



**ANÁLISIS DE PATRONES DE VIOLENCIA ARMADA EN LA FRONTERA
DE COLOMBIA CON VENEZUELA USANDO ALGORITMOS DE
APRENDIZAJE AUTOMÁTICO**

Autor

Daniela Lopera Pai

Director

José Lisandro Aguilar Castro

UNIVERSIDAD EAFIT

Maestría en Ciencia de Datos y Analítica

Medellín

2025

Resumen

La frontera de Colombia con Venezuela se caracteriza por la falta de control estatal efectivo, la presencia de grupos armados, el tráfico de drogas y un constante flujo migratorio, lo que contribuye a crear un entorno marcado por la agudización de la violencia y la inestabilidad. En este sentido, este estudio tiene como objetivo analizar los patrones de violencia armada en esta región mediante el uso de algoritmos de aprendizaje automático y métodos de explicabilidad. Para ello se realizó la integración de fuentes de datos con variables relacionadas con la presencia de grupos armados, economías ilegales y proximidad a la frontera. El enfoque consistió en implementar diferentes algoritmos de agrupamiento para identificar patrones de violencia armada. Estas técnicas se combinaron con el algoritmo A priori para descubrir asociaciones entre las variables en los patrones identificados, y el método SHAP para evaluar la contribución de las variables en esos patrones. Los resultados obtenidos destacan que los municipios con mayores niveles de violencia tienden a estar ubicados en la frontera o a menos de 100 km de un paso fronterizo, además de tener presencia de cultivos ilícitos, actividades de erradicación y presencia de múltiples actores armados. Se obtuvieron reglas que identifican relaciones entre la ocurrencia de hechos victimizantes, la presencia de grupos armados como las AGC, ELN, EPL con la presencia de Disidencias de las FARC, la proximidad a la frontera y los cultivos ilícitos. Las variables más importantes según el método de explicabilidad fueron la presencia de grupos armados como el EPL, y las variables relacionadas con la presencia de cultivos ilícitos. Así, el enfoque metodológico permitió no solo identificar patrones complejos de violencia, sino también ofrecer insumos basados en datos para el diseño de políticas públicas que promuevan la seguridad y la estabilidad en esta región estratégica.

Palabras clave: Patrones de violencia, algoritmos de aprendizaje automático, frontera Colombo-venezolana, análisis de explicabilidad.

1. INTRODUCCIÓN

La frontera colombo-venezolana se extiende por aproximadamente 2.219 km, atravesando cuatro estados en Venezuela (Zulia, Táchira, Apure y Amazona) y siete departamentos en Colombia (Arauca, Boyacá, Cesar, Guainía, La Guajira, Norte de Santander y Vichada). Actualmente, se considera una de las zonas fronterizas más peligrosa del mundo debido a la confluencia de diferentes factores como la presencia de grupos armados, el tráfico de drogas, el contrabando, y un constante flujo migratorio causado por la inestabilidad política y económica en Venezuela [1], [2].

Además, la falta de control estatal efectivo ha permitido que los grupos armados ejerzan control en los territorios, perpetuando acciones criminales y violaciones de los derechos humanos, lo que contribuye a crear un entorno marcado por la agudización de la violencia y la inestabilidad

[2] [3]. Lo anterior se evidencia en la tendencia creciente del número de hechos victimizantes ocurridos en los departamentos que pertenecen a la frontera con Venezuela. Según datos del Registro Único de Víctimas (RUV), proporcionados por la Unidad para las Víctimas (UARIV), en el 2011 se registraron un total de 21.584 hechos victimizantes¹, mientras que en 2022 esta cifra ascendió a 79.782, lo que representa un incremento del 270%.

Las dinámicas de violencia en esta región representan un desafío significativo para la seguridad local y nacional debido a los múltiples factores que impactan en su incidencia [4], [5]. Por ello, resulta crucial el desarrollo de análisis basados en datos que ofrezcan una comprensión más profunda de los patrones del conflicto y sus causas subyacentes a nivel local y global. Si bien a nivel internacional se encuentran una gran variedad de estudios que implementan herramientas de aprendizaje automático para predicción y análisis de conflictos armados, para el caso colombiano, la gran mayoría de los estudios se han centrado en enfoques cualitativos y en áreas de las ciencias sociales [6], [7], [8], lo que puede dificultar el análisis de interacciones complejas entre múltiples variables que impactan el fenómeno de la violencia en el país.

En este sentido, el objetivo de esta investigación es profundizar en el análisis de los patrones de violencia armada, por medio de la implementación de enfoques de aprendizaje automático como lo son las técnicas de agrupamiento de datos (*clustering*) y las reglas de asociación, y los métodos de explicabilidad. A través del cruce de diferentes fuentes de información, se estudian las dinámicas de la violencia en los municipios de la frontera de Colombia con Venezuela, explorando diversas agrupaciones temporales, considerando la influencia de variables como la proximidad a la frontera, economías ilícitas y la presencia de grupos armados, entre otras. Mediante la integración de estos enfoques, se espera identificar patrones de agrupamiento y de interacción entre las variables que permitan contribuir al debate político y académico sobre el conflicto armado.

Los enfoques de aprendizaje automático son capaces de procesar grandes volúmenes de información y detectar patrones ocultos en conjuntos de datos complejos y de alta dimensionalidad. Así, mediante la implementación de técnicas de *clustering* es posible identificar patrones de comportamiento de la violencia, reconocer tendencias y similitudes, que podrían no ser evidentes por medio de una investigación cualitativa y técnicas tradicionales

¹ Según la UARIV, los hechos victimizantes son infracciones al DIH o violaciones graves y manifiestas a las normas internacionales de los Derechos Humanos ocurridas dentro del marco del artículo 3 de la Ley 1448 y los Decretos 4633, 4634 y 4635 de 2011.

como el análisis descriptivo y de correlación; lo que proporciona una visión más detallada del fenómeno de la violencia armada en la región. Por otro lado, la definición de *reglas de asociación* (usando técnicas como el algoritmo *a priori*) permite descubrir relaciones frecuentes entre las variables de un conjunto de datos, generando reglas que revelan cómo ciertos eventos tienden a ocurrir juntos. Finalmente, el uso de *métodos de explicabilidad* facilita la identificación de las características (variables) que tienen mayor impacto en los patrones de violencia. En conjunto, estos enfoques ofrecen un marco estructurado para analizar la interacción entre las variables relacionadas con el conflicto armado, proporcionando una visión más integral y detallada del fenómeno.

Las contribuciones de este trabajo son:

- Aplicar enfoques innovadores basados en aprendizaje automático que complementan los estudios cualitativos tradicionales sobre violencia.
- Identificar patrones de violencia armada en la frontera colombo-venezolana mediante técnicas de agrupamiento de datos.
- Descubrir relaciones frecuentes entre variables relacionadas con el conflicto armado utilizando reglas de asociación.
- Usar métodos de explicabilidad para determinar las variables con mayor impacto.
- Generar insumos para el diseño de políticas públicas enfocadas en la seguridad y estabilidad de la región fronteriza.

En conclusión, esta investigación contribuye al campo del análisis del conflicto armado al mostrar cómo las técnicas de aprendizaje automático pueden ser utilizadas de manera efectiva para identificar patrones complejos de violencia armada. Además, el enfoque adoptado permite un análisis granular a nivel de los municipios de frontera, aportando un mayor entendimiento de cómo factores como la presencia de grupos armados y de economías ilegales, y la distancia a la frontera, entre otros factores, influyen en la violencia en la región. Todo lo anterior enriquece el conocimiento teórico sobre el conflicto, y proporciona un marco analítico basado en datos que pueda ser usado por investigadores, formuladores de políticas públicas, y organizaciones interesadas en el análisis de las dinámicas internas de los conflictos armados.

A continuación, se presenta las diferentes secciones que componen el desarrollo del proyecto de investigación. En la segunda sección se exponen los trabajos relacionados donde se han usado técnicas de aprendizaje automático para el análisis de conflictos. En la tercera sección se desarrolla el marco teórico del presente proyecto, y en la cuarta sección se describe la

metodología usada en el desarrollo del proyecto. Finalmente, en la quinta y sexta sección se muestran los principales resultados obtenidos, y las conclusiones del trabajo.

2. ESTADO DEL ARTE

El análisis de patrones de violencia armada representa un desafío importante debido a su alta heterogeneidad espacial y temporal, la multiplicidad de factores involucrados, y las limitaciones en la medición de algunas características de manera confiable. A pesar de ello, el uso de herramientas de aprendizaje automático para entender y prevenir conflictos armados ha ganado relevancia en diversos campos, principalmente para el análisis de determinantes de la guerra o violencia armada [9], la identificación de patrones subyacentes [10], y la predicción de conflictos [11].

A nivel mundial se han desarrollado estudios que se enfocan en zonas con altos niveles de violencia, como son los países de África [11], [12], [13] y del sur de Asia [14], [15], pero en general, se han analizado datos de un conjunto amplio de países a nivel global [9], [16], [17]. Entre las técnicas empleadas en estos estudios se encuentran técnicas de regresión logística [16], [18], [19], el algoritmo *dynamic elastic net*² [13], [20], técnicas de *clustering* como *k-means* [10], *k-medoids* [21], y de agrupación basada en grafos [15], y modelos de ensamblaje como *random forest* [11], [18], *boosted regression trees* [9] y *gradient boosting* [11]. Estas técnicas han sido usadas principalmente para la predicción de conflictos armados a nivel global y regional [16], [17], [18], [19] y para el análisis de patrones [10], [21] y determinantes de la violencia [9].

Como resultado de los análisis realizados se encuentra que las técnicas de aprendizaje automático muestran una alta precisión para predecir conflictos [9], [11], [12], [19]; logrando resultados en métricas como el Área Bajo la Curva ROC (AUROC) de 0.919, prediciendo incidencia del conflicto a largo plazo [16] y de 0.8612 en *Mean Squared Error (MSE)* para predecir cambios de intensidad en conflictos mayores [16]. Además, se identificaron factores claves como las condiciones socioeconómicas [15], comportamientos políticos [13], [15], [17], variaciones climáticas [9] y localización geográfica [15], como predictores significativos o determinantes de la violencia.

² es una extensión del Elastic Net, un algoritmo que combina las propiedades de las penalizaciones Lasso (L1) y Ridge (L2).

De igual manera, en algunos estudios se implementaron técnicas avanzadas para la detección de patrones en conjuntos de datos complejos usando técnicas como la *convolutional long short-term memory (CLSTM)* [12], [14], las redes neuronales convolucionales (CNN) [22], y las redes neuronales basadas en grafos [14], [15]. Se encuentra que estas técnicas logran una mayor precisión a la hora de predecir conflictos que las técnicas tradicionales. En este sentido, las redes neuronales basadas en grafos logran un ajuste de R^2 de 93%, superando a métodos como *Support Vector Regression (SVR)*, *Gradient Boosting Regressor* y *K-Nearest Neighbors (KNN)*, en la predicción de fatalidades en conflictos armados [15]. De igual manera, el modelo *CLSTM* mostró una alta precisión para predecir la relación entre economía y conflictos armados, logrando en la predicción anual una precisión promedio de 92,4%. Por otro lado, el modelo *CNN* superó en precisión a un predictor espacial basado en proximidad geográfica de eventos previos al conflicto, con un desempeño promedio del 70% en precisión y un área bajo la curva superior al 75% en algunos períodos. Finalmente, los autores de [20] implementaron técnicas de procesamiento de lenguaje natural en conjunto con *Dynamic Elastic Net (DynENet)* para predecir el estallido de violencia armada utilizando un conjunto de datos de noticias de 170 países a nivel global, esto con el fin de predecir inicios repentinos de violencia en países que ha sido pacíficos, logrando una precisión de 80% con tasas de sensibilidad por encima del 50% en la predicción de conflictos armados a 12 meses.

Para el caso colombiano, la mayoría de los análisis sobre el conflicto armado se han centrado en enfoques cualitativos y en áreas de las ciencias sociales [6], [7], [8]. Entre los hallazgos de esos trabajos están que la economía política y el narcotráfico son temas recurrentes en los análisis de la violencia [6]; el conflicto armado es un fenómeno multicausal complejo que resulta de una combinación de factores económicos, sociales, políticos e institucionales [4]. Se destaca que los grupos armados crean instituciones locales para establecer control, y se resalta el papel de la desigualdad, la distribución de la tierra, la historia de la violencia y el narcotráfico, en la duración y la intensidad del conflicto [8].

También se encuentran estudios como los realizados por [23], [24], [25] y [26], donde se implementan técnicas de regresión para el análisis de determinantes y predicción de zonas críticas de violencia. Los resultados más relevantes de esos trabajos muestran que la expansión de la minería de oro, tanto ilegal como legal, estuvo asociada con aumentos significativos en violencia dirigida y desplazamientos forzados [23]. En [25] muestran que las regiones que experimentaron movilizaciones violentas previa tienen seis veces más probabilidades de experimentar una insurgencia en comparación con áreas sin antecedentes de acción colectiva;

y, en [26] se encuentra una asociación significativa entre la violencia en Colombia y las dinámicas de precios de los narcóticos en Estados Unidos.

Asimismo, [25], [27], [28] emplean técnicas de regresión espacial para capturar dependencias espaciales en los datos y analizar relaciones entre variables relacionadas con el conflicto. De esos trabajos se desprenden como resultados que los eventos de violencia ocurridos en un municipio aumentan la probabilidad de conflictos en municipios vecinos [27], y que los flujos de migración forzada derivados de la violencia armada tienen un impacto directo en el crecimiento urbano, particularmente en áreas periféricas y rurales [30]. Otros estudios emplean técnicas más avanzadas como técnicas de procesamiento del lenguaje natural [24], [29], modelos probabilísticos como *hidden markov models* [30], *random forest*, *gradient boosted*, redes neuronales y *ensemble bayesian model average* [24]. En esos trabajos se concluye que el *ensemble bayesian model average* mostró mejor desempeño para identificar áreas de alto riesgo de violencia, superando a cualquiera de los modelos individuales como *random forest* y *gradient boosted* [24]; que los modelos de procesamiento de lenguaje natural combinados con sistemas de información geográfica son eficientes para predecir zonas con presencia de grupos armados en Colombia [31]; y que, durante los conflictos con las Fuerzas Armadas Revolucionarias de Colombia (FARC), se dieron variaciones espaciotemporales significativas en el control territorial especialmente en la frontera con Venezuela y en la región suroeste [30].

Los resultados derivados de la aplicación de estas técnicas de aprendizaje automático proporcionan una comprensión más profunda de las dinámicas de la violencia armada, y los hallazgos pueden ser útiles para diseñar estrategias de prevención de conflictos y para el desarrollo de políticas de seguridad más efectivas [12], [31]. Por ejemplo, el uso de algoritmos de predicción facilita la implementación de sistemas de alerta temprana [12], [29], que anticipen escaladas de violencia y permitan una respuesta preventiva más ágil. De esta manera, las herramientas de aprendizaje automático no solo mejoran la comprensión del fenómeno del conflicto armado, sino que también sirven como guías estratégicas para prevenir futuros conflictos y el desarrollo en la región.

A pesar del potencial que tienen las técnicas de aprendizaje automático para apoyar el análisis de las dinámicas del conflicto, existen desafíos que limitan su aplicación. Uno de los principales problemas es la disponibilidad y calidad de los datos usados en los estudios. De igual manera, los modelos predictivos a menudo dependen de datos históricos que no se encuentran disponibles. Finalmente, el uso de datos regionales específicos limita la generalización de muchos estudios, restringiendo su aplicabilidad en contextos globales; y, la falta de

interpretabilidad de muchos de los algoritmos usados dificulta su adopción para el desarrollo de políticas públicas.

En este contexto, el desarrollo de este proyecto amplía y enriquece la literatura existente al integrar técnicas avanzadas de aprendizaje automático para analizar patrones complejos de violencia armada, específicamente en municipios fronterizos, un área que ha sido subexplorada en el caso colombiano. Este enfoque permite no solo una mayor comprensión de las dinámicas del conflicto, sino que también es una contribución metodológica significativa al aplicar herramientas de explicabilidad que faciliten la interpretación de los patrones de la violencia identificados. Esto complementa la literatura relacionada con el análisis de conflictos al demostrar que las herramientas de aprendizaje automático pueden no solo mejorar la precisión en la detección de patrones de violencia armada, sino también fomentar un análisis más transparente y accesible para los formuladores de políticas y actores sociales.

Finalmente, el desarrollo de este proyecto adopta un enfoque innovador que integra diferentes algoritmos de aprendizaje automático, como lo son las técnicas de agrupación y las técnicas de definición de reglas de asociación, y métodos de explicabilidad, para analizar patrones de violencia. Esta combinación metodológica no solo permite abordar la complejidad del fenómeno, sino que también establece un precedente para futuros estudios que busquen equilibrar la precisión en los resultados con una interpretación clara y accesible.

3. MARCO TEÓRICO

3.1 Contexto político de la frontera de Colombia con Venezuela

Históricamente, las relaciones diplomáticas entre Colombia y Venezuela han estado marcadas por constantes disputas. Según [32], desde la disolución de la Gran Colombia en 1830 la relación entre los dos países entro en conflicto debido a disputas territoriales no resueltas, problemas de deuda, y contrabando en la frontera. Desde la década de los 90, y particularmente tras la llegada de Hugo Chávez al poder, aumentaron las fricciones diplomáticas tras el reconocimiento de las FARC como un actor político legítimo por parte del gobierno de Venezuela [33]. Posteriormente, para el año 2015, el gobierno venezolano bajo el pretexto de seguridad nacional cerró la frontera [1], implementando una política de securitización que se enfoca en la militarización y control rígido de los flujos fronterizos [34], [35], lo que ha llevado a cierres intermitentes de los pasos fronterizos.

Asimismo, las zonas de frontera se han visto caracterizadas por la falta de presencia estatal efectiva, lo cual ha permitido que los grupos armados consoliden su poder en la región [1], [36]. En consecuencia, diversos grupos armados como el Ejército de Liberación Nacional (ELN), Disidencias de las FARC, el Ejército Popular de Liberación (EPL), las Autodefensas Gaitanistas de Colombia (AGC), los Grupos Armados Organizados (GAO) y Bandas Criminales se disputan el control del territorio, generando un entorno marcado por la violencia [36]. Por otro lado, desde la salida de Venezuela de la Comunidad Andina (CAN) en 2006, y el posterior colapso económico, la migración de venezolanos hacia Colombia se ha intensificado. El incremento migratorio ha generado importantes desafíos en las zonas fronterizas, principalmente en materia de seguridad, empleo y acceso a servicios básicos [32].

Las constantes fluctuaciones en las relaciones diplomáticas de los dos países debido a las diferencias en los modelos de gobernanza se han visto reflejadas en los cierres recurrentes de los pasos fronterizos y la interrupción del comercio bilateral. Eso ha impactado directamente en las condiciones económicas de las comunidades que viven en la frontera, quienes se han visto obligadas a recurrir a prácticas ilegales para garantizar sus medios de subsistencia [38], [37], [38]. Esto, sumado a la compleja situación migratoria y a la falta de control estatal, ha ayudado a intensificar la violencia y la inseguridad, ya que los grupos armados han aprovechado la crisis para ampliar sus fuentes de financiamiento a través de actividades ilícitas como tráfico de drogas, el contrabando de combustible, y el cobro de peajes a los migrantes a través de rutas ilegales [8], [37], [38].

La debilidad estatal, la presencia de grupos armados, el flujo migratorio constante, y las oportunidades ilícitas se combinan para transformar a la frontera en una zona de alta inseguridad. La cercanía a la frontera intensifica estos problemas, ya que facilita el tráfico ilegal y otorga a los grupos armados un mayor grado de impunidad, un fenómeno que [37] denomina “efecto frontera”. Este efecto explica cómo las características geográficas y sociales en las zonas fronterizas intensifican la inseguridad y se convierten en puntos críticos que facilitan las operaciones de grupos armados en áreas ya vulnerables con ausencia del estado.

En conclusión, la frontera colombo-venezolana es una zona de intensa disputa, donde convergen guerrillas, bandas criminales, economías ilegales y un constante flujo migratorio. La falta de un control estatal efectivo en ambos países ha facilitado que los grupos armados consoliden su influencia, deteriorando la seguridad y la estabilidad de las comunidades que habitan en la región. Las políticas de seguridad adoptadas por ambos gobiernos no han logrado solucionar las problemáticas subyacentes, y, en muchos casos, han exacerbado la situación.

Además, la interacción entre los modelos de gobernanza divergentes de Colombia y Venezuela, junto con su historial de tensiones diplomáticas, ha resultado en una gestión fronteriza carente de un enfoque coordinado y sostenible.

3.2 Técnicas de agrupamiento

Un algoritmo de agrupamiento se describe como una serie de procedimientos orientados a dividir un conjunto de datos en múltiples grupos con características en común [39]. Su objetivo es reunir patrones similares dentro de un mismo grupo, asegurando al mismo tiempo que los patrones ubicados en diferentes grupos sean lo más disimiles posibles; esta agrupación se fundamenta en una métrica de similitud o proximidad [40]. Las técnicas usadas en este trabajo son las siguientes: dos métodos basados en distancia, k-means y agrupamiento jerárquico aglomerativo; un método basado en densidad, *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN); y un método basado en grafos, *clustering espectral*. A continuación, se describe cada una de las técnicas utilizadas.

3.2.1 Algoritmo k-means: es un método de agrupamiento basado en partición cuyo objetivo es dividir un conjunto de datos en k grupos, donde cada grupo está representado por un centroide [40]. Su objetivo es minimizar la suma de las distancias al cuadrado entre los puntos de datos y el centroide del grupo al que pertenecen. El algoritmo comienza con los centros de los grupos colocados en posiciones arbitrarias, y luego los ajusta de manera iterativa para reducir el error de agrupamiento. Este proceso continua hasta que no hay cambios en las asignaciones o se alcanza un criterio de convergencia, como que el cambio en los centroides es menor que un valor ϵ o se alcanza un número máximo de iteraciones [41].

El algoritmo es sensible a la elección del número de grupos, el cual, puede ser subjetivo y suele requerir experimentación. Asimismo, es sensible a la presencia de datos atípicos, y funciona mejor cuando los grupos tienen forma circular y tamaños similares, ya que utiliza la distancia euclidiana para determinar la similitud en el conjunto de puntos de datos [41].

3.2.2 Agrupamiento jerárquico aglomerativo: es una técnica no supervisada que organiza las observaciones en una jerarquía o árbol conocido como dendograma, según la similitud entre ellos [42]. A diferencia de otros métodos de agrupamiento, como k-means, no se requiere especificar de antemano el número de clústeres [43]. Para determinar la similitud entre los grupos se utilizan diferentes métricas de distancia, tales como: distancia euclídea, distancia Manhattan, distancia de Mahalanobis, y distancia coseno, entre otras.

El algoritmo parte considerando cada observación como un grupo individual, y en cada paso combina los dos grupos más similares hasta que todos los puntos se agrupan en un clúster. Utiliza métodos de vinculación para combinar los grupos [43]. A continuación, se describen los principales métodos de vinculación:

- **Vinculación simple:** usa la menor distancia entre los elementos de los dos grupos.
- **Vinculación completa:** usa la mayor distancia entre los elementos de los dos grupos.
- **Vinculación promedio:** usa la distancia promedio entre todos los pares de elementos de los grupos.

3.2.3 DBSCAN: es un algoritmo de agrupamiento que identifica grupos de puntos en un conjunto de datos según su densidad en el espacio. A diferencia de otros métodos como k-means, no requiere que el número de grupos sea especificado de antemano y es capaz de detectar formas arbitrarias de clústeres, además de identificar datos atípicos [44]. Algunos conceptos claves en el algoritmo DBSCAN son:

- **Punto central:** un punto se considera central si hay al menos un número mínimo de puntos dentro de su vecindario definidos por un radio ϵ . La efectividad del algoritmo DBSCAN depende de la correcta selección de parámetros de puntos mínimos y de ϵ [45].
- **Punto de frontera:** son puntos que pertenecen a un grupo, pero no tiene suficientes vecinos para ser considerados puntos centrales [46].
- **Puntos de ruido:** son puntos que no pertenecen a ningún grupo, ya que no son accesibles por densidad a partir de ningún punto central [46].

