Università di Pisa
Dipartimento di Ingegneria dell'Informazione

Malicious URL Detector

Marco Fresco & Fabio Fadda

Artificial Intelligence for Cybersecurity
Master’s Degree in Cybersecurity
Academic Year 2023/24

Goal Definition

Drive-By Downloads
Execute malicious code on the victim's system

Malware Distribution
Host or redirect to sites that distribute malware

Phishing Attacks
Lead users to fake websites that mimic legitimate ones

Real-Time Protection
Identifying and blocking access to harmful URLs

Early Threat Detection
Preventing users from interacting with dangerous content

Reducing Attack Surface
Prevent potential entry points for cyberattacks

Common Characteristics of Malicious URLs

Misspelled Domain Names

Cybercriminals register domains that are intentionally similar to well-known websites

Unusual Characters

Used to confuse users or evade detection

IP Addresses

Raw IP Addresses instead of domain names to bypass domain registration requirements

Lack of HTTPS

Phishing sites may not have valid SSL certificates (HTTPS)

Long, Random Strings

Attempt to obfuscate the true purpose of the URL

Overuse of Subdirectories

A technique to obscure the final destination

Data Gathering

PhishTank

Data and information about phishing on the Internet.

Kaggle

Platform for data science competitions

Data Pre-Processing

01 - Data Cleaning

Identifying and correcting errors, inconsistencies and inaccuracies

02 – Class Balancing

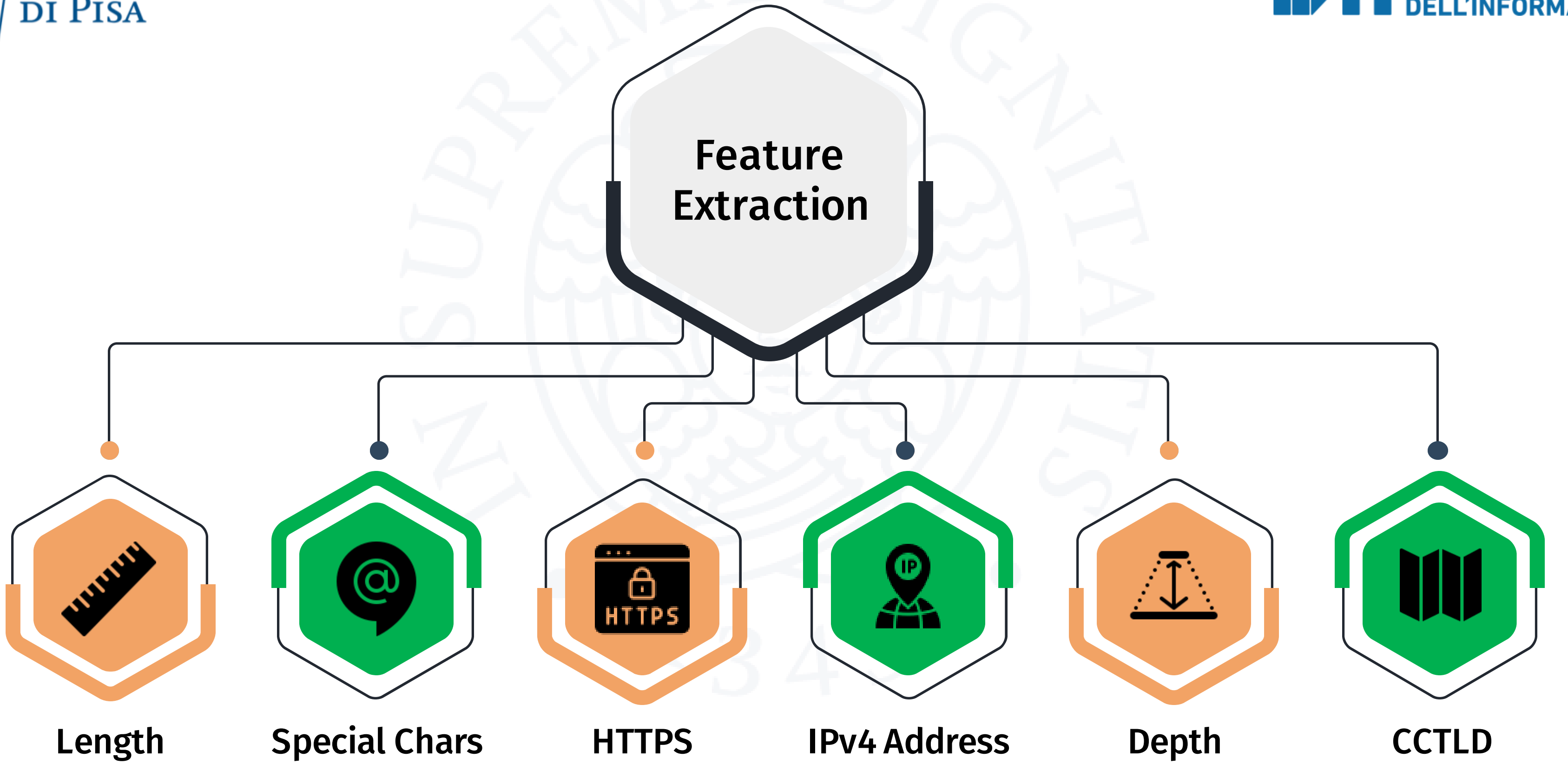
As imbalancing can lead to biased model and misleading accuracy

