**Executive Summary – DSA 2040A Group Project**

***Team Members & Roles:***

Hetal - ETL

Chad – Analyst

Marilyn – Data Mining

Rita – Extract, dashboard & documenter

***1. Project Overview***

This project analyzes cybersecurity intrusion data using both supervised and unsupervised machine learning techniques, as well as association rule mining. The workflow covers data extraction, transformation, exploratory data analysis, modeling, clustering, and association rule mining to uncover patterns and insights related to network attacks.

***2. ETL Process Summary***

**Initial extraction and short data inspection**

* Loaded the cybersecurity intrusion CSV into a pandas DataFrame.
* Ran quick previews (.head(), .info()) to verify column types and sample content.
* Assessed data quality by counting missing values and detecting duplicate rows for subsequent data steps.

**Data transformation**

* Checked the data types for each column and converted to appropriate data types (if necessary).
* Ensured data values in ‘protocol\_type’ column have consistent casing (uppercase only).
* Checked for missing values in each column and filled missing values in ‘encryption\_used’ with “None”.
* Checked and removed (if any) duplicate rows.
* Identified outliers by drawing boxplots for each column. Attacks typically occurred at the outliers, thus kept to aid in analysis and predicting future attacks.
* Saved the cleaned data to ‘data/transformed/transformed\_cybersecurty\_intrusion\_data.csv’.

**Data loading**

* Loaded the cleaned cybersecurity intrusion CSV into a pandas DataFrame.
* Loaded the transformed file to Parquet.
* Previewed the stored results in Parquet.

***3. Exploratory Data Analysis Summary***

This Exploratory Data Analysis (EDA) examined a cleaned cybersecurity dataset (9,537 sessions) to identify patterns distinguishing normal traffic from intrusion attempts. Key findings include:

Top Attack Indicators:

* Failed logins (36% correlation with attacks)
* Excessive login attempts (28% correlation)
* Unknown browsers (73% attack rate)
* Suspicious IP reputation scores (21% correlation)

Data Quality Issues:

* Missing encryption data (21% of records)
* Outliers in network packet size (37 extreme sessions)

Key Insights:

* Attacks tend to have more login attempts but shorter session durations.
* Browser type significantly impacts risk (Unknown browsers = high risk).
* No single feature perfectly detects attacks—multivariate analysis is essential.

**Step 1: Data Loading & Initial Inspection**

Objective: Understand dataset structure, missing values, and basic statistics.

* Loaded the dataset (transformed\_cybersecurity\_intrusion\_data.csv)
* Displayed first 5 rows to inspect structure
* Checked data types & missing values using df.info()
* Generated summary statistics (df.describe())

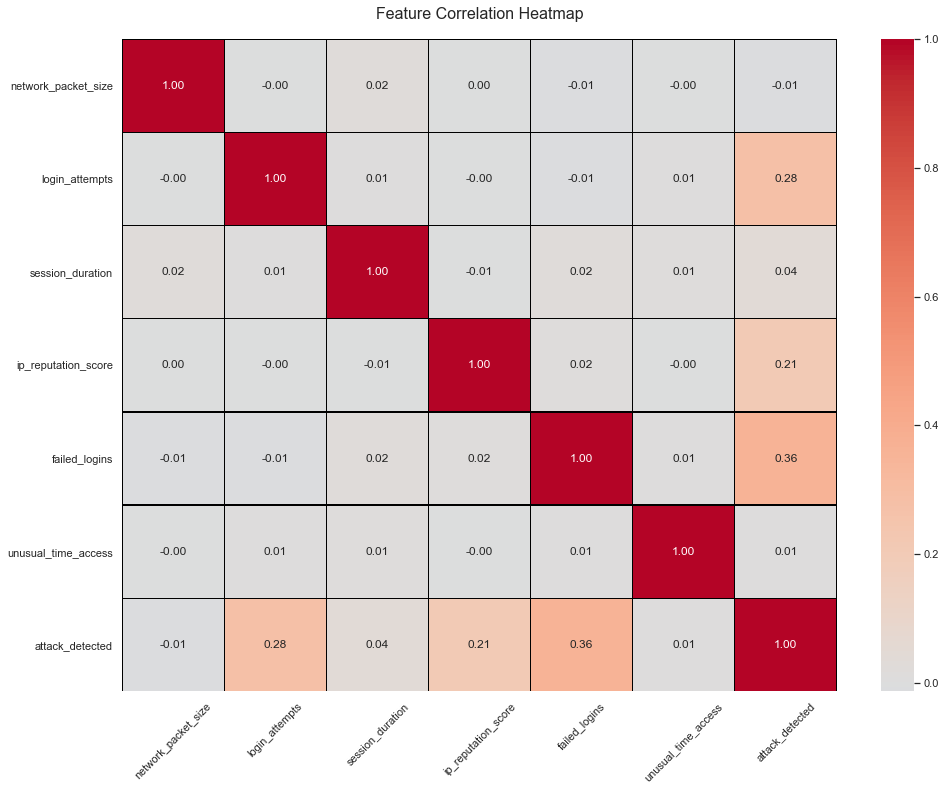
Identified:

* 9,537 records, 11 features
* Missing values in encryption\_used (21%)
* Balanced target (attack\_detected: 45% attacks)

**Step 2: Correlation Analysis**

Objective: Identify linear relationships between numerical features and attacks.

* Calculated Pearson correlation matrix for numerical features
* Plotted a heatmap to visualize relationships

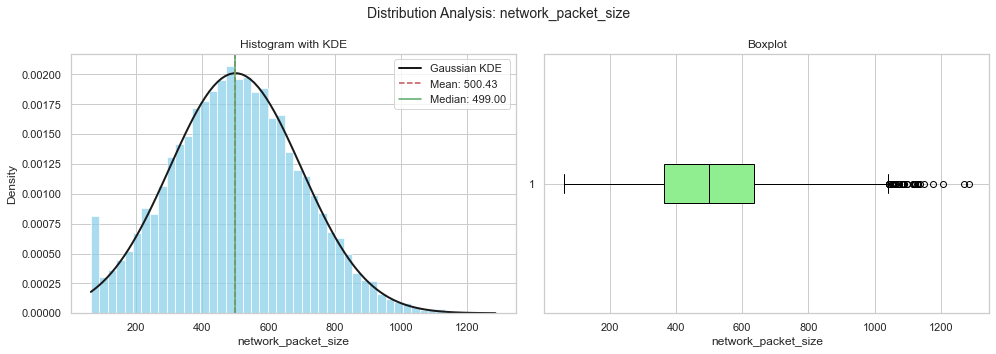


* Identified top correlated features with attacks:
* failed\_logins (0.36)
* login\_attempts (0.28)
* ip\_reputation\_score (0.21)
* Noted weak correlations (e.g., network\_packet\_size)

**Step 3: Distribution Analysis**

Objective: Examine feature distributions, skewness, and outliers.

* Selected numerical features (excluding session\_id)
* Plotted histograms + KDE curves for each feature



* Added boxplots to detect outliers
* Calculated skewness & kurtosis

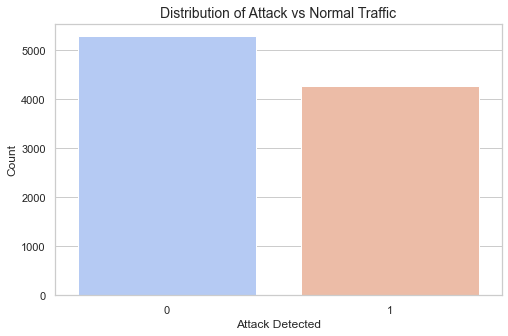
Key findings:

* login\_attempts: Right-skewed (+0.21 skewness)
* failed\_logins: High outliers (1,430 sessions)
* unusual\_time\_access: Binary (85% normal)

