

CIC-Evasive-PDFMaI2022

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)
- Visa corrigir problemas encontrados em datasets muito utilizados (ex. Contagio) para esta área de estudo
 - Duplicatas (dados com mesmas características)
 - Baixa variação (dados semelhantes)
 - Baixa distribuição (características facilmente linearmente separáveis)
- 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 de k-means para clusterização
 - Amostras que caíram no cluster errado entraram para o dataset
- Execução de um modelo de classificação proposto, baseado em empilhamento de classificadores, para a realização da classificação (acurácia, precisão, recall e f1-score maiores que 98% no dataset)



Visão Geral do Dataset - Atributos

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



General 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

Structural features

- 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"

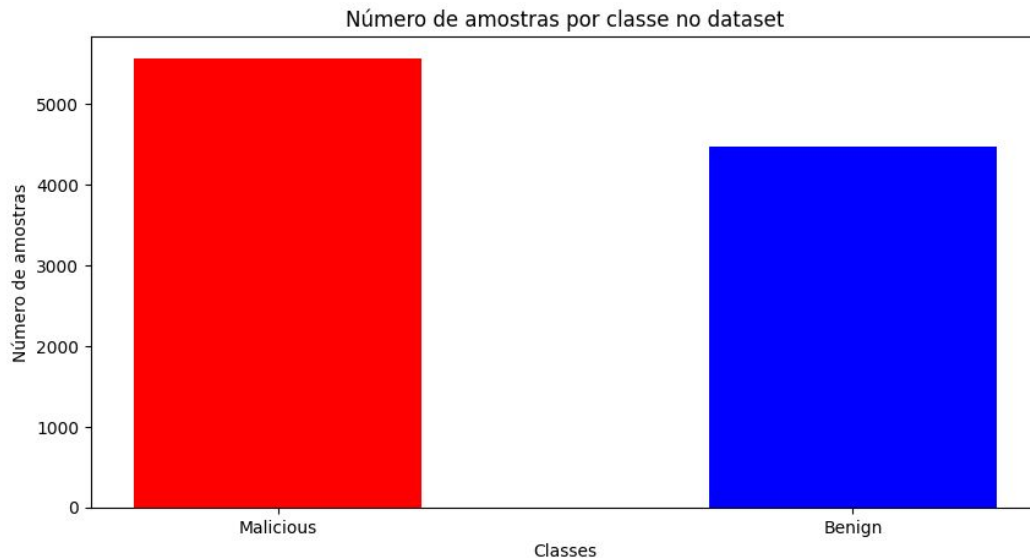
Visão Geral do Dataset - Classes

- 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 linhas:
 - dados nan
 - dados $x(1)$, $x(2)$, etc...
 - echos de erro de linha de comando do bash em certos campos
 - dados inconclusivos:
 - 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



