Network Intrusion Detection System

German University in Cairo

Supervised by Dr.Maggie Ezzat

Presented by :

Nourhan Reda Abdellal

Khalid Badr El Toukhy

Nardeen Shaher

Presentation out line

- Introduction
- Data cleaning
- Data exploration
- Data preprocessing.
- Handling class imbalance
- Preparing for ML model
- Binary classification
- Multi classification

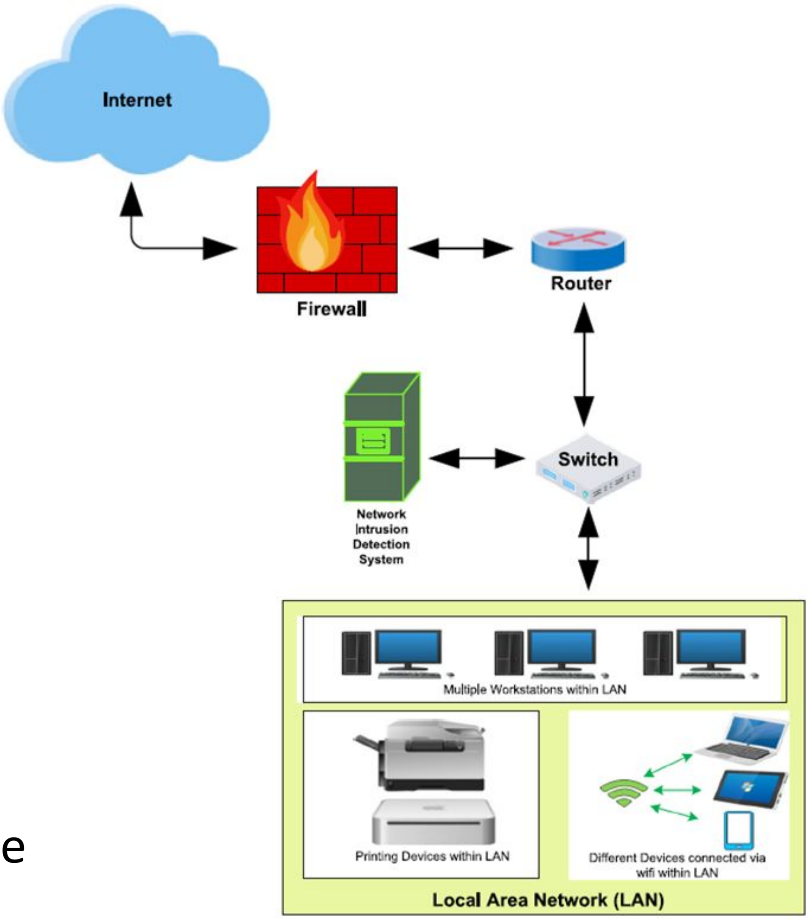
Intro to Intrusion Detection Systems

Abstract

Network intrusion detection systems (NIDS) are the most common tool used to detect malicious attacks on a network. They help prevent the ever-increasing different attacks and provide better security for the network. NIDS are classified into signature-based and anomaly-based detection. The most common type of NIDS is the anomaly-based NIDS which is based on machine learning models and is able to detect attacks with high accuracy.

- NIDS is implemented in the form of a device or software that monitors all traffic passing through a strategic point in the network for malicious activities

It is typically deployed at a single point, for example, it can be connected to the network switch (as in the figure)
If malicious behavior is detected, NIDS will generate alerts to the host or network administrators


Internet
Firewall
Router
Switch
Network Intrusion Detection System
Multiple Workstations within LAN
Printing Devices within LAN
Different Devices connected via wifi within LAN
Local Area Network (LAN)

Abstract

Network intrusion detection systems (NIDS) are the most common tool used to detect malicious attacks on a network. They help prevent the ever-increasing different attacks and provide better security for the network. NIDS are classifed into signature-based and anomaly-based detection. The most common type of NIDS is the anomaly-based NIDS which is based on machine learning models and is able to detect attacks with high accuracy.

