



România
Ministerul Apărării Naționale
Academia Tehnică Militară "Ferdinand I"

Facultatea de Sisteme Informatică și Securitate Cibernetică
Ingineria și Securitatea Sistemelor Informatic Militare

Platformă de Analiză Automată a Aplicațiilor Malițioase prin Utilizarea unor Algoritmi de Inteligență Artificială

Conducător Științific
Lect. Univ. Dr. Ing. Alin PUNCIOIU

Absolvent
Sd. Sg. Maj. George-Andrei IOSIF

București
2021

- ▶ Bancă nou-înființată

- ▶ Bancă nou-înființată
- ▶ Nevoia de a-și apăra infrastructura organizațională

- ▶ Bancă nou-înființată
- ▶ Nevoia de a-și apăra infrastructura organizațională
- ▶ Povara financiară a unei soluții comerciale

- ▶ Bancă nou-înființată
- ▶ Nevoia de a-și apăra infrastructura organizațională
- ▶ Povara financiară a unei soluții comerciale
- ▶ Rigiditatea unei versiuni gratuite a unei soluții comerciale

- ▶ Membrii ai departamentului de securitate cibernetică

- ▶ Membrii ai departamentului de securitate cibernetică
- ▶ Angajați din *front office*

- ▶ Membrii ai departamentului de securitate cibernetică
- ▶ Angajați din *front office*
- ▶ Alte sisteme ale organizației

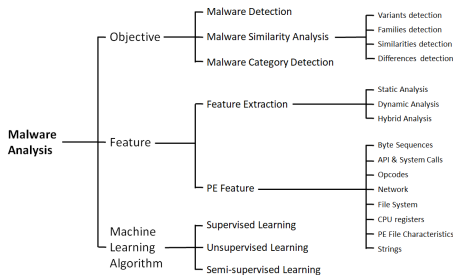
- ▶ Platformă cu sursă deschisă

- ▶ Platformă cu sursă deschisă
- ▶ Înzestrarea cu maleabilitatea inteligenței artificiale

- ▶ Platformă cu sursă deschisă
- ▶ Înzestrarea cu maleabilitatea inteligenței artificiale
- ▶ Condiția unor metrici numerice, obiective

- ▶ Caz de utilizare prezentat, ca prim motiv de creare a lucrării

- Caz de utilizare prezentat, ca prim motiv de creare a lucrării
- Articol " *Survey of machine learning techniques for malware analysis* "



Imagine 1: Taxonomie din Articolul Citat

- ▶ Oferirea numai a unor soluții teoretice, metodologii

- ▶ Oferirea numai a unor soluții teoretice, metodologii
- ▶ Nu și a unora practice, programe sau biblioteci de cod

- Demonstrarea beneficiilor aduse de inteligența artificială analizei de programe malițioase (*academic*)

- ▶ Demonstrarea beneficiilor aduse de inteligența artificială analizei de programe malițioase (*academic*)
 - ▶ Regresia maliției

- ▶ Demonstrarea beneficiilor aduse de inteligența artificială analizei de programe malițioase (*academic*)
 - ▶ Regresia maliției
 - ▶ Clasificarea în familii de programe malițioase

- ▶ Demonstrarea beneficiilor aduse de inteligența artificială analizei de programe malițioase (*academic*)
 - ▶ Regresia maliției
 - ▶ Clasificarea în familii de programe malițioase
 - ▶ Analiza de similaritate

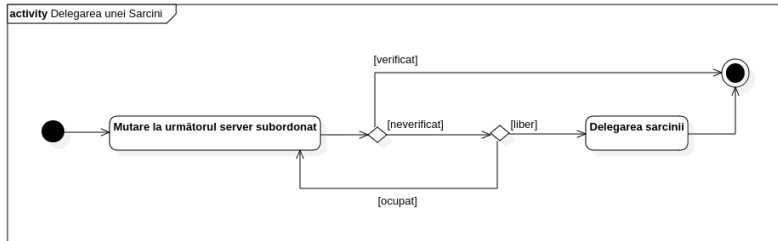
- ▶ Demonstrarea beneficiilor aduse de inteligența artificială analizei de programe malițioase (*academic*)
 - ▶ Regresia maliției
 - ▶ Clasificarea în familii de programe malițioase
 - ▶ Analiza de similaritate
- ▶ Platformă de automatizare a proceselor de ingineria datelor și analiză de programe (*practic*)

