UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Radharani Santos Rocha

ZTA-RAD: Conjunto de Dados de Avaliação de Risco para Ameaças Internas em Arquiteturas Zero Trust

Uberlândia, Brasil

2025

UNIVERSIDADE FEDERAL DE UBERLÂNDIA

Radharani Santos Rocha

ZTA-RAD: Conjunto de Dados de Avaliação de Risco para Ameaças Internas em Arquiteturas Zero Trust

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Orientador: Prof. Dr. Rodrigo Sanches Miani

Universidade Federal de Uberlândia – UFU

Faculdade de Computação

Bacharelado em Sistemas de Informação

Uberlândia, Brasil

2025

Radharani Santos Rocha

ZTA-RAD: Conjunto de Dados de Avaliação de Risco para Ameaças Internas em Arquiteturas Zero Trust

Trabalho de conclusão de curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, como parte dos requisitos exigidos para a obtenção título de Bacharel em Sistemas de Informação.

Trabalho aprovado. Uberlândia, Brasil, 16 de setembro 2025:

Prof. Dr. Rodrigo Sanches Miani
Orientador

Prof. Dr. Claudiney Ramos Tinoco

Prof. Dr. Luís Fernando Faina

Uberlândia, Brasil
2025

Resumo

As ameaças internas, do inglês (*insider threats*), configuram-se como um dos principais riscos de segurança da informação, sobretudo em ambientes corporativos que lidam com dados sensíveis. Diferentemente das ameaças externas, esses ataques exploram credenciais válidas e conhecimento prévio da infraestrutura, tornando insuficientes os modelos tradicionais de segurança baseados em perímetro. Nesse contexto, a Arquitetura de Confiança Zero (ZTA) desponta como um paradigma essencial, ao adotar o princípio do *"nunca confiar, sempre verificar"*, com verificação contínua, microsegmentação e monitoramento em tempo real. Motivado pela ausência de *datasets* específicos para esse cenário, este trabalho propõe o ZTA-RAD (*Zero Trust Architecture – Risk Assessment Dataset*), derivado e expandido a partir do CERT, incorporando métricas de logon, dispositivos e acessos HTTP. O *dataset* foi rotulado por meio de duas abordagens: (i) curadoria de especialista e (ii) análise de risco realizada por cinco LLMs. Para avaliação, foram conduzidos experimentos de treinamento e validação com três algoritmos de aprendizado de máquina (MLP, *Random Forest* e SVM), em cenários balanceados e não balanceados. Os resultados indicam que MLP e *Random Forest* obtiveram desempenho superior, sobretudo após o balanceamento via SMOTE, enquanto o SVM mostrou maior sensibilidade ao desbalanceamento. Além disso, verificou-se que a rotulagem do especialista proporcionou a construção de classificadores mais consistentes em comparação às LLMs, reforçando a importância da curadoria humana. Em síntese, o trabalho contribui para disponibilizar um *dataset* inédito para o estudo de ameaças internas em ZTA, oferecendo subsídios metodológicos e práticos para a evolução de soluções baseadas em aprendizado de máquina.

Palavras-chave:Ameaças Internas, *Dataset*, Rotulagem de Risco, ZTA.

Abstract

Internal threats, often referred to as insider threats, pose significant risks to information security, particularly in corporate environments that manage sensitive data. Unlike external threats, insider attacks utilize valid credentials and insider knowledge of the organization's infrastructure, which renders traditional perimeter-based security models inadequate. In this context, the Zero Trust Architecture (ZTA) becomes a crucial approach, adhering to the principle of "never trust, always verify."This model emphasizes continuous verification, microsegmentation, and real-time monitoring to enhance security. Motivated by the absence of specific datasets for this scenario, this work introduces ZTA-RAD (Zero Trust Architecture – Risk Assessment Dataset). This dataset is derived from and expanded upon the CERT dataset, incorporating metrics related to logon, devices, and HTTP access. The dataset was labeled using two methods: (i) expert curation and (ii) risk analysis conducted by five large language models (LLMs). Evaluation, training, and validation experiments were conducted with three machine learning algorithms (MLP, Random Forest, and SVM) in both balanced and imbalanced scenarios. The results indicate that MLP and Random Forest achieved superior performance, especially after balancing with SMOTE, while SVM showed greater sensitivity to class imbalance. Additionally, it was observed that expert labeling enabled the construction of more consistent classifiers compared to those generated by LLMs, reinforcing the importance of human curation. In conclusion, this study offers a new dataset for investigating insider threats within Zero Trust Architecture (ZTA), providing both methodological and practical insights to enhance machine learning–based security solutions.

Keywords: Dataset, Insider Threats, Risk Labeling, ZTA.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AUC	*Area Under the Curve*
AWS	*Amazon Web Services*
IDS	*Intrusion Detection System*
KNN	*k-Nearest Neighbors*
LLM	*Large Language Model*
ML	*Machine Learning*
MLP	*Multi Layer Perceptron*
NIST	*National Institute Of Standards and Technology*
PEP	*Policy Enforcement Point*
RNA	Rede Neural Artificial
ROC	*Receiver Operating Characteristic*
SDN	*Software Defined Networking*
SMOTE	*Synthetic Minority Over-sampling Technique*
SVM	*Support Vector Machines*
TWOS	*The Wolf of SUTD*
VPN	*Virtual Private Network*
XAI	Inteligência Artificial Explicável
ZTA	*Zero Trust Architecture*
ZTA-RAD	*Zero Trust Architecture – Risk Assessment Dataset*
ZTNA	*Zero Trust Network Access*

Sumário

1 Introdução

No cenário atual de segurança da informação, as ameaças internas, do inglês (*insider threats*), figuram entre os riscos mais críticos para as organizações, uma vez que partem de indivíduos com acesso legítimo a sistemas, redes ou dados sensíveis (NADEAU, 2024).

Ainda de acordo com Nadeau (2024), avaliações dos anos de 2023 e 2024 indicam que incidentes envolvendo ameaças internas representam perdas financeiras anuais da ordem de milhões de dólares em escala global, sendo que, em muitos casos, o impacto operacional e reputacional supera o prejuízo financeiro direto. Além disso, mais de 30% dos incidentes de segurança têm origem em ações internas, sejam elas intencionais ou decorrentes de negligência, como a reutilização de senhas fracas, acessos fora de horário autorizado ou o descumprimento de protocolos estabelecidos.

As ameaças internas podem se manifestar de diferentes formas, abrangendo tanto ações intencionais quanto não intencionais. No primeiro caso, destacam-se situações em que usuários, movidos por insatisfação, benefícios financeiros ou outros fatores motivacionais, utilizam seus acessos legítimos para atividades maliciosas, como a exfiltração de dados sensíveis, a sabotagem de sistemas ou a espionagem corporativa. Exemplos incluem colaboradores que copiam informações estratégicas para repassar a concorrentes ou que provocam deliberadamente a indisponibilidade de serviços críticos (CISA, 2025).

Sob outra perspectiva, encontram-se os comportamentos decorrentes de descuido ou negligência, que embora não apresentem intenção maliciosa, produzem impactos igualmente relevantes. Entre esses, destacam-se a utilização de senhas fracas, o compartilhamento indevido de credenciais, o acesso a sistemas corporativos a partir de dispositivos não autorizados e o descumprimento de protocolos de segurança previamente estabelecidos. Tais práticas, mesmo acidentais, ampliam a superfície de ataque e criam brechas que podem ser exploradas por agentes externos, configurando-se como um risco significativo para a segurança organizacional (COLLINS, 2016; CISA, 2025).

Nesse cenário, observa-se que os mecanismos tradicionais, baseados em perímetro, não são suficientes para lidar com a complexidade e a gravidade de ameaças internas atuais. A exploração de credenciais legítimas e o conhecimento prévio da infraestrutura pelos próprios usuários exigem mecanismos de proteção contínuos e adaptativos. É nesse contexto que a Arquitetura de Confiança Zero (ZTA) se consolida como um novo paradigma de segurança, fundamentado no princípio do *"nunca confiar, sempre verificar"*. A ZTA propõe a substituição do modelo de confiança implícita por controles dinâmicos, com autenticação e autorização constantes, microsegmentação de recursos e monitoramento em tempo real, de forma a reduzir a superfície de ataque e mitigar riscos associados à movi-

mentação lateral, abuso de privilégios e uso indevido de credenciais (Fortinet Inc., 2025; STAFFORD, 2020).

Atualmente tem-se observado a utilização de métodos de aprendizado de máquina supervisionada como estratégia inovadora e amplamente explorada para a criação de modelos preditivos capazes de identificar padrões de comportamento associados a ameaças internas (SARHAN; ALTWAIJRY, 2022). Esses modelos, quando treinados em *datasets* devidamente rotulados e contextualizados, são capazes de aprender relações sutis entre atributos de risco, permitindo não apenas uma detecção mais precisa de atividades maliciosas, mas também maior capacidade de generalização frente a novos incidentes.

Nesse sentido, torna-se imprescindível a existência de um conjunto de dados que reflita cenários alinhados à ZTA, servindo como base para treinar e validar algoritmos de *machine learning* em condições mais próximas da realidade operacional. No contexto de *insider threats* sob a perspectiva da ZTA, a maioria dos *datasets* disponíveis, como o *CERT Insider Threat Test Dataset* (LINDAUER, 2020), foi concebida em contextos de segurança perimetral e não contempla práticas centrais da confiança zero, como microsegmentação, controle de dispositivos e políticas dinâmicas de autenticação. Tal limitação compromete a construção de cenários realistas para avaliação de técnicas de detecção modernas.

Motivado por essa lacuna, este trabalho propõe a construção do ZTA-RAD, abreviação de *Zero Trust Architecture – Risk Assessment Dataset*, um *dataset* orientado à avaliação de risco em ambientes de confiança zero ZTA. O ZTA-RAD foi derivado e expandido a partir do CERT, contemplando métricas de logon, dispositivos e acessos HTTP, e rotulado por meio de duas abordagens complementares: (i) a análise de uma especialista da área e (ii) a classificação de risco realizada por cinco Modelos de Linguagem de Grande Escala (LLMs).

As questões de pesquisa que guiam este trabalho são apresentadas abaixo:

a) Sob a perspectiva da análise de risco, em que medida a rotulagem realizada por um especialista difere daquela produzida por uma LLM?

b) Os classificadores treinados e validados a partir dos *datasets* rotulados por especialista e por LLM apresentam diferenças significativas em termos de desempenho e métricas de avaliação?

c) Como o balanceamento de classes, por meio de técnicas de enriquecimento de dados, influencia o processo de treinamento e validação de modelos de aprendizado de máquina aplicados à detecção de ameaças internas?

1.1 Objetivos

O objetivo principal deste trabalho é instituir um *dataset*, denominado ZTA-RAD, para o estudo de ameaças internas em cenários fundamentados na Arquitetura de Confiança Zero (ZTA), possibilitando a aplicação de algoritmos de aprendizado de máquinas no que tange à detecção de comportamentos de risco.

Para alcançar o objetivo geral, definem-se os seguintes objetivos específicos:

a) Avaliar a contribuição de especialistas humanos e de Modelos de Linguagem de Grande Escala (LLMs) no processo de rotulagem de risco, comparando suas perspectivas na análise de ameaças internas;

b) Investigar o impacto do balanceamento de classes, por meio de técnicas de enriquecimento de dados, no treinamento e validação de modelos de aprendizado de máquina;

c) Analisar criticamente o desempenho de diferentes algoritmos de classificação (MLP, *Random Forest* e SVM) frente ao problema da detecção de ameaças internas em cenários de ZTA;

1.2 Contribuição do Trabalho

Este trabalho contribui em quatro frentes complementares: (i) propõe e descreve o *ZTA-RAD*, um *dataset* inédito orientado à Arquitetura de Confiança Zero (ZTA), estruturado a partir de logs de *logon*, dispositivos e HTTP rotulado em níveis de risco (baixo, médio e alto), informações ampliadas do *dataset* CERT; (ii) investiga um processo híbrido de rotulagem de risco, combinando curadoria de especialista e inferência de cinco LLMs, oferecendo evidências empíricas sobre convergências e limites dessas abordagens, com ênfase no papel da curadoria humana em cenários críticos; (iii) avalia, de forma sistemática, o impacto do balanceamento de classes (via SMOTE) no treinamento e validação de modelos supervisionados, destacando ganhos de sensibilidade nas classes minoritárias e a redução de falsos negativos; e (iv) valida um *pipeline* reprodutível de aprendizado de máquina, contemplando preparação de dados, balanceamento, treinamento e avaliação, com três algoritmos (MLP, *Random Forest* e SVM), oferecendo subsídios metodológicos e práticos para a detecção de ameaças internas sob o paradigma ZTA.

Os experimentos realizados demonstraram que os modelos baseados em MLP e *Random Forest* apresentaram os melhores desempenhos, com acurácia próxima de 1,00 e métricas elevadas de precisão, revocação e F1-score, especialmente quando treinados em *datasets* balanceados. O SVM, por sua vez, mostrou maior sensibilidade ao desbalanceamento, resultando em métricas inferiores e maior instabilidade. Além disso, verificou-se

que a rotulagem realizada pela especialista possibilitou resultados mais consistentes que a rotulagem via LLMs, reforçando o papel da curadoria humana. Por fim, o balanceamento das classes revelou-se essencial para mitigar o viés em favor da classe majoritária, reduzindo falsos negativos e aumentando a confiabilidade na detecção de ameaças internas.

É importante destacar que o *prompt* utilizado para orientar as LLMs no processo de rotulagem de risco encontra-se descrito no Apêndice A deste trabalho. Já os códigos-fonte, o *dataset* produzido e as imagens geradas foram disponibilizados como artefatos no *GitHub*, cujo link de acesso está indicado no Apêndice B.

1.3 Organização da Monografia

Esta monografia está organizada em sete capítulos. O Capítulo 1 apresenta a introdução, contextualizando o problema das ameaças internas, a relevância da Arquitetura de Confiança Zero (ZTA), a motivação, os objetivos e as questões de pesquisa. O Capítulo 2 aborda o arcabouço teórico e os trabalhos correlatos, discutindo os principais conceitos relacionados à ZTA, às ameaças internas e às técnicas de aprendizado de máquina aplicadas à detecção de riscos. O Capítulo 3 descreve os métodos utilizados para alcançar os resultados e responder às questões de pesquisa. O Capítulo 4 detalha o processo de construção do *dataset* ZTA-RAD, os procedimentos de rotulagem de risco e a preparação dos dados para os experimentos. O Capítulo 5 apresenta os resultados obtidos a partir do treinamento e validação dos classificadores, seguido de sua análise crítica. O Capítulo 6 reúne as considerações e discussões sobre os experimentos realizados. Por fim, o Capítulo 7 apresenta a conclusão, destacando as contribuições alcançadas, as limitações identificadas e as perspectivas para trabalhos futuros.

2 Arcabouço Teórico

O arcabouço teórico deste trabalho reúne os conceitos e tecnologias que sustentam o desenvolvimento e a aplicação de soluções voltadas à detecção de ameaças internas em ambientes corporativos. Inicialmente, são apresentados os princípios da Arquitetura de Confiança Zero (ZTA)[1] e sua aplicação prática por meio de ZTNA[2], abordagens que rompem com o modelo tradicional de segurança baseado em perímetro e estabelecem autenticação e verificação contínuas como pilares essenciais.

Em seguida, discute-se o conceito de ameaças internas (*Insider Threats*), explorando suas tipologias, motivações e impactos no contexto organizacional, além da importância de *datasets* específicos para o estudo e modelagem dessas ameaças. Por fim, são exploradas as contribuições da Inteligência Artificial (IA) e do Aprendizado de Máquina (*Machine Learning*) como ferramentas para a detecção de comportamentos anômalos e prevenção de incidentes.

2.1 *Zero Trust Architecture* - ZTA

De acordo com Nascimento e Neves (2024), em uma rede cuja segurança é baseada no controle de perímetro, por exemplo, através de um *Firewall*, presume-se que todo acesso autorizado por esse mecanismo é confiável e, a partir desse ponto, possa trafegar livremente dentro da rede segura conforme o nível de permissão atribuído. Entretanto, uma vez permitido o acesso, muitas vezes não se monitora ou restringe adequadamente ações que possam gerar vulnerabilidades, possibilitando que comportamentos maliciosos ou inadequados, intencionais ou acidentais, representem riscos e comprometam a integridade da rede, um exemplo disso é o atual uso da VPN.

Ainda de acordo com Nascimento e Neves (2024), esse cenário representa um desafio crítico para as organizações, pois a manutenção de uma rede segura exige processos cada vez mais rigorosos, demandando planejamento estratégico, monitoramento constante e ações preventivas contínuas. A complexidade crescente dos ambientes corporativos, somada ao aumento da sofisticação das ameaças, tais como: *Ransomware*, *Spyware* e *Phishing*, dentre outras, torna indispensável a adoção de mecanismos de segurança capazes de realizar verificações contínuas, tanto internamente quanto externamente.

Conforme a norma NIST 800-207, *National Institute of Standards and Technology*, a Arquitetura de Confiança Zero deve ser descrita como um paradigma de segurança que

[1] Do inglês *Zero Trust Architecture*
[2] Do inglês *Zero Trust Network Access*

aplica os conceitos de confiança zero, ou seja, ”nunca confiar, sempre verificar”, abrangendo o relacionamento entre componentes, o planejamento do fluxo de trabalho e as políticas de acesso (STAFFORD, 2020).

Destarte, pode-se entender que o paradigma de confiança zero é definido como o conjunto de princípios destinados a reduzir a incerteza na aplicação de decisões de controle de acesso, assumindo que a rede pode estar comprometida. Sendo assim o princípio fundamental “nunca confiar, sempre verificar” contrasta diretamente com o modelo tradicional de segurança baseado em perímetro, no qual, uma vez autorizado o acesso, o tráfego é tratado como confiável (STAFFORD, 2020; NASCIMENTO; NEVES, 2024).

Em suma, a ZTA exige verificação contínua de identidades, dispositivos e comportamentos, mitigando riscos e fortalecendo a resiliência organizacional. A Figura 1 ilustra os principais componentes lógicos de uma arquitetura *Zero Trust*.

