Predecir la pobreza en Colombia: una propuesta de modelos alternativos costo-eficientes

Mariana Correa, Rodrigo Iriarte, Marcel Montesdeoca, y Juan E. Moncada

Link repositorio: https://github.com/marianacs19/taller_2

1. Introducción

La medición de la pobreza es un desafío para los gobiernos y para algunas instituciones interesadas en la predicción de pobreza para mejorar el bienestar de la población mundial. Si bien los métodos tradicionales permiten obtener una medición de la pobreza, son más difíciles de desarrollar y más costosos. Es por esto que, desarrollar modelos predictivos a partir de un conjunto reducido de variables socioeconómicas, puede ser una herramienta útil, precisa y más costo-eficiente para el diseño de políticas públicas.

Por un lado, Colombia ha logrado avances significativos en la reducción de pobreza; sin embargo, persisten desigualdades en el acceso a oportunidades entre distintos grupos sociales y regiones del país y niveles de pobreza muy diferentes (Banco Mundial, 2024). Según un informe del Banco Mundial "Trayectorias: Prosperidad y reducción de la pobreza en el territorio colombiano" el lugar de nacimiento en Colombia puede determinar el futuro de una persona. Además, "los departamentos con mayores índices de pobreza tienden a tener una menor proporción de empleo formal, por debajo del 20 por ciento para departamentos como Nariño, Sucre, La Guajira y Cauca comparado con el 67 por ciento en Bogotá. (Banco Mundial, 2024).

Por otro lado, estudios como el de (Ariza & Retajac, 2020) argumentan que para estudiar la pobreza, es necesario fijarse en los determinantes micro y macro del nivel de ingreso de los hogares, a través de variables como el nivel educativo, la experiencia, el tipo de vinculación al mercado laboral y la composición del hogar. Los subsidios del gobierno y los régimenes de seguridad social en salud también juegan un rol importante en la predicción de la pobreza, puesto que el papel del Estado es fundamental en la erradicación de la pobreza a través de intrumentos de política social (Henao-Rodríguez, et al;, 2024). En el caso del género, las mujeres cuentan con más restricciones de acceso al mercado laboral por ejemplo, que a su vez tiene un impacto en el bienestar de ellas y de las personas a su cargo (Laverde-Rojas, et al;, 2020).

Para los modelos de predicción de pobreza que se desarrollaron, se utilizó la base de datos proveniente de la "Medición de pobreza monetaria y desigualdad 2018", publicada por el Departamento Administrativo Nacional de Estadística (DANE). Esta base cuenta con variables que permiten caracterizar la composición del hogar, el capital humano, la calidad de la vivienda y la inserción laboral. De esta manera, y teniendo en cuenta la revisión de literatura, se optó por elegir este tipo de variables para los diferentes modelos que se desarrollaron. Particularmente, se le dio importancia a variables demográficas como si el hogar está en una zona rural o urbana y variables socioeconómicas del jefe del hogar como sus años de educación, el oficio y la edad. Además se consideró relevante el uso de otras variables que señalan la composición del hogar como: el número de dormitorios, la cantidad de personas (tal y como lo demuestran Wang et al) y el pago del arriendo. En conjunto, esta batería de datos constituye una fuente robusta para estimar modelos predictivos de pobreza y analizar los determinantes estructurales asociados a ella en el contexto colombiano.

A lo largo del documento, se estiman y comparan distintos algoritmos de predicción como regresión lineal, LOGIT, Elastic Net, CARTs, Random Forest y Boosting con el fin de determinar cuál ofrece el mejor desempeño predictivo evaluado a través del F1-score. Finalmente, el modelo que mejor predice fue un Randon Forest estimado con las variables teóricas y optimizado para maximizar el F1 dentro de muestra. A pesar de que fue un buen modelo, sigue teniendo problemas que se pueden deber a que no se eliminaron los valores atípicos de algunas variables. El F1-score de 0.68 demuestra que aún hay espacio para mejorar el modelo para que pueda ser usado correctamente en el diseño de política pública.

2. Datos

La información proviene de la Medición de Pobreza Monetaria y Desigualdad 2018 del DANE, la cual cuenta con representatividad nacional y su unidad de análisis son los hogares colombianos. En el operativo original, la pobreza se define comparando el ingreso per cápita de la unidad de gasto (IPCG) con dos umbrales: línea de indigencia (LI) y línea de pobreza (LP). Las líneas se construyen con la Encuesta Nacional de Ingreso y Gastos (ENIG) de 2006–2007 (mes base marzo de 2007). Esta encuesta diferencia por dominio urbano–rural y se actualizan con el Índice de Precios al Consumidor (IPC) de alimentos (LI) e IPC total (LP) para ingresos bajos. El IPCG resulta del ingreso agregado del hogar entre personas en la unidad de gasto; antes de su cómputo se depuran faltantes, atípicos (regresiones cuantílicas) y falsos ceros; se imputan casos incompletos (hot deck y reglas por área/estrato) e incorpora arriendo imputado (por donantes, arriendo esperado o diferencia con amortización). La clasificación final es: no pobre (IPCG $\geq$ LP), pobre (IPCG <LP) y pobre extremo (IPCG<LI).

La base utilizada permite caracterizar condiciones de vida y del mercado laboral del hogar (tamaño, hacinamiento, tenencia, zona, localización), además de determinantes de pobreza (educación, inserción laboral, dependencias). Esta característica es adecuada para construir predictores (por ejemplo, hacinamiento, dependencia, costos de vivienda) y estimar modelos con capacidad explicativa y de generalización fuera de Bogotá. Así, se dispone de un insumo robusto para construir y depurar variables predictoras de pobreza y para estimar modelos con capacidad explicativa y poder de predicción. En este orden de ideas, la base final se construyó de la siguiente manera: se integraron cuatro archivos (personas y hogares para entrenamiento y prueba), después se armonizaron los nombres usando diccionarios que detallan el contenido de cada variable y se realizaron dos uniones persona–hogar por left join (una por conjunto), preservando la totalidad de personas y la consistencia del identificador (id).

