DDoS Traffic Clustering using K-Means Report

Ahmed Yasser Ibrahim

222110758

SE495

Introduction

This assignment aims to use the **K-Means clustering** method to analyze network traffic and identify **DDoS-related patterns** by grouping them into clusters. The process involved loading and preprocessing the dataset, engineering new features, determining the optimal number of clusters (**K**), training the model, and evaluating its performance.

Data Preprocessing

We began by loading the dataset and ensuring it was clean, well-structured, and ready for processing. Since the dataset was preprocessed and standardized for evaluating models, it required minimal cleaning. However, one challenge was **normalizing the features** due to the presence of categorical data. To address this, we applied **MinMaxScaler** only to numerical features while keeping categorical features (including the label) unchanged.

Feature Engineering

To enhance clustering accuracy, we introduced **additional features**, including:

- Traffic rate
- TCP flags sum
- Packet size ratio
- Interaction intensity

- Flow entropy

These features aimed to improve the separation of normal and DDoS-related traffic.

Clustering and Model Training

We applied **K-Means clustering** and used the **Elbow Method** to determine the optimal number of clusters. Based on the graph, we observed a sharp decline in inertia until **K=4**, suggesting it as the best choice for clustering.

Model Evaluation

After training the model, we analyzed its performance:

- The model showed **good precision** but **relatively low recall**.
- The **weighted F1-score** was **78%**, indicating room for improvement.
- Possible enhancements include **adding more relevant features** or **encoding categorical features** for better performance.

Conclusion

While the model successfully identified DDoS patterns, further improvements are needed to achieve higher recall. Future work may focus on **feature selection**, **categorical encoding**, and **trying alternative clustering algorithms** like **DBSCAN** or **Hierarchical Clustering** to refine the results.