A Sample Article Using IEEEtran.cls for IEEE Journals and Transactions

IEEE Publication Technology, *Staff, IEEE,*

***Abstract*—El análisis de datos de eventos violentos como los homicidios es crucial para comprender sus causas, contextos y diseñar estrategias de prevención efectivas. Sin embargo, antes de aplicar modelos predictivos o realizar inferencias profundas, es indispensable llevar a cabo un análisis exploratorio exhaustivo y un preprocesamiento cuidadoso de los datos disponibles. Este artículo detalla los primeros pasos de este proceso utilizando un conjunto de datos de homicidios.**

***Index Terms*—Análisis exploratorio de datos, preprocesamiento de datos, homicidios, patrones de violencia.**

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.

II. Descripción del Conjunto de Datos

III. Análisis Exploratorio del Conjunto de Datos

El objetivo de esta sección es realizar un Análisis Exploratorio de Datos (AED) sobre el conjunto de presuntos homicidios en Colombia (2015–2023), con énfasis en las variables categóricas (ya que estas conforman la mayoría de los datos). El notebook adjunto implementa una serie de pasos estándar y avanzados para:

- Conocer la estructura y calidad del dataset.
- Identificar valores faltantes y la manera en la que estos están representados.
- Caracterizar variables numéricas y, especialmente, categóricas.
- Explorar relaciones bivariantes y multivariantes en las variables consideradas como más relevantes.
- Detectar categorías raras que puedan afectar futuros modelos.

Finalmente, a partir de todos estos hallazgos, se propone la variable **Circunstancia del Hecho** como la más adecuada para un futuro modelo de clasificación.

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

A. Estrategias Utilizadas

1) Instalación y carga de librerías:

- Se instaló la librería `tabulate` para presentar tablas de resumen en formato legible dentro de Colab.
- Se importaron `pandas`, `NumPy`, `matplotlib` y `seaborn`.
- Se habilitó `%matplotlib inline` para ver los gráficos directamente en el notebook.

2) Importación y vista inicial del dataset:

- Lectura del CSV desde Google Drive, mediante `pd.read_csv()`.
- Se utilizó `df.head()` para inspeccionar las primeras filas y `df.shape` para conocer su tamaño (111263 filas y 35 columnas).
- Con `df.info()` y `df.dtypes` se validaron los tipos de datos de cada columna (la mayoría como `object`, algunas numéricas como año, ID y Códigos DANE).

3) Detección y tratamiento de valores faltantes:

- Ejecución de `df.isnull().sum().sort_values()` para obtener el conteo de verdaderos “NaN”.
- Según nuestras observaciones, pudimos identificar que muchos faltantes no se codifican como NaN sino como cadenas (“No aplica”, espacios en blanco, etiquetas locales). Teniendo en cuenta esto se realizó una inspección manual de los valores únicos para cada variable, buscando patrones de ausencias no estándar.

4) Análisis univariante:

a) Variables numéricas:

- Uso de `df.describe()` para resumir las variables numéricas, pero de esta solo se encontraron resultados relevantes en el año del hecho.
- Boxplots de edad judicial categorizada por sexo, revelando que la mediana y el rango intercuartílico de la edad difieren ligeramente entre hombres y mujeres.

b) Variables categóricas:

- Identificación automática de variables de tipo objeto con `df.select_dtypes(include='object')`.
- Frecuencias absolutas y relativas impresas con `value_counts(dropna=False)` y presentadas con `tabulate` para mayor legibilidad.
- Gráficos de barras (`sns.countplot` sobre `matplotlib`) para las variables categóricas.

5) Análisis bivariante y pruebas de asociación:

- Tablas de contingencia (`pd.crosstab`) entre pares de categorías (Sexo vs. Circunstancia).

- Prueba de Chi-cuadrado (`chi2_contingency`) en cada tabla para cuantificar si la relación observada es estadísticamente significativa.

