**Data Preprocessing Summary**

**1. Dataset Overview**

* **Original Dataset**: GUIDE (Microsoft Cybersecurity Incidents)
* **File Type**: .xlsx
* **Original Rows**: 1,048,575
* **Original Columns**: 45
* **Target Variable**: IncidentGrade (with classes: True Positive, Benign Positive, False Positive)

**2. Preprocessing Objectives**

* Clean the dataset by removing irrelevant, sparse, or ID-specific columns.
* Handle missing values systematically.
* Engineer new features from existing data to enhance predictive power.
* Encode categorical variables efficiently to make the dataset model-ready.
* Ensure no duplicates or inconsistencies remain.

**3. Steps Executed**

**Column Dropping**

* **Dropped 23 columns** including identifiers (e.g., Id, IncidentId), hash fields (e.g., Sha256, IpAddress), sparse features (ActionGrouped, ActionGranular, ThreatFamily), and other unique or non-predictive attributes.
* **Reason**: These columns either had >95% missing data, were unique per row (hence no predictive power), or acted only as metadata without modeling relevance.

**Feature Engineering**

* **Parsed Timestamp column** into three new features:
  + Hour (0–23)
  + Day (1–31)
  + Month (1–12)
* **Dropped Timestamp** after extraction to avoid redundancy.

**Handling Missing Values**

* Filled missing values in MitreTechniques with 'Unknown'.
* Cleaned invalid values in LastVerdict, retaining only known verdicts (Suspicious, Malicious, ThreatsFound), else labeled as 'Unknown'.
* All remaining missing values across object/categorical fields were imputed with 'Unknown'.

**Rare Category Grouping**

* For Category, ResourceType, and Roles:
  + Categories appearing **less than 100 times** were grouped into an 'Other' class.
* **Reason**: To avoid overfitting and stabilize model learning for rarely seen categories.

**Encoding Categorical Features**

* **Label Encoding** applied to all categorical columns.
* **Reason**:
  + Dataset is very large (close to half a million rows).
  + One-Hot Encoding would create very sparse and memory-heavy matrices.
  + Tree-based models (XGBoost, LightGBM, Random Forest) can natively handle label-encoded integers.

**🗑 Duplicate Handling**

* Identified and removed **all duplicate rows**.
* Final dataset now contains **no duplicate entries**.

**4. Final Dataset Overview**

| **Aspect** | **Value** |
| --- | --- |
| Final Rows | 478,567 |
| Final Columns | 22 |
| Target Variable | IncidentGrade |
| Feature Types | All features are numeric (integer or float) |
| Missing Values | None |
| Duplicate Rows | None |

**5. Final Features Retained**

| **Feature Group** | **Features** |
| --- | --- |
| Categorical Encoded | OrgId, DetectorId, AlertTitle, Category, MitreTechniques, EntityType, EvidenceRole, ApplicationName, ResourceType, Roles, OSFamily, OSVersion, AntispamDirection, SuspicionLevel, LastVerdict, CountryCode, State, City |
| Engineered | Hour, Day, Month |
| Target | IncidentGrade |

**Final Status**

Dataset is fully cleaned.  
All features are encoded and model-consumable.  
Dataset is now fully ready for Machine Learning model building.