- ▶ Arhitectură de tip lider - subordonați

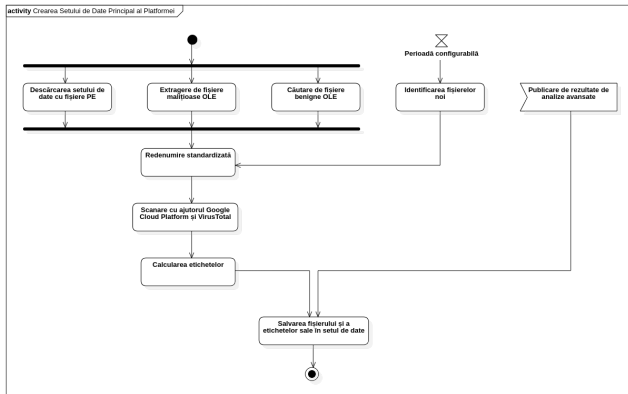
- ▶ Arhitectură de tip lider - subordonați
- ▶ Servere construite pe fundația modulelor

- ▶ Arhitectură de tip lider - subordonați
- ▶ Servere construite pe fundația modulelor
- ▶ Gruparea modulelor ca o linie de asamblare

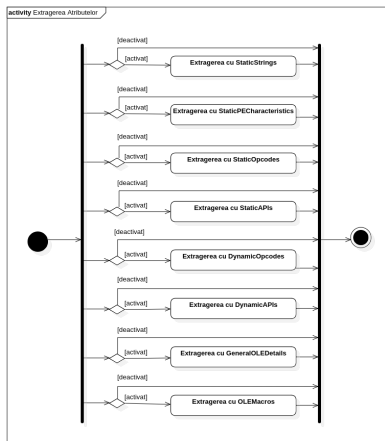
- ▶ Gestionarea serverelor subordonate
- ▶ Etichetarea fișierelor și gestionarea seturilor de date
- ▶ Extragerea atributelor din fișiere
- ▶ Preprocesarea atributelor
- ▶ Gestionarea modelelor de inteligență artificială



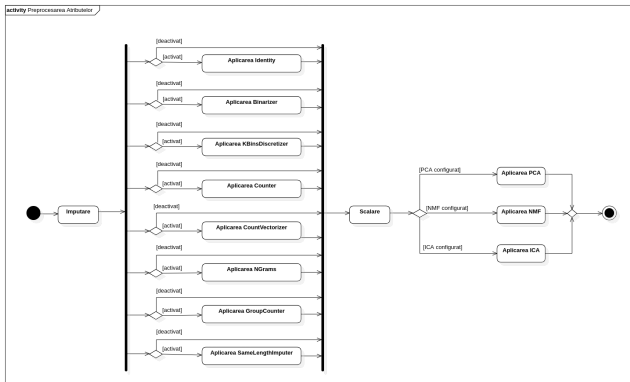
Imagine 2: Arhitectura Modulului pentru Gestionarea Serverelor Subordonate



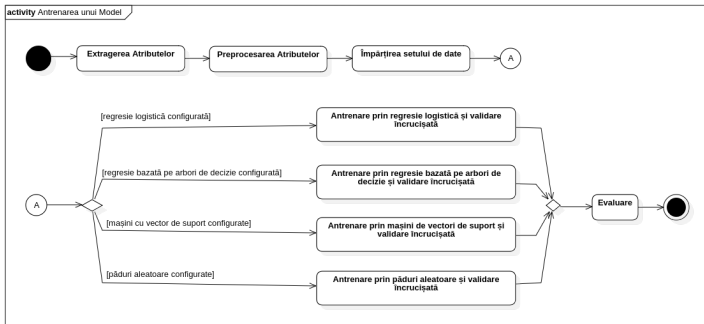
Imagine 3: Arhitectura Modulului pentru Etichetarea Fișierelor și Gestionarea Seturilor de Date