Transformaciones principales: (i) binarios con mapeo 1→1 y 2,9,98,99,NA→0; (ii) estandarización de etiquetas (Urbano/Rural; hombre; `asegurado_salud`); (iii) educación como `años_educ` (a partir de nivel y último grado); (iv) `horas_trabajadas = principal + secundaria` (NA→0); (v) oficio agregado a grupos CIUO-08 (faltantes=99); (vi) `ciudad_cat` que agrega Bogotá y Medellín para controlar diferencias de cobertura entre conjuntos. Agregación por hogar (id): recuentos (personas, menores 5, mayores 65, ocupados, desocupados, afiliados), sumas (horas), y promedios (`años_educ_mean_hogar`). Se propagaron atributos del jefe (sexo, edad, escolaridad, actividad, régimen de salud). Faltantes y atípicos: en `espacios_hogar`, 98→mediana; en estados/beneficios, NA→0 cuando significa ausencia; para `arriendo_sumado` se aplicó cascada (arriendo pagado → arriendo estimado → cuota de amortización) con backfill y descarte de valores ≤ 100; los NA restantes se imputaron con grillas de promedios por celdas (`edad_jefe`, `ciudad_cat`, `mujer_jefe`, `años_educ_jefe`, `actividad_jefe`) con respaldo jerárquico y, en última instancia, media global. No se aplicó depuración general de outliers, solo reglas específicas de vivienda. La muestra final contiene un registro por hogar tras colapsar las variables constantes.

La base final reúne información a nivel hogar y persona, con 57 variables continuas y 49 categóricas. En relación con las continuas, la dispersión es heterogénea. Concretamente, las variables de composición del hogar muestran variabilidad. Personas hogar (media 3,29; desviación 1,78; rango 1–28), Espacios del hogar (3,39; 1,22; 1–43) y Dormitorios del hogar (1,99; 0,90; 1–15). Esta amplitud respalda su uso para capturar hacinamiento y tamaño del hogar, insumos clásicos para perfiles de pobreza. En capital humano, Años de educación del jefe (9,00; 4,91; 0–19) y Años de educación promedio del hogar (8,13; 3,87; 0–19) presentan dispersión suficiente para discriminar entre hogares. La intensidad laboral agregada también es informativa: Horas trabajadas en el hogar (68,37; 49,73; 0–640). En costos de vivienda, Arriendo imputado exhibe alta variabilidad (media 505.413; DE 3,37 millones;), reflejando valores extremos; aunque ya se filtraron montos ≤ 100 e imputó por grillas.

En las categóricas, se observan dos patrones útiles para selección. Primero, variables casi determinísticas aportan poca señal: por ejemplo, Población en edad de trabajar (jefe), Subsidio educativo (jefe), Primas (jefe) tienen modo $\approx 100\,\%$. Segundo, variables con distribución más balanceada o con múltiples categorías sí añaden información: Actividad semana pasada (jefe) y Propiedad de la vivienda capturan heterogeneidad relevante. Sexo del jefe no es perfectamente balanceada. En dominio geográfico,

Cuadro 1: Resumen descriptivo: variables continuas y categóricas

Variables continuas

Variable	N	Media	Desv. Est.	Mín	Máx
Dormitorios del hogar	164 960	1,99	0,90	1	15
Personas hogar	164 960	3,29	1,78	1	28
Asalariados en el hogar	164 960	0,64	0,79	0	9
Desocupados en el hogar	164 960	0,18	0,45	0	6
Menores de cinco en el hogar	164 960	0,24	0,51	0	7
Adolescentes en el hogar	164 960	0,22	0,49	0	5
Ancianos en el hogar	164 960	0,32	0,61	0	6
Educados en el hogar	164 960	0,77	0,97	0	10
Años de educación promedio del hogar	164 960	8,13	3,87	0	19
Ocupados en el hogar	164 960	1,50	1,03	0	14
Horas trabajadas en el hogar	164 960	68,37	49,73	0	640
Edad del jefe de hogar	164 960	49,61	16,39	11	108
Arriendo imputado	164 960	505 413,05	3 371 437,46	280	600 000 000

Variables categóricas

Variable	N	N niveles	Modo	Modo (%)*
Ciudad (categórica)	164 960	24	`BOGOTA_MEDELLÍN`	0,12
Mujer jefe	164 960	2	Hombre_cabeza_hogar	0,58
Ocupado jefe	164 960	2	Ocupado	0,71
Oficio C8 jefe	164 960	11	Desocupado_inactivo	0,29
Pobre	164 960	2	No_pobre	0,80
Propiedad vivienda	164 960	6	Arriendo	0,39
Régimen salud jefe	164 960	4	Contributivo	0,48
Urbano	164 960	2	Urbano	0,91

Notas: N se reporta sin decimales. *En “Modo (%)” los valores están como proporción (0–1). Las variables presentadas es una selección del total.

Departamento, Ciudad categórica y Dominio (urbano/rural) son candidatos centrales para fijar diferencias estructurales y, en particular, para absorber la variación de las líneas LI/LP, que en las continuas muestran variación moderada por dominio.

Con base en esta evidencia descriptiva, la estrategia de selección prioriza: (i) composición y demografía del hogar (personas, menores 5, mayores 65, ocupados/desocupados, escolaridad del jefe y promedio), (ii) condiciones de vivienda y costos (espacios, dormitorios, régimen de tenencia, arriendo imputado y su versión per cápita/transformada), (iii) inserción laboral del jefe (actividad, categoría ocupacional), y (iv) controles territoriales (departamento, dominio, ciudad categórica).Finalmente, dado que el train incluye Bogotá y el test no, las variables territoriales deben entrar de forma que reduzcan sobreajuste (p. ej., `Ciudad_cat` agregada, efectos por dominio/departamento), y conviene validar la estabilidad de los coeficientes/ganancias predictivas fuera de Bogotá mediante regularización y evaluación por subconjuntos geográficos.

