UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Pré-processamento de URLs maliciosos

Regiane de Oliveira Pereira Luz

Trabalho de Conclusão de Curso

MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Pré-processamento de URLs maliciosos

Regiane de Oliveira Pereira Luz

USP - São Carlos

2022

Regiane de Oliveira Pereira Luz

Pré-processamento de URLs maliciosos

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial, Big Data e Segurança.

Orientador: Prof° Marcelo G. Manzato

USP - São Carlos

2022

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

L979p

Luz, Regiane de Oliveira Pereira
Pré-processamento de URLs Maliciosos / Regiane de Oliveira Pereira Luz; orientador Marcelo Manzato; coorientadora Gláucia Maria Saia Cristianini. -- São Carlos, 2022.
37 p.

Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2022.

1. Pré-processamento de dados. 2. Cibersegurança. 3. I.A. Big Data. 4. Ciência de Dados. 5. Análise de Dados. I. Manzato, Marcelo, orient. II. Cristianini, Gláucia Maria Saia, coorient. III. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Dedico este trabalho à renomada Universidade de São Paulo (USP), Fundação de Apoio à Física e à Química (FAFQ), e a Profª Dra. Roseli Aparecida Francelin Romero, pela oportunidade.

A Deus o meu maior mestre, e meu filho Guilherme por seu conhecimento e auxílio.

AGRADECIMENTOS

Aos meus pais e irmão, em memória.

Ao Pastor Joel amado, pela benignidade, amor ao próximo.

Ao meu filho Guilherme Pereira Luz, que esteve lado a lado, acompanhou, dedicou tempo e seu conhecimento. Tenho muito orgulho por sua dedicação em ajudar, incentivar, quando tudo dizia, “não”.
Ao Flávio Luz pelo auxílio.
A Profª Dra. Solange O. Rezende, que nos apoiou em todas as dificuldades.

Ao meu orientador Prof° Marcelo Manzato, professores e tutores agradeço por todo conhecimento, tempo dedicado.

E é claro, a 1º Turma do MBA Inteligência Artificial e Big Data, pessoal dedicado, único, complacente, vocês todos foram demais!

EPÍGRAFE

Foram várias mentes brilhantes por assim dizer, envolvidas com IA, resumidamente a inteligência evoluiu e continuará evoluindo, a tecnologia vem para agregar e não destruir. O que destrói são intensões.

Autora (2022)

RESUMO

LUZ, R. O. P. **Pré-processamento de URLs maliciosos** 2022. 37 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2022.

São numerosos os sites de phishing e ataques maliciosos que transgridem a segurança à privacidade, e à segurança da propriedade dos usuários, a disseminação de programas e scripts maliciosos é um desafio, por sua semelhança aos URL(s) maliciosos e reais e sua demasiada quantidade, será realizado um pré-processamento no banco de dados contendo domínios maliciosos, através da exploração, limpeza e análise, conforme objetivo dessa pesquisa.

Palavras-chave: Pré-processamento, Phishing, Malware, Mineração de dados, DNS.

ABSTRACT

LUZ, R. O. P. **Malicious URLs Preprocessing**. 37f. Completion of course work (MBA in Artificial Intelligence and Big Data) – Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, 2022.

There are numerous phishing sites and malicious attacks that violate security, privacy and the security of users' property, the dissemination of malicious programs and scripts is a challenge, due to their similarity to the malicious and real URL(s) and their excessive amount, a pre-processing will be carried out in the database containing malicious domains, through exploration, cleaning and analysis, according to the objective of this research.

Keywords: Preprocessing, Phishing, Malware, Data Mining, DNS.

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

O contínuo crescimento mundial em redes de computadores e aplicativos, traz diversos benefícios. No entanto, o surgimento de novos ataques cibernéticos, fornece um risco inestimável, que exigem o uso de novos métodos para identificá-los.

Segundo o relatório da Serasa Experian, em 2017, o país registrava uma tentativa de fraude a cada 16 segundos e, de acordo com dados da Axur, entre fevereiro de 2019 a 2020, houve um aumento recorde de 308,17% no volume de phishing. Em 02 de agosto de 2021 a Secretaria da Segurança Pública, Polícia Civil, faz comunicado para prevenção a crimes cibernéticos, o golpe do phishing.

> *"Em 2021, o Brasil pulou de nono para o quarto lugar no ranking de países que mais sofreram tentativas de invasões, atrás somente dos EUA, Alemanha e Reino Unido."* (Serpro, 2021)

Um grande contribuidor com esses acontecimentos é por decorrência da pandemia global do vírus COVID-19. População em geral acessando a internet exposta a todo tipo de vulnerabilidades. Fraudes sempre foram presentes, até mesmo o que é concebido como “phishing”, onde pessoas mal-intencionadas aproveitam oportunidades para tirar proveito ou extorquir suas vítimas na internet.

1.1 Motivação

Privacidade e segurança são coisas essenciais, e atualmente cada vez mais se presencia a invasão de sistemas, a quantidade de malwares dentro da internet é descomunal. Impedir que as pessoas possam roubar informações é de extrema importância. A lei geral de proteção de dados (LGPD), dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade, de privacidade e o livre desenvolvimento da personalidade da pessoa natural (Lgpd *., 2018)[1].

[1] Lei nº 13.709/ 2018 alterada para nº 13.853/2019.

Com a crescente globalização, os recursos digitais, internet, fluxo de dados estão internacionalizados, logo, há necessidade de uma visão holística para uma conformidade de acordo com a expansão da tecnologia no mundo (Pinheiro, 2018).

1.2 Objetivos

Propor que o método de pré-processamento de dados sobre a detecção de sites maliciosos seja baseado na preservação das características das URLs para que, trabalhos como aprendizado de máquina utilizem a técnica de automatizar recursos na representação das URLs, por meio de um conjunto predeterminado coletado de diversos sites, com diferentes amostras no processo de detecção. Ao pré-processar esses dados concentrando-se na relevância dos atributos de phishing e benigno. No **capítulo 1** foi apresentado a introdução, assim como o que motivou o desenvolvimento do trabalho e objetivos desejados; no **capítulo 2** apresenta o embasamento teórico e os trabalhos correlatos utilizados nesta tese; o **capítulo 3** apresenta os trabalhos estudados na literatura sobre descoberta de conhecimento em base de dados, sendo o foco a amostragem de dados e a detecção de agrupamentos, seguido pela justificativa e objetivo; No **capítulo 4** foi apresentado os algoritmos e as técnicas desenvolvidas para a limpeza e organização dos dados, assim como as bibliotecas utilizadas; **Capítulo 5** são apresentados os resultados dos algoritmos, assim como a acurácia; Por fim, o **capítulo 6** apresenta as conclusões chegadas ao final do trabalho.

2 Mineração de Dados

A mineração de dados são aquisição de dados, análise de dados, engenharia de recursos, modelos de treinamento e avaliação de modelos, é usada para transformar dados ruidosos em um formato compreensível, com o processo de limpeza de dados, transformação de dados e redução, cada atividade prosseguindo do processo de pré-processamento de dados, combinação de ferramentas como o aprendizado de máquina e inteligência artificial, com objetivo de análise e propensão em bancos de dados. Na aplicação tal usabilidade é, remover ruídos, dados irrelevantes, valores nulos, similaridade etc.

Este tipo de processo é comumente utilizado para BI (*Business Intelligence*), que consiste em utilizar desses dados para estudar o mercado, como o nome sugere, “inteligência empresarial”. A intensão é tomar decisões da forma mais lógica possível, com o intuito de diminuir prejuízos. Não se pode ter certeza se todas as decisões serão corretas, pois não é possível prever o futuro, e o mercado é bastante imprevisível, porém fazer uso desses dados pode ser de grande valia. Usamos sklearn para o trabalho utilizando para extração de recursos de texto

2.1 Etapas da Mineração de Dados

Iniciado pela coleta de dados, em seguida, pelo pré-processamento, ou seja, a limpeza e organização desses dados, onde são usados os métodos descritos na seção **2.3.1**.

Figure 1 - Etapas da Mineração de dados.

Fonte: Rezende, S. O. Sistemas inteligentes: fundamentos e aplicações. Editora Manole Ltda, 2003.

Identificação do problema:

É realizado uma pesquisa para diagnosticar os problemas existentes ou aqueles que poderão ocorrer, pois em muitos casos o problema é evidente, entretanto a sua causa não.

