INFLUENCIA DE LOS HOMICIDIOS SOBRE LA RELACIÓN ENTRE LA EFICIENCIA ESCOLAR Y LA RESILIENCIA ACADÉMICA: UN ANÁLISIS ESPACIAL PARA SANTIAGO DE CALI (2014-2024)

Nicolás Cardona Londoño y Caris Andrea Chia Amaya

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director Sebastián López Estrada

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE DE 2025

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: "Influencia de los homicidios sobre la relación entre la eficiencia escolar y la resiliencia académica: Un análisis espacial para Santiago de Cali (2014-2024)"

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Economía de la educación y economía del crimen.
4. ESTUDIANTE (S): Caris Andrea Chia Amaya & Nicolás Cardona Londoño
5. CORREO ELECTRÓNICO: carischia@javerianacali.edu.co , nicolascardonal@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Calle 18 N° 118-250 Cali (Colombia) +573504104449
7. DIRECTOR: Sebastián López Estrada
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: sebastian.lopezestra@javerianacali.edu.co
10. PALABRAS CLAVE: análisis espacial, machine learning, ciencia de datos, resiliencia académica, eficiencia educativa, homicidios.
11. ODS QUE APLICA EL PROYECTO (Agenda 2030): 4 - Educación de Calidad
12. FECHA DE INICIO (Desarrollo del proyecto): 1/11/2024
13. RESUMEN.

Este estudio analiza la influencia de los homicidios en la relación entre la resiliencia académica y la eficiencia educativa en las instituciones escolares de Santiago de Cali entre 2014 y 2024. A partir de datos del ICFES, DANE y la Secretaría de Seguridad y Justicia de Cali, se construyó un modelo de machine learning basado en Random Forest con interpretación SHAP para estimar la resiliencia académica, y un modelo condicional de eficiencia (Order-m FDH) para evaluar el desempeño educativo ajustado por factores socioespaciales. Los resultados evidencian una correlación positiva moderada (r = 0.48) entre resiliencia y eficiencia educativa, confirmando que las instituciones más eficientes tienden a generar entornos más favorables para estudiantes resilientes. Sin embargo, la violencia (medida por densidad de homicidios) ejerce una influencia negativa y estadísticamente significativa sobre la eficiencia (-0.12) y la resiliencia (-0.09), especialmente en zonas con alta concentración de delitos y menor infraestructura educativa. El análisis espacial identificó clústeres de baja eficiencia y resiliencia en el oriente y ladera de Cali, coincidentes con mayores tasas de homicidio. Los hallazgos sugieren que la exposición a la

violencia reduce la productividad educativa y limita la capacidad institucional para promover resiliencia, destacando la necesidad de políticas integradas de seguridad y educación en territorios vulnerables.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

La relación entre violencia y educación ha sido ampliamente debatida en las ciencias sociales, destacándose como un fenómeno complejo y multidimensional que impacta tanto en las dinámicas escolares (resultados, asistencia y ausentismo, convivencia) como en los entornos comunitarios [1]. En el caso de Colombia, donde el contexto de violencia urbana y homicidios ha persistido como una problemática estructural, entender cómo estas condiciones están relacionadas con la resiliencia académica y la eficiencia educativa resulta crucial para plantear soluciones desde las políticas y la administración pública. Santiago de Cali, conocida por su alta incidencia de homicidios, presenta un escenario idóneo para explorar esta interacción, especialmente considerando las desigualdades persistentes en el acceso y calidad de la educación.

Este estudio analiza entre el periodo de 2014 y 2024, la incidencia de los homicidios en la eficiencia educativa en Cali considerando la resiliencia académica definida como la capacidad de los estudiantes para obtener resultados sobresalientes en contextos adversos, emerge como un fenómeno de interés académico y social [2]. En tanto, la eficiencia educativa, entendida como el uso óptimo de recursos para maximizar resultados, es clave para evaluar el desempeño del sistema educativo en entornos desafiantes.

La metodología propuesta combina modelos de análisis de eficiencia y uso de modelos de *machine learning* con enfoque espacial. El primero tiene un enfoque no paramétrico donde se utiliza el *Order-m Conditional* para realizar una estimación de eficiencia educativa en relación con la resiliencia académica; mientras que el segundo busca estimar como los outputs (resultados académicos medidos por puntajes en la prueba Saber 11 del ICFES) del modelo de eficiencia pueden cambiar dado la cantidad de homicidios ocurridos en los alrededores de la institución. Se utilizan datos desagregados de la Secretaría de Seguridad y Justicia de Cali, el Departamento Administrativo Nacional de Estadística (DANE) y El Instituto Colombiano para la Evaluación de la Educación (ICFES).

Se plantea como hipótesis preliminar que (1) la resiliencia académica varía estrechamente con la eficiencia educativa; (2) que esta variación sigue un patrón espacial con clusters de barrios contiguos en situaciones similares; y (3) que los homicidios interactúan con diferente intensidad según el nivel de tolerancia interiorizado a nivel institucional. Este proyecto busca visibilizar la relación entre homicidios y eficiencia educativa, generar conocimiento replicable para contextos

urbanos similares y fomentar un debate más amplio sobre la resiliencia académica como concepto clave para la movilidad social y el desarrollo sostenible.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

El bajo desempeño académico de la educación básica y media en Colombia refleja profundos retos estructurales e institucionales [3], entre estos desafíos destacan la falta de acceso adecuado a recursos financieros, materiales y tecnológicos; la brecha entre las políticas públicas formuladas y su efectiva implementación; y la ubicación de muchas sedes educativas en contextos caracterizados por pobreza multidimensional, violencia tanto urbana como derivada del conflicto armado. Además, se enfrentan dificultades propias del acceso y la asistencia escolar debido a las condiciones geográficas del país, que impactan significativamente la calidad y continuidad del proceso educativo.

Estas condiciones no solo limitan las oportunidades de desarrollo de los estudiantes y su movilización social, sino que también ponen en riesgo el cumplimiento del Cuarto Objetivo de Desarrollo Sostenible de las Naciones Unidas [2], que busca garantizar una educación inclusiva, equitativa y de calidad como mecanismo de la reducción global de brechas y desigualdades. El riesgo latente en esta situación no solamente recae sobre el compromiso diplomático explícito, sino que es una señal de potenciales rezagos adicionales.

En este escenario, la resiliencia académica, entendida como la capacidad de los alumnos para lograr resultados sobresalientes en ambientes adversos [2], es no sólo un fenómeno de interés sino una necesidad social. La necesidad de entender cómo un grupo de estudiantes en zonas desfavorecidas logra cumplir las expectativas educativas, a pesar del costo de oportunidad que representa su permanencia escolar para sus familias; lleva inexorablemente a preguntar qué factores inciden para que esta situación suceda y qué se debe hacer para que pueda continuar ocurriendo.

Es así como la literatura señala algunos antecedentes que sugieren que la eficiencia educativa, operacionalizada desde recursos materiales y humanos, puede ser un factor clave para que la resiliencia emerja, se incentive y persista [4], incluso en el contexto colombiano [5]. No obstante, vale la pena preguntarse ¿Cuáles otros factores pueden afectar negativamente estos esfuerzos en cuanto a una creación holística de soluciones públicas? no solo aquello que promueve efectos positivos sino también lo que obstaculiza los efectos o escenarios deseados [6]. Desde una perspectiva similar, en sistemas complejos no basta con evaluar las variables endógenas sino también las exógenas para comprender las variaciones de los fenómenos en su conjunto.

Uno de los factores sociales que a nivel local, nacional e internacional tiende a afectar negativamente los esfuerzos públicos y privados para la movilidad social en general (desde la educación hasta la salud pública, la cultura, la productividad, etcétera) es la violencia en sus diferentes dimensiones. En la ciudad de Cali, de manera particular, la violencia se ha convertido en un fenómeno casi endémico de ciertas zonas impidiendo el florecimiento de las capacidades humanas de manera equitativa [7], [8]. Frente a estas condiciones es donde este trabajo cuestiona si la violencia en Cali (cantidad de homicidios ocurridos) puede estar relacionada y en qué nivel la forma en la cual la resiliencia académica se presenta a pesar de la presencia de diferentes niveles de eficiencia educativa.

La respuesta a este cuestionamiento puede implicar para la ciudad reflexiones en torno a las políticas públicas de educación y seguridad en su articulación y, en general, la comprensión de los efectos de la violencia en procesos estructurales y fundamentales para la sociedad, como la educación. Ahora bien, es de especial relevancia el uso de herramientas de ciencia de datos como los modelos condicionales de eficiencia, los modelos de *machine learning*, las regresiones espaciales, entre otros; para entender la magnitud del problema e invitan a considerar las particularidades zonales que puedan dar cuenta de diferentes situaciones en torno a la problemática descrita, pero que también permitan sugerir generalidades para un debate más amplio acerca de la resiliencia y la violencia en entornos educativos.

La articulación entre problemas sociales y métodos basados en la ciencia de datos, puede favorecer la réplica de la presente investigación en otros contextos nacionales e internacionales, para ahondar en la búsqueda de patrones alrededor de la resiliencia y sus múltiples causas y obstáculos. Precisamente, este trabajo busca determinar la naturaleza de un fenómeno para la ciudad Cali y contribuir con la ampliación de la literatura relacionada con resiliencia académica, eficiencia educativa y homicidios.

1.2 FORMULACIÓN DEL PROBLEMA

Ante la complejidad del problema anterior se hace pertinente abordar las siguientes preguntas:

Pregunta general:

- ¿Cómo aplicar metodologías de ciencia de datos para analizar la interacción entre homicidios, resiliencia académica y eficiencia educativa en las instituciones educativas de Cali entre 2014 y 2024?

Preguntas sistemáticas:

- ¿Qué procesos de limpieza y exploración son necesarios para estructurar los datos de homicidios, rendimiento educativo y factores socioeconómicos de Cali, garantizando su calidad para el análisis?
- ¿Cómo se puede entrenar un modelo de *machine learning* para estimar los niveles de resiliencia académica en las instituciones educativas de Cali, utilizando datos sociodemográficos y variaciones temporales?
- ¿Qué métricas de rendimiento son apropiadas para validar la efectividad del modelo de *machine learning* en la estimación de la resiliencia académica?
- ¿Qué patrones geoespaciales de homicidios se identifican en relación con la ubicación de las instituciones educativas en Cali mediante análisis de autocorrelación y herramientas SIG?
- ¿Cómo están relacionados los homicidios y los recursos institucionales y contextuales la eficiencia educativa y la resiliencia académica según el modelo condicional de eficiencia educativa?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Construir modelos de aprendizaje automático para analizar la incidencia de los homicidios en la interacción entre eficiencia educativa y resiliencia académica en la ciudad de Cali entre los años 2014 y 2024.

2.2 OBJETIVOS ESPECÍFICOS

- Preparar los datos recolectados del ICFES, DANE y Secretaría de Seguridad y Justicia de Cali mediante procesos de limpieza y exploración.
- Realizar análisis geoespaciales y de autocorrelación sobre los patrones geográficos de homicidios en relación con la distribución de las instituciones educativas.
- Entrenar un modelo de *machine learning* que identifique los estudiantes resilientes en las instituciones educativas de Cali con base en variaciones temporales y datos sociodemográficos.
- Validar el modelo de *machine learning* mediante métricas de rendimiento verificando su efectividad para estimar la resiliencia académica.
- Implementar un modelo condicional de eficiencia educativa considerando la incidencia de los homicidios en la relación entre la eficiencia educativa y la resiliencia académica en la ciudad de Cali.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO: EDUCACIÓN Y HOMICIDIOS

La educación es ampliamente reconocida como una forma de inversión en capital humano, un pilar fundamental para el desarrollo individual y la movilidad social y un factor cohesionador para las comunidades y la productividad nacional; este paradigma se ha alineado internacionalmente en los Objetivos de Desarrollo Sostenible [9], específicamente con el ODS 4, donde se establece que la relación entre educación, productividad y crecimiento económico subyace en las necesidades de reducir las brechas de acceso y calidad educativa, especialmente en contextos de vulnerabilidad.

La evidencia respecto al impacto agregado e individual de la educación es amplia. Se ha sugerido que los países que priorizan la educación en sus decisiones de política pública y política de Estado tienden a experimentar mayores niveles de competitividad e innovación y, por tanto, una mejor calidad de vida Buscha y Dickson [10] [11]. El racional en ello es que la educación impulsa el desarrollo de habilidades, conocimientos esenciales para la productividad, competencias ciudadanas y motivaciones personales para la movilidad social [12], [13].

Así, un elemento en común en esta discusión es que la relación entre educación y retornos (en el sentido amplio) no es lineal y está mediada fuertemente por múltiples factores institucionales, estructurales y contextuales. Hanushek y Wößmann [14] señalan que el caso latinoamericano es representativo en este sentido ya que a pesar de tener un grado de cobertura educativa creciente, los países de esta región no reciben los beneficios esperados de la misma. En este contexto, se ha sugerido adicionalmente que en casos como Colombia y México existen variables particulares como la violencia que, en sus diferentes expresiones, se convierten en un factor decisivo en lo que respecta al desempeño académico, el ausentismo [15] y las decisiones presupuestales [16].

Esta investigación se interesa particularmente en la influencia que expresiones concretas de violencia, como los homicidios, pueden llegar a tener sobre la dinámica entre eficiencia educativa y resiliencia académica. Arbona, et. al [17] sugieren que, efectivamente, los homicidios en el marco del conflicto armado se asocian con niveles de eficiencia de la calidad educativa en Colombia y [18] lo confirma específicamente para la ciudad objetivo de esta investigación, Cali (Colombia); mientras que Jamison et al [19] concluye que la evidencia apunta a una afección directa de la violencia en las comunidades sobre el desarrollo de resiliencia en estudiantes en edad escolar. Esta interacción entre violencia y educación requiere de un marco conceptual por sí mismo debido a la transversalidad de su influencia.

Eficiencia educativa y resiliencia

El concepto de eficiencia educativa ha sido influenciado por desarrollos en economía de la educación, necesidades políticas y necesidades sociales; sus inicios se pueden registrar con autores como Coleman [20] quién investigó la calidad educativa y desigualdad de oportunidades en Estados Unidos. Posteriormente, varios autores buscaron conceptualizar el término refiriéndose a él como la capacidad de los sistemas educativos para maximizar los resultados académicos y formativos utilizando la menor cantidad de recursos posibles [21] y para efectos de este trabajo, esta será la definición adoptada. A su vez, según De Witte y López-Torres [22] la medición de la eficiencia educativa permite evaluar, mediante métodos de análisis de frontera, la productividad de las escuelas, sistemas educativos y estudiantes ya que compara los recursos invertidos y los resultados obtenidos.

De igual forma, Perelman y Santin [23] adicionan a la medición al estudiante como ente que se encuentra influenciado por variables como el contexto familiar, entorno escolar y factores externos. A su vez, los sistemas educativos eficientes promueven la inclusión y justicia [24] por lo que no solo se limita a resultados cuantitativos, sino que considera el impacto social que estos pueden tener en diferentes grupos socioeconómicos a los que pertenecen los estudiantes.

Hanushek y Luque [25] introducen una función educativa para modelar el proceso de aprendizaje, donde la educación se entiende como un producto de múltiples insumos integrados en una ecuación matemática, en esta función, se consideran tanto las variables de entorno (como el contexto familiar, los recursos escolares y las condiciones socioeconómicas) como los insumos propios de la escuela, incluyendo la calidad y cantidad de los recursos disponibles. A partir del análisis de datos internacionales, estos autores resaltan que el aumento de recursos no garantiza mejores resultados educativos, sino que es crucial la implementación de políticas orientadas a la eficiencia en el uso de tales recursos, con una visión que integra factores económicos y sociales. Esta perspectiva se vincula estrechamente con el concepto de resiliencia académica, que estudia cómo los factores externos e internos al individuo, denominados “factores protectores”, pueden mitigar las adversidades del entorno desfavorecido y fomentar la eficacia académica de los estudiantes. La capacidad de los alumnos para lograr resultados óptimos en ambientes adversos es la definición de resiliencia académica que se adopta en el presente proyecto.

La resiliencia académica es un concepto que se ha estudiado, en su mayoría desde la psicología, donde autores como Norman Garmezy y Emmy Werner [26] observaban que ciertos niños se desarrollaban de manera positiva a pesar de estar inmersos en un contexto desfavorecedor. Posteriormente, la resiliencia empezó a estudiarse en varios contextos entre ellos el ámbito educativo dando origen al término “resiliencia educativa o académica” impulsado por Michael

Rutter y su estudio sobre los factores protectores escolares y comunitarios que ayudan a los estudiantes a superar adversidades.

A lo largo de los años, la resiliencia académica ha tomado un enfoque holístico, con diferentes investigaciones que consideran la globalización, las tecnologías y la diversidad cultural. Almulla [27] emplea un diseño descriptivo para revisar cómo el apoyo social y la autoconfianza inciden en la resiliencia académica; por otro lado, hay investigaciones como la de Tri y Mardi [28] que analizan la relación entre el optimismo y la resiliencia resaltando que hay una correlación positiva significativa entre ambos factores y aquellos estudiantes optimistas tienen mayor habilidad para afrontar las situaciones adversas.

Además, otras investigaciones de corte psicométrico validan escalas que permiten medir el nivel de resiliencia en estudiantes y concluyen que la motivación y regulación emocional también son factores influyentes [29], [30], [31]; sin embargo, y por el interés de esta investigación, se va a hacer énfasis en estudios cuantitativos con estrecha relación con la Ciencia de Datos.

En línea con la perspectiva cuantitativa y análisis de datos a gran escala, Agasisti y colaboradores [32], [33], [34], [35] han desarrollado investigaciones determinantes sobre la resiliencia académica a partir de bases internacionales como PISA, dichas investigaciones destacan que, incluso entre estudiantes de contextos desfavorecidos, aproximadamente uno de cada cuatro logran resultados sobresalientes, siendo factores como el clima disciplinario escolar, las altas expectativas del profesorado y el tiempo de instrucción variables clave para la resiliencia académica identificadas mediante modelos estadísticos avanzados y análisis comparativos multinacionales. Adicionalmente, [35] profundizan en las características escolares que explican el éxito de estos estudiantes en varios países, encontrando un papel relevante de la calidad del entorno escolar y el apoyo institucional, lo que confirma que la ciencia de datos permite identificar patrones y proporcionar recomendaciones aplicables a diferentes sistemas educativos.

En concordancia, Sandoval-Hernández y Białowolski [36] emplearon un modelo de regresión logística en Asia Oriental que analiza una submuestra de estudiantes desfavorecidos y estiman la probabilidad de desarrollar resiliencia académica en función a varios predictores, sus resultados indican que la actitud positiva de los alumnos hacia las matemáticas, la confianza de los profesores en el rendimiento de alumnos y el idioma del examen se asociaba a mayores probabilidades de éxito académico.

Con una técnica muy similar, Abdillah y Marleni [37] realizaron una regresión logística con estudiantes de enfermería de una Universidad en las Islas Bangka Belitung, allí se aplicaron dos Escalas que miden el estrés académico y nivel de resiliencia, dando como resultado que cuánto mayor sea su nivel de resiliencia, menor es su estrés académico. No obstante, faltan

investigaciones que usen modelos de *Machine Learning* para estudiar la resiliencia académica en territorio Latinoamericano.

Homicidios y educación

La relación entre violencia y educación ha sido objeto de amplio debate académico, destacándose como un fenómeno multifacético que impacta tanto en los entornos escolares como en las dinámicas comunitarias.

Autores como Derriennic [38] proponen que los homicidios permiten abordar la violencia desde los hechos específicos en cuanto este tipo de expresiones impiden el goce efectivo de los derechos de las personas, las llevan a tomar decisiones que en otros escenarios no tomarían e incentivan la interiorización de comportamientos y representaciones sociales propias de estructuras de poder. De esta propuesta, el homicidio genera consecuencias objetivas medibles más allá de las intenciones subjetivas que han motivado el acto, lo cual permite sugerir que tanto los homicidios vinculados al crimen como los homicidios vinculados a otras lógicas sociales podrían afectar las dinámicas educativas y académicas.

En cuanto los homicidios y los actos violentos subyacentes al crimen, sugiere Kalyvas [39] , están intrínsecamente ligado al control territorial de los actores armados, donde la intensidad y el tipo de violencia dependen de la capacidad de los grupos para imponer lealtades entre la población. En contextos urbanos, los homicidios actúan como un mecanismo de control para asegurar la extracción de rentas criminales y la impunidad de los actos. En este sentido, los homicidios orbitan diferentes aspectos de la experiencia de los estudiantes al suceder como producto tiroteos o enfrentamientos de bandas en las rutas hacia la escuela [40] [41], en los alrededor de la escuela misma durante la jornada escolar [42], en los espacios de actividades deportivas y comunitarias [43] o como una posible situación de la que se puede ser víctima (directa o indirecta) de las dinámicas territoriales, como las de las “fronteras invisibles”[1] de Cali [44].

