Universidad de
SanAndrés

Universidad de San Andrés

Departamento de Economía

Maestría en Economía

Predicción de la pobreza en Argentina usando Random Forest

Autor: Cristian Chagalj
DNI: 34.617.087

Mentor de Tesis:
Bruno Cardinale Lagomarsino

Victoria, San Fernando (BA) – Octubre de 2019

Predicción de la pobreza en Argentina usando Random Forest*

Cristian Chagalj**

Maestría en Economía
Departamento de Economía, Universidad de San Andrés

Octubre 2019

Resumen

Este trabajo utiliza el algoritmo de aprendizaje Random Forest (RF) aplicado a la Encuesta Permanente de Hogares (EPH) para encontrar los principales predictores no monetarios de la pobreza en Argentina. El principal resultado es que el algoritmo RF permite identificar a los hogares pobres con un error de predicción del 19 % y a los hogares no pobres con un error del 15 % en la especificación más adecuada del modelo, en la cual se implementa una corrección por desbalanceo muestral de la variable de respuesta. Finalmente, se practican distintos chequeos de robustez y se presentan los resultados preliminares de las estimaciones. Los predictores principales de la situación de pobreza son: la cantidad de miembros del hogar, la edad del principal sostén del hogar, la cobertura médica y el nivel de educación. Finalmente, se discuten ventajas y desventajas de la metodología utilizada.

Palabras claves: Random Forest- Argentina - Pobreza

Códigos JEL: C53 - I32

Abstract

This paper uses a statistical learning technique, the Random Forest algorithm, applied to the Encuesta Permanente de Hogares (EPH) to find the main predictors of poverty in Argentina, leaving aside monetary predictors. The Random Forest algorithm can predict with 85 percent of certainty the poverty status of a household, in the most appropriate specification. Different robustness checks are practiced and preliminary results are presented. Main predictors are: number of household members, head of household's age, medical coverage and level of education. A discussion about advantages and disadvantages of the methodology are presented.

Keywords: Random Forest- Argentina - Poverty

JEL Classification: C53 - I32

*La idea de este trabajo surgió durante la lectura dirigida del curso de Big Data del profesor Walter Sosa Escudero en el ámbito de la Universidad de San Andrés. La presente tesis contó con la indispensable contribución de Noelia Romero.

**cristianchagalj@hotmail.com

1. Introducción

En los años 2000 la gran mayoría de los países de América Latina experimentaron un crecimiento económico, disminuyendo la pobreza y mejorando los indicadores referidos al mercado laboral Guillermo Cruces (2016). En particular, en Argentina, la tasa de pobreza representaba más del 40 % de los hogares a finales de 2003, cayendo a un 9 % para 2015. Sin embargo, las estadísticas del país comenzaron a ser cuestionadas a raíz de la actuación del gobierno en 2006 Cavallo (2013). En este sentido, es que resulta relevante que se desarrollen diversas técnicas que mejoren la identificación de los hogares pobres, los cuales son usualmente el target de las políticas públicas McBride y Nichols (2015).

La introducción de los métodos de *statistical learning* se encuentra en una etapa experimental para la economía. Como mencionan Einav y Levin (2013) los economistas deben desarrollar habilidades computacionales, dar estructura a datos más complejos (cómo organizar y reducir la dimensionalidad), elaborar formas de resumir, describir y analizar la información de los datos. Más aún, la aplicación de dicha metodología ha sido ampliamente utilizada en otras disciplinas pero poco explorada en economía, con excepciones como Caruso, Sosa-Escudero, y Svarc (2015); Keely y Tan (2008).

El objetivo de este trabajo es identificar a los hogares pobres y encontrar las características más relevantes del hogar para reconocerlos. Reconocer las variables más importantes que predicen el estatus de los hogares, puede servir de guía para futuros modelos que expliquen los factores y mecanismos de mejoras del bienestar social con distintas medidas de políticas públicas.

En el presente trabajo se aplica uno de los métodos de statistical learning para clasificar hogares pobres en Argentina. El aprendizaje de *random forest* como clasificador cuenta con la ventaja de no producir *overfitting*, dado que el error de generalización converge a una constante bajo la Ley de Grandes Números. Además de contar con ventajas computacionales con respecto a otros algoritmos, RF genera la descorrelación de los predictores (una ventaja con respecto a *bagging*).

Utilizando los datos de la Encuesta Permanente de Hogares (EPH), se utiliza la metodología del INDEC para establecer el estatus de pobreza de los hogares en el Gran Buenos Aires entre 2003 y 2015. El principal resultado es que este algoritmo permite identificar a los hogares pobres con un error de predicción del 19.2 % y con el 15.1 % a los hogares no pobres con la especificación más adecuada del modelo, donde se implementan correcciones por el desbalanceo muestral de la variable de respuesta. Teniendo en cuenta distintas mediciones, se encuentra que las variables más importantes para predecir el estatus de pobreza de un hogar son: la cantidad de miembros del hogar, edad y cobertura médica del principal sostén del hogar. Se incorporan distintas metodologías de robustez y ejercicios que ayudan a entender mejor la identificación de los hogares pobres.

El trabajo de Thoplan (2014) es uno de los primeros en utilizar el método de Random Forest (RF) para mejorar la clasificación de pobres, definiendo a un individuo como pobre cuando se encuentra por debajo de la línea de la pobreza. En particular, busca distinguir entre los distintos tipos de individuos para el año 2000, donde el nuevo censo de Mauritania no contiene la variable de ingreso. Utilizan la línea establecida en 2001 y 2002 para la clasificación, encontrando que las variables más importantes eran las horas trabajadas, edad, educación y género. Del mismo modo, Sohnesen y Stender (2016) comparan la predicción de pobreza del método de imputación múltiple con RF, encontrando que a nivel rural y urbano RF es más preciso para identificar la pobreza. A diferencia de ellos, en este trabajo se busca clasificar los hogares, a partir de las

características del hogar y de las características del individuo que cumple el rol de principal sostén del hogar.

McBride y Nichols (2015); Otok y Seftiana (2014) se interesan por identificar a los pobres que son el foco de diversas políticas públicas. Los primeros buscan clasificar a los hogares pobres y *crónicamente* pobres a partir del tipo de asistencia social esperada por los hogares en Indonesia.[1] Los segundos se interesan en mejorar la predicción del Test de Medias Próximas (*proxy mean test*) para encontrar los beneficiarios de las políticas públicas. Ambos trabajos emplean los métodos de regresión y clasificación de árboles (CART), RF y Regression Forest por cuantiles para mejorar la predicción. Para evaluar el desempeño, se utilizan distintas medidas de precisión como la tasa de precisión total, de pobreza, de falsos positivos y falsos negativos; medidas que también se utilizan en el presente trabajo. Además, encuentran que las variables más importantes para clasificar los hogares son: el combustible que utilizan para cocinar, el ingreso mensual, el tipo de baño, el tipo de vivienda y el origen del agua que dispone el hogar.

Para reducir el problema de la multidimensionalidad en el análisis de bienestar de los individuos, el trabajo de Caruso y cols. (2015) utiliza el análisis de clusters para clasificar el estatus de pobreza con la encuesta de Gallup. Luego de dicha identificación, seleccionan el conjunto de atributos más pequeños posible capaz de reproducir la clasificación, para definir la dimensión reducida del espacio de bienestar original.[2] Las tres variables (de 15 del espacio original) más importantes son: el ingreso mensual, no haber tenido suficiente dinero para comprar comida tres veces en un año y tener o no una computadora. Este trabajo, enfrenta la desventaja de clasificar los hogares pobres con una sola dimensión, donde los hogares cercanos a la línea de pobreza son indistinguibles entre ellos induciendo una identificación de pobreza que no satisfacen las diferencias requeridas entre ambos grupos. Sin embargo, el propósito es mejorar la predicción de los hogares pobres utilizando características cualitativas no monetarias. Mas aún, en un ejercicio de robustez de la presente investigación, se busca distinguir entre una clasificación de pobres con la línea de pobreza oficial brindada por el INDEC y la misma ajustada por inflación, mostrando sus diferencias antes y después de la intervención del año 2006.

El trabajo está organizado de la siguiente manera. La sección 2 introduce el algoritmo Random Forest. La sección 3 desarrolla en detalle la definición de pobreza y datos utilizados. La sección 4 muestra los resultados preliminares para predecir la pobreza, abordando algunos potenciales problemas. Por último, la sección 5 incluye ejercicios de robustez que aportan a la discusión final.

2. El algoritmo: Random Forest

Para entender el funcionamiento del algoritmo resulta útil comenzar desde la unidad más sencilla del sistema formado por árboles. Si la variable dependiente toma un conjunto finito de valores, se denominan árboles de clasificación (sus hojas son niveles de una categoría). Si es continua, árboles de regresión (sus hojas representan valores de una variable continua). Breiman, Friedman, Stone, y Olshen (1984) introducen un análisis detallado y popularizan el término CART acrónimo para *Classification and Regression Trees*. La idea general es dividir reiteradamente los datos en base a la variable de respuesta de manera que los integrantes de cada grupo sean lo más homogéneos posible en términos de los predictores. Las particiones se

[1] En Otok y Seftiana (2014), el gobierno utiliza la línea de pobreza para identificar a los hogares a los cuales debe brindar un conjunto de programas de asistencia social en áreas de salud, educación, economía y laboral.

[2] Implementan el enfoque *blinding* de Fraiman, Justel, y Svarc (2012) de selección de variables para el análisis de clusters.

realizan utilizando el predictor que mejor particiona los datos (hay distintas formas de definir "mejor").