- NIDS is implemented in the form of a device or software that monitors all traffic passing through a strategic point in the network for malicious activities

Goals of NIDS

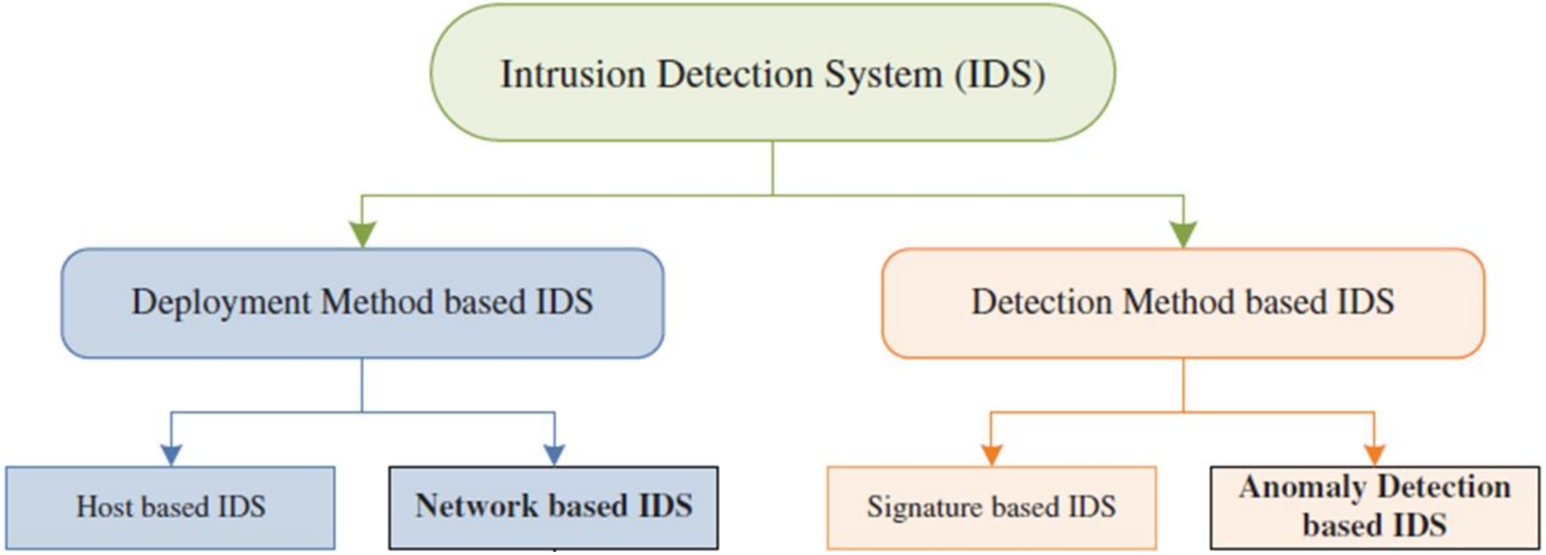
- The main goals of NIDS include:

1. Detect wide variety of intrusions
   - Previously known and unknown attacks
   - Suggests if there is a need to learn/adapt to new attacks or change in behavior
2. Detect intrusions in timely fashion
   - And minimize the time spent verifying attacks
   - Depending on the system criticality, it may be required to operate in real-time, especially when the system responds to (and not only monitors) intrusions
     – Problem: analyzing commands may impact the response time of the system
3. Present the analysis in a simple, easy-to-understand format
   - Ideally as a binary indicator (normal vs malicious activities)
   - Usually the analysis is more complex (than a binary output), and security analysts are required to examine suspected attacks
   - The user interface is critical, especially when monitoring large systems
4. Is accurate
   - Minimize false positives, false negatives

