Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Proyecto Aplicado

Predicción de la Resiliencia Escolar en Municipios Colombianos afectados por el Conflicto Armado

Luz Carime Lucumí Hernández
Luis Carlos Martínez Martínez

Directora
Dra. Gloria Inés Álvarez Vargas

Codirector
Dr. Diego Luis Linares Ospina

18 de Diciembre de 2025

Resumen

Este proyecto tuvo como objetivo predecir la resiliencia escolar en los municipios colombianos afectados por el conflicto armado, utilizando técnicas de aprendizaje automático y análisis espacial. Entendiendo la resiliencia escolar como la capacidad de los estudiantes para mantener un buen rendimiento académico a pesar de las adversidades, dado que, aunque estas regiones enfrentan grandes desafíos, se identificaron estudiantes que lograban sobresalir académicamente, lo que permitió destacar los factores que contribuyen a este éxito.

Se desarrolló un modelo predictivo basado en datos educativos, sociales y económicos, que permitió estimar la resiliencia escolar en distintas zonas del país. El proyecto incluyó la recolección y preparación de los datos, la construcción y evaluación de varios modelos predictivos, y el diseño de un prototipo interactivo para visualizar los resultados. Este enfoque, apoyado en el análisis de grandes volúmenes de datos, permitió una mejor comprensión de la resiliencia escolar en contextos de conflicto y ofreció herramientas útiles para la toma de decisiones en políticas educativas y sociales.

Palabras Clave: Resiliencia escolar, Conflicto Armado, Análisis espacial, Machine Learning, Modelos predictivos.

Índice general

1. Introducción **7**

2. Contextualización del Proyecto **9**

2.1. Definición del Problema . . . 9

2.1.1. Planteamiento del Problema . . . 9

2.1.2. Formulación . . . 10

2.1.3. Sistematización . . . 10

2.2. Objetivos . . . 12

2.2.1. Objetivo General . . . 12

2.2.2. Objetivos Específicos . . . 12

2.3. Marco de Referencia . . . 13

2.3.1. Marco Teórico . . . 13

2.3.2. Antecedentes . . . 19

3. Preparación del conjunto de datos para el modelado **23**

3.1. Fuentes y Consolidación de la Información . . . 23

3.1.1. Integración de bases de datos . . . 23

3.1.2. Criterios de selección de variables . . . 24

3.1.3. Ficha técnica del conjunto de datos . . . 26

3.2. Procesamiento . . . 26

3.2.1. Control de duplicados y coherencia temporal . . . 27

3.2.2. Transformaciones proporcionales y variables derivadas . . . 27

3.2.3. Conversión semántica y armonización de categorías . . . 28

3.2.4. Construcción de la variable objetivo: clase resiliencia . . . 28

3.3. Diseño experimental y partición del conjunto de datos . . . 30

3.3.1. División del conjunto de datos y validación . . . 30

3.4. Transformaciones . . . 30

3.4.1. Imputación de valores faltantes . . . 30

3.4.2. Codificación proporcional por municipio y sede . . . 33

3.5. Análisis Exploratorio de Datos (EDA) . . . 36

3.5.1. Distribución general del conjunto de datos . . . 37

3.5.2. Análisis descriptivo de variables sociodemográficas y familiares . . . 38

3.5.3. Análisis institucional y territorial . . . 39

3.5.4. Correlaciones entre variables . . . 41

3.5.5. Análisis estadístico integrado . . . 42

3.5.6. Síntesis interpretativa del EDA . . . 43

4. Desarrollo de Modelos Predictivos para la Resiliencia Escolar **45**
4.1. Selección de Técnicas de Aprendizaje Automático . . . 46
4.1.1. Justificación de los Métodos Elegidos . . . 46
4.1.2. Criterios de Selección Metodológica . . . 47
4.2. Diseño Experimental . . . 49
4.2.1. Estructura de datos y Preparación del Conjunto Analítico . . . 49
4.2.2. Tratamiento del Desbalance de Clases . . . 50
4.2.3. Métricas de Evaluación . . . 50
4.2.4. Procedimiento de Validación Interna . . . 51
4.2.5. Visualizaciones en el Diseño Experimental . . . 51
4.3. Modelos Base . . . 51
4.3.1. Modelo Base: Regresión Logística . . . 52
4.3.2. Modelo Base: XGBoost . . . 52
4.3.3. Modelo Base: SVM . . . 65
4.3.4. Modelo Base: Perceptrón Multicapa (MLP) . . . 67
4.3.5. Comparación General de los Modelos . . . 71
4.4. Optimización de Hiperparámetros . . . 74
4.4.1. Enfoque general de optimización . . . 74
4.4.2. Optimización de XGBoost . . . 75
4.4.3. Optimización del Perceptrón Multicapa (MLP) . . . 77
4.4.4. Limitaciones en la optimización de SVM . . . 79
4.4.5. Selección de los modelos optimizados para evaluación final . . . 80
4.5. Evaluación Final de los Modelos Optimizados . . . 80
4.5.1. Desempeño cuantitativo en el conjunto de prueba . . . 80
4.5.2. Análisis gráfico: ROC, PR y matrices de confusión . . . 81
4.5.3. Interpretación metodológica y operativa . . . 82

5. Desarrollo del Prototipo de Visualización **85**
5.1. Descripción general del prototipo . . . 85
5.2. Flujo de procesamiento y visualización . . . 85
5.3. Interfaz del prototipo . . . 90
5.4. Resultados y potencial de uso . . . 90

6. Conclusiones **93**

7. Trabajos Futuros **95**

Bibliografía **99**

Capítulo 1

Introducción

El conflicto armado en Colombia ha dejado profundas cicatrices en diversos aspectos de la vida social, económica y cultural del país. Entre las áreas más afectadas se encuentra el sector educativo, que ha experimentado interrupciones significativas en su funcionamiento. Las escuelas en los municipios más golpeados por el conflicto han sufrido daños en su infraestructura, desplazamiento forzado de docentes y estudiantes, y escasez de recursos básicos, lo que ha impactado directamente en la calidad educativa. Como resultado, la deserción escolar es una problemática alarmante en estas regiones, perpetuando las desigualdades sociales y limitando las oportunidades para las nuevas generaciones.

A pesar de las adversidades, se han identificado casos notables de resiliencia escolar, donde algunos estudiantes, incluso en los contextos más difíciles, logran mantener un desempeño académico sobresaliente. Este fenómeno, aunque valioso, no está completamente comprendido. La resiliencia escolar, entendida como la capacidad de los estudiantes para superar adversidades y mantener metas educativas a pesar de las condiciones desfavorables, depende de una variedad de factores. Estos incluyen la calidad del entorno escolar, el apoyo comunitario, la disponibilidad de recursos educativos y la proximidad al conflicto armado, entre otros.

El análisis de estos factores y su impacto en la resiliencia escolar es un área crucial que aún no ha sido suficientemente explorada [1], especialmente en un contexto tan complejo como el colombiano. A pesar de los esfuerzos previos para comprender la resiliencia educativa en contextos de conflicto, existen vacíos significativos en el conocimiento sobre los patrones y determinantes específicos que permiten a ciertos estudiantes sobresalir en medio de la adversidad. Este vacío limita la capacidad de diseñar políticas públicas eficaces que puedan fortalecer el sistema educativo y reducir las desigualdades.

En este contexto, el aprendizaje automático (*machine learning*) se perfila como una herramienta poderosa para enfrentar este desafío. Su capacidad para procesar grandes volúmenes de información, descubrir patrones complejos y anticipar resultados a partir de variables clave lo convierte en un recurso valioso. No obstante, su aplicación en el ámbito educativo implica diversos retos, entre los que destacan la preparación y calidad de los datos, la elección de los modelos más apropiados y la validación rigurosa de los resultados obtenidos.

Este proyecto tiene como objetivo principal predecir la resiliencia escolar en municipios colombianos afectados por el conflicto armado, utilizando técnicas de aprendizaje automático y análisis espacial. Se generó un conjunto de datos preprocesados y listos para el análisis, se desarrollaron modelos predictivos validados que permiten identificar los factores más relevantes que influyen en la resiliencia escolar, y se construirá un prototipo de visualización para facilitar la interpretación y uso de los resultados. Además, se está elaborando un documento técnico que describe el proceso, los métodos y los hallazgos, proporcionando una base sólida para investigaciones futuras.

Capítulo 2

Contextualización del Proyecto

2.1. Definición del Problema

En los municipios colombianos afectados por el conflicto armado, algunos estudiantes logran mantener un rendimiento académico sobresaliente a pesar de las adversidades del entorno. Este fenómeno, conocido como resiliencia escolar, refleja la capacidad de estos estudiantes para adaptarse y superar las dificultades inherentes a contextos desfavorables. En estas regiones, el conflicto armado ha causado la destrucción de infraestructura educativa, el desplazamiento forzoso de docentes y estudiantes, y la interrupción de los procesos de enseñanza, lo cual ha contribuido al aumento de las tasas de deserción escolar y a la disminución de la asistencia estudiantil [2, 3]. A pesar de provenir de contextos socioeconómicos desfavorecidos y de enfrentar graves desafíos, algunos estudiantes logran alcanzar un alto nivel de rendimiento académico, lo que subraya la importancia de comprender los factores que posibilitan esta resiliencia escolar [4].

La resiliencia escolar, entendida como la capacidad de los estudiantes para superar obstáculos y mantener un alto desempeño académico en situaciones adversas, depende de una variedad de factores. Estos incluyen la calidad de la infraestructura educativa, la disponibilidad de personal docente, el apoyo de la comunidad y la cercanía al conflicto [5]. A pesar de la importancia de estos factores, existen vacíos en el conocimiento sobre los patrones espaciales y las variables específicas que determinan la resiliencia en los municipios colombianos afectados por el conflicto armado. Esta falta de información precisa y detallada impide el desarrollo de intervenciones efectivas que optimicen los procesos educativos en estas regiones vulnerables.

Por lo tanto, resulta fundamental realizar un análisis de la resiliencia escolar desde un enfoque geoespacial y predictivo, utilizando técnicas avanzadas de aprendizaje automático para identificar los patrones y factores clave que influyen en este fenómeno.

2.1.1. Planteamiento del Problema

En Colombia, el conflicto armado ha dejado profundas huellas en diversas áreas del desarrollo social y económico, y el sector educativo no ha sido una excepción. En los municipios afectados por el conflicto, las escuelas han enfrentado interrupciones continuas debido a daños en la infraestructura, el desplazamiento forzado de docentes y estudiantes, y la falta de recursos básicos para garantizar una educación de calidad. Estos factores han contribuido a una alta tasa de deserción escolar y han dificultado el acceso a una educación adecuada, lo que perpetúa las desigualdades

sociales en estas regiones [2, 6].

A pesar de este entorno adverso, se han identificado casos de resiliencia escolar, donde algunos estudiantes logran mantener un rendimiento académico sobresaliente a pesar de las difíciles condiciones. Este fenómeno de resiliencia plantea un desafío significativo: entender qué factores específicos permiten que ciertos estudiantes se adapten y logren un alto desempeño en un contexto tan desfavorable. La resiliencia escolar, definida como la capacidad de superar las adversidades y mantener metas educativas, depende de múltiples factores, tales como la calidad del entorno escolar, el apoyo comunitario, la disponibilidad de recursos educativos y la cercanía al conflicto armado [4, 5].

Sin embargo, a pesar de los esfuerzos previos, aún se desconocen con claridad los patrones y factores que determinan la resiliencia educativa en estos contextos. Este vacío de conocimiento limita el diseño de políticas educativas efectivas y dificulta la implementación de prácticas exitosas que puedan replicarse en otras regiones afectadas por el conflicto armado.

El uso de tecnologías avanzadas, como el aprendizaje automático, representa una oportunidad valiosa para abordar estos desafíos. Estas técnicas permiten analizar grandes volúmenes de datos y detectar patrones complejos que no son evidentes a simple vista. Sin embargo, existen retos significativos en la preparación de los datos educativos para su análisis, la selección de los modelos adecuados y la validación de los resultados, lo cual es esencial para garantizar conclusiones confiables y útiles.

Este proyecto busca superar estas limitaciones mediante la aplicación de un enfoque predictivo que combine el aprendizaje automático y el análisis espacial para identificar y analizar los factores clave que influyen en la resiliencia educativa en los municipios colombianos afectados por el conflicto armado. A través de este enfoque, se pretende predecir los niveles de resiliencia escolar y señalar áreas prioritarias de intervención, con el objetivo de desarrollar, a futuro, estrategias educativas basadas en evidencia que fortalezcan el sistema educativo en estas regiones y contribuyan a la reducción de las desigualdades sociales y educativas.

2.1.2. Formulación

¿Cómo predecir la resiliencia escolar en municipios colombianos afectados por el conflicto armado, aplicando métodos de aprendizaje automático y análisis espacial?

2.1.3. Sistematización

- ¿Cuáles son las fuentes oficiales disponibles para recolectar datos relevantes sobre la educación y el conflicto armado en Colombia?
- ¿Cuáles son las características socioeconómicas, educativas y geográficas de los municipios que influencian la resiliencia escolar?

- ¿Cómo preparar los datos para garantizar una adecuada implementación de técnicas de aprendizaje automático?
- ¿Cómo construir modelos usando técnicas de aprendizaje automático que permitan predecir la resiliencia escolar?
- ¿Cómo evaluar el desempeño de los modelos desarrollados?
- ¿Cómo visualizar el resultado de la predicción de la resiliencia escolar en municipios colombianos afectados por el conflicto armado?

2.2. Objetivos

2.2.1. Objetivo General

Predecir la resiliencia escolar en municipios colombianos afectados por el conflicto armado, aplicando métodos de aprendizaje automático y análisis espacial.

2.2.2. Objetivos Específicos

1. Preparar los datos para garantizar una implementación adecuada de las técnicas de aprendizaje automático en el análisis de la resiliencia escolar en los municipios colombianos afectados por el conflicto armado.

2. Construir modelos predictivos mediante el uso de técnicas de aprendizaje automático que permitan estimar el nivel de resiliencia escolar en los municipios afectados por el conflicto armado.

3. Evaluar el desempeño de diferentes modelos de aprendizaje automático, para seleccionar el más adecuado.

4. Desarrollar un prototipo para la visualización del resultado de la predicción de la resiliencia escolar en municipios colombianos afectados por el conflicto armado.

2.3. Marco de Referencia

Este marco de referencia tiene como propósito proporcionar los fundamentos teóricos y antecedentes necesarios para comprender los fenómenos relacionados con la resiliencia escolar en municipios colombianos afectados por el conflicto armado. Se divide en dos secciones: la primera aborda los conceptos clave de la resiliencia escolar, los efectos del conflicto armado en la educación y las tecnologías emergentes para el análisis educativo; y la segunda presenta investigaciones previas relevantes a nivel local e internacional.

2.3.1. Marco Teórico

El marco teórico explora los conceptos fundamentales de la resiliencia escolar, los factores que influyen en su desarrollo y el uso de tecnologías predictivas para identificar áreas vulnerables y optimizar políticas educativas. Se busca establecer una base sólida para entender los factores que determinan la resiliencia escolar en municipios colombianos afectados por el conflicto armado, y cómo esta información puede contribuir a mejorar la educación en contextos vulnerables.

Resiliencia Escolar

La resiliencia escolar se refiere a la capacidad de los estudiantes para enfrentar y superar las adversidades que afectan su rendimiento académico, como la violencia, la pobreza o la inestabilidad. Según Masten, la resiliencia es un proceso de adaptación positiva frente a la adversidad [7]. En el ámbito educativo, no solo implica sobrevivir a un entorno difícil, sino lograr avances significativos en términos de aprendizaje y desarrollo personal.

Este concepto es complejo, ya que no solo involucra al estudiante, sino también a su entorno social, educativo y comunitario. En el caso de los estudiantes colombianos afectados por el conflicto armado, factores como la calidad de la infraestructura educativa, el apoyo social y el nivel de exposición a la violencia son fundamentales para desarrollar resiliencia [5].

En términos educativos, la resiliencia se manifiesta como la capacidad de los estudiantes para obtener buenos resultados académicos a pesar de condiciones socioeconómicas adversas o situaciones difíciles. Los estudiantes resilientes, según estudios como los de la OCDE [4], son aquellos que, a pesar de estar en el cuartil más bajo de su contexto socioeconómico, logran un rendimiento académico superior al 75 % de los estudiantes en pruebas estandarizadas, lo que muestra que el éxito académico es posible incluso en contextos desfavorables [8].

Este enfoque permite identificar casos excepcionales de éxito académico, lo que a su vez puede ser útil para comprender y abordar las desigualdades educativas en contextos vulnerables [7].

Perspectiva Local para el Análisis de la Resiliencia

La resiliencia escolar puede abordarse desde una perspectiva local, que compara a estudiantes de bajos recursos que logran alto rendimiento con sus pares en contextos similares dentro de un país. Este enfoque resulta más adecuado para este estudio, ya que se centra en entender la resiliencia en municipios colombianos afectados por el conflicto armado, sin comparaciones internacionales, las cuales podrían no reflejar las características específicas del contexto local.

Impacto de la Resiliencia Educativa

La resiliencia educativa no solo es un indicador del potencial de superación de los estudiantes, sino también un reflejo de la capacidad del sistema educativo para mitigar desigualdades. En contextos como el colombiano, donde las condiciones socioeconómicas desfavorables afectan significativamente el rendimiento académico, la identificación de estudiantes resilientes y los factores que los favorecen puede orientar la formulación de políticas más efectivas.

Además, al reducir la brecha entre estudiantes resilientes y no resilientes, se contribuye a mejorar la movilidad social. Esto es particularmente relevante porque la educación es una de las principales vías para reducir las desigualdades salariales y promover el desarrollo social [8][9].

Conflicto Armado

El conflicto armado en Colombia ha sido una de las crisis más prolongadas y complejas en América Latina, con profundas repercusiones en la vida de millones de personas, especialmente en el ámbito educativo. Se entiende por conflicto armado las confrontaciones violentas entre actores armados ilegales y el Estado, que han afectado a diversas regiones del país, principalmente en zonas rurales. Este conflicto ha generado desplazamientos forzados, destrucción de infraestructura, y alteración de los procesos sociales y educativos en las áreas más afectadas [2].

El Impacto del Conflicto Armado en la Educación

El conflicto armado en Colombia ha tenido efectos devastadores en el sistema educativo, incluyendo la destrucción de infraestructura escolar, lo que crea entornos inseguros y dificulta la continuidad de las clases. Esto ha reducido la cantidad de estudiantes con buenos resultados en exámenes estandarizados, mientras que ha aumentado el porcentaje de estudiantes con bajos niveles de aprendizaje [6]. Además, los desplazamientos forzados interrumpen los estudios de los niños, especialmente en la educación secundaria [10].

Consecuencias del Conflicto Armado en la Educación

El impacto del conflicto armado en el sistema educativo colombiano es profundo y plantea múltiples impactos:

- **Infraestructura Educativa:** Las escuelas en zonas afectadas por el conflicto han sufrido daños debido a ataques y violencia, lo que crea entornos inseguros y limita la continuidad escolar [10].
- **Continuidad Académica:** La violencia reduce el rendimiento académico, con una disminución en los resultados de exámenes estandarizados y un aumento en los niveles de aprendizaje insuficiente [6].
- **Desplazamiento Forzado:** Los desplazamientos interrumpen los estudios, especialmente para los niños, quienes presentan menos años de escolaridad en comparación con sus pares no desplazados, con una brecha significativa en la educación secundaria [10].
- **Desigualdades Regionales:** La falta de recursos y apoyo educativo en las zonas receptoras de desplazados perpetúa las desigualdades en el rendimiento académico y las tasas de finalización escolar [6].

Referentes sobre el Impacto del Conflicto en la Educación

Diversos estudios han documentado que el conflicto armado reduce la probabilidad de aprobar exámenes y afecta negativamente el rendimiento académico, como se observa en el caso de las pruebas Saber 11 en Colombia [6]. Además, investigaciones en otros países afectados por conflictos, como Guatemala y Ruanda, han mostrado impactos similares en los años de escolaridad y las tasas de matrícula, afectando más a las mujeres y las comunidades rurales [10].

Aprendizaje Automático para la predicción de resiliencia escolar

El aprendizaje automático es un campo de la inteligencia artificial orientado a la construcción de modelos capaces de aprender patrones a partir de los datos y generar predicciones sin requerir instrucciones explícitas. Su fortaleza radica en la capacidad para identificar relaciones complejas, no lineales y multidimensionales que suelen escapar a los métodos estadísticos tradicionales [11]. En fenómenos educativos y sociales, estas características resultan especialmente relevantes, pues las relaciones entre variables suelen depender de umbrales, efectos acumulativos, saturaciones o comportamientos diferenciados según subgrupos poblacionales [12].

Asimismo, es frecuente la presencia de interacciones entre variables que solo emergen cuando dos o más características coexisten, así como niveles de multicolinealidad originados en la interdependencia de factores socioeconómicos, institucionales y familiares. Estas particularidades justifican el uso de modelos capaces de capturar estructuras complejas y representar funciones no lineales de

decisión.

Un aspecto central en el aprendizaje automático es la capacidad de generalización, entendida como el desempeño del modelo frente a datos no observados durante el entrenamiento. Esto exige un equilibrio entre dos riesgos: el subajuste, que ocurre cuando el modelo es demasiado simple para capturar los patrones relevantes, y el sobreajuste, que aparece cuando el modelo se ajusta en exceso al ruido o a particularidades del conjunto de entrenamiento [12]. Por ello, es fundamental emplear algoritmos que incorporen mecanismos de control como la regularización, la maximización del margen o la normalización interna, todos ellos esenciales para comprender el comportamiento comparado de los modelos utilizados.

Representación de datos

En aprendizaje automático, cada observación se representa como un vector en un espacio de características, típicamente expresado como Rn, donde n corresponde al número de variables que describen cada instancia [11]. Esta formalización permite que los modelos operen sobre estructuras algebraicas bien definidas, y posibilita que técnicas como SVM o las redes neuronales apliquen transformaciones geométricas: proyecciones, hiperplanos o funciones de activación, para identificar patrones y separar clases. Entender esta representación es clave para analizar cómo se construyen las fronteras de decisión y cómo la dimensionalidad afecta la dificultad del problema [13].

Importancia de variables

La interpretabilidad es un componente crítico de los modelos predictivos, especialmente en el ámbito educativo, donde las decisiones deben sustentarse en la comprensión de los factores que influyen en los resultados [14]. En modelos lineales, la relevancia de cada predictor puede inferirse a partir de los coeficientes estimados; en métodos de ensamble basados en árboles, como los modelos de boosting, puede analizarse su contribución a la reducción del error [15]; y en redes neuronales pueden utilizarse técnicas que permitan no solo evaluar qué tan bien predice un modelo, sino qué factores explican dichas predicciones.

Aprendizaje supervisado y clasificación binaria

El presente estudio se enmarca en el enfoque de aprendizaje supervisado, una categoría del aprendizaje automático en la que los modelos se entrenan a partir de ejemplos etiquetados con una respuesta conocida. En este caso, el objetivo se formula como un problema de clasificación binaria, en el que cada registro debe asignarse a una de dos categorías mutuamente excluyentes. Este enfoque permite construir funciones de decisión que, a partir de variables socioeducativas, institucionales y contextuales, estiman la probabilidad de pertenecer a una clase y orientan la predicción mediante umbrales de decisión específicos.

Modelos como regresión logística, SVM, XGBoost y redes neuronales, son ampliamente empleados en tareas de clasificación con estructuras multivariadas y relaciones no lineales.

Regresión Logística

La regresión logística es un modelo probabilístico para clasificación binaria que transforma una combinación lineal de predictores en una probabilidad mediante la función sigmoide [16]. Su formulación permite interpretar el efecto de cada variable en el logit de la probabilidad, lo cual facilita analizar relaciones directas entre los predictores y el resultado. Aunque su carácter lineal limita su capacidad para capturar interacciones complejas, su interpretabilidad, eficiencia y solidez estadística la convierten en un modelo base indispensable para comparar el rendimiento de enfoques más complejos.

