Maestría en

INTELIGENCIA ARTIFICIAL APLICADA

Trabajo previo a la obtención de título de
Magister en Inteligencia Artificial

AUTOR/ES:

Jenny Maribel Cuyo Cuyo

Jaime Andres Espinosa Jácome

Anderson David Collaguaso Chorlango

TUTOR/ES:

Fernanda Paulina Vizcaíno Imacaña

Karla Estefanía Mora Cajas

Modelo híbrido de Random Forest y Redes Neuronales para predecir el riesgo de pobreza multidimensional con datos del Censo y la encuesta ENEMDU del INEC

Quito - Ecuador

Julio – 2026

REINVENTEMOS EL FUTURO

Certificación de autoría

Nosotros, **Jenny Maribel Cuyo Cuyo, Jaime Andrés Espinosa Jácome, Anderson David Collaguaso Chorlango,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma
Jenny Maribel Cuyo Cuyo

Firma
Jaime Andrés Espinosa Jácome

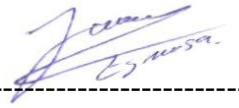
Firma
Anderson David Collaguaso Chorlango

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Jenny Maribel Cuyo Cuyo, Jaime Andrés Espinosa Jácome**, **Anderson David Collaguaso Chorlango** en calidad de autores del trabajo de investigación titulado Modelo híbrido de Random Forest, Gradient Boosting y Redes Neuronales para predecir el riesgo de pobreza multidimensional con datos de la encuesta ENEMDU del INEC, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, Agosto 2026

Firma
Jenny Maribel Cuyo Cuyo

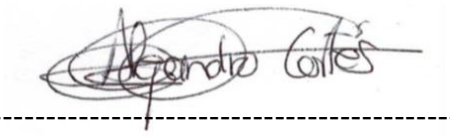


Firma
Jaime Andrés Espinosa Jácome

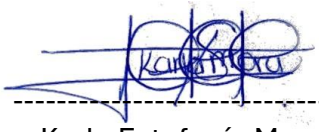
Firma
Anderson David Collaguaso Chorlango

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López Director EIG y Karla Estefanía Mora Cajas Coordinadora UIDE,** declaramos que: **Jenny Maribel Cuyo Cuyo, Jaime Andrés Espinosa Jácome**, **Anderson David Collaguaso Chorlango** son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López
Director de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial



Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Ciencia de datos y máquinas de aprendizaje con mención en Inteligencia Artificial

DEDICATORIA

A Dios, por permitirme alcanzar todas las metas que me propongo. A mi madre, Delfina Cuyo, por ser mi ejemplo de fortaleza, perseverancia y amor incondicional. A mi novio, Antonio Prieto, por acompañarme incondicionalmente en este camino, por su paciencia, comprensión y apoyo constante.

Jenny Maribel Cuyo Cuyo

Este logro se lo dedico a mi esposa Karol y mi hija Mila, porque son mi apoyo incondicional, inspiración y fortaleza que me impulsa a seguir creciendo y logrando nuevos sueños. Este logro también es tuyo.

Jaime Andrés Espinosa Jácome

A Dios, por guiarme y darme la fortaleza para alcanzar esta meta. A mis padres, Patricia y Luis, por su amor, apoyo incondicional y por enseñarme los valores que me han permitido llegar hasta aquí. A mi hermano Jonathan, por su compañía y cariño; y a mi compañera de vida Mayu, por compartir conmigo este camino y hacer de esta etapa una experiencia especial.

Anderson David Collaguaso Chorlango

AGRADECIMIENTOS

Agradezco a todas las mujeres que alzaron su voz, rompieron barreras y abrieron caminos, haciendo posible que hoy pueda acceder a oportunidades de crecimiento, desarrollo profesional, y construir con libertad mi propio camino.

Jenny Maribel Cuyo Cuyo

Agradezco profundamente a mi esposa, Karol, y a mi hija, Mila, por ser mi apoyo, mi inspiración y mi mayor motivación.

Jaime Andrés Espinosa Jácome

A mis padres, Patricia y Luis, por su esfuerzo y apoyo incondicional en cada paso de este proceso. A mi hermano Jonathan, por estar siempre presente alentándome. Y a mi compañera de vida Mayu, por su apoyo constante y por hacer más ligero este camino.

Anderson David Collaguaso Chorlango

RESUMEN

La pobreza multidimensional constituye uno de los desafíos estructurales más persistentes en Ecuador, con brechas significativas entre zonas urbanas y rurales y entre las distintas regiones naturales del país. Los enfoques estadísticos tradicionales presentan limitaciones para capturar las complejas interacciones no lineales entre variables como el nivel educativo, la localización geográfica y los indicadores socioeconómicos que determinan el riesgo de privación. La presente investigación propone el desarrollo, implementación y evaluación de un modelo híbrido de inteligencia artificial que integra Random Forest, Gradient Boosting y Redes Neuronales Artificiales para clasificar el riesgo de pobreza multidimensional en hogares ecuatorianos. El modelo se entrena y valida con microdatos oficiales de la Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU) del Instituto Nacional de Estadística y Censos (INEC) para el período 2018–2025, utilizando como variables predictoras principales el nivel educativo alcanzado y la zona geográfica de residencia, junto con covariables sociodemográficas complementarias. La metodología sigue un enfoque de clasificación supervisada que incluye preprocesamiento avanzado, balanceo de clases, optimización bayesiana de hiperparámetros y validación cruzada estratificada. El desempeño del modelo híbrido se evalúa mediante métricas robustas para datos desbalanceados (precisión, recall, F1-score, AUC-ROC (área bajo la curva ROC) y coeficiente de Gini) y se compara con modelos individuales de referencia (Random Forest, Gradient Boosting, Red Neuronal y regresión logística). Se incorpora un análisis de interpretabilidad mediante valores SHAP para identificar los predictores con mayor poder discriminante. Los resultados preliminares sugieren que el enfoque híbrido mejora la capacidad predictiva respecto a los modelos individuales y confirma que el nivel educativo y la zona geográfica son los determinantes más relevantes del riesgo de pobreza multidimensional. Este trabajo contribuye al campo de la inteligencia artificial aplicada a las ciencias sociales y aporta evidencia útil para la focalización de políticas públicas orientadas a la reducción de la pobreza en Ecuador.

Palabras Claves: pobreza multidimensional, modelo híbrido, Random Forest, Gradient Boosting, redes neuronales, machine learning, ENEMDU, nivel educativo, zona geográfica, SHAP, Ecuador.

ABSTRACT

Multidimensional poverty remains one of the most persistent structural challenges in Ecuador, with significant gaps between urban and rural areas and across the country's natural regions. Traditional statistical approaches have limitations in capturing the complex nonlinear interactions among variables such as educational attainment, geographic location, and socioeconomic indicators that determine the risk of deprivation. This research proposes the development, implementation, and evaluation of a hybrid artificial intelligence model that integrates Random Forest, Gradient Boosting, and Artificial Neural Networks to classify the risk of multidimensional poverty in Ecuadorian households. The model is trained and validated using official microdata from the National Survey of Employment, Unemployment, and Underemployment (ENEMDU) of the National Institute of Statistics and Census (INEC) for the period 2018–2025, employing educational attainment and geographic area of residence as the main predictor variables, along with complementary sociodemographic covariates. The methodology follows a supervised classification approach that includes advanced preprocessing, class balancing, Bayesian hyperparameter optimization, and stratified cross-validation. The hybrid model's performance is assessed using robust metrics for imbalanced data (precision, recall, F1-score, AUC-ROC, and Gini coefficient) and compared against individual benchmark models (Random Forest, Gradient Boosting, Neural Network, and logistic regression). An interpretability analysis using SHAP values is incorporated to identify the predictors with the greatest discriminative power. Preliminary results suggest that the hybrid approach improves predictive performance over individual models and confirms that educational attainment and geographic area are the most relevant determinants of multidimensional poverty risk. This work contributes to the field of artificial intelligence applied to the social sciences and provides useful evidence for targeting public policies aimed at poverty reduction in Ecuador.

Keywords: multidimensional poverty, hybrid model, Random Forest, Gradient Boosting, neural networks, machine learning, ENEMDU, educational attainment, geographic area, SHAP, Ecuador.

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

Capítulo 1. Introducción y planteamiento del problema

1.1. Introducción

La medición convencional de la pobreza se ha basado, tradicionalmente, en un criterio estrictamente monetario, que evalúa el bienestar de los hogares a partir de su nivel de ingreso o consumo. Bajo el enfoque neoclásico, el bienestar se entiende como la satisfacción del consumo, de modo que la pobreza queda definida como la carencia de los recursos monetarios necesarios para alcanzar un ingreso o consumo mínimo (Castillo Añazco y Jácome Pérez, 2017, p. 3). Sin embargo, este criterio no capta privaciones que no se derivan directamente del nivel de ingreso: Sen (2000, p. 116) sostiene que una carencia relativa de ingresos puede traducirse en una privación absoluta cuando el bienestar se mide en el espacio de las capacidades. Un hogar puede superar la línea de pobreza monetaria y, al mismo tiempo, carecer de acceso a agua segura, presentar abandono escolar en alguno de sus miembros o depender de empleo sin afiliación a la seguridad social como se menciona en (Castillo Añazco y Jácome Pérez, 2017, p. 9). Ante estas limitaciones, tanto en el ámbito internacional como en Ecuador se adoptó un enfoque multidimensional que identifica de forma simultánea las privaciones que afectan a los hogares en distintas dimensiones del bienestar (Castillo Añazco y Jácome Pérez, 2017, p. 5).

En el país, esta visión se concreta en el índice de pobreza multidimensional (IPM), que el instituto nacional de estadística y censos (INEC) calcula siguiendo la metodología desarrollada originalmente por Alkire y Foster (2011) y adaptada al caso ecuatoriano por Castillo Añazco y Jácome Pérez (2017, pp. 10-11). El indicador de pobreza multidimensional evalúa cuatro dimensiones: educación; trabajo y seguridad social; salud, agua y alimentación; y hábitat, vivienda y ambiente sano, desagregadas en doce indicadores (Castillo Añazco y Jácome Pérez, 2017, p. 11). Un hogar se clasifica como pobre multidimensional cuando acumula privaciones en un tercio o más de los indicadores ponderados (Castillo Añazco y Jácome Pérez, 2017, p. 17).

No obstante, medir la pobreza únicamente después de que esta se ha manifestado tiene un valor limitado para el diseño de política social, pues impide intervenir antes de que el hogar caiga efectivamente en dicha condición. Por ello, resulta más útil anticiparla, es decir, identificar con antelación a los hogares en riesgo de caer en pobreza multidimensional, de manera que la intervención pública pueda producirse antes de que la privación se consolide.

La presente investigación se propone construir, entrenar y evaluar un modelo híbrido que combina Random Forest (RF), Gradient Boosting (GBT) y Redes Neuronales Artificiales (RNA), con una regresión logística, con el fin de estimar la probabilidad de que una persona se encuentre clasificada en condición de pobreza multidimensional en el plazo de un año. El estudio centra su atención en el capital humano, sin dejar de lado las variables sociodemográficas, sociales y territoriales que permiten diferenciar las realidades urbanas y rurales.

1.2. Antecedentes y contexto del problema

La pobreza puede entenderse como la privación de los recursos y las capacidades necesarias para llevar una vida digna, ya sea por insuficiencia de ingresos, por carencia de acceso a servicios básicos (Alkire y Foster, 2011, p. 476). Se trata de un problema que, lejos de ser marginal, conserva una magnitud considerable a escala mundial: para 2021, alrededor de 736 millones de personas vivían en condición de pobreza extrema, es decir, una de cada diez personas en el planeta (PNUD, 2021, citado en López Idrovo y Sarmiento Moscoso, 2024, p. 171).

Esa magnitud global se refleja, con ciertas particularidades, en América Latina. De acuerdo con el Índice de Pobreza Multidimensional para América Latina (IPM-AL), elaborado conjuntamente por la Comisión Económica para América Latina y el Caribe (CEPAL) y el Programa de las Naciones Unidas para el Desarrollo (PNUD), se presenta la siguiente tabla

Tabla 1: *Pobreza por ingresos en países de la región, 2021*

País	***Pobreza por ingresos (2021)***
Colombia	*36.3%*
Bolivia	*31.2%*
Ecuador	*29.7%*
Argentina	*29.5%*
Perú	*25.1%*

Fuente: CEPAL (2021), citado en López Idrovo y Sarmiento Moscoso (2024, p. 171).

Al interior del país, el contraste territorial es uno de los rasgos más persistentes del fenómeno. Según el boletín técnico más reciente del INEC, la pobreza por necesidades básicas insatisfechas (NBI) se ubicó en 32,4% a nivel nacional, con una incidencia de 23,8% en el área urbana frente a 50,8% en el área rural (INEC, 2025, p. 7). Vivir en una zona rural, entonces, duplica la probabilidad de enfrentar privaciones básicas frente a vivir en una zona urbana, una brecha que se repite de forma consistente en la mayoría de los indicadores sociales del país.

Durante años, el estudio de la pobreza en Ecuador se apoyó principalmente en modelos econométricos como el Logit y el Probit: Morán Chiquito y Lozano (2017, pp. 41-42, 49-50) estimaron los condicionantes de la pobreza rural mediante un modelo Probit aplicado a la ENEMDU de 2014, mientras que López Idrovo y Sarmiento Moscoso (2024, p. 169) aplicaron un modelo Logit para analizar los determinantes regionales de la pobreza por ingresos entre 2007 y 2021, encontrando que la ubicación geográfica, el empleo y la educación son los factores con mayor peso explicativo.

Más cerca todavía del enfoque que persigue esta investigación existe un antecedente ecuatoriano que combina la pobreza multidimensional con la inteligencia artificial. Ochoa, Castro-García, Arias Pallaroso, Machado y Sucozhañay Machado (2021, pp. 220-224) compararon diez modelos de machine learning, para estimar el Índice de Pobreza Multidimensional a nivel de persona en el cantón Cuenca, y encontraron que Random Forest superó de forma consistente a los demás modelos, con un error de 7.5% en validación cruzada.

1.3. Planteamiento del problema

El problema que aborda esta investigación puede formularse en los siguientes términos: ¿es posible predecir, con un grado de precisión adecuado y preservando la interpretabilidad del modelo, qué hogares ecuatorianos se encontrarán en condición de pobreza multidimensional en el plazo de un año, a partir de los microdatos oficiales disponibles y considerando las diferencias estructurales que separan al ámbito rural del urbano?

1.4. Justificación e importancia

La inclusión de variables educativas en un modelo de predicción de pobreza es una decisión natural ya que responde a una relación ampliamente documentada en la literatura.

Salvador Benítez (2008, pp. 237-257) sostiene que el desarrollo humano depende del acceso a la educación en la misma medida que del acceso a la salud, la vivienda o el ingreso, y que la pobreza, más que una carencia material, es la negación de esas oportunidades. En la misma línea, Castillo Viveros y Rojas González (2023, pp. 108-130) documentan, desde un estudio cualitativo con jóvenes mexicanos en contextos de marginación, que la educación actúa como un catalizador de movilidad social cuando las condiciones estructurales lo permiten. Ambos trabajos respaldan la decisión de otorgar un peso particular a las variables de nivel educativo del hogar y de su jefe dentro del modelo propuesto, así como de diferenciarlas según la zona de residencia.

Desde lo científico, este trabajo aporta al campo de la inteligencia artificial aplicada a las ciencias sociales al poner a prueba la efectividad de un modelo de inteligencia artificial frente a un problema real, condiciones bajo las cuales muchos enfoques puramente teóricos suelen fallar.

Desde el ámbito social, un modelo capaz de identificar con mayor precisión qué hogares y territorios enfrentan mayor riesgo de pobreza multidimensional permite focalizar de manera más eficiente los programas de protección social, las políticas de educación inclusiva y las estrategias de desarrollo territorial. Esta capacidad de focalización adquiere, además, una dimensión de política pública más amplia: el estudio se articula directamente con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030, en particular, el ODS 1 (fin de la pobreza), el ODS 4 (educación de calidad) y el ODS 10 (reducción de las desigualdades).

En cuanto a su valor innovador, el presente trabajo permite comprender el peso real que tienen la educación y las condiciones territoriales en la predicción del riesgo. Hasta donde se ha podido establecer, no existe en la literatura ecuatoriana un estudio que combine Random Forest y redes neuronales para clasificar el riesgo de pobreza multidimensional utilizando microdatos de la ENEMDU de varios años juntos.

1.5. Alcance del proyecto

Esta sección delimita el alcance de la investigación, precisando sus dimensiones de cobertura, las fuentes de información empleadas y los límites asumidos en función del tiempo y los recursos disponibles.

1.5.1. Alcances específicos

- **Fuentes de información: microdatos de las Encuestas ENEMDU del INEC para el período 2018–2025**, según disponibilidad y completitud.

- **Variable dependiente:** identificación de pobreza multidimensional como variable binaria, considerando pobre a la persona u hogar cuyo IPM sea mayor o igual al 33 %, conforme a la metodología oficial.

- **Variables independientes principales:** características educativas de la persona y del jefe de hogar junto con variables sociodemográficas, sociales y territoriales asociadas al capital humano.

- **Metodología:** implementación completa de un modelo híbrido supervisado, con preprocesamiento, ingeniería y selección de características, balanceo de clases, optimización de hiperparámetros, entrenamiento, validación cruzada estratificada y evaluación con métricas robustas.

1.5.2. Límites del estudio

- No se contempla la implementación del modelo en un sistema de producción en tiempo real.

- No se realiza recolección primaria de datos ni trabajo de campo.

- El estudio se circunscribe al contexto ecuatoriano y no pretende generalizarse a otros países sin validación adicional.

- No se profundiza en el análisis causal ni en el impacto de otras dimensiones de la pobreza más allá de su uso como variables de apoyo para la predicción.

1.5.3. Flujo metodológico

El desarrollo de este trabajo sigue una secuencia de etapas que va desde la obtención de los microdatos oficiales hasta la construcción, evaluación e interpretación del modelo híbrido. Este ordenamiento no es solo una cuestión de presentación: garantiza que cada paso

pueda rastrearse hasta el anterior, que los resultados sean reproducibles, y que exista coherencia entre cómo se construye la variable objetivo, cómo se entrena el modelo y cómo se leen los hallazgos al final. En síntesis, el flujo de trabajo contempla las siguientes etapas:

1. **Fuentes de datos.** Carga de los archivos de persona y de vivienda/hogar de la ENEMDU (2018–2025), fijando al hogar como unidad de análisis.

2. **Master data table (MDT).** Integración de los doce indicadores del IPM, que constituyen la variable objetivo, junto con los bloques de variables educativas, territoriales, demográficas y de ingresos.

3. **Análisis exploratorio y preprocesamiento (EDA).** Limpieza y validación de los datos, tratamiento de valores faltantes y atípicos, estadística descriptiva, construcción de la variable objetivo e ingeniería de características.

4. **Modelos base.** En paralelo se entrenan dos modelos: un Random Forest (Etapa 3A), que entrega probabilidades e importancia de variables, y una red neuronal tipo perceptrón multicapa (Etapa 3B), que aporta probabilidades y capta patrones no lineales que el primero podría dejar de lado.

5. **Stacking ensemble.** Las probabilidades que entregan los dos modelos base, junto con las características originales, se combinan como entradas del modelo de stacking, encargado de producir la clasificación final. Los resultados de este metamodelo se interpretan posteriormente mediante explicaciones aditivas de Shapley (conocido por sus siglas en ingles, Shapley additive explanations SHAP), tanto a nivel global como local.

6. **Resultados esperados.** Se espera obtener un modelo de stacking cuya capacidad predictiva, sea superior a la de los modelos base (Random Forest y red neuronal) evaluados de manera individual. Asimismo, se busca identificar, mediante los valores SHAP, las variables con mayor peso predictivo en cada estrato geográfico (urbano y rural), así como comparar la contribución relativa de los indicadores educativos frente a los territoriales en la explicación del riesgo de pobreza multidimensional. Finalmente, se pretende traducir estos hallazgos en criterios concretos de focalización, tales como

perfiles de riesgo o umbrales diferenciados por zona, que sirvan de insumo para la priorización de programas de política social.

1.6. Objetivos

1.6.1. Objetivo general

Implementar un modelo híbrido de Random Forest, Gradient Boosting y Redes Neuronales Artificiales para clasificar el riesgo de pobreza multidimensional en hogares ecuatorianos a un año vista, analizando el efecto diferencial de las características educativas de los miembros del hogar y del jefe de hogar como variables predictoras, mediante el uso de los microdatos de la ENEMDU del INEC para el período 2018 - 2025.

1.6.2. Objetivos específicos

1. Recopilar, depurar y procesar los microdatos de la ENEMDU del INEC para los cortes 2018–2025, construyendo un conjunto de datos estructurado, reproducible y con la variable objetivo de pobreza multidimensional binaria calculada según la metodología oficial del INEC.

2. Realizar un EDA para identificar la distribución de la pobreza multidimensional según nivel educativo, zona geográfica y características sociodemográficas y sociales.

3. Diseñar, entrenar y evaluar modelos de stacking que combinen RF y una RNA con una regresión logística como meta-clasificador, para la clasificación del riesgo de pobreza multidimensional un año en el futuro.

4. Desarrollar modelos diferenciados para los estratos urbano y rural, con el fin de reconocer las heterogeneidades territoriales de la pobreza multidimensional en Ecuador.

5. Discutir los resultados del modelo híbrido mediante el análisis de valores SHAP, identificando los predictores educativos, sociodemográficos y territoriales con mayor poder discriminante.

1.7. Hipótesis

El riesgo de que un hogar ecuatoriano se encuentre en situación de pobreza multidimensional está determinado, de manera significativa, por su nivel educativo, la zona geográfica de residencia y un conjunto de características sociodemográficas y sociales. En consecuencia, se espera que los hogares con menor nivel educativo y ubicados en contextos territoriales menos favorecidos presenten una mayor probabilidad de encontrarse en condición de pobreza multidimensional durante el período de estudio.

Capítulo 2. Marco teórico y estado del arte

2.1. Marco conceptual

2.1.1. La pobreza multidimensional

La medición de la pobreza puede descomponerse en dos operaciones distintas: identificar quién se encuentra en situación de pobreza, y agregar esa información en un indicador único para el conjunto de la población. Durante buena parte del siglo XX, la identificación de la pobreza se resolvió con un único criterio: el nivel de ingreso. Bajo este enfoque, se consideraba pobre a toda persona cuyo ingreso se ubicará por debajo de una línea de pobreza determinada, y no pobre a quien la superará. Bajo el enfoque del bienestar, la pobreza queda definida como la carencia de los recursos monetarios necesarios para alcanzar un ingreso o consumo mínimo (Castillo Añazco y Jácome Pérez, 2017, p. 3).

Este criterio resulta cómodo de calcular, pero deja fuera una parte considerable del fenómeno. Un hogar puede superar la línea de pobreza monetaria y, al mismo tiempo, carecer de acceso a agua segura, presentar abandono escolar en alguno de sus miembros, o depender de un empleo sin ninguna afiliación a la seguridad social (Castillo Añazco y Jácome Pérez, 2017, p. 6). Ante estas limitaciones, tanto en el ámbito internacional como en Ecuador se adoptó un enfoque multidimensional que abandona la idea de una sola forma de medir y la reemplaza por una acumulación de privaciones simultáneas en las dimensiones que una sociedad considera esenciales para una vida digna (Castillo Añazco y Jácome Pérez, 2017, p. 3).

La metodología más extendida para llevar esta idea a la práctica y la cual fue adoptada oficialmente en Ecuador, es la desarrollada por Alkire y Foster (2011). Su método conocido como de "doble corte", identifica primero a cada persona y en cuáles indicadores está privada respecto a un umbral mínimo aceptable en cada uno. A continuación, aplica un segundo corte, que determina cuántas privaciones simultáneas debe acumular una persona para ser considerada pobre en el sentido multidimensional. Este segundo corte es lo que distingue al método de dos alternativas más simples y menos satisfactorias: el criterio de unión, que clasifica como pobre a cualquiera con al menos una privación, y el criterio de intersección, que exige privación en absolutamente todos los indicadores. El "doble corte" ofrece un punto

intermedio calibrable, que en Ecuador se fija en un tercio de los indicadores ponderados para la pobreza moderada, y en la mitad para la pobreza extrema (INEC, 2026, p. 15).

A partir de esta identificación se construyen dos componentes que combinados, dan lugar al índice oficial. La tasa de incidencia mide qué proporción de la población es pobre; la intensidad mide, entre quienes son pobres, qué porcentaje promedio de los indicadores ponderados tienen en privación. El Índice de Pobreza Multidimensional (IPM) es, en esencia, el producto de estos dos componentes: no solo cuenta cuántas personas son pobres, sino qué tan profundamente lo son. Esta doble sensibilidad de la extensión y de la intensidad de la pobreza es la razón principal por la que el enfoque multidimensional se prefiere sobre una simple tasa de incidencia, que trataría por igual a una persona apenas por debajo del umbral y a otra con privaciones severas en casi todos los indicadores.

En Ecuador, el IPM se construye sobre cuatro dimensiones con igual peso entre sí (25% cada una): educación; trabajo y seguridad social; salud, agua y alimentación; y hábitat, vivienda y ambiente sano, desagregadas en doce indicadores específicos (INEC, 2026, p. 15; Castillo Añazco y Jácome Pérez, 2017, p. 11). La Tabla 1 reproduce esta estructura oficial de pesos, tal como aparece en el boletín técnico del INEC.

Tabla 2: *Dimensiones, pesos e indicadores del IPM ecuatoriano*

Dimensión (peso)	Indicador	Población aplicable
Educación (25%)	Inasistencia a educación básica y bachillerato	5 a 17 años
	No acceso a educación superior por razones económicas	18 a 29 años
	Logro educativo incompleto	18 a 64 años
Trabajo y Seguridad Social (25%)	Empleo infantil y adolescente	5 a 17 años
	Desempleo o empleo inadecuado	18 años y más
	No contribución al sistema de pensiones	15 años y más
Salud, Agua y Alimentación (25%)	Pobreza extrema por ingresos	Población total
	Sin servicio de agua por red pública	
Hábitat, Vivienda y Ambiente Sano (25%)	Hacinamiento	Población total
	Déficit habitacional	
	Sin saneamiento de excretas	
	Sin servicio de recolección de basura	

Fuente: INEC (2026, p. 15), Boletín Técnico N° 17-2025-ENEMDU.

Para dimensionar la magnitud del fenómeno que esta investigación busca predecir, conviene revisar las cifras más recientes publicadas por el INEC. Según el Boletín Técnico N° 17-2025-ENEMDU (INEC, 2026), con datos de diciembre de 2025, la pobreza por ingresos a nivel nacional se ubicó en 21,4%, con una brecha territorial marcada: 13,8% en el área urbana frente a 37,6% en el área rural (INEC, 2026, p. 5). La pobreza extrema por ingresos, en el mismo corte, fue de 8,3% a nivel nacional, con 3,0% en zona urbana y 19,7% en zona rural (INEC, 2026, p. 5).

Sin embargo, la brecha se amplía de forma considerable cuando se mide con el criterio multidimensional en lugar del monetario. La Tasa de Pobreza Multidimensional (TPM) nacional para diciembre de 2025 fue de 41,7%, con una incidencia de 29,9% en el área urbana frente a 66,9% en el área rural (INEC, 2026, p. 9).

2.1.2. El capital humano y el papel de la educación

Una de las explicaciones más empleadas en la literatura económica respecto a qué es la pobreza multidimensional, se remite al capital humano, definido como: el conjunto de conocimientos, habilidades y capacidades que una persona acumula a lo largo de su vida, principalmente a través de la educación. Becker en (Becker, 1964) formalizó esta idea al tratar la educación como una decisión de inversión, análoga a la inversión en capital físico, cuyo retorno se materializa en mayor productividad, mejores salarios y más estabilidad laboral a lo largo de la vida de una persona.

