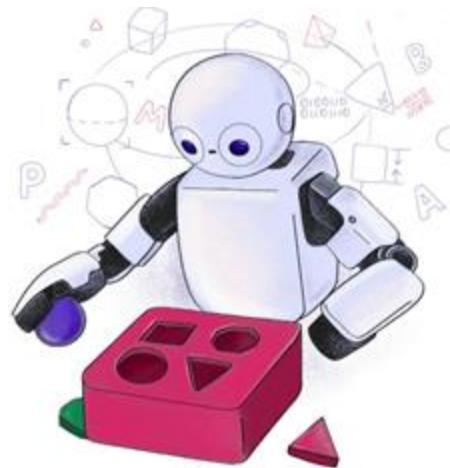


TP558 - Tópicos avançados em Machine Learning:

Detecção de Ameaças IoT: Abordagem Federada e Explicável



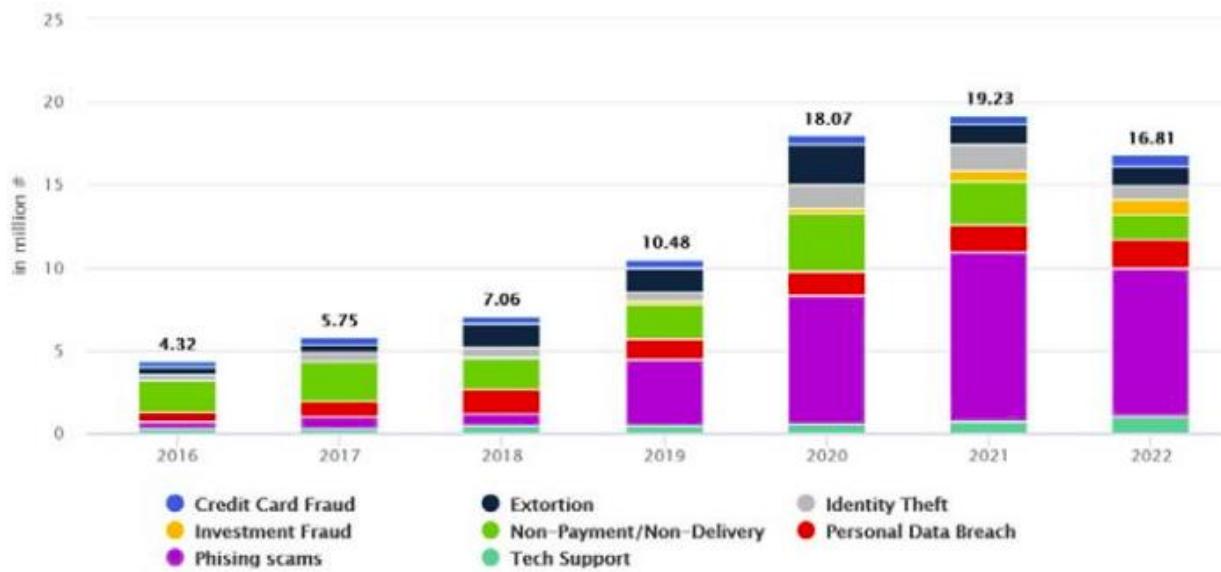
Inatel

Alessandra Carolina Domiciano
Paulo Otavio Luczensky de Souza

Introdução

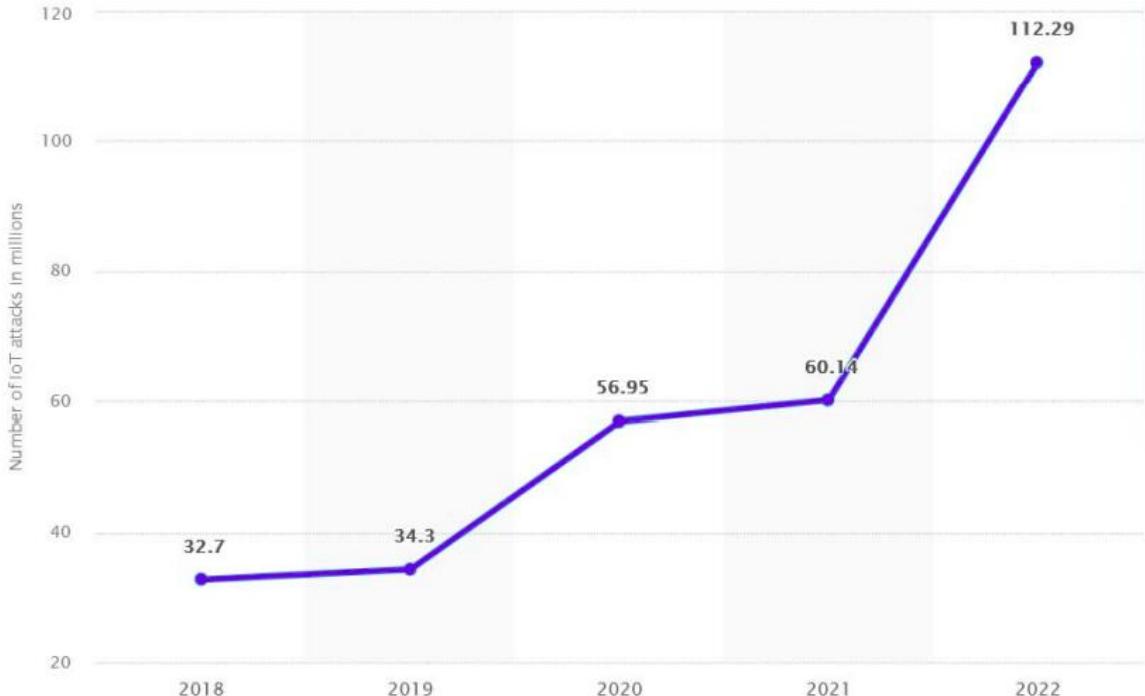
- Nos últimos anos, houve um crescente aumento na quantidade de dados produzidos por aplicações, transmitidos pela rede e armazenados em servidores e data centers.
- Consequentemente, aumentou-se o risco de exposição de dados sensíveis e a incidência de crimes cibernéticos no mundo todo.
- Com isso, nos últimos anos, houve uma maior urgência em prevenir ataques e proteger sistemas críticos.

Introdução - Visão Geral de Ataques Cibernéticos



- Aumento contínuo de ataques entre 2016 e 2021
- Principais tipos de ataques:
 - *Phishing*
 - Fraudes financeiras
 - Golpes com cartão
 - Extorsão
- Queda em 2022, mas forte retomada em 2023 (60+ bilhões de tentativas no Brasil)
- O Brasil gastará cerca de R\$ 104,6 bilhões entre 2025 e 2028 em ciber segurança.

