Universidad de Los Andes
Facultad de Economía
Big Data

Grupo No. 2 - Integrantes:
Alexandra Rizo
Id. 202210094

Danna Camila Bolaños
Id. 201911675

Héctor David Taticuán
Id. 202225884

Carlos David Vergara Díaz
Id. 201414896

Fecha de entrega: 26 de febrero de 2023.

Taller Grupal No. 2
Problem Set No. 2 – Predicting Poverty

1. Introducción:

Por medio de este documento, se presentan los resultados del estudio de la Base de datos de Medición de Pobreza Monetaria y Desigualdad de 2018 para el territorio Colombiano, a través de diferentes modelos de clasificación y regresión. Estos modelos, fueron debidamente cargados a la plataforma de Kaggel en miras de competir con los demás frente a aquel con mejor ajuste.

1.1. La predicción de la pobreza como un asunto relevante para la política pública:

En el contexto económico, político y social, la generación de programas, políticas y proyectos que busquen enfrentar la pobreza, es uno de los principales objetivos de la política pública a nivel nacional, regional e internacional[1]. En el ámbito internacional, 149 miembros de las naciones unidas ratificaron, a inicios del milenio, un compromiso multidimensional tendiente a combatir la pobreza, a partir de la adopción de esquemas gubernamentales y de cooperación comunes a nivel mundial, considerando como finalidad la erradicación del fenómeno[2].

No obstante, la definición de un estándar objetivo donde se pueda determinar cuando una persona o un hogar se encuentra en una situación de pobreza, puede variar según los contextos propios de cada país[3]. De allí, que la definición y cuantificación de como una persona u hogar es pobre, permite generar un marco de información mucho más claro para que las instituciones estatales promuevan las acciones necesarias para enfrentarla.

En el caso colombiano la definición de la pobreza se fundamenta en un estándar a partir de la línea de pobreza, como indicador desarrollado por el DANE. Así, un hogar, cuyos ingresos se encuentre por debajo de este valor, será para estos efectos un hogar "pobre". Cabe aclarar que este indicador, esto es, el de cada hogar, se calcula tomando los ingresos que recibe el hogar, y dividiéndolo entre

---

[1] Janeth Patricia Muñiz Eraso. *La pobreza y las políticas públicas: del referencial global al sectorial.* Cuadernos del CENDES, Vol. 32, No. 88, Caracas, Venezuela. (2015). Disponible en: http://ve.scielo.org/scielo.php?script=sci_arttext&pid=S1012-25082015000100006#:~:text=A%20nivel%20mundial%20existe%20un,metas%20de%20las%20pol%C3%ADticas%20p%C3%BAblicas.

[2] Naciones Unidas, Asamblea General. Resolución No. A/RES/55/2 de 13 de septiembre de 2000, *Declaración del Milenio.* Al respecto, indicó la Asamblea General en los considerandos 11 y siguientes del mencionado acto: "11. *No escatimaremos esfuerzos para liberar a nuestros semejantes, hombres, mujeres y niños, de las condiciones abyectas y deshumanizadoras de la pobreza extrema, a la que en la actualidad están sometidos más de 1.000 millones de seres humanos. Estamos empeñados en hacer realidad para todos ellos el derecho al desarrollo y a poner a toda la especie humana al abrigo de la necesidad.*
12. Resolvemos, en consecuencia, crear en los planos nacional y mundial un entorno propicio al desarrollo y a la eliminación de la pobreza.
13. El logro de esos objetivos depende, entre otras cosas, de la buena gestión de los asuntos públicos en cada país. Depende también de la buena gestión de los asuntos públicos en el plano internacional y de la transparencia de los sistemas financieros, monetarios y comerciales. Propugnamos un sistema comercial y financiero multilateral abierto, equitativo, basado en normas, previsible y no discriminatorio.
[...]"