Diversos estudios coinciden en que el entorno geográfico constituye el factor común que explica la relación entre los homicidios y los resultados educativos, ya que la proximidad de la violencia a los espacios escolares incide directamente en el desempeño académico. Por ejemplo, Davanzo y Justus [45] evidenciaron en la ciudad de São Paulo que la cercanía de los tiroteos y los homicidios a las escuelas y rutas de acceso se relaciona con variaciones en los niveles de competencia escolar. En esta misma línea, Barboza [43] muestra que la proximidad de las escuelas a zonas con alta concentración de disparos en la ciudad de Boston afectó el rendimiento académico, al generar cuadros de estrés y trauma que reducen la concentración y la motivación de los estudiantes.

[1] En la misma investigación se consideran las fronteras invisibles como territorios controlados por pandillas, con límites difusos y en constante competencia.

Debido a lo anterior, los estudios sobre las consecuencias de la violencia en la educación destacan su impacto profundo y multidimensional. Este efecto no es unidireccional, sino que se manifiesta de manera transversal en los procesos personales, institucionales y, de forma agregada, en la inversión en capital humano. Por un lado, la exposición a la violencia tiene repercusiones significativas en el bienestar psicológico y social de los estudiantes, afectando tanto su desarrollo cognitivo como emocional. Investigaciones como la de Duque [46] evidencian que el estrés psicológico asociado a entornos violentos genera ansiedad, trastornos del sueño y dificultades de concentración, lo que compromete el desarrollo integral de los estudiantes.

En términos de los procesos educativos, la violencia incrementa el ausentismo de estudiantes y profesores, reduce los días efectivos de clase y deteriora la calidad académica. Jurges [47] señala que las amenazas a la seguridad interrumpen el funcionamiento regular de las instituciones escolares. Fergusson y Michaelsen [48] subrayan que la inseguridad en los entornos escolares no solo disminuye los puntajes académicos, sino que también afecta el logro de objetivos educativos a largo plazo. Adicionalmente, Bruck et al. [49] destacan que el daño a la infraestructura educativa en zonas afectadas por violencia agrava las desigualdades estructurales, al limitar el acceso a entornos adecuados para el aprendizaje.

Por último, a nivel agregado, estos impactos disminuyen el retorno de la inversión en capital humano y reducen el bienestar social general. La incertidumbre generada por la violencia, como señala Bruck et al. [49] , disminuye las expectativas sobre los beneficios futuros de la educación, desincentivando la permanencia escolar. Además, Salvo et al. [50] muestran cómo los estudiantes, en contextos de violencia, suelen priorizar habilidades de sobrevivencia sobre su formación académica, perpetuando ciclos de exclusión social y limitando las oportunidades de movilidad económica para generaciones enteras.

3.2 MACHINE LEARNING Y RANDOM FOREST

El aprendizaje automático (*machine learning*, ML) puede definirse como un conjunto de métodos computacionales orientados a identificar patrones en los datos con el objetivo de realizar predicciones o clasificaciones, reduciendo la dependencia de supuestos paramétricos estrictos sobre la forma funcional de las relaciones entre variables [51]. A diferencia de los enfoques estadísticos tradicionales, que suelen especificar de antemano la estructura del modelo, los algoritmos de machine learning aprenden dicha estructura directamente a partir de

los datos, lo que resulta especialmente útil en contextos caracterizados por alta complejidad, multicausalidad y no linealidad.

En el campo de las ciencias sociales, la educación y los estudios sobre violencia, estas metodologías han ganado relevancia debido a su capacidad para modelar interacciones complejas entre variables individuales, familiares, institucionales y contextuales. En particular, el machine learning ha sido empleado para la predicción de riesgos, la identificación de poblaciones vulnerables, la detección temprana de eventos adversos y el análisis de procesos sociales que no pueden ser adecuadamente representados mediante modelos lineales simples [52]. En este sentido, su uso no se limita a la predicción, sino que también permite explorar regularidades empíricas que contribuyen a la construcción y validación teórica.

3.2.1 Aprendizaje supervisado: fundamentos conceptuales

Dentro del amplio conjunto de técnicas de machine learning, el aprendizaje supervisado ocupa un lugar central en la investigación aplicada. Este enfoque se refiere a algoritmos que aprenden una función de mapeo entre un conjunto de variables explicativas *X* y una variable objetivo *Y*, utilizando ejemplos en los que el valor de *Y* es conocido [53]. En contextos educativos y de violencia, la variable objetivo puede representar categorías binarias (por ejemplo, presencia o ausencia de violencia, resiliencia académica alta o baja) o variables continuas (como puntajes estandarizados, índices de riesgo o niveles de desempeño).

El objetivo del aprendizaje supervisado no es únicamente reproducir los resultados observados en el conjunto de entrenamiento, sino lograr un desempeño adecuado sobre datos no observados, es decir, generalizar correctamente. Este principio es especialmente relevante en investigaciones orientadas a la toma de decisiones públicas, donde los modelos deben ser capaces de anticipar comportamientos futuros, identificar estudiantes o instituciones en riesgo, y apoyar intervenciones focalizadas basadas en evidencia empírica.

En el marco de esta investigación, el aprendizaje supervisado se articula directamente con objetivos como la identificación de estudiantes resilientes en contextos adversos, la estimación de probabilidades asociadas a determinados resultados educativos y la exploración de cómo variables contextuales, como la violencia homicida, interactúan con los recursos institucionales y socioeconómicos.

3.2.2 Modelos supervisados más comunes en investigación social y educativa

Desde una perspectiva teórica, resulta pertinente situar brevemente los principales modelos supervisados utilizados en la literatura aplicada [54]:

- Regresión lineal y regresión logística

Estos modelos constituyen el punto de partida clásico en la investigación social y educativa. La regresión lineal se emplea para variables continuas, mientras que la regresión logística es ampliamente utilizada para resultados binarios. Ambos enfoques ofrecen interpretabilidad directa y una sólida base inferencial, pero suponen relaciones lineales entre predictores y resultado, lo que limita su capacidad para capturar interacciones complejas y efectos no lineales frecuentes en fenómenos sociales.

- Árboles de decisión

Los árboles de decisión modelan relaciones no lineales mediante particiones recursivas del espacio de variables, generando reglas del tipo “si-entonces” que resultan intuitivas y fácilmente interpretables. No obstante, los árboles individuales tienden a presentar alta varianza y riesgo de sobreajuste, especialmente en conjuntos de datos ruidosos o de alta dimensionalidad, lo que ha motivado el desarrollo de métodos de ensamble.

- Random Forest

El algoritmo de Random Forest consiste en un conjunto de árboles de decisión entrenados sobre muestras bootstrap del conjunto de datos y subconjuntos aleatorios de predictores. La agregación de sus predicciones permite reducir el sobreajuste y mejorar la capacidad de generalización. En investigaciones sobre violencia y educación, este modelo ha demostrado un desempeño sólido frente a datos heterogéneos y altamente correlacionados, además de ofrecer medidas de importancia de variables que facilitan el análisis interpretativo y el diálogo con marcos teóricos existentes.

- Máquinas de gradiente boosting (GBM, XGBoost, LightGBM)

Estos métodos construyen modelos de manera secuencial, donde cada nuevo árbol corrige los errores del conjunto anterior. Suelen alcanzar altos niveles de desempeño predictivo en datos tabulares complejos, aunque a costa de una mayor complejidad computacional y una interpretabilidad menos directa si no se acompañan de herramientas explicativas.

- Máquinas de soporte vectorial (SVM)

Las SVM buscan encontrar un hiperplano que maximice la separación entre clases en un espacio de características potencialmente transformado. Son especialmente útiles en problemas de alta dimensionalidad, como análisis de texto o detección de violencia en narrativas, aunque su interpretación resulta menos intuitiva en contextos de política pública.

- Redes neuronales

Las redes neuronales artificiales constituyen aproximadores universales capaces de capturar patrones altamente complejos. En el ámbito educativo y de violencia, se utilizan con frecuencia en procesamiento de lenguaje natural, análisis de imágenes o videos, y en menor medida sobre

datos estructurados. Su principal limitación radica en la opacidad de sus decisiones, lo que plantea desafíos éticos y metodológicos en investigaciones sociales.

3.2.3 Métricas de evaluación para modelos de clasificación

La evaluación de modelos supervisados requiere el uso de métricas calculadas sobre datos que no han sido utilizados durante el proceso de entrenamiento, con el fin de garantizar que el desempeño observado refleja la capacidad de generalización del modelo y no un ajuste específico a la muestra analizada. En problemas de clasificación binaria o multiclase, una de las métricas más utilizadas es la exactitud (accuracy) [51], [53], [55], la cual mide la proporción de observaciones correctamente clasificadas respecto al total. Si bien esta métrica resulta intuitiva, su interpretación puede ser engañosa en contextos caracterizados por desbalance de clases, como ocurre cuando los eventos de violencia severa o los casos de resiliencia académica representan una fracción reducida de la población total.

Por esta razón, en investigaciones sociales y educativas suele ser necesario complementar la exactitud con métricas que permitan evaluar con mayor precisión los distintos tipos de error del modelo. En este sentido, la precisión, el recall o sensibilidad y el F1-score ofrecen una visión más detallada del desempeño predictivo. La precisión indica la proporción de observaciones correctamente clasificadas entre aquellas que el modelo identifica como positivas, mientras que el recall mide la capacidad del modelo para identificar efectivamente todos los casos positivos existentes. El F1-score, como media armónica entre precisión y recall, permite equilibrar ambos criterios y resulta particularmente útil cuando los costos de los errores de clasificación son asimétricos.

En el análisis de fenómenos como la violencia y la resiliencia educativa, estas métricas adquieren una relevancia no solo técnica sino también ética, dado que los errores del modelo pueden traducirse en consecuencias sociales diferenciadas. En particular, los falsos negativos implican la omisión de individuos o instituciones en situación de alto riesgo, mientras que los falsos positivos pueden generar estigmatización o intervenciones innecesarias, lo que obliga a una evaluación cuidadosa del desempeño del modelo más allá de una única medida agregada.

Adicionalmente, cuando los modelos generan probabilidades o puntajes de riesgo, se hace necesario incorporar métricas que evalúen su capacidad discriminativa a lo largo de distintos umbrales de decisión. En este contexto, la curva ROC representa la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para distintos puntos de corte, mientras que el área bajo la curva (AUC) sintetiza esta información en una única medida. El AUC refleja la probabilidad de que el modelo asigne un puntaje mayor a una observación positiva que a una negativa, independientemente del umbral seleccionado, lo que explica su amplio uso en estudios de predicción de violencia y desempeño educativo y su adopción como métrica principal en el presente trabajo.

3.3 ANTECEDENTES APLICADOS

Debido al interés de esta investigación en aplicar metodologías avanzadas de *machine learning* para comprender cómo los homicidios influyen en la identificación de estudiantes resilientes en un entorno académico y en los modelos de eficiencia educativa, se hace necesario considerar antecedentes al respecto. Gran parte de la literatura previa sobre eficiencia educativa y resiliencia académica proviene de contextos no comparables con los países en desarrollo, lo que limita la aplicabilidad de esos enfoques a realidades como la de Colombia, donde la violencia urbana es persistente y los entornos presentan características muy diferentes. Así, estudios internacionales que combinan directamente la violencia, en particular los homicidios, con eficiencia educativa son escasos.

A grandes rasgos se identifican seis grupos de aplicaciones macro. El primer grupo se centra en la detección de violencia física en entornos educativos que pueden llegar a homicidios y otro tipo de consecuencias como deserción. Mediante el uso de redes neuronales convolucionales (CNN) y redes de memoria a largo plazo (LSTM), se han desarrollado sistemas capaces de analizar datos de video para identificar actos violentos en escuelas [46]. Al capturar características espacio-temporales locales se ha logrado de manera preliminar reducir el tiempo de respuesta institucional en casos de violencia interna.

En contextos digitales, un segundo grupo de aplicaciones se ha enfocado en encontrar los modelos más óptimos para detectar acoso cibernético y casos potenciales de violencia expresados en redes sociales. La investigación realizada por Ogunleye y Dharmaraj [47] concluyen que los modelos de lenguaje grande (LLM) son más robustos que modelos tradicionales como Support Vector Machine o Random Forest para entrenar detectores de publicaciones y contenido alarmante en redes sociales. Específicamente, esta investigación logró un 97% de exactitud a la hora de identificar contenido violento en redes usadas por estudiantes de escuelas experimentales.

Un grupo adicional de aplicaciones se centra en la identificación temprana de jóvenes con riesgo de presentar comportamientos agresivos hacia otros. En esta línea ha sido significativo el uso del algoritmo RIO (Risk of Injury to Others), entrando con datos del interRAI Child and Youth Mental Health Screener (ChYMH-S) en la provincia de Ontario [48]. Este algoritmo, validado con datos de 59 agencias de salud mental, ha demostrado ser un fuerte predictor de comportamientos agresivos en jóvenes. La metodología, que combina herramientas psicométricas y análisis de datos avanzados, permite identificar señales o síntomas asociados con una mayor probabilidad de violencia, facilitando intervenciones tempranas para reducir la agresividad futura.

El cuarto grupo se centra en la identificación de la violencia de género y su impacto en los estudiantes. Un estudio de mapeo sistemático reciente realizado por Pinto-Muñoz et al. [49] en Colombia analizó investigaciones sobre la aplicación de *machine learning* a este problema entre

2018 y 2023. Los hallazgos destacan beneficios significativos en el uso de modelos predictivos y, en particular, el aprendizaje automático ha demostrado ser efectivo para identificar patrones y factores de riesgo, ofreciendo herramientas valiosas para la prevención y mitigación de la violencia de género. Sin embargo, el estudio también evidencia una carencia de modelos o algoritmos específicos aplicados a este fenómeno en contextos latinoamericanos. Este vacío representa una oportunidad para desarrollar soluciones innovadoras que combinan capacidades de procesamiento de datos con enfoques de las ciencias sociales.

Este tipo de aplicaciones también han sido utilizados para evaluar el impacto de la violencia barrial sobre la educación escolar. El estudio citado anteriormente realizado en Brasil [40] es un caso significativo de ello en cuanto se exploró cómo factores externos y la proximidad a áreas con alta criminalidad afectan variables clave como el desempeño académico, el ausentismo escolar y los niveles de estrés estudiantil. Los hallazgos subrayan la importancia de incluir dimensiones geoespaciales debido a que la experiencia escolar de los alumnos incluye momentos como las rutas de transporte o las zonas de dispersión.

Finalmente, y de manera más reciente, múltiples autores se han interesado por la creación y ajuste de modelos de *machine learning* específicamente para evaluar la resiliencia académica no solamente en relación a la violencia sino en relación a fenómenos que pueden incidir significativamente en esta variable: la pandemia del COVID-19 [56], el estrés del profesorado [57] y la atención sociopsicológica personalizada a los estudiantes [58]. En especial, el estudio desarrollado por Cheung et al [56] se convierte en la base metodológica en cuanto al modelo de interacción de homicidios-resiliencia académica se trata. Estos autores realizaron una investigación transversal en 79 países enfocado en las pruebas PISA y sugieren que la resiliencia académica encuentra oportunidades de predicción en modelos de Random Forest con técnicas SHAP de explicación aditivas [56].

En el contexto colombiano, también se han desarrollado investigaciones que relacionan la violencia y el desempeño educativo. Ariza y Saldarriaga [4] identificaron que durante periodos de alta intensidad del conflicto armado, como en 2003 (marcado por ataques terroristas y desplazamientos forzados), se observaron impactos negativos directos e indirectos en los resultados académicos de los estudiantes, con efectos persistentes a largo plazo. De manera similar, Arbona et al. [17], evidenciaron que los homicidios asociados al conflicto redujeron significativamente la eficiencia educativa, afectando tanto el desempeño en las pruebas Saber 11 como las tasas de promoción escolar. Estos hallazgos son sustento empírico de cómo los contextos de violencia prolongada en Colombia pueden alterar las dinámicas institucionales y las trayectorias educativas, especialmente en territorios vulnerables.

No obstante, para el caso urbano contemporáneo, destacan los aportes recientes que analizan la relación entre criminalidad y eficiencia escolar en Santiago de Cali. El estudio de

Muñoz-Galeano, López-Estrada y Arbona [59] representa una contribución relevante, al analizar la relación entre las tasas de homicidio y la eficiencia escolar en 301 instituciones educativas de Cali, Colombia, mediante un enfoque no paramétrico que incorpora variables ambientales y espaciales. Este trabajo aporta evidencia crucial para entender cómo la criminalidad puede impactar la calidad educativa en contextos urbanos vulnerables, lo que representa un valor agregado desde el punto de vista metodológico y contextual para el presente estudio. Aunque no se encuentran antecedentes específicos que unan homicidios con resiliencia académica y eficiencia educativa, sí existen estudios que emplean *machine learning* en el marco general de estas problemáticas, lo que sustenta la novedad y relevancia de la propuesta investigativa.

Esta investigación retoma una aproximación metodológica similar a la del estudio anterior, fortalecida en una tesis que analiza las dinámicas de violencia y eficiencia educativa en Santiago de Cali (sin incluir aún la dimensión de resiliencia académica) [18]. Mediante la aplicación de un modelo condicional no paramétrico de orden-m, dicho estudio identificó una relación negativa y estadísticamente significativa entre los homicidios y la eficiencia educativa, además de evidenciar patrones de tolerancia institucional frente a contextos de criminalidad persistente. Este antecedente local constituye la base empírica y metodológica sobre la cual se desarrolla la presente investigación, que amplía el análisis al incorporar la variable de resiliencia académica. A continuación, se detalla la metodología empleada.

4. METODOLOGÍA

En línea con el objetivo principal de esta investigación, el cuál busca construir modelos de aprendizaje de máquina y aprendizaje estadístico para analizar la incidencia de los homicidios en la interacción entre eficiencia educativa y resiliencia académica en la ciudad de Cali entre los años 2014 y 2024, se identificaron cuatro etapas principales. En la primera se prepararon los datos recolectados del ICFES, DANE y Secretaría de Seguridad y Justicia de Cali mediante procesos de limpieza y exploración. La importancia de esta etapa radica en que permite conocer los datos a profundidad y ajustar inconsistencias que pueden afectar los modelos desarrollados. De manera particular, se adelantaron las siguientes tareas:

1. Recopilar datos históricos del ICFES, DANE y la Secretaría de Seguridad y Justicia de Cali necesarios para esta investigación.
2. Organizar los datos en un formato estructurado para facilitar el análisis.
3. Realizar procesos de limpieza para eliminar duplicados, inconsistencias y valores nulos.
4. Estandarizar las variables para garantizar compatibilidad entre fuentes de datos.
5. Explorar los datos mediante estadísticas descriptivas para identificar tendencias y patrones iniciales.

En coherencia con el marco teórico expuesto y los objetivos de la investigación, se adopta un enfoque de *machine learning* supervisado como herramienta para modelar la resiliencia académica en contextos educativos caracterizados por alta heterogeneidad socioeconómica y territorial. Este enfoque resulta pertinente dado que la relación entre desempeño académico, condiciones socioeconómicas, dotación institucional y violencia no responde a patrones lineales simples, sino que emerge de interacciones complejas entre múltiples dimensiones individuales, escolares y contextuales. En este sentido, el uso de algoritmos supervisados permite estimar dichas relaciones sin imponer supuestos paramétricos restrictivos, privilegiando la capacidad del modelo para generalizar y capturar no linealidades relevantes para el análisis empírico.

Adicionalmente, el objetivo del modelo no se limita a la predicción de resultados, sino que busca contribuir a la comprensión de los factores asociados a la resiliencia académica en entornos adversos. Por esta razón, se opta por un algoritmo de tipo ensamble —Random Forest— complementado con técnicas de interpretabilidad basadas en valores SHAP, lo que permite identificar la contribución relativa de cada variable explicativa en la estimación del fenómeno de interés. Esta combinación metodológica facilita la articulación entre el análisis predictivo y la

discusión sustantiva sobre eficiencia educativa, desigualdad socioeconómica y violencia homicida, asegurando que los resultados del modelo puedan ser interpretados y contextualizados dentro del marco analítico de la investigación.

En la segunda etapa, se buscó entrenar y validar un modelo de *machine learning* que identifique la cantidad de estudiantes resilientes en las instituciones educativas de Cali con base en el contexto socioeconómico del colegio, la disposición interna de recursos y los homicidios cercanos. Para llegar a dichos resultados, primero se evaluó el estado de la resiliencia académica para cada sede educativa a través de la propuesta de López-Estrada et.al [2] que se modela en (1) donde $Globalscore$ mide el desempeño académico; $INSE_{ij}$ representa el Índice Socioeconómico del estudiante, y los términos e_{ij} y δ_{0j} reflejan la variación individual y municipal, respectivamente, por lo que para clasificar como resilientes, los estudiantes deben estar en el cuartil inferior del $INSE$ y en el cuartil superior del puntaje académico.

$$Globalscore = a_{00} + \beta_1 INSE_{ij} + e_{ij} + \delta_{oj} \quad (1)$$

Con la identificación de los estudiantes resilientes, se procedió a integrar esta información con variables institucionales y contextuales, incorporando de manera explícita la dimensión espacial de la violencia. En particular, se estimó la correlación espacial entre la ubicación geográfica de las instituciones educativas y el histórico de homicidios registrados en su entorno inmediato, utilizando buffers de diferentes radios. Este procedimiento permitió construir indicadores anuales y agregados de exposición a la violencia, capturando tanto la intensidad como la persistencia longitudinal de los homicidios en los alrededores de cada sede educativa, elementos que se consideran estructuralmente relacionados con los procesos educativos y el desempeño académico.

