# Predicting poverty

**Daniel Felipe Cortes Cendales**

**Jorge Esteban Gómez Carmona**

**Nicolás Alberto González Gort**

**Sebastián Marín Lombo**

4 de Diciembre del 2023

**Palabras Clave:** Pobreza, medición de pobreza, predicción de pobreza, datamining

**Clasificación JEL:** C81, D12, D61, I32, I38

**Link del repositorio:** https://github.com/BigData-MachineLearning/P-Set3.git

# 1 Introducción

La pobreza es una preocupación política clave, y en la mayoría de los países, particularmente en aquellos con niveles altos, la pobreza se mide con los gastos de consumo. Desafortunadamente, los gastos de consumo son relativamente complejos de medir y, por lo tanto, la pobreza monetaria se considera compleja y costosa de evaluar. Otra forma estándar de identificar la pobreza se basa en los ingresos. Eso significa que las personas con un ingreso inferior a cierto umbral se consideran pobres, y el número de personas por debajo de esa línea determina el índice de recuento. Sin embargo, esta perspectiva es criticada por los teóricos del enfoque del desarrollo humano y las capacidades (Alkire 2005; Nussbaum 2001; Sen 1999, 1985), argumentando ampliamente que el ingreso no es intrínsecamente esencial sino instrumentalmente significativo, y que la medición del ingreso del recuento carece de información. - información sobre la realidad de los pobres (Sen 1985, 1992, 2017).

Por estas razones, el estado de pobreza a menudo se evalúa con indicadores de pobreza. Estos sustitutos son indicadores altamente correlacionados con el consumo, el ingreso y la pobreza, pero más fáciles de observar y recopilar.

Ahora, para monitorear y evaluar cómo el desarrollo económico y la política nacional interactúan y afectan la pobreza, es esencial realizar mediciones periódicas y frecuentes. El método estándar para estimar la tasa de pobreza se basa en datos completos de encuestas sobre el consumo detallado de los hogares. Pero, de igual manera estas encuestas son costosas, consumen mucho tiempo y, a menudo, sólo se realizan cada cuatro o cinco años. Se necesitan métodos más baratos y simples que puedan usarse para evaluar los niveles de pobreza con mayor frecuencia, razón por la cual metodologías de **machine learning** podrían resultar útiles para cubrir esas falencias.

Este texto describe el proceso de construcción de modelos de clasificación para predecir la situación de pobreza en hogares utilizando datos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 en Colombia. La información proviene del Departamento Administrativo Nacional de Estadística (DANE) y se utiliza para crear variables relacionadas con características de los hogares, aspectos laborales, ingresos, gastos y clasificación de pobreza.

Se seleccionaron variables relevantes, como número de personas por hogar, nivel educativo, ingresos, entre otros. Se aplicaron diferentes modelos, incluyendo regresión y clasificación, con el objetivo de maximizar la métrica F1, que evalúa precisión y recuperación. Se detallan los resultados de varios modelos, destacando el uso efectivo de Adaboosting para la clasificación, con un F1 cercano a 0.7.

Se concluye que los modelos de clasificación superaron a los de regresión para predecir la pobreza en hogares. Aunque se logró un desempeño aceptable, se sugiere mejorar la calidad y cantidad de datos, realizar más ingeniería de características, explorar modelos de regresión adicionales y considerar el uso de modelos de aprendizaje profundo para aumentar la precisión en futuros estudios.

# 2 Revisión de Literatura

Varios estudios han destacado la relación entre los indicadores sociales, datos alternativos y las metodologías aplicadas de machine learning para predecir los indicadores de pobreza y focalizar políticas. Cuando se habla de datos alternativos se hace referencia a fuentes distintas a encuestas y censos, principalmente, esto porque hay países o zonas que no tienen datos disponibles para predecir indicadores sociales de pobreza o riqueza, y cuando los hay, es necesario actualizarlos, sin embargo, el costo de actualización es elevado porque se requieren encuestas que abarquen grandes territorios. Siendo los anteriores, los principales argumentos a favor del uso de nuevas tecnologías de inteligencia artificial para estimar indicadores sociales como proxy de la pobreza (Hall & Ohlsson, 2022)

## 2.1 Métodos predictivos

Elbers et al. Propusieron un enfoque para estimar una medida de pobreza mediante el uso de un modelo de regresión (2003). Su enfoque, también conocido como "mapeo de la pobreza", combina datos del censo con un modelo de regresión estimado utilizando datos de una encuesta de gasto. El enfoque de mapeo de la pobreza se ha modificado y utilizado para la imputación de encuesta a encuesta en varios contextos. Mientras que el enfoque de mapeo de la pobreza produce mapas de pobreza detallados, la imputación de encuesta a encuesta produce estimaciones de pobreza a nivel nacional y subnacional. Ambos métodos se basan en un modelo bien definido con parámetros estables. Todo esto fue usado de manera aplicada por Christiaensen et al (2012), utilizando datos de Vietnam y Mongolia, por medio del cual validaron el método de imputación de encuestas logrando estimar datos de pobreza que posteriormente se usaron para focalizar programas sociales.

