Universidad de
SanAndrés

Entrega Final de Exploraciones sobre Pobreza, Desarrollo Humano y Políticas Sociales

Análisis y Predicción de la Dinámica pobreza en Argentina

Alumnos:

- Felipe García Vassallo
- Santiago Peci

Docentes:

- Mariano Tommasi
- Facundo Pernigotti Rebullida
- Joaquín Campabadal

Fecha de entrega: Viernes 10/12 de 2021.

1. Introducción:

Mi motivación principal es el estudio de las personas en situación de vulnerabilidad en la Argentina. En particular, decidí concentrarme en la transición y dinámica de la pobreza. A grandes rasgos, busco realizar un análisis temporal de la variación de la pobreza de ingresos, y hacer una aproximación a la vulnerabilidad y la pobreza crónica que sea a lo largo del tiempo. Más en detalle, mi interés y motivación principal está dado por la utilización de una serie temporal que me permita predecir la dinámica de la pobreza (para cada una de las personas en la muestra), y realizar un análisis ex-post sobre la eficiencia de la predicción, la transición de las personas y las variables que los modelos toman como de mayor relevancia. Las predicciones serían realizadas con alguna de las metodologías de *Machine Learning* que vamos a detallar más adelante.

Creo que este estudio puede ser de gran relevancia en varias dimensiones. En primer lugar, la clasificación de la vulnerabilidad de una persona puede ser mucho más precisa utilizando una serie temporal a corto/mediano plazo. Es decir, si nuestra serie temporal tiene 4 ondas en un plazo de 21 meses, ahora podría analizar la vulnerabilidad de las personas según en cuántos de los 4 intervalos de tiempo estuvieron por debajo de la línea de la pobreza. Más allá de que se sigue teniendo el problema de la unidimensionalidad (solo se tiene en cuenta el ingreso), se puede generar un índice más informativo sobre la vulnerabilidad.

En segundo lugar, en el caso de que la predicción de la dinámica sea efectiva, creo que el análisis ex-post sobre las características de cada uno de los grupos predichos y sobre los distintos coeficientes generados por los modelos predictores (detallado más adelante) puede realizar un aporte significativo y desde una metodología distinta al debate sobre qué características de las personas u hogares son relevantes a la hora de explicar las probabilidades de salida y entrada a la pobreza. Al realizar una predicción sobre si la persona u hogar va a ser pobre o no dentro de 1 año, y luego realizar el análisis mencionado, se puede lograr un análisis diferencial que resulte informativo. Además vamos a realizar un ejercicio comparativo entre 3 predicciones distintas, las cuales van a ir omitiendo variables relacionadas con el ingreso con el objetivo de analizar la variación en la eficiencia de las predicciones.

Por último, creo que el hecho de generar un modelo predictor que sea efectivo es relevante de por sí, debido al alto poder predictivo que tienen los modelos de machine learning por fuera de la muestra de entrenamiento. Esto genera que se puedan extrapolar a series temporales que

contengan los mismos datos pero de otros años y así predecir cuáles serían los niveles de pobreza en el corto/mediano plazo, y en particular, cuáles son los individuos que tienen alta probabilidad de caer por debajo de la línea de la pobreza.

Estas tres dimensiones mencionadas pueden ser de gran utilidad a la hora de realizar lineamientos de políticas públicas, no solo ayudando a generar un diagnóstico más acertado sobre el verdadero número de personas y hogares vulnerables en la Argentina y logrando una predicción efectiva de las tasas de pobreza a futuro, sino también en detectar quienes son los vulnerables que en el indicador clásico utilizado no son detectados (y evitar su caída por debajo de la línea de pobreza). Sin embargo, cabe destacar que estaríamos realizando un esfuerzo por caracterizar a los pobres crónicos con información, dada la estructura particular de datos de la EPH, de tan solo un año y medio (bastante menos de lo deseable).

2. Literatura previa:

Existe un gran esfuerzo en la literatura que trata a la transición de la pobreza y la pobreza crónica para la Argentina. Sin embargo, la gran mayoría de los estudios del tema están dados a partir de la econometría. Encontramos un único trabajo realizado por Lucchetti (2018) que logra en primer lugar la construcción de paneles sintéticos, y luego una predicción con la metodología *Lasso* de la dinámica de la pobreza entre 2014 y 2015 para algunos países de América Latina. Nuestro objetivo es ampliar este estudio, buscando no solo realizar varias predicciones para años posteriores que incluya muchos modelos y distintas variables, sino también un análisis ex-post completo e informativo que pueda introducir conclusiones de características relevantes en la transición de la pobreza a la literatura.

Alejo y Garganta (2014), por ejemplo, estudian la dinámica de la pobreza a partir de paneles sucesivos para el período 1997-2012, lo que les permite descomponer la pobreza (variabilidad del ingreso) en componentes permanentes/crónicos y transitorios. Terminan encontrando que "el factor transitorio está asociado principalmente con las características laborales, mientras que las cualidades estructurales y demográficas del hogar, y la educación del jefe y cónyuge del mismo, poseen incidencia fundamental sobre el componente crónico". En principio, la idea del trabajo es analizar si se llega a conclusiones similares a partir del análisis a posteriori.

En Jorge Paz (2002), se analiza la dinámica de la pobreza en el periodo 1998-2000 para hogares e individuos mediante la utilización de la EPH en forma de panel. Se clasifica la

población observada según si se está por debajo de la línea de pobreza y su persistencia. Se pueden identificar lo que en el *paper* llama "pobres persistentes o pobres estructurales", los cuales son todos aquellos que comenzaron y terminaron siendo pobres sin salir de la pobreza en ninguno de los periodos observados. Estos representan el 14,2% del total de hogares de la muestra y 50% de los alguna vez pobres y en términos de individuos, el 20,9% del total de individuos de la muestra y el 68,4% de los individuos alguna vez pobres. También se observa que el 43% de los hogares estuvo por lo menos una vez por debajo de la línea de pobreza.

Por último, encuentra que las variables que mejor explican entrada y salida de la pobreza son la edad, la educación del jefe de hogar y la desocupación del jefe de hogar. La probabilidad de ser pobre está fuertemente relacionada con si el hogar o individuo fue pobre alguna vez y aún mayormente si fue en un periodo anterior inmediato. Además de las variables ya nombradas, la cantidad de perceptores de ingreso en el hogar aumenta la probabilidad de salir de la pobreza o no entrar en ella y en sentido contrario que el hogar se encuentre en NEA o NOA. En una continuación del trabajo a futuro me gustaría incluir variables como la mencionada a las bases y ver si varían significativamente la eficacia de la predicción.

Por otro lado, para estudiar los efectos de shocks negativos dependiendo de la vulnerabilidad del hogar, Cruces y Wodon (2006) analizaron el periodo 1995-2002 mediante la EPH adaptada a panel. Encuentran que los efectos son heterogéneos en la variabilidad del ingreso dependiendo a qué quintil pertenece, por lo tanto, cuanto más rico el hogar menor es el efecto negativo. Los más afectados son los hogares con mayor vulnerabilidad, como pueden ser los que contienen trabajadores informales, desempleados, migrantes recientes o inactivos. Además, observan un mayor impacto negativo en situaciones de crisis, aumentando los efectos en la pobreza. Es por esto que mi idea es utilizar la EPH en dos años en los que la variación interanual de la actividad económica no haya sido significativa, para que la transición de la pobreza detectada tenga que ver lo menos posible con shocks macroeconómicos temporales.

