Executive Summary: Machine Learning Modeling of Anomaly Detection in Network Traffic

The main objective of this project is to develop an effective anomaly detection system using machine learning algorithms. By leveraging the CICIDS2017 dataset, the project aims to accurately classify network traffic and identify potential anomalies, thereby enhancing network security and mitigating cyber threats.

Methodology:

The project utilizes the CICIDS2017 dataset, which provides a comprehensive collection of network traffic data, including both normal and anomalous instances. PCA is employed for feature selection to reduce the dimensionality of the dataset while retaining the most relevant information. This process enables the identification of key features for classification and improves the overall efficiency of the models. Three machine learning models, namely KNN, random forest, and SVM, are implemented for anomaly detection. These models have been chosen due to their proven effectiveness in classification tasks. They are trained on the preprocessed dataset to learn the patterns and characteristics of normal network traffic. Once trained, the models are used to classify the data as benign or anomalous.

Results:

The project achieved promising results in detecting network anomalies using the proposed machine learning models. We consider two scenarios:

1. Balanced dataset for 'DDos' attack.

The metric used was accuracy score with 5-fold cross validation. The following accuracy scores were recorded: logistic regression - 76%, KNN with tuning - 99%, SVM with tuning - 99%, random forest with GridSearchCV - 99%. Random forest was deployed on the test set and achieved the same accuracy. We considered it as the best model for this problem.

2. Imbalanced dataset for 'All' attacks taken together

The metric used here was f1 score with 5-fold cross validation as in the first case. The following score was recorded: KNN model k = 30 achieved about 97% score. However, we tried fitting other models like SVM, naive Bayes and random forest but the code keeps running forever.

Conclusion:

The project successfully applied machine learning techniques to detect anomalies in network traffic using the CICIDS2017 dataset. By employing PCA for feature selection and implementing KNN, random forest, and SVM models, the project demonstrated the effectiveness of these methods in accurately classifying network traffic. The achieved results highlight the potential of machine learning in enhancing network security and mitigating cyber threats by providing real-time anomaly detection capabilities. Further optimization and fine-tuning of the models could potentially improve the performance and accuracy of the system. Overall, this project provides a solid foundation for future project and development in the field of network traffic anomaly detection using machine learning.