Predicción de la pobreza de los hogares en Colombia para el año 2018 usando aprendizaje automático

Ana Rojas, Juan Rincón, Julián Delgado y Mario Mercado

Universidad de los Andes
Facultad de Economía
Maestría en Economía Aplicada
Bogotá D.C, Colombia
2025

1. Introducción

En el presente documento se tiene como objetivo explorar técnicas de aprendizaje automático para predecir la pobreza en hogares colombianos, utilizando microdatos de la Gran Encuesta Integrada de Hogares (GEIH) del 2018. Para ello, se implementan modelos supervisados, como Random Forest y regresiones penalizadas. Además, se evalúa la capacidad predictiva de los diferentes enfoques y se identifican los factores más determinantes asociados con la condición de pobreza[1].

1.1.1. Contexto

La pobreza es uno de los mayores desafíos de política social que enfrentan los países a nivel mundial. Según datos del World Bank Group (2024), a nivel mundial el 8.5% (o 700 millones de personas) viven con menos de 2.5 dólares por persona al día; para América Latina y el Caribe, la cifra de pobreza y pobreza extrema para el año 2023 fue de 10.6% y 27.3% respectivamente, lo que se traduce en 66 millones de personas (en condición de pobreza monetaria) y 172 millones de personas (en condición de pobreza monetaria extrema) (CEPAL, 2024). Para Colombia, las cifras de pobreza monetaria y pobreza extrema para el año 2023 fueron de 33,0% y 11,4%, respectivamente (DANE, 2024).

1.1.2. Antecedentes

La literatura teórica y empírica sobre los estudios de pobreza monetaria y multidimensional abunda en cientos de artículos escritos por cientos de académicos expertos en la temática (identificación, medición y evaluación de la pobreza, sea directa o indirecta); se destacan autores como (Alkire & Foster, 2011; Banerjee & Duflo, 2012; Ravallion, 1994; Ray, 2002; Sen, 2000).

El economista Amartya Sen (2000) una de las mayores eminencias en temas de medición del bienestar y la pobreza, identifica dos formas de pobreza una indirecta que hace referencia a la pobreza monetaria[2] o por ingresos, esta esencialmente se basa en una línea de pobreza, la CEPAL (2018) dice que " La línea de pobreza representa un valor monetario en que se consideran dos componentes: el costo de adquirir una canasta básica de alimentos y el costo de los demás bienes y servicios, expresado sobre la base de la relación entre el gasto total y el gasto en alimentos", la otra forma es la directa que se refiere a la pobreza multidimensional[3] y esta es más amplia y tiene en cuenta muchas más dimensiones que el

[1] Para más información véase el repositorio alojado en GitHub https://github.com/Judelgadogu/Taller-2-BIG-DATA-Public

[2] Véase Gasparini et al. (2013) un compendio completo sobre la pobreza y desigualdad tanto desde un punto de vista conceptual, teórico y empírico aplicado en América Latina

[3] Lora y Prada (2023) en el capítulo 4 de su destacada obra, analiza en términos teóricos las diferentes mediciones del desarrollo humano, que va desde la pobreza monetaria hasta indicadores de desigualdad.

ingreso o consumo de los hogares, como por ejemplo la estructura del hogar en términos de condiciones de vida (habitabilidad, acceso a servicios esenciales como el agua, o alcantarillado, etc.) (Alkire & Foster, 2011).

Existen muchos estudios a nivel econométrico, como los modelos de *variable dependiente limitada* (Probit y Logit) de determinantes de pobreza, que para el caso colombiano se destacan estudios como los de Núñez y Ramírez (2002), quienes estimaron un modelo de tipo Probit, teniendo en cuenta variables socioeconómicas de los hogares colombianos para los años 1991, 1995 y 2000. Dentro de sus hallazgos encontraron que la variable número de personas en el hogar incrementa el riesgo de pobreza para los diferentes años y para el caso del año adicional promedio del hogar, disminuye la probabilidad de ser pobre. Recientemente, Ariza y Retajac (2020) estimaron un modelo Logit para la pobreza urbana en Colombia y encontraron que, para el jefe de hogar "cuentapropista", el riesgo de pobreza se incrementa entre un 4% y un 10%; para la variable de ingresos del hogar (recepción de remesas), la probabilidad de ser considerado pobre disminuye entre un 6% y un 11%.

La revolución del análisis de datos avanzado (inteligencia artificial, big data y el aprendizaje automático o estadístico) ha puesto de manifiesto la discusión de las nuevas formas de identificación y medición de la pobreza a nivel mundial. Dentro de las investigaciones más interesantes, resaltan las de Blumenstock et al. (2015), Steele et al. (2017) y Chi et al. (2022).

Para el caso de Blumenstock et al. (2015), analizaron datos anonimizados de teléfonos celulares (encuestas de hogares) de Ruanda para predecir la pobreza y la riqueza entrenando un algoritmo de aprendizaje automático. Uno de los mayores expertos en temas sociales (pobreza y desigualdad) y de big data-aprendizaje automático (por sus siglas en inglés machine learning) en la región de América Latina Walter Sosa Escudero ha escrito investigaciones sobre el uso de esta potente herramienta para el entendimiento de la pobreza desde otro enfoque diferente al convencional (econométrico), se destacan estudios como Sosa-Escudero (2018) donde explica las diferentes herramientas del big data y el machine learning para economistas y cómo usarlas para el entendimiento de problemas prácticos como la pobreza, así mismo Sosa-Escudero et al. (2022) estudian a profundidad la aplicabilidad del aprendizaje automático a los estudios de la pobreza, desigualdad y estudios de desarrollo desde un enfoque teórico y práctico y el cómo el usos de estas novedosas y avanzadas técnicas pueden a ayudar al investigador y hacedor de políticas públicas a entender estos temas sociales de gran relevancia para la región de América Latina.

2. Datos

2.1.1. ETL[4] de los microdatos

Los datos anonimizados de la GEIH son datos que recaba el Departamento Administrativo Nacional de Estadísticas (DANE) para conocer las características más relevantes de las personas y hogares en Colombia en temas de empleo, desempleo, vivienda, características generales, entre otros; estos son usados para producir más de una veintena de indicadores sociales, económicos, y son insumo de primera mano para los análisis de la pobreza monetaria, pobreza extrema y desigualdad.