3.2.3 Agrupamiento espectral: es un método de agrupamiento que utiliza la teoría espectral de grafos para identificar grupos en un conjunto de datos. El *clustering espectral* no impone restricciones en la forma de los grupos, lo que lo hace ideal para detectar grupos con formas complejas y no lineales, especialmente en conjuntos no convexos. Trata el problema de la agrupación como una partición de un grafo no dirigido, donde cada punto es un vértice y los valores de similitud entre puntos se convierten en los pesos de las aristas [47].

Su objetivo es dividir el grafo en componentes mediante el uso de métodos de corte de grafos, de manera que las aristas entre los distintos grupos tengan bajo peso y las aristas dentro de un mismo grupo tengan alto peso [48]. La similitud entre los diferentes puntos se halla en base a la matriz de similitud, la cual captura la estructura intrínseca de los datos, a partir de la matriz de similitud se deriva la matriz Laplaciana, la cual se utiliza para el desglose espectral y la selección de vectores propios, que ayudan a identificar los grupos [49].

3.3 Algoritmo A priori

Es un método simple y ampliamente utilizado en minería de datos para extraer conjuntos de elementos frecuentes de conjuntos de datos y generar reglas de asociación que ayuden a descubrir patrones ocultos [50]. El algoritmo sigue un enfoque iterativo y utiliza la propiedad antimonotónica, la cual indica que, si un conjunto de elementos no es frecuente, entonces ninguno de sus superconjuntos puede ser frecuente [51]. Algunos conceptos claves en el algoritmo a priori son:

- **Conjuntos de elementos frecuentes:** son conjuntos de ítems que aparecen juntos en un número significativo de transacciones en la base de datos. Se consideran frecuentes si cumplen con un umbral de soporte mínimo.
- **Reglas de asociación:** son implicaciones del tipo “Si X ocurre, entonces Y ocurre”. Estas reglas se generan a partir de los conjuntos frecuentes.

3.4 Método SHAP (SHapley Additive exPlanations)

Este método es basado en la teoría de juegos coalicionales, específicamente en los valores de Shapley [52]. Se utilizan para explicar modelos de aprendizaje automático de manera interpretativa, lo que es esencialmente útil para solucionar el problema de “caja negra” que presentan algunos algoritmos de aprendizaje automático [52]. El objetivo del método SHAP es explicar la predicción de una instancia al calcular la contribución de cada característica. La técnica asigna un valor a cada característica en función de su contribución a la predicción realizada por un modelo, considerando como cada característica contribuye al resultado final [53]. Los valores de Shapley se basan en la idea de “justa distribución” de la contribución de cada variable, y se calculan considerando todas las posibles combinaciones de variables y midiendo cómo cambian las predicciones del modelo cuando se agregan diferentes variables a la combinación [54].

4. METODOLOGÍA EXPERIMENTAL

La metodología usada en el desarrollo de este proyecto de investigación se basa en CRISP-DM (Cross Industry Standard Process for Data Mining), el cual divide el proceso de desarrollo de proyectos de minería de datos en varias fases interrelacionadas. A continuación, se detalla cada una de las fases implementadas, y se describe en qué sección del documento se desarrolla cada una:

1. **Comprensión del problema:** en esta fase se definieron los objetivos y se estableció el problema de investigación, identificando la relevancia de estudiar los patrones de violencia armada en los municipios de la frontera entre Colombia y Venezuela. Lo anterior se detalla en la Sección 1.
2. **Comprensión de los datos:** en esta etapa se realizó un análisis exploratorio de los datos para evaluar la calidad de los datos disponibles, detectar inconsistencias, y obtener una visión preliminar de los patrones presentes en la información. Esta fase se detalla en la Sección 4.1, donde se describe el conjunto de datos, y en la Sección 4.2, donde se realiza el análisis exploratorio del comportamiento de los datos.
3. **Preparación de los datos:** en esta fase se efectuó la limpieza, transformación y normalización de los datos, además, se seleccionaron las variables más relevantes para el análisis. El proceso realizado se describe en la Sección 4.2.
4. **Modelado:** para abordar el problema de investigación planteado en este documento se aplicaron varios enfoques de aprendizaje automático. Particularmente, se usaron técnicas de *aprendizaje no supervisadas*, tales como k-means, agrupamiento jerárquico, agrupamiento espectral y DBSCAN. Se seleccionaron estas cuatro técnicas en la búsqueda de identificar cuál tenía el mejor ajuste a la hora de identificar patrones de violencia armada en el conjunto de datos de frontera. El proceso llevado a cabo para implementar dichas técnicas se describe en la Sección 5.1. Adicionalmente, se usó el algoritmo a priori para la identificación de *reglas de asociación*, esto con la finalidad de descubrir relaciones latentes y patrones recurrentes entre las variables del conjunto de datos. Esta técnica permite identificar vínculos significativos que, en combinación con los análisis de agrupamiento, enriquecen la interpretación de los comportamientos subyacentes en los municipios fronterizos. Esta implementación se detalla a profundidad en la Sección 6.2. Finalmente, se implementó un *método de explicabilidad* basado en valores SHAP, lo que permitió cuantificar y visualizar la contribución de cada variable en la formación de grupos. Este enfoque facilitó la identificación de los factores más influyentes en el análisis de violencia, proporcionando una interpretación transparente de los resultados obtenidos a partir de las técnicas de agrupación, aportando una capa de análisis adicional que robustece la interpretación de los resultados obtenidos. Este proceso se detalla a profundidad en la Sección 6.3. El proceso de modelado se llevó a cabo utilizando el lenguaje de programación Python, mediante el uso de bibliotecas especializadas como scikit-learn.

5. **Evaluación:** esta etapa se centró en validar y contrastar los resultados obtenidos para cada una de las técnicas implementadas. Las métricas usadas para evaluar los métodos de agrupamiento y las diferentes reglas de asociación se describen en la Sección 4.3. Los resultados obtenidos para las métricas de evaluación de las técnicas de agrupamiento se detallan de la Sección 5.3 a la Sección 5.5; y los valores de las métricas para las reglas de asociación se describen en la Sección 6.2.

4.1 Conjunto de datos

La selección inicial de variables se llevó a cabo en función de la formulación de algunas hipótesis relacionadas con las dinámicas de la violencia, tales como:

- **H1:** Los municipios más cercanos a la frontera internacional tienden a presentar mayor incidencia de la violencia [36].
- **H2:** Hay una concurrencia entre municipios con violencia más crítica y la expansión de economías ilícitas como los cultivos de coca [55], [56].
- **H3:** La presencia de grupos armados incentiva la violencia, es decir, en los municipios con un mayor número de grupos armados activos hay un nivel mayor de violencia.

En la búsqueda de validar la hipótesis anteriormente descritas, se recolectaron datos disponibles de diferentes fuentes de información para los 222 municipios de la frontera de Colombia con Venezuela. Algunas de las fuentes utilizadas son: el RUV de la UARIV³, el Observatorio de Memoria y Conflicto del Centro Nacional de Memoria Histórica (CNMH)⁴, la Defensoría del Pueblo⁵, el Sistema Integrado de Monitoreo de Cultivos Ilícitos del Ministerio de Justicia y del Derecho⁶ y la Agencia Nacional de Minería⁷. Se seleccionaron inicialmente un total de 14 variables para el período 2011 – 2022, entre ellas, el número de hechos victimizantes* (esta variable se toma como medida de la incidencia de la violencia); la presencia de grupos armados como las FARC, las Disidencias de las FARC, el ELN, y el EPL; el número de enfrentamientos; la presencia de economías ilegales; y variables relacionadas con la cercanía a la frontera (Ver Tabla 1 para más detalles).

El conjunto de datos inicial consta de 2.664 registros, organizados en 7 variables de tipo continuo y 7 variables de tipo categórico. Los datos fueron recopilados de fuentes públicas, lo

³ Tomado de <https://www.unidadvictimas.gov.co/publicacion-de-datos-abiertos/>

⁴ Tomado de <https://micrositios.centrodememoriahistorica.gov.co/observatorio/>

⁵ Tomado de <https://alertastempranas.defensoria.gov.co/>

⁶ Tomado de <https://www.datos.gov.co>

⁷ Tomado de <https://www.datos.gov.co>

que facilitó el acceso a la información. Sin embargo, para algunas variables como las relacionadas con la presencia de grupos armados y enfrentamientos, la información se encontraba incompleta para el período analizado en una única fuente. En este sentido, tras seleccionar las variables, se realizó un preprocesamiento inicial para cruzar diferentes fuentes como el RUV, datos del conflicto armado del CNMH y datos de Alertas Tempranas suministrados por la Defensoría del Pueblo. Esto en la búsqueda de completar los datos de presencia de grupos armados y enfrentamientos para el período analizado. Finalmente, basado en las hipótesis planteadas para el análisis se agrupó el conjunto de variables según tres categorías principales:

- **Variables relacionadas con la frontera:** incluye las variables de “distancia mínima” y “municipio de frontera”.
- **Variables relacionadas con la presencia de economías ilegales:** incluye las variables de “hectáreas de coca”, “erradicación de hectáreas de coca”, “destrucción de laboratorios”, “presencia de minería”, “producción de oro” y “producción de plata”
- **Variables relacionadas con la presencia de grupos:** incluye las variables de “presencia de AGC”, “presencia de “ELN”, “presencia de FARC” y “presencia de EPL”.

En conclusión, el conjunto de datos permite analizar la incidencia de la violencia en los municipios de frontera de manera integral, considerando variables internas al conflicto como lo son la presencia de grupos armados, la presencia de economías ilícitas, y variables relacionadas con la cercanía a la frontera. Las variables seleccionadas son relevantes para el análisis dado que permite abordar el problema de la violencia desde una perspectiva multicausal, teniendo en cuenta diferentes indicadores que se han identificado como factores significativos que pueden tener incidencia en que unos municipios presenten mayores niveles de violencia que otros.

Variable	Tipo	Fuente
Hechos victimizantes	Numérica continua	RUV
Presencia de AGC	Categórica binaria	RUV, CNMH, Defensoría del Pueblo
Presencia de Disidencias de las FARC	Categórica binaria	RUV, CNMH, Defensoría del Pueblo
Presencia de ELN	Categórica binaria	RUV, CNMH, Defensoría del Pueblo

Presencia de FARC	Categórica binaria	RUV, CNMH, Defensoría del Pueblo
Presencia de EPL	Categórica binaria	RUV, CNMH, Defensoría del Pueblo
Número de hectáreas de coca	Numérica continua	Sistema Integrado de Monitoreo de Cultivos Ilícitos
Erradicación hectáreas de coca	Numérica continua	Ministerio de Justicia
Destrucción laboratorios	Numérica discreta	Ministerio de Justicia
Presencia de minería	Categórica binaria	Agencia Nacional de Minería
Producción de oro	Numérica continua	Agencia Nacional de Minería
Producción de plata	Numérica continua	Agencia Nacional de Minería
Número de enfrentamientos	Numérica discreta	RUV, CNMH, Defensoría del Pueblo
Municipio de frontera	Categórica binaria	Colombia en mapas

Tabla 1. Conjunto de datos inicial.

4.2 Análisis exploratorio y procesamiento de datos.

El análisis exploratorio tiene como propósito comprender y resumir las características fundamentales del conjunto de datos antes de aplicar técnicas avanzadas de modelado o análisis. Este proceso incluye la identificación de patrones, relaciones y tendencias en las variables, así como la detección de valores típicos, datos faltantes y posibles inconsistencias. En este contexto, para el desarrollo del análisis exploratorio, inicialmente se construyeron nuevas variables como la “distancia mínima a la frontera”, calculada a partir de la distancia euclidia entre el centroide de cada municipio y los pasos fronterizos, tomando la distancia a la frontera más cercana. Adicionalmente, se crearon variables temporales como “post acuerdo” y “período presidencial”. La variable “post acuerdo” agrupa los datos en función de las etapas del acuerdo de paz: antes del acuerdo de paz (2011 – 2016) y después del acuerdo de paz (2017 – 2022). Por su parte, “período presidencial” distingue entre el primer período de Juan Manuel Santos (2011-2014), su segundo período (2015-2018) y la administración de Iván Duque (2019-2022).

Posteriormente, se llevó a cabo un proceso de limpieza de datos y análisis de valores faltantes e inconsistencias en los datos. Se identificaron datos faltantes para la variable de “erradicación

de hectáreas de coca”, correspondiente al año 2011 en cada uno de los municipios del conjunto de datos. Para imputar estos valores, se realizó un ajuste por medio de una transformación logarítmica ($\log(x + 1)$) **para suavizar los valores extremos de la variable**, y se ajustó un modelo de interpolación cúbica por partes. Este modelo se ajusta a los años disponibles dentro de cada grupo y extrapola el valor faltante del primer año.

En la Tabla 2 se presentan las estadísticas descriptivas de las variables continuas. Las variables como “hechos victimizantes”, “erradicación de hectáreas de coca” y “hectáreas de coca” evidencian un rango amplio con alta variabilidad, lo que sugiere la presencia de valores atípicos. En términos generales las variables continuas presentan valores de la media superiores a la mediana y muestran una distribución asimétrica positiva (Ver Anexo A2). Por otra parte, en relación con las variables categóricas, se identifica que la mayoría de los municipios no tienen presencia de grupos armados en el período analizado, tan solo el 15,1% de los municipios tienen presencia del ELN entre el 2011-2022, siendo este el grupo con mayor incidencia en el territorio. De igual manera, se identifica que tan solo el 16,2% (36) de los municipios corresponden a municipios de frontera (Ver Anexo A3).

Estadística	Hechos victimizantes	Hectáreas de coca	Erradicación hectáreas de coca	Destrucción laboratorios	Distancia mínima	Enfrentamientos
Conteo	2664,00	2664,00	2664,00	2664,00	2664,00	2664,00
Promedio	159,77	111,12	86,77	2,89	192,51	0,04
Desviación estándar	723,54	1041,10	3468,17	19,88	94,83	0,44
Mínimo	0,00	0,00	0,00	0,00	4,08	0,00
Percentil 25%	0,00	0,00	0,00	0,00	109,97	0,00
Mediana (50%)	5,00	0,00	0,00	0,00	209,62	0,00
Percentil 75%	49,00	0,00	0,00	0,00	276,32	0,00
Máximo	14858,00	22229,60	178483,23	415,00	345,80	13,00

Tabla 2. Estadística descriptiva de variables continuas.

Para la variable de análisis “hechos victimizantes”, se observa que los años con los niveles más altos de violencia corresponden a 2018 y 2022, registrando un total de 65.271 y 79.782 incidentes, respectivamente (ver Figura 1.a). Por otro lado, como se aprecia en el histograma de distribución de la Figura 1.c, la variable presenta una asimetría positiva pronunciada con colas alargadas, lo que indica que la mayoría de los municipios reportan valores bajos de hechos victimizantes, mientras que un pequeño grupo de municipios concentran valores considerablemente altos. Así, al analizar el gráfico de distribución espacial (Ver Figura 1.b) se puede identificar que para el año 2022, el 46,4% de los municipios no registraron incidentes, mostrando valores de hechos victimizantes cercanos a 0. Este mismo comportamiento se mantiene consistente a lo largo de los años analizados (Ver Anexo A1). Finalmente, el diagrama de caja de la variable hechos victimizantes confirma la alta concentración de datos en valores bajos, reflejado en una caja intercuartil estrecha, junto con la presencia de múltiples valores atípicos que se extienden significativamente más allá de los bigotes (Ver Figura 1.d)

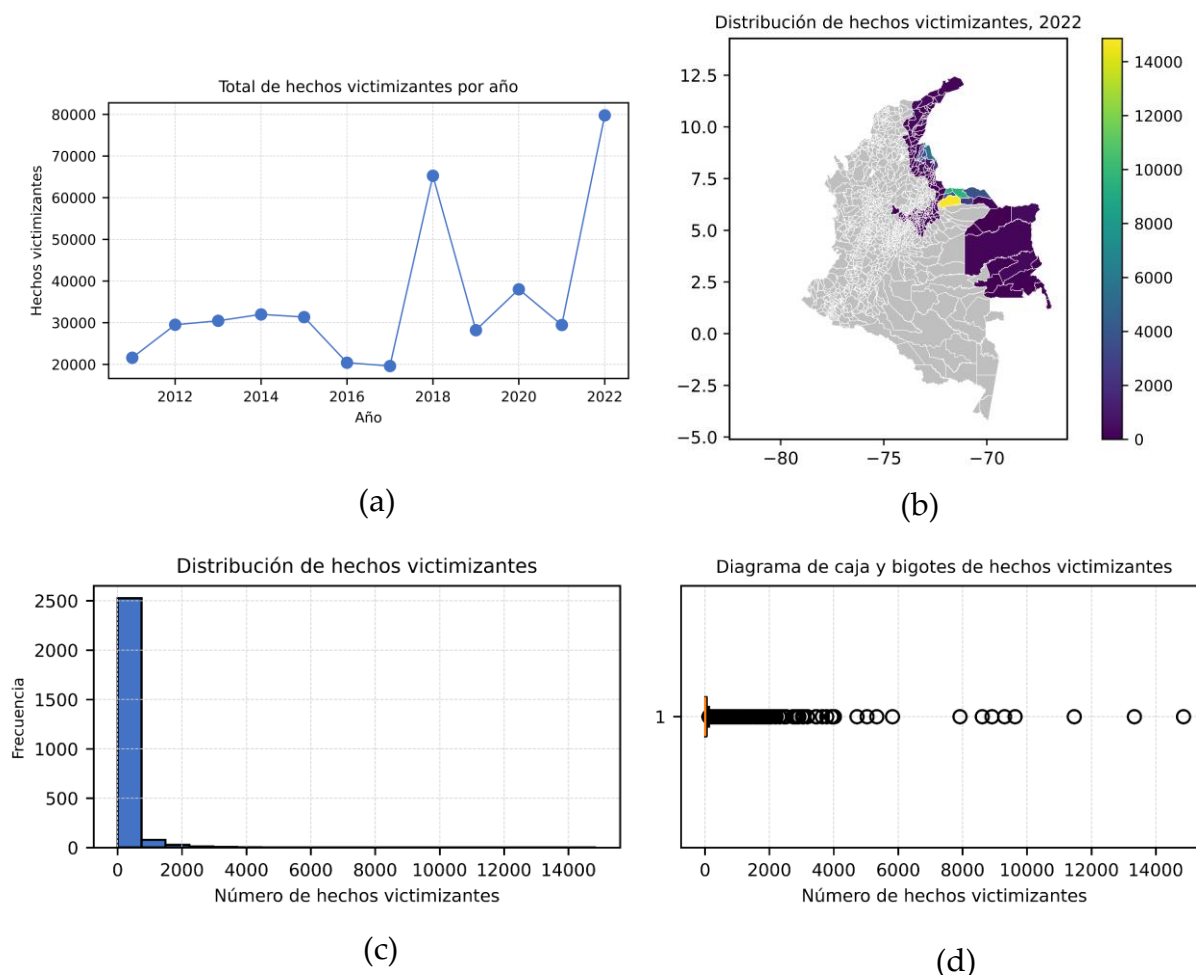


Figura 1. Análisis exploratorio de la variable hechos victimizantes

Debido a que el análisis exploratorio indicó la posible presencia de múltiples valores atípicos en los datos, se implementó un algoritmo *isolation forest* para su detección [57], el cual es un algoritmo de aprendizaje automático no supervisado diseñado para identificar observaciones que se apartan considerablemente del comportamiento general en el conjunto de datos. Como resultado, se identificaron un total de 367 valores atípicos, correspondiente al 13,8% del total de observaciones.

Tras identificar los valores atípicos, se determinó no eliminar ni imputar estos valores, ya que representan municipios con altos niveles de violencia, los cuales son precisamente uno de los focos principales del análisis. Estos valores, aunque atípicos desde el punto de vista estadístico, contienen información valiosa que permite comprender las dinámicas específicas de los municipios más afectados por la violencia. Por lo tanto, se optó por conservarlos y aplicar una transformación logarítmica ($\log(x + 1)$) a todas las variables numéricas con estos valores, con el objetivo de reducir el impacto de los mismos en las técnicas de agrupación. Así, esta transformación permitió suavizar la distribución de todas las variables continuas con este comportamiento, facilitando un mejor desempeño de los algoritmos. Finalmente, para la adecuada implementación de los métodos de agrupación, se normalizaron todas las variables.

Con el conjunto de datos completo, se implementó un análisis de correlación con el objetivo de evaluar la fuerza de la relación entre el conjunto de datos y la variable “hechos victimizantes”. Se utilizaron dos métodos: el coeficiente de correlación de Pearson para identificar relaciones lineales, y el coeficiente de Spearman para capturar relaciones no lineales. En la Figura 2 se muestran los resultados del análisis de correlación del conjunto de datos con la variable “hechos victimizantes”. Este análisis reveló que las variables relacionadas con presencia de grupos armados presentan una correlación positiva, siendo la variable de “presencia de ELN” la que evidencia una mayor correlación con la variable “hechos victimizantes”. Las variables relacionadas con la frontera tienen una correlación no lineal moderada, con una correlación negativa para la variable de “distancia mínima a la frontera” y una correlación positiva para la variable “municipio de frontera”, sugiriendo que los municipios que pertenecen a la frontera y están cerca de los pasos fronterizos tienden a presentar mayores niveles de violencia. Por otro lado, las variables asociadas a la presencia de economías ilegales como “hectáreas de coca”, “erradicación de cultivos” y “destrucción de laboratorios”, tienen una correlación no lineal positiva moderada. En contraste, las variables relacionadas con la presencia de minería mostraron una correlación poco significativa, con valores inferiores a 0.1. Como resultado, se determinó eliminar del análisis las variables que presentaban una correlación baja en la búsqueda de reducir el ruido en la definición de grupos.

