VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Electronics Engineering (SENSE)

J COMPONENT –REPORT

COURSE CODE / TITLE	ECM3001 – Data Analytics and Visualization		
PROGRAM / YEAR/ SEM	B.Tech (ECM)/III Year/ FALL 2020-2021		
LAST DATE FOR REPORT SUBMISSION			
DATE OF SUBMISSION			
	REGISTER NO.	NAME	
	18BLC1092	Shaurya Gupta	
PROJECT TITLE	**Intrusion detection using Machine learning algorithms**		
COURSE HANDLER'S NAME	PROF.SYED IBRAHIM	REMARKS	
COURSE HANDLER'S SIGN			

Certificate

This is to certify that the Project work titled "**Intrusion detection using Machine learning algorithms**" is being submitted by Shaurya Gupta (18BLC1092) for the course Data Analytics and Visualization is a record of bonafide work done under my guidance. The contents of this project work, in full or in parts, have neither been taken from any other source nor have been submitted to any other Institute or University

SYED IBRAHIM

Guide

ABSTRACT :

There has been a tremendous growth of internet based traffic which has led to a wide varieties of vulnerabilities in the corporate networks. This leads to a lot of intrusive attacks which affects the smooth operation of network infrastructure by consuming corporate resources and time. Therefore an efficient way of protection would be to identify the patterns of these attacks, so we can mitigate them to enhance productivity. Intrusion detection system (IDS) is a major part of the network traffic security. Machine learning plays a big role in making IDS automated and efficient over time. Hence, we will be displaying some of those machine learning algorithms for detecting intrusive traffic and comparing them to find out which of them has leads to the best performance. We will be using the NSL-KDD as the dataset for the said comparative study. A good Intrusion detection performance has high accuracy level and minimal error rate which can be determined by precision, recall, F1 score and ROC curve. Four binary classifiers: Stochastic Gradient descent, Random Forests, Logistic regression, Support Vector Machine are tested and validated to come up with the results. The results demonstrated that Random Forest Classifier outperformed the other four classifiers.

Index Terms – IDS, NSL-KDD

INTRODUCTION :

Intrusion is major security problem in the various sectors of our society and causes a severe issue of security breach since, a single instance of intrusion can steal or delete data from the computer and network systems in just a few seconds. And Nowadays, internet-based applications and dependency of cloud-based services are widespread more than ever, almost as a necessity. Organisations and corporates have been focusing on their core businesses and moving their IT services in the cloud. There are many other reasons to why companies are pushed to rely on internet-based services. In a similar trend, the growth of malicious traffic has also exponentially increased. Targeted organisations and companies are being attacked by different techniques by various organised cyber terrorists and script kiddies. Protecting, detecting and managing those intrusion attacks are challenging and costly, as any organisation strive to comply with different standards.

By convention a secure network infrastructure are recommended to have firewalls, intrusion detection and prevention systems, and web content filters to protect internal systems from any type of breach or attack. The advancement in attacking techniques and the intelligence of a competent intruder makes it very hard to fully protect any private or sensitive data from theft , disclosure and denial of service attacks. Researches are studying various machine learning methods to improve the efficiency of intrusion detection systems in order to fully prevent those sort of attacks.

