Evasive PDF Sample

Gonçalo Costa, up202103336

João Correia, up202005015

Ricardo Vieira, up202005091

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Artificial Intelligence 23/24

The chosen dataset is a collection of evasive PDF samples, labeled as malicious (1) or benign (0). Since the dataset has an evasive nature, it can be used to test the robustness of trained PDF malware classifiers against evasion attacks. The dataset contains 500,000 generated evasive samples, including 450,000 malicious and 50,000 benign PDFs.

This resource aims to support researchers and cybersecurity professionals in developing more advanced and robust detection mechanisms for PDF-based malware.

It's now our job to use this dataset to do exactly that: create a detection mechanism for PDF-based malware.

PROBLEM DEFINITION

RELATED WORK AND REFERENCES

- Trad, F.; Hussein, A.; Chehab, A. Leveraging Adversarial Samples for Enhanced Classification of Malicious and Evasive PDF Files. Appl. Sci. 2023, 13, 3472. https://doi.org/10.3390/app13063472
- https://www.kaggle.com/datasets/fouadtrad2/evasive-pdf-samples
- Maryam Issakhani, Princy Victor, Ali Tekeoglu, and Arash Habibi Lashkari1, “PDF Malware Detection Based on Stacking Learning”, The International Conference on Information Systems Security and Privacy, February 2022
- https://www.unb.ca/cic/datasets/pdfmal-2022.html

Programming Language

Development Environment

Data analysis

Algorithms to be used

Neural networks as they offer the capability to capture complex patterns, SVMs because they provide robustness in high-dimensional feature spaces and finally decision trees as they offer interpretability, albeit at the cost of potential limitations in handling complexity.

Dataframes

	pdfsize	pages	title characters	images	obj	endobj	stream	endstream	xref	trailer	...	ObjStm	JS	OBS_JS	Javascript	OBS_Javascript	OpenAction	OBS_OpenAction	Acroform	OBS_Acroform	class
0	644.326	70	0	1	348	351	128	128	1	1	...	0	1	0	1	0	1	0	1	0	1
1	648.050	68	0	1	348	345	124	124	1	1	...	0	1	0	1	0	0	0	1	0	1
2	696.506	68	0	1	353	353	128	125	1	1	...	0	1	0	1	0	0	0	1	0	1
3	715.926	68	0	0	759	667	250	192	1	1	...	0	1	0	1	0	1	0	1	0	1
4	707.102	70	10	2	388	373	141	138	1	1	...	0	1	0	1	0	1	0	1	0	1