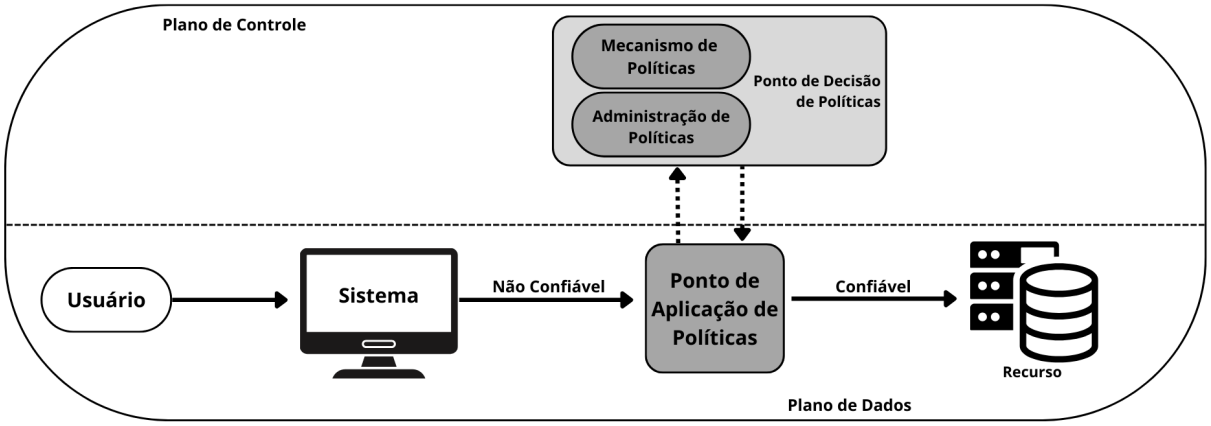

Figura 1 – Principais Componentes Lógicos do Zero Trust
Figura adaptada de (STAFFORD, 2020)

Como pode ser observado na Figura 1, o Mecanismo de Políticas é o componente responsável por avaliar as regras e políticas de segurança, tomando decisões de autorização com base em dados contextuais e critérios definidos. O componente Administração de Políticas gerencia e implementa as decisões do Mecanismo de Políticas, configurando e aplicando as regras nos pontos de controle adequados. E por último, o Ponto de Aplicação de Políticas é o local onde as políticas são efetivamente aplicadas, permitindo ou bloqueando o acesso do solicitante (Usuário) aos recursos corporativos, conforme decisão do Mecanismo de Políticas (STAFFORD, 2020).

É possível observar que a arquitetura ZTA prevista por Stafford (2020) incorpora a filosofia de redes SDN, onde existe um plano de dados e um plano de controle, este último por sua vez compreende os elementos de controle da arquitetura que irão controlar os elementos do plano de dados.

2.1.1 *Zero Trust Network Access* - ZTNA

O ZTNA trata-se da aplicação prática dos princípios da ZTA, voltada especificamente para o controle de acesso remoto ou interno a recursos corporativos (Fortinet Inc., 2025). Diferente de soluções tradicionais, como VPNs, que concedem amplo acesso à rede após a autenticação inicial, o ZTNA estabelece conexões ponto a ponto entre o usuário e o recurso solicitado, aplicando autenticação e verificação contínua da postura de segurança do dispositivo antes e durante a sessão (STAFFORD, 2020; Fortinet Inc., 2025).

Essa abordagem garante que cada solicitação de acesso seja avaliada em tempo real, considerando identidade, contexto, localização e estado do *endpoint*, além de integrar-se a mecanismos de gestão de identidades e sistemas de monitoramento. Assim, o ZTNA reforça o princípio de "nunca confiar, sempre verificar", oferecendo granularidade no controle de permissões e reduzindo significativamente a superfície de ataque, mesmo para usuários autenticados e dispositivos previamente autorizados.

Conforme discutido em Mavroudis (2024), o ZTNA aplica autenticação e verificação contínuas, limitando movimento lateral e reduzindo a superfície de ataque. Esses mecanismos são particularmente relevantes no enfrentamento de ameaças internas, uma vez que estas partem de usuários já autenticados ou dispositivos previamente autorizados.

2.2 Ameaças Internas (*Insider Threats*)

Conforme Tian et al. (2025), a ameaça interna compreende riscos de segurança decorrentes de indivíduos com acesso autorizado a sistemas ou dados de uma organização, que utilizam essas permissões de forma maliciosa ou aé mesmo não intencional, ocasionando vazamento de informações, acessos não autorizados, prejuízos operacionais ou comprometimento da integridade institucional.

Esses indivíduos podem ser funcionários ativos, ex-colaboradores, prestadores de serviços, parceiros comerciais ou qualquer pessoa detentora de credenciais válidas, obtidas de formas legais ou ilegais, podendo atuar de maneira deliberada para fins como espionagem corporativa, roubo de propriedade intelectual, fraude financeira ou divulgação não autorizada de informações sensíveis.

Alternativamente, podem agir de forma negligente, descumprindo normas e protocolos de segurança, o que favorece a exploração de vulnerabilidades internas. A ameaça interna também pode emergir da exploração de um agente interno descuidado, cuja conduta insegura, como a reutilização de senhas fracas, o não cumprimento de procedimentos de autenticação ou a suscetibilidade a ataques de *phishing* e engenharia social, criando um vetor de ataque que compromete diretamente a confidencialidade, a integridade e a disponibilidade dos dados organizacionais.

2.2.1 Tipos de *Insider Threats*

Segundo a Fortinet (Fortinet Inc., 2025), as ameaças internas podem ser classificadas em diferentes categorias:

- **Intencionais (Maliciosas)**: indivíduos que deliberadamente visam causar dano à organização, seja por ganhos financeiros, retaliação ou espionagem corporativa.
- **Não intencionais (Acidentais)**: originadas de ações descuidadas ou erros humanos, como envio de dados para destinatários incorretos, cliques em links maliciosos ou negligência em políticas de segurança.
- **Ameaças de terceiros**: parceiros, prestadores de serviços ou contratados que, por ação maliciosa ou descuido, comprometem a segurança da organização.
- **Ameaças colusivas**: quando *insiders* se associam a agentes externos para roubar propriedade intelectual ou causar dano em troca de benefício financeiro.

Além disso, perfis comportamentais são frequentemente utilizados para caracterizar agentes internos:

- ***Pawns* (Peões)**: funcionários manipulados por agentes externos, muitas vezes sem consciência, via técnicas como *phishing* ou engenharia social.
- ***Turncloaks* (Traidores)**: empregados que voluntariamente prejudicam a organização, incluindo também denunciantes (*whistleblowers*).
- ***Collaborators* (Colaboradores)**: conspiradores que cooperam com criminosos para exfiltrar dados ou sabotar recursos.
- ***Goofs* (Atrapalhados)**: usuários que negligenciam políticas de segurança por comodidade ou incompetência, abrindo portas para riscos.
- ***Lone Wolf* (Lobo Solitário)**: indivíduos que agem sozinhos explorando vulnerabilidades, frequentemente com privilégios administrativos elevados.

Ainda de acordo com Fortinet Inc. (2025), essas categorias evidenciam que as ameaças internas não se limitam a agentes maliciosos, negligência e comprometimentos externos também representam riscos significativos. Nesse sentido, a diferenciação entre perfis manipulados (*pawns*) e traidores deliberados (*turncloaks*) é crucial para estratégias de detecção e resposta.

Neste contexto, a crescente sofisticação deste tipo de ameaça reforça a necessidade de ferramentas eficientes para detecção e mitigação de ameaças internas. Diferentemente

das ameaças externas, que muitas vezes podem ser contidas por barreiras perimetrais tradicionais, os atacantes exploram credenciais legítimas e possuem conhecimento prévio da infraestrutura, o que aumenta significativamente o potencial de dano. Nesse contexto, soluções alinhadas aos princípios da ZTA e sua aplicação prática, o ZTNA, tornam-se fundamentais, uma vez que garantem autenticação contínua, verificação contextual e segmentação granular dos acessos (Fortinet Inc., 2025).

Uma direção promissora para evolução da área consiste na integração de abordagens baseadas em ZTA com técnicas avançadas de análise, como aprendizado de máquina e modelos de linguagem (LLMs). Essa combinação tem potencial para ampliar a detecção de padrões anômalos em tempo real e reduzir a superfície de ataque em cenários corporativos. Além disso, pode possibilitar respostas mais rápidas e contextuais, cobrindo tanto acessos externos quanto ameaças internas. Nesse sentido, o avanço em direção a soluções que conciliem princípios de confiança zero e métodos inteligentes de monitoramento pode representar um passo essencial para enfrentar de forma mais eficaz os riscos associados a usuários mal-intencionados ou negligentes dentro das organizações.

2.3 *Dataset CERT - Insider Threat Test Dataset*

O conceito de *dataset* para Soares et al. (2021) é apresentado como um conjunto estruturado de dados utilizado para conduzir experimentos e análises, com determinados propósitos. O *dataset* reúne registros coletados de fontes relevantes ao estudo proposto e é composto por atributos que descrevem, de forma quantitativa ou qualitativa, as variáveis de interesse.

O *CERT Insider Threat Dataset* Lindauer (2020), desenvolvido pela Carnegie Mellon University, constitui-se em uma das principais bases de dados sintéticos para o estudo de comportamentos associados a ameaças internas. Esse *dataset* foi instituído com o objetivo de suprir a escassez de dados reais em virtude de questões legais, de sigilo e de privacidade. O *dataset* produzido foi amplamente compartilhado com a comunidade de pesquisadores participantes do programa DARPA ADAMS, cujo objetivo foi o desenvolvimento de técnicas avançadas para detecção de ameaças internas. Essa disponibilização possibilitou maior colaboração científica e a validação das abordagens propostas em um ambiente de pesquisa diversificado e multidisciplinar.

Ainda de acordo com Glasser e Lindauer (2013), é ressaltado a importância de assegurar realismo na geração de dados sintéticos para detecção de *insider threats*. O CERT tentou conciliar fidelidade e viabilidade prática, de forma a permitir pesquisas robustas sem comprometer informações sensíveis de organizações reais. Para alcançar esse equilíbrio, os autores utilizaram modelos que refletem elementos típicos de ambientes corporativos, tais como estruturas sociais, interações comunicacionais, acesso a recursos

computacionais e padrões de comportamento individual. O realismo é obtido pela integração de diferentes modelos, incluindo redes sociais organizacionais, perfis psicométricos e simulações de cenários de ameaça, com o intuito de reproduzir a diversidade de comportamentos normais e maliciosos.

O CERT possui 87,23GB de tamanho, os arquivos estão seraparados em 14 partes onde cada parte possui fragmentos de informações que refletem aspectos típicos de um contexto corporativo, tais como: (a) *Logs* de autenticação: registros de logon e logoff; (b) Atividades de e-mail: envio e recebimento de mensagens; (c) Navegação web; (d) Controle de dispositivos físico; (e) Manipulação de arquivos em diretórios compartilhados e, (f) dados psicométicos. Cada uma dessas categorias foi registrada de maneira sistemática, permitindo análises tanto em nível individual quanto coletivo. Além disso, os dados foram enriquecidos com **cenários de ataque previamente modelados**, como roubo de dados, fraude e sabotagem, o que proporciona condições controladas para validação de técnicas de detecção (LINDAUER, 2020).

O *Dataset* disponibilizado pelo CERT é organizado em diferentes arquivos, cada um representando um tipo específico de atividade monitorada dentro do ambiente corporativo (DINARDO; LEMOUDDEN; AHMAD, 2023). Os registros estão divididos em cinco principais categorias conforme ilustrado na Tabela 1.

Tabela 1 – Volumes de registros do dataset CERT - Tabela adaptada de (DINARDO; LEMOUDDEN; AHMAD, 2023).

Nome do arquivo	**Número de campos**	**Entradas de log**
device.csv	6	1.551.828
email.csv	12	10.994.957
file.csv	9	2.014.883
http.csv	7	117.025.216
logon.csv	5	3.530.285
Total	-	135.117.169

Embora o *dataset* CERT seja composto por dados sintéticos, o mesmo foi cuidadosamente construído para reproduzir com qualidade cenários realistas de comportamento em ambientes corporativos. Essa característica o torna uma referência amplamente utilizada na comunidade científica para o treinamento, validação e comparação de soluções voltadas à detecção de ameaças internas. Sua ampla aceitação se alinha com a capacidade de oferecer dados consistentes, organizados e de fácil acesso, permitindo a criação de experimentos reprodutíveis, principalmente no contexto de aprendizado de máquina e construção de modelos de inteligência artificial.

2.4 Inteligência Artificial e Aprendizado de Máquina

No contexto deste trabalho, as ameaças internas representam um grande desafio para a segurança da informação, uma vez que envolvem usuários com acesso legítimo aos sistemas organizacionais. Diversos trabalhos exploram esse conjunto de dados na aplicação de técnicas de Inteligência Artificial e Aprendizado de Máquina, buscando identificar padrões anômalos e caracterizar comportamentos suspeitos de forma automatizada e escalável. Em decorrência do sucesso dos Modelos de Linguagem de Grande Escala (LLMs), a utilização desses modelos como suporte a tomada de decisão se apresenta como um abordagem inovadora.

2.4.1 Modelos de Linguagem de Grande Escala (LLMs) para Análise de Risco

A incorporação de Modelos de Linguagem de Grande Escala (LLMs) no auxílio de detecção de ameaças tem-se mostrado uma estratégia promissora para ampliar a eficácia na detecção e mitigação de ameaças internas em ambientes corporativos (KETHIREDDY, 2022; SONG et al., 2024; LI et al., 2025).

No cenário de suporte ao processo de detecção, a contribuição das LLMs está na capacidade de interpretar dados complexos e heterogêneos, como por exemplo registros de log, e-mails, chats e atividades de sistema, de forma mais contextual e semântica. Torna-se possível evidenciar que os LLMs reduzem falsos positivos ao correlacionar padrões comportamentais sutis (KETHIREDDY, 2022). Observa-se que a colaboração entre agentes baseados em LLMs (decomposição de tarefas, geração de ferramentas e debates baseados em evidências) aumenta a confiabilidade das análises, indo além de classificações estáticas (SONG et al., 2024).

No contexto deste trabalho, foi desenvolvido um novo *dataset* derivado do *CERT Insider Threat Test Dataset*, adaptado para cenários de *Zero Trust Architecture* (ZTA). Esse *dataset* inclui um conjunto ampliado de *features*, construídas a partir da observação de comportamentos típicos em ambientes ZTA, como o monitoramento contínuo de *logins* e *logouts*, acessos condicionados a políticas dinâmicas de confiança, e correlação entre múltiplos vetores de atividade do usuário. Para avaliar essas novas características, foram empregados diferentes LLMs de última geração, incluindo Grok, GPT, Gemini, Copilot e DeepSeek, não apenas para interpretar e correlacionar os padrões extraídos dos dados, mas também para auxiliar na validação de cenários de risco, oferecendo perspectivas complementares.

2.5 Aprendizado de Máquina na Detecção de Ameaças Internas

A aplicação de técnicas de aprendizado de máquina (ML) tem se mostrado essencial para enfrentar a complexidade crescente associada à detecção de ameaças internas. Conforme destacado em (MANOHARAN et al., 2024; RAVAL; GANDHI; CHAUDHARY, 2018; SARHAN; ALTWAIJRY, 2022), os mecanismos tradicionais baseados em regras apresentam limitações significativas para lidar com o grande volume de dados presentes em ambientes corporativos modernos. Nesse contexto, os modelos de ML oferecem maior adaptabilidade ao identificar padrões ocultos e anomalias em comportamentos aparentemente legítimos, favorecendo a detecção precoce de atividades maliciosas.

No contexto de análise e detecção de ameaças, soluções baseadas em ML permitem automatizar o processo de análise de *logs* e eventos, reduzindo a dependência exclusiva de especialistas humanos e aumentando a escalabilidade dos sistemas de monitoramento (SARHAN; ALTWAIJRY, 2022). Algoritmos como *Random Forest*, redes neurais e métodos baseados em aprendizado profundo demonstraram eficácia em diferentes cenários de *insider threats*, fornecendo não apenas detecção, mas também apoio ao processo de decisão (MANOHARAN et al., 2024; RAVAL; GANDHI; CHAUDHARY, 2018).

Ante o exposto, não obstante a automatização do processo de detecção de ameaças internas, soluções baseadas em ML garantem ganho de precisão e eficiência operacional. Essa abordagem complementa métodos tradicionais e representa um passo fundamental para fortalecer a resiliência organizacional contra riscos provenientes de usuários mal-intencionados ou negligentes.

No contexto de ZTA, essa capacidade ganha especial relevância, uma vez que a segurança não é mais dependente de perímetros definidos, mas de uma avaliação contínua de risco em múltiplos pontos de acesso. A incorporação de ML nesse cenário permite analisar padrões de autenticação, movimentação lateral e uso de ativos em tempo real, reforçando a resiliência contra agentes maliciosos ou negligentes dentro da organização.

2.6 Trabalhos Correlatos

No contexto de segurança da informação, a indisponibilidade de dados reais rotulados sobre comportamentos internos maliciosos ou negligentes impõe um desafio significativo ao desenvolvimento e à avaliação de técnicas de detecção de ameaças internas (*insider threats*). O fato é que essas informações estão sujeitas a acesso restrito em grandes corporações, devido à sua elevada criticidade e ao potencial impacto em caso de exposição.

Em resposta a essa lacuna, diversos trabalhos têm proposto a construção de *datasets* específicos, seja por meio de geração sintética, experimentos controlados ou competições gamificadas, visando reproduzir, com o maior realismo possível, cenários que

combinem atividades legítimas e ações maliciosas de usuários.

Neste capítulo, são apresentados estudos que adotam metodologias para a instituição de *datasets* voltadas ao monitoramento, análise e modelagem de ameaças internas, servindo como referência para iniciativas semelhantes, como a proposta deste trabalho. Torna-se importante ressaltar que, o estudo de ameaças internas em ambientes baseados em *Zero Trust Architecture (ZTA)* é relativamente recente, com a maior parte das pesquisas concentrando-se em mecanismos de controle de acesso, segmentação de rede e autenticação contínua, havendo, portanto, espaço para o desenvolvimento de *datasets* específicos que atendam a esse contexto.

Em Camina et al. (2014), os autores apresentam o desenvolvimento de um *dataset* voltado à detecção de ataques de intrusão interna, especificamente projetado para capturar tentativas de intrusão de forma mais realista do que *datasets* clássicos, como o RUU (SALEM; STOLFO, 2011). A proposta baseia-se na premissa de que não são as ações em si que distinguem o comportamento malicioso do legítimo, mas sim o objeto sobre o qual essas ações são realizadas. Dessa forma, a navegação pelo sistema de arquivos reflete, de maneira mais abstrata e rica, os padrões de comportamento do usuário, podendo revelar intrusões com maior eficácia.

Para a construção do *dataset*, os autores coletaram dados em sistemas operacionais Windows (versões XP, 7, 8 e 8.1) utilizando uma ferramenta interna de auditoria do sistema de arquivos, registrando interações típicas como *open*, *read* e *write*. Esses registros incluem dados de aproximadamente 20 usuários reais, enquanto os acessos maliciosos foram simulados por meio de *scripts* que reproduziam o comportamento de intrusos.

Como estudo exploratório, foram treinados classificadores baseados em *Support Vector Machines* (SVM) e *k-Nearest Neighbors* (KNN). Embora a avaliação de desempenho não fosse o foco principal do trabalho, ela foi conduzida para demonstrar a viabilidade e o potencial da proposta.