Trasladada al estudio de la pobreza, esta teoría explica por qué el nivel educativo aparece de forma tan persistente como determinante estructural en la literatura, cada año adicional de escolaridad amplía el abanico de empleos accesibles, mejora la capacidad de un hogar para afrontar cambios económicos bruscos, y modifica las perspectivas de la siguiente generación. Salvador Benítez en (Benitez, 2008, pp. 237-257) sintetiza esta relación al señalar que el desarrollo humano depende del acceso a la educación en la misma medida que del acceso a la salud, la vivienda o el ingreso, y que la pobreza es, ante todo, la negación de esas oportunidades.

Ahora bien, saber qué variables, es solo la mitad del problema de esta investigación. La otra mitad es cómo traducir ese conocimiento teórico en un modelo capaz de predecir, con

datos reales, qué hogares corren mayor riesgo de caer en pobreza. Para eso, el marco conceptual debe salir del terreno de la teoría económica y entrar en el de las herramientas computacionales que permiten operacionalizar esa predicción, que es el tema de las secciones que siguen.

2.1.3. Machine Learning

En (Mitchel, 1997), se define al aprendizaje automático, o machine learning, como una rama de la inteligencia artificial que le enseña a una computadora a resolver una tarea a partir de ejemplos, en lugar de darle todas las reglas escritas de antemano. En otras palabras, en vez de programar paso a paso qué hacer en cada caso, se le muestran al sistema muchos casos ya resueltos para que él mismo descubra el patrón que los explica (Mitchell, 1997).

En un problema de aprendizaje supervisado, como el de esta investigación, esos casos ya resueltos son un historial de hogares de los que se conoce todo: tanto sus características, por ejemplo el nivel educativo o la zona donde viven, como lo que realmente les ocurrió. El algoritmo revisa ese historial y aprende, poco a poco, cómo esas características se relacionan con el resultado. Una vez que aprende ese patrón, puede aplicarlo a hogares nuevos cuyo resultado todavía no se conoce, y así generar una predicción (Hastie, Tibshirani y Friedman, 2009).

Esa es, precisamente, la lógica que define la variable objetivo de este estudio: como lo que interesa es predecir si un hogar caerá o no en pobreza multidimensional dentro de un año, el resultado que el algoritmo debe aprender a predecir tiene solo dos opciones posibles, pobre o no pobre. Por eso, se trata específicamente de un problema de clasificación supervisada.

Antes de que cualquier algoritmo pueda aprender esa relación, hay un problema que resolver primero. De las decenas de variables disponibles en la base construida, no todas aportan información genuina sobre el riesgo futuro de pobreza; algunas, de hecho, contienen de forma encubierta la propia información que define la variable objetivo.

2.1.4. Selección de variables mediante Information Value

El Information Value (IV) es una técnica ampliamente usada para la selección de variables antes de la etapa de modelado, particularmente en el ámbito del scoring de crédito y

áreas afines (Rojas, Alvarez y Rojas, 2023, p. 1). La Tabla 3 presenta el umbral de interpretación convencional recogido por Rojas et al. (2023, Tabla 1).

Tabla 3: *Umbral de interpretación del IV*

Rango de IV	***Interpretación***
< 0,02	*No útil para la predicción*
0,02 – 0,1	*Poder predictivo débil*
0,1 – 0,3	*Poder predictivo medio*
0,3 – 0,5	*Poder predictivo fuerte*
> 0,5	*Sospechosamente alto — posible fuga de información (leakage)*

Fuente: elaboración propia con base en Rojas, Alvarez y Rojas (2023, Tabla 1, p. 1).

Depurado así el conjunto de variables, queda por decidir con qué algoritmo entrenar el modelo.

2.1.5. Random Forest

Random Forest, propuesto por (Breiman, 2001), es un método de ensamble que construye un número elevado de árboles de decisión: cada uno se entrena sobre una muestra distinta de los datos y, en cada división, evalúa solo un subconjunto aleatorio de variables. Las predicciones de todos los árboles se combinan después por votación mayoritaria. El autor en (Breiman, 2001, p. 5) demuestra que, a medida que aumenta el número de árboles, el error de generalización del bosque converge casi con seguridad a un límite fijo, un resultado que explica por qué el modelo no tiende a sobreajustarse simplemente por agregar más árboles.

Formalmente, si se dispone de B árboles entrenados $T_1, T_2, \dots, T_B$, la predicción final del bosque para una observación x se obtiene mediante votación mayoritaria:

$$\hat{y} = \text{moda}\{T_b(x) : b = 1, \dots, B\}$$

Cada árbol, a su vez, elige sus divisiones buscando reducir la impureza de los nodos resultantes. La usual es el índice de Gini, que para un nodo con una proporción de clases $p_k$se calcula como:

$$\text{Gini} = 1 - \sum_{k} p_k^2$$

Un valor de Gini cercano a cero indica un nodo prácticamente puro, mientras que valores cercanos a 0,5 indican una mezcla equilibrada de clases, es decir, una división poco informativa.

El árbol usa este índice como criterio de decisión en cada división: evalúa el Gini resultante de cada corte posible y elige la variable y el punto de corte que más lo reduzca, de modo que cada nueva división separe mejor las clases que la anterior. Este índice no es una invención propia de los árboles de decisión: proviene del coeficiente de Gini, desarrollado originalmente por el estadístico italiano Corrado Gini para medir la desigualdad en la distribución del ingreso, y fue adaptado como medida de impureza para árboles de clasificación por (Breiman, Friedman, Olshen y Stone, 1984) en el algoritmo CART (árboles de clasificación y regresión, por sus siglas en inglés), la base metodológica sobre la que se construye Random Forest.

Para este estudio, su virtud práctica es que maneja con naturalidad variables heterogéneas y relaciones no lineales, sin exigir los supuestos distribucionales que sí demanda un modelo Logit o Probit clásico.

2.1.6. Gradient Boosting

Gradient Boosting, propuesto por Friedman (2001), sigue una lógica distinta: construye los árboles de forma secuencial, y cada nuevo árbol se entrena específicamente para corregir los errores que dejó el conjunto de árboles anteriores. Friedman (2001, p. 1189) enmarca este procedimiento como un problema de optimización en el espacio de funciones, y no en el espacio de parámetros: en cada paso, el algoritmo avanza en la dirección que más reduce el error, de forma parecida a como el descenso de gradiente ajusta los parámetros de un modelo estadístico convencional.

Esta lógica correctiva puede expresarse como una actualización aditiva del modelo: en la iteración m, el modelo acumulado $F_{m-1}(x)$se corrige agregando un nuevo árbol $h_m(x)$, ajustado sobre los errores residuales del modelo anterior, ponderado por una tasa de aprendizaje ν:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)$$

Cuanto menor es ν, más conservador es cada paso de corrección, lo que suele mejorar la generalización del modelo a costa de requerir un mayor número de iteraciones para alcanzar un buen ajuste.

Esta naturaleza correctiva y secuencial suele traducirse en mayor poder predictivo que un bosque aleatorio equivalente, y esta tesis lo confirma con datos propios.

2.1.7. Redes Neuronales Artificiales

El tercer algoritmo que compone el modelo híbrido de esta tesis responde a un paradigma distinto al de los dos anteriores: en lugar de particionar el espacio de variables mediante reglas de decisión secuenciales, las redes neuronales artificiales aproximan la relación entre las variables de entrada y la variable de salida mediante una composición de transformaciones lineales y no lineales, organizadas en capas de unidades de procesamiento interconectadas denominadas neuronas artificiales, inspiradas de forma simplificada en el funcionamiento de las neuronas biológicas (McCulloch y Pitts, 1943). Cada neurona pondera las señales que recibe de la capa anterior, les aplica una función de activación no lineal y transmite el resultado a la capa siguiente, de modo que la red en su conjunto puede aproximar relaciones funcionales más complejas que las que captura un único árbol de decisión, una propiedad respaldada por el teorema de aproximación universal (Hornik, Stinchcombe y White, 1989). Esta arquitectura, conocida como perceptrón multicapa, se volvió entrenable a partir de la formulación del algoritmo de retropropagación del error (Rumelhart, Hinton y Williams, 1986, p. 533), que ajusta iterativamente los pesos de las conexiones de la red con el fin de minimizar la diferencia entre la salida que produce el modelo y la salida esperada.

La transformación que ocurre en cada capa l de la red puede expresarse como:

$$a^{(l)} = f\left(W^{(l)}a^{(l-1)} + b^{(l)}\right)$$

donde $W^{(l)}$y $b^{(l)}$son los pesos y sesgos de la capa, y $f$es una función de activación no lineal. En la capa de salida, se emplea la función sigmoide, que convierte cualquier valor real en una probabilidad entre 0 y 1:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Su contrapartida es que suele comportarse como una caja negra: a diferencia de un árbol de decisión, cuyas reglas de división pueden inspeccionarse directamente, es difícil saber, sin herramientas adicionales, qué peso tuvo cada variable en una predicción concreta de la red. Esta limitación no es exclusiva de las redes neuronales, también aplica a Random Forest y Gradient Boosting, pero es aquí donde se vuelve más evidente; de esta manera, se presenta SHAP como solución común a los tres algoritmos.

2.1.8. Stacking, optimización de hiperparámetros y modelos híbridos

El apilamiento, formalizado por Wolpert (1992), entrena un modelo adicional cuya única tarea es aprender la mejor manera de combinar las predicciones de esos tres modelos base, en lugar de promediarlas con una regla fija. Wolpert (1992, p. 241) lo describe como un esquema para reducir el error de generalización, deduciendo los sesgos particulares de cada modelo base y corrigiéndolos en una segunda etapa de aprendizaje. La idea de fondo es que distintos algoritmos cometen distintos tipos de error, de modo que combinarlos de forma aprendida suele rendir más que cualquiera por separado.

Cada uno de los tres modelos base debe estar correctamente configurado. Random Forest necesita definir, entre otras cosas, cuántos árboles construir y qué tan profundos deben ser; Gradient Boosting depende de su tasa de aprendizaje y del número de iteraciones; la Red Neuronal, de su arquitectura. Buscar manualmente la mejor combinación de estos valores resulta costoso cuando el espacio de opciones crece, y más aún cuando esa búsqueda debe repetirse para cada modelo. Para resolver esto de forma sistemática, este trabajo recurre a Optuna (Akiba et al., 2019), un framework que implementa el algoritmo TPE (Tree-structured Parzen Estimator), originalmente propuesto por Bergstra, Bardenet, Bengio y Kégl (2011). En lugar de evaluar el espacio de búsqueda de forma exhaustiva, el TPE construye un modelo probabilístico de qué combinaciones de hiperparámetros son más prometedoras a partir de las evaluaciones ya realizadas, y concentra ahí la búsqueda siguiente (Akiba et al., 2019, p. 2624).

Con los tres modelos base ya optimizados de esta manera, el modelo de regresión logística puede entonces cumplir su función: aprender la combinación que mejor predice el riesgo de pobreza multidimensional.

2.1.9. Interpretabilidad con valores Shapley Additive exPlanations (SHAP)

Cuando un modelo se entrena y optimiza ofrece una predicción adecauda, pero esa precisión no basta cuando el objetivo final es orientar política pública. Como se advirtió, al hablar de las redes neuronales, un modelo de alto desempeño puede seguir siendo una caja negra: quien deba tomar decisiones de política social necesita saber no solo qué hogares están en riesgo, sino por qué el modelo los clasificó de esa manera.

Los valores SHAP, propuestos por Lundberg y Lee (2017), resuelven este problema apoyándose en un concepto de la teoría de juegos cooperativos: el valor de Shapley, formulado originalmente por Shapley (1953, pp. 307-317) para repartir de forma equitativa la ganancia total de una coalición de jugadores entre sus miembros, según la contribución marginal de cada uno. Trasladado al aprendizaje automático, cada variable predictora se trata como un "jugador" y la predicción del modelo como la "ganancia" que debe repartirse entre ellas.

Aplicado a esta tesis, SHAP permite responder con evidencia empírica una pregunta que hasta ahora solo se había planteado en términos teóricos: cuánto pesa realmente la educación en el riesgo de pobreza multidimensional, a partir del argumento de Becker (1964), pero ahora contestada con la precisión de un modelo entrenado sobre datos reales de los hogares ecuatorianos.

2.2. Estado del arte

Una vez establecido el marco conceptual, corresponde examinar ahora en qué medida este terreno ha sido investigado previamente. Ninguno de los conceptos ni de las técnicas presentados en la sección anterior es original de esta investigación. Por un lado, la metodología Alkire-Foster cuenta con más de una década de aplicación en Ecuador. Por otro lado, Random Forest y las redes neuronales ya han sido empleados para estimar la pobreza en el país, mientras que el apilamiento ha demostrado su utilidad en contextos comparables fuera de la región. Sin embargo, hasta donde ha sido posible establecer, no se ha identificado en la literatura revisada un estudio que combine estos elementos tal como lo propone la presente investigación.

2.2.1. Fundamentos de la medición multidimensional

El punto de partida obligado de esta revisión es el trabajo de Alkire y Foster (2011), que propone la metodología para medir la pobreza multidimensional. Su aporte central no es menor: en lugar de medir la pobreza a partir de una sola variable, la evalúa desde varias dimensiones del bienestar de forma simultánea, lo que permite conocer tanto cuántas personas son pobres como con qué intensidad lo son.

En el caso ecuatoriano, Castillo Añazco y Jácome Pérez (2017) presentan la metodología oficial adoptada por el INEC para medir la pobreza multidimensional del país. Apoyándose directamente en el enfoque de Alkire-Foster, construyen un índice de cuatro dimensiones y doce indicadores, y establecen que un hogar se clasifica como pobre multidimensional cuando presenta privaciones en un tercio o más de los indicadores ponderados. Entre sus resultados, documentan que la pobreza multidimensional en Ecuador pasó de 51.5% en 2009 a 35% en 2015, mientras que el Índice de Pobreza Multidimensional se redujo de 27.2 a 17 puntos en el mismo período. Este antecedente resulta decisivo para la presente investigación, ya que define, de manera exacta, la forma oficial de construir la variable objetivo. Sin embargo, esta metodología describe la pobreza únicamente después de que esta ya se manifestó; corresponde entonces revisar cómo el aprendizaje automático ha sido empleado en Ecuador para anticiparla antes de que ocurra.

2.2.2. Machine learning aplicado a la pobreza en Ecuador

En cuanto al uso de aprendizaje automático aplicado específicamente a la pobreza ecuatoriana, (Ochoa Guaraca et al., 2021) desarrollaron uno de los estudios más cercanos al enfoque de esta tesis. Su objetivo fue estimar el IPM a nivel de persona en un territorio específico del país mediante modelos de aprendizaje automático, para lo cual compararon diez algoritmos distintos. RF obtuvo el desempeño más destacado, con un error de 7,5% en validación cruzada y de 7,48% en el conjunto de prueba; al contrastar el índice estimado con aproximaciones estadísticas del área, la diferencia promedio respecto a los valores oficiales fue de apenas 1%. El trabajo demuestra de forma convincente la utilidad de Random Forest para esta tarea, aunque su alcance fue territorialmente limitado y no incorporó redes neuronales ni estrategias de ensamble de segundo nivel.

Más recientemente, y de manera particularmente relevante para el diseño de esta investigación, Morales Torres (2025) analiza el uso de redes neuronales artificiales para medir y predecir la pobreza multidimensional en Ecuador con microdatos de la ENEMDU de 2024. El autor construye el IPM mediante la metodología Alkire-Foster y entrena un perceptrón multicapa sobre una muestra depurada de 65.979 hogares, incorporando preprocesamiento de datos, balanceo de clases, validación cruzada e interpretabilidad mediante valores SHAP. El modelo alcanzó un desempeño de ROC-AUC cercano a 0,91 y evidenció que la probabilidad de pobreza emerge de interacciones no lineales entre el ingreso per cápita, la educación, las condiciones de vivienda y la estructura del hogar. Este antecedente confirma la utilidad de las redes neuronales para esta tarea específica, pero se limita a un único año de la ENEMDU y no desarrolla un modelo híbrido que combine Random Forest con redes neuronales, que es precisamente el vacío que ocupa la presente tesis.

2.2.3. Modelos de ensamble, híbridos e interpretabilidad

En el plano internacional, Bujiku Sende et al. (2025) analizaron longitudinalmente dos cortes de datos de Tanzania (2014/15 y 2020/21), comparando un amplio conjunto de algoritmos para clasificar el estatus de pobreza multidimensional a nivel de hogar. Su hallazgo principal es que las arquitecturas de ensamble ofrecen mejoras consistentes frente a los clasificadores individuales, un resultado directamente pertinente para esta investigación porque aborda el mismo problema de clasificación mediante una combinación de algoritmos análoga a la que se propone.

Por último, y dentro del propio contexto ecuatoriano, Uquillas y Cruz (2025) proponen un modelo híbrido de redes neuronales para estimar pobreza y desigualdad de ingresos desde una perspectiva espacio-temporal, combinando información económica, imágenes satelitales y series de tiempo mediante redes convolucionales y recurrentes. Entre sus hallazgos destacan variables asociadas a la inversión extranjera, el consumo de los hogares, la productividad agrícola, la infraestructura y la actividad financiera. Aunque este estudio no se centra en el IPM calculado con la ENEMDU sino en indicadores de ingreso y desigualdad, confirma que el uso de modelos híbridos de inteligencia artificial para analizar pobreza en Ecuador ya está en funcionamiento.

2.2.4. Síntesis y vacío en la literatura

En conjunto, los antecedentes revisados en este capítulo permiten sostener tres ideas que orientan directamente el diseño metodológico de esta investigación.

- La primera es que la metodología Alkire-Foster, junto con su adaptación oficial por el INEC, constituye la base indiscutible para medir la pobreza multidimensional en Ecuador.

- La segunda es que Random Forest, Gradient Boosting y las redes neuronales han demostrado, cada uno por su lado, ser herramientas eficaces para estimar y predecir la pobreza con datos oficiales, y que los modelos de ensamble superan de forma consistente a los clasificadores individuales cuando se combinan.

- La tercera es que el nivel educativo y la zona geográfica aparecen, una y otra vez, como determinantes estructurales de las privaciones en el país.

Con todo, persiste un vacío concreto en la literatura ecuatoriana: no se ha identificado un estudio que combine Random Forest, Gradient Boosting y Redes Neuronales Artificiales mediante una estrategia de apilamiento para clasificar el riesgo de pobreza multidimensional, empleando microdatos de la ENEMDU de varios años, con modelos diferenciados por zona urbana y rural e interpretabilidad mediante SHAP. Es exactamente en ese espacio donde se inserta la contribución original de la presente investigación.

Capítulo 3. Metodología

3.1. Flujo metodológico

El desarrollo de esta investigación se organiza en una secuencia de etapas que va desde la obtención de los microdatos oficiales hasta la construcción, evaluación e interpretación del modelo híbrido final. Este encadenamiento no es solo una cuestión de orden expositivo: garantiza la trazabilidad del proceso, la reproducibilidad de los resultados, y la coherencia entre cómo se construye la variable objetivo, cómo se entrena el modelo y cómo se interpretan los hallazgos al final. Las ocho etapas del flujo son las siguientes:

Figura 1: *Flujo metodológico general - Etapas ordenadas del proceso de modelado de la pobreza multidimensional*

Las secciones que siguen desarrollan en detalle las primeras etapas de este flujo; las etapas de entrenamiento, ensamble e interpretación se retoman en los capítulos posteriores.

3.2. Recopilación, depuración y construcción del conjunto de datos

El primer objetivo específico de esta investigación no se limita a la obtención de datos: consiste en construir, de manera trazable y reproducible, la base de datos única que sustenta el análisis exploratorio y el modelo de apilamiento. En un proyecto que integra ocho años de encuestas, un cálculo oficial de pobreza y una proyección hacia el futuro, resulta fácil perder

de vista el origen de cada cifra. Por esta razón, el presente capítulo reconstruye, de manera sistemática, el proceso que conduce desde el microdato original del INEC hasta la base de datos que finalmente entrena el modelo.

De forma esquemática, este proceso consta de los pasos que resume la Figura 2: la descarga de los microdatos oficiales, la validación de su completitud, el cálculo de la pobreza multidimensional de cada hogar, la construcción de una dimensión temporal que permite proyectar el riesgo a un año, y la consolidación de estos elementos en una única base de modelado.

Figura 2: *Cadena de transformación de los datos, de la fuente oficial a la base de modelado*

3.2.1. Fuente de datos: la Encuesta ENEMDU del INEC

La totalidad de la información contenida en la base final proviene de la ENEMDU, instrumento que el Instituto Nacional de Estadística y Censos (INEC) levanta con periodicidad mensual y publica de acceso público (INEC, 2026). Esta fuente se seleccionó dado que es la única que reúne simultáneamente tres condiciones indispensables para el diseño de esta investigación: cobertura nacional, periodicidad mensual y un diseño muestral de panel rotatorio, mediante el cual una fracción de los hogares encuestados vuelve a ser entrevistada doce meses después.

Esta última característica es la que permite construir, en una etapa posterior, una variable objetivo proyectada a un año: sin un diseño de panel no existiría manera de determinar si un hogar que no es pobre en el presente lo será dentro de doce meses, puesto que no habría un registro del mismo hogar en un momento futuro con el cual establecer la comparación.

3.2.2. Descarga automatizada y disponibilidad de los datos

El portal de datos del INEC no ofrece una descarga masiva: cada archivo mensual se obtiene navegando un formulario web. Replicar ese recorrido a mano para ocho años de historia no era razonable, así que se automatizó mediante un script que reproduce la misma secuencia de pasos, garantizando que ningún archivo se edite manualmente entre lo que entrega el INEC y lo que entra al proceso de depuración.

Este proceso permitió recolectar 112 archivos correspondientes a 56 períodos mensuales entre 2018 y 2026. No todos los meses existen en el servidor del INEC: el instituto suspendió por completo el levantamiento durante 2021 por la pandemia y no existe información para los meses de julio desde el 2022. La siguiente tabla documenta estas ausencias, que no son una falla del proceso de descarga sino huecos reales de disponibilidad pública.

Tabla 4: *Períodos disponibles de la ENEMDU (2018-2026)*

Año	Meses disponibles	Total	Motivo de la ausencia
2018	Mar, jun, sep, dic	4	El INEC publicaba cortes trimestrales
2019	Mar, jun	2	Sep y dic no están disponibles
2020	Sep, oct, dic	3	Encuesta reducida por inicio de pandemia
2021	—	0	Año sin levantamiento publicado
2022	Feb-jun, ago-dic	10	Julio no se levantó
2023 – 2025	Ene-jun, ago-dic	11	
2026	Ene-abr	4	Últimos cortes al cierre del estudio

Como resultado del proceso de recolección descrito, se obtuvo el número de registros mensuales disponibles por año, correspondientes a los 112 archivos descargados del portal del INEC entre 2018 y 2026, según se detalla en la siguiente tabla.

Tabla 5: *Número de registros mensuales recolectados, por año (2018-2026)*

Año	Ene	Feb	Mar	Abr	May	Jun	Jul	Ago	Sep	Oct	Nov	Dic
2018	-	-	16594	-	-	16732	-	-	16781	-	-	16736
2019	-	-	16982	-	-	16980	-	-	-	-	-	-
2020	-	-	-	-	-	-	-	-	8587	8770	-	8756
2021	-	-	-	-	-	-	-	-	-	-	-	-
2022	-	8807	8821	8792	8839	8830	-	8962	8965	8864	8857	8852
2023	8879	8867	8834	8809	8799	8781	-	8689	8684	8779	8764	8736
2024	8803	8803	8797	8804	8727	8796	-	8906	8917	8792	8773	8726
2025	8795	8766	8790	8810	8739	8768	-	8607	8584	8757	8767	8748
2026	8791	8765	8798	8733	-	-	-	-	-	-	-	-

En la Tabla 5, se observa el número de registros mensuales recopilados durante el año, lo cual permitió identificar tres patrones esenciales para la creación de la base de datos utilizada en la modelación. En el mes de julio no tiene registros para ninguno de los años 2018 a 2026, lejos de ser una pérdida de información o un problema de calidad de los datos, esta ausencia sistemática es un rasgo estructural de la fuente, posiblemente relacionado con el diseño de la encuesta ENEMDU establecido por el INEC. Por lo tanto, este comportamiento no debe interpretarse como un conjunto de valores faltantes que pueden ser imputados, sino como un mes estructuralmente excluido del diseño muestral.

Por otro lado, a partir de 2020 se identifican cambios estructurales en el tamaño de la muestra, aunque en los meses disponibles 2018-2019 se registraron entre 16.594 y 16.982 observaciones en 2020, disminuyendo los registros mensuales. Lo mencionado se lo atribuye pandemias desde septiembre de 2020, ya que esta medida coincide con cambios metodológicos documentados por el INEC, que modificó el tamaño y distribución de la muestra, la representatividad de las estimaciones y los factores de expansión entre mayo de 2020 y mayo de 2021, tras la introducción de un esquema de encuesta telefónica durante la crisis sanitaria provocada por el COVID-19 (INEC, 2021).

En este sentido, la institución menciona que no fue posible generar la ENEMDU anual correspondiente a 2020, debido a que la información recolectada en entrevistas telefónicas

entre abril y agosto no cumplió con los requisitos metodológicos para crear una muestra anual comparable (INEC, p.F.). Por lo tanto, la ausencia de registros de 2021 se debe a una decisión metodológica de la fuente oficial, más que a problemas con el proceso de recolección o descarga de información.

En cuanto al año 2022 hacia adelante, se observa una clara estabilización del tamaño de la muestra, durante este período, manteniendo una cobertura de diez u once meses por año, pero con ausencia sistemática del mes de julio, lo que sugiere una mayor consistencia en las estadísticas tanto de muestreo como de encuesta.

3.2.3. Depuración y validación de datos

Disponer de 112 archivos descargados no garantiza, que la totalidad de estos resulte utilizable. Previamente al cálculo de cualquier indicador de privación, se verificó que las variables requeridas para la construcción del índice de pobreza estuvieran efectivamente presentes. De este proceso de verificación se derivaron dos decisiones metodológicas que condicionan el desarrollo del resto del capítulo.

La primera consistió en excluir dos meses correspondientes a 2020, período en el cual el INEC levantó la encuesta mediante modalidad telefónica como consecuencia de las restricciones sanitarias derivadas de la pandemia. Esta modalidad de recolección omitió, precisamente, las variables de vivienda e infraestructura indispensables para calcular las privaciones de hábitat y saneamiento.

La segunda decisión respondió a un problema de cobertura temporal: la variable oficial de pobreza monetaria extrema se releva únicamente en los meses de junio y diciembre de cada año. Con el fin de no prescindir de los diez meses restantes, se empleó como el ingreso per cápita del hogar, contrastado frente a la línea de pobreza extrema que el INEC establece anualmente.

3.2.4. Cálculo de privaciones e indicadores IPM de hogares

Con los períodos ya depurados, la etapa siguiente consistió en calcular la pobreza multidimensional conforme a la metodología Alkire-Foster, adoptada oficialmente por el INEC, la cual comprende cuatro dimensiones y doce indicadores de privación. Cada indicador se calcula, en primera instancia, a nivel de persona, y se agrega posteriormente al nivel de hogar

bajo el criterio según el cual un hogar se considera privado en un indicador determinado si al menos uno de sus miembros presenta dicha privación.

