Big Data – 202213
Maestría en Economía Aplicada – Universidad de los Andes

Taller 2
Julio 12, 2022

Integrantes: Ingrid Lorena Molano cód. 200511102
Jorge Eduardo García cód. 201310645
Camilo Villa Moreno cód. 201818624

Repositorio: https://github.com/camilovillam/MECA_BD_PS2

1. Introducción:

En el presente taller se aborda el problema de la clasificación de pobreza para hogares en Colombia, con base en diferentes métodos y algoritmos de aprendizaje de las máquinas. Los datos provienen de extracto de las encuestas de hogares y personas, GEIH del año 2018, elaborada por el DANE. Para este taller no se contó con todas las variables, lo cual representa desafíos adicionales al momento de hacer predicción y clasificación, pero que a su vez se asemeja a las limitaciones que se encontrarían en la práctica, lo cual invita a resolver el problema con pocas variables. La propuesta de solución incluye métodos de clasificación de pobreza y de predicción de ingresos. Así, se encontró que, en nuestro caso y con las variables disponibles, los métodos de clasificación de pobreza funcionan mejor que los de predicción de ingreso. Con 14 variables, nuestro mejor método de clasificación logró identificar correctamente al 82% de pobres y 75% de no pobres. Para mejorar el resultado y como futuro trabajo, se podrían usar algoritmos de *superlearners*, ponderando los diferentes métodos aquí usados, así como explorar otras posibles variables.

2. Datos:

Para seleccionar las variables se hizo una exploración de las bases de datos Train y Test y, dado que la predicción debía realizarse a nivel hogar se realizó un primer filtro de las variables en *train hogares* dejando aquellas que estaban en *test hogares* más las variables *Pobre*[1] e *Ingtotug*[2].

Luego, se identificaron que variables de la base de datos a nivel *personas* podrían agregarse a nivel *hogar* para ayudar a la predicción de un hogar pobre (modelo de clasificación) o para predecir los ingresos (modelo predicción variable continua) y se crearon nuevas variables que se agregaron a la base *Train* a nivel de hogares.

Un segundo filtro fue identificar los NAs en las variables para borrar aquellas variables que tenían más de 6% de NAs. Por último, para las variables que quedaron se identificaron las observaciones con NAs e infinitos y se borraron estas observaciones. Esta decisión se tomó para dejar la base *Train* con observaciones reales sin imputación. En el caso de la base Test, en vez de borrar las observaciones con NAs de las variables con menos de 6% de NAs, se imputaron valores.

Las variables seleccionadas para trabajar con las bases Train y Test según el procedimiento descrito, se presentan en la Tabla 1 del apéndice. Por su parte, las Tablas 2,3 y 4 del apéndice muestran las tablas descriptivas de hogares pobres y no pobres.

En principio, por la teoría de brecha de género en los modelos de estimación de ingresos, se consideraba que si el jefe de hogar era mujer estos hogares tenderían a ser más pobres. Sin embargo, se encontró

[1] Define si el pobre o no Pobre
[2] Ingreso total antes de imputación a arriendo y propietarios

que la participación de hogares pobres es similar si el jefe es mujer o es hombre (Figura A - Ilustración 1).

Tabla 1. Proporción de hogares pobres y no pobres según el género del jefe de hogar

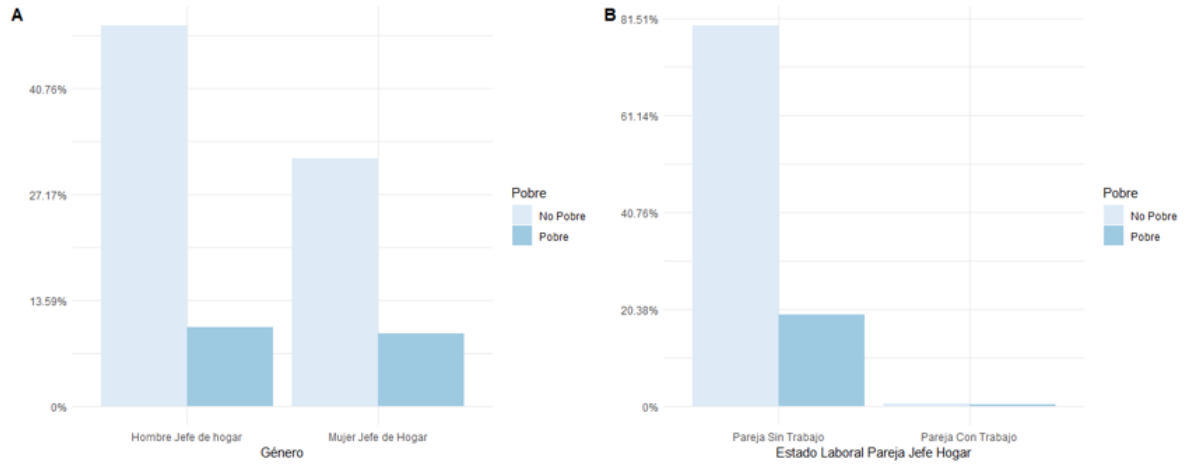
Characteristic	**N**	**Overall**, N = 147,215[1]	**No Pobre**, N = 118,599[1]	**Pobre**, N = 28,616[1]
mujer_jf_h	147,215			
Hombre Jefe de hogar		86,735 (59%)	71,796 (61%)	14,939 (52%)
Mujer Jefe de Hogar		60,480 (41%)	46,803 (39%)	13,677 (48%)

[1] n (%)

Fuente: Elaboración propia.

