

**CSIS 3290-001**

**Fundamentals of Machine Learning**

**Instructor: Dr. Ivan Wong**

**Project 03**

By

**Bryan Rodriguez**

**Hazel Ibasco**

**Nathaniel Pearson**

**Shedrach Ezenwali**

A Final Project Report Submitted to the Department of Computing Studies & Information Systems in Partial Fulfillment of the Requirements for the Degree

**December – 2023**

Table of Contents

[I. Introduction and discovery 2](#_Toc152541679)

[i. Introduction 2](#_Toc152541680)

[ii. Framing the problem 2](#_Toc152541681)

[iii. Initial hypotheses 3](#_Toc152541682)

[II. Data Preparation 3](#_Toc152541683)

[i. Data inventory 3](#_Toc152541684)

[ii. Data processing 3](#_Toc152541685)

[III. Model Planning and Implementation 3](#_Toc152541686)

[i. Proposed model(s) and justification 3](#_Toc152541687)

[ii. Model 3](#_Toc152541688)

[iii. Improvement 3](#_Toc152541689)

[iv. Techniques 3](#_Toc152541690)

[IV. Results Interpretation and Implications 3](#_Toc152541691)

[i. Results 3](#_Toc152541692)

[ii. Assess 3](#_Toc152541693)

[iii. Major insights 3](#_Toc152541694)

[V. Out-of-sample Predictions 4](#_Toc152541695)

[iv. Using new data 4](#_Toc152541696)

# Introduction and discovery

## Introduction

The business domain revolves around traffic data analysis with a focus on encrypted malicious traffic detection. The dataset is a compilation of balanced sizes of encrypted malicious and legitimate traffic from five public datasets. This dataset is intended for machine learning model training in the domain of cybersecurity.

The target organization, in this case, is not explicitly mentioned, but the focus is on creating a dataset that aids in developing models for the detection of encrypted malicious traffic. This is crucial in the context of increasing cyber threats, where the distinction between legitimate and malicious traffic becomes challenging due to encryption.

## Framing the problem

Framing the problem involves addressing questions related to the effective identification of encrypted malicious traffic within network data. The importance of this analysis lies in enhancing cybersecurity measures by developing accurate and efficient models that can distinguish between normal and potentially harmful network activities.

## Initial hypotheses

* The proportion of encrypted malicious traffic within the dataset is expected to be relatively small compared to legitimate traffic due to the nature of cybersecurity threats.
* The balance between malicious and legitimate traffic, along with the inclusion of both conventional and IoT devices, aims to create a dataset that represents real-world scenarios, leading to robust model training.

# Data Preparation

## Data inventory

The dataset is composed of five public traffic datasets, including the Malwares Capture Facility Project dataset and the CICIDS-2017 dataset. The data has been pre-processed to ensure balance, both in terms of the malicious and legitimate traffic and the contributions from individual datasets. Details of each selected public dataset and the final composed dataset are provided in the "Dataset Statistic Analysis Document."

The dataset was obtained from <https://data.mendeley.com/datasets/ztyk4h3v6s/2> searching with <https://datasetsearch.research.google.com/>

## Data processing

P

# Model Planning and Implementation

## Proposed model(s) and justification

Given the constraint that a deep learning approach is not feasible, a more traditional machine learning model will be employed. A Random Forest classifier is a suitable choice. Random Forest is known for its efficiency, interpretability, and ability to handle diverse datasets without overfitting. Its ensemble nature makes it robust against noise and suitable for capturing complex relationships within the data.

## Determination of Model Workflow

The workflow will involve initial data preprocessing, feature selection, and training a Random Forest classifier. Instead of clustering, which might add complexity and computational overhead, we will focus on a single model. The model will be trained on the balanced and preprocessed dataset.

## Efficiency of Workflows

To meet the time constraint, the workflow will be streamlined using pipelines for data preprocessing, feature selection, and model training. This ensures a sequential and efficient process.

## Testing Hypotheses and Modeling Objectives:

The Random Forest model will be evaluated based on its ability to accurately detect encrypted malicious traffic within the given time constraint. Insights gained from the model will contribute to understanding the features crucial for distinguishing between malicious and legitimate traffic.

Random Forest is known for its efficiency, and the use of pipelines will contribute to meeting the 2-minute runtime requirement. Feature selection techniques will be applied judiciously to reduce dimensionality and speed up the training process.

# Results Interpretation and Implications

## Results

In

## Assess

The

## Major insights

The major insights

# Out-of-sample Predictions

## Using new data

In