PS2 - Equipo 03

Catalina Leal Rojas, Lucas Daniel Carrillo Aguirre, Lucas Eduardo Veras Costa

19 de octubre de 2025

El link del repositorio es: https://github.com/lucaseduardoveras/PS2_Equipo3

1. Introducción

Uno de los ejercicios fundamentales en la formulación de política pública consiste en identificar las características socioeconómicas de los individuos y comprender la naturaleza de dichas condiciones. Cada vez que se planea un proyecto de inversión, ya sea público o privado, resulta indispensable realizar una clasificación poblacional basada en indicadores de calidad de vida. Este proceso permite establecer una partición social entre hogares pobres y no pobres —o, de manera más amplia, entre grupos vulnerables y no vulnerables—, lo cual facilita al hacedor de política reconocer los factores más determinantes para cada grupo y, con base en ello, diseñar intervenciones más efectivas orientadas a mejorar su bienestar.

En Colombia, el instrumento de identificación de pobreza más importante es el Sisbén, o Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales, y está administrado por el Departamento Nacional de Planeación, DNP. Este sistema utiliza como insumo encuestas nacionales a hogares y construye una clasificación de cuatro grupos (A a D) a partir de la estimación de su capacidad de generar ingresos y sus características socioeconómicas y demográficas (**?**).

Sin embargo, como todos los instrumentos de medición de la pobreza, el Sisbén ha tenido problemas relacionados con la fiabilidad de sus estimaciones sobre la población vulnerable. Una de las limitaciones de las primeras versiones del instrumento, de hecho, es que los hogares en condiciones relativamente mejores se beneficiaban más de la respuesta estratégica (es decir, la subestimación de sus condiciones materiales) haciendo que la herramienta fuese de cierta forma regresiva (**?**). Esto implica que el problema a resolver sea cómo predecir de forma adecuada la pobreza por medio de nuevas herramientas de medición y, consecuentemente, responder cuál es el modelo de clasificación más pertinente para hacerlo.

En los últimos años, la literatura ha avanzado significativamente en el uso de técnicas de aprendizaje automático y datos no tradicionales para la predicción de la pobreza, con el fin de complementar las mediciones convencionales basadas en encuestas de ingresos y consumo. **?** muestran que modelos de *machine learning* aplicados a datos satelitales y geoespaciales pueden explicar hasta el 60 % de la variación del consumo en áreas donde la información censal es limitada, evidenciando el potencial de estas metodologías para generar estimaciones más oportunas y de bajo costo. De manera complementaria, **?** utilizan enfoques de *microsimulación* y *nowcasting* para predecir tasas de pobreza en América Central, Panamá y República Dominicana, resaltando la importancia de incorporar información de alta frecuencia y validar cuidadosamente los modelos fuera de muestra. Por su parte, **?** enfatizan la necesidad de aprovechar el *big data* y la analítica predictiva para reducir los costos de medición en América Latina y el Caribe, promoviendo la integración de fuentes tradicionales con nuevas tecnologías y conocimiento local. En conjunto, estos estudios coinciden en que las metodologías predictivas pueden mejorar la focalización de

programas sociales y el monitoreo del bienestar, siempre que se acompañen de rigurosas estrategias de validación y transparencia en su uso para la política pública.

Para el presente trabajo fue utilizado como insumo el Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE, (**?**), que contiene datos desagregados a nivel de individuo y de hogar sobre el nivel de ingreso, el nivel educativo, la afiliación a la seguridad social, las características demográficas de los hogares y el tipo de actividad laboral. Estas variables fueron fundamentales para la construcción de un modelo que pudiera estimar de forma veraz la condición de pobreza, especialmente porque presentan información complementaria a la expuesta por el Sisbén, y porque cuenta con un respaldo metodológico robusto por cuenta del DANE.

En cuanto a los modelos de clasificación, seis fueron aplicados a las bases de datos expuestas: Regresión Lineal, Regresión Logística (Logit), Elastic Net, Random Forest, Gradient Boosting y Naive Bayes. De ellos, aquellos que tuvieron el mejor rendimiento en la competencia propuesta en Kaggle, fueron los dos planteados con Random Forest, con un puntaje de 0.63 cada uno. Además, las variables importantes en ambos modelos fueron las características demográficas del hogar, características de inserción laboral y el tipo de vivienda.

Los resultados evidencian que las dimensiones laborales, demográficas y habitacionales -en particular la inserción en el mercado laboral y la composición del hogar- explican de manera sustanical la condición de pobreza, ampliando enfoques anteriores. Desde una perspectiva metodológica, se halló que el desempeño de los modelos de Random Forest es superior a los enfoques lineales o de boosting por cuenta de su robustez frente a la colinealidad y el ruido, además de su capacidad de capturar interacciones no lineales y de calibrar de forma óptima el umbral de decisión.

2. Datos

Los datos empleados en este análisis provienen de la *Medición de Pobreza Monetaria y Desigualdad 2018*, elaborada por el *Departamento Administrativo Nacional de Estadística (DANE)* en el marco de la misión para el *Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE)*. Esta iniciativa surgió con el propósito de unificar y armonizar las series de empleo, pobreza y desigualdad, fortaleciendo así la consistencia y credibilidad de los indicadores sobre calidad de vida en Colombia (**?**). Dada su amplitud y nivel de desagregación, esta fuente resulta especialmente adecuada para el objetivo de predecir la condición de pobreza de los hogares, pues contiene información socioeconómica detallada tanto a nivel individual como del hogar, incluyendo variables sobre ingresos, educación, ocupación, características de la vivienda y composición familiar. Para los fines de este trabajo, los datos fueron obtenidos a través de la plataforma *Kaggle*, donde se desarrolla una competencia orientada a la predicción de pobreza. El conjunto se encuentra estructurado en

cuatro archivos: a nivel de hogar e individual, y divididos en muestras de *training* y *testing*.

