Thesis Exposé

Title: "The Digital Guardian: Leveraging Machine Learning for Identifying and Classifying Cyber Threats in Network Environments"

1. Introduction

This thesis aims to explore the application of machine learning (ML) in analyzing network traffic to detect anomalies and classify malicious activities. It addresses a common challenge faced by a software company experiencing customer complaints about network delays and potential Network traffic data can reveal vulnerabilities, such as unsecured entry points or unusual patterns of data transfer, irregular packet sizes, that can be indicators of a security breach, like intrusion attempts, malware infections, or other risks... The primary objective is to utilize ML techniques to enhance network performance and security, thereby improving customer satisfaction and the company's reputation.

2. The Scenario and Business Challenge

A pressing challenge for a software company involves customer complaints about network delays and bottlenecks in their services, raising concerns about both performance and potential security threats. This situation demands a robust solution to identify the root causes and address them effectively for maintaining customer trust and service quality.

3. Role of the Data Scientist

The data scientist's role involves analyzing network traffic data to pinpoint inefficiencies and detect any malicious activities contributing to the network issues. The task includes processing and interpreting complex network traffic data, encompassing various features related to flow identification, traffic volume, packet sizes, flow duration, and more.

4. Description of the Dataset

The data, sourced from Kaggle, contains 2.704.839 instances of network flow, each with 50 features. These features include flow identification, traffic volume, packet size, flow duration, time features, and more, offering a comprehensive view of network traffic. This dataset, collected from Universidad Del Cauca's network, represents a unique compilation of network data, and the analysis conducted in this thesis is entirely new, with no prior relation to other researchers' work.

5. Machine Learning Solution

The proposed solution is to develop a comprehensive machine learning model that employs techniques like logistic regression and ensemble learning models. These models will be trained on historical network traffic data, focusing on detecting unusual patterns indicative of performance issues or security threats.

6. Methodology

ML Methodology and Steps:

- **Data Preprocessing**: Cleaning, normalization, and transformation of data to prepare it for analysis.
- **Feature Selection and Engineering**: Identifying the most relevant features and potentially creating new features to improve model performance.
- **Model Selection**: Evaluating different ML models, including logistic regression and various ensemble learning techniques like Random Forest and Gradient Boosting.
- **Training and Validation**: Training models on a subset of data and validating their performance using cross-validation techniques.
- **Hyperparameter Tuning**: Adjusting model parameters to find the most effective settings for optimal performance.
- **Model Evaluation**: Assessing the model's performance using metrics such as accuracy, precision, recall, F1-Score, and more.
- **Deployment and Monitoring**: Implementing the model in a real-world setting and continuously monitoring its performance for any necessary adjustments.