

**Summer Internship Report**

**Intrusion-Detection-System-Using-Machine-Learning**

**21-MAY-2023**



SML2029 Research and Consulting Private Limited

Hyderabad, Telangana, India

**Done By: -**

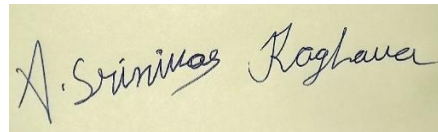
**A Srinivas Raghava**

**Guided by: -**

**Dr. Ajeet Singh**

# ACKNOWLEDGEMENT

I have carried out this summer internship project and completed the report at SML2029 Research and Consulting Pvt. Ltd., Hyderabad, Telangana, India under the guidance of Dr. Ajeet Singh. I'm grateful to him for his assistance, technological inputs and overall guidance during my internship at this company.

A handwritten signature in black ink on a light yellow background. The signature is written in a cursive style and reads "A. Srinivas Raghava".

**Signature of candidate**

**Name: Adduri Srinivas Raghava**

**Date: 19-MAY-2023**

# CERTIFICATE

This is to certify that the internship project entitled, **“Intrusion Detection System Using Machine Learning”**, being submitted by **Mr. Adduri Srinivas Raghava**, is a record of bonafide work carried out by him under my guidance and supervision.

The duration of the internship was 21.04.2023 to 21.05.2023. During this internship period, **Mr. Adduri Srinivas Raghava** was consistent, sincere and showed his good learning capability.



**Signature of Authority**

Dr. Ajeet Singh

(Managing Director)

SML2029 Research and Consulting

Pvt. Ltd., Hyderabad, (T.S.), India

## About the company

**SML2029 Research and Consulting Pvt. Ltd., located in Hyderabad**, is a Deep Tech start-up, founded and registered as Pvt. Ltd. legal entity with Ministry of Corporate Affairs, GOI in Feb 2022. In Apr 2022, the company got registered and recognized with Ministry of Micro, Small and Medium Enterprises (MSME/UDYAM), Government of India. The core competencies of this company are in,

- (i) Promoting technology
- (ii) Knowledge based innovative ventures
- (iii) Research in cutting-edge technologies.

Company is led by an ambitious, energetic and experienced professionals with a unique set of expertise. The core activities/verticals of the company are in:

- a. R&D and Prototyping
- b. Consulting and Software Development
- c. Training/ Internships.

The thrust areas of the company are: Artificial Intelligence Paradigms, Data Science, Quantum Machine Learning, Secure Machine Learning, Cyber Security, Full Stack Development.

## **OUTLINE**

1) Project Statement .....	5
2) Area or Domain of work .....	5
3) Introduction .....	5
4) Approach summary .....	6
5) Experimental Procedure	
i) Data set used .....	8
ii) Language and Frameworks used .....	8
iii) Computational python libraries used .....	8
iv) Statistical Performance .....	13
v) Obtain Result .....	13
6) Our Consolidated Result .....	16
7) Usefulness this problem statement in real time deployment in industries.	16
8) Future scope of this work .....	16

