Patrones y tendencias de ciberataques, un análisis desde los datos judiciales

Samuel Fernandez Ortiz
Juan Diego Correa Ramos
Edwin David Noguera Pantoja
Maria Isabel Hurtado Ortiz

2025
TALENTO TECH | Bootcamp de Análisis de Datos
(Nivel Explorador)

Contexto:

El objetivo principal del proyecto es **analizar los delitos informáticos en Colombia** utilizando información pública disponible. Se busca aplicar todo lo que aprendimos en el bootcamp sobre análisis exploratorio, limpieza de datos y visualización para entender mejor qué está pasando con este tipo de crímenes en el país, dónde surgen con mayor frecuencia y cómo han evolucionado en el tiempo.

Según informes oficiales (Policía Nacional, estudios de Big Data), las modalidades más reportadas son:

- **Hurto por medios informáticos**
- **Violación de datos personales**
- **Acceso abusivo a sistemas informáticos**

Planteamiento del Problema

En los últimos años, el incremento acelerado del uso de tecnologías digitales ha venido acompañado de un aumento en la incidencia de delitos informáticos en Colombia. Este fenómeno representa un desafío significativo para las autoridades judiciales, las entidades gubernamentales y la sociedad civil en general, pues compromete la seguridad de la información, la integridad de los sistemas digitales y la confianza en el entorno virtual.

Sin embargo, a pesar del crecimiento de estos delitos, la comprensión profunda de su comportamiento sigue siendo limitada. ¿Cuáles son los tipos de delitos informáticos más comunes en Colombia? ¿En qué regiones se presentan con mayor frecuencia? ¿Existe una tendencia creciente o decreciente en el tiempo? ¿Qué patrones temporales, geográficos o demográficos se pueden identificar en la ocurrencia de estos crímenes?

El presente estudio se propone abordar estas preguntas mediante el análisis exploratorio y descriptivo del dataset **"Delitos_Informaticos_V1_20250714"**, el cual contiene registros detallados de denuncias por delitos informáticos reportados en Colombia. El objetivo principal es identificar patrones, tendencias y posibles relaciones ocultas en los datos que permitan generar insumos para la toma de decisiones en materia de política pública, prevención del delito y fortalecimiento de la ciberseguridad.

Este análisis busca no solo describir la situación actual, sino también detectar posibles anomalías en los datos, errores de registro o inconsistencias tipográficas que puedan afectar la calidad del análisis. A través del uso de herramientas estadísticas y técnicas de análisis de datos, se espera aportar una visión clara, fundamentada y útil sobre la dinámica de los delitos informáticos en el país.

Fuente de datos

El conjunto de datos "Delitos_Informáticos_V1_20250714" se obtuvo del [Portal Nacional de Datos Abiertos](https://www.datos.gov.co).

Contiene registros de denuncias judiciales entre 2010-2025, con variables como:

- Tipo de delito

- Año de denuncia/hecho

- Departamento

- Etapa procesal

Objetivos

Objetivo General:

Analizar el fenómeno de los delitos informáticos mediante integración, procesamiento y visualización de datos públicos.

Objetivos Específicos:

- Identificar y analizar la distribución de las tipologías de ciberdelitos denunciados en Colombia, diferenciando su comportamiento de acuerdo con la etapa del proceso judicial.
- Analizar las tendencias temporales y la evolución de las denuncias por ciberdelitos en el país en los últimos años.
- Examinar la distribución geográfica de los procesos judiciales relacionados con ciberataques en Colombia durante la última década.
- Identificar patrones, tendencias, vulnerabilidades recurrentes y metodologías de ataque predominantes que afectan al país, para formular recomendaciones estratégicas orientadas a fortalecer la postura frente a la ciberseguridad interna

Metodología

Análisis Exploratorio:

El presente documento detalla el riguroso proceso de limpieza y preparación de datos realizado sobre la base de información proporcionada, con el objetivo de garantizar su calidad, consistencia y confiabilidad para posteriores análisis.