Para la construcción de las bases de datos a utilizar para la estimación de los distintos modelos de clasificación enmarcados en el *aprendizaje supervisado*, como primera medida se importaron al software R los 4 archivos[5] (muestra de entrenamiento y evaluación para personas y hogares). Como segunda medida, se creó una carpeta nombrada *store* donde se alojaron los archivos en formato estándar de R (rds), para entrenar los distintos modelos algorítmicos para predecir la pobreza (tanto por la línea de pobreza como por clasificación ceros-unos). Estos datos de tipo microeconómico con un enfoque en el aprendizaje automático (métricas de evaluación training set y test set) son esenciales para el estudio de los problemas o fenómenos sociales como la pobreza y desigualdad, ya que al ser de carácter individual o agrupado por hogares tienen la esencia de clasificar el resultado (outcome) y en lo posible tienden a minimizar el error de predicción.

El poder que tiene la librería *tidyverse* para la manipulación de datos fue muy importante para el preprocesamiento, esta permitió la limpieza de los datos en términos generales de la base de datos, esto es revisando los valores faltantes y los valores atípicos en las variables socioeconómicas y del hogar de interés[6], renglón seguido se seleccionaron las variables más relevantes (estas en su mayoría de carácter dicotómica referentes al jefe de hogar) tanto desde el punto de vista teórico como empírico para el entrenamiento de los distintos modelos algorítmicos a entrenar y evaluar, después de aplicar la misma lógica para las bases de datos de training set tanto de personas como hogares, se decide hacer la unión o *merge* de las bases de datos para tener una unificación de la base de datos final, por medio del identificador de personas *id* obteniendo así la base de datos test_hogares_full, este mismo ejercicio de realizo para las bases test set[7].

---

[4] En el análisis de datos se refiera a la extracción, transformación y cargue de los microdatos.
[5] Estos archivos se descargaron desde el proyecto de competencia en la plataforma *Kaggle.*
[6] Tomando como referencia las variables incluidas en la documentación oficial del DANE sobre la medición de pobreza monetaria y desigualdad.
[7] Esta presentó la dificultad que no contaba con el *ingreso per cápita de la unidad de gasto*, por lo que se decide aproximar esa variable con la información de ingreso desagregados.

2.1.2. Estadísticas descriptivas

De acuerdo con la tabla 1, se presentan las estadísticas descriptivas de las variables a utilizar en los distintos modelos de clasificación. Se observa que las variables continuas presentan una buena distribución de los datos, mostrando señales de no presentar valores faltantes (missing value) y valores atípicos (outliers); para las discretas, que son en su mayoría dicotómicas, están entre los parámetros estadísticos y no se presentan signos de anomalías en los datos, por lo que son de gran utilidad para el entrenamiento de los modelos de clasificación. Si se observa la figura 2, esta da cuenta de lo importante que es para la clasificación de los distintos modelos el atributo de número de hijos, que tiene una alta explicación de clasificar a un hogar como pobre en Colombia. A medida que en un hogar haya presencia de un mayor número de hijos, este hogar tiende a clasificarse como pobre. Lo mismo sucede en la figura 3 con la composición del hogar esta al presentar un mayor número de miembros dentro del hogar, la clasificación de considerar al hogar pobre tiende a ser uno. Para el caso del ingreso per cápita de la unidad de gasto este en la figura 4 para el caso de los hogares clasificados como pobres tiende a estar distribuidos con una media de $100.000 en términos logarítmicos, para los no pobres de $1.000.000, esto evidenciando la disparidad en el ingreso de los hogares colombianos, siendo así una herramienta importante para la medición de la pobreza por ingresos.

3. Modelos y resultados

3.1.1. Selección del modelo y entrenamiento

Dado que la pobreza es un fenómeno multidimensional e influenciado por interacciones complejas entre variables económicas, sociales y demográficas, se espera que un modelo no lineal como Random Forest ofrezca un mejor desempeño predictivo. A diferencia de los modelos lineales, que asumen relaciones constantes entre cada variable y el resultado, Random Forest permite capturar interacciones no evidentes y efectos no lineales entre las características del hogar y su situación de pobreza. Además, su capacidad para manejar variables correlacionadas, valores atípicos, un amplio número de valores faltantes y distintas escalas de medición sin requerir preprocesamientos intensivos lo hace especialmente útil en contextos como este, donde los determinantes de la pobreza pueden variar significativamente entre hogares. No obstante, se decidió correr cuatro modelos lineales con el fin de comparar resultados y respaldar la elección con evidencia empírica.

Al revisar las Tablas 3 y 4, se identifica que el modelo con mejor predicción y, por ende, el modelo seleccionado es Random Forest. El proceso de entrenamiento y de ajuste de hiperparámetros se realizó en dos etapas, comenzando con un modelo simple de Random Forest que permita entender el comportamiento del modelo (de manera rápida y eficiente) y así expandir el grid a combinaciones más amplias una vez identificado su potencial.

La primera etapa de la construcción del modelo es la selección de variables (a partir de la teoría y un análisis preliminar) para hacer una primera prueba de Random Forest con pocas iteraciones (5) y un grid de parámetros reducido. Fue necesario corregir desequilibrios en la variable objetivo (Pobre), dado que existían más "No_Pobres" que "Pobres". Teniendo en cuenta que el conjunto de datos es grande y de alta dimensión, se usó ranger, una implementación eficiente y rápida del algoritmo Random Forest. Por último, como se observa en la Tabla 4, se evaluó la importancia de las variables de forma inicial, lo que permitió identificar cuáles aportaban más al modelo. De esta forma logramos encontrar hiperparámetros más eficientes y las variables que mejor ayudaban a la predicción.

