Taller 2 - Predicción de Pobreza en Colombia

Harold Stiven Acuña

José David Cuervo
José David Dávila
César Augusto Alfaro

13 de abril de 2025

Resumen

Este documento presenta el análisis de datos y la implementación de modelos de clasificación para la predicción de la pobreza en Colombia.

Palabras clave: pobreza, clasificación, aprendizaje automático
Clasificación JEL: J31, C53, J16

Repositorio GitHub:
https://github.com/alfarocesar/BDML_Predicting_Poverty_Equipo8

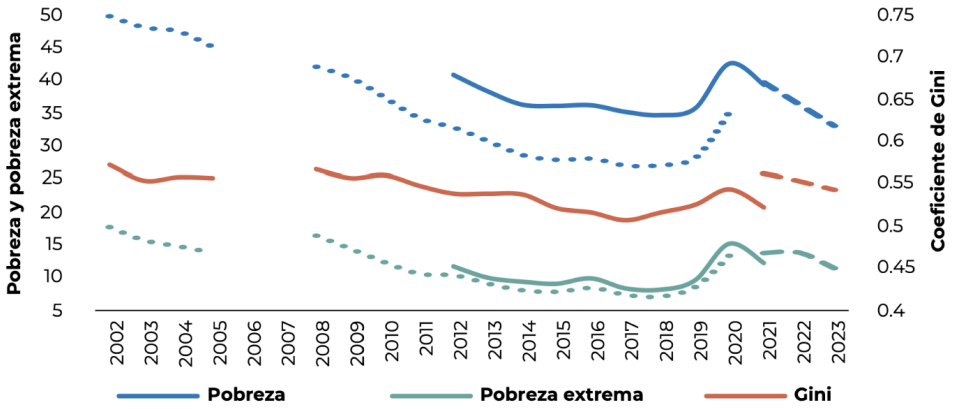
1. Introducción

La pobreza continúa siendo una de las principales barreras para el desarrollo económico y social en Colombia. A pesar de los avances logrados en las últimas décadas, persisten marcadas desigualdades territoriales, sociales y económicas que afectan especialmente a zonas rurales, hogares con jefatura femenina, y comunidades indígenas y afrodescendientes (Banco Mundial, 2024).

Una medición precisa, eficiente y oportuna de la pobreza es clave para diseñar políticas públicas que sean no solo efectivas, sino también costo-eficientes. Sin embargo, los métodos tradicionales de medición —basados en encuestas extensas— implican altos costos y tiempos prolongados de recolección, procesamiento y análisis. En respuesta a esta limitación, el uso de técnicas de *machine learning* ofrece una alternativa prometedora: construir modelos que permitan predecir la condición de pobreza de los hogares utilizando un número reducido de variables y, en consecuencia, realizar evaluaciones más rápidas y baratas.

Este proyecto busca desarrollar modelos de clasificación binaria que permitan identificar si un hogar colombiano se encuentra en condición de pobreza, utilizando microdatos del DANE y la Misión MESE, a nivel de hogar e individuo. La Figura 1 ilustra la evolución reciente de la pobreza en el país, motivando así la necesidad de mejorar las herramientas de diagnóstico.

Figura 1. Evolución reciente de la pobreza monetaria en Colombia.



Fuente: DANE, GEIH 2002-2023.

Nota: Las líneas de pobreza y pobreza extrema se actualizaron en 2019 y hay una interrupción metodológica a partir de 2021.

A lo largo del documento se evalúan diferentes algoritmos de clasificación (como regresión logística, árboles de decisión, random forest, entre otros) y se presenta una comparación sistemática de su desempeño. El modelo con mejor rendimiento logró una puntuación F1 de **[valor]**, utilizando solo **[número]** variables, lo que representa un avance en términos de precisión y simplicidad. Este modelo fue seleccionado como la base para las predicciones entregadas en Kaggle.

Además de identificar el mejor algoritmo, se discute la importancia relativa de las varia-

bles predictoras, destacando aquellas con mayor capacidad explicativa. Estos hallazgos permiten no solo mejorar la focalización de políticas sociales, sino también abrir camino para sistemas de monitoreo más ágiles y adaptativos.

2. Datos

2.1. Adecuación de los datos

Los datos utilizados en este estudio provienen de DANE y la misión .Empalme de las Series de Empleo, Pobreza y Desigualdad (MESE)". Estos datos son adecuados para resolver el problema de predicción de pobreza por varias razones. Primero, contienen información detallada a nivel de hogar e individuo, lo que permite capturar la heterogeneidad socioeconómica de la población colombiana. Segundo, incluyen la variable objetivo de interés (Pobre) que está correctamente definida según el criterio oficial del DANE: un hogar es clasificado como pobre si su ingreso per cápita es menor a la línea de pobreza establecida. Tercero, contienen variables sobre composición demográfica, características laborales, educativas y de vivienda que son teóricamente relevantes para explicar la pobreza.