En cuanto a la limpieza de datos, las bases de hogares y personas se integraron utilizando el identificador único `id` incorporado en los archivos de *Kaggle*. Previo a la construcción de las variables de interés para la predicción, se excluyeron aquellas con una proporción de *missings* superior al 40 %, umbral a partir del cual no se consideró pertinente realizar imputaciones. Las variables restantes fueron seleccionadas con base en un enfoque analítico y en el criterio del equipo de trabajo, priorizando su relevancia teórica y empírica para el fenómeno de estudio. Estas decisiones fueron posteriormente validadas mediante las estadísticas descriptivas que se presentan más adelante en este documento. Finalmente, las variables de ingreso del conjunto de *training* no fueron incluidas, dado que no se encontraban disponibles en el *testing set*, lo que impedía su utilización en la etapa de predicción.

Las 27 variables utilizadas se presentan en el Cuadro 1. Las variables provenientes de la base de personas corresponden a la suma de los individuos en el hogar que cumplen determinada condición, así como a características exclusivas del jefe o jefa del hogar. Por su parte, de la base de hogares se seleccionaron variables que describen las condiciones de la vivienda.

Con relación a las variables relacionadas con el mercado laboral, se construyeron tasas de ocupación e inactividad, bajo la hipótesis de que una mayor participación laboral reduce la probabilidad de pobreza, mientras que una mayor proporción de personas fuera del mercado de trabajo la incrementa. Estas tasas se definieron como el número de personas ocupadas o inactivas sobre el total de personas en edad de trabajar en el hogar. Asimismo, se tuvieron en cuenta las horas totales trabajadas por los miembros del hogar, el número de trabajadores cuenta propia y si el jefe del hogar es cuenta propia. Estos indicadores permiten aproximar la disponibilidad de recursos laborales y la resiliencia del hogar frente a choques económicos.

Respecto a las características socioeconómicas, se incluyeron tanto el número de mujeres como una variable dicotómica que identifica si el jefe o la jefa del hogar es mujer, con el fin de capturar posibles diferencias estructurales de género en la incidencia de la pobreza. Paralelamente, se calcularon indicadores demográficos como la proporción de menores de edad y la tasa de dependencia, definida como la relación entre el número de personas menores de 15 y mayores de 65 años respecto a la población en edad de trabajar. Para esta última, se espera que una mayor proporción de dependientes afecte negativamente la capacidad económica del hogar, al limitar la disponibilidad de miembros que puedan participar en el trabajo remunerado. Igualmente, otra dimensión que también es relevante para predecir la pobreza de un hogar es el nivel educativo de sus integrantes. Se incluyó la proporción de personas con educación superior dentro del hogar, así como el nivel educativo máximo alcanzado por sus miembros. Estas medidas buscan captar el efecto positivo del capital humano sobre la generación de ingresos y la movilidad socioeconómica.

En cuanto a las variables disponibles a nivel de hogar, se consolidó un conjunto de indicadores que permite capturar las condiciones habitacionales y estructurales del entorno doméstico, complementando la información demográfica y laboral previamente agregada desde la base individual. En primer lugar, se incorporó la variable *número de habitaciones*, que permite aproximar el tamaño y la capacidad física de la vivienda. Este indicador es relevante, ya que un menor número de habitaciones por hogar suele asociarse con condiciones de hacinamiento, las cuales representan un componente clave del bienestar material y un reflejo de carencias estructurales. Asimismo, se incluyó la variable *tipo de vivienda*, que distingue entre viviendas propias, arrendadas, en usufructo u ocupadas sin título. Esta variable captura el grado de seguridad en la tenencia y constituye un indicador indirecto de la estabilidad económica y patrimonial del hogar. Se espera que los hogares con vivienda propia —especialmente aquellos que la han adquirido totalmente— presenten menor probabilidad de pobreza en comparación con los hogares en arriendo o en posesión informal.

También se consideró la variable *número de personas por hogar*, que actúa como un control demográfico básico y refleja la carga familiar total. Hogares más numerosos tienden a enfrentar mayores presiones sobre los recursos disponibles, especialmente cuando el número de dependientes es elevado. Además, se mantuvo la variable *Dominio*, que identifica el dominio geográfico de la encuesta y permite capturar diferencias territoriales y regionales en las condiciones socioeconómicas. Su inclusión es relevante, dado que las dinámicas del mercado laboral, los niveles de precios y la disponibilidad de infraestructura difieren sustancialmente entre las regiones del país.

En conjunto, estas variables —de carácter habitacional, demográfico y geográfico— complementan las dimensiones laborales, sociales y educativas incluidas desde la base a nivel de personas y aportan información relevante sobre los determinantes estructurales de la pobreza a nivel de hogar. Su inclusión en el modelo busca capturar tanto las restricciones de capital físico como las condiciones del entorno residencial, las cuales inciden directamente en el bienestar y la vulnerabilidad económica de los hogares. De este modo, se espera mejorar la capacidad predictiva del modelo al aplicar los datos de prueba.