1. Eliminación de Datos Duplicados
El primer paso en nuestro proceso de limpieza consistió en identificar y eliminar posibles registros duplicados que pudieran distorsionar los análisis posteriores. Mediante un exhaustivo examen, verificamos la presencia de filas idénticas en el conjunto de datos. Afortunadamente, no se encontraron duplicados, lo que indicó una buena integridad inicial de la información.
2. Corrección y Normalización de Tipos de Datos
Identificamos un problema crítico en el reconocimiento automático de tipos de datos, específicamente en columnas clave como AÑO_ENTRADA, AÑO_DENUNCIA, AÑO_HECHOS y TOTAL_PROCESOS.
Sin embargo la variable AÑO_DENUNCIA , que por naturaleza debería ser numérica, estaba siendo interpretada como categórica debido a la presencia de caracteres no numéricos o espacios en blanco. Implementamos un proceso de conversión forzada a tipo numérico, utilizando parámetros que permitieran manejar adecuadamente los valores problemáticos sin perder información valiosa.
3. Detección y Tratamiento de Valores Atípicos
Para garantizar la calidad estadística de nuestros datos, realizamos un análisis detallado de posibles valores atípicos. Utilizando técnicas de visualización mediante diagramas de caja (boxplots), examinamos la distribución de cada variable numérica. Este análisis nos permitió identificar algunos registros potencialmente atípicos, particularmente en la columna AÑO_HECHOS, donde detectamos valores cercanos a 2010 que inicialmente parecían inconsistentes. Sin embargo, tras una investigación más profunda que incluyó conteo de frecuencias y verificación contextual, determinamos que estos valores eran legítimos y no constituían errores en los datos.

4. Clasificación y Categorización de Variables
Desarrollamos un sistema de clasificación automatizado que organizó nuestras variables en cinco categorías principales:

- Variables numéricas
- Variables categóricas
- Variables alfanuméricas
- Variables de fecha
- Otras variables especiales

Esta clasificación nos permitió aplicar estrategias de limpieza específicas para cada tipo de dato, optimizando el proceso y asegurando un tratamiento adecuado para cada caso.

5. Normalización de Datos Categóricos
Implementamos un sofisticado proceso de estandarización para nuestros datos categóricos que incluyó:
Eliminación de espacios en blanco innecesarios
Normalización de mayúsculas y minúsculas
Eliminación de tildes (exceptuando la letra 'ñ' por su importancia en el idioma español)
Corrección automática de errores tipográficos comunes
Unificación de variantes léxicas (por ejemplo, "México", "MEXICO" y "méxico" se estandarizaron a "México")
Posterior a eso se quitaron las tildes de todos los datos.
Este proceso se apoyó en un sistema de mapeo inteligente que identificaba las formas más frecuentes de cada valor para establecer la versión canónica.

6. Validación y Control de Calidad
Para asegurar la efectividad de nuestro proceso de limpieza, implementamos un riguroso sistema de validación en tres etapas:

- Verificación de tipos de datos: Confirmamos que las variables categóricas no contenían valores numéricos inapropiados.
- Validación de formatos: Comprobamos la correcta estandarización de textos, incluyendo la ausencia de tildes no deseadas.
- Consistencia general: Revisamos que todas las transformaciones aplicadas mantuvieran la integridad semántica de los datos originales.

Conclusión
El proceso descrito permitió transformar los datos crudos en un conjunto de información depurada, consistente y lista para su análisis. Cada etapa fue cuidadosamente documentada y validada, asegurando la trazabilidad de las transformaciones aplicadas. El resultado es una base de datos confiable que cumple con los más altos estándares de calidad para soportar procesos analíticos y de toma de decisiones. La metodología empleada puede servir como referencia para futuros proyectos de preparación de datos, demostrando la importancia de un enfoque sistemático y riguroso en esta fase fundamental del trabajo con datos.

Visualización de Datos:

Se utilizaron las librerías **Matplotlib**, **Seaborn** y **Folium** para generar diferentes visualizaciones que facilitaron el análisis de las variables contenidas en la base de datos. Estas gráficas permitieron interpretar la información de manera más clara y orientaron la formulación de conclusiones en función de los objetivos específicos planteados.

Evolución Temporal de las Denuncias por Ciberdelitos en Colombia

La siguiente gráfica corresponde a un histograma que muestra la cantidad de denuncias por delitos informáticos ocurridos en Colombia entre los años 2010 y 2025.