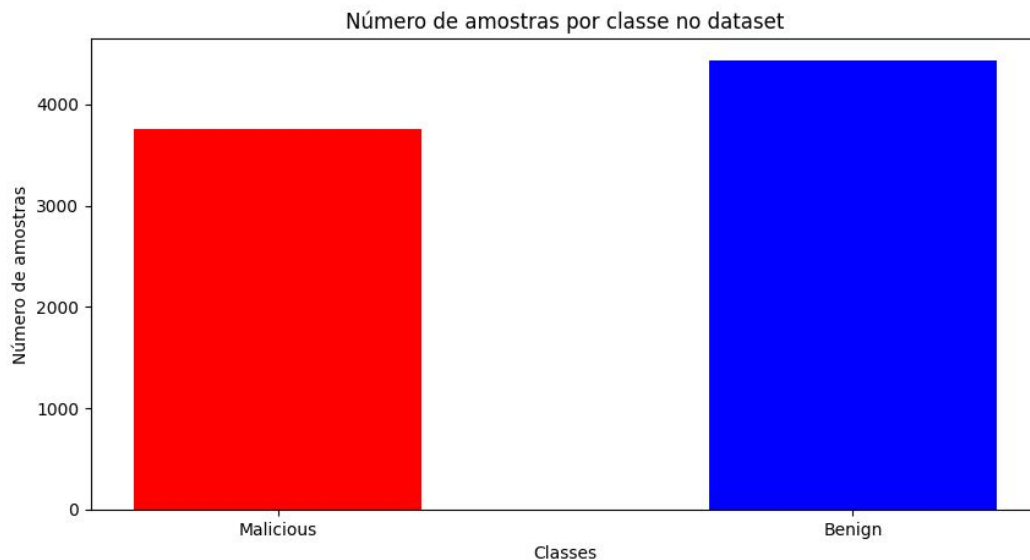
Processamento do Dataset - Limpeza

- Remoção de 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

- 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 como nome de arquivo e das colunas problemáticas)
 - 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 ===
```

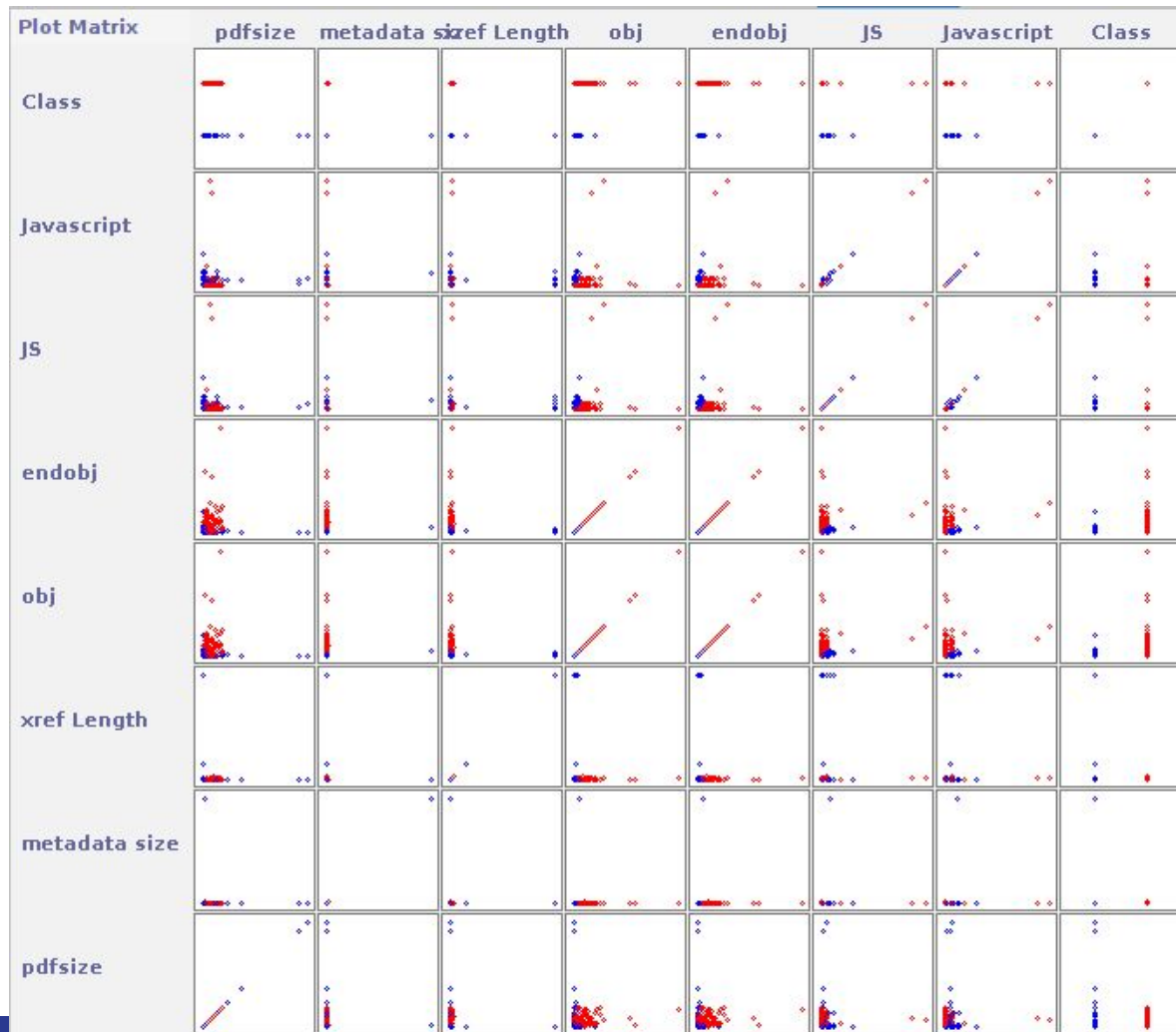
```
Search Method:  
  Attribute ranking.
```

```
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



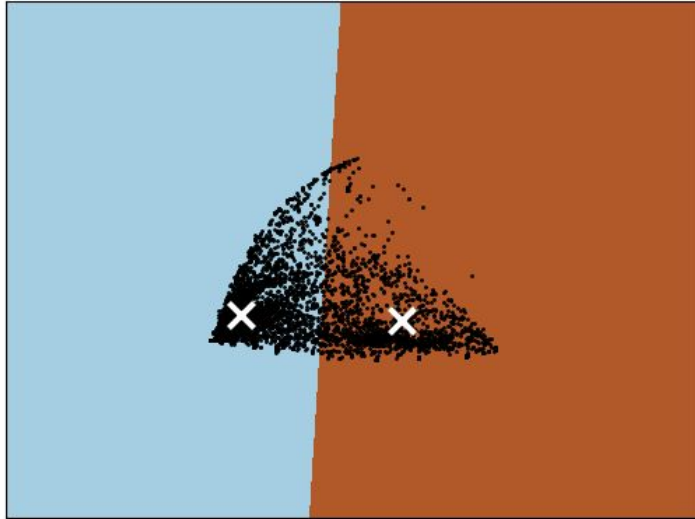
Processamento do Dataset - Extração

- Características selecionadas - 7:
 - 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”