Pasando al análisis de los datos empleados, es importante considerar el desbalance de clases presente en el conjunto de entrenamiento. Aproximadamente el 80 % de los hogares son no pobres en el conjunto de datos de muestreo, un aspecto que será examinado con mayor detalle en la sección siguiente. El conjunto de hogares en el conjunto de datos de entrenamiento posee 164.959 observaciones mientras que el conjunto de datos de prueba posee 66.168 observaciones, alrededor de 40 % del tamaño del conjunto de prueba.

El Cuadro 3 presenta las estadísticas descriptivas (media, mediana y desviación estándar) para los hogares pobres y no pobres en el conjunto de entrenamiento, mientras que el Cuadro 2 presenta los resultados de la diferencia de medias para las variables numéricas. Esta última confirma que nuestra selección de variables es adecuada, pues todas las diferencias entre ambos grupos son estadísticamente significativas.

Estos resultados permiten observar que, en promedio, los hogares clasificados como pobres registran mayores tasas de inactividad y de trabajo por cuenta propia, así como menores horas trabajadas —tanto en total como por parte del jefe o jefa del hogar— y un menor número de personas ocupadas. En conjunto, estos patrones sugieren una inserción laboral más precaria y una menor capacidad del hogar para generar ingresos estables, lo que refuerza la relación entre la vulnerabilidad en el mercado de trabajo y la incidencia de la pobreza. La mayor prevalencia del trabajo por cuenta propia podría reflejar estrategias de subsistencia ante la falta de oportunidades formales, mientras que las elevadas tasas de inactividad evidencian barreras estructurales de acceso al empleo, especialmente en hogares con alta dependencia económica. En el plano socioeconómico y demográfico, los hogares pobres presentan en promedio más integrantes, mayor número de menores de edad, y jefes de hogar más jóvenes, lo que refleja una mayor carga de dependencia y menor acumulación de capital humano. En contraste, los hogares no pobres exhiben niveles educativos más altos, una mayor proporción de miembros con educación superior y tasas de ocupación más elevadas, factores coherentes con una menor vulnerabilidad económica.

El Cuadro 4 presenta las medias de las variables categóricas empleadas en las predicciones. Se observa que la proporción de hogares encabezados por mujeres es superior entre los hogares pobres, lo que sugiere posibles desigualdades de género en el acceso a recursos económicos. Asimismo, la proporción de jefes de hogar ocupados es significativamente mayor entre los hogares no pobres, lo que refleja la estrecha relación entre inserción laboral y condiciones de vida. En cuanto a las condiciones habitacionales, se evidencia que las familias no pobres residen en mayor proporción en viviendas propias y totalmente pagadas, mientras que los hogares pobres presentan una concentración más alta en viviendas en arriendo o en condiciones de tenencia precaria, como la posesión sin título. Finalmente, respecto al nivel educativo, tanto del jefe del hogar como del conjunto de sus miembros, se aprecia un patrón claro: los hogares no pobres exhiben una proporción considerablemente mayor de educación universitaria, mientras que los hogares pobres se concentran en niveles educativos básicos, especialmente primaria y media. Este contraste refuerza la hipótesis de que el capital humano constituye un determinante fundamental de la pobreza en el país.

3. Modelos y Resultados

En esta sección describimos los modelos utilizados para predecir la pobreza a nivel de hogar y el modelo que alcanzó el mejor desempeño. Las estimaciones se realizaron empleando los siguientes algoritmos de clasificación: **Modelo de probabilidad lineal**, **Regresión Logit**, **Naive Bayes**, **Random Forest**, **CART**, **Elastic Net** y **Boosting**.

Para cada modelo, con excepción del modelo lineal y de la regresión Logit, se calibraron los *hiperparámetros* mediante validación cruzada de cinco pliegues (*five-fold cross-*

validation). Dado que la métrica utilizada para evaluar las predicciones en la competencia es la **estadística F**, optamos por emplearla también como criterio de desempeño durante la calibración. Por otro lado, todos los modelos -menos el de probabilidad lineal- incluyeron exactamente las 26 variables que se describieron y analizaron en la sección de datos.

La **estadística F**, o *F1-score*, se define como la media armónica entre la *precisión* (*precision*) y la *sensibilidad* (*recall*), cuya fórmula es la siguiente:

$$F_1 = 2 \cdot \frac{\text{Precisión} \cdot \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}} \tag{1}$$

donde:

$$\text{Precisión} = \frac{TP}{TP + FP}, \qquad \text{Sensibilidad} = \frac{TP}{TP + FN} \tag{2}$$

y TP, FP y FN representan, respectivamente, los verdaderos positivos, falsos positivos y falsos negativos.

Cabe señalar que los datos de entrenamiento presentan un marcado **desbalance de clases**, lo cual tiende a sesgar los modelos hacia la clase más frecuente, afectando negativamente la capacidad de predicción sobre la clase minoritaria —en este caso, los hogares pobres. Para mitigar este problema, se aplicaron *técnicas de remuestreo* (tanto *oversampling* como *undersampling*) en algunos de los modelos. Adicionalmente, el umbral de decisión asociado a la probabilidad estimada p, que por convención se fija en 0.5, fue recalibrado para maximizar la métrica F_1 o el área bajo la curva ROC (AUC) en cada modelo. Para ello, se emplearon dos estrategias a partir de las predicciones del *training set* generadas en cada uno de los pliegues (*folds*) de la validación cruzada: en primer lugar, se construyó una grilla de posibles umbrales de clasificación entre $p = 0{,}05$ y $p = 0{,}95$, seleccionando aquel que maximizara el valor de F_1; en segundo lugar, a partir de la curva ROC, se identificó el umbral que maximizara la proporción de verdaderos positivos y minimizara los falsos negativos. De esta manera, se buscó reducir el sesgo hacia la clase mayoritaria y mejorar la capacidad discriminante de los modelos.