Caso contrario ocurre cuando se hace el análisis por el estado laboral de la pareja del jefe de hogar, donde se encuentra que en su mayoría los hogares pobres las parejas del jefe no tienen trabajo (Figura B - Ilustración 1). Sin embargo, está condición también aplica para los hogares no pobres, lo que indica que en la mayoría de los hogares colombianos la pareja del jefe de hogar no reporta tener trabajo.

Ilustración 1. Participación Hogares pobres según género del jefe de hogar y estado laboral de la pareja del jefe del hogar



Fuente: Elaboración propia.

3. Modelos y resultados:

3.1. Modelos de clasificación:

En primer lugar, se realizó un análisis de importancia de variables, cuyo resultado se presenta en el anexo X. Con base en esto, se seleccionaron algunas variables de interés y se hizo la clasificación con base en diferentes modelos logit que combinaban estas variables. Se analizó el desempeño de estos modelos comparando su sensibilidad, ROC y demás métricas de clasificación. El mejor de éstos, Logit 4, se tomó como base para construir otras variaciones y emplear otras metodologías de clasificación, incluyendo: validación cruzada K-fold; identificación y selección del cut-off óptimo; remuestreo upsample y downsample para corregir el desbalance (80/20) de no pobres / pobres; modelos lasso optimizados para sensibilidad, ROC y precisión; árboles; Boosting (XGBoost), y la combinación de algunas de estas técnicas. La tabla 2 presenta los mejores modelos de acuerdo con las métricas de sensibilidad, especificidad y precisión de los modelos.

Tabla 2. Mejores modelos según criterios de desempeño

Criterio	Mejor modelo	2° mejor modelo
Sensibilidad	Logit 4	Logit 4 CV Cut-off óptimo
Especificidad	XGBoost downsampled	Tree downsampled
Precisión	Tree	XGBoost

Fuente: Elaboración propia.

En nuestro problema, el principal criterio era la sensibilidad del modelo, es decir, la capacidad de predecir correctamente los hogares pobres por la importancia que esto tiene para cualquier política pública. Por lo tanto, elegimos como mejor modelo el Logit 4:

$$Pobre = Npersug : P5010 + P5090 + P5000 + edad_{p1} + mujer_{jf_h} + educ\_p1 + educ\_p3 + ht\_p1 + jf\_sub + hijos + ht\_p1 + pj\_jf\_sintrabajo + jf\_sintrabajo$$

Las variables seleccionadas con base en el análisis de importancia y con justificación económica son: el número de personas por unidad de gasto interactuando con la cantidad de cuartos para dormir (como proxy de hacinamiento); la propiedad de la vivienda; el número de cuartos de la vivienda; la edad del jefe de hogar; si la mujer es jefe de hogar; el nivel educativo del jefe de hogar y sus hijos; las horas trabajadas por el jefe de hogar; si el jefe de hogar pertenece al régimen subsidiado de salud; si el jefe de hogar y su pareja están sin trabajo.

Cabe mencionar que el mejor resultado de verdaderos positivos y verdaderos negativos lo obtuvimos con un punto de corte de 0,18, es decir, aquellos hogares con probabilidad Logit mayor a este valor se clasificaron como pobres. Dicho valor se encontró con la ayuda de un gráfico de cajas y bigotes (ver anexos), y se corroboró luego mediante optimización de ROC.

La tabla 3 muestra la comparación de los criterios de desempeño de los seis mejores modelos mencionados anteriormente. En el anexo se presentan los valores de 14 de los modelos probados.

Tabla 3. Comparación modelos

Modelo	TN	FN	FP	TP	Sensitivity	Specificity	Accuracy
Logit 4	17858	1019	5861	4703	0,82192	0,75290	0,76631
Logit 4 CV Cut-off óptimo	18118	1087	5601	4635	0,81003	0,76386	0,77283
XGBoost	22909	808	3889	1833	0,69406	0,85488	0,84045
Tree	22363	1352	3304	2418	0,64138	0,87127	0,84183
XGBoost downsampled	17794	5921	992	4730	0,44409	0,94719	0,76516
Tree downsampled	17268	6447	1057	4665	0,41982	0,94232	0,74508

Fuente: Elaboración propia.

