Detectarea Aplicațiilor Malițioase folosind Metode de Învățare Automată

COORDONATOR:
Prof. Dr. Viorel Negru
Drd. Mario Reja

STUDENT:
Alexandru-Sebastian
Tufiș-Schwartz

Universitatea de Vest din Timișoara

Introducere

Lucrarea **Detectarea Aplicațiilor Malițioase folosind Metode de Învățare Automată** prezentă o abordare a diferițiilor algoritmi de "machine learning" pentru a depista dacă un fișier este malițios sau curat.

Malware

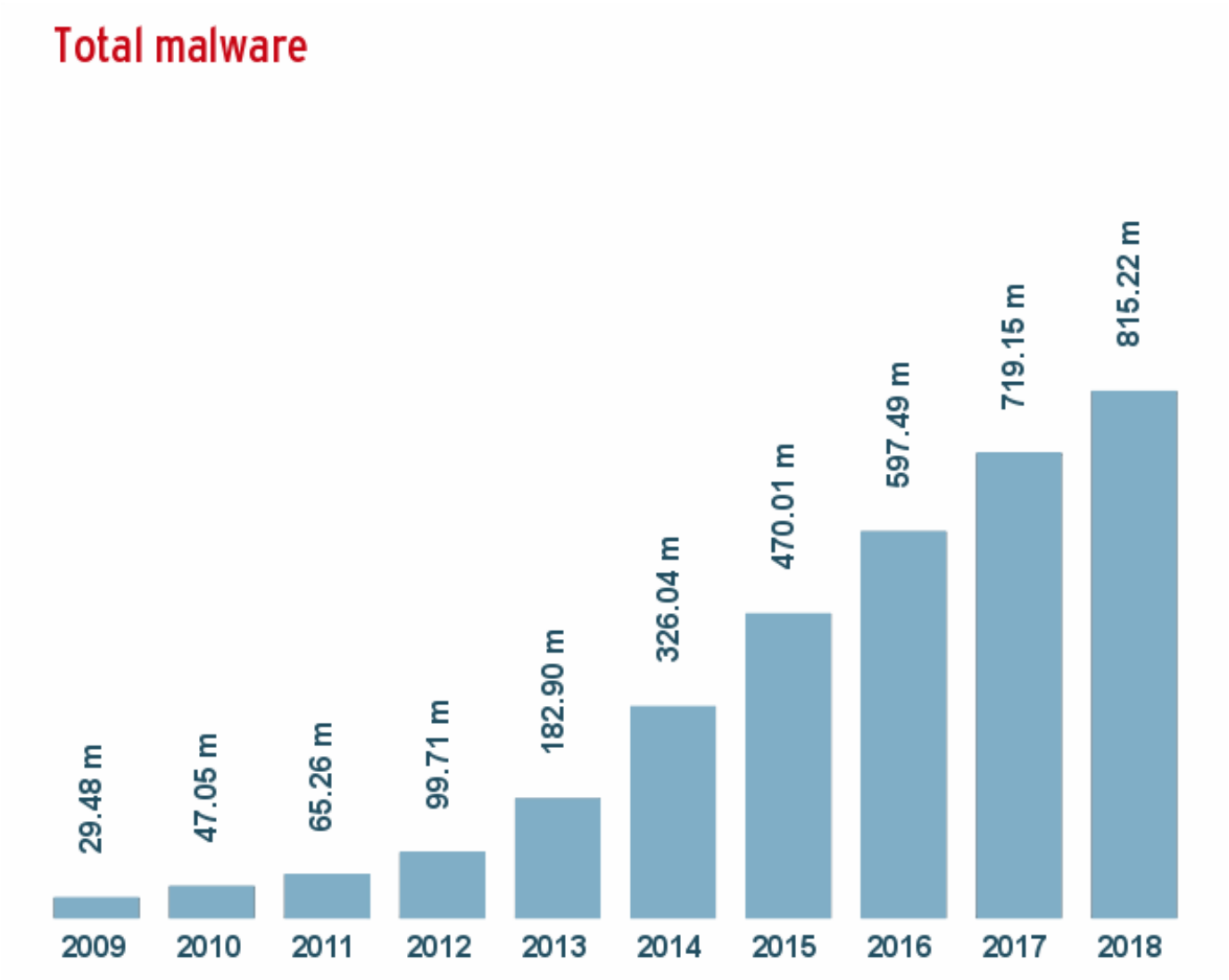

Figura: Creșterea numarului de programe malițioase

Tehnologii

Detectarea Aplicațiilor Malițioase folosind Metode de Învățare Automată a fost creată folosind Python și scikit-learn

Setul de date

Setul de date conține 54 de caracteristici si a fost cteat folosind 10539 PE-files dintre care 6999 malițioase și 3540 curate.