3.1. Modelos utilizados en las Predicciones

Los modelos más sencillos utilizados en la estimación fueron el modelo de**probabilidad lineal** y el modelo **logit**. Para el primero se incluyeron todas las variables de la base más algunos efectos no lineales e interacciones:

$$\begin{aligned} Pobre_i = \beta_0 + \mathbf{X}_i'\boldsymbol{\beta} + \beta_1 Nper_i^2 + \beta_2 edad\_head_i^2 + \boldsymbol{\beta_3}(bin\_headWoman_i \times cat\_educHead_i) \\ + \boldsymbol{\beta_4}(edad\_head_i \times cat\_educHead_i) + \beta_5(Nper_i \times t\_dependencia_i) + \varepsilon_i \end{aligned} \tag{3}$$

donde $\mathbf{X}_i'\boldsymbol{\beta}$ representa los coeficientes asociados a las 26 variables de nuestra base. Dada

la baja complejidad del modelo, buscamos mejorar su capacidad predictiva mediante las 5 variables adicionales. Estas buscan capturar las no linealidades del número de personas en el hogar, el pico de ingresos respecto a la edad y las heterogeneidades producto del nivel educativo.

Para el modelo **Logit** mantuvimos solamente las 26 variables seleccionadas. Este modelo no requiere la definición de hiperparámetros y, en general, busca maximizar la siguiente función de verosimilitud:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \left[\pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i} \right],$$

donde $\pi(x_i) = \frac{1}{1+e^{-x_i'\beta}}$ representa la probabilidad estimada de que el resultado sea igual a uno dado el vector de covariables x_i. En otras palabras, el modelo logit relaciona linealmente los predictores con el logaritmo de las probabilidades relativas del evento de interés. Adicionalmente, para ambos modelos se empleó la estrategia de calibración del punto de corte con la AUC descrita anteriormente.

Otro modelo que utilizamos fue el **Elastic Net**. Este modelo combina las propiedades de los métodos **LASSO** y **Ridge**, introduciendo una regularización sobre los coeficientes de las variables para prevenir el sobreajuste y mejorar la capacidad predictiva. En términos formales, el Elastic Net busca minimizar la siguiente función objetivo:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \left[(1-\alpha)\frac{1}{2}\|\beta\|_2^2 + \alpha\|\beta\|_1 \right] \right\},$$

donde λ controla la intensidad de la penalización total y $\alpha \in [0,1]$ determina el grado de mezcla entre las penalizaciones ℓ_1 (LASSO) y ℓ_2 (Ridge). Valores de α cercanos a 1 se asemejan a LASSO, mientras que valores cercanos a 0 se aproximan a Ridge. Para este modelo , realizamos 2 intentos, uno que mezcla ambos tipos de penalizaciones y otro solamente con penalización tipo LASSO. Al igual que los primeros dos modelos, elegimos el punto de corte basado en la AUC (ROC).

El modelo *Naive Bayes* se basa en el teorema de Bayes, asumiendo independencia condicional entre las variables explicativas dado el estado de la variable objetivo. Formalmente, la probabilidad de pertenecer a la clase $y \in \{0,1\}$ se expresa como:

$$P(Y = y \mid X = x) \propto P(Y = y) \prod_{j=1}^{p} P(X_j = x_j \mid Y = y)$$

y la clase predicha corresponde a $\hat{y} = \arg\max_y P(Y = y \mid X = x)$.

En este modelo se calibraron tres hiperparámetros: `usekernel`, que define si las va-

riables continuas se modelan mediante distribuciones gaussianas o con densidades kernel; `adjust`, que controla la suavidad del kernel; y `laplace`, que aplica un suavizado de Laplace para evitar probabilidades nulas. Para este modelo, decidimos probar el rendimiento de dos tipos de remuestreo que buscan tratar el problema de desbalance de clases. Los resultados se presentan en la Figura 1 y estos comprueban que emplear estas técnicas mejora sustancialmente la sensiblidad del modelo. Asimismo, ajustamos el punto de corte de la clasificación con la técnica de maximización del F1 descrita anteriormente, nuevamente, para probar otras soluciones a través de nuestras estimaciones.

Los últimos 2 métodos empleados corresponden a modelos basados en **árboles**. Implementamos **Random Forest**, el cual es un método de ensamble que combina múltiples árboles de decisión para mejorar la capacidad predictiva y reducir la varianza del modelo individual. Cada árbol se entrena sobre una muestra distinta del conjunto de datos, y las observaciones no incluidas en cada muestra —las llamadas Out-of-Bag (OOB)— se utilizan para estimar el error de generalización del modelo sin necesidad de un conjunto de validación adicional.

$$\hat{y}(x) = \text{mode}\{\, T_b(x) \,\}_{b=1}^{B}$$

donde cada $T_b(x)$ es un árbol de decisión construido a partir de una muestra de *bootstrap* y usando un subconjunto aleatorio de predictores en cada división. Los hiperparámetros que elegimos mediante validación cruzada fueron el número de variables seleccionadas aleatoriamente en cada división (controla la correlación entre árboles), el tamaño mínimo de los nodos terminales y el criterio de división. Teniendo en cuenta los resultados obtenidos con Naive Bayes, en este modelo decidimos tratar el desbalance con *downsampling* y, al igual que el resto de modelos, con la calibración de un punto de corte más óptimo.

