**Project Title:**  
**Anomaly Detection in Encrypted Network Traffic Using Machine Learning**

**Objective:**  
To accurately detect suspicious or malicious activities within encrypted network traffic using machine learning techniques without decrypting it by analysing metadata.

**Dataset Used:**  
**CICIDS2017** – A comprehensive dataset simulating realistic network traffic, including both benign and various attack scenarios.

**Data Preprocessing:**

* All eight files from the CICIDS2017 dataset were preprocessed and merged for model training.
* Irrelevant or redundant features were removed.
* Missing values were imputed and any null or infinite values were replaced with 0.
* The entire preprocessing workflow is documented in data\_preprocessing.ipynb.
* Final processed data is saved as merged.csv.

**Attack Categorization:**

* The statistics.ipynb notebook analyzes the different types of simulated attacks in the dataset.
* Attacks were categorized based on frequency into three groups: **small**, **medium**, and **large**.
* Visualizations such as bar charts depict the distribution of each attack type and compare attack vs. benign traffic.

**Attack Filtering:**

* The attack\_filter.ipynb filters and compiles data based on specific attack types.
* Samples were randomly selected to create a balanced dataset.
* Specialized subsets were created, such as data containing only Web Attacks. Similar filtering can be applied for DoS or other attack categories.

**Feature Extraction:**

* extract\_features\_attack.ipynb:
  + Utilizes **Random Forest Regressor** to compute feature importance for attacks.
  + Identifies and ranks the **top 20 overall features** and **top 5 features per attack type**.
* extract\_features\_all\_data.ipynb:
  + Determines global feature importance across the entire dataset.
  + Plots and highlights the **top 20 most relevant features**.

**Model Building & Evaluation:**

* Implemented in model\_evaluation\_all\_data.ipynb.
* Three machine learning algorithms were evaluated:
  + **Random Forest Classifier**
  + **Logistic Regression**
  + **Support Vector Machine (SVM)**
* Performance was assessed using accuracy and other relevant metrics.
* **Random Forest** consistently outperformed others and was identified as the most effective model.
* All trained models were saved for future deployment.

**Future Goals:** Integrate the trained models with **real-time network traffic** using tools like:

* + **Wireshark** for packet capture
  + **Scapy** for packet manipulation
  + **CICFlowMeter** for flow generation