[3] Erika Yurany Villazón Castro. *Evolución Metodológica de la Medición de la Pobreza en Colombia.* Rev. Inclusión & Desarrollo, No. 4, (2015).

el número de individuos que lo habita, para efectos de calcular el ingreso monetario de cada individuo. En lo que respecta al cálculo de la línea de pobreza, este indicador es definido por el DANE como la cantidad de dinero que necesita una persona para solventar sus necesidades básicas[4]. A esto se le conoce como el indicador de pobreza monetaria, al estar relacionado con los ingresos el hogar y la persona.

Adicionalmente, en Colombia existe otro indicador ampliamente utilizado, relativa a la pobreza multidimensional, no obstante, en este trabajo nos concentraremos en la pobreza monetaria[5].

1.2. Resumen de los resultados:

El estudio se fundamentó en la construcción de un modelo predictivo frente a la pobreza de los hogares a partir de la siguiente ecuación:

$$Poor = I(Income < Pl)$$

Donde I es la función de ingresos del hogar y, reiterando que esta debe ser menor a la línea de pobreza descrita para los hogares en igual condición.

Para el análisis de la pobreza en los hogares se realizan distintos tratamientos a las bases de datos iniciales debido a la presencia de valores nulos, variables con distintas categorías (categóricas) multicolinealidad, desbalance de clases, entre otros. Se realizan dos tipos de modelos predictores:

Método directo: este tipo de modelos incorporo procesos categóricos de distintas variables presentes en la base de datos con el fin de analizar sus características, aprender de estas y predecir si indicasen 1 o 0 en la variable pobre ubicando a los hogares dentro o fuera del estado de pobreza correspondientemente.

Método indirecto: Este proceso difiere del anterior ya que se realiza una estimación en primera medida de los Ingreso total de la unidad de gasto antes de imputación de arriendo a propietarios y usufructuarios, con los cuales se analizarán si son superiores o inferiores a la línea de pobreza y de esta forma indicar a que categoría de nuestra variable pobre pertenecen ya sea 1 = pobre 0= no pobre.

Del análisis realizado se obtuvieron los siguientes resultados:

Para los modelos de clasificación:

Por tratarse de un modelo desbalanceado, se realizó una aproximación por medio de un modelo de regresión lineal y un modelo de clasificación. En clasificación se utilizaron diferentes métricas de

[4] Departamento Nacional de Planeación, Dirección de Desarrollo Social. *Pobreza monetaria y pobreza multidimensional: Análisis 2008-2018.* (2019). Disponible en: https://colaboracion.dnp.gov.co/CDT/Desarrollo%20Social/Documento%20de%20An%C3%A1lisis%20de%20las%20Cifras%20de%20Pobreza%202018.pdf

[5] Según el Departamento Nacional de Planeación: *El Índice de Pobreza Multidimensional (IPM), desarrollado por el Oxford Poverty & Human Development Initiative (OPHI), es un indicador que refleja el grado de privación de las personas en un conjunto de dimensiones. La medida permite determinar la naturaleza de la privación (de acuerdo con las dimensiones seleccionadas) y la intensidad de la misma EL IPM es la combinación del porcentaje de personas consideradas pobres, y de la proporción de dimensiones en las cuales los hogares son, en promedio, pobres).* Ver: Departamento Nacional de Planeación. *Índice de Pobreza Multidimensional (IMP-Colombia). 1997-2008 y meta del PND para 2014.* (s.f.). Disponible en: https://colaboracion.dnp.gov.co/CDT/Estudios%20Econmicos/%C3%8Dndice%20de%20Pobreza%20Multidimensional%20(IPM-Colombia)%201997-2008.pdf

balanceo y regularización, propendiendo una mayor sensibilidad, dado el tema tratado, esto es, la pobreza de los hogares.

Se decide tomar todas las variables disponibles, optimizando el algoritmo para elegir el mejor modelo por medio de regularización con Ridge, Lasso y Elastic net. Por lo tanto, por medio de un algoritmo se evaluó 100 modelos. Del análisis realizado se pudo determinar que:

Con la muestra desbalanceada se obtuvo un SCORE más alto, de 0.865 que corresponde a la precisión de los datos de prueba y un valor más bajo de 0.845 para los datos desbalanceados.

