Problem Set 2: Predicting poverty

Sofia Charry Tobar
s.charry@uniandes.edu.co

Laura Manuela Rodriguez Morales
lm.rodriguezm@uniandes.edu.co

Nicol Valeria Rodríguez Rodríguez
nv.rodriguezr1@uniandes.edu.co

Brahyan Alexander Vargas Rojas
ba.vargas@uniandes.edu.co

El repositorio del ejercicio es: https://github.com/nvrr2028/Taller-2-BDML.git

1. Introducción

La pobreza es una de los grandes preocupaciones del desarrollo económico y la política pública. De hecho, dentro de los Los Objetivos de Desarrollo Sostenible (ODS) establecidos por Naciones Unidas, la eliminación de la pobreza en todas sus formas para 2030 es el primero de ellos, considerando que alrededor del 10 % de la población mundial (700 millones de personas) continúa en situación de pobreza extrema (CEPAL, 2018). De esta manera, en aras de contribuir a la formulación de una política integral y compleja que promueva el bienestar y la generación de oportunidades para la población, la construcción de herramientas que permitan la medición y predicción de la pobreza es de especial importancia.

En esta misma linea, el Departamento Administrativo Nacional de Estadística (DANE) pone a disposición del público el Índice de Pobreza Multidimensional (IPM)[1] y los reportes de pobreza monetaria[2], que tienen como propósito analizar y caracterizar la pobreza en Colombia (DANE, 2014; DANE, 2022). Adicionalmente, trabajos como los desarrolados por Barrientos et al. (2014), Hermes Sabogal y Granados (2021), Escudero y Cornejo (2022) y Ariza y Retajac (2020), entre otros, se han enfocado en profundizar el análisis y la predicción en

[1] https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-multidimensional

[2] https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-monetaria

Colombia y América Latina. Teniendo en cuenta lo anterior, en este documento se presenta un modelo basado en técnicas de machine learning (ML) que permite la predicción de la pobreza en Colombia a nivel hogar y, así, contribuir a la formulación de la política pública.

En el análisis de la pobreza en Colombia, se utilizarán datos provenientes del DANE y del MESEP, ambas instituciones cuentan con cuatro conjuntos de datos divididos en entrenamiento y prueba, tanto a nivel individual como de hogar. Para el estudio en cuestión, se empalmarán las bases de datos a nivel hogar e individual para tener un análisis más integral de las características de los hogares en Colombia.

La misión de la MESEP consiste en evaluar la comparabilidad de las cifras de mercado laboral y pobreza entre la Encuesta Continua de Hogares (ECH) y la Gran Encuesta Integrada de Hogares (GEIH). Para medir la pobreza monetaria en Colombia, la MESEP utiliza un método indirecto que define líneas de pobreza e indigencia y determina cuántos hogares no tienen ingresos suficientes para comprar una canasta básica de alimentos y otros bienes básicos. La metodología empleada por la MESEP incluye cambios en la línea de pobreza y en la construcción del agregado de ingreso del hogar, lo que permite una estimación más actualizada y precisa de la pobreza en Colombia.

Para predecir la pobreza en Colombia, se desarrollaron modelos de clasificación y regresión. Tras comparar su desempeño, se determinó que el modelo *Gradient Boosting Trees* presentaba el mejor rendimiento en la predicción de la pobreza, siendo las variables más importantes la proporción de personas en el hogar que pertenecen al régimen contributivo, la proporción de personas en el hogar que ocuparon la mayor parte del tiempo trabajando, el número de personas que hacen parte de la unidad de gasto, la proporción de personas en el hogar que pertenecen al régimen especial y la proporción de personas que alcanzaron un nivel educativo superior. Adicionalmente, se resalta que los modelos de clasificación presentan un mejor desempeño que los de regresión.

2. Data

2.1. Construcción base de datos: MESE

Para el ejercicio inicialmente manejamos cuatro bases de datos, dos para hogares y dos para personas con un train y test respectivamente. Las variables encontradas en las bases de train y test eran diferentes para cada modelo por lo que en principio identificamos cuáles compartían para trabajar con ellas. Por medio de la función *comparedf* identificamos que las bases de hogares compartían 16 variables y tenían 7 diferentes, en personas comparten 63 variables y difieren en 72.

Nuestro objetivo era determinar si un hogar es pobre, por lo cuál debíamos hacer el proceso

de adaptación para transformar la información de la base de personas. Del proceso anterior encontramos que la mayoría de las variables con las que quedábamos eran de respuesta binaria "sí" o "no", o de clasificación en categorías.

