

Assignment 2

Naive Bayes and Logistic Regression

Dataset: Network Attack - NB15

The raw network packets of the **UNSW-NB 15** dataset was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours.

Tcpdump tool is used to capture the raw traffic (e.g., Pcap files). This dataset has nine types of attacks, namely, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms. The Argus, Bro-IDS tools are used and twelve algorithms are developed to generate totally 49 features with the class label.

[Dataset Link](#)

[Description of all the features](#)

On the above dataset, perform the following tasks.

1. EDA and basic data pre-processing and preparation **[4*0.5 = 2M]**
 - a. Null/Outlier treatment
 - b. Remove non-important features
 - c. Split training and testing data set
 - d. Standardise/Normalise the variables whenever required
2. Train a classification model to classify the 9 different attacks (Use scikit-learn library) **[2+2 = 4M]**
 - a. Using **Naive-Bayes**
 - b. Using **Logistic Regression**
3. Compute **accuracy and F1 scores** for both types of models **[1+1 = 2M]**
4. Apply **feature reduction** techniques **[1M]**
5. Train model on the reduced feature subset using NB and Logistic regression **[1+1 = 2M]**
6. **Comment** on performance of these approaches (NB and Logistic Regression) on reduced dataset (after applying PCA/feature reduction) **[1M]**

Submission Guidelines

1. Submit exactly 1 IPYNB file, with all the tasks, the IPYNB file must be named - **Classification-A2-NBLR-[ID number].ipynb**
2. Follow this template submission Jupyter Notebook - [Template jupyter notebook](#)
3. Not following the submission guidelines will result in penalty.