**IS 734 Midterm Progress Report**

Rithvik Sai Pabba - YK37094

Prashanthi Grace Sandra - JU36815

Catherin Revu - YA55734

Vidyasagar Athikam - XK09178

Sai Anusha Vadlamani - AP46295

**Problem Statement:**

The need for reliable detection and prevention solutions has become essential due to the constant increase in malware attacks via Portable Executable (PE) files. In addition to protecting individual systems, containing large-scale cyber threats and guaranteeing data integrity and inviolability require an extremely effective classification system. We must adapt our strategy to keep up with the always changing cyber threat scenario. Machine learning emerges as a very flexible and innovative solution to this ever-changing dilemma. Its potential is found in its capacity to constantly adapt and fend off new attacks, which makes it a crucial component in the effort to improve cybersecurity.

**Objective**:

Our primary objective is to classify Portable Executable (PE) files as either malicious or benign using various machine learning techniques. As malware frequently uses PE files to compromise systems, the rapid, efficient, and accurate classification of these files becomes crucial.

**Tasks Performed:**

**Data preprocessing:**

**Data Source:**

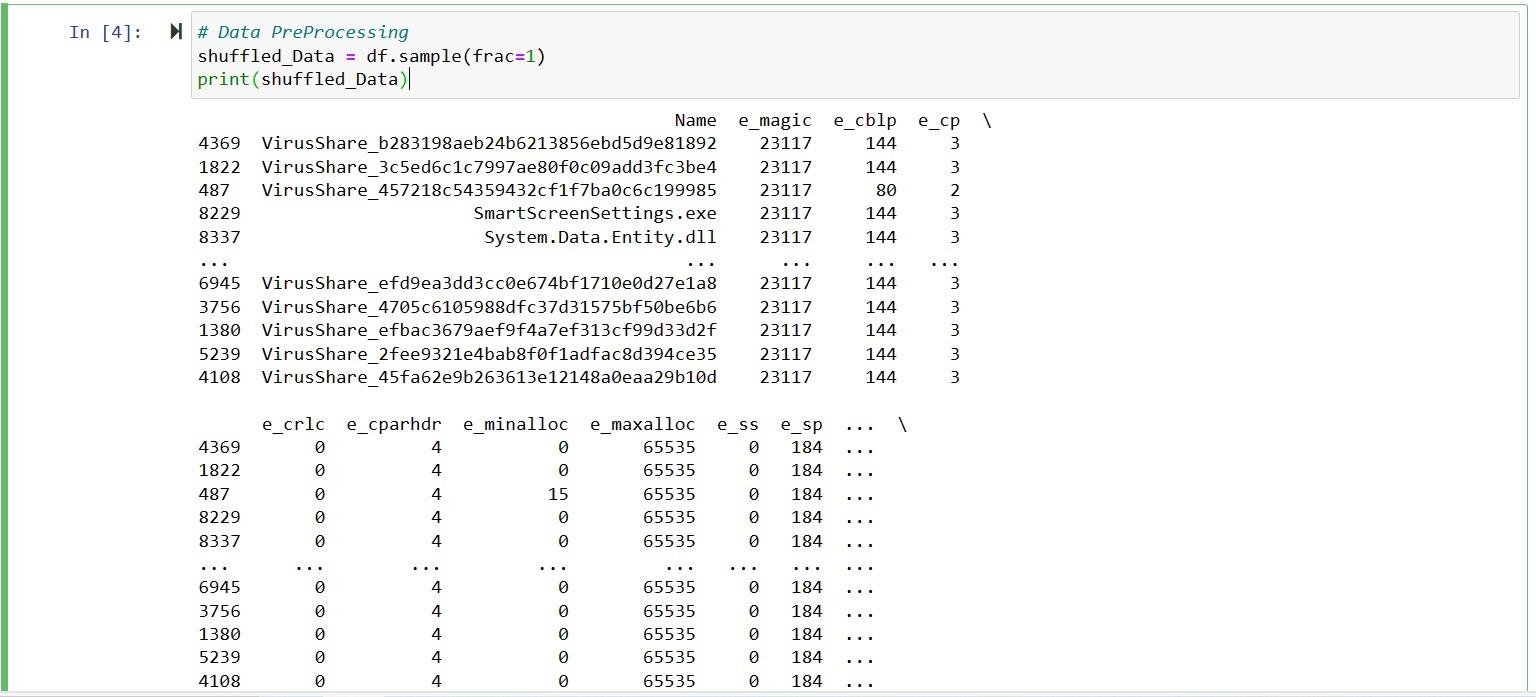
The dataset for our study is taken from Kaggle <https://www.kaggle.com/datasets/amauricio/pe-files-malwares/code?select=dataset_malwares.csv> . It contains both malicious and benign PE files with around 19,000 records and 79 columns.

Since the original dataset had a very large number of records, we reduced it to 10,000 samples for faster processing. The dataset had more malware instances than benign. To balance the classes, we selectively removed some malware records to end up with around 7,500 malware samples while keeping the 2,500 benign samples unchanged. This helped mitigate class imbalance. We then shuffled the reduced dataset using the Pandas library in Python. Shuffling randomizes the order of the records, which helps avoid any sampling bias when creating train-test splits later. The code to perform the balancing and shuffling was as follows:

A computer screen shot of a program code

Description automatically generated

Output:



**Data Feature Extraction:**

The dataset contains 79 features extracted from each PE file. The feature values vary across files, capturing distinctive properties that can differentiate between malware and benign PEs. These features encompass details from various sections of the PE format, including the DOS and COFF headers, optional header, data directories, and imported functions. For instance, features like e-magic and e-cp correspond to fields in the DOS header.

Currently, we are working on compressing the lengthy feature names into two-letter abbreviations for easier reference when building and analyzing models. Assigning concise aliases to features will allow convenient understanding and discussion of the most impactful attributes.

**Challenges Faced & Resolution:**

* **Data Collection**:

The team's first task was to gather relevant data for the analysis that needed to be done for the project. We were having trouble gathering the necessary data qualities since they were sensitive and would give rise to privacy concerns. We are grateful to you, Professor, for giving us access to several trustworthy public data sources where we were able to locate the ideal data set for our study.

* **Understanding complex PE file features:**

Understanding these sophisticated features, their significance, and their relative value for malware identification presents a difficulty. It is challenging to determine which features are most important and should be the focus of attention while developing classification models if one is not domain knowledgeable in the PE file format. Consider how useful it is to have somewhat arcane properties like Subsystem, SizeOfStackReverse, DLL Characteristics, etc. On the other hand, the predictive capability of certain aspects is low, such as Physical Address and e\_magic.Thorough feature selection will aid in identifying the essential PE file characteristics that are most effective in distinguishing between benign and malicious executables. This will result in more efficient and portable models.

* **Identifying dependencies & working synchronously**:

Because there was specific dependency in the duties, the team found it difficult to synchronize and organize after the planning and preparation phase when the work was dispersed among team members. We broke the team up into smaller work groups to solve this issue, allowing those who needed to rely on one another to work closely together. In order to make sure that everyone was aware of the team's progress and working from the same agenda, we also started having stand-up meetings on a regular basis with all of the group members.