Aplicación de técnicas de Machine Learning para la detección de incidentes de ciberseguridad en un SIEM basado en Elastic

Victor Alonso Burgos Torre

El presente análisis exploratorio de datos (EDA) se desarrolla sobre los registros recolectados por los agentes del SIEM basado en Elastic, con el propósito de examinar la calidad, estructura y comportamiento de las variables disponibles. Este estudio permite identificar patrones, fuentes dominantes, distribuciones horarias y posibles riesgos de sesgo o desbalance que puedan comprometer la validez del modelo supervisado. A continuación, se presentan las principales visualizaciones obtenidas y su interpretación analítica.

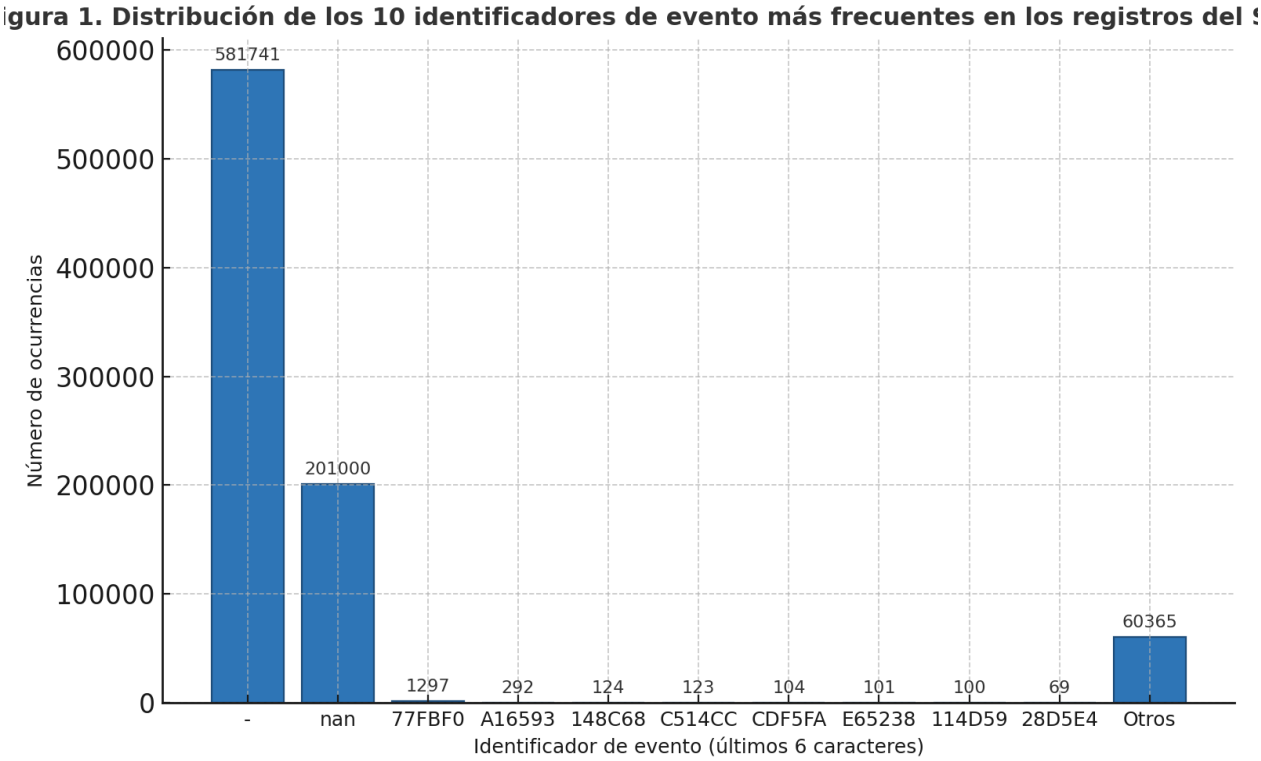

Figura 1. Identificadores de evento (event_id)

La figura presenta los 10 identificadores de evento más frecuentes en los registros del SIEM Elastic. Cada identificador representa un tipo específico de suceso, como autenticaciones, accesos o ejecuciones de procesos. Los valores fueron abreviados para mejorar la legibilidad, mostrando solo los últimos seis caracteres de cada ID. La barra 'Otros' agrupa el resto de identificadores menos frecuentes, brindando una visión global de la distribución.

