

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: 6.5 milhões de registros brutos

Total de Colunas: 45 colunas

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:** 6,516,837 (na camada Bronze)
- **Total de Colunas:** 45
- **Período Observado:** Junho 2024
- **Países Únicos:** 236
- **Organizações Únicas:** 5,300+
- **Detektoren 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	Descrição	Valores Ausentes
AccountSid	SID de conta on-premises	Alto
AccountUpn	UPN/Email (usuario@dominio.com)	Alto
AccountObjectId	ID objeto Azure AD/Entra ID	Alto
AccountName	Nome da conta on-premises	Alto

2.6 Informações de Aplicação (3 colunas)

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

2.7 Informações de Email (2 colunas)

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

2.8 Informações de Registro (3 colunas)

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

2.9 Informações de Recursos Azure (2 colunas)

Coluna	Descrição	Valores Ausentes
ResourceName	Nome do recurso Azure	Muito Alto
ResourceType	Tipo do recurso Azure	99.93%

2.10 Metadados (2 colunas)

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

2.11 Informações do Sistema (2 colunas)

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

2.12 Informações de Segurança (3 colunas)

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

2.13 Informações Geográficas (3 colunas)

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

3. Variável Target (IncidentGrade)

3.1 Distribuição

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

Nota: Distribuição varia por região e organização.

4. Análise de Valores Ausentes

4.1 Colunas com >95% Ausentes (REMOVER)

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

4.2 Colunas com 50-95% Ausentes (CONSIDERAR)

Coluna	% Ausentes	Ação
SuspicionLevel	84.83%	Considerar remoção
LastVerdict	76.52%	Considerar remoção
MitreTechniques	57.46%	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	Contagem	Exemplos
String/Varchar	20	Id, AlertTitle, Sha256, IpAddress
Integer	10	OrgId, IncidentId, DetectorId, DeviceId
DateTime	1	Timestamp
Categorical	8	Category, IncidentGrade, EntityType, OSFamily

6. Principais Observações

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