A la misma conclusión se llegó en un estudio de validación que utilizó siete encuestas de presupuesto familiar de Uganda (Mathiassen, 2013) y en un estudio de Dang y Lanjouw (2018) que utilizó dos encuestas de consumo comparables de la India.

Por otro lado, Newhouse et al. (2014) encontraron que un enfoque similar falla al imputar la pobreza a partir de encuestas de presupuesto familiar en encuestas de fuerza laboral utilizando datos de Sri Lanka. Argumentan que para que tal sistema produzca estimaciones confiables de la pobreza, se debe establecer un sistema de encuestas de seguimiento del bienestar para recopilar predictores de manera consistente

Adicionalmente para predecir métodos, existe una tendencia a utilizar múltiples algoritmos de predicción para comparar desempeños; Un trabajo notable utiliza varios algoritmos como Random Forest, Gradient Boosting Machines, modelos lineales con regularización, modelos de aprendizaje profundo y nuevas estrategias como el proceso de aprendizaje apilado (Pokhriyal et al., 2022). Sin embargo, estudios que buscan avanzar en la capacidad explicativa de los modelos muestran que Random Forest es el algoritmo con mejor rendimiento (Pave & Stender, 2017)

## 2.2 Predicción de datos en Colombia

En Colombia, un estudio reciente del Departamento Nacional de Estadística (DANE) realizó tres experimentos para predecir el IPM de Colombia (Hall & Ohlsson, 2022). Utilizaron la métrica mencionada a nivel de vecino como variable de respuesta. Utilizan redes neuronales convolucionales para la extracción de características, seguidas de un proceso de estimación con algoritmos de aprendizaje automático. Los experimentos parten de una línea de base, donde estiman la pobreza con covariables clásicas en el último censo de 2018. Los detalles de los experimentos que realizaron se compartieron en su repositorio 1 y se pueden resumir de la siguiente manera:

1. Utilizan covariables censales como predictores y el Análisis de Componentes Principales (PCA) como técnica de reducción multidimensional, eligiendo cinco componentes. Informaron que su MPI tiene valores de 0 y 1, que excluyeron del primer experimento. Los métodos fueron la regresión del árbol de aumento de gradiente y el bosque aleatorio. El mejor resultado de rendimiento se obtuvo utilizando el algoritmo Gradient Boosting (r cuadrado igual a 0,6789 y RMSE igual a 0,7818);
2. En el segundo experimento utilizaron las mismas variables que en el primero, pero esta vez incluyeron los valores de 0 y 1 MPI. El mejor resultado fue utilizar el algoritmo Gradient Boosting; el R cuadrado fue igual a 0,6537 y el RMSE igual a 1,1898; mientras tanto, con el segundo mejor algoritmo lograron un R cuadrado de 62,81% y un RMSE de 1,233;
3. En el tercer experimento, utilizaron imágenes Sentinel-2 como funciones de entrada. Utilizaron Resnet34 como modelo previamente entrenado para transferir conocimientos y ajustar los datos. Aplicaron aumento de datos, se realizaron rotaciones de 90 grados en el eje horizontal y vertical y se realizó contraste de imágenes. Extrajeron 512 covariables de la red neuronal (los pesos). Luego de la futura extracción, aplicaron PCA para obtener cinco componentes, que interpolaron con el método del vecino natural a nivel de bloque. Estimaron el MPI utilizando el enfoque 1 (A1), que incluye los valores 0 y 1 del MPI, y el enfoque 2 (A2), que excluye los valores 0 y 1. El mejor resultado reportado alcanzó un RMSE igual a 0.9067 y un R cuadrado igual a 0.5757 utilizando un Random Forest como clasificador y siguiendo el enfoque A2.

La metodología de DANE se basa en un proceso de aprendizaje profundo que utiliza datos y algoritmos de difícil acceso que requieren una potencia computacional considerable. En la propuesta, los datos de acceso abierto se utilizan junto con un preprocesamiento menos complejo y algoritmos de aprendizaje automático fáciles de usar, que permiten avanzar en la interpretabilidad de los modelos de estimación. Ese es un aspecto que se plantea en la literatura como un desafío en los modelos que utilizan herramientas de IA para la estimación de la pobreza (Muñeton & Manrique, 2023).

