

Tabla de contenido

RESUMEN.....	9
ABSTRACT.....	10
PRELIMINARES.....	11
INTRODUCCIÓN.....	13
PLANTEAMIENTO DEL PROBLEMA	17
1. DIAGNÓSTICO DE LA SITUACIÓN ACTUAL.....	17
2. PREDICCIÓN O PRONÓSTICO DE LA SITUACIÓN DESCRITA.....	19
3. ARBOL DEL PROBLEMA.....	20
3.1 <i>Problema central</i>	20
3.2 <i>Causas</i>	21
3.3 <i>Efectos</i>	23
4. ANTECEDENTES.....	24
5. FORMULACIÓN O PREGUNTA DE INVESTIGACIÓN.....	27
OBJETIVOS.....	28
1. OBJETIVO GENERAL	28
2. OBJETIVOS ESPECÍFICOS.....	28
JUSTIFICACIÓN DEL PROYECTO.....	29
ALCANCES Y LIMITACIONES DEL PROYECTO.....	31
1. ALCANCES.....	31
2. LIMITACIONES	36
MARCO REFERENCIAL	38
1. ESTADO DEL ARTE	38
2. MARCO CONCEPTUAL	43
METODOLOGÍA	47
1. TIPO DE ESTUDIO	47
2. MÉTODO O DISEÑO DE INVESTIGACIÓN	47
2.1 <i>Etapa 1: Creación del repositorio virtual de datos</i>	49
2.2 <i>Etapa 2: Aplicación de técnicas de Machine Learning</i>	51
3. PARTICIPANTES O UNIDAD DE ANÁLISIS	55
4. PLAN DE ANÁLISIS DE LA INFORMACIÓN	55
5. CRONOGRAMA Y PRESUPUESTO.....	58
PROCEDIMIENTO O DESCRIPCIÓN DEL TRABAJO DE CAMPO	61
1. DIAGRAMA DE FLUJO GENERAL DE LAS ACTIVIDADES DE CAMPO DEL PROYECTO	61
2. DIAGRAMA DE FLUJO DETALLADO DE LAS ACTIVIDADES DE CAMPO – ETAPA 1.....	62
2.1 <i>Diseño de la base de datos Postgres con PostGIS</i>	64
2.2 <i>Implementación del proceso ETL para poblamiento de la base de datos</i>	69
3. DIAGRAMA DE FLUJO DETALLADO DE LAS ACTIVIDADES DE CAMPO – ETAPA 2.....	72
3.1 <i>Extracción de datos del repositorio virtual</i>	74

3.2 Subir los archivos a Google Drive y cargarlos al entorno de Google Colab	76
3.3 Aplicar transformaciones para machine learning.....	77
3.4 Aplicación de algoritmos de Machine Learning	80
3.5 Evaluación de resultados.....	85
3.6 Informe final	87
RESULTADOS Y DISCUSIONES	88
1. RESULTADOS Y DISCUSIONES DE LA CREACIÓN DEL REPOSITORIO VIRTUAL DE DATOS – ETAPA 1	88
1.1 Diseño de la base de datos.....	88
1.2 Verificación de que las tablas que conforman el core de la base de datos cumplen con el propósito para el que fueron diseñadas.....	90
1.3 Estadísticas y composición de la base de datos	95
1.4 Calidad y consistencia de los datos incorporados	112
1.5 Validación de la Integridad de los Datos.....	112
1.6 Optimización y eficiencia.....	113
1.7 Diccionario de Datos.....	115
1.8 Propuestas de mejora y evolución.....	116
2. RESULTADOS Y DISCUSIONES DE LA SEGUNDA ETAPA DEL PROYECTO	116
2.1 Resultado del enfoque escalonado para la extracción de datos de la base de datos	117
2.2 Integración de los archivos CSV en un solo DataFrame.....	119
2.3 Resultados de la integración.....	121
2.4 Transformaciones orientadas a la aplicación de las técnicas de machine learning.....	121
2.5 Aplicación de los algoritmos Machine Learning en los datasets depurados.....	189
CONCLUSIONES.....	243
REFERENCIAS BIBLIOGRÁFICAS.....	247
ANEXOS.....	254

Lista de tablas

Tabla 1 Presupuesto del proyecto	60
Tabla 2 Reporte de conteo de registros por tabla con total consolidado	98
Tabla 3 Tipos de datos en el repositorio.....	110
Tabla 4 Columnas del dataset inicial resultante de la extracción de información de la base de datos.....	127
Tabla 5 Pares de variables socioeconómicas altamente correlacionadas	139
Tabla 6 Pares de variables socioeconómicas altamente correlacionadas después de ajustes	145
Tabla 7 Variables sobre la violencia después de la depuración inicial.....	169
Tabla 8 Pares de variables sobre violencia altamente correlacionadas	177

Lista de figuras

Figura 1 Arbol de problemas	20
Figura 2 Diagrama general de la arquitectura del proyecto	48
Figura 3 Ciclo metodología CRISP-DM	56
Figura 4 Cronograma general del proyecto	59
Figura 5 Diagrama general de flujo de la ejecución del proyecto.....	61
Figura 6 Diagrama de flujo detallado de la etapa I del proyecto	63
Figura 7 Representación de la estructura de la base de datos del proyecto	67
Figura 8 Representación aproximada y reducida del diagrama Entidad-Relación	68
Figura 9 Diagrama de flujo detallado de la etapa II del proyecto	73
Figura 10 Vista parcial del diagrama ER de la base de datos	89
Figura 11 Consulta SQL para demostrar trazabilidad de la fuente de datos	90
Figura 12 Resultado de consulta SQL que demuestra trazabilidad de la fuente de la base de datos	90
Figura 13 Script Python – Conexión, consulta y generación archivo CSV	92
Figura 14 Script Python - Prepara datos para generar mapas interactivos.....	93
Figura 15 Validación de geolocalización en base de datos con Python.....	94
Figura 16 Mapas generados con la ejecución de los códigos de la figura 15 que demuestran la capacidad geoespacial de la base de datos.....	95
Figura 17 Número total de registros y espacio de almacenamiento de la base de datos	96
Figura 18 Consulta SQL de generación de reporte de número de registros por tabla con total consolidado y exportación a CSV.....	97
Figura 19 Distribución de datasets por fuente de datos	100
Figura 20 Distribución de las fuentes de datos incluidas en repositorio virtual	101
Figura 21 Periodo de Tiempo Cubierto por los Datos.....	102
Figura 22 Dataset con mayor cobertura temporal de datos históricos	103
Figura 23 Conjuntos de datos con rango temporal 1958 - 2022	103
Figura 24 Conjuntos de datos con rango temporal 1996 - 2024	104
Figura 25 Conjuntos de datos con rango temporal 2001 - 2022	104
Figura 26 Conjuntos de datos con rango temporal 2003 - 2024	105
Figura 27 Conjuntos de datos con rango temporal 2007 - 2024	105
Figura 28 Muestra de conjuntos de datos con rango temporal 2010 - 2024	106
Figura 29 Conjuntos de datos con rango temporal 2011 – 2024 y 205 -2024.....	106
Figura 30 Conjuntos de datos con rango temporal 2016-2019	107
Figura 31 Conjunto de datos con rango temporal 2018-2023	108
Figura 32 Conjuntos de datos gestionados con microdatos del Censo de 2018	109
Figura 33 Conjuntos de datos que figuran sin fecha de inicio	109
Figura 34 Consulta SQL para obtener la distribución de tipos datos en el repositorio	110
Figura 35 Distribución tipos de datos.....	110
Figura 36 No. de departamentos y municipios con información en la bd	111
Figura 37 Indices sobre llaves primarias construidos automáticamente por Postgress	114
Figura 38 Script sql para creación de índices en las llaves foráneas de la base de datos	115
Figura 39 Script Python para para procesar las consultas sql y generar los archivo csv.....	118
Figura 40 Script de Python para integrar todas las consultas en un solo dataframe	119
Figura 41 Conteo de valores nulos por columna en el dataset final.....	124
Figura 42 Script de imputación de nulos y verificación de que se aplicó.....	125
Figura 43 Histogramas de las variables socioeconómicas del dataset con curva de densidad superpuesta	131
Figura 44 Variables socioeconómicas con un alto porcentaje de valores en cero	134
Figura 45 Diagramas de caja de las variables socioeconómicas del dataset.....	135

Figura 46 Visualización de la matriz las variables socioeconómicas para altas correlaciones	138
Figura 47 Cálculo del VIF de las variables socioeconómicas presentes en el dataset.....	147
Figura 48 Gráfico de barras del VIF de las variables socioeconómicas del dataset	148
Figura 49 Recalculando el VIF de las variables socioeconómicas que quedaron	149
Figura 50 Matriz de análisis de información muta de todas las combinaciones de las variables socioeconómicas	151
Figura 51 Técnica de la importancia de la permutación con regresión logística a partir de etiquetas artificiales generadas con K-means	153
Figura 52 Importancia de características con Random Forest	155
Figura 53 Población del Departamento del Cauca - Nombres de municipios	157
Figura 54 Habitantes por municipio	157
Figura 55 Tasa de aprobación en educación básica y media.....	158
Figura 56 Atención Infantil en ICBF y Preescolar (número de niños atendidos).....	158
Figura 57 Cobertura neta en educación básica y media (Porcentaje 0-100).....	159
Figura 58 Tasa de desempleo (Censo 2018, Porcentaje 0-100)	159
Figura 59 Tasa de deserción escolar en educación básica y media.....	160
Figura 60 Hectáreas con vocación agrícola condicionada	160
Figura 61 Hectáreas con vocación agrícola no condicionada	161
Figura 62 Hectáreas en explotación ilegal de metales preciosos	161
Figura 63 Índice de Condiciones de Vivienda	162
Figura 64 Personas en Hacinamiento Crítico	162
Figura 65 Personas con Indice de Pobreza Multidimensional	163
Figura 66 Porcentaje de Niños con Bajo Peso al Nacer	163
Figura 67 Porcentaje de Población en Estrato 1	164
Figura 68 Porcentaje de Población en Estrato 1	164
Figura 69 Afiliación a Seguridad Social en Salud	165
Figura 70 Tasa de Alfabetismo.....	165
Figura 71 Tasa General de Mortalidad	166
Figura 72 Tasa de Mortalidad Infantil.....	166
Figura 73 Mapa de hectáreas de hoja de coca reportadas por el Ministerio de Justicia	167
Figura 74 Número de viviendas en el SISBEN	167
Figura 75 Histogramas de las variables sobre violencia del dataset con curva de densidad superpuesta	170
Figura 76 Variables sobre violencia con un alto porcentaje de valores en cero	172
Figura 77 Diagramas de caja de las variables sobre la violencia del dataset.....	173
Figura 78 Visualización de la matriz las variables sobre la violencia para altas correlaciones	175
Figura 79 Cálculo del VIF de las variables sobre la violencia presentes en el dataset.....	181
Figura 80 Nuevos valore del VIF de las variables sobre la violencia después de fusión de variables..	182
Figura 81 Matriz de análisis de información muta de las variables sobre la violencia	183
Figura 82 Script para implementar decisiones de normalizar por población.....	185
Figura 83 Mapa de tasa de armas confiscadas por la Policía Nacional (2019-2023)	186
Figura 84 Mapa de la tasa de delitos registrados por la Fiscalía General (2019-2023).....	187
Figura 85 Mapa tasa incautaciones de drogas ilícitas (kls/10M per) - Min-Defensa (2019-2023)	187
Figura 86 Mapa tasa reclamantes de restitución de tierras – reporte Min-Agricultura	188
Figura 87 Mapa de miembros de la fuerza pública víctimas del conflicto	188
Figura 88 Mapa de tasa víctimas por declaración reportadas por la Unidad de Víctimas	189
Figura 89 Proceso de escalado de datos de variables sobre la violencia	190
Figura 90 Detección de anomalías en datos de violencia con Isolation Forest	192
Figura 91 Mapa de municipios anómalos en el Departamento del Cauca	193
Figura 92 Importancia de las características en la detección de anomalías en los datos con Isolation Forest.....	194

Figura 93 Importancia de variables de violencia en la detección de anomalías	195
Figura 94 Distribución de anomalías por pares de variables	196
Figura 95 Explorando diversas combinaciones de configuración de K-means	197
Figura 96 Segmentación inicial de los municipios. Sin anomalías, sin PCA y con 3 clústeres	198
Figura 97 Listado de municipios por clúster	199
Figura 98 Ubicación de los centroides en el espacio de las variables.....	199
Figura 99 Estadísticas descriptivas por clúster.....	201
Figura 100 “Tesselation de Voronoi” en 2D por cada par de variables.....	202
Figura 101 Análisis comparativo de los clústeres mediante gráficos de boxplots	203
Figura 102 Distribución de los municipios del Cauca según los clústeres identificados	204
Figura 103 Evaluación de K-means con configuración básica: Coeficiente de Silueta e Inercia	205
Figura 104 Visualización de los coeficiente de Silueta por clúster y variable mediante Gráfico de Lollipop	206
Figura 105 Visualización mediante gráfico de silueta	207
Figura 106 Evaluación del número óptimo de clusters mediante el método del codo	208
Figura 107 Análisis de varianza explicada por PCA.....	209
Figura 108 Loadings de las variables en cada componente	210
Figura 109 Grafica de los pesos de las variables en cada componente.....	211
Figura 110 Análisis de sensibilidad: Número de componentes PCA y calidad de la agrupación	213
Figura 111 Gráfica de análisis de sensibilidad componentes PCA y calidad de la agrupación	213
Figura 112 Asignación de clúster a cada municipio con la configuración refinada de K-means.....	215
Figura 113 Lista de municipios por cluster	215
Figura 114 Matriz de centroides en el espacio PCA	216
Figura 115 Estadísticas descriptivas por clúster con K-means con configuración refinada.....	218
Figura 116 Visualización de clústeres con límites de decisión de K-means refinado	220
Figura 117 Distribución de los municipios del Cauca por clústeres modelo refinado	221
Figura 118 Evaluación de K-means con configuración refinada: Coeficiente de Silueta e Inercia	221
Figura 119 Visualización indicadores de desempeño modelo refinado.....	222
Figura 120 Resultado de análisis PCA excluyendo municipios anómalos	223
Figura 121 Método del codo excluyendo anomalías y usando PCA	225
Figura 122 Aplicación de K-means con nueva configuración excluyendo anomalías	225
Figura 123 Distribución de municipios por clúster	226
Figura 124 Distribución de los municipios del Cauca por clústeres modelo con exclusión de anomalías	227
Figura 125 Matriz de centroides del nuevo modelo	227
Figura 126 Estadísticas descriptivas por clúster con k-means refinado excluyendo anomalías.....	229
Figura 127 Visualización de clústeres con límites de decisión de K-means excluyendo anomalías	230
Figura 128 Resumen de las configuraciones de k-means aplicadas a las variables de violencia	231
Figura 129 Script para verificar relación entre variables socioeconómicas y perfiles de violencia en el Cauca, mediante regresión logística	233
Figura 130 Coeficientes de regresión logística por variable y clúster	234
Figura 131 Visualización de los coeficientes de la regresión logística por perfil de violencia.....	235
Figura 132 Coeficientes de la regresión categorizando las variables en función de su impacto	236
Figura 133 Razones de probabilidad de la regresión logística por variable y clúster.....	239
Figura 134 Visualización de las razones de probabilidad de la regresión logística por perfil de cluster	240
Figura 135 Evaluación del modelo de regresión logística	242

Resumen

Este proyecto explora la violencia en el Departamento del Cauca, Colombia, mediante analítica de datos y machine learning. La investigación se desarrolla en dos etapas: la primera abarca la creación de un repositorio virtual en PostgreSQL que integra datos socioeconómicos y de violencia de todos los municipios del país. Este repositorio, con trazabilidad de fuentes y georreferenciación, generó una mega-base de más de 128 millones de registros, 66 tablas y 660 campos, quedando como un recurso académico valioso para proyectos futuros.

En la segunda etapa, del repositorio se extrae la data del Cauca mediante consultas SQL y se organiza en dos conjuntos: variables socioeconómicas y de violencia. Esta información se analiza y depura mediante técnicas tradicionales y métodos de machine learning hasta generar vistas minables. A las variables de violencia se aplican detección de anomalías, PCA y clustering con diferentes configuraciones. El modelo k-means clasifica los municipios en tres clústeres y un grupo anómalo según sus perfiles de violencia, con indicadores de desempeño como el coeficiente de silueta e inercia que evidencian una segmentación moderadamente fuerte para una data tan compleja.

Los clústeres de violencia se utilizan como etiquetas para las variables socioeconómicas y se analizan mediante regresión logística multinomial. A través del estudio de los coeficientes log-odds y las odds ratios, se examina cómo factores socioeconómicos específicos de los municipios se relacionan con distintos niveles de violencia, identificando patrones y relaciones significativas entre las variables para una comprensión más profunda del fenómeno de la violencia en el Cauca.

Palabras clave

- Minería de datos
- Machine learning
- Violencia en Colombia
- Analítica de datos
- Modelos no supervisados
- Modelos supervisados

Abstract

This project explores violence in the Department of Cauca, Colombia, through data analytics and machine learning. The research is conducted in two stages: the first involves creating a virtual repository in PostgreSQL that integrates socioeconomic and violence-related data from all municipalities in the country. This repository, with source traceability and georeferencing, generated a mega-database with over 128 million records, 66 tables, and 660 fields, serving as a valuable academic resource for future projects.

In the second stage, data from Cauca is extracted from the repository using SQL queries and organized into two datasets: socioeconomic variables and violence variables. This information is analyzed and refined using traditional techniques and machine learning methods to generate minable views. Anomaly detection, PCA, and clustering with various configurations are applied to the violence variables. The k-means model classifies the municipalities into three clusters and one anomalous group based on their violence profiles, with performance indicators such as the silhouette coefficient and inertia demonstrating a moderately strong segmentation for such complex data.

The violence clusters are used as labels for the socioeconomic variables and analyzed through multinomial logistic regression. By studying the log-odds coefficients and odds ratios, the analysis examines how specific socioeconomic factors of the municipalities relate to different levels of violence, identifying significant patterns and relationships among the variables for a deeper understanding of the violence phenomenon in Cauca

Keywords

Data mining
Machine learning
Violence in Colombia
Data analytics
Unsupervised models
Supervised models

Introducción

La violencia es un fenómeno complejo que afecta profundamente la sociedad colombiana, y el departamento del Cauca no es la excepción. De hecho, en los últimos años, ha ocurrido un aumento en los índices de violencia en esta región, con graves consecuencias para la vida de sus habitantes y para el desarrollo social y económico del Departamento. Desde las estadísticas se evidencia que solo para el 2023 asesinaron a 38 líderes y lideresas sociales, es decir, que el departamento generó el 20% de este terrible hecho respecto a nivel nacional y, en lo que va del 2024 han asesinado a 8 personas que se encargan de defender los derechos sociales, económicos, ambientales y políticos de sus pobladores. Cifras que permiten visualizar la complejidad de este fenómeno. (González Perafán, Espitia Cueca, & Cabezas Palacios, 2023).

El fenómeno de la violencia ha sido abordado desde múltiples perspectivas en diversas disciplinas como la sociología, la antropología, la psicología, la filosofía, el derecho y la teoría política. Sin embargo, encontrar una definición universalmente aceptada de violencia es una tarea desafiante debido a su naturaleza multifacética y multidimensional (Nateras González, 2021). Cada disciplina aporta su propia comprensión del concepto, lo que refleja la complejidad inherente de este fenómeno.

No obstante, a pesar de las diferentes definiciones existentes, para la presente investigación, la violencia se define como cualquier acción en la que se usa la fuerza para ir en contra de la naturaleza de una persona, constreñir su voluntad, o transgredir lo que una sociedad considera justo según el derecho establecido (Jiménez García, Manzano Chávez, & Mohor Bellalta, 2021), un fenómeno inherente a la condición humana, que conlleva la imposición de la propia voluntad sobre la de otros, ya sea de manera consciente o inconsciente. Se manifiesta en diversos niveles, desde las interacciones cotidianas (micro sociales) hasta las estructuras sociales y políticas (macrosociales). Además del daño físico, abarca también el control y la manipulación como mecanismos de coerción (Vidal, Mejía González, & Curiel Gómez, 2021).

La violencia se manifiesta en una amplia gama de formas y contextos, lo que lleva a algunos autores a preferir el término “violencias”, por reflejar mejor las diversas realidades de este fenómeno. Para estos autores no existe la violencia en singular, sino las violencias en plural (Creittiez, 2009). Es por ello que este fenómeno se manifiesta en una gran variedad de expresiones tales como: física, psicológica o emocional, escolar sexual, intrafamiliar,

política, género, económica, estructural, institucional, racial, religiosa, entre otras, cada una con sus características y efectos específicos.

En este contexto, el presente proyecto de investigación se propone analizar en detalle la problemática de la violencia en el Departamento del Cauca desde una perspectiva multidimensional. Se empleará el análisis de datos como herramienta principal, junto con técnicas no supervisadas y supervisadas de machine learning, para explorar e identificar patrones característicos que nos ayuden a comprender mejor las causas y dinámicas subyacentes de las diversas formas de violencia en la región, para extraer conocimiento que pueda servir como insumo a las autoridades encargadas de tomar las decisiones orientadas a prevenir y mitigar este fenómeno. (Yup de León, 2021) Además de visibilizar las diversas formas de violencia presentes en cada uno de los municipios del Departamento del Cauca, se busca comprender la magnitud del problema y facilitar, por parte de las autoridades correspondientes, el desarrollo de estrategias dirigidas a prevenir y combatir dichas formas de violencia.

Si bien las clasificaciones sobre violencias son múltiples, para la presente investigación se referencia la violencia desde varios ámbitos, una de ellas es la que ofrece la comisión de la verdad, que clasifica a las víctimas a partir de las acciones violentas ocurridas del conflicto armado, en este sentido se tiene: acciones bélicas, asesinatos selectivos, ataques a poblados, atentados terroristas, daño a bienes civiles, desaparición forzada, masacres, reclutamiento y utilización de niños y adolescentes, secuestro, violencia sexual, minas antipersonal y municiones sin explotar. (Comision de la Verdad, 2024), y como víctimas, en este contexto específico, se entiende las definidas en el artículo 3 en la ley 1448 de 2011 (Denominada también ley de víctimas), es decir a aquellas personas que individual o colectivamente han sufrido un daño por hechos ocurridos a partir del 1 de enero de 1985, como consecuencia de infracciones al Derecho Internacional Humanitario o de violaciones graves y manifiestas a las normas internacionales de Derechos Humanos, ocurridas con ocasión del conflicto armado interno. (Congreso de la República de Colombia, 2011).

La tipología de la violencia a partir del conflicto armado es una de las más dicientes por su temporalidad, escalabilidad y el entorno en que se desarrolla, pero no es la única violencia que afecta a la región, también se identifican la violencia de género, aquella se desprende de la condición de ser mujer, personas LGBTQ+ u hombre, que generalmente se orienta de un género hacia el otro (López Gonzalez, 2020); la violencia doméstica, que abarca cualquier forma de abuso físico, psicológico, sexual o económico dentro del ámbito familiar (Vega,

Reynoso, & Corrales, 2022), la violencia racial y étnica, que se comete en contra de una persona a causa de su origen étnico (Lange, 2022), la violencia política, aquella que se orienta para lograr objetivos políticos o de poder y puede ser realizada por el Estado, por organizaciones paralegales o por organizaciones insurgentes (Levalle S. , 2018), entre otras.

El proyecto no se limita a un tipo particular de violencia, sino que busca abarcar las diferentes formas de violencia que están presentes en los municipios del Departamento del Cauca. Una parte de los datos que nos ayudarán en este propósito provienen de la recopilación de estadísticas delictivas acopiadas por autoridades de diferente orden, ya que según González (2021), “ La incidencia delictiva se ha vuelto un indicador de los niveles de violencia que tiene un país (...). La criminalidad hay que describirla y comprenderla, debido a que explica parte de esta violencia” (p. 307), particularmente de los denominados delitos violentos, es decir aquellos que involucran fuerza o amenaza de fuerza y comportamientos intencionalmente amenazantes que causan daño físico” (Johnson Criminal Law Group, 2024). También se considerarán como fuentes de datos otras estadísticas relacionadas con la violencia como son el número de víctimas, las denuncias de delitos, el conteo de actuaciones de las autoridades judiciales y policiales, los reportes de incautación de armas y de estupefacientes, reportes de cultivos ilícitos, reportes de enfrentamientos entre grupos criminales entre sí o con las fuerzas estatales, reporte de desaparecidos, reportes de población desplazada, entre otros.

Todas estas cifras y datos sobre la violencia serán de fuentes oficiales tales como la Fiscalía General de la Nación, La Policía Nacional, El Ministerio de Defensa, El Ministerio del Interior, El Ministerio de Justicia y del Derecho, El Instituto Nacional de Medicina Legal y Ciencias Forenses, el Centro Nacional de Memoria Histórica, La Unidad para la Atención y Reparación Integral a las Víctimas del Departamento de la Prosperidad Social, entre otras, a las que se puede acceder a través de la Plataforma de Datos Abiertos de Colombia (Plataforma Nacional de Datos Abiertos de Colombia, 2024), así como de otras fuentes no oficiales como El Instituto de Estudios para el Desarrollo y la Paz - INDEPAZ (INDEPAZ, 2024), Universidades públicas (humanas.bogota.unal.edu.co, 2024) y privadas (Universidad Javeriana, 2024), observatorios de violencia (SISPRO, 2024), entre otras fuentes.

Además de las cifras que proporcionan información sobre la incidencia de la violencia en Colombia, la vista minable del proyecto se alimentará de otras bases de datos oficiales, como los censos poblaciones del DANE, que ofrecen datos sobre coberturas en servicios públicos, tamaño promedio de los hogares, condiciones de las viviendas, distribución de la población

por edad y sexo, indicadores de calidad de vida, así como otros indicadores sociodemográficos y socioeconómicos por cada municipio (Dane, 2024). También se explorarán las bases de datos del Departamento Nacional de Planeación, que entre otra información relevante, ofrecen muestras anonimizadas de las encuestas del Sisbén, proporcionando información general sobre las personas, condiciones de las viviendas, conformación y activos de los hogares, entre otros aspectos. (Departamento Nacional de Planeación -DNP, 2024). Asimismo se integrarán datos sobre escolaridad y salud pública de la población del Departamento del Cauca.

Todo este proceso de recopilación y extracción de información permitirá construir una base de datos propia del proyecto con la información socioeconómica y de violencia más relevante de los municipios del País. Desarrollar esta base de datos es importante para el proyecto porque permite consolidar la información de diversas fuentes en un único repositorio virtual, lo que facilita el proceso de ingeniería de características y la construcción de la vista minable sobre la que finalmente se aplican las técnicas de machine learning para identificar patrones, tendencias y relaciones en los datos recopilados, que permitan comprender y explicar, desde la data, el fenómeno de la violencia (Clur, 2022). Se espera que este repositorio tenga un alcance que va más allá de este proyecto y se convierta en un recursos académico de valor para futuros proyectos de analítica de datos.

Planteamiento del Problema

1. Diagnóstico de la situación actual

El Departamento del Cauca enfrenta una situación alarmante en términos de violencia. Desde 2019 se ha observado la proliferación de nuevos grupos armados, superando incluso los que estaban antes de la firma del Acuerdo de Paz suscrito entre las Farc y el Gobierno Nacional en septiembre de 2016. El aumento en el asesinato de líderes sociales, el confinamiento, el desplazamiento y reclutamiento de menores es especialmente preocupante. En este departamento confluyen todos los actores armados, lo que exacerba la violencia y la inseguridad (Corredor Rodríguez, 2023).

El Cauca está dividido en siete subregiones: pacífica, norte, oriente, centro, sur, macizo y amazónica, cada una con una diversidad de población, con identidad y desafíos propios, que incluye comunidades indígenas, afrodescendientes, raizales, palenqueros y campesinos. Esta diversidad ha generado una dinámica de violencia diferenciada según el grupo y la región. La institucionalidad ha sido débil en el control de estas violencias, con una falta de programas sociales efectivos y una limitada presencia militar que ha desembocado en un aumento del reclutamiento de menores, ataques contra líderes sociales, desplazamientos masivos y actos terroristas (Torres Erazo , 2023).

Las cifras reflejan esta realidad. El Registro Único de Víctimas (RUV) de la Unidad para la Atención y Reparación Integral a las Víctimas incluye a 538.446 personas oriundas del Departamento del Cauca como víctimas del conflicto armado, por diferentes hechos victimizantes (actos terroristas, atentados, combates, enfrentamientos, hostigamientos, amenazas, delitos contra la libertad y la integridad sexual, desaparición forzada, entre otros), lo que representa el 5.57% del total de víctimas reconocidas por el conflicto armado a nivel nacional, que hasta el 29 de febrero de 2024 sumaba un total de 9.659.204 (Unidad de victimas, 2024), mientras que el Departamento del Cauca alberga apenas el 2,9% del total de la población colombiana (Teleencuestas, 2024).

El Sistema de Información Socioeconómica del Cauca – Tangara, de la Oficina Asesora de Planeación de la Gobernación del Cauca, también presenta indicadores que dan cuenta del grave problema que representa la violencia en el Departamento. La tasa de homicidios por cada 100.000 habitantes del departamento creció de 97.8 en 2019 a 117,3 en 2021. Los hurtos

que habían disminuido de 5.452 en el 2019 a 3.266 en el 2020, volvieron a crecer hasta 5.525 en el 2022. Además, entre el 2019 y 2023, se registraron 205 asesinatos de líderes y 102 masacres (Oficina Asesora de Planeación de la Gobernación del Cauca, 2024).

El Instituto de Estudios para el Desarrollo y la Paz - INDEPAZ, ONG que colabora con actividades por la paz de Colombia, informó sobre 11 actos graves de violencia en el Cauca en tan solo la semana del 16 al 23 de diciembre del 2023, incluyendo la masacre de 5 indígenas y el asesinato del alcalde de Guachené, un asesinato en el municipio de Buenos Aires, un asesinato de un líder social campesino en Silvia, el secuestro de un padre y su hija en Corinto, el asesinato de un menor desmovilizado en Piendamó, y otros hechos similares. La mayoría de estos actos parecen estar relacionados con grupos armados y conflictos por control territorial. Todo este contexto pone de manifiesto la complejidad de los desafíos de seguridad que enfrenta el Departamento (Mejia, 2023).

A pesar de los esfuerzos del Gobierno Nacional por implementar una política de "paz total", que busca abordar no solo la confrontación bélica y la violencia armada, sino también las raíces sociales y económicas que la generan (Camaro, 2022), y de los intentos de la Fuerza Pública por recuperar el control y garantizar la seguridad en el Departamento del Cauca, la situación de violencia sigue siendo extremadamente compleja y aún no se vislumbra una solución clara. Persisten dudas sobre la efectividad de las estrategias actuales. Se observa cierto grado de improvisación y desacuerdo en la implementación de la política de "La paz total", lo que podría llevar a convertirla en una "paz parcial" o, peor aún, en un recrudecimiento de la violencia (García, 2023). De allí que es crucial explorar y adoptar enfoques innovadores que puedan restablecer la paz y el orden en una de las regiones más conflictivas de Colombia.

Existen múltiples fuentes de datos oficiales y no oficiales que documentan tanto las violencias como las características socioeconómicas del departamento, pero la información se encuentra fragmentada y dispersa en diferentes fuentes de datos. Esta falta de centralización y unificación de los datos dificulta un análisis integral y holístico de la problemática de la violencia en el Cauca. Es por ello que se plantea la necesidad de construir una base de datos robusta que combine esta información de manera sistemática. La creación de esta base de datos permitirá integrar indicadores socioeconómicos con registros de diferentes tipos de violencia, facilitando un análisis más preciso y contextualizado de la situación. Esta base de datos será el insumo esencial para aplicar técnicas de machine learning que puedan identificar patrones y tendencias en la violencia a lo largo del periodo de estudio.

2. Predicción o pronóstico de la situación descrita

Se prevé que la violencia en el Cauca continúe, e incluso se intensifique, si persisten y no se abordan las causas subyacentes del conflicto. Entre estas causas están la presencia de múltiples actores armados con diversos intereses, la carencia de oportunidades económicas y sociales para la población, la fragilidad institucional que perpetúa la impunidad, la existencia de cultivos ilícitos y minería ilegal, el tráfico de drogas, entre otras. Las afectaciones del territorio y la población del Departamento continuará, con todas las consecuencias negativas que esto implica: inestabilidad social y escaso desarrollo económico en la región, desconfianza en las instituciones, violación de los derechos humanos, limitación del acceso a servicios públicos básicos, educación y salud, pérdida de vidas humanas tanto de combatientes como de civiles inocentes, desplazamiento forzado interno o incluso exilio, destrucción de infraestructura pública y privada, y profundización de divisiones sociales y étnicas, reclutamiento de menores, daños psicológicos, entre otros (Aguirre Bonilla, Manquillo Gallego , Villota Cabrera , Correa Escobar , & Toro Cano , 2021).

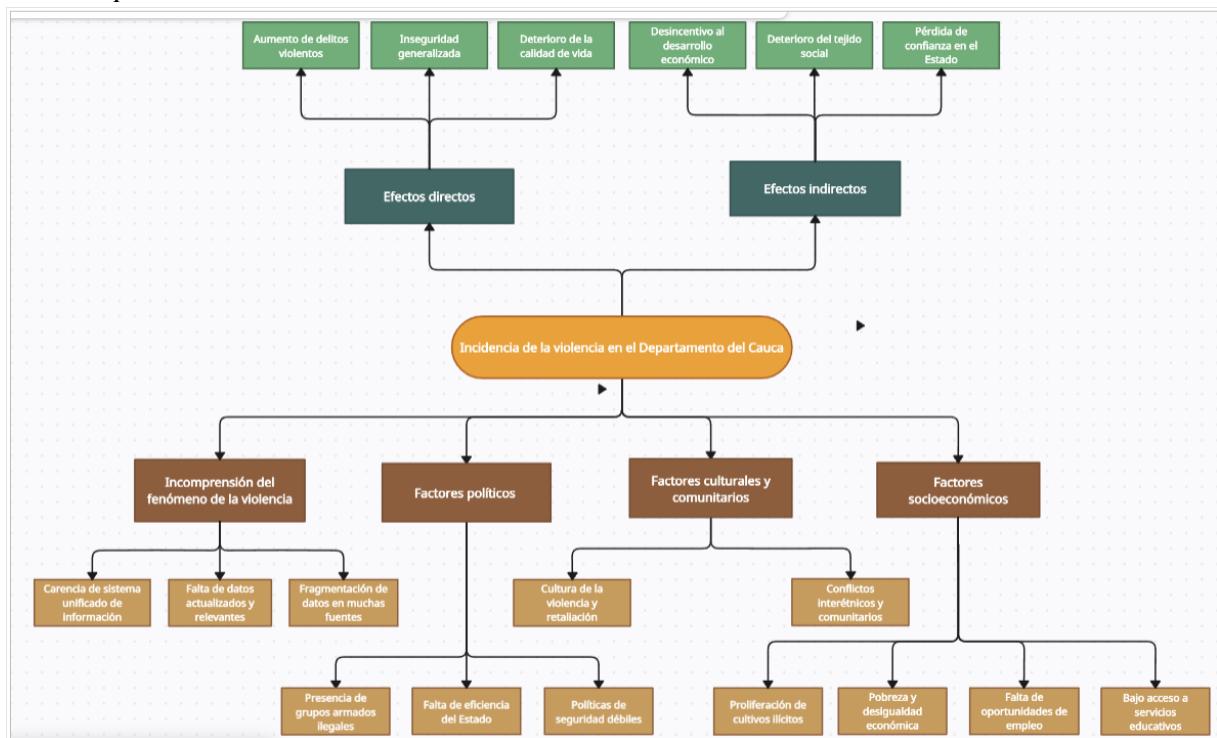
Los grupos armados ilegales en su afán por obtener poder y control sobre las rentas ilícitas, no dudan en utilizar métodos de guerra inhumanos como minas y artefactos explosivos, las masacres, el asesinato de personas protegidas y de la población civil en general (Valencia Muñoz, García Álvarez, Ortega Solarte, & Díaz, 2021).

Mientras las economías ilícitas, como los cultivos de coca y la minería ilegal, sigan siendo rentables, el conflicto y la disputa por el control de estos recursos territoriales se prevé que continúe. Aunque últimamente se ha observado una aparente disminución en la rentabilidad del negocio de la coca debido al aumento en el valor de los insumos y la sobreproducción, los grupos armados están desplazando su interés e incrementando sus actividades en la minería ilegal, especialmente en la extracción de oro. Estas actividades ilícitas conllevan procesos de migración que abarcan tanto a personas que llegan en busca de empleo en las actividades ilícitas como a aquellos que abandonan los territorios debido al aumento de la violencia (Espitia Cueca & González Perafán, 2023). La confluencia entre el conflicto armado y el crimen organizado, particularmente en lo que respecta al narcotráfico y la minería ilegal, es una realidad predominante en el Departamento del Cauca, lo que sugiere que la violencia en todas sus manifestaciones tenderá a persistir e incluso a intensificarse (Valencia Londoño, 2022).

3. Arbol del problema

A continuación se presenta el árbol de problemas que tiene como propósito descomponer y estructurar de manera visual y comprensible las causas y efectos de la violencia en el Departamento del Cauca. Esta herramienta de análisis permite organizar y comprender de manera clara y detallada la compleja problemática de la violencia en la región.

Figura 1
Arbol de problemas



Fuente: construcción propia

Seguidamente se presenta una breve explicación de las causas y efectos relacionados en el árbol de problemas anteriormente graficado.

3.1 Problema central

La alta incidencia de violencia en el Departamento del Cauca. Es el problema central debido a su magnitud y su impacto negativo en la población. Este proyecto de investigación, mediante el uso de la analítica de datos, busca comprender mejor este fenómeno, identificar patrones y tendencias, y proporcionar un insumo para que los tomadores de decisiones formulen políticas públicas más efectivas y coordinadas

3.2 Causas

Existe una diversidad de causas que generan el problema central de la alta incidencia de violencia en el Departamento del Cauca. Estas causas se pueden agrupar en factores socioeconómicos, factores políticos, factores culturales y comunitarios, y aquellos relacionados con la incomprendición del fenómeno de la violencia.

Factores socioeconómicos

- **Pobreza y desigualdad económica:** La alta incidencia de pobreza y desigualdad crea un caldo de cultivo para la violencia, ya que muchas personas se ven forzadas a recurrir a actividades ilegales como una forma de subsistencia (Arzate Salgado, 2022).
- **Falta de oportunidades de empleo:** La escasez de empleos formales y bien remunerados empuja a muchos jóvenes hacia la delincuencia y los grupos armados ilegales, buscando alternativas económicas (Obando Guerrero, Pérez Caicedo, Cuastumal Meneses, & Hernández Narváez, 2020).
- **Bajo acceso a servicios educativos:** La educación insuficiente y de baja calidad limita las oportunidades de desarrollo personal y profesional, perpetuando ciclos de pobreza y violencia (Vidal, Mejía, & Curiel, 2021).
- **Proliferación de cultivos ilícitos:** La economía ilegal basada en el cultivo de drogas financia la violencia y proporciona ingresos a los grupos armados, exacerbando la inseguridad en la región (Miranda & Rodríguez, 2020).

Factores políticos

- **Presencia de grupos armados ilegales:** La existencia de múltiples grupos armados, incluyendo guerrillas y bandas criminales, alimenta el ciclo de violencia. Estos grupos luchan por el control territorial y los recursos, perpetuando el conflicto (García Pinzón & Fernando Trejos, 2021)

- **Falta de eficiencia del Estado:** La debilidad institucional y la ineficiencia en la implementación de políticas de seguridad permiten que los grupos armados operen con relativa impunidad. La falta de presencia estatal en áreas rurales agrava el problema (Behar Villegas, 2021)
- **Políticas de seguridad débiles:** Las estrategias de seguridad ineficaces y mal coordinadas no logran disuadir la violencia ni proteger adecuadamente a la población civil. Las políticas actuales no abordan de manera integral las causas subyacentes del conflicto (Alvarado Mendoza & Padilla Oñate, 2021).

Factores culturales y comunitarios

- **Cultura de la violencia y retaliación:** En algunas comunidades, la violencia se ha normalizado como un medio de resolver conflictos, perpetuando un ciclo de retaliaciones y venganzas que alimenta la criminalidad (Hincapié & Rodríguez, 2021)
- **Conflictos interétnicos y comunitarios:** El Cauca es una región diversa étnicamente, y las tensiones entre diferentes grupos comunitarios y étnicos pueden desencadenar violencia. La competencia por recursos y territorio a menudo se traduce en enfrentamientos (Cortés Landazury, 2024).

Incomprensión del fenómeno de la violencia.

Si bien es cierto que el hecho de comprender el fenómeno de la violencia no causa directamente el problema, si permite que el fenómeno persista y se multiplique, al dificultar la formulación y aplicación de estrategias efectivas de intervención. La falta de comprensión integral del fenómeno de la violencia en el Cauca y la dispersión de la información contribuyen a una respuesta fragmentada e ineficaz por parte de las autoridades y la sociedad en general.

- **Carencia de sistema unificado de información:** La ausencia de un sistema centralizado de información sobre incidentes violentos y condiciones socioeconómicas de los municipios dificulta el análisis y la formulación de políticas efectivas. La fragmentación de datos en múltiples fuentes impide una visión coherente y precisa del problema.

- **Falta de datos actualizados y relevantes:** La escasez de datos actuales y precisos sobre la violencia limita la capacidad de las autoridades y organizaciones para tomar decisiones informadas.
- **Fragmentación de datos en muchas fuentes:** La dispersión de información entre diversas entidades y bases de datos crea inconsistencias y dificulta el seguimiento de tendencias y patrones de violencia.

3.3 Efectos

Efectos directos

- **Aumento de delitos violentos:** La violencia en el Cauca se manifiesta en un incremento de homicidios, secuestros, extorsiones y otros crímenes violentos que afectan a la población diariamente (Aguirre Bonilla, Manquillo Gallego , Villota Cabrera , Correa Escobar , & Toro Cano , 2021).
- **Inseguridad generalizada:** La percepción y la realidad de la inseguridad se extienden por toda la región, afectando la vida cotidiana de sus habitantes y limitando su libertad de movimiento (Noticias RCN, 2024).
- **Deterioro de la calidad de vida:** La violencia impacta negativamente en diversos aspectos de la vida, incluyendo la salud mental y física, la cohesión social y el acceso a servicios básicos (Tarazona & Ríos, 2021).

Efectos Indirectos

- **Desincentivo al desarrollo económico:** La violencia ahuyenta inversiones y limita el crecimiento económico, agravando la pobreza y la desigualdad en la región (Rojas & Maldonado, 2023)
- **Deterioro del tejido social:** La violencia erosiona la confianza y la solidaridad entre los miembros de la comunidad, debilitando las redes de apoyo y cooperación social (Santana & Núñez, 2021).

- **Pérdida de confianza en el Estado:** La incapacidad del Estado para garantizar la seguridad y el bienestar de sus ciudadanos genera desconfianza y deslegitimación de las instituciones públicas (Norza Céspedes, Molano, Harker, & Buitrago Cubides, 2020)

4. Antecedentes

Durante la revisión de distintos gestores bibliográficos tanto nacionales como internacionales, que incluyen artículos, revistas, trabajos de grado y otros documentos, no se encontró un tratamiento similar al enfoque de este trabajo de grado en términos de contexto, conjuntos de datos y metodología. Tampoco se encontraron investigaciones que aborden de manera específica la violencia en el Cauca en un contexto geoespacial detallado por municipios, ni que abarque todas las manifestaciones de violencia que afectan a esta región, pero si existen estudios relevantes que contribuyen al enfoque metodológico de esta investigación. La literatura consultada presenta aplicaciones de técnicas de machine learning en el análisis de la violencia, ya sea para una de sus manifestaciones o en regiones o ciudades distintas al Cauca.

Estos estudios, aunque limitados en cuanto a su alcance geográfico y el tipo de violencia abordado, ofrecen un marco útil para la aplicación de algoritmos no supervisados en el análisis de patrones de violencia. Las investigaciones revisadas han permitido identificar metodologías y enfoques analíticos que pueden ser adaptados para el caso del Cauca, considerando las particularidades de sus municipios y la complejidad de las variables involucradas.

A continuación, se presenta una revisión de algunos estudios, tres para ser exactos, que han utilizado técnicas avanzadas de análisis de datos para estudiar fenómenos de violencia, y que se relacionan de una u otra manera con el presente proyecto.

En primer lugar, se destaca el trabajo de Ordoñez Eraso, Pardo Calvache y Cobos Lozada (2020) titulado *Detection of Homicide Trends in Colombia Using Machine Learning*, que aborda el fenómeno de los homicidios en Colombia entre 1960 y 2019. Este estudio utiliza un modelo de machine learning basado en el algoritmo de random forest (bosque aleatorio) para predecir las tendencias de homicidios en el país hasta el año 2025. El conjunto de datos empleado proviene del Sistema de Información Estadística Delincuencial, Contravencional y Operativo (SIEDCO) de la Fiscalía General de la Nación.

El enfoque metodológico se inicia con el entrenamiento del conjunto de datos, que incluye 233.957 registros de homicidios ocurridos en Colombia entre 1960 y 2019. Tras un filtrado, se seleccionaron 14 columnas con las variables más relevantes. El minado de datos se llevó a cabo utilizando la metodología CRISP-DM, la cual facilita un análisis exploratorio y una comprensión más profunda de la información. Durante este proceso, se realizó una limpieza de los datos, eliminando columnas irrelevantes, valores duplicados y nulos. En la fase de transformación, se normalizaron los datos mediante el método Min-Max, y el modelo se ajustó definiendo los hiperparámetros adecuados. (Ordoñez Eraso, Pardo Calvache, & Cobos Lozada, 2020)

El modelo fue evaluado comparativamente con otros enfoques, como la regresión lineal con máquinas de vectores de soporte (SVR), la regresión lineal normal y la regresión lineal con k vecinos más cercanos (KNN), con el fin de establecer la tendencia de los homicidios hasta 2025. En general, el modelo de random forest ofreció pronósticos más cercanos a los valores reales (Ordoñez Eraso, Pardo Calvache, & Cobos Lozada, 2020).

Finalmente, los autores concluyen que, de acuerdo con el modelo, se prevé una disminución en la cantidad de homicidios en Colombia. Además, sugieren que la metodología utilizada puede aplicarse para predecir la evolución de otras variables relacionadas con la violencia. (Ordoñez Eraso, Pardo Calvache, & Cobos Lozada, 2020).

Otro estudio relevante se denomina “Predicción Geoespacial de Crímenes en Bogotá: Un Enfoque Basado en Machine Learning para Mejorar la Seguridad Ciudadana”. Este trabajo, realizado por Salazar Isairias (2024), busca desarrollar modelos de Machine Learning para proporcionar una comprensión predictiva de los hechos violentos desde una perspectiva espacial, lo que resulta un aporte significativo para este proyecto de investigación. La investigación de Salazar utiliza un modelo de aprendizaje supervisado para predecir el crimen en la ciudad de Bogotá, analizando la distribución espacial y la evolución temporal de delitos como homicidios, hurtos a personas y hurtos a celulares, durante el periodo 2016-2019 (Salazar Isairias, 2024).

El estudio proporciona un marco conceptual que incluye enfoques como el Crime Forecasting, que refiere a la predicción de la ocurrencia de crímenes utilizando modelos de regresión para analizar los crímenes como una variable explicada a partir de un conjunto de datos. Además, se utiliza el análisis de series de tiempo, que consiste en observar los valores

de una variable cuantitativa a lo largo del tiempo, permitiendo predecir futuros valores mediante datos pasados. Otro concepto importante en el estudio es el Deep Learning, el cual simula el comportamiento del cerebro humano capturando gran cantidad de información para "aprender" patrones. Finalmente, se utiliza el método de grafos, que es una estructura no ordenada que modela relaciones espaciales. (Salazar Isairias, 2024)

El dataset sobre crímenes en Bogotá se obtuvo de la plataforma queremosdatos.co, con un total de 9,000 registros. La metodología empleada incluyó: a) la identificación de los crímenes y su distribución espacial y temporal, seguida por la generación de una función de predicción para validar el modelo; b) la recolección de datos centrada en crímenes de alto impacto (homicidios, hurtos a personas y hurtos a celulares) debido a la falta de información histórica sobre otros delitos; c) el tratamiento de datos, priorizando columnas como fecha, barrio, localidad y tipo de crimen; y d) la visualización mediante mapas de calor generados con Python, para mostrar la distribución de los crímenes en la ciudad (Salazar Isairias, 2024).

El modelo de predicción requería la delimitación de regiones espaciales mediante la agrupación de barrios, lo que facilitó la representación geoespacial del Crime Forecasting. Los datos se analizaron como una serie temporal, donde cada reporte de crimen fue tratado como una variable aleatoria que representa la cantidad de incidentes en un momento dado. (Salazar Isairias, 2024)

El resultado del estudio fue una herramienta basada en Deep Learning que genera un modelo predictivo de criminalidad utilizando una estructura de grafo. Este modelo permitió capturar las relaciones espaciales inmersas en los patrones de ocurrencia de crímenes en Bogotá durante el periodo 2016-2019. (Salazar Isairias, 2024)

Como tercera referencia tenemos la investigación denominada “Aplicación de Machine Learning para análisis de los de violencia intrafamiliar en el departamento del Atlántico”, que aborda la violencia intrafamiliar en Colombia, con especial énfasis en el departamento del Atlántico, y utiliza técnicas de Machine Learning para analizar este fenómeno. El trabajo se basa en un dataset de 458,914 registros extraídos de la página de datos abiertos de Colombia, que luego fue filtrado para centrarse en 16,825 registros del Atlántico, con especial interés en Barranquilla y Soledad. Este filtrado incluyó la normalización de datos y la eliminación de columnas irrelevantes para mejorar la calidad del análisis (Chamorro, Laza, Noriega, & Rojano, 2021)

Entre las técnicas de Machine Learning aplicadas se destacan el algoritmo Simple k-means para la creación de clústeres, el árbol de decisión J48 y el método Random Forest. Estas técnicas se utilizaron para detectar patrones y clasificar los registros de violencia intrafamiliar. En el análisis a nivel nacional, Simple k-means identificó clústeres de violencia principalmente en Antioquia y Cundinamarca. Al focalizar el estudio en el Atlántico, los clústeres se centraron en Barranquilla y Soledad, donde se observó que las mujeres eran las principales víctimas, especialmente de violencia con armas contundentes (Chamorro, Laza, Noriega, & Rojano, 2021).

En cuanto a los resultados, las técnicas mostraron una alta precisión en la clasificación de los datos, con Random Forest obteniendo una precisión del 99.96%. Sin embargo, el estudio encontró que los modelos presentaban una limitación al centrarse únicamente en los municipios con mayor cantidad de casos reportados, lo que dificultaba la clasificación precisa de hechos en otros municipios. A pesar de estas limitaciones, se concluyó que las herramientas de Machine Learning ofrecen una visión general del fenómeno de la violencia, aunque su capacidad predictiva podría mejorarse con un dataset más equilibrado (Chamorro, Laza, Noriega, & Rojano, 2021)

5. Formulación o Pregunta de Investigación

La pregunta que guiará la investigación es la siguiente:

¿Cuáles son los patrones socioeconómicos que explican las diferentes manifestaciones de violencia en los municipios del Cauca durante el periodo 2019-2023, aplicando técnicas no supervisadas y supervisadas del Machine Learning?

Objetivos

1. Objetivo General

Analizar la problemática de la violencia en el Departamento del Cauca durante el periodo 2019-2023 mediante el uso de técnicas de Machine Learning, para que surjan patrones entre las características socioeconómicas de la población y las manifestaciones de la violencia en sus diferentes tipos.

2. Objetivos Específicos

- Construir una base de datos relacional que combine información socioeconómica y datos de violencia, con trazabilidad de la fuente y geolocalización por municipio, que centralice y unifique datos dispersos de diferentes conjuntos de datos.
- Aplicar técnicas no supervisadas y supervisadas de Machine Learning a una muestra de datos socioeconómicos y de violencia de los municipios del Departamento del Cauca, que permitan que se conformen grupos de municipios con características socioeconómicas similares y niveles de violencia comparables y municipios que se alejan significativamente de los patrones generales.
- Validar los resultados obtenidos utilizando técnicas de validación y métricas de evaluación para asegurar la robustez de los hallazgos, la coherencia y el significado en el contexto de la violencia en el Departamento del Cauca.

Justificación del Proyecto

La violencia en el Departamento del Cauca es un fenómeno complejo y multifacético que requiere ser abordado desde diversas perspectivas para su comprensión y eventual mitigación. En esta región la violencia ha alcanzado niveles alarmantes en los últimos años, como lo demuestran las estadísticas de homicidios, hurtos, líderes asesinados, y otras formas de violencia reportadas por diversas fuentes oficiales y no oficiales (Red de Derechos Humanos del Sur Occidente de Colombia “Francisco Isaías Cifuentes”, 2020). La presencia de múltiples actores armados, la falta de oportunidades económicas y sociales para la población, la fragilidad institucional, entre otros factores, contribuyen a la persistencia y recrudecimiento de la violencia en la región (Grupo de Redacción de Semana, 2023).

El impacto de la violencia ha sido devastador para el Departamento del Cauca, en la vida de sus habitantes y en el desarrollo social y económico de la zona (Red de Derechos Humanos, 2019), por lo que se demanda respuestas más efectivas de las que hasta ahora se han intentado. Si las soluciones intentadas para mitigar la violencia no han sido efectivas, es tal vez porque no se ha entendido integralmente el problema. Es allí donde cobra importancia el presente anteproyecto de investigación que propone emplear técnicas no supervisadas y supervisadas de análisis de datos, para profundizar en el entendimiento de la violencia en la región durante el periodo 2019-2023.

La complejidad del fenómeno de la violencia demanda un enfoque integral que vaya más allá de las estrategias convencionales de seguridad. Es necesario comprender las causas subyacentes y las dinámicas que alimentan la violencia en sus diversas manifestaciones (Neira, 2021). El análisis de datos y el uso de técnicas no supervisadas y supervisadas de Machine Learning son sin duda herramientas valiosas para este propósito. El análisis de datos ofrece la posibilidad de explorar patrones, identificar correlaciones no evidentes a simple vista, tendencias y relaciones (Greener, Kandathil, & Moffat, 2022) en la información recopilada sobre la violencia en el Departamento del Cauca, al cruzarla con la información estadística sobre las condiciones socioeconómicas y sociodemográficas, coberturas de servicios públicos, tamaño promedio de los hogares, condiciones de las viviendas, distribución de la población por edad y sexo, entre otros indicadores. Esto permitirá una comprensión más completa y objetiva del problema.

El análisis de datos y el uso de técnicas de Machine Learning permitirá explorar y comprender la complejidad de la violencia en el Cauca desde una perspectiva holística,

integrando datos de diferentes fuentes y disciplinas. Esta aproximación posibilitará identificar patrones, tendencias y relaciones en los datos que podrían pasar desapercibidos mediante métodos tradicionales de investigación. Son herramientas que permiten procesar grandes volúmenes de datos de manera eficiente y extraer información significativa (Borjas García, 2020) que puede ser utilizada para comprender mejor las causas y dinámicas subyacentes de la violencia en la región.

Desde un enfoque práctico, la realización de este proyecto tiene el potencial de generar impactos significativos en las políticas públicas dirigidas a prevenir y mitigar la violencia en el Cauca, dado que los resultados obtenidos en la investigación podrían ser utilizados por las autoridades públicas para diseñar mejores estrategias de seguridad y desarrollo que aborden las causas subyacentes de la violencia en la región, y servir como insumo para la toma de decisiones informadas por parte de las autoridades locales, regionales y nacionales, así como de organizaciones de la sociedad civil y la comunidad académica.

Alcances y limitaciones del proyecto

1. Alcances

Para alcanzar los objetivos planteados, el proyecto contempla los siguientes productos y desarrollos específicos:

- Una base de datos integral con información socioeconómica y estadísticas sobre la violencia en los diferentes municipios del País.
- Un informe escrito que describa detalladamente los resultados del análisis de datos.
- La presentación de los resultados a la comunidad académica de la CUN.

Con la creación de la base de datos del proyecto se busca centralizar y unificar en un único repositorio virtual la información socioeconómica y los datos sobre hechos de violencia en Colombia, para que sirva como la fuente primaria de datos para el proyecto y quede como un recurso para futuras investigaciones y proyectos de análisis de datos. Se hará una investigación exploratoria que permita identificar las fuentes de conjuntos de datos socioeconómicos de los municipios de Colombia, por ejemplo el DANE, el Departamento de Planeación Nacional, el IGAC, así como fuentes de conjuntos de datos sobre hechos de violencia, tales como la Fiscalía General de la Nación, La Policía Nacional, El Ministerio de Defensa, El Ministerio del Interior, El Ministerio de Justicia y del Derecho, El Instituto Nacional de Medicina Legal y Ciencias Forenses, el Centro Nacional de Memoria Histórica, La Unidad para la Atención y Reparación Integral a las Víctimas del Departamento de la Prosperidad Social, el Instituto de Estudios para el Desarrollo y la Paz - INDEPAZ, Universidades públicas y privadas, observatorios de violencia, entre otras fuentes.

Se parte del supuesto de que los conjuntos de datos publicados por estas fuentes, ya sea en sus sitios web o a través del portal de datos abiertos de Colombia, son confiables, veraces, completos y consistentes, ya que provienen de entidades con la responsabilidad de recopilar y registrar información socioeconómica y sobre hechos de violencia de manera sistemática y siguiendo metodologías técnicamente establecidas (MinTIC, 2022). Por lo tanto, la base de datos que se construya se alimentará únicamente con la información disponible en esas fuentes.

La construcción de la base de datos está precedida por la implementación de un proceso ETL, que en la etapa de **Extracción**, se concreta en buscar, seleccionar y descargar datasets relevantes que contengan información socioeconómica y de violencia en Colombia, prestando especial atención a aquellos que incluyan datos por municipios. En la etapa de **Transformación**, se adelanta el preprocesamiento de los datos, utilizando el poder de bibliotecas especializadas de Python, como pandas, numpy y scikit-learn, para realizar la limpieza inicial de los datos extraídos, eliminando valores nulos y duplicados, estandarizando variables para asegurar datos consistentes y comparables, suprimiendo columnas irrelevantes y normalizando o escalando variables numéricas donde sea necesario (Tatman, 2024). Finalmente en la etapa de **Carga**, los datos transformados son cargados en una base de datos relacional diseñada en PostgreSQL, con la extensión PostGIS para manejar información geoespacial que permita facilitar la identificación de patrones y relaciones espaciales que no serían visibles con un análisis puramente tabular y posibilitar la georreferenciación de los hechos de violencia por municipios. Esto facilitará el uso de herramientas especializadas como Geopandas de Python, que proporciona métodos para visualizar datos geoespaciales en mapas interactivos y estáticos, lo que permite la exploración y comunicación de información sobre la distribución espacial de los datos (Cook & Li, 2024).

La recolección y almacenamiento de los datos en el repositorio virtual se hará en su formato original, microdatos cuando así se encuentren. No se contempla realizar agregación o agrupación de los datos en esta primera etapa. Al mantener los datos en su estado más granular, se garantiza la escalabilidad del proyecto y se asegura un mejor aprovechamiento de la información para futuras investigaciones (Marchetti, 2024).

La base de datos que se entregará como producto de este proyecto no estará limitada únicamente a la información de violencia en el Departamento del Cauca durante el periodo 2019-2023. Por el contrario, se diseña para ser un recurso robusto y versátil, que incluye datos de múltiples fuentes y de diferentes periodos de tiempo. Algunas tablas contienen información desde el año 2016, otras se remontan a 2010, 2003, e incluso a años anteriores, dependiendo de la disponibilidad y naturaleza de los datos recolectados.

Además, esta base de datos no se restringe geográficamente al Cauca. Si bien el análisis del proyecto se centra exclusivamente en este departamento, se incorpora información de todos los municipios de Colombia. Esto permitirá que futuros investigadores y miembros de la comunidad académica, tanto de la CUN como de otras universidades, puedan utilizar esta base de datos en sus propios proyectos. De este modo, el proyecto no solo entrega un análisis

específico de la violencia en el Cauca, sino que también contribuye al desarrollo de un instrumento de investigación de largo plazo, que puede ser explotado en proyectos de investigación multidisciplinarios en el futuro.

La arquitectura de la base de datos, de naturaleza relacional y flexible, se basa en la integración de múltiples tablas interrelacionadas para facilitar análisis geoespaciales y contextuales. Una tabla central, 'data_source', registra el origen y características de cada conjunto de datos, garantizando la trazabilidad y la integridad de la información. La tabla 'municipalities' contiene información geográfica detallada de los municipios de Colombia, permitiendo obtener los tipos de geometría comúnmente utilizados (punto, polígono, línea), para relacionar los eventos y variables con su ubicación geográfica específica. A través de claves foráneas que vinculan las múltiples tablas de datos con 'data_source' y 'municipalities', se logra una contextualización precisa tanto espacial como por fuente. Esta estructura robusta, escalable y bien definida permite almacenar y consultar eficientemente grandes volúmenes de datos, facilitando análisis detallados a nivel municipal, departamental y regional (Avalos, Castilla , & Colana, 2022).

Como sistema de gestión de bases de datos (DBMS) se eligió PostgreSQL debido a su capacidad para manejar datos geográficos gracias a la extensión PostGIS, su flexibilidad para almacenar distintos tipos de datos, su escalabilidad para grandes volúmenes de información, su gestión eficiente de múltiples usuarios y su compatibilidad con herramientas de Business Intelligence como Power BI. Además, al ser de código abierto, es gratuito y cuenta con una comunidad activa que lo respalda (Pilicita Garrido & Borja López, 2020).

El segundo entregable del proyecto consiste en un informe detallado sobre los resultados obtenidos del análisis de datos realizado para comprender la problemática de la violencia en el Departamento del Cauca.

Este análisis se basará en la base de datos integral diseñada específicamente para este proyecto, previamente mencionada. El proceso comenzará con la implementación de un proceso ETL (Extracción, Transformación y Carga) que permitirá extraer los microdatos almacenados en el repositorio virtual mediante consultas en SQL, optimizando el acceso a la información relevante. Luego, se llevará a cabo la transformación de los datos utilizando Python y sus librerías especializadas, como pandas para la manipulación y limpieza de los datos, numpy para operaciones numéricas eficientes, y scikit-learn para la preparación de los datos orientada a los modelos de Machine Learning. Durante esta fase, se aplicarán diversas

técnicas de preprocessamiento, tales como la imputación de valores nulos, la normalización o estandarización de variables, y la eliminación de columnas innecesarias (Holbrook, 2024).

Adicionalmente, se llevará a cabo un exhaustivo proceso de ingeniería de características, que incluirá la selección de las variables más relevantes para el problema en estudio. Si es necesario, se transformarán y crearán nuevas características derivadas de los datos brutos, con el propósito de mejorar el rendimiento de los modelos de Machine Learning que se aplicarán en las fases posteriores. Este proceso de ingeniería de características permitirá obtener una o varias vistas minables de los datos (Spositto, Blanco, & Matteo, 2022). Las vistas procesadas serán cargadas en scikit-learn, donde serán sometidas a técnicas no supervisadas y supervisadas de Machine Learning, tales como el análisis de componentes principales para identificar las variables socioeconómicas más relevantes que expliquen los hechos de violencia, el clústering para encontrar agrupamientos de municipios con características socioeconómicas y niveles de violencia similares, la detección de anomalías para identificar municipios que se desvén de los patrones generales y la regresión logística para establecer las relaciones entre las variables socioeconómicas y los perfiles de violencia. Con la implementación de estas técnicas se espera descubrir patrones, tendencias y relaciones significativas en el conjunto de datos almacenados (Holbrook, 2024) sobre los municipios del Departamento del Cauca.

El informe también incluirá los resultados de la aplicación de técnicas de validación y las métricas de evaluación correspondientes. Esto se llevará a cabo para garantizar la robustez de los hallazgos, así como la coherencia y el significado de los resultados en el contexto de la problemática de la violencia en el Departamento del Cauca. La validación ayudará a evaluar la capacidad de generalización de los modelos construidos y las métricas de evaluación proporcionarán una medida cuantitativa del desempeño de dichos modelos (Becker, 2024).

Como entorno de desarrollo del trabajo técnico e informático se utilizará Google Colaboratory, comúnmente conocido como Colab, una plataforma de análisis de datos y aprendizaje automático basada en Jupyter Notebook. Colab proporciona un entorno de desarrollo en la nube que permite escribir, ejecutar y compartir código en Python de manera colaborativa y eficiente. Entre sus principales ventajas se encuentra el acceso gratuito a recursos de computación avanzados, como GPUs y TPUs, lo que facilita la ejecución de algoritmos de Machine Learning con altos requerimientos de procesamiento sin necesidad de contar con un hardware especializado localmente. Además, Colab integra perfectamente con

bibliotecas ampliamente utilizadas en el ecosistema de ciencia de datos, como pandas, numpy, scikit-learn, etc., permitiendo realizar análisis complejos, preprocesamiento de datos y entrenamiento de modelos de machine learning en un solo entorno. Otra ventaja clave es su capacidad para almacenar el trabajo en Google Drive, lo que asegura un acceso continuo y la posibilidad de compartir fácilmente el progreso con otros colaboradores. Colab también ofrece la posibilidad de trabajar con notebooks de manera interactiva, lo que resulta ideal para iterar rápidamente sobre diferentes enfoques y ajustar parámetros de los modelos en tiempo real (McKinney, 2022).

La presentación de la investigación ante la comunidad académica de la CUN se hará en los términos y condiciones exigidos en el reglamento de la especialización.

Como trabajos futuros que se pueden derivar del presente proyecto se pueden identificar en otros los siguientes:

- Establecer un proceso automático de actualización de la base de datos a partir de las fuentes primarias de los datos, para mantener la base de datos actualizada y con información precisa para análisis posteriores.
- Incorporar datos de otras fuentes, como encuestas de victimización y estudios cualitativos, para enriquecer la base de datos y obtener una comprensión más profunda de la problemática de la violencia.
- Desarrollar modelos de análisis predictivo para identificar zonas con mayor riesgo de violencia, para la tomar acciones oportunas de prevención y mitigación de la violencia
- Extensión del análisis a otros períodos de tiempo para comprender la evolución de la violencia mediante series de tiempo.
- Diseño de estrategias de intervención para prevenir y mitigar la violencia.
- Diseño de tableros de control o dashboard que sirvan de interfaz visual para presentar los datos recopilados y analizados. Esto permitiría a los usuarios, ya sean investigadores, autoridades locales o tomadores de decisiones, comprender fácilmente la situación de la violencia en Colombia, así como los factores socioeconómicos asociados.

- Creación de un data warehouse, lo cual representaría una evolución natural de la infraestructura de datos construida en la presente etapa

2. Limitaciones

Como en toda investigación, este proyecto enfrenta una serie de limitaciones que podrían influir en el alcance e interpretación de los resultados y en el desarrollo de las actividades planificadas, tales como:

- Aunque la base de datos del proyecto incluirá información de todos los municipios de Colombia y de períodos anteriores a 2019, el análisis de la violencia estará estrictamente limitado al Departamento del Cauca entre los años 2019 y 2023. Los datos de otros departamentos o períodos anteriores no serán considerados en el análisis del presente proyecto, lo que reduce la posibilidad de hacer comparaciones interdepartamentales, identificar tendencias históricas a largo plazo, o generalizar los resultados a otras regiones y contextos.
- Las fuentes de datos utilizadas provienen de diferentes instituciones y períodos, lo que podría llevar a inconsistencias en los registros, vacíos de información o diferencias en los criterios de recolección de datos, posibles sesgos, subregistro de ciertos eventos, enfatizar otros, etc. Esto puede impactar la precisión de los resultados obtenidos, ya que no siempre será posible corregir completamente estas discrepancias. Además, no se llevará a cabo un proceso exhaustivo de validación o imputación de datos faltantes.
- El proyecto se ha propuesto utilizar técnicas de machine learning no supervisadas, como el análisis de componentes principales (PCA), clústering y detección de anomalías y supervisadas como regresión logística, con prelación en las primeras. La prioridad de técnicas no supervisadas conlleva una limitación en cuanto a la interpretación de los resultados, ya que los hallazgos dependerán más de la estructura interna de los datos que de una hipótesis previa.
- Los resultados obtenidos estarán condicionados por la calidad de los datos disponibles y las capacidades de los algoritmos implementados. Dado que los métodos no supervisados dependen de la variabilidad y las relaciones internas entre

los datos, la presencia de sesgos en las fuentes o en los datos preprocesados puede influir en los hallazgos, generando patrones que no necesariamente reflejan la realidad de los fenómenos de violencia.

- A nivel técnico, aunque se ha implementado una base de datos con capacidades geoespaciales usando PostgreSQL y PostGIS, el procesamiento de grandes volúmenes de datos puede estar sujeto a limitaciones de tiempo y recursos computacionales. Adicionalmente las visualizaciones geográficas, pueden tener limitaciones en términos de precisión si los datos geoespaciales no están completos o no son lo suficientemente detallados para ciertas áreas. Además, debido al enfoque del proyecto, no se desarrollarán modelos complejos para estimar la relación causal entre la violencia y las variables socioeconómicas; más bien, se ofrecerán análisis descriptivos y exploratorios de los datos.
- Dado que las técnicas de machine learning no supervisadas proporcionan agrupaciones o patrones, los resultados requieren interpretación. Los patrones encontrados podrían no ser completamente explicables en términos socioeconómicos si no se dispone de todas las variables necesarias o los datos son insuficientes.
- El proyecto tiene un marco de tiempo definido para su desarrollo establecido directamente por la Universidad y, por ende, no será posible abordar todas las dimensiones del fenómeno de la violencia. Si bien se realizará un análisis sólido con los datos disponibles y las técnicas propuestas, factores como la disponibilidad de recursos, tiempo y acceso a información actualizada pueden limitar la profundidad de algunos análisis.

Marco referencial

1. Estado del arte

La violencia, ese desafortunado fenómeno humano, ha sido una constante acompañante de la humanidad desde los albores de la historia, y la literatura, que actúa como un espejo de la sociedad, no ha sido ajena a ello. Desde los relatos épicos de la antigüedad hasta las narrativas contemporáneas, la violencia ha sido un tema recurrente en la literatura y ha sido abordada por diversas disciplinas (Gamez Brambila, 2020), incluyendo la sociología, la psicología, la antropología y la historia, intentando desentrañar su complejidad.

El desafío por comprender la violencia en sus diversas manifestaciones continua vigente, y con el advenimiento de la era de la información y el gobierno de los datos, se ha observado un creciente interés en el uso de la analítica de datos y las técnicas de machine learning para lograr este propósito. Con este nuevo enfoque los investigadores intentan explorar y explicar el fenómeno de la violencia desde nuevas perspectivas, aprovechando el poder del análisis de grandes volúmenes de datos y la capacidad de identificar patrones y correlaciones no evidentes a simple vista (Pinto Muñoz , Zuñiga Sambon, & Ordoñez Erazo , 2023), en busca de desentrañar los factores subyacentes y las dinámicas complejas que alimentan la violencia en comunidades y regiones específicas.

Basta con teclear el tema directamente en un buscador web, o en uno de los sitios especializados en publicaciones científicas y académicas, entre los que encontramos según Barragán (2024), “Bases de datos como Scopus , Redalcy , IEEE, Springer , MDPI, RIP publications , PUBMED, El Sevier , libgen , repositorios de universidades, Google Académico” (p. 11), para encontrar una vasta información que da cuenta de trabajos a nivel internacional y nacional que abordan la violencia en sus diversas manifestaciones. Se encuentran estudios sociológicos, tesis de pregrado y especialización, seminarios, páginas web, informes de organizaciones e incluso libros especializados, que abordan desde las causas y efectos de la violencia hasta las estrategias para prevenirla y combatirla.

En el ámbito del análisis de datos y la violencia, el panorama es igual de prolífico. Se han empleado técnicas y metodologías que permiten analizar grandes conjuntos de datos para identificar patrones, tendencias y factores de riesgo asociados a la violencia. Nombrar cada uno de estos trabajos además de dispendioso sería poco práctico, así que a continuación se

procede a referenciar aquellos antecedentes que se considera, guardan una estrecha vinculación con la investigación propuesta en el presente trabajo:

- La gobernación del Cauca, cuenta con un sistema que consolida información estadística denominado “Sistema de Información Socioeconómica del Cauca – TANGARA”. En este sistema de información se consolidan diversas fuentes de datos con el objeto de convertirse en un instrumento de consulta de datos sociales, económicos y ambientales disponibles para todos los usuarios, principalmente para el sector académico y de investigación. De esta manera al consolidar indicadores sociodemográficos, ambientales y económicos permite acceder a información departamental y municipal, desagregada por grupos poblacionales etarios y étnicos, además de disponer de datos alrededor de las condiciones de vida, seguridad social, recursos naturales entre otros. Acceder a los diferentes datos con que cuenta la plataforma, contribuye de manera propositiva en la construcción de la investigación, en la medida que se brinda información en contexto de las dinámicas propias de los territorios del departamento del Cauca. (Oficina Asesora de Planeación de la Gobernación del Cauca, 2024).
- El Centro Nacional de Memoria Histórica cuenta con el denominado Observatorio de Conflicto y Memoria donde se sistematiza el número de víctimas del conflicto armado en Colombia permitiendo caracterizar a las personas víctimas del conflicto armado de las últimas décadas. Al interior de su base de datos, se muestra las afectaciones en los municipios del territorio nacional especialmente en: acciones bélicas, asesinatos selectivos, ataques a poblados, ataque terroristas, daños a bienes civiles, desaparición forzada, masacres, reclutamiento y utilización de niñas, niños y adolescentes, secuestros violencia sexual, minas antipersonal y municiones sin explotar, entre otros (Centro Nacional de Memoria Historica, 2024).
- El Instituto Nacional de Medicina legal y Ciencias Forenses, que lleva estadísticas oficiales sobre la caracterización de la criminalidad en nuestro país, también tiene desplegado en la Web un observatorio sobre la violencia, donde se aporta datos relevantes que permiten aproximarse a la comprensión de la dimensión del fenómeno de la violencia en Colombia. Este publica de manera permanente datos sobre indicadores de infancia, adolescencia y juventud, cifras de lesiones de causa externa en Colombia, cifras de lesiones contra la mujer, si bien, existe subregistros de las cifras, la información suministrada puede servir para identificar patrones de comportamiento de la violencia en el territorio nacional. (Instituto Nacional de Medicina Legal y Ciencias y Forenses, 2024)

- El Instituto de Estudios para el Desarrollo y la Paz – INDEPAZ, también aporta datos sobre la visibilización de los hechos violentos de Colombia a raíz del conflicto armado, sus informes nos permiten acercarnos a conocer las dinámicas del accionar de los grupos armados. Pero para el presente proyecto lo relevante es la contabilización de las víctimas que llevan, discriminadas por hecho delictivo y ubicación geográfica. El sitio web ofrece una sistematización no oficial de estadísticas y análisis de asesinatos de personas líderes y defensores de derechos humanos en Colombia 2016- 2024. La información suministrada respecto al observatorio de derechos humanos y conflictividades ofrecen elementos importantes para la unificación de las violencias en el proyecto a desarrollar (INDEPAZ, 2024)
- Las universidades privadas también contribuyen a conocer las dinámicas de la violencia en nuestro país mediante diferentes trabajos de investigación enfocados en solucionar problemáticas sociales. La Universidad Externado de Colombia, cuenta con un centro de análisis de datos – Delfos, que incluye temáticas sobre este fenómeno, así como con un dashboard con información actualizada a nivel departamental y municipal; presenta algunos indicadores como la tasa anual de homicidios, muertes por causas externas, suicidios y comparativos entre tasas de homicidios. La temporalidad de la información se desarrollada en las últimas 6 décadas. Para destacar que el centro de datos se alimenta de cuatro fuentes oficiales Nacionales: Ministerio de Salud, Medicina Legal, Policía Nacional y Dane, y dos a nivel internacional: Discover the World’s Health Data y World Health Organization (WHO). Adicionalmente el centro cuenta con una línea de investigación denominada análisis multi, fuente de violencia en Colombia. (Universidad Externado de Colombia, 2024).

Los anteriores son una muestra representativa de sitios web y aplicaciones que abordan el tema de la violencia desde los datos, los indicadores y las estadísticas. En cuanto a autores e investigaciones de diferentes disciplinas que han buscado dimensionar y explicar la violencia en sus múltiples facetas, que se convierten en un antecedente importante para el propósito del presente proyecto de grado, tenemos:

- El trabajo de investigación titulado “Metodología para el Análisis de la Violencia en el Departamento de Bolívar mediante Técnicas de Machine Learning”, que tiene como objetivo desarrollar una metodología para analizar la violencia en el departamento de Bolívar, específicamente en el ámbito de los homicidios, empleando técnicas de machine

learning. Se hace un análisis experimental de los datos disponibles sobre homicidios en ese departamento, identificando variables clave como el tipo de arma, género, ubicación y municipio. Se aplican técnicas de minería de datos para extraer conocimiento de la data, obteniendo resultados que revelaron diferencias significativas entre dos grupos de homicidios clasificados mediante un algoritmo de agrupamiento, resaltando la influencia de la zona y la edad de las víctimas. Se encontró que la mayoría de las víctimas eran hombres, predominantemente trabajadores particulares y solteros, con educación básica secundaria, y que el arma más utilizada fue el arma de fuego. Los hallazgos subrayan la utilidad del machine learning en el análisis de la violencia en Colombia, ofreciendo información crucial para la formulación de estrategias de prevención y reducción de homicidios, en decir de los autores. (Fernández Caraballo & Gómez Franco, 2018)

- Para los autores de “Predicción de las masacres en Colombia empleando inteligencia artificial”, después de la firma del acuerdo de la habana, hubo una disminución de la violencia en términos generales en todo el territorio nacional, puesto que la antigua ex FARC que incidía de manera violenta en estos espacios se retira de sus espacios. Pero después de 2019 se incrementan nuevamente los homicidios, homicidios colectivos y el reclutamiento armado, debido al ocupamiento territorial de nuevos actores armados ilegales. En este entorno aplican modelos predictivos de inteligencia artificial para intentar pronosticar posibles hechos delictivos. Con el uso de técnicas supervisadas de deep learning mediante algoritmos de bosques aleatorios y estimaciones de regresiones binomiales negativas que involucran variables de tasa de homicidios con armas de fuego, los homicidios con otro tipo de arma e incautaciones de cocaína, encuentran evidencia que cada vez que aumenta la tasa de homicidios y los cultivos de uso ilícito existe mayor probabilidad que exista una nueva masacre (Rojas Guerrero & Grautoff, 2022).
- En el estudio denominado “Método de clústering e inteligencia artificial para clasificar y proyectar delitos violentos en Colombia”, los autores de la investigación movidos con el propósito de construir una herramienta de prevención de delitos, construyen clústeres de los delitos violentos en Colombia por departamentos, junto con una estructura de redes neuronales para su clasificación y pronóstico. Para ello, se parte del análisis del método de clústering, la inteligencia artificial y la definición de delitos violentos. Como data del proyecto se usa los datos generados por la Policía Nacional sobre delitos entre 2018 y 2022. Como resultado, se establecieron cuatro clústeres de delitos y factores de violencia que caracterizan grupos de departamentos, lo que permitió identificar regiones con mayor y menor impacto de actos delictivos. Luego se planteó una red neuronal de doble capa

que alcanzó una capacidad de clasificación y predicción de los delitos según su tipo e impacto (Fontalvo Herrera , Vega Hernández , & Mejía , 2023).

- En la investigación titulada “Modelo de machine learning para la predicción de las tendencias de hurto en Colombia”, los investigadores de este trabajo partiendo de la evidencia de que el hurto es un delito significativo que ha tenido un gran impacto en la sociedad colombiana, generando una sensación generalizada de inseguridad debido a sus altas tasas en los últimos años, aplican un modelo de aprendizaje automático basado en Máquinas de Soporte Vectorial para la regresión (SRV), diseñado específicamente para predecir la tendencia de hurto en Colombia en sus tres principales ciudades. El modelo es evaluado utilizando un conjunto de datos extraídos del sistema de información de la Fiscalía Nacional de Colombia, que incluye 2,662,402 registros de delitos cometidos en el país desde 1960 hasta 2019. Los resultados se compararon con un modelo de regresión lineal estándar y un modelo SRV sin ajuste, mostrando, según los autores, resultados prometedores que podrían ser útiles para las autoridades competentes en la lucha contra el hurto. (Ordoñez , Cobos , & Bucheli , 2020)
- Finalmente, se hace alusión del estudio denominado “Comprendiendo la dinámica de los conflictos en América latina: una aproximación desde el Machine learning” que tiene como propósito descubrir tendencias en los conflictos de América Latina desde 1989 hasta el 2022. Se parte del principio de que el agrupamiento puede ofrecer una comprensión más amplia de las conexiones entre la política, la economía y los conflictos. Se asume que las variables en juego están interconectadas de manera compleja y aún no completamente entendida. La clústerización se emplea como herramienta para agrupar entidades en categorías con el fin de buscar explicaciones basadas en características compartidas entre los objetos analizados. De esta manera, se busca una comprensión más clara de las complejas relaciones entre variables como el desarrollo humano, la economía y los conflictos. Para el análisis, se utilizan datos del Programa de Conjuntos de Datos de Conflictos de Uppsala para categorizar los actores involucrados en los conflictos según una serie de criterios específicos. (Villar Roldán & Martín Álvarez , 2023).

2. Marco Conceptual

Violencia

Para la presente investigación, la violencia se define como cualquier acción en la que se usa la fuerza para ir en contra de la naturaleza de una persona, constreñir su voluntad, o transgredir lo que una sociedad considera justo según el derecho establecido (Jiménez García, Manzano Chávez, & Mohor Bellalta, 2021), un fenómeno inherente a la condición humana, que conlleva la imposición de la propia voluntad sobre la de otros, ya sea de manera consciente o inconsciente. Se manifiesta en diversos niveles, desde las interacciones cotidianas (micro sociales) hasta las estructuras sociales y políticas (macrosociales). Además del daño físico, abarca también el control y la manipulación como mecanismos de coerción (Vidal, Mejia González, & Curiel Gómez, 2021).

Machine Learning

Machine learning (ML) es una rama de la inteligencia artificial que permite a las computadoras aprender a partir de datos sin ser programadas explícitamente. A través de algoritmos, el sistema puede identificar patrones y realizar predicciones basadas en los datos de entrenamiento. Es un enfoque en el que las computadoras aprenden a mejorar su desempeño en tareas específicas a partir de datos, en lugar de seguir instrucciones programadas de manera fija. El proceso comienza con un conjunto de datos de entrenamiento, compuesto por entradas y salidas correctas, que el sistema analiza mediante algoritmos de aprendizaje supervisados, no supervisados o de refuerzo. A través de este análisis, se construye un modelo matemático que captura patrones en los datos, permitiendo al sistema hacer predicciones sobre nuevos datos no vistos previamente (Géron, 2023)

Técnicas no supervisadas

Las técnicas no supervisadas de machine learning son métodos que permiten a las máquinas descubrir patrones ocultos y estructuras inherentes en datos sin etiquetar. A diferencia del aprendizaje supervisado, donde se proporcionan ejemplos correctos, aquí los algoritmos exploran los datos por sí mismos para identificar similitudes, diferencias y relaciones entre los puntos de datos. Esto es especialmente útil para tareas como la segmentación de clientes, la detección de anomalías y la reducción de la dimensionalidad de los datos, permitiendo obtener insights valiosos y tomar decisiones más informadas (Bobadilla, 2021).

Clústering

El clústering es una técnica fundamental en el aprendizaje no supervisado que busca agrupar datos en conjuntos homogéneos, llamados clústeres. Estos grupos se forman de manera que los puntos de datos dentro de un mismo clúster sean lo más similares posible entre sí, mientras que los puntos de diferentes clústeres sean lo más distintos posible. La similitud se determina en función de características o atributos específicos de los datos, y puede medirse utilizando diversas métricas de distancia. Esta técnica es ampliamente utilizada en diversas áreas, como la segmentación de clientes, la detección de anomalías y la organización de grandes conjuntos de datos (Arán Godés, 2022).

Detección de Anomalías

La detección de anomalías es un proceso para identificar datos que se desvían de los patrones esperados. Estas anomalías suelen representar eventos inusuales o raros. En el análisis de la violencia, pueden indicar incidentes aislados o brotes de violencia que no siguen los patrones típicos observados en el resto del conjunto de datos (Chandola, Banerjee & Kumar, 2009).

PCA (Análisis de Componentes Principales)

El análisis de componentes principales es una técnica de reducción de dimensionalidad que transforma los datos originales en un nuevo conjunto de variables, llamadas componentes principales, que son combinaciones lineales de las variables originales. Esta técnica permite simplificar la representación de los datos, facilitando su visualización y análisis sin perder información relevante. Los componentes principales se ordenan de forma decreciente según la cantidad de varianza que explican, lo que permite seleccionar un subconjunto de componentes que capture la mayor parte de la variabilidad original, reduciendo así la dimensionalidad de los datos. Esta técnica es ampliamente utilizada en diversas áreas, como la minería de datos, el reconocimiento de patrones y el procesamiento de imágenes (Díaz Ramírez, 2021).

