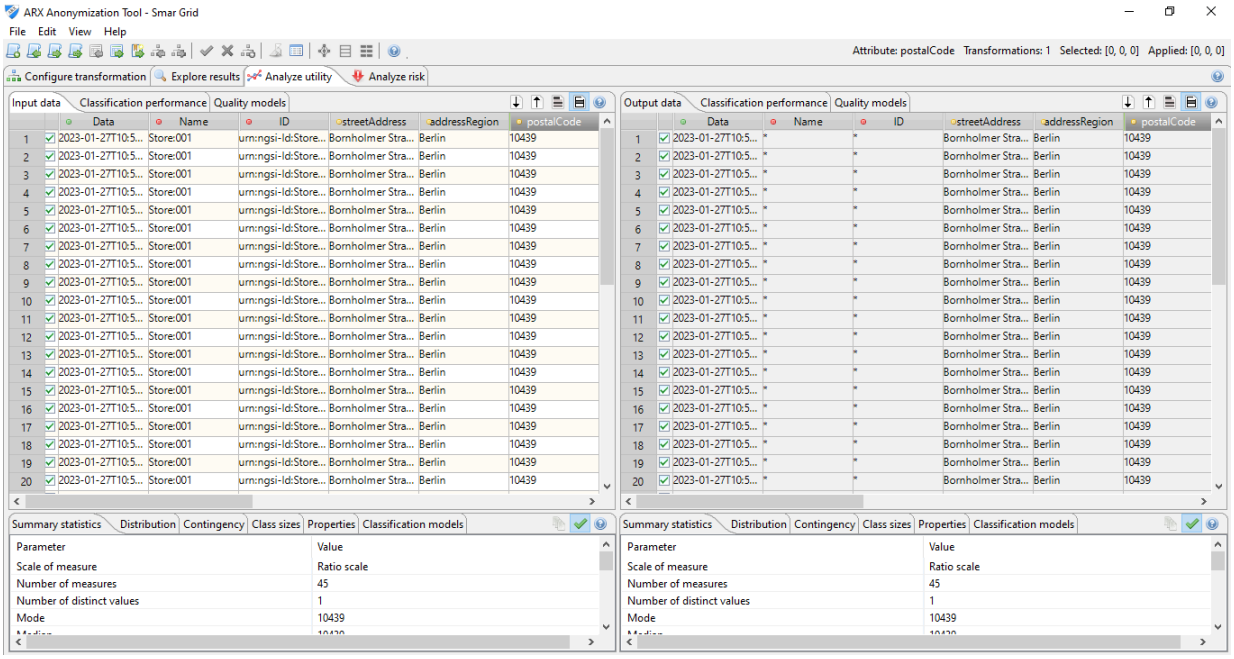


Fonte: Autor.

- TestLight;
- VoltageDetector; e
- WheatstoneBridge.

Na Figura 29, é possível visualizar, do lado esquerdo, os dados antes da anonimização e, do lado direito, os dados após a anonimização. Após o processo de anonimização, o ARX possibilita que os dados anonimizados sejam exportados em um arquivo 'CSV' para utilização futura. Com os anonimizados, eles agora podem ser distribuídos para terceiros sem preocupações relacionadas à exposição de dados privados dos usuários. A Figura 29 mostra os dados do cenário de *Smart Grid* sendo analisados e anonimizados.

Figura 29 – Visualização do ARX anonimizando dados de *Smart Grid*.



Fonte: Autor.

5.3 CONSIDERAÇÕES FINAIS

Neste capítulo, foi apresentada a metodologia utilizada para a avaliação do mecanismo proposto, e também são apresentadas a métricas usadas. A avaliação foi dividida em duas etapas. O processo de avaliação da primeira etapa foi dividido em cinco avaliações focadas no desempenho da classificação. Na segunda etapa, foi realizada uma prova de conceito, e foram criados três cenários de uso da IoT usando a plataforma *FIWARE*. Com essa prova de conceito, é possível demostrar o funcionamento da abordagem proposta, em que foram executados todos os processos, desde a criação dos cenários até a classificação e recomendação do algoritmo de anonimização, finalizando com o uso do ARX para executar o processo de anonimização dos dados.

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste capítulo, é apresentada uma síntese desta Tese. Também é apresentado como os objetivos propostos foram atingidos, bem como as contribuições da pesquisa para a área de privacidade de dados em IoT. Em seguida, são apresentadas as limitações identificadas ao longo do desenvolvimento desta Tese. Por fim, são apresentadas as direções para pesquisas futuras e as conclusões.

6.1 CONCLUSÕES

Como discutido em capítulos anteriores, a IoT está sendo utilizada em várias áreas para trazer comodidade à vida humana. Existe uma grande quantidade de sensores que podem ser utilizados para tornar as atividades cotidianas das pessoas mais práticas e rápidas. Alguns exemplos são sensores que conseguem controlar dispositivos dentro de uma casa como, por exemplo, sensores de temperatura, sensores de luminosidade, sensores de movimento que podem ser usando para ligar ou desligar as lâmpadas, entre outros. Com essa vasta utilização da IoT, estão surgindo várias oportunidades para novos negócios. Entretanto, com a vasta utilização de dispositivos IoT, também surge uma preocupação com a privacidade dos dados gerados por esses dispositivos.

Vários estudos disponíveis na literatura apontam o uso de anonimização de dados como uma solução viável para tratar os problemas relacionados a privacidade de dados em IoT. Porém, dentre os trabalhos analisados, nenhum apresenta uma solução baseada em anonimização que possa ser usada em vários domínios de uso da IoT. Diante disto, esta Tese propõe o desenvolvimento do *Smart Anonymity*, que tem por objetivo recomendar qual algoritmo de anonimização de dados é o mais adequado para o conjunto de dados de acordo com suas características. Para o desenvolvimento dessa solução, foram utilizadas várias tecnologias, como ontologias OWL que usam Lógica de Descrição e também foi utilizada *Machine Learning* para melhorar os resultados alcançados pelas ontologias na classificação dos dados analisados.