Tabla 6: *Indicadores de privación del IPM y su regla de cálculo*

Dimensión	Indicador	Regla de cálculo
Educación	Inasistencia escolar	No asiste a clases, entre 5 y 17 años
	No acceso a educación superior	No accedió por razones económicas, 18-29 años, nivel máximo secundaria o menos
	Logro educativo bajo	Nivel educativo máximo bajo, entre 18 y 64 años
Trabajo	Empleo infantil/adolescente	Ocupado entre 5-14 años; o entre 15-17 años ocupado con remuneración bajo el salario básico unificado (SBU), sin asistir a clases, o trabajando más de 30h semanales
	Desempleo/subempleo	Desempleado o con empleo inadecuado, mayor de 18 años
	Sin aporte a pensiones	No aportó a pensiones ni está jubilado, mayor de 15
Salud/Agua	Pobreza extrema por ingreso	Ingreso per cápita bajo la línea de pobreza extrema
	Sin agua de red pública	El agua no proviene de la red pública
Hábitat	Hacinamiento	Más de 3 personas por dormitorio
	Déficit habitacional	Materiales precarios en piso, pared o techo
	Saneamiento inadecuado	Servicio higiénico inadecuado (criterio varía urbano/rural)
	Eliminación inadecuada de basura	Quema, entierro, pozo o arrojo directo

Cada indicador se multiplica por un peso oficial del INEC, distinto entre zona urbana y rural porque la relevancia de cada privación cambia según el contexto. La suma ponderada de las privaciones de un hogar es su puntaje de privación; cuando este alcanza o supera un tercio, el hogar se clasifica como pobre multidimensional.

3.2.5. Construcción del panel longitudinal con rezagos temporales

Saber si un hogar es pobre en un mes dado, resuelve solo la mitad del problema. El objetivo es clasificar el riesgo de que una persona caiga en pobreza dentro de un año, y para eso se explota el diseño de panel rotatorio de la ENEMDU, que permite encontrar al mismo hogar encuestado doce meses después.

La tabla maestra de datos (MDT) se construyó en tres pasos. Primero, se generaron variables de rezago, que le dan al modelo una noción de trayectoria. Segundo, se construyó la variable objetivo: el estado de pobreza del mismo individuo, observado doce meses después.

Tercero, se filtraron los registros sin ese dato futuro disponible, ya que no todos los individuos permanecen en la muestra doce meses consecutivos.

3.2.6. Estructura de la base final y diccionario de variables

La base de datos final está compuesta por 481.826 observaciones y 43 variables, organizadas en cinco grupos: identificadores y claves de agrupación, variables sociodemográficas, variables de rezago temporal, privaciones e indicadores del IPM, y las dos variables relativas a la condición de pobreza del hogar (presente y futura). A cada variable se le asignó, además, un rol específico dentro del proceso de modelado: predictora, objetivo, clave de identificación, o auxiliar excluida.

Sobre esta última categoría conviene detenerse con mayor detalle. Las variables de ingreso contemporáneo se conservaron en la base de datos con fines de trazabilidad, pero se excluyeron deliberadamente del proceso de entrenamiento, dado que constituyen precisamente los insumos empleados para calcular la privación de pobreza extrema y, por consiguiente, la variable objetivo misma. Su inclusión como predictoras habría implicado que el modelo aprendiera a predecir la condición de pobreza a partir de información que ya la determina de manera directa, lo que habría comprometido la validez del ejercicio de predicción.

El primer objetivo específico se cumple: se construyó una base de datos estructurada y reproducible a partir de 112 archivos de la ENEMDU correspondientes a 56 períodos entre 2018 y 2026, con el hogar como unidad de análisis y la pobreza multidimensional calculada según la metodología oficial de Alkire-Foster que aplica el INEC. El proceso no estuvo libre de decisiones que condicionan el resultado final. La exclusión de los dos meses de 2020 levantados por vía telefónica no fue arbitraria, sino forzada por la ausencia de las variables de vivienda necesarias para calcular las privaciones de hábitat. Algo similar ocurre con el uso del ingreso per cápita como sustituto en los meses donde el INEC no releva directamente la pobreza extrema: es una solución razonable, pero introduce un supuesto que debería quedar explícito como limitación del estudio y no solo como nota de pie de página. El resultado, una base de 481.826 observaciones con 43 columnas, permite construir la variable objetivo proyectada a doce meses gracias al panel rotatorio de la encuesta, que es precisamente lo que hace posible el resto de la investigación.

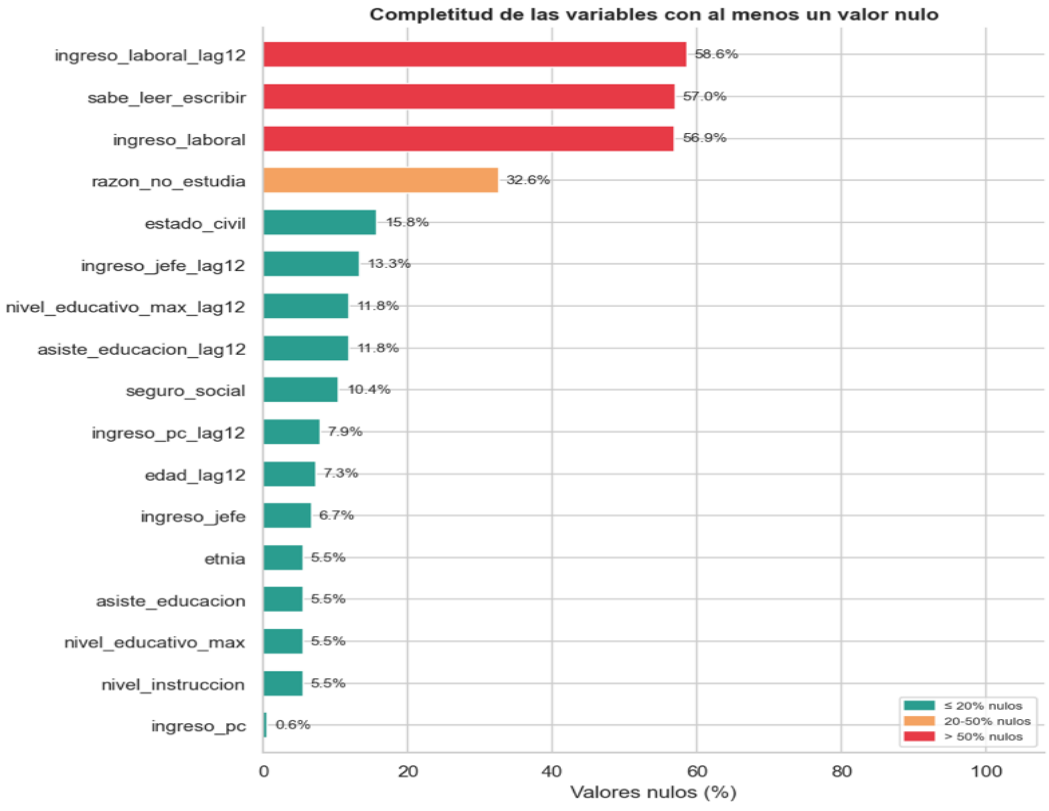
Capítulo 4. Análisis Exploratorio de Datos

El presente capítulo desarrolla el segundo objetivo específico de la investigación: explorar dicha base con el fin de identificar cómo se distribuye la pobreza multidimensional en función del nivel educativo, la zona geográfica y las características sociodemográficas y sociales de los hogares. La totalidad del análisis se realiza directamente sobre la base construida en el capítulo precedente, sin introducir procesamiento adicional alguno sobre los datos.

4.1. Punto de partida y completitud de la base

Antes de realizar cualquier interpretación resulta necesario establecer qué tan completa se encuentra la información sobre la cual se trabaja. Esta verificación previa no es un mero formalismo: un valor faltante puede reflejar un problema real de calidad de datos, o bien puede ser la consecuencia esperada del diseño mismo de la encuesta, y distinguir entre ambos casos es indispensable antes de sacar cualquier conclusión sobre la muestra. La Figura 3 presenta el porcentaje de valores nulos para cada una de las variables de la base que presentan al menos una ausencia.

Figura 3: *Completitud de las variables de la base final con al menos un valor nulo, clasificadas por severidad.*

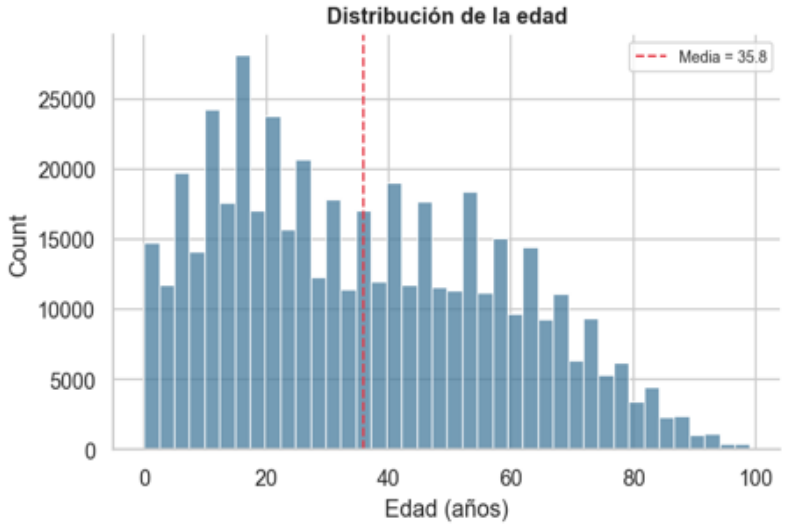


De las 43 variables que componen la base, 17 presentan al menos un valor nulo, aunque únicamente 4 superan el umbral del 20% de datos faltantes: ingreso_laboral_lag12 (58,6%), sabe_leer_escribir (57,0%), ingreso_laboral (56,9%) y razon_no_estudia (32,6%). Ninguna de estas ausencias constituye un error de carga o un problema atribuible al proceso de construcción de la base. Las dos primeras variables faltan de manera sistemática en las personas que no cuentan con empleo o que se encuentran fuera del rango al cual aplica la pregunta correspondiente, mientras que razon_no_estudia únicamente tiene sentido para quienes efectivamente no asisten a un centro educativo, por lo que su ausencia en el resto de la muestra es igualmente esperada. El grupo restante de variables con valores nulos presenta una completitud de entre el 84% y el 95%, cifra coherente con el diseño de panel rotatorio descrito en el capítulo anterior. En conjunto, estos resultados permiten concluir que la incompletitud observada en la base responde a características estructurales del instrumento de recolección y no a fallas en el proceso de depuración, lo cual respalda la validez de los datos para las etapas posteriores del análisis.

4.2. Perfil general de la muestra

Antes de analizar cualquier variable contra la pobreza, conviene establecer con precisión quién compone la muestra sobre la que se trabaja. Las Figuras 4 a 7 completan ese punto de partida con cuatro dimensiones descriptivas de la muestra: la distribución etaria, la composición por sexo, la distribución por zona geográfica y la composición étnica.

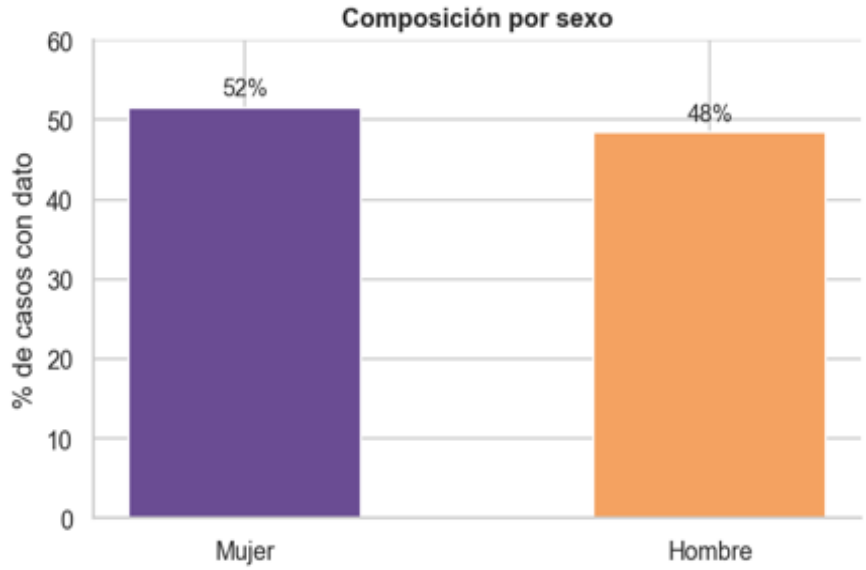
Figura 4: *Perfil general de la muestra: distribución de edad*



La distribución por edades muestra un comportamiento heterogéneo y desigual en toda el área observada. Una mayor concentración de inidividuos se encuentra a una edad temprana, especialmente entre los 10 y 20 años, y alrededor de los 15 años. A partir de este intervalo, la frecuencia de las observaciones muestra una tendencia general a la baja, aunque con fluctuaciones a lo largo de la distribución.

La edad media de la muestra es de 35,8 años, lo que demuestra que, a pesar de la concentración de las observaciones en la juventud, la presencia de adultos y personas mayores tiene un peso significativo en la composición global de la muestra. Igualmente, la presencia de individuos se observa hasta aproximadamente los 100 años de edad, aunque la frecuencia disminuye significativamente con la edad. Esta distribución es esencial para un análisis más detallado de la pobreza multidimensional, ya que la composición por edad puede estar relacionada con diferentes condiciones socioeconómicas y necesidades de los hogares. En particular, la presencia de niños y jóvenes cobra importancia al analizar dimensiones relacionadas con la educación y las condiciones de vida.

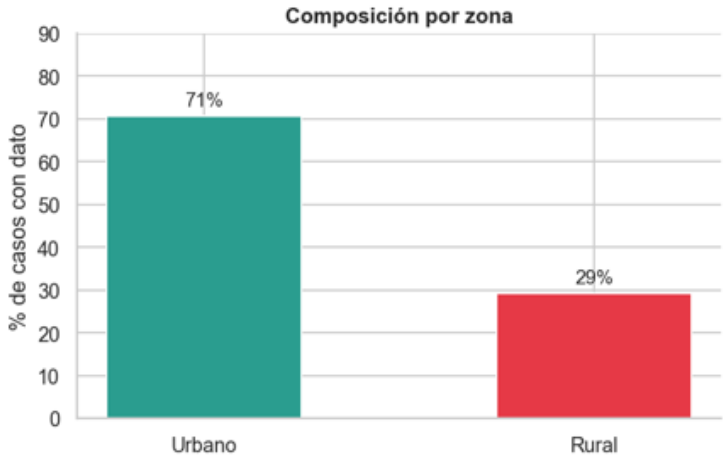
Figura 5: *Perfil general de la muestra: sexo*



En la Figura 5 se observa la concentración de la muestra por género, la figura evidencia una distribución relativamente equilibrada., las mujeres representan el 52% de las observaciones, mientras que los hombres representan el 48%, lo que representa una diferencia de cuatro puntos porcentuales entre ambos grupos. Esta distribución permite observar que ningún género domina las observaciones de la muestra. La participación femenina ligeramente mayor es una característica básica que debe tenerse en cuenta más adelante al analizar las

diferencias multidimensionales de género en el riesgo de pobreza. La variable género es fundamental en el análisis porque nos permitirá determinar si existen diferencias en la distribución de la pobreza multidimensional entre hombres y mujeres, para luego evaluar si esta característica tiene poder discriminante en el modelo de predicción.

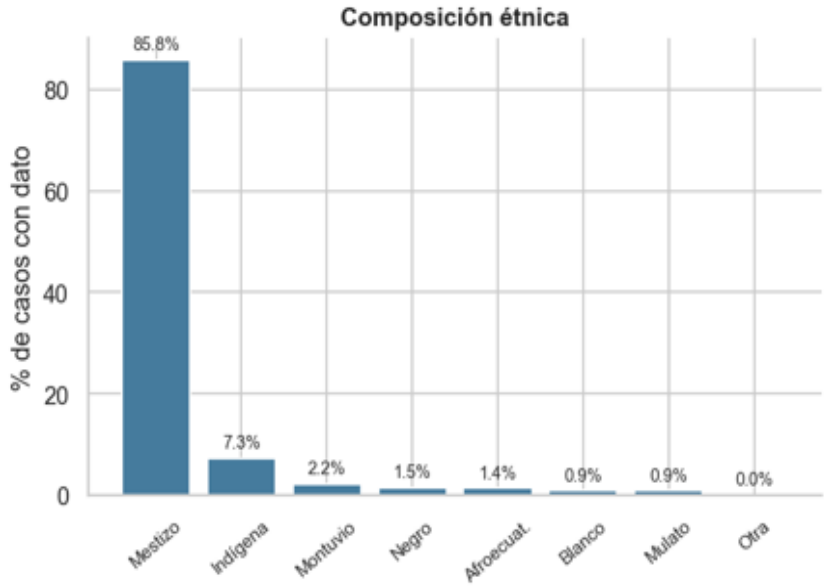
Figura 6: *Perfil general de la muestra: zona geográfica.*



En la Figura 6 observamos la distribución de los individuos por zonas geográficas muestra una mayor concentración en la ciudad, que constituye el 71% de la muestra, mientras que el 29% corresponde al área rural. Por lo tanto, alrededor de siete de cada diez observaciones se refieren a zonas urbanas, mientras que tres corresponden a zonas rurales. Esta diferencia en la composición territorial es un aspecto importante en el análisis de la pobreza multidimensional, ya que luego permite comparar el comportamiento de la variable objetivo en los dos contextos geográficos.

La distinción entre áreas urbanas y rurales es de particular importancia en este estudio, ya que la ubicación territorial es una de las variables más importantes consideradas para identificar diferencias multidimensionales en el riesgo de pobreza. La distribución observada también debe considerarse durante la construcción y evaluación del modelo, especialmente para evitar que las diferencias en el tamaño del grupo influyan en la interpretación del desempeño predictivo o identifiquen patrones específicos en las poblaciones rurales.

Figura 7: *Perfil general de la muestra: etnia.*



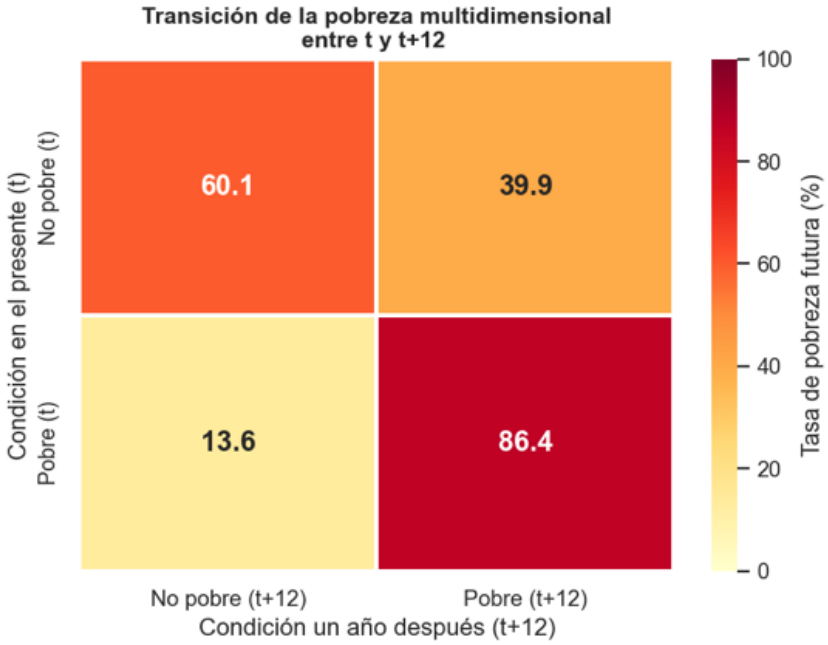
En la Figura 7 observamos la composición étnica, en cambio, el panorama es mucho menos equilibrado: el 85.8% de la muestra se autoidentifica como mestiza, seguida a gran distancia por la población indígena (7.3%) y montuvia (2.2%), mientras que la categoría negra (1.5%), afroecuatoriana (1.4%), blanca y mulata (0.9% cada una) y otra (0.0%) apenas tienen presencia en la base. En conjunto, el perfil describe una muestra equilibrada en sexo, mayoritariamente urbana y étnicamente concentrada en la población mestiza, un punto de referencia necesario para interpretar con cuidado las brechas que se examinan a continuación.

4.3. Dinámica de la pobreza: persistencia, entrada y salida

A continuación, se va a examinar la pobreza no solo en términos de su magnitud, sino de su estabilidad en el tiempo: no basta con saber cuántos hogares son pobres en un momento dado, sino qué tan probable es que esa condición se mantenga o cambie doce meses después.

Para responder a esta pregunta, la Figura 8 cruza la condición de pobreza de cada persona en el momento de la encuesta con su condición doce meses después, ponderando cada caso por el factor de adecuado. Las cuatro celdas de la matriz permiten leer directamente qué tan probable es que un hogar permanezca en su condición inicial o la modifique.

Figura 8: *Matriz de transición de la pobreza multidimensional entre t y t+12, ponderada por el factor de expansión muestral.*

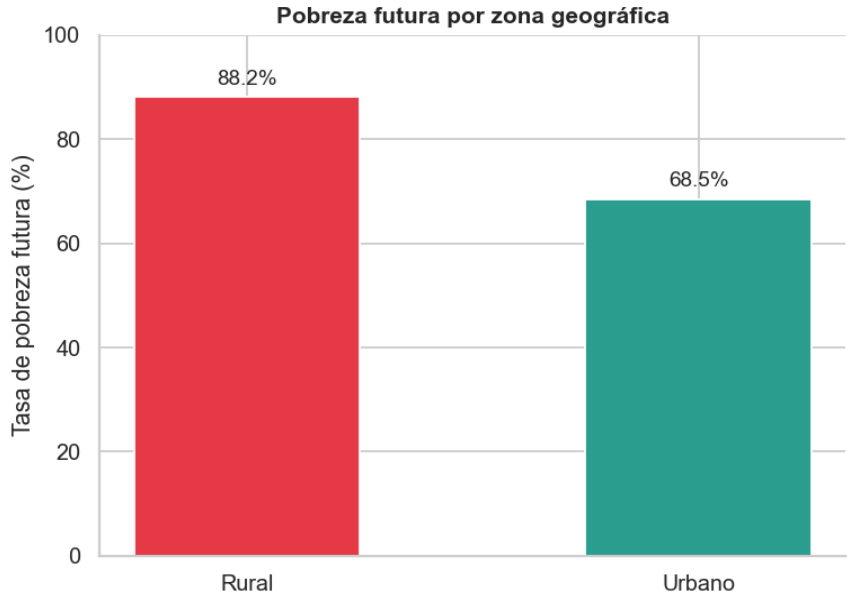


La matriz muestra un resultado contundente. De los hogares que no eran pobres en el presente, el 60.1% se mantiene fuera de la pobreza un año después, pero el 39.9% cae en ella, lo que se puede llamar la tasa de entrada. De los hogares que sí eran pobres, apenas el 13.6% logra salir de esa condición, mientras que el 86.4% permanece pobre, la tasa de persistencia. La conclusión no cambia: perder o mantener la condición de pobreza no se reparte de forma pareja entre ambos grupos, y quedarse en la pobreza es, por un margen amplio, más probable que salir de ella.

4.4. Zona geográfica y las privaciones que explican la brecha

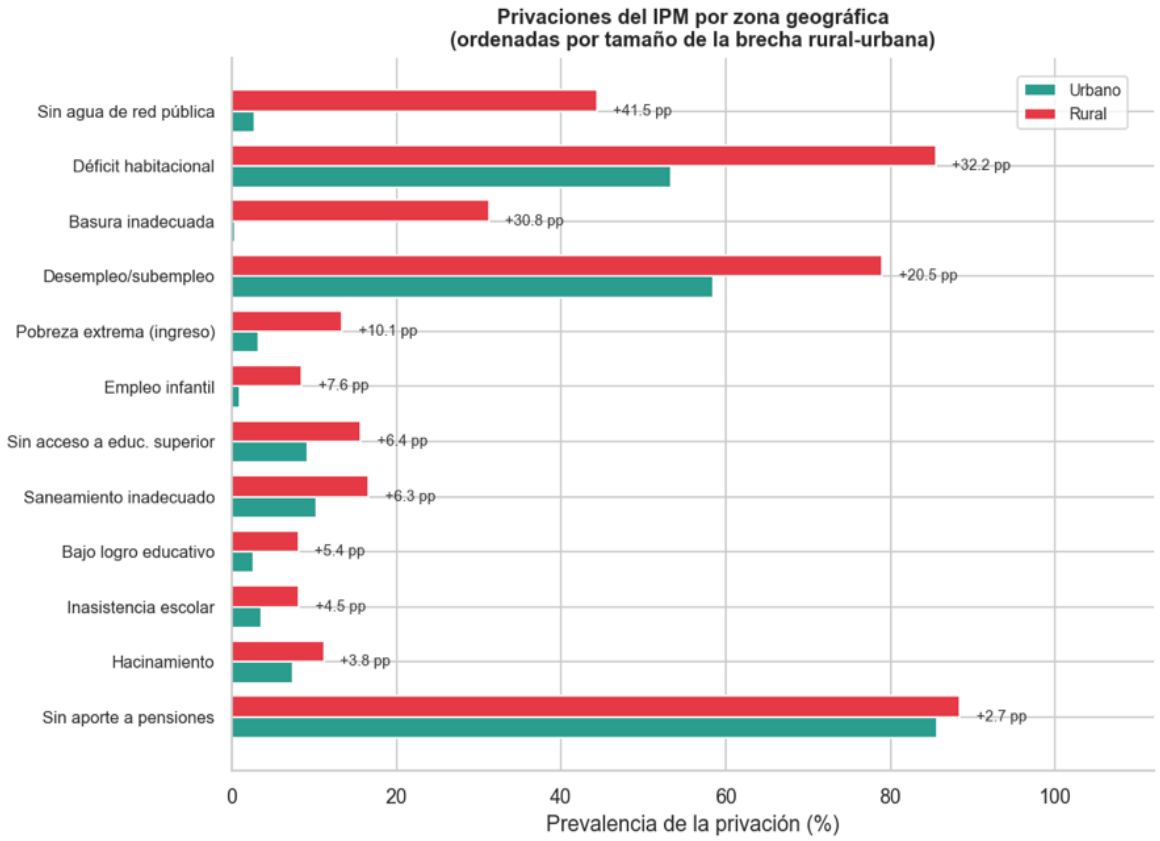
La sección anterior mostró que la pobreza multidimensional, una vez que se instala en un hogar, tiende a persistir con fuerza en el tiempo. Corresponde ahora examinar uno de los factores que más influye en esa persistencia: la zona de residencia. La Figura 9 presenta la tasa de pobreza multidimensional futura, calculada por separado para hogares urbanos y rurales.

Figura 9: *Tasa de pobreza multidimensional futura según zona geográfica, ponderada por el factor de expansión*



La diferencia es amplia: 88.2% de pobreza futura en la zona rural frente a 68.5% en la urbana, una brecha de 19.7 puntos porcentuales. Esta cifra ya se había identificado en el capítulo anterior. Lo que no se había hecho todavía es preguntar qué privaciones específicas construyen esa brecha, y no solo constatar que existe. Para responder eso, la siguiente figura descompone la prevalencia de cada una de las doce privaciones del IPM según la zona geográfica del hogar.

