Análisis Exploratorio Y Preprocesamiento De Datos De Homicidios En Colombia (2015-2023)

Y.Y. Mora Segura, A.D. Ramírez Chiquillo, J.D. Ramírez Castañeda

***Abstract*—Analyzing data on violent events such as homicides is crucial for understanding their causes and contexts and designing effective prevention strategies. However, before applying predictive models or making in-depth inferences, it is essential to conduct a thorough exploratory analysis and careful preprocessing of the available data. This article details the first steps of this process using a homicide dataset.**

***Index Terms*—Exploratory data analysis, homicides, Colombia, missing values, preprocessing, classification, association, data mining.**

I. Introduction

EL análisis de datos de eventos violentos como los homicidios es fundamental para obtener una comprensión profunda de sus raíces, los entornos en los que ocurren y para la formulación de estrategias de prevención que sean efectivas. No obstante, antes de la implementación de modelos predictivos sofisticados o la extracción de conclusiones significativas, se requiere un análisis exploratorio exhaustivo y una preparación meticulosa de los datos existentes. Este artículo describe las etapas iniciales de este procedimiento, utilizando un conjunto de datos específico de homicidios en Colombia.

II. Descripción del Conjunto de Datos

El conjunto de datos analizado proviene del portal de Datos Abiertos Colombia, bajo el título "Presuntos Homicidios. Colombia, 2015 a 2023. Cifras definitivas". Este repositorio contiene registros oficiales de homicidios ocurridos en el territorio colombiano, recopilados por Instituto Nacional de Medicina Legal y Ciencias Forenses. El dataset completo comprende 111.263 registros de presuntos homicidios, cada uno caracterizado por 35 variables que describen aspectos relacionados con:

- **Características de la víctima:** Información sociodemográfica como edad, sexo, escolaridad, estado civil, país de nacimiento y pertenencia grupal.
- **Características espacio-temporales del hecho:** Ubicación geográfica (departamento, municipio, zona), fecha y franja horaria del suceso.
- **Circunstancias y modalidades del homicidio:** Mecanismo causal, diagnóstico de lesiones, escenario, circunstancia y actividad durante el hecho.
- **Información sobre el presunto agresor:** Cuando está disponible, categorización del autor del homicidio con relacion a la víctima.

Es importante destacar que los datos representan "presuntos" homicidios, lo que implica que son casos clasificados preliminarmente como homicidios por las autoridades, pero que podrían estar sujetos a reclasificación posterior durante el proceso judicial [1]. Esta consideración es relevante al momento de interpretar los resultados y establecer conclusiones. La temporalidad del conjunto abarca nueve años (2015-2023), lo que permite observar tendencias y variaciones a lo largo del tiempo. La cobertura geográfica incluye la totalidad del territorio colombiano, con representación de los 32 departamentos y el Distrito Capital, aunque con variaciones significativas en la cantidad de casos registrados por región.

III. Análisis Exploratorio del Conjunto de Datos

El propósito del Análisis Exploratorio de Datos (AED) realizado fue comprender la estructura, calidad y características principales del conjunto de datos, con particular énfasis en las variables categóricas que conforman la mayoría de los atributos disponibles. Este análisis inicial es fundamental para orientar decisiones de preprocesamiento y para la identificación de variables con potencial explicativo.

1) Calidad de los datos: El análisis reveló varios desafíos relacionados con la calidad de los datos:

- **Valores faltantes heterogéneos:** Se identificaron múltiples representaciones de datos ausentes ("No aplica", "Sin información", "Desconocido", espacios en blanco), lo que complicó su cuantificación exacta.
- **Altos porcentajes de datos faltantes en variables clave:** Variables como "Circunstancia del Hecho" (64.16%), "Actividad Durante el Hecho" (38.89%) y "Escolaridad" (21.01%) presentaron porcentajes significativos de valores faltantes.
- **Variables redundantes o no informativas:** Se encontraron siete variables con un único valor ("No aplica") en todos los registros: "Condición de la Víctima", "Medio de Desplazamiento", "Servicio del Vehículo", "Clase o Tipo de Accidente", "Objeto de Colisión", "Servicio del Objeto de Colisión" y "Razón del Suicidio".
- **Inconsistencias en la codificación:** Se detectaron variaciones ortográficas para la misma categoría en variables como "Escenario del Hecho" (por ejemplo, "Vía pública" y "Vía Pública" como entradas separadas).