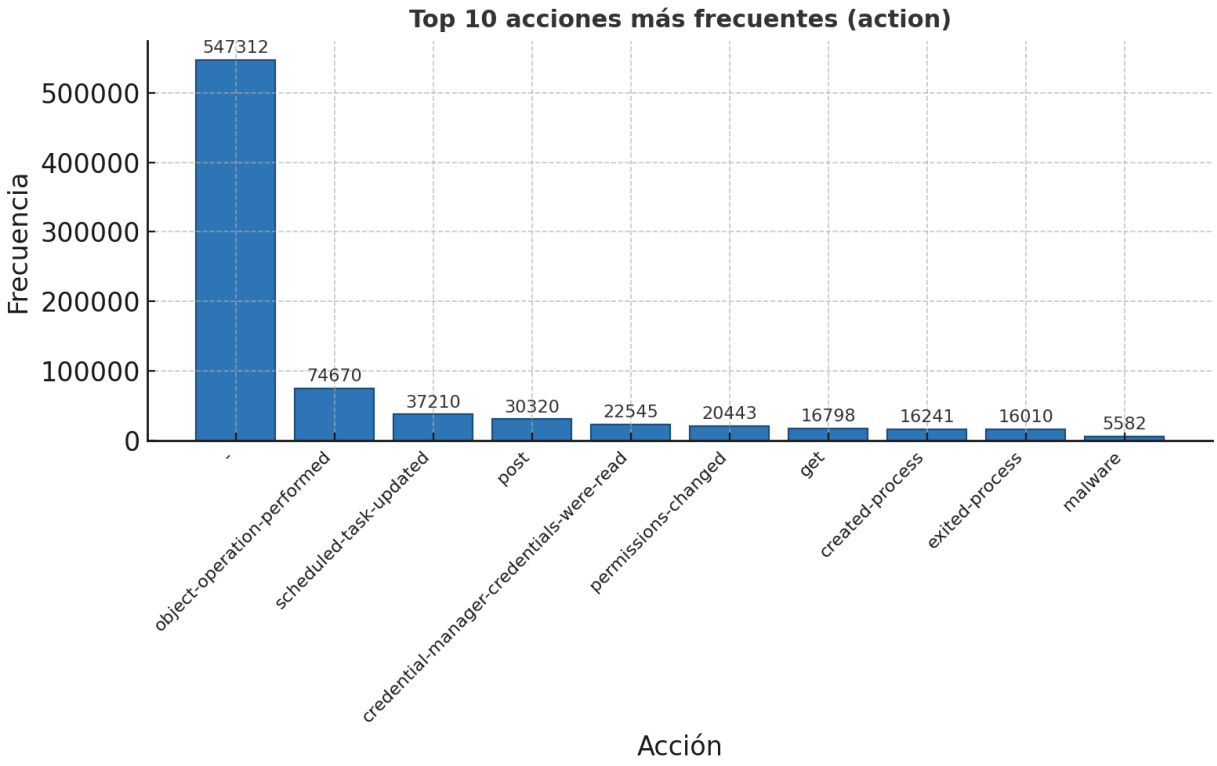

Figura 2. Acciones más frecuentes (action)

Muestra las operaciones registradas por los agentes del SIEM. Las acciones más comunes corresponden a tareas de autenticación, acceso o ejecución, mientras que las de menor frecuencia podrían asociarse a eventos atípicos o de riesgo.

