Centro de Informática

UFPE

IEEE Transactions on Information Forensics and Security - 2022
Zebin, T., and Rezvy, S. and Luo, Y.

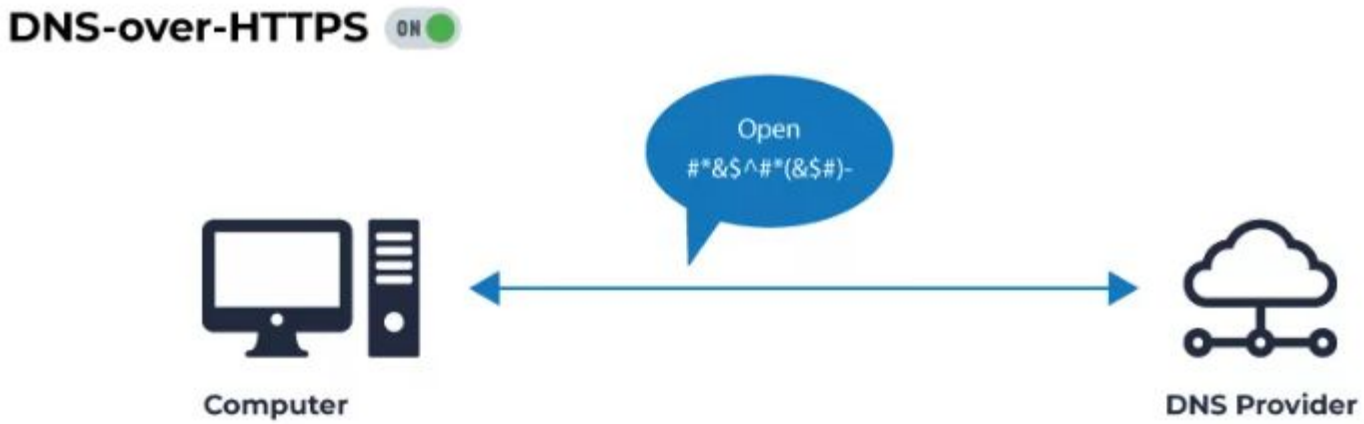
An Explainable AI-Based Intrusion Detection System for DNS Over HTTPS (DoH) Attacks