Já em Harilal et al. (2018), os autores apresentam o desenvolvimento do *dataset The Wolf of SUTD* (TWOS), criado a partir de uma competição gamificada com o objetivo de obter instâncias realistas de ameaças internas maliciosas, contemplando os perfis de *masqueraders* e *traitors*. A proposta foi simular interações de usuários em um ambiente corporativo fictício, no qual comportamentos normais e maliciosos eram incentivados por meio de tarefas e condições específicas, como períodos de acesso indevido a contas de outros times e realocações forçadas de membros entre equipes, fomentando cenários de sabotagem e exfiltração de dados.

Para a construção do *dataset* TWOS, o experimento contou com 24 participantes distribuídos em 6 equipes, competindo durante 5 dias em um ambiente controlado hospedado em *cloud* (AWS), com instâncias Windows gerenciadas por *Domain Controller* e

instrumentadas por múltiplos agentes de coleta. Foram registradas atividades de diversas fontes heterogêneas, incluindo *keystrokes*, movimentos e cliques de mouse, monitoramento de processos e sistema de arquivos, tráfego de rede HTTP interceptado por *proxy*, e-mails (corpo e metadados), eventos de *login/logout* e questionários psicológicos. Técnicas de anonimização e extração de *features* foram aplicadas para preservar a privacidade e viabilizar análises comportamentais.

O *dataset* TWOS resultante compreendeu aproximadamente 320 horas de participação ativa, incluindo 18 horas de dados de *masquerader* e pelo menos duas instâncias de *traitor*. Embora a avaliação de algoritmos não fosse o foco principal do estudo, os autores discutem o potencial de uso do TWOS em múltiplas áreas da segurança da informação, como detecção de ameaças internas, verificação de autoria, autenticação contínua e análise de sentimentos, servindo como base para futuras pesquisas e desenvolvimento de mecanismos de detecção mais robustos.

Em Schonlau et al. (2001), os autores apresentam a criação do *dataset* conhecido como *Masquerading User Data*, instituído para avaliar métodos de detecção de intrusão por *masquerade*. O conjunto de dados contém sequências de comandos Unix geradas por 50 usuários, cada uma com 15000 comandos. Os primeiros 5000 comandos atuam como dados de treinamento, livres de *masqueraders*, enquanto os 10000 comandos seguintes são estruturados em 100 blocos de 100 comandos, nos quais inserções de comandos de outros usuários ocorrem com probabilidade controlada, resultando em aproximadamente 5% de blocos contaminados por intrusões. Para apoiar a avaliação de detectores, também são disponibilizados arquivos com a localização exata dos blocos contaminados e pontuações e limiares de diferentes algoritmos de detecção testados no estudo original. Os perfis de usuário foram gerados a partir dos dados de treinamento, e cada bloco de teste foi avaliado com base em uma pontuação, disparando um alerta quando ultrapassava um limiar predeterminado.

Em Legg et al. (2013), os autores abordam o problema crescente das ameaças internas em organizações, que envolvem indivíduos com acesso legítimo a dados e sistemas, mas que utilizam esse privilégio para fins maliciosos. Os autores argumentam que soluções tradicionais, como sistemas de detecção de intrusão (IDS), são insuficientes nesse contexto, e defendem a necessidade de modelos conceituais mais abrangentes que combinem aspectos tecnológicos, comportamentais e psicológicos. Neste contexto, os autores apresentam uma proposta conceitual para a detecção de ameaças internas. Para isso, os autores desenvolvem um modelo conceitual em três camadas (Mundo Real, Medições e Hipóteses) e estruturado em raciocínio baseado em hipóteses, no qual eventos observados são associados a padrões de comportamento por meio de árvores lógicas e probabilísticas. Diferente dos trabalhos citados, que são focados na construção de datasets, esta proposta é de natureza teórica e arquitetural, enfatizando a integração de múltiplas fontes de dados,

sendo eles: (a) técnicos, (b) comportamentais, (c) físicos e, (d) psicológicos, para oferecer uma visão holística da ameaça interna.

Em Le, Zincir-Heywood e Heywood (2020), foi proposto a instituição de um conjunto padronizado de *features* para apoiar pesquisas em detecção de ameaças internas, buscando superar a fragmentação existente entre diferentes trabalhos que utilizam representações de dados muito heterogêneas. Neste trabalho, os autores apresentaram um repositório contendo um *script* para extração de *features* a partir do *CERT Insider Threat Test Dataset*. A ferramenta foi desenvolvida em Python 3.8 e validada em ambiente Linux, possibilita a geração de diferentes níveis de granularidade temporal, como semana, dia, sessão e subsessão, além de permitir a configuração do número de núcleos de processamento utilizados em paralelo, otimizando o desempenho da execução. Os dados extraídos são organizados em diretórios específicos e incluem rótulos como `insider` (onde 0 indica atividade normal), bem como outras colunas relevantes, tais como `subs_ind`, `starttime`, `endtime`, `sessionid`, `user`, `day` e `week`. Dessa forma, o estudo caracteriza-se como uma proposta de `transformação de dataset`, em que a partir de uma base já existente é possível derivar novas representações enriquecidas e contextualizadas, ampliando o potencial de aplicação em pesquisas sobre detecção de ameaças internas.

2.6.1 Síntese dos Trabalhos Correlatos

Diante dos trabalhos correlatos apresentados, esta proposta de estudo avança em duas direções principais: (a) ao contrário de iniciativas que apenas transformam ou derivam representações do CERT, este trabalho institui um novo dataset de ameaças internas especificamente orientado ao contexto de *Zero Trust Architecture* (ZTA), incorporando informações de *logins*, *logoffs*, acessos a conteúdo e características de dispositivos, de forma a refletir cenários mais alinhados às práticas contemporâneas de segurança e, (b) enquanto a maioria das pesquisas se limita a avaliações técnicas ou conceituais, este trabalho introduziu um processo de avaliação de risco híbrido, combinando a análise de um especialista humano com a inferência de cinco LLM, sendo elas: *GROK, GPT, DeepSeek, Gemini e Copylot*. Essa integração proporciona uma visão mais robusta e multifacetada do risco associado a cada comportamento, abrindo caminho para métodos de detecção que aliam precisão técnica, consistência contextual e apoio de inteligência artificial generativa.

A Tabela 2 apresenta uma síntese comparativa entre os principais trabalhos relacionados e a proposta deste estudo. Observa-se que iniciativas clássicas, como Schonlau et al. (2001) e Camina et al. (2014), concentram-se em dados sintéticos ou cenários restritos, carecendo de diversidade e de aplicabilidade a ambientes contemporâneos de segurança, tais como ZTA.

Já Harilal et al. (2018), por meio da competição TWOS, introduz uma abordagem gamificada, mas que envolve elevado custo experimental e baixa replicabilidade em larga

escala. Por sua vez, Legg et al. (2013) apresenta apenas um modelo conceitual, sem oferecer um *dataset* real para validação prática.

No caso de Le, Zincir-Heywood e Heywood (2020), identifica-se a adaptação do *CERT Insider Threat Dataset*, porém a contribuição limita-se à transformação de dados já existentes, sem a introdução de novos cenários de avaliação.

Tabela 2 – Comparação entre trabalhos correlatos e o presente estudo

Trabalho	**Sintéticos/Reais**	**Ambiente ZTA**	**Validação com LLM**	**Limitações**
Camina et al. (2014)	Reais/sintéticos (20 usuários + ataques simulados)			Restrito a interações no sistema de arquivos (Windows); amostra pequena
Harilal et al. (2018) (TWOS)	Reais/sintéticos (competição gamificada em ambiente AWS)			Alto custo/complexidade; replicabilidade limitada; cenário controlado
Schonlau et al. (2001)	Sintéticos (sequências de comandos Unix)			Dataset antigo; baixa representatividade de cenários atuais
Legg et al. (2013)	Conceitual (sem experimentos práticos)			Modelo teórico; não provê dataset; não específico para ZTA
Le, Zincir-Heywood e Heywood (2020)	Sintéticos (derivados do CERT via extração de *features*)			Transforma o CERT; não cria novos cenários; sem foco em ZTA
Presente Estudo	Reais/sintéticos (derivado do CERT: logins, páginas, dispositivos)	✓	✓	Escopo inicial baseado no CERT; futuras expansões para outras fontes

Ante o exposto, a proposta deste trabalho busca uma adaptação mais robusta e direcionada a ambientes de *Zero Trust Architecture (ZTA)*, contemplando *logs* de autenticação, acessos a páginas e conexões de dispositivos. Além disso, inova ao introduzir um processo de validação com múltiplos modelos de linguagem (*Large Language Models* – LLMs) e a avaliação de um especialista, ampliando a perspectiva analítica e de confiabilidade dos resultados. Embora o escopo inicial estar restrito ao CERT, o estudo aponta caminhos promissores para expansões futuras, incluindo a incorporação de outras fontes de dados e cenários de aplicação.

3 Metodologia

Este capítulo tem como finalidade apresentar em detalhes o método de pesquisa adotado neste trabalho, destacando como principais contribuições: (a) a descrição sistemática das etapas conduzidas na construção da proposta; (b) a delimitação precisa do escopo do estudo, estabelecendo os limites e direcionamentos da investigação e, (c) validação do presente estudo.

O método praticado neste trabalho encontra-se dividido em três macros seções, são elas: (a) Percepção da Problemática, (b) Desenvolvimento e, (c) Validação. A Figura 2 ilustra o método deste trabalho.

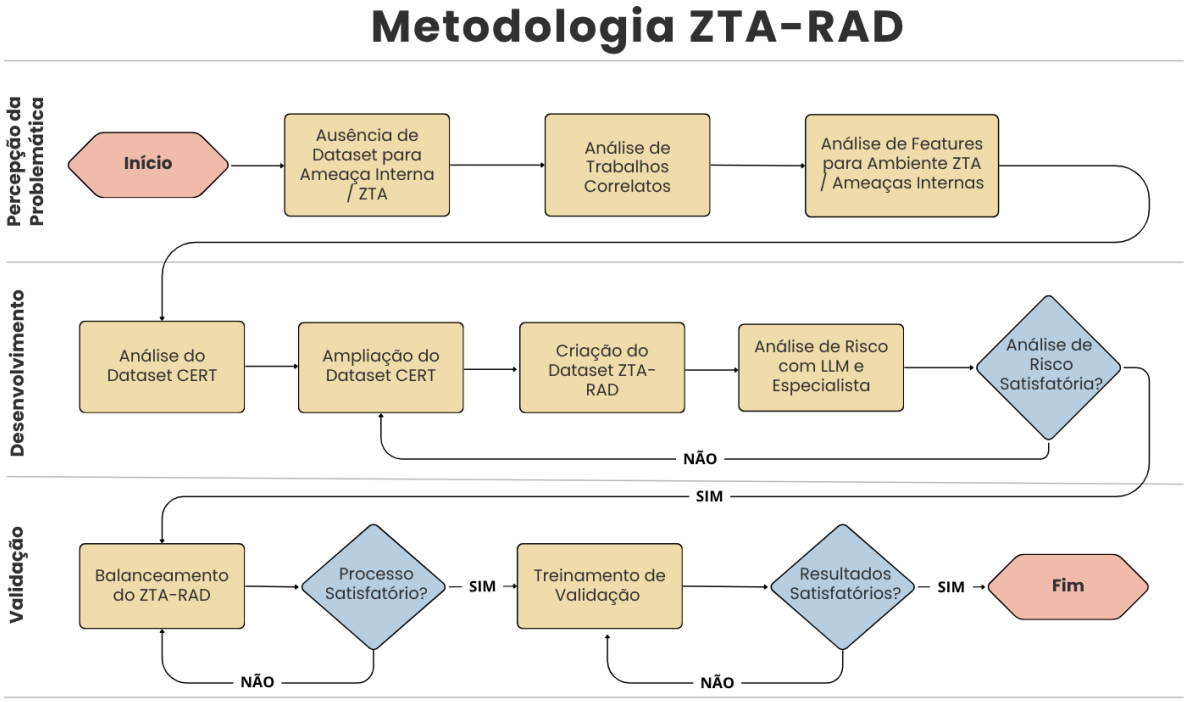

Figura 2 – Fluxograma do Método de Pesquisa - Fonte: Do Autor

3.1 Percepção da Problemática

A percepção da problemática se deu através da lacuna significativa observada no que se refere a segurança de ameaças internas em ambientes concebidos sob os princípios da Arquitetura de Confiança Zero (ZTA). Diferentemente dos modelos tradicionais de segurança baseados em perímetro, a ZTA pressupõe que nenhuma conexão deve ser implicitamente confiável, exigindo monitoramento contínuo e validação em múltiplos pontos

de controle. Essa mudança de paradigma altera de forma profunda os comportamentos a serem observados, bem como as variáveis relevantes para a análise de risco.

Nos ambientes perimetrais clássicos, a detecção de ameaças é geralmente orientada por registros de autenticação (tradicional VPN), tráfego de rede e acesso a arquivos. Em contraste, a ZTA introduz novos elementos de monitoramento, como autenticação e revalidação constantes, controle de dispositivos habilitados por usuário, microsegmentação da rede e restrição de acesso a sistemas e páginas previamente autorizados. Tais características ampliam a granularidade do monitoramento, mas também aumentam a complexidade de identificar padrões suspeitos, uma vez que as atividades legítimas dos usuários se tornam muito mais dinâmicas e multifacetadas.

Nesse sentido, a ausência de um *dataset* de ameaças internas específico para ambientes ZTA representa um obstáculo relevante tanto para a pesquisa acadêmica quanto para o desenvolvimento de soluções práticas de segurança. Os *datasets* de ameaças internas atualmente disponíveis, apesar de úteis, refletem cenários baseados em perímetro e não incorporam as particularidades de ambientes com autenticação contínua e segmentação avançada. Isso limita a capacidade de treinar e validar modelos de aprendizado de máquina para detecção em condições realistas.

Diante do exposto, foi realizada uma revisão da literatura da área com o objetivo de compreender a real necessidade de um *dataset* voltado a ambientes baseados em ZTA. Identificou-se que a construção de um conjunto de dados específico para esse contexto é fundamental, contemplando variáveis derivadas de práticas características da arquitetura, tais como registros detalhados de login e logoff, dispositivos vinculados a identidades, tentativas de acesso a recursos não autorizados e eventos de movimentação lateral restringidos pela microsegmentação.

A criação de um *dataset* com essas particularidades não apenas contribuirá para o avanço das técnicas de detecção de ameaças internas, mas também possibilitará explorar de forma mais efetiva o potencial de integração entre aprendizado de máquina, ampliando a capacidade de análise, validação e tomada de decisão em ambientes de segurança cada vez mais dinâmicos e distribuídos.

3.2 Desenvolvimento

O desenvolvimento do *dataset* ZTA-RAD teve início a partir de uma análise aprofundada do *dataset* CERT, que se consolidou ao longo dos anos como uma das principais referências para estudos de detecção de ameaças internas. No entanto, por se tratar de um *dataset* concebido em um contexto tradicional de segurança baseado em perímetro, suas variáveis não contemplam as particularidades de arquiteturas fundamentadas no paradigma ZTA. Diante dessa lacuna, esta etapa do método foi responsável pela ampliação

do CERT, incorporando novas *features* orientadas ao cenário de ZTA, caracterizado por princípios como monitoramento contínuo, autenticação e autorização granulares, micro-segmentação da rede e gestão rigorosa de identidades e dispositivos. Após este processo, tornou-se necessário realizar uma análise de risco do novo *dataset* com o objetivo de identificar quais instâncias seriam classificadas como maliciosas e benignas.

3.2.1 Ampliação do *Dataset* CERT

Nesta etapa, a experiência profissional da autora, com mais de cinco anos de atuação na área de segurança cibernética, desempenhou papel central, fornecendo subsídios técnicos e práticos para a seleção e a definição das variáveis mais representativas de um ambiente ZTA. As novas *features* foram geradas tanto a partir da combinação de atributos já existentes no CERT quanto pelo desmembramento de variáveis originais em dimensões mais detalhadas, capazes de refletir comportamentos específicos esperados em um modelo de confiança zero. Essa abordagem garantiu a consistência metodológica e preservou a qualidade dos dados, ao mesmo tempo em que ampliou a capacidade do *dataset* de apoiar pesquisas sobre detecção de ameaças internas em ambientes mais próximos da realidade atual das organizações.

Um exemplo representativo de combinação e desmembramento de *features* pode ser observado nas informações de autenticação. No *dataset* CERT, o registro de login é tratado como um único atributo, contendo apenas a data e o horário completos do evento. Entretanto, em um ambiente orientado à ZTA, tal representação mostra-se insuficiente. Nesse contexto, torna-se essencial derivar novas *features* que capturem dimensões adicionais, como a verificação de permissões para o usuário realizar login em determinado dia ou janela de horário específica, a conformidade do evento com o horário comercial atribuído ao colaborador e a existência de mecanismos de janela de acesso, em que *logins* fora do expediente podem ser permitidos mediante solicitação prévia. Além disso, também é relevante correlacionar o volume de *logons* e *logoffs* ao longo do dia, permitindo identificar potenciais anomalias ligadas a sessões não finalizadas ou acessos inconsistentes.

3.2.2 Análise de Risco com LLM e Especialista

Para a análise de risco do *dataset* ZTA-RAD, além da contribuição direta da especialista da área (autora do trabalho), considerou-se o potencial das LLMs (*Large Language Models*) como recurso complementar nesse processo. Essas ferramentas ofereceram uma perspectiva mais robusta e multifacetada na avaliação do risco associado a cada comportamento, permitindo explorar não apenas a precisão técnica, mas também a consistência contextual e a capacidade de inferência de padrões complexos. Tal abordagem abre caminho para métodos de detecção mais sofisticados, que aliam inteligência artificial generativa a práticas consolidadas de segurança cibernética.

Nesta etapa, foram utilizadas diferentes LLMs de destaque no cenário atual: (a) **GROK**, (b) **ChatGPT**, (c) **Gemini**, (d) **Copilot** e (e) **DeepSeek**. Cada uma dessas ferramentas foi aplicada com suas versões específicas, conforme resumido na Tabela 3. Torna-se importante ressaltar que, todas as versão utilizadas são gratuitas.

Tabela 3 – Modelos de Linguagem Utilizados na Análise de Risco e suas Versões (2025)

Modelo	**Versão Utilizada**
GROK	Grok 4
ChatGPT	GPT-3
Gemini	Gemini 2.5 Pro
Copilot	Microsoft Copilot (jun/2025)
DeepSeek	DeepSeek V3

Esta análise de risco teve como objetivo classificar cada instância do *dataset* ZTA-RAD em três categorias: (a)baixo, (b)médio ou (c)alto risco, de forma análoga ao processo de avaliação de ameaças internas em ambientes baseados em ZTA. Essa classificação permite identificar comportamentos potencialmente suspeitos em diferentes níveis, refletindo a criticidade das ações do usuário frente às políticas de segurança.