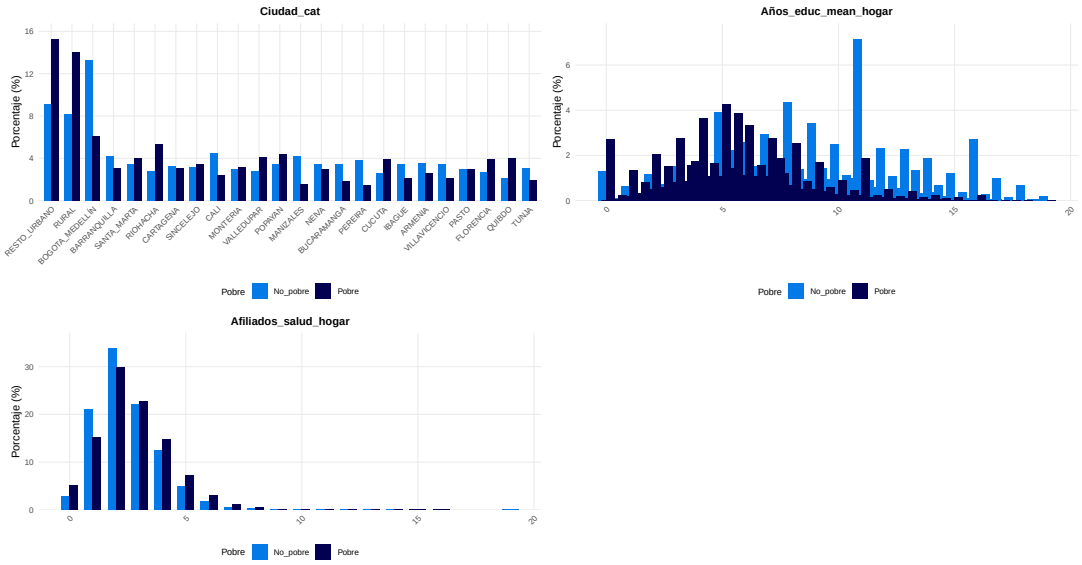

Figura 1: Resumen de variables por su condición de pobreza

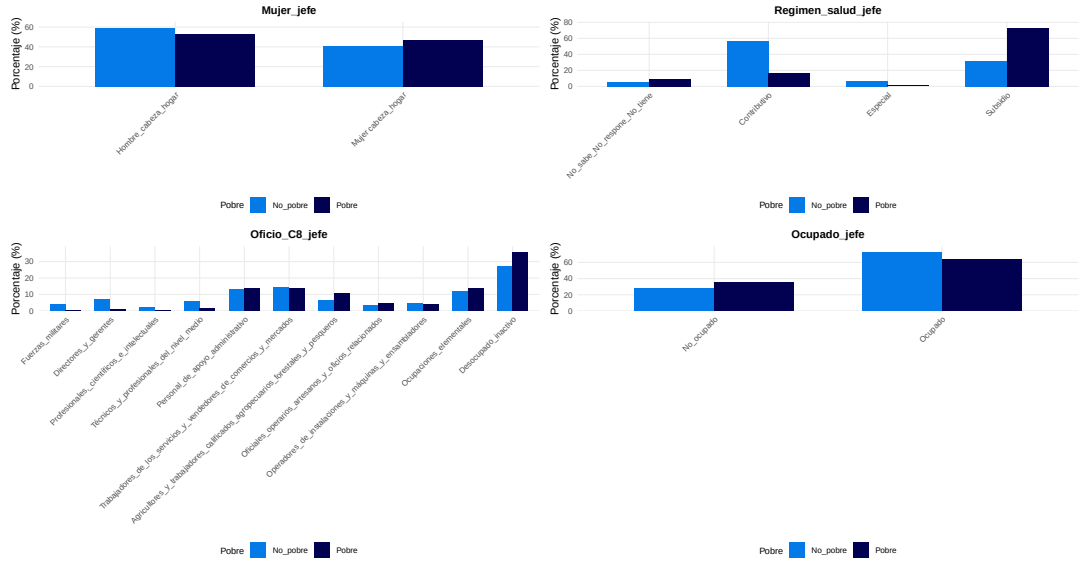

Figura 2: Resumen de variables por su condición de pobreza

3. Modelos y resultados

3.1. Selección del modelo y entrenamiento

El modelo con mayor capacidad predictiva y el que obtuvo el mayor puntaje en Kaggle (F1-score = 0.68) fue un Random Forest. En este algoritmo, se construyen varios árboles sobre una muestra bootstrap del conjunto de entrenamiento y se considera, en cada división, un subconjunto aleatorio de predictores para determinar cada partición. Así mismo, se estarían "descorrelacionando" los árboles (a diferencia de Bagging) y se reduce la varianza sin aumentar significativamente el sesgo. De esta manera, el modelo gana estabilidad y generalización, al evitar que todos los árboles dependan de los mismos predictores dominantes.

En los datos de entrenamiento se incluyeron 22 predictores, como variables que reflejan el capital humano, la composición demográfica, inserción laboral y condiciones de vivienda[1]. La elección de las variables se hizo después de una revisión de literatura y fueron seleccionadas por su relevancia teórica para explicar la pobreza en Colombia.

Con respecto al entrenamiento del modelo, se llevó a cabo 5-fold cross validation. En este proceso, en cada iteración, los datos de entrenamiento se dividieron aleatoriamente en cinco partes del mismo tamaño, donde cuatro se usaron para entrenar y una para validar. El proceso se repitió cinco veces.

Ahora bien, el modelo Random Forest requiere especificar y/o ajustar los siguientes hiperparámetros:

- **mtry**: hace referencia al número de variables en cada partición. Para el ajuste de este hiperparámetro, se llevó a cabo una grilla que evaluó bosques aleatorios considerando 2, 4, 6 y 8 variables posibles en cada partición del árbol.
- **min.node.size**: es el número mínimo de observaciones en cada nodo terminal. Para el ajuste de este hiperparámetro, se llevó a cabo una grilla para ver las diferentes combinaciones con 10 o 20 como número mínimo de observaciones en cada nodos.
- **ntree**: es el número de árboles. Para este hiperparámetro, se estableció un número de 500 árboles. Este número es comúnmente utilizado en los modelos de Random Forest ya que, como fue mostrado por (Oshiro et al., 2012) , las ganancias adicionales de cantidades de árboles mayores son marginales, pero sí implican un aumento en el costo computacional.
- **splitrule** es la medida de impureza en cada partición. En este caso se utilizó la medida por el índice Gini, la cual mide qué tan bien el modelo separa las clases.

