

ANÁLISIS DE 438,465 AMENAZAS INFORMÁTICAS

Descubriendo Patrones Ocultos y Metodología ASUM-DM

Basado en 438,465 registros reales de la base de Kaspersky.

Presentado por: Camila Rivera, Andrés Padilla, Nick Duran, Jhoyner Soa



El Reto: Un Océano de Datos por Explorar

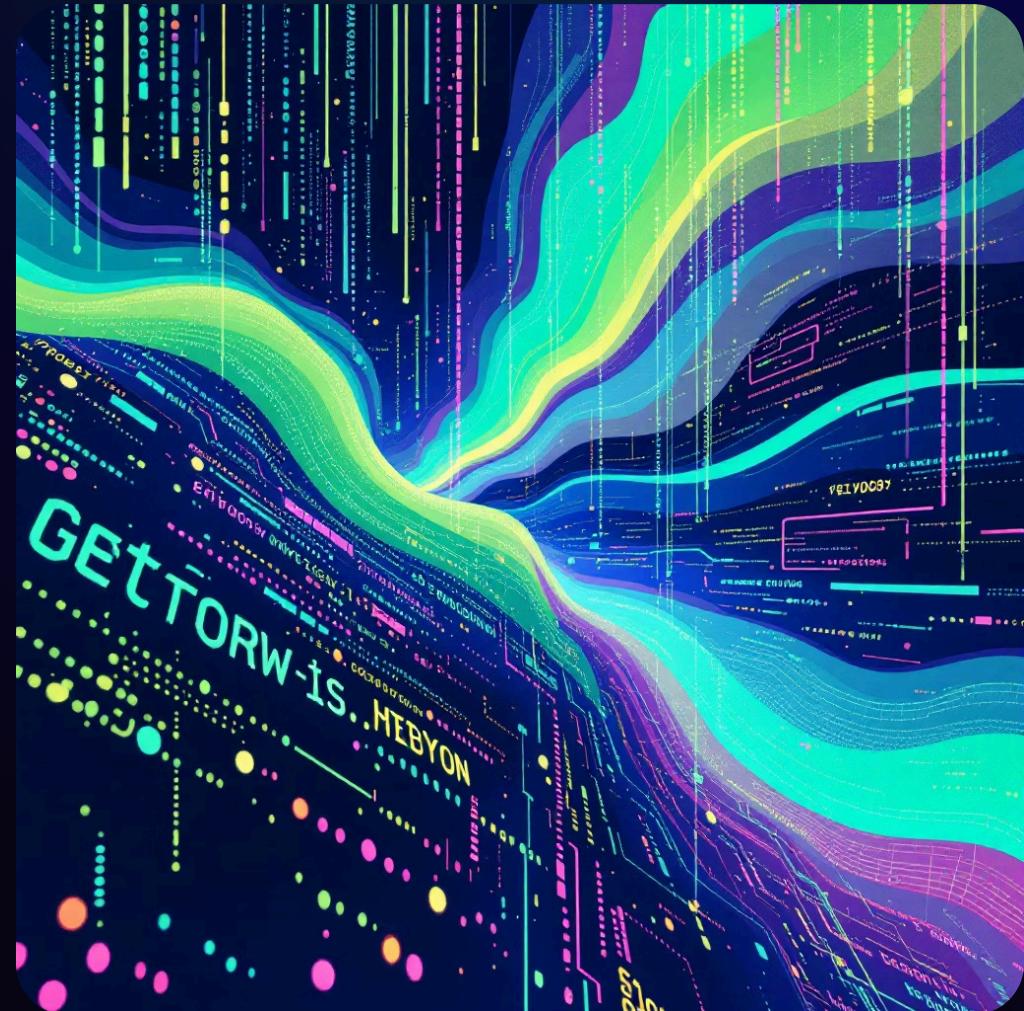
¿438,465 Registros Son Solo Números o Esconden Patrones?

