MINISTÉRIO DA DEFESA
EXÉRCITO BRASILEIRO
DEPARTAMENTO DE CIÊNCIA E TECNOLOGIA
INSTITUTO MILITAR DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS E COMPUTAÇÃO

MADALENA LOPES E SILVA

SEC4ML: ANONIMIZAÇÃO DE DADOS DE INCIDENTES DE SEGURANÇA DA INFORMAÇÃO PARA TAREFAS DE APRENDIZADO DE MÁQUINA

RIO DE JANEIRO
2022

MADALENA LOPES E SILVA

SEC4ML: ANONIMIZAÇÃO DE DADOS DE INCIDENTES DE SEGURANÇA DA INFORMAÇÃO PARA TAREFAS DE APRENDIZADO DE MÁQUINA

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

Orientador(es): Maria Cláudia Reis Cavalcanti, D.Sc.
Kelli de Faria Cordeiro, D.Sc.

Rio de Janeiro

2022

INSTITUTO MILITAR DE ENGENHARIA
Praça General Tibúrcio, 80 – Praia Vermelha
Rio de Janeiro – RJ CEP: 22290-270

Lopes e Silva, Madalena.
Sec4ML: anonimização de dados de incidentes de segurança da informação para tarefas de aprendizado de máquina / Madalena Lopes e Silva. – Rio de Janeiro, 2022.
112 f.

Orientador(es): Maria Cláudia Reis Cavalcanti e Kelli de Faria Cordeiro.

Dissertação (mestrado) – Instituto Militar de Engenharia, Sistemas e Computação, 2022.

1. segurança da informação. 2. anonimização. 3. dados ligados. 4. aprendizado de máquina. 5. inteligência artifical. 6. princípios FAIR. i. Reis Cavalcanti, Maria Cláudia (orient.) ii. de Faria Cordeiro, Kelli (orient.) iii. Título

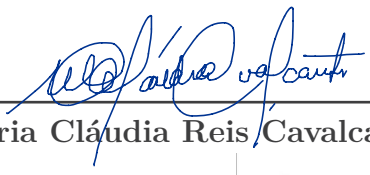
MADALENA LOPES E SILVA

Sec4ML: anonimização de dados de incidentes de segurança da informação para tarefas de aprendizado de máquina

Dissertação apresentada ao Programa de Pós-graduação em Sistemas e Computação do Instituto Militar de Engenharia, como requisito parcial para a obtenção do título de Mestre em Ciências em Sistemas e Computação.

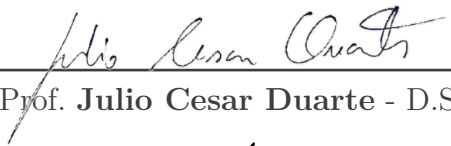
Orientador(es): Maria Cláudia Reis Cavalcanti e Kelli de Faria Cordeiro.

Aprovado em Rio de Janeiro, 12 de julho de 2022, pela seguinte banca examinadora:



Profa. **Maria Cláudia Reis Cavalcanti** - D.Sc. do IME - Presidente

Profa. **Kelli de Faria Cordeiro** - D.Sc. do CASNAV



Prof. **Julio Cesar Duarte** - D.Sc. do IME

Prof. **João Luiz Rebelo Moreira** - Ph.D. da *University of Twente*

Rio de Janeiro

2022

Dedico esta dissertação ao meu marido Alexandre e meus filhos Otávio Henrique e Cesar Augusto, razões da minha vida.

AGRADECIMENTOS

O primeiro agradecimento é a Deus, nosso mestre supremo que me permitiu chegar até aqui. Agradeço profundamente às minhas Orientadoras, professora Maria Claudia e professora Kelli, por todo o apoio, ensinamentos e orientações que recebi. O comprometimento e apoio demonstrados foram fundamentais para o sucesso deste trabalho. Sem essas orientações também não seria possível ter chegado até aqui. Agradeço à todos os professores do IME e à professora Maria Luiza da UFRJ pelos valiosos ensinamentos, troca de experiências e orientação. Agradeço também ao meu marido Alexandre, meu principal incentivador que acreditou sempre que seria possível chegar mais longe, alcançar o que muitas vezes parecia impossível. Aos meus filhos, os meus agradecimentos pelo apoio incondicional e por tolerarem e compreenderem meus períodos de ausência. Por fim, agradeço à todos os meus amigos e colegas de estudo e trabalho que, em algum momento, notadamente os mais difíceis, tiveram uma palavra amiga, conselheira e apoiadora ou me proporcionaram ajuda, tão importantes nesses momentos de dificuldade.

"Sweet dreams are made of this
Who am I to disagree?
I travel the world and the seven seas
Evebody´s looking for something"
Sweet Dreams - Eurythmics

".. I´m gonna have myself a real good time
.. So, (don´t stop me now)
´Cause I´m having a good time
.. I´m travelling at the speed of light"
Don´t Stop Me Now - Queen

RESUMO

Apesar do crescimento exponencial da World Wide Web desde sua criação, ainda há poucos conjuntos de dados disponíveis de incidentes de cibersegurança a serem reutilizados devido a várias questões, tais como preocupações de preservação da privacidade e padronização do formato de publicação de dados. Como resultado, a análise de incidentes de domínio tem um impacto precário no desenvolvimento de Sistemas de Detecção de Intrusão (IDS). As práticas LOD (Linked Open Data), que permitem o compartilhamento de dados na Web como um grafo de dados grande e interligado, juntamente com os princípios FAIR (Findable, Accessible, Interoperable, and Reusable), que orientam a publicação de dados para reutilização, podem apoiar o compartilhamento de conjuntos de dados de incidentes de segurança cibernética. Ademais, técnicas de anonimização podem ser usadas para lidar com preocupações de privacidade. Além disso, as técnicas de Aprendizado de Máquina (AM) podem ser usadas para melhorar a eficácia do IDS. Este trabalho propõe a abordagem Sec4ML que apoia a preparação de conjuntos de dados de incidentes de cibersegurança para técnicas de AM usando práticas LOD e seguindo os princípios FAIR, envolvendo, entre outros, subprocessos de anonimização e pré-processamento, que são ilustrados usando dados de conjuntos de dados públicos.

Palavras-chave: segurança da informação. anonimização. dados ligados. aprendizado de máquina. inteligência artifical. princípios FAIR.

ABSTRACT

Despite the exponential growth of the World Wide Web since its creation, there are still few available datasets of cybersecurity incidents to be reused due to several issues, such as privacy-preserving concerns and data publication format standardization. As a result, the domain incidents analysis are precarious impacting on the Intrusion Detection Systems (IDS) development. The LOD (Linked Open Data) practices, which allows the sharing of data on the Web as a large and interconnected data graph, together with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, which guides the publication of data for reuse, can support the sharing of cybersecurity incidents datasets. Furthermore, anonymization techniques can be used to handle privacy concerns. Moreover, Machine Learning (ML) techniques can be used to improve IDS effectiveness. This work proposes the Sec4ML approach which supports the preparation of cybersecurity incident datasets for ML techniques using LOD practices and following FAIR principles, involving, among others, anonymization and preprocessing subprocesses, which are illustrated using public datasets.

Keywords: cybersecurity. anonymization. linked data. machine learning. artificial intelligence. FAIR principles.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, foi possível notar um aumento considerável no uso de tecnologias de comunicação, seja de novos modelos de negócios através de aplicações ou de novas plataformas, como a Internet das Coisas (IoT), aplicações móveis, indústria 4.0 e redes sociais. Consequentemente, houve um aumento na quantidade de dados gerados por estas tecnologias. Este novo cenário trouxe novas ameaças e vulnerabilidades, tais como vazamento de dados ou invasão de privacidade.

Mesmo com o crescimento exponencial que a rede mundial de computadores apresentou desde a sua criação, ainda há poucos conjuntos de dados disponíveis de incidentes de segurança a serem reutilizados. Uma das principais razões para isso é a preocupação com o compartilhamento de dados de segurança que está relacionada aos desafios de preservação da privacidade e à padronização do formato de publicação de dados. Assim, a escassez destes dados tem um impacto negativo no desenvolvimento e ajuste dos Sistemas de Detecção de Intrusão (IDS) e Sistemas de Prevenção de Intrusão (IPS).

Dentre os principais desafios que o domínio de conhecimento de segurança da informação enfrenta para compartilhar dados podemos destacar (1):

- Não há mecanismos de larga escala que possibilitem a troca de informações;
- Fontes de dados diferentes podem conter dados inconsistentes em alguns casos;
- Existe dificuldade, em alguns casos, de acessar a informação que se deseja, armazenada em larga escala, quer seja em repositórios na Internet ou contida em ferramentas específicas; e
- Muitos protocolos e mecanismos de acesso são proprietários e não interoperáveis.

A nuvem *Linked Open Data* (LOD) [1] (2) surgiu para promover e acelerar o compartilhamento de dados na web, como um grande grafo de dados interconectado. O uso de recursos e formatos semânticos padronizados facilitou a interligação entre os recursos de dados, fornecendo uma melhor maneira de publicar e reutilizar os dados para a pesquisa.

Posteriormente, Wilkinson et al. (3) propuseram os princípios FAIR (*Findable, Accessible, Interoperable e Reusable*). Este conjunto de premissas tem como objetivo garantir que os processos de publicação de dados de pesquisa e metadados agreguem valor, e possibilitem a transparência e a reprodutibilidade, além de mitigar os desafios supracitados. A implementação desses princípios norteadores em projetos de pesquisa e

[1] https://lod-cloud.net/

produção de conteúdo no contexto acadêmico visa: (i) facilitar a descoberta, oferecendo mecanismos de larga escala que possibilitem a troca de informações; (ii) possibilitar o reúso, melhorando a quantidade e qualidade das fontes de dados diferentes disponíveis, da maneira mais fácil possível; e (iii) tornar o conteúdo produzido legível e processável por máquina, reduzindo ou eliminando a necessidade de uso de muitos protocolos e mecanismos de acesso proprietários e não interoperáveis.

Mesmo com o expressivo pioneirismo e crescimento da web semântica voltada para algumas áreas específicas do conhecimento como a de governo e de redes sociais (4), outras áreas também têm demonstrado significativa demanda por disponibilização de *datasets* para fins de pesquisa. Especificamente na área de segurança da informação, há diversos trabalhos evidenciando a necessidade e importância de *datasets* disponíveis e atualizados para o desenvolvimento de modelos de *Aprendizado de Máquina* (AM) e treinamento dos IDS baseados nessa estratégia (5, 6, 7).

Além disso, as técnicas e modelos de AM estão evoluindo para facilitar e acelerar a descoberta do conhecimento, análises de dados de domínio e revelar novos informações obtidas através de inferências, especificamente no domínio de segurança da informação (7, 8, 9), demonstrando que o uso de AM para aumentar a eficácia dos IDS e IPS já é uma realidade.

1.1 Motivação

Dados do "Relatório sobre o prejuízo de um vazamento de dados" [2], mostram que 44% dos vazamentos de dados incluíam informações identificáveis de clientes ou usuários e 28% incluíam informações identificáveis de clientes ou usuários criptografados. De acordo com esse relatório, a quantia de $4,24 milhões de dólares foi o resultado do custo médio global com vazamento de dados. Dados do *Surf Shark*[3] demonstram que o Brasil está em 6^{th} no ranking de vazamentos por país.

Casos concretos ocorridos recentemente ilustram o que os dados estatísticos apontam. Um vazamento de 220 Mi de registros de brasileiros expôs seus dados biográficos, bancários, de crédito e foto, entre outros dados[4]. Outro episódio de vazamento, mas dessa vez envolvendo dados sensíveis de sistemas de saúde de responsabilidade do Ministério da Saúde (MS), expôs dados de 16 Mi de pacientes de COVID-19[5]. O ambiente do judiciário também foi vítima desse tipo de crime, motivado por um ataque de *ransomware*, quando o Supremo Tribunal de Justiça (STJ) teve o acesso a seus sistemas indisponível, bloqueando

[2] https://www.ibm.com/br-pt/security/data-breach
[3] https://surfshark.com/blog/data-breach-statistics-by-country-in-2021
[4] https://tecnoblog.net/404838/ exclusivo-vazamento-que-expos-220-milhoes-de-brasileiros-e-pior-do-que-se-pensava/
[5] https://www.istoedinheiro.com.br/vazamento-expoe-dados-de-16-milhoes-de-pacientes-de-covid-19/

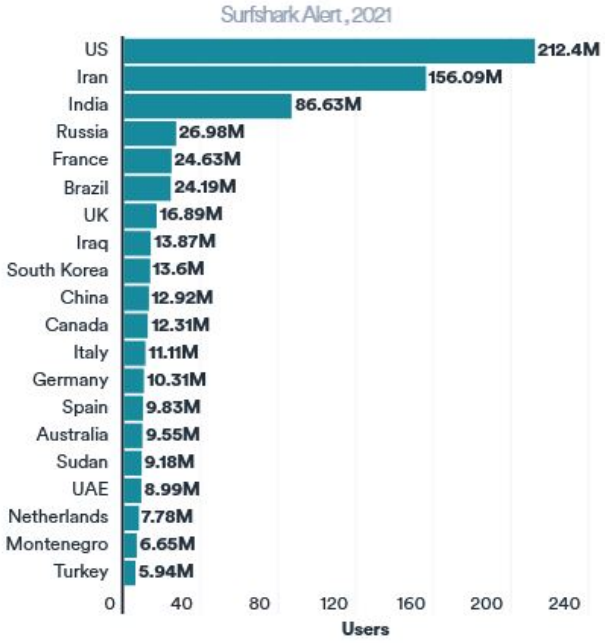


Figura 1 – Estatística de Vazamentos de Dados por País. Fonte: (10).

acesso aos processos e *e-mails* da corte[6]. Indisponibilizados por um ataque hacker[7], o site do MS e a página e o aplicativo do ConecteSUS ficaram indisponíveis. Recentes vazamentos de dados relativos ao sistema de transferência de valores PIX, comunicados pelo Banco Central,[8,9] reforçam a tendência apontada pelos dados estatísticos.

Episódios de vazamentos de dados são uma realidade e para combatê-los é necessário, dentre outras ações, o desenvolvimento e aperfeiçoamento de ferramentas de IDS. Entretanto, para isso é preciso a obtenção de conjuntos de dados de segurança da informação disponíveis para reúso. Um estudo da ENISA (*The European Network and Information Security Agency*) ressalta a necessidade de cooperação e troca de informações de incidentes de segurança da informação além das fronteiras geográficas (11). No cenário atual relativo à publicação de dados, as organizações ainda enfrentam um grande desafio nesta questão pois continuam relutantes em compartilhar dados sobre seus incidentes (12). Trabalhos como o de Oliveira et al. (13) expressam haver dificuldade em obter esta categoria de dados, o que torna quase impossível reproduzir e comparar resultados, dificultando o desenvolvimento de pesquisas e ferramentas de combate e análise de incidentes de segurança da informação. Ademais, para pesquisa e desenvolvimento de ferramentas e algoritmos de um modo geral, os dados disponíveis ou estão em formato que dificulta

[6] https://www.techtudo.com.br/listas/2020/11/ ataque-hacker-ao-stj-seis-coisas-que-voce-precisa-saber-sobre-o-caso.ghtml

[7] https://agenciabrasil.ebc.com.br/saude/noticia/2021-12/sites-e-aplicativo-do-ministerio-da-saude-sofrem-ataque-cibernetico

[8] https://agenciabrasil.ebc.com.br/economia/noticia/2022-01/bc-comunica-vazamento-de-dados-de-1601-mil-chaves-pix

[9] https://agenciabrasil.ebc.com.br/economia/noticia/2022-02/bc-comunica-vazamento-de-dados-de-21-mil-chaves-pix

a manipulação e transformação requeridas em ambientes de pesquisa ou não são disponibilizados publicamente e gratuitamente para uso. No Brasil, o cenário encontrado não é muito promissor. Dentre as iniciativas elencadas pela ENISA (11), não foi encontrada nenhuma no território brasileiro.

O dados compartilhados nesse domínio da informação podem se apresentar com uma grande diversidade. Para que seja possível o reúso é desejável a padronização destes dados. Até onde foi possível investigar, não foi proposto um padrão a ser seguido para a geração e compartilhamento de dados de incidentes de segurança da informação no âmbito global. Entretanto, iniciativas de definição de vocabulários comuns a esse domínio de conhecimento já demonstram a intenção, mesmo que incipiente, de construção de um ambiente conceitualmente comum e abrangente que atenda às necessidades do setor.

Há também outros aspectos relevantes a serem considerados na gestão da publicação e reúso de dados para pesquisas ou desenvolvimento e aperfeiçoamento de ferramentas. Questões relativas à conformidade legal recentemente vem ganhado força motivadas, principalmente, pela possibilidade de aplicação de consideráveis multas e responsabilização penal do *Data Protection Officer* (DPO), papel definido nas principais legislações de proteção de dados no mundo, como a *General Data Protection Regulation* (GDPR[10]) na Europa, e a Lei Geral de Proteção de Dados (LGPD[11]) no Brasil. Outras obrigatoriedades derivadas das legislações citadas, como auditoria e direito do titular dos dados a ser informado quando ocorrer vazamento de seus dados, criam um ambiente mais protetivo para o indivíduo que por ventura venha a sofrer com algum incidente de segurança da informação. Por outro lado, geram demandas até então negligenciadas ou atendidas de maneira precária devido à inexistência de regulamentações fortes de proteção de dados, fazendo com que as organizações se vejam hoje enfrentando desafios para a mitigação de problemas e prejuízos derivados de violações de privacidade ou de segurança.

A despeito das dificuldades citadas na geração, publicação e reúso de dados de segurança da informação, tem-se observado cada vez mais o uso de ferramentas de Inteligência Artificial (IA) na análise de dados e inferência de informações em diversos domínios de conhecimento. Especificamente no domínio de segurança da informação, há diversos trabalhos (8, 7, 14) demonstrando uma realidade diferente, em que novas formas de gerir a segurança da informação demandam novas abordagens que procuram solucionar as dificuldades supracitadas.

[10] https://gdprinfo.eu/
[11] http://www.planalto.gov.br/ccivil_03/_ato20152018/2018/lei/l13709.htm

1.2 Caracterização do Problema

A realidade da sociedade hoje nos apresenta um crescente aumento na geração de dados a partir de diversas tecnologias. Apesar desse aumento, ainda há escassez de conjuntos de dados disponíveis para aperfeiçoamento das ferramentas de IDS e IPS (7). Essa lacuna é ainda mais evidente quando leva-se em consideração a necessidade de obtenção de conjuntos de dados para a criação de modelos de AM com o mesmo objetivo de aperfeiçoamento.

Outro aspecto que dificulta a busca por soluções para o cenário descrito é a obrigatoriedade de estar em conformidade com as principais legislações de proteção de dados hoje existentes, como por exemplo a GDPR e LGPD. Tais legislações visam, principalmente, a proteção dos interesses dos indivíduos e/ou organizações quando na condição de titulares de dados. Dentre os direitos assegurados nestas legislações, por exemplo, está o direito de demandar a exclusão dos seus dados de determinada base de dados ou sistema. Entretanto, para que seja possível atender tal solicitação, os sistemas e/ou bases de dados necessitaram de adaptações em seus processos. Além disso, os sistemas, bases e conjuntos de dados gerados e tratados a partir da vigência de tais legislações devem ser projetados, desenvolvidos e processados sob um novo paradigma de conformidade com as exigências legais.

O emprego de processos de anonimização dos dados pode auxiliar no apoio à publicação de conjuntos de dados de segurança da informação na medida em que possibilita tornar esses conjuntos de dados conformes com legislações de proteção de dados vigentes, através da inclusão de técnicas de anonimização nos processos de geração e publicação destes dados.

Ainda que seja possível encontrar conjuntos de dados sobre segurança da informação disponíveis publicamente para reúso, estes não são, em sua maioria, preparados para o processamento com técnicas de classificação. Assim, uma alternativa para possibilitar a geração e publicação de conjuntos de dados prontos para processamento de AM seria o desenvolvimento de ferramentas e processos de publicação de conjuntos de dados que tenham sido tratados por operadores de *Knowledge Discovery in Databases* (KDD). Moura (15) propôs uma ferramenta que auxilia cientistas de dados a submeter os dados a operadores de pré-processamento tornando assim os conjuntos de dados trabalhados prontos para tarefas de classificação. Entretanto, neste trabalho não foram contempladas questões de preservação da privacidade.

Outra possibilidade que já é uma realidade é a disponibilização de dados na LOD. Dados triplificados publicados na LOD podem ser encontrados, consultados e processados por máquinas. Essa característica torna-se um diferencial no contexto do processamento de dados, possibilitando que sejam feitos outros usos que não seriam possíveis de realizar

caso os conjuntos de dados disponibilizados estejam somente em formato tabular.

Desta forma, as questões de pesquisa que se busca responder são: (i) De que forma é possível apoiar a publicação de dados anonimizados para que possam ser usados para a criação de modelos de AM para tarefas de classificação? (ii) O resultado da solução proposta pode ser usado para avaliar se há perdas significativas no desempenho da criação de modelos de AM quando os dados são anonimizados? (iii) A captura de proveniência pode atender os princípios FAIR?

1.3 Objetivo

O objetivo geral desta dissertação é propor uma abordagem que possibilite submeter os dados dentro de um processo de *Extract, Transform and Load* (ETL) a mecanismos de anonimização, de maneira a garantir o anonimato das entidades geradoras desses dados e a operadores de KDD, tornando os dados processados preparados para uso em tarefas supervisionadas de classificação.

Dentro desta abordagem, são objetivos específicos:

- Desenvolver um processo que contemple mecanismos:
  - de anonimização dos dados a fim de prover privacidade às entidades envolvidas;
  - de pré-processamento voltados à preparação dos conjuntos de dados para tarefas supervisionadas de classificação; e
  - de captura de metadados de proveniência que garantam a reprodutibilidade das pesquisas, conformidade com legislações existentes e reidentificação de indivíduos ou entidades envolvidas, caso necessário.
- Definir uma ontologia que melhor atenda às necessidades do domínio de conhecimento;
- Desenvolver um catálogo de dados que disponibilize na WEB de dados, seguindo princípios FAIR, metadados relativos aos conjuntos de dados publicados no âmbito da segurança da informação; e
- Realizar experimentos de criação de modelos de AM a fim de demonstrar o impacto sobre os dados utilizados em casos de uso.

1.4 Metodologia de Pesquisa

Com o objetivo de buscar soluções de abordagens para a anonimização e publicação de dados de pesquisa já propostas, foram realizadas, inicialmente, pesquisas sobre técnicas

de anonimização, estratégias e abordagens existentes de publicação de dados e metodologias e ontologias usadas na captura e persistência de dados de proveniência.

Foi utilizada a ferramenta Harzing´s Publish or Perish para auxílio nestas pesquisas, a qual possibilita a realização de consultas às seguintes bases de pesquisa: *Google Scholar, Google Scholar Profile, Scopus, Crossref, PubMed, OpenAlex, Semantic Scholar e Web of Science*. Além destas bases de pesquisa, foram utilizados também os canais de pesquisa *IEEE Explore* e *ACM Digital Library*. Os trabalhos pesquisados englobam os diretamente relacionados ao tema desta dissertação, os motivacionais, posicionais e de *survey*.

Procurou-se para os diretamente relacionados à esta dissertação dentro da janela temporal a partir de 2014 até os dias atuais. Já os de *survey* foram buscados dentro da janela de tempo a partir de 2018. Para os demais, que contivessem conceitos relativos ao presente trabalho, foram buscados dentro de uma janela de tempo mais complacente como por exemplo a partir de 2000. Foram utilizados termos de buscas como: *semantic web, privacy preserving, criptography, dataset, cybersecurity, AM operators, anonimyzation e intrusion prevention systems*.

Com base nos resultados dessas pesquisas, foram verificados as técnicas, padrões e estratégias de processamento mais adequados para alcançar o objetivo do trabalho. Após delineado o processo a ser apresentado na abordagem proposta, foram definidas as ferramentas, ambiente de implementação e linguagem de programação adotada nas tarefas específicas de processamento.

Ao final, a abordagem proposta foi aplicada em conjuntos de dados públicos (16) (17) com o objetivo de verificar a sua eficácia quanto a demandas por metadados sobre os dados, quanto à reidentificação de indivíduos, e quanto a seu impacto no desempenho dos processos de AM. Para avaliar esse impacto, foram realizados processamentos de AM sobre os dados resultantes da aplicação da abordagem, e a verificação de alguns parâmetros de desempenho do processamento. Com base nestes processamentos, é possível realizar o comparativo da aplicabilidade em tarefas de AM para classificação com os conjuntos de dados brutos em relação com a aplicabilidade dos dados anonimizados e pré-processados.

1.5 Contribuições Esperadas

Com base nos objetivos especificados e nos problemas evidenciados, as contribuições esperadas para este trabalho são:

(i) Definição de uma ontologia leve (18) que melhor atenda às necessidades de captura e persistência de dados de proveniência;

(ii) Desenvolvimento de um processo que contemple:

(I) mecanismos de anonimização dos dados tratados a fim de prover privacidade às entidades envolvidas;

(II) mecanismos de captura de metadados de proveniência que garantam a reprodutibilidade das pesquisas, conformidade com legislações existentes e facilidade de reuso.

(iii) Implementação de um FAIR *Data Point* que disponibilize na WEB metadados sobre dados relativos à segurança da informação.

1.6 Organização do Trabalho

Este trabalho está dividido em 7 capítulos. Além do corrente Capítulo com a introdução, motivação, caracterização do problema, objetivo, metodologia de pesquisa e contribuições esperadas, estão presentes no Capítulo 2 o *background* necessário ao entendimento deste trabalho. No Capítulo 3 são apresentados os trabalhos relacionados que dão suporte à abordagem proposta. O Capítulo 4 apresenta detalhadamente a abordagem proposta, com suas arquitetura e processo. A implementação da abordagem é apresentada no Capítulo 5. No Capítulo 6 são apresentados dois casos de uso da aplicação da abordagem. Finalmente, a conclusão é apresentada no Capítulo 7.

2 BACKGROUND

Para o perfeito entendimento da abordagem proposta nesta dissertação, se faz necessário a apresentação de alguns conceitos relacionados. A subseções a seguir abordarão alguns desses conceitos.

2.1 Segurança da Informação

Segurança da informação consiste em um conjunto de mecanismos de segurança que podem ser usados para proteger o espaço virtual e os seus ativos contra acessos não autorizados e ataques. O principal objetivo de sistemas de defesas do espaço virtual é manter a integridade, disponibilidade e confiabilidade dos dados de uma organização ou indivíduo. (7).

O termo segurança da informação foi formalmente definido pela ISO/IEC 27032:2012 como sendo a preservação da confidencialidade, integridade e disponibilidade da informação no ambiente virtual, também chamado de princípio CIA (19).

Para se alcançar a segurança no ambiente virtual devemos garantir um ambiente que seja possível também o provimento de autenticação, autorização e o não repúdio. Estas condições devem existir em conjunto com as características supracitadas: a (i) confidencialidade ou a garantia de que a informação não será acessada por indivíduos não autorizados, a (ii) a integridade ou a garantia que a informação armazenada é correta e não sofreu modificações não autorizadas e a (iii) disponibilidade ou a garantia de acesso no momento necessário para os usuários autorizados. Estas características formam um grupo de seis conceitos que, relacionados entre si. possibilitam a implementação de sistemas de informação seguros (20).

Autenticação é a garantia da verificação da origem de uma mensagem ou do indivíduo que se declara o originador. De acordo com o *National Information Assurance Glossary* (NIAG) publicado pelo *Committee on National Security Systems* dos EUA, a autenticação é a medida em que se estabelece a validade de uma transmissão, mensagem ou o seu originador ou ainda a verificação de que um determinado indivíduo tenha autorização para receber determinadas categorias de informação (20).

A autorização, por sua vez, é definida pelo NIAG como o direito de obter acesso privilegiado para um usuário, programa ou processo. Desta forma, enquanto a autenticação define qual usuário está solicitando acesso a um determinado sistema ou recurso, a autorização determina quais são os direitos associados a este usuário, quais sejam, os sistemas, módulos ou recursos que este tem direito a acessar.

O condição de não repúdio é definida pela NIAG como a garantia para o originador da mensagem através de uma prova de entrega desta e ao recebedor a prova da identidade do originador da mensagem. Desta forma é garantido que ambos não possam negar o processamento de mensagem.

Assim, questões relativas à segurança da informação têm se mostrado com importância vital à sobrevivência de empresas, desenvolvimento de negócios e proteção contra violação da privacidade de indivíduos e/ou organizações.

2.2 WEB Semântica

Com o crescimento da Web classicamente como conhecemos, a navegação sem ajuda de buscadores pelo conteúdo contido nas páginas HTML tradicionais não era mais suficiente. Mesmo com o uso de buscadores, ainda há limitações, pois a busca por conteúdo em buscadores tradicionais não leva em consideração a semântica do que se quer buscar. Em 2001, Berners-Lee (21) propôs um formato de dados em triplas orientado pela sua semântica e legível por máquinas e que possuísse a semântica necessária para ser interpretado por computadores, iniciando uma nova era na história da WEB.

Esta nova arquitetura, chamada de WEB de dados, estende a WEB clássica, agregando uma nova estrutura semântica que permite a compreensão dos conteúdos por humanos e máquinas. A WEB de dados é parte constitutiva da WEB Semântica, na medida em que associa semântica aos dados publicados em formato de triplas, de maneira análoga à WEB clássica que associa semântica aos documentos publicados.

Também proposta por Berners-Lee, a Classificação 5 Estrelas é uma forma de classificação onde o grau de abertura do dado é categorizado recebendo uma estrela. Na medida que o dado se torna mais aberto, recebe um maior número de estrelas, como pode ser visualizado na Figura 2.

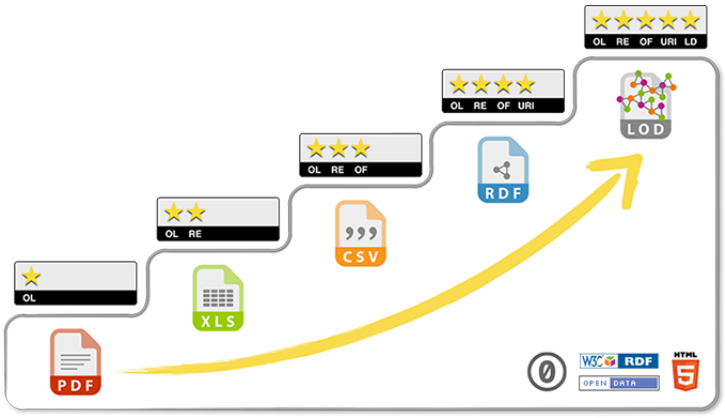


Figura 2 – Classificação dos dados abertos conectados. Fonte: (22)

As 5 estrelas para Dados Abertos significam respectivamente (23):

- Disponível na Internet (em qualquer formato; por exemplo, PDF), desde que com licença aberta, para que seja considerado Dado Aberto;
- Disponível na Internet de maneira estruturada mas ainda em formato proprietário (por exemplo, em um arquivo Excel com extensão XLS);
- Disponível na Internet, de maneira estruturada e em formato não proprietário (CSV em vez de XLS);
- Dados disponíveis na Internet, mas dentro dos padrões estabelecidos pelo W3C (RDF e SPARQL): usar URI para identificar coisas e propriedades; e
- Todos os dados abertos disponíveis e conectados a outros dados, de forma a fornecer um contexto.

A classificação mínima recomendável é de 3 estrelas. Entretanto, quando se trata de web semântica, deve-se focar na classificação a partir de 4 estrelas. Mas, ainda que um determinado conjunto de dados seja publicado com classificação 4 ou 5 estrelas, permanece a necessidade de os repositórios serem ligados semanticamente.

Para que seja possível a criação e/ou preparação de um conjunto de dados para publicação em repositórios é necessário o uso de processos conhecidos como ETL. Esses processos são especificados para a transformação de dados, utilizando-se conjuntamente as estruturas de dados envolvidas (24).

Os processos supracitados podem ser representados como na Figura 3, onde como entradas dos processos pode-se ter uma ou mais fontes de dados. Nas etapas principais podemos ter transformações, limpezas e até mesmo aplicação de técnicas de anonimização. Ao fim do processo, há os destinos onde os dados podem ser armazenados, quer sejam bancos de dados, repositórios institucionais ou conjuntos de dados a serem disponibilizados na LOD.

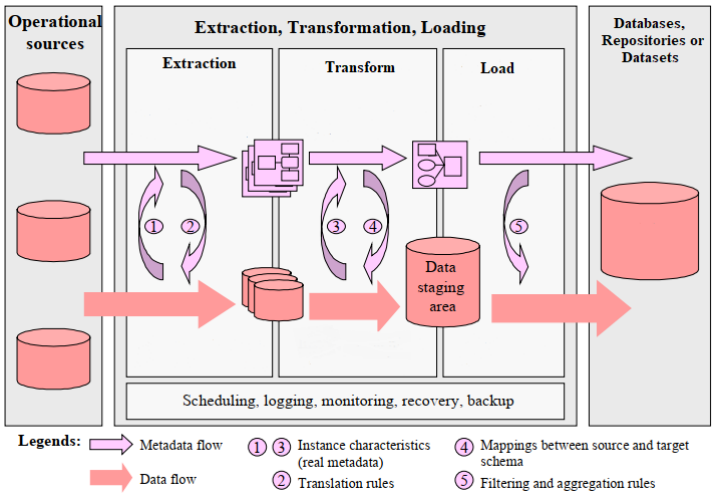


Figura 3 – Arquitetura de um processo ETL típico adaptado (25).

A WEB de dados originou a LOD, rede de dados em que os nós estejam semanticamente ligados, formando um grande grafo global, com informações advindas de várias fontes diferentes ao redor do planeta (26), como pode ser visto na Figura 4.

Iniciada pioneiramente com a DBPedia, nó central da LOD, essa nuvem foi crescendo ao longo do tempo até contabilizar, em maio de 2020, 1.255 conjuntos de dados com 16.174 *links* . É composta por conjuntos de dados expressos por *Resource Description Framework* (RDF) (27), que é um modelo padrão para interligação de recursos na WEB definido pela *World Wide Consortium* (W3C)[1]. Essa base de dados pública é composta por *datasets*, que são conjuntos de dados expressos em RDF que representa os dados estruturados em forma de triplas (sujeito, predicado e objeto), como exemplificado na Figura 5.

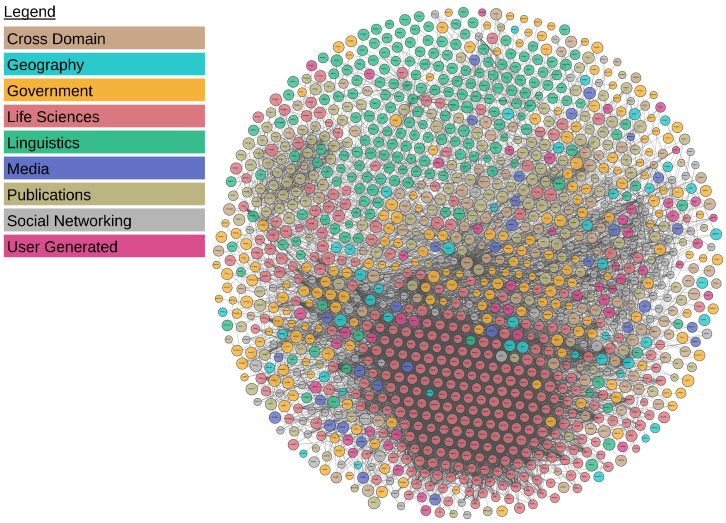


Figura 4 – Nuvem de Dados Abertos Conectados: 1.255 *datasets* com 16.174 links. Fonte: (28)

[1] https://www.w3.org/TR/rdf-sparql-query/

No sentido de aumentar ainda mais a expressividade semântica da Web de Dados, é recomendado o uso de ontologias. Derivado do conceito Aristotélico de ontologia, uma ontologia é a especificação explícita de uma conceitualização compartilhada (29) (30) (31).

Há algumas razões que podem ser citadas para o uso de ontologias (32):

- Compartilhar entendimentos comuns sobre um determinado domínio de informação entre pessoas ou *softwares*;
- Reusar o conhecimento de um determinado domínio, ou seja, usar os mesmos significados de especificações em diferentes projetos e derivá-los para novas variações a partir do modelo inicial;
- Prover explicitamente valores assumidos para um dado domínio, ou seja, assumir parâmetros e relações entre itens pré-definidos;
- Separar o conhecimento de um dado domínio, expressado na sua totalidade, do conhecimento operacional disponível para uso;
- Prover uma análise do domínio de conhecimento compreendendo variantes, riscos, semântica e as relações entre esses valores semânticos.