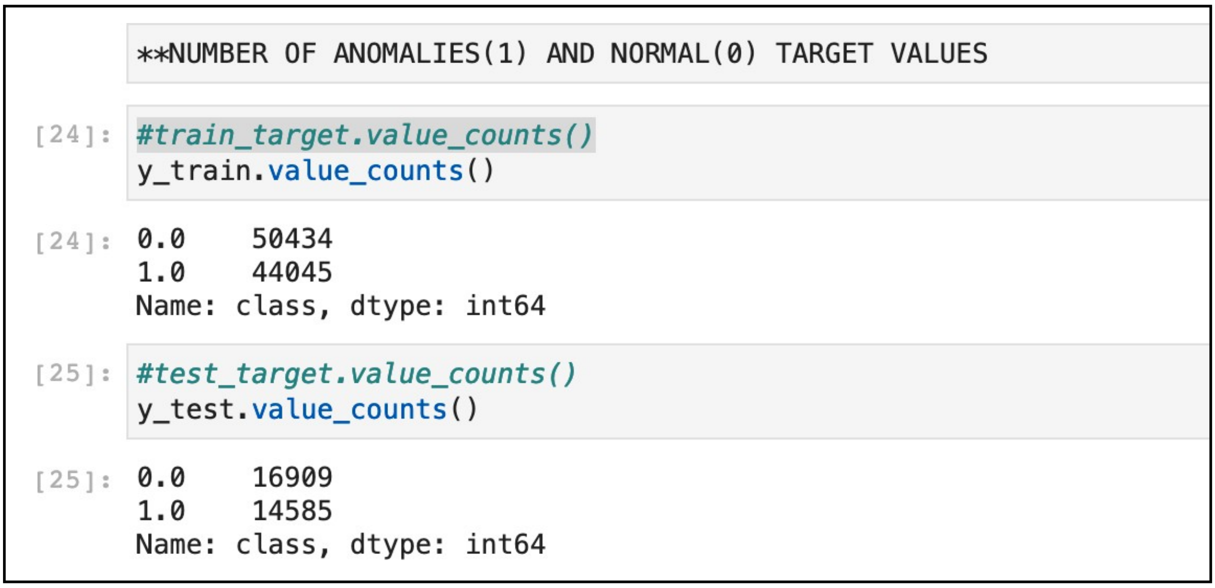
This project targeted IDS analysis with carious machine learning binary classifiers by using the NSL-KDD dataset. Although NSL-KDD dataset is not the perfect representation of the existing system’s traffic, but it is used in various research papers and projects since there is lack of any other public dataset.

DATASET DESCRIPTION:

The selection of dataset has a huge impact on the performance of the machine learning algorithms that we will be applying. The NSL-KDD dataset was selected as it has resolved the problems that are present in the other datasets (like the KDDCUP'99) such as redundant records, which were basically removed. This enables the classifier to produce an unbiased result.

This project uses the training dataset and test dataset that are made up of two target values, normal which is a 0 and an anomaly which is a 1. The known attack types are grouped as anomaly traffics while the remaining traffics were categorised as normal traffic. The original NSL-KDD dataset had 41 features and a label. Pre-processing was done to remove any categorical data (NaN: Not a number), which were present in three features and change those values into numbers before applying the classifiers by using the one hot encoding method. The three features are as follows: 'protocol_type' has 3 unique categories, 'service' has 70 unique categories and 'flag' has 11 unique categories.

After one hot encoding was done to the dataset , the number of features increased to 122 and a label (from 41 features and a label), which is assigned for each instance. The total instances in the dataset are 125,973 that were split into train dataset and test dataset. This was divided, hence train dataset had 100,778 instances and test dataset had 25,195 instances. The figure below depicts the number normal and anomaly instances count in the train and test datasets.



```
**NUMBER OF ANOMALIES(1) AND NORMAL(0) TARGET VALUES

[24]: #train_target.value_counts()
      y_train.value_counts()

[24]: 0.0    50434
      1.0    44045
      Name: class, dtype: int64

[25]: #test_target.value_counts()
      y_test.value_counts()

[25]: 0.0    16909
      1.0    14585
      Name: class, dtype: int64
```

METHOD OF ANALYSIS:

This study employed various techniques of classification and analysed the NSL-KDD dataset in numerous ways. Different performance measures were calculated and compared for insight. The performance measures used were Precision, Recall, F1 score, Receiver operating characteristic (ROC) curve. The precision is calculated by dividing the number of true positive (TP) instances over the sum of the number of true positive and false positive (FP) instances. The recall is calculated by dividing the number of true positive instances over the sum of the number of true positive and false negative (FN) instances. The equation for calculating F1 score is as follows:

