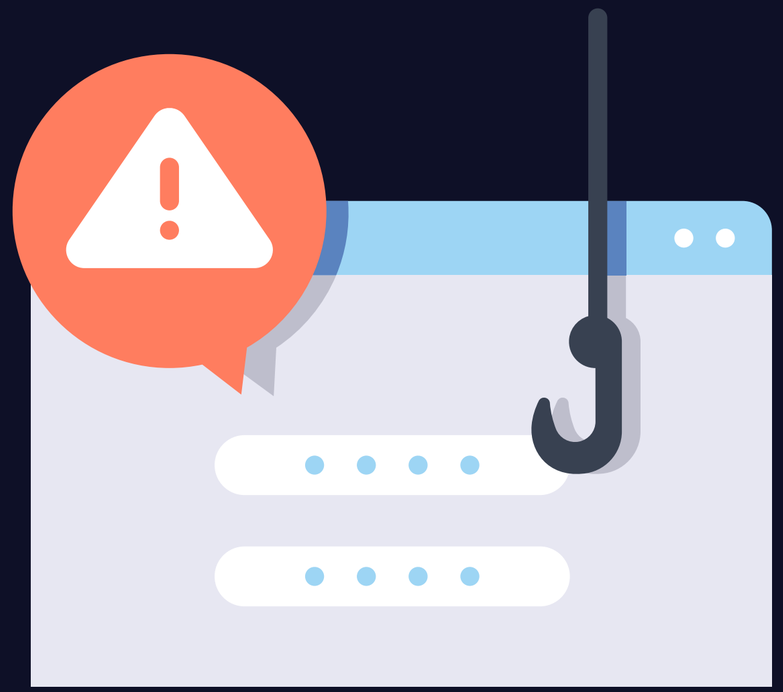
PROJETO C318

SISTEMA DE DETECÇÃO DE PHISHING

AVNER JOSÉ
BERNARDO GALDOLPHO
FÁBIO FIORITA

O QUE É PHISHING

- Ataque cibernético para obter informações confidenciais, como senhas e dados bancários, enganando os usuários por meio de mensagens falsas.

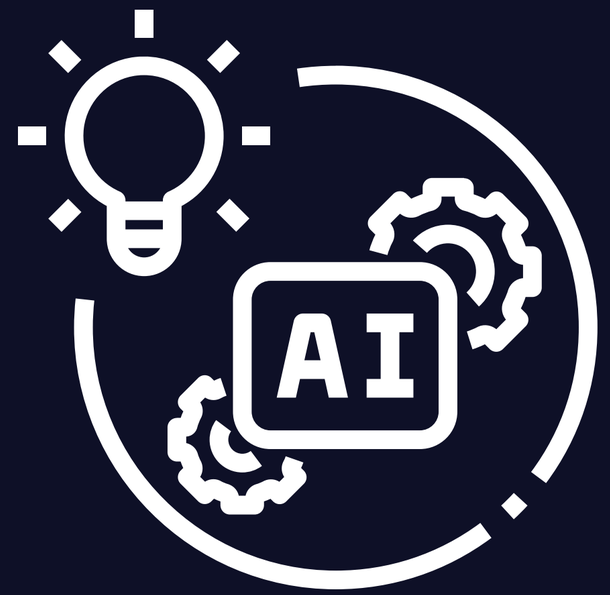


COMO FUNCIONA

- E-mails Falsos
- Sites Falsos
- Engenharia Social
- Spear Phishing
- Formas Variadas

OBJETIVO DO PROJETO

Desenvolver um sistema de detecção de phishing eficaz que ajude a proteger os usuários da internet contra fraudes cibernéticas.



OBJETIVOS DE CIÊNCIA DE DADOS

- Coletar e explorar um conjunto de dados de URLs com características relevantes para a detecção de phishing;
- Desenvolver modelos de Machine Learning capazes de identificar URLs suspeitas e distinguir entre URLs legítimas e URLs de phishing;
- Avaliar o desempenho dos modelos usando métricas apropriadas, como precisão, recall e F1-score.

ENQUADRAMENTO

- **Aprendizagem Supervisionada — Classificação**
- Conjunto de dados que inclui URLs rotuladas como phishing ou legítimas.