Cabe destacar que cada LLM utilizado foi empregado exclusivamente na etapa de análise de risco, produzindo, um *gabarito* de classificação para o *dataset*. Todas as ferramentas foram orientadas pelo mesmo conjunto de dados e *prompt*, garantindo uniformidade no processo avaliativo e permitindo a comparação sistemática entre os modelos.

3.3 Validação

Para o processo de validação do *dataset*, além da checagem de integridade, garantindo a ausência de valores faltantes, duplicados ou inconsistentes, foi realizada uma verificação semântica, de modo a assegurar que os dados produzidos respeitam as regras inerentes a um ambiente ZTA. Em seguida, aplicou-se o balanceamento das classes por meio da técnica SMOTE (*Synthetic Minority Over-sampling Technique*), visando reduzir possíveis vieses de aprendizado. Por fim, foram conduzidos experimentos com dois classificadores distintos, Rede Neural e *Random Forest*, a fim de avaliar a capacidade do *dataset* em sustentar cenários de aprendizado de máquina, confirmando sua aplicabilidade prática no contexto de detecção de ameaças internas.

3.3.1 Balanceamento do ZTA-RAD

O *Synthetic Minority Oversampling Technique* (SMOTE) é uma das técnicas mais consolidadas para lidar com o problema do desbalanceamento de classes em conjuntos de dados de aprendizado de máquina. Em cenários de detecção de ameaças internas, como

no caso do *dataset* ZTA-RAD, a ocorrência de instâncias relacionadas a comportamentos maliciosos (risco médio e alto) tende a ser significativamente menor quando comparada a registros de atividades legítimas (risco baixo), o que pode comprometer a capacidade dos modelos em aprender padrões relevantes da classe minoritária.

O SMOTE criou novas instâncias a partir das instâncias da classe minoritária. Dessa forma, evita-se a simples duplicação de registros já existentes e enriquece-se a representação da classe minoritária, contribuindo para reduzir vieses do modelo em favor da classe majoritária. O objetivo desta etapa foi produzir um *dataset* mais balanceado, favorecendo o treinamento dos classificadores.

3.3.2 Treinamento e Validação

No processo de treinamento, os classificadores Rede Neural e *Random Forest* foram aplicados ao *dataset* ZTA-RAD com o objetivo de avaliar sua eficácia na detecção de ameaças internas em ambientes ZTA.

Para mensurar o desempenho, foram utilizadas métricas: acurácia, para medir a proporção de instâncias corretamente classificadas; precisão, para avaliar a confiabilidade das predições positivas; revocação (ou sensibilidade), que quantifica a capacidade do modelo em identificar corretamente instâncias positivas; e a pontuação F1, que harmoniza precisão e revocação em um único indicador. Além disso, foram analisadas as curvas ROC e a AUC (Área sob a Curva), a fim de fornecer uma visão mais abrangente sobre a robustez dos modelos em cenários com diferentes limiares de decisão.

4 Construção do *Dataset* ZTA-RAD

Este capítulo apresenta os detalhes do processo de construção do *dataset* proposto. Os códigos utilizados, bem como as figuras geradas e o próprio *dataset*, encontram-se disponíveis no repositório indicado no Apêndice B (Artefatos do Trabalho).

É importante destacar que o ZTA-RAD foi desenvolvido a partir do *dataset* CERT, cujos atributos estão descritos na Tabela 1. Considerando o contexto de *Zero Trust Network Access* (ZTNA), o ZTA-RAD concentrou-se na análise dos registros provenientes dos arquivos `device.csv`, `http.csv` e `logon.csv`, por representarem de forma mais direta o comportamento de acesso a dispositivos, sistemas e autenticações. Por outro lado, os registros de `email.csv` e `file.csv` não foram contemplados, uma vez que não se enquadram na modelagem de acesso a recursos de rede prevista pela ZTNA.

Outro aspecto a ser ressaltado diz respeito ao volume de registros do *dataset* CERT, que ultrapassa 80 GB de dados. Em razão dessa magnitude, optou-se por trabalhar apenas com uma fração do conjunto original.

Para tanto, realizou-se inicialmente o download completo do *dataset*, seguido da seleção das informações mais relevantes ao escopo deste trabalho. Assim, foram extraídos os seguintes fragmentos: (a) `device.csv`, consolidado a partir da aglutinação dos fragmentos do *dataset* CERT, com tamanho de 58,2 MB; (b) `http.csv`, também resultante de aglutinação dos fragmentos, com 549,5 MB; e (c) `logon.csv`, com 58,5 MB.

Ressalta-se ainda que os arquivos de maior volume no *dataset* CERT são `http.csv`, `file.csv` e `email.csv`, em virtude da elevada quantidade de registros e do detalhamento das informações que armazenam.

Diante das ações realizadas, o dataset CERT utilizado como base neste trabalho passou a ter 666,2 MB, composto por 2000 registros (instâncias) de dispositivos, acessos e informações de logon. A Tabela 4 apresenta o conjunto de atributos previstos nos arquivos `http.csv`, `file.csv` e `email.csv`. Ressalta-se que cada um desses arquivos foi inicialmente tratado de forma individual e, somente após esse processo, unificado em um único dataset.

4.1 Manipulação e pré-processamento dos dados

Inicialmente, cada um dos arquivos correspondentes, *logon.csv*, *device.csv* e *http.csv*, foi carregado em um ambiente *Python*, utilizando-se bibliotecas adequadas para manipulação e análise de dados. Em seguida, foi realizado o pré-processamento, etapa fundamental para certificar a consistência e qualidade das informações.

Tabela 4 – Atributos de logons, dispositivos e acessos HTTP do *dataset* CERT - Tabela Adaptada de (LINDAUER, 2020)

Atributo	**Descrição**
`id`	Identificador único da sessão de log, gerado em formato alfanumérico.
`date`	Data e horário do evento de log, representando quando ocorreu a atividade.
`user`	Identificação do usuário associado ao evento de log.
`pc`	Identificação do computador ou dispositivo utilizado na sessão.
`activity`	Tipo de atividade registrada, como Logon ou Logoff.
`url`	Endereço eletrônico acessado pelo usuário durante a sessão HTTP.
`content`	Conteúdo ou categoria associada à URL acessada, indicando a natureza da página.

Nessa etapa, o campo de datas de cada conjunto foi convertido para o formato *datetime*, permitindo uma manipulação mais precisa e facilitando a aplicação de operações temporais. A partir dessa conversão, foram extraídos atributos adicionais, como o dia da semana e a hora do evento, possibilitando uma análise mais refinada dos registros.

A Figura 3 ilustra a distribuição das atividades de logon ao longo dos diferentes dias da semana, permitindo identificar a frequência e a regularidade dessas atividades. Nota-se maior concentração de logon em dias utéis, ou seja, na segunda, terça, quarta, quinta e sexta-feira. Essa análise possibilitou não apenas visualizar os dias de maior utilização, mas também começar a inferir padrões de acesso que ocorrem em dias atípicos ou fora dos períodos usualmente permitidos. Dessa forma, a análise fornece subsídios iniciais para a detecção de comportamentos anômalos, que podem indicar potenciais riscos associados ao uso indevido de credenciais. Convém destacar que a Figura 3 representa apenas uma das diversas ampliações conduzidas neste estudo, não englobando a totalidade das explorações realizadas a partir do conjunto de dados.

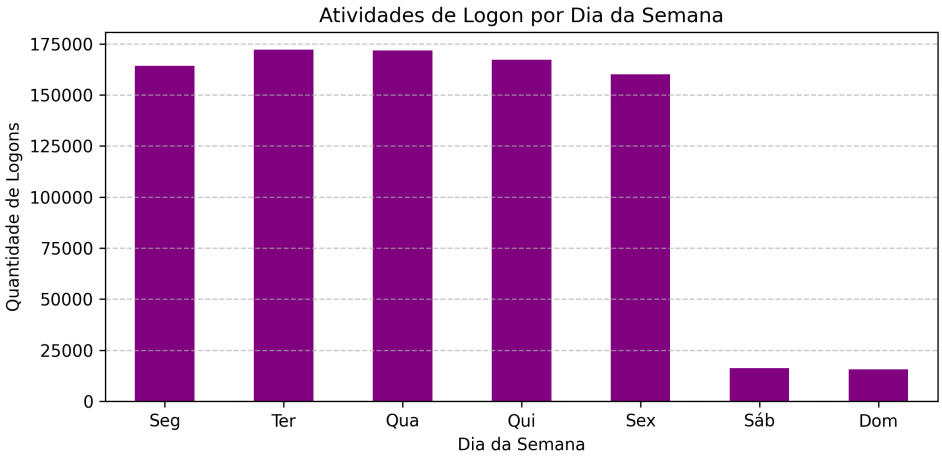

Figura 3 – Atividades de Logons Espalhados nos Dias da Semana - Fonte: Do Autor.

Para o arquivo *logon.csv*, foram obtidas métricas voltadas ao comportamento de autenticação considerando o número total de eventos de *logon* e *logoff*, a identificação dos horários mais recorrentes de acesso e desconexão, e a frequência de acessos em dias úteis e finais de semana. Além disso, foram distinguidos padrões de acesso dentro e fora do

horário comercial, permitindo avaliar desvios de comportamento em relação ao expediente regular.

Como parte dessa análise, definiu-se a chamada *"Janela de Acesso"*, correspondente ao intervalo entre 18h e 22h, utilizada como critério complementar para a detecção de comportamentos potencialmente anômalos. Essa métrica adicional visa capturar atividades atípicas fora do expediente regular, reforçando a capacidade de identificar padrões suspeitos e fornecer subsídios para a análise de risco.

Para o arquivo *device.csv* investigou-se as interações com dispositivo mapeando a variedade de computadores conectados, o total de eventos de conexão e desconexão, a diferença entre esses volumes para detectar acessos persistentes. Adicionalmente, foi analisada a utilização dos dispositivos em diferentes contextos temporais, distinguindo acessos realizados em horários comerciais, não comerciais e durante finais de semana, de modo a evidenciar potenciais desvios de uso legítimo.

A Figura 4 ilustra a quantidade de dispositivos que estabelecem conexões por hora ao longo do dia (0h–23h). Observa-se que a maior concentração de conexões de dispositivos ocorre no período comercial, entre 7h e 18h, com destaque para o intervalo entre 9h e 16h, quando se registra o pico de dispositivos conectados. Ao observar que o comportamento normal concentra-se no horário comercial, torna-se possível estabelecer uma referência que permite diferenciar o uso esperado de dispositivos, desta forma, eventuais desvios em relação a esse padrão, e.g., conexões frequentes fora do horário definido, podem ser caracterizados como anomalias, servindo, assim, como indícios relevantes de risco.

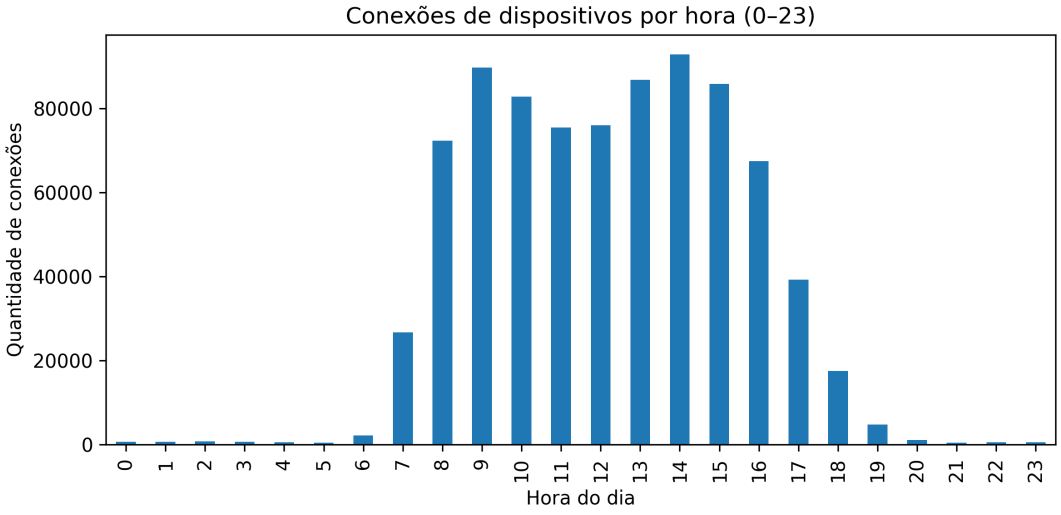

Figura 4 – Quantidade de Conexões de Dispositivos Distintos por Hora do dia - Fonte: Do Autor.

No arquivo *http.csv*, foram mensuradas métricas relacionadas à navegação na web, incluindo o número de sites distintos visitados, o volume total de acessos, os padrões de

navegação ao longo dos diferentes períodos do dia e a frequência de acessos em finais de semana. A Figura 5 ilustra a quantidade de acessos HTTP por hora do dia (0h–23h), evidenciando que a maior concentração de atividades ocorre entre 7h e 17h, com um pico significativo de acessos registrado por volta das 13h.

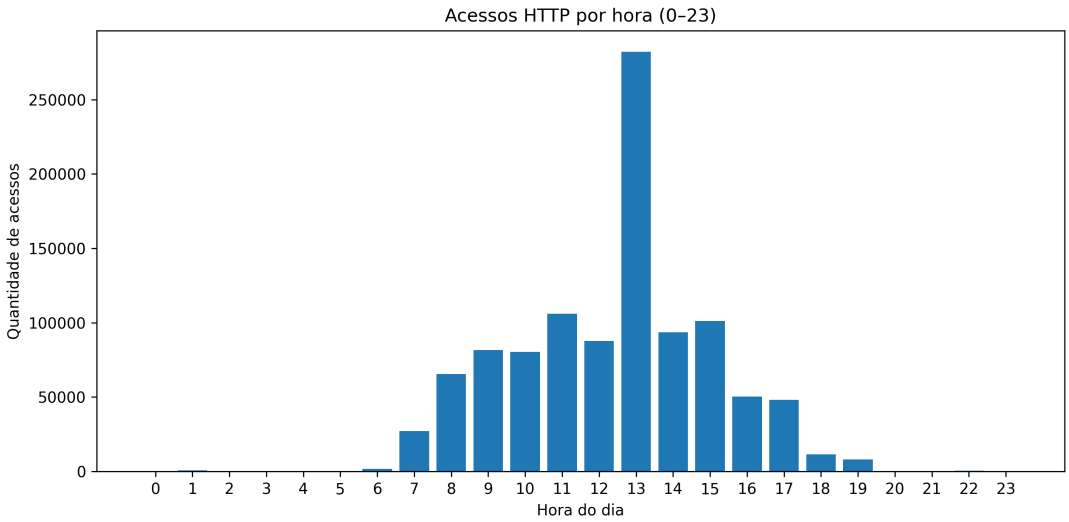

Figura 5 – Quantidade de Acessos HTTP Distintos por Hora do dia - Fonte: Do Autor.

Através da Figura 5 foi possível observar acessos que podem representar risco, já que acessos em horários incomuns podem ser tratados como anomalias e submetidos a verificações adicionais. Além disso, essa análise quando integradas a outras, como logons e conexões de dispositivos, fortalece a detecção de ameaças internas e subsidia a aplicação de respostas adaptativas proporcionais ao risco.

Também foram avaliadas visitas a categorias específicas de sites, como por exemplo: compras, notícias e entretenimento. Foram avaliadas o tamanho médio das URLs acessadas, a presença de termos sensíveis no conteúdo HTTP e a comparação entre acessos realizados dentro e fora do horário comercial.

Todas as análises apresentadas nesta seção mostraram-se essenciais para a definição e construção das características do *dataset* ZTA-RAD, uma vez que permitiram identificar padrões de comportamento e, a partir deles, derivar métricas capazes de evidenciar potenciais ameaças internas. Esse conjunto de características não apenas enriquece a qualidade do *dataset*, mas também fornece subsídios para a aplicação de técnicas de detecção de ameaças alinhadas aos princípios da ZTA.

4.2 Consolidação do *dataset* ZTA-RAD

Após as etapas de manipulação e ampliação de métricas específicas de cada fragmento, foi conduzido o processo de consolidação gerando um novo e único *dataset*, consti-

tuído dos arquivos apresentados acima. Essa etapa teve como objetivo integrar, em nível de usuário, as diferentes perspectivas de comportamento computacional como por exemplo: autenticação, uso de dispositivos e navegação web.

O processo consistiu na junção das tabelas derivadas de cada fragmento, utilizando como chave primária o `username` de identificação do usuário. Essa integração buscou preservar a granularidade das métricas extraídas e ao mesmo tempo garantir a estrutura consistente frente à heterogeneidade dos dados.

Este novo *dataset* contém 36 rótulos (*features*), que contemplam desde métricas de acesso, conteúdo de acesso, contagem de eventos e proporção de acessos em horários atípicos. Para facilitar a compreensão, essas métricas foram agrupadas em três categorias, a saber:

- **Métricas de Logon:** capturam padrões de autenticação, frequência de extitlogins em dias úteis/finais de semana, horários comuns de acesso e desvios em relação ao expediente comercial;
- **Métricas de Dispositivos:** descrevem a variedade e persistência de conexões, diferenciando atividades em contextos comerciais e não comerciais;
- **Métricas de Navegação HTTP:** representam o perfil de utilização de páginas web, incluindo diversidade de sites acessados, visitas a categorias específicas e presença de termos sensíveis.

Para melhor organização, a Tabela 5 apresenta a identificação dos atributos do *dataset* ZTA-RAD e uma breve descrição.

4.2.1 Normalização do ZAT-RAD

Uma etapa importante no processo de construção de *dataset* é a normalização, ou normatização dos dados. Em função do *dataset* ZTA-RAD ter sido construído a partir de uma ampliação e fragmentação e recombinação de dados (ex.: separar logon.csv, http.csv, device.csv e depois criar novas *features* derivadas). Esse processo gera um descompasso entre variáveis: algumas aparecem de forma abundante em determinados usuários, enquanto outras ficam ausentes ou com baixa granularidade.

Para contextualizar, no CERT, o login está registrado em um único campo com data/hora associada a um usuário. No ZTA-RAD, isso foi expandido em várias métricas: logons fora do horário comercial, logons em dias úteis, janela de acesso, proporção de PCs distintos, etc. Esse desmembramento cria colunas em escalas diferentes: algumas são contagens absolutas, outras proporções, outras apenas flags binárias, por exemplo, se um

Tabela 5 – Atributos derivados dos registros de logons, dispositivos e acessos HTTP - Atributos enriquecidos a partir do *dataset* CERT (LINDAUER, 2020).

