Universidad de los Andes

Maestría en Economía Aplicada

Andres Felipe Martinez - 202121008

Angela Paola Morales Guio – 201015503

Oscar Cortes - 200222692

Repositorio: <https://github.com/paolamguio/Problem_Set_2>-G16

Problem Set 2: Predicting Poverty

MECA 4107

# Introduction

A pesar de los avances del país en la reducción de la pobreza, en el año 2018 Colombia se ubicó en términos de pobreza moderada en el puesto 11 de 14 países de Latinoamérica (CEPAL, 2019). Lo anterior resalta la importancia de evaluar el impacto de los programas sociales del Gobierno Nacional enfocados a la reducción de la pobreza, que requieren principalmente de la identificación de los hogares y personas pobres en las diferentes regiones del territorio colombiano para que la entrega de ayudas monetarias y en especie sea dirigidas a quien más lo necesite y les permita superar la línea de pobreza monetaria.

El principal instrumento de focalización individual en Colombia ha sido el Sistema de Identificación de Potenciales Beneficiarios (Sisbén), cuya cuarta versión se comenzó a implementar desde el año pasado. Sin embargo, el sistema más rápido y fácil de clasificación de la pobreza es la estratificación socioeconómica, donde los estratos 1 y 2 se podrían considerar como pobres. El Sisbén en su primera versión clasificaba a las personas pobres a través de un índice de 0 a 100 basado en la exclusión productiva de las personas (ingresos) y se aplicaba a nivel nacional. Posteriormente en su segunda versión, aunque se mantuvo el índice de 0 a 100, se cambió la forma de clasificar las personas a través de la exclusión social (privaciones) de las personas y se introdujo la aplicación del índice a nivel urbano y rural. En la tercera versión, se eliminó la variable del estrato socioeconómico de la persona en su clasificación de la pobreza, se mantuvo como base para la clasificación la exclusión social y se amplió la aplicación del índice de manera diferenciada a las 14 ciudades principales del país.

Como se ha mencionado anteriormente, en los últimos 20 años se ha modificado la forma de identificar a las personas más pobres del país por la importancia de la focalización en la asignación del gasto social para reducir la pobreza. Sin embargo, un error de cálculo en el año 2010 conllevó a la exclusión de 4 millones de pobres en Colombia y durante la crisis del Covid 19, el Departamento Nacional de Planeación (DNP) identificó a 3 millones de hogares pobres que nunca habían recibido una transferencia directa del Estado (Fedesarrollo, 2021). Para evitar que esto vuelva a pasar, DNP ha implementado el Sisbén IV, que tiene un nuevo enfoque conceptual: inclusión social y productiva de las personas para corregir las fallas de exclusión e inclusión de las personas pobres.

En concordancia con lo anterior, para la definición de los modelos de predicción en la identificación de un hogar o persona pobre, se estudiará la metodología del instrumento de focalización individual del Sisbén IV implementado por DNP recientemente y los resultados presentados por el DANE sobre pobreza monetaria 2021.

**Sisbén IV:**

# Datos

El objetivo del trabajo se enfoca en construir modelos predictivos de pobreza a nivel del hogar, para esto se utilizarán datos de la XXXXXXXXXX GEIH año 2018 que contiene información necesaria para el cálculo de la pobreza monetaria por hogar.

Estos datos se encuentran segmentados en dos bases de datos, por un lado se tiene la base que cuenta con la información de la variable pobreza e ingreso, con la cual se entrenarán los modelos y por el otro lado, se tiene la base de datos que no cuenta con estas variables y por lo tanto se realiza la predicción de los hogares pobres sobre esta última considerando el mejor modelo entrenado.

Partiendo de la base de entrenamiento, se realizaron los siguientes ajustes:

1. Se colapsa la base de datos de personas para obtener información agrupada por hogar de las variables relevantes y se consolida con la base de datos de hogares.
2. Se procede a realizar ajuste de los missing values de cada variable, determinando dejar cero en estos valores para las variables: cuota de amortización, arriendo, horas trabajadas promedio, porcentaje de trabajo formal y tasa de desempleo. Lo anterior, debido a que los hogares que no reportaron cuota de amortización y arriendo se asumen que es porque no pagan estos conceptos, las personas que no tienen horas de trabajo, se asume que no se encuentran ocupadas, lo que también sucede para la tasa de trabajo formal, pata la tasa de desempleo, corresponde a hogares en que las PET están inactivas, y por lo tanto se asumió una tasa de desempleo cero ya que no se encuentran desempleadas.

Para la variable años de educación promedio por hogar, se encontraron cuatro valores en missing value, se procedió a eliminarlos al no ser un número de hogares significativos.

Para la base de testing, se replicaron los ajustes mencionados anteriormente, sin embargo, se encontró un missing value para la educación promedio por hogar y se procedió a imputar el valor considerando la educación promedio de los hogares que tenían individuos con la misma edad promedio. Lo anterior se realizó considerando el objetivo del trabajo el cual es la proyección de hogares pobres y no pobres, más allá de la inferencia estadística que se pueda determinar por los datos.

Se eligen las variables XXXXXX.

Inicialmente, la base general de la GEIH contiene 32177 observaciones y una vez realizando el filtro mencionado anteriormente se reduce a 16397 observaciones.

Nota: El proceso anterior se encuentra en los archivos 1. data\_cleaning.R y 2. data\_cleaning2 incluidos en la carpeta 2. Scripts.