En la Figura 5 se muestra la importancia relativa de las variables predictoras en el modelo de Random Forest, medida a través de la disminución en impureza. De esta forma, las cinco principales variables del modelo son: *JH_RSS_S, Hijos, Subsidios, Nper y Educ_prom. JH_RSS_S* es la variable más influyente y está relacionada con el acceso a subsidios en EPS. Su alta importancia sugiere una relación fuerte con la condición de pobreza. *"Hijos"* hace referencia al número de hijos en el hogar y es una de las variables más relevantes, lo cual podría asociarse con mayores cargas económicas en los hogares. *Subsidios* es la tercera variable más relevante en el modelo. El hecho de que los subsidios sean altamente predictivos podría indicar que los hogares que los reciben tienen mayor probabilidad de estar clasificados como pobres, aunque también podría reflejar una política pública bien focalizada. *Nper* (número de personas en el hogar) y *Educ_prom* (escolaridad promedio del hogar) también tienen un rol importante. El tamaño del hogar y el capital educativo promedio parecen estar estrechamente relacionados con las condiciones económicas.

De esta forma, la prueba inicial con el modelo Random Forest revela que las variables más relevantes para predecir la pobreza están relacionadas con: protección social, estructura del hogar y el nivel educativo. Estas variables capturan tanto el acceso a recursos como las condiciones estructurales del hogar, lo cual es coherente con la literatura y con la intuición económica. A pesar de que *AyudasEco, JH_Personas_Trabajo, JH_NoSeguro, P5010*, entre otras, se encuentran en la parte inferior del gráfico, existe una relación con la variable objetivo del modelo.

La segunda etapa fue la optimización del modelo, en donde en primera instancia se seleccionaron las variables más importantes de la primera prueba para simplificar el modelo y reducir ruido. Adicionalmente, se amplió el grid de hiperparámetros y se aumentó ntree a 300 para garantizar una mayor estabilidad y mejor predicción de los resultados.

La Figura 7 de la Curva Característica Operativa del Receptor (ROC), con un Área Bajo la Curva (AUC) de 0.92 del modelo Random Forest, revela un excelente desempeño en la predicción de pobreza en hogares colombianos para el año 2018. La curva ROC, que traza la sensibilidad (tasa de verdaderos positivos) frente a la especificidad (1 - tasa de falsos positivos) en diversos umbrales de clasificación, se aproxima a la esquina superior izquierda

del gráfico. Esto quiere decir que el modelo logra identificar correctamente a la mayoría de los hogares en situación de pobreza, sin confundirlos con los que no lo están. A diferencia de la línea diagonal que representa un clasificador aleatorio, la curva ROC se desvía considerablemente de esa referencia, lo que resalta su alta capacidad para distinguir entre hogares pobres y no pobres.

El valor del AUC de 0.92 proporciona una métrica concisa de esta capacidad discriminatoria. Esto significa que, si se eligen al azar un hogar pobre y uno no pobre del conjunto de datos, hay un 92% de probabilidad de que el modelo Random Forest le asigne un mayor riesgo de pobreza al hogar que realmente lo es. Dentro de la escala de interpretación del AUC, un valor superior a 0.9 se considera indicativo de un rendimiento sobresaliente, lo que sugiere que el modelo en cuestión es altamente efectivo para distinguir entre los hogares pobres y no pobres en el contexto colombiano del año 2018.

3.1.2. Optimización de hiperparametros

Para los modelos penalizados (Ridge, Lasso y Elastic Net), se utilizó la función train del paquete caret en R, aplicando validación cruzada de 10 pliegues y búsqueda en rejilla (grid search) para identificar los hiperparámetros óptimos. En el caso de la regresión logística (Logit), no se requería regularización, por lo tanto, no se ajustaron hiperparámetros.

Se evaluaron rangos amplios para los hiperparámetros lambda (fuerza de penalización) y alpha (tipo de penalización en Elastic Net), incluyendo valores altos. Sin embargo, los mejores resultados se obtuvieron con valores bajos de lambda (0.12) y un alpha cercano a 1 (0.99) en el caso de Elastic Net, lo que sugiere que una penalización moderada fue suficiente para controlar el sobreajuste sin eliminar información relevante.

- Ridge: alpha = 0, lambda = 0.12. La penalización tipo L2 redujo la varianza, pero afectó la capacidad de generalización (F1 Kaggle: 0.5761).
- Lasso: alpha = 1, lambda = 0.12. Aplicó una penalización más agresiva, eliminando algunas variables con poco peso. Obtuvo mejor desempeño en Kaggle (F1: 0.6191).
- Elastic Net: alpha = 0.99, lambda = 0.12. Combinó ambas penalizaciones, privilegiando L1. Igualó el desempeño del modelo Lasso y Logit en Kaggle (F1: 0.6191).

Los tres modelos penalizados mostraron métricas internas similares, pero los que incluyeron penalización L1 (Lasso y Elastic Net) generalizaron mejor al conjunto de prueba externo.

Para optimizar los hiperparámetros del modelo Random Forest, se empezó abordando el desafío del desbalance de clases presente en los datos de pobreza en Colombia, donde los hogares no pobres suelen ser significativamente más numerosos que los hogares pobres. Se implementó la técnica de sobremuestreo SMOTE (Synthetic Minority Over-sampling

Technique) en lugar de simplemente ajustar los pesos de las clases. SMOTE crea nuevas instancias sintéticas (ejemplos artificiales) de la clase minoritaria (hogares pobres) interpolando entre los ejemplos existentes de esa clase. Esta estrategia tiene la ventaja de aumentar explícitamente la representación de la clase minoritaria en el conjunto de entrenamiento, proporcionando al algoritmo más ejemplos concretos sobre los cuales aprender los patrones característicos de la pobreza. Al generar datos nuevos pero realistas, SMOTE mejora la capacidad del modelo para identificar casos de pobreza sin caer en el riesgo de sobreajustar por repetir información.

Por el contrario, el ajuste de pesos de clase asigna una mayor importancia a los errores de clasificación de la clase minoritaria durante el entrenamiento del modelo. Si bien este método puede influir en el algoritmo para que preste más atención a la clase minoritaria, no introduce nueva información al conjunto de datos. El modelo sigue aprendiendo de la misma cantidad de ejemplos originales, aunque penaliza fuertemente los errores en la predicción de la clase minoritaria. Por lo tanto, SMOTE se selecciona, pues busca una mejora en la capacidad del modelo para generalizar e identificar correctamente los casos de pobreza, al ampliar el conjunto de entrenamiento con representaciones sintéticas pero informativas de la clase minoritaria.