La decisión sobre la variable de partición elegida se realiza en base al criterio de *impureza*. Por ejemplo, si una variable es constante, la impureza es cero. La impureza se incrementa con la variabilidad de cada predictor.[3] Estrictamente hablando, maximizar la homogeneidad es lo mismo que minimizar la impureza. Para los dos tipos de árboles, la impureza se obtiene de forma diferente: para clasificación se utiliza habitualmente el índice de Gini, mientras que para regresión la impureza se mide por la suma de los residuos al cuadrado para cada partición.[4]

Si el objetivo es la predicción, CART presenta un problema central: el *overfitting*. Aún utilizando *pruning*,[5] CART no predice bien. Desde CART, el siguiente paso hacia Random Forests es *bagging* Breiman (1996). La idea es eliminar el overfitting de CART utilizando técnicas de remuestreo aleatorio para construir K árboles (desde las K muestras obtenidas), y de alguna forma promediar los votos de clasificación que produce cada árbol. Como resultado adicional, el algoritmo reduce la varianza de la predicción. La rutina tiene los siguientes pasos:

1. Tomar una muestra con reemplazo de tamaño N (utilizando *bootstrapping*).
2. Construir un árbol.
3. Guardar el árbol y asignar una clase a cada observación. En árboles de clasificación se utiliza el voto mayoritario, en árboles de regresión se utiliza la media aritmética de los K promedios de cada árbol, también con criterio de mayoría.
4. Repetir pasos 1 a 3, K veces.

Pero si el objetivo es disminuir la varianza de la predicción y medir la importancia de cada predictor, este algoritmo presenta un inconveniente: al seleccionar las variables que mejoran más la predicción produce árboles que son muy parecidos entre sí en la estructura, árboles muy correlacionados. Este es el primer punto fuerte de RF. El algoritmo que presenta Breiman (2001) utiliza la selección aleatoria de predictores en cada nodo, para cada árbol. De este modo, se introduce aleatoriedad en cada partición de cada árbol (que a su vez proviene de una muestra aleatoria).[6] Para un número de árboles T, el RF tiene el siguiente algoritmo:

1. Tomar una muestra con reemplazo de tamaño N (utilizando bootstrapping).
2. Construir un árbol, en cada nodo:
   a) Se seleccionan aleatoriemente m predictores del conjunto de predictores.[7]
   b) Se realiza una partición binaria utilizando el mejor predictor.

[3] Esta característica genera un sesgo en el algoritmo de Random Forest, ver sección 5.

[4] El Indice de Gini es más pequeño cuando muchas observaciones pertenecen a una misma categoría, de modo que la partición se realiza en la variable individual que minimiza el indice de Gini. Como en un método tradicional de regresión, se realizan las particiones buscando los predictores y umbrales en la variable de respuesta que minimicen la suma de los residuos en cada una de las regiones resultantes.

[5] Intuitivamente, el proceso tiene como objetivo seleccionar un *subárbol* que dé lugar a una menor tasa de error de prueba. Ver James, Witten, Hastie, y Tibshirani (2013).

[6] Utilizando la Ley de Grandes Números Breiman (2001) muestra que cuando aumenta el número de árboles los RF convergen (demuestra *almost sure convergence*).

[7] Usualmente se utiliza $m \approx \sqrt{p}$ para clasificación y $m \approx \frac{p}{3}$ para regresión, en donde p es el número total de predictores.

3. Guardar el árbol y asignar una clase a cada observació. En árboles de clasificación se utiliza el voto mayoritario, en árboles de regresión se utiliza la media aritmética de los K promedios de cada árbol, también con criterio de mayoría.

4. Repetir pasos 1 a 3, K veces.

El valor de m determina sistemas ligeramente diferentes. Si $m = 1$, el ejercicio será una selección aleatoria de una variable de partición por nodo, si m es el número total de predictores estamos realizando el ejercicio de *bagging* de Breiman; mientras que RF requiere un m estrictamente menor al número total de predictores. Para mejorar la precisión, la aleatoriedad inducida en la selección de predictores minimiza la correlación entre los árboles y fortalece la robustez James y cols. (2013).

Las ventajas del algoritmo RF comparado con otros son:

1. Su precisión es tan buena como Adaboost Rätsch, Onoda, y Müller (2001), un algoritmo muy popular de aprendizaje adaptativo que mejora el boosting de Freund y Schapire (1995).

2. Es más robusto al ruido que Adaboost.

3. Es más eficiente computacionalmente que *bagging* ó *boosting*.

4. Proporciona estimaciones internas de error, robustez, correlación e importancia de variables (basadas en la medida de *out of bag error*).

En este trabajo se utiliza el software libre y gratuito R. A menos que se especifique lo contrario, se utiliza la función `randomForest` que pertenece al paquete homónimo.

3. Definición e identificación de hogares pobres

Adoptando la definición del Instituto Nacional de Estadística y Censos (INDEC) que mide la pobreza en el Gran Buenos Aires desde principios de los años noventa utilizando el *enfoque del ingreso* de Estadística y Censos (2003). Se define a un hogar como pobre cuando el ingreso total familiar se encuentra por debajo de la línea de la pobreza. La línea busca representar el costo de los bienes y servicios que se consideran necesarios para que el hogar supla sus necesidades básicas (concepto normativo). Para ello, se computa el valor de la Canasta Básica de Alimentos (CBA) de los productos que satisfacen las necesidades nutricionales y los hábitos de consumo predominantes, valor que representa la línea de indigencia.[8] Luego, se adiciona un factor de expansión (la inversa del coeficiente de Engel) para obtener el ingreso mínimo necesario para cubrir todas la necesidades, lo cual representa la línea de pobreza.

En términos generales, la canasta se construye teniendo en cuenta cantidades mínimas de calorías y nutrientes recomendadas. En Argentina se consideran 2700 calorías como el requerimiento necesario para un hombre adulto entre 30 y 59 años. Los hogares que se encuentran entre el percentil 21 y 40 de la distribución de ingreso per cápita de los hogares de la Encuesta de Ingresos y Gastos de 1985/86, constituyen el grupo de referencia para identificar las pautas de consumo alimentario. El INDEC ajusta la Canasta Básica de Alimentos eliminando los alimentos de costos por calorías alto y reemplazándolos por otros similares y más económicos.

[8] Se utiliza esta segunda medición para el ejercicio de robustez en la clasificación de hogares pobres.

El INDEC toma la lista de precios medios mensuales de 1985 entre julio y octubre, que utiliza para calcular el Índice de Precios al Consumidor (IPC), de modo de establecer el valor de la canasta básica. La canasta se actualiza con el IPC calculado por el INDEC. En este trabajo se emplean la Canasta Básica de Alimentos y la Canasta Básica Total construída por la Fundación de Investigaciones Económicas Latinoamericanas (FIEL) para actualizar correctamente la variación de precios. En las Figuras 1 y 2 se puede notar la significativa diferencia en la medición de las canastas entre INDEC y FIEL. Se espera que las distintas mediciones impacten directamente en el porcentaje de hogares que se consideran pobres y no pobres.

Teniendo en cuenta la línea de pobreza ajustada, se identifican a los hogares pobres del Gran Buenos Aires, a partir de la Encuesta Permanente de Hogares (EPH) del INDEC desde el tercer trimestre de 2003 al segundo trimestre de 2015.[9] Luego de unir el registro de la base hogar y la base personas, se selecciona sólo los hogares del Gran Buenos Aires y la Ciudad Autónoma de Buenos Aires. Además, se limpia la base eliminando los casos especiales en el número de hogar y componentes que representan el servicio doméstico y pensionistas. Siguiendo la metodología del INDEC antes mencionada, se usa el ingreso de todos los miembros del hogar para establecer el estatus de pobreza de un hogar si se encuentra por debajo de la línea de pobreza. Para identificar cada hogar se emplea el concepto de *principal sostén del hogar* (PSH) que es el miembro con mayor ingreso total (laboral y no laboral) en el hogar. De esta forma, se obtienen un conjunto de atributos individuales de la persona con mayor contribución a los ingresos del hogar, y se los agrega al conjunto de variables relacionadas al hogar.

Dado que el objetivo de este trabajo es clasificar los hogares pobres a partir de características no monetarias que contiene la EPH, se eliminan todas las variables referidas a los ingresos, fuentes ocasionales de ingresos, limosnas, etc. Además, se simplifican algunas variables relacionadas a la antigüedad de las personas en sus respectivos trabajos.[10]

A partir de las variables relacionadas con el Clasificador Nacional de Ocupaciones del año 2001 (CNO) y la Clasificación de Actividades Económicas para Encuestas Sociodemográficas del MERCOSUR (CAES) se crean distintas variables de Estadística y Censos (2001, 2011). Con el primer dígito del CNO se generan las variables con el carácter ocupacional, jerarquía (segundo dígito), tecnología (tercer dígito) y calificación ocupacional (cuarto dígito) para los ocupados y desocupados. A partir del CIIU revisión 4, la codificación de CAES en la EPH cambió desde del primer trimestre de 2012. Por este motivo, se aplica una correspondencia de CAES 2000 (basada en CIIU revisión 3) y CAES 1.0 a CIIU revisión 3.2. Por ejemplo, la pregunta *pp*11*b_cod* que le solicita a la persona indicar a qué se dedicaba o qué producía en su trabajo anterior a estar desocupado, tiene un nuevo código según la nueva correspondencia.

Para evitar problemas de sesgo, sólo se seleccionó el primer dígito de dicha clasificación, los cuales se corresponden con las 10 secciones más generales del clasificador de actividades económicas CIIU. Dichas variables tienen el problema de falta de observaciones (*missing values*) por dos causas principales: *i)* la no respuesta de los individuos o entrevistas no realizadas, *ii)* no les corresponde responder dicha pregunta a individuos que no se encuentran en edad de trabajar. A pesar de que el primer problema no tiene solución, se corrige el segundo problema, reemplazando los missing values con un código de 3 dígitos que haga referencia a la condición de actividad (variable *estado* en la EPH). Siguiendo el ejemplo anterior, la pregunta era puntualmente para los desocupados (estado de actividad igual a 2), por lo que a los hogares con el principal sostén del hogar ocupado se les asigna 111. Dado que a la variable *estado* no le falta

[9] Para más información sobre el diseño de la EPH ver Comari (2010)

[10] Por ejemplo, se traducen las variables que indican años, meses, y días a una sola variable general con el total de días.

ninguna observación reducimos el 32.25 % de missing values de estas variables al 0 %. Intuitivamente, a partir de esta nueva forma de categorizar la actividad económica, el algoritmo peude aprender las distintas ramas de actividad y la no respuesta condicional a no requerir contestar esa pregunta según condición de actividad.