Una vez logrados estos datos, se prosiguió a unificar la base de datos panel con la información provista por el DANE en su C600. Así, y llegado a este punto se testaron dos tipos de aproximaciones basadas en la propuesta de [56] random forest con SHAP con y sin iteraciones eliminación recursiva de características. Acá se dividen los datos de entrenamiento y de prueba y se realizan testeos con diferentes hiperparametros.

Una vez consolidada la base de datos final, se entrenó un modelo de machine learning supervisado de tipo Random Forest con el objetivo de estimar la probabilidad de resiliencia académica a nivel institucional. El conjunto de variables explicativas incluyó indicadores

socioeconómicos, características internas de las instituciones educativas relacionadas con recursos humanos y tecnológicos, así como medidas de violencia homicida construidas a partir del análisis espacial. Previo al entrenamiento, se realizó un proceso de preparación de datos que incluyó la imputación de valores faltantes mediante la mediana, con el fin de preservar la robustez frente a distribuciones asimétricas y valores atípicos, así como el balanceo de clases para corregir el desbalance natural entre estudiantes resilientes y no resilientes.

Con el propósito de evaluar la idoneidad del enfoque seleccionado, se exploraron distintas configuraciones del algoritmo, incluyendo aproximaciones basadas en Random Forest Regression con iteraciones sucesivas y análisis de interpretabilidad mediante valores SHAP, así como modelos de Random Forest Classifier acompañados de SHAP sin procesos iterativos. Estas alternativas permitieron contrastar el comportamiento predictivo y la estabilidad del modelo bajo diferentes formulaciones; sin embargo, sus resultados no mostraron mejoras sustantivas en términos de desempeño ni de consistencia interpretativa frente al enfoque finalmente adoptado, aspecto que se analiza con mayor detalle en la sección de resultados.

El modelo seleccionado fue entrenado utilizando particiones independientes de entrenamiento y prueba, y su desempeño fue optimizado mediante un proceso de ajuste de hiperparámetros basado en búsqueda aleatoria (randomized search) con validación cruzada. Este procedimiento permitió explorar de manera eficiente un amplio espacio de configuraciones del algoritmo, incluyendo el número de árboles, la profundidad máxima y el tamaño mínimo de las hojas, priorizando aquellas con mayor capacidad de generalización. La evaluación del modelo se realizó sobre datos no utilizados durante el entrenamiento, empleando métricas adecuadas para problemas de clasificación desbalanceada, lo que garantiza una estimación robusta del desempeño predictivo y sentó las bases para el análisis interpretativo posterior mediante técnicas SHAP.

Así pues esta etapa involucra las tareas:

1. Seleccionar características relevantes para estimar la resiliencia académica.
2. Dividir los datos en conjuntos de entrenamiento y prueba.
3. Implementar algoritmos de *machine learning* adecuados para el problema.
4. Ajustar hiper parámetros para optimizar el rendimiento del modelo.
5. Documentar el proceso de entrenamiento, incluyendo las métricas obtenidas y los parámetros utilizados.

Posteriormente, con el fin de validar el desempeño del modelo de machine learning y verificar su efectividad predictiva, se definió un esquema de evaluación basado en métricas de rendimiento apropiadas para el tipo de problema abordado. En particular, se seleccionaron métricas de clasificación y regresión, tales como precisión, F1-score, error absoluto medio (MAE) y error cuadrático medio (MSE), según la formulación específica del modelo evaluado. Estas métricas fueron calculadas sobre el conjunto de prueba, garantizando que la evaluación se realizara sobre datos no utilizados durante el entrenamiento.

Los resultados obtenidos fueron contrastados con valores de referencia reportados en la literatura especializada, lo que permitió contextualizar el desempeño del modelo dentro de estudios similares en los ámbitos de violencia, educación y resiliencia académica. Adicionalmente, se realizó un análisis orientado a identificar posibles sesgos o patrones sistemáticos de error, con el fin de evaluar la estabilidad del modelo y, cuando fue necesario, ajustar el proceso de entrenamiento. Finalmente, se generaron reportes detallados que documentan el desempeño predictivo, la consistencia de los resultados y las principales limitaciones del modelo, sentando las bases para su interpretación y discusión en las secciones posteriores.

Para evaluar la eficiencia educativa considerando el contexto socio-espacial, se implementa un modelo condicional de orden-m basado en Free Disposal Hull (FDH) propuesto por Cazals et al. [60]. Este método no paramétrico presenta ventajas sustanciales frente a modelos tradicionales de frontera, particularmente en su robustez ante valores atípicos y su capacidad para incorporar factores contextuales no controlables por las instituciones educativas.

La tecnología de producción educativa se define como el conjunto de combinaciones viables de entradas (*inputs*), salidas (*outputs*) y factores contextuales como en la siguiente ecuación (2):

$$\Psi = \{(x, y, \zeta) \in \mathbb{R}_{+}^{p+q+k} : y \in \mathcal{Y}(x; \zeta)\} \text{ donde } y \text{ puede ser producido por } x \text{ dado } \zeta \quad \textbf{(2)}$$

y a su vez:

- $x \in \mathbb{R}_{+}^{p}$ representa los *inputs* educativos (alumnos, personal docente, infraestructura tecnológica)
- $y \in \mathbb{R}_{+}^{q}$ son los *outputs* (resiliencia académica agregada, puntaje global)

- $\zeta \in \mathbb{R}_+^k$ son variables contextuales no discrecionales (homicidios en radio de 250m, nivel socioeconómico vecinal)

El análisis de estos resultados se estructura siguiendo el enfoque two-stage de Daraio y Simar [61], [62] compuesto por dos etapas. La primera de Estimación de eficiencias donde Se calculan scores de eficiencia mediante $\text{order-}m\ \text{FDH}$ con orientación output. Para cada institución educativa k, el score de eficiencia θ_k se estima como (3):

$$\theta_k = \mathbb{E}_m[\max\{\lambda : (x_k, \lambda y_k) \in \Psi_m\}] \quad (3)$$

donde Ψ_m denota la frontera parcial construida con una muestra aleatoria de tamaño m. El parámetro m controla el grado de robustez (valores menores enfatizan comparaciones locales mientras valores mayores se aproximan a la frontera completa). El proceso se replica $B = 200$ veces mediante bootstrap para obtener estimadores estables. Se calculan dos tipos de eficiencias:

- No condicional: utiliza un muestreo uniforme de unidades comparables (peers) del conjunto total de observaciones.
- Condicional: pondera el muestreo según la similitud contextual mediante una función kernel.

La segunda etapa de explicación de eficiencias emplea una regresión no paramétrica de tipo local-linear para modelar el ratio de eficiencias (condicional/no condicional) en función de las variables contextuales:

$$Ratio(\zeta) = m(\zeta) + \varepsilon \quad (4)$$

donde $m(\cdot)$ se estima mediante kernel regression con *bandwidth* óptimo seleccionado por *cross-validation*.

Los efectos marginales $\dfrac{\partial m}{\partial \zeta_j}$ se calculan analíticamente para cuantificar el impacto de cada variable contextual. La significancia estadística se evalúa mediante bootstrap con $B = 100$

repeticiones[2] y para verificar la robustez de los hallazgos, se implementan los siguientes controles:

1. Análisis de sensibilidad del parámetro m: Se comparan resultados para $m \in \{15, 30, 50\}$para evaluar estabilidad de las eficiencias estimadas y efectos contextuales
2. Filtros de calidad de datos:
    - Variables input con $> 70\%$ de valores faltantes son excluidas, las variables restantes fueron incluidas en el análisis.
    - Instituciones con$> 60\%$ de las observaciones incompletas son eliminadas del panel.
3. Validación temporal: Se analiza la evolución de percentiles de eficiencia condicional a través de los 11 años del panel (2014-2024) para identificar tendencias y cambios estructurales
4. Test de significancia: Se aplican tests bootstrap no paramétricos para evaluar si los efectos contextuales estimados $H_0 : \text{Efectos contextuales} = 0$ son estadísticamente rechazables.

Esta estrategia metodológica permite: (i) cuantificar la eficiencia educativa ajustando por factores no controlables, (ii) identificar la relación de la violencia sobre el desempeño educativo, y (iii) proporcionar evidencia robusta para políticas públicas diferenciadas según contexto socio-espacial. Esta etapa integra los resultados obtenidos en las fases anteriores con el fin de analizar en conjunto la influencia de los homicidios sobre la eficiencia educativa y resiliencia académica para generar el reporte final de hallazgos de la investigación.

[2] Una discusión más nutrida sobre los valores óptimos en bootstrap para pruebas de eficiencia se encuentra en los aportes de Aliana, Prior & Tortosa-Ausina [63] de donde se obtuvo este valor de referencia que, de cualquier manera y como se verá, se probó con diferentes escenarios.

5. OBTENCIÓN Y PREPARACIÓN DE DATOS

El proceso de limpieza y extracción de datos inició con la obtención de las bases de datos requeridas para el periodo entre 2014 y 2024, por un lado, la información respecto a los homicidios en Cali proporcionada por la Secretaría de Seguridad y Justicia de Santiago de Cali mediante derecho de petición y por otro lado, las bases de datos relacionadas a educación, las cuáles son de acceso público, y que fueron, los resultados de la Prueba de estado Saber 11 y Caracterización de colegios C600. De ellas se profundizará a continuación.

5.1 DATOS SOBRE EDUCACIÓN

5.1.1 Prueba de estado Saber 11

La Prueba de Estado Saber 11 (Examen de Estado de la Educación Media) es un examen estandarizado oficial aplicado en Colombia al finalizar la educación media (grado 11, último año de bachillerato). Corresponde al examen de egreso de la secundaria y es comparable a las pruebas de acceso universitario en otros países, tales como el SAT en Estados Unidos o la Selectividad en España. Este examen es administrado a nivel nacional por el Instituto Colombiano para la Evaluación de la Educación (ICFES), la entidad gubernamental encargada de las evaluaciones educativas en el país.

El objetivo principal de Saber 11 es comprobar el nivel de desarrollo de las competencias académicas básicas en los estudiantes que están por culminar la educación secundaria [64]; en otras palabras, la prueba evalúa cuánto han aprendido los alumnos en las áreas fundamentales durante su trayectoria escolar y qué tan preparados están para afrontar la educación superior, de hecho el puntaje del examen es determinante en los procesos de admisión a la educación superior. Además, Saber 11 funciona como un instrumento oficial para medir la calidad de la educación media en el país, al evaluar de forma estandarizada los logros educativos de quienes terminan el nivel de bachillerato, aunque no existe un puntaje mínimo aprobatorio, la presentación del Saber 11 es de carácter obligatorio para obtener el título de bachiller en Colombia.

La estructura de Saber 11 abarca cinco áreas o componentes evaluativos del currículo de educación media, y todas las preguntas del examen son de selección múltiple. Dichas áreas son Lectura Crítica, Matemáticas, Sociales y Ciudadanas, Ciencias Naturales e Inglés. A través de estas pruebas se miden competencias académicas básicas indispensables para afrontar la educación superior, tales como la comprensión de textos, el razonamiento cuantitativo, las habilidades cívicas, el razonamiento científico y el dominio de un segundo idioma.

Para obtener acceso a las bases de datos de esta prueba u otras que desarrolla el ICFES para otros ciclos educativos (Saber 3°, 5° y 9°, Saber TyT, Saber Pro u otras) basta con ingresar a la página oficial https://www.icfes.gov.co/investigaciones/data-icfes/ y registrarse en el portal

DataIcfes para que se habilite un usuario que pueda descargar la información de interés en el portal.

La información se encuentra disponible por año y por periodo de aplicación (1 o 2) a nivel nacional. Para los fines de esta investigación, se descargaron los registros correspondientes a todos los años comprendidos entre 2014 y 2024, incluyendo ambos periodos de aplicación de cada año. Dado que los datos se reportan a nivel nacional, la información fue posteriormente filtrada para el municipio de Cali, Colombia, mediante la variable cole_cod_mcpio_ubicacion, seleccionando el valor 7600. Finalmente, se verificó que la variable cole_mcpio_ubicacion presentara de manera consistente el valor "CALI", correspondiente al descriptivo oficial del código municipal "7600".

Como resultado del proceso de depuración y filtrado, la base de datos final de la Prueba Saber 11 para la ciudad de Cali, en el período de interés, quedó conformada por 251.755 registros, correspondientes a estudiantes que presentaron el examen entre 2014 y 2024, abarcando 22 períodos académicos (dos aplicaciones por año). El conjunto de datos incluye información de 497 establecimientos educativos únicos, con un promedio de 506,5 estudiantes por colegio, y una distribución heterogénea que oscila entre 1 y 6.940 estudiantes por institución que presentan el exámen en ese periodo.

El dataset está compuesto por 177 variables, de las cuales 55 son numéricas (continuas y discretas) y 122 categóricas (nominales y ordinales). Estas variables pueden agruparse en cinco grandes categorías: (i) identificación y control, (ii) desempeño académico, (iii) características del estudiante, (iv) características del establecimiento educativo y (v) características familiares y socioeconómicas. En términos de completitud, 23 variables (13%) no presentan valores faltantes, mientras que 154 variables (87%) contienen información incompleta, situación asociada principalmente a preguntas que solo fueron recolectadas en determinados años, algunas en las que no es obligatorio responder debido a ser información sensible o problemas de codificación de los datos.

Dado que el análisis de resiliencia educativa se realiza a nivel de establecimiento, de todo el data set esta investigación se concentra en un subconjunto de variables, el puntaje global del estudiante (punt_global), el período de presentación de la prueba (periodo), el código único del establecimiento educativo (cole_cod_dane_establecimiento) y el Índice Socioeconómico Individual del estudiante (estu_inse_individual).

La variable periodo identifica el año y semestre en que el estudiante presentó la prueba, permitiendo estructurar el análisis temporal de los resultados académicos. Por su parte, cole_cod_dane_establecimiento corresponde al identificador único asignado por el DANE a cada institución educativa, lo que facilita la agregación de resultados a nivel institucional y la vinculación con las otras 2 fuentes de datos de esta investigación (C600 y homicidios en Cali).

El Índice Socioeconómico Individual (INSE) es una variable continua que resume las condiciones socioeconómicas del estudiante a partir de información del hogar y del entorno familiar. En la base analizada, el INSE presenta una media de 54,22, una desviación estándar de 8,70 y valores entre 14,90 y 88,36. Finalmente, el puntaje global representa una medida sintética del desempeño académico del estudiante en las diferentes áreas evaluadas en Saber 11; esta variable muestra una media de 260,06 puntos, una desviación estándar de 51,16 puntos y un rango que va de 0,00 a 495,00 puntos, constituyéndose en el principal indicador de rendimiento académico utilizado en el análisis; estas dos últimas variables van a ser abordadas con mayor detalle en la exploración de los datos.

5.1.2 Formulario C600

La otra base de datos de educación, se construyó con base en las respuestas del formulario C600 de la operación estadística de Educación Formal (EDUC del DANE) el cuál es un censo anual que recopila información detallada sobre la educación preescolar, básica (primaria y secundaria) y media en Colombia, para proporcionar información estadística estratégica que permite formular políticas, planificar y administrar la educación formal en todos los niveles de gobierno (municipal, departamental y nacional), esta información es de acceso público [65].

En el formulario C600 se recolectan un amplio conjunto de variables educativas segmentado por 9 módulos que registra datos sobre matrícula de estudiantes (desagregada por nivel, grado, sexo, zona y sector), aprobación, reprobación y deserción; número y características de docentes y personal; infraestructura y recursos tecnológicos disponibles; poblaciones especiales (étnicas, víctimas, estudiantes con discapacidad, entre otras) y modelos educativos flexibles o de adultos. Tras revisar todas las variables, se seleccionaron las variables de cantidad de personal ocupado (**TABLA I**), cantidad de TIC's (Tecnologías de la Información y la Comunicación) en la institución educativa (**TABLA II**) y cantidad de docentes segmentado por nivel educativo (**TABLA III**), dicha clasificación se presenta en las siguientes tablas ya que gran parte de ellas fueron usadas en el modelo.

TABLA I

CODIFICACIÓN DE PERSONAL OCUPADO

Codificación	Descripción personal ocupado
PO_ID_1	Directivo docente (rector(a) y coordinadores académicos, disciplinarios, etc)
PO_ID_2	Docentes de aula
PO_ID_3	Administrativos (vigilantes, personal de aseo, secretarios, contadores, etc)

PO_ID_4	Docente de apoyo en aula (para estudiantes con discapacidad o con capacidades excepcionales)
PO_ID_5	Personal de apoyo en aula (No docente): profesionales apoyo para estudiantes con discapacidad, maestros bilingües, intérpretes y mediadores
PO_ID_6	Docentes con labores administrativas (sin asignación o carga académica)
PO_ID_7	Docentes orientadores (con enfoque psicosocial, convivencia escolar, incluya psicólogos, trabajador social y demás relacionados)

TABLA II

CODIFICACIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN

Codificación	**Descripción**
TICS_ID_1	Computadores de escritorio
TICS_ID_2	Computadores portátiles
TICS_ID_3	Tabletas

TABLA III

CODIFICACIÓN DE NIVEL MÁXIMO ALCANZADOS POR EL DOCENTE

Codificación	**Descripción de nivel máximo alcanzado por el docente**
DOCNIV_ID_1	Bachillerato pedagógico
DOCNIV_ID_2	Bachillerato técnico
DOCNIV_ID_3	Normalista superior
DOCNIV_ID_4	Etnoeducador
DOCNIV_ID_5	Perito experto o técnico en educación
DOCNIV_ID_6	Tecnólogo en educación
DOCNIV_ID_7	Licenciado
DOCNIV_ID_8	Profesional diferente a licenciado
DOCNIV_ID_9	Posgrado en educación o programa pedagógico
DOCNIV_ID_10	Posgrado en programa no pedagógico
DOCNIV_ID_11	Instructor I, II y A
DOCNIV_ID_12	Instructor III y B

DOCNIV_ID_13	Instructor IV y C
DOCNIV_ID_14	Sin titulación o acreditación educativa
DOCNIV_ID_15	Técnico o tecnólogo diferente a educación

Una vez definidas las variables de interés, se realizó un proceso de depuración orientado a garantizar la consistencia temporal y estructural de la información. Este proceso incluyó la verificación de duplicados, la estandarización de los identificadores institucionales (código DANE de establecimiento) y la armonización de las series anuales, considerando que algunas variables del C600 presentan cambios en su definición o cobertura a lo largo del periodo de estudio. Adicionalmente, se evaluó la presencia de valores faltantes y se aplicaron criterios de exclusión o imputación según el grado de completitud de cada variable, priorizando aquellas con mayor estabilidad y continuidad interanual para evitar sesgos derivados de información incompleta.

Con el fin de facilitar la comparabilidad entre instituciones educativas de distinto tamaño, las variables seleccionadas fueron transformadas en indicadores relativos, tales como razones entre estudiantes y personal docente o administrativo, así como estudiantes por dispositivo tecnológico disponible. Este enfoque permitió capturar de manera más precisa la disponibilidad efectiva de recursos educativos, evitando que los resultados estuvieran dominados por el volumen absoluto de matrícula. Finalmente, la información procesada del C600 fue integrada con los resultados de la Prueba Saber 11, agregando los datos a nivel de institución y año, lo que permitió construir una base panel coherente que articula características institucionales, condiciones socioeconómicas y resultados académicos, sentando las bases para los análisis posteriores de resiliencia y eficiencia educativa.

5.2 DATOS SOBRE HOMICIDIOS

La información sobre homicidios fue obtenida a partir de registros administrativos suministrados por la Secretaría de Seguridad y Justicia de Santiago de Cali, a los cuales se accedió mediante un derecho de petición formal. Esta base de datos contiene información detallada sobre los homicidios ocurridos en el área urbana del municipio durante el periodo de estudio, incluyendo variables temporales (fecha y hora del hecho), espaciales (coordenadas geográficas del evento) y características generales del tipo de violencia.

En una primera fase, se realizó la limpieza de los registros eliminando observaciones con coordenadas faltantes o inconsistentes, así como duplicados y registros con errores evidentes en las variables temporales. Posteriormente, se estandarizaron los formatos de fecha y se extrajeron variables temporales derivadas, como el año y el semestre de ocurrencia del hecho, con el fin de facilitar la agregación y el análisis longitudinal. Este tratamiento permitió construir

series temporales coherentes que capturan la evolución de la violencia homicida en la ciudad a lo largo del periodo 2014–2024.