Invățare automată

Unul din sub-domeniile de bază ale Inteligenței Artificiale, se preocupă cu dezvoltarea de algoritmi și metode ce permit unui sistem informatic să învețe date, reguli, chiar algoritmi

KNN

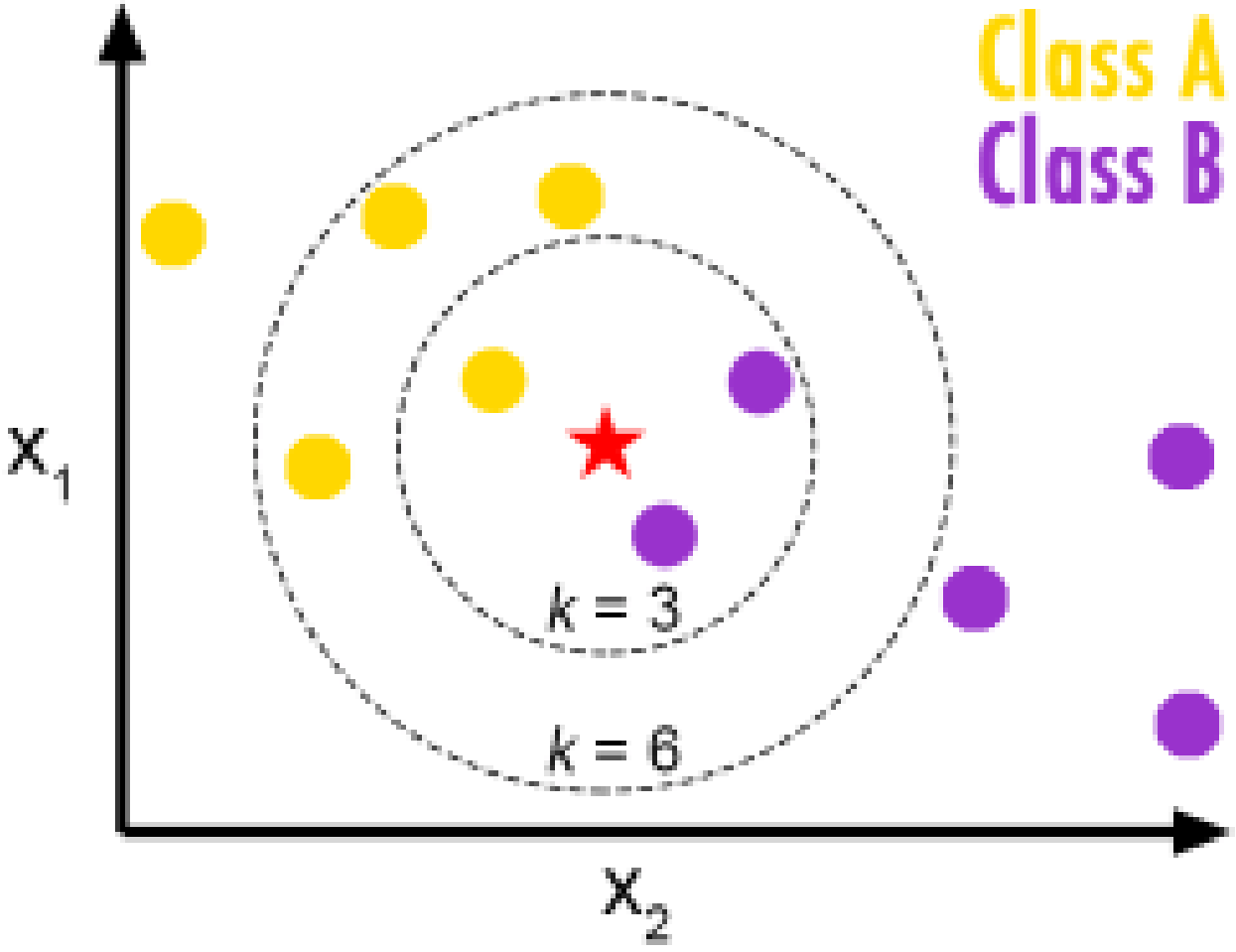

Figura: Vizualizare KNN

Random Forest

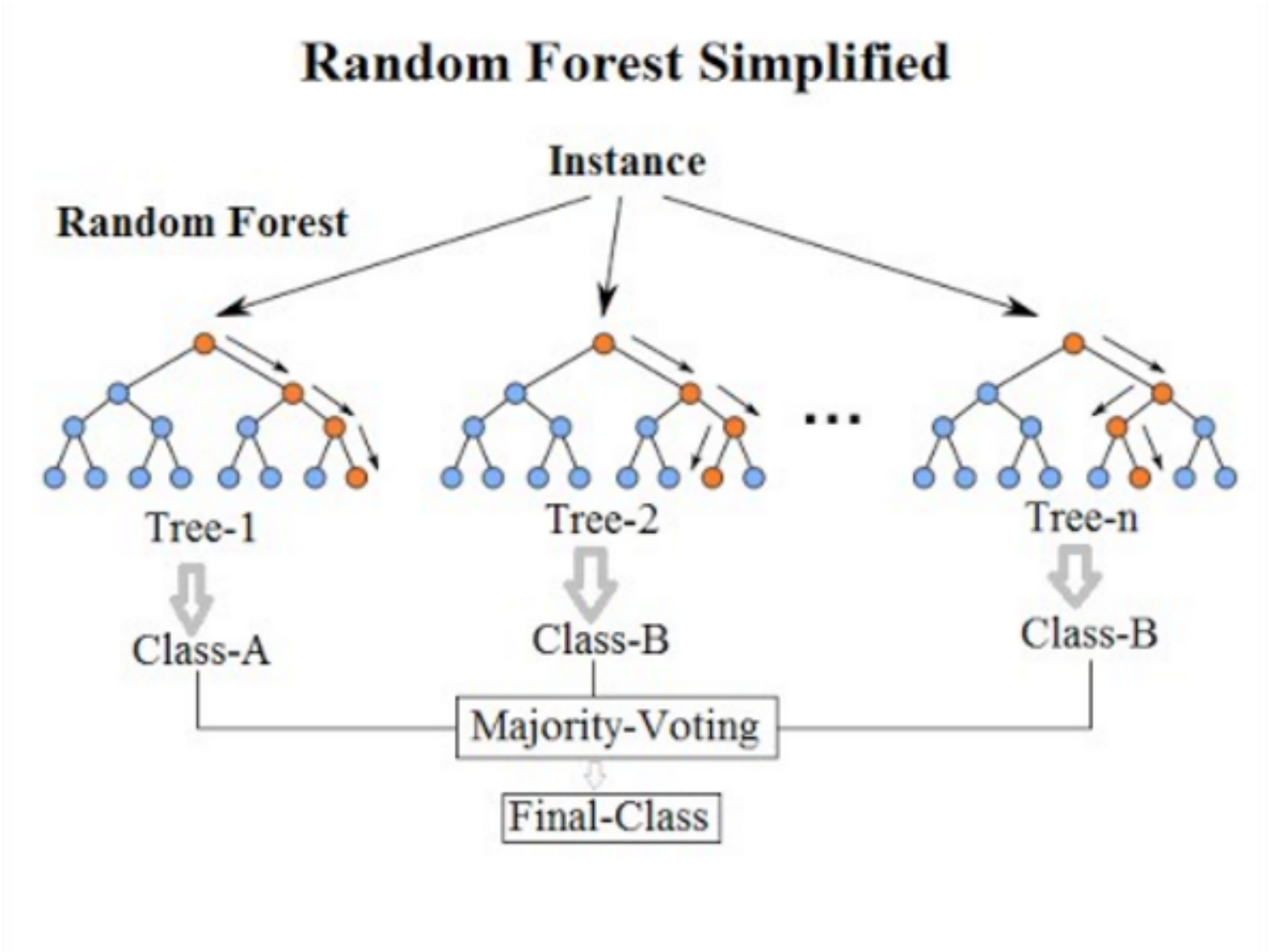

Figura: Vizualizare Random Forest

XGBoost

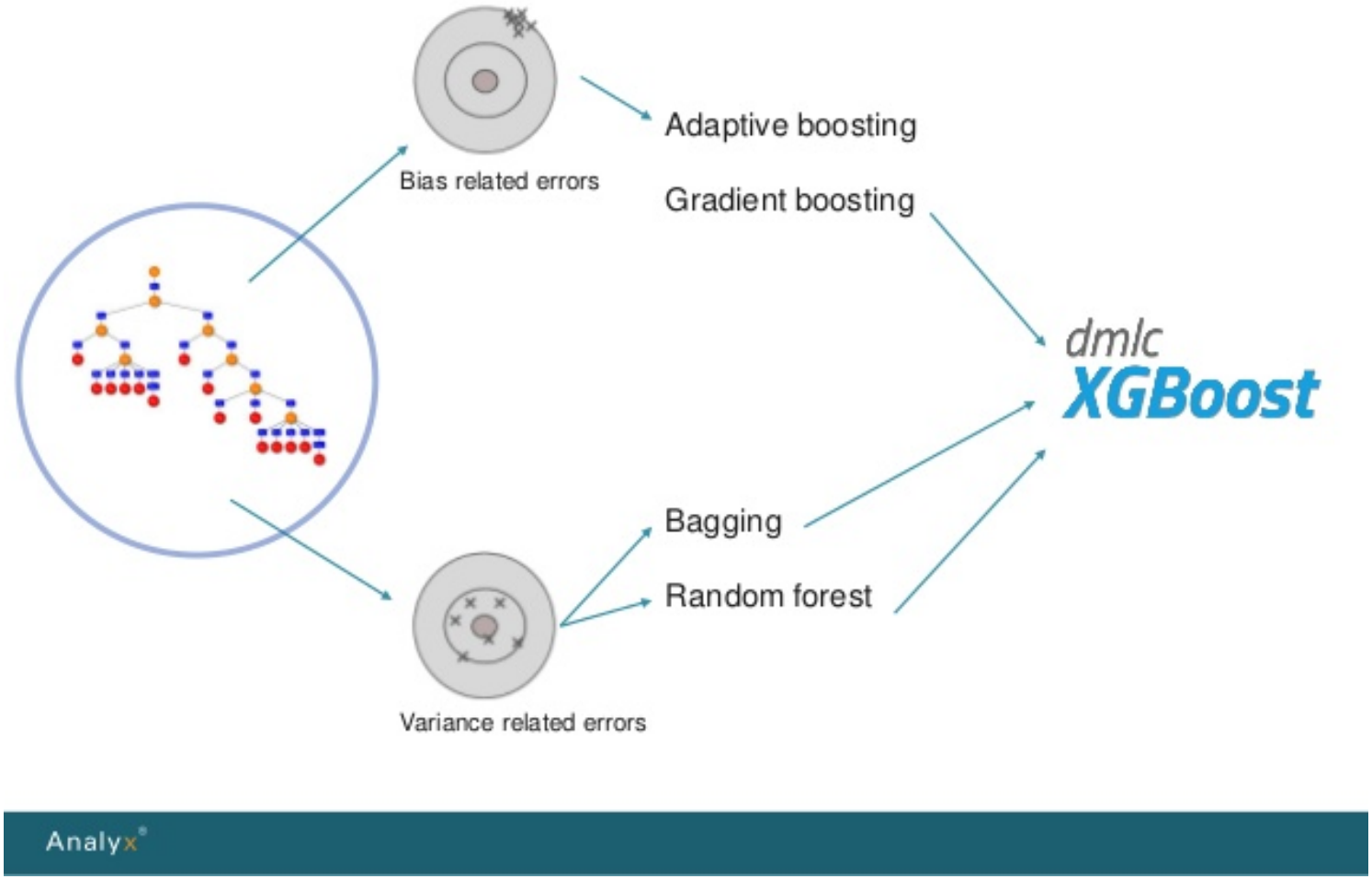

Figura: Vizualizare XGBoost

Rezultate folosind 54 de caracteristici

	0	1
0	1368	49
1	27	664

Figura: Rezultat KNN

2032 preziceri corecte : 76 preziceri gresite. Acuratețe = 96.4%.