Adicionalmente, se realizó un análisis de correlación multivariado, y utilizando el Factor de Inflación de la Varianza (VIF) se filtraron todas las variables que tuvieran un VIF mayor a cinco con el objetivo de identificar problemas de multicolinealidad. Por consiguiente, se identificó que ninguna de las variables presenta problemas de multicolinealidad severa (Ver Anexo A1). En la Tabla 3 se presenta el conjunto final de datos, el cual consta de 15 variables para el período 2011 – 2022⁸

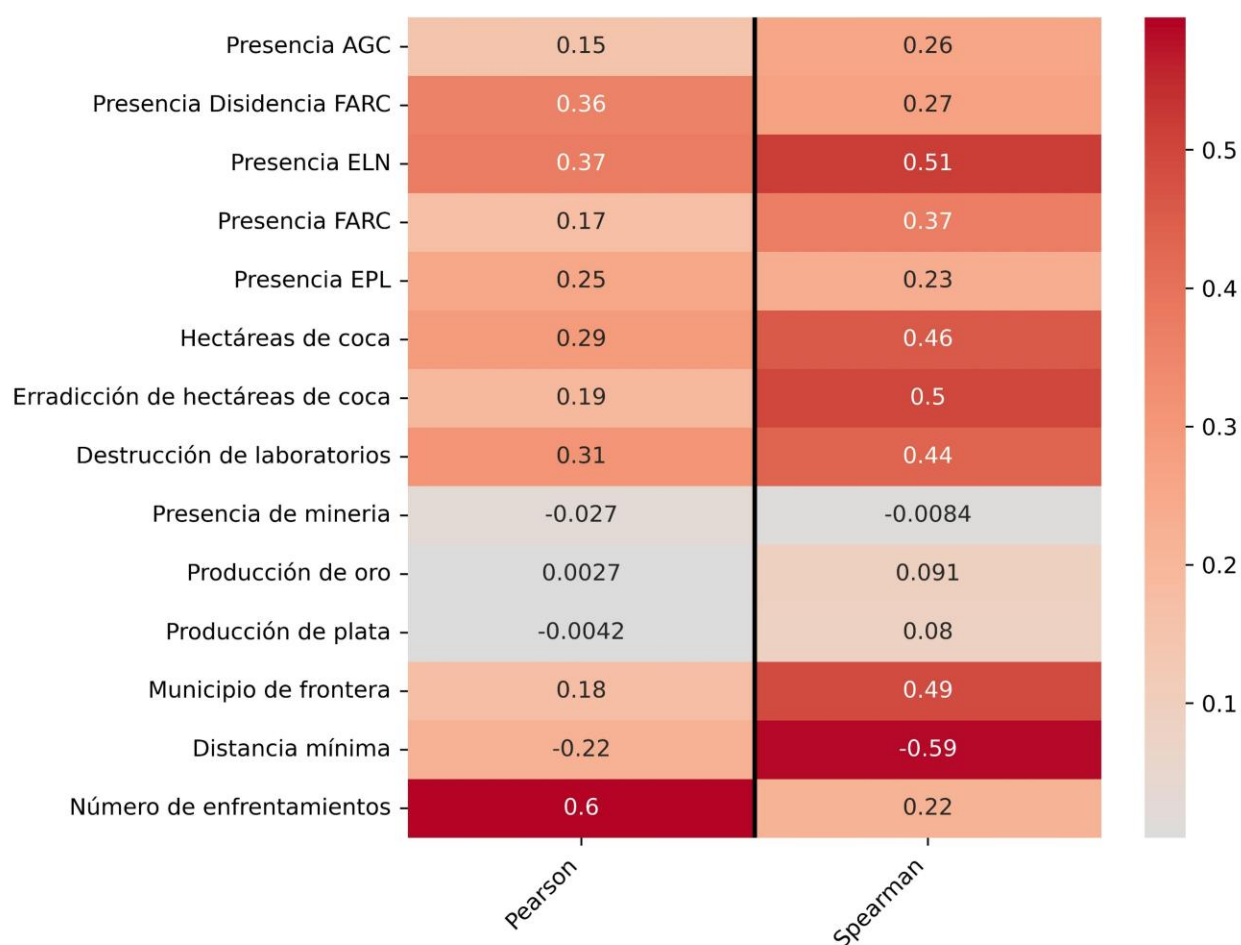


Figura 2. Mapa de calor correlación de Pearson y Spearman

⁸ El conjunto de datos se encuentra en github.com/Danielaloperap/An-lisis-de-patrones-de-violencia-en-los-municipios-de-frontera

Variable	Tipo	Descripción
Hechos victimizantes	Numérica continua	Número de hechos victimizantes
Presencia de AGC	Categórica binaria	1 si hay presencia del grupo armado, 0 de lo contrario
Presencia de Disidencias de las FARC	Categórica binaria	1 si hay presencia del grupo armado, 0 de lo contrario
Presencia de ELN	Categórica binaria	1 si hay presencia del grupo armado, 0 de lo contrario
Presencia de FARC	Categórica binaria	1 si hay presencia del grupo armado, 0 de lo contrario
Presencia de EPL	Categórica binaria	1 si hay presencia del grupo armado, 0 de lo contrario
Número de hectáreas de coca	Numérica continua	Número de hectáreas de coca presentes en el municipio
Erradicación hectáreas de coca	Numérica continua	Número de hectáreas de coca erradicadas en el municipio
Enfrentamientos	Numérica discreta	Número de enfrentamientos entre grupos armados
Destrucción de laboratorios	Numérica discreta	Número de laboratorios destruidos.
Municipio de frontera	Categórica binaria	1 si el municipio es de frontera, 0 de lo contrario
Distancia mínima a la frontera	Numérica continua	Distancia del centroide del municipio al paso fronterizo más cercano
Año	Numérica discreta	Año de medición de la variable
Post acuerdo	Categórica	1 si el año de análisis es posterior al acuerdo de paz, 0 de lo contrario
Período presidencial	Discreta	Agrupar los dos años de los períodos presidenciales de Juan Manuel Santos y el período de Iván Duque

Tabla 3. Descripción del conjunto de datos final.

En resumen, el análisis exploratorio ha revelado patrones claves en la distribución de las variables de interés, destacando la concentración de valores altos para las variables “hechos victimizantes”, “hectáreas de coca” y “erradicación de hectáreas de coca”, en un número reducido de municipios, indicando la presencia de valores atípicos. A pesar de corroborar la presencia de estos valores, se determinó conservarlos, ya que representan observaciones cruciales para el análisis dado que capturan dinámicas relevantes en las zonas más afectadas por el conflicto. De igual manera, se identificó que un porcentaje significativo de los municipios analizados no tienen presencia de grupos armados ni se ubican en zonas fronterizas. Por otro lado, en términos generales, todas las variables seleccionadas para el análisis tienen una correlación significativa con la variable “hechos victimizantes; pero las variables que presentan una mayor correlación son el “número de enfrentamientos”, seguido de “distancia mínima a la frontera” y la “presencia de ELN”. Lo anterior resalta la importancia de estos factores para el análisis de los patrones de violencia en los municipios de frontera.

Estos hallazgos son consistentes con las dinámicas esperadas en territorios afectados por el conflicto armado, donde ciertos territorios, debidos a sus características específicas, tienden a experimentar niveles más altos de violencia. Identificar los patrones y características asociadas a estos territorios constituye uno de los objetivos principales de este estudio. En este contexto, se procede a estructurar el diseño experimental del análisis, el cual, se detalla a continuación:

4.2.1 Diseño experimental

Para llevar a cabo un análisis exhaustivo, se definieron tres escenarios distintos basados en criterios temporales, y cinco escenarios basados en diferentes conjuntos de variables. Cada uno de los escenarios busca capturar aspectos específicos para validar las hipótesis planteadas.

4.2.1.1 Agrupación por criterios temporales

- **Escenario 1. Agrupación por año :** en este escenario se segmenta el conjunto de datos a nivel anual, y se generan los grupos para cada uno de los años en el período 2011-2022.
- **Escenario 2. Agrupación por período del acuerdo de paz:** en este escenario, se agrupa el conjunto de datos según el período del acuerdo de paz. Para el período antes del acuerdo 2011 – 2016 y para el período posterior al acuerdo 2017-2019. Se genera un agrupamiento para cada conjunto de datos.
- **Escenario 3. Agrupación por período presidencial:** en este escenario, se agrupa el conjunto de datos en función del período presidencial de Juan Manuel Santos en su

primer (2011-2014) y segundo período (2015-2018); y el período presidencial de Ivan Duque (2019-2022).

4.2.1.2 Agrupación por conjuntos de variables

- **Escenario 4. Agrupación por hechos victimizantes:** En este escenario se toma únicamente la variable “hechos victimizantes” y se generan los grupos en función de los valores de esta variable por municipio. La razón de este grupo es identificar los grupos de municipios que presentan niveles de violencia similares, evitando que el análisis sea influido por otras dimensiones. Esto permite trazar comparaciones claras y comprender de manera más directa cómo varía la incidencia de la violencia entre municipios.
- **Escenario 5. Agrupación por variables de frontera:** En este escenario se toma la variable “hechos victimizantes” en conjunto con las variables de “municipio de frontera” y “distancia mínima a la frontera”, y se generan grupos en función de los valores de este conjunto de variables por municipio. La razón de este escenario es poder analizar los patrones asociados al contexto fronterizo, esto con el objetivo de identificar si los municipios más cercanos a la frontera tienden a presentar mayor incidencia de la violencia.
- **Escenario 6. Agrupación por variables de presencia de economías ilegales:** En este escenario se agrupan los valores de la variable “hechos victimizantes” con “hectáreas de coca”, “erradicación de hectáreas de coca” y “destrucción de laboratorios”. La razón de este grupo es analizar la relación entre la incidencia de la violencia y las dinámicas asociadas a economías ilegales, específicamente la producción de cultivos ilícitos, las acciones de erradicación de estos cultivos, y las operaciones de destrucción de laboratorios. Este escenario busca determinar si hay una concurrencia entre municipios con violencia más crítica y la expansión de economías ilícitas como los cultivos de coca.
- **Escenario 7. Agrupación por variables de presencia de grupos armados:** En este escenario se agrupan los valores de la variable “hechos victimizantes” con las variables “Presencia FARC”, “Presencia Disidencia FARC”, “Presencia ELN”, “Presencia AGC” y “Presencia EPL”. La razón de este escenario es estudiar la relación de la presencia de grupos armados con la incidencia de la violencia, esto con el objetivo de identificar si la presencia de grupos armados incentiva mayores niveles violencia.
- **Escenario 8. Agrupación por todas las variables:** Incluye todo el conjunto de datos, combinando “hechos victimizantes” con las variables de frontera, economías ilegales y presencia de grupos armados. Este escenario busca identificar patrones de comportamiento general al integrar todo el conjunto de variables.

Para cada uno de los escenarios temporales (año, período post acuerdo, y período presidencial), se implementaron los cinco escenarios de agrupación de variables previamente descritos; por lo tanto, se analizan un total de quince combinaciones posibles.

4.3 Métricas de evaluación

En esta sección se describen las métricas de calidad seleccionadas para evaluar los resultados de las técnicas de agrupamiento (k-means, clustering jerárquico, DBSCAN y clustering espectral) y el algoritmo a priori. Estas métricas permiten evaluar el rendimiento, la precisión y la efectividad de las técnicas usadas.

4.3.1 Métricas de evaluación para k-means y agrupamiento jerárquico: en el caso de k-means, para determinar el número óptimo de clústeres se hace uso del método del codo por medio de la evaluación de la métrica de inercia. Para evaluar los resultados finales de los clústeres creados por los algoritmos de k-means y agrupamiento jerárquico, se usa el coeficiente de silueta. A continuación, se describe brevemente cada una de las métricas.

- **La inercia:** mide cuán compactos son los grupos, es decir, qué tan cerca están los puntos de cada grupo a su centroide. Calcula la suma de las distancias al cuadrado entre cada punto de datos y el centroide del grupo al que pertenece [40]. Se utiliza comúnmente en el método del codo para determinar el número óptimo de grupos.
- **Coeficiente de silueta:** mide la calidad de los grupos al evaluar la cohesión interna y su separación respecto a otros grupos. Su valor varía en un rango de -1 a 1. Un coeficiente de silueta cercano a 1 indica que los grupos son compactos y están bien separados, mientras que un coeficiente cercano a -1 indica que el grupo está mal formado [40]. Su fórmula de cálculo es:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Donde $a(i)$ es la distancia de i a los puntos de su grupo, y $b(i)$ es la distancia promedio a los puntos del grupo más cercano.

4.3.2 Métricas de evaluación para DBSCAN: una métrica robusta para validar la calidad de los grupos en DBSCAN es el *Density-Based Clustering Validation Index* (DBCV) [58]. El DBCV es una combinación de los índices de dispersión de densidad y densidad de separación. El primero se refiere a la baja densidad o dispersión de un punto en un espacio; y, el segundo describe la distancia o el grado de separación entre regiones de alta y baja densidad en un

conjunto de datos. El índice varía entre -1 y 1, donde un valor cercano a 1 indica una alta calidad de los grupos. El DBCV se define como:

$$DBC\,V\,(C) = \sum_{i=1}^l \frac{|C_i|}{|O|} V_c\,(C_i)$$

Donde l es el número total de grupos, C_i es el i -ésimo grupo dentro de la solución C , O es el conjunto completo de datos y $V_c\,(C_i)$ es el índice de validez para el grupo C_i , que mide su calidad con base en la densidad interna y la separación del clúster.

4.3.3 Métricas de evaluación para agrupación espectral: una métrica usada para evaluar la calidad de los grupos en la agrupación espectral es el índice de Calinski-Harabasz, el cual evalúa la calidad de los grupos basándose en la variación entre los grupos y dentro de los mismos. Es utilizado para determinar el número óptimo de grupos al maximizar la proporción entre variación inter-clúster y la intra-clúster. Un mayor valor del índice indica una mejor calidad del grupo [59]. Se calcula como:

$$CH\,(k) = \frac{B(k)\,(N - k)}{W(k)\,(k - 1)}$$

Donde k es el número de grupos, $B(k)$ es la covarianza inter-clúster, $W(k)$ es la covarianza intra-clúster, y N es el número total de muestras en el conjunto de datos. Un valor alto de CH indica una mejor agrupación, porque significa que los puntos de datos están más dispersos entre los grupos que dentro de ellos.

4.3.4 Métricas de evaluación algoritmo a priori: algunas métricas frecuentemente usadas para evaluar la calidad de las reglas de asociación generadas por el algoritmo son:

- **Soporte:** mide la frecuencia con la que una regla de asociación aparece en el total de transacciones. Representa la proporción de transacciones que contiene una regla de asociación. Un valor de soporte alto indica que esa regla de asociación es común, mientras que un valor bajo sugiere lo contrario [60]. Se halla como:

$$\text{Soporte } (A \Rightarrow B) = \frac{\text{Número de eventos que contiene } A \cup B}{\text{Número total de eventos}}$$

- **Confianza:** mide la fuerza de la implicación en una regla de asociación. Indica la probabilidad de que un evento Y se dé, dado que un evento X ya se ha dado. Un valor de confianza alto indica que la regla es fiable, mientras que un valor bajo sugiere que la relación no es fuerte [60].

$$\text{Confianza } (A \Rightarrow B) = \frac{\text{Número de eventos que contiene } A \cup B}{\text{Número total de eventos que contienen } A}$$

- **Elevación (Lift):** mide el incremento en la probabilidad de que B ocurra cuando A ocurre, en comparación con B ocurriendo de manera independiente. Un valor de indicador mayor a 1 indica que los eventos están positivamente relacionados, un valor igual a 0 que son independiente, y menor a 1 que están relacionados negativamente. Se calcula como:

$$\text{Elevación } (A \Rightarrow B) = \frac{\text{Confianza } (A \Rightarrow B)}{\text{Soporte } (A \Rightarrow B)}$$

5. MODELADO

Para el análisis de las dinámicas de la violencia en los municipios de frontera se implementaron diferentes algoritmos de aprendizaje automático no supervisado, incluyendo k-means, agrupamiento jerárquico, DBSCAN y agrupamiento espectral. Estas técnicas se combinaron con el algoritmo A priori y un método de explicabilidad en la búsqueda de caracterizar las dinámicas del conflicto en la región. Esta sección describe los modelos de agrupamiento que se obtuvieron con los algoritmos de aprendizaje automático.

Para la implementación del k-means inicialmente se identificó el número de grupos óptimos por medio del método del codo. Para facilitar la comparabilidad de los grupos a lo largo del tiempo, se seleccionó el número de grupos que maximizará el promedio del coeficiente de silueta para todo el período. Tras determinar el número de grupos se implementó el algoritmo

k-means con una inicialización probabilística avanzada (*k-means++*) para mejorar la convergencia del algoritmo.

En cuanto a la técnica de agrupamiento espectral, se identificó el número de vecinos y grupos óptimos por medio de la evaluación del *Eigen-Gap*, que mide la diferencia entre los valores propios consecutivos de la matriz Laplaciana asociada al grafo de similitud construido a partir de los datos. Se seleccionó el número de vecinos que maximizará el *Eigen-Gap*, y para el número de grupos se optó por el valor que maximizará el promedio de la métrica de calidad Calinski Harabasz. Para el caso donde el número de grupos era muy alto, se determinó un número de grupos intermedio, que fuera coherente con el análisis para la comparación a lo largo del tiempo.

Por otra parte, para el algoritmo BDSCAN, se optimizaron los hiperparámetros definiendo una grilla de valores para los valores de ϵ (tamaño del radio de búsqueda) y el número mínimo de puntos en un grupo. Posteriormente, se evalúan los resultados del método utilizando el DBCV, y se generan los grupos con los valores de los parámetros que maximizan la métrica. Finalmente, en relación con el agrupamiento jerárquico se define una grilla de valores que incluye el número de grupos, la métrica de distancia (euclídea o manhattan), y el método de enlace (simple, completa o promedio). Se evalúa la calidad de los grupos creados para cada combinación de parámetros y se crean los grupos con la combinación de parámetros que maximizan el coeficiente de silueta.

De manera similar al caso de agrupamiento espectral, en los casos donde se obtuvo un número considerablemente alto de grupos en el proceso de optimización, se optó por reducir su cantidad. Esto con el objetivo de garantizar la creación de grupos que brinden información significativa tanto para el algoritmo A priori como para el método de explicabilidad; manteniendo al mismo tiempo un buen desempeño en los valores de las métricas. A continuación, se presentan los resultados promedios de las métricas de evaluación obtenidas, los resultados desagregados se pueden consultar en los Anexos B, C y D.

5.3 Resultados de evaluación de las técnicas de agrupamiento para los escenarios por año

En la Tabla 4 se muestran los resultados de las métricas de calidad evaluadas para cada uno de los métodos de agrupamiento implementados. Para el escenario que agrupa la variable “hechos victimizantes”, la técnica que logró un mayor desempeño fue k-means, con un promedio en coeficiente de silueta de 0,69. Por otro lado, para los escenarios con todo el conjunto de datos y con las variables asociadas a economías ilegales, el agrupamiento jerárquico logro un mejor rendimiento. En el caso de los escenarios con las variables relacionadas con presencia de grupos y variables de frontera, el algoritmo BDSCAN da los mejores resultados, logrando un DBCV de 0,67 y 0,81, respectivamente.

El análisis de agrupamiento arrojó que para los años 2020 y 2021 presenta los mejores resultados los escenarios que consideran las variables de presencia de economías ilícitas y grupos armados, con un coeficiente de silueta de 0,87 y 0,86, respectivamente; lo anterior utilizando agrupamiento jerárquico (Ver Anexo B3 y B4). En el caso del escenario con la variable “hechos victimizantes”, el mejor desempeño se observa en el año 2020, donde DBSCAN obtuvo un 0,83 (Ver Anexo B1). Asimismo, para el escenario de las variables de frontera, DBSCAN obtuvo los mejores resultados en el año 2017, con un valor de 0,73 en su métrica de calidad (Ver Anexo B2). Por otro lado, considerando todas las variables en conjunto, el agrupamiento jerárquico mostró el mejor desempeño en los años 2018 y 2020, con una métrica de calidad de 0,83 (Ver Anexo B5). En términos generales, el desempeño más bajo se encuentra en este mismo escenario en el año 2018, con un valor de 17,18 en la métrica de Calinski-Harabasz, utilizando agrupamiento espectral.

Técnica	Métrica	Hechos victimizantes	Variables de frontera	Variables economías ilegales	Variables Presencia grupos	Todas las variables
k-means	Coeficiente de silueta	0,69	0,56	0,61	0,70	0,64
Agrupamiento espectral	Calinski Harabasz	1072,98	276,49	114,40	100,44	43,35
DBSCAN	DBCV	0,58	0,67	0,66	0,81	0,61
Agrupamiento jerárquico	Coeficiente de silueta	0,67	0,63	0,79	0,77	0,76

Tabla 4. Resultados promedio métricas de evaluación de los métodos de agrupamiento por año.

5.4 Resultados de evaluación de las técnicas de agrupamiento para los escenarios por etapa del acuerdo de paz.

En la Tabla 5 se presentan los resultados promedio de las métricas obtenidas para cada escenario evaluado por etapa del acuerdo de paz. Se tiene que el algoritmo k-means presenta mejores resultados para el escenario con la variable “hechos victimizantes”, con un promedio en coeficiente de silueta de 0,69. Para los escenarios relacionados con las variables de frontera y presencia de grupos, DBSCAN presenta los mejores resultados con un valor en la métrica de 0,72 y 0,82, respectivamente. Por otro lado, para el resto de los escenarios, el agrupamiento jerárquico evidencia el mejor desempeño, logrando para los escenarios con las variables

relacionadas con la presencia grupos armados 0,63, y para el escenario con todo el conjunto de variables 0,64.

En el Anexo C se presentan los resultados desagregados para cada escenario. El escenario con mejor desempeño corresponde al que agrupa las variables relacionadas con la presencia economías ilegales, logrando un coeficiente de silueta de 0,87 para el método de agrupamiento jerárquico en el período posterior a la firma del acuerdo de paz (Ver Anexo C4). Asimismo, el agrupamiento jerárquico obtuvo el mejor desempeño para los escenarios con el conjunto de variables de presencia de economías ilegales y el conjunto completo de variables, con un valor de coeficiente de silueta de 0,87 y 0,80, respectivamente, en este mismo período (Ver Anexo C3 y C5). Por otro lado, para el escenario con el conjunto de variables relacionadas con la frontera, el mejor desempeño lo obtuvo DBSCAN, alcanzando un DBCV de 0.74 (Ver Anexo C2). En cuanto al escenario que incluye la variable “hechos victimizantes”, el algoritmo k-means logró un coeficiente de silueta de 0,72, también en el período posterior al acuerdo de paz (Ver Anexo C1). En términos generales, el peor desempeño se observó en el método de agrupamiento espectral aplicado en el escenario con las variables de presencia de grupos armados, con una métrica CH de 85,84 en el período anterior a la firma del acuerdo.

Técnica	Métrica	Hechos victimizantes	Variables de frontera	Variables economías ilegales	Variables Presencia grupos	Todas las variables
k-means	Coeficiente de silueta	0,69	0,56	0,56	0,73	0,54
Agrupamiento espectral	Calinski Harabasz	3150,50	1216,65	218,02	143,99	215,47
DBSCAN	DBCV	0,63	0,72	0,63	0,82	0,64
Agrupamiento jerárquico	Coeficiente de silueta	0,63	0,63	0,84	0,77	0,76

Tabla 5. Resultados promedio métricas de evaluación de los métodos de agrupamiento por etapa del acuerdo de paz.

5.5 Resultados de la evaluación de las técnicas de agrupamiento para el escenario por período de presidencial.

En la Tabla 6 se presentan los resultados promedios de las métricas de evaluación obtenidos para los grupos creados para el escenario por período presidencial (las métricas desagregadas se pueden consultar en el Anexo D). Para el escenario con la variable “hechos victimizantes”,

la técnica k-means y el agrupamiento jerárquico lograron el mejor desempeño con un promedio en coeficiente de silueta de 0,69. En cuanto a los demás escenarios, el agrupamiento jerárquico obtuvo el mejor desempeño promedio con un coeficiente de silueta de 0,63 para el escenario con las variables de frontera, de 0,84 para el escenario con las variables relacionadas con economías ilegales, y de 0,69 y 0,77 para los escenarios con las variables de presencia de grupos y todo el conjunto de datos, respectivamente.

A nivel de cada período presidencial el método con mejor desempeño fue el agrupamiento jerárquico para el escenario que agrupa el conjunto de variables de economías ilegales, con un coeficiente de silueta de 0,89 en el período de Iván Duque (Ver Anexo D3). De igual manera, el agrupamiento jerárquico logró el mejor rendimiento para el conjunto de todas las variables en el segundo período de Juan Manuel Santos, con un valor en la métrica de 0,81 (Ver Anexo D5). Para el resto de los escenarios, DBSCAN alcanzó el mejor desempeño con un valor en DBCV de 0,77 para el escenario con la variable “hechos victimizantes” en el período de Iván Duque (Ver Anexo D1), de 0,75 para el escenario con las variables relacionadas con la cercanía a la frontera en el segundo período de Juan Manuel Santos (Ver Anexo D2), y de 0,83 para el escenario con las variables relacionadas con la presencia de grupos en el primer período de Juan Manuel Santos (Ver Anexo D4).

Técnica	Métrica	Hechos victimizantes	Variables de frontera	Variables economías ilegales	Variables Presencia grupos	Todas las variables
k-means	Coeficiente de silueta	0,69	0,56	0,57	0,73	0,6
Agrupamiento espectral	Calinski Harabasz	2910,95	847	185,93	786,09	170,71
DBSCAN	DBCV	0,67	0,7	0,55	0,81	0,65
Agrupamiento jerárquico	Coeficiente de silueta	0,69	0,63	0,84	0,69	0,77

Tabla 6. Resultados promedio métricas de evaluación de los métodos de agrupamiento por período presidencial

En términos generales los resultados destacan rendimientos significativos para los métodos basados en distancia. En particular, el agrupamiento jerárquico se destaca como el método con

el mejor desempeño, logrando una separación clara y consistente en los grupos. Por otro lado, aunque el agrupamiento espectral presenta ventajas al adaptarse a estructuras de datos complejos y formas no lineales, mostró limitaciones en la generación de grupos representativos para el análisis. Asimismo, a pesar de que DBSCAN obtuvo el mejor desempeño para el escenario que agrupa las variables relacionadas con presencia de grupos armados, un análisis detallado de los grupos evidenció que para los casos donde se presentan valores significativos de hechos victimizantes (estos municipios son el objetivo de análisis), DBSCAN tiende a generar una gran cantidad de grupos formados por pocos datos (en promedio tres municipios) (ver Anexo B6). En este caso, los grupos formados pueden dificultar la interpretación y reducir su utilidad para el análisis, ya que no capturan patrones generales, lo que resulta en grupos poco significativos. Esto debido a que los grupos de principal interés son los que concentran los municipios con altos niveles de hechos victimizantes.

