Detecção de URLs Maliciosas

Julio Cesar da Silva Rodrigues

Universidade Federal de São João del-rei
Curso de Ciência da Computação

Trabalho Prático 1 - Mineração de Dados

18 de Abril de 2023

Conteúdo

Conteúdo

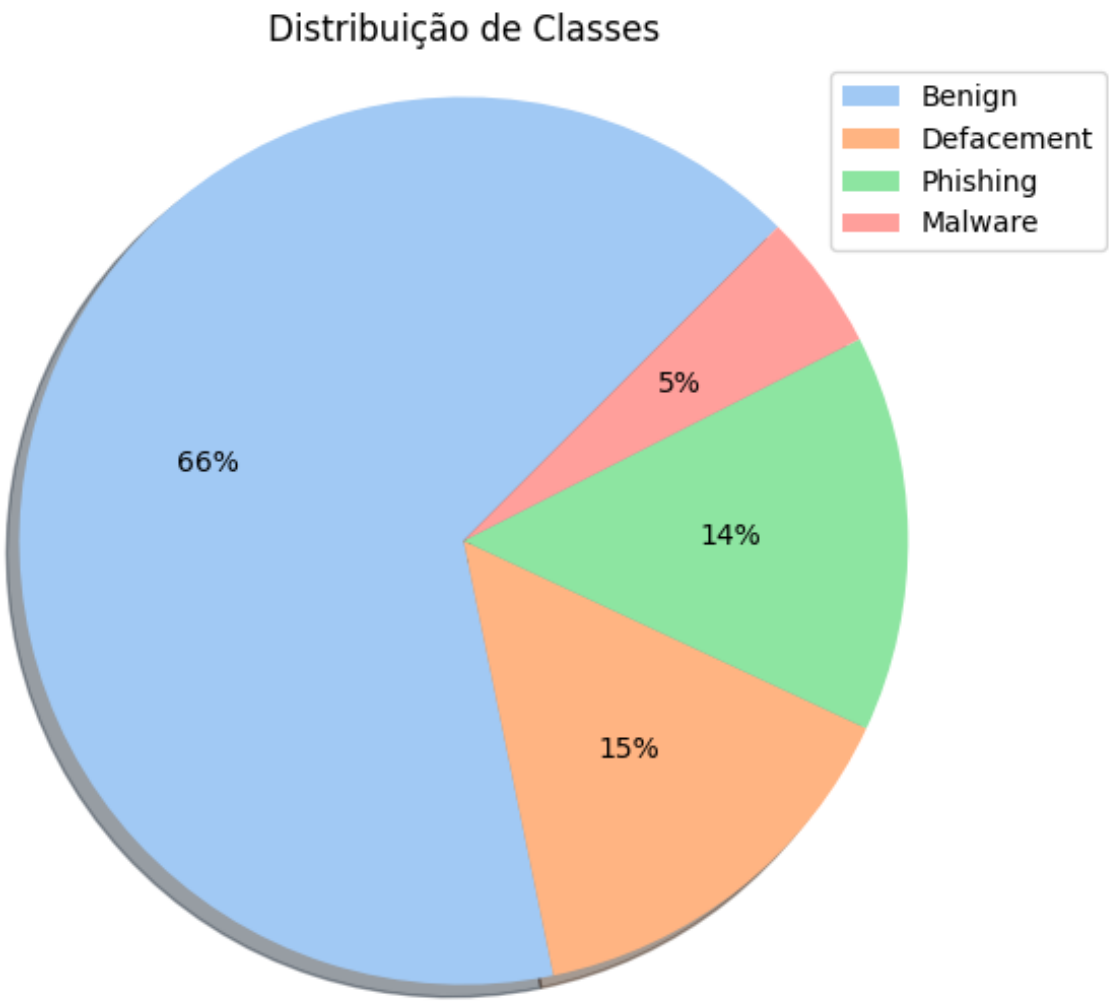
Tecnologias Utilizadas

- Python 3;
- Machine Learning e Manipulação de Dados:
  1. scikit-learn;
  2. xgboost;
  3. pandas.
- Visualização dos Dados:
  1. Matplotlib;
  2. seaborn.