Introdução - Visão Geral dos Ataques em IoT



- Expansão acelerada de dispositivos IoT
- Troca constante de dados → aumento da superfície de ataque
- Crescimento global de ataques IoT entre 2018 e 2022:
 - De **32,7 milhões** → **112,29 milhões**

Referencial Teórico

Referencial Teórico

Este estudo baseou-se em projetos que:

- **Integram segurança desde a concepção**, seguindo a visão de que software é o elemento central da Indústria 4.0 e que IoT/CPS exigem proteção embutida no próprio design.
- **Aplicam XAI para aumentar a transparência na detecção de ataques**, reduzindo a opacidade dos modelos de ML e tornando decisões mais confiáveis em ambientes distribuídos.
- **Buscam equilibrar desempenho, interpretabilidade e eficiência**, usando técnicas leves (como redução de atributos + SHAP) a fim de viabilizar a defesa em dispositivos IoT com baixa capacidade computacional.

Referencial Teórico

Artigo: *The Role of Software Engineering in Industry 4.0* — Samir Lemeš (2020)

Objetivo: Destacar o papel central da engenharia de software como elemento habilitador da Indústria 4.0, evidenciando sua importância para conectividade, automação, análise de dados e segurança.

Conjunto: Aborda a necessidade de projetar sistemas IoT/CPS com segurança desde a concepção, discute desafios de integração com sistemas legados, capacitação técnica, escalabilidade de dados e a lacuna entre teoria e prática em ambientes industriais.

Resultados: O artigo reforça que a Indústria 4.0 depende de software robusto para ser viável, aponta limitações práticas que podem comprometer sua implementação e destaca que segurança e confiabilidade devem ser tratadas como requisitos fundamentais e não opcionais.

Referencial Teórico

Artigo: *Improving IoT Security With Explainable AI: Quantitative Evaluation of Explainability for IoT Botnet Detection* — Kalakoti, Bahsi e Nõmm (2024)

Objetivo: Avaliar métodos de XAI (especialmente LIME e SHAP) aplicados à detecção de botnets em IoT, medindo a qualidade das explicações e sua viabilidade em dispositivos com recursos limitados.

Conjunto: Inclui modelos de classificação treinados em diferentes datasets de ataques IoT, aplicação de técnicas pós-hoc de interpretabilidade e uso de redução de atributos para otimização computacional em dispositivos restritos.

Resultados: O SHAP apresentou explicações mais fiéis, estáveis e coerentes, aumentando a transparência do modelo. O estudo demonstra que combinar desempenho, interpretabilidade e eficiência é essencial para soluções de segurança realmente viáveis em ambientes IoT.

Proposta de projeto

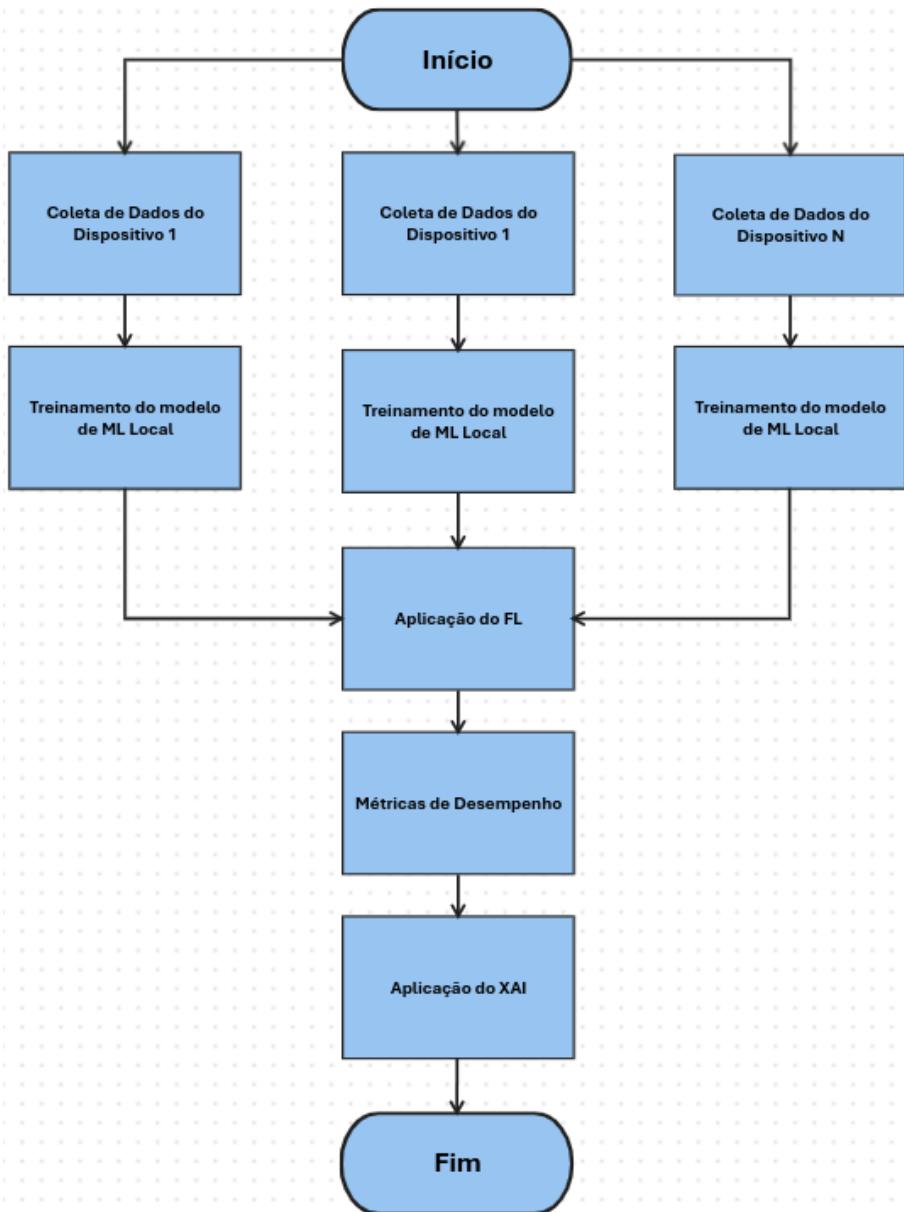
Inatel

Solução Tecnológica

- Desenvolveu-se uma aplicação que utiliza de *Machine Learning* para realizar a detecção desses ataques crescentes em sistemas de IoT.
- Desafios: baixa interpretabilidade e falta de privacidade. Com isso, utilizou-se as técnicas de:
 - *Explainable AI* : aumentando a transparência e compreensão das decisões do modelo.
 - *Federated Learning* : permitindo um treinamento colaborativo sem compartilhamento de dados sensíveis.
- Abordagem proposta: combinar **precisão, explicabilidade e descentralização**.
- Objetivo: tornar sistemas IoT mais seguros, escaláveis e confiáveis.