Pré-processamento:

O intuito é que a base de dados desorganizada e com dados ruidosos, seja limpa, organizada e possivelmente otimizada.

Extração de Padrões:

São retirados e analisados padrões entre os dados coletados, com o intuito de obter maior compreensão sobre a base de dados analisada.

Pós-processamento:

Diferente do pré-processamento que está mais preocupado em limpar a base de dados, este é usado para entender como os padrões obtidos durante as etapas anteriores podem ser utilizados.

Utilização do conhecimento:

O conhecimento obtido através das etapas será utilizado para determinados fins, como tomadas de decisões empresariais.

2.2 PCA

A análise de componentes principais, ou PCA, é uma técnica estatística para converter dados de alta dimensão em dados de baixa dimensão, selecionando os recursos mais importantes que capturam o máximo de informações sobre o conjunto de dados. Os recursos são selecionados com base na variação que causam na saída. O recurso que causa a maior variação é o primeiro componente principal. O recurso responsável pela segunda maior variação é considerado o segundo componente principal e assim por diante. É importante mencionar que os componentes principais não possuem qualquer correlação entre si.

2.3 PRÉ-PROCESSAMENTO

O pré-processamento é um conjunto de atividades que envolvem preparação, organização e estruturação dos dados. Trata-se de uma etapa fundamental que precede a realização de análises e predições.

Detecte e remova dados ruidosos e irrelevantes de conjuntos de dados, processe dados de vulnerabilidade e remova dados em branco, transformação e normalização deve ser realizado. Exemplo: valor ausente, erro, outlier, contradições, etc.

Isso é importante para que seja ocupado menos espaço e seja necessário menos esforço computacional para o armazenamento e processamento dos dados, eles também acabam mais organizados no final do processo, isso porque informações não compatíveis são retiradas, assim, entender o tipo de variável é bastante importante, pois dessa forma pode-se saber o que fazer com cada informação e qual o melhor tipo de tratamento a ser aplicado.

Os dados pré-processados são preparados para outro procedimento de processamento, como mineração de dados, entretanto, recentemente essas técnicas têm sido adaptadas para treinar modelos de aprendizado de máquina e IA (Inteligência Artificial), e para executar interferências sobre eles (Batista, 2003).

Alguns métodos são frequentemente usados, tais quais: **amostragem**, que seleciona um subconjunto representativo de uma grande quantidade de dados, **transformação**, que manipula dados brutos para produzir uma única entrada, "**denoising**", que remove o ruído dos dados, **imputação**, que sintetiza dados estatisticamente relevantes para valores faltantes, **normalização**, que organiza os dados para um acesso mais eficiente e **extração de recursos**, que extrai um subconjunto de recursos relevante que é significativo em um contexto específico. Esses métodos podem ser usados em diferentes fontes de dados, incluindo dados armazenados em arquivos ou bancos de dados.

Melhorar a qualidade dos dados, ajuda na precisão e o desempenho do processo de aprendizado subsequente. Os dados podem ser corrigidos seguindo algumas etapas que serão o foco do trabalho.

- **Limpeza dos dados**
- **Seleção de atributos**
- **Transformação dos dados**

2.3.1 Limpeza dos dados

Remove dados ruidosos, anormais, padroniza, corrige erros, remove dados duplicados, irrelevantes, dessa forma deve-se estabelecer a melhor decisão conforme necessidades. Envolve manuseio de inconsistências.

2.3.2 Seleção de atributos

Tomando a previsão do atributo phishing, dados como defacement e malware podem ser descartados. Melhora a capacidade de generalização do modelo e reduz o overfitting; melhora o entendimento entre recursos e valores de recursos.

2.3.3Transformação dos dados

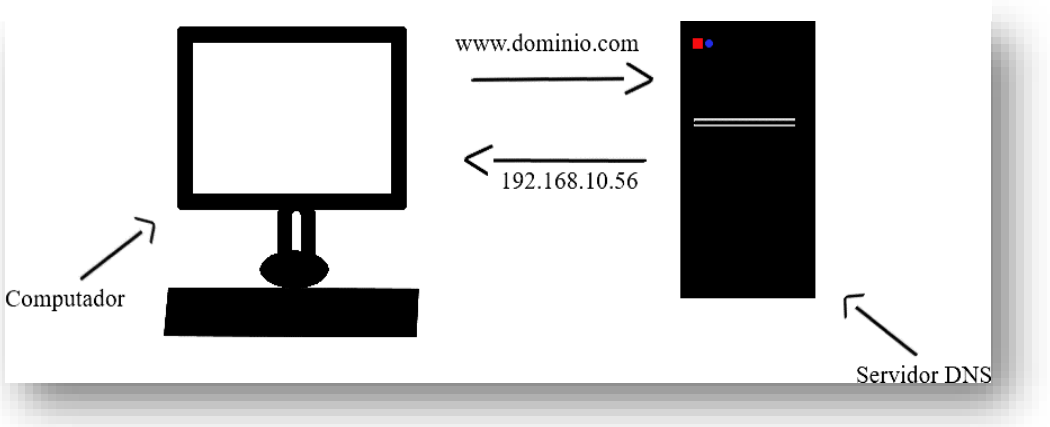
São transformados dados originais e desorganizados em dados estruturados e apropriados para o processo de mineração, são retirados dados iguais, nulos e divergentes.

Dentro dessa etapa são realizados processos como a normalização, que dimensiona os valores dos dados dentro de um intervalo determinado, e a seleção de atributos, onde são gerados novos atributos a partir do conjunto fornecido, com o intuito de ajudar no processo de mineração. Esse processo pode trazer grandes benefícios no que tange a qualidade dos dados e a eficiência do processo

2.4 DNS – Domain Name System

O Sistema de Nomes de Domínio relaciona os nomes de domínios com seus respectivos endereços Ips (*Figura 2*), isso dentro de servidores autorizados. Isso ocorre, pois, é mais simples para humanos memorizar nomes do que sequências de números, porém computadores funcionam apenas com base no sistema binário, ou seja, números,de forma que uma sequência de letras não poderia ser interpretada, para isso, surgiram os servidores dns. Trata-se de um computador com uma base de dados contendo os endereços IPS públicos, estes são associados com o nome do domínio.

Figure 2 - Funcionamento de um servidor DNS.



Fonte: Elaborado pela autora.

O servidor DNS inicia o processo encontrando o endereço IP correspondente para o Localizador Uniforme de Recursos (URL) de um site. Assim que o servidor DNS encontra o endereço IP correto, os navegadores pegam o endereço e o usam para enviar dados para a Rede de Fornecimento de Conteúdo (CDN), após isso, as informações do site podem ser acessadas pelo usuário. Dentre os servidores existem dois tipos principais, os servidores de nome raiz, que são responsáveis por armazenar o nome dos domínios, assim que algum domínio for procurado, o navegador usará desse servidor para fazer o redirecionamento, e os servidores de nomes com autoridade, estes são gerenciados por universidades e grandes empresas que podem optar por construir seu próprio servidor DNS.

URL: significa Uniform Resource Locator. Uma URL é o endereço na Web, cada URL é um recurso podendo ser uma página HTML, um documento CSS, uma imagem e assim por diante. URLs são manipulados pelo servidor web, o proprietário do servidor web precisa cuidar da manutenção do recurso e seu URL associado. Um URL consiste em diferentes partes:

HTTP/ https: Indica o protocolo do navegador, a web requer seu uso.
www.exemplo.com: é o nome do domínio, indica qual servidor web está sendo solicitado. O endereço IP pode ser usado diretamente.
/caminho/para/arquivo.html: é o caminho para o recurso no servidor web, destino.
chave.valor?chave=valor1&chave: são parâmetros adicionais fornecidos ao servidor web, são uma lista de pares chave/valor separados por símbolos, cada servidor web tem suas próprias regras sobre esses parâmetros.

Dentro da estrutura de um URL estão presentes os domínios de alto nível, como: .com, .edu, .net e .gov, e domínios de países como: br, fr, ru, cada qual indica algo a respeito do sistema, por exemplo ".gov" indica algum órgão do governo. Um servidor DNS malicioso é um servidor que está fornecendo respostas incorretas para nomes de domínios.