Atributo	**Descrição**
`user`	Identificação única do usuário.
`computadores_distintos`	Quantidade de computadores diferentes acessados pelo usuário.
`total_logons`	Número total de eventos de autenticação (*logon*).
`total_logoffs`	Número total de encerramentos de sessão (*logoff*).
`computador_mais_usado`	Equipamento mais recorrente associado ao usuário.
`hora_logon_mais_comum`	Horário mais frequente de *logon*.
`hora_logoff_mais_comum`	Horário mais frequente de *logoff*.
`logons_dias_uteis`	Frequência de *logons* em dias úteis.
`logons_fim_de_semana`	Frequência de *logons* em fins de semana.
`logons_em_horario_comercial`	*Logons* realizados entre 08h e 18h.
`logons_fora_horario_comercial`	*Logons* fora do horário padrão (18h–08h).
`logons_janela_acesso`	*Logons* registrados entre 18h–23h (faixa de atenção).
`proporcao_pc_mais_usado`	Proporção de acessos feitos no computador mais utilizado.
`pcs_distintos_dispositivo`	Número de computadores com eventos de conexão de dispositivos.
`total_atividades_dispositivo`	Total de eventos de conexão e desconexão de dispositivos.
`total_connect_dispositivo`	Número de conexões de dispositivos.
`total_disconnect_dispositivo`	Número de desconexões de dispositivos.
`conexoes_sem_desconexao`	Diferença entre conexões e desconexões (persistência).
`connect_em_horario_comercial`	Conexões de dispositivos entre 08h–18h.
`connect_fora_horario_comercial`	Conexões de dispositivos fora do horário padrão.
`connect_dias_uteis`	Conexões em dias úteis.
`connect_fim_de_semana`	Conexões em fins de semana.
`connect_janela_acesso`	Conexões na janela 18h–23h.
`sites_distintos_visitados`	Quantidade de sites únicos visitados.
`total_acessos_http`	Volume total de requisições HTTP.
`acessos_em_horario_comercial`	Acessos HTTP entre 08h–18h.
`acessos_fora_horario_comercial`	Acessos HTTP fora do horário padrão.
`acessos_fim_de_semana`	Acessos HTTP em fins de semana.
`acessos_janela_acesso`	Acessos HTTP entre 18h–22h.
`visita_sites_compras`	Acessos a sites de compras.
`visita_sites_noticias`	Acessos a sites de notícias.
`visita_sites_entretenimento`	Acessos a sites de entretenimento.
`tamanho_medio_urls`	Comprimento médio das URLs acessadas.
`termos_sensiveis_detectados`	Ocorrências de termos sensíveis no conteúdo HTTP.
`tamanho_medio_conteúdo`	Tamanho médio do conteúdo retornado.
`diferenca_acessos_fora_dentro`	Diferença entre acessos fora e dentro do horário comercial.

usuário nunca acessou fora do horário, a coluna “logons_fora_horario_comercial” deve ser zerada e não deixada como valor nulo.

Ainda neste contexto, muitas *features* podem conter valores de ordens de grandeza muito diferentes (ex.: número de acessos varia de 1 a 10.000, enquanto hora do login varia de 0 a 24). Neste cenário precisa-se normalizar os dados em um intervalo fixo, padronizar formato de datas, horários,

Em geral, a normalização não foi apenas um ajuste, mas sim, uma etapa de garantia de coerência para que as novas *features* façam sentido no contexto ZTA.

4.3 Rotulagem de Risco do ZTA-RAD

Como uma das principais características de aprendizado supervisionado (aprendizado de máquina), torna-se importante o *dataset* estar totalmente rotulado, ou seja, os dados precisam ter entradas (*features*) e uma saída esperada (rótulo), por exemplo, sendo

uma entrada informações referentes aos acessos de um usuário, a saída esperada é ”acesso normal” ou ”acesso malicioso”.

Neste contexto, tornou-se necessário realizar um mapeamento entrada–saída, definido neste trabalho como análise de risco aplicada ao novo *dataset* ZTA-RAD. Esse processo foi conduzido em duas etapas complementares: (a) a análise realizada pela especialista, baseada em sua experiência profissional e no conhecimento do domínio, e (b) a utilização de Modelos de Linguagem de Grande Escala (LLMs) para subsidiar e enriquecer a análise de risco, fornecendo uma perspectiva adicional orientada por inteligência artificial. A Figura 6 ilustra o resultado deste processo, complementada pela explicação do procedimento nas subseções abaixo.

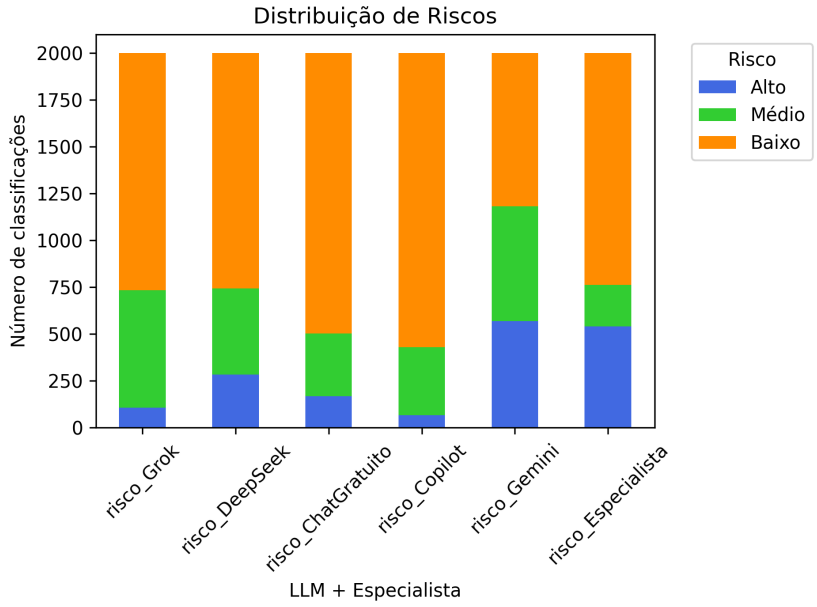

Figura 6 – Rotulagem da Análise de Risco do ZTA-RAD.

4.3.1 Rotulagem pelo Especialista

A rotulagem conduzida pela especialista desempenhou um papel central na validação do *dataset* ZTA-RAD. Baseando-se em sua experiência profissional na área de segurança cibernética, a especialista analisou criticamente as variáveis geradas, verificando sua aderência ao contexto de ambientes orientados pela Arquitetura de Confiança Zero (ZTA). Esse processo envolveu a identificação de padrões de comportamento usuais e atípicos de usuários, a coerência das regras de negócio aplicadas e a consistência entre os registros de diferentes fontes de *logs*. Dessa forma, a avaliação do especialista não apenas garantiu a integridade conceitual do *dataset*, mas também serviu como referência fundamental para a classificação de risco, fornecendo um ”padrão-our” para comparação com os resultados produzidos por técnicas de inteligência artificial. Para cada instância, foram definidos três níveis de risco: (a) baixo, (b) médio e (c) alto.

4.3.2 Rotulagem por LLMs

Com o intuito de assegurar a imparcialidade e ampliar a comparabilidade no processo de rotulagem de risco, além da análise conduzida pela especialista, foram utilizadas 5 diferentes LLMs (Grok, Chat-GPT, Gemini, Copilot e DeepSeek). Para tal, aplicou-se um mesmo *prompt* padronizado em todas as ferramentas, solicitando que cada modelo classificasse os riscos a partir dos atributos do *dataset* ZTA-RAD, tomando como referência os princípios fundamentais do paradigma de segurança da Arquitetura de Confiança Zero (ZTA). Esse procedimento permitiu obter múltiplas perspectivas automatizadas sobre os dados, enriquecendo o processo de rotulagem e fornecendo uma base complementar à avaliação humana.

O *prompt* foi estruturado em quatro seções: **(a)** contextualização teórica, **(b)** descrição dos dados, **(c)** direcionamento da tarefa e **(d)** definição das saídas esperadas. Essa organização teve como objetivo garantir clareza, padronização e reprodutibilidade na avaliação das LLMs, assegurando que as respostas fossem consistentes com os princípios da ZTA e comparáveis entre si. Os *prompts* utilizados encontram-se integralmente disponíveis no Apêndice A (Prompt Utilizado nas LLMs).

4.3.2.1 Contextualização teórica

O *prompt* inicia com uma fundamentação conceitual, apresentando uma breve introdução à ZTA. Nessa etapa, são descritos seus princípios fundamentais, como verificação contínua, privilégio mínimo, segmentação e monitoramento constante e destacada sua relevância no contexto da análise de ameaças internas *Insider Threats*.

Essa contextualização cumpre duas funções principais: (a) estabelecer um referencial teórico sólido que orienta a interpretação dos dados e, (b) assegurar que a LLM compreenda de forma explícita os critérios de segurança que deverão ser aplicados na categorização de riscos. Dessa maneira, cria-se uma base comum de entendimento que favorece respostas consistentes e alinhadas aos paradigmas contemporâneos de cibersegurança.

4.3.2.2 Descrição dos dados

Na sequência, o *prompt* apresenta uma descrição estruturada do *dataset* ZTA-RAD, detalhando as fontes de dados utilizadas (logons, dispositivos e acessos HTTP) e como essas informações foram integradas em um único arquivo consolidado. Essa etapa fornece à LLM uma visão clara sobre a natureza dos registros, as variáveis agregadas e os comportamentos que podem ser observados. Ao explicitar tanto os atributos originais quanto os derivados, garante-se que o modelo compreenda a riqueza e a granularidade do conjunto de dados.

4.3.2.3 Direcionamento da tarefa

O *prompt* orienta explicitamente a função esperada da LLM: atuar como um analista de risco no contexto da ZTA. Esse direcionamento estabelece o papel da LLM indicando que a análise deve ser realizada a partir dos pilares do paradigma de segurança. Dessa forma, reduz-se a ambiguidade na interpretação da tarefa, assegurando que a categorização dos riscos seja conduzida de maneira técnica, fundamentada e comparável entre diferentes modelos.

4.3.2.4 Definição das saídas esperadas

Por fim, o *prompt* especifica claramente as entregas desejadas, estruturadas em três níveis: (a) identificação de comportamentos de risco, (b) explicitação dos critérios, métricas ou heurísticas adotados na avaliação e (c) geração de um código em Python que implemente a lógica de classificação, rotulando cada entrada do *dataset* ZTA-RAD como de ”Alto”, ”Médio” ou ”Baixo” risco.

Essa definição das saídas cumpre um papel essencial de padronização, permitindo avaliar a consistência das respostas entre as LLMs e garantindo que as análises não se restrinjam a descrições genéricas, mas avancem para resultados aplicáveis e verificáveis.

Como resultado desse processo, foram obtidos seis *datasets* distintos, todos compostos pelo mesmo conjunto de atributos originais e complementados, diferenciando-se apenas pelo rótulo da *feature* “risco”, que contempla toda a análise descrita. Em um deles, a rotulagem foi realizada pela especialista em segurança cibernética, estabelecendo a referência humana, enquanto os outros cinco conjuntos foram gerados a partir das análises de risco produzidas individualmente por cada uma das LLMs utilizadas.

Essa abordagem garante a comparabilidade entre os diferentes cenários de rotulagem, permitindo avaliar tanto a proximidade das LLMs em relação ao julgamento especializado quanto a consistência interna entre os modelos, fornecendo assim uma base robusta para experimentação em aprendizado supervisionado.

É importante ressaltar que, além da checagem de integridade, garantindo a ausência de valores faltantes, duplicados ou inconsistentes, foi realizada uma verificação semântica, de modo a assegurar que os dados produzidos respeitam as regras inerentes a um ambiente real ZTA.

4.3.3 Correlação das Rotulagens entre LLMs e Especialista

A Figura 7 evidencia a correlação entre as rotulagens de risco atribuídas pelas LLMs e pela especialista. Observa-se um agrupamento consistente entre Grok, ChatGPT e Copilot, que apresentam correlações muito altas entre si (0,94 a 0,98), sugerindo forte convergência nos critérios adotados para a rotulagem de risco. O DeepSeek, embora tam-

bém apresente valores elevados (0,7 a 0,93), revela certa independência em relação a esse grupo, aproximando-se parcialmente de seu padrão. Por outro lado, destaca-se o comportamento do Gemini, que difere significativamente do trio anterior, apresentando correlações mais baixas com Grok, ChatGPT e Copilot (0,45 a 0,67).

Uma correlação importante pode ser observada com o Gemini (0,96), principalmente com a avaliação da especialista (0,96), configurando-se como a LLM mais próxima da referência humana utilizada neste estudo.

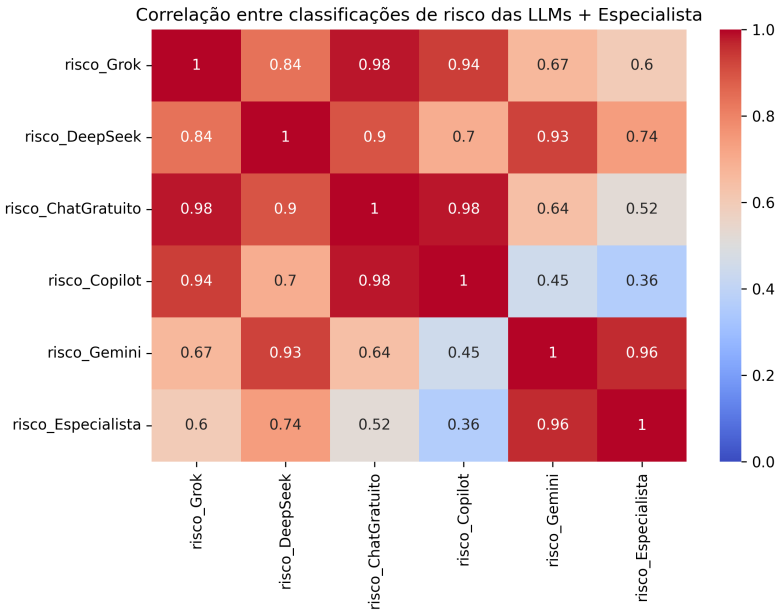

Figura 7 – Correlação entre as avaliações de risco das LLMs e Especialista.

5 Classificadores para Detecção de Ameaças Internas Utilizando o ZTA-RAD

Para a etapa de treinamento dos classificadores, optou-se por utilizar exclusivamente os *datasets* rotulados pela especialista e pela LLM Gemini. Essa decisão fundamenta-se na maior consistência observada entre ambas as análises de risco, principalmente no que tange ao risco alto (comportamento malicioso), uma vez que o modelo Gemini apresentou resultados fortemente alinhados à avaliação humana, de acordo com a Figura 7, configurando-se como a alternativa automatizada mais próxima ao julgamento especializado. Dessa forma, buscou-se assegurar maior confiabilidade no processo de treinamento, preservando tanto a robustez técnica do *dataset* quanto a aderência aos princípios do paradigma de Arquitetura de Confiança Zero (ZTA).

Destarte, foram construídos modelos de detecção de ameaças internas a partir dos *datasets* ZTA-RAD rotulados pela especialista e pelo modelo Gemini. Entretanto, durante o processo de preparação do *dataset*, identificou-se um desbalanceamento entre as classes de risco, o que poderia comprometer o desempenho dos classificadores e enviesar os resultados obtidos.

5.1 Balanceamento do ZTA-RAD

Quando as classes de um *dataset* encontram-se desbalanceadas, os classificadores tendem a privilegiar a classe majoritária, reduzindo significativamente a capacidade de detectar instâncias minoritárias, como comportamentos atípicos associados a ameaças internas. Tal viés pode gerar altas taxas de acerto aparente, mas com desempenho insatisfatório nas métricas de interesse, especialmente nas classes de maior relevância para a segurança da informação, como as de médio e alto risco. Com o intuito de corrigir o desbalanceamento, foi utilizada a técnica *Synthetic Minority Over-sampling Technique* (SMOTE) (CHAWLA et al., 2002).

De acordo com Chawla et al. (2002), o SMOTE em vez de simplesmente duplicar instâncias da classe minoritária, gera exemplos sintéticos a partir da interpolação entre vizinhos mais próximos. Essa abordagem amplia a representatividade das classes sub-representadas sem introduzir redundância artificial, preservando a diversidade dos dados e contribuindo para a formação de fronteiras de decisão mais robustas nos classificadores.

A Figura 8 ilustra, quantitativamente, o resultado do processo de balanceamento dos *datasets* por meio da técnica SMOTE, considerando tanto a rotulação feita pela especialista quanto a gerada automaticamente pelo modelo Gemini.Através da análise

dos registros originais e os sintéticos gerados pelo SMOTE, observa-se que, em ambos os casos, a quantidade de registros artificiais criados é próxima à dos originais, o que indica aumento da representatividade sem grande distorção na distribuição geral. Ao observar os gráficos inferiores, que detalham o balanceamento por classe de risco, nota-se que, no *dataset* rotulado pelo especialista (a), a classe "Médio" recebeu maior quantidade de registros sintéticos, seguida pela classe "Alto", enquanto a classe "Baixo" apresentou menor volume de geração. Por sua vez, no *dataset* rotulado pelo Gemini (b), a distribuição mostrou-se mais uniforme entre as três classes, ainda que com maior ênfase nas classes "Alto" e "Médio".

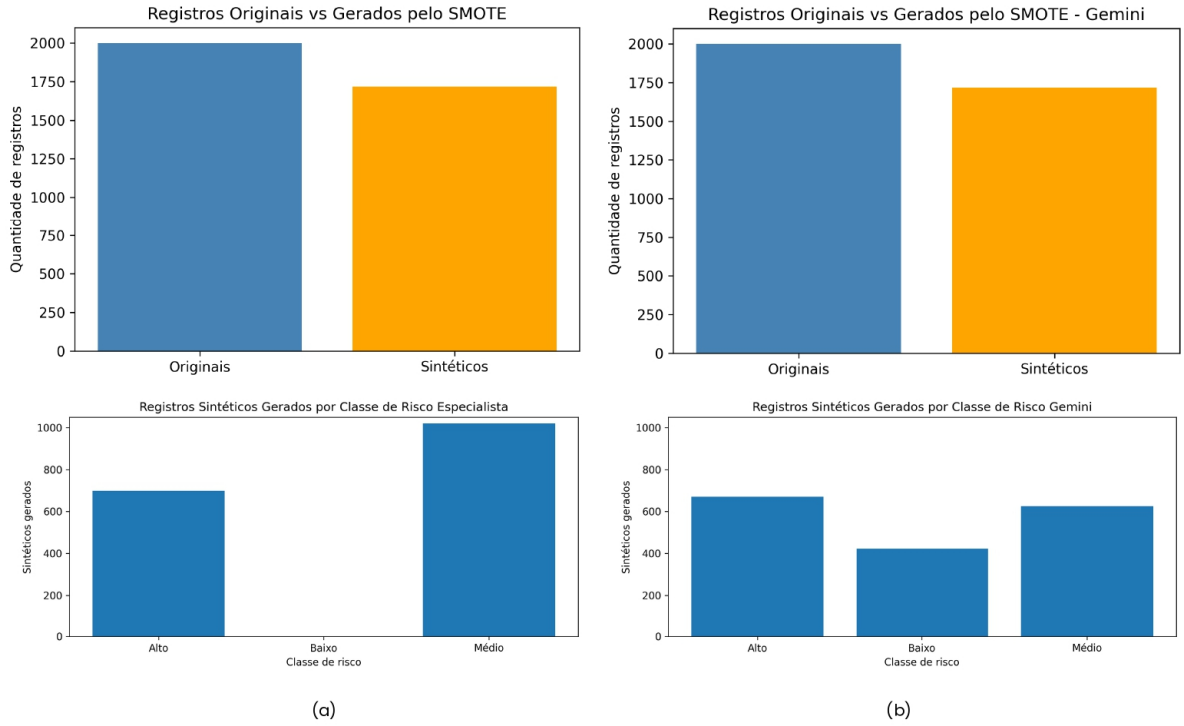

Figura 8 – Balanceamento dos *dataset* - gerado pelo Especialista e Gemini - Fonte: Do Autor.