Metodologia

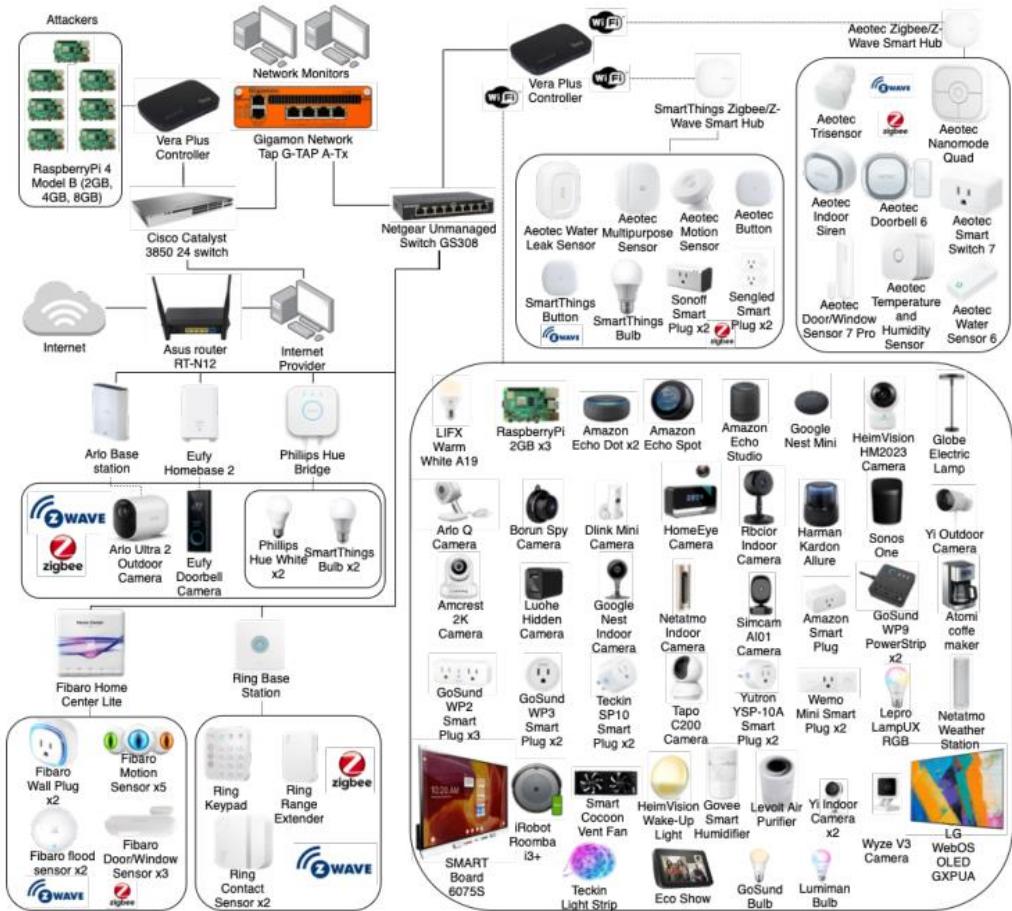
Arquitetura do projeto



- **Coleta distribuída de dados**
 - Cada dispositivo registra atividades normais e ataques cibernéticos.
- **Treinamento local de modelos**
 - Os dispositivos treinam modelos de ML individualmente, evitando o compartilhamento de dados.
- **Aprendizado Federado (FL)**
 - Os modelos locais são agregados para formar um modelo global mais robusto.
- **Otimização automática (Optuna)**
 - Busca dos melhores hiperparâmetros para acelerar experimentos e melhorar desempenho.
- **Avaliação do modelo**
 - Geração de métricas para validar a eficácia na detecção de ataques.
- **Aplicação de XAI (SHAP)**
 - Interpretação das decisões do modelo, destacando a importância de cada atributo.

Dataset

- CICIoT2023:
 - Tráfego capturado em uma rede IoT com 115 dispositivos reais;
 - 33 tipos de ataque organizados em 7 categorias + tráfego benigno;



Dataset

- Para este trabalho, foram filtrados quatro tipos de tráfego: **benigno, DDoS, DoS e reconhecimento.**
- No total, foram utilizadas 664.743 amostras, com a seguinte distribuição:

Classe	Percentual
0	2,5%
1	77,5%
2	18,4%
3	1,6%

Classe	Categoria	Ataque
0	Benign	-
1	DDoS	PSHACK FLOOD ICMP FLOOD TCP FLOOD SYN FLOOD UDP FLOOD SYNONYMOUSIP FLOOD RSTFIN FLOOD SLOWLORIS ICMP FRAGMENTATION ACK FRAGMENTATION UDP FRAGMENTATION HTTP FLOOD
2	DoS	UDP FLOOD TCP FLOOD SYN FLOOD HTTP FLOOD
3	Recon	HOST DISCOVERY VULNERABILITY SCAN PORT SCAN OS SCAN PING SWEEP

Dataset

- Amostras correspondem a fluxos de comunicação processados;
- Cada amostra é descrita por 39 features, derivadas de metadados, estatísticas temporais, comportamento de pacotes e informações de protocolo dos fluxos de comunicação.

Feature	Descrição
Number	Número total de pacotes no fluxo
psh flag number	Valor da flag PSH
IAT	Diferença de tempo em relação ao pacote anterior
fin flag number	Valor da flag FIN
Header Length	Tamanho do cabeçalho
syn flag number	Valor da flag SYN
Rate	Taxa de transmissão de pacotes em um fluxo
ICMP	Indica se o protocolo de rede é ICMP
UDP	Indica se o protocolo de transporte é UDP
HTTPS	Indica se o protocolo de aplicação é HTTPS
rst flag number	Valor da flag RST
AVG	Tamanho médio dos pacotes no fluxo
TCP	Indica se o protocolo de transporte é TCP
Max	Tamanho máximo do pacote no fluxo
ack flag number	Valor da flag ACK
Duration	Tempo de vida (TTL — Time to Live)
syn count	Número de pacotes com flag SYN ativa no mesmo fluxo

XGBoost

Hiperparâmetro	Descrição
eta	Taxa de aprendizado
max_depth	Profundidade máxima da árvore
subsample	Amostragem de linhas por árvore
colsample_bytree	Amostragem de colunas por árvore
lambda	Regularização L2
alpha	Regularização L1

- **Modelo que utiliza ensemble e é baseado em árvores de decisão**, reconhecido pela alta eficiência, precisão e velocidade em tarefas de classificação e regressão.
- **Excelente desempenho com grandes volumes de dados**, incluindo cenários com atributos esparsos e estruturas complexas.
- **Robusto na captura de padrões não lineares**, tornando-se adequado para problemas desafiadores e de alta variabilidade.

