Problem Set 2: Predicting Poverty

By González Galvis, Daniel Enrique; González Junca, Daniela Natalia; Mendoza Potes, Valentina and Rodríguez Pacheco, Alfredo José

I. Introducción

La pobreza sigue siendo un desafío importante en Colombia, que afecta a millones de personas y obstaculiza el desarrollo general de la nación. Predecir la pobreza ofrece una poderosa herramienta en la lucha por reducir las brechas de equidad. Al identificar los factores y los grupos demográficos más susceptibles a caer en la pobreza, se pueden asignar recursos y desarrollar programas de política pública estratégicos y más costo-efectivos para lograr que cada vez menos personas perciban ingresos bajo la línea de pobreza.

Según datos del PNUD en Colombia durante 2023 "la pobreza monetaria pasó de 39,7 % a 36,6 %, lo que significa que 1,3 millones de personas salieron de la pobreza a nivel nacional. En contraste, los resultados de pobreza extrema muestran un leve incremento, al pasar de 13,7 % a 13,8 % a nivel nacional."(PNUD, 2023) Lo anterior debido al incremento de la pobreza extrema en las áreas rurales del país.

Igualmente, algunos de los determinantes de pobreza en el país que explican esta tendencia han sido el crecimiento de la economía que incentivó la disminución de la tasa de desempleo y el crecimiento de la mano de obra en áreas cercanas a la formalidad. (PNUD, 2023) Sin embargo, existen aún múltiples oportunidades en términos de política pública, para garantizar el acceso de los hogares a factores que incrementen la probabilidad de obtener mejores ingresos, tales como acceso a educación, control en el tamaño del hogar, el sexo, la productividad del campo, entre otros.

Los datos empleados para llevar a cabo estas predicciones provienen del DANE y corresponden a la serie ."Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE". Esta contiene datos a nivel hogar y a nivel individuo que permiten hacer una caracterización de los mismos, sus niveles y fuentes de ingreso así como su status de pobreza, por lo cual permiten realizar tanto los modelos de regresión para predecir el ingreso de los hogares a partir de distintos factores, como los modelos de clasificación directa. Igualmente, los datos se recibieron divididos para realizar entrenamiento y test.

II. Datos

Frente a los datos, encontramos en primer lugar que la base contiene observaciones del 2018, lo que supone una limitación al no tener en la base data macroeconómica que podría impactar las predicciones si se quisieran tener en cuenta para intervenciones actuales. Por otra parte, encontramos que las bases a nivel hogar e individuo permiten caracterizar cada uno de estos actores. Para el caso de la limpieza de bases utilizadas en los modelos, se tuvieron en cuenta los factores de ingreso, sexo, regímen de seguridad social, nivel educativo, horas trabajadas por semana y cotización a fondo de pensión (asociado a la formalidad). Se obtuvo un total de 66168 observaciones para la base de test y 164960 observaciones para la base de entrenamiento.

A. Estadísticas descriptivas

La variable principal de interés es "Pobre", variable binaria que identifica si un hogar es pobre (toma el valor de 1) o no es pobre (toma el valor de 0). En la base de datos, existen 33024 hogares pobres y 131936 hogares no pobres. Esto implica que el 20 % de las observaciones son hogares pobres. Aunque los datos están desbalanceados en este sentido, el desbalance no es tan crítico por lo que es posible hacer los análisis.

	Pobres
0	131936
1	33024

Cuadro 1—Número de pobres

Cuadro 2—Estadisticas Descriptivas Variables numéricas

Statistic	N	Mean	St. Dev.	Min	Max
jefe_edad	66,168	49.691	16.369	11	101
nmujeres	66,168	1.750	1.185	0	11
nmenores	66,168	0.655	0.930	0	10
nocupados	66,168	1.510	1.034	0	12
rel_pet	66,168	0.865	0.187	0.000	1.000
horas_trab_por_persona	66,168	22.232	16.084	0.000	126.000

En este sentido seleccionamos las variables numéricas que consideramos de mayor interés (Cuadro 2) para los cuales tenemos dichas características en nuestra base de testeo y de entrenamiento. La edad promedio del jefe del hogar es de 50 años, el número de mujeres promedio es de 1,8, el número de menores es de 0,7, el número de ocupados es de 1,5 personas así como la relación promedio de la población en edad de trabajar PET y la no PET por hogar es de 0,86 y por último la cantidad de horas trabajadas por persona es de 22 horas por semana. Es importante recalcar que estas características son a nivel de hogar y solo muestra una primera aproximación a los datos y las medias impiden observar la distribución completa de variables que pueden tener valores extremos e invalidare en los promedios.