Podemos expressar os dados semanticamente representados por ontologias através do uso da linguagem OWL[2]. A OWL é uma linguagem de ontologias para a WEB Semântica que estende o padrão RDF, acrescentando maior riqueza e sofisticação de axiomas. Utilizada para a definição do domínio, com suas classes e propriedades, permite interoperabilidade com outros conjuntos de dados (23).

Como fatores de importância desse tipo de publicação, Schmachtenberg et al. (4) ressaltam a importância da adoção de boas práticas na construção e disponibilização de conjuntos de dados na LOD, tais como: (i) prover metadados de proveniência juntamente com os dados a serem publicados; (ii) prover informações de licenciamento, mais recentemente definidas nas categorias *Creative Commons*; (iii) prover metadados no nível do conjunto de dados e (iv) prover métodos alternativos de acesso, como *endpoint* SPARQL [3] ou geração de *dump*.

2.3 Gestão de Dados

O tema de Gestão é bastante amplo. Entretanto, para esta introdução de conceitos, abordaremos nas próximas subseções os tópicos mais correlatos com esta dissertação: os princípios FAIR e aspectos sobre legislações e responsabilidade.

[2] https://www.w3.org/TR/owl2-overview/
[3] https://www.w3.org/wiki/SparqlEndpoints

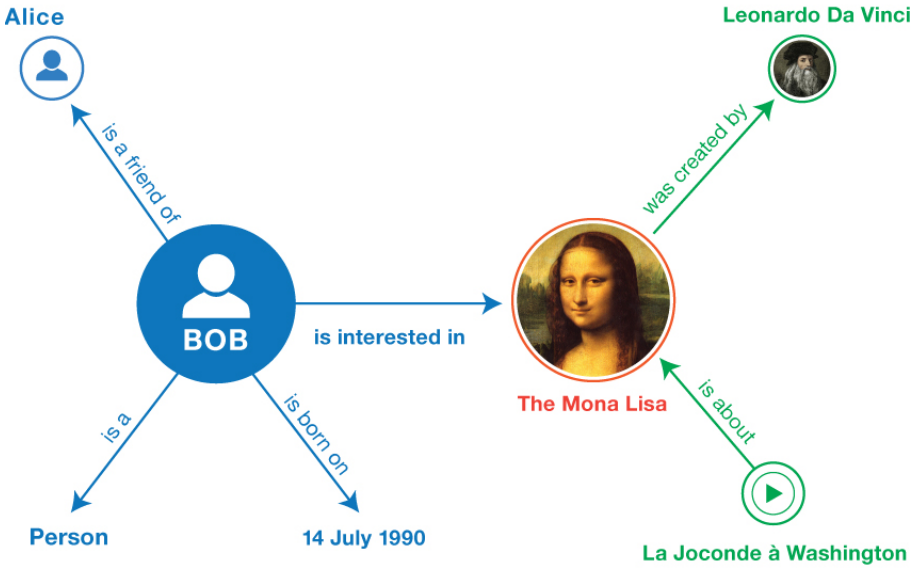


Figura 5 – Exemplo de relação em triplas representado em grafos. Fonte: <https://www.w3.org/TR/rdf11-primer/>.

2.3.1 Princípios FAIR

Os princípios FAIR (*Findable, Accessible, Reusable and Interoperable*) (33), descrevem um conjunto mínimo de requisitos de gestão de dados para lidar com o problema de reusabilidade. São características desejáveis que conferem ao conjunto de dados maior disponibilidade, capacidade de ser acessado e/ou consultado tornando-o, assim, interoperável. Sua aplicabilidade abrange o tratamento de dados realizado por humanos e máquinas. Um ponto importante também é o fato de que esses princípios foram propostos de maneira totalmente independente de tecnologias, ou seja, podem ser aplicados não importa a tecnologia utilizada. Esses princípios são relacionados na Tabela 1.

Importante destacar que a WEB de dados já apresenta determinadas características que possibilitam o atendimento de diversos princípios FAIR, tais como os princípios F1, F3 e F4. A depender dos dados e metadados disponibilizados na WEB de dados é possível atender também aos princípios A2, I1 e I3, por exemplo. Apesar desta condição, a fim de possibilitar que a publicação de determinados recursos na WEB de dados atenda com mais facilidade e de maneira padronizada aos princípios FAIR, o FAIR *Data Point* [4] (FDP) foi proposto como a especificação de uma estrutura de *software* de repositório de dados onde os gestores podem publicar seus metadados para reuso provendo interoperabilidade entre esses repositórios na WEB. Usualmente, os atributos informados de cada *dataset* publicado seguem o padrão de metadados para a descrição semântica *Data Catalog* (DCAT) [5].

[4] https://www.fairdatapoint.org/
[5] https://www.w3.org/TR/vocab-dcat-2/

Tabela 1 – Os princípios FAIR. Adaptado de: (3).

	Princípios
F	F1. Os (meta)dados devem ter identificadores globais, persistentes e identificáveis
	F2. F2. Os dados devem ser descritos com metadados enriquecidos
	F3. Os metadados devem incluir claramente e explicitamente os identificadores dos dados que descrevem
	F4. Os (meta)dados devem ser registrados ou indexados em recursos que ofereçam capacidades de busca
A	A1. Os (meta) dados devem ser recuperáveis pelos seus identificadores usando protocolo de comunicação padronizado
	A1.1. O protocolo deve ser aberto, gratuito e universalmente implementável
	A1.2. O protocolo deve permitir procedimentos de autenticação e autorização, quando necessário
	A2. Metadados devem ser acessíveis, mesmo quando os dados não estão mais disponíveis
I	I1. Os (meta) dados devem ser representados por meio de uma linguagem formal, acessível, compartilhada e amplamente aplicável
	I2. Os (meta) dados devem usar vocabulários que seguem os princípios FAIR
	I3. Os (meta) dados devem incluir referências qualificadas para outros (meta) dados
R	R1. Os (meta) dados são descritos com uma pluralidade de atributos precisos e relevantes.
	R1.1. Os (meta) dados devem ser disponibilizados com licenças de uso claras e acessíveis
	R1.2. Os (meta) dados devem estar associados à sua proveniência
	R1.3. Os (meta) dados devem estar alinhados com padrões relevantes ao seu domínio

O *FAIR Data Point*[6] é um solução que conjuga uma REST API e um Web *client* para a criação, armazenamento e provimento de metadados sobre os *datasets* publicados. Esta solução possibilita armazenar catálogos, *datasets* e distribuições destes *datasets*, gerenciar usuários e direitos de acesso. Há dois níveis de acesso na ferramenta. O primeiro nível de acesso é o que dispensa autenticação e permite o acesso aos catálogos publicados e seus conteúdos. Por outro lado, o acesso autenticado permite a criação de novos itens e o gerenciamento dos conteúdos já publicados. O *FAIR Data Point* foi desenvolvido para ser executado sobre a plataforma Docker e utiliza como componentes: (i) um *proxy* reverso (somente para a configuração de produção), (ii) um *FAIR Data Point Client*, (iii) um *FAIR Data Point*, (iv) o MongoDB para a persistência de dados relativos aos usuários e seus papéis e (v) um *triplestore* para armazenamento de metadados. A estrutura formada por estes componentes e como são interligados pode ser visualizada na Figura 6.

[6] https://fairdatapoint.readthedocs.io/en/latest/

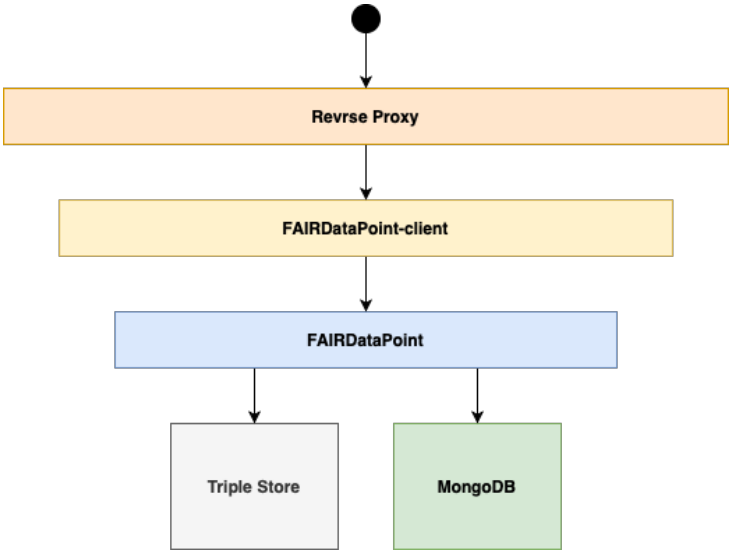


Figura 6 – Componentes do FAIR Data Point. Fonte: <https://fairdatapoint.readthedocs.io/en/latest/about/components.html>.

2.3.2 Legislação e Responsabilidade

Os avanços na automação e digitalização dos sistemas trouxeram muitas vantagens ao dia-a-dia dos indivíduos e organizações mas, por outro lado, também tornaram os seus dados mais expostos a violações de privacidade e de segurança. É necessário ressaltar que segurança e privacidade de dados são dois conceitos relacionados, mas que não devem ser confundidos. A segurança relativa aos dados visa regular o acesso durante todo o ciclo de vida do dado, enquanto a privacidade define como será realizado esse acesso, na maioria das vezes com base em leis e políticas de privacidade (34).

Para compreender as políticas e iniciativas para fornecer segurança de dados e preservar a privacidade, também é necessário conhecer a legislação já consolidada. A GDPR, na Europa, e a LGPD, no Brasil, definem muitos requisitos e qualificam possíveis penalidades para os casos de violações no tratamento dos dados. Entre estas exigências estão garantias como o direito a ser esquecido (exclusão de dados) e a necessidade de recolher o consentimento do usuário para o uso de seus dados (35). A GDPR e a LGPD definiram o conjunto de dados pessoais que pode ser relacionados a um indivíduo ou organização em particular, e que podem levar à identificação destes, direta ou indiretamente. Os dados pessoais são definidos no Art. 4 item 1 da GDPR:

> Dados pessoais, informação relativa a uma pessoa singular identificada ou identificável (titular dos dados); é considerada identificável uma pessoa singular que possa ser identificada, direta ou indiretamente, em especial por referência a um identificador, como por exemplo um nome, um número de identificação, dados de localização, identificadores por via eletrónica ou a um ou mais elementos específicos da identidade física, fisiológica, genética, mental, económica, cultural ou social dessa pessoa singular;

Por sua vez, dados sensíveis estão definidos no Art. 9 item 1:

> ... dados pessoais que revelem a origem racial ou étnica, as opiniões políticas, as convicções religiosas ou filosóficas, ou a filiação sindical, bem como o tratamento de dados genéticos, dados biométricos para identificar uma pessoa de forma inequívoca, dados relativos à saúde ou dados relativos à vida sexual ou orientação sexual de uma pessoa.

Ademais, de acordo com a GDPR Art. 9, itens (h) e (i), e a LGPD Art. 7, item (IV), e Art. 13, os dados utilizados para pesquisa e atividades acadêmicas estão isentos da coleta de consentimento. Em consequência, como resultado da dispensa de consentimento para fins de pesquisa, os projetos também estão isentos de fornecer garantias de exclusão de dados, uma vez que, em princípio, os dados utilizados na pesquisa podem permanecer perpetuamente disponíveis para reutilização.

Muito mais do que a simples adequação legal, hoje o uso, processamento e inferências feitas com dados de pessoas e organizações devem procurar ser realizados à luz de uma gestão de dados responsável (36). Stoyanovich et al. destacam os papéis exercidos hoje pelos gerentes de dados, no que se refere não só à coleta e uso, mas também nas consequências do uso de dados inferidos por mecanismos de AM. Os autores apontam a possibilidade de, por exemplo, uma imputação de dados (um operador de pré-processamento de KDD) realizada de maneira equivocada pode levar um sistema automático de contratação (um tipo de *Automated Decision System* (ADS)) a decorrer em discriminação e tratamento diferenciado entre candidatos homens e mulheres.

Enquanto a conformidade com legislações promove um ambiente legal mais seguro e confiável para o titular de dados, a gestão de dados com responsabilidade vai além: procura conferir a qualquer processo de processamento de AM a garantia do não surgimento de *bias* ou qualquer tipo de distorção durante o pré-processamento. Outras questões, como a existência de bases de dados desbalanceadas entre as categorias apresentadas, podem também gerar algum tipo de distorção em classificações multi-classes. O direito de ser esquecido garantido pela GDPR (direito de exclusão) fez surgir uma outra categoria de processamento, a *Machine Unlearning*. Para os casos em que esse direito seja reinvidicado, se faz necessário realizar todo o rastreamento dos dados do indivíduo no caso de requisição de exclusão e ”re-processamento” de AM, sem os dados do requerente à exclusão, situação esta que pode gerar alto custo de processamento (36).

2.4 Proveniência

Proveniência é definida como toda informação sobre entidades, atividades e pessoas envolvidas na produção de dados ou coisas (37). A captura desse tipo de informação agrega ao ambiente de pesquisa e desenvolvimento: (i) garantia de reprodutibilidade de pesquisas; (ii) reúso; (iii) qualidade; (iv) rastreabilidade; (v) confiabilidade; (vi) transparência e (vii) conformidade Legal.

Os processos de publicação devem capturar e armazenar dados de proveniência a fim de se tornar em conformidade com os princípios FAIR. Considerando-se os dados que resultam de processos computacionais ou fluxos de trabalho, dados de proveniência devem incluir todos os tipos de dados e eventos relacionados a qualquer procedimento que gere, modifique ou trate esses dados. Outra utilidade para os dados de proveniência é tornar possível aos sistemas e bases de dados a eles associados a conformidade com as legislações de proteção de dados como a GDPR e LGPD. A captura e tratamento desse tipo de dado tornou-se fundamental para o atendimento de diversos direitos dos titulares dos dados previstos nas citadas legislações.

Herschel et al. (38) classificam a captura de proveniência em quatro categorias: (a) metadados de procedência, ou seja, o tipo mais geral de procedência, independente de modelos ou método de acesso; (b) procedência de sistemas de informação, onde metadados de sistemas de informação são capturados; (c) procedência de fluxo de trabalho, onde metadados de fluxo de trabalho são capturados e (d) procedência de dados, processo de metadados por dados de sistemas. Estes autores também apontaram diferentes formas de proveniência do fluxo de trabalho: (i) prospectiva, que trata da estrutura e do contexto estático de um determinado fluxo de trabalho; (ii) retrospectiva, que, por outro lado, trata das informações sobre a execução de um determinado fluxo de trabalho, ou seja, as informações disponíveis ao executar o fluxo de trabalho; e (iii) evolução da proveniência, que diz respeito às mudanças feitas entre duas versões do fluxo de trabalho.

Outro aspecto importantíssimo é o papel dos dados de proveniência nos processos de AM. Somente quando esses dados estão disponíveis, todo e qualquer processo de AM se torna "explicável", ou seja, é possível explicar como ocorreu o processo ao qual os dados foram submetidos e, principalmente, de que forma foram encontrados os resultados apresentados (39).

2.5 Preservação da Privacidade

O conceito de preservação da privacidade consiste de abordagens e técnicas que buscam ocultar a identidade e/ou dados sensíveis do titular dos dados (40). Há diversas técnicas e abordagens diferentes com a finalidade de minimizar a probabilidade de reidentificação de indivíduos e organizações. Dentre elas estão as técnicas de anonimização, ilustradas na Figura 7.

Garantir a privacidade dos indivíduos impacta diretamente na qualidade dos dados publicados. Partindo do princípio que a privacidade e a utilidade dos dados são princípios inversamente proporcionais, quanto maior a privacidade, menor será a utilidade dos dados para análise. Assim, é necessário buscar estratégias que garantam a privacidade adequada dos dados sem gerar perda de utilidade dos dados para pesquisas, inferências e/ou tarefas

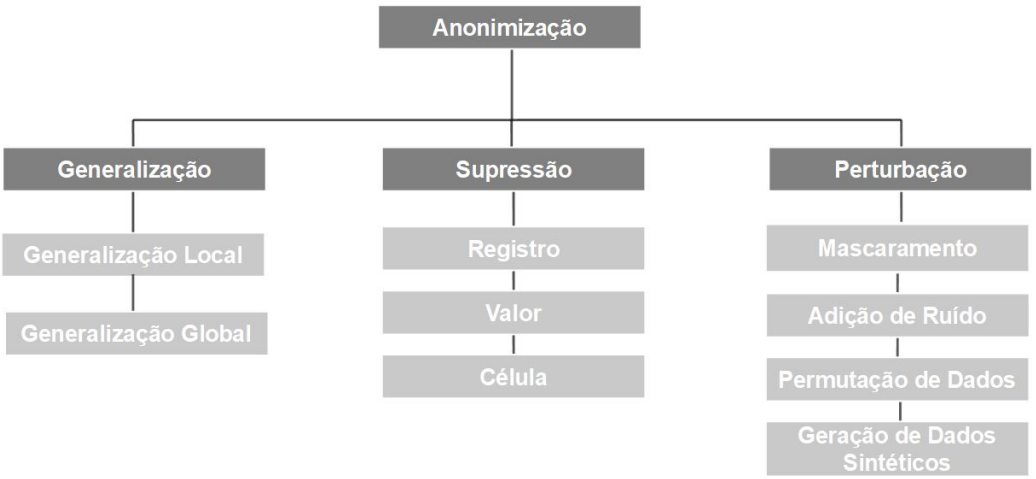


Figura 7 – Categorização de Técnicas de Anonimização. Adaptado de (34).

de AM (34).

Dentre estas estratégias, a mais simples é a supressão de dados, mecanismo no qual atributos definidos são removidos do conjunto de dados, principalmente para atributos identificadores. Essa supressão pode ser de um determinado valor, de apenas uma célula de dados ou até mesmo de um registro inteiro.

Outra abordagem é a generalização, onde valores de atributos semi-identificadores são substituídos por valores mais genéricos, aumentando a incerteza com relação àquele dado. Na generalização global todos os registros de um determinado atributo são mapeados para um mesmo valor generalizado. Já na generalização local um mesmo atributo pode receber diferentes valores para a generalização dentro de uma hierarquia de valores (34).

Já entre as abordagens de perturbação, temos o mascaramento, que é utilizado para a preparação de conjuntos de dados que se deseja publicar, preservando a privacidade dos indivíduos. Muito utilizado na geração de conjuntos de dados para teste ou treinamento. Pode ser realizado por: (i) substituição aleatória de dados por outras informações, (ii) embaralhamento, onde determinado dado também é substituído mas, nesse caso, por algum valor existente no conjunto de dados para esse mesmo atributo, (iii) *Blurring*, onde valores numéricos e datas são alterados por algum percentual aleatório do seu valor real e (iv) a anulação ou truncagem onde os dados a serem anonimizados são substituídos por valores nulos (NULL), como explicitado por Brito e Machado (34).

Umas das abordagens de pertubação dos dados é a adição de ruído, usualmente aplicada em atributos numéricos que recebem seus valores originais perturbados através da soma ou multiplicação por um valor de ruído. De maneira geral, preserva as propriedades estatísticas dos dados embora possa gerar valores sem sentido.

Na estratégia de permutação dos dados são trocados valores do mesmo atributo

Tabela 2 – Exemplo de conjunto de dados antes de aplicadas estratégias de anonimização (Dados Originais). Autoria própria.

ID	Nome	Ocupação	CEP	Telefone
1	João Alves Batista	Médico	22458-323	987551234
2	Maria Claudia Bezerra Souza	Enfermeira	14250-123	976505423
3	Leandro Nogueira da Silva	Analista	21540-660	964081124
4	Luis Claudio Moreira	Programador	15658-234	996801930
5	Carla Sampaio Neves	Arquiteto	16123-456	985542356

Tabela 3 – Exemplo de conjunto de dados depois de aplicadas estratégias de anonimização (Dados Anonimizados). Autoria própria.

ID[1]	Nome[2]	Ocupação[3]	CEP[4]	Telefone[5]
3	João Alves Silveira	Profissional de Saúde	22458	976505423
6	Maria Claudia Lopes Souza	Profissional de Saúde	14250	987551234
9	Leandro Nogueira da Mota	Profissional de TI	21540	996801930
12	Luis Claudio Moreira	Profissional de TI	15658	964081124
15[6]				

Legenda: 1- Adição de Ruído 2- Substituição 3- Generalização
4- Truncagem 5- Embaralhamento 6- Supressão

de registros diferentes e também pode vir a gerar valores sem sentido quando analisados individualmente.

A ideia central da estratégia de geração de dados sintéticos é a construção de um modelo estatístico baseado nos dados e, a partir deste modelo, gerar dados artificiais a serem publicados no lugar dos dados reais, conforme apresentado por Fung et al. (40).

Existem diversos modelos de geração de dados sintéticos, porém o mais conhecido é o k-anonimato proposto por (41), que age na tentativa de proteção a ataques de ligação ao registro. Nesse modelo, k define o nível de privacidade e, simultaneamente, atua sobre a perda da informação. Como não existem abordagens analíticas para determinar o valor de k, a escolha do valor ideal depende de muitos critérios, por exemplo, do requisito de privacidade do proprietário do dado e do requisito de utilidade por parte dos pesquisadores e analistas, de acordo com (34).

Todas as estratégias citadas até aqui visam a preservação da privacidade para dados armazenados de maneira tabulada e que necessitam do cruzamento com outros conjuntos de dados externos para possibilitar a reidentificação de indivíduos. Um exemplo de aplicação dessas técnicas pode ser observado nas Tabelas 2 (Dados Originais) e 3 (Dados Anonimizados).

Além das técnicas de preservação, há ainda uma estratégia que é possibilidade de publicar dados retornados de consultas que tenham a privacidade preservada. Para

tal, são adicionados ruídos ao dados retornados, a fim de evitar ataques probabilísticos. Com a estratégia de Privacidade Diferencial (42) garante-se que o acréscimo ou retirada de um indivíduo do conjunto de dados, retornado por determinada consulta, não afeta o resultado de análises estatísticas feitas com base nesse conjunto de dados. Assim, um suposto atacante não conseguiria obter informações de um determinado indivíduo, as quais já não teria obtido sem o acesso aos dados retornados da consulta.

2.6 Aprendizado de Máquina

IA é uma vertente de ciência da computação que desenvolve técnicas, teorias e aplicações. AM é um sub-ramo de IA e seus algoritmos constroem modelos baseados em dados fornecidos para treinamento que permitem a esses modelos fazerem predições ou apresentarem soluções baseados em dados sem serem especificamente instruídos para tal (7).

A aplicação de técnicas de AM e IA vem se expandindo em diversas áreas como educação, finanças, medicina e, especialmente, no domínio de segurança da informação (7). AM está diretamente relacionado com mineração e análise de dados (ciência de dados), focando em fazer com que os computadores aprendam com os dados, encontrem padrões de dados de interesse, reconheçam ou predigam um determinado comportamento (43).

Desta forma, o uso de técnicas de AM em segurança da informação tem facilitado a descoberta precoce e predição de diferentes tipos de ataques. Essas técnicas podem lidar melhor com novas formas de ataques além de aprender com ações anteriores. Sistemas de detecção e prevenção de ataques baseados em IA podem gerar melhores resultados do que os sistemas baseados em regras, assinaturas ou heurísticas (44).

As técnicas de AM podem ser agrupadas em aprendizado não supervisionado (*unsupervised machine learning*), aprendizado por reforço (*reinforcement learning*) e aprendizado supervisionado (*supervised machine learning*). No aprendizado não supervisionado o agente aprende padrões a partir de dados fornecidos, ainda que não sejam informadas orientações de agrupamento. A técnica mais comum de aprendizado não supervisionado é a clusterização, ou seja, detecção de agrupamentos úteis nos dados informados. No aprendizado por reforço o agente aprende a partir de uma sequência de recompensas ou punições, conforme o alcance de acertos ou erros. No aprendizado supervisionado o agente aprende através da relação dos pares de dados de treinamento como entrada e saída do processamento, o que conduz ao aprendizado da função que determina o dado correto de saída de acordo com a entrada informada (45). Aprendizado supervisionado é aplicado quando há objetivos específicos a serem alcançados a partir da análise de um determinado conjunto de dados. Dentro do espectro de AM, as técnicas mais conhecidas são as de classificação e regressão (46). O escopo desta dissertação se limita às técnicas de

aprendizado supervisionado de classificação. Algumas destas técnicas serão detalhadas a seguir.

O algoritmo *Generalized Linear Model* (GLM) é uma uma técnica avançada de modelagem estatística e extensão dos modelos lineares tradicionais. É um termo guarda-chuva que engloba muitos outros modelos, o que permite que a variável de resposta tenha uma distribuição de erro diferente de uma distribuição normal. Os modelos podem ser *Linear Regression, Logistic Regression, e Poisson Regression* (47).

Deep Learning ou aprendizagem profunda, é um ramo de AM baseado em um conjunto de algoritmos que tentam modelar abstrações de alto nível de dados usando um grafo profundo com várias camadas de processamento, compostas de várias transformações lineares e não lineares. Uma dado (por exemplo, uma imagem), pode ser representado de várias maneiras como, por exemplo, um vetor de valores de intensidade por pixel. Há várias arquiteturas de aprendizagem profunda, tais como redes neurais profundas, redes neurais profundas convolucionais, redes de crenças profundas e redes neurais recorrentes. Estas tem sido utilizadas em áreas como visão computacional, reconhecimento automático de fala, processamento de linguagem natural, reconhecimento de áudio e bioinformática (48).

O *Gradient Boosted Trees* (gradient boosting) é uma técnica de AM para problemas de regressão e classificação, que produz um modelo de previsão na forma de um conjunto de modelos de previsão fracos, geralmente árvores de decisão. Ela constrói o modelo em etapas, como outros métodos de *boosting*, e os generaliza, permitindo a otimização de uma função de perda diferenciável arbitrária (49).

Logistic Regression é frequentemente utilizado para a classificação e análise preditiva. A regressão logística estima a probabilidade de ocorrência de um evento, como votado ou não votado, com base em um determinado conjunto de dados de variáveis independentes. Como o resultado é uma probabilidade, a variável dependente é delimitada entre 0 e 1. Na regressão logística, uma transformação logística é aplicada na probabilidade - isto é, a probabilidade de sucesso dividida pela probabilidade de fracasso (50).

Para que seja possível verificar os resultados obtidos com a criação de modelos de AM para classificação são observadas algumas métricas. Dentre as métricas passíveis de análise, algumas são mais utilizadas em trabalhos existentes. Serão abordadas a seguir algumas destas métricas (48).

A acurácia é, talvez, a mais simples das métricas que pode ser utilizada. Ela é obtida simplesmente dividindo-se o número de predições corretas pelo total de predições e multiplicando-se por 100.

$Acuracia = (Verdadeiro\_Positivo)/(Total\_Predições) * 100$

Entretanto, utilizando-se somente a acurácia pode-se obter análises pobres, podemos

acrescentar à análise a observação das métricas de precisão e *recall*. A precisão é a fração de predições corretas sobre o total de predições.

Precisão = Verdadeiro_Positivo / (Verdadeiro_Positivo + Falso_Positivo)

A métrica de *recall* define a relação entre as predições corretas sobre o total de amostras que deveriam ser preditas como positivas.

Recall = Verdadeiro_Positivo / (Verdadeiro_Positivo + Falso_Negativo)

Ainda que a precisão e o *recall* sejam bons, em alguns casos é necessário realizar a análise da combinação destas duas métricas. Este outra métrica, chamada de F1 *score*, pode ser obtida pela equação:

$$F1\text{-}score = 2 * Precisão * Recall / (Precisão + Recall)$$

3 TRABALHOS RELACIONADOS

Este trabalho propõe-se a apoiar o processo de publicação de dados pré-processados para AM, com a preocupação em atender as leis de proteção à privacidade. Assim sendo, o levantamento de trabalhos relacionados foi organizado de modo a cobrir 3 aspectos. O primeiro diz respeito ao processo de publicação em si, discutido na Seção 3.1. Em seguida, serão apresentados os trabalhos voltados para o tema da preservação da privacidade, na Seção 3.2. Serão comentados os trabalhos relacionados à questões de pré-processamento na Seção 3.3. Por fim, na Seção 3.4, será realizado um comparativo entre esses trabalhos à luz das principais características necessárias a atender ao tema deste trabalho.

3.1 Processos ou *Workflows* para Publicação de Dados

Jacobsen et al. (51) propuseram neste trabalho um *workflow* genérico para tornar a publicação de dados em conformidade com os princípios FAIR, de acordo com a representação na Figura 8. O *workflow* é dividido em 3 fases:

- Pre-FAIRification
  - Identificar o objetivo do processo de *FAIRification*
  - Analisar os dados
  - Analisar os metadados
- *FAIRification*
  - Definir o modelo semântico para os dados
  - Definir o modelo semântico para os metadados
  - Tornar os dados ligáveis
  - Tornar os metadados ligáveis
  - Disponibilizar os dados em formato FAIR
- *Post-FAIRification*
  - Acessar os dados disponibilizados

Por sua vez, no trabalho de Mendonça et al. (52) foi proposta a abordagem ETL4LinkedProv voltada para coletar, ligar e publicar dados de proveniência relativos à tarefas ETL implementadas como um *workflow* assim como uma ontologia que fundamenta a captura de metadados de um determinada execução de um *workflow*.

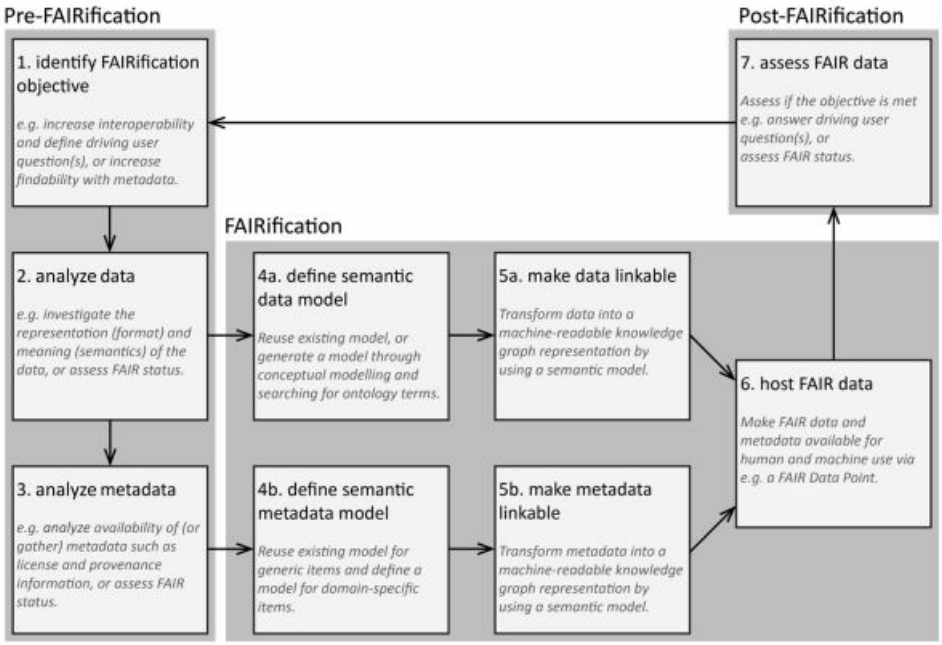


Figura 8 – *Workflow* genérico para o processo de publicação dos dados de acordo com os princípios FAIR (51).

A fim de apoiar a publicação de dados de proveniência coletados, a abordagem usa um conjunto de ontologias existentes. Dentre elas a PROV-O[1] que é usada como a base de representação da semântica dos fluxos de execução. São reusadas também a ontologia OPMW[2] estendendo a PROV-O com o objetivo de diferenciar semanticamente a estrutura do *workflow* (proveniência prospectiva) da execução do *workflow* (proveniência retrospectiva). Acrescenta ainda vocabulários como Dublin Core (DC)[3] e *Friend of a Friend* (FOAF)[4].

Ambos os trabalhos supracitados abordaram a questão de gerência de processos de ETL e a captura de dados de proveniência destes processos. Entretanto, não abordaram questões de preservação da privacidade, ponto importante que esta dissertação contempla. No trabalho de Jacobsen et al. há também a preocupação de possibilitar o atendimento dos princípios FAIR, assim como nesta dissertação.

Outra forma de lidar com a questão de publicação de dados ligados, foi apresentada no trabalho de Salvadori et al. (53), *Data Linking as a Service* (DLaaS) que é uma estrutura para geração e publicação de dados ligados na Web. Essa estrutura tem por objetivo facilitar a execução dos processos necessários à publicação de dados ligados oriundos de conjuntos de dados heterogêneos e distribuídos, através de micro-serviços que proveem enriquecimento semântico, publicação e ligação de dados, alinhamento ontológico e análise

[1] www.w3.org/ns/prov
[2] www.opmw.org/ontology
[3] vocab.deri.ie/cogs
[4] www.foaf-project.org

dos dados.

A infraestrutura proposta em Salvadori et al. busca facilitar a publicação de dados ligados com conjuntos de dados heterogêneos e distribuídos. Essa infraestrutura é composta por *collections* e *containers*. O primeiro é responsável por gerenciar as submissões dos usuários, enquanto o segundo realiza a execução das instâncias assim como dos bancos de dados em triplas. A infraestrutura do DLaaS é ilustrada na Figura 9, compreendendo instâncias da DLaaS *collection*, os quais estão conectadas com múltiplos DLaaS *containers*.

Salvadori et al. (53) propuseram uma abordagem que fosse adaptável ao ambiente da WEB, na medida em que o acesso as dados compartilhados para reúso se faz via serviços, chamados de DLaaS. Entretanto, preocupações como estar em conformidade com os princípios FAIR ou prover a preservação da privacidade, pontos constantes nesta dissertação, ficaram de fora do trabalho citado.

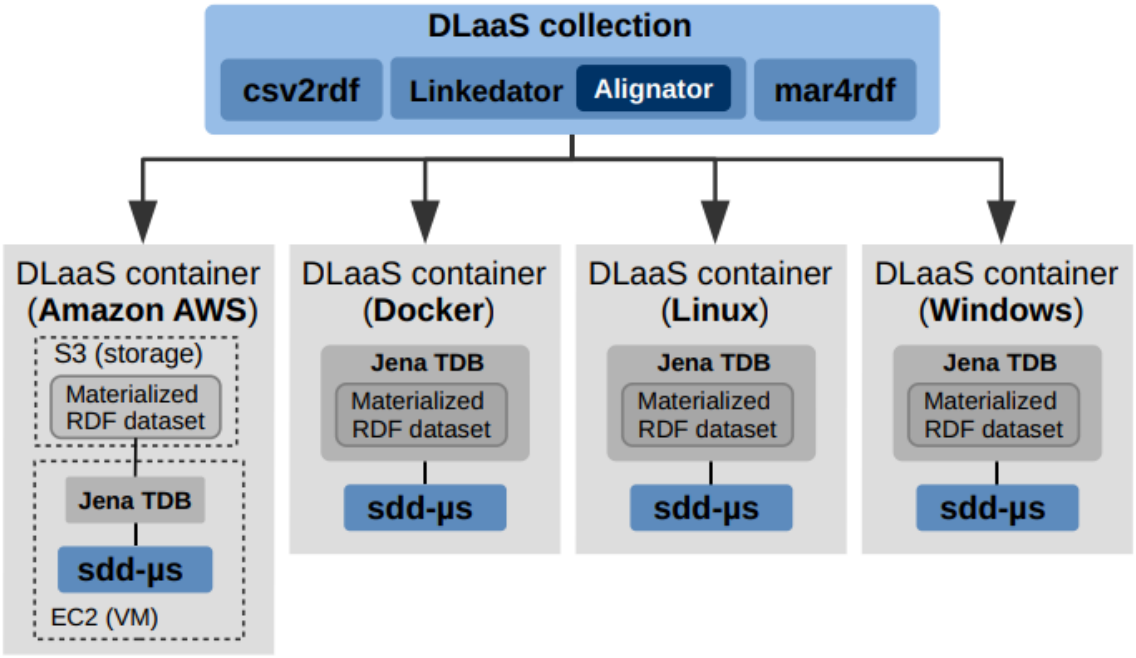


Figura 9 – Componentes da Infraestrutura da solução DLaaS. Fonte: (53)

Rautenberg et al. (54) propuseram um sistema de gerenciamento de *workflow* específico para publicação de dados ligados, o Sistema de Gerenciamento de *Workflow* para processamento de *Linked Data* - LODFLOW. Apresentam também uma ontologia como apresentado na Figura 10, a Linked Data Workflow Project Ontology (LDWPO), baseada nas *Publishing Workflow Ontology* (PWO), *Open Provenance Model Vocabulary* (OPMV) e PROV *Ontology* (PROV-O).

