**פרויקט גמר - הסבר במילים על הקוד:**

DATASET COLLECTORS:

urlScan.py:

אוסף מידע מהשירות urlscan.io

Sniffer.py:

המחלקה Sniffer משתמשת בScapy לצורך איסוף תעבורת הרשת שעוברת בכרטיס הרשת - מסננת לפי הDATA הרצוי (ככה למשל מחלקת CollectDNS מצליחה לאסוף את רשומות הDNS).

CollectDNS.py:

מחלקה שאוספת רשומות של הDNS

מחוץ לתיקייה TOOLS של DATASET COLLECTORS:

append\_whois\_urlscan.py

אחראית להוסיף מידע מurlscan.io לרשומות dataset של DNS.

collectDns.py:

אוסף מידע DNS מקובצי CSV (אקסל)

collect\_vt.py:

אוסף מידע מvirusTotal

vt\_to\_features.py:

מייצא את המידע הרצוי מvirustotal מקובץ JSON.

TOOLS:

## FeaturesExtraction.py

קוד זה אחראי על ייצוא כל 16 הפיצ'רים שלנו וייצוא שלהן לקובץ CSV.

RatioMatrix.py

TESTS:

test\_ann.py:

neural networksבודק

test\_elm.py:

Exteeme learning machineבודק

test\_features\_permutations.py:

בודק את כל ה permutations של הפיצ'רים שלנו

test\_lr.py:

Logistic regression בודק

test\_robustness.py:

בודק רובסטיות של הפיצ'רים שלנו

test\_svm.py:

Support vector machineבודק

MODELS:

המודלים מוסברים בתזה בצורה מעולה, אני מצרף לכאן הסבר שלהם:2 Models

1) **Logistic Regression (LR)** is a method for analyzing a dataset in which there are one or more independent

variables that determine an outcome. In logistic regression, the outcome is a binary dependent variable, i.e.

it only contains data coded as 1 (TRUE, malicious) or 0 (FALSE, benign).

2) **Support-Vector Machine (SVM)** is a learning algorithm that constructs a high dimensional hyperplane used

to solve the classification and regression problem. Given a set of training examples, and their associated

labels, the algorithm trains a model that assigns new examples to one label or the other, making it a

nonprobabilistic binary linear classifier. In addition to performing linear classification, SVMs can efficiently

perform a non-linear classification using a kernel trick implicitly mapping their inputs into high-dimensional

feature spaces. A radial basis function (RBF) is a function that assigns a real value to each input from its

domain, and the value produced by the RBF is always an absolute value.

3) **Artificial Neural Networks (ANN)** are computing systems that are inspired by biological neural networks.

These take the form of a mathematical model which is used to process nonlinear relationships between the

input, hidden and the output layers and the artificial neurons (also called the processing units) in each layer.

4) **Extreme Learning Machines (ELM)** are feed forward neural networks, which means that the data only go

through the layers in one direction (forward). ELM can be configured with a single layer or multiple layers of

hidden nodes, where the parameters of hidden nodes need not be tuned. These hidden nodes can be

randomly assigned. **We decided to use the ELM model to evaluate our results and compare them too.**