Con respecto a la educación, sólo se utiliza la variable de nivel educativo y se genera una variable que considere si asistió o no a un establecimiento educativo, y si el mismo es público o privado.

4. Resultados principales

Para tener una idea general de la tasa de pobreza, las figuras 3 y 4 muestra la evolución de la tasa de pobreza e indigencia.[11] En particular, se observa que la tendencia decreciente del porcentaje de hogares pobres que estima el INDEC es más marcada que la calculada por FIEL, lo cual era esperable dadas las diferencias en el cómputo de las canastas CBT y CBA de ambas fuentes, presentadas en la sección 3.

4.1. Desbalanceo muestral

Por construcción, el ejercicio de clasificación que se intenta realizar está desbalanceado, i.e. la clase que se quiere predecir correctamente constituye sólo una minoría en los datos (en toda la muestra de la EPH, sólo un 15.64 % de la muestra tiene el estatus de *pobre*). Esto es un gran problema a la hora de intentar evaluar cuán bueno es un predictor. Para ilustrarlo, supongamos que solamente el 5 % de la muestra tiene el estatus de pobre. No es difícil darse cuenta que cualquier algoritmo que clasifique a toda la muestra como no-pobre tiene un error de predicción del 5 %, lo cual es demasiado bueno respecto de los resultados disponibles en la literatura. En este trabajo, ese algoritmo tendría un error de predicción del 15.64 %. Afortunadamente existen soluciones viables frente a este inconveniente.

Breiman, Chen, y Liaw (2004) sugieren dos opciones para lidiar con el problema de datos desbalanceados. La primera consiste en utilizar ponderadores para penalizar los errores de clasificación: asignando un costo de error de clasificación tan alto como uno quiera a la clase de interés, y luego minimizar el costo total (*Weighted Random Forest* (WRF)).[12] La segunda opción es utilizar técnicas de muestreo para balancear las categorías. Recordemos que como el algoritmo Random Forest toma una muestra aleatoria para cada árbol, lo único que hay que hacer es agregar un paso que indique qué proporciones se deben respetar en la muestra (sub-muestrando la clase mayoritaria, sobre-muestreando la clase minoritaria, o haciendo un mix entre ambas para crear un *Balanced Random Forest* (BRF)).[13] Los autores concluyen que aunque existen mejoras sustanciales en el error de clasificación con ambos métodos, no queda claro que haya un ganador absoluto. En este ejercicio se utiliza la segunda opción: combinando algoritmos de muestreo y conjuntos de modelos, sub-muestreamos la clase mayoritaria para que

[11] Estimaciones realizadas con la base de hogares de la EPH antes de unir la base de Individuos como se describió en la sección 3 y limpiando los hogares repetidos en un año. Ver diseño muestral en de Estadística y Censos (2003).

[12] Ejemplos de algoritmos que hacen esto se encuentran en Domingos (1999) y Pazzani y cols. (1994).

[13] Es la técnica más utilizada en la literatura y por eso hay mayores avances en esta dirección. Ejemplos de este tipo de algoritmos son SHRINK Kubat, Holte, y Matwin (1997), muestreo unilateral para reducir selectivamente la clase mayoritaria en Kubat y Matwin (1997) o para aumentar la proporción de la clase minoriaria en Ling y Li (1998), y lo más reciente, combinaciones de sub y sobre-muestreo para crear una minoría sintética (SMOTE en Chawla, Bowyer, Hall, y Kegelmeyer (2002) y SMOTEBoost en Chawla, Lazarevic, Hall, y Bowyer (2003))

cada árbol crezca con datos más balanceados. Los motivos de esta elección son dos, en primer lugar, BRF es computacionalmente más eficiente que WRF, y en segundo lugar, WRF es más vulnerable al ruido que BRF.

4.2. Criterios de comparación

Para evaluar el desempeño del Random Forest se usan los criterios de precisión convencionales a partir de la matriz de confusión. La misma tiene en cuenta los distintos casos posibles: *i)* predecir un hogar que verdaderamente es pobre como pobre (TP), *ii)* clasificar un hogar como pobre cuando no es pobre (FP), *iii)* predecir un hogar no pobre cuando verdaderamente no es pobre (TN), y *iv)* clasificar a un hogar como no pobre cuando no es pobre (FN). Según el propósito y contexto, el nombre de los criterios utilizados para evaluar la precisión de los modelos varían en la literatura. Siguiendo a James y cols. (2013); McBride y Nichols (2015), se usan las siguientes medidas de clasificación y testeo de diagnóstico:

- Tasa de error OOB $= \dfrac{FP + FN}{TP + FP + TN + FN}$
- Predicción de pobreza (o *recall*) $= \dfrac{TP}{TP + FN}$
- Predicción no pobreza $= \dfrac{TN}{TN + FP}$
- Error tipo I $= \dfrac{FP}{TN + FP}$
- Error tipo II $= \dfrac{FN}{TP + FN}$
- Precisión $= \dfrac{TP}{TP + FN}$
- Medida F $= \dfrac{2 \times Precision \times Recall}{Precision + Recall}$

La tasa de error OOB da información sobre el porcentaje de toda la muestra que no se predice bien, como muestra la tabla 1, en términos generales, la predicción sin sub-muestreo tiene menor error general para predecir hogares pobres y no pobres, mientras que el error de clasificación de hogares indigentes es del 12 % de la muestra. Sin embargo, al mirar la predicción de los hogares de cada clase (tasa de verdaderos positivos y verdaderos negativos), la predicción con sub-muestreo mejora la predicción.

El error tipo I da la intuición del ratio de hogares pobres predichos equivocadamente sobre el total de hogares no pobres (*tasa de falsos positivos*), mientras que el error tipo II indica el porcentaje de hogares incorrectamente clasificados como no pobres sobre el total de hogares pobres (*tasa de falsos negativos*). La tabla 1 muestra que el error de tipo II de predecir los hogares como no pobres cae del 43 % al 19 % cuando se balancea la muestra. Cuando se mejora la clasificación de los hogares pobres con sub-muestreo, la precisión aumenta de 56 % a 81 %, mientas que el error de tipo I pasa de 4 % a 15 %. De este modo, la técnica de un muestreo que pondere la categoría minoritaria (hogar pobre) y la mayoritaria (hogares no pobres), está acompañada de un *trade off* entre la precisión de predicción de una categoría versus la otra. Sin emabrgo, como el objetivo es identificar los hogares en condición de pobreza, BRF mejora

la clasificación con Random Forest de los hogares que se encuentran por debajo de la línea de pobreza entre 2003 y 2015.

La medida de precisión da una idea sobre el porcentaje de hogares pobres correctamente clasificados con respecto al total de hogares pobres predichos. En otras palabras difiere de la predicción de pobreza antes mencionadas en el denominador: la predicción de pobreza se basa en los hogares pobres observados, mientras que la precisión se basa en los hogares pobres predichos. Al igual que en los casos anteriores la precisión cae cuando se considera el remuestreo ponderado y es más baja para la indigencia. La medida F es un promedio ponderado de la predicción de pobreza (conocido como sensibilidad o recall) y la precisión, mostrando mayor balance entre dichas medidas cuando se aplican a la indigencia que al modelo de predicción de pobreza con y sin muestreo.

Para encontrar las variables más importantes para predecir la condición de pobreza de un hogar, se utilizan dos medidas:

- *Mean Decrease Accuracy*: $1 - \text{máx}_k(\hat{p_{mk}})$ donde k es la categoría que toma la variable de respuesta, m es la partición a la que corresponde la observación y $\hat{p_{mk}}$ es la proporción de observaciones de entrenamiento en la m-ésima región de la clase k-ésima.
- *Mean Decrease Gini* (G): $\sum_{k=1}^{K} \hat{p_{mk}}(1 - \hat{p_{mk}})$ donde K es el número total de clases (en este caso, hay 2 clases: pobre toma el valor 1 y no pobre toma el valor 0).

La primera medida se computa a partir de la permutación de los datos en *out-of-bag* (OOB).[14] Intuitivamente, mean decrease accuracy muestra qué tanto disminuye la media de precisión cuando se elimina una variable, mientras que mean decrease Gini mide la importancia de la variable basada en la impureza del índice de Gini usado en las particiones de los árboles (mide la varianza total entre las clases de la variable de respuesta).

La figura 5 indica que entre todos los árboles considerados en el random forest para ambas medidas de importancia mencionadas, la cantidad de miembros del hogar (IX_Tot) y la edad son la primera y tercera variables más importantes. Para la medida de disminución de la precisión, el régimen de tenencia de propiedad del hogar ($ii7$) es la segunda más importante, mientras que para la medida de Gini, lo cobertura médica ($ch08$) ocupa dicho lugar.

Cuando se tiene en cuenta el desbalanceo que presenta la muestra y se corrige por esto, los resultados presentados en la figura 6, muestran que el número de miembros del hogar y la edad continúan siendo relevantes para la medida de precisión, incorporándose el número de ambientes que tiene el hogar ($ii1$) como variable relevante. Para la medida de Gini, sólo cambia el orden de las tres variables más importantes, siendo la cobertura médica la de mayor relevancia en la predicción. Cuando se considera la clasificación de los hogares indigentes corrigiendo por desbalanceo de la muestra como en el gráfico 7, se incorporan como variables relevantes, la clasificación de actividades económicas del trabajador ocupado (cod_ciiu_1d) y la categoría de inactividad del principal sostén del hogar (cat_inac).