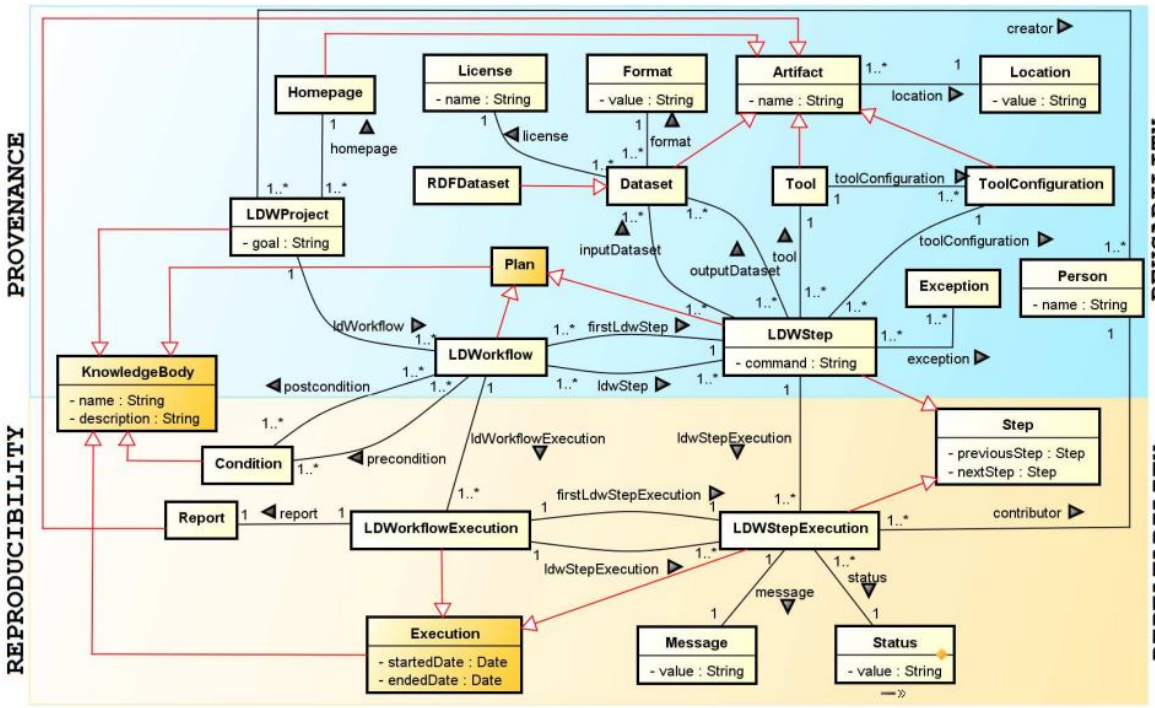


Figura 10 – Ontologia do sistema LODFLOW. Fonte: (54)

Esse trabalho considerou diversos aspectos envolvidos na execução de um ETL para fins de captura de metadados, o que atende às questões de necessidade de captura de proveniência, assim como esta dissertação também atende. Por outro lado, não atendeu questões como preservação da privacidade ou a conformidade com os princípios FAIR.

3.2 Preservação da Privacidade

O PPFSCADA, proposto por Fahad et al. (55), é um *framework* para preservação da privacidade na publicação de dados do sistema SCADA - *Supervisory Control and Data Acquisition*, que é responsável pelo controle e monitoramento de funções de infraestrutura crítica industrial tais como eletricidade, gás, água ou descarte entre outros.

Voltado para aplicação de técnicas de AM, o *framework* PPFSCADA utiliza o conceito de clusterização para aplicação dos algoritmos de adição de ruído de acordo com o tipo do atributo, agrupando os atributos numéricos, categóricos ou qualitativos e hierárquicos. A Figura 11 apresenta o algoritmo usado no processamento de anonimização dos dados tratados pela solução PPFSCADA. No trabalho de Fahad et al. pode-se observar que, no bloco de linhas do Algoritmo 1 na figura supracitada, os dados são particionados de acordo com a característica principal do atributo processado. Nessa forma de categorização dos atributos, o algoritmo não se preocupa com o nível de proteção necessário aos tributos, como destacado na GDPR, dificultando uma escolha adequada de técnica de anonimização.

Algorithm 1: Privacy-preserving framework for SCADA

input :
$Data \longleftarrow \{f_1, f_2, .., f_{n-1}\}$;
$Sim \longleftarrow \{Sim_{NM}, Sim_{HR}, Sim_{CT}\}$;
output:
$[y_{ij}] = [MD_{ij}] \cdot [class_i]$; `// modified data set`
;
$[Nij, Cij, Hij] \longleftarrow PartitionData(Data)$;
if *Data is Partitioned* **then**
  switch *Attribute Type* **do**
    case *Numerical*
      $[NM_{ij}] \longleftarrow NumericalCls(Sim_{NM}, [Nij])$;
    case *Categorical*
      $[CT_{ij}] \longleftarrow CategoricalCls(Sim_{CT}, [Cij])$;
    case *Hierarchical*
      $[HR_{ij}] \longleftarrow HierarchialCls(Sim_{HR}, [Hij])$;
$[MD_{ij}] \longleftarrow$ CombModData($[NM_{ij}]$, $[CT_{ij}]$, $[HR_{ij}]$);

Figura 11 – Algoritmo representativo do processo de anonimização dos dados gerados pelo sistema SCADA. Fonte: (55)

Os autores realizaram experimentos de avaliação com nove conjuntos de dados, além do conjunto de dados gerado pelo próprio sistema SCADA. Por meio desses experimentos foi possível: (i) verificar a acurácia do processo; (ii) verificar a aplicabilidade das técnicas de AM; (iii) quantificar a eficiência computacional e (iv) quantificar a aplicação de técnicas de preservação de privacidade.

Apesar de ser, até o momento, o trabalho encontrado mais completo no sentido de publicação de dados anonimizados voltados para tarefas de pré-processamento de AM, não foram abordadas questões de captura de proveniência e publicação dos dados como dados abertos ligados.

Ødegård (56) descreve e compara vários métodos que possibilitam tornar o tratamento de dados de tráfego de rede em conformidade com a GDPR através de técnicas de anonimização. São tratados dados de logs referentes às camadas de Internet e Transporte, logs no formato NetFlow[5], logs de servidores web e logs de sistema.

No trabalho de Ødegård o processo de anonimização foi orientado às definições de categorias de dados existentes na GDPR[6], dados pessoais e dados sensíveis, de maneira a

[5] https://www.cisco.com/en/US/ technologies/tk648/tk362/technologies_white_paper09186a00800a3db9.html

[6] https://gdprinfo.eu/pt-pt

torná-lo em conformidade com a GDPR. O autor também apresentou diversas possibilidades de uso de técnicas de preservação da privacidade a fim de tornar dados obtidos em diversas fontes em conformidade legal. Apesar disso, não tratou de outras questões como a captura de proveniência, publicação de dados para reúso e a conformidade com os princípios FAIR.

3.3 *Pré-processamento*

Voltada para preparação de dados para tarefas de AM, Moura et al. (15) propôs uma abordagem que contempla uma ferramenta de suporte a cientistas de dados e profissionais de TI na escolha dos operadores a serem aplicados na preparação de conjuntos de dados de treinamento e teste para o uso em algoritmos de classificação. A fim de apoiar a aplicação desses operadores, é proposta também a ontologia *PreProcessing Operators Ontology* (PPO-O), que explicita, além de outros elementos, os operadores de pré-processamento e suas relações.

A ferramenta assistente de pré-processamento de dados proposta por Moura et al. é um software com a finalidade de apoiar a execução de planos de fluxos de trabalho, utilizando metadados do conjunto de dados e das suas colunas. Esses metadados de transformações de dados são capturados e representam a proveniência retrospectiva de cada execução dos operadores. Na Figura 12 são representadas as classes relacionadas com a captura de proveniência de pré-processamento na ontologia PPO-O.

Embora esse trabalho não tenha tido como propósito a preservação da privacidade ou conformidade com legislações de proteção de dados, esta dissertação, com a proposta da abordagem Sec4ML propõe abranger esse aspecto, levando em conta a já desenvolvida ontologia PPO-O, da qual são reusados diversos conceitos, como por exemplo de operadores de pré-processamento.

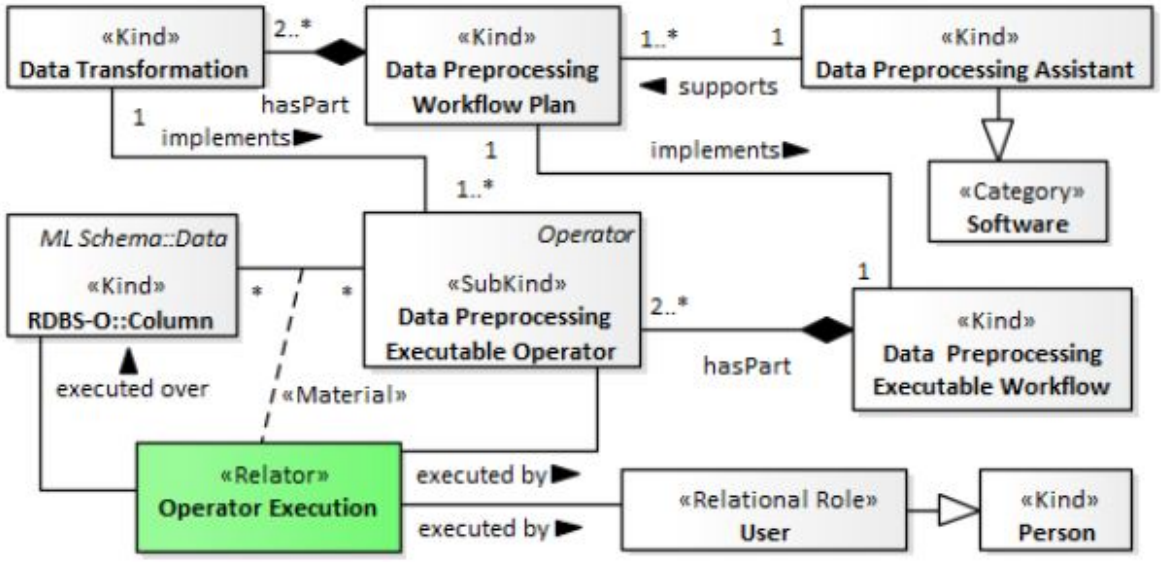


Figura 12 – Representação da Captura de Proveniência do Pré-Processamento da Ontologia PPO-O (parte). Fonte: (15)

3.4 Considerações Finais

A Tabela 4 apresenta uma comparação entre os trabalhos citados, apontando algumas características necessárias para a solução do problema evidenciado. A abrangência de determinado trabalho sobre dados ligados se faz importante na medida em que dados publicados em formato de triplas (RDF) possibilita não somente a ligação de dados oriundos de conjuntos de dados diferentes mas com a mesma semântica, mas também possibilita o processamento por máquina de dados publicados neste formato. Importante ressaltar que outras características como anonimização, adequação aos princípios FAIR e a capacidade de realizar pré-processamento voltado para AM são encontradas em poucos dos trabalhos relacionados. Essa dissertação propõe uma abordagem que pretende contemplar todas as características consideradas.

Tabela 4 – Comparação entre os Trabalhos Relacionados

Trabalhos Relacionados	Características						
	Dados Ligados	Ontologia	Proveniência	Enriquecimento Semântico	Anonimização	FAIR	Pré-Processamento
Salvadori et al. (53)	✓	✓		✓			
Fahad et al. (55)					✓		
de Mendonça et al. (52)	✓	✓	✓	✓			
Rautenberg et al. (54)	✓	✓	✓	✓			
Jacobsen et al. (51)						✓	
Moura et al. (15)		✓	✓				✓
Ødegård (56)					✓		
Sec4ML (esta dissertação)	✓	✓	✓	✓	✓	✓	✓

4 ABORDAGEM SEC4ML

Para apoiar a mitigação do problema da escassez de *datasets* disponíveis para reúso no domínio de segurança da informação, este trabalho propõe a abordagem Sec4ML, que tem como objetivo aplicar em *datasets* tabulares um conjunto de algoritmos de anonimização e pré-processamento para AM, assim como realizar a triplificação dos dados, viabilizando a sua publicação. A abordagem e seus requisitos funcionais está descrita por uma arquitetura que designa seus componentes e interfaces com as quais interage e será detalhada na Seção 4.1. O processo da abordagem Sec4ML define quais são as tarefas executadas sobre os dados e será detalhado na Seção 4.2. Além disso, a Sec4ML utiliza a ontologia Sec4ML-O para apoiar a captura de dados de proveniência durante todo o processo e sua posterior triplificação. Esta ontologia será descrita na Seção 4.3.

4.1 Arquitetura

Para a definição desta arquitetura foram elencados os seguintes requisitos funcionais:

- Ler fontes de dados;
- Realizar limpeza, adaptação e correção de dados;
- Anonimizar e pré-processar dados;
- Capturar metadados retrospectivos;
- Tornar as transformações reprodutíveis, possibilitando a explicação sobre as transformações ocorridas;
- Gerar conjuntos de dados triplificados;
- Tornar possível a reidentificação de entes envolvidos, quer sejam indivíduos ou organizações;
- Publicar dados e metadados; e
- Disponibilizar um catálogo de conjunto de dados.

A fim de atender a esses requisitos, os componentes da arquitetura da abordagem Sec4ML, ilustrada na Figura 13, estão agrupados em duas partes principais: *back-end* e *front-end*. Todos os componentes ilustrados no Sec4ML *Engine* interagem com os componentes do *front-end*, enquanto todas as fontes de dados estão no *back-end*.

O primeiro dos principais componentes do Sec4ML *Engine*, o *Cybersecurity Dataset Processor* é responsável pelo processamento do *dataset*, desde sua leitura, armazenamento na base de dados intermediária, anonimização, pré-processamento até a triplificação. O segundo componente, *Cybersecurity Dataset Publisher*, realiza a publicação dos dados triplificados pelo primeiro componente, utilizando os arquivos gerados com triplas no formato *N-Triple*. Ambos os componentes principais, *Cybersecurity Dataset Processor* e *Cybersecurity Dataset Publisher* interagem com o *Provenance Manager*, fornecendo metadados de proveniência, que alimentam o seu repositório de dados (57). Ao longo de todas estas tarefas a serem executadas, o Sec4ML distingue quatro papéis principais previstos da GDPR: (i) o controlador, que é responsável pelos dados de origem, e que poderá acessar os dados, por exemplo, através de ferramentas de registro de segurança; (ii) o operador, que é responsável por realizar as tarefas de transformação de dados; (iii) o provedor, que é responsável pela publicação e/ou atualização dos dados publicados; e (iv) o consumidor, que pode ser qualquer pesquisador ou desenvolvedor que esteja interessado em reutilizar os conjuntos de dados disponíveis. O Controlador tem acesso aos dados da fonte (conjuntos de dados tabulares) através do módulo *Cybersecurity Dataset Processor*. O Operador transforma os dados também através deste módulo. Por sua vez, o Provedor realiza a publicação dos dados triplificados (bancos de dados em grafo) através do módulo *Cybersecurity Dataset Publisher*.

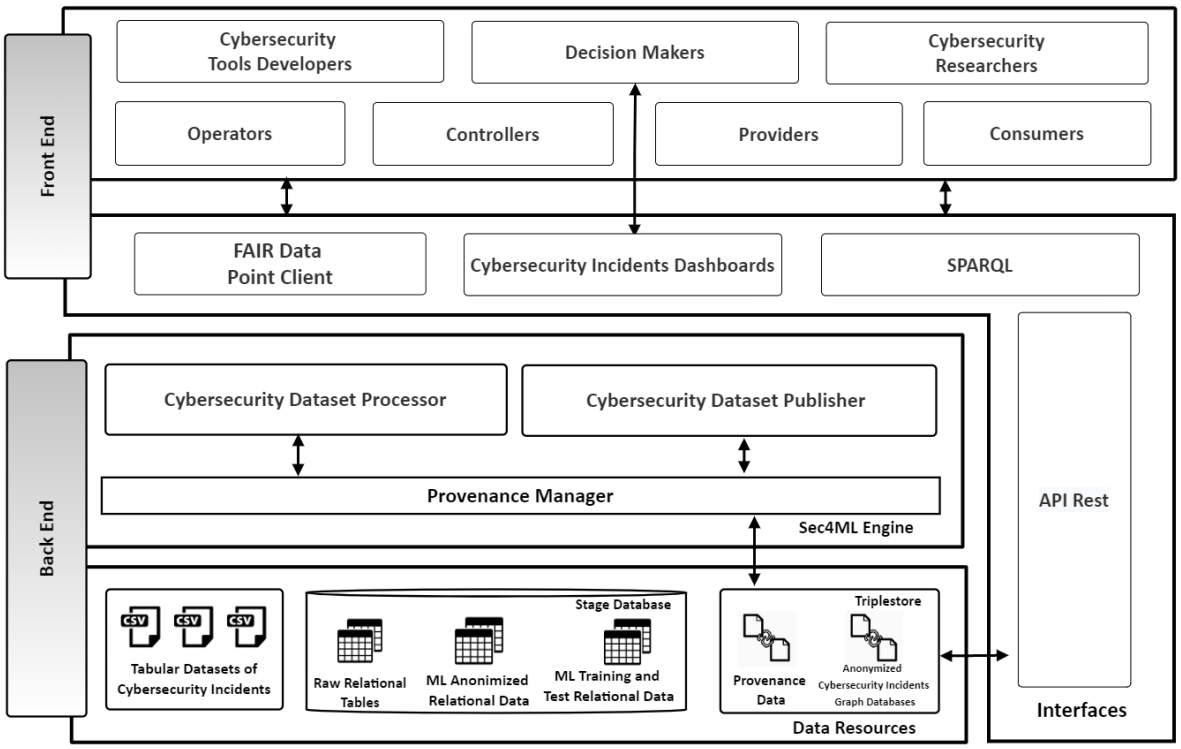


Figura 13 – Arquitetura da abordagem Sec4ML. Autoria própria.

Após triplificados e publicados os dados, o Consumidor é capaz de acessar e reutilizar

não somente os dados triplificados, mas também os dados de proveniência, através de uma variedade de interfaces, tais como FDP *UI endpoints*, *Dashboards* e consultas SPARQL (interface UI). O *Cybersecurity Dataset Processor* é composto por um conjunto de tarefas e será detalhado na próxima subseção. Este módulo apoia o Operador na transformação dos dados brutos em dados anonimizados e pré-processados para tarefas ML. Enquanto estas transformações estão em andamento, o *Provenance Manager* é capaz de registrá-las. Isto fornece dados retrospectivos de proveniência, o que permite: (i) a reprodutibilidade dessas transformações, possibilitando a explicação sobre a transformação sofrida pelos dados transformados e (ii) também a re-identificação como por exemplo a re-identificação de um IP invadido, para fins de investigação.

O componente *Provenance Manager* é o responsável pela captura dos dados sobre a execução em cada etapa, ao longo de toda a execução do fluxo de trabalho (*workflow*). Estes dados capturados tornam possível reproduzir uma parte ou toda a execução do fluxo de trabalho. Ao final da execução deste componente, todos os dados de proveniência e os dados originais do *dataset* estarão disponíveis para triplificação no formato RDF.

4.2 Back-End e seus macroprocessos

O *Back-end* da arquitetura Sec4ML é formado por três componentes principais do Sec4ML *Engine* (vide Figura 13): o *Cybersecurity Dataset Processor*, o *Cybersecurity Dataset Publisher* e o *Provenance Manager*. Os processos representam o fluxo de dados que é trabalhado dentro da arquitetura e são ilustrados na Figura 14. O macroprocesso *Processing and Publication* é responsável pela coleta, transformação e publicação dos dados de incidentes de segurança da informação e corresponde aos componentes *Cybersecurity Dataset Processor* e *Cybersecurity Dataset Publisher*. Por sua vez, o macroprocesso *Provenance Data Capture* é responsável pela captura dos dados de proveniência necessários para apoiar a reutilização do conjunto de dados. Este último macroprocesso representa o fluxo de dados relacionado ao componente *Provenance Manager*.

Durante a execução do macroprocesso *Processing and Publication*, o conjunto de dados é armazenado em formato relacional como apoio para o processamento das demais tarefas, conjuntamente com uma estrutura de descrição de dados processados (dados de proveniência). Alguns trabalhos evidenciam (58, 59) que SGBDs relacionais e orientados a grafos podem se tornar as melhores escolhas a depender da tarefa a ser executada. O formato relacional para armazenamento intermediário foi escolhido por prover recursos para implementações como funções *bulit-in* e gerenciamento de estrutura física entre outras questões. A familiaridade da autora com o ambiente relacional também foi levado em consideração nesta escolha. Entretanto, outros modelos de armazenamento de dados poderiam ser usados. Neste macroprocesso ocorre também a anonimização e o

pré-processamento do conjunto de dados a ser processado. Ao final, estes dados estarão processados pelos subprocessos de anonimização e pré-processamento e triplificados no formato RDF (*n-triple*), prontos para a publicação.

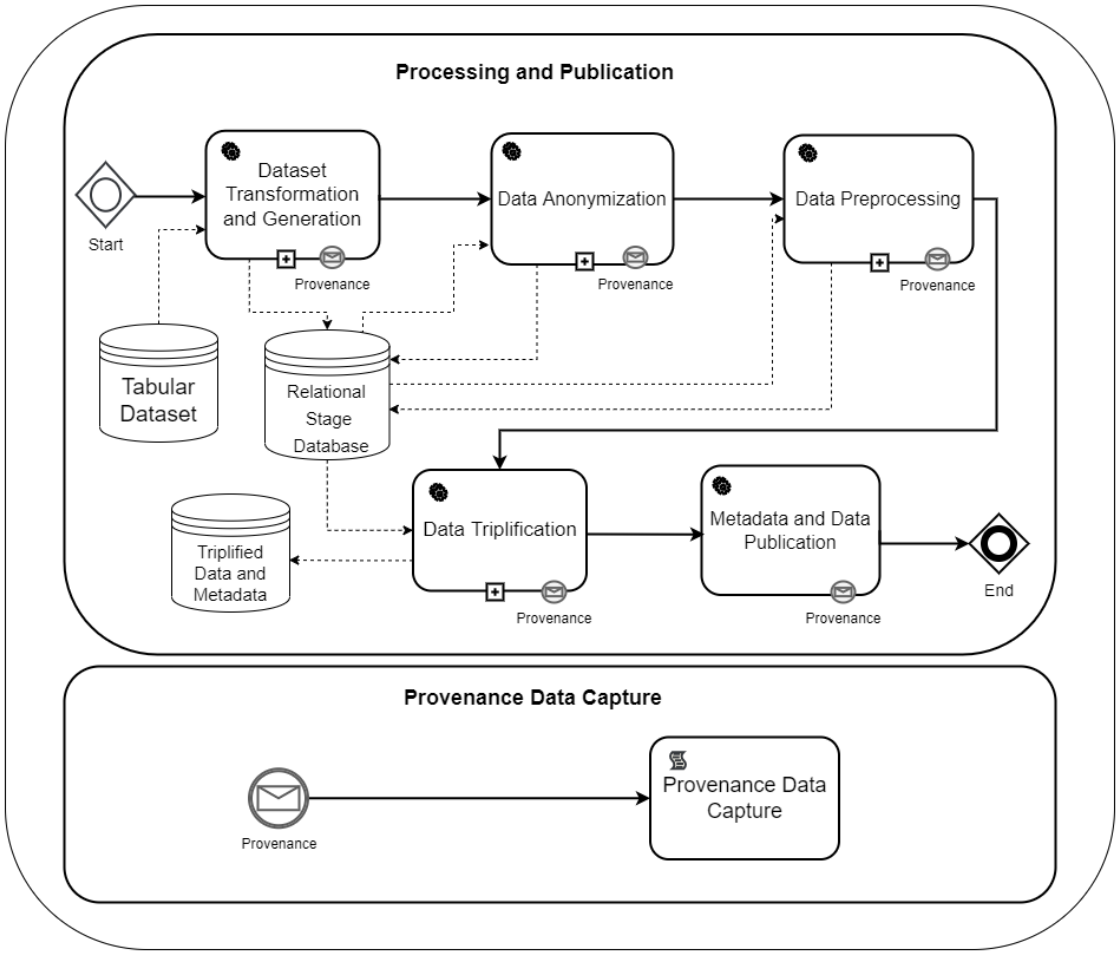


Figura 14 – Processo da abordagem Sec4ML. Autoria própria.

O macroprocesso *Processing and Publication* é constituído pelos cinco processos descritos mais detalhadamente a seguir.

Dataset Transformation and Generation é o processo que inclui subprocessos responsáveis por coletar metadados sobre o *dataset* a ser processado e, dinamicamente, criar uma estrutura de tabela relacional no banco de dados intermediário correspondente aos atributos e tipos do *dataset* original. Este processo também é responsável pela geração de um identificador único (UUID) para cada linha do *dataset*.

Uma vez os dados já carregados na área intermediária, o processo de *Data Anonymization* é realizado por um conjunto de atividades destinado à aplicação de tratamento de anonimização nos atributos do conjunto de dados. Isso é feito permitindo-se que a anonimização possa ser revertida em caso de necessidade de pesquisas, mantendo a utilidade do dados (34), como descrito na seção 2.5. Este processo e o processo *Data Preprocessing*,

por serem o foco do presente trabalho, serão detalhados mais adiante na seção 4.2.1 e na seção 4.2.2, respectivamente.

O processo *Data Triplification* é responsável por transformar os dados relacionais, contidos na base de dados intermediária, em uma estrutura de triplas, tornando possível sua publicação como um recurso RDF. Os dados triplificados publicados, visualizados como um banco de dados gráfico, compõem o LOD.

Ao longo do *Metadata e Data Publication* no final de todos os subprocessos anteriores, os dados triplificados são publicados em um *software* do tipo *triplestore*. Os dados de proveniência capturados são publicados em um FDP.

4.2.1 *Data Anonymization*

O subprocesso *Data Anonymization* é composto por três tarefas essenciais que realizam a anonimização de atributos. Para que esta execução seja possível, se faz necessária a obtenção dos dados necessários à execução do subprocesso. Em especial, é preciso classificar cada atributo do *dataset* de acordo com três categorias principais, conforme previsto na GDPR: identificadores, semi-identificadores e sensíveis. A tarefa *Anonymization Category Definition* realiza a leitura de um arquivo XML contendo como subsídio os nomes dos atributos que forem definidos previamente e a respectiva categoria. Com estes dados capturados, é realizada a leitura da área intermediária, ou seja a tabela criada anteriormente para armazenar os dados oriundos do *dataset* a ser trabalhado. São lidos então os dados relacionados aos atributos informados para cada uma das três categorias de dados. A ilustração deste subprocesso pode ser visualizada na Figura 15.

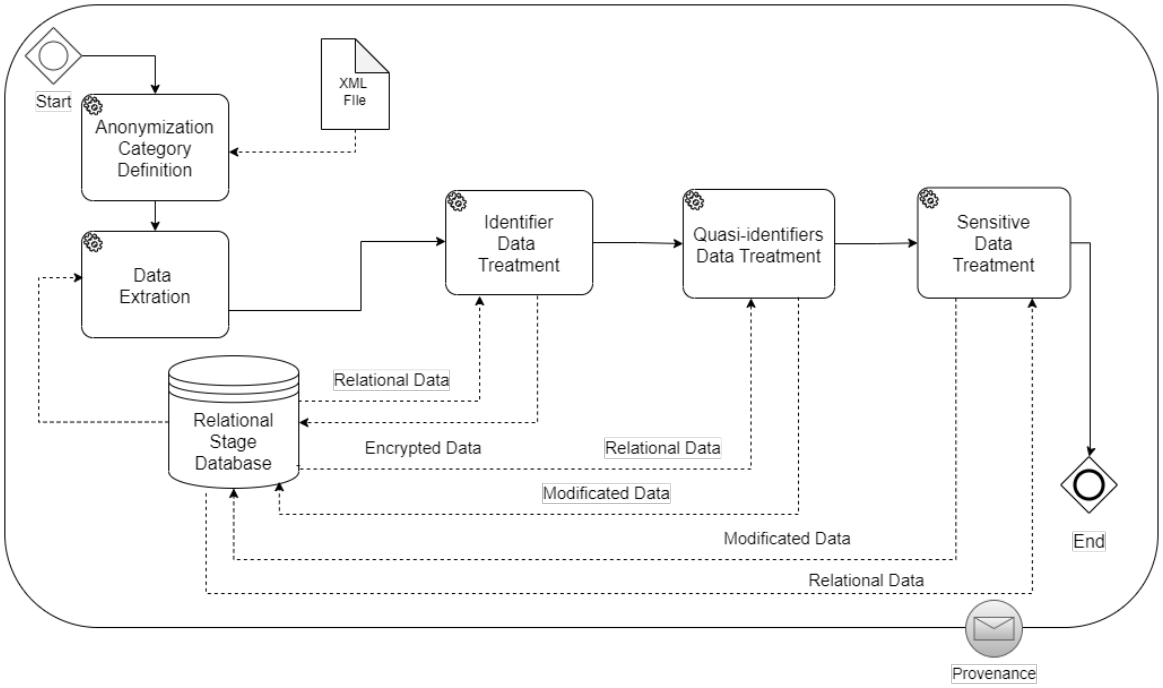


Figura 15 – Subprocesso de Anonimização da abordagem Sec4ML. Autoria própria.

Na tarefa *Identifier Data Treatment* é empregada uma rotina de criptografia simétrica sobre o atributo informado como identificador. Na tarefa seguinte, *Quasi-identifiers Data Treatment* é utilizada, por sua vez, uma técnica de adição de ruído no atributo informado como semi-identificador. Por fim, sobre o atributo informado como sensível é aplicada a estratégia de supressão total deste atributo, eliminando-o do *dataset*.

A escolha de quais técnicas aplicar sobre os atributos informados foi orientada para as mais comumente apontadas como melhores abordagens para cada categoria de atributos relacionada à privacidade ou de acordo com suas características de expressividade (categórica, numérica, etc) (40) (34). Apesar de terem sido escolhidas estas técnicas de anonimização, as três categorias de atributos poderiam ser tratadas por técnicas alternativas, que nesta abordagem não foram consideradas, por limitações de escopo e tempo do trabalho.

4.2.2 *Data Preprocessing*

O subprocesso *Data Preprocessing*, é responsável pela aplicação de determinados operadores de pré-processamento sobre alguns atributos específicos. Uma vez identificados quais atributos necessitam receber tratamento e quais operadores são os mais adequados a serem aplicados, aqueles devem ser informados ao processo de ETL através de arquivo XML. Este arquivo é lido na tarefa *Data Preprocessing Definition*, onde são informados os atributos que receberão tratamento. Na primeira tarefa de tratamento, *Data Normalization*, são processados os atributos informados que necessitam tratamento de operador de

normalização. Na sequência, para os próximos atributos informados, na tarefa *Data Coding* é aplicado um operador de codificação sobre os atributos escolhidos. Por fim, na tarefa *Data Missing* é realizado tratamento de atribuição de valores padrão para dados ausentes, conhecido como operador de *missing values*. É possível visualizar este subprocesso na Figura 16.

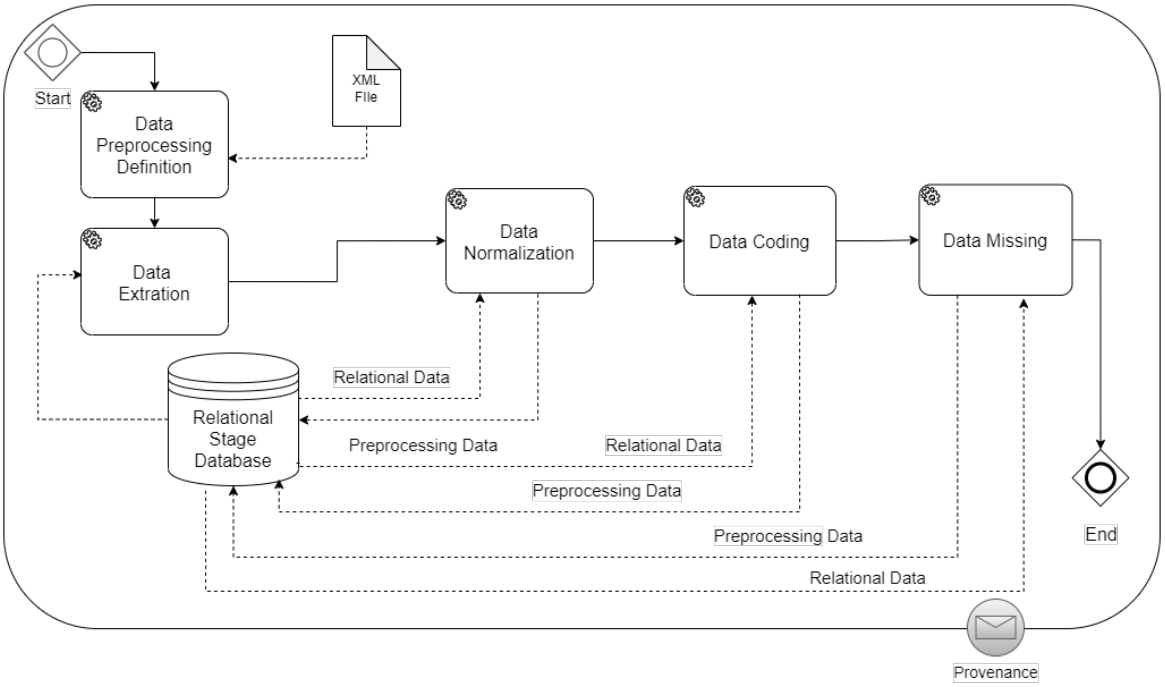


Figura 16 – Subprocesso de Pré Processamento da abordagem Sec4ML. Autoria própria.

4.3 Ontologia Leve Sec4ML-O

A fim de apoiar a captura, processamento e disponibilização dos dados de proveniência, propõe-se uma Ontologia leve (18) chamada Sec4ML-O, que representa os conceitos necessários à captura de dados de proveniência. A conceitualização de ontologia leve se deve ao fato desta ter sido definida sem o formalismo necessário a uma ontologia tipicamente definida. Seus conceitos e relacionamentos foram somente estereotipados pela ontologia PROV-O. Entretanto a Sec4ML-O define os conceitos novos necessários, além dos conceitos reusados, que possibilitam atender às seguintes questões de competência:

1. Características sobre o *dataset* trabalhado;
2. Composição e execução do *workflow* e seus respectivos *steps*;
3. Os operadores e os parâmetros utilizados com os seus respectivos algoritmos;
4. Dados relativos à conformidade com as legislações de proteção de dados.

Tendo como objetivo principal possibilitar o tratamento do dado original através do processo de reidentificação de indivíduos ou organizações utilizando dados de proveniência (38) e ao mesmo tempo atender às exigências da GDPR, essa ontologia leve expressa diversos conceitos reusando classes existentes em ontologias e vocabulários bastante consolidados, tais como Friend of a Friend[1] (FOAF), *Common Warehouse Metamodel*[2] (CWM), *Dublin Core*[3] (DC), *Provenance Ontology* (PROV-O) (37), *GDPR Provenance Ontology*[4] (GDPROV), uma extensão da GDPRov (60) (chamada aqui como *GDPROV-E*), *Data Mining Optimization Ontology* (61) (DMOP), e ontologias propostas em trabalhos recentes como PPO-O (15) e a proposta em (52) (chamada aqui como ETL4LP). Além dos reusos apontados, toda a ontologia é estereotipada com base na ontologia de proveniência PROV[5].