Una vez depurada la base, los homicidios fueron transformados en una capa espacial de eventos puntuales, lo que permitió analizar su distribución geográfica y su relación con la ubicación de las instituciones educativas. A partir de esta georreferenciación, se construyeron indicadores de exposición a la violencia mediante el cálculo del número de homicidios ocurridos dentro de radios de influencia (buffers) de distinto tamaño alrededor de cada institución educativa. En particular, se consideraron múltiples distancias con el fin de evaluar la sensibilidad de los resultados al alcance espacial de la violencia y capturar tanto efectos inmediatos como patrones de concentración en el entorno cercano.

Adicionalmente, se realizaron análisis descriptivos y exploratorios para identificar patrones espaciales y temporales de la violencia homicida, incluyendo medidas de concentración y autocorrelación espacial. Estos análisis están contenidos en el siguiente capítulo.

5.1 UNIFICACIÓN DE DATOS

La unificación se realizó construyendo una base maestra a nivel sede–año, donde el identificador educativo (código DANE) y el periodo permiten integrar los componentes del C600 (recursos, personal y docentes) con las variables educativas derivadas de la Prueba Saber 11 y con la exposición a homicidios estimada espacialmente. En términos operativos, los homicidios se incorporan como un agregado contextual por institución (conteo dentro del buffer definido) y por año, mientras que el C600 aporta los insumos institucionales y dotacionales; el resultado es una base de datos unificada con variables de identificación del establecimiento, ubicación geográfica (latitud y longitud), año, recursos tecnológicos (TIC), personal y docentes, además de los campos necesarios para el resto del análisis, tales como totales y proporciones construidas posteriormente. La definición específica del radio del buffer utilizado para la medición de homicidios se discute en la sección de resultados, en la medida en que constituye en sí misma uno de los hallazgos empíricos de la investigación.

Como resultado de este proceso de integración, se obtuvo una base de datos de tipo panel compuesta por 5.467 observaciones, correspondientes a 497 instituciones educativas únicas, con cobertura temporal entre los años 2014 y 2024. La estructura resultante corresponde a un panel no balanceado, dado que no todas las instituciones cuentan con información completa para cada año del periodo analizado, situación habitual en ejercicios que integran múltiples fuentes administrativas y censales. Esta configuración permite capturar tanto variaciones interinstitucionales como dinámicas temporales en las condiciones educativas y contextuales.

La base unificada incorpora de manera explícita la dimensión espacial de la violencia mediante la variable Homicidios_250m, que cuantifica el número de homicidios ocurridos dentro de un radio de 250 metros alrededor de cada sede educativa en cada año. En términos descriptivos, las

instituciones registran en promedio aproximadamente 2,15 homicidios anuales en su entorno cercano; no obstante, la distribución presenta una alta dispersión, con valores máximos que alcanzan hasta 19 homicidios en un solo año, lo que evidencia la existencia de focos persistentes de violencia urbana y refuerza la pertinencia de incorporar esta variable como un determinante contextual en el análisis educativo.

Desde el punto de vista educativo e institucional, la base refleja una elevada heterogeneidad en el tamaño y la dotación de los establecimientos. La matrícula total reportada presenta valores que oscilan entre 4 y 4.130 estudiantes, lo que pone de manifiesto la coexistencia de instituciones de muy distinta escala dentro del sistema educativo del municipio. Asimismo, la información sobre personal ocupado, docentes y recursos tecnológicos permite capturar diferencias sustantivas en la capacidad institucional, aunque algunas variables altamente desagregadas, especialmente aquellas relacionadas con niveles específicos de formación docente, presentan una alta proporción de ceros o valores faltantes, aspecto considerado explícitamente en los procesos de selección y limpieza de variables.

Finalmente, para asegurar la consistencia del panel y evitar sesgos derivados de información incompleta, se aplicaron filtros de calidad en dos niveles. En primer lugar, se excluyeron variables con una alta proporción de valores faltantes cuando estos no podían ser recuperados de manera consistente entre años. En segundo lugar, se eliminaron observaciones correspondientes a instituciones educativas con un nivel elevado de incompletitud, definido como un umbral del 30% de valores nulos sobre el conjunto de variables retenidas. Este procedimiento se alinea con la estrategia metodológica de control de calidad de datos adoptada en el estudio y permitió garantizar que la base final cuente con información suficiente y consistente para los análisis multivariados y de machine learning desarrollados en las secciones posteriores.

6. ANÁLISIS EXPLORATORIO DE DATOS (EDA)

Si bien en el desarrollo de la investigación se realizó un análisis exploratorio exhaustivo a nivel variable por variable, este documento presenta únicamente los hallazgos más relevantes para los objetivos planteados. Esta decisión responde a criterios de síntesis y claridad, evitando una extensión innecesaria del texto, sin comprometer el rigor analítico ni la trazabilidad metodológica. El análisis exploratorio de datos (EDA) fue fundamental en la comprensión de la estructura, distribución y calidad de la información, así como en la identificación de patrones preliminares que generaron hipótesis y orientaron las decisiones posteriores de modelación y análisis espacial.

En esta sección se sintetizan los resultados más significativos del EDA, enfocándose en aquellas dimensiones que permiten contextualizar la relación entre recursos educativos, desempeño académico, resiliencia y homicidios en Santiago de Cali. Los análisis presentados buscan ofrecer una visión descriptiva integrada del fenómeno de estudio, sirviendo como antelación empírica a los modelos de machine learning y eficiencia educativa desarrollados en las secciones posteriores.

6.1 PANORAMA GENERAL DE LA EDUCACIÓN Y LA RESILIENCIA ACADÉMICA EN CALI

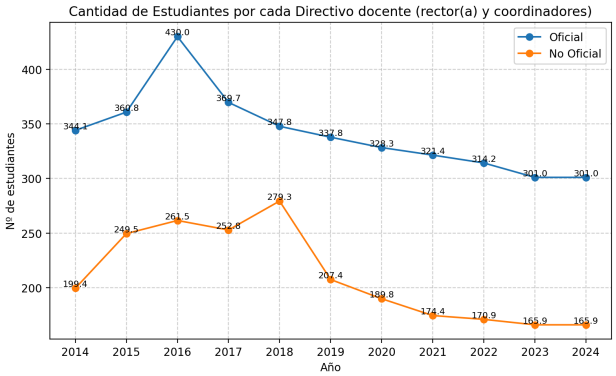
A partir de información proveniente del ICFES y del DANE, a continuación se describen diferencias estructurales entre instituciones oficiales y no oficiales, así como tendencias temporales relevantes que permiten contextualizar los resultados académicos observados en la ciudad. El objetivo de este apartado no es establecer relaciones causales, sino caracterizar el entorno educativo en el cual emergen los patrones de resiliencia académica que serán analizados posteriormente.

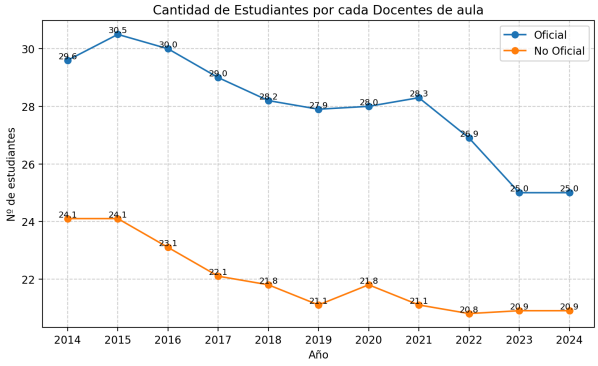
Respecto al C600, donde se pueden generar indicadores asociados al personal educativo, los recursos tecnológicos, la composición institucional y el desempeño agregado de los estudiantes, se evidencia que el personal ocupado (PO) presenta diferencias entre instituciones oficiales y no oficiales, la siguiente figura (**Fig. 1**) permiten comprender por cada categoría de PO cuántos estudiantes hay en promedio de asignación en cada uno de los años; es decir, su asignación relativa al número de estudiantes.

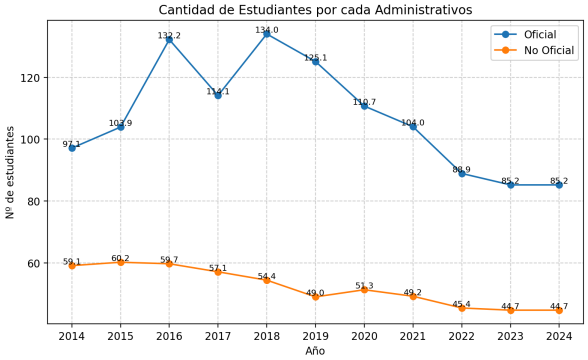
En directivos docentes, los colegios oficiales pasan de 344 estudiantes por directivo en 2014 a 301 en 2024, por lo que hay menor disponibilidad de este personal frente al sector no oficial, que se reduce de 199 a 165 en el mismo período. Con los docentes de aula, el sector oficial mantiene valores ligeramente superiores (29–25 estudiantes por docente) frente al no oficial (24–21 estudiantes por docente). En administrativos, los colegios oficiales tienen mayores

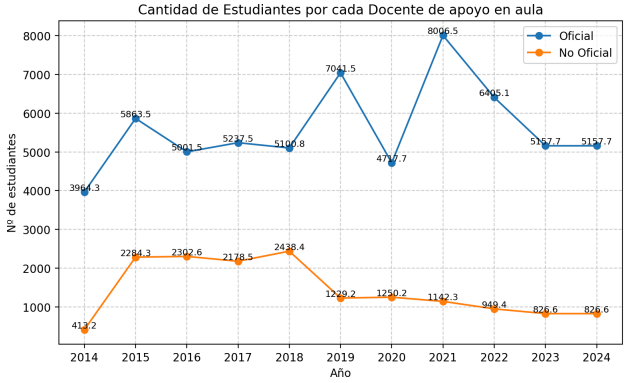
estudiantes por administrativo (97) y la reducen a 85, mientras que en el sector No oficial pasa de 59 a 44, de nuevo, menos estudiantes por cada PO.

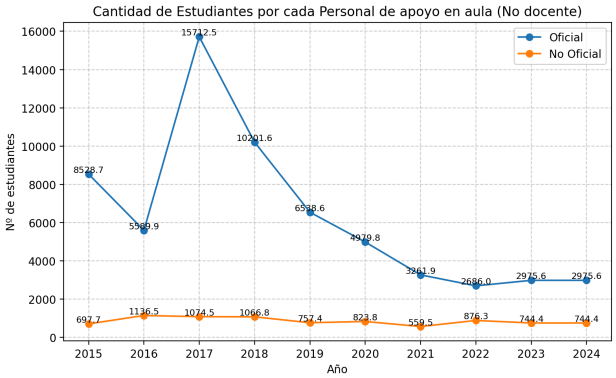
Se observa una gran diferencia con los docentes de apoyo especializado (apoyo en el aula, apoyo no docente y de labores administrativas) ya que en el sector oficial los valores de estudiantes por PO de este tipo llegan a los miles de estudiantes, en general este comportamiento es esperado debido al presupuesto limitado en el sector oficial.











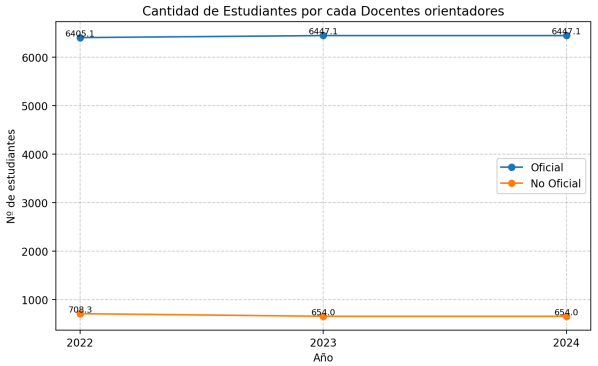


Fig. 1. Evolución del número de estudiantes por tipo de personal educativo y administrativo en instituciones oficiales y no oficiales (2014–2024)

(a) Directivos docentes. (b) Docentes de aula. (c) Administrativos. (d) Docentes de apoyo en aula. (e) Personal de apoyo en aula (no docente). (f) Docentes orientadores

Nota. Algunos gráficos no abarcan todo el rango de años definido, ya que la variable correspondiente solo fue registrada en periodos posteriores.

A continuación se puede ver la relación de estudiantes por equipo TIC entre el 2014 y el 2024 segmentado entre colegios oficiales y no oficiales (**Fig. 2**). En computadores de escritorio (TICS_ID_1) los colegios oficiales parten con 27,3 estudiantes por equipo en 2014 y aumentan hasta valores entre 35 y 37 en los últimos años, su adquisición se mantiene estable pero menor a la del portatil, en contraste el sector no oficial asigna 15,3 a 10,9 estudiantes por computador, menos estudiantes que el sector oficial. En computadores portátiles (TICS_ID_2) el comportamiento es inverso, el sector oficial tiene de 6 a 9 estudiantes por equipo mientras que el no oficial alrededor de 28 a 36 en los últimos años.

En el caso de las tabletas (TICS_ID_3), se evidencia que desde el 2021 se está disminuyendo su adquisición, en el 2014 se ve que este dispositivo no era adquirido por el sector oficial, para el siguiente año se estabiliza y mantiene entorno a 7-10 estudiantes por tableta, el sector no oficial muestra una asignación estable e inferior al no oficial ya que tiene casi 10 veces menos dispositivos por estudiante.

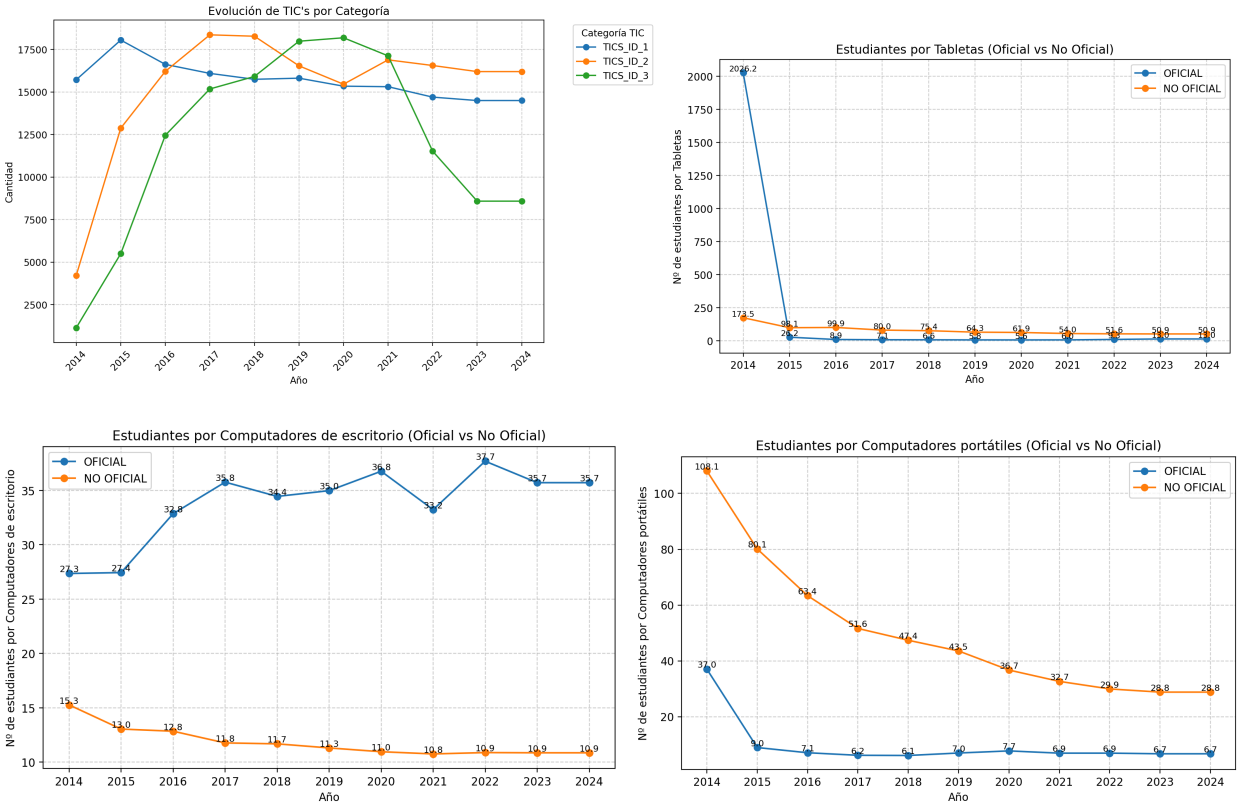


Fig. 2. Evolución de los recursos TIC y de la relación estudiantes–dispositivos en instituciones oficiales y no oficiales (2014–2024)

(a) Evolución de TIC 's por categoría. (b) Estudiantes por tabletas TICS_ID_3. (c) Estudiantes por computadores de escritorio TICS_ID_1. (d) Estudiantes por computadores portátiles TICS_ID_2.

De manera consistente a lo largo de todo el periodo 2014–2024, la mayor proporción de docentes pertenece al nivel Licenciado, seguido por Posgrado en educación y Profesionales no licenciados, por el contrario, niveles como Instructor, Etnoeducador y Posgrado no educativo presentan valores muy bajos o nulos (**Fig. 3**). Hay que hacer énfasis en que hay una ligera disminución a través de los años en el número total de licenciados, acompañado de un incremento en los docentes con posgrado en educación.

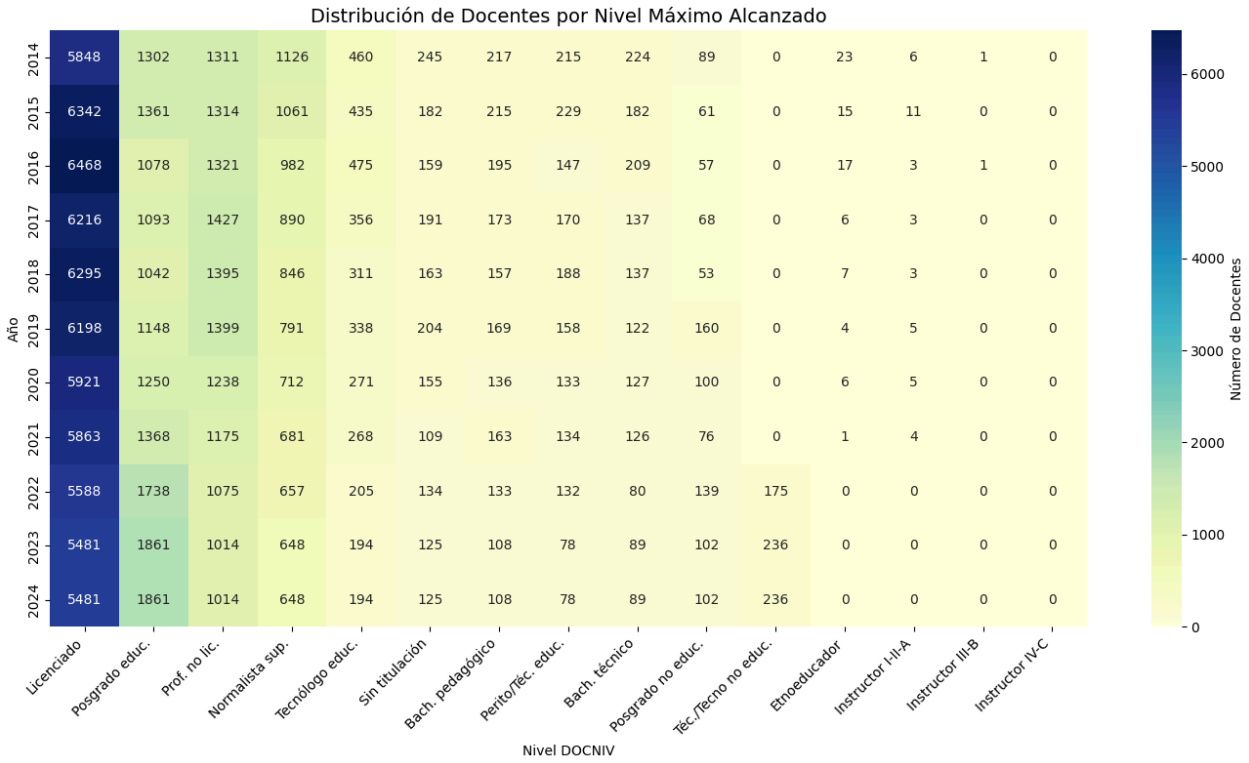


Fig. 3. Distribución de docentes según el nivel máximo de formación alcanzado en el periodo 2014–2024

Resiliencia Académica

Como se comentaba con anterioridad, se recopiló la información de los puntajes de la prueba de Estado Saber 11, elaborada por el Instituto Colombiano para la Evaluación de la Educación (ICFES), que va dirigida a estudiantes que finalizan la educación media (grado 11). Los datos se obtuvieron mediante solicitud a dicha entidad, la cual habilitó el acceso a su repositorio público, para disponer de los resultados año tras año en el periodo de interés (2014-2024) que contienen los resultados de los estudiantes en las cinco áreas evaluadas por la prueba Saber 11, Lectura Crítica, Matemáticas, Ciencias Naturales, Sociales y Ciudadanas, e Inglés. Así mismo, ofrecen y permiten generar indicadores adicionales como el puntaje global o la clasificación de los planteles educativos, que ubica a cada institución en cinco categorías; A+, A, B, C y D, siendo A+ el mejor desempeño académico y D el más bajo [66].