Estos hallazgos señalan la necesidad de un preprocesamiento exhaustivo antes de cualquier modelado predictivo.

2) Distribución temporal de los casos: El análisis temporal de los homicidios mostró patrones importantes:

- **Tendencia anual:** Se observó una ligera tendencia incremental en el número de homicidios durante el periodo estudiado, con la excepción notable del año 2020, donde

se registró una disminución que lo situó en niveles similares a 2017. Esta anomalía podría estar relacionada con las restricciones de movilidad implementadas durante la pandemia de COVID-19, aunque esta hipótesis requeriría análisis adicionales.
- **Estacionalidad mensual:** Los meses con mayor incidencia de homicidios fueron diciembre (10.494 casos) y mayo (9.629 casos), lo que podría relacionarse con periodos festivos y de mayor actividad social.
- **Distribución semanal:** Los días de fin de semana concentraron la mayor cantidad de casos, con el domingo (23.263 casos) y el sábado (17.924 casos) como los días de mayor incidencia.

3) Patrones espaciales: La distribución geográfica de los homicidios reveló concentraciones significativas:

- **Concentración urbana:** Las cabeceras municipales registraron 78.280 casos (70.4% del total), mientras que las áreas rurales reportaron 25.891 casos (23.3%) y los centros poblados 5.666 casos (5.1%).
- **Departamentos con mayor incidencia:** Valle del Cauca (21.124 casos), Antioquia (17.133 casos) y Bogotá D.C. (10.412 casos) concentraron el mayor número de homicidios, coincidiendo con las áreas más pobladas del país pero también con zonas históricamente afectadas por diversas dinámicas de violencia.
- **Municipios críticos:** Bogotá D.C. (10.412 casos), Cali (8.997 casos) y Medellín (4.402 casos) fueron los municipios con mayor número absoluto de casos.

Estos patrones espaciales sugieren la importancia de factores urbanos y regionales específicos en la distribución de la violencia homicida.

4) Perfil demográfico de las víctimas: Las características sociodemográficas de las víctimas mostraron patrones de vulnerabilidad específicos:

- **Distribución por sexo:** El 91.8% de las víctimas fueron hombres (102.174 casos), frente a un 8.1% de mujeres (8.996 casos), evidenciando un desbalanceo.
- **Grupos etarios:** Los jóvenes adultos entre 20-29 años concentraron la mayor proporción de víctimas (42.314 casos, 38%), seguidos por adultos de 30-39 años (28.367 casos, 25.5%).
- **Nivel educativo:** Entre los datos disponibles, predominaron víctimas con educación básica primaria (35.496 casos) y educación media o secundaria (33.319 casos), sugiriendo una posible relación con condiciones socioeconómicas.
- **Ancestro racial:** La categoría "Mestizo" fue predominante (85.983 casos, 77.3%), seguida de "Negro" (9.645 casos, 8.7%).

Estos resultados señalan la existencia de perfiles de riesgo diferenciados, con una clara sobrerrepresentación de hombres jóvenes como víctimas de homicidio.

5) Características de los homicidios: El análisis de las modalidades y circunstancias de los homicidios reveló:

- **Mecanismo causal:** El uso de armas de fuego fue el mecanismo predominante (81.790 casos, 73.5%), seguido por armas cortopunzantes (19.406 casos, 17.4%), lo que refleja patrones de acceso y uso de armas en contextos violentos.
- **Escenario del hecho:** La vía pública fue el escenario más común (57.475 casos, 51.7%), seguido por viviendas (14.924 casos, 13.4%) y espacios abiertos como bosques o playas (7.666 casos, 6.9%).
- **Circunstancia del hecho:** Entre los casos con información disponible, destacaron "Ajuste de cuentas" (8.598 casos), "Riña" (7.089 casos) y "Sicariato" (5.749 casos), señalando la relevancia de la violencia instrumental, interpersonal y organizada.
- **Presunto agresor:** El 52.9% de los casos (58.883) no tenían información sobre el agresor, mientras que un 31.3% (34.861) fueron clasificados como "Agresor desconocido", lo que refleja desafíos importantes para el manejo de datos y uso de futuros modelos de asociación o predicción.

