# Introduction

As the popularity of mobile end devices is growing exponentially, the number of malicious activities directed against them has also surged. People are using their mobile devices for sending, receiving, and storing data. Vulnerabilities against sensitive information are crucial; people can reach and manage financial, medical, and governmental applications by their mobile devices. Because of the mobile devices’ widespread appeal, they became targets of the attackers. If a user's profile and behavior pattern can be correctly identified, then any abnormal activity which deviates from standards, and expected behavior can be identified as an anomaly. Defining user behavior patterns correctly and detecting anomalies is crucial to preventing malicious attacks. For intrusion detection, Machine Learning algorithms are considered efficient in the classification and detection of attacks [1] [2] [3]. To detect malicious behavior on mobile devices, multiple types of research are performed on anomaly detection; these researches focused on the static and dynamic levels of anomaly detection separately: static and dynamic analysis. In this project, dynamic analysis is adopted for anomaly detection, which means, the data is collected with an already developed application, then several models are inspected in terms of accuracy of detecting anomalies. While detecting anomalies, it is observed that most of the selected algorithms for unsupervised learning require evenly distributed time-series data, which means continuous behavior is required. However, since the experiment was based on real-world data, it is realized that for some hours no samples were tracked. For adopting an adequate algorithm, anomaly detection of data with missing values is also covered in this paper.

# Project Requisites

To manage and use data fast, Google Collab Notebook (Jupyter Notebook) is used, it provides 12 GB of RAM. The project’s data science and machine learning parts will be implemented with Python programming language (version 3.7.12). Only LSTM models with different parameters cause problems in terms of usage of RAM, which can be improved with bigger RAM or investigating the reasons of the problems in the execution and changing of the model. To achieve an original solution, literature review is performed multiple days in each week.

# Experiment

There are several machine learning models deployed to reach a proper model for the special case of this project. Models with LSTM, Decision Tree, SARIMAX, ARIMA, K-Means Clustering, Isolation Forest and Interquartile Range Algorithm are developed and tested for dynamic analysis.

## Data

In the beginning of the project, first daily data was provided, and first analysis is performed on the daily data. Based on first instance, component analysis is performed, and the design of the algorithm is selected. Since we had time series data, LSTM was considered as the first choice for model. Due to the feature analysis, in terms of correlations (correlations between columns, hours, weekdays, weeks), memory usage data is selected for bui9lding prediction algorithms.

Second, monthly data (February 2022) is provided which has approximately 1.5 million samples. The LSTM model’s results were not as well as expected therefore, a decision tree model based on mathematical notations, SARIMAX, ARIMA, Interquartile Range, K Means, and Isolation Forest algorithms were deployed. However, since the data is unevenly distributed, research is adopted to deploy algorithms with unevenly distributed data and with filled samples. For filling the missing values, two approaches are maintained:

· Filling with 0: in that case when the network traffic is not tracked for a sampled period, the value is automatically accepted as 0.

· Filling with closest previous value: in that case, the sampled row has not got network traffic information, the sample is filled with closest previously recorded value for given time.

For machine learning models, data is sampled daily. In that way, periodicity between days is aimed to be observed. During the analysis of data, cyclostationary behavior is observed, which means, almost everyday, memory usage in similar hours are tracked in the same trend(increasing or decreasing). For observing these behaviors, Fourier Transform on data is performed. At the end of data analysis, , periodicity of data is taken as 24 hours.

## Machine Learning Models

### LSTM

LSTM, long-short time memory is one of the artificial neural networks that learns from long-term dependencies, and is suitable for predicting time-series data. For that purpose, Vanilla and Stacked LSTM models are deployed for the project.

### Decision Tree

Decision Tree algorithm is adopted with basic Artificial Intelligence rules on given features. In that pèurpose, data is inspected with different window sizes, hourly behavior, daily behavior, week-day behavior, and weekly behavior are inspected separately to create a proper decision tree from scratch. Cumulative Distribution Function of data points is used as separating nodes in decision tree.On the other hand, several process mining tools(ProM, Disco) are used to compare decision trees.

### Auto Regressive Moving Average Models (ARIMA, SARIMAX)

For these models, data points are assumed as lags taken hourly, then average of periodic data points are collected with moving the windows through all data points to make predictions

### Interquartile Range Algorithm

In descriptive statistics, the **interquartile range** tells the spread of the middle half of data distribution.

Quartiles segment any distribution that’s ordered from low to high into four equal parts. The interquartile range (IQR) contains the second and third quartiles, or the middle half of your data set.

Whereas the range gives you the spread of the whole data set, the interquartile range gives you the range of the middle half of a data set.

Calculation of Interquatrtile Range is performed with the formula IQR= Q3-Q1 where Q3 is the 3rd quartile(75% percentile), and Q1is the first quartile(25% percentile).

### K-Means Algorithm

The K-means clustering algorithm computes centroids and repeats until the optimal centroid is found. It is presumptively known how many clusters there are. It is also known as the flat clustering algorithm. The number of clusters found from data by the method is denoted by the letter 'K' in K-means.

### Isolation Forest

The Isolation Forest Algorithm ‘isolates’ observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

Since recursive partitioning can be represented by a tree structure, the number of splittings required to isolate a sample is equivalent to the path length from the root node to the terminating node.

This path length, averaged over a forest of such random trees, is a measure of normality and our decision function.

Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produces shorter path lengths for particular samples, they are highly likely to be anomalies.

# Conclusion

As a conclusion, Interquartile Range Algorithm, K-Means Algorithm, and Isolation Forest observed as best-fitting algorithms therefore, anomalies detected by these three algorithms together are considered to be labeled as “anomaly”.

For developing a stand alone application, permission from the data provider company to upload data on cloud databases can be useful.