Predicting poverty at household level data in Colombia

Páez Iván[1], Niño Melissa[2], Bonilla Jessica[3], Manco Luis[4]

Facultad de Economía de la Universidad de los Andes

Resumen:
El uso de modelos predictivos que permitan el análisis rápido y acertado de la pobreza a un nivel cada vez más detallado surge de la necesidad de contar con herramientas menos costosas y más rápidas, para adelantar política pública que permita la asignación adecuada de recursos permitiendo un mayor beneficio en términos de costo efectividad. En este documento se exploran modelos de clasificación y regresión para la predicción de hogares en condición de pobreza en Colombia, empleando datos oficiales del DANE y estrategias de agregación de datos. Para obtener más detalles y replicar el estudio, se proporciona acceso al repositorio de GitHub *https://github.com/iapaezg/BD_LM_03*

Palabras clave:
Pobreza
Predicción
Precisión

I. Introducción

Durante casi un cuarto de siglo, se ha observado una disminución constante en el número de personas viviendo en pobreza extrema, es decir, con menos de USD 2,15 al día. Sin embargo, esta tendencia se vio afectada en 2020 debido a las alteraciones ocasionadas por la crisis de la COVID-19, así como los impactos de los conflictos y el cambio climático, factores que ya venían ralentizando la reducción de la pobreza. Se estima que entre 75 y 95 millones de personas adicionales podrían encontrarse en situación de pobreza extrema en 2022, en comparación con las proyecciones previas a la COVID-19, debido a los efectos duraderos de la pandemia, la guerra en Ucrania y el aumento de la inflación (Banco Mundial, 2022).

En América Latina y el Caribe, a pesar de haber experimentado una reducción de la pobreza durante la primera década del siglo, ha enfrentado un estancamiento en sus índices debido a estructuras económicas poco productivas y alta informalidad. La pandemia de COVID-19 agravó la situación, dejando a más de un tercio de la población en la pobreza y afectando desproporcionadamente a mujeres, jóvenes y trabajadores informales. La pobreza extrema también aumentó, pasó del 13,1 % en 2020 al 13,8 % en 2021, y el grupo de vulnerables creció, poniendo en riesgo a quienes están al borde de la clase media. Además, bajos niveles de ahorro y la falta de protección social exponen a millones de personas a un riesgo de movilidad social descendente en momentos de crisis (CAF, 2022).

En cuanto a Colombia, entre 2018 y 2022 ha habido un fuerte incremento de la pobreza monetaria. A diciembre de 2021 se registró que 19.621.000 personas (39,3 % de la población) vivían con menos de 11.801 pesos al día, y 6.111.000 personas (12,2 %) con menos de 5.730 pesos (Espitia, 2022). Estas mediciones considerando datos de la Gran Encuesta Integrada de Hogares (GEIH) la cuál mediante un ejercicio de integración de registros administrativos en el 2019, mejoró la precisión en la medición de las ayudas institucionales otorgadas por parte del gobierno nacional y por algunas administraciones

1 Biólogo. Universidad Nacional de Colombia. ia.paez10@uniandes.edu.co
2 Abogada. Universidad del Rosario. i.nino@uniandes.edu.co
3 Administradora Ambiental. Fundación Universitaria Empresarial de la Cámara de Comercio de Bogotá jl.bonillac1@uniandes.edu.co
4 Ingeniero Ambiental. Universidad Distrital Francisco José de Caldas. l.manco@uniandes.edu.co

locales, permitiendo estimar el impacto que estas ayudas han tenido en la reducción de la pobreza (DANE, 2022).

Este tipo de estrategias de integración de registros se ha implementado recientemente para el análisis de pobreza, llegando a estudios como el de Hu et al (2022) en el cual emplearon datos de satélites para reforzar la información de áreas rurales de china, empleando como variable proxy a la pobreza la iluminación artificial nocturna de poblados, permitiendo robustecer y agilizar la toma de decisiones sobre el territorio.