03 – Feature Extraction

Enhancing the model's ability to learn patterns in the data

04 – Heuristics

Comparison with some of the empirical rules found in literature


Feature Extraction
HTTPS
Length
Special Chars
HTTPS
IPv4 Address
Depth
CCTLD

Heuristics Confrontation

Special Chars

'@' Symbol: 93.827%
'-' Symbol: 90.126%
'_' Symbol: 53.846%

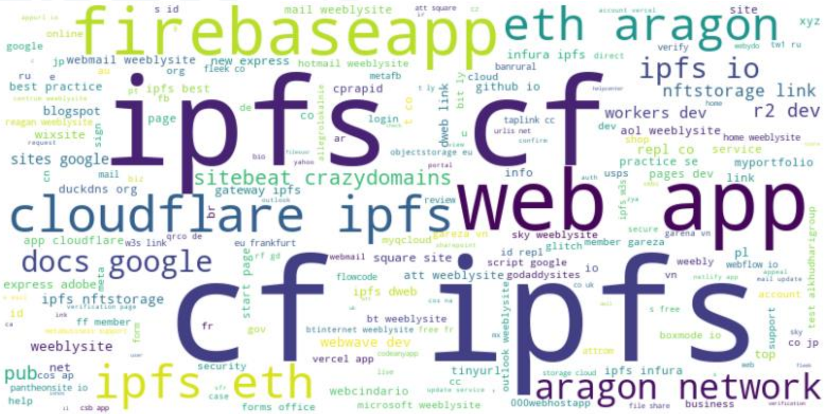
IPv4 Address

100% Malicious URLs

Reference: Youness Mourtaji, Mohammed Bouhorma, Daniyal Alghazzawi, Ghadah Aldabbagh, Abdullah Alghamdi, "Hybrid Rule-Based Solution for Phishing URL Detection Using Convolutional Neural Network".