Alunos:
Camila Barbosa Vieira
Dayane Lira da Silva
José Vinicius de S. Souza

87% sofreram um ou mais ataques

8% a mais em 2021 que no ano anterior

EfficientIP and IDC 2021 Global DNS Threat Report



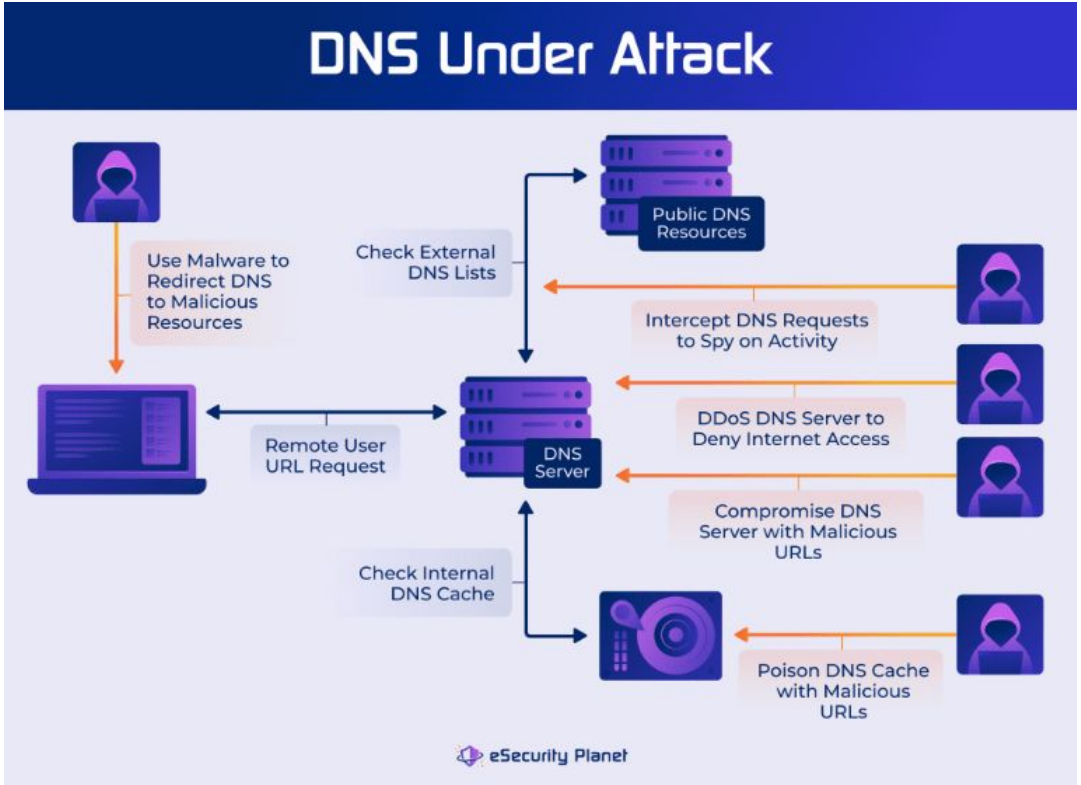
Objetivos

Características de tráfego DoH fáceis de interpretar

Balanced Stacked Random Forest

Visualizações usando métodos de IA explicáveis

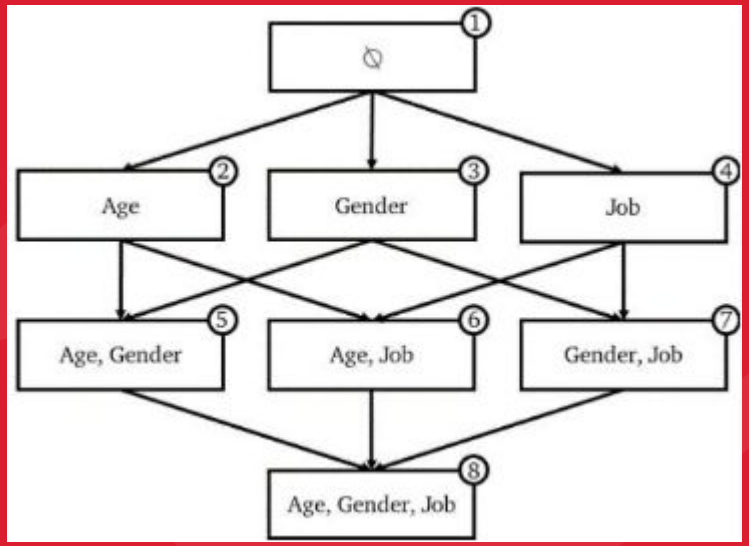
Trabalhos Relacionados



- Análise de tráfego de rede e DNS, blacklist de nomes de domínio e detalhamento do conteúdo da página da Web
- Combina PCA com informações mútuas (MI) para calcular index id
- Técnicas de ML
  - Classificação do tráfego em vez das queries.
  - Pré-processamento, otimização ou métricas pouco claras

Trabalhos Relacionados

Explicabilidade



- Mantendo a explicabilidade em mente, não inclui modelos de aprendizado profundo
- Não é comum insights sobre o comportamento e o raciocínio
- SHAP (SHapley Additive exPlanations) é uma abordagem baseada na teoria dos jogos para explicar a saída de qualquer modelo de ML

Dataset

- *CIRA-CIC-DoHBrw-2020* dataset
- DoH e non-DoH tráfego.
- Non-DoH: acessando diferentes servidores web.
- DoH: Ferramentas de DNS tunnelling e navegadores web como Chrome, Firefox e safari.