En las Figuras 9 y 10 se presenta la dependencia parcial de cada una de las categorías de los 6 principales predictores obtenidos para RF con sub-muestreo, según los dos criterios de importancia, sobre la clase "no pobre". Las preguntas que hacen referencia a esas variables se encuetran en Tablas 2 y 3. Los valores de la dependencia parcial no pueden interpretarse en términos absolutos sino sólo relativos, es decir, entre categorías. La misma mide el efecto parcial

[14] *Out-of-bag* es el tercio de observaciones que no se usan en cada subconjunto de observaciones bootstrapeados por el algoritmo.

de una variable sobre la predicción del modelo, controlando por la influencia media del resto de las variables que forman parte del mismo.

Se puede apreciar que la predicción con éxito de un hogar "no pobre" es menos probable cuando: la variable que refiere a la cantidad de miembros del hogar (IX_Tot) es inferior a 4, la variable edad es menor a 10 años, cuando la variable cobertura médica ($ch08$) indica que el principal sostén del hogar no posee cobertura médica y cuando la variable cat_inac que refiere a la categoría de inactividad indica que el principal sostén del hogar es estudiante, ama de casa o discapacitado.

5. Discusión y ejercicios adicionales

5.1. Pobreza: medición oficial y alternativa

Si bien este trabajo busca identificar características de hogares e integrantes del mismo para predecir pobreza sin tener en cuenta variables monetarias, está latente preguntarse si durante el período de estudio, cambiaron las características estructurales del individuo que la metodología considera como pobre. Por ejemplo, a principios del período una persona jubilada tiene mayor probabilidad de ser pobre que una persona joven al final del período estudiado.

Otro gran cambio que puede identificarse sucede el año 2006 cuando el INDEC es intervenido y se ponen en duda las cifras publicadas de allí en adelante. De hecho, no hay resultados de la EPH disponibles para el primer trimestre del año 2007. De allí en adelante existen mediciones de la inflación oficiales y no–oficales (diversas fuentes, aquí se utiliza la medición de FIEL). Dado que se utiliza la metodología de medición de pobreza por ingresos, las diferentes medidas de inflación se reflejan directamente en diferencias en medidas de la pobreza (a través de la sub–valuación de las CBT y CBA).[15]

Al analizar la importancia de las variables para el período 2003–2006 y usando las CBT y CBA oficiales, se encuentra que la cantidad de miembros en el hogar se mantiene como un fuerte predictor de la pobreza. También son importantes para predecir: la situación de cobertura médica, edad del principal sostén del hogar y una pregunta a trabajadoras domésticas[16] en el período 2003–2006 y 2007–2013 con el IPC oficial (Figuras 12, 13 y 14). En la clasificación con submuestreo de la pobreza, dicha variable ocupaba el séptimo lugar de importancia teniendo en cuenta la disminución en la media de la precisión (Figura 6), y el cuarto lugar cuando se considera la indigencia (Figura 7).

Estos ejercicios sugieren que la especificación usando la metodología de BRF es la más adecuada obteniendo el menor error de predicción del 19.2 % de los hogares pobres y 15.1 % de los hogares no pobres.

5.2. Árboles y bosques de inferencia condicional

El ejercicio presentado podría presentar una debilidad propia del algoritmo de Random Forest. Strobl, Boulesteix, Zeileis, y Hothorn (2007) demuestran que para variables con diferentes unidades de medida, el algoritmo de Breiman (2006) presenta un sesgo hacia variables continuas y variables con muchas categorías. Strobl y Zeileis (2008) desarrollan una implementación del

[15] Ver las Figuras 1, 2, 3 y 4.

[16] "En la casa en la que más horas tiene: ¿Cuántas personas, incluído usted trabajan allí en total?"($pp04c$)

algoritmo que tiene disponible un test de permutación insesgado para evaluar la importancia de las variables en cada paso.[17]

Por sencillez de interpretabilidad y ventajas computacionales, se realizó un árbol de inferencia condicional, como se muestra en la Figura 11.

En estos ejercicios, muchas de las variables más importantes son variables de muchas categorías. Un ejercicio adicional de robustez para realizar en próximas versiones es agrupar a esas variables en menos categorías. Otra posible especificación puede ser trabajar sólo con variables binarias seleccionadas o agrupadas por algún criterio. Strobl y cols. (2007) es un gran alerta para la interpretación de la importancia de las variables, de manera que hay que tratar este punto con especial atención en el futuro.

6. Conclusión y futura investigación

El presente trabajo aplica uno de los métodos de statistical learning para clasificar hogares pobres en Argentina. El aprendizaje de *Random Forest* como clasificador cuenta con la ventaja de no producir *overfitting*, dado que el error de generalización converge a una constante bajo la Ley de Grandes Números. Además de contar con ventajas computacionales con respecto a otros algoritmos (como los bosques con árboles de inferencia condicional de la sección anterior), genera la descorrelación de los predictores (una ventaja con respecto a *bagging* para este caso).

Utilizando los datos de la Encuesta Permanente de Hogares, se siguió la metodología del INDEC para establecer el estatus de pobreza de los hogares en el Gran Buenos Aires entre 2003 y 2015. El principal resultado es que este algoritmo permite identificar a los hogares pobres con un error de predicción del 19.2 % y con un error del 15.1 % a los hogares no pobres con la especificación más adecuada del modelo, donde se implementan correcciones por el desbalanceo muestral de la variable de respuesta elegida (condición de pobreza). Teniendo en cuenta distintas mediciones, se encuentra que las variables más importantes para predecir la condición de pobreza de un hogar son: la cantidad de miembros del hogar, edad y cobertura médica del principal sostén del hogar. Se incorpotan distintas metodologías de robustez y ejercicios que ayudan a entender mejor la identificación de los hogares pobres.

El grado de precisión del algoritmo RF es superior a las regresiones de la econometría clásica, ya que permite tener en cuenta no linealidades entre las variables predictoras. De esta manera, se vuelve interesante poder utilizar el algoritmo de RF para predecir la condición de pobreza de un hogar, lo cual, serviría a su vez, para determinar si un hogar debería recibir o no un subsidio del Estado.

Finalmente, esta metodología podría ser usada para determinar la condición de pobreza de un hogar en encuestas donde no se realizan preguntas sobre el ingreso de las personas o donde la calidad de los datos de ingresos es muy mala. Otra posible aplicación de este algoritmo es brindar un puente entre las clásicas medidas monetarias de la pobreza y las nuevas mediciones de la pobreza en términos multidimensionales.

[17] En el software R, la función para implementar estos bosques es `cforest`, disponible en el paquete llamado `party`.

Referencias

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L. (2006). *randomforest: Breiman and cutler's random forests for classification and regression.* URL https://www.stat.berkeley.edu/ breiman/RandomForests/, R package version.

Breiman, L., Chen, C., y Liaw, A. (2004). Using random forest to learn imbalanced data. *Working paper*.

Breiman, L., Friedman, J., Stone, C. J., y Olshen, R. A. (1984). *Classification and regression trees.* CRC press.

Caruso, G., Sosa-Escudero, W., y Svarc, M. (2015). Deprivation and the dimensionality of welfare: a variable-selection cluster-analysis approach. *Review of Income and Wealth*, *61*(4), 702–722.

Cavallo, A. (2013). Online and official price indexes: measuring argentina's inflation. *Journal of Monetary Economics*, *60*(2), 152–165.

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321–357.

Chawla, N. V., Lazarevic, A., Hall, L. O., y Bowyer, K. W. (2003). Smoteboost: Improving prediction of the minority class in boosting. En *European conference on principles of data mining and knowledge discovery* (pp. 107–119).

Comari, C. (2010). Ponderación de la muestra y tratamiento de valores faltantes en las variables de ingreso en la encuesta permanente de hogares. *Instituto Nacional de Estadísticas y Censos*.

de Estadística y Censos, I. N. (2003). *Acerca del método utilizado para la medición de la pobreza en argentina.* INDEC Buenos Aires.

de Estadística y Censos, I. N. (2001). *Encuesta permanente de hogares. clasificador nacional de ocupaciones.* INDEC.

de Estadística y Censos, I. N. (2011, Octubre). *Clasificación de actividades económicas para encuestas sociodemográficas del mercosur. caes - mercosur 1.0 versión argentina.* INDEC.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. En *Proceedings of the fifth acm sigkdd international conference on knowledge discovery and data mining* (pp. 155–164).

Einav, L., y Levin, J. D. (2013). The data revolution and economic analysis. *National Bureau of Economic Research*.

Fraiman, R., Justel, A., y Svarc, M. (2012). Selection of variables for cluster analysis and classification rules. *Journal of the American Statistical Association*.

Freund, Y., y Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. En *European conference on computational learning theory* (pp. 23–37).

Guillermo Cruces, D. J. y. M. V., Gary S. Fields. (2016, Julio). *The growth-employment-poverty nexus in latin america in the 2000s: cross-country analysis.* CEDLAS.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). Springer.

Keely, L. C., y Tan, C. M. (2008). Understanding preferences for income redistribution. *Journal of Public Economics*, *92*(5), 944–961.

Kubat, M., Holte, R., y Matwin, S. (1997). Learning when negative examples abound. En *European conference on machine learning* (pp. 146–153).

Kubat, M., y Matwin, S. (1997). Addressing the curse of imbalanced training sets: one-sided

selection. En *Icml* (Vol. 97, pp. 179–186).

Ling, C. X., y Li, C. (1998). Data mining for direct marketing: Problems and solutions. En *Kdd* (Vol. 98, pp. 73–79).

McBride, L., y Nichols, A. (2015). Improved poverty targeting through machine learning: An application to the usaid poverty assessment tools. *econthatmatters. com/wp-content/uploads/2015/01/improvedtargeting_21jan2015. pdf, retrieved*, *4*.

Otok, B. W., y Seftiana, D. (2014). The classification of poor households in jombang with random forest classification and regression trees (rf-cart) approach as the solution in achieving the 2015 indonesian mdgs' targets. *International Journal of Science and Research (IJSR) Volume*, *3*.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T., y Brunk, C. (1994). Reducing misclassification costs. En *Proceedings of the eleventh international conference on machine learning* (pp. 217–225).