É importante ressaltar que, ao final do processo de balanceamento, foram produzidos dois *datasets*, cada um contendo aproximadamente 3.700 registros, com distribuição uniforme entre as classes de alto, médio e baixo risco. Estes *datasets* foram utilizados no processo de treinamento dos modelos.

5.2 Treinamento dos Classificadores

Para o treinamento e validação dos modelos, optou-se pela utilização dos algoritmos *Redes Neurais Artificiais (RNAs)*, *Support Vector Machine (SVM)* e *Random Forest*. O processo de treinamento foi conduzido considerando ambas as versões dos *datasets* pro-

duzidos: (a) balanceadas e (b) não balanceadas, com o objetivo de avaliar o impacto do balanceamento nos resultados, bem como nas métricas estáticas. Para a avaliação dos modelos, foram adotadas métricas estáticas, tais como acurácia, *F1-score*, *recall* e *precision*, além da análise da matriz de confusão.

Treinar e avaliar modelos em *datasets* balanceados e não balanceados permite comparar o impacto do balanceamento no aprendizado, analisar os *trade-offs* entre *recall*, *F1-score* e precisão, além de garantir maior robustez na validação. Essa abordagem mostra se o modelo realmente aprende padrões relevantes ou apenas se beneficia da manipulação dos dados e, em contextos críticos como ZTA e segurança da informação, apoia a escolha do modelo que melhor equilibra detecção de ameaças e redução de falsos alarmes.

5.2.1 Treinamento dos Classificadores Utilizando *Datasets* Balanceados

O primeiro treinamento compreendeu a utilização do *dataset* balanceado, tanto a versão da especialista quanto a versão produzida pelo modelo Gemini. A análise das curvas de aprendizado evidencia um desempenho excelente dos modelos MLP treinados.

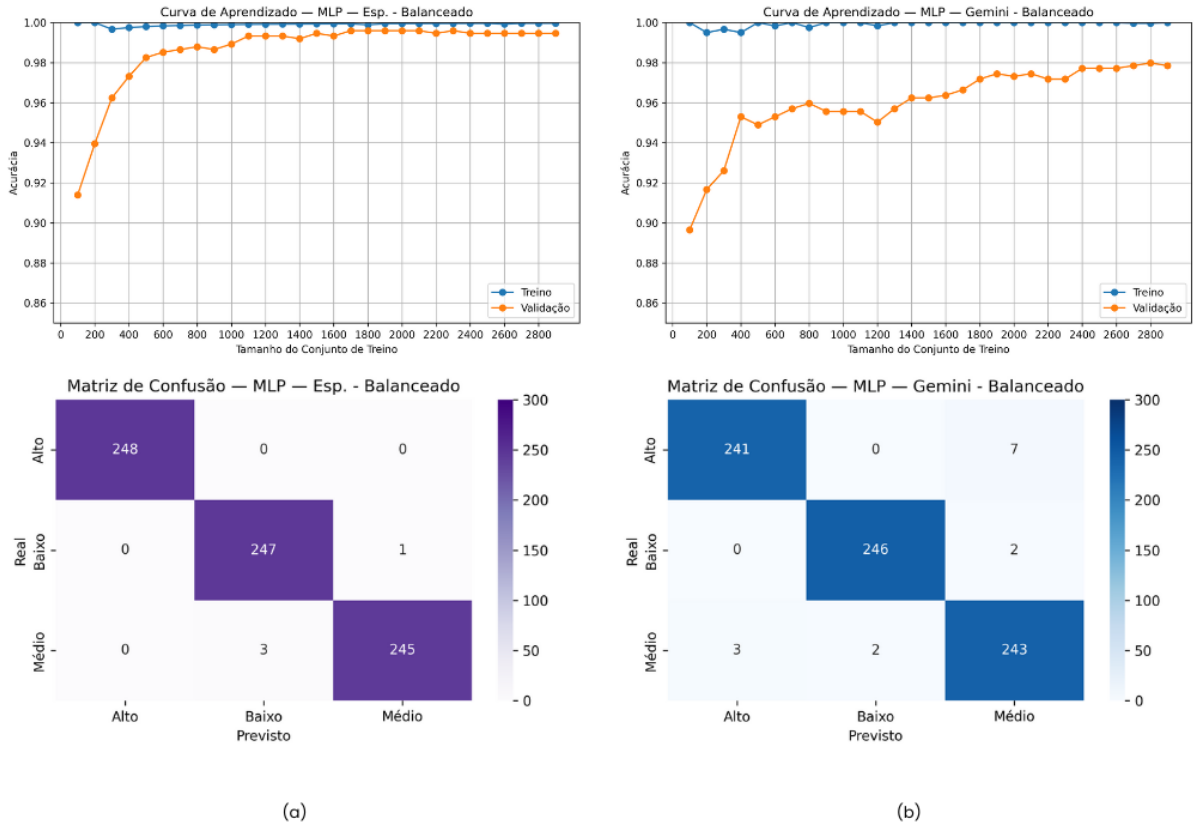

Figura 9 – Treinamento do classificador MLP utilizando os *datasets* balanceados - Fonte: Do Autor.

As Figuras 9 (a) e (b) mostram que ambos os modelos produzidos através da ro-

tulagem feita pela especialista e pela LLM Gemini alcançaram desempenho elevado com os *datasets* balanceados, evidenciando a eficácia do processo de balanceamento. O modelo treinado com o *dataset* produzido pela especialista apresentou curvas mais estáveis, alcançando acurácia próxima de 0,99, além da matriz de confusão com poucos erros de classificação. Já o modelo treinado com o *dataset* rotulado pela LLM Gemini, embora também tenha obtido bons resultados, apresentou uma pequena oscilação durante o treinamento e um maior número de erros de classificação em relação à Figura 9 (a), sobretudo entre as classes de risco "Alto" e "Médio".

As Figuras 10 (a) e (b) ilustram os resultados obtidos durante o treinamento utilizando o algoritmo *Random Forest* em *datasets* balanceados, evidenciam alto desempenho em ambos os cenários (Especialista e a LLM Gemini). As curvas de aprendizado mostram acurácia de validação crescente e estável, atingindo valores próximos de 0,99. A matriz de confusão do modelo produzido a partir do *dataset* gerado pela especialista apresenta classificação quase perfeita, enquanto o modelo produzido a partir do *dataset* Gemini, apesar de também alcançar excelente desempenho, registra pequenos desvios adicionais, especialmente na distinção entre as classes de risco "Alto" e "Médio".

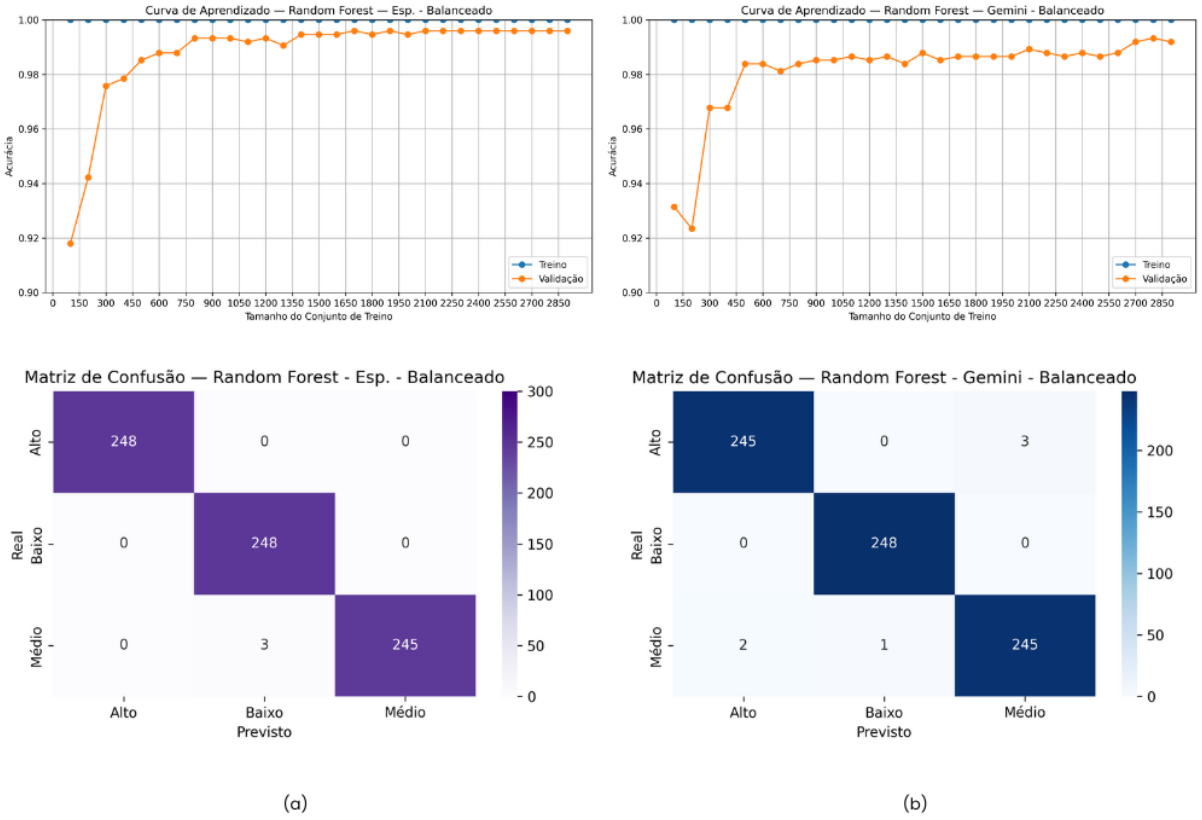

Figura 10 – Treinamento do classificador *Random Forest* utilizando os *datasets* balanceados - Fonte: Do Autor.

De modo geral, de acordo com as Figuras 10 (a) e (b), ambos os modelos demons-

tram alta precisão, mas o treinamento orientado pelo *dataset* do especialista se mostra ligeiramente mais consistente, reduzindo a ocorrência de falsos negativos, aspecto essencial em aplicações críticas como a detecção de ameaças em ambientes de ZTA.

As Figuras 11 (a) e (b) ilustram os resultados obtidos durante o treinamento utilizando o algoritmo SVM em *datasets* balanceados. Observa-se que o desempenho na fase de treinamento e validação foi inferior aos modelos MLP e *Random Forest*. As curvas de aprendizado indicam acurácia de validação estabilizando-se entre 0,80 e 0,82 para o modelo treinado a partir do *dataset* do especialista e em torno de 0,77 a 0,79 para o modelo treinado a partir do *dataset* produzido pelo Gemini, revelando menor poder de generalização e maior dificuldade em separar as classes de risco de forma precisa. As matrizes de confusão confirmam essa tendência: para o modelo produzido a partir da análise da especialista, a classe de risco "Baixo" apresenta confusão significativa com a classe de risco "Médio", enquanto no modelo produzido a partir da análise do Gemini há maior dispersão de erros, especialmente na distinção entre as classes de risco "Alto" e "Médio".

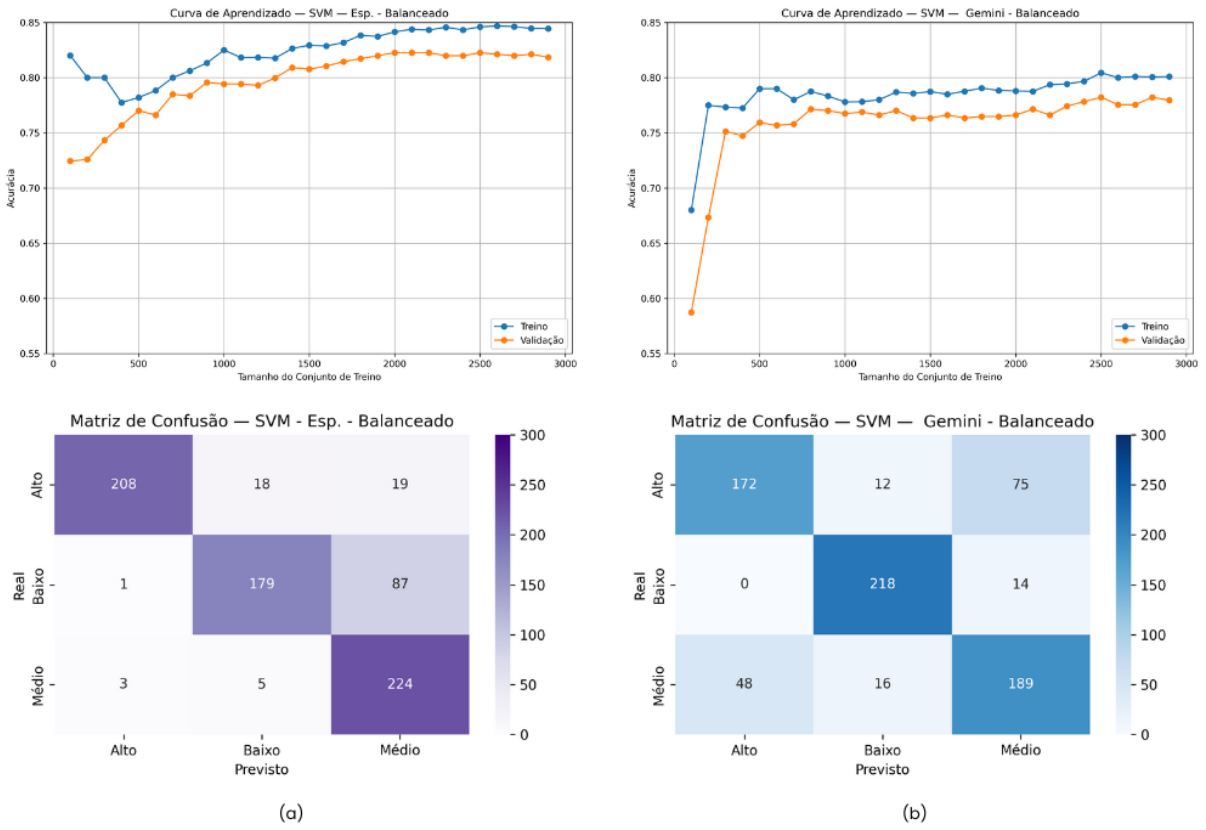

Figura 11 – Treinamento do classificador SVM utilizando os *datasets* balanceados - Fonte: Do Autor.

Assim, embora o SVM alcance resultados bons, sua capacidade de classificação se mostra mais limitada, com maior taxa de falsos positivos e falsos negativos, o que compro-

mete sua confiabilidade em cenários críticos como a detecção de ameaças em ambientes baseados em ZTA.

5.2.2 Treinamento dos Classificadores Utilizando *Datasets* Não Balanceados

Assim como o primeiro treinamento compreendeu a utilização do *dataset* balanceado, tanto a versão da especialista quanto a versão produzida pelo Gemini, neste segundo momento, foram utilizados para o processo de treinamento os *datasets* produzidos pela especialista e a LLM Gemini sem balanceamento.

As Figuras 12 (a) e (b) ilustram os resultados obtidos no processo de treinamento do modelo baseado em MLP a partir dos *datasets* não balanceados construídos a partir da rotulagem de risco do especialista e da LLM Gemini. Os resultados mostram acurácias acima de 0,95 para ambos os cenários (especialista e Gemini), com curvas de aprendizado indicando boa capacidade de generalização.

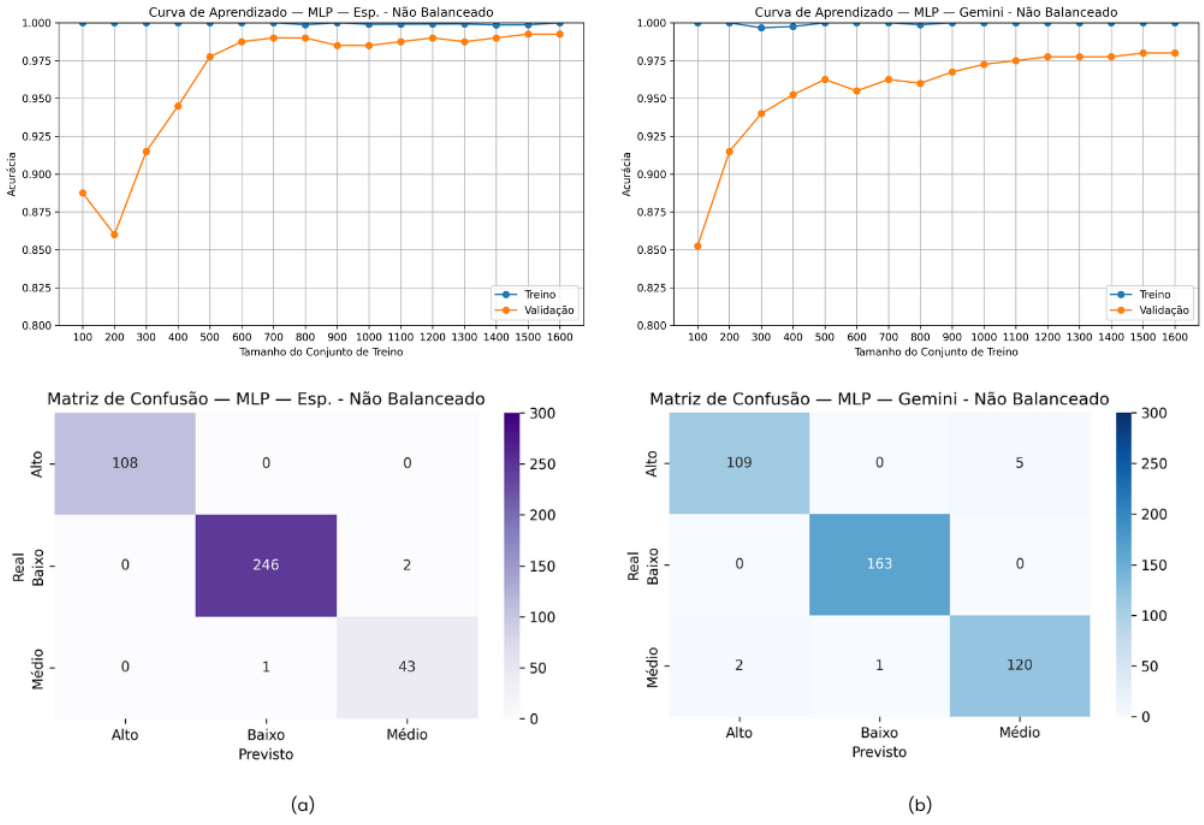

Figura 12 – Treinamento do classificador MLP utilizando os *datasets* não balanceados - Fonte: Do Autor.

