Challenge 8

In this challenge, you are provided with a dataset that contains static analysis data, extracted from the "pe_sections" elements of Cuckoo Sandbox reports. PE malware samples are downloaded from VirusShare, while PE benign samples are downloaded from different places including PortableApps. In particular two different datasets are available in this challenge, including a labeled dataset (i.e., dataset_labeled.csv) and an unlabeled dataset (i.e., dataset_unlabled.csv). The labeled dataset contains 1,500 samples (1,000 benign and 500 malware), while the unlabeled dataset contains 41,777 samples. The list of features for both datasets is the same and is as follows:

- **size_of_data**: The size of section on disk (in Bytes)
- **virtual_address**: Memory address of the first byte of the section relative to the image base
- **entropy**: Calculated entropy of the section
- **virtual_size**: The size of section when loaded into memory

Prepare a report and submit a PDF file by Tuesday (**12/01/2020 before 6 pm**) considering the below details:

1. Evaluate the performance of 5 different supervised learning algorithms (LinearSVC, NuSVC, SGDClassifier, NearestCentroid, KNeighborsClassifier) using 5-fold cross-validation on the labeled data. For each algorithm, report the average root mean square error (i.e., RMSE).
2. For each algorithm of the above algorithms, do the following steps:
    a. Train a model using labeled data.
    b. Use the trained model to predict the label of each unlabeled data (i.e., creating pseudo-labeled data).
    c. Combine the labeled data and pseudo-labeled data and name it merged_data.
    d. Evaluate the performance of each algorithm on the merged data using 5-fold cross-validation and save the average root mean square error of prediction for each algorithm.

3. Now, for each algorithm, compare and discuss the RMSE obtained from cross-validation on merged data (step 2) versus the one obtained from cross-validation on only labeled data (step 1).