Base de Dados

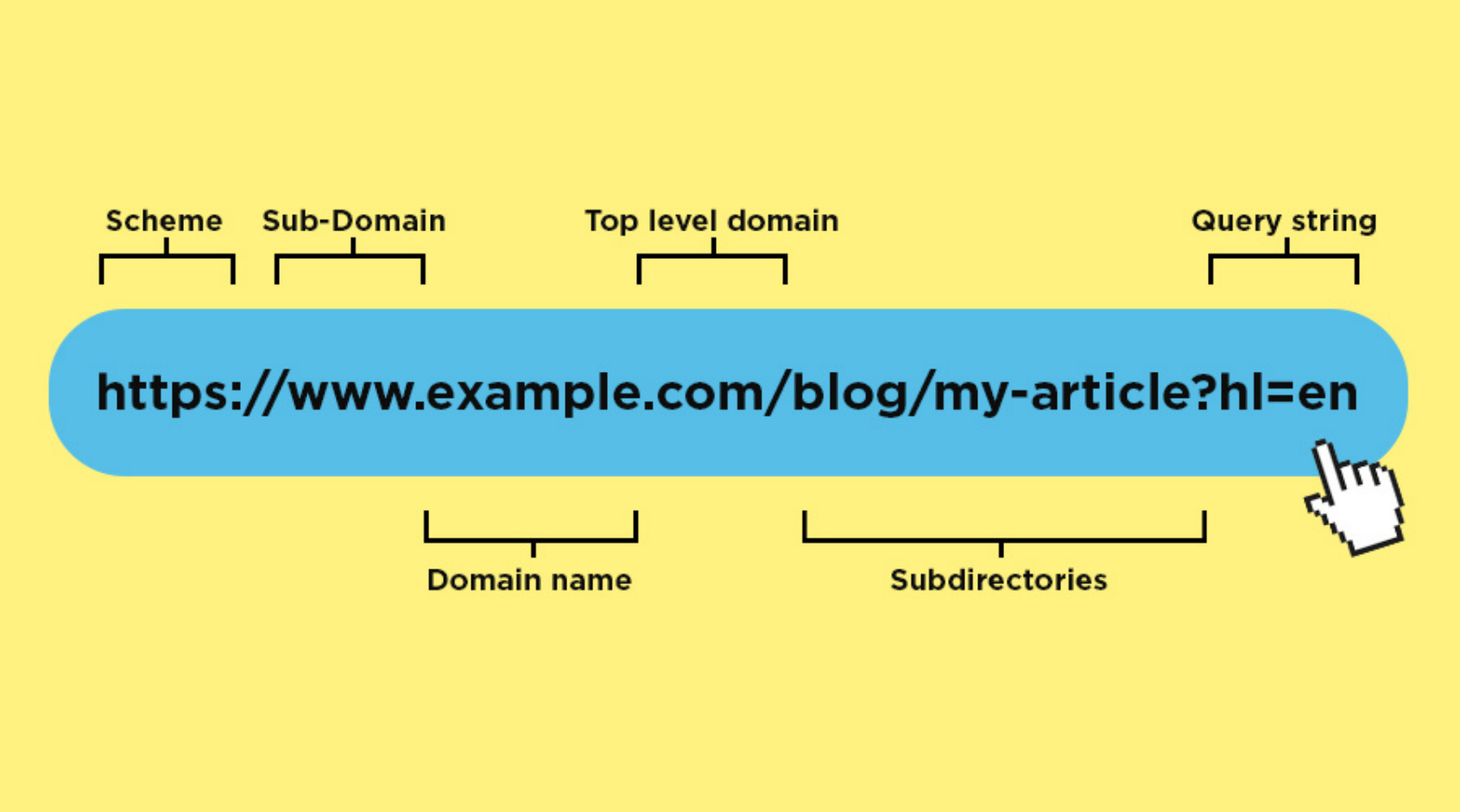
- URLs Maliciosas:
  1. Um atributo;
  2. Uma classe com quatro valores distintos;
  3. Mais de 650 mil instâncias;
  4. Disponível em: `https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset`.
- Objetivos Principais:
  1. Criação de Atributos;
  2. Observar como cada novo atributo criado impacta na classificação.

