The anti-malware product needs to decide in real time if a file is malware \ benign and act upon it. When the engine encounters a file, it scans it. If it looks suspicious, meta data is extracted and sent to the cloud for decision.

In this assignment you have an example to those suspicious encounters. Your aim is to find a way to predict if unlabeled files are malware \ clean.

This folder contains:

1) Train.csv – training set, each row is a labeled file encounter (1= malware, 0=clean)

2) Test.csv – test set (data without the labels)

3) desc.csv – some explanation about the columns

Expected output:

1. *<you\_name>.csv* (without header) containing Sha256, predicted\_label (0 or 1) for each one of the test samples.
2. Feel free to use any language you find comparable, although Python (+Jupyter Notebook) is preferable.

Tips:

1) No need to use any other data (the file hashes are not real anyways 😝)

2) While the final outcome is important, we would like to see the way you approach the problem.

3) with that said, please do not spend me than 2-3 hours on it. The aim is that you will get to know our data, and we will get to know your technical skills a bit.

4) Oh and please DO NOT share these data or your results on any public platform

**Hope you enjoy it!**

Defender researcher team.