CIC-Evasive-PDFMal2022

Ciência de Dados para a Segurança - CI1030

Vinicius Gabriel Machado (GRR20182552)

Visão Geral do Dataset

- Identificar ações maliciosas que fazem o uso de pdfs como meio de transporte e ação, mas que possuem características evasivas, possuem semelhanças com amostras não maliciosas e vice-versa.

- Proposto por: Maryam Issakhani, Princy Victor, Ali Tekeoglu, and Arash Habibi Lashkari

Visão Geral do Dataset

- Junção de pdfs de diferentes fontes
  - 11173 amostras maliciosas do Contagio
  - 20000 amostras maliciosas do VirusTotal
  - 9109 amostras benignas do Contagio
  - remoção de amostras duplicadas
- Criação e execução de um programa para extração de atributos
  - Extração de características gerais e estruturais de um pdf, principalmente das mais propensas a utilização para fins maliciosos
- Execução do kmeans para clusterização
  - Amostras que cairam no cluster errado entram para o dataset

Visão Geral do Dataset - Atributos

- 33 atributos originalmente, incluindo nome do arquivo e a classe em que foi categorizado, malicioso ou benigno.

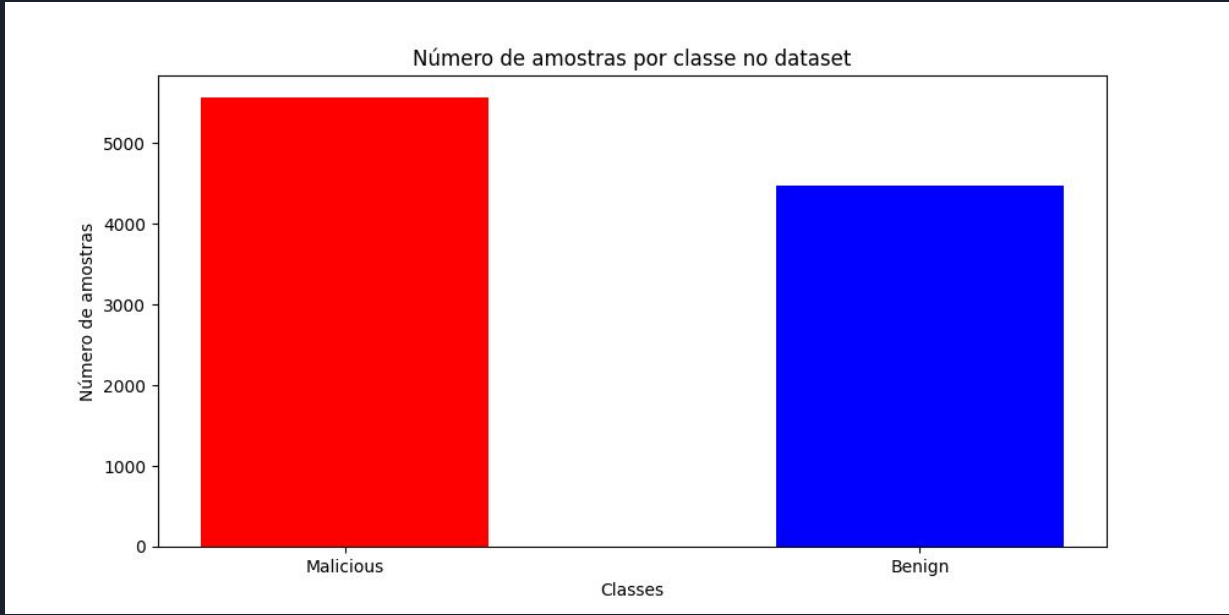
General features	Structural features
• PDF size • title characters • encryption • metadata size • page number • header • image number • text • object number • font objects • number of embedded files • average size of all the embedded media	• No. of keywords "streams" • No. of keywords "endstreams" • Average stream size • No. of Xref entries • No. of name obfuscations • Total number of filters used • No. of objects with nested filters • No. of stream objects (ObjStm) • No. of keywords "/JS", No. of keywords "/JavaScript" • No. of keywords "/URI", No. of keywords "/Action" • No. of keywords "/AA", No. of keywords "/OpenAction" • No. of keywords "/launch", No. of keywords "/submitForm" • No. of keywords "/Acroform", No. of keywords "/XFA" • No. of keywords "/JBig2Decode", No. of keywords "/Colors" • No. of keywords "/Richmedia", No. of keywords "/Trailer" • No. of keywords "/Xref", No. of keywords "/Startxref"

Visão Geral do Dataset - Classes

- Dataset binário:
  - Malicious: Amostras maliciosas com características semelhantes a de benignas;
  - Benign: Amostras benignas com características semelhantes a de maliciosas.

