Identificando URLs maliciosas:

Uma abordagem puramente léxica

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

1. Introdução
2. Trabalhos Relacionados
3. Metodologia
4. Resultados
5. Conclusão

1. Introdução
2. Trabalhos Relacionados
3. Metodologia
4. Resultados
5. Conclusão

Contextualização

- Via rápida e direta para aplicar crimes cibernéticos;
- Potencial de infecção exponencial;
- Brasil no Top 15 com maior número de vítimas, segundo relatório[1] da Abranet;
- Evolução nas técnicas de camuflagem e detecção.

[1] Disponível em: https://www.abranet.org.br/Noticias/Relatorio-aponta-que-cada-URL-maliciosa-no-Brasil-afeta-18-usuarios-2585.html?UserActiveTemplate=site

Objetivos

- Principais características que definem a natureza de uma URL;
- Capacidade de predição analisando somente a estrutura léxica;
- Modelo competitivo com classificação multiclasse.

1 Introdução

2 Trabalhos Relacionados

3 Metodologia

4 Resultados

5 Conclusão

Classificação Multiclasse

- Detecção de URLs baseada somente em atributos léxicos [Saleem Raja et al., 2021];
- Base de dados da UNB[2];
- Extração de 27 atributos;
- Cinco algoritmos de *machine learning* selecionados;
- 99% de acurácia com *random forest*.

[2] University of New Brunswick

Classificação Binária

- Detecção de URLs de *phishing* direcionadas a brasileiros [Ayres et al., 2019];
- Aprendizado baseado em atributos léxicos e relacionados à rede;
- Bases de dados nacionais:
  1. CaUMa[3]: URLs maliciosas;
  2. UFBA[4]: URLs seguras.

[3] Serviço associado ao Catálogo de Fraudes da RNP

[4] Universidade Federal da Bahia

Classificação Binária

- Extração de 117 características;
- Quatro algoritmos de *machine learning* selecionados;
- Empate técnico com *F1 Score* média de 95,85%:
  1. KNN;
  2. SVM;
  3. J48.

