Profesor: Ignacio Sarmiento-Barbieri

Problem Set 2: Predicting Poverty

Big Data y Machine Learning para Economía Aplicada, 2023-1

Integrantes: Jorge Leonardo Rodríguez Arenas[1], Iván Camilo Velásquez Abril[2], Santiago González Cepeda[3] , María José Colmenares Wolff[4]

1. Introducción

El panorama social de América Latina y el Caribe no es para nada alentador cuando de pobreza se habla, a finales de noviembre del 2022 se proyectó que 201 millones de personas (32,1 % de la población total de la región) viven en situación de pobreza, de los cuales 82 millones (13,1 %) se encuentran en pobreza extrema, CEPAL (2022). La pobreza es un problema complejo y transversal en la sociedad y también es estudiada en diferentes disciplinas académicas, puede verse causada por muchos factores como la desigualdad económica, la discriminación y exclusión social, la falta de oportunidades y de acceso a la educación y a la salud, la reducida oferta de empleo y la corrupción, etc. Por consiguiente, el objetivo que tiene el Banco Mundial de acabar con la pobreza extrema en el 2030 parece un objetivo completamente irrealizable ya que, el ejercicio para determinar las estrategias que funcionan para reducir la pobreza suelen ser difíciles, requieren mucho tiempo y son muy costosos, WorldBank (2019). Es por esto que, el Banco Mundial ha organizado recientemente la competencia Pover-T Test, con el fin de incentivar la colaboración por medio de bonificaciones económicas de distintos investigadores y así poder construir modelos más precisos y específicos que ayuden en la medición de la pobreza.

Para el caso específico de Colombia, hacer predicciones de pobreza es relevante por diferentes razones. En primer lugar, el Gobierno tiene un plan de acción en donde se quieren cumplir algunos de los Objetivos de Desarrollo Sostenible (ODS) promovidos por la ONU, entre esos está el ODS 1, correspondiente al fin de la pobreza. Las metas trazadas para este objetivo es eliminar la pobreza extrema, reducir la pobreza en al menos 50 %, implementar sistemas de protección social y garantizar igualdad de acceso a servicios básicos a toda la población. DNP (2017)

Al cierre del 2021, la pobreza monetaria disminuyó puntos porcentuales, llegando a 39.3 % de la población y la pobreza extrema llegó a 15.1 % después de la disminución de 2.9 puntos porcentuales DANE (2022). Como estrategia disminuir la pobreza monetaria, se aumentó el número de beneficiarios a programas sociales como Familias en Acción Jóvenes en Acción e Ingreso Solidario.

Este tipo de programas dan incentivos monetarios después que se cumplan las condiciones de entrega de estos, por ejemplo, para Familias en Acción se debe verificar que los hijos del hogar hayan asistido al colegio, esto para incentivar la asistencia y aumentar los años de escolaridad, ya que esto puede traer beneficios de largo plazo para las familias. Además de que le permite a las familias tener cierto ingreso

[1]Cod. Uniandes: 201715669
[2]Cod. Uniandes: 201114762
[3]Cod. Uniandes: 201719971
[4]Cod. Uniandes: 201811929

extra, que puede ser un impulso para sacarlos de la pobreza o pobreza extrema.

Una de las limitaciones de este tipo de programas es que no hay información completa de todos los hogares del país, lo cual permitiría que existan hogares que necesiten de estos programas y no estén beneficiados o, por el contrario, que existan hogares que tienen ingresos suficientes para no estar clasificados como en pobreza, que de todas formas estén recibiendo el beneficio. Es por eso que trabajos como el presentado a continuación pueden ser un gran motor de desarrollo social, al poder determinar cuáles son los hogares pobres que existen censados en el país de una manera relativamente rápida y de bajo costo, para ayudar en una mejor focalización de los recursos del estado.

De esta manera, la inforamción recolectada en la Misión de Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP), recolectada por el DANE es una herramienta útil debido a que es una nueva metodología para la medición de la pobreza monetaria en Colombia, donde se adoptan cambios en la linea de pobreza y en la construcción del agregado del ingreso de hogar. Esta metodología ofrece una medición actualizada, i) construye una linea de pobreza utilizando una base estadística mas reciente de los hábitos de consumo de los colombianos, ii) incorpora adelantos metodológicos recientes y aceptados por expertos internaciones y iii) utiliza la medición mas precisa del agregado de ingreso DANE (2019). Por lo cual, permite la comparabilidad en el contexto regional debido a que las metodologías vigentes se acercan a los demás países de la región.

