Dicionário de Dados - Microsoft Security Incident Prediction

Camada Bronze - Raw Data

Dataset: Microsoft Security Incident Prediction (GUIDE Dataset)
Fonte: Kaggle - Microsoft Security Incident Prediction
Período dos Dados: Junho 2024
Total de Registros: ~9.5 milhões de registros brutos (antes de limpeza)
Total de Colunas: 45 colunas (44 features + 1 target)

1. Informações Gerais sobre o Dataset

1.1 Descrição Geral

O dataset **GUIDE (Microsoft Security Incident Prediction)** contém dados reais de incidentes de segurança coletados pela Microsoft para identificar padrões de ameaças cibernéticas. Os dados estão estruturados em três níveis hierárquicos:

- **Evidence (Evidência):** Unidade atômica mais básica (IP, Email, User, Machine, File, etc.)
- **Alerts (Alertas):** Consolidação de múltiplas evidências indicando potenciais ameaças
- **Incidents (Incidentes):** Agrupamento coerente de um ou mais alertas representando eventos de segurança

1.2 Objetivo do Dataset

Predizer a **classificação (trilagem) de incidentes de segurança** em três categorias:

- **TruePositive (TP):** Incidentes confirmados como ameaças reais
- **FalsePositive (FP):** Falsos alarmes
- **BenignPositive (BP):** Incidentes benignos ou de baixo risco

1.3 Características do Dataset

- **Total de Registros:** 9,516,837 (na camada Bronze)
- **Total de Colunas:** 45
- **Período Observado:** Junho 2024
- **Países Únicos:** 236
- **Organizações Únicas:** 5,300+
- **Detectores Únicos:** 7,800+

- **Tipos de Entidades:** 33 tipos diferentes
- **Categorias de Alertas:** 20 categorias diferentes

2. Estrutura de Colunas

2.1 Identificadores (7 colunas)

Coluna	amp; **Tipo**	amp; **Descrição**	amp; **Valores Ausentes**
Id	amp; String	amp; Identificador único global do registro	amp; 0%
OrgId	amp; Integer	amp; ID da organização cliente	amp; 0%
IncidentId	amp; Integer	amp; ID único do incidente	amp; 0%
AlertId	amp; Integer	amp; ID único do alerta	amp; 0%
DetectorId	amp; Integer	amp; ID do detector que gerou alerta (7.8k+)	amp; 0%
DeviceId	amp; Integer	amp; ID do dispositivo/máquina	amp; Baixo
ApplicationId	amp; String	amp; ID da aplicação UUID/GUID	amp; Alto

2.2 Informações Temporais

Coluna	amp; **Tipo**	amp; **Descrição**	amp; **Exemplo**
Timestamp	amp; DateTime	amp; Data/hora do alerta (ISO 8601 UTC)	amp; 2024-06-04 06:05:15+00:00

2.3 Classificações e Categorias (6 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
Category	amp; Categoria MITRE ATT&CK do alerta	amp; 0%
MitreTechniques	amp; Técnicas MITRE (ex: T1078, T1566)	amp; 57.46%
IncidentGrade	amp; TARGET: TruePositive, FalsePositive, BenignPositive	amp; 0.54%
AlertTitle	amp; Título descritivo do alerta	amp; 0%
ActionGrouped	amp; Ação remediação agrupada (alto nível)	amp; 99.41%
ActionGranular	amp; Ação remediação granular (detalhada)	amp; 99.41%

2.4 Entidades e Evidências (8 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
EntityType	amp; Tipo de entidade (33 tipos)	amp; 0%
EvidenceRole	amp; Papel na investigação (Impacted, Related)	amp; Baixo
Sha256	amp; Hash SHA-256 de arquivo	amp; Alto
IpAddress	amp; Endereço IP (IPv4/IPv6)	amp; Alto
Url	amp; URL/URI envolvida	amp; Alto
FileName	amp; Nome do arquivo	amp; Alto
FolderPath	amp; Caminho da pasta (ex: C:\ Windows)	amp; Alto
DeviceName	amp; Nome legível do dispositivo	amp; Baixo

2.5 Informações de Conta (4 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
AccountSid	amp; SID de conta on-premises	amp; Alto
AccountUpn	amp; UPN/Email (usuario@dominio.com)	amp; Alto
AccountObjectId	amp; ID objeto Azure AD/Entra ID	amp; Alto
AccountName	amp; Nome da conta on-premises	amp; Alto

2.6 Informações de Aplicação (3 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
ApplicationId	amp; ID único da aplicação	amp; Alto
ApplicationName	amp; Nome legível da aplicação	amp; Alto
OAuthApplicationId	amp; ID aplicação OAuth	amp; Alto