Rätsch, G., Onoda, T., y Müller, K.-R. (2001). Soft margins for adaboost. *Machine learning*, *42*(3), 287–320.

Sohnesen, T. P., y Stender, N. (2016). Is random forest a superior methodology for predicting poverty? an empirical assessment. *An Empirical Assessment (March 18, 2016). World Bank Policy Research Working Paper*(7612).

Strobl, C., Boulesteix, A.-L., Zeileis, A., y Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 1.

Strobl, C., y Zeileis, A. (2008). Danger: high power!–exploring the statistical properties of a test for random forest variable importance. *P. Brito (ed.) COMPSTAT 2008 – Proceedings in Computational Statistics*.

Thoplan, R. (2014). Random forests for poverty classification. *International Journal of Sciences: Basic and Applied Research (IJSBAR), North America*, *17*.

Apéndice A. Tablas

Cuadro 1: Medidas de precisión de Random Forest

	Pobreza		Indigencia
	Sin sub-muestreo	Con sub-muestreo	Con sub-muestreo
Tasa de error OOB	10.57	15.78	11.62
Predicción de pobreza	0.562	0.808	0.840
Predicción de no pobreza	0.960	0.849	0.887
Error tipo I	0.040	0.151	0.113
Error tipo II	0.438	0.192	0.160
Precisión	0.734	0.513	0.344
Medida F	0.636	0.628	0.488

Cuadro 2: Dependencia parcial (Mean Accuracy)

Variable	Definición
IX_Tot	Cantidad de miembros del Hogar
edad	Edad del PSH
ii1	Cuántos ambientes o habitaciones tiene este hogar para su uso exclusivo?
iv2	Cuántos ambientes o habitaciones tiene la vivienda en total?
ii2	De esos, cuántos usan habitualmente para dormir?
ii7	Régimen de tenencia

Cuadro 3: Dependencia parcial (Gini)

Variable	Definición
ch08	¿Tiene algún tipo de cobertura médica por la que paga o le descuentan?
IX_Tot	Cantidad de miembros del Hogar
edad	Edad del PSH
nivel_ed	Nivel educativo
cat_inac	Categoría de inactividad
IX_Men10	Cantidad de miembros del Hogar menores de 10 años

Apéndice B. Gráficos

Figura 1: Canasta Básica Total (CBT)

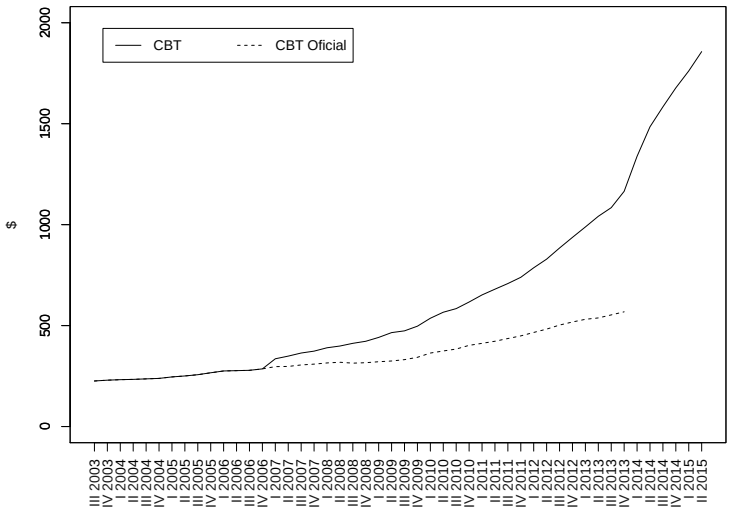

Figura 2: Canasta Básica Alimentaria (CBA)

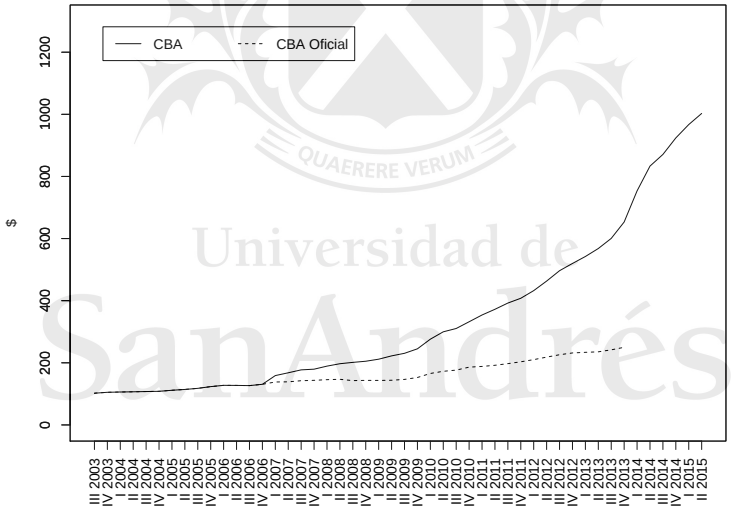

Figura 3: Evolución de la tasa de pobreza

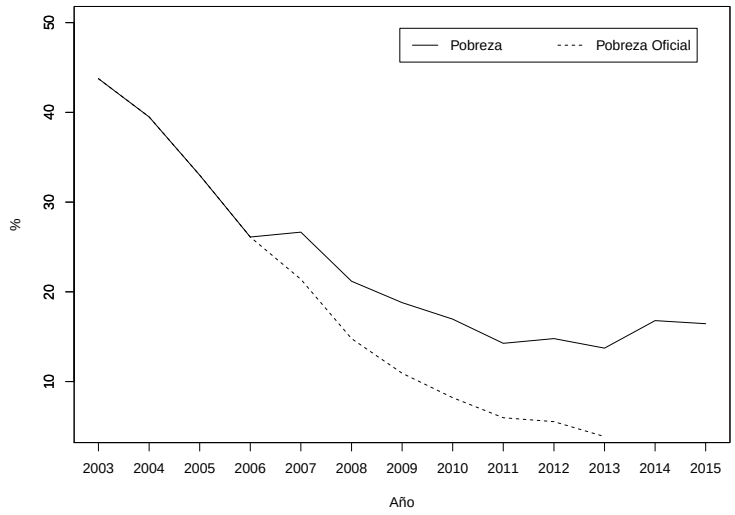

Figura 4: Evolución de la tasa de indigencia

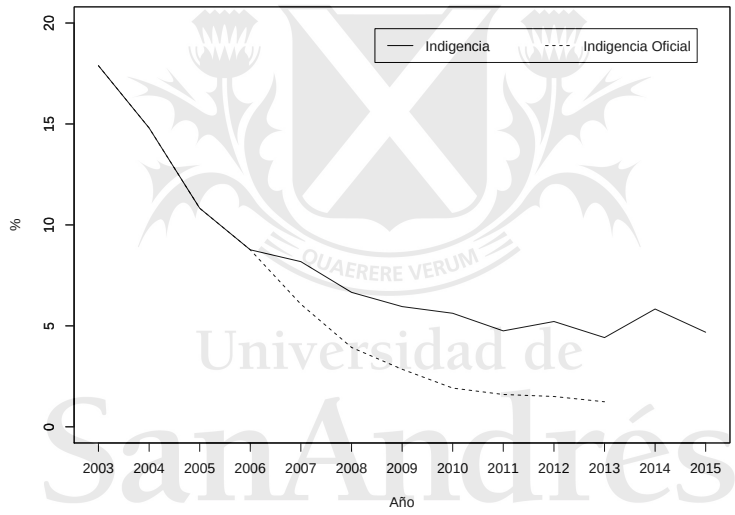

Figura 5: Importancia (2003–2013, sin downsampling)