Por lo tanto, se seleccionó en este caso los resultados generados por el método de agrupamiento jerárquico. La razón es porque esta técnica no solo presenta buen desempeño en términos de las métricas de calidad, sino que también muestra comportamientos más coherentes y estables en cuanto al número de grupos generados. Este criterio es clave para garantizar que los resultados sean significativos y útiles al momento de aplicar el algoritmo A priori y la técnica de explicabilidad.

En la Tabla 7 se presenta un cuadro resumen de los métodos seleccionados por escenario para el análisis de resultados. Para el escenario relacionado con la variable “hechos victimizantes” se seleccionaron los resultados generados por la técnica k-means, debido a que mostró el mejor desempeño en las métricas evaluadas. Por otro lado, en el caso del escenario que agrupa las variables relacionadas con la proximidad a la frontera, se optó por los resultados de DBSCAN, ya que presenta el mejor desempeño en la métrica, y tras un análisis detallado de los grupos formados, se encuentra que en este caso son significativos para el análisis. En contraste, para el escenario que agrupa las variables relacionadas con la presencia de grupos armados, se seleccionó el agrupamiento jerárquico a nivel de año y etapa del acuerdo de paz. Aunque DBSCAN obtuvo el mejor desempeño en las métricas para este escenario, el agrupamiento jerárquico fue preferido por generar resultados más significativos para el análisis. En el caso de los grupos generados a nivel de período presidencial, k-means presentó los mejores resultados, siendo seleccionado como el método más adecuado. Finalmente, para el escenario que incluye la totalidad del conjunto de datos, se eligió el agrupamiento jerárquico, ya que presentó el mejor desempeño según las métricas evaluadas.

Técnica	Año	Etapas del acuerdo de paz	Período presidencial
Hechos victimizantes	k-means	k-means	k-means
Variables de frontera	DBSCAN	DBSCAN	DBSCAN
Variables de economías ilegales	Agrupamiento jerárquico	Agrupamiento jerárquico	Agrupamiento jerárquico
Variables de presencia de grupos	Agrupamiento jerárquico	Agrupamiento jerárquico	k-means
Todas las variables	Agrupamiento jerárquico	Agrupamiento jerárquico	Agrupamiento jerárquico

Tabla 7. Métodos de agrupamiento con mejor desempeño por escenario de análisis

6. ANÁLISIS DE RESULTADOS

Tras identificar los métodos de agrupamiento con mejor desempeño, se realizó el análisis de los patrones característicos de los grupos formados para cada escenario con esos métodos. Posteriormente, se seleccionaron los grupos más significativos (con niveles altos de hechos victimizantes) para generar reglas de asociación mediante la implementación del algoritmo *A priori*, lo que permitió identificar relaciones significativas entre las variables de cada grupo. Adicionalmente, se empleó el método SHAP para determinar las características más influyentes en la formación de esos grupos, esto para brindar una visión interpretativa de cómo cada variable contribuyó al agrupamiento. En esta sección se presentan los hallazgos obtenidos al analizar detalladamente los centroides (patrones) de los grupos formados, y aplicar el algoritmo *A priori* y el método SHAP para explicar los grupos más significativos.

6.1 Análisis de centroides de métodos de agrupamiento

En este apartado se presenta el análisis de los patrones de violencia armada obtenidos a partir de la implementación de los métodos de agrupamiento. Los algoritmos utilizados permitieron identificar grupos de municipios con características similares en términos de sus niveles de violencia y factores asociados, como la proximidad a la frontera, la presencia de economías ilícitas y grupos armados. A continuación, se describen las características principales de los grupos formados para cada escenario.

6.1.1 Escenario hechos victimizantes

Para el escenario basado en la variable que mide los niveles de violencia en los municipios de frontera se obtuvieron tres grupos usando el algoritmo k-means. Este resultado es consistente en los sub-escenarios generados por año, etapa del acuerdo de paz y período presidencial. A continuación, se describe las características generales de cada grupo y se selecciona como ejemplo en cada caso el año 2018 para los grupos por año, esto debido a que este año presenta uno de los mayores niveles de hechos victimizantes.

- **Grupo 1:** contiene la mayor cantidad de municipios y se caracteriza por bajos niveles de hechos victimizantes. Es el grupo que concentra la mayor cantidad de municipios. Para el año 2018, la media de hechos victimizantes para este grupo es de 0 y representa el 52% del total de municipios analizados.
- **Grupo 2:** en términos generales, este grupo está compuesto por municipios con niveles intermedios de violencia. Para el año 2018, la media de hechos victimizantes para los municipios de este grupo es de 3,4 y representa el 34% del total de municipios.
- **Grupo 3:** los municipios que componen este grupo presentan promedios de violencia considerablemente altos. Es el grupo que concentra menos cantidad de municipios. Para el caso del año 2018, la media de hechos victimizantes para los municipios de este grupo es de 2.87, concentrando el 14% de los municipios.

Se evidencia que el promedio de hechos victimizantes es superior a la mediana en el grupo 3, lo que indica la presencia municipios con altos valores de hechos victimizantes que elevan la media. Finalmente, se destaca que, para el período posterior a la firma del acuerdo de paz, los tres grupos muestran promedios de hechos victimizantes superiores en comparación con el período previo a la firma (Ver Anexo E1). Asimismo, durante el gobierno de Iván Duque se observa el mayor promedio de hechos victimizantes entre los períodos analizados para los tres grupos (Ver Anexo E2).

6.1.2 Escenario con variables de frontera

En el escenario del conjunto de variables que miden la cercanía a la frontera, DBSCAN generó dos grupos por año, con excepción de los años 2014, 2015, 2018 y 2020 a 2022, donde se agruparon los datos que el método identificó como ruido como datos atípicos. De igual manera, para el escenario por período del acuerdo de paz, el método formó dos grupos para la etapa anterior a la firma del acuerdo, y para la etapa posterior se generó un tercer grupo con los datos atípicos. Por último, para los grupos formados en el escenario por período presidencial se identificaron dos grupos, con excepción del segundo período de Juan Manuel Santos, donde se agruparon en un tercer grupo los datos identificados como ruido. Las características generales de los grupos son iguales para los escenarios a nivel de año, etapa del acuerdo de paz, y período

presidencial. A continuación, se describen las características principales de los grupos generados, y se selecciona el período tras la firma del período de paz como caso de análisis, esto debido al interés de identificar los patrones de violencia que componen este período.

6.1.2 Escenario con variables de frontera

En el escenario del conjunto de variables que miden la cercanía a la frontera, DBSCAN generó dos grupos por año, con excepción de los años 2014, 2015, 2018 y 2020 a 2022, donde se agruparon los datos que el método identificó como ruido como datos atípicos. De igual manera, para el escenario por período del acuerdo de paz, el método formó dos grupos para la etapa anterior a la firma del acuerdo, y para la etapa posterior se generó un tercer grupo con los datos atípicos. Por último, para los grupos formados en el escenario por período presidencial se identificaron dos grupos, con excepción del segundo período de Juan Manuel Santos, donde se agruparon en un tercer grupo los datos identificados como ruido. Las características generales de los grupos son iguales para los escenarios a nivel de año, etapa del acuerdo de paz, y período presidencial. A continuación, se describen las características principales de los grupos generados, y se selecciona el período tras la firma del período de paz como caso de análisis, esto debido al interés de identificar los patrones de violencia que componen este período.

- **Grupo 1:** este grupo está conformado principalmente por municipios no fronterizos, que se encuentran alejados (a más de 200 km en promedio) de los pasos fronterizos. Los municipios de este grupo se caracterizan por registrar niveles bajos de hechos victimizantes, y representan alrededor del 80% del total de municipios analizados. Por ejemplo, para el período posterior a la firma del acuerdo de paz, este grupo está conformado por municipios que no son de frontera, tienen una media de hechos victimizantes igual a 97, y una distancia mínima a los pasos fronterizos de 211 km. Este grupo concentra el 83,5% del total de municipios analizados.
- **Grupo 2:** se compone principalmente de municipios que son de frontera y se encuentran en promedio a 100 km de los pasos fronterizos. Asimismo, agrupa municipios con niveles altos de violencia. Agrupa aproximadamente el 16% de los municipios analizados. Por ejemplo, en el período posterior a la firma del acuerdo este grupo está compuesto por municipios que son de frontera, y tienen una media de hechos victimizantes igual a 433.
- **Grupo 3:** este grupo se compone de una cantidad reducida de municipios y aparece solamente en ciertos escenarios. En cada escenario se define de una manera diferente. En unos escenarios se define este grupo por municipios que no son de frontera y

presentan niveles de hechos victimizantes por encima del promedio (es el caso del escenario por agrupamiento por período presidencial, año 2015). En otros escenarios representa municipios que son de frontera y presentan niveles de violencia superiores al conjunto de municipios (son los casos de los años 2014, 2018 y 2020 a 2022, y etapa posterior al acuerdo de paz). Por ejemplo, en el caso del período posterior al acuerdo de paz, este grupo se conforma por ocho municipios que son de frontera y tienen una media de hechos victimizantes igual a 5.930.

Se observa un incremento en el promedio de hechos victimizantes tanto en los municipios del grupo 1, caracterizados por bajos niveles de violencia, como en los del grupo 2, con altos niveles de violencia, durante el período posterior a la firma del acuerdo de paz, en comparación con el período anterior a la firma del acuerdo. En el grupo 1, el promedio de hechos victimizantes paso de 83 en el período 2011-2016 a 97 en el período 2017-2022. Por su parte, el grupo 2 mostró una variación aún más significativa, pasando de 375 en el primer período a 433 en el período tras la firma del acuerdo de paz (Ver Anexo E3).

6.1.3 Escenario con variables de economías ilegales

Para este escenario con variables relacionadas con la presencia de economías ilegales, el agrupamiento jerárquico formó dos grupos tanto a nivel de año, etapa del acuerdo de paz y período presidencial. Las características generales de los grupos se describen a continuación, y en detalle, se describen los resultados para el grupo formado para el período posterior a la firma del acuerdo de paz (Ver Anexo E4). Se selecciona este período debido a que se considera relevante para el análisis comprender los patrones asociados al período posterior a la firma del acuerdo teniendo en cuenta las transformaciones que se dieron en los territorios.

- **Grupo 1:** este grupo está caracterizado por municipios con bajos niveles de violencia y limitada actividad relacionada con cultivos de coca, erradicación de cultivos y destrucción de laboratorios, presentando promedios muy bajos. En este grupo se concentran en promedio el 90% de los municipios analizados. Por ejemplo, para el período posterior a la firma del acuerdo de paz, este grupo tiene una media de hechos victimizantes de 92, un promedio de hectáreas de coca sembradas de 12,6, y una media de erradicación de cultivos y destrucción de laboratorios de 5,3 y 0,6, respectivamente.
- **Grupo 2:** representa municipios con altos niveles de violencia y una fuerte relación con economías ilícitas, caracterizados por extensas áreas de cultivos de coca y una mayor actividad de erradicación y destrucción de laboratorios. Para el período posterior a la firma del acuerdo, los municipios que componen este grupo tienen un promedio de

hechos victimizantes de 1936. Por otro lado, el promedio de hectáreas de coca es de 3.392 y de erradicación de 329.

Se observa que los municipios que conforman el grupo 2 presentan un mayor promedio de hechos victimizantes en el período posterior a la firma del acuerdo, acompañado de un aumento en el número de municipios que integran este grupo, en comparación con el grupo con estas mismas características en el período anterior a la firma del acuerdo. Asimismo, se evidencia un incremento en las áreas de coca cultivadas, mientras que se registra una disminución en el promedio de hectáreas de coca erradicadas y laboratorios destruidos. Por otra parte, en el período de Iván Duque se presenta los mayores niveles de hechos victimizantes y hectáreas de coca cultivadas para los municipios del grupo 2.

6.1.4 Escenario con variables de presencia de grupos

Respecto a las variables asociadas a la presencia de grupos armados, el método de agrupamiento jerárquico identificó tres grupos tanto a nivel anual como por período del acuerdo de paz y período presidencial. De manera similar, el método k-means generó tres grupos al agrupar los datos por períodos presidenciales. A continuación, se detallan las características generales de cada grupo y se selecciona como caso de análisis el año 2022 (este año presenta uno de los niveles de violencia más altos para el período analizado):

- **Grupo 1:** este grupo está conformado por municipios con bajos niveles de violencia y una presencia baja o inexistente de grupos armados. Es el grupo con mayor número de municipios. Para el año 2022, este grupo concentra el 80% del total de municipios analizados. Estos municipios tienen una media de hechos victimizantes de 93, y en promedio, el 15% de los municipios tiene presencia de AGC, FARC y ELN (Ver Anexo E5).
- **Grupo 2:** agrupa municipios con niveles moderados de violencia y una presencia de grupos armados, especialmente en los últimos años del análisis. Se caracteriza por una mayor presencia de las Disidencias de las FARC, ELN y AGC. En el caso del año 2022, los municipios que componen este grupo tienen una media de hechos victimizantes de 656, el 75% de los municipios tiene presencia de AGC, y el 38% de Disidencias de las FARC y ELN. Este grupo concentra el 7,2% del total de los municipios (Ver Anexo E5).
- **Grupo 3:** este grupo reúne los municipios con los niveles más críticos de violencia y una fuerte presencia de grupos armados, principalmente de ELN. Adicionalmente, tanto las Disidencias de las FARC como las FARC, tienen también una presencia significativa en este grupo. En cuanto al año 2022, este grupo está conformado por 28 (12,6%)

municipios que tienen una media de hechos victimizantes de 1.884. En relación con la presencia de grupos armados, para el año 2022, el 11% de los municipios tienen presencia de AGC, el 71,4% de Disidencias de las FARC, el 64,3% de ELN, y el 3,6% de EPL (Ver Anexo E5).

En el período posterior a la firma del acuerdo de paz, se observa un incremento en el promedio de hechos victimizantes, acompañado de una disminución en la presencia de las AGC y de las FARC, quienes se reorganizan como Disidencias de las FARC tras la firma del acuerdo. No obstante, se registra un incremento en la influencia de grupos como el ELN en los municipios que conforman el grupo 2.

6.1.5 Escenario con todas las variables

En relación con los grupos generados para el escenario con todo el conjunto de datos, se tiene que el método de agrupación jerárquico formó dos grupos tanto a nivel anual como para la etapa del período de paz y el período presidencial. Las características de los grupos se detallan a continuación:

- **Grupo 1:** este grupo se encuentra conformado por municipios que presentan bajos niveles de hechos victimizantes, baja presencia de grupos armados y economías ilegales. En términos generales, los municipios no son de frontera y presentan una distancia promedio a los pasos fronterizos más cercanos de 190 km, indicando que la mayoría de los municipios no están cerca de la frontera. Este grupo concentra más del 90% del total de los municipios analizados.
- **Grupo 2:** agrupa municipios con altos niveles de hechos victimizantes, una presencia significativa de áreas de cultivos de coca, enfrentamientos y presencia de grupos armados, principalmente de ELN, EPL, FARC y Disidencia FARC posterior a la firma del acuerdo de paz. Estos municipios son principalmente municipios de frontera que se encuentran en promedio a 100 km del paso fronterizo más cercano.

Tras la firma del acuerdo de paz se observa para los municipios del grupo 2 un incremento en el promedio de hechos victimizantes, así como un incremento en las hectáreas de cultivos de coca, esto acompañado de un incremento en la presencia de grupos como el ELN, en comparación con el período anterior a la firma del acuerdo.

6.1.6 Discusión general de los patrones encontrados en los Escenarios

En general, el análisis de patrones de los grupos formados muestra que los municipios con los niveles más altos de hechos victimizantes se encuentran caracterizados por ser municipios de

frontera o estar cercanos a la frontera, presentar extensas áreas de cultivos de coca, y una presencia significativa de múltiples actores armados como el ELN y las Disidencias de las FARC, tras la firma del acuerdo de paz. Aunque se observa una disminución en la influencia de las FARC en el territorio, su transformación en Disidencias y el crecimiento de otros grupos como el ELN, han intensificado los niveles de violencia en el período posterior al acuerdo de paz. Sin embargo, tras la firma del acuerdo también se identificó una disminución en el número de municipios con ocurrencia de hechos victimizantes, lo que sugiere que la violencia se ha intensificado y concentrado en un subconjunto de municipios. Asimismo, el análisis de los diferentes grupos muestra que la violencia tiende a centralizarse en un número limitado de municipios, siendo una proporción más significativa los municipios que presentan niveles bajos de hechos victimizantes.

Los resultados obtenidos apoyan las hipótesis inicialmente planteadas, mostrando que para los municipios que se caracterizan por estar próximos a la frontera, la presencia de múltiples actores armados y economías ilegales como los cultivos de coca, también registran altos niveles de hechos victimizantes. Esta relación refleja la complejidad de las dinámicas de la violencia en esta zona, donde la proximidad a los pasos fronterizos facilita el accionar de grupos armados y el desarrollo de actividades ilícitas.

6.2 Explicabilidad basada en reglas de asociación

En esta sección se exponen las reglas de asociación identificadas mediante la implementación del algoritmo *A priori*. Estas reglas muestran las relaciones entre antecedentes y consecuentes que ocurren con cierta frecuencia en los datos, lo que permite explorar patrones de comportamiento en relación con los niveles de hechos victimizantes. En este sentido, este análisis se enfocó en descubrir patrones frecuentes y relaciones implícitas en las características de los grupos que tenían los niveles más altos de hechos victimizantes.

Aunque se exploraron reglas para múltiples escenarios, los hallazgos principales provienen de los análisis realizados para el escenario con el conjunto completo de variables y, de manera específica, para el subescenario relacionado con la presencia de grupos armados. Se identificaron reglas de asociación para grupos formados para los años 2018 y 2022, ya que estos años registran los niveles más altos de hechos victimizantes en el período analizado; además, se identificaron reglas de asociación para el grupo formado para el período presidencial de Juan Manuel Santos (2015 -2018), esto debido a que para el resto de los escenarios no se identificaron reglas de asociación significativas. Tras identificar las reglas de asociación, se observó que todas compartían los mismos valores en cuanto a las métricas de soporte, confianza y elevación. En consecuencia, se consideraron como equivalentes aquellas reglas en las que el consecuente y el antecedente eran inversos, eliminando así los duplicados.

Para el año 2018 se identificaron un total de 1.303 reglas de asociación para el grupo del escenario con el conjunto total de variables que tenía el mayor promedio de hechos victimizantes (grupo 2). Sin embargo, tras la eliminación de las reglas duplicadas y un análisis de tipo cualitativo, se conservaron un total de 15 reglas significativas para el análisis (Ver Tabla 8). Las reglas de asociación generadas presentan un soporte de 0,33, lo que indica que la relación entre el antecedente y el consecuente se cumple en el 33% de las observaciones analizadas; esto demuestra que las asociaciones no son inusuales dentro del conjunto de datos. Por otra parte, la confianza, con un valor de 1, sugiere que siempre que está presente el antecedente, también lo está el consecuente. Esto señala una relación extremadamente fuerte entre ambos, aunque no implica causalidad. Finalmente, la elevación con un valor de 2 indica que la presencia del antecedente duplica la probabilidad de que ocurra el consecuente. La Tabla 8 presenta una descripción detallada de las reglas seleccionadas.

Se encuentra una relación directa entre la presencia de AGC y Disidencias de las FARC, indicando que en los municipios donde hay presencia de las AGC. Por otro lado, también se encuentra que la presencia de ELN y AGC está asociada con la ocurrencia de hechos victimizantes y la presencia de Disidencias de las FARC, lo que sugiere una relación significativa entre la presencia de estos tres grupos y la incidencia de la violencia. Asimismo, se evidenció una asociación entre la ocurrencia de hechos victimizantes y la presencia de AGC con la presencia de Disidencias de las FARC y proximidad a la frontera. Esto indica que la violencia tiende a intensificarse en las zonas fronterizas donde operan tanto las AGC como las Disidencias de las FARC. Por último, se identificaron reglas más complejas, como aquella que muestra que la ocurrencia de hechos victimizantes en conjunto con la presencia de AGC, ELN y EPL, está asociada con la presencia de Disidencias de las FARC, la proximidad a la frontera, los cultivos de coca y las actividades de erradicación. De manera similar, se observó que estos mismos antecedentes están relacionados con la presencia de Disidencias de las FARC, enfrentamientos, cultivos de coca y erradicación.

Antecedentes	Consecuente	Soporte	Confianza	Elevación
AGC	Disidencia FARC	0,33	1	2
ELN/ AGC	Hechos victimizantes/ Disidencia FARC	0,33	1	2
Hechos victimizantes/ AGC	EPL/ Disidencia FARC	0,33	1	2
Hechos victimizantes/ AGC	Disidencia FARC/ Proximidad a la frontera	0,33	1	2

Hectáreas de coca/ AGC	Hechos victimizantes/ Disidencia FARC	0,33	1	2
Dstrucción de laboratorios / AGC	Hechos victimizantes/ Disidencia FARC	0,33	1	2
Enfrentamientos/ AGC	Hechos victimizantes/ Disidencia FARC	0,33	1	2
AGC/ ELN/ EPL/ Hechos victimizantes	Disidencia FARC/ Proximidad a la frontera/ Erradicación ha de coca/ Hectáreas de coca	0,33	1	2
AGC/ ELN/ EPL/ Hechos victimizantes	Disidencia FARC/ Enfrentamientos/ Erradicación Hectáreas de coca/ Hectáreas de coca	0,33	1	2
AGC/ ELN/ EPL/ Hechos victimizantes	Disidencia FARC/ Dstrucción de laboratorios / Proximidad a la frontera/ Erradicación Hectáreas de coca	0,33	1	2
AGC/ ELN/ EPL/ Hechos victimizantes	Disidencia FARC/ Dstrucción de laboratorios / Proximidad a la frontera/ Enfrentamientos	0,33	1	2
AGC/ ELN/ Hectáreas de coca/ Hechos victimizantes	Disidencia FARC/ Proximidad a la frontera/ Enfrentamientos/ Erradicación Hectáreas de coca	0,33	1	2
AGC/ ELN/ Hectáreas de coca/ Hechos victimizantes	Disidencia FARC/ Dstrucción de laboratorios / Proximidad a la frontera/ Enfrentamientos	0,33	1	2
AGC/ EPL/ Hectáreas de coca/ Hechos victimizantes	Disidencia FARC/ Proximidad a la frontera/ Enfrentamientos/ Erradicación Hectáreas de coca	0,33	1	2
AGC/ EPL/ Hectáreas de coca/ Hechos victimizantes	Disidencia FARC/ Dstrucción de laboratorios / Proximidad a la frontera/ Enfrentamientos	0,33	1	2

Tabla 8. Reglas de asociación para el Grupo 2 del año 2018 del escenario con el conjunto total de variables.

Para el año 2022, se obtuvieron un total de tres reglas de asociación para el grupo del escenario con las variables de presencia de grupos armados conformado por los municipios con mayores promedios de hechos victimizantes (grupo 3) (Ver Tabla 9). Estas reglas presentan un soporte de 0,25, una confianza de 0,67 y una elevación de 1,78, reflejando asociaciones moderadamente significativas. Si bien el soporte es más bajo que en las reglas generadas para el año 2018, la

frecuencia de estas reglas sigue siendo suficiente para considera que representan patrones consistentes en los datos. Las reglas encontradas muestran una asociación directa entre la presencia de Disidencias de las FARC y ELN. Por otro lado, la ocurrencia de hechos victimizantes y la presencia de Disidencias de las FARC está asociada con la presencia de ELN, lo que indica una relación entre la incidencia de la violencia y la presencia de estos dos grupos para ese año.

