*Taller: Problem set 2*

*Lina María Alvarez Ardila, Sara Rocio Rojas Gomez*


*Big Data y Machine Learning para economía aplicada*

*Profesor: Ignacio Sarmiento*

***Repositorio:***
***https://github.com/lialvareza1/Problem-Set-2---Lina-Alvarez-Ardila-y-Sara-Roc-o-Rojas.git***

## Introducción

La medición de la pobreza sigue siendo uno de los desafíos más relevantes para el diseño de políticas públicas en países en desarrollo como Colombia. La pobreza no solo refleja carencias económicas, sino también profundas desigualdades en el acceso a servicios básicos, oportunidades laborales y capital social. En este contexto, contar con herramientas eficientes y precisas para identificar hogares en situación de pobreza es clave para la asignación focalizada de recursos públicos y programas sociales.

Tradicionalmente, las metodologías empleadas para estimar la pobreza en Colombia han estado basadas en indicadores oficiales derivados de encuestas como la Gran Encuesta Integrada de Hogares (GEIH), administrada por el Departamento Administrativo Nacional de Estadística (DANE). Sin embargo, dichas metodologías suelen ser costosas, lentas y dependientes de operativos de campo extensos, lo cual limita su frecuencia de actualización y su capacidad para responder a cambios socioeconómicos en tiempo real.

En los últimos años, el uso de técnicas de aprendizaje automático (*machine learning*) ha emergido como una alternativa prometedora para la predicción de pobreza, debido a su capacidad para modelar relaciones complejas entre múltiples variables, y su versatilidad para ser aplicadas sobre diferentes tipos de datos. Diversos estudios recientes han explorado el potencial de estas técnicas en el caso colombiano. Por ejemplo, Guerrero (2021) utilizó modelos supervisados como XGBoost para identificar factores estructurales asociados a la pobreza, concluyendo que dimensiones como salud, empleo y educación son determinantes clave. Otros enfoques han integrado redes neuronales y modelos de clasificación profunda para predecir niveles de ingreso con base en datos no convencionales, como los extraídos de redes sociales (Patiño & Duque, 2021).

Asimismo, investigaciones como las de Arango (2023) y Guerrero et al. (2022) han propuesto metodologías híbridas que combinan datos estructurados y no estructurados para estimar pobreza multidimensional, expandiendo así las fronteras del monitoreo social mediante el uso de fuentes alternativas de información. En línea con esta tendencia, Martínez y Ramírez (2007) discuten críticamente los distintos indicadores utilizados para medir la pobreza en el país, argumentando que su efectividad depende tanto del contexto como del enfoque estadístico adoptado. Por su parte, Muñoz (2019) ha explorado la pobreza desde una perspectiva subjetiva, resaltando la importancia de variables no monetarias en la autopercepción del bienestar.

Este trabajo tiene como objetivo desarrollar un modelo predictivo de pobreza para los hogares colombianos, utilizando como base los datos provistos por la Gran Encuesta Integrada de Hogares (GEIH) del año 2018. Se trata de un problema de clasificación binaria, en el cual la variable objetivo indica si un hogar se encuentra en situación de pobreza de acuerdo con el ingreso per cápita imputado y la línea de pobreza correspondiente definida por el DANE. La riqueza de la GEIH permite trabajar con información a nivel de hogar e

individuo, lo cual posibilita la construcción de variables agregadas que reflejan dimensiones relevantes del bienestar, como la composición del hogar, el acceso a servicios, las condiciones laborales y el nivel educativo. Esta estructura de datos permite abordar el problema desde una perspectiva multidimensional, integrando tanto factores económicos como sociodemográficos en el proceso de predicción.
La relevancia de esta propuesta radica en su aplicabilidad práctica: si los modelos logran desempeños aceptables con un número reducido de variables, podrían servir como base para instrumentos de focalización de políticas públicas más livianos y económicos, especialmente en zonas con poca cobertura estadística o alta movilidad poblacional. Además, este ejercicio permite contribuir a la discusión académica sobre los límites y oportunidades del aprendizaje automático en contextos de política social, particularmente en países con limitaciones estructurales como Colombia.

## Caracterización y tratamiento de los datos

Para el desarrollo del modelo predictivo se utilizaron dos bases de datos proporcionadas por la Gran Encuesta Integrada de Hogares (GEIH) del año 2018: una a nivel de hogar (train_hogares.csv) y otra a nivel individual (train_personas.csv). La unidad de análisis final es el hogar, por lo cual fue necesario realizar una integración entre ambas fuentes de información utilizando la variable id, que identifica de forma única a cada hogar en ambas bases. La unión se efectuó a través de una operación tipo left join, utilizando como base principal la información contenida en la base de hogares. Previamente, se agruparon los datos de individuos para generar variables agregadas a nivel hogar como el porcentaje de personas inactivas en edad de trabajar, que posteriormente fueron incorporadas a la base original. Este procedimiento permitió enriquecer la base de datos con indicadores sociodemográficos relevantes para la predicción, manteniendo la cobertura completa de los hogares disponibles en la muestra original.

