C3i Hub, Indian Institute of Technology Kanpur
HCL HACK IITK 2021
Submission Round: Programming Challenge
16th Jan 2022 – 23rd January 2022

Question 2: Linux malware detection and classification (LMDC)

Instructions:

1. This round consists of two problems, each carrying 100 marks. Question 1 has already been released, this is Question 2. Both the Questions must be answered.
2. Please read the question very carefully
3. The question contains links to download the dataset provided to develop ML-based malware detector and classifier
4. The test dataset (without any label) will be provided 24hrs prior to the final submission deadline.
5. **Do not try to execute any provide files in dataset. They are malware.**
6. When you extract the downloaded files, make sure that file permissions are set to nonexecutable (run "ls -l"). If it is executable then change its permission (run "chmod -x *" inside every folder)
7. You can make multiple submissions, only the last submission will be considered for evaluation, the previous versions will be automatically erased.
8. The last date for submitting both the questions is 23rd January 2022, midnight.

Question 2: Linux malware detection and classification (LMDC)

Maximum Marks 100

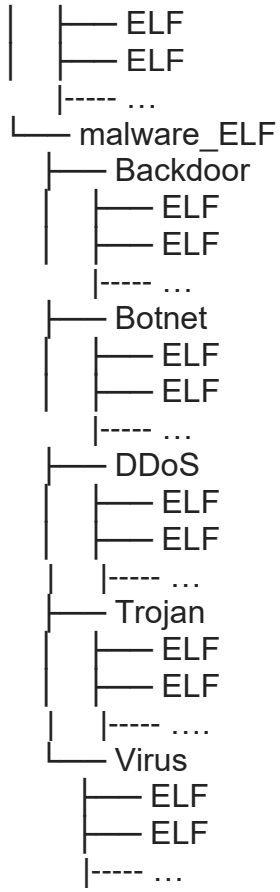
Description: Malware is a malicious software that performs some harmful actions on a target system, such as changing data, deleting files, and exfiltrating vital information available on the target system, change configurations, downloading additional files, making connections to malicious servers etc. Malware threats against Linux have lately increased due to the increasing popularity of Linux, which has been expanding for years, and many popular programs are also available only on Linux. Your objective is to develop a Linux based malware detection and classification tool.

Dataset: You can download the dataset using the following link. The dataset contains **benign ELF** (Executable and Linkable Format) files in benign_ELF folder. In addition, the dataset contains five categories of malware ELF files: Backdoor, Botnet, DDos, Trojan, Virus.

Link: https://drive.google.com/file/d/1HUtqWlEafJmYHKQ5TiNRNk6Nxer10S1K/view?usp=sharing

Tree Structure of malware_traing_data folder:

```
malware_traing_data
├── benign_ELF
```



Steps to follow: These are suggested steps. you are free to use your own methodology.

- **Data collection**: Collect the dataset provided through the link
- **Feature extraction**: Write a script to **e**xtract features from the collected dataset using the script
- **Feature selection:** Select only important features for better training and reducing the prediction time.
- **Classification:** Use machine learning classification models to train the classifiers using extracted features.

Project must fulfill the requirements mentioned below:

The developed tool must have good accuracy, precision, recall, and F1-score for the machine learning model with low false positive and low false negative rates. In case of close competition, a tool having low testing computational cost will be awarded additional marks.

Output of the tool: Any one of the set CLASS: {BENIGN, BACKDOOR, BOTNET, DDOS, TROJAN, VIRUS, NOTELF} in a two-column (FILENAME, CLASS) CSV file. The first column must be file-name and second column the class of the file from the above CLASS as determined by your tool. NOTELF means that the file is not an ELF file.

Note: The test dataset will be provided as folder which will be mixed with ELF files belonging to anyone of the above-mentioned categories. We will mix the test dataset with around 35% benign and around 13% of each malware categories.

Deliverables:

As mentioned before, the test data to generate the result.csv will be provided 24 hours prior to your final submission deadline.

- A result.csv file (a two column csv output file) which is the final result collected by you. We will use this file to compute accuracy by comparing it with the actual labels of the files against what your tool determines. The following deliverable will be used to validate the result.csv file.
- Your LMDC tool must be named LMDC_test.py. This tool takes input from a folder containing ELF files and saves a two-column (FILENAME, CLASS) CSV output file (result.csv). We may consider the computational efficiency of LMDC_test.py to award additional marks. The result.csv file provided by you and generated by us using your submitted LMDC_test.py must be the same.

  python LMDC_test.py absolute_path_for_test_folder
- You have to provide the entire code in a well-structured and commented format which you developed for data processing, feature extraction, feature selection, model training, and testing (whatever you have done) with a README file. This will allow us to verify that you have not done any hard coding of outputs or your model is not biased for the test data.
- All the above three files/folders must be zipped within a single folder named "P2 _Teamname."
- Keep all the folders of training and testing dataset but remove the ELF files (only ELFs, do not change the tree structure folder) to reduce the deliverable size. We will copy the ELFs at appropriate folder during evaluation.

Cheating policy: We will conder the following activities as cheating and any team engaging in any of them will be disqualified.

- Collaboration with other teams
- Intentionally training model to be biased for the test data
- Random output or only one class output
- Anything else that is explicitly not allowed

end