Outlier detection algorithm Isolation Forest – guided by Dr. Liron Allerhand

Goal - Implementing the outlier detection algorithm Isolation Forest in Cython and applying it to measure similarities between malware behaviors.

Project description - The isolation forest algorithm is a well-known anomaly detection algorithm based on iteratively splitting the training data. This splitting defines an anomality metric which is simply the distance of a leaf from the root. While this distance metric is usually implicit and not used for other purposes, we would like to use it to measure the distance between the leafs of the tree. This metric can be trivially generalized to forests by averaging, or otherwise aggregating across trees. We'll extract such distances from isolation forests and use them to find similarities between data points. The algorithm include several statistics save for each node such as the split contribution and several options to customize the model such as choosing the split criteria, the aggregation function, point weight and feature weight, and a soft feature selection limit.

Project implementation -

In the project we will efficiently implement the tree-based algorithm in Cython (an extension of Python) with multithreading. This algorithm will be used to tackle some data science challenges in security like identifying malware and malware types based on the types of system resources that they use.

The project's main topics are - Machine learning, AI, information security, Malware analysis.

The implementation process will include:

- implementing a random decision forest – each tree will be implemented the chosen split criteria, which may be Gini impurity, Lp distance for custom.
- implementing the isolation forest algorithm
- detecting unusual points in our tree
- building roc curves according to our data
- detecting malwares

Project members -

The project will be implemented by Chen hassid and Ron Kozitsa, and will be guided by Dr. Liron Allerhand, a senior data scientist and deep learning expert, currently working in Microsoft and leading this project.