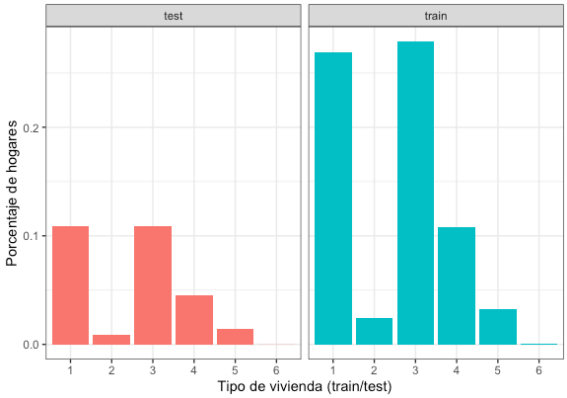
II. Datos

Los datos provienen de la Gran Encuesta Integrada de Hogares - GEIH 2018 realizada por el Departamento Administrativo Nacional de Estadística - DANE para 32.177 individuos pertenecientes a 10.567 hogares de 10.403 viviendas localizadas en la ciudad de la ciudad de Bogotá. Adicionalmente, el DANE construye la línea de pobreza y de indigencia a partir de la aplicación de la metodología de la CEPAL que se basa en la estructuración de una canasta básicos de alimentos siguiendo los requerimientos calóricos, que es posteriormente valorada en términos monetarios.

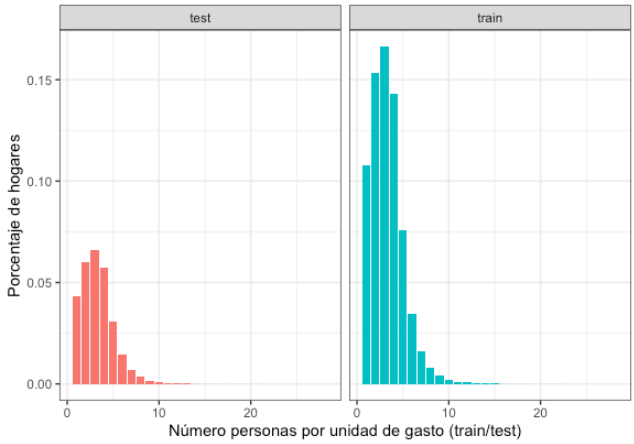
La primera actividad realizada fue la identificación de las muestras de entrenamiento y test mediante la creación de una variable adicional en la que se pudiera determinar en cuál de dichas categorías se encontraba la observación. Posteriormente para la labor de limpieza de los datos, se siguieron los criterios que se describen a continuación. Para las variables del régimen de seguridad social, nivel educativo y las actividades realizadas en la semana anterior, los datos faltantes y la opción de respuesta que indicaba el desconocimiento del encuestado sobre lo preguntado, fueron reemplazados con el valor que presentaba mayor frecuencia. Por su parte, para la variable de los años de educación, se reemplazó el valor 99 que correspondía a los eventos en los que el individuo desconocía sus años de educación con la mediana de los datos. Por último, para la variable de horas trabajadas las respuestas 98 y 9 que corresponden a los eventos en los que la persona no sabe cuántas horas trabajó o desconoce el valor que recibió por dicho concepto, se reemplazaron los valores por ceros. Este mismo proceso se replicó para las variables de afiliación a salud, horas extras, diversos tipos de primas, bonos, subsidio de alimentos, transporte, familia y educación; así como, para variados tipos de salarios en especie como alimentos, vivienda, otros tipos de salario, ayudas de otros hogares nacionales e internacionales, ayudas de instituciones, pensión de invalidez, intereses de inversiones y de cesantías.