La Fig. 4 muestra un incremento en el número de participantes en el segundo semestre, determinado por los registros únicos de la base de datos. Esta tendencia obedece a que la

prueba Saber 11 es exigida tanto como requisito de grado como de ingreso a la educación superior, por lo cual los estudiantes suelen inscribirse antes de culminar su último año escolar. El segundo semestre presenta un pico inicial en 2014 (18.842 registros) y luego mantiene una tendencia descendente, alcanzando uno de sus puntos más bajos en 2020 (15.144), posiblemente asociado a impactos externos como la pandemia, en años posteriores la cantidad de pruebas aplicadas vuelve a estabilizarse. En total hay 497 colegios que participaron en el periodo de tiempo, sin embargo, eso no garantiza que a la actualidad todos sigan vigentes debido que algunos pueden cerrar.

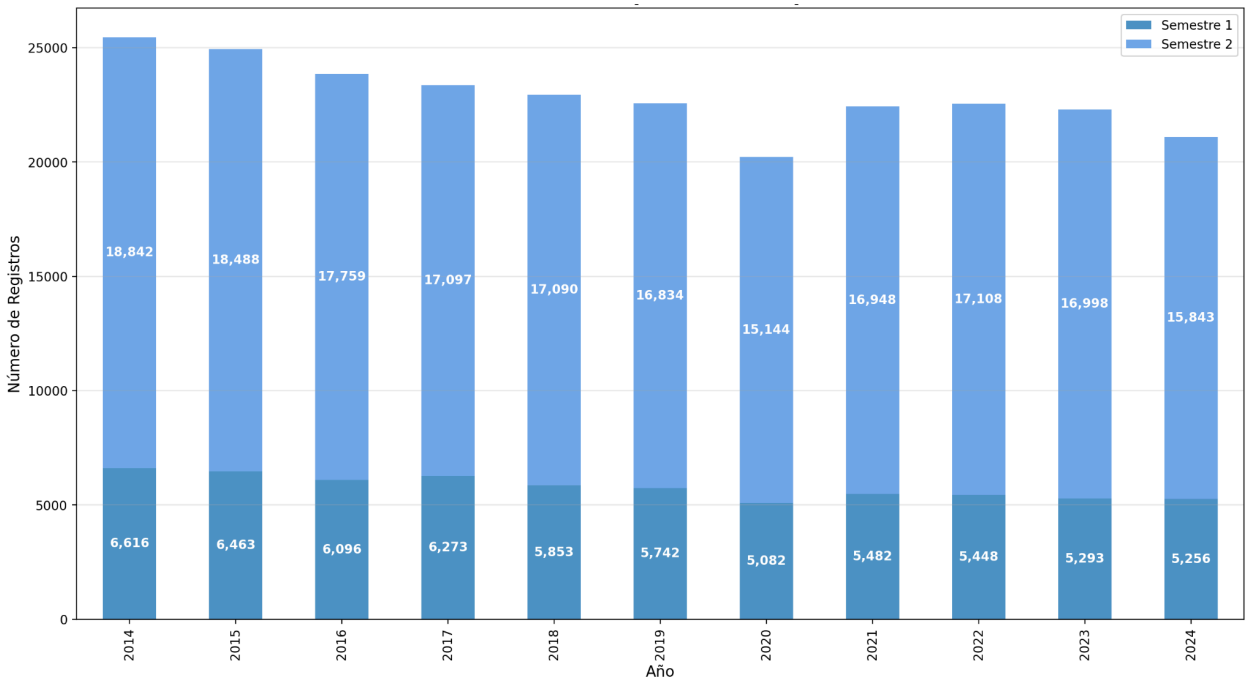


Fig. 4. Conteo absoluto de registros únicos (estudiantes) en la base de datos de la prueba Saber 11 del ICFES por semestre y año en instituciones oficiales y no oficiales (2014–2024)

La prueba Saber 11 no solo recolecta las respuestas a los ítems de conocimiento, también contempla preguntas a nivel socioeconómico para enriquecer análisis a nivel contextual [67]; a nivel de estudiantes extrae información como la edad o cohorte académica, género, zona de residencia (rural/urbana), nivel socioeconómico (estrato 1-6), características del hogar (tipo de vivienda, aseo, servicios), nivel educativo de los padres, autoidentificación étnica, condición de discapacidad, acceso a internet o computador, horas semanales de trabajo o lectura, auto-evaluación de probabilidad de ingreso a educación superior, entre otras.

A nivel de establecimiento, recopila el número de sede o DANE, sector (oficial, no oficial), zona de ubicación (rural/urbana), calendario (A o B), jornada (mañana/tarde/completa), programa bilingüe, entre otras. Por ello, al unir las bases de datos de todos los años, se obtienen en total 177 variables, cabe aclarar que hay preguntas que fueron incluidas o eliminadas en algunos años, por lo que hay ciertas variables que requieren mayor limpieza de valores nulos, sin embargo, de la base de datos general, las de interés para esta investigación son puntaje global 'punt_global', índice socioeconómico individual 'estu_inse_individual', el código único del establecimiento que permite vincular información con otras bases de datos 'cole_cod_dane_establecimiento' y el año de presentación, dichas variables permiten etiquetar a los estudiantes resilientes (1) y no resilientes (2).

A modo de énfasis, el INSE, variable fundamental para este trabajo, es un índice latente que estima el ICFES para representar el nivel socioeconómico del estudiante como constructo psicométrico, este se obtiene mediante las respuestas del cuestionario socioeconómico mencionado con anterioridad e integra análisis estadísticos que obtienen la posición del estudiante en una escala de 1 a 100, con este puntaje se clasifica el NSE (nivel socioeconómico) el cuál va de una escala de 1 (bajo) a 4 (alto) y refleja la condición socioeconómica del estudiante.

Dentro de los descriptores socioeconómicos que componen el INSE se encuentran el nivel educativo de la madre y el padre (incompleto, primaria, secundaria, profesional, posgrado), la ocupación laboral de los padres o encargados y la dotación del hogar utilizada como *proxy* del ingreso familiar (computador, internet, lavadora, horno microondas, automóvil, consola de videojuegos, televisor, etc.) [68] . De esta manera se establecen los siguientes niveles:

- NSE 1: puntos de corte INSE de 0 – 41,11
- NSE 2: puntos de corte INSE hasta 51,18
- NSE 3: puntos de corte INSE hasta 64,08
- NSE 4: puntos de corte INSE hasta 100

La Fig. 5 (a) muestra que la distribución del INSE Individual (métrica proporcionada por el ICFES que es producto de un análisis detallado en [68]) presenta forma casi normal, con ligera asimetría negativa, concentrando la mayoría de los valores entre 45 y 60, su media es de 54,22. El diagrama de caja (**Fig. 5 (b)**) confirma que la mediana está en torno a 53,58, con presencia de valores atípicos tanto en el extremo inferior como en el superior, por lo que existen estudiantes con niveles socioeconómicos muy bajos o muy altos respecto a la media. El INSE individual presenta desviación estándar de 8,70, coeficiente de variación del 16,05%, asimetría de 0,350 y curtosis de 0,345.

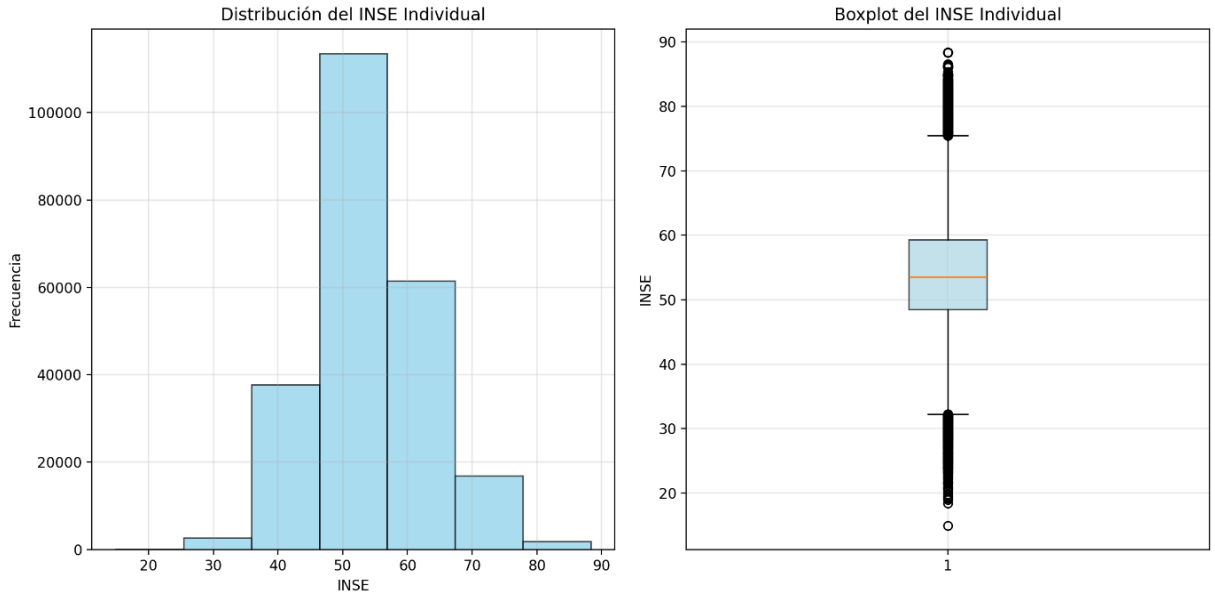


Fig. 5. Distribución y dispersión del INSE individual en la población analizada

(a) Histograma del INSE individual. (b) Boxplot del INSE individual

La otra variable de interés es el puntaje global, que se presenta en una escala de 0 a 500 para facilitar su interpretación y comparación, este se obtiene multiplicando por 5 el índice global (IG), el cual corresponde a una media ponderada de los puntajes en las cinco áreas evaluadas y se encuentra en una escala de 0 a 100.

El puntaje global, junto con el INSE, permite clasificar a los estudiantes resilientes. En esta investigación, se considera estudiante resiliente aquel cuyo puntaje global se ubicó por encima del percentil 66,67 y cuyo INSE estuvo por debajo del percentil 33,33. Esta elección se sustenta en la metodología utilizada en otras investigaciones que, aunque no se desarrollaron en el contexto colombiano, sirven como referencia sobre cómo la literatura define a un estudiante resiliente, en gran parte de estos estudios, se selecciona el cuartil inferior (Q1) del nivel socioeconómico y un puntaje por encima del percentil 75 o cuartil superior (Q4) en la prueba analizada por otro lado, otros trabajos [56] [69] [70] [71] [72] han flexibilizado este criterio al percentil 66, como en el presente estudio, debido a que elegir el Q4 reducía la muestra casi a la mitad (**Fig. 6**), impactando el etiquetado de resilientes y no resilientes, por ende la base de datos final que se usará en el modelo de *machine learning*.

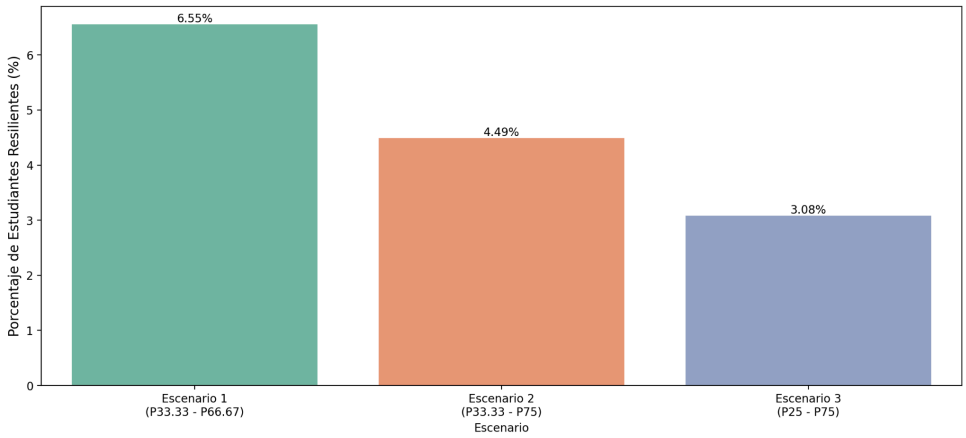


Fig. 6. Porcentaje de estudiantes resilientes bajo diferentes escenarios de referencia

El Puntaje Global también tiene una distribución cercana a la normal con media de 260 puntos, aunque con una ligera asimetría negativa y colas más largas en ambos extremos (**Fig. 7 (a)**) , esto se apoya con el boxplot (**Fig. 7 (b)**) evidencia una mayor cantidad de valores atípicos en la parte baja (puntajes muy bajos) y algunos en la parte alta. Esta variable tiene mediana de 257.00, desviación estándar de 51.16, coeficiente de variación del 19.67%, asimetría de 0.285 y curtosis de -0.166.

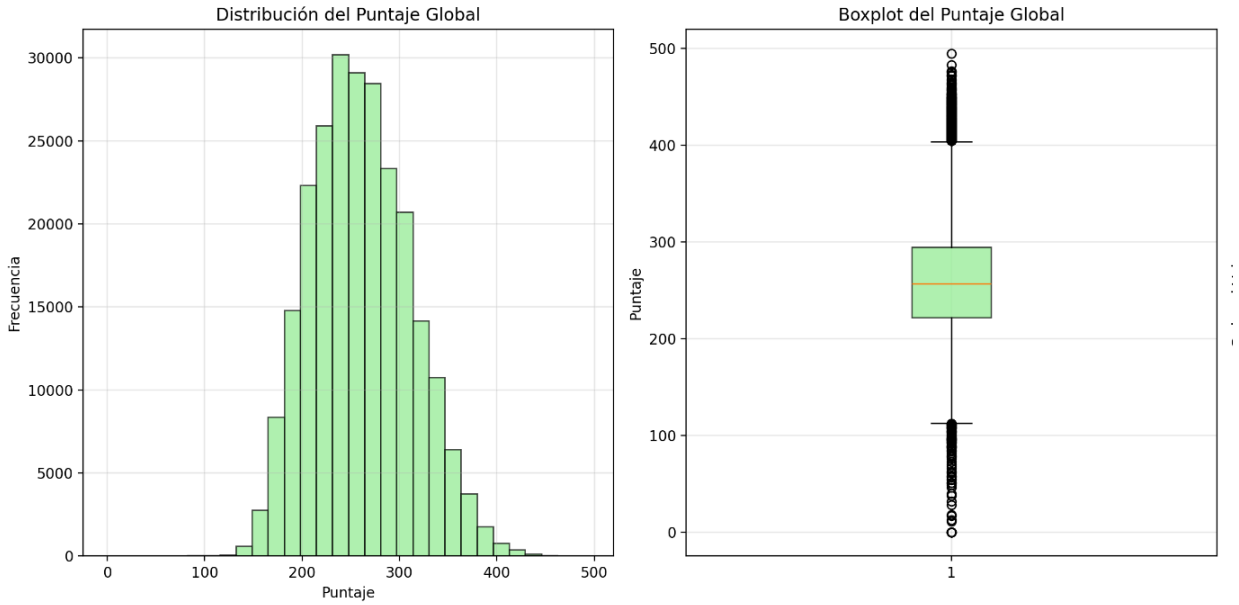


Fig. 7. Distribución y dispersión del puntaje global en la población analizada

(a) Histograma del puntaje global. (b) Boxplot del puntaje global

8.2 PANORAMA DE VIOLENCIA (HOMICIDIOS) EN CALI

Entre 2014 y 2024 se registraron en Cali un total de 13.053 homicidios, con un promedio anual de 1.186 casos, una mediana de 1.171 y un mínimo de 938 casos en el último año disponible. El valor más alto (**Fig. 8**) se presentó en 2014, con 1.561 homicidios en el marco de lo que se considera el coletazo final del conflicto con las extintas FARC, los residuos del Cartel del Norte del Valle y la conformación de nuevos órdenes criminales urbanos [71]. Aunque se observa una tendencia descendente en la serie, el año 2021 registró un rebote en la cifra, alcanzando 1.240 casos, coincidente con el Paro Nacional y las afectaciones sociales derivadas del confinamiento y la crisis post pandemia. Estos datos sugieren que, aunque los homicidios han disminuido a nivel agregado, su comportamiento sigue siendo sensible a coyunturas sociopolíticas y económicas.

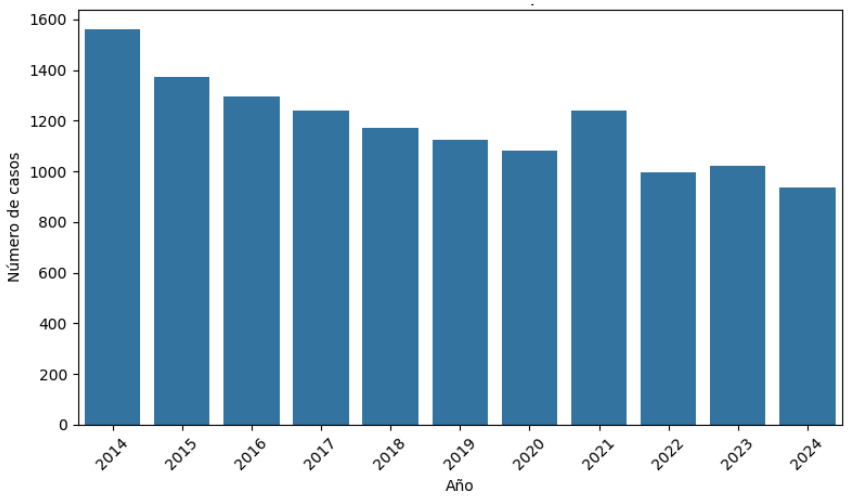


Fig. 8. Número de homicidios registrados por año en el periodo 2014–2024

La distribución semanal de los homicidios en valores absolutos (**Fig. 9**) muestra una mayor concentración los domingos, que representan aproximadamente el 16,6% del total de casos. Le siguen los sábados con un 14,3%, mientras que el resto de los días se mantienen relativamente estables entre el 12% y el 13% cada uno. Este patrón temporal indica una relación entre los picos de homicidios y momentos de esparcimiento comunitario o actividades sociales informales, usualmente sin presencia institucional o control del orden público. A nivel horario (**Fig. 10**), el análisis evidencia que los homicidios aumentan de forma sostenida entre las 17:00 y las 21:00 horas, alcanzando su punto máximo a las 20:00 (con más de 1.200 casos), lo cual refuerza la hipótesis de interacción entre consumo de alcohol, fiestas barriales y conflictos interpersonales o criminales.