3.2. Modelos de predicción de ingresos:

En este punto, se buscó predecir el ingreso y concluir, teniendo en cuenta la línea de pobreza, si una persona se encuentra o no en estado de pobreza. Para tal fin, se plantearon 18 modelos (lasso, ridge, elastic net, regresión lineal) con distintas variables seleccionadas. Por efectos de espacio, solo se mostrarán los siguientes 6 modelos lineales con sus resultados más relevantes:

Tabla 4. Resultados modelos

MODELO	RMSE	R2	SENSITIVITY	ACCURACY
Lineal 1 (modelo1)	4,44E+18	0.399	0.18887	0.8162
Lineal 2 (modelo2)	4,77E+18	0.343	0.22754	0.8139
Lineal 3 (modelo3)	5,04E+18	0.298	0.15886	0.8088
Lineal 4 (otro1)	5,70E+18	0.202	0.028137	0.798
Lineal 5 (otro2)	5,16E+18	0.282	0.12338	0.8021
Lineal 6 (otro3)	4,82E+18	0.333	0.18857	0.8146

Fuente: Elaboración propia.

Como se observa, los modelos presentaron un nivel sensibilidad baja, a pesar de que desempeñaron una fracción de predicciones realizadas correctamente reflejados en la precisión. En este sentido, se escogió el modelo2. En los anexos se presentan las métricas de los demás modelos. A continuación, se presenta el modelo seleccionado:

$$Ingtotug = Dominio + Npersug: P5010 + P5090 + P5000 + edad\_p1 + edadjf\_cua + edad\_p2 + edad\_p3 + edad\_p4 + edad\_p5 + edad\_p6 + edad\_p7 + edad\_p8 + edad\_p9 + mujer\_jf\_h + educ\_p1 + educ\_p3 + ht\_p1 + ht\_p2 + ht\_p3 + ht\_p4 + ht\_p5 + ht\_p6 + ht\_p7 + ht\_p8 + ht\_p9 + jf\_sub + hijos + ht\_p1 + pj\_jf\_sintrabajo + jf\_sintrabajo$$

Estas variables se tomaron teniendo en cuenta la literatura económica que puede explicar el ingreso de una persona, siendo las más relevantes la ciudad donde se reside, teniendo en cuenta las posibilidades y acceso a servicios que ofrecen las ciudades más grandes (Álvarez, 2007), el nivel de educación, especialmente del jefe del hogar que le permite tener mejores ingresos (Muñoz, 2004); la edad del grupo familiar, pues refleja la experiencia del trabajador que incide en el ingreso (Muñoz, 2004); el género, pues hay brechas de ingresos entre hombres y mujeres, dado el diferencial ocupacional, especialmente por limitaciones de tiempo (CEPAL, 2007); el total de horas trabajadas a la semana, pues, en principio, por cada hora de trabajo adicional realizada, es ingreso de la persona aumenta. De esta manera, el modelo predijo de manera efectiva los hogares no pobres (22660); sin embargo, predijo que 4420 hogares no eran pobres cuando en realidad lo eran. Esto excluiría a personas que son pobres dentro de una iniciativa de política pública dirigida hacia personas pobres.

4. Conclusiones y recomendaciones:

Con este trabajo se explora diferentes metodologías y algoritmos de predicción y de clasificación, buscando solucionar un problema real, la clasificación predictiva de pobreza en hogares. La indisponibilidad de variables representa un desafío importante a la hora de diseñar modelos de clasificación y predicción, pero es un reto que corresponde a las restricciones en la práctica; es decir, es una limitación a la que nos tenemos que adaptar como investigadores a la hora de adaptar estos. En nuestro caso, los modelos de clasificación para una variable Y discreta (Pobre / No pobre) tuvieron un mejor desempeño que los de predicción de ingresos. Sin embargo, esto no se puede generalizar, y depende mucho de la información de la que se disponga. Es importante realizar diferentes pruebas con distintos modelos y usando la amplia gama de ajustes disponibles, y basar la selección siempre en criterios objetivos de desempeño en la clasificación, ya que el hecho de emplear modelos más sofisticados o con ajustes o algoritmos más complejos no necesariamente arroja mejores resultados, como se pudo observar en este trabajo. Es importante también tener en cuenta que en la medida en que se empleen más variables y las bases de datos incluyan más observaciones, nos encontramos en la práctica con limitaciones de cómputo que equipos cotidianos no logran suplir. Por último, es interesante continuar explorando métodos adicionales, como los *superlearners*, para hacer una combinación sistemática de diferentes modelos y algoritmos, buscando un mejor desempeño en la predicción.