a. Análisis descriptivo de los datos

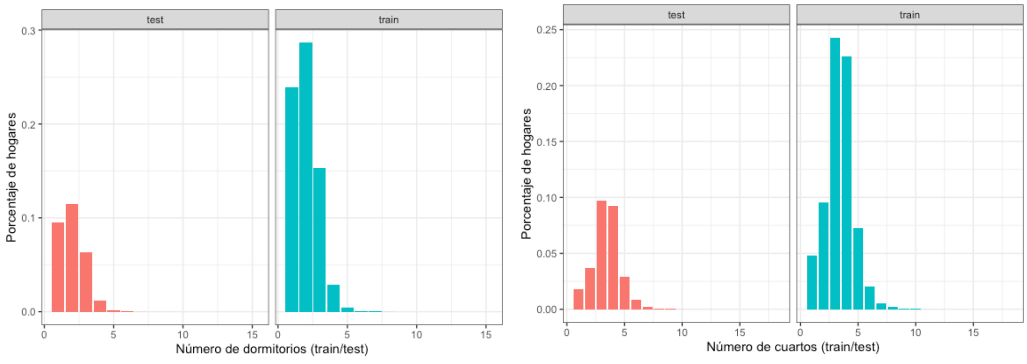
Los tipos de vivienda que fueron indicados por los encuestados corresponden a las siguientes categorías:1. propia, totalmente pagada, 2. propia, la están pagando, 3. en arriendo o subarriendo, 4. en usufructo, 5. en posesión sin título 6. otros. De acuerdo con la partición de la base de datos analizada, en la base de datos de entrenamiento, el 38.79% de los hogares encuestados indicaron que su vivienda es arrendada, seguido por el 37,80% hogares que indicaron que su vivienda es propia y ha sido totalmente pagada. Dicho comportamiento también se puede evidenciar en la base de datos de test.



El 95,27% de los hogares presentan entre 1 y 6 personas por unidad de gastos, con alta concentración de hogares para 2, 3 y 4 personas por unidad de gasto en los que se evidencian 53.715 (21,31%), 49.260 (23,24%) y 46.378 (20,06%) hogares con dicha cantidad de individuos. Este mismo comportamiento se pudo observar en la variable de personas por unidad de hogar, toda vez que los valores de personas por unidad de gasto equivalen a la cantidad de personas por hogar.

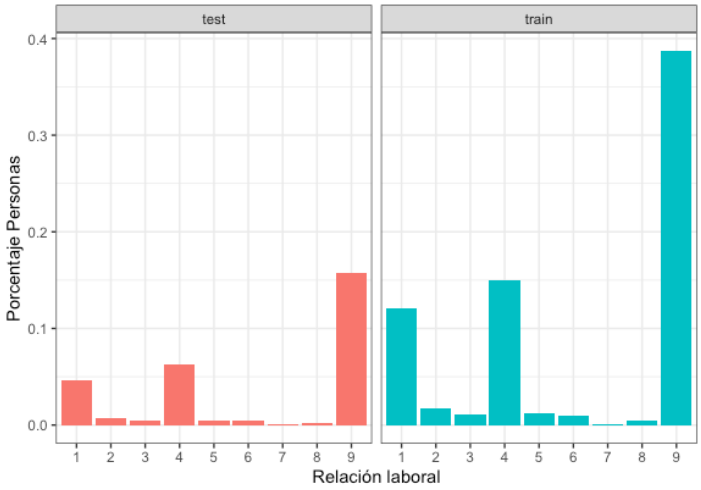


En la encuesta se discriminan los conceptos de dormitorio y de cuarto, siendo este primero el espacio del hogar en el que las personas duermen. Así las cosas, la mayor cantidad de hogares presentan entre 3 y 4 cuartos, ya que se evidencia que el 33,98% y 31,91% de los hogares presentan dicha cantidad de cuartos, respectivamente. Por su parte, la mayor cantidad de dormitorios por hogar se encuentra entre 1 y 2 dormitorios, por cuanto el 33,34% y 40,11% de los hogares presentan dicha cantidad de cuartos, respectivamente.



