**Exploratory Data Analysis (EDA) Report**

**1. Dataset Overview**

* **Dataset**: GUIDE Dataset (Microsoft Cybersecurity Incidents)
* **Format**: Excel (.xlsx)
* **Train Data Size**: 1,048,575 rows × 45 columns
* **Primary Objective**: Predict the incident triage grade (IncidentGrade) — **True Positive (TP)**, **Benign Positive (BP)**, **False Positive (FP)**.

**2. Target Variable (IncidentGrade) Distribution**

| **Class** | **Percentage** |
| --- | --- |
| Benign Positive | 43.3% |
| True Positive | 35.2% |
| False Positive | 21.5% |

**Observation**:

* Mild class imbalance.
* Benign Positive is the majority class.
* False Positive is underrepresented (~21%).

**Action**:

* During model training, handle imbalance with class weights or sampling.

**3. Missing Values Analysis**

| **Column** | **Approx % Missing** | **Notes** |
| --- | --- | --- |
| ActionGrouped | ~99% | Very sparse; consider dropping |
| ActionGranular | ~99% | Very sparse; consider dropping |
| ThreatFamily | ~99% | Very sparse; consider dropping |
| ResourceType | ~99% | Sparse but meaningful values; may need selective treatment |
| Roles, AntispamDirection, LastVerdict, SuspicionLevel | 80–95% | Significant missingness |

**Observation**:

* Some features have overwhelming missingness.

**Action**:

* Drop columns with >95% missing unless critical.
* For partial missingness (e.g., MitreTechniques), impute with "Unknown" or "None".

**4. Duplicate Records**

* **Found**: 396 duplicate rows (~0.04% of dataset)

**Action**:

* Drop duplicates to prevent bias and overfitting.

**5. Timestamp Analysis**

**Converted Timestamp to datetime format** and extracted:

* **Hour**
* **Day**
* **Month**

**Distribution by Hour**

**Observation**:

* Incidents are more frequent during working hours (10:00–18:00).

**Distribution by Day**

**Observation**:

* Highest incident activity from Day 1 to Day 15 each month.
* Very low activity beyond Day 20.

**Distribution by Month**

**Observation**:

* Almost all incidents are concentrated in a single month (June).
* Indicates dataset was captured for a limited time window.

**6. Numerical Features Distribution**

**Observation**:

* Most numerical columns (OrgId, DeviceId, etc.) are not continuous but **ID-like**.
* Heavy skewness in numerical fields.

**Action**:

* Avoid standard scaling for IDs.
* Treat them as categorical if needed.

**7. Correlation Analysis**

**Observation**:

* Very weak correlations between numerical fields.
* Most fields are independent (expected due to ID-based features).

**Action**:

* No multicollinearity concerns.
* No need for dimensionality reduction like PCA based on correlation.

**8. Categorical Feature Analysis**

| **Feature** | **Key Observations** | **Action Plan** |
| --- | --- | --- |
| Category | Highly imbalanced, "InitialAccess" dominates. | Rare categories may be grouped under "Other". |
| EvidenceRole | Balanced between "Related" and "Impacted". | Simple one-hot encoding possible. |
| LastVerdict | Dominated by "Suspicious", but some invalid/random values seen. | Clean invalid values and encode. |
| ResourceType | Dominated by "Virtual Machine". | Group minor types or keep top N categories. |
| Roles | Dominated by "Contextual". | Group rare roles into "Other". |
| SuspicionLevel | Only "Suspicious" and "Incriminated". | Simple binary encoding possible. |

**Summary of Key EDA Insights**

| **Aspect** | **Observations** | **Action** |
| --- | --- | --- |
| Class Imbalance | Moderate (BenignPositive > TruePositive > FalsePositive) | Use class balancing techniques |
| Missing Values | Very high in some columns | Drop >95% missing columns, impute others |
| Duplicate Rows | Minor (396 duplicates) | Drop |
| Timestamp Patterns | Day 1–15, 10AM–6PM active | Create time-based features |
| Categorical Fields | Some dominated by 1–2 values | Group rare values if needed |
| Numerical Fields | Skewed, mostly IDs | Careful treatment without scaling |
| Correlation | Low correlation across features | No need for feature dropping |

**EDA Conclusion**

* The dataset is **large and structured**, but **partially sparse**.
* Several features require **cleaning, engineering, and encoding**.
* Special attention needed for **imbalanced classes** and **rare category handling**.
* Timestamp feature extraction will likely **improve model performance**.