1 Introdução

2 Trabalhos Relacionados

3 Metodologia

4 Resultados

5 Conclusão

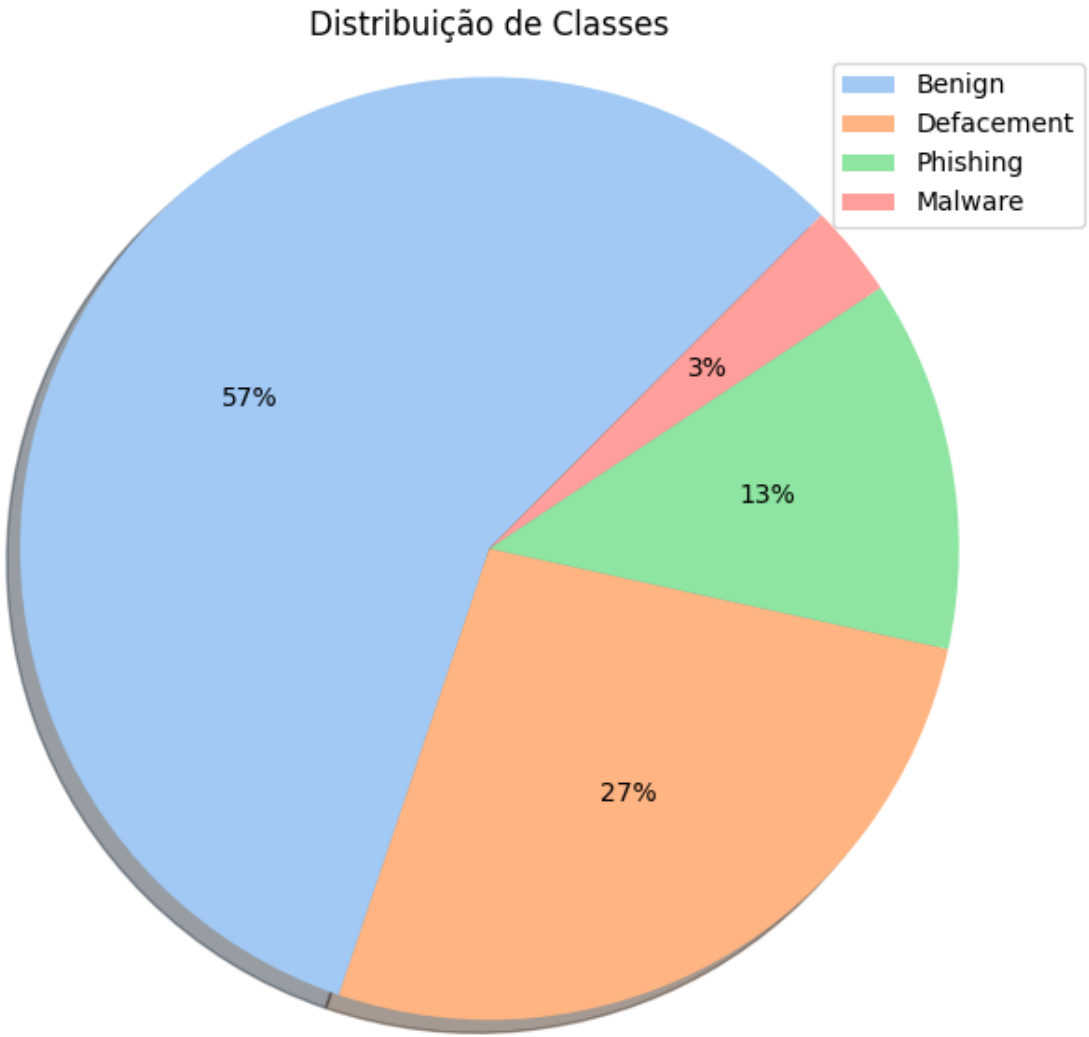
Base de Dados

- Kaggle[5]:
  1. Mais de 600 mil instâncias;
  2. 4 tipos de URL;
- PhishTank[6]:
  1. Mais de 100 mil instâncias;
  2. Somente URLs de *phishing*;

[5] Disponível em: `https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset`