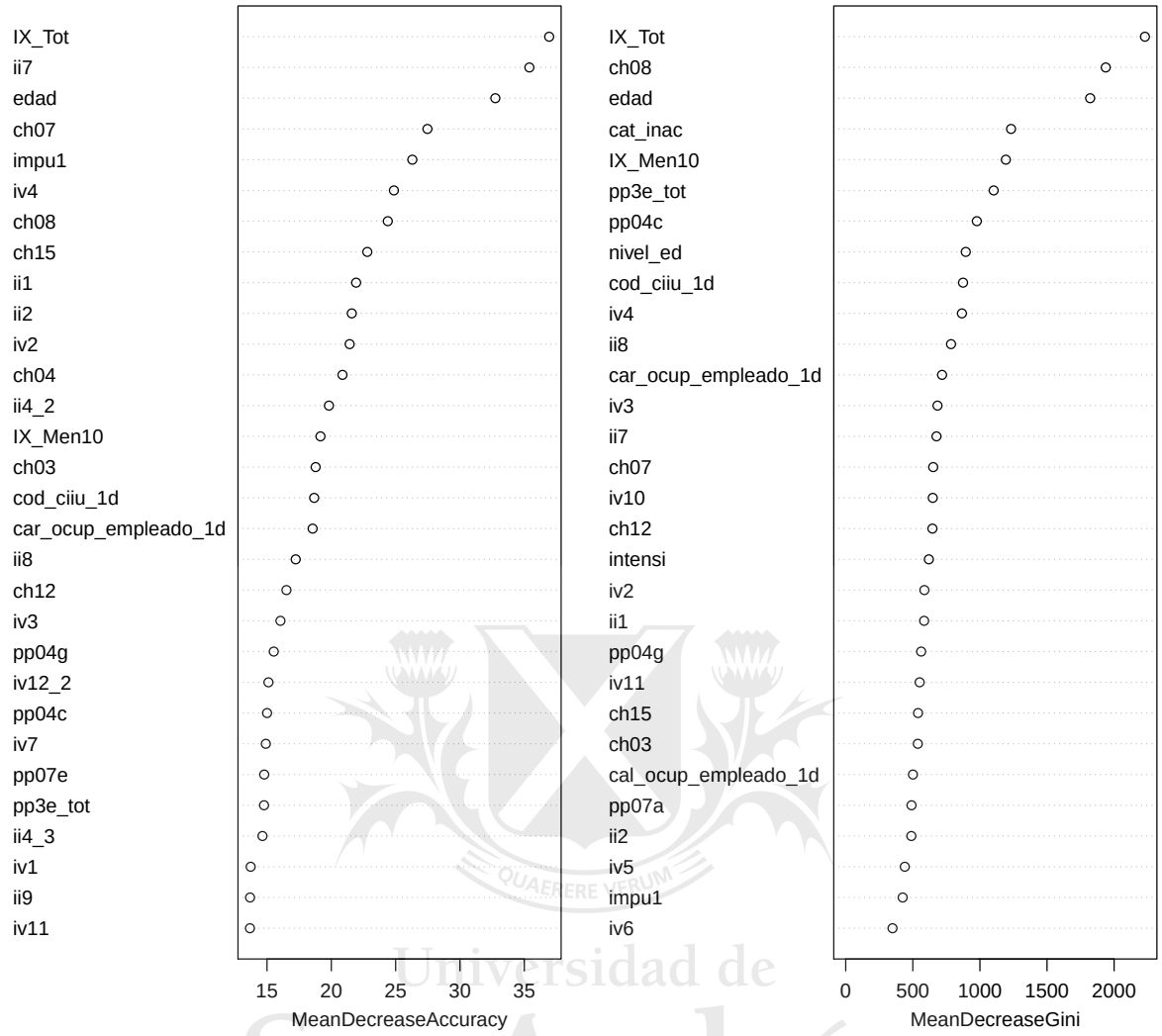

Figura 6: Importancia (2003–2013, con downsampling)

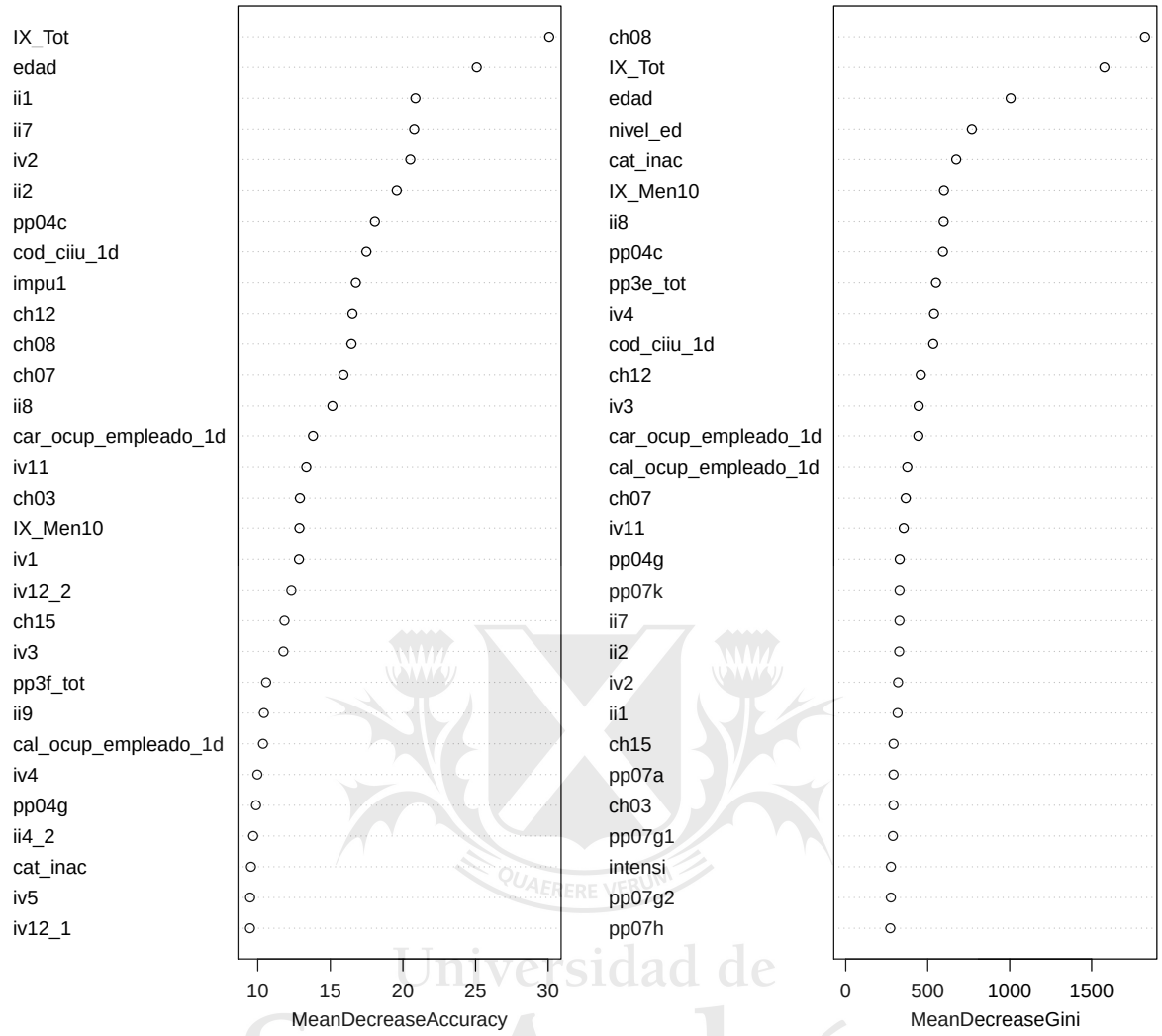

Figura 7: Importancia (2003–2013, para Indigencia con downsampling)

