UNIVERSIDADE VIRTUAL DO ESTADO DE SÃO PAULO

Cintia Izumi Shinoda
Cristiano Gois
Fernando Miguel Escribano Martinez
Jordana Barcala
Juliana de Almeida Gonçalves
Pedro Henrique Faria Cruz
Rogério Gonçalves da Silva
Willy Paulino de Oliveira Gomes

Análise da base de dados IoT-23: Origem, aplicações, aprendizado de máquina e plataforma de visualização

São Paulo - SP
2025

UNIVERSIDADE VIRTUAL DO ESTADO DE SÃO PAULO

Análise da base de dados IoT-23: Origem, aplicações, aprendizado de máquina e plataforma de visualização

Relatório Técnico-Científico apresentado na disciplina de Projeto Integrador IV para os cursos de Bacharelado em Engenharia de Computação e Bacharelado em Ciência de Dados da Universidade Virtual do Estado de São Paulo (UNIVESP).

São Paulo - SP
2025

SHINODA, Cintia Izumi; GOIS, Cristiano; MARTINEZ, Fernando Miguel Escribano; BARCALA, Jordana; GONCALVES, Juliana de Almeida; CRUZ, Pedro Henrique Faria; SILVA, Rogério Gonçalves da; GOMES, Willy Paulino de Oliveira. **Análise da base de dados IoT-23: Origem, aplicações, aprendizado de máquina e plataforma de visualização.** Relatório Técnico-Científico. Engenharia da Computação e Ciência de Dados – **Universidade Virtual do Estado de São Paulo**. Tutor: Iolanda Alves Roque da Fonseca. Parque Bristol, Aricanduva, Jaguaré, Parque Novo Mundo e Jardim Paulistano, 2025.

RESUMO

O presente relatório parcial apresenta os resultados intermediários do Projeto Integrador IV, que tem como foco a aplicação de aprendizado de máquina na análise de tráfego de dispositivos da Internet das Coisas (*IoT*). A rápida expansão desse ecossistema traz benefícios em áreas como automação, saúde e cidades inteligentes, mas também intensifica riscos relacionados à segurança cibernética. A partir da base *IoT-23*, composta por cenários reais com tráfego benigno e malicioso, o projeto desenvolveu um processo estruturado de consolidação, pré-processamento e análise exploratória dos dados, envolvendo estatísticas descritivas, identificação de padrões temporais e avaliação de protocolos predominantes. Além disso, foram aplicados algoritmos supervisionados, como *Random Forest* e *Gradient Boosting*, que alcançaram acurácia superior a 98% na classificação de acessos benignos e maliciosos. Paralelamente, iniciou-se a construção de uma plataforma de visualização interativa, baseada em Flask e tecnologias de front-end, que permite a visualização dinâmica dos resultados e a predição em tempo real de novos registros. Nesta etapa, portanto, o trabalho evidencia tanto a viabilidade técnica da proposta quanto sua relevância prática para a segurança em ambientes *IoT*, consolidando a base metodológica e tecnológica que será expandida na versão final.

PALAVRAS-CHAVE: Internet das coisas; Segurança cibernética; *IoT-23*; Aprendizado de máquina; Visualização interativa.

SUMÁRIO

1 INTRODUÇÃO

A chamada Internet das Coisas (IoT) consolidou-se como uma das principais áreas de inovação tecnológica do século XXI, trazendo novas possibilidades para a integração entre dispositivos, pessoas e sistemas. Com a popularização de sensores, câmeras, dispositivos móveis e eletrodomésticos conectados, cresce também a preocupação com a segurança desses equipamentos. Esses dispositivos, muitas vezes com baixo poder de processamento e recursos de proteção limitados, tornam-se alvos frequentes de ataques cibernéticos que exploram vulnerabilidades e criam riscos tanto para usuários quanto para organizações.

Nesse cenário, a utilização de bases de dados reais sobre o tráfego desses equipamentos em diferentes situações é fundamental para o desenvolvimento de soluções de segurança. Um exemplo importante é o conjunto de dados IoT-23, desenvolvido pela Universidade Técnica Tcheca, que reúne registros legítimos e maliciosos em 23 cenários distintos.