Ao analisar as matrizes de confusão da Figura 12 , observa-se o impacto do desbalanceamento no modelo produzido a partir da rotulagem do especialista, a classe de risco ”Baixo” domina as previsões, resultando em alto acerto para essa categoria, mas

com perda significativa na classe de risco ”Médio” (43 acertos contra 45 erros). Situação semelhante é observada no modelo produzido a partir da rotulagem do Gemini, no qual a classe de risco ”Baixo” concentra 163 acertos, enquanto as classes de risco ”Alto” e ”Médio” apresentam maior confusão, especialmente esta última.

Com base nesses resultados, entende-se que, apesar da alta acurácia geral, o desbalanceamento leva o modelo a priorizar a classe majoritária, reduzindo *recall* e *F1-score* para classes minoritárias, o que compromete a confiabilidade em contextos críticos como a detecção de ameaças em ZTA, onde a correta identificação de casos raros é fundamental.

As Figuras 13 (a) e (b) ilustram os resultados obtidos no processo de treinamento do modelo baseado em *Random Forest* a partir dos *datasets* não balanceados construídos a partir da rotulagem de risco da especialista e da LLM Gemini. Os resultados apresentam curvas de aprendizado com acurácia de validação superior a 0,95 em ambos os cenários (Especialista e Gemini), indicando boa capacidade de generalização e estabilidade do modelo. No entanto, assim como observado no MLP, o desbalanceamento impacta diretamente a matriz de confusão.

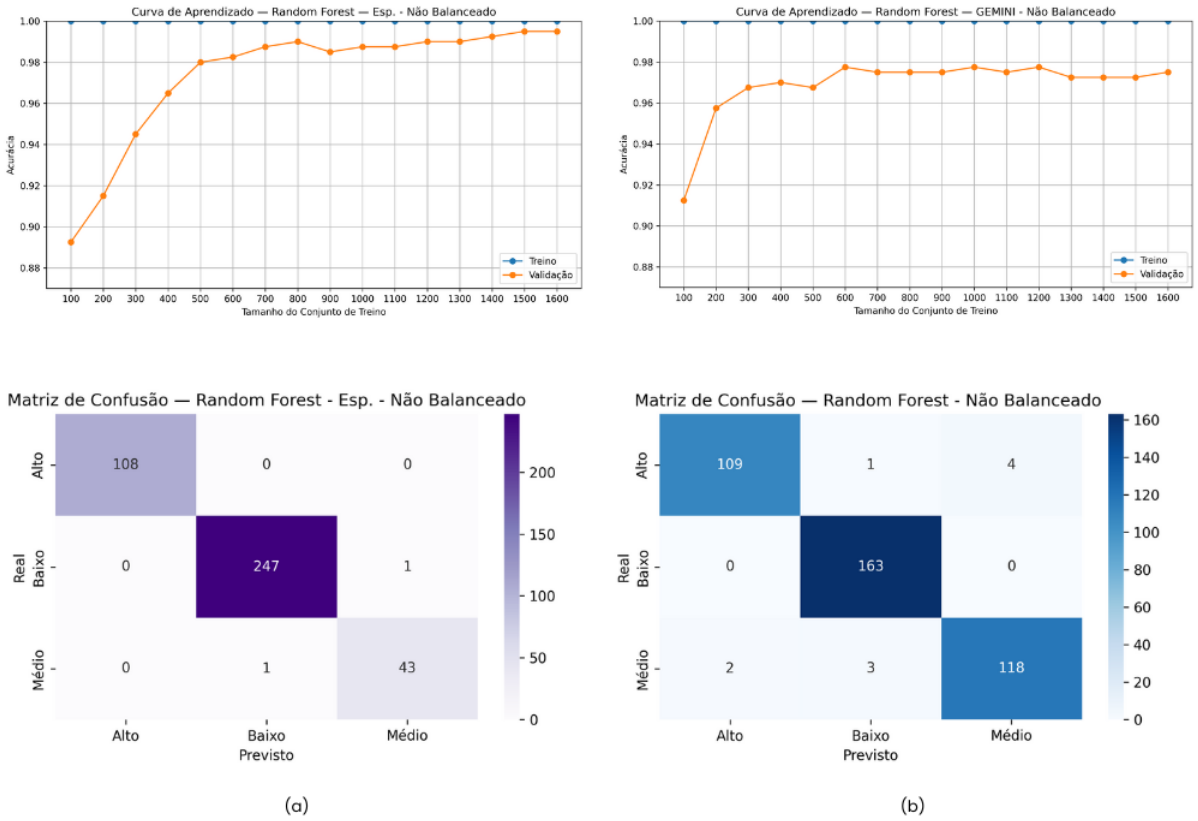

Figura 13 – Treinamento do classificador *Random Forest* utilizando os *datasets* não balanceados - Fonte: Do Autor.

Nos resultados obtidos a partir do modelo produzido pela análise da especialista, a classe Baixo concentra a maior parte dos acertos (247), enquanto a classe Médio so-

fre grande perda, com apenas 43 instâncias corretamente classificadas. Já no modelo produzido pela análise da LLM Gemini, apesar da acurácia geral elevada, observa-se desempenho semelhante, com predominância de acertos na classe Baixo (163) e confusão significativa entre as classes de risco ”Alto” e ”Médio”. Esses resultados demonstram que, embora o classificado *Random Forest* seja robusto em termos de acurácia global, o desbalanceamento leva à priorização da classe majoritária, reduzindo a sensibilidade às classes minoritárias. Em aplicações críticas de segurança, isso pode resultar em maior risco de falsos negativos, reforçando a importância do balanceamento para garantir uma detecção mais justa e confiável entre todas as classes de risco.

As Figuras 13 (a) e (b) ilustram os resultados obtidos no processo de treinamento do modelo baseado em SVM a partir dos *datasets* não balanceados construídos a partir da rotulagem de risco do especialista e da LLM Gemini. Os resultados revelam desempenho inferior em comparação aos demais algoritmos testados. As curvas de aprendizado indicam acurácia de validação estabilizando-se entre 0,75 e 0,82 com menor capacidade de generalização.

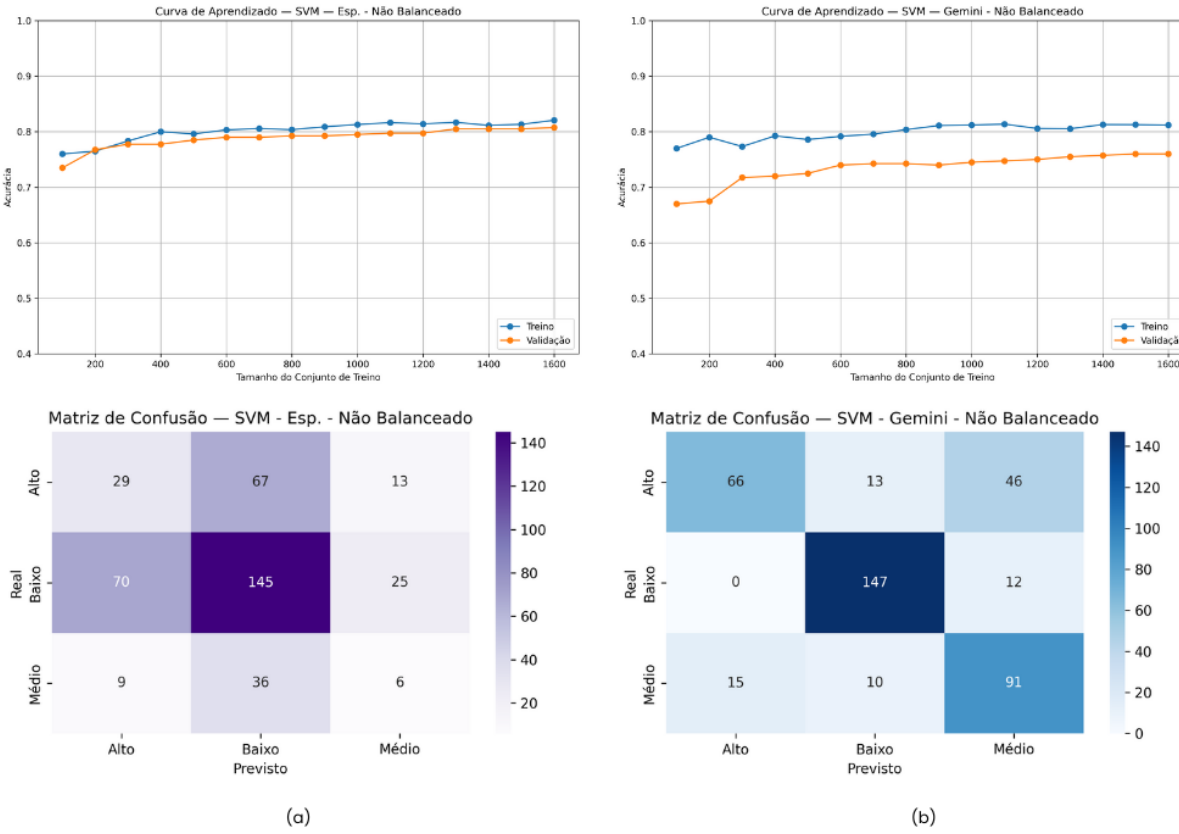

Figura 14 – Treinamento do classificador SVM utilizando os *datasets* não balanceados.

De acordo com com as Figuras 13 (a) e (b), no que tange as matrizes de confusão, observa-se o forte impacto do desbalanceamento, onde, no modelo produzido a partir da rotulagem da especialista, a classe de risco ”Baixo” concentra a maioria das previsões

corretas (145), mas com alta confusão em relação às classes de risco "Alto" e "Médio", que apresentam elevado número de erros. No modelo produzido a partir da rotulagem do Gemini, observa-se melhora na classificação da classe de risco "Alto" (66 acertos), porém ainda com muitos erros distribuídos nas demais classes, sobretudo entre "Alto" e "Médio". Esses resultados evidenciam que, diante de dados não balanceados, o SVM tende a favorecer a classe majoritária e reduzir significativamente o *recall* das classes minoritárias. Em cenários críticos como a detecção de ameaças em ZTA, essa limitação compromete a confiabilidade do modelo, reforçando a necessidade do balanceamento para melhorar a discriminação entre categorias.

A Tabela 6 apresenta as métricas de desempenho dos modelos MLP, *Random Forest* e SVM produzidos a partir dos *datasets* balanceados e não balanceados rotulados pela especialista e a LLM Gemini. Observa-se que MLP e *Random Forest* obtiveram acurácias próximas de 1.00, com alta precisão, revocação e F1-Score em todas as classes, tanto em cenários balanceados quanto não balanceados. Já o SVM apresentou desempenho inferior, com acurácia variando entre 0.45 e 0.82, além de maior variação entre classes e sensibilidade ao desbalanceamento, destacando-se como o algoritmo menos adequado para o problema em análise.

Tabela 6 – Métricas de avaliação dos modelos - *datasets* Balanceados e Não Balanceados - Fonte: Do Autor

Modelos	**Acurácia**	**Precisão**			**Revocação**			**F1-Score**		
		Alto	Médio	Baixo	Alto	Médio	Baixo	Alto	Médio	Baixo
MLP (Esp. Bal.)	0.99	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.99	0.99
MLP (Gemini Bal.)	0.98	0.99	0.96	0.99	0.97	0.98	0.99	0.98	0.97	0.99
MLP (Esp. N. Bal.)	0.99	1.00	0.96	1.00	1.00	0.98	0.99	1.00	0.97	0.99
MLP (Gemini N. Bal.)	0.98	0.98	0.96	0.99	0.96	0.98	1.00	0.97	0.97	1.00
RF (Esp. Bal.)	1.00	1.00	1.00	0.99	1.00	0.99	1.00	1.00	0.99	0.99
RF (Gemini Bal.)	0.99	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.99	1.00
RF (Esp. N. Bal.)	0.99	1.00	0.98	1.00	1.00	0.98	1.00	1.00	0.98	1.00
RF (Gemini N. Bal.)	0.97	0.98	0.97	0.98	0.96	0.96	1.00	0.97	0.96	0.99
SVM (Esp. Bal.)	0.82	0.98	0.68	0.89	0.85	0.97	0.67	0.91	0.80	0.76
SVM (Gemini Bal.)	0.78	0.78	0.68	0.89	0.66	0.75	0.94	0.72	0.71	0.91
SVM (Esp. N. Bal.)	0.45	0.27	0.14	0.58	0.27	0.12	0.60	0.27	0.13	0.59
SVM (Gemini N. Bal.)	0.76	0.81	0.61	0.86	0.53	0.78	0.92	0.64	0.69	0.89

De acordo com Tabela 6, o modelo *Random Forest* apresentou o melhor desempenho geral, alcançando valores de acurácia, precisão, revocação e F1-Score próximos de 1.00 em todos os cenários, tanto com *datasets* balanceados quanto não balanceados, o que demonstra sua robustez e consistência na detecção das diferentes classes de risco. O modelo MLP também obteve resultados satisfatórios, com métricas elevadas e estáveis, ainda que ligeiramente inferiores às do *Random Forest*. Em contrapartida, o SVM apresentou desempenho significativamente inferior, especialmente em *datasets* não balanceados, nos quais a acurácia chegou a 0.45, evidenciando maior sensibilidade ao desbalanceamento das classes.

Em suma, observa-se que os *datasets* rotulados pela especialista resultaram, em geral, em métricas superiores às obtidas com a rotulação do modelo Gemini. Embora ambos os cenários tenham apresentado desempenhos elevados para *Random Forest* e MLP, a análise da especialista mostrou-se mais consistente, especialmente quando considerada a influência do balanceamento nas métricas de precisão, revocação e F1-Score. Essa diferença reforça a importância da curadoria humana na etapa de rotulação, uma vez que contribui para a construção de modelos mais robustos e confiáveis na detecção de ameaças internas.

5.2.3 Síntese dos Resultados

Inicialmente, destacou-se a importância do balanceamento das classes de risco por meio da técnica SMOTE, uma vez que os cenários desbalanceados mostraram tendência de priorização da classe majoritária, resultando em perda de sensibilidade nas classes minoritárias, justamente aquelas de maior relevância para a detecção de ameaças internas.

Os experimentos apresentados neste capítulo demonstraram que os modelos baseados em MLP e *Random Forest* alcançaram os melhores resultados, com elevadas taxas de acurácia, precisão, revocação e F1-score, tanto em *datasets* balanceados quanto não balanceados, ainda que com maior robustez quando aplicado o balanceamento. Já o SVM, embora tenha apresentado desempenho razoável em algumas métricas nos cenários balanceados, mostrou-se mais limitado e sensível ao desbalanceamento, apresentando maiores taxas de erro e, consequentemente, menor confiabilidade em contextos críticos.

De forma geral, os resultados evidenciaram que o balanceamento das classes foi um fator determinante para assegurar a performance dos classificadores, principalmente no que tange à redução de falsos negativos, aspecto essencial em aplicações relacionadas à ZTA. Assim, a utilização de *datasets* balanceados, combinada à escolha de algoritmos tais como MLP e Random Forest, mostrou-se a estratégia mais eficaz para a detecção de ameaças internas em ambientes ZTA.

6 Análise e Discussão dos Resultados

A principal contribuição deste trabalho reside na construção do ZTA-RAD, um *dataset* inédito voltado à análise e à detecção de ameaças internas (*insider threats*) sob a perspectiva da Arquitetura de Confiança Zero (ZTA). Diferentemente de bases amplamente utilizadas em segurança, mas pouco contextualizadas no paradigma de confiança zero, o ZTA-RAD foi estruturado a partir de múltiplas fontes de *logs* (logon, dispositivos e HTTP), permitindo a caracterização de comportamentos de usuários em diferentes dimensões de risco. Essa abordagem viabiliza a criação de cenários realistas para o estudo de ameaças internas, oferecendo não apenas volume e diversidade de registros, mas também uma organização orientada ao mapeamento de riscos em níveis (baixo, médio e alto).

Ao disponibilizar esse conjunto de dados para a comunidade, o trabalho contribui para experimentação, avaliação de algoritmos de aprendizado de máquina e desenvolvimento de soluções práticas de monitoramento em ambientes corporativos que adotam princípios de ZTA.

Sob a óptica do balanceamento dos *datasets*, demonstra-se que o uso do SMOTE exerceu papel fundamental na melhoria do desempenho dos classificadores, sobretudo no tratamento do viés em relação à classe majoritária. Nos cenários onde foi utilizado o *dataset* não balanceados, os modelos tenderam a concentrar suas predições na classe de risco "Baixo" alcançando valores aparentemente elevados de acurácia, mas às custas de redução da revocação e do F1-score nas classes de risco "Médio" e "Alto". Esta condição é colocada como uma limitação em aplicações de segurança, uma vez que tais classes representam os comportamentos mais relevantes para a detecção de atividades maliciosas e potenciais ameaças internas.

Com relação ao treinamento e validação dos classificadores, os experimentos mostraram que os modelos baseados em MLP e *Random Forest* foram os que melhor responderam ao problema de detecção de ameaças internas, apresentando métricas de desempenho consistentemente elevadas. Em ambos os modelos, observou-se acurácia próxima de 1.00, acompanhada de altos valores de precisão, revocação e F1-score em todas as classes, tanto nos cenários balanceados quanto não balanceados. Ainda assim, os resultados foram mais estáveis e robustos após a aplicação do balanceamento via SMOTE.

O SVM, por sua vez, demonstrou limitações mais evidentes: embora tenha alcançado resultados razoáveis com *datasets* balanceados, sua performance caiu consideravelmente em cenários não balanceados, com oscilações significativas entre classes de risco "Médio" e "Alto".

Sob a perspectiva da ZTA, a capacidade de reduzir falsos negativos constitui um

aspecto central para a eficácia dos sistemas de detecção. Isso ocorre porque a não identificação de uma ameaça interna (falso negativo) significa permitir que um comportamento malicioso permaneça ativo, expondo dados sensíveis e comprometendo a integridade do ambiente. Em ZTA isto representa um cenário catastrófico, pois se pratica o princípio de "nunca confiar, sempre verificar", o que exige respostas rápidas a desvios de comportamento.

Em contraste, a ocorrência de falsos positivos, embora indesejada, implica apenas na sinalização equivocada de atividades legítimas como suspeitas, gerando sobrecarga operacional, mas sem o mesmo nível de risco direto à segurança da informação. Dessa forma, os resultados obtidos com MLP e *Random Forest*, que apresentaram desempenho mais consistente na redução de falsos negativos, são mais valiosos em termos práticos do que pequenas variações de precisão.

