NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Advertising – Machine Learning algorithms to detect anomalies

Valentyna Rusinova

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics with specialization in Data

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

ADVERTISING – MACHINE LEARNING ALGORITHMS TO DETECT ANOMALIES

by

Valentyna Rusinova

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics with specialization in Data Science

Advisor: Professor Doutor Vitor Manuel Pereira Duarte dos Santos

June 2022

ACKNOWLEDGEMENTS

I would like to thank and express my gratitude to my thesis supervisor Professor Doctor Vítor Duarte dos Santos from NOVA School of Information Management at NOVA University of Lisbon, because he challenged me, with a theme that captivated my interest, to write this master thesis. I would like to thank him for supporting me at every step of the process, dedicating himself full time and answering all questions, for the friendliness he showed in several meetings, and also for challenging me to publish this work, thank you a lot. He was undoubtedly a person who marked and supported me throughout the realization of this project.

I would like to thank and dedicate this work to my family. I leave here my gratitude for believing and betting on me, making me feel that I was capable of achieving any goal and overcoming any obstacle that might appear.

I also leave here a special thanks to my friends, who were always there whenever I needed them and were a crucial and vital help for this work.

ABSTRACT

Security is a major worry concern nowadays, as the technological development has led to an excessive use of digital devices, where everyone surfs web pages, social networks, blogs, etc. The internet is currently where people spend most of their free time, where they search for and place their information. However, many individuals take advantage of this information for malicious purposes, such as identity, and bank account theft, or even to compromise documents. Furthermore, the methods used not only cause damage to individuals but also affect electronic devices, which become infected with malware, and often become impossible to use again. Consequently, people are becoming progressively worried about digital security, which pushes them to increasingly use software that blocks all kinds of malware. But for the development of Digital Marketing, overcoming these negative consequences becomes a real challenge, as it is affected by ad-blocking software, since most malware is embedded in advertisements, of which criminal minds try to take advantage of. Hence, to overcome the cyber-attacks that can arise from malicious advertisements and prevent internet users from using ad blockers, Digital Marketing will have to find strategies to develop security techniques. With the help of Digital Forensic Science, it is possible to conduct investigations to solve the problems related to digital crime. The expansion of cybersecurity allowed to develop a web extension with which it is possible to block malicious ads, whilst simultaneously allowing for digital advertising not to vanish, but to continue evolving ensuring possible the dissemination of products and information, on which all individuals depend.

KEYWORDS

Machine Learning; Digital Forensics; Cybercrime; Cybersecurity; Advertising; UBlock

PUBLICATIONS

Before the publication and defense of this Master's dissertation, there was the opportunity to submit for approval an article. As such, this work was submitted for approval containing 18 pages in the Journal of Advertising Research, written by the author of this thesis Valentyna Rusinova, together with Professor Dr. Vítor Santos, who is the supervisor of this thesis and work.

INDEX

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

AD Anomaly Detection

AI Artificial Intelligence

ANN Artificial Neural Network

API Application Programming Interface

CNN Convolutional Neural Network

CPU Central Processing Unit

DOM Document Object Model

DoS Denial of Service

DSR Design-Science Research

DT Decision Tree

GAN Generative Adversarial Network

GPU Graphics Processing Unit

HTML Hyper Text Markup Language

ICT Information Communication Technology

IE Information Extraction

IP Internet Protocol

IR Information Retrieval

KNN K-Nearest Neighbors

LSTM Long-short-term Memory Network

ML Machine Learning

MLP Multilayer Perceptron

MSE Mean Squared Error

NLP Natural Language Processing

PPC Pay Per Click

RNN Recurrent Neural Network

ROI Return on Investment

SOM Self-Organizing Maps

SQL Structured Query Language

SVM Support Vector Machine

TTL Time to Live

URL Uniform Resource Locator

1. Introduction

1.1. Context

The technological development allowed the internet to be a part of peoples' lives and, besides that, the number of users increased across the years (Cintra, n.d.). Wide amounts of information are created and propagate at great speed, not just across companies, but around the world. Citizens use and capture information through different means. The most common methods to disseminate and use information is the internet, which is mostly used by individuals on their electronic devices.

The internet started to consume most of people's time, with a vast majority of time spent online on social media, using email, watching videos, news and shopping (broadbandsearch, 2020). Also, advanced image editing software enabled designers to quickly and creatively create posters, flyers and other materials. This succession of events made the internet an essential tool for advertising (Cintra, s.d.), as advertising can be sent by email, reaching a larger number of individuals instantly, without the need for time-consuming distribution.

Nowadays, digital marketing is one of the main methods that companies use to communicate with the public directly, on a personalized manner and at the right time.

Online advertising allows businesses to reach customers using more innovative and cost-effective methods, such as social media, display and paid search, which allows companies to reach a large audience more affordably than traditional advertising would (Herhold, 2018).

One of the main advantages of online marketing is its low operating cost, as sources as social media, blogs and search engine optimization for advertising can be used. There are no travel costs and no printing costs to distribute materials, and server costs are relatively low. Furthermore, digital marketing is faster and easier than traditional marketing, as it is possible to start a marketing campaign and receive customer feedback in real time (RedAlkemi, 2019).

With digital advertising it is easier for companies to reach new customers around the world, without traveling anywhere, businesses can be expanded and can offer personalized experiences to each customer. Marketers can track the success of their campaigns in real time and make adjustments as needed. Advertising tools like online image ads, social media ads like Facebook and Twitter ads, or video ads allow to reach the target audience. These tools will help generate an increasing number of leads and make the campaigns even more effective and profitable (Wroblewski, 2018).

However, these ads affect internet navigation as they appear in the middle of the pages when users try to have access to, or in the middle of a movie, destroying a crucial moment. Let's not forget that ads can equally prompt unwanted windows with inappropriate content and some ads can be illegal.

Some advertisements spread malware by injecting malicious advertisement into legitimate online advertising networks and web pages. This is known as Malvertising, a relatively new cyberattack technique (LIFARS, 2020). It is a form of cybercrime, translating a behavior that violates the law and is punishable, but can only be committed through computers, computer networks or other forms of information communication technology (ICT). It may involve system invasions, virus dissemination, personal data theft, misrepresentation, access to confidential information and many others (Apav, s.d.).

Malware can run without users actively clicking on the malware ad, via visiting the legal page, the attacker looks for vulnerabilities, finding them can redirect the browser to malicious websites or force “drive per download” to execute the code that installs malware on the user's computer. These types of malicious websites look like legitimate websites and users may not be aware of that (Editor, 2016).

Malware can be inserted into advertisements in a number of ways, such as:

- Insertion of malware through third parties, servers can be compromised by an attacker, who can add malicious code to the ad payload.
- After clicking on an advertisement, the URL to a new page may be infected, thus the attacker may execute malicious code.
- When the user clicks on the ad, it redirects to the legitimate web page, but there may be clickable elements that execute malicious code.
- The ad itself may be infected with malware. Videos such as VAST (Video Ad Serving Template) may display a malicious URL at the end of the video.
- Even the Flash video can inject an Iframe ((Inline Frame) is an embedded HTML document inside another HTML document on a website) in the malware download page without clicking on the video.

The companies that publish the ads are often harmed by malvertising, as it damages their reputation, decrease their profits and the ads lose their legal responsibility. It is difficult to test every advertisement that is shown to the user, due to the large amount of published advertisements, and therefore it is difficult to block the malicious advertisement (LIFARS, 2020).

Will we be protected from cyber-attacks when new windows appear? Will viewing a promotional video be safe? Not only our browsing data, but also our computer data can be stolen and endangered.

This problem can be solved with digital forensics which fights cybercrime using forensics to track, to locate and to extract information needed for criminal investigations. Hard drive searches are performed to discover deleted or hidden files using file recovery programs and encryption decryption software. In addition to computers, there is also the collection of relevant information from network servers, databases, smartphones, tablets and other digital devices.

1.2. Motivation

Nowadays, the amount of digital advertisements created is huge and these spread at a rapid pace, which makes it difficult to detect malvertising and prevent its occurrence, both for consumers and publishers. This means that publishers themselves are often unable to directly oversee the process of verification and evaluating of the ad's legality.

It is also very difficult for cybersecurity experts to identify exactly which ad is malicious, as the ads on a webpage are constantly changing. Most malvertising attacks require the user to view the infected ad, which means that not all website visitors will be affected, which makes it even more difficult to detect and reduce fraudulent advertisements (Crowdstrike, 2021).

To prevent cyberattacks and ads appearing adblockers can be used, as uBlock, an extension that works on browsers to prevent ads from being displayed on websites. It uses open databases to prevent banners, pop-ups and other advertisements from being displayed without the user's consent, which makes browsing cleaner (Alves, 2014).

However, the use of adblockers harms marketing campaigns since not only are malicious adverts blocked but all webpage adverts.

How can companies publish their ads without any risk to the users? They will likely need to adopt measures, such as security systems, to implement websites' protection to increase users' safety, specially those surfing without an adblocker. No intruder would be able to insert malware within a website's advertisement and cybercrime would, therefore, be prevented.

Cybersecurity works to implement and maintain a robust information security system in order to defend an organization from cyber-attacks; in case efforts fail and a violation is committed. However, cybersecurity only can be executed if the occurrence of crime patterns is previously studied. Digital forensics, with the help of machine learning algorithms, can identify patterns that help predict cases similar amongst themselves to quickly detect the crime and, consequently, find the intruder. It can detect the crime, aim to understand the source and recover the compromised data (Krakoff, n.d.).

Firstly, it is important to understand how malvertising works to find possible solutions to prevent cybercrime. Secondly, it is crucial to understand how these ads can affect computers and which ads are the most harmful. As seen earlier, in some cases, ads involving Flash can inject malware without the need to click on a malicious link to get infected.

Many websites live off ad revenue, some sites may block or limit access to content if an ad blocker is detected. It is important to determine whether adblockers only block browser's unwanted ads or whether they also protect users from cyber-attacks. uBlock allows users to see which ads have been blocked, thus enabling them to select all the blocked ads in their logs and test whether these contain malware or not (Alves, 2014).

This way, it is possible to study machine learning algorithms and see how harmful an advertisement can be for the consumer.

1.3. Research questions

RQ1 - Which components of the data are related to fraud? Who owns the data and where is it stored? How can the investigator gain access to the data to create detection models? How to prepare data to be able to make good predictions of cyberattacks.

RQ2 - which computer web-oriented advertisement mechanisms and tools exist and that can help preventing digital fraud.

RQ3 - Which model will best detect fraud when AdBlock is not used? The models that are more robust to outliers, (e.g. decision trees and random forests) How to find a distinction between normal and suspicious behavior when downloading, how to recognize patterns in cyber forensics? (e.g. neural networks) How to verify the authorship of an advertising email? (e.g. Bayesian classifiers) Which clusters are needed to isolate advertising anomalies, such as avoidance behavior that might lead to criminal activity. For example, word clustering can be obtained from emails, instant messages, site forms and texts. How to uncover the hidden associations or relations between advertisements - link analysis.

RQ4 - Which technique can best find suspicious behavior and quickly and accurately predict anomalies?

RQ5 - Which rules should be employed to prevent fraud in the future?

1.4. Objectives

The objective of this research is to create a comprehensive framework for the use of ML techniques on the detection of criminal patterns in advertisement and to predict criminal activity connected with advertisement, such as contexts in which crime more frequently happens and when it is likely to happen.

In order to achieve this goal, the following intermediate objectives were defined:

- To study the most relevant techniques and technologies used in machine learning forensics.
- To study the most commonly used computer web-oriented advertisement mechanisms and tools, namely UBlock and Adblock.
- To build a framework that solely blocks the malware advertisement whilst allowing benign advertising still to be visible for the users. In this way companies can promote their new product and captivate new consumers.
- To select a better machine learning model, with a better precision to determine which links are malicious.

- To create a web extension to block websites with malware, that has been detected through the machine learning model. This extension could be used to increase consumers protection against malware attacks.

1.5. Study Relevance and Importance

Years ago, there was not much of an interaction between the company and the customer, but with the evolution of the digital age, the number of internet users has increased significantly, which brought several benefits to companies. Nowadays, companies are increasingly trying to get closer to their customers, and the easiest and fastest way is to achieve that is through technologies. Technological development, and more specifically the internet, has allowed the dissemination of new products or services to be faster and more convenient. As seen before, individuals spend most of their time using the internet, which makes it easier for companies to reach their target audience, strengthening their notoriety and thus increasing their profit (Benetti, 2021).

The company is able to obtain feedback in real-time, to improve customer relationships, to generate sales' opportunities, to reinforce competitive differentials, improve brand awareness, among other advantages.

Online ads allow companies not only to reduce costs but also to segment their market more effectively, reaching solely the audience that really matters to the business. It allows a greatest number of people to be reached, not only in their country but equally abroad. In real-time Feedback allows the company to focus on the strategies that most appeal consumers, immediately outdoing those that are less effective, which allows to increase the number of sales opportunities (Benetti, 2021).

However, although advertising is very important for companies, it has its downsides. Since advertising often involves money, many criminals take advantage of it to commit cybercrimes, as companies pay for their ads to be placed on more sites to reach a higher number of views (Rodrigues, 2016). Criminals take advantage of these advertisements to embed it with fraud. The advertisement which was initially legal, will be placed on a page that is not owned by the company but by a third party, which facilitates intruders to inject malware and commit a cybercrime.

To prevent fraud from occurring, companies need their advertisements to be secure and transparent. For this they need to detect in real time which ones are the malicious advertisements. They equally need to understand the best metric to detect fraud. If any invalid traffic is detected, a software that detects anomalies will automatically report to the advertiser and may even block future visits to the link that contains the malicious ad by the users (Carr, 2020).

Thus, the study of cybercrimes by digital forensics is very important for detecting anomalies, to understand the patterns of how fraud is carried out, and to find solutions, allowing for the development of software that can detect in real time fraud. The software will allow companies to publish ads more securely, preventing fraud

occurrence, which will not only protect their consumers from possible attacks, but will also increase trust between the company and the consumer, strengthening their reputation as a trustworthy and transparent brand, and it will allow the consumer to view ads without worrying about being attacked.

2. Methodology

The methodology used in this thesis is the Design-Science Research (DSR), which seeks to improve human knowledge with the creation of innovative artifacts and the generation of design knowledge. The aim of the research is to create a comprehensive framework for ML techniques usage, it allowing the implementation of the DSR methodology in this dissertation, since the framework will be the artifact.
This methodology is widely used to develop information technology projects, as, through innovative solutions, it allows, the creation of new solutions to relevant design problems whilst solving real world problems (March, Hevner, & Park, 2004).

2.1. Design Science Research (DSR)

A DSR research project aims to extend the limits of human and organizational capabilities by designing new and innovative artifacts represented by constructs, models, methods and instantiations (Gregor & Hevner, 2013). “Artifact” is a pretty broad term when it comes to software development. Most software has multiple artifacts required for its execution (Artifacts, 2020). Some artifacts explain how a software should work, while others allow the program to run (March, Hevner, & Park, 2004).

To carry out research in Information Systems it is necessary to carry out research on the development of theories and research on the development and evaluation of what was created to satisfy the objective. The first refers to behavior and the second to design (Hevner, 2007). Thus, it is possible to consider that the DSR is a research method that aims to generate knowledge on how things can and should be built or designed in an innovative way, to achieve a set of desired goals (March, Hevner, & Park, 2004), allowing the completion of this thesis’ goal, which focuses on ensuring the security of advertisements for users and their companies, so that these will not suffer damage due to possible threats.

The objective is to guarantee the coherence and credibility of the result that is intended to be demonstrated, carrying out a consistent scientific research. The analysis of the fragilities that an advertisement may have will be a crucial point to understand where the anomaly may be, in order to make improvements, avoiding an attack. For this it will be important not to miss any details (March, Hevner, & Park, 2004).

In Figure 1 it is possible to see how the conceptual structure is organized. This structure allows understanding, execution and evaluation of design science research (March, Hevner, & Park, 2004). This structure is divided into three main blocks: the environment, the design and the knowledge base.

The environment defines the problem space in which the phenomena of interest resides. It is composed of people, organizations, and existing or planned technologies. It contains the goals, tasks, problems and opportunities that define the needs as perceived by stakeholders within the organization (March, Hevner, & Park, 2004).

Needs are assessed and evaluated in the context of existing organizational strategies, structure, culture and work processes. They are positioned in relation to existing technology infrastructure, applications, communication architectures and development resources. Together, they define the "research problem" as perceived by the researcher. Framing research activities to meet the real needs of stakeholders ensures the relevance of the research (March, Hevner, & Park, 2004).

The knowledge base provides the raw materials from and through which the DSR is performed. The knowledge base consists of Fundamentals and Methodologies. Past research and results from reference disciplines provide fundamental theories, frameworks, instruments, constructions, models, methods, and instantiations used in the construction phase of a research study. The methodologies provide guidelines used in the assessment phase. Accuracy is achieved by properly applying existing foundations and methodologies (March, Hevner, & Park, 2004).

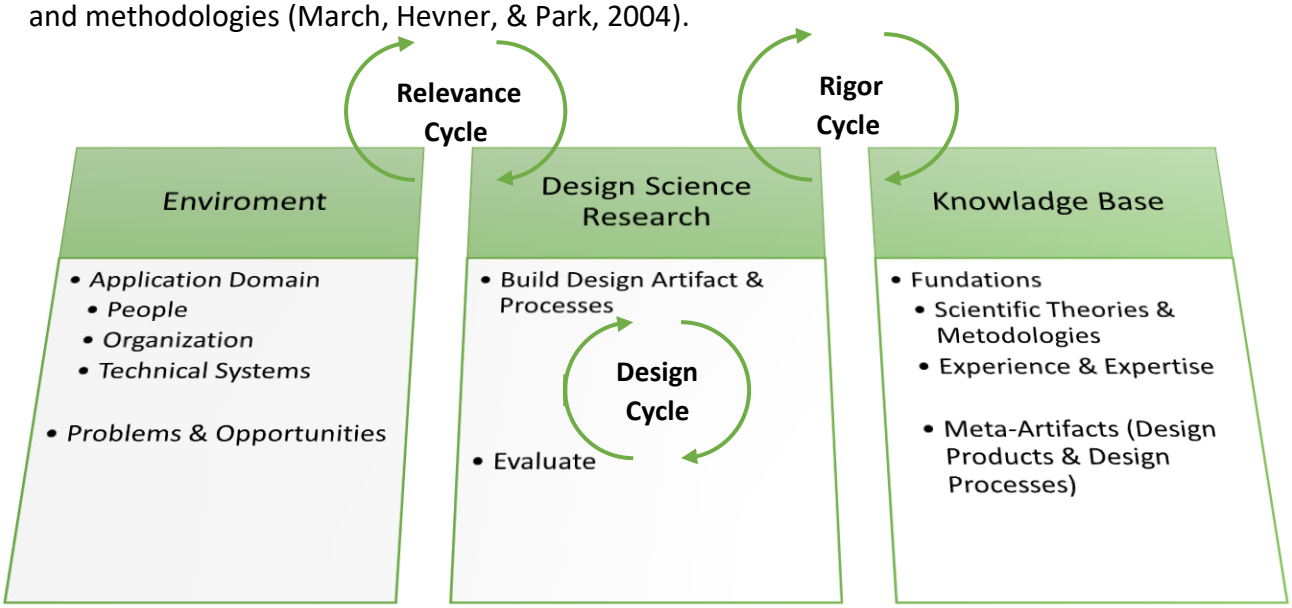


Figure 1: Conceptual Structure
Source: Adapted from (Brocke, Hevner, & Maedche, 2020)

The analysis of the business environment and the derivation of specific needs to be solved build the starting point of a DSR project. However, there are equally situations where the needs have already been studied which can be obtained from existing research. DSR analyzes the (academic) knowledge base as it studies the extent to which design knowledge is already available to solve a problem of interest. Such knowledge can take the form of theories, frameworks, instruments or design artifacts such as constructions, models, methods or instantiations. If knowledge is already available to solve an identified problem, this knowledge can be applied following a “routine project”, which does not constitute DSR. In addition, DSR sets out to create an innovative solution to the problem, which, in most cases, builds on existing parts of a solution and combines, revises, and extends existing design knowledge. Design activities are made up of “build” and “evaluate” activities, typically following several iterations. In the course of a DSR study, a variety of research methods are applied, including those that are well

established in social science research, such as interviews, surveys, literature reviews, or focus groups (March, Hevner, & Park, 2004).

The performance of DSR projects has been based on several process models, such as Nunamaker, Chen, & Purdin (1991), Walls, Widmeyer and El Sawy (1992), Hevner (2007) and Kuchler & Vaishnavi (2008). The most widely referenced model is proposed by Peffers, Tuuanen, Rothenberger, & Chatterjee (2008). For some authors, the DSR process includes six steps: problem identification and motivation, definition of objectives for a solution, design and development, demonstration, evaluation and communication; and four possible entry points: problem-centric initiation, goal-centric solution, design and development-centric initiation, and client/context initiation. As illustrated in the Figure 2, the steps that the DSR research methodology follow are described here:

1. Problem identification and motivation. This activity defines the specific research problem and justifies the value of a solution. Justifying the value of a solution does two things: it motivates the researcher and the research audience to seek the solution, and it helps the audience to appreciate the researcher's understanding of the problem. Resources needed for this activity include knowledge on the state of the problem and the importance of its solution.

2. Objectives of a solution. Infer the objectives of a solution from the definition of the problem. Objectives can be quantitative, for example, terms in which a desirable solution would be better than current ones, or qualitative, for example, where a new artifact is expected to support solutions to problems not addressed so far. Objectives must be rationally inferred from the problem specification. The resources required for this include knowledge of the state of problems and current solutions and their effectiveness, if any.

3. Design and development. An artifact is created. Conceptually, a DSR artifact can be any designed object in which a research contribution is incorporated into the project. This activity includes determining the artifact's desired functionality and architecture and then creating the actual artifact. The resources needed to move from goals to design and development include knowledge of theory that can be used as a solution.

4. Demonstration. Demonstrate the artifact's effectiveness in solving the problem. This may involve its use in experimentation, simulation, a case study, proof or other appropriate activity. Resources needed for the demonstration include effective knowledge on how to use the artifact to solve the problem.

5. Evaluation. The assessment measures how well the artifact supports a solution to the problem. This activity involves comparing the objectives of a solution with the actual results observed from using the artifact in the demonstration. Depending on the nature of the problem site and the artifact, the assessment can take many forms. At the end of this activity, researchers can decide whether to iterate back to step three, to try to improve the artifact's effectiveness, or whether to continue with communication and leave additional improvements for subsequent projects.

6. Communication. Here, all aspects of the problem and the projected artifact are communicated to the relevant stakeholders. Appropriate forms of communication are employed depending on the research objectives and the audience, such as practicing professionals.

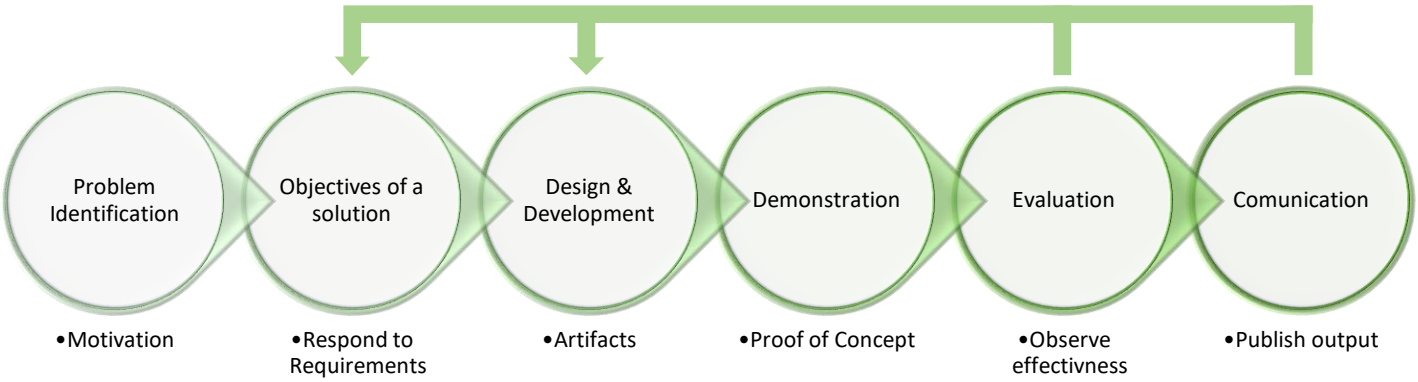


Figure 2: Design Science Research Process Model
Source: Adapted from (Peffers, Tuunanen, Gengler, & Rossi, 2006)

2.2. Research Strategy

The research strategy used in this dissertation presupposes the application of the Design Science Research methodology. The next topics will specify how the DSR was applied in the context of this dissertation and how each step was approached to apply the suggested structure.

1. Problem identification and motivation: The first step refers to the definition of the problem and the objectives defined as part of the project. The chapter 1 of this dissertation identifies the problem that companies face when publishing their advertisements and how these can be used in a cyberattack. Following the identification of the advertisement as a problem, this chapter also aims to explore how much an ad published by advertising companies can be malicious for internet users.

2. Objectives of a solution: The objectives of the dissertation and constructed artifacts are justified in chapter 1 and fully supported by the analysis and investigation carried out in the literature review. The goal of this research is to create a comprehensive framework for the use of ML techniques in detecting patterns of advertising crimes and predicting criminal advertising activity, as well as the different contexts and timings in which crimes are likely to occur. Subsequently, using a web extension, the foreseen process entails blocking URLs that have been determined as malicious, thereby preventing the occurrence of cybercrime.

3. Design and development: This phase of the process consists on creating the artifacts that will be detailed in chapter 4 and will include the development of guidelines based

on the respective lessons learned, and the main conclusions of the work performed. At this stage, the required resource is the theoretical knowledge that will support the investigation.