Finalmente, implementamos un modelo **Gradient Boosting**. En este caso se construye el modelo de manera aditiva y secuencial, combinando múltiples árboles débiles (de poca profundidad) para corregir los errores cometidos por los anteriores. Este proceso iterativo se puede resumir como:

$$\hat{\Theta}_m = \arg\min_{\Theta_m} \left\{ \sum_{i=1}^{n} L\big(y_i, f_{m-1}(x_i) + T(x_i; \Theta_m)\big) \right\}$$

La particularidad de esta técnica para *Boosting*, es que resuelve utilizando gradiente descendente, lo que permite estimar para funciones de pérdida no lineales y mayor escalabilidad. Para el desbalance se ajustó el umbral de clasificación con la AUC (ROC) y los parámetros elegidos por validación cruzada fueron el número de árboles, la profundidad de los árboles, la tasa de aprendizaje y el número mínimo de observaciones por nodo.

En el cuadro 5 se resumen las grillas, hiperparametros y umbrales para cada modelo. En la siguiente sección se profundizará en la elección y consecuencias de nuestros hiper-

parámetros.

3.2. Mejor modelo de predicción (F1-score: 0.63)

El modelo con mejor rendimiento entre los presentados fue el *Random Forest*, estimado en dos versiones. La primera corresponde a un modelo calibrado mediante validación cruzada (*5-fold*), con una búsqueda del número de variables seleccionadas en cada partición (*mtry*) entre 6, 8, 9 y 10, y un entrenamiento de 1,000 árboles, utilizando el índice de Gini para determinar la importancia de las variables. La segunda versión corresponde a un *Random Forest* tipo *Bagging*, es decir, con el número de variables igual al total disponible (26), 500 árboles y la estimación de la importancia por permutación.

En ambos casos se exploraron distintas estrategias de ajuste de hiperparámetros. En el primer modelo, se fijó el número mínimo de observaciones por nodo en uno, con el objetivo de permitir árboles altamente complejos y reducir el sesgo de predicción. En contraste, el modelo tipo *Bagging* permitió seleccionar este parámetro de manera automática e incorporó un mecanismo adicional para manejar el desbalance de clases: la elección entre los criterios de Gini y Hellinger. Este último no evalúa la pureza del nodo, sino la distancia entre las distribuciones de las clases, lo que favorece una mejor identificación de la clase minoritaria. En consecuencia, se confirma la relevancia de tratar explícitamente el problema del desbalance en la predicción de pobreza, dado que el criterio de Hellinger fue el seleccionado y logró compensar la menor complejidad estructural del modelo —con un número menor de árboles (500 frente a 1,000) y sin búsqueda exhaustiva del parámetro *mtry*— obteniendo el mismo puntaje final ($F_1 = 0{,}63$).

Nuevamente, se destaca el uso del F_1 dentro de la muestra de entrenamiento para determinar el umbral de clasificación, que fue de 0.522 para el modelo con Hellinger y de 0.340 para el calibrado con Gini. Estas diferencias sugieren que el umbral óptimo se ajustó de manera más conservadora en el primer caso, privilegiando la identificación de los hogares pobres y confirmando que la calibración conjunta del punto de corte y la métrica de evaluación fue un factor decisivo en el alto desempeño de ambos modelos.

A continuación se presenta la comparación entre este modelo y los cinco modelos mejor calibrados por debajo de él.

RF vs. Lineal. A pesar de que en este modelo fueron seleccionadas interacciones no lineales entre las variables de género y educación, es claro que por el volumen de predictores existentes en la base de datos una agregación manual de las interacciones es insuficiente. El Random Forest lo hace automáticamente. Además, el Random Forest estuvo calibrado a F1 que se alinea con los parámetros de la competencia de Kaggle, mientras que el modelo lineal permanece limitado por su frontera casi lineal. No obstante, cabe resaltar que, siendo un modelo tan sencillo, este superó a modelos teóricamente más complejos. Esto nos da

indicios de la buena selección de las variables que se hizo desde el análisis descriptivo y de la importancia de establecer un punto de corte distinto a 0.5 cuando las clases están desbalanceadas.

RF vs. Logit. El modelo de Random Forest superó al modelo Logit principalmente porque la relación entre las variables explicativas y la pobreza, como se ha expuesto antes, no es necesariamente lineal u homogénea. El modelo Logit impone una frontera de decisión lineal, lo que limita su capacidad de capturar interacciones complejas que el RF hace profundamente. En este sentido, el RF construye múltiples árboles de decisión que dividen el espacio en subregiones que permiten modelar umbrales más flexibles. Por último, no seleccionar los hiperparámetros impide que utilizando el Logit se decida la capacidad de ajuste del modelo acorde a la métrica que evalua la competencia. Al igual que el caso anterior, incluir buenos predictores y buscar soluciones para el desbalance de clases logró que un modelo sencillo superará otras metodologías.

RF vs. Elastic Net Si bien el modelo EN introduce regularización a través de los parámetros λ y α
-que justamente permite reducir la varianza y evita el sobreajuste-, su capacidad de ajuste sigue siendo lineal en los predictores, en contraposición a la naturaleza no lineal del RF. Además, mientras que las penalizaciones de Elastic Net controla la magnitud de los coeficientes, y por ende facilitando interpretabilidad, también puede reducir la contribución a la predicción de las variables altamente correlacionadas como las demográficas y laborales.