[6] Disponível em: `https://phishtank.org/phish_archive.php`

Base de Dados



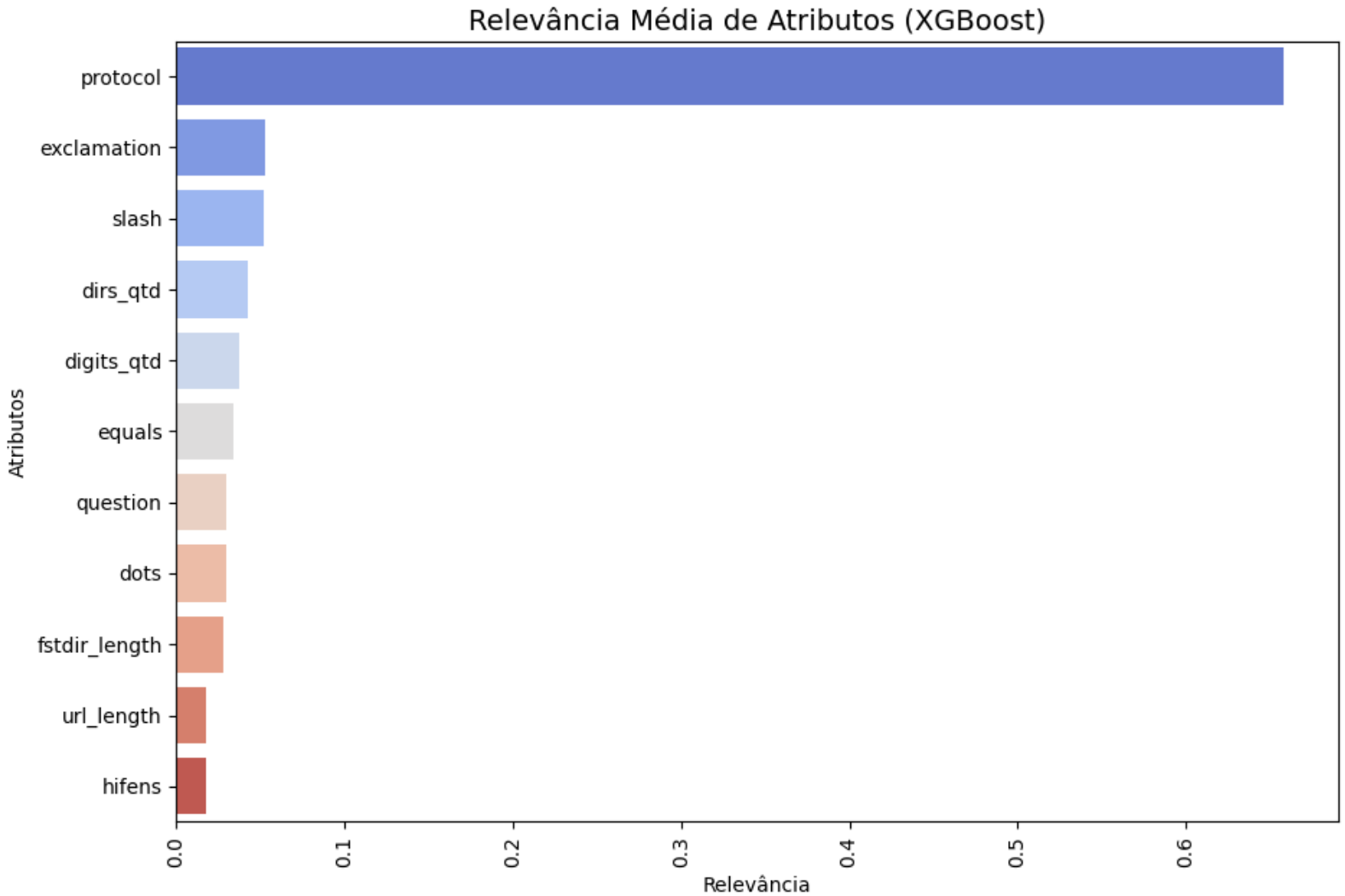
Balanceamento

- *Undersampling* aleatório para instâncias *benign*;
- Redução de 428081 para 200 mil instâncias;
- *Scraper* para coletar mais de 100 mil URLs da PhishTank;
- *SMOTE* [Bowyer et al., 2011] para suprir o déficit de instâncias de *defacement* e *malware*.

Construção de atributos

- 21 atributos criados:
  1. 7 com base em análise estatística da base de dados;
  2. 14 criados utilizando trabalhos relacionados como base;
  3. Exemplos: {Protocolo de comunicação, Comprimento da URL, Quantidade de *tokens*, ...}
- Aplicação de *feature selection* com **RFECV**;

Construção de atributos



Base de Dados Final

- 800 mil instâncias;
- Perfeitamente balanceada;
- 11 atributos e 1 classe (4 valores distintos);
- Composta por URLs de bases de dados do *Kaggle* e *PhishTank*.

Modelos de Aprendizado Supervisionado

- Algoritmos selecionados:
  1. XGBoost;
  2. KNN;
  3. Regressão Logística.
- Validação cruzada com amostragem estratificada (*10-fold*);
- Métrica *Macro F1* e desvio padrão;
- Ajuste fino - *Tripartite* (*5-fold*);
- Teste t de dupla cauda.

1 Introdução

2 Trabalhos Relacionados

3 Metodologia

4 Resultados

5 Conclusão

Comparativo de Algoritmos

Tabela 1: *Macro F1 Scores* alcançadas pelos algoritmos

Resultados		
Algoritmo	Média	Desvio Padrão
Regressão Logística	0,7339	0,0039
XGBoost	0,9476	0,0006
KNN	0,9229	0,0008

- XGBoost e KNN com desempenho similar inicialmente;
- Desvio padrão constantemente baixo é um bom indicativo.