6) Detección de categorías raras:

- Para cada variable categórica, se calcularon las proporciones de cada etiqueta y se listaron aquellas con frecuencia $< 1\%$ del total.
- Estas categorías de baja representación se consideraron para agruparse en un nivel "Otros" antes de cualquier modelado predictivo, con el fin de evitar ruido y problemas de *sparsity*.

7) Análisis multivariante y correlaciones:

- Construcción de la matriz de correlación sobre las variables numéricas y visualización con `heatmap` de `seaborn` para identificar posibles colinealidades.
- Al observar que las variables numéricas no presentaban correlaciones fuertes, se confirmó que el análisis se debía centrar en las categorías.

B. Principales Resultados

1) Calidad de los datos:

- Aunque `df.info()` mostraba pocos NaN reales, se detectaron múltiples cadenas ("No aplica", "Desconocido", "no informado") que requieren un preprocesamiento especial.
- Sólo la variable numérica (Año del hecho) resultó robusta; el resto eran categorías puras.

2) Distribución de casos:

- Se evidencia un aumento en los homicidios en el transcurso de los años. Se ve una tendencia incremental leve, con una excepción en el año 2020, donde se rompió esta tendencia incremental y se mostró valores muy similares a los del año 2017.
- Arma de fuego y arma corto punzante fueron mecanismos y circunstancias dominantes, con más del 60 % de los casos cada uno.

3) Categorías raras:

- Se identificaron en la mayoría de variables etiquetas con frecuencias relativas por debajo del 1 % (por ejemplo, algunas clases de "Objeto de colisión" en homicidios por transporte).
- En estas se debe considerar agrupaciones bajo categorías como "Otros" para mejorar la estabilidad de futuros modelos.

C. Hallazgos

El AED confirmó que la variable **Circunstancia del Hecho** es la más apropiada como objetivo para un modelo de clasificación, porque:

- Tiene múltiples clases significativas (homicidio simple, agravado, riña, feminicidio, etc.).
- Muestra asociaciones estadísticamente robustas con otras variables clave (sexo, zona, escenario).
- Su predicción permitiría orientar políticas de prevención diferenciales según el contexto del homicidio.

Los pasos de detección de valores faltantes no estándar y de agrupación de categorías raras fueron críticos para asegurar la calidad de los datos.

Se debe analizar qué tantas implicaciones e imposibilidades traerá el hecho que muchos atributos contienen valores faltantes hasta del 64%.

IV. Análisis Exploratorio: Un Vistazo a los Datos Crudos

El análisis preliminar del conjunto de datos reveló varias características importantes y desafíos:

- **Desbalance por sexo:** Los datos mostraron un claro desequilibrio en la distribución por sexo de la víctima, con una frecuencia de homicidios considerablemente mayor para hombres.
- **Tendencia temporal:** Se observó que la frecuencia general de homicidios disminuye a medida que avanzan los años cubiertos por el conjunto de datos.
- **Temporalidad específica:** Los meses con mayor incidencia de homicidios fueron diciembre y mayo. Los días de la semana con mayor incidencia fueron domingo y sábado. Se sugiere investigar la relación entre esta temporalidad específica y la incidencia de casos.
- **Ubicación principal:** Las cabeceras municipales se identificaron como los lugares más comunes donde ocurren los hechos. Se propone investigar las características específicas de estas zonas para entender por qué concentran la mayor cantidad de casos.
- **Datos faltantes significativos:** Un desafío importante identificado fue la presencia de un alto porcentaje de datos faltantes en varias variables clave:
  - Nivel de escolaridad: 21.01
  - Actividad durante el hecho: 38.89
  - Circunstancia del hecho: 64.16

  Otras variables, como la pertenencia grupal, también mostraron una cantidad considerable de valores faltantes (16.81

V. Preprocesamiento de Datos

Basándonos en las observaciones preliminares y la naturaleza específica de cada variable, se tomaron varias decisiones clave para preparar el conjunto de datos. El objetivo fue limpiar y organizar los datos para que fueran adecuados para análisis posteriores y modelado, facilitando la identificación de patrones relevantes.