A escolha pela propositura de um modelo que contemplasse os conceitos oriundos da GDPR e não da LGDP se deu pelos seguintes fatores: (i) a GDPR foi promulgada antes da LGPD, (ii) a LGPD foi constituída com base nos conceitos da GDPR, (iii) por ser anterior à LGPD, é possível encontrar mais trabalhos e documentação relativos à GDPR e (iv) por ter sua vigência sobre a União Européia, a GDPR abrange um universo de indivíduos e organizações maior do que a LGPD isoladamente.

A ontologia leve Sec4ML-O pode ser visualizada na Figura 17, e o reuso de vocabulários e ontologias está indicado na Figura 18.

Importante ressaltar que uma das contribuições deste trabalho é a articulação das classes reusadas e das novas classes propostas de maneira a possibilitar a captura dos dados de proveniência de execução, relacionando-os com dados que evidenciam a conformidade com a principal legislação de proteção de dados em vigor no mundo, a GDPR. Esta articulação é obtida ainda que, em algumas situações, o reuso de ontologias não seja para conceitos semanticamente idênticos. Cada subgrupo de classes será descrito detalhadamente nas subseções a seguir.

4.3.1 *Dataset Metadata*

Ilustrado pela Figura 19 está o subgrupo de metadados relativo aos *datasets*. Cada *dataset* é representado pela classe *ppo-o:LabeledDataset*, e é descrito por (*isDescribedBy*) suas colunas (*cwm:Column*). Cada coluna corresponde a um descritor/metadado, que pode ser categorizado de acordo com seu tipo (*cwm:referencedTableType*), representado pela classe *cwm:SQLStructuredType*. Estes tipos são especializados nas duas subclasses *ppo-o:QuantitativeDataType* e *ppo-o:QualitativeDataType*, diferenciando dados quanti-

[1] http://xmlns.com/foaf/spec/
[2] https://www.omg.org/spec/CWM/1.1/AboutCWM/
[3] https://www.dublincore.org/
[4] https://openscience.adaptcentre.ie/projects/CDMM/GDPRov/
[5] https://www.w3.org/TR/2013/REC-prov-o-20130430/

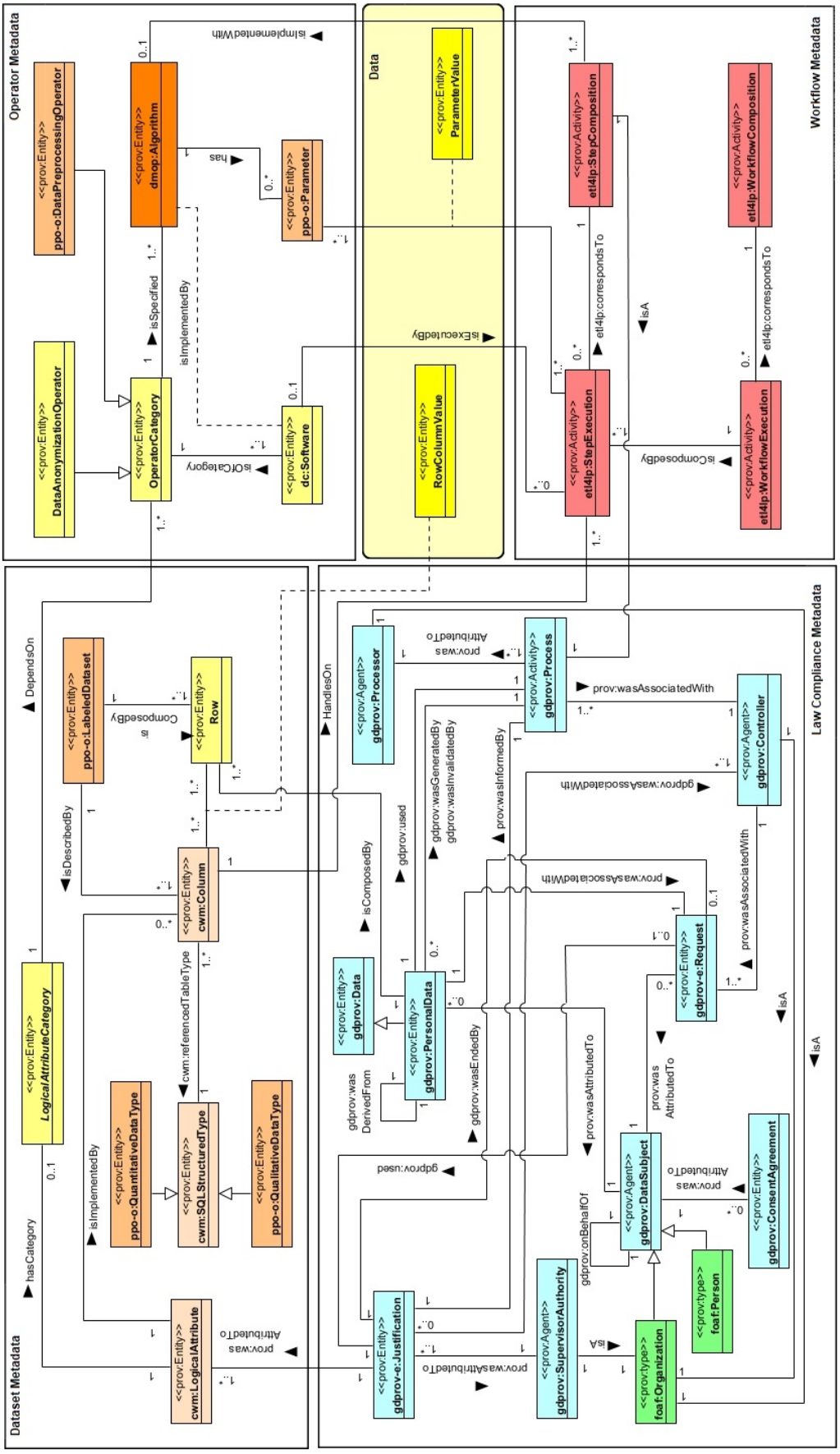


Figura 17 – Ontologia de apoio para a abordagem Sec4ML (Sec4ML-O). Autoria própria.

Figura 18 – Legenda para a Ontologia de apoio Sec4ML-O.

tativos (*e.g. number, float*, etc) e dados qualitativos (*e.g. char, varchar, boolean,* etc, respectivamente. Os dados do *dataset* estando em formato tabular, estão organizados na forma de colunas e linhas. Estas são descritas pelas classes *cwm:Column* e *Row*, respectivamente. Fruto da associação de linhas e colunas, que pode ser da ordem de *n:m*, surge a classe associativa *RowColumnValue*, evidenciando essa congruência. Esta classe é a que representará os próprios dados de cada encontro entre linhas e colunas, comumente chamados de "células", na estrutura tabular.

A classe *cwm:LogicalAttribute*, trazida da especificação CWM, representa os atributos lógicos que podem ser implementados (*isImplementedBy*) na forma de uma ou mais colunas *cwm:Column* de um *dataset*. Conforme a CWN, um atributo lógico pode ser numérico, categórico ou ambos, de acordo com seu uso. Assim, a classe *LogicalAttributeCategory* representa essas categorias.

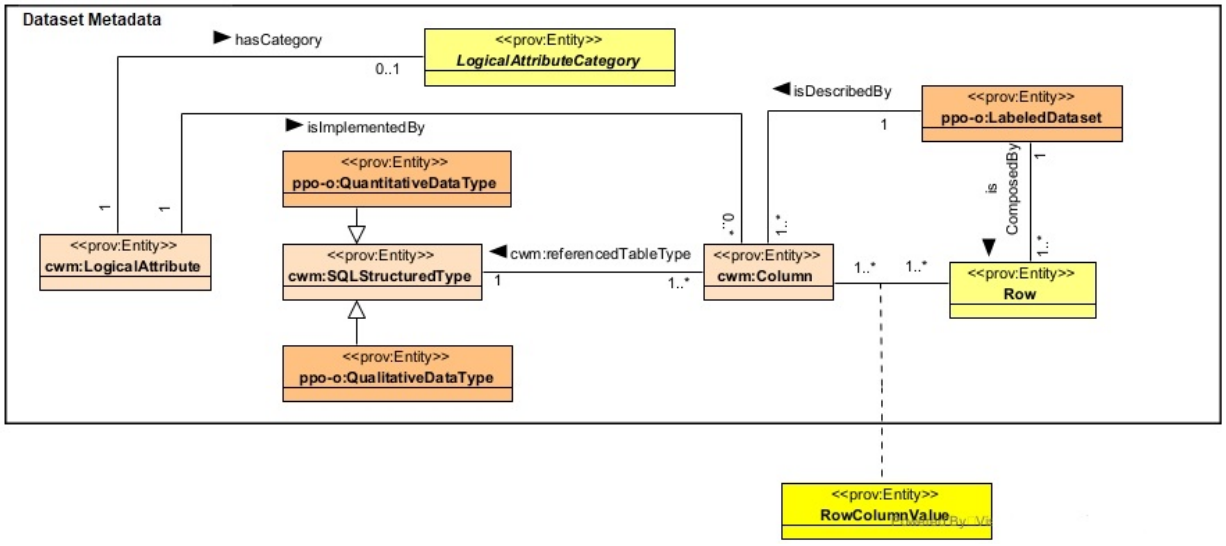


Figura 19 – Ontologia Sec4ML-O para a representação dos metadados sobre o *dataset*. Autoria própria.

4.3.2 *Operator e Workflow Metadata*

Os metadados relacionados aos operadores são capturados de maneira orientada a como são aplicados e algoritmos que estão relacionados, assim como o *software* e parâmetros utilizados. A representação desta parte da ontologia pode ser visualizada na Figura 20. A classe *OperatorCategory* é especializada pelas classes *DataAnonymizationOperator* e *ppo-o:DataPreprocessingOperator*, evidenciando a distinção entre os operadores de anonimização e os operadores de pré-processamento, respectivamente.

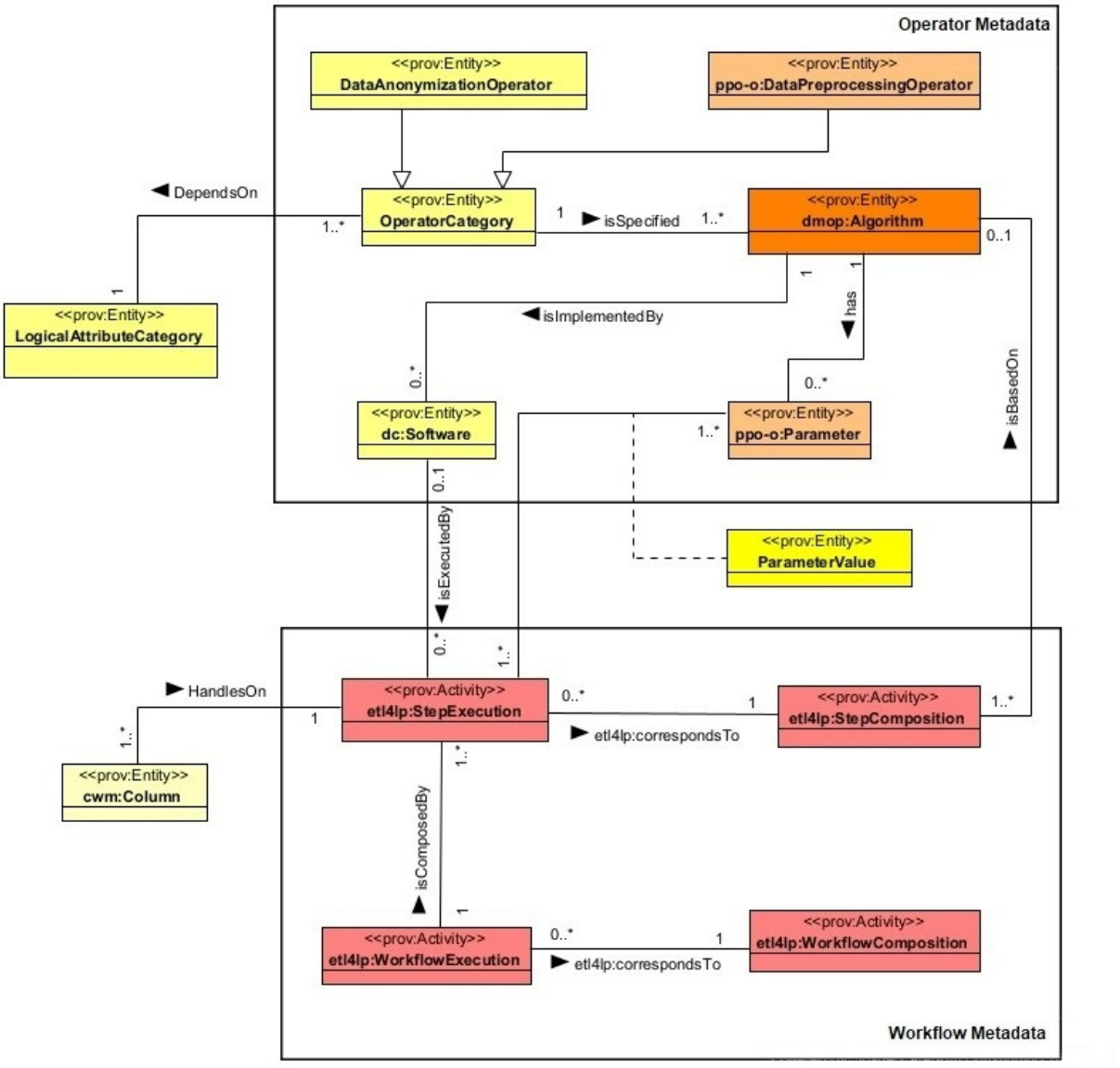


Figura 20 – Ontologia Sec4ML-O para a representação dos metadados sobre os operadores e *workflows* utilizados na execução do ETL. Autoria própria.

Cada instância da classe *OperatorCategory* a ser utilizada é especificada (*isSpecified*) por um ou mais algoritmos evidenciados pela classe *dmop:Algorithm*. Por exemplo, a categoria de operadores de Normalização é especificada pelos algoritmos/fórmulas Zscore,

MinMax, LogNormal, etc. Além disso, cada instância da classe *OperatorCategory* é voltada para (*DependsOn*) atributos da categoria definida pela classe *LogicalAttributeCategory*, do subgrupo de metadados representativos do *dataset*. Por exemplo, a categoria de operadores Normalização trata apenas de atributos numéricos.

Cada software (*dc:software*) pode ter sido usado (*isExecutedBy*) na execução de um *step* (*etl4lp:stepExecution*). Além disso, cada instância de *step* corresponde a um determinado *stepComposition* (*StepComposition*). Essa instância, por sua vez foi é baseada (*isBasedOn*) na chamada de um determinado algoritmo (*dmop:Algorithm*). Por exemplo, o *software* MinMaxScaller da biblioteca Python Scikitlearn pode ser usado (*isExecutedBy*) na execução de *steps*, que por sua vez correspondem a uma determinada instância da classe (*stepComposition*). Essa instância pode ter sido implementada com a chamada ao algoritmo MinMax, que por sua vez é uma especificação do operador de Normalização. Cada algoritmo é implementado por (*isImplementedBy*) um *software* (*dc:Software*).

Se a execução do algoritmo demandar uso de parâmetros, estes são definidos (*has*) pela classe *ppo-o:Parameter*. A partir da relação desta classe com a classe *etl4lp:StepExecution*, surge a classe associativa *ParameterValue* que representa os efetivos valores usados durante determinada execução dos *steps* que necessitam de parâmetros.

Os metadados acerca do processo de ETL são capturados pelo ponto de vista da sua composição e da execução propriamente dita de seus componentes. Para essa primeira visão, as classes *etl4lp:WorkflowComposition* e *etl4lp:StepComposition* armazenam metadados sobre a composição do *workflow* e do *step* respectivamente. Um *workflow* é composto (*isComposedBy*) por *steps*. Os metadados relativos à efetiva execução de cada componente do ETL são capturados pelas classes *etl4lp:WorkflowExecution* e *etl4lp:StepExecution*. As instâncias destas classe são correspondentes (*etl4lp:correspondsTo*) às composições dos *workflows* e dos *steps*, respectivamente. As relações entre estas classes podem ser visualizadas na Figura 20.

4.3.3 *Law Compliance Metadata*

Os metadados relativos à conformidade legal da execução do processo Sec4ML estão orientados sob o ponto de vista dos indivíduos ou organizações. Nessa perspectiva, tanto os primeiros quanto os segundos são representados pela classe *gdprov:DataSubject* daqui em diante referenciada como **ente envolvido**, que se especializa na classe *foaf:Organization* para as organizações e na classe *foaf:Person* para as pessoas. Cada ente envolvido pode ter nenhuma ou muitas declarações de consentimento, representadas pela classe *gdprov:ConsentAgreement*, que lhe foram atribuídas (*prov:WasAttributedTo*). As instâncias da classe *gdprov:ConsentAgreement* representam o que o ente envolvido declara como consentimento relacionado à ações que seus dados podem sofrer. Eventualmente, um ente envolvido pode responder (*gdprov:onBehalfOf*) por outro ente envolvido. A cada ente

envolvido podem ser atribuídas (*prov:wasAttributedTo*) muitas ou nenhuma requisição, e cada uma é representada pela classe *gdprov-e:Request*. Uma instância dessa classe corresponde a alguma ação que pode ser feita por um ente envolvido, como por exemplo, requerer a retirada (*Withdraw*), o acesso (*Access*), a correção (*Correction*), o apagamento (*Erasure* ou a restrição (*Restriction*) aos seus dados. Ainda, ao ente envolvido podem ser atribuídas instâncias da classe *gdprov:PersonalData*. Essa classe representa a especialização da classe *gdprov:Data* para dados pessoais. Por fim, uma instância desses dados pessoais *gdprov:PersonalData* é composta por (*isComposedBy*) uma ou mais instâncias da classe *Row*, classe do subgrupo de metadados de *datasets*.

Cada requisição deve estar associada (*prov:wasAssociatedWith*) a exatamente um controlador representado pela classe *gdprov:Controller*. O controlador (*gdprov:Controller*) é uma organização responsável pela tomada de decisões sobre o gerenciamento de dados das organizações ou pessoas controladas. Uma requisição está associada (*prov:wasAssociatedWith*) a uma instância da classe *gdprov:PersonalData*. Uma requisição pode usar (*gdprov:used*) ou finalizar (*gdprov:wasEndedBy*) uma justificativa representada pela classe *gdprov-e:Justification*. Essa classe representa qualquer justificativa que venha a ser feita com a finalidade de autorizar um determinado processamento sobre dados de indivíduos ou organizações. As justificativas podem ser de: obrigação legal (*Legal Obligation*), interesse vital (*Vital Interest*), interesse público (*Public Interest*), autoridade oficial (*Official Authority*), legítimo interesse (*Legitimate Interest*) e contrato (*Contract*). Uma ou mais justificativas podem ser atribuídas (*prov:wasAttributedTo*) a uma autoridade supervisora representada pela classe *gdprov:SupervisorAuthority*. As justificativas podem ser atribuídas também (*prov:wasAttributedTo*) a um atributo lógico representado pela classe *cwm:LogicalAttribute*. Além disso, uma justificativa deve ser informada (*prov:wasInformedBy*) por um processo representado pela classe *gdprov:Process*.

Um controlador é uma (*isA*) organização que especializa a classe *foaf:Organization*. Um ou mais controladores podem estar associados (*gdprov:wasAssociatedWith*) com justificativas. Um ou mais processos podem ser associados (*prov:wasAssociated With*) a um controlador e atribuídos (*prov:wasAttributedTo*) a um processador. Este, por sua vez, também é uma (*isA*) organização. Uma instância de *gdprov:PersonalData* pode ser gerada (*gdprov:wasGeneratedBy*), usada (*gdprov:used*) ou invalidada (*wasInvalidateBy*) por um processo. Uma instância de *gdprov:PersonalData* pode também ter sido derivada (*gdprov:wasDerivatedFrom*) de outra instância de *gdprov:PersonalData*.

4.4 *Considerações Finais*

Ao longo deste Capítulo foi possível detalhar toda a abordagem Sec4ML, especificando e detalhando sua arquitetura, processos além da ontologia de apoio à aplicação

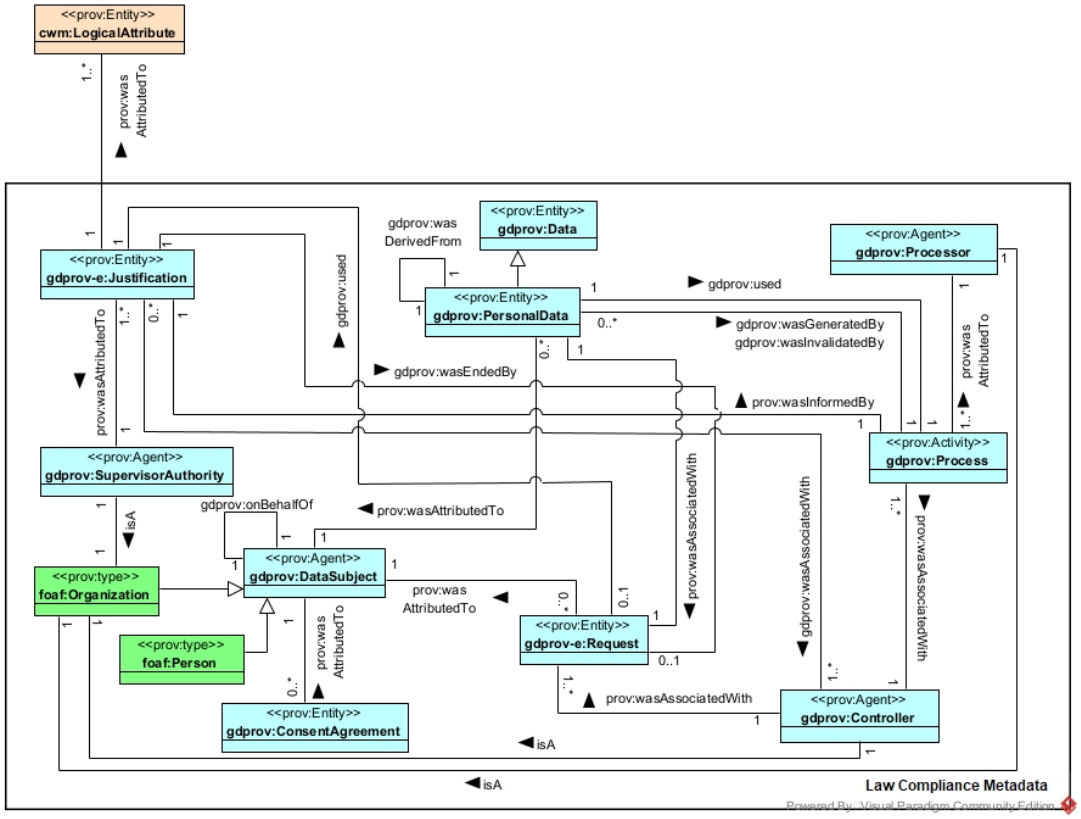


Figura 21 – Ontologia Sec4ML-O para a representação dos metadados sobre a conformidade legal. Autoria própria.

da abordagem. Desta forma possibilitou-se compreender de que forma a arquitetura e o processo estão conectados e qual o papel da ontologia Sec4ML-O na classificação semântica dos dados e definição das estruturas de armazenamento.

5 IMPLEMENTAÇÃO DA ABORDAGEM SEC4ML

Este capítulo aborda todos os aspectos da implementação da abordagem Sec4ML. A implementação da abordagem utiliza recursos de *hardware* e *software*. A escolha destes recursos levou em consideração: (i) o quanto a ferramenta é difundida no universo acadêmico e na indústria; (ii) a usabilidade; (iii) o fato de ser livre de custos de licenciamento e (iv) a familiaridade adquirida em usos anteriores. Na Seção 5.1 são detalhados os recursos de *hardware* e *software* utilizados na abordagem como um todo. Na próxima Seção 5.2 são detalhados os *jobs* e transformações desenvolvidos para a criação do processo de ETL. E, por fim, são abordadas algumas considerações finais na Seção 5.3.

5.1 Infraestrutura de Implementação

Para a implementação do Sec4ML *Engine* foi escolhida como ferramenta de processamento de ETL a *Pentaho Data Integration*[1] (PDI). Esta ferramenta permite a criação de *workflows* ETL e provê uma interface intuitiva, oferecendo um amplo conjunto de recursos. No *workbench* PDI os processos ETL podem ser modelados como grafos orientados, onde as fontes de dados, transformações e processadores de dados são representados como nós chamados de *steps* e *jobs* (estrutura que agrupa *steps*). Ao combinar várias etapas, *workflows* ETL podem integrar fontes de dados heterogêneas e esses dados podem ser submetidos a uma grande variedade de tipos de processamento. A arquitetura instanciada com as ferramentas utilizadas está ilustrada na Figura 22.

Na implementação do componente *Cybersecurity Dataset Processor*, com relação aos processos de anonimização e pré-processamento, foi utilizado, conjuntamente com o PDI, o *plugin CPython Script Executor* que permite o desenvolvimento de *steps* em código Python, possibilitando a agregação de funcionalidades para tratamento e processamento de dados voltados para ML. Este *plugin* utiliza como bibliotecas mínimas o *Pandas, NumPy, Py4J* e *MatPlotLib*. Em uma das implementações, para as tarefas de anonimização e pré-processamento foram utilizadas as bibliotecas *yacryptopan*. Na outra implementação, com criptografia simétrica, foram utilizadas as bibliotecas *cryptography* (com o método *fernet*) e *base64*. As biliotecas *NumPy* e *scikit-learn* também foram usadas em ambas as implementações. Para que seja possível utilizar o *plugin CPython Script Executor* foi necessário a instalação do ambiente Python Anaconda e o JupyterLab como interface de desenvolvimento.

Para a realização da leitura dinâmica de metadados dos *datasets* processados foi utilizado o *plugin File Metadata* que permite realizar a leitura de um determinado arquivo

[1] https://help.hitachivantara.com/Documentation/Pentaho/7.1/0D0/Pentaho_Data_Integration

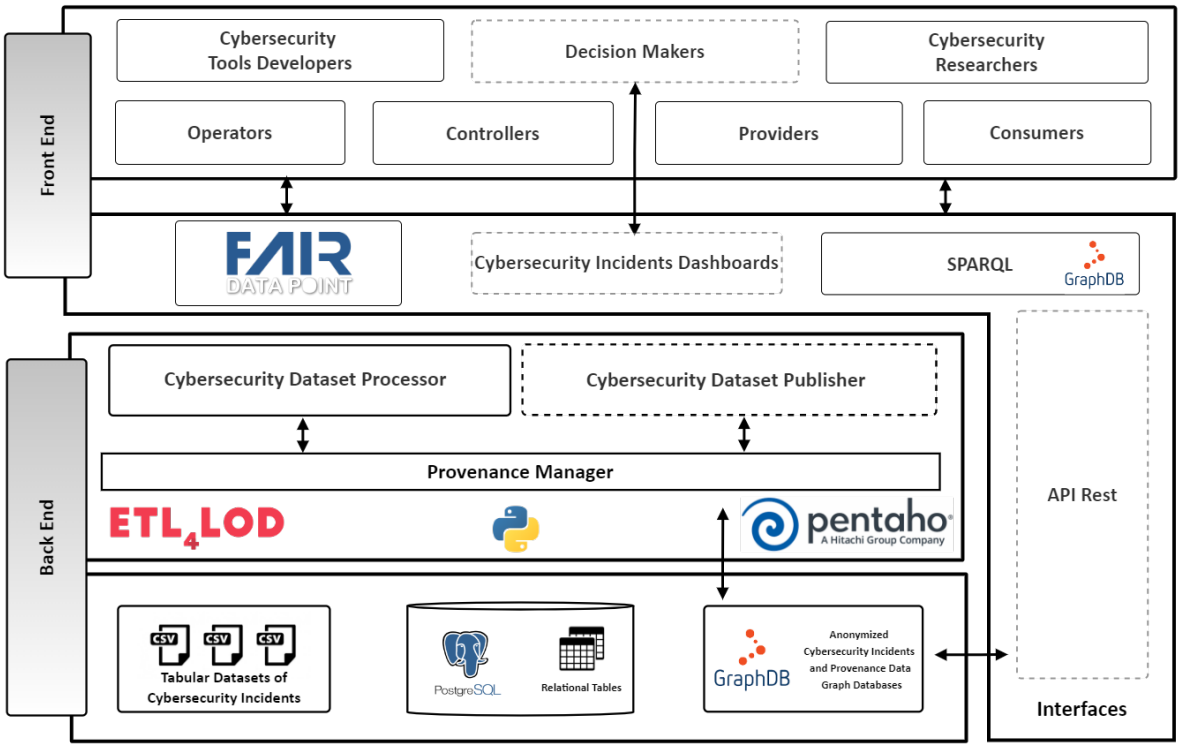


Figura 22 – Ferramentas utilizadas na implementação da arquitetura da abordagem Sec4ML.

tabular e fornecer metadados dinamicamente sobre seus atributos, facilitando a leitura destes e o carregamento dos dados a serem processados.

A fim de implementar a triplificação dos dados e metadados processados foi utilizado também o *plugin* ETL4LOD (visualizado na ferramenta como *LinkedDataBR*) que fornece uma gama de *steps* que possibilitam o processo de triplificação de dados para tipos de dados heterogêneos e fontes distintas. Foram utilizados alguns destes *steps* para a triplificação dos dados.

Como base de dados intermediária, foi adotado o SGBD PostgreSQL. Nesta base de dados são armazenados em área intermediária os dados tratados oriundos de arquivos tabulares (.csv), assim como os metadados gerados e capturados pelo processo *Provenance Data Capture*. O esquema relacional correspondente, desenvolvido. com base na ontologia Sec4ML-O está disponível no Apêndice C. Como interface de trabalho com esta base de dados foi utilizado o *PGAdmin*. Como repositório de triplas, o *triplestore* escolhido foi o GraphDB. Este triplestore foi utilizado para a implementação do banco de dados em grafos, possibilitando também a realização de consultas SPARQL, uma das interfaces propostas na arquitetura Sec4ML. Esta implementação foi realizada no mesmo *hardware* onde foi implementado também o PDI.

Com o objetivo de demonstrar que, após a submissão dos *datasets* aos proces-

sos de anonimização e pré-processamento e sua finalização, os dados não perdem suas características de utilidade para tarefas de classificação em AM, foi escolhida a ferramenta *RapidMiner*. Nesta ferramenta é possível realizar diversas tarefas de AM, tais como predição, clusterização ou detecção de *outliers*.

A implementação da abordagem Sec4ML correspondente ao processamento dos dados foi desenvolvida em um *hardware* (*notebook*) com processador Intel i5, 20Gb de memória RAM e SO Windows 10. Para implementação da ferramenta *RapidMiner* foi utilizado um *hardware* (*notebook*) com processador Intel i5, 12 GB de memória RAM e SO Windows 10. Para esta implementação foram utilizados o PDI 9.0, *Anaconda* 3.8.8, *JupiterLab* 3.0.14, *PostgreSQL* 13, *GraphDB* 9.1.11 e *RapidMiner* 9.10. Nesta instalação a versão do *FAIR Data Point* utilizada foi a versão de instalação local com o GraphDB como *triplestore*. A implementação foi realizada em uma máquina virtual hospedada na Intranet do IME, utilizando processador Intel Xeon E7 (2 CPU), 3Gb de memória RAM e SO Oracle Linux 7.9.

O código da implementação da Sec4ML está disponível no GitHub[2]. A infraestrutura de toda a implementação da abordagem Sec4ML, com todos os componentes necessários e como eles podem se relacionar, pode ser visualizada na Figura 23.

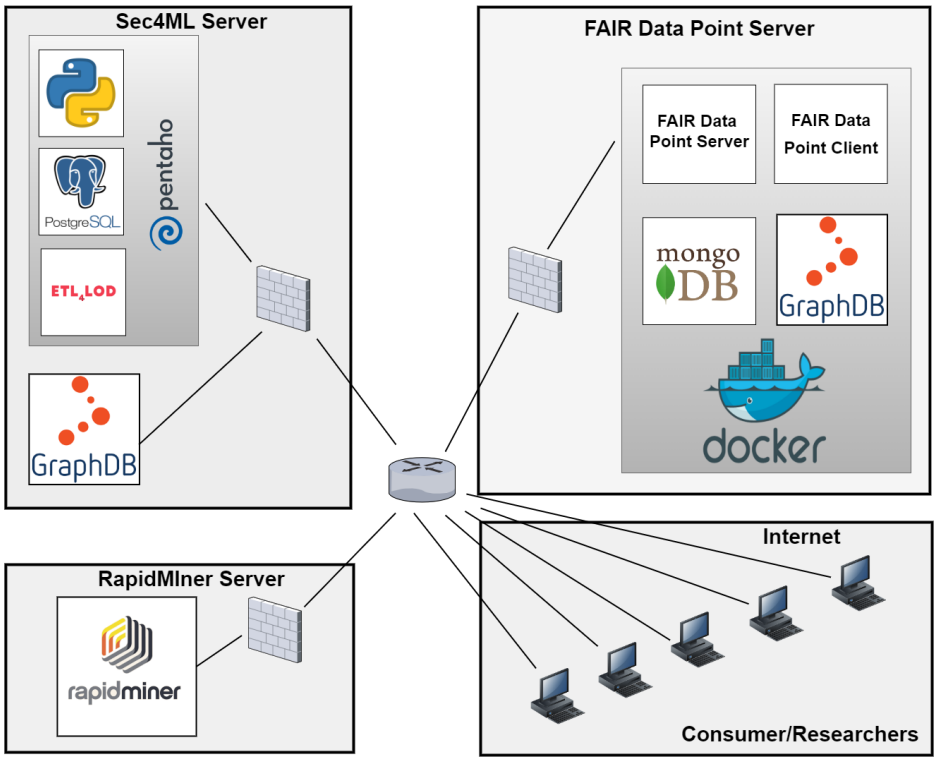


Figura 23 – Infraestrutura da Implementação da Sec4ML.

[2] <https://github.com/madalenals/Sec4ML>

5.2 Processo de ETL

Um dos dois macroprocessos que constituem a abordagem Sec4ML, o processo *Processing and Publication* se desdobra em 5 outros processos. A fim de tornar a implementação mais segmentada de acordo com os processos definidos para esta abordagem, estes foram implementados no workflow como *jobs*, um para cada processo existente. O outro macroprocesso, *Provenance Data Capture*, foi implementado em um único *job* de mesmo nome. A sua execução foi inserida dentro do fluxo do *Processing and Publication* no momento adequado, ou seja após todas as transformações que os dados devem receber. Uma visão geral do ETL pode ser visualizada ma Figura 24.

O *workflow Processing and Publication* é composto de *jobs* e transformações. O primeiro *job Dataset Transformation and Generation* realiza algumas tarefas de tratamento, inserção e modificação da estrutura inicial do *dataset*. A primeira transformação *InsertLog* é responsável pela inserção de um marcador na base de dados intermediária, denotando início de processamento, o qual possibilita que sejam realizadas diversas tarefas relacionadas à captura de dados de proveniência. Foi necessário a criação desta transformação devido ao fato que a ferramenta PDI só persiste os metadados sobre a execução de determinado *step* ou transformação ao final da execução do mesmo. Assim, esse marcador funciona como um controle, pois quando todos os *steps* tiverem sido concluídos, os metadados são inseridos nas tabelas definidas para tal persistência. Essa questão de funcionamento da ferramenta acarretou uma mudança na perspectiva do momento ideal para a captura dos metadados. Na prática, estes só estariam disponíveis para captura após finalizado seu respectivo *job* ou transformação. Desta forma, a captura é realizada dentro de janelas de tempo onde os metadados de determinada execução são persistidos, usando como marcador de início desta janela o registro inserido pelo *step InsertLog*.

Na sequência do *workflow*, as transformações seguintes realizam dinamicamente a captura do nome do arquivo tabular (.csv) informado para o processamento e dos metadados relacionados ao arquivo tabular como, por exemplo, nomes e tipos de atributos existentes. A intenção da captura dinâmica é garantir flexibilidade ao *workflow* para receber qualquer arquivo tabular como entrada de dados sem a necessidade de informar metadados necessários ao processamento explicitamente. Com base nestes metadados, um arquivo contendo um *script* de criação da tabela de apoio correspondente na área intermediária é gerado e as linhas de *inserts* correspondentes à cada registro encontrado no arquivo tabular. Este *script* é executado então criando e populando a tabela na área intermediária, de acordo com o definido na arquitetura.