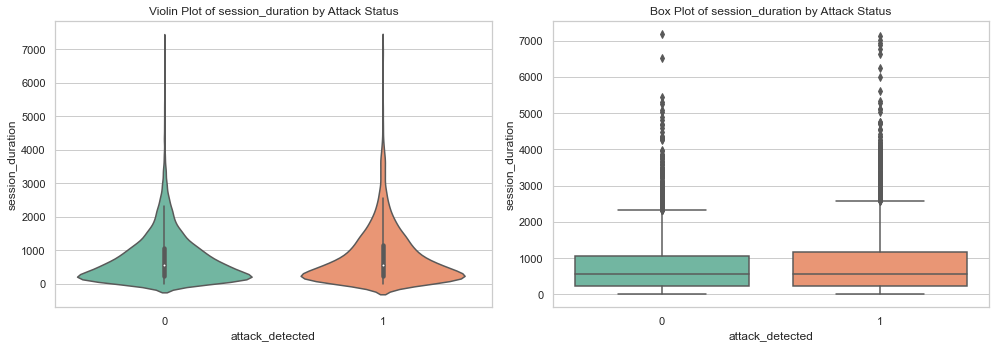
**Step 4: Group Comparisons (Attack vs. Normal)**

Objective: Compare how attack and normal sessions differ statistically.

* Split data by attack\_detected (0 vs. 1)



* Used violin plots & boxplots to compare distributions



* Calculated group means & std. deviations
* Performed t-tests for significant differences

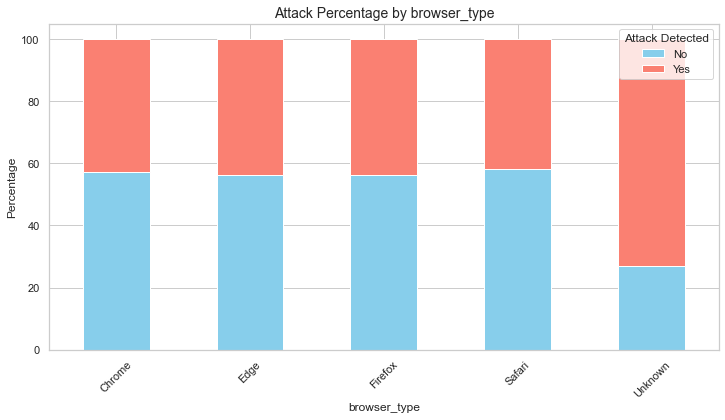
Key insights:

* Attacks have higher login\_attempts (4.8 vs. 3.6)
* Unknown browsers → 73% attack rate
* No difference in encryption type (AES vs. DES)

**Step 5: Advanced Visualizations (Multivariate Analysis)**

Objective: Detect complex patterns across multiple features.

* Analyzed browser\_type, protocol\_type, encryption\_used
* Plotted stacked bar charts for attack percentages



* Ran chi-square tests for categorical associations

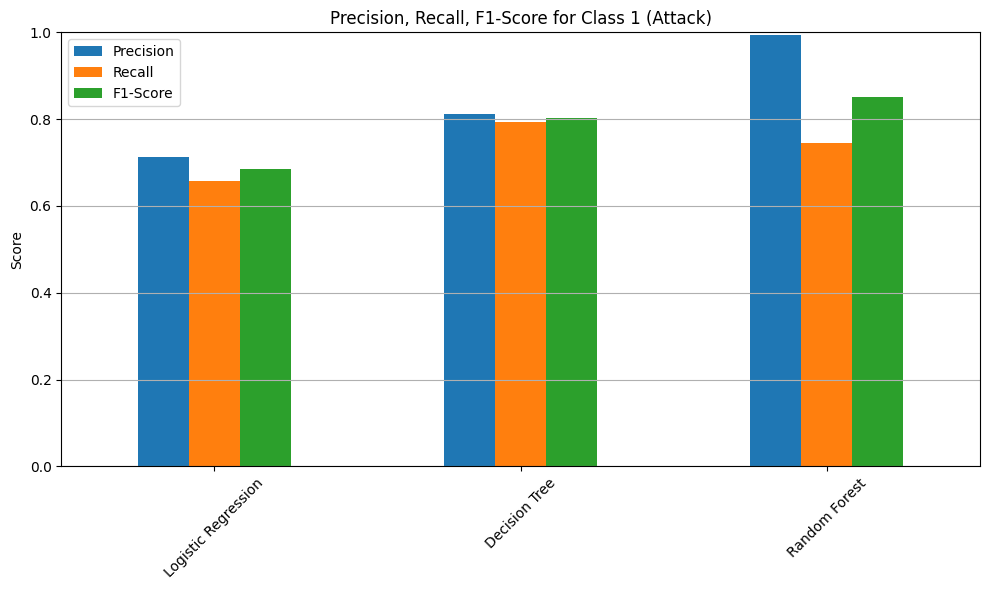
Found:

* Unknown browsers = highest risk
* TCP most common protocol (no major attack bias)

***4. Data Mining Techniques & Findings***

**a. Supervised learning**

Predicted whether a login attempt was an attack or not using logistic Regression, Decision Tree Classifier, and Random Forest Classifier.



* **Precision**: Of all predicted attacks, how many were actually attacks?

High precision = few false positives

* **Recall**: Of all actual attacks, how many did you correctly identify?

High recall = few false negatives

* **F1 Score**: Harmonic mean of precision and recall. Balances the two.

**Best Model: Random Forest**

**Why:**

* **Perfect precision (1.00)** for attacks → Every predicted attack was indeed an attack.
* **Recall (0.75)** is higher than Logistic Regression and close to Decision Tree (0.80), but the **F1-score (0.85)** is significantly higher, which means better balance.
* **Highest accuracy (0.89)** overall.

**b. Unsupervised Learning**

Performed KMean Clustering; Performed elbow technique to identify the optimum number of clusters as 4. Grouping the clusters by login attempts we see that cluster three had the highest mean number of login attempts. It needs to be treated  as a high-priority cluster for flagging or further rule-based detection

**c. Association Rule: Market Basket Analysis**

Using a minimum support of 0.3 and threshold of 0.7 we had these top two rules

|  | **antecedents** | **consequents** | **support** | **confidence** | **lift** |
| --- | --- | --- | --- | --- | --- |
| **0** | (attack\_detected) | (any\_failed\_logins) | 0.390899 | 0.874296 | 1.04764 |
| **1** | (any\_failed\_logins) | (attack\_detected) | 0.390899 | 0.468401 | 1.04764 |

*Interpretation of rule 0*

- When an attack is detected, there is an 87% chance that the session involved at least one failed login.

- Insight:

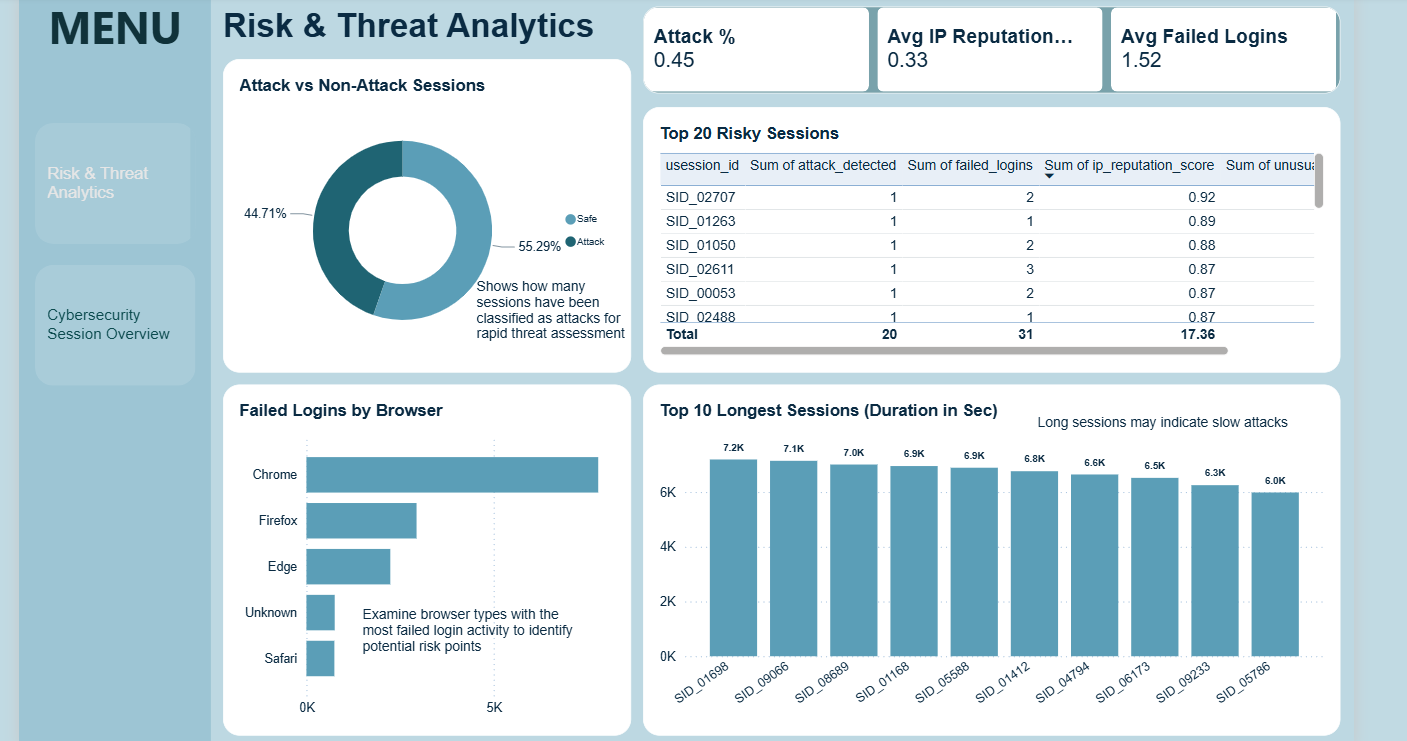
This is a very high confidence level, meaning that failed logins are highly predictive of an attack once it occurs. It suggests that failed login attempts are a strong symptom or precursor during attack sessions.

***5. Insights & Dashboard Overview***

The Power BI dashboard analyzes cybersecurity session data to support intrusion detection and risk assessment. It is organized into two pages:

**Page 1: Risk & Threat Analytics**

Focuses on identifying suspicious sessions and highlighting threat patterns.