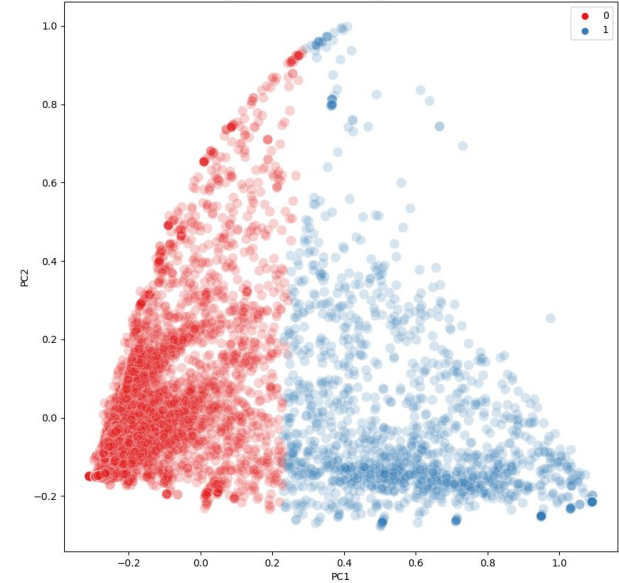


Exploração do Dataset - Clusterização -> Classificado

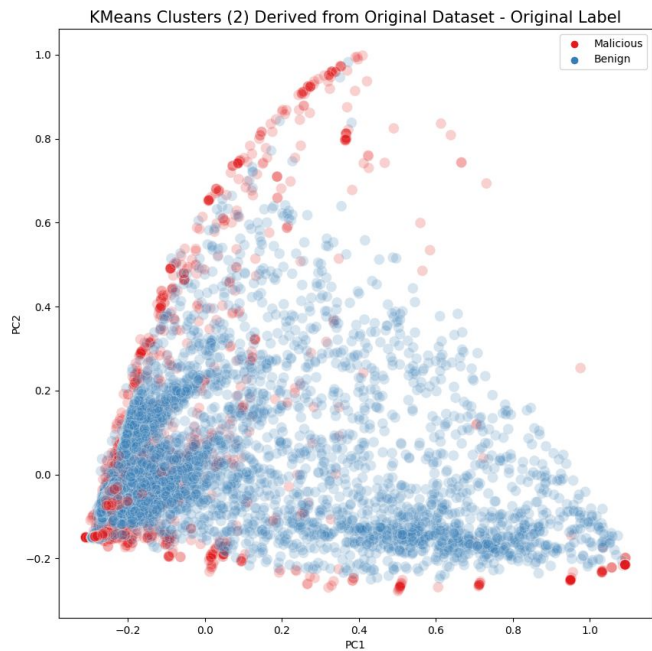
K-means clustering on the dataset (PCA-reduced data)