RF con punto de corte calibrado con F1 vs. RF con punto de corte calibrado con ROC. Como se mencionó al inicio, se experimentó con ambas estrategias para abordar el desbalance de clases. Tal como se muestra en el Cuadro **??**, no hubo cambios en los hiperparámetros entre estas dos alternativas; sin embargo, el cambio de criterio para la calibración del punto de corte generó una leve disminución en el puntaje obtenido. Esto sugiere que resulta ventajoso no solo utilizar la métrica F_1 durante el ajuste de los hiperparámetros en la validación cruzada, sino también al momento de definir el umbral de clasificación, dado que esta métrica está directamente alineada con la función de evaluación empleada en la competencia de *Kaggle*.

RF vs. Gradient Boosting. El *Gradient Boosting* también capturó no linealidades e interacciones mediante la suma secuencial de árboles débiles, pero su desempeño resultó inferior al de los modelos *Random Forest* por varias razones. En primer lugar, la combinación de una tasa de aprendizaje reducida (*shrinkage*) entre 0.01 y 0.001 —valores comúnmente empleados en la literatura— y un número máximo de 150 árboles dio lugar a un modelo subentrenado, limitando su capacidad para refinar los errores de etapas previas. En segundo lugar, el modelo de *Gradient Boosting* no incorporó pesos de clase, lo que condujo a una mayor inclinación hacia la clase mayoritaria. Tanto la restricción en el número de árboles como la ausencia de métodos de remuestreo respondieron a limitaciones computacionales del equipo. Finalmente, el *Gradient Boosting* es más sensible a la especificación de los hi-

perparámetros y, en ausencia de suficientes iteraciones, tiende a una menor capacidad de generalización. En contraste, los modelos de *Random Forest* combinaron un mayor número de árboles y aplicaron submuestreo de variables en cada división, lo que permitió reducir la varianza sin incurrir en *overfitting*, un problema al que los métodos de *Boosting* son especialmente susceptibles.

Finalmente, este ejercicio nos permite establecer cuáles son las variables más importantes a la hora de predecir la pobreza a nivel de hogar en Colombia. Como se trata de modelos Random Forest, se puede calcular la importancia de una variable permutando sus valores y comparando el error *out-of-bag*. Si este error aumenta significa que la variable es importante y es necesaria para mejorar el desempeño del modelo. En las Figuras 2 y 3 se presentan las 20 variables más importantes para el Random Forest estándar y el Random Forest tipo Bagging, respectivamente.

Si bien con cada modelo varía el ranking de importancia para las variables, se pueden notar ciertos patrones que ratifican a algunas variables como buenos predictores de la variable de interés. Los resultados sugieren que los factores laborales y demográficos desempeñan un papel central en la explicación de la pobreza de los hogares. En ambos modelos, destacan entre las variables más influyentes el número total de horas trabajadas por los miembros del hogar, la edad del jefe o jefa del hogar, la presencia de menores de edad y la tasa de ocupación. Estas variables reflejan la relevancia del acceso al mercado laboral y la estructura etaria como determinantes directos del bienestar económico.

En particular, un mayor volumen de trabajo remunerado dentro del hogar tiende a reducir la probabilidad de pobreza, mientras que una elevada proporción de menores implica una mayor carga de dependencia que restringe la capacidad productiva del hogar. La edad del jefe del hogar también aparece como un predictor relevante, lo que sugiere la existencia de un ciclo de ingresos vinculado al momento del ciclo vital: los hogares encabezados por personas en edades intermedias suelen presentar mayor estabilidad económica. Asimismo, la importancia de variables como el tipo de vivienda, la tasa de educación superior y el dominio geográfico resalta la influencia combinada de las condiciones estructurales y territoriales en la generación de pobreza. En conjunto, los resultados confirman que las carencias de capital humano y la inserción laboral precaria son factores decisivos en la explicación de la pobreza monetaria en Colombia.

4. Conclusiones

En conjunto, los resultados confirman que es posible mejorar la identificación de la pobreza integrando información socioeconómica detallada con modelos de aprendizaje automático, como son los modelos de Random Forest. Frente a instrumentos tradicionalmente utilizados -que pueden enfrentar sesgos operativos o estratégicos-, los enfoques predictivos

permiten capturar patrones no lineales, umbrales y heterogeneidades de las variables relevantes para las categorizaciones de los individuos. En nuestro caso, la evidencia muestra que la inserción laboral del hogar, la estructura demográfica (dependencia, presencia de menores en el hogar y la edad del jefe o jefa del hogar) y las condiciones habitacionales concentran gran parte del poder explicativo. Esto es una expansión teórica y metodológica de las primeras aproximaciones de instrumentos como el Sisbén, que concebían la caracterización de la pobreza como un resultado de las capacidades de generación de ingreso; en este sentido se permite entender la pobreza y vulnerabilidad como condiciones más integrales.

Metodológicamente, el mejor desempeño de los modelos Random Forest frente a alternativas lineales o de boosting se explica por su robustez a la colinealidad y al ruido, a su capacidad para modelar interacciones complejas entre las variables y a una calibración explícita del umbral de decisión. Esto podría implicar que, cuando la decisión de política pública depende de distinguir de manera detallada la minoría vulnerable, lo más conveniente es priorizar modelos que reduzcan la varianza, exploren subespacios de predictores y ajusten el umbral a la métrica de interés. Esto, en suma, puede implicar una enseñanza para el ejercicio público: los modelos de aprendizaje pueden complementar las estrategias metodológicas actualmente utilizadas, sobre todo si se acompañan de validaciones fuera de muestra, y se realizan monitoreos de sesgo en los modelos.

