Universidad de los Andes

Maestría en Economía Aplicada

Grupo 16

Andres Felipe Martinez - 202121008

Angela Paola Morales Guio – 201015503

Oscar Cortes - 200222692

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

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 requiere principalmente de la identificación de los hogares y personas pobres en las diferentes regiones del territorio colombiano para la entrega de ayudas monetarias y en especie 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 del país. 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

Treat this section as an opportunity to present a compelling narrative to justify or defend your data choices, walk the reader through your reasoning of how you\cleaned" the data, and describe it accordingly with descriptive stats, graphs, etc.

At a minimum, you should include a descriptive statistics table. Use your professional

knowledge to add value to this section. Do not present it as a \dry" list of ingredients.

En el proceso de data cleaning se seleccionó la variable ocupados la cual define 1= ocupado; 0= no ocupado, de acuerdo con los datos y el directorio, se deduce que esta variable corresponde a las personas que tiene trabajo independientemente de las características de este. Esto quiere decir que cuando la variable ocupado toma el valor de 1 el individuo está empleado, por otro lado cuando la variable toma el valor de 0 el individuo puede estar desempleado y/o también puede estar inactivo, por esta razón la variable “dsi” que considera los desempleados y los no desempleados, no es una variable idónea para elegir las personas empleadas ya que los no desempleados no necesariamente tienen un trabajo sino que pueden estar inactivos.

El objetivo del trabajo se enfoca en los individuos empleados (ocu=1) y mayores de 18 años. Se elige la variable “ocu” ya que relaciona todas las personas que tienen una ocupación y por lo tanto devengan un ingreso salarial, el enfoque en analizar el ingreso de los individuos que se encuentran empleados y esta variable recoge esa información.

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 el archivo 2\_data\_cleaning.R incluido en la carpeta 2. Scripts.

# 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.