Para este análisis, disponemos de cuatro conjuntos de datos divididos en entrenamiento y prueba a nivel de hogar e individuo: `train_hogares.csv`, `train_personas.csv`, `test_hogares.csv` y `test_personas.csv`. Esta división permite realizar predicciones fuera de muestra, lo que representa un desafío adicional ya que algunas variables presentes en los datos de entrenamiento están ausentes en los de prueba, simulando un escenario real donde no toda la información está disponible al momento de realizar predicciones.

Verificamos que la variable objetivo *Pobre* esté correctamente definida siguiendo la metodología del DANE, mediante la comparación con cálculos propios:

$$\text{Pobre} = I(\text{Ingpcug} < \text{Lp}) \tag{1}$$

Al realizar esta validación, obtuvimos una coincidencia del 100 % con la variable original, confirmando su correcta definición. La distribución de la variable objetivo reveló un desbalance importante: el 80 % de los hogares se clasifican como no pobres, mientras que el 20 % son considerados pobres, lo que requirió estrategias específicas durante el modelamiento.

Incluir tabla: Distribución de la Variable Objetivo (Pobreza)

2.2. Construcción de la muestra

El proceso de construcción de la muestra involucró varios pasos clave para garantizar datos limpios y adecuados para el modelamiento:

2.2.1. Análisis de la base train_hogares y comparación con test_hogares

Se inició el trabajo con una revisión detallada de las variables disponibles en la base de entrenamiento de hogares (`train_hogares`), comparándolas con las de la base de prueba (`test_hogares`). Este paso fue crucial para identificar variables que, al estar ausentes en la prueba, debían eliminarse del análisis o de la base consolidada. En particular, se eliminaron las siguientes variables porque contenían información directa del ingreso o porque solo existían en la base de entrenamiento:

`Ingtotug, Ingtotugarr, Ingpcug, Indigente, Npobres, Nindigentes`

Estas variables, aunque relevantes, no estaban disponibles para predicción en la base de prueba, por lo tanto, su inclusión habría implicado una filtración de información inadecuada. Además, se eliminaron variables que no aportan al análisis predictivo como:

`Directorio, Secuencia_p, Mes, P5130, Fex_c, Fex_dpto`

Las variables `P5100` y `P5140` fueron tratadas como gastos que pueden legítimamente tomar valor cero, por lo cual los valores faltantes se imputaron con ceros.

2.2.2. Análisis de la base train_personas y comparación con test_personas

Se realizó una limpieza similar en las bases de individuos. Se eliminaron variables que no están disponibles en ambas bases o que no describen características estables del hogar. En este grupo se encuentran identificadores (`Directorio, Secuencia_p`), variables temporales (`Mes`), ponderadores (`Fex_c, Fex_dpto`) y una lista extensa de variables relacionadas con ingresos individuales, que no están disponibles en la base de prueba.

Además, se excluyeron variables que no aportaban valor predictivo directo a nivel hogar, como `Orden`. Para las variables categóricas binarias que solo tomaban valores 1 y NA (como `Pet, Oc, Des, Ina`), los valores faltantes fueron transformados en ceros, con el fin de estandarizar su uso en los conteos que alimentan la base de hogares.

2.2.3. Construcción de nuevas variables para consolidar la base de hogares

Utilizamos la variable `id` como clave para unir las bases de datos de hogares e individuos. Sin embargo, esto requirió un paso previo de agregación para transformar la información a nivel de individuos en variables a nivel de hogar:

- **Conteos:** Se contó el número de miembros que cumplían ciertas condiciones (por ejemplo, número de ocupados, hombres, mujeres, menores de edad, personas en edad de trabajar, desocupados e inactivos).
- **Proporciones:** Para variables categóricas como afiliación a seguridad social (`P6090`) o deseo de más horas (`P6240`), se calcularon proporciones sobre el total de personas en edad de trabajar (`Pet = 1`).

- **Estadísticas demográficas:** Se calcularon edad promedio, máximo nivel educativo en el hogar, promedio de horas trabajadas usando variables como `P6040`, `P6210s1` y `P6800`.
- **Indicadores de seguridad social:** Proporción de afiliados, cotización a pensión.
- **Características laborales:** Proporciones por tipo de ocupación, tamaño de empresa, actividades adicionales.
- **Características específicas del jefe de hogar:** Edad, sexo, nivel educativo y situación laboral.