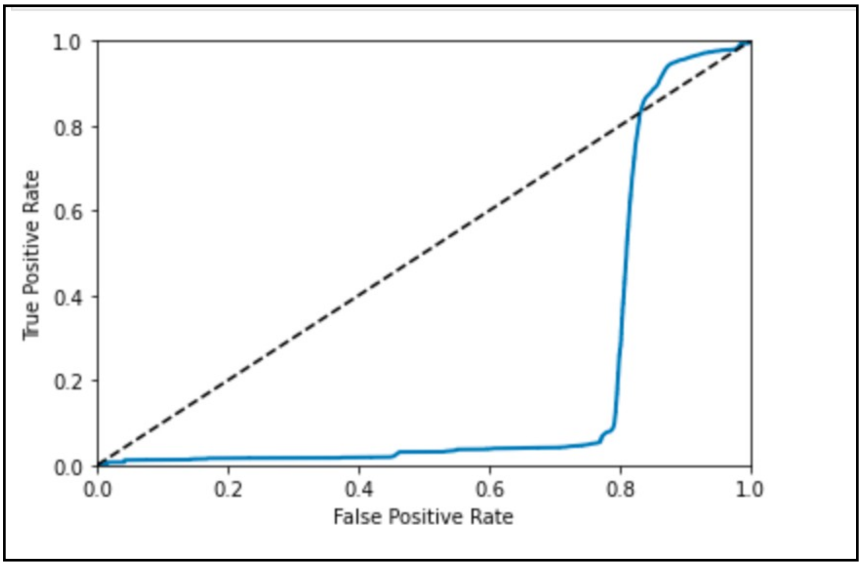
F1 score= 2TP/(2TP+FN+FP)

The classifier can get a high F1 score only if both recall and precision are high. The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR). It is used to illustrate the diagnostic ability of the binary classifier. The performance using the ROC curve can be determined by calculating the area under the curve.

MACHINE LEARNING ALGORITHMS:

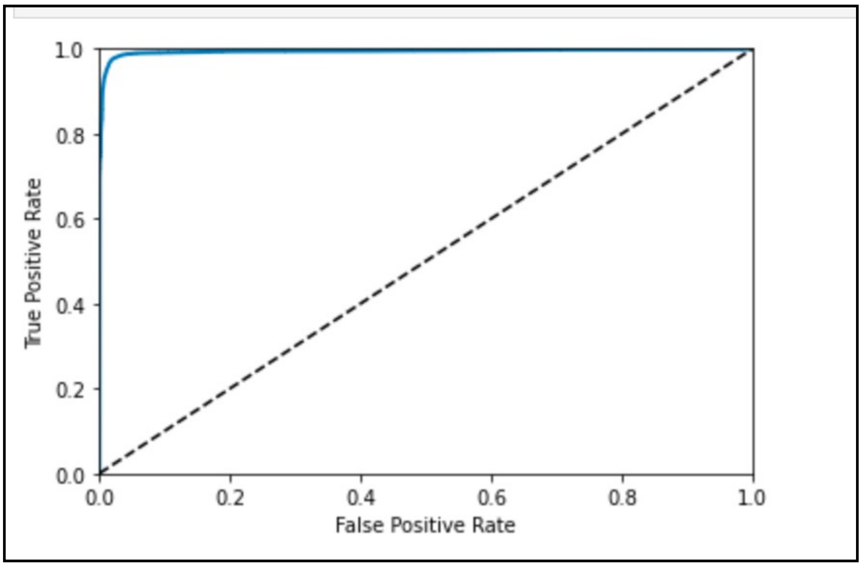
The sklearn library was of great help in our study as it was used import the various machine learning models present in our project.

Stochastic Gradient Descent (SGD) also known as incremental gradient decent classifier has advantages of handling very large dataset and dealing with training instances independently. This classifier demonstrated poor performance initially because features have the large gaps between minimum and maximum values. The ROC curves shows poor performance.



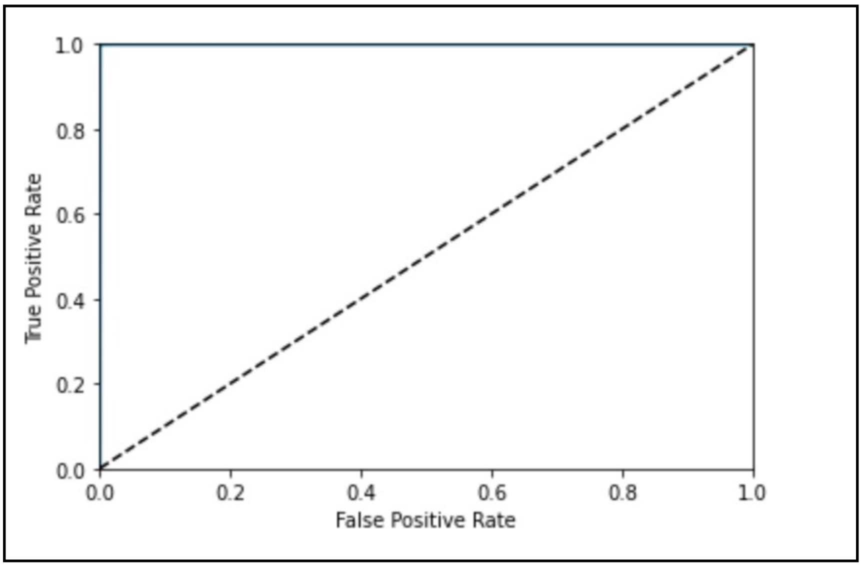
SGD ROC curve.

