IS 734 - DATA ANALYTICS FOR CYBERSECURITY

PROJECT PROPOSAL

SUPERVISED BY: Dr. Faisal Quader

TOPIC: SQL INJECTION DETECTION

TEAM MEMBERS:
1. Asmita Deshpande, OF68333
2. Gayathri Gurram, FR06850
3. Jahnavi Reddy Reddy Reddy, OD30976

- **OBJECTIVE OF THE PROJECT:**

  1.Cybersecurity is becoming increasingly important as there is a vast amount of sensitive information being stored and transmitted and transferred online.

  2. To develop an efficient algorithm for determining the malicious ip address which causes the SQL injection attacks.

  3. To evaluate the algorithm and analyze the characters, pattern of the ip address.

- **WHY THIS PROJECT IS IMPORTANT:**

  1. Cybersecurity is becoming increasingly important as more and more sensitive information is being stored and transmitted online.

  2. SQL Injections are a common way for cyber attackers to gain access to systems, steal data, and launch attacks.

  3. Developing an effective algorithm for detecting the SQL injection attack can help prevent cyber attacks and protect sensitive information.

- **DATASET:**

  1. The dataset used for this project is taken from a research paper which consisted of SQL injection attacks as malicious net flow data.The attacks are mainly focused on Union Query and Blind Query based types of SQL Injection.

2. The network data which is used in this dataset is collected using DOROTHEA; which is a docker framework to capture the network traffic.

3. The entire dataset contains both malicious and benign traffic which is balanced and is decomposed into 2 where the first D1 serves the purpose of training the detection models and the second which is D2 which serves the purpose of testing for generalization.

4. The parameters included in this dataset follow the version of V5 of Netflow data protocol.

Links to the datasets:
SLQ Injection Attack for training (D1).csv
SLQ Injection Attack for Test (D2).csv

- **ALGORITHMS:**

  After initial exploration of the dataset,observations include, the dataset is completely numeric and so the dependent variable is numeric too but is categorical(i.e., binary value). So to best suit the need of the end goal to be achieved classification models(logistic regression,random forest,xgboost etc) serve the purpose right.