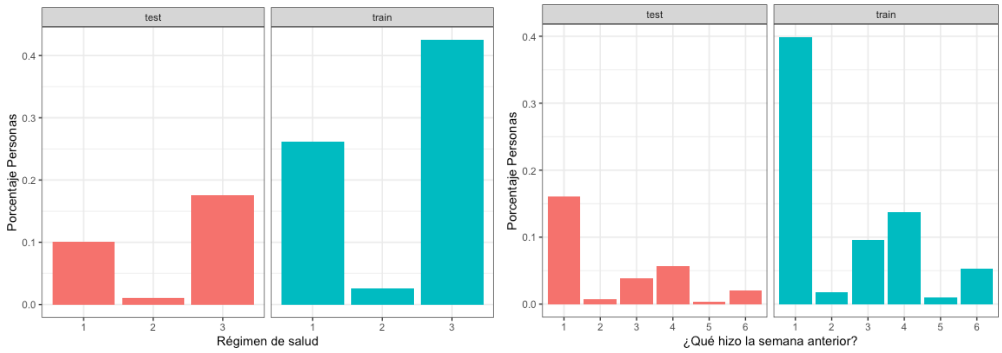
Para la relación laboral de los encuestados se presentan las siguientes categorías: 1. obrero o empleado de empresa particular, 2. obrero o empleado del gobierno, 3. empleado doméstico, 4. trabajador por cuenta propia, 5. patrón o empleador, 6. trabajador familiar sin remuneración, 7. trabajador sin

remuneración en empresas o negocios de otros hogares, 8. Jornalero o peón y 9. otros. El 54,93% de las personas encuestadas reportan otros tipos de relaciones laborales, mientras que el 21,21% son trabajadores por cuenta propia y el 16,71% son obreros o empleados de empresa particular, en este mismo orden.

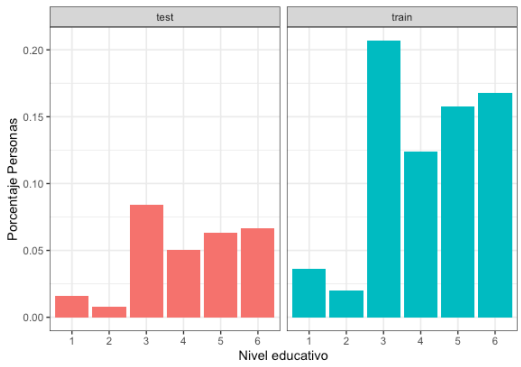


Los regímenes de salud en el que pueden estar afiliados los encuestados son los siguientes: 1. contributivo (eps), 2. especial (fuerzas armadas, Ecopetrol, universidades públicas), y 3. subsidiado. El 60% de los encuestados pertenecen al régimen subsidiado, seguidos, aunque en cantidad ampliamente inferior, por el 36,25% de los encuestados quienes están afiliados al régimen de contributivo.

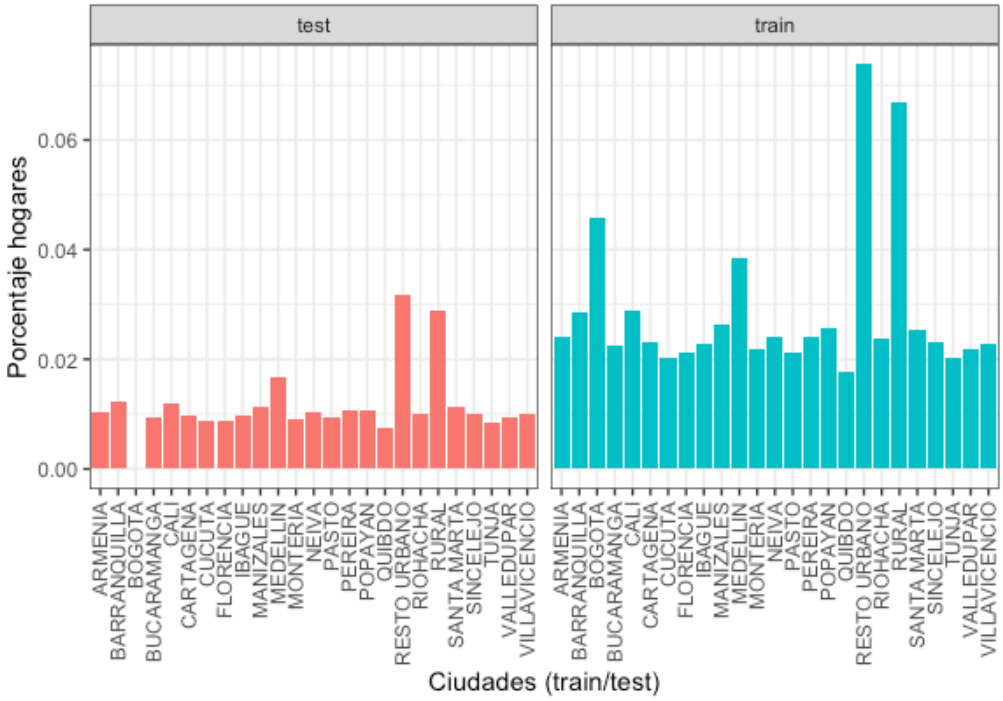
Respecto a la variable identificada como "¿qué hizo la semana pasada?, los encuestados respondieron lo siguiente: 1. trabajando, 2. buscando trabajo, 3. estudiando, 4. oficios del hogar, 5. incapacitado permanente para trabajar, y f. otra actividad. Como se puede evidenciar en la siguiente gráfica, el 55,91% de los individuos encuestados respondieron que la mayor parte de su tiempo la semana anterior fue ocupado en la actividad de trabajar.