La base de hogares contiene variables clave como el ingreso total del hogar con imputación de arriendo (Ingtotugarr), el ingreso per cápita imputado (Ingpcug), el número de personas en la unidad de gasto (Npersug), y la línea de pobreza (Lp) determinada por el DANE, entre otras. A partir de estas variables, se construyó la variable dependiente binaria Pobre, la cual toma el valor de uno si el hogar presenta un ingreso per cápita por debajo de la línea de pobreza (Ingpcug < Lp), y cero en caso contrario. Como validación adicional, también se replicó este criterio utilizando el ingreso total ajustado y el tamaño del hogar (Ingtotugarr < Lp * Npersug), obteniendo resultados completamente consistentes.

Con el fin de enriquecer la información disponible a nivel de hogar, se construyeron nuevas variables agregadas utilizando la base de individuos. En particular, se calculó la proporción de personas inactivas entre 18 y 65 años (Ina) sobre el total de población en edad de trabajar (Pet) en cada hogar, generando la variable h_Inactivosp. Esta variable busca capturar la carga económica interna del hogar y su potencial vínculo con situaciones de pobreza.

El proceso de preparación de la muestra también implicó la verificación y limpieza de variables con valores faltantes o inconsistentes, así como la exclusión de aquellas variables con información redundante o escaso poder explicativo. Para las variables numéricas con valores faltantes, se optó por una imputación simple utilizando la mediana, con el objetivo de preservar la distribución original de los datos sin introducir sesgos extremos. En el caso de las variables categóricas, los valores ausentes fueron reemplazados por una categoría adicional

que permitiera conservar la observación dentro del conjunto de entrenamiento. Esta estrategia de imputación buscó mantener el mayor número posible de registros en la muestra sin comprometer la calidad de la información. Adicionalmente, algunas variables relacionadas con ingresos laborales individuales fueron descartadas para evitar problemas de colinealidad con los ingresos ya agregados a nivel de hogar.

Luego de la integración y transformación de las bases, la muestra final conservó el total de observaciones originalmente disponibles en la base train_hogares.csv, compuesta por 165.960 hogares, de los cuales aproximadamente un 20% fueron clasificados como pobres según los criterios oficiales de línea de pobreza.

## Análisis descriptivo y justificación empírica y teórica de las variables sociodemográficas

**Tabla de estadísticas descriptivas**

| Variable | Media | Mediana | Desv. Est. | Mín. | Máx. |
|---|---|---|---|---|---|
| n_personas | 3.29 | 3.0 | 1.77 | 1.0 | 28 |
| mujer | 1.74 | 2.0 | 1.18 | 0.0 | 14 |
| hombre | 1.55 | 1.0 | 1.12 | 0.0 | 14 |
| niños | 0.59 | 0.0 | 0.88 | 0.0 | 14 |
| mayores | 0.32 | 0.0 | 0.61 | 0.0 | 6 |
| ocupados | 1.26 | 1.0 | 0.94 | 0.0 | 9 |
| educ_sup | 0.0 | 0.0 | 0.02 | 0.0 | 2 |
| edad_promedio | 37.44 | 33.5 | 16.88 | 5.67 | 102 |
| nivel_educativo | 4.31 | 4.33 | 1.08 | 1.0 | 9 |
| prop_ocupados | 0.42 | 0.33 | 0.32 | 0.0 | 1 |

La selección de las variables sociodemográficas empleadas en los modelos de predicción de pobreza responde a una doble motivación: por un lado, a la literatura académica sobre los determinantes estructurales de la pobreza en Colombia; por otro, al comportamiento estadístico observado en la base de datos, lo que permite garantizar tanto su relevancia teórica como su utilidad empírica en el ejercicio predictivo.

El número total de personas por hogar (n_personas) es una variable estructural ampliamente documentada como determinante de la condición de pobreza, al reflejar el tamaño del hogar y, con ello, la carga de consumo que enfrentan los miembros que generan ingresos. En esta muestra, los hogares tienen en promedio 3,29 personas, con un máximo de 28. Este rango amplio sugiere la presencia tanto de núcleos familiares pequeños como de hogares extensos, posiblemente en contextos rurales. Tal como lo señala Guerrero (2021), el tamaño del hogar afecta directamente la capacidad de cubrir necesidades básicas, ya que los recursos deben ser distribuidos entre más individuos, reduciendo el ingreso per cápita disponible (ver también Patiño & Duque, 2021).

La estructura de género del hogar se explora mediante las variables mujer y hombre, que reportan promedios de 1,74 y 1,55 respectivamente. La dispersión en ambos casos es alta (desviaciones estándar superiores a 1), y los máximos alcanzan 14 personas por sexo en un

mismo hogar. Esta composición resulta útil para analizar dinámicas laborales y de cuidado, ya que diversos estudios han destacado que la presencia femenina en el hogar puede estar vinculada con estrategias de ingreso alternativo, especialmente en contextos de informalidad (Muñoz, 2019). Aunque la evidencia no es concluyente, la variable permite capturar diferencias potenciales en la organización productiva del hogar.

En lo que respecta a la carga de dependencia, las variables niños (menores de 12 años) y mayores (personas de 65 años o más) ofrecen información clave. Más de la mitad de los hogares no cuenta con personas en estos rangos etarios, pero su presencia, aunque minoritaria, introduce presiones económicas adicionales. La media de niños es de 0,59, y la de mayores, de 0,32, con máximos de 14 y 6 respectivamente. Estos datos son consistentes con la evidencia de Arango (2023), quien señala que la estructura etaria del hogar incide directamente sobre el riesgo de pobreza multidimensional, en especial en zonas rurales, donde los mecanismos de protección social son más débiles.