Estos datos proporcionan un panorama sobre las dinámicas predominantes de la violencia homicida en Colombia durante el periodo estudiado.

A. Hallazgos Principales del Análisis Exploratorio

El análisis exploratorio reveló varios aspectos críticos sobre la naturaleza y calidad de los datos:

- **Desafíos de calidad:** La presencia significativa de datos faltantes en variables clave constituye un desafío metodológico importante, que requerirá estrategias específicas de imputación o incorporación explícita de la ausencia de información como una categoría analítica.
- **Patrones temporoespaciales:** La distribución de homicidios muestra patrones claros en términos temporales (mayor incidencia en fines de semana y meses específicos) y espaciales (concentración en cabeceras municipales y departamentos específicos), que deben ser considerados en la interpretación y modelado.
- **Perfiles demográficos:** La clara sobrerrepresentación de hombres jóvenes entre las víctimas señala la necesidad de explorar factores específicos de género y edad en la violencia homicida.
- **Modalidades predominantes:** El predominio de armas de fuego, escenarios públicos y circunstancias relacionadas con ajustes de cuentas, riñas y sicariato apunta hacia la coexistencia de diversas formas de violencia (organizada, interpersonal, oportunista).

Estos hallazgos sientan las bases para las decisiones de preprocesamiento y selección de características que se abordan en las siguientes secciones.

IV. Preprocesamiento de Datos

Basados en los hallazgos del análisis exploratorio, se desarrolló una estrategia integral de preprocesamiento para abordar los desafíos identificados y preparar los datos para análisis más avanzados.

A. Estrategia de Limpieza y Transformación

1) Eliminación de variables no informativas: Se identificaron y eliminaron variables que no aportaban información relevante:

- **Variables con un único valor:** Se eliminaron siete variables que contenían "No aplica" en todos los registros: "Condición de la Víctima", "Medio de Desplazamiento", "Servicio del Vehículo", "Clase o Tipo de Accidente", "Objeto de Colisión", "Servicio del Objeto de Colisión" y "Razón del Suicidio".
- **Variables redundantes:** Se eliminaron códigos DANE duplicados y se optó por mantener solo "Edad natural", eliminando "Edad judicial" y "Ciclo Vital" por redundancia informativa.

2) Estandarización de categorías: Se identificaron y corrigieron inconsistencias en la codificación de variables categóricas:

- **Corrección de errores tipográficos:** Se identificaron y unificaron categorías duplicadas en variables como "Escenario del Hecho" y "Estado Civil" (Por ejemplo "Vía pública" y "Vía Pública").
- **Estandarización de valores faltantes:** Se normalizaron las diversas representaciones de datos ausentes ("Sin información", "Desconocido", espacios en blanco) a un formato consistente.

B. Manejo de Datos Faltantes

La estrategia para el tratamiento de valores faltantes varió según la naturaleza y porcentaje de ausencia en cada variable:

1) Eliminación de registros: Se aplicó eliminación selectiva para:

- **Variables críticas con bajo porcentaje de faltantes:** Se eliminaron registros con valores faltantes en "Edad" (0.1%) y "Fecha" (0.21%), ya que estas variables son fundamentales y su porcentaje de ausencia era mínimo.

2) Imputación de valores: Se realizaron imputaciones para:

- **Variables categóricas con niveles moderados de ausencia:** En "Pertenencia Grupal" (16.81% de datos faltantes), se imputó la categoría modal "Ninguno", bajo el supuesto de que la ausencia de información sobre pertenencia grupal suele indicar ausencia de pertenencia específica.

