University Of Engineering and Technology Lahore

Project Title:

Kmean and Naïve Base with RT_IoT2022 Dataset

Subject:

Artificial Intelligence

Submitted To:

Ms. Namra Sheikh

Submitted By:

Saneha Raees (2022-CS-706)

Mahnoor Imran (2022-CS-726)

Submission Date:

15-12-2024

1. Introduction

This project utilizes the RT_IoT2022 dataset, which contains IoT network traffic data with labeled attack types, to build a machine learning model for anomaly detection. The report covers data preprocessing, feature selection, model building using K-Means and Naïve Bayes, and performance evaluation.

2. Dataset Overview

The RT_IoT2022 dataset consists of network traffic features such as protocol type and service type, along with labeled attack types.

- Attack Types: Includes Denial of Service, Information Gathering, and more.
- Features: Numeric and categorical attributes like duration, protocol type, and service type.

Dataset Characteristics:

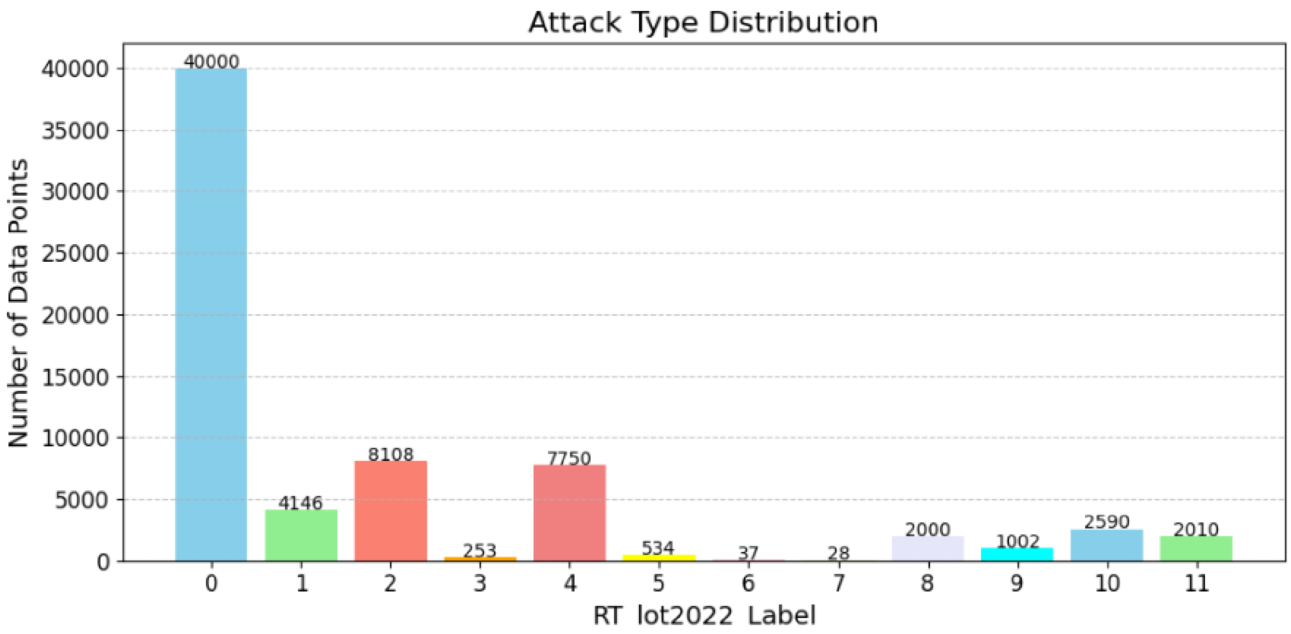
- Samples: 123117
- Features: 84
- Classes: 12

3. Data Preprocessing

- Missing Values: No missing values found.
- Balancing: Down-sampling of the majority class (normal traffic) to balance the dataset.
- Feature Engineering:
    - Removed constant columns.
    - Encoded categorical features (service and protocol type).
    - Normalized features to a range of 0-1.
    - Removed highly correlated features (correlation > 0.9).