Visão Geral do Dataset - Classes

- Originalmente: 10025 amostras, 5557 maliciosas e 4468 benignas.



Processamento do Dataset - Limpeza

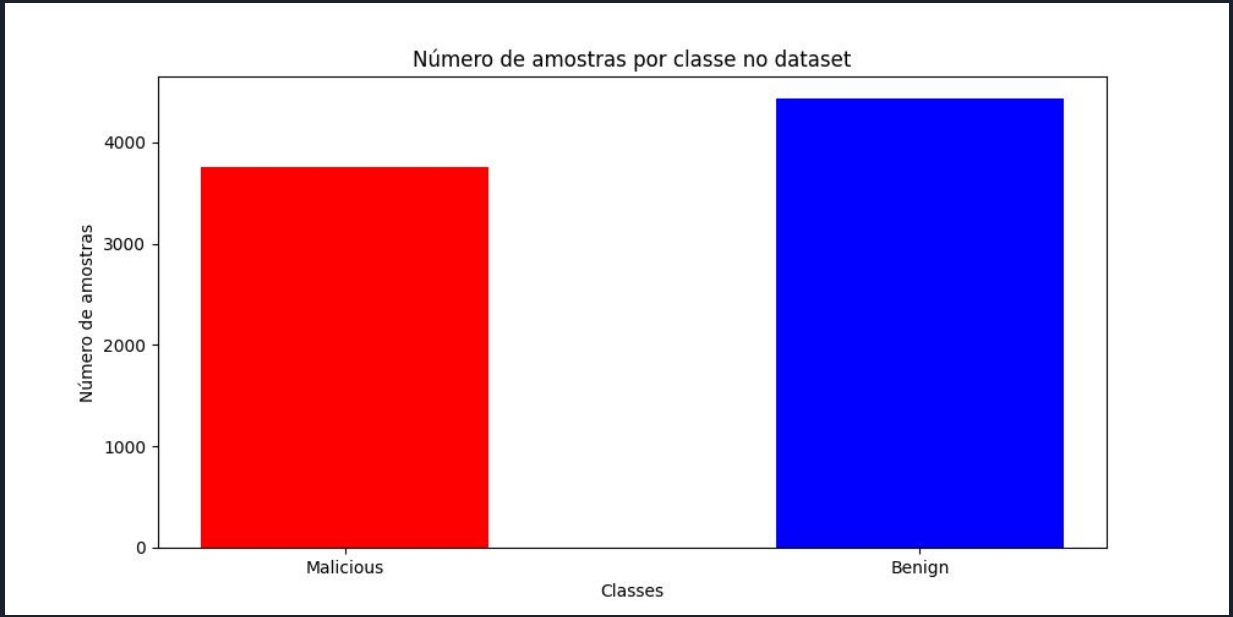
- Remoção de dados “errados” (linhas):
  - nan
  - x(1), x(2), etc...
  - echos de erro de linha de comando do bash em certos campos
- Remoção de dados “estranhos” (colunas):
  - Coluna text removida
    - Ideia: indicar se há texto no conteúdo do pdf
    - Dados: -1, 0, unclear, yes e no -> incerteza do significado do atributo
  - Coluna header removida
    - Ideia: indicar a “versão” do pdf utilizada no arquivo
    - Dados: Falta de padronização no campo e dados sem sentido

Processamento do Dataset - Limpeza

- Remoção de dados inconclusivos (linhas):
  - Exemplo: campo de contagem de uma tag estrutural de um pdf
  - Valores: -1, 0, [1, infinito[
  - Falta de clareza no significado do -1, suposição: informação inconclusiva
  - Remoção de todas as linhas contendo -1 -> grande redução no número de amostras maliciosas

Processamento do Dataset - Limpeza

- Depois da limpeza: 8190 amostras, 3759 maliciosas e 4431 benignas.