El ajuste de estos valores es esencial para lograr un balance entre sesgo y varianza, evitando tanto el sobreajuste como la pérdida de capacidad predictiva. Adicionalmente, la métrica de comparación es el área bajo la curva ROC (AUC), dado que mide la capacidad del modelo para discriminar correctamente entre hogares pobres y no pobres. El uso de esta métrica es pertinente debido a que existe un imbalance de clases y el AUC no se ve afectado por la distribución de la variable Pobre. Es relevante mencionar que se desarrolló otro modelo Random Forest en el que se utilizaron pesos como método para solucionar el imbalance de clases; sin embargo esto no significó ningún aumento en la predictibilidad de los modelos aquí descritos. Como se puede ver en el Cuadro 2, en el proceso de ajuste de hiperparámetros se evaluaron combinaciones de los parámetros mtry y min.node.size.

Cuadro 2: Resultados de validación cruzada del modelo Random Forest

mtry	min.node.size	ROC	Sensibilidad	Especificidad	Desv. ROC
2	10	0.9158	0.9856	0.3323	0.0009
2	20	0.9158	0.9858	0.3299	0.0009
4	10	0.9273	0.9623	0.5501	0.0007
4	20	0.9271	0.9633	0.5449	0.0007
6	10	0.9295	0.9576	0.5902	0.0006
6	20	0.9295	0.9584	0.5852	0.0004
8	10	0.9301	0.9553	0.6072	0.0004
8	20	**0.9303**	0.9562	0.6022	0.0006

Los resultados muestran que un mtry de 8 introduce aleatoriedad y "descorrelaciona" los árboles, reduciendo la varianza. Además un min.node.size de 20 significa que cada hoja del árbol contiene al menos

[1]Número de dormitorios, el tipo de propiedad y el valor del arriendo, el tamaño del hogar, el número de personas que comparten gastos, la presencia de menores de cinco años y de adultos mayores, años promedio de educación del hogar, la educación y edad del jefe, el número de ocupados, asalariados y desocupados, y las horas trabajadas; variables de afiliación al sistema de salud y régimen del jefe como aproximaciones a la formalidad laboral, así como la condición urbana o rural y la categoría de ciudad

20 observaciones, lo cual evita el sobreajuste del modelo. Estos son los hiperparámtros que maximizan el AUC con un valor de 0.93, por lo que son los escogidos para el modelo. En conjunto, estos resultados confirman que el modelo logra un buen balance entre sesgo y varianza, mostrando capacidad para capturar relaciones entre las variables y características de los hogares.

Después de escoger la combinación óptima de hiperparámetros, el modelo final fue entrenado otra vez sobre el conjunto completo de entrenamiento y evaluado utilizando un cutoff de 0,3. Con estas configuraciones de variables, hiperparámetros y cutoff, el modelo Random Forest alcanzó un F1-score de 0.71 en la muestra de entrenamiento y de 0.68 fuera de muestra, logrando un buen equilibrio entre precisión y recall.

El área bajo la curva ROC confirma la alta capacidad discriminatoria del modelo para clasificar correctamente hogares pobres y no pobres. Sin embargo, el F1-score de 0.68 evidencia que aún existen errores de clasificación relevantes, particularmente falsos positivos, lo cual limita su aplicabilidad directa en el diseño de política pública.

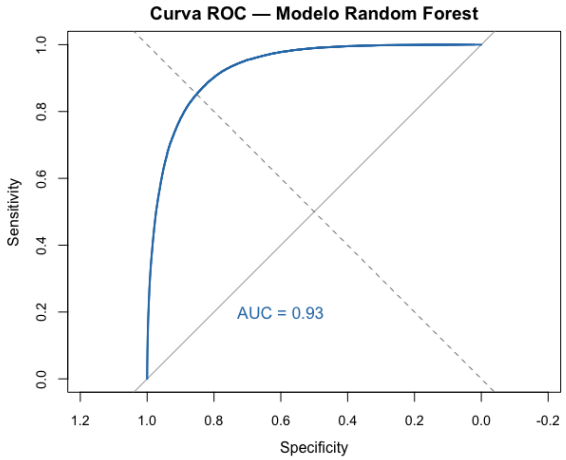

Figura 3: Curva ROC: Modelo Random Forest

3.2. Otros modelos: descripción general y determinación de sus hiperparámetros

A continuación se presenta la descripción de los 5 modelos que tuvieron mejores resultados por fuera de muestra. En cada uno de ellos se muestran los alcances, limitaciones y metodologías escogidas para la definición de los hiperparámetros.

3.2.1. Modelo 4A — Regresión Logística optimizada para el puntaje F_1

Motivación y desarrollo. El modelo parte de una regresión logística parsimoniosa para estimar la probabilidad de pobreza del hogar ($Pobre = 1$) a partir de indicadores clave (educación, empleo asalariado, arriendo y afiliación en salud). Tras validar el modelo base, la especificación se amplió progresivamente mediante *backward selection* y refinamiento teórico, incorporando dimensiones adicionales de vivienda, composición del hogar, capital humano, participación laboral y contexto urbano.

Estrategia de estimación y umbral. La estimación se realizó por Máxima Verosimilitud con enlace logit binomial. El umbral de decisión no se fijó en 0,5, sino que se calibró empíricamente probando puntos de corte en $[0, 1]$ (pasos de 0,01) para maximizar el F_1. El valor óptimo fue $c = 0{,}33$, con $F_1 = 0{,}685$ y AUC $\approx 0{,}91$ en validación interna; fuera de muestra se obtuvo $F_1 = 0{,}67$. Esta calibración mejoró el equilibrio sensibilidad–precisión sin introducir penalización.

Interpretación y rol. El Modelo 4A preserva la interpretabilidad propia del logit (coeficientes como cambios marginales en los *log-odds* de pobreza) y, gracias a la calibración del umbral, alcanza un desempeño competitivo que sirve de referente para contrastar modelos más complejos. Sin embargo, esta simplicidad también es una de sus limitaciones; puede que esté dejando de lado información relevante.