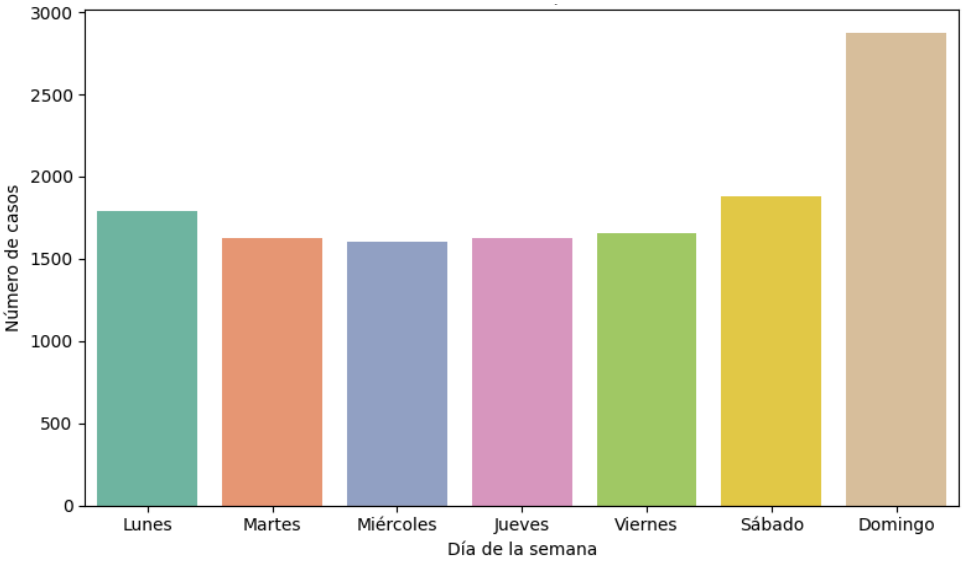


Fig. 9. Distribución de homicidios por día de la semana en el periodo 2014–2024

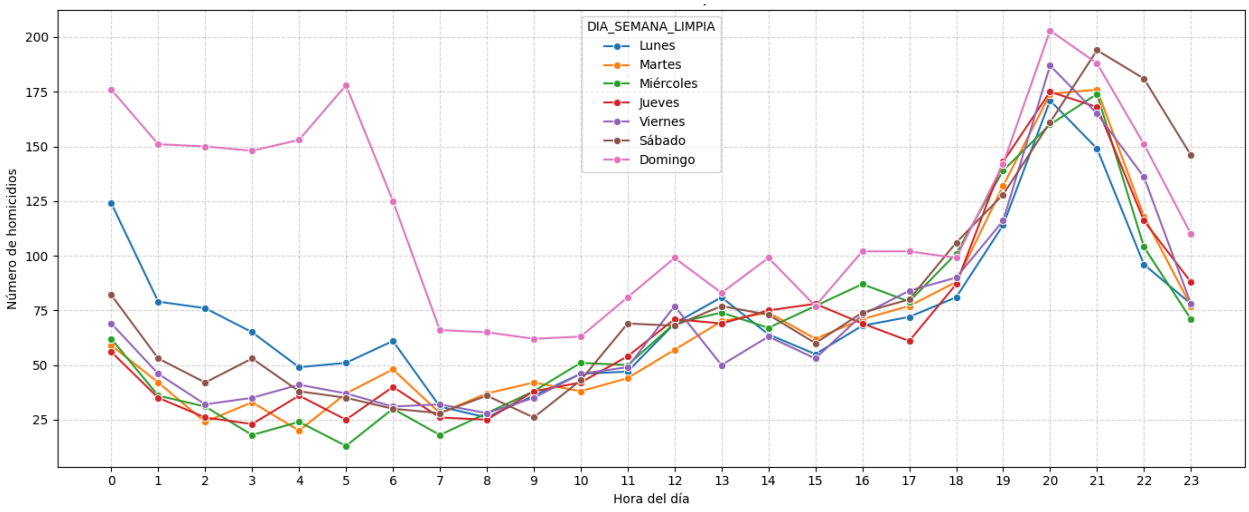


Fig. 10. Distribución horaria de homicidios por día de la semana en el periodo 2014–2024

En cuanto al tipo de arma empleada (**Fig. 11**), el 78% de los homicidios fueron perpetrados con armas de fuego, seguidas por armas cortopunzantes (13%) y otras armas (9%). Esta distribución es consistente en los distintos tipos de violencia. En particular, el 83% de los homicidios clasificados como “delincuencia” involucraron armas de fuego (6.877 de 8.449 casos), mientras que en situaciones de “convivencia” su uso también fue predominante, con un 63% de los casos (2.824 de 4.265). Esta evidencia respalda la hipótesis sobre la expansión del acceso a armamento en contextos urbanos, no restringido únicamente a economías criminales

organizadas como son las riñas, la violencia intrafamiliar o incluso la violencia alrededor de la educación (matoneo, acoso escolar y sexual, etcétera).

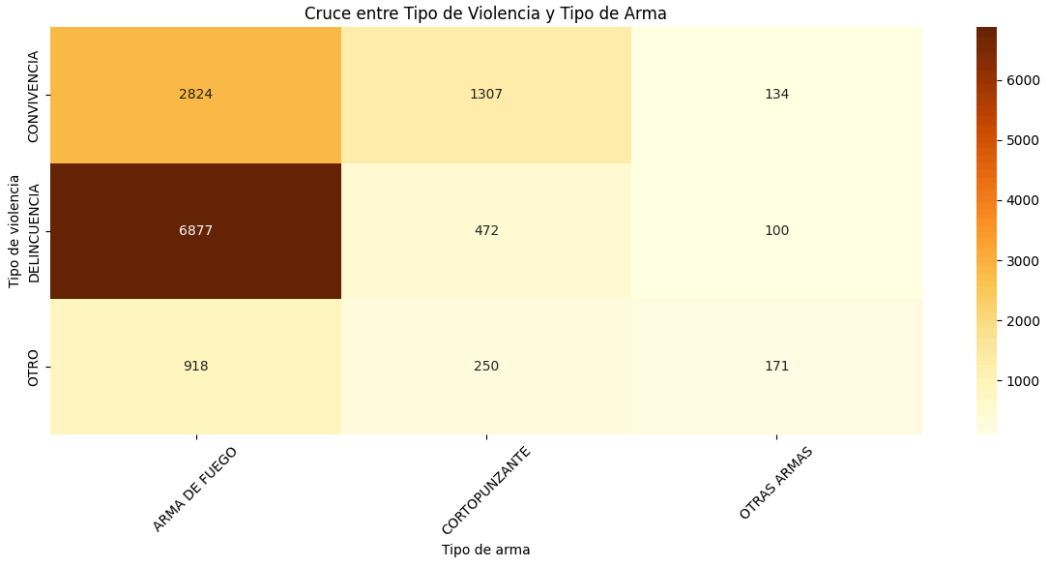


Fig. 11. Relación entre el tipo de violencia y el tipo de arma empleada

La violencia homicida en Cali presenta una concentración espacial significativa. Aunque el promedio por barrio en el periodo fue de 33 homicidios (σ = 46,6; mediana = 16), existen sectores con registros muy superiores. Potrero Grande reportó el mayor número total de homicidios con 308 casos (**TABLA IV**), seguido por Mojica (269), Siloé (274) y Alfonso Bonilla Aragón (212). En el top 10 de barrios más afectados, la media de homicidios por barrio osciló entre 25 y 32, con desviaciones estándar entre 9,8 y 12,6, lo que refleja una persistencia del fenómeno en áreas específicas.

Estos sectores coinciden con zonas históricamente caracterizadas por profundos desafíos estructurales e institucionales: informalidad urbana (de barrios espontáneos a barrios integrados), alejados de los centros productivos, con fuerte presencia de población desplazada o migrante interna y con poca o nula inversión social [73]. Así, el ciclo de violencia, como se ha discutido, es multicausal pero su dinámica surge en el marco de una combinación clara de factores.

TABLA IV

ESTADÍSTICAS DESCRIPTIVAS DEL TOP 10 DE BARRIOS CON MAYOR CONCENTRACIÓN DE HOMICIDIOS

Barrio	Homicidios totales	Promedio	Mediana	SD
Potrero Grande	308	27	23.5	11
Siloe	274	28.5	25	10.7
Mojica	269	28	26	10.7
Comuneros I	241	29	26	12
Manuela Beltrán	227	28	25	12
El Retiro	221	25.5	22	9.8
Alfonso Bonilla	212	30	27	12.6
El Vergel	202	26	24	9.7
Ciudad Córdoba	196	28.5	26	10.7
Sucre	191	31.5	30	12

Por último, el análisis por sexo confirma una sobrerrepresentación masculina sostenida a lo largo del periodo. Del total de homicidios registrados, el 93,4% corresponden a víctimas hombres (12.176 casos), con una media de 30,2 homicidios por barrio y una mediana de 27. En comparación, las mujeres representaron el 6,6% (864 casos), con una media de 33,5 y una mediana de 30 homicidios por barrio. Si bien el volumen de casos femeninos es considerablemente menor, su distribución presenta mayor dispersión, lo cual sugiere dinámicas diferenciadas de victimización por género que podrían no estar concentradas en zonas específicas o asociadas a estructuras delictivas estables.

En términos etarios (**Fig. 12 (a)**), la edad promedio de las víctimas de homicidio en Cali es de aproximadamente 31 años, con una distribución sesgada hacia la izquierda. El grupo más afectado se encuentra entre los 20 y 30 años, acumulando más del 45% de los casos. Este patrón se mantiene para ambos sexos, aunque con matices importantes, los hombres tienden a ser víctimas más jóvenes, con una mediana cercana a los 27 años, mientras que las mujeres presentan una mediana ligeramente superior, en torno a los 30 años. El análisis de cajas por sexo (**Fig. 12 (b)**) evidencia que los hombres tienen una distribución más concentrada y un rango intercuartílico menor, mientras que las mujeres presentan mayor variabilidad en las edades, con

más casos por encima de los 60 años. Este hallazgo podría estar relacionado con la naturaleza de los homicidios femeninos, muchos de los cuales podrían no estar vinculados a dinámicas territoriales delictivas, sino a conflictos intrafamiliares, violencia de género o situaciones contextuales distintas.

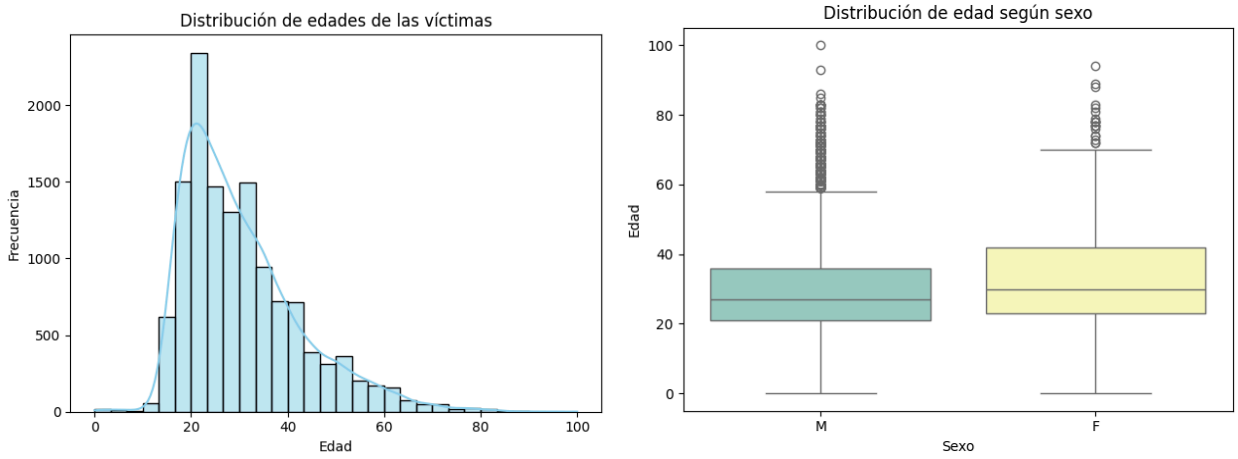


Fig. 12. Distribución de las edades de las víctimas

(a) Histograma de edades de las víctimas. (b) Distribución de edad según sexo

6. ANÁLISIS GEOREFERENCIADO DE LOS HOMICIDIOS Y LAS INSTITUCIONES EDUCATIVAS EN CALI

La georreferenciación de los homicidios en Cali tiene patrones de concentración espacial que justifican un análisis detallado de su proximidad a las instituciones educativas. Al aplicar buffers "polígonos de distancia fija usados para medir proximidad espacial" de 100 m, 250 m y 500 m alrededor de cada IE, se observa un incremento sustancial en la media y la mediana de homicidios capturados (**TABLA V**): desde un promedio de 3.39 homicidios a 100 m (mediana = 2), hasta 87.57 homicidios a 500 m (mediana = 73), con valores máximos que alcanzan los 479 casos. Este comportamiento no lineal sugiere que muchas IEs están inmersas en contextos de alta violencia, especialmente al ampliar el radio de observación. El buffer de 250 m ofrece un equilibrio analítico, capturando una mediana de 17 homicidios, lo que permite identificar focos críticos sin diluir la variabilidad con zonas más extensas.

TABLA V

ESTADÍSTICAS DESCRIPTIVAS DE LOS HOMICIDIOS EN CALI POR BUFFERS EN IE´S

Buffer	Promedio	Mediana	Max
100m	3.4	2	21
250m	22.3	17	113
500m	87.6	73	479

Esta decisión se soporta además a través del análisis de autocorrelación espacial mediante el índice de Moran global (**Fig. 13 (b)**), que arroja un valor de 0.687 con un p-valor prácticamente nulo, lo que confirma una concentración no aleatoria de los homicidios. En otras palabras, los homicidios tienden a agruparse espacialmente, dando lugar a clusters de violencia en la ciudad. Esta dependencia espacial valida el uso de modelos geoestadísticos y justifica la incorporación de variables de entorno inmediato (como la proximidad a IEs) en los modelos explicativos o predictivos de violencia urbana. No se trata solo de la presencia de homicidios en ciertas zonas, sino de la existencia de una estructura espacial sistemática en torno a ellos.

Los buffers (**Fig. 13 (a)**) visualizan de manera clara esta estructura, los sectores con mayores puntos corresponden a zonas del oriente y centro-sur de la ciudad, donde también se localiza una alta densidad de instituciones educativas. El mapa con clusters dibujados confirma cómo varios colegios oficiales y no oficiales están rodeados por un número considerable de homicidios, especialmente en comunas históricamente vulnerables. Esta superposición no es fortuita y refuerza la hipótesis de que los homicidios pueden incidir en la dinámica educativa, particularmente en la percepción de seguridad, la asistencia escolar y la continuidad académica de las y los estudiantes expuestos.

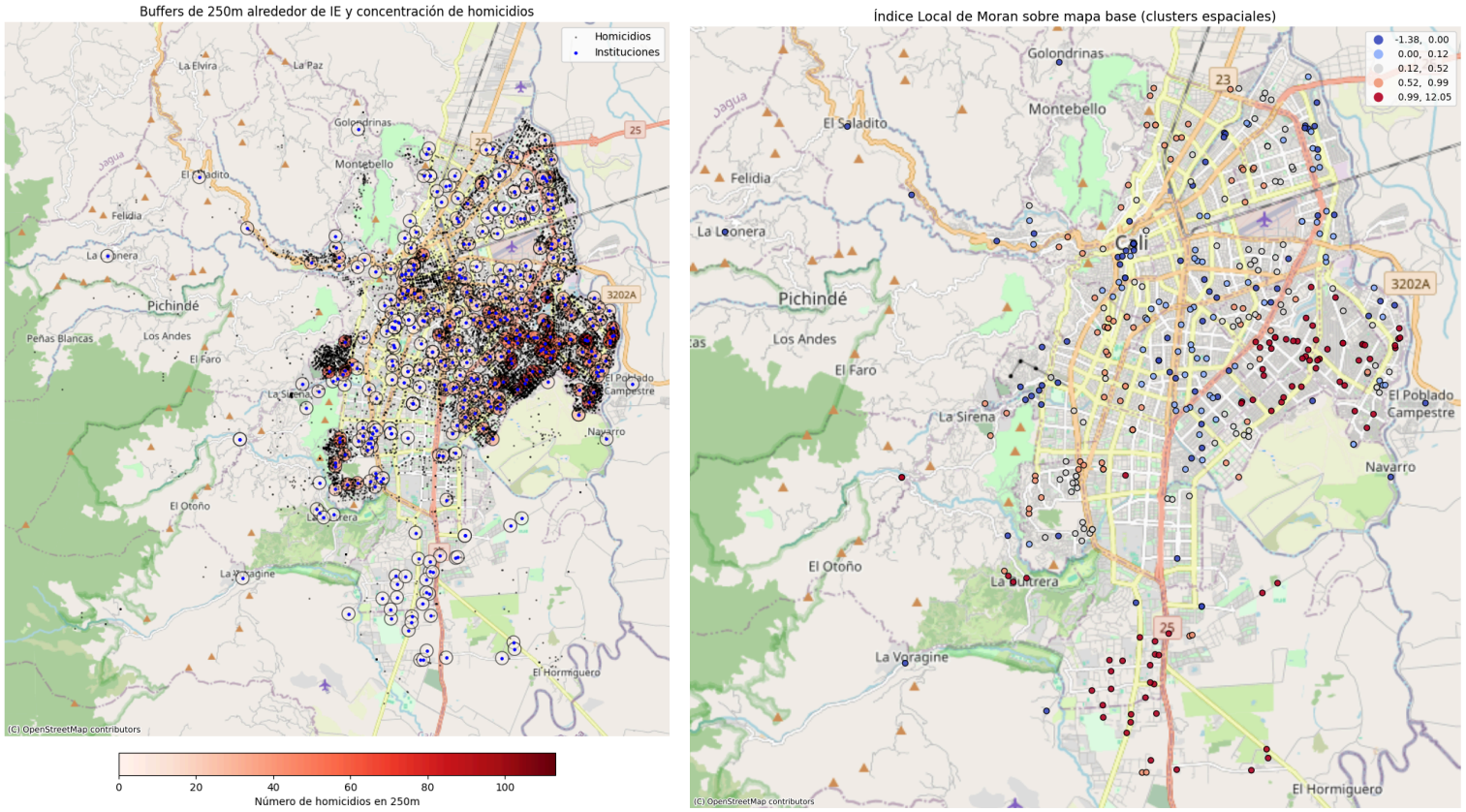


Fig. 13. Análisis espacial de homicidios en el área urbana de estudio

(a) Buffers de 250 m alrededor de instituciones educativas y concentración de homicidios. (b) Índice Local de Moran aplicado a homicidios (clústeres espaciales)

7. APLICACIÓN Y EVALUACIÓN DE RANDOM FOREST SHARP PARA LA ESTIMACIÓN DE LA RESILIENCIA ACADÉMICA

Con el objetivo de predecir la resiliencia académica de las instituciones educativas en relación a los recursos académicos y los homicidios, se desarrolló un modelo de clasificación supervisada que se entrenó sobre una base de datos estructurada, integrando características institucionales (C600 Dane), resultados de pruebas estandarizadas (ICFES - Pruebas Saber 11), y variables contextuales como el nivel de homicidios en un radio de 250 metros alrededor de cada institución. Para la predicción se utilizó el algoritmo XG Boost, ideal para contextos con relaciones no lineales, multicolinealidad y desbalance de clases.

El proceso de modelado inició con la totalidad de las variables descritas en la sección 8.1, a partir de este conjunto inicial, se aplicó un procedimiento de eliminación recursiva de características (RFE) basado en Random Forest y valores SHAP. Este proceso se ejecutó en 10 iteraciones sucesivas, eliminando en cada una las variables con menor contribución predictiva hasta conservar únicamente aquellas con mayor relevancia estadística y explicativa. De esta manera, el subconjunto final de predictores no fue seleccionado de forma arbitraria, sino como resultado de un proceso iterativo de depuración guiado por la importancia de variables.

Posteriormente, y siguiendo la literatura reciente [56], se completó el pipeline de modelado aplicando varias tareas adicionales, imputación de valores faltantes mediante la mediana, balanceo de la clase minoritaria con SMOTE y ajuste de hiper parámetros mediante búsqueda aleatoria con validación cruzada. El rendimiento del modelo se evaluó bajo la métrica F1-macro, y en cada iteración se registraron tanto las variables retenidas como los indicadores de desempeño. Con ello fue posible asegurar que las variables finales no solo fueran estadísticamente relevantes, sino también estables a través del proceso iterativo.

El modelo final fue evaluado con diferentes umbrales de clasificación, que determinan el punto de corte de probabilidad para predecir si una institución es resiliente o no (**TABLA VI**). Se observó que valores bajos del umbral (por ejemplo, 0.3) aumentan significativamente el recall (sensibilidad), identificando correctamente el 84% de las instituciones resilientes, aunque con una baja precisión (27%). Por el contrario, al subir el umbral a 0.5, se logra una mayor precisión (35.6%) pero disminuye la sensibilidad (63%). El área bajo la curva ROC (AUC = 0.75) evidencia que el modelo tiene una capacidad predictiva adecuada (**Fig. 14**), con buen balance entre sensibilidad y especificidad.

