Assignment 2
Malware Detection using PE features

Preface: Malware is short for malicious software, meaning software that can be used to compromise computer functions, steal data, bypass access controls, or otherwise cause harm to the host computer, its applications or data. It is designed to gain access to computer systems, normally for the benefit of some third party, without the user's permission. It's crucial that users know how to recognize the different types of malware in order to help protect yourself, and your business systems, from being compromised.
Malware detection is one of the crucial computer security challenges due to the tremendous growth of new malware and variants of existing malware. But the common commercial antivirus scanners fail to detect zero-day malware. Currently, machine learning is a popular approach to signatureless malware detection because it can generalize to never-before-seen malware families.

Setup

Dataset for the assignment can be downloaded from the below URL.

https://drive.google.com/file/d/1gc4g9qY_NgLEXGWHxg9pRsR18QAUw9v_/view?usp=sharing

Since the dataset contains malware executables, don't try to open the samples in the dataset. Set up an Ubuntu virtual machine (higher than 16.04). There are 3 subparts to this assignment and solution for the first 2 questions should be in a separate python file with the name **"Solution_x.py"** where x denotes the question number.

Question 1

Write a python code to confirm that the file **"Hello_PE"** given in the **"data.zip"** file is a PE executable.

Question 2

Write a python code to print the first 10 bytes in the **".rsrc"** section of **"Hello_PE"** and also extract the value in **"AddressOfEntryPoint"** field**.**

Question 3

Extract the fields from the Windows Portable Executable (PE) format and build your own malware detection system. Solution for this question should be in a folder named **"Solution_3"** and the individual python files should have a meaningful name.

Perform the below tasks to implement the malware detection model.

1. Use the dataset given in the folder **"malware_dataset"** in zipped file "data.zip". The dataset consists of 2 folders 'malware' with 443 samples and 'benignware' with 400 samples.

2. Extract static features from each benign and malicious executable using python utility *pefile* and linux command line utility *strings* to represent the sample. Assign the label as '1' for malware and '0' for benignware.

3. Do research to design good features that will help your machine learning system make accurate inferences. You can start by extracting as many features from the PE header as possible and all the strings with a minimum length of 10. Then, apply a standard feature selection algorithm of your choice to reduce the number of features. You can apply feature selection to the 2 set of features (PE and strings) together or individually.

3. Split dataset into non-overlapping training and test sets, in which the training set consists of 70% of the data (an arbitrarily chosen proportion) and the test set consists of the remaining 30%.

4. Train the machine learning classifiers (your choice. You can try different models and select the best one) to recognize malware using the features you have extracted. Use the inbuilt functions from *sklearn* for implementing the models and calculating metrics.

5. Test your model and calculate the prediction accuracy, precision, recall, false-positive rate, F1-score and confusion matrix.