Universidad de los Andes
Maestría Economía Aplicada: *Big Data and Machine Learning for Applied Economics*
Grupo: Laura Natalia Capacho, Sebastián David Beltrán y Yurani Gonzalez[i]
GitHub URL: https://github.com/sbeltro/G10_PS2

Problem Set 2: Predicting Poverty[1]

1. Introducción:

En Colombia, el índice de pobreza monetaria fue de 39,3% en 2021 y la pobreza multidimensional de 16,0% (DANE, 2022). La ONU (2022) señala la erradicación de la pobreza como un imperativo ético, social, político y económico a nivel mundial, por esto, identificar correctamente los hogares que deben priorizarse es fundamental. El objetivo de este trabajo es construir un modelo predictivo de pobreza en hogares colombianos. Se caracterizó la pobreza desde dos enfoques: un problema de clasificación, donde encontramos que 25,7% de hogares en la base de prueba son pobres, y un modelo de predicción de ingreso comparado con la línea de pobreza, que indica que 32,1% de dichos hogares son pobres.

Destacamos que durante el estudio encontramos que las variables disponibles en la fuente de datos dispuesta para el ejercicio[ii], no permiten una aproximación sobresaliente de pobreza, pues las características del hogar que contiene son limitadas, y no se encuentran variables correspondientes a características de vivienda o condiciones de vida, comúnmente usadas para identificar pobreza (Kambuya, 2020). Para lograr una mejor caracterización de los hogares y enfocar de forma óptima las intervenciones es necesario ampliar las preguntas que contiene la encuesta para capturar otros factores, como lo hace la ELCA[iii].

2. Datos

En el estudio se utilizaron 4 bases[iv] de la GEIH 2018. En la tabla 1 se presentan las principales estadísticas descriptivas de las variables de interés con las que se entrenaron los modelos. Tenemos una muestra de 164,960 hogares, 33,024 clasificados como pobres y 131,936 no pobres. Se observa que en todas sus características existe una diferencia significativa entre los grupos: una mayor proporción de hogares pobres recibe subsidios (82,2% frente al 46% de no pobres), y ayudas monetarias de hogares nacionales e instituciones (29,6% y 31,6% respectivamente, frente al 20,6% y 11,0% en no pobres). Además, vemos que los hogares pobres duermen en promedio 2 personas por cuarto, el doble que en los no pobres.

En la tabla 2 se caracterizan los hogares que buscamos clasificar (*test)*. La muestra se compone de 66,168 hogares, de los cuales, el 56,2% recibe subsidios, 13,2% subsidio familiar y tan solo 0,2% educativo. Asimismo, encontramos que el 23,2% recibió ayuda de hogares nacionales y 15,7% de instituciones, y vemos que en esta muestra duermen cerca de 2 personas en cada cuarto. (En la Gráfica 1 se observan las características de las dos muestras).

[*] Todas las tablas y gráficas mencionadas en el documento se encuentran en la sección de Anexos al final.

3. Modelos y resultados

3.1 Modelos de clasificación

La primera aproximación de un modelo predictivo de pobreza se abordó a través de un problema de clasificación, donde utilizando un clasificador bayesiano[v] se asignan probabilidades a que un hogar sea pobre con base en sus características. En concreto, un hogar se clasificó: $Pobre = I(P > r)$, si la variable de pobreza estimada es mayor a la regla "r", entonces I es igual a 1, indicando que el hogar es pobre.

Estimación de pobreza: revisamos la base disponible y seleccionamos las características que permitirían identificar si el hogar es pobre, dada la escasa disponibilidad de información a nivel del hogar, se construyeron características agregadas a partir de variables individuales. Construimos un conjunto de 12 variables para hacer la estimación (Anexo 1). Posteriormente, hicimos una exploración con el método de *Best Subset Selection* para identificar aquellas que tendrían mayor relevancia y se evaluaron 6 modelos[vi]. En las estimaciones se priorizaron la sensibilidad (buscando un alto nivel en el indicador) y el ratio de falsos negativos (buscando un bajo nivel), y se ajustó una regla $r = 3$ como punto de corte.