La variable nivel educativo corresponde a las siguientes categorías: 1. Ninguno, 2. Preescolar, 3. Básica primaria (1o - 5o), 4. Básica secundaria (6o - 9o), 5. Media (10o -13o), 6. Superior o universitaria, y 7. No sabe, no informa. El 29,11% de las personas encuestadas alcanzaron como grado de escolaridad más alto el de educación básica que equivale a los grados de 1 a 5.



Por último, en cuanto a las ciudades a la que pertenecen los hogares encuestados, en la muestra de entrenamiento se encuentran los datos de la ciudad de Bogotá, pero en la muestra de test no se encuentran datos de hogares localizados en esta ciudad. Igualmente, las ubicaciones de "resto urbano", "rural" y Medellín corresponden a los lugares en los que se encontraban la mayor cantidad de hogares encuestados.



El 80% de la totalidad de los hogares son pobres, a la vez que el restante de los hogares encuestados no está en condiciones de pobreza.

A continuación, se muestran las estadísticas descriptivas de las 45 variables de interés que serán usadas para los análisis propuestos para la predicción de condiciones de pobreza que se formula en los modelos que más adelante se estudian. En esta tabla, se podrá identificar por tipo de tabla el promedio y desviación estándar para cada una de las variables.

Variable	Mean			sd		
	test	train	Variation (%)	test	train	Variation (%)
a_pension	1,18	1,19	0,84	0,41	0,41	0,00
a_salud	0,77	0,77	0,00	0,42	0,42	0,00
anos_edu	5,92	5,96	0,67	3,34	3,33	-0,30
arriendo	0,07	0,08	12,50	0,26	0,26	0,00
ayuda_inst	0,06	0,05	-20,00	0,23	0,22	-4,55
bon_anual	0,00	0,00	0,00	0,02	0,02	0,00