NIDS with Machine Learning

- Enormous increase in network traffic in recent years and the resulting security threats are posing many challenges for detecting malicious network intrusions
- To address these challenges, ML and DL-based NIDS have been implemented for detecting network intrusions
- Anomaly detection has been the main focus of these methods, due to the potential for detecting new types of attacks
- In the remainder of the lecture, we will first overview the datasets that are commonly used for training and evaluating ML-based NIDS, followed by a description of the ML models used for anomaly detection, and followed by adversarial attacks on ML models for NIDS

NIDS with Machine Learning



Intrusion Detection System (IDS)

Deployment Method based IDS

Detection Method based IDS

Host based IDS

Network based IDS

Signature based IDS

Anomaly Detection based IDS

- Machine learning has been used as a technique to detect anomalies in network traffic as well as novel day-zero attacks without having its signature known before.

- machine learning models that appleid

❖ Logistic regression

❖ Random forest

❖ Xgboost at binary classification and Catboost at multi classification

Dataset description

Due to the complexity in obtaining real-world labelled benign and attack network flows, researchers have generated benchmark NIDS datasets. They are made publicly available to be used in the training and testing stages of the proposed ML detection model.

There are more than 15 available NIDS datasets in this field , each containing labelled network data flows.

This dataset reflect real network benign behavior combined with synthetic attack scenarios .

The corresponding packets are captured in their native format packet capture (pcap) and certain network features are extracted

This data is used by University of Queensland, Brisbane QLD 4072, Australia and University of New South Wales, Canberra ACT 2612, Australia

And this data is a subset of big data used , the big data has many data sets inside it , and this data is one type of it t's called TonLot