O presente trabalho tem como objetivo explorar esse material, aplicando técnicas de ciência de dados e aprendizado de máquina para detectar e classificar tráfego malicioso, além de desenvolver uma plataforma de visualização que integre análise exploratória e predição em tempo real.

2 DESENVOLVIMENTO

2.1 OBJETIVOS

2.1.1 OBJETIVO GERAL

Aplicar algoritmos de aprendizado de máquina para detecção e classificação de tráfego malicioso em dispositivos *IoT* a partir da base de dados *IoT-23*, integrando os resultados em uma plataforma de visualização interativa para análise e visualização.

2.1.2 OBJETIVOS ESPECÍFICOS

- Consolidar e pré-processar os arquivos do *IoT-23*, garantindo uniformidade e consistência nos rótulos;
- Realizar análise exploratória dos dados, destacando distribuições de tráfego, categorias de ataques, protocolos e padrões temporais;
- Treinar e avaliar modelos supervisionados, com destaque para *Random Forest* e *Gradient Boosting*, comparando métricas de desempenho;
- Desenvolver uma plataforma baseada em HTML/CSS baseada em Flask que permita a visualização dinâmica dos resultados e a classificação em tempo real de novas conexões; e

Demonstrar a viabilidade prática da integração entre análise de dados, aprendizado de máquina e plataformas de visualização.

2.2 JUSTIFICATIVA E DELIMITAÇÃO DO PROBLEMA

A rápida expansão do IoT trouxe benefícios significativos em áreas como automação residencial, cidades inteligentes, saúde e indústria. Entretanto, esse crescimento também ampliou a superfície de ataque das redes, expondo usuários e organizações a riscos de segurança cibernética cada vez mais sofisticados. Os dispositivos conectados, geralmente projetados com baixo poder computacional e recursos de proteção limitados, acabam se tornando vulneráveis a agentes maliciosos, comprometendo não apenas a privacidade individual, mas também a integridade de sistemas críticos.

A escolha do conjunto de dados IoT-23 como base de estudo se justifica por reunir cenários reais de tráfego benigno e malicioso, possibilitando a avaliação de técnicas de análise e classificação com potencial de aplicação prática. O problema central pode ser definido da seguinte forma: como aplicar algoritmos de aprendizado de máquina para identificar e classificar tráfego malicioso em dispositivos conectados, de forma a apoiar a criação de soluções interativas de monitoramento e visualização?

Como sugestão para o trabalho completo, é possível destacar que a solução aqui desenvolvida poderia futuramente ser adaptada em ferramentas de monitoramento aplicáveis a empresas, residências ou instituições locais, ampliando o impacto prático da proposta.

Além disso, é relevante mencionar o aspecto cultural relacionado à segurança digital. Em um cenário em que tecnologias conectadas se tornam cada vez mais presentes no cotidiano, a conscientização da sociedade sobre os riscos envolvidos e a promoção de uma cultura de proteção cibernética são elementos fundamentais. Projetos como este contribuem não apenas para o avanço tecnológico, mas também para reforçar a importância da educação digital como estratégia de prevenção.

2.3 FUNDAMENTAÇÃO TEÓRICA

A Internet das Coisas é um campo de estudo interdisciplinar que combina computação, redes de comunicação e engenharia de sistemas embarcados. De acordo com autores como Chaffey e Ellis-Chadwick (2019), a *IoT* é parte de um movimento mais amplo de transformação digital, no qual dados gerados por dispositivos interconectados passam a ter papel estratégico em setores como indústria, saúde, transporte e consumo.

No contexto da segurança, Kotler *et al.* (2017) destacam que a ausência de padrões consolidados para a proteção de dispositivos *IoT* contribui para a exposição a ameaças. Estudos recentes identificam ataques baseados em *botnets*, que exploram falhas em câmeras *IP*, roteadores e sensores, transformando-os em agentes de ataques coordenados.

A base *IoT-23* é um exemplo importante de iniciativa voltada ao estudo de segurança em *IoT*, reunindo 23 cenários com tráfego legítimo e malicioso. Ela permite avaliar diferentes categorias de ataques, como *Port Scan*, *DDoS*, Okiru e Gafgyt, além de tráfego benigno. A predominância de dados maliciosos (84% do conjunto consolidado) evidencia a importância de técnicas adequadas de balanceamento e classificação.