KMeans Clusters (2) Derived from Original Dataset - Found Label

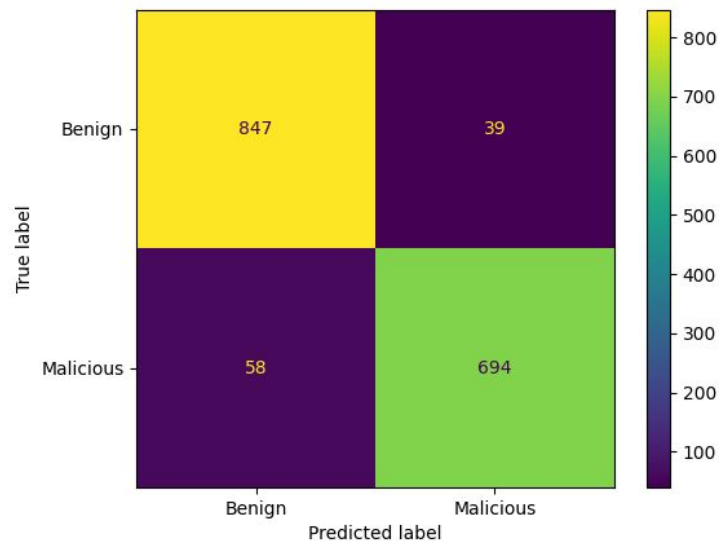


Exploração do Dataset - Clusterização ->Esperado

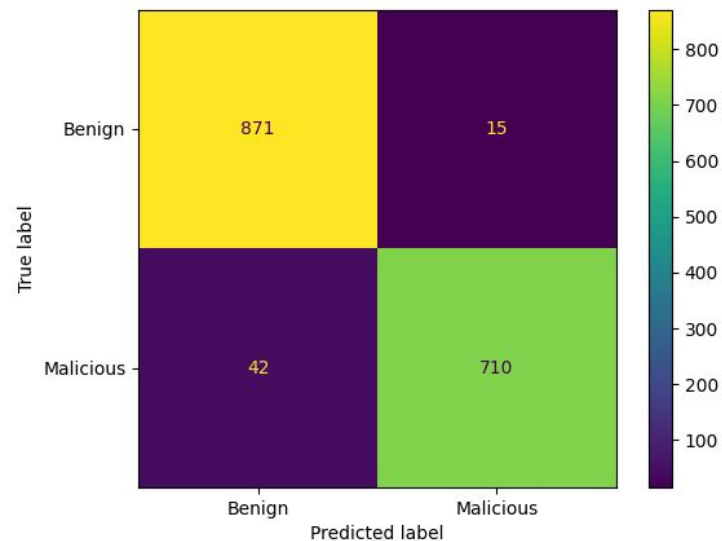


Exploração do Dataset - Classificação de teste inicial

KNN



RF



Ambiente de Testes

- Sistema Operacional: Ubuntu 20.04
- Processador: Ryzen 5 3600 4.2 GHz - 32 MB cache L3
- RAM: 16 GB 3000 MHz
- Disco: HDD 7200 rpm - SSD 550 MB/s (escrita e leitura)



Grid Search

```
=====> KNeighborsClassifier
----- Parameters: {'n_neighbors': [1, 3, 5], 'algorithm': ['brute']}
----- Best parameters: {'algorithm': 'brute', 'n_neighbors': 5}
----- Best score: 0.9775688235000086
----- Confusion Matrix: (pos -> Malicious, neg -> Benign)
      pred_neg pred_pos
neg      3521      42
pos       86     2903

=====> RandomForestClassifier
----- Parameters: {'n_estimators': [50, 100], 'random_state': [42]}
----- Best parameters: {'n_estimators': 100, 'random_state': 42}
----- Best score: 0.9919145805308636
----- Confusion Matrix: (pos -> Malicious, neg -> Benign)
      pred_neg pred_pos
neg      3563      0
pos       0     2989

=====> SVC
----- Parameters: {'max_iter': [1000, 5000], 'random_state': [42], 'tol': [0.01]}
----- Best parameters: {'max_iter': 1000, 'random_state': 42, 'tol': 0.01}
----- Best score: 0.9740528649874088
----- Confusion Matrix: (pos -> Malicious, neg -> Benign)
      pred_neg pred_pos
neg      3510      53
pos       467     2522

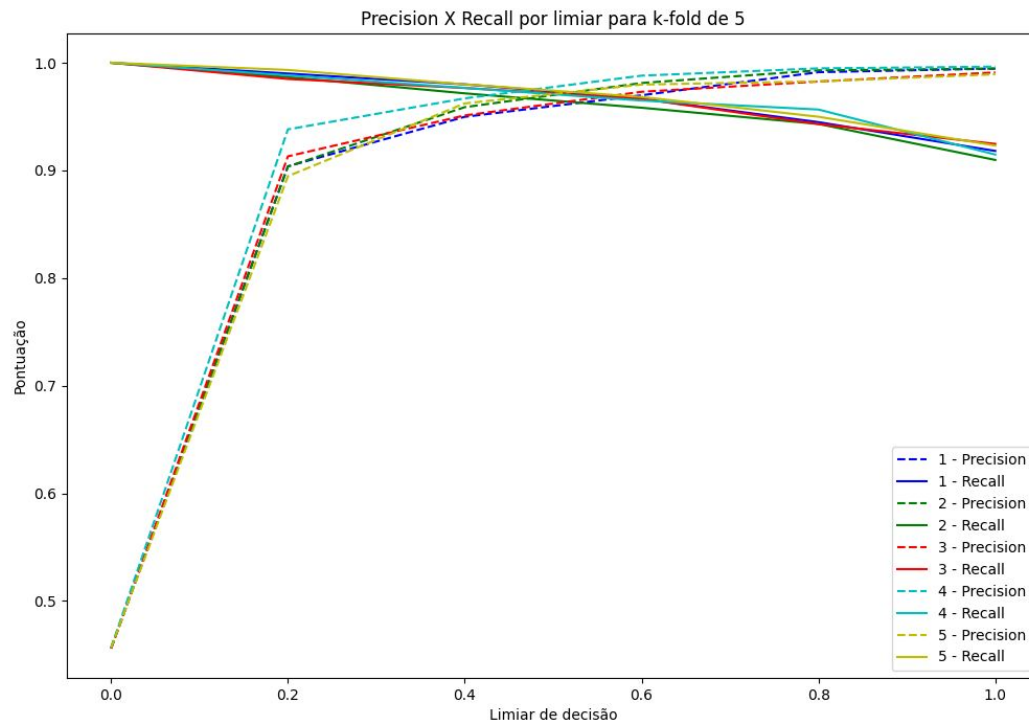
=====> MLPClassifier
----- Parameters: {'max_iter': [1000, 5000], 'random_state': [42], 'learning_rate': ['adaptive'], 'tol': [0.01]}
----- Best parameters: {'learning_rate': 'adaptive', 'max_iter': 1000, 'random_state': 42, 'tol': 0.01}
----- Best score: 0.9755086939667716
----- Confusion Matrix: (pos -> Malicious, neg -> Benign)
      pred_neg pred_pos
neg      3503      60
pos       322     2667
```