- Para el primer tipo "si" o "no" en la base personas convertimos los "sí"== 1 y "no"== 0. Nuestro siguiente paso fue sumar por id de hogar estos 1 y 0 por variable de forma que obtuviéramos cuantas personas del hogar tenían la variable. Por ejemplo, en un hogar de 5 personas, si 3 recibían subsidio alimenticio, encontraríamos 1= 3 y 0= 2, nuestra nueva variable por hogar sumaría 3. Y después calculamos la proporción por hogar como variable final para nuestro modelo.
- Para la clasificación por categorías, creamos variables binarias para cada categoría exceptuando una para no caer en el error de la multicolinealidad perfecta. De la misma forma, sumamos por id de hogar los valores 1 y 0 para obtener la cantidad de personas que cumplían con la característica en cuestión para luego calcular la proporción de personas en el hogar correspondiente con tal característica.

Los datos en la base construida a partir de la base del DANE es adecuada para realizar el ejercicio puesto que las variables predictoras corresponden a características de los hogares que están estrechamente relacionadas con el nivel de ingresos del hogar y con el estado del hogar como pobre o no. En este sentido, tenemos datos sobre las características de la casa que habitan, las cuales están relacionados con medidas multidimensional de la pobreza; si están desempleados; el tipo de trabajo que realizan; si reciben auxilios del gobierno y el régimen de salud al que pertenecen los miembros del hogar, que son variables que demuestran características relacionadas con la pobreza e ingresos. Así, estas variables nos ayudan a clasificar los hogares en pobreza o no y predecir su nivel de ingresos.

Cada una de nuestras nuevas variables de proporción fue creada en un data frame diferente y como estás ya estaban por hogares con ayuda de la función *join_all* construimos nuestra base final.

2.2. Análisis estimación descriptivo

Variables explicativas seleccionadas

El conjunto de variables seleccionadas para el análisis de predicción de la pobreza en Colombia, tanto para modelos de regresión como los de clasificación, se encuentra descrito en el Tabla 1. Lo que representa cada variable está descrito en el Anexo 1.

La base utilizada para el entrenamiento de los modelos cuenta con 164960 observaciones. Con respecto a nuestras variables dependientes, el *Ingreso hogar (Ingtotug)* y *Pobre*, podemos observar que alrededor del 20 % las familias en Colombia son pobres. Además, el

ingreso familiar promedio es de $2'090 895, considerando que el hogar está conformado en promedio por 3 personas. Adicionalmente, la proporción promedio de personas del hogar que recibe algún tipo de apoyo económico o en especie es reducido. Por otro lado, en promedio alrededor del 43,4 % de los miembros de un hogar reportan pertenecer a los regímenes de salud subsidiado y contributivo, mientras que tan solo el 0,4 % indica cotizar a pensión. Con respecto al nivel de educación de los integrantes de la unidad, la proporción promedio que indica tener un nivel de educación primaria y superior es del 13,8 % y 15,7 %, respectivamente. En cuanto al uso el tiempo, en promedio, el 25,0 % de las personas en un hogar señalan estar trabajado. Finalmente, en relación con las modalidades de empleo, la proporción de trabajadores independientes se lleva el mayor peso, con un 13,4 % de los miembros del hogar con esta forma de empleabilidad, seguido por los trabajadores en empresa privada (10,4 %).

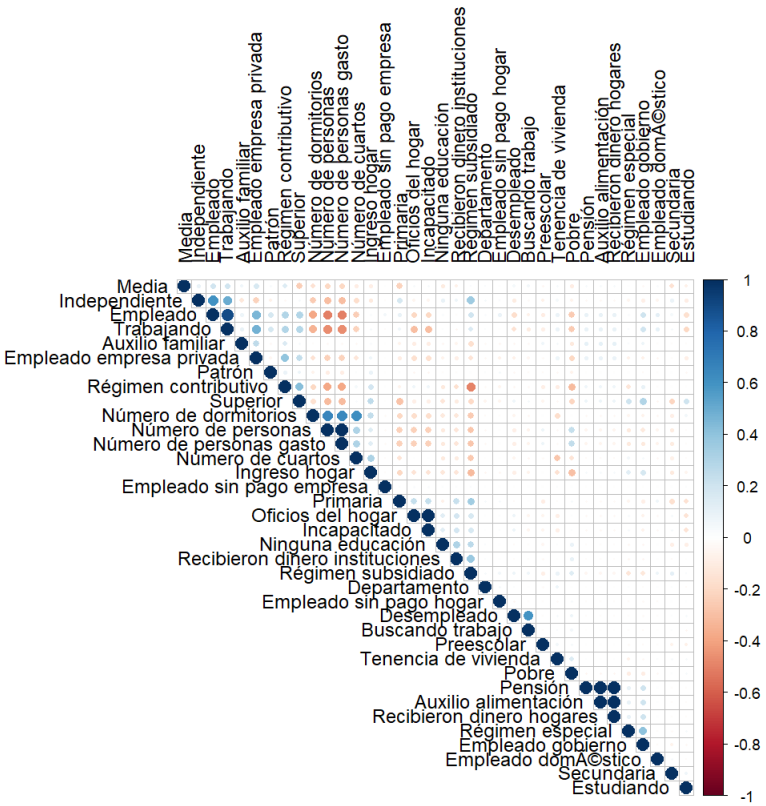
Tabla 1: Resumen descriptivo de la base de datos