Otimização de hiperparâmetros (Optuna)

Hiperparâmetro	Faixa de Valores
eta	0.01 – 0.3
max_depth	6 – 12
subsample	0.6 – 1.0
colsample_bytree	0.6 – 1.0
lambda	10^{-8} – 1.0
alpha	10^{-8} – 1.0

Hiperparâmetro	Valor
eta	0.1673
max_depth	9
subsample	0.83
colsample_bytree	0.72
lambda	0.0124
alpha	0.0019

- **Automação da busca de hiperparâmetros:** O Optuna foi usado para explorar automaticamente combinações de hiperparâmetros do XGBoost, acelerando a identificação das melhores configurações.
- **Avaliação eficiente com pruning:** O framework registrou métricas e interrompeu treinamentos pouco promissores, reduzindo tempo computacional e focando nas combinações mais promissoras.
- **Seleção ótima para o modelo federado:** A busca resultou nos hiperparâmetros que **minimizaram o *mlogloss* (0.8032)**, permitindo aplicar diretamente a melhor configuração ao modelo treinado no FL.

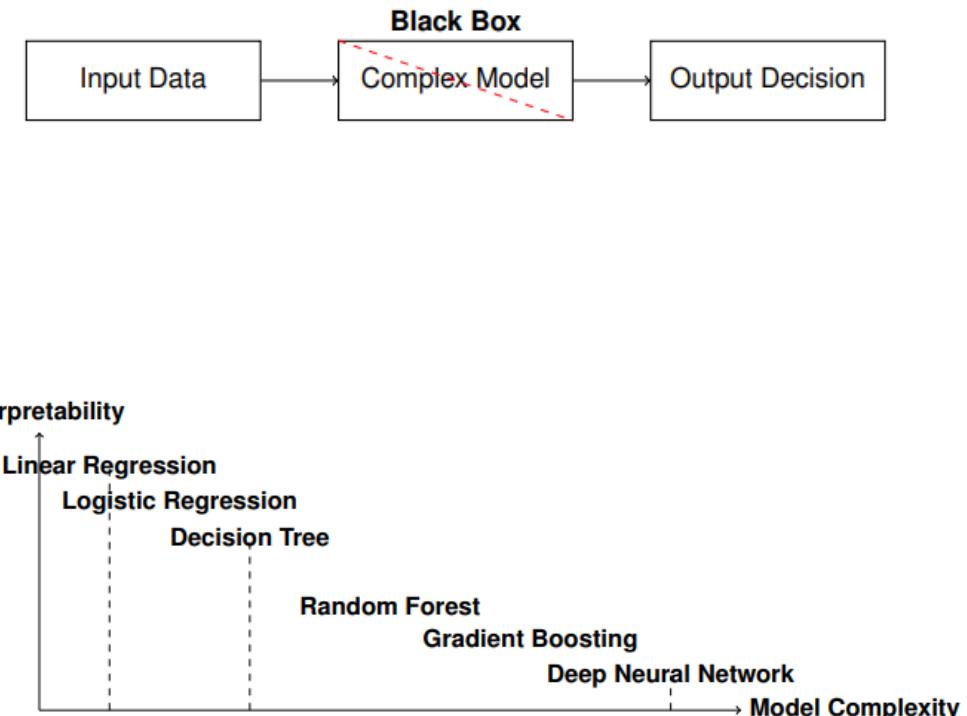
Aprendizado Federado

Parâmetro	Valor
num_clients	4
num_boost_round	50
rounds_per_client	5

- **Treinamento distribuído e descentralizado:** Os modelos são treinados localmente em cada dispositivo, aproveitando dados que permanecem privados e sem necessidade de centralização.
- **Redução de comunicação e maior eficiência:** O FL diminui o tráfego de dados e permite atualizações contínuas do modelo, tornando o processo mais escalável e robusto.
- **Configuração colaborativa do ambiente:** O sistema conta com múltiplos clientes (4), cada um realizando rounds locais (5) dentro de 50 rounds globais, integrando conhecimento distribuído.

IA Explicável (XAI)

- Modelos de aprendizado de máquina estão cada vez mais complexos, o que nos leva ao problema da "caixa preta".
- Em aplicações sensíveis, a **interpretação das decisões do modelo** é requisito para garantir a confiabilidade do sistema.
- Abordagens de XAI permitem **transparência** sobre os mecanismos internos e/ou justificativas locais para cada predição.
- Isso pode ajudar usuários especialistas, usuários finais, desenvolvedores e pesquisadores, auditoria e compliance, etc.



Shapley Additive Explanations (SHAP)

- SHAP é uma abordagem baseada em **Teoria dos Jogos** para explicar a saída de qualquer modelo de *Machine Learning*.
- A técnica atribui um **valor de importância** (valor de Shapley) a cada *feature*, tratando o problema de predição como um **jogo cooperativo**, no qual as *features* são os jogadores e a predição é o prêmio.
- O objetivo é **atribuir de forma justa** a contribuição de cada *feature* para a predição, oferecendo uma interpretação consistente de como cada variável influencia a decisão do modelo.
- Considerado um dos métodos mais **teoricamente fundamentados e estáveis** para explicabilidade local e global.