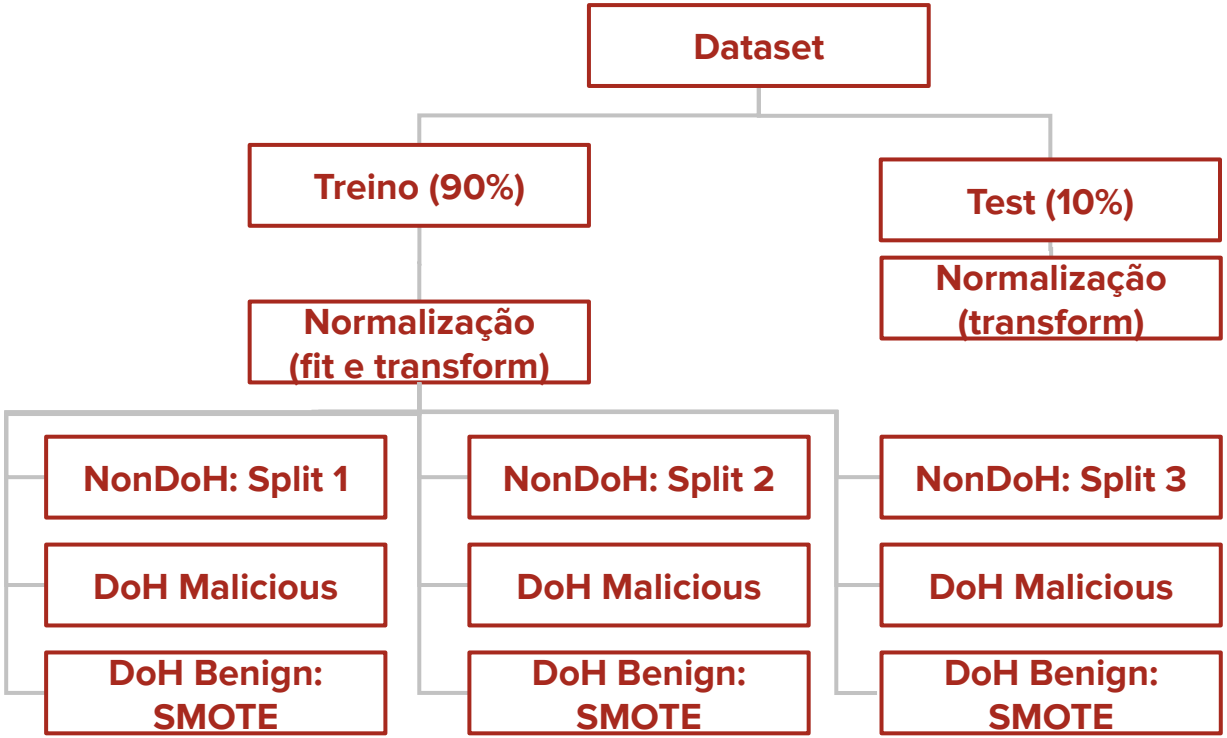
Análise das Características do Tráfego

Categorias:

- *Flow Statistics,*
- *Flow Bytes,*
- *Packet Length,*
- *Packet Time,*
- *Inter-Packet Delay*

- Flow byte e Length distinguem entre DoH malicioso e non-DoH
- Entre benigno e maligno DoH, a variância do último é sempre relativamente alto devido à alternância entre pacotes pequenos e grandes.