# 3 Datos

Para llevar a cabo los modelos de clasificación, se utilizará una base de datos recopilada a partir de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 a nivel de personas. Esta información proviene tanto del Departamento Administrativo Nacional de Estadística (DANE) como de la misión de "Empalme de las series de Empleo, Pobreza y Desigualdad -

MESE". La base de datos abarca diversas categorías, incluyendo características de los hogares, así como datos relacionados con aspectos laborales, educativos, ingresos, gastos, entre otros.

Adicionalmente, se incorporará información sobre ingresos de los hogares, unidades de gasto y el valor de las líneas proporcionadas por el DANE para el cálculo de la incidencia de la pobreza e indigencia, así como el cálculo de la pobreza monetaria. Estos datos específicos fueron utilizados en la elaboración del informe titulado "COLOMBIA - Medición de Pobreza Monetaria y Desigualdad 2018". Es importante destacar que estos datos son de gran relevancia en la medición de indicadores económicos y sociales para la clasificación de la pobreza, convirtiéndose así en una fuente confiable para la investigación y la toma de decisiones políticas.

Después de extraer los datos, se realizó una selección de variables basada en características como el número de personas por hogar, el número de menores de edad, la identificación del jefe del hogar, su nivel educativo, la recepción de subsidios, tipos de ingreso, cantidad de trabajadores formales en los hogares, entre otros. Estas fueron utilizadas para construir diversas variables relacionadas con ingresos, tasas de afiliación, hacinamiento, tipo de vivienda, entre otras. Las variables utilizadas se muestran a continuación:

**Tabla 1**

Estadisticas Descriptivas caracteristicas del hogar

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Total cuartos | 164,959 | 3.4 | 1.2 | 1 | 98 |
| habitaciones | 164,959 | 2.0 | 0.9 | 1 | 15 |
| Tipo de ocupación | 164,959 | 2.5 | 1.3 | 1 | 6 |
| Número de personas | 164,959 | 3.3 | 1.8 | 1 | 28 |
| Número de adultos | 164,959 | 2.4 | 1.2 | 1 | 19 |
| Número de menores de edad | 164,959 | 0.9 | 1.1 | 0 | 15 |
| Cuartos por persona | 164,959 | 1.7 | 0.8 | 0.2 | 16.0 |
| Edad promedio del hogar | 164,959 | 37.4 | 16.9 | 5.7 | 102.0 |
| Cabeza de hogar mujer | 164,959 | 0.4 | 0.5 | 0 | 1 |
| Educacion cabeza de hogar | 164,959 | 4.4 | 1.4 | 1 | 9 |

**Tabla 2**

Estadisticas Descriptivas Variables Laborales

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Ocupados | 164,959 | 0.7 | 1.0 | 0 | 9 |
| Promedio de horas trabajadas | 164,959 | 39.2 | 19.7 | 0.0 | 130.0 |
| Tamaño de la empresa | 117,156 | 3.8 | 3.3 | 1 | 9 |
| Otro trabajo | 164,959 | 0.2 | 0.7 | 0.0 | 9.0 |
| Disponible para trabajar | 164,959 | 0.1 | 0.3 | 0.0 | 4.0 |
| Tasa de afiliación | 164,959 | 0.8 | 0.3 | 0.0 | 1.0 |

| | | | | | |
|---|---|---|---|---|---|
| regimen de salud | 164,959 | 0.5 | 0.6 | 0.0 | 6.0 |

**Tabla 3**

Estadisticas Descriptivas Variables Ingresos y gastos

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Ingreso por trabajo | 164,959 | 0.7 | 1.0 | 0 | 9 |
| Ingreso por arriendos | 164,959 | 39.2 | 19.7 | 0.0 | 130.0 |
| Otros ingresos | 117,156 | 3.8 | 3.3 | 1 | 9 |
| Remesas | 164,959 | 0.2 | 0.7 | 0.0 | 9.0 |
| Recibe ayudas del gobierno | 164,959 | 0.1 | 0.3 | 0.0 | 4.0 |
| Bonificaciones | 164,959 | 0.8 | 0.3 | 0.0 | 1.0 |
| Auxilio de alimentación | 164,959 | 0.5 | 0.6 | 0.0 | 6.0 |

**Tabla 4**

Estadisticas Descriptivas Variables de pobreza

| Statistic | N | Mean | St. Dev. | Min | Max |
|---|---|---|---|---|---|
| Linea de indigencia | 164,959 | 120,395.5 | 7,202.6 | 99,544.8 | 131,125.6 |
| Linea de pobreza | 164,959 | 271,522.3 | 33,657.0 | 167,222.5 | 303,816.7 |
| pobre | 164,959 | 0.2001952 | 0.4001475 | 0 | 1 |