Al final de este estudio, se habrán mostrado cuales son las dos mejores formas que se encontraron para predecir la pobreza, a través de estrategias diferentes. Por un lado, veremos un modelo sencillo en el que se estimará directamente la pobreza, a través de un modelo de predicción de probabilidades Logit. Además, encontramos un modelo en el que, en caso de no tener los suficientes datos para hacerlo de forma directa, hace las predicciones a través de predicciones en los Ingresos por hogar, para después, utilizar otro tipo de modelo para hacer la clasificación del hogar como pobre o no.

2. Datos

Los datos de la encuesta “Medición de Pobreza Monetaria y Desigualdad 2018” creados a partir de la GEIH de 2018 que es usada principalmente por el DANE para realizar las estimaciones del Índice de Pobreza Multidimensional (IPM) y clasificar los hogares en pobres y no pobres con base en la línea de pobreza establecida para el país presentada en el Boletín Técnico Pobreza Monetaria en Colombia Año 2018, que es también un insumo para la Misión de Empalme de las Series de Empleo, Pobreza y Desigualdad (MESEP), y diversas investigaciones que tienen lugar en las ramas de desarrollo económico y mercado laboral. La información es presentada en dos secciones, una corresponde a información del hogar y otra a información de las personas, para crear una base unificada es necesario realizar un emparejamiento a nivel hogar, es decir a cada individuo se le asigna la información correspondiente al hogar. Para fines de este estudio, la unidad de observación será el hogar, lo que significa que se tomará una serie de variables individuales y se harán cálculos a nivel de hogar para saber características dentro del hogar, una utilidad de esto es obtener características de la composición del hogar, ya sea por género, nivel educativo, número de hijos, entre otros.

El universo de estudio cuenta con 231.128 hogares de los cuales 164.960 serán la muestra objeto de entrenamiento de los modelos y los restantes 66.168 hogares corresponden a la muestra de testeo del modelo que se presentara en la siguiente sección. Para consolidar las bases de datos fue necesario realizar imputación de datos para ciertas variables que contenían valores perdidos, esto con el fin de no perder observaciones en ninguna de las muestras del universo de estudio. Para el caso de la variables experiencia y horas trabajadas, a los menores de edad se les imputaba el valor de 0 y para aquellas

observaciones faltantes que reportaban ser mayor de edad se les imputaba el promedio de la experiencia o de las horas trabajadas de los otros individuos del hogar. Para las variables categóricas se les imputaba el valor de cero a los valores faltantes, dado que la comparación tomará el valor de uno si cumple la característica de la categoría y cero si toma otro valor sin importar cual sea la categoría que represente. Dado que la variable ingreso total contiene muchos valores perdidos se hace uso de la variable ingreso total de la unidad del gasto que representa la suma de los ingresos dentro del hogar (unidad de gasto) y esta no presenta valores faltantes.

El cuadro 1, resume los estadísticos descriptivos del universo de estudio discriminando en la muestra de entrenamiento y testeo, dentro de la muestra de entrenamiento se realza la separación entre pobres y no pobres. En la muestra de entrenamiento se cuentan con 164.960 hogares de los cuales 33.024 son pobres y 131.936 no obtienen esta clasificación. Al representar solo el 20 % los hogares pobres del total de hogares, ya se puede tener un indicio del problema del desbalance de clases que se tendrá en cuenta en la siguiente sección. Por otro lado, la muestra de testeo cuenta con 66.168 hogares.

Cuadro 1: Estadísticas Descriptivas de Hogares - Colombia 2018 por clasificación de pobre-no pobre