Variable	Mean			sd		
	test	train	Variation (%)	test	train	Variation (%)
bonos	0,01	0,01	0,00	0,08	0,08	0,00
Clase	0,90	0,90	0,00	0,30	0,29	-3,45
Des	0,05	0,06	16,67	0,23	0,23	0,00
edad	33,50	33,55	0,15	21,69	21,65	-0,18
edad2	1.592,76	1.594,63	0,12	1.731,84	1.729,20	-0,15
exp	23,38	23,38	0,00	21,04	20,99	-0,24
exp2	989,26	987,52	-0,18	1.367,41	1.362,39	-0,37
h_extra	0,01	0,01	0,00	0,11	0,12	8,33
h_tra1	20,29	20,47	0,88	24,61	24,72	0,44
h_tra2	0,30	0,30	0,00	2,37	2,39	0,84
h_trat	20,59	20,77	0,87	24,96	25,07	0,44
hog_int	0,01	0,01	0,00	0,08	0,08	0,00
hog_na	0,08	0,08	0,00	0,27	0,27	0,00
Ina	0,31	0,31	0,00	0,46	0,46	0,00
ing_des	0,01	0,01	0,00	0,09	0,09	0,00
int_cesantias	0,02	0,03	33,33	0,14	0,17	17,65
int_inv	0,00	0,00	0,00	0,05	0,05	0,00
inv_pension	0,05	0,05	0,00	0,21	0,21	0,00
nivel_edu	4,17	4,19	0,48	1,42	1,41	-0,71
Oc	0,46	0,46	0,00	0,50	0,50	0,00
Orden	2,65	2,62	-1,15	1,71	1,69	-1,18
otras_fuentes	0,02	0,02	0,00	0,14	0,15	6,67
otros_esp	0,00	0,00	0,00	0,04	0,04	0,00
pat_pension	0,00	0,00	0,00	0,06	0,07	14,29
prima	0,00	0,00	0,00	0,05	0,05	0,00
prima_nav	0,05	0,05	0,00	0,21	0,21	0,00
prima_ss	0,12	0,13	7,69	0,33	0,33	0,00
prima_vac	0,04	0,04	0,00	0,19	0,19	0,00
q_hizo	2,30	2,30	0,00	1,63	1,64	0,61
reg_salud	2,26	2,23	-1,35	0,95	0,95	0,00
rel_lab	6,26	6,22	-0,64	3,21	3,24	0,93
s_edu	0,00	0,00	0,00	0,03	0,03	0,00
s_fam	0,04	0,05	20,00	0,21	0,21	0,00
s_trans	0,10	0,10	0,00	0,29	0,30	3,33
sal_alim	0,03	0,03	0,00	0,17	0,17	0,00
sal_viv	0,01	0,01	0,00	0,09	0,09	0,00
sexo	0,47	0,47	0,00	0,50	0,50	0,00
viaticos	0,01	0,01	0,00	0,07	0,08	12,50

Como se observa no hay mayor diferencia en la distribución de las variables entre la serie de datos de entrenamiento y de prueba, a excepción de dos variables (ayuda_inst y int_cesantias), esta condición de distribución en la construcción de un modelo de clasificación o predicción, como es el caso de este estudio, puede ser beneficioso para el rendimiento y generalización de modelos. Al tener una distribución similar en ambos conjuntos de datos, el modelo aprenderá patrones y características que son más generalizables y aplicables a datos desconocidos, por lo que se reduce el riesgo de sobreajuste y de sesgo, permitiendo obtener modelos que identifican patrones muy cercanos al mundo real.

A partir de los resultados de la anterior tabla se podría esperar una menor precisión de los modelos de incluir la variable de arriendo la cual presenta una variación del 12,5% entre los datos del entrenamiento y los de prueba. Considerando que, para uno de los análisis propuestos, se predice el ingreso de los individuos para posteriormente compararlo con el valor de la línea de pobreza que les es aplicable, se ha incorporado la variable de experiencia potencial que, de acuerdo con la literatura revisada, es una variable relevante para determinar el salario de los individuos.

III. Modelos

a. Modelos de clasificación

Considerando que en la base de datos se cuenta con la variable pobre o no para los hogares, se plantea el problema de clasificación a partir de dicha variable Pobre que está determinada de la siguiente manera: 0 = No pobre, 1 = Pobre.

Los dos modelos propuestos:

$$\text{Pobre} = \beta_0 + \beta_1 Dominio + \beta_2 \text{Cuartos} + \beta_3 \text{Nper} + \beta_3 \text{Tipo\_Vivienda}$$

$$\text{Pobre} = \beta_0 + \beta_1 Dominio + \beta_2 \text{Cuartos/Persona} + \beta_3 \text{Tipo\_Vivienda}$$

Los modelos antes indicados se corrieron usando Logit, Lasso, Lasso (con ROC) y Elastic Net.