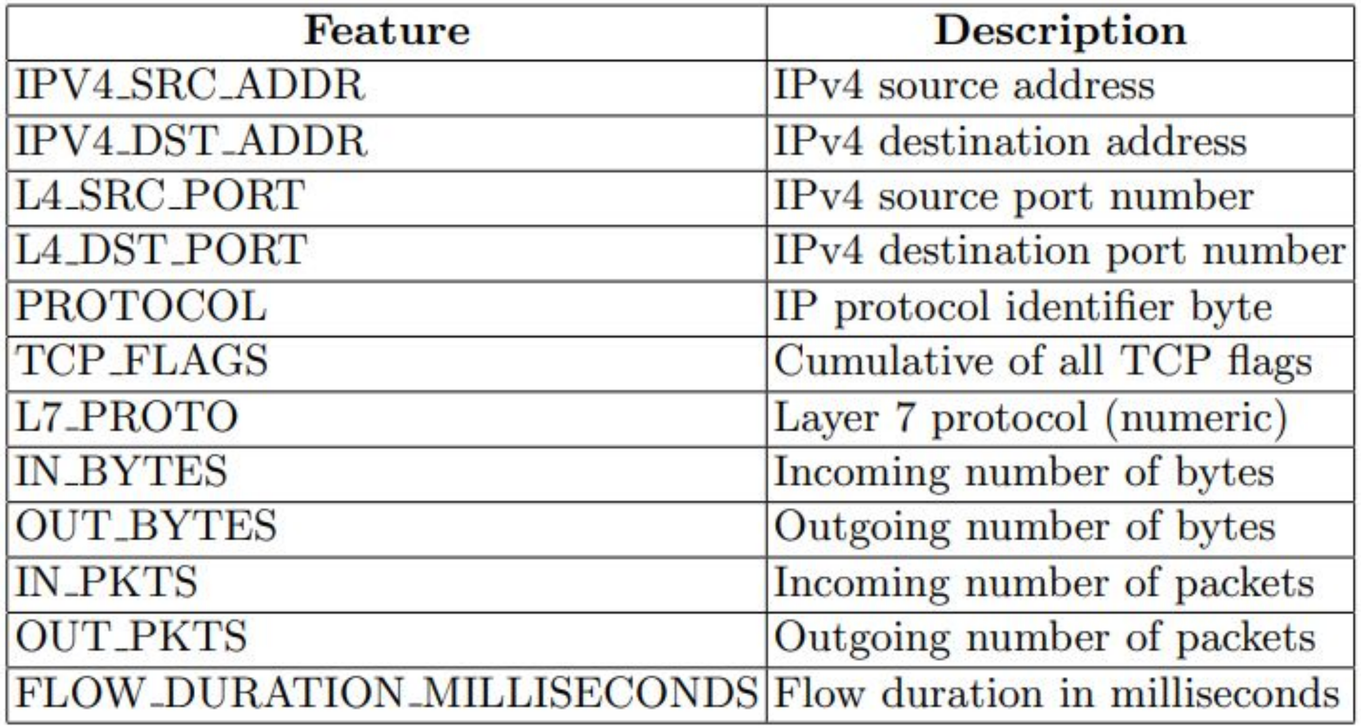
Its about 1,300,000 and 14 feature

Dataset description

ToN-IoT- A recent heterogeneous dataset released in 2020 that includes telemetry data of Internet of Things (IoT) services, network traffic of IoT networks and operating system logs. In this paper, we utilize the portion containing network traffic flows. The dataset is made up of a large number of attack scenarios conducted in a realistic representation of a medium-scale network at the Cyber Range Lab by ACCS.

The total number of data flows is 1,379,274 out of which 1,108,995 (80.4%) are attack samples and 270,279 (19.6%) are benign ones, the table below lists and defines the distribution of the NF-ToN-IoT dataset.