Embora uma análise de similaridade, ilustrada pela Figura 7, tenha demonstrado forte convergência entre a rotulagem realizada pela especialista e a produzida pelo modelo Gemini, os experimentos evidenciaram que os classificadores apresentaram métricas superiores quando treinados com o *dataset* rotulado pela especialista. Esse resultado sugere que, embora a LLM tenha conseguido replicar de forma próxima os padrões de julgamento humano, a rotulagem manual apresenta maior refinamento na definição das fronteiras entre as classes de risco, fornecendo exemplos mais consistentes para o processo de aprendizado.

Destarte, a intervenção humana contribui para reduzir ambiguidades presentes em situações limítrofes, o que se reflete em maior estabilidade das curvas de aprendizado, melhor equilíbrio entre precisão e revocação e redução de erros em classes críticas, como a de alto risco. Assim, conclui-se que, embora as LLMs possam atuar como ferramentas de apoio promissoras para a construção de bases de dados em larga escala, a curadoria especializada ainda se mostra indispensável em aplicações de segurança baseadas no paradigma da Arquitetura de Confiança Zero (ZTA), pois garante maior confiabilidade e robustez na detecção de comportamentos anômalos e ameaças internas.

Consolidar um *dataset* realista para análise de ameaças internas em ambientes corporativos representa um grande desafio para as empresas, principalmente devido à sensibilidade dos dados, que não podem ser expostos por questões legais, éticas e de privacidade. Nesse contexto, a criação e o enriquecimento de um *dataset*, como o proposto neste estudo, assumem relevância prática significativa, pois oferecem uma alternativa segura e representativa para apoiar pesquisas e aplicações em segurança da informação.

Por fim, esta proposta supera os trabalhos correlatos por contemplar um número maior de usuários e por não se restringir apenas a interações no sistema de arquivos Windows. Além disso, apresenta menor complexidade e alta capacidade de replicabilidade, o que favorece sua adoção em diferentes contextos experimentais. Embora seja construída

e enriquecida a partir do *dataset* CERT, mantém elevada representatividade de cenários atuais, com ênfase especial em ambientes baseados em ZTA, ampliando sua relevância prática e acadêmica.

7 Conclusão

Este trabalho apresentou a construção e validação do *ZTA-RAD*, um novo *dataset* voltado à análise de ameaças internas em ambientes fundamentados no paradigma da Arquitetura de Confiança Zero (ZTA). Diferentemente de iniciativas anteriores que se limitam a cenários baseados em perímetro, o *ZTA-RAD* incorpora métricas específicas de logon, dispositivos e acessos HTTP, refletindo práticas contemporâneas de autenticação contínua e monitoramento granular. Essa contribuição preenche uma lacuna significativa observada na literatura e oferece uma base inédita para pesquisas na área de segurança cibernética para ambientes ZTA. Torna-se importante destacar que, o *dataset* ZTA-RAD foi construído a partir do *dataset* CERT ampliado com diversos comportamentos desenhados por uma especialista da área de segurança cibernética.

Outro aspecto inovador deste trabalho foi o processo de rotulagem de risco conduzido de forma híbrida, combinando a análise de uma especialista em segurança com a inferência de cinco LLMs (Grok, GPT, Gemini, Copilot e DeepSeek). Esse arranjo não apenas permitiu avaliar a capacidade das LLMs em reproduzir julgamentos humanos, mas também revelou a relevância da curadoria especializada como "padrão-ouro" para redução de ambiguidades em cenários críticos, especialmente na classificação de risco.

Os experimentos de treinamento e validação realizados com três algoritmos, a saber: (a) MLP, (b) *Random Forest* e (c) SVM, demonstraram que os dois primeiros alcançaram métricas mais consistentes e robustas, com destaque para a redução de falsos negativos em *datasets* balanceados, um fator crucial na mitigação de ameaças internas. Já o SVM apresentou maior sensibilidade ao desbalanceamento, reforçando a necessidade de ajustes adicionais em cenários práticos. A utilização da técnica SMOTE se mostrou essencial para corrigir vieses em favor da classe majoritária, elevando o desempenho e a confiabilidade dos classificadores.

Assim, conclui-se que o presente estudo avança em três dimensões principais: (i) a proposição de um *dataset* inédito e contextualizado à ZTA, (ii) a experimentação de um processo híbrido de rotulagem com apoio de LLMs, e (iii) a validação sistemática de algoritmos de aprendizado de máquina em cenários balanceados e não balanceados. Tais contribuições oferecem subsídios relevantes tanto para a comunidade científica quanto para aplicações práticas de segurança, indicando caminhos promissores para futuras pesquisas, como a ampliação do *dataset* para outras fontes de *logs* e a exploração de modelos mais avançados de inteligência artificial para detecção de ameaças internas.

7.1 Trabalhos Futuros

Como direcionamentos para trabalhos futuros, destaca-se inicialmente a ampliação dos experimentos de treinamento e validação para além das rotulagens realizadas pela especialista e pela LLM Gemini. Embora ambas tenham apresentado correlação de análise de risco, o treinamento dos modelos mostrou diferenças expressivas, sendo que a rotulagem da especialista proporcionou resultados superiores em comparação com a da LLM Gemini. Dessa forma, seria pertinente investigar se o mesmo comportamento se repete ao considerar as demais LLMs utilizadas neste estudo (Grok, GPT, Copilot e DeepSeek).

No intuito de enriquecer o *dataset* e alinhá-lo de forma mais consistente aos princípios da ZTA, propõe-se a inclusão de novas *features*, adicionalmente, a incorporação de informações contextuais, como localização geográfica do acesso, tipo de dispositivo, perfil de risco do usuário e desvios em relação ao comportamento histórico e informações psicométricas para contribuir com um diagnóstico mais preciso e dinâmico de potenciais ameaças internas.

Outro ponto relevante refere-se à exploração de novos algoritmos de aprendizado de máquina. Considerando que o SVM apresentou desempenho consideravelmente inferior em relação ao MLP e ao *Random Forest*, futuras pesquisas podem avaliar classificadores adicionais, como XGBoost ou mesmo arquiteturas mais profundas de redes neurais, buscando observar seu comportamento em cenários balanceados e não balanceados.

Por fim, sugere-se a incorporação de técnicas de *Inteligência Artificial Explicável* (XAI), tais como SHAP ou LIME, a fim de analisar o impacto das *features* no processo de treinamento e validação. Essa abordagem não apenas ampliaria a interpretabilidade dos resultados, mas também forneceria maior transparência para a compreensão dos fatores que influenciam a detecção de ameaças internas, fortalecendo a confiabilidade das soluções propostas no contexto de ZTA.

Referências

CAMINA, J. B.; HERNÁNDEZ-GRACIDAS, C.; MONROY, R.; TREJO, L. The windows-users and-intruder simulations logs dataset (wuil): An experimental framework for masquerade detection mechanisms. **Expert Systems with Applications**, Elsevier, v. 41, n. 3, p. 919–930, 2014. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2013.08.022>. Citado 3 vezes nas páginas 22, 24 e 25.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002. ISSN 1076-9757. Disponível em: <http://dx.doi.org/10.1613/jair.953>. Citado na página 41.

CISA. **Defining Insider Threats**. 2025. Disponível em: <https://www.cisa.gov/topics/physical-security/insider-threat-mitigation/defining-insider-threats>. Acesso em: 24 Ago. 2025. Citado na página 10.

COLLINS, M. **Common Sense Guide to Mitigating Insider Threats, Fifth Edition**. Pittsburgh, 2016. Disponível em: <http://dx.doi.org/10.21236/ada585500>. Acesso em: 24 jul. 2025. Citado na página 10.

DINARDO, K.; LEMOUDDEN, M.; AHMAD, J. Insider threat detection on an imbalanced dataset using balancing methods. In: SPRINGER. **Science and Information Conference**. Cham: Springer Nature Switzerland, 2023. p. 1216–1226. ISBN 9783031377174. ISSN 2367-3389. Disponível em: <http://dx.doi.org/10.1007/978-3-031-37717-4_80>. Acesso em: 25 ago. 2025. Citado 2 vezes nas páginas 6 e 19.

Fortinet Inc. **Zero Trust Network Access Introduction**. 2025. Disponível em: <https://docs.fortinet.com/document/fortigate/7.4.8/administration-guide/855420/zero-trust-network-access-introduction>. Acesso em: 10 de agosto de 2025. Citado 4 vezes nas páginas 11, 16, 17 e 18.

GLASSER, J.; LINDAUER, B. Bridging the gap: A pragmatic approach to generating insider threat data. In: **2013 IEEE Security and Privacy Workshops**. IEEE, 2013. p. 98–104. Disponível em: <http://dx.doi.org/10.1109/spw.2013.37>. Citado na página 18.

HARILAL, A.; TOFFALINI, F.; HOMOLIAK, I.; CASTELLANOS, J. H.; GUARNIZO, J.; MONDAL, S.; OCHOA, M. The wolf of sutd (twos): A dataset of malicious insider threat behavior based on a gamified competition. **J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.**, v. 9, n. 1, p. 54–85, 2018. Disponível em: <http://dx.doi.org/10.1145/3139923.3139929>. Citado 3 vezes nas páginas 22, 24 e 25.

KETHIREDDY, R. R. Ai-powered insider threat detection with behavioral analytics with llm. **International Journal of Science and Research (IJSR)**, 2022. ISSN 2319-7064. Disponível em: <http://dx.doi.org/10.21275/sr221013110718>. Citado na página 20.

LE, D. C.; ZINCIR-HEYWOOD, N.; HEYWOOD, M. I. Analyzing data granularity levels for insider threat detection using machine learning. **IEEE Transactions on Network and Service Management**, IEEE, v. 17, n. 1, p. 30–44, 2020. ISSN 2373-7379. Disponível em: <http://dx.doi.org/10.1109/tnsm.2020.2967721>. Citado 2 vezes nas páginas 24 e 25.

LEGG, P. A.; MOFFAT, N.; NURSE, J. R.; HAPPA, J.; AGRAFIOTIS, I.; GOLDSMITH, M.; CREESE, S. Towards a conceptual model and reasoning structure for insider threat detection. **Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications**, Innovative Information Science & Technology Research Group, v. 4, n. 4, p. 20–37, 2013. Disponível em: <https://doi.org/10.22667/JOWUA.2013.12.31.020>. Citado 2 vezes nas páginas 23 e 25.

LI, C.; ZHU, Z.; HE, J.; ZHANG, X. Redchronos: A large language model-based log analysis system for insider threat detection in enterprises. **arXiv preprint arXiv:2503.02702**, 2025. Disponível em: <https://doi.org/10.48550/arXiv.2503.02702>. Citado na página 20.

LINDAUER, B. **Insider Threat Test Dataset**. Carnegie Mellon University, 2020. Dataset. Disponível em: <https://doi.org/10.1184/R1/12841247.v1>. Acesso em: 17 Jul. 2025. Citado 6 vezes nas páginas 6, 11, 18, 19, 32 e 36.

MANOHARAN, P.; YIN, J.; WANG, H.; ZHANG, Y.; YE, W. Insider threat detection using supervised machine learning algorithms. **Telecommunication Systems**, Springer, v. 87, n. 4, p. 899–915, 2024. ISSN 1572-9451. Disponível em: <http://dx.doi.org/10.1007/s11235-023-01085-3>. Citado na página 21.

MAVROUDIS, V. **Zero-Trust Network Access (ZTNA)**. 2024. Disponível em: <https://doi.org/10.48550/arXiv.2410.20611>. Acesso em: 18 jul. 2012. Citado na página 16.

NADEAU, J. **83% of organizations reported insider attacks in 2024**. IBM Security, 2024. Disponível em: <https://www.ibm.com/think/insights/83-percent-organizations-reported-insider-threats-2024>. Acesso em: 24 de Ago. 2025. Citado na página 10.

NASCIMENTO, E.; NEVES, J. E. D. Arquitetura zero trust: boas práticas de gestão de riscos de segurança da informação. **Revista Brasileira em Tecnologia da Informação**, v. 6, n. 1, p. 69–82, 2024. Acesso em: 20 jul. 2012. Citado 2 vezes nas páginas 14 e 15.

RAVAL, M. S.; GANDHI, R.; CHAUDHARY, S. Insider threat detection: Machine learning way. In: _____. **Versatile Cybersecurity**. Cham: Springer International Publishing, 2018. p. 19–53. ISBN 978-3-319-97643-3. Disponível em: <https://doi.org/10.1007/978-3-319-97643-3_2>. Citado na página 21.

SALEM, M. B.; STOLFO, S. J. Modeling user search behavior for masquerade detection. In: SOMMER, R.; BALZAROTTI, D.; MAIER, G. (Ed.). **Recent Advances in Intrusion Detection**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 181–200. ISBN 978-3-642-23644-0. Disponível em: <https://doi.org/10.1007/978-3-642-23644-0_10>. Citado na página 22.

SARHAN, B. B.; ALTWAIJRY, N. Insider threat detection using machine learning approach. **Applied Sciences**, MDPI, v. 13, n. 1, p. 259, 2022. ISSN 2076-3417. Disponível em: <http://dx.doi.org/10.3390/app13010259>. Citado 2 vezes nas páginas 11 e 21.

SCHONLAU, M.; DUMOUCHEL, W.; JU, W.-H.; KARR, A. F.; THEUS, M.; VARDI, Y. Computer intrusion: Detecting masquerades. **Statistical Science**, Institute of Mathematical Statistics, v. 16, n. 1, p. 58–74, 2001. ISSN 08834237, 21688745. Disponível em: <http://dx.doi.org/10.1214/ss/998929476>. Citado 3 vezes nas páginas 23, 24 e 25.

SOARES, T.; MELLO, J.; BARCELLOS, L.; SAYYED, R.; SIQUEIRA, G.; CASOLA, K.; COSTA, E.; GUSTAVO, N.; FEITOSA, E.; KREUTZ, D. Detecção de malwares android: Levantamento empírico da disponibilidade e da atualização das fontes de dados. In: SBC. **Escola Regional de Redes de Computadores (ERRC)**. Florianópolis/SC, 2021. p. 49–54. Disponível em: <https://doi.org/10.5753/errc.2021.18541>. Citado na página 18.

SONG, C.; MA, L.; ZHENG, J.; LIAO, J.; KUANG, H.; YANG, L. **Audit-LLM: Multi-Agent Collaboration for Log-based Insider Threat Detection**. 2024. Disponível em: <https://doi.org/10.48550/arXiv.2408.08902>. Acesso em: 25 jun. 2025. Citado na página 20.

STAFFORD, V. Zero trust architecture. **NIST special publication**, v. 800, n. 207, p. 800–207, 2020. Disponível em: <https://doi.org/10.6028/NIST.SP.800-207>. Citado 3 vezes nas páginas 11, 15 e 16.

TIAN, T.; ZHANG, C.; JIANG, B.; FENG, H.; LU, Z. Insider threat detection for specific threat scenarios. **Cybersecurity**, Springer, v. 8, n. 1, p. 17, 2025. Disponível em: <https://doi.org/10.1109/InC457730.2023.10263010>. Citado na página 16.

Apêndices

APÊNDICE A – Prompt Utilizado nas LLMs

Como citado no capítulo 4, no item 4.3.2 o *prompt* foi estruturado em quatro seções: **(a)** contextualização teórica, **(b)** descrição dos dados, **(c)** direcionamento da tarefa e **(d)** definição das saídas esperadas. Abaixo seguem os *prompts* utilizados.

(a) *Prompt* - Contextualização teórica

Estou desenvolvendo um novo dataset para análise de ameaças internas, baseado na abordagem da Zero Trust Architecture (ZTA). A ZTA é um modelo de segurança que parte do princípio de que nenhuma entidade, interna ou externa, deve ser automaticamente confiável. Em vez disso, o acesso a recursos deve ser continuamente avaliado com base em identidade, contexto e comportamento. Seus pilares incluem verificação contínua, privilégio mínimo, segmentação e monitoramento contínuo.

(b) *Prompt* - Descrição dos dados

Os dados utilizados foram extraídos do repositório público disponível em: <https://kilthub.cmu.edu/articles/dataset/Insider_Threat_Test_Dataset/12841247>. A partir dele, selecionei e estratifiquei logs de três fontes: (a) Logon: sessões de autenticação e seus horários; (b) Dispositivos: conexões e desconexões de dispositivos utilizados pelos usuários; (c) HTTP: acessos realizados a páginas da web, incluindo categorias de sites, presença de termos sensíveis e padrões de horário. Esses dados foram integrados e transformados em um único CSV (anexado). Cada linha representa um único acesso realizado por um usuário, enriquecido com informações agregadas como: número de logons, logons fora do horário comercial, dispositivos distintos utilizados, sessões sem desconexão, acessos HTTP em horários atípicos, e visitas a sites com conteúdo sensível ou de compras.

(c) *Prompt* - Direcionamento da tarefa

Quero que você atue como um analista de risco baseado nos princípios da ZTA, em especial: (a) Verificação contínua e contextual: identifique desvios de comportamento com base em horário, dispositivo, tipo de acesso e padrões típicos; (b) Privilégio mínimo e microsegmentação: avalie se houve acesso a conteúdos ou recursos que, do ponto de vista comportamental, não condizem com o perfil normal de um colaborador; (c) Integridade de

sessão e dispositivos confiáveis: identifique padrões que indiquem possíveis movimentações laterais ou uso indevido de dispositivos.

(d) Definição das saídas esperadas

Sua tarefa é: (a) Analisar os dados e identificar o que pode ou não ser considerado uma ameaça interna, com base nos campos presentes; (b) Explicar os critérios de análise e como mapeou o risco, indicando as métricas, limiares ou heurísticas utilizadas para classificar os comportamentos; (c) Gerar um código Python que implemente sua lógica de análise, classificando cada linha do dataset como de Alto, Médio ou Baixo risco, adicionando uma nova coluna chamada "risco" ao CSV original e, (d) A explicação deve ser clara, técnica e justificar por que cada comportamento foi considerado suspeito ou não. Esta análise será usada para comparar o desempenho de diferentes LLMs na interpretação de comportamentos suspeitos no contexto de segurança corporativa baseada em ZTA.

APÊNDICE B – Artefatos do Trabalho

Todos os arquivos gerados nesta pesquisa, bem como os códigos utilizados para processamento, análise e experimentação, estão disponíveis em repositório público no GitHub: https://github.com/radharanisr/insider-threats.git .

O repositório foi organizado de modo a garantir a reprodutibilidade dos experimentos, permitindo que outros pesquisadores acessem os datasets derivados, consultem os scripts desenvolvidos e reutilizem as soluções apresentadas em diferentes contextos.

Além do repositório no GitHub, o *dataset* ZTA-RAD também foi disponibilizado no Mendeley Data, podendo ser acessado em:

`https://data.mendeley.com/datasets/d6hvsrxdts/1`