Respecto a la sensibilidad de los modelos con train, se observa que es mayor en la muestra balanceada, pero al querer una sensibilidad de 1, la curva ROC cae a cerca del 50%. Finalmente se cambió el límite de 0.5 a 0.4, permitiendo clasificar, de una mejor manera, a los hogares pobres.

Para los modelos de regresión lineal:

Frente a este tipo de modelos, se planteó un modelo inicial, sustrayendo las variables multicolineales, los *missing values*, etc. Se generó un proceso de sub-muestreo para validación para efectos de compararlos con los datos de prueba. El resultado dio cero (0.0) con una precisión del 79%.

Modelo final:

Se escogió como modelo definitivo el de Regularización: Lasso, Lambda: 0.001, Precisión Kaggle: 0.84u Porcentaje hogares pobres: 25.23%

Como nos enfocamos en este informe en predecir correctamente el mayor número de hogares pobres se elige el modelo balanceado, dado que la precisión es un poco más baja, pero es el que clasifica un número mayor de hogares pobres porcentualmente.

Tabla de resultados del modelo escogido

Muestra balanceada					
ID	**alpha**	**Lambda**	**ROC**	**Precisión**	**Sensibilidad**
6	1.00	0.001	0.9118	0.86812	0.9503

1.3. Estructura del documento:

Para llegar a lo anterior, y considerando lo expuesto, el documento se dividirá en los siguientes acápites, a saber: (i). Presentación del enlace de acceso al repositorio en GitHUB; (ii). Proceso de limpieza de datos y estadísticas descripticas; (iii). Modelo de clasificación; (iv). Modelo de regresiones; (v). Conclusiones.

2. Enlace a repositorio de GitHub:

El enlace de Github donde podrá encontrarse el repositorio con las respuestas del taller es el siguiente:

Enlace al repositorio en GitHub
https://github.com/Carlosvergara1995/Problem_Set_2-_Predicting_Poverty.git

3. Descripción de los datos:

A continuación, se presenta el proceso de limpieza de datos y las correspondientes estadísticas descriptivas:

3.1. Origen de los datos:

Para efectos de este taller, se utilizaron los datos de la Gran Encuesta Integrada de Hogares – GEIH 2018, y dimensión particular para hogares y personas en relación con la pobreza monetaria y desigualdad, ambas creadas por el DANE, según reporte de 2019. Las bases de datos estudiadas se dividían en dos grupos: (i). Bases de datos de testeo (testing) y (ii). Bases de datos de entrenamiento (training), (en adelante los "Grupos"). Cada grupo se componía en dos bases de datos correspondientes a personas y hogares.

3.2. Proceso de limpieza de datos:

Inicialmente, y dado el tamaño de las bases, y las limitaciones de carga del aplicativo GitHub, se modifican las bases de formato cvs a rds.

Considerando las instrucciones del taller, se procedió a tomar la base de hogares y personas de cada uno de los grupos, con el propósito de que la información concerniente a personas complementara la relativa a hogares. Para estos efectos, y considerando que las variables en las bases de entrenamiento y testeo deben ser iguales en cuanto a la existencia de variables independientes, se verificaron las variables de cada una de las bases, se identificó a las variables categóricas y posteriormente se modificaron las variables necesarias para unir la base de datos y de personas en cada Grupo.

Se advierte que, en el caso de la base de testeo, no se contaba con la variable "Pobre", que para efectos de este ejercicio es la variable dependiente. Igualmente, se pudo identificar que existían muchas variables que no existían en la base de testeo, pero si en las de entrenamiento, por lo que se procedió a sustraerlas.

De todo este proceso surgen dos bases unificadas, correspondientes a hogares con el complemento de los datos de las bases de personas, tanto para el grupo de testeo, como para el grupo de entrenamiento.

Luego, y una vez unidas las bases, se identificaron los *missing values* en cada una, para remplazarlos.

Una vez realizado lo anterior, y dada la existencia de variables categóricas, estas se modifican para convertirlas en variables dummy, para efectos de poder ejecutar los análisis de clasificación y regresión. Con esto, se guardan las bases de datos finales: dvf_test de testeo y dvf_train = para entrenamiento.