Máquinas de Vectores de Soporte (SVM)

Las Máquinas de Vectores de Soporte, introducidas por Cortes y Vapnik [17], buscan el hiperplano que maximiza el margen entre dos clases. Esta metodología, basada en los vectores de soporte, garantiza robustez frente al ruido y buen desempeño en espacios de alta dimensionalidad. Además, mediante funciones kernel es posible capturar relaciones no lineales sin aumentar explícitamente la dimensionalidad del espacio [18]. Sin embargo, su desempeño puede verse comprometido cuando el número de muestras es muy elevado o la estructura del fenómeno exige mayor flexibilidad que la que los kernels ofrecen.

XGBoost (Extreme Gradient Boosting)

XGBoost es un algoritmo de gradient boosting desarrollado por Chen y Guestrin [19], ampliamente reconocido por su eficiencia, capacidad de regularización y desempeño sobresaliente. Su construcción secuencial de árboles corregidos mediante gradientes permite capturar interacciones no lineales y patrones complejos en los datos. Además, ofrece mecanismos explícitos para controlar el sobreajuste y proporciona medidas claras de importancia de variables, lo que lo convierte en una herramienta eficaz para problemas multicausales.

Perceptrón Multicapa (MLP)

El Perceptrón Multicapa es un tipo de red neuronal feedforward compuesto por capas ocultas con funciones de activación no lineales, lo que le permite aproximar funciones complejas y capturar relaciones altamente no lineales en los datos [13]. El aprendizaje se realiza mediante backpropagation y descenso de gradiente, ajustando los pesos para minimizar la función de pérdida. Su flexibilidad arquitectónica lo hace adecuado para problemas donde múltiples factores interactúan simultáneamente.

Métricas de evaluación

La evaluación de modelos de clasificación se basa en métricas derivadas de la matriz de confusión (TP, FP, TN, FN), tales como *recall*, *precision*, F1-score, *accuracy* y AUC-ROC. El *recall* mide la capacidad del modelo para identificar correctamente la clase positiva; la *precisión* evalúa la fiabilidad de las predicciones positivas; el F1-score ofrece un equilibrio entre ambos indicadores; la *accuracy* expresa la proporción total de aciertos, aunque puede ser engañosa ante desbalance de clases; y el AUC-ROC resume la capacidad discriminativa del modelo en distintos umbrales [20]. Cada métrica refleja una dimensión complementaria del desempeño y su elección depende del objetivo analítico y operativo del sistema.

La evaluación del desempeño de los modelos predictivos requiere el uso de métricas específicas que permitan analizar distintos aspectos de su comportamiento. En problemas de clasificación binaria, estas métricas se derivan de la matriz de confusión, que organiza las predicciones del modelo según cuatro categorías: verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). Cada métrica proporciona una perspectiva distinta sobre la calidad del modelo y permite comprender su utilidad para distintos tipos de tareas o contextos de decisión.

Una de las medidas fundamentales es el *recall* o sensibilidad, que indica la capacidad del modelo para identificar correctamente los casos pertenecientes a la clase positiva. Su cálculo se define como:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Por su parte, la *precisión* evalúa la proporción de predicciones positivas que son correctas, lo cual es esencial cuando se desea minimizar errores asociados a clasificaciones falsas en la clase positiva:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Ambas métricas se combinan en el F1-score, que representa la media armónica entre precisón y *recall*, ofreciendo un indicador equilibrado del rendimiento cuando se requiere ponderar simultáneamente ambos criterios:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Aunque la exactitud (*accuracy*) expresa el porcentaje total de predicciones correctas, su utilidad es limitada en escenarios donde existe un desbalance entre clases, ya que puede generar una falsa impresión de buen desempeño si el modelo se concentra en la clase mayoritaria. Se calcula como:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Finalmente, el AUC-ROC (Área Bajo la Curva ROC) evalúa la capacidad global del modelo para discriminar entre clases a través de distintos umbrales de decisión. Un AUC cercano a 1 indica una alta capacidad discriminativa, mientras que valores próximos a 0.5 reflejan un comportamiento equivalente al azar [20].

En conjunto, estas métricas permiten analizar la calidad de los modelos desde múltiples ángulos y facilitan la selección del enfoque más adecuado para cada problema. La elección final depende de las prioridades del sistema: ya sea maximizar la identificación correcta de la clase positiva, reducir las falsas alarmas, equilibrar ambas dimensiones o evaluar la capacidad discriminativa del modelo a lo largo de diferentes escenarios operativos.

Análisis Espacial

El análisis espacial es una técnica utilizada para estudiar y comprender patrones, relaciones y fenómenos en función de su distribución geográfica. Este enfoque se basa en la idea de que los datos no son solo puntos aislados, sino que tienen una dimensión espacial que puede influir en su comportamiento. A través del análisis espacial, se busca identificar cómo los elementos en un espacio determinado están relacionados entre sí, lo que es clave para este estudio en el contexto de la movilidad urbana en Cali.

2.3.2. Antecedentes

Educación para la Resiliencia: Un Análisis desde la Perspectiva de Niñas, Niños y Docentes

El estudio realizado en Tabasco, México, se centró en identificar las competencias, prácticas y estrategias utilizadas para fortalecer la resiliencia escolar, empleando un enfoque mixto y descriptivo. La investigación involucró a 230 docentes y 53 estudiantes, utilizando encuestas y grupos focales para comprender las percepciones de la comunidad educativa sobre las estrategias para enfrentar situaciones adversas. Entre los hallazgos se destacaron la importancia de la escucha activa y el vínculo emocional entre docentes y estudiantes, así como el valor de las estrategias socioemocionales como el apoyo psicológico y la resolución de conflictos. A diferencia de este enfoque cualitativo, la investigación propuesta emplea aprendizaje automático y análisis geoespacial para predecir la resiliencia escolar a nivel municipal, considerando factores socioeconómicos, educativos y geográficos que influyen en el desempeño de los estudiantes en contextos de conflicto armado. Mientras que el estudio en Tabasco se enfoca en estrategias de intervención basadas en percepciones cualitativas, el trabajo propuesto utiliza un enfoque cuantitativo y predictivo, con un énfasis en la modelización de los factores que inciden en la resiliencia escolar [8].

La Escuela Rural Colombiana en Medio del Conflicto Armado: Un Análisis desde la Información Publicada en Noticias

Este estudio analiza los impactos del conflicto armado en las escuelas rurales de Colombia, utilizando un análisis textual de noticias para identificar los efectos directos e indirectos del conflicto. Los investigadores clasificaron los impactos en çontingentesz ”deliberados”, resaltando la vulnerabilidad de las escuelas rurales debido a su ubicación aislada y la escasa presencia del Estado. Entre los efectos documentados se incluyen la destrucción de infraestructuras, el desplazamiento forzado y el reclutamiento de menores, lo que contribuye al aumento de la deserción escolar. A diferencia de este enfoque cualitativo basado en noticias periodísticas, la investigación propuesta emplea datos educativos y geoespaciales para predecir la resiliencia escolar a nivel municipal, integrando factores como la ubicación geográfica y los efectos del conflicto armado. En este sentido, la principal diferencia radica en que el estudio analizado utiliza un enfoque cualitativo y retrospectivo, mientras que el proyecto propuesto busca una predicción cuantitativa basada en modelos de aprendizaje automático y análisis espacial [9].

A model for predicting academic performance on standardized tests for lagging regions based on machine learning and Shapley additive explanations

Este estudio utiliza algoritmos de clasificación y valores Shapley para predecir el rendimiento académico de los estudiantes colombianos en el examen Saber 11, con un enfoque especial en las regiones rezagadas del país. A través del uso de nueve algoritmos de clasificación, los investigadores identificaron las variables más influyentes en el desempeño académico, como el nivel socioeconómico y la ubicación geográfica de la institución educativa. A pesar de que ambos estudios comparten el uso de aprendizaje automático, la diferencia principal con la investigación propuesta es que este estudio se enfoca en predecir el rendimiento académico, mientras que la investigación en curso tiene como objetivo predecir la resiliencia escolar en municipios colombianos afectados por el conflicto armado. Mientras que el estudio revisado se concentra en la identificación de variables que afectan el rendimiento en pruebas estandarizadas, la investigación propuesta considera un enfoque más holístico, integrando datos educativos, socioeconómicos y geoespaciales en un modelo predictivo de resiliencia escolar [21].

A machine learning prediction of academic performance of secondary school students using radial basis function neural network

Emplea una red neuronal de función base radial (RBFNN) para predecir el rendimiento académico de estudiantes de secundaria, utilizando datos de registros académicos previos y evaluaciones cognitivas y psicomotoras. Los resultados mostraron que las evaluaciones de habilidades cognitivas y psicomotoras tienen un impacto significativo en el rendimiento escolar, con una precisión global del 86.59 %. La diferencia clave entre este estudio y la investigación propuesta radica en el enfoque de predicción: mientras que el estudio en cuestión se centra en predicciones individuales del rendi-

miento académico a través de un modelo de red neuronal, la investigación en curso se enfoca en la resiliencia escolar a nivel municipal, considerando factores geoespaciales y los efectos del conflicto armado en la comunidad educativa. Así, aunque ambos estudios utilizan aprendizaje automático para predecir variables educativas, el enfoque y las variables consideradas en cada caso son muy diferentes. El estudio de redes neuronales se centra en la predicción individual del rendimiento, mientras que el proyecto propuesto busca predecir la resiliencia escolar a partir de un enfoque más sistémico y contextual [22].

Educational inequality in urban settings: A spatial analysis of school distribution and double-shift system challenges – A case study

Explora las desigualdades educativas en Erbil, Irak, utilizando análisis espacial para evaluar la distribución de escuelas primarias y los problemas del sistema de turnos dobles. Identifica la concentración de escuelas en áreas urbanas centrales y su escasa presencia en las periferias, lo que afecta el acceso a la educación. Aunque ambos estudios comparten el uso de análisis espacial, la diferencia fundamental radica en que este estudio se enfoca en desigualdades educativas urbanas, mientras que la investigación propuesta examina el impacto del conflicto armado en la resiliencia escolar en regiones rurales de Colombia. Mientras que el análisis espacial en Erbil busca identificar disparidades en la distribución de escuelas y el acceso en zonas urbanas, el estudio en curso se enfoca en predecir la resiliencia escolar a nivel municipal, considerando variables educativas, sociales y geográficas en contextos de conflicto armado. Por lo tanto, el estudio de Erbil está más orientado a la distribución geográfica de las escuelas y los problemas logísticos, mientras que la investigación propuesta aborda la predicción de la resiliencia escolar en un contexto social y geopolítico específico [23].

Capítulo 3

Preparación del conjunto de datos para el modelado

3.1. Fuentes y Consolidación de la Información

El proceso de preparación de los datos constituyó una fase crítica en la investigación, al garantizar la validez, consistencia y representatividad del conjunto de datos que serviría como base para el modelado predictivo de la resiliencia escolar. Según el modelo CRISP–DM (Cross Industry Standard Process for Data Mining), esta etapa corresponde a las fases de comprensión de los datos y preparación de los datos, esenciales para transformar información heterogénea en un formato analíticamente útil [24].

La construcción del conjunto de datos se orientó a integrar, depurar y documentar múltiples fuentes oficiales de información educativa, socioeconómica y demográfica [25], bajo criterios de calidad y relevancia analítica. Este proceso permitió representar, de manera multivariada, las condiciones institucionales y territoriales que influyen en la resiliencia escolar en municipios colombianos afectados por el conflicto armado.

En particular, el conjunto de datos se construyó a partir de dos fuentes oficiales principales: los resultados de las pruebas Saber 11 del Instituto Colombiano para la Evaluación de la Educación (ICFES), que aportan información académica y sociodemográfica a nivel del estudiante, y la base de Educación Formal del Departamento Administrativo Nacional de Estadística (DANE – C600), que proporciona variables institucionales y de contexto asociadas a las sedes educativas y a los territorios donde estas se ubican.

3.1.1. Integración de bases de datos

La integración de las fuentes de información se realizó con el propósito de construir una visión sistémica del fenómeno educativo [26], relacionando el desempeño académico individual con factores institucionales y contextuales. Para ello, se emplearon dos bases de datos principales.

Fuente	Descripción	Cobertura / Periodo	Variables relevantes	Referencia
ICFES – Saber 11	Resultados académicos individuales en lectura, matemáticas, ciencias, inglés, sociales y ciudadanas.	2021–2023	Puntajes, características del estudiante, variables familiares, ubicación geográfica.	ICFES (2024)
DANE – Educación Formal (C600)	Información de sedes educativas oficiales y no oficiales.	2021–2023	Docentes, matrícula, infraestructura, servicios TIC.	DANE (2024)

Cuadro 3.1: Fuentes de información.

Nivel	Fuente	Clave de integración	Variables principales	Rol en el modelo
Micro (Estudiante)	ICFES Saber 11	COLE_COD_DANE_SEDE	Puntajes, percentiles, características familiares y socioeconómicas	Variable objetivo y atributos individuales
Meso (Institución / Sede)	DANE C600, TIC, Personal	COLE_COD_DANE_SEDE	Docentes, matrícula, servicios TIC	Condiciones institucionales
Macro (Municipio)	Indicadores DANE y víctimas	DIVIPOLA	Tasa de víctimas, aprobación, jerarquía docente	Condiciones territoriales

Cuadro 3.2: Integración de fuentes de información.

Previo a la aplicación de cualquier técnica de imputación o transformación de los datos, se realizó un proceso de integración y depuración inicial de las bases de información. Esta etapa incluyó la homologación de identificadores territoriales y de sede educativa, la verificación de consistencia temporal entre las fuentes, la eliminación de registros duplicados y la validación de tipos de datos. Solo una vez consolidado el conjunto analítico y verificada su coherencia estructural, se procedió al análisis de valores faltantes y a las etapas posteriores de preprocesamiento.

La integración se realizó mediante el identificador territorial y de sede educativa (COLE_COD_DANE_SEDE), garantizando coherencia geográfica y temporal. El procedimiento se implementó en Python usando `pandas`, priorizando trazabilidad y reproducibilidad. Este enfoque permitió estructurar un análisis multinivel [26].

3.1.2. Criterios de selección de variables

El proceso de selección de variables se basó en criterios teóricos, estadísticos y técnicos, alineados con la literatura sobre Educational Data Mining y Learning Analytics [27].

- **Criterios teóricos**: dimensiones clave como capital educativo familiar, infraestructura, TIC y nivel socioeconómico [7, 1].
- **Criterios estadísticos**: completitud $\geq 90\,\%$, variabilidad, correlación con `clase_resiliencia`.
- **Criterios técnicos**: compatibilidad con algoritmos supervisados.

Variable	Fuente	Tipo	Dimensión	Criterio de Selección
percentil_global*	ICFES	Numérica	Micro	Alta correlación con la resiliencia; base para la construcción de la variable objetivo
cole_area_ubicacion	ICFES	Categórica	Meso	Diferencia estructural urbano–rural
cole_jornada	ICFES	Categórica	Meso	Impacto en el tiempo efectivo de aprendizaje
cole_calendario	ICFES	Categórica	Meso	Diferencias curriculares y temporales
cole_caracter	ICFES	Categórica	Meso	Distinción entre formación académica y técnica
estu_inse_individual	ICFES	Numérica	Micro	Indicador robusto del contexto socioeconómico familiar
fami_estratovivienda	ICFES	Categórica	Micro	Representa las condiciones materiales del hogar
fami_situacioneconomica	ICFES	Categórica	Micro	Percepción directa de vulnerabilidad económica
nivel_capital_familiar	Derivada	Categórica	Micro	Síntesis del entorno educativo y cultural del hogar
nivel_acceso_recursos	Derivada	Categórica	Micro	Acceso a recursos educativos para el aprendizaje
NIVEL_USOTIC	ICFES	Categórica	Micro	Intensidad y frecuencia de uso de tecnologías
Frecuencia_TIC	ICFES	Categórica	Micro	Aproximación conductual al acceso digital
CANT_DOCENTES	DANE	Numérica	Meso	Capacidad institucional en infraestructura humana
ratio_docentes	Derivada	Numérica	Meso	Normalización del personal docente por tamaño de sede
Cant_Alum	DANE	Numérica	Meso	Base para el cálculo de proporciones institucionales
COL_%aprobación	DANE	Numérica	Meso	Indicador del clima académico institucional
TASA_VICTIMAS	DANE	Numérica	Macro	Exposición territorial al conflicto armado
cole_mcpio_ubicacion_te	Derivada	Numérica	Macro	Efecto agregado municipal (target encoding)
cole_cod_dane_sede_te	Derivada	Numérica	Meso	Efecto agregado específico de la sede educativa

Cuadro 3.3: Matriz de criterios de selección de variables explicativas.

Además, se elaboró un diccionario de datos que documenta cada variable, su origen, tipo, rango, descripción y transformación aplicada. Este documento (Anexo A) garantiza la transparencia, trazabilidad y replicabilidad del proceso de modelado, en línea con las buenas prácticas de la gestión de datos para investigación.

3.1.3. Ficha técnica del conjunto de datos

Con el fin de caracterizar el dataset resultante de la consolidación e integración, se elaboró una ficha técnica siguiendo las recomendaciones del DANE para la producción estadística [25] y los principios FAIR (Findable, Accessible, Interoperable, Reusable) para la gestión de datos científicos [28].

Elemento	**Descripción**
Nombre	Dataset Resiliencia Escolar Colombia 2021–2023
Unidades	Estudiantes de grado 11, agregados por sede y municipio
Tamaño	1,607,205 registros
Variables	32 variables transformadas y codificadas
Temporalidad	2021–2023
Geografía	32 departamentos, 1,097 municipios
Formato	CSV UTF-8
Herramientas	Python (Pandas, NumPy)
Variable objetivo	`clase_resiliencia` (1 = resiliente, 0 = no resiliente)

Cuadro 3.4: Características del dataset resultante.

Figura 3.1: Resumen visual del proceso de consolidación e integración.

La consolidación de datos oficiales provenientes del ICFES y el DANE, complementada con información contextual sobre el conflicto armado, permite configurar un conjunto de datos robusto, coherente y representativo del sistema educativo colombiano. Este proceso fortalece la validez externa del estudio y posibilita la modelización de la resiliencia escolar desde una perspectiva territorial y sistémica, superando enfoques centrados exclusivamente en el rendimiento individual [2, 3].

3.2. Procesamiento

El procesamiento y depuración de los datos constituyeron una fase orientada a transformar el conjunto consolidado de información en una estructura analítica limpia, consistente y trazable, apta para la aplicación de técnicas de aprendizaje automático. Esta etapa se centró en la corrección de

inconsistencias, la estandarización de formatos, la generación de variables derivadas y el control de posibles fuentes de error que pudieran afectar el desempeño de los modelos predictivos de resiliencia escolar.

El objetivo fue transformar un conjunto heterogéneo de datos provenientes de diversas fuentes (ICFES Saber 11, DANE Educación Formal, y registros de desplazamiento por conflicto armado) en un dataset integrado, limpio y consistente, libre de valores nulos y sin fuga de información.

3.2.1. Control de duplicados y coherencia temporal

Inicialmente, se ejecutó un proceso de verificación de duplicidad en los microdatos del ICFES (2021–2023) empleando las claves `ESTU_CONSECUTIVO`, `COLE_COD_DANE_SEDE` y `PERCENTIL_GLOBAL`. Los registros duplicados exactos fueron eliminados, y en los casos donde un mismo estudiante presentaba múltiples exámenes en un mismo periodo, se conservó la observación más completa o la de última presentación.

Se calculó además la edad del estudiante a partir de la fecha de nacimiento y el periodo académico de presentación de la prueba. Para cada cohorte, se establecieron fechas estándar de presentación según calendario A o B (abril y agosto, respectivamente) y se excluyeron registros con edades fuera del rango 12 a 70 años, garantizando que la información corresponda con las condiciones reales de los sujetos o situaciones analizadas.

3.2.2. Transformaciones proporcionales y variables derivadas

Con el fin de representar adecuadamente el contexto institucional y municipal, se generaron variables proporcionales normalizadas por la cantidad de estudiantes por sede, tal como se observa en la Ecuación 4.1:

$$tasa_victimas = \frac{CANT_victimasconflicto}{Cant_Alum + 1}$$

$$ratio_docentes = \frac{N_personal_acadmico}{Cant_Alum + 1}$$

$$ratio_admin = \frac{N_PERSONAL_ADMINISTRATIVO}{Cant_Alum + 1}$$

Posteriormente, las variables originales se eliminaron para evitar colinealidad y redundancia. De igual forma, se creó la variable edad del estudiante (EDAD) y se validaron sus rangos de distribución por cohorte.

3.2.3. Conversión semántica y armonización de categorías

Durante la integración de fuentes se detectaron diferencias ortográficas y de codificación en variables categóricas ("rural", "RURAL", "Rural"), por lo que se aplicó un proceso de homogeneización semántica, que incluyó:

- Estandarización de nombres y tildes.
- Conversión a mayúsculas uniformes.
- Consolidación de categorías equivalentes.
- Verificación de coherencia territorial según codificación DIVIPOLA y `COLE_COD_DANE_SEDE`.

Asimismo, las variables originalmente binarias `TIENE_APOYO_INCLUSION` y `COL_Bilingue` fueron mapeadas a categorías "Sí" y "No" y transformadas a tipo categórico, con el propósito de mejorar la interpretación y evitar un tratamiento numérico inadecuado durante el modelado.

3.2.4. Construcción de la variable objetivo: clase resiliencia

La variable objetivo del estudio, denominada clase resiliencia, fue construida con el propósito de operacionalizar el concepto de resiliencia escolar de manera coherente con el enfoque del proyecto "Predicción de la Resiliencia Escolar en Municipios Colombianos Afectados por el Conflicto Armado". Desde una perspectiva del presente proyecto, la resiliencia escolar se concibe como un fenómeno que emerge cuando un estudiante alcanza altos niveles de desempeño académico a pesar de enfrentar condiciones socioeconómicas y contextuales adversas.

En este sentido, la clase resiliencia se definió a partir de la combinación de dos componentes fundamentales:

a) Alto desempeño académico

El primer componente corresponde al resultado educativo observable, representado por el desempeño académico del estudiante en la prueba *Saber 11*. Este se operacionalizó mediante el *percentil global*, indicador que sintetiza el rendimiento relativo del estudiante en el conjunto de áreas evaluadas.

Se estableció como criterio de alto desempeño la pertenencia al percentil global igual o superior al percentil 70, lo cual permite identificar a los estudiantes con resultados académicos.

b) Contexto socioeconómico y territorial adverso

El segundo componente corresponde a la identificación de condiciones adversas, las cuales se definieron a partir de variables socioeconómicas y contextuales disponibles en las fuentes oficiales

integradas en el estudio. En particular, se consideró que un estudiante se encuentra en un contexto adverso cuando cumple al menos una de las siguientes condiciones:

- Presenta un índice socioeconómico individual (INSE) clasificado en nivel medio-bajo o bajo, lo que refleja limitaciones en el capital socioeconómico familiar.
- Ha sido identificado como víctima del conflicto armado, o se encuentra matriculado en una institución educativa que atiende población víctima del conflicto armado, de acuerdo con los registros territoriales y educativos consolidados.

Estas condiciones permiten capturar tanto la dimensión socioeconómica individual como la dimensión territorial e institucional de la adversidad, alineándose con un enfoque multinivel del fenómeno educativo.

3.2.4.1. Definición de la variable objetivo: clase resiliencia

Con base en los componentes descritos, la variable *clase resiliencia* fue definida de forma binaria de la siguiente manera:

- **Clase 1 (Resiliente):** estudiantes que presentan alto desempeño académico (percentil global ≥ 70) y que, simultáneamente, se encuentran en un contexto socioeconómico y/o territorial adverso, caracterizado por un INSE medio-bajo o bajo y/o por la condición de victimización asociada al conflicto armado a nivel individual o institucional.
- **Clase 0 (No resiliente):** estudiantes que no cumplen de manera simultánea con ambas condiciones.

Esta definición permite diferenciar conceptualmente el desempeño académico elevado en contextos favorables de aquel que se alcanza bajo condiciones estructuralmente desfavorables, garantizando que la variable objetivo represente efectivamente el fenómeno de resiliencia escolar.

3.2.4.2. Momento de construcción y control metodológico

La variable *clase resiliencia* fue construida una vez integradas las fuentes de información y antes de la partición del conjunto de datos, utilizando exclusivamente información observada y no imputada. Posteriormente, todas las variables directamente involucradas en su definición y los indicadores de desempeño académico fueron excluidas del conjunto de variables predictoras, evitando así cualquier forma de fuga de información durante el entrenamiento de los modelos.