5. Anexos

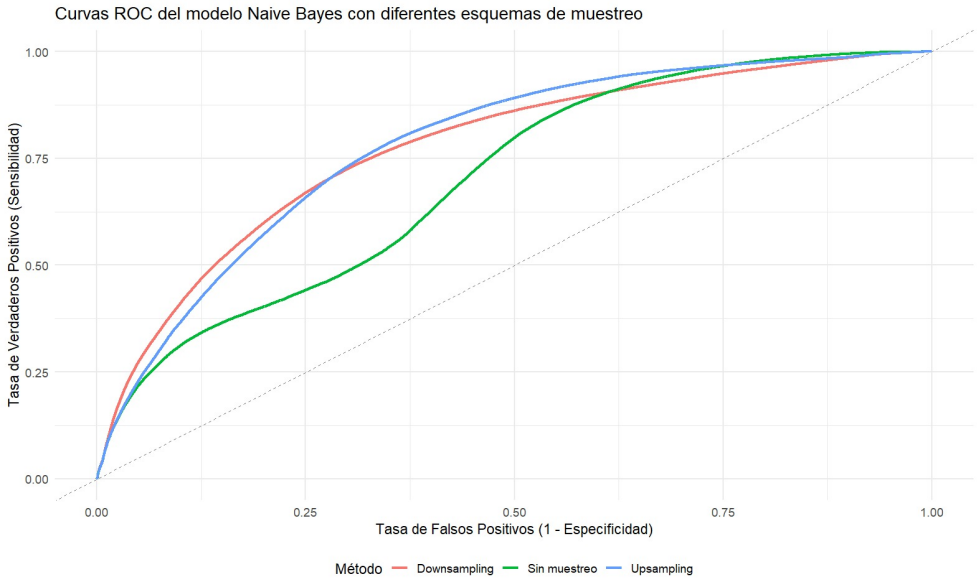
Cuadro 1. Descripción y tipo de las variables utilizadas

Variable	**Tipo**	**Descripción**
Dominio	Categórica	Cada una de las 24 áreas metropolitanas, resto de cabeceras y rural
Pobre	Binaria	1 si el hogar es pobre, 0 de lo contrario
Nper	Numérica	Número de personas en el hogar
n_habitaciones	Numérica	Número de habitaciones en la vivienda
tipo_vivienda	Categórica	Tipo de vivienda
bin_headWoman	Binaria	1 si el jefe del hogar es mujer, 0 de lo contrario
cat_educHead	Categórica	Nivel de educación del jefe del hogar
bin_headSS	Binaria	1 si el jefe del hogar cuenta con seguridad social, 0 de lo contrario
bin_occupiedHead	Binaria	1 si el jefe del hogar está ocupado, 0 de lo contrario
edad_head	Numérica	Edad del jefe del hogar
bin_headCpropia	Binaria	1 si el jefe del hogar tiene casa propia, 0 de lo contrario
h_workedHead	Numérica	Horas trabajadas por el jefe del hogar en la última semana
num_women	Numérica	Número de mujeres en el hogar
num_minors	Numérica	Número de menores de 15 años en el hogar
num_mayores	Numérica	Número de mayores de 65 años en el hogar
cat_maxEduc	Categórica	Nivel máximo de educación en el hogar
num_educ_sup	Numérica	Número de personas con educación superior en el hogar
num_c_propia	Numérica	Número de trabajadores con cuenta propia en el hogar
num_h_trabajadas	Numérica	Número de horas trabajadas por los miembros del hogar
num_ocu	Numérica	Número de ocupados en el hogar
num_pet	Numérica	Número de personas en edad de trabajar en el hogar
num_inac	Numérica	Número de inactivos en el hogar
t_ocu	Numérica	Tasa de ocupación
t_inac	Numérica	Tasa de inactividad
t_educ_sup	Numérica	Tasa de personas con educación superior
t_dependencia	Numérica	Tasa de dependencia (proporción de menores de 15 y mayores de 65 años en el hogar)

Cuadro 2. Diferencia de medias entre hogares pobres y no pobres

Variable	Media Pobre	Media No Pobre	Diferencia	t	p-valor	Signif.
Nper	4.134	3.082	1.052	-87.48	0.0	***
n_habitaciones	3.032	3.48	-0.449	63.27	0.0	***
bin_headWoman	0.468	0.406	0.062	-20.19	0.0	***
bin_headSS	0.909	0.949	-0.04	23.6	0.0	***
bin_occupiedHead	0.643	0.727	-0.084	29.04	0.0	***
edad_head	46.774	50.323	-3.548	35.47	0.0	***
bin_headCpropia	0.473	0.322	0.151	-49.84	0.0	***
h_workedHead	28.621	34.493	-5.872	37.99	0.0	***
num_women	2.236	1.616	0.619	-77.29	0.0	***
num_minors	1.546	0.622	0.924	-119.05	0.0	***
num_mayores	0.275	0.333	-0.058	16.36	0.0	***
num_educ_sup	0.357	0.879	-0.522	111.62	0.0	***
num_c_propia	0.86	0.652	0.208	-40.07	0.0	***
num_h_trabajadas	52.459	71.114	-18.655	67.48	0.0	***
num_ocu	1.257	1.566	-0.309	50.26	0.0	***
num_pet	3.021	2.637	0.384	-41.68	0.0	***
num_inac	1.446	0.923	0.523	-75.06	0.0	***
t_ocu	0.426	0.621	-0.195	101.69	0.0	***
t_inac	0.461	0.327	0.134	-69.49	0.0	***
t_educ_sup	0.127	0.348	-0.222	126.7	0.0	***
t_dependencia	67.415	37.837	29.578	-95.26	0.0	***
t_c_propia	0.298	0.262	0.036	-19.16	0.0	***