Tabla 2. Datos sobre métricas

Métodos	α	λ	ROC	Sensibilidad	Especificidad	Precisión	ϰ
Logit1	NA	NA	0.775095	0.968967	0.204740	0.815974	0.232099
Logit2	NA	NA	0.773779	0.971371	0.189081	0.814761	0.217143
Lasso1	0	1.02329	0.764381	1	0	0.799806	0
Lasso2	0	1.02329	0.769290	1	0.001081	0.800022	0.001728
LassoRoc1	0	0.00943	0.775220	0.974056	0.179997	0.815091	0.210599
LassoRoc2	0	0.01892	0.774014	0.978171	0.154432	0.813263	0.185265
ElasticNet1	0.1	0.01898	0.775359	0.982123	0.140329	0.813601	0.173796
ElascticNet2	1	0.02722	0.742686	0.986487	0.095859	0.808188	0.120861

De acuerdo con lo referenciado en la tabla 2, se obtuvo un buen desempeño en los casos de aplicación Logit1 seguido de LassoRoc1, ambos resultan tener ROC similares, con buena sensibilidad. Se analizan los coeficientes para el modelo Logit1 y se puede verificar que para la predicción de la pobreza:

- La variable número de cuartos tiene un efecto negativo sobre la probabilidad de que el hogar sea pobre, es decir, a mayor número de cuartos la probabilidad de que el hogar sea pobre disminuye.
- Adicionalmente, el efecto de la localización geográfica del hogar es diferencial dependiendo de la ciudad, así, la probabilidad con respecto a la ciudad base (ARMENIA) puede aumentar o disminuir la propensión a que el hogar sea pobre.
- En cuanto al número de personas por hogar, se presenta que al verse aumentado la relación directa sobre la probabilidad de que le hogar sea pobre.

Lo anterior tiene relación con las condiciones de vivienda de los hogares en Colombia, toda vez que hogares de mayores ingresos pueden acceder a inmuebles con áreas más grandes (número de cuartos). Mientras que el número de personas en una vivienda refleja las condiciones socioeconómicas de los hogares, puesto que a menores ingresos existe una tendencia a que haya sobreocupación.

b. Modelos de regresión

Para el modelo de regresión aplicamos las siguientes metodologías regresión lineal con upsampling, regresión lineal con downsampling y árbol de decisión.

El modelo propuesto es:

$$Ingtot = \beta_0 + \beta_1 Oc + \beta_2 Dominio + \beta_3 Clase + \beta_3 sexo + \beta_4 edad2 + \beta_5 reg\_salud + \beta_6 anos\_edu + \beta_7 rel\_lab + \beta_8 h\_trat + \beta_9 exp + \beta_{10} exp2 + \beta_{11} q\_hizo + \beta_{12} a\_pension$$

La variable edad no se incluyó en el modelo por cuanto no resultó ser significativa. A efectos de abordar el problema de clasificación, se tomaron los ingresos totales por persona y se sumaron para cada uno de los hogares, después de lo cual se dividió el valor del ingreso por hogar en la cantidad de personas por hogar. Este ingreso por persona se comparó con la línea de pobreza aplicable a dicho hogar específico para determinar si es pobre o no. Debido a que la muestra presenta un desbalance entre los hogares pobres y no pobres (0.251 y 0.748, respectivamente), se optó por la estrategia de realizar regresiones lineales aumentando el número de observaciones para el factor en subrepresentado (UpSampling) y eliminando observaciones para el factor sobrerrepresentado (DownSampling). En la siguiente tabla se muestran los resultados:

Tabla 3. Datos sobre métricas

Método	Precisión	Sensibilidad	Especificidad
OLS-Up	0.652001	0.884186	0.559488
OLS-Down	0.651449	0.886680	0.557722
Árbol	0.748068	0	1

Para evaluar los modelos implementados, se tomó la precisión (True positive, True negative), la sensibilidad (True positive, predicción correcta) y especificidad (True negative, predicción correcta). De acuerdo con lo expuesto en la tabla, los modelos OLS presentan características similares, con una alta sensibilidad y precisión. Por otra parte, el modelo de árbol presenta la mayor precisión y especificidad. La baja o nula sensibilidad del árbol de decisión tiene sentido en el contexto de la limitación que presenta el usar un único árbol para determinar para la predicción.