De esta forma, podemos observar que los avances de la literatura Argentina sobre la dinámica y la transición de la pobreza sigue un hilo conductor que es la utilización de la EPH en forma de datos de panel con el objetivo de realizar un seguimiento de un mismo individuo/hogar a lo largo del tiempo. Teniendo en cuenta que la EPH no es ideal para los estudios longitudinales, los autores buscan adaptarse a la escasez de datos y utilizan esta base como un *second best*. De esta forma, la estructura de datos que utilizan es similar a la que buscamos

lograr, pero en la literatura se los aproxima a partir de una diferente metodología (no existe demasiada aproximación al problema a partir del *machine learning* para Argentina).

Aparte del estudio de la dinámica de la pobreza, también revisamos la literatura Argentina sobre la pobreza crónica y su aproximación académica. El concepto de pobreza crónica sirve para caracterizar un grupo demográfico que sufre de privaciones de capacidades en múltiples planos persistentemente, estando en una situación de alta vulnerabilidad. La salida de la pobreza en estos casos es compleja y generalmente no es suficiente el crecimiento económico, ya que las características de los individuos en estas condiciones impiden aprovechar en su totalidad los shocks positivos y las políticas públicas inclusivas.

Para Hulme y Shepherd (2003), un pobre crónico es aquel individuo que padece privaciones de capacidad significativas, de por lo menos cinco años. Agregan que para la comprensión de esta definición se deben tener en cuenta ciertos aspectos.

En primer lugar, la cantidad de períodos bajo la línea de pobreza. A mayor cantidad, menor es la probabilidad de salida en el próximo periodo y mayor es la probabilidad de una transmisión intergeneracional. En Yaqub (2000), se encuentra que si un individuo ya padeció cinco años de pobreza es muy probable que lo siga siendo por el resto de su vida.

En segundo lugar, la privación de ingreso o consumo no es suficiente para determinar un pobre crónico, por eso mismo se deben tener en cuenta múltiples aspectos en donde una persona está privada o no y por eso la pobreza multidimensional es un mejor indicador en el largo plazo para el estudio de la pobreza. Esto es debido a que el ingreso o consumo es un factor fluctuante en el corto plazo. A mayor cantidad de dimensiones que un individuo tenga privadas, menor probabilidad de salir de la pobreza.

Por último, tener en cuenta que el estudio de pobreza crónica es distinto en términos relativos o absolutos. Este último se usa sobre todo en países en desarrollo. En Yaqub (2003), se argumenta que los pobres crónicos medidos relativamente como los del último quintil por ingresos tienen igual de difícil la salida de este estado o hasta más compleja que los pobres crónicos medidos de manera absoluta. Por ejemplo, en el caso argentino Gasparini et al. (2019), utilizan por la limitación de datos un concepto de pobreza crónica relativo tomando como pobres crónicos al 10% más vulnerable de la población.

Para la salida de la pobreza crónica en la literatura de países en desarrollo se encuentra una correlación con la adquisición de terrenos, nivel de educación y nivel al inicio de escolaridad.

Para la movilidad hacia abajo hay una alta correlación con el aumento en la cantidad de personas en un hogar y su número de dependientes.

En la práctica se tienen dos alternativas para estimar la pobreza crónica, el "components approach" y el "spells approach". El primero se fija en la variable ingreso o consumo enfocándose relativamente en mayor medida a la profundidad de la pobreza, el segundo en la entrada y salida de la pobreza orientando el concepto de pobreza crónica a pobreza persistente. El estudio que se busca hacer está en principio más enfocado en este último.

3. Armado de base de datos y estadística descriptiva:

Para realizar este ejercicio empírico vamos a armar nuestra base en panel a partir de datos de distintas Encuestas Permanentes de Hogares (EPH-C) realizadas por el INDEC. Esta base va a contener datos de 2 clusters de individuos en un plazo total de 21 meses. El primer cluster va a tener observaciones del segundo y tercer trimestre de dos años consecutivos, y el segundo del tercer y cuarto trimestre. Esto es debido a que la EPH renueva el 25% de la muestra en cada onda trimestral, por lo que te permite mantener el seguimiento de un mismo individuo u hogar en el plazo máximo consecutivo de 4 trimestres. Además, en la EPH-Continua se utilizan datos para hogares en 2 trimestres consecutivos, luego se retiran temporalmente en los dos trimestres siguientes, y finalmente se vuelven a incorporar en dos trimestres adicionales sucesivos. La idea era buscar dos años en los que la variación interanual de la actividad económica no haya sido significativa, para que la transición de la pobreza detectada tenga que ver lo menos posible con shocks macroeconómicos temporales, y a su vez evitar el periodo en el que los datos del INDEC no eran del todo transparentes. Es por esto que nos decidimos por el periodo que va desde el trimestre 2 del 2016 al trimestre 4 del 2017.

A continuación, un diagrama de como queda la base de datos preliminar en panel:

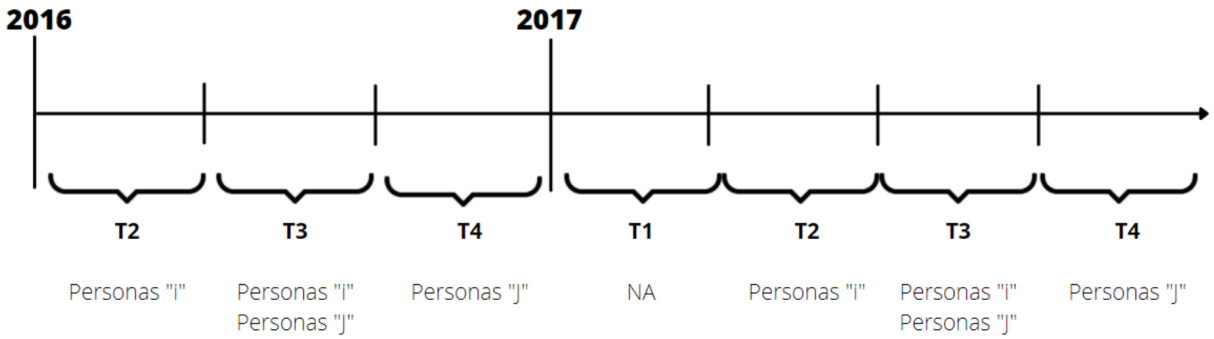


***Figura 1**. Representación base de datos.*

Luego, una vez que tenemos nuestra base preliminar con datos de individuos que reportaron sus ingresos en los 4 periodos, realizamos una limpieza y ciertas modificaciones particulares con el objetivo de realizar la predicción. En primer lugar, reemplazamos ciertos valores que no tenían sentido para que sigan siendo utilizables. Por ejemplo, las edades e ingresos negativos fueron cambiados por el valor promedio de las personas que cumplen el mismo rol en el hogar. Luego agregamos columnas que indican si los individuos son o no pobres en ese periodo utilizando la canasta básica promedio trimestral, el ITF y la tabla de equivalencias. Finalmente, realizamos una selección preliminar de variables explicativas que creemos a priori puedan ser relevantes en la predicción.