PostGIS

PostGIS es una extensión de código abierto que transforma a PostgreSQL en un poderoso sistema de gestión de bases de datos espaciales. Al incorporar tipos de datos geométricos, índices espaciales y funciones avanzadas para el análisis geográfico, PostGIS permite almacenar, consultar y manipular de manera eficiente grandes volúmenes de información georreferenciada. Esta herramienta es ampliamente utilizada en diversas aplicaciones, como sistemas de información geográfica (SIG), análisis de ubicación, gestión de infraestructura y desarrollo de aplicaciones web map. Gracias a su integración con PostgreSQL, PostGIS ofrece una solución robusta y escalable para el manejo de datos espaciales, combinando la flexibilidad de una base de datos relacional con las capacidades específicas de un SIG (Hsu & Obe, 2021).

ETL

Acrónimo de Extracción, Transformación y Carga, es un proceso de integración de datos que implica la extracción de información de diversas fuentes, a menudo heterogéneas, su transformación para garantizar la coherencia, calidad y compatibilidad con el sistema de destino, y finalmente, la carga de los datos transformados en un repositorio virtual de datos. Este proceso es fundamental para consolidar información dispersa en un único repositorio, permitiendo análisis más profundos y toma de decisiones basadas en datos (Inmon, 2022).

Georreferenciación

La georreferenciación es el proceso de asignar coordenadas geográficas precisas (latitud y longitud) a entidades o eventos, vinculándolos a una ubicación específica en la superficie terrestre. Esta técnica fundamental permite transformar datos abstractos en información geográficamente contextualizada, facilitando la visualización, análisis y modelización de fenómenos espaciales. Al asignar coordenadas geográficas, se establece una relación espacial entre los datos y el mundo real, lo que permite realizar análisis espaciales como la medición de distancias, la creación de mapas temáticos y la identificación de patrones espaciales (Buitrago Gómez, 2022).

Variables Socioeconómicas

Las variables socioeconómicas son factores que describen la situación económica y social de una población. En el contexto de la violencia, indicadores como el nivel de ingresos, la tasa de desempleo, la desigualdad económica, el acceso a servicios básicos (educación, salud), la distribución de la riqueza y el grado de urbanización pueden influir significativamente en la dinámica de la violencia en una región. Por un lado, la pobreza y la desigualdad económica pueden generar tensiones sociales y aumentar la probabilidad de conflictos violentos. Por otro lado, la falta de oportunidades económicas y educativas puede llevar a la desintegración social y al reclutamiento en grupos violentos (Vidal, Mejia González, & Curiel Gómez, 2021).

Metodología

La presente investigación se desarrolla bajo un enfoque cuantitativo, empleando técnicas de análisis de datos no supervisadas y supervisadas con el fin de analizar la violencia en el Departamento del Cauca durante el periodo 2019-2023. La metodología se estructura en varias fases que se detallan a continuación.

1. Tipo de estudio

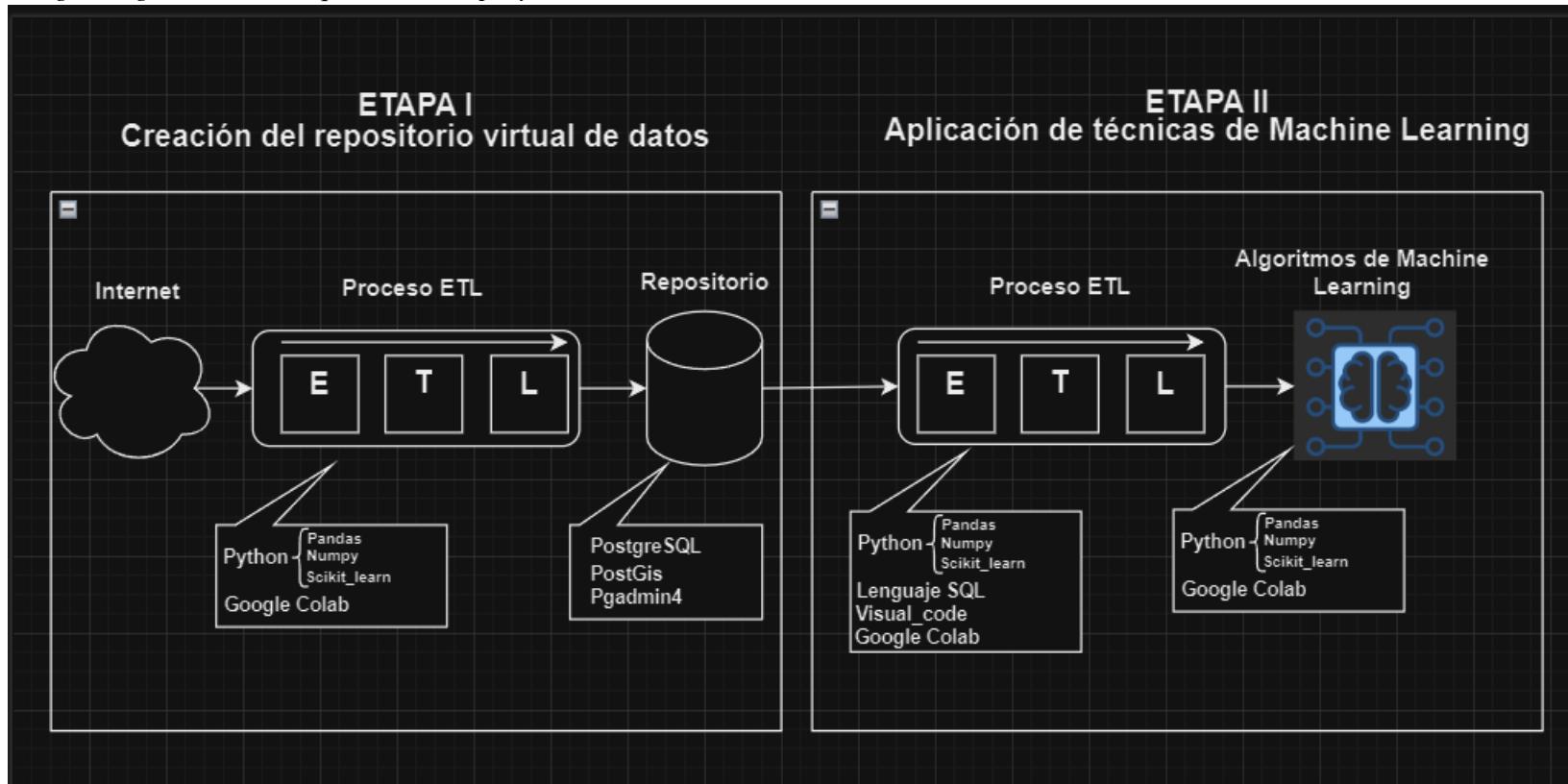
Este proyecto se encuadra dentro de un estudio descriptivo y explicativo, ya que tiene como objetivo principal identificar patrones de violencia en el Departamento del Cauca (2019-2023), integrando información socioeconómica y hechos de diferentes violencias y aplicando técnicas no supervisadas y supervisadas de Machine Learning. El enfoque descriptivo es apropiado porque se busca caracterizar y representar los datos sobre eventos violentos, mientras que el enfoque explicativo se justifica por la necesidad de entender los factores subyacentes y correlaciones entre la violencia y variables contextuales, como las condiciones socioeconómicas y geográficas.

El estudio busca explicar patrones no evidentes en los datos, proporcionando una visión integral del fenómeno de la violencia en el Departamento del Cauca.

2. Método o diseño de investigación

Este proyecto adopta un método empírico y estadístico con un enfoque cuantitativo. El análisis de los datos será realizado a través de técnicas estadísticas avanzadas y algoritmos de Machine Learning, prioritariamente no supervisados, lo cual es adecuado para el tratamiento de grandes volúmenes de datos y para descubrir patrones subyacentes sin la necesidad de hipótesis preconcebidas.

El proceso metodológico se estructura en dos etapas bien diferenciadas que, juntas, garantizan la consecución de los objetivos propuestos y conforman lo que podríamos llamar la arquitectura del proyecto. La primera etapa está orientada a la construcción de un repositorio de datos virtual, robusto y estructurado, mientras que la segunda se enfoca en la aplicación de técnicas avanzadas de Machine Learning para la extracción de patrones significativos. A continuación, se describe cada una de estas etapas.

Figura 2*Diagrama general de la arquitectura del proyecto**Fuente: construcción propia*

2.1 Etapa 1: Creación del repositorio virtual de datos

La primera etapa se centra en el diseño e implementación de una base de datos relacional que contiene información relevante sobre hechos de violencia y variables socioeconómicas en Colombia. Este repositorio será la fuente principal de datos para el análisis con técnicas de Machine Learning.

Diseño de la base de datos

La base de datos está diseñada siguiendo un modelo relacional en PostgreSQL, con soporte geoespacial mediante la extensión PostGIS. Tiene una arquitectura relacional y flexible, orientada tanto al análisis geoespacial como temporal. Su estructura permite realizar consultas precisas sobre fenómenos violentos y su relación con el entorno social y económico a través de una integración robusta de múltiples tablas interrelacionadas.

La arquitectura de la base de datos se construye en torno a una estructura central organizada en torno a dos tablas clave, que garantizan la trazabilidad de los datos y la capacidad de georreferenciación a nivel de municipio, departamento o región de la data almacenada. Estas tablas son esenciales para conectar y contextualizar los hechos de violencia y las variables socioeconómicas:

- **Tabla data_source:** Esta tabla centraliza las fuentes de datos, permitiendo identificar el origen y las características de cada conjunto de información registrado. Todas las demás tablas están relacionadas con data_source, lo que facilita el rastreo del origen exacto de cada dato y asegura la integridad referencial. Esta relación es clave para mantener la coherencia y confiabilidad de los datos a lo largo del tiempo.
- **Tabla municipalities:** Esta tabla contiene información detallada sobre las divisiones geográficas a nivel municipal en Colombia. Proporciona una base sólida para realizar análisis geoespaciales y ofrece datos de geometría común (punto, polígono, línea) que permiten la georreferenciación de eventos y variables socioeconómicas. La relación de esta tabla con todas las demás tablas de la base de datos a excepción de data_source, se realiza a través del campo donde se guarda el código que el Departamento Administrativo Nacional de Estadística – Dane, asigna a cada municipio del País, que es un identificador único. Esto permite agregar datos a diferentes niveles geográficos y realizar análisis espacialmente coherentes.

La otra parte de la arquitectura de la base de datos la conforman las tablas relacionadas que contienen los datos de eventos de violencia y las variables socioeconómicas y están diseñadas para interactuar con las tablas clave `data_source` y `municipalities`, mediante claves foráneas, lo que garantiza que los datos puedan contextualizarse tanto por su ubicación geográfica como por su fuente de origen. Las tablas de eventos o hechos de violencia incluyen detalles de los hechos ocurridos, al nivel de detalle encontrado en la fuente primaria de los datos y las tablas de variables socioeconómicas almacenan información relevante como el nivel de pobreza, tasas de empleo, educación, entre otros indicadores, que se pueden relacionar directamente con los municipios y las fuentes de datos.

Además, se crea una tabla adicional para almacenar los datos de los departamentos, lo que permite ampliar el análisis a nivel departamental y regional. Esta tabla contiene el nombre del departamento y su código Dane, y se relaciona con `municipalities`, facilitando la agregación de datos no solo a nivel municipal, sino también a nivel departamental y regional.

Todas estas tablas de datos mantienen relaciones bien definidas con `data_source` mediante la clave foránea `source_id`, que garantiza el rastreo de la fuente de los datos y con `municipalities`, mediante la clave foránea `dept_mpio_code` que contiene el código Dane del municipio, lo que asegura la ubicación geográfica precisa de cada registro.

Proceso ETL (Extracción, Transformación y Carga)

Para la **Extracción** de los datasets se cargan directamente desde la URL del endpoint de la fuente mediante pandas en el entorno de Google Colab, lo cual elimina el paso intermedio de descarga manual al equipo, facilita el acceso a los datos y permite que en un futuro se automatice la actualización de la base de datos con nuevos datos o versiones de los datasets. Solo se integran datasets que cumplan con los requisitos de estar relacionados con hechos de violencia y variables socioeconómicas de Colombia, ser de una fuente confiable, que se puedan georreferenciar a nivel de municipio, y que cubren al menos el periodo 2019-2023.

Para la **Transformación**, una vez cargados, los datos son sometidos a un proceso de limpieza y transformación utilizando pandas, numpy y scikit-learn. Entre las principales tareas de transformación se incluyen tratamiento de valores nulos, estandarización de variables categóricas, codificación y discretización. Un aspecto clave del proceso de transformación es asegurar la georreferenciación de los datos a nivel municipal. En los casos en que los datasets ya cuentan con códigos de municipios, estos son conciliados para garantizar

consistencia con los códigos oficiales. Para aquellos datasets que no incluyen dicha columna, se creó utilizando técnicas como la coincidencia difusa para emparejar los nombres de los departamentos y municipios con sus respectivos códigos.

Una vez finalizado el proceso de transformación, finalmente se hace la **carga** a la base de datos. El dataset transformado se exporta como un archivo CSV desde el entorno de Google Colab al Google Drive, y luego se descarga al entorno local donde se gestiona la base de datos. Para cargar los datos en PostgreSQL, primero se actualiza la tabla data_source para registrar la fuente del nuevo dataset, posteriormente, se crea una nueva tabla en la base de datos, cuidando que los campos coincidan con los del archivo CSV correspondiente y finalmente, se importan los datos directamente a la base de datos mediante la herramienta pgAdmin4.

Este proceso asegura la integridad y calidad de los datos antes de su inclusión en la base de datos relacional, y facilita su actualización futura de manera más eficiente.

Tecnologías utilizadas

Para la base de datos se seleccionó PostgreSQL como sistema de gestión de bases de datos (DBMS) por su robustez en el manejo de datos geoespaciales a través de la extensión PostGIS, su flexibilidad para soportar diversos tipos de datos, su capacidad para escalar en proyectos de gran envergadura, por ser de código abierto y gratuito y por el respaldo de una comunidad activa que garantiza su constante mejora y soporte (Pilicita Garrido & Borja López, 2020).

En el proceso ETL se utiliza Python y sus librerías especializadas para la manipulación de datos: pandas, numpy y scikit-learn.

Como entorno de desarrollo se utiliza a Google Colaboratory para ejecutar los scripts de Python y gestionar los procesos ETL y para la administración de la base de datos, se utiliza pgAdmin4.

2.2 Etapa 2: Aplicación de técnicas de Machine Learning

En esta segunda etapa, los datos del departamento del Cauca extraídos del repositorio virtual creado en la primera etapa, se someten a un análisis mediante técnicas no supervisadas y

supervisadas de machine learning. El objetivo es descubrir patrones, agrupamientos y anomalías que permitan entender mejor la relación entre las variables socioeconómicas y los hechos de violencia en los municipios del Departamento del Cauca. Esta etapa culminará con un informe detallado que presentará los hallazgos y resultados del análisis, aportando nuevo conocimiento sobre la problemática de la violencia en la región.

Proceso ETL Aplicado al repositorio virtual

El análisis de datos comienza con un proceso ETL (Extracción, Transformación y Carga) que permitirá extraer los microdatos almacenados en la base de datos relacional PostgreSQL creada en la primera etapa del proyecto. A diferencia de la primera fase, donde el objetivo era explorar fuentes externas, en esta etapa las consultas se realizarán directamente sobre el repositorio ya consolidado, utilizando consultas SQL optimizadas.

El objetivo principal de esta etapa es extraer solamente los datos que van a ser utilizados en los modelos de Machine Learning. Para ello, se ejecutarán consultas que filtren y seleccionen datos asociados a los municipios del Departamento del Cauca, durante el periodo de análisis 2019-2023, asegurando que solo se extraigan las variables y períodos de tiempo que sean relevantes para el análisis.

Como en principio se está trabajando con una base de datos alojada localmente, la extracción de los datos se hace ejecutando scripts en Python. El flujo aplicado en esta etapa de extracción es: conexión a la base de datos desde VSCode usando la librería psycopg2 o SQLAlchemy, para ejecutar las consultas SQL desde scripts en Python para extraer los datos necesarios; los resultados de las consultas SQL se salvan en formato CSV, lo que permite una fácil transferencia y acceso desde otros entornos y finalmente se carga el archivo CSV a Google Drive para facilitar el acceso a los datos desde Google Colab.

Si bien el anterior proceso de extracción es un poco artesanal, es porque la base de datos está alojada localmente, en un futuro cuando se aloje este repositorio en un servicio de almacenamiento en la nube ya se puede pensar en aplicar métodos más expeditos y automatizados.

Los datos extraídos del repositorio una vez subidos al entorno de Google Colab, serán sometidos a un nuevo y exhaustivo proceso de preprocesamiento y transformación para

asegurar la calidad y coherencia de los mismos y prepararlos al formato que requieran los algoritmos de Machine Learning a aplicar.

En esta fase incluye las siguientes tareas imputación de valores faltantes mediante técnicas estadísticas como imputación por la media o la mediana o métodos más avanzados, para completar los valores nulos, asegurando que no se pierda información valiosa durante el análisis; normalización/estandarización de variables para ajustar las escalas de las variables numéricas para garantizar el correcto funcionamiento e interpretación por parte de los algoritmos de Machine Learning; aplicación de un proceso de ingeniería de características para seleccionar las variables más relevantes al fenómeno de la violencia en el Cauca. En este paso, se pueden crear nuevas variables derivadas de las originales, con el fin de capturar relaciones o patrones ocultos. Esta transformación permitirá obtener vistas minables del conjunto de datos, optimizando su uso en los modelos de Machine Learning.

Aplicación de Técnicas de Machine Learning

Una vez preprocesados los datos, se aplicarán técnicas de aprendizaje no supervisado, como detección de anomalías mediante Isolation Forest, PCA y clústering k-means, sobre las variables relacionadas con la violencia. Esto permitirá identificar grupos de municipios con patrones de violencia similares. Posteriormente, se utilizará este agrupamiento como etiquetas para las variables socioeconómicas de los municipios, con el objetivo de modelar mediante regresión logística multinomial la relación entre las condiciones socioeconómicas y los distintos perfiles de violencia. A través de este análisis, se espera identificar las variables socioeconómicas más influyentes en cada perfil de violencia y desarrollar modelos para estimar la probabilidad de que un municipio pertenezca a un determinado grupo de riesgo. Las principales técnicas que se prevé, sin descartar otras técnicas cuya necesidad vaya apareciendo en el proceso, son:

- **Análisis de Componentes Principales (PCA):** Para reducir la dimensionalidad del conjunto de datos, permitiendo identificar las variables socioeconómicas más influyentes que puedan explicar los hechos de violencia. Con el PCA, se busca simplificar el análisis sin perder información relevante, facilitando la interpretación de los resultados. (McKinney, 2022)
- **Clústering:** Se aplicarán algoritmos de clústering, como K-means para identificar agrupamientos de municipios que compartan características similares, tanto en términos

socioeconómicos como en sus niveles de violencia. Este análisis permitirá detectar grupos homogéneos de municipios y evaluar si estos grupos presentan comportamientos particulares respecto a la violencia.

- **Detección de Anomalías:** A través de la técnicas de Isolation Forest, se identifican municipios que se desvén de los patrones o promedios generales (McKinney, 2022). Estos municipios, considerados como anomalías, pueden representar zonas donde los niveles de violencia o las condiciones socioeconómicas se comportan de manera atípica, lo cual es fundamental para un análisis más detallado de la problemática de la violencia.
- **Regresión logística:** Esta técnica de análisis estadístico se utilizada para modelar la probabilidad de que una observación pertenezca a una de dos o más categorías discretas. En otras palabras, es una herramienta que nos permite clasificar datos en grupos específicos (Arán Godés, 2022).

Validación y métricas de evaluación

Para garantizar la fiabilidad y validez de los modelos de Machine Learning aplicados, se implementa un proceso de validación y medición del desempeño. Este proceso permitirá evaluar la capacidad de generalización de los modelos, asegurando que los patrones descubiertos no se deben a particularidades de los datos de entrenamiento, sino que pueden generalizarse a situaciones similares en otros contextos (Bobadilla, 2021).

Además, se utilizarán métricas de evaluación específicas para cada técnica implementada. En el caso del clústering, se aplicarán métricas como la *silhouette score* y la inercia. Para la detección de anomalías, se evaluará el rendimiento mediante medidas de precisión, sensibilidad y *F1-score*, buscando obtener un equilibrio entre la detección de verdaderas anomalías y la minimización de falsos positivos. (Arán Godés, 2022)

Tecnologías utilizadas

La segunda etapa del proyecto se llevará a cabo utilizando un conjunto de herramientas tecnológicas que garantizan la eficiencia y escalabilidad del análisis de datos. Las principales tecnologías incluyen:

- Librerías de Python: Se emplearán librerías especializadas como pandas para la manipulación y transformación de los datos, numpy para realizar operaciones numéricas, y scikit-learn para la implementación de los algoritmos de Machine Learning no supervisado.
- PostgreSQL: La base de datos relacional diseñada en la primera etapa seguirá siendo la fuente principal de datos. Las consultas se realizarán mediante SQL para extraer los microdatos georreferenciales de los municipios del Cauca.
- Google Colaboratory: Este entorno permitirá ejecutar los scripts en Python para realizar todo el proceso de análisis, proporcionando un entorno flexible y accesible desde cualquier dispositivo.

3. Participantes o unidad de análisis

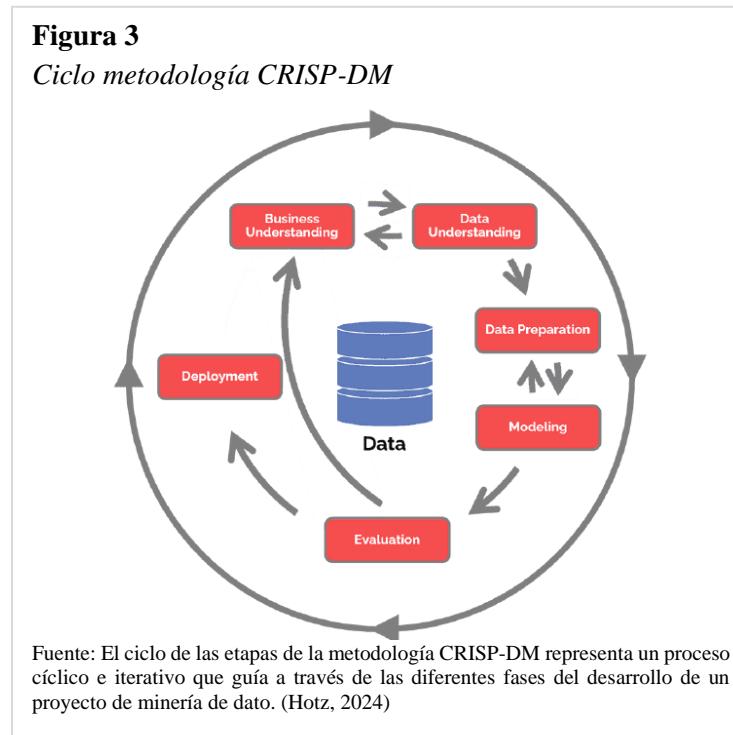
La unidad de análisis de este estudio son los municipios del Departamento del Cauca y los eventos violentos que ocurrieron entre 2019 y 2023. Se analizan diferentes aspectos de la violencia como tipo de evento, frecuencia, y localización geográfica, junto con variables socioeconómicas de los municipios, tales como niveles de pobreza, acceso a servicios básicos y presencia de cultivos ilícitos.

El análisis a nivel municipal permitirá identificar patrones específicos de violencia y su relación con factores contextuales, proporcionando un análisis que permita entender la dinámica de la violencia en la región. Además, aunque el enfoque principal está en el Cauca, los datos abarcan todos los municipios de Colombia, facilitando análisis futuros o ampliaciones del proyecto.

4. Plan de análisis de la información

El plan de análisis se desarrollará a partir de un proceso cílico y estructurado que permitirá organizar, categorizar, y transformar los datos en información valiosa para responder las preguntas del proyecto. Este proceso se adelanta apoyándose en la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining). Este enfoque estructurado y ampliamente utilizado proporciona un marco de trabajo común que servirá como guía a través de las diferentes etapas del proceso de análisis de datos propuesto (Sánchez Trujillo & Pérez Hernández, 2021).

Las seis etapas que propone la metodología CRISP-DM se disponen en un ciclo iterativo, que permite regresar a etapas anteriores en caso de ser necesario para refinar el análisis o abordar nuevos hallazgos, aumentando así las probabilidades de éxito (Pomares Quimbaya & Martínez Bernal, 2023). Las etapas son: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue.



A continuación, se detalla cómo encaja cada una de estas etapas en el contexto del presente proyecto:

Comprensión del negocio

Aunque esta etapa no corresponde a un objetivo específico del presente proyecto, se considera ejecutada con la estructuración del anteproyecto. Ese documento fue elaborado con el objetivo de definir y planificar el proyecto de análisis de datos de manera clara y concisa, ayudando a los involucrados a comprender los objetivos, el problema que se quiere resolver, el alcance, la metodología, los recursos necesarios, los resultados esperados, entre otros aspectos (Vizmanos, Mejía Pérez, & Fránquez Flores, 2022). Para su elaboración, se realizó un detenido análisis del problema de la violencia en el Departamento del Cauca, revisando

literatura y estudios académicos relevantes y explorando posibles fuentes de datos para el proyecto, logrando así el entendimiento del negocio.

Comprendión de los datos

El primer objetivo específico del presente proyecto es construir una base de datos integral que combine información socioeconómica de los municipios del País con datos sobre hechos de diferentes tipos de violencia, provenientes de fuentes oficiales y no oficiales, centralizando y unificando los datos actualmente dispersos. Esta fase incluye actividades como la identificación de las fuentes de datos relevantes, tanto oficiales como no oficiales, la descripción de los datos, la exploración de los mismos y la evaluación de su calidad, confiabilidad y pertinencia para el análisis (Espinosa Zúñiga, 2020).

Preparación de los datos

En esta fase, que también puede enmarcarse dentro del primer objetivo del proyecto la investigación, se seleccionan los datos, estableciendo la razón de inclusión o exclusión de los mismos. Se realiza la recopilación y el preprocesamiento de los datos, respetando las normas éticas y legales de protección de datos. Las actividades incluyen la limpieza, validación y transformación de la data para garantizar su calidad y consistencia (Velásquez, 2023). Posteriormente, se integran los datos en el repositorio virtual elegido para el proyecto con la granularidad adecuada. Además, se lleva a cabo la ingeniería de características, identificando las variables más relevantes y, si es necesario, transformando y creando nuevas características a partir de los datos brutos para mejorar el rendimiento de los modelos de machine learning (Holbrook, 2024). Se utiliza tecnologías como las librerías de pandas y geopandas de Python para estas actividades, dado que los set de datos generalmente están disponibles en archivos de texto delimitado, en archivos json, xml o en archivos geográficos como geojson o shapefiles (Bilogur, 2024).

Modelado

El segundo objetivo específico del proyecto es aplicar técnicas no supervisadas y supervisada de machine learning, como análisis de componentes principales, agrupamiento, detección de anomalías y regresión logística. En esta fase se explorarán en profundidad los datos utilizando estas técnicas de aprendizaje con el propósito de descubrir patrones, tendencias y relaciones significativas (Santana Falcón, 2023) entre los datos socioeconómicos y los hechos de violencia en cada municipio del Departamento del Cauca. Se utilizará Python como la principal herramienta tecnológica.

Evaluación

El tercer objetivo específico del proyecto es validar los resultados obtenidos utilizando técnicas de validación y métricas de evaluación para asegurar la robustez y coherencia de los hallazgos. En esta fase se evaluarán los resultados obtenidos del modelado de datos para determinar su relevancia y coherencia en el contexto de la problemática de la violencia en el Departamento del Cauca. Se utilizarán métricas de evaluación para medir el desempeño de los modelos.

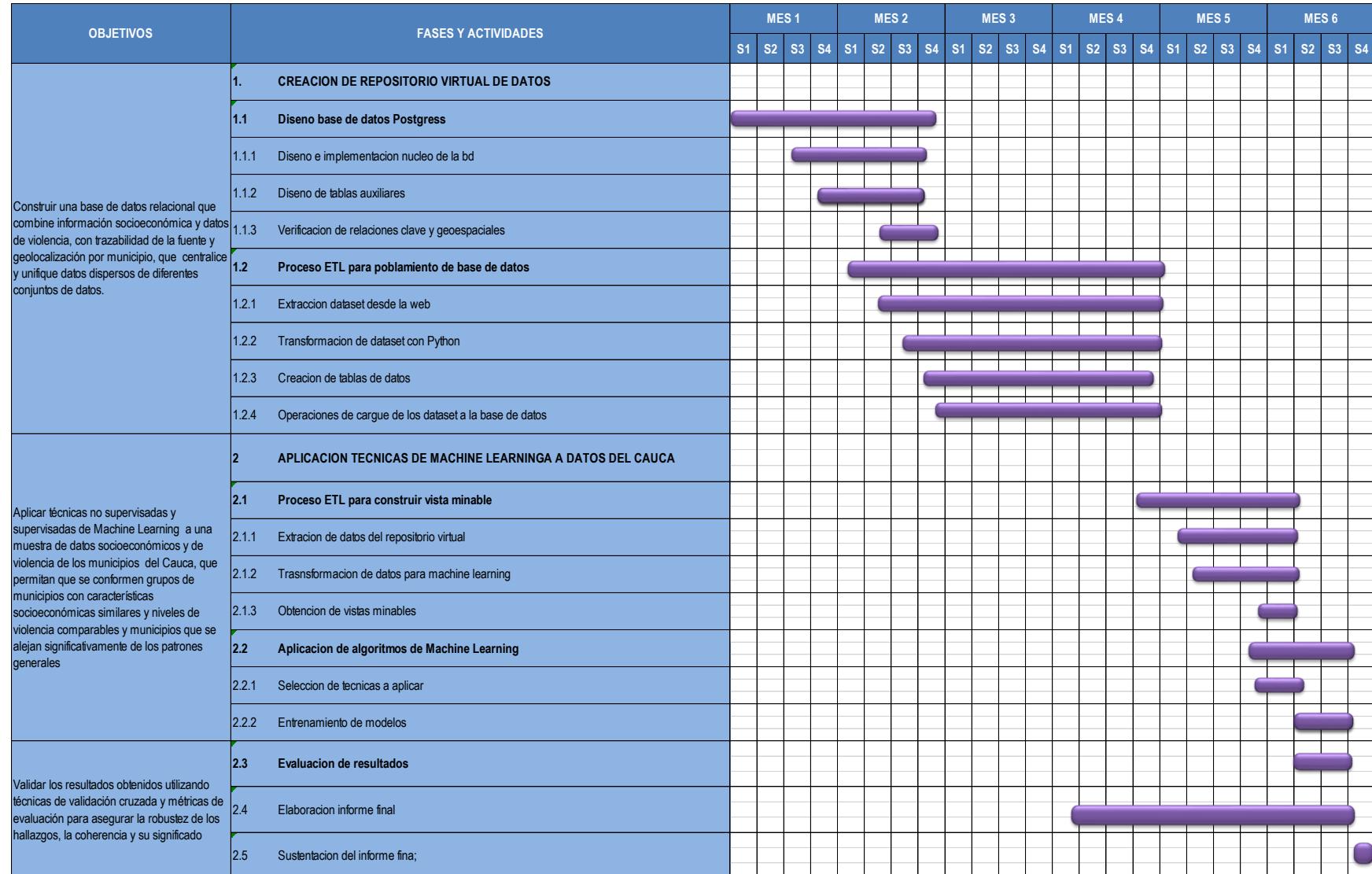
Despliegue

Aunque la fase de despliegue de la metodología CRISP-DM no se corresponde con ninguno de los objetivos específicos expresamente declarados en el presente proyecto, dado que el alcance del mismo no abarca la implementación en un entorno de producción específico, el informe final de resultados puede considerarse como parte de esa fase. Este informe final es un producto tangible que resume y comunica los hallazgos clave del análisis de datos, y puede considerarse como la implementación de los resultados del análisis en un formato útil y accesible para los usuarios finales o los tomadores de decisiones (IBM Corporation, 2021).

Dada la naturaleza cíclica e iterativa de la metodología CRISP-DM, adoptada para la ejecución del presente proyecto, se permite adaptar el proceso de análisis a medida que se avanza en el proyecto. Por ello, en la fase de evaluación se revisan los resultados y el proceso, determinando los pasos a seguir, lo que puede incluir regresar a etapas anteriores para refinar el análisis, contemplar nuevas combinaciones de características para la vista minable, abordar nuevos hallazgos, ajustar el enfoque en función de la información disponible, etc. Esto permite asegurar una mejor calidad de los resultados finales (Caio, 2022).

5. Cronograma y presupuesto

En la figura 4 y en la tabla 1 se presenta el cronograma y el presupuesto respectivamente del proyecto.

Figura 4*Cronograma general del proyecto*

Fuente: Construcción propia

Tabla 1*Presupuesto del proyecto*

NOMBRE DEL PROYECTO DE ANALÍTICA DE DATOS:	Violencia en el Cauca (2019-2024): Análisis mediante Técnicas no Supervisadas de Machine Learning					OBJETIVO GENERAL	Analizar la problemática de la violencia en el Departamento del Cauca durante el periodo 2019-2024 mediante el uso de técnicas no supervisadas de Machine Learning, que permitan identificar patrones entre las características socioeconómicas de la población y las manifestaciones de la violencia en sus diferentes tipos.					Fecha inicio: 11 jun 2024 Fecha final: 11 nov 2024 Vigencia: 6 meses								
Recursos Humanos																				
Obejivos Específicos	Etapas	Producto	Actividades	Item	Sub actividad	Director V/h 200.000	Estudiante V/h 40.000	Horas Director	Horas Estudiante	Horas al mes	Totales									
Construir una base de datos relacional que combine información socioeconómica y datos de violencia, con trazabilidad de la fuente y geolocalización por municipio, que centralice y unifique datos dispersos de diferentes conjuntos de datos.	Etapa1: Creación del repositorio virtual de datos	Una base de datos	Diseño de la base de datos relacional	1.1	Diseno	\$ 200,000	\$ 40,000	34	6	40	\$ 7,040,000									
			Proceso Extracción, Transformación y Carga	1.2	Montaje	\$ 200,000	\$ 40,000	36	4	40	\$ 7,360,000									
				1.3	Extracción	\$ 200,000	\$ 40,000	24	8	32	\$ 5,120,000									
				1.4	Transfromación	\$ 200,000	\$ 40,000	32	0	32	\$ 6,400,000									
				1.5	Carga	\$ 200,000	\$ 40,000	16	0	16	\$ 3,200,000									
Aplicar técnicas no supervisadas y supervisadas de Machine Learning a una muestra de datos socioeconómicos y de violencia de los municipios del Cauca, que permitan que se conformen grupos de municipios con características socioeconómicas similares y niveles de violencia comparables y municipios que se alejan significativamente de los patrones generales	Etapa 2: Aplicación de técnicas de Maching Learnig	Informe escrito	Proceso ETL para M.L.	2.1	ETL repositorio virtual	\$ 200,000	\$ 40,000	24	0	24	\$ 4,800,000									
			Aplicación de técnicas de Maching Learnig	2.2	PCA	\$ 200,000	\$ 40,000	22	2	24	\$ 4,480,000									
				2.3	Clustering	\$ 200,000	\$ 40,000	32		32	\$ 6,400,000									
				2.4	Detención de anomalias	\$ 200,000	\$ 40,000	48	0	48	\$ 9,600,000									
Validar los resultados obtenidos utilizando técnicas de validación cruzada y métricas de evaluación para asegurar la robustez de los hallazgos, la coherencia y su significado		Informe y socialización resultados	Evaluacion resultados	2.6	Validación y métricas de evaluación	\$ 200,000	\$ 40,000	16	0	16	\$ 3,200,000									
			Informe Final	2.7	Informe final	\$ 200,000	\$ 40,000	12	4	63	\$ 2,560,000									
					Total			296	24	367	\$ 60,160,000									
Recursos Físicos																				
						Descripción	Valor Unidad	Cantidad	Total											
						Computadores	2,000,000	2	\$ 4,000,000.00											
						Insumos eléctricos Valor K/H		1,050	\$ 336,000.00											
						Varios	1,000,000	-	\$ 1,000,000.00											
						Totales			\$ 5,336,000.00											
Descripción																				
Total																				
Recursos Humanos																				
\$ 60,160,000.00																				
Recursos Físicos																				
\$ 5,336,000.00																				
Total proyecto																				
\$ 65,496,000.00																				

Fuente: construcción propia

Procedimiento o descripción del trabajo de campo

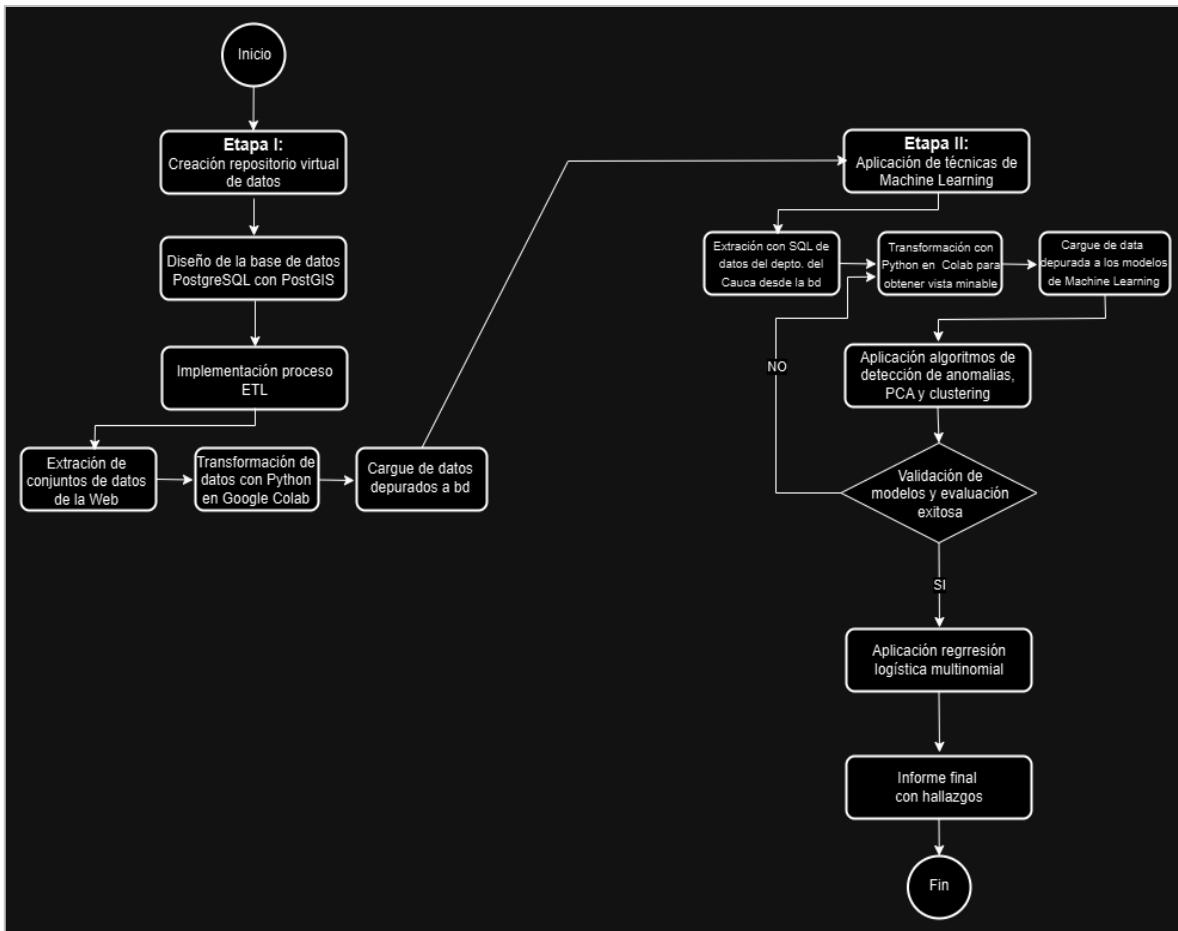
El trabajo de campo está dividido en dos grandes etapas, cada una con objetivos específicos y actividades detalladas que permiten avanzar de manera ordenada hacia la consecución de los objetivos del proyecto.

1. Diagrama de flujo general de las actividades de campo del proyecto

El flujo a alto nivel del proyecto se puede visualizar en el diagrama de flujo general que se presenta a continuación:

Figura 5

Diagrama general de flujo de la ejecución del proyecto



Fuente: Construcción propia

La etapa I, creación del repositorio virtual de datos, comienza con el diseño de la base de datos en el sistema de gestión de bases de datos PostgreSQL con la extensión PostGIS. En esta fase inicial, se define y se implementa una base de datos relacional, la cual incluye soporte para datos geoespaciales a través de la extensión PostGIS. La base de datos está diseñada para almacenar grandes cantidades de datos de eventos de violencia y variables socioeconómicas, relacionados geográficamente con municipios de Colombia. Se implementa una arquitectura que permite la trazabilidad de las fuentes de datos y la posibilidad de realizar consultas geoespaciales.

La siguiente fase para la creación del repositorio de datos es la implementación de un primer proceso ETL. Durante la etapa de extracción, los datos son obtenidos desde diversas fuentes de la web, principalmente en formato CSV o JSON, utilizando el lenguaje Python y la librería pandas.

En la fase de transformación de datos, los datasets extraídos son procesados para garantizar su calidad. Se aplican las técnicas de limpieza y transformación habituales, para el tratamiento de valores nulos, estandarización, codificación, discretización, etc. (Mckinner, 2022). Es importante en esta fase la conciliación de los códigos de municipios establecidos por el Dane. Luego de ser transformados, los datos depurados se exportan a archivos CSV y se cargan a la base de datos PostgreSQL, en la tabla previamente creada.

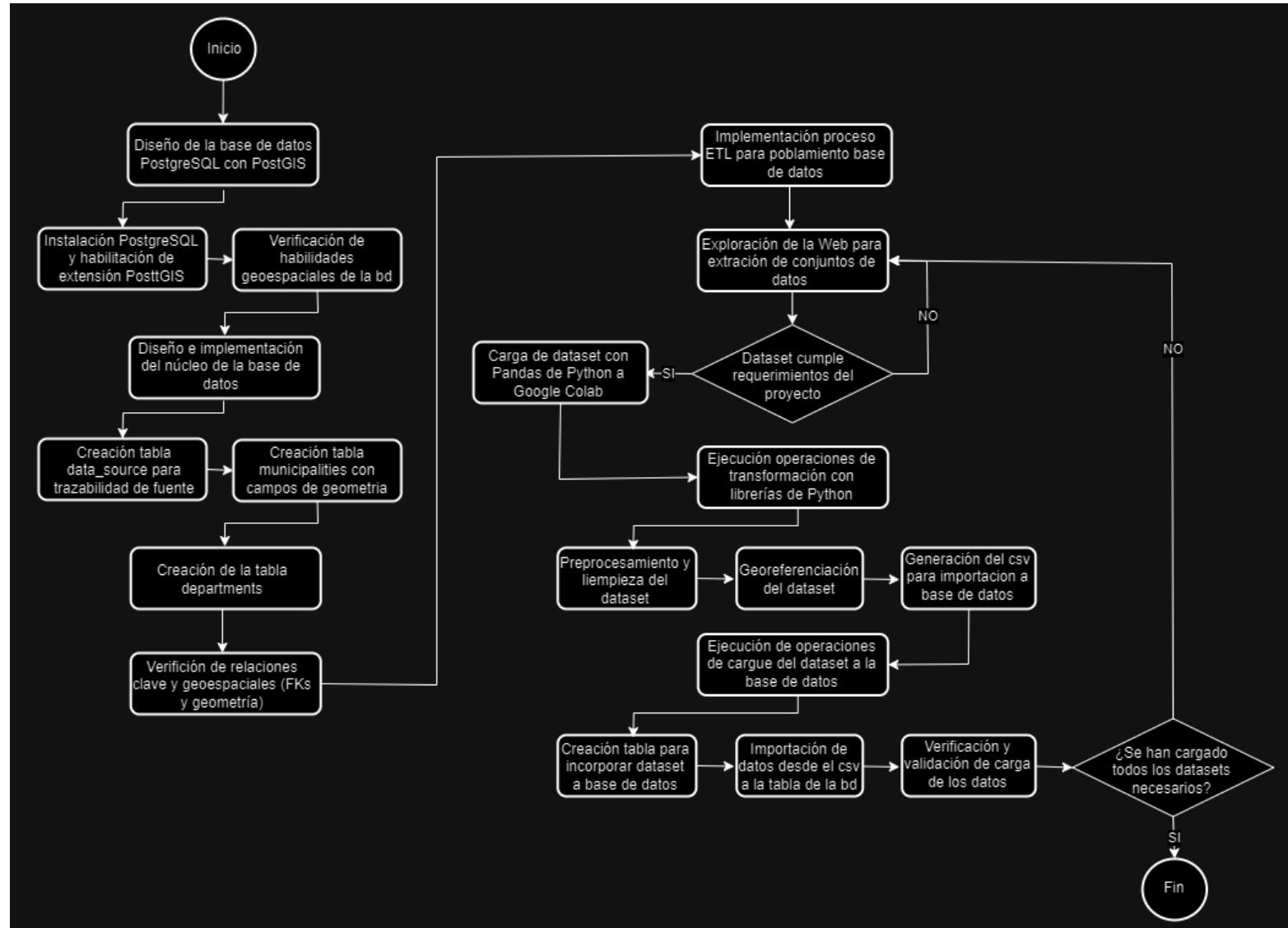
En la Etapa II, se inicia la aplicación de técnicas de Machine Learning. Se extraen datos georreferenciados de la base de datos creada en la Etapa I mediante consultas SQL optimizadas, relacionados específicamente con el departamento del Cauca y correspondientes al periodo 2019-2023. Estos datos son procesados en Google Colaboratory, donde se realizan tareas de limpieza, transformación y selección de características relevantes. Finalmente, se aplican técnicas de clústering, reducción de dimensionalidad (PCA), detección de anomalías y regresión logística multinomial (Arán Godés, 2022).

2. Diagrama de flujo detallado de las actividades de campo – Etapa 1

La figura de la siguiente página presenta el diagrama de flujo detallado de la primera etapa del proyecto, dando una comprensión más profunda y estructurada que la presentada en el diagrama de flujo general de la figura inmediatamente anterior sobre el procedimiento y descripción del trabajo de campo.

Figura 6

Diagrama de flujo detallado de la etapa I del proyecto



Fuente: Construcción propia

Esta primera etapa involucra desde el diseño de la base de datos en PostgreSQL con soporte geoespacial mediante PostGIS, hasta la implementación de un proceso ETL robusto para su poblamiento con datos relevantes y correctamente transformados.

En la Etapa I del proyecto, que se considera crítica para el éxito del proyecto, se destaca los siguientes pasos:

2.1 Diseño de la base de datos Postgres con PostGIS

Instalación PostgreSQL y habilitación de extensión PostGIS

El trabajo de campo de la etapa de diseño de la base de datos PostgreSQL con PostGIS inicia con la instalación del sistema de gestión de bases de datos PostgreSQL, la habilitación de la extensión PostGIS, y la instalación de la herramienta Pgadmin4, con el propósito de preparar un entorno de base de datos relacional con soporte para datos geoespaciales. PostgreSQL es la base de datos principal, PostGIS añade funciones para trabajar con datos espaciales, permitiendo que la base de datos maneje datos geoespaciales, como coordenadas y geometrías, esenciales para el análisis geográfico en este proyecto, y PgAdmin4 proporciona una interfaz gráfica visual para PostgreSQL (Rigg, 2022).

Verificación de habilidades geoespaciales de la base de datos

Para garantizar que las funciones de PostGIS estén correctamente habilitadas y operativas se verifica las capacidades geoespaciales de la base de datos. Esto se logra mediante la creación de una tabla de prueba que contenga un campo de tipo geométrico, seguida de la inserción de datos de ejemplo y la ejecución de consultas SQL especializadas. Estas consultas permiten realizar operaciones como cálculos de distancias, selección geoespacial y visualización de mapas, asegurando que todas las funcionalidades geoespaciales estén plenamente activas y funcionando correctamente (Rigg, 2022).

Diseño e implementación del núcleo de la base de datos

El núcleo de la base de datos se refiere al conjunto de tablas y relaciones que forman la base estructural y funcional más importante del modelo de datos. Son las entidades centrales que sostienen la integridad, lógica y cohesión de la base de datos. Estas tablas son fundamentales porque almacenan los datos esenciales, sobre los cuales otras tablas y funcionalidades dependen directa o indirectamente (Juba & Volkov, 2021).

En el caso del presente proyecto el núcleo central compuesto por dos tablas clave:

- **data_source:** Esta tabla tiene como objetivo principal garantizar la trazabilidad de los datos. Almacena información detallada sobre la fuente original de cada conjunto de datos, como la entidad que lo generó, la fecha de obtención, el periodo que abarca los datos, la url del dataset y la url del endpoint, una descripción del contenido, entre otros. Esta trazabilidad es fundamental para asegurar la calidad y confiabilidad de los datos y la replicabilidad de los análisis.

Saber de dónde provienen los datos es crucial en cualquier investigación, especialmente cuando se integran datos de diferentes fuentes, ya que permite garantizar la calidad, reproducibilidad y validación de los análisis. Al ser una tabla clave para mantener el control y trazabilidad de los datos, facilita la auditoría, el mantenimiento y la validación de la información. Asegura que cualquier investigación futura basada en esta base de datos pueda referenciar correctamente las fuentes de los datos, lo que es esencial para la transparencia y la replicabilidad de los resultados. Los datos almacenados en esta tabla vienen a constituirse en metadatos sobre los conjuntos de datos almacenados en las otras tablas, lo que contribuye a la integridad del modelo de datos.

- **municipalities_dane:** Esta tabla es esencial para la geolocalización precisa de los datos a nivel municipal. Forma parte del núcleo de la base de datos, porque se utiliza para representar las divisiones político-administrativas de Colombia, a nivel de municipios, incluyendo campos de geometría (polígonos para representar la delimitación geográfica de los municipios y puntos para representar la cabecera de cada municipio), que posibilitan análisis geoespaciales y consultas basadas en la localización.

Al ser una base de datos con soporte geoespacial (PostGIS), la tabla municipalities_dane es indispensable para realizar consultas, análisis y visualizaciones de los municipios en un mapa, permitiendo estudiar la distribución espacial de ciertos fenómenos. El código DANE de los municipios de Colombia es clave para relacionar la tabla municipalities_dane con otras tablas. Al asociar los datos a esta tabla con las tablas que contienen los datos sobre violencia y fenómenos socioeconómicos, se garantiza que cada registro esté vinculado a un lugar específico en el territorio.

Otra de las razones del porque la tabla data_source y municipalities_dane conforman el núcleo de la base de datos, es el hecho de que todas las demás tablas de la base de datos están

directamente relacionadas con estas a través de claves foráneas (FK), lo que las convierte en un punto de referencia común para todas las demás entidades en la base de datos.

Creación de la tabla departments

Además del núcleo central, se ha incluido la tabla **departments**. Aunque no forma parte integral del núcleo, esta tabla desempeña un papel importante al facilitar el análisis geográfico a nivel departamental y regional. Al agrupar los municipios por departamento, se pueden realizar análisis más amplios y comparativos entre diferentes regiones.

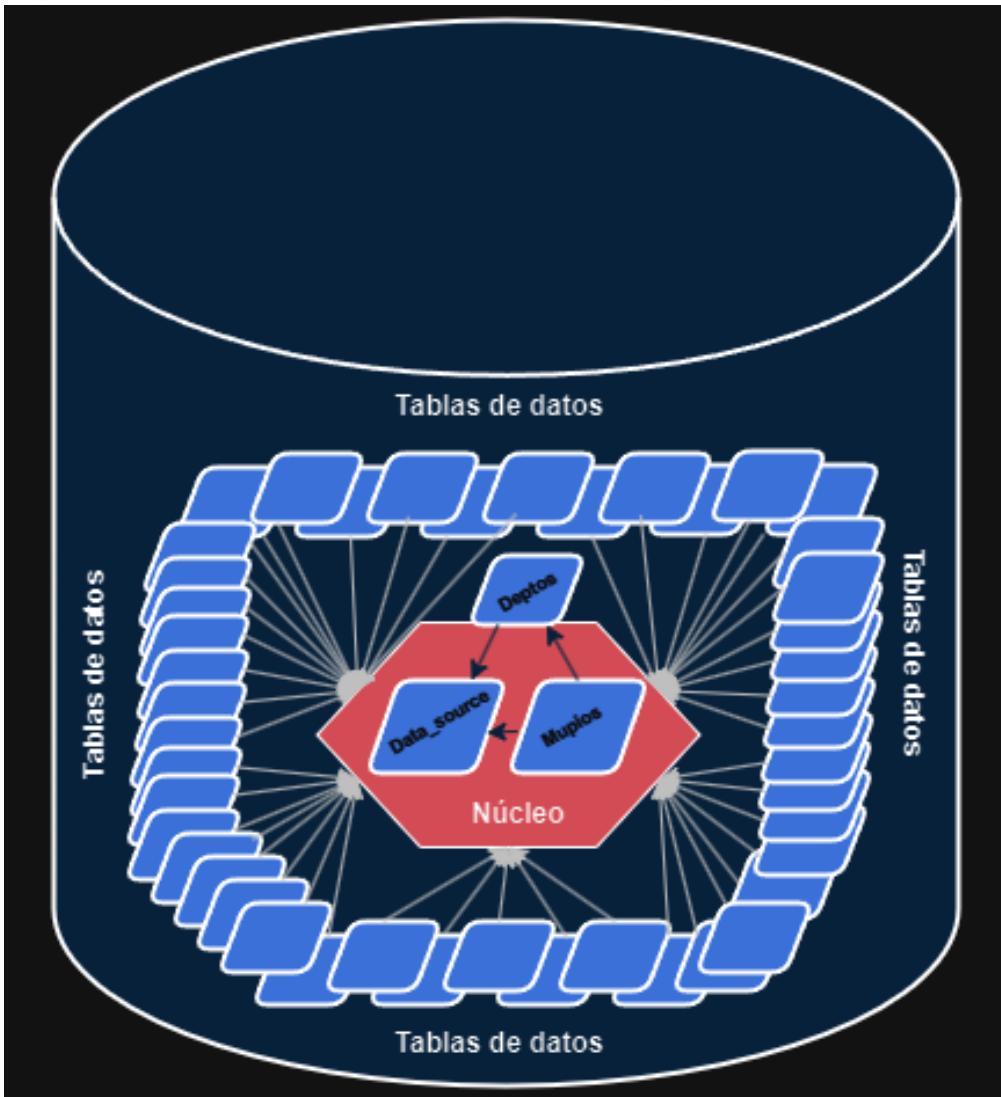
Su propósito es almacenar la información de los departamentos de Colombia (nombre y código dane), permitiendo su respectiva georreferenciación a través de la relación establecida con la tabla *municipalities_dane* a través de claves foráneas, permitiendo mantener relaciones jerárquicas entre departamentos y municipios. Para asociar un municipio a un departamento, se agrega un campo *dept_code* en la tabla *municipalities_dane* que actúa como clave foránea, lo que garantiza que cada municipio esté correctamente asociado a un departamento, y las geometrías almacenadas en la tabla *municipalities_dane* puedan ser usadas para análisis geoespaciales a nivel departamental o regional, si se agregan varios departamentos.

Las tablas que contienen datos específicos sobre hechos de violencia y datos socioeconómicos se encuentran organizadas de manera modular alrededor del núcleo central. Esta estructura modular permite gestionar y analizar cada conjunto de datos de forma independiente, facilitando la flexibilidad y la escalabilidad del sistema (Juba & Volkov, 2021). A pesar de su independencia, todas las tablas de datos están vinculadas al núcleo central a través de relaciones con las tablas *data_source* y *municipalities_dane*, lo que garantiza que todos los datos almacenados en la base de datos cuenten con trazabilidad (gracias a la relación con la tabla *data_source*, se puede rastrear el origen de cada dato, lo que es fundamental para evaluar su calidad y confiabilidad), y geolocalización, garantizada con la vinculación a la tabla *municipalities_dane*, que georreferencia los datos a nivel municipal, permitiendo realizar análisis espaciales precisos y visualizar los resultados en mapas.

El diseño de la base de datos anteriormente descrito se esquematiza en la figura de la siguiente página donde se muestra el diseño en los términos que se acaba de exponer.

Figura 7

Representación de la estructura de la base de datos del proyecto



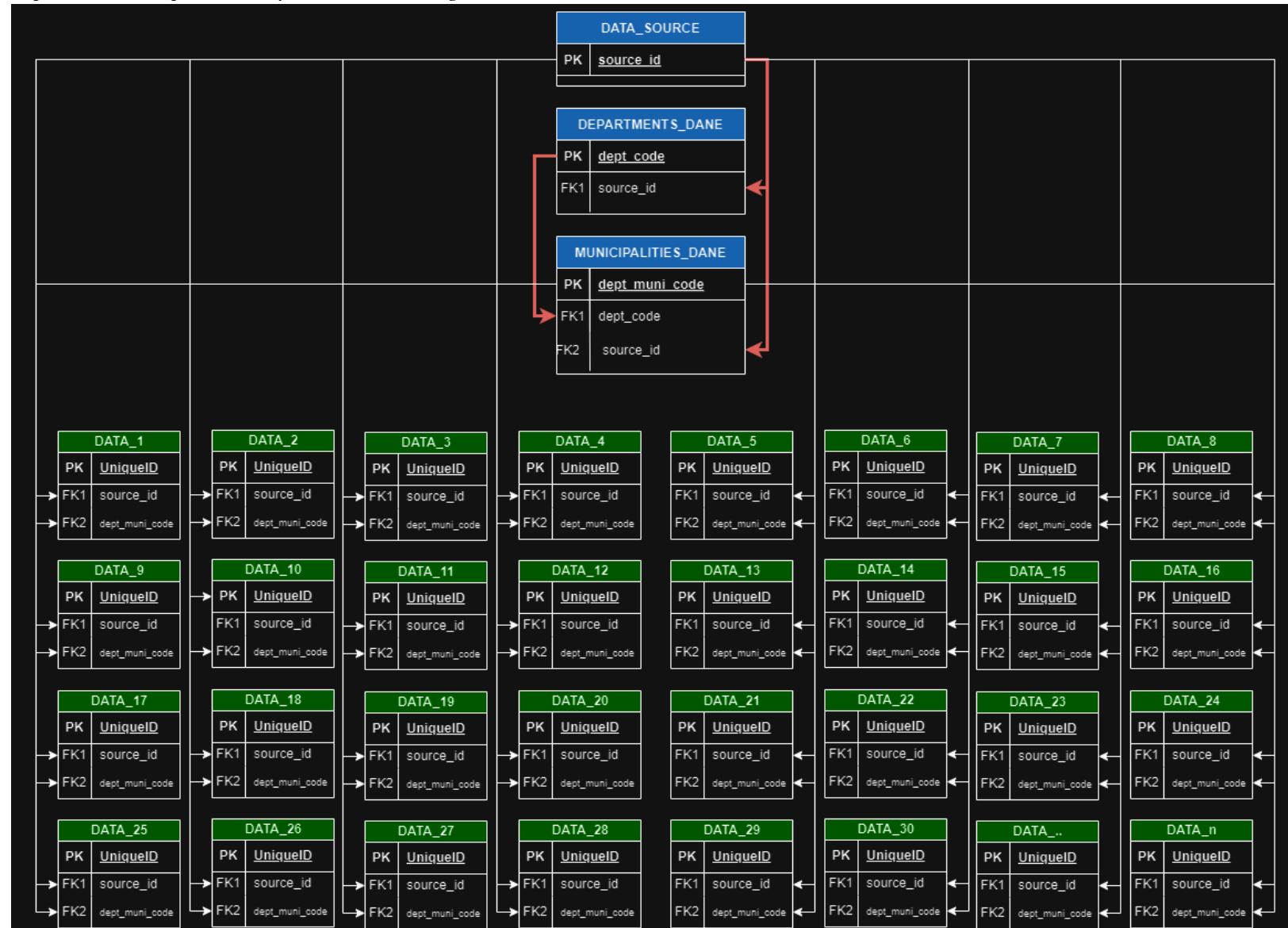
Fuente: Construcción propia

De esta forma, el diseño asegura que cualquier dato almacenado no solo esté asociado a su fuente, sino que también esté correctamente ubicado geoespacialmente, permitiendo realizar análisis a nivel territorial con precisión y consistencia.

Una aproximación al diagrama entidad-relación del diseño de la base de datos que se acaba de explicar se ve en la figura de la página siguiente.

Figura 8

Representación aproximada y reducida del diagrama Entidad-Relación



Fuente: Construcción propia

Cabe destacar que la implementación de la base de datos en este proyecto trasciende el objetivo inmediato de proporcionar un repositorio para el análisis de la violencia en el Cauca entre 2019 y 2023. Su propósito es convertirse en un activo valioso y duradero para la comunidad académica y de investigación, pues ha sido diseñada y está poblada con datos que abarcan no solo los municipios del Cauca, sino todos los municipios de Colombia, además, incluye datos que se remontan más allá de 2019, llegando a 2010, 2003 e incluso a años anteriores del siglo XX.

Este enfoque amplio y exhaustivo no solo permite un análisis detallado y profundo de la situación de violencia en el Cauca, sino que ofrece una plataforma rica en datos históricos y geográficos que podrá ser utilizada en futuros estudios de diferentes ámbitos. Los investigadores tendrán a su disposición un conjunto de datos georreferenciados y trazables, lo que facilita estudios comparativos a lo largo del tiempo, análisis multitemporales, y evaluaciones sobre diversos fenómenos sociales y económicos sobre todo el país.

Así, la base de datos no solo será el principal pilar de nuestro proyecto, sino que también representará un legado duradero para la investigación académica. Al constituirse como un recurso reutilizable, contribuirá a fortalecer las capacidades analíticas de futuras investigaciones, amplificando el impacto de nuestro trabajo más allá del análisis actual sobre la violencia en el Cauca.

Verificación de relaciones clave y geoespaciales (FKs y geometría)

En esta parte del trabajo de campo para la implementación de la base de datos se busca validar la correcta creación de relaciones entre tablas y la integridad de los datos geoespaciales. A través de la ejecución de consultas SQL se comprueba si los municipios están correctamente asignados a un departamento, se verifica que las geometrías de municipios y departamentos estén bien definidas para garantizar que los municipios estén dentro de los departamentos correspondientes. Este paso permite asegurar que los datos geoespaciales no solo están correctamente cargados, sino que las relaciones espaciales sean coherentes (Rigg, 2022). Esto prepara la base de datos para el proceso ETL que sigue a continuación.

2.2 Implementación del proceso ETL para poblamiento de la base de datos

Este primer proceso ETL (Extract, Transform, Load) es crítica en el presente proyecto, ya que permite extraer, transformar y cargar datos en la base de datos de PostgreSQL. Este proceso se lleva a cabo en Python utilizando especialmente, herramientas como pandas y

geopandas, que facilitan la manipulación y transformación de datos, y Google Colaboratory, como plataforma para ejecutar los script.

Exploración de la web para extracción de conjuntos de datos

El propósito de esta parte del trabajo de campo es encontrar fuentes de datos confiables y relevantes para poblar las tablas de la base de datos. Los datasets candidatos a integrarse a la base de datos, deben cumplir con unas condiciones que buscan asegurar la calidad, relevancia y trazabilidad de los datos, tales como:

- Deben estar alineados con los objetivos del proyecto (hechos de violencia y factores socioeconómicos en Colombia), abarcar toda área geográfica de Colombia y tener la cobertura temporal adecuada, como mínimo del 2019 en adelante. A excepción de los set de datos que manejen reportes "con corte a" sin especificar cuándo ocurrió el hecho que dio origen al dato.
- Los datasets deben estar bien estructurados, en formatos estándar (CSV, Excel, JSON), con identificadores consistentes.
- Los conjuntos de datos deben incluir campos que permitan establecer una relación clara con la tabla municipalities_dane, facilitando así su integración con los datos de geometría contenidos en esta última, así la información se puede geolocalizar.
- Los datos deben ser completos, exactos, consistentes a lo largo del tiempo, accesibles y provenientes de fuentes confiables.
- Deben tener un nivel de detalle suficiente, evitando datos agregados que no se puedan desagregar a nivel municipal.

Si se encuentra que el dataset no cumple con los criterios anteriores, se desecha y se regresa a explorar la Web hasta encontrar el siguiente candidato. Si por el contrario el conjunto de datos cumple con los criterios, se selecciona como candidato y se procede al siguiente paso del diagrama de flujo: la carga del dataset al entorno de desarrollo.

Carga de los dataset con Pandas de Python a Google Colab

El propósito de esta parte del trabajo de campo es importar el dataset desde la fuente externa hacia un entorno de Google Colab, utilizando Pandas para cargar y analizar el contenido del archivo. Se prefiere la carga directa desde la URL del endpoint en lugar de la descarga del archivo al entorno local y posterior carga en Colab, ya sea que los datos estén en un formato directo, como CSV o JSON, lo que permite una carga directa y fácil, utilizando las funciones `read_` de pandas, o incluso en el caso de que los datos estén comprimidos (en formato ZIP, TAR, o GZ), donde se hace necesario implementar una descarga más técnica, utilizando funciones adicionales que descompriman el archivo y carguen el contenido correctamente (McMahon, 2022).

La carga desde el endpoint tiene como ventajas que simplifica el flujo de trabajo y reduce el tiempo de procesamiento; ahorra espacio del disco al evitar la descarga al entorno local, lo que es especialmente importante cuando se manejan grandes volúmenes de datos: garantiza que siempre se trabaja con la versión más reciente del dataset, algo que es particularmente útil para datos dinámicos o en constante cambio, ya que se omiten pasos intermedios donde los datos podrían volverse obsoletos (VanderPlas, 2022).

Ejecución de operaciones de transformación con librerías de Python

En esta etapa se realizan las transformaciones necesarias para que los datos sean útiles y estén listos para ser cargados en la base de datos. Se hace el preprocessamiento y limpieza del dataset para garantizar la calidad de los datos antes de la carga en la base de datos. Las tareas habituales de preprocessamiento incluyen el manejo de valores faltantes, para identificar y tratar los datos nulos o vacíos, mediante técnicas como imputación de valores (medias, medianas, modas) o eliminación de registros incompletos; la normalización y estandarización para ajustar los datos a rangos comunes o escalas específicas para asegurar su consistencia; se detectan y remueven registros duplicados y redundantes; se hace la validación y ajuste de los tipos de datos para que coincidan con los requisitos del análisis y la estructura de la base de datos y, se eliminan columnas que se considera innecesarias para la naturaleza de la base de datos, entre otros (VanderPlas, 2022).

Un paso importante en esta fase es lograr la georreferenciación del dataset. Dado que en el núcleo de la base de datos se incorpora la tabla `municipalities_dane`, que contiene datos de geolocalización (geometría de polígonos y puntos) y el código oficial del DANE para cada municipio, es esencial que los datasets importados puedan vincularse correctamente a esta información geoespacial. Para ello si el dataset incluye una columna que contiene los códigos con que se identifican los municipios, se realiza una conciliación con la columna que contiene

el código DANE en la tabla municipalities_dane para garantizar que los registros estén asociados correctamente con su ubicación geográfica oficial. En caso de que el dataset no cuente con ese identificador único del municipio, se genera una nueva columna para almacenar el código DANE, utilizando técnicas avanzadas como la similitud difusa (fuzzy matching), que permiten vincular registros basados en nombres similares, en este caso, se utiliza la combinación del nombre de los municipios y departamentos, para una lograr una mejor coincidencia (McKinney, 2022).

Tras realizar las transformaciones necesarias (preprocesamiento, limpieza y georreferenciación), el dataset se guarda en un archivo CSV, que es un formato eficiente y compatible para la carga masiva de datos en el repositorio virtual de PostgreSQL.

Ejecución de operación de cargue de cada dataset a la base de datos

Una vez procesado y validado, el archivo CSV se descarga de Google Drive. A continuación, se crea una tabla en la base de datos PostgreSQL con una estructura que coincide exactamente con las columnas y tipos de datos del archivo CSV. Utilizando pgAdmin 4, se ejecuta una instrucción SQL de copia para cargar los datos del CSV directamente a la tabla recién creada. Este proceso garantiza una carga eficiente y evita posibles errores de mapeo de datos. Finalmente, se realizan consultas SQL para verificar la integridad y consistencia de los datos cargados, asegurando que se hayan transferido correctamente a la base de datos.

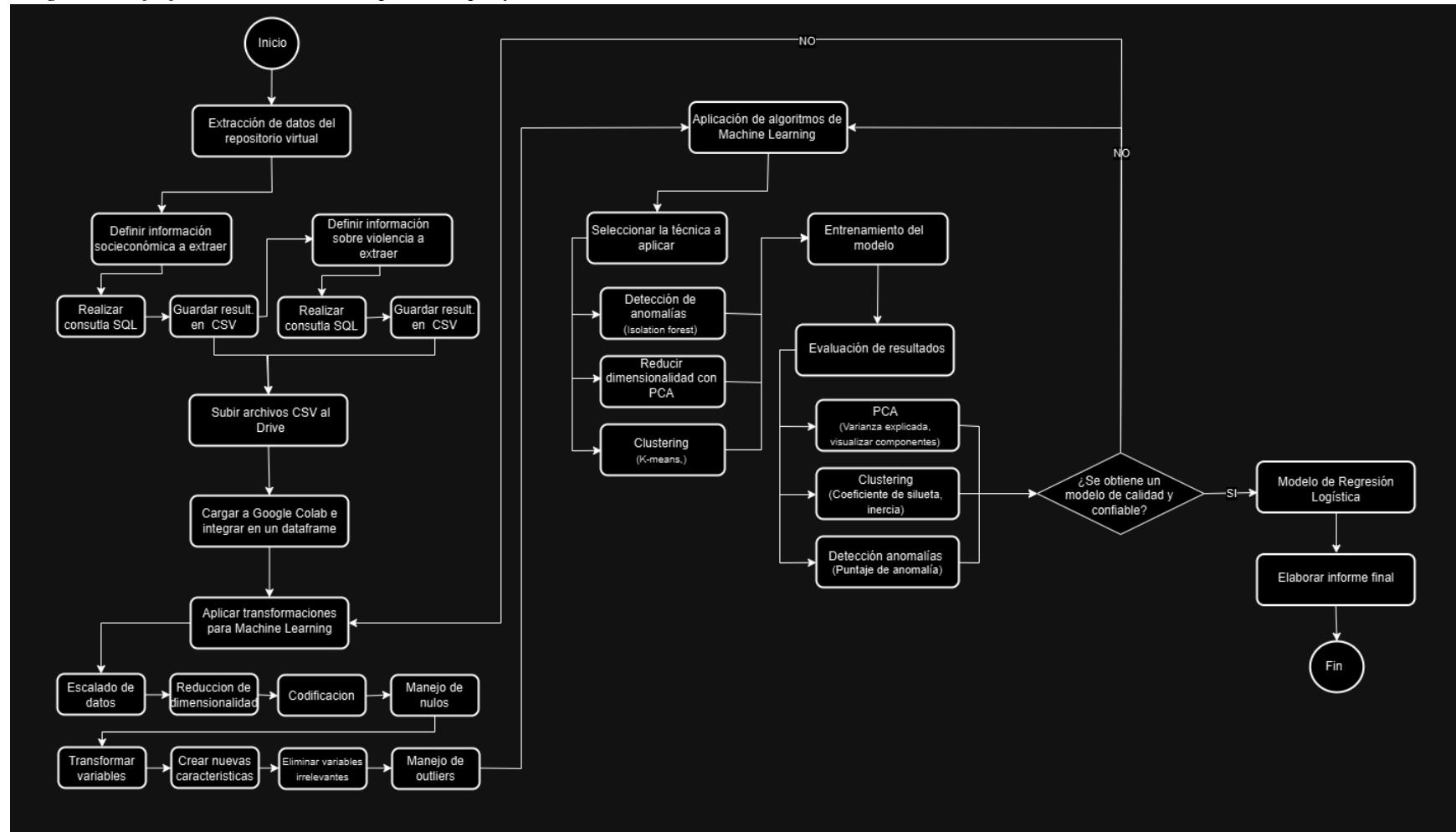
El proceso ETL de esta primera etapa del proyecto se repite cíclicamente, actualizando y enriqueciendo continuamente la base de datos. Una vez se alcanza un nivel de completitud satisfactorio, el proceso entra en un estado de mantenimiento, pero permanece listo para reactivarse en cualquier momento. La arquitectura del sistema está diseñada para incorporar nuevos datasets de manera ágil, asegurando así la evolución y escalabilidad de la base de datos a lo largo del tiempo (Juba & Volkov, 2021).

3. Diagrama de flujo detallado de las actividades de campo – Etapa 2

La siguiente figura presenta el diagrama de flujo detallado de la segunda etapa del proyecto para una comprensión más profunda y estructurada que la presentada en el diagrama de flujo general de la figura 5.

Figura 9

Diagrama de flujo detallado de la etapa II del proyecto



Fuente: Construcción propia

Al igual que en la Etapa I, en la Etapa II del proyecto también se implementa un proceso ETL, pero este se adapta específicamente a las necesidades del análisis que se va a implementar. Los datos, ya almacenados y georreferenciados en la base de datos, se extraen mediante consultas SQL optimizadas, enfocándose en el departamento del Cauca y el periodo 2019-2023. Posteriormente, los datos se procesan en Google Colaboratory, donde se realizan tareas adicionales de limpieza, transformación y selección de características clave, antes de aplicar técnicas como clustering, PCA y detección de anomalías (Grus, 2023).

A continuación, se describe en detalle el procedimiento o trabajo de campo llevado a cabo en cada una de las actividades principales del diagrama de flujo anterior.

3.1 Extracción de datos del repositorio virtual

El propósito de esta etapa del procedimiento es extraer de manera eficiente y organizada la información relevante de la base de datos para construir un dataset adecuado sobre el departamento del Cauca que será sometido a técnicas de machine learning. Dado que la base de datos construida como parte del presente proyecto cuenta con una gran complejidad y tamaño (66 tablas, más de 128 millones de registros, 128 datasets gestionados), el proceso de extracción debe ser cuidadosamente planificado para evitar errores, optimizar tiempos de procesamiento y garantizar la coherencia de la información obtenida.

Enfoque escalonado en la extracción de datos

Debido a la dimensión y complejidad de la base de datos creada, no es recomendable intentar extraer toda la información en una única consulta SQL, pues el diseño de una única consulta que integre todas las tablas y atributos requeridos implicaría realizar múltiples uniones (JOINs), subconsultas y agregaciones, lo que aumentaría significativamente la complejidad de la sintaxis y el riesgo de errores.

A medida que las consultas SQL se vuelven más complejas, optimizarlas para obtener tiempos de respuesta razonables se convierte en un reto, ya que las bases de datos relacionales pueden enfrentar problemas de rendimiento cuando intentan procesar grandes volúmenes de datos en consultas altamente complejas. La probabilidad de introducir errores en la consulta o de generar resultados incoherentes aumenta con la complejidad (Rigg, 2022). Incluso si se logra construir una consulta técnicamente correcta, la cantidad de operaciones necesarias para procesar decenas de millones de registros provenientes de varias tablas podría resultar en tiempos de procesamiento inaceptablemente largos.

Debido a lo anterior se prefiere adoptar un enfoque escalonado en el que las consultas SQL se dividan en partes manejables y específicas. Este enfoque permite extraer datos en secciones temáticas y relacionadas, simplificando las consultas y haciendo que la gestión de errores y la optimización sean más eficientes.

Una estrategia adecuada para la extracción de información desde el repositorio es clasificar los datos por categorías temáticas, comenzando por los datos socioeconómicos. Estos incluyen indicadores clave como la población, los niveles de pobreza, la afiliación a salud, indicadores educativos, coberturas de servicios públicos, etc, los cuales se agrupan por municipio y año, o incluso extraer la información de cada tabla en particular. Una vez organizados, se exportan a archivos CSV, ya sea de manera conjunta o en segmentos específicos, dependiendo de la cantidad y complejidad de los datos.

El mismo proceso se puede aplicar a los datos relacionados con la violencia, que abarcan información sobre delitos de todo tipo, desplazamientos forzados, víctimas y otros eventos violentos. Al igual que con los indicadores socioeconómicos, estos datos se agrupan por municipio y año, y se exportan en archivos CSV para mantener un orden lógico y fácil de manejar.

Además, los datos que garantizan la georreferenciación, como las geometrías de polígonos y puntos, pueden extraerse en un archivo CSV separado. Estos datos espaciales son esenciales para realizar análisis geográficos o visualizaciones en mapas.

Es importante que en todas las consultas SQL se incluyan las columnas que relacionan todas las tablas de la base de datos, siendo los códigos DANE de los municipios el vínculo principal. Esto garantiza la coherencia y facilita la futura integración de la información obtenida en las diferentes consultas dentro de un único dataset consolidado.

Cada consulta ejecutada genera un archivo CSV que representa una parte específica del dataset final, permitiendo trabajar de manera organizada y controlada, con la flexibilidad de procesar, revisar o ajustar los datos según sea necesario. Así no solo se garantiza una extracción precisa y coherente de la información, sino que también se simplifica su posterior análisis e integración.

Agrupación de los datos por municipio y año

La estrategia en esta parte es agrupar los datos que se van extraer por cada uno de los 42 municipios del departamento del Cauca y por año de vigencia de la información, lo que

permitirá analizar tendencias tanto espaciales (entre municipios) como temporales (entre años).

Este enfoque tiene la ventaja de darle claridad y simplicidad a la data, porque se asegura que cada fila represente una unidad territorial clara y al agregar información de varios años, se puede analizar tanto las diferencias entre municipios como la evolución de estos indicadores a lo largo del tiempo. Además al tener los datos organizados por municipio y año, se puede aplicar fácilmente técnicas de Machine Learning para detectar patrones a nivel territorial o temporal.

Tecnologías en la extracción de datos

La extracción de los datos desde el repositorio virtual se hará mediante lenguaje SQL, esencial para interactuar con la base de datos en PostgreSQL y realizar las consultas necesarias que permitan extraer los datos relevantes para el análisis.

SQL permite agrupar y consolidar datos por municipio y año utilizando funciones como GROUP BY, JOIN, SUM, y COUNT; filtrar los datos con WHERE y otras cláusulas, lo que facilita la extracción de los datos relevantes desde las múltiples tablas relacionadas con información socioeconómica, violencia y georreferenciación, para luego procesarlos y consolidarlos en un único DataFrame (Juba & Volkov, 2021).

Aunque las consultas podrían realizarse directamente en la base de datos por medio de la herramienta pgadmin4, las consultas SQL se ejecutarán con Python mediante la librería psycopg2 desde el entorno de en VSCode, dado que la base de datos está montada en local. A través de scripts de Python, las consultas SQL se pueden automatizar en lugar de ejecutarse manualmente consulta por consulta en una interfaz gráfica como pgAdmin4, lo que ahorra tiempo y minimiza errores. Psycopg2 se conecta directamente con PostgreSQL y permite ejecutar consultas SQL desde Python, lo que permite que los datos extraídos sean inmediatamente procesados en el mismo entorno. Python permite manejar grandes volúmenes de datos de manera más eficiente (Mckinner, 2022) y una vez ejecutadas las consultas, los datos se exportan directamente a archivos CSV, lo que facilita la posterior carga y análisis en Google Colab.

3.2 Subir los archivos a Google Drive y cargarlos al entorno de Google Colab

Los archivos CSV generados a partir de las diferentes consultas SQL se suben a Google Drive para facilitar su acceso y manejo desde Google Colab. Esta estrategia permite trabajar con los datos en la nube, evitando la dependencia de los recursos locales y garantizando que los

datos estén organizados y accesibles para el equipo de trabajo. Google Drive se integra de manera eficiente con Google Colab, lo que facilita la carga directa de los archivos en los notebooks de Python.

En Google Colab, los archivos CSV se integrarán en un único DataFrame utilizando Pandas, la biblioteca principal para la manipulación de datos en Python. Esto permite leer y cargar cada archivo de forma eficiente, incluso cuando se trabaja con grandes volúmenes de datos. Además, Pandas ofrece la capacidad de unir o combinar los archivos, ya sea de manera vertical u horizontal, consolidando toda la información en un solo DataFrame (Mckinner, 2022). Los diferentes archivos CSV se combinan utilizando como clave común el código DANE del municipio, lo que facilita realizar las transformaciones adicionales necesarias para preparar los datos con miras al análisis de machine learning.

3.3 Aplicar transformaciones para machine learning

El objetivo principal de esta etapa es preparar los datos para que puedan ser utilizados eficazmente por los algoritmos de machine learning no supervisados (como clústering, PCA y detección de anomalías) y no supervisados (Regresión logística). Estas transformaciones ayudan a garantizar que los datos estén en una forma adecuada, eliminando sesgos y estructuras innecesarias que podrían interferir en la identificación de patrones ocultos (Géron, 2023).

Cada una de las transformaciones que a continuación se especifican son clave para garantizar que los algoritmos de machine learning puedan analizar los datos de manera efectiva. La preparación adecuada de los datos asegura que los algoritmos de machine learning puedan detectar patrones valiosos y generar resultados útiles.

Escalado de datos

El escalado de datos es el proceso de ajustar los valores de las características para que estén dentro de un rango común (generalmente entre 0 y 1, o con media 0 y desviación estándar 1). Esto es necesario porque muchos algoritmos de machine learning son sensibles a las magnitudes de los datos (Bobadilla, 2021).

En PCA una de las técnicas que se tiene previsto aplicar, el escalado es crítico porque este algoritmo maximiza la varianza entre las características, y sin escalado, las variables con mayor magnitud dominarían el análisis. Pero también el escalado se hace necesario en clústering como K-Means, para evitar que las características con diferentes escalas influyan

más en la formación de los clústeres y en la detección de anomalías, para que las anomalías se detecten comparativamente en todas las dimensiones (Géron, 2023).

Reducción de dimensionalidad

Consiste en reducir el número de variables (o dimensiones) en un dataset mientras se conserva la mayor parte de la información relevante. El método más común para esto es el Análisis de Componentes Principales (PCA), y se utiliza cuando el dataset tiene muchas variables, lo que podría hacer que los algoritmos no supervisados sean ineficaces o computacionalmente costosos (Bobadilla, 2021).

Aunque es una transformación clave se debe decidir cuidadosamente cuándo aplicar la reducción de dimensionalidad, ya que puede eliminar variabilidad importante para el clústering o detección de anomalías (McMahon, 2022).

Codificación

La codificación es el proceso de convertir las variables categóricas en un formato que los algoritmos de machine learning puedan entender (generalmente números). Los métodos más comunes son "one-hot encoding" (Convierte categorías en columnas binarias, con 1 si pertenece a la categoría y 0 si no) o "label encoding" (Asigna un número entero único a cada categoría). Para técnicas no supervisadas como las que se van aplicar en este proyecto, es preferible one-hot encoding, que evita la asignación arbitraria de valores numéricos a las categorías (Grus, 2023).

Manejo de nulos

El manejo de valores nulos implica tratar las celdas vacías o faltantes en el dataset para evitar que afecten el funcionamiento de los algoritmos. Antes de aplicar técnicas como PCA, clústering o detección de anomalías, es esencial que el dataset no tenga valores nulos, ya que muchos algoritmos no pueden manejar datos faltantes de manera adecuada (VanderPlas, 2022).

Para gestionar los nulos se utiliza la imputación que consiste en rellenar los valores nulos con la media, mediana o un valor predeterminado o la eliminación que simplemente es remover las filas o columnas que contengan valores nulos, si la cantidad de datos faltantes es considerablemente baja (VanderPlas, 2022).

Transformar variables

Transformar variables implica aplicar modificaciones a los datos originales para que se ajusten mejor a los requisitos de los algoritmos, o para mejorar su estructura. Es conveniente aplicarla cuando las variables presentan distribuciones sesgadas o no normales. Por ejemplo aplicar transformaciones logarítmicas o raíces cuadradas para suavizar las variaciones extremas. Esto es útil en técnicas no supervisadas que se benefician de una distribución más homogénea de los datos (Géron, 2023).

Crear nuevas características

La creación de nuevas características o “variables derivadas” implica combinar o transformar las variables existentes para obtener nuevas variables que mejoren la capacidad del modelo de detectar patrones (Géron, 2023).

Se puede utilizar cuando se cree que las combinaciones de variables podrían ofrecer información adicional para el modelo. Por ejemplo, crear una variable que combine población con indicadores de violencia podría ser útil para identificar patrones más complejos en clústering.

Eliminar variables irrelevantes

Es el proceso de remover variables que no contribuyen a la tarea de machine learning o que introducen ruido en los datos. En algoritmos no supervisados, es crucial eliminar variables irrelevantes que puedan sesgar los resultados. Por ejemplo, identificadores únicos o variables que no aporten variabilidad. Este paso se puede realizar con análisis exploratorio o aplicando técnicas como PCA para observar qué variables tienen poco impacto en la varianza (McMahon, 2022).

Manejo de outliers

Los outliers o valores atípicos son puntos de datos que se alejan significativamente del resto de los datos. Pueden afectar negativamente el rendimiento de los algoritmos. En técnicas de clústering, los outliers pueden desviar la formación de grupos. En detección de anomalías, los outliers son precisamente el objetivo, por lo que identificarlos correctamente es crucial (McMahon, 2022).