Anexos:

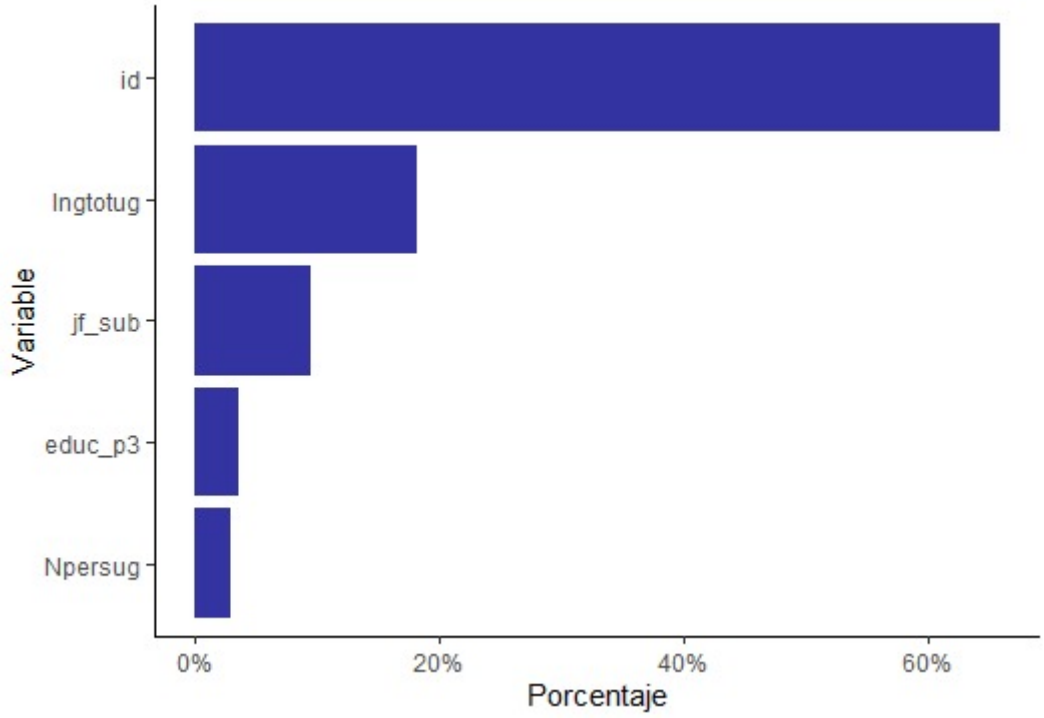
Anexos relacionados la selección de variables y estudio de la base de datos:

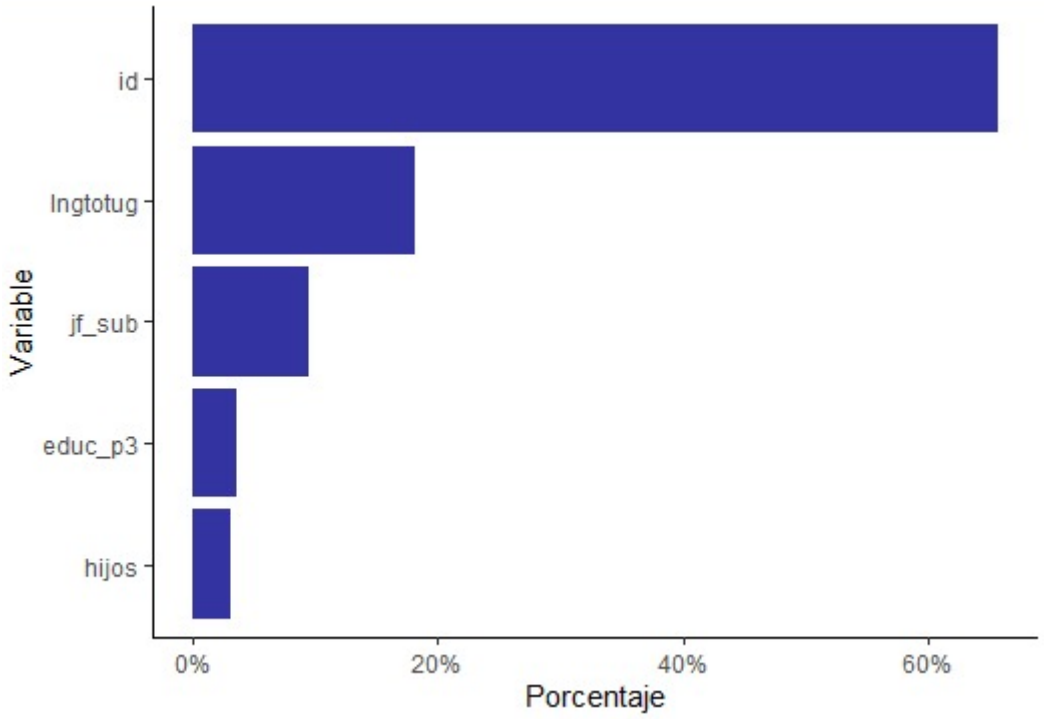
Tabla 1. Variables y etapas de selección