Modelo seleccionado: luego de analizar los modelos y sus resultados en la matriz de confusión[vii] (Tabla 4), seleccionamos el modelo *logit* pues obtuvo los mejores resultados. Este modelo considera 4 variables explicativas: (i) *Subsidiado_hg* que toma el valor de 1 si al menos un miembro de hogar (MdH) pertenece al régimen subsidiado de salud; (ii) *ayudaInstituciones_hg* que indica si al menos un MdH recibió ayuda monetaria de una institución nacional o extranjera (=1), (iii) *personaxCuarto_hg* que indica la ratio entre MdH y cuartos donde duermen; y (iv) *educ_hg* que es el promedio de años de escolaridad de los MdH. Las variables fueron construidas en la base de personas de entrenamiento y se agruparon por hogar (*id) (*Tabla 3 contiene los resultados de la estimación)[viii].

El modelo seleccionado no tiene hiperparámetros alfa o lambda que busquen maximizar alguna métrica. No obstante, 3 de los 6 modelos estimados sí contaban con dichos hiperparámetros, $\alpha = 1$ que corresponde a estimaciones *Lasso*, λ tomó valores $[-3, 10]$ y buscó maximizar la sensibilidad y la curva *ROC*. De igual forma, como se trató de un modelo *logit* convencional, no se aplicó un método de *class imbalances*, sin embargo, 2 de los modelos considerados utilizaron los métodos de *Up-sampling* y *Down-sampling*[ix].

Desempeño del modelo: El modelo *logit* arrojó una sensibilidad (*Sensitivity)*[x], es decir, una proporción de hogares pobres clasificados correctamente de 57.8% y una tasa de 11.1% de falsos negativos[xi], que mide la proporción de verdaderos hogares pobres que el modelo clasificó erróneamente. Además, exhibe un tasa de falsos positivos[xii], que indica la proporción de "falsos pobres" sobre el total de hogares que el modelo clasifica como pobres, de 52.91%, 0.80 en el área bajo la curva (*AUC*[xiii], en inglés), y se observa que su curva ROC[xiv] refleja el mejor desempeño entre los modelos analizados (Gráfica 2).

3.2 Modelos de regresión de ingresos

Una segunda aproximación para construir un modelo predictivo de pobreza de los hogares es a través de su ingreso. Estimamos el ingreso del hogar a partir de la agregación de una estimación individual del ingreso de cada MdH[xv]. Luego, utilizando la línea de pobreza (Lp) definida por el DANE, se hace una comparación y clasifica el hogar, $Pobre = I(Inc < Pl)$. Si el ingreso estimado es menor a la Lp se clasifica como pobre, y la variable $I = 1$.

Estimación del ingreso: Reconocemos que los ingresos, sobre todo en hogares pobres, pueden provenir de diversas fuentes, que además de salario se componen de otros ingresos[xvi]. Sin embargo, dentro de las variables disponibles no contamos con información suficiente para estimar apropiadamente dichos ingresos. Por lo anterior, el ingreso predicho corresponde al ingreso laboral del individuo que logramos aproximar de mejor manera.

Para la estimación se usaron 2 métodos de selección de modelos: *Best Subset Selection* (*BSuS)* y *Backward Stepwise Selection (BSwS)*[xvii], que permiten encontrar las variables de interés (Anexo 2) que mejor ajustan el ingreso. Además, se estimaron 6 modelos usando el método de regularización *Elastic Net*[xviii], que fuerza a que los coeficientes del modelo tiendan a cero, minimizado el riesgo de sobreajuste (*overfitting)*, reduciendo la varianza, mitigando el efecto de correlaciones entre predictores y mejorando la estimación.

Después de realizar las estimaciones, se calculó el error cuadrático medio (*MSE* en inglés) de cada modelo (Tabla 5). En la Gráfica 3 se observa que 5 de los modelos estimados tienen un nivel de ajuste similar (*BSwS y* 4 de *Elastic Net*), mientras el modelo de *BSuS*, y 2 del otro método reportan un ajuste más deficiente frente a los demás. En el modelo (3) es donde encontramos el mejor ajuste, en la medida que presenta el menor error cuadrático medio.