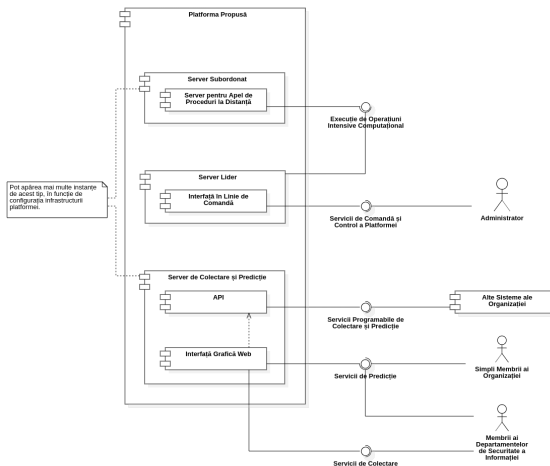
Imagine 4: Arhitectura Modulului pentru Extragerea Atributelor din Fișiere



Imagine 5: Arhitectura Modulului pentru Preprocesarea Atributelor



Imagine 6: Arhitectura Modulului pentru Gestionarea Modelelor de Inteligență Artificială



Imagine 7: Arhitectura de Servere a Platformei

► Tehnologii folosite



► Tehnologii folosite



VirusTotal



pandas



YML



L^AT_EX

► Aspecte considerate

► Tehnologii folosite



Σ VirusTotal



📊 pandas



YAML



L^AT_EX

► Aspecte considerate

- Calcul distribuit și paralel

► Tehnologii folosite



► Aspecte considerate

- Calcul distribuit și paralel
- Containerizare

► Tehnologii folosite



► Aspecte considerate

- Calcul distribuit și paralel
- Containerizare
- Documentare

► Tehnologii folosite



► Aspecte considerate

- Calcul distribuit și paralel
- Containerizare
- Documentare
- Configurabilitate

- ▶ Semi-automată, în timpul dezvoltării

- ▶ Semi-automată, în timpul dezvoltării
 - ▶ De integrare ascendentă

- ▶ Semi-automată, în timpul dezvoltării
 - ▶ De integrare ascendentă
 - ▶ De regresie

- ▶ Semi-automată, în timpul dezvoltării
 - ▶ De integrare ascendentă
 - ▶ De regresie
- ▶ În cadrul unui proiect de cercetare al Academiei Tehnice Militare " *Ferdinand I* " București

1. Crearea de seturi de date variate

1. Crearea de seturi de date variate
2. Deducerea algoritmilor cu cele mai bune rezultate

1. Crearea de seturi de date variate
2. Deducerea algoritmilor cu cele mai bune rezultate
 - 2.1 Analiza componentelor principale pentru reducerea dimensionalității

1. Crearea de seturi de date variate
2. Deducerea algoritmilor cu cele mai bune rezultate
 - 2.1 Analiza componentelor principale pentru reducerea dimensionalității
 - 2.2 *Random Forest* pentru învățare automată

1. Crearea de seturi de date variate
2. Deducerea algoritmilor cu cele mai bune rezultate
 - 2.1 Analiza componentelor principale pentru reducerea dimensionalității
 - 2.2 *Random Forest* pentru învățare automată
3. Antrenarea de modele

1. Crearea de seturi de date variate
2. Deducerea algoritmilor cu cele mai bune rezultate
 - 2.1 Analiza componentelor principale pentru reducerea dimensionalității
 - 2.2 *Random Forest* pentru învățare automată
3. Antrenarea de modele
4. Evaluare automată

Formatul fișierelor	PE
Numărul de exemplare incluse	1000
Raportul dintre fișiere benigne și malițioase	0.5
Maliție minimă	0.9
Familii de programe malițioase	Toate

Tabel 1: Configurația Setului de Date LARGE_PE

Formatul fișierelor	PE
Numărul de exemplare incluse	1000
Raportul dintre fișiere benigne și malițioase	0.5
Maliție minimă	0.9
Familii de programe malițioase	Toate