A fim de atender ao princípio FAIR F3 (Os metadados devem incluir claramente e explicitamente os identificadores dos dados que descrevem), é acrescentada então uma coluna na tabela recém-criada para ser populada com códigos identificadores conhecidos como UUID (*Universally unique identifier*). Essa estrutura de geração de identificadores

únicos é definida pela RFC 4122 [3] e possui quatro versões. A versão utilizada nesta implementação é a geração de UUID através da função *gen_random_uuid* () do SGBD Postgres, a qual gera UUID randomicamente (versão 4). Esta coluna de identificação é populada em seguida, além de acrescentado um índice para esta coluna de UUID. A decisão de acréscimo posterior de uma coluna de UUID deve-se à captura dinâmica dos metadados de cada *dataset* a ser processado, tornando muito mais difícil, neste momento, esta alteração de estrutura.

Neste momento também é populada a tabela de proveniência *rowtable*, que representa as linhas de cada *dataset* processado. Esta última tabela receberá então um UUID para cada linha existente no *dataset* informado. Ao final deste *job* é executado o *job Get Attributes Values* onde é criado um laço de execução a fim de realizar a leitura dos metadados de cada atributo e os dados originais que são capturados na tabela *rowcolumnvalue*.

O próximo *job*, denominado *Data Anonymization*, é responsável pela aplicação de rotinas de anonimização sobre alguns atributos. É constituído por cinco transformações. A primeira delas realiza a leitura dos atributos a serem processados através de arquivo texto (XML) e a segunda transformação é a que realiza a extração dos dados da tabela correlata ao *dataset* criada na área intermediária. Uma vez extraídos os dados a serem processados, estes são submetidos a três transformações correspondentes às estratégias de anonimização: (i) para atributos identificadores (*ID Attributes Prefix Preserving Cryptography*); (ii) para atributos semi-identificadores (*Quasi Identifiers Data Treatment*) e (iii) para atributos sensíveis (*Sensitive Data Treatment*).

A transformação relativa ao tratamento de atributos identificadores foi implementada com aplicação de dois tipos de algoritmo: (i) baseado na estratégia de *prefix-preserving*, onde o prefixo do endereço IP, ou seja, a parte do endereço IP que representa a rede à qual esse IP pertence, é preservado e (ii) algoritmo de criptografia simétrica conhecido como Fernet.

No primeiro caso, a parte restante do endereço IP é anonimizada mas mantendo as características de um endereço IP e a capacidade de correlacionar endereços pertencentes à mesma rede. A escolha deste tipo de criptografia baseou-se na característica da facilidade de leitura humana dos endereços e a manutenção de características estatísticas do dado gerado como saída da rotina de cifração. A motivação principal é gerar uma versão do *dataset* informado criptografado e pré-processado a ser publicado para reúso

Já no segundo caso de implantação, o algoritmo escolhido gera uma cadeia aleatória de bytes sem semântica e de difícil leitura humana. Entretanto, com essa estratégia, é possível demonstrar que os registros originais podem ser alcançados através do uso da função reversa da criptografia, re-identificando os registros com os endereços IP originais.

[3] https://datatracker.ietf.org/doc/html/rfc4122

Este segundo caso de implementação foi gerado especificamente para este fim.

Com relação ao atributo semi-identificador, foi implementada a adição de ruído, um tipo de anonimização típico para atributos semi-identificadores. Por fim, ao atributo designado como sensível é aplicada a estratégia de supressão, onde todos os valores do atributo informado são eliminados. O *job* de anonimização está ilustrado na Figura 25.

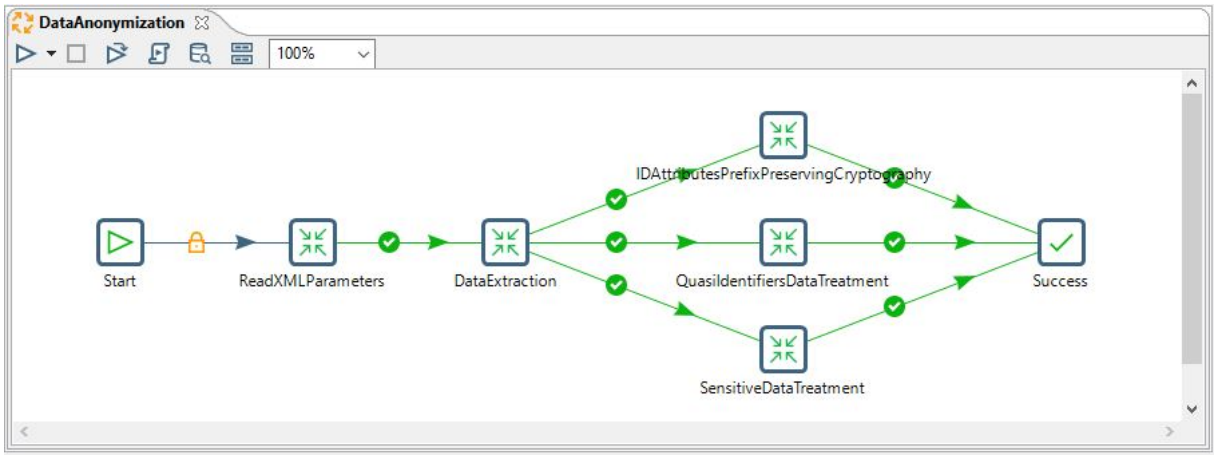


Figura 25 – ETL do Subprocesso *Data Anonymization*.

De maneira análoga ao processo de anonimização, o *job Preprocessing* aplica operadores de pré-processamento aos atributos informados. A leitura dos atributos informados é realizada pela transformação através da leitura de arquivo texto (XML). Existem diversos operadores de pré-processamento mas, por questões de limitações do trabalho, foram escolhidos três operadores: normalização, codificação e supressão. Foi implementada também uma opção de imputação de dados para dados ausentes (comumente referenciados como NaN em Python). A estrutura do *job Preprocessing* pode ser visualizada na Figura 26.

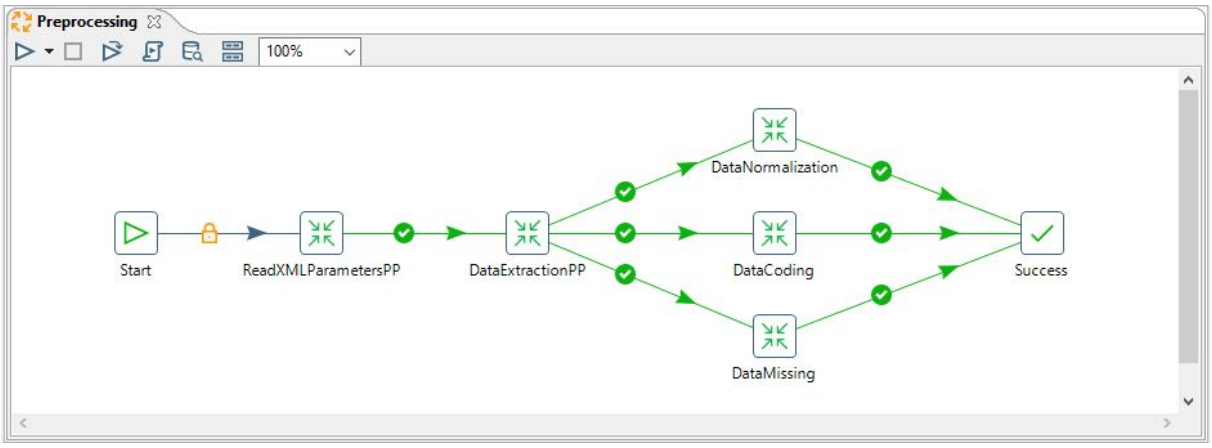


Figura 26 – ETL do Subprocesso *Data Preprocessing*.

A captura dos dados de proveniência gerados pelo processamento do ETL é realizada pelo *job Provenance Data Capture*. Este *job* implementa o processo de mesmo nome e é constituído por cinco etapas. Para a realização desta tarefa foi necessário utilizar como base a rotina de gravação de metadados realizada pelo PDI. Esta ferramenta ao final de cada execução persiste metadados sobre o *workflow* em tabelas em bases relacionais. Estes dados então são lidos e posteriormente gravados em diversas outras tabelas que compõem a estrutura de armazenamento de dados de proveniência.

Com a finalidade de persistir os valores utilizados como parâmetros durante a execução de algoritmos de anonimização, estes valores são persistidos em tabela específica também, como, no caso desta implementação, o valor da chave utilizada em algoritmos de criptografia e o valores definidos randomicamente como ruído a ser acrescido ao atributo designado.

Uma outra atividade que se fez necessária foi a leitura através de arquivo texto (XML) do nome da organização a ser utilizada para processamentos posteriores. Como nas tabelas de armazenamento de dados de proveniência podem haver ocorrências de diversas organizações, a solução encontrada foi a leitura por arquivo. Em seguida, após a coleta de vários dados necessários é realizada a persistência dos dados relativos à classe *PersonalData*. No última tarefa deste *job*, é realizada a persistência os dados de proveniência necessários à classe*LogicalAttribute*. A estrutura deste *job* pode ser visualizada na Figura 27.

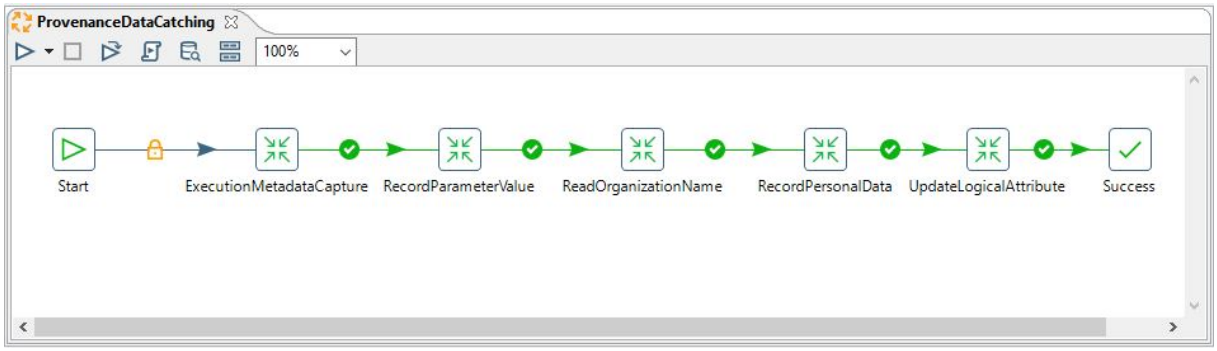


Figura 27 – ETL do Processo *Provenance Data Capture*.

O próximo *job Metadata Triplification* é o responsável pela geração dos dados já capturados e persistidos na área intermediária em formato de triplas (RDF). Esta captura foi organizada em agrupamentos de acordo com o grupo de metadados da ontologia Sec4ML-O. Assim, há quatro *jobs* que triplificam os dados de proveniência de acordo com este agrupamento. O primeiro destes é responsável pela triplificação dos dados de proveniência capturados relacionados com as classes da ontologia Sec4ML-O que representam conceitos de conformidade com a legislação de proteção de dados. O segundo, por sua vez, realiza a triplificação dos dados de proveniência capturados e relacionados com as classes que

representam conceitos inerentes ao *dataset* processado. O próximo *job* realiza a triplificação dos dados de proveniência capturados e relacionados com as classes que representam conceitos inerentes aos operadores utilizados no processamento do *dataset*. Por fim, o último *job* realiza a triplificação dos dados de proveniência capturados e relacionados com as classes que representam conceitos inerentes aos *workflows* executados necessários ao processamento do *dataset*. A geração dos dados triplificados utilizou componentes existentes no *plugin* ETL4LOD e foi realizada no formato *N-Triple*.

O último *job Generate Modular File* é responsável pela geração de um arquivo tabular (.csv) contendo os dados anonimizados e pré-processados armazenados na tabela correspondente ao *dataset* processado na área intermediária. Desta forma temos os dados processados, oriundos do *dataset*, em formato de arquivo tabular (.csv) pronto para a publicação e reúso.

Por restrições de tempo, não foi implementado o *workflow* correspondente ao componente da arquitetura *Cybersecurity Dataset Publisher* e ao processo *Metadata and Data Publication*, que faria a publicação automática dos dados e metadados triplificados. Entretanto, esta tarefa foi realizada de forma semi-automática, utilizando-se os arquivos *N-Triple* gerados pelo job anterior.

5.3 Considerações Finais

Este capítulo apresentou toda a infraestrutura da implementação da abordagem. descrevendo sua infraestrutura e processo de ETL, detalhando-se os recursos de hardware e *software* utilizados e como estes se relacionam.

A informação dos atributos a serem processados em cada tarefa da anonimização e pré-processamento foi implementada através da leitura de um arquivo texto em formato XML visando flexibilizar a entrada de dados. Esta escolha de implementação possibilita também, a qualquer momento, a alteração dos atributos informados para processamento sem a necessidade de alterações de código e/ou implementação da abordagem. Por exemplo, se o atributo identificador de um novo *dataset* for de diferente nomenclatura, este pode ser informado simplesmente alterando-se esta informação no arquivo texto.

No que tange aos algoritmos escolhidos para serem aplicados aos atributos designados, no caso de uma decisão de implementação diferente, faz-se necessário o desenvolvimento de tarefas específicas para cada algoritmo e/ou rotina escolhido. Uma vez implementada a nova tarefa, esta pode ser incorporada ao *workflow* já existente.

Com relação às dificuldades enfrentadas para implementar a abordagem Sec4ML, um dos principais problemas foi a configuração do *FAIR Data Point*, que, durante a primeira tentativa, onde estava sendo implementada a versão de produção, apresentou

a necessidade de serem realizadas algumas alterações necessárias de segurança de borda para que a instalação fosse completada. A máquina virtual que hospeda este serviço está disponibilizada na rede interna do IME (*cloud*) e seriam necessárias diversas solicitações e documentos para que essa configuração fosse realizada com sucesso. Diante deste cenário, optou-se pela instalação mais simples (instalação local) que não demandaria nenhuma intervenção de configuração de segurança de borda. Assim, este serviço de publicação foi disponibilizado somente para a demonstração dos casos de uso com acesso restrito à rede do IME.

Para viabilizar a aplicação desta abordagem foi necessária a adoção de algumas decisões no momento da implementação que viabilizassem, por exemplo, a captura de valores de parâmetros e a posterior reidentificação dos entes envolvidos. Uma destas decisões foi a implementação de anonimização utilizando-se apenas uma chave de criptografia para todo o processamento de cada *dataset*. Este valor de chave é informado dentro do ETL via parametrização em arquivo texto no formato XML. Para cada rotina de criptografia da linguagem escolhida, Python, há o uso de parâmetros diferentes e de maneira diferente. Por este motivo, foi escolhida o formato mais simples, com o algoritmo de Fernet, com a captura de apenas a chave informada previamente dentro do ETL. Outra questão deliberada foi a implementação da supressão sobre todos os valores do atributo informado, ao invés de aplicada a somente determinados valores escolhidos.

Outro grande obstáculo foi entender e gerir corretamente a captura dos metadados pela ferramenta PDI. Esta ferramenta possui uma sistemática em que os metadados são persistidos ao fim de cada etapa do ETL concluída. Desta forma, foi necessária a mudança da forma e o momento ideal para essa coleta, que é realizada a partir das tabelas designadas para armazenamento desses metadados. A intenção inicial era capturar os metadados gerados pela própria ferramenta ao longo de cada etapa executada do PDI. Devido ao exposto, foi necessário a mudança de toda a sistemática de captura dos metadados para que estes pudessem popular a estrutura de persistência de dados de proveniência.

E por fim, a dificuldade de realizar a configuração correta do ambiente do PDI e seus componentes coadjuvantes: Python, ETL4LOD e Postgres. O ambiente de desenvolvimento Python demanda a instalação de diversos partes que a compõem, assim como as bibliotecas necessárias e as utilizadas para a implementação específica de operadores de anonimização e pré-processamento. Após concluída esta etapa, ainda surgiram dificuldades de configuração de *paths* e variáveis de ambiente. Estas questões de configuração muitas vezes demandaram mais tempo que o planejado. Recomenda-se para uma execução do processamento dos *datasets* um ambiente que tenha disponível mais memória do que a utilizada neste ambiente

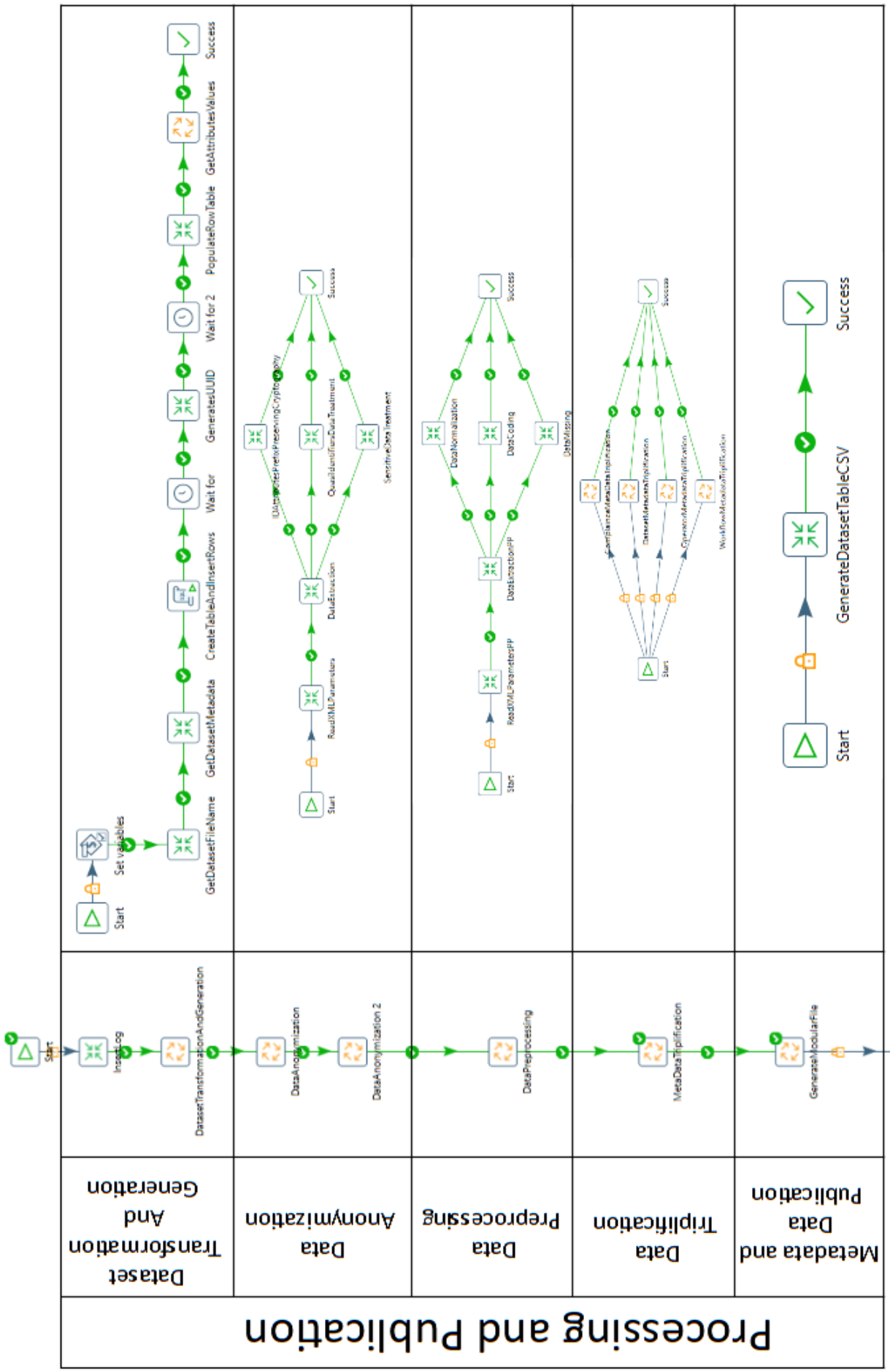


Figura 24 – ETL do Processo *Processing and Publication* e seus subprocessos.

6 APLICAÇÃO DA ABORDAGEM SEC4ML

Com o objetivo de viabilizar a aplicação do processo da abordagem Sec4ML com seus subprocessos de anonimização e aplicação dos pré-operadores de KDD, foram realizadas duas aplicações da abordagem sobre dois conjuntos de dados de *benchmark* públicos e rotulados contendo eventos de incidentes de segurança da informação: (i) o produzido pelo *Australian Centre for Cyber Security* (ACCS) da *University of New South Wales* denominado UNSWNB-15[1] (16) e (ii) o da *University of New Brunswick*, criado pelo *Canadian Institute for Cybersecurity* denominado CSE-CIC-IDS2018[2] (62). Esta estratégia teve como objetivo demonstrar o impacto que a abordagem causa nos resultados de criação de modelos de AM e sua independência com relação aos dados utilizados. Assim, cada conjuntos de dados foi submetido ao AM utilizando-se os dados originais, assim como utilizando-se os dados processados pela aplicação da abordagem Sec4ML, a fim de que seja realizado um comparativo com os resultados de ambos os processos de classificação.

Como metodologia para a escolha dos *datasets* mais apropriados foram levados em consideração alguns aspectos propostos por Gharib et al. (63), tais como a diversidade dos dados coletados, no que concerne a protocolos, o volume desses dados, os tipos de ataques considerados; a presença ou ausência de atributos informativos; se há tráfego real e o tamanho do conjunto de atributos coletados. Os critérios propostos por Sharafaldin et al. (64) também foram observados. Neste trabalho, os autores propuseram 11 critérios a serem considerados na escolha de um *dataset*: (i) diversidade de ataques, (ii) anonimato, (iii) protocolos disponíveis, (iv) captura completa, (v) interação completa, (vi) configuração de rede completa, (vii) tráfego completo, (viii) conjunto de atributos, (ix) heterogeneidade (do tráfego de rede e logs de sistema), (x) etiquetagem correta e (xi) a presença de metadados. Foram levadas em consideração também questões como a quantidade e qualidade de trabalhos que utilizaram esses datasets para experimentos, assim como o ano em que os *datasets* foram criados. Dessa forma, priorizou-se *datasets* bem difundidos na comunidade de pesquisa e que fossem o mais recentes possível.

As seguintes subseções detalham esses conjuntos de dados e explicitam de que forma é possível usá-los na demonstração da aplicabilidade da abordagem Sec4ML.

6.1 Caso de Aplicação do Conjunto de Dados UNSW-NB15

O conjunto de dados UNSW-NB15 é um conjunto de dados de incidentes de segurança da informação gerados sinteticamente pelo *Australian Centre for CyberSecurity*

[1] https://research.unsw.edu.au/projects/unsw-nb15-dataset
[2] https://www.unb.ca/cic/datasets/ids-2018.html

(ACCS)[3]. Na sua criação foi realizada uma combinação de atividades de ataques reais e sintéticos e os dados foram coletados através do uso da ferramenta IXIA PerfectStorm. O *dataset* contém 2.540.044 registros que estão disponibilizados em quatro arquivos CSV, sendo os três primeiros contendo 700.000 registros e o quarto arquivo contendo 440.044 registros. Esses dados coletados são referentes a oito tipos de ataques, conforme elencado na Tabela 5. Apesar das características citadas, este *dataset* é desbalanceado com relação à distribuição das categorias de incidentes.

Tabela 5 – Distribuição dos Registros do *dataset* UNSW-NB15 de acordo com o tipo de ataque. Adaptado de (16).

Tipo de Ataque	Quantidade de Registros
Benigno	2.218.761
Fuzzers	24.246
Analysis	2.677
Backdoors	2.329
DoS	16.353
Exploits	44.525
Generic	215.481
Reconnaissance	13.987
Shellcode	1.511
Worms	174
Total	2.540.044

6.1.1 Atributos

O *dataset* é composto por 49 atributos, os quais são divididos em seis grupos: (i) *Flow Features*, (ii) *Basic Features*, (iii) *Content Features*, (iv) *Time Features*, (v) *Additional Generated Features* e (vi) *Labelled Features*. Entre seus 49 atributos, abaixo encontram-se descritos alguns dos mais relevantes para exemplificação. A relação completa dos atributos existentes encontra-se no Anexo A.

- *srcip*: endereço IP de origem;
- *sport*: porta que originou o evento registrado;
- *dstip*: endereço IP de destino;
- *dsport*: porta de destino relacionado ao evento registrado;
- *proto*: protocolo utilizado no evento registrado;
- *state*: estado relacionado com o evento registrado;
- *dur*: duração total do evento;
- *sbytes*: número de bytes envolvidos no evento desde a origem até o destino;

[3] http://www.accs.unsw.adfa.edu.au/

- *dbytes*: número de bytes envolvidos no evento desde o destino até a origem;
- *attack_cat*: categoria do ataque; e
- *label*: *flag* que apresenta o valor (1) quando o evento registrado é um ataque ou (0) quando o evento é uma atividade normal.

Alguns trabalhos se dedicaram a evidenciar quais atributos seriam os mais relevantes para uso em criação de modelos de AM e predição de incidentes para IDS utilizando-se o *dataset* UNSW-NB15 (65) e (66). Nesse primeiro trabalho foram salientados os atributos [sttl], [ct_dst_src_ltm], [spkts], [dload], [sloss], [dloss], [ct_src_ltm] [ct_srv_dst] como sendo os mais comumente utilizados. Entretanto, outro grupo de atributos foi apontado como mais significantes para AM: [service], [sbytes], [sttl], [smeanz] e [ct_dst_sport_ltm]. Este último conjunto de atributos foi o escolhido para ser usado na criação do modelo de treinamento na ferramenta *RapidMiner* utilizando-se os dados originais, assim como utilizando-se os dados processados pela aplicação da abordagem Sec4ML, a fim de que seja realizado um comparativo com os resultados de ambos os processos de classificação.

6.1.2 Aplicação da abordagem Sec4ML - Estratégia *Prefix-Preserving*

O *dataset* UNSW-NB15 recebeu tratamento de criptografia conhecido como *Prefix-Preserving* (67) (68) (69). Esta estratégia, conhecida como Crypto-PAn, foi aplicada através da biblioteca Python *yacryptopan*. O algoritmo de criptografia foi aplicado sobre o atributo [dstip]. A chave que deve ser utilizada na aplicação deste algoritmo é informada por arquivo de configuração (XML), agregando flexibilidade à execução do ETL. A estrutura do arquivo de parâmetros pode ser visualizada na Figura 28.

Figura 28 – Atributos informados através de arquivo de parâmetros XML para o processamento do *dataset* UNSW-NB15.

O atributo [sttl] recebeu anonimização por adição de ruído através da função *np.random* da biblioteca *Python nunpy*. Já os atributos [srcip] e [service] receberam anonimização por supressão total do atributo. O atributo [ct_dst_sport_ltm] sofreu um processo de normalização como tarefa de pré-processamento. A tarefa de pré-processamento

de codificação foi aplicada sobre o atributo [attack_cat], de forma a transformar os valores categóricos em números, permitindo o uso deste atributo como *label* para tarefas de classificação multi-classes.

Devido a limitações de memória RAM e espaço em disco, os aquivos originais (UNSWNB15_1, UNSWNB15_2, UNSWNB15_3 e UNSWNB15_4) foram divididos em arquivos menores contendo 100.000 registros cada. Foram processados somente os registros contidos nos arquivos UNSWNB15_1 e UNSWNB15_2 (cerca de 1.400.000). Cada segmento de 100.000 registros necessitou de cerca de 5 horas para ser processado totalmente, conforme pode ser visualizado na Figura 29.

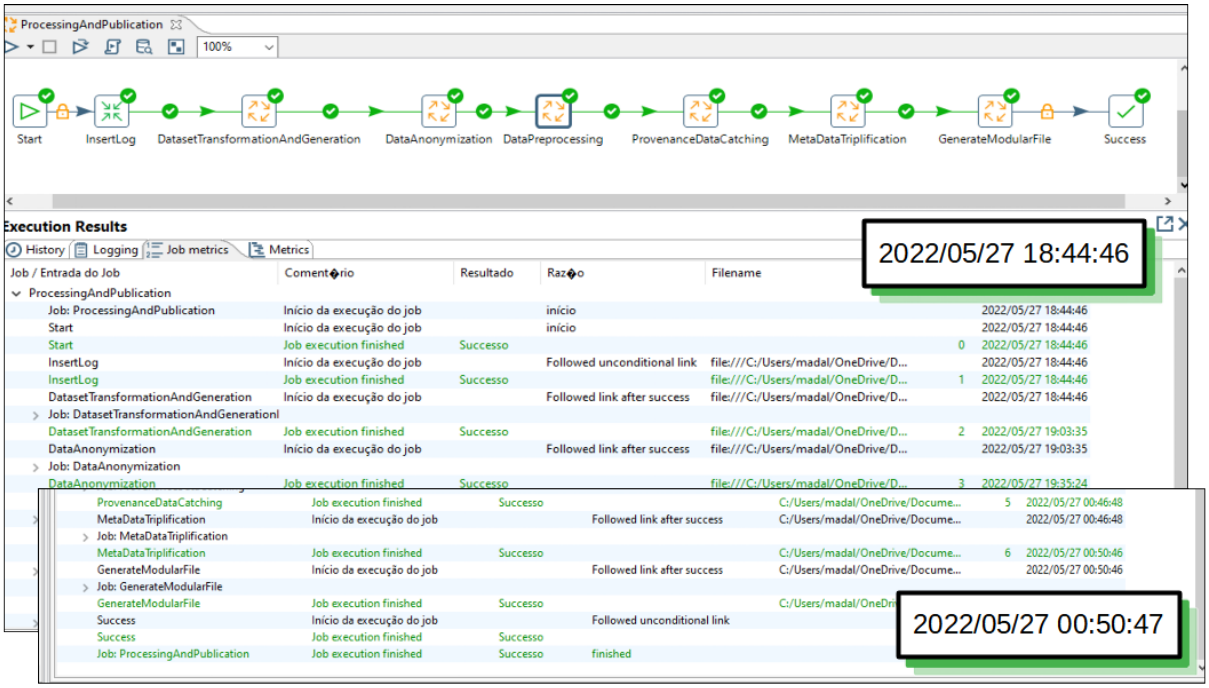


Figura 29 – Métricas de execução do ETL para o *dataset* UNSW-NB15.

6.1.2.1 Resultados observados

Os experimentos de geração de modelos de ML foram realizados em duas etapas: (i) o processamento na ferramenta *RapidMiner* (utilizando-se a opção *Auto Model*) do *dataset* original e (ii) do *dataset* correspondente após o processamento com a implementação da abordagem Sec4ML. Após algumas rodadas do experimento, verificou-se que os modelos que apresentaram melhor desempenho com o *dataset* original foram: (i) *Logistic Regression*, (ii) *Gradient Boosted Trees*, (iii) *Deep Learning* e (iv) *Generalized Linear Model*.

Considerando-se o *dataset* original, o modelo em *Deep Learning* apresentou a melhor acurácia (*Accuracy*) com 98.4%, seguido pelos modelos *Logistic Regression* e *Gradient Boosted Trees* que obtiveram 97.9% de acurácia. O último valor para o indicador de acurácia foi o do modelo *Generalized Linear Model* com 97.8%. O maior valor para o

indicador *Area Under the Curve* (AUC) obtido pela geração deste modelo foi também obtido com o modelo *Deep Learning* com 99.6%, seguido do modelo em *Gradient Boosted Trees* com 99.3% e o modelo em *Generalized Linear Model* com 99.0%. O menor valor para este indicador foi obtido pelo modelo em *Logistic Regression* com 98.5%. O melhor valor para o indicador de precisão (*Precision*) foi do modelo *Deep Learning* com 99.6% seguido dos modelos *Gradient Boosted Trees* e *Logistic Regression* com 99.1%. O pior valor para este indicador foi o do modelo *Generalized Linear Model* com 99.0%. Por sua vez, o indicador 'F' (*F Measure*) obteve valores de 99.1% para o modelo *Deep Learning* seguido dos modelos *Logistic Regression* e *Gradient Boosted Trees* com 98.8%. O menor valor do indicador 'F' foi obtido pelo modelo *Generalized Linear Model* com 98.7%. Um gráfico comparativo com estes indicadores pode ser visualizado na Figura 30.

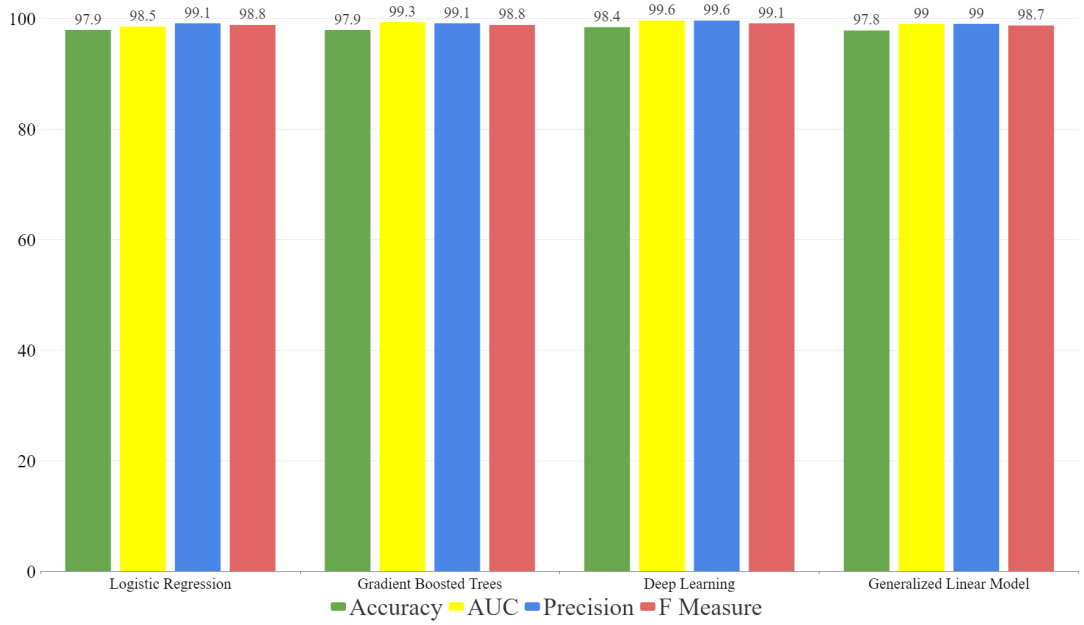


Figura 30 – Indicadores obtidos relacionados com a criação do modelo AM por algoritmo do *dataset* UNSW-NB15 original em percentuais.

Já na criação do modelo de AM com os dados processados pela Sec4ML, os atributos escolhidos como entrada de dados para o modelo se apresentam com as seguintes modificações:

- ***sttl*** - sem modificações;
- ***ct_dst_sport_ltm*** - normalizado;
- ***sbytes*** - sem modificações;
- ***smeansz*** - sem modificações;
- ***service*** - removido por supressão.

Desta forma, para esta criação do modelo com os dados processados foram utilizados quatro atributos ao invés de cinco, em comparação com a criação dos modelos com os dados originais.

O modelo com melhor desempenho apresentado foi o *Deep Learning* obtendo o percentual de 88.4% de acurácia, seguido do modelo *Gradient Boosted Trees* com 88.0% de acurácia. O próximo melhor desempenho foi o do modelo *Generalized Linear Model* com 77.6% de acurácia seguido por último do modelo *Logistic Regression* com 61.0% deste mesmo indicador. O indicador AUC obtido pelos modelos *Logistic Regression* e *Generalized Linear Model* foi 95.0% . O próximo valor obtido para AUC foi do modelo *Gradient Boosted Trees* com o valor de 94.3%. O útlimo resultado do indicador AUC foi o obtido pelo modelo *Deep Learning* com o valor 91.9%. O melhor valor obtido para o indicador de precisão (*Precision*) foi o do modelo *Generalized Linear Model* com 99.0% seguido do modelo *Logistic Regression* com 98.9%. O modelo *Gradient Boosted Trees* obteve precisão de 98.0% seguido por último pelo modelo *Deep Learning* com 88.6%. O indicador 'F' obtido com melhor valor foi o do modelo *Deep Learning* com 93.7% seguido pelo modelo *Gradient Boosted Trees* com 92.8%. Os dois últimos valores do indicador 'F' foram dos modelos *Generalized Linear Model* e *Logistic Regression* com os valores 85.5% e 71.5% respectivamente. Estes indicadores e os valores obtidos podem ser visualizados na Figura 31.