TABLA VI

RESULTADOS DE LA EVALUACIÓN SEGÚN DIFERENTES UMBRALES DE PROBABILIDAD

Umbral de probabilidad	Precisión (promedio ponderado)	Recall (promedio ponderado)	f1-score (promedio ponderado)
0.3	0.79	0.54	0.58
0.4	0.80	0.60	0.68
0.5	0.78	0.70	0.73

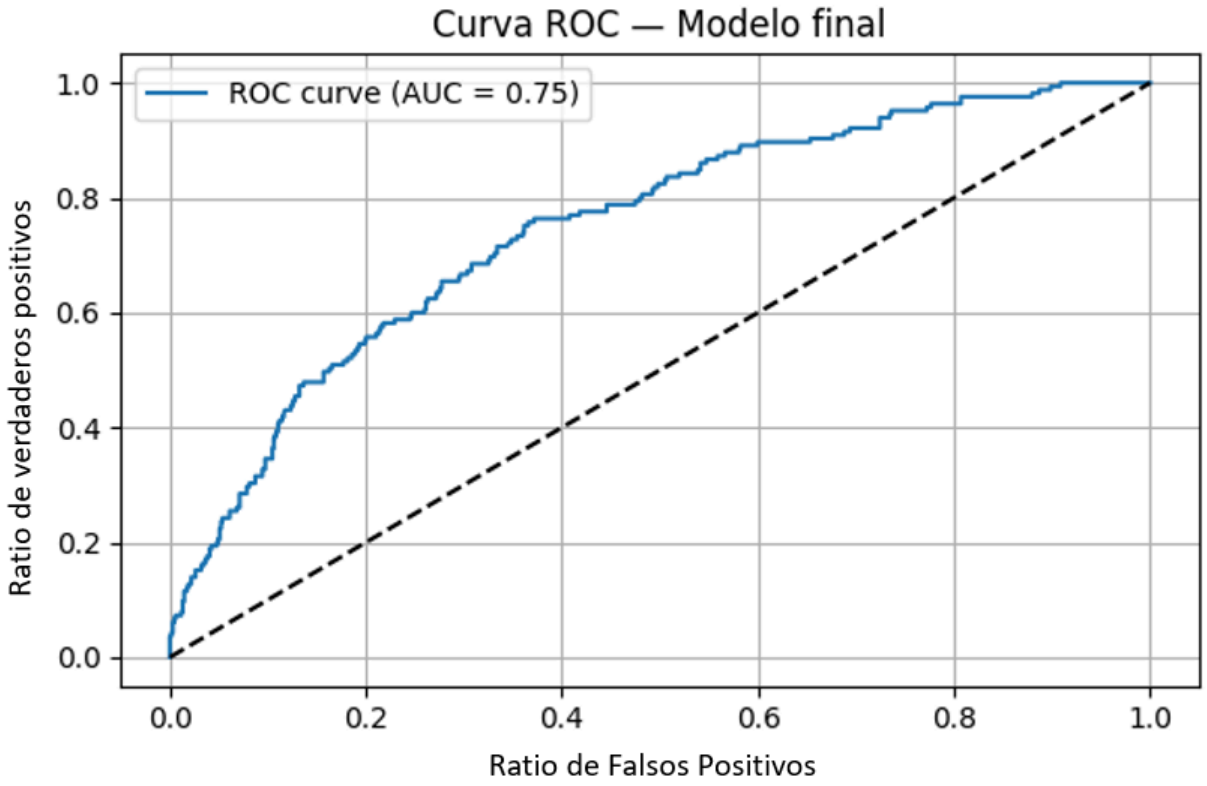


Fig. 14. Curva ROC del modelo final con área bajo la curva (AUC = 0.75)

El análisis de valores SHAP (**Fig. 15**), que explica el impacto de cada variable sobre las predicciones, mostró que las características institucionales como el calendario escolar, la clasificación oficial o no oficial, y el carácter de la institución tienen alta influencia en la probabilidad de ser resiliente. De forma destacada, la variable Homicidios_250m (calculada con base en el análisis de correlación espacial) también figura entre las más importantes, mostrando que niveles elevados de este tipo de violencia en el entorno inmediato se asocian

negativamente con la resiliencia académica. Esta evidencia apoya la hipótesis de que los factores contextuales del territorio inciden de manera crítica en las trayectorias escolares.

En conjunto, los resultados del modelo no solo permiten predecir con una precisión razonable qué instituciones agrupan la mayor cantidad de estudiantes resilientes, sino que también identifican patrones estructurales y contextuales que explican esta condición. Esta herramienta puede servir de base para el diseño de intervenciones focalizadas que busquen promover entornos protectores y mejorar la capacidad de respuesta de las instituciones frente a condiciones adversas. Teniendo en cuenta que este es un fenómeno de naturaleza social, las oportunidades de refinamiento del modelo y de aplicación en otros casos están

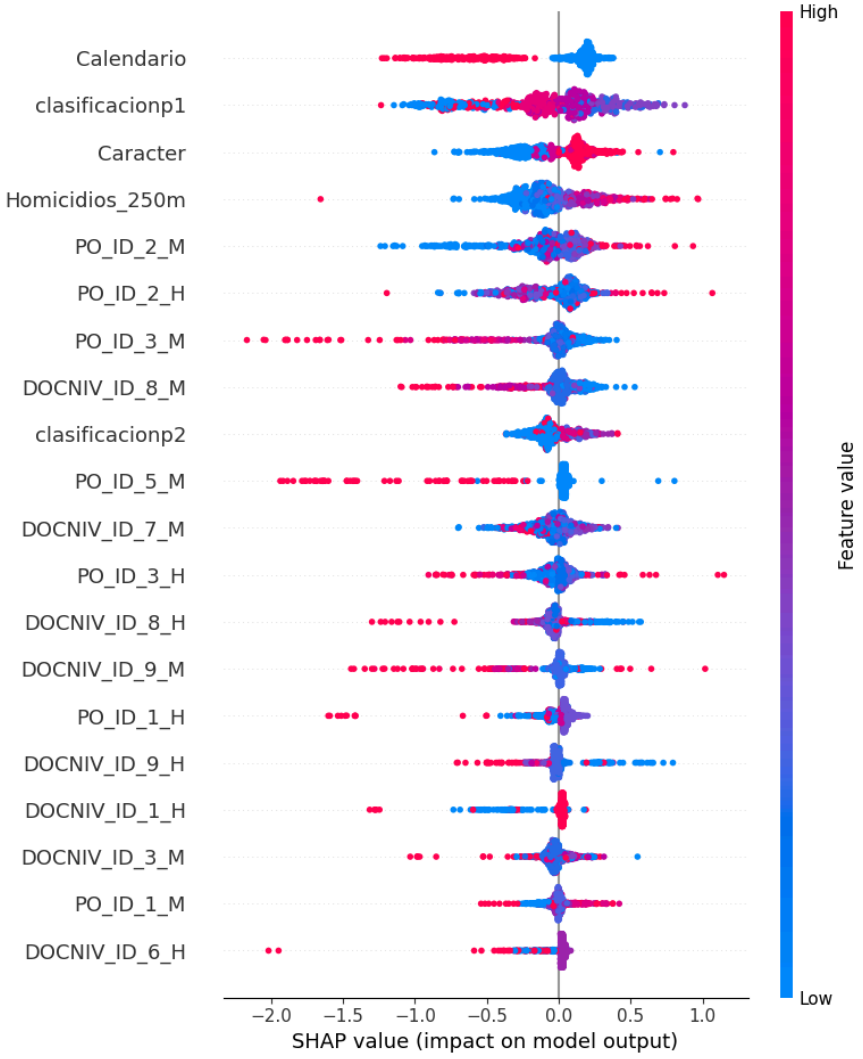


Fig. 15. Importancia y efecto de las variables en el modelo según valores SHAP

8. APLICACIÓN Y ANÁLISIS DEL MODELO DE EFICIENCIA

Posterior al cálculo de resiliencia académica, se procedió a analizar la eficiencia de las instituciones educativas mediante un modelo condicional $\text{order-}m$ FDH que integra tanto el desempeño académico como la capacidad de sostener resultados positivos en contextos adversos.

Como outputs se seleccionaron dos medidas complementarias de desempeño educativo: (i) el puntaje global agregado de las pruebas Saber 11 del ICFES, que captura el rendimiento académico promedio institucional, y (ii) el número de estudiantes resilientes agregado, que refleja la capacidad de la institución para generar resultados destacados en estudiantes provenientes de contextos socioeconómicos desfavorables.

Los inputs considerados corresponden a recursos institucionales disponibles, producto de un proceso de filtrado que elimina variables con más del 70% de valores faltantes. Las 10 variables input finales incluyen: matrícula total (Alum_TOTAL), dotación tecnológica clasificada en tres categorías (TICS_ID_1, TICS_ID_2, TICS_ID_3), estructura de personal ocupado en tres niveles (PO_ID_1_T, PO_ID_2_T, PO_ID_3_T), y nivel educativo del personal docente en tres categorías formativas (DOCNIV_ID_3_T, DOCNIV_ID_7_T, DOCNIV_ID_8_T).

Las variables contextuales incorporadas fueron: (i) densidad de homicidios en un radio de 250 metros alrededor de cada institución, operando como proxy de exposición a violencia urbana, y (ii) INSE promedio de instituciones vecinas en un radio de 500 metros, calculado a partir del promedio del INSE individual reportado por estudiantes en las pruebas Saber 11 del ICFES, que captura el nivel socioeconómico del entorno educativo inmediato.

Tras aplicar filtros de calidad (eliminación de instituciones con más del 60% de observaciones incompletas y casos con valores faltantes en variables críticas), el dataset final consistió en 839 observaciones correspondientes a 125 instituciones educativas a lo largo de 11 años (2014-2024).

Una vez aplicado el análisis de sensibilidad del parámetro m (tamaño de muestra para frontera parcial) con valores de 15, 30 y 50, se observó sensibilidad metodológica sustancial, por lo que se reportan resultados para los tres escenarios:

- Para m=15 (énfasis en comparaciones locales), la eficiencia promedio no condicional alcanzó 0.797, mientras que la eficiencia condicional fue 0.979, generando un ratio promedio de 1.270. Esto indica que, al ajustar por contexto socio-espacial, las instituciones muestran en promedio variaciones del 27% en su eficiencia relativa. El 89.3% de las instituciones presentaron ratios superiores a 1, evidenciando que el

contexto es predominantemente desfavorable para la eficiencia educativa. Aplicando el ajuste tipo 2 (2 - valor crudo) para interpretar como distancia a eficiencia óptima, la eficiencia no condicional ajustada fue 1.203 y la condicional ajustada 1.021.

- Para m=30, el ratio promedio disminuyó a 1.190 (eficiencia no condicional: 0.850; condicional: 0.993), mientras que para m=50 el ratio fue 1.161 (no condicional: 0.871; condicional: 0.997). Esta tendencia decreciente del ratio con m creciente es consistente con la literatura: valores mayores de m aproximan la frontera completa, diluyendo el efecto contextual local.
- Para m=50, el ratio promedio continuó su descenso hasta 1.161, con eficiencia no condicional de 0.871 y condicional de 0.997. Los valores ajustados alcanzaron 1.129 y 1.003 respectivamente, con ratio ajustado de 0.839. Esta convergencia hacia valores cercanos a 1.000 en eficiencia condicional sugiere que, con fronteras más amplias, la mayoría de instituciones se acercan a la frontera de referencia. La tendencia decreciente del ratio con m creciente es consistente con la literatura: valores mayores de m aproximan la frontera completa, diluyendo el efecto contextual local.

La selección del escenario m = 15 se considera óptima por su capacidad para capturar con mayor precisión las heterogeneidades locales y los efectos contextuales en la frontera de eficiencia. Este valor enfatiza comparaciones entre instituciones con condiciones estructurales y socio espaciales similares, evitando la sobre-generalización que ocurre cuando *m* crece y la frontera se aproxima a un comportamiento determinista. En este estudio, *m* = 15 maximizó la sensibilidad del modelo a las variables contextuales, reflejada en un R^2 de 0.106 y en efectos marginales estadísticamente significativos para homicidios (0.0172; $p < 0.05$) e INSE vecinal (0.0037; $p < 0.01$), los cuales se diluyen conforme *m* aumenta. Este escenario permitió, además, mantener la robustez del estimador bootstrap (B = 200) y un balance adecuado entre estabilidad y representatividad local, lo que justifica adoptar m = 15 como configuración base para reportar los resultados del modelo condicional de eficiencia educativa

La regresión no paramétrica local-linear del ratio de eficiencia sobre las variables contextuales reveló efectos marginales positivos y estadísticamente significativos. Para m=15, el R^2 fue 0.106, indicando que las variables contextuales explican aproximadamente 11% de la variación en eficiencia relativa. El efecto marginal de homicidios fue 0.0172 ($p<0.05$), lo que implica que mayor densidad de homicidios se asocia paradójicamente con mayor ratio de eficiencia. Esta relación aparentemente contraintuitiva sugiere que las instituciones que logran mantener buenos resultados en entornos violentos son desproporcionadamente eficientes dado su contexto adverso: el modelo condicional "premia" su capacidad de sobreponerse a condiciones desfavorables.

El efecto marginal de INSE vecinal fue 0.0037 ($p<0.01$), estadísticamente más robusto que el de homicidios. Un aumento de una unidad en el INSE promedio de vecinos se asocia con

incremento de 0.37% en el ratio de eficiencia, reflejando externalidades positivas del entorno socioeconómico sobre el desempeño institucional relativo. Ambos efectos disminuyen con m creciente: para m=50, el efecto de homicidios se reduce a 0.0046 y el de INSE a 0.0011, con R^2 de apenas 0.048. Esto confirma que los efectos contextuales son más pronunciados en evaluaciones locales (m pequeño) y se diluyen al aproximarse a la frontera completa.

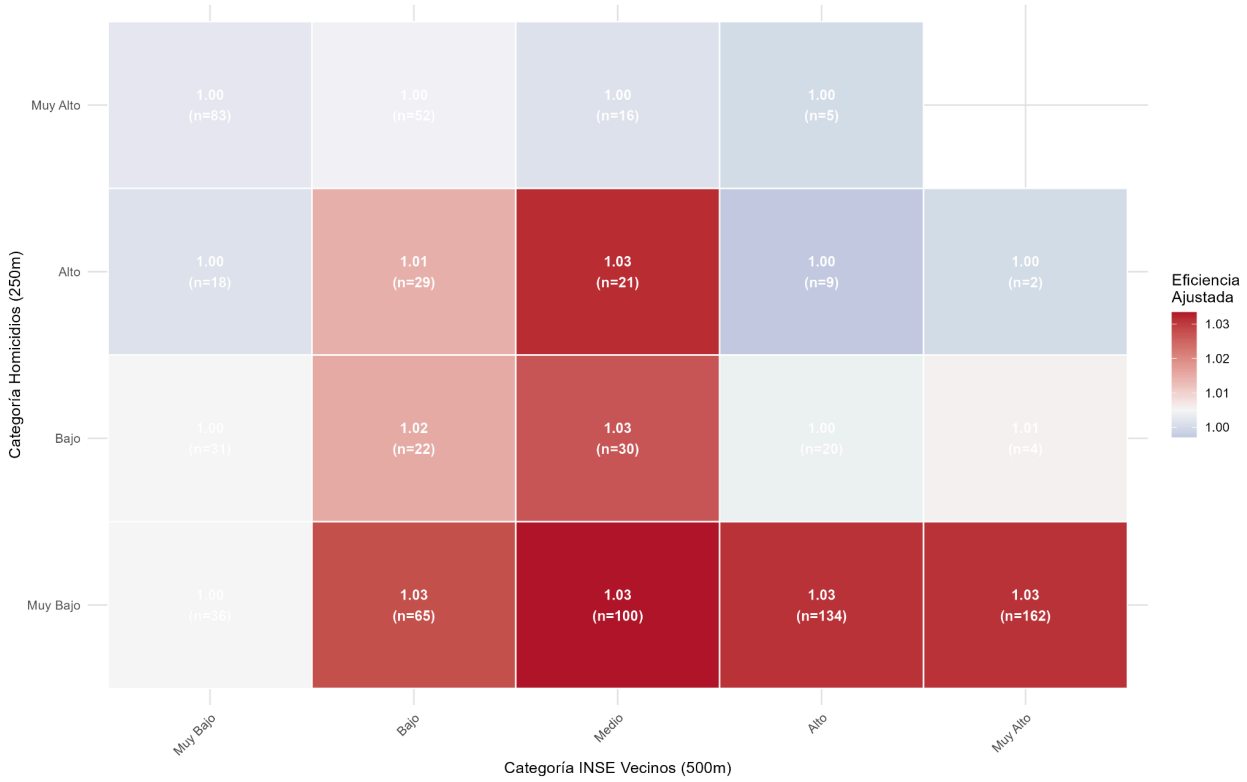


Fig. 16. Interacción entre homicidios e INSE vecinos en el contexto de eficiencia

Así, la eficiencia educativa no depende únicamente de recursos y resultados internos, sino del contexto socio-espacial en el que operan las instituciones. A pesar de ser un resultado contraintuitivo, los homicidios presentaron efecto positivo sobre el ratio de eficiencia, lo que se interpreta como ajuste favorable para instituciones resilientes en entornos adversos. El INSE vecinal mostró el efecto más robusto, confirmando la importancia de externalidades socioeconómicas del entorno inmediato. La sensibilidad al parámetro m subraya la necesidad de reportar múltiples especificaciones en análisis de eficiencia con fronteras parciales.

9. CONCLUSIONES

Este proyecto integró datos educativos (Saber 11 y C600) y criminales (homicidios) de los últimos diez años en Cali, con el fin de analizar cómo los homicidios inciden en la relación entre eficiencia educativa y resiliencia académica. Para ello, se emplearon modelos no paramétricos y técnicas de aprendizaje automático con enfoque espacial, lo que permitió estimar patrones de resiliencia a nivel institucional, identificar determinantes de eficiencia ajustados por contexto socioespacial y cuantificar la influencia de los homicidios sobre ambas dimensiones.

Los modelos implementados permitieron evidenciar con claridad cómo la violencia afecta el desempeño educativo a través de mecanismos diferenciados según la capacidad institucional. El modelo Random Forest con interpretabilidad SHAP, aplicado sobre más de 200.000 observaciones, alcanzó un AUC de 0.75 y confirmó que la resiliencia académica puede predecirse con alta consistencia a partir de variables estructurales y familiares. Entre las variables con mayor importancia se destacan el puntaje global de las pruebas Saber 11 del ICFES y la presencia de personal psicosocial, seguidas por la infraestructura tecnológica y el número de docentes con formación de posgrado. Los valores SHAP indicaron efectos positivos marginales significativos en dichas variables, lo que sugiere que la resiliencia no es un fenómeno aleatorio, sino un resultado acumulativo de recursos y entornos institucionales protectores.

Por su parte, el modelo condicional de eficiencia $\text{order-}m\ \text{FDH}$, replicado 200 veces con técnicas *bootstrap* para garantizar estabilidad, mostró que el número de homicidios tiene un impacto negativo directo sobre la frontera de eficiencia educativa. En las instituciones localizadas en zonas con más de 50 homicidios anuales dentro de un radio de 250 m, el puntaje medio de eficiencia condicional fue 13 % inferior al promedio urbano, con un efecto marginal de -0.12 en la regresión local-lineal de factores contextuales. Los resultados evidencian que, incluso tras controlar por inputs educativos (docentes, recursos TIC, tamaño institucional), la variable de homicidios mantiene una influencia significativa, lo que confirma que la violencia opera como una restricción exógena a la productividad educativa.

En conjunto, ambos enfoques tienen convergencia conceptual, la resiliencia y la eficiencia educativa responden a una misma arquitectura institucional y social. Los colegios con mejor gestión de recursos humanos y tecnológicos logran amortiguar parcialmente los efectos de la violencia, desplazando su frontera de eficiencia hacia niveles comparables a contextos de baja criminalidad. Así, la integración de modelos predictivos e inferenciales no solo permitió identificar patrones de vulnerabilidad educativa, sino también cuantificar dicha incidencia, los homicidios reducen en promedio la eficiencia condicional en -0.12 y se observa una correlación positiva moderada (r = 0.48) entre eficiencia y resiliencia, lo que confirma que ambas dimensiones responden de forma conjunta al contexto violento.

A continuación se presenta un resumen comparativo de las hipótesis planteadas inicialmente en relación con los resultados obtenidos:

TABLA VII
SÍNTESIS DE HALLAZGOS

Hipótesis	**Resultado empírico**
(H1) La resiliencia académica varía estrechamente con la eficiencia educativa.	Confirmada. Correlación positiva moderada (r = 0.48), consistente en todos los años.
(H2) Existen patrones espaciales definidos con clústeres de barrios contiguos con similares niveles de eficiencia y resiliencia.	Confirmada parcialmente. Identificados clústeres significativos (Moran's I = 0.31, p < 0.01) en el oriente y ladera de la ciudad.
(H3) Los homicidios afectan diferencialmente la relación resiliencia–eficiencia según el contexto institucional.	Confirmada. El modelo condicional de eficiencia muestra efectos negativos más altos en instituciones oficiales y zonas con mayor densidad de homicidios.
(H4) La eficiencia educativa se mantiene estable frente a la violencia en instituciones con mayores recursos tecnológicos y directivos con posgrado.	Confirmada parcialmente. Los recursos y la formación del personal reducen la sensibilidad al contexto violento, pero no la eliminan.

Finalmente, y aunque el estudio proporciona evidencia sólida sobre la relación entre homicidios, eficiencia y resiliencia, existen limitaciones relevantes. En primer lugar, la disponibilidad y consistencia de los datos del C600 presentan vacíos interanuales y variaciones en las variables disponibles, lo que afecta la continuidad del panel. En segundo lugar, los homicidios se utilizaron como variable proxy de violencia, sin incorporar otras formas como hurtos, desplazamientos o violencia intrafamiliar que también podrían incidir en los procesos educativos. A su vez, la investigación se centra en una sola ciudad, por lo que su generalización a otros contextos requiere precaución. Futuras investigaciones podrían ampliar el análisis a nivel regional y evaluar el efecto conjunto de múltiples expresiones de violencia sobre la eficiencia del sistema educativo colombiano.