La capacidad productiva del hogar se analiza a partir del número absoluto de personas ocupadas (ocupados) y la proporción relativa respecto al total del hogar (prop_ocupados). Ambas variables reflejan el potencial generador de ingreso. En promedio, los hogares cuentan con 1,26 personas ocupadas y una proporción de 0,42. Sin embargo, no es infrecuente encontrar hogares con solo un miembro que percibe ingresos, lo cual los vuelve más vulnerables ante cualquier choque económico o laboral. Estas variables han sido reconocidas en la literatura como predictores robustos de pobreza (Guerrero & Castellanos, 2022), al capturar la autosuficiencia económica interna del hogar.

Respecto al capital humano, el nivel educativo promedio (nivel_educativo) se ubica en 4,31, lo que corresponde aproximadamente a la educación secundaria completa. Por su parte, el número de personas con educación superior (educ_sup) es notablemente bajo (media de 0), con un máximo de apenas dos personas por hogar. Esta limitada acumulación de capital humano coincide con lo planteado por Martínez y Ramírez (2007), quienes sostienen que el nivel educativo es uno de los principales determinantes del ingreso y del acceso a empleos formales y estables. Así, estas variables permiten aproximar las oportunidades económicas de largo plazo de los hogares.

Finalmente, la edad promedio de los miembros del hogar (edad_promedio) se sitúa en 37,44 años, con una amplia desviación estándar de 16,88. La mediana de 33,5 años indica que buena parte de los hogares se encuentra en etapas activas del ciclo de vida, aunque también se observa la presencia de hogares con personas jóvenes (mínimo de 5,67 años) y de hogares envejecidos (hasta 102 años). Esta heterogeneidad etaria puede incidir tanto en la participación laboral como en la necesidad de asistencia y cuidados, afectando el bienestar general (Guerrero, 2021).

En conjunto, las variables seleccionadas presentan una adecuada combinación de fundamentos teóricos y propiedades estadísticas: presentan variabilidad, distribución asimétrica, y presencia de valores extremos, lo que permite explicar diferencias significativas en los niveles de pobreza observados. Su inclusión permite al modelo capturar aspectos estructurales, demográficos, educativos y laborales que afectan el bienestar económico de los hogares, con una perspectiva integral que va más allá del ingreso monetario y se alinea con los enfoques multidimensionales actualmente promovidos en las políticas públicas.

Para complementar el análisis de la elección de variables, se podrá encontrar en el anexo 1 de este trabajo las gráficas correspondientes de la estádistica descriptiva, con sus respectivas interpretaciones e implicaciones.

## Selección y entrenamiento del modelo

El modelo que logró el mejor desempeño en la plataforma de Kaggle fue un esquema de stacking que combinó las predicciones de dos algoritmos de aprendizaje supervisado no lineal: Random Forest y XGBoost. La elección de este enfoque se fundamentó tanto en la evidencia empírica como en la literatura técnica, que ha demostrado que los métodos de ensamblaje, especialmente el stacking, suelen superar en desempeño a modelos individuales, al integrar de manera estructurada la información contenida en múltiples algoritmos base (Wolpert, 1992; Sagi & Rokach, 2018).

El proceso de entrenamiento del modelo se llevó a cabo en varias fases. En primer lugar, se entrenaron de forma independiente los modelos base sobre la muestra de entrenamiento, utilizando la versión procesada de la base GEIH 2018 que combinaba información a nivel de hogar e individuo. Cada modelo fue entrenado sobre una matriz de variables explicativas construida a partir de indicadores demográficos, laborales y educativos, agregados al nivel del hogar. Posteriormente, se generaron las probabilidades predichas para cada hogar, y estas fueron combinadas mediante un promedio ponderado calibrado para maximizar el F1 Score sobre la validación cruzada.

Los hiperparámetros seleccionados para cada modelo base fueron definidos a partir de un proceso de validación cruzada. Para XGBoost, se utilizó una tasa de aprendizaje (eta) de 0.1, una profundidad máxima de 4 (max_depth), y proporciones de muestreo (subsample y colsample_bytree) de 0.8, lo cual permitió controlar el sobreajuste sin perder capacidad predictiva (Chen & Guestrin, 2016). Para Random Forest, se utilizaron 500 árboles (num.trees), un mtry optimizado empíricamente, y un min.node.size de 20, valores que ofrecieron una buena estabilidad y generalización del modelo (Breiman, 2001).

Un aspecto metodológico central fue el manejo del desbalance de clases, dado que solo una fracción menor de los hogares se encontraba en condición de pobreza. Para abordar este problema, se adoptó una estrategia de ponderación diferencial en el modelo de Random Forest, asignando mayor peso a la clase minoritaria. Esta técnica es recomendada para mejorar la sensibilidad del modelo sin sacrificar precisión (He & Garcia, 2009). En el caso de XGBoost, se utilizó la opción scale_pos_weight, ajustada automáticamente con base en la proporción de clases, lo cual permitió reforzar la penalización de errores sobre la clase pobre.

Adicionalmente, se ajustó el punto de corte (cutoff) utilizado para convertir las probabilidades predichas en clases binarias, seleccionando el valor que maximizaba el F1 Score en la muestra de entrenamiento. Este ajuste es crucial en contextos con desbalance, ya que evita que el modelo prediga sistemáticamente la clase mayoritaria para optimizar métricas tradicionales como la exactitud.

