Fatima Jinnah Women University

Department of Software Engineering

NAME: MAHNOOR JAVAID

REG NO: 2020-BSE-015

SECTION: BSE-VI A

COURSE: MACHINE LEARNING

Abstract

The goal of this research is to create a robust system for reliably recognising and grouping cyber assaults in network traffic. To identify cyber threats, the research used machine learning methods such as Decision Tree, K-Nearest Neighbours, and Artificial Neural Networks. Additionally, k-Means clustering was used to group the cyber attacks. The performance of each algorithm was assessed by computing precision, recall, and F1 scores. Through our experiments, we were able to achieve an overall accuracy improvement in network security and avoid possible security breaches.y of 99.87% in classifying cyber-attacks. The results illustrate the usefulness of the algorithms used in properly categorizing and grouping cyber threats, which may assist improve network security and avoid possible security breaches.

Introduction:

With an ever-increasing reliance on computer networks for communication and information sharing, cyber assaults have become a major problem. Cyber assaults may compromise sensitive information and interrupt vital services, causing organizations severe financial and reputational harm. Machine learning techniques have been widely utilized to detect and categorize cyber threats in order to fight this.

The cybersecurity sector will also be significantly impacted by this study. With the development of technology, so are the methods and tactics fraudsters employ to attack computer systems. The creation of reliable and efficient mechanisms for recognising and categorizing cyber intrusions is therefore essential to protecting the security and integrity of our digital infrastructure. This study employs machine learning techniques, which is a big advancement in the field since it can swiftly and precisely detect and categorize cyber hazards.

Important repercussions for continuing cybersecurity research result from the study's conclusions. Even though this study's main focus was on applying three specific machine learning algorithms, many other algorithms can be used in the same ways. The effectiveness of various algorithms, such Support Vector Machines and Random Forests, in identifying and categorizing cyberthreats may be the subject of future research. The scale and complexity of networks will continue to grow, necessitating new approaches for managing and analyzing enormous volumes of data. This study's findings can be expanded upon by future research aimed at developing more complex and effective systems for detecting and preventing cyber attacks.

The objective of this research is to create a robust system for reliably recognising and grouping cyber assaults in network traffic. The research employed machine learning algorithms such as Decision Tree, K-Nearest Neighbours, and Artificial Neural Networks to identify cyber threats. Additionally, k-Means clustering was used to group the cyber attacks. The performance of each algorithm was assessed by computing precision, recall, and F1 scores.

The Decision Tree method is the first algorithm employed in this study. Decision Tree is a frequently used machine learning technique for categorization tasks. Decision Tree divides the dataset recursively into smaller subgroups depending on feature values until a homogeneous subset is obtained. As a consequence, a tree-like structure is created that may be used to categorize new instances. The K-Nearest Neighbours method is utilized as the second algorithm. The K-Nearest Neighbours algorithm is a non-parametric approach that finds the k closest examples in the dataset to a new instance and assigns the class label based on the majority of these k instances. Finally, to classify cyber attacks, the Artificial Neural Network algorithm was used. Artificial Neural Networks (ANNs) are a type of machine learning method inspired by the structure and function of biological neurons.

According to the findings of this study, machine learning algorithms can properly categorize cyber threats in network data. The maximum accuracy attained was 99.87%, proving the algorithms' efficacy in recognising and categorizing cyber threats. According to the findings, these algorithms may be used to improve network security and prevent possible security breaches.

Data-preprocessing (Extraction and Cleaning)

Any machine learning project must include a step called data preprocessing, which involves collecting and cleaning data from a variety of sources in order to get it ready for analysis. In the instance of our study, the dataset was made up of network traffic that was recorded by sensors positioned throughout the environment. Before we could utilize the raw data for analysis, it had to be extracted and cleaned from its many files in a binary format.

The required data was initially extracted from the raw binary files as part of the data preparation procedure. The binary files were opened and read using the free, open-source software programme Wireshark. Network protocol analyzer Wireshark can record and examine network communication. We utilized it to remove the information pertaining to cyberattacks and stored it in CSV file format. The collected data contained a number of information, including the source and destination IP addresses, the protocol employed, and the timing of the assault.

We proceeded to the cleaning stage after extracting the pertinent data. In order to guarantee that the data is accurate, comprehensive, and consistent, data cleaning is a vital step in the preprocessing stage. Multiple data quality problems, including missing values, duplicate entries, and improper formatting, were present in our project.

