# ANALYSIS REPORT FOR THE KDD CUP DATA FROM 1999

## Overview

The data used for the analysis is the KDD Cup dataset. The dataset contains 494020 rows and 42 columns. Out of these 42 attributes, 41 attributes can be classified into four different classes as discussed below:

1. Basic (B) Features are the attributes of individual TCP connections
2. Content (C) features are the attributes within a connection suggested by the domain knowledge
3. Traffic (T) features are the attributes computed using a two-second time window
4. Host (H) features are the attributes designed to assess attacks which last for more than two seconds.

The tasks performed on this dataset include:

1. Data cleaning and preparation for statistical/exploratory analysis
2. Statistical analysis
3. Exploratory analysis; Insights generation and visualization, and;
4. Development of a prediction model (Intrusion Detection System).

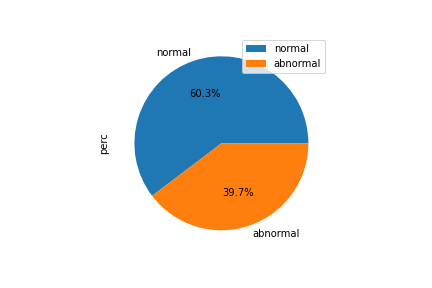
## Data Cleaning Methodology

Data cleaning is one of the most important steps in data analysis and modeling pipeline, and this is evident in the fact that a model and analysis can only be as good as the quality of the data cleaning. The first step taken in the cleaning process was to collect the appropriate column names from <http://kdd.ics.uci.edu/databases/kddcup99/kddcup.names> and rename the columns properly. The main issue faced while performing this was ensuring that each column was paired with its respective and appropriate name. Thereafter, the columns were checked for missing values using the dataframe.info () method. The result of this query shows that there was no missing value in the dataset. Duplicate rows in the dataset were also dropped to avoid redundancy. After dropping the duplicate rows, the number of records(rows) reduced from 494020 to 145585, a proof of the high level of redundancy of the dataset. The formatting of the dataset was lastly checked by using the df.unique () method to check the unique values in the object datatype columns; this is to ensure that the same unique value is not represented with different names. The dataset checks out to be well formatted.

## Statistical Analysis

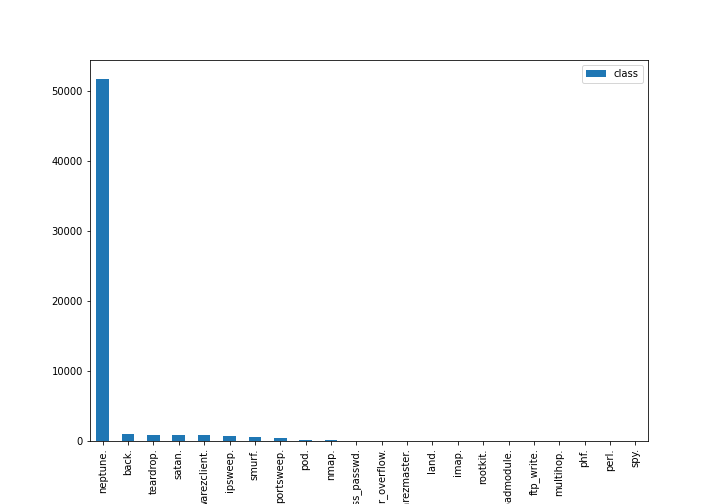
A descriptive statistical analysis was performed on the dataset using the df.describe () method. The correlation between the columns was also checked using the dataframe.corr () method and the result of this query reveals that many of the columns are highly correlated. Lastly, the skewness of the dataset was checked (The skewness reveals the degree of normal/ Gaussian distribution of the columns). 17 of the columns are highly skewed.