Antecedente	Consecuente	Soporte	Confianza	Elevación
Disidencia FARC	ELN	0,25	0,67	1,78
Hechos victimizantes/ Disidencia FARC	ELN	0,25	0,67	1,78
ELN	Hechos victimizantes/ Disidencia FARC	0,25	0,67	1,78

Tabla 9. Reglas de asociación para las variables del Grupo 2 del año 2022 del escenario de la presencia de grupos armados

En cuanto al período presidencial de Juan Manuel Santos, se identificaron 13 reglas de asociación para el grupo del escenario con todo el conjunto de variables que presentaba los mayores niveles de violencia (grupo 2). El soporte para el conjunto de reglas es de 0.21, reflejando una frecuencia moderada pero significativa en el contexto analizado. Por otro lado, la confianza es igual a 1 y la elevación igual a 2, lo que muestra que las reglas identificadas presentan patrones consistentes en los datos. En la Tabla 10 se presentan las reglas identificadas para este conjunto de datos.

Se encuentra que hay una relación directa entre la presencia de Disidencias de las FARC y la ocurrencia de enfrentamientos, lo que resalta a este grupo armado como un factor significativo en la intensificación del conflicto. Asimismo, se identificaron reglas más complejas que muestran una asociación entre la presencia de EPL, Disidencias de las FARC y ELN con la ocurrencia de hechos victimizantes y enfrentamientos. Esto sugiere que en las áreas donde confluyen múltiples grupos armados se experimenta una mayor incidencia de la violencia. Además, se encontró una relación significativa entre la ocurrencia de hechos victimizantes y la presencia de grupos como las FARC, ELN y EPL, con la presencia de cultivos de coca, destrucción de laboratorios, la ocurrencia de enfrentamientos, erradicación de cultivos y proximidad a la frontera. Este hallazgo indica que la presencia de múltiples actores armados,

la existencia de economías ilícitas, y la proximidad a la frontera, son factores relevantes en la intensificación del conflicto.

Antecedente	Consecuente	Soporte	Confianza	Elevación
Disidencia FARC	Enfrentamientos	0,21	1	2
Disidencia FARC/ ELN	Hechos victimizantes/ Enfrentamientos	0,21	1	2
EPL/ Disidencia FARC/ ELN	Hechos victimizantes/ Enfrentamientos	0,21	1	2
Hectáreas de coca/ Disidencia FARC/ ELN	Hechos victimizantes/ Enfrentamientos	0,21	1	2
Proximidad a la frontera/ Disidencia FARC/ ELN	Hechos victimizantes/ Enfrentamientos	0,21	1	2
Proximidad a la frontera/ Disidencia FARC/ ELN	Hechos victimizantes/ Enfrentamientos/ EPL	0,21	1	2
Hectáreas de coca/ EPL/ Disidencia FARC/ ELN	Hechos victimizantes/ Enfrentamientos	0,21	1	2
Hechos victimizantes/ Disidencia FARC/ EPL/ ELN	Hectáreas de coca/ Destrucción de laboratorios/ Erradicación Hectáreas de coca/ Enfrentamientos	0,21	1	2
Hechos victimizantes/ Disidencia FARC/ EPL/ ELN	Hectáreas de coca/ Enfrentamientos/ Erradicación de ha de coca/ Proximidad a la frontera	0,21	1	2
Hechos victimizantes/ Disidencia FARC/ EPL/ ELN	Hectáreas de coca/ Destrucción de laboratorios/ Enfrentamientos/ Proximidad a la frontera	0,21	1	2
Hechos victimizantes/ Disidencia FARC/ EPL/ ELN	Destrucción de laboratorios/ Enfrentamientos/ Erradicación ha de coca/ Proximidad a la frontera	0,21	1	2
Hectáreas de coca/ Hechos victimizantes/ Disidencia FARC/ EPL	Destrucción de laboratorios/ Enfrentamientos/	0,21	1	2

	Erradicación de ha de coca/ Proximidad a la frontera			
Hechos victimizantes/ EPL/ ELN/ Hectáreas de coca/ Disidencia FARC	Destrucción de laboratorios/ Enfrentamientos/ Erradicación ha de coca/ Proximidad a la frontera	0,21	1	2

Tabla 10. Reglas de asociación del Grupo 2 del segundo período presidencial de Juan Manuel Santos del escenario con todo el conjunto de variables.

6.2.1 Discusión general sobre las reglas de asociación

En la Figura 3 se presenta el gráfico de la red de asociación para las reglas del grupo 2 correspondiente al escenario del año 2018. Se encuentra que el nodo central es el relacionados con la presencia de AGC, ELN, EPL y ocurrencia de hechos victimizantes. Este nodo tiene múltiples conexiones, lo que indica que es significativo en las relaciones que se forman. El segundo nodo con mayor número de conexiones se conforma por las variables relacionadas con la presencia de Disidencias de las FARC, destrucción de laboratorios, proximidad a la frontera y enfrentamientos. Este resultado muestra la importancia de estas variables en la incidencia de la violencia para ese año.

Por otro lado, en la

Figura 4 se presenta el gráfico de la red de asociación correspondiente a las reglas del grupo 2 formado para el escenario del segundo período presidencial de Juan Manuel Santos. En este caso, uno de los nodos más relevantes es el asociado a la ocurrencia de hechos victimizantes y la presencia de Disidencias de las FARC, EPL y ELN. Este nodo al ocupar una posición central en la red se encuentra conectado a múltiples nodos, lo que indica la importancia de este conjunto de variables dentro de los patrones de conflicto armado identificados. Es importante resaltar que el segundo período presidencial de Juan Manuel Santos cubre tanto el período anterior a la firma del acuerdo (2014-2016) como al período posterior (2017-2018). Por lo tanto, teniendo en cuenta que las Disidencias de las FARC se consolidaron como grupo armado en el período posterior a la firma del acuerdo, estos resultados estarían abarcando el período 2017 - 2018. Por otro lado, el segundo nodo con mayor número de conexiones es el relacionado con hechos victimizantes y enfrentamientos.

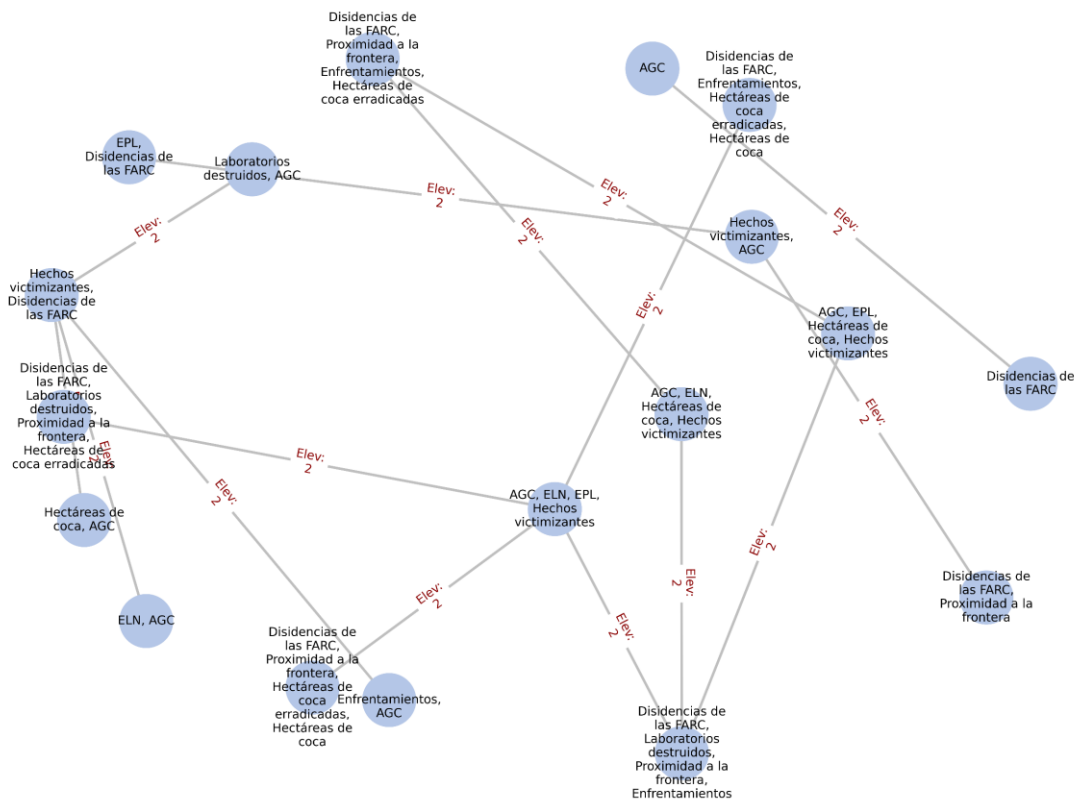


Figura 3. Red de reglas de asociación del año 2018 (Grupo 2)

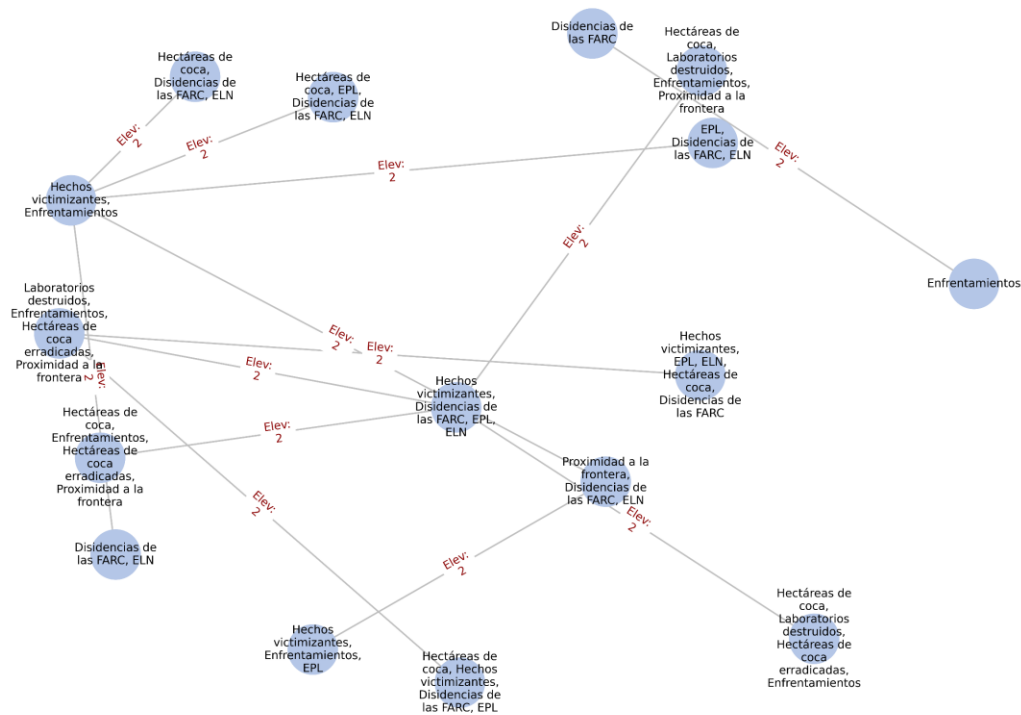


Figura 4. Red de reglas de asociación del segundo período presidencial de Juan Manuel Santos (2014 – 2018) (Grupo)

6.3 Explicabilidad basada en método SHAP

En este apartado se presentan un análisis de los resultados obtenidos mediante la implementación del método SHAP para interpretar los resultados de las técnicas de agrupamiento utilizadas en el análisis. El objetivo principal de esta sección es explicar cómo cada variable contribuye a la configuración de los distintos grupos generados, proporcionando un marco de explicabilidad que facilita la comprensión de las decisiones de las técnicas de agrupamiento.

Para emplear el método SHAP se entrenó un modelo de clasificación supervisado con *random forest*, tomando como variable objetivo las etiquetas asignadas a los grupos como altos o bajos niveles de violencia por el algoritmo de agrupamiento respectivo según la Tabla 7, y como variables explicativas las características originales restantes del conjunto de datos. En esta sección se seleccionó el escenario con el conjunto de todas las variables para cada período analizado, y se optimizaron los hiperparámetros del algoritmo de clasificación. Posteriormente, se utilizaron los valores SHAP para cuantificar el impacto individual de cada característica.

Este enfoque permitió interpretar cómo cada variable contribuyó a la decisión del modelo al replicar las etiquetas de agrupación, proporcionando una explicación detallada mediante la evaluación de la importancia de las características. Específicamente, en este análisis se estudiaron el caso del año 2018, el período posterior a la firma del acuerdo de paz, y el segundo período presidencial de Juan Manuel Santos, el cual coincide con el período anterior y posterior a la firma del acuerdo. Se seleccionaron estos escenarios debido a la relevancia dentro del análisis. En este sentido, el año 2018 presenta uno de los mayores niveles de hechos victimizantes en el período analizado; por otro lado, el período posterior a la firma del acuerdo de paz y el período de gobierno de Juan Manuel Santos permiten identificar las características asociadas a la formación de grupos en un período que trajo consigo importantes transformaciones en los territorios. A continuación, se presentan los resultados obtenidos en cada caso para los grupos que presentan los mayores promedios de hecho victimizantes.

6.3.1 Análisis de explicabilidad para grupos formados para el año 2018

En la Figura 5 se presentan los resultados obtenidos para la clase 1 del grupo del año 2018, esta clase representa los municipios con los mayores promedios de hechos victimizantes para este año. La variable con el mayor promedio SHAP es “presencia de EPL”, esta variable tiene un impacto positivo significativo de 0,38. En segundo lugar se encuentra “destrucción de laboratorios” con 0,024 y “hechos victimizantes” con 0,020. (Ver Anexo F1). Lo anterior indica

que a mayores niveles de estas variables hay una mayor probabilidad de pertenecer a este grupo. Esto evidencia una predominancia de la presencia de EPL en la formación de grupos con altos niveles de hechos victimizantes para este año. Por otra parte, variables como “proximidad a la frontera” y “erradicación de hectáreas de coca”, presencia de grupos como AGC, ELN, y “municipio de frontera” tienen un promedio SHAP de 0, lo que sugiere que no influyen en las asignaciones de los grupos para esta clase (Ver Anexo F1). Este análisis sugiere que este grupo está conformado por municipios donde la presencia del EPL, actividades relacionadas con destrucción de laboratorios y la ocurrencia de hechos victimizantes son predominantes, sin embargo, otros factores de violencia o dinámica fronteriza no son tan relevantes.

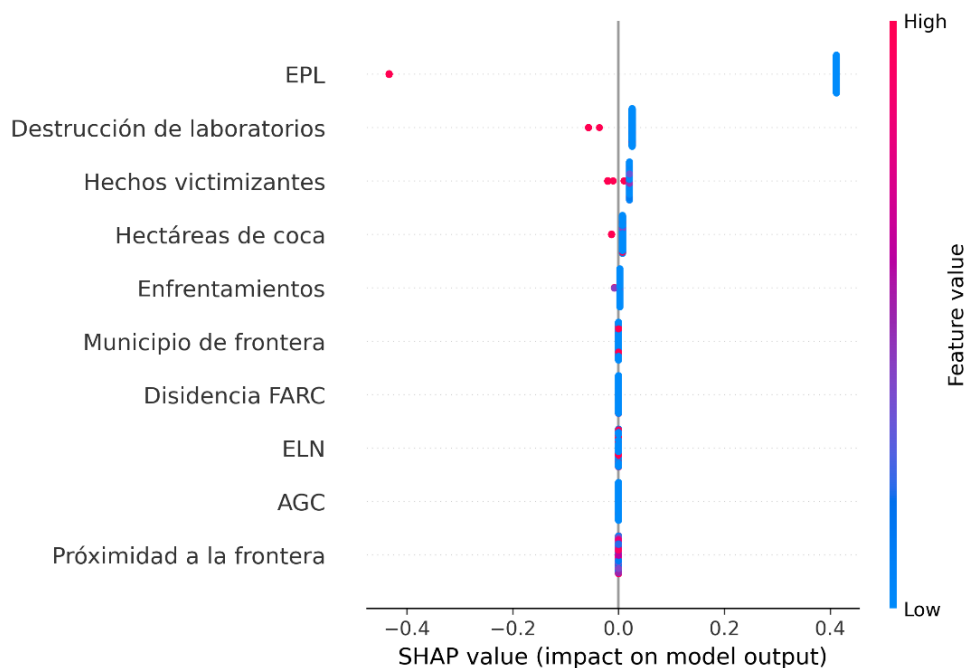


Figura 5. Importancia de cada característica en la formación de la clase 1 del grupo generado para el año 2018.

6.3.2 Análisis de explicabilidad para el período posterior a la firma del acuerdo de paz Los valores SHAP para la clase que presenta mayor promedio de hechos victimizantes para el período posterior a la firma del acuerdo de paz, se presentan en la Figura . Las variables con mayor impacto son “hectáreas de coca” y “hechos victimizantes”, estas variables tienen un impacto positivo de 0.16 y 0.11, respectivamente. Por otro lado, las variables sobre la presencia de grupos armados como el EPL y AGC, en conjunto con “destrucción de laboratorios”, tienen

un impacto moderado. Finalmente, las variables de “proximidad a la frontera” y “municipio de frontera”, entre otros, presentan impactos leves, y las variables relacionadas con “presencia de FARC” y “presencia de Disidencias de las FARC” no tienen relevancia para este grupo (Ver Anexo F2).

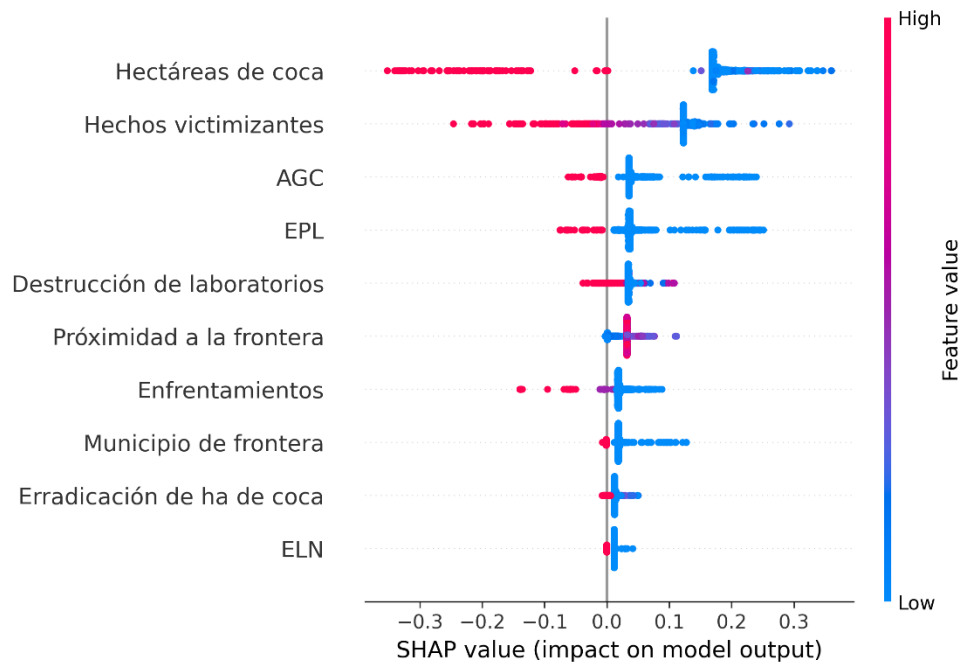


Figura 6. Importancia de cada característica en la formación de la clase 1 del grupo generado para el período posterior al acuerdo de paz (2017-2022)

6.3.3 Análisis de explicabilidad para el segundo período presidencial de Juan Manuel Santos (2014-2018)

Los valores SHAP para la clase con mayores niveles de hechos victimizantes (clase 1) del segundo período presidencial de Juan Manuel Santos se presentan en la Figura 7. Las variables con mayor impacto promedio son la “presencia de EPL”, la “destrucción de laboratorios” y la “hectáreas de coca”, con un SHAP promedio de 0,16, 0,10 y 0,07, respectivamente. Estos valores indican que estas variables tienen un impacto positivo sobre los resultados de la formación de grupos. Por otro lado, las variables de “hechos victimizantes”, “proximidad a la frontera”, y “erradicación de coca” tienen un impacto positivo leve, con valores promedio de 0,06. Finalmente, las variables de “presencia de AGC”, “presencia de ELN” y “municipio de frontera”, entre otros, tienen un promedio SHAP igual o casi 0, lo que indica que no tienen un impacto significativo en la construcción del grupo (Ve Anexo F3).

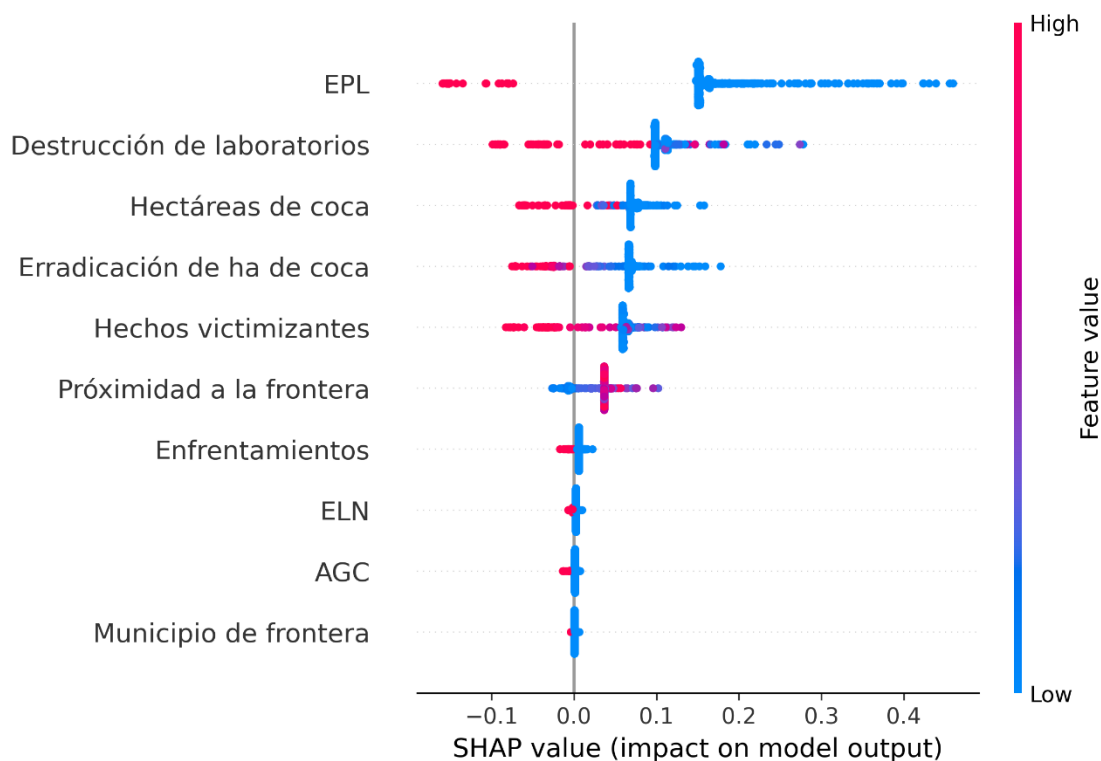


Figura 7. Importancia de cada característica en la formación de la clase 1 del grupo generado para el segundo período presidencial de Juan Manuel Santos (2014-2018)

6.3.4 Discusión general sobre los resultados del método SHAP

En conclusión, los análisis de explicabilidad realizados para los grupos formados en los escenarios que agrupan todo el conjunto de variables, destacan la predominancia de las variables relacionadas con economías ilícitas, como “hectáreas de coca”, “erradicación de cultivos” y “destrucción de laboratorios”, en la formación de grupos con mayores niveles de hechos victimizantes para el período posterior a la firma del acuerdo. Para el año 2018 y el segundo período presidencial de Juan Manuel Santos, la presencia de grupos como el EPL presenta el mayor impacto. Por otro lado, variables relacionadas con la proximidad a la frontera y presencia de grupos armados como el ELN y las AGC mostraron impactos positivos muy leves, mientras que las variables asociadas a la presencia de las FARC y sus disidencias no tuvieron relevancia en la asignación de grupos. Este análisis sugiere que las actividades vinculadas a cultivos ilícitos son un factor determinante en la configuración de los patrones de violencia para el período posterior a la firma del acuerdo, mientras que otros factores, como dinámicas fronterizas o presencia de ciertos grupos armados (con la excepción de EPL), tienen un impacto limitado en estos escenarios.

