Problem Set 3:

Predicting Poverty

**Integrantes:** Juan Diego Duarte[[1]](#footnote-1); Erick Julian Villabon[[2]](#footnote-2)

## **Introducción**

La problemática de la pobreza en América Latina constituye un desafío persistente, especialmente en el contexto de países en vía de desarrollo, como es el caso de países Latinoamericanos. Según datos recientes de la CEPAL, la región enfrenta una tasa de pobreza del 29%, con un 11.2% de la población viviendo en condiciones de pobreza extrema, proyectándose un aumento a 29.1% y 11.4% para el año 2023. Este fenómeno complejo se atribuye a factores como la desigualdad económica, la limitada accesibilidad a oportunidades educativas y al sistema de salud, así como una oferta laboral insuficiente. Como respuesta a esta problemática el Banco Mundial tiene una iniciativa de abordar la medición de la pobreza a través de una competencia “Pover-T tests: Precting Poverty”, la cual tiene el objetivo de construir modelos más eficientes que permitan predecir la pobreza de manera acertada.

Para le caso especifico de Colombia, predecir la pobreza adquiere una relevancia debido a la meta del Gobierno en cumplir con los Objetivos de Desarrollo Sostenible, en los cuales se especifica la erradicación de la pobreza extrema. Según cifras del DANE, para el 2022 se observó una disminución de la pobreza monetaria, alcanzando el 36,6% de la población, y la pobreza extrema se situó en un 13,8%, como parte de las estrategias para reducir la pobreza monetaria, se implementaron medidas que aumentaron el número de beneficiarios en programas sociales, sin embargo, las políticas implementadas para este objetivo tienen limitaciones debido a la información incompleta de los hogares en el país. En este contexto, competencias como “Pover-T tests” ofrece oportunidades de explorar y mejorar herramientas que permitan abordar de manera eficiente este tipo de problemáticas al acceso y costo de la información, cabe destacar la importancia de la precisión en la predicción para maximizar la eficacia de las intervenciones y políticas dirigidas a combatir este fenómeno.

De esta forma, los datos derivados de la Misión de Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP), recopilados por el DANE, se erigen como una herramienta esencial en la evaluación de la pobreza monetaria en Colombia. Esta iniciativa introduce ajustes significativos en la definición de la línea de pobreza y en la construcción del ingreso familiar, permitiendo así la comparación a nivel regional al adoptar métodos de medición en línea con otros países de la región. Por lo que, este estudio utiliza datos provenientes del MESEP a nivel hogar e individual que después de realizar el tratamiento completamos datos a nivel hogar para lograr el objetivo de predecir la pobreza de los hogares. Este estudio se centra en predecir la pobreza de los hogares a través de diferentes metodologías para evaluar cual compone la mejor predicción.

El siguiente trabajo de desarrolla de la siguiente manera: la sección 2 proporciona información sobre la recopilación, tratamiento y descripción de los datos. Posteriormente, la sección 3 presenta las especificaciones y los modelos utilizados en los ejercicios de predicción. Por último, la sección 4 concluye.

## **Datos**

Los datos provienen de la encuesta "Medición de Pobreza Monetaria y Desigualdad 2018", construida a partir de la GEIH de 2018, fundamental para las estimaciones del Índice de Pobreza Multidimensional (IPM) y la clasificación de hogares en pobres y no pobres según la línea de pobreza colombiana establecida en 2018. Utilizada por el DANE y como insumo clave para la MESEP. Los datos correspondientes parten de las bases de entramiento y de prueba, divididas a nivel hogar y personas, por lo que contamos con cuatro bases las cuales requirieren un emparejamiento a nivel de hogar para obtener dos bases de datos (entrenamiento y prueba) a nivel hogar. Este estudio se centra en la unidad de observación del hogar para analizar características propias de cada hogar.

Debido a la importancia de tener variables que cuenten con la totalidad de los valores y no perder observaciones importantes al momento de realizar predicciones, se realizó una serie de pasos para completar la información de cada hogar e incluir variables que consideramos esenciales para calcular los ingresos de los hogares. Dado que las bases de datos a nivel individual poseen un mayor número de variables y observaciones que los datos a nivel de hogares, se realizó una unión utilizando la llave de hogar como referencia. En otras palabras, en la base de datos de personas se identificaron los miembros de los hogares a través de una variable clave común, facilitando la combinación de información. Este proceso permitió, en primer lugar, obtener los datos faltantes a nivel de hogar y, en segundo lugar, adquirir más variables que potencialmente explican los ingresos del hogar. Esto, a su vez, mejora la precisión en la predicción de ingresos y contribuye a una clasificación más certera.

La muestra cuenta con un total de 231,128 observaciones, de las cuales 66,168 hogares se asignaron a la muestra de prueba. Con el objetivo de no perder observaciones, se llevaron a cabo imputaciones de valores faltantes utilizando datos a nivel de personas. La imputación se basó en el promedio de las personas que integran cada hogar. Para las variables categóricas con datos faltantes, se asignó el valor cero, ya que la comparación posterior tomará el valor de uno si cumple con la característica de la categoría y cero en caso contrario. Dada la considerable cantidad de valores perdidos en la variable de ingreso total, se optó por utilizar la variable ingreso total de la unidad de gasto, la cual representa la suma de los ingresos dentro del hogar y no presenta valores faltantes.

La figura 1 muestra la descripción de las variables del estudio, donde se discrimina la muestra de testeo y entrenamiento, y se realiza la separación de pobres y no pobres….

## **Modelos y Resultados**

En este ejercicio se tenia un riesgo alto de sobre ajustar los modelos a la base entrenamiento, por tal razón decidimos partir los datos de entrenamiento en dos, tomando el 75% como una nueva base de entrenamiento la cual es submuestra de la base original y una subbase de prueba del 30% que de igual forma es submuestra de la prueba original. Los mejores modelos se cargan a Kaggle y son probados con la base de prueba original.

**Referencias**

CEPAL. (2023). Panorama Social de América Latina y el Caribe 2023: la inclusión laboral como eje central para el desarrollo social inclusivo.

DANE. (2023). Publicación de pobreza monetaria extrema y pobreza monetaria, Declaración Comité de Expertos en Pobreza.

1. Código: 202011999 [↑](#footnote-ref-1)
2. Código: 201815677 [↑](#footnote-ref-2)