Resultados

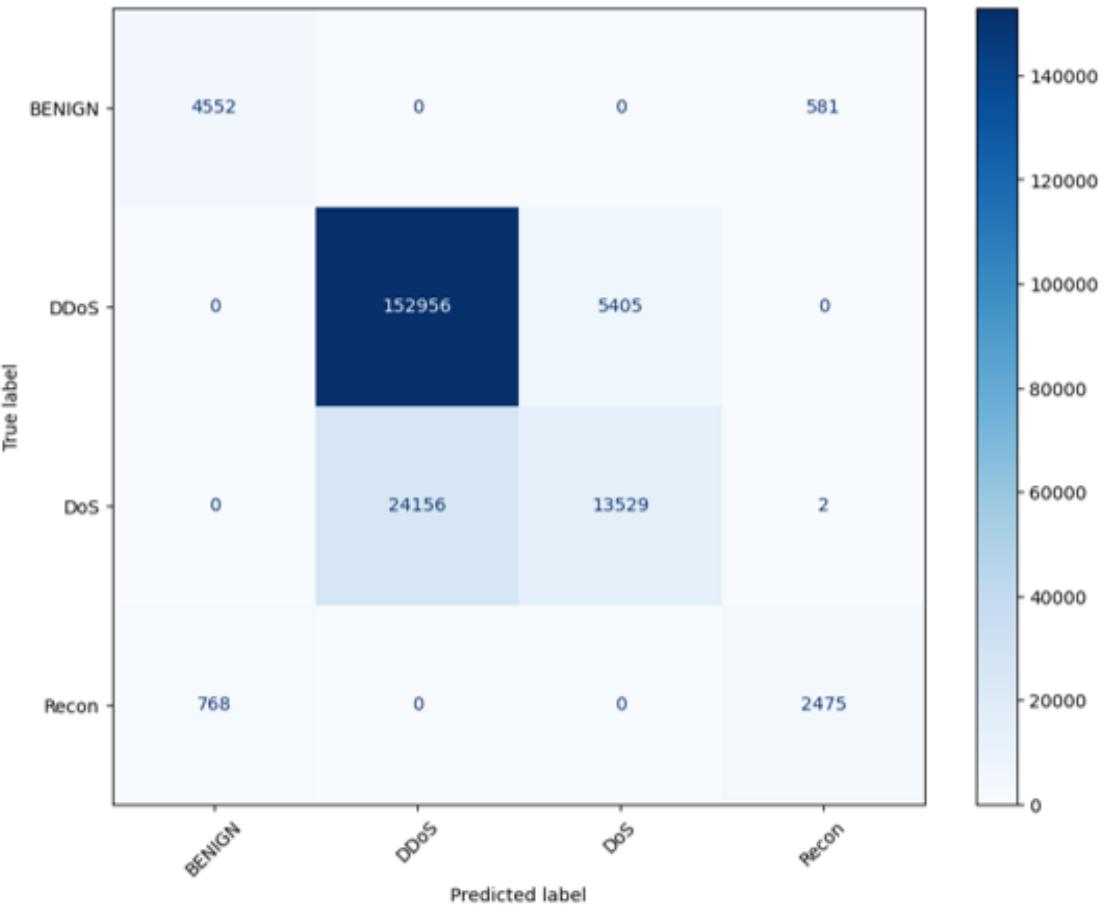
Classificação multiclasse

- As classes 0 e 1 tiveram desempenho alto, com F1-score de 87% e 91%, respectivamente. Isso indica que o modelo consegue identificar bem tráfego benigno e ataques de DDoS.
- A classe 3 apresentou desempenho razoável, com F1-score de 79%.
- Já a classe 2 apresentou desempenho inferior, com F1-score de 78% e recall de 36%, indicando a presença de muitos falsos negativos.
- A acurácia ficou em 85%, o que representa um desempenho razoável para multiclasse.
- Vale lembrar que as classes 1 e 2 são as classes majoritárias e as classes 0 e 3 são as classes minoritárias no dataset.

Classe	Precisão	Recall	F1-score	Suporte
0 (Benign)	0.86	0.89	0.87	5133
1 (DDoS)	0.86	0.97	0.91	158361
2 (DoS)	0.71	0.36	0.48	37687
3 (Recon)	0.81	0.76	0.79	3243
Acurácia			0.8488	
Macro média	0.81	0.74	0.76	204424
Média ponderada	0.84	0.85	0.83	204424

Classificação multiclasse

- Através da matriz de confusão, é possível perceber que praticamente todos os falsos negativos para amostras de DoS são confundidos com tráfego DDoS, o que é compreensível devido a natureza similar de ambos ataques.
- Da mesma forma, os falsos negativos para amostras de DDoS são todos confundidos com tráfego DoS, embora em uma quantidade relativamente pequena.
- Além disso, há algumas amostras de tráfego benigno que são confundidas com ataques de reconhecimento e vice-versa.
- Por fim, não há nenhuma confusão entre tráfego benigno e ataques DDoS e DoS.



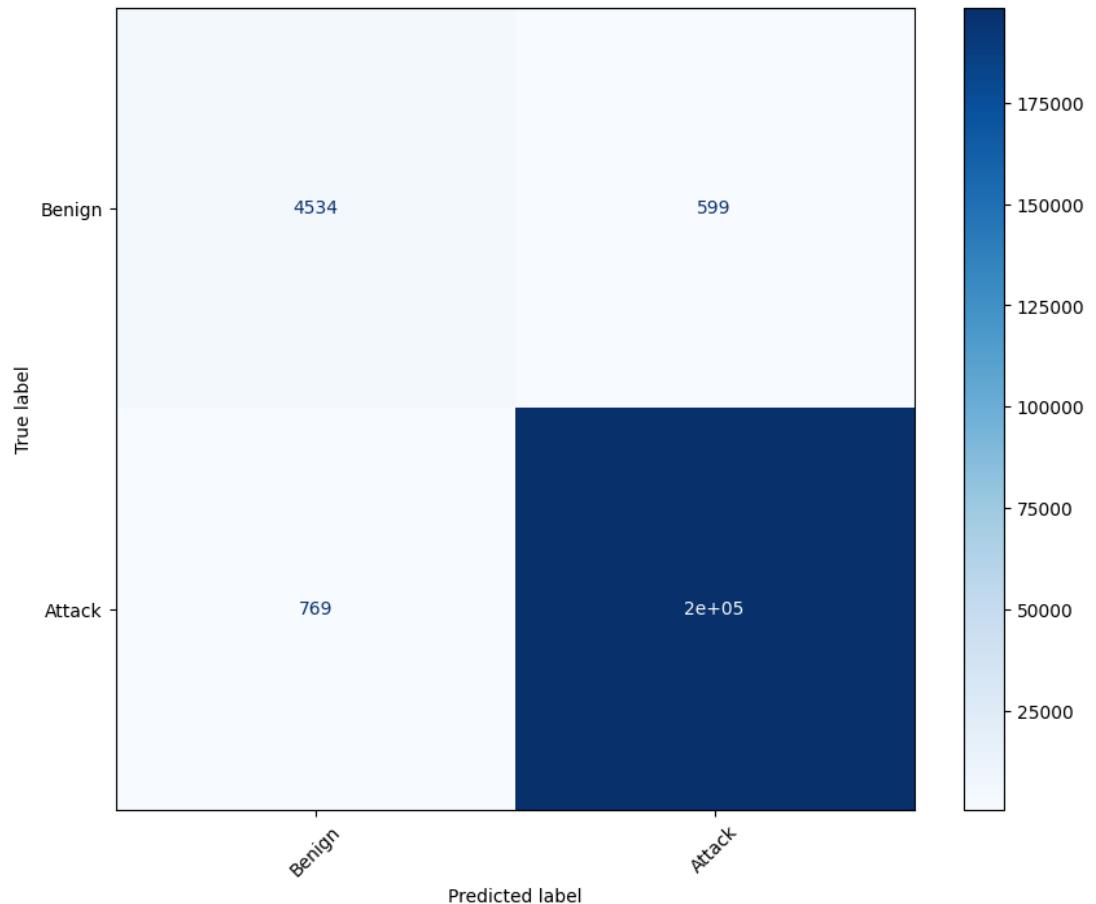
Classificação binária

- O desempenho é substancialmente superior, com acurácia de 99% e F1-score de 100% e 87% para as classes attack e benign, respectivamente.