Este enfoque garantiza que el modelo predictivo no identifique de manera trivial a los estudiantes resilientes, sino que aprenda los patrones socioeconómicos, institucionales y territoriales asociados a la probabilidad de alcanzar alto desempeño académico en contextos adversos, lo cual responde directamente al objetivo central del proyecto.

3.3. Diseño experimental y partición del conjunto de datos

3.3.1. División del conjunto de datos y validación

Tras las etapas de integración, limpieza, construcción de la variable objetivo y depuración de variables, el conjunto de datos consolidado fue organizado mediante un esquema *Hold-Out* con validación interna. Este diseño se estructuró en tres cortes operativos:

1. Separación inicial de un conjunto de prueba independiente.
2. Subdivisión del conjunto de entrenamiento global en entrenamiento y validación.
3. Preparación final de las matrices de datos para el modelado.

En una primera partición, el conjunto de datos fue dividido en un conjunto de entrenamiento global (*X_train_total*) y un conjunto de prueba (*X_test*), reservado exclusivamente para la evaluación final del desempeño. Posteriormente, el conjunto de entrenamiento global fue subdividido en entrenamiento (*X_train*) y validación (*X_val*), utilizados para el ajuste de los modelos y la selección de hiperparámetros. Este diseño garantizó en todo momento la independencia del conjunto de prueba y el control explícito de la fuga de información.

Bajo esta estructura, las proporciones observadas fueron las siguientes:

- **Train total / Test:** 1.285.764 registros (80 %) y 321.441 registros (20 %), respectivamente.
- **Train / Validación (dentro de Train total):** 1.028.611 registros (80 % de *X_train_total*) y 257.152 registros (20 % de *X_train_total*), respectivamente.

Este esquema fue seleccionado por tres razones metodológicas: (i) disponer de un conjunto de prueba completamente aislado para reportar desempeño final sin sesgos; (ii) habilitar una validación interna estable para la selección de hiperparámetros; y (iii) mantener un enfoque computacionalmente eficiente y reproducible, considerando el tamaño del conjunto de datos y el número de modelos evaluados.

En términos de control de fuga de información, ningún procedimiento de aprendizaje de parámetros (imputación, codificación y escalamiento) utilizó información proveniente de *X_val* o *X_test*. Todas las transformaciones fueron ajustadas exclusivamente con *X_train* y posteriormente aplicadas de forma consistente a los conjuntos de validación y prueba.

3.4. Transformaciones

3.4.1. Imputación de valores faltantes

El tratamiento de valores faltantes constituye una etapa significativa del proceso de preparación del conjunto de datos. La identificación inicial de valores ausentes evidenció que estos se concentraban principalmente en variables de carácter institucional y contextual provenientes de la base

de Educación Formal del Departamento Administrativo Nacional de Estadística (DANE – C600), mientras que la base de datos del Instituto Colombiano para la Evaluación de la Educación (ICFES) presentó una mayor completitud en las variables académicas y sociodemográficas a nivel del estudiante.

La imputación se realizó sobre la base de datos consolidada, una vez integradas las fuentes del ICFES y el DANE y verificada su coherencia estructural. Posteriormente, el conjunto de datos fue dividido siguiendo un esquema de partición en dos etapas. En una primera división, se separó un conjunto de entrenamiento global *X_train_total* y un conjunto de prueba *(X_test)*, reservando este último exclusivamente para la evaluación final del modelo. En una segunda división, el conjunto de entrenamiento global fue particionado en un conjunto de entrenamiento *(X_train)* y un conjunto de validación *(X_val)*. Con el fin de prevenir la fuga de información (data leakage), las estrategias de imputación se ajustaron únicamente sobre el conjunto de entrenamiento *(X_train)*, replicando posteriormente los parámetros aprendidos de forma consistente sobre los conjuntos de validación y prueba.

Identificación de valores faltantes

El análisis de completitud evidenció que la mayoría de las variables presentaban porcentajes bajos o moderados de valores faltantes. No obstante, se identificaron algunas variables con niveles relevantes de ausencia, particularmente:

- `estu_tieneetnia`: 13,14 % de valores faltantes.
- `cole_bilingue`: 18,10 % de valores faltantes.

Estas variables, de naturaleza categórica, mostraron un patrón de ausencia no aleatorio, asociado a procesos de autorreporte o a registros administrativos incompletos.

Adicionalmente, otras variables categóricas y numéricas presentaron porcentajes de valores faltantes inferiores al 10 %, entre las que se destacan: `fami_estratovivienda` (7,02 %), `estu_dedicacioninternet` (6,58 %), `estu_dedicacionlecturadiaria` (6,37 %), `fami_situacioneconomica` (4,67 %), `estu_inse_individual` y `estu_nse_individual` (4,30 %), `cole_caracter` (3,96 %), `Frecuencia_TIC` (2,27 %), `NIVEL_USOTIC` (1,94 %), `Cant_Alum` (0,82 %) y `percentil_ingles`, `percentil_global` y `COL_%aprobación` (1 %).

Estrategias de imputación

Las estrategias de imputación se definieron en función del tipo de variable, su nivel de agregación y su fuente de origen. En particular, las variables categóricas familiares, académicas e institucionales fueron imputadas utilizando la moda, calculada a nivel de sede educativa, preservando el patrón más frecuente observado en los datos. Las variables numéricas de contexto institucional fueron

imputadas mediante la mediana, como estadístico robusto frente a valores extremos, calculada a nivel municipal.

En el caso de la variable demográfica `estu_tieneetnia`, se aplicó una imputación proporcional basada en la distribución observada por año y sede educativa, garantizando coherencia con la estructura original del conjunto de datos. La Tabla 3.5 presenta de manera sintética las variables imputadas, el método empleado y el nivel de agregación considerado.

Tipo de variable	**Variable(s)**	**Estrategia de imputación**	**Nivel de agregación**
Categóricas familiares	fami_estratovivienda, fami_situacioneconomica, fami_educacionpadre, fami_educacionmadre, fami_tieneinternet, fami_tienecomputador, fami_tieneserviciotv, fami_comederivados, fami_comecarnepescadohuevo, fami_comecerealfrutoslegumbre	Moda	Por sede
Categóricas académicas	estu_dedicacioninternet, estu_dedicacionlecturadiaria, cole_caracter, Frecuencia_TIC, NIVEL_USOTIC	Moda	Por sede
Numéricas institucionales	PROMEDIO_JERARQUIA_DOCENTE, estu_inse_individual, COL_%aprobación	Mediana	Por municipio
Demográficas	estu_tieneetnia	Proporcional (por año y sede)	Por sede
Variables de contexto	COLE_NATURALEZA, COLE_CALENDARIO, COLE_JORNADA	Moda	Por sede

Cuadro 3.5: Estrategias de imputación aplicadas según tipo de variable y nivel de agregación.

Verificación de la distribución antes y después de la imputación

Con el objetivo de garantizar que el proceso de imputación no alterara la estructura estadística del conjunto de datos, se realizó una verificación de la distribución de las variables antes y después de la imputación. Esta evaluación consistió en la comparación de estadísticas descriptivas. así como la inspección visual de las distribuciones mediante histogramas.

Los criterios utilizados para validar la idoneidad de la imputación incluyeron la estabilidad de la media y la mediana, la conservación del rango intercuartílico y la ausencia de cambios abruptos en la forma de las distribuciones. Como se resume en la Tabla 3.6, las variables evaluadas mantuvieron un comportamiento estadístico estable tras la imputación, sin evidenciar distorsiones relevantes.

Variable	Tipo	% valores faltantes (pre)	Estadístico evaluado	Antes de imputación	Después de imputación	Variación observada	Criterio de validación
estu_inse_individual	Numérica	4,30 %	Media / Mediana / IQR	Estables	Estables	Variación mínima	Conservación de tendencia central
estu_nse_individual	Numérica	4,30 %	Media / Mediana / IQR	Estables	Estables	Variación mínima	Robustez frente a imputación
estu_nse_establecimiento	Numérica	0,05 %	Media / Mediana	Estables	Estables	Sin cambios relevantes	Bajo impacto de imputación
Cant_Alum	Numérica	0,82 %	Mediana / IQR	Estables	Estables	Variación marginal	Preservación de dispersión
fami_estratovivienda	Categórica	7,02 %	Moda / Frecuencia	Moda dominante	Moda dominante	Sin cambio	Preservación del patrón
estu_dedicacioninternet	Categórica	6,58 %	Moda / Frecuencia	Categoría dominante	Categoría dominante	Sin cambio	Consistencia categórica
estu_dedicacionlectura diaria	Categórica	6,37 %	Moda / Frecuencia	Categoría dominante	Categoría dominante	Sin cambio	Estabilidad distributiva
fami_situacioneconomica	Categórica	4,67 %	Moda / Frecuencia	Moda dominante	Moda dominante	Sin cambio	Coherencia socioeconómica
NIVEL_USOTIC	Categórica	1,94 %	Moda / Frecuencia	Moda dominante	Moda dominante	Sin cambio	Preservación del perfil TIC
Frecuencia_TIC	Categórica	2,27 %	Moda / Frecuencia	Categoría dominante	Categoría dominante	Sin cambio	Coherencia en uso TIC
estu_tieneetnia	Categórica	33,14 %	Proporción	“No” predominante	“No” predominante	Ajuste controlado	Imputación coherente con distribución
cole_bilingue	Categórica	18,10 %	Proporción	“No” predominante	“No” predominante	Ajuste controlado	Preservación institucional

Cuadro 3.6: Validación estadística del impacto de la imputación de valores faltantes.

De manera complementaria, se realizó una comparación descriptiva de las principales estadísticas antes y después de la imputación para las variables numéricas imputadas, observándose variaciones mínimas entre ambos escenarios. Estos resultados respaldan la pertinencia de las estrategias empleadas y la preservación de las características originales del conjunto de datos.

3.4.2. Codificación proporcional por municipio y sede

Con el fin de incorporar información contextual de carácter territorial sin introducir fuga de información, se aplicó una estrategia de codificación proporcional (target encoding) sobre variables geográficas de alta cardinalidad. Específicamente, se construyeron las siguientes variables derivadas:

`cole_mcpio_ubicacion_te`: tasa promedio de estudiantes resilientes por municipio.
`cole_cod_dane_sede_te`: tasa promedio de estudiantes resilientes por sede educativa.

Ambas variables se calcularon exclusivamente sobre el conjunto de entrenamiento (`X_train`), utilizando la proporción observada de la clase resiliencia. Posteriormente, estos valores fueron aplicados de forma consistente a los conjuntos de validación y prueba. En los casos en que un municipio

o sede no estuviera presente en el conjunto de entrenamiento, el valor fue imputado utilizando la tasa global de resiliencia, evitando así el uso de información del conjunto de prueba durante el proceso de codificación.

Esta estrategia permitió capturar patrones territoriales relevantes asociados a la resiliencia escolar, manteniendo la integridad metodológica del diseño experimental.

3.4.3 Eliminación de variables con fuga de información o alta correlación

Como parte del proceso de depuración final del conjunto analítico, se eliminaron variables que generaban fuga de información, redundancia estadística o correlaciones directas con la variable objetivo. En particular, se excluyeron:

Variables directamente relacionadas con la definición de la clase resiliencia:

punt_global
percentil_c_naturales
percentil_ingles
percentil_lectura_critica
percentil_matematicas
percentil_sociales_ciudadanas

Identificadores y variables temporales:

estu_fechanacimiento
estu_fechanacimiento_original
fecha_presentacion
cole_codigo_icfes
cole_cod_dane_establecimiento
cole_cod_dane_sede

Variable redundante:

estu_nse_individual, debido a que su información se encuentra contenida en la variable `estu_inse_individual`, construida mediante una metodología más robusta basada en teoría de respuesta al ítem (ICFES, 2024).

La eliminación de estas variables permitió reducir la dimensionalidad del conjunto de datos, mejorar la estabilidad del entrenamiento de los modelos y garantizar que las predicciones obtenidas respondieran a patrones estructurales reales, y no a artefactos derivados de información explícita sobre la variable objetivo.

3.4.4 Estado final de los subconjuntos para modelado

Conjunto	Registros	Variables	Rol en el experimento
`X_train_total`	1.285.764	32	Entrenamiento global (base para generar train y validación)
`X_test`	321.441	32	Prueba independiente, reservada para evaluación final
`X_train`	1.028.611	32	Entrenamiento (ajuste de modelos y aprendizaje de transformaciones)
`X_val`	257.152	32	Validación interna (selección/ajuste de hiperparámetros)

Cuadro 3.7: Estructura final del conjunto de datos según la partición.

Transformaciones

Las transformaciones del preprocesamiento fueron ejecutadas siguiendo la regla de que los parámetros se aprendieron únicamente con el conjunto de entrenamiento (`X_train`) y posteriormente se aplicaron de forma consistente a `X_val` y `X_test`. En particular:

- **Codificación de variables categóricas (One-Hot Encoding):** se ajustó únicamente con las categorías observadas en `X_train`, y luego se transformaron `X_val` y `X_test` respetando el mismo espacio de columnas.

- **Estandarización de variables numéricas (StandardScaler):** para modelos sensibles a escala (Regresión Logística, SVM y MLP), se ajustó sobre `X_train` ya codificado y se aplicó a `X_val` y `X_test`.

- **Modelos basados en árboles (XGBoost):** se entrenaron usando directamente los datos imputados y codificados, sin requerir estandarización.

Este flujo de transformación garantiza que el conjunto de validación y prueba no influya en la estimación de parámetros del preprocesamiento, preservando la integridad de la evaluación del desempeño predictivo.

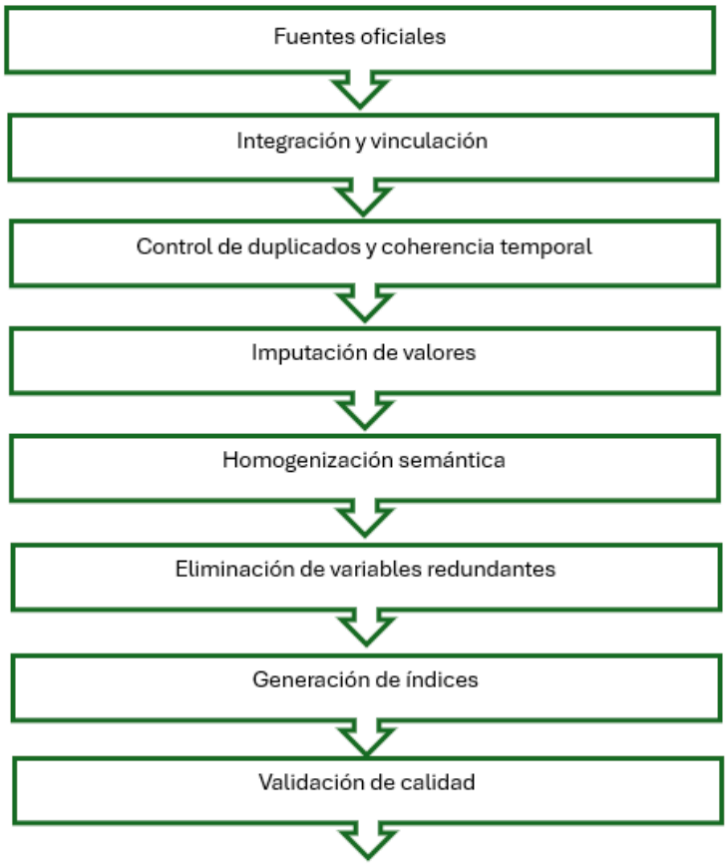

Figura 3.2: Flujo general del procesamiento y depuración del conjunto de datos.

3.5. Análisis Exploratorio de Datos (EDA)

El Análisis Exploratorio de Datos (EDA) constituye una fase fundamental en el proceso de comprensión y caracterización del conjunto de información, orientada a identificar patrones, relaciones, distribuciones y posibles anomalías que pudieran influir en la etapa de modelado. Esta fase permite evaluar la calidad de las imputaciones, validar la coherencia de las transformaciones y detectar relaciones significativas entre variables predictoras y la variable objetivo (`clase_resiliencia`).

De acuerdo con Tukey (1977), el EDA tiene como propósito "comprender lo que los datos pueden revelar más allá de los modelos formales", mediante un enfoque visual, descriptivo y comparativo que guíe el desarrollo de hipótesis y modelos predictivos.

El EDA se desarrolló sobre el conjunto de datos consolidado (ICFES–DANE) correspondiente al periodo 2021–2023, compuesto por 1.607.205 registros. Este análisis se realizó previo a la partición

del conjunto de datos y con un propósito exclusivamente descriptivo y diagnóstico, orientado a caracterizar las distribuciones, relaciones y patrones generales del fenómeno de estudio.

Como resultado de este proceso y de las etapas posteriores de depuración y selección de variables, el conjunto analítico fue reducido a 32 variables finales. Posteriormente, tras la aplicación de la codificación de variables categóricas mediante One-Hot Encoding, dichas variables se expandieron a 62 variables predictoras, las cuales constituyeron la base final para el modelado predictivo.

3.5.1. Distribución general del conjunto de datos

La revisión inicial de la estructura evidenció un conjunto balanceado en cobertura temporal y territorial. Los tres años de observación (2021–2023) presentaron volúmenes similares de registros (aprox. 530,000 por año), correspondientes a estudiantes de grado 11 de todas las regiones del país.

La variable `clase_resiliencia` —definida como 1 para estudiantes con percentil_global 70— permitió clasificar los casos en dos grupos: resilientes (26 %) y no resilientes (74 %). Esta proporción refleja la naturaleza inherentemente desbalanceada del fenómeno, aspecto considerado posteriormente en las estrategias de balanceo y evaluación de modelos (Capítulo 4).

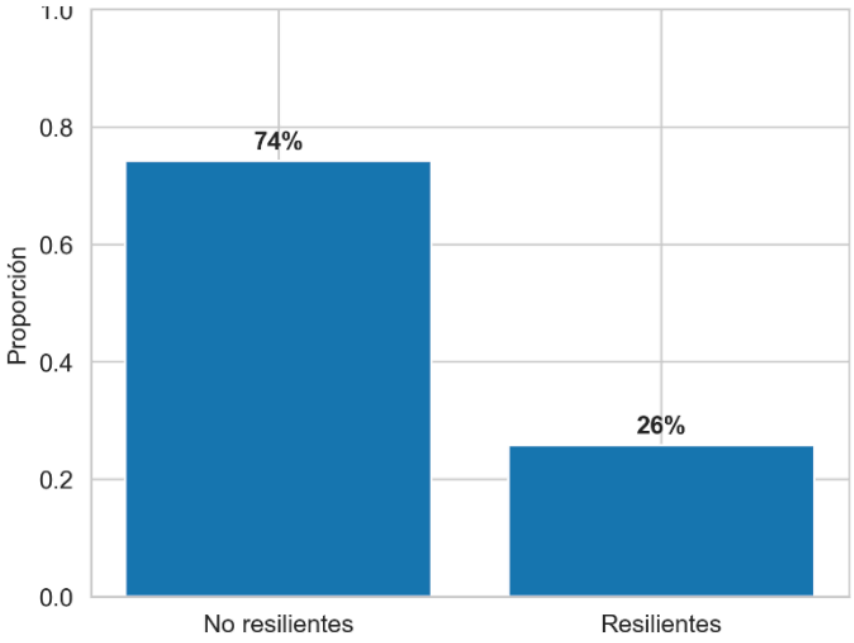

Figura 3.3: Distribución de la variable objetivo.

3.5.2. Análisis descriptivo de variables sociodemográficas y familiares

El conjunto de datos refleja una composición heterogénea de los estudiantes en términos de condiciones familiares, acceso a recursos y contexto socioeconómico.

Las variables incluidas en esta dimensión permiten caracterizar el entorno del hogar y el capital educativo familiar, a partir de indicadores derivados del módulo familiar del ICFES y de índices agregados construidos con base en los datos del DANE.

Los elementos más representativos se pueden organizar en cuatro ejes: estrato socioeconómico, capital familiar, acceso a recursos e infraestructura TIC.

Distribución del estrato socioeconómico

La Figura 3.4 muestra la distribución porcentual de los estudiantes por estrato socioeconómico y área de ubicación de la institución. Los resultados indican que la población estudiantil se concentra principalmente en los estratos 1, 2 y 3, que en conjunto agrupan más del 90 % de los casos.

En el sector rural, predomina el estrato 1 (44 %), seguido por el estrato 2 (38 %), mientras que en el sector urbano las proporciones más altas se registran en los estratos 2 (30 %) y 3 (25 %). Los estratos altos (4, 5 y 6) tienen una representación marginal en ambos contextos (menor al 10 %), evidenciando la fuerte concentración de los estudiantes en contextos de vulnerabilidad socioeconómica.

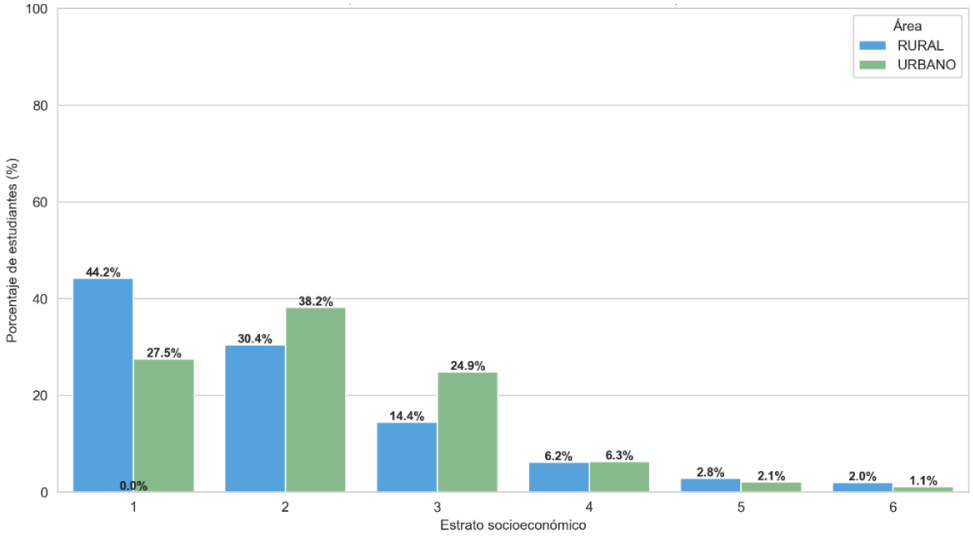

Figura 3.4: Distribución de estrato socioeconómico (urbano-rural).

Capital familiar, acceso a recursos e infraestructura TIC

La Figura 3.5 complementa este análisis con tres dimensiones asociadas al entorno educativo del hogar e institucional:

- **Nivel de capital familiar**: Se observan proporciones similares entre zonas urbanas y rurales en el nivel medio, aunque con una mayor concentración de estudiantes rurales en niveles bajos (aprox. 50 %), frente a una mayor proporción urbana en nivel medio (aprox. 57 %).
- **Nivel de acceso a recursos**: En términos generales, cerca del 40 % de los estudiantes urbanos y 42 % de los rurales se ubican en el nivel medio de acceso a recursos, mientras que los niveles altos no superan el 20 % en ninguno de los dos contextos. Esto confirma que la mayoría de los hogares dispone de recursos limitados para apoyar los procesos educativos fuera del aula.
- **Índice de infraestructura TIC**: Los resultados evidencian una brecha más marcada entre contextos: mientras que el 54,9 % de los estudiantes rurales se ubica en el nivel alto de infraestructura TIC (índice >0,70), esta proporción es inferior al 30 % en áreas urbanas. Sin embargo, los niveles muy altos (>0,90) siguen siendo minoritarios (7 % en urbano y 2,5 % en rural), lo que evidencia que, aunque la cobertura tecnológica ha avanzado, persisten limitaciones en la calidad del acceso y la disponibilidad de equipamiento en las instituciones educativas.