	Dominio	arrienda	jefe_mujer	jefe_Educ_level	jefe_ocupado	jefe_solo	j
	RESTO URBANO: 7371	0:40967	0:38643	NS : 9	0:19145	0:27024	0
	RURAL : 6682	1:25201	1:27525	Ninguno : 3646	1:47023	1:39144	1
	MEDELLIN : 3862			Preescolar : 9			
	BARRANQUILLA: 2835			Primaria :19018			
	CALI : 2755			Secundaria : 8696			
	SANTA MARTA : 2601			Media :16997			
	(Other) :40062			Universitaria:17793			

Por lo mencionado anteriormente analizamos las distribuciones de algunas variables que consideramos pueden ser altamente relevantes a la hora de predecir si un hogar es pobre o no.

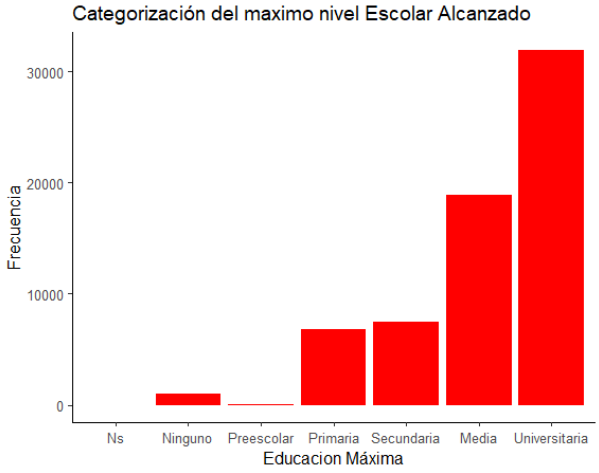

FIGURA 1.

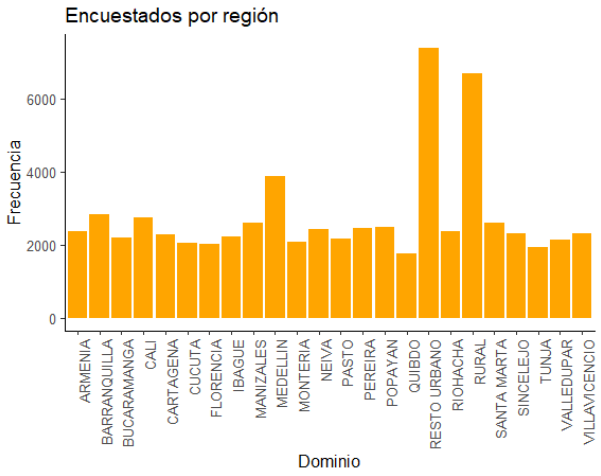

FIGURA 2.

Mientras que en la figura 1 nos muestra que la máxima educación de los individuos se concentra principalmente en la universitaria. Se puede observar que también tiene una tendencia creciente, es decir que a medida que aumenta el grado de escolaridad, aumenta el número de individuos en esa categoría. Por otro lado, en la figura número 2 podemos ratificar la heroicidad y amplia cobertura de las encuestas del DANE lo que nos permite descartar posibles sesgos de selección regionales que nos lleven a predicciones erróneas sobre el ingreso y/o pobreza de los hogares.

B. Otros análisis: análisis de datos exploratorio

Al visualizar la Figura 3 en la distribución del ingreso de los individuos en la base de datos de entrenamiento, encontramos que hay una concentración del ingreso por debajo de 2.5 millones de pesos al mes, habiendo sin embargo valores máximos por encima de los 10 millones, lo que también sugiere algún nivel de inequidad en la población.

En la Figura **??**, encontramos que la distribución de edad del jefe de hogar vs. el status de pobreza, parece sugerir que los hogares con status de pobreza "Sí", tienden a ser más jóvenes, aunque no sea posible establecer esto de manera causal.