Por otro lado, entrando a las especificaciones de todos los modelos contemplados en este análisis con el fin de realizar una comparación y una hoja de ruta para la elección del mejor modelo, nos encontramos con las siguientes características:

## Ajuste de hiperparámetro

**Árboles de decisión (CART):** En el modelo CART, se ajustó el parámetro de complejidad (cp), que controla la poda del árbol. Los valores bajos de cp permiten árboles más complejos, mientras que valores altos favorecen árboles más simples. Se exploró una grilla de valores entre 0.0001 y 0.05, buscando un equilibrio entre la complejidad del modelo y su capacidad de generalización. Esta estrategia es coherente con la literatura, que sugiere que la poda adecuada de árboles evita el sobreajuste (James et al., 2013).

**Random Forest :**para el modelo Random Forest, se ajustaron los hiperparámetros num.trees, mtry y min.node.size. El número de árboles (num.trees) se fijó en 500, ya que un mayor número de árboles puede reducir la varianza del modelo, mejorando su estabilidad (Breiman, 2001). El parámetro mtry, que determina el número de variables consideradas en cada división, se ajustó para optimizar la diversidad entre los árboles. Finalmente, min.node.size se calibró para controlar la profundidad de los árboles, evitando el sobreajuste. Además, se incorporaron pesos de clase para abordar el desbalance en la variable objetivo, una práctica recomendada en contextos con clases desbalanceadas (Chen & Guestrin, 2016).

**Regresión logística:** en la regresión logística, el principal ajuste se centró en el punto de corte (cutoff) para clasificar las probabilidades predichas en clases binarias. Dado que la variable objetivo presenta un desbalance, se exploraron distintos valores de cutoff para maximizar el F1 Score, que equilibra precisión y sensibilidad, siendo adecuado en contextos con clases desbalanceadas (Saito & Rehmsmeier, 2015).

**XGBoost:** para el modelo XGBoost, se ajustaron los hiperparámetros eta, max_depth, subsample y colsample_bytree. La tasa de aprendizaje (eta) se fijó en 0.1, un valor comúnmente utilizado que permite una convergencia estable (Chen & Guestrin, 2016). La profundidad máxima de los árboles (max_depth) se estableció en 4, limitando la complejidad del modelo y reduciendo el riesgo de sobreajuste. Los parámetros subsample y colsample_bytree se fijaron en 0.8, introduciendo aleatoriedad en la selección de muestras y características, lo que mejora la generalización del modelo (Chen & Guestrin, 2016). Se utilizó validación cruzada con parada temprana (early_stopping_rounds) para determinar el número óptimo de iteraciones (nrounds), evitando el sobreajuste al detener el entrenamiento cuando el rendimiento en el conjunto de validación dejaba de mejorar.

**Ensamblaje y Stacking :** se implementaron dos estrategias de combinación de modelos: un ensamblaje simple y un modelo de stacking. El ensamblaje simple combinó las predicciones de la regresión logística y Random Forest, ponderando sus salidas para aprovechar la interpretabilidad del primero y la capacidad de modelar relaciones no lineales del segundo. El modelo de stacking integró las predicciones de XGBoost y Random Forest, utilizando un promedio ponderado de sus probabilidades predichas. Estas estrategias buscan mejorar el rendimiento predictivo al combinar modelos con diferentes fortalezas, una práctica respaldada por la literatura en aprendizaje automático (Wolpert, 1992).

## Análisis comparativo entre modelos

Tras la implementación de los modelos seleccionados y el ajuste de sus respectivos hiperparámetros, se procedió a evaluar su desempeño en la muestra de entrenamiento, utilizando como métrica principal el F1 Score. Esta métrica fue priorizada debido al desbalance de clases en la variable objetivo donde la proporción de hogares en condición de pobreza es significativamente menor, lo cual hace que métricas tradicionales como la exactitud (accuracy) puedan resultar engañosas. El F1 Score permite, en cambio, ponderar de

manera equilibrada la precisión (precision) y la sensibilidad (recall), siendo más informativa en contextos como este (Saito & Rehmsmeier, 2015).

Los modelos base, como la regresión logística y el árbol de clasificación CART, ofrecieron desempeños aceptables y se utilizaron como referencias para evaluar la ganancia marginal obtenida al aplicar algoritmos más sofisticados. El modelo logit, pese a su simplicidad, mostró una capacidad razonable para capturar la estructura de los datos, aunque fue superado en desempeño por modelos no paramétricos. CART, por su parte, evidenció cierta inestabilidad frente al desbalance de clases, a pesar del ajuste cuidadoso del parámetro de complejidad.

El modelo Random Forest mejoró sustancialmente la capacidad predictiva respecto a CART, mostrando mayor estabilidad y capacidad para identificar correctamente los hogares pobres. La inclusión de pesos de clase resultó efectiva para mitigar la tendencia del modelo a sobrepredecir la clase mayoritaria. Sin embargo, su desempeño fue ligeramente inferior al de XGBoost, lo cual es consistente con la literatura, donde los métodos de boosting suelen superar a los de bagging en tareas de clasificación compleja (Chen & Guestrin, 2016).

