Assignment – 4
Network Traffic Analysis
Deadline: 11:59 PM, 10th May 2022

One of the defense wing of Indian government is seeking your help in building a highly efficient network traffic analysis system using Machine Learning for their newly build tactical operation room. For building this classifier, you are provided with the NSL-KDD dataset which contains both training and testing data. The dataset can be downloaded from the following link:

https://drive.google.com/drive/folders/1krOVx8jo9fcECVIF-X93wyy0bAzXnSaf?usp=sharing

NSL-KDD is an improvement to a classic network intrusion detection dataset used widely by security data science professionals. The original 1999 KDD Cup dataset was created for the DARPA Intrusion Detection Evaluation Program, prepared and managed by MIT Lincoln Laboratory. The data was collected over nine weeks and consists of raw tcpdump traffic in a local area network (LAN) that simulates the environment of a typical United States Air Force LAN. Some network attacks were deliberately carried out during the recording period. There were 38 different types of attacks, but only 24 are available in the training set. These attacks belong to four general categories:

i. dos-Denial of service
ii. r2l-Unauthorized accesses from remote servers
iii. u2r-Privilege escalation attempts
iv. probe-Brute-force probing attacks

The labeled training data as comma-separated values (CSV) looks like this:

0,tcp,ftp_data,SF,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,
150,25,0.17,0.03,0.17,0.00,0.00,0.00,0.05, 0.00,normal,20

0,icmp,eco_i,SF,8,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,21,0.00,0.00,
0.00,0.00,1.00,0.00,1.00,2,60,1.00,0.00,1.00,0.50,0.00,0.00,0.00, 0.00,ipsweep,17

The last value in each CSV record is an artifact of the NSL-KDD improvement that we can ignore. The class label is the second-to-last value in each record, and the other 41 values correspond to these features:

1 duration
2 protocol_type
3 service
4 flag
5 src_bytes
6 dst_bytes
7 land
8 wrong_fragment
9 urgent
10 hot

11 num_failed_logins
12 logged_in
13 num_compromised
14 root_shell
15 su_attempted
16 num_root
17 num_file_creations
18 num_shells
19 num_access_files
20 num_outbound_cmd

21 is_host_login
22 is_guest_login
23 count
24 srv_count
25 serror_rate
26 srv_serror_rate
27 rerror_rate
28 srv_rerror_rate
29 same_srv_rate
30 diff_srv_rate

31 srv_diff_host_rate
32 dst_host_count
33 dst_host_srv_count
34 dst_host_same_srv_rate
35 dst_host_diff_srv_rate
36 dst_host_same_src_port_rate
37 dst_host_srv_diff_host_rate
38 dst_host_serror_rate
39 dst_host_srv_serror_rate
40 dst_host_rerror_rate
41 dst_host_srv_rerror_rate

The goal is to build a robust classifier that categorizes each individual sample as one of five classes: benign, dos, r2l, u2r, or probe.

Please note:

I. The training dataset contains samples that are labeled with the specific attack: ftp_write and guess_passwd attacks correspond to the r2l category, smurf and udpstorm correspond to the dos category, and so on. The mapping from attack labels to attack categories is specified in the file **training_attack_types.txt available in the same folder as dataset.**
II. Dataset need preprocessing before it can be feed to machine learning models
III. There is a significant class imbalance. You need to find the ways to minimize the imbalance.
IV. You can use any supervised learning model to design your classifier.

Evaluation Criteria:

I. You are required to submit a report and complete codes along with the dataset that you have used for training and testing (.csv).
II. In the report, you should discuss your strategy on building this model in detail, including steps you took for data exploration, data preparation and identification of classification model.
III. You must also discuss the performance metrics that you targeted for improving your results.
IV. There is going to be a relative grading for this assignment.