Figura 10: *Prevalencia de las doce privaciones del Índice de Pobreza Multidimensional por zona geográfica, con la brecha rural-urbana en puntos porcentuales para cada.*



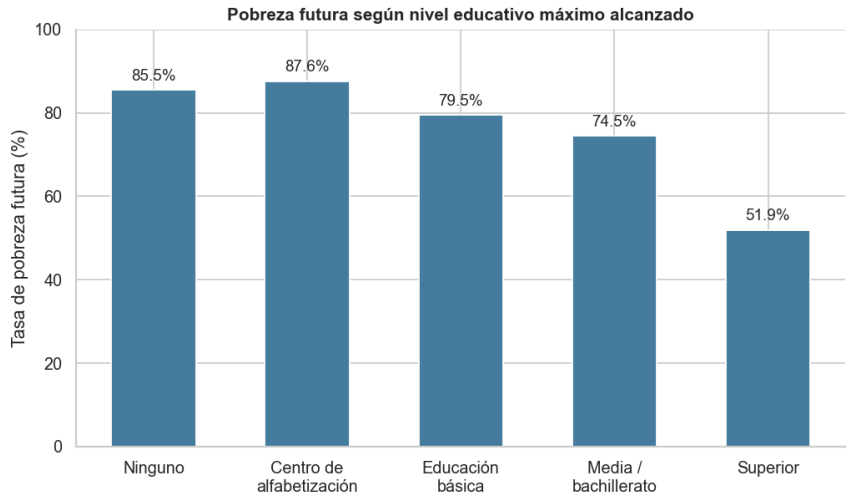
La falta de agua por red pública es, con diferencia, la privación más desigual entre zonas, 44.3% en hogares rurales frente a apenas 2.8% en urbanos, una brecha de 41.5 puntos, seguida por el déficit habitacional (32.2 puntos) y la eliminación inadecuada de basura (30.8 puntos). En el otro extremo, la falta de aporte a un fondo de pensiones, que es la privación más común en términos absolutos, apenas varía entre zonas, con solo 2.7 puntos de diferencia: es un problema casi tan extendido en la ciudad como en el campo. Esto tiene una implicación concreta. La brecha rural-urbana en la pobreza multidimensional es un fenómeno generado principalmente infraestructura de agua y vivienda, no empleo. Junto a la zona, el segundo eje que plantea la hipótesis de esta investigación es el nivel educativo.

4.5. Nivel educativo

El segundo eje que plantea la hipótesis de esta investigación, junto con la zona de residencia, es el nivel educativo del hogar. Con el propósito de examinar esta relación, la Figura 11 presenta la tasa de pobreza multidimensional futura según el nivel educativo máximo alcanzado por los miembros del hogar, clasificado de acuerdo con las categorías oficiales del

INEC: ninguno, centro de alfabetización, educación básica, educación media o bachillerato, y superior.

Figura 11: *Tasa de pobreza multidimensional futura según nivel educativo máximo alcanzado.*



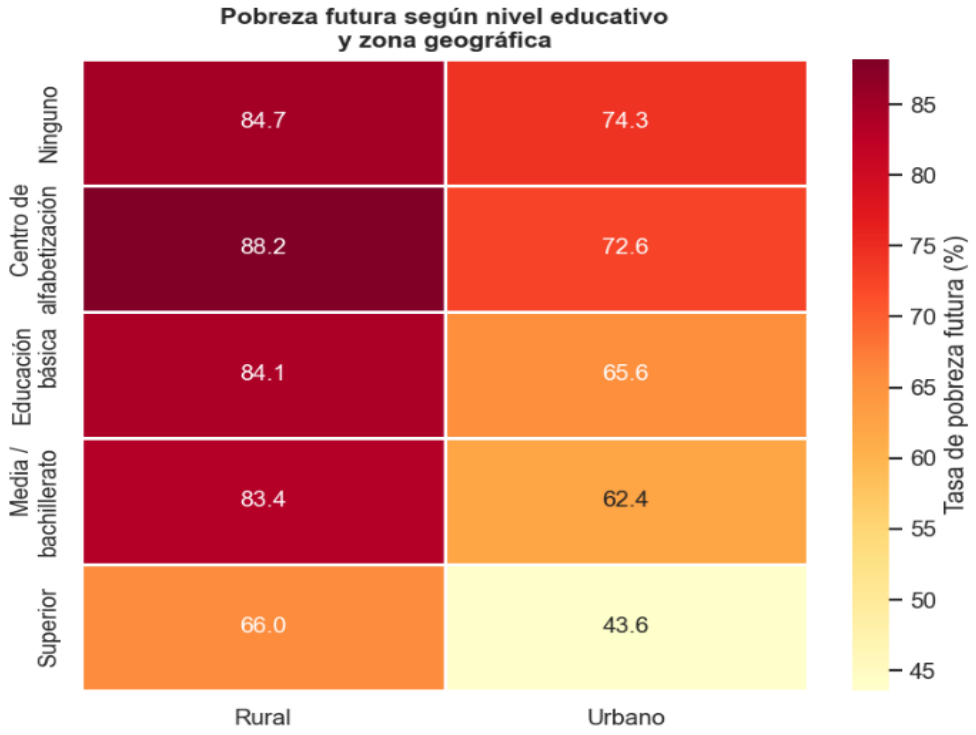
Tal como se observa en la Figura 11, la variable educativa constituye uno de los hallazgos más importantes de todo el capítulo: la tasa de pobreza futura desciende de 85.5% entre los hogares sin ningún nivel de instrucción hasta 51.9% entre aquellos cuyo miembro con mayor nivel educativo alcanzó la educación superior, una reducción de 33.6 puntos porcentuales.

Que la zona geográfica y el nivel educativo se comporten de la manera esperada al analizarse por separado no permite anticipar qué ocurre cuando ambas condiciones actúan de forma conjunta.

4.6. Efecto combinado de zona y educación

Habiendo examinado la zona geográfica y el nivel educativo por separado, corresponde ahora determinar si ambas variables aportan información independiente o si, por el contrario, una de ellas simplemente refleja a la otra. Con este propósito, la Figura 12 cruza el nivel educativo máximo alcanzado con la zona geográfica del hogar, presentando la tasa de pobreza multidimensional futura para cada una de las diez combinaciones posibles.

Figura 12: *Tasa de pobreza multidimensional futura según nivel educativo máximo alcanzado y zona geográfica*



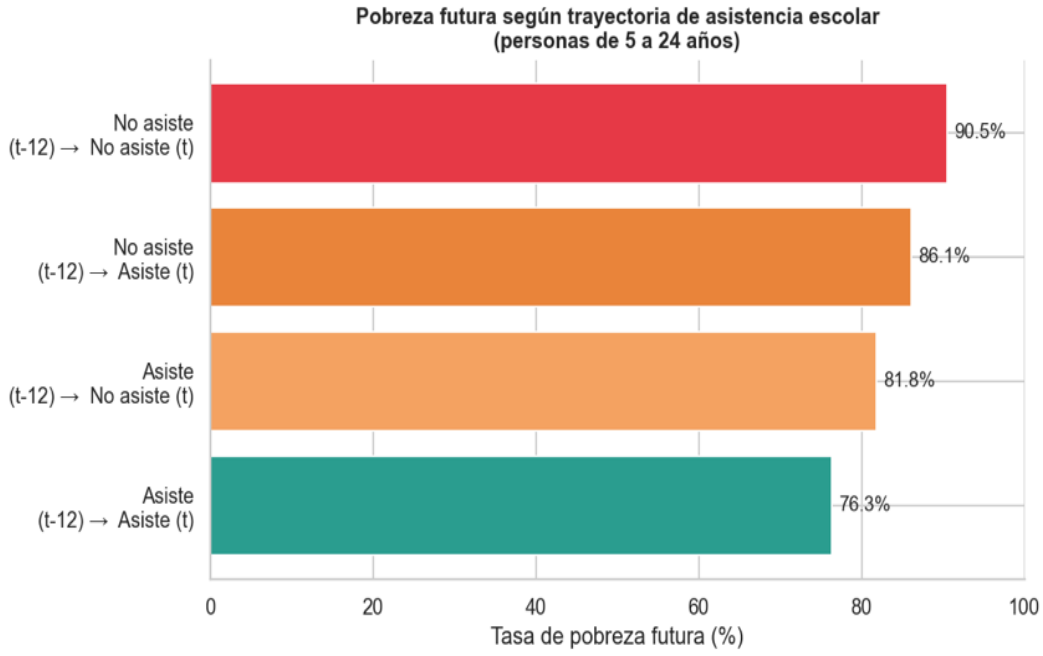
El cruce confirma que ambas variables operan de forma complementaria. Dentro de cada nivel educativo, la zona rural registra sistemáticamente una tasa de pobreza futura más alta que la urbana, sin una sola excepción a lo largo de las cinco categorías. Dentro de cada zona, en cambio, la pobreza tiende a descender a medida que aumenta el nivel educativo, aunque en el caso rural este descenso no es estrictamente monótono: la categoría de centro de alfabetización alcanza 88.2%, por encima incluso de la categoría "ninguno" (84.7%). El extremo más favorable de toda la tabla corresponde a la educación superior en zona urbana, con 43.6%, mientras que el más desfavorable es precisamente el centro de alfabetización en zona rural, con 88.2%, una diferencia de 44.6 puntos porcentuales entre ambos extremos. Que ninguna de las dos variables absorba por completo el efecto de la otra respalda la decisión de incorporarlas al modelo como predictoras independientes, y no como sustitutas la una de la otra.

4.7. Ingresos per cápita y valores atípicos

La variable de asistencia educativa actual apenas diferencia el riesgo de pobreza futura: 75.5% entre quienes asisten a clases frente a 73.9% entre quienes no, una diferencia

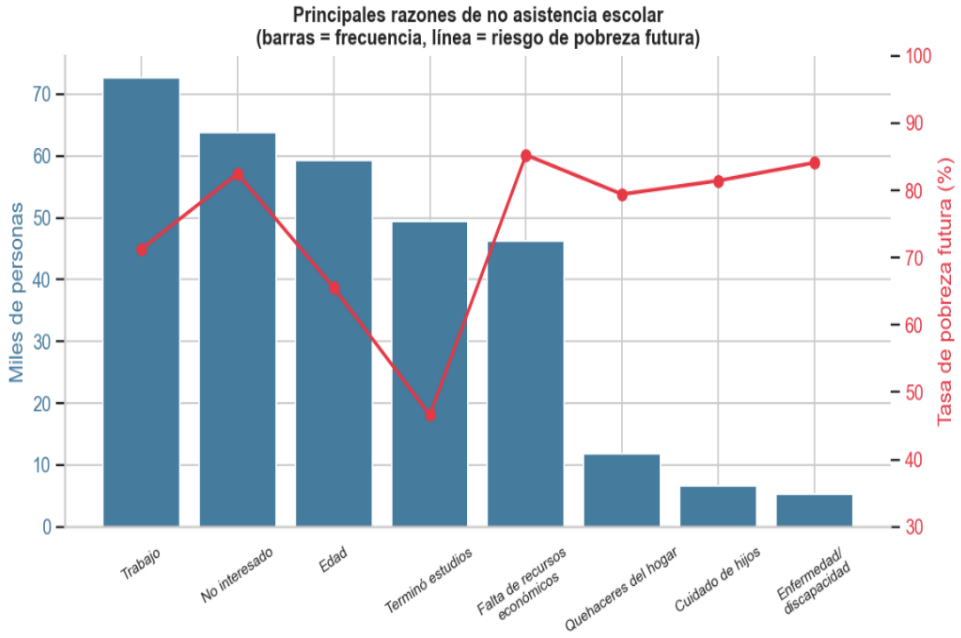
prácticamente nula. Esta escasa capacidad discriminante, no significa que la asistencia escolar carezca de valor predictivo. Para poner a prueba esta idea, la Figura 13 cruza la asistencia educativa en t−12 contra la asistencia en t para personas de 5 a 24 años, distinguiendo así cuatro trayectorias posibles en lugar de dos estados fijos.

Figura 13: *Tasa de pobreza multidimensional futura según trayectoria de asistencia escolar entre t−12 y t, personas de 5 a 24.*



A diferencia de la variable estática, esta desagregación revela una tendencia clara de riesgo. Quienes asistían a clases en ambos momentos registran la tasa más baja, 76.3%. Quienes regresaron a clases tras haber estado fuera del sistema muestran una tasa más alta, 81.8%, y quienes abandonaron los estudios en el último año presentan un riesgo mayor, 86.1%. El riesgo máximo corresponde a quienes llevan al menos un año completo sin asistir, con 90.5%, 14.2 puntos por encima de quienes asisten de forma continua. Este resultado tiene una implicación metodológica que trasciende la variable en sí misma: el rezago no es redundante respecto de la variable actual, sino que aporta información propia y es el que realmente separa a los hogares de mayor riesgo.

Figura 14: *Principales razones de no asistencia escolar: frecuencia y tasa de pobreza multidimensional futura asociada a cada una.*

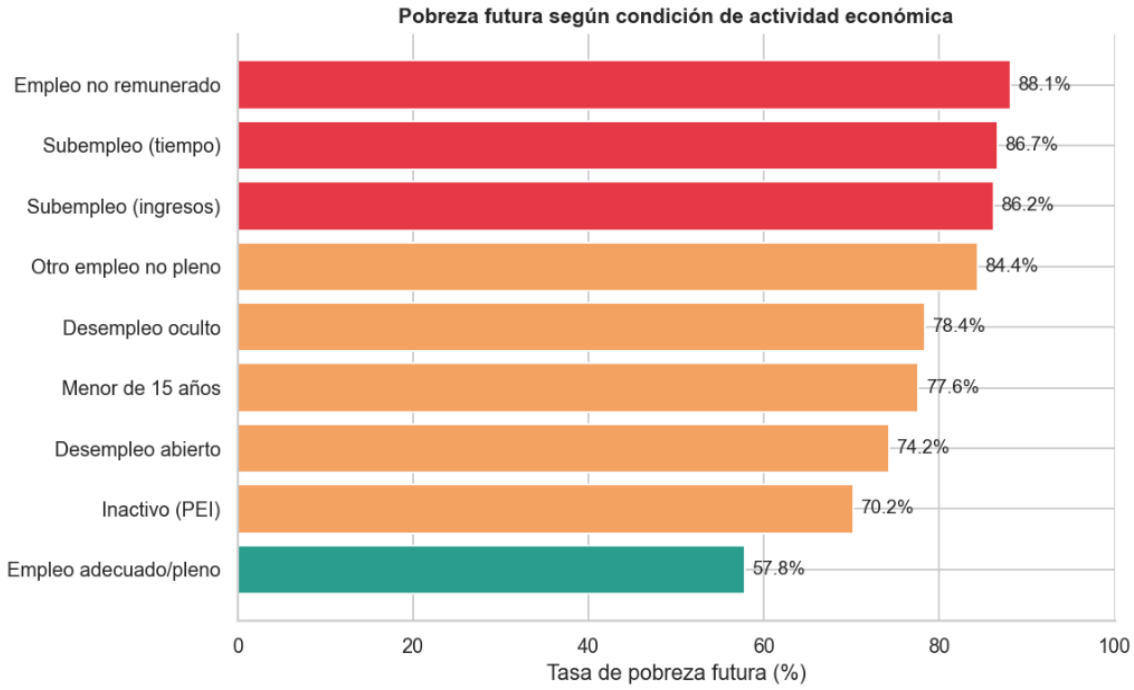


El gráfico presenta que frecuencia y el riesgo no necesariamente coinciden. El trabajo es la razón más citada, con 72,565 personas, pero su tasa de pobreza futura asociada, 71.3%, no es la más alta del conjunto. Esa posición corresponde a la falta de recursos económicos, con 85.3%, a pesar de tratarse apenas de la quinta razón más frecuente, con 46,241 casos. En el extremo opuesto, haber terminado los estudios es la causa con menor riesgo asociado, 46.6%, muy por debajo de cualquier otra categoría, un resultado esperable si se considera que quienes completan su formación suelen encontrarse en una situación económica más favorable que quienes abandonan por necesidad. La educación no actúa sola. Junto con el empleo y la protección social, forma el otro gran bloque de condiciones materiales que separa a los hogares en riesgo.

4.8. Situación laboral y afiliación a la seguridad social

La situación laboral de los miembros del hogar constituye otro determinante material del riesgo de pobreza. Con el fin de examinarlo, la Figura 15 presenta la tasa de pobreza multidimensional futura según la condición de actividad económica, clasificada de acuerdo con la categorización oficial del INEC.

Figura 15: *Tasa de pobreza multidimensional futura según condición de actividad económica.*



La categoría con mayor riesgo asociado no es el desempleo, sino el empleo no remunerado, con 88.1%, seguida de cerca por las dos formas de subempleo: por insuficiencia de tiempo de trabajo (86.7%) y por insuficiencia de ingresos (86.2%). Estas tres categorías superan tanto al desempleo abierto (74.2%) como a la población económicamente inactiva (70.2%), es decir, a quienes se encuentran fuera de la fuerza laboral. El empleo adecuado o pleno es, por lejos, la categoría menos riesgosa, con 57.8%, cerca de treinta puntos por debajo del empleo no remunerado. En conjunto, el patrón indica que trabajar en condiciones precarias se asocia a un riesgo futuro mayor que no trabajar en absoluto. Este resultado resulta coherente si se considera que el desempleo o la inactividad ocurren en hogares que cuentan con otro sostén económico adecuado, mientras que el empleo no remunerado y el subempleo suelen presentarse en hogares donde todos sus miembros deben trabajar, sea cual sea la calidad del empleo disponible, porque no existe otra alternativa.

La afiliación a la seguridad social permite examinar esta misma dinámica desde otro ángulo. No contar con ningún tipo de afiliación, la situación más común en la muestra, se asocia a una tasa de pobreza futura de 83.3%, prácticamente igual a la del seguro campesino, dirigido específicamente a la población rural, con 84.4%. En el extremo opuesto se ubican los regímenes

de afiliación formal: el Instituto Ecuatoriano de Seguridad Social (IESS) general baja el riesgo a 51.8%, el IESS voluntario a 46.0% y el seguro de fuerzas armadas y policía a 44.2%. Por debajo de estos se encuentran los seguros privados, con las tasas más bajas de todo el conjunto: 40.8% sin cobertura de hospitalización y apenas 29.0% con hospitalización incluida, la cifra menor de toda la figura.

4.9. Sexo, etnia y estado civil

Más allá de los ejes estructurales examinados hasta aquí, la zona geográfica, el nivel educativo y la situación laboral, la hipótesis de esta investigación exige revisar también un conjunto de características sociodemográficas adicionales. Con este propósito, Las Figuras 16 a 18 presentan la tasa de pobreza multidimensional futura según tres variables: el sexo de la persona de referencia del hogar, su autoidentificación étnica y su estado civil.

Figura 16: *Tasa de pobreza multidimensional futura según sexo*

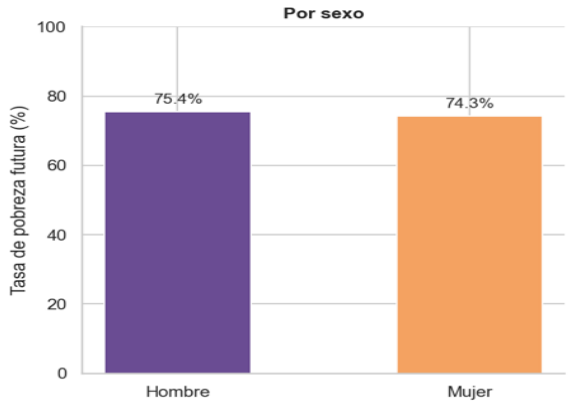
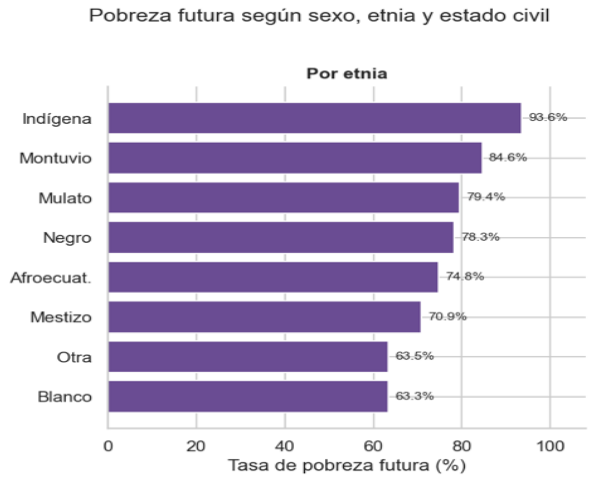
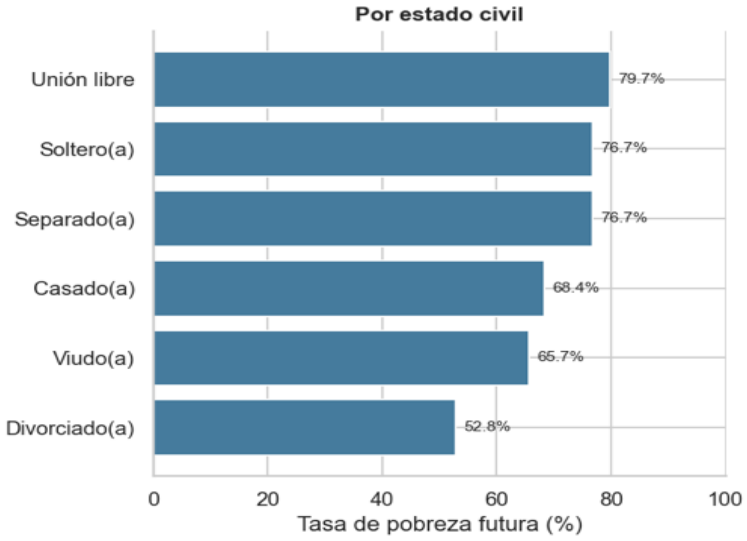


Figura 17: *Tasa de pobreza multidimensional futura según etnia*



El sexo de la persona de referencia aporta poca información: 75.4% de pobreza futura cuando es hombre, frente a 74.3% cuando es mujer, una diferencia de apenas 1.1 puntos porcentuales. La etnia, en cambio, revela una de las brechas más amplias de todo el capítulo: 93.6% entre los hogares que se identifican como indígenas, frente a 63.3% entre los que se identifican como blancos, una diferencia de 30.3 puntos que supera incluso a la brecha territorial examinada en la sección 4.4. Entre ambos extremos se ordenan, de mayor a menor riesgo, la población montuvia (84.6%), mulata (79.4%), negra (78.3%), afroecuatoriana (74.8%) y mestiza (70.9%), esta última mayoritaria en la muestra.

Figura 18: *Tasa de pobreza multidimensional futura según estado civil*



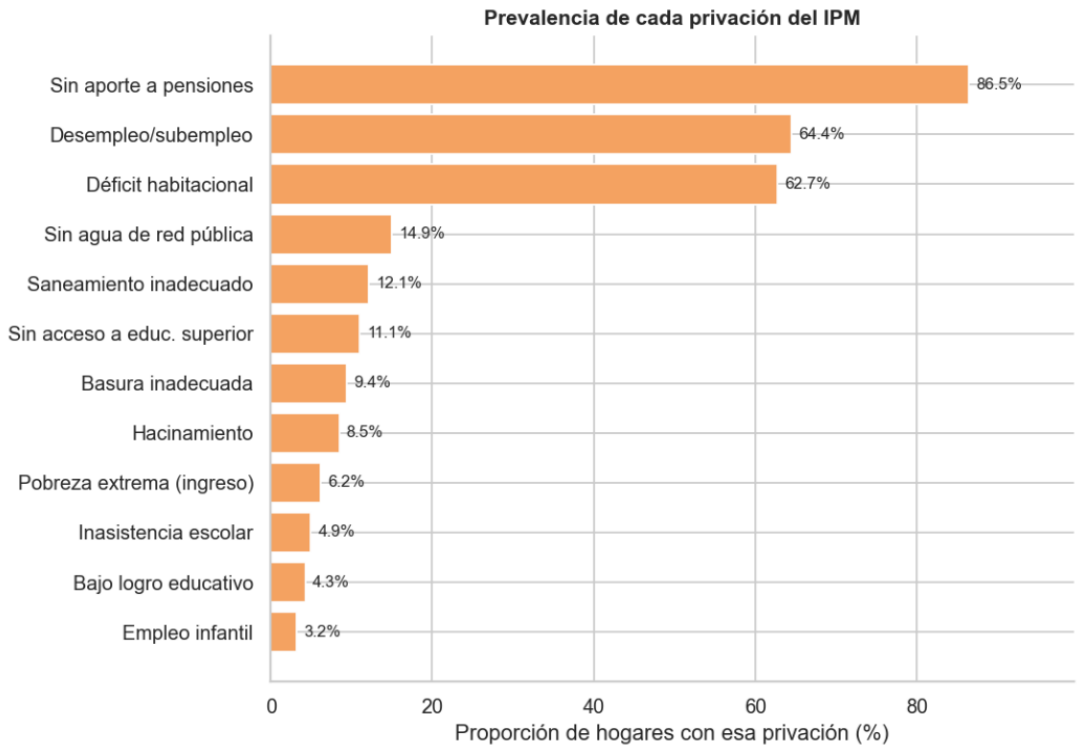
El estado civil, por su parte, introduce un matiz que no se había explorado hasta ahora: la unión libre registra la tasa más alta, 79.7%, mientras que el divorcio registra la más baja, 52.8%, con el resto de categorías, soltería y separación (76.7% cada una), matrimonio (68.4%) y viudez (65.7%), distribuidas entre ambos extremos. Es un contraste que probablemente refleja diferencias de fondo en nivel educativo y zona de residencia más que un efecto directo del estado civil en sí mismo, y que por lo tanto debe interpretarse con cautela si llegara a usarse como predictor independiente en el modelo.

4.10. Las privaciones del IPM: prevalencia, correlación y redundancia

Detrás de las brechas examinadas hasta aquí, por zona, nivel educativo, situación laboral, etnia y estado civil, hay un denominador común: las privaciones específicas que

componen el Índice de Pobreza Multidimensional. La Figura 19 presenta la prevalencia de cada una de las doce privaciones en el conjunto de hogares de la base.