XGBoost se consolidó como uno de los modelos más competitivos en términos de F1 Score, gracias a su capacidad para modelar relaciones no lineales y su buen comportamiento frente a datos desbalanceados. El uso de regularización, subsampling y validación cruzada con parada temprana contribuyó a mejorar su generalización y prevenir el sobreajuste.

No obstante, la mejor performance fue alcanzada a través del modelo de stacking, que combinó las predicciones de XGBoost y Random Forest mediante un promedio ponderado de las probabilidades predichas. Esta estrategia logró capturar patrones complementarios entre ambos modelos y superar las limitaciones individuales de cada uno. El stacking permitió alcanzar un F1 Score (0.75) superior al de cualquier modelo individual, lo cual valida su implementación como técnica de ensamble efectiva en problemas de clasificación binaria.

En términos prácticos, el modelo de stacking fue seleccionado como el modelo final para la predicción sobre la muestra de prueba (test), dado que ofrecía el mejor compromiso entre precisión, sensibilidad y estabilidad en validación cruzada. Además, su desempeño fue consistente a lo largo de múltiples particiones de entrenamiento, lo que sugiere una buena capacidad de generalización.

Por último, dada esta hoja de ruta y las posibilidades de clasfificación que brindó cada uno de los modelos, para la construcción de nuestro mejor modelo , consideramos la elección e importancia de las variables de la siguiente manera

**Importancia de las variables**

Uno de los aspectos clave en la interpretación del modelo de mejor desempeño, el stacking entre XGBoost y Random Forest, fue la evaluación de la importancia relativa de las variables utilizadas en la predicción. Este análisis se realizó a partir de los puntajes de importancia generados por los modelos base, que permiten identificar qué variables contribuyeron más a mejorar la capacidad del modelo para clasificar correctamente los hogares pobres. En el caso de Random Forest, se utilizó el criterio del aumento promedio en la impureza (Mean Decrease in Gini), mientras que para XGBoost se tomó como referencia la ganancia promedio que aporta cada variable cuando es usada para dividir nodos (Gain). En ambos

modelos se observó una fuerte coincidencia respecto a las variables más relevantes, lo cual refuerza la consistencia de los resultados obtenidos.

Es importante aclarar que la decisión sobre qué variables incluir o excluir ya fue justificada en detalle en la sección de análisis descriptivo y justificación empírica de la elección de variables junto con las gráficas del anexo, donde se explicaron tanto los fundamentos teóricos como empíricos detrás de cada elección. En ese proceso, también se descartaron variables que, aunque altamente informativas, estaban directamente relacionadas con la construcción de la variable objetivo. Esto incluyó variables como Ingpcug (ingreso per cápita del hogar), Ingtotugarr (ingreso total con imputación de arriendo), Lp (línea de pobreza) y Li (línea de indigencia), así como cualquier combinación que pudiera implicar una predicción trivial del criterio de pobreza. La intención fue que el modelo aprendiera patrones a partir de características estructurales y sociodemográficas del hogar, y no simplemente reprodujera mecánicamente el umbral de clasificación. Esta decisión permitió construir un modelo más realista y útil desde una perspectiva de política pública, centrado en variables que podrían ser observadas o recolectadas incluso cuando no se dispone de información completa sobre ingresos.

## Conclusión

Este ejercicio permitió aplicar de forma rigurosa diversas técnicas de aprendizaje supervisado al problema de predicción de pobreza en hogares colombianos, utilizando como insumo los datos de la GEIH 2018. A partir de un trabajo cuidadoso de integración y transformación de las bases de datos a nivel hogar e individuo, se construyó una muestra analítica sólida que permitió extraer información relevante sobre las características sociodemográficas, laborales y educativas de los hogares.

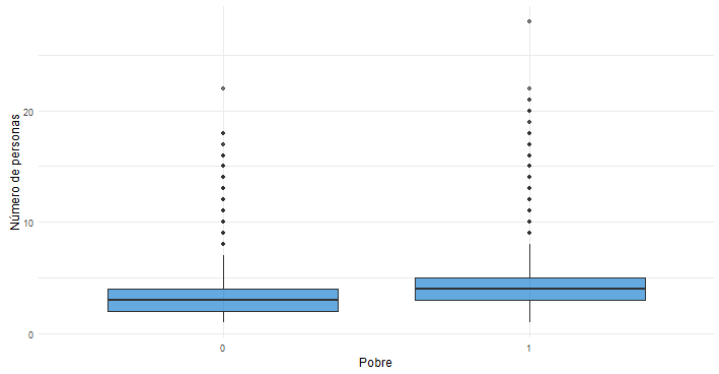
El análisis exploratorio mostró patrones claros en la relación entre pobreza y variables como el tamaño del hogar, la proporción de ocupados, el nivel educativo promedio y la presencia de personas dependientes. Estos hallazgos guiaron la selección de variables en los modelos predictivos y sirvieron de base para justificar la exclusión de variables directamente asociadas al ingreso, con el fin de evitar redundancias en la predicción.