## Exploratory Analysis (Insight Generation and Visualization)



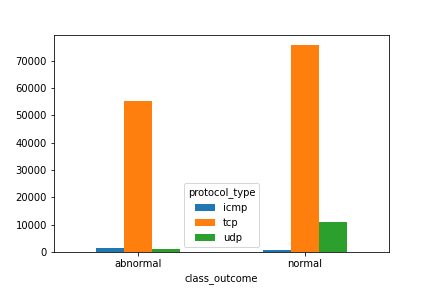
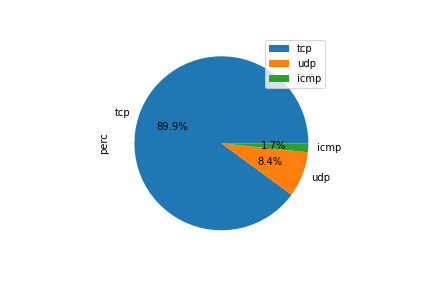
Figure

The pie-chart in Fig. 1 above shows the distribution of the connection types of the dataset. 60% of the connections are normal while 40% of the connections are attacks (abnormal).



Figure

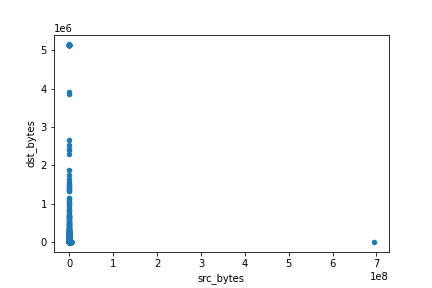
The Bar chart in Fig. 2 above shows the distribution of the different attack connection types. There are 22 different forms of attack types. The highest attack type is Neptune with a frequency of 51820 (89.7%).



Figure

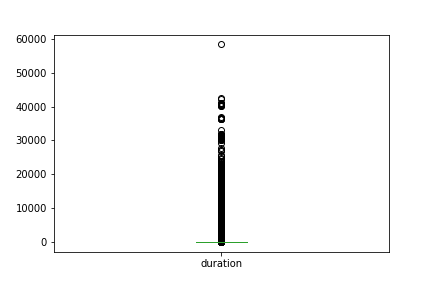
The pie chart above in Fig. 3 shows the distribution of the protocol types in the record. 89.9% of the record is tcp protocol type, 8.4% is udp and 1.7% is icmp. The grouped Bar chart explains the distribution of the different protocol types relative to the connection types, majority of the protocol type in both normal and attack connections are tcp while there are more udp occurrences in normal connections than attack connections.

The analysis of the service column reveals that there are 66 different services, with the highest used being ‘http’ at 42.6% which is followed by ‘private’ at 33.6%. The least used services are pm\_dump, red\_i, tftp\_u.



Figure

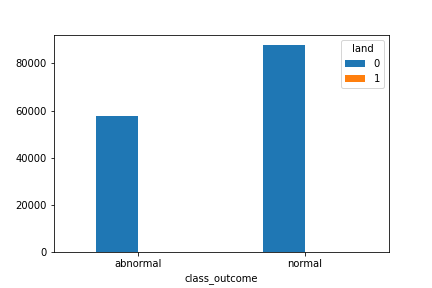
The chart above is a scatter plot showing a comparison of the total bytes going out from the source IP address to those going to each destination IP address. The maximum source byte is 6.933756e+08 while the maximum destination byte is 5.155468e+06.



Figure

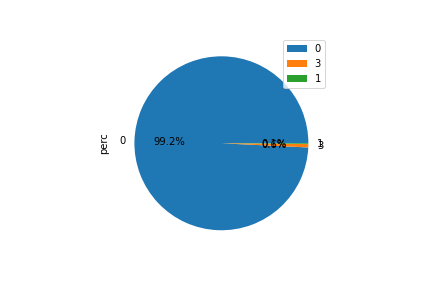
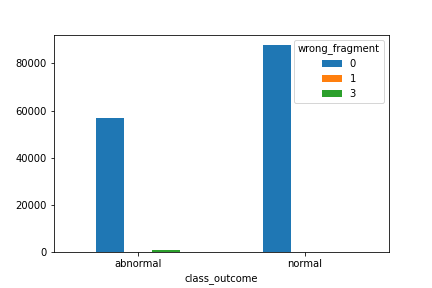
The box plot in Fig. 5 above gives a descriptive statistics (summary distribution) of the duration column. The maximum connection time recorded is 58329 seconds and the mean connection time is 132 seconds.

