Weekly Report of Week 1

Name: Punyawat Jaroensiripong

Topic: Dataset Analysis

Description: In the project group, we have an idea to design the machine learning model to predict the attack by the behavior of the network flow (network log). Therefore, the input of the model that we'll design will be network flow and the output will be binary form (0, 1) Since we have arrived at JAIST, Prof. Prarinya has assigned our group to analysis the dataset. I'm choosing NSL-KDD dataset to analyzed. The dataset was improved from KDD Cup 1999 Dataset that came from University of California, Irvine. The dataset is about the flow of network that contain multiple attack in the system which already labeled into multiple class of attack. The attack can categories in to 4 class

1. Dos (is an attack that tries to shut down traffic flow to and from the target system. The Intrusion Detection Systems is flooded with an abnormal amount of traffic, which the system can't handle, and shuts down to protect itself. This prevents normal traffic from visiting a network.)
2. Probe (is an attack that tries to get information from the network system.)
3. U2R (is the attack that start from user account and tries to get the super-user (root))
4. R2L (is the attack that can gain the root priority)

In these 4 classes, it can be divided in to 39 attack such as worm, land, teardrop, spy, rootkit etc. And in this dataset contain 43 features, 40 features are numerical data, and 3 features are string data. In this case, I must change those 3 features (protocol_type, service, flag) into numerical data by using OnehotEncoder in sklearn preprocessing library. So, the feature will be change in to 124 features. At the end of the preprocessing, I change the outcome of the dataset which is the categories of the attack into binary form (0, 1)

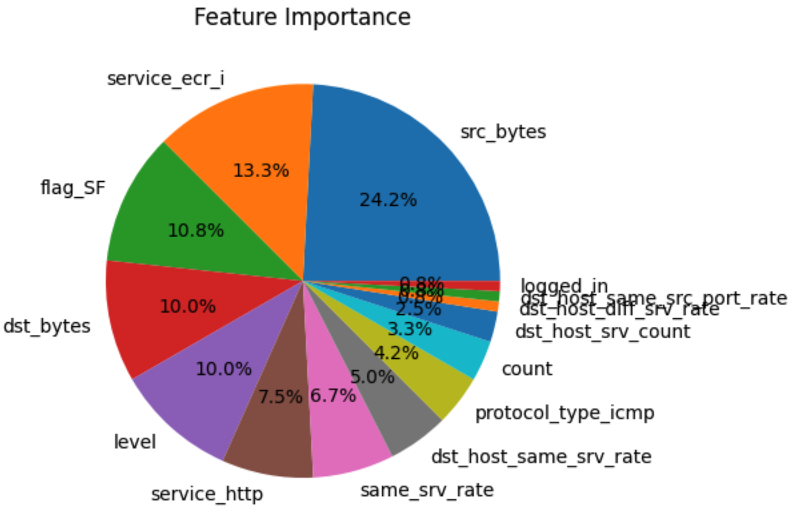
In the part of data analysis, I must analyze the feature importance by using 4 methods.

1. Feature_importances_ (The build-in feature of the machine learning model in sklearn)
2. L1 Regression (Lasso)
3. Correlation analysis
4. Permutation importance

Each of the methods will show the weight of each feature that have influence on the outcome of the data. So, I use top 5 of each method and labeled into priority (1-5) and make a summation for every method together. Conclusion, the top 5 of the features that have influence on the outcome is:

1. src_bytes (Number of data bytes transferred from source to destination in single connection)
2. service_ecr_i (ecr_i service)

3. flag_SF (Normal establishment and termination. Note that this is the same symbol as for state S1. You can tell the two apart because for S1 there will not be any byte counts in the summary, while for SF there will be)
4. dst_bytes (Number of data bytes transferred from destination to source in single connection)
5. level (Difficulty level)



Next week, I have been assigned the assignment to find the state of art of the machine learning model or the paper work that will be fit into the NSL-KDD dataset.