However, by applying standard feature scaling, the problem was solved. The ROC curve for SGD technique is shown in figure below.



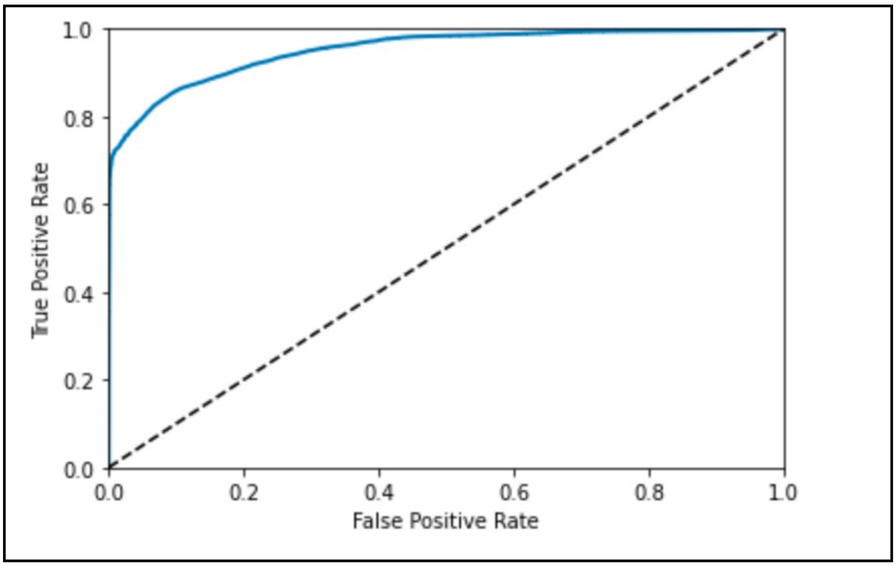
Improved SGD ROC curve.

Random Forests classifier works by training many Decision Trees on random subsets of the features, then averaging out their predictions. Random Forests classifier demonstrate good performance. The accuracy level achieved in Random Forests is near to perfection. Figure below depicts ROC curve of Random Forests classifier.



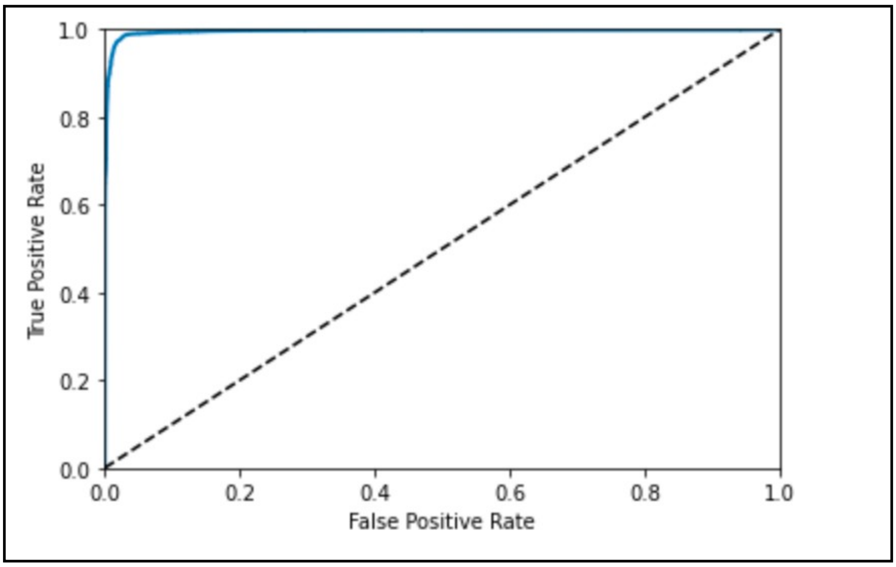
Random Forest ROC curve.