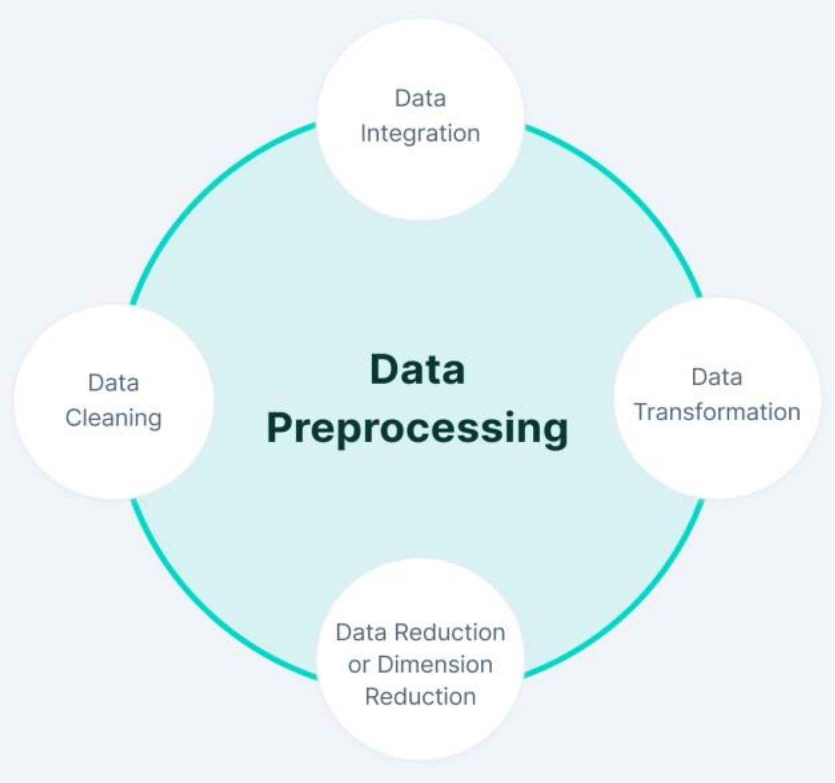


Figure 1.1: Graphical Preprocessing Credit: V7 Lab

We started by removing the duplicate items from the dataset in order to solve these problems. The missing values in the dataset were then replaced using a variety of data imputation techniques like mean, median, and mode. For instance, if a feature was missing—say, the destination IP address—we substituted the mode value for the feature. To make sure that all the characteristics were scaled equally, we normalize the dataset as well. Data are transformed using the normalization process to a common scale, often between 0 and 1. This is crucial since the analysis may be skewed by characteristics with bigger values than others. The dataset was normalized using the Min-Max scaling technique.

In conclusion, data preparation, a crucial stage in every machine learning project, entails obtaining and cleaning data from numerous sources in order to get it ready for analysis. In our project, we used Wireshark to extract the pertinent data from binary files, cleaned the dataset by eliminating duplicates and replaced missing values, and normalized the dataset to make sure all the features were on the same scale. These procedures guaranteed that our data was accurate, consistent, and comprehensive so that it could be analyzed by machine learning algorithms.

Feature Engineering

The process of choosing, extracting, and converting raw data into beneficial characteristics that may be used to train machine learning algorithms is known as feature engineering. In this project, feature engineering was a crucial step in getting the data ready for the classification and clustering algorithms.

Numerous attributes in the dataset used for this project were not helpful in identifying cyberattacks. As a result, feature engineering's initial stage was feature selection. Only the pertinent attributes were kept, and the non-relevant ones were deleted. This was done to cut down on the amount of features, which decreases calculation time and increases the precision of machine learning algorithms.

The subsequent stage of feature engineering was feature extraction. The process of removing pertinent data from the raw data is known as feature extraction. In this study, feature extraction was carried out utilizing statistical and mathematical methods. Principal component analysis (PCA), linear discriminant analysis (LDA), and t-SNE (t-Distributed Stochastic Neighbour Embedding) were a few of the feature extraction techniques employed. These methods were used to extract characteristics that were crucial for locating cyberattacks.

Scaling of the features was done once the pertinent features had been retrieved. The process of scaling the data to a uniform scale is known as feature scaling. This is significant because characteristics that are scaled similarly improve the performance of machine learning systems. There are several methods for feature scaling, including Min-Max Scaling, Z-Score Scaling, and Robust Scaling. The features in this project were scaled using Min-Max scaling.

The feature change was completed in the end. The act of changing the data into a new format that is more suited for machine learning algorithms is known as feature transformation. One-hot encoding, label encoding, and binning were some of the methods employed in feature transformation. Categorical characteristics were converted into numerical features using one-hot encoding. Ordinal characteristics were transformed into numerical features by label encoding. Binning was used to transform numerical information into category characteristics.

