**SOLUÇÕES EM MINERAÇÕES DE DADOS PARA**

**BASE DE DADOS DO UCI MACHINE LEARNING: PHISHING WEBSITES DATA**

Relatório final com objetivo de obtenção de notas da Disciplina Soluções em Mineração de Dados do Curso de Ciências de Dados Analytics – Poli - UPE

Prof Dr Leandro Maciel Almeida

ALUNOS DO GRUPO 3:

* DIOGO DIAS
* ANA BEATRIZ DE ALMEIDA
* ISABELLA PAZ
* LAÍS VIDOTO
* HELMER PAIVA
* MARIA REGINA MENEZES

**RECIFE – PE**

**13/07/2019**

**SUMÁRIO**

[**1. I**](#_3fxrbqfh7jey)**ntrodução**

[**2.**](#_jnkbz08y3p7g) **Etapas do KDD – Scripts em Python**

**2.1 Seleção**

**2.2. Pré-processamento e Limpeza**

**2.3. Transformação dos dados**

**2.4. Data Mining**

**2.4.1. Arvore de Decisão Simples**

**2.4.2. Random Forest**

**2.4.3. Rede Neural MLP**

**2.4.4. Comitê de Redes Neurais em Python**

**2.5. Interpretação e Avaliação**

**2.5.1. Gráficos de caixa (Boxplot)**

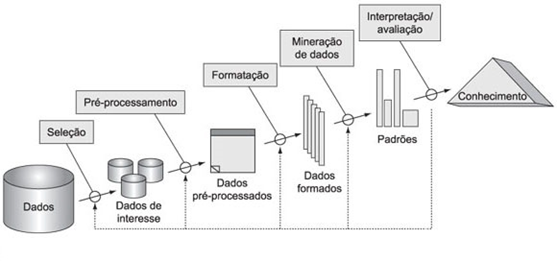
**2.5.2. Analise dos Resultados**

**3. Aspectos Relevantes**

[**Referências**](#_56h4futan5g4)

**1** [**I**](#_3fxrbqfh7jey)**IIGBIIIIntrodução**

Estudo abaixo é baseado pelo KDD.



A base de dados apresenta os recursos mais importantes e que se mostraram sólidos e eficazes na previsão de sites de phishing. Além disso, foi proposto alguns novos recursos, atribuímos experimentalmente novas regras a alguns recursos conhecidos e atualizamos alguns outros recursos.

Quando um site é considerado **SUSPEITO**, isso significa que ele pode ser ou falso ou legítimo, o que significa que o site possui alguns recursos legítimos e fraudulentos.

* Analise realizada na base de dados, escolhida abaixo:

https://archive.ics.uci.edu/ml/datasets/phishing+websites#

* arquivo de utilização-> traning dataset.arf
* Resultados encontrados para cada atributo:
  + 1 – Sim
  + 0 – Talvez/Desconhecido/Não informado
  + -1 – Não
* Numero total de atributos: 31
* y é o campo: Result

**Atributos de sua base:**

|  |  |
| --- | --- |
| **Nome do atributo** | **Descrição** |
| having\_IP\_Address | Ter endereço de IP |
| URL\_Length | Comprimento do URL |
| Shortining\_Service | Serviço de Shortining |
| having\_At\_Symbol | URL com o símbolo@ |
| double\_slash\_redirecting | Redirecionamento de barra dupla |
| Prefix\_Suffix | Prefixo sufixo |
| having\_Sub\_Domain | Ter subdomínio |
| SSLfinal\_State | Estado SSLfinal |
| Domain\_registeration\_length | Comprimento de registro de domínio – expiração do domínio |
| Favicon | a imagem gráfica (ícone) associada a uma página da Web específica |
| port | Portas abertas |
| HTTPS\_token | Token HTTPS na parte do domínio do URL |
| Request\_URL | URL de solicitação, a página contém objetos externos do tipo imagens. |
| URL\_of\_Anchor | URL da âncora - as tags <a> e o site possuem nomes de domínio diferentes. |
| Links\_in\_tags | Links em tags - as tags <Meta>, <Script> e <Link> estão vinculadas a mesma página da Web. |
| SFH | examina se o Server Form Handler (SFH) é uma string vazia ou em branco. |
| Submitting\_to\_email | Envio para email - as funções ‘mail” ou “mailto” são usadas no código-fonte da página da Web para enviar as informações do usuário para o e-mail pessoal do phisher. |
| Abnormal\_URL | o nome do host foi incluído no URL do site examinado. |
| Redirect | quantas vezes um site foi redirecionado |
| on\_mouseover | se foi mostrado para o usuário um URL falso na barra de status. |
| RightClick | se a função de clicar com o botão direito foi desativada para impedir que os usuários visualizem e salvem o código-fonte da página da Web. |
| popUpWidnow | se o site solicita que o usuário envie suas informações pessoais através de uma janela pop-up. |
| Iframe | se a tag HTML IFram é usada para exibir uma página da Web adicional. |
| age\_of\_domain | verifica se os sites vivem por um curto período de tempo. |
| DNSRecord | se nenhum registro DNS (mapeamento entre o endereço IP e o domínio associado) existe para o domínio. |
| web\_traffic | popularidade do site. |
| Page\_Rank | o page rank. |
| Google\_Index | se o wesite é indexado pelo Google. |
| Links\_pointing\_to\_page | número de links que apontam para a página da Web. |
| Statistical\_report | se o site é relatado como um site de phishing por várias partes que rastreie phishing como phishTank |
| Result | resu |