Base de Dados



Conteúdo

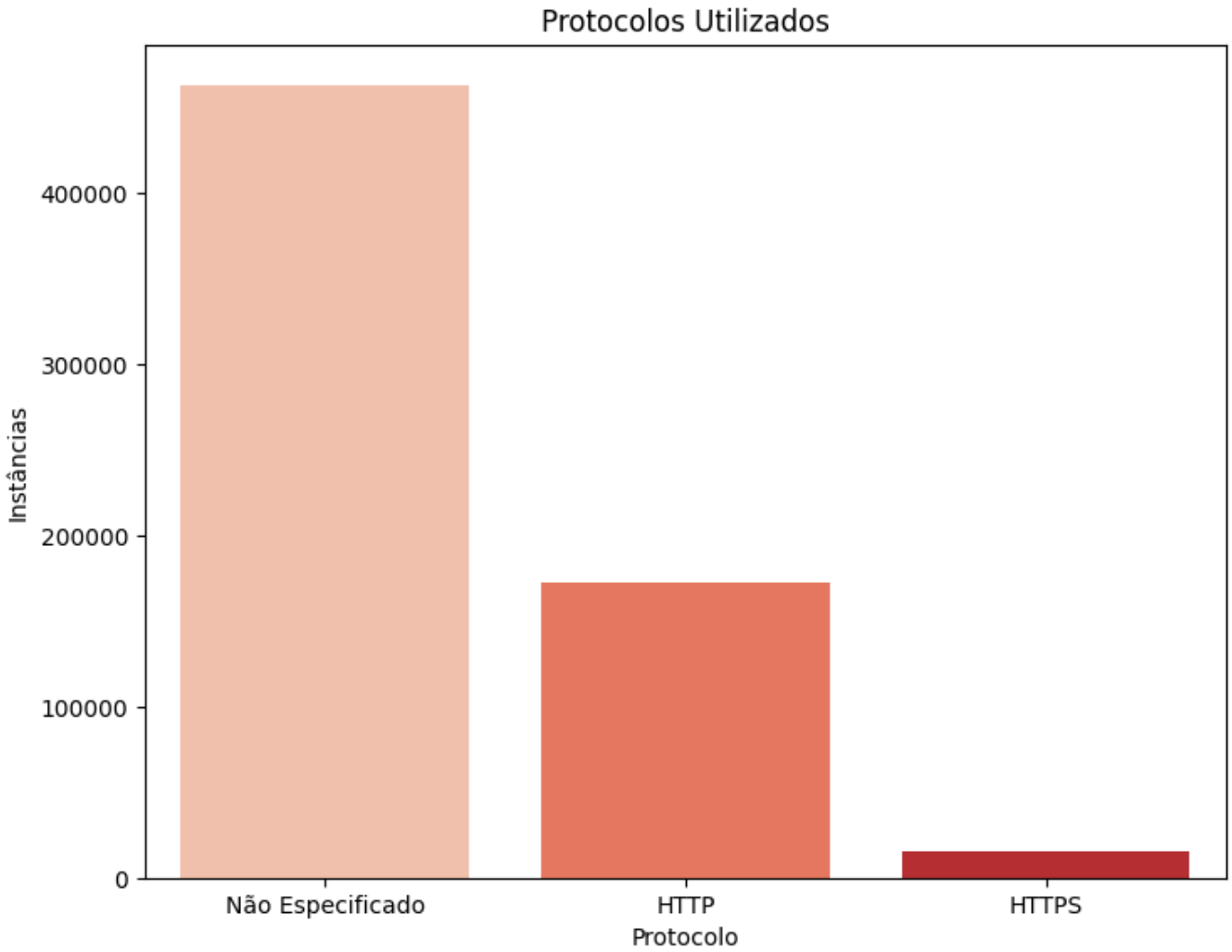
Análise Léxica



Protocolo de Comunicação

- Grande potencial para influenciar na classificação;
- Protocolos das URLs presentes na base são distribuídos em:
  1. HTTPS;
  2. HTTP;
  3. Não Especificado.
- Somente 2,4% das URLs utilizam HTTPS (explícito);
- 85,7% das URLs HTTPS são maliciosas (*phishing* e *malware*);
- Todas as URLs de *defacement* utilizam HTTP.