3.2.2. Modelo Logit penalizado — Regularización Elastic Net

Motivación y evolución. Aunque el Modelo 4A mostró alta precisión, su simplicidad podía obviar relaciones importantes sin explorar. Para ampliarla sin sobreajustar ante predictores correlacionados, se estimó un logit penalizado con Elastic Net, incorporando tres predictores adicionales teóricamente relevantes (p. ej., edad promedio ponderada del hogar, presencia de adolescentes y actividad laboral de la semana pasada del/de la jefe/a). La regularización permitió ampliar el conjunto de variables controlando la flexibilidad y resguardando la estabilidad fuera de muestra.

Regularización y ajuste. Elastic Net combina penalizaciones LASSO (L_1) y Ridge (L_2), equilibrando selección de variables y *shrinkage*. Los hiperparámetros λ (intensidad de penalización) y α (balance L_1–L_2) se ajustaron vía validación cruzada de 5 pliegues para estimar de forma robusta el error fuera de muestra. Se optimizó el F_1 manteniendo el umbral fijo en $c = 0{,}33$ para poder facilitar la comparabilidad directa con el Modelo 4A. La configuración óptima fue $\alpha^* = 0{,}25$ y $\lambda^* = 0{,}01$, lo que sugiere una ligera tendencia Ridge: la mayoría de predictores aportaban señal y no fue necesaria una penalización intensa para estabilizar coeficientes.

Resultados e interpretación. El desempeño fuera de muestra fue prácticamente indistinguible del Modelo 4A tanto en F_1 como en AUC. Esta equivalencia indica baja redundancia o ruido en los predictores y que el 4A ya capturaba los determinantes estructurales de la pobreza. El λ pequeño respalda que la multicolinealidad era moderada y que el *shrinkage* requerido fue leve.

Discusión: El Elastic Net funcionó más como validación robusta que como mejora sustantiva frente al logit no penalizado. Sus parámetros ($\alpha = 0{,}25$, $\lambda = 0{,}01$) muestran que la penalización aportó estabilidad sin modificar resultados predictivos: la especificación simple y guiada por teoría del Modelo 4A ya generalizaba adecuadamente sin requerir restricciones adicionales. Este hecho sin embargo otorga información valiosa: el modelo Logit con las variables escogidas es una poderosa herramienta. Se destaca que estos dos modelos fueron hicieron parte del los tres mejores modelos estimados.

3.2.3. XGBoost optimizado para F1

Motivación: Se han documentado casos de pronósticos con un alto nivel de precisión en distintos campos usando XGBoost. Por ejemplo, (Shen, 2025) demostró la utilidad de este tipo de modelos para predecir el precio de acciones en el mercado bursátil estadounidense, obteniendo mejores resultados que otros modelos de predicción tradicionales. De manera similar, (Abhinaya et al., 2024) compararon el modelo XGBoost frente a otros modelos de predicción de granizados, obteniendo una mejora significativa en sus pronósticos del clima. Más interesante aún, (Zhang et al., 2025) exploraron la idea de usar el XGBoost para predecir la pobreza en Estados Unidos. La conclusión de su estudio demostró una precisión del 88 %, demostrando su utilidad también para la economía. Frente a los buenos resultados del XGBoost en la literatura, se decidió explorar la idea de implementar este modelo para predecir la pobreza en Colombia.

Proceso de estimación de hiperparámetros: Para llevar a cabo este modelo, se utilizaron 22 variables, las cuales consideramos podían ser las mejores predictoras de pobreza con base en la literatura y en los resultados de nuestra selección (como se ha mencionado anteriormente). Para evitar sobreajustar el modelo a los datos, se escogió hacer entre 250 y 500 árboles con una profundidad máxima de dos nodos. La razón de este número bajo de nodos es el menor costo computacional, la búsqueda de menor variabilidad, evitar sobreajustar el modelo y mayor interpretabilidad. Una característica del XGBoost es su tasa de aprendizaje, la cual le da un peso a cada árbol de acuerdo con su aporte al pronóstico. Así, se escogió una tasa de aprendizaje del 0.01, la cual, a pesar de ser lenta, permite una mayor estabilidad

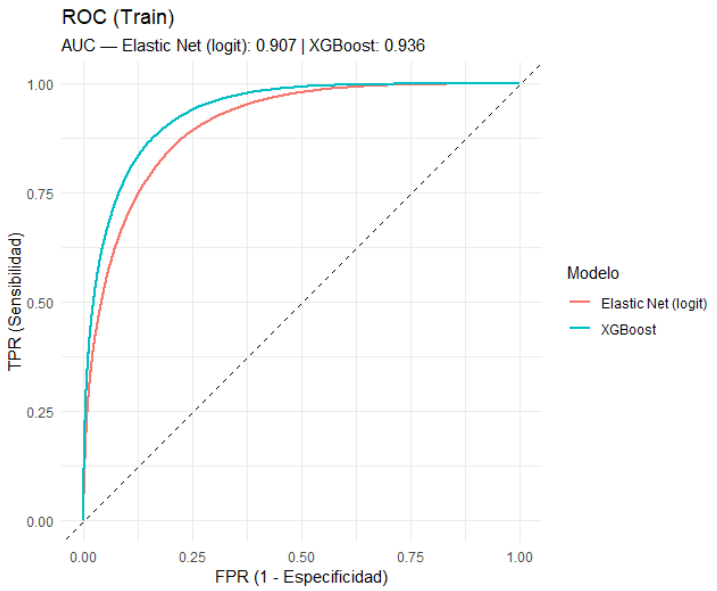

Figura 4: ROC para Elastic Net y XGBoost

en el aprendizaje. El resto de los parámetros se escogieron pensando en reducir correlaciones entre los árboles, mejorar la robustez y conseguir una muestra subset para hacer validación cruzada.

3.2.4. Elastic Net optimizado para F1

Motivación: A pesar de ser un modelo comparativamente sencillo, el Elastic Net también ha demostrado una gran utilidad para apoyar los modelos predictivos en distintos campos. Por ejemplo, Uniejewski, Nowotarski, y Weron (2016) demostraron la utilidad del Elastic Net como un método para seleccionar variables para pronósticos del precio de la energía. Más relevante aún, (Tansuchat et al., 2023) consiguieron pronosticar el PIB de Tailandia con una precisión del 99 % usando Elastic Net. Con estos buenos resultados, se decidió explorar el Elastic Net para pronosticar la pobreza en Colombia.