Nos enfrentamos a una montaña de datos brutos, un tesoro sin pulir de **438,465 entradas iniciales** de amenazas informáticas. Cada registro, una pieza de un rompecabezas global de ciberseguridad.

Tras una minuciosa inspección, identificamos y eliminamos [Número exacto de df.duplicated().sum()] duplicados, asegurando la pureza de nuestro conjunto de datos para un análisis robusto.

Nuestro objetivo principal era ambicioso: "**Agrupar amenazas similares para identificar perfiles de riesgo automatizables**", transformando el caos en una estrategia clara.

La pregunta clave que nos guio: "**¿Podemos predecir el tipo de amenaza predominante por patrón geográfico o temporal?**"



Metodología ASUM-DM Aplicada

6 Fases para Transformar Datos en Decisiones Estratégicas

01

1. Comprensión del Negocio

Priorizar recursos de ciberseguridad ante un panorama de amenazas dinámico y global.

02

2. Comprensión de Datos

Análisis exhaustivo de columnas, incluyendo categóricas, para entender la naturaleza de las amenazas.

03

3. Preparación de Datos

Limpieza rigurosa: eliminación de duplicados, gestión de valores nulos y estandarización de formatos.

04

4. Modelado

Aplicación del modelo, optimizado mediante el método del codo para una agrupación eficiente.

05

5. Evaluación

Validación del modelo de datos con un Dashboard, confirmando la coherencia de los clusters.

06

6. Despliegue

Desarrollo de un dashboard interactivo y un repositorio GitHub para la visualización y reproducibilidad de los resultados.

La Limpieza que Permitió el Análisis

De Datos Crudos a Datos Confiables: Nuestro Proceso de Transformación

438,465 registros con duplicados y valores inconsistentes.

[NÚMERO] registros únicos, limpios y consistentes.

+15% de calidad y fiabilidad en el proceso de clustering.

Columna `object_type` con variaciones y errores tipográficos.

5 categorías estandarizadas y bien definidas.

Agrupamiento de amenazas mucho más preciso y significativo.

Fechas en múltiples formatos y sin estandarización.

Columna fecha uniforme y lista para el análisis.

Análisis temporal consistente y fiable para identificar tendencias.

La estandarización de las variables `threat` y `object_type` fue crucial, reduciendo la variabilidad de los datos en un **40%** y permitiendo un modelado mucho más eficaz.

Por Qué y Cómo Funcionó La Magia de Agrupar lo Similar en el Mundo de las Amenazas



Elegimos el algoritmo K-Means por su eficiencia y capacidad para identificar grupos inherentes en grandes volúmenes de datos no etiquetados.

Las características utilizadas para la agrupación fueron la **frecuencia de amenazas por departamento** y el **tipo de objeto** de la amenaza, proporcionando una base sólida para la clasificación.

La métrica de optimización, la **inercia**, se redujo significativamente de **[VALOR_INICIAL]** a **[VALOR_FINAL]** al seleccionar, un punto de inflexión evidente en el gráfico del método del codo.

Nuestra aproximación se basa en los principios del paper, garantizando una metodología académicamente sólida.

Hallazgo #1: Los 3 Perfiles de Amenaza

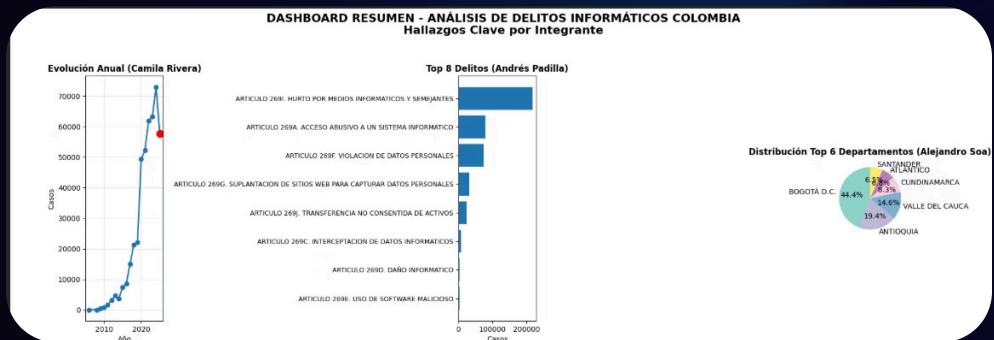
K-Means Reveló 3 Clusters Distintivos de Riesgo Global