Modelo seleccionado: El modelo (3) seleccionado corresponde a la estimación utilizando *Elastic Net*, y validación cruzada en 5 conjuntos[xix], de un modelo donde la variable dependiente es $y\_laboral$, que denota el ingreso laboral del individuo y las variables explicativas son: $microEmpresa$, una variable categórica que indica si la persona trabaja en microempresa (=1), $ocupado$, que determina si la persona se encuentra ocupada (=1), $educ$, es el nivel educativo medido en años y $oficio$, denota la ocupación del individuo.

En la estimación las variables fueron centradas y escaladas con media 0 y desviación estándar 1, para asegurar que la penalización que impone el modelo se aplique por igual sobre cada coeficiente. La estimación utilizó los hiperparámetros[xx] óptimos $\alpha = 0.1$ y $\lambda = 1{,}042.2$, que minimizan el *MSE*, valores que se encontraron empleando validación cruzada en 5 conjuntos.

El modelo de predicción del ingreso individual se entrenó con la base *train*, y luego se agregaron los ingresos predichos de los MdH para obtener un estimado del ingreso agregado del hogar, que comparamos frente a la Lp[xxi] y clasificamos los hogares en "pobre" y "no pobre". Encontramos que, de un total de 164,960 hogares, 52,035 son clasificados como pobres, mientras 112,925 son clasificados como no pobres.

Desempeño del modelo: Dado que conocemos la clasificación de los hogares en la base con la que entrenamos el modelo, podemos evaluar su desempeño. En la matriz de confusión para esta estimación (Tabla 6), se tiene que la sensibilidad (*Sensitivity)* es de 65.6%, y la tasa de falsos negativos, que mide la proporción de verdaderos pobres que el modelo clasifica como

no pobres, es de 10.0%. La tasa de falsos positivos es 58,3%, la especificidad (*Specificity*[xxii]) de 76.9% y la precisión (*accuracy)* de 74.7% con un $IC\ (95\%) = [74.5\%\ ,74.9\%]$.

4. Conclusiones y recomendaciones

El problema de medir correctamente la pobreza de los hogares ha sido abordado desde diferentes perspectivas, que dependen principalmente de la disponibilidad de información y el concepto de pobreza que se busque entender. En este trabajo, se hizo una aproximación de la pobreza desde dos enfoques: un problema de clasificación de los hogares en pobres y no pobres, a partir de características construidas con los datos disponibles, y un problema de predicción del ingreso, que al comparar con una línea de pobreza determinada para cada hogar permite clasificarlo como pobre o no pobre.

Desde el primer enfoque encontramos que variables que capturan la recepción de subsidios o ayudas, el espacio disponible para dormir en la vivienda y el nivel educativo promedio de los MdH, son las características que mejor permiten perfilar el hogar. De otro lado, características tales como trabajar en microempresa, estar ocupado, el tipo de ocupación y el nivel educativo, son los determinantes del ingreso individual que permitieron un mejor ajuste del ingreso agregado del hogar para caracterizar pobreza. En concreto, con el método de clasificación encontramos que el 25,7% de hogares en *test* son pobres, y con el modelo de predicción del ingreso comparado con la línea de pobreza, que el 32,1% de dichos hogares se clasifican como pobres.

Los hallazgos señalan que cerca del 30% de los hogares colombianos analizados en la muestra se encuentran en condiciones de pobreza, un porcentaje importante, si además consideramos que dentro de estos hogares se encuentran aquellos que además están en condición de pobreza extrema e indigencia. Resaltamos que la información disponible se enfoca más en las fuentes de ingresos por trabajo de los individuos, y no de otros ingresos, que representan un porcentaje importante de las entradas de hogares de bajos recursos. Asimismo, carecemos de información sobre las condiciones y calidad de vida de las personas, que suelen funcionar mejor como indicadores de pobreza.