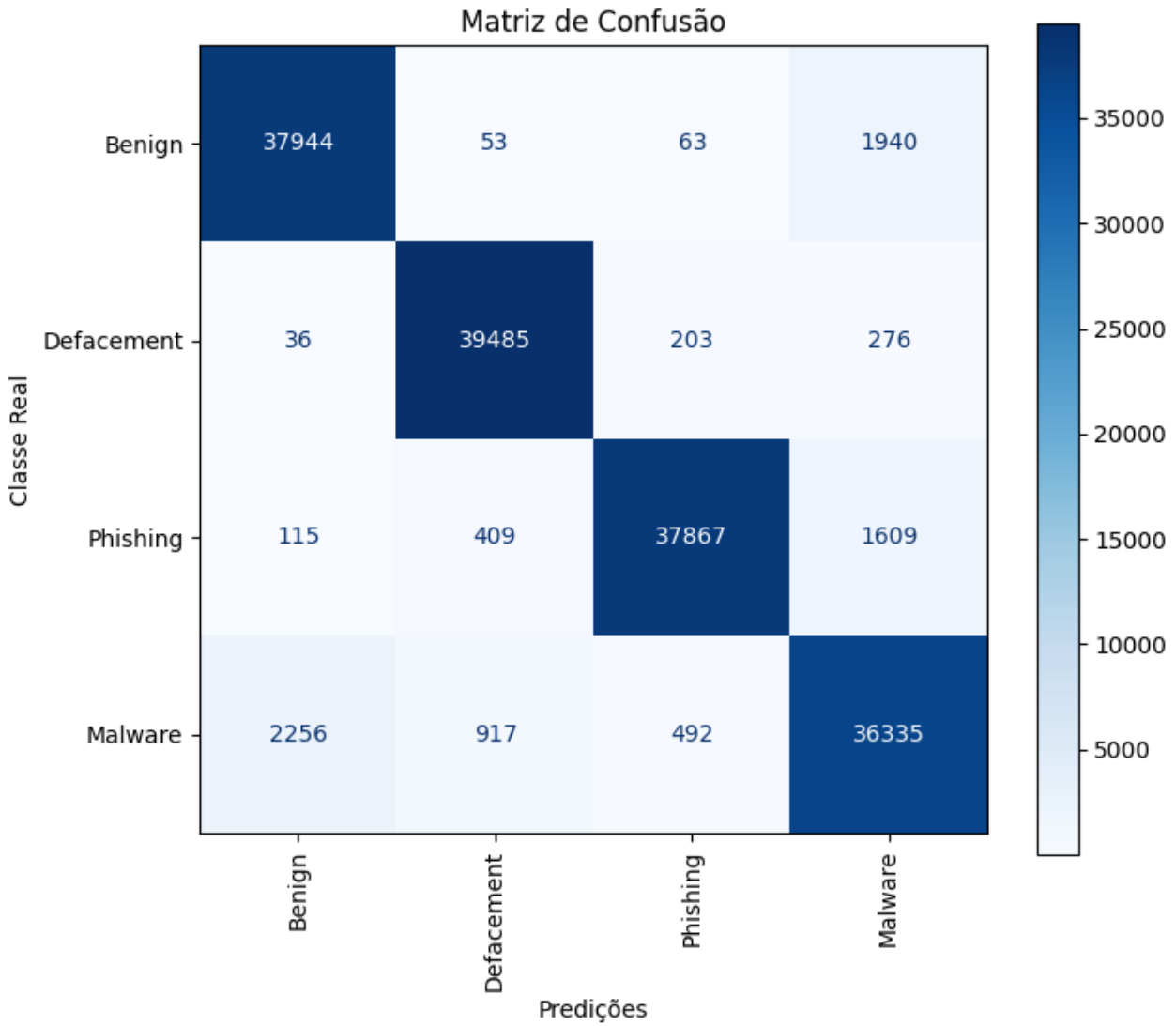
Teste Estatístico

- Nível de significância $\rightarrow \alpha = 0,05$;
- XGBoost x KNN:
  1. Hipótese nula rejeitada;
  2. Modelos estatisticamente distintos.
- XGBoost x Regressão Logística:
  1. Hipótese nula rejeitada;
  2. Modelos estatisticamente distintos.