Pré-Processamento

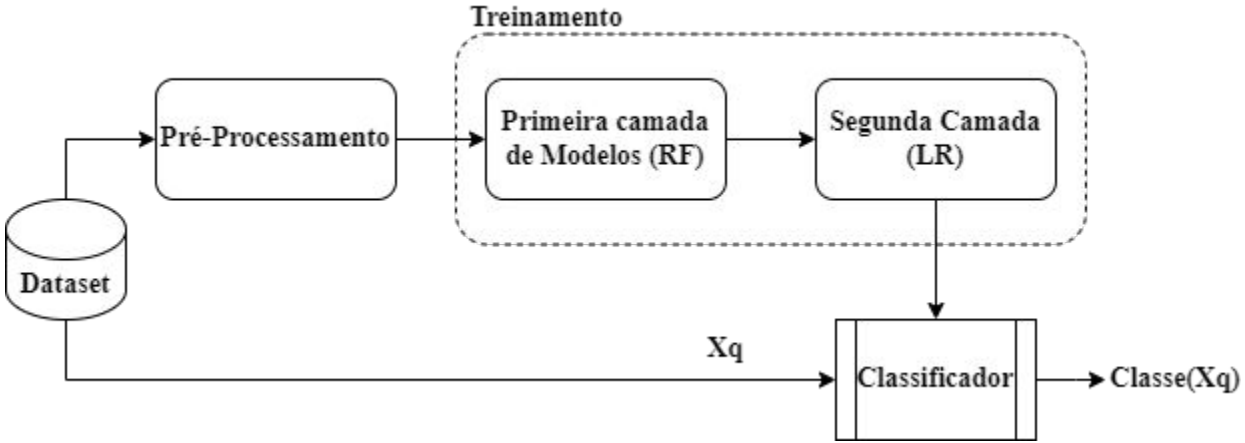


- 45:1:12 - 15:12:12
- Redução do tempo (paralelismo)
- GridsearchCV (CV = 10)

Solução Proposta

Sistema de Detecção de Intrusões

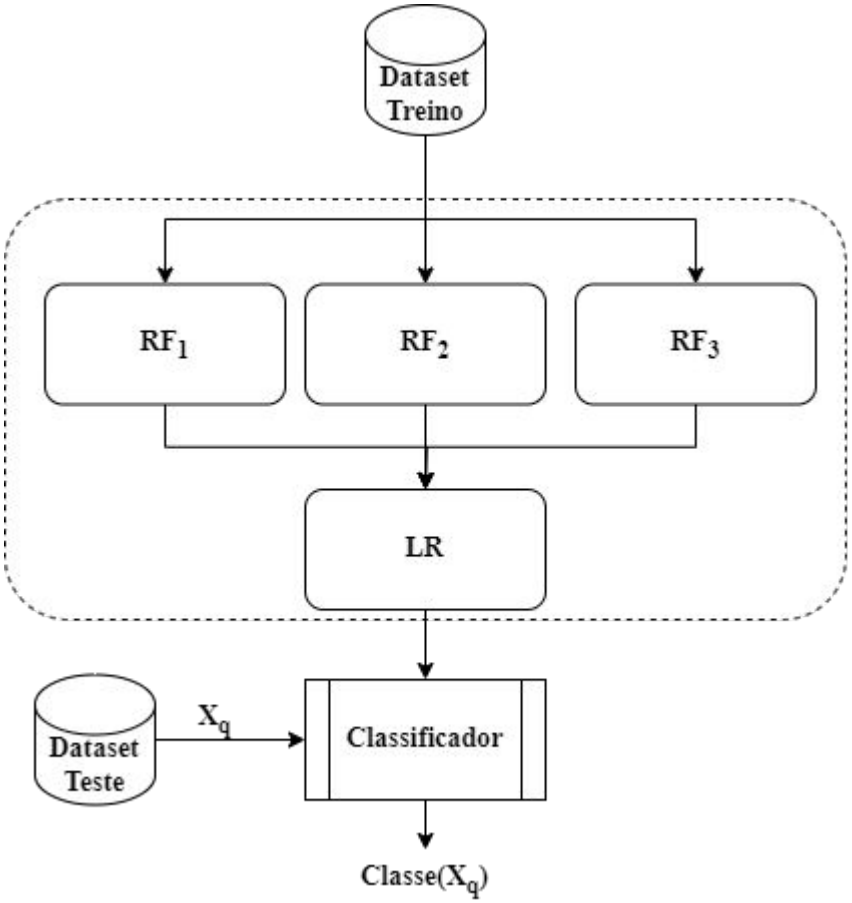
- Utilizando Balanced and Stacked Random Forest (BSRF)
- Focado em DoH