3.4 Aplicación de algoritmos de Machine Learning

Esta parte del procedimiento incluye la selección de las técnicas de análisis más adecuadas para el dataset, el entrenamiento de los modelos y la posterior evaluación de los resultados. Se aplican técnicas de aprendizaje no supervisado sobre el dataset depurado que contiene variables de violencia. Los grupos resultantes (perfils específicos de violencia de los municipios) se utilizan como etiquetas en el dataset de variables socioeconómicas, aplicando una regresión logística lo que permitirá explorar la relación entre condiciones socioeconómicas y los diferentes perfils de violencia.

El análisis de datos no supervisados para analizar las variables de violencia es particularmente adecuado en este proyecto porque los datos sobre violencia de los municipios del Cauca no tienen una etiqueta predefinida de "resultado correcto". En lugar de buscar clasificaciones predeterminadas de violencia, primero se busca entender la distribución de los datos, descubrir posibles relaciones entre variables y municipios, e identificar tendencias o casos atípicos que podrían estar ocultos en la estructura compleja del conjunto de datos.

Selección de las técnicas a aplicar

Las técnicas no supervisadas, como PCA, clústering y detección de anomalías, son idóneas para este tipo de análisis porque nos permiten explorar la estructura subyacente de los datos y encontrar agrupaciones naturales, reducir la dimensionalidad para simplificar el análisis y detectar puntos anómalos que podrían ser de interés (Géron, 2023). Esto es especialmente relevante en el contexto de los datos de violencia y socioeconómicos, donde los patrones pueden no ser evidentes a simple vista, y las relaciones entre las variables pueden ser complejas.

Se seleccionan las técnicas de clústering, PCA y detección de anomalías porque son técnicas no supervisadas que permiten extraer información valiosa de conjuntos de datos sin etiquetas predefinidas, como en el caso de los datos de violencia de los municipios del departamento del Cauca. Estas técnicas ayudan a descubrir patrones subyacentes en los datos que podrían no ser evidentes a simple vista, proporcionando así un enfoque integral para el análisis exploratorio (Becker, 2024).

Análisis de Componentes Principales (PCA). Es una técnica estadística de reducción de dimensionalidad que transforma un conjunto de variables posiblemente correlacionadas en

un conjunto de componentes no correlacionados, ordenados según la cantidad de varianza que explican. Esto permite visualizar y trabajar con los datos en un espacio reducido, reteniendo la mayor cantidad de información posible (Becker, 2024).

Los datos sobre violencia suelen tener muchas variables correlacionadas, son muchas las variables sobre este tema que se pueden extraer de la colosal base de datos que se construyó para el proyecto. El PCA permite reducir la cantidad de dimensiones, facilitando la visualización de los datos y su posterior análisis sin perder de vista las relaciones más significativas (Géron, 2023).

El Clústering o agrupamiento. Es una técnica que agrupa los datos en subconjuntos (clústeres), donde los datos dentro de un clúster son más similares entre sí que con los datos en otros clústeres. Esto se utiliza para identificar grupos naturales dentro de los datos sin necesidad de etiquetas predeterminadas (Grus, 2023).

Dada la diversidad de los municipios del Cauca y sus variaciones en términos de violencia, el clústering puede ayudar a identificar grupos de municipios que comparten características similares.

El clústering podría revelar, por ejemplo, que ciertos municipios con niveles similares presentan patrones de violencia comparables, o que existen grupos de municipios que, aunque parezcan similares en ciertos indicadores, difieren significativamente en términos de violencia.

La detección de anomalías. Es útil para identificar puntos de datos que se desvían significativamente del patrón general. En este contexto, puede ayudar a identificar municipios cuyos niveles de violencia son excepcionales, en comparación con el resto del departamento (VanderPlas, 2022).

En un conjunto de 42 municipios del departamento del Cauca, es muy probable que existan municipios con características extremas en términos de violencia. La detección de anomalías permite resaltar estos casos atípicos, que podrían ser objeto de estudios adicionales. Esta técnica podría identificar municipios con niveles de violencia atípicamente altos o bajos en comparación con otros municipios con condiciones aparentemente similares.

La regresión logística: Cumple un papel importante al servir como un modelo de aprendizaje supervisado que permite cuantificar y entender la relación entre las condiciones socioeconómicas y los perfiles de violencia identificados en los municipios del Cauca. Tras aplicar técnicas no supervisadas (K-means, PCA y detección de anomalías) al dataset de violencia y agrupar los municipios en perfiles específicos, estos grupos se convierten en etiquetas de clasificación para cada municipio. Con estas etiquetas, la regresión logística se utiliza para analizar en qué medida las variables socioeconómicas predicen la pertenencia a cada perfil de violencia, otorgando así un marco de interpretación predictiva a los datos.

En este proceso, la regresión logística funciona como un ensamblador que incorpora los hallazgos no supervisados, convirtiéndolos en una referencia de clasificación multietiqueta. Esto facilita el análisis del impacto potencial de cada variable socioeconómica en el riesgo o la probabilidad de que un municipio exhiba características de ciertos perfiles de violencia. Este enfoque en ensamble permite construir modelos más completos y específicos sobre cómo las condiciones sociales y económicas de cada municipio están vinculadas a sus niveles y tipos de violencia, añadiendo un componente analítico clave para la toma de decisiones basadas en estos hallazgos.

Entrenamiento de los modelos

En esta etapa, cuando los datos ya han pasado por el preprocesamiento necesario (escalado, manejo de nulos, reducción de dimensionalidad, etc.), se procede a entrenar los modelos seleccionados: Clústering (K-Means), PCA (Análisis de Componentes Principales), Detección de Anomalías con Isolation Forest y regresión logística, usando la biblioteca Scikit-learn en Python. Dado que todas las transformaciones previas de los datos ya se han completado, esta fase se centra en ajustar y entrenar los modelos sobre los datos preprocesados.

Es importante resaltar que las técnicas seleccionadas, como clústering, PCA y detección de anomalías, por ser no supervisadas, no requieren dividir los datos en los conjuntos tradicionales de entrenamiento y validación. A diferencia de los algoritmos de aprendizaje supervisado, que requieren una variable objetivo (etiqueta) para predecir, las técnicas no supervisadas no buscan predecir una etiqueta específica, sino que se enfocan en descubrir patrones, estructuras y relaciones subyacentes en los datos (Raschka & Mirjalili, 2020). Debido a esta naturaleza, no es necesario dividir los datos en conjuntos de entrenamiento y validación, ya que el objetivo es analizar la totalidad del dataset para identificar agrupaciones,

reducir la dimensionalidad (PCA), o detectar anomalías sin tener un valor conocido de referencia.

Proceso de clústering utilizando K-Means. El modelo agrupa los datos en un número predefinido de clústeres, donde cada punto se asigna al clúster cuyo centroide está más cercano, minimizando la distancia intra-clúster. En un entrenamiento inicial, se puede emplear un valor de k definido empíricamente o mediante un análisis previo, con el método del codo, que ayuda a determinar el número óptimo de clústeres (Géron, 2023).

El entrenamiento inicial con K-Means sigue varios pasos. Primero, se asigna un valor predeterminado para k, típicamente 3 o 5, dependiendo de las características de los datos o su naturaleza. Luego, el modelo entrena en el conjunto de datos, iterando hasta que los centroides converjan, es decir, hasta que los movimientos de los centroides entre iteraciones sean mínimos y la configuración de los clústeres se stabilice (Becker, 2024).

Si después de este entrenamiento inicial los resultados son prometedores pero aún no óptimos, es común proceder con el ajuste de hiperparámetros. Los hiperparámetros clave a ajustar en K-Means incluyen el número de clústeres k y el criterio de inicialización de los centroides, como "k-means++" o inicialización aleatoria. A través de técnicas como la validación o el análisis de la curva del codo, se puede ajustar el número óptimo de clústeres para mejorar tanto la cohesión dentro de los clústeres como la separación entre ellos, logrando un mejor rendimiento del modelo (Géron, 2023).

En Python, utilizando la librería scikit-learn, la función KMeans permite implementar este algoritmo, proporcionando opciones para ajustar tanto el valor de k como la estrategia de inicialización de los centroides, optimizando así el modelo según la estructura de los datos (McKinney, 2022).

Entrenamiento para la reducción de dimensionalidad con PCA. El objetivo es proyectar los datos en un espacio de menor dimensionalidad, capturando la mayor cantidad de variabilidad posible en las primeras componentes principales. Inicialmente, se lleva a cabo el proceso sin especificar el número de componentes, lo que permite que el algoritmo calcule y retenga todas las componentes principales posibles (Becker, 2024).

Durante este entrenamiento, los datos son descompuestos en componentes principales, y se calcula la varianza explicada por cada una de ellas. Esto proporciona una visión clara de cómo se distribuye la información en el nuevo espacio reducido.

Si es necesario reducir la dimensionalidad de manera más significativa, el hiperparámetro clave a ajustar es el número de componentes a retener. Para esto, se puede definir un umbral de varianza explicada acumulada, por ejemplo, seleccionando las componentes que, en conjunto, expliquen el 90% de la varianza total de los datos. Ajustar este parámetro permite encontrar un balance adecuado entre la simplicidad del modelo (al reducir la dimensionalidad) y la preservación de información, garantizando que se sigan capturando los patrones más relevantes dentro de los datos (Becker, 2024).

En Python, utilizando la librería sklearn, se aplica la función PCA para llevar a cabo este proceso, especificando el número de componentes o el umbral de varianza explicada según sea necesario. Esto permite un control preciso sobre la cantidad de información que se retiene en el análisis final.

Detección de anomalías mediante Isolation Forest. Es una técnica efectiva diseñada para identificar observaciones que se comportan de manera diferente al resto. El algoritmo funciona mediante la construcción de "árboles de aislamiento", que dividen aleatoriamente los datos hasta que las observaciones quedan completamente aisladas. Las observaciones que requieren pocas divisiones para ser separadas suelen ser consideradas anomalías, ya que su comportamiento es notablemente distinto al del resto del conjunto de datos (Bobadilla, 2021).

El entrenamiento inicial de modelo de Isolation Forest se entrena usando el conjunto de datos completo, aplicando los hiperparámetros por defecto, como `n_estimators`, que define el número de árboles en el bosque de aislamiento (por defecto, 100) y `max_samples`, que especifica el número máximo de muestras utilizadas para entrenar cada árbol.

Si las anomalías no se detectan adecuadamente o si se producen falsos positivos, se puede ajustar el modelo usando varios hiperparámetros:

- `n_estimators`: aumentar el número de árboles puede mejorar la precisión del modelo, aunque esto incrementa también el tiempo de procesamiento (Géron, 2023).
- `max_samples`: variar la proporción de datos que utiliza cada árbol permite ajustar el modelo según la complejidad del conjunto de datos (Géron, 2023).

- contamination: este parámetro establece la proporción esperada de anomalías en los datos, mejorando así la precisión al ajustar el modelo a la realidad del dataset (Géron, 2023).

La regresión logística, en este contexto, se presenta como una técnica especialmente efectiva para intentar modelar la relación entre los factores socioeconómicos y los perfiles de violencia identificados en los municipios del Cauca. Tras agrupar estos municipios mediante técnicas no supervisadas (K-means, PCA y detección de anomalías) en función de sus patrones de violencia, se asignan etiquetas de clasificación a cada uno. La regresión logística multietiqueta aprovecha estos grupos para explorar de forma precisa las asociaciones entre las condiciones socioeconómicas y las distintas categorías de violencia, brindando una comprensión de los factores que más inciden en cada perfil.

Esta metodología es particularmente adecuada en el análisis del contexto regional, ya que facilita una interpretación clara de los elementos socioeconómicos que aumentan la probabilidad de que un municipio pertenezca a un perfil de violencia específico. Al combinar técnicas de aprendizaje no supervisado con un enfoque supervisado, el modelo de regresión logística se convierte en una herramienta ideal para estudiar la relaciones entre las características socioeconómicas y la variable objetivo (perfiles de violencia).

3.5 Evaluación de resultados

Evaluar los resultados de las técnicas de aprendizaje no supervisado es fundamental para garantizar la calidad y fiabilidad de los modelos. Cada técnica tiene formas específicas de evaluar su rendimiento, y los resultados se consideran satisfactorios cuando cumplen ciertos criterios. Además, cuando los resultados no son satisfactorios, es importante ajustar los parámetros, mejorar el preprocesamiento de los datos o intentar con enfoques alternativos.

Clustering

Para evaluar los algoritmos de clustering como K-means en Python, se utilizan métricas internas y externas. En el caso de K-means, la función KMeans de sklearn.clúster permite obtener la inercia, que mide la suma de las distancias al cuadrado de cada muestra al centro de su clúster. Los valores más bajos de inercia indican mejor agrupación (Bobadilla, 2021).

Otra métrica es el coeficiente de silueta, obtenido mediante sklearn.metrics.silhouette_score, que mide qué tan bien asignada está una muestra a su clúster. Un valor cercano a 1 es deseable, ya que indica que los clústeres están bien definidos. El índice de Davies-Bouldin también se puede usar con sklearn.metrics.davies_bouldin_score, donde valores bajos indican clústeres bien separados (McMahon, 2022).

Si las métricas indican que los clústeres son compactos, bien separados y consistentes con los datos, los resultados se consideran satisfactorios. De ser así, los siguientes pasos incluyen la interpretación de los clústeres, visualizaciones como gráficos de dispersión y la aplicación de los resultados al contexto específico. Sin embargo, si los resultados no son satisfactorios, es posible ajustar parámetros, como el número de clústeres, probar con otro algoritmo o asegurarse de que los datos estén correctamente normalizados.

Análisis de Componentes Principales (PCA)

El análisis de componentes principales (PCA) se evalúa principalmente mediante la proporción de varianza explicada por cada componente. Con la función PCA de sklearn.decomposition, se puede obtener la varianza explicada acumulada, que idealmente debería ser del 80-90% en las primeras componentes. Un gráfico de sedimentación (scree plot) puede mostrar un "punto de codo", donde agregar más componentes no incrementa significativamente la varianza explicada (Becker, 2024).

Cuando los primeros componentes explican la mayor parte de la variabilidad de los datos, se considera que los resultados son satisfactorios. Los siguientes pasos incluyen interpretar las componentes para identificar las variables más influyentes, visualizar los datos en un espacio de menor dimensión y usar las componentes para reducir la dimensionalidad de los datos en análisis posteriores. Si los resultados no son satisfactorios, se puede intentar incluir más componentes, probar técnicas no lineales como t-SNE, o transformar los datos antes de aplicar PCA.

Detección de Anomalías con Isolation Forest

La detección de anomalías para técnicas de machine learning no supervisado con Isolation Forest se basa en asignar a cada muestra un puntaje que indica su probabilidad de ser una anomalía. El algoritmo funciona creando árboles de decisión que aislan observaciones individuales, donde las anomalías suelen ser aisladas más rápidamente debido a su rareza. Para evaluar los resultados, es importante observar la coherencia de las anomalías detectadas con las expectativas del análisis, asegurándose de que correspondan a comportamientos inusuales en los datos (Géron, 2023).

Si los resultados no son satisfactorios, se pueden ajustar los umbrales de detección, probar otros algoritmos no supervisados o realizar un análisis adicional sobre las características de los datos para asegurar que capturen correctamente los patrones anómalos.

Regresión logística

La evaluación de la regresión logística en este caso se centra en la precisión y la capacidad del modelo para identificar correctamente los perfiles de violencia en función de los factores socioeconómicos de cada municipio. Para lograrlo, se emplean el análisis de los coeficientes log-odds y las razones de probabilidad (odds rates), además de las métricas de desempeño como la exactitud (accuracy), la sensibilidad (recall), y el valor F1, las cuales permiten medir tanto la cantidad de predicciones correctas del modelo (Géron, 2023) como su capacidad de capturar los perfiles de violencia más críticos en el contexto del Cauca.

Adicionalmente, se emplea la matriz de confusión para identificar errores específicos en la clasificación, analizando qué perfiles de violencia son más difíciles de predecir en función de los factores socioeconómicos. Esto permite un ajuste fino del modelo, destacando aquellos aspectos de las condiciones socioeconómicas que podrían estar sub- o sobre-representados. La evaluación, así, no solo determina la precisión del modelo en el presente conjunto de datos, sino que también proporciona información para refinar el análisis y mejorar su capacidad para explicar la relación entre las condiciones socioeconómicas y los perfiles de violencia en el departamento del Cauca.

3.6 Informe final

Una vez obtenidos resultados satisfactorios, el paso final es la interpretación y presentación de los hallazgos en un informe técnico detallado. Este informe debe proporcionar un análisis profundo de los patrones identificados a través de las técnicas de clústering, PCA, detección de anomalías y regresión logística, destacando las agrupaciones relevantes de municipios, las dimensiones más significativas que explican la variabilidad en los datos socioeconómicos y de violencia, así como las anomalías detectadas en ciertos municipios o períodos.

El informe también incluirá visualizaciones claras que faciliten la comprensión de los resultados, como gráficos de componentes principales, mapas de calor de las agrupaciones, y descripciones cualitativas de los hallazgos más importantes.

Resultados y discusiones

En esta sección se presentan los resultados obtenidos en las dos etapas del proyecto. La primera, dedicada a construir un robusto repositorio virtual de datos, y la segunda, donde se emplean técnicas de aprendizaje automático para identificar patrones y relaciones en los datos sobre violencia y condiciones socioeconómicas del Cauca entre 2019 y 2023.

Para cada etapa, se incluyen tablas, figuras, gráficos, diagramas y anexos que complementan la explicación, proporcionando una visión clara de los resultados obtenidos.

1. Resultados y discusiones de la creación del repositorio virtual de datos – Etapa 1

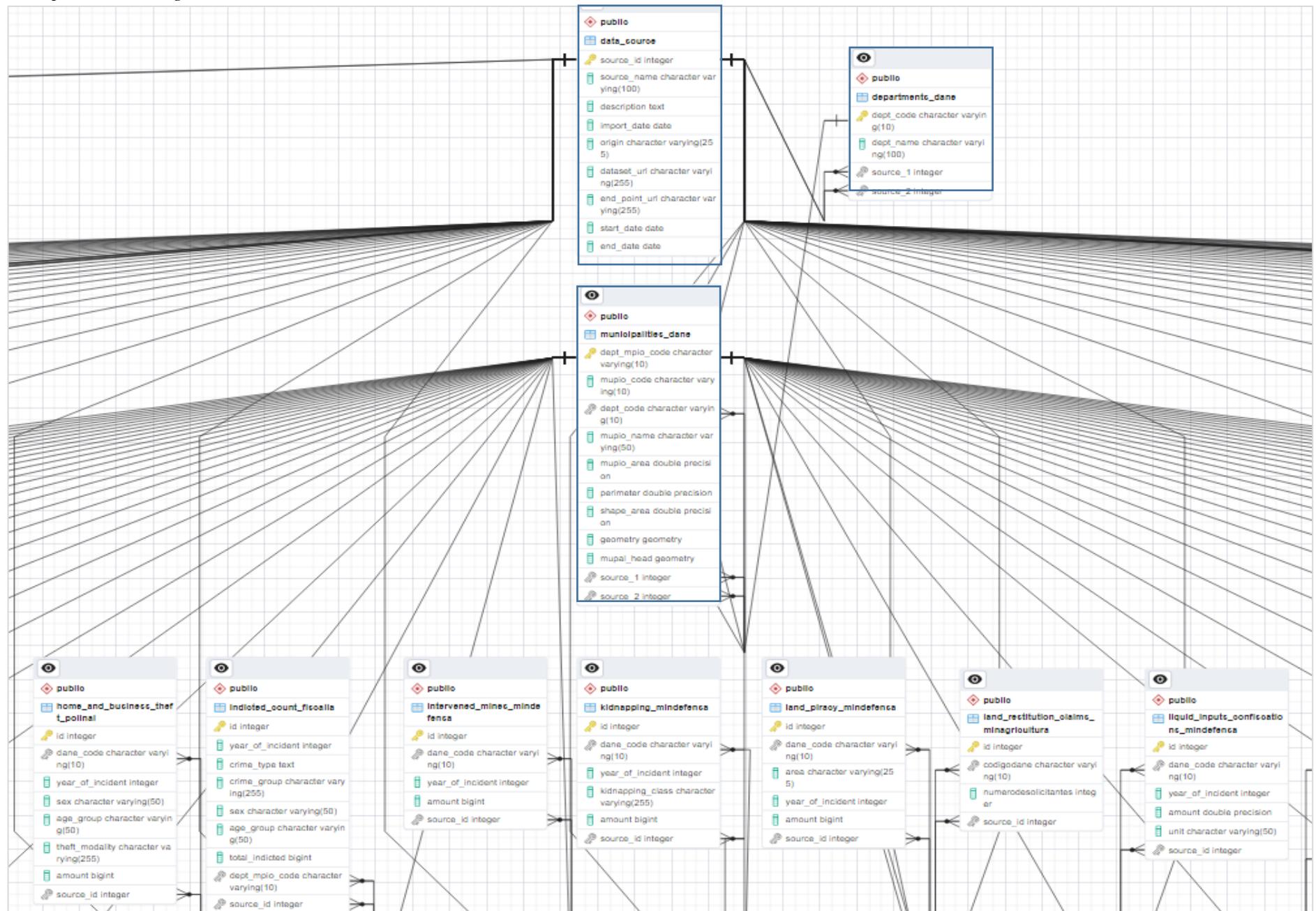
La primera etapa del proyecto se centra en la construcción de una base de datos relacional geoespacial en PostgreSQL con PostGIS, diseñada para almacenar datos sobre eventos de violencia y variables socioeconómicas en los municipios de Colombia. A través de un proceso ETL robusto y juiciosamente implementado con Python, se logró unificar información de diversas fuentes en un repositorio virtual accesible y estructurado. Los resultados tangibles de esta etapa se muestran a continuación.

1.1 Diseño de la base de datos

En la figura presentada en la siguiente página se muestra una vista parcial del diagrama Entidad-Relación (ER) de la base de datos implementada, generado directamente en pgAdmin4, donde se distingue el núcleo (*core*) de la base de datos, compuesto por las tablas principales *data_source* y *municipalities_dane*, junto con la tabla auxiliar *departments*. Debido al gran número de tablas (66 en total), no es posible visualizar el diagrama completo, pero se destacan las relaciones clave entre el núcleo central y varias tablas de datos. Se observa que todas las tablas de datos, a pesar de no estar relacionadas entre sí, están vinculadas directamente a las tablas del núcleo, lo que asegura tanto la trazabilidad de las fuentes de los datos a través de *data_source* como su geolocalización a nivel municipal mediante *municipalities_dane*. Esta estructura garantiza una sólida integridad relacional y permite realizar análisis geoespaciales y multitemporales.

Figura 10

Vista parcial del diagrama ER de la base de datos



Fuente: Generado desde pgAdmin4

1.2 Verificación de que las tablas que conforman el core de la base de datos cumplen con el propósito para el que fueron diseñadas

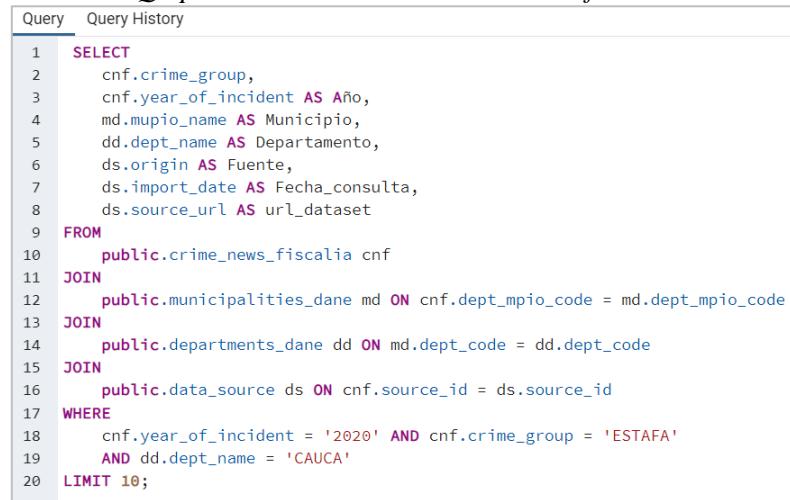
A continuación se demuestra que el sistema cumple con los requisitos de trazabilidad y georreferenciación anunciados a lo largo de este informe, permitiendo seguir la procedencia de los datos y visualizarlos de manera efectiva en el espacio geográfico.

Trazabilidad de la fuente en la base de datos

La siguiente consulta SQL ejecutada sobre la base de datos construida sirve para rastrear la procedencia de los datos relacionados con incidentes de estafa en el departamento del Cauca durante el año 2020.

Figura 11

Consulta SQL para demostrar trazabilidad de la fuente de datos



```

1  SELECT
2      cnf.crime_group,
3      cnf.year_of_incident AS Año,
4      md.mupio_name AS Municipio,
5      dd.dept_name AS Departamento,
6      ds.origin AS Fuente,
7      ds.import_date AS Fecha_consulta,
8      ds.source_url AS url_dataset
9  FROM
10     public.crime_news_fiscalia cnf
11  JOIN
12     public.municipalities_dane md ON cnf.dept_mpio_code = md.dept_mpio_code
13  JOIN
14     public.departments_dane dd ON md.dept_code = dd.dept_code
15  JOIN
16     public.data_source ds ON cnf.source_id = ds.source_id
17  WHERE
18      cnf.year_of_incident = '2020' AND cnf.crime_group = 'ESTAFA'
19      AND dd.dept_name = 'CAUCA'
20  LIMIT 10;

```

Fuente: Pantallazo de la consulta obtenido desde la aplicación PgAdmin4

La consulta SQL ejecutada en pgAdmin4 devuelve los siguientes resultados, mostrando los metadatos de la tabla data_source: fuente, fecha_consulta y url_dataset.

Figura 12

Resultado de consulta SQL que demuestra trazabilidad de la fuente de la base de datos

	crime_group character varying (255) 	año integer 	municipio character varying (50) 	departamento character varying (100) 	fuente character varying (255) 	fecha_consulta date 	url_dataset character varying (255) 
1	ESTAFA	2020	POPAYAN	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv
2	ESTAFA	2020	EL TAMBO	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv
3	ESTAFA	2020	POPAYAN	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv
4	ESTAFA	2020	POPAYAN	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv
5	ESTAFA	2020	PUERTO TEJADA	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv
6	ESTAFA	2020	SILVIA	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv
7	ESTAFA	2020	POPAYAN	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv
8	ESTAFA	2020	PADILLA	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv
9	ESTAFA	2020	PIENDAMO - TUNIA	CAUCA	Fiscalía General de la Nación	2024-11-09	https://www.datos.gov.co/resource/6d52-qyqg.csv

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Georreferenciación de los datos

A continuación se demuestra la georreferenciación de los datos almacenados en la base de datos. El propósito de esta demostración es realizar una visualización geoespacial de los homicidios registrados en diferentes municipios de Colombia entre el periodo 2015 y 2020, utilizando técnicas que permitan la visualización de estos datos en mapas interactivos. El flujo de trabajo empieza por la extracción de datos de PostgreSQL mediante un script ejecutado localmente en VSCode, seguido de la exportación de los datos a un archivo CSV, que luego se procesa y visualiza en Google Colab.

Se inicia conectándose a la base de datos PostgreSQL del proyecto mediante un script de Python ejecutado en Vscode, cuya imagen se muestra en la página siguiente. Se usa la librería psycopg2, para establecer la conexión entre el script y la base de datos.

El código accede a la base de datos alojada localmente, y mediante la ejecución de una consulta SQL sobre las tablas que contienen la información de municipios, departamentos y noticias de crímenes en la Fiscalía General de la Nación, recupera información sobre el número de delitos agrupados como "HOMICIDIOS" cometidos entre los años 2015 y 2020 a nivel municipal, además de la geometría (Polígono) y la ubicación de la cabecera municipal (geometría de punto) para la visualización geoespacial de los datos en los municipios. Los resultados de la consulta, se transforman en un DataFrame utilizando pandas, que finalmente se salva en un archivo CSV, que se guarda en el dispositivo local.

Aunque es posible generar visualizaciones de mapas directamente en VSCode utilizando bibliotecas como matplotlib y geopandas, se opta por subir los datos en formato CSV a Google Drive para generar los mapas de geolocalización desde el entorno de Google Colab utilizando las librerías de Python. Este entorno de ejecución da acceso a recursos en la nube de mayor potencia que los de la máquina local, lo que es necesario para tareas como el procesamiento de grandes conjuntos de datos. Además Google Colab permite compartir cuadernos fácilmente con otros, lo cual es ideal cuando se trabaja en equipo, se integra con Google Drive, lo que elimina la necesidad de mover archivos de un lugar a otro manualmente, y permite el acceso a muchas de las bibliotecas necesarias para trabajar con datos, lo que ahorra tiempo en la configuración de entornos y facilita la ejecución inmediata del código (Colab.Google, n.d.).

Figura 13*Script Python – Conexión, consulta y generación archivo CSV*

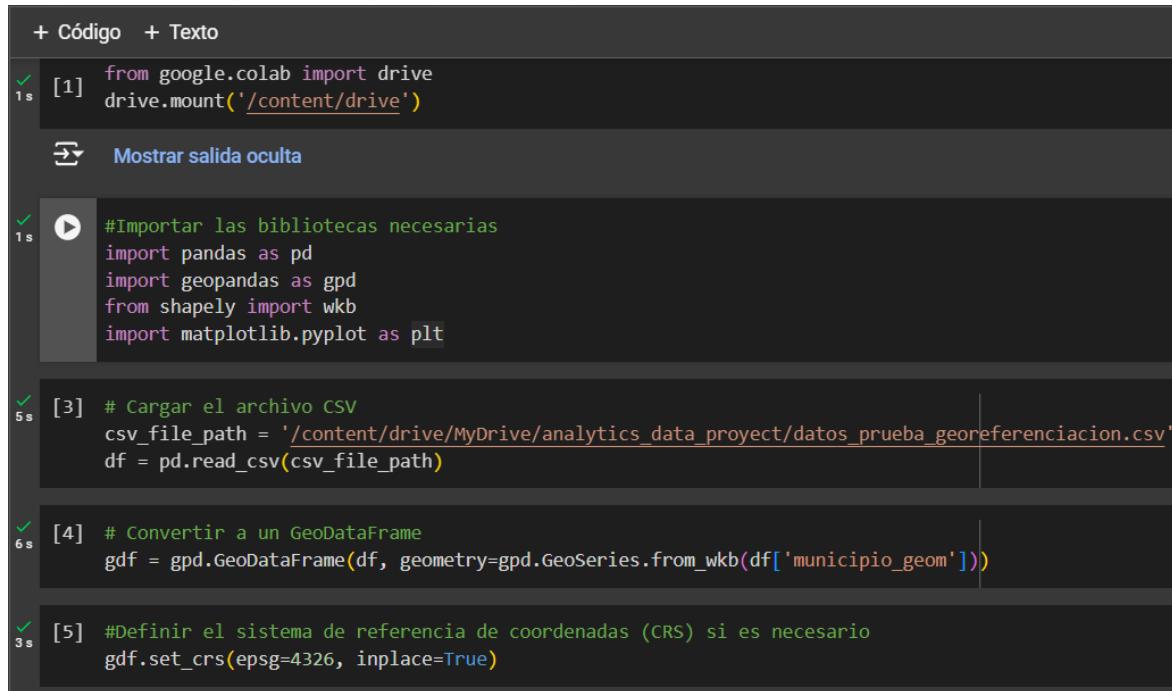
```

proyecto_grado > ejecucion_proyecto > prueba_georeferenciacion_datos.py > ...
1 import psycopg2
2 import pandas as pd
3
4 # Configuración de la conexión
5 conn_params = {
6     "host": "localhost",
7     "port": "5432",
8     "database": "db_proyecto",
9     "user": "postgres",
10    "password": "cauca"
11 }
12
13 # Consulta optimizada
14 query = """
15 SELECT
16     d.dept_name AS departamento,
17     m.mupio_name AS municipio,
18     m.mupal_head AS cabecera_mpal,
19     COUNT(c.id) AS num_homicidios,
20     m.geometry AS municipio_geom
21 FROM
22     public.municipalities_dane m
23 LEFT JOIN
24     public.departments_dane d ON m.dept_code = d.dept_code
25 LEFT JOIN
26     public.crime_news_fiscalia c ON m.dept_mpio_code = c.dept_mpio_code
27 WHERE
28     c.crime_group IN ('HOMICIDIO CULPOSO', 'HOMICIDIO DOLOSO') |
29     AND c.year_of_incident BETWEEN 2015 AND 2020
30 GROUP BY
31     d.dept_name, m.mupio_name, m.mupal_head, m.geometry
32 ORDER BY
33     d.dept_name, num_homicidios DESC;
34 """
35
36 # Usar el manejador de contexto para la conexión
37 with psycopg2.connect(**conn_params) as conn:
38     df = pd.read_sql_query(query, conn)
39
40 # Generar archivo CSV
41 path ='D:/CUN/Analitica_de_datos_especializacion/proyecto_grado/ejecucion_proyecto/'
42 df.to_csv(path + 'datos_prueba_georeferenciacion.csv', index=False)

```

Fuente: Pantallazo obtenido desde la aplicación Vscode

Una vez que el archivo CSV es cargado en Google Drive, se accede a él desde Google Colab. A partir de aquí, se utilizan bibliotecas como pandas y geopandas para manipular los datos, transformar coordenadas y preparar los datos geoespaciales para la visualización. (Becker, 2024). Tal como se muestra el siguiente fragmento de código.

Figura 14*Script Python - Prepara datos para generar mapas interactivos*


```
+ Código + Texto
1 s [1] from google.colab import drive
drive.mount('/content/drive')

Mostrar salida oculta

1 s ⏴ #Importar las bibliotecas necesarias
import pandas as pd
import geopandas as gpd
from shapely import wkb
import matplotlib.pyplot as plt

5 s [3] # Cargar el archivo CSV
csv_file_path = '/content/drive/MyDrive/analytics_data_proyect/datos_prueba_georeferenciacion.csv'
df = pd.read_csv(csv_file_path)

6 s [4] # Convertir a un GeoDataFrame
gdf = gpd.GeoDataFrame(df, geometry=gpd.GeoSeries.from_wkb(df['municipio_geom']))

3 s [5] #Definir el sistema de referencia de coordenadas (CRS) si es necesario
gdf.set_crs(epsg=4326, inplace=True)
```

Fuente: Pantallazo obtenido desde la aplicación Vscode

Los datos procesados con el script anterior quedan listos para ser visualizados en mapas utilizando geopandas o matplotlib. Puesto que los datos pasan a un GeoDataFrame, se puede generar mapas a diferentes niveles geográficos: todo el país, por regiones, por departamento y por municipios, como se demuestra a continuación, donde a través de código Python se generan “mapas de coropletas”, que son un tipo de mapa temático en el que las áreas (en este caso, los municipios de Colombia) están sombreadas o coloreadas de acuerdo con el valor cuantitativo (GeoPandas, n.d.) que representa el número de homicidios para el periodo 2015 -2020.

La siguiente figura muestra cuatro script de Python: el primero genera el mapa de todo el país dividido por municipios, el segundo el mapa de los departamentos de la costa Caribe, el tercero el mapa del departamento de Antioquia. Todos estos son ‘mapas de coropletas’ donde los municipios están coloreados de acuerdo al número de homicidios para el periodo 2015 - 2020. El cuarto script representa solamente el Municipio de El Tambo en el Cauca, y resalta el número de homicidios cometidos durante el periodo 2015-2020. Los mapas generados, todos a partir de datos extraídos directamente de la base de datos Postgress, se muestran a continuación de los códigos.

Figura 15*Validación de geolocalización en base de datos con Python*

```

▶ # Visualizar la totalidad de municipios de Colombia
gdf.plot(column='num_homicidios', cmap='Reds', legend=True, figsize=(6, 5), vmin=0, vmax=gdf['num_homicidios'].mean())
plt.suptitle('Distribución de Homicidios por Municipio')
plt.title('Periodo 2015-2022', fontsize=10)
# Ocultar los ejes
plt.axis('off')
# Mostrar mapa
plt.show()

▶ # Visualizar departamentos del Caribe Colombiano
departamentos_caribe = ['ATLANTICO', 'BOLIVAR', 'SUCRE', 'CORDOBA', 'MAGDALENA', 'CESAR', 'LA GUAJIRA', ]

# Filtrar el GeoDataFrame para mostrar solo los departamentos del Caribe
caribe_gdf = gdf[gdf['departamento'].isin(departamentos_caribe)]

# Visualizar el resultado
caribe_gdf.plot(column='num_homicidios', cmap='Reds', legend=True, figsize=(7,4), vmin=0, vmax=caribe_gdf['num_homicidios'].mean())
plt.suptitle('Homicidios en los Departamentos del Caribe')
plt.title('Distribución periodo 2015 - 2020', fontsize = 10)

# Ocultar los ejes
plt.axis('off')

# Mostrar el mapa
plt.show()

▶ # Filtrar el GeoDataFrame para mostrar solo el departamento Antioquia
cauca_gdf = gdf[gdf['departamento'] == 'ANTIOQUIA']

# Visualizar solo el departamento del Cauca
cauca_gdf.plot(column='num_homicidios', cmap='Reds', legend=True, figsize=(6, 5), vmin=0, vmax=cauca_gdf['num_homicidios'].mean())
plt.suptitle('Homicidios en el Departamento del Antioquia')
plt.title('Distribución periodo 2015 - 2020', fontsize = 10)
# Ocultar los ejes
plt.axis('off')

# Mostrar el mapa
plt.show()

▶ # Filtrar el GeoDataFrame para mostrar solo el municipio de El Tambo en el departamento del Cauca
tambo_gdf = gdf[(gdf['municipio'] == 'EL TAMBO') & (gdf['departamento'] == 'CAUCA')]

# Reproyectar el GeoDataFrame a un CRS proyectado
tambo_gdf = tambo_gdf.to_crs(epsg=3116)

# Visualizar solo el municipio de El Tambo
ax = tambo_gdf.plot(column='num_homicidios', cmap='Blues', legend=True, figsize=(5, 4), vmin=0, vmax=tambo_gdf['num_homicidios'].mean())

# Agregar título principal y subtítulo
plt.suptitle('Homicidios en el Municipio de El Tambo')
plt.title('Periodo 2015 - 2020', fontsize=10)

# Ocultar los ejes
plt.axis('off')

# Mostrar el número de homicidios directamente en el mapa
for x, y, label in zip(tambo_gdf.geometry.centroid.x, tambo_gdf.geometry.centroid.y, tambo_gdf['num_homicidios']):
    plt.text(x, y, str(label), fontsize=14, color='white', ha='center', va='center')

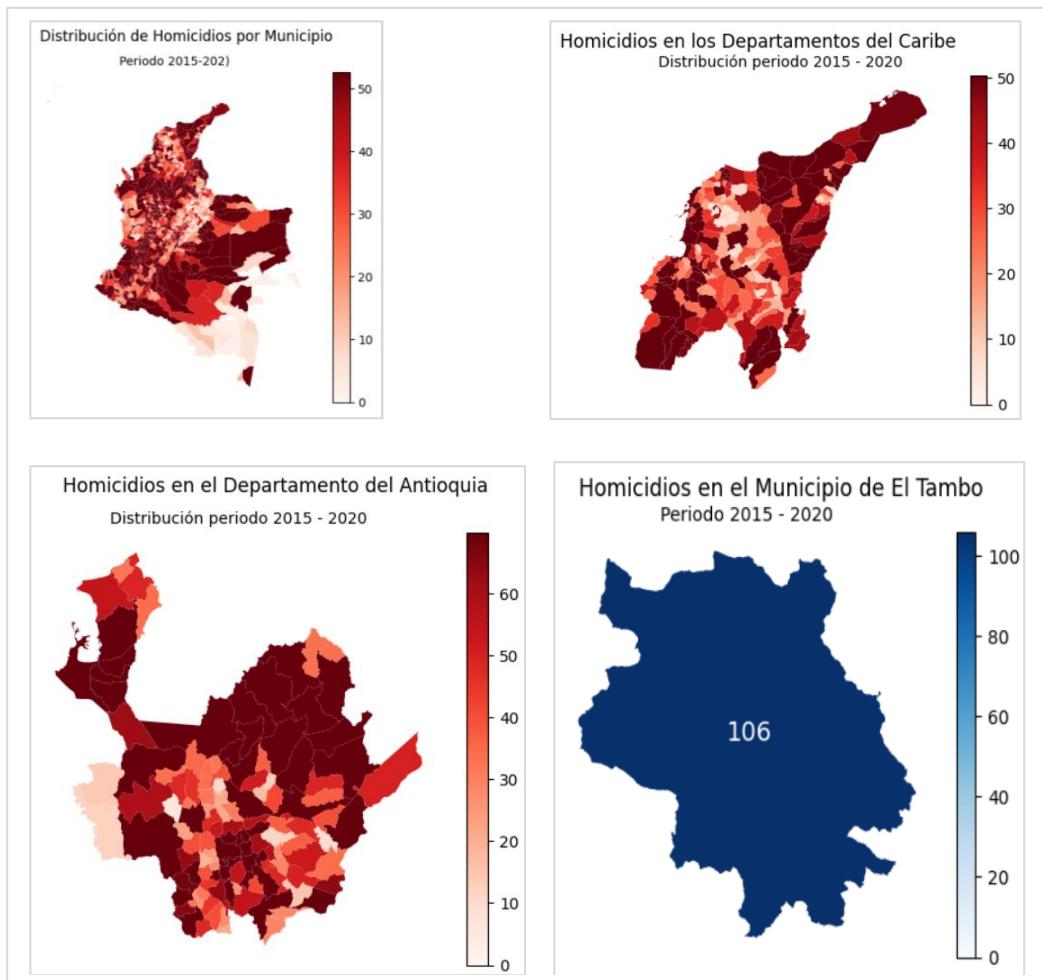
# Mostrar el mapa
plt.show()

```

Fuente: Pantallazo obtenido desde la aplicación Vscode

Figura 16

Mapas generados con la ejecución de los códigos de la figura 15 que demuestran la capacidad geoespacial de la base de datos



Fuente: Pantallazo obtenido desde la aplicación Vscode

1.3 Estadísticas y composición de la base de datos

Las información que se presenta a continuación ofrece una visión clara y cuantitativa de la estructura y contenido de la base de datos creada como producto del presente proyecto, y de paso permite evaluar su adecuación para el análisis que se pretende realizar. Este apartado proporciona una descripción detallada del volumen de datos, su distribución en las diferentes tablas y categorías, y su cobertura temporal y geográfica, lo cual es importante para establecer el contexto y las capacidades analíticas del sistema. Además, resalta las fuentes de los datos, su calidad, y las optimizaciones implementadas, para que los usuarios comprendan el

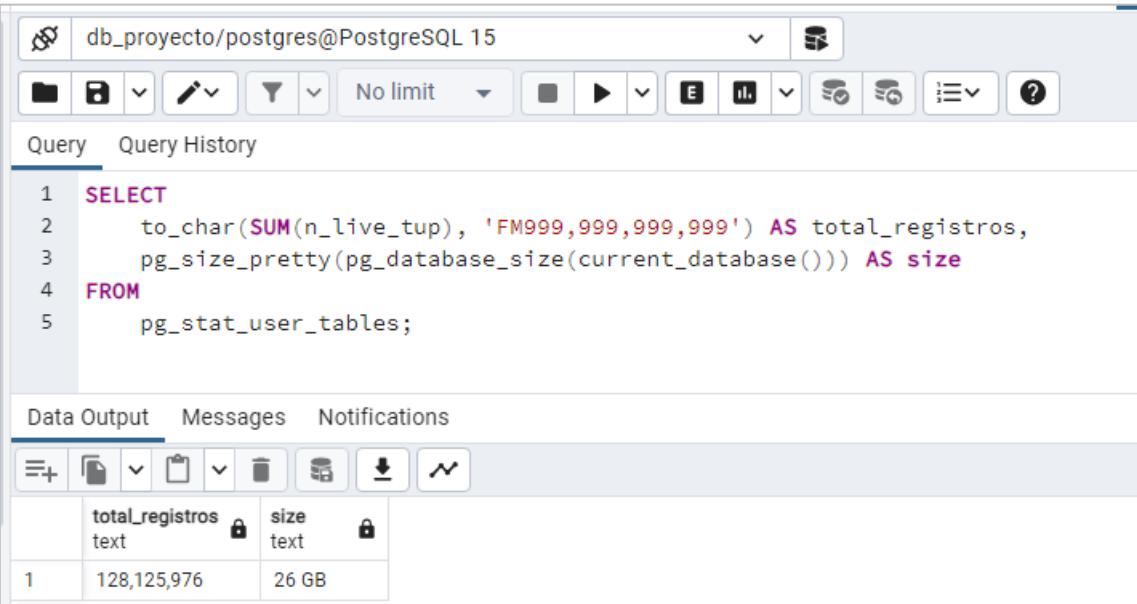
rendimiento y la eficiencia de las consultas, facilitando así una interpretación informada y sólida de los resultados.

Tamaño de la base de datos

El número total de registros y el espacio de almacenamiento de la base de datos nos da una primera aproximación sobre el tamaño de la misma. La figura siguiente muestra un consulta SQL ejecutada en pgAdmin4 que nos devuelve directamente estos valores.

Figura 17

Número total de registros y espacio de almacenamiento de la base de datos



The screenshot shows the PgAdmin4 interface. The top bar displays the connection information: db_proyecto/postgres@PostgreSQL 15. Below the toolbar, there are tabs for 'Query' (which is selected) and 'Query History'. The main area contains a SQL query:

```

1 SELECT
2     to_char(SUM(n_live_tup), 'FM999,999,999,999') AS total_registros,
3     pg_size_pretty(pg_database_size(current_database())) AS size
4 FROM
5     pg_stat_user_tables;

```

Below the query, there are tabs for 'Data Output', 'Messages', and 'Notifications'. The 'Data Output' tab is selected, showing a table with two columns: 'total_registros' and 'size'. The data row shows values: 128,125,976 and 26 GB respectively.

	total_registros	size
	text	text
1	128,125,976	26 GB

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

128,13 millones de registros es una cantidad considerable de datos, su almacenamiento en 26 GB puede considerarse eficiente, en cuanto al tamaño podemos decir que PostgreSQL está diseñado para manejar bases de datos mucho más grandes, tanto en términos de número de filas como de tamaño en disco.

Número total de tablas

La siguiente consulta aplicada a la base de datos desde la herramienta pgAdmin4 genera un archivo CSV que lista todas las tablas de usuario de la base de datos, incluyendo el número de registros en cada una, y presenta un total consolidado de los registros. Además, se agrega una columna numérica que incrementa de forma secuencial para identificar las tablas, en lugar de utilizar el nombre del esquema. El listado se ordena alfabéticamente por el nombre de las tablas, y el resultado se exporta a un archivo con encabezados, facilitando así su análisis externo.

Figura 18

Consulta SQL de generación de reporte de número de registros por tabla con total consolidado y exportación a CSV

```

Query  Query History
1 COPY(
2   WITH table_counts AS (
3     SELECT
4       ROW_NUMBER() OVER () AS row_number,
5       pg_class.relname AS table_name,
6       n_live_tup AS row_count,
7       (
8         SELECT COUNT(*)
9         FROM pg_attribute
10        WHERE attrelid = pg_class.oid
11        AND attnum > 0
12        AND NOT attisdropped
13       ) AS column_count
14     FROM
15       pg_stat_user_tables
16     JOIN
17       pg_class ON pg_stat_user_tables.relid = pg_class.oid
18   )
19   SELECT
20     row_number,
21     table_name,
22     row_count,
23     column_count
24   FROM
25     table_counts
26   UNION ALL
27   SELECT
28     NULL AS row_number,
29     'Total' AS table_name,
30     SUM(row_count) AS row_count,
31     NULL AS column_count
32   FROM
33     table_counts
34 )
35 TO 'D:/CUN/Analitica_de_datos_especializacion/db_tables.csv' WITH CSV HEADER;

```

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

El contenido del archivo CSV generado por la anterior consulta SQL se muestra en la siguiente tabla. Son en total 66 tablas, con 660 campos o columnas y que están pobladas por 128.125.976 registros.

Tabla 2

Reporte de conteo de registros por tabla con total consolidado

row_number	table_name	row_count	column_count
1	data_source	128	9
2	spatial_ref_sys	0	5
3	municipalities_dane	1,121	11
4	departments_dane	33	4
5	sexual_crimes_polinal	319,529	8
6	vehicle_theft_polinal	376,433	6
7	theft_from_financial_institutions_mindefensa	2,442	5
8	home_and_business_theft_polinal	669,314	8
9	missing_people_med_legal	186,366	7
10	blowing_up_of_pipelines_mindefensa	1,292	5
11	confiscation_of_marijuana_mindefensa	905,029	6
12	cocaine_confiscations_mindefensa	194,164	6
13	cocaine_base_confiscations_mindefensa	291,808	6
14	basuco_confiscations_mindefensa	439,542	6
15	liquid_inputs_confiscations_mindefensa	99,671	6
16	destroyed_labs_mindefensa	32,470	6
17	eradication_of_illicit_crops_mindefensa	143,224	7
18	land_piracy_mindefensa	9,814	6
19	intervened_mines_mindefensa	11,738	5
20	brigge_blowing_up_mindefensa	1,550	6
21	fatal_injuries_med_legal	221,225	12
22	non_fatal_injuries_med_legal	1,891,147	14
23	crime_news_fiscalia	2,889,477	7
24	victim_count_fiscalia	4,327,402	9
25	indicted_count_fiscalia	4,167,165	9
26	theft_by_modality_polinal	40,824	8
27	weapons_confiscation_polinal	281,479	6
28	domestic_violence_polinal	670,522	7
29	drug_seizure_polinal	1,929,610	6
30	terrorism_crimes_polinal	4,591	6
31	threat_crimes_polinal	412,607	7
32	personal_injury_polinal	924,842	7
33	murders_mindefensa	313,454	7
34	massacres_mindefensa	623	6
35	deaths_in_public_forces_mindefensa	20,229	6

36	theft_from_people_mindefensa	588,117	5
37	kidnapping_mindefensa	27,753	6
38	extortion_mindefensa	85,718	5
39	environmental_crimes_mindefensa	71,362	7
40	arrests_for_illegal_mining_mindefensa	6,423	5
41	terrorism_mindefensa	5,694	6
42	victims_municipal_events_unidadvictimas	4,607,723	12
43	victims_armed_conflict_sievcac	408,581	10
44	victims_by_type_of_crime_unidadvictimas	410,122	13
45	censo_hogares_2018_dane	14,220,487	10
46	censo_viviendas_2018_dane	16,172,455	23
47	cases_armed_conflict_sievcac	349,503	8
48	sisben_iv_hogares_dnp	1,621,798	35
49	sisben_iv_viviendas_dnp	1,446,237	17
50	sisben_iv_personas_dnp	4,465,984	50
51	health_insurance_affiliated_adres	1,263,971	15
52	mortality_and_morbidity_rates_minsalud	265,983	6
53	preschool_primary_secondary_education_mineducacion	12,150,247	17
54	land_restitution_claims_minagricultura	2,160	4
55	familias_en_accion_prosperidad_social	3,958,793	15
56	early_childhood_indicators_mineducacion	10,107	8
57	educational_indicators_mineducacion	14,573	38
58	coca_plantations_minjusticia	319	25
59	censo_2018_personas_dane	43,224,520	22
60	livestock_count_minagricultura	11,125	6
61	water_quality_risk_insalud	1,050	7
62	alluvial_gold_mining_minminas	298	8
63	mobile_coverage_mintic	407,281	13
64	municipal_performance_measure_dnp	22,020	7
65	agricultural_frontier_minagricultura	455,143	6
66	municipal_population_projections_dane	69,564	6
Total		128,125,976	660

Fuente: Construcción propia

La misma información de la tabla anterior pero incluyendo una columna con la descripción del contenido de cada tabla puede consultarse en el anexo 1.

Fuentes de datos

Para rastrear el origen de los datos en las tablas de la base de datos, se utiliza la siguiente consulta SQL sobre la base de datos que permite identificar los conjuntos de datos incorporados y las fuentes de donde se obtuvieron.

Figura 19*Distribución de datasets por fuente de datos*

The screenshot shows a PostgreSQL query results interface in PgAdmin4. At the top, there's a toolbar with icons for copy, paste, refresh, and other database operations. Below the toolbar is a menu bar with 'Query' and 'Query History' tabs. The main area contains a query editor with the following SQL code:

```

1 SELECT origin, COUNT(*) AS count
2 FROM data_source
3 GROUP BY origin
4 UNION ALL
5 SELECT 'Total' AS origin, COUNT(*) AS count
6 FROM data_source
7 ORDER BY count DESC;

```

Below the query editor is a results grid with three tabs: 'Data Output' (selected), 'Messages', and 'Notifications'. The results grid has columns for 'origin' (character varying) and 'count' (bigint). The data is as follows:

origin	count
Total	128
Departamento Administrativo Nacional de Estadísticas - DANE	37
Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVCAC)	22
Ministerio de Defensa Nacional	21
Instituto Nacional de Medicina Legal y Ciencias Forenses	15
Policía Nacional de Colombia	10
Departamento Nacional de Planeación	4
Fiscalía General de la Nación	3
Ministerio de Educación Nacional	3
Unidad para la Atención y Reparación Integral a las Víctimas del Departamento de la Prosperidad Social	2
Administradora de los Recursos del Sistema General de Seguridad - ADRES	2
Unidad de Planificación de Tierras Rurales, Adecuación de Tierras y Usos Agropecuarios	1
Ministerio de Tecnologías de la Información y las Comunicaciones	1
Unidad Administrativa Especial de Gestión de Restitución de Tierras Despojadas	1
Ministerio de Salud y Seguridad Social	1
Ministerio de Justicia y del Derecho	1
Unidad de Planificación Rural Agropecuaria - UPRA de MinAgricultura	1
Departamento Administrativo para la Prosperidad Social	1
Ministerio de Minas y Energía	1
Instituto Nacional de Salud	1

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Los resultados muestran un total de 128 conjuntos de datos procesados, provenientes de 20 fuentes originales. Al analizar la cantidad de conjuntos de datos aportados por cada fuente, se observa que el DANE es el mayor contribuyente, con 37 datasets, lo que representa el 29% del total procesado. Le sigue el Sistema de Información de Eventos de Violencia del

Conflictos Armados en Colombia (SIEVCAC), con 22 datasets (17,2%), el Ministerio de Defensa con 21 (16,4%), Medicina Legal con 15 (11,7%) y la Policía Nacional con 10, equivalente al 7,8%.

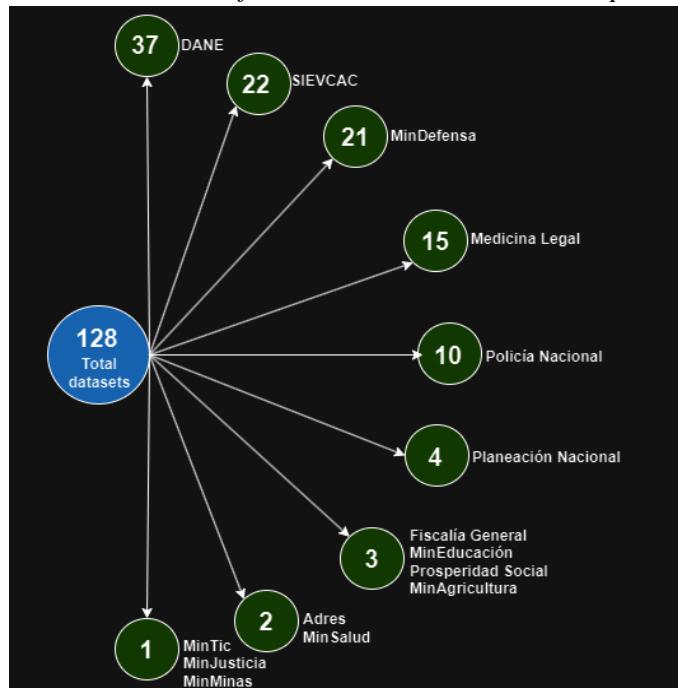
El Departamento Nacional de Planeación aporta 4 datasets (3,1%), mientras que la Fiscalía General de la Nación y el Ministerio de Educación Nacional contribuyen con 3 cada uno. La Unidad para la Atención y Reparación Integral a las Víctimas del Departamento de la Prosperidad Social y la Administradora de Recursos del Sistema de Seguridad Social, aportan 2 datasets cada una. Finalmente, con un dataset cada una, encontramos a: la Unidad de Planificación de Tierras Rurales, Adecuación de Tierras y Usos Agropecuarios; el Ministerio de Tecnologías de la Información y las Comunicaciones; la Unidad Administrativa Especial de Gestión de Restitución de Tierras Despojadas; el Ministerio de Salud y Seguridad Social; el Ministerio de Justicia y del Derecho; la Unidad de Planificación Rural Agropecuaria (UPRA) del Ministerio de Agricultura; el Departamento Administrativo para la Prosperidad Social; el Ministerio de Minas y Energía, y el Instituto Nacional de Salud.

Este análisis ofrece una visión clara de las principales fuentes utilizadas en el proceso de poblamiento de las tablas.

Si clasificamos las fuentes teniendo en cuenta el Ministerio al que pertenecen algunas de las entidades que originaron los datos, tenemos la distribución de fuentes que se aprecia en la gráfica siguiente.

Figura 20

Distribución de las fuentes de datos incluidas en repositorio virtual



Fuente: Construcción propia

Composición temporal de los datos

A partir de los campos start_date y end_date de la tabla data_source de la base de datos se puede describir el rango temporal cubierto por la base de datos, tal como se muestra en la siguiente figura que presenta la consulta SQL aplicada a la base de datos, junto con los resultados obtenidos.

Figura 21

Periodo de Tiempo Cubierto por los Datos

1	SELECT start_date, COUNT(*) AS count
2	FROM data_source
3	GROUP BY start_date
4	ORDER BY start_date;
5	

Data Output		Messages	Notifications

	start_date	count
	date	bigint
1	1930-01-01	1
2	1958-01-01	22
3	1996-01-01	2
4	2001-01-01	1
5	2003-01-01	8
6	2005-01-01	1
7	2007-01-01	2
8	2010-01-01	21
9	2011-01-01	1
10	2015-01-01	1
11	2016-01-01	10
12	2018-01-01	1
13	2018-10-01	33
14	2018-12-31	1
15	2020-01-01	3
16	2020-09-08	1
17	2021-01-01	1
18	2021-11-02	1
19	2022-02-21	1
20	2023-01-01	2
21	2024-04-20	1
22	[null]	13

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Como se observa en los resultados de la anterior consulta SQL hay una fuente que incluye datos desde 1930 hasta 2024 y corresponde a un dataset originado por el Instituto Nacional de Medicina Legal que presenta los datos históricos del Sistema de Información Red de desaparecidos y cadáveres, tal como se muestra en la siguiente consulta.

Figura 22

Dataset con mayor cobertura temporal de datos históricos

```

Query   Query History
1  SELECT
2      origin,
3      SUBSTRING(description FROM 1 FOR 100) AS short_description,
4      start_date,
5      end_date
6  FROM
7      data_source
8  WHERE
9      start_date = '1930-01-01';

Data Output  Messages  Notifications
origin          short_description          start_date          end_date
character varying(255)  text                date              date
1  Instituto Nacional de Medicina Legal y Ciencias Forens... Contiene datos históricos en el Sistema de Información Red de Desaparecidos y Cadáveres. - SIRDEC... 1930-01-01 2024-04-30

```

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

A continuación con el siguiente rango temporal más antiguo están los 22 dataset originados por el Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia -SIEVCAC, que tienen información de víctimas del conflicto armado desde el año 1958 hasta el 2022.

Figura 23

Conjuntos de datos con rango temporal 1958 - 2022

```

1  SELECT
2      origin,
3      SUBSTRING(description FROM 1 FOR 100) AS short_description,
4      start_date,
5      end_date
6  FROM
7      data_source
8  WHERE
9      start_date = '1958-01-01' ;

Data Output  Messages  Notifications
origin          short_description          start_date          end_date
character varying(255)  text                date              date
1  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
2  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
3  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
4  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
5  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
6  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
7  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
8  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
9  Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20
10 Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVC... Contiene microdata del Sistema de Información de Eventos de Violencia del Conflicto Armado en Colo... 1958-01-01 2024-04-20

```

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Siguen dos conjuntos de datos provenientes del Ministerio de Defensa Nacional. Estos datasets contienen información sobre el número de casos de víctimas de secuestro y extorsión, abarcando desde el 1 de enero de 1996 hasta el año 2024, tal como se observa en los resultados de la consulta aplicada a la base de datos.

Figura 24

Conjuntos de datos con rango temporal 1996 - 2024

```

Query   Query History
1  SELECT
2    origin,
3      SUBSTRING(description FROM 1 FOR 100) AS short_description,
4    start_date,
5    end_date
6  FROM
7    data_source
8  WHERE
9    start_date = '1996-01-01';

```

Data Output Messages Notifications

	origin character varying (255)	short_description text	start_date date	end_date date
1	Ministerio de Defensa Nacional	Contiene información sobre el total del número de casos y la relación de víctimas de secuestro simpl...	1996-01-01	2024-04-30
2	Ministerio de Defensa Nacional	Contiene información sobre el casos de extorsión que, según el art. 244 del Código Penal Colombi...	1996-01-01	2024-04-30

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Con un recorrido temporal que va desde el 2001 al 2022, se tiene el dataset originado por el Ministerio de Justicia con información sobre la cuantificación de los cultivos de coca en el País, como se observa en el resultado de la siguiente consulta SQL. Desafortunadamente estos datos no se han actualizado para el 2023 y 2024.

Figura 25

Conjuntos de datos con rango temporal 2001 - 2022

```

Query   Query History
1  SELECT
2    origin,
3      SUBSTRING(description FROM 1 FOR 100) AS short_description,
4    start_date,
5    end_date
6  FROM
7    data_source
8  WHERE
9    start_date = '2001-01-01';

```

Data Output Messages Notifications

	origin character varying (255)	short_description text	start_date date	end_date date
1	Ministerio de Justicia y del Derecho	Contiene información sobre la cuantificación de las hectáreas de los cultivos de coca existentes ...	2001-01-01	2022-12-31

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Con un rango temporal que abarca desde el 2003 al 2024 tenemos 8 conjuntos de datos sobre conteo de diferentes delitos. Siete provienen del Ministerio de Defensa Nacional y uno de la Policía Nacional, como se muestra en la siguiente figura.

Figura 26

Conjuntos de datos con rango temporal 2003 - 2024

1	SELECT						
2	origin,						
3	SUBSTRING(description FROM 1 FOR 100) AS short_description,						
4	start_date,						
5	end_date						
6	FROM						
7	data_source						
8	WHERE						
9	start_date = '2003-01-01';						
Data Output Messages Notifications							
1	Policía Nacional de Colombia	Contiene información del delito de Lesiones personales y Lesiones en accidente de Tránsito desde...	2003-01-01	2024-04-30			
2	Ministerio de Defensa Nacional	Contiene información sobre homicidios. Toda muerte (civiles, miembros de la Fuerza Pública y per...	2003-01-01	2024-04-30			
3	Ministerio de Defensa Nacional	Contiene información sobre aquellos hechos en los cuales resultan muertos cuatro (4) o más pers...	2003-01-01	2023-12-31			
4	Ministerio de Defensa Nacional	Contiene información sobre la modalidad de hurto donde el victimario utiliza diferentes medios co...	2003-01-01	2024-04-30			
5	Ministerio de Defensa Nacional	Contiene información sobre los delitos contra los recursos naturales y el medio ambiente descrito...	2003-01-01	2024-04-30			
6	Ministerio de Defensa Nacional	Contiene información sobre casos de hurto a entidades financieras. Delito definido como: Accione...	2003-01-01	2024-04-30			
7	Ministerio de Defensa Nacional	Contiene información sobre piratería terrestre: actividad delictiva encaminada al hurto de mercancía	2003-01-01	2024-04-30			
8	Ministerio de Defensa Nacional	Contiene información sobre voladuras de puentes y vías. Vías públicas y puentes vehiculares públ...	2003-01-01	2024-04-30			

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Con rango temporal que va desde el 1 de enero de 2007 al 30 de abril del 2024, se procesaron dos dataset del Ministerio de Defensa Nacional que contienen información sobre voladura de oleoductos y erradicación de cultivos ilícitos, como se muestra en la siguiente figura.

Figura 27

Conjuntos de datos con rango temporal 2007 - 2024

1	SELECT						
2	origin,						
3	SUBSTRING(description FROM 1 FOR 100) AS short_description,						
4	start_date,						
5	end_date						
6	FROM						
7	data_source						
8	WHERE						
9	start_date = '2007-01-01';						
Data Output Messages Notifications							
1	Ministerio de Defensa Nacional	Contiene información sobre la afectación a la infraestructura crítica: voladura de oleoductos. ...	2007-01-01	2024-04-30			
2	Ministerio de Defensa Nacional	Contiene información sobre erradicación de cultivos ilícitos que incluye la eliminación física o ...	2007-01-01	2024-04-30			

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Conjuntos de datos con una cobertura temporal del 2010 al 2024 se gestionaron 21 en la base de datos, 9 generados por la Policía Nacional, 3 por la Fiscalía General de la Nación y 9 por el Ministerio de Defensa Nacional. En la siguiente figura visualiza una muestra de los mismos.

Figura 28

Muestra de conjuntos de datos con rango temporal 2010 - 2024

<pre> 1 SELECT 2 origin, 3 SUBSTRING(description FROM 1 FOR 100) AS short_description, 4 start_date, 5 end_date 6 FROM 7 data_source 8 WHERE 9 start_date = '2010-01-01'; </pre>					
	origin character varying (255)	short_description text	start_date date	end_date date	
1	Policía Nacional de Colombia	Contiene información de Incautación de Armas de Fuego desde 01 de enero del año 2010 al 31 de en...	2010-01-01	2024-04-30	
2	Policía Nacional de Colombia	Contiene información del delito de hurto en Colombia a través de las modalidades de abigeato, bancos	2010-01-01	2024-04-30	
3	Policía Nacional de Colombia	Contiene información del delito de hurto en Colombia a través de las modalidades de residencias y en	2010-01-01	2024-04-30	
4	Policía Nacional de Colombia	Contiene información del delito Amenazas del 01 de enero del año 2010 al 30 de abril del año 2024	2010-01-01	2024-04-30	
5	Fiscalía General de la Nación	Contiene noticias criminales por delito registrados en el Sistema Penal Oral Acusatorio en la Ley 90	2010-01-01	2024-09-02	
6	Fiscalía General de la Nación	Contiene total de víctimas según las entradas de noticias criminales por delito al Sistema Penal Ora	2010-01-01	2024-09-02	
7	Fiscalía General de la Nación	Contiene total de Indicados por delito según entradas de noticias criminales al Sistema Penal Oral	2010-01-01	2024-09-02	

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Con periodo que abarca del 2011 al 2024 y del 2015 al 2024 se gestionaron dos dataset originados por el Ministerio de Educación Nacional que contienen respectivamente indicadores de educación prescolar, básica y media e información estadística de primera infancia.

Figura 29

Conjuntos de datos con rango temporal 2011 – 2024 y 2015 -2024

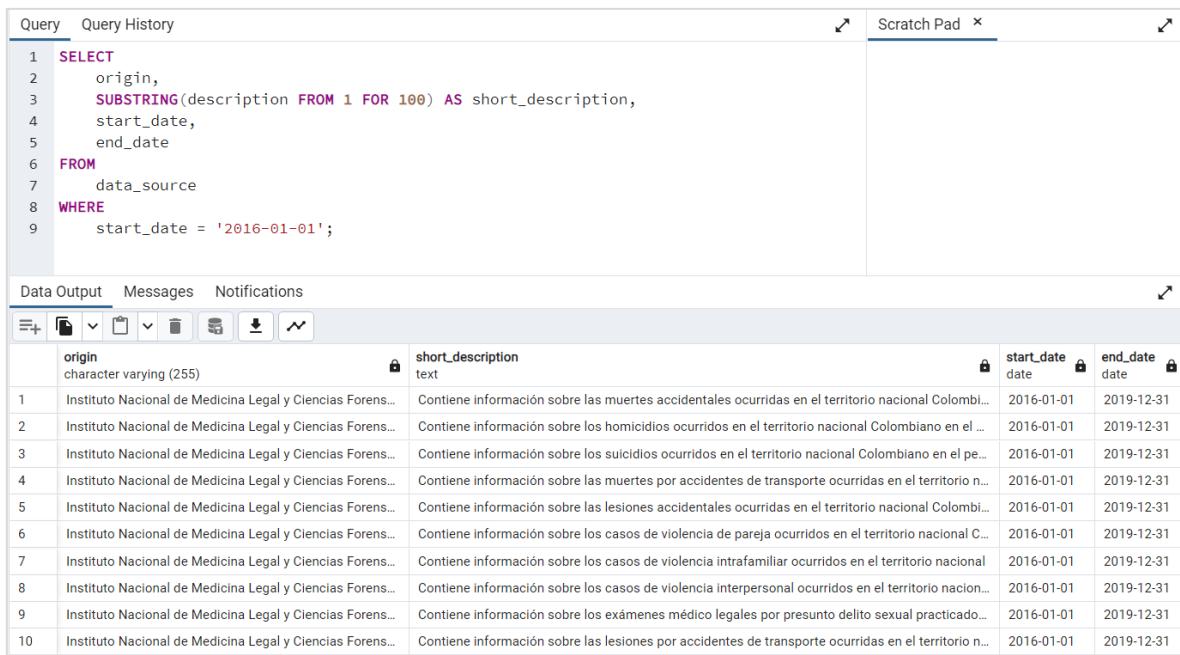
<pre> 1 SELECT 2 origin, 3 SUBSTRING(description FROM 1 FOR 100) AS short_description, 4 start_date, 5 end_date 6 FROM 7 data_source 8 WHERE 9 start_date = '2011-01-01' OR start_date = '2015-01-01' ; </pre>					
	origin character varying (255)	short_description text	start_date date	end_date date	
1	Ministerio de Educacion Nacional	Contiene información estadística de primera infancia a relacionada con indicadores de (EDUCACI...)	2015-01-01	2024-08-15	
2	Ministerio de Educacion Nacional	Contiene información estadística (indicadoares educativos) de los niveles preescolar, básica y me...	2011-01-01	2024-08-15	

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Se cuenta con 10 conjuntos de datos cuya cobertura temporal abarca desde el 1 de enero de 2016 hasta el 31 de diciembre de 2019. Todos estos datos provienen del Instituto Nacional de Medicina Legal y Ciencias Forenses. Aunque inicialmente parece que estos datasets no cumplen con el requisito de cubrir al menos el periodo 2019-2024, se identificó que Medicina Legal publica sus estadísticas en periodos específicos: 2016-2019, 2020-2022 y 2023-2024. En particular, las estadísticas del periodo 2016-2019 se encontraban en un formato desagregado, lo que hizo necesario un proceso de integración vertical. Esta integración permitió unificar los datos de manera coherente antes de incorporarlos al repositorio, logrando así que la cobertura se extendiera hasta el año 2024.

Figura 30

Conjuntos de datos con rango temporal 2016-2019



The screenshot shows the PgAdmin4 interface with two main panes. The top pane is the Query History tab, containing the following SQL code:

```

1 SELECT
2     origin,
3     SUBSTRING(description FROM 1 FOR 100) AS short_description,
4     start_date,
5     end_date
6 FROM
7     data_source
8 WHERE
9     start_date = '2016-01-01';

```

The bottom pane is the Data Output tab, displaying the results of the query. The results are presented in a table with four columns: origin, short_description, start_date, and end_date. The data consists of 10 rows, each representing a different type of death or injury recorded by the Instituto Nacional de Medicina Legal y Ciencias Forenses. The start_date for all rows is 2016-01-01, and the end_date is 2019-12-31.

origin	short_description	start_date	end_date
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre las muertes accidentales ocurridas en el territorio nacional Colomb...	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre los homicidios ocurridos en el territorio nacional Colombiano en el ...	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre los suicidios ocurridos en el territorio nacional Colombiano en el pe...	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre las muertes por accidentes de transporte ocurridas en el territorio n...	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre las lesiones accidentales ocurridas en el territorio nacional Colomb...	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre los casos de violencia de pareja ocurridos en el territorio nacional C...	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre los casos de violencia intrafamiliar ocurridos en el territorio nacional	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre los casos de violencia interpersonal ocurridos en el territorio nacion...	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre los exámenes médico legales por presunto delito sexual practicado...	2016-01-01	2019-12-31
Instituto Nacional de Medicina Legal y Ciencias Forens...	Contiene información sobre las lesiones por accidentes de transporte ocurridas en el territorio n...	2016-01-01	2019-12-31

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

La información sobre matricula de educación prescolar, básica y media de la población colombiana distribuida Institución educativa y municipio abarca el espacio temporal comprendido entre el 2018 a diciembre de 2023, como se aprecia en el resultado de la siguiente consulta SQL.

Figura 31*Conjunto de datos con rango temporal 2018-2023*


The screenshot shows a PostgreSQL client interface (PgAdmin4). In the top-left corner, there are tabs for 'Query' and 'Query History'. Below these, a code editor displays the following SQL query:

```

1  SELECT
2      origin,
3      SUBSTRING(description FROM 1 FOR 100) AS short_description,
4      start_date,
5      end_date
6  FROM
7      data_source
8  WHERE
9      start_date = '2018-01-01' ;
10

```

Below the code editor, there are three tabs: 'Data Output', 'Messages', and 'Notifications'. The 'Data Output' tab is selected and shows a table with the following data:

	origin	short_description	start_date	end_date
1	Ministerio de Educacion Nacional	Contiene información sobre la matrícula estadística de Educación Preescolar Básica y Media de Colo...	2018-01-01	2023-12-31

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Los 33 datasets que abarcan el rango temporal de 2018 corresponden a los microdatos del Censo de Población y Vivienda, con información detallada de los hogares y viviendas en cada uno de los departamentos de Colombia, más Bogotá. Estos conjuntos de datos se encontraron desagregados por departamento en el portal del Dane. Para su integración en la base de datos, se aplicó un proceso de consolidación mediante una integración vertical, lo que permitió unificar los registros en dos tablas: una que agrupa las variables de los hogares y otra que almacena las de las viviendas, ya que ambos conjuntos contienen variables distintas que ofrecen una visión complementaria sobre la situación social y económica de la población.

La tabla de viviendas incluye variables como el tipo de vivienda, condiciones de ocupación, materiales predominantes en paredes y pisos, acceso a servicios básicos como energía eléctrica, acueducto y gas natural, así como la presencia de servicios sanitarios y acceso a Internet. Estos datos permiten evaluar las condiciones habitacionales de los colombianos en distintas áreas del país.

Por otro lado, la tabla de hogares recoge variables como el número total de cuartos en la vivienda, el número de cuartos destinados al descanso, las fuentes de agua para la preparación de alimentos, y el número de personas residentes. Estos microdatos permiten analizar las características internas de los hogares y las condiciones de vida de las personas dentro de cada unidad familiar, proporcionando un panorama detallado de las estructuras y dinámicas del hogar en diferentes regiones del país. Si consultamos a la base de datos los datasets que contienen información a partir de 2018, obtendremos lo siguiente:

Figura 32*Conjuntos de datos gestionados con microdatos del Censo de 2018*

```

1 SELECT
2     origin,
3     SUBSTRING(description FROM 1 FOR 100) AS short_description,
4     start_date,
5     end_date
6 FROM
7     data_source
8 WHERE
9     start_date = '2018-10-01';

```

Data Output Messages Notifications

	origin character varying (255)	short_description text	start_date date	end_date date
1	Departamento Administrativo Nacional de Estadísticas - DA...	Contiene los microdatos de los hogares y viviendas del territorio nacional del censo de población...	2018-10-01	2024-08-31
2	Departamento Administrativo Nacional de Estadísticas - DA...	Contiene los microdatos de los hogares y viviendas del territorio nacional del censo de población...	2018-10-01	2024-08-31
3	Departamento Administrativo Nacional de Estadísticas - DA...	Contiene los microdatos de los hogares y viviendas del territorio nacional del censo de población...	2018-10-01	2024-08-31
4	Departamento Administrativo Nacional de Estadísticas - DA...	Contiene los microdatos de los hogares y viviendas del territorio nacional del censo de población...	2018-10-01	2024-08-31
5	Departamento Administrativo Nacional de Estadísticas - DA...	Contiene los microdatos de los hogares y viviendas del territorio nacional del censo de población...	2018-10-01	2024-08-31
6	Departamento Administrativo Nacional de Estadísticas - DA...	Contiene los microdatos de los hogares y viviendas del territorio nacional del censo de población...	2018-10-01	2024-08-31
7	Departamento Administrativo Nacional de Estadísticas - DA...	Contiene los microdatos de los hogares y viviendas del territorio nacional del censo de población...	2018-10-01	2024-08-31

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Los 13 dataset que aparentemente carecen de temporalidad, ya que no presentan una fecha en el campo start_date, se muestran en la siguiente gráfica. Estos conjuntos pertenecen al Dane y fueron utilizados para obtener la información de las tablas de departamentos y municipios. Los otros conjuntos corresponden a reportes "con corte a", donde generalmente no se especifica cuándo ocurrió el evento. Entre estas están los reportes de afiliados a seguridad social, los reportes de beneficiarios de familias en acción, los reportes del SISBEN IV, etc.

Figura 33*Conjuntos de datos que figuran sin fecha de inicio*

```

1 SELECT
2     origin,
3     SUBSTRING(description FROM 1 FOR 100) AS short_description,
4     start_date,
5     end_date
6 FROM
7     data_source
8 WHERE
9     start_date IS NULL ;

```

Data Output Messages Notifications

	origin character varying (255)	short_description text	start_date date	end_date date
1	Departamento Administrativo Nacional de Estadísticas - DANE	Este archivo proporciona una detallada descripción de los centros poblados en Colombia, incluyendo...	[null]	[null]
2	Departamento Administrativo Nacional de Estadísticas - DANE	Este archivo proporciona una descripción completa de los municipios colombianos, incluyendo sus...	[null]	[null]
3	Unidad para la Atención y Reparación Integral a las Víctimas del Departamento de la Prosperidad So...	Contiene información sobre cifras de Víctimas de 32 reportes con las siguientes fechas de corte: 2...	[null]	2024-08-31
4	Unidad para la Atención y Reparación Integral a las Víctimas del Departamento de la Prosperidad So...	Contiene información sobre cifras de Víctimas del conflicto armado por tipo de hecho o tipo de crí...	[null]	2024-08-31
5	Administradora de los Recursos del Sistema General de Seguridad - ADRES	Contiene la información la población activa del régimen contributivo que se encuentra almacenada...	[null]	2024-09-06
6	Administradora de los Recursos del Sistema General de Seguridad - ADRES	Contiene la información la población activa del régimen subsidiado que se encuentra almacenada ...	[null]	2024-09-06
7	Departamento Administrativo para la Prosperidad Social	Contiene datos anonimizados del programa Más Familias en Acción, determinando características ...	[null]	2024-07-31
8	Unidad Administrativa Especial de Gestión de Restitución de Tierras Despojadas	Contiene el conteo de las solicitudes de restitución de tierras discriminadas por departamento, mun...	[null]	2024-10-03
9	Unidad de Planificación Rural Agropecuaria - UPRA de MinAgricultura	Contiene información de la EVA (una investigación que se realiza mediante un mecanismo subjetiv...	[null]	2024-03-26
10	Unidad de Planificación de Tierras Rurales, Adecuación de Tierras y Usos Agropecuarios	Contiene información sobre la identificación de la frontera agrícola y frontera agrícola condicionad...	[null]	2024-10-08
11	Departamento Nacional de Planeación	Contiene la información de hogares de la base de datos del Sisbén IV, del Departamento Nacional d...	[null]	2024-04-20
12	Departamento Nacional de Planeación	Contiene la información de viviendas de la base de datos del Sisbén IV, del Departamento Nacional ...	[null]	2024-04-20
13	Departamento Nacional de Planeación	Contiene la información de personas de la base de datos del Sisbén IV, del Departamento Nacional ...	[null]	2024-04-20

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Campos y tipos de datos

Para obtener una visión general de la distribución de los tipos de datos en las columnas de las tablas del esquema public de la base de datos, podemos realizar la siguiente consulta sql:

Figura 34

Consulta SQL para obtener la distribución de tipos datos en el repositorio

```
Query   Query History
1 COPY (
2   SELECT
3     data_type,
4     COUNT(*) AS type_count
5   FROM
6     information_schema.columns
7   WHERE
8     table_schema = 'public'
9   GROUP BY
10    data_type
11   ORDER BY
12    type_count DESC
13 )
14 TO 'D:/CUN/Analitica_de_datos_especializacion/data_types_summary.csv'
15 WITH CSV HEADER;
```

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Los resultados se presentan en la tabla siguiente y ofrecen una vista clara de los tipos de datos predominantes en la base de datos.

Tabla 3

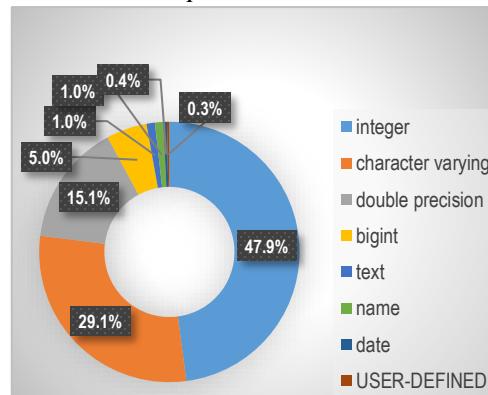
Tipos de datos en el repositorio

data_type	type_count
Integer	323
character varying	196
double precision	102
Bigint	34
Text	7
Name	7
Date	3
USER-DEFINED	2

Fuente: Construcción propia

Figura 35

Distribución tipos de datos



Los tipos de datos integer y character varying son los tipos de datos más utilizados en el esquema público, representando el 77% de las columnas. Los de double precision y bigint son usados en menor medida, principalmente para valores numéricos más específicos o de

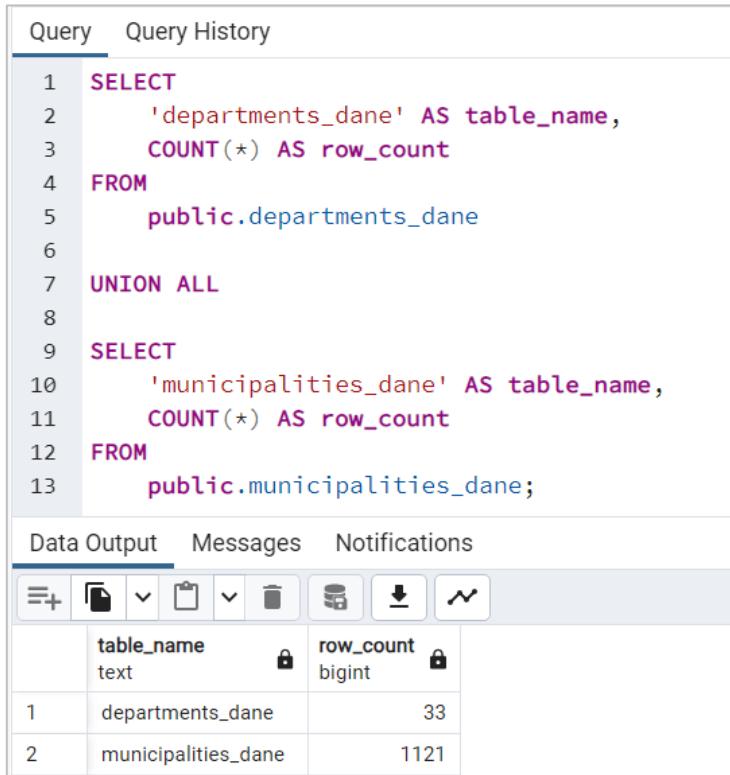
gran tamaño, en conjunto representan el 20% de los campos. Los tipos de datos como text, name, y date aparecen ocasionalmente en el esquema, conjuntamente representan el 5% del total. Los tipos de datos USER-DEFINED, que indican tipos de datos personalizados, para este caso en particular, probablemente corresponden a las columnas que almacenan las geometrías de georeferencia incluidas en la tabla municipalities, dado que no se ha definido explícitamente ningún tipo de dato personalizado, pero sí se tiene instalado PostGIS, que introduce tipos como geometry.

Distribución geográfica de los datos

La totalidad de los departamentos y municipios de Colombia están representados en la base de datos.

Figura 36

No. de departamentos y municipios con información en la bd



```

Query   Query History
1 SELECT
2     'departments_dane' AS table_name,
3     COUNT(*) AS row_count
4 FROM
5     public.departments_dane
6
7 UNION ALL
8
9 SELECT
10    'municipalities_dane' AS table_name,
11    COUNT(*) AS row_count
12 FROM
13    public.municipalities_dane;

```

Data Output		Messages	Notifications			
	table_name	row_count				
1	departments_dane	33				
2	municipalities_dane	1121				

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

En la base de datos se almacenan datos geoespaciales utilizando PostGIS, específicamente en la tabla municipalities_dane, que contiene dos columnas clave: geometry (que almacena los polígonos de los municipios) y mupal_head (que guarda los puntos de las cabeceras municipales). Aunque solo esta tabla tiene datos geográficos, gracias al código DANE de

cada municipio mapeado en todas las tablas de datos, estas son vinculadas geográficamente al territorio. Esto permite que cualquier dato en la base de datos pueda ser georreferenciado y representado en mapas para análisis espacial.