Statistic	N	Mean	St. Dev.	Min	Max
Número de personas	164,960	3.292	1.775	1	28
Número de personas gasto	164,960	3.280	1.772	1	28
Ingreso hogar	164,960	2,090,895.000	2,512,488.000	0.000	85,833,333.000
Número de pobres	164,960	0.828	1.886	0	28
Auxilio alimentación	164,960	0.004	0.044	0.000	1.000
Auxilio familiar	164,960	0.021	0.079	0.000	1.000
Recibieron dinero instituciones	164,960	0.035	0.134	0.000	1.000
Recibieron dinero hogares	164,960	0.004	0.044	0.000	1.000
Pensión	164,960	0.004	0.044	0.000	1.000
Desempleado	164,960	0.029	0.103	0.000	1.000
Empleado	164,960	0.293	0.287	0.000	1.000
Régimen subsidiado	164,960	0.199	0.276	0.000	1.000
Régimen contributivo	164,960	0.235	0.294	0.000	1.000
Régimen especial	164,960	0.025	0.116	0.000	1.000
Ninguna educación	164,960	0.030	0.122	0.000	1.000
Preescolar	164,960	0.009	0.036	0.000	1.000
Primaria	164,960	0.138	0.228	0.000	1.000
Secundaria	164,960	0.081	0.162	0.000	1.000
Media	164,960	0.120	0.204	0.000	1.000
Superior	164,960	0.157	0.253	0.000	1.000
Trabajando	164,960	0.250	0.280	0.000	1.000
Buscando trabajo	164,960	0.014	0.076	0.000	1.000
Estudiando	164,960	0.055	0.120	0.000	1.000
Oficios del hogar	164,960	0.112	0.191	0.000	1.000
Incapacitado	164,960	0.112	0.191	0.000	1.000
Empleado empresa privada	164,960	0.104	0.203	0.000	1.000
Empleado gobierno	164,960	0.018	0.098	0.000	1.000
Empleado doméstico	164,960	0.008	0.058	0.000	1.000
Independiente	164,960	0.138	0.238	0.000	1.000
Patrón	164,960	0.012	0.080	0.000	1.000
Empleado sin pago hogar	164,960	0.006	0.036	0.000	1.000
Empleado sin pago empresa	164,960	0.001	0.023	0.000	1.000

En el Gráfico 1 se presenta el mapa de correlación entre las variables seleccionadas. Por un lado, para el *Ingreso hogar (Ingtotug)*, se puede observar que esta variable mantiene una correlación negativa con la proporción de integrantes del hogar que: no tiene un nivel educativo o que solo llega a preescolar, primaria y secundaria, reporta haber realizado oficios del hogar, estar incapacitados o estudiando la semana pasada, pertenece al régimen subsidiado y recibió un apoyo económico de alguna institución. Asimismo, la variable presenta una correlación positiva con la proporción de personas que tiene un nivel educativo de categoría superior y pertenece al régimen contributivo, así como también con el número de cuartos y dormitorios en la vivienda y el número de personas integrantes de la unidad de gasto. Finalmente, la variable *Ingreso hogar (Ingtotug)* no tiene correlación significativa al 5 % con la proporción de personas que recibieron un auxilio familiar, empleadas y que son trabajadores/as domésticos/as.

Por otro lado, para la variable *Pobre*, esta presenta una correlación negativa con la proporción de integrantes del hogar que: tienen un nivel educativo medio y superior, reporta haber estar trabajando, pertenece al régimen contributivo, recibió un auxilio familiar y es empleado del gobierno, de una empresa privada o independiente. Asimismo, la variable presenta una correlación positiva con la proporción de personas que tiene no tiene ningún nivel educativo y pertenece al régimen subsidiado, así como también con el número de personas del hogar e integrantes de la unidad de gasto. Finalmente, la variable *Pobre* no tiene correlación significativa al 5 % con el departamento y la proporción de personas que trabajan sin pago en una empresa.

