INTERNSHIP REPORT

UNDER GUIDANCE OF

DR.ARUN RAJ KUMAR P.

AT CSED,NITC

BY MUMMANA SANJAY

BTECH
MAJOR - MECHANICAL ENGINEERING [NITK]
MINOR - COMPUTER SCIENCE [NITK]

TASK 5 ASSIGNED ON 7TH JUNE:

Task Overview

The primary objective of this task is to improve the classification accuracy and reduce false positives in a network intrusion detection system using Logistic Regression enhanced by Multi-Agent Reinforcement Learning (MARL). The project follows a systematic approach as outlined below:

1. Data Preparation

- **Combine and Clean Datasets**:
  - Load and concatenate data from multiple CSV files.
  - Remove duplicate entries and handle non-numeric columns.
- **Handle Missing Values and Scale Features**:
  - Impute missing values using mean imputation.
  - Scale features using standard scaling techniques.

2. MARL Agent Definition

- **Define MARL Framework**:
  - Create a Multi-Agent Reinforcement Learning framework with agents capable of dynamically adjusting model parameters.
  - Agents use Q-learning to update actions based on rewards received from the environment (model performance).

3. Model Training and Evaluation

- **Implement K-Fold Cross-Validation**:
  - Use 5-fold cross-validation to ensure robust model evaluation.
- **Train Logistic Regression Models**:
  - Train models using Logistic Regression both with and without MARL enhancement.
  - Agents in the MARL framework adjust model parameters during training to optimize performance.
- **Collect Performance Metrics**:
  - Focus on key metrics such as accuracy and false positives.
  - Evaluate the models on test data to compare their performance.

4. Performance Comparison and Analysis

- **Present Comparative Analysis**:
  - Compare the performance metrics of Logistic Regression models with and without MARL.

    - Highlight improvements in accuracy and reductions in false positives achieved by integrating MARL.
- **Theoretical Analysis of Top Features**:
    - Identify top features using information gain.
    - Discuss the relevance of these features in detecting SYN traffic patterns.

Performance Comparison Table

Fold	Test Accuracy (Without MARL)	False Positives (Without MARL)	Test Accuracy (With MARL)	False Positives (With MARL)
1	0.6759	4496	0.9617	803
2	0.6750	4525	0.9562	927
3	0.6779	4447	0.9616	809
4	0.6775	4368	0.9582	896
5	0.6809	4257	0.9572	912
Mean	0.6774	4419	0.9590	869

Summary

- **Test Accuracies**:
    - Mean Test Accuracy without MARL: 0.6774
    - Mean Test Accuracy with MARL: 0.9590
- **False Positives**:
    - Mean False Positives without MARL: 4419
    - Mean False Positives with MARL: 869

Relevance of Top Features Identified Using Information Gain in Delineating the SYN Traffic Pattern

Information gain is a feature selection method that measures the reduction in entropy (uncertainty) when a feature is used to split the data. It helps identify features that provide the most information about the class labels. In the context of delineating SYN traffic patterns, the top features identified using information gain are crucial for several reasons:

Rank	CIC DDoS 2019 (Feature)	Information Gain
1	Average Packet Size	1.562453
2	Packet Length Mean	1.550324
3	Fwd Packet Length Mean	1.539658
4	Avg Fwd Segment Size	1.539144
5	Max Packet Length	1.536757

1. **Average Packet Size**:
    - SYN flood attacks often generate packets of specific sizes. By monitoring average packet size, it's possible to detect anomalies that indicate the presence of SYN traffic.
2. **Packet Length Mean**:
    - Similar to average packet size, the mean packet length helps in identifying deviations from normal traffic patterns. SYN packets typically have a fixed size, and significant variations in this feature can signal an attack.
3. **Fwd Packet Length Mean**:
    - This feature captures the mean length of forward packets. In SYN attacks, the forward packets (SYN packets sent to a server) often have a consistent size. Variations from this norm can be indicative of malicious activity.
4. **Avg Fwd Segment Size**:
    - The average size of segments in the forward direction can help identify unusual traffic patterns. SYN packets, being part of a connection initiation process, have specific segment sizes that can be used as markers.
5. **Max Packet Length**:
    - The maximum length of packets in the traffic flow can highlight anomalies. SYN flood attacks may produce unusually large packets to overwhelm the target, making this feature particularly relevant.

Conclusion

By focusing on these top features identified through information gain, the detection system can more accurately identify SYN traffic patterns, distinguishing between normal and malicious activities. This enhances the overall effectiveness of the intrusion detection system, making it more resilient to SYN flood attacks and other related threats.

-----THANK YOU-----