En primer lugar, en la Tabla 1 se pueden observar las variables de características de los hogares. De esta tabla, la mayoría de las variables fueron sacadas directamente de la encuesta. Sin embargo, la variable cuartos por persona si fue una variable construida utilizando la cantidad de personas en el hogar y dividiéndolas por la cantidad de cuartos en el mismo. Esta variable fue creada con el objetivo de usarla como proxy de hacinamiento. De manera similar, la variable edad promedio del hogar también fue construida a partir de la base de datos utilizando la edad de cada persona en el hogar y dividiéndola por el total de personas en el mismo. En general, las variables de la Tabla 1 buscan recapitular la composición demográfica de los hogares al interior.

Por otro lado, en la Tabla 2 se pueden observar las variables seleccionadas que hacen parte de las características laborales de los hogares. En esta, las variables promedio de horas trabajadas, otro trabajo, disponible trabajar y tasa de afiliación fueron creadas con la base de datos original. Particularmente, las variables promedio de horas trabajadas, otro trabajo y disponible para trabajar buscaban capturar si los hogares deseaban trabajar más horas o si

deseaban conseguir otro trabajo. Análogamente, la variable de tasa de afiliación al régimen de salud tenía el objetivo de ser un indicador de acceso a los servicios de salud, el impacto en la productividad laboral y actuar un indicador socioeconómico.

Asimismo, la Tabla 3 presenta las variables de ingresos y gastos de los hogares. En ella se detallan los ingresos obtenidos a través del trabajo, así como los provenientes de arrendamientos y otras fuentes como intereses, dividendos, utilidades o cesantías. De igual manera, se registra la recepción de remesas, en caso de haber recibido dinero desde el extranjero. Además, se destaca la presencia de primas, bonificaciones, auxilio de transporte, junto con la recepción de auxilio de alimentación y ayudas del gobierno en el hogar. Cabe destacar que también se incluye una variable de acceso al sistema financiero, que mide si el hogar recibió ingresos por inversiones, préstamos, depósitos, entre otros. Estas variables son de especial importancia para evaluar la distribución de ingresos entre los hogares y facilitar la clasificación por categoría de pobreza. Un hogar que requiere auxilio de alimentación o ayudas del gobierno es más probable que sea catalogado como pobre, así como un hogar con acceso al sistema financiero tiene más probabilidad de no ser considerado pobre.

Por último, en la Tabla 4 se pueden observar las variables de clasificación de pobreza. En principio, se encuentra la línea de indigencia, que define el umbral por debajo del cual una persona o familia se considera incapaz de cubrir sus necesidades básicas de alimentación. De igual forma, se presenta la línea de pobreza, que es un umbral económico utilizado para determinar el límite monetario por debajo del cual se considera que una persona o una familia se encuentra en situación de pobreza. Finalmente, se incluye la variable "pobre", que clasifica cada hogar como pobre si su valor es igual a 1 y como no pobre si es igual a 0.

Por otra parte, se llevó a cabo un análisis gráfico como preámbulo para la clasificación de hogares en la categoría de pobreza. En primer lugar, en la Gráfica 1 se presenta la distribución de los hogares considerando el número de habitaciones por persona y el costo de la vivienda. Se puede observar que los hogares clasificados como pobres tienden a tener un mayor número de personas por habitación, indicando un mayor nivel de hacinamiento. Al mismo tiempo, estos hogares pagan valores de arriendo más bajos. En contraste, los hogares no pobres muestran menos hacinamiento y tienen la capacidad de pagar arriendos más elevados. De igual manera, en el Gráfico 2 se observa que los hogares clasificados como pobres presentan una menor tasa de afiliación a la seguridad social. Esto podría indicar que los integrantes del hogar tienen empleos informales o se encuentran desempleados, lo cual podría traducirse en ingresos más bajos y estar asociado con pagos más reducidos por concepto de arriendo.

Finalmente, en el gráfico 3 se observa que la mediana del número de horas trabajadas para el grupo de personas pobres es ligeramente más alta que la del grupo no pobre. Esto podría sugerir que aquellos que son pobres necesitan trabajar más horas, en promedio, posiblemente debido a empleos con salarios más bajos que requieren jornadas laborales más extensas para cubrir necesidades básicas.