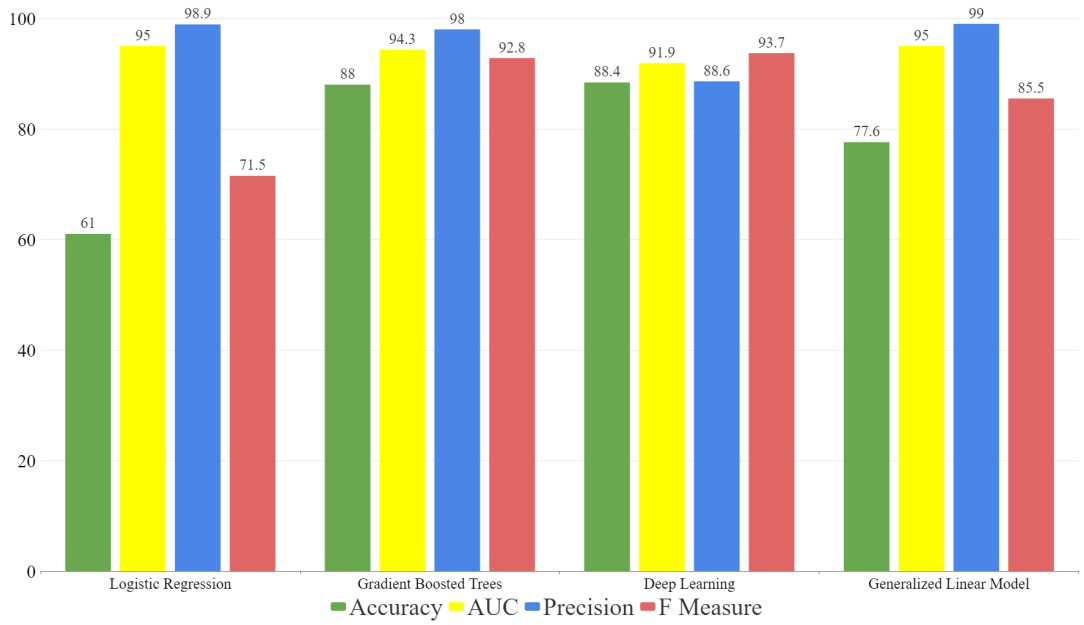


Figura 31 – Indicadores obtidos com a criação do modelo AM por algoritmo do *dataset* UNSW-NB15 processado em percentuais.

No cap 6 linha 143 sugiro reescrever parte do parágrafo da seguinte forma: ...

Pode-se observar também que a criação de modelos de AM a partir do *dataset* UNSW-NB15 processado foi realizada com um atributo a menos (*service*). Não foi possível

repetir o experimento com o *dataset* contendo este atributo pois, ao final do processamento da ETL da abordagem Sec4ML, este atributo não existia mais. Desta forma, há uma possibilidade de que a perda observada nos indicadores está relacionada à retirada deste atributo. Entretanto não foi possível demonstrar que tal relação é verdadeira dentro do escopo deste trabalho. Podem ser visualizados na Tabela 6 os valores obtidos dos indicadores pela criação de modelos AM com o *dataset* original e com os dados processados após a execução do ETL.

Tabela 6 – Tabela comparativa dos resultados obtidos com a criação do modelo para o *dataset* UNSW-NB15 original e processado em percentuais.

	Raw				Processed			
Model	Accuracy	AUC	Precision	F Measure	Accuracy	AUC	Precision	F Mesasure
Logistic Regression	97.9	98.5	99.1	98.8	61.0	95.0	98.9	71.5
Gradient Boosted Trees	97.9	99.3	99.1	98.8	88.0	94.3	98.0	92.8
Deep Learning	**98.4**	**99.6**	**99.6**	**99.1**	**88.4**	**91.9**	**88.6**	**93.7**
Generalized Linear Model	97.8	99.0	99.0	98.7	77.6	95.0	99.0	85.5

6.1.3 Aplicação da abordagem Sec4ML - Estratégia Criptografia Simétrica

Com o objetivo de demonstrar a possibilidade de, através de dados persistidos de proveniência, realizar a reidentificação de um determinado ente envolvido, neste experimento o atributo identificador criptografado foi o [dst_ip] ou endereço IP de destino do evento registrado. Foram escolhidos dez registros para este experimento. O algoritmo aplicado foi *Fernet* utilizando-se a biblioteca *cryptography* com o método *Fernet*. Após o processamento pela abordagem Sec4ML este atributo é transformado em uma cadeia de *bytes* sem semântica. Os valores de identificador do registro [uuid] e [dst_ip] resultantes do processamento podem ser visualizados na Figura 32.

	uuid uuid	dstip character varying (1000)
1	70984266-f766-4afc-aea6-c9d61941b2ed	b'gAAAAABirkNRF1JmcGR4yOPmPbpMYm5C5IRP-4TndVvB0l-uk7kNILPpiYX5NJm4Kq_US-SkQmTJuRAkYZxt2VAMFhkovU6o7Q=='
2	13c16a79-8ec8-4472-88e2-f7dfd0265e2f	b'gAAAAABirkNRcyf0jhW1ofxBtTWhZlCTzbQ2ayhVp1HXvHij-SBRVFyYoJ6JHzcSBeuRm9Q4ftCfbCLCT3PI7eiZgdfdYT0zwg=='
3	211eb855-5a91-42b4-b25b-a60d29de4522	b'gAAAAABirkNRrkqciH0lhMgLiS8Idt9BNpTBi5FS6iyKNNy00hy2roqYTk4pNvkY1-Ee6QpyRMDWw3UD93U8-k0ocfBmLjOTdA=='
4	b487f38b-5d49-40c7-a0e3-e2e673e0473a	b'gAAAAABirkNRpgQKwqbQrm6HtTEXqd07CM5gsU0lNrkG9Sbnt9HQ5QbdFcKw9rrVaJDhlTwukAyKTVQStZPkzGMHvAon3knpTA=='
5	caf18dc6-860f-427f-a54f-6bdf9fc7d056	b'gAAAAABirkNRBA_bNowV9g7v1lx7lZoTWKHxQn9OkLH0FeQqreoYiAaagIAKtrCbyd0dgiQ0g7wdyySnQ1Md7e8AMQVLsWG7fA=='
6	dc137b2a-fb37-4f9e-ab54-913e02b2cbfe	b'gAAAAABirkNRqh0bkjJC9bY3LvdG45G2o3p7HbWZW5AP82p2OTaYtw0r6qa8npThj2XPwzdb9zbh5GMITqujf5SFBWVL6mKLJw=='
7	e3c39b8d-aba8-4845-a027-0bf09b67501f	b'gAAAAABirkNRh_GRsAssvcQW4nGzxwFsv6NszYMYTLbknsZC6PDfN_hx4OmKrRfOEa0u3xky-Ap2pneXTPVjVVdPnctM_xFDiw=='
8	ef123609-1837-4e54-bf68-6acfed4b06ca	b'gAAAAABirkNRBKTv0ZRUSEuXiNy4b4ljRhjenXITMUVnqZJE-ylef5-w8jjjGGR1DcTzjgsUVt1cHyM7Fy9lMaMlkN7RZkw2hA=='
9	f167504c-0e50-4bec-b6e3-d10d0de1f3dc	b'gAAAAABirkNS4xV5NYsgzM_DBAC9zzzTo7WJ0A9cri4VF34kOlOFpE04XIi-XeXFuNPnSb4SHzE7LplB7X8rJM3kmbATEJAggg=='
10	6ff3cda7-bbe9-4c27-b476-47b08786ae8d	b'gAAAAABirkNRtnNC3pXBrD3LWP_fyMG3ZjkjYxmjJTsUV6ii-DCqxZFmdE-X55L-vTyJ9q1Zn7iqctGO8fSDnmFoBx1LeJkwdA=='

Figura 32 – Registros dos *dataset* UNSW-NB15 selecionados para o experimento de reidentificação.

Utilizando-se dos dados obtidos através da Consulta 3 referenciada na Seção 6.3, podemos obter a chave utilizada em determinado processamento do ETL. De posse desta

chave, é possível realizar a reidentificação do endereço IP envolvido em determinado incidente de segurança da informação.

O processo de criptografia foi realizado no ambiente de desenvolvimento *Python JupiterLab* através da construção de um *data frame* Python contendo os dados referenciados na Figura 32. O código Python e os dados originais descriptografados podem ser visualizados na listagem abaixo.

```
01 | from cryptography.fernet import Fernet
02 | from base64 import b64encode
03 | import pandas as pd
04 |
05 | dfSec4ML = pd.DataFrame({'A':
06 | ['70984266-f766-4afc-aea6-c9d61941b2ed',
07 | '13c16a79-8ec8-4472-88e2-f7dfd0265e2f',
08 | '211eb855-5a91-42b4-b25b-a60d29de4522',
09 | 'b487f38b-5d49-40c7-a0e3-e2e673e0473a',
10 | 'caf18dc6-860f-427f-a54f-6bdf9fc7d056',
11 | 'dc137b2a-fb37-4f9e-ab54-913e02b2cbfe',
12 | 'e3c39b8d-aba8-4845-a027-0bf09b67501f',
13 | 'ef123609-1837-4e54-bf68-6acfed4b06ca',
14 | 'f167504c-0e50-4bec-b6e3-d10d0de1f3dc',
15 | '6ff3cda7-bbe9-4c27-b476-47b08786ae8d'],
16 |
17 | 'B':
18 | [b'gAAAAABirkNRF1JmcGR4yOPmPbpMYm5C5IRP-4TndVvB0I......',
19 | b'gAAAAABirkNRcyf0jhW1ofxBtTWhZlCTzbQ2ayhVp1HXvHi......',
20 | b'gAAAAABirkNRrkqciH0IhMgLiS8Idt9BNpTBi5FS6iyKNNy......',
21 | b'gAAAAABirkNRpgQKwqbQrm6HtTEXqd07CM5gsU0INrkG9Sb......',
22 | b'gAAAAABirkNRBA_bNowV9g7v1lx7IZoTWKHxQn9OkLH0FeQ......',
23 | b'gAAAAABirkNRqh0bkjJC9bY3LvdG45G2o3p7HbWZW5AP82p......',
24 | b'gAAAAABirkNRh_GRsAssvcQW4nGzxwFsv6NszYMYTLbknsZ......',
25 | b'gAAAAABirkNRBKTv0ZRUSEuXiNy4b4ljRhjenXITMUVnqZJ......',
26 | b'gAAAAABirkNS4xV5NYsgzM_DBAC9zzzTo7WJ0A9cri4VF34......',
27 | b'gAAAAABirkNRtnNC3pXBrD3LWP_fyMG3ZjkjYxmjJTsUV6i......']
28 | })
29 |
30 | key = 'The-sky-is-blue-but-some-times-1'
31 | bkey = bytearray(key, 'utf-8')
32 | password = b64encode(bkey)
33 |
34 | fernet= Fernet(password)
35 |
36 | defernet= Fernet(password)
37 |
38 | for i in range(0,len(dfSec4ML.index)):
39 |
40 |         dfSec4ML.iloc[i][1] =
```

```
            str(defernet.decrypt(dfSec4ML.iloc[i][1]))

dfSec4ML

                    A                                       B
0       70984266-f766-4afc-aea6-c9d61941b2ed    b'149.171.126.9'
1       13c16a79-8ec8-4472-88e2-f7dfd0265e2f    b'149.171.126.4'
2       211eb855-5a91-42b4-b25b-a60d29de4522    b'10.40.182.3'
3       b487f38b-5d49-40c7-a0e3-e2e673e0473a    b'149.171.126.0'
4       caf18dc6-860f-427f-a54f-6bdf9fc7d056    b'149.171.126.6'
5       dc137b2a-fb37-4f9e-ab54-913e02b2cbfe    b'149.171.126.9'
6       e3c39b8d-aba8-4845-a027-0bf09b67501f    b'149.171.126.5'
7       ef123609-1837-4e54-bf68-6acfed4b06ca    b'149.171.126.6'
8       f167504c-0e50-4bec-b6e3-d10d0de1f3dc    b'149.171.126.4'
9       6ff3cda7-bbe9-4c27-b476-47b08786ae8d    b'149.171.126.7'
```

Desta forma, é possível demonstrar que, através da captura e persistência de dados de proveniência das execuções dos *workflows* é possível, caso necessário, realizar a reidentificação de entes envolvidos que tenham dados processados pela abordagem Sec4ML. Esta reidentificação possibilita o resgate dos dados originais através dos atributos identificadores para fins de investigação, por exemplo.

6.2 Caso de Aplicação Conjunto de Dados CSE-CIC-IDS2018

O conjunto de dados CSE-CIC-IDS2018[4] (64) é um *dataset* derivado da evolução do *dataset* ISCXIDS2012, criado pelo *Information Security Centre of Excellence* (ISCX) na *University of New Brunswick* (UNB). Em 2017, os criadores do ISCXIDS2012 e o *Canadian Institute of Cybersecurity* (CIC) trabalharam para o desenvolvimento do *dataset* CICIDS2017, corrigindo diversos problemas que o *dataset* anterior apresentava. Em 2018, com a participação do *Communications Security Establishment* (CSE) foi gerada uma nova versão do CICIDS2017, o CSE-CIC-IDS2018, mais completo e contendo mais instâncias. Considerado hoje o mais atual dos *datasets* de *benchmark*, ele é usado em muitos trabalhados recentes (5) (62), principalmente os voltados para a melhoria de IDS através do uso de modelos de AM. A distribuição dos registros deste dataset encontra-se na Tabela 7.

[4] https://www.unb.ca/cic/datasets/ids-2018.html

Tabela 7 – Distribuição dos Registros do *dataset* CSE-CIC-IDS2018 de acordo com o tipo de ataque. Adaptado de (5).

Tipo do Incidente	Quantidade de Registros
Benigno	2.856.035
BOT	286.191
Brute Force	286.191
DoS	1.289.544
Infiltration	286.191
Total	5.290.343

6.2.1 Atributos

O *dataset* é composto pela coletânea de 10 arquivos, sendo 9 deles apresentando 79 atributos e o décimo arquivo contendo 83 atributos que representam 7 categorias de ataques: Benigno, *Brute Force*, *Botnet*, *DoS*, *DDoS*, *Web attacks* e *Infiltration*.

Dentre esses 83 atributos, abaixo encontram-se descritos alguns atributos para exemplificação. A relação completa dos atributos existentes encontra-se no Anexo B.

- flow_duration: Duração do fluxo de dados;
- total_fwd_packet: Total de pacotes no sentido do fluxo;
- total_bwd_packets: Total de pacotes no sentido inverso do fluxo;
- total_len_fwd_packet: Tamanho total do pacote no sentido do fluxo;
- total_len_bwd_packet: Tamanho total do pacote no sentido inverso do fluxo:
- fwd_packet_len_min: Tamanho mínimo do pacote no sentido do fluxo;
- fwd_packet_len_max: Tamanho máximo do pacote no sentido do fluxo;
- fwd_packet_len_mean: Tamanho médio do pacote no sentido do fluxo;
- fwd_packet_length_std: Desvio padrão do tamanho do pacote no sentido do fluxo;
- bwd_packet_len_min: Tamanho mínimo do pacote no sentido inverso do fluxo;
- bwd_packet_len_max: Tamanho máximo do pacote no sentido inverso do fluxo;
- bwd_packet_len_mean: Tamanho médio do pacote no sentido inverso do fluxo;
- bwd_packet_len_std: Desvio padrão do tamanho do pacote no sentido inverso do fluxo.

6.2.2 Aplicação da abordagem Sec4ML - Estratégia *Prefix-Preserving*

Dentre os dez arquivos que compõem o *dataset* CSE-CIC-IDS2018, para os experimentos foram usados 1.500.000 registros oriundos do arquivo *Thuesday20022018_TrafficForML_CICFlowMeter.csv* contemplando 1.400.000 registros de tráfego benigno e 100.000 de ataques DDoS. Assim, como somente este arquivo continha atributos categóricos identificadores, para fins de equivalência com a aplicação da abordagem Sec4ML refenciada na Seção 6.1, foram utilizados somente os registros contidos neste arquivo.

Até onde foi possível investigar, não foram encontrados trabalhos que explicitaram os melhores atributos a serem aplicados na criação de modelos de AM. Após diversas rodadas deste experimento, os melhores resultados foram obtidos com a aplicação de quase todos os atributos do *dataset* exceto *flow_id*, *src_ip*, *src_port*, *dst_ip*, *dst_port*, *protocol* e *timestamp*.

De maneira análoga ao primeiro caso de aplicação, este total de registros foi dividido em partes menores contendo 100.000 registros devido a restrições de memória do ambiente utilizado.

O *dataset* CSE-CIC-IDS2018 recebeu tratamento de criptografia conhecido como *Prefix-Preserving* . Esta estratégia foi aplicada através da biblioteca Python *yacryptopan*. O algoritmo de criptografia foi aplicado sobre o atributo [dst_ip]. A chave que deve ser utilizada na aplicação deste algoritmo é informada por arquivo de configuração (XML), agregando flexibilidade à execução do ETL. A estrutura do arquivo de parâmetros pode ser visualizada na Figura 33.

Figura 33 – Atributos informados através de arquivo de parâmetros XML para o processamento do *dataset* CSE-CIC-IDS2018.

O atributo [flow_duration] recebeu anonimização por adição de ruído através da função *np.random* da biblioteca *Python nunpy*. Já os atributos [src_ip], [src_port] e [fwd_pkt_len_mean] receberam anonimização por supressão total do atributo. O atributo [flow_duration] sofreu um processo de normalização como tarefa de pré-processamento. A tarefa de pré-processamento de codificação foi aplicada sobre o atributo [label], de forma a transformar os valores categóricos em números.

6.2.2.1 Resultados observados

Considerando-se o *dataset* original, o modelo *Generalized Linear Model* apresentou a melhor acurácia com 93.3%, seguido pelo modelo *Generalized Linear Model* com 78.6%. O modelo *Deep Learning* apresentou o valor de 73.2% e, por último, o modelo *Logistic Regression* com 6.7%. O maior valor para o indicador AUC obtido foi o modelo *Deep Learning* com 100% seguido dos modelos *Gradient Boosted Trees* e *Generalized Linear Model* com 99.8% cada um. Por último, o modelo *Logistic Regression* apresentou 50% de AUC. O melhor valor para o indicador de precisão foi obtido pelos modelos *Logistic Regression*, *Deep Learning* e *Generalized Linear Model* com 100%. O modelo *Generalized Linear Model* apresentou AUC de 93.3%. Por sua vez, o indicador 'F' obteve valores de 96.6% para o modelo *Gradient Boosted Trees*, seguido de 87.1% para o modelo *Generalized Linear Model*, 83,3% para o modelo *Deep Learning* e 0% para o modelo *Logistic Regression*. Um gráfico comparativo com estes indicadores pode ser visualizado na Figura 34.

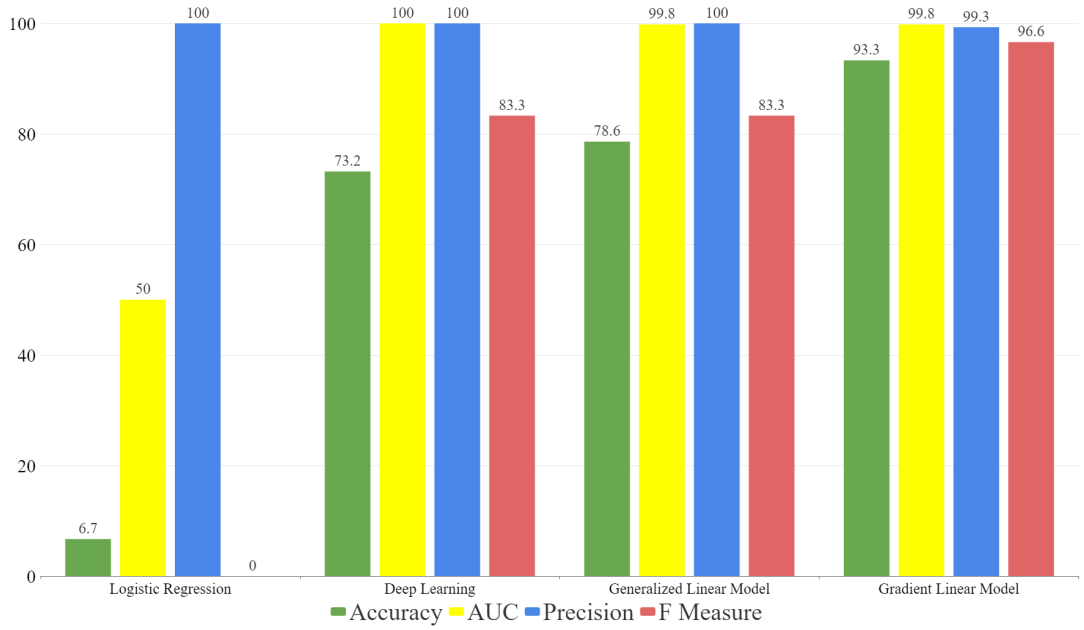


Figura 34 – Indicadores obtidos com a criação do modelo AM por algoritmo do *dataset* CSE-CIC-IDS2018 original em percentuais.

Por outro lado, na criação do modelo de AM com os dados processados pela Sec4ML, os atributos escolhidos como entrada de dados para o modelo são os mesmos escolhidos no processamento do *dataset* original, ou seja, quase todos excetuando-se os já citados e o atributo *uuid*.

O maior valor obtido de acurácia na criação do modelo com os dados do *dataset* CSE-CIC-IDS2018 foi apresentado pelo modelo *Generalized Linear Model* com 95.3%, seguido pelo modelo *Logistic Regression* com 95.0%. A acurácia apresentada pelo modelo *Gradient Boosted Trees* foi 82.4% seguido pelo modelo *Deep Learning* com 23.6%. O maior valor para o indicador AUC obtido também foi o modelo *Generalized Linear Model*

com 86.9% seguido dos modelos *Deep Learning* e *Logistic Regression* com 75.9% e 67.9% respectivamente. O menor valor obtido para o indicador AUC foi o do modelo *Gradient Boosted Trees* com 64.7%.

O melhor valor para o indicador de precisão foi obtido pelo modelo *Deep Learning* com 99.3% seguido pelo modelo *Generalized Linear Model* com 95.8%. O menor valor obtido pelo indicador foi o dos modelos *Logistic Regression* e *Gradient Boosted Trees* com 94.9%. O melhor valor para o indicador 'F' foi obtido pelo modelo *Generalized Linear Model* com 97.6% seguido do modelo *Logistic Regression* com 97.4%. O valor seguinte é de 90.1% obtido pelo modelo *Gradient Boosted Trees* e o último valor foi o do modelo *Deep Learning* com 30.8%. Um gráfico comparativo com estes indicadores pode ser visualizado na Figura 35.

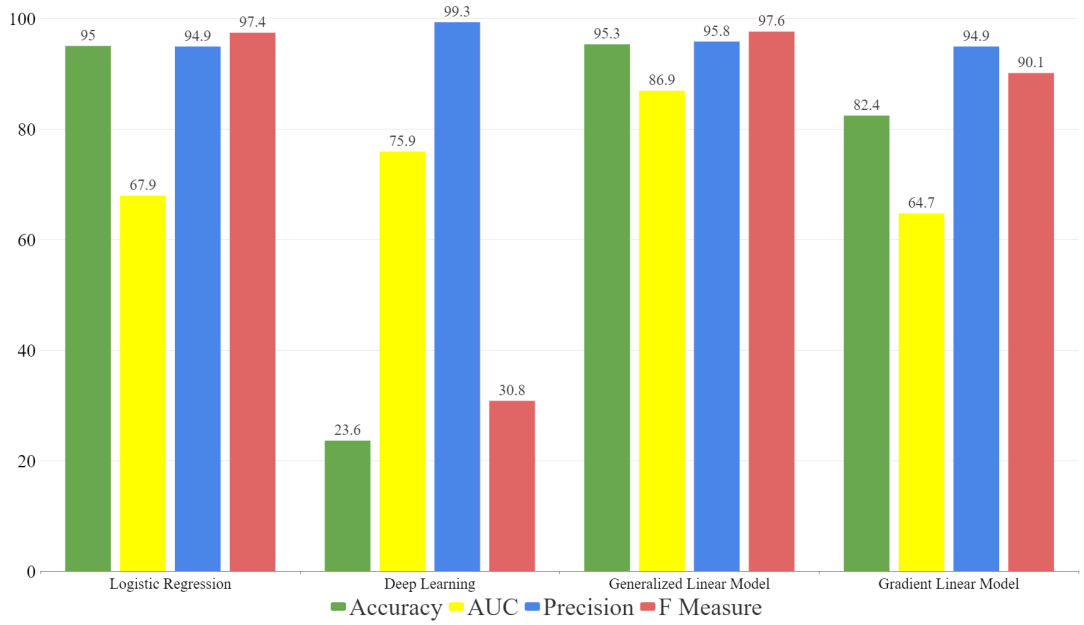


Figura 35 – Indicadores obtidos com a criação do modelo AM por algoritmo do *dataset* CSE-CIC-IDS2018 processado em percentuais.

Os indicadores obtidos pelos processos de construção de modelos de AM estão compilados na Tabela 8.

Tabela 8 – Tabela comparativa dos resultados obtidos com a criação do modelo para o *dataset* CSE-CIC-IDS2018 original e processado em percentuais.

	Raw				Processed			
Model	Accuracy	AUC	Precision	F Measure	Accuracy	AUC	Precision	F Mesasure
Logistic Regression	6.7	50	100	0	95.0	67.9	94.9	97.4
Gradient Boosted Trees	**93.3**	**99.8**	**93.3**	**96.6**	82.4	64.7	94.9	90.1
Deep Learning	73.2	100	100	83.3	23.6	75.9	99.3	30.8
Generalized Linear Model	78.6	99.8	100	87.1	**95.3**	**86.9**	**95.8**	**97.6**

6.3 Consultas aos Dados de Proveniência

Após o processamento dos arquivos correspondentes aos dois *datasets* utilizados como casos de uso, as tabelas especificadas para armazenar os dados de proveniência já possuem dados passíveis de serem consultados. Através destas consultas podem ser atendidas as questões de competência apresentadas na Seção 4.3.

Na primeira questão são apresentados os atributos e suas características tais como o nome do *dataset* processado, a sua versão, o nome e o tipo das colunas e atributos e a categoria do atributo. Um exemplo do consulta e linhas retornadas pode ser visualizado na Figura 36.

- Questão 1 - Características sobre o *dataset* trabalhado: Qual o nome e características dos atributos anonimizados do *dataset* processado?

Consulta1.txt

```
SELECT LABEL AS DATASET_NAME, VERSION AS VERSION_OF_DATASET,  INITIAL_VALIDITY AS BEGINNIG_OF_VALIDITY,
C.COLUMN_LABEL AS LABEL_OF_COLUMN, S.STRUCTUREDTYPE_LABEL AS LABEL_OF_COLUMN_TYPE,
LA.LOGICALATTRIBUTE_LABEL AS LABEL_OF_ATTRIBUTE,
LAC.LOGICALATTRIBUTECATEGORY_LABEL AS LABEL_OF_ATTRIBUTE_CATEGORY

FROM LABELEDDATASET D, COLUMNTABLE C, SQLSTRUCTUREDTYPE S, LOGICALATTRIBUTE LA, LOGICALATTRIBUTECATEGORY LAC

WHERE D.DATASET_ID = C.DATASET_ID
AND   C.COLUMNTABLE_ID = LA.COLUMNTABLE_ID
AND   LA.LOGICALATTRIBUTECATEGORY_ID = LAC.LOGICALATTRIBUTECATEGORY_ID
AND   C.STRUCTUREDTYPE_ID = S.STRUCTUREDTYPE_ID
```

	dataset_name character varying (	version_of_dataset integer	beginnig_of_validity timestamp without time zone	label_of_column character varying (	label_of_column_type character varying (500)	label_of_attribute character varying (500)	label_of_attribute_category character varying (500)
1	UNSWNB1512	1	2022-05-24 18:40:53.691476	sttl	String	sttl	QUASI-IDENTIFIER
2	UNSWNB1513	1	2022-05-27 07:27:10.689285	sttl	String	sttl	QUASI-IDENTIFIER
3	UNSWNB1512	1	2022-05-24 18:40:53.691476	srcip	String	srcip	SENSITIVE
4	UNSWNB1513	1	2022-05-27 07:27:10.689285	srcip	String	srcip	SENSITIVE
5	UNSWNB1512	1	2022-05-24 18:48:18.867387	sttl	String	sttl	QUASI-IDENTIFIER
6	UNSWNB1512	1	2022-05-24 18:48:18.867387	srcip	String	srcip	SENSITIVE
7	UNSWNB1513	1	2022-05-25 13:48:58.529301	srcip	String	srcip	SENSITIVE
8	UNSWNB1523	1	2022-05-31 18:28:00.253069	srcip	String	srcip	SENSITIVE
9	UNSWNB1512	1	2022-05-25 06:11:33.853375	srcip	String	srcip	SENSITIVE
10	UNSWNB1512	1	2022-05-25 09:14:57.714517	srcip	String	srcip	SENSITIVE

Figura 36 – Questão 1 aos dados capturados de proveniência e extrato do resultado retornado.

Assim, os dados visualizados na figura citada representam, por exemplo, que o *dataset* UNSWNB1512, aqui representando a nomenclatura dada à parte 1.2 do arquivo 1 do *dataset* UNSW-NB15, versão 1, atributos [sttl] e [srcip], armazenados em formato *String* e classificados como *Quasi-Identifier* e *Sensitive* respectivamente.

Na questão de número 2 são apresentadas as justificativas e os dados relacionados a elas tais como o identificador e nome do *dataset* processado, o nome e identificador dos

atributos relacionados, o identificador e nome da organização consultada e a justificativa. Um exemplo da consulta e os dados retornados pode ser visualizada na Figura 37.

- Questão 2 - Composição e execução do *workflow* e seus respectivos *steps*: Qual a justificativa para o processamento dos dados contidos em um dataset relativo à determinada organização?

Consulta2.txt

```
SELECT T.ID_OF_DATASET, T.DATASET_NAME, T.ATTRIBUTE, T.ID_OF_ROW, T.ID_OF_ORGANIZATION, T.NAME_OF_ORGANIZATION,
       CA.CONSENTAGREEMENT_TEXT AS AGREEMENT_TEXT, T.KIND_OF_JUSTIFICATION
FROM

(
SELECT LD.DATASET_ID AS ID_OF_DATASET, LD.LABEL AS DATASET_NAME, LA.LOGICALATTRIBUTE_LABEL AS ATTRIBUTE,
       R.ROWTABLE_ID AS ID_OF_ROW, O.ORGANIZATIONNAME AS NAME_OF_ORGANIZATION,
       DS.DATASUBJECT_ID AS ID_OF_ORGANIZATION, J.JUSTIFICATION_TYPE AS KIND_OF_JUSTIFICATION


FROM LABELEDDATASET LD, LOGICALATTRIBUTE LA, COLUMNTABLE CT, ROWTABLE R, PERSONALDATA PD,DATASUBJECT DS,
     ORGANIZATION O, JUSTIFICATION J



WHERE   LD.DATASET_ID        = CT.DATASET_ID
AND     LD.DATASET_ID        = R.DATASET_ID::UUID
AND     R.ROWTABLE_ID        = PD.ROWTABLE_ID
AND     PD.DATASUBJECT_ID    = DS.DATASUBJECT_ID::VARCHAR
AND     DS.ORGANIZATION_ID   = O.ORGANIZATION_ID::UUID
AND     CT.COLUMNTABLE_ID    = LA.COLUMNTABLE_ID
AND     LA.JUSTIFICATION_ID  = J.JUSTIFICATION_ID

) T

LEFT JOIN CONSENTAGREEMENT CA
ON CA.DATASUBJECT_ID = T.ID_OF_ORGANIZATION
```

	id_of_dataset uuid	dataset_name character varying	attribute character varying	id_of_row character varying (1000)	id_of_organization uuid	name_of_organization character varying (1000)	agreement_text text	kind_of_justification character varying (100)
1	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	attack_cat	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
2	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	ct_src_dport_ltm	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
3	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	ct_dst_ltm	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
4	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	ct_ftp_cmd	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
5	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	ct_dst_src_ltm	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
6	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	ct_srv_dst	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
7	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	is_ftp_login	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
8	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	ct_state_ttl	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
9	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	ackdat	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest
10	ac25d074-1386-4bb4-a566-f3f4cc07bd45	UNSWNB1511	Label	6eaaf494-dad8-4212-bd2e-9aabdcaa682c	f24e8339-0a79-4289-8a1f-cbcc7f64e7f7	Viva	[null]	Public Interest

Figura 37 – Questão 2 aos dados capturados de proveniência e os dados retornados.

Importante ressaltar que não há termo de consentimento nos registros retornados. Isto se deve ao fato de que, para este tipo de justificação informado para este processamento, não há a necessidade de termo de consentimento de acordo com a GDPR. Assim, os dados visualizados na figura citada representam, por exemplo, que o *dataset* UNSWNB1511, aqui representando a nomenclatura dada à parte 1.1 do arquivo 1 do *dataset* UNSW-NB15 com o seu identificador (*id_of_dataset*) e seu nome (*dataset_name*), os atributos como [attack_cat] e [ct_src_dport_ltm] e o identificador da linha da qual esses atributos fazem parte (*id_of_row*), são relacionados à organização (*name_of_organization*) 'Viva', são processados com a justificativa (*kind_of_justification*) de *'Public Interest'* mas não possuem declarações de consentimento (*agrrement_text*) relacionados a estes atributos.

Na questão 3 são consultados os valores dos parâmetros que foram utilizados para o processamento do *dataset*, especificamente nas tarefas de anonimização e pré-processamento, tais como nome e identificador do *workflow* e do *step*, nome e valor do

parâmetro e o algoritmo e *software* associados. Através desta estrutura de captura de dados de proveniência é possível, por exemplo, realizar um processo de reidentificação de entes envolvidos. Um exemplo da consulta e os dados retornados pode ser visualizada na Figura 38.

- Questão 3 - Os operadores e os parâmetros utilizados com os seus respectivos algoritmos: Quais os valores dos parâmetros aplicados em determinada execução de determinado *dataset*?

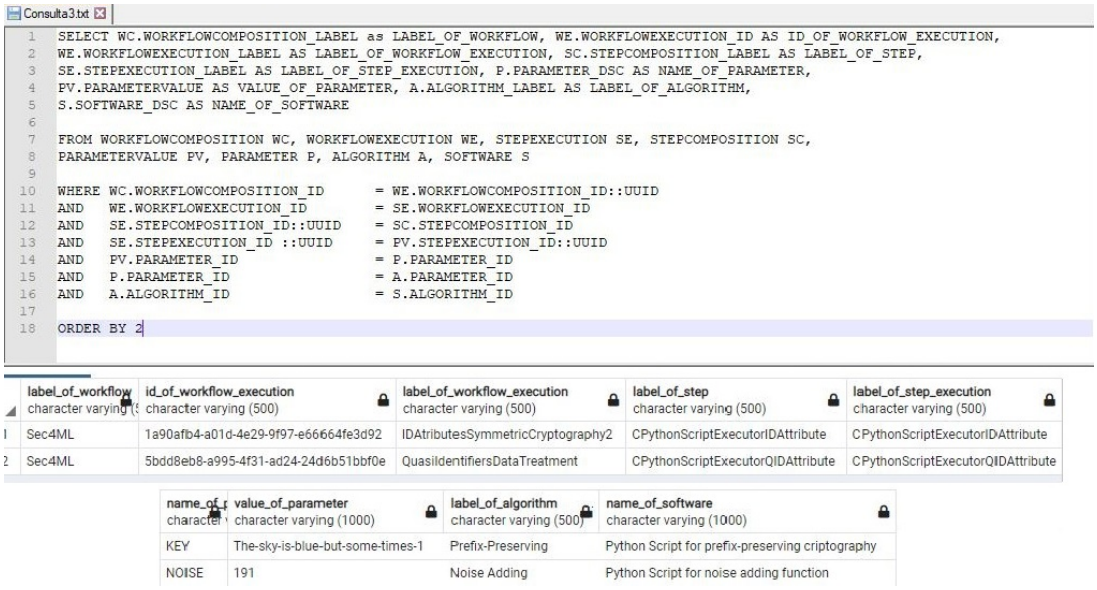


Figura 38 – Questão 3 aos dados capturados de proveniência e os dados retornados.