Para optimizar el rendimiento del modelo Random Forest, se llevó a cabo un proceso de ajuste de hiperparámetros mediante validación cruzada con 5 folds, utilizando la función train del paquete caret en R. Se exploraron dos configuraciones del modelo: una completa, diseñada para robustez y captura de patrones complejos (con ntree en 300, mtry probando valores de 2 a 20, splitrule considerando Gini y ExtraTrees, y min.node.size entre 1 y 20), y una más rápida para evaluaciones iniciales (con ntree en 5, mtry en 4 y 8, splitrule solo Gini, y min.node.size en 1 y 10). Se empleó una búsqueda en cuadrícula para evaluar sistemáticamente todas las combinaciones de estos hiperparámetros, midiendo su impacto en las métricas de rendimiento del modelo durante la validación cruzada.

Se prestó especial atención a métricas sensibles al desbalance de clases, como el F1-Score y el AUC, además de la precisión, sensibilidad, especificidad y precisión. Tras la búsqueda en cuadrícula, los hiperparámetros óptimos se seleccionaron basándose en el mejor equilibrio entre el AUC y el F1-Score, priorizando la capacidad del modelo para discriminar entre hogares pobres y no pobres y su rendimiento general en la identificación de la clase minoritaria. Los parámetros finales elegidos fueron mtry = 4, splitrule = gini y min.node.size = 20. Este proceso de ajuste, en conjunto con la aplicación de SMOTE para mitigar el desbalance de clases, resultó en una mejora significativa en la capacidad predictiva del modelo Random Forest para la identificación de la pobreza en el contexto colombiano. La rigurosa validación cruzada con 5 folds aseguró la selección de hiperparámetros robustos y generalizables, evidenciando un sólido rendimiento en el AUC, una métrica clave para evaluar modelos de clasificación en escenarios con clases desbalanceadas.

3.1.3. Análisis comparativo

Al comparar el desempeño de los modelos implementados por el equipo (Logit, Ridge, Lasso y Elastic Net), se observaron diferencias importantes tanto en las métricas internas como en los resultados obtenidos en la competencia de Kaggle.

En términos generales, los modelos Lasso y Elastic Net obtuvieron los mejores resultados, tanto en las métricas internas (F1-Score ≈ 0.919) como en Kaggle (F1-Score = 0.6191), igualando el rendimiento del modelo Logit. Por otro lado, el modelo Ridge tuvo el rendimiento más bajo en Kaggle (F1-Score = 0.5761), a pesar de tener una sensibilidad ligeramente mayor en la validación interna (0.9574). Esto sugiere que el modelo Ridge probablemente presentó cierto sobreajuste al conjunto de entrenamiento.

La regresión logística simple (Logit) logró un rendimiento competitivo, lo cual indica que los atributos o variables socioeconómicas logran predecir de forma efectiva la variable pobre sin incluir regularización. Sin embargo, los modelos penalizados ofrecieron una ventaja al reducir el riesgo de sobreajuste, especialmente en presencia de variables correlacionadas.

En cuanto a la precisión y especificidad, todos los modelos penalizados mostraron una leve caída en comparación con Logit, aunque se mantuvieron en niveles aceptables. Esto puede deberse a que la penalización tiende a favorecer la sensibilidad a costa de una menor especificidad, priorizando la identificación de casos positivos (pobres) frente a los negativos (no pobres).

Al realizar una comparación entre el modelo final de Random Forest y los distintos modelos lineales (Logit, Ridge, Lasso y Elastic Net), considerando múltiples métricas de desempeño. El modelo Random Forest obtuvo un valor de accuracy de 0.8724, ligeramente superior al de los modelos lineales, cuyo valor osciló entre 0.860 y 0.866. Esto indica un mejor rendimiento general en la clasificación binaria analizada.

Por otro lado, Random Forest presentó una especificidad de 0.9375, superando ampliamente a los modelos lineales, cuyos valores estuvieron por debajo de 0.54. Esto significa que tiene una mayor capacidad para identificar correctamente a los hogares no pobres, lo cual resulta útil si el objetivo es reducir los errores de inclusión, es decir, evitar que se asignuen subsidios a hogares que no los necesitan.

Sin embargo, al evaluar la sensibilidad o recall, que mide la capacidad del modelo para identificar correctamente los hogares pobres, los modelos lineales mostraron un desempeño considerablemente superior (alrededor de 0.950), mientras que el modelo seleccionado obtuvo un valor notablemente más bajo (0.6124). Esta diferencia sugiere que los modelos lineales tienden a sobreidentificar hogares pobres, lo cual puede ser deseable si se busca evitar omitir beneficiarios potenciales de una política pública.

En cuanto a la precisión, los modelos lineales obtuvieron mejores resultados (alrededor de 0.890) frente al 0.7103, lo que refleja que, aunque el Random Forest clasifica correctamente a muchos hogares como no pobres, comete más errores al identificar hogares pobres. Esto también se refleja en el F1-score, métrica que combina precisión y sensibilidad, donde los modelos lineales obtienen un valor de 0.919, mientras que el modelo seleccionado alcanza solo 0.6577.

Al considerar el puntaje de Kaggle, que resume el desempeño del modelo en un entorno de evaluación externo, Random Forest alcanzó el valor más alto (0.6703), lo cual sugiere una mejor capacidad de generalización en comparación con los otros modelos (cuyo valor máximo fue 0.619). De esta forma, se identifica que, si el objetivo es evitar la exclusión de hogares pobres, los modelos lineales son preferibles por su alta sensibilidad. Por el contrario, si se busca evitar errores de inclusión y mejorar la precisión en la asignación de beneficios, el modelo de Random Forest ofrece ventajas importantes, destacándose además por su mejor desempeño general en la métrica de Kaggle.

3.1.4. Importancia de los atributos

En la Figura 6 se observan las variables más importantes en el modelo Random Forest para la predicción de pobreza en Colombia en 2018; muestra cuáles son los factores clave. En primer lugar, el hecho de que un hogar esté vinculado al sistema de salud subsidiado (*JH_RSS_S*) es el indicador más fuerte, lo que refleja una directa relación entre esta condición y la situación de pobreza. En otras palabras, depender del régimen subsidiado parece ser un claro reflejo de vulnerabilidad económica.