Asimismo, el gráfico 4 muestra una disparidad en los niveles de empleo entre los dos grupos. Los no pobres parecen tener más oportunidades de empleo o un mayor número de empleos por adulto en comparación con los pobres. Por último, los resultados del gráfico 5 indican que las familias pobres tienden a tener un número medio de hijos más alto y una mayor variabilidad en el tamaño de la familia en comparación con las familias no pobres. La presencia de valores atípicos en ambas categorías señala casos excepcionales en ambas situaciones económicas, lo que podría complicar la clasificación al tener en cuenta estas variables.

# 4 Modelo y resultados

Con el objetivo de abordar la problemática de prever la situación de pobreza en hogares a través de datos censales, se optó por implementar diversos modelos con enfoques de regresión y de clasificación. Por un lado, para los modelos de regresión se implementaron modelos que buscaban estimar el ingreso total de los hogares e implementar una clasificación directa del estatus de pobreza mediante el uso de la línea de pobreza disponible en los datos. Por el otro lado, para los modelos con enfoque de clasificación se buscó categorizar directamente a los hogares. En términos de desempeño, objetivo primordial consistió en maximizar la métrica F1, la cual evalúa la precisión y recuperación de los modelos. Por esta razón, se evaluaron distintos modelos con especificaciones diferentes, esto con el objetivo de fortalecer la robustez de los resultados. Además, se implementaron técnicas de ingeniería de características y gestión del desbalanceo de clases. A continuación, se presenta un resumen de los resultados obtenidos para los enfoques mencionados anteriormente.

Inicialmente, para el enfoque de clasificación se buscó implementar un modelo de predicción logístico, con una selección de variables preliminar que arrojó un valor F de 0.22. Con esto en mente, se concluyó que existía una pobre ingeniería de características y que el modelo presentaba limitaciones al no explorar de manera exhaustiva variables y transformaciones que pudieran capturar la complejidad de la relación entre las características y la condición de pobreza. Seguidamente, se implementó para el mismo enfoque un modelo de árboles de clasificación y regresión (CART) con un mayor número de variables explicativas con el objetivo de mejorar la capacidad descriptiva. No obstante, este modelo resultó sumamente deficiente ya que no pudo identificar las sutilezas que caracterizan a los hogares pobres, llevándolo a clasificar a los hogares de manera generalizada como pobres.

En adición a estos modelos, se implementó un modelo de Adaboost que se mostró significativamente más eficiente a la hora de clasificar. Particularmente, el F que arrojó este modelo fue de 0.69, mostrando una mejoría sustancial en el poder predictivo del modelo con respecto al modelo logístico y al de árboles de clasificación y regresión. Aún más, se continuaron buscando alternativas en las estrategias implementadas para mejorar la capacidad predictiva. Por esta razón, se diseñó un modelo logístico con estrategia de

submuestreo (undersampling) que se mostró igual de deficiente que el CART, ya que presentó un F de 0, situación que ocurrió debido a que esta técnica redujo excesivamente el tamaño de la muestra y eliminó información relevante de la clase mayoritaria. Adicionalmente, se exploraron dos modelos de Adaboost adicionales, uno incorporando la técnica de manejo del desbalanceo con Smote y otro introduciendo una variable adicional (oficio) en la base de datos. Sin embargo, aunque ambos modelos incorporaron elementos novedosos, ninguno logró al primer modelo adaptado con la metodología Adaboost, la cual resultó siendo la más eficiente dada su capacidad para ajustarse gradualmente a los errores, posiblemente compensando el desbalanceo.

Para este punto del trabajo, se quiso explorar un enfoque por regresión. Para esto, se implementaron modelos de regularización como Lasso y Elastic Net, esto bajo la sospecha de presencia de multicolinealidad en las variables explicativas del modelo. Además, se deseó implementar la metodología de boosting, ya que esta mostró buen desempeño para los modelos de clasificación. Sin embargo, los resultados para estas metodologías se mostraron inferiores en la capacidad predictiva de los modelos. Particularmente, el Lasso y el elastic net tuvieron un F de 0.29 cada uno. Por otro lado, el modelo que utilizó la metodología de Boosting obtuvo un F de 0.13, mostrándose sumamente deficiente con respecto a sus pares en el enfoque de clasificación. La hipótesis planteada para explicar este comportamiento es que la variable de ingreso requiere una ingeniería de características más detallada y la exploración de especificaciones más detalladas para capturar relaciones no lineales más complejas. Además, limitaciones temporales, la restricción de realizar únicamente 5 envíos por día, llevaron al equipo a priorizar el abordaje del problema de clasificación. Más aun teniendo en cuenta que el procesamiento de modelos con metodologías de Boosting son significativamente demandantes computacionalmente, llegando a tardar hasta 9 horas en procesarse.