The analysis of the land column shows that 99% of the connections are not from/to the same host/port. The grouped bar chart below shows the relationship between the connection types and land (1 if connection is from/to the same host/port; 0 otherwise).



Figure

Of the 20 connections that are from/to the same host/port, 19 are of attack connection class.



Figure

The pie chart above gives information on the distribution of the number of ``wrong'' fragments. 99% of the connection records contain no wrong fragment. There are 910 (0.6%) with 3 wrong fragments and 211 (0.1%) with 1 wrong fragment. The grouped bar chart shows that all of the connection records with wrong fragments are attacks.

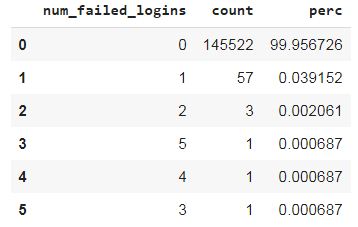
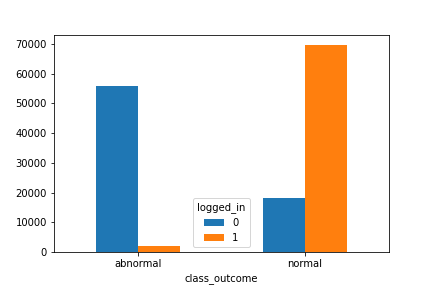


Table 1

Figure : 1 means successfully logged in; 0 otherwise

The table above shows the number of failed log-in attempts with 99% of the connection records having zero failed log-in attempts. From the grouped bar chart of Fig. 8, we see that majority of the successful log-in were normal and majority of the unsuccessful log-in were attacks. However, out of the total successful log-in, 1936 connections were attacks while out of the failed log-in, 18214 connections were normal.

Root shell is a **shell** with administrator privileges. There are 145530 connections without root shell being obtained and 55 with root shells obtained, 32 connections out of the 55 ended up being attacks as shown in Table 2 below.

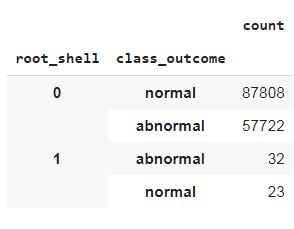


Table 2: 1 if root shell is obtained; 0 otherwise.

Analysis of the ‘is\_host’ and ‘is\_guest’ log-in columns showed that all of the connection records were not hosts and 99% were not guest log-in.

## Conclusion and Surprises

In this report, we statistically analyzed the KDD data set. The analysis showed that there are two important issues in the dataset which are *high redundancy in the row records* and *high level of correlation amidst many columns*. Several insights have also been generated from the dataset and these insights have been presented with the aid of various visualization charts. A predictive model of an intrusion detection system was also built, with majority of the classifiers attaining 99% accuracy. The table below shows the accuracy attained on several models after training and testing.

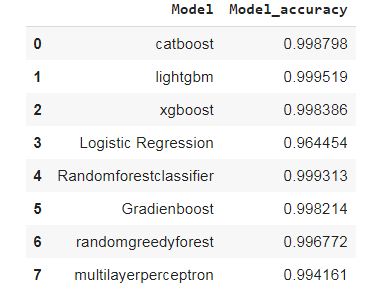


Table 3

The surprises encountered during the process of analysis were from the analysis of the hot, su\_attempted and ‘is\_host’ / ‘is\_guest’ log-in columns. It is believed that the higher the hot indicator, the more likely the connection record is an attack. However, there are connection records with 22-24 hot indicators with all being normal connections. Also it is expected that the su\_attempted column will have just 2 unique values which are 1 and 0 (1 if ``su root'' command attempted; 0 otherwise), but however, a third unique value was discovered. Analysis of the is\_host log-in column shows that none of the connection records were from hosts *meaning* they were all guests, however on analyzing ‘’ is\_guest’’ column it was discovered that 99% are not guests; this two results hence appear highly contradictory and surprising.