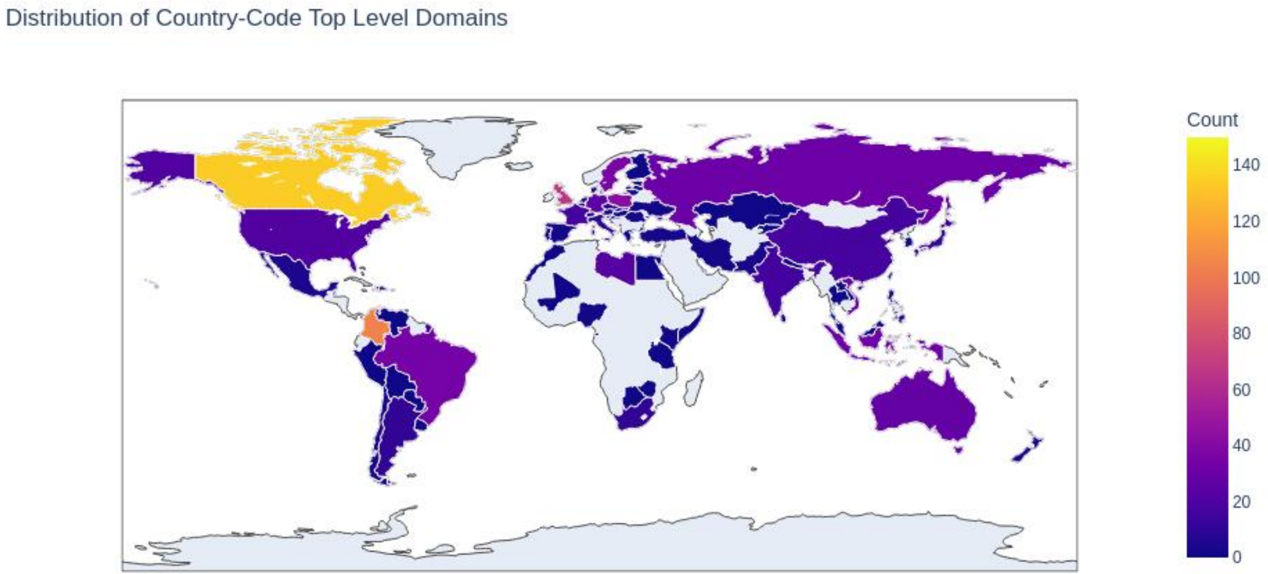
Preliminary Data Exploration

Benign



Malicious

Preliminary Data Exploration

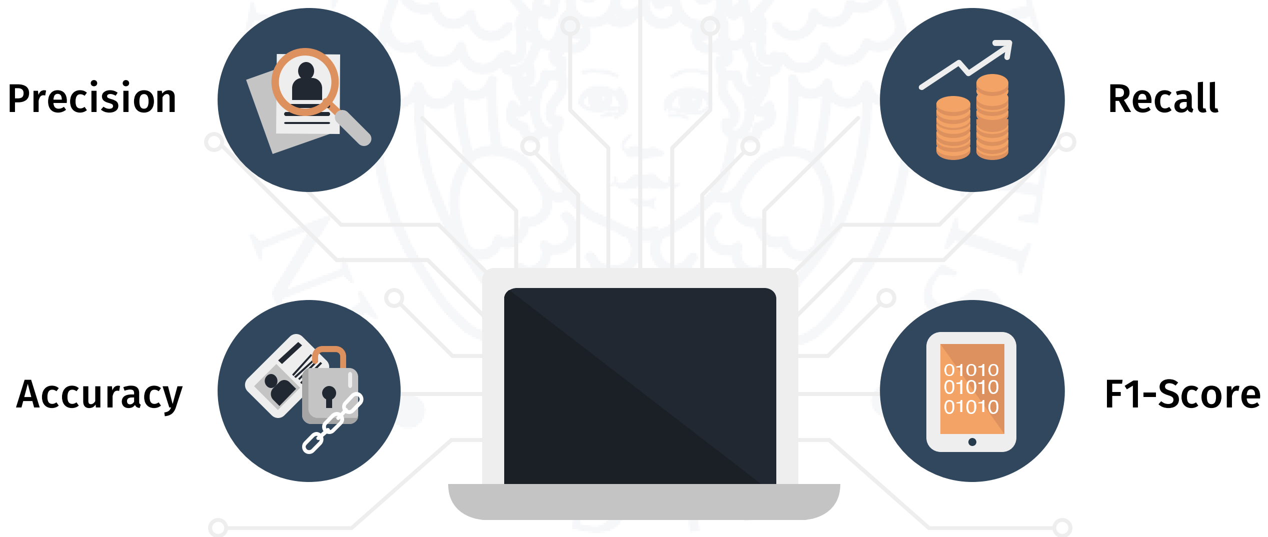


Normalization process