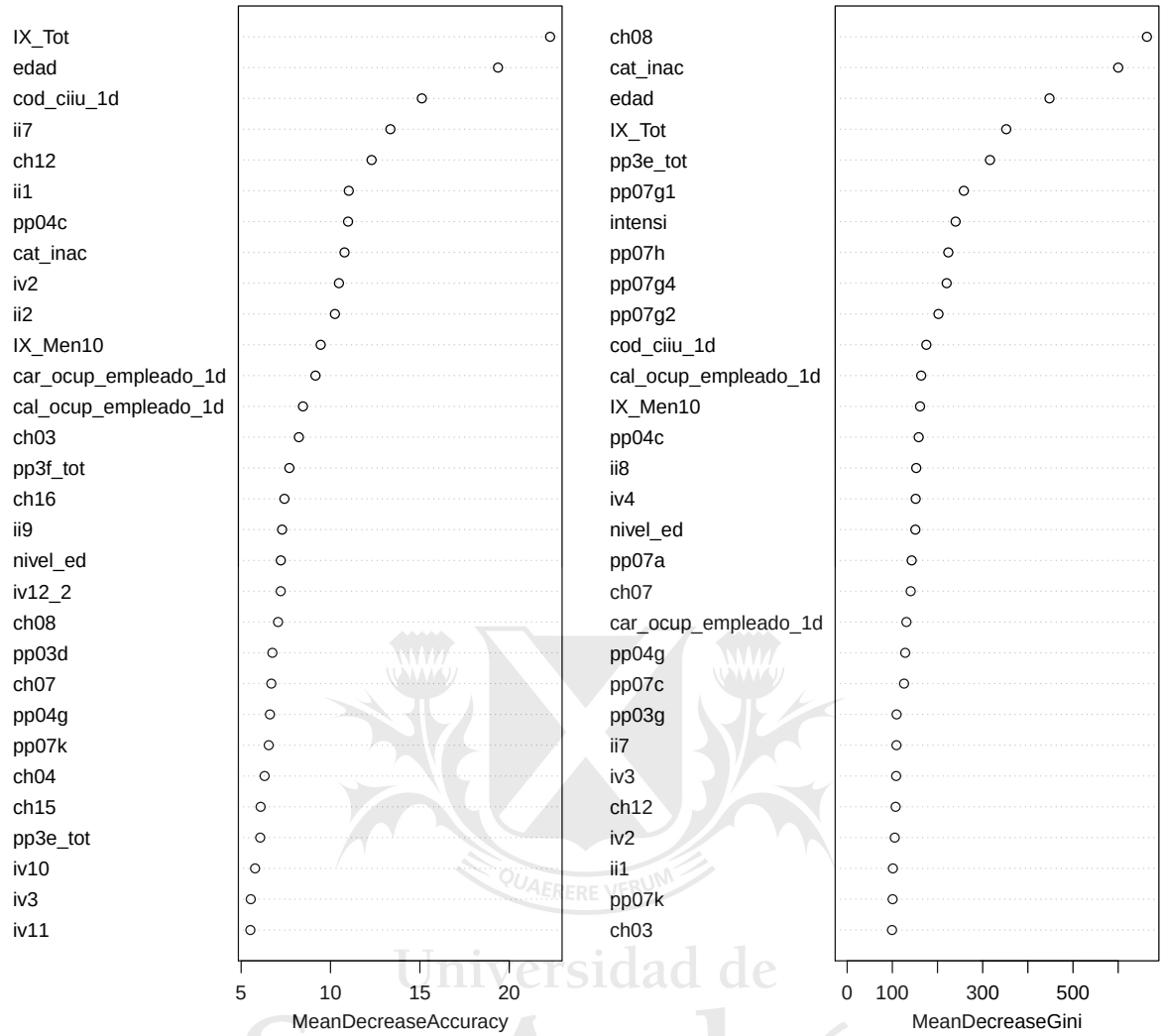

Figura 8: Matriz de correlación

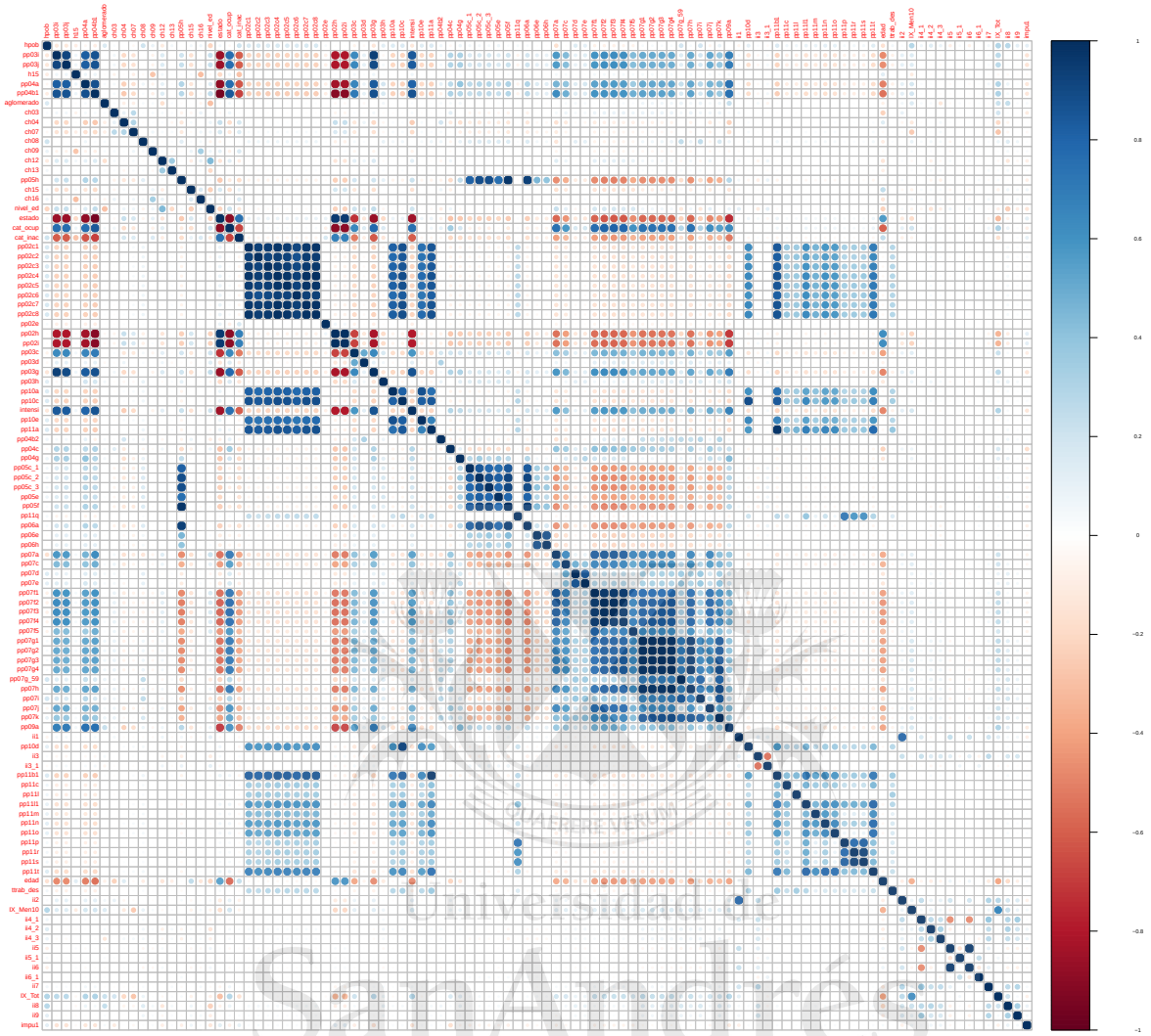

Figura 9: Dependencia parcial (Mean Accuracy)

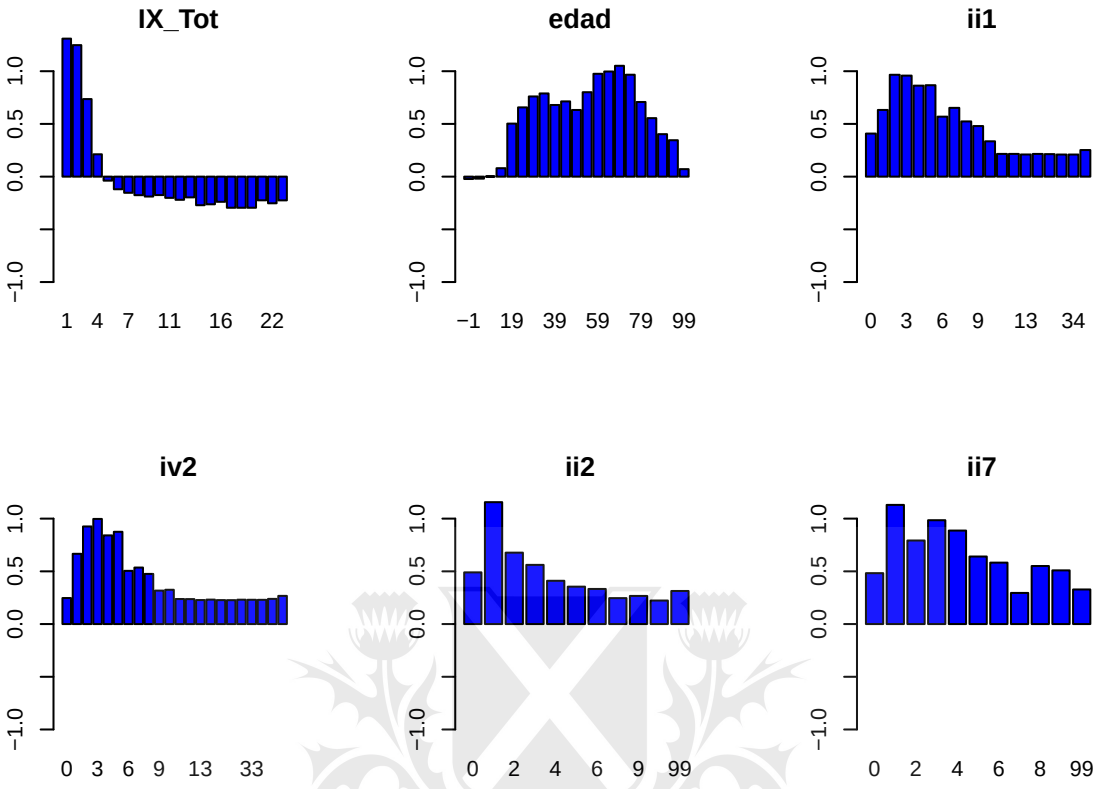

Figura 10: Dependencia parcial (Gini)