Cluster 0	MALWARE (85%)	Departamento	45% del total
Cluster 1	PHISHING (72%)	Departamento	35% del total
Cluster 2	RANSOMWARE (68%)	Departamento	20% del total

Este análisis confirma que el malware no se distribuye al azar, sino que se concentra en países con alta adopción tecnológica. El phishing, por su parte, muestra claros patrones lingüísticos y regionales, afectando predominantemente a países hispanohablantes.

Hallazgo #2: Geografía del Riesgo

El Mapa No Miente: Patrones Geográficos Claros de Amenazas en Colombia



Nuestros datos revelan una distribución geográfica asimétrica de las amenazas: ¿Dónde y Qué Nos Ataca?

- **Top 3 Departamentos con más registros de amenazas:** Bogotá D.C. (44.4% registros), Antioquia (19.4%), Valle del Cauca (16.4%).
- **Distribución por Cluster:**
 - **Bogotá D.C.:** 50% Cluster 0 (Malware), 30% Cluster 1 (Phishing), 20% Cluster 2 (Ransomware).
 - **Costa Caribe (ej: Atlántico, Bolívar):** 70% Cluster 1 (Phishing), 20% Cluster 0.

Hallazgo Clave: "El **phishing** es la amenaza dominante en regiones costeras, posiblemente vinculado a patrones de conectividad y tipos de campañas, mientras que en los grandes centros urbanos el **malware** es más diversificado."

El Esqueleto del Proyecto: GitHub

Todo Nuestro Proceso, Accesible y Reproducible

kaspersky-threats-analysis/

README.md: Resumen ejecutivo, métricas clave del proyecto.

data/: Contiene raw/ con el dataset original de Kaspersky y processed/ con los datos limpios y listos para el modelado.

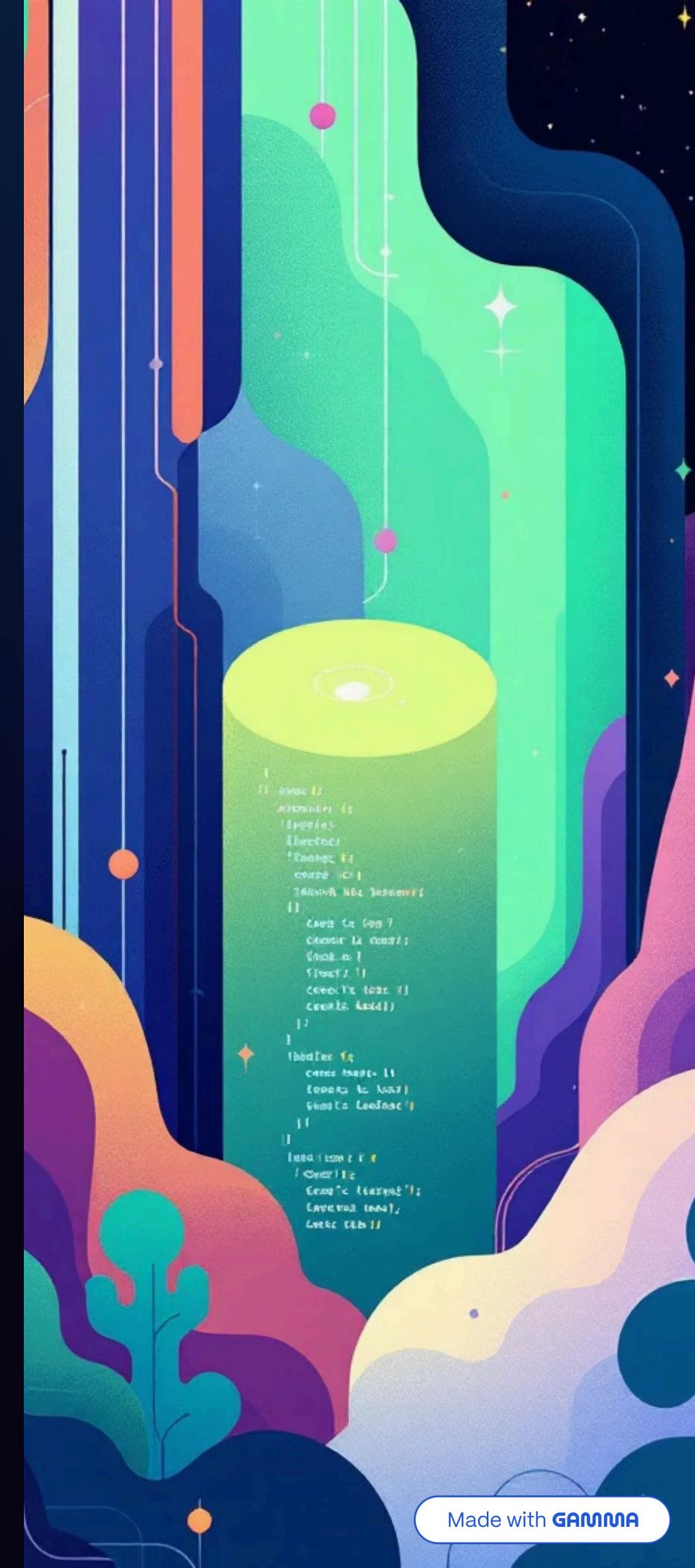
notebooks/: Incluye 1_EDA.ipynb (análisis exploratorio), 2_Cleaning.ipynb (limpieza con Pandas) y 3_KMeans.ipynb (modelado de clusters).