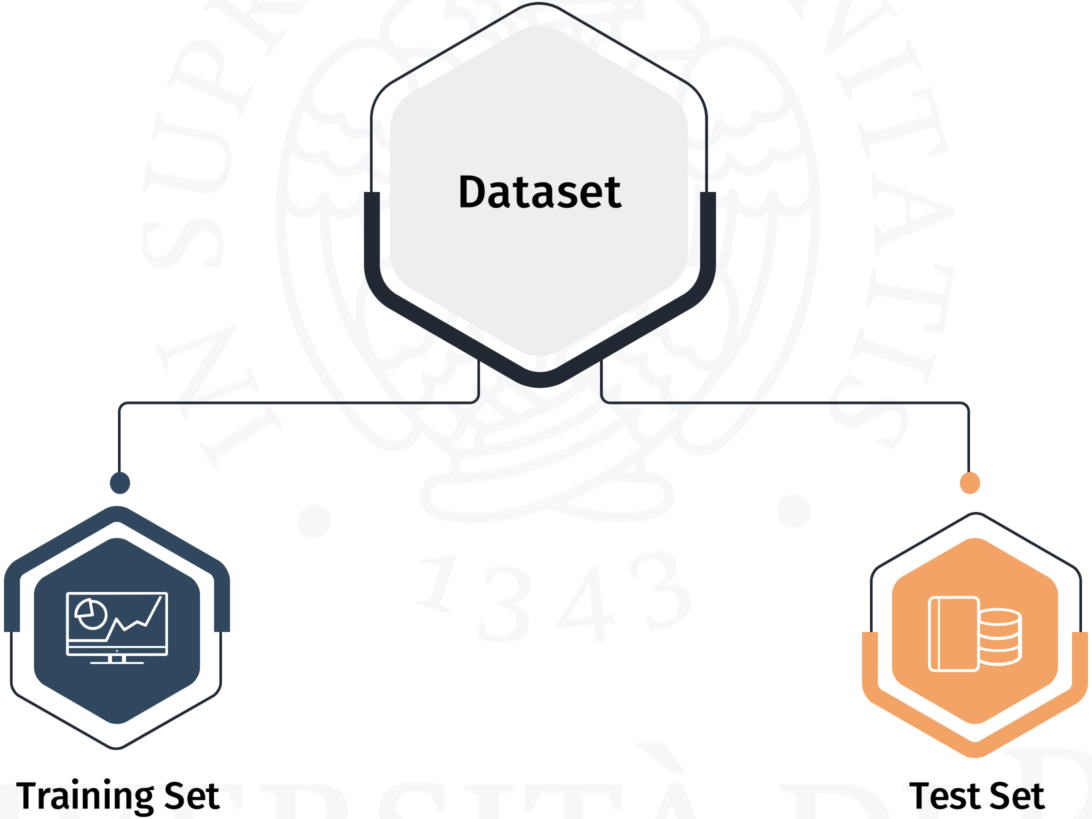
Data Mining

- Decision Tree Classifier
- Random Forest Classifier
- Logistic Regression
- Gaussian Naive Bayes
- Linear SVC

