**Data Preprocessing Plan (for all columns)**

| **Column Name** | **What We Will Do** | **Why** |
| --- | --- | --- |
| **Id** | Drop | Unique identifier, no predictive value |
| **OrgId** | Keep (Label Encode) | Organizational grouping might influence incident patterns |
| **IncidentId** | Drop | Unique event ID, not meaningful |
| **AlertId** | Drop | Unique ID, not useful |
| **Timestamp** | Parse datetime + Extract Hour, Day, Month | Strong pattern by working hours/days/months seen in EDA |
| **DetectorId** | Keep (Label Encode) | Device type/source may be predictive |
| **AlertTitle** | Keep (Label Encode) | May carry useful context |
| **Category** | Keep (Group rare values + Label Encode) | Highly imbalanced, but dominant classes meaningful |
| **MitreTechniques** | Impute missing as 'Unknown', Label Encode | Important for attack categorization, sparse |
| **IncidentGrade** | Target Variable | No action needed |
| **ActionGrouped** | Drop | >99% missing |
| **ActionGranular** | Drop | >99% missing |
| **EntityType** | Keep (Label Encode) | Type of entity affected (user, device) is meaningful |
| **EvidenceRole** | Keep (Label Encode) | Role of evidence (Related, Impacted) meaningful |
| **DeviceId** | Drop | Unique device identifier, no predictive power |
| **Sha256** | Drop | File hash, unique, no modeling value |
| **IpAddress** | Drop | Highly unique, no modeling value |
| **Url** | Drop | Highly unique, too sparse for modeling |
| **AccountSid** | Drop | Unique ID |
| **AccountUpn** | Drop | Unique user ID |
| **AccountObjectId** | Drop | Unique user ID |
| **AccountName** | Drop | Unique user names, high cardinality |
| **DeviceName** | Drop | Device identifier |
| **NetworkMessageId** | Drop | Unique |
| **EmailClusterId** | Drop | Sparse and unique |
| **RegistryKey** | Drop | Sparse and mostly unique |
| **RegistryValueName** | Drop | Sparse |
| **RegistryValueData** | Drop | Sparse |
| **ApplicationId** | Drop | Mostly unique application ID |
| **ApplicationName** | Keep (Label Encode) | Application may influence incident type |
| **OAuthApplicationId** | Drop | Unique |
| **ThreatFamily** | Drop | >99% missing |
| **FileName** | Drop | Mostly unique |
| **FolderPath** | Drop | Mostly unique |
| **ResourceIdName** | Drop | Mostly unique |
| **ResourceType** | Keep (Group rare values + Label Encode) | Some predictive value in type (e.g., Virtual Machine) |
| **Roles** | Keep (Group rare values + Label Encode) | Entity roles may influence threats |
| **OSFamily** | Keep (Label Encode) | Important (e.g., Windows vs Linux) |
| **OSVersion** | Keep (Label Encode or bin versions) | May influence attack pattern |
| **AntispamDirection** | Keep (Label Encode) | Direction (Inbound/Outbound) important |
| **SuspicionLevel** | Keep (Label Encode) | Suspicious vs Incriminated |
| **LastVerdict** | Clean + Label Encode | Remove weird hash values; encode verdict type |
| **CountryCode** | Keep (Label Encode) | Location info may help in modeling |
| **State** | Keep (Label Encode) | Sub-location info |
| **City** | Keep (Label Encode) | Sub-location info |

**Summary:**

| **Step** | **How Many Columns** | **Action** |
| --- | --- | --- |
| Drop (ID-like / high missing / useless) | ~23 columns | Drop them |
| Keep and Label Encode | ~17 columns | Important context or attributes |
| Timestamp | 1 column | Parse + Feature Engineering (Hour, Day, Month) |
| Target Variable | 1 column | IncidentGrade (No preprocessing needed) |

**Data Preprocessing Tasks Grouped:**

**Handling Missing Data:**

* Drop columns with >95% missingness.
* Impute MitreTechniques with 'Unknown'.
* Clean LastVerdict (remove invalid hash codes).

**Feature Engineering:**

* Extract Hour, Day, Month from Timestamp.
* Bin rare categories in Category, ResourceType, Roles into "Other" if needed.

**Encoding:**

* Label Encoding for all categorical fields (since one-hot would explode columns — dataset too large).

**Column Breakdown**

| **Category** | **Count** | **Explanation** |
| --- | --- | --- |
| **Dropped** | 23 columns | Unnecessary IDs, hashes, >95% missing columns |
| **Retained (processed)** | 21 columns | Useful predictive features kept |
| **Newly Added** | 3 columns | Extracted from Timestamp → Hour, Day, Month |

**Final Numbers:**

| **Action** | **Number of Columns** |
| --- | --- |
| **Dropped** | 23 |
| **Retained** | 21 |
| **Newly Added (Feature Engineering)** | 3 |

**Simple math:**

* **Original Columns** = 45
* **Dropped Columns** = 23
* **Retained Columns** = 21
* **New Columns Added** = 3

**Final columns after preprocessing = 21 (retained) + 3 (new) = 24 columns**

**What the final dataset contains:**

* 21 original useful features (cleaned and encoded)
* 3 new features: **Hour, Day, Month**
* **Target column** (IncidentGrade) is among the retained (so model-ready!)

**Short Executive Summary:**

| **Item** | **Details** |
| --- | --- |
| Starting Columns | 45 |
| Columns Dropped | 23 |
| Columns Retained | 21 |
| Features Created | 3 (Hour, Day, Month) |
| Final Columns | **24 columns total** |