Luego de realizar esta limpieza, con el objetivo de controlar que nuestra base de datos siga siendo representativa (y no hayamos incurrido en algún sesgo), calculamos la tasa de pobreza de hogares por trimestre y la comparamos con la del INDEC . Los resultados son muy similares:

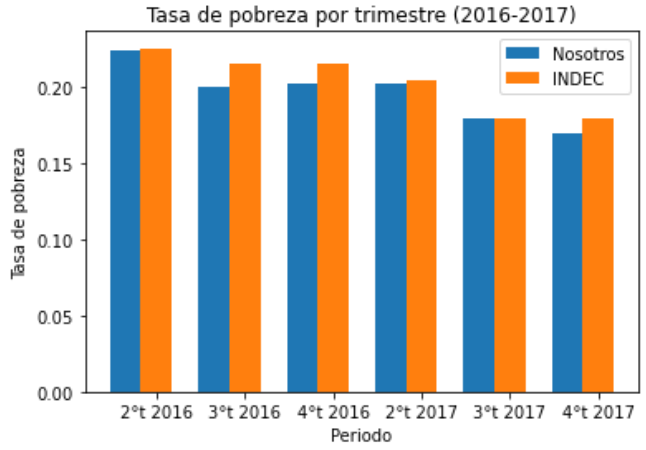


***Gráfico 1**. Tasa de pobreza por trimestre (2016 - 2017).*

Una vez que ya tenemos nuestra base preliminar armada, y comprobamos que sea representativa, realizamos cierta estadística descriptiva. En primer lugar, realizamos una primera aproximación a la dinámica de la pobreza. En azul se mide el porcentaje de pobres que logró salir de la pobreza entre periodos, y en naranja el porcentaje de no pobres que cayó por debajo de la línea de la pobreza. Los resultados son que entre el 20 y el 25% de los pobres logran salir de la pobreza entre períodos, mientras que entre el 10 y 15% de los no pobres cae en la pobreza.

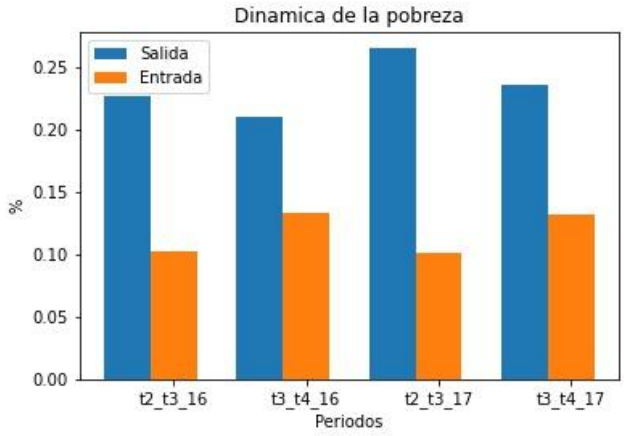


***Gráfico 2**. Dinámica de la pobreza.*

De todas formas, dado que bastante menos del 50% de los individuos eran pobres entre 2016 y 2017, este gráfico no muestra necesariamente que la tasa de pobreza de individuos disminuye considerablemente trimestre a trimestre. Dado que esta representación puede derivar en esta confusión, también representamos la dinámica entre periodos en valores absolutos:

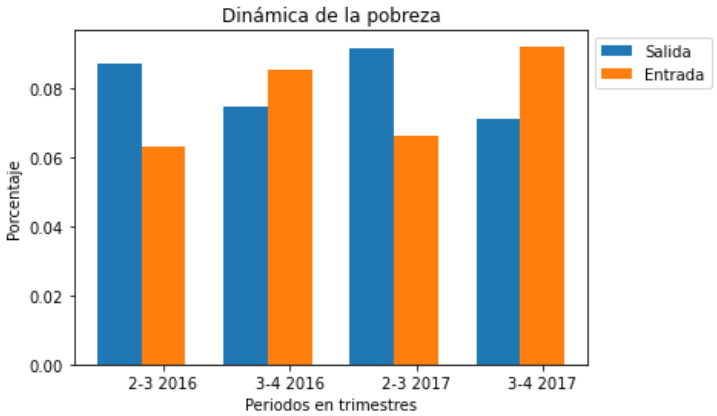


***Gráfico 3**. Dinámica de la pobreza 2.*

Acá se puede observar que del tercer al cuarto trimestre de los dos años hubo una mayor cantidad de personas que entraron a la pobreza, mientras que del segundo al tercero fue mayor la cantidad que salieron.

Luego, realizamos dos gráficos más. Uno que clasifica a los individuos como crónicos, muy vulnerables, vulnerables o no vulnerables según si estuvieron 4, 3 o 2, 1 o 0 períodos por debajo de la línea de la pobreza respectivamente, y el otro que discrimina el nivel educativo entre estos grupos formados.