Com base nos resultados das avaliações realizadas no decorrer desta Tese, é possível concluir que o *Smart Anonymity* apresenta resultados promissores para a classificação e recomendação dos algoritmos de anonimização para dados gerados por dispositivos IoT. Ainda com base nos resultados das avaliações, é possível concluir que o uso de *Machine Learning* traz melhorias no

processo de classificação dos algoritmos de anonimização.

De acordo com os dados coletados e analisados nas avaliações, é possível concluir que o *Smart Anonymity* pode ser usado em cenários reais de IoT. Também foi possível concluir que a plataforma *FIWARE* usa a mesma estrutura usada nas ontologias para estruturar os sensores e os dados gerados pelos sensores.

Por fim, o código-fonte e todos dos dados usados nas avaliações estão disponíveis no *Gitub*: <https://github.com/flavionevesfsn/SmartAnonymity>.

6.2 CONTRIBUIÇÕES

Esta Tese apresenta um avanço na área de privacidade de dados em IoT. Este avanço busca permitir que as pessoas usuárias de dispositivos IoT não precisem se preocupar com a privacidade dos dados que são gerados e coletados por esses dispositivos. Para tratar a privacidade dos dados coletados, eles são analisados e classificados de acordo com suas características para que seja recomendado o algoritmo de anonimização mais adequado.

A anonimização dos dados pessoais pode reduzir a preocupação das pessoas com sua privacidade, pois os dados anonimizados não permitem mais a identificação do seu proprietário. Após a anonimização, os dados podem ser armazenados em algum servidor da empresa responsável por manter os dados. Esses dados agora podem ser disponibilizados para terceiros, se for de interesse da empresa e se o usuário autorizar, ou até mesmo se os dados caírem em mãos erradas, não irão colocar a privacidade de seus donos em risco.

A principal contribuição desta Tese foi o desenvolvimento do *Smart Anonymity*. Esse mecanismo faz a classificação e recomendação de algoritmos de anonimização, levando em consideração as características dos dados. Para alcançar o objetivo desta pesquisa, são utilizadas várias tecnologias. Inicialmente, os dados recebidos como entrada pelo *Smart Anonymity* são classificados em sensores; depois são classificados em categorias, essa classificação é feita pela aplicação usando as bases de conhecimento das ontologias. Após esse momento, os algoritmos de *Machine Learning* recebem os dados classificados em categorias e finalizam a classificação dos dados em subdomínios da IoT, dessa forma atendendo ao objetivo geral desta Tese. Após a classificação dos dados, é feita a recomendação do algoritmo de anonimização mais adequado para os dados que estão sendo analisados. Após a recomendação do algoritmo de anonimização, os dados podem ser submetidos ao algoritmo recomendado. Essa etapa é feita por um software externo chamado ARX.

Outra contribuição desta pesquisa foi a definição dos critérios para classificação considerando as características dos Datasets, assim, atendendo ao objetivo específico **OE1**. O uso de ontologias e Lógica de Descrição para classificação e recomendação de algoritmos de anonimização também é considerado uma contribuição, dessa forma, atendendo ao objetivo específico **OE2**. O uso de *Machine Learning* para melhorar a classificação realizada pelas ontologias é mais uma das contribuições desta Tese, assim, atende o objetivo específico **OE3**. Por fim, o processo de avaliação e validação atende ao objetivo específico **OE4**.

6.3 AMEAÇAS À VALIDAÇÃO

Ao longo do desenvolvimento desta pesquisa, foram encontradas várias dificuldades para garantir sua validade. As potenciais ameaças de validade do estudo e as estratégias para superá-las estão listadas abaixo:

- **Validade Interna:** para o desenvolvimento da aplicação, foram desenvolvidas duas ontologias. A divisão da abordagem de duas ontologias foi devido a questões relacionadas ao desempenho e velocidade de processamento da análise dos dados. Em testes iniciais com o uso de apenas uma Ontologia para classificação dos dados, foi possível observar alto consumo do poder de processamento da máquina usada, porém, quando a abordagem foi dividida em duas ontologias, não foi mais notado o alto consumo de poder de processamento. Para a primeira avaliação, foram desenvolvidos *Datasets* sintéticos, porém, a estrutura dos *Datasets* foi elaborada seguindo a mesma estrutura usada na plataforma *FIWARE*. Para validação da estrutura usada nos *Datasets*, foi criada uma prova de conceito com três cenários de uso da IoT. A prova de conceito está descrito na seção 5.2.

- **Validade de Constructo:** a escolha das métricas usadas para avaliação desta pesquisa foi baseada em trabalhos publicados na áres de Inteligência Artificial e Sistemas de Recomendação, e são métricas já validadas na área, ou seja, não foram escolhidas de forma aleatória.

- **Validade Externa:** com relação às avaliações descritas nas seções 5.1.3 e 5.2, os resultados são promissores. Porém, vale ressaltar que a pesquisa ainda está em fase experimental, possibilitando diversas oportunidades para melhorias, visto que uma das

vantagens de usar Ontologia é justamente a possibilidade de estender sua base de conhecimento. O uso de *Machine Learning* pode ser estendido, tanto com uso de outros algoritmos, quanto com o ajuste dos que estão sendo usados. Um dos próximos passos desta pesquisa é justamente avaliar o uso de novos algoritmos de *Machine Learning* e validar a solução proposta em um ambiente real.

- **Validade da Conclusão Estatística:** para a avaliação, foram utilizadas métricas amplamente aceitas e usadas na área. Foram feitas avaliações para medir o desempenho da abordagem proposta, além de estudos comparativos para identificar qual dos algoritmos de *Machine Learning* usados obtém melhores resultados quando são usados juntos com as ontologias.

6.4 PUBLICAÇÕES

Esta Tese proporcionou a publicação de dois artigos: o primeiro foi no *Workshop de Teses e Dissertações (Wtdsoft)*, e o segundo no *Journal of Computer Security*, periódico com importante fator de impacto. Essa segunda publicação ocorreu no final segundo semestre de 2022.