4. Demonstration: After the design and development phase is complete, it will be necessary to test the resulting model, performing experiments (classification and prediction of threats), and simulate how the model allows to detect anomalies.

5. Evaluation: In the evaluation phase the performance of the ML algorithms will be evaluated. The objectives of a solution will be compared with the actual results observed using the artifact in the demonstration. With the results obtained it will be possible to decide the effectiveness of the artifact and what improvements can be made to best achieve the objectives.

6. Communication: Finally, after completing of the evaluation stage, there should be a reflection on the work developed and the constructed artifacts. In this last step of the framework, we must not only make an overview of the work carried out, but also declare the limitations found during the process, as well as define suggestions for possible future improvements. The overall objective is to share these learnings with professionals in the marketing area, to provide possible contributions to innovate advertising and reduce possible cyber-attacks.

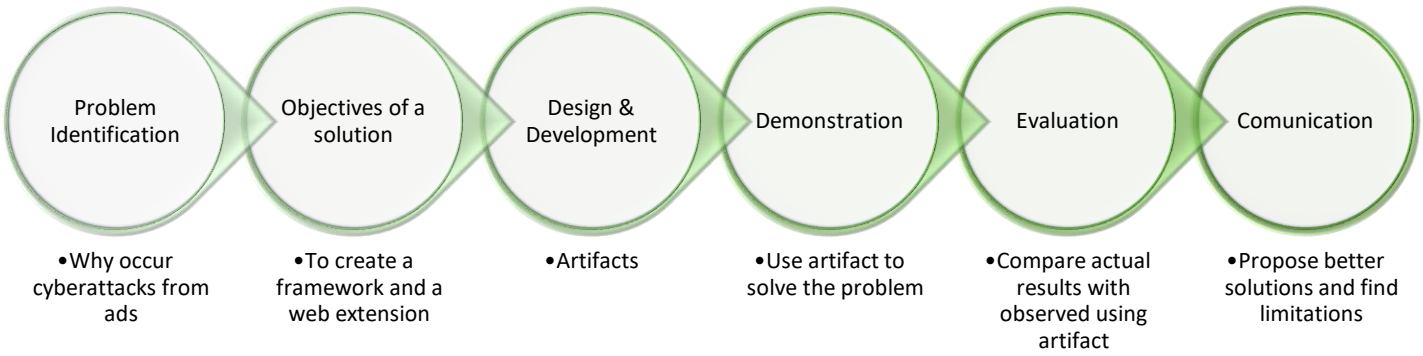


Figure 3: DSR Model

3. Literature Review

In this chapter, the theoretical bases that will be used for the development of this project will be presented. Initially it will be analysed what digital advertising is, what its benefits are and why many internet users choose to use Adblockers. The role of cybersecurity in digital advertising will be investigated. The Literature Review aims to identify which areas of Cybersecurity can be supported by Machine Learning techniques. Therefore, in the literature review, the main concepts related to the areas associated with Cybersecurity and Machine Learning will be investigated.

3.1. Web-oriented advertisement

3.1.1. Context

Online advertising is simply a form of internet promotion, i.e., the process of using the Internet to do business, where marketing manages to attract more consumers. It has a stronger impact on the consumer's mind as it allows companies to advertise in more places than other marketing tools and can take many forms, such as banner ads, email promotions and social media postings (Katke, 2007).

Advertising plays a very important role in dissemination and developing brand awareness. Online advertising is able to create a strong relationship with consumers as they can find out all about the product or service whenever they want. In addition, online advertising is more cost-effective and allows the advertiser have more control over what the ad is, when, where and for how long the ad should be posted, as this information is very relevant to keep the advertisements up to date (Mishra & Mahalik, 2017).

Many companies consider digital marketing the main tool in lead generation. The main advantage of online advertising is the quick delivery of a product/service to the target audience across the world (Chornyy, 2021).

The internet allows to increase sales with online advertising. Since it allowed changing consumer behavior at the time of purchase and their preferences, it also allowed to increase the purchasing power of new products (Kumar & Shah, 2004).

For an advertisement to be successful, the most important thing is not only to be able to sell the product or service, but also to be able to capture the consumer to keep them satisfied and continue to buy the brand's products or services (Mishra & Mahalik, 2017).

The web is a distribution and communication channel that facilitates communication between the brand and the consumer. With online advertising, marketers are able to reach new consumers in a more interactive way and captivate existing audiences. Online advertising is increasingly used as the Internet allows consumers to hold their attention

for longer due to multimedia which exposes content in a more entertaining and exciting way (Mishra & Mahalik, 2017).

Given the development of technologies, digital or virtual business is used more and more with the intention of focusing on online advertising. Marketers have more tools that they can use to communicate with the consumer through the Internet. These tools play a vital role in creating an effective brand. Information and data can be displayed in various forms such as text, images, videos and sound, resulting in a flexible medium (Mishra & Mahalik, 2017).

The advantages of online advertising can be defined as awareness efficiency, localization, contact efficiency, conversion efficiency and retention efficiency. It can be seen that the entertainment capacity of a consumer is positively correlated with the perceived value of advertising (Ducoffe, 1996).

3.1.2. Benefits

Digital marketing offers unique benefits that other means of advertising and promotion do not. Some of these benefits are: cost-effective, focused, builds relationships, easy to adapt and edit, measurable and easy to determine ROI (Standberry, 2019).

Web marketing is cost-effective: Internet ads are cheaper compared to traditional advertising, posting an ad on social media is cheaper than posting an ad on radio, television or print, and digital ads reach a greater number of consumers.

Web marketing is focused: Digital advertising makes it possible to segment the market effectively and efficiently with the evolution of the digital age and it is possible to obtain data from the target audience. With this data companies can reach consumers who will be interested in the product, captivating different age groups around the world. This is one of the great strengths that digital marketing provides, since the location of the consumer will not be a disadvantage for brand promotion.

Web marketing builds relationships: With online advertising, it is possible to create strong relationships between the consumer and the brand, as digital marketing allows reaching the target audience that is interested in the products and services that are being advertised. In this way, it is possible to resolve problematic points more quickly, as it will be easier to build trust between the brand and the interested consumer, avoiding the persuasion of consumers who do not show any interest in the product or service.

Web marketing is easy to adapt and edit: Ads on social media, banners or other online campaigns can be quickly adjusted and revised, and it is possible to edit an ad in real time if an error is found or if any content changes. New prices, additional products and urgent sales can be easily addressed with web marketing initiatives.

Web marketing is measurable and easy to determine ROI: One of the best benefits of web marketing is the ability to quantify results. Often with traditional marketing it is necessary to wait some time to check if the ad works or not, with online advertising, it

is possible to check in real time if the ad we publish has an effect on our target audience, if it is not working as it was planned, it is possible to change quickly and make changes with the new needs of consumers.

In the Figure 4 below, we can see some differences between digital marketing and traditional marketing.

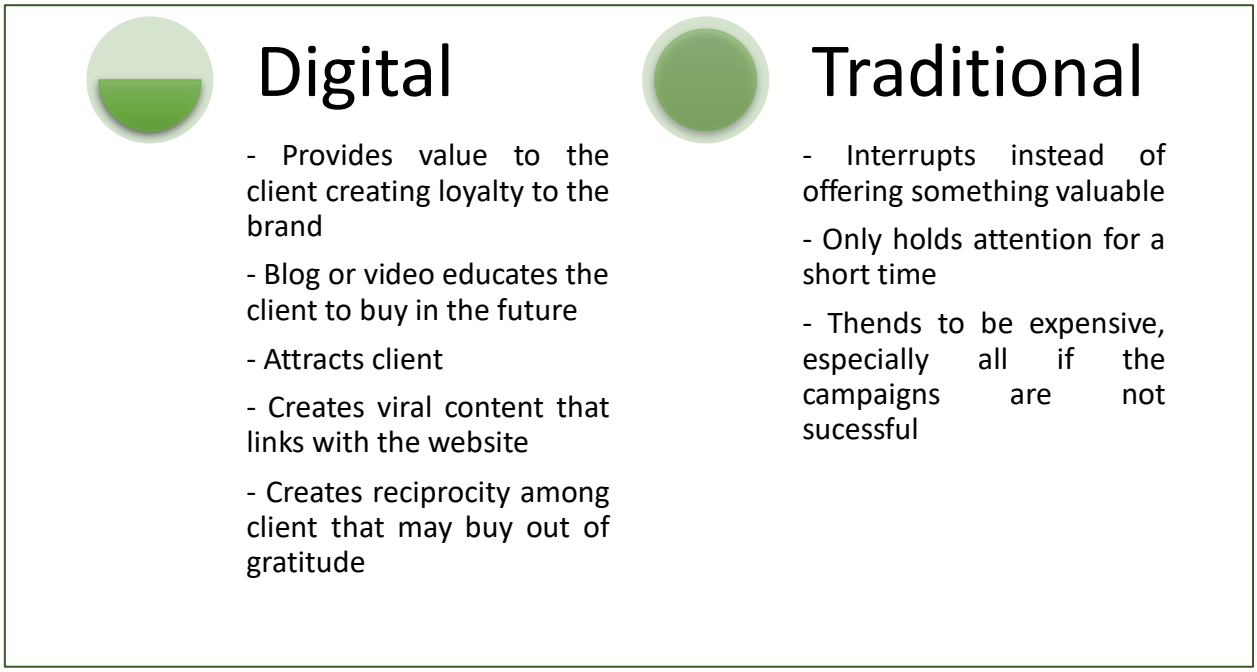


Figure 4: Digital Marketing VS Traditional Marketing

3.1.3. Mechanisms / Types of Online Advertising

Digital marketing has improved a lot over the years, nowadays, there are many options that companies can use to publish their ads and communicate with the consumer. Here are some of the types of online advertising that can be used by marketers to build relationships with their target audience (Standberry, 2019).

Email Marketing: Advertising of new products and services can be done by sending emails to customers. This is one of the most widely used forms of marketing on the web and it is an inexpensive web marketing form. However, it has a downside, consumers receive tons of emails in their inboxes every day, which can be ignored by the customer. In order for this not to happen, it is necessary to define an appealing message, so that the client when receiving this e-mail does not ignore it or throw it in the trash folder.

Social Media Marketing: Publishing on social media is a simple and quick way to publicize new products and services that are being used by many brands. In addition to being simple and reaching a large number of consumers, it is a very cheap means of advertising. However, companies need to constantly captivate the consumer given the high competition.

Content Marketing: Content marketing is exploding as an effective web marketing tool. Companies that post blogs regularly have four times more website traffic than

companies that don't. Consumers use the internet to find out more about products or services, companies that are able to provide useful and quick answers are able to achieve consumer trust and loyalty. A steady stream of strategically distributed high-quality content greatly increases a company's sales.

PPC Advertising (Google Ads): It is the one of the most widespread and popular ad types. Searches are often performed through Google, the most popular one is Google Ads. It is completely safe and legal. In order for an ad to be viewed by a greater number of consumers, companies can place their ads in front of an audience that is already looking for information about their product or service. For this to happen, companies need to place their advertising at the top of the search, before the organic results returned.

Display Ads: Companies can choose to advertise on certain websites that their target audience is likely to visit. The key to making successful display ads is knowing what can attract potential customers and then being able to capture their information once they click on the landing page.

Retargeting: Retargeting is a form of web marketing that can bring consumers back to a company's website, it allows tracking the website users and shows them ads again as a reminder. For example, a person accesses the website. The website drops a cookie which is a trail that will take you back to the website. When customers visit other websites, our retargeting ad is displayed as a banner, reminding them of our company. If they click on it, they'll be taken back to one of our landing pages so that that customer can complete their purchase or find new products of interest. Retargeting is essential for re-engaging potential customers who, for one reason or another, have not completed the purchase.

Online video ads: Digital advertising can be used in shopping ads. This advertising can be placed on google appearing at the time the search is performed. These advertisements allow to view short videos about the proposed products. Ads are also placed in videos, movies, overviews and other things on YouTube. In this way, companies are able to introduce a short video about the product or service, providing important information to the potential customer as they manage to reach a good part of the population. However, these videos may not be to the customer's liking as they may disturb the viewing of the video. One way to overcome this problem could be to place the ad only at the beginning or at the end of the video, movie, etc. (Chornyy, 2021).

Internet location-based ads: Another way to advertise on the internet is by location, if a user looks for something on a map, the application shows similar places that the consumer can visit and plan his visit with other places close to the place of interest. This allows to increase the coffee attendance of these diners (Chornyy, 2021).

3.1.4. Tools

As stated earlier, the most popular forms of online advertising are banners, called “conventional ads”, which consist of placing graphic elements of various shapes and

sizes on a website. These advertising elements redirect the consumer to the advertiser's website, which allows the company to establish its online presence and increase its awareness. The least invasive ads are static banners. Audible announcements are considered aggressive announcements. Many internet users want to avoid any kind of advertisement, as they can harm their searches (Sołtysik-Piorunkiewicz, Strzelecki, & Abramek, 2019).

Many internet users want to avoid any kind of advertising since the advertising can harm their searches. In addition to the advertisement being something that disrupts internet browsing or the use of social networks, it can also track our personal and behavioral data. To avoid interruptions and not completely ruin our Internet experience, it is possible to use browser extensions to slow down or stop this type of data collection. Sometimes a browser extension can cause a website to display text strangely, prevent images from displaying correctly, or remove the little social media buttons that facilitate sharing an article. But the extension will make it harder for malicious parties to track personal data and view what the individual does online (Klosowski, 2021).

On the other hand, there is so-called non-intrusive advertising, in these advertisements the user hardly notices the difference between the advertisements and the organic content, the advertisements are practically camouflaged (Bolina, 2018).

Ad blocking systems rely on the help of crowdsourcing in terms of creating filtering lists. The analysis of crowdsourcing actions shows that there are often false positive errors, i.e. blocking legitimate content and ad publisher attacks against ad blockers (Alrizah, Zhu, Xing, & Wang, 2019). Millions of web users rely on filter lists to protect their privacy and improve their browsing experience. Filter lists are maintained by a small number of contributors, who use a variety of undocumented heuristics to determine which rules to include (Snyder, Vastel, & Livshits, 2017). Further comparisons of online ad blocking lists revealed that blacklists can be reactive or proactive in combating online ad and tracking services (Hashmi, Ikram, & Kaafar, 2019).

Ad blocking is often analysed in terms of digital advertising, preservation of privacy, measurement of ad effect, and advertising fraud (Sołtysik-Piorunkiewicz, Strzelecki, & Abramek, 2019).

Nowadays, internet users are increasingly resistant to different forms of advertising. Users are better prepared to receive promotions of products and services on the Internet. They seem to be more focused on reviewing certain ad criteria, i.e. location, contrast, and ad unit size, than during the early period of Internet development (Sołtysik-Piorunkiewicz, Strzelecki, & Abramek, 2019). Ad forms are still very popular, but they are no longer as efficient as they used to be.

Users express a strong negative feeling about ads and a moderate positive feeling if they can subscribe to a fee-financed, ad-free site (Tudoran, 2018). The adoption of ad

blockers by users is positively influenced by the level of knowledge of their advantageous features (Sołtysik-Piorunkiewicz, Strzelecki, & Abramek, 2019).

There are several different types of web browser plugins that are currently popular and block ads: AdBlock, AdBlock Plus, uBlock and uBlock Origin (Strzelecki, Abramek, & Sołtysik-Piorunkiewicz, 2019).

AdBlock is considered to be one of the best software in terms of ad blocking ability. However, uBlock is considered the best performing plugin, in terms of ad and third-party filtering, and the least privacy tracking (Garimella, Kostakis, & Mathioudakis, 2017).

Below, these two softwares will be analysed.

3.1.4.1. UBlock

uBlock Origin is not just an "ad blocker", it is a broad-spectrum content blocker whose main advantage is efficient use of CPU and memory. It is a free, open-source cross-platform browser extension for content filtering. The main task of the application is to effectively neutralize privacy invasion threats with ease of use. Open source ad blockers are a potentially effective energy-saving technology. Efficiently utilizes CPU and memory. The uBlock Origin extension is available for many of the most popular browsers, including Chrome, Chromium, MS Edge, Opera, Firefox and all Safari versions up to 13 (uBlock, n.d.).

History:

In 2014, founder, original author and lead developer Raymond Hill created the first version of the uBlock extension, whose development began by forking the HTTP Switchboard codebase into uBlock and uMatrix, a standalone blocking extension for advanced users (uBlock, n.d.).

The original uBlock was designed so that the user community can start building and maintaining block lists, and Raymond Hill can simultaneously add new features to the extension and update code quality to meet official release standards. The first version of uBlock was released in 2014 as an exclusive extension for Chrome and Opera browsers, and in late 2015 the extension became available for other browsers under its current name - uBlock Origin (uBlock, n.d.).

The uBlock Origin Firefox version quickly gained popularity, so in December 2016, developer Nick Rolls officially released the version of uBlock Origin for the Microsoft Edge browser (uBlock, n.d.).

The uBlock Origin extension remains the leading open source cross-platform plug-in designed specifically for use across multiple browsers. Since 2021, uBlock Origin has been available for all popular browsers including Chrome, Chromium, Edge, Opera, Firefox and all Safari versions up to 13. Currently, the uBlock Origin project still

fundamentally refuses donations and instead encourages all of its clients, users and supporters to send funds to those who maintain and build the block list (uBlock, n.d.).

Blocking and filtering:

uBlock Origin (or uBlock$_o$) is not an ad blocker; is a general-purpose blocker. uBlock Origin blocks ads through its support for Adblock Plus filter syntax. uBlock Origin extends the syntax and is designed to work with custom rules and filters. In addition, advanced mode allows uBlock Origin to work in default deny mode, which will cause all third-party network requests to be blocked by default unless allowed by the user (Hill, n.d.).

The main objective of uBlock Origin is helping users to neutralise privacy-invasive devices in a simple way, for those users who do not want to use more technical and involved means (Hill, n.d.).

EasyList, EasyPrivacy, Peter Lowe's, Online Malicious URL Blocklist and uBO's own lists are enabled by default when installing uBlock Origin. Several other lists are readily accessible to block trackers, analytics, and more. Host files are also supported (Hill, n.d.).

After installing uBlock Origin, users can easily deselect any of the pre-selected filter lists if they feel that uBlock Origin is blocking excessively (Hill, n.d.).

uBlock Origin includes a growing list of features not available in uBlock, including: A way to help people with color vision impairment (Hill, n.d.).

Performance:

uBlock Origin and user reviews considered the extension to be less intensive than extensions that provide feature sets similar to it, such as Adblock Plus. uBlock Origin researches which style features are needed for an individual web page, rather than relying on a universal style sheet. The extension snaps a snapshot of the filters which the user has activated, helping to speed up the browser's launch speed in comparison to retrieving filters from the cache every time (Hill, n.d.).

3.1.4.2. AdBlock

AdBlock, or ad blocker, is a web browser extension that was originally created in 2002 by Henrik Aasted Sorensen to block unwanted advertisements and advertisements on web pages (Gonçalves, 2019).

With the arrival of AdBlock, the pop-up advertising internet boom slowed dramatically in 2004. The ads that appeared in separate web browser windows were considered the most unpleasant and intrusive features, the original AdBlock was able to hide these ads, but did not prevent the ads from being downloaded (Master, 2018).

In 2006 AdBlock Plus was created by Wladimir Palant, he took the AdBlock code and rewrote the code to make advertisements completely blocked from downloading and

not simply blocked from being displayed. The original code was modified almost completely, and left little of the original until today. Adblock Plus has been the most downloaded and used ad unit extension since its creation in 2006. It is an open source project that blocks annoying ads on the web (Master, 2018).

Wladimir Palant said that the reason behind Adblock Plus was to bring back to internet users by allowing them to select and block annoying ads of their own choices (Master, 2018).

In 2009, Michael Gundlach created AdBlock, one of the most popular browser tools available today. The open source software is designed to give each user full control over what they want to view on their browser (AdBlock, n.d.). This plugin performs blocking of advertisements such as banners, advertisements on YouTube videos, advertisements on Facebook, pop-ups on websites and blogs and any other type of advertisement that is invasive (Gonçalves, 2019).

According to Gonçalves, people choose to use an AdBlock for the following reasons (Gonçalves, 2019):

- **Experience:** be reading an interesting article and being bombarded with ads, a pop-up, a quick video, or all at the same time. This ends the content consumption experience;
- **Security:** Internet ads are not regulated and can be used to distribute malware (a harmful program that can damage devices and use personal information);
- **Economy:** The ads can be expensive if the user is a person who likes to purchase things online.

3.1.5. The future of advertising when using ad blockers

Adblockes are able to solve user's dissatisfaction, however, companies that invested in that advertisement or ad believing they would get some return are affected (Gonçalves, 2019). What about those sites and blogs that distribute free content and use ads to increase reach?

The popularization of AdBlockers has generated losses in companies, which can significantly impact in terms of lost revenue. E-commerces are greatly impacted, as ad blockers act precisely on banners on websites both on the web and on mobile devices. With that, they prevent possible sales. In addition, websites and blogs that distribute free content to help the public learn even more about a particular subject are harmed (Gonçalves, 2019).

This is the reason why advertisers are feeling haunted. Advertising on websites has a high advertising cost. However, users are not being reached as expected, which reduces the Click Through rate. And this is one of the metrics used to know the effectiveness of outreach (Gonçalves, 2019).

The market will need to adapt to these tools and understand how they can work in the advertisers' interest. Everything indicates that when promotions are announced, it will be necessary to do more than answer questions. It will be necessary to satisfy users' needs; thus, the use of ads will become far more strategical. In addition, they can have a much better range, which can cause AdBlockers to lose their real function (Gonçalves, 2019).

3.1.5.1. How Preventing Ad Blocking

To prevent the use of adblockers, some websites prevent viewing the content when it finds an active adblock. Thus, it establishes a law on its website that does not allow the blocking of ads since the content of the website can be changed due to the use of adblokers. Another way that some companies have found to avoid being harmed by adlboking, has been to use software that creates ads that are not detected by ad blocking software filters (Sołtysik-Piorunkiewicz, Strzelecki, & Abramek, 2019).

Nowadays, many programming techniques to detect very effectively mechanisms that block the ads and then prevent access to the website when the ad blocking software is active and configured to block ads (Singh & Potdar, 2009).

Companies that offer content on the Internet understand that the user cannot resort to the simplest method, that is, the total ban on access to the content, as they will search for the content on other free sites. Therefore, the best solution is to deliver advertisements in a way tailored to the expectations of Internet users. A surfer must be able to decide and choose how they want to receive ads. Therefore, it is necessary to create IT solutions that will accommodate the expectations of ad blocking publishers and users (Sołtysik-Piorunkiewicz, Strzelecki, & Abramek, 2019).

Ad-blocking software threatens the revenue of many sites and raises fears about the viability of digital advertising as a whole. (Sołtysik-Piorunkiewicz, Strzelecki, & Abramek, 2019).

Adblockers are popular solutions to increase privacy on the web. By enabling ad blocking software, not only ads, but also many tracking scripts to track user activities, are blocked. Adblockers are often considered web privacy tools that block third-party advertising. They are very effective in reducing third-party tracking (Ajdari, Hoofnagle, Stocksdale, & Good, 2013).

The main reasons for blocking ads using ad blocking software were: security, interruption, speed, lots of ads, privacy, low frequency camp and others (Haddadi, Nithyanand, Khattak, & Javed, 2016).

Comparing the user's gender, women often mention that they are afraid of viruses and malware, and men say that the biggest annoyance is the interference of advertising in the continuous browsing of online content. There is more than one reason to block ads. Users don't care about the ads themselves, but they are bothered by their aggressive form, like a sudden sound or an ad suddenly covering the browsed content, and furthermore, the ad doesn't allow itself to be skipped or closed. Users are most

disturbed by: interruption of advertising for various web content, a large number of ads (over-advertising) and slow page loading (speed) (Sołtysik-Piorunkiewicz, Strzelecki, & Abramek, 2019).

In order for internet users to stop using ad blockers and for agencies to be able to publish their ads, it is necessary that the ads in the first place are safe and ensure consumer confidence, so that they are not afraid to click on an ad from yours interest and being the victim of a cyber-attack. Secondly, advertisements published on pages should not oblige consumers to view their content if they do not wish to do so, for this reason the advertisements must be placed in plain sight but not harming the user's navigation. This will allow advertising companies to have views, earn revenue and protect their customers from cyber-attacks.

To enable consumer security, it is possible to analyze with the help of adblokers which sites are blocked and for what reason, or just because they are advertisements or because they are harmful to users, because there is a possibility of data theft or it is malware. With this data, it is possible, through ML techniques, to identify the causes and develop protection within the sites so that agencies can advertise without being afraid that their ad will be targeted where malware may be hidden.

3.2. Digital Forensics

3.2.1. Context

It is a new branch of forensic science of study and most digital forensic investigation involves providing digital information stored on computers, mobile devices, game consoles and other media storage media for civil and criminal investigation purposes. Digital forensic investigation involves steps involving the conventional process such as Identification, Acquisition, Preservation, Examination and Presentation of tools to the principal investigator, court of law and other parties made by forensic investigators where decisions are about the outcome of an investigation (Satpathy, Mallick, & Pradhan, 2018).

Data collected from various investigations of computer crimes, cyber frauds and crimes, for example tools that facilitate the efficient management, analysis, evaluation, visualization and dissemination of data, preserve the intrinsic value of the data and the original copy of the unmodified data. The magnitude of data generated and shared by various sectors, such as businesses, public administrations, numerous industrial and non-profit sectors and scientific research, social media sites, sensor networks, cyber-physical systems and the Internet of Things, has increased immeasurably (Satpathy, Mallick, & Pradhan, 2018).

Digital expertise comprises four main processes (Satpathy, Mallick, & Pradhan, 2018):

Identification: The first step in a digital forensic investigation is identification, in which an investigator identifies incidents that are important in processing litigation and identifies evidence related to those incidents.

Collection: After identifying it as evidence, an investigator needs to collect evidence from various digital media such as cell phone, hard drive, router, etc.