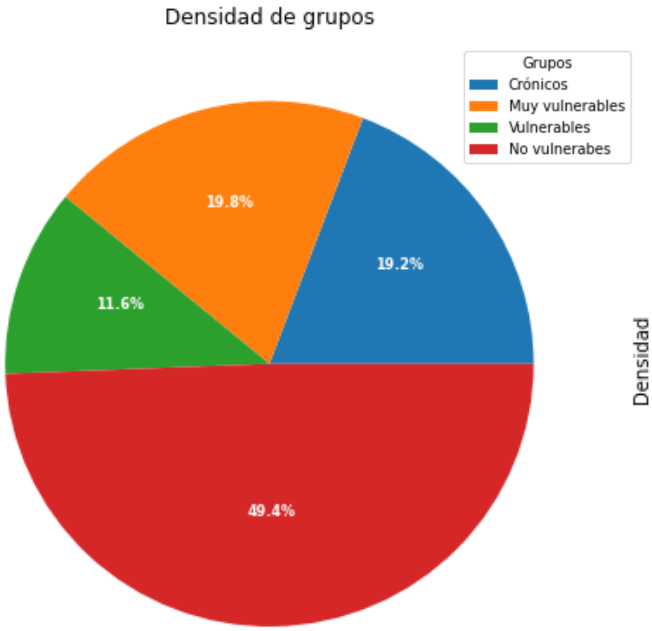


***Gráfico 4**. Densidad de grupos.*

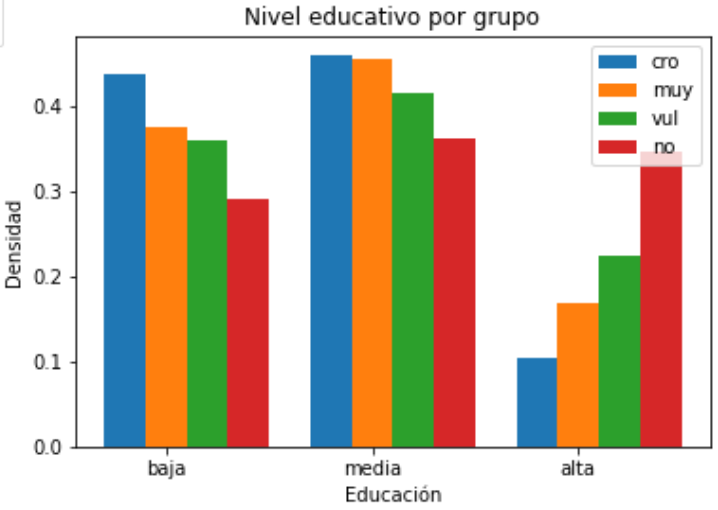


***Gráfico 5**. Nivel educativo por grupo.*

En el gráfico 4 notamos que más del 50% de los individuos fue pobre en al menos uno de los 4 periodos, y que casi el 20% estuvo por debajo de la línea de la pobreza en todos ellos. El gráfico 5 lo que hace es discriminar qué porcentaje dentro de cada uno de los grupos formados corresponde a los distintos niveles educativos. Más en detalle, formamos cada nivel educativo a partir de una división arbitraria sobre el nivel educativo más alto que el individuo cursa actualmente o alguna vez cursó:

- **Bajo**: Jardín/preescolar, Primario 3 o EGB
- **Medio**: Secundario o Polimodal

- **Alto**: Terciario, Universitario, Posgrado universitario o Educación especial (discapacitado).

Este gráfico nos pareció muy interesante e informativo ya que muestra muy bien como la vulnerabilidad de las personas tiene una correlación negativa muy marcada con el nivel educativo.

4. Metodología de las predicciones:

Esta parte del trabajo consiste en realizar 3 predicciones interanuales para cada individuo, las cuales utilicen las características de su segunda observación para predecir si van a encontrarse en situación de pobreza de ingresos o no en su última observación. Por ejemplo, para el caso de un individuo tipo "i", con sus características del trimestre 3 de 2016 se busca predecir si va a ser pobre o no en el trimestre 3 de 2017 (ver Figura 1 para mayor comprensión).

Como comentamos, se van a realizar tres predicciones. La primera va a ser realizada incluyendo todas las variables de la EPH selectas como relevantes (y la dummy generada de pobreza actual). La segunda se realiza excluyendo las variables definidas como "variables ingreso" en el Diseño de Registro realizado por el INDEC. Esto es con el objetivo de observar cómo varía la eficacia de la predicción, y analizar en qué medida la transición de las personas en el corto/mediano plazo es fundamentalmente producto del nivel de ingreso presente y también cómo varían los coeficientes de ciertos modelos. De todas formas, esta última predicción sigue teniendo ciertas variables que de manera indirecta denotan ingresos, por lo que se hace una tercera y última predicción con variables que hacen referencia exclusivamente a características habitacionales, nivel educativo y composición del grupo familiar.

Entonces, las variables predictoras serían todas las que contiene la EPH y seleccionamos como relevantes al limpiar la base de datos (primero con y quitando las variables de ingreso), y la variable a predecir sería una dummy de si el individuo es o no pobre en su última observación (un año a futuro). Es decir, no se predice el nivel de ingreso futuro y luego se lo

clasifica como pobre o no pobre, sino que se predice directamente si el individuo va a ser pobre o no. (Yi = {0, 1}).

Los modelos a utilizar son los más utilizados en la literatura de *machine learning*: Lasso, Ridge, Elastic Net, Cart, Bagging, Boosting, Random Forest, Knn, Logit, Discriminante Lineal y SVM. Vamos a definir por cross-validation los parámetros óptimos para cada uno, y luego seleccionar el modelo que mejor predice.

Vamos a incluir una breve explicación de lo que hace cada modelo para un mejor entendimiento del trabajo:

1. **Lasso:**

$$R_l(\beta) = \sum_{i=1}^{n} (y_i - x'_i\beta)^2 + \lambda \sum_{s=2}^{p} |b_s|, \quad \lambda > 0$$

Esta es la función objetivo que el modelo Lasso busca minimizar. Notamos que el primer término penaliza la falta de ajuste (igual que el método MCO), pero ahora se le agrega un segundo término que penaliza la cantidad de variables predictoras. Esto logra que ciertos coeficientes sean llevados hacia cero (con respecto a las de MCO), generando en ellos un sesgo, pero haciendo que el modelo termine ganando en materia predictiva dada la disminución de la varianza. El parámetro de penalización (λ *o* $Shrinkage$) óptimo que minimice el error de pronóstico es elegido por la metodología Cross-Validation. Notamos que esta penalización, la elección del parámetro y el cómputo de error por cross validation son cruciales, ya que evitan que se genere el problema de overfit del modelo y logran que este mantenga una buena performance por fuera de la muestra (lo que cumple nuestro objetivo de que sea utilizable para predecir pobreza futura).

2. **Ridge:**

$$R_l(\beta) = \sum_{i=1}^{n} (y_i - x'_i\beta)^2 + \lambda \sum_{s=2}^{p} (b_s)^2, \quad \lambda > 0$$

Misma lógica que Lasso pero ahora la penalización está elevando al cuadrado, lo que genera que las soluciones de los coeficientes tiendan a ser interiores y no necesariamente de esquina. Es decir, ahora no sucede que muchos coeficientes toman valores iguales o muy cercanos a cero, por lo que genera un mayor interpretabilidad con respecto a Lasso. Se puede demostrar que existe λ tal que Ridge predice mejor que MCO.

3. **Elastic Net:**

$$R_l(\beta) = \sum_{i=1}^{n} (y_i - x'_i\beta)^2 + \lambda_1 \sum_{s=2}^{p} (b_s)^2 + \lambda_2 \sum_{s=2}^{p} |b_s| \quad , \quad \lambda > 0$$

Toma lo mejor de Lasso (la elección de variables predictoras significativas y la reducción del modelo) y lo mejor de Ridge (resuelve un problema técnico de inestabilidad por agrupamiento muy común en bases como la eph que cuentan con variables muy correlacionadas). Tiene gran poder predictivo.

4. **Vecinos Cercanos (KNN):**

Es un método muy sencillo que va a pedir si la persona va a ser pobre o no según si las K observaciones más similares son pobres o no. La cantidad de vecinos a utilizar (K) se elige por cross-validación. En bases de datos con muchas variables como la EPH puede tener problemas prácticos al definir la cercanía entre observaciones.

5. **CART:**

Cart es un método muy utilizado para capturar efectos no lineales de las variables predictoras en la variable a predecir. En particular, se generan predicciones a partir del armado de un árbol de decisión que contiene en cada uno de sus nodos particiones generadas a partir de valores de las variables predictivas. En los nodos terminales se encuentran las predicciones efectivas. Notamos que el árbol tiende a poner en los primeros nodos a las variables más relevantes, por lo que podemos realizar un análisis ex post de la predicción y ver que variables fueron detectadas por el modelo como las más relevantes en términos de minimizar error de pronóstico (y cómo se relacionan entre ellas).