3) Inclusión explícita de valores faltantes: Para variables con alto porcentaje de datos ausentes:

- **Incorporación como categoría analítica:** En variables como "Circunstancia del Hecho" (64.16% de ausencia) y "Actividad Durante el Hecho" (38.89% de ausencia), la categoría "Sin información" se mantuvo como una categoría válida para el análisis, reconociendo que la ausencia de datos puede ser informativa en sí misma.

C. Agrupación de Categorías

Para abordar la alta cardinalidad y la presencia de categorías raras, se implementaron estrategias de agrupación:

1) Agrupación por frecuencia:

- **Umbral de frecuencia:** Se estableció un umbral del 1% para identificar categorías poco frecuentes, que fueron candidatas para agrupación.
- **Categoría "Otros":** Se creó una categoría "Otros" para agrupar valores con frecuencia inferior al umbral establecido en variables como "País de Nacimiento de la Víctima", donde las categorías distintas a "Colombia" y "Venezuela" representaban menos del 2% de los casos totales.

2) Agrupación conceptual: Para la variable "Circunstancia del Hecho", se realizó una agrupación basada en criterios conceptuales:

- **Violencia interpersonal:** Agrupando categorías como "Riña", "Violencia de pareja" y "Celos".
- **Crimen organizado/sicariato:** Unificando "Sicariato", "Ajuste de cuentas" y "Hurto".
- **Conflicto armado/terrorismo:** Incluyendo "Acción grupos alzados al margen de la ley", "Enfrentamiento armado" y "Acto terrorista".
- **Violencia institucional/Estado:** Agrupando "Acción militar", "Intervención Legal" y "Retención legal".
- **Violencia contra población vulnerable:** Unificando "Feminicidio", "Violencia a niños, niñas y adolescentes" y "Violencia al adulto mayor".

Esta agrupación redujo las 50 categorías originales a un conjunto más manejable de 7 categorías principales, facilitando el análisis y potencial modelado posterior.

D. Codificación para Modelado

Para preparar los datos para su uso en modelos estadísticos o de aprendizaje automático, se aplicaron estrategias de codificación:

- **One-hot encoding:** Se aplicó a variables categóricas con número limitado de categorías (menos de 10), como "Sexo", "Estado Civil" y "Grupo Mayor Menor de Edad". Este método permite transformar variables categóricas en representaciones numéricas manteniendo toda la información sin establecer relaciones de ordinalidad.

V. Selección de Características

Durante esta fase se realizó un análisis exploratorio detallado con el objetivo de identificar las variables categóricas más relevantes para el análisis posterior. Se utilizaron técnicas estadísticas y métricas de información que permiten evaluar tanto la variabilidad interna de las variables como las relaciones entre ellas. Las principales técnicas empleadas fueron el análisis de entropía, la prueba de Chi-cuadrado y el coeficiente V de Cramer, seleccionadas por su capacidad para evaluar variables categóricas de forma objetiva y cuantificable [4].

A. Distribuciones y Entropía

Se comenzó con un análisis de las distribuciones de frecuencia de las variables categóricas utilizando conteos (`value_counts`) para detectar concentraciones en ciertas categorías. Por ejemplo, se observaron distribuciones muy

dominadas por categorías como "Hombre" en "Sexo de la víctima" o "Colombia" en "País de nacimiento de la víctima", lo cual sugiere baja variabilidad.

Para cuantificar esta variabilidad se utilizó la entropía, una medida derivada de la teoría de la información [5]. La entropía se define como:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

donde $p(x_i)$ es la probabilidad de ocurrencia de la categoría x_i de una variable categórica X. Una entropía más alta indica que los valores están más distribuidos entre las categorías posibles, lo cual sugiere una mayor capacidad de la variable para diferenciar entre observaciones.

Esta métrica permitió identificar variables con alta riqueza informativa, tales como:

- Municipio del hecho DANE (entropía: 5.198)
- Departamento del hecho DANE (2.899)
- Mes del hecho (2.483)
- Escenario del hecho (1.849)
- Escolaridad (1.798)
- Actividad durante el hecho (1.753)
- Rango de hora del hecho (1.539)

Estas variables, por su distribución uniforme de valores, poseen un alto potencial para aportar información relevante en análisis descriptivos o modelos predictivos.