Variable	Descripción	Primer Filtro	Creación variable	Segundo Filtro
Clase	Cabecera	x		x
Dominio	Dominio	x		x
Pobre	Define si el hogar tiene ingresos menores a la Línea de Pobreza	x		x
Ingtotug	Ingreso total antes de imputación a arriendo y propietarios	x		x
P5000	Cuartos totales (incluye sala-comedor)	x		x
P5010	Cuartos para dormir	x		x
P5090	Propiedad de la vivienda	x		x
P5100	Cuota amortización	x		
P5130	Estimación arriendo	x		
P5140	Valor arriendo	x		
Nper	Número de personas en el hogar	x		x
Npersug	Número de personas en la unidad de gasto	x		x
Li	Línea de indigencia	x		x
Lp	Línea de pobreza	x		x
Fex_c	Factor de expansión	x		x
Depto	Departamento	x		x
Fex_depto	Factor de expansión departamental	x		x
edad_pi	Edad por tipo de persona que conforma el hogar		x	x
ht_pi	Horas trabajadas por tipo de persona que conforma el hogar		x	x
of_pi	Oficio de las personas que conforman el hogar		x	x
educ_pi	Nivel educativo de las personas que conforman el hogar		x	x
mujer_jf_h	mujer jefe de hogar		x	x
jf_10_18_h	jefe de hogar entre los 10 y 18 años		x	x
jf_19_28_h	jefe de hogar entre los 19 y 28 años		x	x
jf_29_59_h	jefe de hogar entre los 29 y 59 años		x	x
jf_60_h	jf con 60 años o más		x	x
jf_sub	jefe de hogar en régimen de salud subsidiado		x	x
jf_afiliado	jefe de hogar no afiliado a entidad de seguridad social en salud		x	x
jf_sintrabajo	jefe de hogar sin trabajo		x	x
pj_jf_ofhogar	pareja del jefe de hogar dedicada a oficio del		x	x
pj_jf_sintrabajo	pareja del jefe hogar si trabajo		x	x
jf_nc_pension	jefe de hogar no cotiza pensión		x	
jf_P7422	jefe de hogar recibió ingresos el mes pasado por concepto de trabajo (desocupados)		x	
pj_jf_P7422	Pareja del hogar recibió ingresos el mes pasado por concepto de trabajo (desocupados)		x	
jf_P7472	jefe de hogar recibió ingresos el mes pasado por concepto de trabajo		x	
pj_jf_P7472	Pareja del hogar recibió ingresos el mes pasado por concepto de trabajo		x	
hijos	Número de hijos		x	x

Fuente: Elaboración propia.

Ilustraciones 1 y 2. Análisis preliminares de importancia de variables (VarImportance) basado en un árbol de predicción.





Fuente: Elaboración propia.

Tabla 2. Proporción de hogares pobres y no pobres por ciudad

Characteristic	**N**	**Overall**, N = 147,215[1]	**No Pobre**, N = 118,599[1]	**Pobre**, N = 28,616[1]
Dominio	147,215			
ARMENIA		4,748 (3.2%)	4,066 (3.4%)	682 (2.4%)
BARRANQUILLA		6,120 (4.2%)	5,214 (4.4%)	906 (3.2%)
BOGOTA		9,249 (6.3%)	8,464 (7.1%)	785 (2.7%)
BUCARAMANGA		4,602 (3.1%)	4,098 (3.5%)	504 (1.8%)
CALI		5,845 (4.0%)	5,233 (4.4%)	612 (2.1%)
CARTAGENA		5,008 (3.4%)	4,034 (3.4%)	974 (3.4%)
CUCUTA		4,032 (2.7%)	2,903 (2.4%)	1,129 (3.9%)
FLORENCIA		4,247 (2.9%)	3,136 (2.6%)	1,111 (3.9%)
IBAGUE		4,643 (3.2%)	4,067 (3.4%)	576 (2.0%)
MANIZALES		5,411 (3.7%)	4,989 (4.2%)	422 (1.5%)
MEDELLIN		7,981 (5.4%)	7,197 (6.1%)	784 (2.7%)
MONTERIA		4,625 (3.1%)	3,671 (3.1%)	954 (3.3%)
NEIVA		5,047 (3.4%)	4,204 (3.5%)	843 (2.9%)
PASTO		4,441 (3.0%)	3,550 (3.0%)	891 (3.1%)
PEREIRA		4,954 (3.4%)	4,496 (3.8%)	458 (1.6%)
POPAYAN		5,132 (3.5%)	3,952 (3.3%)	1,180 (4.1%)
QUIBDO		3,723 (2.5%)	2,533 (2.1%)	1,190 (4.2%)
RESTO URBANO		15,210 (10%)	10,789 (9.1%)	4,421 (15%)
RIOHACHA		4,770 (3.2%)	3,323 (2.8%)	1,447 (5.1%)
RURAL		13,716 (9.3%)	9,551 (8.1%)	4,165 (15%)
SANTA MARTA		5,308 (3.6%)	4,099 (3.5%)	1,209 (4.2%)
SINCELEJO		5,092 (3.5%)	3,985 (3.4%)	1,107 (3.9%)
TUNJA		4,027 (2.7%)	3,514 (3.0%)	513 (1.8%)
VALLEDUPAR		4,567 (3.1%)	3,398 (2.9%)	1,169 (4.1%)
VILLAVICENCIO		4,717 (3.2%)	4,133 (3.5%)	584 (2.0%)

[1] n (%)

Fuente: Elaboración propia.