De esta forma, es fundamental mejorar las fuentes de información para lograr un mejor perfilamiento de los hogares, para que los recursos puedan dirigirse a la población más vulnerable, se logre trabajar en iniciativas que fomenten la educación y el empleo, al tiempo que se reduzcan las disparidades en oportunidades, de modo que se logren incrementar los ingresos de dichos hogares y se fortalezcan sus medios de vida. Lo que, sumado a mejores condiciones de salubridad, nutrición, protección social, conectividad, productividad, entre otros, mejoraría sus condiciones de vida, bienestar y permitiría salir progresivamente de la pobreza.

Anexos

Anexo 1: variables de interés (clasificación de pobreza)

Adicional a las variables usadas en el modelo *logit* (*Subsidiado_hg, ayudaInstitucions_hg, personaxCuarto_hg y educ_hg*), se consideraron otras características: **subsidio familiar** - *subFamiliar_hg* una variable *dummy* que indica si algún MdH había recibido un subsidio familiar en el último mes (=1); **subsidio educativo** - *subEducativo_hg* variable *dummy* que denota si algún MdH había recibido un subsidio educativo en el último mes (=1); **ayuda de hogares nacionales** - *ayudaHogaresnal_hg* una *dummy* que indica si algún MdH había recibido ayuda monetaria de otro hogar o persona a nivel nacional en los últimos dos meses (=1); **ayuda de hogares extranjeros** - *ayudaHogaresext_hg dummy* que señala si algún MdH había recibido ayuda monetaria de otro hogar o persona del extranjero en los últimos dos meses (=1); **profesional en el hogar** - *profesional_hg* toma el valor de 1 si al menos un MdH tiene más de 12 años de educación, indicando que terminó bachillerato y tiene la posibilidad de aplicar a un título técnico o universitario; **trabajador de microempresa en el hogar** - *microempresa_hg* indica si al menos un MdH trabaja en una empresa con máximo 10 empleados (=1); **trabajador formal en el hogar** - *formal_hg* determina si al menos un MdH era cotizante o beneficiario de alguna entidad de seguridad social en salud, y por lo tanto, se considera trabajador formal; y **miembros del hogar** - *Nper* que indica el número de personas en el hogar.

Anexo 2: variables de interés (ingreso laboral)

Para la estimación del ingreso laboral de cada miembro del hogar se identificaron y seleccionaron un conjunto de variables que de acuerdo con la teoría económica tienen una influencia significativa sobre nuestra variable de interés (el ingreso laboral). En concreto, se consideró un conjunto de variables continuas: **edad** - *edad*, **edad al cuadrado** - *edad2* (para capturar el efecto decreciente de la edad sobre el ingreso), **experiencia potencial** - *experiencia* (construimos un *proxy* de experiencia laboral[xxiii], debido a que no se tiene información reportada de los años de experiencia laboral, utilizamos el concepto de "experiencia potencial", que se crea a partir de la edad, los años de educación y los años de iniciación en el mercado laboral (Aristizábal & Ángel, 2017)), **experiencia potencial al cuadrado** – *experiencia2* (para capturar el efecto decreciente que viene con la edad sobre el ingreso), y **nivel educativo** - *educ* (variable de educación medida en años que construimos a partir del máximo nivel educativo alcanzado que se reportó en la encuesta. En efecto, si este nivel es ninguno o no sabe se asignaron 0 años, si es preescolar 3 años, si es básica primaria 8 años, si es básica secundaria 12 años, si es media 16 años y si es superior o universitaria 20 años).

De otro lado, un conjunto de variables categóricas o *dummies*: **género** - *mujer* (variable *dummy* que toma el valor de 1 si el individuo es mujer y 0 en otro caso), **jefe de hogar** - *jefeHogar* (variable *dummy* que caracteriza si la persona es jefa del hogar o no), una **interacción entre jefe de hogar y género** – *jefeHogar_mujer*, **trabajo formal** - *formal* (variable *dummy* que se creó a partir de la variable p6090 "¿Está afiliado, es cotizante o es beneficiario de alguna entidad de seguridad social en salud?", puesto el empleo informal se