Assim, os dados visualizados na figura citada representam, por exemplo, que o *workflow* denominado 'Sec4ML' possui uma instância de execução (*label_of_workflow_execution*) com *label* 'IDAtributesSymmetricCryptography2'. Esta instância está relacionada a um instância de execução de *step* denominada 'CPythonScriptExecutorIDAttribute' que utilizou nesta determinada execução o valor 'The-sky-is-blue-but-some-times-1' como chave (*key*) para o algoritmo implementado (*label_of_algorithm*) como 'PrefixPreserving' pelo *software* (*name_of_software*) 'Python Script for prefix-preserving criptography'.

Na quarta e última questão são verificados quais entes envolvidos possuem requisições de exclusão de dados e, se possuem, relativos a quais dados. São consultados dados como a organização envolvida, o nome do *dataset*, o nome da coluna e o tipo e data da requisição. Esta consulta e os dados retornados pode ser visualizada na Figura 39. Importante ressaltar que, como pode ser visto na última figura citada, não havia no momento da consulta, requisições para os atributos consultados. Entretanto, esta estrutura possibilita a exclusão de dados de determinado ente envolvido no futuro (direito de ser esquecido).

- Questão 4- Dados relativos à conformidade com as legislações de proteção de dados: Quais entes envolvidos (*DataSubject*) possuem requisições de exclusão de seus dados? Se existem, são relativos a quais dados?

Consulta4.txt

```
SELECT T.ORGANIZATION, T.LABEL_OF_DATASET, T.PERSONALDATA_IDENTIFIER, T.COLUMNTABLE_IDENTIFIER, T.LABEL_OF_COLUMN,
       T.ROWTABLE_IDENTIFIER, T.VALUE_OF_CELL, R.REQUEST_TYPE AS TYPE_OF_REQUEST, R.REQUEST_DATE AS DATE_OF_REQUEST
FROM

(
SELECT O.ORGANIZATIONNAME AS ORGANIZATION, LD.LABEL AS LABEL_OF_DATASET, DS.DATASUBJECT_ID,
       PD.PERSONALDATA_ID AS PERSONALDATA_IDENTIFIER, CT.COLUMNTABLE_ID AS COLUMNTABLE_IDENTIFIER,
       CT.COLUMN_LABEL AS LABEL_OF_COLUMN, RT.ROWTABLE_ID AS ROWTABLE_IDENTIFIER, RC.ROWCOLUMN_VALUE AS VALUE_OF_CELL

FROM ORGANIZATION O,  DATASUBJECT DS, PERSONALDATA PD,
ROWTABLE RT, ROWCOLUMNVALUE RC, COLUMNTABLE CT, LABELEDDATASET LD

WHERE O.ORGANIZATION_ID::UUID   = DS.ORGANIZATION_ID
AND   DS.DATASUBJECT_ID         = PD.DATASUBJECT_ID::UUID
AND   PD.ROWTABLE_ID            = RT.ROWTABLE_ID
AND   RT.ROWTABLE_ID            = RC.ROWTABLE_ID
AND   RC.COLUMNTABLE_ID::UUID   = CT.COLUMNTABLE_ID
AND   CT.DATASET_ID             = LD.DATASET_ID

) T

LEFT JOIN REQUEST R
ON   R.DATASUBJECT_ID           = T.DATASUBJECT_ID
```

	organization character vary	label_of_dataset character varying (5	personaldata_identifier character varying (500)	columntable_identifier uuid	label_of_column character varying (50	rowtable_identifier character varying (1000)	value_of_cell character varying (	type_of_request character varying	date_of_request date
1	Viva	UNSWNB1527	24920776-a13b-45a2-9137-d67b632d52f6	2e60e952-99ae-4ab9-bd7e-db50e3357f39	srcip	0fa32c0a-446d-48fe-a270-8635eb59e07b	59.166.0.1	[null]	[null]
2	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	db59b5df-6abe-4fa0-ae34-98732af8c58b	srcip	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	59.166.0.3	[null]	[null]
3	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	35bc1e29-26dd-4a61-b309-2813f405a7b4	sport	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	38429	[null]	[null]
4	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	17c62039-2463-4e55-82a1-da9c9f198a84	dstip	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	149.171.126.3	[null]	[null]
5	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	56386cbb-1cee-4b6b-a352-bb909d4dedee	dsport	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	111	[null]	[null]
6	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	ef3ca8ce-34ed-4ae7-9759-c5eaba12ee9a	proto	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	udp	[null]	[null]
7	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	7bd9bcf3-5cb1-473d-a7c4-3e9b1d56bca5	state	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	CON	[null]	[null]
8	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	86601a9c-928d-4ce1-837d-eb412846362b	dur	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	0.004517	[null]	[null]
9	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	8556ae51-a4b3-4e37-a310-65df8739092a	sbytes	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	568	[null]	[null]
10	Viva	UNSWNB1516	08b18acb-2fcd-4e57-a1ce-3b960347750a	cd4c2b52-6c59-47a8-81e0-7fee4b5c2e77	dbytes	5081dd0d-d854-44d5-b8eb-b6c9407c79b9	304	[null]	[null]

Figura 39 – Questão 4 aos dados capturados de proveniência e os dados retornados.

Na consulta apresentada na figura 39 há, por exemplo a organização 'Viva', ente envolvido com os *datasets* processados. Para cada *dataset* consultado há algumas colunas descritas (*label_of_column*) e os valores a elas associados (*value_of_cell*). Entretanto, neste exemplo, para as colunas consultadas dos *datasets* processados, não há nenhum tipo de requisição (*type_of_request*) associado.

6.4 Sec4ML *FAIR Data Point*

A estrutura possível para publicação de *datasets* no *Sec4ML FAIR Data Point* contempla a criação de catálogos de *datasets*. Estes catálogos podem possuir um ou mais *datasets*. Uma visualização das interfaces de consulta dos catálogos e seus *datasets* é apresentada pela Figura 40.

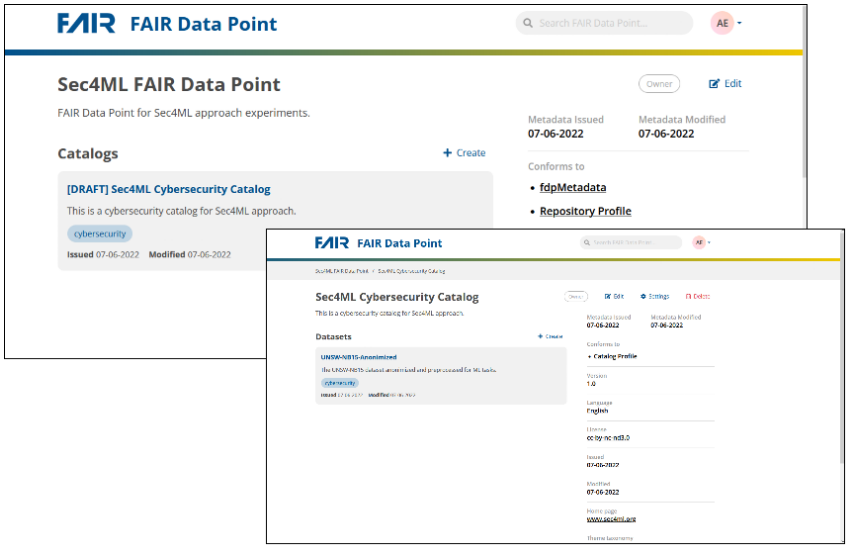


Figura 40 – Interfaces do FAIR Data Point (catálogo).

Para cada *dataset* catalogado, é possível criar uma ou mais distribuições. Estas distribuições, por sua vez, podem possuir uma ou mais versões a serem publicadas. Uma visualização das interfaces de consulta das distribuições e suas versões é apresentada pela Figura 41.

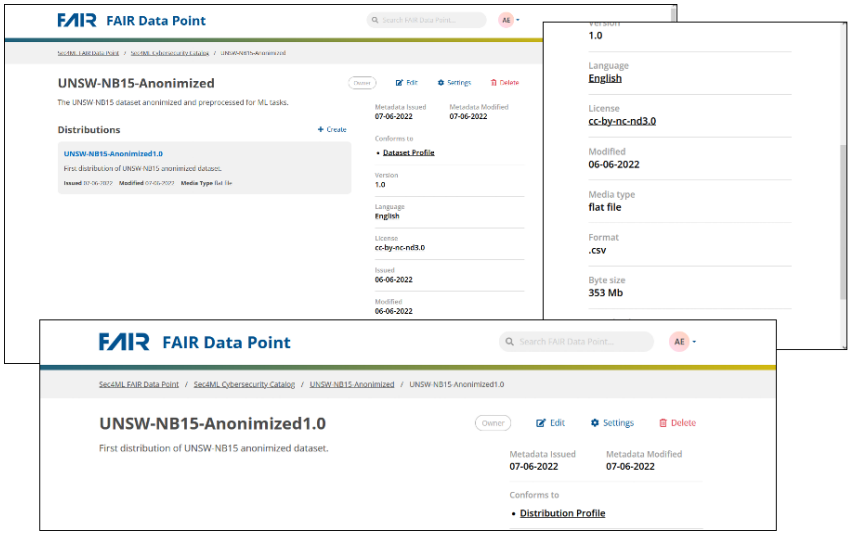


Figura 41 – Interfaces do FAIR Data Point (*dataset* e distribuição).

6.5 Considerações Finais

Um dos principais problemas enfrentados foi quanto à configuração de uso de memória RAM pela máquina Java do PDI. A implementação foi desenvolvida com uso de um pequeno grupo de 12 registros oriundos do *dataset* UNSW-NB15. Entretanto quando

foi iniciada a execução do ETL sobre um dos arquivos .csv componentes ocorria erros de estouro de memória alocada pelo Java.

Foram realizados diversas mudanças de parâmetros de chamada do PDI que usa parâmetros de configuração do Java sem sucesso. Esta situação permaneceu até que através de uma orientação recebida, foi verificado que havia uma variável de memória (PENTAHO_DI_JAVA_OPTIONS) que estava restringindo a alocação de memória pelo Java e este valor (1Gb) se sobrepunha aos valores indicados no arquivo .bat que realiza a chamada da execução do PDI.

A memória RAM do *notebook* foi estendida para 20Gb e foi configurado o uso de 8Gb de memória a ser utilizado pelo Java. Ainda assim, com essa quantidade de memória dedicada foi possível executar somente *subsets* de 100.000 registros. Desta forma tornou-se necessário dividir cada arquivo .csv (4 arquivos contendo cerca de 700.000 registros cada) em 7 *subsets* de 100.000 para que fosse possível a execução do ETL. Posteriormente os dados gerados separadamente foram reagrupados em um único arquivo.

Outro problema enfrentado foi a necessidade de tratar alguns valores de lixo encontrados durante as primeiras execuções do ETL. Foi necessário para cada ocorrência deste tipo de valores acrescentar um step de substituição do valor encontrado.

Foi possível realizar experimentos a fim de observar o quanto o processamento realizado pela implementação da abordagem Sec4ML pode impactar em resultados de criação de modelos em AM para tarefas de classificação. Entretanto, é necessário o refino e aprofundamento dos experimentos, tornando-os mais controlados e parametrizados, para que se possa realizar uma análise mais fundamentada sobre os possíveis impactos do processamento nos *datasets* trabalhados, realizando: (i) acréscimo da abrangência na apresentação dos gráficos, (ii) formalização das métricas, (iii) análise da viabilidade de reduzir o experimento, dentre outras ações.

As transformações sofridas pelos atributos dos dois *datasets* e seus valores de exemplo estão relacionadas na Tabela 9.

Tabela 9 – Tabela dos atributos processados e técnicas empregadas.

	Atributo	Anonimização	Pré-processamento	Descrição	Dado Não Processado	Dado Processado pela Sec4ML	Resultado Sec4ML
UNSW-NB15	dstip	Crypto-Pan (prefixpreserving)		Endereço IP de destino	149.171.126.6	213.151.98.62	Endereço IP cifrado
	sttl	Adição de ruído		*Time to Live* da origem para o destino	31	229	*Time to Live* + ruído
	srcip	Supressão		Endereço IP de origem	59.166.0.0	-	Suprimido
	service	Supressão		Serviço envolvido no incidente registrado	dns	-	Suprimido
	ct_dst_sport_ltm		Normalização	Número de linhas com o mesmo dstip e sport em 100 linhas	1	0.230769230769230 78	Número normalizado
	attack_cat		Codificação	Categoria do ataque	Reconnaissance	1	Categoria codificada
CSE-CIC-IDS2018	dst_ip	Crypto-Pan (prefixpreserving)		Endereço IP de destino	172.31.66.49	227.223.77.186	Endereço IP cifrado
	fwd_pkt_len_mean	Supressão		Média do tamanho do pacote	75.83333333333333	-	Suprimido
	src_ip	Supressão		Endereço IP de origem	195.175.55.106	-	Suprimido
	flow_duration	Adição de ruído	Normalização	Duração do fluxo	812620	2.8333517028968742e-06	Duração do fluxo + ruído + normalização
	label		Codificação	*Label* de classificação	Benign	0	*Label* codificado

7 CONCLUSÃO

Nos últimos anos, foi possível observar um crescimento significativo na geração de dados através de novas tecnologias de TI. Junto com esta onda surgiu também uma nova realidade, com diferentes ameaças e vulnerabilidades que também podem dar origem a um vazamento de dados ou invasão de privacidade. Entretanto, os recursos e ferramentas destinados a combater as ameaças não se tornaram melhores na mesma velocidade dos desafios dentro do domínio da segurança da informação.

Ao mesmo tempo, modelos e conceitos de AM surgiram facilitando e acelerando a aquisição e o aprendizado do conhecimento. Alguns conjuntos de dados de eventos de segurança da informação rotulados surgiram para tarefas de AM. Neste contexto, ainda há muitos desafios a enfrentar, tais como questões culturais e desenvolvimento de ferramentas para gerar conjuntos de dados de segurança cibernética anônimos. Devido a estas questões, há uma disponibilidade limitada desses conjuntos de dados e, consequentemente, a necessidade de aumentar o número de conjuntos de dados de segurança da informação disponíveis para pesquisa.

A necessidade de construir processos e bases de dados em conformidade com legislações de proteção de dados ao redor do mundo também veio à tona no momento da entrada em vigor de tais legislações, demandando a adaptação dos processos e bases de dados existentes. Desta forma, novos processos e bases de dados necessitam ser, a partir de agora, projetados sob a ótica da aplicação de técnicas de preservação da privacidade.

Diante deste cenário, esta dissertação propôs a abordagem Sec4ML que visa atender a esta demanda com as diversas vertentes apontadas, tornando possível preparar, anonimizar e publicar dados e metadados de incidentes de cibersegurança para tarefas de KDD e, ao mesmo tempo, gerar conjuntos de dados reutilizáveis prontos na Web de Dados, em conformidade com as legislações de proteção de dados.

A aplicação da abordagem Sec4ml mostrou que é (i) uma forma de apoiar a publicação de dados anonimizados para que possam ser usados para a criação de modelos de AM para tarefas de classificação e que (ii) o resultado da solução proposta pode ser usado para avaliar o desempenho da criação de modelos de AM quando os dados são anonimizados. Além disso, a ontologia criada permitiu (iii) a captura de proveniência que atende os princípios F1, F2, F3, F4, A2, I1, I2, I3 e R1.2 do conjunto de princípios FAIR.

Outra importante contribuição foi a implementação desta abordagem através do desenvolvimento de um *workflow* em PDI, que demonstrou ser viável agregar, além de tratar e processar os dados, a captura de dados de proveniência relacionados não somente com o *dataset*, mas também com os *steps* executados e parâmetros utilizados. Este *workflow*

foi desenvolvido de maneira flexível, através do uso de parametrização para a informação de atributos a serem processados, assim como a possibilidade de mudança e/ou acréscimo de rotinas dentro deste *workflow*. Com essa característica, é possível realizar mudanças nas rotinas e algoritmos aplicados sem a necessidade de mudanças significativas no restante da definição do *workflow*.

Para apoiar essa implementação foi necessária a proposta também da Ontologia leve denominada Sec4ML-O e de seu respectivo modelo de dados relacional. Essa estrutura possibilita: (i) o armazenamento dos dados em ambiente relacional caracterizado como uma área de dados intermediária possibilitando, desta forma, a realização de consultas que respondam às questões de competência; (ii) a triplificação dos dados de proveniência, ação necessária para que esses dados possam ser lidos por humanos e máquinas, uma vez disponibilizados na WEB de Dados e (iii) a captura de diversos metadados relativos à conceitos originários da GDPR, possibilitando a conformidade legal.

Pode-se destacar ainda como contribuição a instalação de um ambiente para publicação de dados de proveniência sobre dados de segurança de informação, que ficam disponíveis para consulta e pesquisa com a finalidade de reúso e aplicação de procedimentos de AM. Como ambiente para a publicação dos dados de proveniência foi escolhido o *FAIR Data Point*, ferramenta disponibilizada para a comunidade científica. Através desta publicação, os dados de proveniência ficam disponíveis para consulta e pesquisa com a finalidade de reuso. Esta solução de *software* permite a criação de catálogos de *datasets*. Estes *datasets*, por sua vez podem ser publicados com quantas distribuições forem necessárias permitindo o versionamento destes dados dentro do catálogo.

Além disso, foi possível responder às questões de pesquisa apresentadas. Foi demonstrado que é possível apoiar a publicação de dados anonimizados para que possam ser usados para a criação de modelos de AM para tarefas de classificação através da aplicação da abordagem Sec4ML. O resultado da solução proposta pode ser usado sim para avaliar se há perdas significativas no desempenho da criação de modelos de AM quando os dados são anonimizados, na medida em que sejam realizados novos experimentos mais controlados e parametrizados. Foi demonstrado também, que a captura de proveniência, além de atender aos princípios FAIR, pode responder às questões de conformidade legal, assim como possibilitar a reidentificação de entes envolvidos.

É importante ressaltar que a abordagem e sua implementação podem ser ajustadas ou adaptadas para outros domínios, além do domínio de segurança da informação, como por exemplo para dados clínicos (70).

Dentre as dificuldades enfrentadas, as mais desafiadoras foram a implementação do *FAIR Data Point* e a construção de uma ontologia integrando diversos conceitos necessários para a captura de dados de proveniência suficientes para que todas as questões de competência pudessem ser respondidas.

Na questão da implementação do *FAIR Data Point* podemos destacar o grande número de tecnologias que precisam ser instaladas e configuradas para que o repositório seja disponibilizado. Muitas dessas tecnologias estão em processo de amadurecimento e de configuração detalhada e complexa, o que dificultou ainda mais a conclusão dessa tarefa.

O processo de construção da ontologia demandou a análise de diversas outras ontologias passíveis de serem reusadas. Entretanto, o possível reuso ficaria condicionado à aplicabilidade da mesma semântica definida originalmente, o que muitas vezes não é verdadeiro. Assim, somente evidenciando-se os conceitos já definidos mas que não seriam aplicáveis ao desenvolvimento da abordagem, foi possível identificar quais conceitos eram necessários mas ainda não propostos por ontologias existentes.

Durante a aplicação da abordagem, na realização dos experimentos foram enfrentadas algum problemas de desempenho da infraestrutura computacional. Problemas como dificuldade de encontrar a configuração correta de parâmetros de uso de memória pela ferramenta e uso excessivo de carga de processador quando foram processados *subsets* grandes também impactaram o conclusão dos experimentos.

Como trabalhos futuros podemos apontar algumas demandas identificadas mas que não foram incluídas no escopo dessa dissertação. Estas possibilidades de trabalhos futuros são listadas a seguir:

1. A adaptação ou evolução do ETL a fim de prover a cobertura de mais tarefas ML, além da classificação/predição dos conjuntos de dados;
2. A implementação de mais opções de operadores de anonimização e pré-processamento;
3. A implementação de operadores de anonimização e pré-processamento em valores de atributos específicos (um único valor de um atributo) das estruturas tabulares e não somente em colunas inteiras (todos os valores de um atributo);
4. A implementação da anonimização utilizando-se chaves distintas para cada valor do mesmo atributo processado;
5. A possibilidade de diferenciação da execução do ETL, no que tange à anonimização, para mais de um domínio específico, ou seja, tornando possível a utilização de valores diferentes de parâmetros para diferentes domínios, de forma a obter resultados específicos;
6. Adequação desta abordagem em relação ao paradigma *"privacy by design"* (PbD) (71) e seus princípios;
7. A verificação da aderência da Sec4ML-O à LGPD, identificando possíveis lacunas e/ou conceitos ausentes;

8. Generalização da implementação da abordagem Sec4ML para tornar-se independente de domínio, podendo ser melhor utilizada em outras áreas de conhecimento, como por exemplo dados de saúde;

9. O aperfeiçoamento da ontologia Sec4ML-O a fim de torná-la bem fundamentada pela Unified Foundational Ontology (UFO); e

10. A análise da ontologia Sec4ML a fim de identificar sua relação com outras ontologias existentes e possíveis reusos.

REFERÊNCIAS

1 DANDURAND, L.; SERRANO, O. S. Towards improved cyber security information sharing. In: *n/a*. [S.l.: s.n.], 2013. p. 16.

2 GANDON, F. A survey of the first 20 years of research on semantic web and linked data. *Ingénierie des systèmes d'information*, v. 23, n. 3–4, p. 11–38, Aug 2018. ISSN 16331311.

3 WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J.-W.; SANTOS, L. B. da S.; BOURNE, P. E.; AL. et. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, v. 3, n. 1, p. 160018, Dec 2016. ISSN 2052-4463.

4 SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. Adoption of the linked data best practices in different topical domains. In: _____. *The Semantic Web – ISWC 2014*. Springer International Publishing, 2014. (Lecture Notes in Computer Science, v. 8796), p. 245–260. ISBN 978-3-319-11963-2. Disponível em: <http://link.springer.com/10.1007/978-3-319-11964-9_16>.

5 KARATAS, G.; DEMIR, O.; SAHINGOZ, O. K. Increasing the performance of machine learning-based idss on an imbalanced and up-to-date dataset. *IEEE Access*, v. 8, p. 32150–32162, 2020. ISSN 2169-3536.

6 GONI, I.; GUMPY, J. M.; MAIGARI, T. U. Cybersecurity and cyber forensics: Machine learning approach. *Semiconductor Science and Information Devices*, v. 2, n. 2, Dec 2020. ISSN 2661-3212. Disponível em: <https://ojs.bilpublishing.com/index.php/ssid/article/view/2495>.

7 SHAUKAT, K.; LUO, S.; VARADHARAJAN, V.; HAMEED, I. A.; XU, M. A survey on machine learning techniques for cyber security in the last decade. *IEEE Access*, v. 8, p. 222310–222354, 2020. ISSN 2169-3536.

8 DASGUPTA, D.; AKHTAR, Z.; SEN, S. Machine learning in cybersecurity: a comprehensive survey. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, p. 154851292095127, Sep 2020. ISSN 1548-5129, 1557-380X.

9 MISHRA, P.; VARADHARAJAN, V.; PILLI, E. S. A detailed investigation and analysis of using machine learning techniques for intrusion detection. v. 21, n. 1, p. 43, 2019.

10 NEWS, S. in. *Data breach statistics by country in 2021*. 2022. (Acessado em 25/05/2022). Disponível em: <https://surfshark.com/blog/data-breach-statistics-by-country-in-2021>.

11 NETWORK, E.; AGENCY., I. S.; BEDRIJFSREVISOREN., D. *Cyber security information sharing: an overview of regulatory and non regulatory approaches.* Publications Office, 2015. Disponível em: <https://data.europa.eu/doi/10.2824/43639>.

12 FIGUEIREDO, G. B. de; MOREIRA, J. L. R.; CORDEIRO, K. de F.; CAMPOS, M. L. M. Aligning dmbok and open government with the fair data principles. In: GUIZZARDI, G.; GAILLY, F.; MACIEL, R. S. P. (Ed.). *Advances in Conceptual Modeling*. [S.l.]: Springer International Publishing, 2019. p. 13–22. ISBN 978-3-030-34146-6.

13 OLIVEIRA, F. T. de; CAVALCANTI, M. C.; SALLES, R. M. Towards effective reproducible botnet detection methods through scientific workflow management systems. *Anais do XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, v. 1, p. 14, 2017.

14 TORRES, J. M.; COMESAñA, C. I.; GARCíA-NIETO, P. J. Review: machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, v. 10, n. 10, p. 2823–2836, Oct 2019. ISSN 1868-8071, 1868-808X.

15 MOURA, L.; SILVA, M. da; CORDEIRO, K.; CAVALCANTI, M. A well-founded ontology to support the preparation of training and test datasets:. In: *Proceedings of the 23rd International Conference on Enterprise Information Systems*. SCITEPRESS - Science and Technology Publications, 2021. p. 99–110. ISBN 978-989-758-509-8. Disponível em: <https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010460000990110>.

16 MOUSTAFA, N.; SLAY, J. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: *2015 Military Communications and Information Systems Conference (MilCIS)*. IEEE, 2015. p. 1–6. ISBN 978-1-4673-7007-3. Disponível em: <http://ieeexplore.ieee.org/document/7348942/>.

17 CYBERSECURITY, C. I. for. *CSE-CIC-IDS2018 on AWS*. 2018. (Acessado em25/05/2022). Disponível em: <https://www.unb.ca/cic/datasets/ids-2018.html>.

18 GIUNCHIGLIA, F.; ZAIHRAYEU, I. *Lightweight Ontologies*. Trento, 2007. 10 p.

19 TORRES, J. M.; COMESAñA, C. I.; GARCíA-NIETO, P. J. Review: machine learning techniques applied to cybersecurity. *International Journal of Machine Learning and Cybernetics*, v. 10, n. 10, p. 2823–2836, Oct 2019. ISSN 1868-8071, 1868-808X.

20 GRAHAM, J. *Cyber Security Essentials*. [S.l.: s.n.], 2011. 331 p.

21 BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. In: *Scientific American*. [S.l.]: Scientific American, 2001. p. 1–3.

22 BERNERS-LEE, T. *As 5 estrelas dos Dados Abertos*. 2015. (Acessado em 07/11/2019). Disponível em: <https://5stardata.info/pt-BR/>.

23 ISOTANI, S.; BITTENCOURT, I. I. *Dados abertos conectados*. [S.l.]: Novatec, 2015. ISBN 978-85-7522-449-6.

24 KIMBALL, R.; ROSS, M. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd. ed. [S.l.]: Wiley Publishing, 2013. ISBN 1118530802.

25 RAHM, E.; DO, H. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, v. 23, p. 3–13, 01 2000.

26 SEGUNDO, J. E. S. Web semântica, dados ligados e dados abertos: uma visão dos desafios do brasil frente às iniciativas internacionais. In: *XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação*. [S.l.: s.n.], 2015. p. n/a.

27 WEB, W. S. *Resource Description Framework*. 2014. (Acessado em 11/11/2020). Disponível em: <https://www.w3.org/RDF/>.

28 CLOUD, T. L. O. D. *The Linked Open Data Cloud*. 2019. (Acessado em 05/11/2019). Disponível em: <https://lod-cloud.net/>.

29 GUARINO, N.; OBERLE, D.; STAAB, S. What is an ontology? In: _____. *Handbook on Ontologies*. Springer Berlin Heidelberg, 2009. p. 1–17. ISBN 978-3-540-70999-2. Disponível em: <http://link.springer.com/10.1007/978-3-540-92673-3_0>.

30 STUDER, R.; BENJAMINS, V.; FENSEL, D. Knowledge engineering: Principles and methods. *Data Knowledge Engineering*, v. 25, n. 1–2, p. 161–197, Mar 1998. ISSN 0169023X.

31 GUARINO, N. Formal ontology and information systems. In: *n/a*. [S.l.: s.n.], 1998. p. 13.

32 CHIANG, T. J.; KOUH, J. S.; CHANG, R.-I. Ontology-based risk control for the incident management. In: *n/a*. [S.l.: s.n.], 2009. v. 9, n. 11, p. 10.

33 MOREIRA, J. L. R.; BONINO, L.; PIRES, L. F.; SINDEREN, M. V.; HENNING, P. Towards findable, accessible, interoperable and reusable (fair) data repositories: Improving a data repository to behave as a fair data point - repositórios para dados localizáveis, acessíveis, interoperáveis e reutilizáveis (fair): adaptando um repositório de dados para se comportar como um fair data point. *Liinc em Revista*, v. 15, n. 2, Dec 2019. ISSN 1808-3536. *. Disponível em: <http://revista.ibict.br/liinc/article/view/4817>.

34 BRITO, F. T.; MACHADO, J. C. Preservação de privacidade de dados: Fundamentos, técnicas e aplicações. In: _____. *Jornadas de Atualização em Informática 2017*. Brazil: Sociedade Brasileira de Computação, 2014. p. 70–87.

35 FERREIRA, A. Gdpr: What's in a year (and a half)? In: *Proc. 22nd Int. Conf. on Enterprise Inf. Syst.* [S.l.: s.n.], 2020.

36 STOYANOVICH, J.; ABITEBOUL, S.; HOWE, B.; JAGADISH, H. V.; SCHELTER, S. Responsible data management. *Communications of the ACM*, v. 65, n. 6, p. 64–74, Jun 2022. ISSN 0001-0782, 1557-7317.

37 L., M.; P., M.; K., B. Prov-dm: The prov data model. In: *s/n*. [s.n.], 2013. (Acessado em 13/01/2021). Disponível em: <https://www.w3.org/TR/2013/REC-prov-dm-20130430/.>

38 HERSCHEL, M.; DIESTELKäMPER, R.; LAHMAR, H. B. A survey on provenance: What for? what form? what from? *The VLDB Journal*, v. 26, n. 6, p. 881–906, Dec 2017. ISSN 1066-8888, 0949-877X.

39 SOUZA, R.; AZEVEDO, L. G.; LOURENçO, V.; SOARES, E.; THIAGO, R.; BRANDãO, R.; CIVITARESE, D.; BRAZIL, E. V.; MORENO, M.; VALDURIEZ, P.; MATTOSO, M.; CERQUEIRA, R.; NETTO, M. A. S. Workflow provenance in the lifecycle of scientific machine learning. *Concurrency and Computation: Practice and Experience*, Aug 2021. ISSN 1532-0626, 1532-0634. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/cpe.6544>.

40 INTRODUCTION to privacy-preserving data publishing: concepts and techniques. [S.l.]: Chapman Hall/CRC, 2011. (Chapman Hall/CRC data mining and knowledge discovery series). ISBN 978-1-4200-9148-9.

41 SWEENEY, L. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, v. 10, n. 05, p. 557–570, Oct 2002. ISSN 0218-4885, 1793-6411.

42 DWORK, C. Differential privacy. In: BUGLIESI, M.; PRENEEL, B.; SASSONE, V.; WEGENER, I. (Ed.). *Automata, Languages and Programming*. [S.l.]: Springer Berlin Heidelberg, 2006. p. 1–12. ISBN 978-3-540-35908-1.

43 SARKER, I. H.; KAYES, A. S. M.; BADSHA, S.; ALQAHTANI, H.; WATTERS, P.; NG, A. Cybersecurity data science: An overview from machine learning perspective. p. 28, 2020.

44 GELUVARAJ, B.; SATWIK, P. M.; KUMAR, T. A. A. The future of cybersecurity: Major role of artificial intelligence, machine learning, and deep learning in cyberspace. In: _____. *International Conference on Computer Networks and Communication Technologies*. Springer Singapore, 2019. (Lecture Notes on Data Engineering and Communications Technologies, v. 15), p. 739–747. ISBN 978-981-10-8680-9. Disponível em: <http://link.springer.com/10.1007/978-981-10-8681-6_67>.

45 RUSSELL, S.; NORVIG, P. Artificial intelligence: a modern approach. 2002.

46 SARKER, I. H.; KAYES, A. S. M.; WATTERS, P. Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. *Journal of Big Data*, v. 6, n. 1, p. 57, Dec 2019. ISSN 2196-1115.

47 GENERALIZED Linear Models. Wiley Journal of the Royal Statistical Society, v. 135, n. 3, p. 370–384, 1972.

48 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

49 FRIEDMAN, J. H. Greedy function approximation. p. 39, 1999.

50 LAVALLEY, M. P. Logistic regression. *Circulation*, Am Heart Assoc, v. 117, n. 18, p. 2395–2399, 2008.

51 JACOBSEN, A.; KALIYAPERUMAL, R.; SANTOS, L. O. B. da S.; MONS, B.; SCHULTES, E.; ROOS, M.; THOMPSON, M. A generic workflow for the data fairification process. *Data Intelligence*, v. 2, n. 1–2, p. 56–65, Jan 2020. ISSN 2641-435X.

52 MENDONçA, R. R. de. Etl4linkedprov: Managing multigranular linked data provenance. v. 7, n. 2, p. 16.

53 SALVADORI, I.; HUF, A.; SIQUEIRA, F. Data linking as a service: An infrastructure for generating and publishing linked data on the web. In: *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2020. p. 262–271. ISBN 978-1-72817-303-0. Disponível em: <https://ieeexplore.ieee.org/document/9202833/>.

54 RAUTENBERG, S.; ERMILOV, I.; MARX, E.; AUER, S.; NGOMO, A.-C. N. Lodflow: a workflow management system for linked data processing. In: *Proceedings of the 11th International Conference on Semantic Systems*. ACM, 2015. p. 137–144. ISBN 978-1-4503-3462-4. N/a. Disponível em: <https://dl.acm.org/doi/10.1145/2814864.2814882>.

55 FAHAD, A.; TARI, Z.; ALMALAWI, A.; GOSCINSKI, A.; KHALIL, I.; MAHMOOD, A. Ppfscada: Privacy preserving framework for scada data publishing. *Future Generation Computer Systems*, v. 37, p. 496–511, Jul 2014. ISSN 0167739X.

56 RD, P. Ø degå. *Data Anonymization for Research*. Noruega: [s.n.], 2019. 99 p. (1).

57 SILVA, M. L. e.; CORDEIRO, K. de F.; CAVALCANTI, M. C. Sec4ml: An approach to support cybersecurity data publishing for machine learning tasks. In: *2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)*. IEEE, 2021. p. 226–235. ISBN 978-1-66544-488-0. Disponível em: <https://ieeexplore.ieee.org/document/9626339/>.

58 VICKNAIR, C.; MACIAS, M.; ZHAO, Z.; NAN, X.; CHEN, Y.; WILKINS, D. A comparison of a graph database and a relational database: a data provenance perspective. In: *Proceedings of the 48th Annual Southeast Regional Conference on - ACM SE '10*. Oxford, Mississippi: ACM Press, 2010. p. 1. ISBN 978-1-4503-0064-3. Disponível em: <http://portal.acm.org/citation.cfm?doid=1900008.1900067>.

59 CHENG, Y.; DING, P.; WANG, T.; LU, W.; DU, X. Which category is better: Benchmarking relational and graph database management systems. *Data Science and Engineering*, v. 4, n. 4, p. 309–322, Dec 2019. ISSN 2364-1185, 2364-1541.

60 UJCICH, B. E.; BATES, A.; SANDERS, W. H. A provenance model for the european union general data protection regulation. In: _____. *Provenance and Annotation of Data and Processes*. Cham: Springer International Publishing, 2018. (Lecture Notes in Computer Science, v. 11017), p. 45–57. ISBN 978-3-319-98378-3. Disponível em: <http://link.springer.com/10.1007/978-3-319-98379-0_4>.

61 KEET, C. M.; ŁAWRYNOWICZ, A.; D'AMATO, C.; KALOUSIS, A.; NGUYEN, P.; PALMA, R.; STEVENS, R.; HILARIO, M. The data mining optimization ontology. *Journal of Web Semantics*, v. 32, p. 43–53, May 2015. ISSN 15708268.

62 LEEVY, J. L.; KHOSHGOFTAAR, T. M. A survey and analysis of intrusion detection models based on cse-cic-ids2018 big data. *Journal of Big Data*, v. 7, n. 1, p. 104, Dec 2020. ISSN 2196-1115.

63 GHARIB, A.; SHARAFALDIN, I.; LASHKARI, A. H.; GHORBANI, A. A. An evaluation framework for intrusion detection dataset. In: *2016 International Conference on Information Science and Security (ICISS)*. Pattaya, Thailand: IEEE, 2016. p. 1–6. ISBN 978-1-5090-5493-0. Disponível em: <http://ieeexplore.ieee.org/document/7885840/>.