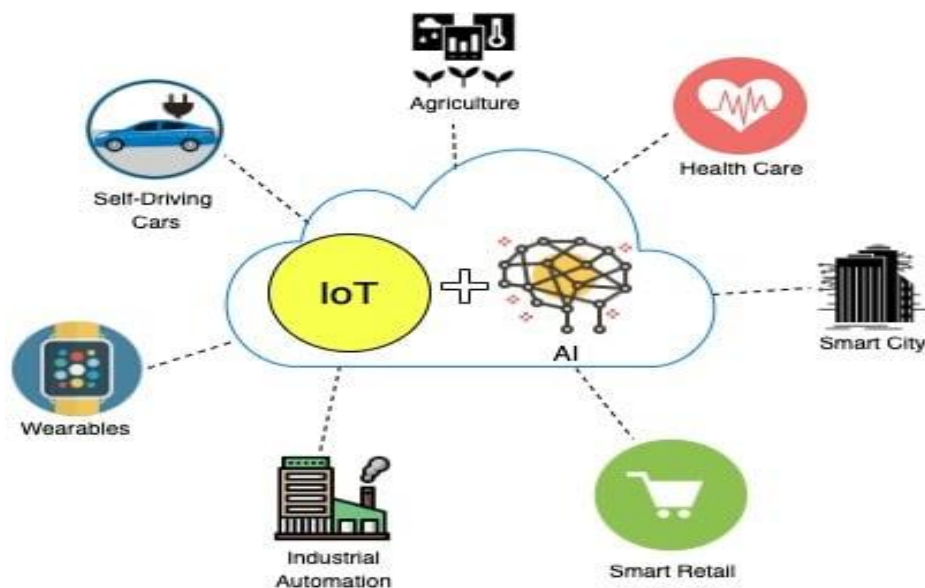
## **Problem Statement: -**

In today's digital age, wireless devices have become an integral part of our daily lives. From smartphones to laptops, we rely on these devices for communication, entertainment, and work. However, as the number of wireless devices increases, so does the risk of cyberattacks. Hackers can exploit vulnerabilities in these devices to gain access to sensitive information or even take control of them. By using machine learning algorithms, we are trying to predict cyberattacks in this case.

## **Area or domain of work: -** Machine Learning

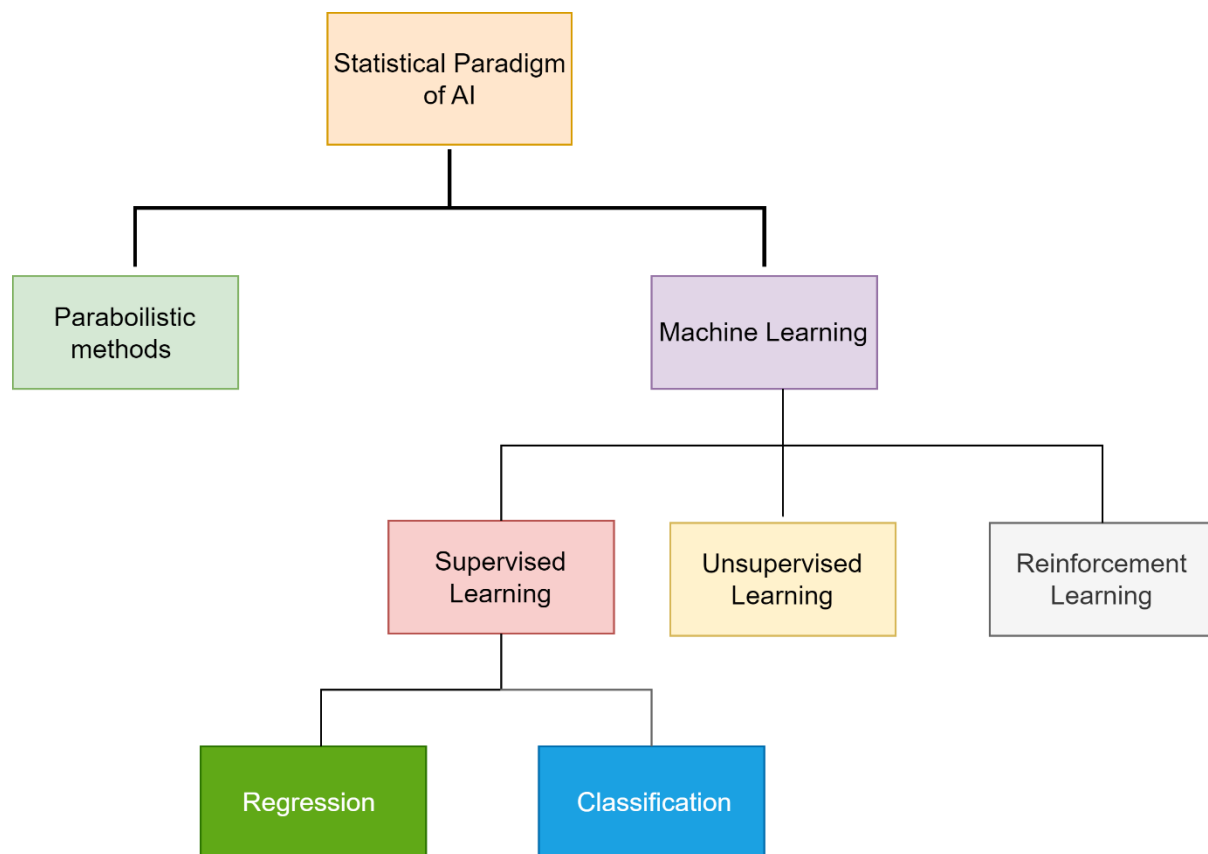
## **Introduction: -**

As we were deploying the network traffic capture by IoT devices, we knew that this was a crucial step in managing the traffic data through IoT devices. With the increasing number of IoT devices being used, it has become essential to have a system in place that can handle the vast amount of data generated by them. By capturing network traffic, we could analyse and monitor the data flow, identify any potential issues, and optimise the performance of the devices. This would help us ensure that our network remained secure and efficient.



