Detecção de URLs Maliciosas

Mineração de Dados Aplicada

Julio Cesar da Silva Rodrigues[1]

[1]Universidade Federal de São João del-Rei
Curso de Ciência da Computação
julio.csr.271@aluno.ufsj.edu.br

25 de Maio de 2023

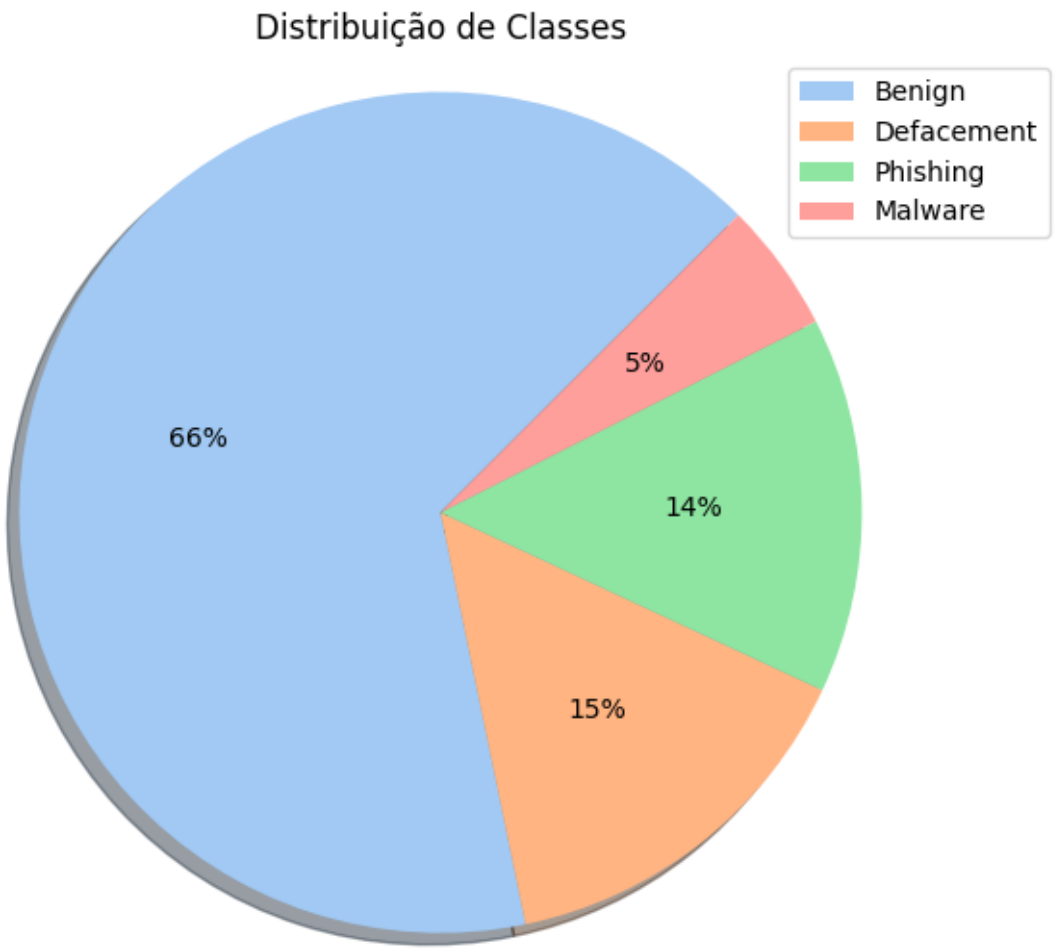