El número de hijos (*hijos*) se destaca como factor fundamental. Tener más personas a cargo implica mayores gastos y una mayor presión sobre los recursos económicos del hogar, lo que aumenta el riesgo de caer o permanecer en situación de pobreza. La recepción de ayudas directas para la alimentación (*subsidios*) es una variable clave en la predicción. Aquellos hogares que necesitan subsidios son altamente propensos a ser clasificados como pobres, evidenciando que una necesidad básica no se satisface completamente.

De la misma manera, la dimensión del hogar en términos del número total de integrantes (*Nper*) refuerza la idea de que familias más grandes pueden enfrentar mayores desafíos para satisfacer sus necesidades básicas con los ingresos disponibles. El nivel educativo promedio (*Educ_prom*) dentro de los hogares es una variable sumamente importante, puesto que muestra cómo una menor formación académica puede limitar las oportunidades de empleo y, por ende, los ingresos potenciales de las familias, aumentando la probabilidad de pobreza.

Por último, el valor estimado del arriendo de la vivienda (*P5130*) juega un papel en la predicción. Esta variable puede estar reflejando el valor del activo vivienda o su costo de oportunidad. Por ejemplo, vivir en una casa con un alto valor de arriendo podría asociarse a

mejores condiciones socioeconómicas o zonas con mayor costo de vida, lo que también implicaría mayores ingresos. En cambio, un valor bajo podría indicar una vivienda de menor calidad o ubicada en una zona más vulnerable, lo que tiende a estar relacionado con un mayor riesgo de pobreza.

En conjunto, estas variables muestran que la pobreza en los hogares colombianos durante 2018 está fuertemente relacionada con la falta de recursos, el tamaño de la familia y las oportunidades limitadas derivadas de un bajo nivel educativo. Son estos los factores que el modelo identificó como más determinantes para explicar la situación de pobreza.

4. Conclusiones

Este trabajo tiene como objetivo explorar cómo el uso de técnicas de aprendizaje automático, y en particular el modelo Random Forest, puede mejorar la predicción de la pobreza en los hogares colombianos. Para ello, se comparó el desempeño de Random Forest con el de modelos estadísticos más tradicionales, como los modelos lineales, con el fin de evaluar sus fortalezas y limitaciones relativas. El análisis se basó en datos detallados provenientes de encuestas de hogares en Colombia, los cuales permitieron entrenar y evaluar varios modelos predictivos

Los resultados mostraron que Random Forest superó a los modelos tradicionales en términos de precisión general, especialmente al momento de identificar correctamente a los hogares que no son pobres y al adaptarse mejor a datos nuevos. Aunque algunos modelos más simples fueron más sensibles al detectar hogares en situación de pobreza, Random Forest logró un equilibrio más adecuado entre sensibilidad y especificidad. Esta ventaja radica en su capacidad para capturar relaciones no lineales y complejas entre las distintas características del hogar, lo que lo hace especialmente útil en contextos sociales diversos y multidimensionales como el colombiano.

El análisis de las variables más influyentes en el modelo reveló patrones coherentes con la literatura existente. Entre los factores más determinantes se encontraron la afiliación al sistema de salud subsidiado, el número de hijos y personas en el hogar, la recepción de ayudas alimentarias y el nivel educativo promedio de los miembros del hogar. Estos elementos reflejan distintos ejes de la vulnerabilidad económica: desde el acceso limitado a servicios formales hasta la presión sobre los recursos familiares y las barreras estructurales al empleo y al ingreso.

Desde una perspectiva de política pública, estos hallazgos refuerzan la idea de que las herramientas de análisis basadas en aprendizaje automático pueden complementar los enfoques tradicionales y contribuir a diseñar intervenciones más focalizadas. Por ejemplo, podrían ayudar a priorizar la asignación de subsidios, optimizar el uso de recursos o

identificar con mayor precisión a los hogares que necesitan apoyo. Sin embargo, también plantean desafíos éticos y prácticos, como definir si se prefiere reducir los errores al excluir a hogares realmente pobres o al evitar incluir a quienes no lo son.

En resumen, este estudio demuestra que el uso de modelos como Random Forest puede aportar valor al análisis de la pobreza en Colombia, al permitir una comprensión más profunda de sus determinantes y al ofrecer herramientas más precisas para la toma de decisiones. Estas capacidades pueden traducirse en políticas sociales más efectivas, equitativas y basadas en evidencia.

5. Bibliografía

[1]. 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

[2]. Ariza, J. F., & Retajac, A. (2020). Descomposición y determinantes de la pobreza monetaria urbana en Colombia. Un estudio a nivel de ciudades. *Estudios Gerenciales*, 36(155), 167-176. https://doi.org/10.18046/j.estger.2020.155.3345

[3]. Banerjee, A., & Duflo, E. (2012). *Repensar la pobreza: Un giro radical en la lucha contra la desigualdad global*. Taurus.

[4]. Nuñéz, J., & Ramírez, J. C. (2002). *Determinantes de la pobreza en Colombia: Años recientes.* Cepal (Serie Estudios y Perspectivas 1). https://hdl.handle.net/11362/4789

[5]. Ravallion, M. (2003). *Las líneas de pobreza en la teoría y en la práctica*. Comisión Económica para América Latina y el Caribe.

[6]. Ray, D. (2002). *Economía del desarrollo*. Antoni Bosch.

[7]. Sen, A. (2000). *Desarrollo y libertad*. Planeta.

[8]. Lora, E. y Prada, S. I. (2023). *Técnicas de medición económica: metodología y aplicaciones en Colombia (Sexta edición)*. Editorial Universidad Icesi. https://doi.org/10.18046/EUI/tme.6

[9]. Gasparini, L., Cicowiez, M., & Sosa-Escudero, W. (2013). *Pobreza y desigualdad en América Latina: Conceptos, herramientas y aplicaciones.* Temas Grupo Editorial.

