Relatório CRISP-DM [1,2,3]

Tema:	Análise de metadados de sites Phishing
Alunos(as):	Augusto Grandini Serrano, Caio dos Santos Lopes e Eduardo Pires Carvalho
Curso:	Sistemas para internet
Disciplina:	Mineração de Dados
Professor:	Prof. Dr. Mário Popolin Neto

1. Domínio e Situação Problema (Business Understanding)

Com a evolução dos sistemas de informação, das ferramentas de segurança e dos processos organizacionais, os ataques cibernéticos [7] que exploram vulnerabilidades técnicas têm ficado cada vez mais difíceis de serem executados. No entanto, diferentemente de ataques que exploram vulnerabilidades técnicas, a engenharia social [8], um tipo de ataque popular que utiliza manipulação psicológica para enganar alvos, foca no elo mais fraco de qualquer sistema: o fator humano. [8] Dentro da engenharia social existe o Phishing [9] - em português: Pescaria -, que utiliza e-mails, mensagens ou sites fraudulentos para enganar as vítimas e "pescar" dados confidenciais, como senhas, dados de cartão de crédito, dados bancários ou até mesmo instalar malwares.

Em 2020, o Brasil foi o país mais atingido por tentativas de roubo de dados pessoais ou financeiros de pessoas na internet. O levantamento foi feito pela empresa de segurança da informação Kaspersky sobre práticas de phishing e spam no mundo. [9] Os golpes foram aplicados por meio de links em mensagens ou sites falsos, que se passam por empreendimentos conhecidos, como grandes cadeias de varejo online - Amazon e outras. [10] Aplicativos de comunicação, especialmente o Whatsapp, tornaram-se os principais canais para aplicar esses golpes. Usuários receberam mensagens com promessas de prêmios com links que levavam a sites falsos destinados a roubar informações da vítima. [10]

O percentual de usuários brasileiros que tentou abrir pelo menos uma vez *links* enviados para roubar dados representa 19,9% dos internautas do país. Em

segundo lugar no *ranking* de países vem Portugal (19,7%), seguido da França (17,9%), Tunísia (17,6%), de Camarões (17,3%) e da Venezuela (16,8%). [10] Apesar do alto índice, vale destacar uma queda importante em relação a 2019. Naquele ano, mais de 30% dos brasileiros haviam tentado, ao menos uma vez, abrir um link que levava a uma página de phishing, dez pontos percentuais a mais do que em 2020. Isso mostra que as campanhas e alertas sobre esse tipo de golpe têm deixado as pessoas mais atentas, mas não significa que não precisamos evoluir, pois as estatísticas permanecem muito ruins - avalia Fabio Assolini, analista sênior de segurança da Kaspersky no Brasil. [10] No devido contexto torna-se oportuno desenvolver um modelo capaz de reconhecer possíveis links fraudulentos.

O presente trabalho se debruça sobre o *Phishing* [9] que utiliza de sites mal intencionados para enganar suas vítimas. Será analisado um conjunto de metadados - quantidade de caracteres da URL, possíveis campos escondidos, ofuscação, protocolo de navegação etc. - de diversos sites maliciosos e não maliciosos, com o objetivo de entender quais conjuntos de padrões podem definir possíveis golpes. Dessa forma, esses dados ajudarão a formar um modelo capaz de reconhecer possíveis sites falsos.

2. Descrição dos Dados (Data Understanding)

O conjunto de dados utilizado foi retirado do site archive.ics (UC Irvine) [6], possui 54 variáveis e 235.795 instâncias de URLs. Sendo 134,850 legítimas e 100,945 phishing.

Dicionário de Dados		
Variável	**Descrição**	**Tipo**
FILENAME	Nome do arquivo	Categórico
URL	URL do site	Categórico
URLLength	Quantidade de caracteres da URL	Numérico
Domain	Domínio do site	Categórico
DomainLength	Quantidade de caracteres da Domínio	Numérico
IsDomainIP	Se o domínio é apenas um IP	Binário
TLD	Top-Level Domain, último segmento do domínio	Categórico
URLSimilarityIndex	Grau similaridade de URL comparado com URLs	Numérico