refiere a los trabajadores que, entre muchos factores, pertenecen a una empresa o desempeñan un trabajo sin contrato laboral y sin aportes a seguridad social (Departamento Administrativo Nacional de Estadística [DANE], 2009)), **microempresa -** *microEmpresa* (variable *dummy* que indica si la persona trabaja en una microempresa o no, a partir del tamaño de la empresa, medido por el número de trabajadores, se denotan como microempresas las compañías que tengan personal no superior a 10 trabajadores (Ministerio de Comercio, Industria y Turismo, 2007)), **ocupado -** *ocupado* (variable *dummy* que indica si la persona se encontraba dentro de la población ocupada en el momento de la encuesta), **segundo trabajo -** *segundoTrabajo* (variable *dummy* que indica si la persona tenía un segundo trabajo en el momento de la encuesta) y **oficio -** *oficio* (variable categórica que denota la ocupación de la persona y toma valores de 1 a 99).

Tablas

Tabla 1. Estadísticas descriptivas (base *train)**

	Hogares		
Variables	**Pobres**	**No pobres**	**p-valor****
Subsidiado	27,815	60,721	< 0.001
(=1)	(84.2)	(46.0)	
Subsidio familiar	1,481	20,759	< 0.001
(=1)	(4.5)	(15.7)	
Subsidio educativo	11	359	< 0.001
(=1)	(0.0)	(0.3)	
Ayuda hogares nal.	9,761	27,160	< 0.001
(=1)	(29.6)	(20.6)	
Ayuda hogares extr.	443	2,755	< 0.001
(=1)	(1.3)	(2.1)	
Ayuda instituciones	10,452	14,468	< 0.001
(=1)	(31.6)	(11.0)	
Profesional	27,975	117,955	< 0.001
(=1)	(84.7)	(89.4)	
Personas x cuarto	2.25	1.60	< 0.001
	(1.15)	(0.67)	
N	33,024	131,936	

*El dato corresponde al número de hogares que cumplen con dicha característica, y el valor entre paréntesis indica la proporción sobre el total. Únicamente para la variable de número de personas por cuartos donde duermes el número corresponde al promedio y el valor entre paréntesis a la desviación estándar.

** El p-valor corresponde a una prueba de diferencia de medias entre los dos grupos

Fuente: GEIH 2018, cálculos propios.

Tabla 2. Estadísticas descriptivas (base *test)**

Variables	Proporción
Subsidiado	37,175
(=1)	(56.2)
Subsidio familiar	8,723
(=1)	(13.2)
Subsidio educativo	140
(=1)	(0.2)
Ayuda hogares nal.	15,430
(=1)	(23.3)
Ayuda hogares extr.	1,326
(=1)	(2.0)
Ayuda instituciones	10,398
(=1)	(15.7)
Profesional	58,355
(=1)	(88.2)
Personas x	1.73
cuarto	(0.83)
N	66,168

*El dato corresponde al número de hogares que cumplen con dicha característica, y el valor entre paréntesis indica la proporción sobre el total. Únicamente para el número de personas por cuartos donde duermes el número corresponde al promedio y el valor entre paréntesis a la desviación estándar.

Fuente: GEIH 2018, cálculos propios.

Tabla 3. Modelo *logit*

	Pobre
	(2)
Intercepto	-2.05***
	(0.03)
Subsidiado	1.22***
(=1)	(0.02)
Ayuda instit.	0.46***
(=1)	(0.02)
Personas x	0.65***
cuarto	(0.01)
Educación	-0.12***
promedio	(0.00)
AIC	131,898.54
BIC	131,948.61
N	164,960

****p < 0.001; **p < 0.01; *p < 0.05*

Fuente: GEIH 2018, cálculos propios.

Tabla 4. Comparación de *AUC*, Falsos-positivos y Falsos negativos

Modelo	*AUC*	Falsos-positivos	Falsos-negativos
(1)	0.804897	0.529122	0.111858
(2)	0.733349	0.811092	0.699973
(3)	0.751427	0.801512	0.437553
(4)	0.752866	0.801332	0.428702
(5)	0.614599	0.838825	0.419662
(6)	0.645174	0.82601	0.459785

Fuente: Cálculos propios.