Tabla 35. Proporción de hogares pobres y no pobres según el grupo etario del jefe de hogar

Characteristic	N	**Overall**, N = 147,215[1]	**No Pobre**, N = 118,599[1]	**Pobre**, N = 28,616[1]
Jefe del hogar entre los 10 y 18 años	147,215			
0		146,861 (100%)	118,382 (100%)	28,479 (100%)
1		354 (0.2%)	217 (0.2%)	137 (0.5%)
Jefe del hogar entre los 19 y 28 años	147,215			
0		133,463 (91%)	107,947 (91%)	25,516 (89%)
1		13,752 (9.3%)	10,652 (9.0%)	3,100 (11%)
Jefe del hogar entre los 29 y 59 años	147,215			
0		54,331 (37%)	44,779 (38%)	9,552 (33%)
1		92,884 (63%)	73,820 (62%)	19,064 (67%)
Jefe con 60 años o más	147,215			
0		106,990 (73%)	84,689 (71%)	22,301 (78%)
1		40,225 (27%)	33,910 (29%)	6,315 (22%)

[1] n (%

Fuente: Elaboración propia.

Tabla 4. Proporción de hogares pobres y no pobres según características de la vivienda

Characteristic	N	**Overall**, N = 147,215[1]	**No Pobre**, N = 118,599[1]	**Pobre**, N = 28,616[1]
Cuartos totales (incluye sala-comedor)	147,215			
1		8,126 (5.5%)	5,914 (5.0%)	2,212 (7.7%)
2		18,419 (13%)	13,252 (11%)	5,167 (18%)
3		50,386 (34%)	38,871 (33%)	11,515 (40%)
4		48,384 (33%)	40,971 (35%)	7,413 (26%)
5		15,661 (11%)	13,865 (12%)	1,796 (6.3%)
6		4,462 (3.0%)	4,098 (3.5%)	364 (1.3%)
7		1,226 (0.8%)	1,120 (0.9%)	106 (0.4%)
8		382 (0.3%)	354 (0.3%)	28 (<0.1%)
9		96 (<0.1%)	87 (<0.1%)	9 (<0.1%)

Characteristic	N	Overall, N = 147,215[1]	No Pobre, N = 118,599[1]	Pobre, N = 28,616[1]
10		41 (<0.1%)	38 (<0.1%)	3 (<0.1%)
11		16 (<0.1%)	15 (<0.1%)	1 (<0.1%)
12		8 (<0.1%)	8 (<0.1%)	0 (0%)
13		1 (<0.1%)	1 (<0.1%)	0 (0%)
14		3 (<0.1%)	2 (<0.1%)	1 (<0.1%)
15		1 (<0.1%)	1 (<0.1%)	0 (0%)
16		1 (<0.1%)	1 (<0.1%)	0 (0%)
18		1 (<0.1%)	0 (0%)	1 (<0.1%)
98		1 (<0.1%)	1 (<0.1%)	0 (0%)
Cuartos para dormir	147,215			
1		42,847 (29%)	35,322 (30%)	7,525 (26%)
2		62,349 (42%)	48,627 (41%)	13,722 (48%)
3		34,176 (23%)	28,060 (24%)	6,116 (21%)
4		6,521 (4.4%)	5,488 (4.6%)	1,033 (3.6%)
5		1,037 (0.7%)	860 (0.7%)	177 (0.6%)
6		217 (0.1%)	184 (0.2%)	33 (0.1%)
7		50 (<0.1%)	43 (<0.1%)	7 (<0.1%)
8		15 (<0.1%)	12 (<0.1%)	3 (<0.1%)
9		1 (<0.1%)	1 (<0.1%)	0 (0%)
10		1 (<0.1%)	1 (<0.1%)	0 (0%)
15		1 (<0.1%)	1 (<0.1%)	0 (0%)
Propiedad de la vivienda	147,215			
Propia, totalmente pagada		56,957 (39%)	48,590 (41%)	8,367 (29%)
Propia, la están pagando		5,337 (3.6%)	4,852 (4.1%)	485 (1.7%)
En arriendo o subarriendo		55,886 (38%)	44,055 (37%)	11,831 (41%)
En usufructo		21,856 (15%)	17,346 (15%)	4,510 (16%)