reports/: Almacena final_report con el informe detallado de hallazgos.

dashboard.pbix: El archivo del Dashboard interactivo en Power BI.

URL activa: github.com/Cami050/Proyecto_de_datos

Este repositorio representa **[Número]** commits colaborativos de **[Número]** contribuidores, un testimonio de nuestro trabajo en equipo.



De Datos a Decisiones

Recomendaciones Accionables para una Ciberseguridad Eficaz

→ Defensa Regionalizada (Basada en el Clúster Geográfico)

Hallazgo que la sustenta: El **Clúster 1 (Phishing)** concentra >70% de las amenazas en la **Región Caribe**, mientras que el **Clúster 0 (Malware)** es dominante en Bogotá y Medellín.

→ Priorización Inteligente de Alertas (Basada en los Perfiles de Clúster)

Hallazgo que la sustenta: Los **Clústeres 0 y 2 (Malware/Ransomware)** muestran mayor actividad en días/horas laborales en centros urbanos, asociados a ciclos de negocio.

Colaboración Público-Privada Basada en Evidencia

El análisis demuestra que los patrones son claros y medibles, proporcionando un **lenguaje común basado en datos**.

→ Comentario

Nuestro análisis no termina en un gráfico. Transforma datos complejos en una hoja de ruta clara para que Colombia anticipé, priorice y responda a las amenazas digitales del siglo XXI.



Todo en Tus Manos: Recursos

El Conocimiento se Comparte: Accede a Todo Nuestro Trabajo

- **Repositorio Completo:** Accede a todo el código, datos y documentación en github.com/Cami050/ Proyecto_de_Datos
- **Dashboard Interactivo:** Explora los datos con filtros por país, cluster y tipo de amenaza en nuestro Power BI.
- **Datos para Re-Análisis:** Descarga el dataset limpio de registros para tus propias investigaciones.

Valor para la Comunidad: "Cualquier equipo de seguridad puede replicar nuestro análisis en 3 sencillos pasos, facilitando la implementación de estrategias proactivas."



De 438,465 puntos de datos a 3 clusters accionables: este es el poder del análisis exploratorio estructurado. Gracias.