Task 4: Evaluating Classification Models

Diana Pham

Progress Report:

For Task 4 of my final project, I evaluated the classification models that were built in the previous task by constructing a confusion matrix and calculating their Accuracy, True Positive Rate/Recall, False Positive Rate, Precision, F1 Score, Percentage of Incorrectly Classified, and Execution Time. The execution time is measured during the training time (the time measured between the classification process starting and ending). Overall, in terms of Accuracy and F1 Score, the models that performed the best across all feature subsets were Random Forest and Random Tree. However, for the Naive Bayes classifier, it produced NaN values which may be due to the 10-fold cross validation during the training period since the occurrence of some attacks are relatively small, therefore, some folds do not contain those types of attacks. These evaluation metrics are looking at the dataset as a whole, but within the next week, I will look deeper and evaluate how well each model can classify specific attacks for each feature group.

Challenges:

While following the implementation of the original research paper titled, CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection, I encountered the following challenges:

1. Since they built their Bayesian Network in Weka and did not describe the network structure in the paper, I was still unable to recreate it in python due to memory issues while processing the data.
    a. Instead of randomly selecting 20% of the dataset, I will try looking at the data's traffic and associated attacks to select a single day to focus on and see how it applies.
    b. If this still does not work, then I will do some research to figure out what a suitable replacement for this model on the data can be.
2. Since the researchers were using Weka instead of Python, I had to research how to implement the J48 Classifier in Python which works in Google Colab, but is now not running properly on my local machine.
    a. This may be due to the java path for the JAVA_HOME variable on my local machine since I got an error that said the path didn't exist so I will investigate this further and aim to have everything running on my local machine by my final presentation next week.