1.4 Calidad y consistencia de los datos incorporados

El proceso de extracción y transformación de datos para poblar la base de datos presentó resultados positivos en términos de la calidad de los datos incorporados. Siguiendo los criterios establecidos en la etapa de exploración web, se lograron integrar datasets alineados con los objetivos del proyecto, garantizando una cobertura geográfica y temporal adecuada para los municipios colombianos.

Los datasets seleccionados estaban bien estructurados y en formatos estándar (CSV, Excel), facilitando su manipulación y transformación con las herramientas utilizadas en Python. Además, se verificó que todos los datasets incluyeran identificadores consistentes, particularmente el código DANE, lo cual permitió su integración eficiente con los datos georreferenciados de la tabla municipalities_dane.

Uno de los aspectos críticos fue la vinculación correcta de los datos con su ubicación geográfica oficial mediante el código DANE. Los datasets que ya contenían el código DANE se integraron sin mayores inconvenientes, solo haciendo el mapeo a los códigos reales que ya se tenía en la base de datos en la tabla municipalities_dane. En los casos donde no estaba disponible, se implementó un proceso de similitud difusa (fuzzy matching), lo cual permitió una coincidencia precisa basada en los nombres de los municipios y departamentos. Este método probó ser eficaz en la mayoría de los casos, con un bajo margen de error en la asignación de códigos.

Los datos incorporados cumplieron con los requisitos temporales del proyecto, con un énfasis particular en el periodo 2019-2024. A excepción de los dataset que se manejan como reporte ‘con corte a’, donde generalmente solo se contaba con el último reporte, que fue el que se incorporó a la base de datos previa las transformaciones que hemos documentado en este informe.

1.5 Validación de la Integridad de los Datos

Durante el proceso de carga de los datasets a la base de datos en PostgreSQL, se implementaron mecanismos de control de calidad que incluyeron la revisión de tipos de datos y formatos de las columnas, asegurando su correspondencia con la estructura de las tablas en la base de datos. Las verificaciones poscarga mostraron que los datos se transferían

correctamente, sin alteraciones en los valores y con todos los registros georreferenciados adecuadamente.

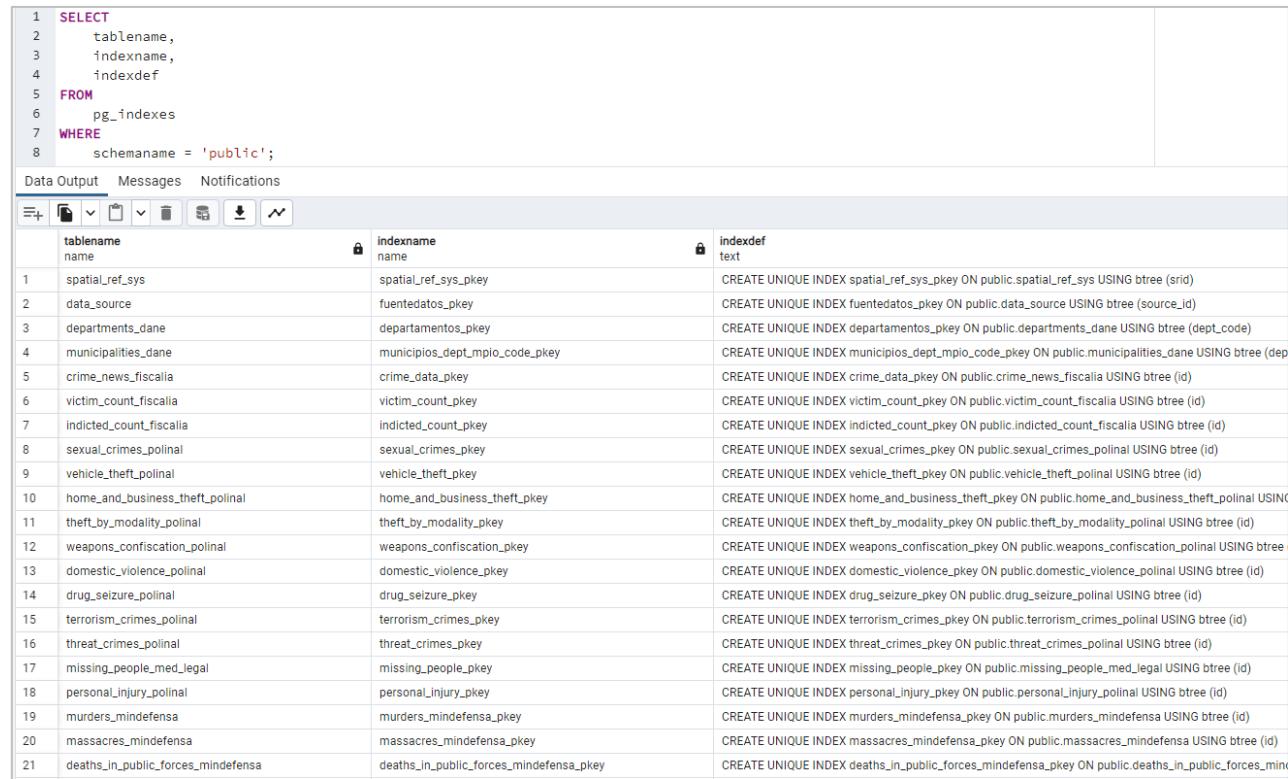
Como parte del preprocesamiento, se manejaron los valores faltantes utilizando métodos de imputación, asegurando que no se comprometiera la integridad de los análisis posteriores. También se identificaron y eliminaron registros duplicados, cuando se encontraron, garantizando así que los datos fueran consistentes y no redundantes.

Se llevaron a cabo consultas SQL para validar la integridad de los datos cargados. Estas consultas incluyeron la revisión de la correspondencia entre los códigos DANE y las geometrías municipales, así como el conteo de registros y la comparación de los totales de poblaciones y otros indicadores clave entre los datasets, asegurando que las relaciones y los cálculos derivados fueran coherentes.

El proceso ETL fue iterativo y permitió realizar mejoras constantes en la calidad de los datos. A medida que se incorporaban nuevos datasets, estos se validaban de acuerdo con los mismos criterios, logrando una base de datos robusta y escalable, preparada para adaptarse a futuros conjuntos de datos sin comprometer la consistencia general.

1.6 Optimización y eficiencia

En términos de optimización del rendimiento, PostgreSQL automáticamente crea índices sobre las llaves primarias de cada tabla durante la creación de la base de datos. Estos índices son estructuras de datos que permiten acelerar las consultas de selección y filtrado basadas en las claves primarias, mejorando significativamente el tiempo de respuesta para operaciones como la búsqueda y actualización de registros. Estos índices se generaron sin intervención manual y proporcionan un acceso eficiente a los datos al garantizar que las búsquedas directas por clave primaria se realicen en tiempo casi constante. Esta funcionalidad predeterminada es fundamental para mantener la eficiencia en la manipulación de grandes volúmenes de datos, especialmente en las consultas que involucran múltiples tablas relacionadas (Juba & Volkov, 2021).

Figura 37*Indices sobre llaves primarias construidos automáticamente por Postgress*


The screenshot shows a PostgreSQL database interface in PgAdmin4. A SQL query window at the top displays:

```

1 SELECT
2     tablename,
3     indexname,
4     indexdef
5 FROM
6     pg_indexes
7 WHERE
8     schemaname = 'public';

```

Below the query window is a table titled "Data Output" showing the results of the query. The table has three columns: "tablename", "indexname", and "indexdef". The "indexdef" column contains the CREATE INDEX statements for each index.

tablename	indexname	indexdef
spatial_ref_sys	spatial_ref_sys_pkey	CREATE UNIQUE INDEX spatial_ref_sys_pkey ON public.spatial_ref_sys USING btree (srdf)
data_source	fuentedatos_pkey	CREATE UNIQUE INDEX fuentedatos_pkey ON public.data_source USING btree (source_id)
departments_dane	departamentos_pkey	CREATE UNIQUE INDEX departamentos_pkey ON public.departments_dane USING btree (dept_code)
municipalities_dane	municipios_dept_mpio_code_pkey	CREATE UNIQUE INDEX municipios_dept_mpio_code_pkey ON public.municipalities_dane USING btree (dep
crime_news_fiscalia	crime_data_pkey	CREATE UNIQUE INDEX crime_data_pkey ON public.crime_news_fiscalia USING btree (id)
victim_count_fiscalia	victim_count_pkey	CREATE UNIQUE INDEX victim_count_pkey ON public.victim_count_fiscalia USING btree (id)
indicted_count_fiscalia	indicted_count_pkey	CREATE UNIQUE INDEX indicted_count_pkey ON public.indicted_count_fiscalia USING btree (id)
sexual_crimes_polinal	sexual_crimes_pkey	CREATE UNIQUE INDEX sexual_crimes_pkey ON public.sexual_crimes_polinal USING btree (id)
vehicle_theft_polinal	vehicle_theft_pkey	CREATE UNIQUE INDEX vehicle_theft_pkey ON public.vehicle_theft_polinal USING btree (id)
home_and_business_theft_polinal	home_and_business_theft_pkey	CREATE UNIQUE INDEX home_and_business_theft_pkey ON public.home_and_business_theft_polinal USING
theft_by_modality_polinal	theft_by_modality_pkey	CREATE UNIQUE INDEX theft_by_modality_pkey ON public.theft_by_modality_polinal USING btree (id)
weapons_confiscation_polinal	weapons_confiscation_pkey	CREATE UNIQUE INDEX weapons_confiscation_pkey ON public.weapons_confiscation_polinal USING btree
domestic_violence_polinal	domestic_violence_pkey	CREATE UNIQUE INDEX domestic_violence_pkey ON public.domestic_violence_polinal USING btree (id)
drug_seizure_polinal	drug_seizure_pkey	CREATE UNIQUE INDEX drug_seizure_pkey ON public.drug_seizure_polinal USING btree (id)
terrorism_crimes_polinal	terrorism_crimes_pkey	CREATE UNIQUE INDEX terrorism_crimes_pkey ON public.terrorism_crimes_polinal USING btree (id)
threat_crimes_polinal	threat_crimes_pkey	CREATE UNIQUE INDEX threat_crimes_pkey ON public.threat_crimes_polinal USING btree (id)
missing_people_med_legal	missing_people_pkey	CREATE UNIQUE INDEX missing_people_pkey ON public.missing_people_med_legal USING btree (id)
personal_injury_polinal	personal_injury_pkey	CREATE UNIQUE INDEX personal_injury_pkey ON public.personal_injury_polinal USING btree (id)
murders_mindefensa	murders_mindefensa_pkey	CREATE UNIQUE INDEX murders_mindefensa_pkey ON public.murders_mindefensa USING btree (id)
massacres_mindefensa	massacres_mindefensa_pkey	CREATE UNIQUE INDEX massacres_mindefensa_pkey ON public.massacres_mindefensa USING btree (id)
deaths_in_public_forces_mindefensa	deaths_in_public_forces_mindefensa_pkey	CREATE UNIQUE INDEX deaths_in_public_forces_mindefensa_pkey ON public.deaths_in_public_forces_min

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

Adicionalmente, para optimizar aún más el rendimiento de las consultas que involucran relaciones entre tablas, se implementaron índices sobre las llaves foráneas. Aunque PostgreSQL no crea estos índices de forma automática, su inclusión se hizo para mejorar la velocidad de las consultas particularmente las que involucran JOINs entre tablas relacionadas a través de las llaves foráneas. Estos índices reducen la cantidad de tiempo necesario para buscar coincidencias entre tablas, disminuyendo los tiempos de ejecución de las consultas complejas que integran datos de múltiples fuentes (Rigg, 2022). Así, se logra una mayor eficiencia en la ejecución de las operaciones ETL y en las consultas analíticas posteriores, asegurando un desempeño robusto del sistema.

Figura 38

Script sql para creación de índices en las llaves foráneas de la base de datos

```

1 DO
2 $$ 
3 DECLARE
4     r RECORD;
5 BEGIN
6     FOR r IN
7         SELECT
8             tc.table_schema,
9             tc.table_name,
10            kcu.column_name,
11            ccu.table_name AS foreign_table_name,
12            ccu.column_name AS foreign_column_name
13        FROM information_schema.table_constraints AS tc
14        JOIN information_schema.key_column_usage AS kcu
15        ON tc.constraint_name = kcu.constraint_name
16        AND tc.table_schema = kcu.table_schema
17        JOIN information_schema.constraint_column_usage AS ccu
18        ON ccu.constraint_name = tc.constraint_name
19        WHERE tc.constraint_type = 'FOREIGN KEY'
20    LOOP
21        EXECUTE format('CREATE INDEX IF NOT EXISTS idx_%I_%I ON %I.%I (%I);',
22                         r.table_name, r.column_name, r.table_schema, r.table_name, r.column_name);
23    END LOOP;
24 END
25 $$;
```

Data Output Messages Notifications

```

NOTICE: relation "idx_cocaine_base_confiscations_mindefensa_source_id" already exists, skipping
```

Fuente: Pantallazo obtenido desde la aplicación PgAdmin4

1.7 Diccionario de Datos

El diccionario de datos se elaboró con el propósito de documentar de manera detallada la estructura de la base, definiendo claramente el significado, el tipo de datos, las relaciones y las restricciones de cada campo, así como los índices utilizados para optimizar el acceso a los datos. Este recurso sirve para garantizar la trazabilidad y la comprensión de los datos por parte de los usuarios, y facilita futuras expansiones y el mantenimiento del sistema (Juba & Volkov, 2021).

El proceso de creación de este diccionario implicó una revisión de todas las tablas, asegurando que cada campo estuviera correctamente documentado. Dado el volumen de la base de datos y la diversidad de los conjuntos de datos que almacena, esta documentación

fue un componente importante para asegurar la consistencia y coherencia en el manejo de los datos. El diccionario completo de la base de datos se incluye en el Anexo 3, donde se puede consultar en detalle la estructura de cada tabla, los campos que la componen, y las relaciones entre ellas.

1.8 Propuestas de mejora y evolución

Con el objetivo de garantizar la sostenibilidad y escalabilidad a largo plazo de la base de datos, se plantean varias propuestas de mejora que optimizarían su rendimiento y permitirían una evolución acorde con las demandas futuras del sistema.

Automatización del proceso de actualización. Uno de los principales retos de una base de datos de gran escala es mantener actualizados los datos sin requerir intervención manual constante. Para abordar este desafío, se propone la implementación de un pipeline ETL automatizado que ejecute actualizaciones periódicas de forma programada. Esto permitiría reducir errores humanos, optimizar el tiempo de actualización y garantizar que la base de datos siempre contenga la información más reciente, contribuyendo a su estabilidad y confiabilidad a largo plazo.

Evolución hacia un Data Warehouse. Dado el volumen de datos que ya contiene la base de datos y ante la perspectiva de que sigan creciendo, se hace evidente la necesidad de evolucionar hacia un modelo de Data Warehouse, que permita gestionar eficientemente grandes volúmenes de datos históricos y proporcionar un entorno robusto para el análisis avanzado de datos mediante operaciones OLAP. La migración hacia este modelo también mejoraría la capacidad de respuesta del sistema y aumentaría su escalabilidad, preparándolo para un crecimiento exponencial en los datos.

2. Resultados y discusiones de la segunda etapa del proyecto

En esta parte del informe se procede a informar sobre los resultados de la segunda etapa del proyecto, la aplicación de técnicas no supervisadas de Machine Learning sobre una muestra de datos extraída de la base de datos, específicamente sobre el departamento del Cauca. El objetivo de esta fase es identificar patrones, tendencias y anomalías en los datos que puedan contribuir al análisis socioeconómico y de violencia en los municipios del Cauca durante el periodo de los últimos años.

La etapa comienza con la implementación de un proceso de ETL, Extract, Transform, Load por sus siglas en inglés, que permita extraer los datos pertinentes de la base de datos, transformarlos para su adecuada estructuración y limpieza, y cargarlos en un entorno adecuado para su posterior análisis mediante Machine Learning. El enfoque en esta fase del ETL se centra en asegurar la coherencia y completitud del dataset para garantizar que el conjunto de datos esté preparado para la aplicación de técnicas avanzadas de análisis.

2.1 Resultado del enfoque escalonado para la extracción de datos de la base de datos

El proceso de extracción de datos se realizó mediante un enfoque escalonado, lo que permitió gestionar de manera eficiente la gran cantidad de información almacenada en la base de datos, la cual, como ya se ha documentado, contiene 66 tablas con más de 128 millones de registros. Este enfoque consistió en dividir las consultas SQL en partes más manejables y específicas, para lo cual se diseñó un total de 52 consultas que permitieron extraer los microdatos y agruparlos a nivel de municipio para los años 2019-2023. Cada consulta se enfocó en obtener información temática clave —como indicadores socioeconómicos, datos de violencia y geometrías georreferenciadas—, siempre agrupando los datos por municipio y año. La intención fue la de capturar la mayor cantidad posible de información relevante desde la base de datos de los municipios del departamento del Cauca.

Este enfoque escalonado evitó los riesgos asociados a consultas excesivamente complejas, que pudieran terminar en tiempos de respuesta inaceptablemente largos o resultados inconsistentes debido a la complejidad sintáctica de las uniones (JOINS) y agregaciones entre tablas. Las consultas fueron cuidadosamente optimizadas para minimizar el tiempo de procesamiento y garantizar la coherencia de los datos extraídos. Todas las consultas SQL generadas se organizaron en un diccionario de Python y, mediante un script ejecutado en VSCode, se automatizó la extracción de los datos en 52 archivos CSV, mediante el script que se muestra en la figura de la siguiente página.

Este enfoque no solo simplificó el proceso, sino que también permitió una mayor consistencia y trazabilidad en la extracción, asegurando que cada archivo estuviera correctamente estructurado y alineado con los objetivos de la fase de análisis.

Figura 39

Script Python para procesar las consultas sql y generar los archivo csv

```

1 import psycopg2
2 import pandas as pd
3 import os
4 from consultas_para_dataset_final import consultas # Importa el diccionario de consultas
5
6 # Configuración de la conexión
7 conn_params = {
8     "host": "localhost",
9     "port": "5432",
10    "database": "db_proyecto",
11    "user": "postgres",
12    "password": "cauca"
13 }
14 # Directorio de salida
15 output_dir = 'D:/CUN/Analitica_de_datos_especializacion/proyecto_grado/ejecucion_proyecto/CSV'
16
17 # Asegurar que el directorio de salida exista
18 if not os.path.exists(output_dir):
19     os.makedirs(output_dir)
20
21 # Manejo de la conexión
22 try:
23     with psycopg2.connect(**conn_params) as conn:
24         for consulta_nombre, query in consultas.items():
25             try:
26                 # Ejecutar la consulta
27                 df = pd.read_sql_query(query, conn)
28                 print(f"Consulta '{consulta_nombre}' ejecutada exitosamente.")
29
30                 # Definir el nombre del archivo de salida
31                 output_file = os.path.join(output_dir, f'{consulta_nombre}.csv')
32
33                 # Guardar resultado de la consulta en archivo CSV
34                 df.to_csv(output_file, index=False)
35                 print(f"Archivo guardado en: {output_file}")
36
37             except Exception as e:
38                 print(f"Error al ejecutar la consulta '{consulta_nombre}': {e}")
39                 continue # Continuar con la siguiente consulta si ocurre un error
40     except Exception as e:
41         print(f"Error al conectar con la base de datos: {e}")
42     raise

```

Fuente: Pantallazo obtenido desde la aplicación Vscode

Es importante destacar que, aunque algunos conjuntos de datos incluían información parcial para el año 2024, se decidió no incorporarlos en esta fase, dado que su inclusión podría haber afectado la precisión del análisis al tratarse de un periodo incompleto. Por ello, previa consulta con el tutor del proyecto se decidió dejar por fuera del análisis el año del 2024. En los casos en los que no se disponía de información para todos los años (2019-2023), se tomó

la decisión de replicar los datos de años cercanos, aunque esta decisión es revisada durante la etapa de transformación de los datos.

2.2 Integración de los archivos CSV en un solo DataFrame

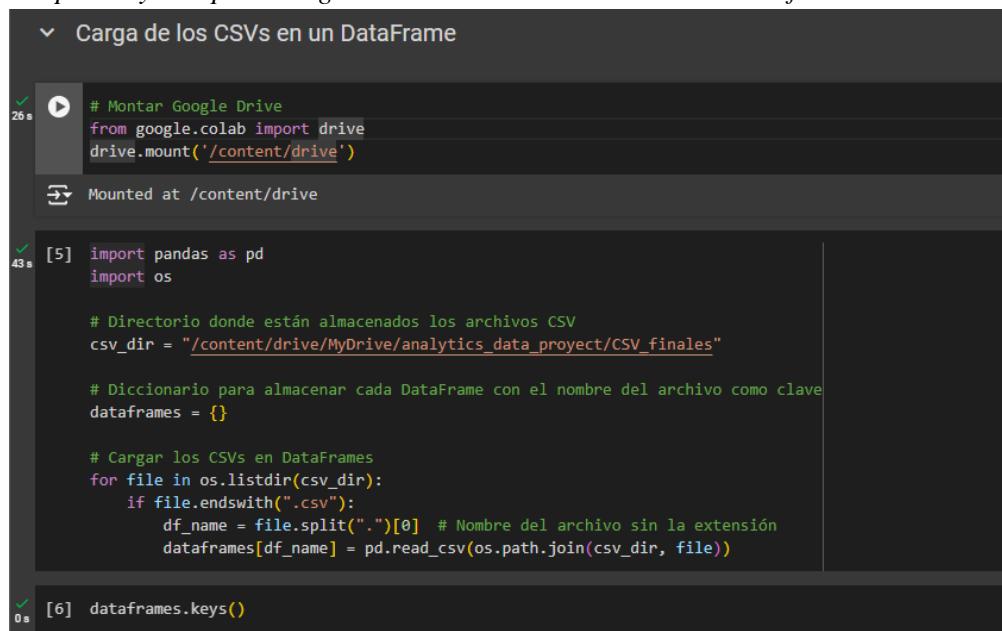
Una vez generados los 52 archivos CSV correspondientes a las consultas SQL, estos se subieron a Google Drive y se cargaron en el entorno de Google Colaboratory, aprovechando la capacidad de este entorno para manejar grandes volúmenes de datos de manera eficiente en la nube. Mediante la biblioteca Pandas en Python, los archivos fueron organizados en un diccionario de DataFrames.

Antes de proceder a la integración horizontal de los diferentes archivos CSV en un solo DataFrame, fue necesario realizar un proceso de estandarización de los nombres de las columnas, para evitar errores durante el proceso de combinación de los archivos y garantizar una estructura coherente para los análisis posteriores.

Posteriormente, los archivos se integraron utilizando como clave las columnas que el código DANE, que identifica de manera única a cada municipio y el año. Columnas comunes a todos los dataframes. El script que implementa lo anterior se muestra en la figura siguiente.

Figura 40

Script de Python para integrar todas las consultas en un solo dataframe



```

26 s   ✓  Carga de los CSVs en un DataFrame
26 s   [4] # Montar Google Drive
         from google.colab import drive
         drive.mount('/content/drive')
         ➔ Mounted at /content/drive

43 s   ✓  [5] import pandas as pd
         import os

         # Directorio donde están almacenados los archivos CSV
         csv_dir = "/content/drive/MyDrive/analytics_data_proyect/CSV_finales"

         # Diccionario para almacenar cada DataFrame con el nombre del archivo como clave
         dataframes = {}

         # Cargar los CSVs en DataFrames
         for file in os.listdir(csv_dir):
             if file.endswith(".csv"):
                 df_name = file.split(".")[0] # Nombre del archivo sin la extensión
                 dataframes[df_name] = pd.read_csv(os.path.join(csv_dir, file))

0 s     ✓  [6] dataframes.keys()

```

▼ Proceso para integrar en un solo df

Ajuste de los nombres de la columnas para que sean mas descriptivos

```
[7] # Diccionario donde almacenaremos los mapeos de columnas para cada archivo
column_mapping = {}

# Iterar sobre cada DataFrame en el diccionario
for df_name, df in dataframes.items():
    # Crear un subdiccionario donde cada columna se mapea a sí misma inicialmente
    column_mapping[df_name] = {col: col for col in df.columns}

# Mostrar el diccionario generado para revisarlo o editararlo manualmente
import pprint
pprint.pprint(column_mapping)
```

Mostrar salida oculta

▼ Ajuste manual del diccionario de mapeo

```
▶ column_mapping = {
    'afiliados_sss_adres': {'afil_contributivo': 'afil_contributivo',
                            'afil_subsidiado': 'afil_subsidiado',
                            'codigo_dane': 'codigo_dane',
                            'year': 'anio'},
    'armas_confiscadas_polinal': {'armas_confis_polinal': 'armas_confis_polinal',
                                  'año': 'anio',
                                  'codigo_dane': 'codigo_dane'},
```

▼ Aplicar mapeo

```
[6] # Aplicar el mapeo específico a cada DataFrame
for df_name, df in dataframes.items():
    if df_name in column_mapping:
        df.rename(columns=column_mapping[df_name], inplace=True)

[7] # imprimir columnas
for df_name, df in dataframes.items():
    print(f'{df_name}: {df.columns}')
```

Mostrar salida oculta

```
▶ # Revisar prerequisites para integrar
for name, df in dataframes.items():
    print(f'{name}:')
    print(len(df))
    print(df['codigo_dane'].nunique())
    print(df['anio'].nunique())
    print(df[['codigo_dane', 'anio']].isnull().sum())
```

Mostrar salida oculta

▼ Integrar horizontalmente todos los df contenidos en dataframes

```
[9] # Establecemos 'geoloc_poblacion' como el dataframe base
df_final = dataframes['geoloc_poblacion']

# Iteramos sobre los otros dataframes en el diccionario (excluyendo 'geoloc_poblacion')
for name, df in dataframes.items():
    if name != 'geoloc_poblacion':
        # Realizamos un merge en base a las columnas 'codigo_dane' y 'anio'
        df_final = pd.merge(df_final, df, on=['codigo_dane', 'anio'], how='outer')
```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

2.3 Resultados de la integración

El proceso de integración resultó en un DataFrame final con 210 filas (correspondientes a los 42 municipios durante los 5 años analizados) y 143 columnas, representa un extenso conjunto de datos sobre los 42 municipios del departamento del Cauca para el periodo 2019-2023. Cada fila del DataFrame corresponde a un municipio en un año determinado, mientras que las columnas incluyen información variada sobre condiciones socioeconómicas, eventos violentos y geometrías geoespaciales, entre otros aspectos.

Como era previsible, algunas columnas presentaban valores nulos. Esto se debe a la falta de datos para ciertos municipios o periodos en algunos de los conjuntos de datos originales. La presencia de estos valores nulos refleja la naturaleza incompleta de algunos registros, lo cual es abordado en la fase de preprocesamiento, donde se evaluará la mejor estrategia para gestionar estos datos ausentes, ya sea mediante la imputación de valores o la exclusión de variables específicas. No obstante, este comportamiento no compromete la integridad general del DataFrame, que proporciona una representación amplia y detallada de los indicadores relevantes para el análisis.

Como se aprecia el proceso de extracción e integración de datos se desarrolló con éxito gracias a una planificación cuidadosa y al uso de herramientas automatizadas que permitieron manejar eficientemente la gran cantidad de información disponible. Este conjunto de datos final, robusto y bien estructurado, constituye la base sobre la cual se realizarán las etapas de análisis de Machine Learning, permitiendo un estudio profundo de los patrones socioeconómicos y de violencia en los municipios del Cauca. Las consultas SQL ejecutadas, detalladas en el Anexo 4, ofrecen plena trazabilidad sobre los pasos seguidos en este proceso, garantizando transparencia y reproducibilidad del análisis.

2.4 Transformaciones orientadas a la aplicación de las técnicas de machine learning

El DataFrame obtenido tras la integración de las consultas SQL, con 210 filas (representando los 42 municipios del departamento del Cauca durante los cinco años del periodo 2019-2023) y 143 columnas (atributos relacionados con datos socioeconómicos, hechos de violencia y georreferenciación), constituye una base sólida para el análisis. Sin embargo, debido a su tamaño y alta dimensionalidad, es necesario aplicar un riguroso proceso de depuración y selección de características antes de proceder con las técnicas de Machine Learning. Estas técnicas incluyen la visualización de distribuciones, análisis de correlación, evaluación del factor de inflación de la varianza (VIF), análisis de información mutua, importancia por permutación, y Random Forest para la evaluación de la importancia de las características. Dichos análisis se ejecutan secuencialmente.

Este proceso es clave para abordar la "maldición de la dimensionalidad", un fenómeno que afecta particularmente a algoritmos de Machine Learning no supervisados, como el clústering y el análisis de componentes principales (PCA). La maldición de la dimensionalidad se refiere a la dificultad que surge cuando hay un número elevado de características en relación con el número de observaciones. En tales casos, los puntos de datos tienden a dispersarse en el espacio, dificultando que los algoritmos identifiquen patrones significativos (Géron, 2023).

A partir de la revisión bibliográfica, se identificaron algunas recomendaciones cruciales para preparar datasets adecuados para las técnicas de clústering con K-means, PCA, detección de anomalías con Isolation Forest y regresión logística, que son uno de los objetivos de este proyecto.

En primer lugar, los datos deben ser normalizados o escalados, dado que estas técnicas son sensibles a la escala de las variables, y una gran variabilidad entre ellas podría distorsionar los resultados. Es esencial que el conjunto de datos contenga variables relevantes, eliminando aquellas que solo aportan ruido (Mckinner, 2022).

Además, es fundamental gestionar adecuadamente los outliers extremos, ya que pueden sesgar los resultados, especialmente en K-means y PCA. Las variables categóricas deben ser transformadas a representaciones numéricas, y es crucial evitar correlaciones excesivas entre las variables, que podrían afectar el rendimiento de técnicas como PCA (VanderPlas, 2022).

Una distribución balanceada de los datos es también necesaria, evitando concentraciones excesivas en ciertas zonas del espacio de características, que podrían dificultar la identificación de patrones claros. Asimismo, el tamaño de la muestra debe ser lo suficientemente grande como para que los algoritmos detecten patrones sin comprometer la fiabilidad de los resultados (Bobadilla, 2021).

Finalmente, es clave gestionar correctamente los valores faltantes, ya sea mediante imputación o eliminación, para evitar sesgos, y reducir la dimensionalidad a un nivel adecuado que permita capturar la información suficiente sin introducir complejidad innecesaria en el análisis (McMahon, 2022).

Con el fin de obtener un dataset alineado con estas recomendaciones, se implementan diversas transformaciones sobre los datos integrados a partir de las consultas SQL, como se detalla a continuación.

2.4.1 Exploración inicial del dataset resultante de la extracción

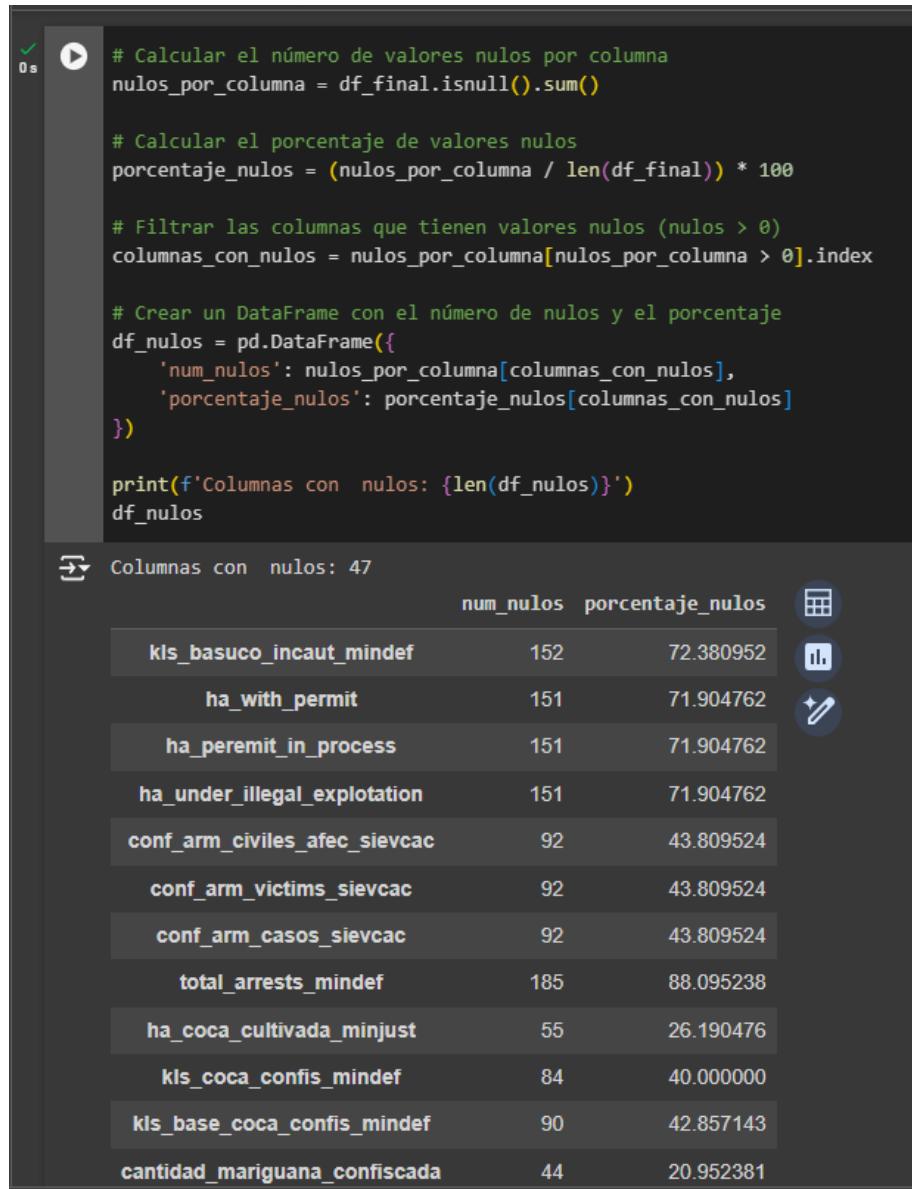
Antes de realizar transformaciones profundas, es esencial obtener una visión clara del dataset para entender su estructura y detectar problemas como valores nulos, redundancias o relaciones no evidentes entre variables.

Detección y gestión de valores nulos. A pesar de que en la etapa de integración de los datos a la base de datos se eliminaron los valores nulos para garantizar la coherencia y calidad de la información almacenada, al integrar horizontalmente los datos resultantes de las 52 consultas SQL se generaron valores nulos en varias columnas. Este fenómeno no refleja una falta de calidad en los datasets fuente, sino la realidad de ciertos eventos. Por ejemplo, en algunos municipios simplemente no hubo reportes de ciertos eventos de violencia durante el periodo analizado, simplemente porque no ocurrieron. Lo mismo ocurre con ciertos datos socioeconómicos: la ausencia de un indicador no implica que no se haya recolectado información, sino que esos eventos o condiciones no se dieron en ese contexto.

Además, el proceso de integración horizontal de las diferentes consultas en Pandas produce la aparición automática de valores nulos cuando los datasets no tienen información para todos los municipios o periodos de tiempo. Al combinar columnas de diferentes tablas o años en un único DataFrame, Pandas introduce nulos allí donde ciertos municipios no cuentan con datos para determinadas variables en ese año, reflejando de manera automática la falta de ocurrencia de eventos o la ausencia de ciertos indicadores. Esto es una consecuencia natural del proceso de consolidación y no un defecto en la calidad de los datos.

A través del script Python que se muestra en la figura de la siguiente página se calcula cuántos valores nulos tiene cada columna y el porcentaje que representan respecto al total de filas.

En este caso, dado que los valores nulos en las columnas representan la ausencia de hechos en ciertos municipios o períodos de tiempo, la imputación de estos con el valor 0 parece ser una decisión adecuada. Al hacerlo, se está respetando la semántica de los datos, donde un valor nulo no es un indicio de falta de calidad, sino de que ese evento no ocurrió.

Figura 41*Conteo de valores nulos por columna en el dataset final*


```

# Calcular el número de valores nulos por columna
nulos_por_columna = df_final.isnull().sum()

# Calcular el porcentaje de valores nulos
porcentaje_nulos = (nulos_por_columna / len(df_final)) * 100

# Filtrar las columnas que tienen valores nulos (nulos > 0)
columnas_con_nulos = nulos_por_columna[nulos_por_columna > 0].index

# Crear un DataFrame con el número de nulos y el porcentaje
df_nulos = pd.DataFrame({
    'num_nulos': nulos_por_columna[columnas_con_nulos],
    'porcentaje_nulos': porcentaje_nulos[columnas_con_nulos]
})

print(f'Columnas con nulos: {len(df_nulos)}')
df_nulos

```

Columnas con nulos: 47

	num_nulos	porcentaje_nulos
kls_basuco_incaut_mindef	152	72.380952
ha_with_permit	151	71.904762
ha_permit_in_process	151	71.904762
ha_under_illegal_explotation	151	71.904762
conf_arm_civiles_afec_sievcac	92	43.809524
conf_arm_victims_sievcac	92	43.809524
conf_arm_casos_sievcac	92	43.809524
total_arrests_mindef	185	88.095238
ha_coca_cultivada_minjust	55	26.190476
kls_coca_confis_mindef	84	40.000000
kls_base_coca_confis_mindef	90	42.857143
cantidad_mariguana_confiscada	44	20.952381

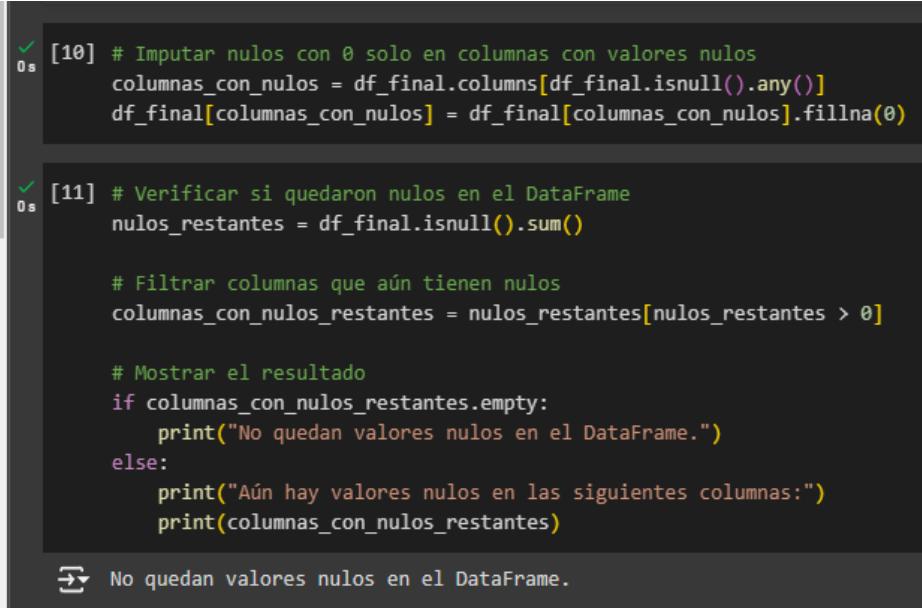
Fuente: Pantallazo obtenido desde el entorno de Google Colab

Aunque algunas columnas tienen un alto porcentaje de valores nulos, como en el caso de variables con más del 80% de nulos, es prudente no eliminarlas inmediatamente. Estos datos podrían tener valor en la generación de nuevas características o para ciertos análisis específicos, especialmente porque reflejan eventos poco comunes pero potencialmente relevantes. Eliminar las columnas ahora podría reducir la riqueza del conjunto de datos para futuros procesos de transformación y modelado.

Por lo tanto, el siguiente paso lógico es proceder con la imputación de los nulos a 0, dejando abierta la posibilidad de reevaluar estas variables más adelante en el proceso de depuración

Figura 42

Script de imputación de nulos y verificación de que se aplicó



```

[10] # Imputar nulos con 0 solo en columnas con valores nulos
columnas_con_nulos = df_final.columns[df_final.isnull().any()]
df_final[columnas_con_nulos] = df_final[columnas_con_nulos].fillna(0)

[11] # Verificar si quedaron nulos en el DataFrame
nulos_restantes = df_final.isnull().sum()

# Filtrar columnas que aún tienen nulos
columnas_con_nulos_restantes = nulos_restantes[nulos_restantes > 0]

# Mostrar el resultado
if columnas_con_nulos_restantes.empty:
    print("No quedan valores nulos en el DataFrame.")
else:
    print("Aún hay valores nulos en las siguientes columnas:")
    print(columnas_con_nulos_restantes)

```

No quedan valores nulos en el DataFrame.

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Clasificación de las variables. Dado que el dataset con el que se está trabajando consta de 143 columnas, es fundamental adoptar un enfoque estructurado para su análisis. En este conjunto de datos podemos distinguir claramente tres grupos principales de variables: variables de georreferenciación, variables socioeconómicas y variables relacionadas con hechos de violencia. Ante la gran cantidad de columnas y la diversidad de los datos, se procede a segmentar el análisis según estos grupos, ya que cada uno tiene características y comportamientos únicos.

Las variables de georreferenciación no requieren un análisis de distribución, ya que su propósito es facilitar la integración de análisis geoespaciales futuros. Estas variables actúan como un marco de referencia para la localización geográfica de los datos, por lo que no forman parte del proceso de depuración en términos de análisis estadístico.

Por otro lado, las variables socioeconómicas tienden a seguir patrones más estables y predecibles, con distribuciones que generalmente presentan menos variabilidad extrema a lo largo del tiempo. Estas variables se ven afectadas por factores estructurales como políticas

públicas o contextos económicos, y suelen tener una relación clara con el comportamiento de los municipios a lo largo del periodo analizado. Analizar estas distribuciones de manera independiente nos permitirá comprender mejor si siguen distribuciones normales o están sesgadas, facilitando su transformación y normalización para el posterior modelado (VanderPlas, 2022).

En contraste, las variables relacionadas con los hechos de violencia pueden mostrar comportamientos más erráticos y heterogéneos. Es común que estas variables estén influenciadas por circunstancias contextuales específicas, lo que puede generar picos o irregularidades en las distribuciones. Separar su análisis permite centrarse en la identificación de outliers y patrones atípicos, los cuales son relevantes para la detección de anomalías y otros análisis más especializados en el contexto de la violencia en el Cauca.

Por lo anterior se segmenta el análisis de las distribuciones según estas categorías, ya que permite abordar de manera más efectiva la diversidad de los datos.

Del resultado de la clasificación de las 143 columnas del dataset en tres grupos principales de variables se obtuvo: 5 variables de georreferenciación que permiten situar espacialmente la información y son esenciales para análisis geoespaciales futuros, 50 variables socioeconómicas que proporcionan una visión sobre las condiciones demográficas y estructurales de los municipios del Cauca, y 88 variables sobre hechos de violencia que los distintos tipos de delitos, víctimas y eventos relacionados con la violencia en la región. Los nombres de las variables por grupo, su significado y la fuente de donde provienen se presentan en la tabla 3.

Esta clasificación facilita la comprensión de los datos, permitiendo que las transformaciones y el análisis exploratorio se lleven a cabo de manera más eficiente y coherente con los objetivos del proyecto.

Tabla 4*Columnas del dataset inicial resultante de la extracción de información de la base de datos*

Clase	Nombre en dataset	Significado	Fuente
Georreferenciación	codigo_dane	Código Dane del municipio - 5 dígitos	DANE
	nombre_municipio	Nombre oficial del municipio	DANE
	geometria_mpio	Información geográfica del polígono de los límites del municipio	DANE
	geometria_cabecera	Información geográfica de la ubicación de la cabecera municipal	DANE
	Anio	Año de vigencia de la información	DANE
Socioeconómicas	Poblacion	Número de habitantes	DANE
	ha_with_permit	Hectáreas con permiso para explotación de metales preciosos	Ministerio de Minas
	ha_peremitt_in_process	Hectáreas con permiso en trámite para explotación de metales preciosos	Ministerio de Minas
	ha_under_illegal_explotation	Hectáreas en explotación ilegal de metales preciosos	Ministerio de Minas
	ha_condicionada_upra	Hectáreas con vocación agrícola condicionada	Ministerio de Agricultura
	ha_no_condicionada_upra	Hectáreas con vocación agrícola no condicionada	Ministerio de Agricultura
	desempleo_cens_2018	Tasas de desempleo calculada del censo 2018	DANE
	tasa_alfab_cens_2018	Tasa de alfabetas según censo 2018	DANE
	tasa_analf_cens_2018	Tasa de analfabetas según censo 2018	DANE
	numero_hogares_dane	Número de hogares según censo 2018	DANE
	hacinamiento_promedio_dane	Promedio Número de hogares sobre Número de dormitorios	DANE
	densidad_habitacional_dane	Total personas sobre Número de cuartos	DANE
	tasa_acceso_a_cocina_dane	Tasa de viviendas con cocina	DANE
	tasa_acceso_agua_cocinar_dane	Tasa de viviendas con acceso al agua para cocinar	DANE
	cob_elect_cens_2018	Cobertura servicio de energía eléctrica	DANE
	cob_acued_cens_2018	Cobertura de servicio de acueducto	DANE
	cob_alcant_cens_2018	Cobertura de servicio de alcantarillado	DANE
	cob_gas_cens_2018	Cobertura de servicio de gas	DANE
	cob_rec_bas_cens_2018	Cobertura de servicio de recolección de basura	DANE
	cob_internet_cens_2018	Cobertura de servicio de Internet	DANE
	total_viviendas_cens_2018	Número de viviendas censadas	DANE
	prom_hog_vivienda_cens_2018	Número promedio de hogares por vivienda	DANE
	porc_est_1_cens_2018	Porcentaje de población en estrato 1	DANE
	porc_est_2_cens_2018	Porcentaje de población en estrato 2	DANE
	porc_est_3_cens_2018	Porcentaje de población en estrato 3	DANE
	porc_pared_adecuada_cens_2018	Porcentaje de vivienda con paredes adecuadas	DANE
	porc_pared_inadecuada_cens_2018	Porcentaje de viviendas con paredes inadecuadas	DANE
	porc_piso_adecuado_cens_2018	Porcentajes de viviendas con pisos adecuados	DANE

porc_piso_inadecuado_cens_2018	Porcentajes de viviendas con pisos inadecuados	DANE
tasa_matricula_5_16_mined	Tasa de matrícula entre población de 5 a 16 años	Ministerio de Educación Nacional
cobertura_neta_mined	Cobertura neta de educación básica y media	Ministerio de Educación Nacional
cobertura_bruta_mined	Cobertura bruta de educación básica y media	Ministerio de Educación Nacional
desercion_mined	Tasa de deserción escolar educación básica y media	Ministerio de Educación Nacional
poblacion_5_16_mined	Total población entre 5 a 16 años	Ministerio de Educación Nacional
aprobacion_mined	Tasa de aprobación educación básica y media	Ministerio de Educación Nacional
ninos_icbf_mined	Número de niños atendidos por el ICBF	Ministerio de Educación Nacional
ninos_prescolar_mined	Número de niños matriculas en prescolar	Ministerio de Educación Nacional
afil_contributivo	Número de afiliados al régimen contributivo en salud	Ministerio de Educación Nacional
afil_subsidiado	Número de afiliados al subsidiado contributivo en salud	Ministerio de Educación Nacional
tasa_mort_desnutricion_minsal	Tasa de mortalidad por desnutrición infantil	DAPRE
tasa_mort_general_minsal	Tasa general de mortalidad	DAPRE
tasa_mort_diarr_minsal	Tasa de mortalidad por enfermedades diarrea aguda	Ministerio de Salud
tasa_mort_ninez_minsal	Tanas de mortalidad infantil	Ministerio de Salud
porc_bajo_peso_nacer_minsal	Porcentaje de niños con bajo peso al nacer	Ministerio de Salud
per_ipm_sisben	Personas con índice de pobreza multidimensional	DNP - Sisben IV
casos_trab_infan_sisben	Número de casos de trabajo infantil	DNP - Sisben IV
per_tab_inf_sisben	Personas con trabajos informales	DNP - Sisben IV
per_haci_critico_sisben	Personas en hacinamiento critico	DNP - Sisben IV
hogares_sisbenizados	Número de hogares en el Sisben	DNP - Sisben IV
viviendas_sisbenizados	Número de viviendas en el Sisben	DNP - Sisben IV
kls_basuco_incaut_mindef	Kilos de basuco incautados	Ministerio de Defensa
conf_arm_civiles_afec_sievcac	civiles afectados por conflicto armado	SIEVCAC
conf_arm_victims_sievcac	Víctimas conflicto armado	SIEVCAC
conf_arm_casos_sievcac	casos documentados conflicto armado	SIEVCAC
total_arrests_mindef	Total arrestos	Ministerio de Defensa
ha_coca_cultivada_minjust	Hectáreas de coca cultivada	Ministerio de Justicia
kls_coca_confis_mindef	Kilos de coca incautados	Ministerio de Defensa
kls_base_coca_confis_mindef	Kilos de base de coca incautados	Ministerio de Defensa
cantidad_mariguana_confiscada	Kilos de marihuana incautados	Ministerio de Defensa
amenazas_casos_fiscalia	Casos delitos de amenaza	Fiscalía General de la Nación
delitos_sexuales_casos_fiscalia	Casos delitos sexuales	Fiscalía General de la Nación
desaparicion_forzada_casos_fiscalia	Casos delitos desaparición forzada	Fiscalía General de la Nación
desplazamiento_casos_fiscalia	Casos delitos desplazamiento	Fiscalía General de la Nación
estafa_casos_fiscalia	Casos de delitos de estafa	Fiscalía General de la Nación
extorsion_casos_fiscalia	Casos de delitos de extorsión	Fiscalía General de la Nación

hurto_casos_fiscalia	Casos de delito de hurto	Fiscalía General de la Nación
personas_bienes_dih_casos_fiscalia	Casos de delito contra personas y bienes protegidos por el DIH	Fiscalía General de la Nación
reclutamiento_ilicito_casos_fiscalia	Casos de delitos reclutamiento ilícito	Fiscalía General de la Nación
violencia_intrafamiliar_casos_fiscalia	Casos de delitos violencia familiar	Fiscalía General de la Nación
corrupcion_casos_fiscalia	Casos de delitos corrupción	Fiscalía General de la Nación
homicidios_casos_fiscalia	Casos de delitos de homicidio	Fiscalía General de la Nación
lesiones_personales_casos_fiscalia	Casos de delitos de lesiones personales	Fiscalía General de la Nación
secuestro_casos_fiscalia	Casos de delitos de secuestro	Fiscalía General de la Nación
otros_delitos_casos_fiscalia	Casos de otros delitos	Fiscalía General de la Nación
indiciados_fisc_amenazas	Indiciados por delitos de amenazas	Fiscalía General de la Nación
indiciados_fisc_delitos_sexuales	Indiciados por delitos sexuales	Fiscalía General de la Nación
indiciados_fisc_desp_forzada	Indiciados por delitos de desaparición forzada	Fiscalía General de la Nación
indiciados_fisc_desplazamiento	Indiciados por delitos de desplazamiento	Fiscalía General de la Nación
indiciados_fisc_estafa	Indiciados por delitos de estafa	Fiscalía General de la Nación
indiciados_fisc_extorsion	Indiciados por delitos de extorsión	Fiscalía General de la Nación
indiciados_fisc_hurto	Indiciados por delitos de hurto	Fiscalía General de la Nación
indiciados_fisc_dih	Indiciados por delitos contra personas y bienes protegidos por el DIH	Fiscalía General de la Nación
indiciados_fisc_reclut_ilicito	Indiciados por delitos reclutamiento ilícito	Fiscalía General de la Nación
indiciados_fisc_viol_intraf	Indiciados por delitos de violencia intrafamiliar	Fiscalía General de la Nación
indiciados_fisc_corrupcion	Indiciados por delitos de corrupción	Fiscalía General de la Nación
indiciados_fisc_homicidios	Indiciados por delitos de homicidios	Fiscalía General de la Nación
indiciados_fisc_lesiones_personales	Indiciados por delitos de lesiones personales	Fiscalía General de la Nación
indiciados_fisc_secuestro	Indiciados por delitos de secuestro	Fiscalía General de la Nación
indiciados_fisc_otros_delitos	Indiciados por otros delitos	Fiscalía General de la Nación
crimenes_amb_mindef	Delitos ambientales	Ministerio de Defensa
gal_insum_liq_incau_mindef	Galones de insumos líquidos incautados	Ministerio de Defensa
asesinatos_mindef	Número de asesinatos	Ministerio de Defensa
masacres_mindef	Número de masacres	Ministerio de Defensa
violenc_domest_polinal	Número de casos de violencia doméstica	Policía Nacional
pirateria_terrest_mindef	Casos de piratería terrestre	Ministerio de Defensa
casos_terror_mindef	Casos de terrorismo	Ministerio de Defensa
victim_fisc_amenazas	Víctimas de delitos de amenaza	Fiscalía General de la Nación
victim_fisc_delitos_sexuales	Víctimas de delitos sexuales	Fiscalía General de la Nación
victim_fisc_desaparicion_forzada	Víctimas de delitos de desaparición forzada	Fiscalía General de la Nación
victim_fisc_desplazamiento	Víctimas de delitos de desplazamiento	Fiscalía General de la Nación
victim_fisc_estafa	Víctimas de delitos de estafa	Fiscalía General de la Nación
victim_fisc_extorsion	Víctimas de delitos de extorsión	Fiscalía General de la Nación

victim_fisc_hurto	Víctimas de delitos de hurto	Fiscalía General de la Nación
victim_fisc_personas_bienes_dih	Víctimas de delitos contra personas y bienes protegidos por el DIH	Fiscalía General de la Nación
victim_fisc_reclutamiento_ilicito	Víctimas de delitos de reclutamiento ilícito	Fiscalía General de la Nación
victim_fiscViolencia_intrafamiliar	Víctimas de delitos de violencia intrafamiliar	Fiscalía General de la Nación
victim_fisc_corrupcion	Víctimas de delitos de corrupción	Fiscalía General de la Nación
victim_fisc_homicidios	Víctimas de delitos de homicidios	Fiscalía General de la Nación
victim_fisc_lesiones_personales	Víctimas de delitos de lesiones personales	Fiscalía General de la Nación
victim_fisc_secuestro	Víctimas de delitos de secuestro	Fiscalía General de la Nación
victim_fisc_otros_delitos	Víctimas de otros delitos	Fiscalía General de la Nación
robo_entid_finan_mindef	Número de robo a entidades financieras	Ministerio de Defensa
robo_a_personas_mindef	Número de robo a personas	Ministerio de Defensa
secuestro_mindef	Número de secuestro	Ministerio de Defensa
lab_destruidos_mindef	Número de laboratorios de coca destruidos	Ministerio de Defensa
casos_extorsion_mindef	Número de extorsiones	Ministerio de Defensa
minas_interv_mindef	Número de minas antipersonales	Ministerio de Defensa
vict_fuerza_pub_mindef	Número de Víctimas de la fuerza publica	Ministerio de Defensa
terrorismo_polinal	Casos de terrorismo	Policía Nacional
robo_viv_neg_polinal	Casos de robo a vivienda	Policía Nacional
robos_polinal	Casos de robos a personas	Policía Nacional
les_personales_polinal	Casos de lesiones personales	Policía Nacional
incautacion_drogas_poli	Kilos de incautación de drogas	Policía Nacional
delito_sexual_polinal	Casos de delitos sexuales	Policía Nacional
num_muertes_medlegal	Número de muertes	Instituto de Medicina Legal
armas_confis_polinal	Número de armas confiscadas	Policía Nacional
robo_vehiculos_polinal	Casos de robo de vehículos	Policía Nacional
reporte_desaparecidos_ml	Reporte de desaparecidos	Instituto de Medicina Legal
amenazas_polinal	Casos de amenazas policía	Policía Nacional
total_reclam_rest_tierr_minagr	Total reclamantes de restitución de tierras	Ministerio de Agricultura
violencia_intraf_ml	Casos de violencia intrafamiliar	Instituto de Medicina Legal
violencia_interper_ml	Casos de violencia interpersonal	Instituto de Medicina Legal
lesiones_accid_ml	Casos de lesiones personales	Instituto de Medicina Legal
presunto_delit_sexual_ml	Casos de presunto delito sexual	Instituto de Medicina Legal
vict_por_declarac_uv	Número de Víctimas del conflicto armado con declaración	Unidad de victimas
vict_ubicadas_uv	Número de Víctimas del conflicto armado con por ubicación	Unidad de victimas
vict_en_atencion_uv	Número de Víctimas del conflicto armado sujetos de atención	Unidad de victimas
vict_por_eventos_uv	Número de Víctimas del conflicto armado por eventos registrados	Unidad de victimas

2.4.2 Análisis y transformación de las variables socioeconómicas del dataset

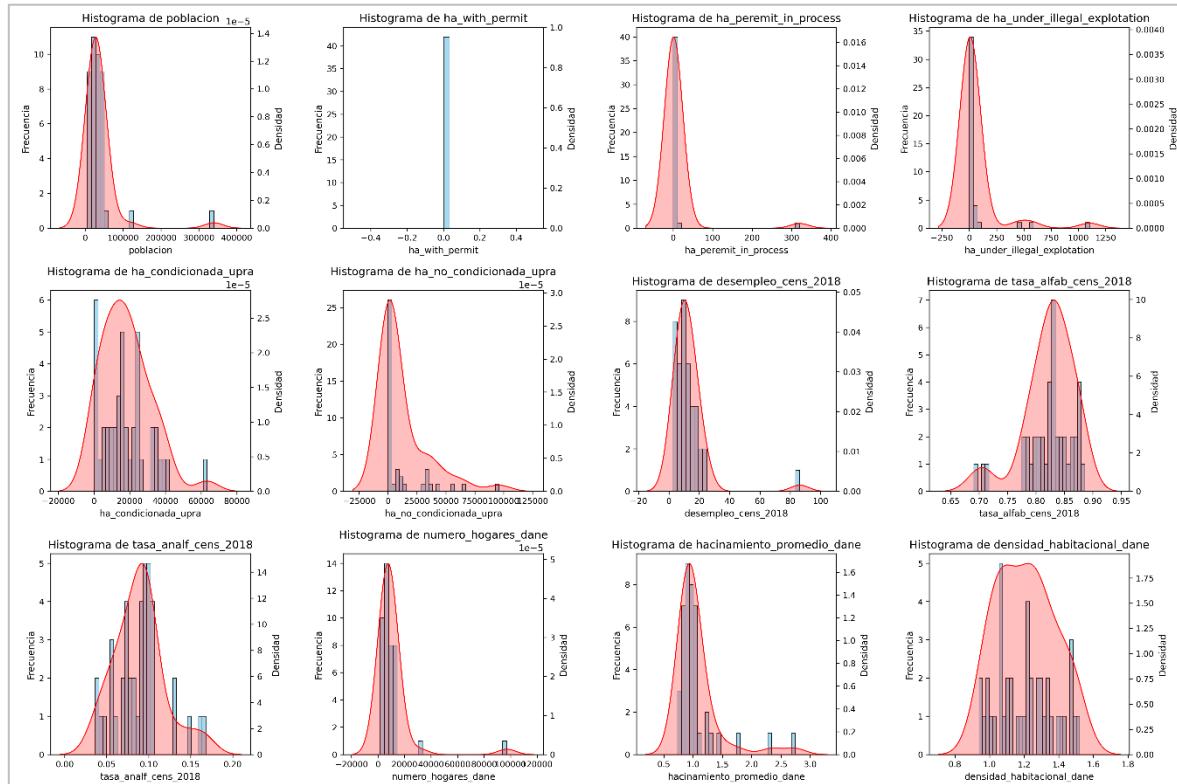
Revisión de distribuciones de las variables. Para asegurar una correcta interpretación de las distribuciones de las variables, es importante considerar cómo se manejan los datos correspondientes a diferentes años (2019-2023). Si se visualizan todas las observaciones de forma conjunta, se corre el riesgo de distorsionar los gráficos debido a la mezcla de tendencias anuales. Por ejemplo, un histograma que combine datos de varios años puede ocultar patrones específicos de un periodo en particular o suavizar cambios importantes en un gráfico de densidad, dificultando la comprensión de la evolución temporal de las variables.

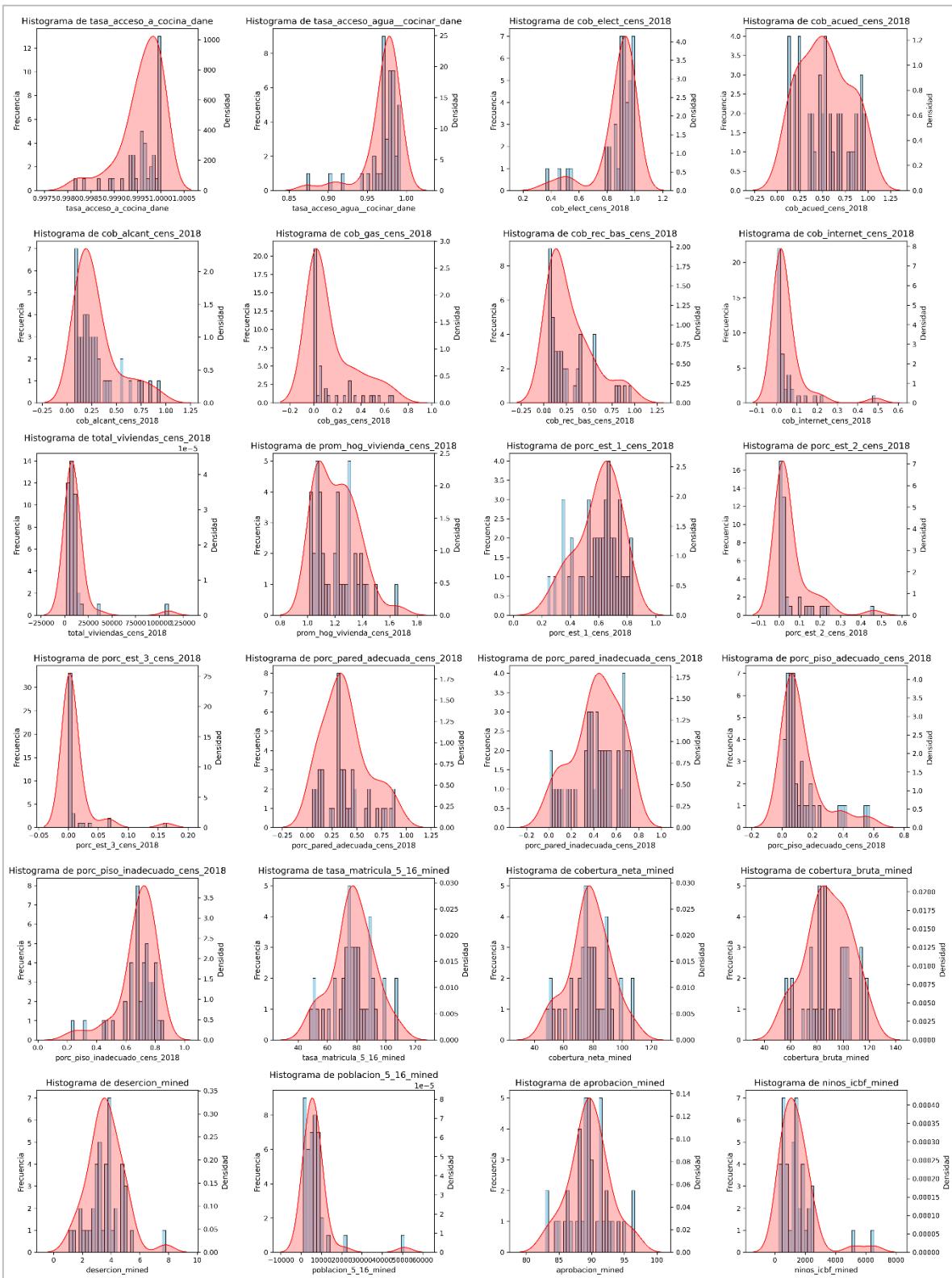
Por esta razón, se decide filtrar los datos por año antes de generar las visualizaciones. Esto permite analizar de manera clara y precisa las distribuciones de cada variable en un periodo específico, evitando la interferencia de otros años.

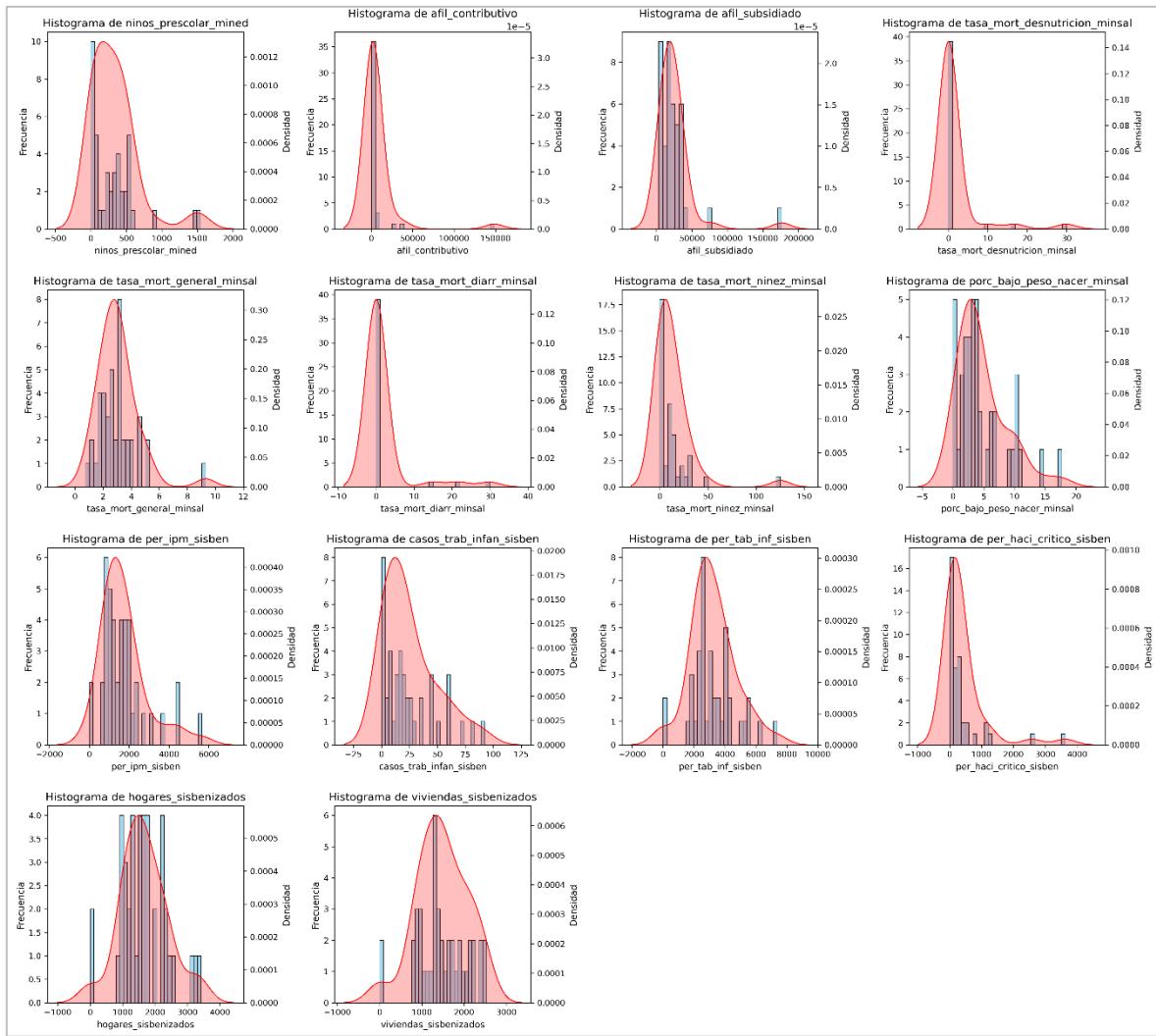
A continuación se presenta los histogramas de las variables socioeconómicas del dataset, con una curva de densidad superpuesta en el mismo gráfico para una mejor comparación visual.

Figura 43

Histogramas de las variables socioeconómicas del dataset con curva de densidad superpuesta







Fuente: Pantallazo obtenido desde el entorno de Google Colab

El análisis de las distribuciones de las variables socioeconómicas para los 42 municipios del departamento del Cauca revela una tendencia marcada hacia la asimetría positiva en varias de ellas. Variables como la población y las hectáreas bajo explotación ilegal presentan una alta concentración de valores bajos, mientras que unos pocos municipios destacan con valores extremadamente altos, lo que genera colas largas en los histogramas. Esta dispersión es especialmente notable en variables como la población, lo que sugiere la presencia de unos pocos municipios que concentran la mayor parte de la población.

Asimismo, se observan valores concentrados en cero en diversas variables lo que refleja una falta de actividad o de datos no registrados en muchos municipios. Esta concentración afecta la forma de las distribuciones, ya que los histogramas muestran picos muy pronunciados en torno al valor cero. Sin embargo, en otras variables como el desempleo o la tasa de

alfabetización, la variabilidad entre municipios es más homogénea, lo que permite inferir que, en ciertos aspectos, los municipios comparten características más uniformes.

Algunas variables como las tasas de mortalidad general y mortalidad infantil exhiben mayor dispersión, lo que podría sugerir distribuciones bimodales en las que ciertos municipios tienen condiciones significativamente diferentes. Estos resultados, que destacan por la presencia de valores atípicos y distribuciones sesgadas, sugieren la necesidad de un análisis más cuidadoso al interpretar promedios y tendencias, dado que los extremos juegan un papel importante en la caracterización de los datos.

En cuanto a las variables cuyos histogramas muestran una concentración significativa de valores en cero, el comportamiento se debe a la imputación de valores nulos que se realizó anteriormente, donde los ceros representan la ausencia de datos en algunos municipios. Dicho fenómeno indica que estas variables no cuentan con información para un alto número de municipios, lo que afecta su relevancia para el análisis. Las variables identificadas en esta situación incluyen: ha_with_permit, ha_peremt_in_process, ha_under_illegal_explotation, tasa_mort_desnutricion_minsal y tasa_mort_diarr_minsal.

Para verificar este análisis, se filtran las variables socioeconómicas con más del 60% de ceros como valores y se le aplica a la muestra la función describe() de pandas. Los resultados junto con el script de Python se muestran en la figura siguiente.

Figura 44

Variables socioeconómicas con un alto porcentaje de valores en cero

```
# Definir el umbral de porcentaje de ceros para filtrar
threshold_percentage = 0.6

# Filtrar las columnas que son de tipo socioeconómico
df_socioeconomico = df_final[var_socieconomicas]

# Calcular el porcentaje de ceros en cada columna
zero_percentage = (df_socioeconomico == 0).mean()

# Filtrar las columnas que tienen un porcentaje de ceros mayor al umbral
high_zero_columns = zero_percentage[zero_percentage > threshold_percentage].index

# Filtrar el DataFrame por esas columnas
df_high_zeros = df_socioeconomico[high_zero_columns]

# Aplicar describe() para ver las estadísticas de estas columnas
df_high_zeros.describe()
```

	ha_with_permit	ha_peremt_in_process	ha_under_illegal_explotation	tasa_mort_desnutricion_minsal	tasa_mort_diarr_minsal
count	42.0	42.000000	42.000000	42.000000	42.000000
mean	0.0	8.365569	58.474966	1.346186	1.581403
std	0.0	49.452438	198.365506	5.401418	6.007825
min	0.0	0.000000	0.000000	0.000000	0.000000
25%	0.0	0.000000	0.000000	0.000000	0.000000
50%	0.0	0.000000	0.000000	0.000000	0.000000
75%	0.0	0.000000	6.493179	0.000000	0.000000
max	0.0	320.559333	1101.373169	29.868578	29.868578

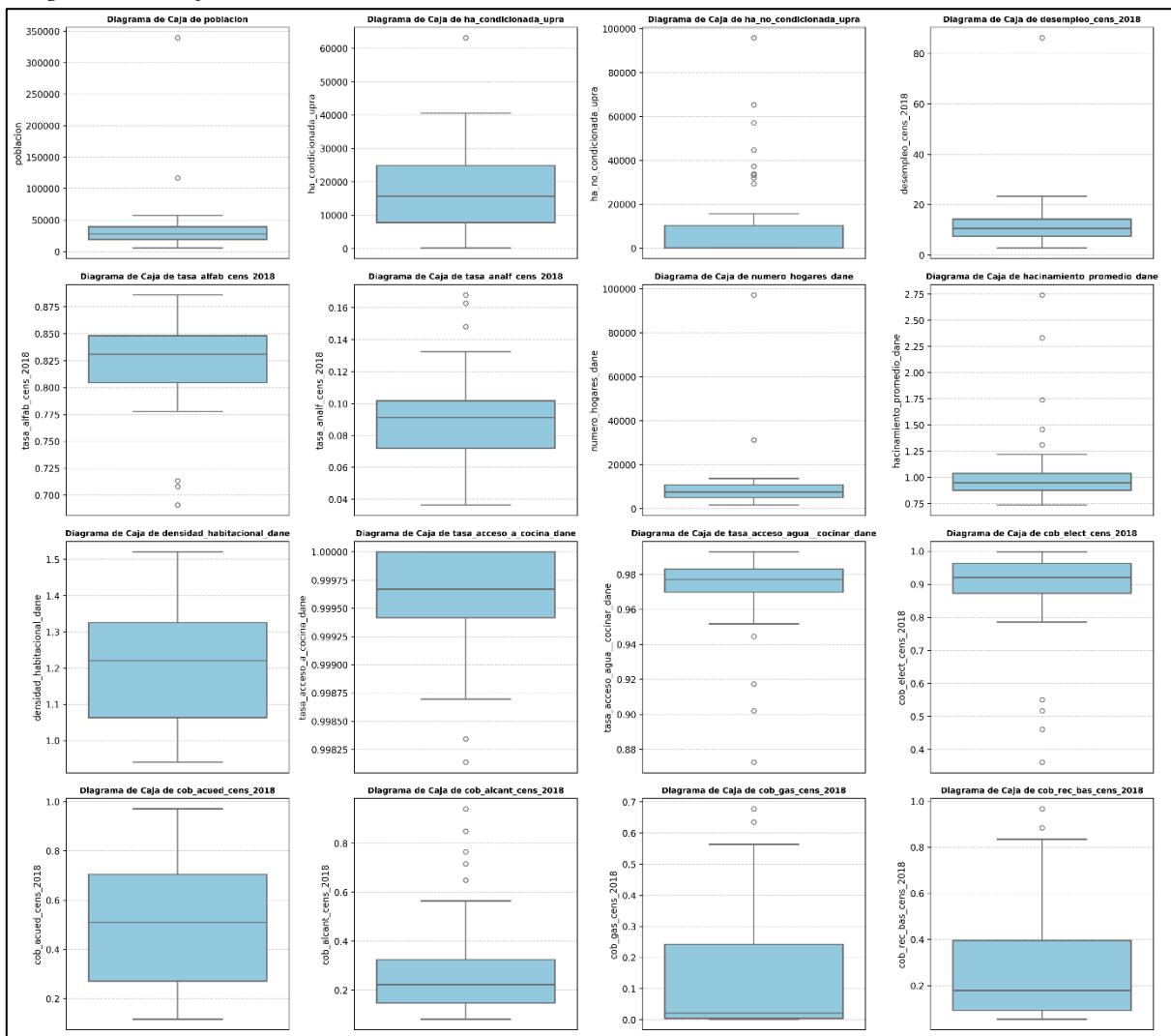
Fuente: Pantallazo obtenido desde el entorno de Google Colab

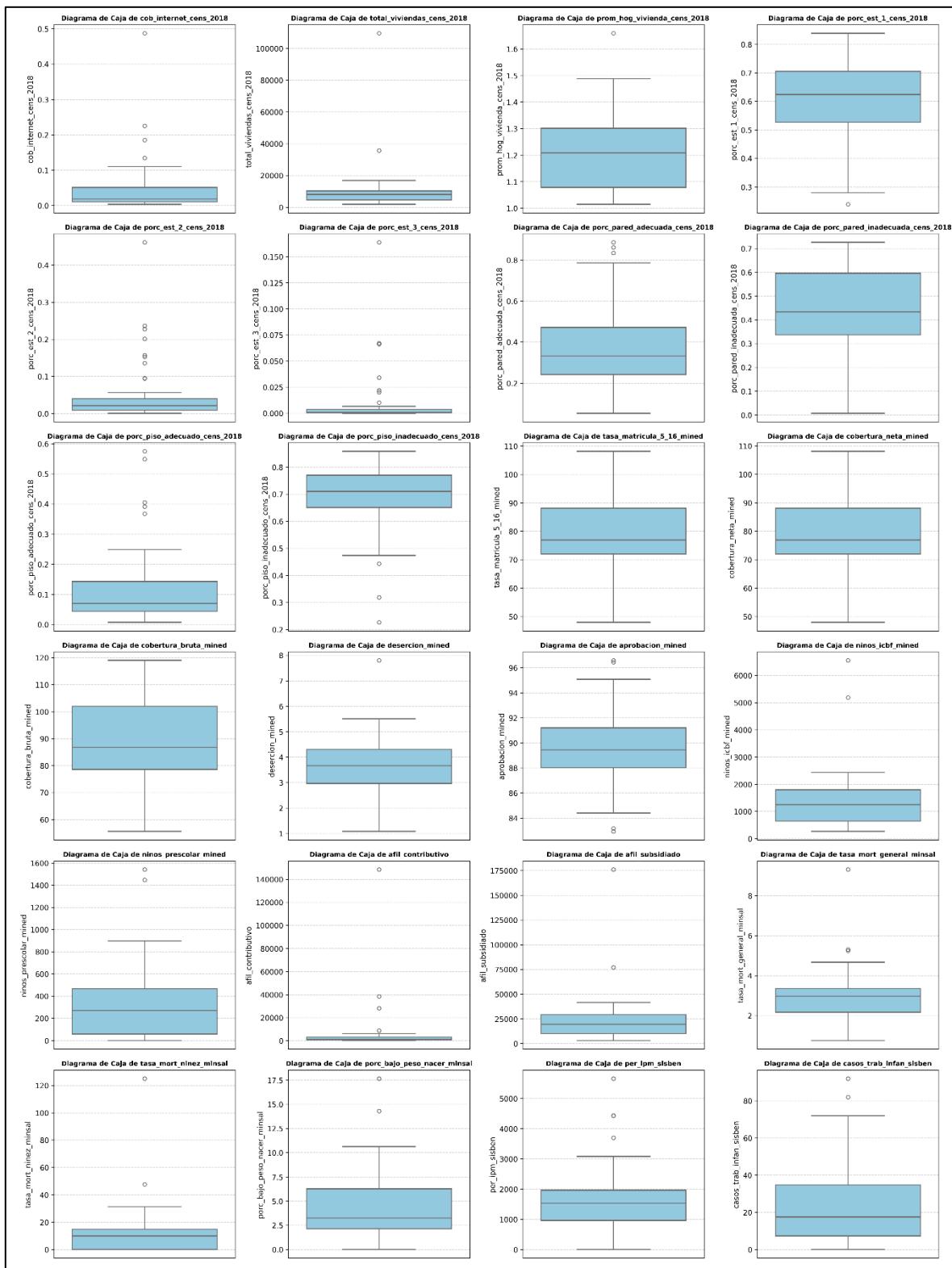
Respecto a estas variables se decide marginarlas del análisis por su baja utilidad informativa. No obstante, se conserva la variable ha_under_illegal_explotation, ya que, aunque más del 60% de los municipios no reportan información en esta variable, porque en ellos no se presenta la esta actividad, es un indicador relevante para el análisis, ya que la explotación ilegal de minerales preciosos se asocia frecuentemente como desencadenantes de hechos de violencia.

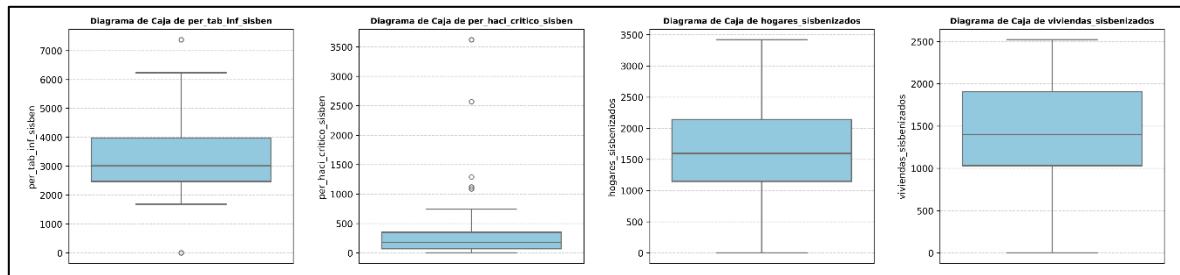
Para evaluar la dispersión, simetría y presencia de valores atípicos en las variables socioeconómicas del estudio, a continuación se generan los diagramas de caja (*boxplots*) para cada una de ellas.

Figura 45

Diagramas de caja de las variables socioeconómicas del dataset







Fuente: Pantallazo obtenido desde el entorno de Google Colab

Apoyándose en la visualización de los diagramas de cajas se aprecia que las variables socioeconómicas del departamento del Cauca contempladas en el dataset, correspondiente a los 42 municipios, revela una presencia significativa de valores atípicos en la mayoría de las variables. De las 45 variables estudiadas, 31 muestran valores atípicos por encima de la cola superior de su distribución, lo que sugiere que la distribución puede estar influenciada por algunos municipios con características atípicas en términos de población, desempleo, tasa de alfabetización, entre otros factores socioeconómicos.

Por otro lado, 7 variables presentan valores atípicos en la cola inferior, indicando que ciertos municipios tienen valores significativamente más bajos en estas variables, lo que también podría reflejar disparidades socioeconómicas importantes dentro del departamento.

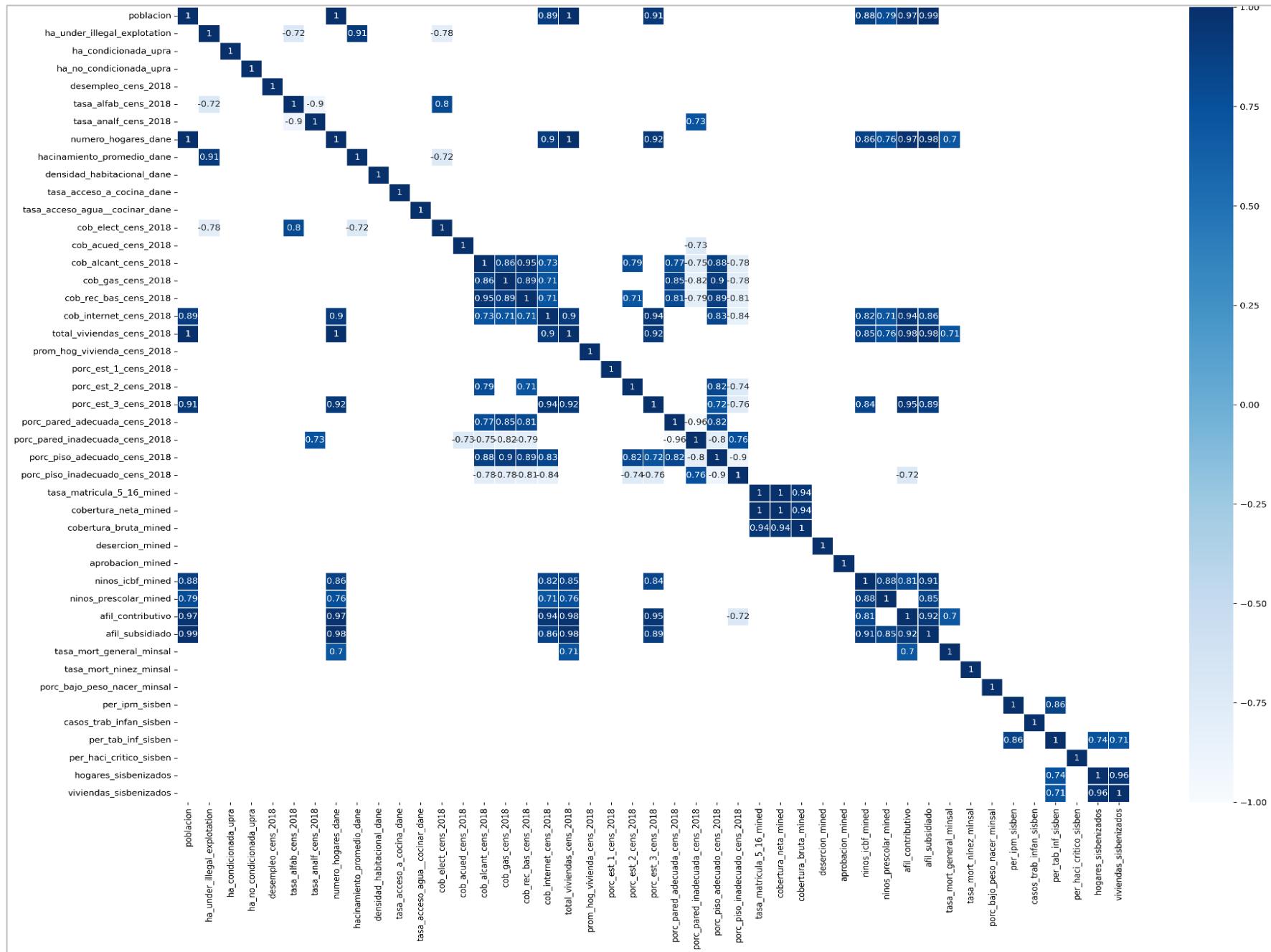
Además, 2 variables presentan valores atípicos tanto en la cola superior como en la inferior. Esto indica una distribución más dispersa con extremos en ambos lados, reflejando una variabilidad considerable en estas características entre los municipios. El resto de las variables no presentan valores atípicos visibles en los diagramas de caja, lo que sugiere que sus distribuciones son más homogéneas o tienen una dispersión moderada sin puntos extremos.