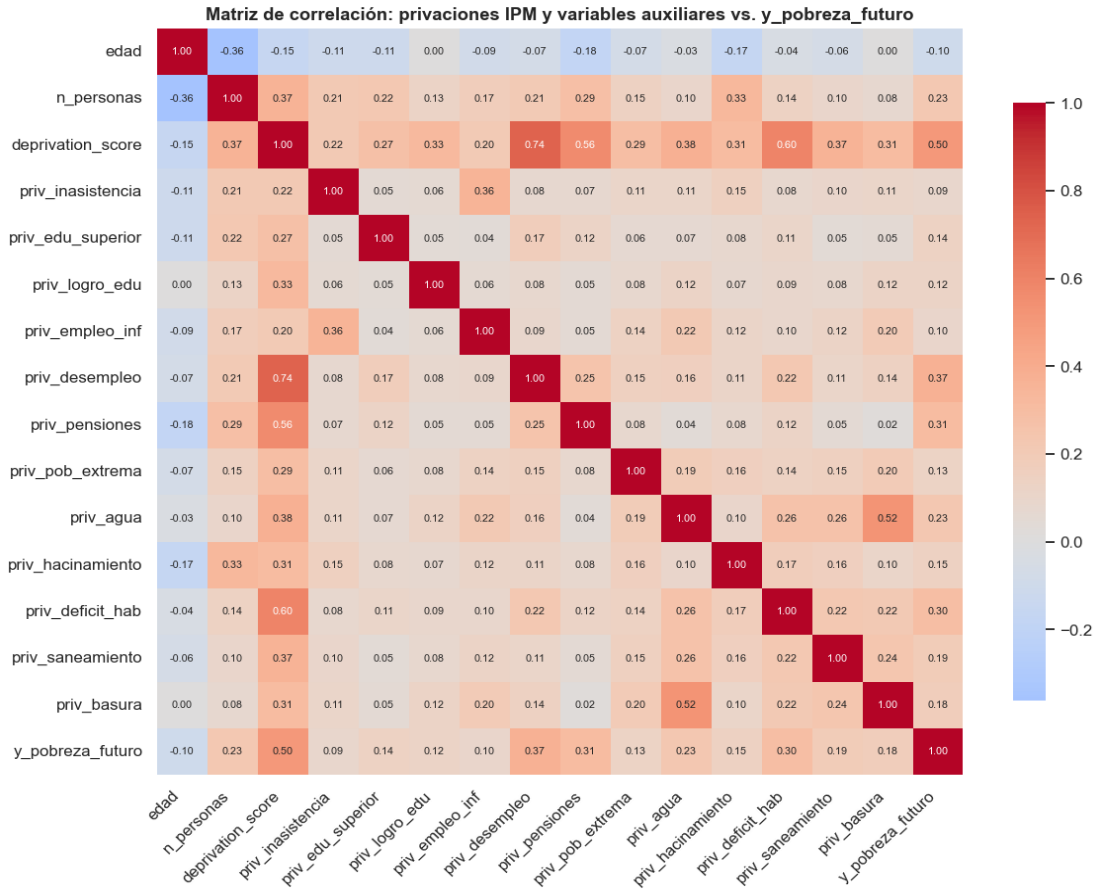
Figura 19: *Prevalencia de cada privación del Índice de Pobreza Multidimensional.*



La distribución es marcadamente desigual. La falta de aporte a un fondo de pensiones es, con amplio margen, la privación más extendida, presente en 86.5% de los hogares, seguida por el desempleo o empleo inadecuado (64.4%) y por el déficit habitacional (62.7%). En el extremo opuesto, el trabajo infantil (3.2%) y el bajo logro educativo (4.3%) son las privaciones menos frecuentes en la muestra.

Conocer qué tan extendida está cada privación es solo la mitad del panorama. Para entender cómo se relacionan entre sí, y no solo con la pobreza futura, la Figura 20 presenta la matriz de correlación entre las doce privaciones, dos variables auxiliares (edad y tamaño del hogar), el puntaje de privación y la propia variable objetivo.

Figura 20: *Matriz de correlación entre las privaciones del IPM, variables auxiliares y la pobreza multidimensional futura.*



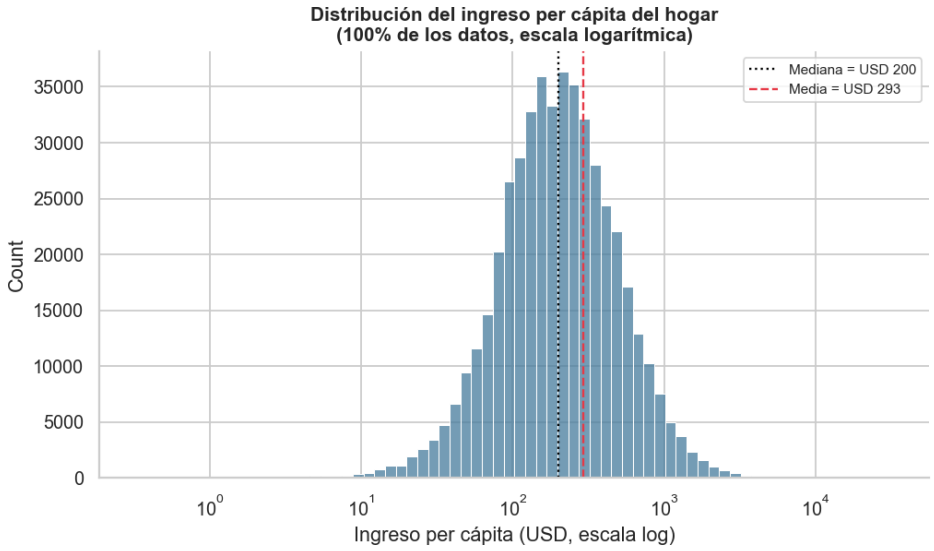
La matriz revela un problema práctico para el modelado que no era evidente al mirar cada privación por separado. El puntaje de privación correlaciona 0.74 con la privación por desempleo, 0.60 con el déficit habitacional y 0.56 con la falta de aporte a pensiones, un resultado esperable si se considera que ese puntaje se construye, precisamente, sumando esas mismas privaciones con sus pesos correspondientes. Incluir el puntaje de privación junto con las doce privaciones individuales como predictores del modelo introduce, en consecuencia, una redundancia considerable. Esta redundancia no compromete el funcionamiento de un modelo de árboles como Random Forest, pero sí reparte la importancia de las variables entre predictores que describen la misma información, dificultando la interpretación de cuál de ellos pesa más.

4.11. Ingresos: comportamiento y calidad del dato

El ingreso ocupa un lugar aparte en este análisis por su papel en la construcción del índice. La Figura 21 presenta la distribución completa del ingreso per cápita del hogar, la única variable de ingreso que interviene directamente en el cálculo del índice a través de la privación

por pobreza extrema, representada en escala logarítmica para incluir la totalidad de las observaciones sin necesidad de recortar los valores extremos.

Figura 21: *Distribución del ingreso per cápita del hogar, 100% de las observaciones, escala logarítmica.*



La distribución muestra una asimetría marcada hacia la derecha: la mitad de los hogares vive con menos de USD 200 mensuales por persona, la mediana, mientras que el promedio se ubica en USD 293, por encima de la mediana, porque un grupo reducido de hogares reporta ingresos varias veces superiores al típico.

El análisis confirma los dos factores centrales de la hipótesis, zona rural y bajo nivel educativo, pero el aporte real de este capítulo está en lo que añade más allá de esa confirmación. La pobreza multidimensional es un estado persistente: ocho de cada diez hogares pobres lo siguen siendo un año después. La brecha rural-urbana tiene una causa identificable y concentrada en agua y vivienda, no en empleo. La precariedad laboral, no solo el desempleo, es un factor de riesgo tan fuerte como la falta de educación.

Capítulo 5. Diseño, entrenamiento y evaluación de modelos de stacking para la clasificación del riesgo de pobreza multidimensional

En este capítulo se describe el diseño, experimentación y evaluación de los diferentes modelos de machine learning e iteraciones realizadas para predecir la probabilidad de pobreza multidimensional, en la población ecuatoriana, con una temporalidad de 12 meses. Cada prueba se enfoca en optimizar las métricas de rendimiento, abordar variables que inciden en la pobreza futura y analizar el conjunto de características permite representar de manera más eficaz la diversidad territorial entre áreas urbanas y rurales. Para esto, la MDT se construyó a partir de los datos levantados de la encuesta de ENEMDU para los periodos 2018-2025, con una base de datos de 481.826 registros a nivel de individuo y 35 variables de prueba. La variable objetivo *y_pobreza_futuro*, es una variable de tipo binaria que evidencia si el hogar al que pertenece el individuo será clasificado como pobre 12 meses en el futuro, a través de la metodología de Alkire-Foster ajustada al entorno ecuatoriano, de acuerdo con lo que se registra en los boletines del INEC.

5.2. Preparación y partición temporal del conjunto de datos

Para particionar el conjunto de datos para el modelo, se tomó la decisión metodológica de utilizar una partición temporal, dado que la base de datos de la ENEMDU contiene un panel longitudinal de varios periodos. El conjunto de entrenamiento está conformado por los periodos hasta noviembre de 2024, el conjunto de validación lo conforman meses de diciembre de 2024 a mayo de 2025 y para la prueba se toman los meses posteriores a mayo de 2025. A continuación, se detalla en la tabla.

Tabla 7: *Partición del conjunto de datos*

Partición	Registros	%
Entrenamiento	324,287	67%
Prueba	82,925	17%
Validación	74,614	15%
Total	481,826	100%

Para complementar el análisis y tratamiento de la información se realizó una transformación logarítmica (log1p) a todas las variables de ingreso, para estabilizar la escala de las variables y posteriormente se aplicó un StandardScaler para normalizar.

5.3. Selección de variables para entrenamiento de los modelos mediante Information Value (IV)

Para fundamentar la selección de variables predictoras de la pobreza multidimensional se utilizó el IV, el cual, es utilizada ampliamente en los modelos de riesgo de crédito (Siddiqi, 2006), esta técnica mide el poder discriminativo de cada variable para dividir la clase positiva, hogares que serán clasificados como pobres de la clase negativa.

Su aplicación fue sobre el conjunto de 35 variables, con la siguiente formulación:

$$IV = \Sigma\ (\%\ Eventos_i - \%\ No\ Eventos_i) \times ln(\%\ Eventos_i / \%\ No\ Eventos_i)$$

Tabla 8: *Interpretación del Information Value*

Rango de IV	Poder Predictivo	Decisión adoptada
< 0,02	Bajo	Evaluar caso por caso
0,02 – 0,10	Débil	
0,10 – 0,30	Medio	Incluir en el modelo
0,30 – 0,50	Fuerte	Incluir con análisis de endogeneidad
> 0,50	Sospechosamente alto	Excluir, leakage probable

Nota: Interpretación basada en Siddiqi (2006, pág. 79)

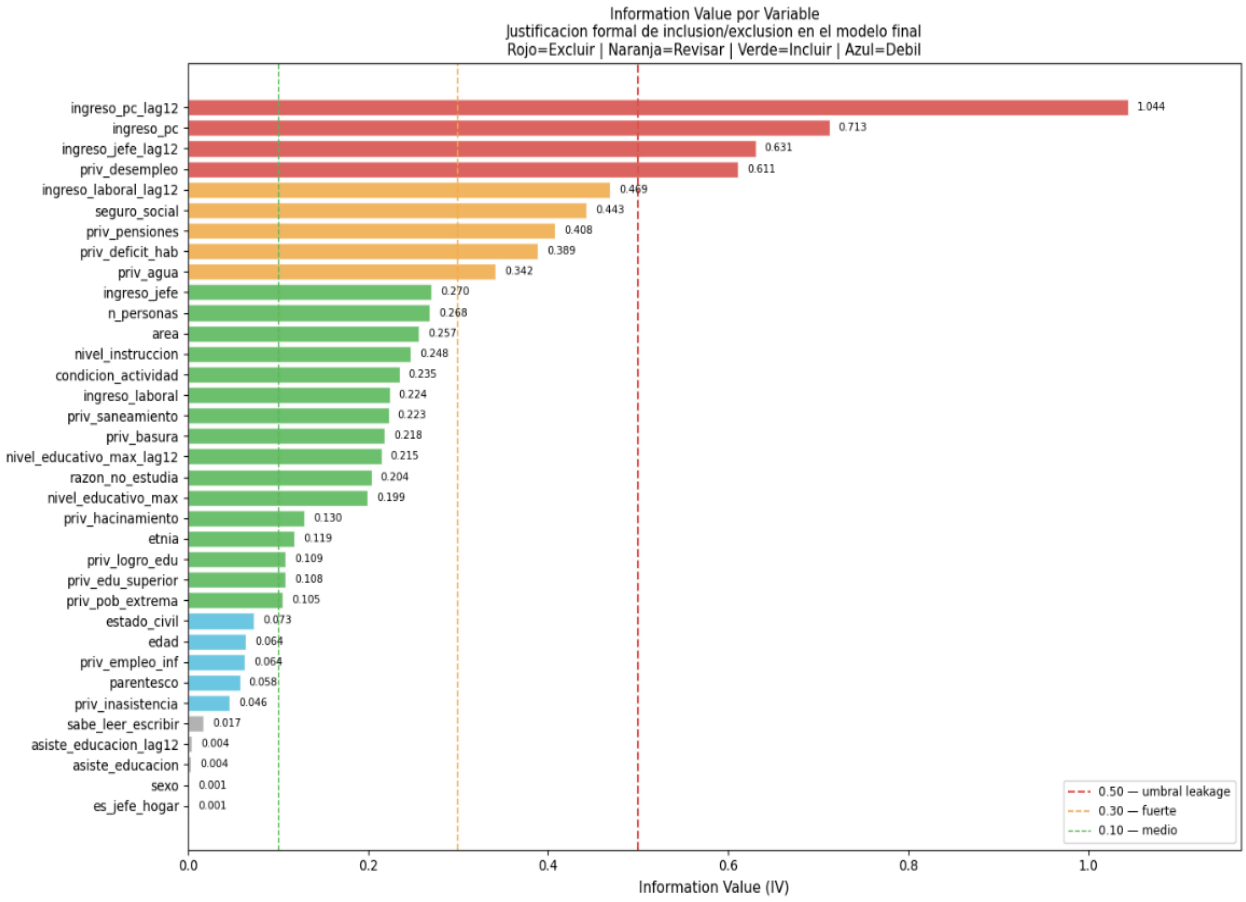
Tabla 9: *Resultados completos del Information Value*

Information Value (IV) – Selección de Variables Predictoras			
ENEMDU-INEC 2018-2026			
N	Variable	IV	Poder predictivo
1	ingreso_pc_lag12	1,0444	EXCLUIR – Leakage probable
2	ingreso_pc	0,7127	EXCLUIR – Leakage probable
3	ingreso_jefe_lag12	0,6306	EXCLUIR – Leakage probable
4	priv_desempleo	0,6114	EXCLUIR – Leakage probable
5	ingreso_laboral_lag12	0,4686	Fuerte
6	seguro_social	0,4431	Fuerte
7	priv_pensiones	0,4079	Fuerte
8	priv_deficit_hab	0,3887	Fuerte
9	priv_agua	0,3418	Fuerte
10	ingreso_jefe	0,2704	Medio – Incluir
11	n_personas	0,2683	Medio – Incluir
12	area	0,2567	Medio – Incluir
13	nivel_instruccion	0,2477	Medio – Incluir

14	condicion_actividad	0,2350	Medio – Incluir
15	ingreso_laboral	0,2241	Medio – Incluir
16	priv_saneamiento	0,2231	Medio – Incluir
17	priv_basura	0,2183	Medio – Incluir
18	nivel_educativo_max_lag12	0,2152	Medio – Incluir
19	razon_no_estudia	0,2041	Medio – Incluir
20	nivel_educativo_max	0,1994	Medio – Incluir
21	priv_hacinamiento	0,1296	Medio – Incluir
22	etnia	0,1186	Medio – Incluir
23	priv_logro_edu	0,1086	Medio – Incluir
24	priv_edu_superior	0,1082	Medio – Incluir
25	priv_pob_extrema	0,1051	Medio – Incluir
26	estado_civil	0,0732	Débil – Evaluar
27	edad	0,0641	Débil – Evaluar
28	priv_empleo_inf	0,0635	Débil – Evaluar
29	parentesco	0,0578	Débil – Evaluar
30	priv_inasistencia	0,0464	Débil – Evaluar
31	sabe_leer_escribir	0,0170	Débil – Evaluar
32	asiste_educacion_lag12	0,0045	Débil – Evaluar
33	asiste_educacion	0,0036	Débil – Evaluar
34	sexo	0,0012	Débil – Evaluar
35	es_jefe_hogar	0,0012	Débil – Evaluar

En la Figura 22 se evidencian los IV de cada variable, en la cual se identificó que 4 variables presentan IV mayor a 0.50, en línea con la literatura estas variables son sospechosamente altas y pueden causar posible fuga de datos: ingreso_pc_lag12, ingreso_pc, ingreso_jefe_lag12 y priv_desempleo. Lo cual alinea el comportamiento de las variables y la naturaleza del ingreso para el cálculo del indicador de priv_pob_extrema, mismo que pertenece a privaciones utilizadas para identificar si un hogar es pobre a través de factores multidimensionales. En base al IV, se excluyen las variables con IV mayor a 0,5 para evitar proceso de sobre ajuste. Además, se utilizan variables con poder predictivo débil (sobre todo de carácter educativo) para identificar la viabilidad de las variables educativas en el proceso de predicción.

Figura 22: *Information Value por cada variable de la base de datos*



A pesar del IV de cada variable, la selección del set de variables utilizadas en cada experimentación responde a agrupaciones de concepto para identificar tipos de variables que puedan dar una interpretación social a la predicción de pobreza. En base al objetivo del proyecto, no se descartan variables educativas, para poder estudiar tanto su interpretabilidad como las relaciones que pueden presentar en conjunto con otras variables. En la sección 5.3 se detallarán las variables utilizadas para cada iteración del modelo.

5.3. Experimentaciones sobre la base de datos global

5.3.1. Iteración de modelos sin optimización de hiperparámetros

Para evaluar los modelos utilizados, se realizó un primer experimento con un modelo stacking con Random Forest y un perceptrón multicapa (MLP), utilizando el grupo de variables sociodemográficas y de educación, los resultados se presentan a continuación:

Figura 23: *Modelo con variables de Educación – RF + MLP*

```
EXPERIMENTO 1: Educacion pura con RF + MLP

============================================================
EXPERIMENTO: Educacion_RF_MLP_v1
Variables: 17 | Balance: class_fexp
Base: ['rf', 'mlp'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.7391 0.6898    0.6878
           mlp     Base 0.7402 0.6546    0.6473
Stacking_logit Stacking 0.7433 0.6890    0.6853
```

Tabla 10: *Variables de Iteración 1*

Variables Experimentación Iteración 1
nivel_instruccion
nivel_educativo_max
nivel_educativo_max_lag12
asiste_educacion
asiste_educacion_lag12
sabe_leer_escribir
priv_inasistencia
priv_edu_superior
priv_logro_edu
edad
sexo
es_jefe_hogar
n_personas
area
parentesco
estado_civil
etnia

El modelo RF alcanzó un AUC de 0.739, por otro lado, el MLP logró obtener un AUC de 0.74, con las predicciones de estos dos modelos se llevó a cabo el modelo Stacking con regresión logística (RL) obteniendo un AUC de 0,7433. Esta primera iteración evidencia que la pobreza y la educación están estrechamente relacionadas, especialmente en regiones donde el acceso a la educación de calidad es limitado.

Figura 24: *Iteración 2: Modelo Completo con Variables Contaminadas (Benchmark)*

```
EXPERIMENTO 2: Modelo completo — BENCHMARK (leakage incluido)
ADVERTENCIA: Este modelo NO es el reportado en la tesis.

============================================================
EXPERIMENTO: Completo_RF_MLP_Logit
Variables: 35 | Balance: class_fexp
Base: ['rf', 'mlp', 'logit'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando LOGIT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.8388 0.7584    0.7537
           mlp     Base 0.8489 0.7549    0.7494
         logit     Base 0.8240 0.7419    0.7367
Stacking_logit Stacking 0.8495 0.7667    0.7623
```

Para la iteración 2 se incluye las 35 variables independientemente del valor del IV identificado, este modelo se entrenó como punto de referencia para cuantificar la diferencia que genera el leakage informacional. Los resultados de este modelo con relación a la primera iteración reflejan una diferencia de 10.6% (mayor) de AUC, aunque no presenta una mejora real en el poder predictivo, demuestra el efecto de incluir variables predictoras que directa o indirectamente contienen información sobre la variable predictiva (predicción sobre ajustada), lo cual refuerza el enfoque que tomaremos con variables de educación, al estudiar las características finales.

5.3.2. Iteraciones con Optimización de hiperparámetros mediante Optuna

Al identificar que el stacking con 3 modelos presenta mejor desempeño y con el objetivo de eliminar los sesgos que causan la selección arbitraria de hiperparámetros dentro del entrenamiento de cada modelo, RF, GBT, MLP. Se implementó un proceso de optimización sistemática utilizando la biblioteca Optuna (Akiba, Sano, Yanase, Ohta, & Koyama, 2019) el cual utiliza el algoritmo de Estimador Parzen estructurado en árbol (TPE) para explorar el espacio de hiperparámetros de manera bayesiana. La optimización se realizó con 20 ensayos para cada modelo en los que se estimó el AUC en el conjunto de validación para evitar sobreajuste de los hiperparámetros. A continuación, se presenta los rangos de búsqueda, los valores predeterminados y los resultados óptimos.

Tabla 11: *Comparativa de Hiperparámetros: Valores por Defecto vs. Óptimos (Optuna)*

Modelo	Hiperparámetro	Valor por defecto	Valor óptimo (Optuna)	AUC en validación
Logit	regParam	0,01	0,00115	0,7287
Logit	elasticNetParam	0	0,485	-
RF	numTrees	100	122	0,7447
RF	maxDepth	8	12	-
RF	minInstancesPerNode	1	25	-
GBT	maxIter	80	130	0,7511
GBT	maxDepth	5	5	-
GBT	stepSize	0,1	0,162	-
MLP	Arquitectura	[n,32,16,2]	[14,54,28,2]	0,7452
MLP	maxIter	80	185	-

Nota: n = número de variables de entrada (14). Fuente: elaboración propia con base en resultados del notebook.

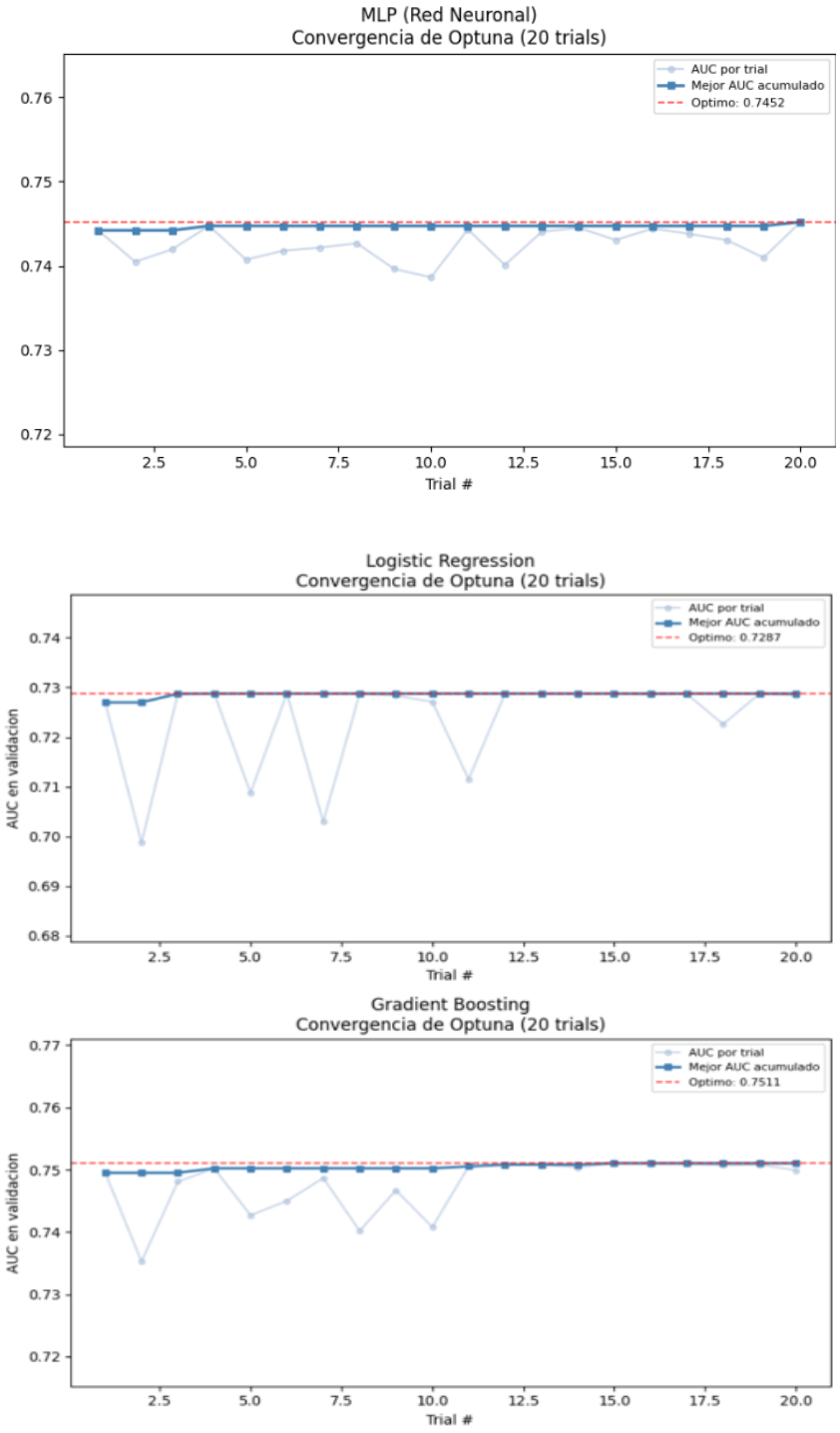
La Tabla 11 muestra que al incluir Optuna para la optimización de los hiperparámetros de los diferentes modelos, se obtuvo una mejora no significativa. La diferencia de AUC entre modelos con parámetros estándar y modelos optimizados por Optuna fue menor a 0,01. Adicionalmente, se evidencia que, si la selección de las variables se realiza con IV, los resultados de las métricas son robustas, tomando en cuenta que las variables que se seleccionan son de educación.

Figura 25: *Resumen de optimización de hiperparámetros – Optuna*

```
======================================================================
RESUMEN DE OPTIMIZACION DE HIPERPARAMETROS — Optuna (20 trials)
======================================================================

LOGISTIC REGRESSION:
  Default  : regParam=0.01, elasticNetParam=0.0
  Optimo   : regParam=0.001153, elasticNet=0.4850
  AUC VAL  : 0.7287

RANDOM FOREST:
  Default  : numTrees=100, maxDepth=8, minInstancesPerNode=1
  Optimo   : numTrees=122, maxDepth=12, minInst=25
  AUC VAL  : 0.7447

GRADIENT BOOSTING:
  Default  : maxIter=80, maxDepth=5, stepSize=0.1
  Optimo   : maxIter=130, maxDepth=5, stepSize=0.1618
  AUC VAL  : 0.7511

MLP:
  Default  : layers=[n,32,16,2], maxIter=80
  Optimo   : layers=[14,54,28,2], maxIter=185
  AUC VAL  : 0.7452
```

En la **Figura 25** se observa el resumen de los 20 trials a través de optuna, en la cual se identifica una regresión logística con AUC 0.7287, mientras que RF obtiene un AUC de 0.74, a diferencia de GBT con AUC de 0.75 (destacándose como el más alto) y finalmente MLP con AUC de 0.74.

Figura 26: *Convergencia de la Optimización de hiperparámetros de cada técnica*



En la **Figura 26** observamos la convergencia de la optimización de hiperparámetros de cada técnica, en el cual, se identifica que a través del numero de iteraciones el AUC se va maximizando en cada una de las metolodologías aplicadas.

Además, de la experimentación global, se realizaron iteraciones por zona (urbana, rural) para identificar diferencias en los procesos de predicción, este análisis se realizó en base a los resultados del capítulo 4, en el cual se observa en la sección 4.4, y que se vuelve a abordar en la sección 5.4, donde se demuestra la existencia de una diferencia entre la tasa de malos para el sector rural como para el sector urbano, con una brecha de 20 puntos porcentuales, se muestra un resumen de los experimentos realizados, en los cuales se probaron varios tipos de modelos, varios conjuntos de variables.