[10]. Sosa-Escudero, W., Anauati, M. V., & Brau, W. (2022). Poverty, inequality and development studies with machine learning. In F. Chan & L. Mátyás (Eds.), *Econometrics with machine learning: Vol. 53. Advanced studies in theoretical and applied econometrics* (pp. 291-335). Springer Cham. https://doi.org/10.1007/978-3-031-15149-1

[11]. Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata, *Science*, 350(6264), 1073-1076. https://doi.org/10.1126/science.aac4420

[12]. Blumenstock, J. (2018). Estimating economic characteristics with phone data, *AEA Papers and Proceedings,* 108, 72–76. https://doi.org/10.1257/pandp.20181033

[13]. Steele, J.E., Sundsøy, P.R., Pezzulo, C., Alegana, V.A., Bird, T.J., Blumenstock, J., Bjelland, J., Monsen, K.E., Montjoye, Y.A., Iqbal, A.M., Hadiuzzaman, K.N., Lu, X., Wetter, E., Tatem, A.J & Bengtsson, L. (2017). Mapping poverty using mobile phone and

satellite data, *Journal of the Royal Society Interface*, 14(127). https://doi.org/10.1098/rsif.2016.0690

[14]. World Bank. (2024). *Poverty, prosperity, and planet report 2024: Pathways out of the polycrisis.* World Bank. doi:10.1596/978-1-4648-2123-3.

[15]. Comisión Económica para América Latina y el Caribe. (2018). *Medición de la pobreza por ingresos: actualización metodológica y resultados*, Metodologías de la CEPAL, N° 2. https://repositorio.cepal.org/server/api/core/bitstreams/60b5f962-5ec5-4b6c-b36a-e0545ce6c2f4/content

[16]. Comisión Económica para América Latina y el Caribe. (2024). *Panorama social de América Latina y el Caribe 2024: Desafíos de la protección social no contributiva para avanzar hacia el desarrollo social inclusivo*, CEPAL. https://www.cepal.org/es/publicaciones/80858-panorama-social-america-latina-caribe-2024-desafios-la-proteccion-social

[17]. Departamento Administrativo Nacional de Estadísticas. (2024). *Pobreza monetaria en Colombia 2023*, Boletín técnico, DANE.

[18]. Ahumada, H., Gabrielli, M.F., Herrera, M., & Sosa-Escudero, W. (2022). *Una nueva econometría: Automatización, big data, econometría espacial y estructural*. Editorial de la Universidad Nacional del Sur.

6. Apéndice

6.1.1. Tablas y figuras

Tabla 1 *Estadísticas descriptivas de los atributos*

Statistic	N	Mean	St. Dev.	Min	Max
JH_Mujer	164,960	0.418	0.493	0	1
JH_Edad	164,960	49.612	16.390	11	108
JH_NoSeguro	164,960	0.058	0.235	0	1
JH_RSS_Subsidiado	164,960	0.402	0.490	0	1
JH_BasicaSecundaria	164,960	0.467	0.499	0	1
JH_Media	164,960	0.261	0.439	0	1
JH_Trabaja	164,960	0.631	0.483	0	1
JH_HorasExt	164,960	0.588	0.492	0	1
JH_Independiente	164,960	0.291	0.454	0	1
JH_Microempresa	164,960	0.182	0.386	0	1
JH_Pequena	164,960	0.052	0.222	0	1
JH_Mediana_Grande	164,960	0.185	0.388	0	1
Hijos	164,960	0.919	1.121	0	15
JH_RSS_S	164,960	0.402	0.490	0	1
Subsidios	164,960	1.397	1.649	0	20
Nper	164,960	3.292	1.775	1	28
Educ_prom	164,960	4.315	1.076	1.000	9.000
P5130	164,960	304,543.500	3,258,725.000	0	600,000,000
TGP	164,960	0.353	0.327	0.000	1.000
Trabajadores	164,960	1.346	0.967	0	10
JH_CotizaPension	164,960	0.278	0.448	0	1
Lp	164,960	271,522.300	33,656.890	167,222.500	303,816.700
JH_NEduc	164,960	4.367	1.404	1	9
tasa_desempleo	164,960	0.089	0.228	0.000	1.000
P5000	164,960	3.390	1.239	1	98
P5090	164,960	2.456	1.262	1	6
JH_Personas_Trabajo	164,960	33.317	24.892	0	130
AyudasEco	164,960	0.439	0.496	0	1

Nota. Elaboración propia

Tabla 2 Modelos de *clasificación de pobreza*

Variable	Logit	Ridge	Lasso	ElasticNet
(Intercept)	-2.54	-2.13	-2.54	-2.54
JH_RSS_S	0.743	0.582	0.742	0.742
Hijos	0.442	0.525	0.443	0.443
Subsidios	-0.821	-0.587	-0.820	-0.820
Nper	0.895	0.520	0.892	0.892
Educ_prom	-0.338	-0.308	-0.338	-0.338
P5130	-5.03	-0.152	-4.98	-4.98
TGP	0.295	0.275	0.295	0.295
Trabajadores	-0.549	-0.339	-0.547	-0.547
JH_CotizaPension	0.104	-0.0596	0.103	0.103
Lp	0.152	0.0867	0.151	0.151
JH_NEduc	-0.168	-0.170	-0.168	-0.168
tasa_desempleo	0.340	0.308	0.340	0.340
P5000	-0.341	-0.310	-0.341	-0.341
JH_Edad	-0.205	-0.226	-0.205	-0.205
P5090	0.194	0.209	0.194	0.194
JH_NoSeguro	0.320	0.242	0.320	0.320
JH_Personas_Trabajo	-0.0960	-0.0680	-0.0955	-0.0955
AyudasEco	-0.0667	-0.0133	-0.0663	-0.0663
Alpha		0	1	0.99
Lambda		0.12	0	0

Nota. Elaboración propia

Tabla 3 *Resultados modelos lineales*

Modelo	Accuracy	Sensibilidad (Recall)	Especificidad	Precisión	F1-Score	Kaggle
Logit	0.866	0.950	0.532	0.890	0.919	0.619
Ridge	0.860	0.957	0.469	0.878	0.916	0.576
Lasso	0.866	0.950	0.532	0.890	0.919	0.619
Elastic Net	0.866	0.950	0.532	0.890	0.919	0.619