Organizing: Organizing how efficiently collected evidence leads to the facts of a criminal incident. First, an investigator inspects the data and its characteristics. After that, the investigator interprets and correlates the available data to determine the facts.

Presentation: In the final stage, an investigator prepares an organized report to expose their actions on the case, which must be admitted to court.

3.2.2. Computer forensics

With the large volume of cases related to data leakage in the corporate environment, companies have been able to identify electronic crimes and punish those responsible for fraud through Computer Forensic techniques. The tools used in Forensic Commutation allow the collection and analysis of data in digital media for the investigation and identification of the criminals involved in the crime (Ipog, 2017).

Computer Forensics aims to investigate, through its techniques and methods, ways to solve possible digital crimes, for this, it performs a whole process of analysis, data collection, organization of facts, identification and detection of evidence found in computers and electronic equipment (Silva, 2021).

Regarding the document (files) analysis in Computer Forensics, it aims to analyze and investigate from where, the location and what is the original source of the document, data or information, it should mainly observe whether it has authenticity (Silva, 2021).

Secondly comes the collection of data to be analysed, which aims to identify the main and possible data, which have a higher level of evidence in order that they can be analysed. Following this reasoning, the organisation attempts to determine the best solutions found and classify by order of importance the main evidence in relation to what happened and bring more precise reports so that these can be analysed by forensic specialists (Silva, 2021).

3.3. Cybersecurity

3.3.1. Context

The rise of the digital age has led to the existence of numerous pieces of information that are stored in large amounts in computer data. This data is mostly confidential and information which, if stolen, can have negative consequences for businesses. Not only large companies can become victims of cyber-attacks, but also small businesses can be

targeted by external threats, which puts the future of the organisation at risk. In addition, consumers can suffer a cyber-attack and be targeted with the possible loss of personal information, such as bank accounts, which can lead to negative consequences (Cyber, n.d.).

Nowadays, with the increasing use of the internet, it is increasingly important to be protected to avoid being the victim of a cyber-attack, because most security flaws are intentionally created by malicious people trying to obtain some benefit through the technologies (Teles, 2015).

The number of cyber-attacks is growing more frequent currently and the number of devices connected to each other via the Internet is increasing all the time, which also increases the risk of attacks (Teles, 2015).

Many cybercriminals can access to our information without us realizing it, they still manage to convince us to submit our information willingly and without us realising that we are being deceived. The Internet can be a scary and dangerous place to browse alone, so it is necessary to ensure that the information we access is secure, consistent and reliable (Solms & Niekerk, 2013). Cyber-attacks and crimes can cause devastating financial losses and affect organizations and individuals, implementing a strong cyber security approach is essential for companies to be able to mitigate the loss. It is important to effectively identify multiple cyber incidents and intelligently protect relevant systems against these cyber-attacks (Oliveira, 2021).

To ensure the safety of users while they are browsing the internet, businesses need to have their websites secure to ensure the protection and safety of consumers to avoid being the target of a cyber-attack (Cyber, n.d.).

In this way a concept of cyberspace security was created, commonly known as “cybersecurity”. Cybersecurity aims to keep users and computer systems safe, as it is concerned with ensuring that third parties cannot read or modify messages intended for other recipients (Teles, 2015).

Cyber security encompasses everything that aims to protect organisations and individuals from intentional attacks, breaches and incidents as well as their consequences. It essentially deals with the types of attacks, breaches or incidents that are targeted, sophisticated and difficult to detect or manage. Cybersecurity's primary focus is on advanced persistent threats and the impact that they have on businesses and individuals. However, cybersecurity is not just about protecting cyberspace, but also about protecting what works in cyberspace and any of its assets that may have a direct or indirect relationship with cyberspace (Couto, 2018).

Craigen et al. defined “cybersecurity as a set of tools, practices and guidelines that can be used to protect computer networks, software programs and data from attack, damage or unauthorized access” (Craigen, Diakun-Thibault, & Purse, 2014). In general, cybersecurity is concerned with understanding diverse cybernetics and preparing a

correct defense that preserves several properties defined below (Jang-Jaccard & Nepal, 2014):

- Confidentiality is a property used to prevent access and disclosure of information to unauthorized entities or systems.
- Integrity is a property used to prevent any unauthorized modification or destruction of information.
- Availability is a property used to ensure timely and reliable access to assets and information systems to an authorized entity.

3.3.2. Cybersecurity Risks

There are several areas in Cybersecurity that should be considered as risks, such as (Zúquete, 2018):

Intrusion: An intrusion is any set of actions intended to compromise the integrity, confidentiality or availability of a resource. An intrusion results from the execution of one or more attacks on the systems that manage this resource. These attacks may or may not cause permanent changes to the information stored on these systems. It is a difficult risk to assess, since it does not need to involve exactly one piece of data, however, it grants access to something that is normally denied to the intruder.

Access to reserved or confidential information: Computers store information, therefore, all unauthorized accesses are defined as risks.

Loss or theft of information: It accommodates all situations where information is lost or stolen by unauthorised individuals, and may even pass into your possession.

Impersonation: Occurs when an individual subverts an authentication system, impersonating another person or when a behavior defined for an application undergoes a change. Sometimes it is used as a deception (hiding the true identity of a machine) or appropriation (using someone else's identity).

3.3.3. Cybersecurity Attacks

The risks typically associated with any attack, consider three security factors, such as threats (who is attacking), vulnerabilities (the weaknesses they are attacking), and impacts (what the attack does). A security incident is an act that threatens the confidentiality, integrity or availability of assets and information systems. There are several types of cyber-attacks that occur today and with which Information Systems users must be aware and be careful not to be targeted by them (Sarker, et al., 2020). These are:

Malware: The concept of Malware or malicious software encompasses any and all software that has been altered with the aim of damaging devices, stealing information and taking control, whether at an individual or organizational level. There are several types of Malware, such as backdoors, spyware, trojans, viruses, among others. For example, spyware is a special type of Malware that is installed on the target computer,

with or without the user's permission, and is used to acquire confidential and sensitive information (Pande, 2017).

How malvertising works: The attackers pay for the ads through third-party ad networks. It is the responsibility of the ad network to review the code for malicious content. Once the malicious ad is approved by the network, it is placed on a trusted website. An unsuspecting user then visits this site. Even if the user does not click on the ad, they are redirected to a malicious page, resulting in a drive-by-download (an unintentional download) of information-stealing malware. Unfortunately, because the ad networks are so vast, it is extremely difficult to identify the person responsible for implementing malvertising (Din, 2021).

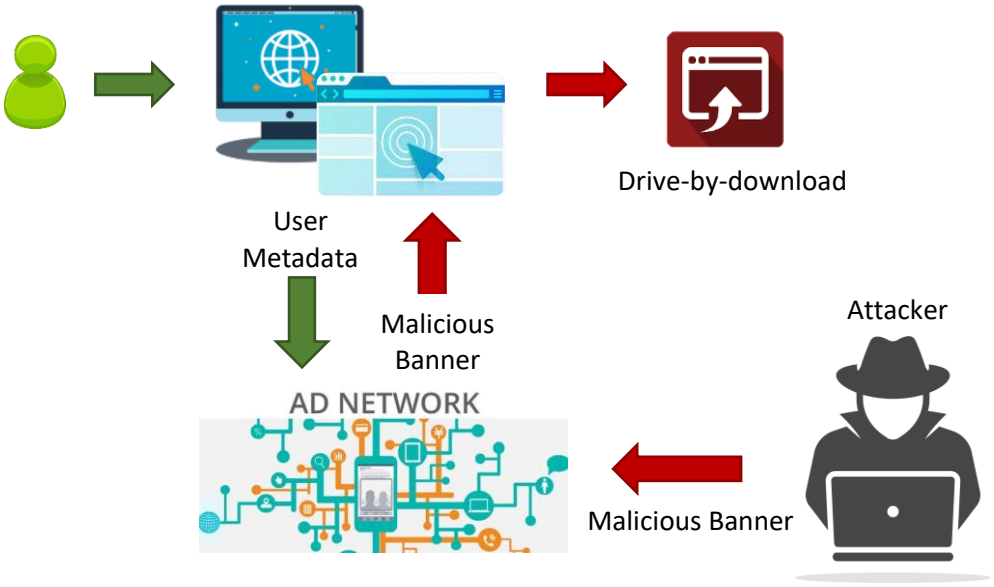


Figure 5: How malvertising works
Source: Adapted from *(Din, 2021)*

Scarewares: Software that at first glance is harmless, however, is quite harmful as it sometimes misleads the user by indicating that a website is reliable, however, it has Malware and infects the machine used to access the website. This software is sometimes sold as a solution to protect the user from attacks, however, its purpose is to steal personal information from the person who purchased it (Landage & Wankhade, 2013).

Denial of Service (DoS) Attacks: A DoS attack is an attack that has the main objective of inactivating a machine or a network, so that it becomes inaccessible to users, this is performed by blocking traffic, namely, filling it ordering or triggering a system failure (Oliveira, 2021).

Phishing: Phishing is a type of cyberattack that uses email as a weapon, the recipient of the message believes in the sender's viability, opens the email, selects the link that is normally provided and the information about the user is obtained without your permission. Can obtain information such as passwords or credit card details (Jagatic, Johnson, Jakobsson, & Menczer, 2007).

Ransomware: This type of cyber-attack restricts access to the computer system or device that is intended to be attacked and tries to demand a ransom from the owner, in order to free up access. Ransomware can reach a machine by receiving an attachment sent by email or via the browser, usually when visiting a page that has already been infected with the virus (Pope, 2016).

SQL Injection: It is a type of cyber-attack that takes advantage of failures in connected systems or systems that interact with databases. The attack is executed through SQL commands, where the attacker inserts a custom SQL statement within a query (Ahmad, Shekhar, & Yadav, 2010).

Cross-site Scripting: It is a defined vulnerability in a computer system, usually present in web applications that trigger malicious attacks by inserting scripts into web pages that are accessed by other users. These scripts allow attackers to escape the control performed during access (Oliveira, 2021).

Credential Stuffing: This type of cyberattack consists of stealing access credentials, usually users and email addresses, as well as the corresponding passwords, these are later used to gain unauthorized access to web applications (Pal, Daniel, Chatterjee, & Ristenpart, 2019).

3.3.4. Cybersecurity Forensics

Cyber security forensics is a branch of cyber forensics that primarily focuses on protecting digital assets as well as responding to a cybersecurity breach. Like any other type of machine learning forensic investigation, with cybersecurity forensic science it is necessary to follow methodical techniques to solve crimes that involve machines attacking other machines (Mariyann, 2018).

Cybersecurity forensic analysis involves searching for specific clues through artifacts that help detect a cyber-crime that has occurred on networks, servers and the internet. To obtain an incident response, it is necessary to carry out forensic investigations that require the preservation, identification, extraction, documentation and interpretation of digital evidence. Incident response is a term used to describe the procedure by which an organization handles a data breach or cyber-attack, including how the organization attempts to manage the consequences of the attack or breach or incident. The ultimate goal is to manage the incident effectively so that damage is limited and recovery time and costs, as well as collateral damage such as brand reputation, are kept to a minimum (Mariyann, 2018).

Cybersecurity expertise is much more than simply discovering and prosecuting cybercrime and hackers. It also serves to ensure that the same attacks do not happen again, and thus guarantee security. Cybersecurity forensic methods consist of multidisciplinary approaches, including the following tasks (Mariyann, 2018):

- Investigation of an incident

- Incident reconstruction
- Digital evidence collection
- Analysis of evidence
- Establish links, associations and reconstructions
- Use evidence to prosecute perpetrators

3.4. Machine Learning Forensics

3.4.1. Machine learning overview

Machine learning is a subset of Artificial Intelligence (AI) that consists of the automatic acquisition of knowledge by machines, without the explicit need to be programmed, to do what is natural for humans and animals. It allows computers to learn from historical data, identify distinct patterns, and make predictions about the data with minimal human intervention (Oliveira, 2021).

In the machine learning model, an algorithm is trained using a set of known input-output data to predict sets of unknown output data. The ML model makes a prediction for the new input data entered, and the model is evaluated for its accuracy. If accuracy is not acceptable the ML algorithm is retrained with more training data. The most important thing for an algorithm to obtain good results in the forecast, it will be necessary to build a ML model that can best perform a good treatment of training data (Mariyann, 2018).

Over the years, the application of ML has become popular in many areas, not only for having cheap and powerful hardware and software applications, but also due to the increasing availability of free and open source software, which allows machine learning to be easily applied (Mariyann, 2018).

Machine Learning methods or techniques are usually classified into 4 categories:

Supervised Learning: Supervised learning is an approach to machine learning defined by the use of labeled data sets. These data sets are designed to train or "supervise" algorithms to accurately classify data or predict results. Using labeled inputs and outputs, the model can measure its accuracy and learn over time. Generally used in familiar environments, with prior knowledge regarding its characteristics. Examples of supervised learning: regression, decision tree, random forest, KNN, ANN, RNN, etc. (Delua, 2021).

Unsupervised Learning: Unlike supervised learning, unsupervised learning is the ability to solve problems using only input data. The goal of the unsupervised machine learning algorithm is to be able to find the structure or relationships between different inputs. An example of these algorithms is clustering (a set of techniques for data mining). Examples of unsupervised learning: Clustering, Association, dimensionality reduction (Haq, et al., 2015).

Semi-Supervised Learning: is a mixture of Supervised and Unsupervised Learning, the data may or may not be qualified. It can be used in methods such as classification, regression and prediction. It is useful when the cost related to qualification is too high to allow for a fully qualified process (Delua, 2021) .

Reinforcement Learning: Reinforcement Learning is a type of Machine Learning to automatically determine the actions appropriate to a given situation, in order to maximize the reward within a specific context. Reinforcement algorithms are not given explicit goals; instead, there is an interaction between the computer and the environment in order to achieve a certain objective, for example, asking a user for a certain classification for an instance that can be part of a set of unidentified instances (Haq, et al., 2015).

In order to obtain the desired goals, and to be able to create a correct machine learning model, below are the steps you can take to achieve consistent results (Mariyann, 2018):

Selecting the Machine Leaning Approach: Before starting any step, it is necessary to define which problem has to be solved, since the final result will depend on the selected algorithm.

Collecting Data: Data can exist in a variety of forms, such as written on paper, text files and spreadsheets or stored in an SQL database. It is necessary to collect the relevant data in an appropriate electronic format to be used for problem analysis.

Exploring and Preparing the Data: Knowing what data the company is dealing with is essential to building an effective solution. The quality of any ML project is based on the quality of the data it uses. This step requires a lot of manual effort as the effective construction of solutions depends on the quality of data preparation.

Training the Model on the Data: This step in the machine learning approach involves selecting an appropriate algorithm. Cleaned data is divided into training and testing datasets. Training data is used to train and develop the model, while test data is used to validate the model.

Evaluating Model Performance: It is very important to assess how well the algorithm learns from experience, this helps to ensure the accuracy and precision of the algorithm.

Improving Model Performance: To improve the performance of the built model, it is necessary to use more advanced strategies, or switch to a different model, supplement with extra data, and do additional preparation work on the data. After the above steps have been completed, if the model appears to perform satisfactorily, it can be deployed for the intended task.

Choosing a Machine Learning Algorithm: The choice of machine learning algorithm depends on the data that is available and the proposed task at hand. It is important to be careful with this process when collecting, exploring, and cleaning up raw input data.

3.4.2. Machine Learning Forensics Strategies

Information technology is constantly changing due to technological advances, bringing new challenges to forensic science. The increase in cyber-attacks on digital systems has made this science very important due to the need to protect the privacy of data, which are now stored in large amounts in different digital systems (Mariyann, 2018).

The digital age allows to collect information from different media and in real time, from security cameras, telephones, email, messages, etc., which can be used to detect cybercrimes. When using the internet, cyber criminals leave a digital trace that allows detecting the occurrence of a cybercrime. This data can be collected and analysed, allowing to model models to anticipate cybercrimes. To succeed in this goal, a behavioral forensic investigator's strategy is to recognize and identify where and how to retrieve, organize, and leverage these historical behaviors for modeling to detect and prevent crimes (Mariyann, 2018).

Machine learning-based forensic science offers enormous potential for dealing with many of the problems that currently exist in cyber forensic science. Machine learning forensic science is the application of software technology that can analyze a large amount of digital data to analyze and identify behavior that has criminal intent and is triggered based on automatic recognition of specific patterns in the underlying data. The process of finding evidence of malicious intent in large amounts of data is complicated and time-consuming. Machine learning speeds up the process, as it allows to automate the analysis process (Mariyann, 2018).

Machine learning forensic science can be broadly classified into three categories, extractive, inductive and deductive of forensic machine learning (Mariyann, 2018):

Extractive: Extractive machine learning forensic analysis is an unsupervised technique that allows to discover hidden associations or relationships between entities buried under large amounts of structured or unstructured data or both to intelligently extract evidential content. Association rules, KNN, link analysis or unattended text mining are some of the machine learning algorithms suitable for extractive evidence analysis.

Inductive: Inductive machine learning forensics is an unsupervised technique and based on the principle of inductive reasoning, the conclusions arrived are implicative by nature and may go beyond what is contained within the forensic dataset. The conclusions arrived at using inductive machine learning are not necessarily true, although may be true. It involves cluster analysis and self-Organizing maps.

Deductive: Deductive forensic analysis is a supervised learning technique that is based on the principle of deductive reasoning, where conclusions are drawn based on information from the forensic dataset. The conclusions reached on the basis of deductive machine learning are necessarily true, which means they must be true. The investigator must strategically plan on what and how criminal behavior will be detected, continually measuring its importance and detection value. This technique involves algorithms such as Decision Trees, Nearest Neighbors K, Artificial Neural Networks,

Convolution Neural Networks, Recursive Neural Networks, Long Term Memory, Support Vector Machines, etc. that require "training" of a model to identify or predict specific behavior such as fraud or intrusion.

Building a correct ML model is very important to get desired results. Data collection and preparation is a set of procedures that helps to make the dataset more suitable for machine learning, this means that it allows to improve the quality of the data allowing to gain an understanding and insight into the criminal's method of operation (Mariyann, 2018).

Articulation: The first step in building a correct model is to have a good understanding of what direction you are going and what are the most valuable data to collect. Select the method that will produce good results and detect a crime.

Dataset Preparation: The investigator is responsible for strategically selecting the correct data inputs and outputs for the forensic system. One of the initial tasks an investigator needs to do when gaining access to datasets is to obtain counts and summary statistics for each field, as well as counts with the number of different values for all variables. This knowledge would allow the investigator to access the necessary data preparation activities. It is common in the data preparation stage before modeling to make several adjustments to the raw data before analysis. Subsequently, the model is built using adjusted, synthetic and prepared data.

Data Consistency: Maintaining a consistent data format is a vital data preparation activity. Aggregating data from different sources, or if the dataset has been manually extracted by different people, it is good to make sure that all variables within a given attribute are written consistently. The input format must be the same across the entire dataset.

Data Reduction: The best models are created from the best data quality, not quantity. It is essential to identify the most relevant variable that influences the decision of a certain behavior that is being modeled with machine learning algorithms.

Missing Values: Missing values significantly reduce the accuracy of model results. To avoid these situations, missing values can be replaced by approximate values or deleting records with missing values can also be a solution, depending on the goals to be achieved.

Data Decompose: Depending on the problem at hand, in some situations it is important to add new data resulting from the data that has already been captured, that is, to decompose them into several parts. This separation can be useful in some analyses, such as breaking down the address into postal code, street, country, etc.

Normalize data: It is important to normalize the data, putting all variables in the data on the same scale, to improve the quality of a dataset, reducing dimensions and avoiding the situation where some of the values overlap others.

Discretize data: Sometimes it can be more effective to turn numeric values to categorical values and vice versa for better forecasting results. This can be achieved, for example, by dividing the entire range of values into multiple groups.

3.4.3. Machine Learning Forensics for Cybersecurity

Today, it's impossible to deploy effective cybersecurity technology without relying heavily on machine learning, and it's also impossible to deploy machine learning effectively without a comprehensive, rich, and complete approach to the underlying data (Perlman, n.d.).

Machine learning has become important to cybersecurity because cybersecurity systems can analyze patterns and learn from them, thereby helping to prevent similar attacks and being able to respond to changes in behaviour. It can help cybersecurity teams be more proactive in preventing threats and responding to active attacks in real time. It can reduce time spent on routine tasks and allow organizations to use their resources more strategically (Perlman, n.d.).

In short, machine learning can make cybersecurity simpler, more proactive, less expensive, and much more effective. But these things can only be achieved if the underlying data supporting machine learning provides a complete picture of the environment (Perlman, n.d.).

The focus on data is critical to machine learning success in cybersecurity as it involves developing patterns and manipulating those patterns with algorithms. To develop patterns, it is necessary to have a lot of rich data from everywhere because the data needs to represent as many potential outcomes from as many possible scenarios as possible, however this data must be of high quality in order to be used in the analysis (Perlman, n.d.).

Thus, according to Giora Engel, the data we collect should contain information relevant to our analysis, not only about cyber-attack threats, but also information about everything that happened. It is necessary to collect as much quality data as possible in order to create correlations and have a complete picture of what is happening. This will allow to build different models, model different aspects of behavior and then use algorithms to make decisions about when to issue alerts, when to act to respond to potential threats, when to build preventive protection. An integrated approach is needed between machine learning and data collection, organisation and structuring (cited in Perlman, n.d.).

The main goal of machine learning for cybersecurity is to find out how intrusions have occurred on the system (computer) and how to protect from future attacks. It involves the task of modeling and simulating computer behaviors and attacks against other computer systems, servers and networks. This approach relies heavily on machine learning-based pattern recognition technology to study digital evidence regarding criminal events and scenes involving illegal burglaries and robberies through analysis of

file systems, mobile devices, log files and digital repositories related to the internet (Mariyann, 2018).

Machine learning models can be created based on different network behaviors, such as sequential patterns that represent highly repetitive activities, abnormal activations, etc., and that can automatically learn new patterns of deviant behavior in the network and block attacks (Mariyann, 2018).

Analyzing cybersecurity data and building the right tools and processes to successfully protect against cybersecurity incidents goes beyond a simple set of functional requirements and knowledge about risks, threats or vulnerabilities. To extract insights or patterns from security incidents effectively, various machine learning techniques such as resource engineering, data grouping, classification and association analysis, or neural network-based deep learning techniques can be used. These learning techniques are able to find anomalies or malicious behavior and patterns based on data from associated security incidents to make an intelligent decision. So, based on the concept of data-driven decision making, aims to focus on cybersecurity data science, where data is collected from relevant cybersecurity sources such as network activity, database activity, application activity or user activity and analytics supplement the latest data-based standards to provide corresponding security solutions (Sarker, et al., 2020).

As seen earlier, ML has different methods, now let's look at how these can be used to solve machine learning tasks and how they are related to cybersecurity tasks (Sarker, et al., 2020):

Supervised learning: Supervised learning is performed when specific targets are defined to be achieved from a given set of inputs, a task-oriented approach. In machine learning, the most popular supervised learning techniques are known as classification and regression methods. These techniques are popular for classifying or predicting the future of a specific security issue. For example, to predict denial of service attacks (yes/ no) or to identify different classes of network attacks such as scanning and spoofing, classification techniques can be used in the cybersecurity domain. Regression techniques are useful for predicting the continuous or numeric value, for example, total phishing attacks over a period of time or predicting network packet parameters. Regression analyzes can also be used to detect the causes of cybercrime and other types of fraud. Random Forest learning that generates several decision trees allows solving a particular security task.

Unsupervised learning: In unsupervised learning problems, the main task is to find patterns, structures or knowledge in unlabeled data, data-based approach. In the area of cyber security, cyber-attacks like malware remain hidden in a number of ways, including changing behavior dynamically and autonomously to evade detection. Grouping techniques can help uncover the hidden patterns and structures of datasets. Likewise, in identifying anomalies, policy violations, detecting and eliminating noisy instances in data, grouping techniques can be helpful. Learning association rules can prevent cyber-attacks.

Semi-Supervised Learning: Semi-supervised learning can be described as a hybridization of supervised and unsupervised techniques, as it works with tagged and untagged data. In the area of cybersecurity, it can be useful, when it requires labeling data automatically, without human intervention, to improve the performance of cybersecurity models.

Reinforcement Learning: Reinforcement techniques are another type of machine learning that characterizes an agent by creating their own learning experiences through direct interaction with the environment, that is, an environment-oriented approach, where the environment is typically formulated as a process of Markov decision and makes decisions based on a reward function. Another technique, as Neural Networks are essentially a part of Deep Learning, which in turn is a subset of Machine Learning, let's look at how this technique can be used in cyber security (Sarker, et al., 2020).

Neural networks and deep learning: Deep learning is a part of machine learning in the field of artificial intelligence, which is a computational model inspired by the biological neural networks of the human brain. Artificial Neural Network (ANN) is used in deep learning and the most popular neural network algorithm is backpropagation. It performs learning in a multi-layered progressive feed neural network that consists of an input layer, one or more hidden layers, and an output layer. The main difference between deep learning and classic machine learning is its performance in increasing the amount of safety data. Typically, deep learning algorithms perform well when data volumes are large, while machine learning algorithms perform comparatively better on data sets. In terms of extracting resources to build models, deep learning reduces the effort of designing a resource extractor for each problem than classical machine learning techniques. In addition to these features, deep learning takes longer to train an algorithm than a machine learning algorithm, however, the testing time is just the opposite. Thus, deep learning depends more on high-performance machines with GPUs than classical machine processing algorithms. The most popular deep neural network learning models include multilayer perceptron (MLP), convolutional neural network (CNN), recurrent neural network (RNN) or long-short-term memory network (LSTM) Researchers use these deep learning techniques for different purposes such as network intrusion detection, malware traffic detection and classification, etc. in the field of cybersecurity.