Tabla 12: *Tabla consolidada de todos los experimentos*

```
================================================================================================
REGISTRO COMPLETO DE EXPERIMENTOS
================================================================================================
                 experimento estrato                   enfoque    auc     f1  accuracy
       Completo_RF_MLP_Logit  Global       Completo con leakage 0.8495 0.7667    0.7623
         Educacion_RF_MLP_v1  Global            Solo Educacion 0.7433 0.6890    0.6853
     Rural_Socioecon_Sin_IPM   Rural          Socioecon Sin IPM 0.8168 0.7549    0.7250
       Rural_Solo_Educacion    Rural            Solo Educacion 0.7399 0.7102    0.6732
Final_Rural_Educacion_Limpio   Rural        Educacion Final IV 0.7367 0.7118    0.6754
    Urbano_Socioecon_Sin_IPM  Urbano          Socioecon Sin IPM 0.8023 0.7293    0.7293
       Urbano_Solo_Educacion  Urbano            Solo Educacion 0.7132 0.6599    0.6644
           Control_2_Modelos  Urbano          2 Modelos RF+GBT 0.7096 0.6567    0.6614
Final_Urbano_Educacion_Limpio Urbano        Educacion Final IV 0.7095 0.6571    0.6608
           Control_3_Modelos  Urbano      3 Modelos RF+GBT+MLP 0.7095 0.6571    0.6608
           Control_4_Modelos  Urbano 4 Modelos RF+GBT+MLP+Logit 0.7095 0.6574    0.6603

Guardado: registro_experimentos_tesis.csv
```

Como se observa en la Tabla 12, existen dos experimentaciones determinadas como finales. Estas iteraciones, corresponden a la metodología determinada en el capítulo 3, donde se indica que el modelo utilizado es un stacking de 2 capas, una capa base de 3 modelos (RF, GBT, MLP) y una capa de salida de RL, que utiliza 13 variables finales correspondientes a características de educación (todas) y sociodemográficas (con IV mayor a 0.02).

5.4. Segmentación Geográfica y Análisis por Estrato

Un análisis inicial de la distribución de la variable objetivo por área geográfica evidenció una brecha estructural, lo cual, demostró la decisión de desarrollar modelos independientes para los estratos urbano y rural. La tasa de pobreza futura en el área urbana es

del 59,3%, mientras que en el área rural se eleva al 82,7%, una diferencia de 23,4 puntos porcentuales.

Esta brecha es un reflejo de diferentes condiciones estructurales en los mecanismos de transmisión de la pobreza entre los dos territorios (rurales y urbanos), lo mencionado se respalda con los boletines técnicos del ENEMDU. La segmentación de los modelos en áreas urbanas y rurales señala las ventajas de utilizar datos diferenciados por tipo de área. Los modelos de aprendizaje automático utilizados sin diferenciarlos (segmentarlos) sobreestiman la región de pobres y rurales y lo subestiman en economías más estables y urbanas (Saito & Nomura, 2024). A continuación, se observan los resultados de los modelos de la experimentación para el área rural y urbano:

Figura 27: *Modelo Stacking (RF+MLP+GBT) para el área urbano con variables de Educación seleccionados con IV*

```
>>> EXPERIMENTO DEFINITIVO: URBANO <<<

============================================================
EXPERIMENTO: Final_Urbano_Educacion_Limpio
Variables: 14 | Balance: class_fexp
Base: ['rf', 'mlp', 'gbt'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.6955 0.6434    0.6515
           mlp     Base 0.6999 0.6303    0.6285
           gbt     Base 0.7097 0.6523    0.6626
Stacking_logit Stacking 0.7095 0.6571    0.6608
```

Figura 28: *Modelo Stacking (RF+MLP+GBT) para el área rural con variables de Educación seleccionados con IV*

```
>>> EXPERIMENTO DEFINITIVO: RURAL <<<

============================================================
EXPERIMENTO: Final_Rural_Educacion_Limpio
Variables: 14 | Balance: class_fexp
Base: ['rf', 'mlp', 'gbt'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.7310 0.6789    0.6358
           mlp     Base 0.7270 0.6630    0.6177
           gbt     Base 0.7431 0.7780    0.8263
Stacking_logit Stacking 0.7367 0.7118    0.6754

============================================================
RESUMEN DEFINITIVO — Resultados para la tesis
============================================================
  URBANO: AUC=0.7095 | F1=0.6571 | Acc=0.6608
  RURAL: AUC=0.7367 | F1=0.7118 | Acc=0.6754
```

Las variables utilizadas para la experimentación por zona se pueden observar a continuación:

Tabla 13: *Variables Experimentación: enfoque educación*

Variables Experimentación: enfoque educación
nivel_instruccion
nivel_educativo_max_lag12
razon_no_estudia
asiste_educacion
asiste_educacion_lag12
sabe_leer_escribir
priv_logro_edu
priv_edu_superior
priv_inasistencia
n_personas
edad
edad_lag12
sexo
etnia

En la Tabla 13, se muestra las 14 variables utilizadas para la experimentación de educación con las variables seleccionadas, sin embargo, para el modelamiento final se excluye la variable edad_lag12, ya que no aporta información significativa al modelo, al representar la edad de la persona hace 12 meses y además integrar la variable de edad en el punto de análisis. El resumen

de los resultados del proceso de identificación de hiper parámetros se puede observar a continuación:

Figura 29: *Modelo Stacking - Experimento con hiperparámetros óptimos (Optuna)*

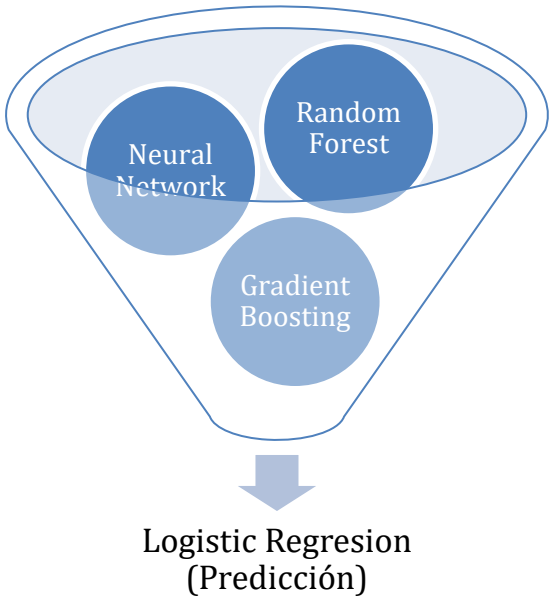
```
============================================================
COMPARACION: Hiperparametros por defecto vs Optuna
============================================================
  Estrato    Modelo                          AUC
  ------------------------------------------------
  Urbano     Default (Sin Optuna)       0.7095
  Rural      Default (Sin Optuna)       0.7367
  Urbano     Optuna (Optimo)            0.7096
  Rural      Optuna (Optimo)            0.7412

Si Optuna mejora > 0.005 AUC: la optimizacion es significativa.
Si la mejora es menor: los hiperparametros por defecto eran razonables.
```

Se observa en la Figura 29, tanto para el modelo rural como para el modelo urbano, la optimización de hiperprámetros utilizando Optuna, donde se presenta una ligera mejora en el rendimiento de la predicción. Esta metodología se encuentra respaldada por análisis estadísticos, por lo tanto, se utilizan estos resultados para el entrenamiento de los modelos finales. Además, se identificó el uso de una variable que no tiene un sentido.

Capítulo 6. Desarrollo de modelos diferenciados para los estratos urbano y rural

Con base en la experimentación realizada en la sección anterior, se identificó que el mejor esquema de modelamiento para predecir la pobreza multidimensional futura corresponde a un esquema de stacking de 2 capas, aplicado de forma diferenciada para cada zona (rural, urbana). Para lo cual, se utiliza para cada zona, una primera capa de 3 modelos base (RF, GBT y MLP) y un metamodelo de salida (RL). Este esquema de predicción permite utilizar la técnica de bagging y boosting, complementando la reducción de ruido del árbol de decisión y la identificación granular del gradient boosting, que, en complemento con una red neuronal, permite identificar relaciones complejas, que presentan resultados robustos, los cuales son interpretados por la regresión logística de salida, suavizando y procesando los resultados de los modelos base.

Figura 30: *Esquema Modelo Stacking – Modelos finales*



Se replicó el mismo esquema y conjunto de variables educativas y sociodemográficas en ambas zonas, de manera que las diferencias que de desempeño e importancia de variables se deben a la heterogeneidad estructural de las dos poblaciones urbana y rural. A continuación, se presentan las variables utilizadas para el modelamiento:

Tabla 14: *Variables finales para entrenamiento*

Tipo de Variable	Nombre de Variable
Educación	nivel_instruccion
	nivel_educativo_max_lag12
	razon_no_estudia
	asiste_educacion
	asiste_educacion_lag12
	sabe_leer_escribir
	priv_logro_edu
	priv_edu_superior
	priv_inasistencia
Sociodemográficas	n_personas
	edad
	sexo
	etnia

Como se puede observar en la Tabla 14, se incluyen las 9 variables educativas y 4 sociodemográficas que presentan un IV menor a 0.5. Además, se incluyen variables con baja capacidad predictiva, como el "*sexo"* y *"saber leer y escribir"*, para su análisis y comparativa de resultados de los modelos por zona. Con base a esto, se muestra los resultados sobre las variables, su importancia y los resultados del entrenamiento de los modelos finales.

6.1. Importancia de Variables: Hallazgos sobre el Capital Humano y la Desigualdad Territorial

El modelo final es un Stacking Ensemble, donde una regresión logística aprende a combinar las predicciones de 3 algoritmos base: Random Forest, Gradient Boosting Trees, y Multilayer Perceptron para obtener una predicción final. El resultado se obtiene mediante la siguiente ecuación:

$$P(y = 1) = \sigma(\beta_0 + \beta_1 RF + \beta_2 GBT + \beta_3 MLP)$$

Dónde:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Esta ecuación permite que la regresión logística aprenda y confía en cada modelo, ya que no combina variables originales, sino combina probabilidades. Lo cual, es importante porque hace que el stacking disminuya el sesgo que puede tener un solo algoritmo (Hastie, Tibshirani, & Friedman, 2008).

Modelo de Stacking – Área Urbana– Interpretación de Betas

$$P(\text{Pobre}) = \frac{1}{1 + e^{-[(-2.4700+0.7440P_{RF}+4.2687P_{GBT}+0.7440P_{MLP})]}}$$

Dónde:

P_{RF}: Probalidad estimada por modelo Random Forest

P_{GBT}: Probabilidad estimada por modelo Gradient Boosting

P_{MLP}: Probabilidad estimada por modelo Red Neuronal

La ecuación presenta que el coeficiente asociado al modelo Gradient Boosting (4,2687) es superior al de los demás modelos. Lo cual evidencia que, la combinación de predicciones, el metamodelo asigna la mayor confianza hacia las probabilidades obtenidas con el modelo de Gradient Boosting. En otras palabras, cuando este modelo aumenta la probabilidad de que un hogar sea multidimensionalmente pobre, el modelo final responde con un aumento mucho más pronunciado en la probabilidad estimada que si dicho aumento provenga de Random Forest o la Red Neuronal.

Modelo de Stacking – Área Rural – Interpretación de Betas

$$P(\text{Pobre}) = \frac{1}{1 + e^{-[(-2.5239+0.8579P_{RF}+2.7376P_{GBT}+0.2.8589P_{MLP})]}}$$

Dónde:

P_{RF}: Probalidad estimada por modelo Random Forest

P_{GBT}: Probabilidad estimada por modelo Gradient Boosting

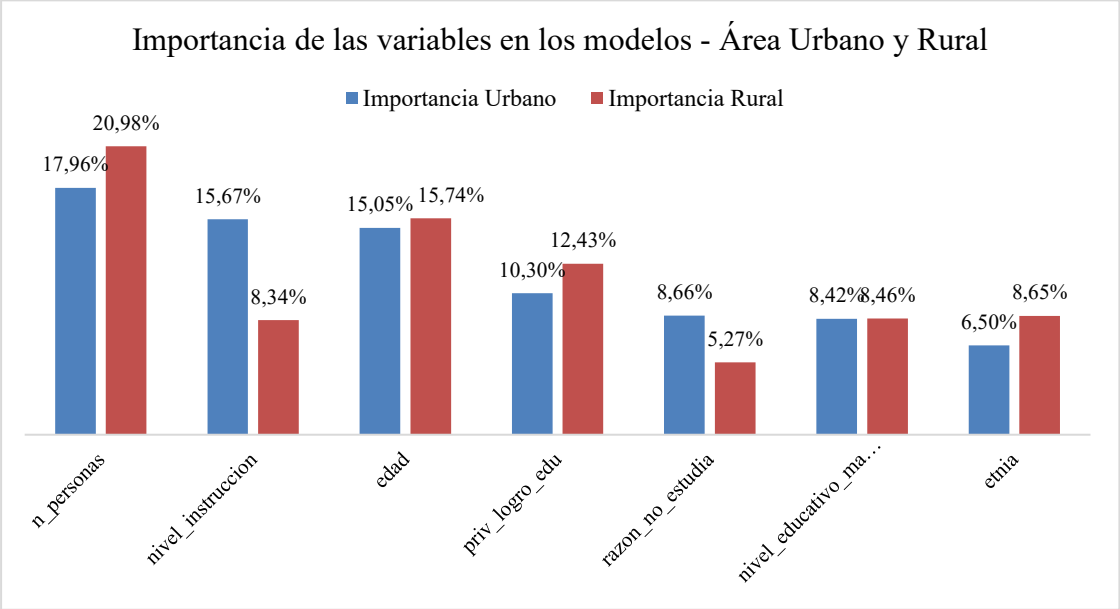
P_{MLP}: Probabilidad estimada por modelo Red Neuronal

En comparación al modelo urbano, el metamodelo del área rural marca importancia tanto en el modelo de Gradient Boosting (2,7376), como en el modelo De redes Neurales

(2,8589). Lo mencionado denota que la pobreza multidimensional crea relaciones mucho más complejas y no lineales entre las variables independientes, por lo tanto, el modelo debe combinar simultáneamente modelos Gradient Boosting para capturar reglas jerárquicas y la capacidad de las redes neuronales para aprender patrones complejos.

Interpretación de la importancia de variables

Figura 31: *Importancia de las variables en el modelo - Área Urbano y Rural*



En el modelo del área urbano se observa que la variable número de personas aporta con 17.96%, mientras que para el área rural aporta con 20.98%, con 3.02% de diferencia. Lo mencionado tiene un respaldo teórico basado en el estudio de (The Global MPI, 2024),el bienestar de un hogar no depende solamente de factores como el ingreso del hogar, sino también por la capacidad que tiene para distribuir recursos entre el número de integrantes, ya que los integrantes de un hogar deben compartir, alimentación, vivienda, educación, salud, tiempo de cuidado entre otros factores de bienestar. En el Ecuador, existen estudios en los que muestran que los hogares con un mayor número de integrantes presentan una mayor relación de pobreza multidimensional, en especial en el área rural (predomina familias extensas) por lo tanto, existe una dependencia económica entre sus integrantes (Morales & Mideros, Análisis de la pobreza multidimensional en los hogares de la agricultura familiar campesina en el Ecuador, 2009- 2019, 2021). Como se observa en los resultados el predictor de la pobreza multidimensional es el número de personas, lo que sugiere que las presiones demográficas

sobre los recursos del hogar aumentan la probabilidad de privaciones concurrentes, la diferencia de la importancia de esta variable de 3,02 puntos porcentuales refleja que las familias de las zonas rurales que son numerosas enfrentan mayores limitaciones para acceder a oportunidades de educación, empleo e infraestructura.

En cuanto a la variable nivel de educación su importancia para el modelo urbano es de 15.67% y en el modelo rural es de 8,34%, disminuye 8%. En las ciudades la educación es un mecanismo de movilidad social mucho más fuerte, la escolaridad se asocia a más oportunidades laborales, mayor formalidad en los empleos, ingresos estables y acceso a seguridad social. Por otro lado, en las zonas rurales, aunque sigue siendo importante, la pobreza se ve agravada por la infraestructura, el acceso a los servicios, el aislamiento geográfico y la actividad agrícola. Es decir, el nivel de educación explica una parte de la pobreza multidimensional; el nivel de educación es uno de los principales mecanismos de acumulación de capital humano en las ciudades donde el mercado laboral premia más intensamente la educación académica. Aunque sigue siendo un factor importante en las zonas rurales, su poder predictivo disminuye porque la pobreza también está relacionada con las limitaciones estructurales de la zona (Mora, Pobreza Multidimensional, 2019).

Por otro lado, la edad en el modelo Urbano tiene 15.05% de aporte, mientras que en el modelo rural 15.74%, en este escenario podemos identificar que la edad afecta de manera transversal a ambas áreas, la literatura evidencia que los niños, jóvenes y adultos mayores, presentan mayor vulnerabilidad de pobreza multidimensional, debido a su dependencia económica o menor capacidad de general ingresos (Mora, Defining and measuring multidimensional poverty. Exploring poverty in Ecuador 2006-2010, 2011).

6.2. Interpretación de los valores SHAP del modelo Stacking por área

Metodología aplicada

Para complementar el análisis de significancia global obtenido con los modelos RF y GBT, Lundberg y Lee (2017) se propuso el método SHAP (SHapley Additive Explanations), este método permite interpretar las contribuciones individuales de cada variable a las predicciones del modelo stacking completo (preprocesamiento + 3 modelos base + metamodelo), tratando como una función de caja negra utilizando el algoritmo de KernelExplainer. Este análisis es necesario dado que los modelos de árboles implementados en

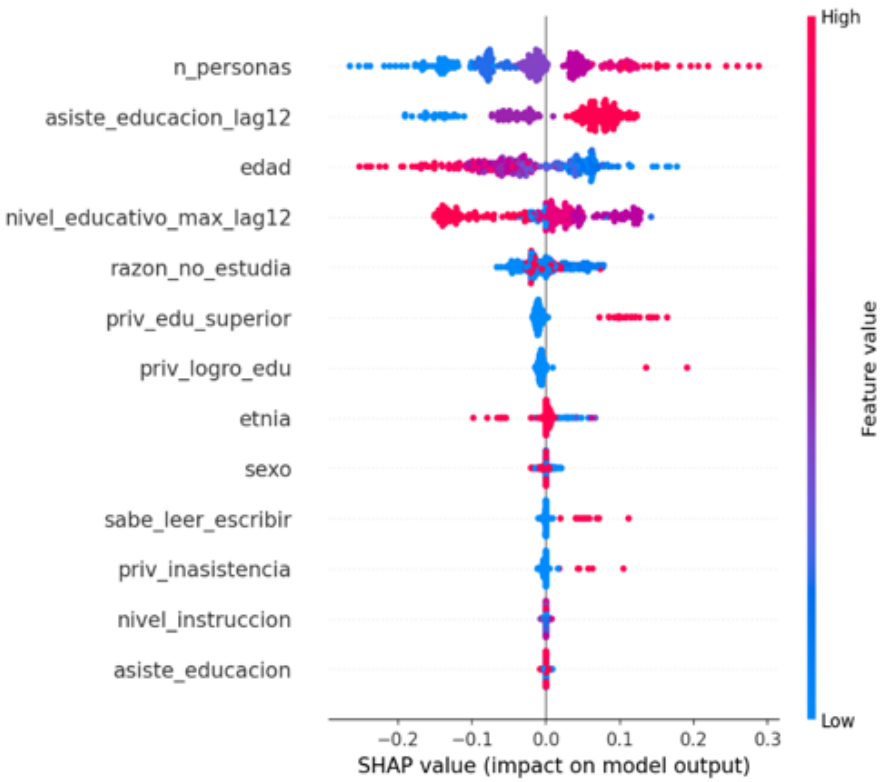
Spark ML no admiten explicadores específicos como TreeExplainer. Para cada segmento se utilizó una muestra de referencia (background) de 100 observaciones de train, y se explicó una muestra de 300 observaciones de test, con 100 coaliciones sintéticas por observación.

El tamaño de las muestras utilizadas para este cálculo responde a un balance explícito entre precisión de la aproximación y factibilidad computacional. La muestra de referencia o background, compuesta por 100 observaciones de train, corresponde al tamaño sugerido por los propios autores del método como punto de partida habitual para KernelExplainer (Lundberg y Lee, 2017), y ha sido documentada en estudios posteriores como un tamaño de background comúnmente utilizado en aplicaciones con restricciones computacionales (Yuan et al., 2022). El número de coaliciones sintéticas por observación, fijado en 100, se ubica por debajo del valor por defecto que sugiere la implementación estándar de SHAP, equivalente a 2M+2048 coaliciones, donde M es el número de variables, lo que para las 13 variables consideradas en este estudio representaría alrededor de 2074 coaliciones por observación. Esta reducción se adoptó de forma deliberada como una decisión de compromiso computacional, y no como un valor arbitrario, dado el costo de cada evaluación del modelo completo de stacking. Finalmente, la muestra de 300 observaciones explicadas por segmento se definió con el propósito de obtener estimaciones agregadas suficientemente estables (importancia media y distribución de valores SHAP por variable), sin que el tiempo de cómputo resultara prohibitivo para los recursos disponibles en esta investigación.

Los valores SHAP se calcularon directamente sobre las variables originales previamente especificadas y sin escalar, facilitando la interpretación directa de la contribución de cada predictor a su escala natural.

6.3. Zona urbana

Figura 32: *SHAP beeswarm - segmento urbano*



En la Figura 32 se presentan los resultados para la zona urbana, las cuatro variables con mayor significancia promedio corresponden a la variable "número de personas" (0.0782, presentando la mayor dispersión de los valores SHAP), "participación educativa rezagada" (0.0734), "edad" (0.0672) y "nivel máximo de educación del hogar" (0.0647). Este análisis corrobora los resultados del estudio de significancia de las variables, obtenidos previamente utilizando algoritmos (RF y GBT), lo que confirma la solidez de estos predictores con diferentes enfoques interpretativos.

Además, se muestra que los hogares con mayor número de miembros (puntos rojos) tienden a tener valores SHAP positivos, lo que aumenta la probabilidad de pobreza multidimensional, mientras que los hogares más pequeños (puntos azules) cambian la predicción a valores negativos. Este comportamiento es consistente con el enfoque de Alkire-Foster y la evidencia proporcionada por la Oxford Poverty and Human Development Initiative (OPHI) que hemos revisado a lo largo de este estudio, según la cual, los hogares grandes tienen más dificultades para distribuir los recursos escasos entre todos sus miembros.

Respecto a la variable "participación educativa rezagada" es el segundo predictor más importante, en el contexto educativo, muestra que el historial de participación en educación en el período anterior afecta significativamente la clasificación final de pobreza multidimensional,

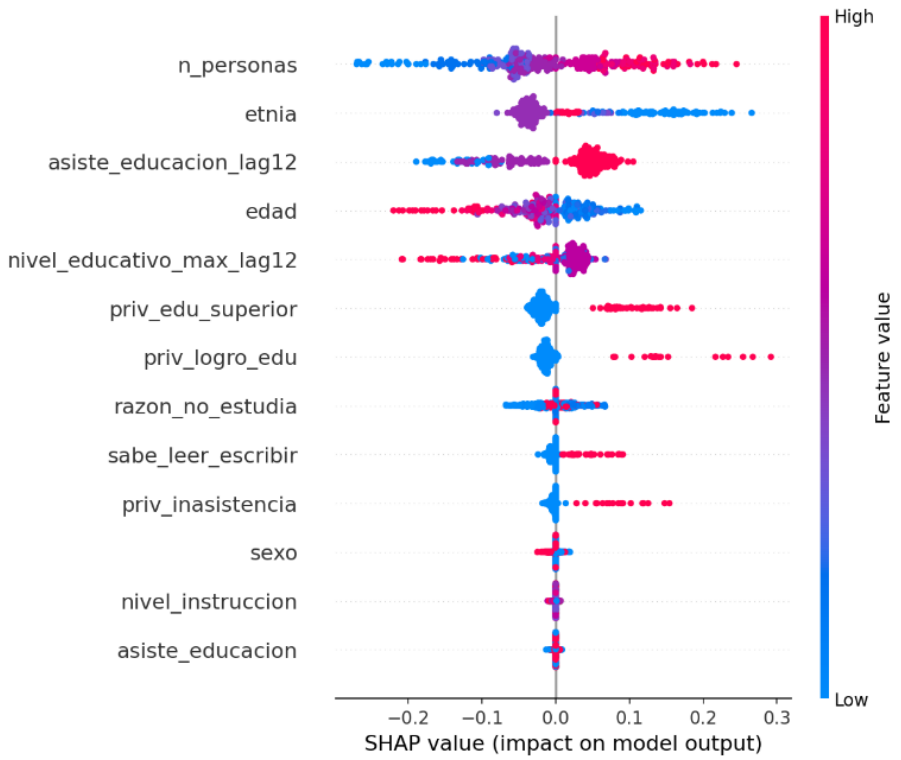
lo que demuestra que la resiliencia en el sistema educativo es un factor importante contra la pobreza urbana multidimensional. El modelo muestra que las trayectorias educativas, reflejan procesos de acumulación de capital humano que promueven el bienestar en el hogar.

El comportamiento observado para la variable “edad” también es importante, los valores de edad bajos (puntos azules) se concentran principalmente en la región positiva del eje SHAP, mientras que las edades más altas (puntos rojos) tienden a sesgar las predicciones hacia valores negativos, por lo tanto, en el segmento urbano, los hogares con miembros más jóvenes tienen más probabilidades de experimentar pobreza multidimensional, posiblemente debido a una menor estabilidad laboral y acumulación de capital humano.

Por otra parte, “nivel máximo de educación del hogar” presentan un comportamiento inverso, los hogares con niveles más altos de educación tienden a registrar valores SHAP negativos, lo que reduce la probabilidad de pobreza multidimensional. Finalmente, en contraste, con las variables como género, alfabetización, nivel educativo y participación actual en educación muestran valores SHAP cercanos a cero y dispersión reducida, lo que indica que su contribución individual es limitada cuando se incluye en el modelo información de más variables explicativas.

6.4. Zona rural

Figura 33: *SHAP beeswarm - segmento rural*



En la Figura 33, en la zona rural se observa una estructura similar en la variable "número de personas del hogar" (0.0826), es el predictor más relevante del modelo al igual que en el área urbano. A pesar de ello, a diferencia del contexto urbano, la segunda variable en importancia corresponde a la "etnia" (0,0645), que desplaza a la "edad" y al "nivel educativo". Este es uno de los hallazgos principales del estudio, ya que la etnicidad juega un rol importante para explicar la pobreza multidimensional en la zona rural. Esto demuestra que etnicidad tiene una contribución positiva y significativa al modelo, reflejando desigualdades que afectan principalmente a los pueblos indígenas que viven en áreas rurales, como lo vimos en el análisis de los capítulos anteriores de exploración de datos.

La variable "participación educativa rezagada" tiene una significancia moderada (0.0629), lo que confirma que la continuidad de la educación es un factor importante tanto en áreas urbanas como rurales. Sin embargo, en este último contexto cobran mayor relevancia variables relacionadas con la falta de educación, como la "privación por educación superior" (0,0320) y la "privación por nivel educativo" (0,0222). Este comportamiento está en línea con la evidencia proporcionada por (CEPAL, 2026), que destacan que las zonas rurales de América Latina enfrentan mayores dificultades para acceder y completar niveles superiores de educación debido a restricciones geográficas, económicas y de oferta educativa.

En este contexto educativo, los efectos de las variables relacionadas con la privación educativa, como la "privación por educación superior" (0,0320) y la "privación por nivel educativo" (0,0222), también son importantes para explicar la pobreza rural. La variable "edad" es menos importante respecto al segmento urbano (0,0452), aunque mantiene un patrón explicativo similar y las personas mayores reducen parcialmente el riesgo de pobreza multidimensional, mientras que los hogares conformados por personas más jóvenes tienen una mayor tendencia a aumentar la probabilidad de pobreza multidimensional.