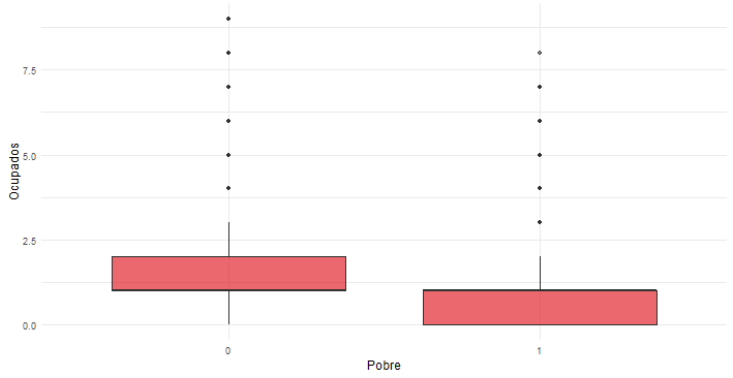
Se implementaron múltiples modelos de clasificación binaria, destacando la regresión logística, CART, Random Forest y XGBoost. El modelo de mejor desempeño fue una estrategia de stacking que combinó Random Forest y XGBoost, logrando el mayor F1 Score sobre la base de entrenamiento. Esta combinación fue efectiva para capturar tanto relaciones no lineales como estructuras generales en los datos, integrando fortalezas de ambos algoritmos.

Además, se abordó de manera explícita el desbalance de clases, mediante la inclusión de pesos de clase y la calibración del umbral de clasificación, lo que permitió mejorar la sensibilidad del modelo hacia la clase minoritaria (los hogares pobres). La evaluación de importancia de variables reforzó la relevancia de las dimensiones estructurales, laborales y educativas como predictores de pobreza, validando tanto los resultados empíricos como las decisiones metodológicas adoptadas.

En conjunto, este trabajo pone en evidencia el valor del aprendizaje automático como herramienta para la focalización de política social y la identificación de hogares vulnerables, siempre y cuando se combine con una adecuada comprensión del fenómeno social subyacente y una elección informada de variables e indicadores.

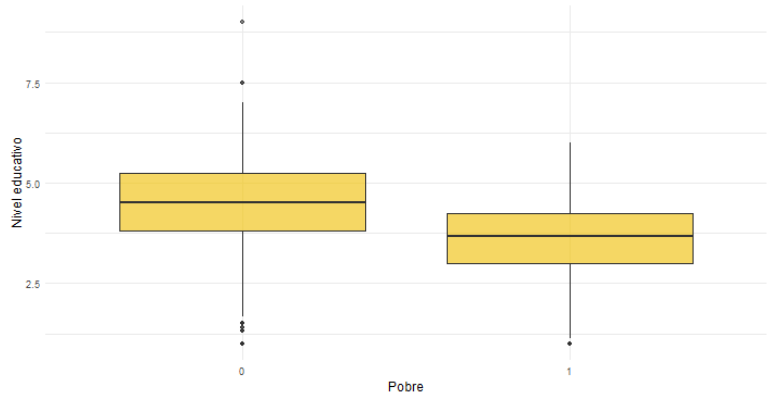
# Anexo 1

| Boxplot tamaño del hogar y clasificación de pobreza |
| --- |
|  |
| A partir del gráfico se observa que los hogares en condición de pobreza tienden a ser más numerosos que aquellos no pobres, como lo evidencia su mayor mediana y rango intercuartílico. |

| Boxplot ocupados y clasificación de pobreza |
| --- |
|  |
| El gráfico muestra que los hogares pobres tienen en promedio menos personas ocupadas que los no pobres, con una mediana cercana a uno. Esta diferencia refleja una menor capacidad de generación de ingresos en los hogares en condición de pobreza, lo cual es coherente con el enfoque clásico del mercado laboral como determinante central del |

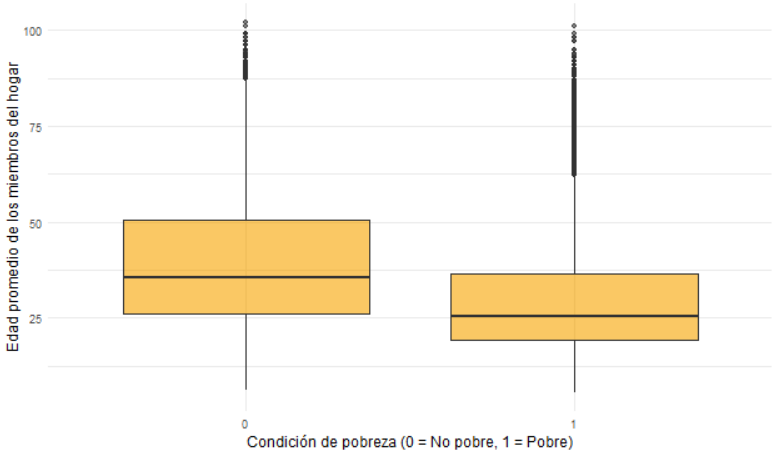
bienestar económico.

**Boxplot Nivel educativo y clasificación de pobreza**




El gráfico evidencia una brecha clara en el nivel educativo promedio entre hogares pobres y no pobres. Los hogares no pobres presentan una mediana de escolaridad más alta y mayor concentración en niveles superiores.