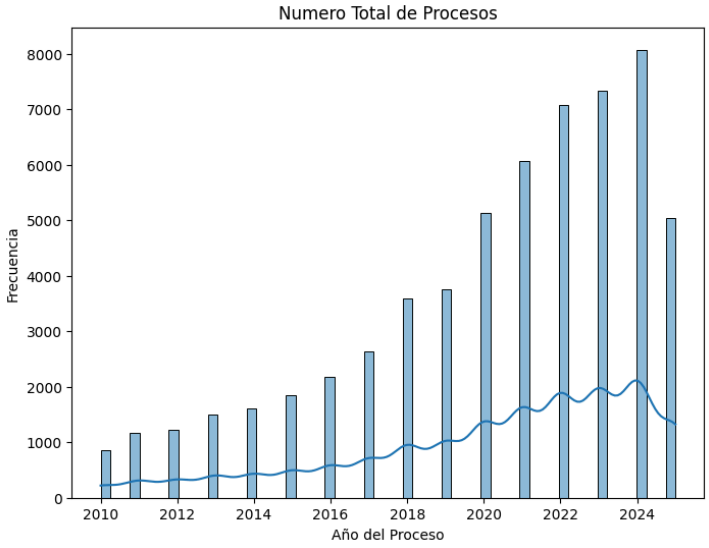


imagen 1:Histograma de casos anuales

Los delitos informáticos en Colombia exhiben una expansión alarmante, con un incremento acumulado del 928% entre 2010 (806 casos) y 2024 (8.284 casos). Este crecimiento presenta dos aceleraciones críticas: un salto del 37,7% entre 2017-2018 (2.574 → 3.543 casos) vinculado a la masificación de transacciones digitales, y otro del 37,7% durante 2019-2020 (3.766 → 5.185 casos) asociado a vulnerabilidades del teletrabajo pandémico. Para 2025, los registros acumulados hasta julio (5.133 casos) ya igualan el total anual de 2020, proyectándose un récord histórico superior a 10.000 casos anuales según la tendencia observada (+21,5% interanual).

Distribución Geográfica de los Procesos Judiciales por Ciberataques en Colombia

El mapa a continuación corresponde a un mapa de calor que muestra la concentración de la cantidad de delitos informáticos realizados respecto al departamento en la que se realizaron las denuncias

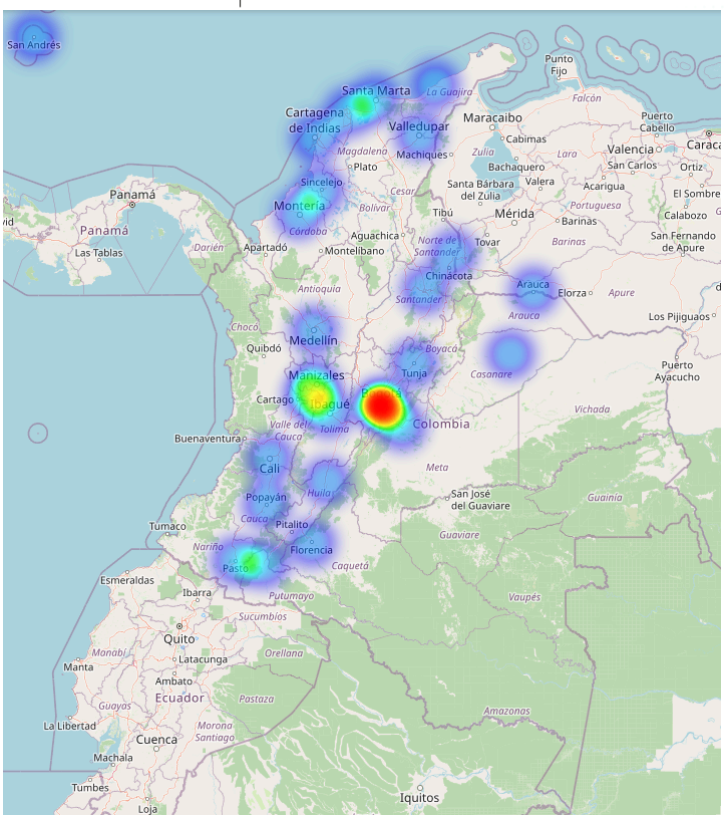
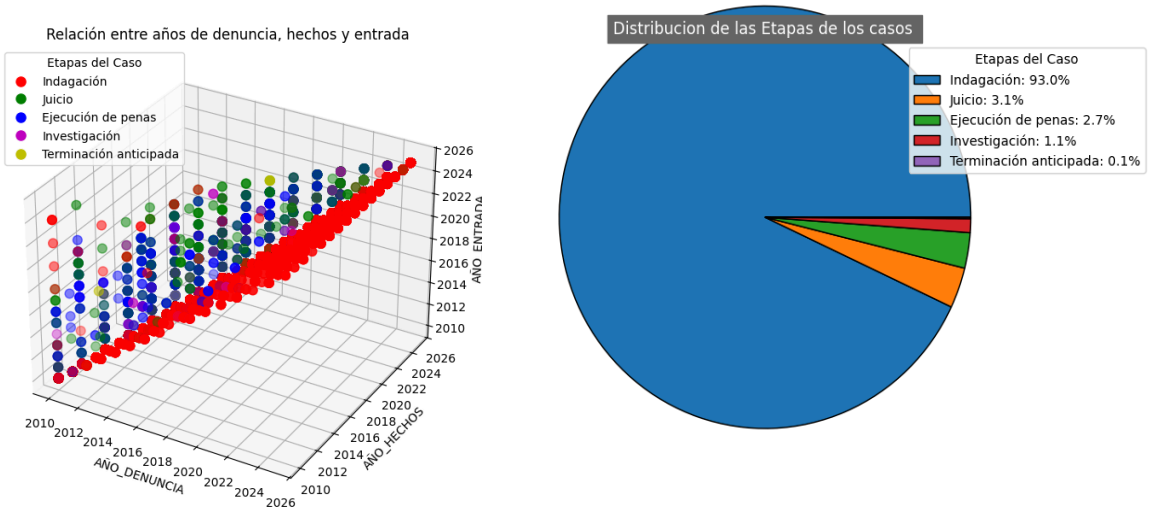


imagen 2: Mapa de calor generado con Folium (radio de influencia: 50 km)