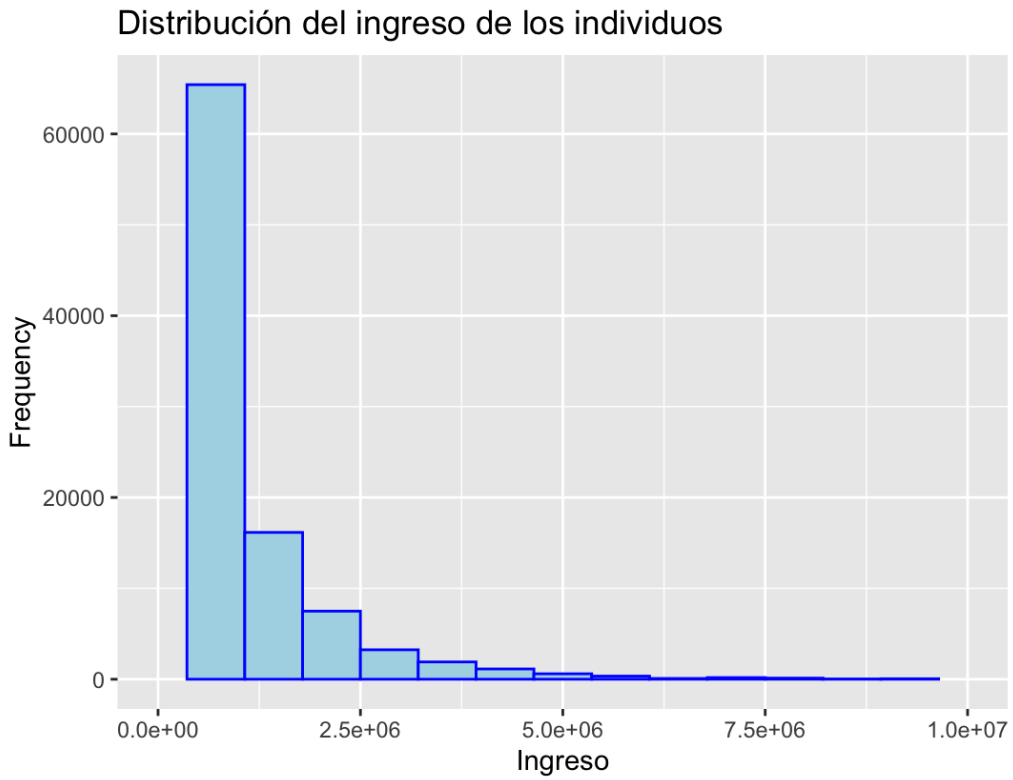

Figura 3. Distribución de los ingresos de los individuos

III. Modelo y resultados

A. Modelos de Clasificación

En primer lugar, utilizamos un modelo Logit. La ventaja de este modelo es que, a la vez de funcionar en aprendizaje supervizado para clasificar, también permite interpretar el peso de las variables en la probabilidad de pertenecer a una clase. Además, este tipo de modelos funciona especialmente bien para predecir variables binarias, como es el caso de la pobreza en los datos suministrados.

Como se ve en la tabla 3, dado el modelo logit, muchas de las variables predictoras son significativas. Sin embargo, cabe destacar que las variables que tienen un efecto mayor sobre la probabilidad de ser pobre son la educacion del jefe de hogar, siendo de -2.15, -1.41 y -1.11 cuanto tiene educación universitaria, media y secundaria respectivamente.

Posteriormente, utilizamos un modelo de CARTs (Aprendizaje basado en árboles de decisión) basados en las características de las personas de cada hogar, así

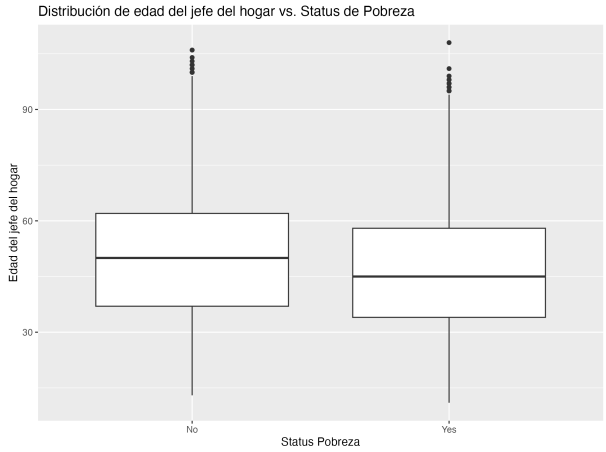

Figura 4.

