ENGLISH SUMMARY - INTRUSION DETECTION SYSTEM

- Context

Cyber-attacks are a new ranged weapon targeting critical infrastructure. Nowadays, we are increasingly looking to secure sensitive information in network infrastructures.

- Goal

The task of the project consists in determining, from a set of statistical information of a network flow, whether this flow is a benign traffic or an intrusion, based on various supervised learning models formed from its static information.

We will use machine learning to anticipate whether the protocol is a Doh or not. While with Deep Learning we will determine if the traffic is benign or malicious.

- Methods

Here are the methods we used to predict an intrusion with DoH connection data:

  - *Data processing*

First, we processed the data provided to us by the Canadian University of New Brunswick. We had a file that contained the DoH protocol information and another that contained the intrusions.

We analyzed this data and then removed the empty columns as well as the unnecessary columns. We have also added a target column to know if the data presented is DoH or not Doh and Intrusion or Benin. We also balanced the data so that the Machine Learning and Deep Learning models were not subsequently distorted, because the number of data representing benign traffic was much higher than the number of intrusions. In addition, we had input IPs which are of object type and therefore unrecognizable by our models, so we converted them to float. Finally, we normalized the data between 0 and 1, then rounded the values and recorded all this processing and cleaning in a new csv file for use it in our models.

  - *Machine Learning*

Machine Learning is a branch of artificial intelligence that aims to empower computers to learn. A computer is not smart, it just performs tasks. Machine Learning deals with complex subjects where traditional programming finds its limits. Building a program that drives a car would be very complex if not impossible. This being due to the infinite number of possible cases to be treated. Machine Learning treats this problem differently. Instead of describing what to do, the program will learn on its own how to conduct by “observing” experiments.

For the second step of our project, we therefore implemented machine learning models to find out if the data represented a DoH protocol or not DoH. For this, we have chosen to instantiate the Decision Tree Classifier, the Random Forest Classifier, the K nearest neighbors, the Gaussian Naives Bayes and the Perceptron. Before implementing them, we varied several of their hyperparameters to optimize our models. Finally, we trained our models on the dataset and saved their training to use them later.

- *Deep Learning*

Deep Learning is a type of machine learning based on artificial neural networks in which multiple layers of processing are used to extract progressively higher-level characteristics from the data. New concepts are present in neural networks: Activation functions and back propagation. The activation function is a mathematical function that processes the information that arrives at an artificial neuron in machine learning, as do those in the brain with the electrical signals they receive. Backpropagation allows a deep learning algorithm to learn from its errors by minimizing the cost function at each epoch traversed by the algorithm.

For the third step of our project, we implemented deep learning models to determine if we had data representing intrusive or benign traffic. We chose to study the Convolutional Neural Network 1D and the Convolutional Network 2D. We followed the same path as for the machine learning models, we first varied their parameters then those with the best results we had implemented them, trained on the dataset, and saved their training. to use them later.

- *Interfaces*

We have created two interfaces, a local one with PySimple GUI and a Web one with Flask. These are interfaces in which the user either inserts new network data or loads a csv file containing the network data. With these data inserted we will also process them, so clean them, normalize them and convert to float, then we will make predictions on them from the pre-trained models. The user will directly receive the response if his data is a DoH or a non-DoH, and if the data is DoH then he will know if the recorded traffic is benign or intrusive.

- <u>Conclusion</u>

To summarize, we receive datasets containing connection information via the DoH protocol as well as intrusions. We process them to inject DoH data into our Machine Learning models on the one hand and intrusions into our Deep Learning models on the other. Then, we save the training of these models and then make a prediction on new data entered by the user, either via the graphical interface of the local application, or via the graphical interface of the web application. For both applications, we show the user the result of the predictions for each model.

- <u>Perspectives</u>
  - *Applications*

The goal of our local application would eventually become a real-time security system against attacks on the DoH protocol.

The web application could subsequently become a network analysis application.

- *Implementation and cloud*

Decentralization of database servers by leaving a response time adapted to a customer's request thanks to Edge Computing.

Enrichment of Deep Learning models thanks to users with Federated Learning.