Estas restricciones metodológicas y de cobertura generan varias líneas de investigación futura, entre ellas, la ampliación del análisis a nivel intraurbano incluyendo variables como la distancia hogar-escuela o el tipo de transporte estudiantil; la incorporación de otras expresiones de violencia (hurtos, extorsión, violencia intrafamiliar); y la replicación del modelo en otros municipios o regiones del país para evaluar la estabilidad geográfica de los resultados. También se plantea la posibilidad de integrar diseños cuasiexperimentales (como diferencias en diferencias) o modelos longitudinales que permitan aproximaciones más sólidas a relaciones causales.

En términos de implicaciones para la política pública, los resultados obtenidos evidencian que la eficiencia escolar no puede entenderse al margen de las condiciones territoriales en las que operan las instituciones educativas. La reducción promedio de -0.12 en la eficiencia condicional asociada a los homicidios y la correlación positiva moderada entre resiliencia y eficiencia ($r = 0.48$) muestran que la violencia urbana no solo afecta el desempeño individual de los estudiantes, sino que reduce la eficiencia institucional en el uso de los recursos escolares, ampliando la distancia respecto a la frontera eficiente. Esto sugiere la necesidad de avanzar hacia políticas educativas diferenciadas por territorio, donde la asignación de recursos humanos, tecnológicos y psicosociales priorice las zonas con mayor densidad delictiva o trabajen en reducir la violencia. Así mismo, la evidencia indica que la resiliencia académica no debe abordarse únicamente como una característica individual, sino como un resultado institucional dependiente de la gestión escolar y del entorno comunitario. En conjunto, los hallazgos plantean la importancia de articular políticas educativas y de seguridad urbana, de modo que la escuela deje de ser tratada como un sistema aislado y se reconozca como un actor inmerso en ecologías criminales específicas.

10. REFERENCIAS BIBLIOGRÁFICAS

[1] S. Perkins y S. Graham-Bermann, “Violence Exposure and the Development of School-Related Functioning: Mental Health, Neurocognition, and Learning”, *Aggress Violent Behav*, vol. 17, núm. 1, pp. 89–98, 2012, doi: 10.1016/j.avb.2011.10.001.

[2] S. López-Estrada, T. Agasisti, V. Giménez, y D. Prior, “Students’ resilience and school efficiency in one of the most unequal countries in the world: empirical evidence from Colombia”, *Education Economics*, vol. 0, núm. 0, pp. 1–18, doi: 10.1080/09645292.2023.2298845.

[3] J. C. H. Perez, “Evaluación de la calidad en la educación básica y media en Colombia”, *Cultura Educación Sociedad*, vol. 11, núm. 2, Art. núm. 2, jul. 2020, doi: 10.17981/cultedusoc.11.2.2020.08.

[4] J. Ariza y J. P. Saldarriaga, “Armed conflict and academic performance. A spatial approach for Colombia”, *Crime Law Soc Change*, vol. 80, núm. 5, pp. 503–523, dic. 2023, doi: 10.1007/s10611-023-10098-7.

[5] H. Bayona Rodríguez y D. C. López Vera, “Factores asociados a la resiliencia académica: evidencia para Colombia”, 2021, doi: 10.57784/1992/48061.

[6] B. Mark, “Assessing Policy Outcomes: Social and Political Biases”, en *Understanding Policy Fiascoes*, Routledge, 1996.

[7] R. Escobedo, “Violencia Homicida en Cali: focos y organizaciones criminales.”, mar. 2013.

[8] POLIS, *Cali Da Miedo [Datos en Breve No. 47]*, 1a ed. Cali: Universidad ICESI, 2022. [En línea]. Disponible en: https://www.icesi.edu.co/polis/images/2023/publicaciones/db/pdf/polis-db47-seguridad.pdf

[9] U. N. D. of E. and S. Affairs, *Informe de los Objetivos de Desarrollo Sostenible 2023: Edición especial*. United Nations, 2023. doi: 10.18356/9789210024938.

[10] M. Dickson y F. Buscha, “Returns to education - individuals”, en *Handbook of Labor, Human Resources and Population Economics*, K. Zimmermann, Ed., Springer Nature, 2023.

[11] S. Baidybekova, S. Sauranbay, y D. Yermekbayeva, “Investment in Education as a Factor of Economic Growth of the Country”, *Eurasian Journal of Economic and Business Studies*, vol. 4, pp. 105–114, dic. 2022, doi: 10.47703/ejebs.v4i66.194.

[12] A. Rothomi y M. Rafid, “RELATIONSHIP ANALYSIS AND CONCEPT OF HUMAN CAPITAL THEORY AND EDUCATION”, *EDUCATUM: Scientific Journal of Education*, vol. 1, pp. 26–31, feb. 2023, doi: 10.59165/educatum.v1i1.14.

[13] “Education Investment and Human Capital Development in India”, *ResearchGate*, oct. 2024, doi: 10.47604/jpid.2587.

[14] E. A. Hanushek y L. Wößmann, “Education and economic growth”, *Chapters in Economics*, Consultado: el 14 de noviembre de 2024. [En línea]. Disponible en: https://ideas.repec.org//h/lmu/muench/20460.html

[15] M. Padilla-Romo y C. Peluffo, “Persistence of the Spillover Effects of Violence and Educational Trajectories”, el 11 de agosto de 2023, *Social Science Research Network, Rochester, NY*: 4538038. doi: 10.2139/ssrn.4538038.

[16] F. Barrera Osorio y A. M. Ibáñez Londoño, “Does violence reduce investment in education?: a theoretical and empirical approach”, 2004, doi: 10.57784/1992/7882.

[17] A. Arbona, V. Giménez, S. López-Estrada, y D. Prior, “The relationship between homicides from armed conflict and efficiency of educational quality in Colombia”, *International Journal of Educational Development*, vol. 110, p. 103120, oct. 2024, doi: 10.1016/j.ijedudev.2024.103120.

[18] J. M. Muñoz Galeano, “Effects of high-crime environments on educational efficiency, a spacial case”, 2021, Consultado: el 23 de noviembre de 2024. [En línea]. Disponible en: https://vitela.javerianacali.edu.co/handle/11522/2709

[19] L. Jamison, K. Howell, K. Campbell, K. Thomsen, y A. Hasselle, “Exploring Multisystemic Resilience among Youth of Color Exposed to Direct and Indirect Violence”, *International Journal of Child and Adolescent Resilience*, vol. 10, núm. 1, Art. núm. 1, may 2023, doi: 10.54488/ijcar.2023.321.

[20] J. S. Coleman, “EQUALITY OF EDUCATIONAL OPPORTUNITY”, 1966. Consultado: el 14 de noviembre de 2024. [En línea]. Disponible en: https://eric.ed.gov/?id=ED012275

[21] A. Arbona, V. Giménez, S. López-Estrada, y D. Prior, “Efficiency and quality in Colombian education: An application of the metafrontier Malmquist-Luenberger productivity index”, *Socio-Economic Planning Sciences*, vol. 79, p. 101122, feb. 2022, doi: 10.1016/j.seps.2021.101122.

[22] K. De Witte y L. López Torres, “Efficiency in education: a review of literature and a way forward”, *Documents de Treball (Universitat Autònoma de Barcelona. Departament d’Economia de l’Empresa)*, núm. 1, p. 1, 2015.

[23] S. Perelman y D. Santin, “Measuring educational efficiency at student level with parametric stochastic distance functions: an application to Spanish PISA results”, *Education Economics*, vol. 19, núm. 1, pp. 29–49, feb. 2011, doi: 10.1080/09645290802470475.

[24] D. Stumbrienė, R. Želvys, J. Žilinskas, R. Dukynaitė, y A. Jakaitienė, “Efficiency and effectiveness analysis based on educational inclusion and fairness of European countries”, *Socio-Economic Planning Sciences*, vol. 82, p. 101293, ago. 2022, doi: 10.1016/j.seps.2022.101293.

[25] E. A. Hanushek y J. A. Luque, “Efficiency and equity in schools around the world”, *Economics of Education Review*, vol. 22, núm. 5, pp. 481–502, oct. 2003, doi: 10.1016/S0272-7757(03)00038-4.

[26] M. B. Shean, *Current Theories Relating to Resilience and Young People: A Literature Review*. VicHealth, 2015.

[27] M. O. Almulla, “Academic Resilience and its Relationships With Academic Achievement Among Students of King Faisal University in Saudi Arabia”, *Revista de Gestão Social e Ambiental*, vol. 18, núm. 9, pp. e07391–e07391, jun. 2024, doi: 10.24857/rgsa.v18n9-134.

[28] A. M. Tri y M. N. M. Rahayu, “Staying Optimistic in the Middle of Academic Challenges: A Correlational Study of Optimism with Academic Resilience in Bidikmisi/KIP Students”, *Psikoborneo*, vol. 12, núm. 1, p. 35, abr. 2024, doi: 10.30872/psikoborneo.v12i1.12863.

[29] N. R. Ahern, E. M. Kiehl, M. L. Sole, y J. Byers, “A Review of Instruments Measuring Resilience”, *Issues in Comprehensive Pediatric Nursing*, ene. 2006, doi: 10.1080/01460860600677643.

[30] L. Campbell-Sills y M. B. Stein, “Psychometric analysis and refinement of the connor–davidson resilience scale (CD-RISC): Validation of a 10-item measure of resilience”, *Journal of Traumatic Stress*, vol. 20, núm. 6, pp. 1019–1028, 2007, doi: 10.1002/jts.20271.

[31] H. S. Herbert y M. Manjula, “Stress-Coping and Factors Contributing to Resilience in College Students: An Exploratory Study from India”, *Indian Journal of Clinical Psychology*, 2017, Consultado: el 20 de noviembre de 2024. [En línea]. Disponible en: https://www.semanticscholar.org/paper/Stress-Coping-and-Factors-Contributing-to-in-An-Herbert-Manjula/d2090f48c27d77005b29d75a87a83d74891fdb60

[32] T. Agasisti, M. Soncin, y R. Valenti, “School factors helping disadvantaged students to succeed: empirical evidence from four Italian cities”, *Policy Studies*, vol. 37, núm. 2, pp. 147–177, mar. 2016, doi: 10.1080/01442872.2015.1127341.

[33] T. Agasisti y S. Longobardi, “Equality of Educational Opportunities, Schools’ Characteristics and Resilient Students: An Empirical Study of EU-15 Countries Using OECD-PISA 2009 Data”, *Soc Indic Res*, vol. 134, núm. 3, pp. 917–953, dic. 2017, doi: 10.1007/s11205-016-1464-5.

[34] T. Agasisti, F. Avvisati, F. Borgonovi, y S. Longobardi, “Academic resilience. What schools and countries do to help disadvantaged students succeed in PISA.”, *OECD Publishing*, vol. 167, 2018, doi: https://doi.org/10.1787/e22490ac-en.

[35] T. Agasisti, F. Avvisati, F. Borgonovi, y S. Longobardi, “What School Factors are Associated with the Success of Socio-Economically Disadvantaged Students? An Empirical Investigation Using PISA Data”, *Soc Indic Res*, vol. 157, núm. 2, pp. 749–781, sep. 2021, doi: 10.1007/s11205-021-02668-w.

[36] A. Sandoval-Hernandez y P. Białowolski, “Factors and conditions promoting academic resilience: a TIMSS-based analysis of five Asian education systems”, *Asia Pacific Education Review*, vol. 17, jul. 2016, doi: 10.1007/s12564-016-9447-4.

[37] H. Z. Abdillah y Marleni, “Cultivating Resilience: A Key to Managing Academic Stress among Health Students in Online Learning”, *Psyche 165 Journal*, pp. 304–309, dic. 2023, doi: 10.35134/jpsy165.v16i4.294.

[38] J.-P. Derriennic, “Theory and Ideologies of Violence”, *Journal of Peace Research*, vol. 9, núm. 4, pp. 361–374, dic. 1972, doi: 10.1177/002234337200900406.

[39] S. Kalyvas, “La violencia en medio de la guerra civil: esbozo de una teoría”, *Análisis Político*, núm. 42, Art. núm. 42, ene. 2001.

[40] V. Sanfelice, “Are safe routes effective? Assessing the effects of Chicago’s Safe Passage program on local crimes”, *Journal of Economic Behavior & Organization*, vol. 164, pp. 357–373, ago. 2019, doi: 10.1016/j.jebo.2019.06.013.

[41] M. Foureaux Koppensteiner y L. Menezes, “Violence and Human Capital Investments”, *Journal of Labor Economics*, vol. 39, núm. 3, pp. 787–823, jul. 2021, doi: 10.1086/711001.
[42] E. Chang y M. Padilla-Romo, “The Effects of Local Violent Crime on High-Stakes Tests”, *Working Papers*, Art. núm. 2019–03, jul. 2019, Consultado: el 14 de noviembre de 2024. [En línea]. Disponible en: https://ideas.repec.org//p/ten/wpaper/2019-03.html
[43] G. Barboza, “A Secondary Spatial Analysis of Gun Violence near Boston Schools: a Public Health Approach”, *J Urban Health*, vol. 95, núm. 3, pp. 344–360, jun. 2018, doi: 10.1007/s11524-018-0244-8.
[44] D. Jiménez y M. Castillo, “¿Cuántas fronteras toca superar? El caso de adolescentes y jóvenes de Cali, Colombia, en condiciones de vulnerabilidad social”, *Papeles de Población*, vol. 29, p. 169, mar. 2024, doi: 10.22185/24487147.2023.116.17.
[45] E. S. Davanzo y M. Justus, “Violence, Spatial Effects, and Education: Exploring the Relationship Between Exposure to Neighborhood Violence and Student Performance”, *Urban Rev*, sep. 2024, doi: 10.1007/s11256-024-00712-w.
[46] V. Duque, “Violence and Children’s Education: Evidence from Administrative Data”, *Working Papers*, Art. núm. 2019–16, oct. 2019, Consultado: el 14 de noviembre de 2024. [En línea]. Disponible en: https://ideas.repec.org//p/syd/wpaper/2019-16.html
[47] H. Jürges, L. Stella, S. Hallaq, y A. Schwarz, “Cohort at risk: long-term consequences of conflict for child school achievement”, *J Popul Econ*, vol. 35, núm. 1, pp. 1–43, ene. 2022, doi: 10.1007/s00148-020-00790-6.
[48] N. T. N. Ferguson y M. M. Michaelsen, “Money changes everything? Education and regional deprivation revisited”, *Economics of Education Review*, vol. 48, núm. C, pp. 129–147, 2015.
[49] T. Brück, M. Di Maio, y S. H. Miaari, “Learning The Hard Way: The Effect of Violent Conflict on Student Academic Achievement”, *Journal of the European Economic Association*, vol. 17, núm. 5, pp. 1502–1537, 2019.
[50] S. Salvo Garrido, H. Miranda Vargas, O. G. Vivallo Urra, J. L. Gálvez Nieto, y E. Miranda Zapata, “Estudiantes resilientes en el área de matemática: Examinando los factores protectores y de riesgo en un país emergente”, *Revista iberoamericana de diagnóstico y evaluación psicológica*, vol. 2, núm. 55, pp. 43–57, 2020.
[51] A. C. M. Guido Sarah, *Introduction to Machine Learning with Python*. Consultado: el 21 de diciembre de 2025. [En línea]. Disponible en: https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/
[52] I. Lundberg, J. E. Brand, y N. Jeon, “Researcher reasoning meets computational capacity: Machine learning for social science”, *Social Science Research*, vol. 108, p. 102807, nov. 2022, doi: 10.1016/j.ssresearch.2022.102807.
[53] J. D. Kelleher, B. M. Namee, y A. D’Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, Massachusetts, 2015.
[54] M. Kuhn y K. Johnson, *Applied Predictive Modeling*.
[55] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, y J. Foulds, *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2025.

[56] K.-C. Cheung, P.-S. Sit, J.-Q. Zheng, C.-C. Lam, S.-K. Mak, y M.-K. Ieong, "A machine-learning model of academic resilience in the times of the COVID-19 pandemic: Evidence drawn from 79 countries/economies in the PISA 2022 mathematics study", *Br J Educ Psychol*, vol. 94, núm. 4, pp. 1224–1244, dic. 2024, doi: 10.1111/bjep.12715.

[57] G. K. Mahalakshmi, "Machine Learning Approaches for Analyzing Stress Adaptation and Resilience Among Teachers in Higher Education", *Journal of Informatics Education and Research*, vol. 5, núm. 2, Art. núm. 2, abr. 2025, doi: 10.52783/jier.v5i2.2550.

[58] "First-year students' psychological resilience and college adjustment: A person-oriented approach - Gao - 2024 - Psychology in the Schools - Wiley Online Library". Consultado: el 4 de agosto de 2025. [En línea]. Disponible en: https://onlinelibrary.wiley.com/doi/10.1002/pits.23253

[59] A. Munoz-Galeano, S. López-Estrada, y A. Arbona, "High-crime environments and educational efficiency: A spatial case study", *International Journal of Educational Research*, vol. 129, p. 102509, ene. 2025, doi: 10.1016/j.ijer.2024.102509.

[60] C. Cazals, J.-P. Florens, y L. Simar, "Nonparametric frontier estimation: a robust approach", *Journal of Econometrics*, vol. 106, núm. 1, pp. 1–25, ene. 2002, doi: 10.1016/S0304-4076(01)00080-X.

[61] C. Daraio y L. Simar, "Conditional nonparametric frontier models for convex and nonconvex technologies: a unifying approach", *J Prod Anal*, vol. 28, núm. 1, pp. 13–32, oct. 2007, doi: 10.1007/s11123-007-0049-3.

[62] C. Daraio y L. Simar, "Introducing Environmental Variables in Nonparametric Frontier Models: a Probabilistic Approach", *J Prod Anal*, vol. 24, núm. 1, pp. 93–121, sep. 2005, doi: 10.1007/s11123-005-3042-8.

[63] M. Aliana, D. Prior, y E. Tortosa-Ausina, "Environmental factors in cross-country productivity growth: A conditional Malmquist Index", *Working Papers*, Art. núm. 2025/01, 2025, Consultado: el 3 de noviembre de 2025. [En línea]. Disponible en: https://ideas.repec.org//p/jau/wpaper/2025-01.html

[64] "Decreto 869 de 2010 - Gestor Normativo". Consultado: el 21 de diciembre de 2025. [En línea]. Disponible en: https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=39636

[65] DANE, "Educación formal". Consultado: el 4 de agosto de 2025. [En línea]. Disponible en: https://www.dane.gov.co/index.php/estadisticas-por-tema/educacion/poblacion-escolarizada/educacion-formal

[66] Alcaldia de Bogotá, "El ICFES cambia la clasificación de resultados de planteles | Bogota.gov.co". Consultado: el 4 de agosto de 2025. [En línea]. Disponible en: https://bogota.gov.co/mi-ciudad/educacion/el-icfes-cambia-la-clasificacion-de-resultados-de-planteles

[67] S.-M. Carrillo-Sierra, O. González-Gutiérrez, A. F. Pereira Mora, D. Y. Arenas Tarazona, A.-E. Arévalo-Batista, y J. D. Hernández Lalinde, "Determinantes del desempeño en pruebas Saber 11 del 2020 en Cúcuta: un análisis multinivel", 2021, Consultado: el 4 de agosto de 2025. [En línea]. Disponible en: https://hdl.handle.net/20.500.12442/11794

[68] ICFES, “Saber al detalle - Boletín 4 -¿Cómo se construye el Índice de Nivel Socioeconómico (INSE) en el contexto de las pruebas Saber?” Consultado: el 4 de agosto de 2025. [En línea]. Disponible en: https://www.icfes.gov.co/publicaciones-icfes/documentos-tecnicos-y-metodologicos/saber-al-detalle/

[69] A. Fernández Aráuz, “Análisis de la resiliencia educativa de los estudiantes costarricenses con datos de la prueba de lectura de la evaluación pisa 2009”, *Revista de ciencias económicas*, vol. 31, núm. 2, pp. 75–99, 2013.

[70] Universidad del Norte, “Diseño de un modelo de machine learning para predicción de rendimiento académico a partir de la resiliencia”, Consultado: el 4 de agosto de 2025. [En línea]. Disponible en: https://manglar.uninorte.edu.co/handle/10584/10497

[71] Y. Zhang y M. Cutumisu, “Predicting the Mathematics Literacy of Resilient Students from High-performing Economies: A Machine Learning Approach”, *Studies in Educational Evaluation*, vol. 83, p. 101412, dic. 2024, doi: 10.1016/j.stueduc.2024.101412.

[72] Y. Choi y J. Sung, “Do Key Predictors of Academic Resilience Differ Across Cultures? Evidence From Korea and the US”, *Youth & Society*, vol. 56, núm. 7, pp. 1237–1262, oct. 2024, doi: 10.1177/0044118X241227563.

[73] P. Lasso Toro, “Cuando se vive el desarraigo. Educación y desplazamiento forzado: una mirada desde el Distrito de Aguablanca, Cali, Colombia”, *Revista Guillermo de Ockham*, vol. 11, núm. 2, pp. 35–51, dic. 2013, doi: 10.21500/22563202.608.