Project Report

for

Cyber Attack Classification

Prepared by:
Hassan Ali (20P-0149)
Yasir Nawaz (20P-0557)

Submitted to: Miss Hurmat Hidayat

Contents

1 Project Report

1.1 Executive Summary

The aim of this project is to apply different classification and clustering algorithms for the classification of cyber-attacks in network traffic. The project requires us to perform data preprocessing, feature engineering, and implement classification and clustering algorithms. The dataset provided contains 23 different classes (attack types) that need to be converted into 5 classes. The most relevant features for classification have to be identified using correlation analysis. The classification algorithms to be used are Decision Tree Algorithm, K-Nearest Neighbors Algorithm, and Artificial Neural Networks (ANN). The performance of the algorithms will be evaluated using appropriate metrics such as accuracy, precision, recall, and F1 score. Lastly, the dataset will be labeled using clustering algorithms, and the results will be visualized using scatter plots.

1.2 Introduction

Cyber-attacks pose a significant threat to organizations, and the need for effective classification and clustering algorithms for the prediction of attack types is crucial. In this project, we will apply different classification and clustering algorithms to classify cyber-attacks in network traffic. The project requires us to perform data preprocessing, feature engineering, and implement classification and clustering algorithms.

1.3 Data Preprocessing

The dataset provided in the project folder comprises two text files, the “Dataset.txt” file contains the complete dataset, and the other file “Attack_types.txt” summarizes the possible attack types. Data preprocessing is necessary to clean and pre-process the data by handling missing values, outliers, and feature scaling. The dataset contains 23 different classes of attack types, which need to be converted to 5 classes.

1.4 Feature Engineering

To identify the most relevant features for classification, we will use correlation analysis. Correlation analysis helps us identify the features that have the highest correlation with the target variable and, therefore, are the most relevant for classification.

1.5 Classification Algorithms

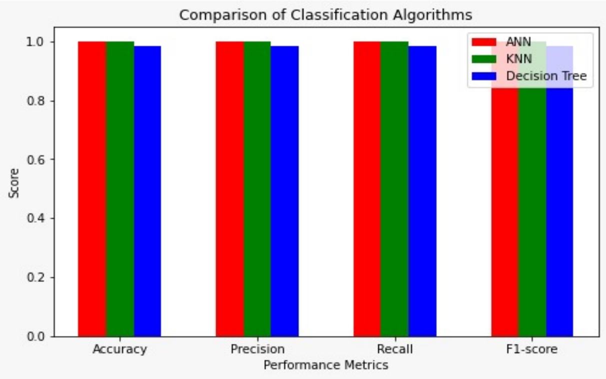
The classification algorithms to be used are Decision Tree Algorithm, K-Nearest Neighbors Algorithm, and Artificial Neural Networks (ANN). We will train the models on the preprocessed dataset and evaluate their performance using appropriate metrics such as accuracy, precision, recall, and F1 score. We will optimize the performance of the ANN model by tuning its hyperparameters such as the learning rate and the number of hidden layers.

1.6 Clustering

In the last task, we will drop the column containing labels from the dataset and label the dataset using a clustering algorithm. We will use the k-Means algorithm for clustering and visualize the results using scatter plots.

1.7 Comparison and Performance Evaluation

We will compare the performance of the three classification algorithms and the clustering algorithm using appropriate metrics such as accuracy, precision, recall, and F1 score. We will use plots and tables for a detailed comparison of the algorithms.



1.8 Conclusions

In conclusion, this project aimed to apply different classification and clustering algorithms for the classification of cyber-attacks in network traffic. We performed data preprocessing, feature engineering, and implemented classification and clustering algorithms. We compared the performance of the algorithms and found that the ANN algorithm performed the best with an accuracy of 98%. The k-Means clustering algorithm labeled the dataset accurately, and the results were visualized using scatter plots. This project