Proceso de estimación de hiperparámetros: Para pronosticar, se utilizaron 22 variables escogidas mediante revisión de literatura y backward selection. El Elastic Net, al ser una combinación de Ridge y Lasso, cuenta con dos parámetros para ser estimados: la penalización Lambda y Alpha. Así, se utilizó una grilla que combinaba distintos valores de Alpha y Lambda para dar con la mejor estimación usando como criterio maximizar el F1. Entonces, se escogió un Alpha entre 0 y 1 que aumentara de a 0.1 unidades, resultando en 100 modelos. Con relación al Lambda, se utilizó un Lambda que fuera desde 0.001 hasta 1000 con aumentos de a 10, lo que resultó en otros 100 modelos. Al usr validación cruzada de 5 pliegues, se están estimando alrededor de 550 modelos.

Discusión: Al estimar varios modelos, no se impone uno a la fuerza (considerando solo Ridge o solo Lasso) y se permite una combinación de parámetros sobre las cuales el modelo puede aprender y resultar en la combinación óptima que maximice el F1. Para evaluar el desempeño del modelo, se utilizó una validación cruzada de cinco niveles. Esta decisión fue impulsada por el trade-off entre variabilidad y sesgo. En concreto, se consideró que un nivel más bajo podría dividir los folds en muestras más grandes y esto generaría variabilidad en los resultados. Si se escogía un nivel más alto, los folds serían más pequeños. A pesar de ofrecer menor sesgo, se presentaría mayor varianza.

3.2.5. Random forest con pesos

Motivación: El modelo corresponde a un bosque aleatorio binario entrenado. La especificación incluye 58 variables. En términos conceptuales, se combinaron atributos demográficos y de composición del hogar,

intensidad de trabajo, educación, condiciones del mercado laboral, entorno urbano-rural y categorías geográficas/ocupacionales, junto con un conjunto amplio de interacciones con la jefatura femenina y una interacción adicional entre condición urbana y la categoría geográfica. Esta estructura buscó capturar no linealidades y heterogeneidades relevantes sin recurrir a variables textuales o identificadores directos.

Estimación de hiperparámetros: El entrenamiento se realizó usando 500 árboles, criterio de partición Gini, y validación cruzada estratificada de cinco pliegues. Se fijó semilla (123) para reproducibilidad. Para atender el desbalance de clases: a la clase positiva se le asignó un peso igual al cociente entre la frecuencia de la clase mayoritaria y la minoritaria. La selección de hiperparámetros se hizo mediante rejilla (mtry = 5, 8, 11 y min.node.size = 30, 50), eligiendo la combinación que maximizó F1 en validación cruzada. Los valores finales utilizados fueron: mtry = 11, splitrule = gini y min.node.size = 30. Adicionalmente, se eligió el umbral de decisión optimizando F1 sobre probabilidades, explorando cortes cada 0,005.

Discusión: Con el cut-off 0,655, el F1 en entrenamiento fue 0,86 y en prueba 0,65. La ganancia de F1 frente a alternativas lineales es consistente con la capacidad del bosque para modelar interacciones y no linealidades en presencia de muchas variables discretas y continuas. La estrategia de ponderación también ayudó a recuperar verdaderos positivos sin deteriorar en exceso la precisión.

¿Por qué fue "uno de los mejores" y no "el mejor"? El diferencial entre F1 de entrenamiento (0,86) y de prueba (0,65) sugiere cierto sobreajuste. Hay tres fuentes plausibles: i) el número y la naturaleza de las interacciones (incluida una triple) multiplican el espacio efectivo de particiones, permitiendo que algunos árboles capturen idiosincrasias del fold de entrenamiento; ii) el umbral óptimo se calibró sobre el conjunto de entrenamiento (o out-of-fold), y trasladarlo tal cual puede inflar el ajuste in-sample frente a out-of-sample; iii) la rejilla fue relativamente estrecha, lo que limita el control fino de la complejidad: con muchos predictores relevantes, un mtry grande combinado con un min.node.size no lo bastante alto puede permitir particiones muy profundas en subpoblaciones poco representadas. En conjunto, el modelo captura bien la señal (de ahí que esté entre los mejores), pero la brecha de desempeño sugiere sensibilidad a ruido y a la calibración del corte, lo que impidió que fuera el mejor fuera de muestra.

3.3. Análisis comparativo intermodelos

Comprender las métricas de desempeño es fundamental para interpretar la capacidad predictiva de los modelos. Ese es el propósito del Cuadro 3 que muestra algunas métricas clave (en orden descendiente):

Cuadro 3: Resultados generales de desempeño de los modelos

Modelo	F_1 **(muestra)**	F_1 **(fuera de muestra)**	**AUC ROC**
Random Forest	0.71	0.68	0.93
Logit (umbral optimizado)	0.68	0.67	0.91
Logit penalizado (Elastic Net)	0.67	0.67	—
Random Forest Agrandado	0.86	0.65	0.92
XGBoost	0.70	0.66	0.94
Regresión Elastic Net	0.62	0.63	0.91

En este caso, se observa una relación positiva entre el área bajo la curva ROC (AUC) y el puntaje (F1) fuera de muestra, lo que sugiere que los modelos con mejor capacidad de discriminación también logran un balance adecuado entre precisión y sensibilidad. Asimismo, se evidencia que el score objetivo dentro de muestra es ligeramente superior al obtenido en el conjunto de prueba, un comportamiento esperable dado que el modelo tiende a ajustarse con mayor precisión a las características del conjunto de entrenamiento. No obstante, la brecha relativamente pequeña entre ambos valores indica una buena capacidad de generalización: los modelos no solo capturan las particularidades del conjunto de entrenamiento, sino también los patrones estructurales que explican la pobreza en nuevos datos. Siendo esto al mismo tiempo un indicativo positivo de la forma en la que se está manejando el sesgo-varianza.