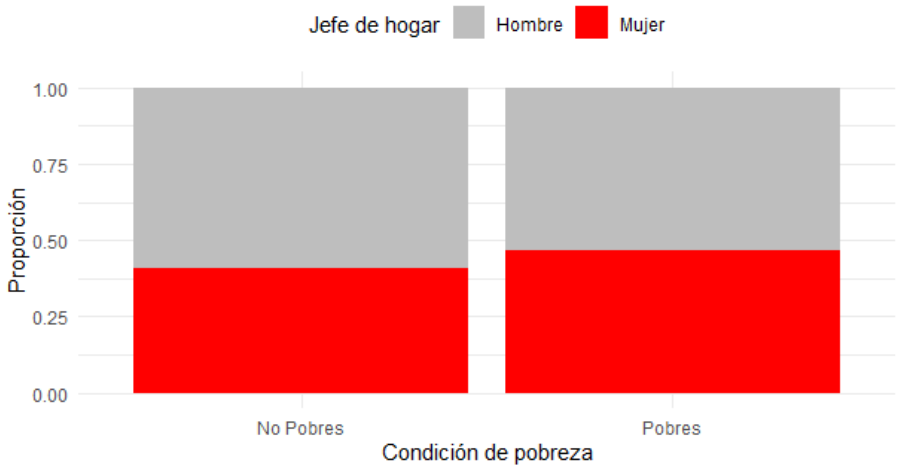
Nota. Elaboración propia

Tabla 4 *Resultados modelo de clasificación Random Forest*

	Métrica	Valor
Accuracy	Accuracy	0.8724
Sensitivity	Sensibilidad (Recall)	0.6124
Specificity	Especificidad	0.9375
Precision	Precisión	0.7103
F1	F1-Score	0.6577
	Kaggle	0.6703