como del los hogares en conjunto para clasificar a los hogares como pobres o no. Este modelo en particular ofrece diferentes ventajas significativas, ya que es de fácil interpretación al tener la importancia de las variables organizadas jerárquicamente más arriba en el árbol y es capaz de manejar variables categóricas y numéricas. Así mismo, los árboles se pueden trazar gráficamente siempre y cuando una baja complejidad lo permita. Esto proporciona una comprensión detallada de los factores que influyen en este caso en la pobreza y como a partir de la base de datos de Medición de Pobreza Monetaria y Desigualdad del DANE se puede llegar a esta predicción.

La metodología que utilizamos involucró un árbol único que fue ajustado mediante prunning para evitar el sobre ajuste. En este sentido, evaluamos si era necesario reducir la complejidad del árbol limitando el número mínimo de observaciones en cada nodo terminal para garantizar que no tengamos un árbol muy complejo con nodos finales con muy pocas observaciones que nos lleven a malas predicciones fuera de muestra. De esta forma utilizamos como mecanismo de regularización el método de k-fold-Cross Validation para determinar el mejor valor de alpha para penalizar la función de error. A partir del k-fold-Cross Validation (con k=5) encontramos que el cp que maximiza el ROC es de 0 para el que se tiene un ROC de 0,85. De esta forma, el modelo resultante es extremadamente complejo y cuya interpretación gráfica poco productiva. A partir del entrenamiento del modelo, la predicción resultante evaluada en la plataforma Kaggle, usando el F1 score que mide la precisión y el recall de la predicción, sugiere que existe una tasa de éxito baja del 0,31.

En segundo lugar, utilizamos la metodología Random Forest estimando una ma-

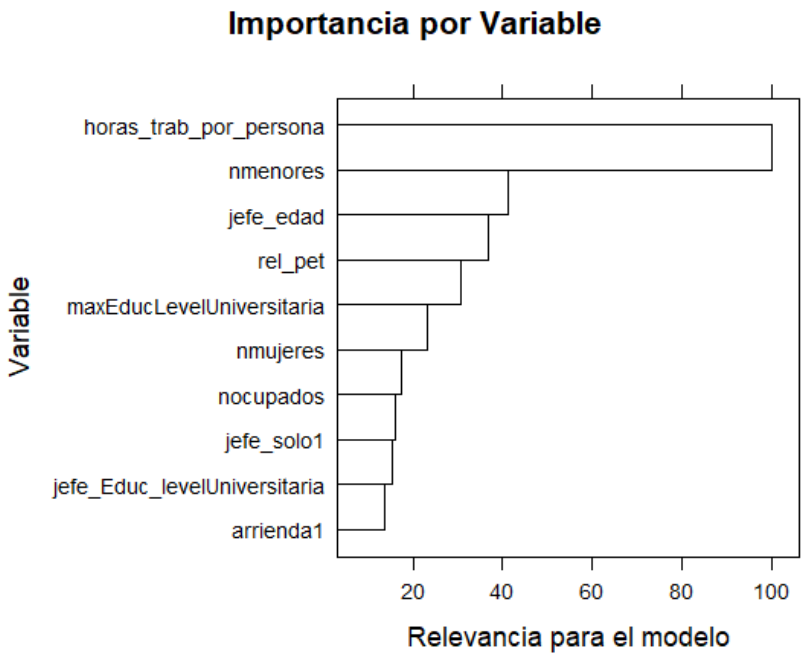

FIGURA 5.