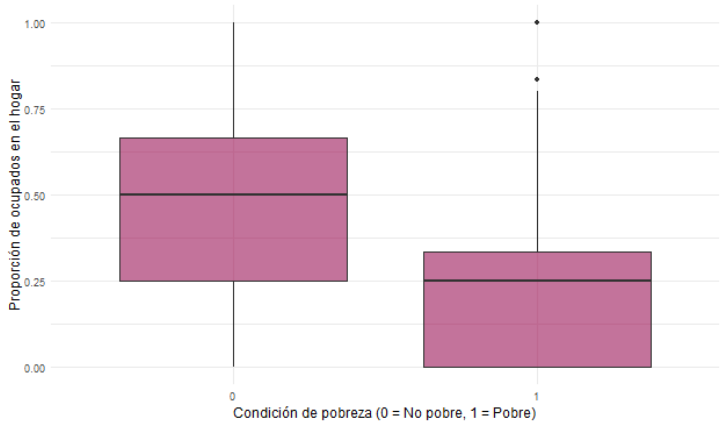
**Boxplot edad y clasificación de pobreza**




El gráfico muestra que los hogares no pobres presentan una edad promedio más alta que los hogares pobres, tanto en la mediana como en el rango intercuartílico. Además, se observa una mayor dispersión en los hogares no pobres, con una proporción mayor de hogares con

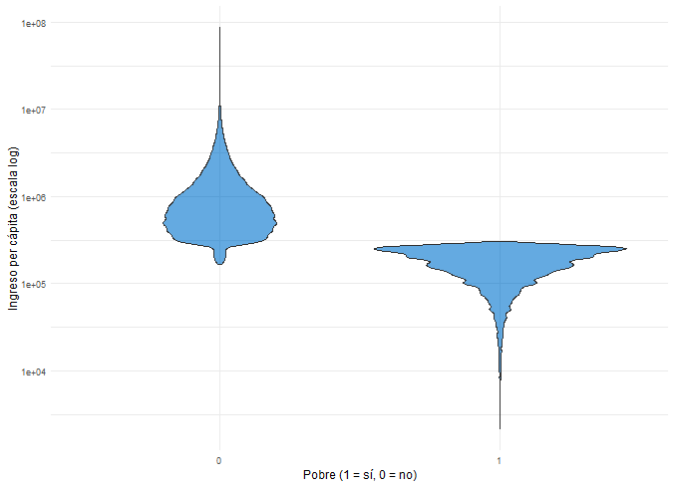
miembros de edad avanzada. En contraste, los hogares pobres tienden a concentrarse en edades más jóvenes, con una distribución más compacta y menor variabilidad.

**Boxplot proporción de ocupados en el hogar y clasificación de pobreza**



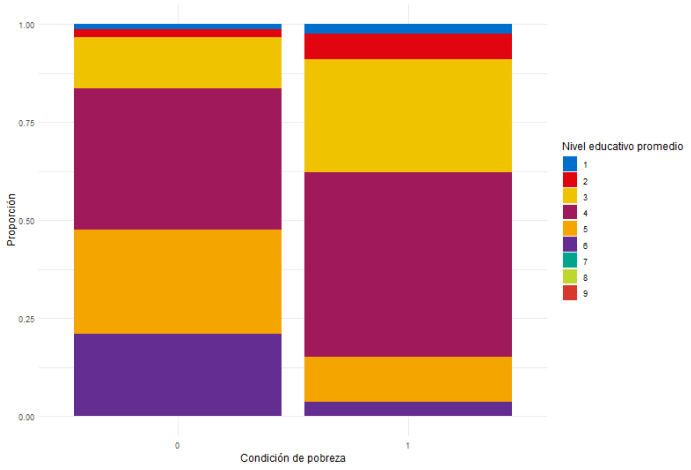

El gráfico evidencia que la proporción de personas ocupadas dentro del hogar es mayor en los hogares no pobres, con una mediana cercana al 0.6, mientras que en los hogares pobres esta mediana se reduce a aproximadamente 0.3. Además, la dispersión en los hogares pobres es considerablemente menor, con la mayoría de los valores concentrados en proporciones bajas. Esto sugiere una diferencia clara en la intensidad de participación laboral entre los dos grupos poblacionales.

**Violin plot de ingreso per cápita y clasificación de pobreza**



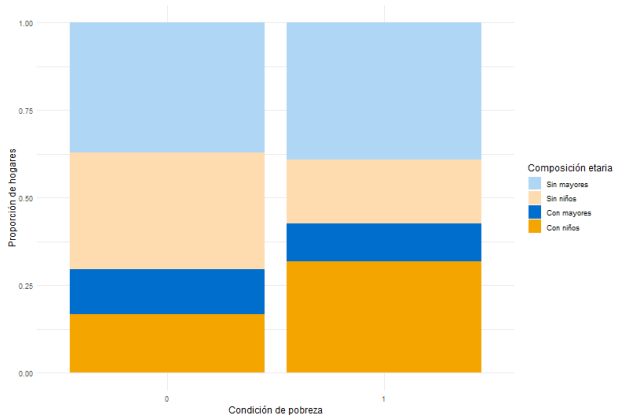

El gráfico muestra una clara separación en las distribuciones de ingreso per cápita entre los hogares pobres y no pobres. Mientras los ingresos de los no pobres presentan mayor dispersión y densidad en niveles altos, los ingresos de los pobres se concentran en los valores más bajos, con una distribución mucho más angosta. Esta diferencia es especialmente evidente en la escala logarítmica, donde la brecha entre ambos grupos se visualiza como una diferencia prácticamente completa en las densidades centrales.