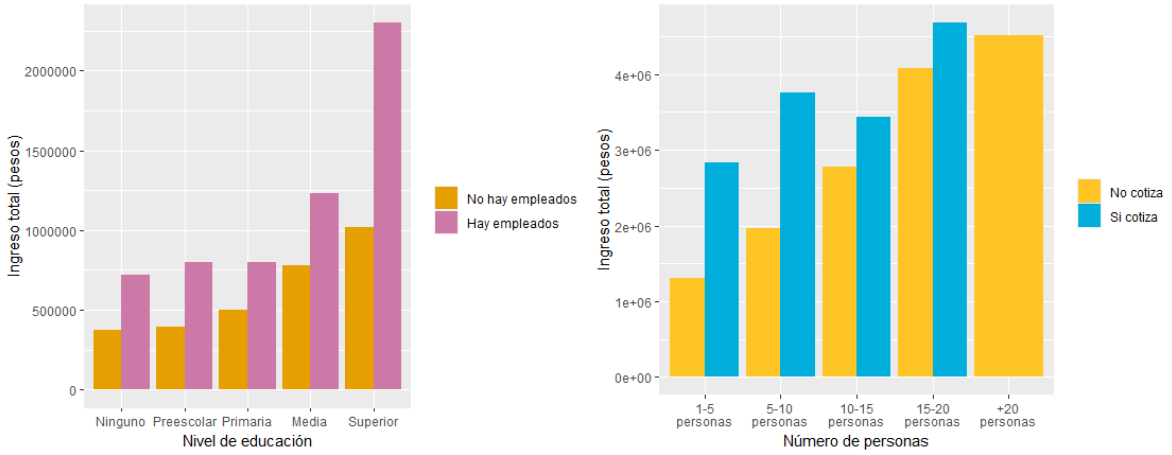
Gráfico 1: Mapa de correlaciones del conjunto de variables seleccionadas



Nota: El tamaño de los círculos indica la magnitud de la correlación entre las variables. Además, se excluyen los coeficientes de correlación que no sean signficativos al 5 %. Fuente: cálculos de los autores.

De acuerdo con el Gráfico 2 Panel (a), conforme hay presencia de personas con un mayor nivel educativo en el hogar, se va incrementando el nivel de ingreso. Además, el ingreso mediano total es siempre mayor cuando hay ocupados en el hogar, sin importar el nivel educativo. Por otra parte, el Panel (b) nos indica que, entre más personas hagan parte de la unidad de gasto del hogar, el ingreso es superior. Igualmente, si en el hogar hay personas que cotizan a pensión, se podría esperar un mayor nivel de ingreso mediano.

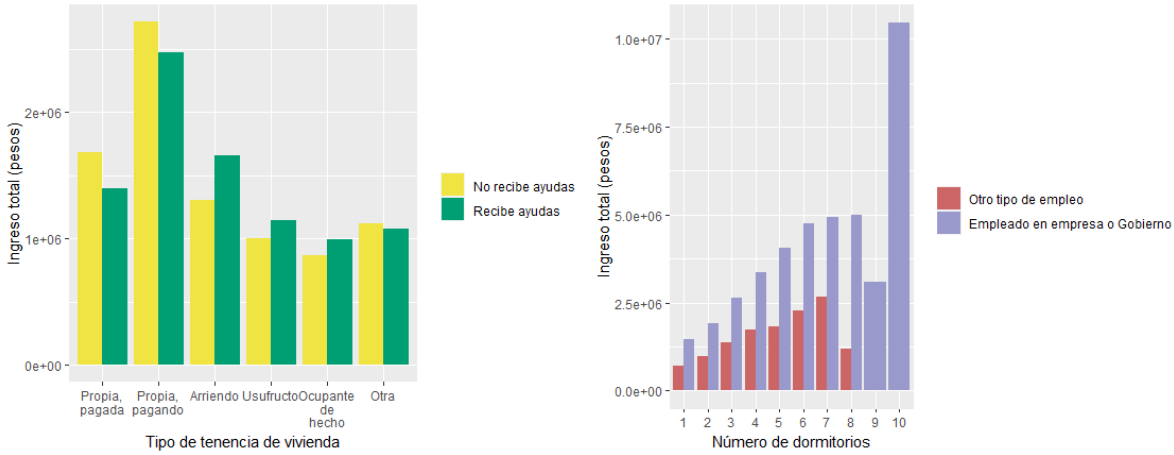
Gráfico 2: Ingreso mediano total del hogar de acuerdo con ciertas características



(a) (b)

Por último, el Gráfico 3 Panel (a) nos indica que los hogares que cuentan con una vivienda propia y que la están pagando, presentan un mayor nivel de ingreso mediano entre todas las formas de tenencia. Por demás, al tener en cuenta si el hogar recibe algún tipo de ayuda (subsidio de alimentación/familiar o apoyo económico de alguna institución u hogar), se observa que esto suele beneficiar a los hogares bajo un esquema de propiedad diferente al de vivienda propia. En cuanto al Panel (b), la presencia de empleados de empresas privadas o del Gobierno en el hogar suele favorecer el ingreso mediano total, lo cual también va incrementando con el número de dormitorios en la vivienda.

Gráfico 3: Ingreso mediano total del hogar de acuerdo con ciertas características



(a) (b)

3. Modelos: predicción de la pobreza

A continuación, se describen los modelos implementados para la predicción de la pobreza a nivel hogar en Colombia, a partir de la base de datos MESEP del DANE. La evaluación de los modelos se realizo de la siguiente manera:

1. La muestra de entrenamiento **train_hogares** se dividió en un dos conjuntos adicionales, de tal manera que el 70 % se utilizara en el entrenamiento de los modelos, y el 30 % restante en su evaluación. Este procedimiento se debe a que en la base de **test_hogares** no están las variables dependientes *Ingreso hogar (Ingtotug)* y *Pobre*.