Do ponto de vista de ciência de dados, algoritmos supervisionados de classificação são amplamente empregados em problemas semelhantes. O *Random Forest*, descrito por Breiman (2001), utiliza múltiplas árvores de decisão *ensemble* para aumentar a precisão e reduzir o risco de *overfitting*. Já o *Gradient Boosting*, segundo Friedman (2002), constrói modelos de forma sequencial, otimizando erros residuais e alcançando alta performance, ainda que com maior custo computacional.

Além disso, a análise exploratória de dados (*EDA*) desempenha papel essencial no processo, permitindo a identificação de padrões, correlações e comportamentos anômalos. Como defendem Turban *et al.* (2018), a *EDA* não apenas prepara os dados para o aprendizado de máquina, mas também fornece insights valiosos para a interpretação dos resultados.

2.4 METODOLOGIA

A metodologia adotada neste projeto foi orientada por uma abordagem iterativa, unindo práticas de *Design Thinking* e *Scrum* para organização do trabalho em ciclos curtos, com entregas parciais e revisões constantes.

Os procedimentos técnicos podem ser divididos em quatro etapas principais:

- Pré-processamento dos dados;

- Consolidação da base *IoT-23* em arquivos CSV e Parquet;
- Normalização de rótulos e padronização de atributos numéricos e categóricos;
- Tratamento de valores ausentes e inconsistências; e
- Criação de subconjuntos balanceados para treino e teste.

- Análise Exploratória de Dados (*EDA*):

- Estatísticas descritivas para compreender a distribuição de tráfego benigno e malicioso;
- Identificação das categorias de ataques mais frequentes (como *Port Scan*, Okiru e *DDoS*);
- Estudo dos protocolos predominantes e de padrões temporais de ocorrência; e
- Visualização gráfica com ferramentas como Matplotlib e Seaborn.

- Treinamento de algoritmos supervisionados:

- Seleção de dois modelos de aprendizado de máquina: *Random Forest* e *Gradient Boosting*;
- Definição das variáveis de entrada (características numéricas e categóricas);
- Validação por métricas como acurácia, precisão, recall e ROC-AUC; e
- Comparação entre os modelos para escolha da solução mais robusta.

- Implementação da plataforma:

- Utilização do Flask no *backend*, integrado a bibliotecas Python de análise de dados;
- Desenvolvimento de interfaces em HTML, CSS e Chart.js para visualização interativa;

- Criação de formulários para permitir inserção manual de dados e consulta em tempo real; e
- Estruturação de rotas para visualização de estatísticas, gráficos e predição automática com os modelos treinados.

2.5 RESULTADOS PRELIMINARES: SOLUÇÃO INICIAL

Com base nos procedimentos descritos na metodologia, já foi possível alcançar resultados parciais relevantes, que demonstram a viabilidade da proposta. A análise exploratória revelou que aproximadamente 84% do tráfego consolidado era malicioso, sendo a categoria de *Port Scan* a mais frequente, seguida pelos ataques Okiru e *DDoS*. Além disso, a investigação dos protocolos confirmou a predominância do TCP nos cenários de ataque e identificou padrões temporais característicos em determinados horários do dia, reforçando a existência de comportamentos recorrentes em atividades maliciosas.

Na etapa de aprendizado de máquina, os modelos aplicados apresentaram desempenho expressivo. O *Random Forest* alcançou 98,79% de acurácia, enquanto o *Gradient Boosting* obteve 98,53%, ambos com área sob a curva ROC acima de 0,99. Esses números atestam a qualidade da base *IoT-23* para a tarefa de classificação e a eficiência dos algoritmos escolhidos para diferenciar tráfego benigno e malicioso.

A plataforma de visualização prototipada consolidou os avanços obtidos até esta etapa. A aplicação reúne, em uma interface intuitiva, um painel inicial com estatísticas resumidas, gráficos interativos que permitem explorar os dados de acordo com rótulos, protocolos e serviços, e um formulário de inserção de parâmetros para a previsão em tempo real. Essa última funcionalidade possibilita classificar novas conexões como benignas ou maliciosas a partir do modelo *Random Forest*, oferecendo ao usuário uma experiência prática e direta com os resultados da pesquisa.

