# Predicción de pobreza

Carlos Ayala
Mateo Contreras
Federico Meneses

link repositorio: https://github.com/Mateocontrerass/Predicting-Poverty

## Introducción

El primer objetivo de desarrollo sostenible propuesto por las Naciones Unidas en 2015 es la erradicación de la pobreza[1]. Este objetivo es especialmente importante dado que indica explícitamente la cantidad de personas que no pueden satisfacer sus necesidades básicas (en términos calóricos y de consumo). Una de las formas más populares para medir la cantidad de población que vive en esta condición es la pobreza monetaria, lo cuan establece que una persona se puede clasificar en situación de pobreza monetaria si su ingreso es menor a cierto umbral, el cual es $354,031[2] para Colombia. Así, la predicción de pobreza se hace imperante dados los retos sociales que con daba vez más relevancia se presentan en la sociedad colombiana.

El problema radica en determinar si una persona califica como pobre o no pobre es una tarea costosa porque requiere de una evaluación por parte de una entidad administrativa que registre sus ingresos y con ello se pueda determinar su situación. En este proyecto busca saltarse la necesidad de ejecutar encuestas y poder predecir los ingresos de las hogar, y consecuentemente establecer si el hogar de esa persona es pobre. Para ello, se extrae infromación de la Gran Encuesta Integrada de Hogares (GEIH) del año 2018, la cual es la base de datos más grande que se tiene con respecto a hogares y personas con información como sexo, educación, ubicación, características de la vivienda y de condiciones de vida que permite caracterizar de mejor manera a las personas.

El modelo de predicción seleccionado busca minimizar la probabilidad de establecer como no pobre un hogar en situación de pobreza, motivo por el cual se utiliza este criterio como principal forma de selección del modelo. Así mismo, el modelo no deja de lado la predicción como verdadero pobre, donde se busca no sobre estimar la pobreza con el objetivo de tener presente el costo asociado a una mayor cantidad de pobreza en futuros programas sociales. En este orden de ideas, el modelo escogido es un modelo de regresión logística, con el cual las predicciones de pobreza cumplen los criterios mencionados, y presenta mejor desempeño que las demás especificaciones planteadas.

## Datos

### Preprocesamiento

La primera modificación que se le hizo a las bases de datos fue cambiar las variables categóricas a (n-1) variables binarias para evitar la trampa de la variable dicótoma. Esto con el objetivo de facilitar la implementación de los modelos. Por otro lado, otro elemento clave

[1] UN. (2015). Objetivo 1: Poner fin a la pobreza en todas sus formas en todo el mundo.
[2] DANE.(2022). Pobreza monetaria y grupos de ingreso en Colombia.

en el procesamiento de los datos fue la imputación de valores nulos. Para este procedimiento se utilizó un algoritmo de imputación múltiple basado en XGBoost, Boostraping y *Predictive Mean Matching* con la librería ***mixgb.*** Se utilizó este método dado que se buscó un modelo más complejo que la imputación de las medias de la variable en caso de baja *Skewness* y mediana para alta *Skewness.* También se omitió la implementación de KNN debido al tamaño de la base de datos y la cantidad de memoria y tiempo que requiere ese algoritmo.

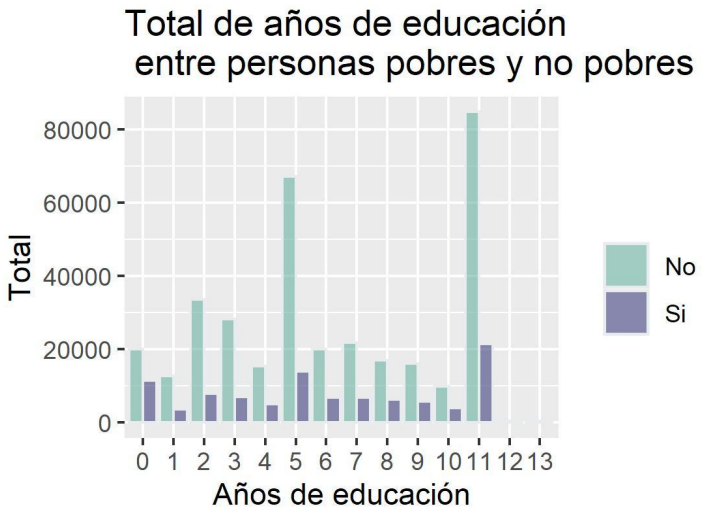
**Análisis descriptivo**

En primer lugar, es importante saber como está la distribución de personas que hacen parte del grupo de interés, en este caso en situación de pobreza.