Logistic Regression is one of the regression algorithms that can also be used for classification. Logistic Regression also called Logit Regression is used to estimate the probability that the instance belongs to a particular class. This classifier had less performance compared to the other classifiers applied in the dataset. Figure below shows the ROC curve of Logistic Classifier.



Logistic regression ROC curve.

A Support Vector Machine (SVM) is a powerful and versatile Machine Learning model, capable of performing linear or nonlinear classification and regression. SVM is well suited for classification of complex but small or medium sized datasets. The result of applying SVM classifier in NSL-KDD dataset demonstrated a good performance and is comparable to the result obtained in case of Random Forests classifier. Figure below shows the ROC curve of Support Vector Machine.



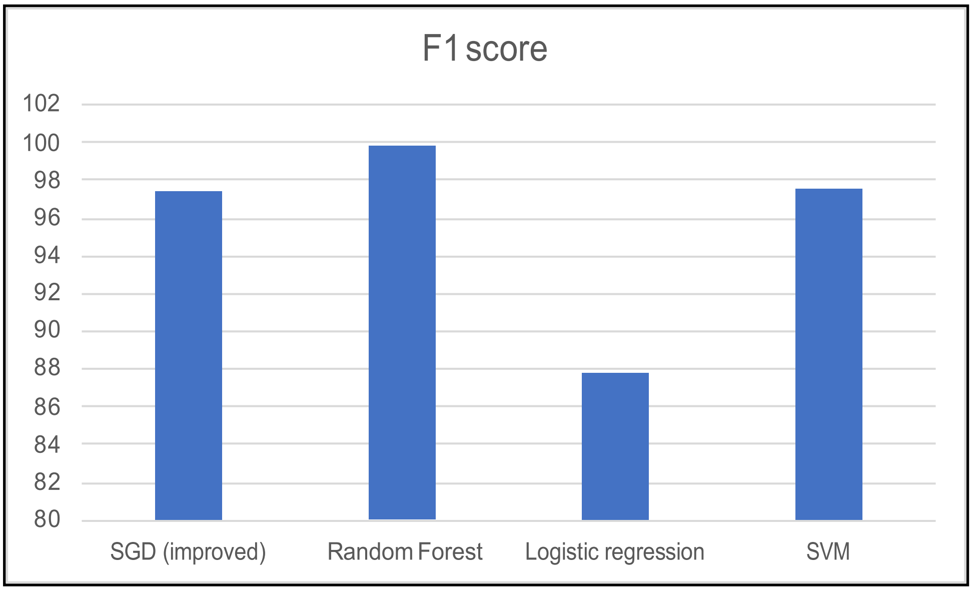
SVM ROC curve.

EXPERIMENTAL RESULT:

The experimental results have been shown in the table below, it includes a summary of precision, recall and F1 score.

Score Type	SGD (improved)	Random Forest	Logistic regression	SVM
Precision	0.9805	0.9993	0.8757	0.97779
Recall	0.9693	0.9976	0.8766	0.9730
F1 score	0.9748	0.9984	0.8787	0.9755

The accuracy results of the four classifiers are shown in the graph below.



CONCLUSION :

The comparative study of different machine learning models on intrusion detection systems using the NSL-KDD dataset was conducted. The dataset was pre-processed using the one hot encoding method to convert the categorical data present in the dataset into numerical values. The study was conducted on four classifiers namely, Stochastic gradient descent, Random Forest, Logistic regression, Support vector machine. The results clearly depicted that Random Forest outperformed the other three classifiers. The overall results of Random Forest classifier are near to perfection. In our future work, we plan to integrate and analyze various artificial neural networks to classify different class types or attacking techniques in intrusion detection systems dataset.

REFERENCES:

[1] IFTIKHAR AHMAD, MOHAMMAD BASHERI, MUHAMMAD JAVED IQBAL, AND ANEEL RAHIM, "*Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection"*

[2] GOOGLE MACHINE LEARNING COURSE, https://developers.google.com/machine-learning/crash-course/training-and-test-sets/splitting-data

[3] SKLEARN DOCUMENTATION, https://scikit-learn.org/stable/model_selection.html

[4] UNDERSTANDING AUC – ROC curve, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

[5] Why One-Hot Encode Data in Machine Learning, https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/