2.4.1 Ataques ao DNS

O propósito de um servidor DNS malicioso *(Gráfico 1)*, é direcionar os usuários para sites falsos, como parte de ataques de pharming *(Cert.br, 2021).*

No geral, os cibercriminosos fazem uso desse método para roubar dados de suas vítimas, como por exemplo, nome, CPF, RG, número do cartão etc. Existem formas diferentes de se executar esse ataque:

- **DNS Hijacking:**

Direciona o usuário para domínios não autorizados, iss faz com que o usuário pense que está no lugar desejado, mas na verdade está no domínio do invasor, pois a página ainda é a mesma, entretanto o código é diferente, ele é programado para obter os dados sem consentimento.

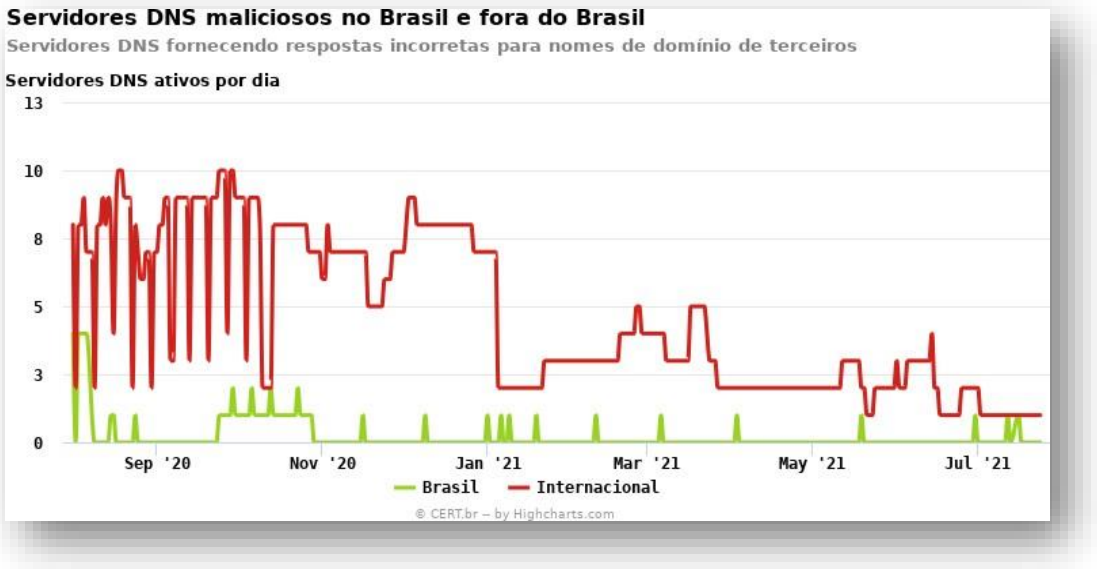
- **DNS Cache Poison**:

Nesse caso o servidor é infectado, mudando o direcionamento dos domínios, de forma que o IP será alterado, porém o domínio será o mesmo, assim que o usuário entrar, seus dados serão roubados.

- **Man-in-the-middle:**

A comunicação entre o servidor e o usuário é interceptada, assim, o criminoso pode observar tudo o que ocorre durante a conexão, e também, enviar e receber dados, isso sem ser detectado. Instituições vítima, em geral instituições financeiras, de comércio eletrônico, redes sociais ou domínios. Em sua maioria estes servidores são instalados pelo próprio invasor, contratandoserviços de hospedagem ou de nuvem.

Gráfico 1 - DNS maliciosos



Fonte: (Cert.br, 2021)

2.5 Phishing

É um processo fraudulento onde o objetivo é obter informações da vítima, comumente feito por e-mails, mensagens de texto, sms ou sites, o invasor faz com que a página acessada pela vítima seja visualmente semelhante à página original, porém o back-end delas é programado para coletar dados. Está diretamente ligado a engenharia social, visto que, ela se relaciona com a manipulação psicológica de pessoas. De forma que, o invasor pode se passar por alguém conhecido da vítimapara que ela seja mais facilmente manipulada. Tanto empresas quanto a própria comunidade têm se interessado em classificar quais domínios podem ser maliciosos, com o intuito de impedir que mais pessoas caiam nesse tipo de golpe. Dentre os métodos e técnicas de phishing se destacam:

- **Spear phishing**

Comumente executado por e-mail ou outro tipo de comunicação eletrônica, pode ser direcionado a um indivíduo, organização ou empresa, porém o que o diferencia é ser focado alguém específico e, normalmente inclui informação de interesse da vítima, como documentos financeiros por exemplo;

- **Watering hole**

Diferente do spear phishing este não faz uso de comunicação, são observados quais sites são normalmente acessados pela(s) vítima(s), de forma que alguns desses sites serão infectados com malware, eventualmente a vítima será infectada;

- **Whaling**

Semelhante ao spear phishing, entretanto os alvos são pessoas importantes, como CEOs por exemplo, são utilizados métodos como a falsificação de e-mails e sites com malware, pode ser feito uso da engenharia social.

3 TRABALHOS RELACIONADOS

Durante as pesquisas, foram encontrados trabalhos que visam pré-processamento e mineração de dados. A seguir, são apresentados alguns dos mais relevantes. A Seção 3.1 apresenta o motivo para o desenvolvimento deste trabalho, já a Seção 3.2 apresenta trabalhos relacionados a este.

3.1 Justificativa

Conforme a tecnologia avança pode-se notar como cada vez surgem mais ataques e mais exploits de forma que, se faz necessário que a defesa avance junto, a estatística de phishing apenas aumenta a cada pesquisa. A realização do pré-processamento em um Dataset contendo dados de URLs maliciosos e benignos, trata-se de uma etapa fundamental tornando a classificação e reconhecimento desses dados mais rápida e eficaz. O objetivo é que ao final da classificação seja possível reconhecer quais domínios podem conter ameaças e quais são confiáveis. O pré-processamento surge como uma etapa fundamental para o aumento da produtividade e qualidade dos dados, ele pode tornar dados desorganizados e irregulares em dados estruturados. Batista G., descreve em sua Tese “Pré-processamento de Dados em Aprendizado de Máquina Supervisionado” o tratamento de um conjunto de dados malicioso, e define as limitações quanto ao tratamento de valores desconhecidos.

> *"Podem existir diversos objetivos na fase de pré-processamento. Um deles é solucionar problemas nos dados, tais como identificar e tratar dados corrompidos, atributos irrelevantes e valores desconhecidos" (BATISTA, 2003, p.3).*

Datasets são bases de dados, estes podem ser usados para diversas finalidades, no Trabalho de Conclusão de Curso de Martini I. "Geração de Datasets com Ataques Criptografados para IDS", foram gerados *Datasets* cifrados como medida de impedir o contínuo crescimento de ataques, juntamente com os IDSs (Sistemas de Detecção de Intrusão) que tentam reconhecer uma invasão ao sistema (MARTINI, 2017, p.9).

"*Python Data Science Handbook*" escrito por *Vander J. P.*, descreve o uso de python para ciência de dados, junto com formas de fazê-lo, explicando as principais bibliotecas utilizadas, como pandas, numpy, matplotlib, entre outros. Estas informações foram bastante úteis durante o desenvolvimento do trabalho.

"*Técnicas de Invasão*" criado por *Fraga B.*, é focado em segurança da informação e hacking, explica como a maioria dos ataques acontecem e diversas técnicas de efetuar ataques, discute a respeito da ética hacker, *White Hat*, *Black Hat* e *Gray Hat*. Informa como redes funcionam e como hackers fazem uso dela para efetuar seus ataques.

4 METODOLOGIA

Ao analisar o conjunto de dados de URLs maliciosos, temos um total de 651191 titularizadas *url* e *type,* na coluna type estão atribuidos por benign, phishing, defacement e malware. Trata-se de um banco de dados semiestruturado coletado de várias fontes de todo tipo e lugar, dados codificado e extraído com ruídos. Foi necessário interpretar essas URLs para estar em conformidade, decodificando, codificando o tipo de caracteres, extraindo caracteres, convertendo dados inconsistentes, granularidade. Detalhe importante foi a observação aos caracteres especiais, nomes de domínio internacionalizados (IDNs), texto bidirecional devendo ser tratados com cuidado para evitar a descaracterização.

Dados como *defacement* e *malware* foram removidos para definir o reconhecimento realizando um processo de limpeza *(Gráfico 4),* transformação e normalização, que dimensiona os valores dos dados dentro de um intervalo determinado, os dados a ser dimensionados são os dados da coluna *type*, chamados *benign* e *phishing*, pelo qual, foram convertidos em recursos numéricos *(0 e 1)*, 0 para benign e 1 para phishing. É injuntivo mencionar que um conjunto de recursos deve ser normalizado antes de aplicar o PCA, por ser um método estatístico e só poder ser aplicado a dados numéricos.