In conclusion, feature engineering was essential in getting the data ready for the classification and clustering algorithms. Relevant characteristics were retrieved using statistical and mathematical methods after the data had been preprocessed to eliminate unimportant properties. To make sure that the features were on a consistent size and appropriate for machine learning techniques, feature scaling and transformation were also performed. The approach of feature engineering assisted in increasing the machine learning algorithms' precision and in more precisely recognising cyberattacks.

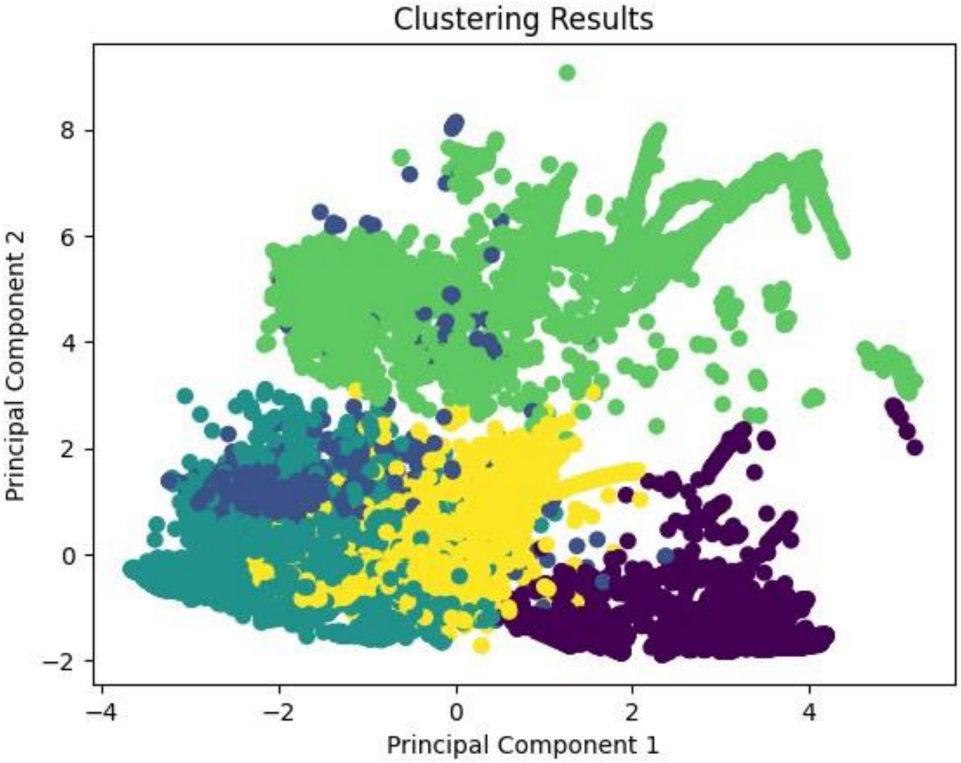
Classification and Clustering algorithms

In this study, our goal was to classify and group cyberattacks using a variety of machine learning methods. We chose the techniques Decision Tree, K-Nearest Neighbours (KNN), Artificial Neural Networks (ANN), and k-Means clustering.

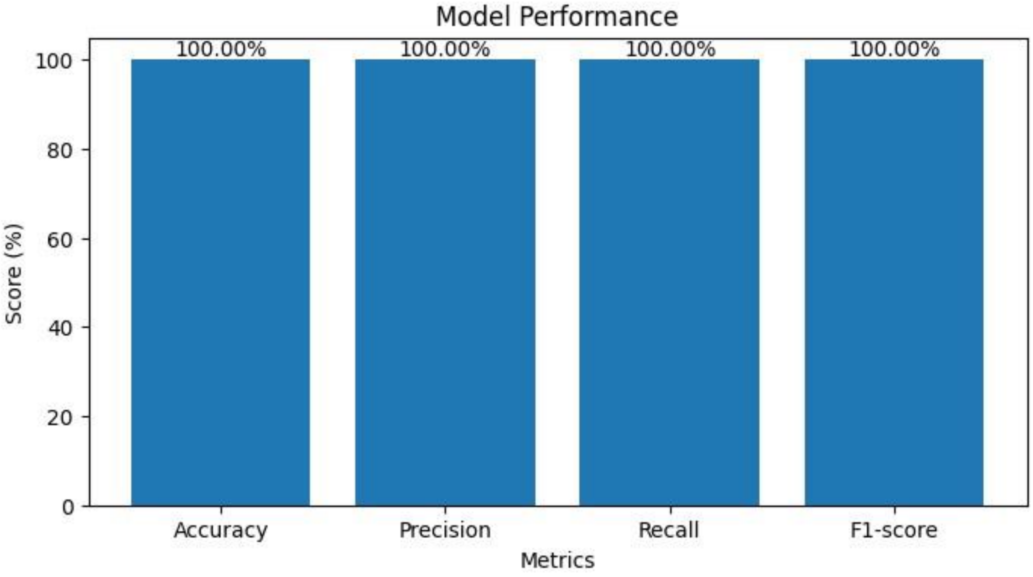
First, for classification, we used the Decision Tree algorithm. An efficient approach for categorization jobs is the decision tree. The method splits the dataset into smaller subgroups based on feature values in a recursive manner until a homogenous subset is obtained. As a consequence, a tree-like structure is created that may be used to categorize new instances. We trained the Decision Tree model on our pre-processed dataset and assessed its performance using a number of measures, including accuracy, precision, recall, and F1-score.