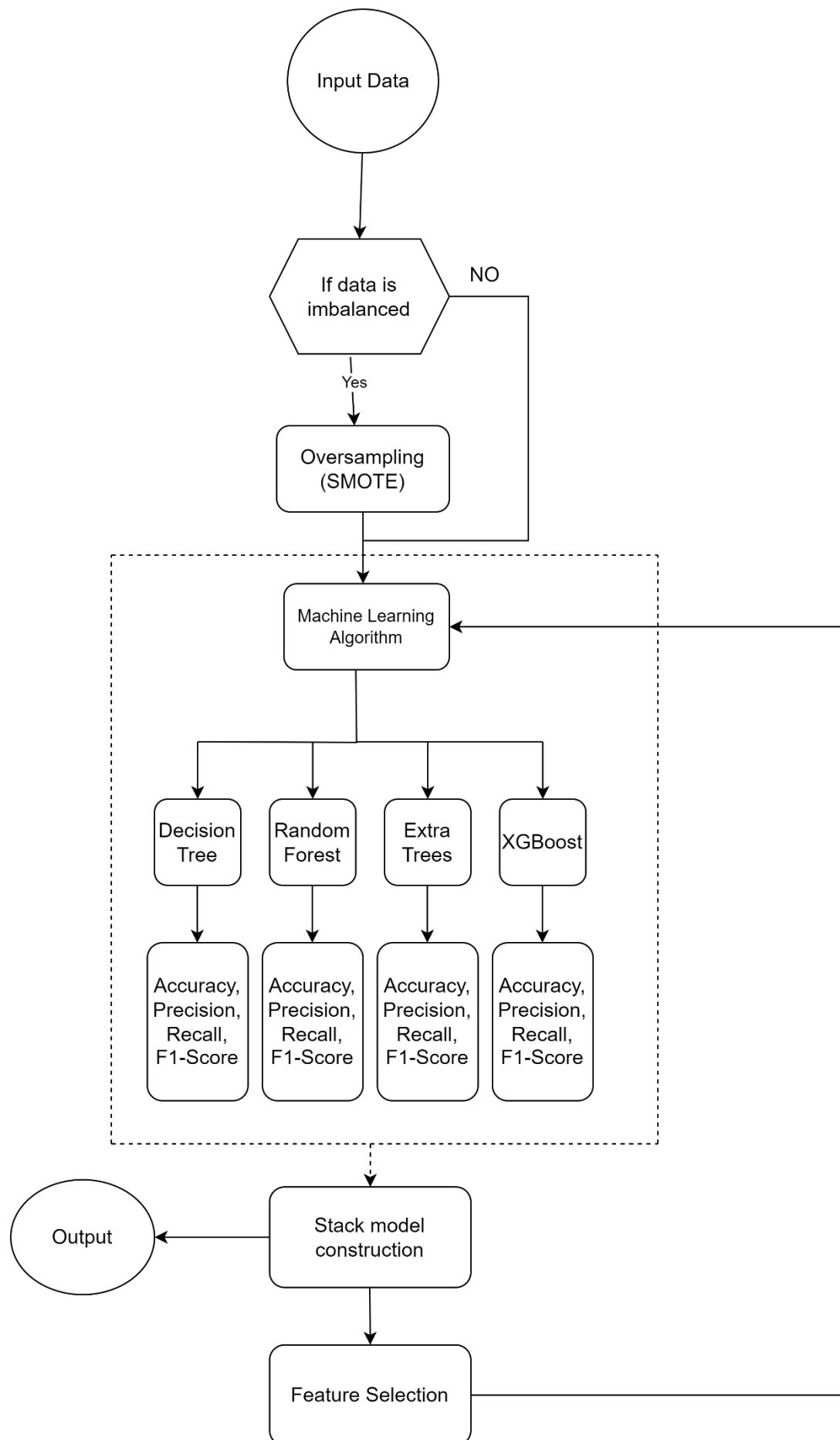
### **Approach summary: -**

In the statistical paradigm of Artificial Intelligence, it contains mainly two parts: Paraboilistic methods and Machine Learning. Now here we are, using machine learning. Machine learning can be classified into three types: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. We were using a supervised learning data set. Supervised learning is classified into two types: regression and classification. Our problem comes under classification.