Usamos sklearn com o método transform para transformar os recursos, a transformação é não supervisionada que se refere à transformação que utiliza apenas informações estatísticas de recursos, incluindo média, desvio padrão, limite, etc., como padronização, redução de dimensionalidade do método PCA. O PCA diminui o número de recursos elegendo a dimensão dos recursos com maior variação, é indicado para visualizar melhor os dados e aumentar a precisão, converte os dados do espaço de alta dimensão em espaço de baixa dimensão, selecionando os atributos mais importantes. O principal trabalho é obter informações de características e informações de valor alvo específico ao url malicioso do tipo phishing.

4.1 Considerações iniciais

Com o objetivo de pré-processar conjunto de dados os mantendo o mais fiel possível para proveito de apuração de suas características. O conjunto de dados selecionados com 651191 de urls divididos entre benigno, phishing, desfigurado e malware (*Gráfico 2*), são típicos do atual cenário que convivemos.

Dataset: Malicious_Phish, extraído da comunidade de cientistas de dados Kaggle.

Kaggle: é uma empresa subsidiária da Google, é uma comunidade de cientistas de dados e programadores, nela são compartilhadas diversas bases de dados que podem ser utilizadas para aprendizado, treinamento e outros fins.

Ambiente: Jupyter Notebook, é um software gratuito, organização que não possui fins lucrativos, foi criada com a intenção de desenvolvimento, computação interativa de código aberto, para ajudar a comunidade de programadores. Possui suporte para diversas linguagens de programação.

4.2 Base de Dados

O conjunto de dados coletado é um tipo específico de URL malicioso, conhecido como phishing, localizado em: https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset para download *(acesso em 2022).*

Seu conteúdo são explorações em vários ambientes vulneráveis, para detecção e categorização de URLs maliciosos de acordo com seu tipo de ataque (ISCX-URL-2016).

O Dataset está dividido entre duas colunas, uma contém o nome dos domínios, a outra contém seus tipos, que podem ser: benign, phishing, defacement e malware. O objetivo é pré-processar e classificar de acordo com seu tipo específico.

4.3 Bibliotecas: descrição

- **Seaborn**

É uma biblioteca usada para visualização de dados, é baseado na biblioteca matplotlib. Ele traz uma interface que pode ser usada para desenhar gráficos estatísticos.

- **Numpy**

É um sucessor do Numeric, que foi criado por *Jim Hugunin*, com ajuda de outros desenvolvedores. Em 2005, *Travis Oliphant* criou o Numpy, que tem suporte para arranjos e matrizes, e possui bastantes funções matemáticas de alto nível que operam sobre elas, também pode ser usado para álgebra linear.

- **Pandas**

Foi criado por *Wes McKinney*, que começou o seu desenvolvimento de 2008, é utilizado para manipulação e análise de dados, oferece estrutura para manipulação de tabelas numéricas e séries temporais, é uma biblioteca de código aberto.

- **Scikit-learn**

Em 2007 começou como um projeto de *David Cournapeau*, que iniciou dentro do programa Google Summer of Code, e no mesmo ano *Matthieu Brucher* trabalhou no projeto como parte de sua tese de doutorado. É uma biblioteca de código aberto que pode ser usado para aplicação prática de aprendizado de máquina, comumente usados para mineração e análise de dados baseado em Numpy, Scipy e Maltplotlib.

- **Matplotlib**

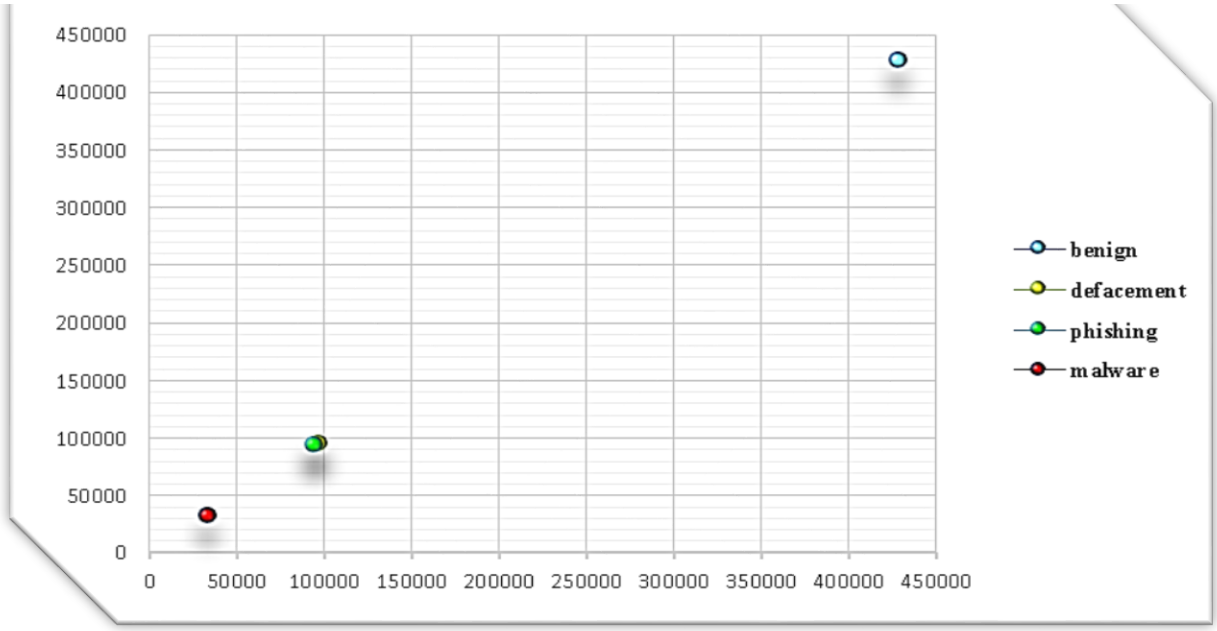
Foi originalmente criado por John D. Hunter, é uma biblioteca utilizada para criação de gráficos e visualização de dados, pode ser usada para criar visualizações estáticas, animadas e interativas em python. Matplotlib é uma biblioteca de código aberto.

4.4 Aplicação do pré-processamento

Leitura do Dataset *(Gráfico 3)*, em uma breve análise nota-se que a maior parte das URLs são benignos, mesmo em relação aos outros três tipos. Os domínios benignos ultrapassam mais de 400.000 unidades, e pode-se notar que dentre todos, o malware é o de menor quantidade, o que evidencia a grande quantidade de phishing(*Gráfico 2*).