6.4 Discusión general sobre los resultados

6.4.1 Resumen de los hallazgos obtenidos

Los resultados obtenidos en este estudio proporcionan una visión integral de los patrones de violencia armada en los municipios de la frontera colombo-venezolana, revelando *patrones* distintivos entre municipios con altos y bajos niveles de violencia, así como de interacción entre las variables. En este sentido, los resultados obtenidos a través del análisis de los diferentes escenarios y los *patrones* que aparecen, muestran que los municipios con los niveles más altos de hechos victimizantes comparten características específicas, como ser municipios de frontera o estar cercanos a los pasos fronterizos, presentan áreas extensas de cultivos de coca y registran una presencia significativa de múltiples actores armados, principalmente del ELN, FARC y AGC para el período anterior a la firma del acuerdo de paz y del ELN, Disidencia FARC y AGC para el período posterior a la firma del acuerdo. A pesar de que se observa una reducción en la influencia de grupos como las FARC, tras su reorganización en Disidencias, se evidencia un aumento en la influencia del ELN, lo cual ha desencadenado en un incremento en los niveles de hechos victimizantes para el grupo de municipios con altos niveles de hechos victimizantes tras la firma del acuerdo si se compara con el período anterior a la firma. De igual manera, se evidencia un patrón de concentración de la violencia, donde un porcentaje menor de municipios presenta niveles críticos de hechos victimizantes, mientras que la mayoría registra niveles bajos.

Por otro lado, el análisis de *reglas de asociación* permitió identificar patrones relevantes sobre la relación entre variables en los grupos de municipios con altos niveles de hechos victimizantes para los años 2018 y 2022, y el segundo período presidencial de Juan Manuel Santos. En este sentido, en el año 2018 se evidenció una fuerte asociación entre la ocurrencia de hechos victimizantes, la presencia de grupos como las AGC, el ELN y el EPL, con la presencia de Disidencias de las FARC, la proximidad a la frontera, los cultivos de coca y las actividades de erradicación. Asimismo, tras el análisis del gráfico de red se encuentra que el nodo conformado por la presencia de AGC, el ELN, EPL y ocurrencia de hechos victimizantes es central y tiene el mayor número de conexiones, lo que indica que estos elementos son significativos en las relaciones para este año. En el caso del escenario con las variables de presencia de grupos en el año 2022, se destaca como resultado relevante que la ocurrencia de hechos victimizantes está asociada con la presencia de las Disidencias de las FARC y del ELN. Por otro lado, durante el segundo período presidencial de Juan Manuel Santos, se identifica una relación significativa entre la ocurrencia de hechos victimizantes y la presencia de grupos armados como las FARC,

ELN y EPL, junto con factores como la existencia de cultivos de coca, la destrucción de laboratorios, la ocurrencia de enfrentamientos, la erradicación de cultivos y la proximidad a la frontera. Asimismo, el análisis del gráfico de la red de asociación revela que el nodo central, compuesto por la ocurrencia de hechos victimizantes y la presencia de las FARC, ELN y EPL, presenta múltiples conexiones con otras variables, lo que subraya la importancia de este conjunto de factores en la configuración de los patrones de conflicto armado durante este período.

Finalmente, el *análisis de explicabilidad* con el *método SHAP* desarrollado para los grupos formados en los escenarios que agrupan todo el conjunto de variables a nivel de año y para el segundo período presidencial de Juan Manuel Santos, destacan la predominancia de la presencia de EPL, la destrucción de laboratorios y la ocurrencia de hechos victimizantes en la formación de grupos con altos niveles de hechos victimizantes, siendo la presencia de EPL la variable con mayor relevancia. Por otro lado, para el período posterior a la firma del acuerdo de paz, se destaca el impacto de las variables relacionadas con economías ilícitas, como “hectáreas de coca”, “erradicación de cultivos” y “destrucción de laboratorios”.

Cada uno de los escenarios analizados presentan resultados diversos en relación con la importancia de las características y reglas de asociación identificadas. Esto sugiere que la incidencia de la violencia en diferentes períodos está influenciada por una variedad de variables, lo que refleja la naturaleza multicausal y compleja de los conflictos. En la búsqueda de sintetizar los resultados obtenidos, a continuación, se presenta un cuadro comparativo de los resultados obtenidos para el escenario con todo el conjunto de variables en los diferentes períodos analizados para los grupos de municipios con altos niveles de hechos victimizantes (Ver Tabla 11). Los resultados reflejan que la violencia en los municipios analizados está impulsada principalmente por la presencia de grupos armados y economías ilícitas, aunque la intensidad y variables más influyentes varían entre períodos. Variables como “hectáreas de coca”, “destrucción de laboratorios” y “erradicación de cultivos” son recurrentes en los tres escenarios analizados, evidenciando una relación directa entre las actividades económicas ilícitas y la incidencia del conflicto armado. Por otro lado, se observa una relación significativa entre la presencia de grupos como el EPL, ELN y Disidencias de las FARC con la ocurrencia de hechos victimizantes, destacando la influencia territorial y operativa de estos actores en los municipios más afectados. En particular, la presencia de EPL y la destrucción de laboratorios destacan como variables clave durante el año 2018 y el segundo período presidencial de Juan Manuel Santos. Finalmente, en el período posterior a la firma del acuerdo de paz, las hectáreas de coca y hechos victimizantes ganan protagonismo, reflejando una expansión de cultivos ilícitos y una posible reconfiguración del control territorial tras el acuerdo. A pesar de no ser considerada una variable directamente influyente en la formación de grupos, la proximidad a

la frontera es una constante en los patrones de caracterización de los municipios con altos niveles de violencia. En conjunto, estos resultados confirman la naturaleza multicausal del fenómeno de violencia en los municipios fronterizos, donde interactúan factores territoriales, económicos y la presencia de múltiples actores armados, cuya influencia varía en función del contexto temporal y político.

Escenario	Valores de las variables de los centroides	Reglas de asociación	Variables más influyentes (SHAP)
Todas las variables año 2018	<ul style="list-style-type: none"> • Grupo 2 (con el 2,7% de municipios): "Hechos victimizantes =6.080; "AGC" = 0,3; "Disidencia FARC" = 0,5; ELN = 1,0; "FARC" = 0,0; "EPL" = 1,0; "Hectáreas de coca" = 4.389; "Erradicación de ha de coca" = 165,5; "Destrucción de laboratorios" = 94,0; "Municipio de frontera" = 0,7; "Distancia mínima" = 71,5; Enfrentamientos = 5,0 	<ul style="list-style-type: none"> • Si ocurrencia de hechos victimizantes, AGC Presente, ELN Presente y EPL Presente, entonces Disidencias de las FARC Presente, Proximidad a la frontera, cultivos de coca y actividades de erradicación. 	<ul style="list-style-type: none"> • Presencia de EPL • Destrucción de laboratorios • Hechos victimizantes
Todas las variables segundo período presidencial de Juan Manuel Santos	<ul style="list-style-type: none"> • Grupo 2 (con el 1,6% de municipios): "Hechos victimizantes =3.102; "AGC" = 0,36; "Disidencia FARC" = 0,21; ELN = 0,86; "FARC" = 0,07; "EPL" = 1,0; "Hectáreas de coca" = 4.357; "Erradicación de ha de coca" = 132,5; "Destrucción de laboratorios" = 106,4 "Municipio de frontera" = 0,57; "Distancia mínima" = 71,72; Enfrentamientos = 2,21 	<ul style="list-style-type: none"> • Si ocurrencia de hechos victimizantes, Disidencia FARC Presente, ELN Presente y EPL Presente, entonces cultivos de coca, destrucción de laboratorios, enfrentamientos, erradicación de cultivos y proximidad a la frontera 	<ul style="list-style-type: none"> • Presencia de EPL • Destrucción de laboratorios • Hectáreas de coca

Todas las variables período posterior a la firma del acuerdo de paz	<ul style="list-style-type: none"> • Grupo 2 (con el 1,8% de municipios): "Hechos victimizantes = 2.559; "AGC" = 0,64; "Disidencia FARC" = 0,57; ELN = 0,96; "FARC" = 0,0; "EPL" = 0; "Hectáreas de coca" = 4.656; "Erradicación de ha de coca" = 474,76; "Destrucción de laboratorios" = 83,86; "Municipio de frontera" = 0,68; "Distancia mínima" = 68,6; Enfrentamientos = 2,2 	Sin resultados significativos	<ul style="list-style-type: none"> • Hectáreas de coca • Hechos victimizantes
---	---	-------------------------------	---

Tabla 11. Cuadro resumen de resultados relevantes para escenario con todas las variables en diferentes períodos.

En conclusión, los resultados de este estudio evidencian la complejidad y la evolución de los patrones de violencia armada en los municipios de la frontera colombo-venezolana, resaltando la influencia de factores estructurales y contextuales como la presencia de actores armados, la proximidad a la frontera y las dinámicas asociadas a las economías ilícitas. La identificación de variables predominantes, como la presencia del EPL, la destrucción de laboratorios y las hectáreas de coca, y su relación con los hechos victimizantes, permite comprender cómo estos elementos configuran los niveles críticos de violencia, especialmente en períodos y escenarios específicos. Además, el análisis comparativo entre el período previo y posterior al acuerdo de paz revela cambios significativos en la dinámica de influencia de los grupos armados, destacando el papel creciente del EPL, ELN y las Disidencias de las FARC en el incremento de la violencia. Estos hallazgos no solo aportan una visión detallada de los factores asociados a la concentración de la violencia en determinados municipios, sino que también proporcionan información clave para el desarrollo de políticas públicas y estrategias focalizadas que permitan mitigar las dinámicas del conflicto en regiones fronterizas.

6.4.2 Análisis estadístico de los resultados

Para reforzar estos hallazgos con evidencia estadística, se realizaron pruebas de hipótesis que validan la *relación de cercanía a la frontera, la presencia de grupos armados y la expansión de economías ilícitas con un mayor número de hechos victimizantes*. Estas pruebas permiten validar las hipótesis formuladas en la Sección 4.1 de este documento. Para seleccionar los métodos estadísticos más adecuados para el análisis, se realizó previamente una prueba de normalidad para las variables "Hechos victimizantes", "Distancia mínima", "Hectáreas de coca" y "Número de grupos" (esta variable fue calculada sumando el número de grupos presentes en cada municipio). Así, se realizó la prueba de Shapiro-Wilk (ver Tabla 12), que se usa para contrastar la normalidad de un conjunto de datos. Los resultados de la prueba muestran que ninguna de las variables

analizadas es normal, por lo tanto, se seleccionaron pruebas no paramétricas para la validación de las hipótesis.

Variable	p-valor	Normalidad
Hechos victimizantes	7,71e-69	No normal
Distancia mínima	1,88e-24	No normal
Hectáreas de coca	7,35e-72	No normal
Número de grupos	7,36e-59	No normal

Tabla 12. Prueba de normalidad Shapiro- Wilk

A continuación, se presentan las pruebas planteadas para cada una de las hipótesis.

H1: Los municipios más cercanos a la frontera internacional tienden a presentar mayor incidencia de la violencia: los resultados obtenidos por las técnicas de agrupamiento muestran que los municipios que están más cercanos a la frontera presentan mayores niveles de hechos victimizantes. Para validar estadísticamente estos resultados se planteo la siguiente prueba de hipótesis para la variable “Distancia mínima”:

H₀ : La distancia a la frontera no está relacionada con la incidencia de la violencia.

H₁ : Los municipios más cercanos a la frontera presentan mayores niveles de violencia.

Para validar esta prueba de hipótesis se selecciono el coeficiente de Spearman, dado que la distribución de los datos no cumplía con los supuestos de normalidad. El *p-valor* fue obtenido a partir de la distribución t de Student para la correlación rangos, lo que permite evaluar la significancia de la correlación. El análisis arrojó un coeficiente de Spearman de -0,59 con un *p-valor* de 7,26e-234. Dado que el *p-valor* es menor a 0.05, se rechaza la hipótesis nula, lo que indica que existe una correlación negativa y estadísticamente significativa entre las variables analizadas. Es decir, a media que aumenta la distancia a la frontera, el número de hechos victimizantes tiende a disminuir.

De igual manera, se plantearon las siguientes hipótesis para la variable “Municipio de frontera”:

H₀ : No hay diferencia significativa en la incidencia de violencia entre municipios de frontera y no frontera.

H₁: Los municipios de frontera presentan una mayor incidencia de violencia en comparación con los municipios que no son de frontera.

Para validar esta hipótesis se aplicó una prueba Mann-Whitney U, una prueba no paramétrica adecuada para compara distribuciones cuando los datos no siguen una distribución normal. Como resultado de la prueba se encuentra que el *p-valor* es igual a 6.05e-141. Dado que el *p-valor* es significativamente menor que 0.05 se rechaza la hipótesis nula, lo que indica que existe una diferencia estadísticamente significativa en la incidencia de hechos victimizantes entre municipios de frontera y no frontera. Estos resultados refuerzan la evidencia de que la proximidad a la frontera está asociada con una mayor incidencia de la violencia, lo que concuerda con los hallazgos previos encontrados en el análisis de agrupamiento.

H2: Hay una concurrencia ente municipios con violencia más crítica y la expansión de economías ilícitas como los cultivos de coca: para validar esta hipótesis se dividió el conjunto de datos entre municipios con alta y baja violencia por medio del percentil 75. Posteriormente, se planteo la siguiente prueba de hipótesis:

H₀: No hay diferencia significativa en la cantidad de hectáreas de coca entre los municipios con alta y baja violencia.

H₁: Los municipios con alta violencia tienen una cantidad significativamente mayor de hectáreas de coca en comparación con los municipios con baja violencia.

Teniendo en cuenta que las variables no son normales se evaluó la prueba de Mann-Whitney U, obteniendo un *p-valor* de 5.78e-128. Dado que el *p-valor* es significativamente menor que 0.05, se rechaza la hipótesis nula, lo que indica que existe una diferencia estadísticamente significativa en la cantidad de hectáreas de coca entre municipios de alta y baja violencia. Este resultado refuerza los hallazgos realizados por las técnicas de agrupamiento donde se logró identificar que los municipios con altos niveles de violencia también tienden a presentar una mayor cantidad de hectáreas de coca.

H3: En los municipios con un mayor número de grupos armados activos hay un mayor nivel de violencia: para comprobar esta hipótesis, se clasificaron los municipios en dos grupos (alta y baja violencia), utilizando el percentil 75 como umbral. A partir de la segmentación, se formuló la siguiente prueba de hipótesis:

H₀: No hay diferencia significativa en el número de grupos armados activos entre municipios con alta y baja violencia.

H₁: Los municipios con alta violencia tienen un mayor número de grupos armados activos en comparación con los municipios con baja violencia.

Para evaluar si los municipios con alta violencia tienen un mayor número de grupos armados activos en comparación con los municipios con baja violencia, se aplicó la prueba de Mann-Whitney U. Como resultado de la prueba se obtuvo un *p-valor* de $1.51e-222$, dado que el *p-valor* es menor que 0.05, se rechaza la hipótesis nula. Es decir, los municipios con alta incidencia de la violencia presentan una diferencia significativa mayor de grupos armados que los que tienen baja incidencia de la violencia. Estos resultados respaldan la hipótesis de que los municipios con más grupos armados presentan una mayor incidencia de hechos victimizantes, lo cual es consistente con los resultados de las técnicas de agrupamiento en los diferentes escenarios.

7. Comparación con trabajos previos

En esta sección se presenta un análisis comparativo entre el enfoque desarrollado en el presente trabajo y otros estudios previamente realizados en el campo del análisis de conflictos armados y violencia. Para ello se seleccionaron una serie de criterios de comparación que reflejan aspectos clave en el desarrollo de este trabajo, los cuales son:

- A) Estudian patrones de violencia.
- B) Hacen un análisis de explicabilidad sobre los resultados obtenidos.
- D) Usan diversas fuentes de datos

En la Tabla 12 se presenta un cuadro resumen que compara los criterios establecidos con trabajos previos desarrollados. Se observa que los estudios desarrollados por [10], [15], [21], [25], [29], [31] y [32] cumplen con el criterio de análisis de patrones de violencia. En particular, [10] utiliza el algoritmo k-means para definir patrones de violencia y paz en diferentes países. De manera similar, [15] y [21] emplean técnicas de agrupamiento para analizar patrones de eventos violentos. Por otro lado, [29], [31] y [32], utilizan métodos de análisis espacial para identificar patrones de conflictos armado dentro del contexto colombiano. A su vez, los autores de [25] mediante la implementación de herramientas econométricas analiza la relación entre la expansión minera y eventos del conflicto armado. Finalmente, nuestra propuesta hace un análisis de los diversos patrones de violencia considerando una multiplicidad de escenarios espacio-temporales.

En relación con los criterios vinculados al análisis de explicabilidad, se observa que ninguno de los trabajos analizados cumple con estos criterios. Solamente nuestro trabajo mezcla varias técnicas (reglas de asociación y método SHAP) para comprender las variables más influyentes en los patrones descubiertos. Por otro lado, en cuanto a la integración de diversas fuentes de información, el estudio [10] utiliza múltiples fuentes como Our World in Data y World Bank

para obtener variables relevantes como el PIB per cápita, el gasto militar, la inflación, entre otros. Asimismo, [31] utiliza datos del sistema Noche y Niebla de CINEP que incluye narrativas verificadas de violaciones de derechos humanos y eventos violentos en Colombia, recolectadas de múltiples fuentes como periódicos y testimonios. En nuestro caso, también usamos diferentes fuentes de datos (para más detalles, ver sección 4.1).

Trabajos vs criterios	Estudian patrones de violencia (A)	Análisis de explicabilidad (B)	Diversas fuentes de datos (D)
[10]	X		X
[15]	X		
[21]	X		
[25]	X		
[29]	X		
[31]	X		X
[32]	X		
Presente trabajo	X	X	X

Tabla 12. Cuadro comparativo con trabajos previos.

El análisis comparativo evidencia una brecha importante entre nuestro trabajo y previos enfoques, particularmente, nosotros integramos metodológicamente técnicas de aprendizaje automático con métodos de explicabilidad para interpretar los patrones de violencia descubierto desde perspectivas multidimensionales. Esto es fundamental para poder hacer un análisis cualitativo y diseñar políticas públicas. Por otro lado, la integración de diversas fuentes de información es un punto relevante, lo que subraya la importancia de utilizar datos heterogéneos para enriquecer los análisis.

8. CONCLUSIONES

El análisis de los patrones de violencia armada en los municipios de frontera entre Colombia y Venezuela es un tema relevante en el marco del análisis de los conflictos, teniendo en cuenta el contexto político y social complejo que caracteriza la región. Históricamente, la frontera ha sido escenario de tensiones diplomáticas recurrentes y una marcada vulnerabilidad debido a la falta de control estatal. Esto sumado a la crisis migratoria, la presencia de economías ilícitas, y el

control territorial ejercido por grupos al margen de la ley, han consolidado a la frontera como un epicentro de violencia. En este contexto, resulta fundamental el desarrollo de herramientas analíticas basadas en datos que complementen los enfoques cualitativos tradicionales de las ciencias políticas, que permitan una comprensión más profunda de las dinámicas subyacentes asociadas a la violencia armada, esto de cara a la prevención y resolución de los conflictos.

En este sentido, este trabajo tuvo como objetivo contribuir al campo del análisis de conflictos mediante la implementación de un enfoque basado en datos que complementara los estudios cualitativos tradicionales; integrando métodos de aprendizaje automático orientados a identificar patrones de violencia armada en la frontera colombo-venezolana, junto con métodos de explicabilidad que facilitarán la interpretación de las relaciones entre las múltiples variables asociadas al conflicto.

El estudio logró cumplir con su contribución al combinar técnicas avanzadas de aprendizaje automático, como lo son los métodos de agrupamiento *k-means*, agrupamiento jerárquico, BDSCAN y agrupamiento espectral, con métodos de explicabilidad basados en reglas de asociación y el método SHAP. A través de los algoritmos de agrupamiento, se generaron grupos representativos que reflejan dinámicas de violencia en los municipios de frontera, mientras que el uso de reglas de asociación y el método SHAP permitió identificar relaciones frecuentes entre las variables y las características con mayor impacto dentro de la formación de grupos. Este enfoque metodológico integrador facilitó no solo la detección de patrones de violencia armada asociados a factores como la proximidad a la frontera, la presencia de grupos armados y las economías ilícitas, sino también, la generación de interpretaciones que pueden ser directamente aplicables al diseño de políticas públicas.

Los análisis revelaron que técnicas como *k-means*, agrupamiento jerárquico y DBSCAN fueron más efectivas en la caracterización de los patrones de violencia para diferentes escenarios, siendo el agrupamiento jerárquico el método con mejor desempeño en promedio. La implementación de estos métodos de agrupamiento permitió abordar diferentes escenarios relacionados con la violencia armada en la frontera colombo-venezolana, generando agrupaciones significativas para cada escenario y el contexto temporal analizado. En este sentido, *k-means* demostró ser eficiente para identificar grupos para el escenario con la variable “hechos victimizantes”, presentando el mejor desempeño en la métrica de evaluación (coeficiente de silueta). Por otro lado, DBSCAN destacó en escenarios con variables de proximidad a la frontera, gracias a su capacidad para detectar grupos con formas irregulares y manejar valores atípicos, identificando municipios con características extremas. Por otra parte, el agrupamiento jerárquico fue el método con mejor desempeño en la formación de grupos para

los escenarios con variables relacionadas con economías ilegales y presencia de grupos armados, así como en el análisis del conjunto completo de datos.

En cuanto al resultado obtenido, se observó que los municipios que pertenecen a la frontera o están a una distancia promedio a los puntos fronterizos de 100 km, tienden a presentar mayores niveles de violencia. Asimismo, los municipios que tienen extensas áreas de cultivos de coca y una presencia significativa de múltiples actores armados como el ELN, el AGC y las Disidencias de las FARC, evidencian niveles de hechos victimizantes por encima del promedio. Adicionalmente, el análisis permitió observar que tras la firma del acuerdo de paz se presentó una reducción en los municipios con ocurrencia de hechos victimizantes; sin embargo, se evidencia un aumento en los niveles de violencia para este grupo de municipios, lo que sugiere que la violencia se ha intensificado y concentrado tras la firma del acuerdo en un subconjunto de municipios.

Por otro lado, el análisis de reglas de asociación reveló que los grupos armados como las Disidencias de las FARC, las AGC, el ELN y el EPL, están significativamente relacionadas con la incidencia de hechos victimizantes, especialmente en zonas caracterizadas por presencia de cultivos de coca y actividades de erradicación. Asimismo, la proximidad a la frontera, combinada con la presencia de economías ilícitas y grupos armados, emergió como un factor clave asociado a la ocurrencia de hechos victimizantes. Adicionalmente, el análisis permitió identificar reglas complejas que subrayan la naturaleza multicausal del conflicto en la región, donde factores económicos, geográficos y sociales interactúan para perpetuar los niveles de violencia.

El análisis de explicabilidad realizado para los escenarios que integran el conjunto completo de variables, destaca la predominancia de las variables relacionadas con economías ilícitas y la presencia del EPL en la formación de grupos conformados por municipios con altos niveles promedios de hechos victimizantes. Estas variables mostraron un impacto positivo a lo largo de las diferentes dimensiones temporales analizadas. Este hallazgo, sugiere que, al considerar la totalidad de las variables estudiadas, los factores asociados a la presencia de economías ilícitas tienen un impacto preponderante en los patrones de violencia en la región, posicionándolos como determinantes clave en la configuración de las dinámicas de conflicto. Por otro lado, las variables relacionadas con la proximidad a la frontera y la presencia de grupos como el ELN y las AGC evidenciaron impactos positivos leves, lo que indica que, aunque relevantes, su influencia es menos determinante.