2. Una vez se hayan entrenado los modelos y verificado que no hay problemas de *overfitting*, se predicen los hogares pobres y no pobres a partir de la base de **test_hogares**, para luego introducir los resultados en Kaggle y obtener la medida de evaluación *accuracy*.

3.1. Modelos de regresión

En el caso de los modelos de regresión, el objetivo es predecir la variable *Ingreso hogar (Ingtotug)* a partir del conjunto de variables explicativas descritas en la Sección 2.2, para luego comparar tal predicción con la *Línea de pobreza* establecida por el DANE. De esta manera, se logra identificar qué hogares son pobres (pobre=1) o no (pobre=0), lo cual también permitiría comparar con los modelos de clasificación desarrollados.

Teniendo en cuenta lo anterior se estimaron 4 técnicas: regresión lineal, Lasso, Elastic net y Gradient Boosting. En particular, la regresión utilizada para todos los modelos está consignada en el Anexo 2. De esta manera, se utilizan más de 33 regresores en la predicción dado que algunas de las variables son categóricas y se introducen como tal para evitar la trampa de las dummies. Cabe mencionar que, para todos los escenarios, las variables son estandarizadas.

Por otro lado, se utiliza la técnica de remuestreo *cross-validation* con k=10, con el objetivo de tener una mejor medida del desempeño de los modelos. Junto a lo anterior, la grilla utilizada para la identificación de los hiperparámetros utilizados para Lasso, Elastic Net y GBM son:

```
# Lasso:
expand.grid(alpha = 1, lambda = seq(0.001,0.02,by = 0.001))
# Elastic Net:
expand.grid(alpha = seq(0,1,by = 0.01), lambda = seq(0.001,0.1,by = 0.001))
# GBM
```

```
expand.grid(n.trees=c(200,300,500),interaction.depth=c(1,2,3),
shrinkage=c(0.01,0.001),n.minobsinnode=c(10,30))
```

3.2. Modelos de clasificación

En el caso de los modelos de clasificación, el objetivo es predecir si nuestra observación debe ser categorizada como *Pobre* a partir del conjunto de variables explicativas descritas en la Sección 2.2. De esta manera, se logra identificar qué hogares son pobres (pobre=1) o no (pobre=0).

Teniendo en cuenta lo anterior se estimaron las siguientes técnicas: regresión lineal, Lasso, Elastic net, Remuestreo, Cambio de pesos en la función de pérdida, optimización del umbral de decisión y Gradient Boosting. En particular, la regresión utilizada para todos los modelos es como se muestra en el Anexo 2, que es el mismo utilizado para los modelos de clasificación. De acuerdo con lo anterior, se utilizan más de 33 regresores en la predicción dado que algunas de las variables son categóricas y se introducen como tal para evitar la trampa de las dummies. Cabe mencionar que, para todos los escenarios, las variables son estandarizadas.

Por otro lado, se utiliza la técnica de remuestreo *cross-validation* con k=10, con el objetivo de tener una mejor medida del desempeño de los modelos. Junto a lo anterior, los hiperparámetros de los mejores modelos para Lasso, Elastic Net y GBM son:

```
# Lasso:
alpha= 0, lambda = 0.006056
# Elastic Net:
alpha = 1, lambda = 0.000251
# Elastic Net. Variación Regresión:
alpha = 1  , lambda = 0.000251
# Remuestreo Upsample
alpha = 1, lambda = 0.000438
# Cambio de pesos en la función de pérdida
alpha = 1, lambda = 0.00025174
# Random Forest
mtry= 5, splitrule= gini, min.node.size= 70
# GBM
n.trees= 300,interaction.depth= 3,
shrinkage= 0.01, n.minobsinnode= 30
```

3.3. Modelo final

Tras un exhaustivo análisis de los modelos de predicción de ingreso y clasificación, se determinó que el modelo de *Gradient Boosting Trees* de los modelos de clasificación era el más adecuado para ser el modelo final. En términos generales, el modelo de Gradient Boosting Trees va construyendo modelos más simples secuencialmente al predecir el error que deja el modelo anterior .

Tabla 2: Modelo principal vs Modelos de regresión

Estadística	Modelo principal	Regresión lineal	Lasso	Elastic Net	GBM
Accuracy (Kaggle)	0.8583	0.75349	0.75349	0.75462	0.78810
Balanced Accuracy	0.7268	0.5202	0.5202	0.5202	0.5000
Sensitivity	0.9516	0.9332	0.9334	0.9334	1.0000
Specificity	0.5021	0.1072	0.1069	0.1069	0.0000
Kappa		0.0527	0.0527	0.0527	0.0000
RMSE		2148874		2148724	2047868