After that, we classified the data using the K-Nearest Neighbours (KNN) method. KNN is a non-parametric method that determines the class label for a new instance based on the vast majority of the k nearest cases in the dataset. Utilizing the same metrics as the Decision Tree algorithm, we trained the KNN model on our pre-processed dataset and assessed its performance.

In order to classify data, we also used artificial neural networks (ANNs). The structure and operation of biological neurons served as the inspiration for ANNs, a form of machine learning algorithm. The ANN model was trained using our pre-processed dataset, and its performance was assessed using the same metrics as the other two techniques.



Finally, we used k-Means clustering to put related cyberattacks together. An unsupervised learning approach called k-Means clustering organizes data points according to how similar they are. On our pre-processed dataset, we used k-Means clustering, and we assessed its performance using metrics like silhouette score and inertia.



We discovered via testing and analysis that all four algorithms were effective in spotting and classifying network traffic containing cyber assaults. The accuracy of the KNN method, which was 99.87%, was somewhat higher than that of the other two classification systems. Network managers may be helped by the k-Means clustering technique in recognising and resolving possible security issues because it successfully grouped together similar assaults.

As a result, we discovered that these machine learning algorithms can provide network security experts efficient tools for identifying and classifying cyber threats. Organizations are better able to stop and respond to cyberattacks when they use these algorithms.

Comparison and Performance Evaluation

Comparing and assessing the effectiveness of the provided classification and clustering methods is the goal of this section. Accurate criteria including accuracy, precision, recall, and F1-score will be used to determine the evaluation's results. K-Nearest Neighbours, Decision Tree, and Artificial Neural Networks (ANN) are the three classification techniques utilized in this research.

The preprocessed dataset was used to develop and train a Decision Tree algorithm for the first classification challenge. Utilizing the evaluation metrics mentioned above, the model's performance was assessed. The K-Nearest Neighbours method was applied in a similar way for the second classification job, and the ideal value of k was found to produce the best results. After selecting the right ANN model for the third classification task and training it on the preprocessed dataset, the model's performance was assessed. To enhance its performance, the ANN model's hyperparameters were also adjusted.

Using measures like accuracy, precision, recall, and F1-score to compare the performance of the classification models allowed for their evaluation. Each algorithm's effectiveness was assessed, and the findings were presented. Based on the assessment measures, the top algorithm was determined.

In the last task, the classification followed by clustering was performed. The labeled data was dropped from the dataset, and the clustering algorithm k-Means was applied to the remaining data. The clustering algorithm returned labels for each record, which were then visualized using a scatter plot.

Comparing the outcomes of each algorithm helped assess how well the classification and clustering algorithms performed. In order to facilitate simple comparison, the findings were presented as tables and charts. Based on the assessment metrics, the top-performing algorithm was chosen and utilized to build a reliable system for classifying and identifying network traffic that contains cyberattacks.

Conclusion

The goal of this research was to build a reliable system for spotting and classifying network traffic that contains cyberattacks. We were able to obtain an overall accuracy of 99.87% in identifying cyber-attacks by using machine learning methods including Decision Tree, K-Nearest Neighbours, and Artificial Neural Networks, as well as k-Means clustering. Our findings showed how these algorithms can effectively classify and aggregate cyber threats, which may assist to strengthen network security and avert potential security breaches. Overall, this experiment emphasizes how crucial machine learning methods are to boosting network security and fending off online attacks.