El modelo que fortalece este argumento es el Random Forest agrandado, cuya distancia entre el valor F1 en muestra y el F1 fuera de muestra es grande. En ese sentido, seguramente el modelo aprendió las dinámicas de las variables dentro de la muestra en lugar de sus versiones generalizables. Otro aspecto que soporta esta visión es que el Random Forest con menos variables obtuvo tanto una distancia menor entre sus métricas dentro y fuera de la muestra, como un mejor resultado

Ahora bien, la inclusión de dos modelos logit dentro de la estimación (uno de ellos con optimización de umbral), refuerza la idea de que la simplicidad metodológica puede alcanzar resultados comparables con algoritmos más complejos. La posición destacada del modelo logit con umbral optimizado (segundo en desempeño general) sugiere que un ajuste cuidadoso del punto de decisión puede ser tan efectivo como la regularización o el uso de técnicas más sofisticadas. De igual forma, la presencia de un modelo lineal dentro del conjunto evaluado, con resultados consistentes, respalda la relevancia de los enfoques más clásicos y fundamentados teóricamente frente a métodos que pueden priorizar la complejidad.

Cuadro 4: Detalle de hiperparámetros, método de ajuste y tiempos de estimación

Modelo	**Hiperparámetros ajustados**	**Método de ajuste**	**# Variables**	**Tiempo estimación**
Random Forest	1) Vars. en cada partición. 2) Num. mínimo de obs. nodo.	Búsqueda en grilla para maximizar el AUC ROC.	22	1 hora
Logit (umbral optimizado)	1) Punto de decisión (c) que maximiza el F_1.	Búsqueda en grilla para maximizar F_1.	23	2 minutos
Logit penalizado (Elastic Net)	1) Penalización λ y balance Ridge–LASSO α.	Validación cruzada (*cross-validation*) para maximizar el AUC ROC.	26	10 minutos
XGBoost	1) Num. de rondas. 2) Profundidad máxima. 3) Proporción de vars. muestreadas.	Búsqueda en grilla para maximizar F_1.	23	1 horas
Random Forest Agrandado	1) Num. de rondas. 2) Profundidad máxima.	Five fold cross validation para maximizar F_1.	58	2 horas
Regresión Elastic Net	1) Penalización λ y balance Ridge–LASSO α.	Búsqueda en grilla para maximizar F_1.	22	45 minutos

Los modelos evaluados pueden agruparse en dos grandes familias: (i) modelos basados en árboles como los Random Forest y XGBoost, y (ii) modelos de naturaleza lineal, como las regresiones logísticas penalizadas y no penalizadas. Cada grupo presenta ventajas distintas en términos de flexibilidad, interpretabilidad y costo computacional.

En primer lugar, los resultados muestran que no existe evidencia concluyente de una superioridad sistemática de un tipo de modelo sobre el otro en términos de desempeño predictivo. Sin embargo, sí se observan diferencias marcadas en los costos computacionales: los modelos de tipo bosque, al requerir la estimación de múltiples árboles y la búsqueda exhaustiva de combinaciones de hiperparámetros, demandan tiempos de entrenamiento significativamente mayores. En contraste, los modelos logit penalizados o con umbral optimizado alcanzan resultados comparables en cuestión de minutos, lo que los convierte en alternativas más eficientes cuando se dispone de recursos limitados o se requiere escalabilidad.

Un segundo hallazgo relevante es que los modelos ajustados optimizando el área bajo la curva ROC, siempre que la estimación no tuviera muchas variables, obtuvieron resultados muy similares a los calibrados para maximizar el puntaje F_1. Esta consistencia sugiere que la estructura del problema de clasificación; es decir, la relación entre las variables predictoras y la condición de pobreza es estable frente a distintas métricas de evaluación. En consecuencia, tanto la precisión en la discriminación (ROC) como el equilibrio entre sensibilidad y precisión (F_1) convergen hacia un mismo conjunto de parámetros óptimos.

Finalmente, la comparación entre las dos familias de modelos revela que los enfoques lineales, pese a su simplicidad, logran capturar adecuadamente la relación funcional entre las variables socioeconómicas y la probabilidad de pobreza. Esto se alinea con la naturaleza del fenómeno: la pobreza, entendida como estar por encima o por debajo de un umbral monetario, puede representarse adecuadamente mediante funciones logísticas. Por su parte, los modelos basados en árboles ofrecen mayor flexibilidad para capturar

relaciones no lineales o interacciones complejas, pero en este caso esa flexibilidad adicional no se tradujo en mejoras sustanciales en desempeño fuera de muestra.

3.4. Importancia de las variables

3.4.1. Justifiación Teórica de las variables

Varias de las variables empleadas en la estimación están respaldadas por la literatura sobre los determinantes estructurales de la pobreza, que subraya la importancia del capital humano, la composición del hogar, las condiciones de vivienda, la inserción laboral y las desigualdades territoriales (Banco Mundial, 2024; Henao-Rodríguez et al., 2024; Laverde-Rojas et al., 2020).

Las variables educativas, como los años de educación promedio del hogar y los años de educación del jefe, se sustentan en la evidencia que vincula la educación con una mayor productividad y mejores ingresos. Este componente del capital humano es central para la movilidad social y la estabilidad económica. De forma similar, las variables laborales, entre ellas ocupados en el hogar, asalariados, horas trabajadas, oficio del jefe y régimen de salud, reflejan la relevancia del empleo formal y estable para la generación de ingresos. El Banco Mundial (2024) destaca que la baja formalización laboral explica buena parte de las diferencias en pobreza entre regiones del país, lo cual justifica su inclusión en el modelo.

Por otra parte, las características del hogar, como el número de personas, los menores de cinco años o los ancianos en el hogar, permiten capturar las cargas demográficas y las dependencias económicas internas. En ese sentido, variables como mujer jefe de hogar incorporan dimensiones de género que influyen en la vulnerabilidad económica.

Por último, variables como propiedad de la vivienda, número de dormitorios y condición urbano/rural (Urbano) representan las condiciones materiales y las brechas territoriales en acceso a oportunidades, factores ampliamente documentados como determinantes del bienestar.

3.4.2. Justifiación Empírica de las variables

Para la selección de variables del modelo Random Forest se hizo una pre-selección de todas las variables disponibles con base en la teoría descrita previemanete y la revisión de literatura. A partir de las variables pre seleccionadas se llevó a cabo un proceso de backward selection para determinar las que mejor predicen la pobreza.