Classe	Precisão	Recall	F1-score	Suporte
0 (Benign)	0.85	0.88	0.87	5133
1 (Attack)	1.00	1.00	1.00	199291
Acurácia		0.9933		
Macro média	0.93	0.94	0.93	204424
Média ponderada	0.99	0.99	0.99	204424

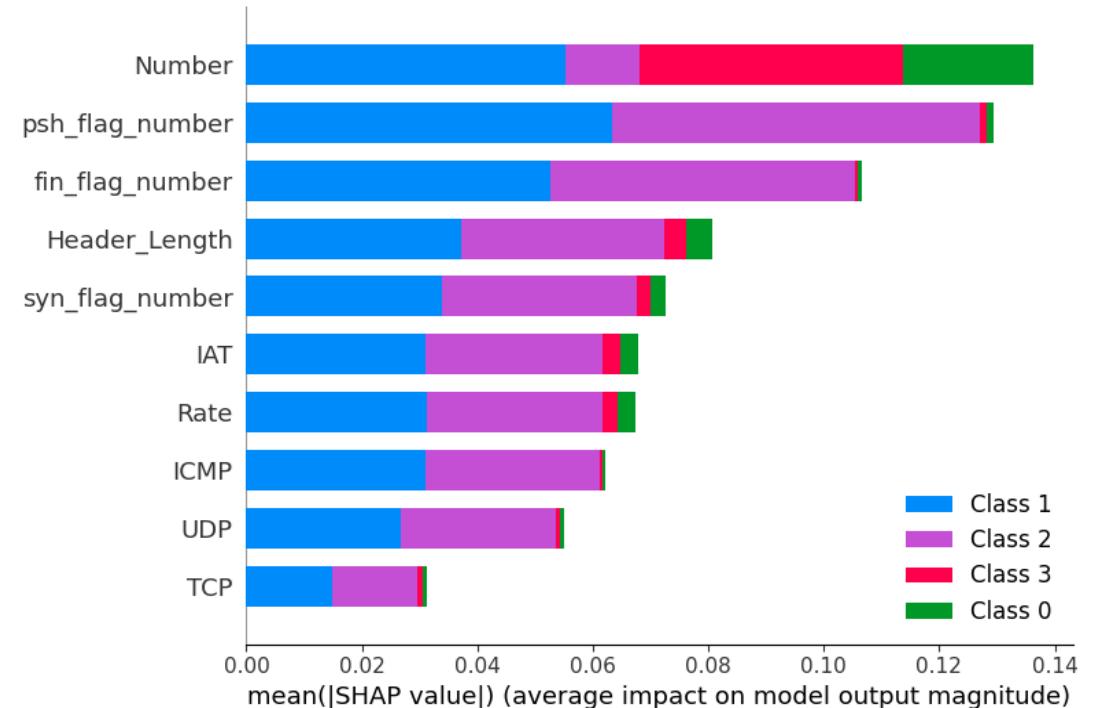
Classificação binária

- A matriz de confusão evidencia que, ao agrupar todos os ataques em uma única categoria, a confusão entre classes é reduzida, uma vez que os erros entre diferentes tipos de ataques deixam de aparecer.
- Em cenários práticos, esse comportamento é geralmente aceitável, pois a distinção entre tráfego benigno e malicioso costuma ser mais relevante do que a classificação detalhada de cada tipo específico de ataque.



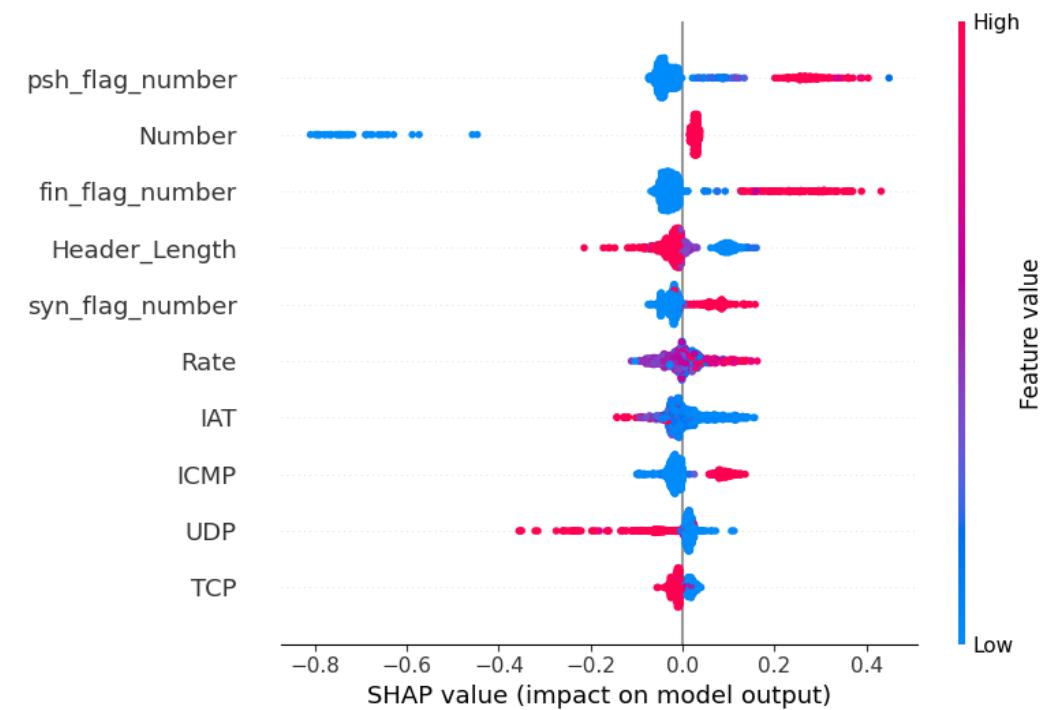
Explicabilidade com SHAP

- Características de fluxo com maior peso nas classificações:
 - número de pacotes
 - valor da flag PSH
 - valor da flag FIN
 - comprimento de cabeçalho
 - valor da flag SYN
 - intervalo entre a chegada de pacotes
 - taxa de transmissão
- Isso demonstra coerência com o comportamento típico de tráfego malicioso, especialmente em ataques de negação de serviço (DoS e DDoS), que tendem a gerar volumes altos de pacotes, intervalos curtos entre transmissões e uso recorrente de flags como SYN, PSH e FIN para sobrecarregar ou manipular o estado da conexão.