- **Eliminación de variables no relevantes o redundantes:** Se decidió eliminar variables que tenían un único valor ("No aplica" en todos los registros), así como códigos DANE ('Código Dane Municipio', 'Código Dane Departamento'), y variables redundantes como 'Edad judicial' y 'Ciclo Vital' en favor de 'Edad natural'. La razón es que estas variables no proporcionaban información útil o única para identificar patrones o variabilidad en los homicidios. Variables como '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' tenían un solo valor ('No aplica') en todos los casos y fueron identificadas para posible eliminación.

- **Manejo de datos faltantes:** La estrategia varió según la variable y el porcentaje de datos faltantes.
  - Para la Edad y la Fecha, los registros con valores faltantes fueron eliminados. Razón: Representaban una proporción muy pequeña del total de datos (0.1
  - Para la Pertenencia Grupal, que tenía un 16.81
  - Para variables con porcentajes muy altos de datos faltantes como Escolaridad (21.01%), Actividad Durante el Hecho (38.89%) y Circunstancia del Hecho (64.16%), se identificó el desafío. Razón: Dada la magnitud de los faltantes, la eliminación de filas no era viable sin perder la mayoría del conjunto de datos. En el caso de "Circunstancia del Hecho", la categoría "Sin información" se mantuvo como una categoría válida para el análisis, dada su alta frecuencia.
- **Corrección de errores y agrupación:**
  - Se identificaron y eliminaron categorías duplicadas en la variable "Escenario del Hecho", causadas por errores de digitación. Esto asegura que cada categoría represente un escenario único y correcto.
  - Se evaluó la posibilidad de agrupar categorías amplias en subgrupos más manejables para simplificar el análisis y mejorar el rendimiento de modelos predictivos. Ejemplos incluyeron agrupar el País de Nacimiento en "Colombia" y "Otros". También se evaluó agrupar categorías en el Estado Civil (que tenía 8 categorías únicas) antes de aplicar técnicas como one-hot encoding.
  - En cuanto a las edades, se descartó el uso de 'Edad judicial' en favor de 'Edad natural' debido a que esta última ofrece una mayor granularidad. Además, se planteó la categorización de las víctimas como mayores o menores de edad, y la aplicación de one-hot encoding a esta variable categórica. Se observaron también agrupaciones de edad en ciclos vitales como "Juventud" (18 a 28 años) y "Adultez" (29 a 59 años).
  - Específicamente para la variable "Circunstancia del Hecho", se realizó una reagrupación de las múltiples categorías existentes (50 valores únicos inicialmente) en subgrupos más amplios y significativos. Razón: Esto fue crucial para hacer la variable más manejable para el modelado posterior y para capturar patrones a un nivel más general. Este mapeo permitió consolidar las categorías, por ejemplo, agrupando 'Riña', 'Violencia de pareja' y 'Celos' bajo "Violencia interpersonal", o 'Sicariato', 'Ajuste de cuentas' y 'Hurto' bajo "Crimen organizado / sicariato". Otras categorías se agruparon en temas como "Conflicto armado / terrorismo", "Violencia institucional / Estado", "Violencia estructural o sociopolítica", y "Violencia contra población vulnerable". Las categorías con muy baja frecuencia pudieron ser agrupadas en una categoría "Otros".

  Este proceso de preprocesamiento no solo limpió los datos, sino que también los transformó (agrupación de categorías, imputación, creación de nuevas variables como mayor/menor de edad) para ser más adecuados para análisis posteriores y para mejorar la efectividad de la selección de características.