- Um mecanismo para recomendação de algoritmos de anonimização de dados baseados no perfil dos dados para ambientes IoT. DOI: https://doi.org/10.5753/cbsoft estendido.2021.17286. Workshop de Teses e Dissertações (Wtdsoft) - Anais Estendidos do XII Congresso Brasileiro de Software: Teoria e Prática - 2021;

- *Data Privacy in the Internet of Things Based on Anonymization: A Review. DOI: 10.3233/JCS-210089 - Journal of Computer Security, vol. Pre-press, no. Pre-press, pp. 1-31, 2023* – Disponível em: https://content.iospress.com/articles/journal-of-computer-security/jcs210089. (Qualis: A4).

Por fim, um terceiro artigo está em processo de submissão:

- *Smart Anonymity*: *A Mechanism for Recommending Data Anonymization Algorithms Based on Data Profiles for IoT Environments.* The Journal of Supercomputing, 2023 (Qualis: A2).

6.5 TRABALHOS FUTUROS

Além desta Tese apresentar as contribuições citadas na seção 6.2, outras perspectivas são possíveis. Diante disso, a seguir, são listadas algumas direções para trabalhos futuros desta pesquisa:

- Avaliar o *Smart Anonymity* usando uma maior quantidade de *Datasets*. Nesse caso será usado um maior número de *Datasets* para avaliar se os resultados vão sofrer alteração;
- Avaliar o *Smart Anonymity* usando *Datasets* gerados por ambientes IoT reais que usem a mesma estrutura adotada nesta Tese. Para esta avaliação, será necessária a criação de ambientes reais de IoT usando os 11 subdomínios da IoT descritos na seção 4.1;
- Adicionar aprendizagem automática para atualizar a base de conhecimento das ontologias, para estender a base de conhecimento das ontologias. Usar Inteligência Artificial para atualizar a base de conhecimento das ontologias à medida que novos dados sejam analisados;
- Usar a combinação de algoritmos de *Machine Learning* e avaliar seu desempenho na classificação com as ontologias. Testar a combinação de vários algoritmos de *Machine Learning* para avaliar se os resultados alcançados são melhorados em relação ao uso de apenas um algoritmos, conforme foi feito na versão atual do *Smart Anonymity*;
- Adicionar outros algoritmos de *Machine Learning* e avaliar seu desempenho na classificação junto com as ontologias, além dos algoritmos utilizados para melhorar os resultados da classificação. Existem vários outros algoritmos disponíveis na literatura que poderiam ser usados para avaliar se podem ser obtidos melhores resultados em comparação com os algoritmos que já foram avaliados;
- Avaliar o uso de outros algoritmos de anonimização no *Smart Anonymity*. Nesse caso, para expandir o *Smart Anonymity* podem ser adicionados outros algoritmos de anonimização. Para essa expansão do *Smart Anonymity*, pode ser feita uma revisão da literatura sobre algoritmos de anonimização que já são usados na área de IoT ou até mesmo algoritmos de anonimização que são usados em outras áreas;

- Fazer a integração do *Smart Anonymity* com a API do ARX. O ARX disponibiliza uma API que pode ser usada por outras aplicações, isso pode ser útil para automatizar todo o processo, desde a classificação até a anonimização dos dados; e
- Fazer uma versão do *Smart Anonymity* usando apenas *Machine Learning* e comparar os resultados com a versão atual. Atualmente, o *Smart Anonymity* utiliza ontologias juntamente com *Machine Learning* para fazer a classificação dos dados. Porém, outra possibilidade é usar apenas *Machine Learning* para fazer todo o processo. Nesse caso, é interessante ser feita uma avaliação de desempenho comparando as duas abordagens.

REFERÊNCIAS

ABOMHARA, M.; KØIEN, G. M. Security and privacy in the internet of things: Current status and open issues. In: IEEE. *2014 international conference on privacy and security in mobile systems (PRISMS)*. [S.l.], 2014. p. 1–8.

AL-FUQAHA, A.; GUIZANI, M.; MOHAMMADI, M.; ALEDHARI, M.; AYYASH, M. Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE communications surveys & tutorials*, IEEE, v. 17, n. 4, p. 2347–2376, 2015.

ALAM, M. R.; ST-HILAIRE, M.; KUNZ, T. A modular framework for cost optimization in smart grid. In: IEEE. *2014 IEEE World Forum on Internet of Things (WF-IoT)*. [S.l.], 2014. p. 337–340.

ALDEEN, Y. A. A. S.; SALLEH, M. Privacy preserving data utility mining architecture. In: *Smart Cities Cybersecurity and Privacy*. [S.l.]: Elsevier, 2019. p. 253–268.

ALGHAMDI, A.; SHETTY, S. Survey toward a smart campus using the internet of things. In: IEEE. *2016 IEEE 4th international conference on future internet of things and cloud (FiCloud)*. [S.l.], 2016. p. 235–239.

ALMEIDA, M. B.; BAX, M. P. An overview about ontologies: survey about definitions, types, applications, evaluation and building methods. *Ciência da Informação*, SciELO Brasil, v. 32, n. 3, p. 7–20, 2003.

AMARAN, M. H.; NOH, N. A. M.; ROHMAD, M. S.; HASHIM, H. A comparison of lightweight communication protocols in robotic applications. *Procedia Computer Science*, Elsevier, v. 76, p. 400–405, 2015.

ANDERSON, C. *Creating a data-driven organization: Practical advice from the trenches*. [S.l.]: "O'Reilly Media, Inc.", 2015.

ARX. *ARX – Data Anonymization Tool*. 2023. Disponível em: <https://arx.deidentifier.org/>. Acesso em: 17 de maio 2023.

ASHTON, K. et al. That 'internet of things' thing. *RFID journal*, v. 22, n. 7, p. 97–114, 2009.

ATALLAH, D. M.; BADAWY, M.; EL-SAYED, A.; GHONEIM, M. A. Predicting kidney transplantation outcome based on hybrid feature selection and knn classifier. *Multimedia Tools and Applications*, Springer, v. 78, p. 20383–20407, 2019.