3.5. Anomaly detection methodologies and algorithms

3.5.1. Methodologies and algorithms

Nowadays, there is a wide variety of ML techniques, which can be used for the right purposes and it is important to understand what the underlying algorithms are learning and to understand how the various algorithms are learning the data patterns to be used in the most efficiently way. These algorithms can be applied to almost any data problem in a variety of domains (Ayyadevara, 2018):

Linear Regression: Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a prediction value of target based on independent variables. It is mostly used to find out the relationship between the variables and the prediction. Different regression models differ based on - the number of independent variables used and the type of relationship between the dependent and independent variables, which they are considering (Mohit, 2018).

Logistic Regression: Logistic regression is essentially a supervised classification algorithm. In a classification problem, the target (or output) variable can be given only discrete values for a specified set of features (or inputs). Like linear regression, it assumes that the data follow a linear function, logistic regression models the data using the sigmoid function. Logistic regression is only considered as a classification technique when a decision threshold exists. The definition of the threshold value is an extremely important aspect of logistic regression and depends on the classification problem itself. The decision of the threshold value is greatly affected by the precision and recall values. For this reason, it is tried to approximate them to 1, but this rarely happens (GeeksforGeeks, 2021).

Decision Tree: In the classification process, a decision tree (DT) can be used as a statistical model. This algorithm classifies data into classes and generates a tree structure. A DT algorithm sorts data in a dataset through a query structure, going from the bottom up from the root to the leaf that represents a class. In classification, the attribute that plays the main role is represented by the root and the class is represented by the leaf (Han, Kamber, & Pei, 2012).

Random Forest: Random forest is a data construct applied to machine learning that develops a large number of random decision trees that analyze sets of variables. This type of algorithm helps improve the ways technologies analyze complex data. In general, decision trees are popular for machine learning tasks. In a random forest, random sets of decision trees are built to isolate data mining knowledge more carefully, with different matrices of variables applied (Techopedia, 2019).

Extra Trees: Extra Trees is an ensemble machine learning algorithm which combines the predictions of multiple decision trees. It is closely related to the random forest algorithm that is widely used. It can frequently achieve good or better performance than the random forest algorithm, despite the fact that it uses a simpler algorithm to construct the decision trees used as members of the ensemble. It operates by creating a huge number of unpruned decision trees from the training data set. The predictions are averaged over the decision trees in case of regression or using majority voting in case of classification (Brownlee, How to Develop an Extra Trees Ensemble with Python, 2020).

SVM: Algorithm that tries to learn by training over a data set to correctly make a prediction and perform a generalisation over the remaining data. SVM is often used to make predictions about two classes, to make a binary classification. SVMs can be categorized as supervised learning models along with related algorithms that are used for pattern recognition, data analysis, regression analysis, and classification. The SVM

algorithm begins training on the known data, building a model that will be able to determine to which categories the new data will belong (Jordan, Kleinberg, & Scholkopf, 2008).

Naive Bayes: The Naïve Bayes classification can be defined as a probabilistic classifier that is derived from the application of Bayes' theorem. In other words, it is an equation in statistical quantities that defines the conditional probability relationship. In high dimensional datasets, Naïve Bayes classification can be very useful since it is a fast and simple classification algorithm, and it is also a baseline for the classification problem since it is a fast and dirty algorithm based on naïve assumption about the data. There are different naive Bayes classifiers (VanderPlas, 2017), such as:

a. Gaussian Naïve Bayes: the easiest way to understand a naïve Bayes classifier is through this algorithm. This classifier works on the assumption that for each label, the data is drawn from a simple Gaussian distribution.
b. Multinomial Naïve Bayes: in this type of naïve Bayes classifier, a simple multinomial distribution generates the assumed characteristics. Out of several categories, the multinomial distribution determines the probability of counting the observation.

KNN: The k-nearest neighbor algorithm can be defined as a non-parametric method, used for regression and classification. In both cases k is the input which refers to the nearest training example that lives in the feature space and if the classification or regression is used it will determine the output of KNN (Mohammed, Khan, & Bashier, 2017):

a. In KNN is a classification process, class membership will be the output. If k = 1, the object is assigned to the most common class where the value of k is positive and usually has a small value.
b. In KNN regression, the property value for the object would be the output since this value is the mean k-NN value.

K-Means: K-means clustering is a simple unsupervised learning algorithm used to solve clustering problems. K-means clustering is a method used for cluster analysis, especially in data mining and statistics. It follows a simple procedure for sorting a given dataset into a series of clusters, defined by the letter "k", which is fixed in advance. The clusters are then placed as points and all observations or data points are linked to the nearest cluster, calculated, adjusted and then the process is restarted using the new adjustments until the desired output is obtained. (Techopedia, n.d.).

Dimensionality Reduction Algorithms: Dimensionality reduction is a series of machine learning and statistical techniques to reduce the number of random variables to be considered. It involves selection and extraction of resources. Reducing dimensionality makes data analysis much effortless and faster for machine learning algorithms without foreign variables to process, making machine learning algorithms faster and simpler. Dimension reduction attempts to reduce the number of random variables in the data. Dimensionality reduction techniques are divided into two main categories: feature

selection and extraction. Feature selection techniques find a smaller subset of a multidimensional dataset to create a data model. The main strategies for the feature set are filter, wrapper (using a predictive model), and embedded, which accomplish feature selection when building a model. Resource extraction involves transforming high-dimensional data into smaller spaces. The methods include principal component analysis, core PCA, graph-based core PCA, linear discriminant analysis, and generalised discriminant analysis (Techopedia, 2018).

Gradient Boosting algorithms: The gradient increase algorithm is one of the most powerful algorithms in the field of machine learning. Errors in machine learning algorithms are widely classified into two categories, respectively, bias error and variance error. Gradient augmentation is one of the augmentation algorithms that is used to minimize the model's bias error. Unlike the AdaBoost algorithm, the basis estimator in the Gradient Boost algorithm cannot be selected by us, since the base estimator for the Gradient Boost algorithm is fixed. It is possible, as in the AdaBoost model, to adjust the number of estimators, if this is not mentioned, the default value will be 100. The gradient-increasing algorithm can be used to predict not only the continuous target variable (such as a regressor), but also the categorical target variable (such as a classifier). When used as a regressor, the cost function is the mean squared error (MSE) and when used as a classifier, the cost function is the log loss (Tarbani, 2021).

ANN: A neural network or an artificial neural network is one of the machine learning algorithms derived from the model or system that exists in the human brain. Made up of millions of neurons, the human brain uses electrical and chemical signals to communicate and then process them. Special structures, known as synapses, connect these neurons, allowing signals to pass through. A neural network, as one of the machine learning algorithms, simulates the behaviour of 'neurons' in the biological system, it has the ability to recognise patterns, besides being used in machine learning, because it has a group of interconnected 'neurons' that act on the input to provide an output value, consisting of three layers (Mohammed & Varol, 2020).

- Input layer: this layer aims to receive the input values of the explicative variable, in the majority of cases the number of input nodes is equal to the number of the explicative variable.
- Hidden layer: One or more layers can be formed in the hidden layer, and this occurs when the actual processing is done through a system of weighted links.
- Output layer: All the hidden layers connect to the output layer, which generates an output value based on the prediction of the response variables. In the classification problem, the output layer is usually represented by a single node.

LSTM: Short-term long-term memory units or blocks (LSTM) are part of a recurrent neural network structure. Recurrent neural networks are made to use certain types of artificial memory processes that may help these artificial intelligence programs to emulate human thinking more effectively. The recurrent neural network uses long-term to short-term memory blocks in order to provide context for how the program takes inputs and generates outputs. The short-term long memory block is a complex unit with

several components such as weighted inputs, activation functions, inputs from previous blocks and eventual outputs. The unit is called a short-term to long-term memory block because the program utilizes a structure based on short-term memory processes to generate long-term memory. These systems are often used in natural language processing. The recurrent neural network uses the long blocks of short-term memory to take a specific word or phoneme and evaluate it in the context of others in a chain, where the memory can be useful for classifying and categorising these types of inputs. In general, LSTM is an accepted and common concept in pioneering recurrent neural networks (Techopedia, b.d.).

Stacking: Stacking or Stacked Generalization is an ensemble machine learning algorithm. This uses a meta-learning algorithm to learning the optimal way to combine the predictions from two or more baseline machine learning algorithms. The benefit of stacking is that it can take advantage of the abilities of a range of well performing models in a classification or regression task and make predictions that perform even better than any other model in the ensemble (Brownlee, Stacking Ensemble Machine Learning With Python, 2020).

Voting Classifier: Voting is an ensemble machine learning algorithm. On regression, a voting set involves making a prediction that is the average of multiple other regression models. For classification, a hard-voting set involves adding up the votes for crisp class labels from other models and predicting the class with the highest number of votes. A soft voting set involves summing the predicted probabilities of class labels and predicting the class label with the highest sum probability (Brownlee, How to Develop Voting Ensembles With Python, 2020).

3.5.2. Machine Learning algorithms to detect anomalies

Machine learning forensics has the ability to recognise criminal patterns and predict criminal behaviour, this includes the ability to predict where and when crimes might happen. In order to make this type of digital forensic analysis occur, a framework needs to be able to capture and analyse servers, on the Internet or over a wireless connection, and many other types of data for link association, visualisation, segmentation and clustering of criminal activities (Mena, 2011).There are many techniques and technologies used in machine learning forensics (Mariyann, 2018), some of these are:

Association rule mining: Association rule mining is an unsupervised machine learning method based on detecting frequent item sets from the attributes of items in the dataset and revealing the relationships between those items. The association rules created by the A priori algorithm is used to extract the relationships between the characteristics of criminal records. The relationships between the attributes of different criminal records make it possible to establish relationships between new and old incidents. Thus, this method can be used to predict the unknown attributes of possible future cases by analyzing previous cases. It allows profiling of user behaviour and identifying irregularities in huge machine-generated log files, which can help locate evidence that could be crucial to a cyber investigation.

Link analysis: Link analysis allows discovering valuable information through data visualization in order to perform better analysis, particularly in the context of links. Link analysis often gives investigators the ability to discover patterns of association, allowing them to shape an immediate visual picture of communications and help them understand the relationship between the people involved in a case. Example: webpage links or relationships between people or any other entities. The relationship between multiple identities will determine how these entities are linked to each other. The investigative method of link analysis enables examiners to develop critical clues during the investigation, which is most crucial. New technologies allow to detect critical communications through texts, calls, emails and social media applications, accomplish something that with more traditional forms of forensic investigations would never be possible.

Text analytics: Text mining can be used for in-depth patent analysis, blogs, reports, emails, surveys, applications and other documents that would require a lot of time to run manually. Text analytics provides investigators with an automated solution for organising main concepts to identify textual evidence. Text analysis can use information retrieval (IR) and information extraction (IE) as well as natural language processing (NLP) techniques to organize and prioritize documents on any subject. These text analytics techniques can be used by investigators to gain new insights into unstructured content data sources in their legacy and operational systems.

Clustering: Machine learning forensic clustering can be used to create clusters based on human behavior and words, such as in text analytics. It is a type of exploratory analysis performed to identify anomalies in data, to avoid behavior that indicates some criminal activity, such as the suspicious use of a computer port for a network attack or a questionable financial transaction that could indicate money laundering. This method allows grouping the data by the number of websites visited, aiming to verify which websites are most frequently visited by the internet user. Clustering can also be applied to text content such as phone call lists, emails, SMS, chats, etc. for detecting anomalies.

Self-Organizing Maps: Self-Organizing Maps (SOM) are another type of inductive modeling technique used to review interesting patterns in forensic data. It allows the mapping of elevated dimensional data onto a two-dimensional map. SOMs are used to aid investigators in drawing a graphical snapshot of large forensic data, enabling them to make better decisions about where to focus to conduct cyber forensic analysis on a large dataset. By doing this, the examiner can conduct the forensic analysis process more efficiently and effectively. Maps generated by a SOM application create visualizations that allow finding information of interest quickly and efficiently. Comparing component maps with each other allows to identify correlations between them. In forensic cyber science the SOM technique can be used to identify correlations (associations) in data, discovering and sorting data into groups based on similarity (classification), visually locating and presenting groups of latent facts (clustering), and discovering patterns in data that can lead to useful predictions (prediction).

Decision Tree: The decision tree algorithm can be used to model criminal behavior in order to develop models for specific types of crimes and criminals. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. It divides a dataset into smaller and smaller subsets, while at the same time an associated decision tree is incrementally developed. The end result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and a leaf node represents a ranking or decision. The top decision node in a tree that matches the best predictor called the root node. Decision trees can handle both categorical and numerical data. The decision tree machine learning algorithm can be applied to detect criminal activities in email.

4. Framework for the detection of advertisement criminal patterns

After studying how cybersecurity forensics can help prevent digital crime, focusing on digital advertisements that nowadays can be seen very frequently on different digital platforms, web pages or social networks. It has also been studied that with machine learning techniques it is possible to make predictions to detect if a link will contain malware or not, which will allow to guarantee a safety for Internet users to browse safely.
The purpose of this thesis is to analyse through machine learning techniques, especially supervised learning, since it is aimed to demonstrate with certainty which links will contain malware and then block them. In order to achieve this goal, this chapter will present a comprehensive framework with the use of ML techniques to detect criminal patterns in advertising and predict criminal activity through the patterns to ensure safety and prevent a crime happening again.

4.1. Assumptions

4.1.1. Online Advertising

As seen previously in the literature review, nowadays the dissemination of products is done especially digitally, through the internet. This allows companies to reach their potential customers faster (Chornyy, 2021). Since this method of advertising allows to reduce costs, reaching a greater number of consumers and obtaining feedback in real time, that allows to interact with customers in the most appropriate way to captivate and retain their attention for the advertised product (Mishra & Mahalik, 2017).

Thus, through effective and fast communication via the internet, companies are able to create strong relationships with their consumers and increase their notoriety, maintaining consumer satisfaction, in order to keep them buying their products (Mishra & Mahalik, 2017).

Online advertising allows to obtain several benefits over traditional advertising, such as cost-effective, focused, builds relationships, easy to adapt and edit, measurable and easy to determine ROI (Standberry, 2019). Furthermore, with online advertising it is possible to communicate with the client not only through the newspaper, as in traditional marketing, but through emails, social media, PPC Advertising, Retargeting, video ads, Internet location-based ads, etc. (Standberry, 2019).

However, companies cannot guarantee their consumers that their advertisements are 100% safe, as many criminals can take advantage of these to place malware to attack the consumer. With the increasing use of the internet and the increase in the number

of devices connected to each other via the internet, criminals are increasingly taking advantage of technology to attack users by stealing their personal data or damaging their electronic devices with viruses (Teles, 2015).

Many users choose to use adblockers, not only because some advertisements make it difficult to view the content, because they are too noisy and impossible to close, but also because they are protected against malware attacks (Jang-Jaccard & Nepal, 2014).

Thus, when using tools such as Adblock or Ublock, they can select the sites on which they want to block ads, as well as select the ads on a site they want to block, once these tools allow them to view only the advertisements that they are interested in and block those that disturb them.

The use of these tools by consumers is harmful to companies because they lose revenue when their products are no longer seen by potential customers, resulting in a decrease in the number of buyers. To prevent their ads from being blocked, many companies choose to put on their website the adblockers detector that prevents them from viewing the website's content if it is active. In order for consumers to have access to the content, they will have to turn off adblockers. However, the website has to allow security to its users so that they are not victims of malware and so that they have a good satisfaction when visiting the site, it will only need to have adverts that do not disrupt consumers' browsing.

To provide security on their websites, companies need to analyse the adverts they place on their pages, checking that they do not have any harmful malware in their link. This is possible due to the emergence of digital forensic science that analyzes similar cases of cyberattacks, analyzing how digital information is stored on computers, mobile devices, game consoles and other media storage media for civil and criminal investigation purposes (Satpathy, Mallick, & Pradhan, 2018).

Not forgetting that the companies themselves can suffer cyber-attacks since more and more data is stored in databases that can be the target of theft by criminals who want to obtain information about the company, its products and also about its customers.

It is important to effectively identify multiple previously seen and unseen cyber incidents and intelligently protect relevant systems from these cyber-attacks (Oliveira, 2021).

4.1.2. Computer Forensics

Computer Forensics aims to investigate, through its techniques and methods, ways to solve possible digital crimes, performs an entire process of analysis, data collection, organization of facts, identification and detection of evidence found in computers and electronic equipment (Silva, 2021).

Computer Forensics allows identifying malware, while cybersecurity allows them to be re-identified in order to protect both companies and their customers, keeping users and computer systems safe, as it is concerned with ensuring that third parties cannot read or modify messages destined for other recipients (Teles, 2015).

As seen previously, cybersecurity allows to prepare a correct defense that preserves several properties, such as Confidentiality, Integrity and Availability (Jang-Jaccard & Nepal, 2014).

Cybersecurity forensic analysis allows to look for specific clues through artifacts that help detect a cybercrime that has occurred. For this it is necessary to carry out forensic investigations that require the preservation, identification, extraction, documentation and interpretation of digital evidence (Mariyann, 2018).

The identification of malware allows companies to effectively manage the damage that cyberattacks can cause, reducing recovery costs and maintaining its reputation as a brand, as it will reduce to a minimum the cyberattacks that could be suffered by your customers (Mariyann, 2018).

Based on what was studied in the literature review on advertising, digital forensics and cybersecurity, it is possible to define that machine learning techniques can help companies to define whether advertising can be malicious or benign, which allows them to protect their websites allowing its consumers a safe navigation.

4.1.3. Anomaly Detection

Anomaly detection is a tool to identify unusual or interesting occurrences in the data. However, it is important to analyze how anomalies are detected from a domain / business perspective before removing them (Thakur, 2021).

These anomalies can point to unusual network traffic, discover a faulty sensor, detect ecosystem disturbances, or simply identify data for cleaning prior to analysis (Johnson, 2020).

According to Bram Steenwinckel: “Anomaly detection (AD) systems are either built manually by experts who set limits on the data or built automatically, learning from available data through machine learning (ML).” (Johnson, 2020)

Building an anomaly detection system manually requires mastery knowledge, furthermore it is difficult to make predictions. To detect digital fraud, detecting anomalies manually is not a good solution since the data changes over time, that is, new frauds appear every day. When the system fails, builders need to go back and manually add other security methods (Johnson, 2020).

Manually building anomaly detection is not favorable, once it is necessary to go back and manually add other security methods when the system crashes, as the system can adapt to new anomalies (Johnson, 2020).

In this way, machine learning allows to adapt to the ecosystem in real time, allows to deal with large data sets and works better.

Supervised Learning

Popular ML Algorithms for Structured Data:

- SVM
- KNN
- Bayesian networks
- DT
- Random Forests

Unsupervised Learning

In unstructured data, the main objective is to create clusters from the data and then find the few groups that don't belong. In fact, all anomaly detection algorithms are some form of approximate density estimation.

Popular ML algorithms for unstructured data are:

- SOM
- K-means
- Clustering
- Association Rules

Table 1 presents a summary of the main techniques that could be used in the framework for malware detection, for the framework, will be selected, the techniques that allow to detect in the best way the malware related to advertising.

Machine Learning Technics	Purpose
SVM	• Features selection, intrusion detection and classification • To build intrusion detection system • To build network intrusion detection systems • To classify various attacks (DoS, Probe, U2R, R2L)
KNN	• Network intrusion detection system • To reduce the false alarm rate • To build intrusion detection system • Anomaly intrusion detection system
ANN	• To build network intrusion detection systems
Clustering	• To build intrusion detection system
K-means	• To build intrusion detection system
Naive Bayes	• To build an intrusion detection system for multi-class classification
Decision Trees	• To detect the malicious code's behavior information by running malicious code on the virtual machine and analyze the behavior information for intrusion detection • Anomaly intrusion detection system

	• To solve the problem of small disjunct in the decision tree-based intrusion detection system
Random Forests	• To build network intrusion detection systems
Association Rules	• To build network intrusion detection systems
Supervised Learning	• For malware detection and analysis
Semi-supervised Learning Adabost	• For network anomaly detection
Deep Learning (CNN, GAN, Multilayer perceptions, …)	• To build anomaly intrusion detection system and attack classification • Malware traffic classification system • Malicious activities and intrusion detection system
Deep Learning - LSTM & GRU	• Temporal analysis

Table 1: Machine Learning Technics Purpose
Source: Adapted from (Sarker, et al., 2020)

4.2. Framework

The proposed framework is a framework that allows the detection of advertising criminal patterns and predicts advertising criminal activity through machine learning techniques.

The application of this framework will allow users to view only those advertisements that will be determined to be benign, safe and malware-free. In this way companies will be able to promote their products in the best way possible, not only protecting the customer but also their data from possible cyber-attacks.

Accordingly, the framework is composed by six sequential steps. As follow: identifying the needs of companies to advertise their products/services, collecting important data via the website URL, selecting the model that best predicts malware, implementing the model and predicting whether the ad has malware, deciding, blocking the ad that has malware, and then providing a secure malware-free web page to the customer.

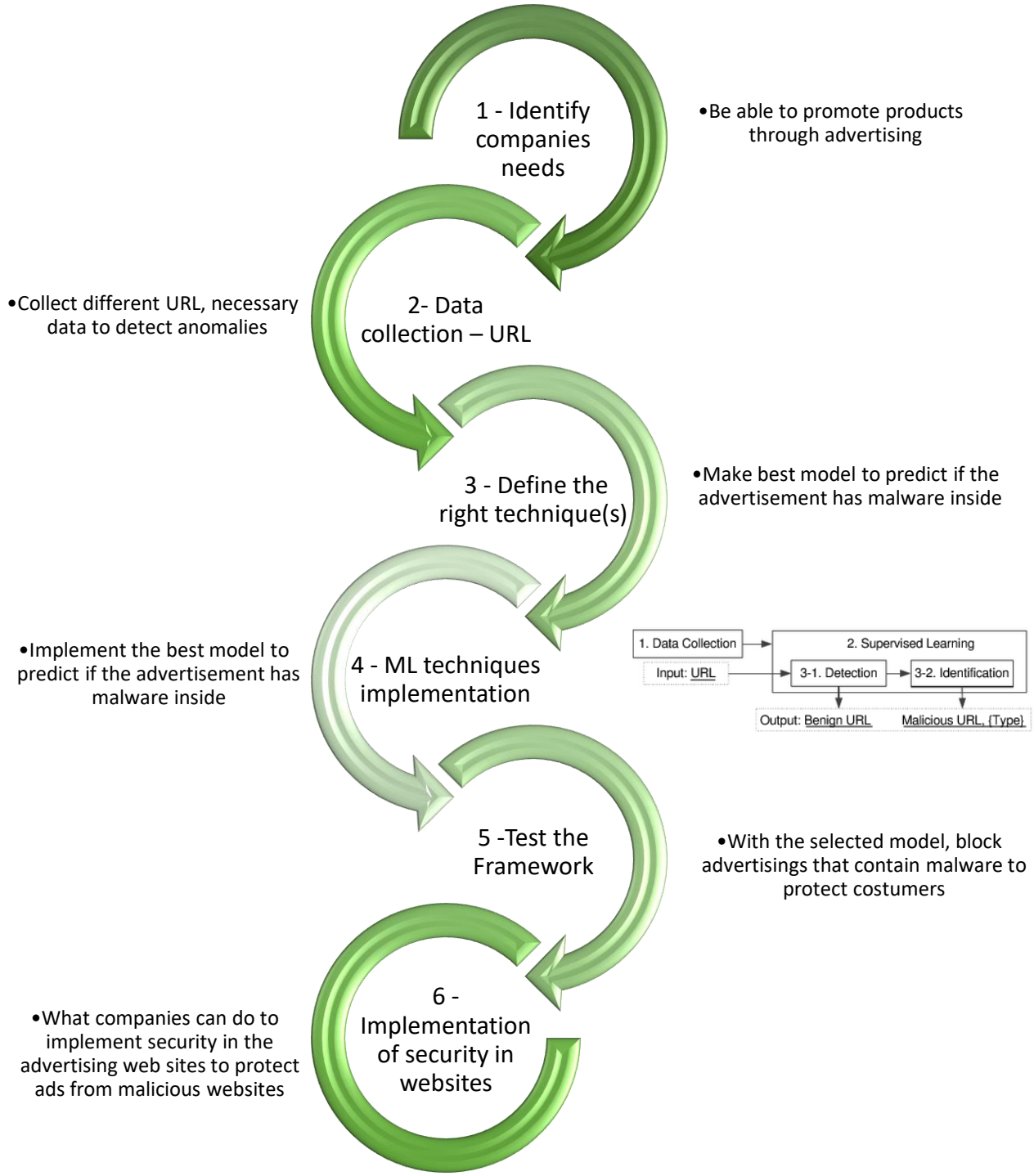


Figure 6: Framework Components

In the next section each framework step is explained in detail.

4.2.1. Identify companies needs

The implementation of this framework consists, first of all, to enable companies to guarantee the safety of their consumers when they visit their website. However, it should not forget that the principal revenue of companies comes from the sale of products or the provision of services. In order for companies to get more customers,

they need to advertise their products more effectively. Nowadays, as seen in previous chapters, the most effective way of spreading advertising is through ads on social media or other web pages visited by consumers. For an advertisement to be successful, the customer must see it, however, as seen before, with the development of technology, consumers are increasingly attacked by cyber criminals, forcing individuals to use ad blockers. In this way, this framework will allow customers to not be afraid to remain with active advertisements on web pages, as the framework will detect malicious advertisements and block them, leaving only safe advertisements. Thus, companies will be able to get potential customers to view their advertisements and be interested in their products, which will allow them to run their business, make a profit and ensure consumer safety.

4.2.2. Data collection

Universal Resource Locator:

URLs, also known as Universal Resource Locator as the name implies are used to locate a particular resource on the Internet, are also known as web addresses. A URL is formed by the resource access protocol, by the location of the server to be accessed, which can be in the form of the domain or in the form of the IP address and the path where the resource is located (Ferreira, 2019).

