TRABAJO DE GRADO
Opción Seminario

Análisis de datos sobre la Violencia Doméstica en Colombia

Corporación Universitaria Remington

Facultad de Ingenierías

Ingeniería de Sistemas

Cristhian Felipe Obregón Muñoz[1]

Luisa María Ciro Silva[2]

Ivonne Castaño Osorio[3]

John Edison Amórtegui Granada [4]

Seminario

2024

[1] Estudiante noveno semestre de Ingeniería de Sistemas Uniremington sede Pereira. E-mail: cristhian.obregon.0423@miremington.edu.co
[2] Estudiante noveno semestre de Ingeniería de Sistemas Uniremington sede Pereira. E-mail: luisa.ciro.8914@miremington.edu.co
[3] Tutor temático seminario Big Data y Ciencia de Datos. E-mail: ivonne.castano@uniremington.edu.co
[4] Tutor metodológico seminario Big Data y Ciencia de Datos. E-mail: john.amortegui@uniremington.edu.co

Dedicatoria

Este trabajo está dedicado a aquellas personas que sufren violencia doméstica; debe ser difícil enfrentar este tipo de situaciones y aún más difícil denunciarlas. Por medio del análisis de datos, se pretende identificar qué clasificación de departamento tiene una alta tasa de violencia doméstica, con el objetivo de poder incentivar iniciativas que ayuden a prevenir estas situaciones en la familia.

Tabla de contenidos

Resumen

En el documento se identifica la aplicabilidad de la metodología CRISP-DM de IBM, la cual "proporciona una visión general del ciclo de vida de la minería de datos" (IBM SPSS Modeler CRISP-DM Guide, 2021, p. 9). Se analizarán datos extraídos de un DataSet[5] sobre la violencia doméstica en Colombia. Se optó por usar esta metodología ya que consta de diez pasos estructurados y de fácil comprensión. El objetivo es aplicar y comprender el proceso de levantamiento de datos para cumplir con todas las fases que ofrece la metodología, con la finalidad de obtener datos de valor que proporcionen un amplio conocimiento sobre el tema principal. Además, la metodología busca ofrecer un enfoque integral y real sobre los diversos factores y patrones que afectan durante el análisis de datos. El DataSet está compuesto por grupo etario, sexo, año en que ocurrieron los hechos y arma utilizada. Con estos datos, se esperan identificar patrones de comportamiento por regiones y así determinar la alta incidencia de violencia familiar por departamento.

Palabras clave

Violencia Domestica.

Analítica de Datos.

Metodología CRIPS- IBM.

Ciencia de Datos.

Big Data.

---

[5] "Colección estructurada de información gestionada de forma organizada y sistemática, donde cada elemento tiene relación con los demás" (Lab, Redacción The Information, 2023, párrafo segundo).

Pregunta orientadora de la búsqueda

A continuación, se busca analizar la problemática del maltrato doméstico que aún persiste en Colombia, este análisis de hará por medio de la metodología de analítica de datos CRIPS-IBM, con este método se identificarán patrones de violencia intrafamiliar[6] en todos los departamentos. Además, se estará implementado, conociendo una parte integral del concepto que se tiene de Big Data[7].

Se buscó una base de datos que permitiría identificar en que departamentos tiene una mayor tasa de violencia categorizado por grupo etario y sexo esto con el fin de analizar el patrón de violencia, si se presenta con mayor frecuencia en departamentos grandes, medianos y pequeños, y por grupo etario identificar donde corre más riesgo a sufrir dicha violencia. A continuación, se visualiza la pregunta orientadora para resolver las dudas planteadas en el texto descrito: ¿Desde la metodología CRISP-DM vivir en un departamento grande aumenta la posibilidad de sufrir violencia intrafamiliar en Colombia?