| label | variable | Pobre | | Total |
|---|---|---|---|---|
| | | No | Si | |
| Sexo | Hombre | 193506 (75%/48%) | 62800 (25%/46%) | 256306 (47%) |
| | Mujer | 212967 (74%/52%) | 73668 (26%/54%) | 286635 (53%) |
| | Total | 406473 (75%) | 136468 (25%) | 542941 (100%) |

*Tabla 1. Tabla cruzada entre Sexo y Pobre. Elaboración propia.*

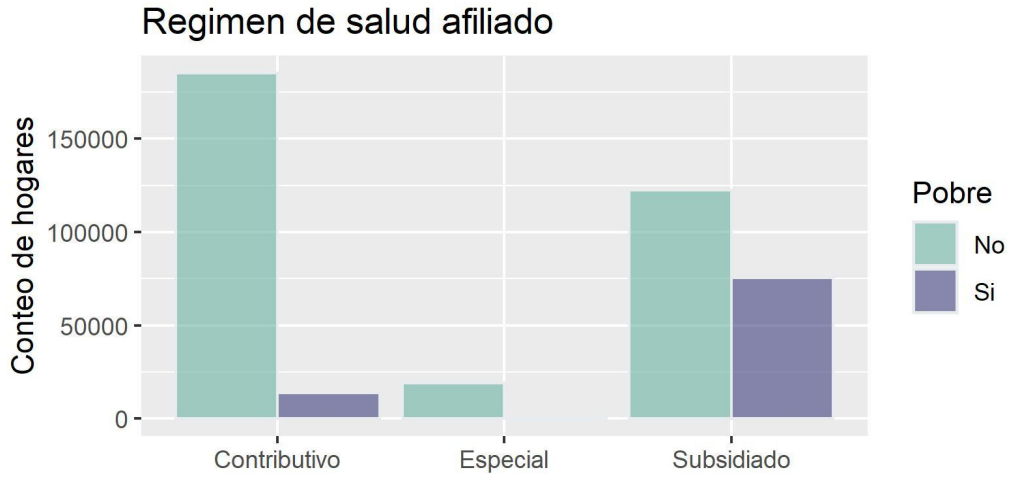
En la *tabla 1* se presenta una tabla cruzada entre las variables Sexo y Pobre. De la interpretación de la última fila se puede determinar que el 75% de la muestra no está marcada como Pobre, mientras que el 25% si lo está. Esto anterior indica que la base de datos está desbalanceada. Por otro lado, la última columna nos indica que la información proviene de una población con 47% de hombres y 53% de mujeres. Los otros valores de la tabla muestran la proporción de hombres y mujeres pobres y no pobres. Sin embargo, se mantiene la proporción 75/25 de no pobre y pobre entre ambos géneros.




*Grafica 1. Total de años de educación entre personas pobres y no pobres. Elaboración propia.*

Ahora bien, es común identificar en la literatura que al tener más años de educación se mejoren los ingresos y consecuentemente se llegue a una situación mejor a la de pobreza[3]. En la *gráfica 1* se plantea un histograma que muestre los años de estudio para cada persona y discriminar los grupos entre pobres y no pobres. Si bien, se puede ver que para cualquier año de estudio hay más personas no pobres que pobres, esto puede ser efecto del desbalance entre clases. La tendencia que se puede plantear es entre los años 5 y 10, donde la incidencia en pobreza tiende a reducirse.