Tabla 5. Error cuadrático medio

Modelo	*MSE*
(1)	860,960
(2)	853,635
(3)	844,780
(4)	923,149
(5)	1,003,958
(6)	866,619
(7)	844,958
(8)	862,336

Fuente: Cálculos propios.

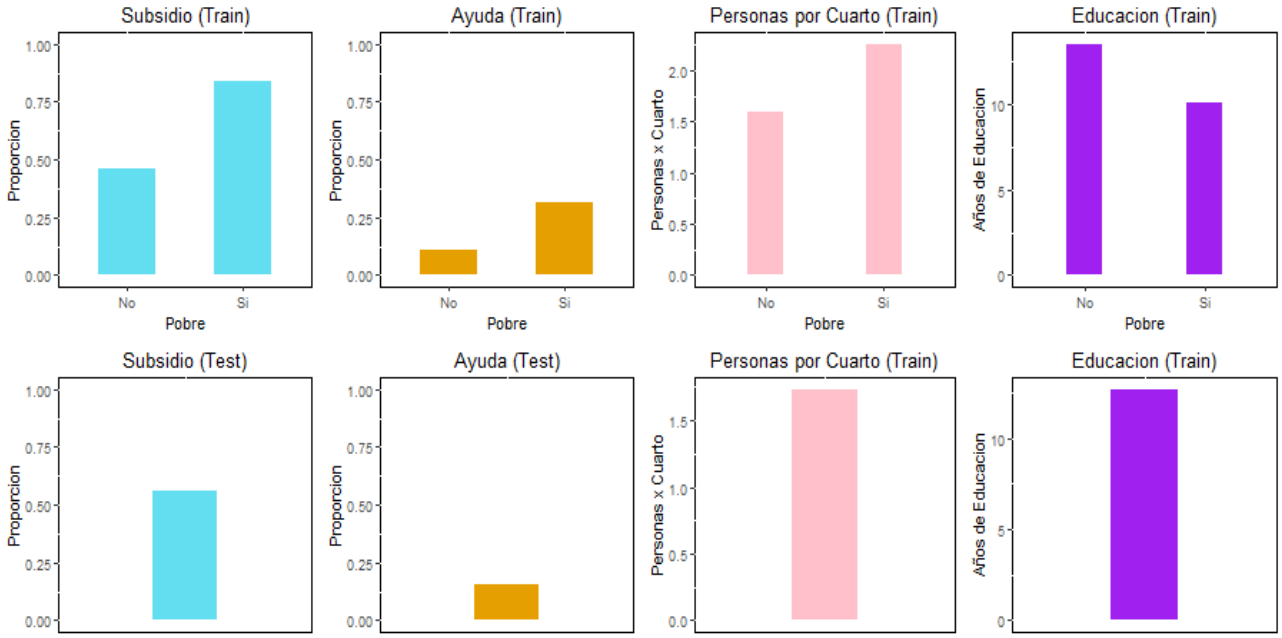
Tabla 6. Clasificación real y predicha de pobreza

		Clasificación real	
		No pobre	Pobre
Predicción	No pobre	101,581	11,344
	Pobre	30,355	21,680

Fuente: GEIH 2018, cálculos propios.