Metodología de búsqueda de la información

La metodología de ciencia de datos que se implementará es una variación de la metodología CRISP-DM de IBM, CROSS-INDUSTRY STANDARD PROCESS FOR DATA MINING (proceso estándar entre industria para minería de datos). Esta variación

---

6 “Se divide en violencia intrafamiliar, que se produce dentro de la familia, enfocada hacia parejas, cónyuges, hijos y, por norma general, suele acontecer dentro del hogar, afectando a los niños, a los ancianos, a la pareja.” (Sevillano & Rivera, 2023, p. 11)

7 Colección de datos grandes, complejos, muy difíciles de procesar a través de herramientas de gestión y procesamiento de datos tradicionales. Son datos cuyo volumen, diversidad y complejidad requieren nueva arquitectura, técnicas, algoritmos y análisis para gestionar y extraer valor y conocimiento oculto en ellos. (Varona & Madera, 2018, p. 2)

incluye 10 preguntas básicas que permite mejorar la conceptualización general del problema:

1. Se debe identificar la compresión del negocio esto con el fin de identificar el problema a analizar.
2. Definir enfoque analítico el cual permite definir el camino (descriptivo, predictivo, estadístico, machine learning y asociación de clustering) más apropiado.
3. Se definen los requisitos de datos identificando la información necesaria para el análisis de la problemática según el enfoque seleccionado. Allí se validará la estructura de datos como por ejemplo si los datos son cualitativos, cuantitativos, tipo de archivo, tipo de dato Etc.
4. La recopilación de datos permite validar que la información recolectada sea la necesaria para la problemática y validar que la información sea acorde a la necesidad revisando la calidad de los datos como por ejemplo que no existan columnas o registros vacíos, que tenga el formato correcto, respetar el formato de los campos según el tipo de datos.
5. La comprensión de datos el cual es todas las actividades que se llevarán a cabo para la construcción del conjunto de datos y evaluar si los datos son representativos para dar solución a la problemática inicial.
6. Preparación de datos, esta parte es la más importante por que conlleva el mayor tiempo del trabajo, debido a que allí se realiza un tratamiento a los datos para que así los resultados del proyecto se respalden y mantenga una alta calidad.

7. Modelado de datos se define algoritmos y estrategias que permita garantizar que la información suministrada sea realmente necesaria. Allí usan modelos tales como el de Regresión que “Determina si existe, o no, relación causal entre una variable dependiente (Y) y un conjunto de otras variables explicativas (X)” (Roldán, 2020, párrafo primero), y Covarianza “Permite analizar la relación lineal entre dos variables.” (Cosio, 2021, párrafo tercero).
8. Evaluación del modelo, evalúa la calidad del modelo seleccionado durante su ejecución e identifica que tan buena es la solución al problema planteado. Esta se compone de dos fases las cuales son: fase de medida y aplicación de una prueba de significancia de estadística.
9. Implementación después de realizar la validación y ejecución del modelo se visualizará el resultado arrojado por el análisis y socializando el producto final con las partes interesadas y validar si se cumple o no con la solución del planteamiento inicial.
10. Retroalimentación allí se refina el modelo según la retroalimentación suministrada por los usuarios finales donde se evalúa el impacto y el rendimiento.

A continuación, se describirá el proceso realizado mediante la metodología CRISP-DM de IBM aplicados para poder responder la pregunta planteada:

Sustentación teórica de la pregunta

Comprensión del Negocio

La violencia doméstica es un problema que debe ser tratado con mucho cuidado. En el presente documento se quiere dar relevancia a dicha problemática, que por años ha sido un pilar de preocupación, ya que el maltrato familiar en Colombia es significativo. Existen cifras que muestran que Este tipo de violencia es una de las principales causas de lesiones y muertes, especialmente entre mujeres y niños. Las estadísticas recientes reflejan su gravedad. Según la Policía Nacional, a mediados de mayo de 2023 se reportaron más de 38,000 casos de violencia intrafamiliar, lo que equivale a aproximadamente 323 casos diarios (Radio Nacional de Colombia) (Escobar, 2023)

Enfoque Analítico.

Es un enfoque descriptivo ya que se busca describir y cuantificar la tasa de maltrato por departamento. Allí se busca tener un enfoque centrado en proporcionar una comparativa de datos que demuestren detallada y claramente la problemática a través de la recolección de datos específicos.

Requisitos de Datos

Se requiere poder demostrar estadísticas de reportes de violencia intrafamiliar por departamento por lo cual se necesita tener datos como departamento, cantidad de registros presentados, fecha de suceso y otros datos que permitan hacer un análisis para informar que departamento tiene la mayor tasa de maltratos y así saber si la categoría por departamento influye o no sobre esta problemática.