Correlación entre variables. Se realiza un análisis utilizando el coeficiente de correlación de Pearson, inicialmente sobre las 45 variables socioeconómicas que aún permanecen en el dataset, se busca sobre todo detectar variables fuertemente relacionadas (con umbrales de ± 0.7 o superior), para que facilite decisiones sobre eliminar, fusionar o crear nuevas características que capturen mejor la información relevante.

En la matriz de correlación de las 45 variables socioeconómicas que se muestra en la gráfica de la siguiente página, se verifican varias relaciones de alta correlación, aquellas superiores a 0.7, que sugieren la posible redundancia de información entre variables. Variables como la población total, número de hogares, y viviendas censadas muestran correlaciones superiores a 0.99, lo cual indica que están capturando dimensiones muy similares del tamaño poblacional y la infraestructura básica.

Figura 46

Visualización de la matriz las variables socioeconómicas para altas correlaciones



Fuente: Pantallazo obtenido desde el entorno de Google Colab

De manera similar, las variables relacionadas con la cobertura de servicios como internet y electricidad también presentan alta correlación, lo que podría estar describiendo un mismo nivel de desarrollo tecnológico y urbano.

La tabla siguiente resume de manera más clara los pares de variables socioeconómicas que presentan un correlación elevada, mayor a 0.7 (positiva o negativa) en el dataset.

Tabla 5

Pares de variables socioeconómicas altamente correlacionadas

Variable 1	Variable 2	Correlación
tasa_matricula_5_16_mined	cobertura_neta_mined	0.999999
numero_hogares_dane	total_viviendas_cens_2018	0.997812
Poblacion	numero_hogares_dane	0.997454
Poblacion	total_viviendas_cens_2018	0.995763
Poblacion	poblacion_5_16_mined	0.987708
poblacion_5_16_mined	afil_subsidiado	0.986740
Poblacion	afil_subsidiado	0.985029
numero_hogares_dane	afil_subsidiado	0.979557
numero_hogares_dane	poblacion_5_16_mined	0.977587
total_viviendas_cens_2018	afil_subsidiado	0.975723
total_viviendas_cens_2018	afil_contributivo	0.975320
numero_hogares_dane	afil_contributivo	0.973249
total_viviendas_cens_2018	poblacion_5_16_mined	0.971390
Poblacion	afil_contributivo	0.967007
hogares_sisbenizados	viviendas_sisbenizados	0.955475
porc_est_3_cens_2018	afil_contributivo	0.952205
cob_alcant_cens_2018	cob_rec_bas_cens_2018	0.945798
cobertura_neta_mined	cobertura_bruta_mined	0.941962
tasa_matricula_5_16_mined	cobertura_bruta_mined	0.941893
cob_internet_cens_2018	afil_contributivo	0.940302
cob_internet_cens_2018	porc_est_3_cens_2018	0.939249
poblacion_5_16_mined	afil_contributivo	0.931263
afil_contributivo	afil_subsidiado	0.921388
numero_hogares_dane	porc_est_3_cens_2018	0.919506
total_viviendas_cens_2018	porc_est_3_cens_2018	0.917931
poblacion_5_16_mined	ninos_icbf_mined	0.915511
Poblacion	porc_est_3_cens_2018	0.914687
ha_under_illegal_explotation	hacinamiento_promedio_dane	0.911835
ninos_icbf_mined	afil_subsidiado	0.909469
cob_internet_cens_2018	total_viviendas_cens_2018	0.903040
numero_hogares_dane	cob_internet_cens_2018	0.900610
cob_gas_cens_2018	porc_piso_adecuado_cens_2018	0.900545

cob_rec_bas_cens_2018	porc_piso_adecuado_cens_2018	0.894831
Poblacion	cob_internet_cens_2018	0.894140
porc_est_3_cens_2018	afil_subsidiado	0.891444
porc_est_3_cens_2018	poblacion_5_16_mined	0.890426
cob_gas_cens_2018	cob_rec_bas_cens_2018	0.888799
ninos_icbf_mined	ninos_prescolar_mined	0.882281
cob_alcant_cens_2018	porc_piso_adecuado_cens_2018	0.878721
Poblacion	ninos_icbf_mined	0.876453
cob_internet_cens_2018	poblacion_5_16_mined	0.865621
cob_alcant_cens_2018	cob_gas_cens_2018	0.864895
numero_hogares_dane	ninos_icbf_mined	0.861782
per_ipm_sisben	per_tab_inf_sisben	0.858381
cob_internet_cens_2018	afil_subsidiado	0.858362
total_viviendas_cens_2018	ninos_icbf_mined	0.852536
cob_gas_cens_2018	porc_pared_adecuada_cens_2018	0.852200
poblacion_5_16_mined	ninos_prescolar_mined	0.847895
ninos_prescolar_mined	afil_subsidiado	0.845678
porc_est_3_cens_2018	ninos_icbf_mined	0.842140
cob_internet_cens_2018	porc_piso_adecuado_cens_2018	0.833495
porc_est_2_cens_2018	porc_piso_adecuado_cens_2018	0.824946
cob_internet_cens_2018	ninos_icbf_mined	0.823755
porc_pared_adecuada_cens_2018	porc_piso_adecuado_cens_2018	0.822315
cob_rec_bas_cens_2018	porc_pared_adecuada_cens_2018	0.806057
ninos_icbf_mined	afil_contributivo	0.805300
tasa_alfab_cens_2018	cob_elect_cens_2018	0.800485
Poblacion	ninos_prescolar_mined	0.788687
cob_alcant_cens_2018	porc_est_2_cens_2018	0.785656
cob_alcant_cens_2018	porc_pared_adecuada_cens_2018	0.773823
numero_hogares_dane	ninos_prescolar_mined	0.761613
porc_pared_inadecuada_cens_2018	porc_piso_inadecuado_cens_2018	0.755407
total_viviendas_cens_2018	ninos_prescolar_mined	0.755064
per_tab_inf_sisben	hogares_sisbenizados	0.743946
cob_alcant_cens_2018	cob_internet_cens_2018	0.728062
tasa_analf_cens_2018	porc_pared_inadecuada_cens_2018	0.727857
porc_est_3_cens_2018	porc_piso_adecuado_cens_2018	0.717899
total_viviendas_cens_2018	tasa_mort_general_minsal	0.714740
cob_rec_bas_cens_2018	porc_est_2_cens_2018	0.714601
per_tab_inf_sisben	viviendas_sisbenizados	0.714149
cob_gas_cens_2018	cob_internet_cens_2018	0.713667
cob_rec_bas_cens_2018	cob_internet_cens_2018	0.712797
cob_internet_cens_2018	ninos_prescolar_mined	0.708696
numero_hogares_dane	tasa_mort_general_minsal	0.703115
afil_contributivo	tasa_mort_general_minsal	0.702688

ha_under_illegal_explotation	tasa_alfab_cens_2018	-0.715312
porc_piso_inadecuado_cens_2018	afil_contributivo	-0.717999
hacinamiento_promedio_dane	cob_elect_cens_2018	-0.723262
cob_acued_cens_2018	porc_pared_inadecuada_cens_2018	-0.725872
porc_est_2_cens_2018	porc_piso_inadecuado_cens_2018	-0.739793
cob_alcant_cens_2018	porc_pared_inadecuada_cens_2018	-0.752070
porc_est_3_cens_2018	porc_piso_inadecuado_cens_2018	-0.755117
ha_under_illegal_explotation	cob_elect_cens_2018	-0.781636
cob_gas_cens_2018	porc_piso_inadecuado_cens_2018	-0.782501
cob_alcant_cens_2018	porc_piso_inadecuado_cens_2018	-0.782572
cob_rec_bas_cens_2018	porc_pared_inadecuada_cens_2018	-0.791348
porc_pared_inadecuada_cens_2018	porc_piso_adecuado_cens_2018	-0.796208
cob_rec_bas_cens_2018	porc_piso_inadecuado_cens_2018	-0.805930
cob_gas_cens_2018	porc_pared_inadecuada_cens_2018	-0.823558
cob_internet_cens_2018	porc_piso_inadecuado_cens_2018	-0.836662
porc_piso_adecuado_cens_2018	porc_piso_inadecuado_cens_2018	-0.897670
tasa_alfab_cens_2018	tasa_analf_cens_2018	-0.902823
porc_pared_adecuada_cens_2018	porc_pared_inadecuada_cens_2018	-0.964497

Fuente: Construcción propia

Con base en los datos de correlación presentados en la tabla anterior, se procede a tomar decisiones encaminadas a reducir la dimensionalidad del dataset. En este proceso, se analiza cada variable en conjunto con todas sus correlaciones relevantes, y a medida que se deciden eliminaciones o fusiones de variables, se actualiza la información para evitar redundancias en el análisis. De esta forma, si una variable es eliminada, se excluye de futuras revisiones para no caer en repeticiones innecesarias.

Al revisar las correlaciones de la variable población con otras variables incluidas en la tabla, se pueden extraer las siguientes conclusiones:

- numero_hogares_dane y total_viviendas_cens_2018 presentan correlaciones extremadamente altas de 0.997 y 0.995, respectivamente, con población. Dado que ambas variables describen aspectos similares de la estructura poblacional, se decide eliminarlas y conservar población debido a su mayor capacidad para capturar una interpretación más amplia y representativa del fenómeno demográfico.
- Las variables afil_subsidiado y afil_contributivo tienen correlaciones de 0.985 y 0.967 con población. Aunque estas variables son clave para entender la cobertura en salud por régimen, es más eficiente fusionarlas en una sola que represente la afiliación total al sistema de salud, sin comprometer la riqueza informativa pero simplificando la estructura del análisis.

- El porcentaje de personas en estrato 3 (porc_est_3_cens_2018) presenta una correlación de 0.914 con población, lo que indica una fuerte relación con el tamaño poblacional. Se procede a eliminarla dado que su inclusión podría generar redundancia con otras variables que ya aportan información poblacional.

En la tabla también se revela una fuerte relación (0.942) entre la cobertura neta (cobertura_neta_mined) y bruta (cobertura_bruta_mined) en educación básica y media vocacional. Ambas variables miden el acceso a la educación, pero la cobertura neta ofrece una visión más precisa al considerar únicamente a los estudiantes de la edad adecuada para cada nivel educativo. Por lo tanto, se procede a conservar la cobertura neta como indicador principal.

La correlación extremadamente alta (0.999999) entre la variable tasa_matricula_5_16_mined y cobertura_neta_mined sugiere una redundancia significativa en la información, ya que ambas miden aspectos similares de la educación básica y media. Se procede a mantener cobertura neta y eliminar la tasa de matrícula.

El análisis de correlación de las variables socioeconómicas extraídas del SISBEN para los municipios del departamento del Cauca que presentan alta correlación, revela pares de variables con alta redundancia:

- hogares_sisbenizados vs. viviendas_sisbenizadas (0.955): Dada la alta correlación, se decide eliminar hogares_sisbenizados y conservar viviendas_sisbenizadas, que representa mejor la información de los municipios.
- per_ipm_sisben vs. per_tab_inf_sisben (0.858), que corresponde a personas con índice de pobreza multidimensional y personas con empleo informal, se opta por mantener solo la variable de pobreza multidimensional, ya que la informalidad laboral probablemente está capturada como uno de los factores que contribuyen a la pobreza multidimensional

La alta correlación de 0.882281 entre el número de niños atendidos por el ICBF (ninos_icbf_mined) y el número de niños matriculados en prescolar (ninos_prescolar_mined) que se muestra en la tabla, sugiere una relación significativa entre ambas variables, pero como representan diferentes grupos etarios, para optimizar el análisis y reducir la dimensionalidad del dataset, se procede fusionar estas variables en una nueva variable compuesta, que resulte de la suma de las dos variables, lo que permitiría capturar la atención infantil en el contexto de servicios sociales y educación prescolar en cada municipio.

La correlación observada entre las hectáreas de explotación ilegal de metales preciosos (ha_under_illegal_explotation) y el hacinamiento promedio según el censo del DANE

(hacinamiento_promedio_dane) calculado entre el número de personas en el hogar sobre el número de dormitorios, es notablemente alta (0.911835). A primera vista parece extraña, pero como la correlación sugiere una posible interrelación entre la explotación ilegal y las condiciones socioeconómicas de los hogares, se sugiere mantenerlas para analizar más adelantes su relación en un modelo de Machine Learning.

La correlación observada entre porc_est_2_cens_2018, porcentaje de viviendas en estrato 2 y porc_piso_inadecuado_cens_2018, porcentaje de viviendas con piso inadecuado es de menos -0.739793, mientras que la relación con porc_piso_adecuado_cens_2018, porcentaje de viviendas con piso adecuado es de 0.824946. Estas correlaciones sugieren que las condiciones del piso son inversamente proporcionales al estrato socioeconómico, además de un alto grado de redundancia entre las variables. Se opta por fusionar las variables de piso (inadecuado y adecuado) en una nueva variable que refleje la calidad del piso, así:

$$ind_cal_piso = \frac{porc_piso_adecuado_cens_2018}{porc_piso_adecuado_cens_2018 + porc_piso_inadecuado_cens_2018}$$

Siguiendo con el análisis de las variables socioeconómicas altamente correlacionadas según la tabla 4 y la matriz de correlaciones de figura 43, en las condiciones estructurales de las viviendas, se observan las siguientes relaciones:

- porc_pared_adecuada_cens_2018 y porc_pared_inadecuada_cens_2018 presentan una correlación negativa muy fuerte (-0.964497), lo que indica que ambas variables capturan prácticamente la misma información desde perspectivas opuestas. Se procede a eliminar porc_pared_inadecuada_cens_2018, ya que la versión positiva de la medición suele ser más intuitiva para interpretar en los análisis.
- porc_pared_adecuada_cens_2018 y porc_piso_adecuado_cens_2018 tienen una correlación significativa positiva (0.822315), lo que sugiere que ambas variables reflejan, en gran medida, las condiciones de infraestructura de las viviendas. Sin embargo, dado que la calidad de las paredes y el piso son dimensiones diferentes de habitabilidad, ambas variables son importantes para una evaluación completa.
- porc_piso_adecuado_cens_2018 y porc_piso_inadecuado_cens_2018 tienen una correlación negativa muy fuerte (-0.897670), lo que indica una duplicación clara en la información. De nuevo, se procede a conservar porc_piso_adecuado_cens_2018 y eliminar porc_piso_inadecuado_cens_2018 por las mismas razones que en el caso de las paredes.

Deteniéndose en la variable tasa_alfab_cens_2018, se observa que está altamente correlacionada con tasa_analf_cens_2018 (-0.902823) y con cob_elect_cens_2018 (0.800485). En el primer caso se muestra una fuerte redundancia entre ambas variables, dado que una puede deducirse directamente de la otra, se procede a eliminar la variable tasa_analfab_cens_2018, para mantener un enfoque positivo en la educación. En el segundo caso, aunque la correlación es alta, son variables que capturan fenómenos socioeconómicos diferentes, la tasa de alfabetización y la cobertura eléctrica, por lo tanto se mantienen las dos en el dataset.

Para finalizar este primer análisis de altas correlaciones de las variables socioeconómicas, se observa que varias de las tasas que miden la cobertura de servicios públicos domiciliarios, tales como cob_elect_cens_2018, cob_acued_cens_2018, cob_alcant_cens_2018, cob_gas_cens_2018, cob_rec_bas_cens_2018 y cob_internet_cens_2018, presentan correlaciones elevadas entre sí o con otras variables del conjunto de datos. Esto sugiere una redundancia en la información aportada por estas variables, lo cual abre la posibilidad de fusionarlas en una métrica compuesta que represente de manera general el acceso a servicios públicos en cada municipio.

Con el objetivo de reducir la dimensionalidad del dataset y simplificar el análisis sin perder una medida representativa del acceso a estos servicios, se procede a la creación de un índice de cobertura de servicios públicos. Este índice se construye a partir de un promedio ponderado de las coberturas individuales, donde se asignan pesos diferenciados según la importancia relativa de cada servicio en el contexto socioeconómico del departamento. Por ejemplo, servicios como electricidad, agua y alcantarillado reciben un mayor peso en la ponderación debido a su impacto crítico en la calidad de vida, mientras que internet o la recolección de basuras podrían tener un peso menor.

$$\text{La fórmula es: } ind_serv_pcos_domic = \frac{\sum(\text{cobertura individual} \times \text{peso correspondiente})}{\sum \text{pesos}}$$

Este enfoque permite reducir la complejidad del análisis sin perder la capacidad de evaluar el acceso general a servicios básicos en cada municipio, aunque se sacrifica el detalle específico de cada uno de estos servicios.

Después de hacer todos los ajustes anteriormente referidos a las variables socioeconómicas derivados del análisis de correlación volvemos a generar un resumen de los pares de variables socioeconómicas que persisten con un correlación elevada, mayor a 0.7 (positiva o negativa) en el dataset, las que se muestran en la siguiente tabla.

Tabla 6

Pares de variables socioeconómicas altamente correlacionadas después de ajustes

Variable 1	Variable 2	Correlación
Poblacion	afil_total_salud	0.99637
ha_under_illegal_explotation	hacinamiento_promedio_dane	0.91184
afil_total_salud	atencion_inf_prescolar	0.87574
Poblacion	atencion_inf_prescolar	0.87497
porc_pared_adecuada_cens_2018	ind_cob_serv_pcos_dom	0.85389
ind_cal_piso	ind_cob_serv_pcos_dom	0.84016
porc_est_2_cens_2018	ind_cal_piso	0.81396
porc_pared_adecuada_cens_2018	ind_cal_piso	0.80756
tasa_alfab_cens_2018	ind_cob_serv_pcos_dom	0.73294
porc_est_2_cens_2018	ind_cob_serv_pcos_dom	0.71571
ha_under_illegal_explotation	tasa_alfab_cens_2018	-0.71531

Fuente: Construcción propia

Del análisis de correlaciones altas presentado en la tabla anterior se desprende:

- población vs. afil_total_salud (0.99637): La correlación extremadamente alta entre estas dos variables sugiere una redundancia significativa. Esto es porque el número de afiliados a la seguridad social está directamente relacionado con la población total. Se procede a sustituir afil_total_salud que solo hace un conteo de población afiliada a salud por una tasa de afiliación a la seguridad social en salud (afil_total_salud / ‘poblacion’), que puede reflejar de manera más precisa el acceso a servicios de salud, independientemente del tamaño de la población.
- población vs. atencion_inf_prescolar (0.87497): Dado que la correlación entre estas variables es alta, pero miden aspectos diferentes (población total vs. atención infantil), no se recomienda una eliminación o fusión en este punto. Ambas son relevantes para el análisis.
- porc_pared_adecuada_cens_2018, ind_cob_serv_pcos_dom y ind_cal_piso: La alta correlación superior al 0.84 indica que la calidad de las paredes y de las paredes en las viviendas está relacionada con la cobertura de servicios públicos. Se procede a fusionar estas variables en un índice de condiciones de la vivienda, calculado como el promedio simple de las variables estandarizadas.
- porc_est_2_cens_2018 vs. ind_cal_piso (0.81396): Aunque existe una alta correlación, estas variables miden aspectos diferentes (estrato socioeconómico vs. calidad del piso de las viviendas). Se mantienen ambas variables por separado.

- tasa_alfab_cens_2018 vs. ind_cob_serv_pcos_dom (0.73294): Aunque la correlación es alta, estas variables miden factores muy diferentes (alfabetización y acceso a servicios públicos). Se mantienen las variables.
- porc_est_2_cens_2018 vs. ind_cob_serv_pcos_dom (0.71571): Esta correlación indica una relación entre el estrato socioeconómico y la cobertura de servicios públicos. No obstante se conservan ambas variables por separado, ya que el estrato refleja un aspecto más amplio que no se captura completamente en el acceso a servicios.
- ha_under_illegal_explotation vs. tasa_alfab_cens_2018 (-0.71531): La correlación negativa indica una relación inversa entre hectáreas con explotación ilegal de metales preciosos y la tasa de alfabetización. Ambas variables reflejan fenómenos distintos y críticos para el análisis, por lo que deben mantenerse por separado.

Factor de Inflación de la Varianza (VIF). Para profundizar en el análisis de la multicolinealidad presente en el conjunto de datos, se procede a calcular el Factor de Inflación de la Varianza (VIF). A pesar de haber realizado un análisis de correlación previo, el VIF nos proporciona una perspectiva adicional y más precisa sobre la colinealidad entre las variables. Mientras que la correlación mide la relación lineal entre dos variables, el VIF evalúa el impacto de todas las variables independientes en la varianza de un coeficiente de regresión en particular. De esta manera, el VIF puede detectar problemas de multicolinealidad que no son evidentes en la matriz de correlaciones (Quintana Romero & Mendoza. , 2019).

En la Figura de la siguiente página, se presenta el código Python utilizado para calcular el VIF de las 26 variables socioeconómicas restantes y los resultados obtenidos, ordenados de mayor a menor VIF. Esta información es importante para identificar aquellas variables que contribuyen significativamente a la multicolinealidad y tomar decisiones informadas sobre su inclusión o exclusión en el modelo final.

En la figura se observan varias variables con valores altos del Factor de Inflación de la Varianza (VIF), que sugieren multicolinealidad, lo cual podría afectar la robustez de los modelos de machine learning. A continuación, se discuten las variables a eliminar o fusionar, teniendo en cuenta además la posible redundancia de algunas variables:

- hacinamiento_promedio_dane (VIF: 28.85) y per_haci_critico_sisben (VIF: 7.14): Ambas variables reflejan aspectos de hacinamiento en los municipios. La primera capture el promedio de hogares sobre dormitorios, mientras que la segunda se enfoca en las personas en hacinamiento crítico según el SISBEN IV. Dado el alto VIF de hacinamiento_promedio_dane se procede a eliminarla y a mantener per_haci_critico_sisb.

Figura 47*Cálculo del VIF de las variables socioeconómicas presentes en el dataset*

```

# Aplicar VIF
# Extraer solo las columnas socioeconómicas
X = df_final[var_socieconomicas]

# Añadir una constante al modelo
X = sm.add_constant(X)

# Calcular el VIF para cada variable
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

# Eliminar la fila correspondiente a la constante
vif_data = vif_data[vif_data["Variable"] != "const"]

# Ordenar el VIF de mayor a menor
vif_data = vif_data.sort_values(by="VIF", ascending=False)

# Mostrar el VIF para cada variable
print(vif_data)

```

	Variable	VIF
7	hacinamiento_promedio_dane	28.848236
2	ha_under_illegal_explotation	27.569961
1	poblacion	15.447699
26	indice_condiciones_vivienda	14.767771
24	atencion_inf_prescolar	13.887879
6	tasa_alfab_cens_2018	12.768679
13	porc_est_2_cens_2018	11.667544
23	viviendas_sisbenizados	10.228181
14	cobertura_neta_mined	9.739930
20	per_ipm_sisben	8.602870
12	porc_est_1_cens_2018	8.440188
25	tasa_afiliacion_salud	8.181436
22	per_haci_critico_sisben	7.141483
16	aprobacion_mined	4.874000
17	tasa_mort_general_minsal	4.841941
21	casos_trab_infan_sisben	4.571345
9	tasa_acceso_a_cocina_dane	4.184704
11	prom_hog_vivienda_cens_2018	4.126722
8	densidad_habitacional_dane	3.724352
15	desicion_mined	3.574193
3	ha_condicionada_upra	3.417139
19	porc_bajo_peso_nacer_minsal	2.864843
5	desempleo_cens_2018	2.447001
10	tasa_acceso_agua_cocinar_dane	2.405006
4	ha_no_condicionada_upra	2.018694
18	tasa_mort_ninez_minsal	1.710122

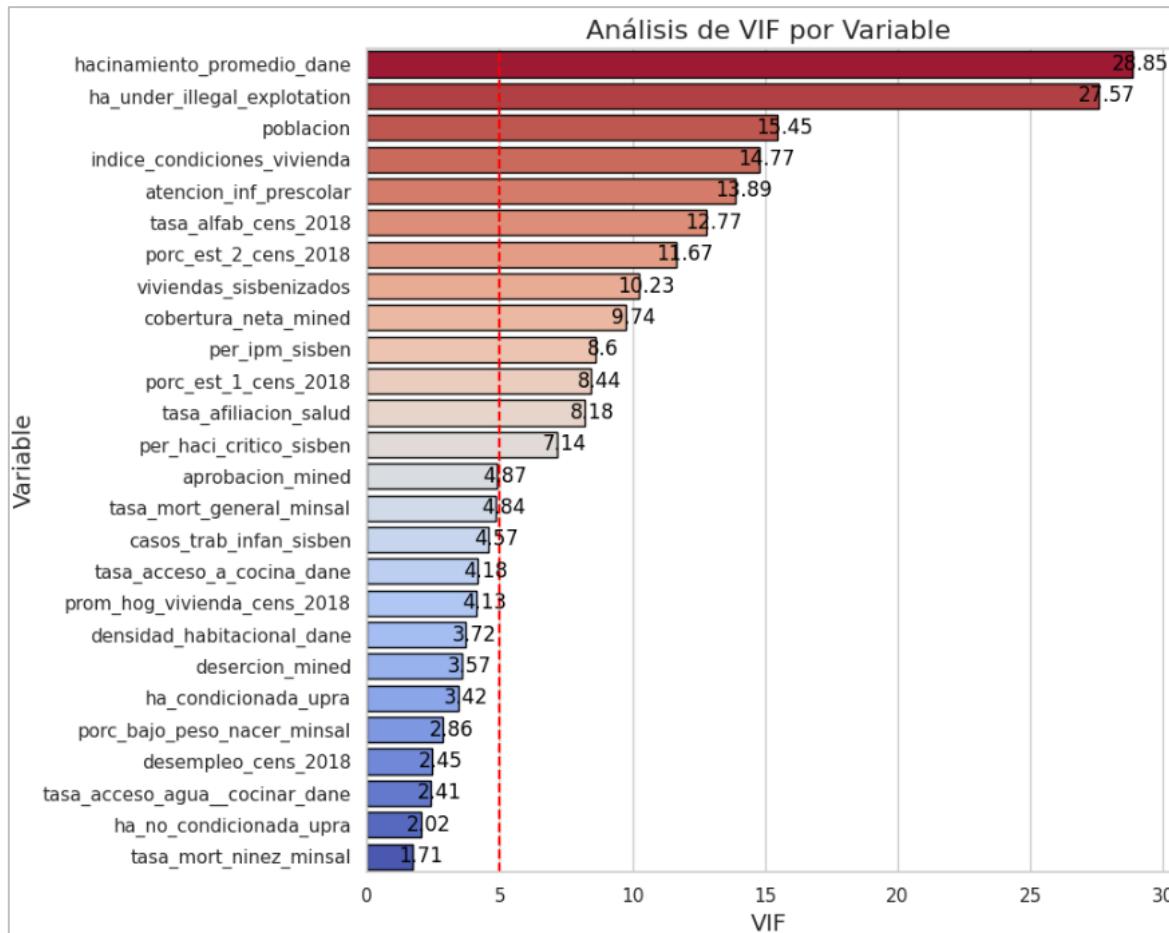
Fuente: Pantallazo obtenido desde el entorno de Google Colab

- ha_under_illegal_explotation (VIF: 27.57): Aunque presenta un VIF elevado, esta variable es de alta relevancia en el análisis, dado que refleja un fenómeno económico y social crítico en el contexto de la violencia (explotación ilegal de metales preciosos). Su importancia en el análisis de hechos de violencia justifica su retención, pese a la multicolinealidad, como ya se había establecido en el análisis de correlación.
- ‘poblacion’ (VIF: 15.45): Es una variable que inevitablemente está correlacionada con otras variables agregadas, como atencion_inf_prescolar (VIF: 13.88) y con

tasa_afiliacion_salud (VIF:8.18). Sin embargo la variable ‘poblacion’ se retiene como referencia base para ponderaciones o tasas y como principal indicador demográfico.

Figura 48

Gráfico de barras del VIF de las variables socioeconómicas del dataset



Fuente: Pantallazo obtenido desde el entorno de Google Colab

- indice_condiciones_vivienda (VIF: 14.76), tasa_acceso_agua_cocinar_dane (VIF: 2.41), densidad_habitacional_dane (VIF: 3.72), tasa_acceso_a_cocina_dane (VIF: 4.18), prom_hog_vivienda_cens_2018 (VIF: 4.12): Estas variables describen las condiciones estructurales de las viviendas y servicios básicos. Dado que el índice_condiciones_vivienda ya capture una medida compuesta, que fusiona esa clase de variables de condiciones de infraestructura y servicios se procede a mantenerla a pesar del elevado VIF y desechar las otras.
- atencion_inf_prescolar (VIF: 13.88): Esta variable que incluye tanto la atención del ICBF como la matrícula en prescolar, además de tener un elevado VIF está altamente correlacionada con variables como ‘poblacion’, pero se conserva dado que su carácter agregado la hace relevante.

- tasa_alfab_cens_2018 (VIF: 12.76) y cobertura_neta_mined (VIF: 9.74): Ambas variables están relacionadas con la educación, aunque desde diferentes perspectivas (alfabetismo vs. cobertura educativa). Se propone retener ambas, ya que aportan información complementaria relevante en el análisis socioeconómico y educativo.
- Se prioriza el indicador de pobreza multidimensional (per_ipm_sisben – VIF: 8.6) en lugar de los casos de trabajo infantil (casos_trab_infan_sisben – VIF: 4.6). Aunque ambos son indicadores de vulnerabilidad social, el índice de pobreza multidimensional proporciona una perspectiva más completa de la situación, al considerar múltiples dimensiones de la pobreza. Esta decisión se basa en la necesidad de un enfoque más general en el análisis de la vulnerabilidad social.

Una vez aplicadas al dataset las decisiones tomadas a partir del análisis del VIF anteriormente presentado, el número de variables socioeconómicas se reduce a 20 y si calculamos los nuevos valores VIF sobre las mismas se aprecia que los valores de los factores disminuyen notablemente.

Figura 49

Recalcularando el VIF de las variables socioeconómicas que quedaron

```
# Aplicar VIF nuevamente
X = df_var_socieconomicas[var_socieconomicas]
X = sm.add_constant(X)
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif_data = vif_data[vif_data["Variable"] != "const"]
vif_data = vif_data.sort_values(by="VIF", ascending=False)
print(vif_data)
```

	Variable	VIF
20	indice_condiciones_vivienda	12.108004
18	atencion_inf_prescolar	11.786351
1	poblacion	10.857337
6	tasa_alfab_cens_2018	9.163933
17	viviendas_sisbenizados	8.941216
8	porc_est_2_cens_2018	8.001834
15	per_ipm_sisben	7.386931
19	tasa_afiliacion_salud	5.993253
9	cobertura_neta_mined	5.669347
2	ha_under_illegal_explotation	5.487263
16	per_haci_critico_sisben	4.209474
7	porc_est_1_cens_2018	3.699205
12	tasa_mort_general_minsal	3.614158
11	aprobacion_mined	3.419998
10	desercion_mined	3.048213
3	ha_condicionada_upra	2.974959
14	porc_bajo_peso_nacer_minsal	2.627862
4	ha_no_condicionada_upra	1.853980
5	desempleo_cens_2018	1.594362
13	tasa_mort_ninez_minsal	1.404730

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Análisis de Información Mutua. A continuación, se aborda el análisis de información mutua sobre estas variables, el cual busca evaluar la dependencia estadística entre ellas. Aunque la información mutua es más comúnmente utilizada en problemas supervisados, su aplicación en este contexto no supervisado puede ayudar a detectar redundancias o similitudes significativas entre variables, contribuyendo a optimizar el desempeño de los algoritmos al eliminar atributos que no añaden valor informativo relevante (Becker, 2024).

Los resultados de aplicar el cálculo de la información mutua sobre todas las combinaciones de las 20 variables que todavía permanecen en el dataset se presentan en la matriz de la siguiente página. Los valores de información mutua representan la cantidad de dependencia o "información compartida" entre dos variables. Valores altos indican que las dos variables tienen una fuerte dependencia entre sí, es decir, que conocen mucha información una de la otra. Si el valor es cercano a 1, implica que gran parte de la variabilidad de una variable puede explicarse por la otra. Un valor bajo, por el contrario indica poca dependencia entre las variables, lo que sugiere que las dos variables son prácticamente independientes. Si la información mutua es cercana a 0, significa que conocer una (Becker, 2024) variable no ayuda a predecir la otra.

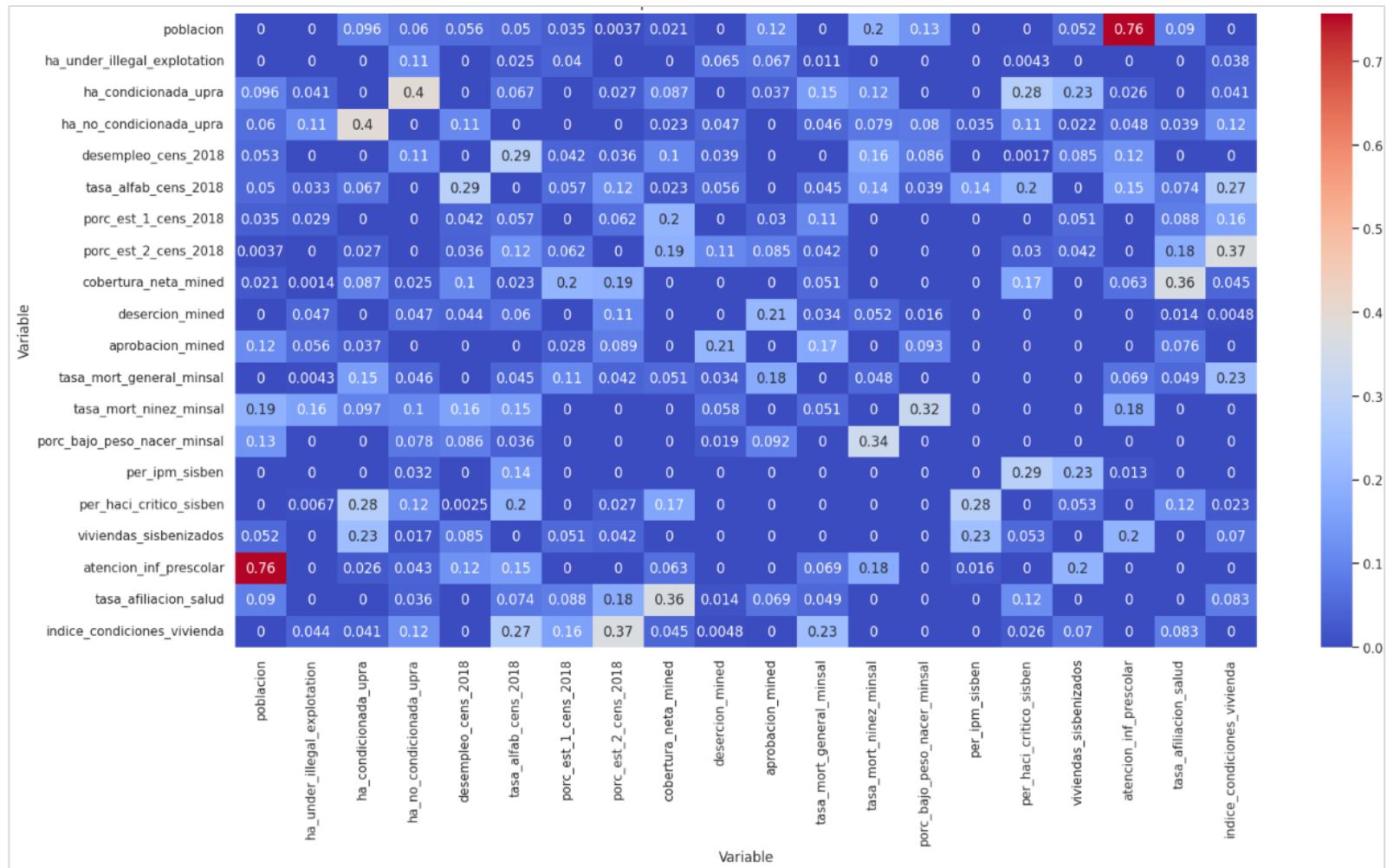
Si dos variables tienen un valor de información mutua muy alto (digamos 0.7 en adelante), podría considerarse eliminar una de ellas, ya que ambas aportan prácticamente la misma información. Si dos variables tienen una información mutua alta, pero ambas aportan valor desde una perspectiva interpretativa, una opción es fusionarlas para crear una nueva característica que combine la información que aportan. Si una variable no está correlacionada ni muestra alta información mutua con otras, puede ser valiosa para aportar diversidad de información al modelo (Bobadilla, 2021).

Los resultados del análisis de información mutua revelan que la mayoría de las variables socioeconómicas que quedan presentan valores relativamente bajos de dependencia mutua entre ellas, lo que indica que la información compartida es baja. Sin embargo, destaca la relación entre las variables 'poblacion' y 'atencion_inf_prescolar', con un valor de 0.75. Esta relación, que ya se había identificado previamente a través de la correlación y el VIF, refleja la dependencia entre el tamaño poblacional de los municipios y el número de niños atendidos por el ICBF y en prescolar.

Dada la importancia conceptual de ambas variables en el contexto socioeconómico del estudio, se ha decidido no eliminarlas, ya que cada una captura aspectos clave y distintos sobre la dinámica poblacional y el acceso a servicios para la infancia, factores que podrían tener implicaciones en el análisis de patrones de violencia y otros fenómenos sociales.

Figura 50

Matriz de análisis de información mutua de todas las combinaciones de las variables socioeconómicas



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Importancia de la permutación. La importancia de la permutación permite evaluar la contribución de cada variable a un modelo, midiendo cómo cambia la métrica de rendimiento cuando se permuta aleatoriamente una variable. Si al permutar una variable el rendimiento del modelo no se ve significativamente afectado, entonces esa variable puede considerarse menos relevante o incluso redundante, facilitando su posible eliminación (Becker, 2024).

En nuestro proyecto, aunque no contamos con datos etiquetados, podemos aplicar la técnica de importancia de la permutación generando primero etiquetas artificiales que representen los patrones o grupos de datos. Para ello, se usa el algoritmo de K-Means que agrupa los municipios del Cauca en clústeres según su similitud socioeconómica. Estas etiquetas de clúster se usan luego como el objetivo en un modelo de clasificación, como Regresión Logística, que nos permitirá calcular la importancia de las variables para explicar estos grupos (Bobadilla, 2021).

El proceso detallado es el siguiente:

- Escalado de las variables socioeconómicas: Antes de aplicar K-Means, las variables socioeconómicas se escalan utilizando StandardScaler para garantizar que todas las características estén en la misma escala y no sean afectadas por diferencias de magnitud.
- Generación de etiquetas con K-Means: Se ejecuta K-Means con un número predeterminado de clústeres (en este caso, 5 clústeres), lo que genera etiquetas artificiales que identifican los diferentes grupos de municipios en perfiles socioeconómicos.
- Entrenamiento del modelo supervisado (Regresión Logística): Estas etiquetas artificiales se usan como el objetivo (`y_kmeans`) en un modelo de clasificación, en este caso, regresión logística. Este modelo trata de predecir a qué clúster pertenece cada municipio en función de las variables socioeconómicas.
- Cálculo de la Importancia de la Permutación: Una vez entrenado el modelo, aplicamos la técnica de importancia de la permutación. Consiste en medir cómo cambia el rendimiento del modelo (es decir, su capacidad para predecir correctamente los clústeres) cuando se permuta una variable. Si el rendimiento no cambia significativamente, se considera que la variable tiene poca relevancia para el modelo.

Los cálculos en Python se muestran en la figura siguiente. Antes de entrar a la interpretación de los resultados obtenidos es preciso aclarar que el hecho de no tener etiquetas reales del dataset y utilizar etiquetas artificiales generadas aplicando K-means sobre los mismas variables limita el alcance de los resultados, como se explica más adelante.

Figura 51

Técnica de la importancia de la permutación con regresión logística a partir de etiquetas artificiales generadas con K-means

```

from sklearn.cluster import KMeans
from sklearn.linear_model import LogisticRegression
from sklearn.inspection import permutation_importance
from sklearn.preprocessing import StandardScaler

# Solo variables socioeconómicas
X = df_final[var_socieconomicas]

# Escalar los datos
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Aplicar K-Means para obtener etiquetas artificiales
kmeans = KMeans(n_clusters=5, random_state=42)
y_kmeans = kmeans.fit_predict(X_scaled)

# Entrenar un modelo supervisado LogisticRegression
log_reg = LogisticRegression(max_iter=1000, random_state=42)
log_reg.fit(X_scaled, y_kmeans)

# Aplicar la importancia de la permutación
result = permutation_importance(log_reg, X_scaled, y_kmeans, n_repeats=10, random_state=42, n_jobs=-1)

# Obtener las importancias y ordenar
sorted_indices = result.importances_mean.argsort()[:-1:-1]

# Mostrar los resultados ordenados
for idx in sorted_indices:
    print(f'{var_socieconomicas[idx]}: {result.importances_mean[idx]:.3f}')

```

Variable	Importancia
ha_condicionada_upra	0.076
porc_est_2_cens_2018	0.074
aprobacion_mined	0.074
viviendas_sisbenizados	0.048
desercion_mined	0.038
per_haci_critico_sisben	0.036
ha_under_illegal_explotation	0.026
tasa_afiliacion_salud	0.026
ha_no_condicionada_upra	0.017
atencion_inf_prescolar	0.005
tasa_mort_ninez_minsal	0.005
tasa_alfab_cens_2018	0.000
desempleo_cens_2018	0.000
indice_condiciones_vivienda	0.000
porc_est_1_cens_2018	0.000
cobertura_neta_mined	0.000
tasa_mort_general_minsal	0.000
porc_bajo_peso_nacer_minsal	0.000
per_ipm_sisben	0.000
poblacion	0.000

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Los resultados muestran que las variables con la más alta importancia de permutación son ha_condicionada_upra, porc_est_2_cens_2018 y aprobacion_mined, es decir que para el modelo los perfiles socioeconómicos que crea k-means en este código, están mayormente influenciados por la vocación agrícola de la tierra, por la pertenencia al estrato 2 de la población que en cierta medida refleja niveles de ingresos de los municipios y por las tasas

de aprobación de los niveles de educación básica y media que quizá diferencia los municipios con distintas oportunidades educativas.

Las variables con importancia media, tales como viviendas_sisbenizados, desercion_mined, per_haci_critico_sisben, ha_under_illegal_explotation, tasa_afiliacion_salud, y ha_no_condicionada también contribuyen a la formación de los clústeres, pero no tanto como las primeras.

Llama la atención el elevado número de variables con poca o ninguna importancia atencion_inf_prescolar, tasa_mort_ninez_minsal, tasa_alfab_cens_2018, desempleo_cens_2018, indice_condiciones_vivienda, porc_est_1_cens, cobertura_neta_mined, tasa_mort_general_minsal, porc_bajo_peso_nacer_minsal, per_ipm_sisben, poblacion, lo que indica que estas variables no tienen importancia alguna y por lo tanto no contribuyeron a mejorar la predicción del modelo.

En un escenario de datos con etiquetas reales se tendría que proceder, sin dudarlo, a la eliminación de las variables que no tienen importancia para el modelo. Pero hay que recordar que los cálculos en este caso se están haciendo con etiquetas artificiales provenientes de las agrupaciones que hace k-means sobre las mismas variables, que solo son agrupaciones basadas en similitudes internas de las variables. Estas etiquetas actúan como una aproximación para evaluar la distribución de los datos y obtener indicios de posibles relaciones, pero no representan el fenómeno de interés real y final.

El cálculo de importancia de la permutación en este contexto de etiquetas artificiales, entonces, sirve principalmente para observar qué variables socioeconómicas contribuyen a distinguir estos grupos aproximados de manera interna. No obstante, eliminar variables basándose exclusivamente en estas etiquetas artificiales conlleva el riesgo de descartar características que podrían ser relevantes cuando el modelo se entrena con las etiquetas reales de violencia. La relación entre las variables socioeconómicas y la violencia puede diferir notablemente de las agrupaciones artificiales, y algunas variables que aquí parecen irrelevantes podrían resultar fundamentales en el análisis final.

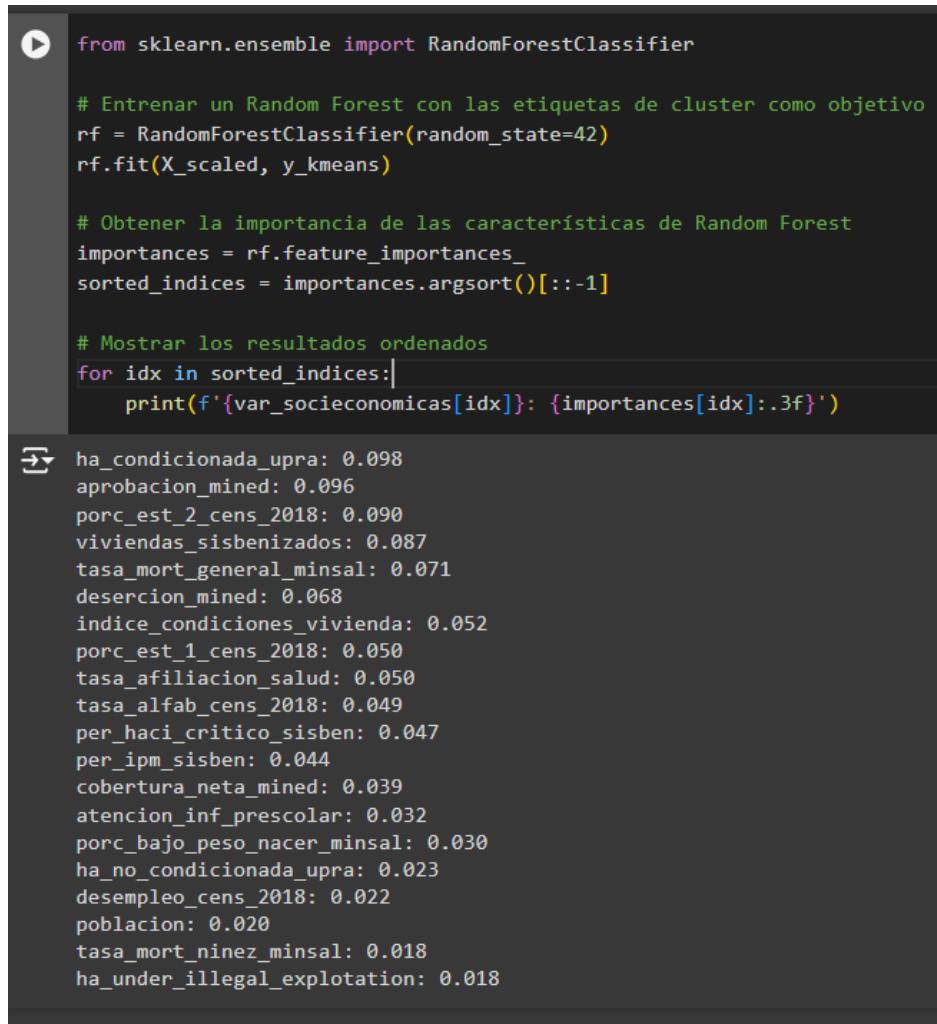
Importancia de características basada en la frecuencia de uso de cada variable en el modelo de Random Forest. Esta técnica mide la importancia de cada variable observando cuánto contribuye a la reducción de la impureza (como la impureza Gini) en los árboles de decisión que componen el modelo Random Forest. En cada árbol, las divisiones que utilizan una característica importante mejoran la pureza de los nodos. Por lo tanto, las variables más importantes son aquellas que aparecen con mayor frecuencia en las divisiones clave (Géron, 2023).

En nuestro caso, aunque estamos trabajando con un conjunto de datos no supervisado, utilizamos etiquetas artificiales generadas por K-Means (las mismas generadas para la técnica de importancia de las características ya comentada). Luego, aplicamos Random Forest como un modelo supervisado, usando estas etiquetas de clústeres como el objetivo, para calcular la importancia de las características. Esto nos ayuda a identificar las variables que más

contribuyen a la formación de los clústeres, aunque se debe advertir que las etiquetas no sean "reales" sino generadas por el algoritmo de clustering, de allí que tenga que tenerse cuidado con la interpretación de los resultados. El siguiente código Python implementa este proceso y muestra los resultados:

Figura 52

Importancia de características con Random Forest



```

from sklearn.ensemble import RandomForestClassifier

# Entrenar un Random Forest con las etiquetas de cluster como objetivo
rf = RandomForestClassifier(random_state=42)
rf.fit(X_scaled, y_kmeans)

# Obtener la importancia de las características de Random Forest
importances = rf.feature_importances_
sorted_indices = importances.argsort()[:-1]

# Mostrar los resultados ordenados
for idx in sorted_indices:
    print(f'{var_socieconomicas[idx]}: {importances[idx]:.3f}')

```

Variable	Importancia (aprox.)
ha_condicionada_upra	0.098
aprobacion_mined	0.096
porc_est_2_cens_2018	0.090
viviendas_sisbenizados	0.087
tasa_mort_general_minsal	0.071
desercion_mined	0.068
indice_condiciones_vivienda	0.052
porc_est_1_cens_2018	0.050
tasa_afiliacion_salud	0.050
tasa_alfab_cens_2018	0.049
per_haci_critico_sisben	0.047
per_ipm_sisben	0.044
cobertura_neta_mined	0.039
atencion_inf_prescolar	0.032
porc_bajo_peso_nacer_minsal	0.030
ha_no_condicionada_upra	0.023
desempleo_cens_2018	0.022
poblacion	0.020
tasa_mort_ninez_minsal	0.018
ha_under_illegal_explotation	0.018

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Al comparar los resultados de la importancia de características obtenidos con Random Forest y la importancia de la permutación, observamos una notable coherencia en el orden de las variables. Las variables aparecen consistentemente en el mismo orden o en un lugar muy cercano, lo cual sugiere que en ambos métodos las características tienen un peso muy similar entre los grupos generados por el algoritmo K-Means, lo que puede deberse al hecho de que se están utilizando las mismas etiquetas artificiales.

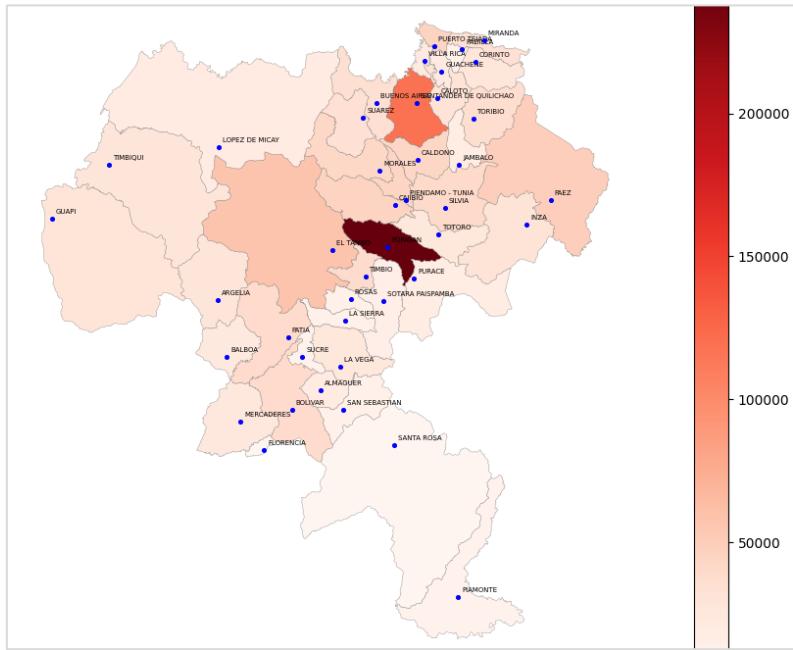
Sin embargo, se debe recordar que este análisis se basa en etiquetas artificiales, derivadas de un proceso no supervisado. Estas etiquetas no reflejan patrones observados en datos reales, sino que son agrupaciones creadas artificialmente, lo cual puede sesgar la interpretación de la importancia de cada variable. Debido a esto, resulta prudente evitar decisiones definitivas de eliminación de variables.

En lugar de eliminar características en este punto, es más adecuado posponer la reducción de dimensionalidad hasta aplicar PCA (Análisis de Componentes Principales) en un siguiente paso. De esta manera, se logrará una reducción más objetiva basada en la estructura intrínseca de los datos socioeconómicos, preservando la información relevante y minimizando el riesgo de perder variables que podrían ser fundamentales en el análisis posterior de correlación entre factores socioeconómicos y hechos de violencia en los municipios del Cauca.

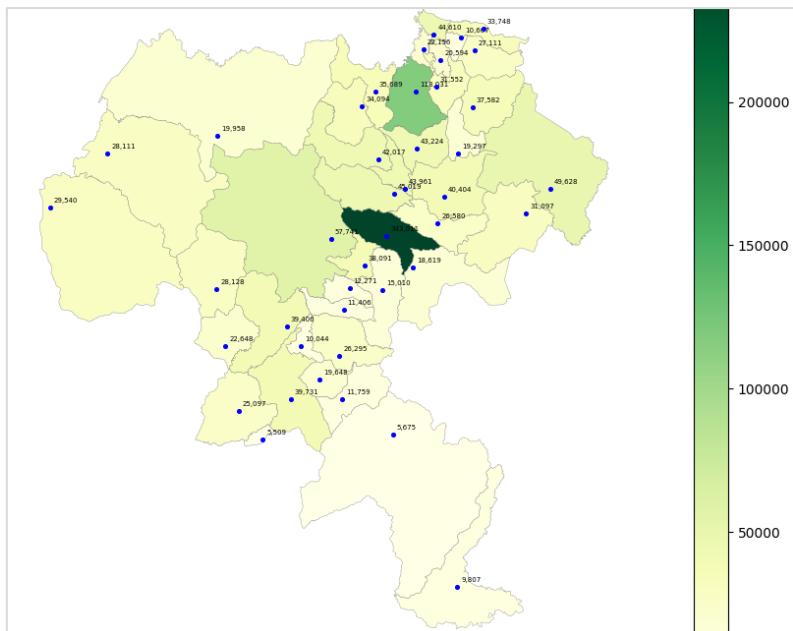
Análisis geoespacial de las variables socioeconómicas. A continuación se presentan las variables socioeconómicas que permanecen en el dataset tras el exhaustivo proceso de depuración, visualizadas mediante mapas georreferenciados del departamento del Cauca. Esta representación permite observar la distribución espacial y los patrones territoriales de cada variable, proporcionando un enfoque integral que contextualiza los fenómenos socioeconómicos en cada municipio.

El uso de datos georreferenciados, cuidadosamente integrados y estructurados en la base de datos relacional del proyecto, aporta un valor analítico significativo. La base de datos ha sido diseñada para incluir geometrías detalladas y consistentes de cada municipio y su cabecera, facilitando una representación precisa y visualmente intuitiva de cada área. Esta georreferenciación permite realizar comparaciones espaciales entre variables a nivel territorial, lo que resulta en una mejor identificación de tendencias y áreas críticas en el contexto regional.

En esta serie de mapas, creados con geopandas de Python, cada variable socioeconómica es visualizada en su propio título, facilitando la autocomprensión de los datos sin necesidad de comentarios adicionales. Se espera que cada título sirva de guía suficiente para que el lector interprete el contenido espacial y territorial de cada variable, maximizando la claridad y comprensión de los fenómenos observados.

Figura 53*Población del Departamento del Cauca - Nombres de municipios*

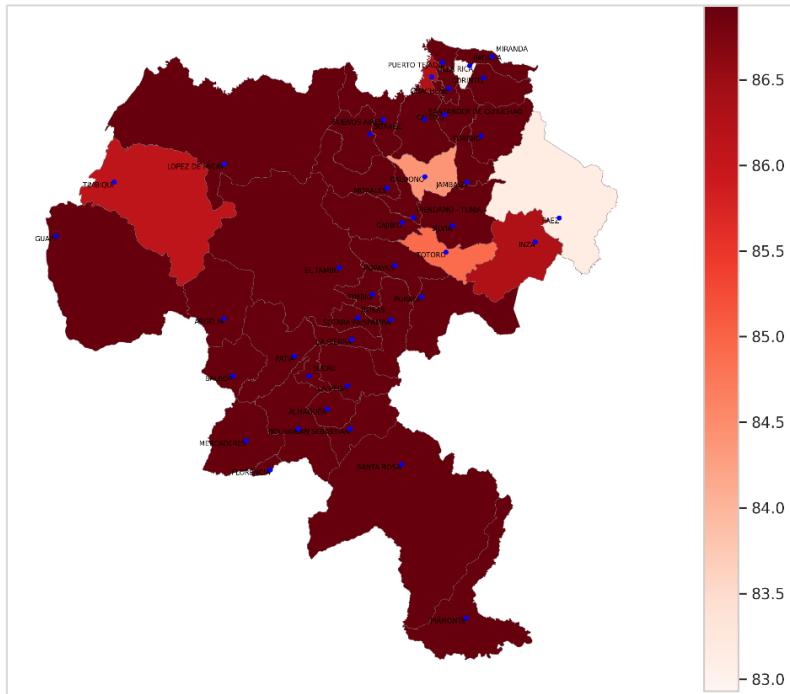
Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 54*Habitantes por municipio*

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 55

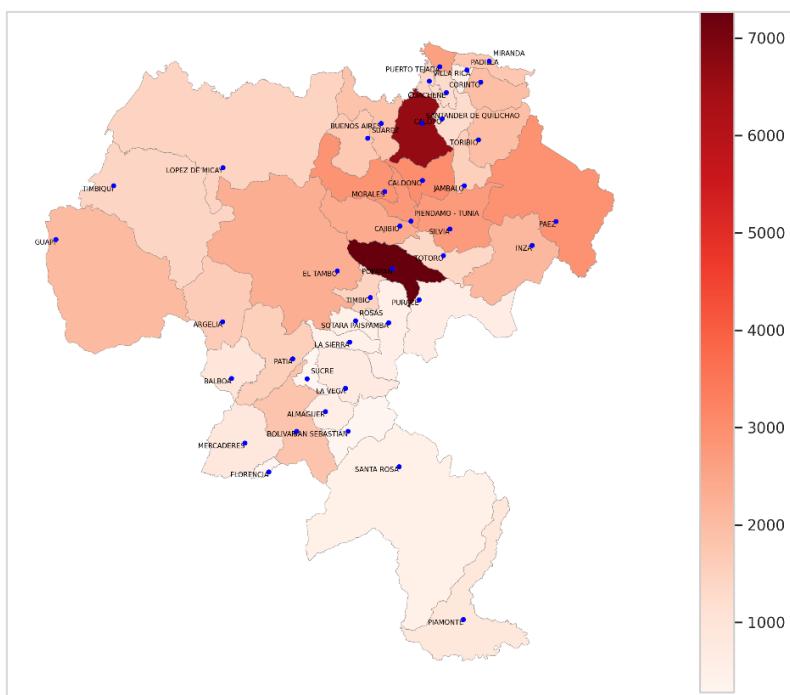
Tasa de aprobación en educación básica y media



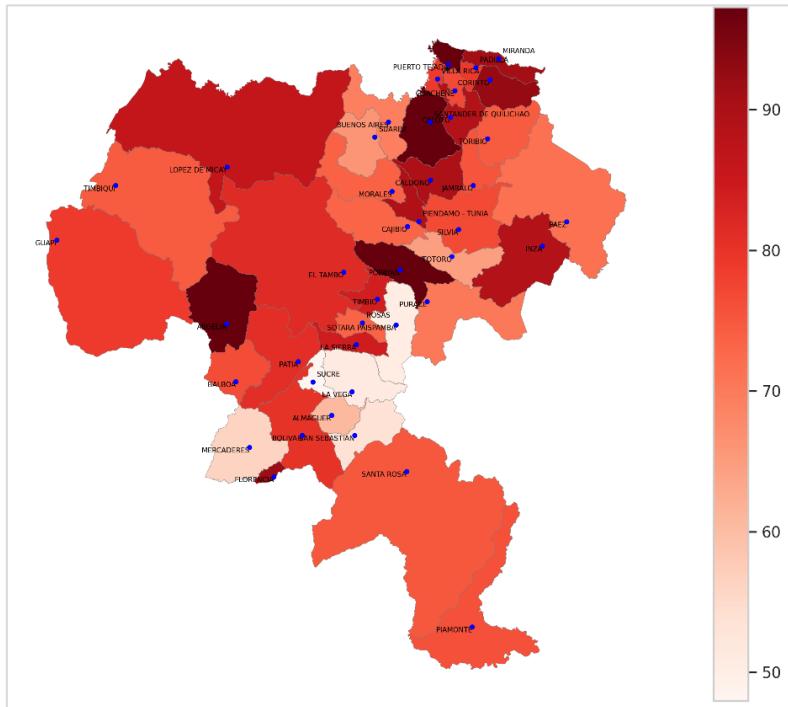
Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 56

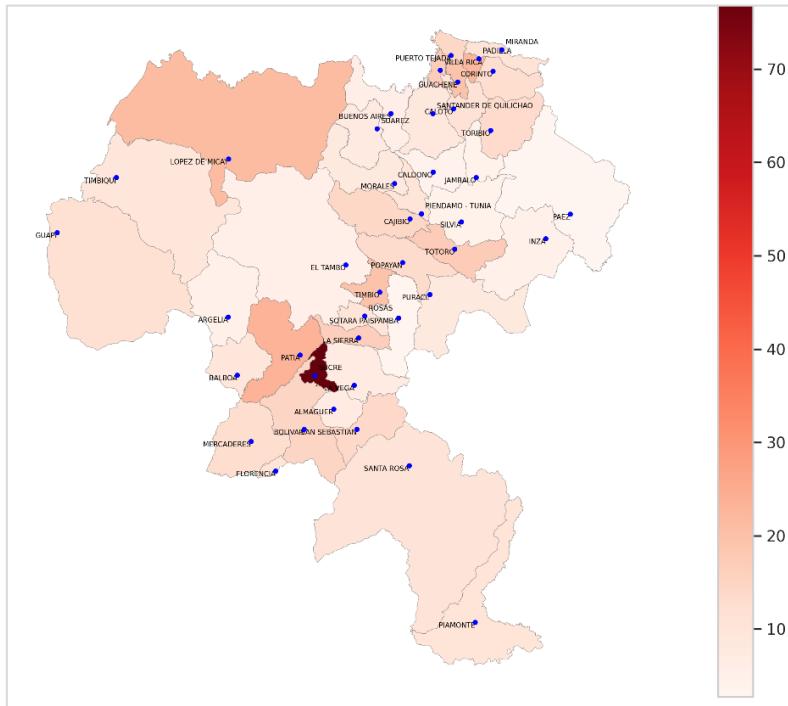
Atención Infantil en ICBF y Preescolar (número de niños atendidos)



Fuente: Pantallazo obtenido desde el entorno de Google Colab

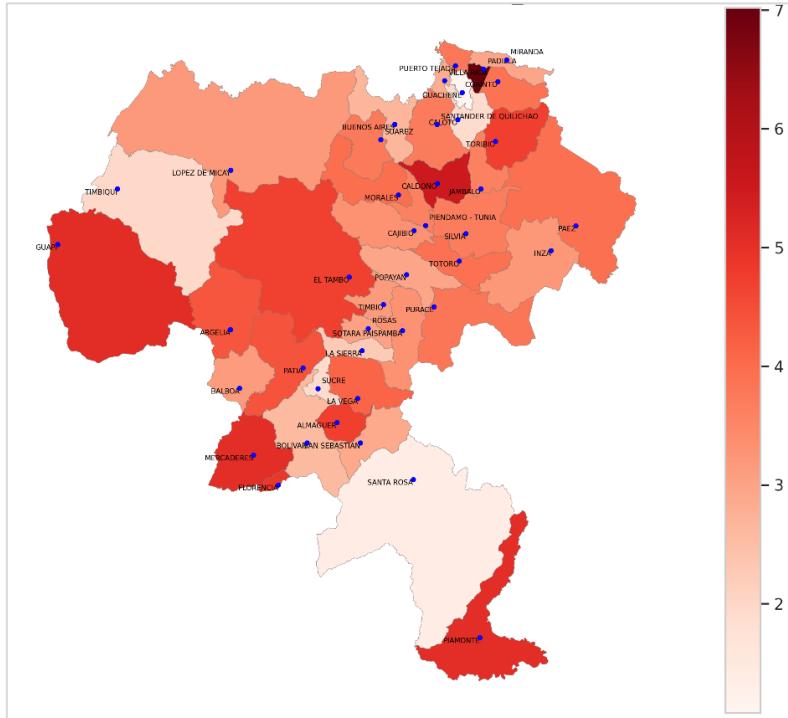
Figura 57*Cobertura neta en educación básica y media (Porcentaje 0-100)*

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 58*Tasa de desempleo (Censo 2018, Porcentaje 0-100)*

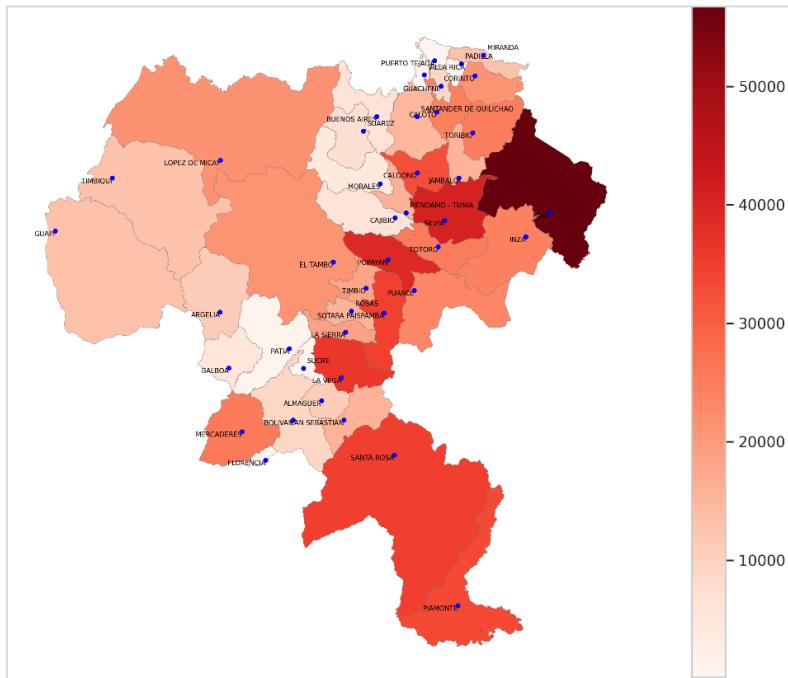
Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 59
Tasa de deserción escolar en educación básica y media



Fuente: Pantallazo obtenido desde el entorno de Google Colab

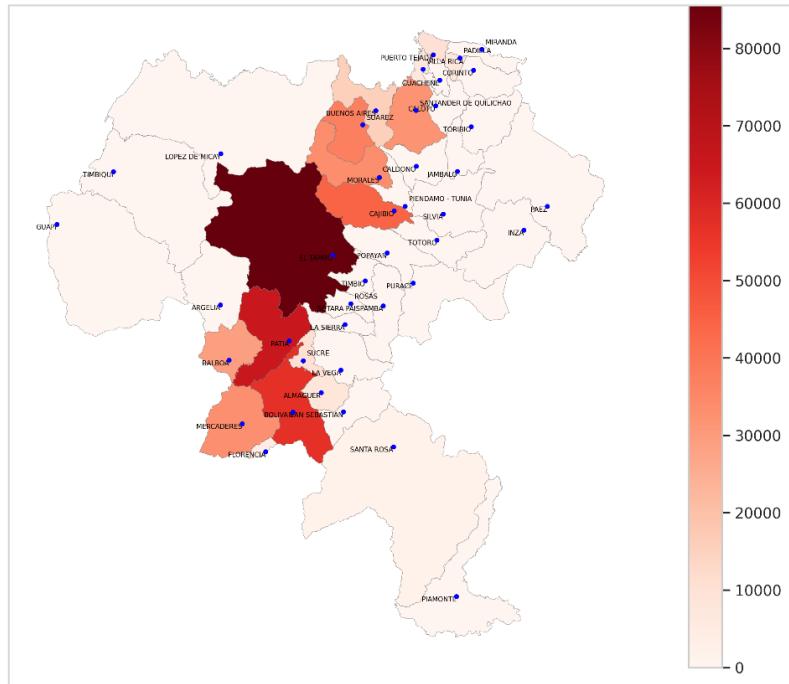
Figura 60
Hectáreas con vocación agrícola condicionada



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 61

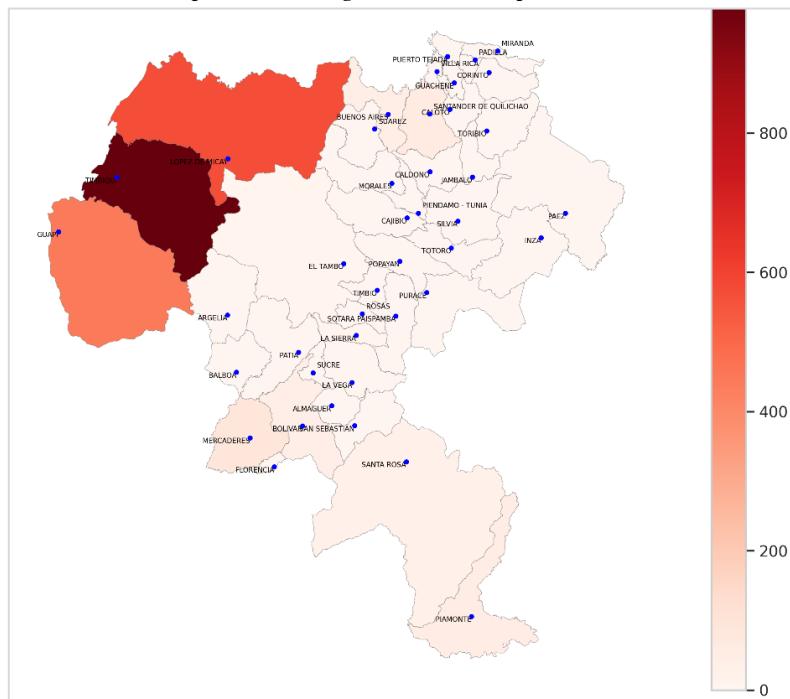
Hectáreas con vocación agrícola no condicionada



Fuente: Pantallazo obtenido desde el entorno de Google Colab

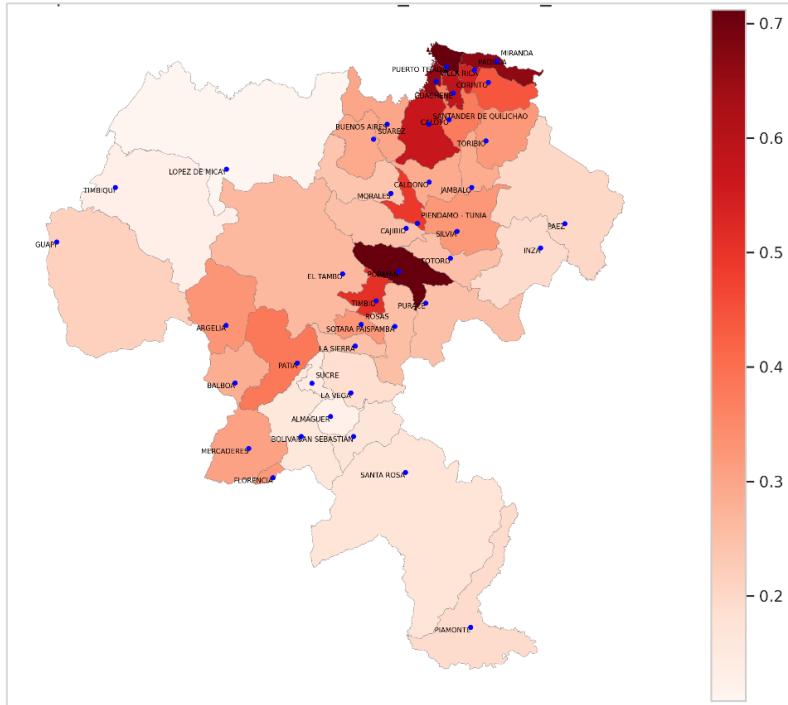
Figura 62

Hectáreas en explotación ilegal de metales preciosos



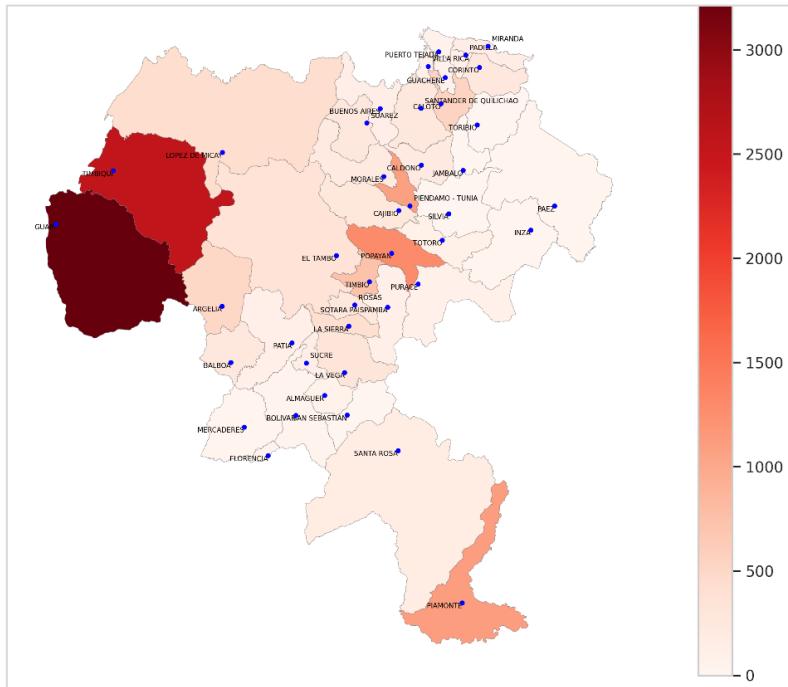
Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 63
Índice de Condiciones de Vivienda



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 64
Personas en Hacinamiento Crítico



Fuente: Pantallazo obtenido desde el entorno de Google Colab

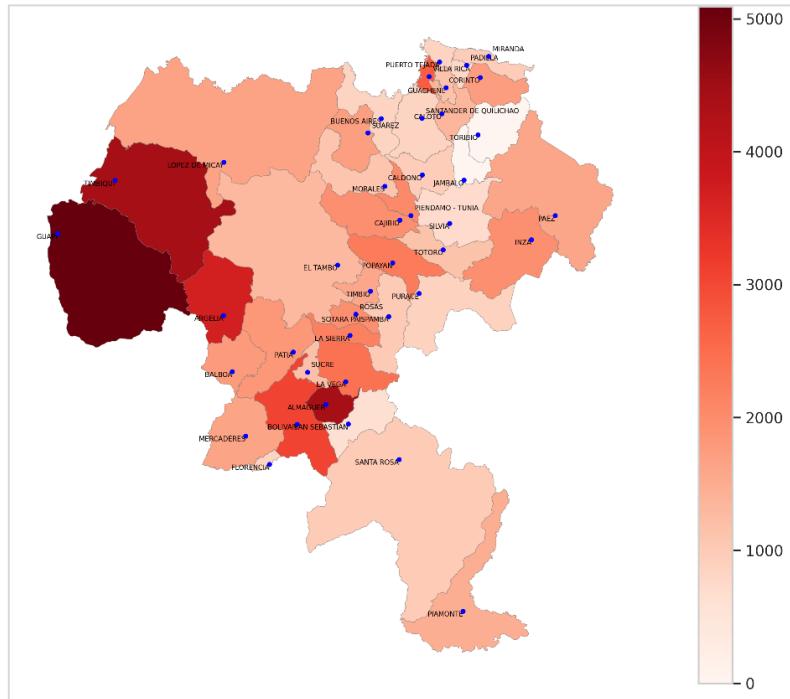
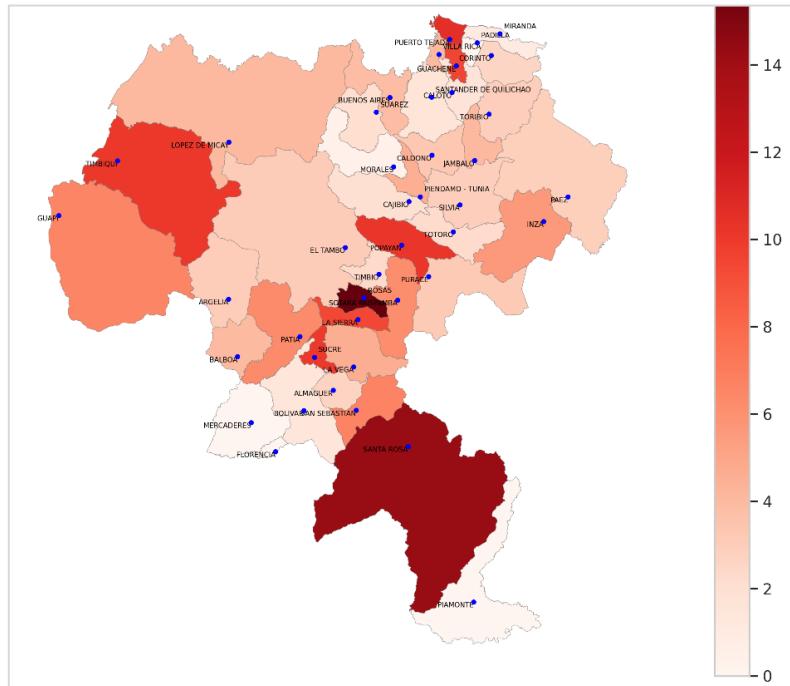
Figura 65*Personas con Indice de Pobreza Multidimensional**Fuente: Pantallazo obtenido desde el entorno de Google Colab***Figura 66***Porcentaje de Niños con Bajo Peso al Nacer**Fuente: Pantallazo obtenido desde el entorno de Google Colab*

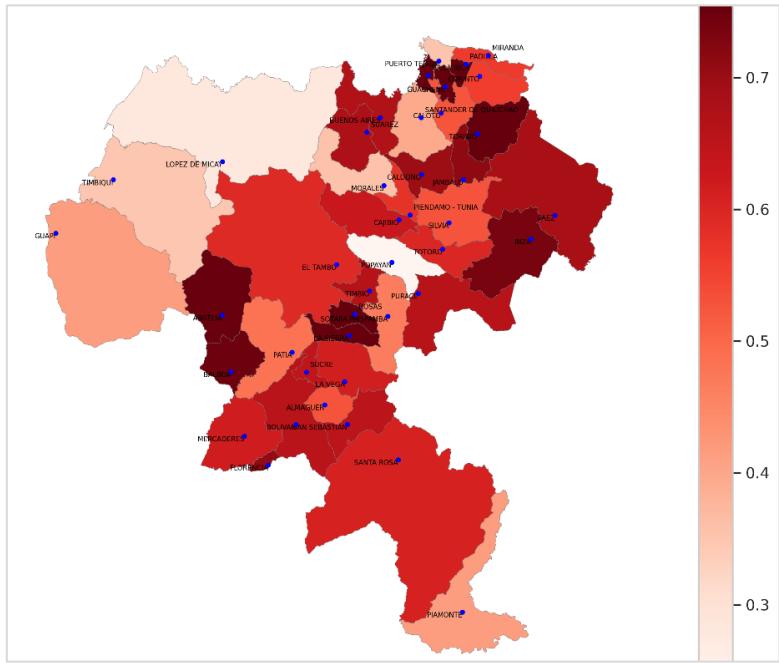
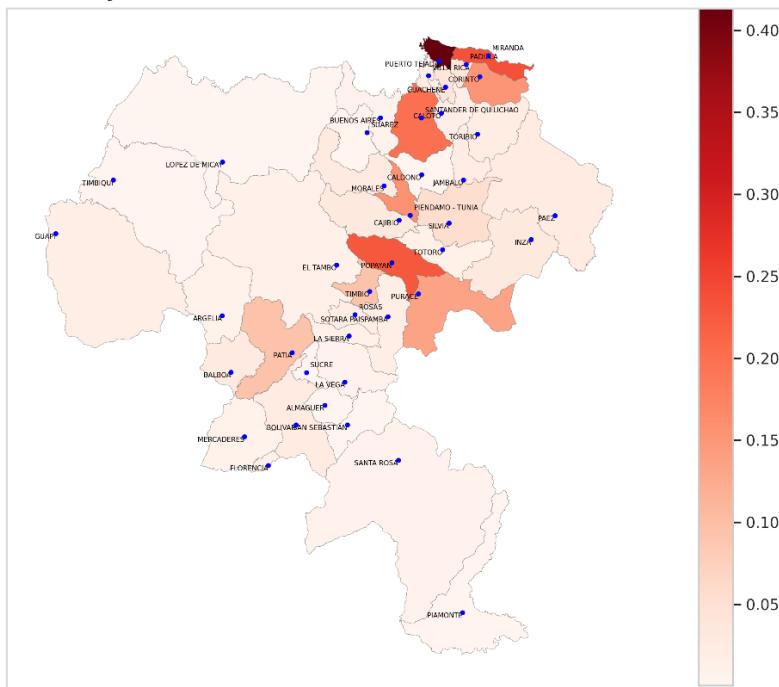
Figura 67*Porcentaje de Población en Estrato 1**Fuente: Pantallazo obtenido desde el entorno de Google Colab***Figura 68***Porcentaje de Población en Estrato 1**Fuente: Pantallazo obtenido desde el entorno de Google Colab*

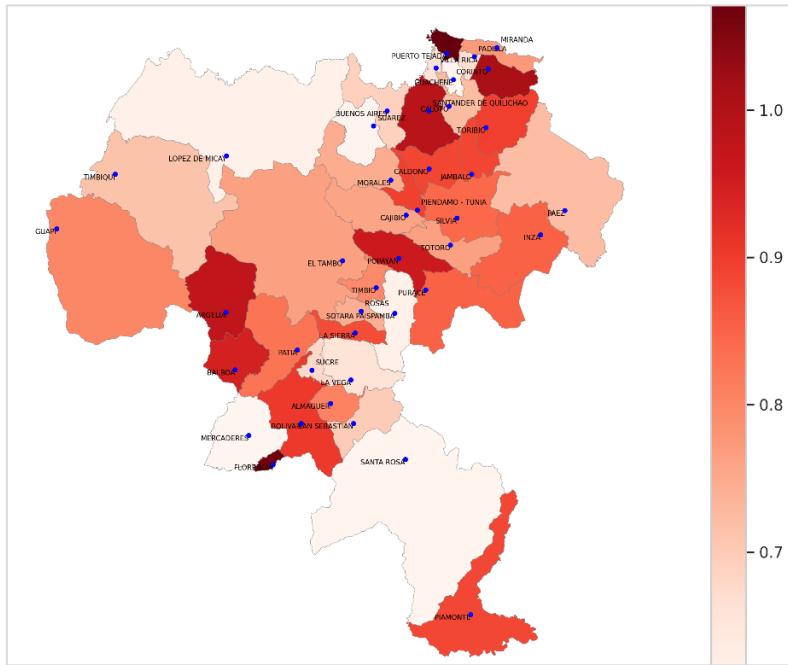
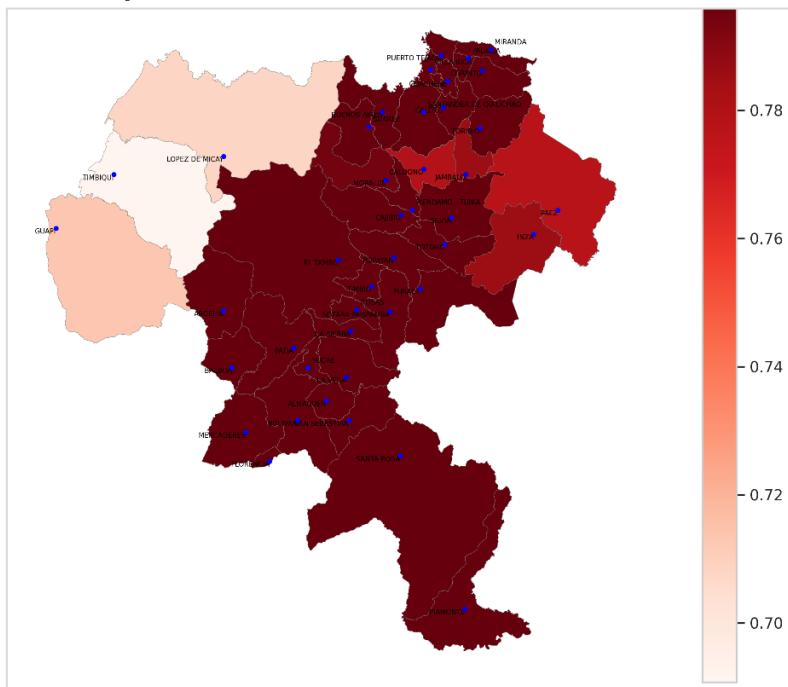
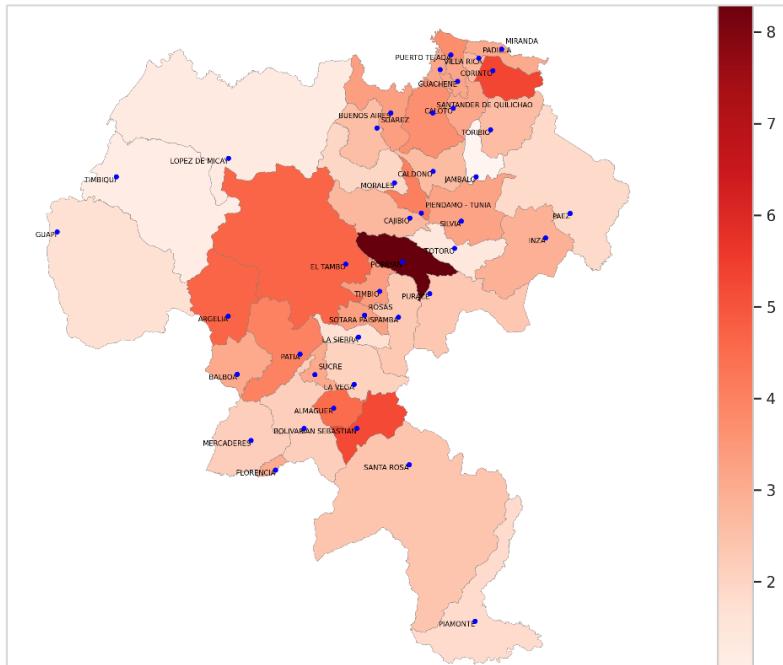
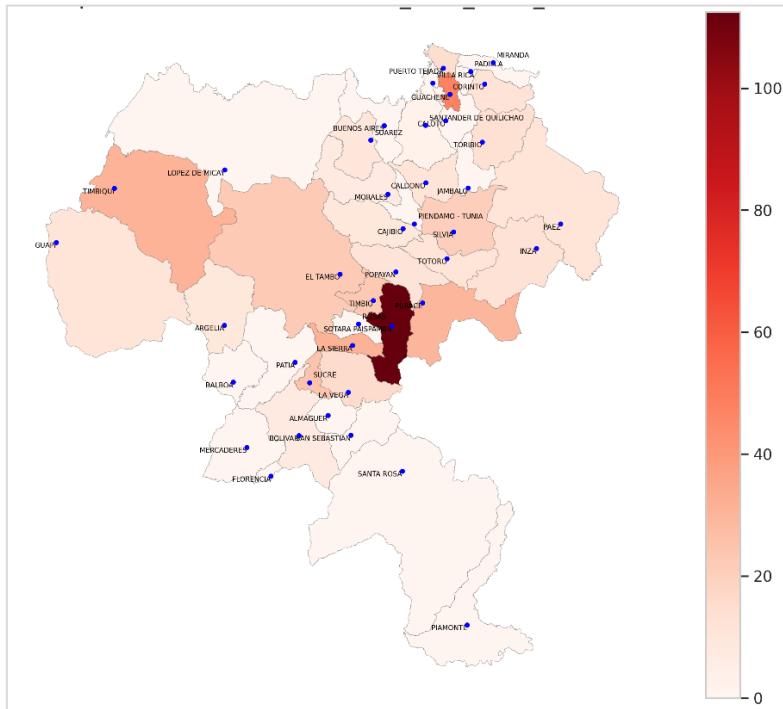
Figura 69*Afiliación a Seguridad Social en Salud**Fuente: Pantallazo obtenido desde el entorno de Google Colab***Figura 70***Tasa de Alfabetismo**Fuente: Pantallazo obtenido desde el entorno de Google Colab*

Figura 71
Tasa General de Mortalidad

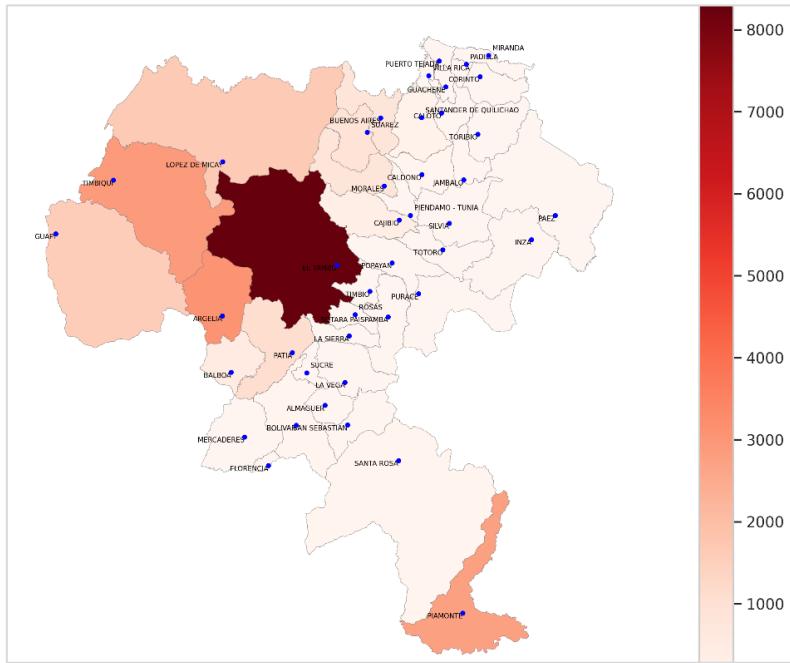


Fuente: Pantallazo obtenido desde el entorno de Google Colab

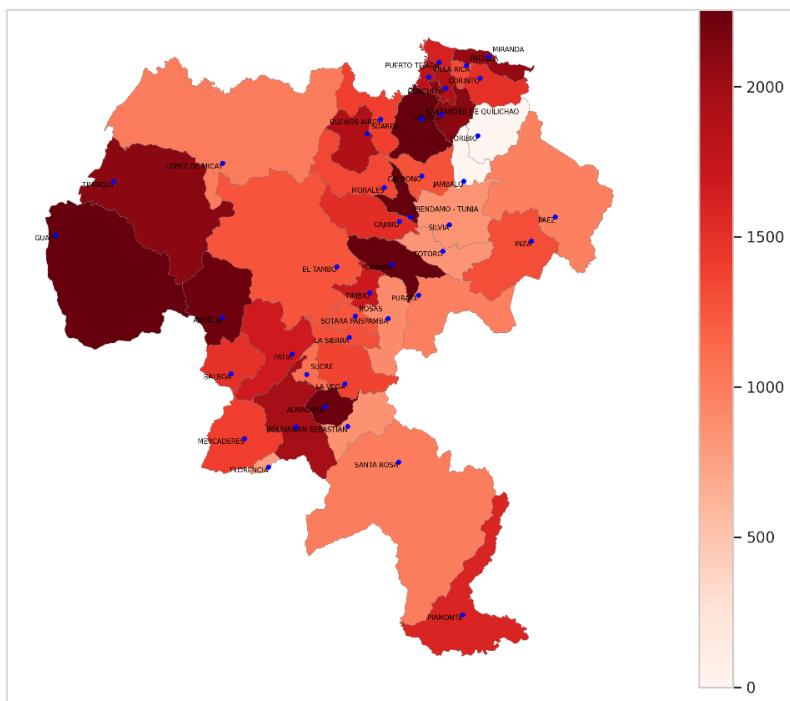
Figura 72
Tasa de Mortalidad Infantil



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 73*Mapa de hectáreas de hoja de coca reportadas por el Ministerio de Justicia*

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 74*Número de viviendas en el SISBEN*

Fuente: Pantallazo obtenido desde el entorno de Google Colab

2.4.3 Análisis y transformación de las variables sobre la violencia del dataset

Depuración inicial de las variables. En la tabla 3 se describen las 88 variables sobre hechos de violencia que contiene el dataset resultante del proceso de extracción. Se observa la existencia de múltiples fuentes que documentan los mismos hechos, como la Fiscalía, la Policía Nacional, el Ministerio de Defensa y Medicina Legal, lo cual genera redundancia. Dado que el objetivo del análisis es estudiar la violencia en el departamento del Cauca y no comparar fuentes, resulta más coherente concentrarse en una única fuente. Esto no solo optimiza la estructura del dataset, sino que también facilita un análisis más consistente de los patrones de violencia.