Tempo de Treinamento

```
=====> KNeighborsClassifier_50-50
Training time: 0.00023102760314941406s
=====> RandomForestClassifier_50-50
Training time: 0.18396258354187012s
=====> SVC_50-50
Training time: 0.07034015655517578s
=====> MLPClassifier_50-50
Training time: 0.25839686393737793s
=====> KNeighborsClassifier_80-20
Training time: 0.0003724098205566406s
=====> RandomForestClassifier_80-20
Training time: 0.27854013442993164s
=====> SVC_80-20
Training time: 0.1661815643310547s
=====> MLPClassifier_80-20
Training time: 0.3571891784667969s
```

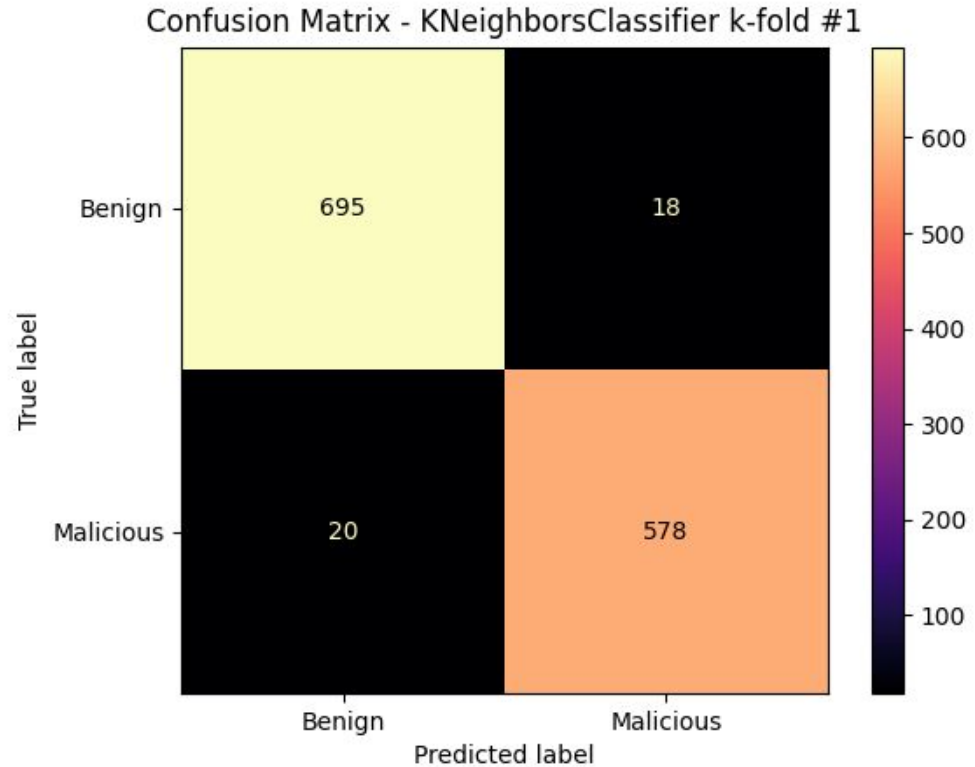
Curvas Precision/Recall

KNeighborsClassifier



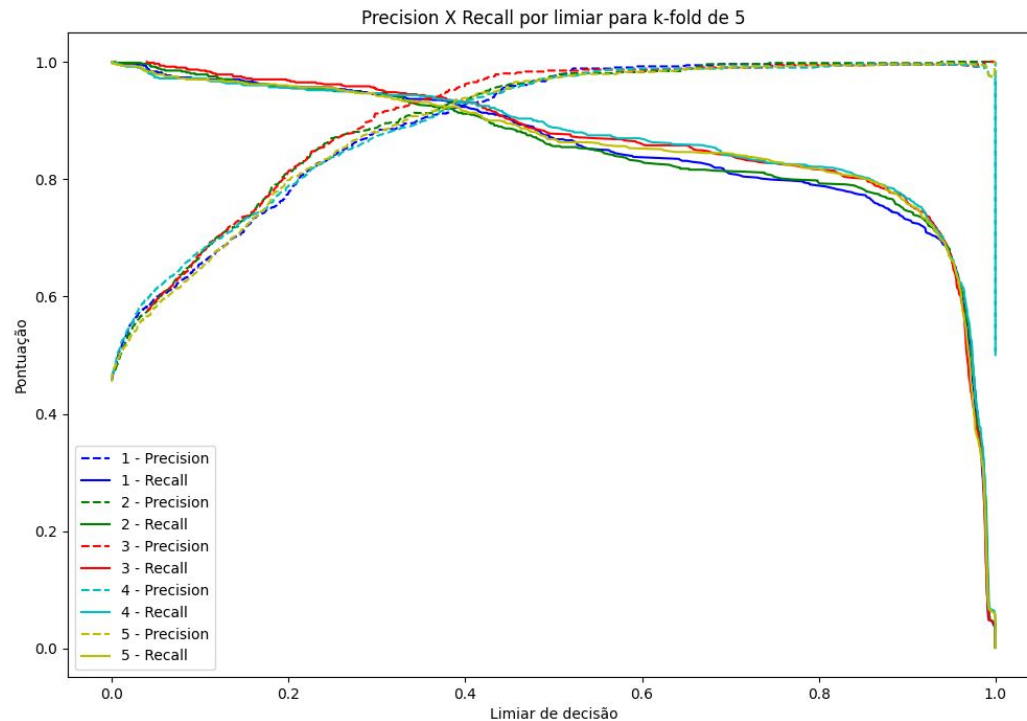
Curvas Precision/Recall

KNeighborsClassifier



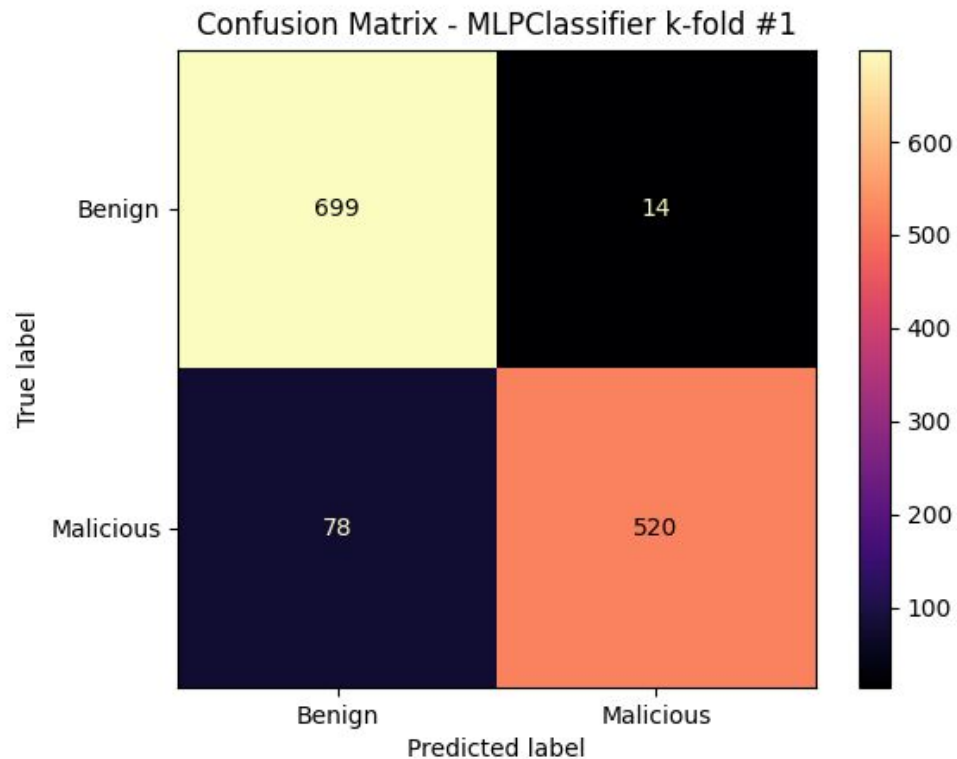
Curvas Precision/Recall

MLPClassifier



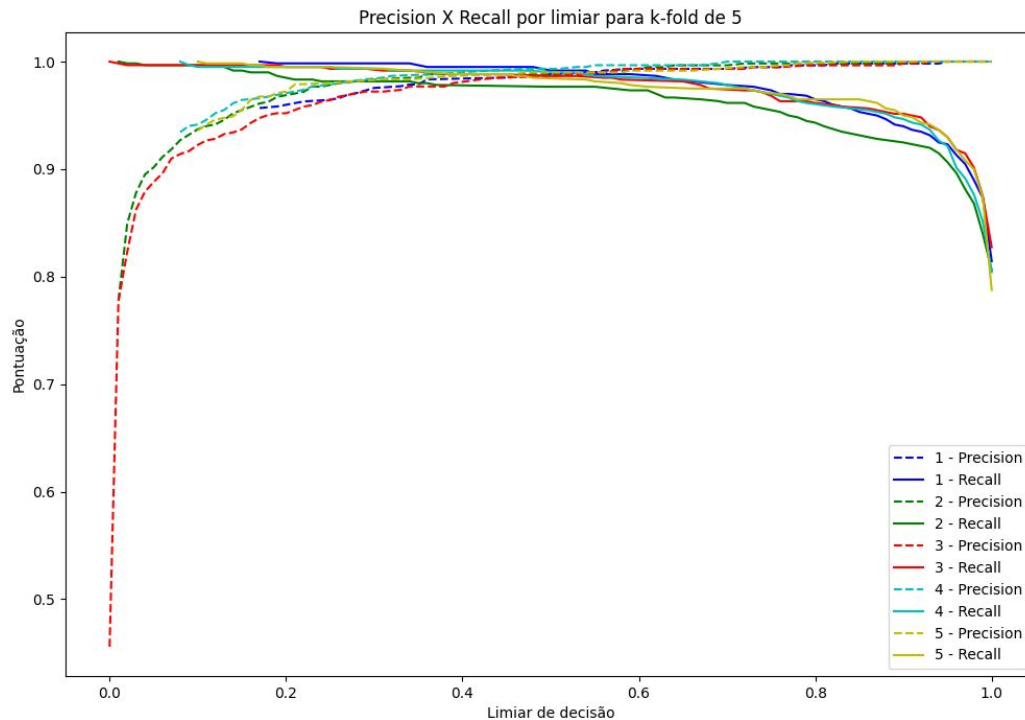