yor cantidad de árboles con el propósito de evaluar las variables que efectivamente son determinantes en la clasificación de pobre/no pobre para los hogares como se muestra a e la figura 1. Como se puede observar la variable de horas trabajadas por miembro del hogar fue la que más importancia tuvo, probablemente porque representa directamente una cantidad de ingreso que se distribuye sobre el total de miembros del hogar. De igual forma otras variables relevantes fueron el número de menores en el hogar, la edad del jefe, la relación de miembros del hogar pertenecientes a la población en edad de trabajar, si el nivel maximo de educación fue universitaria, si el jefe trabaja solo, si el nivel de educación del jefe es universitaria y por último si la vivienda es arrendada.Para esta metodología elegimos mediante prueba y error los hiperparámetros mtry=8 y min.node.size= 50 de tal forma que obtenemos 8 variables predictoras y un mínimo número de nodos terminales del árbol de 50. El "splitrule" utilizado fue el gini, para categorizar la efectividad de predicción de cada árbol. El resultado en Kaggle para esta predicción fue de 0,55 representando una mejora significativa frente al resultado del modelo CARTs.

También se utilizó la metodología Elastic Net, la cual ajusta modelos de regresión lineal con penalizaciones lasso y ridge. Haciendo uso de cross validation con k = 5 se determinaron los parámetros alpha = 0.4 y lambda = 0.01, obteniendo

una precisión de 0.846 en los datos de entrenamiento. El resultado en Kaggle de este modelo fue de 0.49, un desempeño promedio al ser comparado con los demás modelos evaluados. De igual manera, se utilizó Elastic Net para ajustar un modelo Logit. Una vez más, haciendo uso de cross-validation con k = 5 se determinaron los parámetros alpha = 0.1 y lambda = 0.01, los cuales arrojaron una precisión en los datos de entrenamiento de 0.845. En este modelo al igual que en el anterior, se obtuvo una puntuación en Kaggle de 0.49. Ambos modelos resultan igual de competitivos para la determinación de la pobreza en Colombia, sin embargo, su desempeño al ser comparados con otros modelos de clasificación no es el mejor.

B. Modelos de Regresión

Una vez predicha la probabilidad de ser pobre de un individuo basado en sus características, procedimos a predecir el ingreso de cada hogar para evaluar si su nivel de ingreso era inferior a la linea de pobreza establecida para ese hogar. Para esto utilizamos la definición de DANE en la que "un hogar es clasificado pobre si el Ingreso percápita de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios es menor a la línea de pobreza que le corresponde al hogar.", por ende la predicción de ingreso por hogar se realizó sobre la variable "Ingreso percápita de la unidad de gasto con imputación de arriendo a propietarios y usufructuarios".

En primer lugar, utilizamos nuevamente el modelo CARTs en el que a partir de un k=5-fold-cross validation model definimos un cp de penalización que maximiza el RMSE. Sin embargo, obtuvimos una discrepancia muy elevada entre el error predicho y la linea de pobreza de cada hogar lo que llevó a que la predicción del modelo predijera que ningún hogar estaba por debajo de la línea de la pobreza. Por lo tanto, agregamos la variable de estimación de costo de arriendo de la vivienda con el propósito de mejorar el ajuste de la predicción del modelo. Una vez agregada esta variable, y realizando el procedimiento anterior, obtuvimos un resultado de predicción de Kaggle de 0,51, lo que representó una mejora determinante.

En segundo lugar, utilizamos la metodología Random Forest para estimar múltiples árboles y obtener una regularización del modelo con las mejores variables predictoras, como se muestra a en la figura 2. Para este caso, la nueva variable de costo del arriendo resulto ser determinante para predecir el ingreso de los hogares, posiblemente porque hay una relación muy estrecha entre el costo del arriendo o la valoración del arriendo de la vivienda habitada y el verdadero nivel de ingresos del hogar. De igual forma, el resto de variables se asemejan a las descritas para el Random Forest de clasificación.

Para esto, mediante prueba y error definimos los hiperparámetros mtry= 15 y min.node.size= 35 de tal forma que obtenemos 15 variables predictoras y un