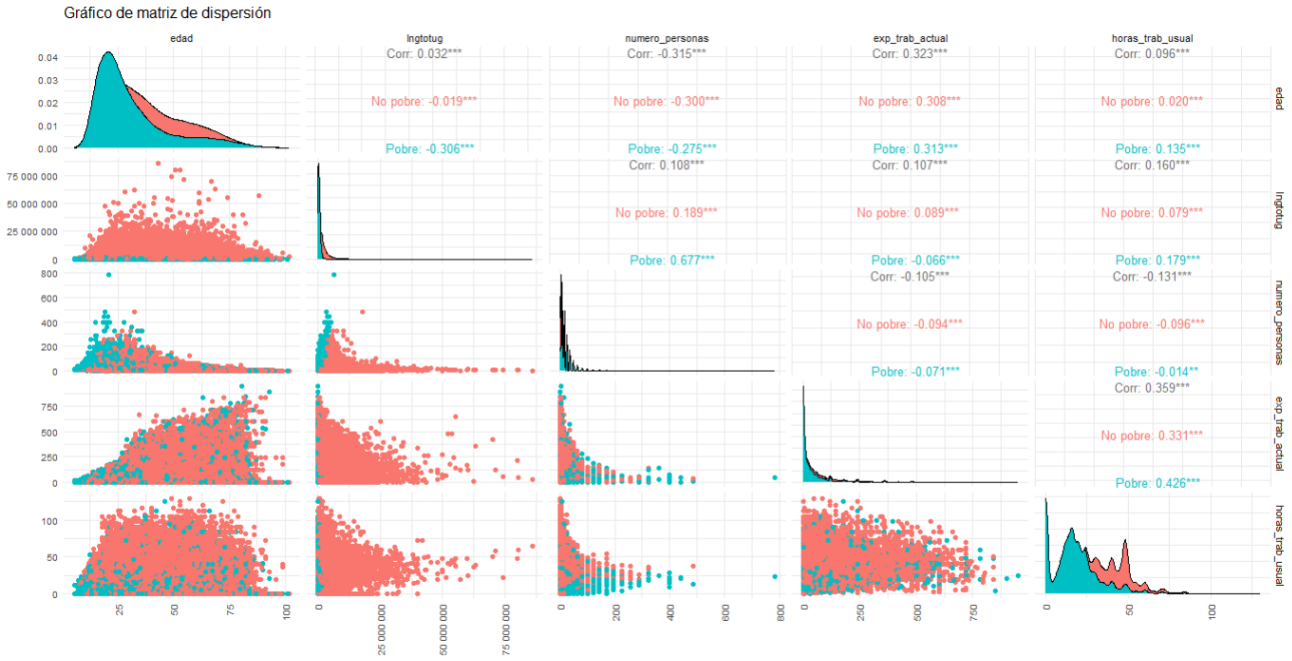
	Muestra de entrenamiento							**Muestra de testeo**	
	Todos		**Pobres**		**No probres**			**Todos**	
	Media	Desv. Estándar	Media	Desv. Estándar	Media	Desv. Estándar	**Diff**[1]	Media	Desv. Estándar
A. Composición									
Prop. Mujeres	0,526	0,277	0,546	0,241	0,526	0,277	***	0,523	0,276
Edad	37,44	16,88	30,63	16,23	37,44	16,88	***	37,47	16,88
Prop. Amo(a) de casa	0,200	0,243	0,244	0,219	0,200	0,243	***	0,204	0,244
Prop. hijos en el hogar	0,200	0,243	0,244	0,219	0,200	0,243	***	0,204	0,244
Número de menores de edad	0,919	1,121	1,733	1,415	0,919	1,121	***	0,938	1,133
Número de personas	13,99	16,91	21,20	23,86	13,99	16,91	***	14,24	17,08
B. Educación[2]									
Estudiantes	0,116	0,193	0,152	0,203	0,116	0,193	***	0,116	0,193
Primaria	0,052	0,159	0,096	0,199	0,052	0,159	***	0,056	0,164
Secundaria	0,159	0,236	0,202	0,229	0,159	0,236	***	0,160	0,236
Media	0,221	0,278	0,187	0,223	0,221	0,278	***	0,219	0,278
Terciaria	0,265	0,335	0,096	0,199	0,265	0,335	***	0,259	0,332
C. Mercado laboral									
Ingreso Total UG	2.090.895	2.512.488	633.200	462.872	2.090.895	2.512.488	***	-	-
Experiencia trab. Actual[3]	59,4	89,0	39,8	76,3	59,4	89,0	***	60,4	90,2
Horas trabajadas en la semana	28,4	18,3	17,4	14,2	28,4	18,3	***	28,3	18,2
Observaciones	164.960		33.024		131.936			66.168	