Es un método muy intuitivo (que nos va a permitir interpretar qué variables son relevantes), pero no suele ser de los mejores predictores debido a que puede llegar a tener una alta varianza. Bagging, Random Forest y Boosting buscan resolver ese problema (más allá de que no son tan interpretables o intuitivos).

6. **Bagging:**

Modelo que sigue la misma lógica que CART, pero aplican un método llamado Bootstrap que predice el árbol "B" veces pero tomando muestra con reemplazo de las observaciones, lo que potencialmente disminuye la varianza. Esto no sucede cuando hay un predictor muy fuerte, por lo que los árboles ponderados en la predicción final siguen siendo muy parecidos.

7. **Random Forest:**

Lo que hace es el mismo procedimiento de Bagging pero en cada una de las "B" iteraciones elegir "m<p" predictores, lo que genera que se resuelva el problema de alta varianza incluso en situaciones en las que haya un predictor muy fuerte.

8. **Boosting:**

Es un algoritmo que genera "M" modelos predictores débiles a partir de árboles CART que tengan pocas ramas, pero ponderando en mayor medida las observaciones que fueron mal predichas en la iteración anterior. Esto permite derivar una predicción a partir del promedio ponderado de las predicciones de los M árboles (ponderado por el error de pronóstico de cada modelo). Suele ser una técnica de gran performance predictiva, pero con baja interpretabilidad).

9. **SVM:**

Este método es un algoritmo que sitúa a las observaciones en un plano de acuerdo a sus variables predictoras y luego genera un hiperplano separador que realiza la predicción según si las observaciones quedaron de un u otro lado. Es muy bueno en términos predictivos, pero por supuesto que es imposible de interpretar un plano con tantas variables (como lo sería uno que utilice las de la eph).

10. Logit:

Utiliza la siguiente función:

$$P(Y = 1 / X) = \frac{e^{X.B}}{1+e^{X.B}},$$

donde B es el vector de los "p" coeficientes y es estimado por máxima verosimilitud. Luego con esto computa la probabilidad de que un individuo sea pobre dentro de un año dadas las características presentes, y utiliza el Clasificador de Bayes:

$$P(Y = 1 / X) \geq 1/2 \quad \rightarrow \widehat{Yi} = 1$$
$$P(Y = 1 / X) < 1/2 \quad \rightarrow \widehat{Yi} = 0$$

Sabemos de antemano que algunos de los modelos pueden incurrir en ciertos problemas prácticos. Por ejemplo, cuando la cantidad de observaciones es mayor a la cantidad de variables y además hay una alta correlación entre predictores, Ridge tiende a funcionar mejor que Lasso (debido a que este último tiende a eliminar varias de estas variables quedándose solo con unas pocas). Elastic net, a su vez, suele predecir bien bajo estas circunstancias, reduciendo la dimensionalidad y eligiendo correctamente los grupos de variables significativas. KNN y CART también tienen sus limitaciones prácticas en esta predicción. Vecinos Cercanos tiende a perder eficiencia cuando el problema tiene tantas dimensiones debido a que comienza a tener dificultades para definir la cercanía. Por su parte, Cart, tiene un problema de alta varianza que puede ser corregido por modelos como Random Forest y Bagging (como ya mencionamos).

Sin embargo, más allá de estos detalles teóricos que a priori nos dan una idea sobre la eficacia relativa de los distintos modelos, la elección final va a ser realizada con la comparación de los outputs como el error cuadrático medio (ECM) y la accuracy (ACC).

Como comentamos, luego de realizar la predicción, vamos a analizar la eficacia de los modelos y detectar cual y con qué parámetros predice mejor. En el caso de que la predicción del mejor modelo haya sido efectiva, no solo tendríamos un modelo que pueda predecir la pobreza interanual, sino que también vamos a poder realizar un análisis a posteriori de algunos de los modelos mencionados que resulte informativo y detectar si efectivamente encontramos las correlaciones que la literatura existente marca como relevantes en la transición y la dinámica de la pobreza.

En Lasso, Ridge y Elastic Net, buscaremos analizar y describir qué coeficientes no parecen ser significativos, cuales sí lo son, y qué signo tienen. En los modelos CART, analizar las ramas generadas y el orden de los nodos.

5. Resultados:

Como habíamos mencionado, en primer lugar entrenamos a cada uno de los modelos con una serie distinta de parámetros y luego calculamos el ECM promedio de cada iteración por cross validation. En la tabla 1 observamos cual es el mejor parámetro para cada modelo y que error de pronóstico contiene cuando se utilizan todas las variables (incluidas las de ingreso).
Luego, en la tabla 2 repetimos el procedimiento pero excluyendo las variables de ingreso.
Recordamos que tanto el ECM como la accuracy son calculados a partir de observaciones no utilizadas en la muestra de entrenamiento, por lo que representan ajuste por fuera del modelo entrenado.

	modelo	ecm	parámetro	accuracy
0	Lineal	0.230699	0.4	0.769301
0	KNN	0.199818	3.0	0.800182
0	SVM	0.244021	3.0	0.755979
0	Bagging	0.212534	3.0	0.787466
0	Logit	0.174992	10000.0	0.825008
0	CART	0.085377	20.0	0.914623
0	RandomForest	0.099606	20.0	0.900394
0	Boosting	0.033000	10.0	0.967000
0	Lasso	0.162277	100000.0	0.837723
0	Ridge	0.159552	1000.0	0.840448

***Tabla 1**. Predicciones con ingreso.*

	modelo	ecm	parámetro	accuracy
0	Lineal	0.217984	0.01	0.782016
0	KNN	0.270360	17.00	0.729640
0	SVM	0.238874	1.00	0.761126
0	Bagging	0.227066	3.00	0.772934
0	Logit	0.212534	1000.00	0.787466
0	CART	0.175295	20.00	0.824705
0	RandomForest	0.158644	20.00	0.841356
0	Boosting	0.093854	10.00	0.906146
0	Lasso	0.220103	1000.00	0.779897
0	Ridge	0.210718	0.10	0.789282

***Tabla 2**. Predicciones sin ingreso.*

En primer lugar, notamos que en ambas predicciones todos los modelos cuentan con una accuracy por encima del 70%, teniendo como promedio un 83.5% cuando se utilizan las variables de ingreso, y un 79.2% cuando no. Es decir, los modelos no solo tienen una eficacia bastante alta, sino también que en promedio no parecería caer significativamente cuando no se utilizan las variables de ingreso.

El modelo que mejor predice en ambos casos es el de Boosting con parámetro igual a 10. Con las variables de ingreso logra una accuracy extremadamente alta del 96%, con una tasa de falso pronóstico de pobreza (fp) del 5% y de falsos pronóstico de no pobreza (fn) del 2% (ver en tabla 3 la matriz de confusión). Este modelo resulta de alta relevancia, ya que si se cuenta con toda la información proporcionada por la EPH, se puede pronosticar la tasa de pobreza interanual con una eficacia promedio mayor al 95%.