6.5. Resultados de modelamiento

El metamodelo de cada zona se entrenó utilizando las probabilidades que generan los modelos base sobre la partición de validación (que no fue vista durante el entrenamiento de los modelos base), mientras que la partición de test se mantuvo completamente aislada, siendo usada únicamente para evaluar la capacidad de generalización final del modelo. El umbral de clasificación de cada modelo (0/1) se determinó de forma independiente mediante el índice de Youden. Este índice es una medida ampliamente utilizada en la literatura biomédica y de ciencias sociales para seleccionar el umbral que maximiza la capacidad discriminativa del modelo, al equilibrar la sensibilidad y la especificidad (Youden, 1950). Matemáticamente, el Índice de Youden se define como:

$$J = \text{Sensibilidad} + \text{Especificidad} - 1$$

Donde el valor óptimo de J, se encuentra entre 0 y 1, y el umbral que maximiza este índice representa el punto que ofrece el mejor comportamiento entre verdaderos positivos y verdaderos negativos, minimizando tanto los falsos negativos como los falsos positivos en términos globales (Fluss et al., 2005; Perkins & Schisterman, 2006).

La elección del Índice de Youden resulta particularmente adecuada en contextos de predicción de riesgo social como la pobreza multidimensional, donde se busca un equilibrio razonable entre la detección de hogares en riesgo (sensibilidad) y la correcta clasificación de aquellos que no lo están (especificidad), evitando tanto la subestimación como la sobreestimación del riesgo.

Modelo: Zona Urbana

El threshold (indice de Youden) del modelo urbano final es 0.5943.

A continuación, se presentan las métricas de modelamiento final, resultados:

Tabla 15: *Resultados del modelo de zona urbana.*

Modelo		**RF**			**GBT**			**MLP**			**Stacking Final**	
particion	**train**	**val**	**test**	**train**	**val**	**test**	**train**	**val**	**test**	**train**	**val**	**test**
auc	0.714	0.711	**0.702**	0.716	0.71	**0.703**	0.708	0.712	**0.704**	0.716	0.715	**0.707**
gini	0.428	0.422	**0.405**	0.432	0.42	**0.405**	0.416	0.424	**0.407**	0.431	0.429	**0.414**
accuracy	0.641	0.635	**0.627**	0.65	0.642	**0.635**	0.649	0.647	**0.641**	0.646	0.64	**0.633**
precision	0.752	0.758	**0.741**	0.739	0.741	**0.727**	0.727	0.736	**0.723**	0.744	0.751	**0.736**
recall	0.593	0.567	**0.552**	0.637	0.613	**0.6**	0.658	0.634	**0.622**	0.618	0.592	**0.577**
specificity	0.712	0.734	**0.731**	0.669	0.686	**0.685**	0.636	0.666	**0.667**	0.687	0.711	**0.711**
f1	0.663	0.649	**0.633**	0.684	0.671	**0.657**	0.691	0.681	**0.668**	0.676	0.662	**0.647**

Se muestra en la Tabla 15 que el modelo final alcanza, en la partición de test, un AUC de 0.7069 (Gini = 0.4138), lo que supera al mejor de los tres modelos base (RF, GBT, MLP). Es decir, la combinación de los tres algoritmos mediante el metamodelo RL aporta una ganancia moderada de poder discriminante, respecto al mejor modelo base (MLP). El nivel de discriminación alcanzado por el modelo, tanto en su versión de stacking como en cada uno de los modelos base, se ubicaría en un rango moderado (valores de AUC entre 0.70 y 0.80 suelen clasificarse como aceptables o moderados, y no como excelentes o sobresalientes), por lo que el modelo no debe interpretarse como una herramienta de alta precisión predictiva en términos absolutos, sino como un instrumento con una capacidad de discriminación superior al azar y consistente entre sus particiones.

El modelo presenta una utilidad práctica, que debe evaluarse en función del uso específico que se le dé y no exclusivamente a partir del valor de las métricas. Se observa también que la caída del AUC entre train (0.7157) y test (0.7069) es de apenas 0.0088 puntos, lo que indica un nivel de sobreajuste bajo y una capacidad de generalización temporal adecuada.

En cuanto al balance entre precisión (precision) y sensibilidad (recall), el modelo de stacking prioriza la precisión (0.7355) y la especificidad (0.7107) por sobre la sensibilidad (recall 0.5771), lo cual es consistente con el peso que el metamodelo termina asignando a GBT

y, sobre todo, al MLP, que es el modelo base con mejor comportamiento de sensibilidad entre los tres, dentro de la combinación final. Este resultado no debería leerse únicamente como un rasgo técnico del modelo, sino que tiene una implicación práctica que merece atención explícita: una sensibilidad de 0.5771 significa que el modelo deja de identificar correctamente a poco más del 42 por ciento de las personas que efectivamente transitarán a la pobreza multidimensional en el horizonte de doce meses considerado, error que en la matriz de confusión de la partición de test se traduce en 14943 falsos negativos. De acuerdo con esto se presenta la matriz de confusión:

Tabla 16: *Matriz de confusión modelo urbano.*

		Predicho	
		No pobre (0)	***Pobre (1) modelo***
Real	***No pobre (0)***	*18013*	*7332*
	Pobre (1)	*14943*	*20390*

En la Tabla 16 se observa un total de 60,678 observaciones en test, el modelo identifica correctamente al 67% de los casos. Se observa que el número de falsos negativos (14,943 personas clasificadas como no pobres, que en realidad transitarán a la pobreza multidimensional) es superior al de falsos positivos (7,332 personas clasificadas como pobres por el modelo, que en realidad no poseen la característica de pobre multidimensional), por lo que, con el umbral óptimo obtenido, el modelo tendería a subestimar el riesgo de pobreza futura en el área urbana.

Para poder evaluar el desempeño de la predicción, se realizó un análisis de la tabla de ganancias respecto a la probabilidad resultado del modelo final. Las tablas de ganancia constituyen una herramienta fundamental en la evaluación de modelos de clasificación binaria, particularmente en contextos de riesgo social como la predicción de pobreza multidimensional. A diferencia de métricas agregadas como el AUC-ROC, estas tablas permiten examinar el comportamiento del modelo a lo largo de todo el espectro de riesgo, evaluando simultáneamente poder de discriminación, estabilidad, calibración y utilidad operativa para la focalización de políticas (Siddiqi, 2012). A continuación, se muestran los resultados:

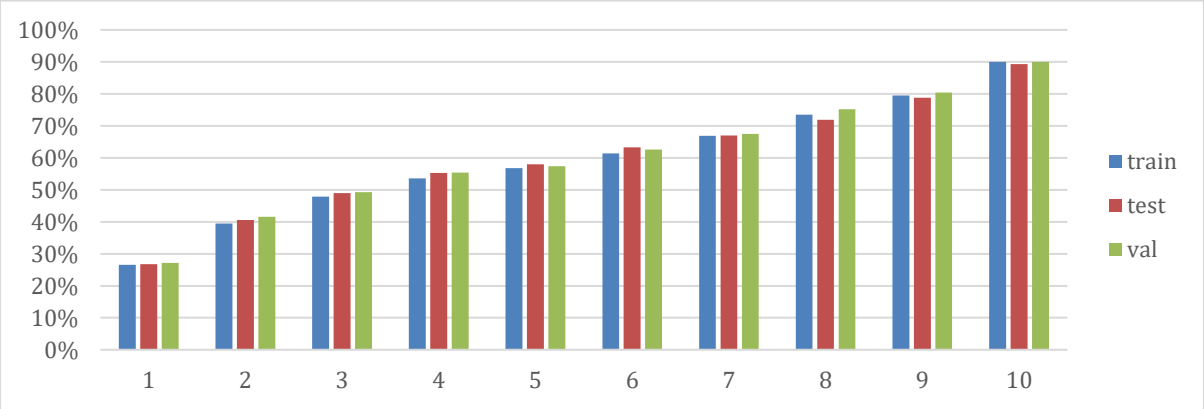
Tabla 17: *Tabla de ganancias del modelo urbano*

Partición	Decil	conteo	Malos	Buenos	Tasa malos	Odds	Malos acum	N acum	Tasa malos acum	Prob prom
test	1	7222	1933	5289	26.77%	2.74	1933	7222	26.80%	22.40%
test	2	6680	2709	3971	40.55%	1.47	4642	13902	33.40%	34.50%
test	3	6391	3128	3263	48.94%	1.04	7770	20293	38.30%	43.50%
test	4	6394	3533	2861	55.25%	0.81	11303	26687	42.40%	50.80%
test	5	5936	3440	2496	57.95%	0.73	14743	32623	45.20%	56.30%
test	6	6098	3859	2239	63.28%	0.58	18602	38721	48.00%	61.50%
test	7	5916	3966	1950	67.04%	0.49	22568	44637	50.60%	67.70%
test	8	5565	3999	1566	71.86%	0.39	26567	50202	52.90%	74.60%
test	9	5604	4416	1188	78.80%	0.27	30983	55806	55.50%	80.30%
test	10	4872	4350	522	89.29%	0.12	35333	60678	58.20%	86.90%
train	1	22600	5997	16603	26.54%	2.77	5997	22600	26.50%	22.80%
train	2	22665	8948	13717	39.48%	1.53	14945	45265	33.00%	34.60%
train	3	22521	10778	11743	47.86%	1.09	25723	67786	37.90%	43.50%
train	4	22618	12121	10497	53.59%	0.87	37844	90404	41.90%	50.80%
train	5	22579	12816	9763	56.76%	0.76	50660	112983	44.80%	56.30%
train	6	22593	13866	8727	61.37%	0.63	64526	135576	47.60%	61.50%
train	7	22644	15138	7506	66.85%	0.5	79664	158220	50.40%	67.80%
train	8	22545	16581	5964	73.55%	0.36	96245	180765	53.20%	74.70%
train	9	22592	17963	4629	79.51%	0.26	114208	203357	56.20%	80.30%
train	10	22594	20341	2253	90.03%	0.11	134549	225951	59.50%	87.00%
val	1	6229	1693	4536	27.18%	2.68	1693	6229	27.20%	22.60%
val	2	5843	2428	3415	41.55%	1.41	4121	12072	34.10%	34.60%
val	3	5573	2746	2827	49.27%	1.03	6867	17645	38.90%	43.50%
val	4	5720	3170	2550	55.42%	0.8	10037	23365	43.00%	50.80%
val	5	5241	3010	2231	57.43%	0.74	13047	28606	45.60%	56.30%
val	6	5602	3507	2095	62.60%	0.6	16554	34208	48.40%	61.50%
val	7	5122	3455	1667	67.45%	0.48	20009	39330	50.90%	67.70%
val	8	4971	3738	1233	75.20%	0.33	23747	44301	53.60%	74.60%
val	9	5158	4146	1012	80.38%	0.24	27893	49459	56.40%	80.30%
val	10	5018	4516	502	90.00%	0.11	32409	54477	59.50%	87.00%

En la Tabla 17 se observa que, en la partición de test, la tasa de malos (ratio de pobreza) pasa de 26.8% en el primer decil a 89.3% en el décimo, lo que representa una brecha de 62.5 puntos porcentuales entre el segmento de menor y de mayor pobreza. Este comportamiento es similar en las particiones de entrenamiento y validación, con rangos de 26.5% a 90.0% y de 27.2% a 90.0% respectivamente. La ausencia de alteraciones de tendencia entre deciles consecutivos y el correcto ordenamiento de riesgo es consistente en todo el rango de

probabilidades, esta propiedad es más exigente que el AUC agregado y esencial para la confianza en aplicaciones de ranking (Siddiqi, 2012). A continuación, se presenta la tasa de pobreza por decil.

Figura 34: *Tasa de malos para el modelo Urbano*

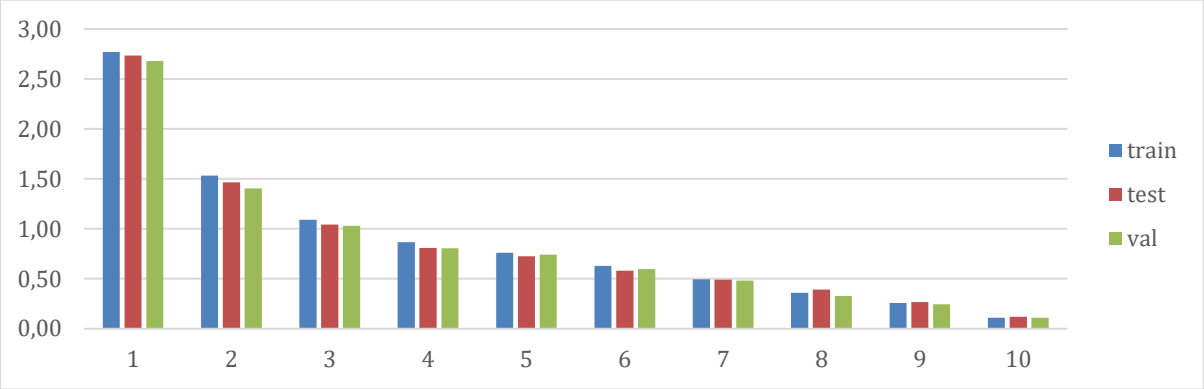


También se estudia el comportamiento del Kolmogorov Smirnov (KS), esta métrica mide la diferencia máxima entre las distribuciones acumuladas de las clases "malos" y "buenos" (pobres y no pobres en el estudio), constituyendo un indicador robusto de la capacidad del modelo para separar los extremos de la distribución de riesgo (Thomas et al., 2017).

En la partición de test, esta distancia máxima alcanza un valor de 28.8 puntos porcentuales, y se ubica en el decil 5, punto en el cual el 45.2% de todas las personas pobres de la muestra ya ha sido acumulado, frente a apenas el 16.4% de las personas no pobres acumuladas hasta ese mismo decil.

La tabla de ganancias muestra además la evolución de odds, definidos como la razón entre el número de no pobres y el número de pobres dentro del decil, permiten interpretar la clasificación binaria, especialmente en ámbitos de riesgo crediticio, médico y de políticas sociales (Siddiqi, 2012). A continuación, se muestra su evolución:

Figura 35: *Odds para el modelo Urbano*



En la Figura 35, se observa que los odds siguen una misma tendencia para todas las particiones, donde, para test comienza con 2.74 en el primer decil y llega a 0.12 en el décimo, mostrando una gran capacidad de discriminación de riesgo, además, se interpreta que, desde el cuarto decil, la mayoría de las personas dentro de cada decil pasa a ser pobre.

Modelo: Zona Rural

El threshold (índice de Youden) del modelo urbano final es 0.5927

A continuación, se presentan las métricas de modelamiento resultados:

Tabla 18: *Resultados de métricas modelo rural*

Modelo		**RF**			**GBT**			**MLP**			**Stacking Final**	
particion	**train**	**val**	**test**	**train**	**val**	**test**	**train**	**val**	**test**	**train**	**val**	**test**
auc	0.749	0.735	**0.737**	0.757	0.734	**0.739**	0.741	0.737	**0.741**	0.756	0.74	**0.743**
gini	0.499	0.47	**0.473**	0.514	0.467	**0.478**	0.481	0.474	**0.482**	0.512	0.479	**0.486**
accuracy	0.643	0.635	**0.622**	0.652	0.636	**0.627**	0.649	0.633	**0.62**	0.632	0.622	**0.61**
precision	0.92	0.917	**0.916**	0.922	0.915	**0.917**	0.914	0.915	**0.916**	0.926	0.921	**0.923**
recall	0.624	0.612	**0.596**	0.633	0.615	**0.602**	0.636	0.61	**0.593**	0.604	0.591	**0.574**
specificity	0.737	0.744	**0.745**	0.739	0.735	**0.745**	0.71	0.736	**0.745**	0.766	0.766	**0.776**
f1	0.744	0.734	**0.722**	0.751	0.736	**0.727**	0.75	0.732	**0.72**	0.731	0.72	**0.708**

En la Tabla 18 se observa que el modelo final alcanza, en la partición de test, un AUC de 0.743 (Gini de 0.486), superando a los tres modelos base considerados de forma individual (RF con 0.737, GBT con 0.739 y MLP con 0.741). El modelo de stacking muestra una ganancia de poder discriminante moderada, del orden de dos a seis milésimas de AUC frente al mejor modelo base individual. Al igual que en el segmento urbano, este valor de AUC se ubicaría en

un rango moderado de acuerdo con los criterios generales de la literatura de clasificación, por lo que, pese a tratarse del mejor desempeño discriminativo entre los cuatro modelos y los dos segmentos analizados en este estudio, no debe interpretarse como evidencia de una capacidad predictiva sobresaliente, sino como una mejora incremental y consistente. La diferencia entre las métricas de train, validación y test refleja un nivel de sobreajuste bajo y una capacidad de generalización adecuada.

En cuanto al balance entre precisión y recall, el modelo de stacking rural prioriza de forma marcada la precisión (0.923) y alcanza la especificidad más alta de los cuatro modelos evaluados (0.776). Esta priorización de la precisión y la especificidad ocurre, sin embargo, a costa de la sensibilidad (0.574), la más baja de los cuatro modelos, y resulta consistente con el peso que el metamodelo asigna a RF como el algoritmo base más influyente dentro de la capa meta, seguido de GBT, que es, de los tres modelos base, el que presenta el mejor comportamiento de recall en la partición de test (0.602). Una sensibilidad de 0.574 implica que el modelo deja de identificar correctamente a cerca del 43 por ciento de las personas que efectivamente se encuentran en situación de pobreza multidimensional futura en el área rural, lo que en la matriz de confusión de la partición de test corresponde a 7794 falsos negativos sobre un total de 18314 casos positivos reales. Esta proporción de exclusión resulta, en términos relativos, incluso mayor que la observada en el área urbana, y adquiere una relevancia particular en el contexto rural, donde la prevalencia estructural de la pobreza es considerablemente más alta (82.3 por ciento en la muestra de test).

Esta mejora en el AUC, el Gini y la especificidad no se traduce en una mejora generalizada del resto de las métricas. Tanto la exactitud (0.610) como el F1 (0.708) del modelo de stacking resultan los más bajos entre los cuatro modelos comparados en test, por debajo de los valores obtenidos por cualquiera de los tres modelos base de manera individual. Esto indica que, en el segmento rural, el metamodelo optimiza principalmente la capacidad de discriminación global y la especificidad, en lugar de maximizar el número total de aciertos o el balance entre precisión y sensibilidad (F1), un comportamiento distinto al observado en el segmento urbano, donde el modelo de stacking mantiene un F1 intermedio entre los tres modelos base en lugar de ubicarse por debajo de todos ellos. El modelo rural, pese a discriminar mejor en términos relativos, subestimaría el riesgo de pobreza futura. De acuerdo con esto se presenta la matriz de confusión:

Tabla 19: *Matriz de confusión modelo rural (test)*

		Predicho	
		No pobre (0)	***Pobre (1) modelo***
Real	***No pobre (0)***	*3052*	*881*
	Pobre (1)	*7794*	*10520*

En la Tabla 19 se observa un total de 22,247 observaciones en test, se identifican correctamente al 61% de los casos. A diferencia del segmento urbano, el número de falsos negativos es proporcionalmente mayor en el modelo de zona rural, donde 7,794 personas son clasificadas como no pobres cuando en realidad transitarán a la pobreza multidimensional, frente a apenas 881 falsos positivos, es decir, personas clasificadas como pobres que en realidad no lo serán. Esto indica que el modelo tendería a subestimar el riesgo de pobreza futura.

Expresado en términos relativos, el modelo deja de identificar correctamente al 42.6% de las personas que efectivamente se encuentran en situación de pobreza multidimensional futura, mientras clasifica de forma errónea a 22.4% de las personas como pobres, que en realidad no lo son. Esta asimetría en los errores de la estimación de pobreza son consecuencia del comportamiento natural de la pobreza en la zona rural y urbana. La primera se caracteriza por un 82.3% de casos positivos, al tratarse de una población donde la pobreza es mayoritaria. Por esto, el modelo tiende a ser conservador, favoreciendo la precisión (0.923) y la especificidad (0.776) por sobre la sensibilidad (0.574). Para identificar con mayor precisión las características de predicción y la separación de clases se realiza una tabla de ganancias. A continuación, los resultados:

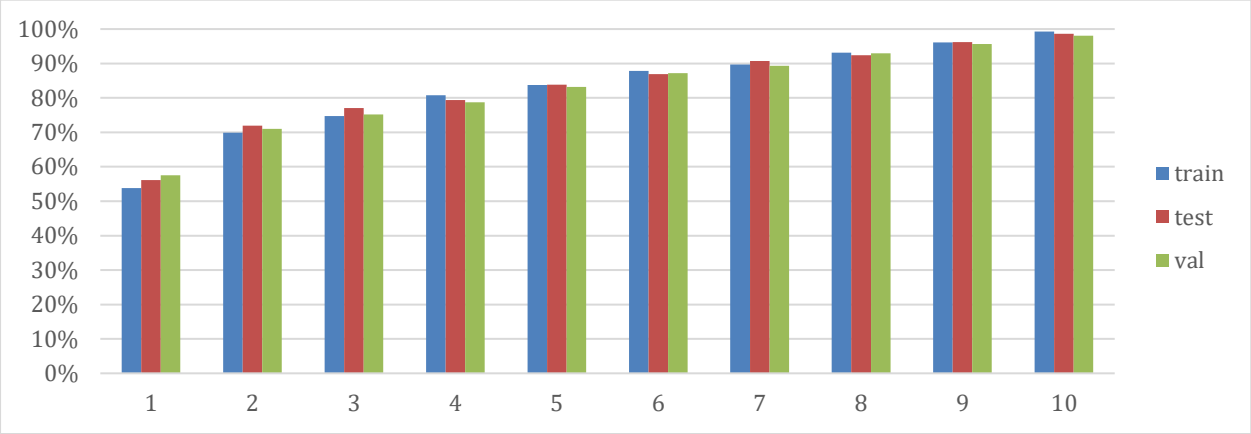
Tabla 20: *Tabla de ganancias del modelo rural*

Partición test	Decil	conteo	Malos	Buenos	Tasa malos	Odds	Malos acum	N acum	Tasa malos acum	Prob prom
	1	2708	1519	1189	56.10%	0.783	1519	2708	56.10%	21.50%
test	2	2351	1691	660	71.90%	0.39	3210	5059	63.50%	34.20%
test	3	2368	1825	543	77.10%	0.298	5035	7427	67.80%	43.40%
test	4	2105	1671	434	79.40%	0.26	6706	9532	70.40%	51.70%
test	5	2243	1881	362	83.90%	0.192	8587	11775	72.90%	58.60%

test	6	2183	1898	285	86.90%	0.15	10485	13958	75.10%	67.00%
test	7	1959	1777	182	90.70%	0.102	12262	15917	77.00%	73.70%
test	8	2249	2078	171	92.40%	0.082	14340	18166	78.90%	80.00%
test	9	2112	2032	80	96.20%	0.039	16372	20278	80.70%	86.50%
test	10	1969	1942	27	98.60%	0.014	18314	22247	82.30%	90.60%
train	1	9881	5317	4564	53.80%	0.858	5317	9881	53.80%	22.00%
train	2	9787	6837	2950	69.90%	0.431	12154	19668	61.80%	34.40%
train	3	9866	7374	2492	74.70%	0.338	19528	29534	66.10%	43.60%
train	4	9804	7921	1883	80.80%	0.238	27449	39338	69.80%	51.70%
train	5	9832	8238	1594	83.80%	0.193	35687	49170	72.60%	58.60%
train	6	9837	8637	1200	87.80%	0.139	44324	59007	75.10%	67.20%
train	7	9829	8819	1010	89.70%	0.115	53143	68836	77.20%	73.70%
train	8	9833	9155	678	93.10%	0.074	62298	78669	79.20%	79.90%
train	9	9833	9450	383	96.10%	0.041	71748	88502	81.10%	86.50%
train	10	9834	9760	74	99.20%	0.008	81508	98336	82.90%	90.60%
val	1	2400	1380	1020	57.50%	0.739	1380	2400	57.50%	21.50%
val	2	2092	1485	607	71.00%	0.409	2865	4492	63.80%	34.30%
val	3	2068	1555	513	75.20%	0.33	4420	6560	67.40%	43.30%
val	4	1875	1476	399	78.70%	0.27	5896	8435	69.90%	51.80%
val	5	1926	1603	323	83.20%	0.201	7499	10361	72.40%	58.80%
val	6	1942	1693	249	87.20%	0.147	9192	12303	74.70%	67.20%
val	7	1918	1714	204	89.40%	0.119	10906	14221	76.70%	73.80%
val	8	2029	1887	142	93.00%	0.075	12793	16250	78.70%	80.00%
val	9	1920	1837	83	95.70%	0.045	14630	18170	80.50%	86.40%
val	10	1967	1929	38	98.10%	0.02	16559	20137	82.20%	90.70%

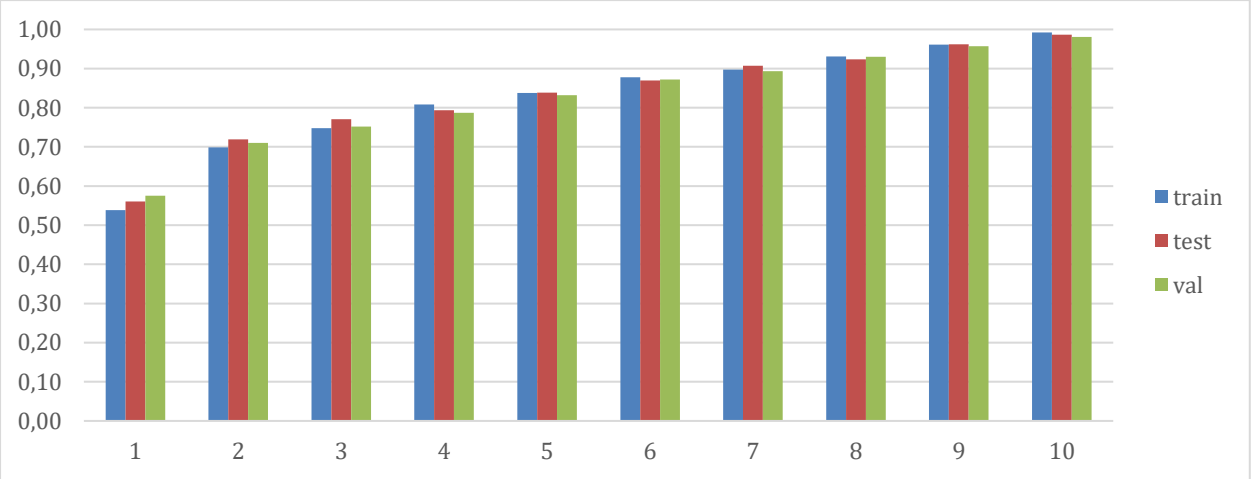
En la Tabla 20 se puede observar que, la tasa de malos (RP) también se aumenta constantemente a lo largo de los deciles, el primer decil muestra ya una diferencia significativa con los resultados de la zona urbana, con una tasa de malos de 56.1% en el primer decil hasta 98.6% en el décimo (test).

Figura 36: *Tasa de pobreza del sector rural (RP)*



Además, se observa que el KS alcanza un valor de 35.2 puntos porcentuales, siendo superior al de zona urbana, y ubicándose en el decil 4, con una tasa de malos acumulada del 36.6. Este resultado del modelo muestra que en la zona rural logra una separación adecuada entre pobres y no pobres, con estadísticos mejores que los del modelo urbano.

Figura 37: *Resultado de odds para zona rural.*



La Figura 37 presenta un ordenamiento adecuado, iniciando en 0.78 para el primer decil (por debajo del punto de equilibrio), es decir, desde el primer decil se presenta una mayor proporción de personas en estado de pobreza, y descendiendo hasta 0.01 en el décimo decil. Mostrando que en ningún decil se logra encontrar una proporción favorable a la no pobreza.