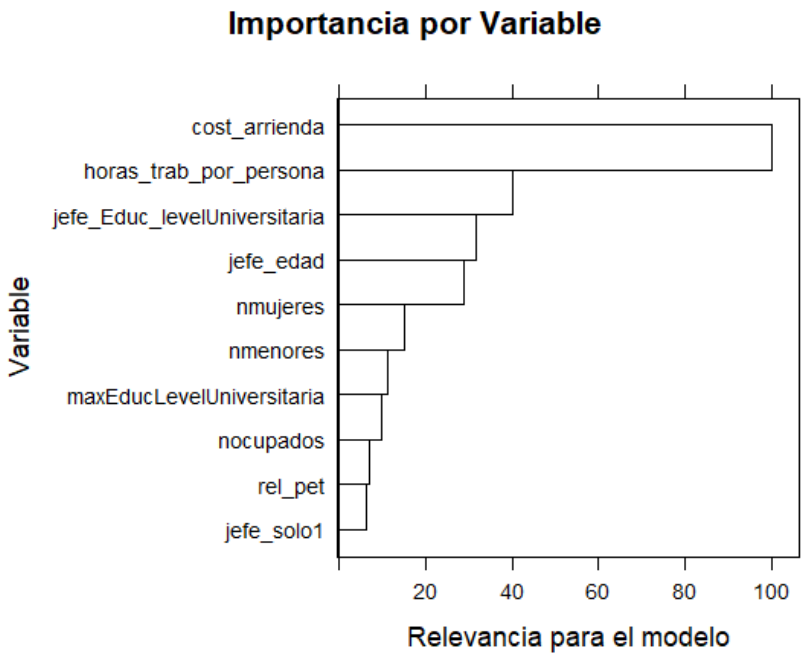

Figura 6.

mínimo número de nodos terminales del arbol de 35. El "splitrule" utilizado fue el variance, para categorizar la efectividad de predicción de cada árbol con base en la homogeneidad de los sub nodos. El resultado de predicción de Kaggle fue de 0,53 lo que representó una leve mejora frente al modelo estimado por CARTs.

C. Modelos Finales

En términos generales podemos afirmar que los modelos no divergen mucho en el éxito de predicción de pobreza en el problema de clasificación. El único modelo que cuanta con un muy bajo nivel predictivo es el de CARTs, posiblemente por la no regularización e inclusión de todas las variables. Por otro lado, el modelo que cuenta con una mayor capacidad predictiva el estimado mediante random forest, al que le atribuimos este éxito por la selección que tiene de las variables más relevantes y la penalización que tiene por sobreajuste. Por su parte los modelos de elastic net y logit pueden tener la desventaja del ajuste en mayor medida lineal y por ende predecir con una menor tasa de éxito esta muestra en particular.

Finalmente, para la regresión, en la que el modelo primero debía predecir el ingreso y posteriormente el nivel de pobreza de los hogares obtenemos un resultado similar, el modelo de Random Forest es el que cuenta con una mejor tasa de éxito por las razones mencionadas anteriormente.

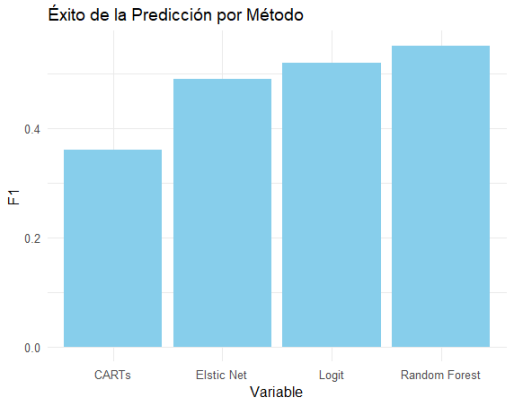

Figura 7. Tasa de éxito para cada modelo de clasificación

IV. Conclusiones

Para la predicción de la pobreza en Colombia se evaluaron una serie de modelos, tanto de clasificación como de regresión, con el fin de obtener aquel modelo con mejor desempeño fuera de muestra. Como se mencionó anteriormente, entre esas metodologías se encuentran Random Forest, Elastic Net, Logit y CARTs. Los 3 modelos con mejores resultados de predicción corresponden a modelos de clasificación utilizando la metodología Random Forest.

La metodología Random Forest proporciona modelos más robustos con una alta tasa de precisión, ya que se reduce el problema de sobreajuste. La metodologia Logit, la cual parece ser relativamente más sencilla, tambien presenta resultados competitivos, y en terminos de interpretación presenta ventajas para analizar los coeficientes de cada variable en el contexto de pobreza.