**Barras apiladas de proporción de nivel educativo de clasificación de pobreza**



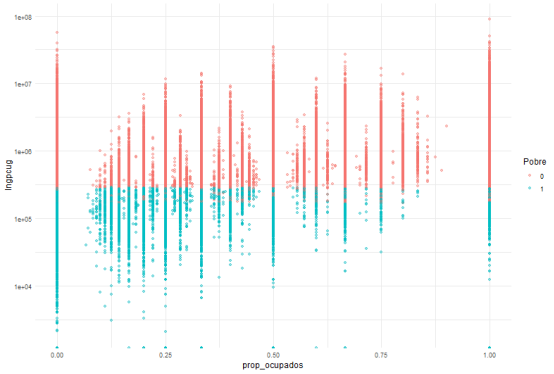

El gráfico muestra que los hogares no pobres concentran una mayor proporción de niveles educativos promedio entre 4 y 6, mientras que los hogares pobres se agrupan mayoritariamente en niveles más bajos, especialmente entre los niveles 2, 3 y 4. La presencia de niveles superiores (7, 8 y 9) es prácticamente nula en ambos grupos, pero se observa una ligera representación únicamente entre los no pobres. Esta distribución revela una clara brecha en el perfil educativo promedio de los hogares según su condición económica, con una mayor acumulación de escolaridad en los hogares que no se encuentran en situación de pobreza..

**Barras apiladas de número de integrantes del hogar y clasificación de pobreza**



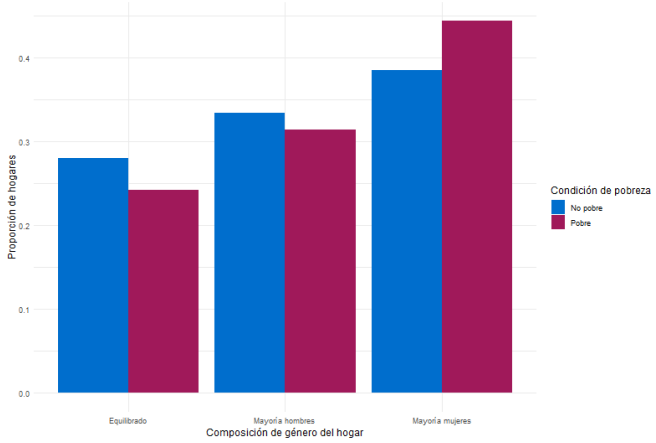

El gráfico muestra que la proporción de hogares con al menos un niño es notablemente mayor entre los hogares pobres, mientras que la presencia de adultos mayores no presenta diferencias sustanciales entre ambos grupos. En particular, los hogares pobres concentran cerca del 35% de hogares con niños, frente a aproximadamente 20% en los no pobres. Esto sugiere que la presencia de menores de edad es un rasgo distintivo de los hogares en situación de pobreza, lo cual puede implicar una mayor carga de dependencia y demanda de recursos básicos como educación, salud y cuidado.

**Scatterplot de ingreso percapita y proporción de ocupados**




El gráfico revela una asociación positiva entre la proporción de personas ocupadas en el hogar y el nivel de ingreso per cápita, especialmente entre los hogares no pobres. A medida que aumenta prop_ocupados, se observa una mayor concentración de puntos con ingresos altos en el grupo no pobre (color rojo), mientras que los hogares pobres (color azul) tienden a permanecer en niveles bajos de ingreso independientemente de su nivel de ocupación. Esta diferencia sugiere que, si bien la participación laboral es un factor relevante, no garantiza por sí sola la superación de la pobreza, posiblemente debido a la informalidad o precariedad de los empleos.

**Barras apiladas de composición de género en el hogar por condición de pobreza**




El gráfico muestra que los hogares con mayoría de mujeres tienen una mayor representación dentro del grupo de hogares pobres, con una proporción cercana al 45%, superior a la observada en los no pobres. En contraste, los hogares con mayoría de hombres o composición equilibrada presentan proporciones más altas en el grupo no pobre. Esta diferencia sugiere que la composición femenina del hogar podría estar asociada a una mayor vulnerabilidad económica, lo que refuerza la necesidad de considerar enfoques diferenciales de género en el análisis de pobreza.

## Referencias


- Arango, F. (2023). Identificación de dimensiones y estimación de la pobreza multidimensional en Colombia mediante métodos de aprendizaje estadístico. Universidad de Antioquia.

- Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., ... & Boulesteix, A.-L. (2021). Hyperparameter optimization: Foundations, algorithms, best practices and open challenges.

- Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).

- Guerrero, A. (2021). Análisis de la pobreza en Colombia basado en aprendizaje automático. Universidad de Bogotá Jorge Tadeo Lozano.

- Guerrero, A., & Castellanos, J. (2022). Un modelo de estimación de pobreza a partir de datos no estructurados y machine learning. Universidad de los Andes.

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R. Springer.

- Martínez, C. A., & Ramírez, J. (2007). ¿Cuál es el mejor indicador de pobreza en Colombia? Revista de Economía Institucional, 9(16), 165–190.

- Muñoz, J. (2019). Análisis de la pobreza subjetiva en Colombia. Universidad de los Andes.

- Patiño, M., & Duque, D. (2021). Modelo de predicción del nivel de ingresos basado en Machine Learning y Deep Learning. Universidad de los Andes.

- Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432.

- Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259.