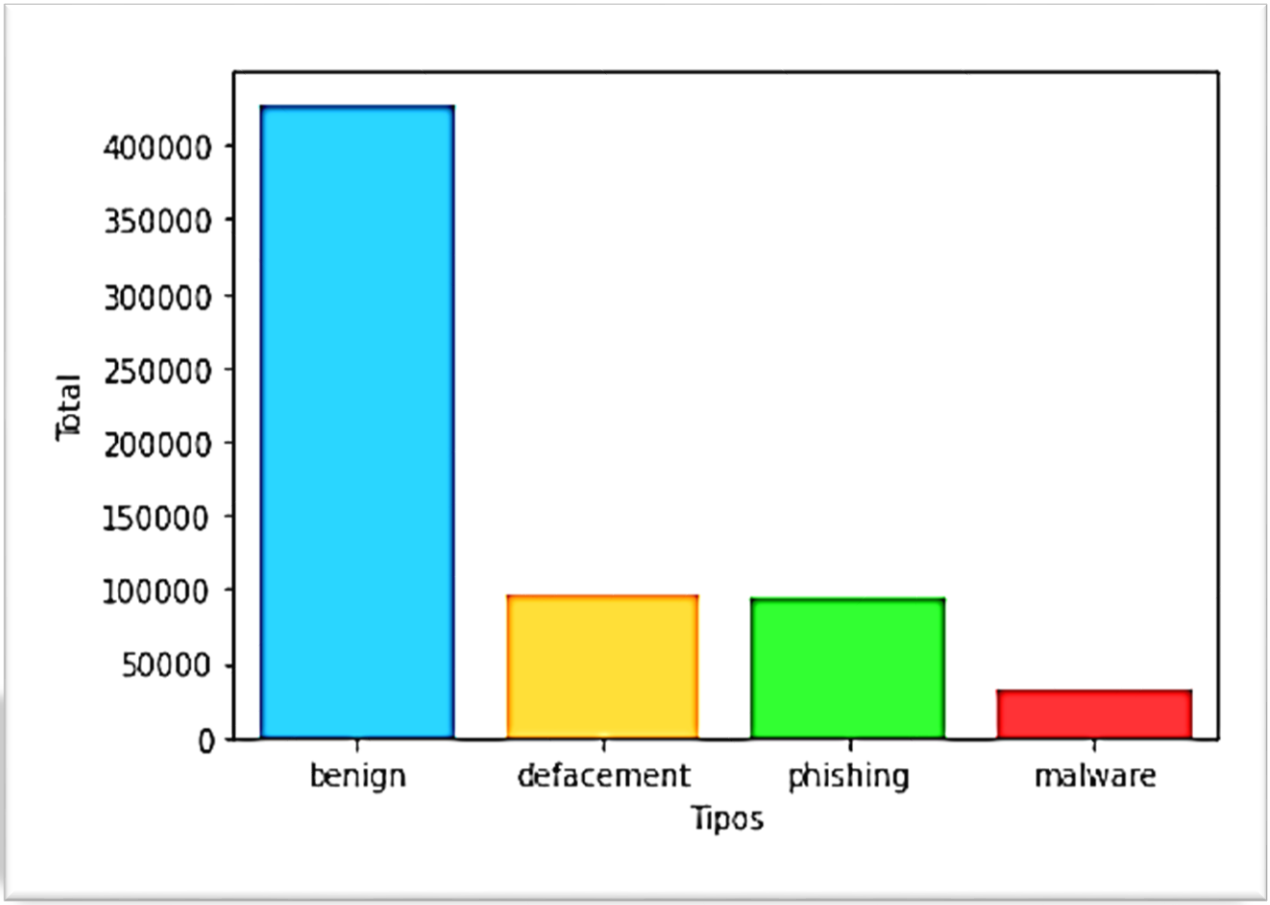
Gráfico 2 - Totais tipos URLs.

.



Fonte: Elaborado pela autora.

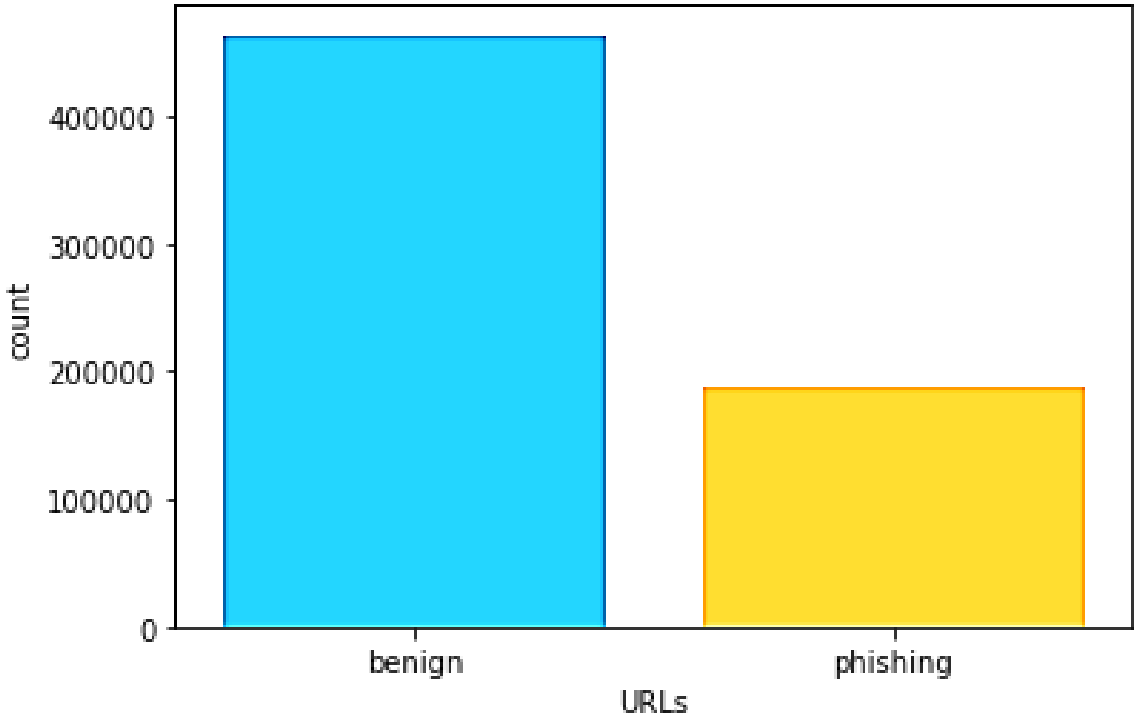
Gráfico 3 - Banco de Dados



Fonte: Elaborado pela autora.

O propósito está no reconhecimento de phishing, portanto no processo de limpeza foi extraído as URLs defacement e malware (*Gráfico 4*), Essas URLs são elementos, com ou sem atributos de definições, formulários diferentes, imagem, vídeos, jogos, etc. a decodificação é necessária para a validade da arquitetura das URLs. Não constando espaços vazios ou ambiguidade, mas com muitos caracteres de codificações distintas. A decodificação de um URL codificado em UTF-8 converte bytes que representam caracteres não ASCII nos caracteres especiais apropriados, sendo assim, todos os caracteres não ASCII são codificados em UTF-8.

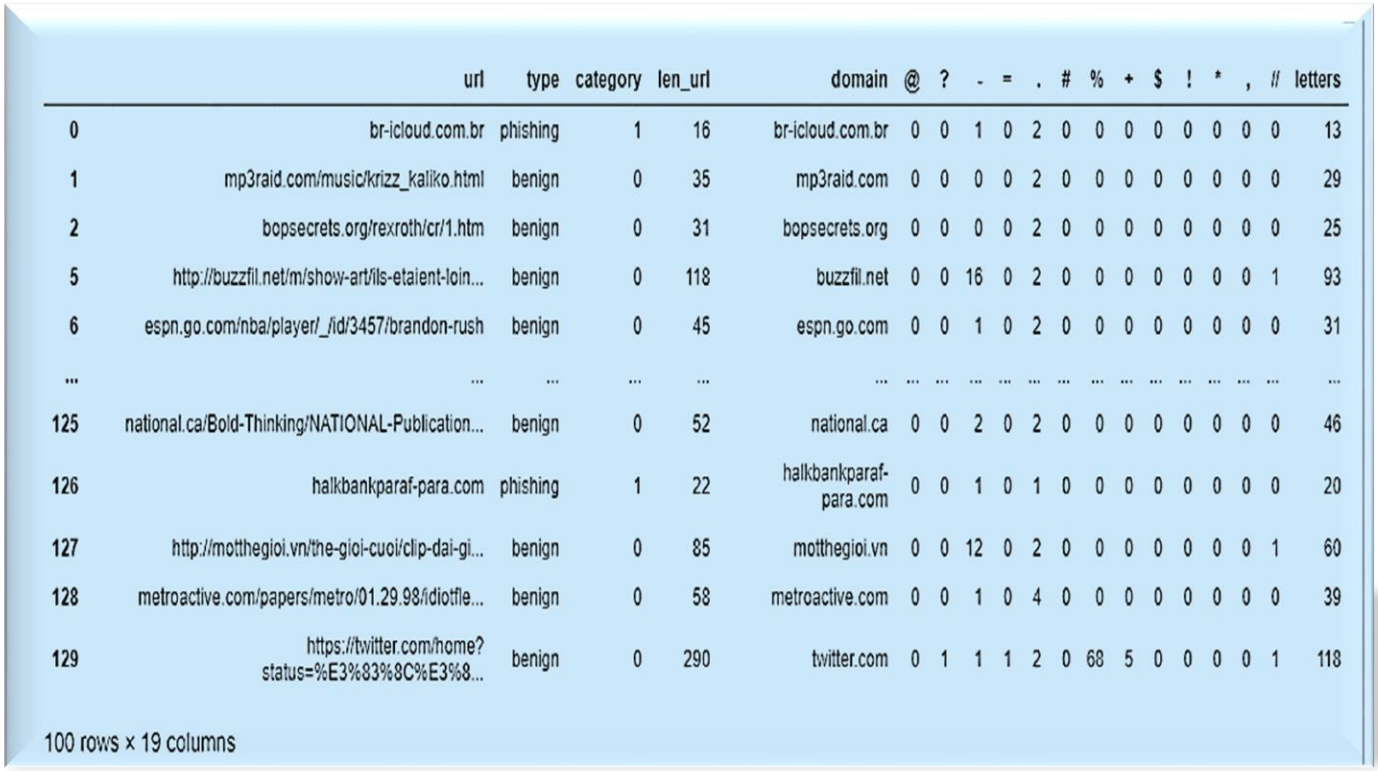
Gráfico 4 – URLs benigno e phishing.