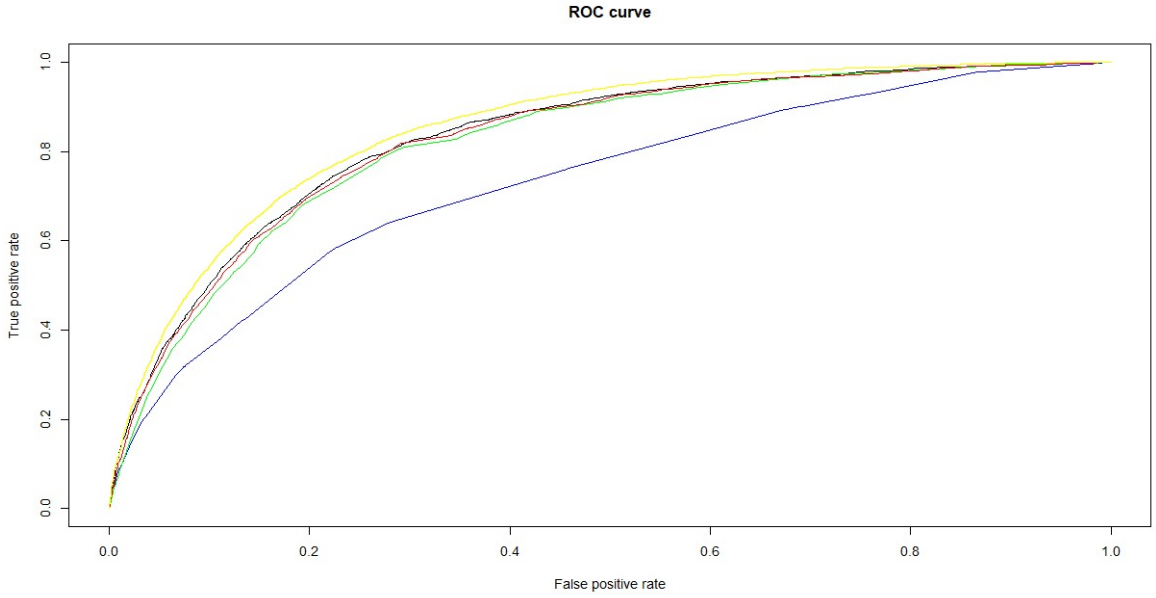
Characteristic	N	Overall, N = 147,215[1]	No Pobre, N = 118,599[1]	Pobre, N = 28,616[1]
Posesión sin título (ocupante)		7,049 (4.8%)	3,656 (3.1%)	3,393 (12%)
Otra		130 (<0.1%)	100 (<0.1%)	30 (0.1%)

[1] n (%)

Fuente: Elaboración propia.

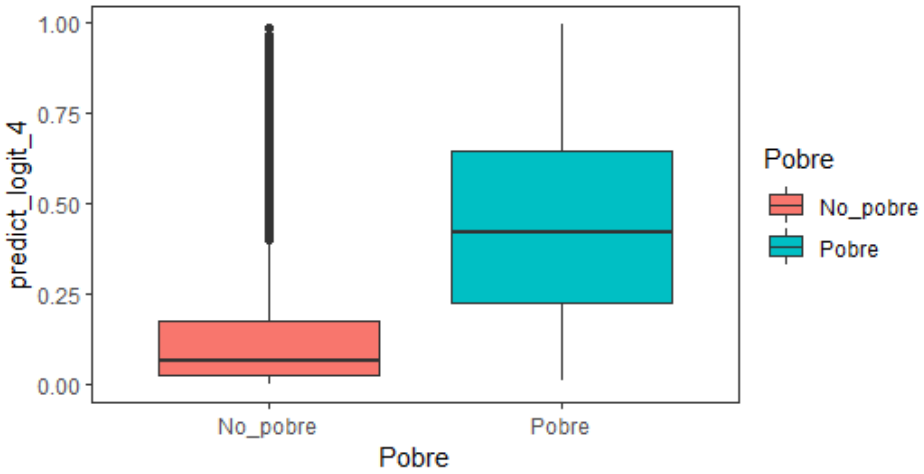
Anexos relacionados con los modelos de clasificación:

Ilustración 3. Curvas ROC para los 4 modelos logit probados. La curva amarilla corresponde a Logit 4.



Fuente: Elaboración propia.

Ilustración 4. Diagrama de caja Modelo Logit4



Fuente: Elaboración propia.

Tabla 5. Métricas de desempeño de todos los modelos de clasificación probados.