	famosas	
CharContinuationRate	O quanto os caracteres da URL seguem um padrão linguístico	Numérico
TLDLegitimateProb	Probabilidade do TLD ser legítimo	Numérico
URLCharProb	Probabilidade de naturalidade dos caracteres da URL	Numérico
TLDLength	Quantidade de caracteres do TLD	Numérico
NoOfSubDomain	Quantidade de subdomínios	Numérico
HasObfuscation	Se possui ofuscação	Binário
NoOfObfuscatedChar	Quantidade de caracteres ofuscados	Numérico
ObfuscationRatio	Taxa de ofuscação	Numérico
NoOfLettersInURL	Quantidade de letras na URL	Numérico
LetterRatioInURL	Taxa de letras na URL	Numérico
NoOfDegitsInURL	Quantidade de dígitos numéricos na URL	Numérico
DegitRatioInURL	Taxa de dígitos numéricos na URL	Numérico
NoOfEqualsInURL	Quantidade de iguais (=) na URL	Numérico
NoOfQMarkInURL	Quantidade de interrogações (?) na URL	Numérico
NoOfAmpersandInURL	Quantidade de (&) na URL	Numérico
NoOfOtherSpecialCharsInURL	Quantidade de outros caracteres especiais na URL	Numérico
SpacialCharRatioInURL	Taxa de caracteres especiais na URL	Numérico
IsHTTPS	Se o domínio utiliza de criptografia de segurança	Binário { 1 = HTTPS, 0 = HTTP }
LineOfCode	Quantidade de linhas de código da página	Numérico
LargestLineLength	Quantidade de caracteres da maior linha de código da página	Numérico
HasTitle	Se a página possui título	Binário { 1 = Possui, 0 = Não possui }
Title	Título da página	Categórico

DomainTitleMatchScore	Taxa de similaridade entre o domínio e o título da página	Numérico
URLTitleMatchScore	Taxa de similaridade entre a URL e o título da página	Numérico
HasFavicon	Se a página possui favicon	Binário { 1 = Possui, 0 = Não possui }
Robots	Se possui o arquivo robots.txt acessível no domínio	Binário { 1 = Possui, 0 = Não possui }
IsResponsive	Se a página é responsiva	Binário { 1 = Responsiva, 0 = Não responsiva }
NoOfURLRedirect	Quantidade de redirecionamentos automáticos	Numérico
NoOfSelfRedirect	Quantidade de redirecionamentos automáticos para o mesmo domínio	Numérico
HasDescription	Se a página possui descrição	Binário { 1 = Possui, 0 = Não possui }
NoOfPopup	Quantidade de janelas pop-ups	Numérico
NoOfiFrame	Quantidade de elementos iFrames	Numérico
HasExternalFormSubmit	Se possui envio de informações para formulário externo	Binário { 1 = Possui, 0 = Não possui }
HasSocialNet	Se a página possui link para rede social	Binário { 1 = Possui, 0 = Não possui }
HasSubmitButton	Se a página possui botão que envia informações	Binário { 1 = Possui, 0 = Não possui }
HasHiddenFields	Se a página possui campos escondidos	Binário { 1 = Possui, 0 = Não possui }
HasPasswordField	Se a página possui campo de senha	Binário { 1 = Possui, 0 = Não possui }
Bank	Se possui relação com sistemas bancários	Binário { 1 = Possui, 0 = Não possui }
Pay	Se possui relação com sistemas de pagamentos	Binário { 1 = Possui, 0 = Não possui }
Crypto	Se possui relação com sistemas de criptomoedas	Binário { 1 = Possui, 0 = Não possui }
HasCopyrightInfo	Se a página possui informações de copyright	Binário { 1 = Possui, 0 = Não possui }
NoOfCSS	Quantidade de scripts CSS	Numérico

NoOfJS	Quantidade de scripts JS	Numérico
NoOfSelfRef	Quantidade de links internos que apontam para o mesmo domínio da página	Numérico
NoOfEmptyRef	Quantidade de elementos html <a> com a tag href em branco	Numérico
NoOfExternalRef	Quantidade de links que apontam para páginas externas	Numérico
label	Variável alvo, se a instância é Phishing ou legítima	Categórico { 0 = Phishing, 1 = Legitimo }

3. Pré-Processamento (Data Preparation)

Técnica		
Nome	**Descrição**	**Objetivo**
Adequação nome variável alvo	Alterado campo **label** para **IsPhising**	Melhorar a compreensão do objetivo da coluna
Adequação valores binários	Alterado todos os valores binários **1** e **0** para **VERDADEIRO** E **FALSO**	Melhorar o entendimento em futuros gráficos
Normalização dos dados	Normalizado todos os valores numéricos para **[0 - 1]**	Uniformizar os dados na mesma escala
Definir cores variáveis booleanas	Alterado as cores das variáveis binárias	Melhorar a visualização em futuros gráficos

4. Modelagem (Modeling)

Técnica		
Nome	**Descrição**	**Objetivo**

5. Resultados e Avaliação (Evaluation)

Apresente de forma gráfica e textual os resultados obtidos com aplicação das técnicas descritas na seção “4. Modelagem (Modeling)”, discutindo e explicando tais resultados. [8 a 12 parágrafos]