Se prioriza la Fiscalía como fuente de información sobre delitos, ya que sus datos suelen ser más rigurosos y fidedignos, al estar basados en investigaciones judiciales formales o denuncias, lo que confiere un mayor nivel de validación frente a otras fuentes, que podrían basarse en reportes preliminares sin investigación profunda. Medicina Legal, aunque precisa en términos forenses, se enfoca en lesiones y muertes, lo que puede no captar la totalidad de los eventos violentos si no son denunciados formalmente.

En cuanto a los datos de la Fiscalía, se dispone de tres aspectos clave de cada delito: los casos reportados, las personas indiciadas y las víctimas. De estos, se selecciona el reporte de "casos" por ser generalmente el más representativo para capturar la magnitud de la violencia en una región, ya que refleja la ocurrencia del delito. Los "indiciados" aportan información sobre el avance procesal y la efectividad de la justicia, mientras que las "víctimas" son fundamentales para un análisis social más detallado, aunque pueden estar sujetas a subregistro.

Adicionalmente, el dataset incluye información sobre víctimas del conflicto armado de dos fuentes: el Sistema de Información de Eventos de Violencia del Conflicto Armado en Colombia (SIEVCAC) y la Unidad para la Atención y Reparación Integral a las Víctimas. Se prefiere esta última por ofrecer datos más completos y detallados, seleccionándose la variable "número de víctimas del conflicto armado con declaración", que refleja un mayor nivel de formalidad en el registro de las víctimas.

Con esta depuración inicial, las 88 variables se reducen a 24, como se muestra en la tabla siguiente.

Tabla 7*Variables sobre la violencia después de la depuración inicial*

nombre_variable	Significado	Fuente
kls_basuco_incaut_mindef	Kilos de basuco incautados	Min-Defensa
ha_coca_cultivada_minjust	hectáreas de coca cultivada	Min-Defensa
kls_coca_confis_mindef	Kilos de coca incautados	Min-Defensa
kls_base_coca_confis_mindef	Kilos de base de coca incautados	Min-Defensa
cantidad_mariguana_confiscada	Kilos de marihuana incautados	Min-Defensa
amenazas_casos_fiscalia	Casos delitos de amenaza	Fiscalía General
delitos_sexuales_casos_fiscalia	Casos delitos sexuales	Fiscalía General
desaparicion_forzada_casos_fiscalia	Casos delitos desaparición forzada	Fiscalía General
desplazamiento_casos_fiscalia	Casos delitos desplazamiento	Fiscalía General
estafa_casos_fiscalia	Casos de delitos de estafa	Fiscalía General
extorsion_casos_fiscalia	Casos de delitos de extorsión	Fiscalía General
hurto_casos_fiscalia	Casos de delito de hurto	Fiscalía General
personas_bienes_dih_casos_fiscalia	Casos de delito contra personas y bienes protegidos por el DIH	Fiscalía General
reclutamiento_illicito_casos_fiscalia	Casos de delitos reclutamiento ilícito	Fiscalía General
violencia_intrafamiliar_casos_fiscalia	Casos de delitos violencia familiar	Fiscalía General
corrupcion_casos_fiscalia	Casos de delitos corrupción	Fiscalía General
homicidios_casos_fiscalia	Casos de delitos de homicidio	Fiscalía General
lesiones_personales_casos_fiscalia	Casos de delitos de lesiones personales	Fiscalía General
secuestro_casos_fiscalia	Casos de delitos de secuestro	Fiscalía General
otros_delitos_casos_fiscalia	Casos de otros delitos	Fiscalía General
vict_fuerza_pub_mindef	Número de víctimas de la fuerza pública	Min-Defensa
armas_confis_polinal	Numero de armas confiscadas	Policía Nacional
total_reclam_rest_t ierr_minagr	Total reclamantes de restitución de tierras	Min-Agricultura
vict_por_declarac_uv	Número de víctimas del conflicto armado con declaración	Unidad de víctimas

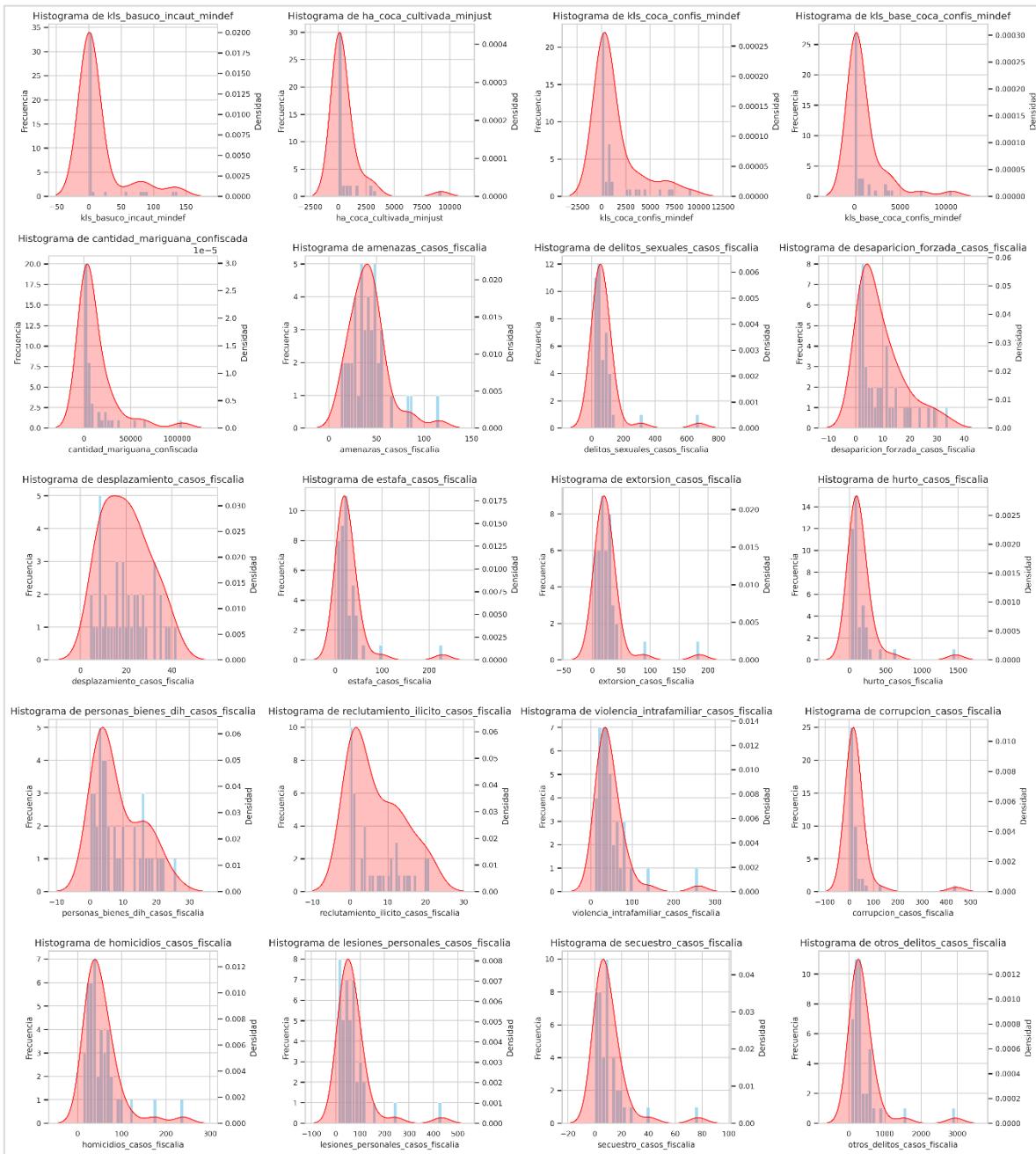
Fuente: Construcción propia

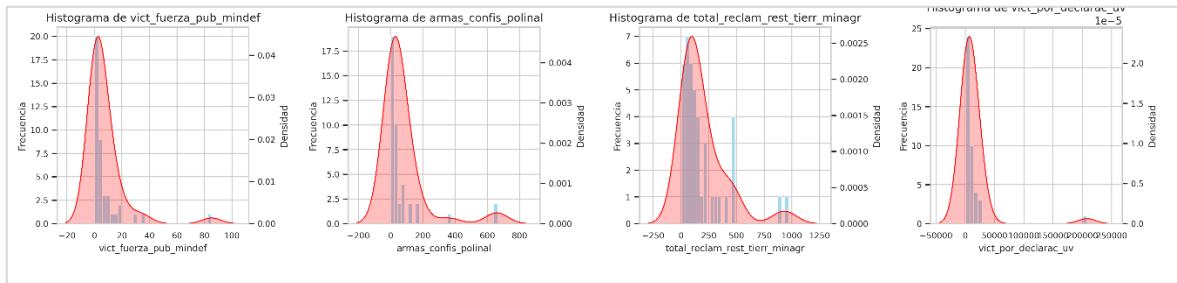
Antes de iniciar el análisis de las 24 variables resultantes de la depuración inicial, se aplican algunos ajustes al dataset. Las variables que representan la ocurrencia anual o el conteo de casos, como las relacionadas con delitos e incautaciones de narcotráfico, se acumulan para el periodo 2019-2023. Esto permite equilibrar su representatividad temporal frente a variables de reportes "con corte a", como 'vict_por_declarac_uv', 'ha_coca_cultivada_minjust' y 'total_reclam_rest_t ierr_minagr', las cuales ya reúnen datos de varios años. Este enfoque facilitará una comparación coherente entre fenómenos de diferente duración temporal, permitiendo un análisis balanceado de los datos y evitando que los eventos de corta duración sean subestimados frente a aquellos con periodos más largos.

Revisión de distribuciones de las variables. A continuación se presenta los histogramas de las variables sobre la violencia que quedaron después de la depuración inicial , con una curva superpuesta en el mismo gráfico para una mejor comparación visual.

Figura 75

Histogramas de las variables sobre violencia del dataset con curva de densidad superpuesta





Fuente: Pantallazo obtenido desde el entorno de Google Colab

El análisis de las distribuciones de las variables de violencia tras la depuración inicial revela patrones de alta asimetría y dispersión entre los 42 municipios del Cauca. Variables como las hectáreas de coca cultivada, los kilos de cocaína y marihuana confiscada presentan distribuciones extremadamente dispersas, con valores concentrados en unos pocos municipios y largas colas hacia la derecha. Esto indica que mientras la mayoría de los municipios reportan valores bajos o nulos, unos pocos exhiben cifras considerablemente altas, lo que afecta la interpretación de las medias y refleja la desigual distribución de estas actividades en el territorio.

Asimismo, se observa una alta concentración de ceros en variables relacionadas con incautaciones de drogas, cultivos ilegales y víctimas de la fuerza pública, lo que sugiere una actividad reducida o nula en varios municipios para estas variables, ya sea por falta de operaciones o por ausencia de registros. Estas características refuerzan la idea de que algunos municipios concentran los episodios más graves de violencia y narcotráfico.

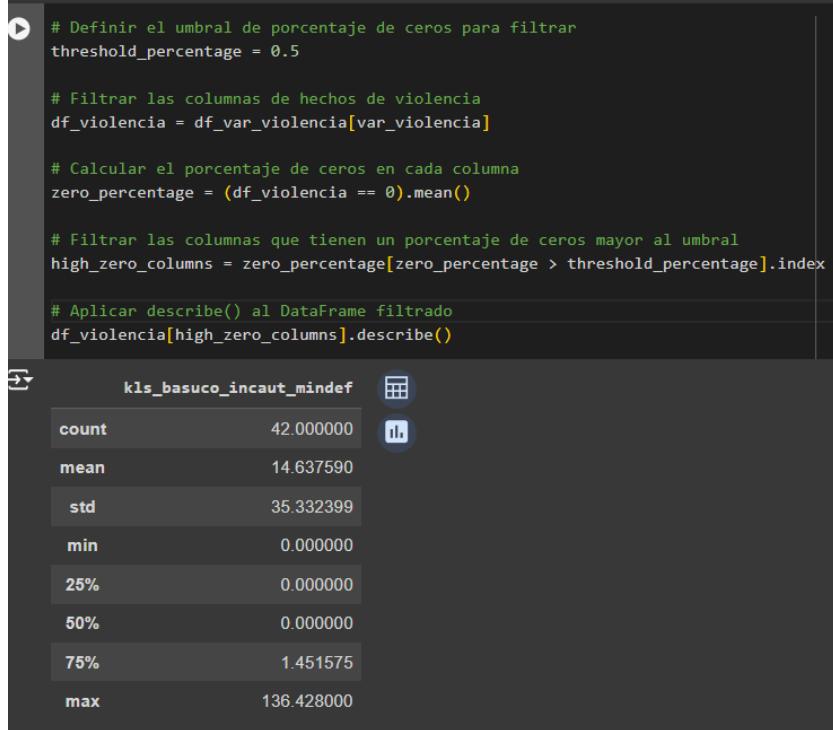
Por otro lado, variables como los homicidios y las lesiones personales reportadas por la Fiscalía muestran mayor variabilidad entre municipios, reflejando una distribución más dispersa en el territorio y valores extremos que sugieren diferencias significativas en las dinámicas locales de violencia. La presencia de estos valores extremos y la naturaleza sesgada de las distribuciones en varias variables sugieren la necesidad de un análisis más detallado, considerando la influencia desproporcionada de algunos municipios sobre la caracterización general de la violencia en el Cauca.

En las variables con una alta proporción de ceros, como las incautaciones de basuco y los cultivos ilegales, los ceros reflejan la falta de actividad reportada en algunos municipios, como se explicó cuando se hizo el proceso de imputación de nulos del dataset.

Para verificar este análisis, se filtran las variables sobre violencia con más del 50% de ceros como valores y se le aplica a la muestra la función describe() de pandas. Los resultados se muestran en la figura siguiente.

Figura 76

Variables sobre violencia con un alto porcentaje de valores en cero



```
# Definir el umbral de porcentaje de ceros para filtrar
threshold_percentage = 0.5

# Filtrar las columnas de hechos de violencia
df_violencia = df_varViolencia[varViolencia]

# Calcular el porcentaje de ceros en cada columna
zero_percentage = (df_violencia == 0).mean()

# Filtrar las columnas que tienen un porcentaje de ceros mayor al umbral
high_zero_columns = zero_percentage[zero_percentage > threshold_percentage].index

# Aplicar describe() al DataFrame filtrado
df_violencia[high_zero_columns].describe()
```

	kls_basuco_incaut_mindef
count	42.000000
mean	14.637590
std	35.332399
min	0.000000
25%	0.000000
50%	0.000000
75%	1.451575
max	136.428000

Fuente: Pantallazo obtenido desde el entorno de Google Colab

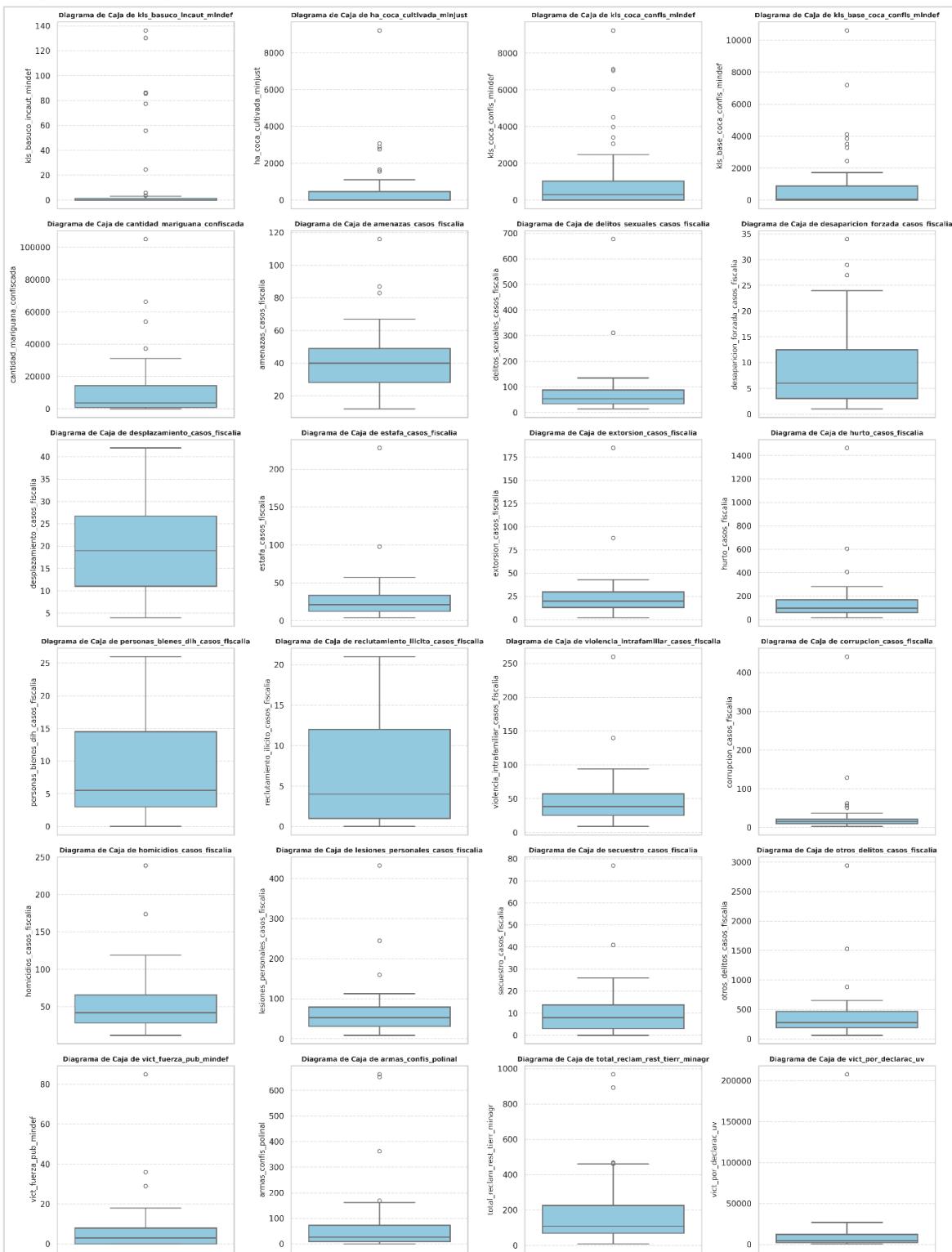
El análisis muestra que solo la variable 'kls_basuco_incaut_mindef' con alta proporción de ceros, lo que refleja que en la mayoría de los municipios no se registraron eventos asociados a esta categoría, lo que genera una distribución altamente concentrada en torno a la ausencia de eventos. Ante este escenario, sería razonable eliminar esta variable del análisis por su bajo nivel informativo y la escasa representatividad en los datos, sin embargo se decide mantenerla aún para evaluar la opción de fusionarla con otras relacionadas a la confiscación de drogas (coca, marihuana, etc.), lo cual podría permitir capturar un panorama más amplio de los esfuerzos de control de sustancias ilícitas en la región.

Para evaluar la dispersión, simetría y presencia de valores atípicos en las variables sobre la violencia del estudio, a continuación se generan los diagramas de caja (boxplots) para cada una de ellas. Los resultados se muestran en la figura de la página que sigue.

Las 24 variables de violencia presentan características particulares en sus distribuciones, con patrones que indican variaciones notables en la concentración y dispersión de los valores.

Figura 77

Diagramas de caja de las variables sobre la violencia del dataset



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Las variables de narcotráfico, tales como kilos de basuco incautados, hectáreas de coca cultivada, kilos de coca y base de coca incautados, y cantidad de marihuana confiscada, muestran cajas extremadamente pequeñas o invisibles en los diagramas de caja, junto con múltiples valores atípicos hacia el eje superior. Este patrón sugiere una alta concentración de valores bajos en la mayoría de los municipios, con un pequeño número de puntos donde la magnitud de la actividad es considerablemente mayor.

Las variables de ocurrencia de delitos y criminalidad, como homicidios, lesiones personales, otros delitos, y armas confiscadas, también presentan valores atípicos significativos en el extremo superior, acompañados de cajas pequeñas que indican una prevalencia de valores bajos en la mayoría de los casos y un número reducido de puntos con ocurrencias notablemente altas. En el ámbito de la violencia interpersonal y el conflicto armado, las variables de desaparición forzada, delitos contra personas y bienes protegidos por el derecho internacional humanitario, reclutamiento ilícito, desplazamiento forzado, víctimas del conflicto armado y reclamantes de restitución de tierras, presentan cajas de tamaño mediano a grande con una mediana centrada o desplazada, lo que sugiere una distribución menos sesgada. Sin embargo, los valores atípicos presentes en estas variables pueden estar asociados con municipios específicos donde el impacto del conflicto armado es marcadamente elevado.

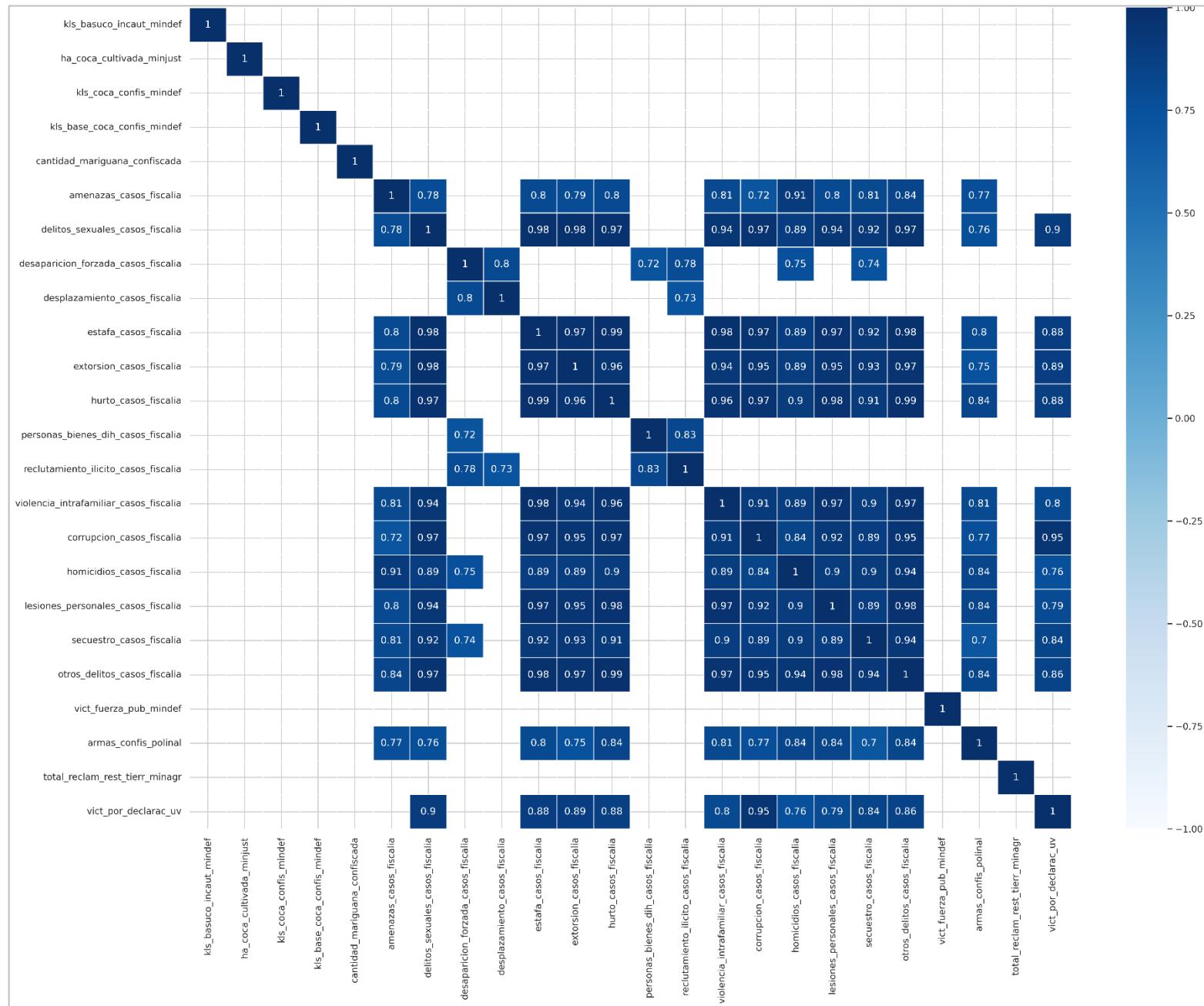
Finalmente, las variables económicas y de delitos contra la propiedad, como corrupción, estafa y hurto, se caracterizan por cajas pequeñas con valores concentrados y varios valores atípicos en el eje superior. Esta distribución sugiere una disparidad en la incidencia de estos delitos, reflejando una alta concentración de casos en ciertos municipios. La identificación de estos patrones en las distribuciones es importante para realizar los ajustes estadísticos adecuados, como la transformación logarítmica o el escalado, de forma que se minimice la influencia desproporcionada de los valores atípicos y se permita una representación equilibrada en el análisis de conglomerados y en los indicadores compuestos proyectados.

Correlación entre variables. Se realiza un análisis utilizando el coeficiente de correlación de Pearson, sobre las 24 variables sobre la violencia que aún permanecen en el dataset, se busca sobre todos detectar variables fuertemente relacionadas (con umbrales de ± 0.7 o superior), para que facilite decisiones sobre eliminar, fusionar o crear nuevas características que capturen mejor la información relevante.

En el resultado de la matriz de correlación significativas de las 24 variables sobre la violencia que se muestra en la gráfica de la siguiente página, se verifica varias relaciones de alta correlación, aquellas superiores a 0.7, que sugieren la posible redundancia de información entre variables o que podrían representar fenómenos estrechamente relacionados.

Figura 78

Visualización de la matriz las variables sobre la violencia para altas correlaciones



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Las variables relacionadas con la incautación de narcóticos (cls_basuco_incaut_mindef, ha_coca_cultivada_minjust, cls_coca_confis_mindef, cls_base_coca_confis_mindef, y cantidad_mariguana_confiscada) presentan una falta de correlación reportada con otras variables, lo que sugiere un comportamiento aislado de estos fenómenos en la región. Esto podría indicar que los incidentes de incautación y cultivo de narcóticos no tienen una relación directa con otros delitos o eventos de violencia documentados en el dataset.

La matriz sugiere una correlación elevada entre amenazas, delitos sexuales, homicidios, lesiones personales y violencia intrafamiliar, todos reportados por la Fiscalía. Estas variables presentan correlaciones superiores a 0.7 en varios casos, lo que podría indicar que en municipios con altas tasas de uno de estos delitos también se observa una frecuencia considerable de otros delitos interpersonales. Variables como corrupción y secuestro también muestran correlaciones moderadas con estos delitos, lo que podría sugerir una interrelación entre fenómenos de victimización directa y otros tipos de violencia interpersonal.

Las variables de desplazamiento y desaparición forzada tienen correlaciones fuertes entre sí y con casos de reclutamiento ilícito y personas y bienes protegidos por el Derecho Internacional Humanitario (DIH), lo cual es consistente con la dinámica observada en áreas afectadas por conflicto armado. Esto refuerza la idea de que ciertos municipios presentan patrones de violencia relacionados con el conflicto armado.

Variables como estafa y corrupción muestran correlación con delitos sexuales, extorsión, y violencia intrafamiliar, lo que podría estar indicando que estos fenómenos tienden a coexistir en contextos específicos o pueden estar asociados con condiciones socioeconómicas desfavorables en la región.

Reclamación de restitución de tierras muestra independencia de las formas de violencia y su distribución parece responder a dinámicas históricas sin una conexión aparente con las tasas de violencia actuales. Por su parte víctimas del conflicto (vict_por_declarac_uv) si tiene correlaciones significativas con varios tipos de violencia, indicando una posible relación con contextos de alta conflictividad que aún impactan a las comunidades en términos de violencia directa y otros delitos.

La tabla siguiente resume de manera más clara los pares de variables sobre la violencia que presentan un correlación elevada, mayor a 0.7 (positiva o negativa) en el dataset.

Tabla 8*Pares de variables sobre violencia altamente correlacionadas*

Variable 1	Variable 2	Correlación
estafa_casos_fiscalia	hurto_casos_fiscalia	0.9867
hurto_casos_fiscalia	otros_delitos_casos_fiscalia	0.9854
estafa_casos_fiscalia	otros_delitos_casos_fiscalia	0.9836
delitos_sexuales_casos_fiscalia	extorsion_casos_fiscalia	0.9799
delitos_sexuales_casos_fiscalia	estafa_casos_fiscalia	0.9776
estafa_casos_fiscalia	violencia_intrafamiliar_casos_fiscalia	0.9769
hurto_casos_fiscalia	lesiones_personales_casos_fiscalia	0.9762
lesiones_personales_casos_fiscalia	otros_delitos_casos_fiscalia	0.9756
estafa_casos_fiscalia	lesiones_personales_casos_fiscalia	0.9749
violencia_intrafamiliar_casos_fiscalia	lesiones_personales_casos_fiscalia	0.9736
estafa_casos_fiscalia	extorsion_casos_fiscalia	0.9698
delitos_sexuales_casos_fiscalia	hurto_casos_fiscalia	0.9690
violencia_intrafamiliar_casos_fiscalia	otros_delitos_casos_fiscalia	0.9686
extorsion_casos_fiscalia	otros_delitos_casos_fiscalia	0.9671
hurto_casos_fiscalia	corrupcion_casos_fiscalia	0.9671
estafa_casos_fiscalia	corrupcion_casos_fiscalia	0.9669
delitos_sexuales_casos_fiscalia	corrupcion_casos_fiscalia	0.9657
delitos_sexuales_casos_fiscalia	otros_delitos_casos_fiscalia	0.9651
hurto_casos_fiscalia	violencia_intrafamiliar_casos_fiscalia	0.9633
extorsion_casos_fiscalia	hurto_casos_fiscalia	0.9586
corrupcion_casos_fiscalia	otros_delitos_casos_fiscalia	0.9519
extorsion_casos_fiscalia	corrupcion_casos_fiscalia	0.9514
corrupcion_casos_fiscalia	vict_por_declarac_uv	0.9492
extorsion_casos_fiscalia	lesiones_personales_casos_fiscalia	0.9480
delitos_sexuales_casos_fiscalia	lesiones_personales_casos_fiscalia	0.9429
delitos_sexuales_casos_fiscalia	violencia_intrafamiliar_casos_fiscalia	0.9425
homicidios_casos_fiscalia	otros_delitos_casos_fiscalia	0.9384
secuestro_casos_fiscalia	otros_delitos_casos_fiscalia	0.9374
extorsion_casos_fiscalia	violencia_intrafamiliar_casos_fiscalia	0.9373
extorsion_casos_fiscalia	secuestro_casos_fiscalia	0.9332
delitos_sexuales_casos_fiscalia	secuestro_casos_fiscalia	0.9203
estafa_casos_fiscalia	secuestro_casos_fiscalia	0.9187
corrupcion_casos_fiscalia	lesiones_personales_casos_fiscalia	0.9186
violencia_intrafamiliar_casos_fiscalia	corrupcion_casos_fiscalia	0.9119
hurto_casos_fiscalia	secuestro_casos_fiscalia	0.9096
amenazas_casos_fiscalia	homicidios_casos_fiscalia	0.9074
homicidios_casos_fiscalia	lesiones_personales_casos_fiscalia	0.9012
hurto_casos_fiscalia	homicidios_casos_fiscalia	0.9009
homicidios_casos_fiscalia	secuestro_casos_fiscalia	0.9009
delitos_sexuales_casos_fiscalia	vict_por_declarac_uv	0.9003

violencia_intrafamiliar_casos_fiscalia	secuestro_casos_fiscalia	0.8977
violencia_intrafamiliar_casos_fiscalia	homicidios_casos_fiscalia	0.8936
extorsion_casos_fiscalia	homicidios_casos_fiscalia	0.8919
corrupcion_casos_fiscalia	secuestro_casos_fiscalia	0.8901
delitos_sexuales_casos_fiscalia	homicidios_casos_fiscalia	0.8892
estafa_casos_fiscalia	homicidios_casos_fiscalia	0.8891
lesiones_personales_casos_fiscalia	secuestro_casos_fiscalia	0.8888
extorsion_casos_fiscalia	vict_por_declarac_uv	0.8851
estafa_casos_fiscalia	vict_por_declarac_uv	0.8844
hurto_casos_fiscalia	vict_por_declarac_uv	0.8769
otros_delitos_casos_fiscalia	vict_por_declarac_uv	0.8585
corrupcion_casos_fiscalia	homicidios_casos_fiscalia	0.8444
secuestro_casos_fiscalia	vict_por_declarac_uv	0.8442
lesiones_personales_casos_fiscalia	armas_confis_polinal	0.8430
hurto_casos_fiscalia	armas_confis_polinal	0.8392
homicidios_casos_fiscalia	armas_confis_polinal	0.8391
amenazas_casos_fiscalia	otros_delitos_casos_fiscalia	0.8370
otros_delitos_casos_fiscalia	armas_confis_polinal	0.8351
personas_bienes_dih_casos_fiscalia	reclutamiento_illicito_casos_fiscalia	0.8272
amenazas_casos_fiscalia	secuestro_casos_fiscalia	0.8121
amenazas_casos_fiscalia	violencia_intrafamiliar_casos_fiscalia	0.8116
violencia_intrafamiliar_casos_fiscalia	armas_confis_polinal	0.8088
amenazas_casos_fiscalia	lesiones_personales_casos_fiscalia	0.8035
amenazas_casos_fiscalia	estafa_casos_fiscalia	0.8027
estafa_casos_fiscalia	armas_confis_polinal	0.8020
amenazas_casos_fiscalia	hurto_casos_fiscalia	0.8005
desaparicion_forzada_casos_fiscalia	desplazamiento_casos_fiscalia	0.7993
violencia_intrafamiliar_casos_fiscalia	vict_por_declarac_uv	0.7989
lesiones_personales_casos_fiscalia	vict_por_declarac_uv	0.7923
amenazas_casos_fiscalia	extorsion_casos_fiscalia	0.7861
amenazas_casos_fiscalia	delitos_sexuales_casos_fiscalia	0.7844
desaparicion_forzada_casos_fiscalia	reclutamiento_illicito_casos_fiscalia	0.7842
amenazas_casos_fiscalia	armas_confis_polinal	0.7717
corrupcion_casos_fiscalia	armas_confis_polinal	0.7688
delitos_sexuales_casos_fiscalia	armas_confis_polinal	0.7612
homicidios_casos_fiscalia	vict_por_declarac_uv	0.7556
desaparicion_forzada_casos_fiscalia	homicidios_casos_fiscalia	0.7526
extorsion_casos_fiscalia	armas_confis_polinal	0.7467
desaparicion_forzada_casos_fiscalia	secuestro_casos_fiscalia	0.7359
desplazamiento_casos_fiscalia	reclutamiento_illicito_casos_fiscalia	0.7316
desaparicion_forzada_casos_fiscalia	personas_bienes_dih_casos_fiscalia	0.7193
amenazas_casos_fiscalia	corrupcion_casos_fiscalia	0.7171
secuestro_casos_fiscalia	armas_confis_polinal	0.7042

Fuente: Pantallazo obtenido desde el entorno de Google Colab

La elevada correlación entre las variables de violencia, observada en la tabla 8 y la matriz de la figura 78, refleja fenómenos interrelacionados que forman parte de una misma dinámica social y criminal. No obstante, simplemente eliminar variables altamente correlacionadas para evitar colinealidad podría simplificar el análisis en exceso y limitar la comprensión del fenómeno. En su lugar, una estrategia más eficaz sería fusionar variables correlacionadas en indicadores compuestos, especialmente para delitos con patrones comunes, lo cual facilita el análisis y reduce redundancias sin perder profundidad.

La fusión de variables de violencia en indicadores compuestos permite optimizar la estructura de los datos sin perder riqueza informativa. Variables con alta correlación capturan dimensiones similares del fenómeno, pero al eliminar una de ellas se corre el riesgo de omitir detalles relevantes para comprender el contexto violento en los municipios del Cauca. La creación de estos indicadores simplifica el análisis sin reducir la complejidad de la violencia en sus diversas manifestaciones, preservando el impacto diferenciado de cada delito en la sociedad y sus posibles vínculos con dinámicas locales.

Para construir los indicadores se pueden emplear sumas, promedios, o índices ponderados según la relevancia de cada variable. A continuación, se sugieren algunos posibles indicadores para concretar la fusión de estas variables:

- **Indice de narcotráfico:** se incluye las variables kls_basuco_incaut_mindef, kls_coca_confis_mindef, kls_base_coca_confis_mindef, cantidad_mariguana_confiscada, que tienen la misma unidad de medida, kilogramos de sustancias incautadas y su valor corresponde a la suma de las cantidades incautadas para cada año del periodo 2019-2023. Como método de fusión se emplea la suma total de las cantidades incautadas de cada sustancia, representando así un índice de incautación en kilogramos. Este índice reflejará el volumen de sustancias ilícitas incautadas para el periodo de análisis.
- **Indice de coacción y extorsión:** se incluye las variables: amenazas_casos_fiscalia, extorsion_casos_fiscalia, secuestro_casos_fiscalia, que tiene la misma unidad de medida, número de casos y cuyo valor corresponde a la suma acumulada para el periodo 2019-2023. Como método de fusión se emplea la suma del total de casos reportados en cada tipo de delito para reflejar el grado de coacción y extorsión. La variable compuesta representará la incidencia de estos delitos en términos absolutos.
- **Indice de delitos económicos:** se incluye las variables estafa_casos_fiscalia, hurto_casos_fiscalia, corrupcion_casos_fiscalia que tienen como unidad de medida el

número de casos acumulados. El método de fusión sugerido en la suma de casos acumulados para obtener un índice de delitos económicos.

- **Indice de violencia interpersonal:** se incluye las variables violencia_intrafamiliar_casos_fiscalia, homicidios_casos_fiscalia, lesiones_personales_casos_fiscalia, delitos_sexuales_casos_fiscalia cuya unidad de medida es el número de casos acumulados. Se utiliza el método de sumar el total de casos de violencia interpersonal, reflejando la incidencia de estos delitos sobre el tejido social.

Existe un conjunto de variables derivadas de la afectación por el conflicto armado: desaparicion_forzada_casos_fiscalia, desplazamiento_casos_fiscalia, personas_bienes_dih_casos_fiscalia, reclutamiento_illicito_casos_fiscalia, total_reclam_rest_tierr_minagr y vict_por_declarac_uv. Dado que las variable de víctimas por declaración de la Unidad de Víctimas incluye en su contabilidad todos los casos de relacionados con el conflicto armado se procede a eliminar del conjunto de datos las variables: desaparicion_forzada_casos_fiscalia, desplazamiento_casos_fiscalia, personas_bienes_dih_casos_fiscalia, reclutamiento_illicito_casos_fiscalia y se mantienen por separado los campos total_reclam_rest_tierr_minagr y vict_por_declarac_uv

No se incluyen en los indicadores compuestos sino que se mantiene por separado las variables otros_delitos_casos_fiscalia, armas_confis_polinal, ha_coca_cultivada_minjust y vict_fuerza_pub_mindef.

Una vez hechas las fusiones y eliminaciones anteriormente documentadas las variables en el conjunto de datos sobre hechos de violencia quedan reducidas a 10 columnas así:

- 0: ha_coca_cultivada_minjust
- 1: otros_delitos_casos_fiscalia
- 2: vict_fuerza_pub_mindef
- 3: armas_confis_polinal
- 4: total_reclam_rest_tierr_minagr
- 5: vict_por_declarac_uv
- 6: indice_narcotrafico
- 7: indice_coaccion_extorsion
- 8: indice_delitos_economicos
- 9: indice_violencia_interpersonal

Factor de Inflación de la Varianza (VIF). Para profundizar en el análisis de la multicolinealidad presente en el conjunto de datos, se procede a calcular el Factor de Inflación de la Varianza (VIF).

En la Figura siguiente se presenta el código Python utilizado para calcular el VIF de las 10 variables sobre la violencia restantes y los resultados obtenidos, ordenados de mayor a menor VIF.

Figura 79

Cálculo del VIF de las variables sobre la violencia presentes en el dataset

```

[67] # Aplicar VIF
      # Extraer solo las columnas socioeconómicas
      X = df_var_violencia[var_violencia]

      # Añadir una constante al modelo
      X = sm.add_constant(X)

      # Calcular el VIF para cada variable
      vif_data = pd.DataFrame()
      vif_data["Variable"] = X.columns
      vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

      # Eliminar la fila correspondiente a la constante
      vif_data = vif_data[vif_data["Variable"] != "const"]

      # Ordenar el VIF de mayor a menor
      vif_data = vif_data.sort_values(by="VIF", ascending=False)

      # Mostrar el VIF para cada variable
      print(vif_data)

```

	Variable	VIF
9	indice_delitos_economicos	174.226482
2	otros_delitos_casos_fiscalia	113.413677
10	indice_violencia_interpersonal	103.383956
8	indice_coaccion_extorsion	30.544057
6	vict_por_declarac_uv	12.310729
4	armas_confis_polinal	4.686906
1	ha_coca_cultivada_minjust	2.127569
7	indice_narcotrafico	2.000379
5	total_reclam_rest_tierr_minagr	1.886667
3	vict_fuerza_pub_mindef	1.476728

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Los valores de VIF permiten detectar problemas de multicolinealidad, es decir, situaciones en las que una variable puede ser explicada en gran medida por otras variables en el dataset. En general, un VIF por encima de 10 indica una multicolinealidad fuerte, mientras que valores extremadamente altos (como los que superan 100) sugieren una redundancia significativa. Es claro que indice_delitos_economicos, otros_delitos_casos_fiscalia, y indice_violencia_interpersonal, tiene un VIF extremadamente alto (más de 100), sugiriendo una multicolinealidad muy fuerte entre ellos, lo que podría llevar a problemas en los modelos predictivos y afectar la interpretación.

Si bien es cierto que el proyecto no pretende implementar un modelo predictivo sobre las variables de violencia, sino que el objetivo es realizar clústering, y además se tiene previsto aplicar PCA, donde la alta multicolinealidad reflejada en los valores VIF entre variables de violencia es menos problemática. De todas estos altos valores VIF están identificando una redundancia informativa significativa entre ciertos indicadores específicos, en particular, las

variables “otros_delitos_casos_fiscalia”, “indice_coaccion_extorsion”, “indice_delitos_economicos” e “indice_violencia_interpersonal” lo que sugiere que comparten gran parte de su variabilidad y capturan aspectos interrelacionados de la criminalidad y violencia. Para optimizar el modelo de datos, se propone la creación de un índice de criminalidad general que fusione estas variables en una métrica unificada, reflejando de manera robusta las diferentes dimensiones de la criminalidad sin duplicación informativa.

Este índice de criminalidad general se construye a partir de la suma de los valores de los indicadores previamente mencionados, los cuales fueron previamente constituidos como agregados de grupos de delitos afines. Estos valores representan el conteo de casos documentados entre 2019 y 2023. La integración de estos indicadores en una única variable compuesta permite preservar la riqueza informativa sobre los diferentes tipos de criminalidad, optimizando la eficiencia analítica del conjunto de datos y mejorando su rendimiento cuando se aplique el análisis de componentes principales (PCA). Este índice fusionado conserva la profundidad de información al captar aspectos de coacción, extorsión, delitos económicos y violencia interpersonal, facilitando el análisis de patrones complejos de criminalidad en el Cauca en un solo indicador representativo.

Una vez implementa la fusión de variables se vuelve a calcular el VIF para las variables que quedan y se obtienen valores ostensiblemente más bajos.

Figura 80

Nuevos valore del VIF de las variables sobre la violencia después de fusión de variables

```
# Aplicar VIF
# Extraer solo las columnas socioeconómicas
X = df_var_violencia[var_violencia]

# Añadir una constante al modelo
X = sm.add_constant(X)

# Calcular el VIF para cada variable
vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]

# Eliminar la fila correspondiente a la constante
vif_data = vif_data[vif_data["Variable"] != "const"]

# Ordenar el VIF de mayor a menor
vif_data = vif_data.sort_values(by="VIF", ascending=False)

# Mostrar el VIF para cada variable
print(vif_data)
```

	Variable	VIF
7	indice_criminalidad_general	12.680041
5	vict_por_declarac_uv	5.624048
3	armas_confis_polinal	4.219029
4	total_reclam_rest_tierr_minagr	1.825622
1	ha_coca_cultivada_minjust	1.671117
6	indice_narcotrafico	1.316026
2	vict_fuerza_pub_mindef	1.190009

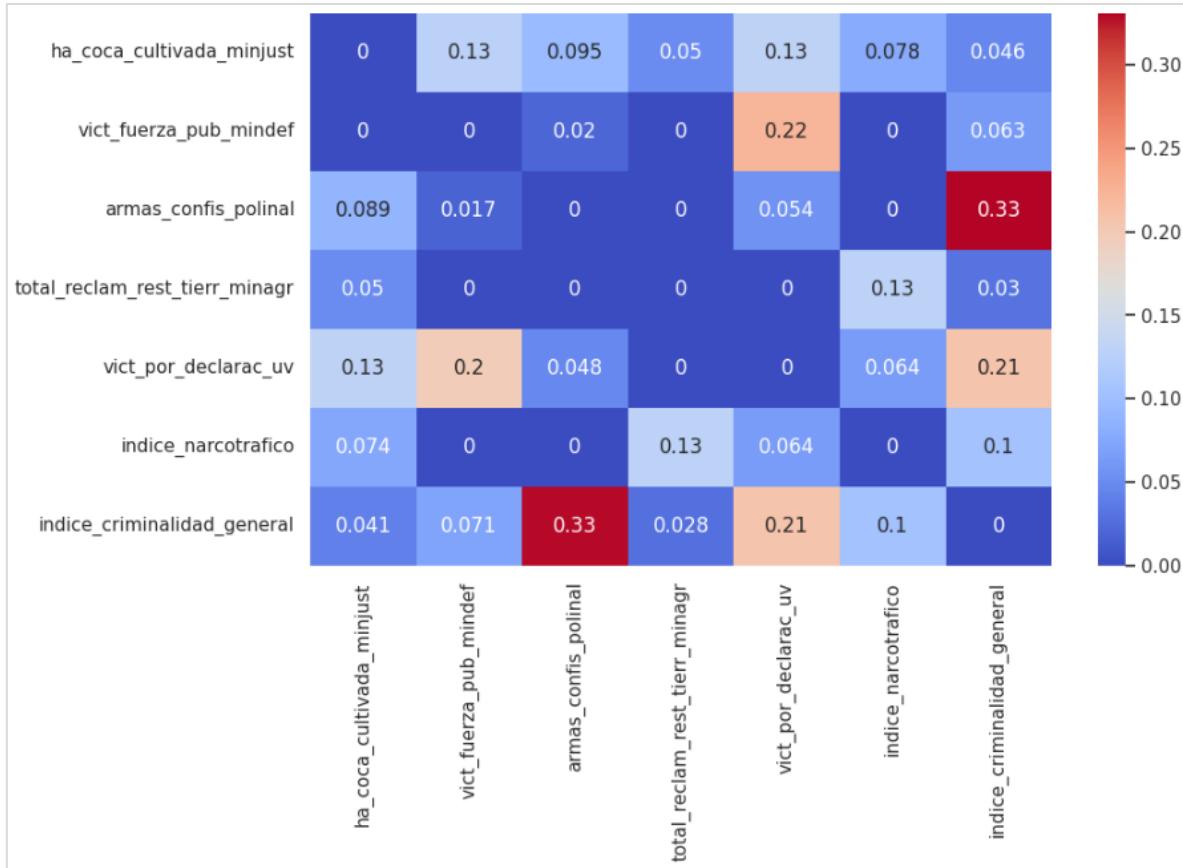
Fuente: Pantallazo obtenido desde el entorno de Google Colab

Análisis de Información Mutua. A continuación, se aborda el análisis de información mutua sobre las variables de violencia para evaluar la dependencia estadística entre ellas.

Los resultados de aplicar el cálculo de la información mutua sobre todas las combinaciones de las 7 variables se presentan en la matriz siguiente.

Figura 81

Matriz de análisis de información mutua de las variables sobre la violencia



El análisis de información mutua realizado sobre las variables de violencia revela que, en general, los valores de dependencia informativa son bajos, lo cual indica que cada variable proporciona información exclusiva, sin redundancia significativa en el contexto del dataset. Este patrón de baja dependencia es especialmente evidente en la mayoría de las interacciones entre variables, mostrando valores cercanos a cero.

Sin embargo, se observaron valores de dependencia moderados en dos relaciones específicas: entre *armas_confis_polinal* e *indice_criminalidad_general* (0.33), y entre *vict_por_declarac_uv* tanto con *vict_fuerza_pub_mindef* (0.22) como con

indice_criminalidad_general (0.21). Estas interacciones moderadas sugieren una leve superposición de información entre la criminalidad y aspectos específicos de la violencia en el conflicto armado, aunque esta redundancia no es suficientemente fuerte como para justificar la eliminación de una de las variables.

Dado que el dataset será procesado mediante K-means para identificar patrones de agrupamiento y que se ha previsto una etapa de análisis de componentes principales (PCA) para optimizar la estructura dimensional, no se recomienda realizar ninguna modificación en este momento. El PCA se encargará de reducir la redundancia preservando la mayor varianza posible, permitiendo que el conjunto de variables conserve su riqueza informativa sin necesidad de ajustes adicionales.

Con base en el análisis de información mutua, se concluye que el conjunto de datos ha alcanzado un nivel adecuado de depuración en términos de selección de variables, mostrando baja redundancia y preservando la riqueza informativa necesaria para un análisis robusto de los fenómenos de violencia en la región. Las variables restantes son complementarias y esenciales para representar diferentes aspectos de este fenómeno, lo que justifica la decisión de no eliminar ninguna en esta fase.

Dado este nivel de optimización, no se considera necesario aplicar métodos adicionales de selección de características, como la evaluación de importancia por permutación o la medición de importancia de variables mediante algoritmos como Random Forest, dado que las variables actuales han sido filtradas y seleccionadas cuidadosamente, conservando solo aquellas que aportan información valiosa y diferencial al modelo.

Normalizar las variables de violencia por la población de los municipios. Dado que las variables de violencia en el dataset son valores de conteo agregados a nivel de municipio, surge un tema importante: municipios con mayor población pueden mostrar recuentos de eventos violentos más altos simplemente porque tienen más habitantes y, en consecuencia, una mayor probabilidad de ocurrencia de dichos eventos. Este efecto podría crear sesgos, haciendo que los municipios más poblados se detecten como anómalos en el análisis sin necesariamente reflejar un mayor nivel de violencia proporcional a su tamaño.

Sino se normalizan las variables de violencia por la población, los municipios grandes tienden a aparecer más violentos en términos absolutos, aunque su nivel de violencia relativa pueda ser similar al de municipios más pequeños. Para abordar este posible sesgo, sería prudente realizar una normalización de las variables de violencia para reducir la influencia del tamaño poblacional. Por lo tanto se procede a dividir cada variable de conteo por la población del municipio y convertirlas en tasas por cada 10,000 habitantes, para que se ajusten los valores

de violencia a un estándar poblacional, permitiendo que se comparan de manera más equitativa entre municipios de distintos tamaños. Así eliminando el efecto de la población, los algoritmos de Machine Learning estarán menos sesgados y podrán centrarse en patrones reales de violencia, independientemente del tamaño del municipio.

Se toma la decisión de normalizar seis de las variables de violencia mediante la tasa por cada 10,000 habitantes para eliminar el sesgo que introduce el tamaño poblacional de los municipios: `armas_confis_polinal`, `indice_criminalidad_general`, `indice_narcotrafico`, `total_reclam_rest_tierr_minagr`, `vict_fuerza_pub_mindef` y `vict_por_declarac_uv`. Adicionalmente, se decide reclasificar la variable `ha_coca_cultivada_minjust`, que mide las hectáreas de hoja de coca reportadas, como una variable socioeconómica en lugar de una variable de violencia. Este cambio responde a que el área de cultivo de coca refleja tanto factores de desarrollo económico como ocupación de tierras, ofreciendo un contexto más amplio que el de un indicador de violencia.

Figura 82

Script para implementar decisiones de normalizar por población

```
# Lista de variables de violencia a normalizar
variables_a_normalizar = [
    'armas_confis_polinal',
    'indice_criminalidad_general',
    'indice_narcotrafico',
    'total_reclam_rest_tierr_minagr',
    'vict_fuerza_pub_mindef',
    'vict_por_declarac_uv'
]

# Unir datasets para acceder a la columna de población en df_var_violencia
df_var_violencia = df_var_violencia.merge(df_var_socieconomicas[['codigo_dane', 'poblacion']],
                                            on='codigo_dane', how='left')

# Normalizar dividiendo por la población y multiplicando por 10,000
for variable in variables_a_normalizar:
    df_var_violencia[f'{variable}_tasa'] = (df_var_violencia[variable] / df_var_violencia['poblacion']) * 10000

# Actualizar lista de variables de violencia normalizadas
var_violencia = [f'{variable}_tasa' for variable in variables_a_normalizar]

# Mover ha_coca_cultivada_minjust a var_socieconomicas
var_socieconomicas.append('ha_coca_cultivada_minjust')

[81] df_var_violencia[var_violencia].head(3)

   armas_confis_polinal_tasa indice_criminalidad_general_tasa indice_narcotrafico_tasa total_reclam_rest_tierr_minagr_tasa
0          19.207315           208.098888            21362.35969             468.0
1          1.542416            163.496144             0.021594              107.0
2          14.736010           255.903389             5.256802              461.0
```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

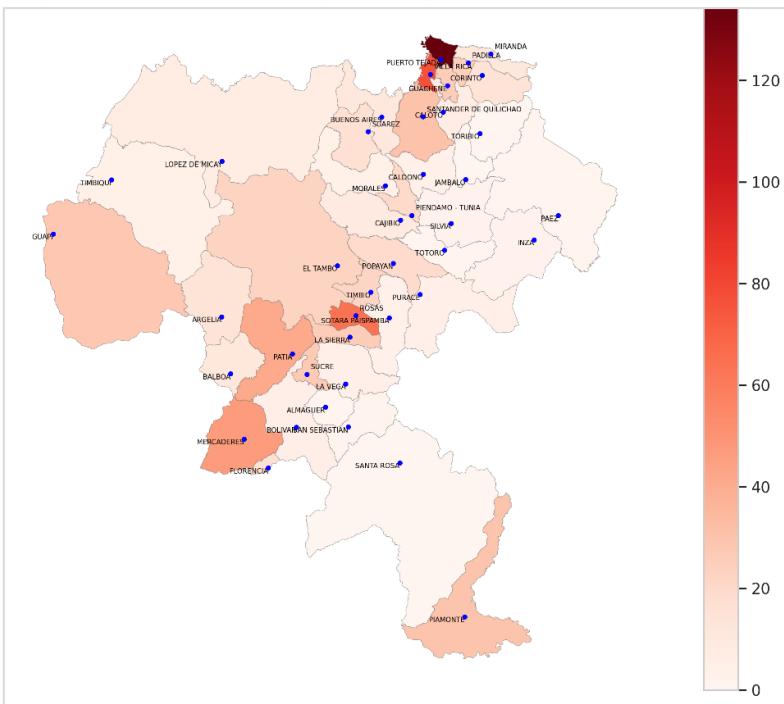
A partir de este punto, los únicos ajustes necesarios antes de someter el dataset a K-means son de tipo técnico y enfocados a la normalización y el escalado de las variables. Estos procesos asegurarán que el modelo interprete correctamente los patrones inherentes a los datos, maximizando la eficacia del agrupamiento en el análisis de violencia.

Análisis geoespacial de las variables sobre la violencia. A continuación se presentan las variables sobre la violencia que permanecen en el dataset tras el exhaustivo proceso de depuración, visualizadas mediante mapas georreferenciados del departamento del Cauca. Esta representación permite observar la distribución espacial y los patrones territoriales de cada variable, proporcionando un enfoque integral que contextualiza los fenómenos socioeconómicos en cada municipio.

En esta serie de mapas, creados con geopandas de Python, cada variable de violencia es visualizada en su propio título, facilitando la autocomprendión de los datos sin necesidad de comentarios adicionales. Se espera que cada título sirva de guía suficiente para que el lector interprete el contenido espacial y territorial de cada variable, maximizando la claridad y comprensión de los fenómenos observados.

Figura 83

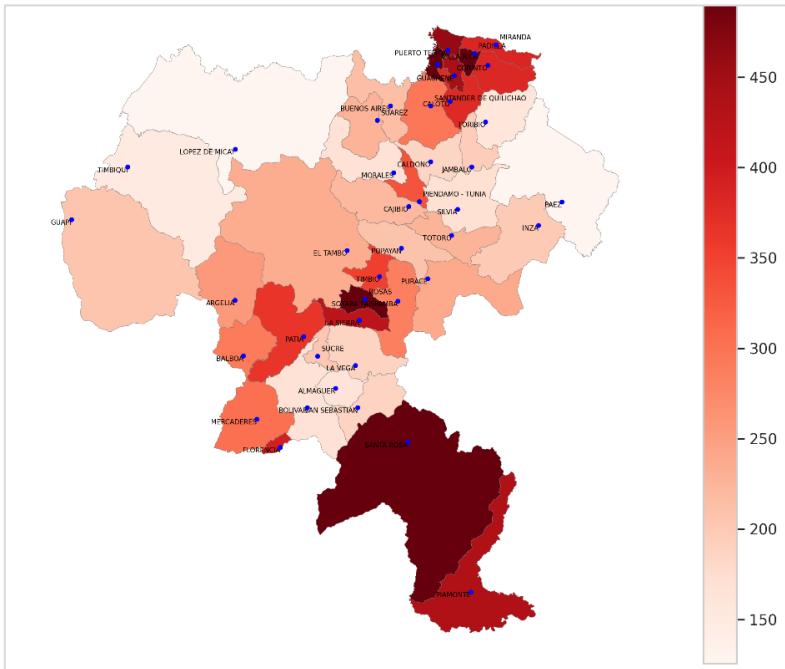
Mapa de tasa de armas confiscadas por la Policía Nacional (2019-2023)



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 84

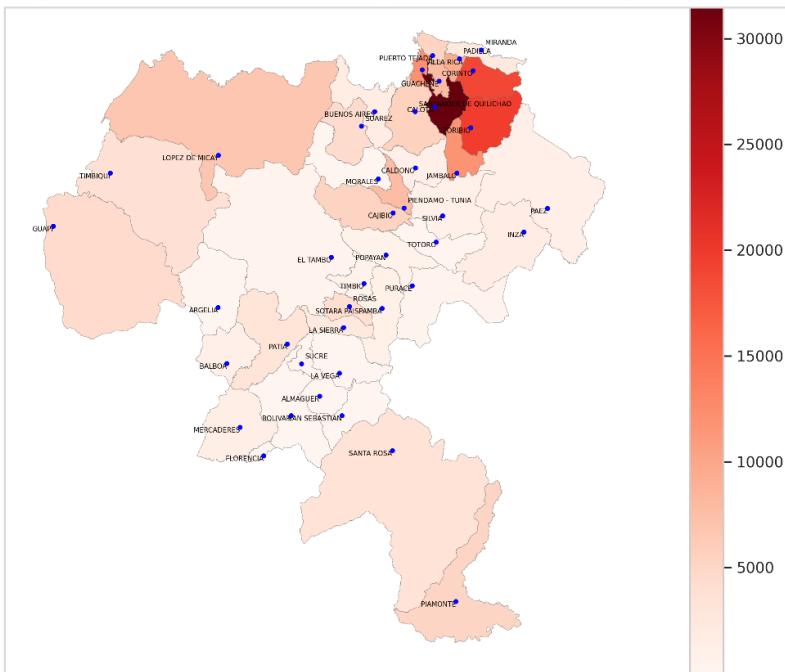
Mapa de la tasa de delitos registrados por la Fiscalía General (2019-2023)



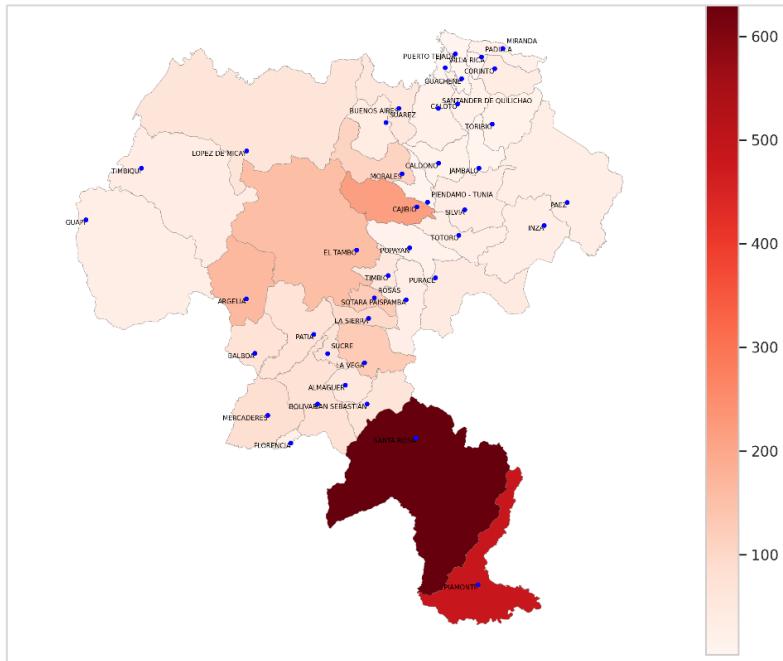
Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 85

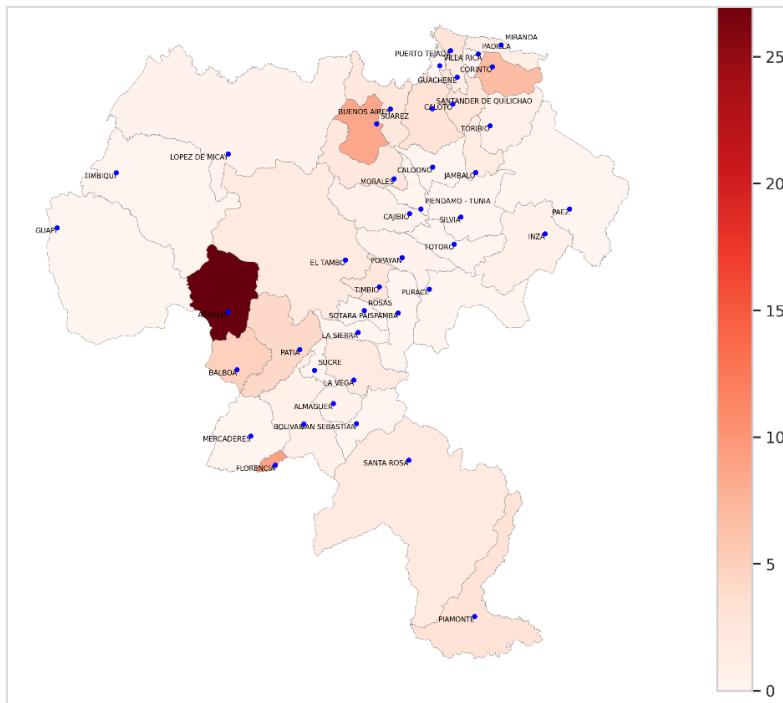
Mapa tasa incautaciones de drogas ilícitas (kls/10M per) - Min-Defensa (2019-2023)



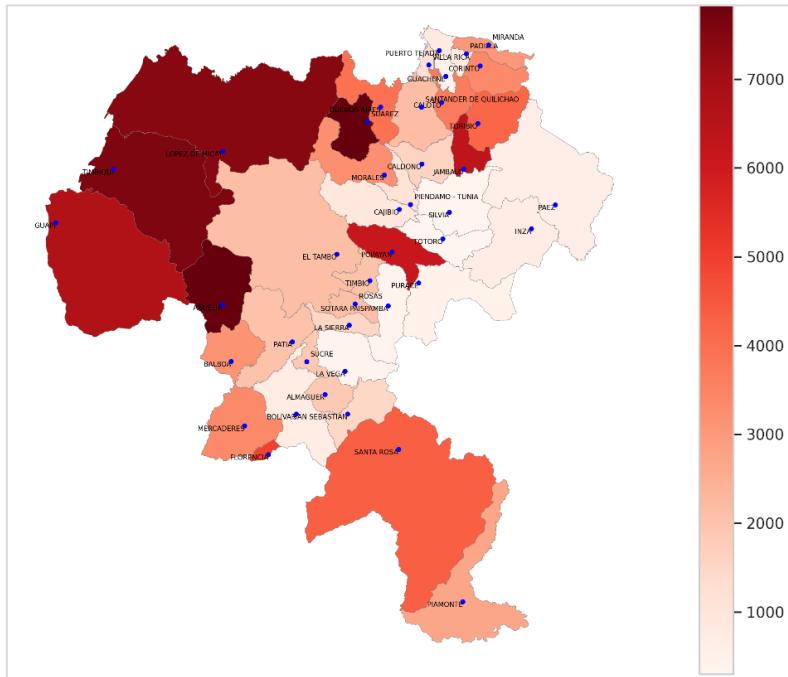
Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 86*Mapa tasa reclamantes de restitución de tierras – reporte Min-Agricultura*

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 87*Mapa de miembros de la fuerza pública víctimas del conflicto*

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 88*Mapa de tasa víctimas por declaración reportadas por la Unidad de Víctimas*

Fuente: Pantallazo obtenido desde el entorno de Google Colab

2.5 Aplicación de los algoritmos Machine Learning en los datasets depurados

Para la etapa final del proyecto, se aplican técnicas de aprendizaje no supervisado sobre el dataset depurado que contiene variables de violencia de los municipios del Departamento del Cauca. Este análisis busca identificar patrones y agrupaciones latentes mediante métodos de clústering, como K-means y PCA, así como detectar posibles anomalías. La estructura de agrupación resultante permitirá clasificar a los municipios en perfiles específicos de violencia, basados en características comunes. Utilizando estos grupos como etiquetas, se procederá a un análisis supervisado sobre el dataset de variables socioeconómicas, aplicando una regresión logística multinomial. Este enfoque permitirá explorar la relación entre condiciones socioeconómicas y los diferentes perfiles de violencia, proporcionando una perspectiva comprensiva sobre la caracterización de los municipios en función de sus niveles de violencia.

Para avanzar en el análisis, se crearon dos datasets finales, cuidadosamente depurados, uno diseñado para capturar las dimensiones socioeconómicas y el otro las de violencia de los municipios del departamento del Cauca. El primero de estos datasets, denominado df_socieconomico_final, contiene las 21 variables socioeconómicas que quedaron tras el proceso de depuración, junto con datos adicionales como el código DANE de cada municipio

y el nombre del municipio para facilitar el análisis y su interpretación, además de la geometría tanto del municipio como de su cabecera, los cuales permitirán la representación geoespacial. Las variables socioeconómicas depuradas se almacenaron en una lista, var_socieconomicas, para facilitar su manipulación en el análisis posterior.

El segundo dataset, df_violencia_final, alberga un total de 6 variables de violencia seleccionadas tras el proceso de depuración y también cuenta con la identificación del código DANE, el nombre del municipio y la geometría territorial, proporcionando así una base sólida para la georreferenciación y el análisis espacial. Las variables de violencia depuradas se incluyeron en una lista, var_violencia.

Escalado de Variables

Dado que las variables en ambos datasets presentan escalas y unidades de medida heterogéneas (como porcentajes, conteos y superficies en hectáreas), es necesario normalizar o escalar los valores antes de someterlos a técnicas de Machine Learning. Este paso garantiza que todas las variables contribuyan de manera equilibrada al proceso de agrupación y facilita la convergencia de los algoritmos al evitar que las variables con valores absolutos mayores dominen la formación de los clústeres. Se emplea **StandardScaler** de scikit-learn que transforma las variables a una distribución con media 0 y desviación estándar 1, garantiza que todas las características tengan una influencia similar en los modelos de aprendizaje automático. La siguiente imagen muestra los resultados del proceso.

Figura 89

Proceso de escalado de datos de variables sobre la violencia

```
# Escalado de variables de violencia
scalerViolencia = StandardScaler()
dfViolencia_scaled = dfViolencia_final.copy()
dfViolencia_scaled[varViolencia] = scalerViolencia.fit_transform(dfViolencia_final[varViolencia])

# Datos antes de escalarlos
dfViolencia_final.head(3)

  armas_confis_polinal_tasa indice_criminalidad_general_tasa indice_narcotrafico_tasa total_reclam_rest_tierr_minagr_tasa vict_fuerza_pub_mindef_tasa
0      19.207315          208.098888           629.315303            13.786846        0.500804
1      1.542416           163.496144            0.021594            55.012853        0.514139
2     14.736010           255.903389            5.256802            165.690256       30.550264

os siguientes: Generar código con dfViolencia_final | Ver gráficos recomendados | New interactive sheet

# Datos escalados
dfViolencia_scaled.head(3)

  armas_confis_polinal_tasa indice_criminalidad_general_tasa indice_narcotrafico_tasa total_reclam_rest_tierr_minagr_tasa vict_fuerza_pub_mindef_tasa
0      -0.011613          -0.658272           -0.590319            -0.531666        -0.365650
1      -0.679668          -1.033793           -0.684163            -0.205530        -0.362945
2      -0.180710          -0.255795           -0.683383            0.670029         5.728880
```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Análisis de anomalías con Isolation Forest

Para profundizar en los patrones de violencia en el departamento del Cauca, se implementa el algoritmo Isolation Forest como un paso de preprocesamiento antes de la aplicación de técnicas de clústerización. Este enfoque permite identificar y comprender las observaciones anómalas que podrían distorsionar el análisis de clústering, lo cual es particularmente relevante en datos de violencia, donde ciertos municipios pueden presentar valores extremos en comparación con el resto de la región.

El objetivo de Isolation Forest en este contexto es identificar aquellos municipios con características de violencia excepcionalmente altas o bajas, diferenciando así patrones que pueden influir significativamente en la interpretación de los clústeres. Identificar y analizar las anomalías en esta etapa permite tomar decisiones informadas sobre si dichas observaciones deberían excluirse o manejarse de forma diferenciada antes de aplicar K-means, para evitar que estos puntos anómalos definan los centros de los clústeres, distorsionando los grupos formados.

Una vez identificadas las anomalías, se puede decidir si remover los municipios que el Isolation Forest identifica como anómalos, para que los clústeres resultantes puedan reflejar patrones de violencia más representativos de la mayoría de los municipios. También es posible realizar la clústerización en ambos escenarios (con y sin anomalías) para comparar resultados y entender cómo las anomalías afectan la estructura de los clústeres, o los municipios identificados como anómalos pueden ser sujetos de un análisis específico y separado, que aporte detalles relevantes sobre los eventos y factores de violencia en estos territorios.

La siguiente figura muestra el proceso de aplicación y los resultados del algoritmo de Isolation Forest a las variables de violencia, el cual, además de señalar los municipios que clasifica como anómalos, también incluye la configuración de la opción `feature_importance` para evaluar la relevancia de cada variable en la identificación de estas anomalías. Esta configuración permite obtener un desglose de cómo cada variable contribuye a la detección de comportamientos atípicos en los datos, lo que facilita la interpretación de los resultados y permite comprender mejor el contexto y las dinámicas de violencia en los municipios estudiados.

Figura 90*Detección de anomalías en datos de violencia con Isolation Forest*

```

from sklearn.ensemble import IsolationForest

# Entrenar Isolation Forest con todas las variables
isolation_forest = IsolationForest(random_state=42, contamination=0.05)
isolation_forest.fit(df_violencia_scaled[varViolencia])

# Obtener los puntajes de anomalía con todas las variables
full_scores = isolation_forest.decision_function(df_violencia_scaled[varViolencia])

# Identificar las anomalías
predictions = isolation_forest.fit_predict(df_violencia_scaled[varViolencia])
anomaly_mask = predictions == -1
anomalies = df_violencia_scaled[anomaly_mask]

# Crear un diccionario para almacenar la importancia de cada variable
feature_importance = {}

# Calcular el cambio en el puntaje de anomalía al eliminar cada variable individualmente
for variable in varViolencia:
    # Crear un dataframe sin la variable actual
    subset = df_violencia_scaled[varViolencia].drop(columns=[variable])

    # Re-entrenar el Isolation Forest con el subconjunto de variables
    isolation_forest.fit(subset)

    # Obtener los puntajes de anomalía con el subconjunto de variables
    subset_scores = isolation_forest.decision_function(subset)

    # Calcular la importancia como la diferencia promedio entre los puntajes completos y los puntajes sin la variable
    importance = np.mean(np.abs(full_scores - subset_scores))
    feature_importance[variable] = importance

# Ordenar la importancia de las variables de mayor a menor
feature_importance = dict(sorted(feature_importance.items(), key=lambda item: item[1], reverse=True))

# Mostrar los municipios considerados como anomalías
print("\nMunicipios detectados como anomalías:")
anomalies[['nombre_municipio']+varViolencia]

```

Municipios detectados como anomalías:

nombre_municipio	armas_confis_polinal_tasa	indice_criminalidad_general_tasa	indice_narcotrafico_tasa	total_reclam_rest_tierr_minagr_tasa
2 ARGELIA	-0.180710	-0.255795	-0.683383	0.670029
27 PUERTO TEJADA	4.943214	1.479397	0.110663	-0.585165
32 SANTA ROSA	-0.737999	1.844848	-0.152546	4.942780

vict_fuerza_pub_mindef_tasa	vict_por_declarac_uv_tasa
5.728880	2.545635
0.084236	-0.872649
-0.107553	0.642662

Fuente: Pantallazo obtenido desde el entorno de Google Colab

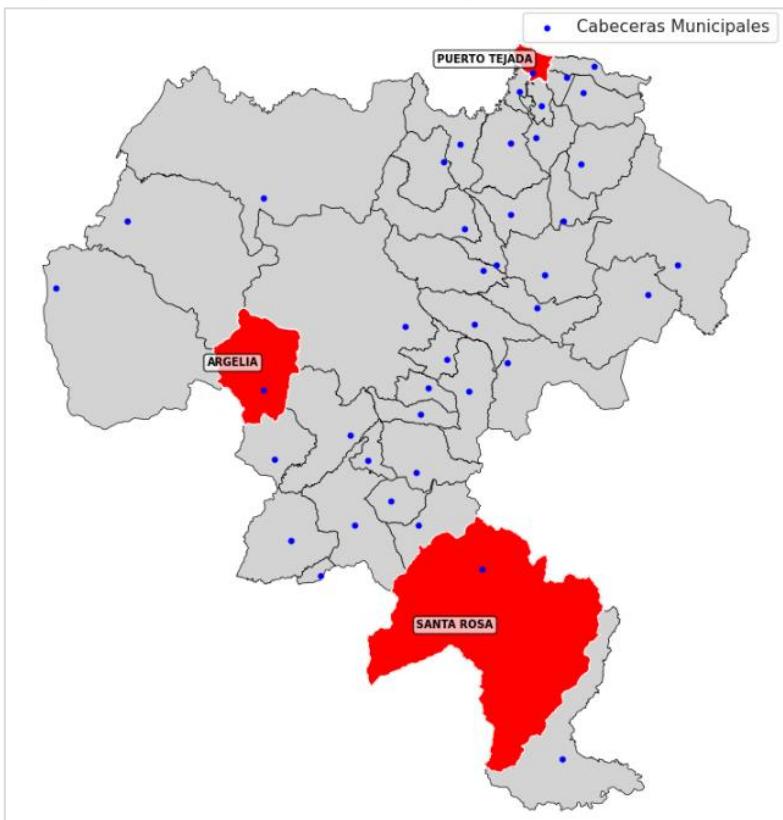
Los resultados del análisis de anomalías mediante Isolation Forest identifican tres municipios en el departamento que presentan patrones atípicos en relación con las tasas de variables de violencia normalizadas. A continuación, se describe la situación particular de cada municipio.

Argelia muestra valores anómalos particularmente altos en las tasas de víctimas de la fuerza pública y de victimización por declaración. Estos elevados puntajes pueden indicar un contexto de alto impacto en la población y la fuerza pública, lo que podría estar asociado a condiciones de inseguridad persistente y enfrentamientos armados en la región. La tasa

elevada de víctimas por conflicto y declaración refleja un entorno donde la población ha sido especialmente afectada por dinámicas de violencia estructural y conflicto armado.

Figura 91

Mapa de municipios anómalos en el Departamento del Cauca



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Puerto Tejada se destaca por su alta tasa de armas confiscadas y un índice general de criminalidad elevado. Estos indicadores sugieren una actividad delictiva intensa y la circulación de armas en el municipio, elementos que podrían relacionarse con fenómenos de violencia urbana y posibles redes de microtráfico. A pesar de una baja incidencia en víctimas de conflicto y declaración, su perfil de riesgo se ve reforzado por la actividad criminal y el acceso a armas, lo cual podría requerir atención en políticas de desarme y control de crimen organizado.

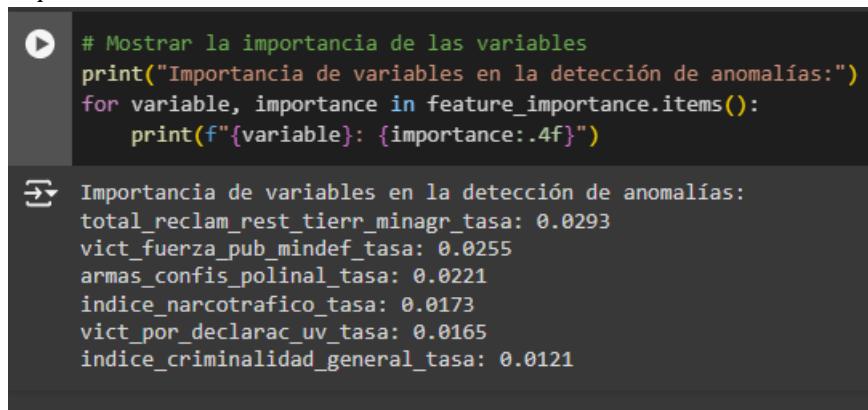
Santa Rosa presenta valores altos en las tasas de reclamantes de restitución de tierras y en el índice de criminalidad general. Estos resultados podrían estar vinculados a conflictos territoriales y una alta demanda de restitución de tierras, lo que sugiere un entorno de desplazamiento y disputa territorial histórica. La situación podría implicar tensiones en el

acceso a tierras y condiciones de inseguridad persistente, asociadas también a altos niveles de actividad delictiva, lo cual demanda estrategias integrales que atiendan tanto los conflictos agrarios como la criminalidad.