Los resultados obtenidos en este estudio se enmarcan en un contexto político donde la violencia en la frontera colombo-venezolana representa un desafío central para la seguridad y la estabilidad en la región. La concentración de la violencia en municipios fronterizos con

presencia de grupos armados y economías ilícitas refleja la falta de presencia estatal y el limitado control ejercido sobre estos territorios, lo que facilita el accionar de grupos armados y economías criminales. Estos hallazgos subrayan la necesidad de fortalecer políticas de seguridad enfocadas en desarticular las estructuras económicas ilegales y reducir el control territorial de grupos armados. De igual manera, los patrones de violencia identificados tras la firma del acuerdo de paz evidencian la necesidad de ajustar las estrategias de postconflicto para evitar el recrudecimiento y concentración de la violencia.

Por otra parte, el desarrollo de este trabajo enfrenta varias limitaciones. En primer lugar, la calidad y disponibilidad de los datos representa un desafío, especialmente para las variables relacionadas con la presencia de grupos armados en algunos períodos, lo que requirió el cruce de varias fuentes de información, lo cual pudo influir en la calidad de los datos. Además, la alta heterogeneidad espacial de los municipios que pertenecen a la frontera colombo-venezolana y la heterogeneidad temporal del estudio limitan la posibilidad de generalizar los resultados a otros municipios de frontera o áreas fronterizas del país, donde las dinámicas de frontera pueden estar impulsadas por factores completamente distintos. Adicionalmente, se tiene que los conflictos armados y la violencia son fenómenos intrínsecamente complejos, influenciados por una combinación de factores económicos, sociales, culturales, políticos y geográficos. Los patrones identificados a través de la metodología empleada en este trabajo pueden capturar correlaciones válidas en los datos disponibles, pero no necesariamente explican causalidades subyacentes.

Futuras investigaciones podrían enfocarse en ampliar el análisis incluyendo las cinco fronteras del país (Venezuela, Brasil, Perú, Ecuador y Panamá), esto permitiría captar la complejidad de las dinámicas de las zonas fronterizas y generar un análisis a nivel general de los patrones que caracterizan las diferentes zonas de frontera, identificando así diferencias y similitudes. De igual manera, incluir nuevas variables relacionadas con aspectos socioeconómicos como niveles de pobreza, acceso a servicios básicos, educación, empleo, ingresos y desarrollo de infraestructura, permitirían analizar de manera integral cómo otros factores estructurales contribuyen a la configuración de patrones de violencia en los municipios fronterizos. Estas variables también podrían dar lugar a un entendimiento más profundo de las interacciones entre violencia, desigualdad y oportunidades económicas en estas zonas. Asimismo, incluir todos los municipios del país podría ampliar el alcance del estudio para identificar si los patrones observados en las zonas fronterizas tienen características únicas o si, por el contrario, comparten similitudes con otros municipios no pertenecientes a la frontera.

REPOSITORIO

Daniela Lopera, "Análisis de patrones de violencia en municipios de frontera", GitHub, 2025. [En línea]. Disponible en: <https://github.com/Danielaloperap/An-lisis-de-patrones-de-violencia-en-los-municipios-de-frontera>

REFERENCIAS

- [1] V. García Pinzón and J. Mantilla, "Contested borders: organized crime, governance, and bordering practices in Colombia-Venezuela borderlands," *Trends Organ Crime*, vol. 24, no. 2, pp. 265–281, Jun. 2021, doi: 10.1007/s12117-020-09399-3.
- [2] L. Marcela and P. Arias, "Money Laundering: Risks and Socioeconomic Consequences in the Colombo-Venezuelan Border," 2024.
- [3] A. Hernández, "Cross-border Dynamics and Informal Markets between Colombia and Venezuela: The Case of Cúcuta and San Antonio," *J Borderl Stud*, 2023, doi: 10.1080/08865655.2023.2289104.
- [4] L. Yaffe, "Conflicto armado en Colombia: análisis de las causas económicas, sociales e institucionales de la oposición violenta," *Revista CS*, pp. 187–208, Jul. 2011, doi: 10.18046/recs.i8.1133.
- [5] C. A. Aparicio *et al.*, "Análisis y perspectivas Problemas transversales Algunos diagnósticos específicos."
- [6] C. Rojas-Granada and R. Cuesta-Borja, "Los estudios sobre el conflicto armado y la construcción de paz en Colombia desde una perspectiva territorial: abordajes y desafíos," *Revista CS*, no. 33, pp. 205–235, Mar. 2021, doi: 10.18046/recs.i33.3995.
- [7] L. R. Chaparro Moreno and A. M. Jiménez Pava, "Bibliometric study of scientific production on conflict-related sexual violence, with emphasis in Colombia," *Revista Republicana*, vol. 2021, no. 30, pp. 47–71, 2021, doi: 10.21017/rev.repub.2021.v30.a96.
- [8] J. Tappe Ortiz, "100 Years of Solitude Revisited: A Critical Analysis of 25 Years of Scholarship on Colombia's Civil Conflict," *Civil Wars*, vol. 25, no. 2–3, pp. 398–427, 2023, doi: 10.1080/13698249.2023.2249321.

- [9] Q. Ge *et al.*, “Modelling armed conflict risk under climate change with machine learning and time-series data,” *Nat Commun*, vol. 13, no. 1, Dec. 2022, doi: 10.1038/s41467-022-30356-x.
- [10] B. Žmuk and H. Jošić, “k-means clustering of countries according to war activities: common characteristics and differences,” *Proceedings of FEB Zagreb International Odyssey Conference on Economics and Business*, vol. 5 No. 1, no. Proceedings of FEB Zagreb International Odyssey Conference on Economics and Business, pp. 493–505, 2023.
- [11] F. Ettensperger, “Forecasting conflict using a diverse machine-learning ensemble: Ensemble averaging with multiple tree-based algorithms and variance promoting data configurations,” 2022, *Routledge*. doi: 10.1080/03050629.2022.1993209.
- [12] B. J. Radford, “High Resolution Conflict Forecasting with Spatial Convolutions and Long Short-Term Memory *,” 2022. [Online]. Available: <https://dvn.iq.harvard.edu/dvn/dv/internationalinteractions>.
- [13] F. Attinà, M. Carammia, and S. M. Iacus, “Forecasting Change in Conflict Fatalities with Dynamic Elastic Net.”
- [14] M. Hao, J. Fu, D. Jiang, F. Ding, and S. Chen, “Simulating the Linkages Between Economy and Armed Conflict in India With a Long Short-Term Memory Algorithm,” *Risk Analysis*, vol. 40, no. 6, pp. 1139–1150, Jun. 2020, doi: 10.1111/risa.13470.
- [15] S. P. Singha, M. M. Hossain, M. A. Rahman, and N. Sharmin, “Investigation of graph-based clustering approaches along with graph neural networks for modeling armed conflict in Bangladesh,” *Int J Data Sci Anal*, vol. 18, no. 2, pp. 187–203, Aug. 2024, doi: 10.1007/s41060-024-00572-3.
- [16] H. Hegre, H. M. Nygård, and P. Landsverk, “Can We Predict Armed Conflict? How the First 9 Years of Published Forecasts Stand up to Reality,” *International Studies Quarterly*, vol. 65, no. 3, pp. 660–668, Sep. 2021, doi: 10.1093/isq/sqaa094.
- [17] M. D. Ward *et al.*, “Learning from the past and stepping into the future: Toward a new generation of conflict prediction,” 2013, *Blackwell Publishing Inc*. doi: 10.1111/misr.12072.
- [18] Havelange, ““Predicting armed conflicts : a machine learning approach.”” [Online]. Available: <http://hdl.handle.net/2078.1/thesis:30471>
- [19] H. Vard Hegre, J. Karlsen, H. Vard Strand, and H. Urdal, “Predicting Armed Conflict, 2010-2050 1.” [Online]. Available: <http://muparser.source>

- [20] H. Mueller, C. Rauh, and B. Seimon, "Introducing a global dataset on conflict forecasts and news topics," *Data Policy*, vol. 6, Mar. 2024, doi: 10.1017/dap.2024.10.
- [21] J. J. Villar Roldán and J. M. Martín Álvarez, "Comprendiendo la dinámica de los conflictos en américa latina," *Iberoamerican Business Journal*, vol. 7, no. 1, pp. 47–75, Jul. 2023, doi: 10.22451/5817.ibj2023.vol7.1.11076.
- [22] S. Goodman, A. BenYishay, and D. Runfola, "Spatiotemporal Prediction of Conflict Fatality Risk Using Convolutional Neural Networks and Satellite Imagery," *Remote Sens (Basel)*, vol. 16, no. 18, Sep. 2024, doi: 10.3390/rs16183411.
- [23] L. Hotte, L.-P. Morin, M. Mohnen, L.-P. Béland, and P. Jaimes, "Land Conflicts in the Wake of Gold Mining Expansion in Colombia *," 2024.
- [24] S. Bazzi, R. A. Blair, C. Blattman, O. Dube, M. Gudgeon, and R. Peck, "The Promise and Pitfalls of Conflict Prediction: Evidence from Colombia and Indonesia," 2019.
- [25] S. Z. Daly, "Organizational legacies of violence: Conditions favoring insurgency onset in Colombia, 1964-1984," *J Peace Res*, vol. 49, no. 3, pp. 473–491, May 2012, doi: 10.1177/0022343311435801.
- [26] F. Murtagh, M. Spagat, and J. A. Restrepo, "Ultrametric wavelet regression of multivariate time series: Application to Colombian conflict analysis," *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, vol. 41, no. 2, pp. 254–263, Mar. 2011, doi: 10.1109/TSMCA.2010.2064301.
- [27] B. Lepri, Q. Jiang, E. Letouzé, and S. Fiandrino, "Combining environmental and socioeconomic data to understand determinants of conflicts in Colombia."
- [28] G. Camargo, A. M. Sampayo, A. P. Galindo, F. J. Escobedo, F. Carriazo, and A. Feged-Rivadeneira, "Exploring the dynamics of migration, armed conflict, urbanization, and anthropogenic change in Colombia," *PLoS One*, vol. 15, no. 11 November, Nov. 2020, doi: 10.1371/journal.pone.0242266.
- [29] J. Osorio, M. Mohamed, V. Pavon, and S. Brewer-Osorio, "Mapping Violent Presence of Armed Actors in Colombia," *Advances in Cartography and GIScience of the ICA*, vol. 1, pp. 1–9, Jul. 2019, doi: 10.5194/ica-adv-1-16-2019.
- [30] T. Anders, "Territorial control in civil wars: Theory and measurement using machine learning," *J Peace Res*, vol. 57, no. 6, pp. 701–714, Nov. 2020, doi: 10.1177/0022343320959687.

- [31] T. Obukhov and M. A. Brovelli, "Identifying Conditioning Factors and Predictors of Conflict Likelihood for Machine Learning Models: A Literature Review," *ISPRS Int J Geoinf*, vol. 12, no. 8, Aug. 2023, doi: 10.3390/ijgi12080322.
- [32] M. Alba-Niño *et al.*, *Trabajo social: Frontera, flujo migratorio y capitalismo global*. 2020.
- [33] C. Santacruz Bravo *et al.*, "La frontera en la agenda de seguridad: los efectos transversales en el caso colombo - venezolano," *Dossier de Análisis*, 2014, [Online]. Available: <http://cei.uniandes.edu.co/>
- [34] Luis Gómez Calcaño, *La disolución de las fronteras: sociedad civil, representación y política en Venezuela*. Caracas: Universidad Central de Venezuela, 2009.
- [35] R. Linares, "Seguridad y política fronteriza: una mirada a la situación de la frontera entre Venezuela y Colombia," *OPERA*, no. 24, pp. 135–156, Feb. 2019, doi: 10.18601/16578651.n24.08.
- [36] Annette Idler, *Borderland Battles: violence, crime, and governance at the edges of colombias's war*. New York: Oxford Academic, 2019.
- [37] R. Linares, "Seguridad y política fronteriza: una mirada a la situación de la frontera entre Venezuela y Colombia," *OPERA*, no. 24, pp. 135–156, Feb. 2019, doi: 10.18601/16578651.n24.08.
- [38] L. O. Machado, A. R. Novaes, and L. do Rego Monteiro, "Building walls, breaking barriers: Territory, integration and the rule of law in frontier zones," *J Borderl Stud*, vol. 24, no. 3, pp. 97–114, 2009, doi: 10.1080/08865655.2009.9695742.
- [39] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, Jun. 2015, doi: 10.1007/s40745-015-0040-1.
- [40] R. Xu and D. Wunsch, "Survey of clustering algorithms," May 2005. doi: 10.1109/TNN.2005.845141.
- [41] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," 2003. [Online]. Available: www.elsevier.com/locate/patcog
- [42] F. Nielsen, "Hierarchical Clustering," in *Introduction to HPC with MPI for Data Science*, Springer International Publishing, 2016, ch. 8, pp. 195–211. [Online]. Available: <http://www.springer.com/series/7592>

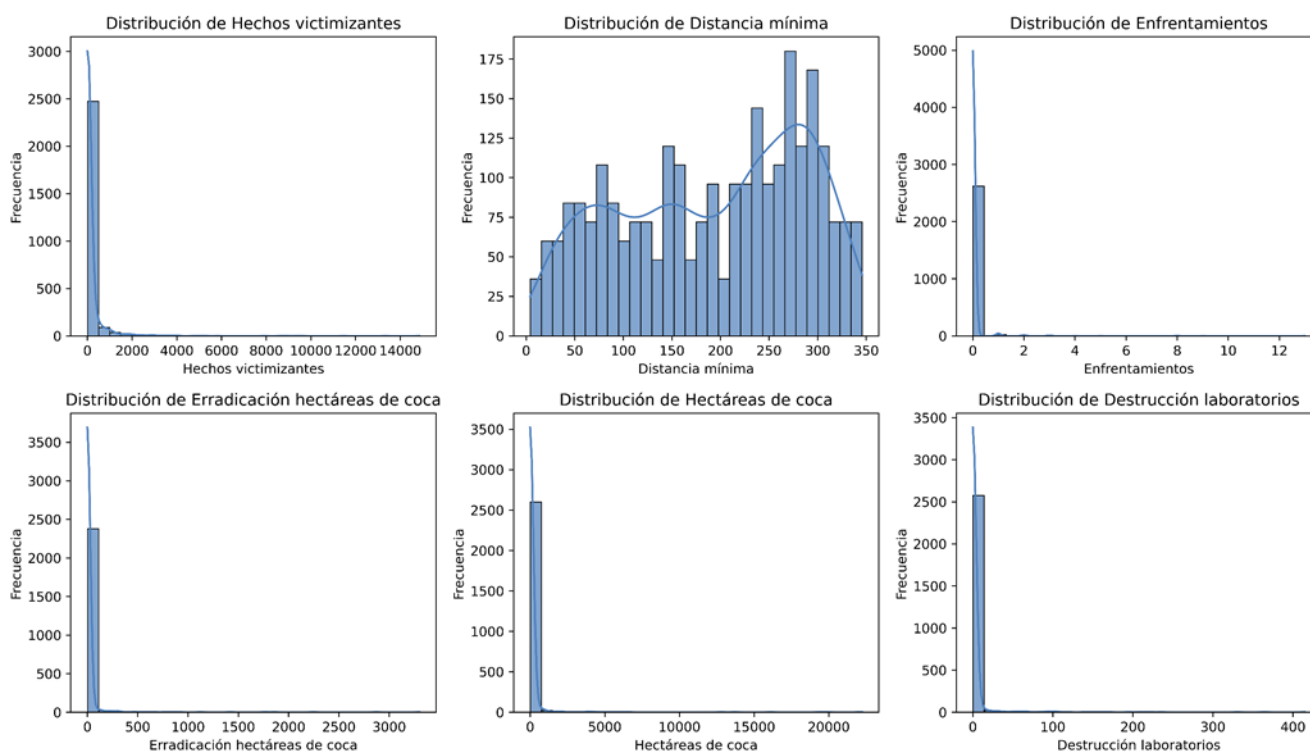
- [43] Reddy Chandan K. and Vinzamuri Bhanukiran, "A Survey of Partitional and Hierarchical Clustering Algorithms," 1st Edition., Aggarwal Charu. C and Reddy Chandan K, Eds., Chapman and Hall/CRC, 2014, pp. 88–107.
- [44] Martin Ester, "Density-Based Clustering," in *Data Clustering. Algorithms and Application*, Aggarwal Charu. C and Reddy Chandan K, Eds., Chapman & Hall/CRC, 2014, ch. 9.
- [45] D. Deng, "DBSCAN Clustering Algorithm Based on Density," in *Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.
- [46] Khan Kamran, Rehman Saif Ur, Aziz Kamran, Fong Simon, and Sarasvady S., "DBSCAN: Past, Present and Future," in *The Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2014)*, Bangalore: IEEE, Feb. 2014.
- [47] M. T. Law, R. Urtasun, and R. S. Zemel, "Deep Spectral Clustering Learning," 2017.
- [48] H. Jia, S. Ding, X. Xu, and R. Nie, "The latest research progress on spectral clustering," 2014, *Springer-Verlag London Ltd*. doi: 10.1007/s00521-013-1439-2.
- [49] U. Von Luxburg, "A tutorial on spectral clustering," *Stat Comput*, vol. 17, no. 4, pp. 395–416, Dec. 2007, doi: 10.1007/s11222-007-9033-z.
- [50] M. Al-Maolegi and B. Arkok, "An Improved Apriori Algorithm For Association Rules," *International Journal on Natural Language Computing*, vol. 3, no. 1, pp. 21–29, Feb. 2014, doi: 10.5121/ijnlc.2014.3103.
- [51] *ICEOE 2011 : Proceedings, 2011 International Conference on Electronics and Optoelectronics : 29-31, July 2011, Dalian, China*. IEEE, 2011.
- [52] J. C. L. Bedoya and J. L. A. Castro, "Explainability analysis in predictive models based on machine learning techniques on the risk of hospital readmissions," *Health Technol (Berl)*, vol. 14, no. 1, pp. 93–108, Jan. 2024, doi: 10.1007/s12553-023-00794-8.
- [53] D. Benito, C. Quintero, J. Aguilar, J. M. Ramírez, and A. Fernández-Anta, "Explainability analysis: An in-depth comparison between Fuzzy Cognitive Maps and LAMDA," *Appl Soft Comput*, vol. 164, Oct. 2024, doi: 10.1016/j.asoc.2024.111940.
- [54] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," Aug. 2019, [Online]. Available: <http://arxiv.org/abs/1908.08474>

- [55] T. Reitano, "Making war: Conflict zones and their implications for drug policy.," *Drug Policies and Development*, Brill | Nijhoff, pp. 127–141, Oct. 28, 2020. doi: 10.1163/9789004440494_008.
- [56] Defensoría del Pueblo, "Informe especial de riesgo: economías ilegales, actores armados y nuevos escenarios de riesgo en el posacuerdo," Bogotá D.C, May 2018.
- [57] Z. Cheng, C. Zou, and J. Dong, "Outlier detection using isolation forest and local outlier," in *Proceedings of the 2019 Research in Adaptive and Convergent Systems, RACS 2019*, Association for Computing Machinery, Inc, Sep. 2019, pp. 161–168. doi: 10.1145/3338840.3355641.
- [58] D. Moulavi, P. A. Jaskowiak, R. J. G. B. Campello, A. Zimek, and J. Sander, "Density-Based Clustering Validation," in *Proceedings of the 2014 SIAM International Conference on Data Mining*, 2014. [Online]. Available: <https://epubs.siam.org/terms-privacy>
- [59] F. Fouedjio, "A spectral clustering approach for multivariate geostatistical data," *Int J Data Sci Anal*, vol. 4, no. 4, pp. 301–312, Dec. 2017, doi: 10.1007/s41060-017-0069-7.
- [60] R. Srikant and R. Agrawal, "Mining Generalized Association Rules."

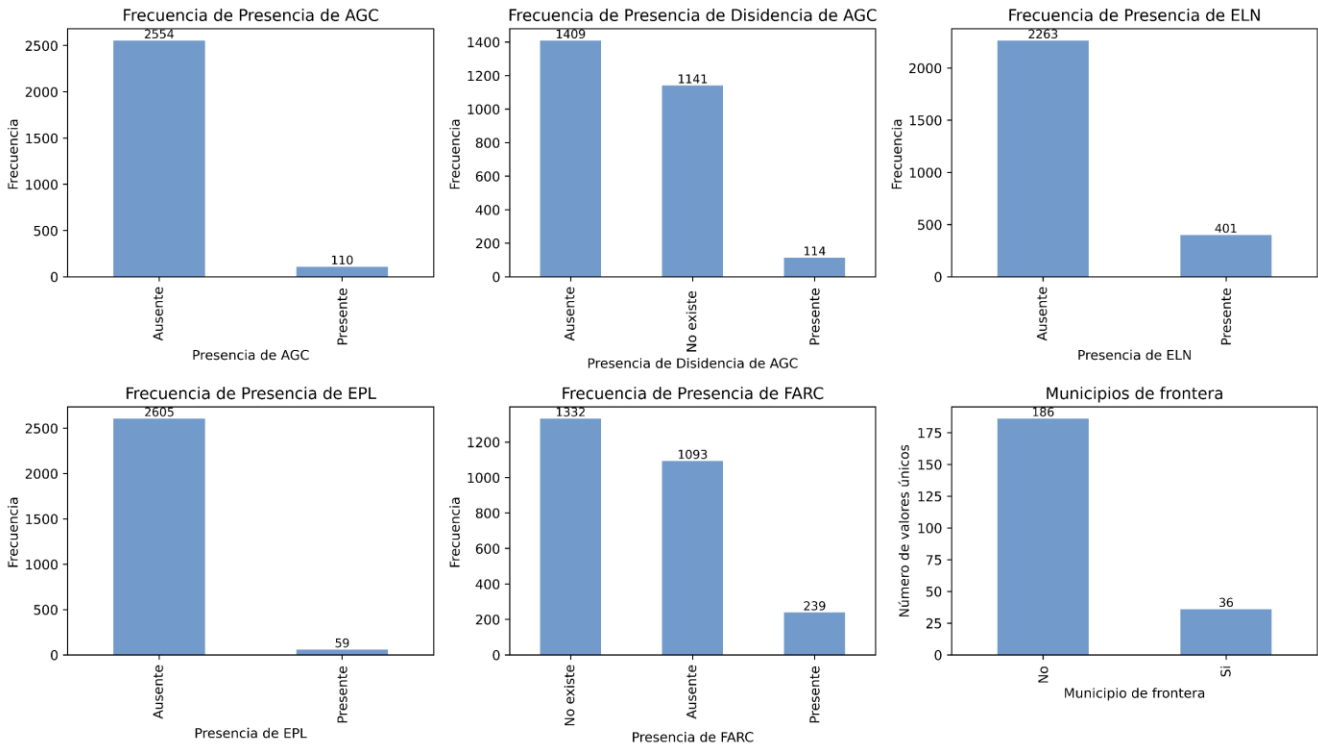
ANEXOS

Anexo A. Análisis exploratorio

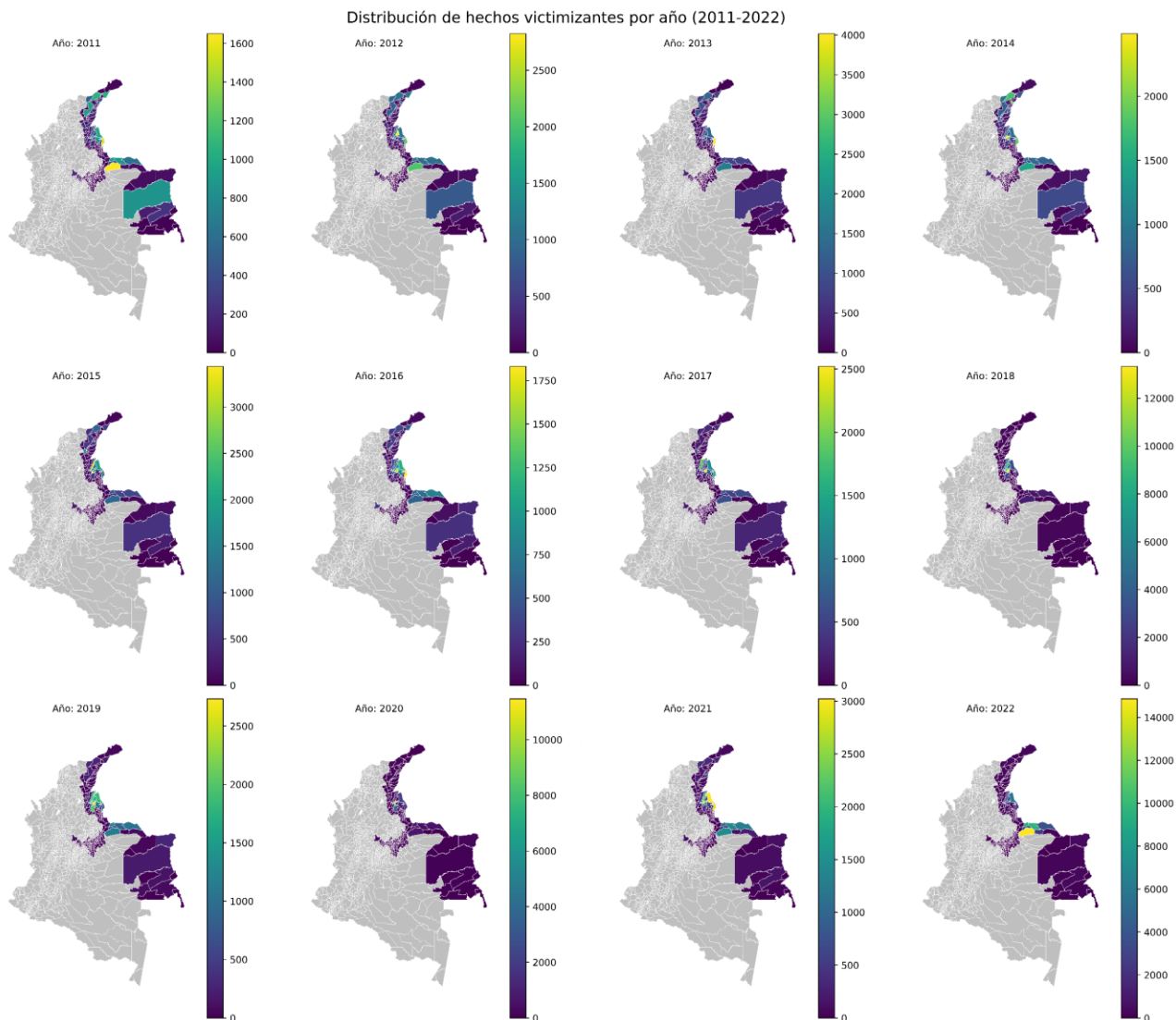
Anexo A1. Histogramas de distribución variables continuas.