Fonte: Elaborado pela autora.

Com scikit-learn foi analisado a quantidade de repetições de letras e números, tamanho da URL, com *feature_names_in_ ndarray de forma (n_features_in_),* preparamos a base de dados para extrair os resultados dos tipos de dados e quantas vezes caracteres como “@”, “?”, “-“, “=”, “.”,“#”, “%”, “+”, “$”, “!”, “*”, “//” se repetem *(Tabela 1).* Geramos objetos X e Y para armazenar os dados.

Tabela 1 - Caracteres URLs.



	url	type	category	len_url	domain	@	?	-	=	.	#	%	+	$	!	*	,	//	letters
0	br-icloud.com.br	phishing	1	16	br-icloud.com.br	0	0	1	0	2	0	0	0	0	0	0	0	0	13
1	mp3raid.com/music/krizz_kaliko.html	benign	0	35	mp3raid.com	0	0	0	0	2	0	0	0	0	0	0	0	0	29
2	bopsecrets.org/rexroth/cr/1.htm	benign	0	31	bopsecrets.org	0	0	0	0	2	0	0	0	0	0	0	0	0	25
5	http://buzzfil.net/m/show-art/ils-etaient-loin...	benign	0	118	buzzfil.net	0	0	16	0	2	0	0	0	0	0	0	0	1	93
6	espn.go.com/nba/player/_/id/3457/brandon-rush	benign	0	45	espn.go.com	0	0	1	0	2	0	0	0	0	0	0	0	0	31
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
125	national.ca/Bold-Thinking/NATIONAL-Publication...	benign	0	52	national.ca	0	0	2	0	2	0	0	0	0	0	0	0	0	46
126	halkbankparaf-para.com	phishing	1	22	halkbankparaf-para.com	0	0	1	0	1	0	0	0	0	0	0	0	0	20
127	http://motthegioi.vn/the-gioi-cuoi/clip-dai-gi...	benign	0	85	motthegioi.vn	0	0	12	0	2	0	0	0	0	0	0	0	1	60
128	metroactive.com/papers/metro/01.29.98/idiotfle...	benign	0	58	metroactive.com	0	0	1	0	4	0	0	0	0	0	0	0	0	39
129	https://twitter.com/home?status=%E3%83%8C%E3%8...	benign	0	290	twitter.com	0	1	1	1	2	0	68	5	0	0	0	0	1	118

100 rows × 19 columns

Fonte: Elaborado pela autora.

. Seguimos para normalização com *StandardScaler* para dimensionar os dados tornando a média igual a*"0"* e desvio padrão igual a "*1"*, é criado o objeto *std_scl* esse processo foi feito para usar *standardScaler* e executar operações sobre a base de dados. Aplicamos o PCA importando de sklearn.decomposition que faz o pré-processamento não supervisionado. O PCA é baseado na "transformação linear ortogonal", que é uma técnica matemática para projetar os atributos de um conjunto de dados em um novo sistema de coordenadas, fazendo análise de componente principal, converte assim, os dados do espaço de alta dimensão em espaço de baixa dimensão, selecionando os atributos mais importantes que capturam o máximo de informações sobre o conjunto de dados.

De modo geral, o resultado que esperamos é projetar o espaço de características dos dados originais (amostras n d-dimensionais) em um subespaço menor e expressá-lo da melhor maneira possível sem que perdemos informações relevantes, uma aplicação comum é no reconhecimento de padrões. Podemos expressar melhor nossos dados reduzindo a dimensão do espaço de características e extraindo dados do subespaço, reduzindo assim o erro de estimação de parâmetros.

5 RESULTADOS OBTIDOS

Observamos que um invasor com a intenção de evitar a análise estática em recursos lexicais de URL usa técnicas de ofuscação para que URLs maliciosos se tornem estatisticamente semelhantes aos benignos. Grande quantidade de dados de URL maliciosos mostra que os invasores geralmente usam URLs muito longos na tentativa de mascarar partes suspeitas do URL, podemos ter variáveis que não apresentam informação útil, isso faz com que o modelo selecionado possua muitos parâmetros, o que pode causar *overfitting*, para evitar utilizamos técnicas de redução de dimensionalidade. Geralmente resulta em uma matriz de covariância e uma matriz de correlação, essas matrizes podem ser calculadas a partir de dados brutos. A matriz de covariância contém a soma dos produtos quadrados e vetoriais. A matriz de correlação é semelhante à matriz de covariância, mas a primeira variável, a primeira coluna, são os dados normalizados. Se a variância entre as variáveis for grande, ou as dimensões das variáveis não forem uniformes, devemos primeiro padronizar e depois realizar a análise de componentes principais para que os componentes principais não seja confundidos com pequena variância e cause algum descarte.

Matriz de confusão

A matriz de confusão traz o resultado do desempenho de algoritmo de classificação, calculando a quantidade de falso positivo e falso negativo, e de verdadeiro positivo e verdadeiro negativo, assim como a acurácia. Esse passo já consideramos o treinamento para visualizarmos em nossos testes a porcentagem em reconhecimento do mal phishing, aplicando esse conceito foi possível chegar em determinados valores (*Tabela 2*).