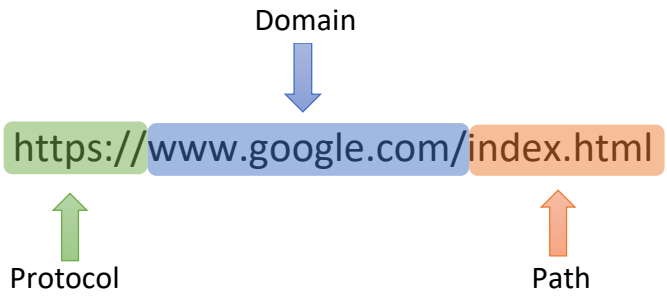


Figure 7: Parts that constitute an URL

When a URL directs the browser to a file that can be opened, such as images or PDFs, the browser displays the content without having to download the file, but many other file types require a download. Because of all the complexity and diversity of functions, URLs can also be made to do harm and attack the user (Ferreira, 2019).

Attacks using URL:

As seen before, cybercriminals use ads to place malware inside these links, through phishing attacks, the main goal of the attack is to trick the user into providing their data, usually login data – the attack consists in making the target click on a link that takes them to a page similar to the one the target intended, but when they enter, the attackers get their login data; or download-driven – which consists of unintentionally downloading malicious code, this attack does not require the user to click on something or do

something. These are the attacks that more and more affect our computer and that through these attacks the cyber criminals are able to access our computer without our authorization and this way steal our personal data such as bank accounts, e-mail passwords, etc.

As this malware is formed from a URL that directs to another site, it is possible to analyse this URL and detect whether or not it contains malware inside, i.e. through the link to which it is directed.

Data collection:

Since the URL itself does not provide relevant information for ML analysis, it is necessary to decompose it to obtain important data to perform the analysis and detect whether the URL is malicious or not.

This way, some URLs were selected from Kaggle, containing 503 benign URLs and 502 malign URLs. Through these URLs 66 features were created, with their help it will be possible to find what malicious URLs have in common and what distinguishes them from benign URL, so therefore it will be possible to detect which URLs need to be blocked to prevent users from becoming victims of a cyber-attack. The features that were created can be found in Annex I.

The extracted features from the URLs can be divided in three groups, below are the characteristics of each group:

URL String Characteristics (Lexical Features): Resources derived from the URL string itself. The motivation for including lexical features is based on the simple fact that malicious URLs look different from benign URLs, so that statistical properties can be extracted that quantify the differences in these appearances. For example, length of the URL string, number of digits, number of parameters in its query part, if the URL is encoded, etc.

Page Content Characteristics (Content Features): Resources extracted from the page's URL. They are obtained from the HTML code downloaded from the web page and the javascript (example in Annex II) content of the web page. These features capture the structure of the web page and the content embedded in it. This will include information about script tags, embedded objects, executables, hidden elements, etc. The logic behind including these content-based features is to capture the characteristics of the page content found on compromised pages, for example the presence of injected content or pages designed to contain malicious code, for example the presence of certain script tags or suspicious HTML elements.

URL Domain Characteristics (Host-Based Features): Domain characteristics of domain URLs. This includes whois information (example in Annex III) and shodan information (example in Annex IV). This set of features allows to capture certain characteristics of 'who', 'where', 'when' and 'how a website is hosted'. They provide information about the web page host, e.g. country of registration, domain name properties, open ports, named servers, connection speed, lifetime from registration, etc. The motivation behind

including these parameters is that there is a difference in website deployment tactics, the longevity of existence, and the reputation for malicious and benign sites. A classic example is malicious sites hiding registration information, the average number of days between an update to websites.

- To extract meaningful host-based features, it needs two libraries in python written to:

    - Interact with domain registration information (python's whois Library). This library provides information about the domain name associated with the host IP. For example, when it was registered, who registered it, when the registration expires, etc.
    - Access host information on the web (python shodan API). This library provides information about the host, i.e. the machine where the site is hosted (if available). Shodan provides an open API to access information such as open ports and operating systems.

After some preprocessing, it was analysed that some features are correlated, therefore some of these features are not used in the model. Also, it is analysed (Annex V) that some features don't have characteristics what help to distinguish benign URL from malign URL, for this reason these features are not included to train the model.

Below a brief description of the selected features can be seen:

Entroty: Entropy is the randomness collected by an operating system or application for use in cryptography or other uses that require random data. A lack of entropy can have a negative impact on performance and security. Benign URLs may have slightly higher entropy than malicious strings. (entroty_html, entroty_script)

Digits: Total number of digits in URL string. After extract this feature, and when we analyses it, we can conclude that benign strings also record a higher number of digits counts with a higher number of encoded characters.

urlLength: Total number of characters in URL string. Malicious URLs are generally shorter in length than benign URLs. Malicious URL strings are shorter in length than benign strings.

Path: Malicious URLs have generally shorter path than benign URLs.

numSubDomains: We noted that, frequently, malicious web pages refer to the domains serving malware with-out specifying a subdomain. This feature keeps track of whether a subdomain is present in the URL.

is_encoded: URL-encoded using a '%' character and a two character hew value corresponding to their UTF-8 character. Benign URL is frequently encoded that malign URL.

num_encoded_char: Number of '%' characters in the URL. Benign url have higher number of encoded characters.

number_of_periods: Number of '.' characters in the URL. Malign url have higher number of '.' characters in the URL.

Malicious URL sometimes have keywords, as 'login', 'client', 'server', 'admin', which may relate to keywords attackers use when trying to spoof a legitimate page or keywords that relate to popular nomenclature of security settings on a website that a hacker will try to manipulate. For example, the keyword 'admin' in a URL usually indicates an authentication page for a site administrative user. Some features are created to analyses these keywords. (has_client, has_admin, has_server, has_login)

get_tld: Number of known malicious patterns. This feature counts the number of occurrences of specific patterns commonly found in drive-by-download campaigns. The pattern list is compiled and updated by a human analyst. We currently identify only one of such patterns: the presence of a meta tag that causes the refresh of the page, pointing it to index.php?spl=, as this is very common in pages redirecting to exploit servers.

get_html: Structure of web pages, javascript, it is important to extract for analyses and extract other characteristics.

get_pq: A query string is a part of a uniform resource locator (URL) that assigns values to specified parameters. A query string commonly includes fields added to a base URL by a Web browser or other client application, for example as part of an HTML form.

len_html: Total number of characters in URL's HTML page. Malicious URL HTML are shorter in length than benign HTML.

number_script: Scripts are lists of commands executed by certain programs or scripting engines. They are usually text documents with instructions written using a scripting language. They are used to generate Web pages and to automate computer processes. Also, benign URL have more scripts than the malicious URL.

n_sentences: Total number of sentences on page as separated by '.' excluding tags. Benign HTML have more sentence than the malicious HTML.

n_punctuations: Total number of punctuations in the page. Benign HTML have more sentence than the malicious HTML.

distinct_tokens: Total number of distinct words separated by ' '. Benign HTML have more words than the malicious HTML.

n_capitalizations: Total number of upper-case characters in the page content. Benign HTML have more capitalizations than the malicious HTML.

n_html_tags: HTML tags are like keywords which defines that how web browser will format and display the content. Total number of HTML tags in page. Benign HTML have more tags than the malicious HTML.

n_iframe: Number of “iframe” strings. This feature counts how many strings containing “iframe” are present in a script. This feature is motivated by the fact that malicious scripts often inject several iframes into a web page, and, if the script is not obfuscated, it is possible to identify when the script modifies the DOM to inject an iframe element.

n_elements: Number of included URLs. This feature counts the number of elements which, being not inline, are included specifying their source location. Elements such as ‘*script’, ‘iframe’, ‘frame’, ‘embed’, ‘form’, ‘object’* are considered in computing this feature, because they can be used to include external content in a web page. The ‘img’ elements and other elements are not considered, as they cannot be used to include any executable code.

n_objects, n_embeds, n_hyperlinks, n_images, n_whitespace, n_elements, n_eval_functions: Number of suspicious tag strings. Similarly, to the previous feature, this feature counts the number of times that certain tag names appear inside strings declared in JavaScript code. In fact, instead of injecting iframes, sometimes malicious scripts write other scripts or objects inside the page. This feature counts the appearance of ‘*script’, ‘object’, ‘embed’, and ‘frame’* inside JavaScript strings. The goal of this group of feature extraction is to look for ‘suspicious’ content on the page. Also some popular script functions used in cross-site scripting attacks are *escape(), eval(), link(), unescape(), exec() and search()*.

n_double_documents: Presence of double documents. This feature indicates whether a web page contains two or more ‘*head’, ‘title’ or ‘body’* elements. This is not allowed by the HTML specification, but can be seen in certain malicious web pages as a side-effect of the com-promise of a web site.

url_creation_year: Registration year. This feature examines the registration year for the host name (domain), if it is available via the Whois service. Registration dates are commonly used to distinguish between benign and malicious domains, since most of the command-and-control and exploit servers reside on domains whose registration date is recent and/or whose expiration date is in the near future. This is because attackers often buy domain names for short time frames, since they expect that those names will be blocked quickly.

url_age: Number of days since registration date. Malicious URLs are more recent than the benign URL.

url_intended_life_span: Number of days from registration to expiration. This feature examines the Time to Live (TTL). Shorter TTLs are usually associated with services that are likely to be moved to another IP address in the near future.

url_life_remaining: Number of days left until expiration. Shorter time-to-live domains are likely to move to other IP addresses or domains in the future. As a result, as malicious URLs have shorter lifetimes than benign URLs, they also tend to have fewer total website updates throughout their lifetime than benign URLs.

url_connection_speed: In the analise benign URL get the connection faster than malign URL.

4.2.3. Define the right technique(s)

As seen in the literature review, there are several ML techniques. To define which of these techniques can detect more accurately which advertisement may contain malware, some benign and malign URLs were selected to perform tests. These URLs were taken from Kaggle, containing in total 1000 URLs. The URLs were analysed and through pre-processing the data mentioned in the previous point was removed in order to be able to train the models and find which of these techniques can determine more accurately the malware in the URL.

Firstly, will be analysed some of the advantages and disadvantages of each model, and the results obtained after testing each model will be discussed, in the Table 2:

Technique	Advantages and Disadvantages	Results
Linear Regression	• Simple to implement and easier to interpret the output coefficients. • It is often quite prone to noise and overfitting. • Diversely, linear regression assumes a linear relationship between dependent and independent variables. That means it assumes that there is a straight-line relationship between them. It assumes independence between attributes. • Linear regression is quite sensitive to outliers.	• Very sensitive to noise outliers, which decreases the score. • Very linear which makes it difficult to analyse with different features.
Logistic Regression	• Performs well when the dataset is linearly separable. • Is less prone to over-fitting but it can overfit in high dimensional datasets. • Is easier to implement, interpret and very efficient to train. • Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.	• All the data used are discrete, so the model predict well and have a high score. • Despite being linear it presents good results. • Do not overfit as the dataset is not too big.

	• Logistic Regression can only be used to predict discrete functions. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set. This restriction itself is problematic, as it is prohibitive to the prediction of continuous data.	
Decision Tree	• No feature scaling required • Decision Tree can automatically handle missing values • Is usually robust to outliers and can handle them automatically • Overfitting: This is the main problem of the Decision Tree. • Little bit of noise can make it unstable which leads to wrong predictions.	• Sometimes predict always that the URL is malware, we can think that the model is overfitting.
SVM	• SVM works relatively well when there is a clear margin of separation between classes. • SVM is more effective in high dimensional spaces. • SVM is effective in cases where the number of dimensions is greater than the number of samples. • SVM algorithm is not suitable for large data sets. • SVM does not perform very well when the data set has more noise i.e. target classes are overlapping. • In cases where the number of features for each data point exceeds the number of training data samples, the SVM will underperform.	• The dataset contains some noise which makes the analysis difficult and predicts many false positives which is not very favorable in anomaly detection. • Since the target classes overlap, it is difficult to make correct predictions.
Naive Bayes	• Naive Bayes is suitable for solving multi-class prediction problems • Naive Bayes is better suited for categorical input variables than numerical variables. • Naive Bayes assumes that all predictors (or features) are independent, rarely happening in real life. This limits the applicability of this algorithm in real-world use cases.	• Its estimations can be wrong in some cases, due to interpreting each resource independently, which leads to wrong predictions.

KNN	• Very easy to understand when there are few predictor variables. • Useful for building models that involve non-standard data types, such as text. • Have large storage requirements. • Sensitive to the choice of the similarity function that is used to compare instances. • Computationally-expensive technique. • Lack a principled way to choose k.	• Despite being difficult to select k, the model manages to present a high score in the predictions. • It is favorable because it has large storage requirements.
Random Forest	• It reduces overfitting in decision trees and helps to improve the accuracy. • It is flexible to both classification and regression problems. • It works well with both categorical and continuous values. • It automates missing values present in the data. • Normalizing of data is not required as it uses a rule-based approach. • It requires much computational power as well as resources as it builds numerous trees to combine their outputs. • It also requires much time for training as it combines a lot of decision trees to determine the class.	• Despite detecting classes with high accuracy, the model requires a lot of computational power, which causes it to take longer to perform predictions.
Extra Trees	• Show same results as Random Forest but it is faster. • Nodes are split based on random splits among a random subset of the features selected at every node.	• It presents a high score, and similar results to the random forest, but unlike the Random Forest, the Extra Trees are faster in making their predictions.
ANN	• A neural network can perform tasks that a linear program cannot. • The neural network needs training to operate. • Is quite robust to noise in the training data • The architecture of a neural network is different from the	• Although it has a high score, and is noise resistant, it requires a lot of processing time to run the neural networks.

	architecture of microprocessors therefore needs to be emulated. • Requires high processing time for large neural network.	
Gradient Boosting algorithms	• Often provides predictive accuracy that cannot be trumped. • Lots of flexibility - can optimize on different loss functions and provides several hyper parameter tuning options that make the function fit very flexible. • No data pre-processing required - often works great with categorical and numerical values as is. • Handles missing data - imputation not required. • Gradient Boosting Models will continue improving to minimize all errors. This can overemphasize outliers and cause overfitting. • Computationally expensive, usually requires many trees which can consume a lot of time and memory.	• Presents a high score, the model predicts more false positives, which is not favorable for anomaly detection. • It takes some time in the execution, which is not beneficial for a fast detection of anomalies.
AdaBoost	• Less prone to overfitting as the input parameters are not jointly optimized. • The accuracy of weak classifiers can be improved by using Adaboost. • It needs a quality dataset. • Noisy data and outliers have to be avoided before adopting an Adaboost algorithm.	• Presents a high score since the dataset has favorable data for the analysis. • It predicts few false positives, which allows for more accurate detection of malware.
Bagging Classifier	• Bagging offers the advantage of allowing many weak learners to combine efforts to outdo a single strong learner. It also helps in the reduction of variance, hence eliminating the overfitting of models in the procedure. • The resultant model can experience lots of bias when the proper procedure is ignored. Despite bagging being highly	• Base estimator – KNN • Since KNN is considered weak because it is hard to set k, if k is set incorrectly, bagging can improve the model by combining the efforts. • Since KNN has large storage requirements it is a good model for determining anomalies.

	accurate, it can be computationally expensive, which may discourage its use in certain instances.	
Stacking	• Produces improvements in model performance. • Reduces variance and generates a more robust model by combining predictions from multiple models. • It can potentially take considerably longer to train than simpler models and will require more memory.	• Always selects the best model, getting the same score. This makes it better to implement only the model in question because of the long-time stacking takes to train.
Voting Classifier	• Every machine learning classification model has its own advantages and disadvantages. By aggregating many models, we can overcome the disadvantages of each model to generalize the classification model.	• Models: Logistic Regression, KNN with Bagging, Extra Trees and AdaBoost • As several models can correctly predict the URL classes, these models were selected since it is possible to overcome the disadvantages that each model has, combat overfitting and perform the analysis in a short period of time. • Voting is able to group these models, predicting with more accuracy the URL classes.

Table 2: Supervised techniques

Next, some unsupervised learning techniques were analysed, Table 3, with the following results:

Technique	Advantages and Disadvantages	Results
K-Means	• If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls. • K-Means produce tighter clusters than hierarchical clustering, especially if the clusters are globular. • Difficult to predict K-Value. • With global cluster, it didn't work well.	• Difficult to predict k value. • Predict labels is difficult because this model divides the dataset not in two clusters, so we cannot predict exactly 2 classes.

	• Different initial partitions can result in different final clusters. • It does not work well with clusters (in the original data) of Different size and Different density.	
Self-Organizing Maps	• Data mapping is easily interpreted. • Capable of organizing large, complex data sets. • Difficult to determine what input weights to use. • Mapping can result in divided clusters. • Requires that nearby behave similarly.	Predict labels is difficult because this model divides the dataset not in two clusters, so we cannot predict exactly 2 classes.

Table 3: Unsupervised Techniques

It was verified that according to these techniques, the best model would be to create 3 different clusters, and not only 2 clusters, one with the benign URLs and another with the malicious URLs. Another decision that can be reached is that some ads that are benign can be considered as malware due to some characteristics. Blocking a benign ad is not harmful to the user, however, not blocking a malware ad could seriously affect the user. As it is not possible to guarantee which ads are benign or malign as it is unsupervised leaning, it was decided not to use these techniques since the accuracy is low and no risk should be taken in leaving a malign ad active.

4.2.4. ML techniques implementation

The Voting Classifier model was created with the following models Logistic Regression, KNN with Bagging, Extra Trees, and AdaBoost, due to providing better precision.

Unsupervised learning techniques were not selected because they only aggregate the data and it is not possible to determine with any certainty if that group contains malware, since it was determined with the tests performed that with this data it would be preferable to create 3 or 4 clusters instead of only 2 (to determine benign or malign URLs).

With this, one can draw a conclusion, some URLs, despite being malignant, may not be as dangerous as others, however, for our security is preferable to block these URLs than run the risk of being attacked by malware. This way it runs mistake type II, having more false negatives than false positives, since, to be safe it is preferable to block a benign ad than to let a malign ad escape.

4.2.5. Test the Framework

After deciding which template to use, it will be used in the framework to detect whether an advertisement contains malware or not. If it is defined that the URL contains

malware, it will be blocked, and both the content and the link to another page will not be visible to the user browsing the web page selected by him.

By implementing this framework, the companies can guarantee the security for their costumers, since the cybercriminals won't be able to take advantage of the ads to attack the users. Thus, unlike ad blockers that block all ads, by blocking only malicious ads, companies can continue to advertise their products and get new customers who show interest in the product or service after viewing the ad.

4.2.6. Replication - Implementation of security in websites

The problem that ad blockers face is blocking all ads, starting with blacklists, on which not only malware but also all other types of ads have been defined.

Each time a malware is detected, it is blacklisted so that when it appears, it does not need to be analysed again, allowing faster access to the website. To combat these malicious URLs, the framework must detect them quickly and effectively, hence if the URL has already been defined as malicious by the framework, it is important to add it to blacklists in order to be remembered as malicious and to no longer appear on the webpage. Blacklists are databases where the data of URLs that have already been confirmed as malicious are stored and new URLs detected as malicious are added to the list over time. Whenever a new URL is visited, a database query is performed. If the URL is blacklisted, it is considered malicious and then a warning will be generated; otherwise it is considered benign (Ferreira, 2019).

Although it seems to be a secure and effective method, the previous lists are slow because they cannot keep up with the growing number of URLs, which means that these databases can never have all the malicious URLs that exist because new ones are created every day and new ways to bypass the blacklists are created (Ferreira, 2019).

Analysing this method with blacklists for malicious ads only, it was detected that this method is not useful since every day new malicious URLs appear and the lists become heavy not being able to support all the URLs.

For companies to be able to have their websites safe for their customers to use without being victims of cybernetic attacks, it is necessary to find an effective method to detect malicious URLs, such as seeking a solution in machine learning algorithms (Sahoo, Liu, & Steven, 2019).

For malicious ads to be detected more accurately it is necessary to expand the database on which the ML model will be based, because for the ML model to show the desired results it is important to have a large dataset and high-quality data. This will be possible if data on URLs recognised as malware is fed into the database so that the model can be retrained and allow malware to be detected more quickly and accurately.

4.3. Demonstration

An application was created that, after analysis from the ML model, blocks URLs where it was predicted that could contain malware inside. This application has the intention of not blocking all ads, but only the ads that are considered malware, as defined above.

Many websites allow companies to place adverts on their websites so that users can view new products and if they show an interest, be able to visit the company's website.

Some websites, which contain advertisements, were selected to verify the functioning of the application and the detection of malware.

From the selected website, with the help of Ublock, the blocked ads were selected, as this application blocks all ads, whether benign or not. The intention will be to check which of these are truly malignant and block only these. Thus, through Ublock, data can be obtained for future analysis.

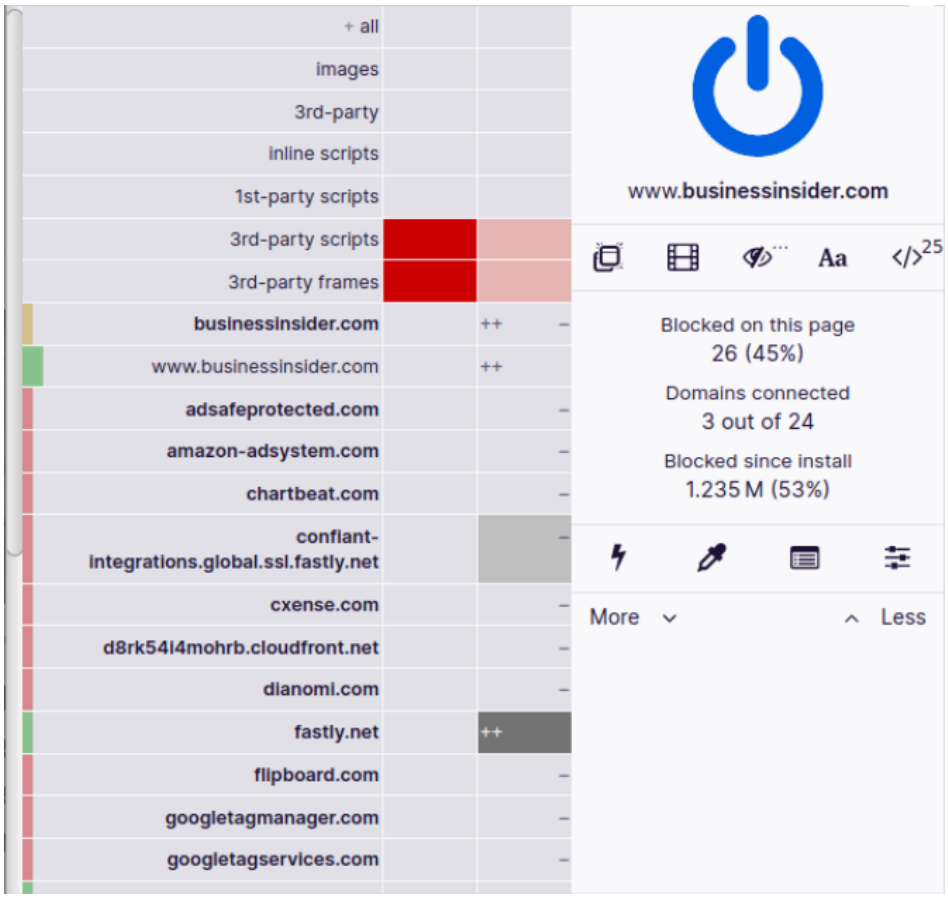


Figure 8: URLs that AdBlock block (with red line)

To train the model, it was used data, available in Kaggle, of URLs already detected as malign and benign. Thus, from these URLs, features were extracted in order to be able to perform the model training. As stated earlier, the features that can accomplish well the distinction of malign URLs from benign URLs have been selected, however, these features are difficult to create from the URLs since malign URLs most of the time have

their script blocked, which makes an analysis more difficult, and therefore it is difficult to detect URLs that are truly malign.

Thus, to ensure that malicious URLs do not appear on the page the user visits, it was defined that it would be more important to get fewer false positives than false negatives.

Several models were trained, and the model was selected as seen in the previous section, the one that can get fewer false positives, and that can train in a short period of time.

After selecting the URLs that were blocked by Ublock, the necessary features were created to perform the predictions. Once the predictions were obtained, the malicious URLs were selected, and only these will be blocked when the user accesses the site. Thus, these URLs are placed in the application so that it blocks the URLs that are malicious. The application has been created as a Google Chrome extension.

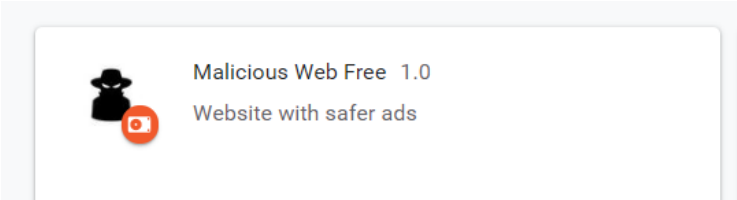


Figure 9: Program that block malicious ads

When we access the site, we notice that this application blocks the URLs that have been predicted as malicious in the ML model. In this way, this application only allows benign ads to appear, preventing contagion by any malware.

Figure 10: Website without adblocker

Figure 11: Website with malicious adblocker

In the images above, it is possible to verify that the ads that may contain malware are blocked. Furthermore, the content of the page is not affected, allowing the consumer to navigate through the page without running any undesired risks related to cybernetic attacks, and to enjoy the content of the page without any disturbance.

Next, in the image below, is shown the implementation of the framework, the steps that are followed to achieve the objectives previously defined. In other words, from the

identification of the companies' needs, to the detection of malicious URLs and their blocking to provide security to the consumer.

