

Vysvetliteľná klasifikácia rodín škodlivého kódu







Autor: Bc. Radovan Slíž
Vedúci práce: Ing. Peter Švec, PhD.



Motivácia

- Experimentálne porovnanie výsledkov klasifikačných modelov a na nich aplikovaných vysvetliteľných metód
- Pochopiť rozhodnutie klasifikačných modelov
- Vysvetlenie klasifikácie

Úlohy a ciele zadania

1.  Naštudovať si klasifikáciu a dynamické vlastnosti malwaru
2.  Dáta na tréningovanie (Dynamické vlastnosti, rodiny malwaru)
3.  Naštudovať si vysvetliteľné metódy
4.  Implementácia klasifikačných modelov
5.  Implementácia vysvetliteľných metód
6.  Metodika vyhodnocovania a samotné vyhodnotenie

CCCS-CIC-AndMal-2020 Dataset

Canadian Institute for Cybersecurity (CIC) project in collaboration with Canadian Centre for Cyber Security (CCCS)

- Android Malware
- 12 000 záznamov v našom datasete (1000 vzoriek pre každú z 12 rodín)
- 175 vlastností, atribútov správania
- Dynamické vlastnosti:
 - Pamäťové operácie
 - API
 - Sieťové operácie
 - Batéria
 - Logcat
 - Počet procesov



Canadian
Institute for
Cybersecurity



Klasifikačné modely

- SVM
- Random Forest
- Bagging Classifier
- Extra Trees Classifier
- Light Gradient-Boosting Machine
- ADABoost
- XGBoost

Explainable Artificial Intelligence - XAI

- Zlepšenie schopnosti ľudí pochopiť, interpretovať výstupy AI modelov a dôverovať im
- Môžeme napríklad riešiť otázky:
 - „Prečo bol malware priradený k rodine X?“
 - „Ktoré vlastnosti najviac ovplyvnili rozhodnutie modelu?“
 - „Aké sú najväčšie rozdiely medzi jednotlivými rodinami malwaru?“

Explainable Artificial Intelligence - XAI

- Možnosti interpretácie:
 - Lokálna - Vysvetlenie konkrétnej predikcie
 - Globálna - Vysvetlenie celkového správania modelu
- Na každý klasifikačný model aplikujeme lokálny aj globálny interpretačný model
- Očakávaných 14 rôznych výstupov na vyhodnotenie (7 klasifikátorov s 2 vysvetliteľnými metódami)

Ďalšie kroky

1. Vyladenie a dotrénovanie 7 klasifikačných modelov
2. Implementácia vysvetliteľných metód
3. Definovať metodiku vyhodnocovania a zrealizovať vyhodnotenie výsledkov