Recopilación de datos

Se utilizará la página “Kaggle” en dicha plataforma se obtiene información desde el año 2010 hasta el 2023 en COLOMBIA para realizar la analítica de datos de distintos

temas de interés. En dicha herramienta se obtuvo un dataset con la siguiente estructura que permitirá identificar la solución de la problemática planteada:

- **Departamento:** este campo esta de tipo texto y está correctamente diligenciado.
- **Municipio:** este campo esta de tipo testo y está correctamente diligenciado.
- **Código DANE:** este campo es numérico, tiene registros que dice NO REPORTA, ya que no es relevante para el análisis no se homologara.
- **Armas:** este campo es de tipo lista, hay campos que no tienen valores, se definió como valor por defecto "NO DEFINIDO". Esto con el fin de analizar y validar cuantos casos existen sin la tipología del arma.
- **Fecha Hecho**: este campo es de tipo fecha.
- **Género**: este campo es de tipo lista, hay campos que no tienen valores, se definió como valor por defecto "NO DEFINIDO". Esto con el fin de analizar y validar cuantos casos existen sin la tipología del arma.
- **Grupo Etario:** este campo es de tipo lista, hay campos que no tienen valores, se definió como valor por defecto "NO DEFINIDO". Esto con el fin de analizar y validar cuantos casos existen sin la tipología del arma.
- **Cantidad**: este campo es numérico e indica la cantidad de afectados.

La cantidad de datos analizar es de 575.720.

Comprensión de Datos

En este paso se identificó las columnas que se requieren para el análisis, el resto de los datos fue borrado (grupo etario, sexo y municipio), además a ello, en el DATASET la

columna departamento tenía valores no reportados, se tomó la decisión de eliminarlos ya que no afectaba en los indicadores a analizar.

Preparación de Datos:

Se agrega dos columnas nuevas, para realizar el análisis de datos se le asignó al campo id departamento, población con dicho campo se definió la clasificación (Grande, Mediano o Pequeño) de los departamentos en una nueva columna, estos datos de población se sacaron de la página "GOV.CO" (DANE, 2023) y se tiene una sección donde se obtiene el informe de categorización departamentos por población se validó que los registros estuvieran diligenciados de manera correcta y el campo cantidad lo convertimos a numérico para que así garanticemos su correcta diligencia. Se definido no agrupar los datos ya que si se agrupaba se perdería la congruencia de los datos. Además, a ello la variable dependiente es la columna cantidad (cantidad de personas afectadas).

Para poder evaluar el comportamiento de los datos por cada departamento se creó una gráfica de mapa de árbol (utiliza rectángulos de diferentes tamaños para transmitir valores numéricos a cada rama (Wood & Dykes, 2008)) por departamento, identificando cuales son los departamentos con mayores casos de violencia doméstica, allí Cundinamarca, Antioquia y valle son los departamentos con mayor número de casos reportados y a su vez estos se encuentran en la categoría de departamento grandes:

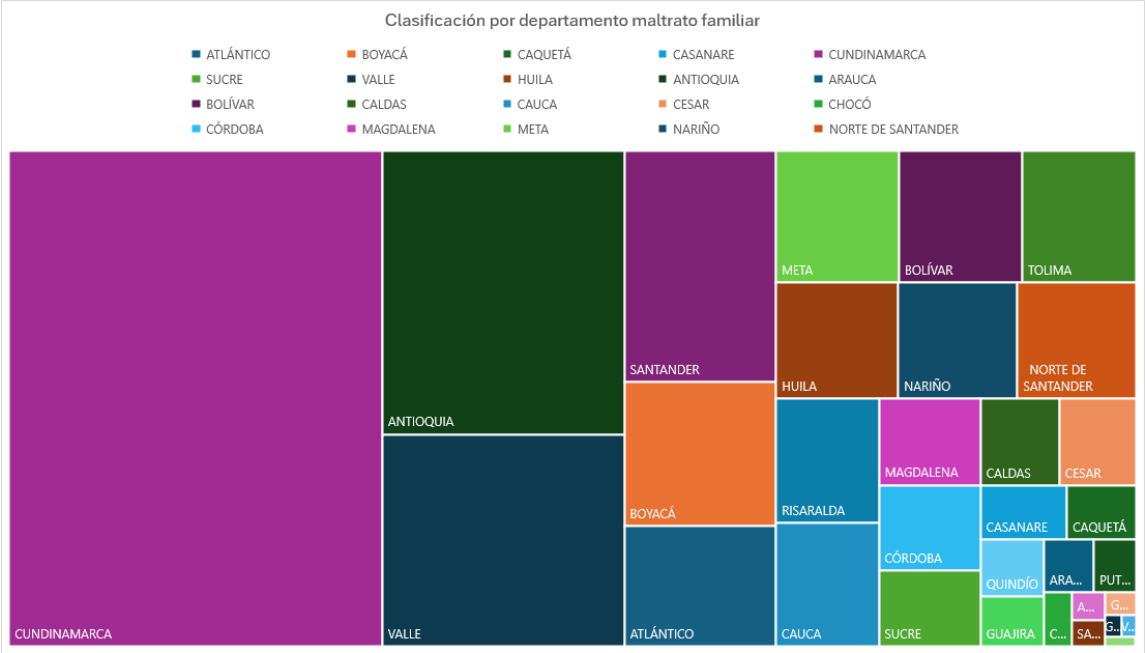

Figura 1. Clasificación departamentos

Modelado

Se realiza el modelado por medio del proceso de regresión, partiendo que la variable dependiente es la clasificación departamento (pequeño, mediano y grande) y la independiente es la cantidad de registros presentados en el set de datos. Allí se puede identificar que existe una relación significativa entre las variables escogidas, pero cabe aclarar que la relación entre dichas variables es débil. Determinando así que la cantidad de registros por violencia doméstica no es un buen predictor para la clasificación del departamento, allí lo que se puede hacer es incluir otras variables que si sean relevantes para el modelo evaluado.

Para el modelado se realzará diferentes métodos los cuales son:

- Modelo de regresión: en este se tomará dos datos principales para la solución de la problemática el cual es la clasificación de departamento y la cantidad de casos que

se obtuvo, en ese caso nuestra variable Y es clasificación y la variable X es cantidad de casos registrado como violencia doméstica.

Tabla 1. Estadística de regresión

Regression statics	
Multipler R	0.131615885
R Squere	0.017322741
Adjusted Rsquare	0.017321034
Standard Error	0.521120223
Obcevation	575715

Tabla 2. Estadística de regresión detalle

	df	SS	MS	F	Significance F
Regression	1	2756.053244	2756.053244	10148.73118	0
Residual	575713	156344.2418	0.271566287		
Total	575714	159100.295			

- Modelo de correlación[8]: En este modelo al igual que el anterior se tomó las variables la clasificación de departamento y la cantidad de casos que se obtuvo el cual la nuestra variable Y es clasificación y la variable X es cantidad de casos registrados.

Tabla 3. Estadística de correlación

	Clasificación	Cantidad
Clasificación	1	
Cantidad	-0.13162	1

[8] "Estudia si existe asociación entre dos variables cuantitativas y establece cuál es la dirección y la magnitud de esa asociación" (Arias, Sangrador, & Páez, 2021, párrafo cuarto).

En el modelo de correlación se puede ver que es una correlación negativa débil, Según lo obtenido por este modelo se entiende que los datos no son suficientemente fuertes para garantizar un buen análisis a la problemática.

- Modelo de covarianza: En este modelo al igual que el anterior se tomó las variables la clasificación de departamento y la cantidad de casos que se obtuvo el cual la nuestra variable Y es clasificación y la variable X es cantidad de casos registrados. A continuación, se visualiza la ejecución de su formulación:

Tabla 4. Estadística de correlación

	Columna 1	Columna 2
X	0.276353	
Y	-0.27068	15.3046

En el modelo de covarianza permite ver que la relación entre las variables relacionada es muy baja y no es muy útil para darle respuesta a la problemática planteada.

Evaluación del modelo

Las variables dependientes contra la independiente no tienen una buena relación, según el coeficiente de correlación múltiple puesto que este da un valor de 0.1316 siendo este un valor muy bajo, adicional el coeficiente de la variable dependiente en comparación con las variables independientes no tiene mucha variabilidad según el coeficiente de determinación, en cuanto el r ajustado indica que no existen variables que no ayuden al análisis de las estadísticas.

