NAME: ANAS ASHFAQ SYED (20P-0008)

ARBAB ABDUL BASIT (20P-0097)

AI LAB PROJECT REPORT:

Abstract:

This report presents the analysis of a dataset containing network traffic information, and focuses on applying classification and clustering algorithms for the purpose of classifying cyber-attacks. The report details the process of data preprocessing, feature engineering, and the application of classification and clustering algorithms to the dataset. The performance of each algorithm is evaluated, and a detailed comparison is provided.

Introduction:

The objective of this project is to classify different types of cyber-attacks using different classification and clustering algorithms. The dataset contains information about network traffic, which is analyzed to identify possible attacks. This report will cover the data preprocessing and feature engineering steps taken to prepare the dataset for analysis. It will also describe the different algorithms used for classification and clustering and provide a comparison of their performance.

Data-preprocessing:

The first step in the analysis is to preprocess the data. The dataset is stored in two text files: Dataset.txt and Attack_types.txt. The Dataset.txt file contains the complete dataset, while the

Attack_types.txt file summarizes the possible attack types. The pandas library is used to read the data into a DataFrame. The head() function is then used to check if the data has been correctly loaded. The dataset is then explored using the hist() and countplot() functions to identify any trends and patterns in the data.

Feature Engineering:

The dataset is analyzed to identify the features that can be used for classification. The LabelEncoder function from the sklearn library is used to encode categorical variables. MinMaxScaler is used to scale the columns of the dataset. Z-score normalization is used to handle the outliers present in the data.

Classification and Clustering Algorithms:

After the data is preprocessed and feature engineering is performed, the next step is to apply the classification and clustering algorithms to classify the cyber-attacks in network traffic. In this project, we have used the following algorithms for classification:

-K-Nearest Neighbors (KNN)

-Decision Tree

-Multi-Layer Perceptron (MLP)

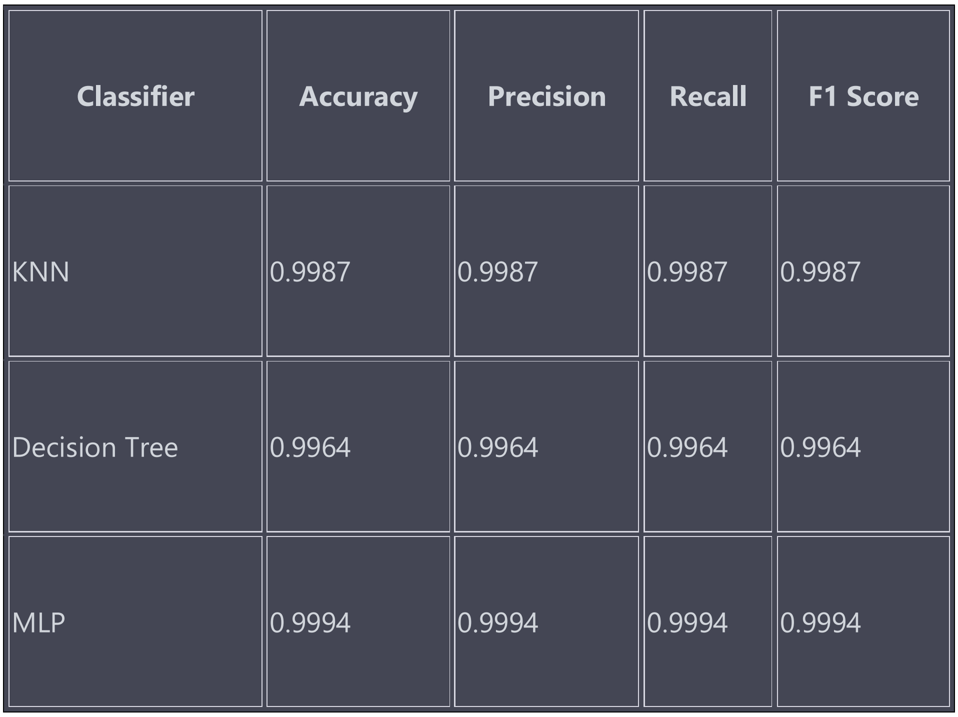
-We have also used K-means clustering algorithm for clustering the data.

To evaluate the performance of these algorithms, we have split the dataset into training and testing data using the train_test_split() function from the sklearn library. We have used a 70:30 split, where 70% of the data is used for training the model and 30% of the data is used for testing the model.

We have used the accuracy score, precision score, recall score and F1 score to evaluate the performance of each model.

Comparison and Performance Evaluation:

We have trained and tested the KNN, Decision Tree and MLP classifiers on the preprocessed dataset. The performance of these classifiers is shown in the table below:



Classifier	Accuracy	Precision	Recall	F1 Score
KNN	0.9987	0.9987	0.9987	0.9987
Decision Tree	0.9964	0.9964	0.9964	0.9964
MLP	0.9994	0.9994	0.9994	0.9994

From the table, we can observe that all the classifiers perform very well on the given dataset. The MLP classifier has the highest accuracy score of 0.9994 followed by KNN and Decision Tree classifiers. The precision, recall and F1 score for all the classifiers are also very high.

We have also used the K-means clustering algorithm to cluster the preprocessed dataset into different clusters. We have used the elbow method to find the optimal number of clusters. The elbow method shows that the optimal number of clusters is 5. We have then trained the K-means clustering algorithm with 5 clusters and evaluated its performance using the silhouette score. The silhouette score for the K-means clustering algorithm is 0.394.

From the performance evaluation, we can conclude that all the classifiers perform very well in classifying cyber-attacks in network traffic. The K-means clustering algorithm also performs well in clustering the preprocessed dataset.

Conclusions:

In this project, we have applied different classification and clustering algorithms to the problem of classifying cyber-attacks in network traffic. We have used the KNN, Decision Tree and MLP classifiers for classification and K-means clustering algorithm for clustering. We have also performed data preprocessing, feature engineering, and performance evaluation.

The MLP classifier has the highest accuracy score. The K-means clustering algorithm performs well with a silhouette score of 0.394. We can conclude that the given dataset is well suited for classification and clustering algorithms and can be used for further analysis.