Here we were using main four Machine Learning algorithm they are - decision tree (DT), random forest (RF), extra trees (ET), and extreme gradient boosting (XGBoost). After applying it we will apply feature selection to get more accuracy. As you see more clarity in flow chart which was given below.

With help of confusion matrix we will analysis the obtained output.





## **Experimental Procedure: -**

### **I. Data set used**

Here we have use **CICIDS2017** dataset contains benign and the most up-to-date common attacks, which resembles the true real-world data (PCAPs).

The CICIDS2017 dataset is publicly available at: <https://www.unb.ca/cic/datasets/ids-2017.html>

### **II. Language and Frameworks used**

*Python Language and Jupiter Notebook (Anaconda Navigator 2.4.0)*

### **III. Computational Python libraries used**

- **import** numpy as np
- **import** pandas as pd
- **import** seaborn as sns
- **import** matplotlib.pyplot as plt
- **from** sklearn.preprocessing **import** LabelEncoder
- **from** sklearn.model\_selection **import** train\_test\_split
- **from** sklearn.metrics **import** classification\_report, confusion\_matrix, accuracy\_score, precision\_recall\_fscore\_support
- **from** sklearn.metrics **import** f1\_score
- **from** sklearn.ensemble **import** RandomForestClassifier, ExtraTreesClassifier
- **from** sklearn.tree **import** DecisionTreeClassifier
- **import** xgboost as xgb
- **from** xgboost **import** plot\_importance
- **from** imblearn.over\_sampling **import** SMOTE

***Numpy***: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python.

***Pandas***: The last library is the Pandas library, which is one of the most famous Python libraries and is used for importing and managing datasets. It is an open-source data manipulation and analysis library.

***Seaborn*** library is a widely popular data visualization library that is commonly used for data science and machine learning tasks

***Matplotlib***: The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot. This library is used to plot any type of charts in Python for the code.

***Scikit-learn*** is an open-source data analysis library, and the gold standard for Machine Learning (ML) in the Python ecosystem. Key concepts and features include: Algorithmic decision-making methods, including: Classification: identifying and categorizing data based on patterns

***Label Encoder***: - Encode target labels with value between 0 and n\_classes-1. This transformer should be used to encode target values

***Training Set***: A subset of the dataset to train the machine learning model, and we already know the output.

***Test set***: A subset of the dataset to test the machine learning model, and by using the test set, model predicts the output.

A ***Classification report*** is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False.

A **confusion matrix** is a visual representation of the performance of a machine learning model. It summarizes the predicted and actual values of a classification model to identify misclassifications.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Where TP is True Positive

FP is False Positive

TN is True Negative

FN is False Negative

**Accuracy** defined as the ratio correct prediction and total number of cases.

$$\text{Accuracy} = \frac{\text{correct Prediction}}{\text{Total cases}} \times 100 \%$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \times 100\%$$