Assim, mesmo em caráter parcial, a solução inicial já demonstra o funcionamento do sistema como protótipo operacional, constituindo uma base sólida para futuras melhorias e expansões que deverão ser realizadas na versão final do trabalho.

REFERÊNCIAS

BREIMAN, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

CIOS, Krzysztof J.; PEDRYCZ, Witold; SWINIARSKI, Roman W.; KURTZ, Lisa A. Data Mining: A Knowledge Discovery Approach. New York: Springer, 2007.

CHAFFEY, D., & ELLIS-CHADWICK, F. (2019). Digital marketing: Strategy, implementation, and practice (7th ed.). Pearson Education Limited.

CHAWLA, N. V. *et al.* SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, v. 16, p. 321-357, 2002.

CRESWELL, J. W. (2014). Investigação qualitativa e projeto de pesquisa: escolhendo entre cinco abordagens (3. ed.). Porto Alegre: Penso.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.

GARCIA, S. *et al.* An empirical evaluation of botnet behavior and detection. Stratosphere Laboratory, Czech Technical University in Prague, 2020. p. 2–7. Disponível em: https://www.stratosphereips.org/datasets-iot23. Acesso em: 26 jul. 2025.

GIBBONS, S. (2016). Design Thinking 101. Disponível em: https://www.nngroup.com/articles/design-thinking/?platform=hootsuite. Acessado em 26 de setembro de 2024.

HAN, J.; PEI, Jian; KAMBER, Micheline. Data Mining: Concepts and Techniques. 4. ed. Cambridge, MA: Morgan Kaufmann, 2022.

HAN, X. *et al.* IoT Network Intrusion Detection Based on Supervised Machine Learning. IEEE Access, v. 8, p. 217027-217037, 2020. DOI: 10.1109/ACCESS.2020.3042462.

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. (2017). The elements of statistical learning: Data mining, inference, and prediction. Springer.

HE, H.; GARCIA, E. A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, v. 21, n. 9, p. 1263-1284, 2009.

KE, G., MENG, Q., FINLEY, T., WANG, T., CHEN, W., MA, W., … & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.

KOTLER, P., KARTAJAYA, H., & SETIAWAN, I. (2017). Marketing 4.0: Moving from traditional to digital. Wiley.
KORONIOS, N. et al. Machine Learning Techniques for Cybersecurity: A Comprehensive Review. Future Internet, v. 14, n. 3, p. 1-23, 2022. DOI: 10.3390/fi14030100.

KUMAR, A.; SINGH, P.; TYAGI, H. Malicioso traffic classification in IoT using machine learning. Procedia Computer Science, v. 198, p. 800–807, 2022. DOI: 10.1016/j.procs.2022.06.105.

LAUDON, K. C., & LAUDON, J. P. (2020). Management information systems: Managing the digital firm (16th ed.). Pearson.

MECHELLI, A. *et al*. A Comprehensive IoT Traffic Analysis Dataset for Machine Learning. Data in Brief, v. 42, p. 1-10, 2022. DOI: 10.1016/j.dib.2022.108353.

MOUSTAFA, Nour; SLADEK, Ross; TURNBULL, Ben. An evaluation of IoT intrusion detection systems and their impact on network traffic. Journal of Network and Computer Applications, v. 144, p. 19-31, 2019.

PROKHORENKOVA, L., GUSEV, G., VOROBEV, A., DOROGUSH, A. V., & GULIN, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.

SOLOKOVA, M., & LAPALME, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437.

SOMMER, R.; PAXSON, V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. IEEE Symposium on Security and Privacy, p. 26-40, 2010. DOI: 10.1109/SP.2010.11.

TURBAN, E., Pollard, C., Wood, G., & Gill, P. (2018). Information technology for management: On-demand strategies for performance, growth, and sustainability (11th ed.). Wiley.

YAVUZ, T.; CELIK, A.; KAYA, H. A comparative study on anomaly detection for IoT networks based on machine learning algorithms. Computer Networks, v. 197, p. 1–11, 2021. DOI: 10.1016/j.comnet.2021.108281.

ZHU, Y. *et al.* Network Traffic Classification for IoT Devices Using Machine Learning. Sensors, v. 21, n. 1, p. 110-125, 2021. DOI: 10.3390/s21010110.