**Tabla XX**. Estadísticas descriptivas

| Descriptivas principales por hogar | | Pobre | |
| --- | --- | --- | --- |
| Si | No |
| No. Hogares | 164,955 (100%) | 33,022 (20%) | 131,933 (80%) |
| No. Personas | 3.00 (2.00, 4.00) | 4.14 (2.03) | 3.08  (1.64) |
| Edad promedio | 34  (24, 48) | 31  (16) | 39  (17) |
| Mujer como jefe de Hogar | 68,684  (42%) | 15,501  (47%) | 53,183  (40%) |
| No. De hijos | 1.00  (0.00, 2.00) | 1.74  (1.32) | 1.02  (1.01) |
| Años educación | 10.5 (7.5, 13.0) | 8.0  (3.5) | 10.9  (4.1) |
| No. Personas con trabajo formal |  |  |  |
| 0 | 93,346 (57%) | 29,060  (88%) | 64,286  (49%) |
| 1 | 51,434 (31%) | 3,822  (12%) | 47,612  (36%) |
| 2 | 17,320 (10%) | 134  (0.4%) | 17,186  (13%) |
| Tipo de vivienda |  |  |  |
| Propia, totalmente pagada | 62,118 (38%) | 9,051  (27%) | 53,067 (40%) |
| Propia, la están pagando | 5,616  (3.4%) | 537  (1.6%) | 5,079  (3.8%) |
| En arriendo o subarriendo | 64,341  (39%) | 14,435  (44%) | 49,906 (38%) |
| En usufructo | 25,000  (15%) | 5,169  (16%) | 19,831  (15%) |
| Posesión sin titulo | 7,717  (4.7%) | 3,786  (11%) | 3,931  (3.0%) |
| No. de personas por cuarto | 1.50  (1.00, 2.00) | 2.25  (1.15) | 1.60  (0.67) |
| Ingreso Total | 1,187,717  ($600,000, $2,200,000) | $554,163  ($467,797) | $2,113,626  ($2,510,813) |
| Horas trabajadas | 44  (32, 49) | 34  (22) | 40  (19) |
| tasa ocupación | 0.50  (0.33, 1.00) | 0.43  (0.31) | 0.62  (0.33) |
| tasa desempleo | 0.00  (0.00, 0.00) | 0.17  (0.31) | 0.07  (0.20) |
| tasa participación | 0.67  (0.50, 1.00) | 0.54  (0.31) | 0.67 (0.32) |
| 1 Mean (SD); n (%) |  |  |  |

Fuente: Elaboración propia

De la tabla XX se infiere que del total de 164.955 observaciones (hogares), el 20% corresponde a hogares clasificados como pobres y el 80% restante a hogares no pobres, el 57% de los hogares en promedio sus individuos no cuentan con trabajo formal, el 42% de los hogares tienen como jefe de hogar a una mujer, se tiene una edad promedio de 34 años y 10,5 años de estudio en promedio, sin embargo, respecto a este último dato se evidencia brecha entre los hogares pobres y no pobres, teniendo como años promedio de educación 8 y 11 respectivamente, situación que esta enlazada con los ingresos promedio de los hogares como se observa en la siguiente gráfica:

**Gráfico XXX.** Dispersión entre años de educación

e ingreso total entre hogares pobres y no pobres

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Fuente: Elaboración propia

Adicionalmente, se observa que a mayores años de educación disminuye la participación de los hogares pobres, siendo más notoria la brecha de ingresos, la cual también es evidente entre los hogares pobres y no pobres con igual número de años de educación.

Aunado a esto, se considera que existe una relación entre ser pobre, tener más hijos, trabajar menos horas a la semana, no contar con trabajo formal y no contar con vivienda propia, esto a su vez se estaría viendo reflejado en menos ingresos promedio de los hogares. Esto, se evidencia en los gráficos relacionados en el Anexo 1.

# Modelos y resultados

• Models and Results. Present the models that you are using. When selecting the best

models, argue why these are the best models for your task.

{ Classication Models. This section should include (not necessarily in this order):

\* A detailed explanation of the final chosen model. The explanation must include how the model was trained, hyper-parameters selection, and other relevant information.

\* Include comparisons to at least 5 other models. You can compare them in terms of ROC, AUC, False Positives, and/or False Negatives.

**Tabla xx.** Resultados modelos

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Modelo** | **Descripción** | **ROC** | **AUC** | **Sensibilidad** | **Especificidad** | **SI** | | | **NO** |
| **Xx** | xxxx |  |  |  |  | SI |  | xxx | Xx |
| NO |  |
| **Xx** | xxxx |  |  |  |  |  | | |  |
|  |  |  |  |  | | |  |
| **xxx** | xxx |  |  |  |  |  | | |  |
|  |  |  |  |  | | |  |

Fuente: Elaboración propia

\* Describe the variables that you used in the model and a measure of their relative importance in the prediction.

\* Describe any sub-sampling approach used to address class imbalances.

{ Income regression Models. This section should include (not necessarily in this order)

\* A detailed explanation of the \_nal chosen model. The explanation must include how the model was trained, hyper-parameters selection, and other relevant information.

\* Include comparisons to at least 5 other models. Compare them in terms of MSE.

\* Convert the predicted income to a binary indicator and show the performance in terms of the False Positives, and/or False Negatives.

\* Describe the variables that you used in the model and a measure of their relative importance in the prediction.

Conclusions and recommendations. In this section, you state the main takeaways of

your work.