B. Relaciones Entre Variables

Para evaluar la relación entre variables categóricas, se recurrió a la prueba de Chi-cuadrado y al coeficiente V de Cramer [6].

Estas herramientas permiten identificar asociaciones estadísticas entre pares de variables.

a) Chi-cuadrado: La prueba de Chi-cuadrado (χ^2) evalúa si existe una relación estadísticamente significativa entre dos variables categóricas. Se calcula a partir de una tabla de contingencia y compara las frecuencias observadas con las esperadas bajo la hipótesis de independencia:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

donde O es la frecuencia observada y E la esperada. Un valor alto de χ^2 y un p-valor bajo sugieren una relación significativa entre las variables.

b) Coeficiente V de Cramer: El V de Cramer se basa en el estadístico Chi-cuadrado, pero normaliza el resultado en una escala entre 0 y 1, facilitando la comparación de la fuerza de asociación:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

donde n es el tamaño de la muestra y k es el menor número de categorías entre las dos variables. Un valor cercano a 0 indica independencia, mientras que valores más cercanos a 1 indican una asociación fuerte [6].

El análisis reveló asociaciones significativas, entre ellas:

- Municipio del hecho DANE y Departamento del hecho DANE: V de Cramer = 0.990 (alta redundancia)
- Circunstancia del hecho con: "Municipio del hecho" (0.304)

Estas relaciones sugieren que ciertas variables capturan contextos comunes o dependencias importantes entre dimensiones del evento, lo que puede ser aprovechado en tareas de modelado.

C. Selección Basada en Métricas

La decisión de qué variables conservar se basó en una combinación de alta entropía (variabilidad interna) y alta asociación (dependencia estadística con otras variables relevantes). Este enfoque permite reducir la redundancia y seleccionar un subconjunto de características más informativo y eficiente.

Las variables prioritarias identificadas incluyen:

- Municipio del hecho DANE: Máxima entropía (5.198), útil para patrones espaciales.
- Escenario del hecho y Actividad durante el hecho: Alta entropía y correlaciones significativas, aportan contexto situacional.
- Circunstancia del hecho: Aunque con datos faltantes, muestra diversidad y asociaciones relevantes.
- Grupo de edad de la víctima y Mes del hecho: Representan dimensiones demográficas y temporales con buena variabilidad.

Su combinación facilita una selección de características objetiva, interpretable y basada en evidencia, lo cual es crucial para el éxito de modelos predictivos y para generar análisis descriptivos robustos.

VI. Conclusión

El análisis exploratorio y el preprocesamiento descritos son pasos fundamentales para transformar un conjunto de datos crudo, que a menudo presenta imperfecciones y desafíos como desbalances o datos faltantes, en una base sólida para análisis predictivos y descriptivos más avanzados. Se identificaron y abordaron problemas clave, como la presencia de datos faltantes significativos y la necesidad de consolidar o agrupar categorías para un mejor manejo. Las técnicas de evaluación de relaciones entre variables, como el V de Cramer y Chi-cuadrado, junto con la entropía, fueron cruciales durante la fase exploratoria para entender la estructura de los datos e informar las decisiones sobre qué variables son potencialmente más relevantes para el análisis y la modelización.

References

[1] "Presuntos Homicidios. Colombia, 2015 a 2023. Cifras definitivas — Datos Abiertos Colombia," Dec. 05, 2024. https://www.datos.gov.co/Justicia-y-Derecho/Presuntos-Homicidios-Colombia-2015-a-2023-Cifras-d/vtub-3de2.

[2] J. McHugh, "The Chi-square test of independence," *Biochemia Medica*, vol. 23, no. 2, pp. 143–149, Jun. 2013.

[3] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, 5th ed. Boca Raton, FL: CRC Press, 2011.

[4] P. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. Pearson, 2018.

[5] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.
[6] K. Backhaus, B. Erichson, W. Plinke, and R. Weiber, *Multivariate Analysemethoden: Eine anwendungsorientierte Einführung*, 14th ed. Springer Gabler, 2016.