Processamento do Dataset - Extração

- Utilização do arquivo de dados final como entrada para o WEKA
  - 29 atributos de entrada (após a remoção dos desnecessários)
  - 7 atributos selecionados como características significativas para a saída

Processamento do Dataset - Extração

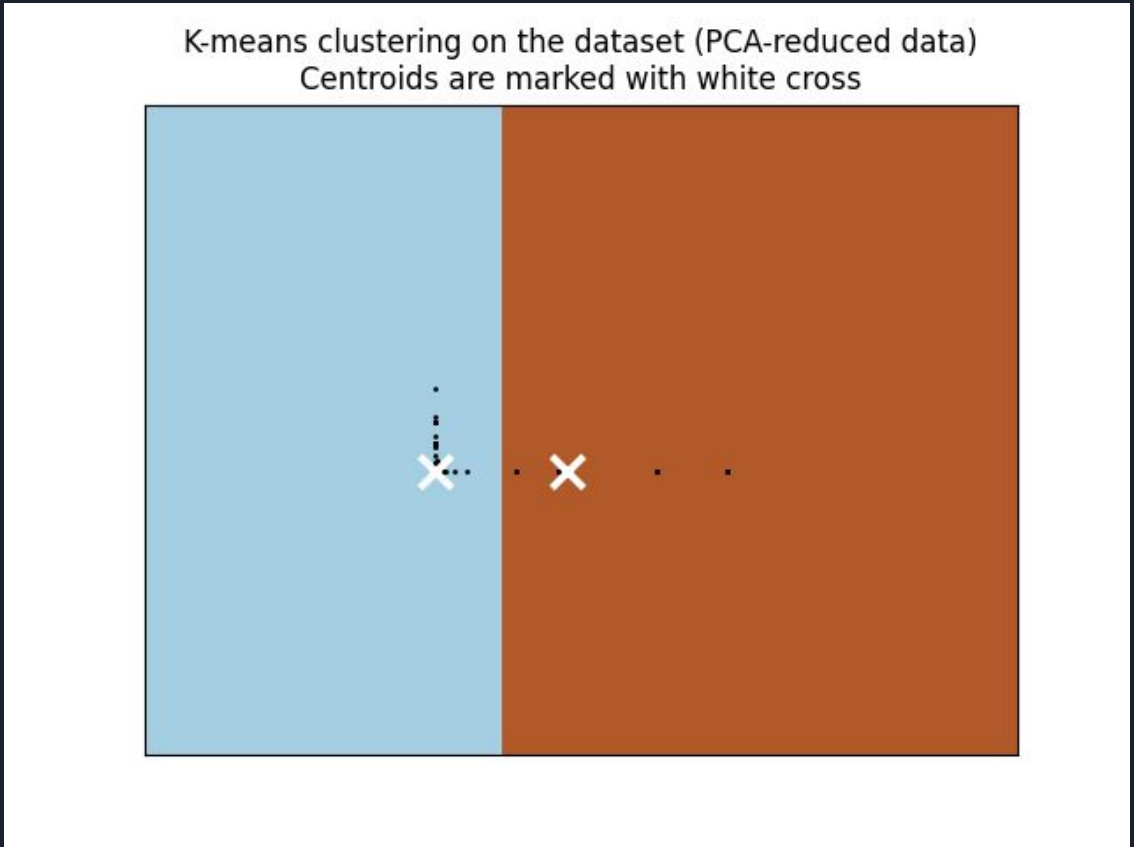
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 30 Class):
        Information Gain Ranking Filter