4. Target Labeling

The Attack_type column was label-encoded to convert attack types into numeric labels.



5. Data Splitting

The dataset was split into an 80/20 training-test ratio.

6. Model Training and Evaluation

Naïve Bayes Classification
The Gaussian Naive Bayes model was trained and evaluated on accuracy, precision, recall, and F1-score.

- Cross-validation was used for model robustness.

Evaluation:

- Confusion Matrix: Analyzed for performance.
- Classification report heatmap
- ROC Curve: Plotted and AUC calculated.

7. K-Means Clustering

Elbow Method
K-Means was applied to identify patterns, with the Elbow method determining the optimal number of clusters.

- Elbow Plot: Identified the optimal cluster number.

K-mean UI

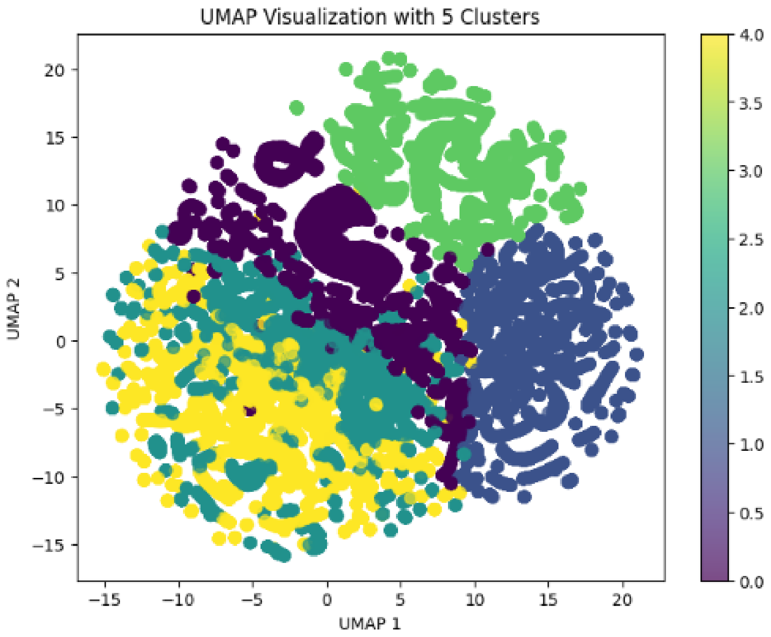
This program allows the user to select the number of clusters for K-Means and choose a plot type (Scatter Plot, Elbow Method, or 3D Plot).

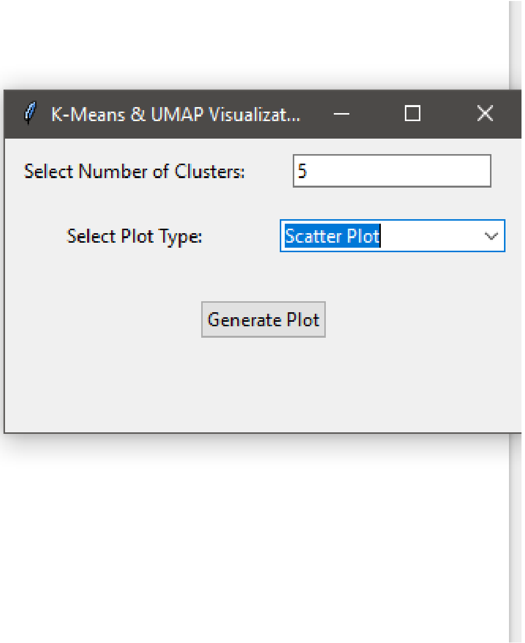
Evaluation

Silhouette Score

Visualization

It visualizes the clustering results using UMAP and PCA.




K-Means & UMAP Visualizat...

Select Number of Clusters: 5

Select Plot Type: Scatter Plot

8. Code and Report Generation

The profiling report generated using ydata_profiling provides a detailed overview of the dataset’s statistics, distributions, and visualizations.