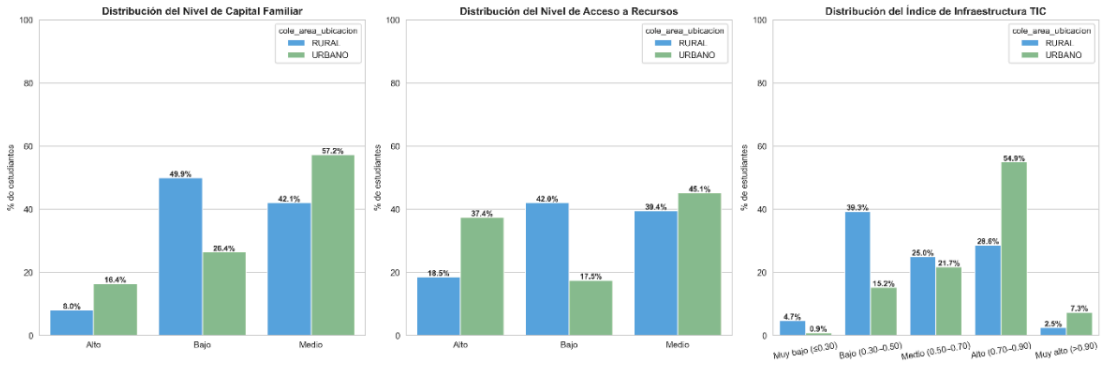

Figura 3.5: Capital familiar, acceso a recursos e infraestructura TIC (urbano-rural).

3.5.3. Análisis institucional y territorial

El análisis a nivel institucional permitió examinar variables que reflejan la disponibilidad y distribución de recursos humanos, así como la incidencia territorial del conflicto armado. A nivel institucional (meso), se analizaron variables provenientes del DANE Educación Formal y de los módulos TIC e Infraestructura En particular, se consideraron los indicadores jerarquía docente, razón docente/estudiante (`ratio_docentes`) y tasa de víctimas del conflicto armado (`tasa_victimas`), los cuales aportan evidencia sobre las condiciones estructurales de los entornos educativos en que

se desenvuelven los estudiantes.

- El promedio de jerarquía docente (PROMEDIO_JERARQUIA_DOCENTE) osciló entre 0.8 y 3.6, con valores superiores en zonas urbanas. Esta distribución sugiere que los docentes en municipios urbanos cuentan, en promedio, con mayor cualificación profesional o experiencia docente, lo cual se traduce en una mayor capacidad institucional para sostener procesos pedagógicos de calidad y generar entornos de aprendizaje más estables.

- La razón docente/estudiante (`ratio_docentes`) evidenció una media general de 0.050, lo que equivale aproximadamente a un docente por cada 19,8 estudiantes. Este indicador mostró diferencias por área: en las zonas rurales, el promedio fue 0.0588 (aprox. 17 estudiantes por docente), mientras que en las zonas urbanas se ubicó en 0.0489 (aprox. 20 estudiantes por docente). Estas variaciones reflejan la mayor dispersión geográfica y menor densidad de matrícula en las instituciones rurales, frente a una mayor concentración de estudiantes por docente en los contextos urbanos. Aunque el promedio nacional sugiere una proporción docente–estudiante relativamente equilibrada, la variabilidad por tamaño y ubicación institucional indica que los recursos humanos no se distribuyen de manera homogénea, lo cual puede incidir en la atención pedagógica y en los niveles de resiliencia escolar observados en los distintos territorios.

- La variable tasa_victimas, que representa la proporción de estudiantes con registro de víctimas del conflicto armado, presenta una distribución marcadamente asimétrica, concentrada en valores cercanos a cero ver Figura 3.6. Más del 80 % de los municipios se ubican por debajo de una tasa de 0,1 (10 %), lo que refleja que, aunque la mayoría de las instituciones educativas se localizan en territorios con baja afectación directa, existe un grupo minoritario de municipios donde la incidencia del conflicto es significativamente mayor.

El histograma evidencia esta concentración, con una larga cola hacia la derecha que representa a los municipios más afectados, mientras que el mapa de calor por deciles muestra una estructura altamente desigual: los tres primeros deciles (Q1–Q3) agrupan tasas promedio inferiores al 5 %, el cuarto decil se aproxima al 10 %, y el último decil (Q5) concentra las tasas más elevadas, con promedios superiores al 30 %.

Este comportamiento confirma la heterogeneidad territorial del impacto del conflicto armado, en la que un número reducido de municipios concentra la mayor proporción de víctimas.

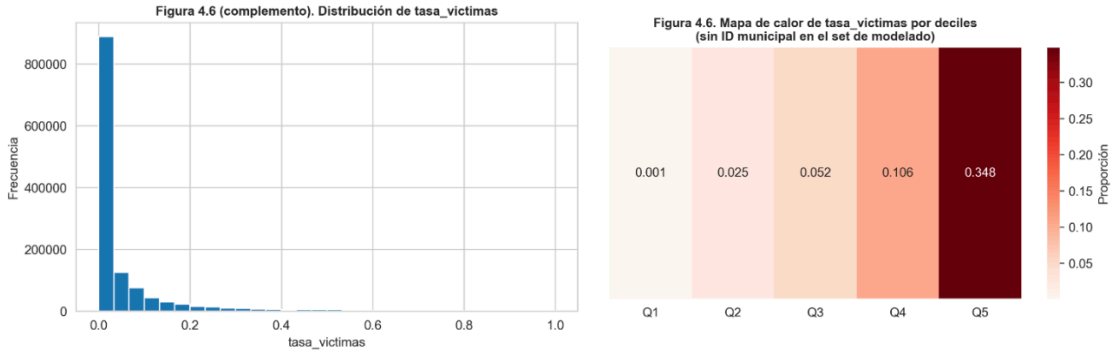

Figura 3.6: Tasa de víctimas del conflicto armado.

De forma complementaria, la variable COL_%aprobación, considerada como un indicador del rendimiento académico institucional, registró un promedio nacional de 87.4 %, con una desviación estándar de 9.9 puntos porcentuales. Las tasas de aprobación resultaron similares entre zonas rurales y urbanas, lo que sugiere una distribución relativamente homogénea del desempeño académico promedio a nivel nacional.

No obstante, los valores mínimos observados (inferiores al 3 % en algunos casos) se asocian a instituciones con mayores niveles de vulnerabilidad social y educativa, lo que apunta a una posible relación entre baja aprobación académica, inequidad territorial y menor resiliencia escolar en contextos de alta exclusión o afectación por el conflicto armado.

3.5.4. Correlaciones entre variables

El análisis de correlaciones entre los principales atributos numéricos permitió identificar las relaciones internas del conjunto de datos y evaluar la pertinencia de cada variable para el modelado predictivo (ver Figura 3.7).

Las variables asociadas al desempeño académico percentiles global, matemáticas, lectura crítica y ciencias naturales mostraron correlaciones muy altas entre sí (r >0.75), lo cual era esperable, dado que el percentil global se calcula a partir de los resultados obtenidos en dichas áreas. Asimismo, la variable objetivo clase_resiliencia, definida con base en el percentil global, presentó correlaciones elevadas con estos indicadores, lo que confirmaba la existencia de una dependencia directa y redundante entre ellos.

Por esta razón, y con el fin de evitar fuga de información (data leakage) y redundancia en las representaciones, se decidió conservar únicamente el percentil global como medida representativa del desempeño académico, eliminando los percentiles por área en la versión final del dataset. De esta manera, se aseguró que las variables explicativas del modelo midieran factores contextuales, institucionales y familiares independientes del criterio usado para definir la resiliencia escolar.

En contraste, las variables socioeconómicas (estu_inse_individual, estu_nse_establecimiento) mostraron correlaciones moderadas con el desempeño académico (r entre 0.35 y 0.45), mientras que

las variables institucionales como el promedio de jerarquía docente, el índice de infraestructura TIC y el porcentaje de aprobación institucional (COL_%aprobación) evidenciaron correlaciones bajas (r <0.25) con la variable objetivo, indicando efectos indirectos sobre la resiliencia.

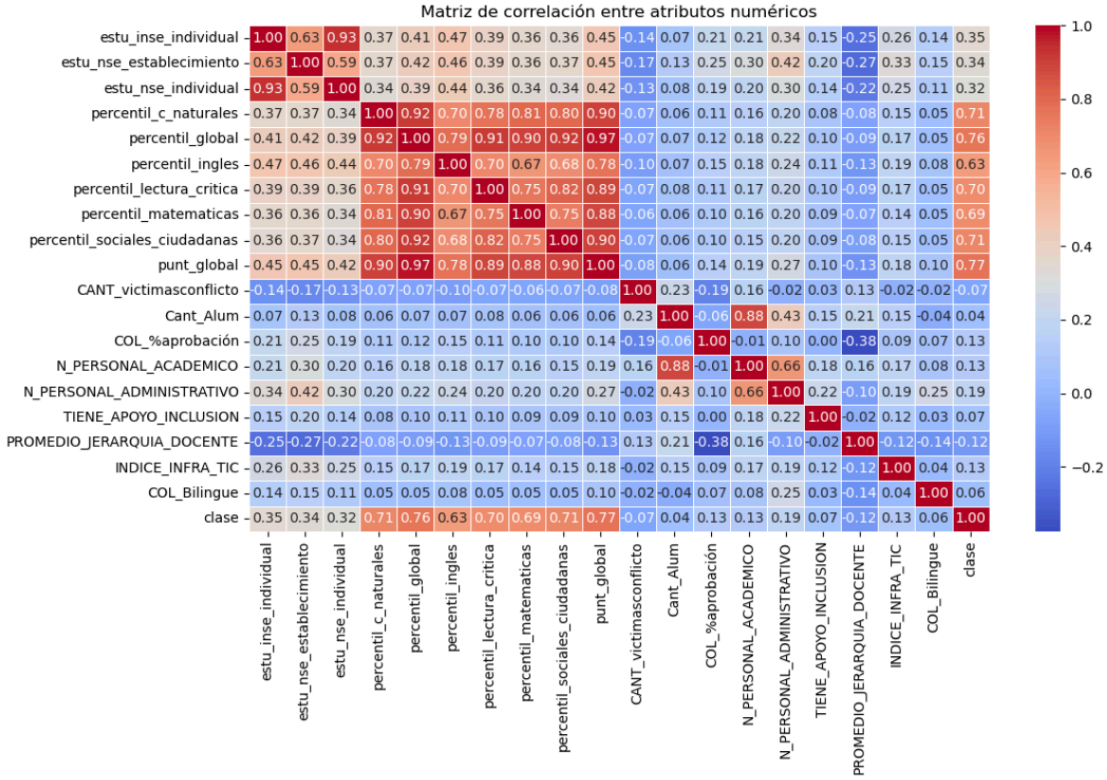

Figura 3.7: Mapa de calor de correlaciones entre variables numéricas.

3.5.5. Análisis estadístico integrado

Como resultado del procesamiento y análisis del conjunto de datos consolidado, se obtuvo un dataset de prueba representativo conteniendo 321,441 registros procesados que permitieron generar un análisis estadístico integral de la resiliencia escolar a nivel nacional.

El análisis automatizado del conjunto de prueba reveló patrones significativos en la distribución territorial y institucional de la resiliencia educativa:

- **Resiliencia nacional:** De los 321,441 estudiantes evaluados, 83,276 fueron clasificados como resilientes, representando un 25.9 % del total. Este resultado es consistente con la definición de resiliencia basada en el percentil 70 y confirma la naturaleza selectiva del fenómeno.

- **Cobertura territorial:** Se identificaron 355 municipios con al menos una sede educativa resiliente de un total de 1,121 municipios evaluados, equivalente al 31.67 % de cobertura territorial. Esta distribución evidencia la concentración geográfica del fenómeno de resiliencia educativa.

- **Municipios con mayor resiliencia:** Cartagena de Indias lidera con 18,216 estudiantes resilientes (38.2 % de tasa local), seguido por municipios del departamento de Bolívar como Achí (4,388 resilientes, 27.1 %) y El Carmen de Bolívar (3,059 resilientes, 23.8 %). Esta concentración sugiere factores institucionales y territoriales específicos que favorecen la resiliencia.

- **Identificación de brechas:** El análisis permite la identificación sistemática de municipios con menor resiliencia educativa, facilitando la focalización de intervenciones y políticas educativas diferenciadas según el contexto territorial.

- **Análisis por sede educativa:** A nivel institucional, se identificaron las sedes con mejor desempeño en términos de estudiantes resilientes, proporcionando insumos para el análisis de mejores prácticas educativas y factores institucionales exitosos.

Este análisis estadístico constituye la base empírica para la generación de visualizaciones geoespaciales y representa un insumo fundamental para la toma de decisiones en política educativa. La metodología desarrollada permite replicar estos análisis con cualquier conjunto de datos que mantenga la estructura de 62 variables predictoras, facilitando estudios comparativos temporales y evaluaciones de impacto de intervenciones educativas específicas.

Los resultados obtenidos confirman la hipótesis de desigualdad territorial en la resiliencia educativa y proporcionan evidencia cuantitativa para orientar estrategias de mejoramiento educativo focalizadas en los contextos de mayor vulnerabilidad identificados.

3.5.6. Síntesis interpretativa del EDA

El análisis exploratorio de datos (EDA) permitió caracterizar la integridad, coherencia y representatividad del conjunto de datos consolidado, proporcionando una visión descriptiva y diagnóstica del fenómeno de la resiliencia escolar en el contexto colombiano.

Esta fase aportó una comprensión integral del comportamiento de las variables, al identificar patrones recurrentes, relaciones generales entre dimensiones académicas, socioeconómicas e institucionales, así como disparidades estructurales asociadas a los contextos territoriales.

Entre los hallazgos más relevantes se destacan:

- Balance temporal y territorial adecuado para el análisis multianual (2021–2023), garantizando comparabilidad entre cohortes y regiones.

- Desbalance moderado en la variable objetivo (clase_resiliencia), con una proporción aproximada del 26 % de estudiantes resilientes y 76 % no resilientes, característica inherente al fenómeno observado.
- Asociaciones descriptivas consistentes entre resiliencia escolar, capital familiar y condiciones institucionales, lo que evidencia la relevancia de los factores socioeconómicos y organizativos en el desempeño académico en contextos adversos.
- Persistencia de desigualdades regionales, particularmente en territorios históricamente afectados por el conflicto armado, reflejando la influencia del entorno territorial sobre los resultados educativos.
- Coherencia en las distribuciones y relaciones internas del conjunto de datos, lo que permitió identificar redundancias y variables altamente correlacionadas para su posterior tratamiento metodológico.

En conjunto, el EDA consolidó la comprensión empírica del fenómeno y aportó insumos fundamentales para la definición del diseño experimental y las estrategias de modelado, que se desarrollan en el capítulo siguiente, sin introducir ajustes paramétricos ni decisiones predictivas en esta etapa.

Capítulo 4

Desarrollo de Modelos Predictivos para la Resiliencia Escolar

---

Este capítulo describe el proceso de construcción, entrenamiento y evaluación de los modelos predictivos, a partir del conjunto de variables familiares, institucionales y territoriales integradas y depuradas en el Capítulo 3. A diferencia de enfoques centrados en la estimación de puntajes o probabilidades continuas, el presente estudio adopta una perspectiva binaria, en la que cada estudiante es asociado a una categoría: Resiliente, No Resiliente.

La transición hacia el modelado supervisado se fundamenta en los hallazgos del Análisis Exploratorio de Datos (EDA), el cual evidenció relaciones significativas entre la resiliencia y dimensiones como el capital familiar, el acceso a recursos educativos, las condiciones institucionales y el contexto territorial del conflicto armado. Asimismo, el procesamiento y depuración permitió verificar la calidad estadística del dataset final, garantizar la ausencia de fugas de información y comprender la estructura del desbalance en la variable objetivo, elementos críticos para la selección y evaluación de los modelos de clasificación.

La organización del capítulo sigue la lógica del proceso CRISP–DM:

- Primero, se justifica la selección de las técnicas de aprendizaje automático empleadas.
- Se describe el diseño experimental, la división del conjunto de datos y las métricas utilizadas.
- Posteriormente se presentan los modelos base y sus resultados iniciales.
- Se realiza un proceso sistemático de optimización de hiperparámetros.
- Finalmente, se evalúan los modelos optimizados para seleccionar el modelo con mejor desempeño y estabilidad.

El propósito final es identificar el modelo que ofrece la mejor capacidad para distinguir entre estudiantes resilientes y no resilientes con base en factores multivariados, proporcionando una herramienta predictiva que contribuya a la toma de decisiones educativas, la focalización territorial y la comprensión de los determinantes de la resiliencia escolar en contextos afectados por el conflicto armado.

4.1. Selección de Técnicas de Aprendizaje Automático

La selección de las técnicas de aprendizaje automático constituye un paso crucial dentro del proceso de modelado, pues determina la capacidad del sistema para capturar patrones significativos y clasificar adecuadamente a los estudiantes como resilientes o no resilientes. La elección de los modelos responde, además de los requisitos metodológicos establecidos para garantizar interpretabilidad, desempeño y estabilidad en el proceso de clasificación, así como a una evaluación cuidadosa del problema, de las transformaciones aplicadas y de las condiciones operativas del entorno de desarrollo.

4.1.1. Justificación de los Métodos Elegidos

La predicción de la resiliencia escolar constituye un problema de clasificación binaria con variables heterogéneas (socioeconómicas, institucionales y de desempeño educativo) y un desbalance inherente en la clase positiva (resiliencia). En este contexto, la selección de técnicas se fundamentó en tres criterios principales:

1. **Capacidad para capturar relaciones no lineales y efectos de interacción**
   La resiliencia escolar no emerge de una relación lineal simple entre variables; por tanto, modelos como XGBoost y SVM permiten capturar patrones complejos y fronteras de decisión no lineales.

2. **Robustez estadística y estabilidad interpretativa**
   Métodos como Regresión Logística y MLP ofrecen un equilibrio entre interpretabilidad, rigor estadístico y capacidad para modelar estructuras más profundas en los datos.

3. **Capacidad para manejar el desbalance de clases y manejo de ruido estructural**
   Técnicas basadas en árboles, como XGBoost, incorporan parámetros nativos para manejar clases desbalanceadas y toleran mejor valores atípicos o relaciones no estrictamente lineales.

Con base en lo anterior, se seleccionaron cuatro familias de modelos representativas y complementarias:

Modelo	Justificación principal
Regresión Logística (LR)	Punto de partida estadístico, interpretable y estable. Actúa como línea base para contrastar la ganancia marginal de modelos más complejos.
Máquinas de Vectores de Soporte (SVM)	Especialmente útiles cuando las fronteras de decisión requieren transformaciones no lineales mediante kernels.
XGBoost	Modelo de referencia en problemas tabulares. Buen ajuste a relaciones no lineales, manejo del desbalance y alto desempeño empírico en datasets similares.
Perceptrón Multicapa (MLP)	Arquitectura neuronal capaz de aproximar funciones complejas conservando un nivel moderado de interpretabilidad estructural.

Cuadro 4.1: Modelos de aprendizaje automático seleccionados.

Cada uno de estos modelos aporta una perspectiva diferente sobre el fenómeno predictivo, enriqueciendo la comprensión del comportamiento de las variables explicativas frente a la resiliencia.

4.1.2. Criterios de Selección Metodológica

La elección de las técnicas de aprendizaje automático se sustentó en un conjunto de criterios metodológicos derivados de las características del conjunto de datos preparado en el Capítulo 3, los objetivos del proyecto y la naturaleza del problema de clasificación binaria. Estos criterios orientaron la selección de algoritmos que permitieran capturar patrones complejos en los datos, minimizar el impacto del desbalance de clases y garantizar un proceso de modelado reproducible, transparente y estadísticamente sólido. Los criterios considerados fueron los siguientes:

1. **Pertinencia para datos tabulares de alta dimensionalidad**

   El conjunto de datos integrado, después de combinar atributos numéricos normalizados, proporciones institucionales, codificaciones categóricas binarias y variables derivadas a nivel territorial, incluye un total de 62 variables.

   Dada su robustez frente a este tipo de estructuras, se priorizaron algoritmos ampliamente utilizados en problemas educativos y en ciencia de datos aplicada con bases multivariadas: Regresión Logística, SVM, XGBoost y Perceptrón Multicapa (MLP).

2. **Capacidad para modelar relaciones no lineales y efectos de interacción**

   El análisis exploratorio evidenció patrones complejos entre capital familiar, rasgos institucionales y variables territoriales. Modelos como XGBoost, SVM y MLP permiten capturar fronteras de decisión no lineales y dependencias cruzadas que la regresión logística por sí sola

no puede representar.
Este criterio fue clave para explorar modelos con diferentes capacidades expresivas y complementariedad teórica.

3. **Capacidad para manejar el desbalance de clases y la presencia de ruido estructural**

   La variable objetivo presenta una distribución aproximada de 76 % estudiantes no resilientes y 24 % resilientes, lo que exige técnicas capaces de reducir sesgos hacia la clase mayoritaria. Los modelos seleccionados permiten integrar estrategias como:

   - Undersampling
   - Ajuste de scale_pos_weight,
   - Ajuste de umbral de decisión,
   - Optimización por recall o F1,
   - Regularización en redes neuronales (dropout, batch normalization).

   Estas características favorecen modelos más sensibles a los casos resilientes, aspecto fundamental para los objetivos predictivos del estudio.

4. **Escalabilidad computacional y viabilidad de entrenamiento**

   Dadas las dimensiones del dataset final (más de 1,6 millones de estudiantes, 2021–2023), se priorizaron algoritmos que pudieran entrenarse de forma eficiente en hardware convencional:

   - LinearSVC resultó viable, SVC con kernels no, debido a su costo computacional.
   - XGBoost ofrece entrenamiento incremental y manejo nativo de grandes volúmenes de datos.
   - MLP permitió entrenamiento con estructuras neuronales moderadamente profundas.

   Este criterio permitió excluir o ajustar métodos con alto costo computacional (p. ej., SVC con kernel RBF).

5. **Transparencia, interpretabilidad y coherencia técnica con CRISP–DM**

   La metodología exige un equilibrio entre desempeño y explicabilidad.

   - Regresión Logística ofrece interpretabilidad base.
   - XGBoost permite analizar importancia de variables y SHAP values.
   - MLP y SVM, aunque menos interpretables, aportan capacidad predictiva superior.

   Este enfoque permitió equilibrar la interpretación del modelo con el rendimiento, favoreciendo el análisis comparativo.

6. **Compatibilidad con procesos de optimización sistemática**

   Los modelos seleccionados permiten optimizar hiperparámetros mediante ParameterGrid. Esta capacidad fue esencial para explorar:

   - Profundidad, tasa de aprendizaje y número de árboles en XGBoost.
   - Penalización C en SVM.
   - Arquitectura, dropout y tasa de aprendizaje en MLP.

   El desarrollo del proyecto requería modelos con suficiente flexibilidad para ajustarse al comportamiento del dataset sin riesgo de sobreajuste.

4.2. Diseño Experimental

El diseño experimental constituyó la estructura operativa del proceso de modelado, garantizando la correcta preparación del dataset, la ausencia de fuga de información (data leakage) y la comparabilidad entre los diferentes modelos evaluados. Esta sección describe las decisiones metodológicas adoptadas para la construcción de la estructura de datos, la división de conjuntos, el tratamiento del desbalance de clases, las métricas empleadas y los criterios de validación necesarios para asegurar la rigurosidad del enfoque metodológico.

4.2.1. Estructura de datos y Preparación del Conjunto Analítico

Con base en el Capítulo 3, se estableció una estructura que define las reglas sobre el tipo de variables, su escala, codificación, momento de aplicación de transformaciones y restricciones de uso en los conjuntos de entrenamiento, validación y prueba.

Elementos principales:

- **Número total de variables:** 62 características finales.
- **Tipos de variables incluidos:**
  - Numéricas estandarizadas (percentiles Saber 11, tasas institucionales, indicadores familiares).
  - Categóricas transformadas mediante One-Hot Encoding.
  - Variables derivadas territoriales mediante target encoding controlado para evitar fuga de información.
- **Variables no permitidas:**
  - Cualquier atributo derivado del resultado del estudiante o agregados del conjunto completo que puedan inducir leakage.

- **Reglas de preprocesamiento:**
  - El escalado (StandardScaler) se ajustó únicamente con el conjunto de entrenamiento.
  - La estructura de columnas generadas en One-Hot Encoding se fijó con el conjunto de entrenamiento para garantizar compatibilidad con validación y prueba.
  - No se desarrolla reingeniería posterior del dataset después de entrenar un modelo base.

Estas reglas proporcionaron coherencia y reproducibilidad en toda la fase de modelado.