Notas: Cálculos propios usando datos DANE Medición de Pobreza Monetaria y Desigualdad 2018. [1] Indica el nivel de significancia estadística de la diferencia de medias entre hogares pobres y hogares no pobres bajo la $Ho : \mu(Pobre) - \mu(No\ pobre) = 0$. [2] Los valores representados en esta sección corresponden a la proporción de personas que esta dentro de cada categoría.[3] La experiencia en el trabajo actual esta medida en meses. La unidad monetaria de las variables referentes al ingreso de la unidad de gasto son pesos colombianos nominales. *** p<0.01, ** p<0.05, * p<0.1

Entre la muestra de testeo y entrenamiento no hay mayor diferencia entre las diversas características del hogar, las mujeres son más del 50 % de las personas dentro del hogar, la edad promedio ronda los 37 años, y menos del 10 % de las personas del hogar son menores de edad. En términos de niveles educativos la educación terciaria es la categoría que prima con respecto a las otras y tan solo el 5 % corresponde a educación primaria. En cuanto a mercado laboral la experiencia promedio de ambas muestras es de aproximadamente 60 meses y las horas trabajadas en promedio de 28 horas.

Al centrarnos en la muestra de entrenamiento y al discriminar los hogares en pobres y no pobres, se destaca que existe una diferencia estadísticamente significativa del 1 % entre ambas categorías en todas las características presentadas en el cuadro 1, la edad promedio de los hogares no pobres es mayor (7 años más en promedio que los pobres), es decir que los hogares pobres son mayoritariamente compuestos de personas jóvenes, y adicional son estos hogares los que presentan una mayor proporción de hijos dentro del hogar. En términos educativos, los hogares pobres tienen una menor cantidad de personas con educación terciaria frente a los no pobres y el nivel educativo predominante es la educación media. La diferencia en salarios es grande ya que el ingreso promedio de la unidad de gasto de los hogares no pobres es de alrededor de 3 veces más que el ingreso obtenido por los hogares pobres; estas

diferencias pueden estar siendo impulsadas por la experiencia y las horas trabajadas, ya que los pobres tienen una menor experiencia promedio en el mercado laboral y trabajan alrededor de 11 horas menos en promedio a la semana que los no pobres.



(a) Muestra de entrenamiento

Figura 1: Matriz de dispersión

En la figura 1, se muestra una matriz de dispersión para algunas variables presentadas en el cuadro 1, la distribución de la edad tiene mayor concentración en edades superiores para los no pobres, esto sucede también en las horas trabajadas para los no pobres, las correlaciones entre variables son bajas, sin embargo existe una correlación positiva y grande entre el ingreso total de la unidad del gasto y el número de personas en los hogares pobres.

3. Modelos y Resultados

En este ejercicio se tenía 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 70 % como una nueva base train la cual es submuestra del de la base train original y una sub base test del 0.30 que de igual forma es submuestra de la train original. Los mejores modelos se cargan a Kaggle y son pobrados con base de test original.

3.1. Ingresos

En esta sección, nos vamos a dedicar a realizar un paso intermedio para la predicción de la Pobreza, y esto es a través de los Ingresos. Esta es una manera de realizar estimaciones sin tener los datos directos para estimar la pobreza, sino que, a través de diferentes modelos y sus respectivas reglas de decisión, aporta información fundamental para realizar estimaciones de pobreza.

Para esto, tenemos diferentes modelos, los cuales se deben probar su capacidad predictiva de los ingresos totales por hogar. A continuación, se describirán los modelos que fueron propuestos en esta sección.

- **Modelo 1:** Esta especificación es una regresión lineal:

$$Ingtotug_i = \beta(X) + u_i$$

  La cual **X** incluye las siguientes variables explicativas: la edad, la edad al cuadrado, la proporción de mujeres en el hogar, de estudiantes de primaria, secundaria, media y superior y finalmente la experiencia de trabajo al momento de la encuesta, para finalmente predecir el Ingreso total por unidad de gasto.

- **Modelo 2:**Esta especificación es una regresión lineal:

$$Ingtotug_i = \beta(X) + u_i$$

  La cuál contiene algunos controles diferentes que el primer modelo. En este se incluyen dos variables adicionales, las cuales son horas de trabajo promedio a la semana y la proporción de las personas en búsqueda de trabajo.

- **Modelo 3:** Gradient Boosting Machine, esta es una herramienta que se utiliza enfocada a los Random Forest, en la cual se entrenan a diferentes tipo de modelos de forma gradual y consecutiva, optimizando así la función de pérdida.

