Universidad del Rosario | Escuela de Ingeniería, Ciencia y Tecnología | MACC Matemáticas Aplicadas y Ciencias de la Computación | HINNT Hub de INNovación y Transferencia

Malware detection through Portable Executable information

Sara Gallego y Daniel Forero

sara.gallego@urosario.edu.co
daniel.forero@urosario.edu.co

@MACC_URosario @MACC.URosario macc_ur

Agenda

1. Introduction.
2. Problem Statement.
3. References.

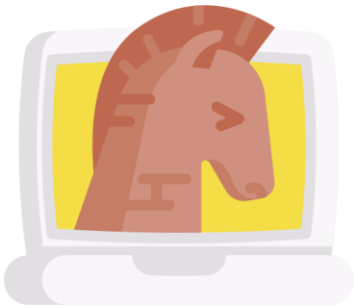
INTRODUCTION

Nowadays technology has become fundamental in people's lives, and this is why more and more users are using electronic devices to access it. However, due to this, the risks of using electronic devices have also increased.

For this reason, currently, many people are dedicated to the development of software and protection methods for these devices and systems. However, some people see an opportunity to exploit possible vulnerabilities that systems may have for malicious purposes, these are malware developers.

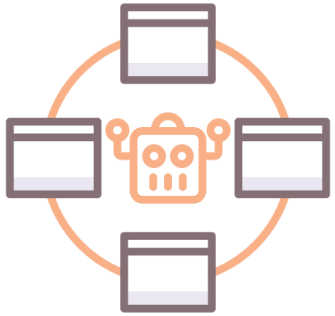
Malware characterization

There are many types of malware. Its classification depends on their purpose and what they do on the victim's machine.



Troyan

Botnet

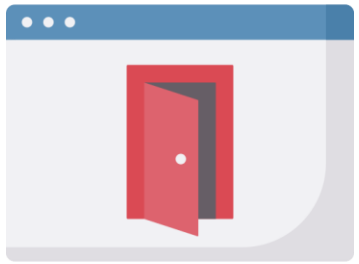


Worm

Ransomware

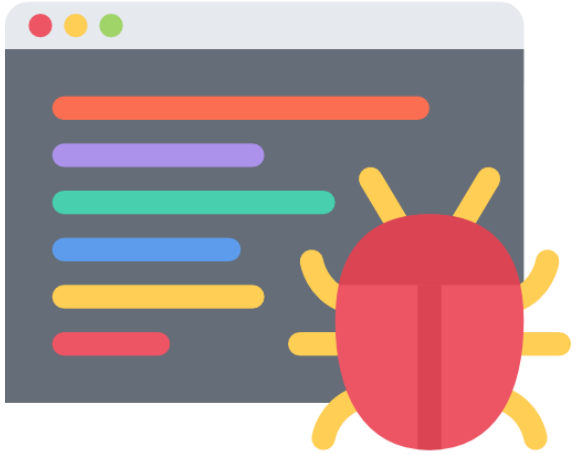
Keylogger

Backdoor



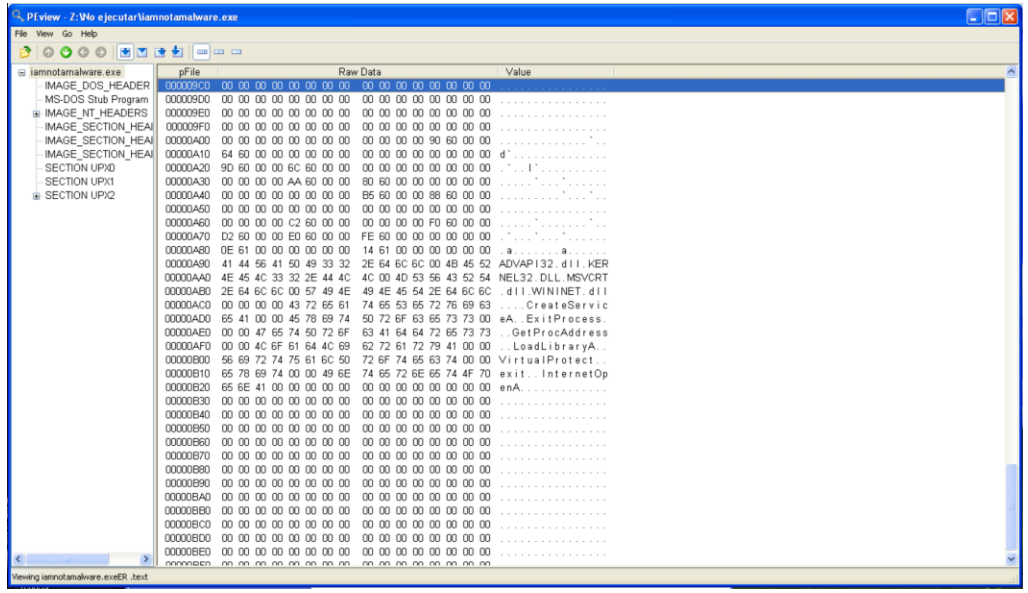
Malware characterization

Malware, like any type of file, has certain information that facilitates its recognition. An example of this could be the HASH of a program, which allows us to recognize if it is a malware or not, by searching for it on a platform that analyzes files and web pages through antivirus (e.g. VirusTotal).



As well as HASH, there is the Portable Executable (PE) format (this format exists exclusively for 32-bit or 64-bit Windows OS) that refers to the object code executable file format and dynamic link libraries (DLL). A PE is a data structure that contains the information necessary for the Windows loader to run the program.

PE as malware recognition tool



Example of a PE format of a malware obtained using Peview.