Figure 3 - Matriz de confusão

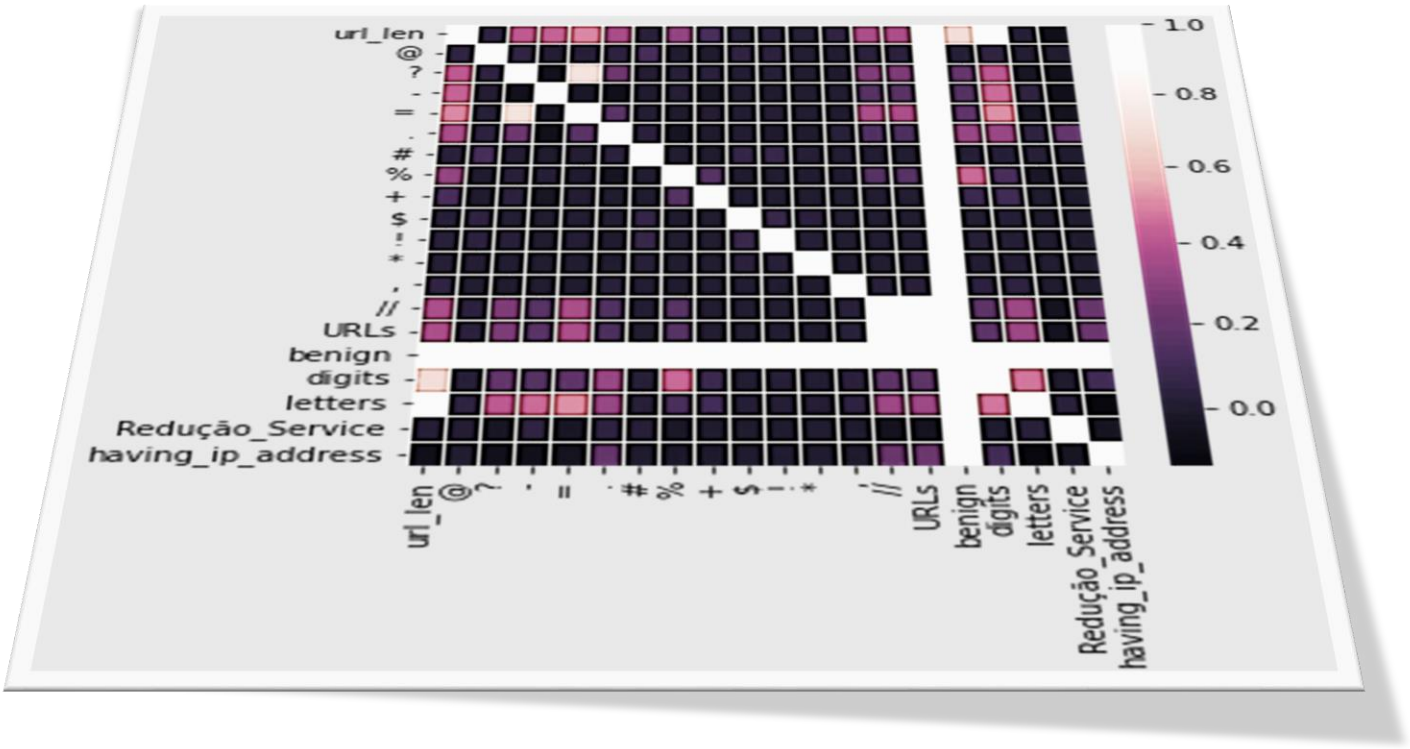
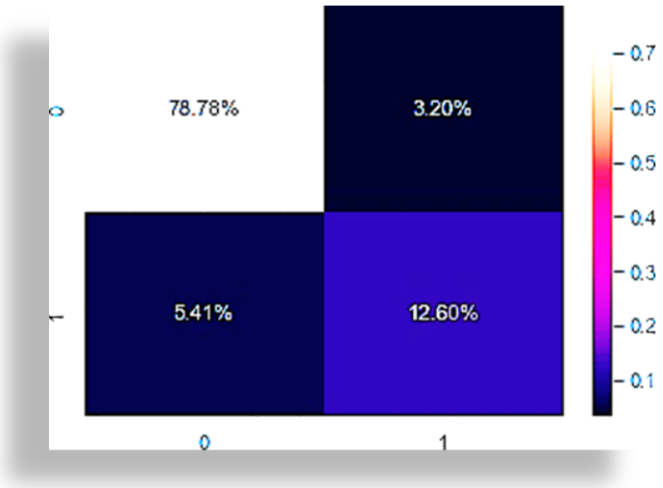
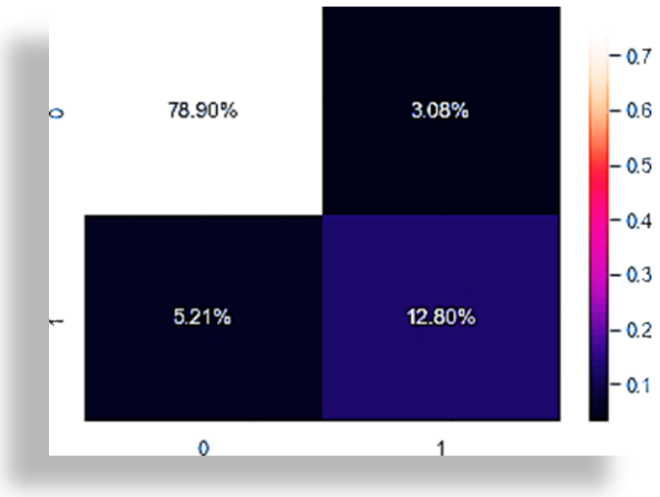


Gráfico 5 – Decision Tree Classifier.



Test Accuracy : **91.39%**

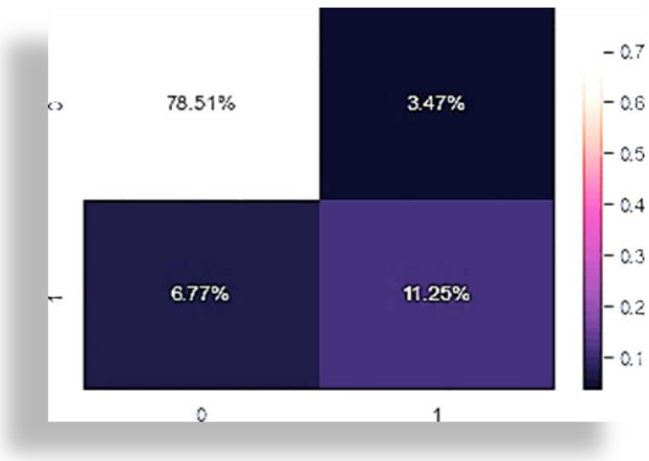
Gráfico 6 - Random Forest Classifier.



Teste Accuracy: **91.71%**

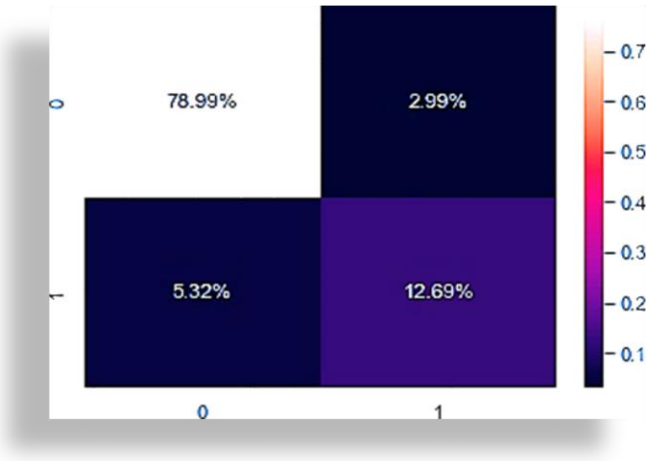
Fonte: Elaborado pela autora.

Gráfico 7 - KNeighbors Classifier.



Test Accuracy: 89.76%

Gráfico 8 - Extra Trees Classifier.



Test Accuracy: 91.68%

Tabela 3 – Resultados Classificações.

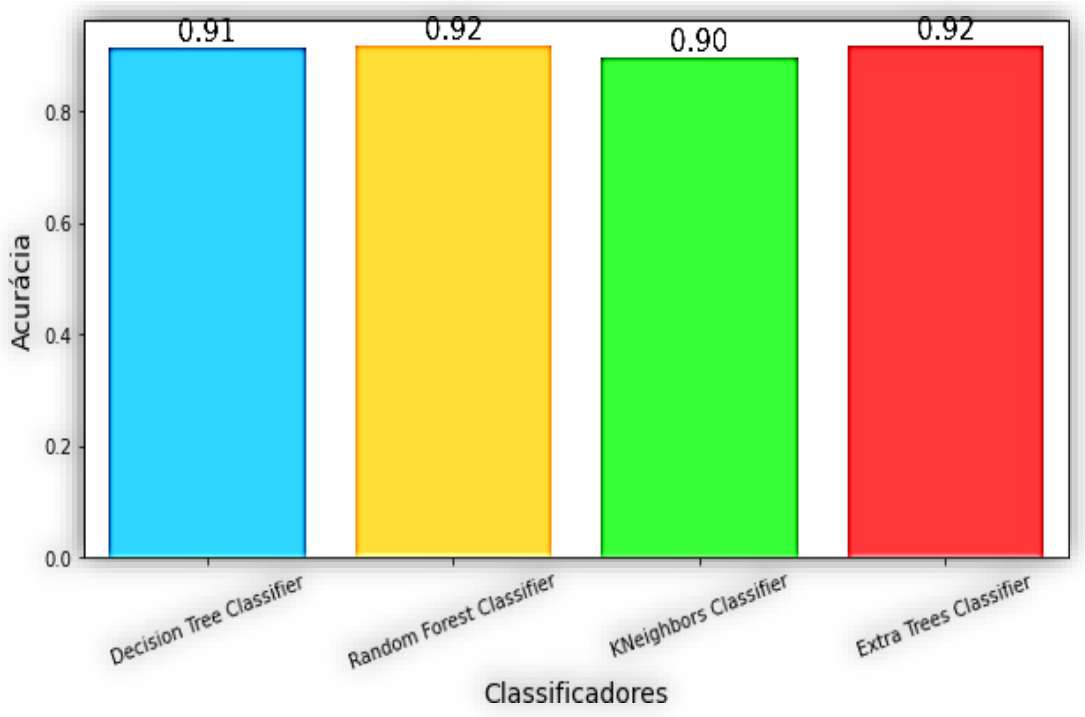
Decision Tree Classifier	**91.39%**
Random Forest Classifier	**91.71%**
KNeighbors Classifier	**89.76%**
Extra Trees Classifier	**91.68%**