BAADER, F.; CALVANESE, D.; MCGUINNESS, D.; PATEL-SCHNEIDER, P.; NARDI, D. et al. *The description logic handbook: Theory, implementation and applications*. [S.l.]: Cambridge university press, 2003.

BAADER, F.; LUTZ, C.; SUNTISRIVARAPORN, B. Cel — a polynomial-time reasoner for life science ontologies. Springer Berlin Heidelberg, Berlin, Heidelberg, p. 287–291, 2006.

BABADI, A. N.; NOURI, S.; KHALAJ, S. Challenges and opportunities of the integration of iot and smart grid in iran transmission power system. In: IEEE. *2017 Smart Grid Conference (SGC)*. [S.l.], 2017. p. 1–6.

BAJAJ, N. S.; PATANGE, A. D.; JEGADEESHWARAN, R.; PARDESHI, S. S.; KULKARNI, K. A.; GHATPANDE, R. S. Application of metaheuristic optimization based support vector machine for milling cutter health monitoring. *Intelligent Systems with Applications*, Elsevier, p. 200196, 2023.

BERREHILI, F. Z.; BELMEKKI, A. Privacy preservation in the internet of things. In: SPRINGER. *International Symposium on Ubiquitous Networking*. [S.l.], 2016. p. 163–175.

BHATTACHARYA, S.; PANDEY, M. Significance of iot in india's e-medical framework: A study. In: IEEE. *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*. [S.l.], 2020. p. 321–324.

BLASCH, E.; XU, R.; NIKOUEI, S. Y.; CHEN, Y. A study of lightweight dddas architecture for real-time public safety applications through hybrid simulation. In: IEEE. *2019 Winter Simulation Conference (WSC)*. [S.l.], 2019. p. 762–773.

BORGIA, E. The internet of things vision: Key features, applications and open issues. *Computer Communications*, Elsevier, v. 54, p. 1–31, 2014.

BORST, W. N. *Construction of engineering ontologies. 1997. 243 f.* Tese (Doutorado) — Tese (Doutorado).–University of Twente, Enschede, 1997. Disponível em:< http . . . , 2006.

BUTUN, I.; EROL-KANTARCI, M.; KANTARCI, B.; SONG, H. Cloud-centric multi-level authentication as a service for secure public safety device networks. *IEEE Communications Magazine*, IEEE, v. 54, n. 4, p. 47–53, 2016.

CEPEDA-PACHECO, J. C.; DOMINGO, M. C. Deep learning and internet of things for tourist attraction recommendations in smart cities. *Neural Computing and Applications*, Springer, v. 34, n. 10, p. 7691–7709, 2022.

CHANDRASEKARAN, B.; JOSEPHSON, J. R.; BENJAMINS, V. R. What are ontologies, and why do we need them? *IEEE Intelligent Systems and their applications*, IEEE, v. 14, n. 1, p. 20–26, 1999.

CHANG, S.; NAM, K. sook. Spatial design direction of smart home in iot paradigm. In: IEEE. *2019 2nd World Symposium on Communication Engineering (WSCE)*. [S.l.], 2019. p. 74–77.

CHIBANI, S.; COUDERT, F.-X. Machine learning approaches for the prediction of materials properties. *Apl Materials*, AIP Publishing, v. 8, n. 8, 2020.

COLLINA, M.; BARTOLUCCI, M.; VANELLI-CORALLI, A.; CORAZZA, G. E. Internet of things application layer protocol analysis over error and delay prone links. In: IEEE. *2014 7th Advanced Satellite Multimedia Systems Conference and the 13th Signal Processing for Space Communications Workshop (ASMS/SPSC)*. [S.l.], 2014. p. 398–404.

COSTA, A. F. d. Arandu, um chatbot para construção de ontologias guiado por uma ontologia de topo. Universidade Federal de Pernambuco, 2020.

CUI, Z.; XU, X.; FEI, X.; CAI, X.; CAO, Y.; ZHANG, W.; CHEN, J. Personalized recommendation system based on collaborative filtering for iot scenarios. *IEEE Transactions on Services Computing*, IEEE, v. 13, n. 4, p. 685–695, 2020.

DAGAR, R.; SOM, S.; KHATRI, S. K. Smart farming–iot in agriculture. In: IEEE. *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. [S.l.], 2018. p. 1052–1056.

DAI, C.; GHINITA, G.; BERTINO, E.; BYUN, J.-W.; LI, N. Tiamat: a tool for interactive analysis of microdata anonymization techniques. *Proceedings of the VLDB Endowment*, VLDB Endowment, v. 2, n. 2, p. 1618–1621, 2009.

DAVOLI, L.; PROTSKAYA, Y.; VELTRI, L. An anonymization protocol for the internet of things. In: IEEE. *2017 International Symposium on Wireless Communication Systems (ISWCS)*. [S.l.], 2017. p. 459–464.

DENG, A.; SHI, X. Data-driven metric development for online controlled experiments: Seven lessons learned. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2016. p. 77–86.

DESIMA, M. A.; FAISHAL, M.; PERMANA, Y. E. et al. Design of smart mobility application to realize sukabumi smart cities. In: IEEE. *2017 International Conference on Computing, Engineering, and Design (ICCED)*. [S.l.], 2017. p. 1–4.

DHOLU, M.; GHODINDE, K. Internet of things (iot) for precision agriculture application. In: IEEE. *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*. [S.l.], 2018. p. 339–342.

DRESCH, A.; LACERDA, D. P.; JÚNIOR, J. A. V. A. *Design science research: método de pesquisa para avanço da ciência e tecnologia*. [S.l.]: Bookman Editora, 2015.

DU, S.; MENG, F.; GAO, B. Research on the application system of smart campus in the context of smart city. In: IEEE. *2016 8th International Conference on Information Technology in Medicine and Education (ITME)*. [S.l.], 2016. p. 714–718.

DYBÅ, T.; DINGSØYR, T. Empirical studies of agile software development: A systematic review. *Information and software technology*, Elsevier, v. 50, n. 9-10, p. 833–859, 2008.