Con lo anterior en mente, se concluye que, para el enfoque de clasificación, el modelo que obtuvo la mayor capacidad predictiva fue aquel utilizado con la metodología de Adaboosting. En particular, este modelo se construyó mediante el uso de 100 árboles para el proceso de Boosting y un tamaño mínimo para las hojas de 10. Cabe resaltar que estos parámetros fueron los que maximizaron la métrica f. En comparación con el siguiente mejor modelo, que también fue un Adaboosting, la diferencia radica en que el mejor modelo no incluyó la variable "oficio", y que el siguiente se realizó con 93 árboles y un tamaño mínimo para las hojas de 8. Por otro lado, la penalidad utilizada en el modelo logístico fue de 0.01.

Adicionalmente, para el enfoque de regresión, los modelos Lasso y Elnet se mostraron superiores a los demás utilizados. Específicamente, para el modelo Lasso y Elnet, se utilizó una penalidad de 0.01. Seguidamente, los parámetros utilizados en el Boosting para el enfoque de regresión fueron determinados a partir de una grilla de prueba de entre 300-400 árboles, un N mínimo para división pequeño de 3 y un learn rate estándar de entre 0.01 y 0.1, bajo para poder corregir los errores en medición. Es importante resaltar que todos los modelos

utilizaron las mismas variables descritas en la sección de datos del presente artículo, con la excepción del modelo al que se le incluyó la variable "oficio", lo cual redujo la capacidad predictiva del mismo.

# 5 Conclusiones y Recomendaciones

De acuerdo con los resultados obtenidos, se puede concluir que los modelos con enfoque de clasificación presentaron un mejor desempeño que los de regresión para predecir la situación de pobreza en hogares a partir de los datos proporcionados. Esto se debe a que los modelos de clasificación permiten diferenciar entre clases o categorías discretas claras, mientras que los modelos de regresión buscaban predecir una variable que no sólo requiere una ingeniería de características sumamente sutil y compleja, sino que a su vez contiene errores de medición de forma recurrente. Adicionalmente, dentro de los modelos de clasificación, los que utilizaron la metodología de adaboosting (adpatative boosting) fueron los más eficientes, alcanzando un valor de F1 cercano a 0.7 de forma consistente, aunque parece que para superar tal umbral se debe probar otros modelos o usar nueva información de las personas en la base de datos. No obstante, la naturaleza de incrementos leves de la metodología y su adaptación a los errores de medición parecen haber sido claves en lidiar con los principales retos de los datos, la falta deliberada de variables clave y el desbalance de clases.

Por otro lado, el valor de F1 obtenido sigue siendo bajo para el objetivo de prever la situación de pobreza en hogares, lo que sugiere que hay un margen de mejora en la calidad y cantidad de los datos, así como en la selección y transformación de las variables. Por lo tanto, se recomienda que en futuros estudios se realice un mayor trabajo de ingeniería de características, que consiste en crear, seleccionar o modificar las variables explicativas para mejorar el rendimiento de los modelos. Asimismo, se recomienda explorar más modelos de regresión y comparar sus resultados con los de clasificación. Finalmente, sería recomendable contar con el tiempo y la infraestructura computacional adecuada para lograr implementar modelos de aprendizaje profundo (redes neuronales), que podrían captar mejor las relaciones no lineales entre las variables y compensar la falta de información relevante en la base de datos. Estas acciones podrían aumentar la precisión y la recuperación de los modelos y, por ende, la capacidad de identificar correctamente a los hogares pobres.

# 6 Bibliografía


Christianensen, L., Lanjouw, P., & Luoto, J. (2012). Small area estimation-based prediction methods to track poverty: Validation and applications. *Journal of Economic Inequality, 10*(2), 267-297.

Dang, H., & Lanjouw, P. (2018). Updating poverty estimates at fequent intervals in the absence of consumption data methods and ilustation with reference. *World Bank Policy Research*.

Elbers, C., Lanjouw, J., & Lanjouw, P. (2003). Micro-level estimation of poverty and inequiality. *Econometrica, 71*(1), 355-364.

Hall, O., & Ohlsson, M. (2022). A review of explainable ai in the satellite data, deep machine learning, and human poverty domain. *Patterns*(3).

Mathiassen, A. (2013). Testing prediction performance of poverty models. Empirical evidence from Uganda. *Review of Income and Wealth*, 91-112.

Muñeton, G., & Manrique, L. (2023). Predicting Multidimensional Poverty with Machine Learining Algorithms: An Open Data Source Approach Using Spatial Data. *Social Sciencias, 12*(296).

Newhouse, D., & et al. (2014). How survey-to-survey imputation can fail. *World Bank Policy Research*.