VI. Selección de Características

Durante esta fase exploratoria, se examinaron las distribuciones de variables individuales, a menudo utilizando conteos de valores (`value_counts`) para entender la frecuencia y proporción de cada categoría. Por ejemplo, se observó la alta concentración en categorías como "Hombre" para Sexo de la víctima o "Colombia" para País de Nacimiento de la Víctima.

Para comenzar a explorar las relaciones entre variables categóricas, se emplearon tabulaciones cruzadas. Por ejemplo, se analizaron cruces entre la Zona del Hecho y el Sexo de la víctima, o la Circunstancia del Hecho y el Sexo de la víctima.

Para cuantificar la fuerza de la asociación entre estas variables categóricas y evaluar su potencial relevancia (un paso inicial hacia la selección de características), se emplearon medidas estadísticas como la Chi-cuadrado y el coeficiente V de Cramer.

A. Chi-cuadrado

Esta estadística se utiliza para determinar si existe una asociación significativa entre dos variables categóricas. Evalúa si las frecuencias observadas en una tabulación cruzada difieren significativamente de las frecuencias esperadas bajo la hipótesis de independencia. Un valor de Chi-cuadrado grande y un p-valor pequeño sugieren que las variables están asociadas. Los resultados del análisis mostraron cómo ciertas variables categóricas tenían puntuaciones altas de Chi-cuadrado en relación con otras variables categóricas (consideradas como "pseudo-objetivos" en el contexto del análisis), lo que indica una fuerte dependencia y las identifica como características potencialmente importantes.

B. Coeficiente V de Cramer

Es una medida de asociación entre dos variables nominales (categóricas). Se basa en la estadística Chi-cuadrado, pero su valor se ajusta para estar entre 0 y 1. Un valor de 0 indica que no hay asociación entre las variables, mientras que un valor de 1 indica una asociación perfecta. En el análisis exploratorio y como parte de la evaluación de características, el V de Cramer es valioso para identificar qué variables categóricas muestran una relación más fuerte con otras variables o una posible variable de interés. Esto puede informar la selección de características para futuros modelos predictivos al centrarse en variables que tienen relaciones más fuertes con lo que se desea analizar o predecir.

También se utilizó la entropía como medida de la variabilidad o "riqueza de información" en las variables categóricas. Variables con mayor entropía (dentro de las no dominadas por valores faltantes o "No aplica") sugieren que sus categorías

están más uniformemente distribuidas, lo que puede hacerlas más informativas. Por ejemplo, variables como Escenario del Hecho, Escolaridad, Actividad Durante el Hecho y Rango de Hora mostraron entropías más altas. La variable ”Circunstancia del Hecho”, a pesar de tener muchos valores faltantes, mostró una entropía de 1.107589, sugiriendo suficiente diversidad en sus categorías no faltantes para ser considerada.

Estas medidas permitieron identificar qué variables categóricas mostraban las asociaciones más fuertes entre sí o con posibles variables de interés. Particularmente, el Chi-cuadrado fue usado para medir la dependencia entre las características y potenciales variables objetivo y la entropía ayudó a entender la distribución de la información dentro de cada variable, destacando aquellas con mayor variabilidad y potencial para diferenciar entre casos.

La información obtenida de estos análisis de relaciones y distribuciones es vital para guiar la selección de un subconjunto óptimo de características para la construcción de modelos predictivos o para centrar análisis descriptivos más profundos. Variables con relaciones fuertes con fenómenos de interés o con distribuciones informativas son candidatas más fuertes para ser incluidas en futuros análisis.

VII. 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 el desbalance de género, 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. Este enfoque sistemático allana el camino para investigaciones subsiguientes y el desarrollo de modelos que puedan informar eficazmente la toma de decisiones orientadas a la prevención de homicidios.

References

[1] *Mathematics Into Type*. American Mathematical Society. [Online]. Available: https://www.ams.org/arc/styleguide/mit-2.pdf

[2] T. W. Chaundy, P. R. Barrett and C. Batey, *The Printing of Mathematics*. London, U.K., Oxford Univ. Press, 1954.