**Precision** is defined as ratio of true positive and sum of true positive and false positive.

$$\text{Precision} = \frac{TP}{(TP+FP)}$$

**Recall** is defined as ratio of true positive and sum of true positive and false negative.

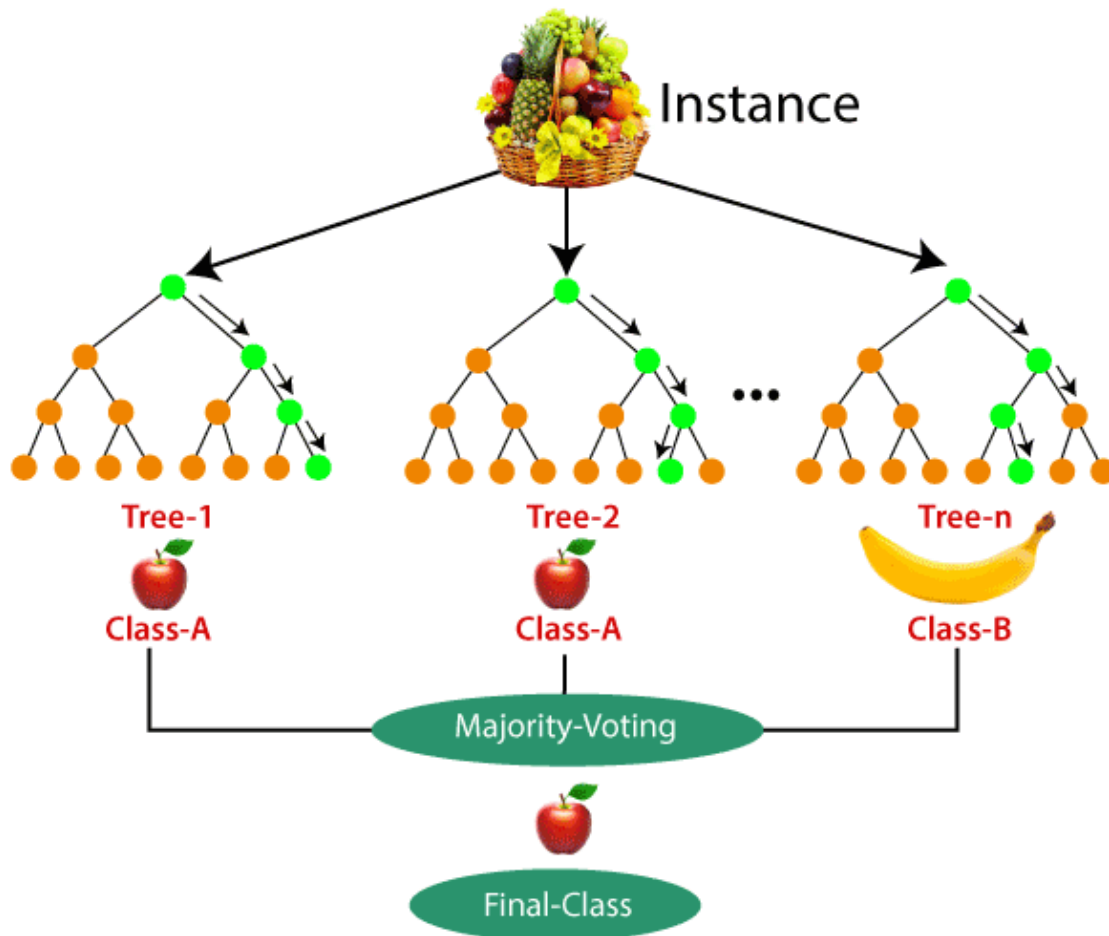
$$\text{Recall} = \frac{TP}{(TP+FN)}$$

**F1 score** - An F-score is the harmonic mean of a system's precision and recall values. It can be calculated by the following formula

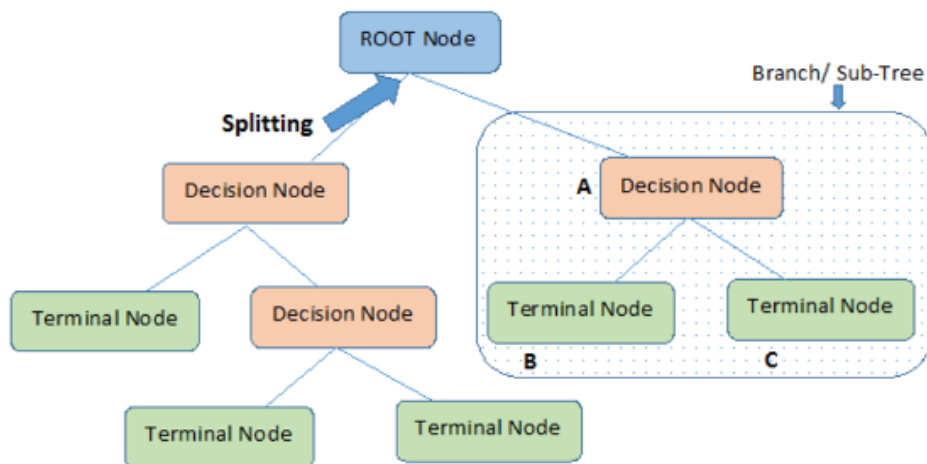
$$\text{F1 score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