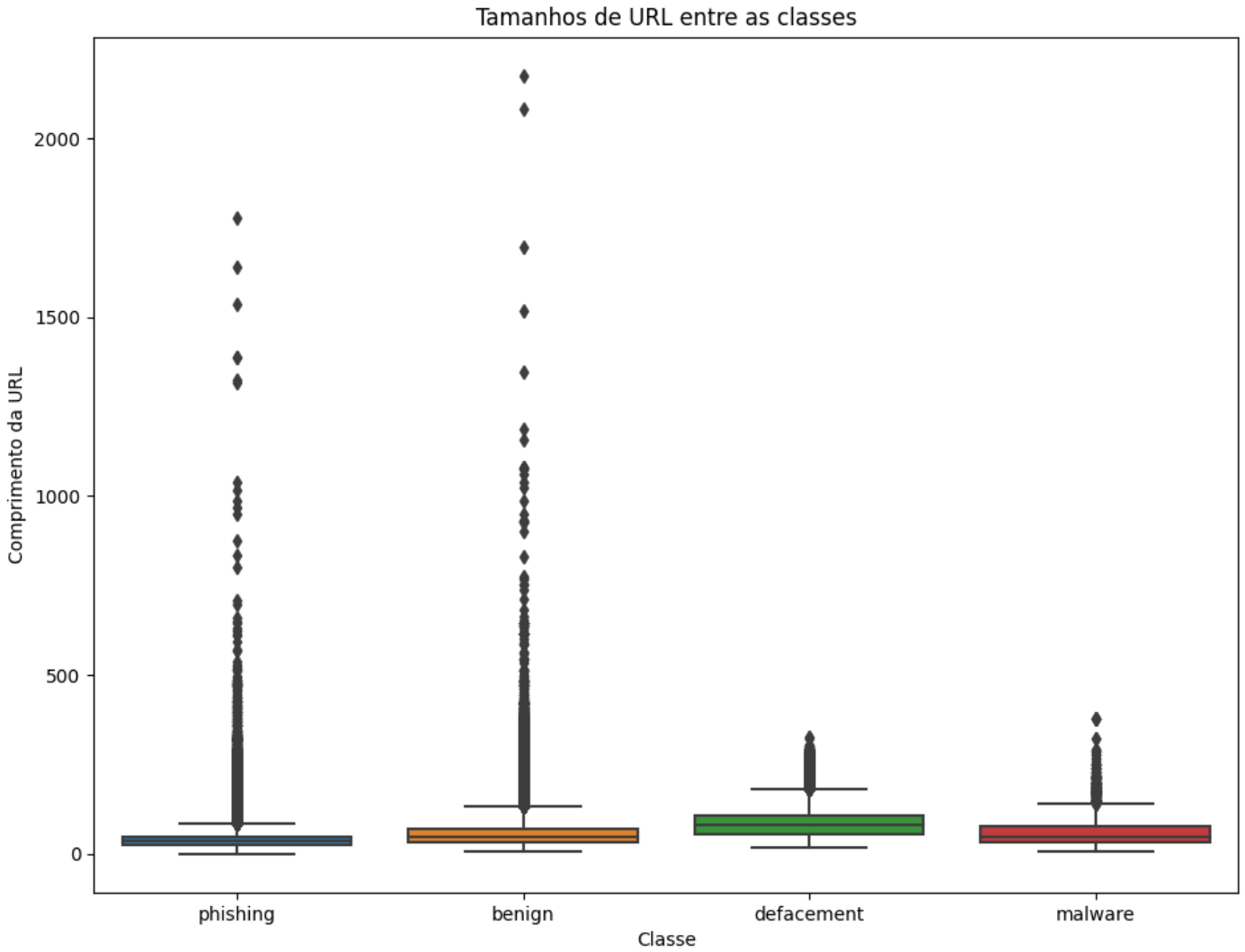
Protocolo de Comunicação



Comprimento de URLs

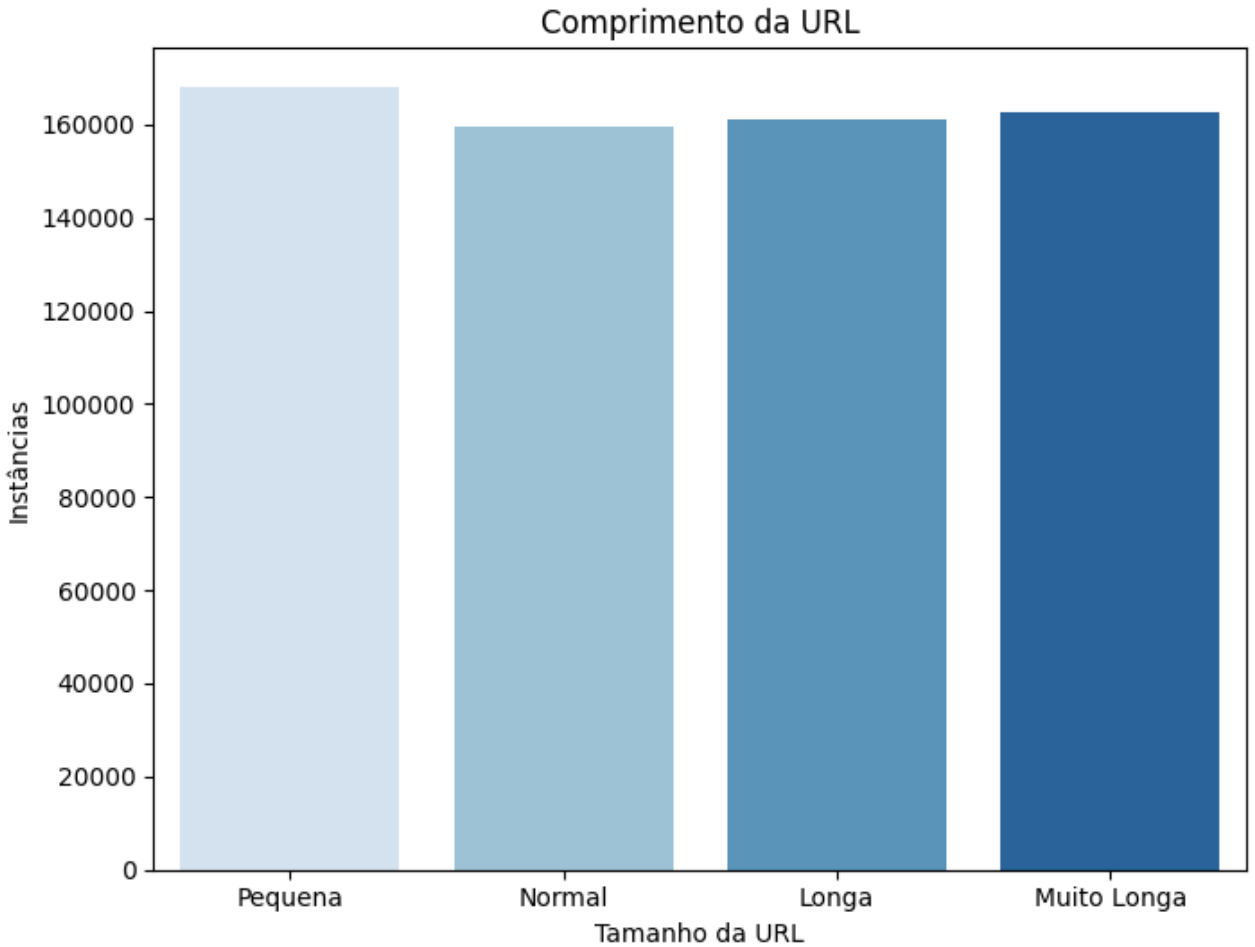
- Apresentaram ligeiras disparidades entre as classes;
- URLs de *defacement* são, em média, 50% maiores que URLs seguras;
- URLs de *phishing* são, em média, 25% menores que URLs seguras;
- URLs HTTPS e HTTP são, em média, de 63% a 72% maiores que as de protocolo não explícito.

Comprimento das URLs



Comprimento das URLs

- Equal-Frequency Binning



Tamanho do Primeiro Diretório e Quantidade de Dígitos

- Tamanho do primeiro diretório de URLs de *malware* é, em média, o dobro de URLs seguras;
- Tamanho do primeiro diretório de URLs de *phishing* é, em média, 25% menores que de URLs seguras;
- URLs de *malware* possuem, em média, mais que o dobro de dígitos de URLs seguras;
- URLs de *phishing* possuem, em média, 35% menos dígitos que URLs seguras.

