# DATS6301 - Project Report

William Arliss, warliss@gwu.edu

June 2022

## Introduction

This project is focused on detecting network intrusions in server telemetry data. This is a classification task where each observation is a collection of measurements from a given network flow. Each flow is either identified to be "malicious" or "benign". This problem was selected out of personal interest.

Different ensembling techniques for decision trees are tested. This includes bagging (Breiman, 2001), adaptive boosting (Freund and Schapire, 1997), and gradient boosting (Friedman, 2001). The Scikit-Learn library will be used for running each algorithm. Model performance will mainly be measured by Receiver Operating Characteristic (ROC) —specifically the area under the curve (ROC-AUC). This metric is chosen instead of simple accuracy because it is better suited for tasks where class imbalance is a problem. The data used contain roughly 62% "benign" samples and 38% "malicious samples".

The first day of work on this project was devoted to exploratory data analysis, data processing, and establishing a classification baseline. The next day of work involved hyper-parameter tuning and model selection. After that will came continued hyper-parameter refinement and model evaluation on the hold-out data.

The rest of this report is structured as follows.

## Data

The data used in this project come from the "LUFlow Network Intrusion Detection Data Set". This is a publically available dataset created by Lancaster University for the purpose of researching "detection mechanisms suitable for emerging threats". The data are available through Kaggle or a GitHub repository. The dataset is composed of 9 months of network flow data from 2020-2021 amounting to roughly 150 million observations. Due to limited computing resources and time restrictions, only a small sample (1%) of the data are used —this amounts to 1,441,018 observations used for training and 337,186 held out for testing.

# References

Breiman, L. (2001, 10). Random forests. *Machine Learning 45*, 5–32.

Freund, Y. and R. E. Schapire (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55*(1), 119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*(5), 1189 – 1232.