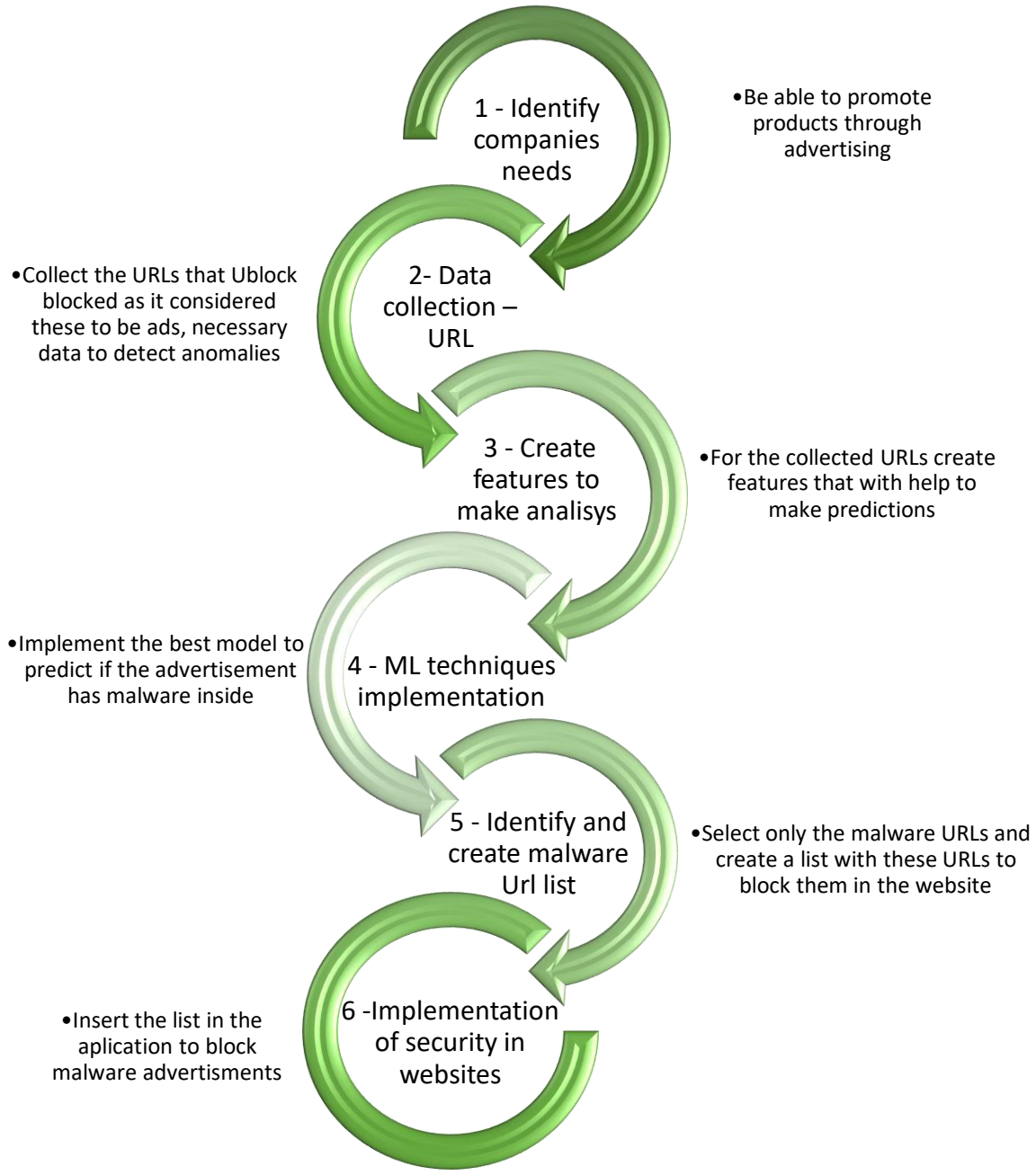


Figure 12: Framework Implementation

4.4. Evaluation

It has been found that for malware analysis it is very important to contain a solid data set, with all URLs that are required to be benign, because the intention is to block all malignant ones, obtaining the lowest possible number of false positives.

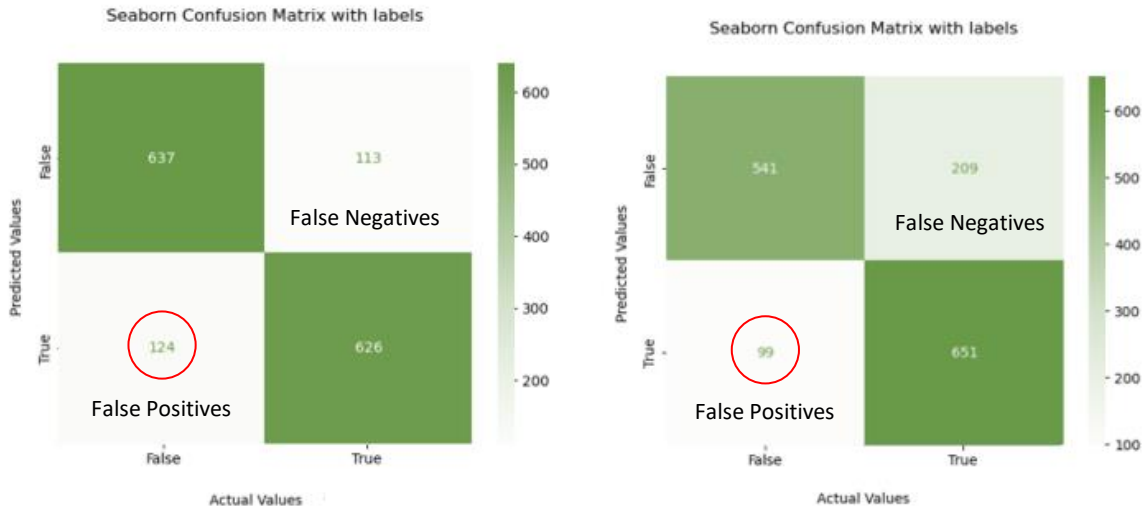


Figure 13: Confusion Matrixes

Several tests were performed, with the dataset that was contained 1000 URLs (sum of benign and malign URLs), to verify which would be the most adequate model for anomaly detection. Afterwards, for future analysis, other URLs were selected, in this way two databases were used, one with the same number of benign and malign URLs (1000 URLs in total), the other with more benign than malign URLs, being tested with different amounts of URLs (1000, 5000 and 20000 URLs). In the table below, the results obtained can be analysed:

ML Model	1000 URLs	5000 URLs	20000 URLs	20000 URLs without normalization	20000 URLs without URLs with no data
Logistic Regression	0.9	0.72	0.84	0.81	0.85
K-Nearest Neighbours - bagging	0.98	0.82	0.93	0.92	0.93
Decision Tree-bagging	0.99	0.82	0.94	0.94	0.94
Random Forest	0.99	0.85	0.95	0.94	0.95
Extra trees	0.98	0.81	0.92	0.92	0.93
Neural Network	0.96	0.83	0.94	0.73	0.94
AdaBoost	0.99	0.81	0.94	0.94	0.95
Gradient Boosting	0.99	0.84	0.95	0.95	0.96
Stacking	0.99	0.85	0.95	0.95	0.95
Voting Classifier	0.99	0.83	0.94	0.93	0.94

Table 4: Accuracy results from training data

A database with the same number of benign and malicious URLs can be seen to be overfitting the results. The database with more benign than malicious URLs has a better result when there are more URLs to analyze, works better with normalization and after all lines without data are removed in the 'request' feature. The features that are obtained with the techniques of 'request' library were found to be important to detect malicious URLs more accurately, for this reason this data cannot be null/zero to train the model. Thus, model selection is also important, in this way, several models were selected within Voting so that there is no risk of error. In this way, overfitting can also be avoided.

Next, 25 websites were selected to check how the selected model manages to detect malicious URLs and allow benign URLs to continue to appear so that users can be sure that these URLs have malware-free advertising.

Through these websites, 200 URLs were removed to carry out the predictions. Through 'https://www.virustotal.com/', it was verified whether or not these URLs are malignant, however, it cannot be guaranteed that this website accurately determines the malignant URLs.

In the table below, the results obtained through various models can be seen:

ML Model	1000 URLs	5000 URLs	20000 URLs	20000 URLs without normalization	20000 URLs without URLs with no data
Logistic Regression	0.23	0.41	0.56	0.48	0.61
K-Nearest Neighbours - bagging	0.25	0.42	0.64	0.22	0.65
Decision Tree-bagging	0.24	0.46	0.69	0.21	0.65
Random Forest	0.29	0.53	0.57	0.21	0.74
Extra trees	0.27	0.43	0.62	0.37	0.59
Neural Network	0.18	0.30	0.45	0.43	0.52
AdaBoost	0.21	0.46	0.54	0.21	0.52
Gradient Boosting	0.23	0.42	0.44	0.18	0.41
Stacking	0.21	0.53	0.57	0.21	0.74
Voting Classifier	0.24	0.38	0.63	0.63	0.69

Table 5: Accuracy results from testing data

In the table above, it can be seen that the models that previously showed better results, with 1000 URLs, now with real data show very low results due to overfitting. Thus, it is possible to verify that the models with better results are Random Forest, KNN, Decision Tree and Logistic Regression. As Random Forest and Decision Tree are similar models, only Random Forest was selected for the model in Voting, with the following results:

	20000 URLs	20000 URLs without URLs with no data
Voting: LR + RF + KNN	0.63	0.69
Voting: RF + KNN	0.64	0.71

Table 6: Accuracy results from testing data in the voting model

The model that scores best for predicting benign and malignant URLs is the model with Random Forest and KNN in the voting classifier, which is trained with all the features of the 'request' filled in, as this is the data that allows it to make predictions with the highest accuracy.

Accordingly, this framework will allow cybersecurity managers to display benign ads on their website with reduced probability of having malware inside them and ensure the safety of their users when they browse their website.

However, to ensure greater security and greater accuracy in detecting URLs that contain malware, it would probably be necessary to refresh the database to train the model every time new malware is detected, this will allow it to detect new malware that may appear and detect it more reliably. URLs that are benign and that marketing managers want to display on the website without being blocked by adblockers, will also need to be added to the database to train the model.

The renewal of the database to train the model will further increase the accuracy of detection of malicious URLs and increase the security of the web page, making it no longer possible for an unwanted cyber-attack to occur.

5. Conclusions

To conclude this dissertation, it is important to start by mentioning that the previously defined objectives were achieved. The proposed framework allows, through the browser extension, to detect malicious links, blocking them, which will provide more security to the users of the website, and will also allow companies to place advertising on different web pages without any fear of their ads being blocked through an ad blocker, which will ensure greater confidence between the company and the consumer.

5.1. Synthesis of the developed work

The research begun with a systematic literature review on the different fields that were considered relevant for this dissertation, namely Cybersecurity, Digital Forensic Science and Advertising, allowing to characterize the danger that digital media can provide nowadays if security measures are not implemented. From this analysis it was possible to design a framework to detect malicious URLs and block them, in order to protect Internet users, since this framework will allow companies to promote their products through digital advertising with increased security.

Below the questions asked at the beginning of the thesis will be answered.

RQ1: From the URLs extracted through Ublock, the components that are necessary to detect fraud were extracted, it was verified that it is possible to divide the characteristics into URL String Characteristics, Page Content Characteristics (obtained from the HTML code) and URL Domain Characteristics (whois information and shodan information). Thus, based on these characteristics it became possible to distinguish which components are more important for the detection of fraud, which allowed to prepare the data for the achievement of good predictions regarding the existence or lack of fraud in a certain URL.

RQ2: It was found that there are several web extensions that block advertisements, however these extensions block all types of advertisements, both malicious and benign, disabling companies from promoting their products with safe advertisements, without the existence of fraud or malware.

RQ3: Several machine learning models have been analysed, both supervised learning and unsupervised learning techniques. However, for the analysis of malware it was assumed that, in order to make more accurate predictions, it is very important to have real data from which the predictions will be performed, since the detection of malware will be based on similar characteristics detected by known URLs, which will allow a more precise determination of whether or not a URL contains malware.

RQ4: In Chapter 4, it was demonstrated which ML techniques can predict more accurately which URLs contain malware. It was determined that the best model for

detecting suspicious behaviour in a URL was the combination of Random Forest and KNN into Voting model, since voting allows to obtain the best advantages of both models.

RQ5: The goal of this thesis was to create a framework that would help companies to safely exhibit their advertisements to their clients. Therefore, from the ML model implemented, will be possible to detect the malicious URLs, and then block them through the web extension created. This framework, allows to prevent frauds and avoid consumers to suffer from malware attacks in the future.

In this way, it can be confirmed that the questions were answered along the thesis and the objective of this thesis was reached, several ML models were analysed and the most adequate one for malware detection was created, having been created the framework that, through the web extension, allows to guarantee more security to the consumer protecting him from malware attacks.

5.2. Research Limitations

Acknowledging what are the limitations of this research, as well as those that this science faces, is crucial to ensure greater security and greater accuracy in detecting URLs that may contain malware.

Previously it was mentioned that the larger the dataset the greater are the results obtained. However, this research was limited to the number of URLs to perform the training, since the larger the dataset, the longer it takes to train the model.

Another limitation results from the risk that can be run when extracting features from malicious URLs, since to obtain the features it is necessary that, through the code, the URL be opened, which could infect the computer in question if precautionary measures like antivirus were not used. This difficult to obtain data regarding the malicious URLs because when a dangerous malware was detected, the computer system, due to the use of antivirus, did not allow the analysis of the URL, so the fields regarding these URLs remained empty, which made the dataset less consistent.

The biggest limitation to run this framework is the difficulty in obtaining good features that allow a real distinction between benign and malign URLs, as mentioned earlier, technological advances have allowed criminals to infiltrate malware in advertising, this technological development has also allowed them to improve techniques with which criminals can insert a malware with a code very similar to benign ads. Thus, it is increasingly difficult to perform malware detection since it is necessary to implement new and more expensive techniques to obtain good features in order to perform the analysis in the ML model.

5.3. Future Work

Considering the limitations pointed out above, in the future work it would be important to focus on obtaining a larger dataset as possible to perform the training, avoiding blocking benign URLs, since a larger database will allow to obtain a larger number of observations, in order to obtain more accurate predictions regarding malicious URLs.

Likewise, it is also considered relevant to obtain the features of malicious URLs without the use of any antivirus, this will be possible if the gathering of the features is done on a virtual machine, on which it will then be possible to perform the formatting of the system to avoid information theft.

To be able to extract more relevant information about what can distinguish URLs, it would be important to perform analysis of the URL script through Text Mining techniques, since this model will allow to perform predictions based on what is encrypted in the URL code itself and then it would be possible to perform a clear differentiation between benign and malign URLs.

Finally, as a last consideration for future work, it is important to follow and adapt to innovations, since the rapid evolution of cyberattacks is increasingly able to adapt to the detection systems and cause even more damage. It will be important to develop tools, more powerful and trustworthy to make predictions faster and to be able to adapt constantly to the updates.

References

AdBlock. (n.d.). *About AdBlock*. Retrieved from AdBlock: https://getadblock.com/en/

Ahmad, K., Shekhar, J., & Yadav, K. (2010). Classification of SQL Injection Attacks. *VSRD Technical & Non-Technical Journal*, 235-242.

Ajdari, D., Hoofnagle, C., Stocksdale, T., & Good, N. (2013). Web Privacy Tools and Their Effect on Tracking and User Experience on the Internet.

Alrizah, M., Zhu, S., Xing, X., & Wang, G. (2019). Errors, Misunderstandings, and Attacks: Analyzing the Crowdsourcing Process of Ad-blocking Systems. *IMC'19*, 230–244.

Alves, P. (2014, August 27). *Use uBlock para bloquear anúncios de sites no seu navegador*. Retrieved from techtudo: https://www.techtudo.com.br/tudo-sobre/ublock.html

Apav. (n.d.). *Crime Ciberdependente*. Retrieved from Apav: https://apav.pt/cibercrime/

Artifacts. (2020, May 8). *artifacts*. Retrieved from What Is an Artifact? Everything You Need to Know: https://artifacts.ai/what-is-an-artifact/

Ayres, L., Brito, I., & Souza, R. (2019). Utilizando Aprendizado de Máquina para Detecção Automática de URLs Maliciosas Brasileiras. *Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, (pp. 972-985).

Ayyadevara, A. (2018). *Pro Machine Learning Algorithms.* Copyright .

Benetti, R. (2021, March 10). *Marketing Digital em 2021: o que é e como funciona?* Retrieved from Organica digital: https://www.organicadigital.com/blog/afinal-como-funciona-o-marketing-digital/

Bolina, L. (2018, February 14). *Publicidade Nativa: tudo o que você precisa saber sobre essa estratégia*. Retrieved from Rockcontent: https://rockcontent.com/br/blog/publicidade-nativa/

broadbandsearch. (2020). *How people use the internet in 2020*. Retrieved from broadbandsearch: https://www.broadbandsearch.net/blog/most-common-uses-internet-daily-life

Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. In *Design Science Research. Cases* (pp. 1-13).

Brownlee, J. (2020, April 22). *How to Develop an Extra Trees Ensemble with Python.* Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/extra-trees-ensemble-with-python/

Brownlee, J. (2020, April 17). *How to Develop Voting Ensembles With Python.* Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/voting-ensembles-with-python/

Brownlee, J. (2020, April 10). *Stacking Ensemble Machine Learning With Python.* Retrieved from Machinee Learning Mastery: https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/

Carr, S. (2020, May 15). *How Digital Fraud Will Eat 30% Of Media Budgets In 2020*. Retrieved from digital marketing magazine:

https://digitalmarketingmagazine.co.uk/articles/how-digital-fraud-will-eat-30-of-media-budgets-in-2020/5165

Chornyy, A. (2021, August 19). *17 Types of Online Advertising*. Retrieved from Plerdy: https://www.plerdy.com/blog/online-advertising-main-types-and-peculiarities/

Cintra, A. (n.d.). *Qual é a história da Publicidade Online?* Retrieved from postdigita: https://www.postdigital.cc/blog/artigo/qual-e-a-historia-da-publicidade-online

Couto, J. (2018). *Auditoria de Cibersegurança: um caso de estudo.* Porto.

Craigen, D., Diakun-Thibault, N., & Purse, R. (2014). Defining Cybersecurity. *Technology Innovation*, 13–21.

Crowdstrike. (2021, May 6). *What is Malvertising?* Retrieved from crowdstrike: https://www.crowdstrike.com/cybersecurity-101/malware/malvertising/

Cyber. (n.d.). *O que é a Cibersegurança?* Retrieved from Crispus: https://cyber-security.pt/blog/o-que-e-a-ciberseguranca/

Delua, J. (2021, March 12). *Supervised vs. Unsupervised Learning: What's the Difference?* Retrieved from IBM: https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Din, A. (2021, April 7). *What Is Malvertising and How to Remove It?* Retrieved from Heimdal Security: https://heimdalsecurity.com/blog/what-is-malvertising-and-how-to-remove-it/

Ducoffe, R. (1996). Advertising Value and Advertising on the Web. *Journal of Advertising Research*, 21-35.

Editor. (2016, December 23). *O Malvertising é a evolução do adware?* Retrieved from welivesecurity: https://www.welivesecurity.com/br/2016/12/23/malvertising-evolucao-adware/

Ferreira, M. (2019). Malicious URL Detection using Machine Learning Algorithms. *Proceedings of the Digital Privacy and Security Conference 2019*, (pp. 114-122). Porto.

Garimella, k., Kostakis, O., & Mathioudakis, M. (2017). Ad-blocking: A Study on Performance, Privacy and Counter-measures. *WebSci '17*, 259-262.

GeeksforGeeks. (2021, July 26). *Understanding Logistic Regression*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/understanding-logistic-regression/

Gonçalves, T. (2019, August 28). *AdBlock: o que é, para que serve e qual o seu futuro*. Retrieved from Compara Plano: https://comparaplano.com.br/blog/adblock/

Gregor, S., & Hevner, A. (2013). Positioning and Presenting Design Science Research for . *Mis Quarterly*, 337-355.

Haddadi, H., Nithyanand, R., Khattak, S., & Javed, M. (2016). The Adblocking Tug-of-War. *Login, vol. 41, no. 4*, 41-43.

Han, J., Kamber, M., & Pei, J. (2012). Classification: Basic Concepts. In J. Han, M. Kamber, & J. Pei, *Data Mining Concepts and Techniques* (pp. 330-349). Elsevier Inc.

Haq, N., Onik, A., Avishek, M., Hridoy, K., Rafni, M., Shah, F., & Farid, D. (2015). Application of Machine Learning Approaches in Intrusion Detection System: A Survey. *International Journal of Advanced Research in Artificial Intelligence*, 9-18.

Hashmi, S., Ikram, M., & Kaafar, M. (2019). Longitudinal Analysis of Online Ad-Blocking Blacklists. *IEEE 44th LCN Symposium on Emerging Topics in Networking*, (pp. 1-9).

Herhold, K. (2018, August 14). *How Businesses Use Online Advertising in 2018*. Retrieved from themanifest: https://themanifest.com/digital-marketing/how-businesses-use-online-advertising

Hevner, A. (2007). A three cycle view of design science research . *Scandinavian journal of*, 87–92.

Hill, R. (n.d.). *Ublock*. Retrieved from Gitplanet: https://gitplanet.com/project/ublock1618783248

Ipog. (2017, May 11). *Perícia digital é a solução para a investigação de crimes cibernéticos em empresas*. Retrieved from Ipog: https://blog.ipog.edu.br/tecnologia/computacao-forense-pericia-digital-e-a-solucao-para-a-investigacao-de-crimes-ciberneticos-em-empresas/

Jagatic, T., Johnson, N., Jakobsson, M., & Menczer, F. (2007). Social Phishing. *Communication of* .

Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in cybersecurity. *Journal of Computer and System Sciences*, 973-993.

Johnson, J. (2020, September 16). *Anomaly Detection with Machine Learning: An Introduction*. Retrieved from bmc: https://www.bmc.com/blogs/machine-learning-anomaly-detection/

Jordan, M., Kleinberg, J., & Scholkopf, B. (2008). *Information Science and Statistics: Support Vector Machines.* Germany: Springer.

Katke, K. (2007). The Impact of television advertising on child health and family spending-A Case Study. *International Marketing Conference on Marketing & Society*, 283-286.

Klosowski, T. (2021, March 11). *Our Favorite Ad Blockers and Browser Extensions to Protect Privacy*. Retrieved from Wirecutter: https://www.nytimes.com/wirecutter/reviews/our-favorite-ad-blockers-and-browser-extensions-to-protect-privacy/

Krakoff, S. (n.d.). *What's the Difference Between Cybersecurity and Computer Forensics?* Retrieved from Champlain College Online: https://online.champlain.edu/blog/difference-between-cybersecurity-and-computer-forensics

Kumar, V., & Shah, D. (2004). Pushing and Pulling on the Internet. *Marketing research*, 29-33.

Landage, J., & Wankhade, M. (2013). Malware and Malware Detection Techniques: A Survey. *International Journal of Engineering Research & Technology*, 61-68.

LIFARS. (2020, July 7). *What is Malvertising and How to Protect and Mitigate from it?* Retrieved from LIFARS: https://lifars.com/2020/07/what-is-malvertising/

Mamun, M., Rathore, M., Lashkari, A., & Stakhanova, N. (2016). Detecting Malicious URLs Using Lexical Analysis. *International Conference on Network and System Security*, (pp. 467–482).

March, S., Hevner, A., & Park, J. (2004). Design Science in Information Systems Research. *Mis Quarterly*, 75-105.

Mariyann, S. (2018). *Machine Learning for Cyber Forensics and Judicial Admissibility.* Bengaluru: National Law School of India University Nagarbhavi.

Master, E. (2018, January 17). *The Evolution of Ad Blocking*. Retrieved from Digital Media Knowledge: https://digitalmediaknowledge.com/medias/the-evolution-of-ad-blocking/

Mena, J. (2011). *Machine learning forensics for law enforcement, security, and intelligence.* Auerbach Publications.

Mishra, A., & Mahalik, D. (2017). IMPACT OF ONLINE-ADVERTISING ON CONSUMERS. *International Journal of Advanced Research*, 1935-1939.

Mohammed, A., & Varol, A. (2020). The Role of Machine Learning in Digital Forensics. *2020 8th International Symposium on Digital Forensics and Security (ISDFS).*

Mohammed, M., Khan, M., & Bashier, E. (2017). *Machine Learning Algorithms and Applications.* CRC Press.

Mohit, G. (2018, September 13). *ML | Linear Regression*. Retrieved from Geeks for Geeks: https://www.geeksforgeeks.org/ml-linear-regression/

Oliveira, V. (2021). *Cibersegurança e Inteligência Artificial .*

Pal, B., Daniel, T., Chatterjee, R., & Ristenpart, T. (2019). Beyond Credential Stuffing: Password Similarity Models using Neural Networks. *IEEE Symposium on Security and Privacy*, 417-434.

Pande, J. (2017). Introduction To Cyber Security. *Uttarakhand Open University, Haldwani*.

Peffers, K., Tuunanen, T., Gengler, C., & Rossi, M. (2006). The design science research process: A model for producing and presenting information systems research. *DESRIST*, 84-106.

Perlman, A. (n.d.). *The Growing Role of Machine Learning in Cybersecurity*. Retrieved from Security Roundtable: https://www.securityroundtable.org/the-growing-role-of-machine-learning-in-cybersecurity/

Pope, J. (2016). Ransomware: Minimizing the Risks. *Innovation in Clinical Neuroscience*, 37-40.

RedAlkemi. (2019, May 9). *Importance of Online Advertising*. Retrieved from RedAlkemi: https://www.redalkemi.com/blog/post/importance-of-online-advertising

Rodrigues, R. (2016, September 10). *Por que os anúncios na Internet atraem o cibercrime?* Retrieved from kaspersky: https://www.kaspersky.com.br/blog/internet-ads-101/6530/

Sahoo, D., Liu, C., & Steven, C. H. (2019). Malicious URL Detection using Machine Learning: A Survey. *arXiv:1701.07179v3*, 1-37.

Sarker, I., Kayes, A., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*.

Satpathy, S., Mallick, C., & Pradhan, S. (2018). Big Data Computing Application in Digital Forensics Investigation and Cyber Security. *International Journal of Computer Science and Mobile Applications*, 129-136 .