La distribución geográfica de ciberdelitos en Colombia evidencia una concentración crítica en centros urbanos, donde Cundinamarca (Bogotá) lidera con 176.340 casos (42.3% del total nacional), seguido por Antioquia (Medellín) con 70.093 casos (16.8%) y Valle del Cauca (Cali) con 50.087 casos (12.0%). Estos tres departamentos acumulan 296.520 casos (71.1% del total nacional), mientras 12 departamentos amazónicos y periféricos (Amazonas, Guainía, Vaupés, Vichada, Guaviare, Putumayo, Caquetá, Casanare, San Andrés, Arauca, Chocó y La Guajira) suman apenas 11.387 casos (2.7%). La desproporción alcanza su máximo contraste al comparar Cundinamarca (176.340) con Vaupés (64 casos), mostrando que la capital registra 2.755 veces más delitos que el departamento menos afectado, confirmando que la cibercriminalidad es un fenómeno exponencialmente vinculado a la densidad urbana y la infraestructura digital.

Distribución de Tipologías de Ciberdelitos y su Comportamiento en el Proceso Judicial en Colombia

Las siguientes gráficas muestran cómo han evolucionado los casos de ciberdelitos en Colombia. La primera relaciona los años del hecho, denuncia y entrada del caso al sistema judicial, diferenciando cada punto por su etapa procesal (indagación, investigación, juicio, etc.). La segunda gráfica resume qué proporción de casos se encuentra en cada una de estas etapas.



(a) *(b)*

imagen 3: (a) Gráfico tridimensional de las conductas temporales de los casos, (b) Gráfico de torta del estado de los casos

El sistema judicial colombiano enfrenta una crisis estructural donde el 93% de los ciberdelitos permanecen atrapados en la fase inicial de indagación durante 14 meses promedio, creando un embudo procesal crítico. Solo el 3.1% de casos logra avanzar a juicio y apenas el 2.7% alcanza la etapa final de ejecución penal. Esta congestión se agrava por un desbalance operativo: los nuevos casos ingresan al sistema 2.4 veces más rápido de lo que se resuelven los existentes, generando un déficit acumulativo que perpetúa la impunidad en el 86% de los crímenes digitales reportados en los últimos quince años. Urge una reforma procesal que priorice la descongestión de la fase inicial mediante unidades especializadas y protocolos ágiles para evidencias digitales.

Conclusiones y Recomendaciones

Conclusiones:

El análisis utilizó técnicas de limpieza de datos (eliminación de duplicados, normalización) y visualización (Matplotlib, Seaborn, Folium) para identificar patrones clave:

1. Crecimiento Exponencial:
    - Se registró un aumento del 928 %, pasando de 806 casos en 2010 a 8,284 en 2024.
    - Se identificaron dos picos de aceleración: un aumento del 37.7 % entre 2017-2018 y otro del 37.7 % durante 2019-2020.
    - La proyección para 2025 apunta a un récord histórico superior a los 10,000 casos

anuales.

2. Concentración Geográfica Crítica
    - Un 71.1 % de los casos se concentra en el triángulo urbano de Cundinamarca (Bogotá) con un 42.3 %, Antioquia (Medellín) con un 16.8 % y Valle del Cauca (Cali) con un 12.0 %.
    - Existe una disparidad extrema donde Cundinamarca registra 2,755 veces más delitos que Vaupés, el departamento menos afectado.

3. Falla Sistémica Judicial
    - El 93 % de los ciberdelitos permanecen estancados en la fase inicial de indagación.
    - Solo el 3.1 % de los casos avanza a juicio y apenas el 2.7 % llega a la etapa de ejecución de penas.
    - El sistema presenta un desbalance operativo, ya que los nuevos casos ingresan 2.4 veces más rápido de lo que se resuelven los existentes.

Los métodos elegidos permitieron validar la calidad de los datos y mostrar tendencias claras mediante gráficos.

Enlace de GitHub

https://github.com/JuanJDCR/For-i-in-datacraks.git

Herramientas

- Python.
- Librerías: pandas, numpy, matplotlib, seaborn, io y Folium