En cuanto a la predicción sin las variables de ingreso, se logra una accuracy del 90%, con una tasa de falso pronóstico de pobreza (fp) del 15% y de falsos pronóstico de no pobreza del 8% (ver en tabla 4 la matriz de confusión). Esto resulta también muy relevante, ya que la predicción continua teniendo una eficacia promedio mayor al 90%, lo que respalda la literatura econométrica que afirma que la probabilidad de salida o entrada a la pobreza en el mediano plazo depende no sólo del monto de ingresos actual, sino también de otras dimensiones que vamos a estar analizando a continuación.

	Negativos	Positivos
Negativos	1921	47
Positivos	76	1256

***Tabla 3**. Matriz de confusión con ingreso.*

	Negativos	Positivos
Negativos	1807	159
Positivos	203	1134

***Tabla 4**. Matriz de confusión sin ingreso.*

Notamos que en las filas de la matriz de confusión tenemos el pronóstico realizado, y en las columnas la verdadera realización. Para obtener una interpretación gráfica de estos resultados, utilizamos la curva ROC, la cual muestra como varía la tasa de verdaderos positivos en relación a la tasa de falsos positivos a medida que va variando el umbral de decisión. En los casos extremos de la curva, todas las observaciones son pronosticadas como 0 o 1, dependiendo de sí se penaliza infinitamente el error de tipo 1 o de tipo 2. En los puntos intermedios, se puede notar que la concavidad de la curva muestra como crece en mayor medida la tasa de verdaderos positivos con respecto a la de falsos positivos (caso contrario el modelo sería peor que lanzar una moneda). Entonces, a mayor área por debajo de la curva (AUC), mejor es la predicción. Observamos:

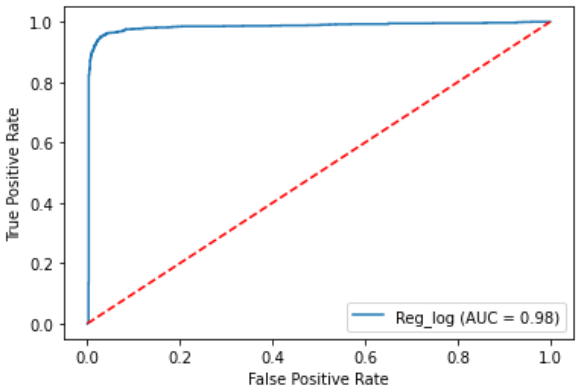


***Gráfico 6**. Curva ROC predicción con ingreso.*

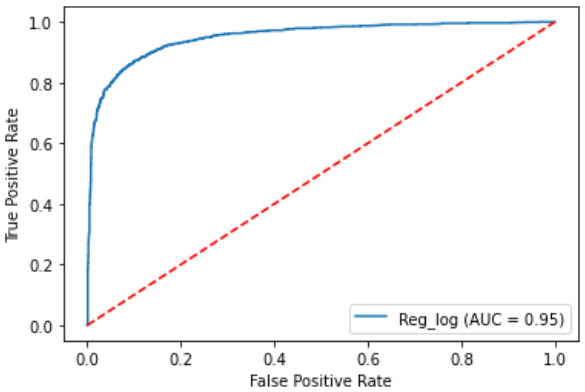


***Gráfico 7**. Curva ROC predicción sin ingreso.*

Notamos que en el gráfico 6 la AUC es de 0.98, mientras que en el gráfico 7 la AUC es de 0.95 (en ambos casos considerablemente alta).

En el caso de los modelos Ridge, que logran una eficacia del 84% y 78%, buscamos analizar los coeficientes de los modelos entrenados. La tabla 1 muestra los coeficientes que parecerían ser más importantes en la predicción con las variables de ingreso (aquellos que son mayores a

0.10 en valor absoluto), mientras que la tabla 2 muestra los coeficientes más importantes en la predicción sin variables de ingreso (mayores a 0.70 en valor absoluto).

Variable	Coeficiente
CAT_INAC	0.104660
IV2	-0.110483
II8	0.152548
V5	-0.130616
IXTOT	0.343028
DECIFR	-0.178307
DECCFR	-0.338159
Pobre	0.227681

***Tabla 5**. Coeficientes Ridge con ingreso*

***Tabla 6**. Coeficientes Ridge sin ingreso*

Variable	Coeficiente
ESTADO	-0.702831
PP10C	-0.735159
PP07G1	0.908467
V5	-0.969670
V4	0.949485
IV8	2.103017
V8	0.828879
V9	2.867172
V10	2.284494
V18	0.804712

El detalle de cada variable se encuentra en el anexo del trabajo.
Recordamos que la interpretabilidad de estos coeficientes es limitada, dado que están sesgados debido al *shrinkage* mencionado, y que ciertas variables no están normalizadas.

Sin embargo, se pueden desprender algunas conclusiones interesantes. En primer lugar, notamos que los 3 coeficientes de mayor valor en la predicción con variables de ingreso son los correspondientes a las variables de ingreso total familiar, el número de decil del ingreso per cápita familiar del total EPH y la dummy generada de si el individuo es pobre o no actualmente. Es decir, el modelo predictivo le da alta preponderancia a las variables de ingreso, y es consistente con las conclusiones mencionadas de los trabajos de Alejo y Garganta (2014) y Paz (2002). Sin embargo, se destacan como variables significativas e interesantes las de cuántos ambientes/habitaciones tiene la vivienda en total, la del qué combustible es utilizado en el hogar para cocinar, la de la categoría de inactividad del individuo y la de la cantidad de miembros en el hogar.

En cuanto a la predicción Ridge sin las variables de ingreso, notamos que ahora se le da una mayor preponderancia a muchas variables que antes no la tenían. Una de las variables de mayor coeficiente es la de si el hogar cuenta con baño o letrina, mientras que se destaca como

variables relevante la composición de la cubierta exterior del techo del hogar. Cabe destacar que el modelo tomó como importantes a un grupo de variables que no estaban en la sección de variables de ingreso en el diseño de registro y estructura para bases preliminares realizado por el INDEC, pero que están denotando de alguna forma un ingreso (Ej: V18, V8, etcétera). Es por esto que para la continuación del trabajo para la tesis vamos a realizar una tercera predicción que incluya sólo variables de identificación, características habitacionales del hogar y características del miembro del hogar (tratando de aislar lo mayor posible el efecto ingreso en esta predicción sobre la transición).

Finalmente, vamos a realizar un análisis sobre las ramas superiores de los árboles de decisión generados por el modelo CART, los cuales lograron una accuracy del 91% y 82%. En la figura 2 observamos el árbol del modelo con variables de ingreso, y en la figura 3 el árbol del modelo sin las variables de ingreso.