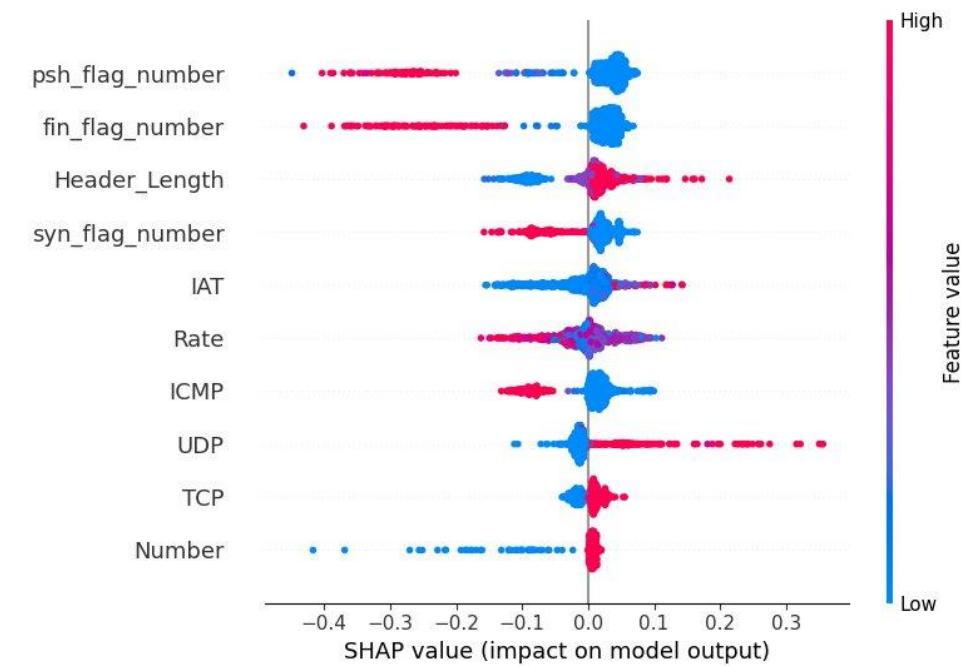
Análise global por classe – DDoS

- As variáveis com maior contribuição positiva incluem as flags PSH, FIN e SYN. Isso pode indicar padrões de manipulação e sinalização atípica no fluxo TCP visando esgotar recursos.
- De forma similar, valores altos de taxa de transmissão e valores baixos de intervalo entre a chegada de pacotes favorecem a predição por essa classe, o que é compatível com o perfil de saturação de ataques de DDoS.
- Em contraste, a variável com maior contribuição negativa é o número de pacotes no fluxo: quanto menor seu valor, menos provável será a classificação de uma instância como DDoS.



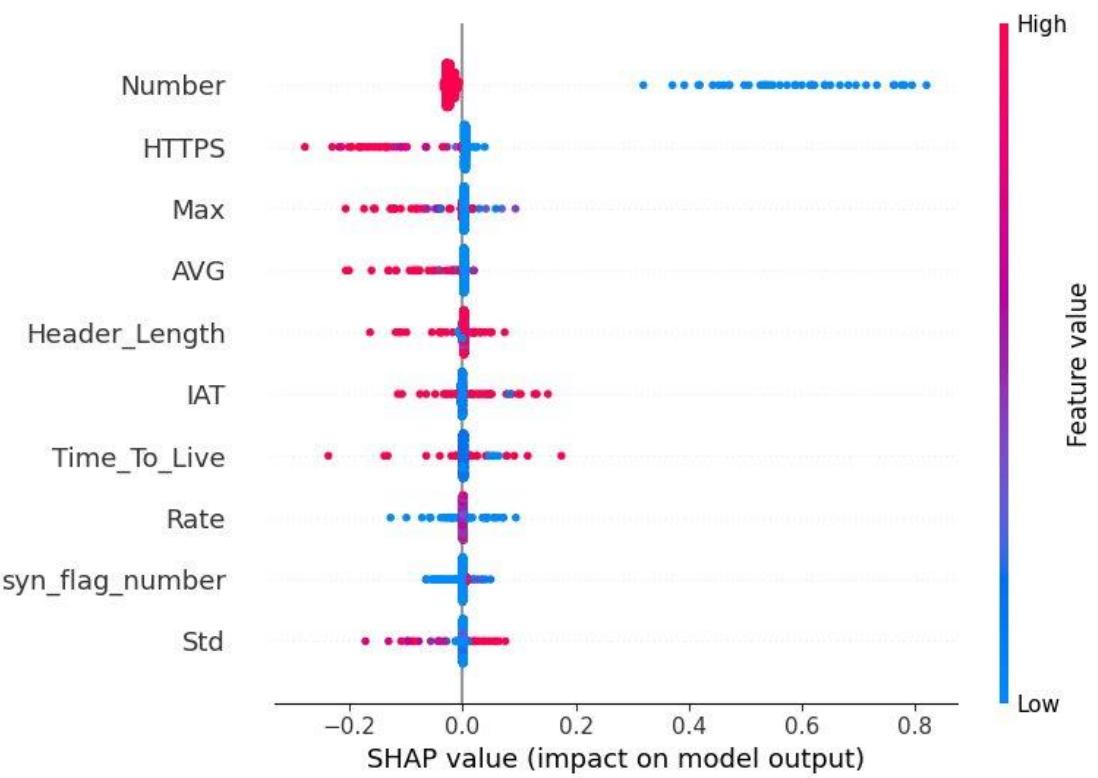
Análise global por classe – DoS

- As variáveis com maior impacto negativo são as flags **PSH** e **FIN**, com valores altos reduzindo a probabilidade de classificação de uma instância como DoS. Isso é coerente com a composição dos ataques constantes no dataset.
- Por outro lado, o valor da flag SYN, a taxa de transmissão e o intervalo entre a chegada de pacotes apresentam contribuições negativas com valores baixos para a última e valores altos para as demais. Isso reflete um padrão impreciso para DoS aprendido pelo modelo, que pode explicar a alta taxa de falsos negativos dessa classe, pois muitas instâncias de DoS acabam sendo absorvidas pela classe DDoS devido à semelhança dos padrões e ao forte desequilíbrio entre as classes.