DATASET

- Entradas: 11,430
- Colunas: 89
- Não possui valores nulos
- A variável “Status” está balanceada

Web page Phishing Detection Dataset

Detect Phishing in Web Pages

Data Card | Code (27) | Discussion (4)

About Dataset

Context

Phishing continues to prove one of the most successful and effective ways for cybercriminals to defraud us and steal our personal and financial information.
Our growing reliance on the internet to conduct much of our day-to-day business has provided fraudsters with the perfect environment to launch targeted phishing attacks. The phishing attacks taking place today are sophisticated and increasingly more difficult to spot. A study conducted by Intel found that 97% of security experts fail at identifying phishing emails from genuine emails.

Content

The provided dataset includes 11430 URLs with 87 extracted features. The dataset is designed to be used as benchmarks for machine learning-based phishing detection systems. Features are from three different classes: 56 extracted from the structure and syntax of URLs, 24 extracted from the content of their correspondent pages, and 7 are extracted by querying external services. The dataset is balanced, it contains exactly 50% phishing and 50% legitimate URLs.

Acknowledgements

Hannousse, Abdelhakim; Yahiouche, Salima (2021), “Web page phishing detection”, Mendeley Data, V3, doi: 10.17632/c2gw7fy2j4.3

- The Source of the dataset is available here.

Usability
9.41

License
Attribution 4.0 International (CC ...

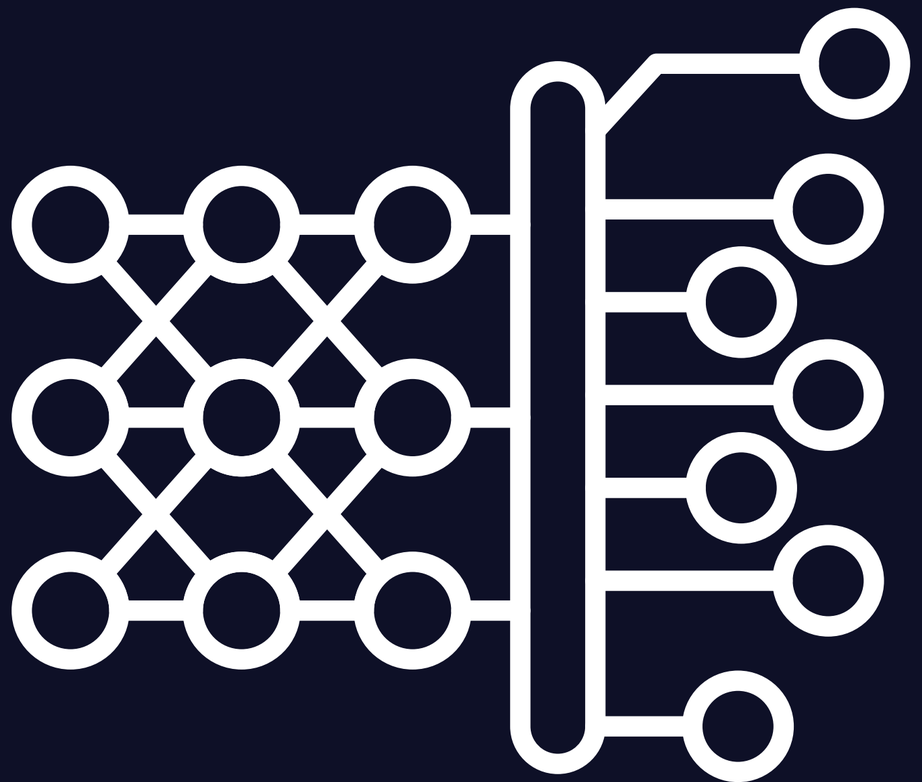
Expected update frequency
Never

Tags

Tabular | Classification | Exploratory Data Analysis | Binary Classification | Websites

ETAPA DE ANÁLISE

ETAPA DE MODELAGEM



RESULTADOS

- Melhores parâmetros:
  - RandomForestClassifier
  - StandardScaler
  - RFECV
    - cv = 5
    - estimator = RandomForestClassifier(n_estimators=350)

RESULTADOS

- Acurácia: 96,44%

Matriz de Confusão

		Valor Predito	
		Sim	Não
Valor Real	Sim	1647	68
	Não	54	1660

Relatório de Classificação

	Precision	Recall	F1-Score	Support
0	0,97	0,96	0,96	1715
1	0,96	0,97	0,96	1714