**2 Etapas do KDD – Scripts em Python**

**2.1 Seleção**

**Definição:** A etapa de seleção é a primeira etapa do KDD, é uma etapa muito importante, pois é nela que serão decididos quais os conjuntos de dados que serão relevantes para que sejam obtidos resultados com informações uteis.

possue valores 0 para seus atributos->

**URL\_Length, having\_Sub\_Domain, SSLfinal\_State, URL\_of\_Anchor, Links\_in\_tags, SFH, Redirect, web\_traffic, Links\_pointing\_to\_page**

Utilizamos todos os atributos listados na introdução

**2.2 Pré-processamento e Limpeza**

**Definição**: Acontece a limpeza dos dados e seleção de atributos. Nesta etapa informações ausentes, errôneas ou inconsistentes nas bases de dados devem ser corrigidas de forma a não comprometer a qualidade dos modelos de conhecimento a serem extraídos ao final do processo de KDD.

Convertemos os dados vindos no arquivo em inteiro, pois o mesmo está vindo como string.

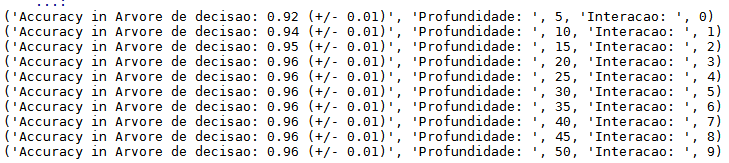
**2.3 Transformação dos dados**

**Definição**: Analisa os dados obtidos da etapa anterior e os reorganiza de uma forma específica para que possam ser interpretados na etapa seguinte.

**2.4. Data Mining**

**Definição**: É onde tudo acontece, os dados depois de transformados serão lidos e interpretados. A mineração faz com que meros dados sejam transformados em informações, tais informações são indicadas através de força bruta, ou seja, lendo regra por regra e as interpretando.

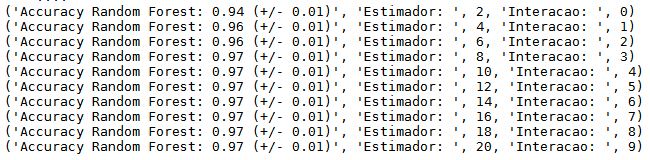
**2.4.1 Arvore de Decisão Simples**



Para esse tipo de classificador, a melhor representação da imagem acima é:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classificador Utilizado** | **Acurácia** | **Profundidade** | **Iteração** |
| tree.DecisionTreeClassifier | 0.96 | 20 | 3 |

**2.4.2 Random Forest**



Para esse tipo de classificador, a melhor representação da imagem acima é:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classificador Utilizado** | **Acurácia** | **Estimador** | **Iteração** |
| RandomForestClassifier | 0.967 | 8 | 3 |

**2.4.3 Rede Neural MLP**

Para esse tipo de classificador, a melhor representação da imagem acima é:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classificador Utilizado** | **Acurácia** | **Profundidade** | **Iteração** |
|  |  |  |  |

**2.4.4 Comitê de Redes Neurais em Python**

Para esse tipo de classificador, a melhor representação da imagem acima é:

|  |  |  |  |
| --- | --- | --- | --- |
| **Classificador Utilizado** | **Acurácia** | **Profundidade** | **Iteração** |
|  |  |  |  |

**2.5. Interpretação e Avaliação**

**Definição**: É onde as regras indicadas pelo processo anterior serão interpretadas e avaliadas. Após a interpretação poderão surgir padrões, relacionamentos e descoberta de novos fatos, que podem ser utilizados para pesquisas, otimização e outros.

**2.5.1 Gráficos de caixa (Boxplot)**

**2.5.2 Analise dos Resultados**

**3** **Aspectos Relevantes**

[**RRReferências**](#_56h4futan5g4)

* **xx**
* **xx**