## ***Random Forest Classifier***

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

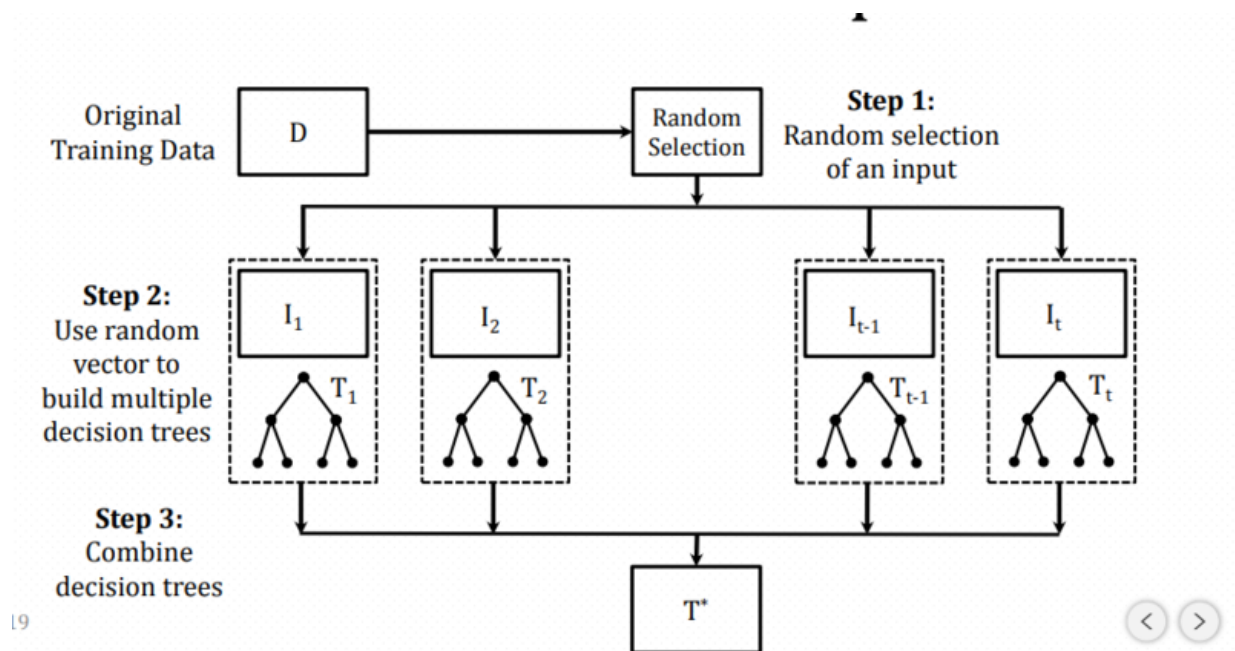


## Decision Tree Classifier



A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

**Extra Trees Classifier:** The purpose of the Extra Trees Classifier is to fit a number of randomized decision trees to the data, and in this regard is a form of ensemble learning. Particularly, random splits of all observations are carried out to ensure that the model does not overfit the data.



- Random Forests build multiple decision trees over bootstrapped subsets of the data, whereas Extra Trees algorithms build multiple decision trees over the entire dataset
- Decision tree learns from one path while extra tree learns from multiple tree

**XG-Boost** is an ensemble learning algorithm meaning that it combines the results of many models, called base learners to make a prediction.



**SMOTE** – Synthetic Minority Over Sampling Technique. It is used to train the classifier and obtain synthetically class balance or nearly class balanced training set

### **III Statistical Performance**

- We will train a 4-ML base classification for intrusion detection, and we will find the accuracy, precision, recall, and F1 score.
- The data was pre-processed in the next stage, and oversampling was used if the data were unbalanced.
- On the next stage, we applied function selection, and we will once again train all ML basis classification and the difference.

### **IV. Obtain Result**

In the given data set, there are 7 classes: -

- BENIGN
- DoS
- PortScan

- BruteForce
- WebAttack
- Bot
- Infiltration

In below table represents the total number of classes repeating in the dataset before splitting the dataset .

Classes	Total number of classes repeating in the dataset before splitting the dataset
BENIGN	2273097
Dos	380699
PortScan	158930
BruteForce	13835
WebAttack	2180
Bot	1966
Intiltration	36

Before factor selection, there are 78 attributes, and after factor selection, there are 38 attributes. We were performing the 4 ML Base algorithm to find the accuracy, precision, recall, and F1 score for before and after feature selection as shown below.