Esta agregación nos permitió generar 37 nuevas variables derivadas que capturan la estructura y características socioeconómicas del hogar, enriqueciendo significativamente el conjunto de datos. El enfoque utilizado garantiza que las variables incorporadas describan características estructurales del hogar y puedan ser calculadas en la base de prueba, utilizando únicamente la información disponible en ambas bases.

2.3. Limpieza de datos y tratamiento de valores faltantes

El análisis de valores faltantes reveló patrones importantes:

- La mayoría de variables tienen menos del 15 % de valores faltantes, siendo procesables con técnicas de imputación.
- Identificamos variables con alta proporción de valores faltantes (>33 %) como `jefe_tiempo_traba` y varias proporciones de características específicas que fueron candidatas a eliminación.
- Las variables relacionadas con ingresos complementarios presentaban patrones de valores faltantes no aleatorios, siendo más frecuentes en hogares no pobres.

Como parte de la limpieza, se eliminaron variables con más del 33 % de valores faltantes en la base consolidada. Esta decisión se basó en las buenas prácticas revisadas en clase y en los tutoriales de preprocesamiento de datos.

Para las variables restantes con valores faltantes, se adoptaron las siguientes estrategias:

- **Imputación con la mediana:** Para variables numéricas como `promedio_horas_trab`, `prop_cotiza_pension`, `prop_actividad_adicional` y `prop_desea_mas_horas`, se imputó la mediana dentro de cada base (train o test).
- **Imputación con la moda:** Variables categóricas como `max_nivel_educativo` y discretas como `max_años_educ` fueron imputadas con su moda, debido a su naturaleza y su importancia como predictores.

- **Para variables como `prop_afiliados_ss` y `prop_busca_trabajo`:** También se utilizó la mediana para la imputación.

Estas decisiones fueron implementadas en el script `03_data_cleaning_Fast.R`.

2.3.1. Manejo de variables categóricas

Un análisis especial se realizó para la variable `Oficio`, que presentaba más de 80 categorías diferentes. Mediante un análisis de asociación con la variable objetivo, agrupamos los oficios en tres categorías según su relación con la pobreza:

- **Grupo 1:** Oficios con baja tasa de pobreza (promedio 17 %)
- **Grupo 2:** Oficios con tasa media de pobreza (promedio 37 %)
- **Grupo 3:** Oficios con alta tasa de pobreza (promedio 59 %)

Esta agrupación simplificó el modelamiento y mejoró la interpretabilidad manteniendo la relevancia predictiva.

2.4. Análisis descriptivo

2.4.1. Distribución de la pobreza

La variable objetivo presenta un desbalance notable: el 79.9 % de los hogares se clasifican como no pobres y el 20.1 % como pobres, con un ratio de desbalance de aproximadamente 4:1.

Incluir tabla: Distribución de la Variable Objetivo (Pobreza)

2.4.2. Características por estado de pobreza

El análisis reveló diferencias significativas entre hogares pobres y no pobres:

Incluir figura: Distribución de horas trabajadas por estado de pobreza

Los hogares pobres tienden a reportar menos horas trabajadas en promedio, lo que se relaciona con condiciones laborales más precarias y menor estabilidad en el empleo.

Encontramos variables altamente correlacionadas con la pobreza, destacando:

- **Negativas (menor valor asociado a mayor pobreza):** Proporción de cotizantes a pensión (-0.51), proporción de trabajadores en empresas grandes (-0.47), y máximo nivel educativo (-0.39).

- **Positivas (mayor valor asociado a mayor pobreza):** Proporción de oficios del grupo 3 (0.43), número de menores (0.37) y proporción de inactivos (0.31).

Incluir figura: Variables con mayor correlación con la pobreza

Este análisis de correlaciones ayudó a identificar las variables más relevantes para la predicción, permitiéndonos crear un conjunto parsimonioso y predictivo.

2.4.3. Importancia de la composición familiar y laboral

El análisis chi-cuadrado para variables categóricas reveló asociaciones significativas entre pobreza y diversas características:

- La presencia de jefes de hogar ocupados reduce significativamente la probabilidad de pobreza (V de Cramér = 0.23)
- Los hogares con mayor proporción de miembros en oficios del grupo 3 (alta tasa de pobreza) tienen mayor probabilidad de ser pobres (V de Cramér = 0.31)
- El nivel educativo del jefe de hogar muestra una fuerte asociación con la pobreza (V de Cramér = 0.27)

Estos hallazgos confirman la importancia de variables relacionadas con capital humano, estructura familiar y características laborales para predecir la pobreza.