2.7 Informações de Email (2 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
NetworkMessageId	amp; ID único da mensagem email	amp; Alto
EmailClusterId	amp; ID cluster de emails similares	amp; 98.98%

2.8 Informações de Registro (3 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
RegistryKey	amp; Chave do Registro do Windows	amp; Alto
RegistryValueName	amp; Nome do valor do Registro	amp; Alto
RegistryValueData	amp; Dados do valor do Registro	amp; Alto

2.9 Informações de Recursos Azure (2 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
ResourceIdName	amp; Nome do recurso Azure	amp; Muito Alto
ResourceType	amp; Tipo do recurso Azure	amp; 99.93%

2.10 Metadados (2 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
ThreatFamily	amp; Família de malware	amp; 99.21%
Roles	amp; Metadados do papel da evidência	amp; 97.71%

2.11 Informações do Sistema (2 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
OSFamily	amp; Família do SO (Windows, Linux, macOS)	amp; Baixo
OSVersion	amp; Versão do SO (Windows 10 Build 19042)	amp; Baixo

2.12 Informações de Segurança (3 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
AntispamDirection	amp; Direção antispam (Inbound/Outbound)	amp; 98.14%
SuspicionLevel	amp; Nível de suspeita (Low, Medium, High)	amp; 84.83%
LastVerdict	amp; Veredito final (Suspicious, Malware)	amp; 76.52%

2.13 Informações Geográficas (3 colunas)

Coluna	amp; **Descrição**	amp; **Valores Ausentes**
CountryCode	amp; Código ISO (US, BR, GB)	amp; Baixo
State	amp; Estado/Região	amp; Médio
City	amp; Cidade	amp; Médio

3. Variável Target (IncidentGrade)

3.1 Distribuição

Classe	amp; **Proporção**	amp; **Descrição**
BenignPositive (BP)	amp; 42-46%	amp; Incidentes informativos, baixo risco
TruePositive (TP)	amp; 19-35%	amp; Ameaças reais confirmadas
FalsePositive (FP)	amp; 21-35%	amp; Falsos alarmes, detecções incorretas

Nota: Distribuição varia por região e organização.

4. Análise de Valores Ausentes

4.1 Colunas com >95% Ausentes (REMOVER)

Coluna	amp; **% Ausentes**	amp; **Ação**
ResourceType	amp; 99.93%	amp; Remover
ActionGrouped	amp; 99.41%	amp; Remover
ActionGranular	amp; 99.41%	amp; Remover
ThreatFamily	amp; 99.21%	amp; Remover
EmailClusterId	amp; 98.98%	amp; Remover
AntispamDirection	amp; 98.14%	amp; Remover
Roles	amp; 97.71%	amp; Remover

4.2 Colunas com 50-95% Ausentes (CONSIDERAR)

Coluna	amp; **% Ausentes**	amp; **Ação**
SuspicionLevel	amp; 84.83%	amp; Considerar remoção
LastVerdict	amp; 76.52%	amp; Considerar remoção
MitreTechniques	amp; 57.46%	amp; Manter com cuidado

4.3 Colunas Completas (0% Ausentes)

Id, OrgId, IncidentId, AlertId, Timestamp, DetectorId, AlertTitle, Category, IncidentGrade, EntityType, EvidenceRole, CountryCode, OSFamily, OSVersion

5. Tipos de Dados

Tipo	amp; **Contagem**	amp; **Exemplos**
String/Varchar	amp; 20	amp; Id, AlertTitle, Sha256, IpAddress
Integer	amp; 10	amp; OrgId, IncidentId, DetectorId, DeviceId
DateTime	amp; 1	amp; Timestamp
Categorical	amp; 8	amp; Category, IncidentGrade, EntityType, OSFamily

6. Principais Observações

1. **Estrutura Hierárquica:** Evidence → Alerts → Incidents
2. **Dados Esparsos:** Muitos campos específicos de certos tipos de incidentes
3. **9.3M Duplicatas:** Requer deduplicação antes da análise
4. **Desbalanceamento:** Moderado, não extremo
5. **Período:** Junho 2024 (~2 semanas)
6. **Cobertura MITRE:** 441 técnicas diferentes mapeadas
7. **Geográfico:** 236 países, 5.3k+ organizações

7. Recomendações para Limpeza (Bronze → Silver)

1. Remover 7 colunas com >95% valores ausentes
2. Considerar remover SuspicionLevel e LastVerdict
3. Remover 9.3M registros duplicados
4. Remover 0.54% registros sem IncidentGrade
5. Converter timestamps para datetime
6. Categorizar explicitamente variáveis categóricas
7. Estratificar split treino/validação/teste pela target