Tabla 3: Modelo principal vs Modelos de clasificación

Estadística	Modelo principal	Random Forest	Lasso	Elastic Net 1	Elastic Net 2	Upsample	Función de pérdida
Accuracy (Kaggle)	0.858	0.836	0.000	0.853	0.856	0.000	0.813
Balanced Accuracy	0.726	0.847	0.847	0.86	0.79	0.79	0.811
Sensitivity	0.951	0.68	0.400	0.54	0.40	0.87	0.0823
Specificity	0.502	0.39	0.711	0.69	0.69	0.48	0.517
F1		0.49	0.5121	0.59	0.60	0.62	0.0635

El modelo final escogido es ***Gradient Boosting Trees***. Fue entrenado a partir del método que consiste en crear varios predictores en secuencia. El primer predictor usa la media de la variable Y para predecir, luego el segundo predictor explica los errores del primer predictor, el tercer predictor explicar los errores del segundo predictor y así sucesivamente. Construye el modelo de forma escalonada, basado en modelos de predicción más "débilesz los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable.

Los hiperparámetros escogidos son 300 árboles, 30 observaciones por nodo, 0.01 shrinkage y 3 en máxima profundidad del árbol. La regresión utilizada en el modelo está descrita en el Anexo 2. Igualmente, las variables con mayor importancia son, en orden: proporción de personas en el hogar que pertenecen al régimen contributivo, proporción de personas en el hogar que ocuparon la mayor parte del tiempo trabajando, número de personas que hacen parte de la unidad de gasto, proporción de personas en el hogar que pertenecen al régimen especial y proporción de personas que alcanzaron un nivel educativo superior. Para ver la importancia de las demás variables, ver Anexo 3.

Frente a los modelos de regresión, (ver Tabla 2), se evidencia que estos presentan un desempeño considerablemente inferior al modelo principal, tanto para las métricas de *accuracy* como *balanced accuracy*. Además, los modelos de regresión cuentan con un sesgo a favor

de la sensibilidad, es decir, los modelos tienen a predecir correctamente los hogares que no son pobres. Por el contrato, la especificidad tiende a ser inferior, por lo que los modelos presentan dificultades a la hora de predecir los hogares que son pobres.

En lo que respecta a los modelos de clasificación (ver Tabla 3), todas las alternativas registran *accuracy* superiores al $0{,}8$, siendo señal de que la estrategia de clasificación puede ser más apropiada que la de regresión, para este problema. Además, con respecto al *balanced accuracy*, se obtienen registros de alrededor del $0{,}8$, indicando un buen desempeño de los modelos. Con respecto a la sensibilidad y especificidad, el modelo principal presenta un sesgo a favor de la primera, mientras que para las demás opciones existe heterogeneidad.

4. Conclusiones y recomendaciones

Con el propósito de predecir la pobreza a nivel hogar, se estimaron múltiples modelos basados en técnicas de regresión y clasificación de Machine Learning. De tal análisis, resaltamos que los modelos de *Gadrient Boosting Trees*, *Elastic Net* y *Random Forest* son los que presentan mejor desempeño en la tarea de pronóstico. En específico, el modelo *Gadrient Boosting Trees* sería la de mayor la alternativa de utilidad en la formulación de políticas públicas enfocadas en la pobreza en Colombia.

Además, identificamos que las variables de mayor utilidad en la caracterización de la pobreza a nivel hogar son la proporción de personas en el hogar que pertenecen al régimen contributivo, la proporción de personas en el hogar que ocuparon la mayor parte del tiempo trabajando, el número de personas que hacen parte de la unidad de gasto, la proporción de personas en el hogar que pertenecen al régimen especial y la proporción de personas que alcanzaron un nivel educativo superior, por lo que el seguimiento de estos determinantes es fundamental. Cabe mencionar que la literatura ha reconocido a la educación como el mejor mecanismo de movilidad social y reducción de pobreza, así como también se resalta el papel de transferencias monetarias condicionadas o no condicionadas y otros instrumentos, como herramientas convenientes para atender trampas de pobreza.

Referencias

Ariza, J. y Retajac, A. (mayo de 2020). "Descomposición y determinantes de la pobreza monetaria urbana en Colombia. Un estudio a nivel de ciudades". En: *Estudios Gerenciales* 36, págs. 167-176. DOI: 10.18046/j.estger.2020.155.3345.

Barrientos, J., Ramírez, S. y Tabares, E. (dic. de 2014). "El patrón de crecimiento económico y la pobreza en Colombia". En: *Perfil de Coyuntura Económica*. DOI: 10.17533/udea.pece.n24a01.

CEPAL (2018). *The 2030 Agenda and the Sustainable Development Goals: An opportunity for Latin America and the Caribbean*. Inf. téc. CEPAL. URL: https://repositorio.cepal.org/bitstream/handle/11362/40156/25/S1801140_en.pdf.

DANE (jun. de 2014). *Metodología índice de pobreza multidimensional (IPM)*. Inf. téc. DANE. URL: https://microdatos.dane.gov.co/index.php/catalog/735/related-materials.

— (Abril de 2022). *Medición de Pobreza Monetaria y Desigualdad 2021*. Inf. téc. DANE. URL: https://microdatos.dane.gov.co/index.php/catalog/733/related-materials.