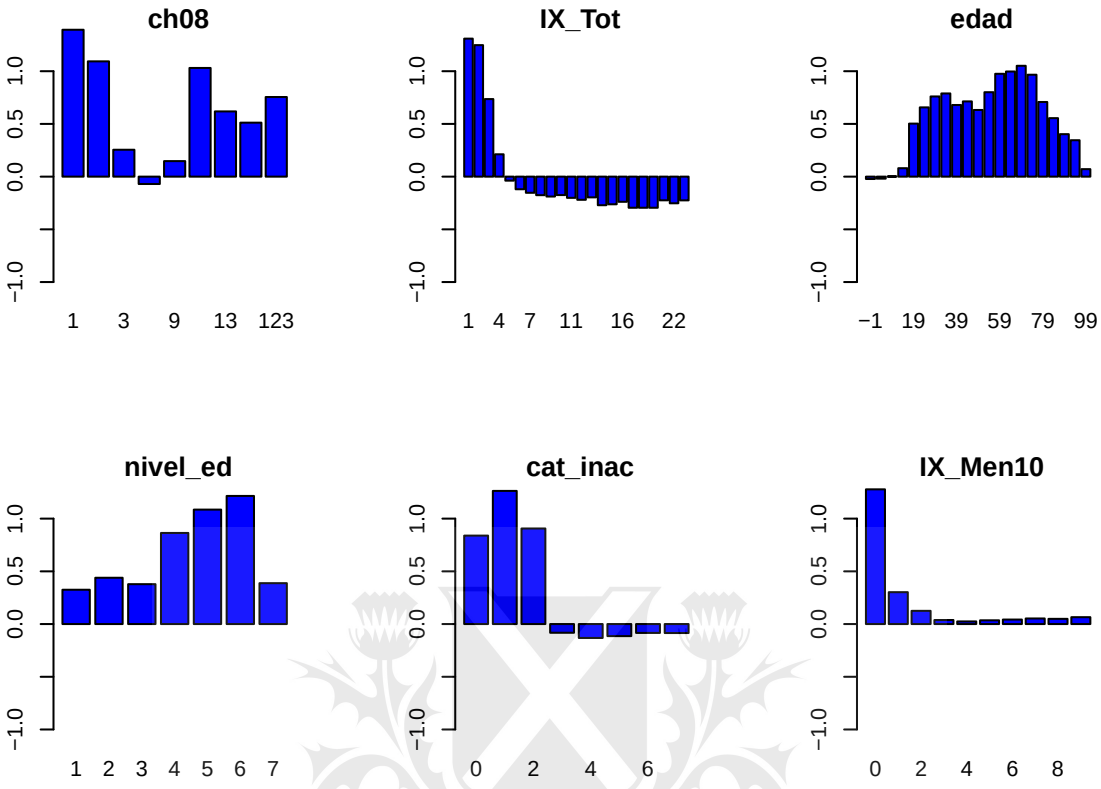

Figura 11: Árbol de inferencia condicional

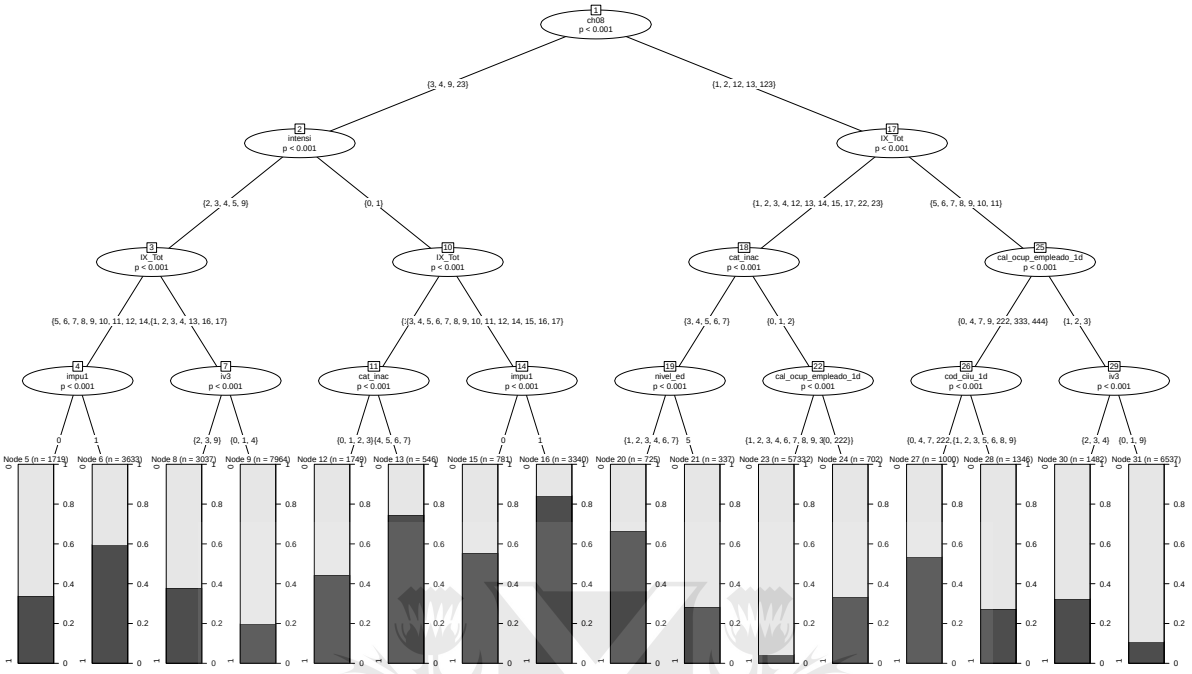

Figura 12: Importancia de variables (Período 2003–2006)

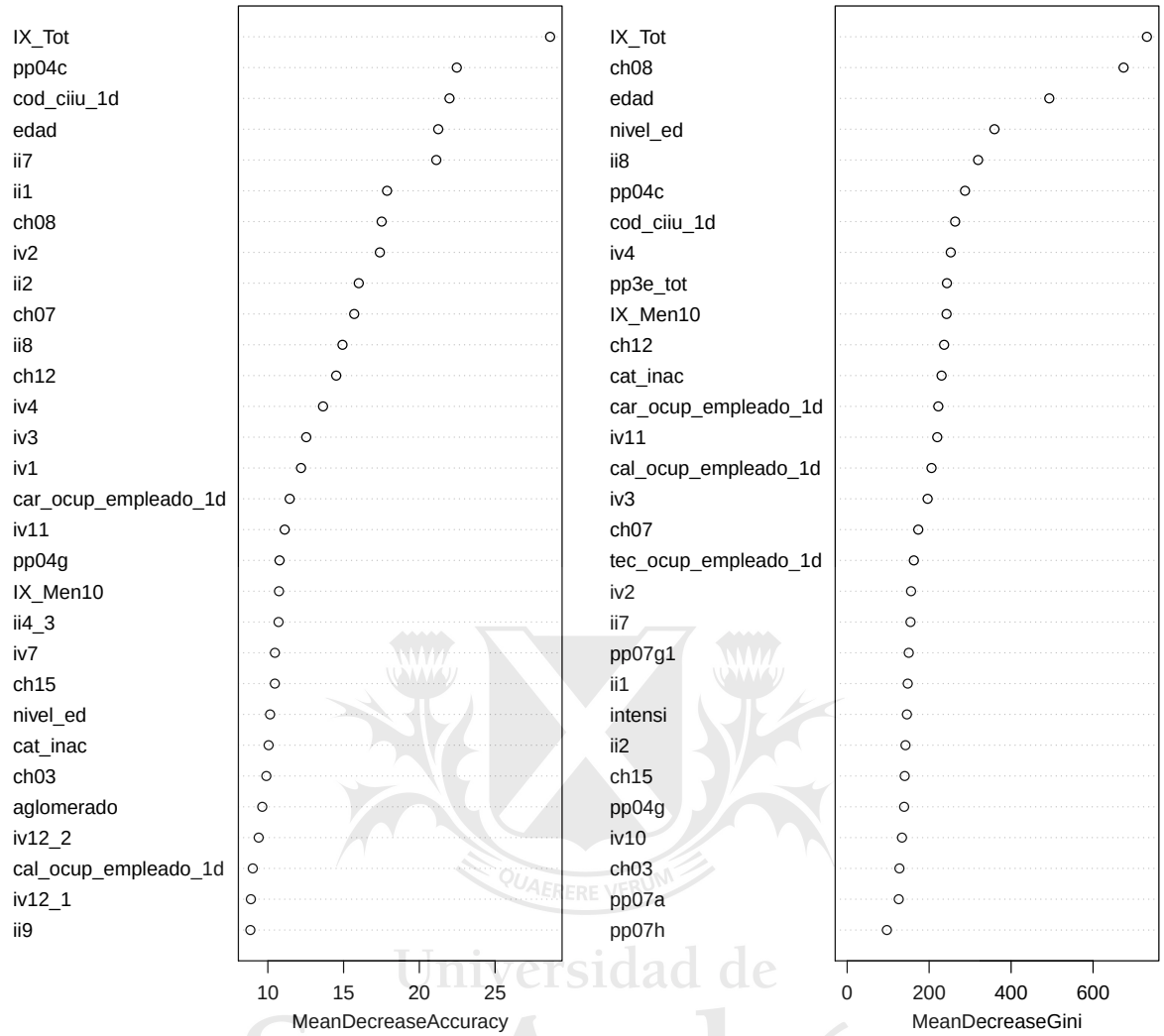

Figura 13: Importancia de variables (Período 2007–2013, IPC oficial)

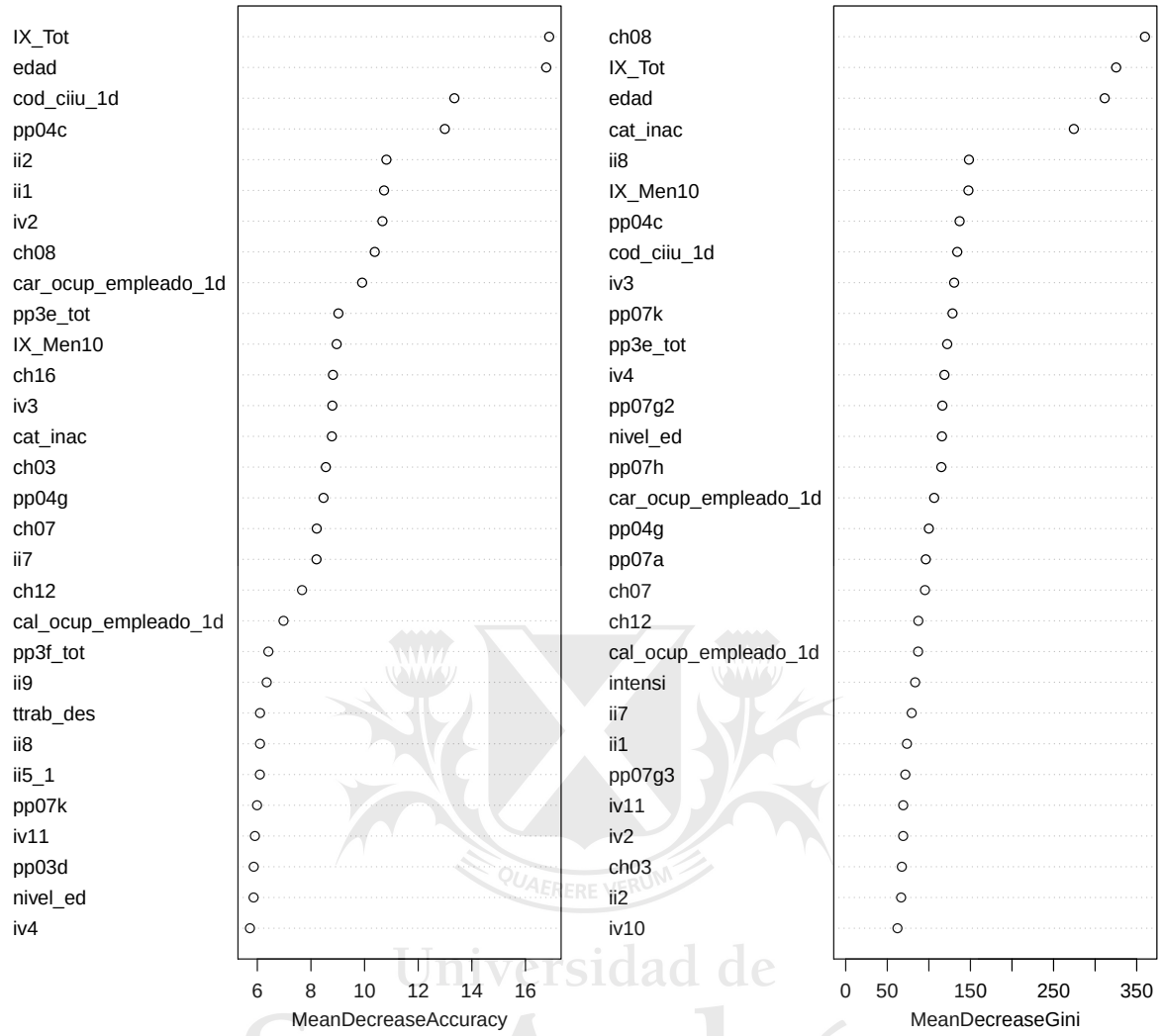

Figura 14: Importancia de variables (Período 2007–2013, IPC FIEL)