Silva, I. (2021, August 11). *Computação Forense - Conceitos básicos*. Retrieved from Linkedin: https://pt.linkedin.com/pulse/computa%C3%A7%C3%A3o-forense-conceitos-b%C3%A1sicos-idal%C3%A9cio-silva

Singh, A., & Potdar, V. (2009). Blocking Online Advertising – A State of the Art . *IEEE International Conference on Industrial Technology*, 1-10.

Snyder, P., Vastel, A., & Livshits, B. (2017). Who Filters the Filters: Understanding the Growth, Usefulness and Efficiency of Crowdsourced Ad Blocking. Washington.

Solms, R., & Niekerk, V. (2013). From information security to cyber security. *Computers & Security*, 97-102.

Sołtysik-Piorunkiewicz, A., Strzelecki, A., & Abramek, E. (2019). Evaluation of Adblock Software Usage. *Complex Systems Informatics and Modeling Quarterly*, 51-61.

Standberry, S. (2019, August 6). *What is Web Marketing? (And How to Use Web Marketing to Make Money)*. Retrieved from Life Marketing: https://www.lyfemarketing.com/blog/what-is-web-marketing/

Strzelecki, A., Abramek, E., & Sołtysik-Piorunkiewicz, A. (2019). Adblock Usage in Web Advertisement in Poland. In A. Strzelecki, E. Abramek, & A. Sołtysik-Piorunkiewicz, *Advances in Information and Communication Networks* (pp. 13-23). Springer, Cham.

Tarbani, N. (2021, April 19). *How the Gradient Boosting Algorithm works?* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/04/how-the-gradient-boosting-algorithm-works/

Techopedia. (2018, March 8). *Dimensionality Reduction*. Retrieved from Techopedia: https://www.techopedia.com/definition/30392/dimensionality-reduction

Techopedia. (2019, February 6). *Random Forest*. Retrieved from Techopedia: https://www.techopedia.com/definition/32935/random-forest

Techopedia. (b.d.). *Long Short-Term Memory (LSTM)*. Retrieved from Techopedia: https://www.techopedia.com/definition/33215/long-short-term-memory-lstm

Techopedia. (n.d.). *K-Means Clustering*. Retrieved from Techopedia: https://www.techopedia.com/definition/32057/k-means-clustering

Teles, T. (2015). *Cibersegurança Deteção de outliers.*

Thakur, N. (2021, May 22). *Anomaly Detection in Python — Part 2; Multivariate Unsupervised Methods and Code*. Retrieved from Towards Data Science: https://towardsdatascience.com/anomaly-detection-in-python-part-2-multivariate-unsupervised-methods-and-code-b311a63f298b

Tudoran, A. (2018). Why do Internet consumers block ads? New evidence from consumer opinion mining and sentiment analysis. *Internet Research*, 144-161.

uBlock, O. (n.d.). *uBlock Origin - Free, open-source ad content blocker.* Retrieved from uBlock Origin: https://ublockorigin.com/

VanderPlas, J. (2017). *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media.

Wroblewski, M. (2018, November 8). *Advantages of Online Marketing*. Retrieved from smallbusiness.chron: https://smallbusiness.chron.com/advantages-online-marketing-20438.html

Zúquete, A. (2018). *Segurança em Redes Informáticas.* Lisboa: Editora de Informática.

Annexes

Annex I – Created Features

Feature Name	Group	Description
Entroty	String Characteristics	Entropy of URL string
digits	String Characteristics	Total number of digits in URL string
urlLength	String Characteristics	Total number of characters in URL string
path	String Characteristics	Length of the URL path
netloc	String Characteristics	Total number of characters in the domain or host name
numParameters	String Characteristics	Total number of query parameters in URL
numFragments	String Characteristics	Total number of fragments in URL
numSubDomains	String Characteristics	Total number of subdomains in URL
domainExtension	String Characteristics	Domain Extension
hasHttp	Domain Characteristics	Website domain has http protocol
hasHttps	Domain Characteristics	Website domain has https protocol
bodyLength	String Characteristics	Total number of characters in URL's HTML page
url_host_is_ip	String Characteristics	The host or domain part of the URL string is an IP address rather than a domain name
port_in_string	String Characteristics	Verify if port is a string
is_encoded	String Characteristics	URL-encoded using a '%' character and a two character hew value corresponding to their UTF-8 character
num_encoded_char	String Characteristics	Number of '%' characters in the URL
number_of_periods	String Characteristics	Number of '.' characters in the URL
has_client	String Characteristics	URL string contains the keyword 'client'
has_admin	String Characteristics	URL string contains the keyword 'admin'

has_server	String Characteristics	URL string contains the keyword ‘server’
has_login	String Characteristics	URL string contains the keyword ‘login’
get_tld	String Characteristics	URL string has a port number in it
get_ip	Content Characteristics	URL IP
get_html	Content Characteristics	Structure of web pages
get_pq	Content Characteristics	HTML pyquery
get_scripts	Content Characteristics	Get all scripts in HTML
len_html	Content Characteristics	Total number of characters in HTML page excluding tags
entroty_html	Content Characteristics	Entropy of raw page content excluding HTML tags
number_script	Content Characteristics	Total number of scripts included in the page
script_ratio	Content Characteristics	The ratio of the total number of characters in the script tags to the total number of characters in all other tags
number_tokens	Content Characteristics	Total number of words on page as separated by ‘’ excluding tags
n_sentences	Content Characteristics	Total number of sentences on page as separated by ‘.’ excluding tags
n_punctuations	Content Characteristics	Total number of punctuations in the page
distinct_tokens	Content Characteristics	Total number of distinct words separated by ‘’
n_capitalizations	Content Characteristics	Total number of upper-case characters in the page content
average_n_tokens_in_sentence	Content Characteristics	Average number of words in all sentence
n_html_tags	Content Characteristics	Total number of HTML tags in page
n_hidden_tags	Content Characteristics	Total number of tags with class or id as hidden or attributes of visibility or display as none
n_iframes	Content Characteristics	Total number of iframe tags on page

n_objects	Content Characteristics	Total number of objects tags on page
n_embeds	Content Characteristics	Total number of embed tags on page
n_hyperlinks	Content Characteristics	Total number of hyperlinks on page
n_images	Content Characteristics	Total number of images tags on page
n_whitespace	Content Characteristics	Total number of whitespaces in page content
n_elements	Content Characteristics	Total number of tags that are not valid HTML tags
n_double_documents	Content Characteristics	Total numbrt of HTML structural tags that are repeated
n_eval_functions	Content Characteristics	Total number of eval() functions across all scripts on page
average_script_length	Content Characteristics	Average lenght off all script tag contents
entroty_script	Content Characteristics	Average entropy off all script tag contents
get_whois_dict	Domain Characteristics	
get_shodan_dict	Domain Characteristics	
n_of_subdomains	Domain Characteristics	Number of registered sub domains associated with the host
url_creation_date	Domain Characteristics	Domain registration date from whois data
url_expiration_date	Domain Characteristics	Domain expiration date from whois data
url_last_updated	Domain Characteristics	Last updated date
url_age	Domain Characteristics	Number of days since registration date
url_intended_life_span	Domain Characteristics	Number of days from registration to expiration
url_life_remaining	Domain Characteristics	Number of days left until expiration
url_registrar	Domain Characteristics	Domain registrar from whois data
url_regist_country	Domain Characteristics	Country of registration
url_open_ports	Domain Characteristics	Ports open on host
url_n_open_ports	Domain Characteristics	Total number of open ports
url_is_live	Domain Characteristics	If the host is online

url_isp	Domain Characteristics	Internet service of open on host
url_connection_speed	Domain Characteristics	Connection speed to host

Annex II – Example of HTML javascript

```
<html>

<head>

<script>

var forwardingUrl =
"/page/bouncy.php?&amp;bpae=Gbiud70ipVx%2Fj%2FNWspYHMpXcH0oKq166Hb4qpmQtoR
HsoWvyXaCLruubJU%2BSV624GhUBSuFZm2XkaU%2BvxaDjb6ELwB9%2FCBPlvlf5EbcJugwYP76
GBTN3VLrp69LMBTqEa%2FjGdDbzsUCuDsBYg97lNc%2BIoexvIZ%2B5%2FxhClO5pWFfS%2FITvx
hCXP9vascpTplPNwdRBEjM%2FWjed%2FZwy9mbymG7YA%2FJLJkqjr2D%2FEzgXuhG4BBBgOdc
8kS9WRgHHpR0Z62FswbUkNypyndI37OaWK9eTvG%2BwHr3HHP85aNba6L1ARAZyVmrW0HvV
wygGEXefzIf6ZJP4xlQCfjakAAuXlI55VJlT89hrKg0gzXMBqq9lpI91lATmrjm2zZdVBx8TZ%2BJ30NR
HF1CxRwvwDF%2BGyezfCXF8h%2Bkb&amp;redirectType=js";

var destinationUrl =
"/page/bouncy.php?&amp;bpae=Gbiud70ipVx%2Fj%2FNWspYHMpXcH0oKq166Hb4qpmQtoR
HsoWvyXaCLruubJU%2BSV624GhUBSuFZm2XkaU%2BvxaDjb6ELwB9%2FCBPlvlf5EbcJugwYP76
GBTN3VLrp69LMBTqEa%2FjGdDbzsUCuDsBYg97lNc%2BIoexvIZ%2B5%2FxhClO5pWFfS%2FITvx
hCXP9vascpTplPNwdRBEjM%2FWjed%2FZwy9mbymG7YA%2FJLJkqjr2D%2FEzgXuhG4BBBgOdc
8kS9WRgHHpR0Z62FswbUkNypyndI37OaWK9eTvG%2BwHr3HHP85aNba6L1ARAZyVmrW0HvV
wygGEXefzIf6ZJP4xlQCfjakAAuXlI55VJlT89hrKg0gzXMBqq9lpI91lATmrjm2zZdVBx8TZ%2BJ30NR
HF1CxRwvwDF%2BGyezfCXF8h%2Bkb&amp;redirectType=meta";

var addDetection = true;

if (addDetection) { var inIframe = window.self !== window.top; forwardingUrl +=
"&amp;inIframe=" + inIframe;

var inPopUp = (window.opener !== undefined &amp;&amp; window.opener !== null
&amp;&amp; window.opener !== window); forwardingUrl += "&amp;inPopUp=" + inPopUp; }

window.location.replace(forwardingUrl);

</script>

<noscript>

<meta http-equiv="refresh"
content="1;url=/page/bouncy.php?&amp;bpae=Gbiud70ipVx%2Fj%2FNWspYHMpXcH0oKq166
Hb4qpmQtoRHsoWvyXaCLruubJU%2BSV624GhUBSuFZm2XkaU%2BvxaDjb6ELwB9%2FCBPlvlf5
EbcJugwYP76GBTN3VLrp69LMBTqEa%2FjGdDbzsUCuDsBYg97lNc%2BIoexvIZ%2B5%2FxhClO5p
WFfS%2FITvxhCXP9vascpTplPNwdRBEjM%2FWjed%2FZwy9mbymG7YA%2FJLJkqjr2D%2FEzgXu
hG4BBBgOdc8kS9WRgHHpR0Z62FswbUkNypyndI37OaWK9eTvG%2BwHr3HHP85aNba6L1ARA
ZyVmrW0HvVwygGEXefzIf6ZJP4xlQCfjakAAuXlI55VJlT89hrKg0gzXMBqq9lpI91lATmrjm2zZdVBx
8TZ%2BJ30NRHF1CxRwvwDF%2BGyezfCXF8h%2Bkb&amp;redirectType=meta"/>

</noscript>

</head>

</html>
```

Annex III – Example of whois

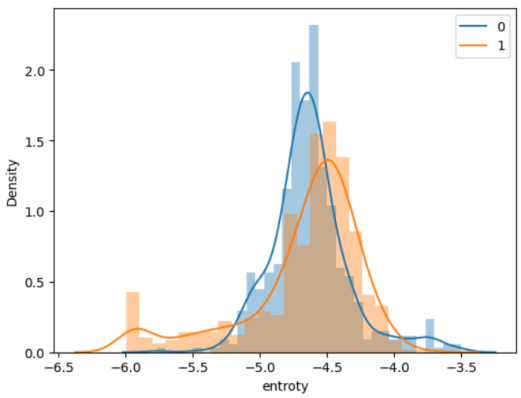

```
{
  "domain_name": "SENDGRID.NET",
  "registrar": "GoDaddy.com, LLC",
  "whois_server": "whois.godaddy.com",
  "referral_url": null,
  "updated_date": "2021-04-19 15:49:45",
  "creation_date": "2009-04-20 09:09:23",
  "expiration_date": "2026-04-20 09:09:23",
  "name_servers": [
    "NS10.DNSMADEEASY.COM",
    "NS11.DNSMADEEASY.COM",
    "NS12.DNSMADEEASY.COM",
    "NS13.DNSMADEEASY.COM",
    "NS14.DNSMADEEASY.COM",
    "NS15.DNSMADEEASY.COM" ],
  "status": [
    "clientDeleteProhibited https://icann.org/epp#clientDeleteProhibited",
    "clientRenewProhibited https://icann.org/epp#clientRenewProhibited",
    "clientTransferProhibited https://icann.org/epp#clientTransferProhibited",
    "clientUpdateProhibited https://icann.org/epp#clientUpdateProhibited"],
  "emails": "abuse@godaddy.com",
  "dnssec": "unsigned",
  "name": null,
  "org": null,
  "address": null,
  "city": null,
  "state": null,
  "zipcode": null,
  "country": null
}
```

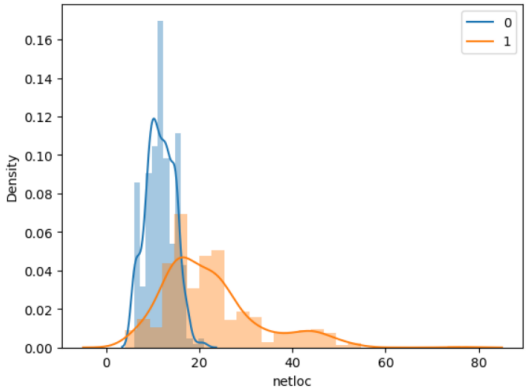
Annex IV – Example of Shodan

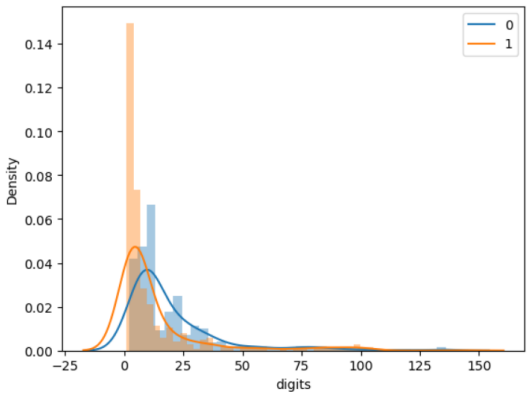
{'region_code': '14', 'tags': ['cloud'], 'ip': 232911387, 'area_code': None, 'domains': ['cloudfront.net'], 'hostnames': ['server-13-225-242-27.lis50.r.cloudfront.net'], 'postal_code': None, 'dma_code': None, 'country_code': 'PT', 'org': 'Amazon.com, Inc.', 'data': [{'ip': 232911387, 'asn': 'AS16509', 'http': {'robots_hash': None, 'redirects': [], 'securitytxt': None, 'title': 'ERROR: The request could not be satisfied', 'sitemap_hash': None, 'robots': None, 'server': 'CloudFront', 'host': '13.225.242.27', 'html': '<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">\n<HTML><HEAD><META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=iso-8859-1">\n<TITLE>ERROR: The request could not be satisfied</TITLE>\n</HEAD><BODY>\n<H1>403 ERROR</H1>\n<H2>The request could not be satisfied.</H2>\n<HR noshade size="1px">\nBad request.\nWe can\'t connect to the server for this app or website at this time. There might be too much traffic or a configuration error. Try again later, or contact the app or website owner.\n<BR clear="all">\nIf you provide content to customers through CloudFront, you can find steps to troubleshoot and help prevent this error by reviewing the CloudFront documentation.\n<BR clear="all">\n<HR noshade size="1px">\n<PRE>\nGenerated by cloudfront (CloudFront)\nRequest ID: NwPbRv5DPiguX7Xslbf0RJZhnTUYiOuJsE4I1Texc41AW4Xe6uFUfw==\n</PRE>\n<ADDRESS>\n</ADDRESS>\n</BODY></HTML>', 'location': '/', 'components': {}, 'securitytxt_hash': None, 'sitemap': None, 'html_hash': -16696367}, 'tags': ['cloud'], 'timestamp': '2021-10-18T06:37:18.646482', 'org': 'Amazon.com, Inc.', 'isp': 'Amazon.com, Inc.', 'data': 'HTTP/1.1 403 Forbidden\r\nServer: CloudFront\r\nDate: Mon, 18 Oct 2021 06:37:15 GMT\r\nContent-Type: text/html\r\nContent-Length: 915\r\nConnection: keep-alive\r\nX-Cache: Error from cloudfront\r\nVia: 1.1 9286764bc0c8327719870fa33a225c9a.cloudfront.net (CloudFront)\r\nX-Amz-Cf-Pop: LIS50-C1\r\nX-Amz-Cf-Id: NwPbRv5DPiguX7Xslbf0RJZhnTUYiOuJsE4I1Texc41AW4Xe6uFUfw==\r\n\r\n', 'port': 80, 'cloud': {'region': 'GLOBAL', 'service': 'AMAZON', 'provider': 'Amazon'}, 'hostnames': ['server-13-225-242-27.lis50.r.cloudfront.net'], 'transport': 'tcp', 'ip_str': '13.225.242.27', 'domains': ['cloudfront.net'], 'hash': 2013972049, 'os': None, '_shodan': {'crawler': 'bfb295b1dac9b1783126e88777b186a5006c26b0', 'ptr': True, 'id': '0178f469-63d1-4a33-977a-090237882dad', 'module': 'http', 'options': {}}, 'opts': {}, 'location': {'city': 'Lisbon', 'region_code': '14', 'area_code': None, 'longitude': -9.13333, 'country_code3': None, 'latitude': 38.71667, 'postal_code': None, 'dma_code': None, 'country_code': 'PT', 'country_name': 'Portugal'}}, {'hash': 406313240, 'tags': ['cloud'], 'timestamp': '2021-10-11T17:09:07.097137', 'org': 'Amazon.com, Inc.', 'isp': 'Amazon.com, Inc.', 'data': 'HTTP/1.1 400 Bad Request\r\nServer: CloudFront\r\nDate: Mon, 11 Oct 2021 17:09:07 GMT\r\nContent-Type: text/html\r\nContent-Length: 915\r\nConnection: close\r\nX-Cache: Error from cloudfront\r\nVia: 1.1 20aad0efbdc15ee6c121141c606f1781.cloudfront.net (CloudFront)\r\nX-Amz-Cf-Pop: LIS50-C1\r\nX-Amz-Cf-Id: 7DWSBuon_00GBbEh5UdNWs9vY2z3BmelFbKhzuZi6vmxsKEUB9Dbrg==\r\n\r\n<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">\n<HTML><HEAD><META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=iso-8859-1">\n<TITLE>ERROR: The request could not be satisfied</TITLE>\n</HEAD><BODY>\n<H1>400 ERROR</H1>\n<H2>The request could not be satisfied.</H2>\n<HR noshade size="1px">\nBad request.\nWe can\'t connect to the server for this app or website at this time. There might be too much traffic or a configuration error. Try again later, or contact the app or website owner.\n<BR clear="all">\nIf you provide

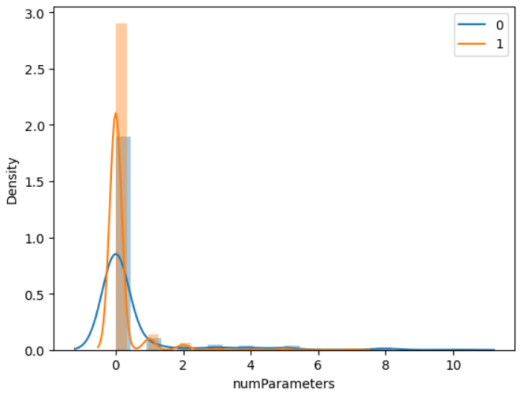
content to customers through CloudFront, you can find steps to troubleshoot and help prevent this error by reviewing the CloudFront documentation.\n<BR clear="all">\n<HR noshade size="1px">\n<PRE>\nGenerated by cloudfront (CloudFront)\nRequest ID: 7DWSBuon_00GBbEh5UdNWs9vY2z3BmelFbKhzuZi6vmxsKEUB9Dbrg==\n</PRE>\n<ADDRESS>\n</ADDRESS>\n</BODY></HTML>', 'asn': 'AS16509', 'port': 443, 'cloud': {'region': 'GLOBAL', 'service': 'AMAZON', 'provider': 'Amazon'}, 'hostnames': ['server-13-225-242-27.lis50.r.cloudfront.net'], 'transport': 'tcp', 'ip': 232911387, 'domains': ['cloudfront.net'], 'ip_str': '13.225.242.27', 'os': None, '_shodan': {'crawler': 'cdd92e2d835a37d2798fa6c7105171f4d214012f', 'ptr': True, 'id': '592a74c9-b898-4633-9fdc-349488869fa3', 'module': 'https', 'options': {}}, 'opts': {}, 'location': {'city': 'Lisbon', 'region_code': '14', 'area_code': None, 'longitude': -9.13333, 'country_code3': None, 'latitude': 38.71667, 'postal_code': None, 'dma_code': None, 'country_code': 'PT', 'country_name': 'Portugal'}}], 'asn': 'AS16509', 'city': 'Lisbon', 'latitude': 38.71667, 'isp': 'Amazon.com, Inc.', 'longitude': -9.13333, 'last_update': '2021-10-18T06:37:18.646482', 'country_code3': None, 'country_name': 'Portugal', 'ip_str': '13.225.242.27', 'os': None, 'ports': [80, 443]}