Si bien los modelos de clasificación tuvieron los desempeños más altos, los modelos de regresión con las metodologías tanto de CARTs como Random Forest obtuvieron capacidades predictivas de una magnitud similar a los modelos de clasificación. En conclusión, la metodología Random Forest fue la más efectiva para predecir la pobreza en Colombia, tanto en los modelos de clasificación como en los modelos de regresión. Dado que en este caso el objetivo era predecir, las ventajas de interpretación y explicabilidad, que pueden presentar otras metodologías como Logit o CARTs pueden resultar mas utiles al buscar una interpretación profunda entre la pobreza y las variables correlacionadas.

A. Anexos

Tabla de estimaciones por medio de logit para el problema de clasificación:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.5593	0.3052	1.83	0.0669
DominioBARRANQUILLA	-0.0545	0.0593	-0.92	0.3581
DominioBOGOTA	-0.3742	0.0568	-6.58	0.0000
DominioBUCARAMANGA	-0.3455	0.0653	-5.29	0.0000
DominioCALI	-0.3550	0.0610	-5.82	0.0000
DominioCARTAGENA	0.2371	0.0595	3.99	0.0001
DominioCUCUTA	0.5172	0.0581	8.90	0.0000
DominioFLORENCIA	0.4279	0.0580	7.38	0.0000
DominioIBAGUE	-0.1960	0.0631	-3.11	0.0019
DominioMANIZALES	-0.6710	0.0670	-10.01	0.0000
DominioMEDELLIN	-0.2545	0.0577	-4.41	0.0000
DominioMONTERIA	0.2154	0.0596	3.61	0.0003
DominioNEIVA	0.0939	0.0594	1.58	0.1139
DominioPASTO	0.4392	0.0600	7.32	0.0000
DominioPEREIRA	-0.7456	0.0682	-10.93	0.0000
DominioPOPAYAN	0.7113	0.0561	12.68	0.0000
DominioQUIBDO	1.0557	0.0604	17.48	0.0000
DominioRESTO URBANO	0.6700	0.0476	14.09	0.0000
DominioRIOHACHA	0.9964	0.0566	17.61	0.0000
DominioRURAL	0.2229	0.0493	4.52	0.0000
DominioSANTA MARTA	0.4807	0.0569	8.44	0.0000
DominioSINCELEJO	0.4370	0.0590	7.41	0.0000
DominioTUNJA	-0.0494	0.0653	-0.76	0.4498
DominioVALLEDUPAR	0.5607	0.0577	9.72	0.0000
DominioVILLAVICENCIO	-0.4355	0.0635	-6.86	0.0000
arrienda1	0.7024	0.0175	40.21	0.0000
jefe_mujer1	-0.0967	0.0166	-5.82	0.0000
jefe_Educ_levelPreescolar	0.4738	0.8160	0.58	0.5615
jefe_Educ_levelPrimaria	-0.6804	0.0342	-19.87	0.0000
jefe_Educ_levelSecundaria	-1.1142	0.0391	-28.47	0.0000
jefe_Educ_levelMedia	-1.4145	0.0393	-36.01	0.0000
jefe_Educ_levelUniversitaria	-2.1525	0.0456	-47.20	0.0000
jefe_ocupado1	0.8032	0.0230	34.98	0.0000
jefe_solo1	0.8798	0.0195	45.11	0.0000
jefe_contrib1	-0.3932	0.2866	-1.37	0.1701
jefe_edad	-0.0366	0.0007	-49.91	0.0000
jefe_sin_pension1	0.7294	0.0246	29.60	0.0000
nmujeres	0.2325	0.0089	26.21	0.0000
nmenores	0.4854	0.0166	29.24	0.0000
maxEducLevelPreescolar	0.1067	0.5112	0.21	0.8347
maxEducLevelPrimaria	0.0898	0.0618	1.45	0.1462
maxEducLevelSecundaria	0.3081	0.0626	4.92	0.0000
maxEducLevelMedia	0.0885	0.0624	1.42	0.1563
maxEducLevelUniversitaria	-0.5159	0.0642	-8.04	0.0000
nocupados	-0.1355	0.0126	-10.73	0.0000
rel_pet	0.5339	0.0743	7.19	0.0000
horas_trab_por_persona	-0.0739	0.0011	-70.17	0.0000

Cuadro 3—Modelo Logit