3.3. Estadísticas descriptivas:

A continuación, se presenta de manera breve las estadísticas de las bases de datos finales:

Frente a la base de datos de entrenamiento (dvf_train), se pudo evidenciar que es una base compuesta por setenta (70) columnas y 164,960 observaciones. En relación con el número de personas pobres se encontró que el 20% es pobre y el 80% no es pobre.

A continuación, se presentan las estadísticas descriptivas, aplicables para las bases de datos de entrenamiento (dvf_train), y de testeo (dvf_test), según el resultado de la limpieza y preparación de datos, a saber:

Tabla No. 1 – Estadísticas descriptivas de entrenamiento

Variable	Pobre	
	0, N = 131,936 (No es pobre)	**1**, N = 33,024 (Es pobre)
Nper	3 (2)	4 (2)
Npersug	3 (2)	4 (2)
tipo_vivienda		
Propia	53,196 (40%)	9,080 (27%)
En proceso de pago	5,077 (3.8%)	549 (1.7%)
An arriendo	49,999 (38%)	14,454 (44%)
Sub-arrendada	19,634 (15%)	5,231 (16%)
Usufructo	3,917 (3.0%)	3,657 (11%)
Posesión de título	113 (<0.1%)	53 (0.2%)
Nro_cuartos	3 (1)	3 (1)
Nro_personas_cuartos	1.60 (0.67)	2.25 (1.15)
arriendo	179,219 (1,034,194)	138,667 (207,860)
Nro_mujeres	1.62 (1.10)	2.24 (1.35)
edad_promedio	39 (17)	31 (16)
jefe_hogar_mujer	53,555 (41%)	15,446 (47%)
Nro_hijos	1.02 (1.01)	1.74 (1.32)
Nro_personas_trabajo_formal		
0	64,481 (49%)	28,992 (88%)
1	47,401 (36%)	3,903 (12%)
2	17,159 (13%)	123 (0.4%)
3	2,490 (1.9%)	6 (<0.1%)
4	351 (0.3%)	0 (0%)
5	46 (<0.1%)	0 (0%)
6	3 (<0.1%)	0 (0%)

Tabla No. 1 – Estadísticas descriptivas de entrenamiento

Variable	Pobre	
	0, N = 131,936 (No es pobre)	**1**, N = 33,024 (Es pobre)
7	3 (<0.1%)	0 (0%)
8	2 (<0.1%)	0 (0%)
edu_promedio	7.08 (2.77)	6.72 (2.91)
Nro_personas_subsidio_familiar	1 (2)	1 (1)
horas_trabajadas_promedio	40 (19)	34 (22)
Nro_personas_empleo_propio	7 (6)	3 (3)
Nro_personas_segundo_trabajo		
0	122,497 (93%)	31,204 (94%)
1	8,675 (6.6%)	1,677 (5.1%)
2	704 (0.5%)	130 (0.4%)
Nro_personas_arriendos		
0	97,353 (74%)	31,473 (95%)
1	30,078 (23%)	1,507 (4.6%)
2	4,161 (3.2%)	44 (0.1%)
3	315 (0.2%)	0 (0%)
4	28 (<0.1%)	0 (0%)
5	1 (<0.1%)	0 (0%)
Nro_personas_pensiones		
0	109,387 (83%)	32,592 (99%)
1	19,887 (15%)	427 (1.3%)
2	2,489 (1.9%)	5 (<0.1%)
3	159 (0.1%)	0 (0%)
Nro_personas_pension_alimenticia		