Figura 1. ROC$_N aive$



Cuadro 3. Estadísticas descriptivas para las variables numéricas

Variable	**Media**		**Mediana**		**Desv. Est.**	
	Pobre	No Pobre	Pobre	No Pobre	Pobre	No Pobre
N. de Personas	4.13	3.08	4.00	3.00	2.03	1.64
N. de Habitaciones	3.03	3.48	3.00	3.00	1.13	1.25
Edad del Jefe	46.77	50.32	45.00	50.00	16.24	16.35
H. Trab. del Jefe	28.62	34.49	32.00	42.00	25.23	24.67
N. Mujeres	2.24	1.62	2.00	1.00	1.35	1.10
N. Menores	1.55	0.62	1.00	0.00	1.34	0.86
N. Mayores	0.27	0.33	0.00	0.00	0.57	0.61
N. Educ. Sup.	0.36	0.88	0.00	1.00	0.69	1.00
N. Cuenta Propia	0.86	0.65	1.00	0.00	0.85	0.80
Horas Trab. Total	52.46	71.11	48.00	60.00	43.68	49.59
N. Ocupados	1.26	1.57	1.00	1.00	0.99	1.03
N. Pet	3.02	2.64	3.00	2.00	1.53	1.35
N. Inactivos	1.45	0.92	1.00	1.00	1.17	0.96
Tasa Ocup.	0.43	0.62	0.50	0.67	0.31	0.34
Tasa Inac.	0.46	0.33	0.50	0.33	0.31	0.33
Tasa Educ. Sup.	0.13	0.35	0.00	0.25	0.25	0.38
Tasa Dependencia	67.42	37.84	60.00	33.33	52.86	39.45
Tasa Cuenta Propia	0.30	0.26	0.25	0.00	0.29	0.33

Cuadro 4. Promedios de variables categóricas por condición de pobreza

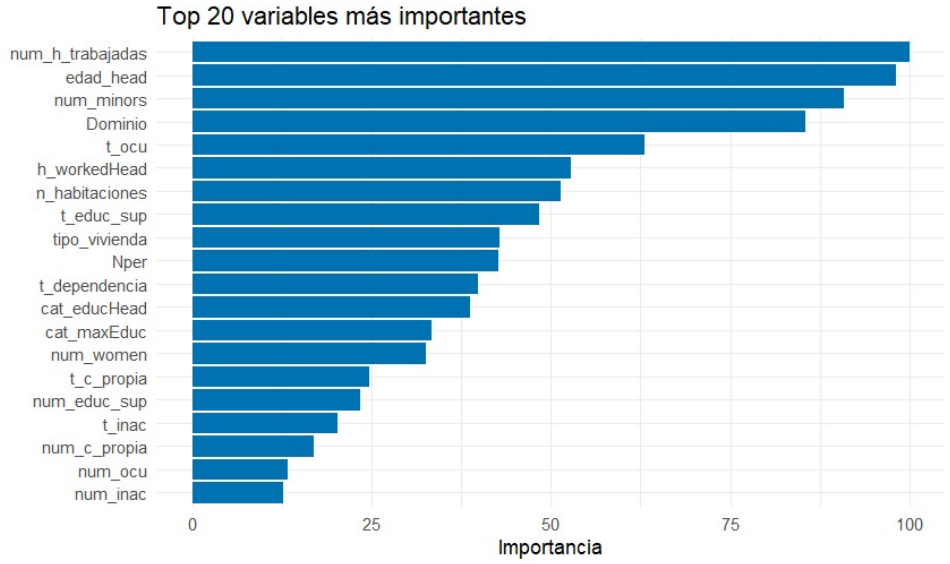
Variable	**Pobre**	**No Pobre**
Jefe del Hogar Mujer	0.468	0.406
Jefe del Hogar Ocupado	0.643	0.727
Jefe del Hogar Cuentra Propia	0.473	0.322
Tipo de vivienda		
Propia Totalmente Pagada	0.275	0.403
Propia la Están Pagando	0.017	0.038
En arriendo o subarriendo	0.438	0.379
En usufructo	0.158	0.149
posesión sin título	0.111	0.030
Nivel educativo del jefe del hogar		
Preescolas	0.0002	0.00003
Primaria	0.378	0.259
Medias	0.253	0.263
Secundaria	0.162	0.124
Universitaria	0.103	0.315
Máximo nivel educativo del hogar		
Preescolar	0.001	0.0001
Primaria	0.128	0.093
Media	0.385	0.264
Secundaria	0.201	0.087
Universitaria	0.261	0.544

Cuadro 5. Comparación de modelos y configuración de hiperparámetros

Modelo	**Grilla**	
Random Forest	node.size = [1, 5, 15, 50], split.rule = {Gini, Hellinger}	node.siz
Random Forest	num.vars = [6, 8, 9, 10]	no
Logit	NA	
Random Forest	num.vars = [6, 8, 9, 10]	no
Lineal	NA	
Elastic Net	α = (0,1) saltos de 0.1, $\lambda = (10^{-3}, 10^{3})$ diez valores	
Gradient Boosting	num.trees = [50, 100, 150], depth = [2, 5, 7], tasa.aprendizaje = [0.01, 0.001], num.node = [5, 10]	num.
Elastic Net	$\lambda = (10^{-3}, 10^{3})$, diez valores	
Naive Bayes (Down)	kernel = {TRUE, FALSE}, laplace = {0, 1}	

Figura 2. Importancia calculada con Random Forest (Gini)

Figura 3. Importancia calculada con Bagging (Permutación)



Referencias