ELKHODR, M.; SHAHRESTANI, S.; CHEUNG, H. A review of mobile location privacy in the internet of things. In: IEEE. *2012 Tenth International Conference on ICT and Knowledge Engineering*. [S.l.], 2012. p. 266–272.

ESPÍNDOLA, R.; EBECKEN, N. On extending f-measure and g-mean metrics to multi-class problems. *WIT Transactions on Information and Communication Technologies*, WIT Press, v. 35, 2005.

FARIA, R.; BRITO, L.; BARAS, K.; SILVA, J. Smart mobility: A survey. In: IEEE. *2017 International Conference on Internet of Things for the Global Community (IoTGC)*. [S.l.], 2017. p. 1–8.

FILHO, D. M.; FREITAS, F.; OTTEN, J. Raccoon: A connection reasoner for the description logic alc. *EPiC Series in Computing*, EasyChair, v. 46, p. 200–211, 2017.

FIWARE. *Core Context Management: Processing History Management*. 2023. Disponível em: <https://fiware-tutorials.readthedocs.io/en/latest/time-series-data.html>. Acesso em: 17 de maio 2023.

FORESTIERO, A. Heuristic recommendation technique in internet of things featuring swarm intelligence approach. *Expert Systems with Applications*, Elsevier, v. 187, p. 115904, 2022.

FORKAN, A. R. M.; MONTORI, F.; GEORGAKOPOULOS, D.; JAYARAMAN, P. P.; YAVARI, A.; MORSHED, A. An industrial iot solution for evaluating workers' performance via activity recognition. In: IEEE. *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. [S.l.], 2019. p. 1393–1403.

FRIGIERI, E. P.; MAZZER, D.; PARREIRA, L. M2m protocols for constrained environments in the context of iot: A comparison of approaches. In: *International Telecommunications Symposium*. [S.l.: s.n.], 2015.

GARTNER. *IoT Security Primer: Challenges and Emerging Practices*. 2020.

GLIMM, B.; HORROCKS, I.; MOTIK, B.; STOILOS, G.; WANG, Z. Hermit: an owl 2 reasoner. *Journal of Automated Reasoning*, Springer, v. 53, n. 3, p. 245–269, 2014.

GOEL, R. et al. Flood damage analysis using machine learning techniques. *Procedia Computer Science*, Elsevier, p. 78–85, 2020.

GRUBER, T. R. Towards principles for the design of ontologies used for knowledge sharing. 1993.

GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: *Handbook on ontologies*. [S.l.]: Springer, 2009. p. 1–17.

GÄRDENFORS, P. Belief revision: A vade-mecum. Springer, 1992.

HAARSLEV, V.; MÖLLER, R. Racer system description. In: GORÉ, R.; LEITSCH, A.; NIPKOW, T. (Ed.). *Automated Reasoning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001. p. 701–705.

HARADAT, K.; OHNOT, Y.; NAKAMURAT, Y.; NISHIT, H. Anonymization method based on sparse coding for power usage data. In: IEEE. *2018 IEEE 16th International Conference on Industrial Informatics (INDIN)*. [S.l.], 2018. p. 571–576.

HARMELEN, F. V.; LIFSCHITZ, V.; PORTER, B. *Handbook of knowledge representation*. [S.l.]: Elsevier, 2008.

HEVNER, A. R.; MARCH, S. T.; PARK, J.; RAM, S. Design science in information systems research. *MIS quarterly*, JSTOR, p. 75–105, 2004.

HORRIDGE, M.; JUPP, S.; MOULTON, G.; RECTOR, A.; STEVENS, R.; WROE, C. A practical guide to building owl ontologies using protégé 4 and co-ode tools edition1. 2. *The university of Manchester*, v. 107, 2009.

HOSSAIN, I.; DAS, D.; RASHED, M. G. Internet of things based model for smart campus: Challenges and limitations. In: IEEE. *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*. [S.l.], 2019. p. 1–4.

HSU, A. P.; LEE, W.-T.; TRAPPEY, A. J.; TRAPPEY, C. V.; CHANG, A.-C. Using system dynamics analysis for performance evaluation of iot enabled one-stop logistic services. In: IEEE. *2015 IEEE International Conference on Systems, Man, and Cybernetics*. [S.l.], 2015. p. 1291–1296.

HUANG, Z.; XU, X.; NI, J.; ZHU, H.; WANG, C. Multimodal representation learning for recommendation in internet of things. *IEEE Internet of Things Journal*, IEEE, v. 6, n. 6, p. 10675–10685, 2019.

IEEE. IEEE Approved Draft Standard for an Architectural Framework for the Internet of Things (IoT). *IEEE P2413/D0.4.6, March 2019*, n. March, p. 1–265, 2019. Disponível em: <https://standards.ieee.org/content/ieee-standards/en/standard/2413-2019.html>.

ISLAM, S. R.; KWAK, D.; KABIR, M. H.; HOSSAIN, M.; KWAK, K.-S. The internet of things for health care: a comprehensive survey. *IEEE access*, IEEE, v. 3, p. 678–708, 2015.

JAHN, G. F. *Uma proposta de arquitetura para tratamento de dados não estruturados no âmbito dos institutos federais de educação*. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2017.

JAIN, R. *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. [S.l.]: john wiley & sons, 1990.

JAMALI, M. A. J.; BAHRAMI, B.; HEIDARI, A.; ALLAHVERDIZADEH, P.; NOROUZI, F.; JAMALI, M. A. J.; BAHRAMI, B.; HEIDARI, A.; ALLAHVERDIZADEH, P.; NOROUZI, F. Iot architecture. *Towards the Internet of Things: Architectures, Security, and Applications*, Springer, p. 9–31, 2020.

JEON, M.; TEMUUJIN, O.; AHN, J.; IM, D.-H. Distributed l-diversity using spark-based algorithm for large resource description frameworks data. *The Journal of Supercomputing*, Springer, p. 1–17, 2021.

JUNG, M.; KIM, J.; WI, H.; KIM, S.; KOVATSCH, M. Things–to–cloud communication: technology overview and design considerations. In: *Proceeding of IEEE 5th International Conference on the Internet of Things (IoT), Seoul, Korea*. [S.l.: s.n.], 2015. p. 1–2.