4.2.2. Tratamiento del Desbalance de Clases

El desbalance original (76 % clase 0; 24 % clase 1) afectaba directamente la capacidad de los modelos para identificar estudiantes resilientes. Por ello, se pusieron a prueba múltiples estrategias complementarias:

1. **Undersampling**
   Se utilizaron tres esquemas principales:
   - **Balance 1:1**, reduciendo drásticamente la clase mayoritaria.
   - **Balance 60/40**, estrategia intermedia que conserva mayor representatividad de la clase 0 y ofrece un mejor equilibrio entre información y estabilidad.
2. **Ajustes específicos por modelo**
   - **XGBoost:** uso de `scale_pos_weight` y evaluación en datasets balanceados.
   - **SVM:** balanceo previo obligatorio debido a limitaciones computacionales.
   - **MLP:** uso selectivo de datasets 60/40 y control del sesgo mediante ajuste de umbral.
   - **Regresión Logística:** uso de `class_weight="balanced"` y entrenamiento sobre datasets balanceados.

4.2.3. Métricas de Evaluación

Dado el énfasis en detectar correctamente casos de resiliencia, las métricas principales fueron:

- **AUC-ROC:** Permite comparar modelos independientemente del umbral de decisión.
- **Recall (Sensibilidad) – Clase 1:** Fundamental para minimizar falsos negativos (estudiantes resilientes mal clasificados).
- **Precision – Clase 1:** Relevante en escenarios donde el costo de falsos positivos es significativo.
- **F1-score:** Métrica balanceada entre sensibilidad y precisión.
- **Accuracy:** Utilizada solo como métrica secundaria dado el desbalance.

Matriz de confusión En cada modelo se analizó la distribución de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN) para comprender:

- Riesgo de sobrealertas (FP).
- Riesgo de omitir resiliencia (FN).
- Exactitud operativa del modelo.

4.2.4. Procedimiento de Validación Interna

Para garantizar la consistencia en el proceso de selección de modelos, se aplicó un enfoque de validación por etapas:

1. Entrenamiento con dataset balanceado o ponderado según el algoritmo.
2. Evaluación en el conjunto de validación no balanceado, simulando condiciones reales.
3. Selección del modelo con mejor combinación de AUC, Recall y F1.
4. Ajuste del umbral de decisión, especialmente crítico en LR, XGBoost y MLP.
5. Reentrenamiento final del modelo con todos los datos disponibles del conjunto de entrenamiento.
6. Evaluación independiente en el conjunto de prueba como medida definitiva.

4.2.5. Visualizaciones en el Diseño Experimental

Para apoyar la selección de modelos, se generaron:

- Curvas ROC comparando distintos modelos y variantes.
- Gráficos de barrido de umbral mostrando cómo cambia el recall, precisión y F1.
- Matrices de confusión para cada modelo optimizado.

Estas visualizaciones serán integradas en las secciones 5.3, 5.4 y 5.5 según corresponda.

4.3. Modelos Base

La fase de modelos base constituye un momento clave dentro del proceso de modelado, pues permite entender cómo se comporta cada técnica antes de aplicar balanceo, optimización y/o ajustes. En este apartado se presentan los resultados iniciales de los cuatro modelos evaluados: Regresión Logística, XGBoost, SVM y Perceptrón Multicapa (MLP), con un análisis detallado de sus patrones de error, sensibilidad, capacidad discriminativa y estabilidad ante el desbalance de clases.

4.3.1. Modelo Base: Regresión Logística

La Regresión Logística fue el primer modelo evaluado debido a su interpretabilidad, estabilidad y rapidez de entrenamiento. Aunque es un modelo lineal, constituye un punto de referencia para comprender la estructura del fenómeno antes de recurrir a algoritmos más complejos.

Los resultados iniciales mostraron un equilibrio adecuado entre las principales métricas de desempeño:

- AUC ≈ 0.75
- Recall ≈ 0.62
- Precisión ≈ 0.59
- F1 ≈ 0.61
- Accuracy ≈ 0.79

La regresión logística es capaz de capturar patrones lineales en el conjunto de datos, pero presenta limitaciones para modelar interacciones más complejas entre las variables familiares, institucionales y territoriales. Además, tiende a ser conservadora al clasificar la clase resiliente, lo que ayuda a reducir los falsos positivos, pero a costa de una menor detección de casos. La matriz de confusión revela un número moderado de falsos negativos, lo que indica que este modelo "pierde" una proporción significativa de estudiantes resilientes.

4.3.2. Modelo Base: XGBoost

El modelo XGBoost constituyó uno de los enfoques centrales de este proyecto debido a su capacidad para modelar relaciones no lineales, manejar datos tabulares y ofrecer interpretabilidad mediante la estimación de importancia de variables. Esta sección expone el proceso de desarrollo, evaluación y análisis del modelo base antes de la optimización, integrando evidencia numérica y visual.

A. Desarrollo del modelo base XGBoost

La primera aproximación consistió en entrenar un modelo sin balanceo de clases y con hiperparámetros por defecto.

Para evitar fuga de datos y garantizar un flujo reproducible, se definió un Pipeline con preprocesamiento mediante:

- `OneHotEncoder` para variables categóricas.
- Las variables numéricas se integran al modelo sin aplicar transformaciones.
- `XGBClassifier` como estimador final.

Este pipeline fue posteriormente envuelto en un `RandomizedSearchCV` para futuras iteraciones, pero en el modelo base se evaluó la configuración estándar.

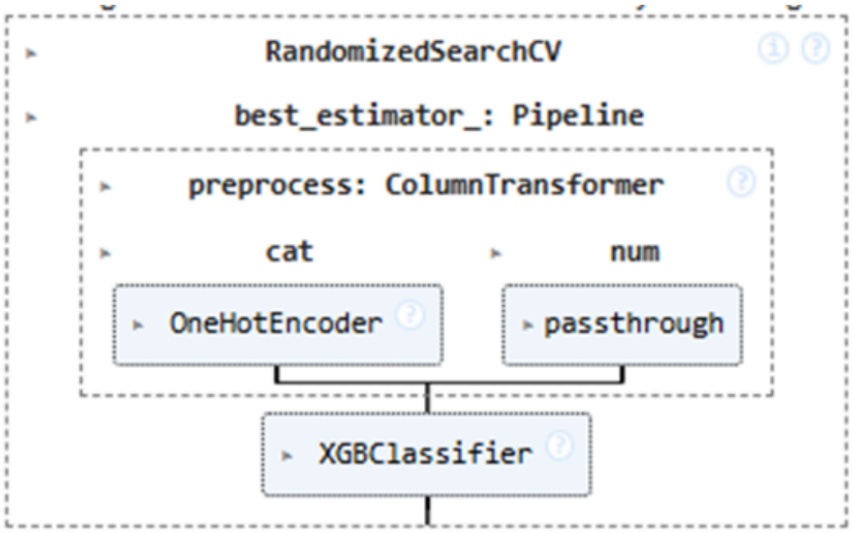

Figura 4.1: Estructura del Pipeline con RandomizedSearchCV.

Los resultados del modelo XGBoost sin técnicas de balanceo mostraron un comportamiento característico:

- Accuracy: 0.80
- AUC: 0.832
- Recall (clase resiliente): 0.45
- Precisión clase resiliente: 0.69
- F1-score clase resiliente: 0.54

Los indicadores muestran un modelo conservador, que clasifica correctamente la clase mayoritaria (no resilientes), pero deja sin identificar una proporción importante de estudiantes resilientes (alto número de falsos negativos).

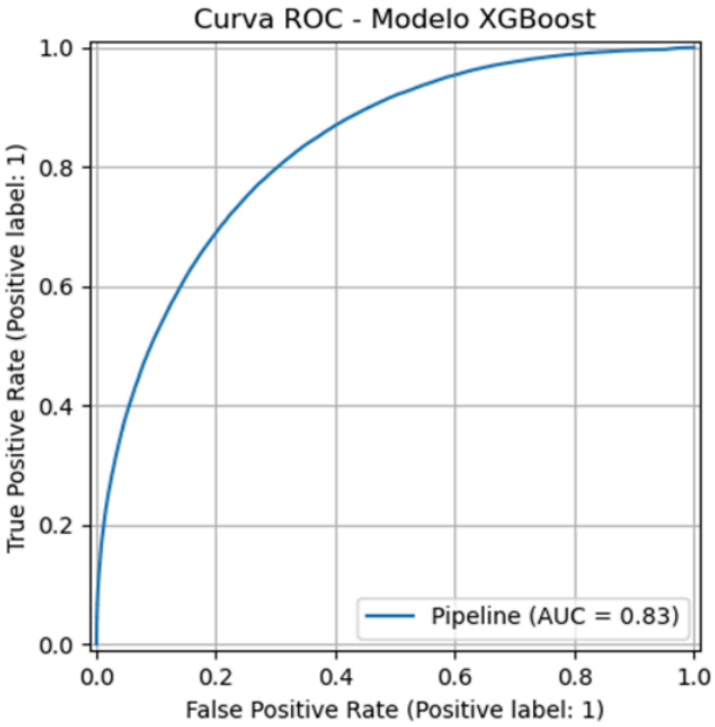

Figura 4.2: Curva ROC - Modelo XGBoost Base.

La curva ROC del modelo base se aproxima al vértice superior izquierdo, pero sin alcanzarlo de forma consistente.

El AUC de 0.83 indica buena capacidad discriminativa general, aunque no se traduce en un desempeño adecuado en recall.

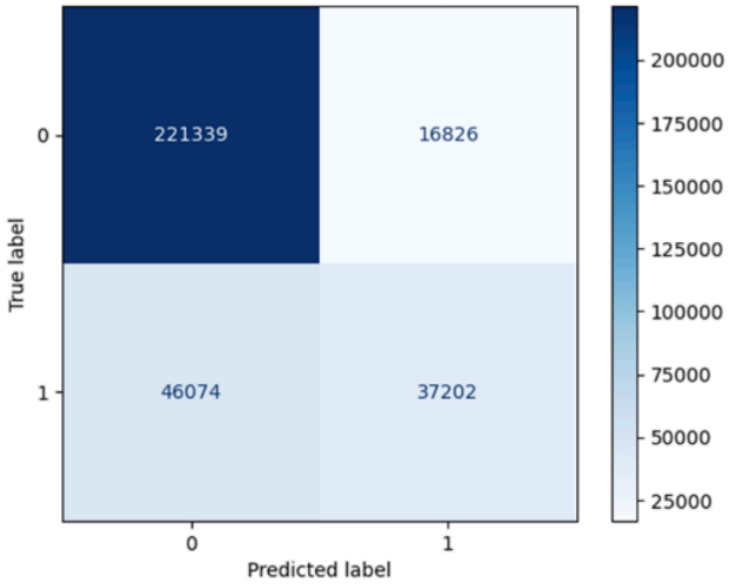

Figura 4.3: Matriz de Confusión

La matriz evidencia:

- Muy alta tasa de verdaderos negativos
- Bajos verdaderos positivos
- Muchos falsos negativos

La matriz de confusión del modelo base evidencia una clasificación correcta para la mayor parte de la clase 0 (no resilientes), pero poca sensibilidad hacia la clase minoritaria.

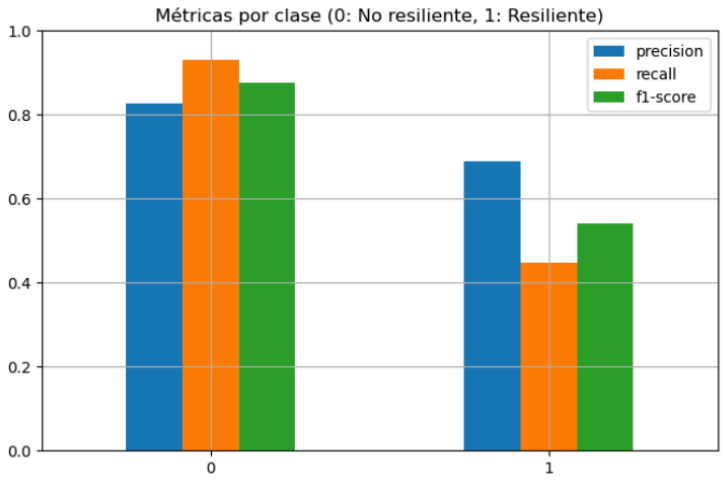

Figura 4.4: Métricas por Clase - Modelo XGBoost Base.

El gráfico de métricas por clase evidencia la asimetría:

- Clase 0 (no resiliente): precisión y recall altos
- Clase 1 (resiliente): recall bajo, F1 moderado

La precisión de clase 1 (0.69) demuestra que los casos etiquetados como resilientes suelen serlo, pero esto ocurre a costa de no identificar a la mayoría.

Lo cual evidencia que debido al desbalance del dataset, la versión sin balanceo resulta insuficiente cuando el objetivo es no perder estudiantes resilientes, aunque es un punto de partida válido.

B. Modelo XGBoost con ajuste de `scale_pos_weight`

Considerando la proporción entre clases, se ajustó el hiperparámetro `scale_pos_weight = 2.88` como una aproximación proporcional al ratio 76/24. Esta variante modificó el énfasis del modelo para penalizar más fuertemente los errores asociados a la clase resiliente.

Resultados:

- Accuracy: 0.74
- AUC: 0.832
- Recall (resiliente): 0.7
- Precisión: 0.50
- F1-score: 0.61

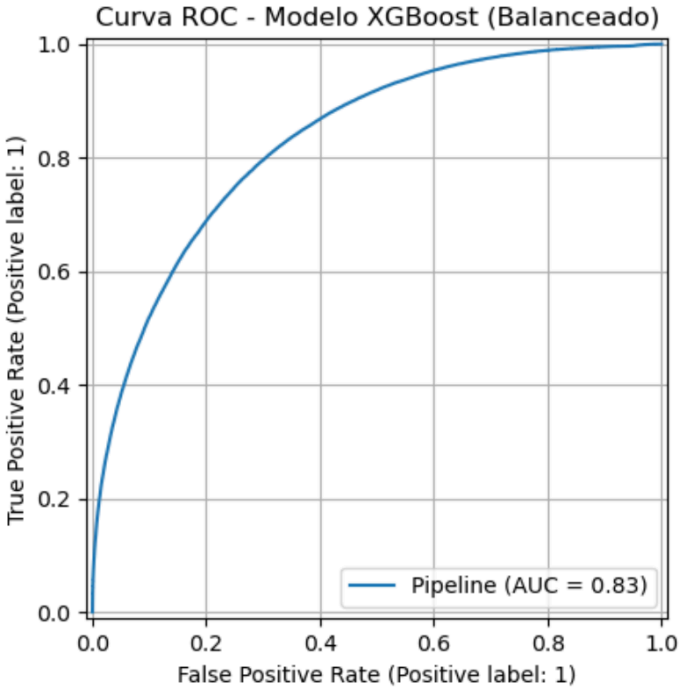

Figura 4.5: Curva ROC - XGBoost Balanceado.

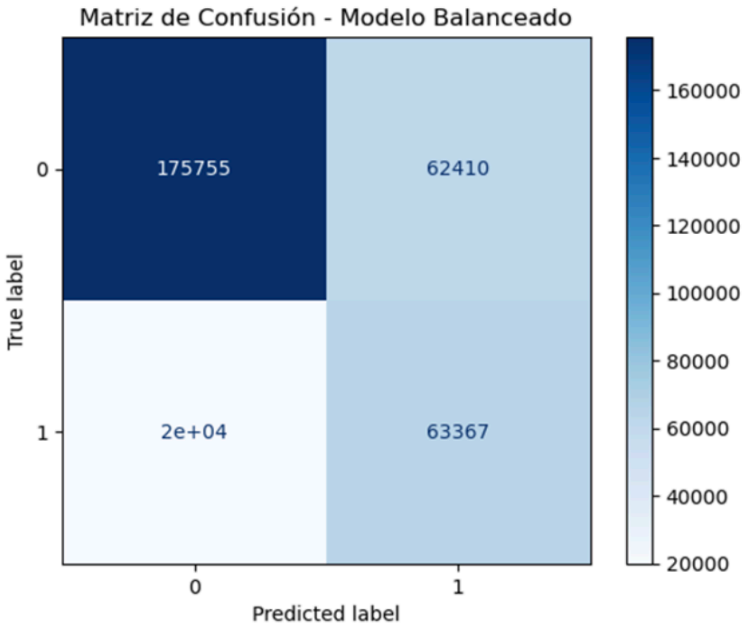

Figura 4.6: Matriz de Confusión - Modelo Balanceado.

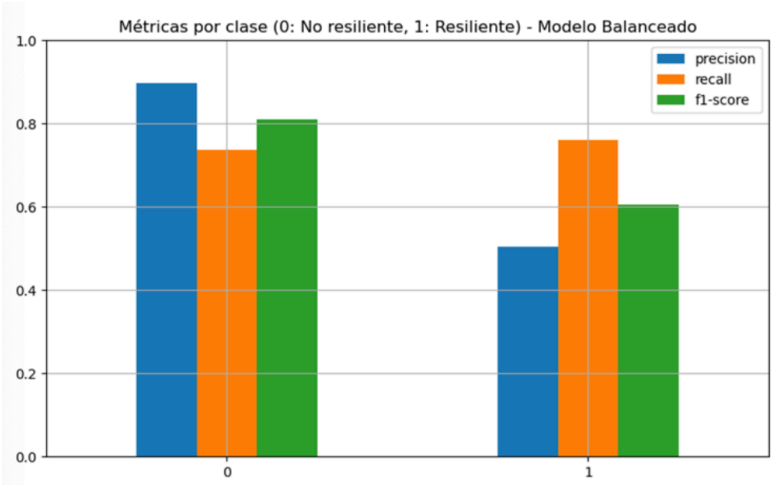

Figura 4.7: Métricas por Clase - XGBoost Balanceado.

El comportamiento del modelo con `scale_pos_weight` muestra:

- Incremento sustancial en recall, pasando de 0.45 en el modelo base a 0.7. Esto se traduce en un aumento significativo en detección de resiliencia: de aproximadamente 38.576 casos identificados en la versión base a 63.367 en esta variante.

- La precisión disminuye (0.50), lo cual es esperable dado el cambio en la función de pérdida que favorece la clase minoritaria.
- El F1-score se convierte en el más alto entre las variantes evaluadas (0.61), indicando un mejor equilibrio entre sensibilidad y precisión.
- La curva ROC mantiene prácticamente el mismo AUC que el modelo base (0.832), pero la parte superior izquierda muestra un mejor comportamiento, reflejando mayor TPR (True Positive Rate) para el mismo nivel de FPR.

El modelo con `scale_pos_weight` es una variante efectiva para la detección de resiliencia y la que mejor se alinea con el propósito del estudio: minimizar la pérdida de estudiantes resilientes.

C. Modelo XGBoost con Balanceo Undersampling 60/40

Posterior a los avances obtenidos con el ajuste del hiperparámetro `scale_pos_weight`, el comportamiento del modelo XGBoost reveló importantes mejoras en sensibilidad (recall) para la clase resiliente. Sin embargo, el análisis previo mostró dos retos clave:

1. **Persistencia de patrones dominados por la clase mayoritaria:** Aunque `scale_pos_weight` aumentó el recall, el modelo seguía generando probabilidades concentradas en rangos medios, y la frontera no se desplazaba lo suficiente como para capturar patrones minoritarios más complejos.
2. **Necesidad de calibración y estabilidad:** Las predicciones mostraban variabilidad dependiendo de la composición del conjunto de validación. El objetivo de este ciclo fue estabilizar la frontera de decisión y obtener predicciones más confiables, especialmente para intervenir académicamente estudiantes resilientes.

Por esta razón, el balanceo 60/40 se eligió como una estrategia intermedia entre un equilibrio total (1:1) y un modelo completamente desbalanceado, permitiendo mejorar sensibilidad sin perder demasiada información estructural de la clase 0.

Este enfoque buscó reducir de manera controlada la clase mayoritaria (no resilientes), conservando la mayor cantidad de información relevante sin depender exclusivamente del ponderado de clases (como ocurre con `scale_pos_weight`). El objetivo fue construir una frontera de decisión más estable y calibrada, capaz de mejorar tanto la detección de resiliencia como la coherencia de las predicciones.

Configuración del conjunto balanceado

El proceso de undersampling produjo un dataset con la siguiente distribución:

- 60 % no resilientes

- 40 % resilientes

Con un total de 829,015 registros, esta proporción permitió entrenar un modelo menos sesgado, mejorando su capacidad de explorar zonas intermedias del espacio de características donde suele ubicarse la resiliencia.

Resultados del modelo

Las pruebas realizadas (48 combinaciones en la grilla de hiperparámetros) mostraron un comportamiento más equilibrado del modelo, con mejoras importantes en estabilidad y desempeño global.

Principales métricas obtenidas:
- AUC: 0.8485
- Recall clase resiliente: 0.6720
- Precisión: 0.5781
- F1-score: 0.6215

1. El AUC aumentó de 0.832 (modelo base) a 0.8485, señalando una mejor capacidad discriminativa global.
2. El F1-score es superior al del modelo base y comparable al del modelo con `scale_pos_weight`.
3. Aunque el recall es menor que el obtenido con `scale_pos_weight` (0.7 a 0.672), el modelo balanceado 60/40 ofrece mayor coherencia y estabilidad.

Evaluación de métricas en función del umbral

Aunque XGBoost produce valores logísticos que pueden interpretarse como puntajes de probabilidad, estos no corresponden a probabilidades calibradas. En este estudio no se aplicaron técnicas de calibración probabilística como Platt scaling o isotonic regression. Por ello, el análisis de umbrales se basó en los puntajes de salida del modelo, los cuales permiten clasificar las observaciones según la fuerza del patrón asociado a la clase positiva, pero no deben entenderse como probabilidades calibradas. En consecuencia, los resultados del barrido de umbral se interpretan como variaciones relativas en el comportamiento del modelo y no como decisiones sustentadas en probabilidades absolutas.

Hallazgos clave del barrido de umbral:

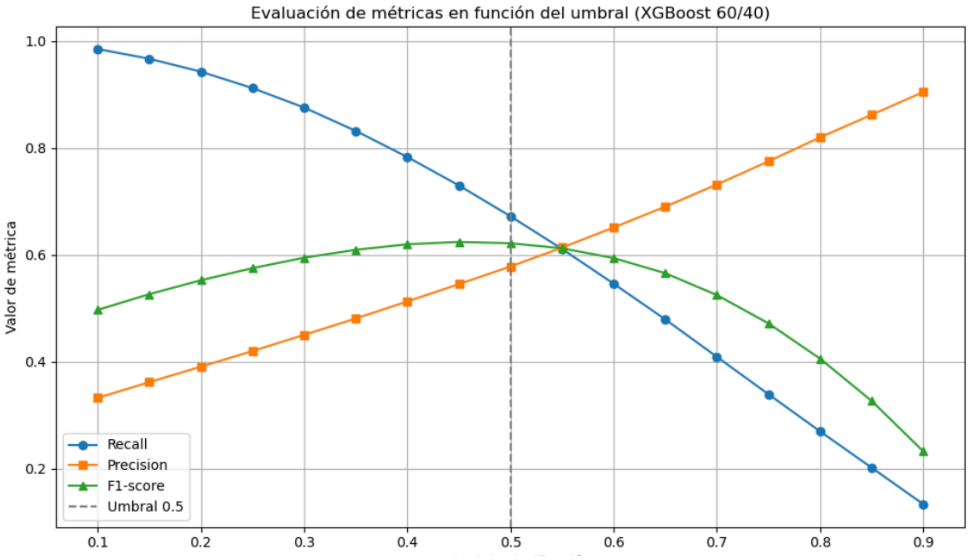

Figura 4.8: Evaluación de métricas según el umbral – XGBoost 60/40

El recall disminuye progresivamente a medida que el umbral aumenta: Con umbrales bajos, el modelo clasifica una mayor proporción de observaciones como positivas, lo que incrementa la sensibilidad. Por ejemplo:

- Umbral 0.10 ; recall $\approx$ 0,98
- Umbral 0.40 ; recall $\approx$ 0,78
- Umbral 0.50 ; recall $\approx$ 0,67
- Umbral 0.80 ; recall $\approx$ 0,30

Este comportamiento refleja la relación natural entre sensibilidad y especificidad.

La precisión aumenta de manera continua conforme el umbral se hace más estricto: A medida que el modelo exige mayor evidencia para clasificar un caso como positivo, disminuyen los falsos positivos, mejorando la precisión:

- Umbral 0.10 ; precisión $\approx$ 0,33
- Umbral 0.40 ; precisión $\approx$ 0,51
- Umbral 0.50 ; precisión $\approx$ 0,58
- Umbral 0.80 ; precisión $\approx$ 0,85

El F1-score alcanza su máximo alrededor del umbral 0.45–0.50.