Tomando de guía el valor de F que es el nivel de confianza del modelo se determina que dicho modelo es bueno ya que el resultado arrojado es alto “10.148”,

además el estadístico T también dio un buen resultado en la estadística de intercepción con un valor de 2.224.

Implementación

Este proceso no aplica ya que no se realizó el uso práctico, solo se realizó el análisis de datos haciéndolo por medio de un análisis descriptivo.

Retroalimentación

Este método no aplica ya que no se realizó ninguna retroalimentación con expertos en el tema dado la naturaleza del trabajo a presentar.

Conclusiones.

La implementación de la metodología de CRIPS-DM ha servido como guía para hacer un buen análisis de datos, esta ha proporcionado una visión mucho más amplia a la hora de querer analizar datos, además de que permite gestionar y ejecutar un buen plan de análisis para cualquier entorno o problemática que se desee evaluar, cubriendo distintos escenarios como lo es la recopilación de datos, evaluación de los datos compresión del negocio.

Esta metodología permite abordar un enfoque estructurado además de permite explorar un poco sobre la analítica de datos y el uso que se le da en cuanto a Big data dando así un enfoque sistemático resaltando la importancia y ofreciendo una base sólida que facilita la toma de decisiones. Se recomienda que al momento de hacer uso de un data set se evalué si la información que contiene es necesaria y cubre la problemática a investigar, revisar los datos y la cantidad de registros, adicional la calidad de dichos datos. Hacer un buen uso y siguiendo el paso a paso de la metodología esto con el fin de realizar un análisis adecuado y con buenos resultados, abarcando la solución al problema o tema a tratar.

Lista de referencias

Arias, M., Sangrador, O., & Páez, O. (9 de 6 de 2021). *evidenciasenpediatria*. Obtenido de evidenciasenpediatria: https://evidenciasenpediatria.es/articulo/7827/correlacion-modelos-de-regresion

Corporation, International Business Machines. (2021). *IBM SPSS Modeler CRISP-DM Guide*.

Cosio, N. A. (10 de 12 de 2021). *medium*. Obtenido de medium: https://medium.com/@nicolasarrioja/covarianza-y-correlaci%C3%B3n-7f16e59445b4

DANE. (30 de 10 de 2023). *Contaduria General de la nación* . Obtenido de GOV.CO: https://www.contaduria.gov.co/categorizacion-de-departamentos-distritos-y-municipios

Escobar, J. (12 de 05 de 2023). *Radio Nacional*. Obtenido de https://www.radionacional.co/actualidad/violencia-intrafamiliar-en-colombia-2023-cifras-de-la-policia

ESTIVEN0507. (13 de 3 de 2023). *kaggle*. Obtenido de kaggle: https://www.kaggle.com/datasets/estiven0507/domestic-violence-in-colombia

Lab, Redacción The Information. (27 de 11 de 2023). *theinformationlab*. Obtenido de theinformationlab: https://www.theinformationlab.es/blog/que-es-un-dataset/

Roldán, P. N. (1 de 9 de 2020). *economipedia*. Obtenido de economipedia: https://economipedia.com/definiciones/modelo-de-regresion.html

Sevillano Cabel, A. M., Obando-Peralta, C., & Rivera Gamarra, A. C. (1 de 2023). Violencia intrafamiliar: Una problemática humana actual. *CORPORACION UNIVERSITARIA REMINGTON*, pág. 11.

Varona, r. L., YoanMartínez-López, & Madera, J. (12 de 08 de 2018). Una revisión de la tecnología “BigData” para grandes volúmenes de datos. *Información, Instituto de Información Científica Técnicay Sociedad Cubana de Ciencias de la*, pág. 2.

Wood, J., & Dykes, J. (6 de 12 de 2008). *arcgis*. Obtenido de arcgis: https://doc.arcgis.com/es/insights/latest/create/treemap.htm#:~:text=Un%20mapa%20de%20%C3%A1rbol%20es,alto%20ser%C3%A1%20el%20valor%20num%C3%A9rico.