Tabla No. 1 – Estadísticas descriptivas de entrenamiento

Variable	Pobre	
	0, N = 131,936 (No es pobre)	**1**, N = 33,024 (Es pobre)
0	130,026 (99%)	32,577 (99%)
1	1,871 (1.4%)	437 (1.3%)
Nro_personas_otros_ingresos		
0	77,335 (59%)	15,210 (46%)
1	41,139 (31%)	13,616 (41%)
2	11,030 (8.4%)	3,470 (11%)
3	2,003 (1.5%)	622 (1.9%)
4	342 (0.3%)	92 (0.3%)
Nro_personas_otros_ingresos_pais		
0	104,544 (79%)	23,318 (71%)
1	23,703 (18%)	8,272 (25%)
2	3,206 (2.4%)	1,235 (3.7%)
3	395 (0.3%)	175 (0.5%)
Nro_personas_otros_ingresos_otros_paises		
0	129,177 (98%)	32,571 (99%)
1	2,492 (1.9%)	430 (1.3%)
Nro_personas_otros_ingresos_instituciones		
0	117,512 (89%)	22,521 (68%)
1	12,595 (9.5%)	8,721 (26%)
2	1,685 (1.3%)	1,567 (4.7%)
3	125 (<0.1%)	192 (0.6%)
Nro_personas_otras_ganancias		
0	130,516 (99%)	32,939 (100%)
1	1,340 (1.0%)	84 (0.3%)
2	77 (<0.1%)	1 (<0.1%)
3	3 (<0.1%)	0 (0%)
Nro_personas_PET	3 (1)	3 (2)
Nro_personas_ocupadas	2 (1)	1 (1)
Nro_personas_desempleadas		
0	114,197 (87%)	24,292 (74%)
1	16,100 (12%)	7,215 (22%)
2	1,489 (1.1%)	1,292 (3.9%)

Tabla No. 1 – Estadísticas descriptivas de entrenamiento

Variable	Pobre	
	0, N = 131,936 (No es pobre)	**1**, N = 33,024 (Es pobre)
3	131 (<0.1%)	191 (0.6%)
Nro_personas_inactivas	1 (1)	1 (1)

Elaboración propia con fundamento en las bases de datos.

Tabla No. 2 – Estadísticas descriptivas base de datos de testeo

Variable	**N = 66,168**
Nper	3 (2)
Npersug	3 (2)
Pobre	33,024 (20%)
tipo_vivienda	
Propia	25,079 (38%)
En proceso de pago	2,138 (3.2%)
An arriendo	25,201 (38%)
Sub-arrendada	10,394 (16%)
Usufructo	3,321 (5.0%)
Posesión de título	35 (<0.1%)
Nro_cuartos	3 (1)
Nro_personas_cuartos	1.73 (0.83)
arriendo	421,771 (1,016,769)
Nro_mujeres	1.75 (1.19)
edad_promedio	37 (17)
jefe_hogar_mujer	27,525 (42%)
Nro_hijos	1.17 (1.12)
Nro_personas_trabajo_formal	
0	38,160 (58%)
1	20,311 (31%)
2	6,649 (10%)
3	900 (1.4%)
4	127 (0.2%)
edu_promedio	6.96 (2.87)
Nro_personas_subsidio_familiar	1 (2)
horas_trabajadas_promedio	45 (13)
Nro_personas_empleo_propio	6 (6)
Nro_personas_segundo_trabajo	

Tabla No. 2 – Estadísticas descriptivas base de datos de testeo

Variable	**N = 66,168**
0	61,570 (93%)
1	4,231 (6.4%)
2	332 (0.5%)
Nro_personas_arriendos	
0	52,232 (79%)
1	12,188 (18%)
2	1,621 (2.4%)
3	113 (0.2%)
Nro_personas_pensiones	
0	57,154 (86%)
1	7,997 (12%)
Nro_personas_pension_alimenticia	
0	65,338 (99%)
1	819 (1.2%)
Nro_personas_otros_ingresos	
0	37,670 (57%)
1	21,770 (33%)
2	5,519 (8.3%)
3	977 (1.5%)
4	189 (0.3%)

Tabla No. 2 – Estadísticas descriptivas base de datos de testeo

Variable	**N = 66,168**
Nro_personas_otros_ingresos_pais	
0	64,856 (98%)
1	1,199 (1.8%)
Nro_personas_otros_ingresos_instituciones	
0	55,777 (84%)
1	8,838 (13%)
2	1,399 (2.1%)
3	138 (0.2%)
Nro_personas_otras_ganancias	
0	65,652 (99%)
1	500 (0.8%)
2	16 (<0.1%)
Nro_personas_PET	3 (1)
Nro_personas_ocupadas	2 (1)
Nro_personas_desempleadas	
0	55,630 (84%)
1	9,254 (14%)
2	1,154 (1.7%)
3	113 (0.2%)
Nro_personas_inactivas	1 (1)

Elaboración propia con fundamento en las bases de datos.