Fonte: Elaborado pela autora

5.1 Acurácia

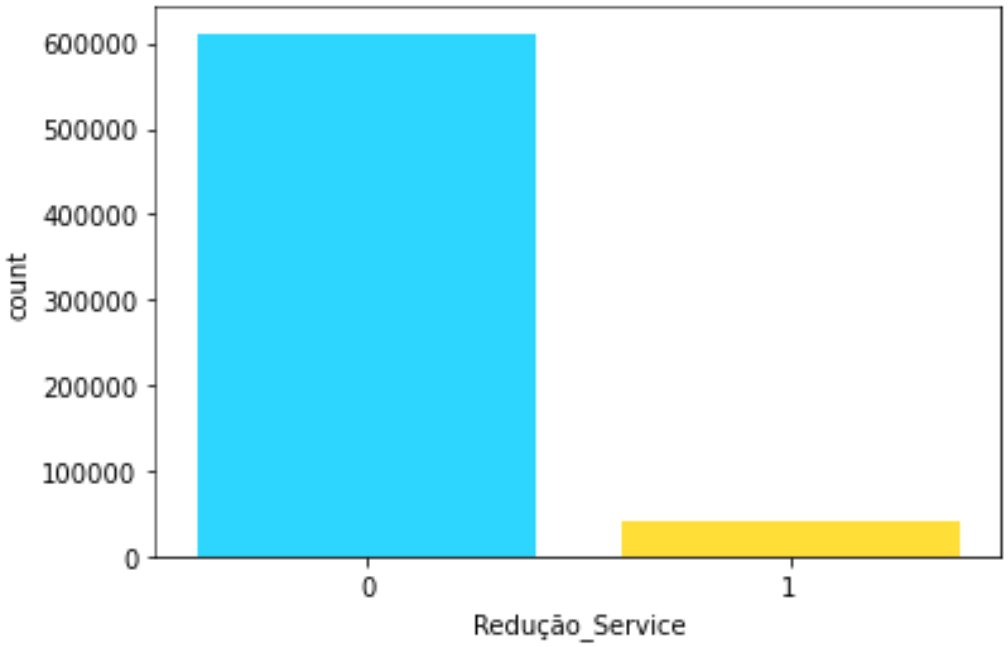
Quanto maior a acurácia, mais autêntico é o resultado, ela mede a proximidade entre o valor obtido experimentalmente e o valor verdadeiro.

Gráfico 9 - Acurácia.



Fonte: Elaborado pela autora.

Gráfico 10 - Redução



Fonte: Elaborado pela autora.

Com a redução dos caracteres, esses seriam os dados de phishing que a máquina reconheceria, notável a perda significante do malicioso phishing. O que nos faz observar que com a redução dos caracteres perdemos dados que devem ser reconhecidos e não reduzidos.

CONCLUSÃO

As bases de dados comumente são grandes aglomerados de dados que podem conter diversas inconsistências, portanto, o pré-processamento surge como uma forma eficiente de se melhorar o desempenho de um aprendizado de máquina, por exemplo. Como forma de diminuir a disseminação de phishing, se faz útil uma base de dados limpa, onde pode-se executar testes para compreender similaridades entre eles.

URLs contém estruturas normalizadas, de forma que não se fez necessário exclusão de itens incompletos, entretanto, foi importante analisar algumas de suas características, como seu tamanho, quantidade de letras, números e caracteres especiais. Com os gráficos, e figuras temos uma melhor compreensão visual desses dados, portanto bibliotecas como pandas, seaborn e matplotlib ajudaram no desenvolvimento.

Mesmo com valores diferenciados não se fez necessário excluir dados faltantes, pois é importante que durante o processo de limpeza de dados, esses dados sejam mantidos o mais fiel possível, visto que grande parte dos trabalhos correlacionados descaracterizam a base de dados. Em uma supervisão automatizada resultaria. Há de se pontuar a importância de diferenciar os tipos de dados maliciosos, para que, em aprendizado de máquina, por exemplo, seja possível distinguir com precisão.

REFERÊNCIAS

Aggarwal, C. C. (2018). *Neural Networks and Deep Learning: The Textbook* ((1ª ed.) ed.). USA: Springer International Publishing. doi:https://doi.org/10.1007/978-3-319-94463-0

Apwg. (2022). *Relatório de tendências de atividade de phishing, 1º trimestre.* Acesso em 07 de junho de 2022, disponível em www. reportphishing@apwg.org

Ashby, W. R. (1970). *Introdução à Cibernética,* (U. d. Paulo, Ed.) Perspectiva S.A.

Batista, G. (2003). Pré-processamento de Dados em Aprendizado de Máquina Supervisionado. *Tese - Doutorado*, 232. São Carlos, S.P., Brasil: Instituto de Ciências Matemática e de Computação - ICMC.

Berners-Lee, Masinter & McCahill. (dezembro de 1994). *Rfc 1738.* (T. B. Lee, L. Masinter, & M. McCahill, Eds.) Fonte: URL:http://www.acl.lanl.gov/URI/archive/uri-archive.index.html

Cert.br. (2021). *Servidores DNS maliciosos no Brasil e fora do Brasil.* Fonte: https://www.cert.br/stats/dns-malicioso/dns-malicioso.png

chenjiandongx. (2022). *Pyecharts*. Acesso em 2022, disponível em Github: https://github.com/pyecharts/pyecharts/blob/master/README.en.md

Civil, P. (02 de agosto de 2021). *Secretaria da Segurança Pública.* Acesso em 2022, disponível em https://www.policiacivil.sp.gov.br/portal/faces/pages home/noticias/

Fraga, B. (2019). *Técnicas de Invasão* (2017 ed., Vol. 253). (T. VANGLER, Ed.) Londres: Labrador.

J., V. (04 de março de 2021). *Brasil é o país com maior número de phishing na internet.* (D. Griesinger, Ed.) Acesso em 2022, disponível em https://agenciabrasil.ebc.com.br/geral/noticia/2021-03/brasil-e-o-pais-com-maior-numero-de-vitimas-de-phishing-na-internet

Keras. (2022). *Plataforma de código aberto.* Fonte: Keras: https://Keras.io/

Lgpd. (08 de julho de 2019). *Lei nº 13.853.* Acesso em 2022, disponível em http://www.planalto.gov.br/ccivil 03 ato2019-2022/2019lei/13853.htm/

Lgpd, *. (2018). *Lei nº 13.709 (Alterada pela nº 13.853).* Fonte: http://www.planalto.gov.br/ccivil 03 ato2019-2022/2019lei/13853.htm/

Marquesone, R. (2017). *Big Data técnicas e tecnologias para extração de valor dos dados.* Casa do Código. Fonte: www.casadocodigo.com.br

Martini, H. I. (2017). Geração de Datasets com Ataques Criptografados para IDS. *Bacharelado*, 75. Santa Cruz do Sul.

Matlab. (1984 - 2016). *Revised.* Acesso em 2022, disponível em Math Works: www.mathworks.com

Matplotlib. (2022). *Plataforma de código aberto*. Fonte: Matplotlib: https://matplotlib.org/

Mertz, D. (2003). *Text Processing in Python* (CRS 0706050403 - Copyright ed.). (P. E. Inc, Ed.) Addison Wesley Pearson. Acesso em 2022

Mitchell, R. (2018). *Web Scraping with Python* (2º ed.). (R. Prates, Ed., & L. A. Kinoshita, Trad.) O'Reilly Media, Inc. Fonte: www.novatec.com.br

Numpy. (2022). *Plataforma de código aberto*. Fonte: Numpy: https://numpy.org/news/

Oliveira, M. D., & Araújo, R. M. (1985). Sistemas de informações seletivas especializadas. *Revista de Biblioteconomia de Brasília, 13*(1). Acesso em 18 de julho de 2022, disponível em http://hdl.handle.net/20.500.11959/brapci/76915

ORG, S. (2022). *SEABORN: statistical data visualization*. Acesso em 2022, disponível em SEABORN: https://seaborn.pydata.org/

Org, S. L. (2022). Acesso em 2022, disponível em https://scikit-learn.org/stable/

Pinheiro, P. P. (2018). *Proteção de dados pessoais.* São Paulo: Saraiva Educação.

Serpro. (2021). *Dificuldades reais do mundo virtual.* Acesso em 2022, disponível em Serpro campanha Soc: https://campanhas.serpro.gov.br/segurança/cubo/?utm source=portal&utm medium=materia&utm campaign=vem-pro-cubo&utm content=thursday0707--soc

Siddhartha, M. (2022). *Conjunto de dados de URLs maliciosos*. Acesso em 2022, disponível em Kaggle: https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset

Team, P. D. (2022). Acesso em 2022, disponível em https://pyecharts.org/#/

Tensorflow. (2022). Acesso em 2022, disponível em https://www.tensorflow.org/?hl=pt-br