- **Modelo 4:** Árbol de Decisión, este es un modelo de predicción basado en diferentes reglas de elección que representan los diferentes pasos jerárquicos, con los que se realiza la estimación.

- **Modelo 5:** Random Forest, en este modelo se crearon mil arboles de manera aleatoria, el resultado de este método fue un modelo que donde cada árbol contenía tres ramas y la cantidad de nodos era de cien, lo anterior fue tomado porque era la opción que minimizaba el Error Cuadrático Medio (RMSE) mediante validación cruzada.

Al evaluar estos modelos en los diferentes sets de datos disponibles, encontramos las siguientes métricas, medidas en el subset de validación, sacado del set de entrenamiento. En la Tabla siguiente, se tiene una recopilación de las 4 principales métricas, que serán decisivos a la hora de seleccionar el mejor modelo para realizar la predicción.

Cuadro 2: Criterios

Modelo	RMSE	R^2	MAE	MedianAPE
Regresión Lineal 1	2275433	0.18	1263290	802783.7
Regresión Lineal 2	2300923	0.18	1250251	794982.6
Random Forest GBM	2118218	0.29	1107930	661483.2
Arbol de Decisión	2270356	0.18	1300000	869671.9
Random Forest	2074955	0.32	1071345	618345.2

Las métricas que se escogieron para medir la calidad del modelo en términos de su capacidad predictiva outsample son las siguientes: i) MAE, esta indica en promedio, cuánto fue la diferencia entre los valores reales y los valores predichos en cada observación del subset de validación. ii) MEdianAPE, indica cual es el error promedio en el percentil cincuenta. iii) RMSE, evalúa la precisión del modelo de predicción ya que, representa la desviación típica de las predicciones del modelo con respecto a los valores reales de la variable de interés y iv) el R^2 que muestra la proporción de la variabilidad en el ingreso que es explicada por el modelo de regresión.

Como se puede observar, el Modelo con las mejores métricas es el modelo de Random Forest, ya que incluye el menor valor de MAE, MedianAPE, RMSE y el mayor R^2 en comparación con los otros modelos que fueron diseñados en este estudio para la predicción de los ingresos.

3.2. Pobreza

Para poder realizar la predicción de si un hogar es pobre o no, se debe de recurrir a la estimación de modelos de clasificación binaria, en este caso si la variable pobre toma el valor de uno indica que el hogar es pobre 0 de lo contrario. Para esta estimación de modelos se tendrá un vector de predictores $\mathbf{X}$ que serán usados en la estimación, este vector incluye variables acorde con la teoría económica que dan cuenta de la composición del hogar en distintas características.

Se realizan dos ejercicios en paralelo, 1) estimar pobreza directamente; 2) estimar primero ingreso y posteriormente pobreza; por lo cual se está prediciendo pobreza en dos pasos. Los modelos 1 y 2, corresponden al ejercicio 1. Los modelos 3 predicen pobreza secuencialmente.

- **Modelo 1:** La primera especificación será un modelo Logit tradicional, cuya ecuación a estimar es:

  $$Pr(Pobre_i = 1|\mathbf{X}) = \Lambda(\mathbf{X})$$

  Donde el vector $\mathbf{X}$ contiene la edad, la edad al cuadrado, la proporción de mujeres, estudiantes, buscadores de empleo, amos(as) de casa, hijos en el hogar, adicional incluye información sobre la composición de niveles educativos dentro del hogar (primaria, secundaria, media, terciaria) y finalmente incluye variables de mercado laboral como experiencia, y horas trabajadas. Para esta especificación se probo distintos modelos que consistió en adicionar controles hasta consolidar el modelo estimado. A este modelo se le hacen las variaciones de CARET y ridge.

- **Modelo 2 - modelos no lineales:** explorando relaciones no lineales de los datos utilizamos modelos CART. Por tal razón, se estiman árboles de decisión, árboles con baggin y random forest. De esta estos modelos encontramos que los mejores modelos eran bagging y random forest, ya que corregían el problema de varianza del los árboles normales. Sin embargo, esta diferencia no es significativa ya que pasamos de un accuracy de 0.82 a 0.83 en los últimos dos modelos mencionados.