Capítulo 7. Discusiones y Conclusiones

7.1. Discusiones:

Esta investigación evidenció que los algoritmos de aprendizaje automático constituyen una alternativa robusta para la predicción de la pobreza multidimensional en el Ecuador, logrando desempeños satisfactorios tanto en la zona rural como en la urbana. Para ambas zonas, el modelo de stacking logró integrar la capacidad predictiva de RF, GBT y MLP, demostrando un rendimiento superior al de los modelos individuales y sustentando que la combinación de algoritmos permite capturar patrones que ninguno de los modelos base identifica de forma completa por separado.

En el modelo del área urbana se observa que la variable número de personas aporta con 7.8%, mientras que para el área rural aporta con 8.2%, con una diferencia de 0.4 puntos porcentuales entre ambas zonas. Este hallazgo tiene un respaldo teórico en el estudio de The Global MPI (2024), según el cual el bienestar de un hogar no depende solamente del ingreso, sino también de la capacidad de distribuir recursos entre el número de integrantes, dado que estos deben compartir alimentación, vivienda, educación, salud y tiempo de cuidado, entre otros factores de bienestar. En el caso ecuatoriano, Morales y Mideros (2021) documentan que los hogares con mayor número de integrantes presentan una relación más estrecha con la pobreza multidimensional, particularmente en el área rural, donde predominan las familias extensas y, con ellas, una mayor dependencia económica entre sus miembros. Los resultados de esta tesis confirmarían de forma directa este planteamiento, dado que el número de personas del hogar se posiciona como el predictor de mayor peso en ambos segmentos, lo que sugeriría que las presiones demográficas sobre los recursos del hogar incrementan la probabilidad de privaciones concurrentes. La diferencia de 0.4 puntos porcentuales en la importancia de esta variable entre zonas, aunque modesta en magnitud, sería consistente con la lectura de Morales y Mideros (2021), en el sentido de que las familias numerosas del área rural enfrentarían limitaciones adicionales, y no solamente equivalentes, para acceder a oportunidades de educación, empleo e infraestructura, en comparación con las familias numerosas del área urbana.

En el ámbito educativo, las variables asistencia a educación rezagada, máximo nivel educativo rezagado y razón de no asistencia concentran, en conjunto, el 18% de la importancia

del modelo urbano, mientras que en el modelo rural, las variables asistencia a educación rezagada, máximo nivel educativo rezagado y privación en educación superior concentran el 16%. Esta diferencia de cuatro puntos porcentuales, aunque no muy amplia en términos absolutos, resultaría coherente con el planteamiento de Mora (2019), según el cual la educación constituye un mecanismo de movilidad social considerablemente más fuerte en las ciudades, donde la escolaridad se asocia de forma más directa con oportunidades laborales, mayor formalidad en el empleo, ingresos estables y acceso a la seguridad social. En las zonas rurales, en cambio, aunque la educación conserva relevancia, la pobreza estaría más condicionada por limitaciones estructurales adicionales, como la infraestructura, el acceso a servicios básicos, el aislamiento geográfico y la actividad agrícola, factores que el conjunto de variables utilizado en este estudio no captura de forma directa. Los resultados de esta tesis, en este sentido, no solamente confirmarían la hipótesis de Mora (2019), sino que la cuantificarían: el poder predictivo del componente educativo disminuye, aunque de forma moderada, al pasar del contexto urbano al rural, lo que sería consistente con la coexistencia de determinantes estructurales adicionales en el ámbito rural que compiten con la educación como explicación de la pobreza multidimensional futura.

Por otro lado, la variable edad aporta 6% en el modelo urbano y 4% en el modelo rural, diferencia que permitiría identificar que esta variable afecta de manera transversal a ambas zonas, sin constituir un eje diferenciador territorial tan marcado como otras variables analizadas. Este comportamiento sería consistente con la literatura, que documenta que niños, jóvenes y adultos mayores presentan una mayor vulnerabilidad frente a la pobreza multidimensional, debido a su dependencia económica o a su menor capacidad de generación de ingresos (Mora, 2011). Un factor que sí resalta con fuerza en el modelo rural, y que no encuentra un correlato de magnitud comparable en el modelo urbano, es la etnia, que se posiciona como el segundo predictor más importante de dicho segmento. Este hallazgo sería consistente con lo señalado por la CEPAL (2022), organismo que documenta que los pueblos indígenas continúan experimentando mayores niveles de exclusión social, derivados de desigualdades históricas en el acceso a la educación, la salud, el empleo y la infraestructura básica. La contribución específica de esta tesis frente a dicha literatura residiría en haber cuantificado, mediante un modelo predictivo y no solamente descriptivo, la magnitud de esta brecha étnico-territorial, y en haber mostrado que dicha brecha es prácticamente inexistente en el área urbana, lo que sugeriría que los mecanismos de exclusión documentados por la CEPAL

(2022) operarían de forma particularmente intensa en el ámbito rural ecuatoriano, y no de manera homogénea a nivel nacional.

La diferencia en la prevalencia de pobreza multidimensional futura entre dominios, de 58.2% en el área urbana frente a 82.3% en el área rural, replicaría con notable cercanía la brecha estructural documentada oficialmente por el INEC (2025) para diciembre de dicho año, de 29.9% en el área urbana frente a 66.9% en el área rural. Esta correspondencia, más allá de confirmar un resultado ya expuesto en la sección de resultados, permitiría sostener que el desbalance de clases observado en la muestra de modelamiento no constituiría un artefacto propio del diseño muestral de esta investigación, sino un reflejo razonablemente fiel de la heterogeneidad territorial de la pobreza que documentan las fuentes oficiales, lo que otorgaría mayor validez externa a las diferencias de desempeño encontradas entre ambos modelos.

Precisamente en relación con este desbalance estructural, el contraste entre segmentos revela una tensión que la literatura sobre calibración de modelos de clasificación ha documentado de forma recurrente: el modelo rural exhibe una mejor discriminación relativa, expresada en un mayor AUC, Gini y estadístico KS, pero presenta al mismo tiempo una calibración notablemente peor que la del modelo urbano. Esta descalibración sistemática se atribuiría, principalmente, al esquema de balanceo de clases utilizado durante el entrenamiento, el cual, al ponderar las observaciones para compensar el desbalance entre pobres y no pobres, tiende a desplazar las probabilidades predichas hacia un punto cercano al 50%, con independencia de la prevalencia real de cada dominio. Este desplazamiento resultaría comparativamente inocuo en el área urbana, donde la prevalencia real (58.2%) ya se encuentra relativamente próxima a dicho punto medio, pero se volvería considerablemente más perjudicial en el área rural, donde la prevalencia real (82.3%) se aleja de forma sustancial del 50%, generando con ello la brecha de calibración de 23.3 puntos porcentuales identificada en el análisis de la tabla de odds de dicho segmento.

El umbral óptimo utilizado para la clasificación binaria de cada modelo se determinó mediante el índice de Youden (Youden, 1950), calculado sobre la partición de validación. Es importante señalar que este criterio maximiza la suma de la sensibilidad y la especificidad, pero no garantiza por sí solo una buena calibración de las probabilidades predichas, en la medida en que se trata de un criterio orientado al ordenamiento y la clasificación, y no a la exactitud de la escala de probabilidad. En escenarios donde se requiera estimar volúmenes absolutos de

pobreza, o asignar recursos con base en las probabilidades predichas y no únicamente en la clasificación binaria resultante, se recomendaría complementar el análisis con procedimientos de recalibración, como Platt scaling o regresión isotónica, con especial atención al segmento rural, dado que es en dicho segmento donde la brecha entre probabilidad predicha y frecuencia observada resulta más pronunciada.

En esta misma línea, el número de falsos negativos identificado en el área urbana (14,943 personas clasificadas como no pobres que en realidad transitarán a la pobreza multidimensional), superior al número de falsos positivos (7,332 personas clasificadas como pobres que en realidad no lo serán), no debería interpretarse como una simple constatación numérica, sino contrastarse con la literatura sobre los riesgos de exclusión de los mecanismos de focalización basados en modelos predictivos. Development Pathways (2017) documenta que los métodos de focalización de tipo Proxy Means Test tienden a sesgar sus estimaciones hacia la media de la distribución, subestimando el bienestar, o en este caso el riesgo, de los hogares en los extremos de la distribución, y generando con ello errores de exclusión que afectarían de forma desproporcionada a quienes más requerirían la intervención. El patrón de falsos negativos identificado en esta tesis, tanto en el área urbana como, de forma aún más marcada, en el área rural, sería consistente con esta preocupación documentada en la literatura internacional sobre focalización, y reforzaría la recomendación de no utilizar el modelo de forma automática para la exclusión de hogares de eventuales programas de protección social, sin una instancia adicional de revisión.

Finalmente, dado que tanto el estadístico KS como el índice de Gini miden, desde formulaciones distintas, la separación entre clases resulta coherente que ambos indicadores señalen a la zona intermedia de la distribución de riesgo como el punto de mayor capacidad de diferenciación del modelo entre pobres y no pobres, más que a los extremos de dicha distribución (Gini de 0.4138 y KS de 28.8 puntos porcentuales en el segmento urbano, ubicado este último en el decil 5). Este resultado sería plenamente coherente con la metodología de entrenamiento empleada, pues, al tratarse de un entrenamiento balanceado mediante pesos de clase, la probabilidad predicha tendería a concentrarse en torno a un punto de corte cercano a 0.5, extremo que se confirma en el valor del umbral óptimo obtenido mediante el índice de Youden para el modelo urbano (0.5943).

7.2. Conclusiones:

La construcción de la base de datos se realizó a partir de los microdatos de la Encuesta Nacional de Empleo, Desempleo y Subempleo del Instituto Nacional de Estadística y Censos, aplicando un proceso de depuración estructurado. La variable objetivo de pobreza multidimensional se calculó de forma binaria, siguiendo la metodología oficial del Instituto Nacional de Estadística y Censos. El resultado de este proceso es un conjunto de datos reproducible, sobre el cual se construyeron de manera consistente los modelos de predicción de pobreza multidimensional a un año.

El análisis exploratorio de los datos permitió identificar diferencias claras en la distribución de la pobreza multidimensional según el nivel educativo, la zona geográfica y las características sociodemográficas de los hogares. Dentro de la muestra utilizada para el modelamiento, brecha de pobreza multidimensional futura de 58.2% en el área urbana frente a 82.3% en el área rural. La exploración de las variables sociodemográficas mostró, además, que el tamaño del hogar y el nivel educativo máximo alcanzado se relacionan de manera sistemática con la condición de pobreza en ambos dominios, mientras que la etnia mostró una asociación considerablemente más marcada con la pobreza en el área rural que en la urbana, siendo la población indígena la que concentra el mayor riesgo relativo dentro de dicho segmento. Estos patrones, identificados desde la fase exploratoria, se confirmaron posteriormente de manera cuantitativa en las etapas de modelamiento y de explicabilidad del estudio.

El modelo urbano, entrenado sobre 225951 observaciones de entrenamiento, 54477 de validación y 60678 de prueba, alcanzó en la partición de prueba un área bajo la curva ROC de 0.7069, un índice de Gini de 0.4138 y un estadístico KS de 28.8%, mientras que el modelo rural, entrenado sobre 98336 observaciones de entrenamiento, 20137 de validación y 22247 de prueba, obtuvo un área bajo la curva ROC de 0.7431, un índice de Gini de 0.486 y un estadístico KS de 35.2% en el cuarto decil. Estas diferencias en el desempeño reflejan la prevalencia estructural de la pobreza documentada oficialmente por el Instituto Nacional de Estadística y Censos.

La matriz de confusión del modelo urbano presenta una exactitud de 63.29%, mientras que la matriz del modelo rural, una exactitud de 61.0%. Por otro lado, la comparación de las tablas de odds por decil, demuestra que en el área urbana se puede encontrar un punto de corte

para la identificación de pobreza multidimensional, ya que, en los dos primeros deciles de riesgo, la condición de no pobreza predomina sobre la de pobreza, mientras que en el área rural ningún decil llega a superar dicho punto de equilibrio. Es por ello que, a pesar de no tener una exactitud superior a 70%, los modelos muestran una gran capacidad de discriminación de la pobreza multidimensional.

Además, la composición interna de las variables de cada metamodelo resultó en modelos base de importancia opuesta para ambas zonas, dado que, la red neuronal concentró el mayor peso para geografía urbana con un coeficiente de 3.0319, mientras que, para rural destacó Random Forest con un coeficiente de 2.7138. Por lo que el desarrollo de modelos diferenciados no solamente cumplió el propósito metodológico, sino que resultó indispensable para visibilizar las dinámicas propias de cada estrato territorial.

El análisis permitió identificar que "número de personas del hogar", es la variable con mayor poder discriminante en ambos segmentos, con una importancia SHAP promedio de 0.0782 en el área urbana y de 0.0826 en el área rural. Entre los predictores de naturaleza educativa, "asistencia a educación doce meses atrás" se posicionó como la segunda variable más relevante en el área urbana, con una importancia de 0.0734, y como la tercera en el área rural, con una importancia de 0.0629, mientras que "nivel educativo máximo alcanzado en el hogar" mostró una importancia de 0.0647 en el área urbana frente a 0.0381 en el área rural, lo que indicaría que el nivel educativo discrimina con mayor fuerza la pobreza futura en el contexto urbano que en el rural.

Entre los predictores de naturaleza sociodemográfica, la variable "etnia" ejerce una influencia sustancialmente mayor sobre la probabilidad de pobreza futura en el área rural, con una importancia SHAP de 0.0645, que, en el área urbana, donde dicha importancia se reduce a 0.0067. El análisis gráfico de los valores SHAP permitió, adicionalmente, precisar que este efecto se concentra en la categoría indígena, hallazgo consistente con la evidencia documentada sobre la relación entre etnicidad y pobreza multidimensional en el Ecuador rural.

Los modelos desarrollados logran responder al problema de investigación planteado inicialmente, al predecir la pobreza multidimensional futura con un nivel de precisión adecuado. Además, mediante el uso complementario de las importancias de variables y de los

valores SHAP, se logró un grado satisfactorio de interpretabilidad, sin renunciar por ello a las diferencias estructurales que separan al ámbito rural del urbano en el Ecuador.

Anexos:

Anexo 1: Segmentación Geográfica y Experimentos por Estrato

```
>>> ESTRATO URBANO <<<

============================================================
EXPERIMENTO: Urbano_Solo_Educacion
Variables: 17 | Balance: class_fexp
Base: ['rf', 'mlp', 'gbt'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.7014 0.6459    0.6580
           mlp     Base 0.7030 0.6302    0.6284
           gbt     Base 0.7135 0.6532    0.6642
Stacking_logit Stacking 0.7132 0.6599    0.6644

============================================================
EXPERIMENTO: Urbano_Socioecon_Sin_IPM
Variables: 23 | Balance: class_fexp
Base: ['rf', 'mlp', 'gbt'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.7827 0.7161    0.7157
           mlp     Base 0.7830 0.7003    0.6982
           gbt     Base 0.8030 0.7292    0.7313
Stacking_logit Stacking 0.8023 0.7293    0.7293
```

```
>>> ESTRATO RURAL <<<

============================================================
EXPERIMENTO: Rural_Solo_Educacion
Variables: 17 | Balance: class_fexp
Base: ['rf', 'mlp', 'gbt'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.7325 0.6718    0.6277
           mlp     Base 0.7251 0.6650    0.6199
           gbt     Base 0.7421 0.7777    0.8255
Stacking_logit Stacking 0.7399 0.7102    0.6732

============================================================
EXPERIMENTO: Rural_Socioecon_Sin_IPM
Variables: 23 | Balance: class_fexp
Base: ['rf', 'mlp', 'gbt'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.8085 0.7515    0.7210
           mlp     Base 0.7978 0.6803    0.6375
           gbt     Base 0.8195 0.8276    0.8463
Stacking_logit Stacking 0.8168 0.7549    0.7250

================================================================================
COMPARATIVA POR ESTRATO Y ENFOQUE — solo resultados del Stacking
================================================================================
Estrato           Enfoque         modelo    auc     f1  accuracy
 Urbano    Solo Educacion Stacking_logit 0.7132 0.6599    0.6644
 Urbano Socioecon Sin IPM Stacking_logit 0.8023 0.7293    0.7293
  Rural    Solo Educacion Stacking_logit 0.7399 0.7102    0.6732
  Rural Socioecon Sin IPM Stacking_logit 0.8168 0.7549    0.7250
```

Anexo 2: Experimento Definitivo -Variables Depuradas por IV

```
>>> EXPERIMENTO DEFINITIVO: URBANO <<<

============================================================
EXPERIMENTO: Final_Urbano_Educacion_Limpio
Variables: 14 | Balance: class_fexp
Base: ['rf', 'mlp', 'gbt'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.6955 0.6434    0.6515
           mlp     Base 0.6999 0.6303    0.6285
           gbt     Base 0.7097 0.6523    0.6626
Stacking_logit Stacking 0.7095 0.6571    0.6608

>>> EXPERIMENTO DEFINITIVO: RURAL <<<

============================================================
EXPERIMENTO: Final_Rural_Educacion_Limpio
Variables: 14 | Balance: class_fexp
Base: ['rf', 'mlp', 'gbt'] | Meta: logit
============================================================

  Entrenando RF...
  Entrenando MLP...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

======================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
======================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.7310 0.6789    0.6358
           mlp     Base 0.7270 0.6630    0.6177
           gbt     Base 0.7431 0.7780    0.8263
Stacking_logit Stacking 0.7367 0.7118    0.6754

============================================================
RESUMEN DEFINITIVO — Resultados para la tesis
============================================================
  URBANO: AUC=0.7095 | F1=0.6571 | Acc=0.6608
  RURAL: AUC=0.7367 | F1=0.7118 | Acc=0.6754
```

Anexo 3: Experimento con Hiperparámetros Óptimos (Optuna)

```
•••  Ejecutando experimentos definitivos con hiperparametros Optuna...

  Urbano — Final_Urbano_Optuna:
    AUC-ROC : 0.7096
    F1      : 0.6571
  Rural — Final_Rural_Optuna:
    AUC-ROC : 0.7412
    F1      : 0.7154

============================================================
COMPARACION: Hiperparametros por defecto vs Optuna
============================================================
  Estrato    Modelo                          AUC
  ------------------------------------------
  Urbano     Default (Sin Optuna)       0.7095
  Rural      Default (Sin Optuna)       0.7367
  Urbano     Optuna (Optimo)            0.7096
  Rural      Optuna (Optimo)            0.7412

Si Optuna mejora > 0.005 AUC: la optimizacion es significativa.
Si la mejora es menor: los hiperparametros por defecto eran razonables.
```

Anexo 4: Prueba de Hipótesis - Número Óptimo de Modelos Base

```
>>> PRUEBA DE HIPOTESIS: ARQUITECTURA OPTIMA DEL STACKING <<<

================================================================
EXPERIMENTO: Control_2_Modelos
Variables: 14 | Balance: class_fexp
Base: ['rf', 'gbt'] | Meta: logit
================================================================

  Entrenando RF...
  Entrenando GBT...
  Entrenando Meta-learner (logit)...

================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.6955 0.6434    0.6515
           gbt     Base 0.7097 0.6523    0.6626
Stacking_logit Stacking 0.7096 0.6567    0.6614

================================================================
EXPERIMENTO: Control_3_Modelos
Variables: 14 | Balance: class_fexp
Base: ['rf', 'gbt', 'mlp'] | Meta: logit
================================================================

  Entrenando RF...
  Entrenando GBT...
  Entrenando MLP...
  Entrenando Meta-learner (logit)...

================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.6955 0.6434    0.6515
           gbt     Base 0.7096 0.6523    0.6626
           mlp     Base 0.6999 0.6303    0.6285
Stacking_logit Stacking 0.7095 0.6571    0.6608
```

```
================================================================
EXPERIMENTO: Control_4_Modelos
Variables: 14 | Balance: class_fexp
Base: ['rf', 'gbt', 'mlp', 'logit'] | Meta: logit
================================================================

  Entrenando RF...
  Entrenando GBT...
  Entrenando MLP...
  Entrenando LOGIT...
  Entrenando Meta-learner (logit)...

================================================================
RESULTADOS (evaluados en TEST — datos no vistos durante entrenamiento)
================================================================
        modelo     tipo    auc     f1  accuracy
            rf     Base 0.6955 0.6434    0.6515
           gbt     Base 0.7097 0.6523    0.6626
           mlp     Base 0.6999 0.6303    0.6285
         logit     Base 0.6825 0.6361    0.6392
Stacking_logit Stacking 0.7095 0.6574    0.6603

================================================================================
EVIDENCIA: RENDIMIENTO POR COMPLEJIDAD DEL ENSAMBLE
================================================================================
                          Arquitectura         modelo    auc     f1  accuracy
    2 Modelos (RF + GBT) — homogeneos Stacking_logit 0.7096 0.6567    0.6614
3 Modelos (RF + GBT + MLP) — diversos Stacking_logit 0.7095 0.6571    0.6608
   4 Modelos (RF + GBT + MLP + Logit) Stacking_logit 0.7095 0.6574    0.6603

INTERPRETACION:
  2→3 modelos: ganancia esperada por diversidad algorítmica (árboles+neuronal)
  3→4 modelos: rendimiento decreciente si Logit agrega poca informacion nueva
```

Bibliografía

Corral, P., et al. (2025). [Referencia citada de forma secundaria a través de Gonzales Martinez y Cooray (2025) — ver nota al final].

Development Pathways. (2017). Exclusion by design: An assessment of the effectiveness of the proxy means test poverty targeting mechanism. https://www.developmentpathways.co.uk/wp-content/uploads/2017/03/Exclusion-by-design-An-assessment-of-the-effectiveness-of-the-proxy-means-test-poverty-targeting-mechanism-1-1.pdf

Gonzales Martinez, R., & Cooray, M. (2025). Enhancing poverty targeting with spatial machine learning: An application to Indonesia (arXiv:2503.04300). arXiv. https://arxiv.org/html/2503.04300v1

Instituto Nacional de Estadística y Censos. (2025). Encuesta Nacional de Empleo, Desempleo y Subempleo (ENEMDU): Indicadores de pobreza y desigualdad, diciembre 2025. INEC. https://www.ecuadorencifras.gob.ec/documentos/web-inec/POBREZA/2025/Diciembre/202512_PobrezayDesigualdad.pdf

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. En I. Guyon et al. (Eds.), Advances in Neural Information Processing Systems 30 (pp. 4765-4774). Curran Associates. https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

Mideros Mora, A. (2011). Defining and measuring multidimensional poverty. Exploring poverty in Ecuador 2006-2010 [Documento de trabajo]. Maastricht Graduate School of Governance, UNU-MERIT, Maastricht University.

Mideros Mora, A. (2019, 5 de septiembre). Pobreza multidimensional. Primicias. https://www.primicias.ec/noticias/firmas/analisis-pobreza-multidimensional/

Morales, M., & Mideros, A. (2021). Análisis de la pobreza multidimensional en los hogares de la agricultura familiar campesina en el Ecuador, 2009-2019. Revista Economía, 73(118), 7-21. https://doi.org/10.29166/economa.v73i118.3379

Oxford Poverty and Human Development Initiative (OPHI), & Programa de las Naciones Unidas para el Desarrollo (PNUD). (2024). Global Multidimensional Poverty Index 2024: Poverty amid conflict. Programa de las Naciones Unidas para el Desarrollo y Oxford Poverty and Human Development Initiative. https://ophi.org.uk/global-mpi/2024

Turbau, I., & Arroyo, D. (2025, 6 de mayo). Abrir la caja negra de la inteligencia artificial: Qué es la explicabilidad o IA explicable y por qué es importante para construir modelos más transparentes. Maldita.es. https://maldita.es/malditatecnologia/20250506/ia-caja-negra-explicabilidad-explicable/

Wobcke, W., & Mariyah, S. (2023). Machine learning and data augmentation in the proxy means test for poverty targeting. Statistical Journal of the IAOS. https://doi.org/10.3233/SJI-230033

Yuan, H., Kang, Y., Feng, M., Ang, Y., Duan, X., Xie, F., Ning, Y., Li, Y., Ong, M. E. H., & Liu, N. (2022). An empirical study of the effect of background data size on the stability of SHapley Additive exPlanations (SHAP) for deep learning models (arXiv:2204.11351). arXiv. https://arxiv.org/abs/2204.11351

Youden, W. J. (1950). Index for rating diagnostic tests. Cancer, 3(1), 32-35. https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631. https://doi.org/10.1145/3292500.3330701

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics*, *95*(7–8), 476–487. https://doi.org/10.1016/j.jpubeco.2010.11.006

Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. National Bureau of Economic Research.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in Neural Information Processing Systems*, *24*, 2546–2554.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Castillo Añazco, R., & Jácome Pérez, F. (2017). *Medición de la pobreza multidimensional en Ecuador*. Instituto Nacional de Estadística y Censos. https://www.ecuadorencifras.gob.ec/documentos/web-inec/POBREZA/2017/Pobreza_Multidimensional/ipm-metodologia-oficial.pdf

CEPAL, & PNUD. (2025). *Índice de pobreza multidimensional para América Latina. Resumen*. Comisión Económica para América Latina y el Caribe. https://hdl.handle.net/11362/81426

Fhaldian, W., & Fahmi, A. (2025). Explainable machine learning for poverty prediction in Central Java regencies and cities. *Sinkron: Jurnal dan Penelitian Teknik Informatika*, *9*(4). https://doi.org/10.33395/sinkron.v9i4.15312

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2.ª ed.). Springer.

López Idrovo, D., & Sarmiento Moscoso, L. S. (2024). Las determinantes de la pobreza en Ecuador, aplicación de un modelo LOGIT. Un estudio Regional para los años 2007 y 2021. *UDA AKADEM*, (13), 168–198. https://doi.org/10.33324/udaakadem.vi13.757

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Morales Torres, I. F. (2025). Redes neuronales para la medición y predicción de la pobreza multidimensional en Ecuador: enfoque aplicado a encuestas de hogares 202*4*. *Nexus Research Journal*, 4(2), 297–318. https://doi.org/10.62943/nrj.v4n2.2025.414

Morán Chiquito, D., & Lozano, C. (2017). Condicionantes de la Pobreza Rural en el Ecuador 2007-2014: Una estimación de modelos Probit. *REICE: Revista Electrónica de Investigación en Ciencias Económicas, 5*(10), 38–53. https://doi.org/10.5377/reice.v5i10.5529

Ochoa Guaraca, M. E., Castro, R., Arias Pallaroso, A., Machado, A., & Sucozhañay, D. (2021). Machine learning approach for multidimensional poverty estimation. *Revista Tecnológica ESPOL*, *33*(2), 205–225. https://doi.org/10.37815/rte.v33n2.853

Rojas, H., Alvarez, C., & Rojas, N. (2023). *Statistical hypothesis testing for information value (IV)* (arXiv:2309.13183). arXiv. https://arxiv.org/abs/2309.13183

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Salvador Benítez, L. (2008). Desarrollo, educación y pobreza en México. *Papeles de Población, 14*(55), 237–257.

Sen, A. (2000). *Development as freedom*. Anchor Books.

Sende, N. B., Saha, S., Ruganzu, L., & Kar, S. (2025). Prediction of multidimensional poverty status with machine learning classification at household level: Empirical evidence from Tanzania. *IEEE Access*, *13*, 23461–23471. https://doi.org/10.1109/ACCESS.2025.3537807

Shapley, L. S. (1953). A value for n-person games. En H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games* (Vol. 2, pp. 307–317). Princeton University Press.

Uquillas, A., & Cruz, S. (2025). Hybrid neural network with image caption generator for spatio-temporal income inequality and poverty estimation: A case study in Ecuador. *Review of Regional Research*, *45*(4), 573–607. https://doi.org/10.1007/s10037-025-00241-3

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1