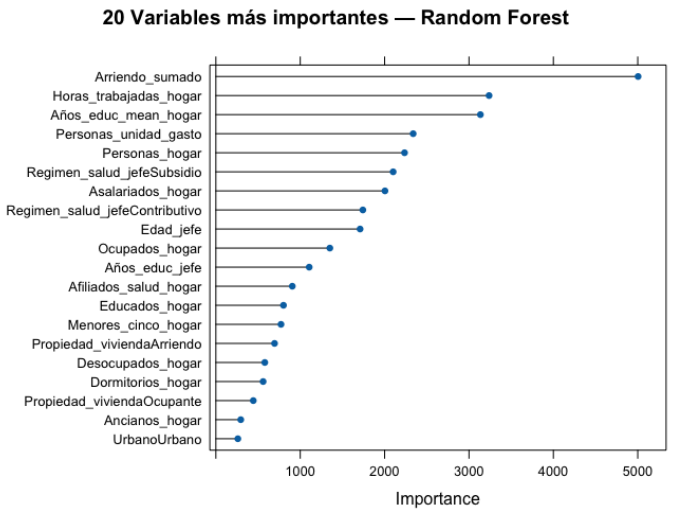

Figura 5: 20 variables más importantes en Random Forest

La 5 muestra las 20 variables más relevantes según el índice Gini, que mide cuánto contribuye cada predictor a mejorar la pureza de los nodos terminales en los árboles. La variable de arriendo aparece como la de mayor importancia, lo cual sugiere que el costo de la vivienda es el factor más influyente para determinar la condición de pobreza. Luego hay variables relacionadas al capital humano y a la inserción laboral, como horas trabajadas por hogar, promedio de años de educación y asalariados por hogar, lo que indica que la intensidad y calidad del trabajo, así como la educación promedio del hogar, son determinantes para superar la línea de pobreza. Esto tiene sentido, después de todo, estas variables están estrechamente relacionadas con el ingreso que reciben los hogares.

Asimismo, variables como personas como unidad de gasto y personas por hogar capturan la dimensión demográfica, reflejando el tamaño del hogar. Lo que representa aquello en que los hogares se gastan lo que han ganado. El modelo también otorga un peso relevante a las variables relacionadas con el acceso a servicios y protección social, como régimen de salud del jefe del hogar, afiliados a salud en cada hogar evidenciando la importancia de la cobertura de salud como una manera de medir de formalidad laboral y estabilidad económica.

3.5. Conclusión

La medición de la pobreza tiende a ser un proceso complicado y costoso. Los resultados de este estudio demostraron la viabilidad de métodos alternativos para predecir la pobreza con unos resultados acertados. La revisión de literatura enseña que, para el caso colombiano, las variables relacionadas con el ingreso, el empleo, la educación, las características del hogar, el género, y el departamento de pertenencia en el país, son fuertes predictores de pobreza en los modelos, lo que puede ser interpretado como fuertes determinantes de pobreza en el contexto colombiano. En relación con los modelos, se puede concluir que son una oportunidad costo-efectiva viable para el diseño de políticas públicas con un alto nivel de precisión.

Por un lado, el modelo de Random Forest permitió la precisión más alta de todos los modelos presentados, llegando a tener un F_1 de 0.68 evaluado fuera de la muestra y un AUC de 0.93. Este modelo, a pesar de sufrir de un costo computacional alto, permite capturar relaciones complejas que otros modelos no alcanzan a identificar, ofreciendo un pronóstico altamente acertado y robusto con un bajo nivel de variables predictoras. Por otro lado, los modelos de Logit, Elastic Net y XGBoost permitieron también obtener predicciones altamente acertadas, al estar entre un rango de 0.63 y 0.66. Sin embargo, es oportuno rescatar que los modelos más complejos no necesariamente pronosticaron más acertadamente, enseñando que los modelos lineales corectamente especificados pueden capturar adecuadamente la relación entre variables socioeconómicas y pobreza con un menor costo computacional. Estos resultados, si bien tienen espacio para mejoras, evidencian una potencial para pronosticar acertadamente variables que pueden ser de interés para políticas públicas a un costo comparativamente más bajo.

Por úlimo, a pesar de tener resultados prometedores, es importante mencionar que los modelos ofrecen limitaciones. Particularmente, uno de los mayores retos es el sobreajuste, lo que se apreció al tener mediciones más altas de F_1 en la muestra que fuera de ella. Este caso particular se presentó para todos los modelos, de manera que es preciso mencionar que estos modelos pueden estar sufriendo de un sobreajuste a pesar de haber hecho un proceso de estimación de hiperparámetros.

Referencias

Abhinaya, P., Reddy, C. K. K., Ranjan, A., & Ozer, O. (2024). *Explicit monitoring and prediction of hailstorms with XGBoost classifier for sustainability*. https://doi.org/10.4018/979-8-3693-3896-4.ch006

Ariza, J. F., & Retajac, A. (2020). Decomposition and determinants of urban monetary poverty in Colombia. A study at the city level; [Decomposição e determinantes da pobreza monetária urbana na Colômbia. Um estudo no nível de cidades]; [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

Banco Mundial. (2024). Informe sobre pobreza del Banco Mundial destaca desigualdades persistentes en Colombia [Recuperado de la página oficial del Banco Mundial]. https://www.bancomundial.org/es/news/press-release/2024/12/03/informe-sobre-pobreza-del-banco-mundial-destaca-desigualdades-persistentes-en-colombia

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7376 LNAI*, 154-168. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84864950013&doi=10.1007%2f978-3-642-31537-4_13&partnerID=40&md5=3971cf8aaeb75b64a6b8ca8fc69a20ec

Shen, C. Stock forecasts based on the XGBoost model. En: 2025, 1198-1202. https://doi.org/10.1145/3708036.3708234

Tansuchat, R., Rakpho, P., & Klinlampu, C. (2023). Variable Selection Methods-Based Analysis of Macroeconomic Factors for an Enhanced GDP Forecasting: A Case Study of Thailand. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *14375 LNAI*, 213-223. https://doi.org/10.1007/978-3-031-46775-2_19

Zhang, Q., Nizovksy, D., Zeng, T. H., & Shalaginov, M. Y. Predicting Poverty in the Us Using Machine Learning on Demographic and Socioeconomic Data. En: 2025. https://doi.org/10.1109/ICAD65464.2025.11114072