**CSCI 5622 - Data & Baseline Report**

**Team Members**: Byron Becker

Kevin Holligan

Saim Khan

Ishita Srivastava

Victoria Slattum

**Project Scope:** Based on the data we have retrieved, we have decided to solely look at various supervised classifiers and evaluate how they predict the category of certain network traffic based on the networking data from our dataset. The goal of our model is to block types of networking traffic when they originate from unusual sources (e.g. embedded medical equipment accessing bittorrent data).

**Data:** The dataset we are using for this project was prepared by MIT Lincoln Labs by simulating a typical U.S Air Force local area network. The dataset contains millions of tuples and each tuple is labeled as either normal or a particular kind of intrusion (attack). The four major types of attack categories are:

1. DOS (denial of service)
2. R2L (unauthorized access from a remote machine)
3. U2R (unauthorized access to root)
4. Probing (Surveillance)

Each record has a total of 41 features. The features are of the following three types:

1. Basic features of individual TCP connections
2. Content features suggested by domain knowledge
3. Traffic features computed using a two second time window

For more information on dataset refer to <http://kdd.ics.uci.edu/databases/kddcup99/task.html>

For reference, the training data contains 494,021 training examples, and the test dataset contains 311,029 data points.

The training dataset contains data that result in 23 different classifications of attacks, whereas the test dataset contains 38 different classifications of attacks, so we hope that we will have greater success in classifying these different new attacks than in the baseline Naive Bayes model. Looking into the data more we can see that 98.2% of the training data has a result of one of 3 classifications - “smurf”, “neptune”, and “normal”, and 90.9% of the test data falls into these classifications. Smurf and Neptune are both denial of service attacks, and normal would represent no attack. On both the test and train datasets, the “normal” data represent between 19-20% of the data, whereas then 79-80% of the data represents some form of attack. We therefore feel unlike regular network traffic, that the dataset we have been given enough “attack” examples, although we may need to alter whatever bias it may have based on the fact that normal network traffic is not nearly as malicious.

**Baseline:** The baseline classifier we used for the data is a Naïve Bayes model. We feel that this is a good baseline because it assumes all features in detecting attack traffic are independent, and therefore doesn’t account for the associative impact of various features on the traffic classification. For our baseline model, we used Sklearn’s builtin GaussianNB and MultinomialNB frameworks to fit both the continuous and discrete features respectively. To do this, we parsed the data into a discrete or continuous subset, formatted the respective subsets and ran the discrete feature data through CountVectorizer to obtain counts that we could run through the MultinomialNB model. We then took all of the class assignment probabilities from both models and fit a new Gaussian model on this training data. We then repeated the process with the test data on the trained Gaussian model to the following results.

Train accuracy: - 0.94740 – 94.7%

Test accuracy: 0.77363 – 77.3%

This will be our Baseline which we will compare against the results of other models to determine their relative improvement throughout the rest of the project.