Solução Proposta

BSRF

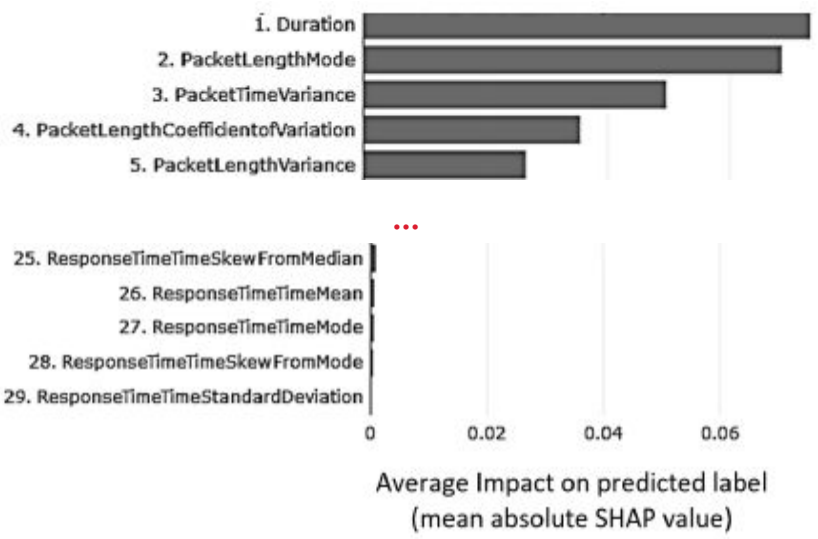
- Cada subconjunto de treino foi usado para treinar uma Random Forest
- As previsões desses três modelos foram combinadas em um meta-classificador de Regressão Logística



Solução Proposta

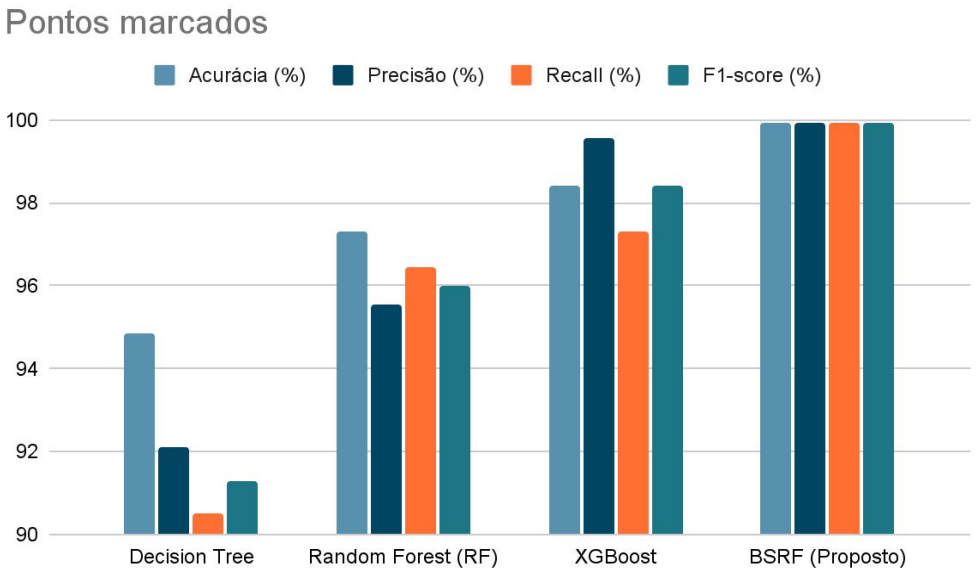
Explicabilidade

- Utilizaram SHAP (SHapley Additive Explanations) para interpretar as decisões do modelo.
- Identificaram as features mais importantes, como duração do fluxo e comprimento dos pacotes.