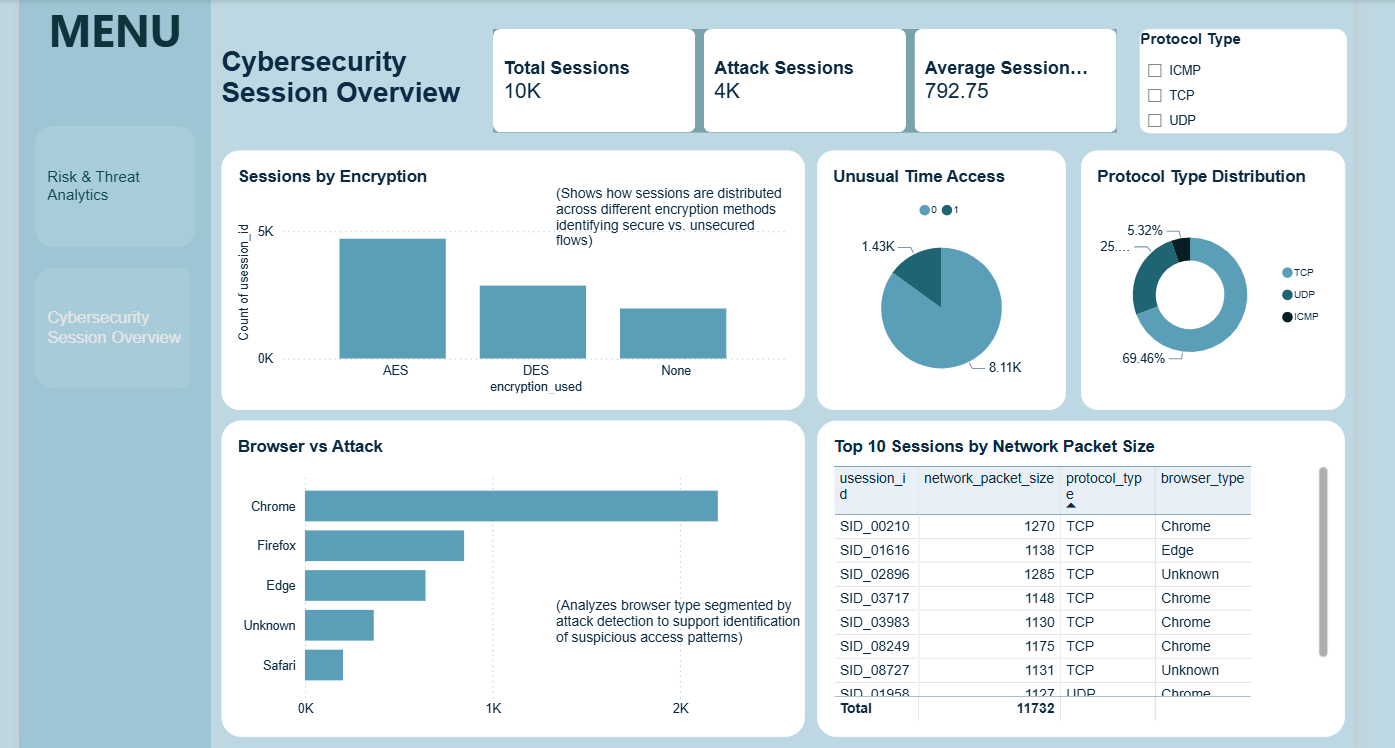


**Key Visuals:**

* **Attack vs. Non-Attack Sessions (Donut chart) -** Shows the overall distribution of attack-labeled vs safe sessions.
* **Top 20 Risky Sessions (Table) -** Lists sessions with the most critical risk factors (e.g., failed logins, low IP reputation).
* **Failed Logins by Browser (Stacked bar chart) -** Identifies browsers frequently used in failed login attempts.
* **Top 10 Longest Sessions (Clustered column chart) -** Highlights potentially stealthy or persistent session activity.
* **KPI Cards:** Summarize average attack rate, failed login attempts and IP reputation scores.

**Page 2: Cybersecurity Session Overview**

Provides broader trends across protocols, encryption and session behaviors.



**Key Visuals:**

* **Sessions by Encryption Type (Stacked column chart) -** Displays counts of sessions using AES, DES or no encryption.
* **Unusual Time Access (Pie chart) -** Shows the proportion of sessions occurring at non-standard hours.
* **Protocol Type Distribution (Donut chart) -** Compares session counts by protocol (TCP, UDP, ICMP).
* **Browser vs. Attack (Stacked bar chart) -** Relates browser usage to attack frequency.
* **Top 10 Sessions by Network Packet Size (Table) -** Displays the largest data sessions with browser and protocol details.
* **KPI Cards -** Show total sessions, number of attack sessions and average session duration.
* **Filters & Interactivity:** Includes a slicer to filter by protocol type for deeper drill-down.

***6. Tools & Libraries Used***

* pandas
* numpy
* matplotlib
* seaborn
* scikit-learn
* mlxtend

***7. Conclusion***

This project demonstrates the power of combining data preprocessing, exploratory analysis and machine learning to detect and understand cybersecurity threats.

By leveraging both supervised and unsupervised techniques along with association rule mining, the team uncovered valuable insights into attack patterns such as the significance of failed logins, unusual browsers and session behaviors. The findings not only support real-time intrusion detection but also offer a foundation for building more robust, data-driven cybersecurity strategies.

The interactive Power BI dashboard further empowers stakeholders to monitor and respond to threats effectively. Together, these tools and insights provide a comprehensive framework for enhancing cybersecurity resilience.