# Intursion Detection Data

## Abstract

Coburg Intrusion Detection Data Sets (CIDDS) is a concept to create evaluation data sets for anomaly-based network intrusion detection systems. The goal of this project was to use classification models to predict the suspicious or normal or unknown in the network in order to help improve prevent cyberattacks and maintain stability. leveraging categorical feature engineering along with a random forest model to achieve promising results for this multiclass problem. After refining a model, I built an interactive visualization and communicate my results using the seaborn library.

## Problem Statement

This project aims to create a classifier to identify the suspicious or normal or unknown. The external server attack logs are the most interesting part. These days sophisticated systems are being built to encounter Server attacks and suspicious content. Working in building a model to predict an attack session.

## Design

This project uses data provided by Kaggle. This data contains features that Server-Logs Suspicious ,These Logs are categorized as suspicious or normal or unknown, Server-Logs are a very large suspicion of this dataset.  This project follows five stages.

The five stages adopted for this project are:

1. Problem Definition
2. Data Collection
3. Exploratory Data Analysis
4. Modeling
5. Evaluation

## Data

Dataset I obtained was from Kaggle https://www.kaggle.com/kartikjaspal/server-logssuspicious. this dataset is uploaded to check what factors contribute to server anomalies. Dataset: 172838 rows, 16 columns, Included columns: Time and duration of attack, Source and destination IP, Packets, bytes, flows, and flags, Type, ID, and label/class I am planning to use deep learning model such KNN, RNN and

Logistic Regression. I will plan to conduct this model in which differentiate between normal and suspicious attacks.

## Algorithms

Based on the initial analysis, it is evident that both text and numeric data is to be used for final modeling. Before data modeling a final dataset is determined. This project will use a dataset with these features for the final analysis:

This project will use a dataset with these features for the final analysis:

1.Duration
2.Dst IP Addr
3.Src IP Address
4.Packets
5.Bytes
6. Flags (P, R, F)

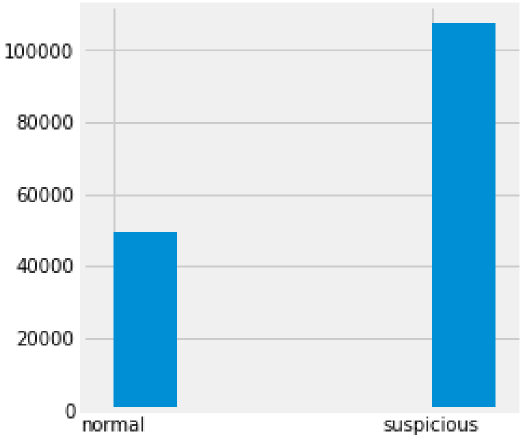
The algorithms and techniques used in the project are:
1.K Nearest Neighbor(KNN)
2.Multi-layer Perceptron classifier
3.Gaussian Naive Bayes
4.Support Vector Machines
5.Decision Tree
6. Random Forest

Random Forest is the baseline model, and it is used because The best model and the high Accuracy.

## Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting
- Tableau for interactive visualizations

## Communication


normal
suspicious