	0	1
0	1408	9
1	28	663

Figura: Rezultat Random Forest

2071 preziceri corecte : 37 preziceri gresite. Acuratețe = 98.24%.

	0	1
0	1402	15
1	21	670

Figura: Rezultat XGBoost

2072 preziceri corecte : 36 preziceri gresite. Acuratețe = 98.29%.

Selectare de caracteristici

Lista cu caracteristicile pastrate si importanta acestora folosind Tree-based feature selection

1. feature MajorSubsystemVersion (0.155935)
2. feature Characteristics (0.118596)
3. feature MajorOperatingSystemVersion (0.110355)
4. feature ImageBase (0.108349)
5. feature Machine (0.068662)
6. feature DllCharacteristics (0.050951)
7. feature SectionsMaxEntropy (0.048521)
8. feature LoadConfigurationSize (0.038567)
9. feature ResourcesMaxEntropy (0.038000)
10. feature MajorLinkerVersion (0.032729)
11. feature ResourcesMinSize (0.022452)

Rezultate folosind 11 caracteristici

	0	1
0	1387	30
1	24	667

Figura: Rezultat KNN

2054 preziceri corecte : 54 preziceri gresite. Acuratețe = 97.44%.

	0	1
0	1405	12
1	24	667

Figura: Rezultat Random Forest

2072 preziceri corecte : 36 preziceri gresite. Acuratețe = 98.29%.

	0	1
0	1409	8
1	27	664

Figura: Rezultat XGBoost

2073 preziceri corecte : 35 preziceri gresite. Acuratețe = 98.34%.