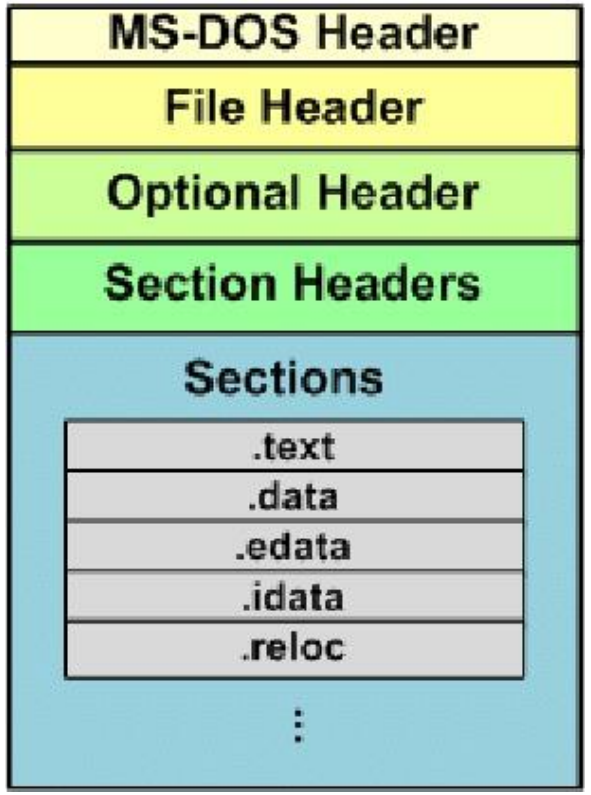
By analyzing the PE associated with a file or program, it is possible to recognize certain features that make it achievable to identify whether a file is malicious or not.

Behaviors such as the appearance of OS folder names, domains, or import and use of libraries and functions that may compromise the integrity of the machine, make a program a potential malware.

PROBLEM STATEMENT

Universidad del Rosario | Escuela de Ingeniería, Ciencia y Tecnología | MACC | HINNT

By means of the information that is obtained from file's or program's PE, we seek to standardize the features that allow the categorization of a file as a malware.

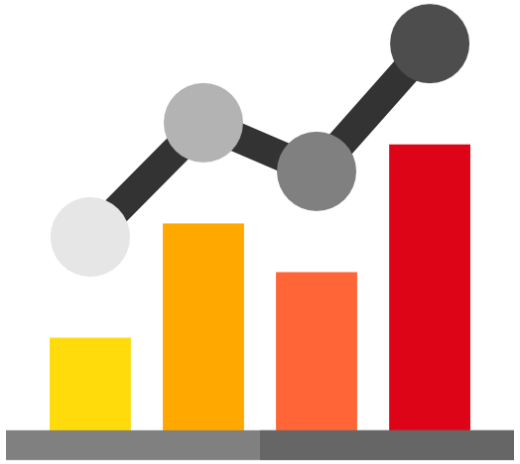


Modelo ML para abordar el problema

Universidad del Rosario | Escuela de Ingeniería, Ciencia y Tecnología | MACC | HINNT

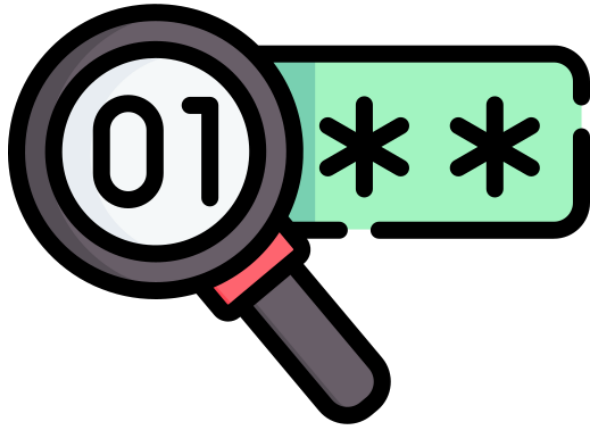
Weka, a software platform for machine learning written in Java and developed at the University of Waikato, was used for the development of the project.

Through this platform, a system will be trained in order to recognize potentially malicious files through certain frequent features or characteristics.



Datasets utilizados para el entrenamiento del modelo de ML

Initially, there are two datasets used for the development of the project. Those two datasets contain information related to the PE format. The databases allow the proper construction of the ML model providing data that allow us to identify and analyze specific features of the PE that could lead the model to categorize it as malware or not.



PE-imports and PE section headers

The information contained on the datasets previously mentioned is:

- TOP-1000 PE-imports
- PE section headers

Each one of those contains the analysis of at least 45k different hashes.

The first dataset allows the model to filter and recognize the imports commonly found on malware files out of the 1000 most frequent imports on PE format files.

The second dataset allows the model to have information about the headers of the .text, .code, and CODE sections. This dataset relates the data with the fact of it being malware or goodware. It'll be useful to understand the header's behavior and having a solid foundation on the recognition capability of the ML model.

REFERENCES

- https://www.kaggle.com/ang3loliveira/malware-analysis-datasets-pe-section-headers
- https://www.kaggle.com/ang3loliveira/malware-analysis-datasets-top1000-pe-imports
- https://blog.kowalczyk.info/articles/pefileformat.html

Universidad del Rosario | Escuela de Ingeniería, Ciencia y Tecnología | MACC Matemáticas Aplicadas y Ciencias de la Computación | HINNT Hub de INNovación y Transferencia

GRACIAS

Sara Gallego y Daniel Forero

sara.gallego@urosario.edu.co
daniel.forero@urosario.edu.co

@MACC_URosario | @MACC.URosario | macc_ur