Aunque los modelos OLS predijeron con menor precisión con respecto al Árbol, tomamos el OLS-Down dada su sensibilidad para participar en Kaggle, con lo que se generó un mayor score en la competencia.

c. Modelo final

El modelo final propuesto para la competencia en Kaggle fue el de OLS-Down, donde se eliminaron datos asociados a la muestra sobrerrepresentada con el objetivo de balancearla. Para el entrenamiento del modelo propuesto se utilizó el 70% de los datos y el restante para la muestra de testeo. Se utilizaron las variables por persona para predecir los ingresos de las personas por unidad de gasto, posteriormente se sumaron y se dividieron por el número de personas por unidad de gasto, posterior a lo cual se evaluó si el hogar era pobre o no de acuerdo con la línea de pobreza reportada. Se realizó validación cruzada con k=5, 13 predictores y de donde se puede obtener que RMSE=866353.

Aunque el modelo de regresión resulta generar valor de predicción buenos, cabe resaltar que el supuesto de distribución normal puede no cumplirse para las variables incluidas como predictores, por lo cual podría ser apropiado realizar un random forest. En general, con la predicción por OLS, se obtuvieron resultados sobre el ingreso que corresponden a la teoría económica y social:

Existe una brecha de género que se identifica con un mayor ingreso en hombres con respecto a las mujeres.

Para el caso de las ciudades o áreas metropolitanas, se encontró que todas estas presentan valores positivos frente a la ciudad de referencia (ARMENIA), sin embargo, pertenecer al sector rural tiene un efecto negativo sobre el ingreso de las personas. Lo anterior afecta directamente la probabilidad de los hogares a ser pobres o no.

Como es de esperarse, para las personas ocupadas presentaron un mayor nivel de ingresos que las desocupadas.

IV. Conclusiones

Las conclusiones de este análisis son de especial relevancia para la formulación de políticas públicas pues permite identificar aquellos individuos que, con base en las características analizadas, se encuentran o encontrarán en condiciones de pobreza, y con ello, se dirigirán las medidas de las políticas a la población de interés. Hemos evaluado diversos modelos teniendo en cuenta criterios propios de los hogares y de las personas que los conforman, tales como el tipo de vivienda, la cantidad de personas por hogar y unidad de gasto, así como, la cantidad de cuartos y habitaciones, y su proporcionalidad con los ocupantes. Para el enfoque basado en el ingreso de los individuos, se analizaron todas las variables que afectaban ingresos ordinarios o extraordinarios para las personas encuestadas, así como, las comúnmente estudiadas en la ecuación de Mincer, estas son, nivel educativo, experiencia y edad. El modelo con mejor desempeño fue el de la regresión lineal, el cual presenta un RMSE de 866353.

V. Bibliografía

Banco Mundial. (2022). Poverty and shared prosperity. In *Managing Automation* (Vol. 24, Issue 5). https://openknowledge.worldbank.org/server/api/core/bitstreams/b96b361a-a806-5567-8e8a-b14392e11fa0/content

Hu, S., Ge, Y., Liu, M., Ren, Z., & Zhang, X. (2022). Village-level poverty identification using machine learning, high-resolution images, and geospatial data. *International Journal of Applied Earth Observation and Geoinformation*, *107*, 102694. https://doi.org/10.1016/j.jag.2022.102694

CAF. (2022). *5 datos sobre pobreza en América Latina y el Caribe*. Obtenido de https://www.caf.com/es/actualidad/noticias/2022/04/5-datos-sobre-pobreza-en-america-latina-y-el-caribe/

DANE. (2022). *Publicación de pobreza monetaria extrema y pobreza monetaria*. Obtenido de https://www.dane.gov.co/files/investigaciones/condiciones_vida/pobreza/2021/Comunicado-expertos-pobreza-monetaria_2021.pdf

Espitia, G. y. (2022). *Población pobre en Colombia pierde 10 % de su ingreso por la inflación*. Obtenido de https://periodico.unal.edu.co/articulos/poblacion-pobre-en-colombia-pierde-10-de-su-ingreso-por-la-inflacion/