De las anteriores tablas se puede apreciar lo siguiente:

Primero, en la base de datos de testeo existen 66.188 observaciones, mientras que en la base de datos de entrenamiento existen y 164,960.

Segundo, en la base de datos de entrenamiento, se pueden apreciar los cambios entre las observaciones, esto es, cuando un hogar es pobre y cuando no, y como esto se relaciona con diferentes variables como el número de personas que conforman el hogar, los ingresos en cada caso, el nivel educativo, entre otros.

Tercero, entre otros aspectos, en el 57% de los hogares en promedio sus individuos no cuentan con trabajo formal, el 42% de los hogares tienen como jefe de hogar a una mujer, se tiene una edad promedio de 34 años y 10,5 años de estudio en promedio, sin embargo, respecto a este último dato se evidencia brecha entre los hogares pobres y no pobres, teniendo como años promedio de educación 8 y 11 respectivamente.

Cuarto, se observa que a mayores años de educación disminuye la participación de los hogares pobres, siendo más notoria la brecha de ingresos, la cual también es evidente entre los hogares pobres y no pobres con igual número de años de educación.

Quinto, comparando estas estadísticas generales de la base training respecto a la base testing, se concluye que no existen diferencias significativas en la media de las variables presentadas, lo que quiere decir que ambas bases son muy similares. Lo anterior, en punto a los valores porcentuales de las variables, considerando que el número de observaciones en una y otra base son diferentes.

Sexto, finalmente, se debe advertir que se sustrajeron varios de los datos en las Tablas No. 1 y No. 2 por ser inferiores al 1%, no obstante, se tuvieron en cuenta en el análisis.

4. Modelo de clasificación:

Para el modelo de clasificación se eligió todas las variables que podían afectar el modelo, eliminando dominio, dado que esta variable esta correlacionada con la línea de pobreza. La línea de pobreza depende directamente de la ciudad en que se encuentre. Se decide tomar todas las variables disponibles sin agregar sesgo por selección de variables y se busca optimizar el algoritmo para elegir el mejor modelo por medio de regularización con Ridge, Lasso y Elastic net. Esto se realizó por medio de una grilla de interacción donde Alpha tomó valores entre 0 (Ridge) y 1 (Lasso) y lamba valores entre 0.001 y 1 (OLS-MCO). Por lo tanto, por medio de un algoritmo se evaluó 100 modelos. Adicionalmente se trabajó con un segundo conjunto de datos, donde se balanceo la muestra minoritaria (pobre) por medio de la función Rose en R, la cual se utiliza para realizar un sobremuestreo aleatorio de una clase minoritaria en un conjunto de datos desbalanceados. Se escogió este tipo de balanceo al usar la técnica SMOTE, una de las más utilizadas actualmente para tal finalidad, dado que, como plan de gobierno, este debe enfocarse en los pobres; además al tener una data desbalanceada la precisión del modelo se puede ver sesgado, dado que, si dice que nadie es pobre, se asegura de tener un R2 de al menos el 79%.

Tabla No. 3

	Balanceada	Desbalanceada
Pobre	40 %	20 %
No Pobre	60 %	80 %

Los resultados de los modelos más destacados con base en su curva ROC se presenta en la siguiente tabla:

Tabla No. 4 - con datos de entrenamiento

Desbalanceada					
ID	**alpha**	**Lambda**	**ROC**	**Precisión**	**Sensibilidad**
1	0.66	0.001	0.8610	0.77626	0.8326
2	0.55	0.001	0.8610	0.77622	0.8325
3	0.77	0.001	0.8610	0.77622	0.8327
4	0.44	0.001	0.8610	0.77620	0.8325
5	0.88	0.001	0.8609	0.77616	0.8327
Balanceada					
ID	**alpha**	**Lambda**	**ROC**	**Precisión**	**Sensibilidad**
6	1.00	0.001	0.9118	0.86812	0.9503
7	0.88	0.001	0.9115	0.86792	0.9501
8	0.77	0.001	0.9114	0.86785	0.9500