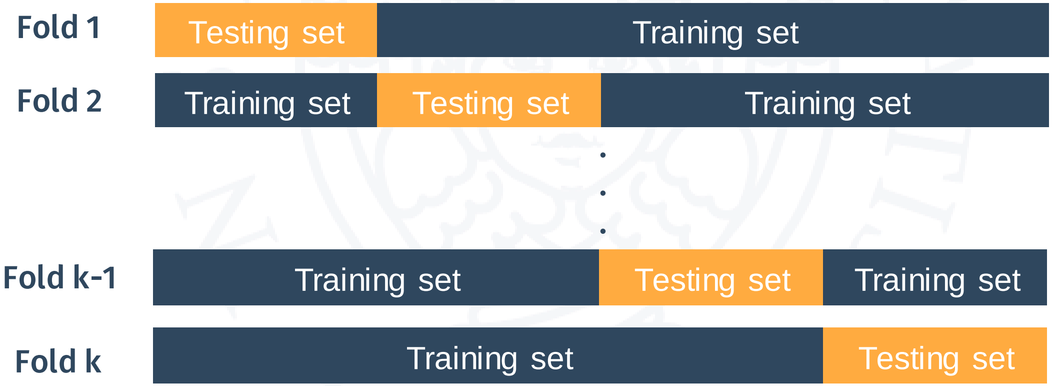
Performance Evaluation



Holdout Method

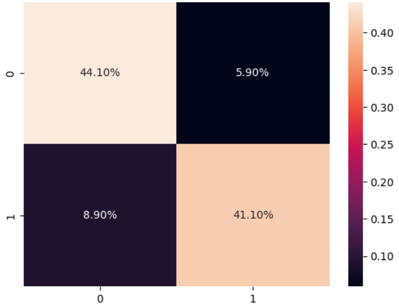


Stratified K-Fold



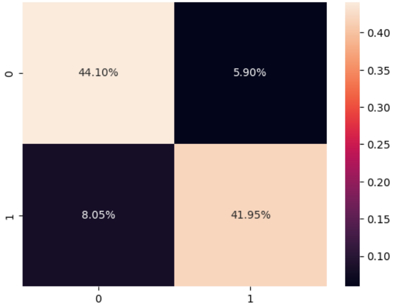
Results – Hold out

Decision Tree Classifier



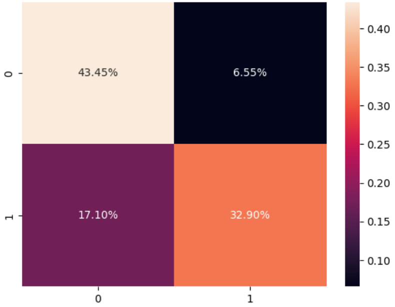
85.20%

Random Forest Classifier



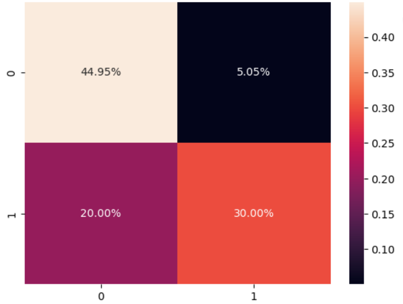
86.05%

Logistic Regression



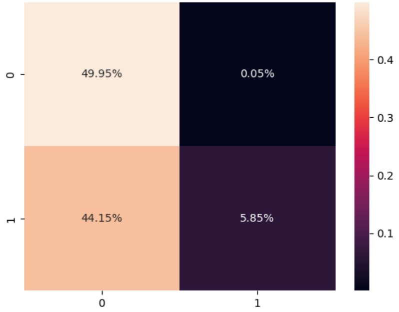
78.35%

Linear SVC



74.95%

Gaussian Naive Bayes

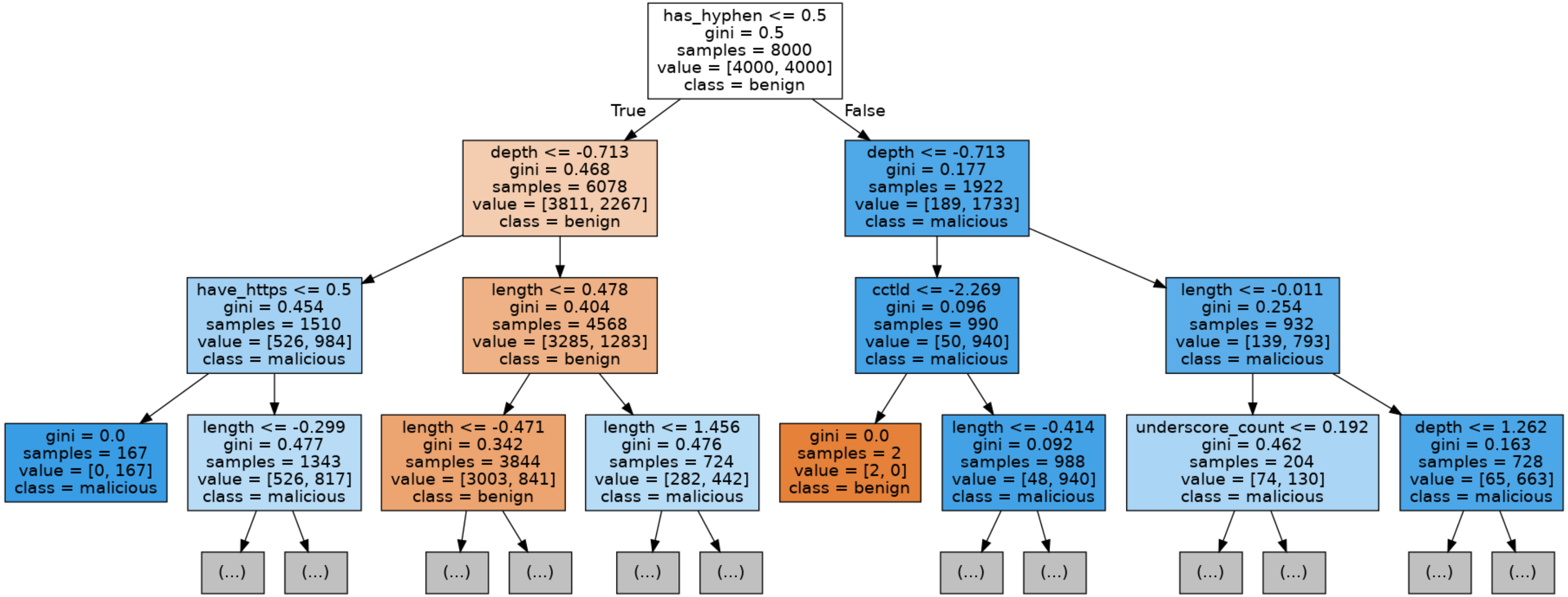


55.80%

Results – Accuracy

	Model	Accuracy	Accuracy with SKF	Accuracy with Feature Selection
0	DecisionTreeClassifier	0.8520	0.842	0.7770
1	RandomForestClassifier	0.8605	0.852	0.7785
2	LogisticRegression	0.7635	0.736	0.6655
3	LinearSVC	0.7495	0.740	0.6700
4	GaussianNB	0.5580	0.561	0.5520

Decision Tree Representation



System Improvement

HTML encoding

JavaScript code

Request parallelization

Updated/Tested Dataset

Thanks For The Attention!