La importancia relativa de las características de cada variable en la detección de anomalías se muestra en la siguiente figura.

Figura 92

Importancia de las características en la detección de anomalías en los datos con Isolation Forest



```
# Mostrar la importancia de las variables
print("Importancia de variables en la detección de anomalías:")
for variable, importancia in feature_importance.items():
    print(f"{variable}: {importancia:.4f}")

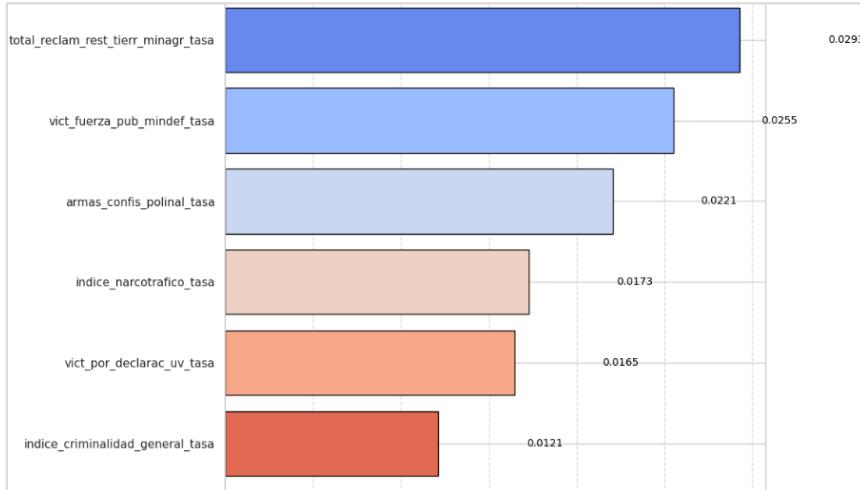

```

Importancia de variables en la detección de anomalías:

- total_reclam_rest_tierr_minagr_tasa: 0.0293
- vict_fuerza_pub_mindef_tasa: 0.0255
- armas_confis_polinal_tasa: 0.0221
- indice_narcotrafico_tasa: 0.0173
- vict_por_declarac_uv_tasa: 0.0165
- indice_criminalidad_general_tasa: 0.0121

El análisis de importancia de características con *Isolation Forest* aplicado a los datos de violencia en el departamento del Cauca revela qué variables tienen mayor peso en la identificación de patrones anómalos entre los municipios. A continuación, se presenta la interpretación de los resultados obtenidos y su posible impacto en las dinámicas de violencia:

- Tasa de reclamantes de restitución de tierras (total_reclam_rest_tierr_minagr_tasa) es la que presenta el mayor valor de importancia (0.0293), esta variable destaca el papel fundamental de los conflictos de tierra en la configuración de patrones anómalos. Las altas tasas de reclamantes pueden reflejar una historia de desplazamientos y despojo, vinculada al impacto que estos conflictos tienen en la seguridad y estabilidad de ciertos municipios.
- Tasa de víctimas de la fuerza pública (vict_fuerza_pub_mindef_tasa), es la segunda variable en importancia (0.0255) refleja la violencia contra la fuerza pública en el contexto de conflicto armado y seguridad pública. Esta incidencia resalta el papel de las dinámicas de confrontación armada y conflictos de orden público en ciertos municipios, aportando una dimensión crucial en la identificación de patrones anómalos relacionados con el riesgo y la seguridad en las zonas afectadas.

Figura 93*Importancia de variables de violencia en la detección de anomalías*

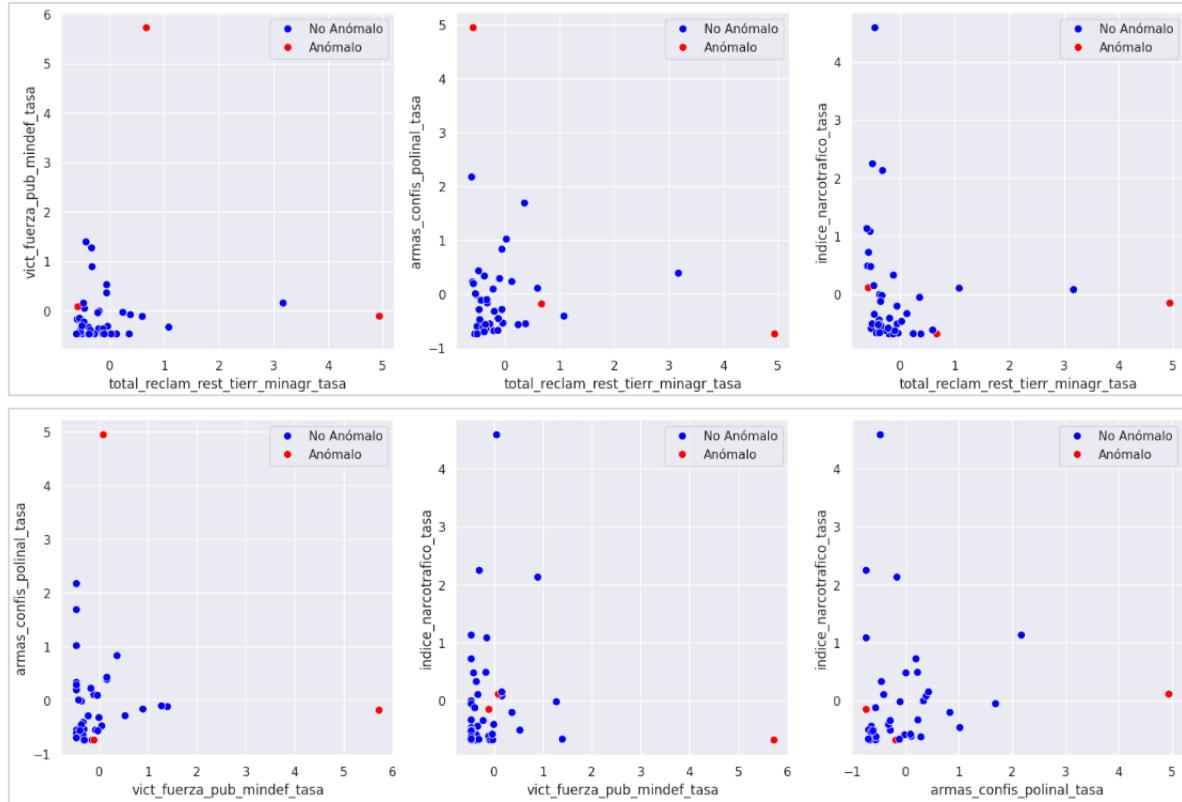
Fuente: Pantallazo obtenido desde el entorno de Google Colab

- Tasa de armas confiscadas (armas_confis_polinal_tasa), es la tercera variable en importancia (0.0221), la cantidad de armas confiscadas es indicativa de la actividad de grupos armados en los municipios. Esta variable sugiere la presencia de actores armados y redes delictivas, vinculadas a un entorno de violencia y criminalidad que afecta la seguridad pública y contribuye a la detección de anomalías.
- Tasa de narcotráfico (indice_narcotrafico_tasa), viene a representar un valor de importancia de 0.0173, cuarto lugar de importancia, la incidencia de narcotráfico destaca el impacto de esta actividad ilegal en ciertos territorios, frecuentemente asociado a violencia territorial y conflicto armado. Este fenómeno resalta cómo el narcotráfico contribuye a patrones anómalos de violencia y tensión en los municipios afectados.
- Tasa de víctimas por declaración (vict_por_declarac_uv_tasa), con una importancia de 0.0165 representa el quinto lugar de importancia. La presencia de víctimas registradas por declaración en la Unidad de Víctimas aporta una perspectiva humanitaria, reflejando el impacto del conflicto en la población civil. Esta variable permite identificar municipios con altas tasas de afectación en su población, debido al conflicto y la violencia prolongada.
- Tasa de criminalidad general (indice_criminalidad_general_tasa), tiene la menor importancia relativa de las variables (0.0121). Esta tasa de criminalidad general agrupa diferentes delitos reportados en los municipios y reportados por la Fiscalía General de la

Nación. Esta variable permite una visión amplia de la actividad delictiva general y refuerza el análisis multifactorial de la violencia.

Figura 94

Distribución de anomalías por pares de variables



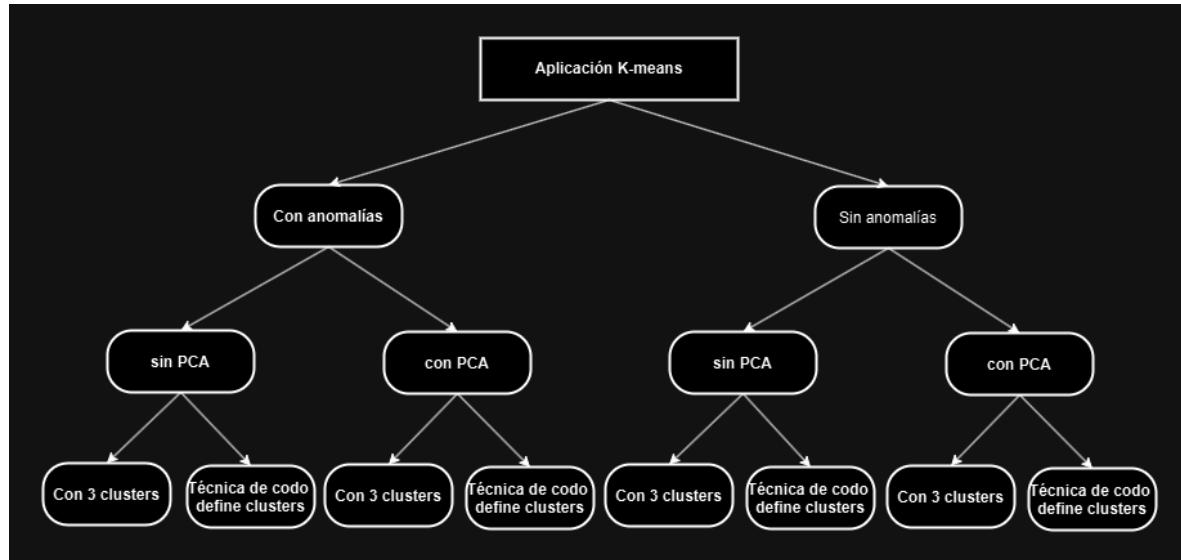
Fuente: Pantallazo obtenido desde el entorno de Google Colab

La distribución de las importancias de las variables muestra una cercanía relativa entre sus valores, lo que sugiere que no existe una única dimensión dominante en la detección de anomalías. Esto indica que las dinámicas de violencia en los municipios anómalos son multifactoriales, resultado de la interacción compleja entre distintas variables.

Enfoque progresivo y exhaustivo para la clusterización de las variables de violencia

Para analizar las variables de violencia en los municipios del Cauca, se emplea un enfoque de clustering K-means de forma progresiva. Inicialmente, se realiza un análisis básico sin considerar anomalías ni reducción de dimensionalidad, utilizando 3 clusters. Posteriormente, se refina el análisis incorporando técnicas como PCA y la selección del número óptimo de clusters mediante el método del codo. Finalmente, se evaluará el impacto de excluir los municipios identificados como anómalos por Isolation Forest. Estas son solo tres de las ocho posibles combinaciones de configuración que podrían aplicarse, tal como se muestra en la figura siguiente:

Figura 95
Explorando diversas combinaciones de configuración de K-means



Fuente: Construcción propia

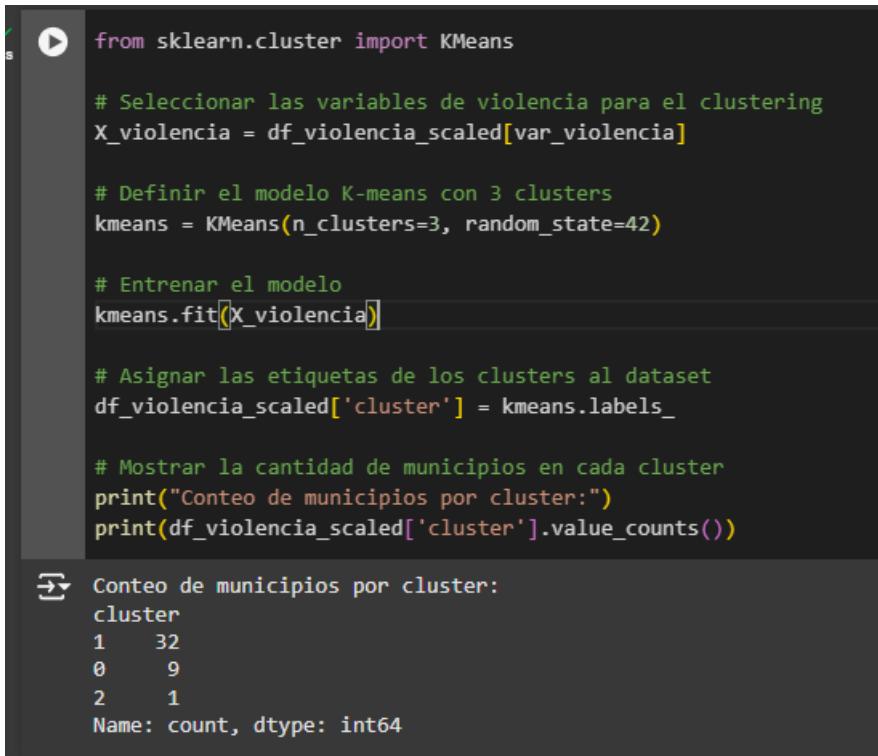
K-means sin reducción de dimensionalidad, con anomalías y con 3 de clústeres. Para iniciar con el estudio de violencia en el Cauca mediante K-means, primero se implementa una configuración básica de tres clústeres sin anomalías ni reducción dimensional. Para facilitar un análisis detallado y progresivo, cada paso del proceso se organiza en etapas o parcelaciones de código, lo cual permite verificar los resultados en cada fase de la agrupación. Estas parcelaciones incluyen la asignación de clústeres a cada municipio, la identificación e impresión de los municipios que componen cada clúster, el cálculo de centroides, el análisis de estadística descriptiva de las variables al interior de cada clúster, la representaciones gráficas como la tesselación de Voronoi y la distribución de variables clave mediante boxplots.

Adicionalmente se incorporan visualizaciones geoespaciales que representan los clústeres sobre el mapa del Cauca, permitiendo apreciar la distribución geográfica de los perfiles de violencia, y la evaluación del modelo, donde se emplean métricas de cohesión y separación como el coeficiente de silueta y la inercia, cuyo propósito es garantizar la efectividad y optimización del número de clústeres de acuerdo con la estructura interna de los datos.

Para iniciar, el siguiente código aplica el algoritmo de K-means para segmentar los municipios del Cauca en función de sus características de violencia. El código selecciona el subconjunto de variables de violencia ya escaladas para el clustering, crea un modelo K-means con 3 clústeres y una semilla aleatoria para garantizar la reproducibilidad de los resultados. Entrena o ajusta el modelo calculando los centros de los clústeres y asignando a cada municipio una etiqueta de clúster; las etiquetas resultantes se agregan al DataFrame en una nueva columna llamada 'clúster' y finalmente imprime el número de municipios en cada clúster.

Figura 96

Segmentación inicial de los municipios. Sin anomalías, sin PCA y con 3 clústeres



```

from sklearn.cluster import KMeans

# Seleccionar las variables de violencia para el clustering
X_violencia = df_violencia_scaled[var_violencia]

# Definir el modelo K-means con 3 clusters
kmeans = KMeans(n_clusters=3, random_state=42)

# Entrenar el modelo
kmeans.fit(X_violencia)

# Asignar las etiquetas de los clusters al dataset
df_violencia_scaled['cluster'] = kmeans.labels_

# Mostrar la cantidad de municipios en cada cluster
print("Conteo de municipios por cluster:")
print(df_violencia_scaled['cluster'].value_counts())

```

Conteo de municipios por cluster:

cluster	count
1	32
0	9
2	1

Name: count, dtype: int64

Fuente: Pantallazo obtenido desde el entorno de Google Colab

El conteo de municipios por clúster muestra la siguiente distribución:

- Clúster 1: 32 municipios
- Clúster 0: 9 municipios
- Clúster 2: 1 municipio

Esta distribución revela una fuerte concentración de municipios en el clúster 1, mientras que el clúster 2 contiene solo un municipio (Argelia). Este resultado sugiere que la mayoría de los municipios del Cauca comparten características de violencia similares, mientras que un grupo reducido presenta perfiles marcadamente diferentes. Esto puede indicar perfiles de violencia atípicos o características singulares en los municipios de los clústeres minoritarios. Los siguientes pasos en el análisis profundizarán en las características de cada clúster para validar y comprender mejor estas agrupaciones.

Figura 97

Listado de municipios por clúster

```
# Agrupar por 'cluster' y crear una lista de municipios en cada cluster
municipios_por_cluster = df_violencia_scaled.groupby('cluster')[['nombre_municipio']].agg(list)

# Calcular el conteo de municipios por cluster
conteo_municipios = df_violencia_scaled['cluster'].value_counts().sort_index()

# Crear un DataFrame con los resultados
cluster_summary = pd.DataFrame({
    'municipios_en_cluster': municipios_por_cluster,
    'cantidad_municipios': conteo_municipios
})
cluster_summary
```

cluster	municipios_en_cluster	cantidad_municipios
0	POPAYAN,CALOTO,CORINTO,GUAPI,JAMBALO,LOPEZ DE MICAY,SUAREZ,TIMBICUI,TORIBIO	9
1	ALMAGUER,BALBOA,BOLIVAR,BUENOS AIRES,CAJIBIO,CALDONO,EL TAMBO,FLORENCIA,GUACHENE,INZA,LA SIERRA,LA VEGA,MERCADERES,MIRANDA,MORALES,PADILLA,PAEZ,PATIA,PIAMO NTE,PIENDAMO - TUNIA,PUERTO TEJADA,PURACE,ROSAS,SAN SEBASTIAN,SANTANDER DE QUILICHOA,SANTA ROSA,SILVIA,SOTARA PAISPAMBA,SUCRE,TIMBIO,TOTORO,VILLA RICA	32
2	ARGELIA	1

Fuente: Pantallazo obtenido desde el entorno de Google Colab

A continuación, se procede a calcular los centroides de cada clúster, los cuales representan los puntos medios de los municipios asignados a cada grupo. Los centroides ofrecen una visión promedio del perfil de violencia para cada clúster, ayudando a identificar las características distintivas de los municipios agrupados en cada uno.

Figura 98

Ubicación de los centroides en el espacio de las variables

```
[106] # Convertir los centroides en un DataFrame con los nombres de las variables
centroids_df = pd.DataFrame(kmeans.cluster_centers_, columns=varViolencia)
print("Centroides de los clusters:")
centroids_df
```

	armas_config_polinal_tasa	indice_criminalidad_general_tasa	indice_narcotrafico_tasa	total_reclam_rest_tierr_minagr_tasa	vict_fuerza_pub_mindef_tasa	vict_por_declarac_uv_tasa
0	-0.321563	-0.508071	1.070229	-0.397539	0.019675	1.318455
1	0.096087	0.150889	-0.279646	0.090870	-0.184561	-0.450367
2	-0.180710	-0.255795	-0.683383	0.670029	5.728880	2.545635

Fuente: Pantallazo obtenido desde el entorno de Google Colab

En la matriz de centroides anterior, cada fila representa un vector que define las coordenadas del centroide de un clúster específico, donde cada valor son las coordenadas exactas de un punto en el espacio multidimensional (seis dimensiones en este caso) y, cada columna representa una de las características o variables del dataset de violencia en el Cauca. Cada coordenada indica el valor promedio de cada variable en particular para todos los puntos asignados a ese clúster.

Los resultados obtenidos para los centroides muestran variaciones significativas entre los clústeres en función de las seis variables de violencia consideradas:

- Clúster 0: se caracteriza por un valor positivo en el indice_narcotrafico_tasa (1.07) y una alta vict_por_declarac_uv_tasa (1.32), lo que indica que los municipios en este grupo tienden a presentar una incidencia superior en temas de narcotráfico y alta en victimizaciones reportadas por la Unidad de Víctimas. Sin embargo, el resto de sus tasas están cerca de cero o en valores negativos, lo cual podría representar un perfil de violencia moderada pero con problemas específicos de narcotráfico y victimización.
- Clúster 1 muestra valores en su mayoría cercanos a cero, o negativos, con un valor notablemente bajo en vict_por_declarac_uv_tasa (-0.45). Esto sugiere que los municipios en este grupo tienen un perfil de violencia relativamente bajo en comparación con otros clústeres, sin picos significativos en ninguna de las variables. Este perfil parece representar la norma o el nivel de violencia más estable y moderado entre los municipios analizados.
- Clúster 2 destaca por presentar valores excepcionalmente altos en vict_fuerza_pub_mindef_tasa (5.73) y vict_por_declarac_uv_tasa (2.54). Esto indica una concentración de violencia relacionada con la fuerza pública y reportes de victimización, configurando un perfil atípico y extremo de violencia. Este clúster contiene un solo municipio, que muestra características únicas en términos de conflicto y victimización.

El siguiente código calcula estadísticas descriptivas (media, mediana y desviación estándar) para cada variable de violencia en los diferentes clústeres. Estas estadísticas permiten caracterizar y comparar la variabilidad y la tendencia central de cada clúster en relación con las variables de estudio.

Figura 99*Estadísticas descriptivas por clúster*

# Estadísticas descriptivas de cada cluster																		
cluster_mean_std = df_violencia_scaled.groupby('cluster')[varViolencia].agg(['mean', 'median', 'std'])																		
print("Media, mediana y desviación estándar de variables de violencia por cluster:")																		
cluster_mean_std																		
Media, mediana y desviación estándar de variables de violencia por cluster:																		
armas_confis_polinal_tasa indice_criminalidad_general_tasa indice_narcotrafico_tasa total_reclam_rest_tierr_minagr_tasa vict_fuerza_pub_mindef_tasa vict_por_declarac_uv_tasa																		
mean median std																		
cluster																		
0	-0.321563	-0.450758	0.362030	-0.508071	-0.672092	0.786174	1.070229	0.327759	1.657107	-0.397539	-0.378001	0.134693	0.019675	-0.303510	0.630393	1.318455	1.491716	0.740902
1	0.086087	-0.283536	1.131553	0.150889	0.018599	1.046318	-0.279646	-0.487446	0.466766	0.090870	-0.212143	1.131932	-0.184561	-0.333035	0.397868	-0.450367	-0.546227	0.538600
2	-0.180710	-0.180710	NaN	-0.255795	-0.255795	NaN	-0.683383	-0.683383	NaN	0.670029	0.670029	NaN	5.728880	5.728880	NaN	2.545635	2.545635	NaN

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Los resultados muestran que el Clúster 0 tiene una variabilidad notable en el indice_narcotrafico_tasa, evidenciada por su alta desviación estándar (1.66), junto con valores medios y medianas negativos en otras variables como indice_criminalidad_general_tasa y total_reclam_rest_tierr_minagr_tasa. Esto sugiere una concentración en actividades de narcotráfico con menor presencia de otros tipos de violencia.

En contraste, el clúster 1 presenta mayor dispersión general en las variables, con una desviación estándar especialmente alta en armas_confis_polinal_tasa y total_reclam_rest_tierr_minagr_tasa, lo cual indica una diversidad de situaciones de violencia dentro de este grupo.

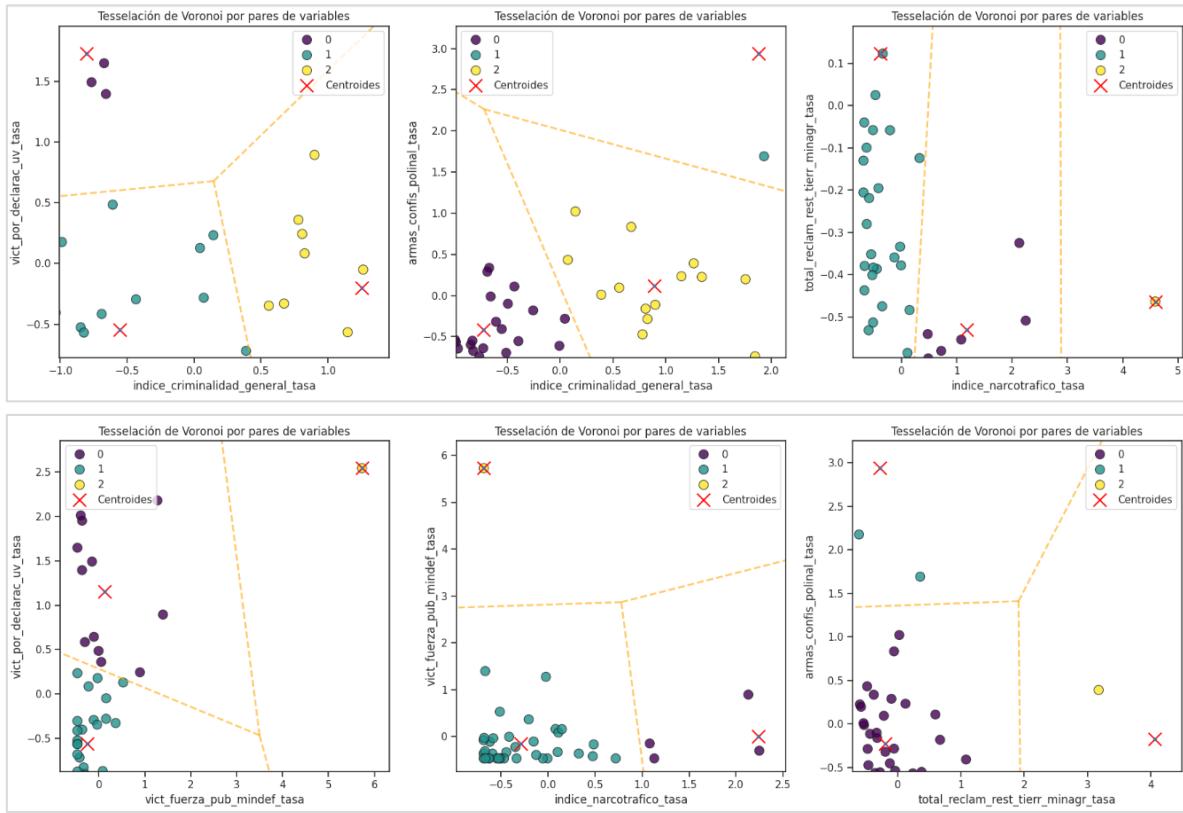
El clúster 2, que agrupa a un solo municipio, presenta valores extremadamente altos en vict_fuerza_pub_mindef_tasa y vict_por_declarac_uv_tasa, reflejando la intensidad de estos tipos de violencia en ese municipio específico. La desviación estándar en este clúster no está definida porque para calcularla se necesita al menos dos datos y este clúster está conformado únicamente por un municipio.

Estas estadísticas refuerzan los perfiles identificados en los centroides, indicando que cada clúster posee características distintas y evidencian la variabilidad en los tipos de violencia entre los municipios analizados en el Cauca.

A continuación se muestra las gráficas “Tesselación de Voronoi” en 2D utilizando los centroides, lo que permite visualizar los límites de decisión entre clústeres en el espacio de por cada par de variables (solo algunos pares).

Figura 100

“Tesselation de Voronoi” en 2D por cada par de variables



Fuente: Pantallazo obtenido desde el entorno de Google Colab

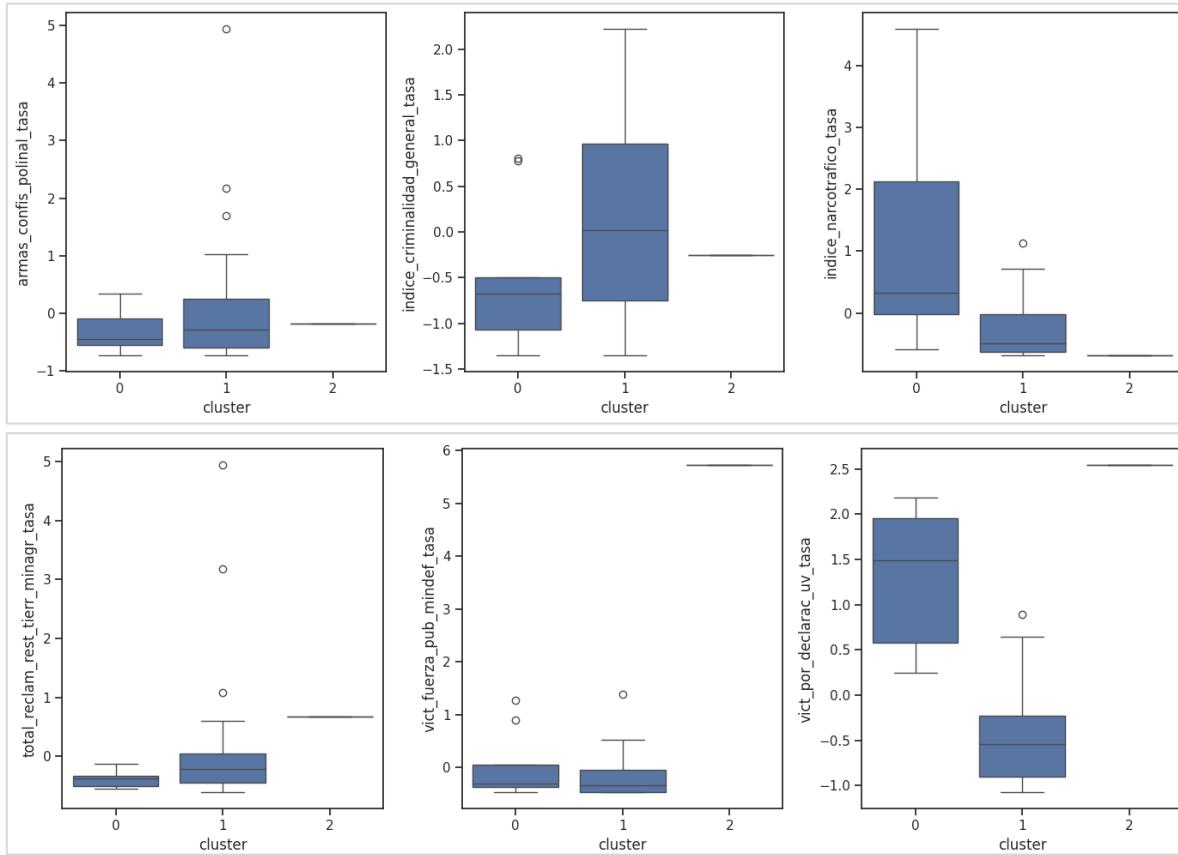
En el gráfico cada punto representa un municipio, coloreado según el clúster al que es asignado. Los límites de Voronoi indican las regiones del espacio de variables que están más cerca de cada centroide, y marcan visualmente las zonas donde cambia la asignación de clúster entre municipios. Los centroides se destacan con marcas rojas, y las líneas de Voronoi en naranja e ilustran los límites naturales de cada clúster (Géron, 2023).

La gráfica permite observar la distribución de los municipios en función de las dos variables y su proximidad a cada clúster. Sin embargo, como solo se analiza un par de variables, no refleja la complejidad completa de los datos multivariados de violencia, solo es útil para entender relaciones específicas entre las dos variables y los clústeres.

El siguiente análisis explora la distribución de cada variable de violencia en función de los clústeres mediante gráficos de boxplot, permitiendo una interpretación de las características y patrones internos de cada grupo. Cada gráfico ilustra las variaciones en la dispersión, la mediana y la presencia de outliers, revelando diferencias en los perfiles de violencia entre clústeres.

Figura 101

Análisis comparativo de los clústeres mediante gráficos de boxplots



Fuente: Pantallazo obtenido desde el entorno de Google Colab

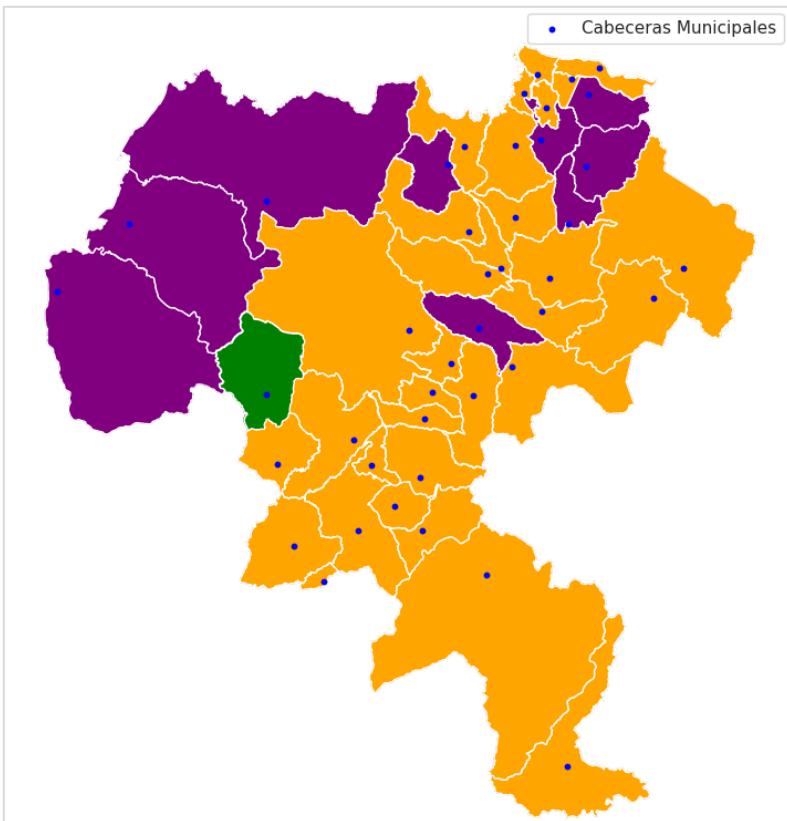
Los boxplots muestran diferencias notables en el tamaño de las cajas, que reflejan la variabilidad interna de cada clúster en cada variable, así como en la posición de la mediana, indicando la concentración relativa de los valores. En términos generales, el clúster 1 tiende a presentar cajas más grandes, señalando una mayor dispersión de sus datos, mientras que el clúster 0 muestra cajas pequeñas en varias variables, sugiriendo una mayor homogeneidad en su composición. La presencia de outliers es otra característica distintiva, especialmente en el clúster 1, donde se observan valores atípicos en la parte superior, indicando casos con niveles de violencia que destacan por encima del rango general del clúster. En cambio, el clúster 2, representado por un único municipio, aparece solo como una línea en los gráficos, lo que subraya la particularidad de este clúster y su diferencia frente a los demás.

Utilizando Python y su librería geoespacial geopandas, se genera el mapa siguiente que muestra la distribución de los municipios del Cauca según los clústeres identificados. El mapa permite observar la disposición territorial de cada clúster, proporcionando una visión

visualmente intuitiva de los resultados de la agrupación. El mapa es una muestra del poder de la estructura geoespacial de la base de datos creada para este proyecto.

Figura 102

Distribución de los municipios del Cauca según los clústeres identificados



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Para evaluar la calidad y robustez de los resultados obtenidos en el modelo de clustering K-means anteriormente expuesto, se procede aplicar métricas de evaluación ampliamente utilizadas en la literatura de aprendizaje automático, como son el coeficiente de silueta y la inercia del modelo. El coeficiente de silueta brinda una medida de cuán bien asignados están los municipios a sus respectivos clusters, reflejando el grado de cohesión interna y separación entre los grupos. Por su parte, la inercia del modelo cuantifica la dispersión de los datos dentro de cada cluster, dando indicios de qué tan compactos o definidos están los grupos. El análisis de estos indicadores, tanto a nivel global del modelo como desagregado por cluster y por variable, permite evaluar de manera integral la calidad del agrupamiento generado, identificar áreas de mejora y orientar el refinamiento del análisis en las siguientes etapas del estudio (Géron, 2023).

El código Python para implementar el proceso de evaluación junto con los resultados obtenidos sobre este modelo K-means inicial se muestra en la siguiente figura.

Figura 103

Evaluación de K-means con configuración básica: Coeficiente de Silueta e Inercia

```
from sklearn.metrics import silhouette_samples, silhouette_score

# Calcular el coeficiente de silueta para cada punto
silhouette_vals = silhouette_samples(X_violencia, kmeans.labels_)

# Añadir el coeficiente de silueta al DataFrame original
df_violencia_scaled['silhouette'] = silhouette_vals

# Calcular los coeficientes de silueta promedio
silhouette_by_cluster = df_violencia_scaled.groupby('cluster')['silhouette'].mean()
silho_by_var = X_violencia.apply(lambda col: silhouette_samples(col.values.reshape(-1, 1), kmeans.labels_)).mean()
silhouette_avg = silhouette_score(X_violencia, df_violencia_scaled['cluster'])

# Imprimir los resultados
print(f"Coefficiente de Silueta promedio por cluster:\n{silhouette_by_cluster}")
print(f"\nCoefficiente de silueta promedio por variable:\n{silho_by_var}")
print(f"\nCoefficiente de Silueta promedio para el modelo: {silhouette_avg:.3f}")
print(f"\nInercia del modelo: {kmeans.inertia_:.3f}")

Coefficiente de Silueta promedio por cluster:
cluster
0    0.252349
1    0.304148
2    0.000000
Name: silhouette, dtype: float64

Coefficiente de silueta promedio por variable:
armas_configis_polinal_tasa      -0.437475
indice_criminalidad_general_tasa -0.298720
indice_narcotrafico_tasa         -0.338402
total_reclam_rest_tierr_minagr_tasa -0.321960
vict_fuerza_pub_mindef_tasa       0.156848
vict_por_declarac_uv_tasa        0.439480
dtype: float64

Coefficiente de Silueta promedio para el modelo: 0.286
Inercia del modelo: 169.681
```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

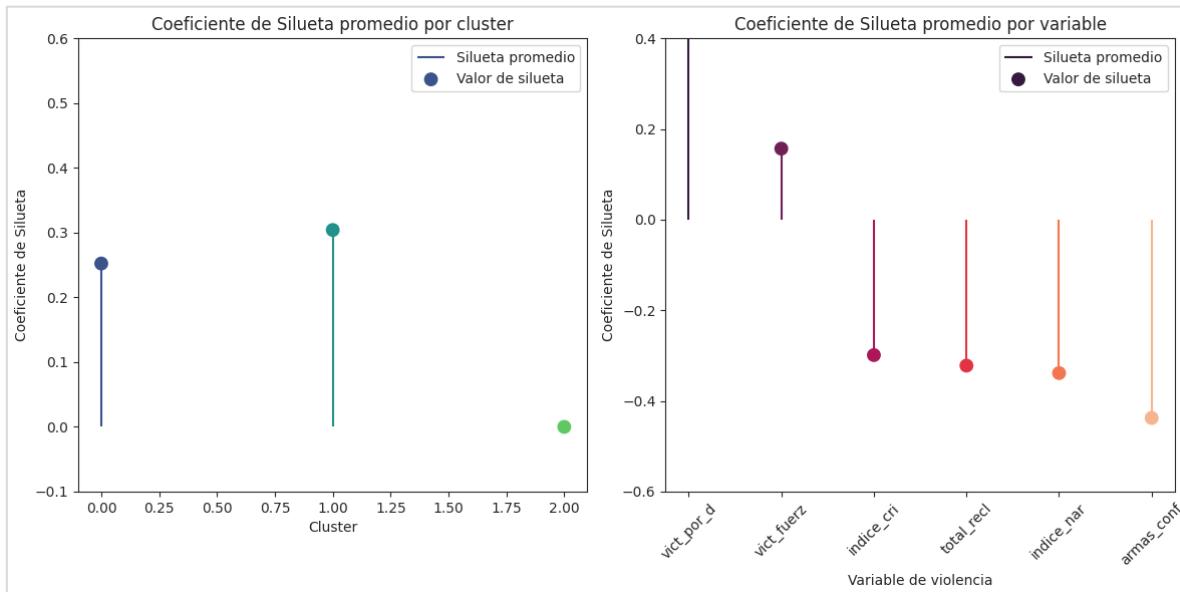
La evaluación inicial del modelo K-means en la configuración básica arroja resultados bajos en cuanto a la cohesión y separación de los clusters. El coeficiente de silueta promedio global, de 0.286, indica que la segmentación lograda es moderadamente débil. Este valor sugiere que, en general, los municipios dentro de un mismo cluster no presentan una gran cercanía entre sí, y la separación entre clusters es limitada. Estos hallazgos son característicos de una estructura con cierto solapamiento, lo que sugiere que los clusters actuales no son del todo compactos ni claramente distintos.

Observando los coeficientes de silueta por cluster, el cluster 1 presenta el promedio más alto (0.304), lo cual apunta a una mayor homogeneidad en esta agrupación en comparación con las demás. En contraste, el cluster 0 muestra un coeficiente menor (0.252), lo que indica una menor cohesión interna. La situación es aún más marcada en el cluster 2, que tiene un

coeficiente de silueta igual a 0, representando un único municipio. Este valor nulo sugiere que el municipio en cuestión es difícil de agrupar y podría presentar características únicas o fuera del rango de los otros clusters, reduciendo la consistencia y robustez de la partición.

Figura 104

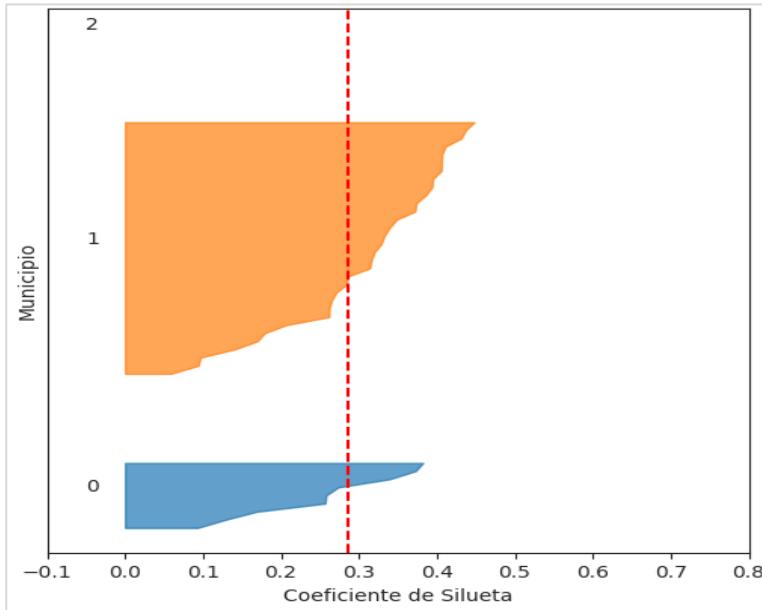
Visualización de los coeficiente de Silueta por clúster y variable mediante Gráfico de Lollipop



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Al analizar la silueta promedio por variable, se destaca un desempeño heterogéneo. Las variables 'vict_por_declarac_uv_tasa' (0.439) y 'vict_fuerza_pub_mindef_tasa' (0.156) exhiben coeficientes de silueta positivos, indicando que aportan diferenciación entre clusters. Sin embargo, las restantes variables, especialmente 'armas_confis_polinal_tasa' (-0.437) e 'indice_narcotrafico_tasa' (-0.338), muestran coeficientes negativos, lo que sugiere que en estas dimensiones los clusters presentan una dispersión notable o ruido. Esto debilita la capacidad del modelo para delinejar una segmentación clara en base a estas características.

Por otra parte, la inercia del modelo es de 169.681, un valor elevado que refleja una significativa dispersión de los puntos dentro de los clusters, lo cual se traduce en un menor nivel de cohesión interna. Este nivel de inercia respalda las observaciones previas sobre la solidez moderada de la segmentación inicial.

Figura 105*Visualización mediante gráfico de silueta*

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Para fortalecer la estructura de los clusters, se procede con una reducción de dimensionalidad mediante PCA, con el fin de concentrar la variabilidad esencial y reducir el impacto de las variables menos diferenciadoras y a la aplicación de la técnica del codo para seleccionar el número óptimo de clusters para una segmentación más precisa y representativa (Becker, 2024).

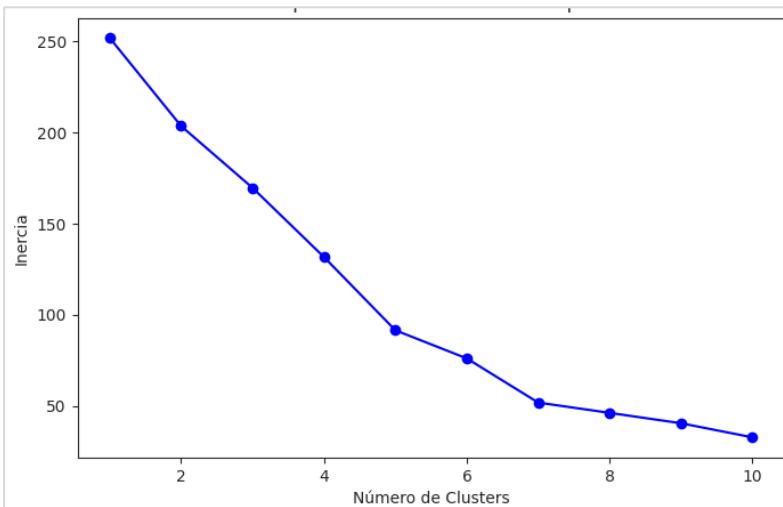
K-means con reducción de dimensionalidad, con anomalías y con cálculo óptimo del número de clústeres. Dado que el modelo que se ha analizado hasta el momento se basó en una configuración inicial básica sin reducción de dimensionalidad, sin filtrado de anomalías y sin búsqueda optima del número de clústeres, a continuación se va a refinar el enfoque aplicando la técnica del codo para determinar el número óptimo de clústeres, implementando PCA (Análisis de Componentes Principales) para reducir la dimensionalidad y mitigar la posible redundancia en las variables con el propósito de obtener una segmentación más precisa de los municipios en función de sus perfiles de violencia.

Para aplicar la técnica del codo, se entrena varios modelos K-means con un rango de valores de k y se calcula la inercia de cada uno, al graficar los resultados, el gráfico resultante muestra la inercia en función del número de clusters, y el "codo" debería ser visible en el punto donde la tasa de disminución se estabiliza, lo que indicará el número óptimo de clusters (Bobadilla, 2021).

En la siguiente figura se presenta la gráfica de la técnica de codo generada a partir de un Script de Python:

Figura 106

Evaluación del número óptimo de clusters mediante el método del codo



Fuente: Pantallazo obtenido desde el entorno de Google Colab

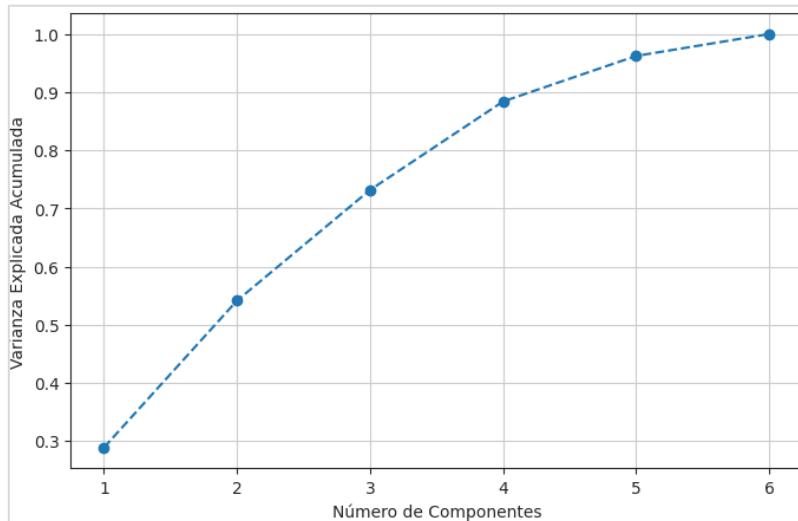
Al graficar la inercia para un rango de valores de k , se observa una disminución pronunciada que comienza a estabilizarse alrededor de los valores 5, 6 y 7, donde se identifica un posible "codo" en el gráfico. Si bien el punto de estabilización no se aparece inequívocamente definido en la gráfica anterior, el número de clústeres en 6 parece ser una elección adecuada, porque puede proporcionar un balance entre la reducción de inercia y la complejidad del modelo. Se espera que esta selección optimice la cohesión interna de los clústeres y la diferenciación entre ellos.

Como parte del proceso de refinamiento, además de la selección del número óptimo de clústeres, también se procede a implementar la técnica de Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos y mitigar la redundancia entre las variables. Este enfoque permite transformar el conjunto de variables originales en un nuevo espacio de componentes principales ordenadas según la varianza explicada, maximizando así la información retenida en menos dimensiones.

La figura siguiente muestra los resultados obtenidos después de ajustar las variables de violencia mediante la clase PCA de la librería sklearn.decomposition, mediante la gráfica de varianza explicada.

Figura 107

Análisis de varianza explicada por PCA



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Los valores de la varianza explicada en la gráfica son:

- 1 componente: 28.88%
- 2 componentes: 54.15%
- 3 componentes: 73.18%
- 4 componentes: 88.40%
- 5 componentes: 96.23%
- 6 componentes: 100.00%

A la luz de estos resultados, lo que se aconseja es seleccionar las primeras cuatro componentes, ya que explican aproximadamente el 88.4% de la varianza total. Este nivel de retención de información es considerado adecuado para el análisis, ya que logra un balance entre simplificación del modelo y precisión representativa. La elección de las primeras cuatro componentes ofrece una representación compacta de los datos mientras mantiene la mayoría de las relaciones presentes en el conjunto original, evitando así el riesgo de sobreajuste que podría surgir al incluir demasiadas componentes.

Es importante subrayar que una mayor proporción de varianza explicada no garantiza necesariamente una mejor calidad de agrupación. La efectividad de la segmentación mediante K-means depende también de la estructura inherente de los datos en el espacio transformado y de cuán bien esta estructura respalde la formación de clústeres coherentes (Géron, 2023).

Antes de continuar con el proceso es importante detenerse en interpretar las dimensiones transformadas obtenidas mediante PCA, es importante analizar los pesos o *loadings* de cada variable en cada componente. Los *loadings* indican la contribución de cada variable original a las componentes principales, permitiendo identificar qué variables dominan en cada una de ellas. Esta información ayuda a interpretar el contenido informativo de cada componente y a entender cómo representan los patrones de los datos originales.

Figura 108

Loadings de las variables en cada componente

# Obtener los pesos de cada variable en cada componente principal						
loadings = pd.DataFrame(pca_full.components_.T, columns=[f'PC{i+1}' for i in range(len(pca_full.components_))], index=X_violencia.columns)						
# Mostrar los loadings						
print("Pesos de las variables en cada componente:")						
loadings						
Pesos de las variables en cada componente:						
PC1	PC2	PC3	PC4	PC5	PC6	
armas_confis_polinal_tasa	0.582487	0.168271	-0.069036	-0.513712	0.406968	-0.445090
indice_criminalidad_general_tasa	0.641138	0.315204	0.111265	0.097816	-0.185897	0.658090
indice_narcotrafico_tasa	0.217834	0.216469	-0.649472	0.606946	-0.100402	-0.324672
total_reclam_rest_tierr_minagr_tasa	0.044184	0.314065	0.717848	0.488837	0.176438	-0.337660
vict_fuerza_pub_mindef_tasa	-0.221654	0.638510	0.018840	-0.344859	-0.614417	-0.215369
vict_por_declarac_uv_tasa	-0.388738	0.564925	-0.213027	-0.016117	0.617333	0.320940

Fuente: Pantallazo obtenido desde el entorno de Google Colab

El análisis de los pesos de las variables en las componentes principales ofrece una visión sobre cómo las variables se agrupan en función de relaciones similares u opuestas en la misma componente, revelando la estructura latente en el conjunto de datos. Al analizar estos patrones, es posible interpretar cada componente en términos de los aspectos de violencia que están más fuertemente representados.

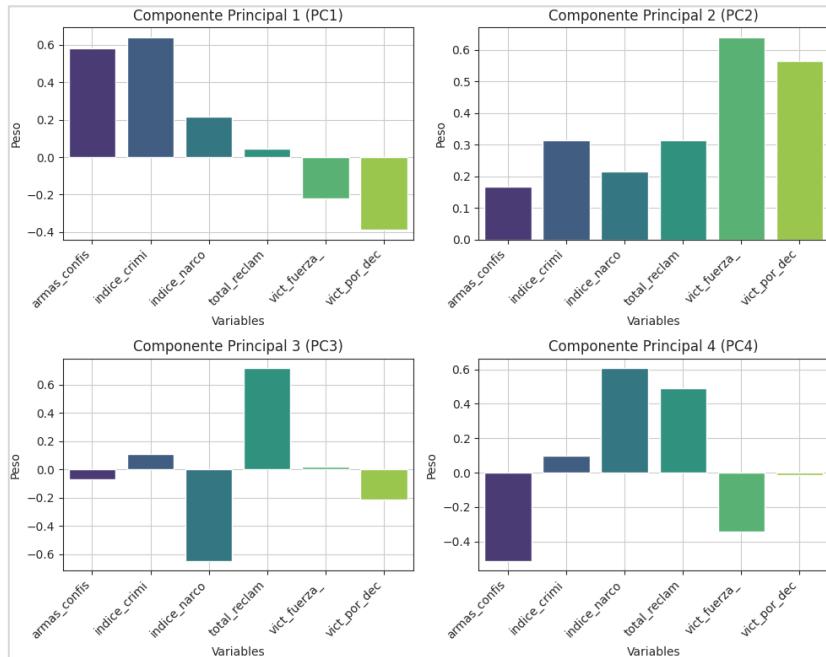
Los resultados muestran que la primera componente principal (PC1) concentra una carga significativa en las variables de armas_confis_polinal_tasa e indice_criminalidad_general_tasa, con pesos positivos de 0.582 y 0.641, respectivamente. Esto sugiere que PC1 representa una dimensión de la violencia asociada principalmente a la criminalidad y las intervenciones policiales contra armas, capturando un perfil en el que la violencia criminal y la acción represiva juegan roles predominantes.

La segunda componente principal (PC2) tiene sus mayores pesos en vict_fuerza_pub_mindef_tasa y vict_por_declarac_uv_tasa, con 0.639 y 0.565,

respectivamente. Estas variables están asociadas a las víctimas en situaciones de violencia, tanto por parte de la fuerza pública como en el marco de desplazamientos, lo que posiciona a PC2 como un eje de victimización en contextos de conflicto. Este componente captura un perfil de violencia relacionado con las consecuencias para la población civil en entornos del conflicto interno.

Figura 109

Grafica de los pesos de las variables en cada componente



Fuente: Pantallazo obtenido desde el entorno de Google Colab

En la tercera componente principal (PC3), destaca la carga de total_reclam_rest_tierr_minagr_tasa y indice_narcotrafico_tasa, con pesos de 0.718 y -0.649, respectivamente. La combinación de estas variables en sentidos opuestos indica un eje que involucra disputas por tierras y actividades de narcotráfico.

La cuarta componente principal (PC4) presenta pesos elevados y opuestos en indice_narcotrafico_tasa y armas_confis_polinal_tasa, con valores de 0.607 y -0.514, respectivamente. Esto puede interpretarse como un componente que captura la relación inversa entre narcotráfico y confiscaciones policiales, sugiriendo que, en algunos casos, el incremento de una de estas variables podría implicar la reducción de la otra, probablemente por razones de concentración de recursos o políticas locales.

Como se ve cada componente principal agrupa un conjunto específico de variables, reflejando distintos aspectos del fenómeno de la violencia. A través de estas dimensiones, el

análisis de componentes principales (PCA) facilita una representación más compacta de los datos, resaltando patrones clave que simplifican la estructura de las variables y pueden mejorar la interpretación en la segmentación mediante K-means.

Ahora se procede a refinar el modelo K-means con una configuración más optimizada, utilizando los dos parámetros clave que se acaba de definir: el número de clústeres y las dimensiones transformadas mediante PCA. El análisis de la técnica del codo señala que el número óptimo de clústeres es 6 y el análisis de componentes principales en el PCA, dice que con 4 componentes se abarca el 88.4% de la varianza total.

Seis clústeres y cuatro componentes, deberían producir la mejor segmentación de los datos. No obstante, el hecho de seleccionar componentes que explican una alta proporción de la varianza no garantiza una agrupación óptima, pues los patrones detectados pueden no reflejar todos los matices de la relación entre las variables de violencia en diferentes contextos municipales (Géron, 2023).

Para comprobar si esta selección de cuatro componentes es la configuración que produce la mejor segmentación, se procede a realizar un análisis de sensibilidad en el número de componentes. Este análisis incluye combinaciones de 6 clústeres con distintas configuraciones del PCA: 4 componentes (88.4% de la varianza), 5 componentes (96.2%), 3 componentes (73.2%) y 2 componentes (54.2%). En cada caso, se evalúa la calidad de la agrupación mediante el coeficiente de silueta, que mide la coherencia interna de los clústeres, permitiendo determinar la configuración que optimiza la cohesión y separación entre los clústeres. El propósito de esta análisis es que en lo sucesivo se avance con el modelo que proporciona la segmentación más robusta y representativa de los datos. El script Python que implementa este análisis se muestra en la siguiente figura junto con los resultados obtenidos.

Los resultados del análisis de sensibilidad en el número de componentes sorpresivamente muestran una relación inversa entre el número de componentes y la calidad de la agrupación medida por el coeficiente de silueta. Específicamente, al usar solo dos componentes principales, el modelo logra un coeficiente de silueta de 0.4989, indicando una aceptable cohesión y separación de los clústeres. Sin embargo, a medida que se incrementa el número de componentes, el coeficiente de silueta disminuye de manera notable, pasando a 0.3703 con tres componentes, 0.3199 con cuatro, y 0.2857 con cinco componentes.

Estos resultados confirman la advertencia previamente discutida en el informe: un mayor número de componentes principales, aunque represente una fracción acumulada más alta de la varianza de los datos, no garantiza necesariamente una mejora en la calidad de la agrupación. En el contexto de clustering, la clave está en balancear la información contenida

en los datos con la simplificación de las dimensiones, ya que el aumento de componentes podría incluir ruido o redundancia que perturbe la estructura subyacente de los clústeres (Géron, 2023).

Figura 110

Análisis de sensibilidad: Número de componentes PCA y calidad de la agrupación

```

from sklearn.metrics import silhouette_score

# Definir el rango de componentes a evaluar
num_components = [2, 3, 4, 5]
num_clusters = 6 # Basado en el resultado del método del codo
best_score = -1
best_num_components = None

# Almacenar los resultados
results = {}

for n_components in num_components:
    # Aplicar PCA con el número de componentes especificado
    pca = PCA(n_components=n_components)
    X_pca = pca.fit_transform(X_violencia)

    # Ajustar el modelo K-means con el número de clusters óptimo
    kmeans_pca = KMeans(n_clusters=num_clusters, random_state=0)
    labels = kmeans_pca.fit_predict(X_pca)

    # Calcular el coeficiente de silueta
    silhouette_avg = silhouette_score(X_pca, labels)
    results[n_components] = silhouette_avg

    # Verificar si es el mejor score
    if silhouette_avg > best_score:
        best_score = silhouette_avg
        best_num_components = n_components

print(f"Número de componentes: {n_components}, Coeficiente de Silueta: {silhouette_avg:.4f}")

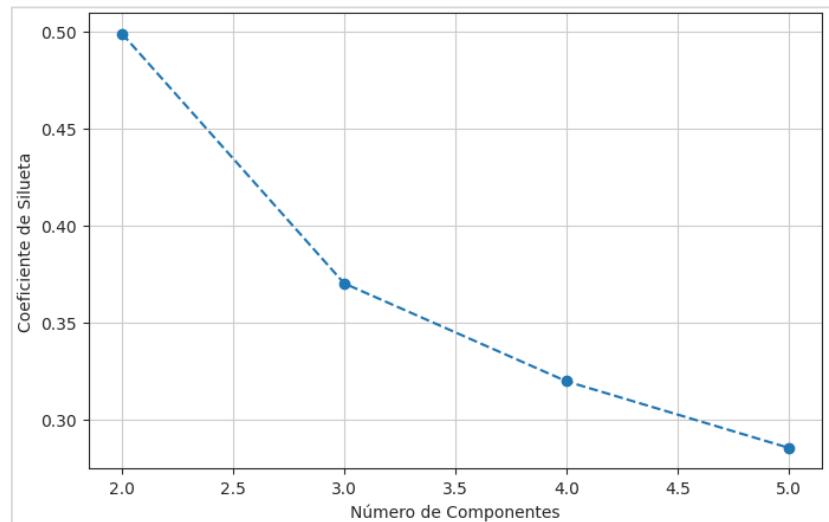
Número de componentes: 2, Coeficiente de Silueta: 0.4989
Número de componentes: 3, Coeficiente de Silueta: 0.3703
Número de componentes: 4, Coeficiente de Silueta: 0.3199
Número de componentes: 5, Coeficiente de Silueta: 0.2857

```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Figura 111

Gráfica de análisis de sensibilidad componentes PCA y calidad de la agrupación



Fuente: Pantallazo obtenido desde el entorno de Google Colab

En este caso, parece que la reducción a dos componentes no solo preserva un grado suficiente de la varianza (54.15%), sino que además facilita la identificación de clústeres bien definidos y distinguibles. Esta simplificación, aunque no maximice la varianza acumulada, permite que el modelo K-means resalte las diferencias esenciales entre grupos sin ser afectado por dimensiones adicionales que podrían no aportar valor a la segmentación en términos de cohesión y separación.

Dado lo anterior se continua el análisis con dos componentes principales para optimizar la calidad de la agrupación. Esta elección asegura que el modelo mantenga una estructura de clústeres más nítida. Además basándose en los dos primeros componentes principales, el modelo de segmentación propuesto capture de manera efectiva las dos dimensiones clave de la violencia en los municipios estudiados: la criminalidad activa con implicaciones en el control de armas, y la victimización de la población civil en contextos de conflicto, tal como se vio cuando se analizó los pesos de las variables en cada uno de los componentes PCA.

Como ya se estableció en este informe, la primera componente principal (PC1) destaca áreas con altos niveles de criminalidad y acciones policiales contra armas, sugiriendo zonas de intervención orientadas a la criminalidad estructural y el control armamentístico. Por otro lado, la segunda componente principal (PC2) refleja territorios donde la violencia afecta directamente a la población civil, caracterizada por victimización en interacciones con el conflicto interno.

Esta representación en dos dimensiones proporciona un modelo claro donde el primer componente principal (PC1) describe aspectos estructurales y de criminalidad, mientras que el segundo componente principal (PC2) se centra en la victimización y los impactos en la población. La elección de estos dos componentes no solo permite obtener los mejores resultados de la agrupación, sino que permite una interpretación significativa y distintiva de los patrones de violencia en los municipios.

La nueva distribución de municipios en los clústeres obtenidos con el modelo refinado de K-means, utilizando dos componentes principales, muestra lo siguiente:

- Clúster 0: 10 municipios
- Clúster 1: 3 municipios
- Clúster 2: 1 municipio
- Clúster 3: 13 municipios
- Clúster 4: 5 municipios
- Clúster 5: 10 municipios

Figura 112

Asignación de clúster a cada municipio con la configuración refinada de K-means

```
# Reducir a las 2 componentes principales
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_violencia)

# Definir y entrenar el modelo K-means con 5 clusters
kmeans_refinado = KMeans(n_clusters=6, random_state=42)
kmeans_refinado.fit(X_pca)

# Asignar etiquetas de clusters al dataset original
df_violencia_scaled['cluster_pca'] = kmeans_refinado.labels_

# Mostrar la cantidad de municipios en cada cluster
print("Conteo de municipios por cluster:")
print(df_violencia_scaled['cluster_pca'].value_counts())

Conteo de municipios por cluster:
cluster_pca
3    13
0     10
5     10
4      5
1      3
2      1
Name: count, dtype: int64
```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Este resultado refleja una estructura de agrupación donde algunos clústeres contienen una mayor cantidad de municipios con características de violencia similares, mientras que el clúster 2, al igual que en el análisis inicial, está representado por un solo municipio, Argelia. Esto indica diferencias significativas entre Argelia y el resto de los municipios, pero seguiremos profundizando en el análisis para comprender estas posibles singularidades y la estructura interna de cada grupo.

Figura 113

Lista de municipios por cluster

cluster_pca	municipios_en_cluster	cantidad_municipios
0	POPAYAN,BALBOA,BUENOS AIRES,GUAPI,JAMBALO,LOPEZ DE MICAY,MORALES,SUAREZ,TIMBIQUI,TORIBIO	10
1	PUERTO TEJADA,ROSAS,VILLA RICA	3
2	ARGELIA	1
3	ALMAGUER,BOLIVAR,CAJIBIO,CALDONO,INZA,LA VEGA,PAEZ,PURACE,SAN SEBASTIAN,SILVIA,SOTARA PAISPAMBA,SUCRE,TOTORO	13
4	CALOTO,CORINTO,FLORENCIA,PIAMONTE,SANTA ROSA	5
5	EL TAMBO,GUACHENE,LA SIERRA,MERCADERES,MIRANDA,PADILLA,PATIA,PIENDAMO - TUNIA,SANTANDER DE QUILICHAO,TIMBIO	10

Fuente: Pantallazo obtenido desde el entorno de Google Colab

A partir de esta agrupación, el estudio se detiene en un examen detallado de las variables dentro de cada clúster, examinando las tendencias y patrones de cada grupo. Esto permite caracterizar las dimensiones de violencia y victimización con mayor precisión y comprender mejor la dinámica interna en el Departamento del Cauca.

En el análisis de los centroides en el espacio de las componentes principales (PCA), cada vector en la matriz que se muestra en la figura siguiente representa la posición central de un clúster en el plano definido por las dos principales dimensiones de violencia identificadas: la criminalidad e intervenciones policiales (PC1) y la victimización de la población en el contexto del conflicto (PC2).

Figura 114

Matriz de centroides en el espacio PCA

	PCA_1	PCA_2
0	-1.140383	0.302880
1	3.292396	0.551593
2	-2.647932	5.047507
3	-0.509841	-1.203523
4	0.643716	1.549609
5	0.758392	-0.183334

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Estos centroides proporcionan una referencia para interpretar el perfil de cada grupo en función de la carga de cada componente. Los centroides para cada clúster reflejan la siguiente estructura en el espacio de las componentes:

- Clúster 0: (-1.140383, 0.302880). Este clúster presenta un valor negativo en la dimensión PC1 y un valor levemente positivo en PC2, sugiriendo un perfil con menores incidencias de criminalidad e intervenciones policiales, y una victimización moderada. Este punto central podría representar municipios con un nivel de violencia algo controlado y una incidencia de victimizaciones no extrema.
- Clúster 1: (3.292396, 0.551593). La alta positividad en PC1 indica que los municipios de este clúster tienen una fuerte presencia de criminalidad y actividades represivas, siendo el grupo más destacado en esta dimensión. El valor positivo moderado en PC2

también sugiere un nivel significativo de victimización, aunque menos marcado que en otras agrupaciones. Este clúster refleja municipios donde la acción represiva y el crimen son factores predominantes.

- Clúster 2: (-2.647932, 5.047507). La posición en el extremo positivo de PC2 y en el extremo negativo de PC1 muestra un perfil de victimización excepcionalmente alto en el contexto de violencia, particularmente en relación con el conflicto armado interno, mientras que la criminalidad en sí es menos intensa. Este clúster, que contiene un solo municipio (Argelia), representa un perfil atípico y extremo, donde las víctimas de la violencia constituyen el rasgo distintivo más relevante.
- Clúster 3: (-0.509841, -1.203523). Con valores negativos en ambas dimensiones, los municipios en este clúster muestran niveles bajos tanto en criminalidad como en victimización. Este perfil podría ser característico de municipios con una relativa estabilidad y menor exposición a las dinámicas de violencia observadas en otras agrupaciones.
- Clúster 4: (0.643716, 1.549609). La posición en valores positivos en ambas dimensiones indica una combinación moderada de criminalidad y victimización en estos municipios. Este clúster muestra un perfil de violencia mixto, donde ambos ejes de violencia presentan una incidencia intermedia en comparación con los extremos observados en otros clústeres.
- Clúster 5: (0.758392, -0.183334). Este grupo también tiene un valor positivo en PC1, pero un valor casi neutral en PC2, lo cual sugiere que los municipios en este clúster presentan una tendencia moderada hacia la criminalidad y acciones represivas, sin una marcada incidencia de victimización. Podría representar un grupo de municipios con desafíos en criminalidad, pero menos afectados por los impactos directos en la población civil del conflicto armado.

Figura 115*Estadísticas descriptivas por clúster con K-means con configuración refinada*

# Seleccionar solo las columnas numéricas de interés para el análisis df_numerical = df_violencia_scaled[['varViolencia', 'cluster_pca']]																		
# Calcular estadísticas descriptivas (media, mediana, desviación estándar) por cada clúster en las columnas numéricas cluster_mean_std_pca = df_numerical.groupby('cluster_pca').agg(['mean', 'median', 'std']) print("Media, mediana y desviación estándar de las variables numéricas por clúster:") cluster_mean_std_pca																		
Media, mediana y desviación estándar de las variables numéricas por clúster:																		
cluster_pca																		
0	-0.342837	-0.384684	0.342559	-0.771145	-0.719154	0.395079	0.130923	-0.073469	0.907686	-0.280249	-0.346508	0.249567	-0.027424	-0.226094	0.542938	1.205158	1.443314	0.789219
1	2.936795	2.175735	1.754402	1.877412	1.930362	0.374360	0.395424	0.110663	0.641874	-0.280311	-0.585165	0.551321	-0.283402	-0.467221	0.318384	-0.621249	-0.683839	0.287845
2	-0.180710	-0.180710	NaN	-0.255795	-0.255795	NaN	-0.683383	-0.683383	NaN	0.670029	0.670029	NaN	5.728880	5.728880	NaN	2.545635	2.545635	NaN
3	-0.537347	-0.610717	0.261304	-0.749063	-0.823411	0.335212	-0.543739	-0.624866	0.210997	-0.131640	-0.280131	0.430394	-0.396685	-0.467221	0.114749	-0.822519	-0.915497	0.251045
4	-0.218252	-0.159653	0.424315	1.119606	0.900455	0.449162	1.195697	0.080268	2.174352	1.379047	-0.325004	2.526646	0.478768	0.159078	0.640888	0.417383	0.359208	0.364202
5	0.287546	0.211753	0.386945	0.647479	0.615929	0.651436	-0.072199	-0.269569	0.490342	-0.221053	-0.346833	0.391591	-0.184137	-0.196633	0.287500	-0.412763	-0.339268	0.395770

Fuente: Pantallazo obtenido desde el entorno de Google Colab

El análisis de estadísticas descriptivas en la configuración refinada (ver figura 115) permite profundizar en los perfiles de violencia de cada clúster en el espacio reducido de PCA, en función de las variables de violencia.

- Clúster 0: Este grupo presenta una variabilidad significativa en variables como vict_por_declarac_uv_tasa, con un valor medio elevado (1.21) y una mediana similar (1.44), lo cual sugiere que los municipios en este clúster tienden a reportar una alta victimización en el contexto de desplazamiento forzado. En contraste, las tasas de otras variables, como indice_criminalidad_general_tasa y armas_confis_polinal_tasa, son negativas tanto en la media como en la mediana, indicando una menor incidencia de criminalidad y decomisos de armas en comparación con los otros clústeres.
- Clúster 1: El Clúster 1 se distingue por sus valores positivos en armas_confis_polinal_tasa y indice_criminalidad_general_tasa, con medias de 2.94 y 1.88 respectivamente. Estas estadísticas reflejan una mayor intervención policial y niveles de criminalidad en estos municipios. Sin embargo, la desviación estándar elevada en armas_confis_polinal_tasa (1.75) sugiere una amplia variabilidad en la intervención policial dentro del grupo, lo que podría indicar una heterogeneidad en las prácticas de seguridad entre los municipios.
- Clúster 2: Agrupando únicamente a un municipio, este clúster muestra valores extremadamente altos en vict_fuerza_pub_mindef_tasa (5.73) y vict_por_declarac_uv_tasa (2.54). La ausencia de desviación estándar en este clúster se debe a su composición unitaria, lo que enfatiza el perfil extremo y particular de violencia

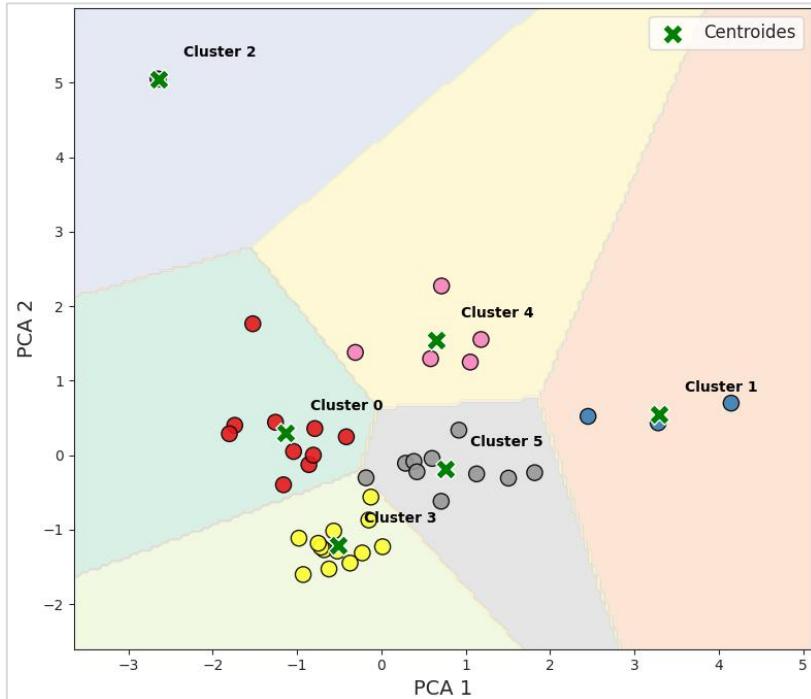
de este municipio, en el cual la victimización por parte de la fuerza pública es altamente significativa y persistente.

- Clúster 3: Caracterizado por valores medianamente negativos en varias variables, refleja un perfil de violencia bajo en términos generales. Las variables como indice_criminalidad_general_tasa y armas_confis_polinal_tasa tienen medias y medianas negativas, lo cual sugiere una incidencia relativamente baja de criminalidad y decomisos de armas. La desviación estándar moderada, especialmente en total_reclam_rest_tierr_minagr_tasa, sugiere cierta diversidad en las reclamaciones de tierras.
- Clúster 4: Este clúster presenta un perfil único de violencia con una media positiva destacable en indice_narcotrafico_tasa (1.19) y en total_reclam_rest_tierr_minagr_tasa (1.38), reflejando una prevalencia de actividades relacionadas con el narcotráfico y conflictos en la restitución de tierras. La desviación estándar alta en ambas variables sugiere que existen diferencias significativas en los niveles de violencia dentro de los municipios de este clúster, posiblemente debido a contextos diversos de violencia y narcotráfico.
- Clúster 5: Presenta valores medios y medianas moderados en casi todas las variables, lo cual lo posiciona en un perfil de violencia menos extremo. La variabilidad en indice_criminalidad_general_tasa (desviación estándar de 0.65) indica un nivel de criminalidad intermedio en los municipios de este clúster. La estabilidad general en la mayoría de las variables sugiere que este grupo agrupa municipios con incidencias moderadas de violencia en el Cauca.

A continuación se muestra las gráfica “Tesselación de Voronoi” en 2D utilizando los centroides.

Figura 116

Visualización de clústeres con límites de decisión de K-means refinado



Fuente: Pantallazo obtenido desde el entorno de Google Colab

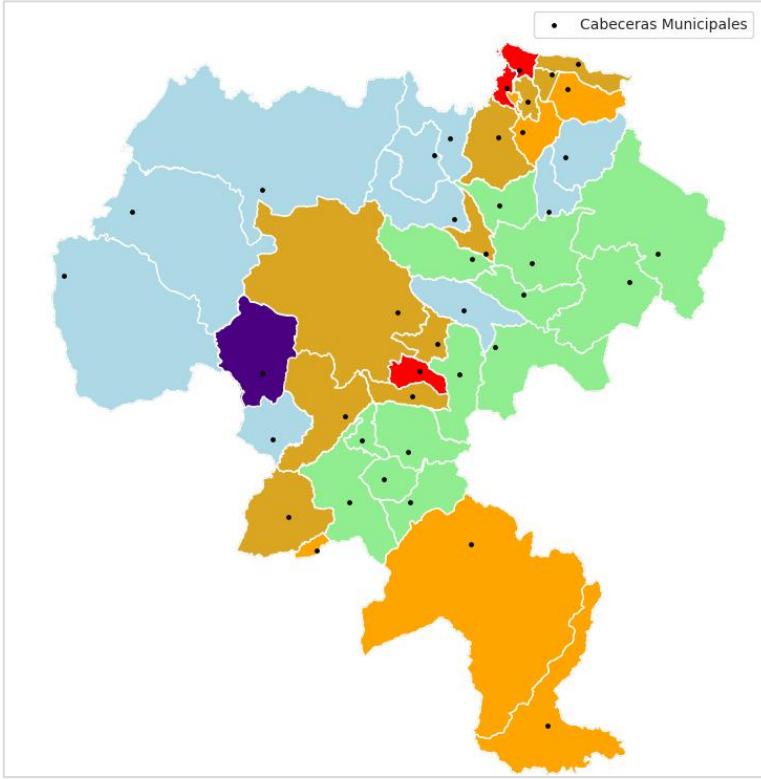
La gráfica permite observar los límites de decisión del modelo K-means en el espacio bidimensional de PCA_1 y PCA_2, mostrando cómo se distribuyen los municipios entre los clusters y la ubicación de los centroides en el plano. Los límites de decisión demarcan áreas de influencia de cada cluster, aunque se observan algunos traslapes en los clusters 3 y 5, lo que indica que algunos municipios presentan características intermedias entre estas agrupaciones. Este solapamiento sugiere proximidad en el espacio PCA, indicativo de similitudes en ciertos patrones de violencia.

Algunos puntos se ubican directamente en los límites entre clusters, como uno asignado al cluster 5 y otro al cluster 3, lo cual refleja la complejidad de estos perfiles intermedios. La visualización de los centroides, etiquetados y claramente diferenciados, facilita la identificación de cada agrupación y resalta las relaciones entre clusters cercanos.

El siguiente mapa generado con código Python con su librería geopandas muestra la distribución de los municipios del Cauca según los clústeres identificados con el modelo k-means refinado. El mapa permite observar la disposición territorial de cada clúster, proporcionando una visión visualmente intuitiva de los resultados de la agrupación.

Figura 117

Distribución de los municipios del Cauca por clústeres modelo refinado



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Para evaluar la calidad y robustez de los resultados obtenidos en el modelo de clustering K-means con parámetros refinados, se procede a aplicar métricas de evaluación el coeficiente de silueta y la inercia del modelo. Los resultados obtenidos sobre este modelo K-means inicial se muestra en la siguiente figura.

Figura 118

Evaluación de K-means con configuración refinada: Coeficiente de Silueta e Inercia

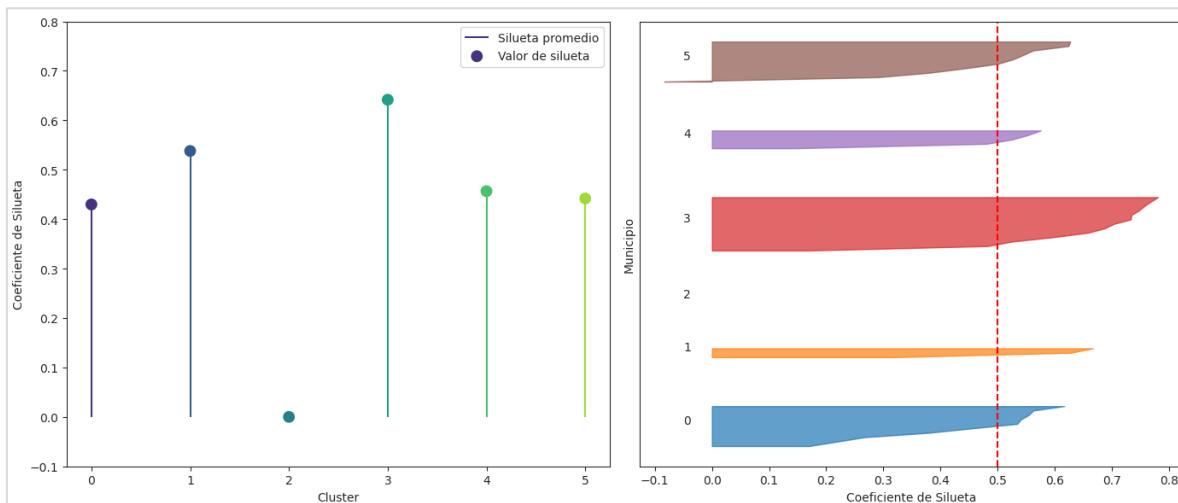
```
Coeficiente de Silueta promedio por cluster:  
cluster_pca  
0    0.429772  
1    0.537898  
2    0.000000  
3    0.641535  
4    0.456541  
5    0.441778  
Name: silhouette_pca, dtype: float64  
  
Coeficiente de Silueta promedio por componente principal:  
[0.09630547605545978, 0.29052090927275265]  
  
Coeficiente de Silueta promedio para el modelo refinado: 0.499  
Inercia del modelo refinado: 14.227
```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

La evaluación del modelo refinado de K-means, aplicada en el espacio reducido de componentes principales (PCA), muestra una mejora significativa en la cohesión y separación de los clusters en comparación con la configuración básica. El coeficiente de silueta promedio global es de 0.499, lo que indica una segmentación moderadamente fuerte. Este valor sugiere que los municipios dentro de cada cluster presentan una mayor cercanía entre sí y una separación más clara entre clusters, reflejando una partición más definida en el contexto de las características principales.

Figura 119

Visualización indicadores de desempeño modelo refinado



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Al analizar los coeficientes de silueta por cluster, el cluster 3 destaca con un valor promedio de 0.642, indicando una alta homogeneidad y cohesión interna. Los clusters 1, 4, y 5 también presentan valores de silueta positivos, aunque más bajos (0.538, 0.457, y 0.442, respectivamente), sugiriendo que mantienen cierta cohesión interna y separación respecto a otros clusters, aunque con menos definición que el cluster 3. Por el contrario, el cluster 2 tiene un coeficiente de silueta de 0, ya que está compuesto por un único municipio, lo cual señala un perfil singular que no se asimila fácilmente a los demás grupos.

En cuanto a los coeficientes de silueta promedio por componente principal, PCA_1 y PCA_2 presentan valores de 0.096 y 0.291, respectivamente. Estos valores reflejan que PCA_2 aporta mayor claridad en la diferenciación de clusters en comparación con PCA_1.

La inercia del modelo se reduce a 14.227, un descenso notable frente a la configuración básica inicial, lo cual respalda la cohesión interna más elevada de los clusters en el modelo refinado y confirma una segmentación mejor definida en el análisis de violencia en el Cauca.

K-means con reducción de dimensionalidad, excluyendo anomalías y con cálculo óptimo del número de clústeres: El análisis de clustering avanza hacia una tercera configuración, en la que se incorpora la exclusión de municipios identificados como anómalos para observar cómo este refinamiento afecta la segmentación en el contexto de violencia en el Cauca. En las dos configuraciones previas, primero aplicamos un modelo K-means básico sin PCA, ni exclusión de anomalías, y luego optimizamos este modelo usando PCA con dos componentes principales y la técnica del codo, sin considerar los municipios anómalos. En esta nueva configuración, se excluyen los municipios detectados como anómalos —ARGELIA, PUERTO TEJADA y SAN ROSA— y se repite el análisis con K-means, con técnica del codo, y PCA, para evaluar el impacto de estos ajustes en el desempeño del modelo.

Al excluir los municipios anómalos, resulta prudente recalcular tanto el codo como el PCA. Pero en esta oportunidad se aplica la técnica del codo sobre las dos componentes principales que se van a utilizar, en lugar de las variables originales.

Los resultados del script de Python que filtra los datos excluyendo las filas clasificadas como anomalías del dataframe, aplica PCA con 2 componentes a los datos filtrados, extrae la proporción de varianza explicada y muestra los pesos (cargas) de las variables originales en las dos componentes principales, se presenta en la figura:

Figura 120

Resultado de análisis PCA excluyendo municipios anómalos

Varianza explicada por el PCA sin anomalías: [0.36414249 0.29708542]
Varianza total explicada (suma): 0.661
Pesos de las variables en las componentes principales:
Componente 1 Componente 2
armas_confis_polinal_tasa 0.344417 -0.215066
indice_criminalidad_general_tasa 0.702900 -0.345947
indice_narcotrafico_tasa 0.616704 0.527325
total_reclam_rest_tier_minagr_tasa 0.011190 -0.168767
vict_fuerza_pub_mindef_tasa 0.079770 0.123746
vict_por_declarac_uv_tasa -0.022284 0.715681

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Con la nueva configuración del modelo, que excluye los municipios que se detectaron como anomalías, se observan resultados que reflejan un ajuste significativo en la estructura latente del conjunto de datos. La varianza total explicada por las dos primeras componentes principales PCA aumenta a un 66.1%, en comparación con el 54.15% de la configuración anterior, lo cual indica una mejora en la capacidad del modelo para capturar la variabilidad subyacente de las variables de violencia con estas dos componentes.