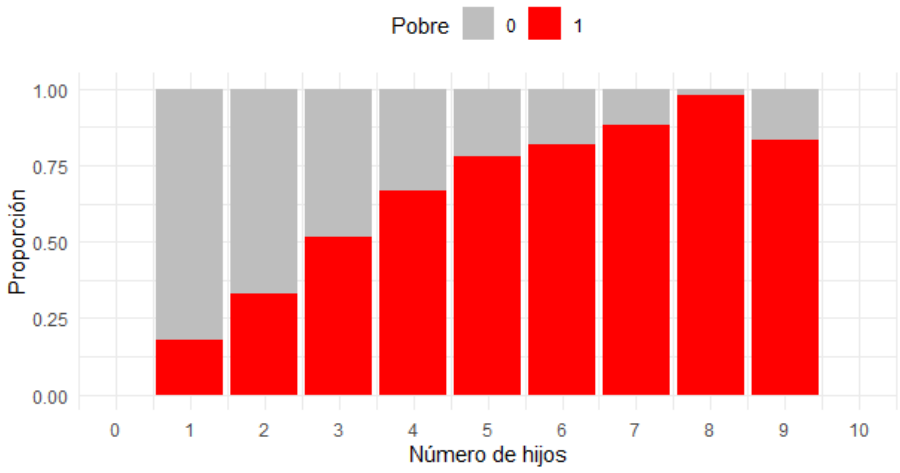
Nota. Elaboración propia

Figura 1 *Distribución de la pobreza por sexo del jefe de hogar*



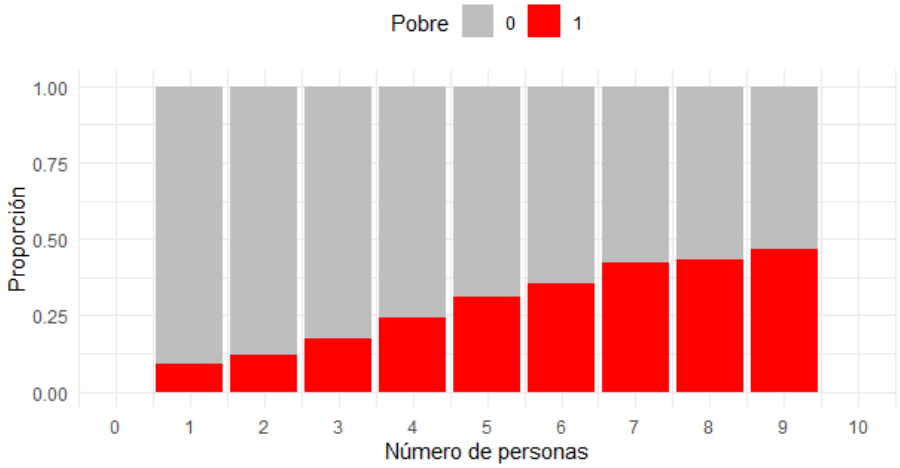
Nota. Elaboración propia

Figura 2 *Distribución de la pobreza por número de hijos*



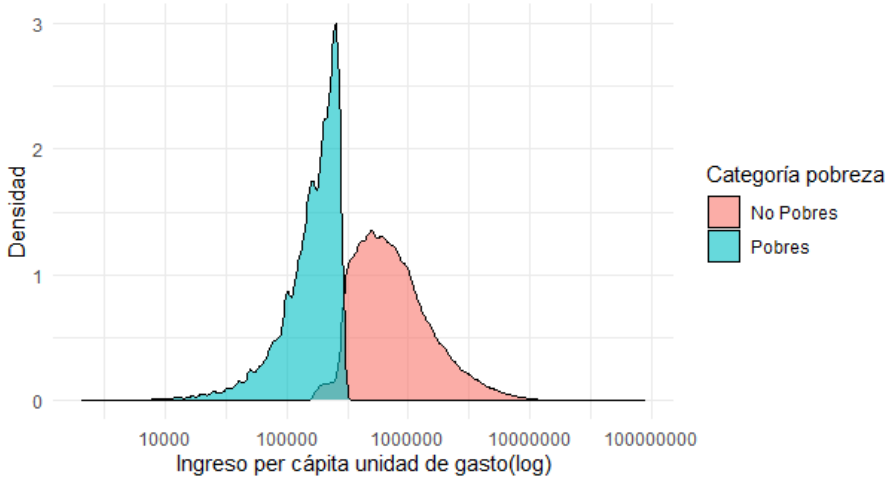
Nota. Elaboración propia

Figura 3 *Distribución de la pobreza por composición del hogar*



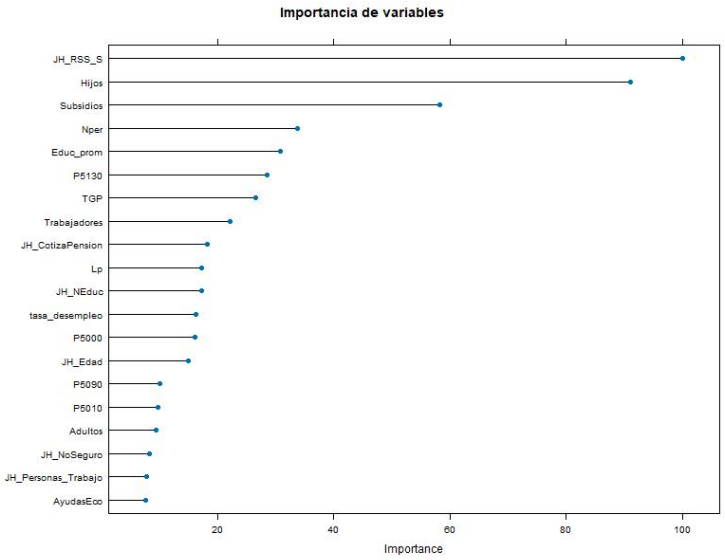
Nota. Elaboración propia

Figura 4 *Distribución del ingreso per cápita de la unidad de gasto*



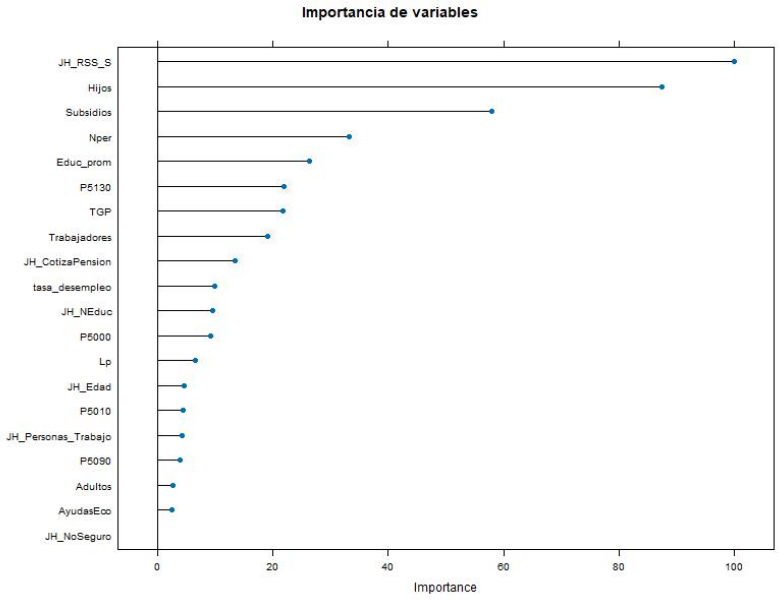
Nota. Elaboración propia

Figura 5 *Importancia de los atributos inicial (Random Forest)*



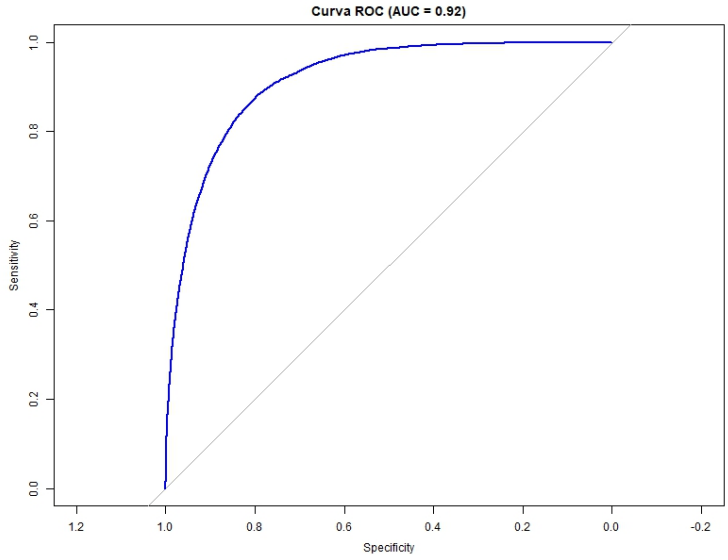
Nota. Elaboración propia

Figura 6 *Importancia final de atributos (Random Forest)*



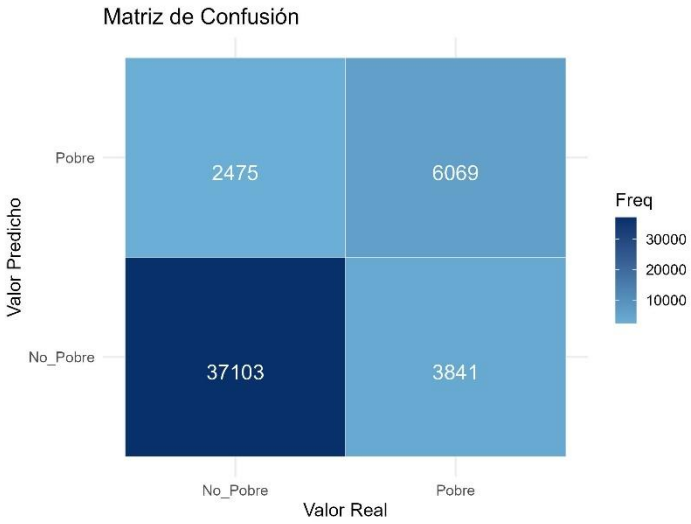
Nota. Elaboración propia

Figura 7 *Curva ROC (Random Forest)*



Nota. Elaboración propia

Figura 8 *Matriz de confusión (Random Forest)*



Nota. Elaboración propia