KADIYALA, E.; MEDA, S.; BASANI, R.; MUTHULAKSHMI, S. Global industrial process monitoring through iot using raspberry pi. In: IEEE. *2017 International Conference on Nextgen Electronic Technologies: Silicon to Software (ICNETS2)*. [S.l.], 2017. p. 260–262.

KANWAL, T.; ANJUM, A.; MALIK, S. U.; SAJJAD, H.; KHAN, A.; MANZOOR, U.; ASHERALIEVA, A. A robust privacy preserving approach for electronic health records using multiple dataset with multiple sensitive attributes. *Computers & Security*, Elsevier, p. 102224, 2021.

KHAN, F.; SIDDIQUI, M. A. B.; REHMAN, A. U.; KHAN, J.; ASAD, M. T. S. A.; ASAD, A. Iot based power monitoring system for smart grid applications. In: IEEE. *2020 International Conference on Engineering and Emerging Technologies (ICEET)*. [S.l.], 2020. p. 1–5.

KHAN, R.; KHAN, S. U.; ZAHEER, R.; KHAN, S. Future internet: the internet of things architecture, possible applications and key challenges. In: IEEE. *2012 10th international conference on frontiers of information technology*. [S.l.], 2012. p. 257–260.

KIRYAKOV, A.; OGNYANOV, D.; MANOV, D. Owlim–a pragmatic semantic repository for owl. In: SPRINGER. *International Conference on Web Information Systems Engineering*. [S.l.], 2005. p. 182–192.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Citeseer, 2007.

KITCHENHAM, B. A.; BUDGEN, D.; BRERETON, P. *Evidence-based software engineering and systematic reviews*. [S.l.]: CRC press, 2015. v. 4.

KJELLBY, R. A.; CENKERAMADDI, L. R.; FRØYTLOG, A.; LOZANO, B. B.; SOUMYA, J.; BHANGE, M. Long-range & self-powered iot devices for agriculture & aquaponics based on multi-hop topology. In: IEEE. *2019 IEEE 5th World Forum on Internet of Things (WF-IoT)*. [S.l.], 2019. p. 545–549.

KOMPELLA, K. *A Guide to A/B Testing Tools*. [S.l.]: ONLINE INC 213 DANBURY RD, WILTON, CT 06897-4007 USA, 2015.

KURANI, A.; DOSHI, P.; VAKHARIA, A.; SHAH, M. A comprehensive comparative study of artificial neural network (ann) and support vector machines (svm) on stock forecasting. *Annals of Data Science*, Springer, v. 10, n. 1, p. 183–208, 2023.

LAPLANTE, P. A.; LAPLANTE, N. The internet of things in healthcare: Potential applications and challenges. *It Professional*, IEEE, v. 18, n. 3, p. 2–4, 2016.

LI, C.; PALANISAMY, B. Reversible spatio-temporal perturbation for protecting location privacy. *Computer Communications*, Elsevier, v. 135, p. 16–27, 2019.

Li, T.; Li, N.; Zhang, J.; Molloy, I. Slicing: A new approach for privacy preserving data publishing. *IEEE Transactions on Knowledge and Data Engineering*, v. 24, n. 3, p. 561–574, 2012.

LIAO, D.; SUN, G.; LI, H.; YU, H.; CHANG, V. The framework and algorithm for preserving user trajectory while using location-based services in iot-cloud systems. *Cluster Computing*, Springer, v. 20, n. 3, p. 2283–2297, 2017.

LIM, Y.-s.; SRIVATSA, M.; CHAKRABORTY, S.; TAYLOR, I. Learning light-weight edge-deployable privacy models. In: IEEE. *2018 IEEE International Conference on Big Data (Big Data)*. [S.l.], 2018. p. 1290–1295.

LIU, Y.-N.; WANG, Y.-P.; WANG, X.-F.; XIA, Z.; XU, J.-F. Privacy-preserving raw data collection without a trusted authority for iot. *Computer Networks*, Elsevier, v. 148, p. 340–348, 2019.

MAHANAN, W.; CHAOVALITWONGSE, W. A.; NATWICHAI, J. Data anonymization: a novel optimal k-anonymity algorithm for identical generalization hierarchy data in iot. *Service Oriented Computing and Applications*, Springer, v. 14, n. 2, p. 89–100, 2020.

MAIER, J. Anonymity: Formalisation of privacy–k-anonymity. In: *Seminar paper, Technische Universität München, Munich*. [S.l.: s.n.], 2013.

MALCHE, T.; MAHESHWARY, P. Internet of things (iot) for building smart home system. In: IEEE. *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*. [S.l.], 2017. p. 65–70.

MALEKZADEH, M.; CLEGG, R. G.; CAVALLARO, A.; HADDADI, H. Mobile sensor data anonymization. In: *Proceedings of the International Conference on Internet of Things Design and Implementation*. [S.l.: s.n.], 2019. p. 49–58.

MARTINS, R. F. d. V. *Sistema de Recomendação de Tutoriais*. Tese (Doutorado), 2016.

MEREGE, D. A.; UEDA, E. T. Hamra—a middleware for data traffic management in public safety networks. In: IEEE. *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*. [S.l.], 2018. p. 464–469.

MURTHY, S.; BAKAR, A. A.; RAHIM, F. A.; RAMLI, R. A comparative study of data anonymization techniques. In: IEEE. *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*. [S.l.], 2019. p. 306–309.

NAGARAJA, G.; SOPPIMATH, A. B.; SOUMYA, T.; ABHINITH, A. Iot based smart agriculture management system. In: IEEE. *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*. [S.l.], 2019. v. 4, p. 1–5.

NAKAMURA, T.; NISHI, H. Tmk-anonymity: Perturbation-based data anonymization method for improving effectiveness of secondary use. In: IEEE. *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. [S.l.], 2018. p. 3138–3143.

NAYAHI, J. J. V.; KAVITHA, V. Privacy and utility preserving data clustering for data anonymization and distribution on hadoop. *Future Generation Computer Systems*, Elsevier, v. 74, p. 393–408, 2017.

