Luke Kong, 12311386

Project Title
nPrint OS Detection

Project Summary
The goal of this project is to identify the operating system of an endpoint based on patterns in its network traffic. Each sample contains 100 packets. I believe the importance of this task lies in security measures. For instance, older versions of Windows or Linux might have known security vulnerabilities.

Data
The dataset is drawn from the CICIDS 2017 dataset and has been relabelled for the OS Detection project. Each sample consists of 100 sequential packets from a source IP address. The dataset size is under 1 GB uncompressed, covers 13 OS classes, and uses IPv4/TCP traffic.

Machine Learning
I will start by using NetML or nPrint to turn each packet sample into a set of numeric features that can be used for training. Then I plan to test a few basic models such as a Ridge classifier, logistic regression, and a small neural network. These models are simple to train and should be enough to see clear patterns in the data. I'll compare how well each model can correctly identify the operating system and pick whichever performs best.

Evaluation
I will judge my models using balanced accuracy, which looks at how well the model performs on each operating system separately and then averages the results. The dataset has 13 different operating systems, so balanced accuracy gives each one an equal say in the final grade. I'll also look at simple accuracy and a confusion matrix to see which systems the model confuses most often. My goal is to get close to or higher than the benchmark result of about 77% balanced accuracy that the nPrint project achieved.

Learning Objective
From this project I expect to learn how to transform network traffic into features as well as how to apply standard ML workflows in network traffic analysis. I will also gain experience in handling large datasets, interpreting results for security-relevant inference, and documenting reproducible results for a benchmark task.