	Decision Tree	Random forest	Extra Tree	XGBoost	Stack model construction
Accuracy %	99.6029	99.2411	99.2058	99.4793	99.6029
Precision %	99.6012	99.2464	99.2019	99.4785	99.6012
Recall %	99.6029	99.2411	99.2058	99.4793	99.6029
F1 Score	0.9960	0.9923	0.9920	0.9947	0.9960

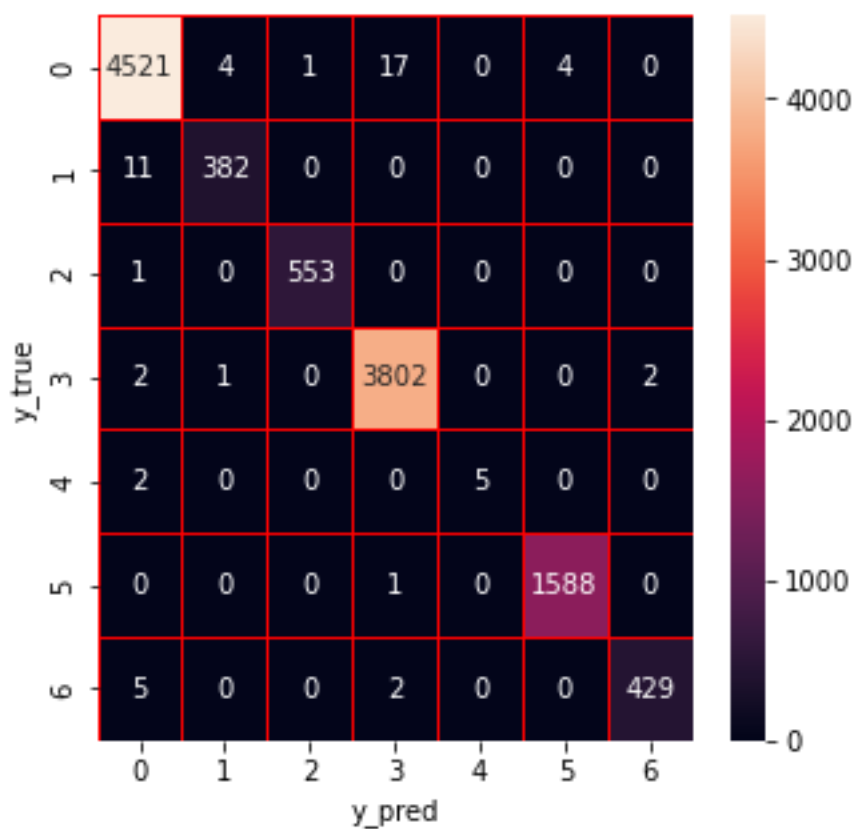
***Before Feature selection***

	Decision Tree	Random forest	Extra Tree	XGBoost	Stack model construction
Accuracy %	99.5941	99.6382	99.574	99.4529	99.5588
Precision %	99.5942	99.6363	99.5753	99.4503	99.5594
Recall %	99.5941	99.6382	99.5764	99.4529	99.5588
F1 Score	0.9951	0.3363	0.9957	0.9944	0.9955

### *After Feature Selection*

On above both table we had observe that after feature selection the random forest and extra tree as increased and rest two were decreased.

By comparing all 4 classifiers, before feature selection Decision tree has highest accuracy. But after feature selection random forest has highest accuracy.





## **Our Consolidated Observation: -**

Here, we were working on IoT network traffic data. And used to detect cyber-attacks by sending large numbers of applications at once (Dos) and changing the GPS location (spoofing). Our team was thrilled to be working on such a cutting-edge project as IoT network traffic data analysis. Our goal was to develop a system that could detect cyberattacks, such as DoS and GPS location spoofing, before they caused any significant damage. To achieve this, we utilised advanced algorithms and machine learning techniques to analyse the massive amounts of data generated by these networks. By identifying patterns and anomalies in the traffic data, we were able to quickly identify potential threats and take proactive measures to prevent them from causing harm.

## **Usefulness this problem statement in real time deployment in industries**

- Banking Sector
- Healthcare
- Self-driving system
- Defence sector

## **Future scope of this work**

Following are computational possibilities to extend this work and explore various perspectives.

- ❖ Utilising the other oversampling techniques
- ❖ While training a model test with some other base learners, such as Neural network version with various hyperparameters such as
  - Total number of layers
  - Total number of neurons at each layer
  - Mechanism used for weights and bias
  - Generation allotment
  - Optimized algorithm used
  - Activation function used
  - Learning rate used
  - Number of batches used
  - Number of epox used
- ❖ Test the system's performance behaviour by varying hyperparameters.
- ❖ In the simulation, test whether deep neural network mathlogy is performing better than the ensemble learning approach.