Los pesos de las variables, muestran que la primera componente (PC1) concentra una carga significativa en las variables indice_criminalidad_general_tasa (0.703) e indice_narcotrafico_tasa (0.617), además de una contribución menor de armas_confis_polinal_tasa (0.344). Este perfil indica que PC1 representa una dimensión de la violencia asociada principalmente a la criminalidad general y el narcotráfico, con un aporte adicional de las incautaciones de armas, sugiriendo que PC1 captura un contexto de violencia donde la criminalidad y el tráfico de drogas son factores dominantes.

La segunda componente (PC2) muestra altos pesos en las variables vict_por_declarac_uv_tasa (0.716) e indice_narcotrafico_tasa (0.527), lo que indica una dimensión de victimización y desplazamiento, en la que los efectos sobre la población del conflicto armado tienen una fuerte representación. El peso positivo de vict_por_declarac_uv_tasa junto con el aporte de indice_narcotrafico_tasa en esta componente sugiere que PC2 captura una dinámica de violencia donde las víctimas civiles se ven impactadas por contextos de narcotráfico y conflicto.

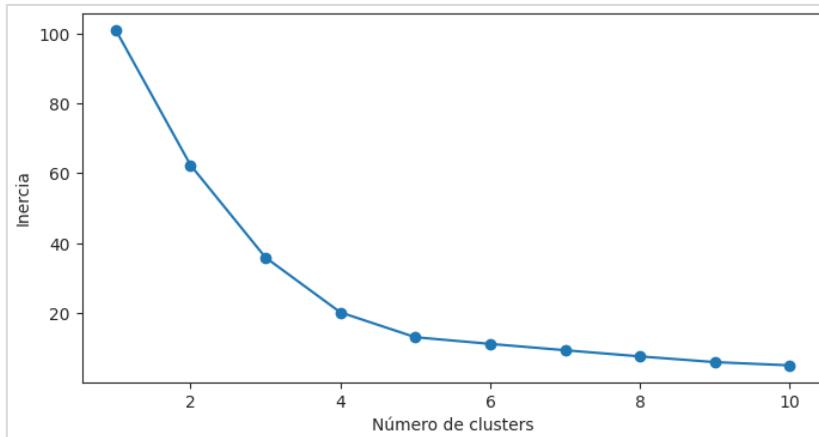
En la configuración previa, PC1 estaba dominada por las variables armas_confis_polinal_tasa e indice_criminalidad_general_tasa, indicando una relación entre criminalidad y acciones policiales en confiscación de armas. En esta nueva configuración, sin embargo, el índice de narcotráfico tiene un peso mayor, desplazando parcialmente la importancia de las intervenciones policiales en la incautación de armas. Asimismo, PC2 en la configuración anterior estaba asociada a las víctimas de la violencia en entornos de conflicto, mientras que en la configuración actual, se refuerza la relación de vict_por_declarac_uv_tasa con el narcotráfico, destacando un perfil de impacto social vinculado al tráfico de drogas y desplazamiento forzado.

La exclusión de las anomalías permite un ajuste más preciso del modelo, capturando un perfil más matizado del fenómeno de violencia en el Cauca. Esto no solo incrementa la varianza explicada sino que también ofrece una estructura de componentes principales que refleja con mayor claridad las dinámicas específicas de criminalidad, narcotráfico y victimización en la región.

Los resultados de aplicar la técnica del codo a los dos componentes principales obtenidos excluyendo anomalías se muestra en la siguiente figura.

Figura 121

Método del codo excluyendo anomalías y usando PCA



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Con el nuevo enfoque, excluyendo anomalías, el análisis de la gráfica de la técnica del codo muestra un "codo" claramente marcado en cuatro clústeres. Vale la pena mencionar que además de estos 4 clústeres, los municipios excluidos se agrupan conceptualmente como una categoría especial, lo que se podría llamar un “cluster de anomalías”, en el sentido de que poseen niveles de violencia o indicadores de conflictividad que los diferencian significativamente del resto. Pero hay que aclarar que aunque estos municipios puedan considerarse interpretativamente como un grupo único, no son un cluster en el sentido estricto del modelo K-means, pues no participan en su proceso de formación. La aplicación del modelo con los nuevos parámetros y el resultado de evaluación se muestra en la figura.

Figura 122

Aplicación de K-means con nueva configuración excluyendo anomalías

```
# Aplicar K-means con el número óptimo de clusters y asignar etiquetas de cluster
n_clusters_opt_pca_excl_anomalos = 4

kmeans_final_pca_excl_anomalos = KMeans(n_clusters=n_clusters_opt_pca_excl_anomalos, random_state=42)
kmeans_final_pca_excl_anomalos.fit(X_pca_excl_anomalos)

# Asignar etiquetas de cluster al DataFrame sin anomalías
dfViolencia_scaled_excl_anomalos = dfViolencia_scaled[dfViolencia_scaled['is_anomalous'] == 0].copy()
dfViolencia_scaled_excl_anomalos['cluster_pca_final'] = kmeans_final_pca_excl_anomalos.labels_

# Evaluar el modelo
silhouette_avg_pca_excl_anomalos = silhouette_score(X_pca_excl_anomalos, kmeans_final_pca_excl_anomalos.labels_)
print(f"Coefficiente de Silueta para el modelo con PCA (excluyendo anomalías): {silhouette_avg_pca_excl_anomalos:.3f}")
print(f"Inercia del modelo refinado (con PCA y sin anomalías): {kmeans_final_pca_excl_anomalos.inertia_:.3f}")

Coefficiente de Silueta promedio para el modelo refinado con PCA (excluyendo anomalías): 0.520
Inercia del modelo refinado (con PCA y sin anomalías): 20.190
```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

El modelo refinado de K-means, aplicando PCA y excluyendo municipios atípicos, muestra una mejora en la calidad de la segmentación de los municipios según sus características de violencia. El coeficiente de silueta promedio global ahora es de 0.52, lo que representa una segmentación moderadamente fuerte, ligeramente superior al valor obtenido en el modelo anterior (0.499). Este incremento sugiere una mayor cohesión interna y una separación más clara entre clusters, reflejando que los municipios dentro de cada grupo son más similares entre sí y más distintos en relación a los otros clusters. Es decir que la exclusión de municipios anómalos, como Argelia, Puerto Tejada y Santa Rosa, permite una partición más robusta y representativa del conjunto de datos principal.

La inercia del modelo se sitúa en 20.190, un valor superior en comparación con la configuración previa (14.227), pero sigue siendo una cifra indicativa de una buena cohesión interna, asegurando que los datos de cada cluster estén bien agrupados en el espacio reducido de las componentes principales.

La nueva distribución de municipios en los clústeres obtenidos con el modelo refinado de K-means, utilizando dos componentes principales y excluyendo las anomalías, puede apreciarse en la tabla y mapa siguientes:

Figura 123

Distribución de municipios por clúster

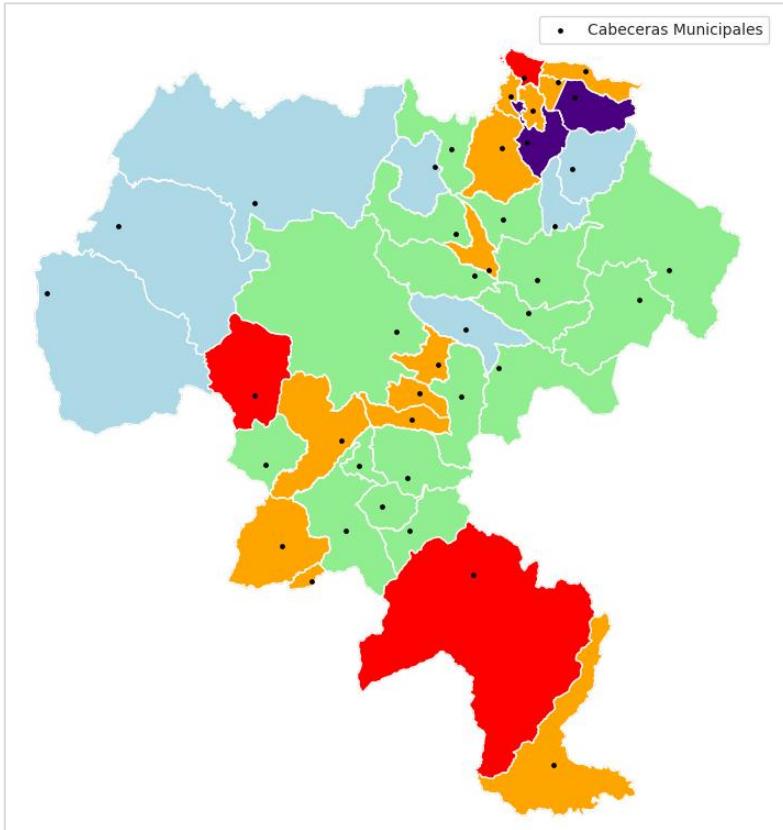
cluster_pca_final	municipios	numero_municipios
0	POPAYAN, GUAPI, JAMBALO, LOPEZ DE MICAY, SUAREZ, TIMBIQUI, TORIBIO	7
1	ALMAGUER, BALBOA, BOLIVAR, BUENOS AIRES, CAJIBIO, CALDONO, EL TAMBO, INZA, LA VEGA, MORALES, PAEZ, PURACE, SAN SEBASTIAN, SILVIA, SOTARA PAISPAMBA, SUCRE, TOTORO	17
2	CALOTO, CORINTO	2
3	FLORENCIA, GUACHENE, LA SIERRA, MERCADERES, MIRANDA, PADILLA, PATIA, PIAMONTE, PIENDAMO - TUNIA, ROSAS, SANTANDER DE QUILICHAO, TIMBIO, VILLA RICA	13
Anómalos	ARGELIA, PUERTO TEJADA, SANTA ROSA	3

Fuente: Pantallazo obtenido desde el entorno de Google Colab

En la configuración final de clustering, los municipios del departamento del Cauca se agruparon en cuatro clústeres principales y un clúster de anomalías. El clúster 0 incluye 7 municipios, entre los que se encuentra Popayán la capital. El clúster 1, con 17 municipios, como Almaguer y Buenos Aires, agrupa una variedad de localidades con características de violencia y seguridad similares, mientras que el clúster 2 está conformado por Caloto y Corinto. En el clúster 3 se encuentran 13 municipios, incluidos Florencia y Mercaderes. Por último, el grupo de **anómalos** contiene 3 municipios (Argelia, Puerto Tejada y Santa Rosa), caracterizados por patrones inusuales que los diferencian de los demás.

Figura 124

Distribución de los municipios del Cauca por clústeres modelo con exclusión de anomalías



Fuente: Pantallazo obtenido desde el entorno de Google Colab

En el modelo con exclusión de anomalías, los centroides de los clusters se expresan en función de dos componentes principales (PCA1 y PCA2), que capturan las dimensiones clave de la violencia y los factores socioeconómicos relacionados.

Figura 125

Matriz de centroides del nuevo modelo

	PCA_1	PCA_2
0	-0.431189	1.778474
1	-0.903236	-0.421277
2	2.638868	1.884456
3	1.007354	-0.696656

Fuente: Pantallazo obtenido desde el entorno de Google Colab

Las posiciones de los centroides en el espacio PCA permiten interpretar la estructura y el perfil de cada grupo en relación con estos factores:

- Clúster 0: (-0.431, 1.778). Este clúster se caracteriza por una posición levemente negativa en PCA1 y un valor alto y positivo en PCA2. Esto sugiere que los municipios agrupados en este clúster presentan un nivel moderado de criminalidad, con una victimización destacada, indicando una mayor vulnerabilidad de la población sin una alta correlación con actos delictivos intensos. Este perfil es indicativo de zonas afectadas principalmente por factores externos de violencia del conflicto armado. Por lo que se podrían llamar zonas de vulnerabilidad por conflicto.
- Clúster 1: (-0.903, -0.421). La posición negativa en ambas dimensiones (PCA1 y PCA2) implica un perfil de baja incidencia tanto en criminalidad como en victimización. Los municipios de este clúster muestran una relativa estabilidad, posiblemente caracterizados por una menor exposición a la violencia directa y a factores de victimización. Este clúster representa áreas con niveles bajos de conflicto y afectación por violencia. Por lo que se podrían caracterizar como áreas de baja violencia y estabilidad.
- Clúster 2: (2.639, 1.884). Con valores altos y positivos en ambas dimensiones, este clúster sobresale por su alta criminalidad y elevada victimización. Los municipios en este grupo probablemente enfrentan una combinación significativa de actividades delictivas y de impacto en la población civil, indicando áreas donde la violencia y las intervenciones son altamente visibles. Este perfil representa zonas de conflicto activo y de alto riesgo social, donde la respuesta estatal y el conflicto armado pueden estar interconectados. A estos municipios se puede catalogar como enclaves de narcotráfico, alta violencia y victimización.
- Clúster 3: (1.007, -0.697). Este clúster muestra un valor positivo en PCA1 y un valor moderadamente negativo en PCA2. Los municipios agrupados aquí presentan indicios de criminalidad relativamente elevada, pero una menor incidencia de victimización directa. Este perfil puede representar áreas en las que la actividad delictiva es visible, aunque la afectación sobre la población no es tan pronunciada, sugiriendo una menor exposición al impacto directo en los habitantes al conflicto armado. Estos municipios se pueden perfilar como regiones de criminalidad activa y baja victimización del conflicto interno.

A lo anterior se suma la situación de los tres municipios con patrones de violencia atípicos en el Cauca según el análisis mediante Isolation Forest: **Argelia** con altas tasas de víctimas, con un contexto de inseguridad severa y conflictos armados que impactan a la población. **Puerto Tejada** por una alta tasa de armas confiscadas y un índice elevado de criminalidad,

lo que sugiere un entorno de violencia urbana y circulación de armas, posiblemente asociado a redes de microtráfico. **Santa Rosa** con altas tasas de reclamantes de restitución de tierras y criminalidad, posiblemente reflejando conflictos agrarios y disputas territoriales.

El análisis descriptivo de las variables de violencia en la configuración refinada del modelo con exclusión de anomalías que se resume en la figura, revela distintos perfiles de violencia para cada clúster en el espacio de PCA, capturando diferencias significativas en la incidencia de criminalidad, victimización y otras variables relacionadas.

Figura 126

Estadísticas descriptivas por clúster con k-means refinado excluyendo anomalías

cluster_pca_exc_anom	Media, mediana y desviación estándar de las variables numéricas por cluster:																	
	armas_confis_polinal_tasa			indice_criminalidad_general_tasa			indice_narcotrafico_tasa			total_reclam_rest_tierr_minagr_tasa			vict_fuerza_pub_mindef_tasa			vict_por_declarac_uv_tasa		
	mean	median	std	mean	median	std	mean	median	std	mean	median	std	mean	median	std	mean	median	std
0	-0.323210	-0.450758	0.408183	-0.880171	-0.766216	0.311638	0.416275	-0.005639	0.955594	-0.398457	-0.378001	0.150278	-0.109861	-0.364524	0.618522	1.609211	1.648172	0.535676
1	-0.473061	-0.565235	0.282970	-0.689530	-0.728864	0.361774	-0.546497	-0.617136	0.189716	-0.066394	-0.195338	0.420371	-0.280842	-0.362945	0.268804	-0.600058	-0.830593	0.487247
2	-0.315800	-0.315800	0.220825	0.794278	0.794278	0.020163	3.359069	3.359069	1.737208	-0.394328	-0.394328	0.098039	0.473048	0.473048	0.595983	0.300809	0.300809	0.082589
3	0.531661	0.233760	0.721574	1.017400	0.900455	0.673401	0.029511	-0.054804	0.544406	-0.024700	-0.437217	1.012416	-0.085256	-0.224290	0.531289	-0.306212	-0.330510	0.513980
Anomalo	1.341502	-0.180710	3.131596	1.022816	1.479397	1.122285	-0.241755	-0.152546	0.404470	1.675881	0.670029	2.897990	1.901854	0.084236	3.315689	0.771883	0.642662	1.712801

Fuente: Pantallazo obtenido desde el entorno de Google Colab

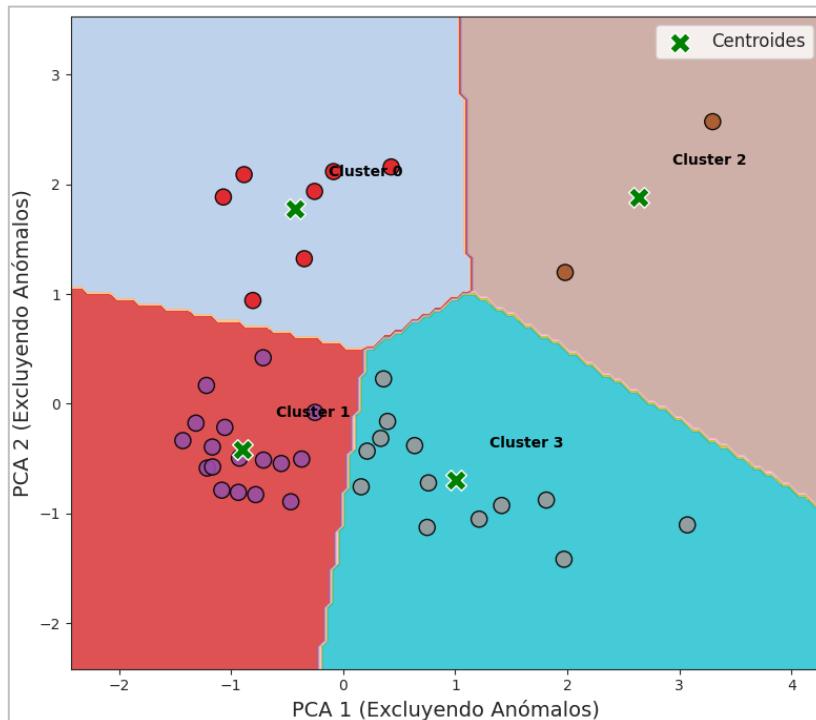
- Clúster 0: Este clúster contiene siete municipios con una media y mediana negativas en variables como indice_criminalidad_general_tasa y armas_confis_polinal_tasa, sugiriendo una menor incidencia de criminalidad y decomisos de armas. Sin embargo, destaca por un valor elevado en vict_por_declarac_uv_tasa (media de 1.61), lo que indica una mayor victimización asociada con el conflicto armado. La consistencia relativa de este tipo de victimización es evidenciada por una desviación estándar moderada (0.54).
- Clúster 1: Con 17 municipios, este clúster tiene valores negativos en la mayoría de las variables, reflejando niveles reducidos de violencia y victimización en general. La desviación estándar es baja en armas_confis_polinal_tasa y indice_criminalidad_general_tasa, sugiriendo una estabilidad en los perfiles de baja violencia de estos municipios. La variable vict_por_declarac_uv_tasa presenta una variabilidad ligeramente mayor, aunque sigue siendo consistente en su baja incidencia.
- Clúster 2: Con dos municipios (Caloto y Corinto), este clúster muestra valores notablemente altos en indice_narcotrafico_tasa (media de 3.36), lo cual sugiere una alta incidencia de actividades relacionadas con el narcotráfico en estas áreas. La desviación estándar de 1.74 en esta variable indica variabilidad entre los municipios del clúster, sugiriendo contextos distintos dentro del ámbito del narcotráfico.

- Clúster 3: Agrupando 13 municipios, este clúster se caracteriza por una media positiva en indice_criminalidad_general_tasa (1.01), lo que indica una criminalidad relativamente alta en comparación con otros grupos. La variabilidad es alta en total_reclam_rest_tierr_minagr_tasa, indicando diversidad en los contextos de reclamación de tierras. Este grupo representa un perfil intermedio de violencia, con ciertas variables mostrando una incidencia moderada.
- Grupo Anómalo: Este grupo contiene tres municipios (Argelia, Puerto Tejada, y Santa Rosa) con características individuales destacables. Sus valores extremos en variables como armas_confis_polinal_tasa y total_reclam_rest_tierr_minagr_tasa, junto con una alta desviación estándar, reflejan un contexto intenso de violencia y narcotráfico. Este grupo se analiza como una agrupación provisional, ya que cada municipio tiene un perfil único, lo que subraya sus diferencias con los otros clústeres.

La gráfica de teselación de Voronoi para los clusters obtenidos en el modelo K-means refinado en el espacio bidimensional de las primeras dos componentes principales (PCA_1 y PCA_2) excluyendo las anomalías se muestra en la figura.

Figura 127

Visualización de clústeres con límites de decisión de K-means excluyendo anomalías



Fuente: Pantallazo obtenido desde el entorno de Google Colab

En la gráfica se muestra de manera clara y diferenciada las áreas de influencia de cada cluster. Los límites de decisión del modelo delimitan regiones bien definidas, sin traslapos ni puntos ambiguos en los bordes. Cada municipio se encuentra sólidamente ubicado dentro del área correspondiente a su cluster asignado, lo que sugiere que los perfiles de violencia entre los municipios se agrupan de manera coherente y distintiva en el espacio PCA. Los centroides de cada cluster están claramente marcados y etiquetados, lo que facilita la identificación de las agrupaciones y refuerza la segmentación alcanzada en este modelo refinado.

En la figura se muestra un resumen de las tres configuraciones de k-means que se ha aplicado a las seis variables de violencia depuradas

Figura 128

Resumen de las configuraciones de k-means aplicadas a las variables de violencia

	Selección No. clústeres	Arbitraria - 3	Técnica del codo - 6	Técnica del codo - 4
Configuración PCA	NO	SI	SI	
Anomalías	Incluidas	Incluidas	Excluidas	
Clúster 0	9 mpios - Perfil de narcotráfico y victimización moderada	10 mpios - Moderada victimización y baja criminalidad	7 mpios - Zonas de Vulnerabilidad por conflicto	
Clúster 1	32 mpios - perfil de violencia estable	3 mpios - Alta criminalidad, confiscación armas, moderada victimización	17 mpios - Areas de baja violencia y estabilidad	
Clúster 2	1 mpio - Perfil de alta violencia y conflictividad	1 mpio - Extrema victimización y conflicto armado	2 mpios - Enclaves de narcotráfico, alta violencia y victimización	
Asignación clústeres	Clúster 3	---	13 mpios - Estabilidad y baja exposición a violencia	13 mpios - Regiones de criminalidad activa y conflicto agrario
	Clúster 4	---	5 mpios - Violencia mixta moderada	---
	Clúster 5	---	10 mpios - Criminalidad moderada con baja victimización directa	---
Evaluación	Cof. Silueta	28.6	49.9	52
	Inercia	169.68	14.27	20.19
				3 mpios anómalos

Fuente: Construcción propia

Análisis de la asociación entre perfiles de violencia y las variables socioeconómicas. Con las etiquetas de los clusters generadas a partir de la última configuración refinada de K-means, se procede a explorar la relación entre los perfiles de violencia identificados y las variables socioeconómicas de los municipios. Para este análisis, se emplea un modelo de regresión logística multinomial, utilizando el dataset de variables socioeconómicas como predictor. La idea central es investigar si los factores socioeconómicos pueden explicar, en parte, los distintos perfiles de violencia obtenidos mediante clustering.

Se espera que este enfoque híbrido, que combina técnicas de clustering no supervisado para etiquetar perfiles y análisis supervisado mediante regresión logística, permita identificar patrones y asociaciones significativas. Así, se podrá obtener una visión más clara sobre cómo los determinantes socioeconómicos podrían estar influyendo en los niveles de violencia. Para ejecutar este análisis, se siguen los pasos descritos a continuación:

- Se toman las 21 variables que están en el dataset ‘df_socieconomico_final’, producto del proceso de depuración descrito en este informe, las cuales incluyen aspectos como educación, empleo, condiciones de vivienda, entre otros, para ser empleadas como variables predictoras en el modelo.
- La variable de salida está conformada por las etiquetas de clusters obtenidas en la configuración final de K-means, almacenadas en la columna ‘cluster_pca_exc_anom’ de df_violencia_kmeans. Estas etiquetas representan diferentes perfiles de violencia en los municipios. Los municipios que fueron detectados como anomalías por Isolation Forest se consideran como si fuesen un grupo, ‘cluster de anomalos’.
- Para evitar que las diferencias de escala entre las variables afecten el modelo, se realiza una normalización o estandarización de las variables predictoras antes de alimentar el modelo. Esto asegura que cada variable tenga un impacto equivalente en la regresión logística.
- Con las variables predictoras y la variable de salida, se entrena un modelo de regresión logística multinomial. Pero dado que el propósito del análisis no es predictivo sino descriptivo y dado el tamaño reducido de la muestra (42 municipios), se entrena el modelo utilizando todos los datos disponibles sin dividirlos en conjuntos de entrenamiento y validación. Este enfoque, si bien es proclive a generar sobreajuste del modelo quitándole su capacidad de generalización, permite aprovechar la totalidad del dataset para examinar en detalle la relación entre las variables predictoras y los perfiles de violencia, mediante los coeficientes de regresión del modelo que nos brindan información sobre la significancia estadística de cada variable socioeconómica en los perfiles de violencia. Se espera que el modelo permita evaluar la asociación entre las características socioeconómicas y los perfiles de violencia, revelando qué factores socioeconómicos tienen una mayor influencia en la probabilidad de pertenencia de un municipio a cada perfil de violencia.
- A pesar del propósito exploratorio y descriptivo del análisis, se calculan las métricas como el accuracy, F1-score, precisión y recall para describir el ajuste del modelo. En este contexto, los indicadores de desempeño tienen un alcance limitado en cuanto a la generalización del modelo, ya que entrenar con todos los datos generalmente lleva a un ajuste excesivo.

El script de Python que se muestra en la figura de la siguiente página implementa los pasos descritos anteriormente.

Figura 129

Script para verificar relación entre variables socioeconómicas y perfiles de violencia en el Cauca, mediante regresión logística.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, accuracy_score

# Seleccionar y escalar las variables socioeconómicas
X = df_socieconomico_final[var_socioeconomicas]
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Definir las etiquetas de clusters como variable de salida
y = df_violencia_kmeans['cluster_pca_exc_anom']

# Crear y entrenar el modelo de regresión logística
model = LogisticRegression(multi_class='multinomial', max_iter=1000, random_state=42)
model.fit(X_scaled, y)

# Obtener y visualizar los coeficientes del modelo
coef_df = pd.DataFrame(model.coef_, columns=var_socioeconomicas, index=model.classes_)
coef_df.T.plot(kind='bar', figsize=(12, 8), title="Coeficientes de la Regresión Logística")

# Imprimir los coeficientes para interpretar
print("Coeficientes de cada variable en la predicción de perfiles de violencia:")
print(coef_df)

# Predecir los perfiles de violencia en los datos de entrenamiento
y_pred = model.predict(X_scaled)

# Evaluar el rendimiento del modelo
print("Evaluación del Modelo de Regresión Logística:")
print(classification_report(y, y_pred, target_names=[f"Cluster {c}" for c in model.classes_]))
print(f"Exactitud global del modelo: {accuracy_score(y, y_pred):.2f}")
```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

El código anterior arroja como primer resultado los coeficientes de la regresión logística. En modelos de regresión logística multinomial, estos coeficientes representan la magnitud y dirección del impacto que cada variable predictora ejerce sobre la probabilidad de clasificación en cada una de las categorías de la variable de respuesta. Estos coeficientes se interpretan en términos de “probabilidades logarítmicas relativas” (“log-odds”, “una forma de expresar probabilidades en el contexto de la regresión logística, utilizada para relacionar la probabilidad de que ocurra un evento respecto a la probabilidad de que no ocurra”), y cada valor positivo o negativo refleja si la variable aumenta o disminuye, la probabilidad de pertenencia a una categoría específica, manteniendo las demás variables constantes (Agresti, 2019).

La utilidad de estos coeficientes radica en su capacidad para ofrecer una interpretación detallada de cómo los cambios en las variables predictoras afectan las categorías de la variable dependiente. Los coeficientes positivos alejados de cero indican una relación directa fuerte, donde un aumento en la variable predictora incrementa la probabilidad de clasificación en una categoría específica. En contraste, los coeficientes negativos alejados de cero revelan una relación inversa, donde un aumento en la variable predictora reduce esta

probabilidad. Aquellos coeficientes cercanos a cero, ya sean positivos o negativos, sugieren una influencia baja o nula, indicando que esa variable no contribuye significativamente a la clasificación (Agresti, 2019).

En el contexto de nuestro análisis de la violencia en el Cauca, los coeficientes de la regresión logística multinomial permiten explorar la relación entre los perfiles de violencia, obtenidos mediante clustering, y las variables socioeconómicas de los municipios. A través de estos coeficientes, se identifica qué factores socioeconómicos incrementan o disminuyen la probabilidad de que un municipio pertenezca a un perfil de violencia determinado.

Es decir que los coeficientes son útiles para identificar tanto factores de riesgo como factores protectores asociados con la violencia en distintas zonas (Agresti, 2019). Las variables con coeficientes positivos alejados de cero indicarían características socioeconómicas que incrementan la predisposición de un municipio a ciertos tipos de violencia, mientras que las variables con coeficientes negativos relevantes se interpretarían como factores que reducen esta probabilidad. Aquellos coeficientes próximos a cero señalarían factores socioeconómicos cuya variación tiene poca o nula influencia en los niveles de violencia, ayudando a distinguir las variables realmente relevantes de las que no lo son en este contexto.

Figura 130

Coeficientes de regresión logística por variable y clúster

	poblacion	ha_under_ill	ha_condicion	ha_no_condic	desempleo_ce	tasa_alfab_c	porc_est_1_c	porc_est_2_c	cobertura_ne	desercion_mi	aprobacion_m
0	0.879040	0.516359	-0.272861	-0.031873	-0.095381	-0.675605	0.224859	-0.111014	-0.254645	0.393438	0.212366
1	0.122116	-0.534787	0.019132	0.355078	-0.022615	-0.423432	0.093188	-0.199177	-0.938839	-0.534308	-0.829476
2	-0.380700	-0.109155	0.313213	-0.295816	0.058438	-0.207926	-0.323959	0.043277	0.428819	-0.436773	0.309086
3	-0.237236	0.165633	-0.027535	0.221697	0.363708	1.173427	0.210810	-0.250629	0.090957	0.723644	-0.209806
Anomalo	-0.383220	-0.038050	-0.031949	-0.249087	-0.304150	0.133535	-0.204899	0.517543	0.673708	-0.146001	0.517831

	tasa_mort_ge	tasa_mort_ni	porc_bajo_pe	per_ipm_sisb	per_haci_cri	viviendas_si	atencion_inf	tasa_afiliac	indice_condi	ha_coca_cult
	-0.254598	-0.009787	-0.065491	-0.095379	0.454888	-0.685467	0.396503	-0.486185	0.244443	-0.052179
	0.390204	0.503769	-0.273222	0.034638	-0.513024	-0.025894	0.411073	0.205607	-1.091261	0.019493
	0.480471	-0.119566	-0.593218	-0.103109	-0.036390	0.232318	-0.324838	0.121808	0.130223	-0.349525
	-0.585017	-0.155029	0.287909	-0.107115	0.333309	0.392250	-0.254309	-0.137493	0.756089	-0.211703
	-0.031060	-0.219387	0.644022	0.270965	-0.238784	0.086793	-0.228429	0.296263	-0.039493	0.593914

Fuente: Pantallazo obtenido desde el entorno de Google Colab

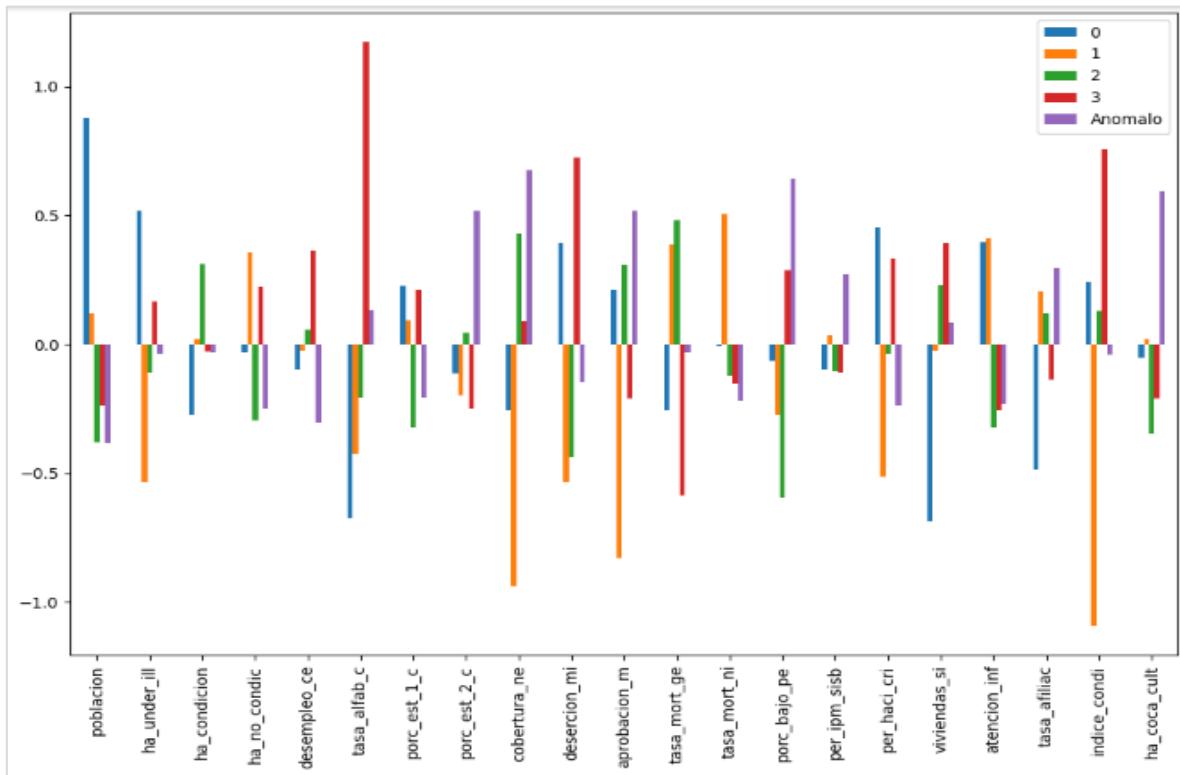
Cada coeficiente nos muestra cómo cambia la probabilidad de pertenencia a un cluster de violencia cuando “la variable aumenta en una unidad, manteniendo las demás constantes”. Por lo tanto un primer acercamiento a la interpretación de los coeficientes es identificar los coeficientes positivos alejados de cero (que indican una relación directa significativa entre la variable y la pertenencia a ese cluster), los coeficientes negativos alejados de cero (que señalan una relación inversa, es decir, estas variables reducen la probabilidad de pertenencia

al cluster específico) y los coeficientes cercanos a cero, ya sean positivos o negativos (que sugieren que la variable no tiene una influencia considerable en la clasificación del cluster) (Agresti, 2019).

Dada las escala y la variabilidad de las variables predictoras, en nuestro caso la mayoría de los coeficientes tomaron valores en el rango de -1 a 1, por lo tanto para facilitar la interpretación de los coeficientes y su relación con los perfiles de violencia en los municipios del Cauca, se definieron tres rangos de magnitud para reflejar la influencia de cada variable en la probabilidad de pertenencia a un clúster específico: a) Coeficientes positivos alejados de cero (mayores de 0.1), indican una relación directa entre la variable y la probabilidad de pertenencia al clúster, un aumento en el valor de la variable se asocia con una mayor probabilidad de que el municipio pertenezca a dicho perfil de violencia. b) Coeficientes negativos alejados de cero (menores de -0.1), reflejan una relación inversa importante entre la variable y el clúster, a medida que la variable aumenta, la probabilidad de pertenencia al perfil de violencia disminuye y c) Coeficientes cercanos a cero (de -0.1 a 0.1), que se interpretan como no significativos en cuanto a su influencia sobre la pertenencia al clúster, y por lo tanto cambios en estas variables no afectan sustancialmente la probabilidad de clasificación en un perfil de violencia específico.

Figura 131

Visualización de los coeficientes de la regresión logística por perfil de violencia



Fuente: Pantallazo obtenido desde el entorno de Google Colab

Para facilitar el análisis de los coeficientes la tabla a continuación sintetiza el análisis de los coeficientes obtenidos en la regresión logística, categorizando las variables en función de su impacto en la clasificación de los perfiles de violencia en municipios del Cauca. Esta organización facilita la interpretación de los factores determinantes para cada perfil de violencia y ayuda a identificar patrones específicos asociados a cada grupo.

Figura 132

Coeficientes de la regresión categorizando las variables en función de su impacto

Clúster	Perfil (nombre del clúster)	Variables con coeficientes positivos alejados de cero	Variables con coeficientes negativos alejados de cero	Variables con coeficientes cercanos a cero
0	Zonas de vulnerabilidad por conflicto	Número de habitantes: 0.87904	Porcentaje de población en estrato 2: -0.110104	Tasa de mortalidad infantil: -0.009787
		Hectáreas en explotación ilegal de metales preciosos: 0.516359	Tasa general de mortalidad: -0.254598	Hectáreas con vocación agrícola no condicionada: -0.031873
		Personas en hacinamiento critico segun el sisben: 0.454888	Cobertura neta de educación básica y media: -0.254645	Hectáreas de hoja de coca reportadas por el Ministerio de Justicia: -0.052179
		Número de niños atendidos por el ICBF mas niños en prescolar : 0.396503	Hectáreas con vocación agrícola condicionada: -0.272861	Porcentaje de niños con bajo peso al nacer: -0.065491
		Tasa de deserción escolar educación básica y media: 0.393438	Afilados a seguridad social en salud: -0.486185	Personas con indice de pobreza multidimensional: -0.095379
		Producto de la fusión de todas las condiciones de vivienda medidas en el dane: 0.244443	Tasa de alfabetas según censo 2018: -0.675605	Tasas de desempleo calculada del censo 2018: -0.095381
		Porcentaje de población en estrato 1: 0.224859	Número de viviendas en el Sisben: -0.685467	
		Tasa de aprobación educación básica y media: 0.212366		
1	Áreas de baja violencia y estabilidad	Tasa de mortalidad infantil: 0.503769	Porcentaje de población en estrato 2: -0.199177	Porcentaje de población en estrato 1: 0.093188
		Número de niños atendidos por el ICBF mas niños en prescolar : 0.411073	Porcentaje de niños con bajo peso al nacer: -0.273222	Personas con índice de pobreza multidimensional: 0.034638
		Tasa general de mortalidad: 0.390204	Tasa de alfabetas según censo 2018: -0.423432	Hectáreas de hoja de coca reportadas por el Ministerio de Justicia: 0.019493
		Hectáreas con vocación agrícola no condicionada: 0.355078	Personas en hacinamiento critico segun el sisben: -0.513024	Hectáreas con vocación agrícola condicionada: 0.019132
		Afilados a seguridad social en salud: 0.205607	Tasa de deserción escolar educación básica y media: -0.534308	Tasas de desempleo calculada del censo 2018: -0.022615
		Número de habitantes: 0.122116	Hectáreas en explotación ilegal de metales preciosos: -0.534787	Número de viviendas en el Sisben: -0.025894
			Tasa de aprobación educación básica y media: -0.829476	
			Cobertura neta de educación básica y media: -0.938839	
2	Enclaves de narcotráfico, alta violencia y victimización	Producto de la fusión de todas las condiciones de vivienda medidas en el dane: -1.091261		
		Tasa general de mortalidad: 0.480471	Personas con índice de pobreza multidimensional: -0.103109	Tasas de desempleo calculada del censo 2018: 0.058438
		Cobertura neta de educación básica y media: 0.428819	Hectáreas en explotación ilegal de metales preciosos: -0.109155	Porcentaje de población en estrato 2: 0.043277
		Hectáreas con vocación agrícola condicionada: 0.313213	Tasa de mortalidad infantil: -0.119566	Personas en hacinamiento critico segun el sisben: -0.03639
		Tasa de aprobación educación básica y media: 0.309086	Tasa de alfabetas según censo 2018: -0.207926	
		Número de viviendas en el Sisben: 0.232318	Hectáreas con vocación agrícola no condicionada: -0.295816	
		Producto de la fusión de todas las condiciones de vivienda medidas en el dane: 0.130223	Porcentaje de población en estrato 1: -0.323959	
		Afilados a seguridad social en salud: 0.121808	Número de niños atendidos por el ICBF mas niños en prescolar : -0.324838	
			Hectáreas de hoja de coca reportadas por el Ministerio de Justicia: -0.349525	
			Número de habitantes: -0.3807	
			Tasa de deserción escolar educación básica y media: -0.436773	

	Tasa de alfabetas según censo 2018: 1.173427	Personas con índice de pobreza multidimensional: -0.107115	Cobertura neta de educación básica y media: 0.090957
	Producto de la fusión de todas las condiciones de vivienda medidas en el dane: 0.756089	Afiliados a seguridad social en salud: -0.137493	Hectáreas con vocación agrícola condicionada: -0.027535
	Tasa de deserción escolar educación básica y media: 0.723644	Tasa de mortalidad infantil: -0.155029	
	Número de viviendas en el Sisben: 0.39225	Tasa de aprobación educación básica y media: -0.209806	
3	Regiones de criminalidad activa y conflicto agrario	Tasas de desempleo calculada del censo 2018: 0.363708	Hectáreas de hoja de coca reportadas por el Ministerio de Justicia: -0.211703
	Personas en hacinamiento critico segun el sisben: 0.333309	Número de habitantes: -0.237236	
	Porcentaje de niños con bajo peso al nacer: 0.287909	Porcentaje de población en estrato 2: -0.250629	
	Hectáreas con vocación agrícola no condicionada: 0.221697	Número de niños atendidos por el ICBF mas niños en prescolar : -0.254309	
	Porcentaje de población en estrato 1: 0.21081	Tasa general de mortalidad: -0.585017	
	Hectáreas en explotación ilegal de metales preciosos: 0.165633		

Fuente: Construcción propia

Teniendo en cuenta la información sobre los coeficientes de la regresión logística resumida anteriormente, junto con la caracterización de cada uno de los clusters obtenidos con la ultima configuración refinada de kmeans, el análisis de los componentes PCA empleados, la interpretación de la posición de los centroides y el análisis descriptivo de las variables de violencia dentro de cada cluster, a continuación, se realiza se identifican las variables socioeconómicas con influencia significativa en cada cluster.

Clúster 0: Zonas de vulnerabilidad por conflicto

Este clúster donde se agrupa los municipios con una victimización destacada asociada al conflicto armado y un nivel moderado de criminalidad, donde la violencia es principalmente un efecto externo del conflicto, incluye como variables socioeconómicas con coeficientes positivos alejados de cero el número de habitantes, las hectáreas en explotación ilegal de metales preciosos, las personas en hacinamiento crítico y la tasa de deserción escolar. Estos factores sugieren una relación directa con este perfil de violencia, indicando que aumentos en estas variables incrementan la probabilidad de pertenencia a este clúster y contribuyen a la vulnerabilidad del municipio. Por otro lado, variables como el porcentaje de población en estrato 2, afiliación a seguridad social en salud y la tasa de alfabetización presentan coeficientes negativos alejados de cero, lo cual indica que estos factores pueden actuar como protectores, atenuando la probabilidad de pertenencia a este clúster y a los problemas asociados al conflicto. Las variables cercanas a cero incluyen la tasa de mortalidad infantil, la población en estrato 1, las personas con índice de pobreza multidimensional, el número de viviendas en el SISBEN, las hectáreas de hoja de coca, entre otras, no presentan una influencia significativa en la probabilidad de pertenencia a este clúster.

Clúster 1: Areas de baja violencia y estabilidad

Este perfil representa municipios caracterizados por bajos niveles de criminalidad y victimización, son áreas con estabilidad y menores niveles de exposición a la violencia. Las variables con coeficientes positivos alejados de cero, como la tasa de mortalidad infantil, el número de niños atendidos por el ICBF y las hectáreas con vocación agrícola no condicionada, muestran una relación directa con el perfil, sugiriendo que estos factores contribuyen a la estabilidad del municipio. En contraste, variables como el porcentaje de población en estrato 2, la tasa de alfabetización y la tasa de deserción escolar tienen coeficientes negativos alejados de cero, indicando que estos factores podrían tener un papel protector contra la violencia en estas áreas. Las variables cercanas a cero incluyen el porcentaje de población en estrato 1 y el índice de pobreza multidimensional, que no tienen una influencia considerable en la clasificación dentro de este clúster.

Clúster 2: Enclaves de narcotráfico, alta violencia y victimización

Este clúster agrupa municipios con altos niveles de criminalidad y victimización. El análisis de los componentes principales sugiere que estos municipios enfrentan una combinación intensa de actividades delictivas y afectación a la población civil, siendo áreas donde el narcotráfico y el conflicto activo son altamente visibles. Las variables socioeconómicas como la tasa general de mortalidad, la cobertura neta de educación básica y media, y las hectáreas con vocación agrícola condicionada muestran coeficientes positivos alejados de cero, sugiriendo que estos factores están relacionados con la alta violencia y victimización en estas áreas. En contraste, variables como el índice de pobreza multidimensional, la tasa de alfabetización y el porcentaje de niños con bajo peso al nacer presentan coeficientes negativos alejados de cero, lo que implica que pueden reducir la probabilidad de que el municipio se clasifique en este perfil de alta violencia. Entre las variables cercanas a cero se encuentran el desempleo y el porcentaje de población en estrato 2, las cuales no aportan una influencia significativa en la clasificación dentro de este clúster.

Clúster 3: Regiones de criminalidad activa y conflicto agrario

Este clúster comprende municipios que muestran criminalidad elevada y niveles significativos en temas como la reclamación de tierras, pero con una menor incidencia de victimización directa, lo que sugiere una exposición moderada al impacto directo del conflicto armado. Entre las variables con coeficientes positivos alejados de cero, se destacan la tasa de alfabetización, el índice de condiciones de vivienda y la tasa de deserción escolar, indicando que estos factores están estrechamente vinculados con este perfil de violencia. Las variables con coeficientes negativos alejados de cero incluyen el índice de pobreza

multidimensional, la afiliación a seguridad social en salud y la tasa de aprobación en educación básica y media, lo que sugiere que estos factores podrían atenuar la intensidad del conflicto y la criminalidad en estos municipios. Las variables cercanas a cero, como la cobertura neta de educación y las hectáreas con vocación agrícola condicionada, muestran un impacto limitado en la probabilidad de clasificación dentro de este perfil.

Para complementar el análisis de los coeficientes de la regresión logística, se presentan a continuación el análisis de los odds ratios (razones de probabilidad) derivados de los coeficientes de la regresión logística realizada en cada clúster de violencia. Los odds ratios, calculados como la exponencial de cada coeficiente, indican la magnitud de cambio en las probabilidades de pertenencia a un clúster específico ante un incremento unitario en una variable socioeconómica. Un odds ratio superior a 1 sugiere una relación positiva entre la variable y la clasificación en el clúster, mientras que un valor inferior a 1 implica una relación inversa. En contraste, los valores cercanos a 1 reflejan un impacto limitado o nulo en la clasificación.

Figura 133

Razones de probabilidad de la regresión logística por variable y clúster

Odds Ratios:											
	poblacion	ha_under_i	ha_condici	ha_no_cond	desempleo_	tasa_alfab	porc_est_1	porc_est_2	cobertura_	desercion_	
Cluster 0	2.408586	1.675915	0.761199	0.968630	0.909026	0.508849	1.252147	0.894926	0.775192	1.482067	
Cluster 1	1.129886	0.585794	1.019316	1.426292	0.977639	0.654796	1.097668	0.819405	0.391082	0.586075	
Cluster 2	0.683383	0.896591	1.367812	0.743925	1.060179	0.812267	0.723280	1.044227	1.535443	0.646118	
Cluster 3	0.788805	1.180140	0.972840	1.248193	1.438655	3.233053	1.234678	0.778311	1.095221	2.061934	
Cluster Anómalo	0.681663	0.962665	0.968556	0.779512	0.737750	1.142862	0.814730	1.677901	1.961497	0.864157	
aprobacion tasa_mort_ tasa_mort_ porc_bajo_ per_ipm_si per_haci_c viviendas_ atencion_i tasa_afili indice_con ha_coca_cu											
	1.236601	0.775228	0.990261	0.936607	0.909029	1.575998	0.503855	1.486618	0.614968	1.276909	0.949159
	0.436278	1.477282	1.654947	0.760924	1.035244	0.598682	0.974439	1.508436	1.228270	0.335793	1.019684
	1.362179	1.616836	0.887305	0.552546	0.902029	0.964264	1.261520	0.722644	1.129537	1.139082	0.705023
	0.810741	0.557096	0.856390	1.333636	0.898422	1.395579	1.480308	0.775452	0.871541	2.129929	0.809205
	1.678383	0.969417	0.803011	1.904124	1.311229	0.787585	1.090671	0.795783	1.344824	0.961276	1.811064

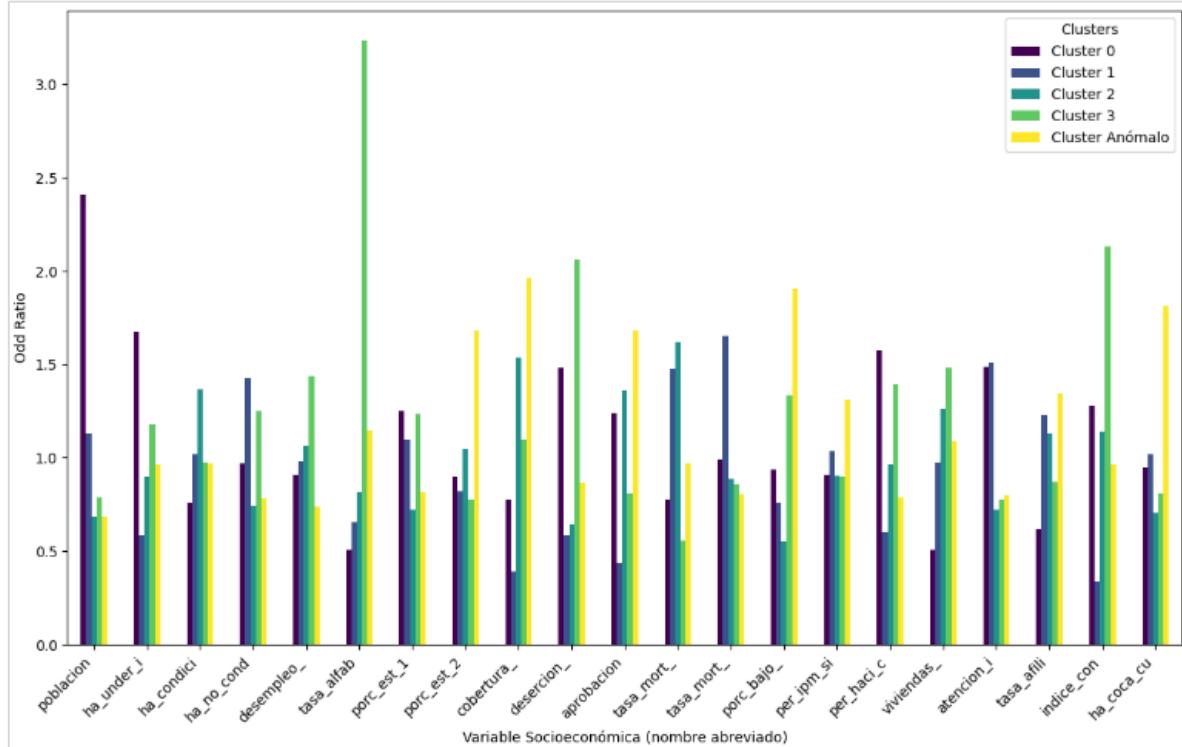
Fuente: Pantallazo obtenido desde el entorno de Google Colab

El análisis de los odds ratios confirma y complementa las relaciones ya observadas en los coeficientes de regresión logística, ilustrando la magnitud del cambio en la probabilidad de pertenencia a cada clúster ante un incremento en una variable específica. Ambos métodos, al

derivar del mismo modelo, reflejan relaciones consistentes entre las variables socioeconómicas y los perfiles de violencia, permitiendo una interpretación sólida de las influencias significativas y la dirección de cada variable.

Figura 134

Visualización de las razones de probabilidad de la regresión logística por perfil de cluster



Fuente: Pantallazo obtenido desde el entorno de Google Colab

En el Clúster 0, los odds ratios altos en variables como el número de habitantes y la tasa de deserción escolar, así como los valores menores a 1 en afiliación a salud y tasa de alfabetización, ratifican el perfil de vulnerabilidad en el que estos factores aumentan o reducen la probabilidad de pertenencia a este grupo. En el Clúster 2, donde se observa un perfil de alta criminalidad y victimización, los odds ratios respaldan el hallazgo de que factores como la cobertura educativa y mortalidad general incrementan la probabilidad de clasificación en este clúster, mientras que variables vinculadas a pobreza multidimensional y salud pueden atenuarla. Lo mismo sucede con el Clúster 3, donde los odds ratios confirman la relación que ya habían develado los coeficientes de la regresión logística.

En el análisis anterior, tanto los coeficientes de la regresión logística como los odds ratios deben interpretarse como patrones de asociación y no como relaciones de causa-efecto. Un coeficiente positivo o un odds ratio mayor a 1 sugiere que un aumento en esa variable está

asociado con una mayor probabilidad de pertenencia a un clúster de violencia específico, mientras que un coeficiente negativo o un odds ratio menor a 1 indica lo contrario. Sin embargo, esto no significa que estas variables sean factores causales o desencadenantes directos del perfil de violencia; simplemente reflejan asociaciones dentro del contexto particular del dataset analizado.

Estos resultados proporcionan información útil para identificar posibles áreas de intervención, como mejoras en infraestructura social o programas de seguridad. No obstante, es esencial tener cautela al interpretar estos patrones como objetivos de intervención aislados, ya que las variables analizadas pueden estar influenciadas por factores estructurales más amplios. Un enfoque de políticas públicas basado en estos datos debe ser integral, considerando múltiples factores socioeconómicos y evitando centrarse en ajustar individualmente una sola variable.

Los patrones de asociación encontrados ayudan a identificar factores que coexisten con ciertos tipos de violencia y vulnerabilidad en los municipios, ofreciendo pistas sobre áreas prioritarias para intervenciones. Sin embargo, estas asociaciones no implican que modificar una variable específica reducirá automáticamente los niveles de violencia. Es fundamental que las estrategias de intervención se diseñen considerando la complejidad del fenómeno y no se simplifiquen a ajustes directos de las variables predictoras. Interpretar los coeficientes y odds ratios con esta perspectiva evitará conclusiones incorrectas y permitirá un análisis más matizado y realista de los datos.

A continuación, se presenta la evaluación del modelo de regresión logística aplicado sobre el conjunto completo de datos, una decisión tomada de manera consciente con el fin de que el modelo aprenda de manera exhaustiva los patrones específicos de los municipios analizados. Dado que el objetivo de este estudio no es generalizar los resultados a otras muestras de municipios, sino capturar con precisión las características de violencia en este dataset específico, se emplearon todos los datos disponibles. Este enfoque permite que el modelo se ajuste de manera óptima a la totalidad del conjunto de datos, logrando una representación detallada y fiel de las asociaciones entre las variables socioeconómicas y los perfiles de violencia de cada clúster.

Figura 135*Evaluación del modelo de regresión logística*

```

from sklearn.metrics import classification_report, accuracy_score

# Predecir los perfiles de violencia en los datos de entrenamiento
y_pred = model.predict(X_scaled)

# Evaluar el rendimiento del modelo
print("Evaluación del Modelo de Regresión Logística:")
print(classification_report(y, y_pred, target_names=[f"Cluster {c}" for c in model.classes_]))
print(f"Exactitud global del modelo: {accuracy_score(y, y_pred):.2f}")

Evaluación del Modelo de Regresión Logística:
      precision    recall  f1-score   support

Cluster 0      1.00     0.86     0.92       7
Cluster 1      0.94     1.00     0.97      17
Cluster 2      1.00     1.00     1.00       2
Cluster 3      1.00     1.00     1.00      13
Cluster Anomalo 1.00     1.00     1.00       3

      accuracy          0.98      42
     macro avg      0.99     0.97     0.98      42
weighted avg     0.98     0.98     0.98      42

Exactitud global del modelo: 0.98

```

Fuente: Pantallazo obtenido desde el entorno de Google Colab

El análisis de los resultados de la evaluación del modelo de regresión logística confirma que el modelo ha aprendido de forma detallada las características del conjunto completo de municipios. La exactitud global del 98% indica que el modelo clasifica correctamente casi todos los municipios en su respectivo clúster. Dado que el propósito no es la generalización, sino un ajuste óptimo a estos datos, este alto desempeño es positivo y deseado, pues implica que el modelo refleja fielmente los patrones específicos del dataset.

Analizando los valores de precisión, recall y f1-score para cada clúster, se observa una precisión y recall perfectos (1.00) en la mayoría de los grupos, excepto en el Clúster 0, que presenta una leve reducción en el recall (0.86). Este valor implica que el modelo identifica correctamente la mayoría de los municipios en el Clúster 0, aunque no todos, resultando en un f1-score de 0.92, aún alto y adecuado para los fines de este análisis. Los valores perfectos de f1-score para los clústeres 2, 3 y el grupo anómalo indican un desempeño excelente en la identificación de municipios con perfiles de violencia específicos, tanto en términos de precisión (evitar falsos positivos) como de recall (evitar falsos negativos).

La media macro y la media ponderada muestran coherencia entre las clases y ratifican la estabilidad del modelo en términos de precisión y recall. En conclusión, el modelo se ha adaptado bien a los datos, ofreciendo una representación precisa de los perfiles de violencia para los 42 municipios del Cauca, lo que lo convierte en una herramienta sólida para análisis y estudios en este conjunto particular de datos.

Conclusiones

El repositorio virtual de datos construido representa una plataforma de datos bien estructurada y preparada para análisis avanzados. Este sistema no solo responde a las necesidades actuales del proyecto, sino que también está diseñado para adaptarse y ampliarse a medida que se disponga de nuevos datos, brindando una base sólida para futuras investigaciones y aplicaciones en el análisis de variables socioeconómicas y hechos de violencia en Colombia.

La creación del repositorio virtual de datos sobre variables socioeconómicas y hechos de violencia en los municipios colombianos destacan logros significativos en cuanto a cobertura, integridad y escalabilidad de la base de datos, logrando una estructura robusta para futuros análisis geoespaciales y temporales. La base de datos se caracteriza por:

- **Cobertura temporal y geográfica completa:** la base de datos relacional logró incorporar conjuntos de datos de diferentes entidades gubernamentales con rangos temporales amplios, en promedio mayor a 10 años hasta el 2024, cubriendo todos los departamentos y municipios de Colombia. Este alcance asegura que la base de datos ofrece un contexto histórico y actualizable que permite realizar análisis longitudinales y comparativos.
- **Integración geoespacial sólida:** La utilización de PostGIS permitió una integración precisa de datos geoespaciales, específicamente de los municipios colombianos, permitiendo que cualquier dato en el repositorio pueda asociarse a un lugar en el mapa. Esto hace posible no solo visualizar datos en mapas, sino también realizar análisis espaciales críticos para comprender las relaciones entre variables socioeconómicas y eventos de violencia en diferentes regiones.
- **Estandarización y consistencia de identificadores:** Se incorporó y estandarizó el código DANE en todas las tablas de la base de datos. En los casos donde este código no estaba presente, se implementó un proceso de coincidencia mediante similitud difusa (fuzzy matching), lo cual resultó efectivo para la asignación de códigos de manera precisa. Este paso fue clave para la integración exitosa y sin errores en el repositorio, ya que el código DANE actúa como un identificador único a nivel nacional, permitiendo una vinculación precisa de información entre diferentes fuentes de datos. Además, su implementación garantiza que los análisis y resultados puedan ser replicables y comparables a lo largo del tiempo y en otros estudios. Al contar con un sistema de identificación estandarizado, se facilita no solo el análisis de patrones en el contexto del Cauca, sino también la posibilidad de escalabilidad y compatibilidad de la base de datos

a nivel regional y nacional, lo cual es esencial para investigaciones futuras y procesos de toma de decisiones informadas.

- **Procesos ETL y control de calidad rigurosos:** Los procesos de Extracción, Transformación y Carga (ETL) para poblamiento de la base de datos se ejecutaron con controles estrictos que incluyeron validación de formatos y verificación de integridad de los datos post-carga, asegurando que los datos se mantuvieran sin duplicaciones ni errores. Además, se aplicaron imputaciones para valores faltantes y se eliminó cualquier duplicado, optimizando la consistencia y fiabilidad de la información en el repositorio.
- **Optimización de consultas y desempeño:** La creación manual de índices en llaves foráneas, además de los índices automáticos en las llaves primarias, mejora la eficiencia de las consultas, especialmente en aquellas que involucran grandes volúmenes de datos y múltiples tablas. Esto asegura que el sistema mantenga un desempeño ágil para consultas analíticas y operaciones ETL.
- **Documentación apropiada:** La creación de un diccionario de datos detallado garantiza la trazabilidad y claridad en el uso de cada campo y tabla en la base de datos, facilitando tanto la comprensión como el mantenimiento y expansión futura de la base.

En lo que respecta a la segunda parte del trabajo, la aplicación de técnicas de Machine Learning sobre una muestra de datos extraída de la base de datos sobre el departamento del Cauca, sin haber todavía terminado el informe, se puede concluir:

- **Eficiencia en el procesamiento de datos que permitió la conformación de un dataset completo:** La implementación del proceso ETL y el enfoque escalonado para la extracción de datos demostró ser una estrategia eficaz, permitiendo manejar un volumen masivo de información de manera ordenada y trazable. La automatización con Python y la integración en un DataFrame consolidado demostró la robustez y coherencia del repositorio de datos. Se logró implementar un exitoso proceso ETL que organizó datos dispersos de una base con más de 128 millones de registros en un formato estructurado y trazable, lo que permitió gestionar eficazmente la información y obtener un dataset con 143 variables de información socioeconómica y de violencia de los 42 municipios del Cauca para los años 2019-2023.
- **Segmentación de variables y transformaciones relevantes para análisis avanzados:** Las 143 variables fueron agrupadas en 3 categorías: 5 de información general y georreferenciación, 50 variables socioeconómicas y 88 variables sobre hechos de violencia. Con las variables socioeconómicas y de violencia se aplicó un robusto proceso

de reducción de dimensionalidad y de colinealidad aplicando diferentes técnicas desde la visualización de distribuciones, análisis de correlación, evaluación del factor de inflación de la varianza (VIF), análisis de información mutua, importancia por permutación, y Random Forest para la evaluación de la importancia de las características, lo que permitió finalmente obtener dos vistas minables, una con 20 variables socioeconómicas y otra con 7 variables sobre la violencia, manteniendo la riqueza de la información y proporcionando una base sólida para aplicar técnicas avanzadas de machine learning.

- **Aplicación del algoritmo de detección de anomalías sobre variables de violencia:** Con el algoritmo de Isolation Forest, aplicado sobre la vista minable de datos sobre la violencia, se detectaron municipios con características de violencia atípicas, destacando a Argelia, Puerto Tejada y Santa Rosa. Estos municipios presentaron valores elevados en variables específicas relacionadas con altas tasas de víctimas, un contexto de inseguridad severa y conflictos armados que impactan a la población (Argelia); una alta tasa de armas confiscadas y un índice elevado de criminalidad, lo que sugiere un entorno de violencia urbana y circulación de armas, posiblemente asociado a redes de microtráfico (Puerto Tejada) y altas tasas de reclamantes de restitución de tierras y criminalidad, posiblemente reflejando conflictos agrarios y disputas territoriales (Santa Rosa).
- **Perfiles de violencia identificados con K-means:** La configuración de K-means que ofreció el mejor desempeño permitió definir 4 clústeres que podemos vincular con perfiles específicos que diferencian los municipios del Cauca por su intensidad y tipo de violencia, así: zonas de vulnerabilidad por conflicto (7 municipios), áreas de baja violencia y estabilidad (17 municipios), enclaves de narcotráfico, alta violencia y victimización (2 municipios) y regiones de criminalidad activa y baja victimización del conflicto interno (13), más los tres municipios detectados como atípicos por Isolation forest.
- **Coeficiente de silueta de 0.52 en la mejor configuración de kmeans:** El coeficiente de silueta obtenido indica una separación moderada entre los clústeres y cohesión interna suficiente. Aunque no alcanza niveles excelentes (mayores a 0.7), este resultado es adecuado dado el carácter complejo y multidimensional de los datos de violencia. La clasificación obtenida es útil para interpretar tendencias generales, aunque se reconoce que existen casos limítrofes entre clústeres, lo que podría requerir más refinamiento en futuras etapas.

- **Uso de clústeres como etiquetas en regresión logística multinomial:** La utilización de los clústeres generados por K-Means como etiquetas en el modelo de regresión logística multinomial proporcionó una forma efectiva de relacionar las variables socioeconómicas con los perfiles de violencia. Esto permitió evaluar qué variables están más asociadas con cada perfil, destacando la influencia de condiciones socioeconómicas sobre la configuración de los patrones de violencia.
- **Sobreajuste intencionado en la regresión logística multinomial:** El modelo de regresión logística se entrenó deliberadamente para sobreajustarse, aprovechando todo el dataset con el objetivo de maximizar la identificación de relaciones entre variables específicas del de los municipios del departamento del Cauca, más que para evaluar su capacidad predictiva. Esta estrategia fue exitosa, permitiendo que el modelo "aprenda" de forma explícita las asociaciones subyacentes entre las variables socioeconómicas y los perfiles de violencia de los municipios del departamento del Cauca definidos por K-Means.
- **Relaciones de variables socioeconómicas con perfiles de violencia:** Los coeficientes de regresión logística y las razones de probabilidad obtenidos del modelo de regresión logística multinomial mostraron asociaciones clave entre las variables socioeconómicas y los perfiles de violencia que pueden utilizarse como insumos útiles para identificar posibles áreas de intervención de política pública. Pero hay que tener claro que son solo patrones de asociación y no relaciones de causa-efecto, su interpretación requiere de sumo cuidado y conocimiento de la realidad de la zona.

Referencias bibliográficas

- Aguirre Bonilla, D., Manquillo Gallego , N., Villota Cabrera , L., Correa Escobar , L., & Toro Cano , N. (9 de 8 de 2021). La imagen y la narrativa como herramientas para el abordaje psicosocial en escenarios de violencia–Departamentos de Cauca, Nariño, Quindío y Valle del Cauca. Bogota, Colombia: UNAD.
- Alvarado Mendoza, A., & Padilla Oñate, S. (2021). Organización policial y debilidad institucional: balance de las capacidades de las policías estatales. *Revista de ciencias sociales y humanidades*, 11-47.
- Arán Godés, P. (2022). *Métodos de aprendizaje automático y aplicaciones*. Zaragoza: Universidad de Zaragoza.
- Arzate Salgado, J. (2022). Herramientas para la comprensión sociológica del bienestar:: analítica de las formas de precariedad social y visibilización del continuo desigualdad (es) violencia (s). *Ánfora*, 29(53), 42-62., 42-62.
- Avalos, C., Castilla , F., & Colana, M. (2022). Trazabilidad de operaciones en base de datos para mitigar riesgos en los procesos de auditoría. *Innovación y Software*, 40-51.
- Barragán Vargas, D. (04 de 2024). Revisión del Estado del Arte. Bogotá, Colombia.
- Becker, D. (04 de 2024). *kaggle.com*. Obtenido de kaggle.com/learn: <https://www.kaggle.com/learn/machine-learning-explainability>
- Behar Villegas, E. (2021). Culturas del malgasto público: ineficiencia estatal y narrativas de política pública. *Revista de Administración Pública*, 662-678.
- Bilogur, A. (03 de 2024). *kaggle.com*. Obtenido de kaggle.com/learn: <https://www.kaggle.com/learn/pandas>
- Bobadilla, J. (2021). *Machine learning y deep learning: usando Python, Scikit y Keras*. . Bogotá: Ediciones de la U.
- Borjas García, J. E. (2020). Validez y confiabilidad en la recolección y análisis de datos bajo un enfoque cualitativo. *Trascender, contabilidad y gestión*, 79-97.
- Buitrago Gómez, A. F. (2022). *Los sistemas de información geográfica están transformando las tecnologías de la información y las comunicaciones*. Bogotá: Universidad Militar de Nueva Granada.
- Caio, R. (2022). Crisp-DM: las 6 etapas de la metodología del futuro. *Data Science e Analytics*.
- Camaro, A. (03 de 12 de 2022). *La Silla Vacía*. Obtenido de Dudas sobre la paz total: https://www.lasillavacia.com/opinion/dudas-sobre-la-paz-total-en_ca/

Centro Nacional de de Memoria Historica. (04 de 2024). *Observatorio de Memoria y Conflicto*. Obtenido de Centro Nacional de de Memoria Historica: <https://micrositios.centrodememoriahistorica.gov.co/observatorio/>

Chamorro, K., Laza, N., Noriega, H., & Rojano, R. V. (2021). Aplicación de Machine Learning para análisis de los fenómenos de violencia intrafamiliar en el departamento del Atlántico. *Investigación y desarrollo en TIC*, 12.

Clur, M. (2022). *Violencia Doméstica en Argentina: un Modelo de Evaluación de Riesgos Aplicando Técnicas de Machine Learning*. Buenos Aires: Itba.

Colab.Google. (n.d. de n.d de n.d.). *Colab.Google*. Obtenido de Colab.Google: <https://colab.google/>

Comision de la Verdad. (2024). *Hay futuro si hay verdad*. Obtenido de Hay futuro si hay verdad: <https://www.comisiondelaverdad.co/por-que-persiste-el-conflicto-en-el-cauca>

Congreso de la República de Colombia. (10 de junio de 2011). Por la cual se dictan medidas de atención, asistencia y reparación integral a las víctimas del conflicto armado interno y se dictan otras. *Ley 1448*. Bogotá DC: Imprenta nacional .

Cook, A., & Li, J. (04 de 2024). *kaggle.com*. Obtenido de kaggle.com/learn: <https://www.kaggle.com/learn/geospatial-analysis>

Corredor Rodríguez, S. (17 de 04 de 2023). Radiografía de la violencia en Cauca, donde ni el cese al fuego paró el conflicto. *El Espectador, Edicion digital*.

Cortés Landazury, R. (2024). Instituciones, etnicidad y conflicto: cohesión y fragmentación social en el departamento del Cauca (Colombia, 1990-2012). *Revista de Economía Institucional*, 175-202.

Creittiez, X. (2009). *Las formas de la violencia*. Buenos Aires: Waldhuter.

Dane. (2024). *Sistema Estadístico Nacional*. Obtenido de Sistema Estadístico Nacional: <https://www.dane.gov.co/index.php/sistema-estadistico-nacional-sen/planificacion-estadistica>

Departamento Nacional de Planeación -DNP. (2024). *Transparencia y acceso a la información pública*. Obtenido de Transparencia y acceso a la información pública: <https://www.dnp.gov.co/transparencia-y-acceso-informacion-publica>

Díaz Ramírez, J. (2021). Aprendizaje Automático y Aprendizaje Profundo. *Revista chilena de ingeniería*, 182.

Espinosa Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, investigación y tecnología*, , 21.

Espitia Cueca, C. E., & González Perafán, L. (2023). Economías de los conflictos armados en Colombia: acercamiento a la cadena de valor del narcotráfico. *Punto de encuentro No. 73 - Indepaz*, 4-39.

Fernández Caraballo, E., & Gómez Franco, Y. (09 de 2018). Metodología para el Análisis de la Violencia en el Departamento de Bolívar mediante Técnicas de Machine Learning. Cartagena, Bolívar, Colombia: Universidad Tecnologica.

Fontalvo Herrera , T. J., Vega Hernández , M. A., & Mejía , Z. F. (2023). Método de clústering e inteligencia artificial para clasificar y proyectar delitos violentos en Colombia. *Revista Científica General José María Córdova*, 551-572.

Gamez Brambila, M. (2020). La violencia en la historia. El papel de la memoria frente al trauma y la guerra. *Utopía y Praxis Latinoamericana*, 63-76.

García Pinzón, V., & Fernando Trejos, L. (2021). Las tramas del conflicto prolongado en la frontera colombo-venezolana: un análisis de las violencias y actores armados en el contexto del posacuerdo de paz. *Colombia Internacional*, 89-115.

García, M. (2023). *Revista 100 dias vistos por CINEP/PPP*.

GeoPandas. (n.d. de n.d. de n.d.). *Geopandas.org*. Obtenido de Geopandas.org: <https://geopandas.org/en/stable>

Géron, A. (2023). *Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-TensorFlow*. Sebastopol: O'Reilly Media.

González Perafán, L., Espitia Cueca, C. E., & Cabezas Palacios, J. V. (2023). *Cifras de la Violencia en Colombia*. Bogotá: Indepaz.

Greener, J. G., Kandathil, S. M., & Moffat, L. (2022). A guide to machine learning for biologists. *Nature reviews Molecular cell biology* - Vol. 23, 40-55.

Grupo de Redacción de Semana. (2023). Caucagiztán: este es el escalofriante panorama del Cauca, azotado por las Farc de 'Mordisco'. *Semana*.

Hincapié, L., & Rodríguez, C. (2021). Violencia basada de en género: conceptualización y análisis de su desarrollo en el conflicto colombiano. *Misión Jurídica*, 14-21.

Holbrook, R. (20 de 4 de 2024). *kaggle.com*. Obtenido de kaggle.com/learn: <https://www.kaggle.com/learn/feature-engineering>

Hotz, N. (28 de 04 de 2024). *datascience-pm.com*. Obtenido de Data science process alliance: <https://www.datascience-pm.com/crisp-dm-2/>

Hsu, L. S., & Obe, R. O. (2021). *PostGIS in Action*. New York: Manning Publications.

humanas.bogota.unal.edu.co. (2024). *Universidad Nacional de Colombia*. Obtenido de Universidad Nacional de Colombia: <https://www.humanas.unal.edu.co/observapazyconflicto/>

IBM Corporation. (17 de 8 de 2021). *ibm.com/docs/da/spss-modeler*. Obtenido de Deployment Overview: <https://www.ibm.com/docs/da/spss-modeler/saas?topic=deployment-overview>

- INDEPAZ. (2024). *Instituto de Estudios para el Desarrollo y la Paz*. Obtenido de INDEPAZ: <https://indepaz.org.co/>
- Inmon, W. H. (2022). *Building a data warehouse*. New York: John Wiley & Sons, Inc.
- Instituto Nacional de Medicina Legal y Ciencias y Forenses. (02 de 05 de 2024). *Observatorio de la Violencia*. Obtenido de <https://www.medicinallegal.gov.co/observatorio>
- Jiménez García, W. G., Manzano Chávez, L., & Mohor Bellalta, A. (2021). Medición de la vulnerabilidad social: propuesta de un índice para el estudio de barrios vulnerables a la violencia en América Latina. *Papers. Revista de Sociología*, 381-412.
- Johnson Criminal Law Group. (2024). *Johnson Criminal Law Group*. Obtenido de Johnson Criminal Law Group: <https://www.californiacriminaldefender.com>
- Lange, M. (2022). *Matar al otro: Una historia natural de la violencia étnica*. Fondo de Cultura Económica.
- Levalle , S. (2018). Resistencia a la violencia política y defensa de la territorialidad comunitaria en el departamento del Cauca, Colombia (1971-2012. *Sociedad y Economía* , 251-266.
- López Gonzalez, E. (2020). *Violencia de género en adolescentes*. Almeria.
- Marchetti, J. (2024). *Análisis de herramientas de ETL y reporting en entornos Big Data*. Buenos Aires: Universidad Nacional de La Plata.
- McKinney, W. (2022). *Python for Data Analysis_Data Wrangling with pandas, NumPy, and Jupyter-Wes McKinney*. Sebastopol: O'Reilly Media.
- Mejia, M. (23 de 12 de 2023). *Infobae*. Obtenido de Indepaz alerta sobre 11 actos de violencia en el Cauca en tan solo una semana de diciembre: <https://www.infobae.com/colombia/2023/12/23/indepaz-alerta-sobre-11-actos-de-violencia-en-el-cauca-en-tan-solo-una-semana-de-diciembre>
- MinTIC. (07 de 2022). *gobiernodigital.mintic.gov.co*. Obtenido de Guia de estandares de calidad e interoperatividad de Datos abiertos: https://gobiernodigital.mintic.gov.co/692/articles-179118_recuso_5.pdf
- Miranda, F., & Rodríguez, G. (2020). La inestabilidad social en El Catatumbo desde la óptica de la violencia estructural (2010-2018). *Revista Colombiana de Ciencias Sociales*, 562-585.
- Moreno, A., Armengol, E., Béjar, J., Belanche, L., Cortés, U., Ravaldá, R., . . . Sánchez, M. (1994). *Aprendizaje Automático* . Barcelona : Ediciones UPC.
- Nateras González, M. E. (2021). Aproximación teórica para entender la violencia desde un enfoque crítico. *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales*, 305-324.
- Neira, A. (26 de 04 de 2021). Una herida abierta que atraviesa el departamento. *El Tiempo*.

Norza Céspedes, E., Molano, A., Harker, A., & Buitrago Cubides, J. (2020). Trayectorias de la violencia homicida y desempeño estatal en Colombia. *Colombia Internacional*, 91-120.

Noticias RCN. (21 de 04 de 2024). *Violencia en el Cauca: ciudadanos temen por su vida ante el recrudecimiento del conflicto*. Obtenido de RCN Noticias: <https://www.noticiasrcn.com/colombia/violencia-en-el-cauca-ciudadanos-temen-ante-el-recrudecimiento-del-conflicto-684908>

Obando Guerrero, L., Pérez Caicedo, C., Cuastumal Meneses, R., & Hernández Narváez, E. (2020). La violencia urbana como fenómeno multicausal: un estudio en tres comunas de la ciudad de San Juan de Pasto. *Psicogente*, 102-120.

Oficina Asesora de Planeación de la Gobernación del Cauca. (2024). *Sistema de Información Socieconómica del Cauca - Tangara*. Obtenido de Sistema de Información Socieconómica del Cauca - Tangara: <https://tangara.gov.co/indicadores/>

Ordoñez , H., Cobos , C., & Bucheli , V. (2020). Modelo de machine learning para la predicción de las tendencias de hurto en Colombia. *Sistema e Tecnologías de información*, 494 - 506.

Ordoñez Eraso, H. A., Pardo Calvache, C. J., & Cobos Lozada, C. A. (2020). Detection of Homicide Trends in Colombia Using Machine Learning . *Revista Facultad de Ingenierí* , Volúmen 29, número 54.

Pilicita Garrido, A., & Borja López, Y. (2020). Rendimiento de MariaDB y PostgreSQL. *Revista Científica y Tecnológica UPSE (RCTU)*, 9-16.

Pinto Muñoz , C., Zuñiga Sambon, J., & Ordoñez Erazo , H. (2023). Machine Learning aplicado a la violencia de genero. *Revista Facultad de Ingeniería.*, 64.

Plataforma Nacional de Datos Abiertos de Colombia. (2024). *Datos Abiertos de Colombia*. Obtenido de Datos Abiertos de Colombia: <https://www.datos.gov.co/>

Pomares Quimbaya, A., & Martínez Bernal, O. (03 de 2023). Metodología CRISP-DM. Bogota: Universidad Javeriana.

Red de Derechos Humanos. (2019). *Impacto del conflicto social y armado 2018 –2019. Violaciones a los derechos humanos en el Departamento del Cauca*. Popayán.

Red de Derechos Humanos del Sur Occidente de Colombia “Francisco Isaías Cifuentes”. (2020). *Situación de emergencia por vulneraciones a los Derechos Humanos en el Departamento del Cauca*. Popayán.

Rojas Guerrero , M., & Grautoff, M. (2022). *Predicción de las masacres en Colombia empleando inteligencia artificial*. Bogotá DC.

Rojas, C., & Maldonado, E. (2023). *Civilización y violencia: la búsqueda de la identidad en el siglo XIX en Colombia*. Bogota: Pontificia Universidad Javeriana.

- Salazar Isairias, S. D. (23 de septiembre de 2024). *Predicción Geoespacial de Crímenes en Bogotá: Un Enfoque Basado en Machine Learning para Mejorar la Seguridad Ciudadana*. Obtenido de Universidad de los Andes : <https://repositorio.uniandes.edu.co/entities/publication/03397bcf-f2a5-477c-aebf-12d4747c1d3f>
- Sánchez Trujillo, M., & Pérez Hernández, J. (2021). *Metodología CRISP-DM en la gestión de proyecto de Data Mining. Caso enfermedades dermatológicas*. Bogota.
- Santana Falcón, F. (2023). *Análisis de datos abiertos de fútbol aplicando la metodología CRISP-DM*. Madrid: Universidad Politécnica.
- Santana, D., & Núñez, A. (2021). La vulneración de derechos, su incidencia en la salud mental de mujeres víctimas de violencia. *Sociedad & Tecnología*, 624-637.
- SISPRO. (2024). *Observatorio Nacional de Violencias de Género*. Obtenido de Observatorio Nacional de Violencias de Género: <https://www.sispro.gov.co/observatorios/onviolenciasgenero/Paginas/home.aspx>
- Spositto, O., Blanco, G., & Matteo, L. (2022). Técnicas de preprocesamiento de datos en modelos no supervisados aplicados al estudio genético de la raza Aberdeen Angus. *Revista digital del departamento de Ingeniería e investigaciones tecnológicas de la Universidad de Matanza*.
- Tarazona, A., & Ríos, A. (2021). Efectos de la inseguridad Ciudadana en el bienestar de la población. *Ciencia Latina Revista Científica Multidisciplinar*, 3341-3352.
- Tatman, R. (10 de 04 de 2024). *kaggle.com*. Obtenido de kaggle.com/learn: <https://www.kaggle.com/learn/data-cleaning>
- Teleencuestas. (2024). *Teleencuestas*. Obtenido de Teleencuestas: <https://telencuestas.com/censos-de-poblacion/colombia/2024>
- Tom, M. (1997). *Maching Learning*. New York : McGraw-Hill.
- Torres Erazo , J. M. (2023). *Razonpublica*. Obtenido de Razonpublica: <https://azonpublica.com/la-violencia-cauca-tras-acuerdo-paz/>
- Unidad de victimas. (29 de 02 de 2024). *Registro Unico de Victimas*. Obtenido de <https://datospaz.unidadvictimas.gov.co/registro-unico-de-victimas/>
- Universidad Externado de Colombia. (02 de 05 de 2024). *Centro de analisis de datos - Celfos*. Obtenido de Centro de analisis de datos - Celfos: <https://www.uexternado.edu.co/delfos-centro-analisis-datos/balance-de-la-violencia-letal-en-colombia-2020-2021-2/>
- Universidad Javeriana. (2024). *Instituto de Estudios Interculturales* . Obtenido de Instituto de Estudios Interculturales : <https://www.javerianacali.edu.co/intercultural>

Valencia Londoño, P. A. (2022). Retos del posconflicto frente a la respuesta humanitaria: entre la persistencia de las consecuencias humanitarias y las limitaciones al mandato de los actores humanitarios. *Opinion jurídica vol.19 no.38.*

Valencia Muñoz, M. L., García Álvarez, L., Ortega Solarte, N. I., & Díaz, D. F. (2021). *La imagen y la narrativa como herramientas para el abordaje psicosocial en escenarios de violencia. Departamentos de Cauca y Valle del Cauca.* Cali: Universidad Nacional Abierta y a Distancia - UNAD. Obtenido de Repositorio Institucional UNAD: <https://repository.unad.edu.co/handle/10596/42204>

Vega, A., Reynoso, A., & Corrales, C. (2022). Vega, A. A. Z., Reynoso, A. M., Corrales, C. R. F. Un análisis objetivo en los últimos 5 años de la violencia doméstica en el Perú - Una Revisión Sistemática. *Revista de derecho*, 99-109.

Velásquez, J. (2023). Evolución de las Preferencias de Ingreso Universitario, Un Enfoque de Minería de Datos con CRISP-DM. *Revista de Investigación*, 13.

Vidal, J. E., Mejía González, L., & Curiel Gómez, R. Y. (2021). La violencia como fenómeno social: Dimensiones filosóficas para su evaluación. *Revista de filosofía*, 38(99), 179-189., 38-99.

Vidal, J., Mejía, L., & Curiel, R. (2021). La violencia como fenómeno social: Dimensiones filosóficas para su evaluación. *Revista de filosofía*, 179-189.

Villar Roldán, J. J., & Martín Álvarez , J. M. (2023). Comprendiendo la dinámica de los conflictos en América latina: una aproximación desde el Machine learning. *Iberoamerican Business Journal*, 47-75.

Vizmanos, B., Mejía Pérez, J., & Fránquez Flores, Y. (2022). Bases para fundamentar un anteproyecto de investigación. . *Revista Educación y Desarrollo*, 81-96.

Yup de León, P. D. (2021). Análisis espacial de violencia homicida en la región norte de Centroamérica (2019-2020). *Revista Latinoamericana, Estudios de la Paz y el Conflicto*, 99-114.

Anexos