Escudero, W. S. y Cornejo, M. (2022). "Predicciones agregadas de pobreza con información a escala micro y macro: evaluación, diagnóstico y propuestas". En: *Serie Estudios Estadísticos* 103, pág. 76. ISSN: 1680-8789. URL: https://hdl.handle.net/11362/33066.

Hermes Sabogal, O. G.-B. y Granados, O. M. (2021). "Un análisis de la pobreza en Colombia basado en aprendizaje automático". En: *International Economic Review* 14. URL: http://hdl.handle.net/20.500.12010/22282.

5. Anexos

Anexo 1

- **Ingtotug (Ingreso hogar):** variable continua que representa el ingreso total de la unidad de gasto antes de imputación de arriendo a propietarios y usufructuarios.
- **Pobre (Pobre):** variable discreta que toma el valor de 1 cuando el hogar es pobre, y 0 en otro caso.
- **Número de cuartos (P5000):** variable discreta que responde a la pregunta Ïncluyendo sala-comedor ¿de cuántos cuartos en total dispone este hogar?".
- **Número de dormitorios (P5010):** variable discreta que responde a la pregunta "¿En cuántos de esos cuartos duermen las personas de este hogar?".
- **Tenencia de vivienda (P5090):** variable discreta que responde a la pregunta "La vivienda ocupada por este hogar es".
- **Número de personas (Nper):** variable continua que representa el número de personas en el hogar.
- **Número de personas gasto (Npersug):** variable continua que representa el número de personas en la unidad de gasto.
- **Departmanento (Depto):** variable discreta que representa el departamento.
- **Auxilio alimentación (prop_P6585s1h):** variable que representa la proporción de personas que recibieron un auxilio de alimentación en el hogar.
- **Auxilio familiar (prop_P6585s3h):** variable que representa la proporción de personas que recibieron un auxilio familiar en el hogar.
- **Recibieron dinero instituciones (prop_P7510s3h):** variable que representa la proporción de personas que recibieron dinero de instituciones en el hogar.
- **Recibieron dinero hogares (prop_P7505h):** variable que representa la proporción de personas que recibieron dinero de otros hogares en el hogar.
- **Pensión (prop_P6920h):** variable que representa la proporción de personas en el hogar que estarían cotizando a pensión.
- **Desempleado (prop_Desh):** variable que representa la proporción de personas en el hogar desempleadas.
- **Empleado (prop_Och):** variable que representa la proporción de personas en el hogar empleadas.

- **Régimen subsidiado (prop_subsidiado):** variable que representa la proporción de personas en el hogar cotizantes al régimen subsidiado.
- **Régimen contributivo (prop_contributivo):** variable que representa la proporción de personas en el hogar cotizantes al régimen contributivo.
- **Régimen especial (prop_especial):** variable que representa la proporción de personas en el hogar cotizantes al régimen especial.
- **Ninguna educación (prop_ningunoeduc):** variable que representa la proporción de personas en el hogar con ningún nivel de educación.
- **Preescolar (prop_preescolar):** variable que representa la proporción de personas en el hogar con nivel de educación preescolar.
- **Primaria (prop_basicaprimaria):** variable que representa la proporción de personas en el hogar con nivel de educación básica primaria.
- **Secundaria (prop_basicasecundaria):** variable que representa la proporción de personas en el hogar con nivel de educación básica secundaria.
- **Media (prop_media):** variable que representa la proporción de personas en el hogar con nivel de educación media.
- **Superior (prop_superior):** variable que representa la proporción de personas en el hogar con nivel de educación superior.
- **Trabajando (prop_mayoriatiempotrabajo):** variable que representa la proporción de personas en el hogar que pasaron la mayor parte del tiempo trabajando.
- **Buscando trabajo (prop_mayoriatiempobuscandotrabajo):** variable que representa la proporción de personas en el hogar que pasaron la mayor parte del tiempo buscando trabajo.
- **Estudiando (prop_mayoriatiempoestudiando):** variable que representa la proporción de personas en el hogar que pasaron la mayor parte del tiempo estudiando.
- **Oficios del hogar (prop_mayoriatiempooficiohogar):** variable que representa la proporción de personas en el hogar que pasaron la mayor parte del tiempo haciendo oficios del hogar.
- **Incapacitado (prop_mayoriatiempoincapacitado):** variable que representa la proporción de personas en el hogar que pasaron la mayor parte del tiempo incapacitado.
- **Empleado empresa privada (prop_obreroemplempresa):** variable que representa la proporción de personas en el hogar que son empleados de una empresa privada.
- **Empleado gobierno (prop_obreroemplgobierno):** variable que representa la proporción de personas en el hogar que son empleados del gobierno.