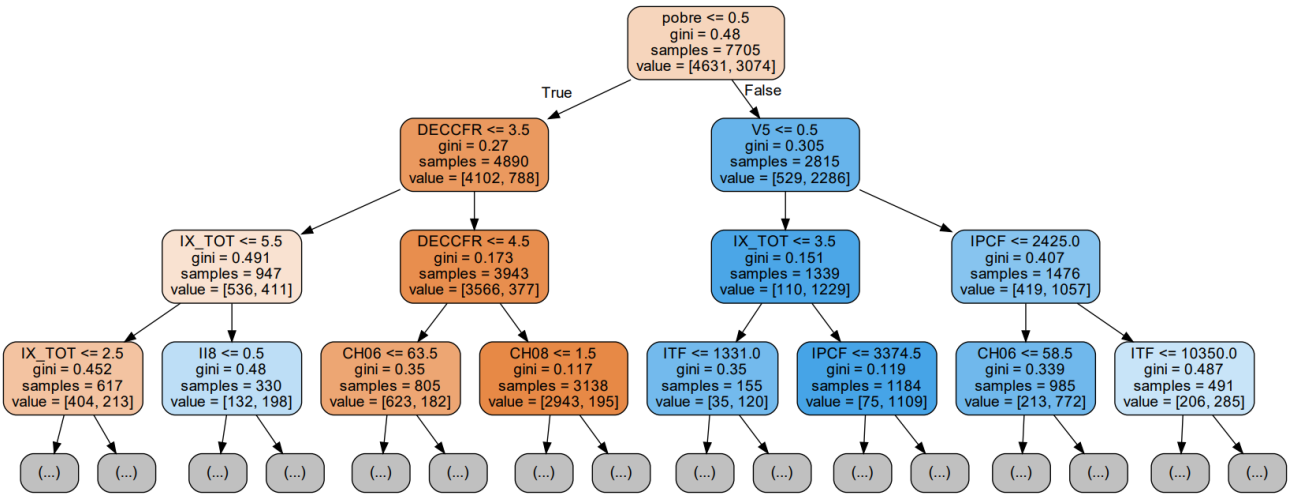


***Figura 2**. Árbol CART con variables de ingreso.*

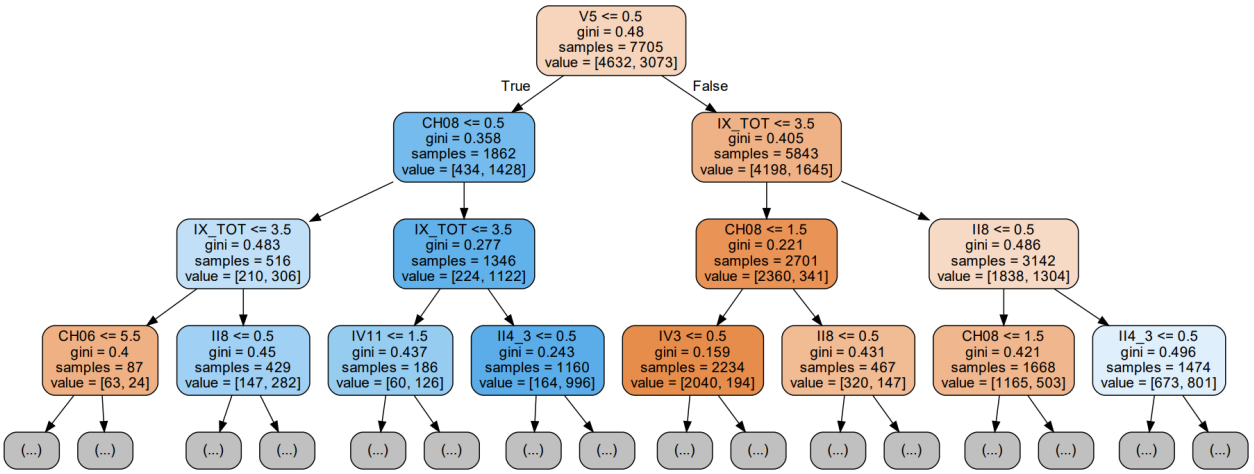


***Figura 3**. Árbol CART sin variables de ingreso.*

En la figura 2, analizando los dos nodos principales, notamos que el modelo marca como las tres variables más relevantes a la dummy de pobreza en la actualidad, al número de decil del ingreso per cápita familiar (en el caso de que pobre = 0), y a la que indica si en los últimos 3 meses las personas del hogar han recibido ayuda social o un subsidio (en el caso de que pobre = 1). Luego se podría realizar un análisis más minucioso de la incidencia de todas las variables, pero nuevamente vemos la alta preponderancia de las variables de ingreso.

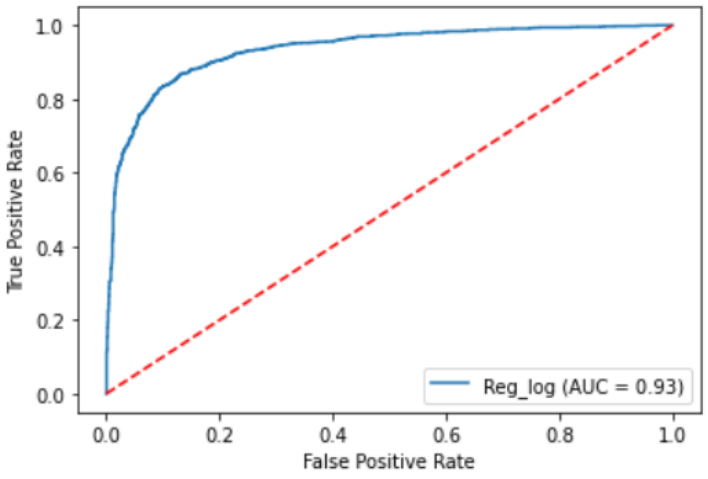
En la figura 3, en cambio, vemos que las variables más relevantes pasa a ser la de ayuda social o subsidio del gobierno y/o iglesia, y luego se incluye la cantidad de miembros del hogar por un lado (IX_TOT), y si tiene cobertura médica o no por el otro(CH08).

Sin embargo, notamos que en la segunda predicción, más allá de haber limpiado las variables que el INDEC denotaba como variables ingreso, seguían habiendo variables que indirectamente estaban denotando un ingreso corriente (como por ejemplo una dummy que indicaba si el hogar recibía un subsidio o apoyo de la iglesia). Es por esto que para desagregar aún más el efecto de las variables ingreso en la efectividad del modelo decidimos hacer una tercera predicción con exclusivamente las variables que denotan características particulares de las personas y características del hogar.

Los resultados fueron los siguientes:

	modelo	ecm	parámetro	accuracy
0	Lineal	0.230699	0.01	0.769301
0	KNN	0.275507	15.00	0.724493
0	SVM	0.235241	1.00	0.764759
0	Bagging	0.228883	3.00	0.771117
0	Logit	0.234030	0.01	0.765970
0	CART	0.199818	20.00	0.800182
0	RandomForest	0.176809	20.00	0.823191
0	Boosting	0.123524	10.00	0.876476
0	Lasso	0.218286	10.00	0.781714
0	Ridge	0.233121	100000.00	0.766879

Notamos que a pesar de sacar una gran cantidad de variables que denotaban ingreso de manera indirecta, la eficacia de las predicciones se mantiene bastante alta. El mejor modelo predictor fue Boosting, logrando un 87% de accuracy (apenas 3 puntos porcentuales menos que la predicción anterior).




```
array([[1799,  166],
       [ 253, 1085]], dtype=int64)
```

5. Conclusiones y limitaciones:

En primer lugar, realizamos estadística descriptiva relevante para analizar la transición de la pobreza en los años 2016 y 2017 para Argentina. Encontramos resultados interesantes como

que más del 50% de los individuos habría estado al menos uno de los 4 periodos analizados por debajo de la línea de la pobreza, que entre un 20 y un 25% de los pobres logran salir de la pobreza entre periodos y un gráfico que ilustra muy bien la correlación entre los niveles educativos y la vulnerabilidad de las personas.