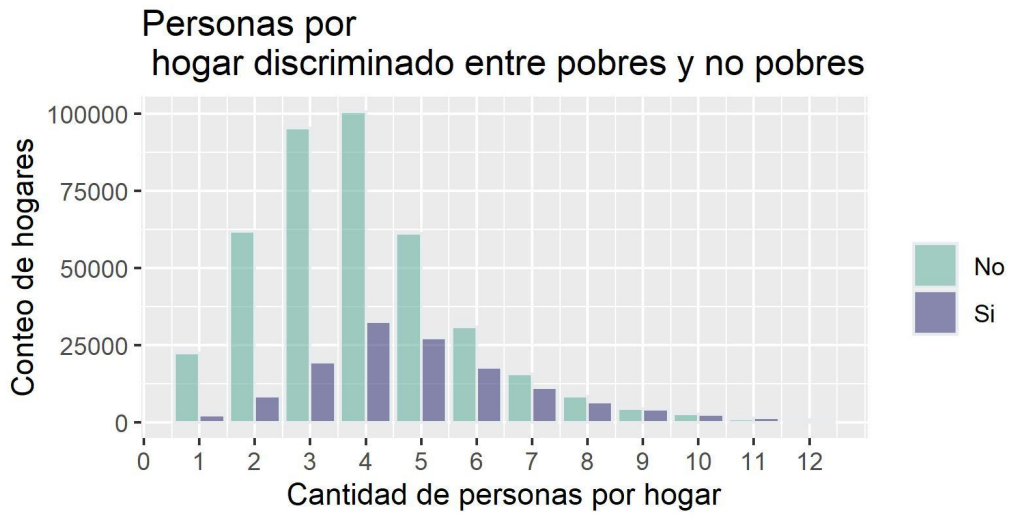
Por otro lado, se consideró interesante ver como es la distribución de las personas pobres y no pobres en los regímenes de seguridad social. En la *gráfica 2* se puede apreciar como las personas pertenecientes al régimen contributivo son en su mayoría no pobres. Se puede plantear que las personas pobres que también hacen parte de este régimen tienen una gran cantidad de personas a su cargo, lo que diluye el ingreso per cápita del hogar y consecuentemente los identifica como pobres. Por su parte, las personas que pertenecen a un régimen especial como el de las fuerzas militares o educativo no están clasificadas como pobres. Por último, se tiene que en el régimen subsidiado hay un mejor balance entre ambas clases, se plantea que está subestimada la participación de las personas pobres en este régimen dado el desbalance de la base de datos.




*Grafica 2. Distribución de personas pobres y no pobres en los regímenes de seguridad social. Elaboración propia.*

Ahora bien, también se considera importante ver el comportamiento de la variable de interés (pobreza) sobre la cantidad de personas por hogar. Esto debido a que entre más personas haya por hogar, el ingreso del mismo se ve diluido en mayor medida y consecuentemente puede llegar a un punto donde todas las personas sean clasificadas en situación de pobreza.

[3] Lagarza, O. & Villezca, P. (2006). Efecto de la sobre-educación en el ingreso de personas con estudios de nivel superior en México.




*Grafica 3. Cantidad de personas por hogar discriminado entre pobres y no pobres. Elaboración propia.*

De la gráfica 3 se puede observar que la distribución para ambas clases se asemeja a una normal. Sin embargo, la de personas pobres está corrida hacia la derecha. También se puede observar como la proporción de personas pobres aumenta al incrementar la cantidad de personas por hogar.

**Modelos y Resultados**

Para poder predecir la condición de pobreza del hogar se hizo necesario dividir el problema en 2 partes. El primero, predecir ingreso, requirió utilizar modelos de predicción para variables continuas; el segundo, predecir pobreza, requirió utilizar modelos de clasificación que permitiera establecer si el hogar se podía considerar como pobre dadas unas características. Así, se establecen 5 modelos de predicción de ingreso, y 8 modelos de clasificación. Para clasificar se utilizaron los siguientes modelos: 1. Logit - Regla de Bayes (p=0.5), 2. Logit - gráfica (p=0.35, como resultado de observar las predicciones del modelo de forma gráfica), 3. Logit - óptimo (p=0,28), 4. Logit - SMOTE, 5. Logit - Undersampling, 6. Arbol de clasificación y 7. XGBoost de clasificación.