Ranked attributes:
 0.65792    2 metadata size
 0.56649    4 xref Length
 0.52958    9 obj
 0.52682   19 JS
 0.52333   20 Javascript
 0.51406   10 endobj
 0.50434    1 pdfsize
 0.44454   15 startxref
 0.38983   14 trailer
 0.3793    12 endstream
 0.37826   11 stream
 0.31401   13 xref
 0.27414   22 OpenAction
 0.16103    8 images
```

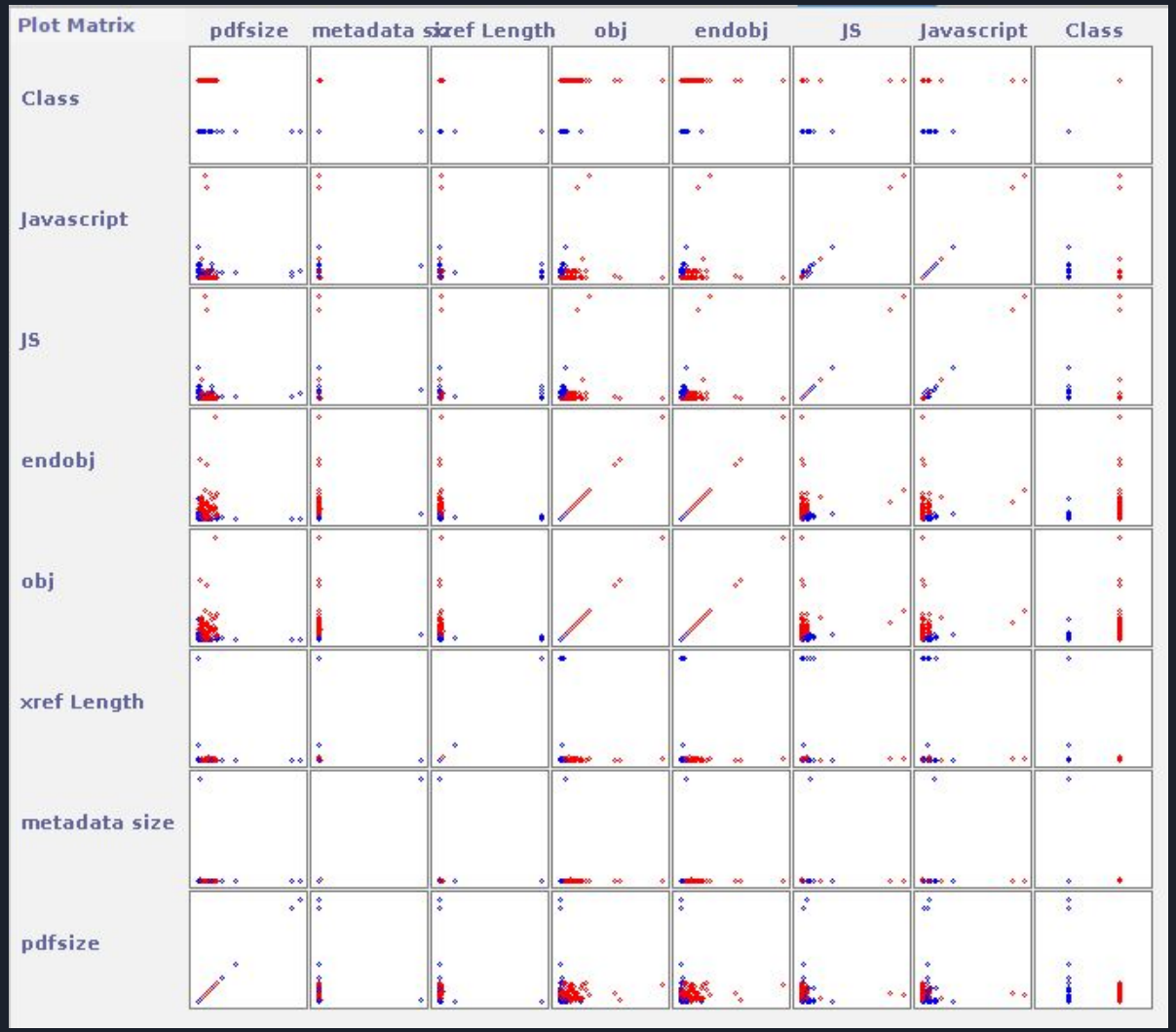
```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 30 Class):
        OneR feature evaluator.

        Using 10 fold cross validation for evaluating attributes.
        Minimum bucket size for OneR: 6

Ranked attributes:
90.72     2 metadata size
89.194   19 JS
89.109   20 Javascript
87.717    4 xref Length
87.179    9 obj
86.813    1 pdfsize
86.667   10 endobj
```

Processamento do Dataset - Extração

- Características atuais - 7 (pode ser alterado de acordo com a necessidade):
  - pdfsize: Tamanho do arquivo
  - metadata size: Tamanho da região de metadata
  - xref Length: Número de Xrefs
  - obj: Número de palavras-chave indicando o início de objetos
  - endobj: Número de palavras-chave indicando o fim de objetos
  - JS: Número de palavras-chave “/JS”
  - Javascript: Número de palavras-chave “/JavaScript”

Exploração do Dataset




Plot Matrix
pdfsize
metadata size
xref Length
obj
endobj
JS
Javascript
Class
Class
Javascript
JS
endobj
obj
xref Length
metadata size
pdfsize