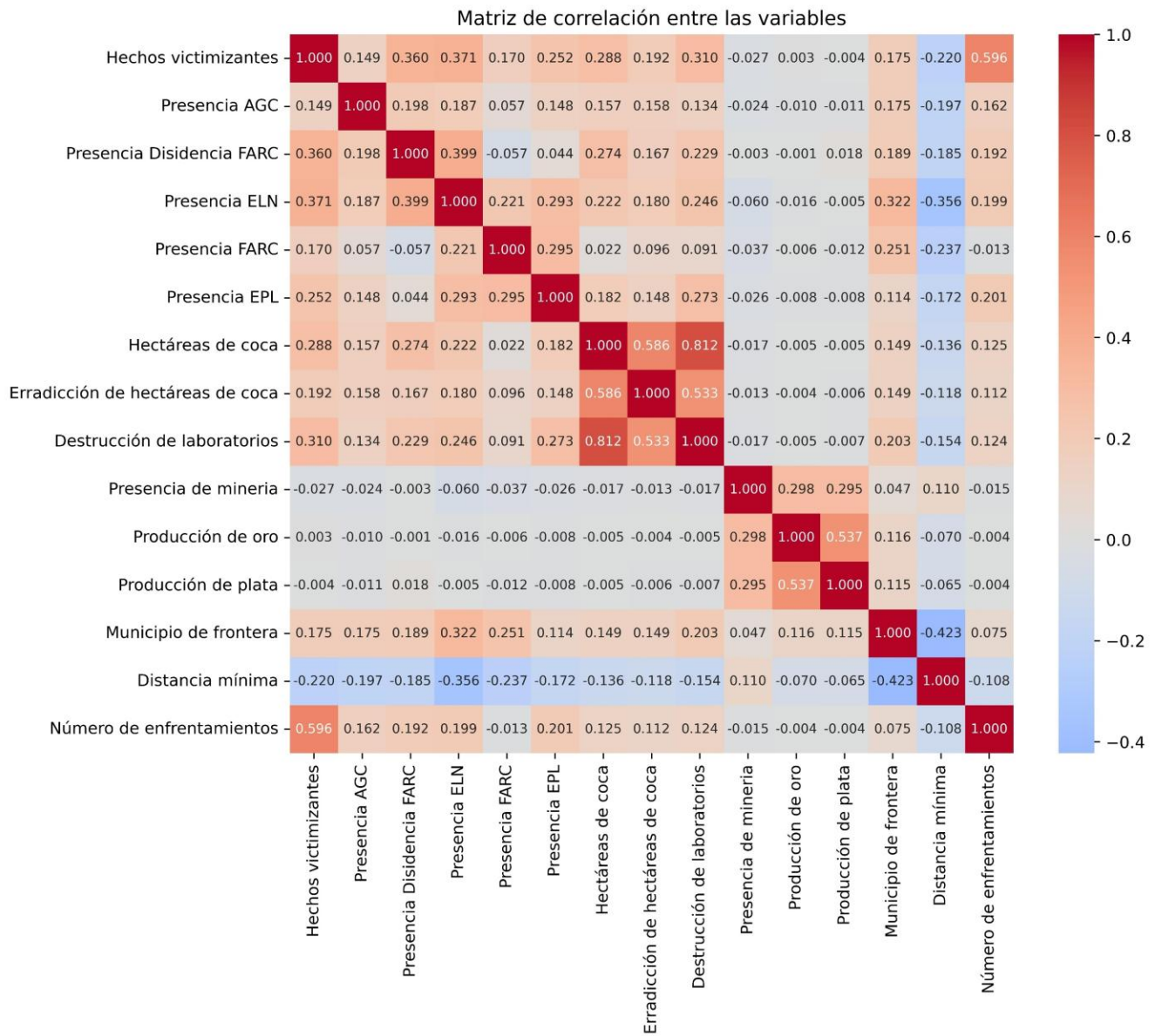


Anexo A2. Gráficos de barras variables categóricas.



Anexo A3. Dsistribución espacial de los hechos victimizantes por año.



Anexo A4. Matriz de correlación.

Anexo B. Resultados métricas de evaluación agrupamiento por año

Anexo B1. Resultados detallados de evaluación por año para el escenario con la variable hechos victimizantes.

Año	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCV	Coefficiente de silueta
2011	0,68	1194,89	0,58	0,70
2012	0,67	984,09	0,71	0,65
2013	0,66	1124,74	0,20	0,66
2014	0,65	1134,24	0,33	0,65
2015	0,66	1134,14	0,58	0,67
2016	0,66	1190,95	0,40	0,64
2017	0,72	1554,78	0,75	0,66
2018	0,69	634,56	0,69	0,57
2019	0,73	1023,43	0,43	0,72
2020	0,77	882,01	0,83	0,74
2021	0,74	1151,37	0,76	0,73
2022	0,71	866,56	0,72	0,69

Anexo B2. Resultados detallados de evaluación por año para el escenario con las variables de cercanía a la frontera.

Año	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCV	Coefficiente de silueta
2011	0,56	324,55	0,71	0,63
2012	0,54	297,51	0,64	0,62
2013	0,54	291,63	0,71	0,62
2014	0,54	300,58	0,70	0,62
2015	0,54	298,01	0,67	0,63
2016	0,56	307,78	0,68	0,63
2017	0,57	302,94	0,73	0,64
2018	0,54	248,39	0,62	0,63
2019	0,58	214,22	0,59	0,64
2020	0,58	230,93	0,69	0,65
2021	0,58	236,23	0,69	0,65
2022	0,57	265,14	0,66	0,63

Anexo B3. Resultados detallados de evaluación por año para el escenario con variables de presencia de economías ilegales.

Año	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCv	Coefficiente de silueta
2011	0,53	121,70	0,57	0,74
2012	0,63	107,32	0,57	0,75
2013	0,52	100,91	0,56	0,75
2014	0,53	120,71	0,51	0,73
2015	0,55	101,95	0,56	0,76
2016	0,59	105,37	0,64	0,78
2017	0,62	113,10	0,66	0,80
2018	0,65	75,07	0,81	0,83
2019	0,65	110,41	0,69	0,78
2020	0,71	300,38	0,86	0,87
2021	0,66	64,19	0,78	0,87
2022	0,68	51,67	0,68	0,84

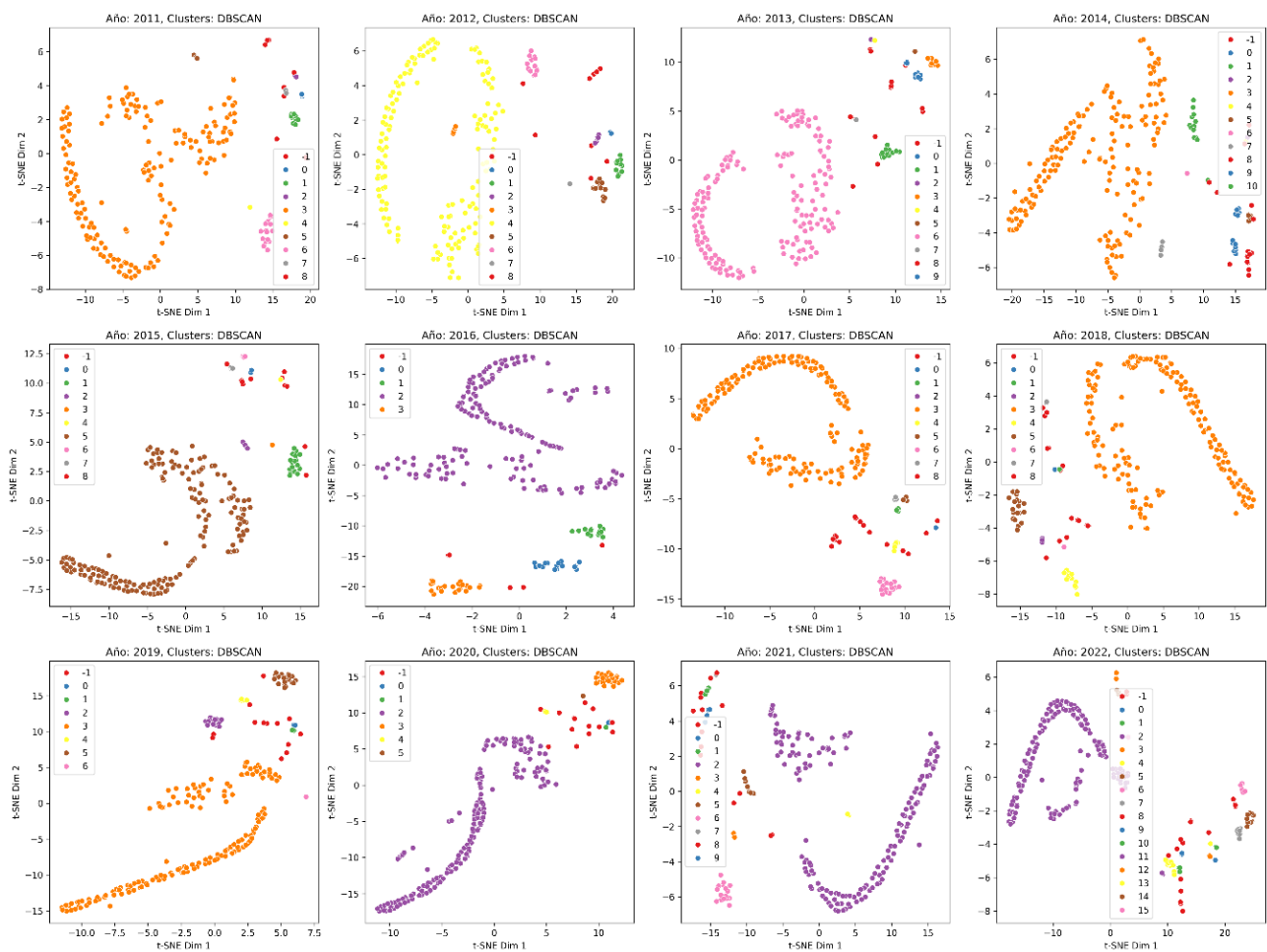
Anexo B4. Resultados detallado de evaluación por año para el escenario con variables de presencia de grupos.

Año	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCv	Coefficiente de silueta
2011	0,61	93,48	0,81	0,81
2012	0,70	82,27	0,84	0,70
2013	0,58	81,01	0,83	0,76
2014	0,76	83,71	0,83	0,72
2015	0,66	79,05	0,79	0,81
2016	0,75	133,36	0,84	0,82
2017	0,65	54,29	0,78	0,79
2018	0,75	44,02	0,81	0,79
2019	0,71	57,72	0,79	0,80
2020	0,66	48,52	0,79	0,86
2021	0,76	170,01	0,81	0,78
2022	0,78	277,80	0,82	0,65

Anexo B5. Resultados detallados de evaluación por año para el escenario con todas las variables.

Año	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCV	Coefficiente de silueta
2011	0,60	36,74	0,52	0,72
2012	0,65	81,37	0,56	0,74
2013	0,59	33,36	0,53	0,73
2014	0,57	83,89	0,48	0,72
2015	0,58	27,36	0,60	0,72
2016	0,53	69,92	0,62	0,71
2017	0,70	25,95	0,64	0,79
2018	0,79	17,18	0,62	0,83
2019	0,67	28,64	0,67	0,78
2020	0,75	25,07	0,68	0,83
2021	0,70	29,50	0,68	0,80
2022	0,54	61,23	0,67	0,71

Anexo B6. Gráfico t-SNE clustering DBSCAN de variables de presencia de grupos por año.



Anexo C. Resultados métricas de evaluación clustering por período del acuerdo de paz

Anexo C1. Resultados detallados de evaluación por período del acuerdo de paz para el escenario con la variable hechos victimizantes.

Período	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCV	Coefficiente de silueta
2011 - 2015	0,67	1332,44	0,58	0,62
2016 - 2022	0,72	4968,56	0,67	0,63

Anexo C2. Resultados detallados de evaluación por período del acuerdo de paz para el escenario con las variables de cercanía a la frontera.

Período	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCV	Coefficiente de silueta
2011 - 2015	0,55	1014,75	0,70	0,61
2016 - 2022	0,57	1418,54	0,74	0,64

Anexo C3. Resultados detallados de evaluación para el escenario con variables de presencia de economías ilegales por período del acuerdo de paz

Período	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCV	Coefficiente de silueta
2011 - 2015	0,51	226,83	0,50	0,80
2016 - 2022	0,61	209,20	0,75	0,87

Anexo C4. Resultados detallados de evaluación por período del acuerdo de paz para el escenario con las variables de presencia de grupos.

Período	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCV	Coefficiente de silueta
2011 - 2015	0,75	85,84	0,82	0,75

2016 - 2022	0,71	202,14	0,82	0,79
-------------	------	--------	------	------

Anexo C5. Resultados detallados de evaluación por período del acuerdo de paz para el escenario con todas las variables.

Período	k-means	Agrupamiento espectral	DBSCAN	Agrupamiento jerárquico
	Coefficiente de silueta	Calinski Harabasz	DBCV	Coefficiente de silueta
2011 - 2015	0,58	237,36	0,59	0,71
2016 - 2022	0,50	193,57	0,69	0,80

Anexo D. Resultados métricas de evaluación clustering por período presidencial .

Anexo D1. Resultados detallados métricas de evaluación por período presidencial para el escenario con la variable hechos victimizantes.

Período	k-means	Clustering Espectral	DBSCAN	Clustering Jerárquico
	Coefficiente de silueta	Calinski harabasz	DBCV	Coefficiente de silueta
Juan Manuel Santos (2011-2014)	0,67	3108,51	0,70	0,64
Juan Manuel Santos (2015 - 2018)	0,68	2700,90	0,55	0,67
Iván Duque (2019 - 2022)	0,73	2923,45	0,77	0,59

Anexo D2. Resultados detallados de evaluación por período preidencial para el escenario con las variables de cercanía a la frontera.

Período	k-means	Clustering Espectral	DBSCAN	Clustering Jerárquico
	Coefficiente de silueta	Calinski harabasz	DBCV	Coefficiente de silueta
Juan Manuel Santos (2011-2014)	0,55	766,12	0,72	0,56
Juan Manuel Santos (2015 - 2018)	0,55	859,02	0,75	0,55
Iván Duque (2019 - 2022)	0,58	915,85	0,64	0,61

Anexo D3. Resultados detallados de evaluación por período presidencial para el escenario con las variables de presencia de economías ilegales.

Período	k-means	Clustering Espectral	DBSCAN	Clustering Jerárquico
	Coeficiente de silueta	Calinski harabasz	DBCV	Coeficiente de silueta
Juan Manuel Santos (2011-2014)	0,53	236,92	0,56	0,69
Juan Manuel Santos (2015 - 2018)	0,57	262,86	0,52	0,61
Iván Duque (2019 - 2022)	0,67	235,53	0,60	0,76

Anexo D4. Resultados detallados de evaluación por período presidencial para el escenario con las variable de presencia de grupos.

Período	k-means	Clustering Espectral	DBSCAN	Clustering Jerárquico
	Coeficiente de silueta	Calinski harabasz	DBCV	Coeficiente de silueta
Juan Manuel Santos (2011-2014)	0,75	236,92	0,83	0,69
Juan Manuel Santos (2015 - 2018)	0,70	262,86	0,80	0,61
Iván Duque (2019 - 2022)	0,76	234,91	0,81	0,78

Anexo D5. Resultados detallados de evaluación por período presidencial para el escenario con todas las variables.

Período	k-means	Clustering Espectral	DBSCAN	Clustering Jerárquico
	Coeficiente de silueta	Calinski harabasz	DBCV	Coeficiente de silueta
Juan Manuel Santos (2011-2014)	0,58	190,02	0,81	0,72
Juan Manuel Santos (2015 - 2018)	0,42	164,46	0,78	0,83
Iván Duque (2019 - 2022)	0,61	360,38	0,80	0,76

Anexo E. Resultados centroides de métodos de agrupamiento.

Anexo E1. Resultados agrupación escenario con variable hechos victimizantes por período del acuerdo de paz.

Período	Grupo	Municipios	Media de hechos victimizantes	Mediana hechos victimizantes	Desviación estándar
2011 - 2016	1	415	0,18	0	0,5
	2	401	18,05	12	15,4
	3	294	467,95	235,5	581,4
2017 - 2022	1	795	0,25	0	0,7
	2	511	26,65	18	23,3
	3	248	1076,28	376,5	2022,7

Anexo E2. Resultados agrupación escenario con variable hechos victimizantes por período presidencial

Periodo presidencial	Grupo	Municipios	Media de hechos victimizantes	Mediana hechos victimizantes	Desviación estándar
Juan Manuel Santos (2011 -2014)	1	327	0,2	0	0,52
	2	306	16,2	11	13,21
	3	255	425,6	204	530,25
Juan Manuel Santos (2015 -2018)	1	440	0,6	0	1,20
	2	299	28,2	19	23,70
	3	149	858,7	365	1647,87
Iván Duque (2019-2022)	1	495	0,3	0	0,96
	2	270	34,3	22,5	29,94
	3	123	1349,8	523	2231,57

Anexo E3. Resultados agrupación escenario de variables relacionadas con cercanía a la frontera por etapa del acuerdo de paz.

Período	Grupo	Municipios	Media hechos victimizantes	Media municipios de frontera	Media distancia mínima a la frontera
2011-2016	1	930	83	0	210,1
	2	180	375	1	101,5
2017-2022	1	1299	97	0	210,2

	2	247	433	1	101,0
	-1	8	5930	1	146,3

Anexo E4. Resultados agrupación escenario de variables relacionadas con presencia de economías ilícitas por etapa del acuerdo de paz.

Período	Grupo	Municipios	Media hechos victimizantes	Media hectáreas de coca	Media erradicación ha de coca	Media destrucción de laboratorios
2011-2016	0	1091	115	12,6	5,3	0,6
	1	19	1003	1353,8	464,2	88,9
2017 - 2022	0	1479	92	1,4	1,1	0,2
	1	75	1936	3392,3	328,8	68,9

Anexo E5. Resultados agrupación escenario de variables relacionadas con presencia de grupos armados para el año 2022.

Año	Grupo	Municipios	Media de hechos victimizantes	Media de presencia de AGC	Media de presencia de Disidencia FARC	Media de presencia de ELN	Media de presencia de FARC	Media de presencia de EPL
2022	1	178	93	19,1%	10,1%	18,5%	0,0%	0,0%
	2	16	656	75,0%	37,5%	37,5%	0,0%	0,0%
	3	28	1884	10,7%	71,4%	64,3%	0,0%	3,6%

Anexo F. Resultados valores SHAP

Anexo F1. Estadísticas descriptivas de los valores SHAP para la clase 1 del grupo formado para todo el conjunto de variables en el año 2018

Variable	Promedio SHAP	Desviación Estándar SHAP	Percentil 25 SHAP	Percentil 75 SHAP
EPL	0,389	0,14	0,41	0,41
Destrucción de laboratorios	0,024	0,01	0,03	0,03

Hechos victimizantes	0,020	0,01	0,02	0,02
Hectáreas de coca	0,007	0,00	0,01	0,01
Enfrentamientos	0,002	0,00	0,00	0,00
Proximidad a la frontera	0,000	0,00	0,00	0,00
Erradicación de ha de coca	0,000	0,00	0,00	0,00
FARC	0,000	0,00	0,00	0,00
AGC	0,000	0,00	0,00	0,00
ELN	0,000	0,00	0,00	0,00
Disidencia FARC	0,000	0,00	0,00	0,00
Municipio de frontera	0,000	0,00	0,00	0,00

Anexo F2. Estadísticas descriptivas de los valores SHAP para la clase 1 del grupo formado para todo el conjunto de variables en el período posterior al acuerdo de paz (2017-2022)

Variable	Promedio SHAP	Desviación Estándar SHAP	Percentil 25 SHAP	Percentil 75 SHAP
Hectáreas de coca	0,16	0,09	0,17	0,18
Hechos victimizantes	0,11	0,06	0,12	0,12
Presencia de EPL	0,04	0,03	0,04	0,04
Presencia de AGC	0,04	0,03	0,04	0,04
Destrucción de laboratorios	0,03	0,01	0,03	0,04
Proximidad a la frontera	0,03	0,01	0,02	0,03
Enfrentamientos	0,02	0,01	0,02	0,02
Municipio de frontera	0,02	0,01	0,02	0,02
Erradicación de ha de coca	0,01	0,00	0,01	0,01
Presencia de ELN	0,01	0,00	0,01	0,01
Disidencia FARC	0,00	0,00	0,00	0,00
Presencia de FARC	0,00	0,00	0,00	0,00

Anexo F3. Estadísticas descriptivas de los valores SHAP para la clase 1 del grupo formado para todo el conjunto de variables en el segundo período presidencial de Juan Manuel Santos (2014-2018)

Variable	Promedio SHAP	Desviación Estándar SHAP	Percentil 25 SHAP	Percentil 75 SHAP
EPL	0,16	0,06	0,15	0,16
Destrucción de laboratorios	0,10	0,04	0,10	0,10
Hectáreas de coca	0,07	0,02	0,07	0,07

Maestría en Ciencia de Datos y Analítica

Erradicación de ha de coca	0,06	0,03	0,07	0,07
Hechos victimizantes	0,06	0,03	0,06	0,06
Proximidad a la frontera	0,03	0,02	0,03	0,04
Enfrentamientos	0,01	0,00	0,01	0,01
ELN	0,00	0,00	0,00	0,00
AGC	0,00	0,00	0,00	0,00
Municipio de frontera	0,00	0,00	0,00	0,00
Disidencia FARC	0,00	0,00	0,00	0,00
FARC	0,00	0,00	0,00	0,00