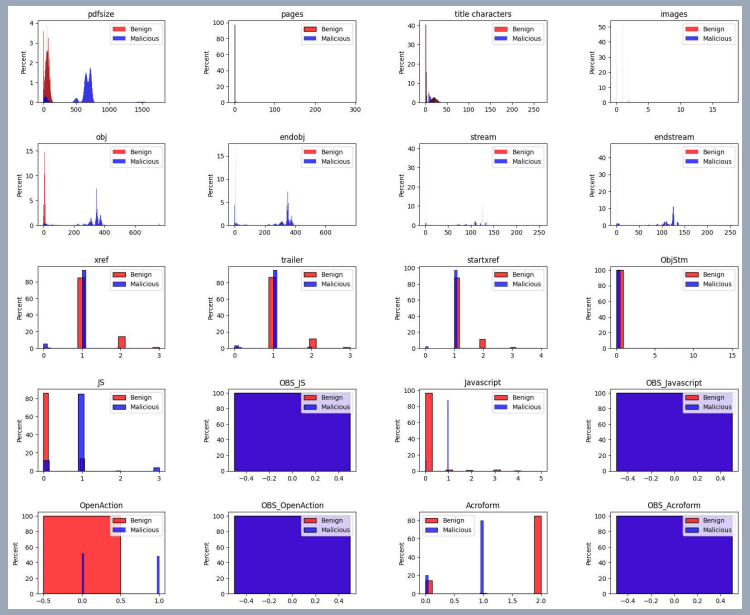
TOOLS AND ALGORITHMS

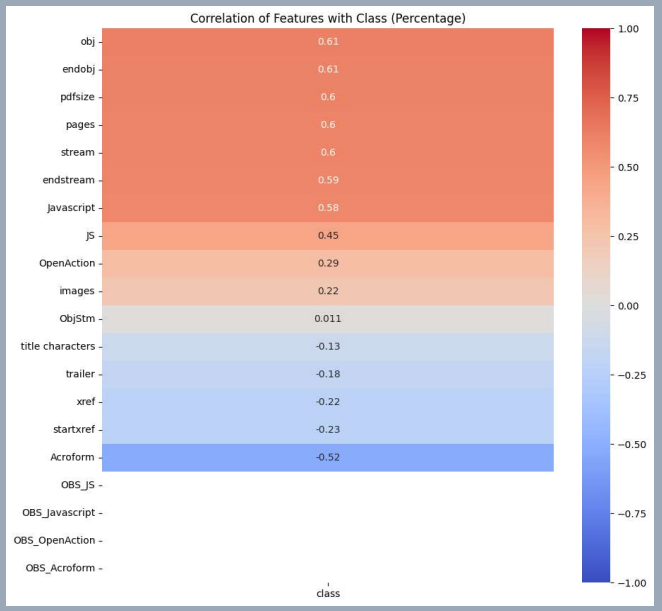
```
df.isna().any()
```

```
pdfsize            False
pages              False
title characters   False
images             False
obj                False
endobj             False
stream             False
endstream          False
xref               False
trailer            False
startxref          False
ObjStm             False
JS                 False
OBS_JS             False
Javascript         False
OBS_Javascript     False
OpenAction         False
OBS_OpenAction     False
Acroform           False
OBS_Acroform       False
class              False
dtype: bool
```

```
df.describe()
```

	pdfsize	pages	title characters	images	obj	endobj	stream	endstream	xref	trailer	...	ObjStm	JS	OBS_JS	Javascript	OBS_Javascript	OpenAction	OBS_OpenAction	Acroform	OBS_Acroform	class
count	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	500000.000000	...	500000.000000	500000.000000	500000.0	500000.000000	500000.0	500000.000000	500000.0	500000.000000	500000.0	500000.0
mean	563.363772	55.101686	5.617004	1.041594	273.595072	273.472290	95.115512	95.331500	0.969714	1.001358	...	0.008572	0.873134	0.0	0.795662	0.0	0.436600	0.0	0.887564	0.0	0.9
std	280.213763	30.233062	6.501397	0.734654	142.333280	142.734185	51.683914	52.094421	0.263349	0.244811	...	0.198168	0.547981	0.0	0.416932	0.0	0.495965	0.0	0.519314	0.0	0.3
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.000000	0.0	0.0
25%	585.864250	67.000000	0.000000	1.000000	266.000000	266.000000	85.000000	87.000000	1.000000	1.000000	...	0.000000	1.000000	0.0	1.000000	0.0	0.000000	0.0	1.000000	0.0	1.0
50%	657.841000	68.000000	4.000000	1.000000	346.000000	345.000000	123.000000	122.000000	1.000000	1.000000	...	0.000000	1.000000	0.0	1.000000	0.0	0.000000	0.0	1.000000	0.0	1.0
75%	708.503250	69.000000	9.000000	2.000000	355.000000	354.000000	126.000000	126.000000	1.000000	1.000000	...	0.000000	1.000000	0.0	1.000000	0.0	1.000000	0.0	1.000000	0.0	1.0
max	1761.042000	287.000000	267.000000	18.000000	760.000000	760.000000	254.000000	254.000000	3.000000	3.000000	...	15.000000	3.000000	0.0	5.000000	0.0	1.000000	0.0	2.000000	0.0	1.0





WHAT WE HAVE DONE SO FAR