F1 máximo 0.624

Este resultado evidencia que el umbral predeterminado de 0.50 no necesariamente coincide con el punto óptimo para el equilibrio entre recall y precisión.

Interpretación educativa del análisis de umbral

La selección del umbral en modelos educativos no es una decisión técnica, sino estratégica, pues implica:

- identificar estudiantes para intervención,
- asignar recursos institucionales,
- diseñar apoyos diferenciales.

Por tanto, el análisis de umbral permite adaptar el modelo a distintos contextos:

1. Si la prioridad es no perder resilientes (alta sensibilidad): Es coherente usar umbral 0.30–0.40, para alcanzar Recall aprox. 0.87–0.78.
2. Si se requiere mayor precisión (menos falsos positivos): Sería conveniente subir el umbral a 0.55–0.60.
3. Si se busca equilibrio para decisiones generales: Umbral óptimo 0.45. Máximo F1, Recall 0.70, Precisión 0.55

En conclusión, el modelo XGBoost con balanceo 60/40 mejora la estabilidad del aprendizaje, ofrece una frontera de decisión más coherente que la versión ajustada únicamente con `scale_pos_weight`, permite realizar un análisis más fino del desempeño mediante el ajuste del umbral de clasificación y se posiciona como una alternativa sólida en escenarios donde el equilibrio entre *recall* y precisión resulta más relevante que maximizar exclusivamente la sensibilidad. En conjunto, esta variante complementa de manera adecuada al XGBoost potenciado por `scale_pos_weight` y amplía la capacidad del sistema para diseñar estrategias de intervención más matizadas, calibradas y alineadas con las necesidades institucionales.

D. Importancia de Variables en el Modelo XGBoost

La interpretación de la importancia de variables constituye un componente crucial del análisis, pues permite comprender cómo el modelo construye la clasificación entre estudiantes resilientes y no resilientes. XGBoost ofrece esta lectura a través de la ganancia promedio obtenida por cada variable al participar en divisiones de los árboles, permitiendo identificar los factores con mayor

contribución a la predicción.

El análisis se basa en el ranking de las 20 variables más importantes obtenidas del modelo, representado en la siguiente figura:

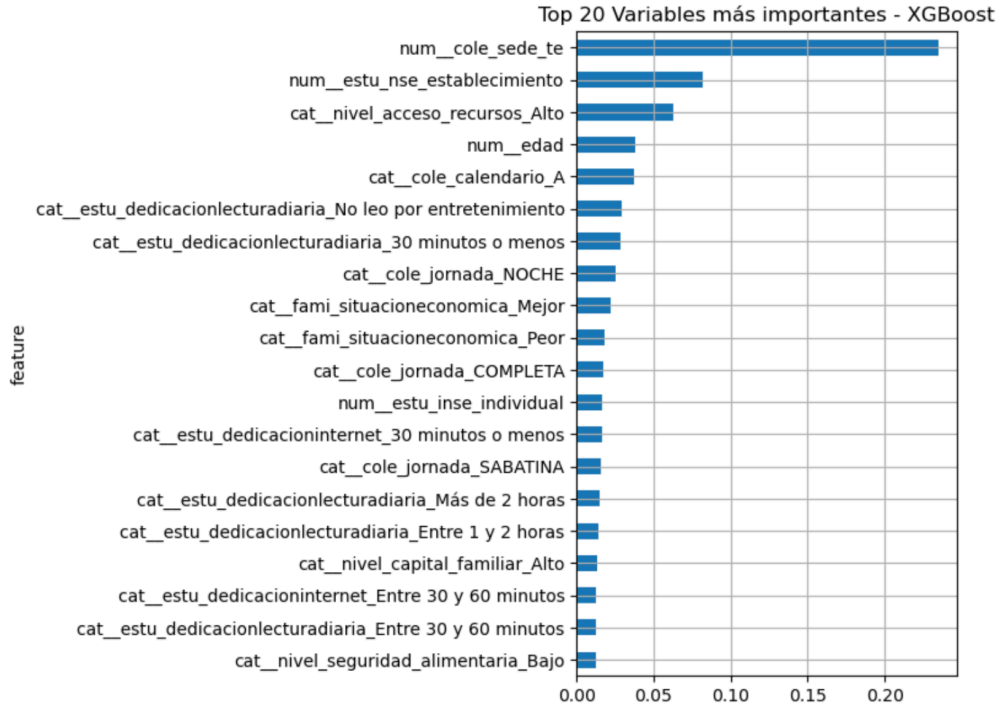

Figura 4.9: Top 20 Variables Importantes – XGBoost

El modelo XGBoost identifica una combinación de factores institucionales, familiares y estudiantiles como los más relevantes para predecir resiliencia académica. Esto evidencia que la resiliencia no depende únicamente de características individuales, sino que se configura como un fenómeno multidimensional, alineado con los planteamientos teóricos del proyecto.

Entre las variables de mayor peso destacan:

- **num__cole__sede__te:** Indicador institucional clave: refleja la magnitud, complejidad organizativa y diversidad de sedes de la institución educativa.
  Las sedes tienden a diferir en calidad, recursos y clima escolar.
  El modelo sugiere que la resiliencia es sensible a estas variaciones intra institución.

- **num__estu__nse__establecimiento:** Un proxy del nivel socioeconómico agregado del colegio, asociado a:
  - Mayor acceso a recursos,

- Mayor estabilidad institucional,
- Menor vulnerabilidad estructural.
Su alta influencia indica que el contexto colectivo pesa fuertemente sobre la resiliencia individual.

- **cat_nivel_acceso_recursos_Alto:** Una de las variables más explicativas: mayor acceso a materiales educativos, tecnología, biblioteca y conectividad.
Reafirma que la disponibilidad de recursos pedagógicos es un componente crítico para sostener trayectorias educativas resilientes.

- **num_edad:** Refleja aspectos de madurez, reprobación y trayectoria escolar. Edades mayores pueden indicar interrupciones educativas; edades regulares suelen asociarse a mayor continuidad escolar.
Su relevancia concuerda con estudios sobre rezago escolar y resiliencia.

- **cat_cole_calendario_A:** Indicador institucional con carga estructural significativa:
- Los colegios de calendario A tienden a ubicarse en contextos más vulnerables.
- Los de calendario B suelen tener mayor inversión y recursos.
El modelo detecta esta polaridad.

Factores relacionados con hábitos de estudio y uso del tiempo

El grupo de variables asociadas a dedicación lectora e internet aparece de manera consistente en la jerarquización:

- cat_estu_dedicacionlecturadiaria_No leo por entretenimiento.
- cat_estu_dedicacionlecturadiaria_30 minutos o menos.
- cat_estu_dedicacionlecturadiaria_Entre 1 y 2 horas.
- cat_estu_dedicacioninternent_30 minutos o menos.
- cat_estu_dedicacioninternet_Entre 30 y 60 minutos.

Estas variables indican que la resiliencia está asociada a:

- Patrones de uso del tiempo.
- Hábitos de lectura,
- Regularidad en las actividades académicas y recreativas,
- Equilibrio entre tecnologías e involucramiento escolar.

No se observa una relación lineal simple (más horas $\neq$ más resiliencia), sino un patrón complejo que combina intensidad, intención y regulación del tiempo.

Factores familiares: capital y situación económica

El modelo también identifica como relevantes:

- cat_fami_situacioneconomica_Mejor
- cat_fami_situacioneconomica_Peor
- cat_nivel_capital_familiar_Alto
- cat_nivel_seguridad_alimentaria_Bajo

La resiliencia se articula con dos dimensiones clave:

- **Capital familiar:** acceso a educación de adultos, nivel de escolaridad, estabilidad laboral, capital social.
- **Situación económica:** Influye no solo por acceso a recursos, sino por estabilidad emocional, expectativas familiares y acompañamiento en la trayectoria educativa.

Estas variables no solo tienen impacto directo, sino que interactúan con factores escolares y personales, justificando su presencia entre las primeras posiciones.

Variables relacionadas con la jornada escolar y el entorno institucional

Variables como:

- cat_cole_jornada_NOCHE
- cat_cole_jornada_SABATINA
- cat_cole_jornada_COMPLETA

Muestran que las modalidades de jornada influyen significativamente. La jornada escolar determina:

- Disponibilidad de tiempo libre para estudio.
- Acceso a alimentación escolar.
- Interacción con pares.
- Exposición a riesgos comunitarios.

Los colegios nocturnos y sabatinos suelen estar vinculados a poblaciones más vulnerables o estudiantes con responsabilidades adicionales (trabajo, cuidado del hogar), lo que explica su peso en la predicción.

4.3.3. Modelo Base: SVM

El uso de Máquinas de Vectores de Soporte (SVM) en este estudio se planteó como una alternativa valiosa para explorar fronteras de decisión más definidas que las de los modelos lineales tradicionales, especialmente considerando que la resiliencia escolar es un fenómeno influenciado por múltiples factores socioeducativos que pueden interactuar de manera no lineal. Sin embargo, la implementación práctica de esta técnica se vio profundamente condicionada por limitaciones computacionales, restricciones de memoria, tiempos de ejecución superiores a 24 horas y fallos recurrentes de compilación, lo que obligó a replantear el alcance del modelado con SVM.

En consecuencia, gran parte del trabajo con SVM tuvo que ser realizado sobre submuestras reducidas, y únicamente con la versión lineal de la técnica, lo que condicionó la profundidad del análisis y la magnitud de conclusiones posibles.

Los experimentos se ejecutaron en Google Colaboratory pro y en un equipo local de características equivalentes, con las siguientes especificaciones generales:

- Procesador: CPU multinúcleo
- Memoria RAM: entre 25 GB
- Aceleración: GPU disponible (NVIDIA Tesla T4).
- Sistema operativo: entorno Colab
- Lenguaje: Python 3.12.12
- Tiempo máximo de procesamiento 24 horas

A. Restricciones computacionales y su impacto metodológico

El conjunto de datos de esta investigación superior al millón de registros en su versión completa representa un desafío extremo para algoritmos como SVM, especialmente en sus variantes no lineales. Tal como se evidenció durante la experimentación:

1. Imposibilidad de ejecutar SVM con balanceo 60/40

Aunque el balanceo 60/40 fue exitosamente aplicado en XGBoost y MLP, en SVM: el entrenamiento con aproximadamente 830.000 muestras no pudo ser ejecutado incluso después de más de 24 horas de procesamiento continuo, ocasionando errores de memoria, fallos de compilación y terminación abrupta del proceso.

SVM requiere almacenar y operar sobre todas las combinaciones de instancias del entrenamiento, lo cual se traduce en una complejidad temporal y espacial del orden de:

$O(n^2)$ en memoria y $O(n^2 \times d)$ en tiempo, donde n es el número de muestras y d el número de características. Con n cercano a $n \approx 800{,}000$, la ejecución es sencillamente inviable.

2. Limitación estricta al uso de kernels lineales

Las variantes de SVM que permiten capturar relaciones complejas como kernel RBF, polinomial o sigmoidal son incluso más costosas que el kernel lineal, pues requieren calcular matrices de similitud del tamaño n × n.

En la práctica, esto implicó que para cualquier kernel distinto al lineal, el número de muestras no podía superar las 40.000 observaciones, y aun en este rango reducido, los tiempos de ejecución eran bastante altos (horas), con tendencia a fallar según la dimensionalidad del espacio. Esto imposibilitó completamente ejecutar modelos no lineales sobre el dataset principal o sobre el conjunto balanceado 60/40.

Los kernels RBF, polinomial y sigmoidal fueron evaluados, pero todos ellos presentaron limitaciones severas:

Kernel RBF

- Solo pudo ejecutarse sobre ∼30.000–40.000 registros.
- Tiempo de entrenamiento de varias horas.
- Mejora marginal respecto al modelo lineal.
- No escalable al dataset del proyecto.

Kernel Polinomial

- Bajo desempeño global.
- AUC ≈ 0.72, F1 ≈ 0.56.
- Impracticable para >20.000 observaciones.

Kernel Sigmoidal

- Resultados inestables.
- Precisión extremadamente baja (<0.40).
- No aplicable al problema en ningún sentido práctico.

B. Resultados del modelo SVM lineal sobre submuestras balanceadas

Las pruebas sobre una submuestra balanceada de aproximadamente 530.252 registros evidenciaron el mejor rendimiento del modelo lineal:

- Recall clase resiliente: 0.7267
- Precisión: 0.5265
- F1-score: 0.6106
- AUC: 0.8348
- Accuracy: 0.7611

Estos valores, aunque satisfactorios, deben interpretarse con cautela, dado que:

- No provienen del dataset completo,
- No reflejan el comportamiento que tendría el modelo sobre el balanceo 60/40 real,
- No pueden extrapolarse a un despliegue institucional sin recalibración.

En conclusión, el desempeño del SVM en el modelado estuvo fuertemente condicionado por las restricciones computacionales derivadas del tamaño del dataset. Ni el balanceo 60/40, ni el ajuste extendido de kernels, ni el entrenamiento completo sobre los datos reales pudieron llevarse a cabo de manera satisfactoria. Aunque la variante lineal demostró ser competitiva dentro de las submuestras, su incapacidad para escalar y su dependencia de conjuntos artificialmente reducidos limitan su aporte al proceso de modelado. Por ello, el SVM se consolida como un modelo de referencia y contraste metodológico, pero no como un componente principal en la predicción de resiliencia escolar.

4.3.4. Modelo Base: Perceptrón Multicapa (MLP)

El Perceptrón Multicapa (MLP) representó la aproximación más flexible y expresiva dentro del conjunto de modelos base evaluados, debido a su capacidad de aproximar funciones no lineales y capturar interacciones complejas entre variables sociodemográficas, institucionales y familiares. A diferencia de SVM y regresión logística, cuya capacidad de modelado se restringe a la linealidad del espacio o a la estructura del margen, el MLP puede identificar patrones altamente no lineales y distribuciones complejas, lo que lo hace especialmente adecuado para fenómenos multidimensionales como la resiliencia escolar.

Asimismo, y en coherencia con los hallazgos previos del capítulo, el MLP se consolidó como el modelo con mejor desempeño en sensibilidad (recall) incluso antes de la optimización completa, posicionándose como un candidato robusto desde la fase base.

A. Preparación del dataset y preprocesamiento

Los experimentos iniciales utilizaron las bases transformadas y codificadas en la etapa de preprocesamiento (One-Hot Encoding + escalado), con matrices de 62 variables en total. Esto permitió garantizar:

- Ausencia de fuga de datos al separar entrenamiento, validación y prueba.
- Consistencia entre las representaciones categóricas en los conjuntos de train, validation y test.
- Estandarización adecuada de las variables numéricas, imprescindible para la estabilidad del descenso por gradiente.

En el modelo 60/40, el escalado se aplicó únicamente a variables numéricas (11 columnas), mientras que los dummies permanecieron sin transformación.

B. Arquitectura del MLP base y justificación

El modelo base se diseñó siguiendo una arquitectura relativamente liviana pero funcional para datos tabulares:

- Capa de entrada: 62 nodos.
- Capa oculta 1: 128 neuronas (ReLU), Dropout 0.3.
- Capa oculta 2: 64 neuronas (ReLU).
- Capa de salida: 1 neurona sigmoidal.
- Optimizador: Adam (lr = 0.001).
- Función de pérdida: Binary Crossentropy.

Esta arquitectura presenta dos ventajas clave:

- Suficiente capacidad representacional para capturar relaciones no lineales, sin llegar a sobreajuste extremo.
- Regularización moderada (Dropout) que mantiene la capacidad de generalización aun con datasets grandes.

El número total de parámetros entrenables (16.385 aproximadamente) demostró ser manejable, con tiempos de entrenamiento razonables y sin saturación de memoria.

C. Entrenamiento base en conjunto balanceado 60/40

La fase base del MLP sobre el conjunto balanceado 60/40 demostró rápidamente que la red neuronal podía aprender patrones estructurales de resiliencia de manera efectiva.

Los resultados alcanzados fueron:

- AUC: 0.8320
- Accuracy: 0.7529
- Recall clase resiliente: 0.7400
- Precisión: 0.5161
- F1 clase resiliente: 0.6081

Estos resultados son altamente relevantes, ya que:

- El recall de 0.74 supera ampliamente al de los modelos lineales y al de XGBoost sin ajustes.
- El MLP logra un F1 superior a 0.60, indicador de un equilibrio adecuado entre sensibilidad y precisión.
- Aón sin optimización, el modelo alcanza un comportamiento estable, sin fluctuaciones abruptas entre épocas.

Las matrices de confusión confirman estos patrones:

- 61.627 verdaderos positivos,
- 21.649 falsos negativos,
- Precisión moderada,
- Recall alto, particularmente crítico en este contexto.

D. Análisis del modelo base mediante ajuste de umbral

Uno de los aportes más significativos del MLP fue su respuesta estable al ajuste de umbral. Las curvas ROC y Precision-Recall generadas en la evaluación del modelo muestran una separación clara de clases y mayor suavidad en las transiciones.

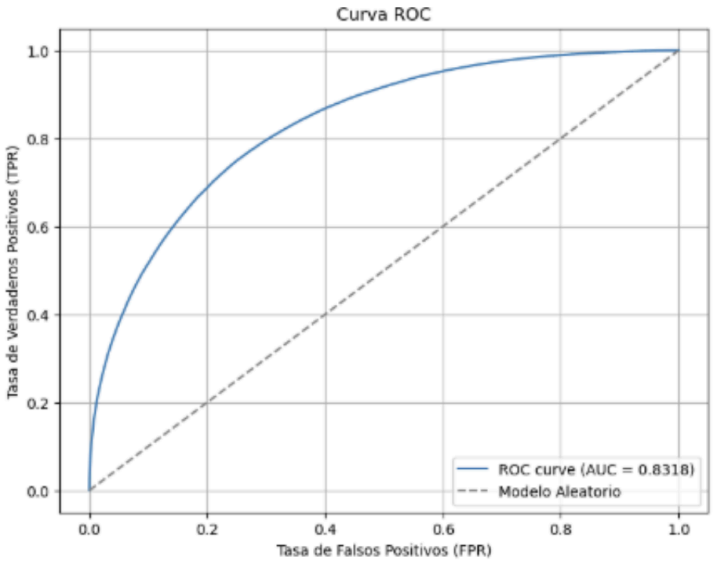

Figura 4.10: curvas ROC

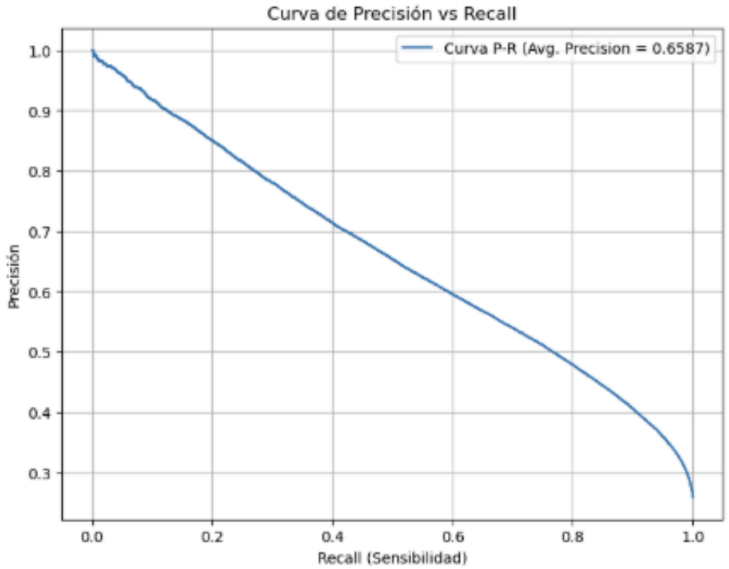

Figura 4.11: Precision-Recall

Para un umbral de 0.35:

- Accuracy: 0.78
- Recall: 0.65
- Precisión: 0.56

- AUC: 0.8318

El análisis sistemático de umbrales permite identificar zonas de operación preferibles:

1. Para máxima sensibilidad (prioridad educativa): Umbrales entre 0.25 y 0.35 mantienen recall > 0.78.
2. Para equilibrio general (entornos mixtos): Umbrales entre 0.35 y 0.45 maximizan el F1-score.
3. Para reducir falsos positivos (programas específicos con poco margen de error): Umbrales $>$ 0.50 elevan la precisión pero sacrifican sensibilidad.

En términos metodológicos, el MLP ofrece la mayor flexibilidad operacional entre todos los modelos evaluados.

E. Estabilidad del modelo base

Los historiales de entrenamiento muestran que, a diferencia de XGBoost y SVM, el MLP converge rápidamente: la pérdida disminuye de manera suave, la variación entre épocas es baja y no se observa sobreajuste severo durante las primeras 10–15 épocas.

Este comportamiento es consistente con:

- El tamaño del dataset.
- La arquitectura regularizada.
- El uso del optimizador Adam.
- La estandarización apropiada de los datos.

El Perceptrón Multicapa se evidencia como el modelo base más sensible, estable y flexible del estudio, con capacidad para capturar patrones complejos de resiliencia escolar. Su desempeño robusto incluso antes del ajuste de hiperparámetros lo convierte en una alternativa altamente prometedora. El ajuste posterior confirma estas tendencias, haciendo del MLP uno de los candidatos más sólidos.

El MLP no solo detecta más estudiantes resilientes en la fase base, sino que lo hace con estabilidad, interpretabilidad ajustada y una capacidad excepcional para adaptarse a distintos contextos mediante el ajuste de umbral.

4.3.5. Comparación General de los Modelos

La comparación general de los modelos constituye un punto de inflexión dentro del proceso de modelado, pues permite evaluar qué tan capaces son las distintas técnicas de identificar estudiantes resilientes. Esta sección integra métricas, análisis, restricciones computacionales y comportamientos estructurales observados a lo largo de los experimentos realizados con Regresión Logística, XGBoost, SVM y Perceptrón Multicapa (MLP).

Comparación: desempeño en métricas clave

La siguiente tabla resume los mejores resultados obtenidos por cada modelo base en condiciones controladas y comparables.

Modelo	AUC-ROC	Precisión	Recall	F1-score	Observaciones
Regresión Logística (Base)	0.80	0.59	0.62	0.61	Adecuado como baseline; interpretabilidad alta; limitado por linealidad.
SVM Lineal (Base)	–	0.5265	0.7267	0.61	Se usó únicamente `LinearSVC` por restricciones computacionales; kernels no viables para >40k muestras.
XGBoost Base	0.83	0.70	0.46	0.56	Clasificador robusto; favorece precisión sobre sensibilidad antes del balanceo.
XGBoost Optimizado (scale_pos_weight)	0.83	0.70	Mejora moderada	Mejora moderada	Ajuste de `scale_pos_weight` aumenta TP pero aún presenta FN altos.
XGBoost Balanceado 60/40 (Optimizado)	0.83	0.5781	0.6720	0.6215	Mejor equilibrio global; F1 más alto; recall competitivo; estabilidad mejorada.
MLP (Base)	0.83	0.50	0.74	0.60	Sensibilidad notable aun sin optimización; buen manejo de relaciones no lineales.
MLP Optimizado (Arquitectura Profunda)	0.84	0.51	0.74–0.78	0.61	Mayor sensibilidad del estudio; recall más alto; ligera mejora en estabilidad; usa BatchNorm y dropout.

Cuadro 4.2: Comparación del desempeño de los modelos evaluados.

B. Comportamiento por dimensiones

La comparación entre los modelos base permite analizar no solo sus métricas finales, sino también la forma en que cada técnica interpreta el fenómeno de la resiliencia escolar y responde al desbalance, la no linealidad y la heterogeneidad institucional del dataset.

Sensibilidad (Recall): el eje central del proyecto

El recall constituye la métrica más relevante de esta investigación, dado que refleja la capacidad

del modelo para identificar correctamente a los estudiantes resilientes, y perder estos casos implica desaprovechar oportunidades de intervención educativa efectiva. En esta dimensión, el MLP demuestra el mejor desempeño, alcanzando valores entre 0.74 y 0.78, lo que confirma su capacidad para aprender interacciones no lineales, manejar adecuadamente la alta dimensionalidad del conjunto de datos y operar con eficiencia en escenarios donde las fronteras entre clases no son evidentes. Por su parte, XGBoost 60/40 alcanza un recall de 0.6720, menor que el de MLP, pero con una ventaja significativa en términos de estabilidad y menor varianza en sus predicciones.