Tabla No. 4 - con datos de entrenamiento

Desbalanceada					
9	0.66	0.001	0.9113	0.86784	0.9499
10	0.55	0.001	0.9113	0.86778	0.9498

Se destaca que ambos casos se tiene un mejor modelo con Lambda muy bajos de 0.001 y el valor de Alpha es el que vario, iniciando con un modelo Lasso fuerte para los datos desbalanceada y disminuyendo. Para los datos desbalanceados Alpha no es el mayor.

Con estos resultados se procede a realizar la validación con la información de prueba de Kaggle, encontrando que el modelo con la muestra desbalanceada tiene una precisión más alta de 0.865, que corresponde a la precisión con los datos de prueba y un valor más bajo de 0.845 para los datos desbalanceados; de contar con los datos de prueba se procedería a calcular la sensibilidad del modelo y con esto elegir el mejor dado que como se mencionó se busca identificar los hogares pobres.

Respecto a la sensibilidad de los modelos con train, se observa que es mayor en la muestra balanceada, pero al querer una sensibilidad de 1, la curva ROC cae a cerca del 50%. Finalmente se cambia el límite de 0.5 a 0.4, esto ocasiona que existan más falsos positivos, pero nos permite clasificar mejor a los hogares pobres.

5. Modelo de regresiones:

Para los modelos de regresión se plantea un modelo inicial que incluya las variables previamente tratadas por multicolinealidad, estacionalidad y missing values; posteriormente se realizó el proceso de submuestreo con la división de train en train y validación con el fin de entrenar los datos con la submuestra train, y evaluarla dentro con la muestra de validación. Posteriormente, se evalúan los resultados fuera de muestra con el data de test.

A partir de estos datos, se analizan diferentes escenarios donde se observa la pobreza de manera indirecta, proporcionando predicciones evaluadas en nuestra muestra de prueba del ingreso total de la unidad de gasto del hogar (Ingtotug).

Posteriormente, con los resultados de ingresos calculamos el promedio y la mediana de la variable línea de pobreza (Lp) la cual nos indicara el punto de corte para nuestras predicciones de tal forma que si el Ingreso total de la unidad de gasto antes de imputación de arriendo a propietarios y usufructuarios predicho, es menor a la línea de pobreza, el hogar se encuentra en estado de pobreza y dado el caso contrario no es un hogar pobre.

Los resultados con los distintos puntos de corte fueron idénticos y se evidencia la tendencia de los modelos lineales a volverse “perezosos” debido al desbalance de los dataframe, dando como resultado una matriz que contiene únicamente valores de 0, categorizando a todas las observaciones como no pobres, de tal forma que el modelo presenta una exactitud del 79 % aproximadamente, fuera de muestra y una tasa de error equivalente al restante de 21 % aproximadamente.

Lo anterior, demostraría que los modelos realizados no están capturando la relación entre las variables de entrada y la variable de salida de manera efectiva; en otras palabras, los modelos de regresión no están aprendiendo de los datos provisionados y están generando valores constantes sin tener en cuenta las características de las observaciones a pesar de las distintas modificaciones realizadas.

En nuestro proceso de selección del mejor modelo predictor se descartaron los modelos de regresión generados debido a su básico desempeño tanto en las evaluaciones dentro de muestra y como fuera de la misma, aunque a pesar de esto se encuentran un porcentaje de exactitud del 79 %.

6. Modelo final:

A continuación, se presentan los datos de precisión y sensibilidad de validación en el modelo final de manera comparada.