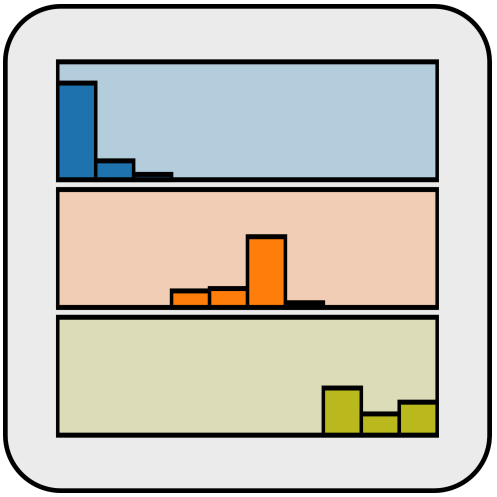

Figura 1. Logo do método de Visual Analytics VAX (multiVariate dAta eXplanation) [4], fazendo alusão aos histogramas por classe e padrão empregados no método.

Figuras devem estar centralizadas, legíveis, com numeração e legenda explicativa, e devem ser mencionadas e explicadas dentro do texto, como a Figura 1, que com propósito didático, apresenta o logo do método de Visual Analytics VAX [4]. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

a. Análise Descritiva

Faça uso de técnicas de medidas de resumo e representações visuais (gráficos), como: correlação entre variáveis, redução de dimensionalidade, gráfico de dispersão, histograma e coordenadas paralelas.

Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento

entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

b. Análise de Agrupamento

Faça uso de técnicas de agrupamento e representações visuais (gráficos), como: k-means, redução de dimensionalidade, gráfico de dispersão, histograma e coordenadas paralelas.

Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

c. Análise Preditiva

Faça uso de técnicas de predição para variável algo categórica (classificação) e/ou real (regressão).

Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

6. Implantação (Deployment)

Quais são os desdobramentos, realizações, frutos, soluções, respostas, impactos, repercussões, consequências, implicações, conclusões, resultantes, desfechos, registros, efeitos e/ou hipóteses? [2 a 4 parágrafos]

Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento

entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50. Texto em Arial 12, com recuo de início de parágrafo. Texto em Arial 12, com recuo de início de parágrafo e espaçamento entre linha de 1,50.

Referências

1. SCHRÖER, C.; KRUSE, F.; GÓMEZ, J. M. A Systematic Literature Review on Applying CRISP-DM Process Model. Procedia Computer Science, v. 181, n. 1, p. 526–534, 2021.

2. WIRTH, R; HIPP, J. CRISP-DM: Towards a Standard Process Model for Data Mining. In: Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining. 2000. p. 29-39.

3. CHAPMAN, P; CLINTON, J; KERBER, R; KHABAZA, T; REINARTZ, T; SHEARER. C; WIRTH, R. CRISP-DM 1.0. Step-by-step Data Mining Guide. 2000.

4. POPOLIN NETO, M.; PAULOVICH, F. V. Multivariate Data Explanation by Jumping Emerging Patterns Visualization. IEEE Transactions on Visualization and Computer Graphics, v. 30, n. 2, p. 1549–1563. 2024

5. DAMASCENO, Heitor et al. Monitoramento e Identificação de Páginas de Phishing. In: SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS (SBRC), 39. , 2021, Uberlândia. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 378-391. ISSN 2177-9384. DOI: https://doi.org/10.5753/sbrc.2021.16734.

6. PhiUSIIL Phishing URL (Website) [Dataset]. Disponível em: <https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset>.

7. IBM. Ataque cibernético. Disponível em: <https://www.ibm.com/br-pt/think/topics/cyber-attack>.

8. NUNES, E. O que é engenharia social: a face humana na cibersegurança. Disponível em: <https://especializacao.ccec.puc-rio.br/blog/engenharia-social>.

9. IBM. Phishing. Disponível em: <https://www.ibm.com/br-pt/think/topics/phishing>.

10. Brasil é o país com maior número de vítimas de phishing na internet. Disponível em: <https://agenciabrasil.ebc.com.br/geral/noticia/2021-03/brasil-e-o-pais-com-maior-numero-de-vitimas-de-phishing-na-internet>.

11. REDBELT SECURITY. Fraudes no setor financeiro: por hora, são criados 14 sites falsos no Brasil; conheça as iscas. Disponível em: <https://www.redbelt.com.br/blog/fraudes-no-setor-financeiro-por-hora/>. Acesso em: 31 out. 2025.