Conteúdo

Conteúdo

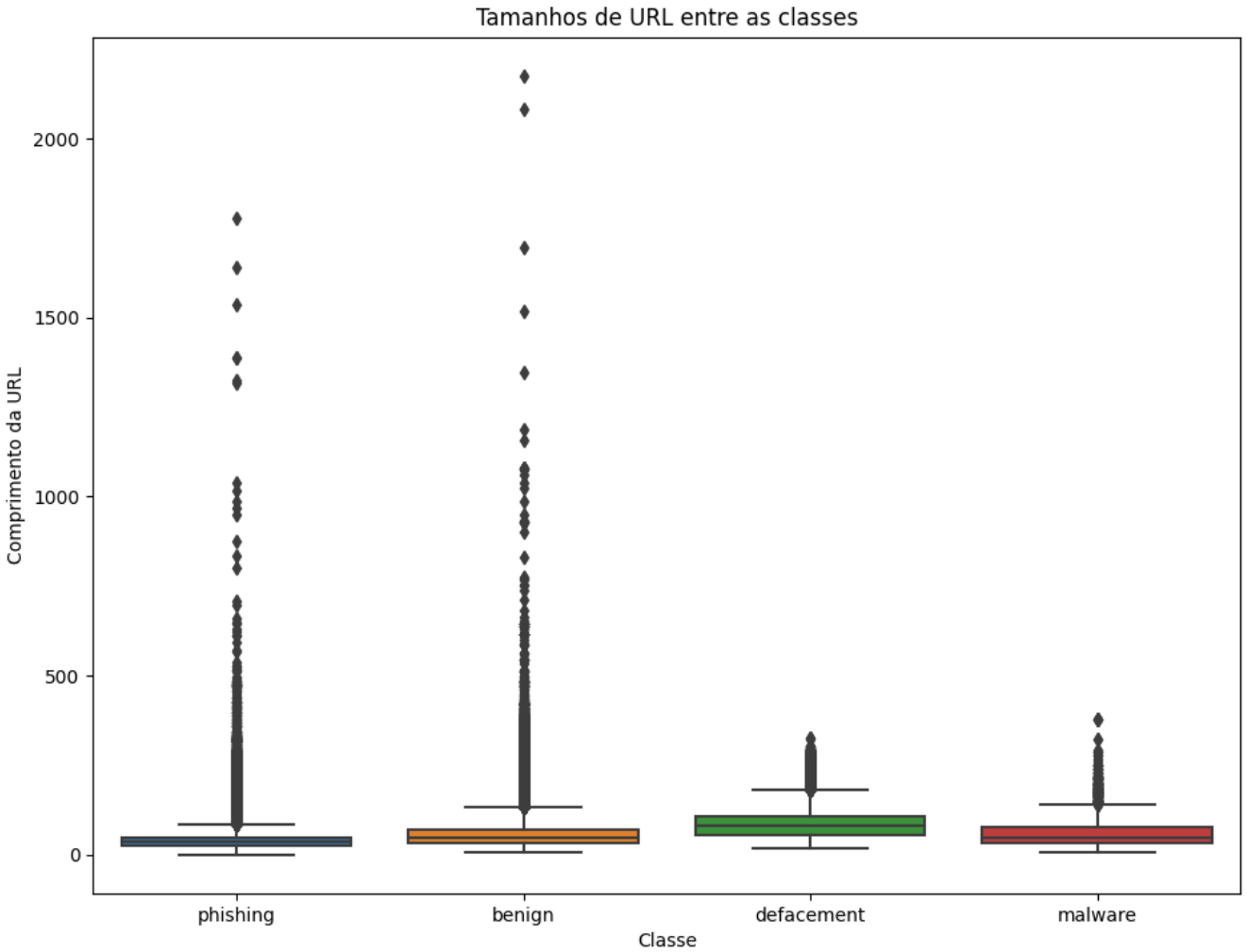
Recapitulando

- 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. Análise exploratória da base;
  2. Observar o quanto cada atributo criado define a natureza de uma URL.