Luego, a partir de la utilización de metodologías de machine learning obtuvimos un modelo eficaz predictor de la pobreza interanual. Utilizando la información proporcionada por la EPH se pudo predecir con un 96% de accuracy la transición de la pobreza interanual a nivel individuo en periodos desde 2016 hacia el 2017. Luego, demostramos que este ejercicio se puede realizar sin perder demasiado poder predictivo sin utilizar las variables directamente relacionadas con el monto de ingreso familiar. Esto va en línea con la literatura mencionada, la cual marca que ciertas variables de la composición del hogar suelen tener alta relevancia en estas transiciones. Sin embargo, notamos que en el plazo interanual en el trabajamos, el monto de ingresos actual sigue siendo el factor más influyente a la hora de definir la probabilidad de pobreza a futuro. Por último, realizamos una tercera predicción anulando toda variable que denote ingreso directa o indirectamente para ratificar los resultados encontrados.

Cabe destacar que en las diversas predicciones, aparte de las variables de ingreso corriente, se destacó como relevante la de ayuda social o subsidio del gobierno y/o iglesia, la cantidad de miembros del hogar y si se tiene cobertura médica o no.

Las principales limitaciones de esta metodología son que, dada la inexistencia de datos de un mismo individuo en un plazo mayor al de 4 ondas temporales, es difícil sacar conclusiones de gran robustez sobre la pobreza crónica o realizar predicciones a un mediano o largo plazo que potencialmente le den mayor preponderancia a variables que no sean de ingreso.

Anexo:

Detalle de variables EPH:

ESTADO: Condición de actividad: 0 = Entrevista individual no realizada (no respuesta al cuestionario individual) 1 = Ocupado 2 = Desocupado 3 = Inactivo 4 = Menor de 10 años.

PP10C: ¿Durante el tiempo de desocupación hizo algún trabajo / changa? 1= Sí 2= No

PP07G1: Su trabajo le proporciona vacaciones pagas? 1 = Si 2 = No

V4: La cubierta exterior del techo es de: 1. membrana / cubierta asfáltica 2. baldosa / losa sin cubierta 3. pizarra / teja 4. chapa de metal sin cubierta 5. chapa de fibrocemento / plástico 6. chapa de cartón 7. saña / tabla / paja con barro / paja sola. 9. N/S. Departamento en propiedad horizontal

IV8: ¿Tiene baño / letrina? 1 = Sí 2 = No

V8: ¿En los últimos tres meses, las personas de este hogar han vivido algún alquiler (por una vivienda, terreno, oficina, etc.) de su propiedad? 1 = Sí 2 = No

V9: ¿En los últimos tres meses, las personas de este hogar han vivido ganancias de algún negocio en el que no trabajan? 1 = Sí 2 = No.

V18: Tuvieron otros ingresos en efectivo (limosnas, juegos de azar, etc.) 1 = Sí 2 = No.

CAT_INAC: Categoría de inactividad: 1 = Jubilado / Pensionado 2 = Rentista 3 = Estudiante 4 = Ama de casa 5 = Menor de 6 años 6 = Discapacitado 7 = Otros.

IV2: ¿Cuántos ambientes/habitaciones tiene la vivienda en total? (sin contar baño/s, cocina, pasillo/s, lavadero, garaje).

II8: Combustible utilizado para cocinar: 01 = Gas de red 02 = Gas de tubo / garrafa 03 = Kerosene / leña / carbón

V5: ¿En los últimos tres meses, las personas de este hogar han vivido de subsidio o ayuda social (en dinero) del gobierno, iglesias, etc.? 1 = Sí 2 = No

IX_Tot: Cantidad de miembros del hogar.

DECIFR: Número de decil del ingreso total del hogar del total EPH.

DECCFR: Número de decil del ingreso per cápita familiar del total EPH

CH08: ¿Tiene algún tipo de cobertura médica por la que paga o le descuentan? 1 = Obra social (incluye PAMI) 2 = Mutual / prepaga / servicio de emergencia 3 = Planes y seguros públicos 4 = No paga ni le descuentan 9 = Ns/Nr 12 = Obra social y mutual / prepaga / servicio de emergencia 13 = Obra social y planes y seguros públicos 23 = Mutual / prepaga / servicio de emergencia/ Planes y seguros públicos 123 = obra social, mutual/prepaga/ servicio de emergencia y planes y seguros públicos

Pobre: Dummy generada que indica si el individuo se encuentra en pobreza de ingresos (=1) o no (=0).

Bibliografía

Alejo, Osvaldo Javier; Garganta, Santiago; Pobreza crónica y transitoria: evidencia para Argentina 1997-2012; Centro de Estudios Distributivos, Laborales y Sociales; Documentos de trabajo (CEDLAS); 175; 12-2014; 1-36. **1**

Cruces, Guillermo Antonio & Wodon, Quentin, 2006. "Risk-adjusted poverty in Argentina: measurement and determinants," Financiamiento para el Desarrollo 182, Naciones Unidas Comisión Económica para América Latina y el Caribe (CEPAL). **2**

Foster, James and Santos, Maria Emma (2013), "Measuring Chronic Poverty", in Betti, G. and Achille, L. Poverty and Social Exclusion. New Methods of Analysis, Cap. 7, pp. 143-165. Routledge. **3**

Gasparini, L., Tornarolli, L., Gluzmann, P. (2019), "El desafío de la pobreza en Argentina: diagnóstico y perspectivas", CEDLAS, CIPPEC y PNUD. **4**

Hulme, D. and Shepherd, A. (2003) "Conceptualizing Chronic Poverty," World Development 31:403-423. **5**

Jalan, J and Ravallion, M. (2000) "Is Transient Poverty Different? Evidence for Rural China," Journal of Development Studies 36: 82-89. **6**

Jorge A. Paz, 2002. "Una introducción a la dinámica de la pobreza en la Argentina," CEMA Working Papers: Serie Documentos de Trabajo. 226, Universidad del CEMA. **7**

Lucchetti, Leonardo; Corral, Paul; Ham, Andrés; Garriga, Santiago (2008). Lassoing welfare dynamics with cross-sectional data. Policy Research Working Paper. World Bank. **8**

Yaqub, S. (2000). Intertemporal welfare dynamics: extents and causes. In 'Globalization: new opportunities, new vulnerabilities' workshop. Brookings Institution, Carnegie Endowment. **9**

Yaqub, S. (2003). Chronic poverty: scrutinising patterns, correlates, and explorations. CPRC Working Paper 21. Manchester: IDPM, University of Manchester.