Feature Descriptions

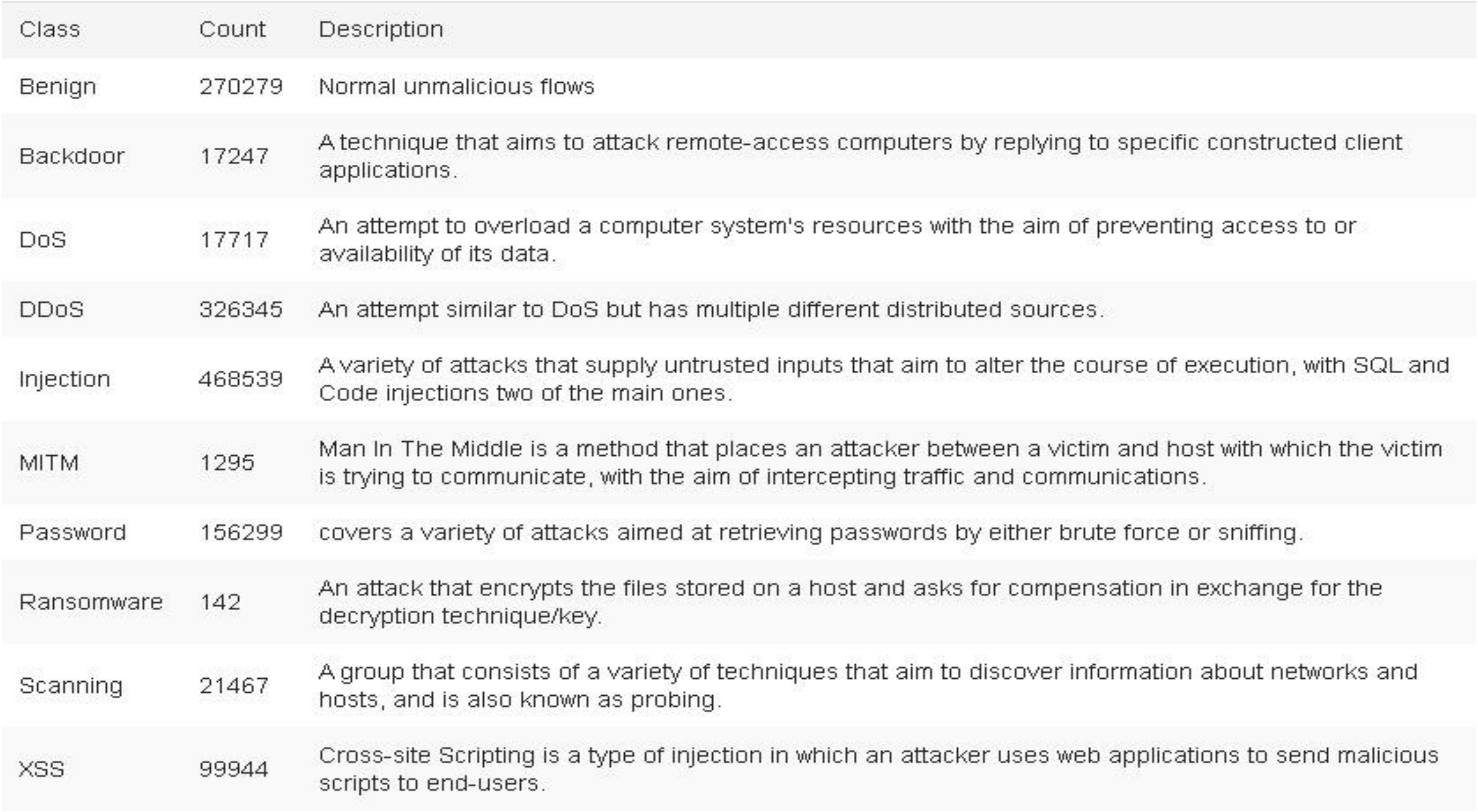


Feature	Description
IPV4_SRC_ADDR	IPv4 source address
IPV4_DST_ADDR	IPv4 destination address
L4_SRC_PORT	IPv4 source port number
L4_DST_PORT	IPv4 destination port number
PROTOCOL	IP protocol identifier byte
TCP_FLAGS	Cumulative of all TCP flags
L7_PROTO	Layer 7 protocol (numeric)
IN_BYTES	Incoming number of bytes
OUT_BYTES	Outgoing number of bytes
IN_PKTS	Incoming number of packets
OUT_PKTS	Outgoing number of packets
FLOW_DURATION_MILLISECONDS	Flow duration in milliseconds

Collecting and recording network traffic is necessary to monitor and analyze networks. There are two main trends for this process, capturing the complete network traffic, i.e. traffic packets, and capturing a summary of network packets in the form of flows. While packet capturing provides full access to traffic history for the network and security analysis.

Feature Descriptions

the table below lists and defines the distribution of the NF-ToN-IoT dataset.



Class	Count	Description
Benign	270279	Normal unmalicious flows
Backdoor	17247	A technique that aims to attack remote-access computers by replying to specific constructed client applications.
DoS	17717	An attempt to overload a computer system's resources with the aim of preventing access to or availability of its data.
DDoS	326345	An attempt similar to DoS but has multiple different distributed sources.
Injection	468539	A variety of attacks that supply untrusted inputs that aim to alter the course of execution, with SQL and Code injections two of the main ones.
MITM	1295	Man In The Middle is a method that places an attacker between a victim and host with which the victim is trying to communicate, with the aim of intercepting traffic and communications.
Password	156299	covers a variety of attacks aimed at retrieving passwords by either brute force or sniffing.
Ransomware	142	An attack that encrypts the files stored on a host and asks for compensation in exchange for the decryption technique/key.
Scanning	21467	A group that consists of a variety of techniques that aim to discover information about networks and hosts, and is also known as probing.
XSS	99944	Cross-site Scripting is a type of injection in which an attacker uses web applications to send malicious scripts to end-users.