Curvas Precision/Recall

MLPClassifier



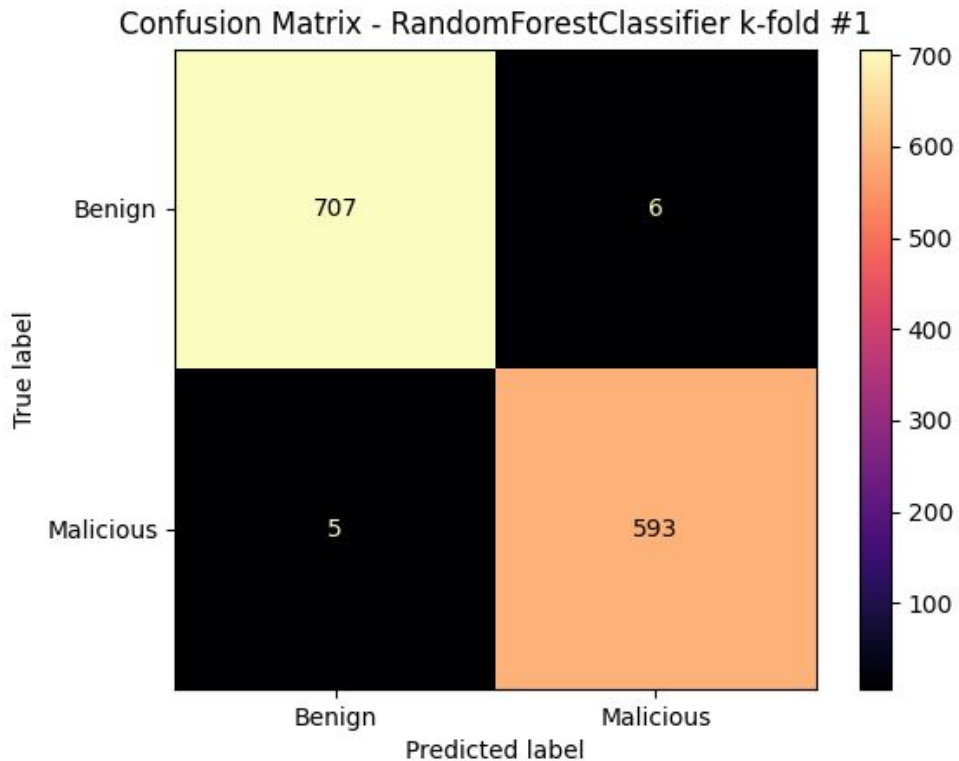
Curvas Precision/Recall

RandomForestClassifier



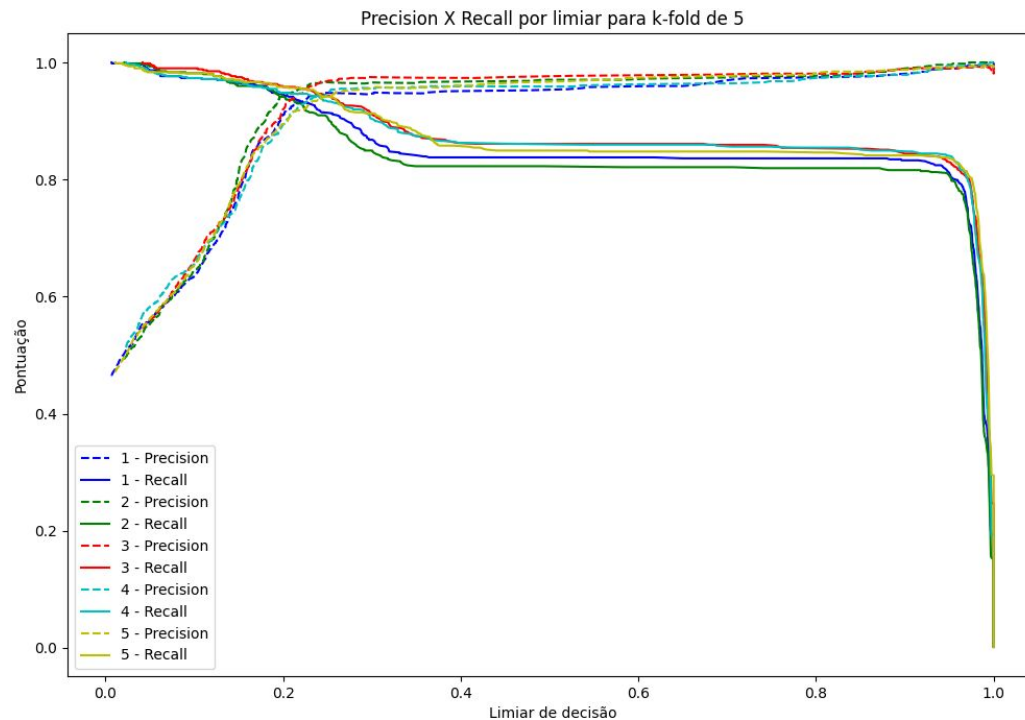
Curvas Precision/Recall

RandomForestClassifier



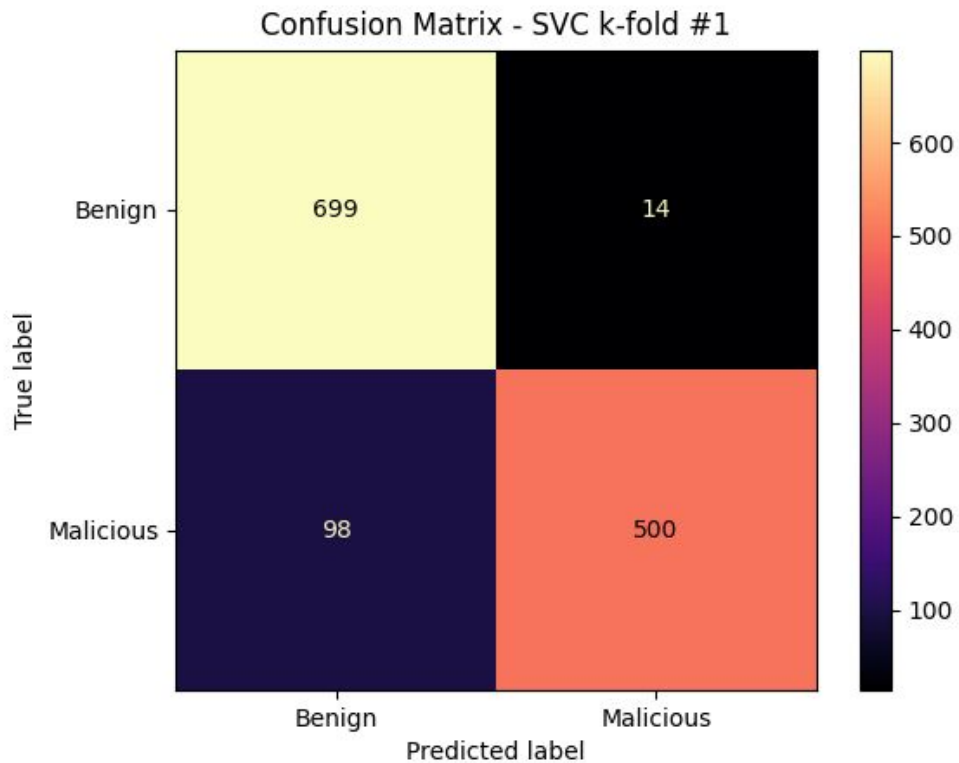
Curvas Precision/Recall

SVC



Curvas Precision/Recall

SVC



Resultados dos Autores do Dataset

Table 5: Proposed model results on our dataset.

Accuracy	Precision	Recall	F1 score
98.69	98.88	98.87	98.77

Referências

- Artigo sobre o dataset:
<https://www.scitepress.org/PublicationsDetail.aspx?ID=VibglHYeOxw=&t=1>
- Visão geral do dataset e link para fazer o download:
<https://www.unb.ca/cic/datasets/pdfmal-2022.html>
- WEKA: <https://www.cs.waikato.ac.nz/ml/weka/>
- Bibliotecas Python: <https://numpy.org>, <https://matplotlib.org>,
<https://scikit-learn.org/stable/> e <https://pandas.pydata.org>