  Se subió a Kaggle el modelo bagging con las variables edad, edad al cuadrado, mujer, estudiante, busca, trabajo, amo de casa, hijos hogar, primaria, secundaria, media, superior, experiencia en el trabajo actual y horas de trabajo semanal , considerando que era uno de los mejores modelos. Sin embargo, ninguno de los modelos CARTS supero en Kaggle al modelo 1 (logit). Esto nos dio un primer indicio de que era mas importante explorar las relaciones lineales de la base.

- **Modelo 3:** La segunda especificación será un modelo Logit con el método de regularización de LASSO con CV, cuya ecuación a optimizar es:

  $$\min_\beta\{-\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\Lambda(\mathbf{X}_i)) + (1-y_i)\log(1-\Lambda(\mathbf{X}_i))] + \alpha\sum_{j=1}^{p}|\beta_j|\}$$

  donde y_i es la variable respuesta binaria para el individuo i y $\mathbf{X}$ es el mismo vector de covariables del modelo 1.

  Dado que la base no esta balanceada en la variable pobreza, utilizamos las técnicas de balanceo de clase SMOTE y upsampling. Al comparar el accuracy de los tres modelos nos da que el modelo con mayor accuracy es el lasso simple, con un acurracy de 0.823 en comparación con SMOTE – 0.778 , upsampling 0.697. Por esta razón, se deicidio correr los modelos sin ninguna técnica de balanceo.

Al evaluar estos modelos se obtienen las siguientes métricas:

Cuadro 3: Criterios

Modelo	Recall	Accuracy	F1_Score	Precision
Logit	0.957	0.824	0.897	0.843
Logit Lasso Upsampling	0.679	0.695	0.781	0.918
Logit Caret	0.958	0.825	0.897	0.844
Logit Ridge	0.972	0.819	0.896	0.831
SMOTE	0.789	0.778	0.850	0.921
LDA	0.968	0.821	0.896	0.834

Conforme a los modelos y resultados obtenidos previamente, se procede a elegir el modelo ganador, es decir el que presentó un mejor accuracy:

Cuando se quiere estimar directamente la pobreza el modelo que obtiene un mayor Accuracy es el modelo 1, este es un Logit estimado únicamente con máxima verosimilitud. Esto implica que para predecir la pobreza directamente la mejor opción es un modelo lineal sencillo, que tiene en cuenta un vector de predictores acorde con la teoría económica. Cabe resaltar que la diferencia en Accuracy no es muy grande con los otros métodos de estimación, con un árbol de decisión simple se obtiene un valor de 0.831. Sin embargo, para predecir fuera de muestra el modelo Logit sencillo es quien mejor desempeño tiene.

Por otra parte, cuando se quiere estimar la pobreza de forma indirecta a través de los ingresos utilizando el modelo Random Forest con cien nodos y tres ramas, métricas que reducen el Error Cuadrático Medio comparadas con otros escenarios, . Posteriormente se utilizó una regresión Logit con el paquete CARET para hacer la clasificación de los hogares entre Pobre y No pobre. De este modo, el Accuraccy obtenido es de 0.825 y la calificación que arrojó la competencia de Kaggle fue de 0.813.

4. Conclusiones

Al final de este estudio, vemos que, a pesar de tener diferentes especificaciones de la estimación de la pobreza de manera directa e indirecta, el mejor modelo de predicción de la misma es el modelo mas sencillo sin métodos de regularización en donde solo se tienen en cuenta un vector de predictores que va acorde la teoría económica. Esto a su vez nos indica que para el año 2018 la pobreza en Colombia presenta una relación linealizable con los diferentes predictores tenidos en cuenta en el modelo uno.

5. GitHub

Mediante este link :
namedDarkOrchid=1 GitHub grupo 9

Referencias

CEPAL (2022): «Tasas de pobreza en América Latina se mantienen en 2022 por encima de los niveles previos a la pandemia, alerta la CEPAL,» .

DANE (2019): «Pobreza Monetaria en Colombia: Nueva metodología y nuevos resultados,» .

——— (2022): «Comunicado de Prensa: Pobreza en 2021,» .

DNP (2017): «La Agenda ODS 2030,» .

WORLDBANK (2019): «DrivenData - World Bank Poverty Prediction,» https://www.drivendata.org/competitions/50/worldbank-poverty-prediction/page/97/, accedido el 23 de febrero de 2023.