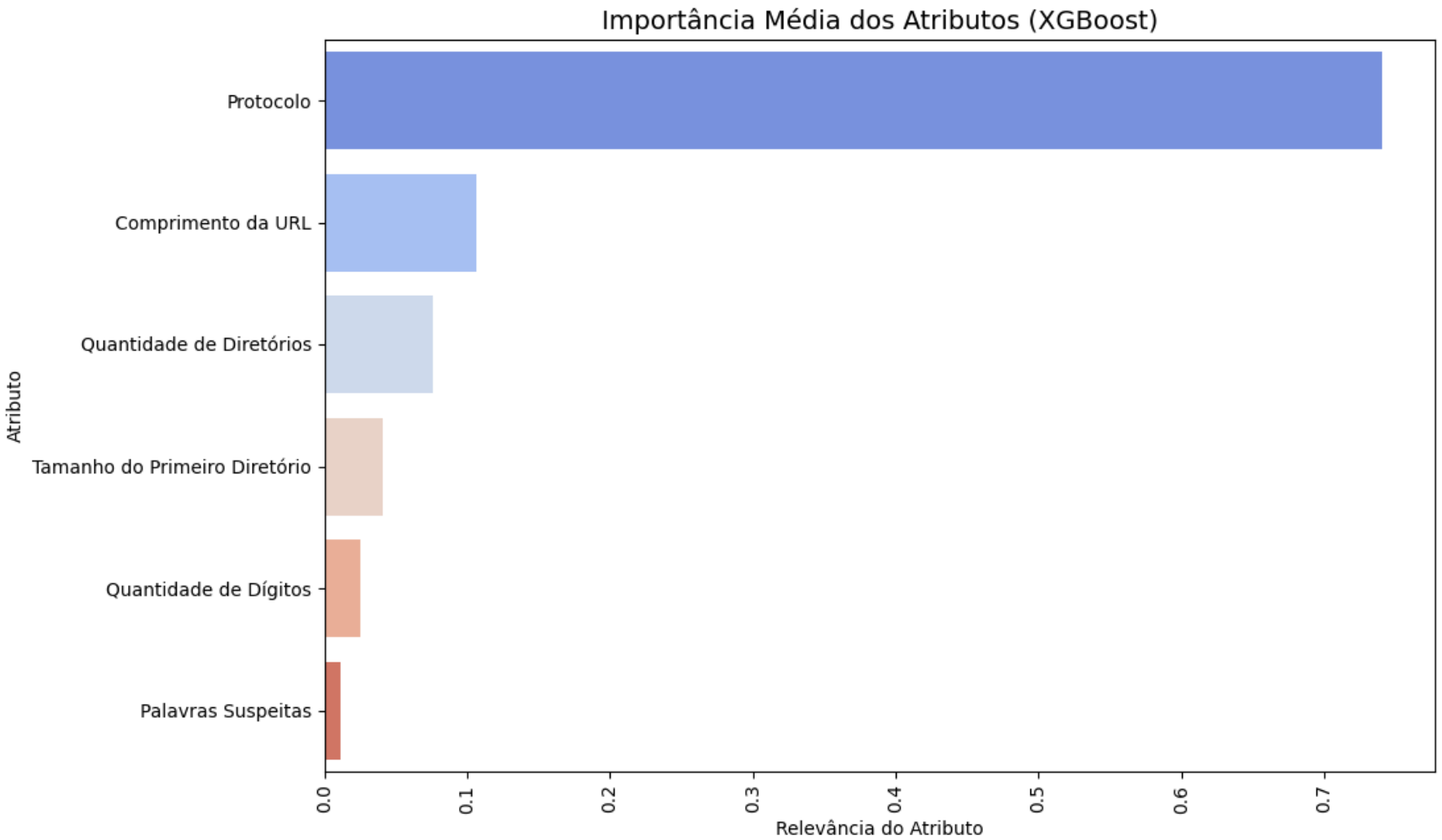
Palavras Suspeitas

- Foco em destacar URLs de *phishing*;
- Impacto muito abaixo do esperado.