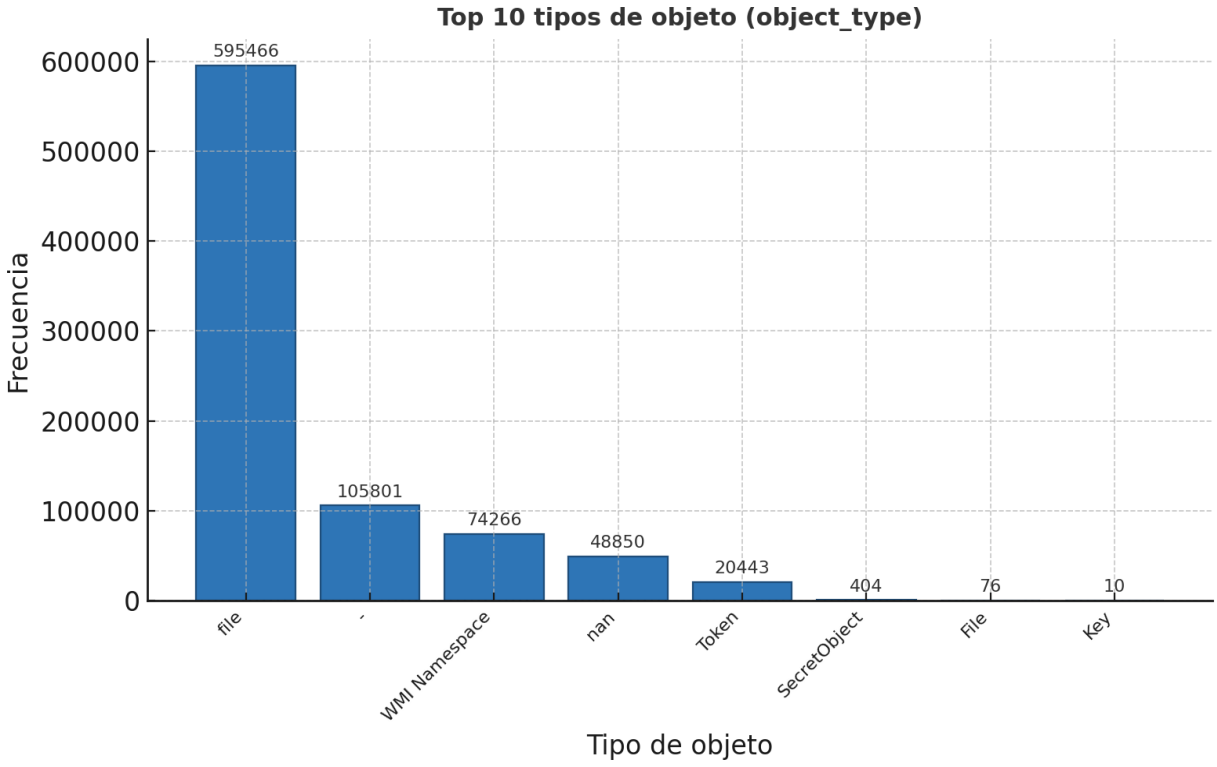

Figura 3. Tipos de objeto (object_type)

Representa los recursos del sistema sobre los que se ejecutan las acciones registradas. La concentración en ciertos objetos indica las áreas con mayor supervisión de seguridad.

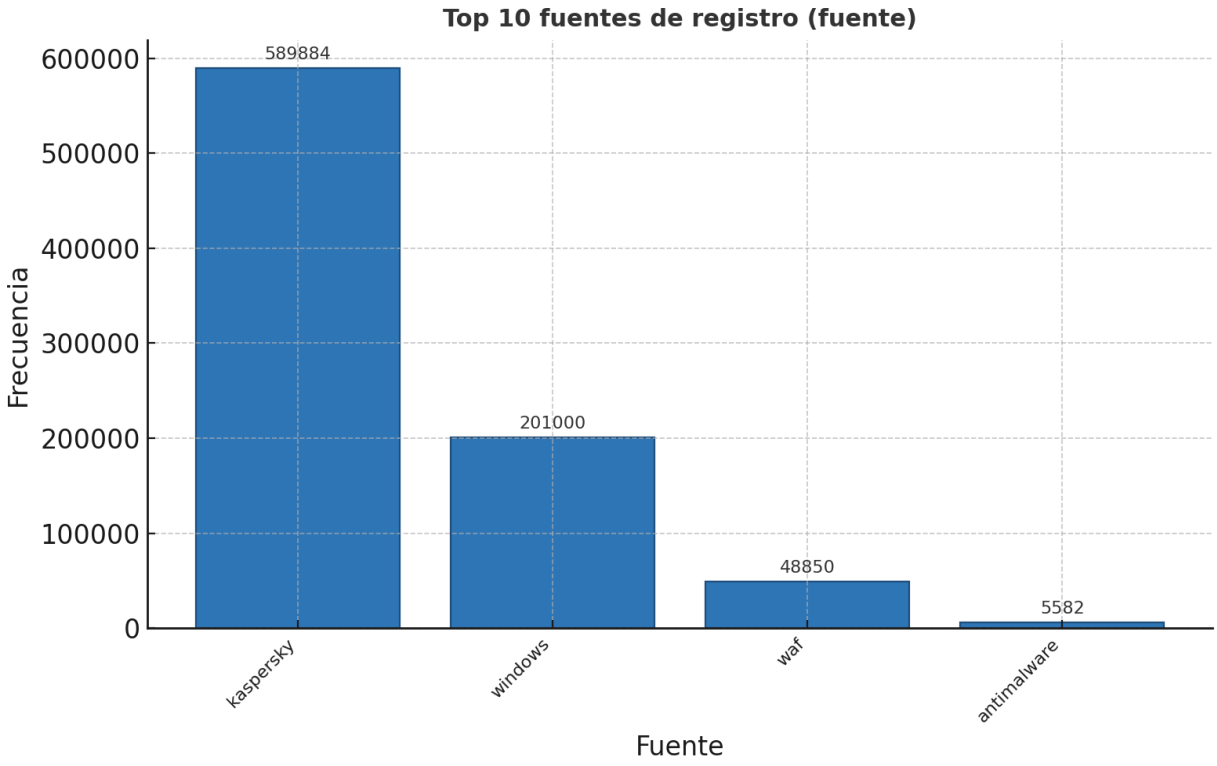

Figura 4. Fuentes de registro (fuente)

Describe la proporción de logs generados por cada agente o fuente. La distribución desigual sugiere la necesidad de balancear la representatividad de las fuentes en el modelado.