Tabel 1: Configurația Setului de Date LARGE_PE

Descriere	Regresia maliției pentru fișiere PE, cu extractori ce asigură o procesare rapidă a fișierelor
ID-ul setului de date	LARGE_PE
Extractori	STATIC_STRINGS, STATIC_PE_CHARACTERISTICS
Preprocesoare	IDENTITY, COUNTER, K_BINS_DISCRETIZER, N_GRAMS

Tabel 2: Configurația Modelului PE_STATIC_FAST_MALICE

Eroare maximă	0.7169
Eroare medie absolută	0.0303
Rădăcina erorii medie pătratice	0.0743
Scorul R^2	0.974

Tabel 3: Evaluarea Modelului PE_STATIC_FAST_MALICE

- ▶ Întrebări la care am răspuns

- ▶ Întrebări la care am răspuns
 - ▶ De ce?

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?
 - ▶ Cum?

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?
 - ▶ Cum?
 - ▶ Cu ce rezultate?

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?
 - ▶ Cum?
 - ▶ Cu ce rezultate?
- ▶ Realizări

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?
 - ▶ Cum?
 - ▶ Cu ce rezultate?
- ▶ Realizări
 - ▶ Crearea unui set de date etichetate

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?
 - ▶ Cum?
 - ▶ Cu ce rezultate?
- ▶ Realizări
 - ▶ Crearea unui set de date etichetate
 - ▶ Crearea de metode automate de extragere a unor atribute

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?
 - ▶ Cum?
 - ▶ Cu ce rezultate?
- ▶ Realizări
 - ▶ Crearea unui set de date etichetate
 - ▶ Crearea de metode automate de extragere a unor atribute
 - ▶ Crearea unui proces automatizat de inginerie a datelor

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?
 - ▶ Cum?
 - ▶ Cu ce rezultate?
- ▶ Realizări
 - ▶ Crearea unui set de date etichetate
 - ▶ Crearea de metode automate de extragere a unor atribute
 - ▶ Crearea unui proces automatizat de inginerie a datelor
 - ▶ Integrarea celor menționate în cadrul unei platforme

- ▶ Întrebări la care am răspuns
 - ▶ De ce?
 - ▶ Ce?
 - ▶ Cum?
 - ▶ Cu ce rezultate?
- ▶ Realizări
 - ▶ Crearea unui set de date etichetate
 - ▶ Crearea de metode automate de extragere a unor atribute
 - ▶ Crearea unui proces automatizat de inginerie a datelor
 - ▶ Integrarea celor menționate în cadrul unei platforme
 - ▶ Integrarea soluției software dezvoltate în cadrul unui proiect de cercetare

► Îmbunătățiri

- ▶ Îmbunătățiri
 - ▶ Echilibrarea setului de date

- ▶ Îmbunătățiri
 - ▶ Echilibrarea setului de date
 - ▶ Extinderea funcționalității extractorilor dinamici

- ▶ Îmbunătățiri
 - ▶ Echilibrarea setului de date
 - ▶ Extinderea funcționalității extractorilor dinamici
- ▶ Funcționalități noi

- ▶ Îmbunătățiri
 - ▶ Echilibrarea setului de date
 - ▶ Extinderea funcționalității extractorilor dinamici
- ▶ Funcționalități noi
 - ▶ Crearea unor extractori noi cu tehnici de analiză dinamică

- ▶ Îmbunătățiri
 - ▶ Echilibrarea setului de date
 - ▶ Extinderea funcționalității extractorilor dinamici
- ▶ Funcționalități noi
 - ▶ Crearea unor extractori noi cu tehnici de analiză dinamică
 - ▶ Studiarea scalării soluției la organizațiile mari

- ▶ Soluția software dezvoltată¹
- ▶ Set de date²
- ▶ Lucrarea scrisă și prezentare³

¹<https://github.com/iosifache/dike>

²<https://github.com/iosifache/DikeDataset>

³<https://github.com/iosifache/BachelorThesis>