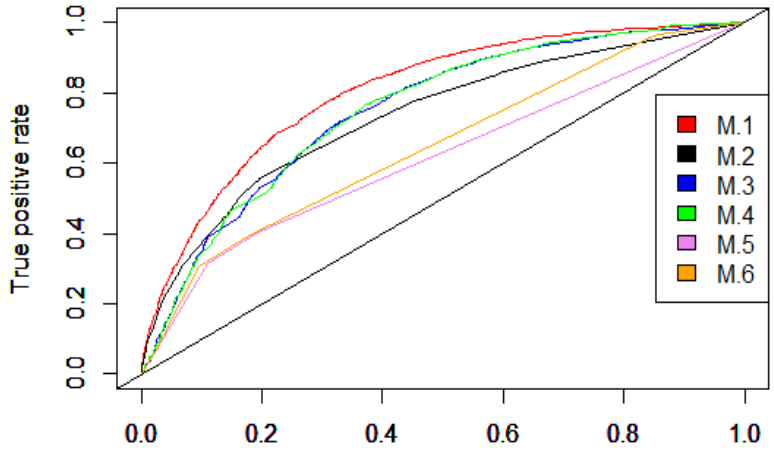
Gráficas

Gráfica 1. Características de los hogares (*Train* y *Test)*



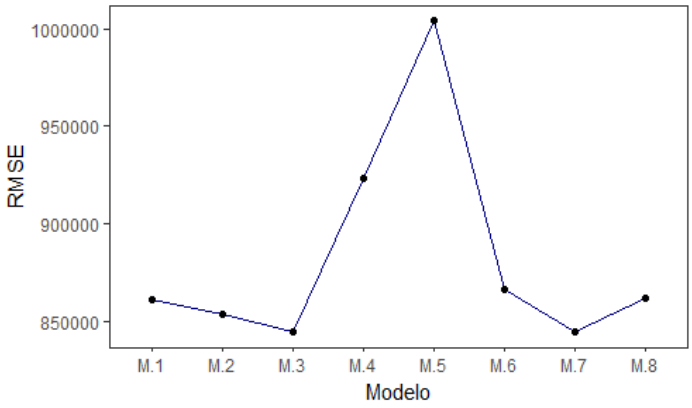
Fuente: Construcción propia.

Gráfica 2. Curva de ROC



Fuente: Construcción propia.

Gráfica 3. Error cuadrático medio*



** M.1 es la estimación con Best Subset Selection, M.2 con Backward Stepwise Selection, M.3 – M.8 son las estimaciones usando Elastic Net con distintas covariables.*

Fuente: Construcción propia.

Bibliografía

Aristizábal, T. & López, E. (2017). *Efectos de los aumentos en la escolaridad en el mercado laboral colombiano entre 2008 y 2016.* Ecos de Economía: A Latin American Journal of Applied Economics, 21(44).

Ayuda en Acción. (2020). *Seis formas de luchar contra la pobreza*. Fundación Ayuda en Acción [AeA].

Departamento Administrativo Nacional de Estadística [DANE] (2009). *Metodología informalidad Gran Encuesta Integrada de Hogares – GEIH*. Dirección de Metodología y Producción Estadística – DIMPE.

Departamento Administrativo Nacional de Estadística [DANE] (2022). *En 2021, la pobreza multidimensional en el país fue de 16,0%, 2,1 puntos porcentuales menos que en 2020 (18,1%).* Pobreza Multidimensional 2021 [Comunicado de presa].

Departamento Administrativo Nacional de Estadística [DANE] (2022). *Pobreza Monetaria 2021. Enfoque diferencial.*

Food and Agriculture Organization of the United Nations [FAO]. (2022). *Poverty Eradication.*

Hughey, J. & Butte, A. (2015). *Robust meta-analysis of gene expression using the elastic net*. Nucleic Acids Research, 43(12). DOI: 10.1093/nar/gkv229

Kambuya, P. (2020). *Better Model Selection for Poverty Targeting through Machine Learning: A Case Study in Thailand.* Thailand and The World Economy, 38(1).

Ministerio de Comercio, Industria y Turismo (2007). *Guía de contratación pública para micro y pequeñas empresas – MIPYME.* Ministerio de Comercio, Industria y Turismo.

United Nations (2022). *Poverty Eradication*. Department of Economic and Social Affairs. United Nations.

United Nations (2022). *Peace, dignity and equality on a healthy planet*. Department of Economic and Social Affairs. United Nations.

Zou, H. & Hastie, T. (2005). *Regularization and variable Selection via the elastic net*. Journal of the Royal Statistical Society: Series B (Statistical Methdology), 67(2). DOI: 10.1111/j.1467-9868.2005.00503.x

Notas

[i] Los códigos de los estudiantes del grupo son 202121025, 202121021 y 201212100, respectivamente.
[ii] (Gran Encuesta Integrada de Hogares 2018 del DANE).
[iii] Encuesta Longitudinal Colombiana de la Universidad de Los Andes.