Nussbaum, Martha C. 2001. Women and Human Development: The Capabilities Approach. Cambridge: Cambridge University Press, vol. 3.

Pave, T., & Stender, N. (2017). Is Random Forest a Superior Methodology for Predicting Poverty? An Empirical Assesment. *Poverty & Public Policy*.

Pokhriyal, N., Zmabrano, O., & Linares, J. (2022). *Estimating and Forecasting Income Poverty and Inequality in Haiti.* Whashington D.C: Inter-American Development Bank.

Sen, Amartya. 1985. Commodities and Capabilities. Oxford: Oxford University Press.

Sen, Amartya. 1992. Inequality reexamined. Cambridge: Harvard University Press.

Sen, Amartya. 1999. Development as Freedom. New York City: Anchor Books.

Sen, Amartya. 2017. Collective Choice and Social Welfare. Cambridge: Harvard University Press.

# 7 Anexos

**Gráfico 1**

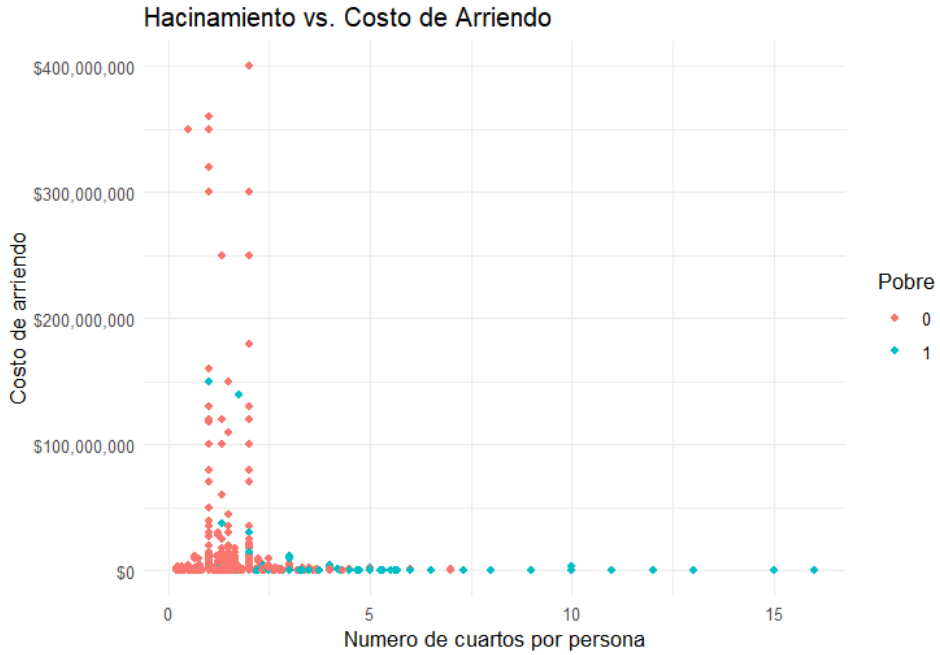

**Gráfico 2**

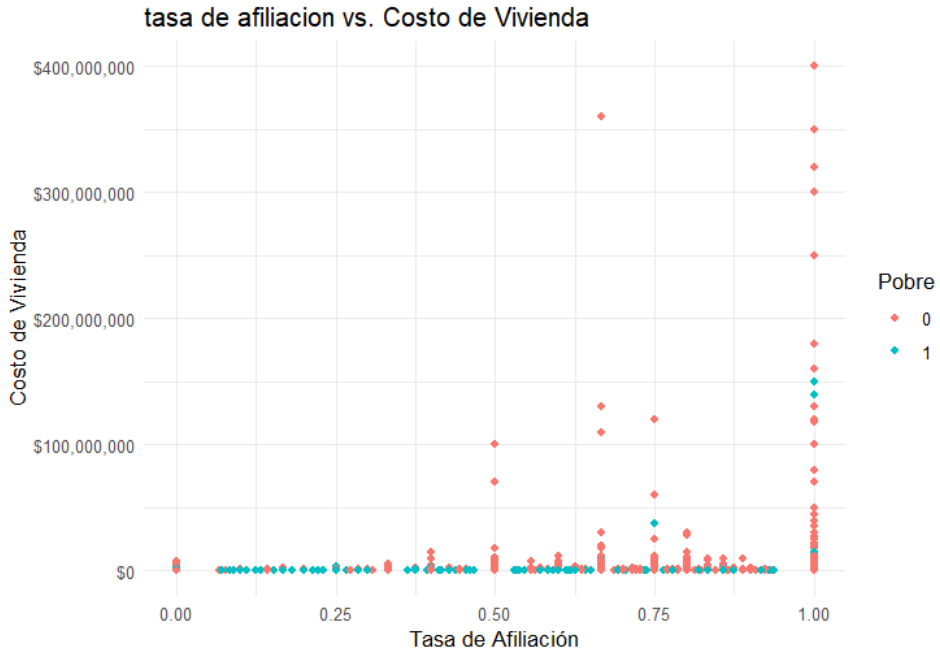


**Gráfico 3**

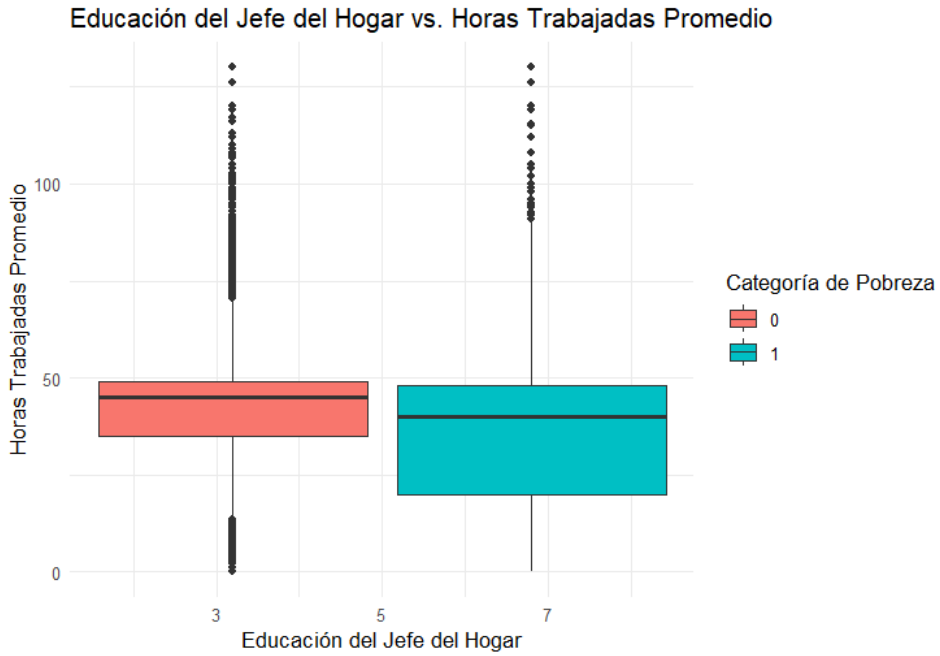

**Gráfico 4**

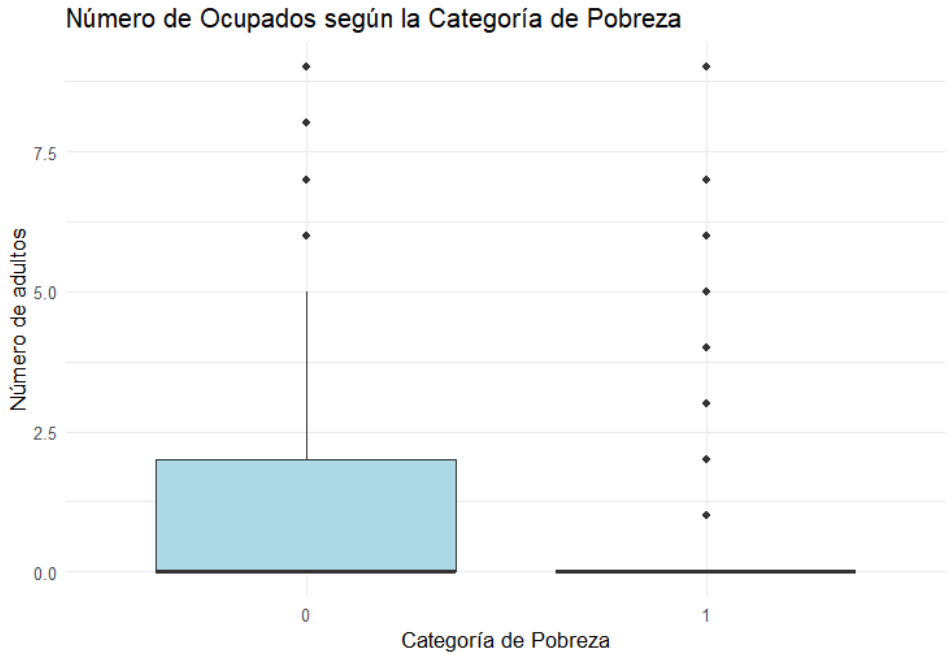


**Gráfico 5**