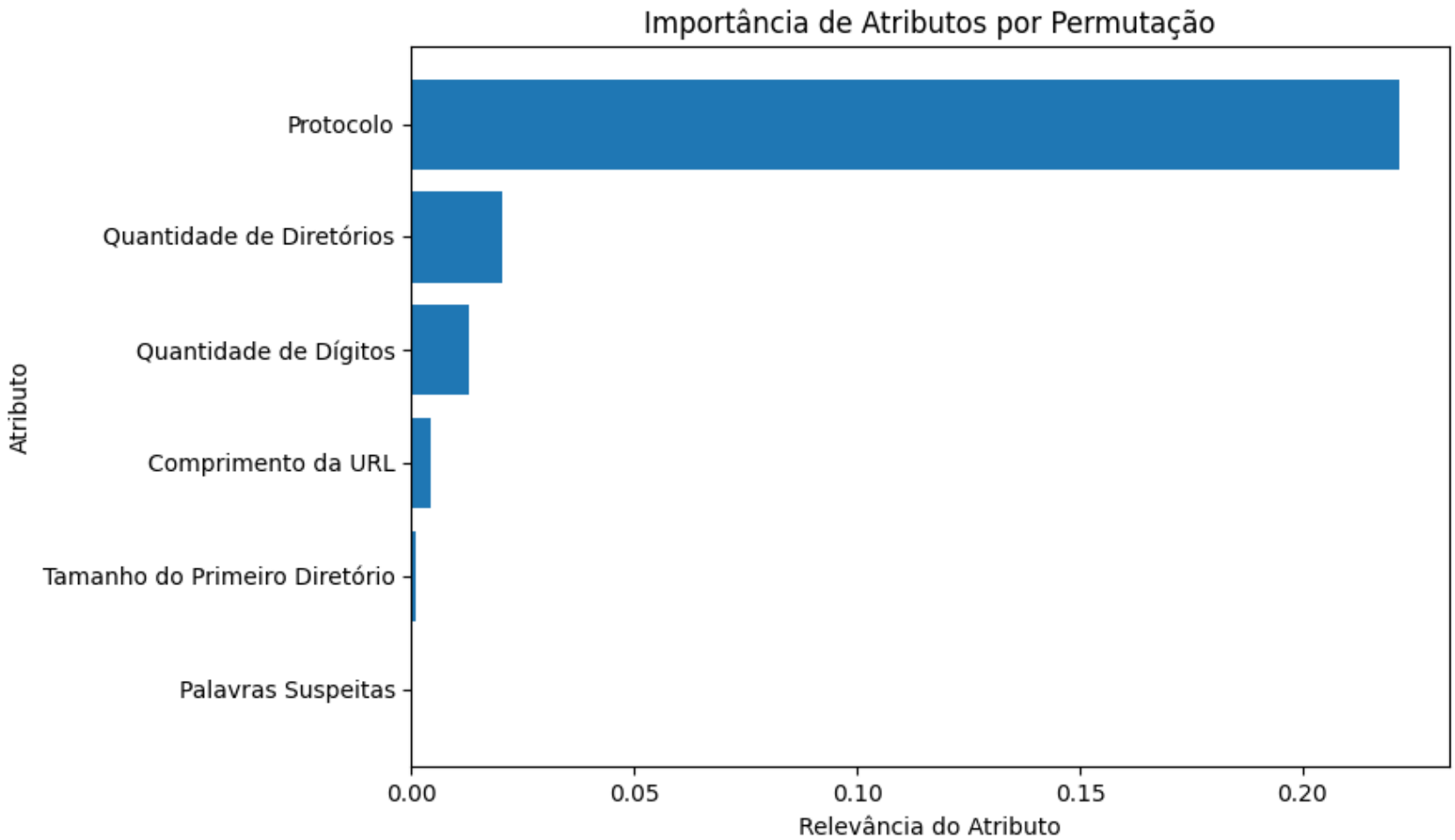
```
def odd_words(url):

    # Search for suspicious words related to phishing in each url
    pattern = re.search('free|account|signin|bonus\
                        |lucky|extra|payment|details', url)
    if pattern:
        return 1
    else:
        return 0
```