2.5. Justificación de la selección de variables

La selección final de variables para nuestros modelos se basó en tres criterios principales:

1. **Relevancia predictiva:** Utilizamos correlaciones con la variable objetivo y pruebas chi-cuadrado para identificar las variables más predictivas.
2. **Disponibilidad en datos de prueba:** Garantizamos que todas las variables seleccionadas estuvieran disponibles tanto en el conjunto de entrenamiento como en el de prueba.
3. **Parsimonia:** Buscamos un conjunto mínimo de variables que maximizara el poder predictivo.

Eliminamos variables con más del 33 % de valores faltantes y aquellas con alta redundancia (correlacionadas entre sí a más de 0.7). Para variables con correlación alta, mantuvimos la que presentaba mayor asociación con la variable objetivo.

En el caso de variables categóricas como la ocupación, optamos por transformaciones que preservaran su poder predictivo al tiempo que simplificaban el modelamiento (agrupación en 3 categorías).

Las variables finales incluyeron:

- **Características del hogar:** Número de miembros, relación de dependencia, proporción de ocupados.
- **Capital humano:** Nivel educativo máximo y del jefe de hogar, proporción de afiliados a seguridad social.
- **Características laborales:** Distribución por grupos de ocupación, horas trabajadas, tamaño de empresa.
- **Características del jefe de hogar:** Sexo, edad, ocupación.
- **Vivienda y ubicación:** Características de la vivienda, departamento.

El flujo de trabajo seguido en esta etapa responde a varios objetivos:

- **Evitar fuga de información:** Se eliminaron variables disponibles solo en el entrenamiento.
- **Garantizar la coherencia entre bases:** Se preservaron únicamente las variables comunes entre train y test.
- **Convertir la información individual en predictores agregados:** Esto permite maximizar el uso de la información disponible sin violar las restricciones impuestas por la estructura del problema.
- **Preparar una base limpia y funcional para modelado:** Las bases `train_cleaned.csv` y `test_cleaned.csv` contienen las variables seleccionadas e imputadas, listas para el entrenamiento de modelos.

Este conjunto final de variables balanceó el poder predictivo, la disponibilidad en los datos de prueba y la parsimonia, permitiéndonos construir modelos robustos y generalizables para la predicción de pobreza.

2.6. Análisis descriptivo

Exploración de la variabilidad de los datos con tablas y gráficos.

3. Modelos y Resultados

3.1. Metodología

Descripción de los algoritmos utilizados: Regresión Logística, Random Forest, etc.

3.2. Entrenamiento y validación de modelos

Explicación del proceso de entrenamiento y evaluación.

3.3. Comparación de modelos

Tabla comparativa con métricas de desempeño.

3.4. Importancia de variables

Análisis de las características más relevantes en la predicción.

4. Conclusión

Resumen de los hallazgos principales y posibles mejoras futuras.

Referencias

Anthropic (2023). Conversaciones con claude ai assistant. Utilizado para estructura de documentos LaTeX, y mejoramiento del código en R.

Banco Mundial (2024). Trayectorias: Prosperidad y reducción de la pobreza en el territorio colombiano. Technical report, Banco Mundial, Washington, D. C., Estados Unidos de América. Publicado el 3 de diciembre de 2024.

Castillo-Álvarez, G. and Sarmiento-Barbieri, I. (2024a). Data pre-processing: Visualizing and handling missing values (part 1). Notebook proporcionado como material del curso.

Castillo-Álvarez, G. and Sarmiento-Barbieri, I. (2024b). Data pre-processing: Visualizing and handling missing values (part 2). Notebook proporcionado como material del curso.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, NY, 2 edition.

OpenAI (2023). Conversaciones con chatgpt. Utilizado para estructura de documentos LaTeX, y mejoramiento del código en R.

Sarmiento-Barbieri, I. (2024). Limpieza de datos con tidyverse. Notebook proporcionado como material del curso.

Sarmiento-Barbieri, I. and Castillo-Álvarez, G. (2024a). Caret para clasificación. Material del curso BDML, Universidad de los Andes. Notebook proporcionado como material del curso.

Sarmiento-Barbieri, I. and Castillo-Álvarez, G. (2024b). Classification - cuaderno de clase. Notebook proporcionado como material del curso.

Sarmiento-Barbieri, I. and Castillo-Álvarez, G. (2024c). Problem set 2: Instrucciones del taller - predicción de pobreza en colombia. Material del curso BDML, Universidad de los Andes. Notebook proporcionado como material del curso.

Sarmiento-Barbieri, I. and Castillo-Álvarez, G. (2024d). Rúbrica de evaluación del taller - predicción de pobreza en colombia. Material del curso BDML, Universidad de los Andes. Notebook proporcionado como material del curso.

Sarmiento-Barbieri, I. and Castillo-Álvarez, G. (2024e). Uniendo bases y calculando pobreza. Material del curso BDML, Universidad de los Andes. Notebook proporcionado como guía para el taller.