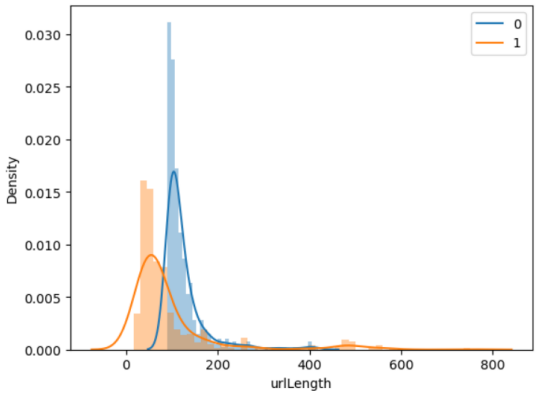
Annex V – Feature Analisis

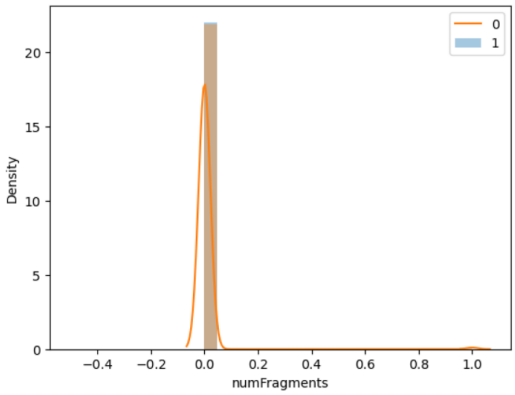


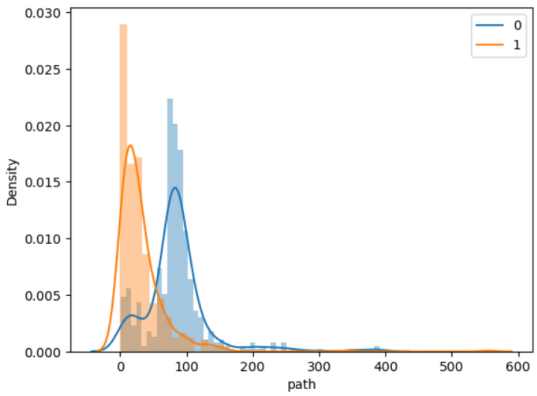


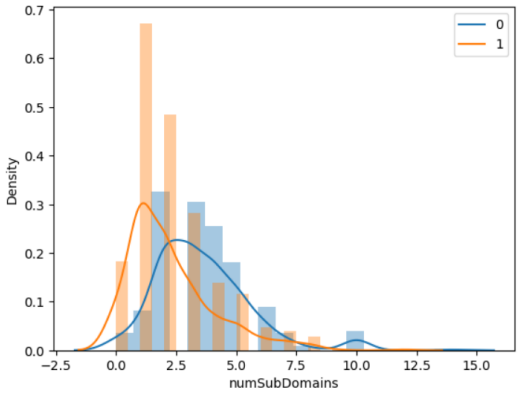


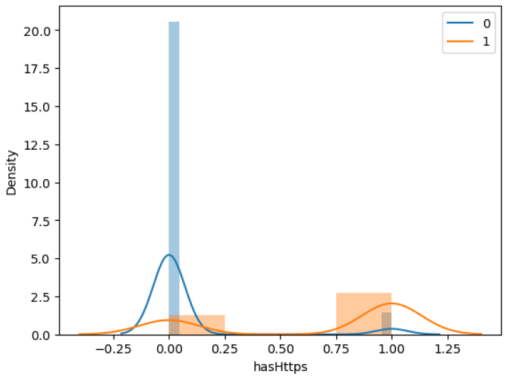


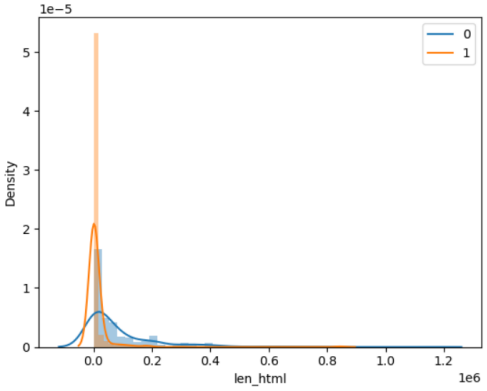


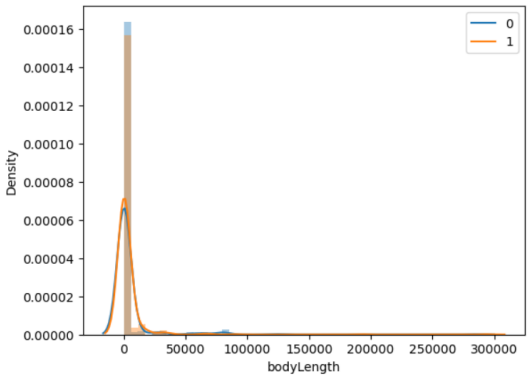


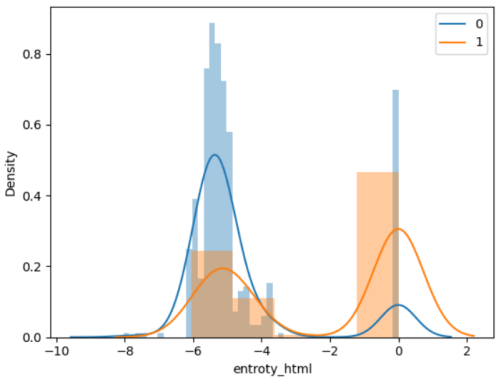


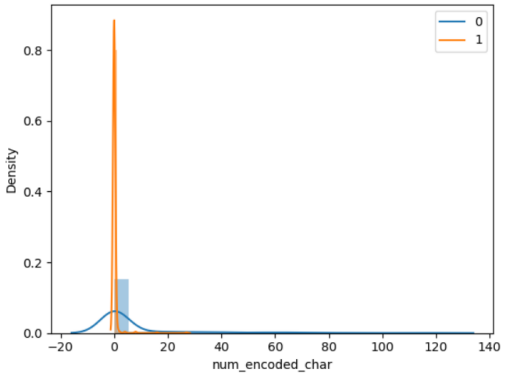


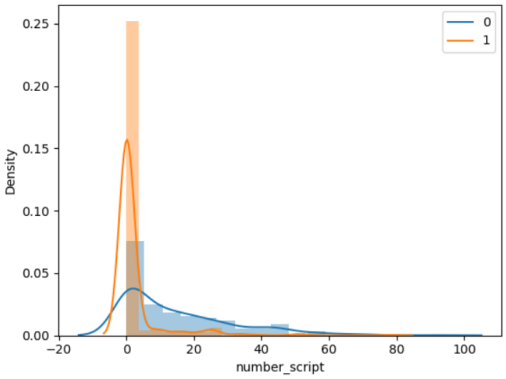

0
1
Density
hasHttps

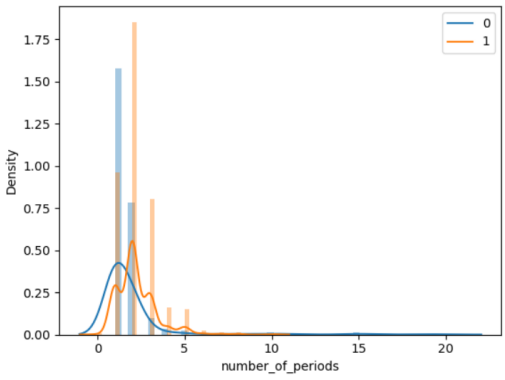

1e−5
0
1
Density
len_html
1e6

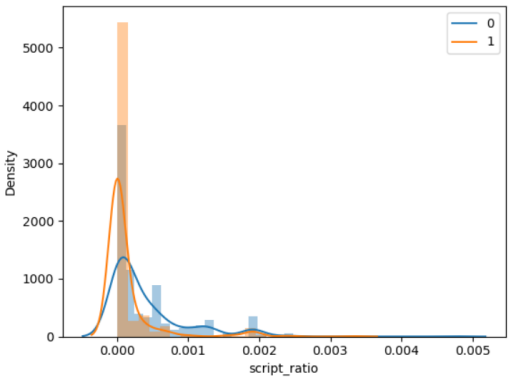

0
1
Density
bodyLength

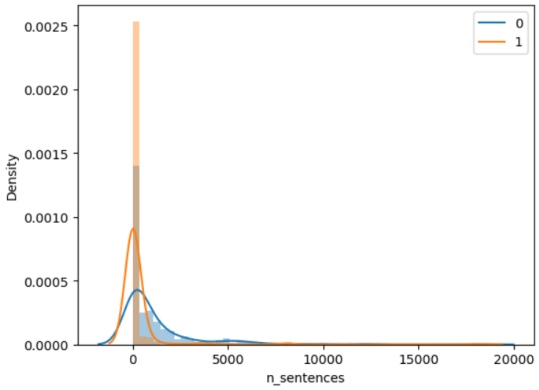

0
1
Density
entroty_html

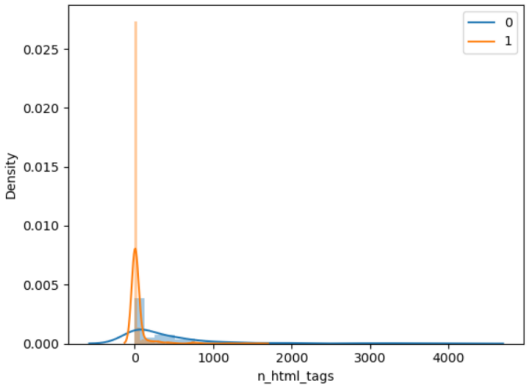

0
1
Density
num_encoded_char

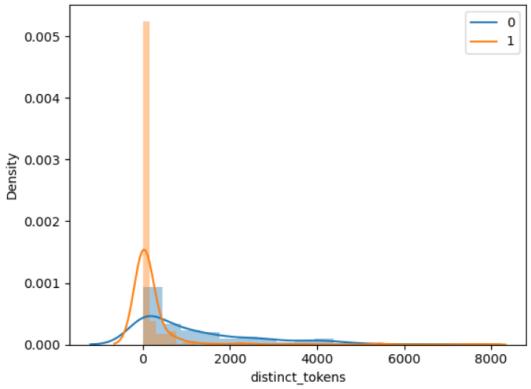

0
1
Density
number_script

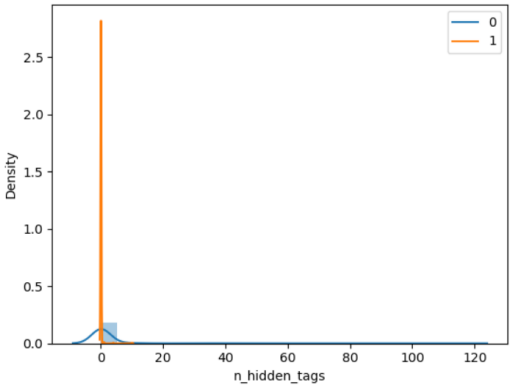

0
1
Density
number_of_periods

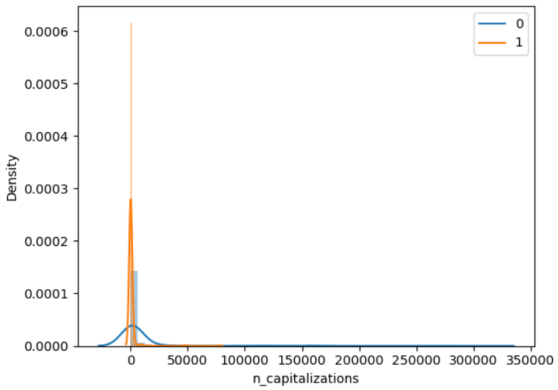

0
1
Density
script_ratio

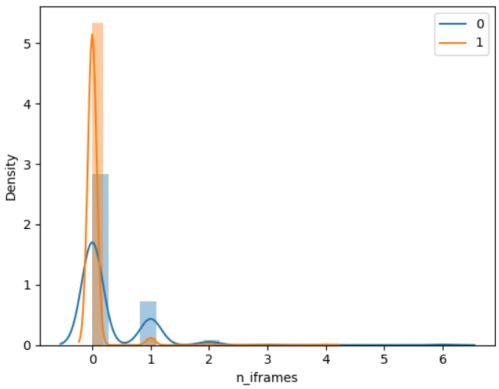

0
1
Density
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
5000
10000
15000
20000
n_sentences

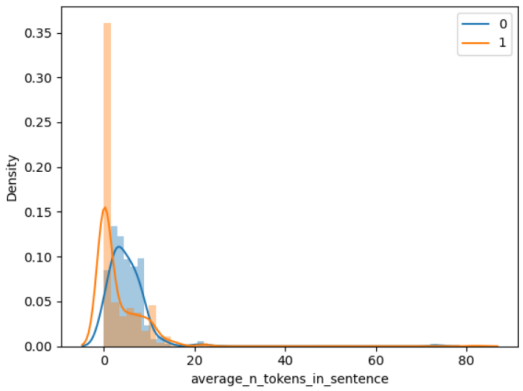

0
1
Density
0.025
0.020
0.015
0.010
0.005
0.000
0
1000
2000
3000
4000
n_html_tags

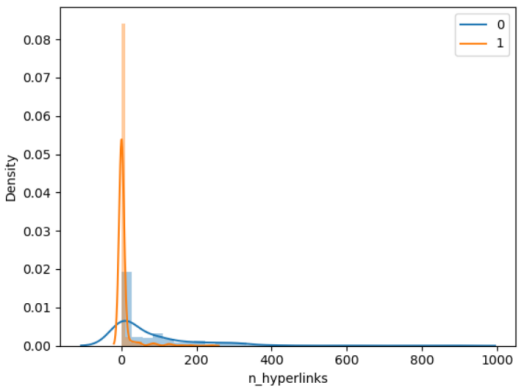

0
1
Density
0.005
0.004
0.003
0.002
0.001
0.000
0
2000
4000
6000
8000
distinct_tokens

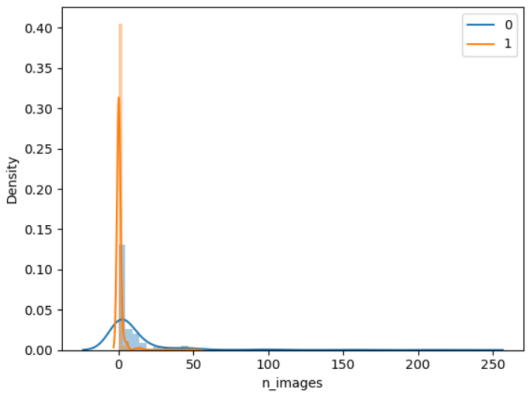

0
1
Density
2.5
2.0
1.5
1.0
0.5
0.0
0
20
40
60
80
100
120
n_hidden_tags

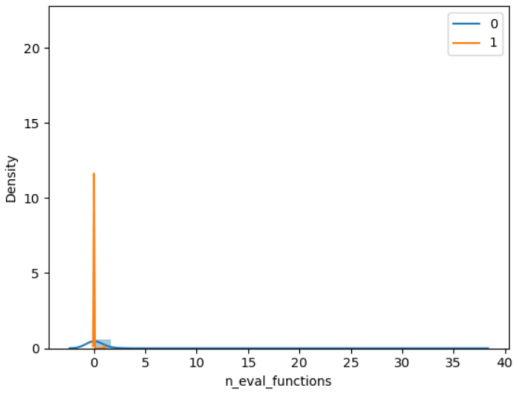

0
1
Density
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
50000
100000
150000
200000
250000
300000
350000
n_capitalizations

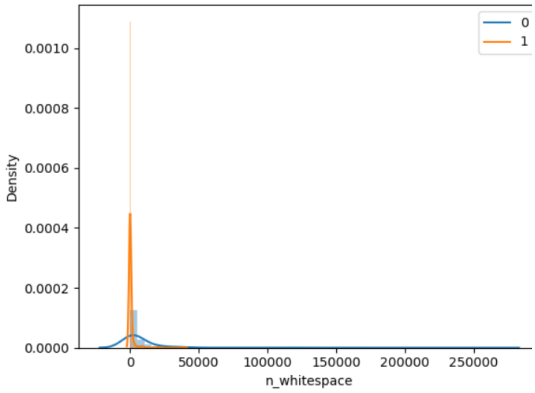

0
1
Density
5
4
3
2
1
0
0
1
2
3
4
5
6
n_iframes

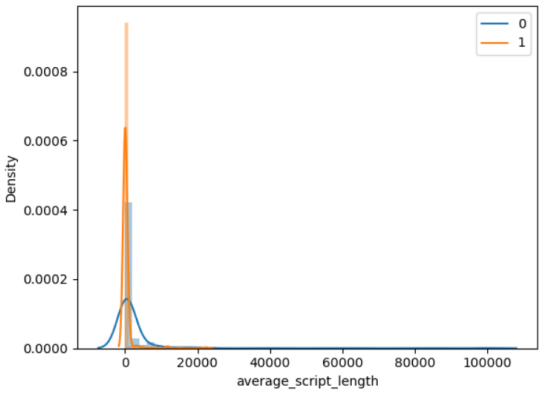

0
1
Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
average_n_tokens_in_sentence

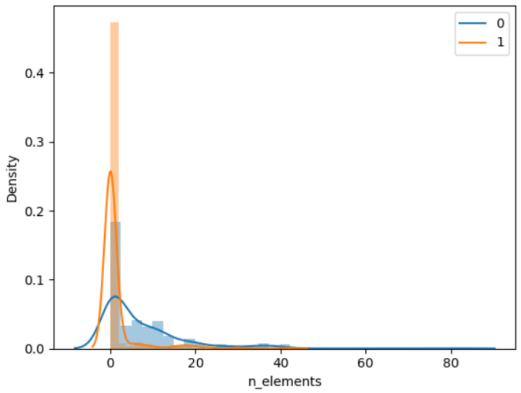

0
1
Density
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
200
400
600
800
1000
n_hyperlinks

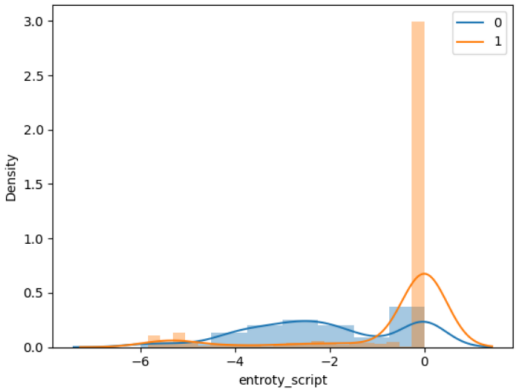

0
1
Density
n_images

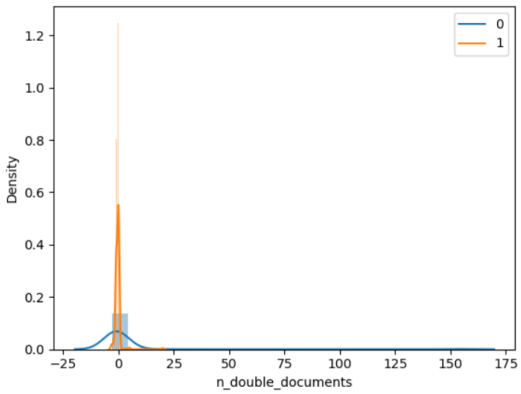

0
1
Density
n_eval_functions

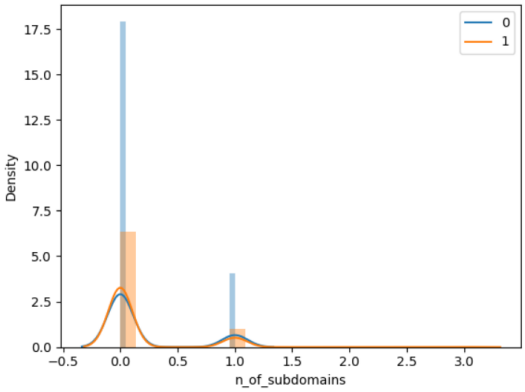

0
1
Density
n_whitespace

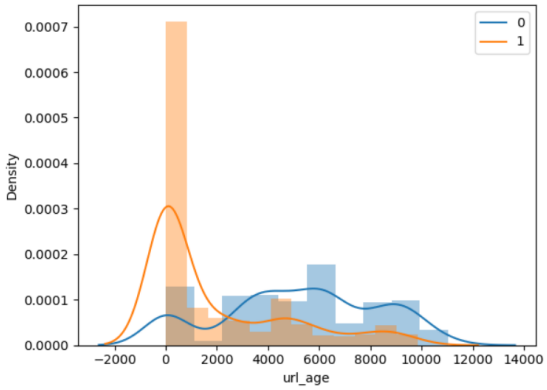

0
1
Density
average_script_length

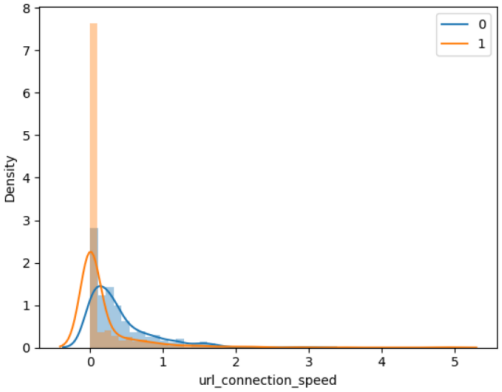

0
1
Density
n_elements

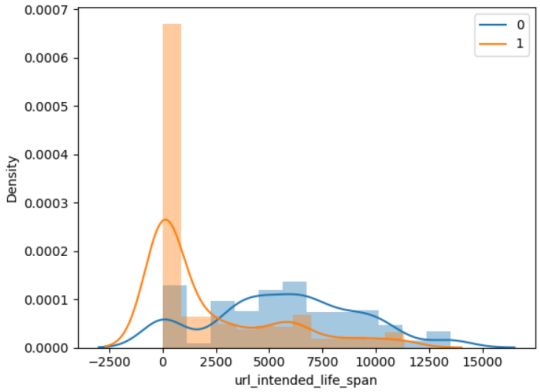

0
1
Density
entroty_script

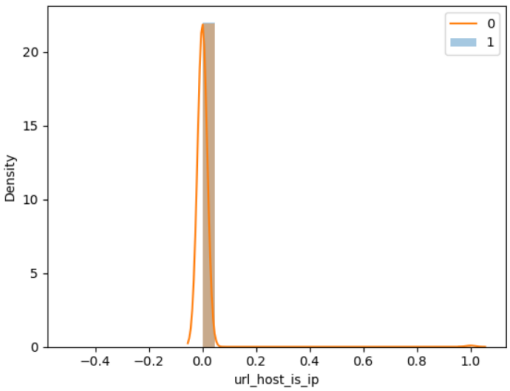

0
1
Density
n_double_documents

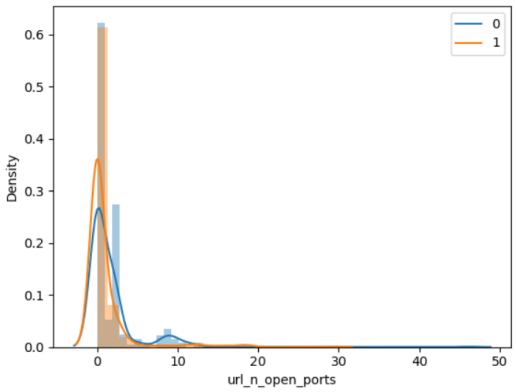

0
1
Density
n_of_subdomains

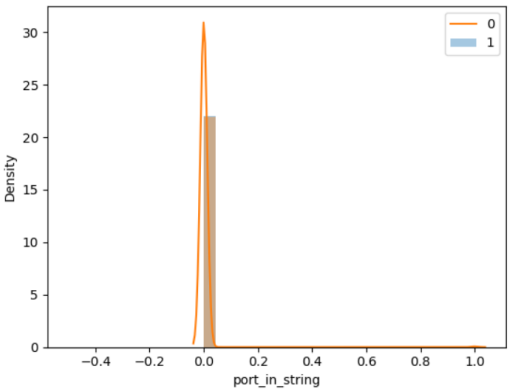

0
1
Density
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
−2000
0
2000
4000
6000
8000
10000
12000
14000
url_age

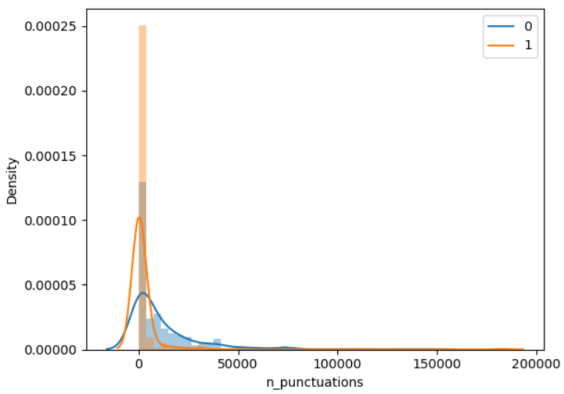

0
1
Density
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
url_connection_speed

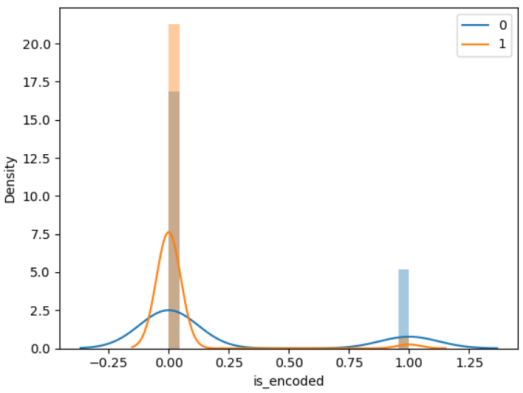

0
1
Density
0.0007
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
−2500
0
2500
5000
7500
10000
12500
15000
url_intended_life_span

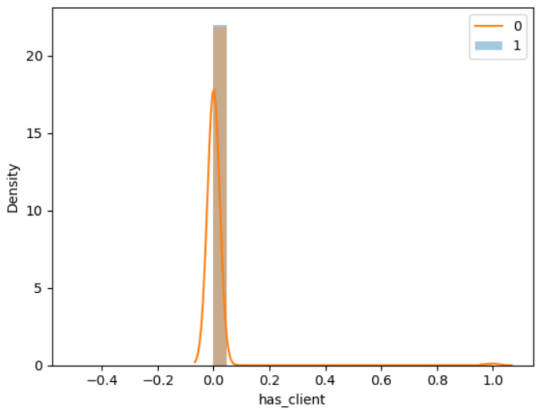

0
1
Density
20
15
10
5
0
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
url_host_is_ip

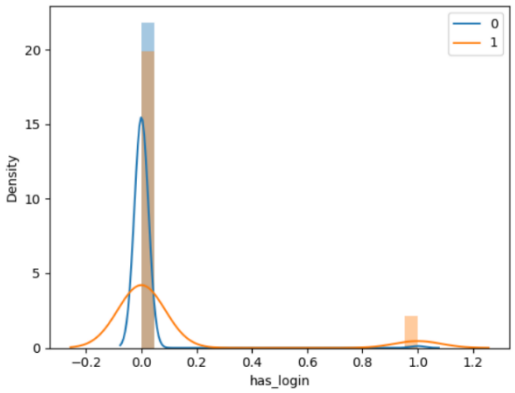

0
1
Density
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
url_n_open_ports

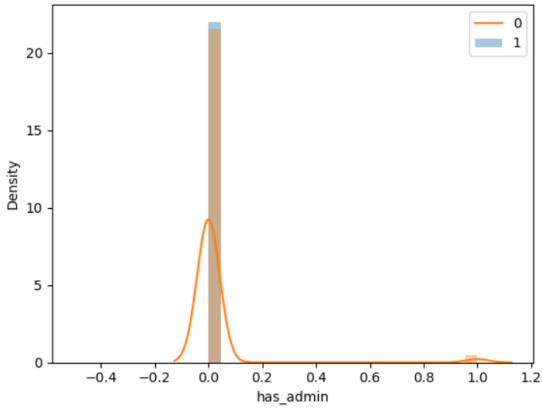

0
1
Density
30
25
20
15
10
5
0
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
port_in_string

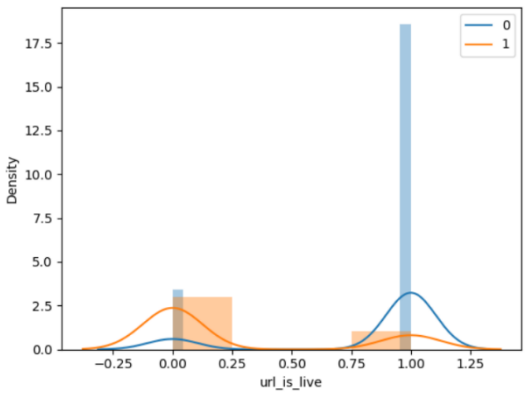

0
1
Density
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
50000
100000
150000
200000
n_punctuations

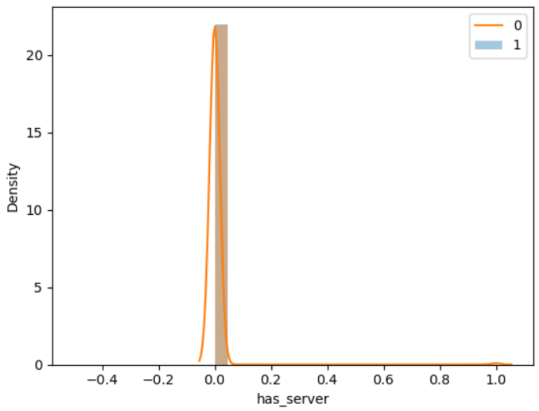

0
1
Density
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
is_encoded


0
1
Density
has_client


0
1
Density
has_login


0
1
Density
has_admin


0
1
Density
url_is_live


0
1
Density
has_server