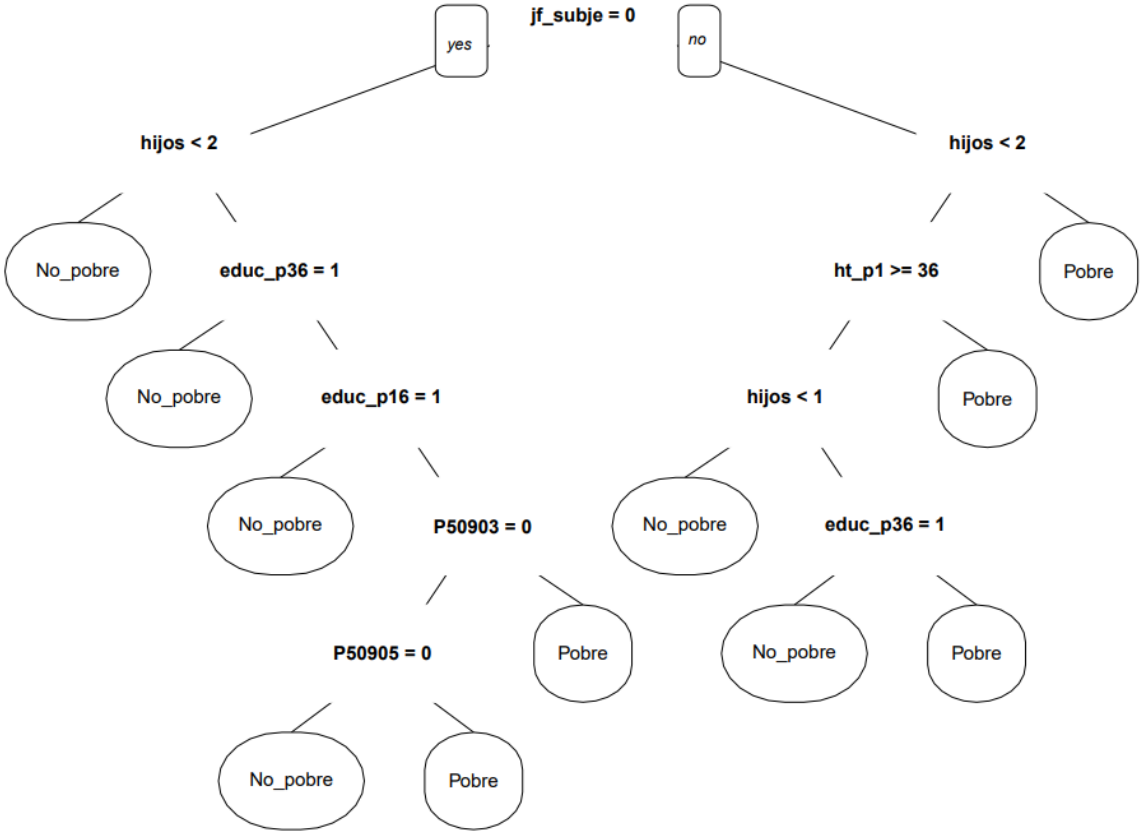
Modelo	TN	FN	FP	TP	Sensitivity	Specificity	Accuracy
XGBoost downsampled	17794	5921	992	4730	0,4440898	0,9471947	0,7651595
Tree downsampled	17268	6447	1057	4665	0,4198164	0,9423192	0,7450827
Tree upsampled	18818	4897	2008	3714	0,4313088	0,9035821	0,7654313
Tree	22363	1352	3304	2418	0,6413793	0,8712744	0,8418317
XGBoost	22909	808	3889	1833	0,6940553	0,8548772	0,8404497
Logit 2	18878	1336	4841	4386	0,7665152	0,7959020	0,7901906
Logit 4 CV upsampled	18354	1159	5361	4563	0,7974484	0,7739405	0,7785100
Logit 4 CV downsampled	18325	1146	5390	4576	0,7997204	0,7727177	0,7779665
Logit 3	18325	1183	5394	4539	0,7932541	0,7725874	0,7766041
Logit 5	18308	1189	5411	4533	0,7922055	0,7718707	0,7758228
Lasso Sens upsample	18150	1151	5565	4571	0,7988466	0,7653384	0,7718518
Logit 1	18121	1531	5598	4191	0,7324362	0,7639867	0,7578547
Logit 4 CV Cut-off óptimo	18118	1087	5601	4635	0,8100315	0,7638602	0,7728338
Logit 4	17858	1019	5861	4703	0,8219154	0,7528985	0,7663123

Fuente: Elaboración propia.

Tabla 6. Comparación métricas de desempeño modelos basados en predicción de ingresos.

Modelo	TN	FN	FP	TP	Sensitivity	Specificity	Accuracy
Linear model 1	22951	4642	768	1080	0,18875	0,96762	0,8162
Linear model 2	22660	4420	1059	1302	0,22754	0,95535	0,8139
Linear model 3	22902	4813	817	909	0,15886	0,96556	0,8088
Lasso 1	22971	4667	748	1055	0,18438	0,96846	0,8161
Lasso 2	22677	4435	1042	1287	0,22492	0,95607	0,814
Lasso 3	22906	4825	813	897	0,15676	0,96572	0,8085
Ridge 1	23065	4755	654	967	0,169	0,97243	0,8163
Ridge 2	22825	4564	894	1158	0,20238	0,96231	0,8146
Ridge 3	23026	4949	693	773	0,13509	0,97078	0,8084
Elastic Net 1	22967	4660	752	1062	0,1856	0,9683	0,8162
Elastic Net 2	22674	4433	1045	1289	0,22527	0,95594	0,8139
Elastic Net 3	22909	4830	810	892	0,15589	0,96585	0,8084

Ilustración. Árbol de predicción con Downsample



Fuente: Elaboración propia.