NEVES, F.; SOUZA, R.; SOUSA, J.; BONFIM, M.; GARCIA, V. Data privacy in the internet of things based on anonymization: A review. *Journal of Computer Security*, IOS Press, n. Preprint, p. 1–31, 2023.

NGUYEN-HOANG, P.; VO-TAN, P. Development an open-source industrial iot gateway. In: IEEE. *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*. [S.l.], 2019. p. 201–204.

OLIVEIRA, A. B. F.; WERNECK, V. M. B. Ontologias. 2003.

ONG, A. K. S.; ZULVIA, F. E.; PRASETYO, Y. T. "the big one" earthquake preparedness assessment among younger filipinos using a random forest classifier and an artificial neural network. *Sustainability*, Multidisciplinary Digital Publishing Institute, v. 15, n. 1, p. 679, 2023.

OTGONBAYAR, A.; PERVEZ, Z.; DAHAL, K. Toward anonymizing iot data streams via partitioning. In: IEEE. *2016 IEEE 13th International conference on mobile ad hoc and sensor systems (MASS)*. [S.l.], 2016. p. 331–336.

OTGONBAYAR, A.; PERVEZ, Z.; DAHAL, K.; EAGER, S. K-varp: K-anonymity for varied data streams via partitioning. *Information Sciences*, Elsevier, v. 467, p. 238–255, 2018.

OZDEMIR, M. E.; ALI, Z.; SUBESHAN, B.; ASMATULU, E. Applying machine learning approach in recycling. *Journal of Material Cycles and Waste Management*, Springer, v. 23, p. 855–871, 2021.

PEARSON, K. On the probability that two independent distributions of frequency are really samples of the same population, with special reference to recent work on the identity of trypanosome strains. *Biometrika*, JSTOR, v. 10, n. 1, p. 85–143, 1914.

PEKAR, A.; MOCNEJ, J.; SEAH, W. K.; ZOLOTOVA, I. Application domain-based overview of iot network traffic characteristics. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 53, n. 4, p. 1–33, 2020.

PIMENTEL, M.; FILIPPO, D.; SANTOS, T. M. Design science research: pesquisa científica atrelada ao design de artefatos. *RE@ D-Revista de Educação a Distância e Elearning*, v. 3, n. 1, p. 37–61, 2020.

POULIS, G.; GKOULALAS-DIVANIS, A.; LOUKIDES, G.; SKIADOPOULOS, S.; TRYFONOPOULOS, C. Secreta: A system for evaluating and comparing relational and transaction anonymization algorithms. OpenProceedings. org, University of Konstanz, 2014.

PRASSER, F.; EICHER, J.; SPENGLER, H.; BILD, R.; KUHN, K. A. Flexible data anonymization using arx—current status and challenges ahead. *Software: Practice and Experience*, Wiley Online Library, v. 50, n. 7, p. 1277–1304, 2020.

Protégé . *Protégé*. 2023. Disponível em: <https://protege.stanford.edu/software.php>. Acesso em: 30 de julho de 2023.

PURI, V.; KAUR, P.; SACHDEVA, S. Data anonymization for privacy protection in fog-enhanced smart homes. In: IEEE. *2020 6th International Conference on Signal Processing and Communication (ICSC)*. [S.l.], 2020. p. 201–205.

QI, L.; LIU, Y.; ZHANG, Y.; XU, X.; BILAL, M.; SONG, H. Privacy-aware point-of-interest category recommendation in internet of things. *IEEE Internet of Things Journal*, IEEE, v. 9, n. 21, p. 21398–21408, 2022.

QUIST-APHETSI, K.; XENYA, M. C. Securing medical iot devices using diffie-hellman and des cryptographic schemes. In: IEEE. *2019 International Conference on Cyber Security and Internet of Things (ICSIoT)*. [S.l.], 2019. p. 105–108.

RAMOS, F. E. d. C. S. *Anonimização Automática de Dados Estruturados*. Tese (Doutorado), 2022.

RAY, A. K.; BAGWARI, A. Iot based smart home: Security aspects and security architecture. In: IEEE. *2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT)*. [S.l.], 2020. p. 218–222.

RODRIGUEZ-GARCIA, M.; CIFREDO-CHACÓN, M.-á.; QUIRÓS-OLOZÁBAL, Á. Cooperative privacy-preserving data collection protocol based on delocalized-record chains. *IEEE Access*, IEEE, v. 8, p. 180738–180749, 2020.

SALMAN, L.; SALMAN, S.; JAHANGIRIAN, S.; ABRAHAM, M.; GERMAN, F.; BLAIR, C.; KRENZ, P. Energy efficient iot-based smart home. In: IEEE. *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*. [S.l.], 2016. p. 526–529.

SAMANI, A.; GHENNIWA, H. H.; WAHAISHI, A. Privacy in internet of things: A model and protection framework. In: *ANT/SEIT*. [S.l.: s.n.], 2015. p. 606–613.

SANTOS, B. P.; SILVA, L.; CELES, C.; BORGES, J. B.; NETO, B. S. P.; VIEIRA, M. A. M.; VIEIRA, L. F. M.; GOUSSEVSKAIA, O. N.; LOUREIRO, A. Internet das coisas: da teoriaa prática. *Minicursos SBRC-Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuıdos*, v. 31, 2016.

SAPUTRA, R. H.; SURYANEGARA, M. On developing the model of blockchain technology for logistic services in indonesia. In: IEEE. *2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129)*. [S.l.], 2019. p. 269–274.

SASTRA, N. P.; WIHARTA, D. M. Environmental monitoring as an iot application in building smart campus of universitas udayana. In: IEEE. *2016 International Conference on Smart Green Technology in Electrical and Information Systems (ICSGTEIS)*. [S.l.], 2016. p. 85–88.

SCHREIBER, M.; NADAL, Z. I.; VALENTE, S. A. Smart grid tools: Iot device-managed power storage through local consumer demand control. In: IEEE. *2019 IEEE PES Innovative Smart Grid Technologies Conference-Latin America (ISGT Latin America)*. [S.l.], 2019. p. 1–6.