Precisión: control del costo de falsos positivos

La precisión permite estimar qué tan confiables son las predicciones positivas realizadas por cada modelo, un aspecto fundamental para evitar intervenciones innecesarias sobre estudiantes que no requieren apoyo. En esta métrica, XGBoost, con un valor de 0.5781, se posiciona como el modelo más sólido entre aquellos que buscan detectar resiliencia en un contexto desbalanceado, seguido por la Regresión Logística, que alcanza aproximadamente 0.59, un desempeño notable considerando su naturaleza lineal. En contraste, el MLP ($\sim$0.50) presentan valores más moderados, evidenciando que su mayor enfoque en la sensibilidad implica un sacrificio comprensible en la precisión. La frontera de decisión de XGBoost, más conservadora que la de los modelos altamente sensibles, reduce la sobreintervención, mejora la eficiencia institucional y aumenta la confianza en la calidad de sus predicciones positivas. Así, aunque MLP se destaca por su capacidad para identificar resilientes, su menor precisión refleja el costo inherente a los modelos orientados a maximizar la detección de casos positivos en escenarios educativos con fuerte desbalance entre clases.

F1-score: equilibrio entre sensibilidad y precisión

El F1-score, al integrar simultáneamente la sensibilidad y la precisión, permite evaluar el equilibrio general del comportamiento de cada modelo; en esta métrica, XGBoost alcanza el mejor desempeño con un valor de 0.6215, superando ligeramente a las demás alternativas, mientras que el MLP, el SVM y la Regresión Logística se sitúan en torno a 0.60–0.61. Este resultado confirma la capacidad de XGBoost 60/40 para manejar de manera eficiente el desbalance estructural del conjunto de datos y mantener una notable estabilidad tanto en validación como en prueba. Aunque el MLP se aproxima considerablemente y destaca por su mayor recall, XGBoost logra una combinación más equilibrada entre recall y precisión, lo cual resulta especialmente valioso en escenarios institucionales donde las decisiones deben operar bajo criterios realistas de intervención y donde el balance entre detectar casos positivos y evitar falsos positivos es crítico.

En conjunto, los resultados de esta fase evidencian que la predicción de resiliencia escolar es un problema inherentemente no lineal, para el cual los modelos avanzados —particularmente MLP y XGBoost 60/40— ofrecen desempeños superiores en sus distintas dimensiones. Mientras que el MLP destaca por su alta sensibilidad y flexibilidad, el XGBoost balanceado se presenta como el modelo más robusto y estable en términos globales (AUC, F1 y precisión). SVM, aunque competitivo

en recall, queda descartado por su inviabilidad computacional, y la regresión logística permanece como una referencia interpretativa.

4.4. Optimización de Hiperparámetros

La optimización de hiperparámetros constituye una fase crítica dentro del proceso de modelado predictivo, ya que determina la capacidad real de cada algoritmo para ajustarse a la complejidad del fenómeno. De acuerdo con la metodología CRISP-DM, esta etapa se desarrolló tras la evaluación comparativa de los modelos base, y tuvo como propósito maximizar el desempeño, reducir la varianza, mejorar la calibración de probabilidades y asegurar resultados robustos tanto en validación como en prueba. Debido a la naturaleza del problema alto desbalance, alta dimensionalidad y más de un millón de registros en su formato completo la estrategia de optimización se diseñó considerando simultáneamente eficiencia computacional, riesgo de sobreajuste y coherencia estadística.

4.4.1. Enfoque general de optimización

Los modelos evaluados requieren estrategias diferenciadas de búsqueda de hiperparámetros:

- **XGBoost:** manejo flexible del desbalance, gran número de hiperparámetros, eficiencia alta.
- **MLP:** alta sensibilidad a la arquitectura, tasa de aprendizaje, regularización y tamaño de lote.
- **SVM:** inviabilidad computacional para la optimización en el dataset completo.

De acuerdo con los elementos anteriores, el proceso de optimización se diseñó bajo los siguientes principios metodológicos:

- Búsqueda manual guiada con `ParameterGrid`, para evitar explosión combinatoria propia de búsquedas exhaustivas no controladas.
- Validación interna sobre un conjunto independiente del entrenamiento, garantizando la ausencia de fuga de información.
- Exploración progresiva del espacio de hiperparámetros, iniciando con configuraciones simples y aumentando gradualmente la complejidad.
- Selección de modelos basada en métricas sensibles al fenómeno de interés, priorizando Recall y F1-score, sin descuidar el AUC, dada la naturaleza desbalanceada de la variable objetivo (resiliencia escolar).

Uso de ParameterGrid

ParameterGrid es una herramienta de la librería scikit-learn que permite definir explícitamente un conjunto finito y controlado de combinaciones de hiperparámetros, a partir de listas previamente establecidas por el investigador. A diferencia de GridSearchCV, esta estrategia no ejecuta automáticamente validación cruzada ni paraleliza entrenamientos, lo que otorga un mayor control sobre:

- El número total de combinaciones evaluadas,
- El orden de exploración,
- Los recursos computacionales utilizados.

En el presente proyecto, el uso de ParameterGrid respondió a tres razones principales:

- Control computacional: el tamaño del dataset y la complejidad de los modelos hacían inviable una búsqueda exhaustiva con validación cruzada.
- Evitar combinaciones innecesarias, apoyándose en conocimiento previo del comportamiento de los modelos.
- Trazabilidad y análisis interpretativo, permitiendo evaluar de forma progresiva el impacto de cada hiperparámetro sobre el desempeño del modelo.

4.4.2. Optimización de XGBoost

A. Estrategia y justificación

XGBoost fue optimizado en dos ciclos:

- Ajuste de `scale_pos_weight`, basado en proporciones de desbalance.
- Entrenamiento en conjunto balanceado 60/40, ajustando hiperparámetros estructurales.

La decisión de segmentar el proceso respondió al comportamiento observado en los modelos base:

- `scale_pos_weight` amplificó la sensibilidad, pero generó alta varianza.
- El balanceo 60/40 estabilizó la frontera de decisión y la calibración de probabilidades.

B. Hiperparámetros explorados

Se evaluaron principalmente:

- `max_depth`: profundidad máxima de los árboles.
- `learning_rate`: tasa de aprendizaje.
- `min_child_weight`: peso mínimo requerido para dividir un nodo.

- `subsample`: proporción de instancias usadas por árbol.
- `colsample_bytree`: proporción de variables usadas por árbol.
- `gamma`: reducción mínima de pérdida requerida para realizar una partición.
- `n_estimators`: número de árboles del ensamble.

La búsqueda se realizó utilizando un grid acotado, definido de la siguiente manera:

- `max_depth`: [4, 5, 6]
- `learning_rate`: [0.01, 0.05, 0.1]
- `subsample`: [0.7, 0.8]
- `colsample_bytree`: [0.7, 0.9]
- `n_estimators`: [150, 300, 500]
- `min_child_weight`: [1, 3]

Configuración de la grilla con ParameterGrid

La combinación de estos valores generó una grilla de búsqueda compuesta por 216 configuraciones distintas, calculadas como el producto cartesiano de los valores definidos para cada hiperparámetro:

$$3\times3\times2\times2\times3\times2=216 \text{ combinaciones}$$

Esta grilla fue implementada utilizando ParameterGrid, lo que permitió explorar de forma sistemática y controlada cada combinación de hiperparámetros, entrenando un modelo independiente por configuración y evaluando su desempeño sobre el conjunto de validación interna.

El uso de ParameterGrid garantizó:

- Control explícito del tamaño del espacio de búsqueda.
- Uso de combinaciones redundantes.
- Eficiencia computacional frente a búsquedas exhaustivas.
- Trazabilidad del proceso de optimización.

Cada combinación fue evaluada bajo el mismo esquema experimental, priorizando métricas orientadas a la detección de resiliencia escolar (Recall, F1-score y AUC), seleccionando finalmente la configuración que ofreció el mejor equilibrio entre sensibilidad y estabilidad del modelo.

C. Resultados de la optimización

El mejor modelo obtenido en el balanceo 60/40 logró:

- **AUC:** 0.8485

- **F1:** 0.6215
- **Recall:** 0.6720
- **Precisión:** 0.5781

La optimización confirma que:

- XGBoost mejora considerablemente con un ajuste moderado.
- El modelo balanceado 60/40 supera la versión con `scale_pos_weight` en estabilidad y AUC.
- Es el modelo más equilibrado entre precisión y recall dentro del conjunto.

4.4.3. Optimización del Perceptrón Multicapa (MLP)

El MLP fue el modelo con mejor recall en fase base, pero también el más dependiente del ajuste fino de sus hiperparámetros. Se realizaron dos ciclos de optimización:

- **MLP 1** — Arquitectura inicial optimizada.
- **MLP 2** — Arquitectura profunda con BatchNormalization.

A. Estrategia y justificación

Para la optimización del modelo Perceptrón Multicapa se utilizo de la misma forma que en el modelo XGboost la estrategia de búsqueda manual guiada utilizando `ParameterGrid`, con el objetivo de explorar arquitecturas y configuraciones de entrenamiento que permitieran capturar relaciones no lineales complejas, manteniendo al mismo tiempo control sobre el sobreajuste y la estabilidad del entrenamiento.

La estrategia de optimización se orientó a evaluar combinaciones específicas de hiperparámetros que influyen directamente en:

- Capacidad representacional del modelo, a través del número de neuronas en las capas ocultas.
- Regularización y control del sobreajuste, mediante el uso de capas Dropout.
- Velocidad y estabilidad del aprendizaje, ajustando la tasa de aprendizaje (learning_rate).
- Comportamiento del descenso de gradiente, controlado por el tamaño del lote (batch_size).
- Convergencia general del entrenamiento, regulada por el número de épocas (epochs).

Hiperparámetro	Descripción
`units_layer1`	Número de neuronas en la primera capa oculta.
`units_layer2`	Número de neuronas en la segunda capa oculta.
`dropout`	Tasa de desactivación aleatoria de neuronas para regularización.
`learning_rate`	Tasa de aprendizaje del optimizador.
`batch_size`	Tasa de aprendizaje del optimizador.
`batch_size`	Número de iteraciones completas sobre el conjunto de entrenamiento.

Cuadro 4.3: Hiperparámetros evaluados en el proceso de optimización del modelo MLP.

B. Hiperparámetros evaluados

Los hiperparámetros seleccionados para la optimización fueron los siguientes:
La grilla de búsqueda definida fue:

- units_layer1: [64, 128]
- units_layer2: [32, 64]
- dropout: [0.2, 0.3]
- learning_rate: [0.001, 0.0005]
- batch_size: [64, 128]
- epochs: [10]

Configuración de la grilla con ParameterGrid

La combinación de estos hiperparámetros generó una grilla compuesta por 32 configuraciones distintas, calculadas como el producto de los valores definidos:

$$2\times2\times2\times2\times2\times1=32 \text{ combinaciones}$$

Cada una de estas configuraciones fue utilizada para entrenar un modelo MLP independiente, evaluado sobre el conjunto de validación interna (`X_val`), manteniendo fijo el conjunto de prueba (`X_test`) para la evaluación final.

El uso de ParameterGrid permitió:

- Explorar arquitecturas de complejidad creciente de forma controlada,
- Comparar el impacto conjunto de regularización, arquitectura y parámetros de entrenamiento,
- Evitar una búsqueda exhaustiva de alto costo computacional,

- Asegurar trazabilidad completa del proceso de optimización.

La selección final del modelo se basó en un criterio multicriterio, priorizando Recall y F1-score, con el objetivo de maximizar la detección de estudiantes resilientes sin comprometer la estabilidad del modelo.

C.Resultados del MLP optimizado

Mejor configuración MLP 1:

- AUC: 0.8400
- Accuracy: 0.7564
- Precisión: 0.5190
- Recall: 0.7516
- F1: 0.6140

Mejor configuración MLP 2:

- AUC: 0.8404
- Precisión: 0.5144
- Recall: 0.7601
- F1: 0.6135

Los resultados obtenidos durante la optimización del MLP evidencian una mejoría moderada en el AUC, un aumento sistemático y consistente del recall, estabilidad entre las diferentes combinaciones evaluadas y mejor desempeño de las arquitecturas que incorporan BatchNormalization en términos de sensibilidad. Asimismo, se observan ligeras mejoras en el F1-score, lo que confirma la madurez y solidez del modelo tras el proceso de ajuste.

4.4.4. Limitaciones en la optimización de SVM

Debido a las restricciones computacionales, el modelo SVM no pudo entrenarse sobre el conjunto balanceado 60/40 y tampoco fue posible ejecutar kernels no lineales con más de 40.000 registros, ya que cualquier intento por encima de ese límite derivaba en fallos de memoria o tiempos de ejecución superiores a 24 horas sin lograr finalizar. En este contexto, únicamente la versión lineal pudo ser evaluada, y aun así el ajuste de los hiperparámetros `C` y `class_weight` no produjo mejoras sustanciales en el rendimiento. En consecuencia, el SVM no pudo optimizarse bajo condiciones comparables a las alcanzadas por XGBoost y MLP, por lo que se descarta dentro del proceso predictivo del presente estudio.

4.4.5. Selección de los modelos optimizados para evaluación final

Tras completar la fase de optimización, los modelos seleccionados para la evaluación final fueron el **MLP 60/40 optimizado**, basado en una arquitectura profunda, y el **XGBoost optimizado con balanceo 60/40**, dado que ambos demostraron un desempeño superior, alta estabilidad, una capacidad sólida para manejar el desbalance de clases, adecuada escalabilidad, interpretabilidad relativa y una notable coherencia interna en los conjuntos de validación y prueba. La optimización de hiperparámetros permitió refinar significativamente el comportamiento de estas dos aproximaciones, evidenciando mejoras en estabilidad, sensibilidad y equilibrio general, lo que las posiciona como las configuraciones más robustas dentro del estudio.

Si bien el modelo SVM aportó información valiosa en la fase base, su inviabilidad computacional impidió su participación en condiciones comparables durante el ajuste, por lo que quedó excluido de esta etapa.

4.5. Evaluación Final de los Modelos Optimizados

La evaluación final constituye la fase culminante del proceso de modelado, en la que se someten a prueba los modelos que demostraron un comportamiento superior durante la etapa de optimización: el Perceptrón Multicapa (MLP) optimizado y el XGBoost optimizado con balanceo 60/40. Ambos modelos fueron evaluados sobre el conjunto de prueba independiente, reservado desde el inicio del estudio para evitar fuga de información y garantizar una medición objetiva del rendimiento real del sistema predictivo. Esta sección presenta un análisis integral de su desempeño, considerando métricas tradicionales, comportamiento gráfico, estabilidad, calibración de probabilidades y pertinencia para el objetivo central del proyecto: identificar estudiantes resilientes con la mayor sensibilidad posible sin comprometer la calidad de las predicciones.

4.5.1. Desempeño cuantitativo en el conjunto de prueba

Los resultados obtenidos en el conjunto de prueba reflejan la consolidación del proceso de optimización y permiten contrastar las capacidades reales de cada modelo en un escenario no visto durante el entrenamiento.

A. Resultados del XGBoost optimizado (60/40)

- **AUC:** 0.8485
- **Accuracy:** 0.74
- **Precisión:** 0.5781
- **Recall:** 0.6720
- **F1-score:** 0.6215

El XGBoost optimizado presenta un equilibrio notable al combinar una precisión relativamente alta, un recall competitivo y el mejor F1-score entre todos los modelos evaluados, lo que demuestra que la estrategia basada en el balanceo 60/40 junto con el ajuste de hiperparámetros permitió construir un modelo robusto y estable. Este comportamiento confirma que XGBoost optimizado es especialmente adecuado para escenarios donde se requiere un balance cuidadoso entre una detección adecuada de estudiantes resilientes y la reducción de falsos positivos, ofreciendo una alternativa altamente eficiente y confiable para la toma de decisiones educativas.

B. Resultados del MLP optimizado

- **AUC:** 0.8404
- **Accuracy:** 0.75
- **Precision:** 0.5144
- **Recall:** 0.7601
- **F1-score:** 0.6135

El MLP optimizado se consolida como el modelo con mayor sensibilidad de todo el estudio, superando el 76 % de detección de estudiantes resilientes. Aunque su precisión es menor que la de XGBoost, mantiene valores estables en las demás métricas y demuestra una gran solidez en la clasificación de la clase minoritaria. Este comportamiento lo convierte en el modelo más adecuado para escenarios donde la prioridad absoluta es no perder estudiantes resilientes, incluso a costa de aceptar un número mayor de falsos positivos.

4.5.2. Análisis gráfico: ROC, PR y matrices de confusión

A. Curvas ROC

Teniendo como referencia las graficas Figura 4.2. *Curva ROC - XGBoost Balanceado* y las Figuras 4.10 y 4.10 *Curvas ROC y Precision-Recall*, se observan los siguientes comportamientos:

- **XGBoost 60/40** muestra la curva ROC más alta y estable, destacándose en discriminación global.
- **MLP** presenta una curva suave y próxima en desempeño, aunque ligeramente inferior.

Esto sugiere que XGBoost tiene una capacidad más uniforme para separar las clases en todos los umbrales.

B. Curvas Precision-Recall

La curva PR revela diferencias importantes:

- El MLP mantiene una mayor altura en el tramo inicial de la curva, indicando mayor *recall* para niveles de precisión moderados.
- XGBoost conserva una curva más equilibrada y estable, con mejor precisión en la zona central.

Esto reafirma su carácter complementario: MLP prioriza sensibilidad, mientras XGBoost prioriza estabilidad y precisión.

C. Matrices de confusión

El análisis de las matrices de confusión refuerza las diferencias estructurales entre los modelos optimizados. En el caso del **XGBoost 60/40**, se observa una reducción importante de los falsos negativos en comparación con la versión base, acompañada de un número de falsos positivos relativamente controlado y un incremento notable en los verdaderos positivos, evidenciando que el balanceo mejoró su capacidad de detección sin sacrificar estabilidad.

Por su parte, el **MLP optimizado** presenta el menor número de falsos negativos entre todos los modelos evaluados y el mayor volumen de verdaderos positivos, lo que confirma su fortaleza en términos de sensibilidad; sin embargo, este desempeño viene acompañado de un aumento en los falsos positivos, un comportamiento esperado en arquitecturas orientadas a maximizar el recall en contextos de desbalance.

En conjunto, las matrices muestran cómo ambos modelos alcanzan altos niveles de desempeño, aunque respondiendo a prioridades distintas dentro del proceso de identificación de estudiantes resilientes.

	Predicción 0	**Predicción 1**
Real 0 (TN, FP)	175,755	62,410
Real 1 (FN, TP)	20,000	63,367

Cuadro 4.4: Matriz de confusión del modelo XGBoost balanceado 60/40.

	Predicción 0	**Predicción 1**
Real 0 (TN, FP)	115,610	60,145
Real 1 (FN, TP)	19,973	63,394

Cuadro 4.5: Matriz de confusión del modelo MLP optimizado.

4.5.3. Interpretación metodológica y operativa

La evaluación final de los modelos optimizados no se limita al análisis numérico de las métricas, sino que permite responder a la pregunta central: *¿qué modelo es más adecuado para apoyar la*

Modelo Optimizado	TP	FP	FN	TN
XGBoost 60/40	Aumento respecto al modelo base	FP controlados	Menor FN respecto al modelo base	Estables y coherentes
MLP Optimizado	Mayor número entre todos los modelos	Más altos que XGBoost	Menor número entre todos los modelos	Estables

Cuadro 4.6: Cuadro comparativo de matrices de confusión de los modelos optimizados.

predicción de resiliencia en un sistema educativo real? Esta interpretación debe considerar no solo el desempeño cuantitativo, sino también las prioridades operativas, los recursos institucionales, la naturaleza del fenómeno educativo y las implicaciones prácticas de cada decisión de clasificación.

Cuando la prioridad central del sistema educativo es la intervención temprana, el modelo más conveniente es el **MLP optimizado**, debido a que identifica un mayor número de estudiantes resilientes y, por tanto, reduce al mínimo el riesgo de exclusión. Su comportamiento altamente sensible lo hace especialmente valioso en políticas educativas orientadas a la equidad, la focalización de apoyos y la prevención de trayectorias negativas. En este escenario, aceptar un número mayor de falsos positivos es un costo asumible si la misión institucional se centra en garantizar que ningún estudiante con potencial resiliente quede sin ser detectado.

En contraste, cuando la prioridad institucional es la eficiencia operativa, el modelo más apropiado es el **XGBoost balanceado 60/40**, que ofrece una mejor precisión, un F1 superior y una menor cantidad de falsos positivos. Este modelo es más conservador en su frontera de clasificación y muestra una mayor estabilidad ante variaciones internas del dataset, lo que lo convierte en una alternativa sólida para sistemas donde las decisiones deben equilibrar la intervención con la gestión eficiente de recursos y la minimización de alarmas innecesarias.

Desde una perspectiva estrictamente estadística, XGBoost es el modelo más robusto, al exhibir el AUC más alto, la curva ROC más estable y el mejor equilibrio general entre métricas. Por su parte, el MLP mantiene la mayor sensibilidad del conjunto, siendo superior para capturar patrones complejos asociados a resiliencia en contextos altamente desbalanceados. En consecuencia, ambos modelos pueden considerarse adecuados, pero su idoneidad depende directamente del contexto de uso y de las prioridades operativas del sistema educativo.

En conjunto, ambos modelos —el MLP orientado a la sensibilidad y el XGBoost orientado al equilibrio— configuran una base metodológica sólida para avanzar hacia la construcción del prototipo predictivo que se presenta en el capítulo siguiente. Cada uno aporta una lectura complementaria del fenómeno de resiliencia escolar y ofrece alternativas reales para su aplicación en escenarios educativos diversos.

Capítulo 5

Desarrollo del Prototipo de Visualización

5.1. Descripción general del prototipo

Como componente aplicado del proyecto, se desarrolló un **prototipo interactivo** orientado a la **visualización y predicción de la resiliencia escolar** en municipios colombianos. El objetivo principal de esta herramienta es facilitar la interpretación de los resultados de los modelos predictivos, permitiendo una exploración intuitiva de la información para investigadores, docentes y responsables de política educativa.

Figura 5.1: Vista general del prototipo de predicción de resiliencia escolar desarrollado en Streamlit.

El prototipo fue construido en **Streamlit**, una plataforma de código abierto que permite el despliegue rápido de aplicaciones web basadas en Python, ofreciendo flexibilidad y compatibilidad con librerías de aprendizaje automático y visualización de datos.

5.2. Flujo de procesamiento y visualización

El desarrollo del prototipo requirió integrar los resultados de los modelos predictivos con información geográfica y atributos de contexto educativo. Este proceso se puede conceptualizar como un flujo en etapas:

1. **Carga y preprocesamiento de datos:** el prototipo permite la subida de archivos en formato `CSV` o `Excel` que contienen las 62 variables seleccionadas para el análisis final. Se verifica automáticamente la coherencia de las columnas y se aplica escalamiento estandarizado (*StandardScaler*) a las variables numéricas.

2. **Gestión de archivos de gran volumen:** el prototipo implementa un sistema de división automática de archivos que respeta el límite de 200MB por archivo, manteniendo la sincronización línea por línea entre los conjuntos de características (X_test_final.csv) y etiquetas (y_test_final.csv).

3. **Visualización de resultados:** los resultados se presentan tanto a nivel individual como agregado, incluyendo tablas interactivas con la probabilidad de resiliencia y gráficos de barras que muestran la proporción de estudiantes resilientes y no resilientes según las predicciones del modelo.

