Data Analytics using Machine Learning

Administrative Info	
Mode of delivery	Report and Q&A Session
Due	April 19, at 11:59 pm
Group	4-5 students
Weight	20%

PREAMBLE

Data Analytics (DA) and Machine Learning (ML) applications in cybersecurity are gaining a lot of momentum. Machine learning based classifiers are becoming more ubiquitous in detecting intrusions, malware, phishing, among many other types of attacks.
In this project, you will build a simplified machine learning system that should be capable of handling a cybersecurity dataset to produce a useful application.

OBJECTIVE

- Download and Pre-process a dataset
- Train 3 machine learning classifiers
- Test and explain the results of the three classifiers

ASSIGNMENT

1. Download the dataset from DC Connect and extract the following information from the dataset:
   a. Number of instances
   b. Number of features
   c. Number of instances from each class (malware/benign)
2. Load and split the dataset into two parts (training and testing), assuring that it is balanced.
3. EDA on the data set
4. Create 3 different machine learning classifiers of your choice.
5. Train the three classifiers.
6. Test the three classifiers and generate a confusion matrix, with a confusion matrix plot
7. Discuss the results of the three sets.

DELIVERABLE

- Report
- Code (.py, .ipynb, .html)
- Discussion session (Q&A)
- Submit to your github repository.

Report Rubric (Max 100 Points)

Criteria	**Points**
Title page: Contains name, course section, date, professor name	5
Information about the dataset	10
Proper split of dataset	15
Building the three classifiers	15
Training the classifiers	15
Testing the three classifiers	20
Explaining and comparing the results	5
Q&A session	15
Total	**100**