SCHURGOT, M. R.; SHINBERG, D. A.; GREENWALD, L. G. Experiments with security and privacy in iot networks. In: IEEE. *2015 IEEE 16th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. [S.l.], 2015. p. 1–6.

SHOHATA, S.; NAKAMURA, Y.; NISHI, H. Hardware for accelerating anonymization transparent to network. In: IEEE. *2018 Sixth International Symposium on Computing and Networking (CANDAR)*. [S.l.], 2018. p. 181–187.

SHU-WEN, W. Research on the key technologies of iot applied on smart grid. In: IEEE. *2011 International Conference on Electronics, Communications and Control (ICECC)*. [S.l.], 2011. p. 2809–2812.

SILVA, H. d. O. et al. Uma abordagem baseada em anonimização para privacidade de dados em plataformas analíticas. [sn], 2019.

SIRIN, E.; PARSIA, B.; GRAU, B. C.; KALYANPUR, A.; KATZ, Y. Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, v. 5, n. 2, p. 51 – 53, 2007. ISSN 1570-8268. Software Engineering and the Semantic Web.

STENERSEN, H. W. *Anonymization of Health Data*. Dissertação (Mestrado), 2020.

SUNDHARAM, S. M.; FEJOZ, L.; NAVET, N. Connected motorized riders—a smart mobility system to connect two and three-wheelers. In: IEEE. *2016 Sixth International Symposium on Embedded Computing and System Design (ISED)*. [S.l.], 2016. p. 345–348.

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

TAKBIRI, N.; LI, K.; PISHRO-NIK, H.; GOECKEL, D. L. Statistical matching in the presence of anonymization and obfuscation: Non-asymptotic results in the discrete case. In: IEEE. *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*. [S.l.], 2018. p. 1–6.

THIRUMALAISAMY, M.; BASHEER, S.; SELVARAJAN, S.; ALTHUBITI, S. A.; ALENEZI, F.; SRIVASTAVA, G.; LIN, J. C.-W. Interaction of secure cloud network and crowd computing for smart city data obfuscation. *Sensors*, MDPI, v. 22, n. 19, p. 7169, 2022.

ToolBox. *UTD Anonymization ToolBox*. 2023. Disponível em: <http://www.cs.utdallas.edu/dspl/cgi-bin/toolbox/index.php>. Acesso em: 30 de julho de 2023.

TSARKOV, D.; HORROCKS, I. Fact++ description logic reasoner: System description. In: FURBACH, U.; SHANKAR, N. (Ed.). *Automated Reasoning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 292–297.

ULLAH, I.; SHAH, M. A. A novel model for preserving location privacy in internet of things. In: IEEE. *2016 22nd International conference on automation and computing (ICAC)*. [S.l.], 2016. p. 542–547.

VIEIRA, R.; SANTOS, D. A. dos; SILVA, D. M. da; SANTANA, M. R. W eb semântica: ontologias, lógica de descrição e inferência. *Web e Multimidia: Desafios e Soluções (WebMedia 2005-Minicursos)*, v. 1, p. 127–167, 2005.

WANG, G.; NIXON, M.; BOUDREAUX, M. Toward cloud-assisted industrial iot platform for large-scale continuous condition monitoring. *Proceedings of the IEEE*, IEEE, v. 107, n. 6, p. 1193–1205, 2019.

WANG, K.; WANG, Y.; SUN, Y.; GUO, S.; WU, J. Green industrial internet of things architecture: An energy-efficient perspective. *IEEE Communications Magazine*, IEEE, v. 54, n. 12, p. 48–54, 2016.

WANG, P.; LU, X.; SUN, H.; LV, W. Application of speech recognition technology in iot smart home. In: IEEE. *2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*. [S.l.], 2019. p. 1264–1267.

WU, M.; LU, T.-J.; LING, F.-Y.; SUN, J.; DU, H.-Y. Research on the architecture of internet of things. In: IEEE. *2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)*. [S.l.], 2010. v. 5, p. V5–484.

XU, R.; NIKOUEI, S. Y.; CHEN, Y.; BLASCH, E.; AVED, A. Blendmas: A blockchain-enabled decentralized microservices architecture for smart public safety. In: IEEE. *2019 IEEE International Conference on Blockchain (Blockchain)*. [S.l.], 2019. p. 564–571.

YADAV, V.; BORATE, S.; DEVAR, S.; GAIKWAD, R.; GAVALI, A. Smart home automation using virtue of iot. In: IEEE. *2017 2nd International Conference for Convergence in Technology (I2CT)*. [S.l.], 2017. p. 313–317.

ZENKERT, J.; DORNHOFER, M.; WEBER, C.; NGOUKAM, C.; FATHI, M. Big data analytics in smart mobility: Modeling and analysis of the aarhus smart city dataset. In: IEEE. *2018 IEEE Industrial Cyber-Physical Systems (ICPS)*. [S.l.], 2018. p. 363–368.

ZHAO, J.-c.; ZHANG, J.-f.; FENG, Y.; GUO, J.-x. The study and application of the iot technology in agriculture. In: IEEE. *2010 3rd International Conference on Computer Science and Information Technology*. [S.l.], 2010. v. 2, p. 462–465.

ZHOU, C.; ZHANG, X. Toward the internet of things application and management: A practical approach. In: IEEE. *Proceeding of IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks 2014*. [S.l.], 2014. p. 1–6.

ZIEGELDORF, J. H.; MORCHON, O. G.; WEHRLE, K. Privacy in the internet of things: threats and challenges. *Security and Communication Networks*, Wiley Online Library, v. 7, n. 12, p. 2728–2742, 2014.

ZUO, Z.; WATSON, M.; BUDGEN, D.; HALL, R.; KENNELLY, C.; MOUBAYED, N. A. Data anonymization for pervasive health care: systematic literature mapping study. *JMIR medical informatics*, JMIR Publications Toronto, Canada, v. 9, n. 10, p. e29871, 2021.