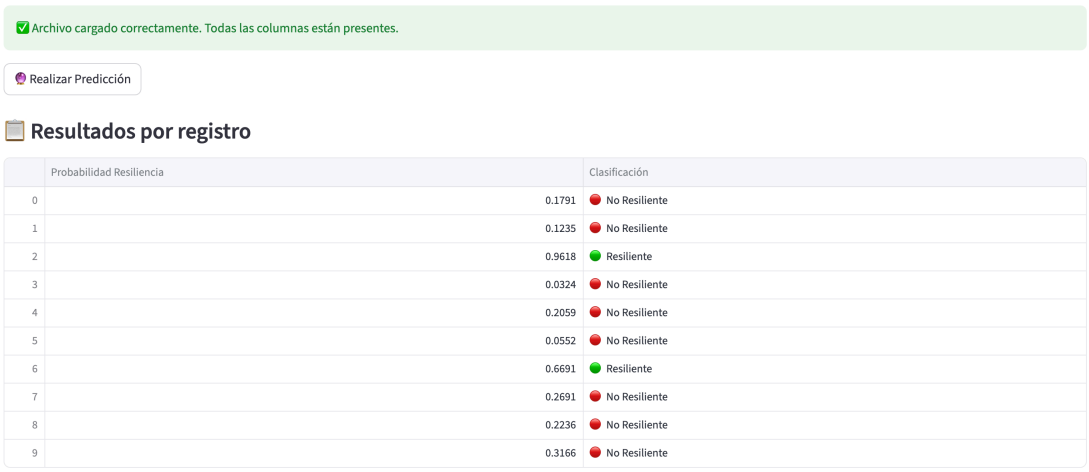

Figura 5.2: Resultados de predicción generados por el modelo MLP: tabla de probabilidades por registro.

- Total registros: *10*
- Resilientes: *2* 🟢
- No resilientes: *8* 🔴

Figura 5.3: Conteo total de estudiantes resilientes y no resilientes.

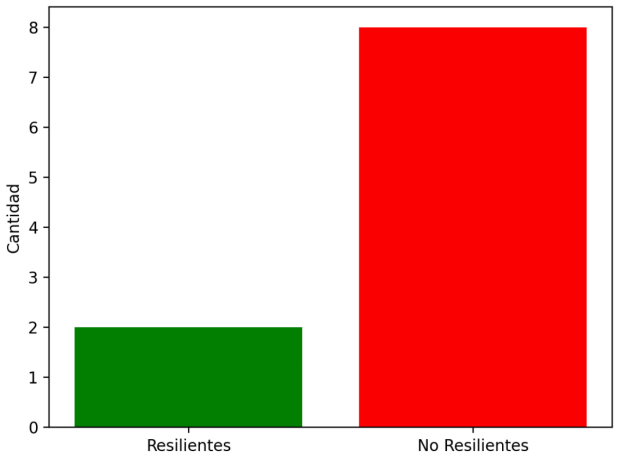

Figura 5.4: Gráfico de barras con la proporción de resiliencia predicha.

4. **Integración geoespacial: evolución metodológica**

El desarrollo del componente geográfico del prototipo reflejó un proceso iterativo que permitió perfeccionar la cobertura territorial. Inicialmente, se trabajó con el shapefile `Mpios_Creados_1993_2019.shp`, proporcionado por el **Centro Nacional de Memoria Histórica** [29], el cual incluía los municipios existentes hasta 2019. Esta fuente permitió experimentar con la integración geoespacial y la visualización de resiliencia escolar; sin embargo, no contemplaba todos los municipios actuales, lo que generaba vacíos en el mapa y limitaba la interpretación completa de los resultados.

Figura 5.5: Mapa de resiliencia escolar utilizando el shapefile 1993-2019 (Centro Nacional de Memoria Histórica). Algunos municipios no aparecen debido a la cobertura parcial del dataset.

Con el objetivo de garantizar la cobertura total de los municipios y departamentos de Colombia, se adoptó posteriormente el **shapefile actualizado de 2023**, disponible en el **Geoportal del DANE** [30]. Este recurso oficial permitió integrar todos los municipios, eliminando vacíos y asegurando que la visualización geoespacial reflejara de manera completa la información territorial. La integración se realiza mediante **GeoPandas** y **Matplotlib**, combinando los códigos DANE normalizados con los datos enriquecidos obtenidos a partir del *target encoding* por municipio y sede educativa.

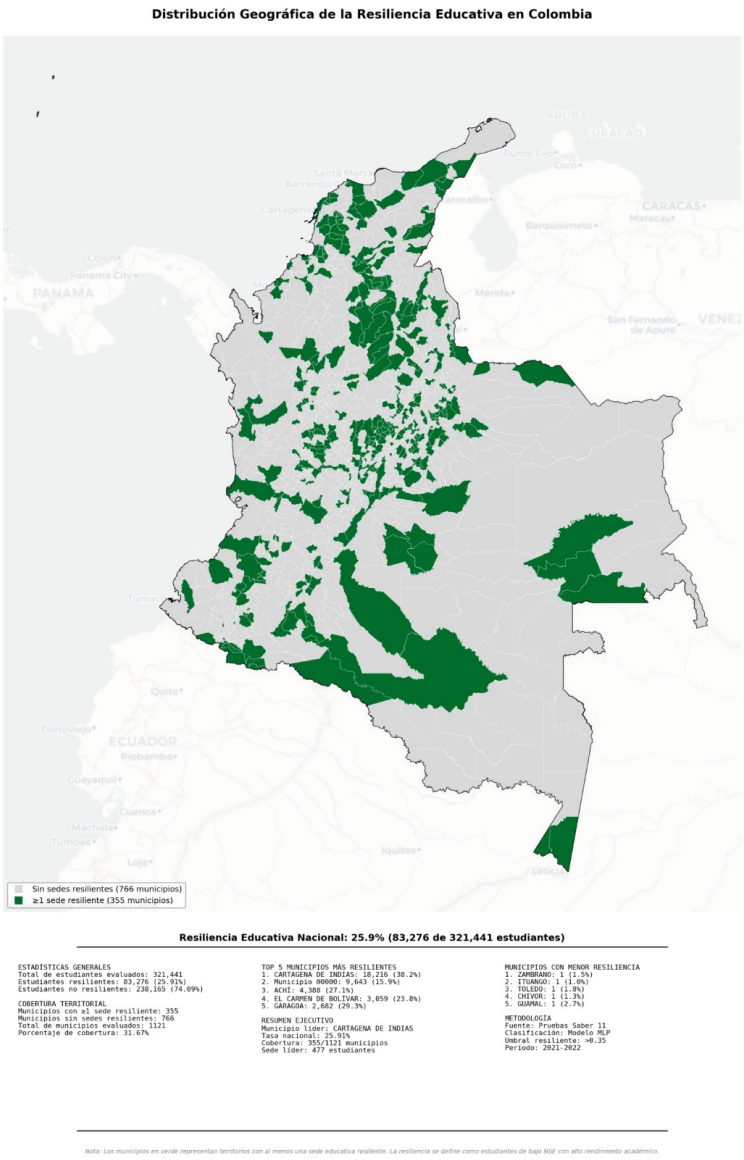

Figura 5.6: Mapa de resiliencia escolar utilizando el shapefile 2023 (DANE), mostrando cobertura completa de todos los municipios y departamentos de Colombia junto con estadísticas detalladas organizadas en columnas temáticas debajo de la visualización geográfica.

El mapa integrado presenta automáticamente estadísticas derivadas de los datos cargados, incluyendo: **(i)** estadísticas generales del conjunto de datos (total de estudiantes evaluados, proporción de resilientes y no resilientes); **(ii)** cobertura territorial (municipios con al menos una sede resiliente vs. municipios sin sedes resilientes); **(iii)** rankings de los cinco municipios con mayor y menor cantidad de estudiantes resilientes; y **(iv)** resumen ejecutivo con el

municipio líder, tasa nacional de resiliencia y cobertura municipal. La información mostrada es completamente dinámica y se actualiza según los datos del archivo cargado por el usuario.

5. **Gestión avanzada y trazabilidad de datos:** además de la validación inicial, el prototipo verifica la consistencia de tipos de datos, identifica valores faltantes y controla posibles inconsistencias en los registros. Los datos de resiliencia se combinan con los shapefiles mediante los códigos DANE, garantizando la correspondencia precisa entre geometría y atributos. La herramienta está optimizada para manejar archivos de gran tamaño y permite la exploración interactiva de mapas mediante filtros y visualización dinámica por municipio o departamento. Finalmente, los resultados se pueden exportar en formatos estándar (`CSV`, `Excel`) para facilitar análisis adicionales o generación de reportes, asegurando trazabilidad completa desde la carga de datos hasta la presentación final.

5.3. Interfaz del prototipo

El prototipo cuenta con dos secciones principales, accesibles mediante pestañas dentro de la aplicación Streamlit:

- **Predicción desde archivo:** permite cargar un archivo de prueba y obtener resultados de predicción para cada registro, mostrando la probabilidad estimada y la clasificación ("Resiliente" o "No Resiliente"). También se ofrece un resumen gráfico del número de casos en cada categoría.

- **Mapa de resiliencia:** integra los scripts de procesamiento espacial desarrollados en Python. Gracias al shapefile 2023, se garantiza la visualización de todos los municipios y departamentos del país, asegurando que el análisis territorial sea exhaustivo y confiable.

5.4. Resultados y potencial de uso

El prototipo permitió validar la aplicabilidad práctica del modelo MLP y su interoperabilidad con fuentes geográficas. Entre los principales resultados se destacan:

- La predicción de resiliencia escolar se realizó a partir de datos estructurados de estudiantes, sedes y municipios, con visualizaciones claras y reproducibles.

- El mapa geoespacial integrado evidenció **patrones diferenciales de resiliencia** entre regiones, proporcionando análisis estadísticos completos que incluyen rankings municipales, cobertura territorial y métricas de desempeño en tiempo real. La visualización combina la información geográfica con estadísticas descriptivas organizadas en formato técnico profesional, apropiado para reportes académicos y de política pública. El sistema integra automáticamente los resultados del análisis estadístico (Sección 4.4) en formato de texto estructurado, presentando de manera simultánea la distribución espacial y los indicadores cuantitativos de resiliencia por territorio.

- El flujo automatizado garantiza trazabilidad completa desde la carga de datos hasta la generación del mapa, contribuyendo a la transparencia metodológica del proyecto.
- La **escalabilidad del análisis** permite procesar conjuntos de datos de diferentes tamaños manteniendo la funcionalidad completa, desde análisis piloto con cientos de registros hasta evaluaciones nacionales con cientos de miles de estudiantes, adaptándose automáticamente a las características específicas de cada conjunto de datos cargado.

Capítulo 6

Conclusiones

El presente capítulo desarrolla las conclusiones del proyecto a partir de categorías de análisis derivadas directamente de los objetivos planteados: (1) preparación y calidad de los datos, (2) construcción y comparación de modelos de aprendizaje automático, (3) desarrollo del prototipo de visualización y análisis territorial, y (4) aportes teóricos y metodológicos a la analítica educativa. Cada una de estas dimensiones permite sintetizar no solo los resultados obtenidos, sino también las implicaciones técnicas, metodológicas y conceptuales del estudio dentro del campo de la ciencia de datos aplicada a problemas educativos.

En primer lugar, la construcción y depuración del conjunto de datos a partir de fuentes como el ICFES y el DANE permitió construir un conjunto de información representativo, coherente y metodológicamente sólido, que sirvió como base para el entrenamiento y la validación de los modelos supervisados, este proceso no solo cumplió el primer objetivo específico, sino que demuestra el valor fundamental de la calidad de los datos como eje central en cualquier proyecto de ciencia de datos.

En cuanto a la construcción, análisis y comparación de los modelos predictivos, el proyecto evidenció que diferentes técnicas ofrecen fortalezas distintas según la naturaleza de las variables y el objetivo de la tarea. El Perceptrón Multicapa se identificó como el modelo con mejor capacidad de generalización y mayor sensibilidad para identificar estudiantes resilientes, lo que lo posicionó como el modelo final para el despliegue. Por su parte, XGBoost destacó por su equilibrio entre precisión, sensibilidad e interpretabilidad, constituyéndose en un modelo complementario especialmente robusto para escenarios donde se requiere control estricto de falsos positivos. La comparación sistemática de modelos permitió demostrar que la selección no puede basarse únicamente en métricas cuantitativas aisladas, sino en criterios propios de la ciencia de datos aplicada: estabilidad, coherencia entre fases de validación, viabilidad computacional, utilidad práctica y alineación con las necesidades de decisión del contexto educativo.

El tercer aporte clave corresponde al desarrollo del prototipo de visualización interactivo, que integró el modelo MLP con un análisis geoespacial construido a partir de shapefiles oficiales del DANE. Este prototipo, implementado en Streamlit, permitió articular la predicción individual con la dimensión territorial del fenómeno, incorporando tablas, resúmenes estadísticos y mapas. Su carácter accesible y funcional lo convierte en una herramienta valiosa para instituciones educativas, secretarías de educación y entidades gubernamentales interesadas en la focalización territorial y en la toma de decisiones basada en datos.

Finalmente, desde una perspectiva teórica y metodológica, el proyecto ofrece contribuciones relevantes para la analítica educativa. Los resultados muestran que los fenómenos educativos multicausales como la resiliencia no pueden comprenderse mediante variables aisladas, sino a través de configuraciones complejas caracterizadas por interdependencias entre factores socioeducativos, institucionales y territoriales. Esto refuerza el marco conceptual que concibe la resiliencia como un fenómeno emergente y contextual, no reducible a características individuales. Asimismo, el proyecto evidencia el valor de los modelos de aprendizaje automático como herramientas conceptuales que permiten explorar interacciones estructurales y patrones no lineales, ofreciendo capacidades explicativas además de predictivas. La integración entre ciencia de datos, teoría educativa y análisis territorial constituye una contribución significativa a enfoques sistémicos y contextualizados en educación.

Desde la perspectiva de la ciencia de datos, este proyecto demuestra el potencial del aprendizaje automático para abordar problemas educativos complejos mediante un proceso riguroso que integra calidad de datos, modelado supervisado, evaluación comparativa y despliegue de prototipos. Asimismo, los resultados muestran el valor de comparar modelos, pues cada uno aporta fortalezas distintas según su capacidad de generalización, interpretabilidad o desempeño. La selección del modelo debe responder al mejor balance entre rendimiento técnico, coherencia metodológica y utilidad práctica para la toma de decisiones, demostrando cómo los modelos pueden transformarse en herramientas operativas con valor real para la gestión, la analítica y la formulación de intervenciones.

En síntesis, este proyecto evidencia que el manejo integrado de volúmenes de datos provenientes de fuentes estandarizadas, administrativas, territoriales y educativas constituye un campo todavía poco explorado en la investigación educativa, pero con un enorme potencial transformador. La capacidad de relacionar millones de observaciones, múltiples capas de información y modelos avanzados de aprendizaje automático abre posibilidades inéditas para comprender fenómenos educativos de alta complejidad, superar las limitaciones de análisis tradicionales y diseñar intervenciones más precisas, equitativas y contextualizadas. Este trabajo demuestra que la ciencia de datos aplicada a la educación favorece el avance hacia sistemas educativos capaces de aprovechar la riqueza informacional disponible y orientar decisiones políticas y pedagógicas realmente fundamentadas en evidencia.

Capítulo 7

Trabajos Futuros

El desarrollo de este proyecto abre un conjunto amplio de posibilidades para continuar avanzando tanto en la comprensión del fenómeno de la resiliencia escolar como en la consolidación de la ciencia de datos como un campo emergente y estratégico dentro de la investigación educativa. A partir de los resultados obtenidos y de los retos metodológicos enfrentados, se identifican dos líneas principales de proyección: (1) trabajos futuros orientados directamente al fenómeno analizado, y (2) trabajos futuros enfocados en fortalecer las capacidades de la ciencia de datos en el sector educativo, más allá del caso específico de la resiliencia.

Trabajos futuros orientados directamente al fenómeno analizado

Futuras investigaciones pueden profundizar en el modelado avanzado mediante la incorporación de redes neuronales profundas, arquitecturas híbridas que integren información tabular con datos espaciales, modelos de boosting de nueva generación y grafos de aprendizaje para representar dependencias territoriales. Estas aproximaciones permitirían capturar patrones más complejos e interacciones de mayor nivel entre factores socioeducativos, institucionales y comunitarios, ampliando la capacidad explicativa y predictiva del sistema.

Otro eje de trabajo relevante consiste en fortalecer la validación temporal, territorial y contextual del modelo. Para ello, es necesario evaluarlo con nuevas cohortes de Saber 11, con el fin de medir su capacidad de generalización, su estabilidad frente a cambios contextuales y su comportamiento diferencial por zonas. La realización de pilotos con secretarías de educación e instituciones permitiría ajustar el modelo a condiciones operativas reales.

Dado que la resiliencia escolar tiene una importante dimensión territorial, también resulta pertinente avanzar en análisis espaciales más sofisticados. La incorporación de técnicas como autocorrelación espacial, modelos geoestadísticos, regresión geográficamente ponderada y grafos espaciales permitiría representar con mayor precisión las dinámicas regionales y comprender cómo influyen las condiciones del territorio en los niveles de resiliencia.

En términos de aplicación operativa, el prototipo desarrollado puede evolucionar hacia sistemas de analítica educativa escalables, con actualización en tiempo real, paneles longitudinales, alertas automáticas y mecanismos de priorización basados en umbrales dinámicos. La consolidación de estas herramientas fortalecería la toma de decisiones basada en evidencia y facilitaría su adopción

institucional.

Finalmente, futuros trabajos pueden integrar información cualitativa y variables socioemocionales como encuestas de bienestar, narrativas estudiantiles, reportes institucionales y datos de clima escolar para enriquecer la comprensión del fenómeno. De igual manera, evaluar el impacto de políticas educativas informadas por el modelo permitirá comprender cómo las predicciones influyen en la equidad y la permanencia, generando evidencia causal sobre la efectividad de las intervenciones derivadas del sistema predictivo.

Trabajos Futuros en Ciencia de Datos para la Educación

Más allá del estudio específico sobre resiliencia escolar, este proyecto abre oportunidades amplias para consolidar la ciencia de datos como un componente estratégico en la investigación y la gestión educativa. Una primera línea de trabajo consiste en el desarrollo de sistemas integrales de alerta temprana capaces de predecir riesgos educativos diversos, como deserción, bajo desempeño, ausentismo, tránsito a la educación superior. Estos sistemas no solo deben incorporar modelos de predicción, sino también mecanismos de interpretación, priorización y acción automatizada, permitiendo que los resultados se traduzcan en intervenciones oportunas y eficaces.

Otra proyección relevante es la exploración de representaciones de datos multimodales y multifuente. La educación genera una amplia variedad de registros textos, imágenes, datos de plataformas, secuencias temporales y relaciones entre actores educativos que usualmente no se integran de manera sistemática. El uso de técnicas de procesamiento de lenguaje natural, visión por computador, análisis de redes y aprendizaje multimodal permitiría obtener comprensiones más profundas de los procesos educativos, trascendiendo las limitaciones de los datos tabulares tradicionales.

Asimismo, investigaciones futuras pueden profundizar en el análisis del sesgo algorítmico y la equidad en las predicciones. El estudio de métricas de equidad algorítmica, la detección de sesgos entre grupos poblacionales y la implementación de mecanismos correctivos permitirá diseñar sistemas predictivos que no reproduzcan inequidades históricas, sino que contribuyan a su reducción. Este enfoque es crucial en países con altos niveles de desigualdad.

En paralelo, es indispensable fortalecer la infraestructura técnica que permita implementar analítica educativa a gran escala. Esto implica construir pipelines reproducibles, garantizar almacenamiento seguro y anonimizado, desarrollar APIs de consulta para instituciones y articular sistemas educativos como SIMAT, SISTEMA DE ALERTAS, APOYO ESCOLAR en un ecosistema robusto de datos. Esta infraestructura constituye un desafío persistente en Colombia y representa un campo estratégico para el desarrollo tecnológico y académico.

Finalmente, la ciencia de datos tiene un papel creciente en la formulación de políticas educativas basadas en evidencia. Modelos predictivos, simulación de escenarios territoriales, análisis prospecti-

vo de brechas y optimización de recursos pueden convertirse en herramientas estructurales para la toma de decisiones pública. La consolidación de estas capacidades permitirá avanzar hacia sistemas educativos más eficientes, equitativos y contextualizados, alineados con una visión moderna de la gestión educativa apoyada en datos.

Bibliografía

[1] H. Bayona Rodríguez and D. C. López Vera, "Factores asociados a la resiliencia académica: evidencia para colombia," Universidad de los Andes, Centro de Estudios sobre Desarrollo Económico (CEDE), Tech. Rep. 01, 2021, disponible en: http://hdl.handle.net/1992/48061. [Online]. Available: https://www.researchgate.net/publication/348842667_Factores_asociados_a_la_resiliencia_academica_evidencia_para_Colombia_Actors_Associated_with_Academic_Resilience_Evidence_for_Colombia

[2] Centro Nacional de Memoria Histórica, *La guerra en las aulas: impacto del conflicto armado en la educación en Colombia*. CNMH, 2021.

[3] Indepaz, "Los focos del conflicto en colombia: informe sobre presencia de grupos armados," 2021, disponible en: https://www.indepaz.org.co/wp-content/uploads/2021/10/INFORME-DE-GRUPOS-2021.pdf.

[4] OCDE, *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*. OCDE, 2019. [Online]. Available: https://doi.org/10.1787/acd78851-en

[5] Save the Children, "Impact of the armed conflict on children and youth," 2017, disponible en: https://resourcecentre.savethechildren.net/es/document/colombia-impact-of-the-armed-conflict-on-children-and-youth/.

[6] A. Gaviria and C. Vázquez, "Exposición al conflicto armado y logro académico," Universidad de los Andes, Bogotá, Colombia, Tech. Rep., 2013.

[7] A. Masten, "Ordinary magic: Resilience processes in development," *American Psychologist*, vol. 56, no. 3, pp. 227–238, 2001.

[8] A. Q. Barahona, E. B. Sosa, J. O. Torres, C. G. Gallardo, and C. B. Noygues, "Educación para la resiliencia, un análisis desde la perspectiva de niñas, niños y docentes," *Revista Latinoamericana de Estudios Educativos*, vol. 53, no. 1, pp. 155–178, 2023. [Online]. Available: https://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S2448-878X2023000100155

[9] . C. Moreno and F. C. Torres, "La escuela rural colombiana en medio del conflicto armado: un análisis desde la información publicada en noticias," *Andamios*, vol. 18, no. 47, pp. 493–518, 2021.

[10] R. Oyelere and K. Wharton, "The impact of conflict on education attainment and enrollment in colombia: Lessons from recent idps," Households in Conflict Network, Tech. Rep. 141, 2013. [Online]. Available: https://ideas.repec.org/p/hic/wpaper/141.html

[11] C. Bishop, *Pattern recognition and machine learning*. Springer, 2006. [Online]. Available: https://link.springer.com/book/9780387310732

[12] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, 1st ed. Springer, 2001. [Online]. Available: https://web.stanford.edu/~hastie/ElemStatLearn//

[13] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016. [Online]. Available: http://alvarestech.com/temp/deep/Deep%20Learning%20by%20Ian%20Goodfellow,%20Yoshua%20Bengio,%20Aaron%20Courville%20(z-lib.org).pdf

[14] C. Molnar, *Interpretable Machine Learning*. Leanpub, 2020. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[15] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[16] D. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. Wiley, 2000.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[18] B. Schölkopf and A. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[19] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[20] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[21] M. Suaza-Medina, R. Peñabaena-Niebles, and M. Jubiz-Diaz, "A model for predicting academic performance on standardized tests for lagging regions based on machine learning and shapley additive explanations," *Scientific Reports*, vol. 14, no. 25306, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-76596-3

[22] O. Olabanjo, A. Wusu, and M. Mazzara, "A machine learning prediction of academic performance of secondary school students using radial basis function neural network," *Technology in Education and Learning*, vol. 1, p. 100190, 2022.

[23] D. Zangana, A. Ibrahim, H. Yuan, and M. Amani-Beni, "Educational inequality in urban settings: A spatial analysis of school distribution and double-shift system challenges – a case study," *Journal of Urban Management*, vol. 13, no. 3, pp. 1–16, 2024.

[24] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth, "Crisp-dm 1.0: Step-by-step data mining guide," CRISP–DM Consortium, Tech. Rep., 2000. [Online]. Available: https://www.the-modeling-agency.com/crisp-dm.pdf

[25] DANE, "Normas técnicas de calidad para la producción estadística," 2018, disponible en: https://www.dane.gov.co.

[26] U. Bronfenbrenner, *The Ecology of Human Development*. Cambridge, MA: Harvard University Press, 1979.

[27] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020.

[28] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, and et al., "The fair guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, p. 160018, 2016.

[29] Centro Nacional de Memoria Histórica, "Mpios_creados_1993_2019," disponible en: https://serviciosgis.centrodememoriahistorica.gov.co.

[30] DANE, "Datos geoestadísticos de municipios 2023," disponible en: https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/datos-geoestadisticos/.