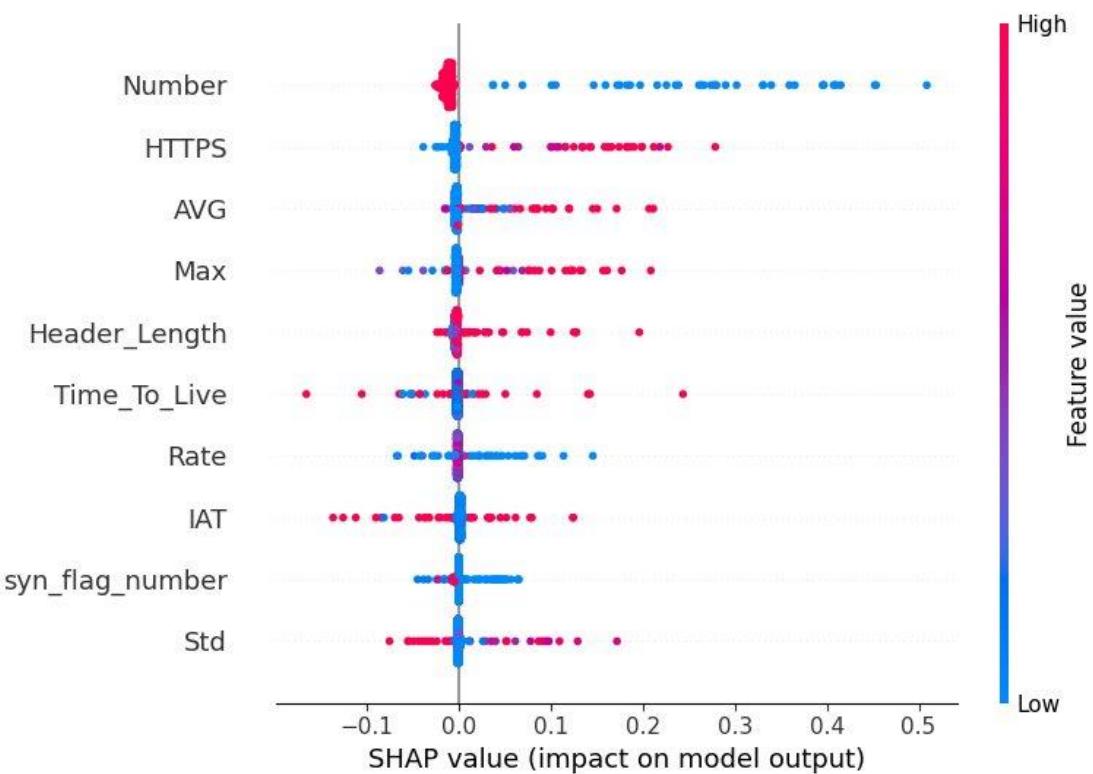
Análise global por classe – Reconhecimento

- A feature com maior impacto positivo é o número de pacotes no fluxo, com valores menores aumentando a probabilidade de classificação como ataque de reconhecimento.
- Já as features com maior impacto negativo são protocolo HTTPS e os tamanhos máximo e médio dos pacotes, com valores elevados reduzindo a probabilidade de escolha pela classe.
- Esses são padrões claramente distintos das classes anteriores e compatíveis com a natureza exploratória, esparsa e de baixo volume típica de ataques de reconhecimento.



Análise global por classe – benigno

- A feature com maior impacto positivo é o número de pacotes no fluxo, com valores menores aumentando a probabilidade de classificação como tráfego benigno.
- Da mesma forma, as features HTTPS e tamanhos médio e máximo de pacotes têm contribuição positiva, mas com valores maiores favorecendo a escolha pela classe.
- Esse é um padrão compatível com o cenário IoT, caracterizado por fluxos curtos e de baixa volumetria.



Conclusão

Conclusão

- Este trabalho apresentou uma aplicação de aprendizado federado para a tarefa de detecção de ameaças no contexto de IoT, utilizando o modelo XGBoost e o dataset CICIoT2023. A análise contemplou tanto o cenário de classificação multiclasses quanto o binário, além da aplicação de SHAP para compreender o comportamento interno do modelo.
- No cenário multiclasses, embora o modelo apresente desempenho razoável, ainda existem limitações relevantes na discriminação entre tipos específicos de ataques, devido a semelhança entre os padrões de tráfego e ao forte desbalanceamento presente no dataset.
- No cenário binário, o modelo alcança desempenho quase perfeito. A agregação das classes de ataque reduz drasticamente os erros de confusão observados no cenário multiclasses.
- A análise interpretável via SHAP complementou a avaliação quantitativa, revelando as variáveis mais relevantes na discriminação entre as classes e elucidando as causas da principal limitação observada na classificação multiclasses — a sobreposição de características entre DDoS e DoS.
- Esses achados destacam tanto a eficácia quanto os desafios da utilização de aprendizado de máquina para detecção de ameaças, ao mesmo tempo em que demonstram o valor da explicabilidade para compreender e aprimorar soluções baseadas em IA em cibersegurança.

Trabalhos Futuros

Inatel

Trabalhos Futuros

- Como trabalhos futuros, propõe-se expandir a avaliação para outros modelos de aprendizado de máquina, incluindo arquiteturas de Deep Learning capazes de capturar padrões mais complexos do tráfego.
- Além disso, recomenda-se testar a abordagem em diferentes datasets de ameaças cibernéticas, que apresentem conjuntos de features distintos e cenários mais diversificados.
- A aplicação de métodos de XAI nesses novos contextos permitirá comparar modelos e bases de dados não apenas do ponto de vista quantitativo, mas também qualitativo, oferecendo uma compreensão mais profunda sobre como cada modelo distingue as classes de tráfego e revelando quais atributos são mais relevantes para a caracterizar diferentes tipos de ataques.

Perguntas?

Inatel

Referências

- [1] Samir Lemes. “The Role of Software Engineering in Industry 4.0”. Em: *ResearchGate* (2023). Disponível em ResearchGate – ID de publicação 374532017.
- [2] Vitor Fantin. *Cibersegurança em Alta: Tendências e Investimentos até 2028*. URL: <https://www.linkedin.com/pulse/ciberseguran%C3%A7a-em-alta-tend%C3%AAncias-e-investimentos-at%C3%A9-2028-ecake/> (acesso em 20/08/2025).
- [3] R. Kalakoti, H. Bahsi e S. Nõmm. “Improving IoT Security With Explainable AI: Quantitative Evaluation of Explainability for IoT Botnet Detection”. Em: *IEEE Internet of Things Journal* 11.10 (2024), pp. 18237–18254. DOI: 10.1109/JIOT.2024.3360626.
- [4] K. Fatema et al. “Federated XAI IDS: An Explainable and Safeguarding Privacy Approach to Detect Intrusion Combining Federated Learning and SHAP”. Em: *Future Internet* 17.6 (2025), p. 234. DOI: 10.3390/fi17060234.

Referências

- [5] Euclides Carlos Pinto Neto, Sajjad Dadkhah, Raphael Ferreira, Alireza Zohourian, Rongxing Lu e Ali A. Ghorbani. “CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment”. Em: *Sensors* 23.13 (2023), p. 5941. DOI: 10.3390/s23135941. URL: <https://www.mdpi.com/1424-8220/23/13/5941>
- [6] W. Hsieh, Z. Bi e C. et al. Jiang. “A Comprehensive Guide to Explainable AI: From Classical Models to LLMs”. Em: *arXiv preprint* (2024). arXiv: 2412.00800 [cs.LG].

Repositório GitHub

- Link de acesso ao repositório:
- <https://github.com/PauloLuczensky/Projeto-TP-558/tree/main>

Obrigado!