Recapitulando

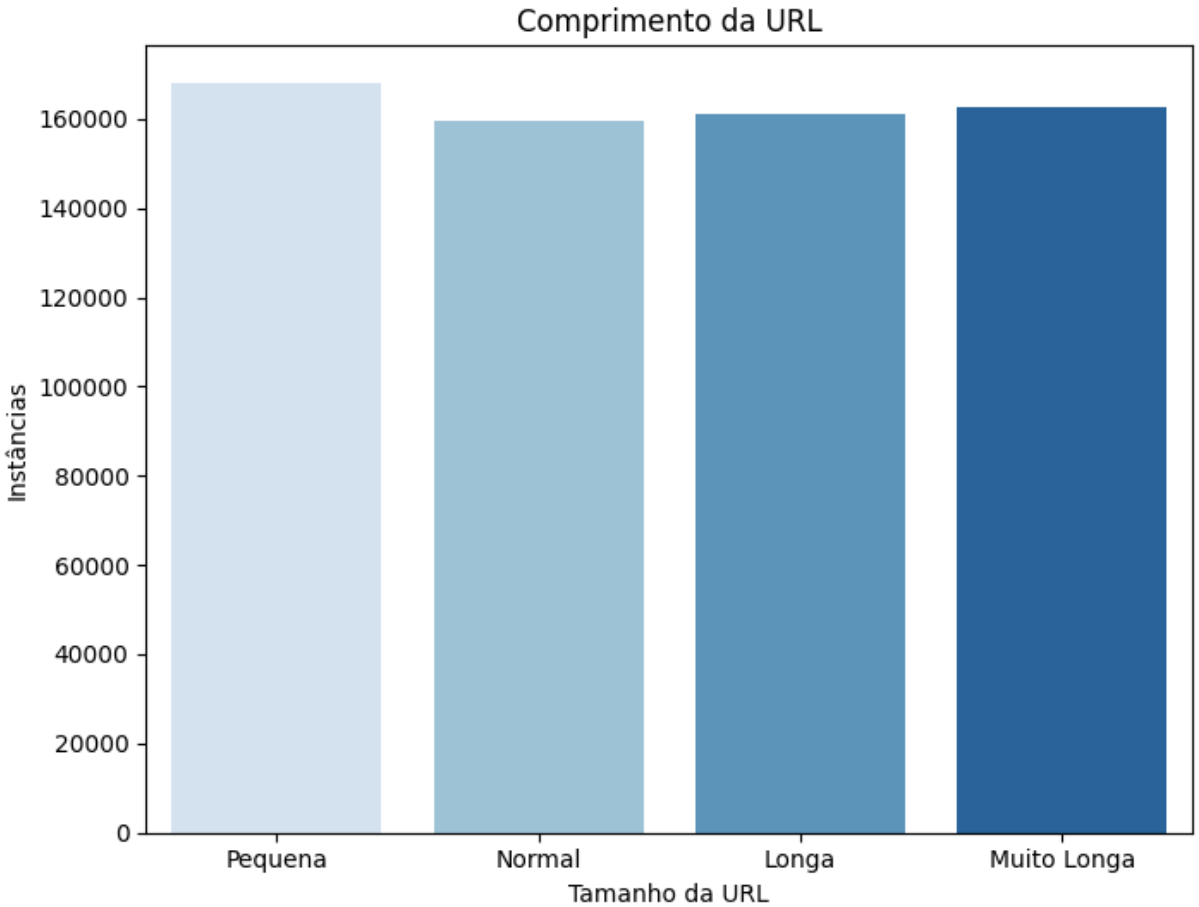


Recapitulando



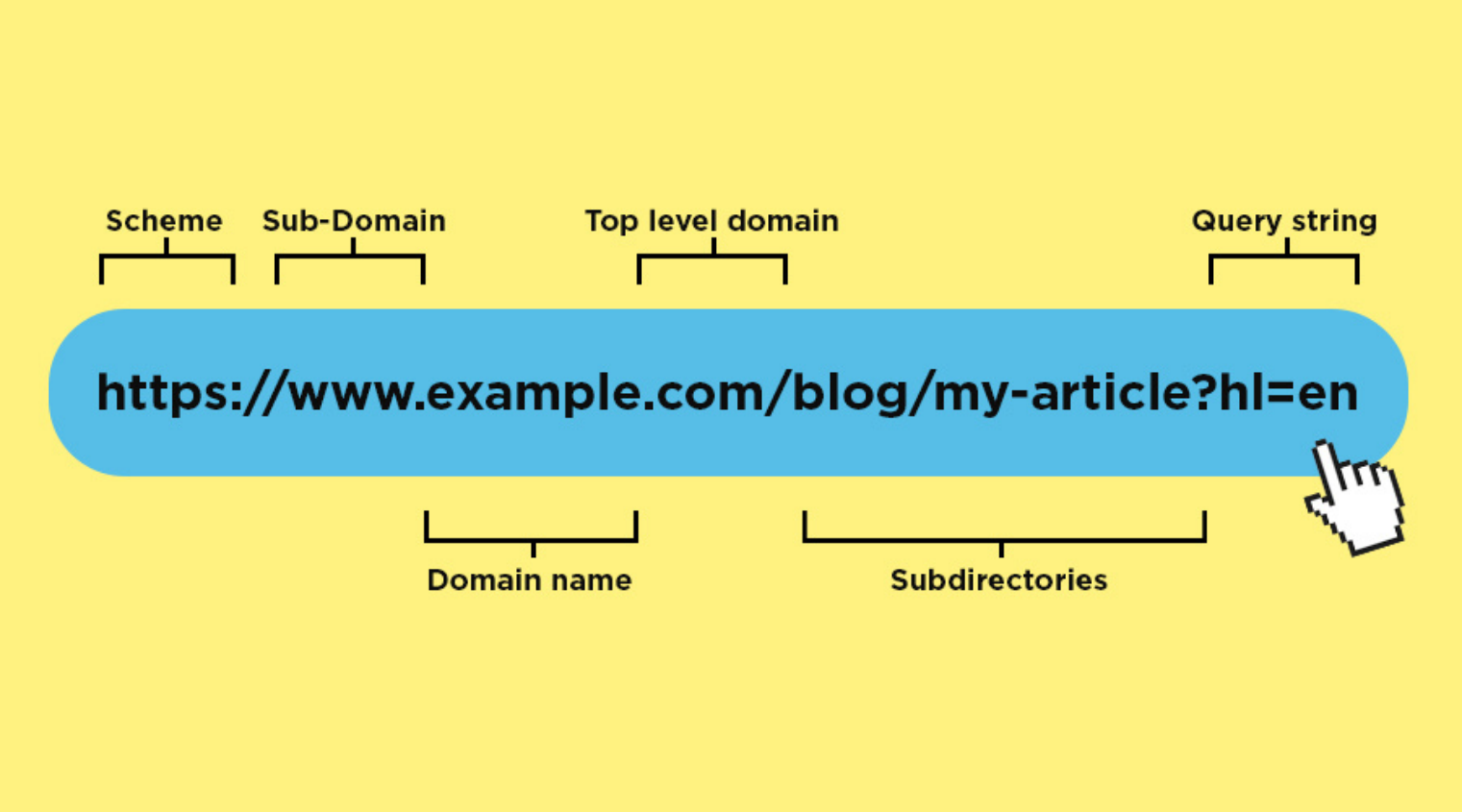
Recapitulando

- Equal-Frequency Binning



Conteúdo

Estrutura de uma URL

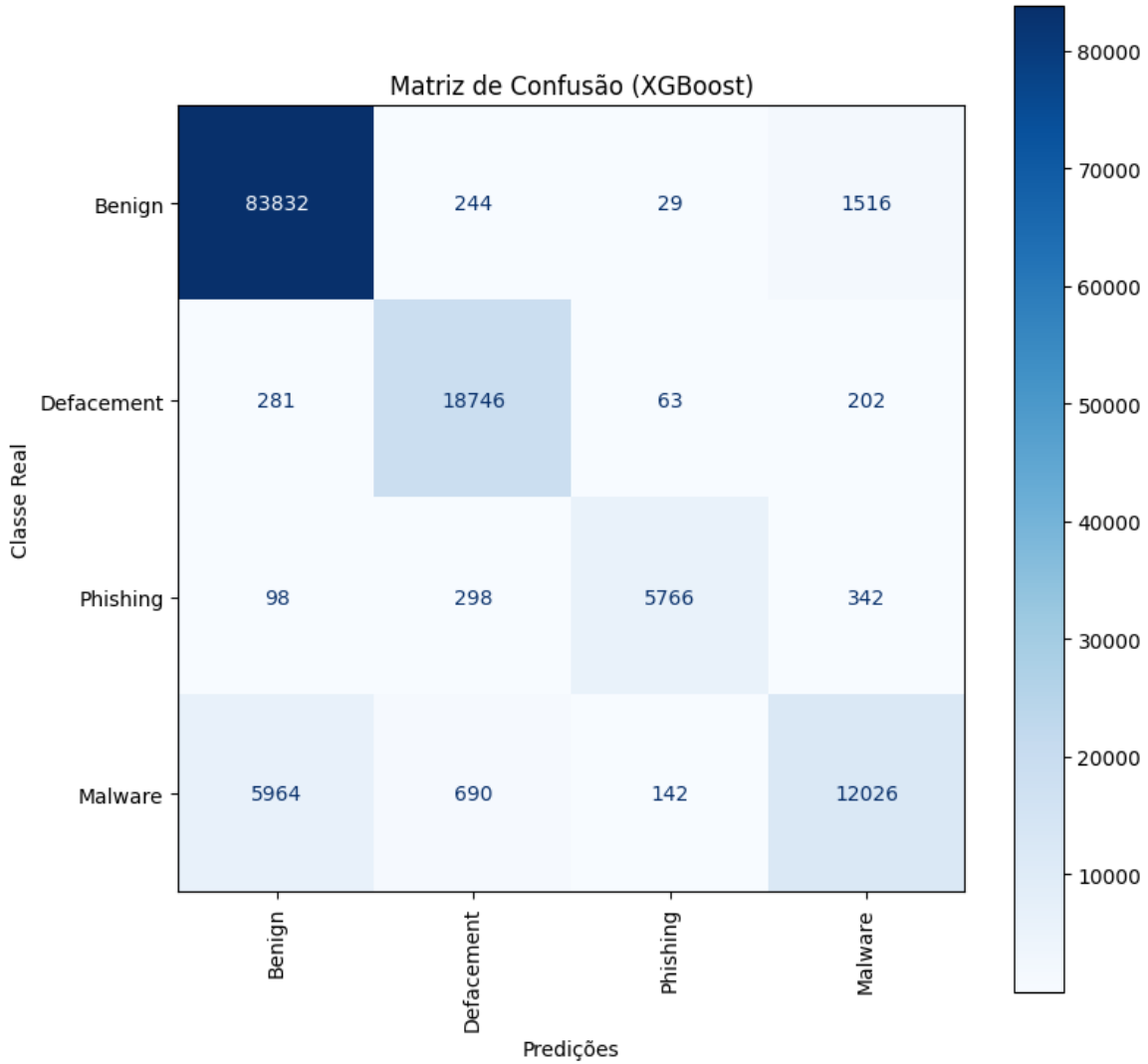


Tipos de Caracteres

- Disparidade na quantidade de caracteres não-alfanuméricos;
- Em relação às URLs seguras:
  1. URLs de *defacement* possuem, em média, 75% mais caracteres do tipo;
  2. URLs de *malware* possuem, em média, o dobro de caracteres do tipo;
  3. URLs de *phishing* possuem, em média, 25% menos caracteres do tipo.

Conteúdo

Matriz de Confusão



F1 Score

TP 1		
Algoritmo	Média	Desvio Padrão
Regressão Logística	0,4343240025832924	0,0014418608790156475
XGBoost	0,8218246353658317	0,0019078924494051 27

TP 2 - Parcial I		
Algoritmo	Média	Desvio Padrão
Regressão Logística	0,5780295368328738	0,0021293237223429956
XGBoost	0,856891093617907 9	0,0017065729226258411

Conteúdo

Finalização da Parcial I

- Balanceamento da base:
  1. PhishTank;
  2. Kaggle;
  3. *Oversampling?*.
- Novos atributos:
  1. Conteúdo das páginas;
  2. Dados de rede.

Parcial II

- Seleção de 2 a 4 algoritmos mais recentes;
- Formulação do modelo com a base polida;
- Comparativo com trabalhos relacionados.