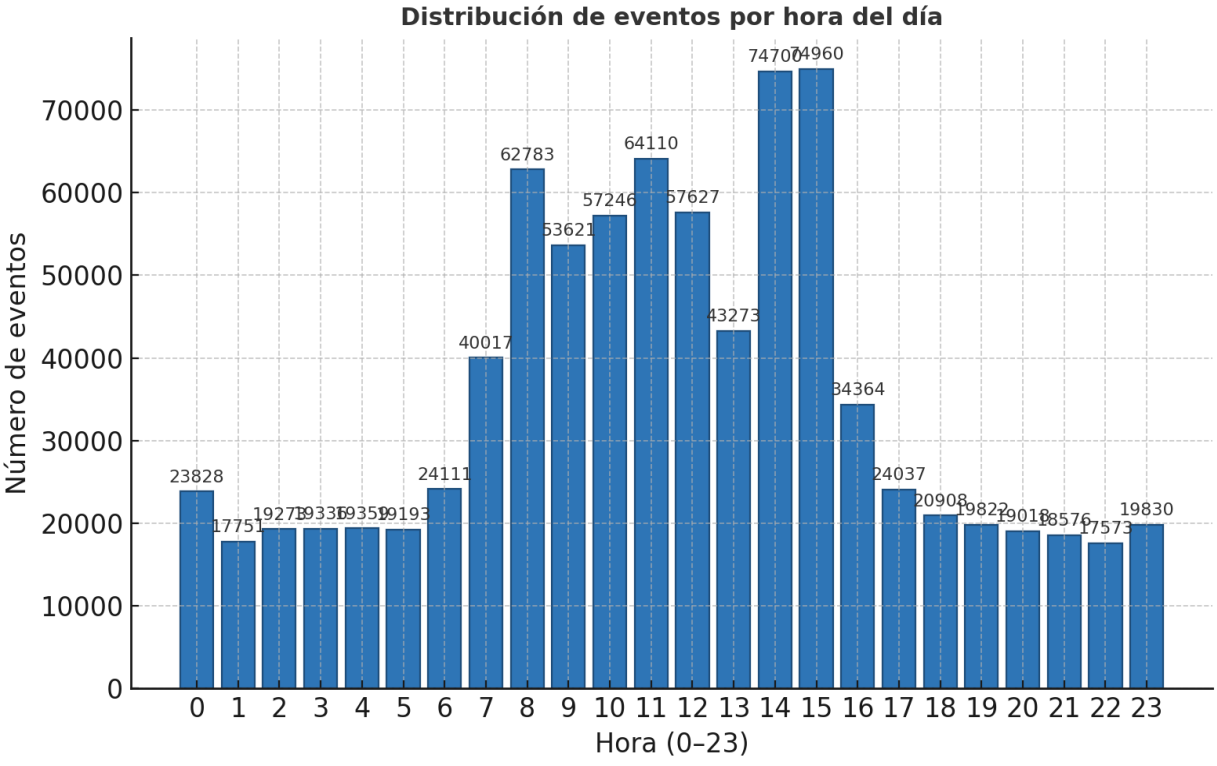

Figura 5. Distribución temporal de eventos por hora del día

Expone la variación del volumen de eventos a lo largo del día. Los picos coinciden con horarios de mayor actividad operativa y los mínimos con periodos de inactividad.

Riesgos identificados

El análisis identifica factores que podrían comprometer la estabilidad del modelo y la validez del proceso de aprendizaje. Entre ellos destacan la representación desigual de fuentes, el riesgo de fuga de información (leakage) por inclusión de variables derivadas, la posible deriva temporal (drift) causada por cambios operativos, y el desbalance de clases al incorporar la etiqueta supervisada. Estos riesgos deberán mitigarse mediante validación temporal, rebalanceo y control estadístico continuo.

Conclusiones accionables

Se establecen dos acciones prioritarias: (1) aplicar un preprocesamiento integral con depuración de valores atípicos, codificación eficiente de variables categóricas y generación de atributos temporales derivados; y (2) implementar una validación temporal con monitoreo de estabilidad estadística (PSI o KS), junto con técnicas de rebalanceo estratificado como SMOTE o undersampling al incorporar la variable objetivo. Estas medidas garantizarán la robustez, generalización y sensibilidad del modelo ante incidentes reales.

Conclusión general

El EDA confirma que el dataset proveniente de los agentes del SIEM Elastic posee una estructura coherente y suficiente representatividad para el entrenamiento de modelos supervisados. Las visualizaciones y análisis realizados consolidan una comprensión detallada de la composición del conjunto de datos y establecen bases sólidas para el desarrollo de un pipeline de aprendizaje automático orientado a la detección temprana de incidentes en entornos operativos.