Tabla No. 5 – Comparación de modelos datos de testeo

No.	Modelo	Precisión	Sensibilidad
1	Máximo Roc desbalanceado	0.8659748	0.9501591
2	Máximo Accuracy desbalanceado	0.8661300	0.9496976
3	Máximo sens desbalanceado	0.7987003	1.0000000
4	Máximo Roc balanceado	0.8443453	0.8757621
5	Máximo Accuracy balanceado	0.8443453	0.8757621
6	Máximo sens balanceado	0.7986615	1.0000000

Para tener un mejor análisis de los datos, se divide la data de prueba en prueba y en validación, así se analizó el modelo con datos que el modelo no conocía, obteniendo los resultados de la tabla 5, se destaca que los valores de la precisión son cercanos a los valores entregados por Kaggle en las prueba, adicionalmente la tabla 4 y 5 presentan diferentes valores de precisión dado que los valores de la tabla 4 son los valores entregados por el modelo, probados bajo los mismos datos de prueba y validado mediante validación cruzada.

Finalmente, el modelo elegido es el No. 4 de la Tabla No. 5, que corresponde al modelo No. 6 de la Tabla No. 4:

- Regularización: Lasso
- Lambda: 0.001
- Precisión Kaggle: 0.84
- Porcentaje hogares pobres: 25.23%

Como nos enfocamos en este informe en predecir correctamente el mayor numero de hogares pobres se elige el modelo balanceado, dado que la precisión es un poco más baja, pero es el que mayor porcentaje correctos de hogares pobres clasifica.

Tabla No. 6

	Estimación negativa	Estimación positiva
Real negativo	36 056	2 909
Real positivo	5 115	7 470

Se observa como reducimos el error tipo I aumentando el error tipo II, pero aumentando las predicciones correctas de hogares pobres.

7. Conclusiones:

Definir cuando un hogar es pobre o no, es una de las grandes cuestiones de la política pública, esto, en función de generar programas que puedan, de la mejor manera posible, modificar la situación de las personas en estado de vulnerabilidad. Por medio de este trabajo, y fundamentados en la información de la GEIH de 2018, se buscó realizar un análisis de las variables que definen a un hogar como pobre, a partir de una aproximación de modelos de clasificación y de regresión lineal, a partir de diferentes técnicas de machine learning.

Los modelos de clasificación calculados nos permiten tener una precisión alta, pero dado que la data era desbalanceada una precisión alta no nos asegura que el modelo clasifique correctamente en los datos de prueba, por lo cual realizar balanceo de información nos permite tener modelos más sensibles. Adicionalmente el cambio de parámetro límite de 0.5 a 0.4 disminuye la precisión dado que aumenta los eventos de hogares pobres y con esto los falsos positivos. La función grid search permite encontrar valores óptimos de modelos, con varios parámetros permitiendo elegir el modelo óptimo de manera sencilla, a pesar de tener una base de datos desbalanceada.

Balancear los datos aumenta las métricas en los modelos evaluados con los datos de prueba, pero bajada significativamente con los datos de validación. El modelo puede tender a predecir ceros y se encontró que, en el modelo lineal, predijo que todos los valores eran 0 (no pobres) y a pesar de esto la prueba le dio una precisión de cerca de 80%, pero una sensibilidad de cero debido a que no predijo bien ningún positivo, esto demuestra la dificultad para trabajar bases de datos desbalanceadas, dado que nuestro mejor modelo aumento únicamente en un 4% su precisión, pero aumenta significativamente en la sensibilidad.

Los modelos seleccionados pueden utilizarse para clasificar hogares de los que no se tenga información del ingreso promedio y servirán para orientar y dar línea a los gobiernos en la focalización adecuada de las políticas frente a esta población.

Los anterior, permite concluir que los modelos desbalanceados son difíciles de trabajar, dado que necesitan preprocesamiento. Se utilizó regresión lineal y clasificación. En clasificación se usaron diferentes métricas de balance cero y se utilizaron técnicas de regularización, buscando siempre una mayor sensibilidad para que de esta forma, las políticas públicas se dirijan hacia las personas de bajos recursos, debido a esto se probaron modelos de alta sensibilidad, la página en kaggle solo medía la precisión se decidió subir modelos que estaban con alta sensibilidad pero con bajo precisión para enfocarnos más en la gente pobre, a pesar de esto solo se logró subir al rededor del 6% a pesar de todas las pruebas hechas solo se logró aumentar en un 7%.