Conteúdo

Ganho de Informação



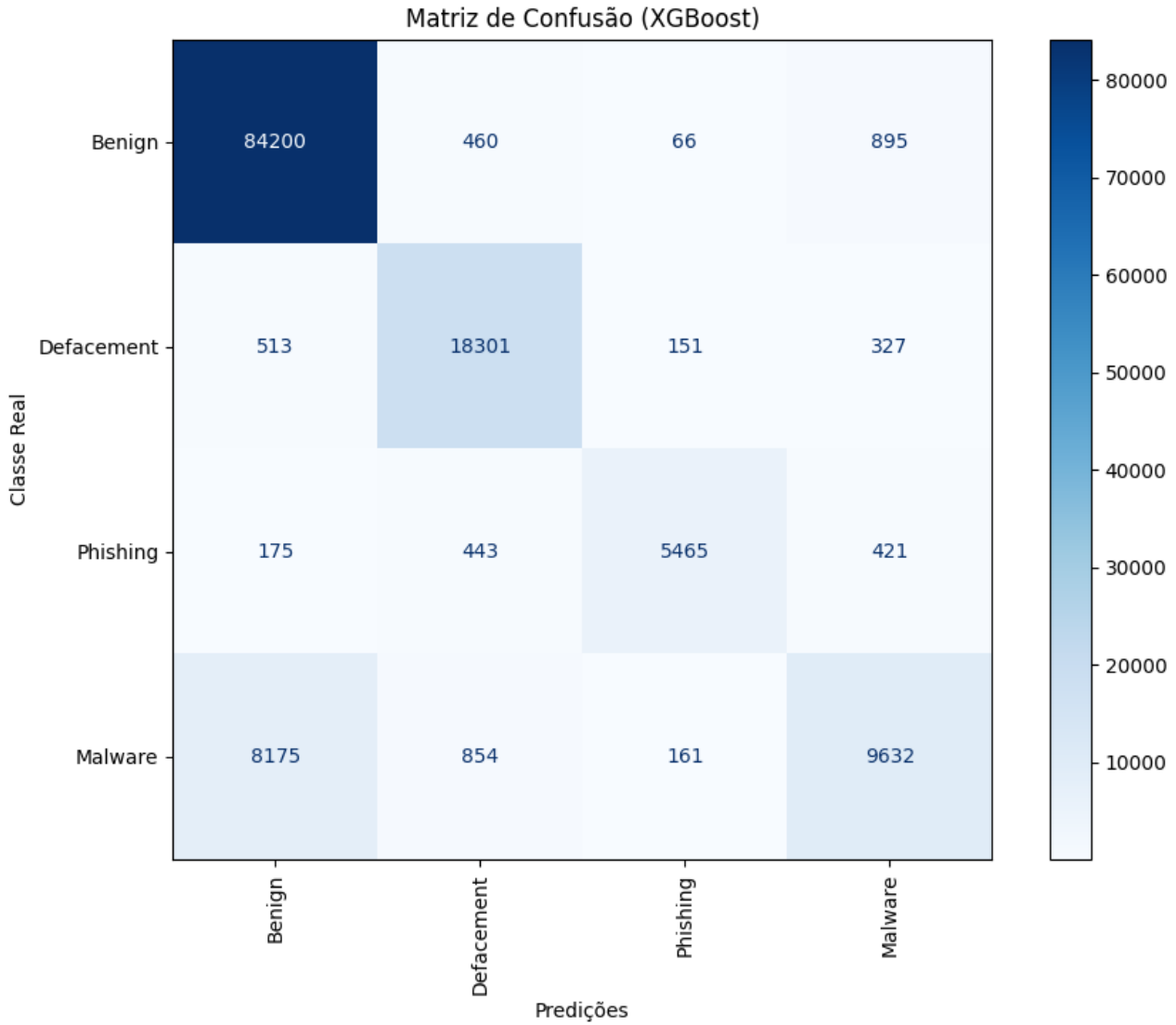
Ganho de Informação



Testes Incrementais

XGBoost				
Class	Precision	Recall	F1-Score	Support
Benign	0.90	0.98	0.94	85621
Defacement	0.91	0.95	0.92	19292
Phishing	0.94	0.84	0.89	6504
Malware	0.85	0.51	0.64	18822
Accuracy			0.90	130239
Macro Avg	0.90	0.82	0.85	130239
Weighted Avg	0.90	0.90	0.89	130239

Matriz de Confusão



Conteúdo

Modelos de Machine Learning

- Regressão Logística;
- XGBoost;
- Validação Cruzada (*k-fold*):
  1. 10 partes;
  2. Amostragem estratificada;
  3. Métrica *Macro F1*.

Comparativo de Modelos

Macro F1		
Modelo	Média	Desvio Padrão
Regressão Logística	0.4343240025832924	0.0014418608790156475
XGBoost	0.8218246353658317	0.0019078924494051 27

- Calculada para cada uma das classes;
- Não leva em conta possível desbalanceamento presente na base de dados;
- XGBoost exige mais processamento, mas entregou resultados melhores;
- Ainda há grande margem para melhorias.

Conteúdo

Próximos Passos

- Construção de novos atributos;
- Expandir análises:
  1. Atributos relacionados à conteúdo;
  2. Atributos relacionados à rede.

Próximos Passos

- Construção de novos atributos;

- Expandir análises:
  1. Atributos relacionados à conteúdo;
  2. Atributos relacionados à rede.

Próximos Passos

- União de classes pode favorecer os modelos;

Extração de Classes

{Defacement, Phishing, Malware} $\rightarrow$ {Malicious}

- Balanceamento da base de dados:
  1. Oversampling;
  2. Instance Selection.

- Ajuste fino dos hiperparâmetros dos modelos.

Próximos Passos

- União de classes pode favorecer os modelos;

Extração de Classes

{Defacement, Phishing, Malware} $\rightarrow$ {Malicious}

- Balanceamento da base de dados:
  1. Oversampling;
  2. Instance Selection.

- Ajuste fino dos hiperparâmetros dos modelos.

Próximos Passos

- União de classes pode favorecer os modelos;

Extração de Classes

$\{\text{Defacement, Phishing, Malware}\} \rightarrow \{\text{Malicious}\}$

- Balanceamento da base de dados:
  1. Oversampling;
  2. Instance Selection.

- Ajuste fino dos hiperparâmetros dos modelos.