[iv] Las bases se encuentran divididas en dos muestras, una de entrenamiento (*train*) y otra de prueba (*test*), de cada una se contó con una base de datos a nivel de personas y otra a nivel hogar. Dado que la estimación que nos interesa es pobreza del hogar, se exponen las características en esta agregación.
[v] El clasificador se entrenó en una base de entrenamiento donde conocemos la "clasificación correcta" de los hogares.
[vi] (1) *logit*; (2) *logit* con un control de *Two-Class Summary* y *5-fold Cross Validation (CV)*; (3) *logit* con un control de *5 Stats* y *5-fold CV*; (4) *logit* con control de *5 Stats* y *5-fold CV*, maximizando con *Lasso* la *ROC*; (5) *logit* con control de *5 Stats* y *5-fold CV*, maximizando con *Lasso* la sensibilidad y usando *Up-sampling*; y (6) *logit* con control de *5 Stats* y *5-fold CV*, maximizando con *Lasso* la sensibilidad y usando *Down-sampling.*
[vii] La matriz de confusión brinda información sobre la sensibilidad, ratio de falsos negativos y positivos y demás indicadores
[viii] Se observa que todas las variables resultaron significativas a un nivel del 1%, y además los coeficientes tienen los signos esperados.
[ix] *Up-sampling* es un método que simula puntos adicionales de la clase minoría (pobre) para balancear entre dos clases y *Down-sampling* reduce de manera aleatoria la clase mayoritaria (no pobre) para balancear las clases.
[x] *Sensitivity* es la ratio entre el total de hogares clasificados como pobres y el número real de dichos hogares.
[xi] La tasa de falsos negativos es la ratio entre el número falsos negativos (hogares mal clasificados como no pobres) sobre el total de hogares clasificados como no pobres por el modelo.
[xii] La tasa de falsos positivos es la ratio entre el número falsos positivos (hogares mal clasificados como pobres) sobre el total de hogares clasificados como pobres por el modelo.
[xiii] El *AUC* toma un valor entre 0 y 1, donde un valor cercano a 1 indica que el modelo está seleccionado a los verdaderos positivos y tiene un ratio bajo de falsos positivos.
[xiv] La *Receiver Operating Characteristic Curve (ROC)* es una curva que mide la predicción del modelo frente al ratio de verdaderos positivos y falsos positivos.
[xv] $Ingreso\_h = \sum_{i:1}^{N} Ingreso_i$, donde N es el número de miembros del hogar e $Ingreso_i$ es una función $f(x)$ del ingreso individual.
[xvi] Otros ingresos como subsidios, auxilios y ayudas del Estado, instituciones y otros hogares.
[xvii] *Best Subset Selection* combina los predictores disponibles y *Backward Stepwise Selection* (introduce todas las variables en la ecuación y excluye secuencialmente una tras otra)
[xviii] Este método es una combinación de las penalizaciones que imponen sobre los coeficientes lasso y ridge, dos métodos de regularización.
[xix] *5-fold Cross Validation* en inglés
[xx] En el modelo existen dos hiperparámetros: alfa (α) que controla el grado en que influye cada penalización y toma valores [0,1], así, si $\alpha = 1$ se aplica *Lasso* y si $\alpha = 0$ se aplica *Ridge;* y lambda (λ) que es el hiperparámetro de regularización.
[xxi] Se multiplicó la línea de pobreza (Lp) por el número de personas en la unidad de gasto, puesto que la Lp está en términos per cápita, y el ingreso estimado es agregado del hogar.
[xxii] *Specificity* es la ratio entre el total de hogares clasificados como no pobres y el número real de dichos hogares.
[xxiii] Para las personas con educación se utiliza la siguiente aproximación de experiencia (X): Si 18 < edad < 22, X = edad – 18; si edad > 22, X = edad – educación – 6. Y para personas sin educación terciaria se aproxima como sigue: si edad > 18, X = edad – 18.