Resultados

- BSRF superou os outros classificadores em acurácia, precisão e recall
- Redução de tempo de treinamento
- Menos falsos negativos



Conclusões

- A Balanced and Stacked Random Forest (BSRF) é altamente eficaz para detectar ataques DoH.
- A abordagem de divisão balanceada reduziu o tempo de treinamento sem comprometer a precisão.
- O uso de SHAP melhorou a interpretabilidade, permitindo entender decisões do modelo.

- Falsos positivos em Benign-DoH: Algumas amostras benignas foram confundidas com Non-DoH.
- Detecção limitada a ataques conhecidos: O modelo foi treinado apenas com ataques de dns2tcp, DNSCat2 e Iodine.
- Pode não generalizar bem para novos ataques DoH.

Discussão

- Pré-Processamento cuidadoso
- Atenção ao desbalanceamento
- Explicabilidade dos modelos

- **Novas técnicas de balanceamento**
- Diferentes classificadores
- Investigar explicabilidade e performance com um SVM linear

DREBIN

- 129.013 amostras, sendo apenas 5.560 malware
- 22,20 de razão de desequilíbrio

Características extraídas incluem:

- Permissões
- Chamadas de API suspeitas
- Registros de rede
- Componentes do aplicativo

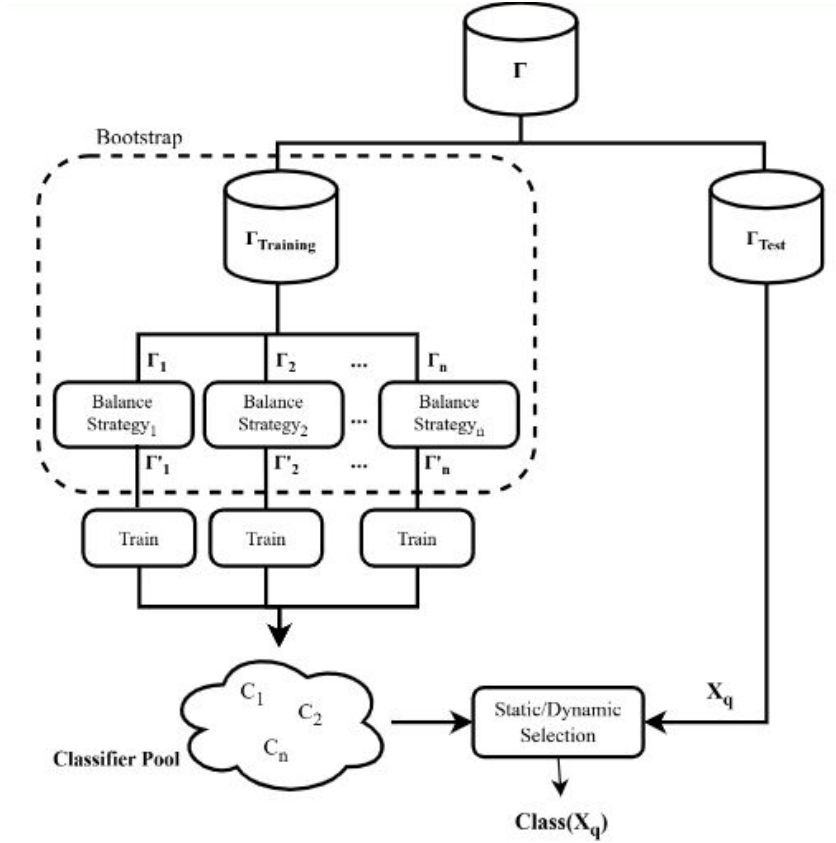
Desequilíbrio de Classes

Falsos negativos são perigosos

Classificadores tradicionais tendem a favorecer a classe majoritária

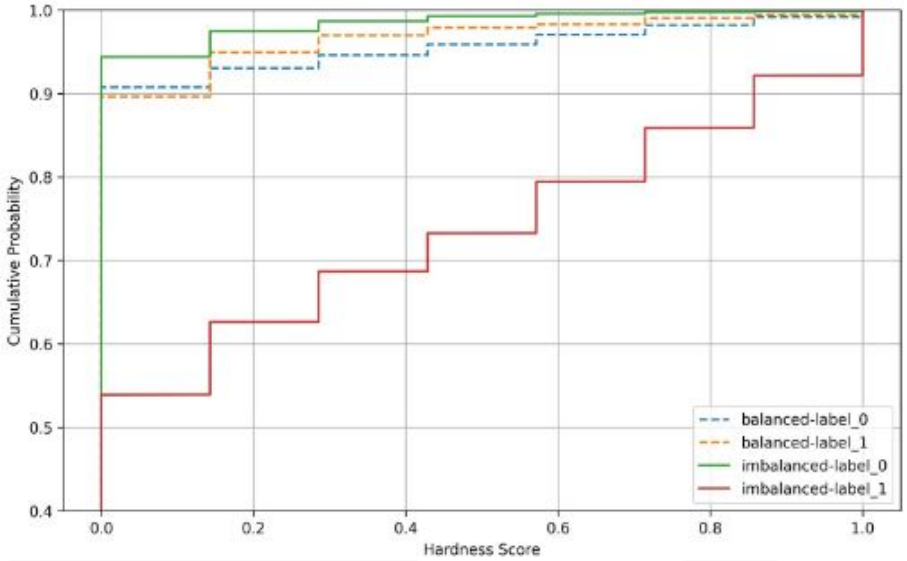
Problema de detecção de malware é inerentemente desbalanceado

Melhoria

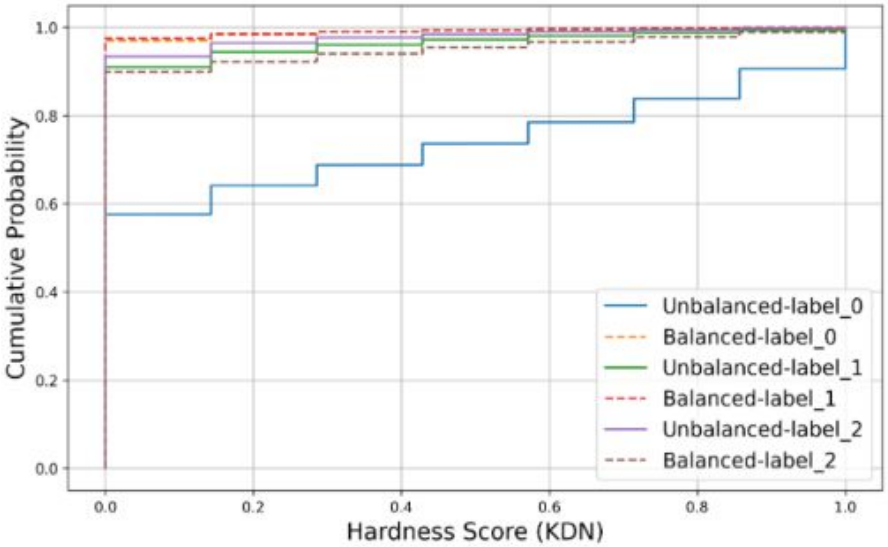


- Implementamos uma abordagem de balanceamento individual para cada Bootstrap do Bagging
- Bagging Decision Tree
- Objetivo: Aumentar a variabilidade dos dados, consequentemente, dos classificadores, tornando o *ensemble* mais robusto.

Instance Hardness



Drebin



CIRA-CIC-DoHBrw-2020

Resultados e Conclusão

Zebin

Metric	CIRA-CIC-DoHBrw-2020	Drebin
Accuracy	99.49	98.22
Precision	99.56	98.41
Recall	99.49	98.22
F1-Score	99.51	98.29
G-Mean	99.59	93.18
MCC	98.61	80.44

HBBB

Metric	CIRA-CIC-DoHBrw-2020	Drebin
Accuracy	99.67	98.39
Precision	99.70	78.61
Recall	99.67	86.23
F1-Score	99.68	82.23
G-Mean	99.71	92.59
MCC	99.11	92.37

Centro de Informática

UFPE