**Modelos de clasificación:**

Para la selección de un modelo de clasificación, se utilizaron varias posibilidades. Entre estas optamos por comparar los modelos de clasificación y predicción que se observan en la *tabla.2*. Es interesante observar que para muchos de los modelos, como el árbol o el Logit entero, tienen una buena acertividad y sensibilidad; sin embargo, muy mala especificidad, lo que puede indicar overfitting. Esto conlleva en rechazar a estos dos modelos, así como el LM o en Logit con regla de Bayes que presenta características similares. De esta manera, para quedarnos con solo un modelo de predicción y uno de clasificación, observamos que el Ridge es tan solo un poco mejor que sus los otros modelos de predicción como regularización por Lasso, Elastic Net y XG boost. Por parte de los modelos de clasificación, se escogió al umbral óptimo, ya que en características como acertividad, sensibilidad y especificidad demuestra ser más balanceado.

| Modelo | Accuracy | Sensitivity | Specificity | Base |
|---|---|---|---|---|
| Logit - Regla de Bayes | 0.79 | 0.38 | 0.93 | Clasificación |
| Logit - Gráfica | 0.78 | 0.61 | 0.83 | Clasificación |
| Logit - Umbral óptimo | 0.75 | 0.72 | 0.76 | Clasificación |
| Logit - SMOTE | 0.73 | 0.76 | 0.72 | Clasificación |
| Logit - Undersampling | 0.73 | 0.77 | 0.72 | Clasificación |
| Lasso | 0.70 | 0.79 | 0.45 | Predicción |
| Ridge | 0.71 | 0.79 | 0.45 | Predicción |
| Elastic Net | 0.70 | 0.79 | 0.45 | Predicción |
| XGboost | 0.64 | 0.64 | 0.64 | Predicción |
| LM | 0.79 | 0.97 | 0.23 | Predicción |
| Arbol | 0.80 | 0.93 | 0.39 | Clasificación |
| XGBoost | 0.64 | 0.64 | 0.64 | Clasificación |
| Logit Entero | 0.79 | 0.97 | 0.23 | Clasificación |

**Modelo clasificación: Logit-Umbral óptimo**

Un modelo logit está diseñado para realizar predicciones en un rango entre cero y uno, por lo que es necesario establecer una regla de asignación mediante la cual se determine si el individuo es pobre o no. Así, surge la necesidad de establecer dicho umbral, y determinar si ese umbral es óptimo en términos de lo que nos interesa: reducir la probabilidad de predecir como no pobre un hogar que es pobre. Así, se prueban diferentes métodos para establecer ese umbral óptimo de decisión. Los diferentes umbrales utilizados son: Regla de Bayes p=0.5; p=0.35, como resultado de observar las predicciones del modelo de forma gráfica, y p=0,28 como resultado de ejecutar un algoritmo donde se prueban 100 umbrales aleatorios entre 0 y 1. El umbral óptimo se establece en p=0.28, como resultado de observar las predicciones del modelo con este umbral. La tabla 2 muestra las medidas Accuracy, Sensitivity y Stability para los 3 modelos en cuestión, y se observa que el modelo Logit - umbral óptimo presenta un alto Sensitivity, lo que nos garantiza que hay una alta probabilidad de predecir bien pobreza. Además, el modelo también presenta un alto Accuracy, el cual garantiza que la probabilidad de predecir mal pobreza sea menor. Al comparar estas dos medidas con los otros modelos, se observa que este presenta la mejor relación predicción bien vs predicción mal. Por consiguiente, este modelo se elige para realizar la predicción. Resultados de la predicción en el archivo .csv adjunto.

Referencias Bibliográficas

UN. (2015). Objetivo 1: Poner fin a la pobreza en todas sus formas en todo el mundo. https://www.un.org/sustainabledevelopment/es/poverty/

DANE.(2022). Pobreza monetaria y grupos de ingreso en Colombia. https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2021/Presentacion-pobreza-monetaria_2021.pdf

DANE. (2022). Empleo informal y seguridad social. https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-informal-y-seguridad-social#:~:text=Para%20el%20total%20nacional%2C%20en,proporci%C3%B3n%20fue%2044%2C3%25.