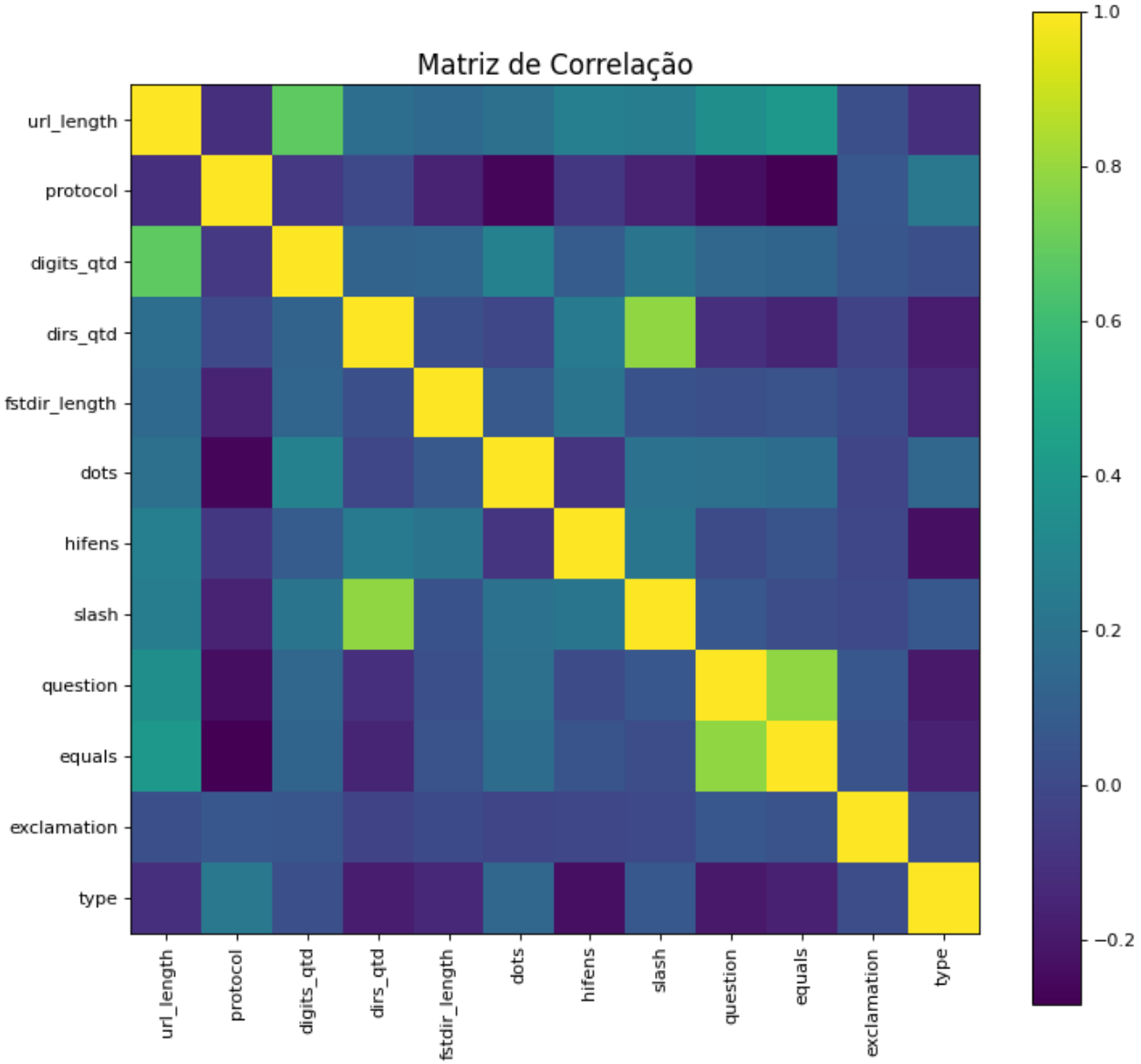
Teste Estatístico

- KNN x Regressão Logística:
  1. Hipótese nula rejeitada;
  2. Modelos estatisticamente distintos.
- XGBoost na dianteira como o melhor desempenho;
- Alta disparidade em custo computacional.

Limitações


Matriz de Confusão
Benign
Defacement
Phishing
Malware
37944
53
63
1940
36
39485
203
276
115
409
37867
1609
2256
917
492
36335
Classe Real
Predições
35000
30000
25000
20000
15000
10000
5000

Limitações


Matriz de Correlação
url_length
protocol
digits_qtd
dirs_qtd
fstdir_length
dots
hifens
slash
question
equals
exclamation
type
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

1 Introdução

2 Trabalhos Relacionados

3 Metodologia

4 Resultados

5 Conclusão

Considerações Finais

- Distinção entre URLs maliciosas;
- Potencial para competir com modelos que utilizam outros tipos de características;
- Generalização e nível de confiabilidade na classificação.

Passos Futuros

- Exploração mais profunda na construção de atributos;
- Investigação minunciosa sobre o que separa *malware* de outros tipos de URL;
- Possíveis reduções no volume da base de dados (*instance selection?*).

Referências

[Ayres et al., 2019] Ayres, L., Brito, I. V. S., and e Souza, R. G. (2019).
Utilizando aprendizado de máquina para detecção automática de urls maliciosas brasileiras.
In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 972–985, Porto Alegre, RS, Brasil. SBC.

[Bowyer et al., 2011] Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer, W. P. (2011).
SMOTE: synthetic minority over-sampling technique.
CoRR, abs/1106.1813.

[Saleem Raja et al., 2021] Saleem Raja, A., Vinodini, R., and Kavitha, A. (2021).
Lexical features based malicious url detection using machine learning techniques.
Materials Today: Proceedings, 47:163–166.
NCRABE.