- **Empleado doméstico (prop_empldomestico):** variable que representa la proporción de personas en el hogar que son empleados domésticos.
- **Independiente (prop_trabajadorcuentapropia):** variable que representa la proporción de personas en el hogar que son cuenta propia.
- **Patrón (prop_patronempleador):** variable que representa la proporción de personas en el hogar que son patrón.
- **Empleado sin pago hogar (prop_trabajadorsinremunfamilia):** variable que representa la proporción de personas en el hogar que son trabajadores de la familia sin remuneración.
- **Empleado sin pago empresa (prop_trabajadorsinremunempresa):** variable que representa la proporción de personas en el hogar que son trabajadores de una empresa sin remuneración.

Anexo 2

Ecuación modelos de regresión

$$
\begin{aligned}
Ingtotug_i = \beta_1 P5000_i + \beta_2 P5010_i + \beta_3 P5090_i + \beta_4 Nper_i + \beta_5 Npersug_i + \beta_6 Depto_i \\
+ \beta_7 prop\_P6585s1h_i + \beta_8 prop\_P6585s3h_i + \beta_9 prop\_P7510s3h_i + \beta_{10} prop\_P7505h_i \\
+ \beta_{11} prop\_P6920h_i + \beta_{12} prop\_Desh_i + \beta_{13} prop\_subsidiado_i + \beta_{14} prop\_contributivo_i \\
+\beta_{15} prop\_especial_i + \beta_{16} prop\_ningunoeduc_i + \beta_{17} prop\_preescolar_i + \beta_{18} prop\_basicaprimaria_i \\
+\beta_{19} prop\_basicasecundaria_i + \beta_{20} prop\_media_i + \beta_{21} prop\_superior_i + \beta_{22} prop\_mayoriatiempotrabajo_i \\
+ \beta_{23} prop\_mayoriatiempobuscandotrabajo_i + \beta_{24} prop\_mayoriatiempoestudiando_i \\
+ \beta_{25} prop\_mayoriatiempooficiohogar_i + \beta_{26} prop\_mayoriatiempoincapacitado_i \\
+\beta_{27} prop\_obreroemplempresa_i + \beta_{28} prop\_obreroemplgobierno_i + \beta_{29} prop\_empldomestico_i \\
+ \beta_{30} prop\_trabajadorcuentapropia_i + \beta_{31} prop\_patronempleador_i + \beta_{32} \\
prop\_trabajadorsinremunfamilia_i + \beta_{33} prop\_trabajadorsinremunempresa_i
\end{aligned}
$$

Ecuación modelos de clasificación

$$
\begin{aligned}
Pobre_i = \beta_1 P5000_i + \beta_2 P5010_i + \beta_3 P5090_i + \beta_4 Nper_i + \beta_5 Npersug_i + \beta_6 Depto_i \\
+ \beta_7 prop\_P6585s1h_i + \beta_8 prop\_P6585s3h_i + \beta_9 prop\_P7510s3h_i + \beta_{10} prop\_P7505h_i \\
+ \beta_{11} prop\_P6920h_i + \beta_{12} prop\_Desh_i + \beta_{13} prop\_subsidiado_i + \beta_{14} prop\_contributivo_i \\
+\beta_{15} prop\_especial_i + \beta_{16} prop\_ningunoeduc_i + \beta_{17} prop\_preescolar_i + \beta_{18} prop\_basicaprimaria_i \\
+\beta_{19} prop\_basicasecundaria_i + \beta_{20} prop\_media_i + \beta_{21} prop\_superior_i + \beta_{22} prop\_mayoriatiempotrabajo_i \\
+ \beta_{23} prop\_mayoriatiempobuscandotrabajo_i + \beta_{24} prop\_mayoriatiempoestudiando_i \\
+ \beta_{25} prop\_mayoriatiempooficiohogar_i + \beta_{26} prop\_mayoriatiempoincapacitado_i \\
+\beta_{27} prop\_obreroemplempresa_i + \beta_{28} prop\_obreroemplgobierno_i + \beta_{29} prop\_empldomestico_i \\
+ \beta_{30} prop\_trabajadorcuentapropia_i + \beta_{31} prop\_patronempleador_i + \beta_{32} \\
prop\_trabajadorsinremunfamilia_i + \beta_{33} prop\_trabajadorsinremunempresa_i
\end{aligned}
$$

Anexo 3

Importancia de las variables

```
> prop_contributivo              100.00000
> prop_mayoriatiempotrabajo      84.1311524
> persug prop_especial           20.44407
> prop_superior                  15.54652
> prop_obreroemplempresa         9.71021
> P50903                         9.49986
> P50905                         8.13923
> Nper                           3.96584
> prop_subsidiado                3.21927
> prop_ningunoeduc               1.59694
> prop_Desh                      1.40373
> prop_obreroemplgobierno        1.16507
> prop_patronempleador           1.08008
> Depto27                        0.98622
> prop_P7510s3h                  0.79340
> Depto19                        0.79319
> prop_mayoriatiempoestudiando   0.43764
> prop_basicaprimaria            0.14804
> prop_mediaOverall.09364        0.09091
```