Project Overview: Data Science in Cybersecurity

# Overview

This project applies data science and machine learning techniques to cybersecurity incident data to identify behavioral anomalies, detect rare attack patterns, and visualize global cyber threats from 2015–2024. We used Python, SQL, and Power BI to carry out end-to-end analysis, from preprocessing to reporting.

# PHASE 1: Preprocessing & EDA (Python – Jupyter Notebook)

• Data Cleaning: Checked for missing values, duplicates, normalized column names, converted data types.

• EDA with Pandas, Seaborn, Plotly: Analyzed attack types, yearly trends, financial loss, source patterns, resolution time, and heatmaps.

# PHASE 2: ETL into SQL Database

• Schema: Countries, Industries, Attack\_Types, Cyber\_Incidents

• Loaded into MySQL via Workbench.

• SQL Queries: Top industries by loss, YOY trends, rare attacks, avg resolution time by defense.

# PHASE 3: Modeling (Anomaly Detection & Behavioral Analytics)

• Feature Engineering: Encoded categorical features, normalized continuous features.

• Models Used:

- Isolation Forest – Outlier detection (PyOD)

- One-Class SVM – Behavioral deviation (Scikit-learn)

- KMeans – Cluster incidents by similarity (Scikit-learn)

# PHASE 4: Power BI Dashboard

• Visuals: Line Chart, Bar Chart, Pie Chart, Gauge Chart

• Showcased financial loss, trends and sources.

# PHASE 5: Interpretation & Reporting

• Key Questions Answered:

- What are the most damaging attack types?

- Which industries are most targeted?

- Which countries show increasing trends?

- What patterns suggest anomalies?

# Tools Summary

• Python: Pandas, Seaborn, Plotly

• ML: Scikit-learn, PyOD

• SQL: MySQL

• Dashboard: Power BI

# Conclusion

This project proves the value of applying machine learning in cybersecurity. By using anomaly detection and clustering, we flagged risky incidents that deviate from normal patterns. With SQL and Power BI, we transformed data into real-time insights for proactive defense.