64 SHARAFALDIN, I.; LASHKARI, A. H.; GHORBANI, A. A. Toward generating a new intrusion detection dataset and intrusion traffic characterization:. In: *Proceedings of the 4th International Conference on Information Systems Security and Privacy*. Funchal, Madeira, Portugal: SCITEPRESS - Science and Technology Publications, 2018. p. 108–116. ISBN 978-989-758-282-0. Disponível em: <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0006639801080116>.

65 JANARTHANAN, T.; ZARGARI, S. Feature selection in unsw-nb15 and kddcup'99 datasets. In: *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*. Edinburgh, United Kingdom: IEEE, 2017. p. 1881–1886. ISBN 978-1-5090-1412-5. Disponível em: <http://ieeexplore.ieee.org/document/8001537/>.

66 MOUSTAFA, N.; SLAY, J. The significant features of the unsw-nb15 and the kdd99 data sets for network intrusion detection systems. In: *2015 4th International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*. Kyoto, Japan: IEEE, 2015. p. 25–31. ISBN 978-1-4673-8944-0. Disponível em: <http://ieeexplore.ieee.org/document/7809531/>.

67 XU, J.; FAN, J.; AMMAR, M.; MOON, S. Prefix-preserving ip address anonymization: measurement-based security evaluation and a new cryptography-based scheme. In: *10th IEEE International Conference on Network Protocols, 2002. Proceedings.* Paris, France: IEEE Comput. Soc, 2002. p. 280–289. ISBN 978-0-7695-1856-5. Disponível em: <http://ieeexplore.ieee.org/document/1181415/>.

68 COULL, S. E.; WRIGHT, C. V.; MONROSE, F.; COLLINS, M. P.; REITER, M. K. Playing devil's advocate:. p. 14, 2007.

69 FAN, J.; XU, J.; AMMAR, M. H.; MOON, S. B. Prefix-preserving ip address anonymization: measurement-based security evaluation and a new cryptography-based scheme. *Computer Networks*, v. 46, n. 2, p. 253–272, Oct 2004. ISSN 13891286.

70 SILVA., M.; CAVALCANTI., M.; CAMPOS., M. Privacy-preservation and the use of data for research: A covid-19 use case in randomly generated healthcare records. In: INSTICC. *Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 2: ICEIS,*. [S.l.]: SciTePress, 2022. p. 317–324. ISBN 978-989-758-569-2. ISSN 2184-4992.

71 CAVOUKIAN, A. The 7 foundational principles. p. 12, 2006.

72 LIU, L.; WANG, P.; LIN, J.; LIU, L. Intrusion detection of imbalanced network traffic based on machine learning and deep learning. *IEEE Access*, v. 9, p. 7550–7563, 2021. ISSN 2169-3536.

APÊNDICE A – ESQUEMA RELACIONAL BASEADO NA ONTOLOGIA SEC4-ML-O

```
01 |
02 | /***********************************************
03 | *
04 | *    Schema Sec4ML-O Prov
05 | *
06 | *    Versao: 1.0
07 | *
08 | ***********************************************/
09 |
10 |
11 | --DROP TABLE LOGICALATTRIBUTECATEGORY;
12 | CREATE TABLE LOGICALATTRIBUTECATEGORY
13 | (
14 |   logicalattributecategory_id    UUID primary key,
15 |   logicalattributecharacteristic varchar(500) not null,
16 |   logicalattributecategory_label varchar(500) not null
17 | );
18 |
19 | CREATE INDEX logicalattributecategory_id_idx ON
20 | logicalattributecategory (logicalattributecategory_id);
21 |
22 |
23 | --DROP TABLE LOGICALATTRIBUTE;
24 | CREATE TABLE LOGICALATTRIBUTE
25 | (
26 |  logicalattribute_id         UUID primary key,
27 |  logicalattributecategory_id UUID,
28 |  logicalattribute_label      varchar(500) not null,
29 |  columntable_id              UUID   not null,
30 |  justification_id            UUID
31 | );
32 |
33 | CREATE INDEX logicalattribute_id_idx ON logicalattribute
34 | (logicalattribute_id);
35 |
36 |
37 | --DROP TABLE COLUMNTABLE;
38 | CREATE TABLE COLUMNTABLE
39 | (
40 | columntable_id              UUID primary key,
41 | dataset_id                  UUID not null,
```

```
42 |  column_label                varchar(500) not null,
43 |  structuredtype_id           UUID not null,
44 |  column_lenght                   float
45 |  );
46 |
47 |  CREATE INDEX COLUMNTABLE_id_idx ON COLUMNTABLE (COLUMNTABLE_id);
48 |
49 |
50 |  --DROP TABLE ROWTABLE;
51 |  CREATE TABLE ROWTABLE
52 |  (
53 |   rowtable_id                    varchar(1000) primary key,
54 |   dataset_id                     varchar(1000)
55 |   );
56 |
57 |  CREATE INDEX rowtable_id_idx ON ROWTABLE (rowtable_id);
58 |
59 |
60 |  --DROP TABLE ROWCOLUMNVALUE;
61 |  CREATE TABLE ROWCOLUMNVALUE
62 |  (
63 |   columntable_id               varchar(100),
64 |   rowtable_id                      varchar(100),
65 |   rowcolumn_value            varchar(1000),
66 |   rowcolumn_modifiedvalue    varchar(1000)
67 |   PRIMARY KEY (columntable_id, rowtable_id)
68 |   );
69 |
70 |  CREATE INDEX ROWCOLUMNVALUE_id_idx ON ROWCOLUMNVALUE
71 |  (columntable_id, rowtable_id);
72 |
73 |
74 |  --DROP TABLE SQLSTRUCTUREDTYPE;
75 |  CREATE TABLE SQLSTRUCTUREDTYPE
76 |  (
77 |   structuredtype_id                 UUID primary key,
78 |   structuredtype_label              varchar(500),
79 |   qualitativedatatype_id            UUID,
80 |   quantitativedatatype_id           UUID
81 |   );
82 |
83 |  CREATE INDEX SQLSTRUCTUREDTYPE_id_idx ON SQLSTRUCTUREDTYPE
84 |  (structuredtype_id);
85 |
86 |  --DROP TABLE LABELEDDATASET;
87 |  CREATE TABLE LABELEDDATASET
88 |  (
```

```
89 |  dataset_id                              UUID primary key,
90 |  version                                 integer not null,
91 |  label                                   varchar(500),
92 |  initial_validity                        timestamp,
93 |  final_validity                          timestamp
94 |  );
95 |
96 |  CREATE INDEX LABELEDDATASET_id_idx ON LABELEDDATASET
97 |  (dataset_id);
98 |
99 |
100 |  --DROP TABLE QUALITATIVEDATATYPE;
101 |  CREATE TABLE QUALITATIVEDATATYPE
102 |  (
103 |    qualitativedatatype_id                UUID primary key,
104 |    nominal_flag                              boolean,
105 |    ordinal_flag                              boolean
106 |  );
107 |
108 |  CREATE INDEX QUALITATIVEDATATYPE_id_idx ON QUALITATIVEDATATYPE
109 |  (qualitativedatatype_id);
110 |
111 |  --DROP TABLE QUANTITATIVEDATATYPE;
112 |  CREATE TABLE QUANTITATIVEDATATYPE
113 |  (
114 |    quantitativedatatype_id               UUID primary key,
115 |    discreet_flag                                 boolean,
116 |    continuos_flag                                boolean
117 |  );
118 |
119 |  CREATE INDEX QUANTITATIVEDATATYPE_id_idx ON QUANTITATIVEDATATYPE
120 |  (quantitativedatatype_id);
121 |
122 |
123 |  --DROP TABLE DATAPREPROCESSINGOPERATOR;
124 |  CREATE TABLE DATAPREPROCESSINGOPERATOR
125 |  (
126 |    datapreprocessingoperator_id          UUID primary key,
127 |    datapreprocessingoperator_label   varchar(1000),
128 |    datapreprocessingoperator_dsc         varchar(1000),
129 |    operatorcategory_id                           UUID
130 |  );
131 |
132 |  CREATE INDEX DATAPREPROCESSINGOPERATOR_id_idx ON
        DATAPREPROCESSINGOPERATOR (DATAPREPROCESSINGOPERATOR_id);
133 |
134 |
```

```
135 | --DROP TABLE OPERATORCATEGORY;
136 | CREATE TABLE OPERATORCATEGORY
137 | (
138 |   operatorcategory_id                        UUID primary key,
139 |   logicalattributecategory_id       UUID,
140 |   operatorcategory_dsc                       varchar(500)
141 |  );
142 |
143 | CREATE INDEX OPERATORCATEGORY_id_idx ON OPERATORCATEGORY
144 | (OPERATORCATEGORY_id);
145 |
146 |
147 | --DROP TABLE DATAANONYMIZATIONOPERATOR;
148 | CREATE TABLE DATAANONYMIZATIONOPERATOR
149 | (
150 |   dataanonymizationoperator_id          UUID primary key,
151 |   dataanonymizationoperator_dsc         varchar(1000),
152 |   dataanonymizationoperator_label   varchar(1000),
153 |   operatorcategory_id                           UUID
154 |  );
155 |
156 |  CREATE INDEX DATAANONYMIZATIONOPERATOR_id_idx ON
        DATAANONYMIZATIONOPERATOR (DATAANONYMIZATIONOPERATOR_id);
157 |
158 | --DROP TABLE PARAMETER;
159 | CREATE TABLE PARAMETER
160 | (
161 |  parameter_id                   UUID primary key,
162 |  parameter_dsc                  varchar(1000)
163 | );
164 |
165 | CREATE INDEX PARAMETER_id_idx ON PARAMETER (PARAMETER_id);
166 |
167 |
168 | --DROP TABLE SOFTWARE;
169 | CREATE TABLE SOFTWARE
170 | (
171 |  software_id                            UUID primary key,
172 |  software_dsc                       varchar(1000),
173 |  software_modificationdate      timestamp,
174 |  software_version                       integer,
175 |  operatorcategory_id            UUID,
176 |  algorithm_id                           UUID
177 | );
178 |
179 | CREATE INDEX SOFTWARE_id_idx ON SOFTWARE (SOFTWARE_id);
180 |
```

```
181 | --DROP TABLE ALGORITHM;
182 | CREATE TABLE ALGORITHM
183 | (
184 |   algorithm_id                              UUID,
185 |   algorithm_label                           varchar(500),
186 |   algorithm_version                         integer,
187 |   parameter_id                              UUID,
188 |   operatorcategory_id               UUID,
189 |   primary key (algorithm_id, operatorcategory_id)
190 | );
191 |
192 | CREATE INDEX ALGORITHM_id_idx ON ALGORITHM (ALGORITHM_id);
193 |
194 | --DROP TABLE PARAMETERVALUE;
195 | CREATE TABLE PARAMETERVALUE
196 | (
197 |   parameter_id                              UUID,
198 |   stepexecution_id                  varchar(1000),
199 |   parametervalue                            varchar(1000)
200 |  );
201 |
202 | CREATE INDEX PARAMETERVALUE_id_idx ON PARAMETERVALUE
203 | (PARAMETER_id);
204 |
205 | --**** ETL4LP ****
206 |
207 | --DROP TABLE STEPCOMPOSITION;
208 | CREATE TABLE STEPCOMPOSITION
209 | (
210 |  stepcomposition_id                 UUID primary key,
211 |  stepcomposition_label              varchar(500),
212 |  stepcomposition_previousstep       UUID,
213 |  stepcomposition_nextstep           UUID,
214 |  stepcomposition_modificationdate   timestamp,
215 |  stepcomposition_version            integer,
216 |  operatorcategory_id                UUID
217 |  );
218 |
219 | CREATE INDEX STEPCOMPOSITION_id_idx ON STEPCOMPOSITION
220 | (STEPCOMPOSITION_id);
221 |
222 |
223 | --DROP TABLE STEPEXECUTION;
224 | CREATE TABLE STEPEXECUTION
225 | (
226 |  stepexecution_id                       UUID primary key,
227 |  stepexecution_label                    varchar(500),
```

```
228 |   stepexecution_starttime                date,
229 |   stepexecution_endtime                  date,
230 |   stepexecution_errormessage          text,
231 |   stepexecution_status                   varchar(100),
232 |   workflowexecution_id                   varchar(1000),
233 |   stepcomposition_id                     varchar(1000),
234 |   software_id                                    UUID
235 |  );
236 |
237 |  CREATE INDEX STEPEXECUTION_id_idx ON STEPEXECUTION
238 |  (STEPEXECUTION_id);
239 |
240 |  --DROP TABLE WORKFLOWCOMPOSITION;
241 |  CREATE TABLE WORKFLOWCOMPOSITION
242 |  (
243 |   workflowcomposition_id                 UUID primary key,
244 |   workflowcomposition_label              varchar(500),
245 |   workflowcomposition_modificationdate   timestamp,
246 |   workflowcomposition_version            integer,
247 |   workflowcomposition_contributor        varchar(100)
248 |   );
249 |
250 |  CREATE INDEX WORKFLOWCOMPOSITION_id_idx ON
251 |  WORKFLOWCOMPOSITION (WORKFLOWCOMPOSITION_id);
252 |
253 |
254 |  --DROP TABLE WORKFLOWEXECUTION;
255 |  DROP TABLE WORKFLOWEXECUTION;
256 |  CREATE TABLE WORKFLOWEXECUTION
257 |  (
258 |   workflowexecution_id            varchar(500) primary key,
259 |   workflowcomposition_id      varchar(500),
260 |   workflowexecution_label         varchar(500),
261 |   worflowexecution_unit           varchar(500),
262 |   workflowexecution_starttime timestamp without time zone,
263 |   workflowexecution_endtime       timestamp without time zone,
264 |   workflowexecution_status        varchar(100)
265 |   );
266 |
267 |  CREATE INDEX WORKFLOWEXECUTION_id_idx ON WORKFLOWEXECUTION
268 |  (WORKFLOWEXECUTION_id);
269 |
270 |
271 |  --DROP TABLE PERSON;
272 |  CREATE TABLE PERSON
273 |  (
274 |  person_id                          varchar(500) primary key,
```

```
familyName                varchar(500) not null,
firstName                 varchar(500) not null,
lastName                  varchar(500) not null,
surname                   varchar(500) not null,
gender_concept_id         integer not null,
year_of_birth             integer not null,
month_of_birth            integer not null,
death_date            date,
race_concept_id           integer not null,
ethnicity_concept_id  integer not null,
location_id           integer,
plan                  varchar(1000),
geekcode              varchar(1000),
publications          varchar(1000),
pastProject           varchar(500),
currentProject        varchar(500)
);

CREATE INDEX PERSON_id_idx ON PERSON (PERSON_id);

--DROP TABLE ORGANIZATION;
CREATE TABLE ORGANIZATION
(
organization_id                           varchar(500) primary
   key,
organizationName                          varchar(1000),
organizationnature_id              integer,
location_id                               varchar(500)
);

CREATE INDEX ORGANIZATION_id_idx ON ORGANIZATION
(ORGANIZATION_id);

--DROP TABLE LOCATION;
CREATE TABLE LOCATION
(
location_id                               varchar(500) primary key
   ,
address_1                         varchar(1000),
address_2                         varchar(1000),
city                              varchar(500),
state                             varchar(500),
zip                               varchar(500),
country                           varchar(500)
);

CREATE INDEX LOCATION_id_idx ON LOCATION (LOCATION_id);
```

```
320 |
321 |
322 |  DROP TABLE DATASUBJECT;
323 |  CREATE TABLE DATASUBJECT
324 |  (
325 |    datasubject_id                              UUID primary key,
326 |    person_id                                       UUID,
327 |    organization_id                             UUID
328 |  );
329 |
330 |  CREATE INDEX DATASUBJECT_id_idx ON DATASUBJECT
331 |  (DATASUBJECT_id);
332 |
333 |  DROP TABLE CONSENTAGREEMENT;
334 |  CREATE TABLE CONSENTAGREEMENT
335 |  (
336 |    consentagreement_id           UUID PRIMARY KEY,
337 |    consentagreement_text         text,
338 |    datasubject_id                               UUID,
339 |    rowtable_id                                  UUID
340 |   );
341 |
342 |  CREATE INDEX CONSENTAGREEMENT_id_idx ON CONSENTAGREEMENT
343 |  (CONSENTAGREEMENT_id);
344 |
345 |
346 |  DROP TABLE DATATABLE;
347 |  CREATE TABLE DATATABLE
348 |  (
349 |    data_id                                          UUID primary key,
350 |    datakind_id                             integer,
351 |    datakind_dsc                          varchar(1000)
352 |  );
353 |
354 |  CREATE INDEX DATATABLE_id_idx ON DATATABLE (DATA_id);
355 |
356 |  --DROP TABLE PERSONALDATA;
357 |  CREATE TABLE PERSONALDATA
358 |  (
359 |    personaldata_id                         varchar(500),
360 |    rowtable_id                       varchar(500),
361 |    datakind_id                                  integer,
362 |    consentagreement_id           varchar(500),
363 |    datasubject_id                               varchar(500),
364 |    primary key (personaldata_id, rowtable_id)
365 |  );
366 |
```

```
367 |
368 |  CREATE INDEX PERSONALDATA_id_idx ON PERSONALDATA
369 |  (personaldata_id, rowtable_id);
370 |
371 |
372 |  --DROP TABLE SUPERVISORAUTHORITY;
373 |  CREATE TABLE SUPERVISORAUTHORITY
374 |  (
375 |   supervisorauthority_id             UUID primary key,
376 |   organization_id                    UUID,
377 |   effective_date                     date
378 |   );
379 |
380 |  CREATE INDEX SUPERVISORAUTHORITY_id_idx ON SUPERVISORAUTHORITY
381 |  (SUPERVISORAUTHORITY_id);
382 |
383 |  CREATE DOMAIN JUSTIFICATIONTYPE AS VARCHAR(100)
384 |  NOT NULL
385 |  CHECK (VALUE IN ('Legal Obligation','Vital Interest','Public
         Interest','Official Authority','Legitimate Interest','Contract
         '));
386 |
387 |  --DROP TABLE JUSTIFICATION;
388 |  CREATE TABLE JUSTIFICATION
389 |  (
390 |   justification_id                          UUID,
391 |   controller_id                             UUID,
392 |   supervisorauthority_id                 UUID,
393 |   justification_start                       date,
394 |   justification_end                             date,
395 |   justification_type                        justificationtype,
396 |   request_id                                        UUID,
397 |   primary key (justification_id, controller_id)
398 |   );
399 |
400 |
401 |  CREATE INDEX JUSTIFICATION_id_idx ON JUSTIFICATION
402 |  (justification_id, controller_id);
403 |
404 |  CREATE DOMAIN REQUESTTYPE AS VARCHAR(100)
405 |  NOT NULL
406 |  CHECK ( VALUE IN ('Consent','Withdraw','Access','Correction','
         Erasure','Restriction'));
407 |
408 |
409 |  --DROP TABLE REQUEST;
410 |  CREATE TABLE REQUEST
```

```
411 | (
412 |  request_id                                    UUID,
413 |  datasubject_id                                UUID,
414 |  request_date                                  date,
415 |  request_type                                  requesttype,
416 |  controller_id                                 UUID,
417 |  personaldata_id                               UUID,
418 |  primary key (request_id, datasubject_id)
419 |  );
420 |
421 |  CREATE INDEX REQUEST_id_idx ON REQUEST (REQUEST_id);
422 |
423 | --DROP TABLE CONTROLLER;
424 | CREATE TABLE CONTROLLER
425 | (
426 |  controller_id                                 UUID,
427 |  process_id                                    UUID,
428 |  organization_id                               UUID,
429 |  primary key (controller_id,process_id)
430 |  );
431 |
432 |   CREATE INDEX CONTROLLER_id_idx ON CONTROLLER (CONTROLLER_id);
433 | --DROP TABLE PROCESSOR;
434 | CREATE TABLE PROCESSOR
435 | (
436 |  processor_id                                  UUID,
437 |  process_id                                            UUID,
438 |  organization_id                                   UUID,
439 |  primary key (processor_id, process_id)
440 |  );
441 |
442 | CREATE INDEX PROCESSOR_id_idx ON PROCESSOR (PROCESSOR_id);
443 |
444 |
445 | --DROP TABLE PROCESS;
446 | CREATE TABLE PROCESS
447 | (
448 |  process_id                                                UUID primary
        key,
449 |  process_kind                                  varchar(100),
450 |  stepcomposition_id                            UUID,
451 |  processor_id                                  UUID,
452 |  controller_id                                 UUID,
453 |  justification_id                              UUID
454 | );
455 |
456 |  CREATE INDEX PROCESS_id_idx ON PROCESS (PROCESS_id);
```

ANEXO A – ATRIBUTOS DO CONJUNTO DE DADOS UNSW-NB15

Tabela 10 – Atributos do conjunto de dados UNSW-NB15. Adaptado de (66)

#	Nome	Descrição
	1. Atributos de Fluxo	
1	srcip	Endereço IP de origem
2	sport	Número da porta de origem
3	dstip	Endereço IP de destino
4	dsport	Número da porta de destino
5	proto	Protocolo
	2. Atributos Básicos	
6	state	O estado e seu protocolo dependente
7	dur	Duração total
8	sbytes	Bytes enviados da origem para o destino
9	dbytes	Bytes enviados do destino para a origem
10	sttl	*Time to Live* da origem para o destino
11	dttl	*Time to Live* do destino para a origem
12	sloss	Pacotes da origem retransmitidos os perdidos
13	dloss	Pacotes do destino retransmitidos os perdidos
14	service	Serviço envolvido como FTP, SMTP e DNS
15	sload	Bits por segundo da origem
16	dload	Bits por segundo do destino
17	spkts	Somatório de pacotes da origem para o destino
18	dpkts	Somatório de pacotes do destino para a origem
	3, Atributos relativos ao conteúdo	
19	swin	Valor do anúncio de janela TCP da origem
20	dwin	Valor do anúncio de janela TCP do destino
21	stcpb	Número de sequência de base TCP da origem
22	dtcpb	Número de sequência de base TCP do destino
23	smeansz	Média do tamanho do pacote transmitido pelo IP de origem
24	dmeansz	Média do tamanho do pacote transmitido pelo IP de destino
25	trans_depth	Transação de conexão de http *request/response*
26	res_bdy_len	Tamanho do conteúdo do dado transferido pelo http
	4. Atributos de Tempo	
27	sjit	*Jitter* da origem
28	djit	*Jitter* do destino
29	stime	Hora do início da fila
30	ltime	Hora do fim da fila
31	sintpkt	Hora de chegada entre pacotes da origem
32	dintpkt	Hora de chegada entre pacotes do destino
33	tcprtl	Somatório do *synack* e *ackdat* (configuração de tempo de ida e volta)
34	synack	Período de tempo compreendido entre os pacotes *SYN* e *SYN_ACK*
35	ackdat	Período de tempo compreendido entre os pacotes *SYN_ACK* e *SYN*
36	is_sm_ips_ports	Se *srcip* ¯dstip e *sport* ¯dsport, assinala 1 senão 0
	5. Atributos gerados adicionalmente	
37	ct_state_ttl	Número de cada estado de acordo com os valores de sttl e dttl
38	ct_fwl_http_mthd	Número de *Get* e *Post* no serviço http
39	is_ftp_login	Se a sessão FTP é acessada com usuário e senha atribui 1 senão 0
40	ct_ftp_cmd	Número de fluxos que tem comandos em uma sessão FTP
41	ct_srv_src	Número de linhas com o mesmo *service* e o mesmo *srcip* em 100 linhas
42	ct_srv_dst	Número de linhas com o mesmo *service* e o mesmo dstip em 100 linhas
43	ct_dst_ltm	Número de linhas com o mesmo *dstip* em 100 linhas
44	ct_src_ltm	Número de linhas com o mesmo *srcip* em 100 linhas
45	ct_src_dport_ltm	Número de linhas com o mesmo *srcip* e *dsport* em 100 linhas
46	ct_dst_sport_ltm	Número de linhas com o mesmo *dstip* e *sport* em 100 linhas
47	ct_dst_src_ltm	Número de linhas com o mesmo *srcip* e *dstip* em 100 linhas
	6. Atributos de rotulagem	
48	attack_cat	Nome de cada categoria de ataque
49	label	0 para registro normal e 1 para registro de ataque

ANEXO B – ATRIBUTOS DO CONJUNTO DE DADOS CSE-CIC-IDS2018

Tabela 11 – Atributos do conjunto de dados CSE-CIC-IDS2018. Adaptado de (72) e (17).

#	Nome	Descrição
		1.Atributos básicos de conexão
1	flow_id	Identificador do registro
2	src_ip	Endereço IP de origem
3	src_port	Número da porta de origem
4	dst_ip	Endereço IP de destino
5	dst_port	Número da porta de destino
6	protocol	Protocolo
7	timestamp	Data e hora do registro
		2. Atributos de pacotes de rede
8	total_fwd_packet	Total de pacotes no sentido do fluxo
9	total_bwd_packets	Total de pacotes no sentido inverso do fluxo
10	total_len_fwd_packet	Tamanho total do pacote no sentido do fluxo
11	total_len_bwd_packet	Tamanho total do pacote no sentido inverso do fluxo
12	fwd_packet_len_min	Tamanho mínimo do pacote no sentido do fluxo
13	fwd_packet_len_max	Tamanho máximo do pacote no sentido do fluxo
14	fwd_packet_len_mean	Tamanho médio do pacote no sentido do fluxo
15	fwd_packet_length_std	Desvio padrão do tamanho do pacote no sentido do fluxo
16	bwd_packet_len_min	Tamanho mínimo do pacote no sentido inverso do fluxo
17	bwd_packet_len_max	Tamanho máximo do pacote no sentido inverso do fluxo
18	bwd_packet_len_mean	Tamanho médio do pacote no sentido inverso do fluxo
19	bwd_packet_len_std	Desvio padrão do tamanho do pacote no sentido inverso do fluxo
		3. Atributos de fluxo de rede
20	flow_duration	Duração do fluxo
21	flow_byte/s	Número de bytes no fluxo por segundo
22	flow_packets/s	Número de pacotes no fluxo por segundo
23	flow_iat_mean	Tempo médio entre dois pacotes enviados no fluxo
24	flow_iat_std	Desvio padrão no tempo entre dois pacotes enviados no fluxo
25	flow_iat_min	Tempo mínimo entre dois pacotes enviados no fluxo
26	fwd_iat_min	Tempo mínimo entre dois pacotes enviados no sentido do fluxo
27	fwd_iat_max	Tempo máximo entre dois pacotes enviados no sentido do fluxo
28	fwd_iat_mean	Tempo médio entre dois pacotes enviados no sentido do fluxo
29	fwd_iat_std	Desvio padrão no tempo entre dois pacotes enviados no sentido do fluxo
30	fwd_iat_total	Tempo total entre dois pacotes enviados no sentido do fluxo
31	bwd_iat_min	Tempo mínimo entre dois pacotes enviados no sentido inverso do fluxo
32	bwd_iat_max	Tempo máximo entre dois pacotes enviados no sentido inverso do fluxo
33	bwd_iat_mean	Tempo médio entre dois pacotes enviados no sentido inverso do fluxo
34	bwd_iat_std	Desvio padrão no tempo entre dois pacotes enviados no sentido inverso do fluxo
35	bwd_iat_total	Tempo total entre dois pacotes enviados no sentido inverso do fluxo
36	fwd_psh_flag	Número de vezes que o 'flag' PSH foi enviado em pacotes trafegando no sentido do fluxo (0 para UDP)
37	bwd_psh_flag	Número de vezes que o 'flag' PSH foi enviado em pacotes trafegando no sentido inverso do fluxo(0 para UDP)
38	fwd_urg_flag	Número de vezes que o 'flag' URG foi enviado em pacotes trafegando no sentido do fluxo (0 para UDP)
39	fwd_header_len	Total de bytes usados em cabeçalhos no sentido do fluxo
40	bwd_header_len	Total de bytes usados em cabeçalhos no sentido inverso do fluxo
41	fwd_packets/s	Número de pacotes por segundo no sentido do fluxo
42	bwd_packets/s	Número de pacotes por segundo no sentido inverso do fluxo
43	min_packet_len	Tamanho mínimo de um pacote
44	max_packet_len	Tamanho máximo de um pacote
45	packet_len_mean	Tamanho médio de um pacote
46	packet_len_std	Desvio padrão do tamanho de um pacote
47	packet_len_variance	Varância do tamanho do pacote

Tabela 12 – Atributos do conjunto de dados CSE-CIC-IDS2018. (continuação)

#	Nome	Descrição
48	FIN_flag_count	Número de pacotes com FIN
49	SYN_flag_count	Número de pacotes com SYN
50	RST_flag_count	Número de pacotes com RST
51	PSH_flag_count	Número de pacotes com PSH
52	ACK_flag_count	Número de pacotes com ACK
53	URG_flag_count	Número de pacotes com URG
54	CWR_flag_count	Número de pacotes com CWR
55	ECE_flag_count	Número de pacotes com ECE
56	down/up_ratio	Taxa de *download* e *upload*
57	average_packet_size	Tamnho médio do pacote
58	avg_fwd_segment_size	Tamanho médio observado no sentido do fluxo
59	avg_bwd_segment_size	Número de bytes médio no sentido do fluxo
60	fwd Header Length	Tamanho do pacote de cabeçalho
61	fwd_avg_bytes/bulk	Número de bytes médio em massa no sentido inverso do fluxo
62	fwd_avg_packet/bulk	Número de pacotes médio em massa no sentido do fluxo
63	fwd_avg_bulk_rate	Número médio de taxa de massa no sentido do fluxo
64	bwd_avg_bytes/bulk	Número médio de bytes de taxa de massa no sentido inverso do fluxo
65	bwd_avg_packet/bulk	Número médio de pacotes de taxa de massa no sentido inverso do fluxo
66	bwd_avg_bulk_rate	Número médio de taxa de massa no sentido inverso do fluxo
67	subflow_fwd_packets	O número médio de pacotes em um subfluxo no sentido do fluxo
68	subflow_fwd_bytes	O número médio de bytes em um subfluxo no sentido do fluxo
69	subflow_bwd_packets	O número médio de pacotes em um subfluxo no sentido inverso do fluxo
70	subflow_bwd_bytes	O número médio de bytes em um subfluxo no sentido inverso do fluxo
71	init_win_bytes_forward	Número total de bytes enviados na janela inicial no sentido do fluxo
72	init_win_bytes_backward	Número total de bytes enviados na janela inicial no sentido inverso do fluxo
73	act_data_pkt_forward	Quantidade de pacote com pelo menos 1 byte de TCP "payload" no sentido do fluxo
74	min_seg_size_forward	Tamanho mínimo do segmento observado no sentido do fluxo
75	active_min	Tempo mínimo que um fluxo esteve inativo antes de se tornar ocioso
76	active_mean	Tempo médio que um fluxo esteve inativo antes de se tornar ocioso
77	active_max	Tempo máximo que um fluxo esteve inativo antes de se tornar ocioso
78	active_std	Desvio padrão do tempo que um fluxo esteve inativo antes de se tornar ocioso
79	idle_min	Tempo mínimo que um fluxo esteve ocioso antes de se tornar inativo
80	idle_mean	Tempo médio que um fluxo esteve ocioso antes de se tornar inativo
81	idle_max	Tempo máximo que um fluxo esteve ocioso antes de se tornar inativo
82	idle_std	Desvio padrão do tempo que um fluxo esteve ocioso antes de se tornar inativo

ANEXO C – ATRIBUTOS UTILIZADOS NA CRIAÇÃO DO MODELO ML COM O CONJUNTO DE DADOS CSE-CIC-IDS2018

Tabela 13 – Atributos Utilizados na criação do modelo de ML com o conjunto de dados CSE-CIC-IDS2018.

#	Nome	Descrição
	1.Atributos básicos de conexão	
1	protocol	Protocolo
	2. Atributos de pacotes de rede	
2	total_fwd_packet	Total de pacotes no sentido do fluxo
3	total_len_fwd_packet	Tamanho total do pacote no sentido do fluxo
4	fwd_packet_len_min	Tamanho mínimo do pacote no sentido do fluxo
5	fwd_packet_len_max	Tamanho máximo do pacote no sentido do fluxo
6	fwd_packet_len_mean	Tamanho médio do pacote no sentido do fluxo
7	fwd_packet_length_std	Desvio padrão do tamanho do pacote no sentido do fluxo
8	bwd_packet_len_min	Tamanho mínimo do pacote no sentido inverso do fluxo
9	bwd_packet_len_max	Tamanho máximo do pacote no sentido inverso do fluxo
10	bwd_packet_len_mean	Tamanho médio do pacote no sentido inverso do fluxo
11	bwd_packet_len_std	Desvio padrão do tamanho do pacote no sentido inverso do fluxo
	3. Atributos de fluxo de rede	
12	flow_duration	Duração do fluxo
13	flow_byte/s	Número de bytes no fluxo por segundo
14	flow_packets/s	Número de pacotes no fluxo por segundo
15	flow_iat_mean	Tempo médio entre dois pacotes enviados no fluxo
16	flow_iat_std	Desvio padrão no tempo entre dois pacotes enviados no fluxo
17	flow_iat_min	Tempo mínimo entre dois pacotes enviados no fluxo
18	fwd_iat_min	Tempo mínimo entre dois pacotes enviados no sentido do fluxo
19	fwd_iat_max	Tempo máximo entre dois pacotes enviados no sentido do fluxo
20	fwd_iat_mean	Tempo médio entre dois pacotes enviados no sentido do fluxo
21	fwd_iat_std	Desvio padrão no tempo entre dois pacotes enviados no sentido do fluxo
22	fwd_iat_total	Tempo total entre dois pacotes enviados no sentido do fluxo
23	bwd_iat_min	Tempo mínimo entre dois pacotes enviados no sentido inverso do fluxo
24	bwd_iat_max	Tempo máximo entre dois pacotes enviados no sentido inverso do fluxo
25	bwd_iat_mean	Tempo médio entre dois pacotes enviados no sentido inverso do fluxo
26	bwd_iat_std	Desvio padrão no tempo entre dois pacotes enviados no sentido inverso do fluxo
27	bwd_iat_total	Tempo total entre dois pacotes enviados no sentido inverso do fluxo
28	fwd_header_len	Total de bytes usados em cabeçalhos no sentido do fluxo
29	fwd_packets/s	Número de pacotes por segundo no sentido do fluxo
30	bwd_packets/s	Número de pacotes por segundo no sentido inverso do fluxo
31	min_packet_len	Tamanho mínimo de um pacote
32	max_packet_len	Tamanho máximo de um pacote
33	packet_len_mean	Tamanho médio de um pacote
34	packet_len_std	Desvio padrão do tamanho de um pacote
35	packet_len_variance	Varância do tamanho do pacote

Tabela 14 – Atributos Utilizados na criação do modelo de ML com o conjunto de dados CSE-CIC-IDS2018 (continuação).

#	Nome	Descrição
36	RST_flag_count	Número de pacotes com RST
37	PSH_flag_count	Número de pacotes com PSH
38	ACK_flag_count	Número de pacotes com ACK
39	ECE_flag_count	Número de pacotes com ECE
40	down/up_ratio	Taxa de *download* e *upload*
41	average_packet_size	Tamnho médio do pacote
42	avg_fwd_segment_size	Tamanho médio observado no sentido do fluxo
43	avg_bwd_segment_size	Número de bytes médio no sentido do fluxo
44	fwd Header Length	Tamanho do pacote de cabeçalho
45	subflow_fwd_packets	O número médio de pacotes em um subfluxo no sentido do fluxo
46	subflow_fwd_bytes	O número médio de bytes em um subfluxo no sentido do fluxo
47	init_win_bytes_forward	Número total de bytes enviados na janela inicial no sentido do fluxo
48	init_win_bytes_backward	Número total de bytes enviados na janela inicial no sentido inverso do fluxo
49	act_data_pkt_forward	Quantidade de pacote com pelo menos 1 byte de TCP "payload" no sentido do fluxo
50	min_seg_size_forward	Tamanho mínimo do segmento observado no sentido do fluxo
51	active_min	Tempo mínimo que um fluxo esteve inativo antes de se tornar ocioso
52	active_mean	Tempo médio que um fluxo esteve inativo antes de se tornar ocioso
53	active_max	Tempo máximo que um fluxo esteve inativo antes de se tornar ocioso
54	idle_min	Tempo mínimo que um fluxo esteve ocioso antes de se tornar inativo
55	idle_mean	Tempo médio que um fluxo esteve ocioso antes de se tornar inativo
56	idle_max	Tempo máximo que um fluxo esteve ocioso antes de se tornar inativo