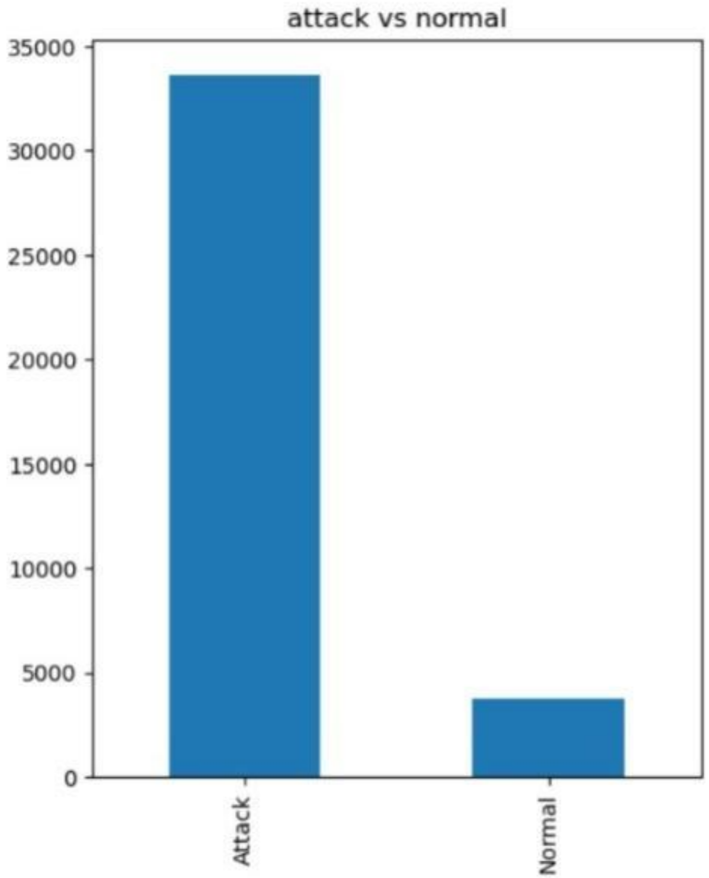
Addressing the class imbalance problem in network intrusion detection systems using data resampling

when a class is underrepresented in a dataset, this causes the dataset to be imbalanced. Detecting the minority class becomes a difficulty which lowers the performance of the intrusion detection system

The TON-LOT dataset has much more attack samples than the normal samples

Techniques used to handle class imbalance

- Random oversampling
- Smote
- Adasyn



Addressing the class imbalance problem in network intrusion detection systems using data resampling

Random over sampling

this is the technique in which we select random points from the minority class and duplicate them to increase the number of data points in the minority class. But is considered to be the most robust out of all. This method is considered to be the most basic oversampling technique. Since the selection of the points is random this artificially creates a reduction in the variance of the dataset

SMOTE over sampling

SMOTE stands for Synthetic Minority Oversampling Technique. This technique generates new observations by interjecting a point between observations of the original dataset. It makes use of the K-Nearest Neighbors strategy, an easy to use model selection framework is supplied to enable the rapid evaluation of oversampling techniques on unseen datasets

- Adasyn

the Adaptive Synthetic Sampling Approach. This approach is based on the methodology of SMOTE but instead of generating points from nearby points its uses outlier points i.e. “harder-to-learn-from” to generate new points. But the drawback of this is that due to its adaptive nature its precision gets affected. Also for minority examples that are scattered, each neighborhood may only contain 1 minority example.

Data Preprocessing

Addressing the class imbalance problem in network intrusion detection systems using data resampling

Random over sampling

this is the technique in which we select random points from the minority class and duplicate them to increase the number of data points in the minority class. But is considered to be the most robust out of all. This method is considered to be the most basic oversampling technique. Since the selection of the points is random this artificially creates a reduction in the variance of the dataset

SMOTE over sampling

SMOTE stands for Synthetic Minority Oversampling Technique. This technique generates new observations by interjecting a point between observations of the original dataset. It makes use of the K-Nearest Neighbors strategy, an easy to use model selection framework is supplied to enable the rapid evaluation of oversampling techniques on unseen datasets

- Adasyn

the Adaptive Synthetic Sampling Approach. This approach is based on the methodology of SMOTE but instead of generating points from nearby points its uses outlier points i.e. “harder-to-learn-from” to generate new points. But the drawback of this is that due to its adaptive nature its precision gets affected. Also for minority examples that are scattered, each neighborhood may only contain 1 minority example.